Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms

Pedro R. Romero*†, Saima Zaidi*, Ya Yin Fang†, Vladimir N. Uversky†, Predrag Radivojac‡, Christopher J. Oldfield†, Marc S. Cortese†, Megan Sickmeier†, Tanguy LeGall†, Zoran Obradovic§, and A. Keith Dunker*†¶

*School of Informatics, Indiana University–Purdue University Indianapolis, 535 West Michigan Street, IT475, Indianapolis, IN 46202; †Department of Biochemistry and Molecular Biology and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 714 North Senate Avenue, Suite 250, Indianapolis, IN 46202; ‡School of Informatics, Indiana University, Eigenmann Hall 1005, 1900 East 10th Street, Bloomington, IN 47406; and §Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-24), 1805 North Broad Street, Philadelphia, PA 19122

Edited by Richard Henderson, Medical Research Council, Cambridge, United Kingdom, and approved April 14, 2006 (received for review September 12, 2005)

Alternative splicing of pre-mRNA generates two or more protein isoforms from a single gene, thereby contributing to protein diversity. Despite intensive efforts, an understanding of the protein structure–function implications of alternative splicing is still lacking. Intrinsic disorder, which is a lack of equilibrium 3D structure under physiological conditions, may provide this understanding. Intrinsic disorder is a common phenomenon, particularly in multicellular eukaryotes, and is responsible for important protein functions including regulation and signaling. We hypothesize that polypeptide segments affected by alternative splicing are most often intrinsically disordered such that alternative splicing enables functional and regulatory diversity while avoiding structural complications. We analyzed a set of 46 differentially spliced genes encoding experimentally characterized human proteins containing both structured and intrinsically disordered amino acid segments. We show that 81% of 75 alternatively spliced fragments in these proteins were associated with fully (57%) or partially (24%) disordered protein regions. Regions affected by alternative splicing were significantly biased toward encoding disordered residues, with a vanishingly small *P* value. A larger data set composed of 558 SwissProt proteins with known isoforms produced by 1,266 alternatively spliced fragments was characterized by applying the PONDR VSL1 disorder predictor. Results from prediction data are consistent with those obtained from experimental data, further supporting the proposed hypothesis. Associating alternative splicing with protein disorder enables the time- and tissue-specific modulation of protein function needed for cell differentiation and the evolution of multicellular organisms.

evolution | natively unfolded | intrinsically unstructured | protein structure

The splicing of pre-mRNA (1) was first described in 1977. Soon thereafter, Gilbert (2) coined the terms "intron" (intragenic region) and "exon" (expressed region) for the noncoding and coding regions, respectively. Alternative splicing occurs when different mRNAs are assembled from a single gene by joining exons in different ways. Alternative splicing is proposed to generate complexity in multicellular eukaryotes by increasing protein diversity, and thus proteome size, from a relatively small number of genes (3). Estimates indicate that between 35 and 60% of human genes yield protein isoforms by means of alternatively spliced (AS) mRNA (4). Furthermore, complexity in higher organisms is also brought about by signaling and regulatory networks that enable robustness (5). The importance of alternative splicing as a regulatory process (3, 6) has been highlighted by the high occurrence of such splicing in the pre-mRNAs of regulatory and signaling proteins (7).

Alternative splicing can bolster organism complexity, not only by effectively increasing proteome size and regulatory and signaling network complexity, but also by doing so in a time- and tissue-specific manner, supporting cell differentiation, developmental pathways, and other processes associated with multicellular organisms. Indeed, alternative splicing is only prevalent in multicellular eukaryotes (8). This relation suggests that the appearance of alternative splicing was seminal in bringing about the development of multicellular life.

Arguably, splicing within a structured protein domain would have catastrophic effects on the structure of the remaining protein (9), leading to misfolding and aggregation. Fig. 1 depicts a schematic view of this problem: The left-hand side shows the expression path (orange arrows) of a protein of known structure (isoform 1) whose pre-mRNA is subject to alternative splicing. An AS mRNA is also generated as seen in the right-hand side (purple arrows), so that a sizable region (colored red) is missing from the resulting isoform 2. The removal of this "AS region" (region affected by alternative splicing) will no doubt have an important effect on isoform 2's structure. Despite this evident connection, few studies have examined the relationship between alternative splicing sites and protein structure in detail.

Comparing the 3D structures of protein isoforms could provide insight into the mechanisms by which alternative splicing can avoid the catastrophic disruption of protein structure. To date, the 3D structures for only five isoform pairs have been reported (10–14). In the two tumor necrosis factors, EDA-A1 and EDA-A2 (11), alternative splicing leads to the removal of a very small segment containing only 2 aa. In the 216-residue glutathione *S*-transferase isozymes AdGST1-3 and AdGST1-4 from *Anopheles dirus* (10), the entire C-terminal domains (residues 46–216) arise from different exons as a result of alternate splicing, whereas the N-terminal domains (residues 1–45) arise from the same exon. Nevertheless, the two sequences show high similarity where the most striking difference is the insertion of nine extra residues in AdGST1-4 that adds a small helix in the middle of a short loop. In these two pairs of examples, because the affected protein regions are small ordered segments, the folded protein domains are able to undergo slight adjustments to compensate for the deleted or altered segments. The effect in both cases is that of slight structural rearrangements that alter binding specificity, thus modulating function.

The remaining three structurally characterized isoform pairs provide a different picture of alternative splicing. For these three pairs [human pyrophosphorylase AGX1 and AGX2, which differ by removal/insertion of 19 residues (12); human GTPase Rac1

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: AS, alternatively spliced; ASED, Alternative Splicing and Experimental Disorder; ASSP, Alternative Splicing in SwissProt; ASG, Alternative Splicing Gallery.

¶To whom correspondence should be addressed. E-mail: kedunker@iupui.edu.

© 2006 by The National Academy of Sciences of the USA

www.pnas.org/cgi/doi/10.1073/pnas.0507916103

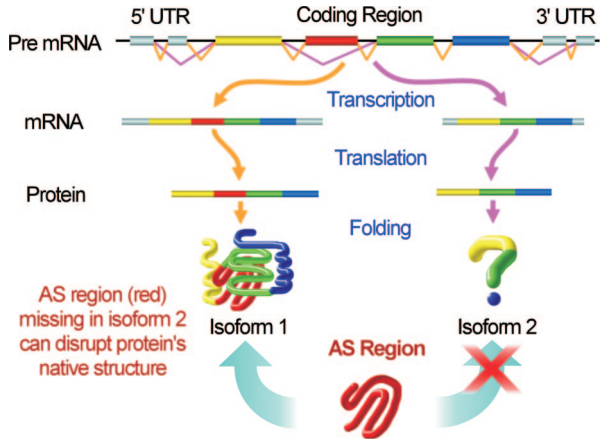

Fig. 1. Alternative splicing and its potential effects on protein structure. During transcription, RNA is generated from the chromosome's genetic material. This piece of RNA forms the basis for the mRNA that will be translated into proteins, hence the designation ''pre-mRNA.'' mRNA is composed of terminal UTRs, which help guide and regulate the translation process, and a central coding region, which is translated into the corresponding protein product. Eukaryotic genes contain fragments that are not used in the translation process, called introns, and regions that will become part of the final mRNA, called exons. Here exons are shown as colored cylinders. The mRNA to be translated is assembled from pre-mRNA by splicing the introns away, so that only the exons remain, as shown in both sides of the image. During alternative splicing, some exons also can be partially or completely spliced away, as can be seen in the path marked by purple arrows to the right of the figure, where the red exon (and two of the UTR exons) are ''skipped,'' generating a shorter mRNA and hence a shorter protein product. The different products of alternative splicing are called isoforms. In this work, the region missing in isoform 2 because of alternative splicing is called an AS region. The generation of isoforms by alternative splicing poses potential problems for the structural stability of globular proteins, because protein folding is a strongly cooperative process. Indeed, the absence of the red AS region in isoform 2 can potentially disturb the protein's 3D structure very strongly, but few systematic studies of this effect have been carried out. (Note: Although we have described this process in terms of removal of protein regions, the same argument can be made about insertion of AS regions, without loss of generality.)

and Rac1b, which differ by removal/insertion of 17 residues (14); and two isoforms of human cholesterol sulfotransferase, which differ by distinct 8- or 23-residue N-terminal sequences (13)], the entire amino acid segments associated with alternative splicing are disordered and missing from the x-ray electron density maps. For these protein examples, the structured regions remain basically unchanged. Thus, in three of the five pairs structurally characterized to date, the region removed or inserted by alternative splicing is intrinsically disordered. In the remainder of this work, we will refer to these segments affected by alternative splicing of the corresponding mRNA as AS regions or segments.

Intrinsically disordered proteins or regions (15) also have been called rheomorphic (16), natively unfolded (17), intrinsically unstructured (18), and various combinations of these terms. Numerous examples of proteins that are totally or partially disordered in physiological buffers have been reported in the literature for decades, and the number of experimentally characterized examples is growing at an ever-increasing pace (19). Interest in intrinsic disorder is increasing in recognition that such regions are commonly responsible for important protein functions (15, 20, 21). In fact, disordered regions were shown to be associated with signaling and regulation (15, 21, 22). It has been emphasized that several characteristics of disordered regions, e.g., decoupled specificity and affinity, binding diversity, binding commonality, the ability to form large interaction surfaces, their fast association and dissociation rates, polymorphism in the bound state, and the reduced lifetimes of disordered proteins in the cell, are ideal for molecules involved in signaling and regulation (23).

The latter three pairs of examples, AGX1/AGX2, SULT2B1a/SULT2B1b, and Rac1/Rac1b, suggest a previously undescribed mechanism for reconciling protein structure with alternative splicing, specifically that AS regions code for segments of intrinsic disorder. This mechanism has two obvious advantages over the alternative, structure-based mechanism. First, structural problems are avoided. Second, signaling and regulation functions can be modulated directly through inclusion or exclusion of recognition regions, without reliance on subtle conformational changes. The second point is supported by observations that correlate both intrinsic disorder (15, 21, 22) and alternative splicing (3) with signaling and regulation functions. The correspondence of AS regions with intrinsically disordered regions provides a straightforward mechanism for developing functional and regulatory diversity. Here we provide evidence that regions of pre-mRNA that are removed by alternative splicing frequently code for regions of intrinsic disorder in the corresponding protein products and that this association leads to functional and regulatory diversity through the various isoforms.

Results and Discussion

Testing the Proposed Linkage Between Alternative Splicing and Intrinsic Disorder. ***Analysis of proteins with experimentally characterized ordered and disordered regions.*** Alternative splicing annotations from both the Alternative Splicing Gallery (ASG) (24) and SwissProt (25) were combined with disordered protein annotations from DisProt (26) to construct the Alternative Splicing and Experimental Disorder (ASED) data set. This data set consists of human proteins that arise from AS pre-mRNA and contain both structured and disordered regions that have been experimentally determined, which allows the measure of the preference, if any, of AS regions to encode for disordered or ordered regions. Because of the scarcity of experimentally characterized disordered proteins of human origin, ASED contains only 46 proteins with a total of 19,643 structurally characterized amino acid residues. Of these residues, 34% reside in disordered regions, and the remaining 66% reside in regions of known 3D structure.

The 46 proteins in the ASED data set contained 75 AS regions located entirely within structurally characterized areas. Fig. 2 shows that most of these 75 AS fragments correspond to entirely disordered segments, whereas less than one-fifth are entirely structured.

We also compared the distribution of disordered residues in the AS fragments to that of the entire data set and found that 57% of the residues in AS regions was characterized as disordered. This strong bias was in stark contrast to the ASED data set's disorder distribution: Only 34% of ASED's residues are characterized as disordered. Another 18 AS regions were partially characterized structurally, whereas 5 more AS regions had no structural characterization whatsoever (see Table 1, which is published as supporting information on the PNAS web site, for details).

To estimate the significance of this correlation, the observed frequency of disordered residue in AS regions was compared with the null hypothesis that ordered and disordered residues occur with the same frequency in the studied AS regions as in the entire ASED proteins. We used every 15th residue for this experiment on the assumption that such a separation insures that the residues are structurally independent of one another, thus enabling the use of the χ^2 statistical test. With these assumptions, the null hypothesis can be rejected with a P value of 5.68×10^{-30}.

Expanding the data set by means of disorder prediction. Although the statistical analysis of the ASED proteins indicate that AS regions

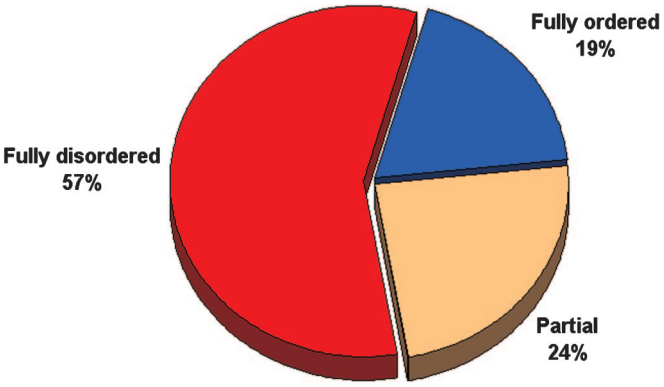

Fig. 2. Distribution of 75 AS regions according to the incidence of disorder. Regions of proteins encoded by sections of pre-mRNA that are spliced out in different protein isoforms were compared with regard to their structural characterization. The largest proportion of the protein segments (43 of 75, or 57%) is entirely disordered. These fragments encompass a total of 4,929 aa residues and range in length from 1 to 1,183. The 18 partially disordered segments (24%) range in length from 8 to 570 and include a total of 820 disordered and 1,581 ordered residues. Only 14 segments, or 19%, correspond to fully ordered regions. These segments range in length from 1 to 147 and comprise 613 aa residues. To obtain these results, a set of isoform transcripts corresponding to each protein in the ASED data set was obtained from the ASG database. Regions that are spliced out from pre-mRNA were found by translating those isoform transcripts and aligning the resulting protein sequences to that of the original ASED protein. Any segment from the original ASED protein that is missing from an isoform sequence is considered a spliced-out region. Only spliced-out regions that were fully characterized structurally were used to construct this graph.

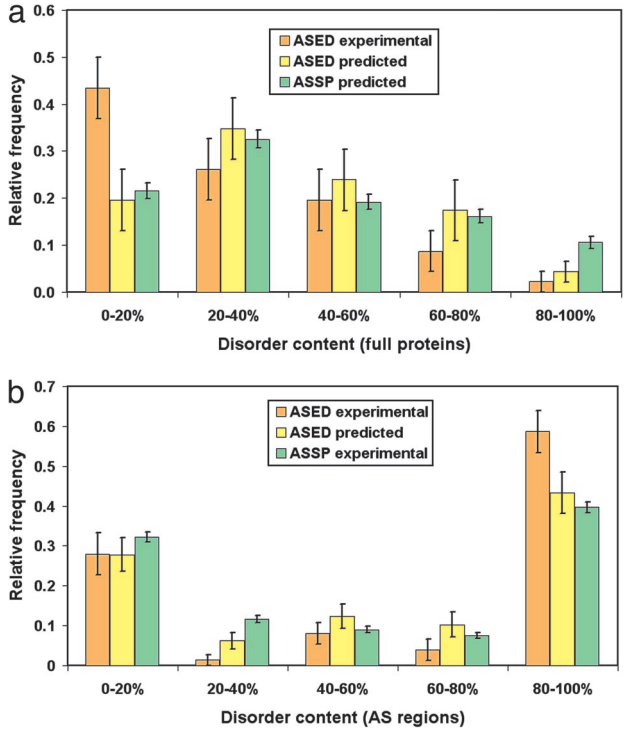

Fig. 3. Disorder content distributions of the studied data sets and their corresponding regions affected by alternative splicing. For proteins or regions in the ASED data set, the plots show bars for both experimentally determined disorder and predicted disorder. For the ASSP data set, the chart shows the predicted disorder content distribution. (*a*) Histogram of disorder content per protein. (*b*) Histogram of disorder content per region affected by alternative splicing. The error bars represent a 68% confidence interval or 1SD.

more often code for disorder, the number of proteins in ASED is small. Can the current results be extrapolated to eukaryotic proteins in general? One study of alternative splicing and protein structure found that, of 1,780 proteins with isoforms produced by alternative splicing, only 48, or 2.7% have a known 3D structure (27). The study used 4,804 known isoforms from higher organisms with fully sequenced genomes (*Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) extracted from the SwissProt database (27).

The small size of the ASED data set seems to be a consequence of the scarcity of structurally characterized proteins arising from alternative splicing. To provide a larger data set in which to test the hypothesis, the currently most accurate disorder predictor, PONDR VSL1 (28), was used. The analysis was carried out on two data sets: (*i*) the ASED data set described above, and (*ii*) a set of 558 proteins from SwissProt with experimentally confirmed isoforms arising from alternative splicing, which we called Alternative Splicing in SwissProt (ASSP). The latter data set included 1,266 regions affected by alternative splicing.

The VSL1 predictor has an average error rate of 20%, and very short regions of sequence can be erroneously predicted as either ordered or disordered. These errors produce small fluctuations in prediction that prevent many regions modified by alternative splicing from being predicted as fully ordered or fully disordered. As a result, a region-based comparison with the earlier ASED results, as shown in Fig. 2, is not useful, because almost all of the regions arising from alternative splicing are predicted as partially disordered. Because of the preponderance of partially disordered regions among predicted disorder, we measure the proportion of disorder in AS regions and entire proteins, as shown in Fig. 3. The figure shows the distribution of disorder content on full proteins from both data sets (Fig. 3*a*) and regions affected by alternative splicing (Fig. 3*b*). In the case of the ASED data set, we included the distributions of both predicted and experimentally determined disorders.

The distributions of disorder between the ASED and ASSP data sets are very similar, regardless of whether we use predicted or experimentally determined disorder content in the ASED data set. This result provides a strong validation for the analyses based on predicted disorder. From comparison of the distributions of predicted and experimental disorder in the full protein and AS regions only, it is clear that AS regions are biased toward higher disorder content, whereas the entire proteins are not highly biased toward higher disorder content. Thus, the results from the analysis of the larger ASSP data set using disorder predictions are consistent with those obtained from the smaller ASED data set using experimentally characterized disorder and thus support the hypothesis.

An analysis of amino acid compositions provided additional support for the relationship between alternative splicing to intrinsic disorder. Structured and intrinsically disordered regions have been shown to contain distinct amino acid compositions (29). To test for compositional biases, we analyzed the amino acid compositions of the following protein data sets: (*i*) a general, experimentally characterized, disorder data set containing all of the disordered regions in DisProt plus the unobserved regions from a set of nonredundant structures in the Protein Data Bank (PDB); (*ii*) all of the AS regions in the ASED data set; and (*iii*) all of the AS regions in the ASSP data set. These compositions were compared with those of a data set of ordered regions extracted from the PDB by amino acid composition profiling. The three sets display similar, statistically significant reductions in the structure-promoting amino acids and similar, statistically significant enrichments in the disorder-promoting amino acids (the results are shown graphically in Fig. 5, which is published as supporting information on the PNAS web site).

These data are consistent with the proposition that protein regions affected by alternative splicing of the corresponding genes are biased toward intrinsic disorder, further supporting the hypothesized relationship between intrinsic disorder and alternative splicing.

Functional Connection Between Intrinsic Disorder and Alternative Splicing. Although alternative splicing can lead to functional modification in different ways, the most common way appears to be the modification of the final protein product (4). A large-scale study found that 30% of a group of 1,300 AS genes showed differential profiles of conserved functional motifs; that is, different isoforms were shown to preserve different sets of functional domains (30). Many protein interaction domains are removed by alternative splicing with high frequency, which correlates with the fact that they are a very common type of functional domain (31). This finding would have the effect of modulating signaling and regulation pathways. Indeed, a computational study of alternative splice variants in SwissProt shows that 46% of the studied proteins are involved in signal transduction, gene expression, or regulation (7), and another large-scale study concluded that AS proteins are predominantly involved in signaling, cell communication, development, and apoptosis (32).

Intrinsic disorder is also predominantly related to regulation and signaling (22). Indeed, a diverse collection of regulatory elements has been experimentally associated with structurally characterized regions of disorder including the following examples: (*i*) binding targets for calmodulin and SH3; (*ii*) sites for phosphorylation, acetylation, methylation, fatty acylation, and ubiquitination; (*iii*) autoinhibitory peptides for inhibiting enzyme function; and (*iv*) binding regions for DNA and RNA. The preceding are only a few examples of ≈30 disorder-associated functions identified so far (15, 20).

The functional and regulatory elements in disordered regions are necessarily localized along the sequence. This arrangement is different from the mechanism of active site formation in ordered proteins, which relies on stable tertiary structure elements that bring together nonlocal residues. Also, many disordered regions contain multiple functional elements in tandem (15). Given these two features of disordered regions, alternative splicing could readily generate a set of protein isoforms having a highly diverse collection of regulatory elements. Indeed, arbitrarily large numbers of insertions and deletions arising from alternative splicing could be tolerated in disordered regions, thus readily accounting for the very large number of splicing isoforms that are sometimes indicated. Furthermore, removal of some, but not all, key residues could have the effect of modulating rather than completely eliminating the localized function. These observations provide a mechanism by which intrinsic disorder could explain the diversity of protein function that arises from alternative splicing.

An interesting, well characterized example of functional modulation by removal of functional domains and covalent modification sites is provided by the tumor suppressor protein BRCA1. This protein participates in many different cellular pathways, including transcription, apoptosis, and DNA repair, through direct or indirect interaction with a variety of partners (33). A well studied isoform of this protein has 1,863 aa and comprises a long central region flanked by ordered domains at the two termini. At the N terminus is a RING finger domain of 103 residues. This domain is reported to form a heterodimer with BRCA1 associated RING domain 1 (BARD1) and to bind to the ubiquitin C-terminal hydrolase BAP1. At the C terminus are two tandem copies of the BRCA1 C-terminal domain with a total of 218 residues making up the two domains. These two domains are reported to bind with transcriptional activators and repressors like CtIP.

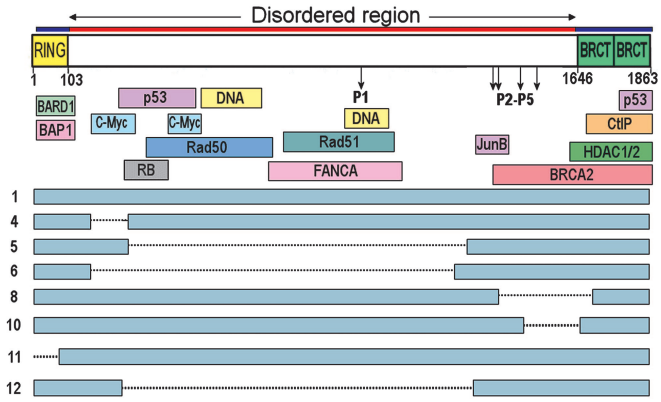

Fig. 4. Functional domains of BRCA1 and representative isoforms. The red line marked with the horizontal arrows on top indicates the extent of the 1,500-aa-long disordered central region. The ordered RING and BRCT domains are also shown at the termini. The vertical arrows show the location of phosphorylation sites, and the colored rectangles represent binding domains, including the binding partner name (see text). Isoforms are shown at the bottom, with dotted lines representing the regions missing from the translated protein products due to alternative splicing. The numbers at the left are the ASG isoform identifiers. The missing numbers (2, 3, 7, and 9) refer to isoforms that involve insertions or additions for which we have no structural information.

Mark *et al.* (34) recently undertook the structural characterization of the 1,500-aa central region of BRCA1. NMR and CD spectroscopy, protease sensitivity, and order/disorder predictions were used to examine 27 overlapping fragments spanning this central region. No structured segments were found, and all of the methods concurred in their indications of disorder over almost the entire central region. Several fragments formed stable complexes when mixed with their biological partners, namely DNA and p53, thus providing evidence that coupled folding and binding is the likely mechanism for the functions of these regions (15, 21, 35). Coupled folding and binding enables interactions with high specificity and low affinity (36) and also permits binding diversity (37). Overall, the Mark *et al.* study (34) provides strong evidence that the long central region of BRCA1 is disordered.

The disordered central region of BRCA1 contains molecular recognition domains for both DNA and several protein-binding partners, including the following: (*i*) tumor suppressors such as p53, retinoblastoma protein (RB), and BRCA2; (*ii*) oncogenes like c-Myc and JunB; (*iii*) DNA damage repair proteins such as Rad50 and Rad51; and (*iv*) the Fanconi anemia protein (FANCA). Fig. 4 displays a schematic view of the BRCA1 sequence, showing the terminal domains, the central disordered region, the published binding domains, the phosphorylation sites as reported in the Mark *et al.* study (34), and the protein regions found in different isoforms, as annotated in the ASG and SwissProt databases. Notice the absence of functional regions in different isoforms. This absence has the effect of creating diverse functional profiles for the transcribed gene products. Taking into account that several more partners of BRCA1 are known (33), and that the number of its known functional isoforms is greater than shown here [at least 24 variants have been reported (38)], the potential for a much larger number of functional profiles is clear.

Because regulatory and signaling elements in disordered regions can be comprised of just a few more or less continuous amino acids, a high density of functionally important segments can be located in regions of intrinsic disorder, which is observed for BRCA1. Indeed, previous work suggests that collections of regulatory functions are common in disordered regions (29).

Furthermore, disorder allows for the removal of functional segments without any deleterious effect on the stability of structured domains. Evidence for differential functional profiles arising from the disordered regions was found for all of the other proteins in the ASED data set. However, many of these functional regions were not experimentally verified like those of BRCA1 but rather were inferred by homology to conserved motifs and domains in functional databases.

Discussion

A hypothesis to explain how alternative splicing can produce multiple protein isoforms, which often differ in their functions but retain their original structural integrity, has been sorely lacking in the literature. Here we propose and provide support for the ideas that splicing sites generally occur within regions of intrinsic disorder and that splicing within these regions allows for functional and regulatory diversity with little or no disruption in protein structure. However, the implications of this apparent relationship between alternative splicing and intrinsic disorder go beyond the structural stability of isoforms and give insight into processes underlying the evolution of multicellular life forms.

Single-cell eukaryotes like *Schizosaccharomyces pombe* have an abundance of splicing for conversion of pre-mRNAs to mRNAs but apparently no alternative splicing, which appears to be common only in multicellular organisms. This relation suggests a link between alternative splicing and the higher-order complexity needed for multicellular life (8). Furthermore, alternative splicing has been proposed to reduce the selective pressure on genes, allowing organisms to experiment with new gene products fashioned by differential splicing. Indeed, novel gene products produced by the inclusion of new genetic material are free to evolve with less selective pressure because the original gene product is still present to carry out its function (39).

Similarly, disordered proteins are found in all kingdoms but are predicted to be much more common in multicellular eukaryotes than in archaea, eubacteria, and even single-celled eukaryotes (40, 41), an observation that corresponds intriguingly to the finding that alternative splicing occurs almost exclusively in multicellular eukaryotes. With regard to human proteins, 35–60% are estimated to be AS (4). For comparison, using methods described in ref. 40, 35% of human proteins are estimated to have at least one region of disorder that is 55 or more residues long and 60% are estimated to have at least one region of disorder that is 35 or more residues long. Thus, intrinsic disorder is estimated to be sufficiently prevalent in multicellular eukaryotes for regions of alternative splicing to code primarily for regions of protein disorder as proposed by our hypothesis.

Complex higher organisms in general have longer life cycles than their simpler counterparts, such as prokaryotes. This trait means that adaptation to new environments, and thus protein evolution, has to occur over far fewer generations. Indeed, a microbe usually goes through thousands of generations during the growth of an equivalent single-animal generation. Both alternative splicing and intrinsic protein disorder facilitate more rapid evolution of gene products, not only through functional profiling, which generates new functional variants with a higher probability and less selective pressure, but also by the fact that sequences of disordered protein evolve faster than those of ordered ones (42), thanks to less-restrictive amino acid substitutions and the lack of structural ramifications. Indeed, a recent study confirms that AS exons evolve faster than constitutively spliced exons in mammals (43).

It is believed that multicellular organisms could not have appeared until a suitable "genetic toolkit" had evolved. This toolkit has been suggested to be based on genes required for signaling and regulation (44) and, more specifically, homeotic genes (45, 46), which direct the building of multicellular bodies according to a master plan through genetic regulation. We propose an expanded developmental toolkit that includes alternative splicing, acting especially on genes encoding for intrinsically disordered proteins, which, as we have shown, are predominantly involved in signaling and regulation. Indeed, developmental regulation genes have been shown to generate isoforms through alternative splicing: In a particular example, a T-box gene of a sponge, considered the earliest example of metazoans, has been shown to produce isoforms through alternative splicing that affect phosphorylation and glycosylation states, which, in turn, affect gene regulation (46). The combination of these mechanisms also could have facilitated explosive speciation events in the evolutionary history of multicellular organisms, by tolerating multiple mutations that led to changes in regulation and in signaling networks. Taking into account that alternative splicing allows for the generation of not only enlarged proteomes, but also more complex and more precisely modulated protein interaction networks (3), it therefore follows that the proposed explosion of mutations could bring about a steep rise in regulatory complexity. Indeed, it has been proposed that these rapid diversification events are related to changes in genetic regulation and cell–cell signaling (47), and at least one modern case of explosive diversification in Lake Victoria cichlids has been related to alternative splicing (48).

The tolerable variations brought about by alternative splicing, facilitated by an increasing abundance of protein disorder in early eukaryotes, then would have provided an avenue for natural selection to enable the evolution of multicellular organisms and even facilitated their diversification over relatively short periods on the geological time scale.

Materials and Methods

We constructed this ASED data set to contain proteins having both ordered and disordered regions, with at least 10% of each molecule being in one of the two states. For such proteins, the ordered and disordered regions share a common evolutionary history, so that if alternative splicing has a preference to be associated with ordered or disordered regions, this preference would be detectable through comparison of ordered and disordered regions within each protein. By including protein with both structured and disordered regions, we avoided the complication of comparing proteins with different evolutionary histories. Additionally, it has been shown that the rates at which ordered and disordered regions evolve are different, with disordered regions usually evolving more rapidly than ordered regions (42).

For each ASED protein, the corresponding UniGene or SwissProt ID was used to obtain the sequence of its isoform from the ASG database or SwissProt, respectively. The isoform sequence then was aligned with the corresponding DisProt sequence to find structurally characterized regions that were removed by alternative splicing. Regions that were inserted or added to the structurally characterized protein were not included in our analysis, because these added fragments lack structural characterization annotations.

The ASSP data set was assembled as follows: 2,800 human proteins with reported isoforms produced by alternative splicing were downloaded from SwissProt through the SRS (Sequence Retrieval System). The set was processed to eliminate redundancy, so that 580 nonredundant proteins with $<25\%$ identity were obtained. After removing all of the proteins that did not have clearly defined regions affected by alternative splicing, as well as any proteins that were already included in the ASED data set, the final ASSP data set consisted of 558 proteins.

The PONDR VSL1 (28) disorder predictor was applied to all sequences in the ASED and ASSP data sets to characterize all their residues as either predicted disordered or predicted ordered. Notice that when using disorder predictions there are no uncharacterized residues, i.e., all residues are labeled as either disordered or ordered.

We thank Roderic Guigo for noticing the similarity in the values estimated for the frequencies of alternative splicing and for the frequencies of long regions of intrinsic disorder in eukaryotic cells.

This work was supported by National Institutes of Health Grant 5R01LM007688-02 and the Indiana Genomics Initiative (supported in part by the Lilly Endowment).

1. Sambrook, J. (1977) *Nature* **268,** 101–104.
2. Gilbert, W. (1978) *Nature* **271,** 501.
3. Lareau, L. F., Green, R. E., Bhatnagar, R. S. & Brenner, S. E. (2004) *Curr. Opin. Struct. Biol.* **14,** 273–282.
4. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A. & Soreq, H. (2005) *Gene* **344,** 1–20.
5. Csete, M. E. & Doyle, J. C. (2002) *Science* **295,** 1664–1669.
6. Lopez, A. J. (1998) *Annu. Rev. Genet.* **32,** 279–305.
7. Boue, S., Vingron, M., Kriventseva, E. & Koch, I. (2002) *Bioinformatics* **18,** Suppl. 2, S65–S73.
8. Ast, G. (2004) *Nat. Rev. Genet* **5,** 773–782.
9. Demchenko, A. P. (2001) *J. Mol. Recognit.* **14,** 42–61.
10. Oakley, A. J., Harnnoi, T., Udomsinprasert, R., Jirajaroenrat, K., Ketterman, A. J. & Wilce, M. C. (2001) *Protein Sci.* **10,** 2176–2185.
11. Hymowitz, S. G., Compaan, D. M., Yan, M., Wallweber, H. J., Dixit, V. M., Starovasnik, M. A. & de Vos, A. M. (2003) *Structure (London)* **11,** 1513–1520.
12. Peneff, C., Ferrari, P., Charrier, V., Taburet, Y., Monnier, C., Zamboni, V., Winter, J., Harnois, M., Fassy, F. & Bourne, Y. (2001) *EMBO J.* **20,** 6191–6202.
13. Lee, K. A., Fuda, H., Lee, Y. C., Negishi, M., Strott, C. A. & Pedersen, L. C. (2003) *J. Biol. Chem.* **278,** 44593–44599.
14. Fiegen, D., Haeusler, L. C., Blumenstein, L., Herbrand, U., Dvorsky, R., Vetter, I. R. & Ahmadian, M. R. (2004) *J. Biol. Chem.* **279,** 4743–4749.
15. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002) *Biochemistry* **41,** 6573–6582.
16. Holt, C. & Sawyer, L. (1993) *J. Chem. Soc. Faraday Trans.* **89,** 2683–2692.
17. Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A. & Lansbury, P. T., Jr. (1996) *Biochemistry* **35,** 13709–13715.
18. Wright, P. E. & Dyson, H. J. (1999) *J. Mol. Biol.* **293,** 321–331.
19. Iakoucheva, L. M. & Dunker, A. K. (2003) *Structure (London)* **11,** 1316–1317.
20. Uversky, V., Gillespie, J. & Fink, A. (2000) *Proteins* **41,** 415–427.
21. Dyson, H. J. & Wright, P. E. (2005) *Nat. Rev. Mol. Cell Biol.* **6,** 197–208.
22. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. (2002) *J. Mol. Biol.* **323,** 573–584.
23. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. (2005) *J. Mol. Recognit.* **18,** 343–384.
24. Leipzig, J., Pevzner, P. & Heber, S. (2004) *Nucleic Acids Res.* **32,** 3977–3983.
25. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.* (2003) *Nucleic Acids Res.* **31,** 365–370.
26. Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., Cortese, M. S., Lawson, J. D., Brown, C. J., Sikes, J. G., *et al.* (2005) *Bioinformatics* **21,** 137–140.
27. Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S. & Sunyaev, S. (2003) *Trends Genet.* **19,** 124–128.
28. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P. & Dunker, A. K. (2005) *Proteins* **61,** Suppl. 7, 176–182.
29. Dunker, A. K., Brown, C. J. & Obradovic, Z. (2002) *Adv. Protein Chem.* **62,** 25–49.
30. Loraine, A. E., Helt, G. A., Cline, M. S. & Siani-Rose, M. A. (2003) *J. Bioinform. Comput. Biol.* **1,** 289–306.
31. Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R. & Lee, C. (2004) *J. Proteome Res.* **3,** 76–83.
32. Liu, S. & Altman, R. B. (2003) *Nucleic Acids Res.* **31,** 4828–4835.
33. Deng, C. X. & Brodie, S. G. (2000) *BioEssays* **22,** 728–737.
34. Mark, W. Y., Liao, J. C., Lu, Y., Ayed, A., Laister, R., Szymczyna, B., Chakrabartty, A. & Arrowsmith, C. H. (2005) *J. Mol. Biol.* **345,** 275–287.
35. Spolar, R. S. & Record, M. T., Jr. (1994) *Science* **263,** 777–784.
36. Schulz, G. E. (1979) in *Molecular Mechanism of Biological Recognition*, ed. Balaban, M. (Elsevier/North–Holland, New York), pp. 79–94.
37. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 11504–11509.
38. Orban, T. I. & Olah, E. (2003) *Mol. Pathol.* **56,** 191–197.
39. Boue, S., Letunic, I. & Bork, P. (2003) *BioEssays* **25,** 1031–1034.
40. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000) *Genome Inform. Ser. Workshop Genome Inform.* **11,** 161–171.
41. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004) *J. Mol. Biol.* **337,** 635–645.
42. Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J. & Dunker, A. K. (2002) *J. Mol. Evol.* **55,** 104–110.
43. Chen, F. C., Wang, S. S., Chen, C. J., Li, W. H. & Chuang, T. J. (2006) *Mol. Biol. Evol.* **23,** 675–682.
44. King, N. & Carroll, S. B. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 15032–15037.
45. Slack, J. M., Holland, P. W. & Graham, C. F. (1993) *Nature* **361,** 490–492.
46. Adell, T., Grebenjuk, V. A., Wiens, M. & Muller, W. E. (2003) *Dev. Genes Evol.* **213,** 421–434.
47. Sole, R. V., Fernandez, P. & Kauffman, S. A. (2003) *Int. J. Dev. Biol.* **47,** 685–693.
48. Terai, Y., Morikawa, N., Kawakami, K. & Okada, N. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 12798–12803.