NOT THAT RIGID MIDGETS AND NOT SO FLEXIBLE GIANTS: ON THE ABUNDANCE AND ROLES OF INTRINSIC DISORDER IN SHORT AND LONG PROTEINS

MARK HOWELL

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
mhowell1@health.usf.edu

RYAN GREEN

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
rgreen1@health.usf.edu

ALEXIS KILLEEN

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
akillee1@health.usf.edu

LAMAR WEDDERBURN

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
lwedderb@health.usf.edu

VINCENT PICASCIO

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
vincentpicascio@gmail.com

ALEJANDRO RABIONET

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
arabion2@health.usf.edu

ZHENLING PENG

*Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 2V4, Canad*a
zhenling@ualberta.ca

MAYA LARINA

Department of Mathematics and Informatics, College of Medical Biochemistry, Volgograd State Medical University, 400131 Volgograd, Russia
m_larina2@pochta.ru

BIN XUE

Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA
bxue@health.usf.edu

LUKASZ KURGAN

*Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 2V4, Canad*a
lkurgan@ualberta.ca

VLADIMIR N. UVERSKY

Department of Molecular Medicine, USF Health Byrd Alzheimer's Research Institute, College of Medicine, University of South Florida, Tampa, FL 33612, USA; Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia
vuversky@health.usf.edu

Received (09 August 2012)
Accepted (16 November 2012)

Intrinsically disordered proteins or proteins with disordered regions are very common in nature. These proteins have numerous biological functions which are complementary to the biological activities of traditional ordered proteins. A noticeable difference in the amino acid sequences encoding long and short disordered regions was found and this difference was used in the development of length-dependent predictors of intrinsic disorder. In this study, we analyze the scaling of intrinsic disorder in eukaryotic proteins and investigate the presence of length-dependent functions attributed to proteins containing long disordered regions.

Keywords: Intrinsically disordered protein; Protein function; Eukaryotic proteome.

1. Introduction

It is clear now that any given proteome contains significant number of highly flexible proteins that possess crucial biological functions.[1-5] The discovery of these intrinsically disordered proteins (IDPs) and protein regions (IDPRs) challenged and changed the traditional protein structure-function paradigm which states that well-defined three-dimensional structure is required for the correct functioning of a protein and that the structure defines the function of the protein.[6-8] IDPs and IDPRs are involved in numerous biological processes, where they play different roles in regulation of the function of their binding partners and in promotion of the assembly of supra-molecular complexes.[6-39] It was further suggested that many protein functions indeed require disordered regions of flexible, dynamic conformations instead of rigid ordered regions.[12] The conformational plasticity associated with intrinsic disorder provides IDPs/IDPRs with a wide spectrum of exceptional functional advantages over the functional modes of ordered proteins and domains.[8; 9; 11-13; 20; 21; 25; 30; 31; 33; 34; 39]

IDPs/IDPRs possess complex “anatomy” (they contain multiple, relatively short functional elements), which contributes to their unique “physiology” (an ability to be involved in interaction with, regulation of and control by multiple structurally unrelated partners). Given the existence of multiple functions in a single disordered protein, and given that each functional element is typically relatively short, alternative splicing could readily generate a set of protein isoforms with a highly diverse set of regulatory elements.[40] The complexity of the disorder-based interactomes is further increased due to the ability of a single IDPR to bind to multiple partners gaining very different structures in the bound state.[34; 41] Often, dysfunction and dysregulation of IDPs are associated with the development of various pathological conditions and intrinsic disorder is commonly seeing in proteins from pathogenic microbes and viruses.[42-52]

IDPs and IDPRs differ from structured globular proteins and domains with regard to many attributes, including amino acid composition, sequence complexity, flexibility, charge, hydrophobicity, and type and rate of amino acid substitutions over evolutionary time.[8; 26; 53; 54] This makes IDPs and IDPRs recognizable and opens numerous opportunities for the development of specific predictors of intrinsic disorder that can be used to evaluate the propensity for intrinsic order and disorder from the amino acid sequence.[55] The success of these predictors strongly supports the hypothesis that intrinsic disorder, like globular structure, is also encoded by the amino acid sequence.[7; 8]

It was also proposed that there could be several subtypes (flavors) of intrinsic disorder that could be distinguished by amino acid compositions and sequence properties and therefore resemble the structural classification of ordered proteins, e.g. α-helix and β-sheet at the secondary structure level, and all α, all β, α/β and α+β classes at the tertiary structure level.[56] Structurally, entirely disordered proteins are commonly classified as proteins possessing extended disorder (native coils and native “pre-molten globules”) and proteins with compact disordered conformations (native “molten globules”).[10; 15-17; 39] Furthermore, it was recognized that there is a noticeable difference in the amino acid sequences encoding long and short disordered regions.[57-59] These differences were exploited in the development of length-dependent predictor of intrinsic disorder, which is an accurate meta-predictor which was trained to integrate the specialized predictors for short (≤30 residues) and long disordered regions (>30 residues) into the final predictor model.[59] The resulting meta-predictor was shown to be applicable for identification of disordered regions of any length, being able to accurately identify the short disordered regions that are often misclassified by our previous disorder predictors.[59]

In this study, we first analyzed the peculiarities of the disorder distribution in short and long eukaryotic proteins and investigated their functional priorities. Next, from a list of all know short (20-100 amino acids) and long (over 3,000 amino acids) UniProt proteins, random protein samples were selected, and functional and structural information was gathered from literature for each protein in these sample sets. For each protein, this analysis was further supplemented by the evaluation of the intrinsic disorder propensity using a set of well-known disorder predictors. These data were used to determine if intrinsic disorder play a role in the functioning of a given protein.

2. Materials and Methods

2.1. *Datasets*

We analyze all 110 complete eukaryotic proteomes, which total to 1,901,810 proteins, from UniProt release 2011_08.[60] The proteomes are assigned to their abbreviated taxonomic lineage based on NCBI,[61] where the lowest taxonomic level, which we refer to as “species”, could be the genus, family or species. Two protein sets were extracted from the resulting dataset: (1) 73,261 short-proteins with up to 75 amino acids; and (2) 50,090 long proteins that include chains with at least 1500 residues. Three subsets were generated from these two sets: (1) short-proteins that contain long disordered segments with at least 30 consecutive disordered residues; (2) long-proteins with long disordered segments with at least 30 consecutive disordered residues; and (3) long-proteins with long disordered segments of at least 100 residues.

2.2. *Methods*

We applied two fast and accurate disordered predictors, IUPRED[62; 63] and Espritz,[64] to obtain putative disordered residues and segments. We used two versions of IUPRED that were designed for predictions of long and short disordered segments, respectively, and three versions of Espritz that consider disorder annotations by false positive rate and based on NMR structures, X-ray crystal structures, and experimental annotations from DisProt database.[65] Espritz and IUPRED were shown to be competitive in terms of their predictive quality[63; 66] and they cover the main characteristics of the disorder including the three annotation types and two types of disordered segments. The resulting five predictions were combined together using the majority vote consensus, which is motivated by the fact that consensus-based approaches provide improved predictive quality.[67] The putative disorder was used to calculate the disorder content (fraction of disordered residues in a given chain) and to characterize proteins with long disordered segments that consist of at least 30 or 100 consecutive disordered amino acids.

Based on the Gene Ontology (GO)[68] terms that are provided in UniProt, we investigate relations between protein length, intrinsic disorder, and protein functions; the latter are expressed as biological processes or molecular functions. We removed functional annotations with less than 50 annotated chains, as they would not provide sufficient sample size to produce statistically sound results. We utilize all available annotations including those based on experimental evidence, curator and author statement-based evidence, evidence based on computational analysis, and evidence inferred automatically from electronic annotations. Significant majority of the annotations is computational, based either on sequence/structural similarity or inferred from electronic annotations, and thus relatively broad and statistically sound analysis was possible only when we included these annotations into the analysis. We note that we include all GO terms in the analysis, including both leaf and internal terms. We contrast the abundance of a given functional annotation between a given set of chains (say short-proteins with long, >30, disordered segments) and a background abundance of this annotation in the entire set of eukaryotic proteins. We define this functional abundance as a ratio between the number of proteins annotated with a given functions and the total number of functionally annotated proteins in a given protein set. We randomly select half of the proteins from the considered protein subset and compare them with the same number of chains drawn at random from the entire eukaryotic proteome dataset using the functional abundance. This is repeated 10 times and we evaluate significance of the differences in these two functional abundance vectors of a given function. If the measurements are normal, as evaluated with the Anderson-Darling test[69] at 0.05 significance, then we utilize the t-test; otherwise we use the non-parametric Wilcoxon rank sum test.[70] We consider only the functions with abundance of at least 2% in a considered protein subset.

2.3. *Functional and disorder annotations for random protein samples*

From a list of all know short (20-100 amino acids) and long (over 3,000 amino acids) UniProt proteins (www.uniprot.org), random protein samples were selected, and functional and structural information was gathered from literature for each protein in these sample sets. For each protein, this analysis was further supplemented by the evaluation of the intrinsic disorder propensity using a set of well-known disorder predictors.

The intrinsic disorder propensities of the proteins from random samples were evaluated by two different disorder predictors. The first one is PONDR® VLXT,[71] which applies various compositional probabilities and hydrophobic measures of amino acid as the input features of artificial neural networks for the prediction. Although it is no longer the most accurate predictor, it is very sensitive to the local compositional biases. Hence, it is capable of identifying potential molecular interaction motifs.[22; 72] The second predictor is a meta-predictor PONDR®FIT,[73] that combines six individual predictors, which are PONDR® VLXT,[71] VSL2,[74] VL3,[59] FoldIndex,[75] IUPred,[62] and TopIDP.[76] This meta-predictor is moderately more accurate than each of the component predictors.

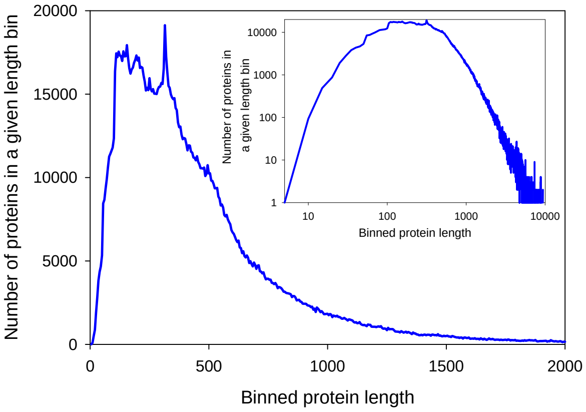

Figure 1. Length distribution of the eukaryotic proteins analyzed in this study. Plot shows data for limited range of protein lengths. Inset represents the length distribution of all the eukaryotic proteins in the double logarithmic scale. In this plot, protein chain length is binned into intervals of 5 residues.

3. Results and Discussion

3.1. *Amino acid compositions of short and long eukaryotic proteins*

Figure 1 represents length distribution of 1,901,810 proteins from 110 complete eukaryotic proteomes found in UniProt release 2011_08, whereas inset to this figure represents data in the double logarithmic scale. It is clearly seeing that the number of proteins first increases with the increase in the protein length, approaching maximum at a length of about 150 residues, and then decreases steadily as the protein length further increases. Typically, bins corresponding to the proteins longer than 2,000 residues contain only a few counts. The presence of the optimal protein length in a range of 150-200 residues is further evidenced from the inclusion to Figure 1, which shows the length distribution of eukaryotic proteins in the double logarithmic scale.

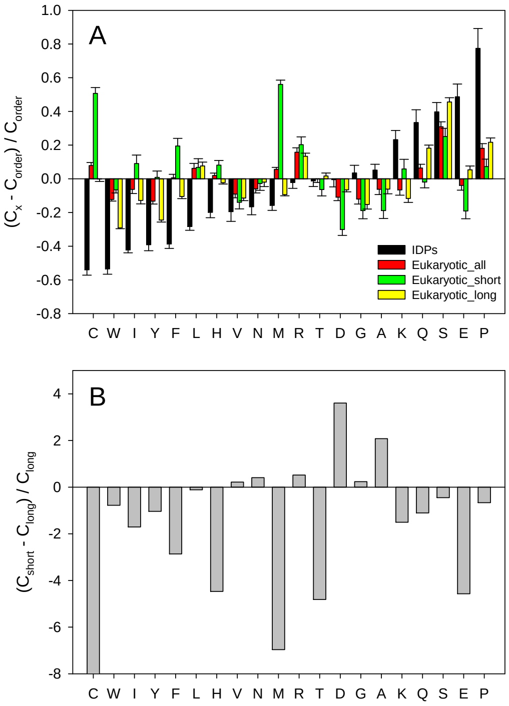

Figure 2. **A**. Fractional difference in the amino acid composition between the different eukaryotic proteins and a set of completely ordered proteins calculated for each amino acid residue (compositional profiles). The fractional difference was evaluated as $(C_x\text{-}C_{order})/C_{order}$, where C_x is the content of a given amino acid in a query set, and C_{order} is the corresponding content in the dataset of fully ordered proteins. Composition profiles for all eukaryotic proteins, as well as short and long eukaryotic proteins are shown by red, green and yellow bars, respectively. Composition profile of typical intrinsically disordered proteins from the DisProt database is shown for comparison (black bars). Positive bars correspond to residues found more abundantly in histones, whereas negative bars show residues, in which histones are depleted. Amino acid types were ranked according to their increasing disorder-promoting potential. B. Fractional difference in the amino acid composition between the short and long eukaryotic proteins. For this plot, the fractional difference was evaluated as $(C_{short}\text{-}C_{long})/C_{long}$, where C_{short} is the content of a given amino acid in a set of short eukaryotic proteins, and C_{long} is the corresponding content in the dataset of long proteins.

Analysis of the amino acid composition biases can provide interesting information on the nature of a protein. For example, the amino acid compositions of extended IDPs are characterized by some global biases, where low mean hydropathy is combined with high mean net charge. These global biases determine the highly unstructured and extended state of these proteins, since high net charge leads to strong electrostatic repulsion, and low hydropathy prevents efficient compaction.[7] In agreement with these global observations, IDPs were shown to be significantly depleted in so-called order-promoting amino acids, C, W, I, Y, F, L, H, V, and N, and substantially enriched in disorder-promoting residues, A, G, R, T, S, K, Q, E, and P.[8; 26; 71; 77; 78] We use a computational tool, Composition Profiler,[78] to investigate the compositional biases in short and long eukaryotic proteins. This approach is based on the calculation of a normalized composition of a given protein or protein dataset in the $(C_s - C_{order})/C_{order}$ form, where C_s is a content of a given residue in a query protein or dataset, and C_{order} is the corresponding value for the set of ordered proteins from PDB Select 25.[79]

Figure 2 shows that, in comparison with typical ordered proteins, eukaryotic proteins from all 110 complete eukaryotic proteomes are moderately depleted in some order-promoting residues (e.g., W, I, Y, V, and N, see red bars in Figure 2A) and are moderately enriched in some disorder-promoting residues (e.g., Q, S, and P). Long eukaryotic proteins are depleted in order-promoting W, I, F, V, N, and M, and enriched in disorder-promoting Q, S, E, and P. Both depletion in order-promoting residues and enrichment in disorder-promoting residues are generally more pronounced that those calculated for eukaryotic proteins in general (see yellow bars in Figure 2A). Short eukaryotic proteins are depleted in order-promoting W and V, being noticeably enriched in C, I, F, H, and M, and are enriched in disorder-promoting K, S and P, being depleted in such disorder-promoting residues as D, G, A, and E (see green bars in Figure 2A). Figure 2 also clearly shows that amino acid compositions of short and long eukaryotic proteins are rather different. This is most evident from Figure 2B, which represents the relative composition of short eukaryotic proteins calculated as $(C_{short} - C_{long})/C_{long}$. This plot clearly indicates that in comparison with long eukaryotic proteins, short eukaryotic proteins are dramatically depleted in almost all order-promoting residues (C, W, I, Y, F, H, and M) and some disorder-promoting residues (K, Q, S, E, and P). However, some disorder-promoting residues (R, D, G, and A) are noticeably more common in short proteins. Based on these observations, one could expect that short and long eukaryotic proteins might possess different overall propensities for intrinsic disorder. In agreement with this hypothesis, the subsequent analysis revealed very peculiar protein length scaling of intrinsic disorder in eukaryotic proteins.

3.2. *Abundance of intrinsic disorder in short and long eukaryotic proteins*

Figure 3A shows that the protein length-dependence of the average disorder content possess an intriguing shape. In fact, short proteins are predicted to have significant amount of disorder. The amount of predicted disorder decreases as protein length increases and reaches minimum at ~ 15% for proteins with the length of 250-300 residues. Then, the amount of intrinsic disorder start to increase, reaches a plateau at the level of 24-28% for proteins with length of ~1,000-2,000 residues, and then again starts to decrease for longer proteins. Since the number of very long proteins is relatively small, that part of the plot corresponding to proteins longer than 5,000 residues is very noisy. Importantly, some long proteins contain very significant amount of predicted disorder, up to 90-95%. Inset to Figure 3A represents length distribution of predicted disorder for proteins shorter than 1,500 residues to better illustrates the peculiarities of scaling for shorter proteins.

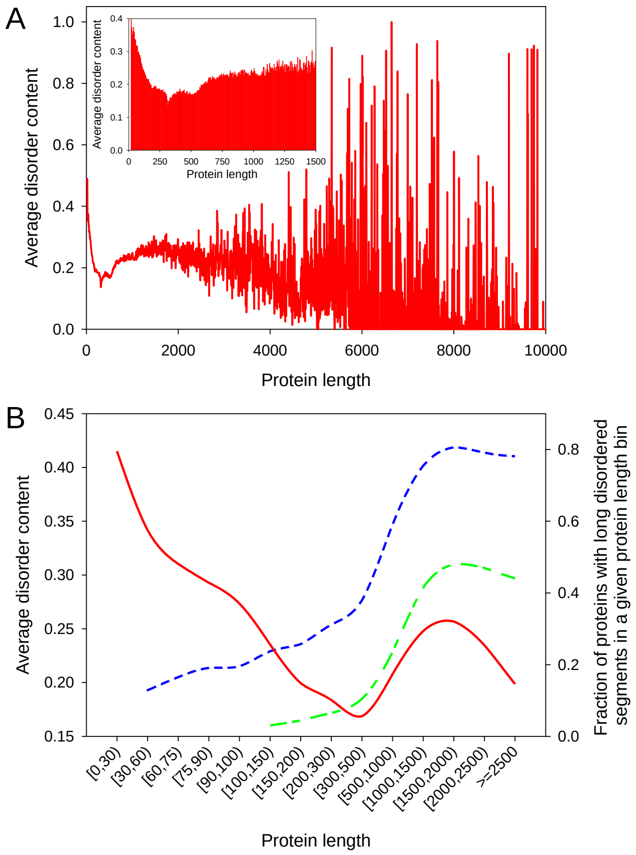

Figure 3. **A**. Length dependence of the intrinsic disorder in eukaryotic proteomes represented as an average disorder content versus protein length plot in a wide range of protein length (up to 10,000 residues). Inset represents an expanded dependence for proteins with the length up to 1500 residues. In this plot, protein chain length is binned into intervals of 5 residues. **B**. Average disorder content (solid red line) and fraction of proteins with long disordered segments composed of 30 (short dashed blue line) or 100 (long-short dashed green line) consecutive disordered residues vs. protein chain length that is binned into intervals shown on the x-axis.

Figure 3B further summarizes the results of this analysis and shows the relations between the disorder content, fraction of proteins with long disordered segments; i.e., proteins with at least 30 or at least 100 consecutive disordered residues, and protein chain size. It is seen that the fraction of proteins with long disordered regions increases for proteins with the length from 30 to ~2,000 residues and then reaches plateau or even starts to slightly decrease. Figure 3B also shows that ~80% of proteins longer than 1,500 contain disorder regions longer than 30 residues and ~50% of such proteins contain disordered regions longer 100 residues.

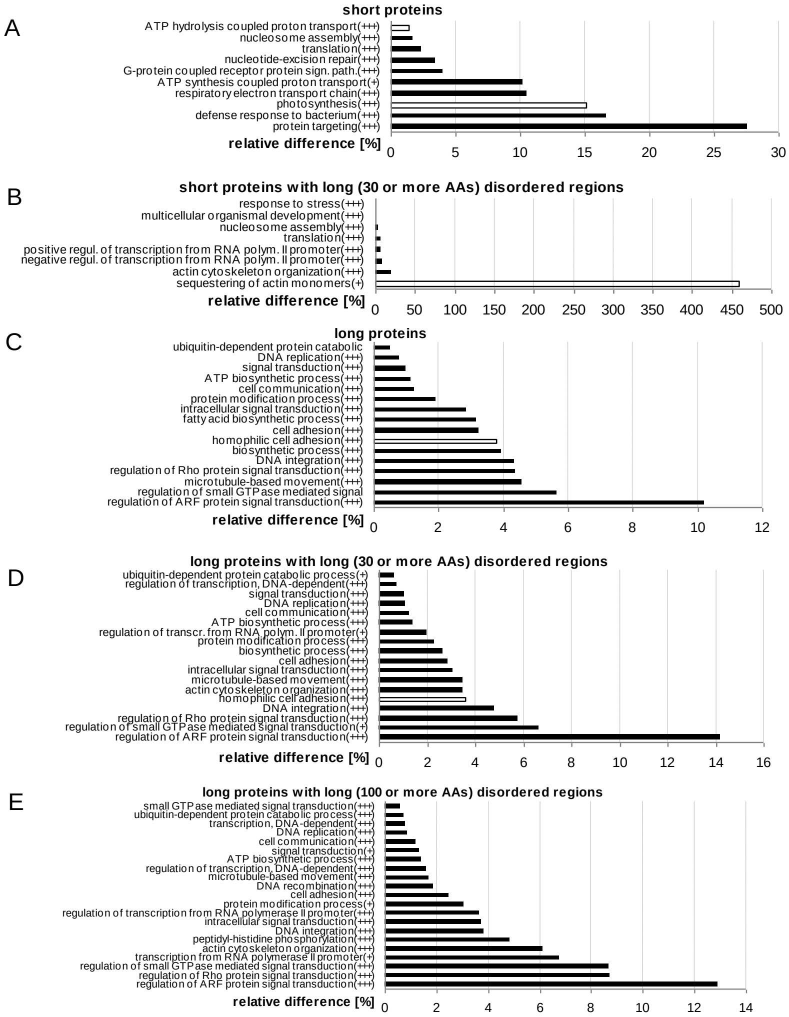

Figure 4. Significantly enriched biological processes for (Panel **A**) short proteins (75 or fewer amino acids), (Panel **B**) short proteins that have long (30 or more consecutive amino acids) disordered segments, (Panel **C**) long proteins (1500 or more residues), (Panel **D**) long proteins that have long (30 or more consecutive amino acids) disordered segments, (Panel **E**) long proteins that have long (100 or more consecutive amino acids) disordered segments. Hollow/solid bars denote processes that correspond to GO terms from leaf/internal nodes. The x-axis shows relative difference defined as the difference of the functional abundance for a considered protein subset relative to the abundance on the entire protein dataset. "+/++/+++" denotes the differences that were found significant with p-value below 0.05/0.001/0.0001.

In agreement with the data shown in Figure 3A, proteins with the length of 200-500 residues possess the least amount of predicted disorder. It is of interest that this length interval is within the optimal length of a protein domain. It is also interesting that it seems that there is an optimal length of long proteins to have high disorder content. In fact, Figure 3 clearly shows that long protein with the length ranging from 1,500 to 2,000 residues are predicted to have most disorder, with ~25-30% of their residues being predicted to be disordered.

3.3. *GO-based functional annotation of short and long eukaryotic proteins*

Data shown in Figure 3 (especially in its part describing the length dependence of the average disorder content) are a bit counter-intuitive. In fact, based on the simple probability evaluations, one can expect that short proteins would contain less disorder than long proteins, and therefore the disorder content would increase with the protein length. In fact, this expectation is fulfilled for proteins with the length of their polypeptide chain ranging from ~200 to ~1,500 residues (see Figure 3A). The fact that short proteins (shorter than 150 residues) contain highest amount of predicted disorder and the fact that long disordered proteins seem to have some optimal length (1,500-2,000 residues) with relatively high disorder content (25-30%) may potentially have some functional explanations. To check this hypothesis, we analyzed the distribution of GO-annotated biological processes and functions in short and long proteins. Figure 4 represents the results of such analysis for biological processes, whereas the results of the analysis for biological functions are shown in Figure 5.

Figures clearly show that both biological processes and biological functions ascribed to short proteins (shorter than 75 residues) are very different from those of long proteins (1,500 or more residues). Furthermore, some interesting observations can be made in relation to the functions of proteins with long disordered regions. Most abundant biological processes for short proteins with long regions of disorder are multicellular organism development, nucleosome assembly, translation, positive and negative regulation of transcription, and actin cytoskeleton organization (see Figure 4B). Most abundant biological functions for these proteins are transcription factors, DNA binding, receptor activity, protein binding, and structural constituent of ribosome (see Figure 5B). Sets of biological processes and molecular functions for long proteins with long disordered regions are very different those discussed above (see Figures 4D, 4E, 5D and 5E). The general functional themes here are binding of small molecules, metal ions, cofactors, DNA, other proteins, regulation and control activity in various biological processes such as transcription, DNA replication, signal transduction, cell communication, DNA integration and many others. Importantly, these observations are in a good agreement with earlier studies, since many of functions and processes shown in Figures 4 and 5 were earlier attributed to IDPs and IDPRs. We also note that the enriched GO terms for short and long chains include similar proportions of annotations from leaf (hollow bars) and internal (solid bars) nodes, which suggests that generality of the GO terms did not bias our length dependent analysis.

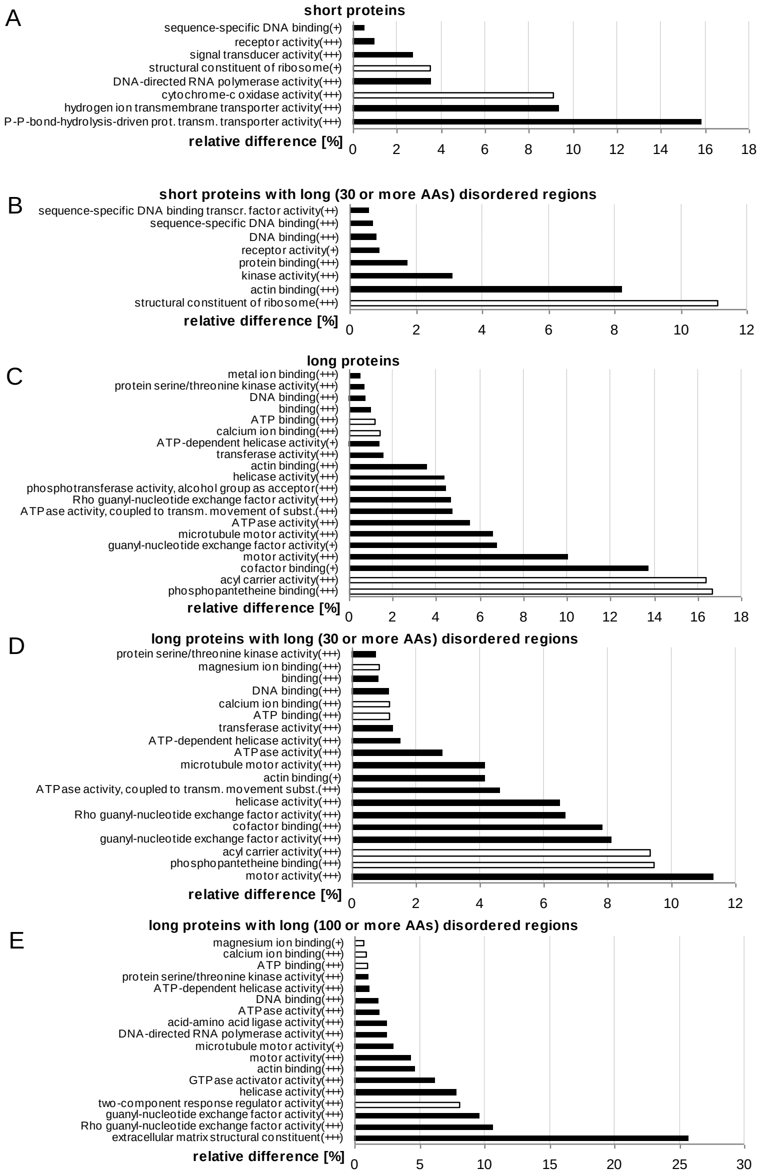

Figure 5. Significantly enriched molecular functions for (Panel **A**) short proteins (75 or fewer amino acids), (Panel **B**) short proteins that have long (30 or more consecutive amino acids) disordered segments, (Panel **C**) long proteins (1500 or more residues), (Panel **D**) long proteins that have long (30 or more consecutive amino acids) disordered segments, (Panel **E**) long proteins that have long (100 or more consecutive amino acids) disordered segments. Hollow/solid bars denote functions that correspond to GO terms from leaf/internal nodes. The x-axis shows relative difference defined as the difference of the functional abundance for a considered protein subset relative to the abundance on the entire protein dataset. "+/++/+++" denotes the differences that were found significant with p-value below 0.05/0.001/0.0001.

3.4. *Structure, function, and intrinsic disorder of some short proteins*

At the next stage, we analyzed available structural and functional information for several random samples of short and long proteins. To this end, 16 short (containing 20-100 amino acids) and 12 long proteins (containing ≥ 3,000 residues) were randomly selected from the UniProt database (www.uniprot.org). For these 27 proteins, functional and structural information was gathered from literature and disorder propensity was evaluated using a set of well-known disorder predictors. This functional, structural and disorder information was then used to see if intrinsic disorder play a role in the functioning of a given protein. Results of this analysis are represented below.

3.4.1. *HEPN-1*

The product of the putative cancer susceptibility gene, hepatocellular carcinoma down-regulated 1 (*HEPN1*), is a 10.305 Da protein comprised of 88 amino acids, which is predominantly found in the cytoplasm of hepatocytes.[80] *HEPN1* is mapped to chromosome 11q24.2. More specifically, *HEPN1* is located on the antisense strand of the 3'-noncoding region of the *HEPACAM* gene.[81] This protein retain "Putative" status since its gene maps to the 3'-noncoding region of the *HEPACAM* gene and since it is not entirely clear if this DNA sequence is translated into an actual protein within the cells. However, there are several abnormalities associated with chromosome 11q dysfunction, such as chromosome breakage, rearrangement and loss on the long arm.[80] These abnormalities have been indicated in a number of human cancers, including colorectal, breast, melanoma, acute lymphoblastic leukemia, and ovarian cancer, signifying the likely presence of tumor suppressor genes in this region.[80]

Transient transfection studies showed that exogenous *HEPN1* is capable of suppressing cell growth and inducing apoptosis in Hepatocellular carcinoma (HepG2) cells.[80] Since expression of *HEPN1* is down-regulated or lost in hepatocellular carcinomas (HCC) patients and cell lines, it was suggested that loss of this gene can be involved in carcinogenesis of hepatocytes.[80] However, researchers have failed to identify any correlation between the loss of *HEPN1* expression and the events related to the development and progression of HCC, as well as failed to identify the mechanisms of its antiproliferative effect.[80]

HEPN1 does not resemble any known human tumor suppressor genes, and the product of this gene, HEPN1, has no significant homology to any known proteins or peptides and does not contain any known functional or sequence motifs.[80] Disorder propensity analysis revealed that the N-terminal segment comprising the first 13 amino acids of HEPN1 is expected to be disordered, but the rest of the protein (14-88) is predicted to be ordered.

3.4.2. *MALAT-1*

MALAT-1 is a gene located on chromosome 11q13. The RNA transcript of more than 8000 nucleotides, with an ORF of 940 nucleotides encodes for a theoretical protein with a length of 57 amino acids and a molecular mass of 6.29 kDa. High levels of MALAT-1 RNA transcript expression have been reported in the pancreas and lung, with intermediate expression levels reported in the prostate, ovary, colon, placenta, spleen, small intestine, kidney, heart, liver, testis and brain.[82] High levels of expression have also been detected in both non-small cell lung cancer cell lines and patient samples.[82] High levels of MALAT-1 RNA transcript expression in lung adenocarcinoma have been associated with the subsequent metastasis of the cancer.[82]

MALAT-1 RNA transcript expression has been shown to be significantly up regulated in placenta previa increta/percreta and strongly associated with the trophoblast-like cell invasion *in vitro*.[83] These results suggest that MALAT-1, already having been linked to invasive malignancies, may also play a functional role in the development of abnormally invasive placentation.[83] It has been shown that there is a striking similarity of invasion ability between placental and cancer cells, suggesting a general role for MALAT-1 in abnormal cellular invasion.[84]

In yet another role, MALAT-1 has been identified as a regulator of the tumor suppressor protein, PSF.[85] Other cellular functions of MALAT-1 include pre-mRNA metabolism via an intimate association with SC35 splicing domains within the mammalian nucleus,[86] and a role in nucleoli as a riboregulator controlling expression of its targeted gene,[87] as well as regulating alternative splicing by modulating the levels of active serine/arginine splicing factor proteins.[88] These results suggest that MALAT-1 plays a role in regulating cellular interactions and functions.[84-87] Much more information is needed to determine the exact molecular mechanisms of how MALAT-1 is involved in normal cellular functioning and abnormal cellular invasion.

According to UniProtKB, there is evidence at the transcript level for the theoretical protein that MALAT-1 RNA transcript would encode for. This theoretical protein discloses no substantial sequence motifs. The MALAT-1 RNA transcript lacks credible open reading frames and does not contain a valid Kozak sequence, suggesting the unlikelihood of translation and its role as a non-coding RNA.[82; 83] Furthermore, significant signal of MALAT-1 is not observed outside the nucleus.[86] Disorder predictions for the MALAT-1 show a protein that is intrinsically disordered in the segments containing the first 8 amino acids and the last 7 amino acids. The protein is predicted to be highly structured throughout its central part. Due to the fact that MALAT-1 RNA transcript is most likely not translated into a protein, intrinsic disorder will not play a role in the functioning of this molecule.

3.4.3. *Ovarian cancer-related protein 1*

Ovarian cancer-related protein 1 is a protein composed of 76 amino acids, with a molecular mass of 8,369 Da. The protein contains a phosphorylated threonine residue at position 61. A literature search for this protein turned up no useful information, suggesting that very little research has been done on this protein. The protein discloses no substantial sequence motifs. Disorder prediction for this protein shows a protein that is intrinsically disordered in the segments containing the first 12 amino acids and the last 8 amino acids, being highly structured throughout the rest of the protein. Due to the fact that the function of this protein is unknown it is difficult to determine if these short segments of intrinsic disorder will play a role in the functioning of this molecule. More studies need to be done to determine the role this protein plays within human cells.

3.4.4. *Up-regulated during skeletal muscle growth protein 5*

Up-regulated during skeletal muscle growth protein 5, also known as hepatitis C virus (HCV) F- transactivated protein 2, is a recently described, frame-shift product of the HCV core encoding sequence of genotype 1a.[89] Its function and antigenic properties are currently unknown. The protein sequence is 58 amino acids long, the sequence status is complete, and its existence has the evidence at the protein level. This protein is located is in the mitochondrion membrane and is thought to be a single-pass membrane protein. According to the disorder predictions, this protein is expected to have disordered N- and C-termini (the first 9 and the last 11 residues), being mostly structured throughout its central region.

3.4.5. *APM2*

Adipose most abundant gene transcript 2 (APM2), also known as C10orf116, is a 76 amino acid protein that is expressed in liver, cornea and adipose tissue and is encoded by a gene which maps to human chromosome 10. Chromosome 10 houses over 1,200 genes and comprises nearly 4.5% of the human genome. Several protein-coding genes, including those that encode for chemokines, cadherins, excision repair proteins, and early growth response factors (Egrs) and fibroblast growth receptors (FGFRs) are located on chromosome 10. Defects in some of the genes that map to chromosome 10 are associated with Charcot-Marie Tooth disease, Jackson-Weiss syndrome, Usher syndrome, nonsyndromatic deafness, Wolman’s syndrome, Cowden syndrome, multiple endocrine neoplasia type 2 and porphyria. It was shown that APM2 might be responsible for the promotion of the cisplatin resistance in the HCT116 cell line irrespectively of the status of two well-established determinants of cisplatin resistance, p53 and Mismatch-repair (MMR) protein.[90] This is an important finding since the cisplatin resistance hinders the effectiveness of this one of the most widely used chemotherapeutics in the world today. Intriguingly, APM2 is predicted to be mostly disordered suggesting an important role of intrinsic disorder in promoting the cisplatin resistance.

3.4.6. *SMPX*

Small muscular protein X-linked (SMPX), also known as Stretch-responsive skeletal muscle protein, is an 88 long amino acid protein that plays a role in the regulatory network through which muscle cells coordinate their structural and functional states during growth, adaptation, and repair.[91; 92] SMPX is preferentially and abundantly expressed in heart and skeletal muscle. Defects in SMPX are the cause of deafness X-linked type 4 (DFNX4), which is a non-syndromic form of progressive hearing loss.[93; 94] In affected males, the auditory impairment affects high frequency hearing first. It later becomes a severe form, which affects all frequencies. Carrier females manifest moderate hearing impairment in the high frequencies. Disorder predictions reveled that this protein is expected to be mostly disordered, suggesting that natural disorderedness might play a role in its functioning.

3.4.7. *Bg-II toxin*

The Bg-II toxin is a short polypeptide found within the sea anemone *Bunodosoma granulifera*.[95] Bg-II consists of 48 amino acid residues and shares a 71% sequence identity with the 49 residue-long polypeptide anthopleurin-A (also known as AP-A) from the sea anemone *Anthopleura xanthogrammica* (Giant Green Sea Anemone).[96] These two proteins are expected to possess similar compact structure consisting of four short strands of antiparallel β-sheet (in AP-A, residues 2-4, 20-23, 34-37, and 45-48) connected by three loops.[96] The structure of both AP-A and Bg-II is stabilized by three disulfide bonds that function to condense the molecule to some degree. The first loop, that accounts for ~30% of the protein (residues 5-19), was shown to be the least well-defined region of AP-A.[96] This poorly defined loop contains at least some of the residues considered to be essential for activity.[97] The other region that shows increased mobility involves residues 39-44, which forms a loop linking the third and fourth β-strands of the β-sheet.[96]

Despite some degree of disorder, Bg-II is a biologically functional cardio- and neuro-toxin. This toxin holds open voltage gated sodium channels, prolonging action potentials such that regeneration of new signals is slowed. This is accomplished via binding to site 3 on the alpha subunit of the channel, locking it into a position that increases its permeability to the sodium ion.[98] Searching for similarities between Bg-II and other cardiotoxins, such as those found in scorpions, has helped to build a greater understanding of the molecular basis behind such peptide-channel interactions. Although there was not enough sequence similarity between the two proteins to suggest relatedness, those homologies observed in the active sites appear to indicate convergent evolution.[95] When the functionally important regions were overlaid, they revealed that 4 basic residues and 1 tyrosine are similarly located within their sequences.[95] These 4 residues (asparagine, lysine, lysine, and asparagines) in both toxins form two groups of positive charges at either end of the protein. When the asparagine was changed to aspartic acid, toxicity decreased.[95] It can be assumed that the positive poles are responsible for the efficacy of the toxin. This may be because the IVS4 segment of site 3 on the channel to which Bg-II binds, is negatively charged.[98]

In agreement with the NMR data for homologous AP-A protein, disorder prediction revealed that the N-terminal half of Bg-II is expected to be highly disordered. This is where the arginine residues proposed responsible for the formation of the polar poles mentioned earlier are located. The disorder seen within the peptide could assist in protein function by increasing flexibility. Given that a channel will open and close, it only makes sense to have a binding molecule that exhibit a similar degree of movement. Rigidity would decrease its ability to properly bind and therefore, its toxicity. Disorder decreases towards the center of the sequence around hydrophobic residues like leucine, tryptophan, and tyrosine. Increased order in conjunction with hydrophobic residues, suggest that these amino acids may be internalized from the hydrophilic environment. Flexibility in the form of disorder may enhance functionality of these poles.

3.4.8. *Cytotoxin 2*

The cytotoxin 2 CX2_NAJME is a polypeptide of 61 amino acid residues found within the venom of *Naja melanoleuca* (Black-lipped cobra).[99] This particular toxin characteristically causes hemolysis and cardiac cell death.[100] It does so by binding to both the lipid bilayer of normal cells and voltage gated ion channels in cardiac cells.[101] Hemolysis is the phenomenon of bursting cells, which, being left untreated, results in tissue necrosis. Binding of the toxin to the ion channels in cardiomyocytes causes constant depolarization within the cell, preventing repolarization from taking place. Hence, new action potentials cannot be generated and without proper treatment, the heart may cease to beat entirely.

Cytotoxin-2 is a member of the snake three-finger toxin family.[100] Structurally, it contains a three-loop β-sheet motif resulting in a sort of hydrophobic tripod structure responsible for penetrating the surface of the cell membrane.[101] Additionally, preceding the hydrophobic tips, there are positively charged amino acids that are attracted to the negatively charged fatty acids in the lipid bilayer, increasing the binding force between the two surfaces.[101] The extent to which the toxin will permeate the membrane depends on the belt of polar lysine and arginine residues around the peptide.[100]

Interestingly, this particular toxin has a substitution of neutral glycine and leucine at positions 25 and 27 in place of methionine.[99] The substitution is a proposed cause of its reduced toxicity, a viable solution given that one of the hydrophobic residues is being changed to a neutral charge. This may ultimately reduce the affinity of the three-loop structure for the cell membrane, thereby decreasing its ability to disrupt the lipid bilayer.

Disorder prediction for Cytotoxin-2 revealed that this toxin is expected to be disordered at the very beginning and very end of its sequence. The grand majority of the sequence seems to exhibit a very high degree of order. This central part of the sequence contains a handful of hydrophobic lysine residues. These residues may function to create the previously mentioned tips that assist in penetrating the cell membrane. The disorder at the beginning and end of the sequence may lend a degree of flexibility, which aids not only in binding but also in penetrating the lipid bilayer. It could allow the rigid hydrophobic structure of toxin to better open up the cell membrane resulting in hemolysis.

3.4.9. *HIRV2*

The small protein hirudin variant-2 (HIRV2) consists of 72 amino acid residues and is found in the medicinal leech *Hirundo medicinalis* (medicinal leech). It functions as a protease inhibitor that modulates clotting by binding to α-thrombin.[102] The thrombin protein is essential to the liberation of fibrin monomers from fibrinogen, a process ultimately responsible for blood coagulation. HIRV2 residues Ile-8, Thr-9, and Tyr-10 form bonds with the thrombin active site. Residues 62 through 72 interact with the fibrinogen binding site of thrombin.[102] Interestingly, although HIRV2 is a protease inhibitor, it has a unique way of inhibiting thrombin function independent of active site interaction. It is the *amount* of interaction rather than the location that is responsible for such strong inhibition and specificity.

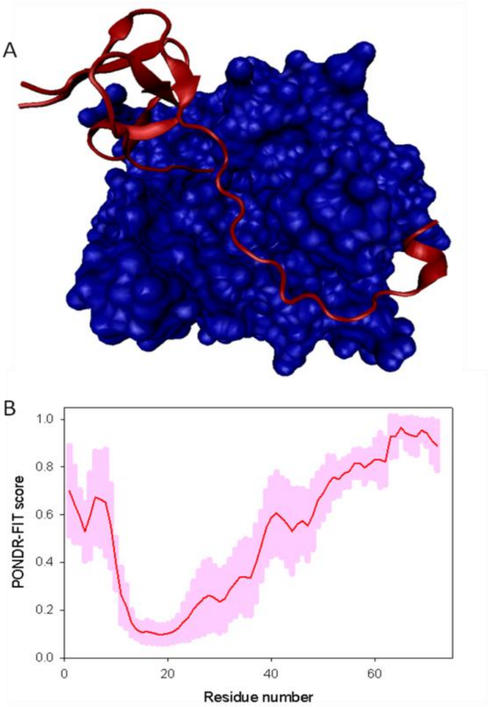

Figure 6. Intrinsic disorder in hirudin. **A**. Crystal structure of the hirudin-thrombin complex (PDB ID: 4HTC), where hirudin binds to thrombin non-covalently over an area of approximately 1400 Å, therefore representing an illustrative example of wrapping-upon-binding mode of an IDP interaction with its binding partner (see the text for more details). **B**. Intrinsic disorder propensity distribution in hiruding evaluated by PONDR-FIT (red line). Light pink shadow around PONDR-FIT prediction shows the statistical error of PONDR-FIT prediction.

Unlike other members of this protein family, which have binding areas of roughly 400-500 Å.[102], hirudin binds to thrombin non-covalently over an area of approximately 1400 Å.[102] Figure 6A illustrates this wrapping-upon-binding mode which covers an area that is almost three times as much as other members of this protein family. Such kind of wrapping interaction is typical for the intrinsic disorder-based binding.[103] Figure 6A shows that the bound structure of the HIRV2 contains a long tail and shows the remaining part of the protein as four short β-strands connected by short bends. In agreement with this structure, Figure 6B shows that HIRV2 is predicted to have an ordered central domain (residues 10-38) and possess long disordered tails (residues 1-9 and 39-72). The asymmetry and disorder of this HIRV2 bound structure may lend to its intricate binding properties. The disordered nature of this protein confers a considerable amount of flexibility useful for blanketing the thrombin. Hirundin can therefore physically restrain the thrombin from breaking down fibrinogen. The flexibility in conjunction with the primary sequence also function together to make this blanketing (or wrapping) inhibition highly specific to the thrombin protein. Because this mode of interaction involves covering thrombin rather than reacting with its active site, this is a proposed reason why hirundin inhibits no other serine proteases.[102]

3.4.10. *Small proline-rich protein 2A*

Small proline-rich protein 2A is a 72 amino acid keratinocyte protein that first appears in the cell cytosol before, ultimately, becoming cross-linked to the membrane proteins by transglutaminase. This, in turn, results in the formation of an insoluble envelope underneath the plasma membrane. The most distinctive feature of the small proline-rich protein gene family lies in the central segments of the encoded polypeptides that are assembled from tandemly repeated units of either eight or nine amino acids with the general consensus *K*PEP**.[104] Sequencing data of the different members of this family combined with their chromosomal organization strongly indicated that this gene family evolved from a single ancestor gene as a result of several intra- and intergenic duplications.[105] Analysis of the different subfamilies indicated that a process of homogenization has acted on the different members of one subfamily. At the levels of both protein structure and gene regulation however, the different subfamilies appear to have diverged from one another.[105] Human small proline-rich protein 2A is predicted to be mostly disordered.

3.4.11. *SUMO4*

Small ubiquitin-related modifier 4 (SUMO4) is a 95 amino acid ubiquitin-like protein, which can be covalently attached to lysines of target proteins as a monomer. It is primarily expressed at various levels in immune tissues, with highest expression being found in the lymph nodes and the spleen. One particular polymorphism of interest related to disease is the variant Val-55, which is believed to be associated with insulin-dependent diabetes mellitus. It does not seem to be involved in protein degradation and may modulate protein subcellular localization, stability or activity. When cell undergoes oxidative stress, SUMO4 demonstrates the ability to conjugate with various anti-oxidant enzymes, chaperones, and stress defense proteins. In addition to these proteins, it may also conjugate with NFKBIA, TFAP2A and FOS, which negatively regulate transcriptional activity. Another protein it conjugates with, NR3C1, is able to positively regulate its transcriptional activity. Covalent attachment to its substrates requires prior activation by the E1 complex SAE1-SAE2 and linkage to the E2 enzyme UBE2I.[106] This ability to covalently bond with and interact with a diverse array of different proteins indicates that the SUMO4 may possess intrinsic disorder. In agreement with this hypothesis the disorder prediction revealed that 20 N-terminal and 30 C-terminal residues of SUMO4 are expected to be disordered, shown that ~50% of this protein is characterized by the highly flexible structure.

3.4.12. *TRIA1*

p53 is a well-known tumor suppressor whose mutation is implicated in more than half of all cancers. As a result, a lot of research has been focused on characterizing p53 and the proteins it interacts with that are involved in the cancer pathway. Despite the efforts made, the mechanisms of the p53-dependent activities that determine cellular survival or death are still not fully understood. A 76 amino acid novel p53 target protein was recently found and named the TP53-regulated inhibitor of apoptosis (TRIA1).107 It is believed that TRIA1 (also known as p53-inducible cell-survival factor) is one of the important players in the p53-mediated cell survival mechanism. This protein functions by facilitating cell survival through the inhibition of caspase-9 activation. By preventing this activation, this protein is successfully able to prevent the initiation of apoptosis 1. This would be a favorable protein to interact with the p53 protein because it would aid in tumor suppression by aiding in the elimination of identified cancerous cells. It is induced in response to low levels of DNA damage, however, and not when damage is severe, which may prevent the killing of potentially dangerous further developed cancer cells.107 In order to interact with the highly intrinsically disordered p53 protein.34, this inhibitor of apoptosis likely contains some regions of intrinsic unfolding. This hypothesis was supported by the results of the analysis of disorder distribution within TRIA1 which showed the disordered nature of its N- and C-terminal regions (residues 1-11 and 61-76).

3.4.13. *Vpr*

The HIV-1 viral protein R (Vpr) is a 96 amino acid protein with the molecular mass of 14kDa. The specific sequence analyzed here was isolated from group M subtype B (isolate BRU/LAI) (HIV-1). The function of Vpr in HIV parthenogenesis is to regulate transport of the HIV pre-integration complex into the nucleus of the host cell.[108] It contains two separate nuclear localization activities associated with the 1-73 and 73-96 regions.[109] These activities do not depend on the traditional nuclear localization signals but instead give viral protein R direct access to the nuclear pore complex.[109] Vpr is also known to create pores in the membranes of neurons that function as cation channels and to permeabilize the mitochondrial membrane through a pore forming interaction with the adenine nucleotide translocator.[110; 111] In addition to its pore forming activities, Vpr also functions to arrest the cell cycle of proliferating cells at the G_2/M transition.[112] It does this by activating Wee1 and inactivating Cdc2 which both result in the production of inactive Cdk-1.[113]

Vpr is a pleiotropic protein that has a variety of roles in determining HIV-1 infectivity, as it is involved in apoptosis, cell cycle arrest, and dysregulation of immune functions.[114-117] Among other important biological functions of Vpr are the modulation of transcription of the virus genome,[118] the induction of defects in mitosis,[119] the facilitation of reverse transcription,[120] the suppression of immune activation,[121] the reduction of the HIV mutation rate,[122] and ion channel formation and cytopathogenicity.[110; 123]

The solution structure of Vpr was determined by NMR.[124] It consists of three well-defined α-helices (residues 17-33, 38-50, and 56-77) surrounded by flexible N- and C-terminal domains. Ordered helical structures are common in proteins that cross or form pores in plasma membranes so this is not unexpected. The helical region closest to the C terminus has been determined to be important in the protein's nuclear localization and apoptosis mediating functions.[113; 125] This helix may also allow the protein to form homodimers.[126]

Disorder predictions by several algorithms revealed that both the N-terminal (residues 1-17) and the C-terminal tails (residues 75-96) are likely to be disordered. These regions correspond to the disordered flexible domains detected in the NMR structure of Vpr. This disorder is probably a contributing factor to the ability of Vpr to bind to multiple partners and carry out the multiple cellular functions described above. Various roles of intrinsically disordered regions in function of Vpr were considered in a recent review dedicated to the analysis of the structural peculiarities of HIV-1 proteins, the abundance of intrinsic disorder in viral proteins, and the role of intrinsic disorder in their functions.[52]

3.4.14. *TIM10*

Mitochondrial import inner membrane translocase subunit Tim10 (TIM10) is a 93 amino acid protein with the molecular mass of 10.3 kDa. The specific sequence analyzed in this study was taken from *Saccharomyces cerevisiae* (strain ATCC 204508 / S288c). As its name implies, the mitochondrial import inner membrane translocase complex is found on the inner membrane of the mitochondria. It is a heterohexamer composed of 3 TIM9 subunits and 3 TIM10 subunits.[127; 128] The complex functions as a chaperone, guiding the import and insertion of transmembrane proteins into the inner mitochondrial membrane.[127] It binds to the hydrophobic membrane spanning regions of these proteins preventing them from aggregating in the intramembrane space.[127] The TIM9 and TIM10 proteins can also exist outside of the complex and still facilitate the integration of intramembrane proteins through interactions with the TIM22 complex.[129] The TIM9/TIM10 complex is also essential for the assembly of TOM40, a component of the TOM complex which facilitates protein transport across the outer mitochondrial membrane.[130]

The TIM10 protein itself contains one N-terminal and one C-terminal α-helical region joined by a loop.[128] The central loop of each subunit interacts with an α-helical region of another subunit forming the quaternary structure of the complex as shown in the figure below.[128] The PONDR-FIT analysis reveals that this loop region is predicted to be more disordered than adjacent regions. This area corresponds to a twin CX3C motif in which the ends of the loop come together in 3D space and form two disulfide bonds.[131] The increased disorder of the loop may impart flexibility to the structure that allows for necessary movement during complex formation. The N-terminal region of TIM10, in the area before the α-helix structure begins, functions as a substrate sensor for the complex.[132] This also corresponds to a disordered region as predicted by PONDR-FIT. The disorder in this region may function in the ability of the protein to the different transmembrane proteins that the TIM9/TIM10 complex is known to associate with.

3.4.15. *Ice structuring protein A*

Ice structuring protein A is an 82 amino acid protein with the molecular mass of 7.7kDa. The specific sequence analyzed here was taken from the winter flounder (*Pseudopleuronectes americanus*). The protein is cleaved *in vivo* to produce its active 37 amino acid form.[133] This sequence is composed of amino acids 45-81 from the original protein. Its function *in vivo* is the binding of ice crystals as they form in the blood of the fish in order to lower the blood's freezing point. In order to carry out its function, the protein must be secreted into the bloodstream. Its first 23 amino acids at the N-terminal region compose a secretory signal sequence that is part of the region which is removed during cleavage to the active form.[134]

Many different proteins with antifreeze properties have been isolated from coldwater fish.[135] The proteins can be placed into five categories based on their structure. Ice-structuring protein A has been labeled a type one antifreeze protein because of its α-helical structure.[135] In the energy optimized structure determined by molecular modeling simulations, ice-structuring protein A is seeing as an alanine rich α-helix containing four threonine resides that take on a specific spatial orientation. The side chains of these threonine residues are able to hydrogen bond with the water molecules in newly forming ice crystals and interrupt the crystal growth process.[133] Because of this, the 3-D structure of the entire helix that, in turn, determines the specific orientation of the threonine resides must be maintained for the protein to function. However, another study involving NMR spectroscopy has determined that significant bending of the helix and rotation of the side chains actually takes place when the protein is in solution.[136]

Disorder predictions for this protein reveal that at least one region of the protein is likely to be disordered, the sequence spanning amino acids 19-46. It is worth noting that residue number 45 is where the functional α-helical structure of the protein begins. Therefore, any intrinsic disorder present in the adjacent 19-46 region may be involved in the proteolytic cleavage that takes place when the N terminal half of the protein is removed and the final active form of the protein is generated. The significant disorder was also predicted in the C terminal region of the protein, where the α-helix is located. A possible explanations for this somewhat counterintuitive result is that the α-helix structure is not known to be completely stable. It contains a central bend that may provide flexibility in solution and is exaggerated in the X-ray crystallographic structure of the protein.[136]

3.4.16. *Acylphosphatase*

Acylphosphatase is a small cytosolic enzyme. It belongs to the acylphosphatase family and contains one acylphosphatase-like domain. Acylphosphatase catalyzes the hydrolysis of the carboxyl-phosphate bond of acylphosphates. The reaction involves acylphosphate and water, which yields a carboxylate and a phosphate. Two isoenzymes have been isolated, muscle acylphosphatase and erythrocyte acylphosphatase. They are named on the basis of their tissue localization. This gene encodes the erythrocyte acylphosphatase isoenzyme. Alternatively spliced transcript variants that encode different proteins were identified through data analysis. Acylphosphatase has active sites at amino acids 24 and 42 and is 99 amino acids in length. Its physiological role is not completely clear as of yet. Organ-common type isozyme is found in many different tissues. Disorder prediction for this protein reveals disordered N- and C-terminal regions, with the central region being predicted to be highly ordered, an expected behavior for the small enzyme.

3.5. ***Structure, function, and intrinsic disorder of some long proteins***

3.5.1. *FAT tumor suppressor homolog 4*

FAT tumor suppressor homolog 4 (Fat4) is a protein that is encoded for by the Fat4 gene. It is a large protein, which is encoded for by 4,981 amino acids and has a molecular mass of 542.687 kDa. Fat4 is located on chromosome 4 in the human genome.[137] The protein has three isoforms, which are produced by alternative splicing. Isoform 3 is significantly smaller than the other two isoforms, containing only 3,222 amino acids and having a molecular mass of only 351.191 kDa. Fat4 is a member of the Fat family that belongs to the cadherin family, which is involved in cell adhesion and consists of more than 80 members in mammalian species.[137] There are four members of Fat family in mice and humans, which share high homology to the *Drosophila* Fat gene.[137] Fat is most likely to function at the apical point of the Hippo signaling as a potential transmembrane receptor in Drosophila.[138] The Hippo signaling pathway plays an important role in controlling organ size.[139] Fat is also involved in the planar polarity formation, which refers to the asymmetry of a cell within the plane of the epithelium.[139]

Several small regions from the cytoplasmic region of Fat4 are highly conserved in the *Drosophila* Fat, suggesting that human Fat4 is a homolog of the *Drosophila* Fat.[137] A critical role of Fat4 in tumorigenesis has been demonstrated, with the loss of Fat4 expression having been found to occur in some primary breast tumors and breast cell lines.[137] It was found that the loss of Fat4 in non-tumorigenic mammary epithelial cells transformed the cells into tumorigenic cells and restoring Fat4 in these cells inhibited their ability to form tumors, implicating its important role in breast tumorigenesis.[137]

There is evidence that Fat4 is essential gene with a key role in vertebrate planar cell polarity (PCP).[140; 141] The loss of Fat4 has been shown to disrupt oriented cell divisions and tubule elongation during kidney development, leading to cystic kidney disease.[140; 141] Fat4 genetically interacts with the PCP genes Vangl2 and Fjx1 in cyst formation, suggesting that Fat4 plays a role in regulating vertebrate PCP and that loss of PCP signaling may help cause cystic diseases in humans.[140; 141]

The Fat4 protein contains three different types of conserved domains. Cadherin repeat domains, laminin G domains, and calcium-binding EGF-like domains. All of these domains are involved with cell-to-cell calcium signaling. These domains suggest that Fat4 is a transmembrane protein with both intracellular and extracellular domains . Disorder predictions showed that this protein has long regions of intrinsic disorder, with high disorder scores expected for the first 43 amino acids and throughout the last 70 amino acids. In agreement with its transmembrane nature, the central part of this protein is predicted to be highly ordered. Though the exact molecular mechanisms of this protein are still unknown, the fact that it is likely a transmembrane protein with long disordered N- and C-terminal tails make it likely that intrinsic disorder plays a role in the functioning of this protein by giving flexibility to the intracellular and extracellular ends and allowing them to perform their specific functions while being incorporated into the membrane.

3.5.2. *Ovarian cancer-related tumor marker CA125*

Ovarian cancer-related tumor marker CA125 is a mucin protein that is encoded by the MUC16 Mucin gene.[142] MUC16 has been shown to map to chromosome 19p13.2.[142] The protein has a length of 22,152 amino acids and a molecular weight of 2,353.428 kDa. CA125 is a tumor antigen that forms the basis for a serum assay that has been widely used in the monitoring of ovarian since its discovery in 1981.[143] A radioimmunoassay for the antigen was subsequently developed, application of which showed that serum CA125 levels are elevated in about 80% of women with epithelial ovarian cancer, but is elevated in less than 1% of healthy women.[144] A rise in CA125 levels usually precedes clinical detection of ovarian cancer by about 3 months.[142] Therefore, CA125 is considered to be one of the best available cancer serum markers.[142] Furthermore, in chemotherapy, rising CA125 levels during or post treatment often indicate an insufficient response to the therapy and an unfavorable prognosis.[145; 146] However, CA125 has shown to be of limited use in the initial diagnosis of ovarian cancer because of its elevation in some benign conditions.[145; 146] It has also been found that the induction of anti-CA125 responses in ovarian cancer patients leads to prolonged survival time compared with untreated patients.[147] This suggests that CA125 could act as a targeting antigen to elicit antibody-dependent, cell-mediated cytotoxicity against ovarian cancer cells or that it plays a key physiological role that promotes tumor development.[147] Furthering the evidence for CA125 playing a key role in tumorigenesis, it has been shown to bind selectively to mesothelin, a protein that is expressed on all normal mesothelial cells and that could provide a point of contact between metastasizing ovarian cancer cells and the normal mesothelial cells which line the peritoneum.[148; 149]

The normal function of CA125 remains unknown, but it is found normally in seminal fluid, the fallopian tube and endometrium, which is consistent with a role in reproduction.[149] CA125 is a membrane spanning molecule with a heavily glycosylated amino terminal extracellular domain, an extracellular tandem repeat domain, and a carboxy terminal domain with a transmembrane region and a short cytoplasmic tail.[148-150]

The repeats in CA125 are characteristic of mucins and are each composed of 156 amino acids. Different Ca125 variants can have 7, 20, or 60 of these repeats.[148] The repeats house varying types of SEA domains, which are so named for the organisms that they were first found in: S = sea urchin, E = enterokinase, and A = agrin.[150] These SEA domains are believed to comprise sites for proteolytic cleavage, but unlike the well-characterized MUC1, the SEA domains within CA125 are not all similar and do not all have the proteolytic domain sequence typical of most mucins.[150] The N-terminal domain is of particular interest in mucin proteins, as it contains glycosylation sites and the heavy glycosylation renders CA125 extremely greasy and therefore able to lubricate surfaces and modulate cellular adhesion.[150] The addition of the phosphate group at the C-terminal tyrosine phosphorylation site creates a conformational change, which switches CA125 from its inactive state to its active state.[150] CA125 is phosphorylated when it is in the cell and dephosphorylated prior to its release from the cell.[150] Due to the fact that there are so many possible cleavage sites for this protein, it is also possible that there are many different phenotypes of the ovarian cancer it is implicated in. Determining how these variations affect the prognosis and treatment of ovarian cancer may help to allow for more precise decision making in the future.

Disorder predictions for CA125 show a high percentage of disordered residues in this protein. The first 12,071 residues corresponding to the extracellular N-terminal and tandem repeat domains form a large, continuous, highly disordered segment. The rest of the protein shows some smaller segments of disorder, but this region of the protein is much more structured than the N-terminal part. This is most likely due to the fact that this part of the protein contains a transmembrane domain, which would have to be structured to span the membrane. The tandem repeat domain most likely shows a very high abundance of disorder due to the fact that it contains many potential splice sites that would have accessible to cleavage proteins.

3.5.3. *Reelin*

Reelin is a protein with a length of 3,460 amino acids and a molecular mass of 388.388 kDa. Gene that encodes Reelin, *RELN*, is mapped to the human chromosome 7q22.[151] *RELN* is expressed in fetal and postnatal brain as well as in liver, with the expression of *RELN* in postnatal human brain being high in the cerebellum.[151] In mice, Reelin is an extracellular matrix molecule, a large glycoprotein secreted by Cajal-Retzius cells during early cortical development, which is required to control proper migration and positioning of cortical neurons.[152] It has been thought that Reelin functions as a neuron-specific extracellular protein that controls neuronal migration in the developing brain and stabilizes the architecture of laminar structures.[151]

Lack of Reelin expression in mice results in the reeler phenotype, which is associated with a both pronounced neurological symptoms and striking abnormalities in the architecture of the telencephalic and cerebellar cortices.[151; 152] The apolipoprotein E receptor 2 (Apoer2) and the very low-density lipoprotein receptor (Vldlr) were identified as Reelin receptors in mice.[153] With each individual receptor exerting a different function in neuronal positioning.[154] Research has revealed that the predicted mouse and human proteins are similar in size and amino acid composition, with the protein and nucleotide sequences having 94.2% and 87.2% sequence identity, respectively.[151]

In the human cerebral cortex, Reelin deficiency is accompanied by neuronal migration defects leading to the phenotype of lissencephaly or smooth brain.[155] Reelin has been shown to promote extension of dendritic processes and maturation of dendritic spines during cortical development.[156] and to play a role in modulating synaptic function in the mature brain.[157] A variety of neurological and psychiatric disorders, including schizophrenia, autism, major depression, temporal lobe epilepsy and Alzheimer's disease have been associated with decreased Reelin expression.[158] Reelin is obviously a complex molecule with a role in both developmental and adult network functions. Progress is being made in the understanding Reelin function and signaling at different developmental stages, but much more research needs to be done before we fully understand this molecule.

The Reelin protein appears to be extracellular matrix molecule containing an amino-terminal region with homology to F-spondin, which is a protein expressed and secreted at high levels in a cell group, the floor plate, that plays a critical role in the control of neural cell pattern and axonal growth in the developing nervous system.[151-158] The amino terminal region is followed by a hinge region upstream from eight repeats of 350–390 amino acids, with each repeat being composed of two sub repeats separated by an epidermal growth factor (EGF)-like motif.[151] Reelin ends with a highly basic C-terminus.[151] Disorder predictions of this protein show a highly ordered structure throughout the whole amino acid sequence. In fact, according to different algorithms, Reelin has from 2.5 to 7% disordered residues, with the longest disordered region being of 55 residues (in the 2142-2196 region). Thus, it is highly unlikely that intrinsic disorder plays any role in its function. This is an exciting molecule that, with more understanding, may lead to important breakthroughs in our understanding of various neurodevelopmental and neurodegenerative disorders.

3.5.4. *Dystrophin*

The mouse dystrophin protein consists of 3,678 amino acid residues and functions to attach the extracellular matrix to the cytoskelaton within muscle tissue.[159] It is also present in varying isoforms during different developmental stages and in other tissues including the heart and murine brain tissue.[160] Dystrophin has a domain of residues 1-240 that interacts with F-actin to reinforce the structure of a muscle fiber cell, preventing it from being altered by the force of its own contraction.[159]

Defects in the formation of a functional dystrophin protein may result in diseases such as muscular dystrophy affecting but not limited skeletal muscle, heart, and brain tissue.[160] Symptoms include muscle weakness, cardiovascular abnormalities, and shortness of breath. Some forms of mental retardation have also been associated with defects in the dystrophin protein.[160] These symptoms are a direct result of the diminished structural integrity of cells arising from these defects.

According to the disorder predictors, more than half of the dystrophin molecule exhibits a high degree of order. The disordered regions in the dystrophin protein may aid in binding to intracellular elements. Given that the protein is particularly large, it adopts a fairly rigid structure that confers cellular stability. Hence, disordered regions of loops connecting the various helices within the dystrophin allow it greater leniency when trying to find a bond with another molecule within the cell.

3.5.5. *Filaggrin*

Filaggrin found in humans is a 4,061 amino acid residue sequence. Interacting with keratin, it functions to create the epithelial barrier of the skin.[161] At the cellular level, filaggrin can be seen concentrated towards the top of the epidermis, while keratin is distributed throughout the underlying layers of skin. This differentiation is evidence for the suppression of filaggrin transcription and translational modification as it ascends to the epithelial surface.[161] The filaggrin molecule itself is functional in its ability to interact with keratin only after its linker regions have been removed via proteolysis.[161] Defects concerning filaggrin tend to result in forms of hyperkeratinosis in which the epidermal layer becomes thicker and tougher in appearance.[162]

Examination of the primary sequence has revealed the first 92 amino acid residues sharing 78% identity and 86% homology with the last 92 residues.[161] There also exist conserved aliphatic regions at residues 23-38 and 347-362 followed by sequences homologous to mouse filaggrin.[161] The homology may imply a conservation of function. The aliphatic regions may play a role in interaction with keratin.

According to the intrinsic disorder prediction, the first third of the filaggrin protein is extremely disordered. The last two thirds however exhibit a high degree of order. As filaggrin is a protein associated with the formation of the epidermal layer, a disordered region would grant it the flexibility for physical movement associated with the elastic nature of skin. The other, longer, ordered region may provide the structural support via interaction with keratin.

3.5.6. *Twitchin*

The protein twitchin consists of 7,158 amino acid residues and can be found in the worm *Caenorhabditis elegans*.[163] Twitchin is a large titin-like protein made of many immunoglobin and fibronectin type domains. It also contains one kinase domain near the C-terminal end of the protein, proposed to be both a "regulator of muscle contraction" as well as a sensor of muscle movement.[164] It affects muscle contraction by prolonging relaxation of muscle movement. In terms of spatial structure, this kinase domain is shoved between two subdomains of the “catalytic core” to which it strongly interacts, resulting in a considerable degree of inhibition.[164] This type of inhibition is called autoinhibition- where structures within the molecule physically block it from its active conformation.[164]

One proposed cause of “giant kinase” activation is the sheer force of mechanical movement.[164] The extent and intensity of movement may be ebough to physically detach the kinase domain from its “wedged” position.[164] Research has found that this protein is able to handle physical stretching without breaking apart and is still able to return to its original shape. This kinase domain is also positioned with the “molecular spring” of the muscle to act as an ideal force sensor.[164] It is the position of the two sets of β-sheets within the molecule that is responsible for this type of flexibility and structural stability. Those sheets resisting a “pulling force” are positioned parallel to said force; while those sheets exposing the kinase domain are located perpendicular to this force.[164]

According to the disorder prediction algorithms, the N-terminal region of twitchin appears to be significantly ordered with some noticeable peaks of predicted disorder. These interspersed peaks of disorder may indicate regions that confer a degree of flexibility to the protein. According to the current model of the twitchin action, there is a significant amount of unwinding and stretching that occurs at the two ends of the protein. Disordered regions may give the twitchin the flexibility to move while still having an order in its primary sequence that would allow it to return to its original shape after the force is removed. This change in shape may be the driving impulse of the ‘sensor’ function.

3.5.7. *Nucleoporin RanBP2*

RanBP2 is a nucleoporin that has E3 SUMO-protein ligase activity and mediates SUMO1 and SUMO2 conjugation by UBE2I.[165] RanBP2 is comprised of 3,224 amino acid residues and it has multiple unrelated functions. For example, this protein is involved in transport factor (Ran-GTP, karyopherin)-mediated protein import via the FG repeat-containing domain which acts as a docking site for substrates.[166] RanBP2 is believed to be a component of the nuclear export pathway and contain a specific docking site for the nuclear export factor exportin-1.[166] It contains a GTPase-binding domain.[167], promotes modification of the HDAC4 deacetylase.[168], possesses isomerase and/or chaperone activity.[169] and may act as a tumor suppressor.[170]

According to various disorder predictors, RanBP2 is expected to contain up to 50% disordered residues. This finding is in a good agreement with recent temperature-dependent enzyme kinetic analysis that has revealed that E3 SUMO-protein ligase RanBP2 confers unusually large and favorable activation entropy to lower the activation energy of the reaction.[171] This entropy-driven mechanism of the E3 is in line with the lack of sequential and structural conservation among E3’s despite their similar functions.

These experiments also illustrate that intrinsically disordered proteins could enhance thermodynamic chemistry and have a role in determining the protein–protein interactions.[171]

3.5.8. *E3 ubiquitin-protein ligase HERC2*

E3 ubiquitin-protein ligase HERC2 controls ubiquitin-dependent retention of repair proteins on damaged chromosomes.[172] HERC2 ubiquitin ligase is involved in the circadian control of XPA and excision repair of cisplatin-DNA damage.[173]

HERC2 contains 5,233 amino acids and is encoded by the *HERC2* gene. Genetic variations in this gene are associated with skin, hair and eye pigmentation variability. For example, a mutation in the HERC2 gene adjacent to the OCA2 gene, which affects OCA2's expression in the human iris, is found in almost all people with blue eyes.[174] HERC2 is predicted to contain ~30% disordered residues forming a number of short, medium and long disordered regions. In fact, HERC2 possesses up to 75 disordered regions ranging in length from 5 to 115 residues.

3.5.9. *BRCA2*

BRCA2 is one of the two major breast cancer susceptibility genes. The product of this gene is a large protein that consists of 3,418 amino acid residues. In the cell, BRCA2 severs as breast cancer tumor suppressor being involved in the repair of double strand breaks and broken replication forks by homologous recombination through its interaction with DNA repair protein Rad51. Here, BRCA2 aids assembly of Rad51 onto single-stranded DNA and acts by stabilizing Rad51 and single-stranded DNA filaments via the blocking of ATP hydrolysis.[175] Furthermore, C-terminal domain of BRCA2 is involved in interaction with Fanconi anemia (FA, which is an autosomal recessive cancer susceptibility syndrome) complementation group D2 (FANCD2) protein.[176; 177] There are eight FA proteins that cooperate in a common pathway, and of the eight genes related to the FA, the *FANCD1* gene is identical to the breast cancer susceptibility gene, BRCA2,178 which is able to interact with both non-ubiquitinated and mono-ubiquitinated FANCD2. BRCA2 has been observed in both monomeric and dimeric form and is also found as a part of a trimeric complex, which consists of BRCA1, BRCA2 and PALB2.[175]

Defects in BRCA2 have been known to lead to malignancy originating from breast epithelial tissue.[179] and also to be a cause of susceptibility to breast-ovarian cancer familial type 2, which is a condition associated with familial predisposition to cancer of the breast and ovaries.[180; 181] BRCA2 mutations are also associated with the onset of pancreatic cancer type 2, which is a malignant neoplasm of the pancreas developed from both the exocrine and endocrine portions of the pancreas, with the 95% of tumors developing from the exocrine portion.[182]

The *BRCA2* tumor suppressor gene is implicated in many cellular pathways including transcription; cell-cycle checkpoint control, apoptosis and DNA repair.[175] BRCA2 is predicted to contain up to 50% disordered residues. According to this analysis, there are at least 20 long disordered regions in this protein ranging in length from 50 to 250 residues.

3.5.10. *BRCA1*

Talking about BRCA2 it is difficult not to mention another famous member of the BRCA family, the breast cancer type 1 susceptibility protein (BRCA1). This is because of the fact that the differential response of this protein to different types of DNA damage represents an excellent example of how IDPs can modify signal flow. BRCA1 is involved in many diverse biological signaling processes such as DNA damage response (DDR), transcription and cell-cycle checkpoint control, tumor suppression, oncogenesis, stress response and apoptosis.[183; 184] Concomitant with its role in DDR, BRCA1 has been implicated in a variety of different cancers.[184]

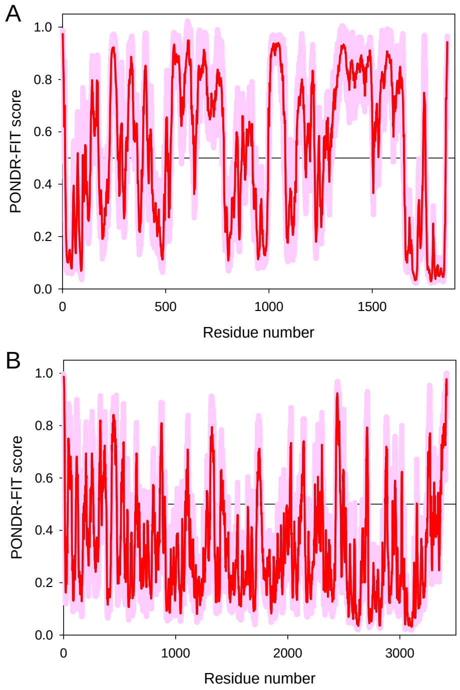

Figure 7. Intrinsic disorder propensity distribution in BRCA1 (**A**) and BRCA2 (**B**) evaluated by PONDR-FIT (red line). Light pink shadow around PONDR-FIT predictions show the statistical errors of PONDR-FIT predictions.

BRCA1 is a 1,863 residue-long protein that consists of two structured domains located at the N- and C-termini that are separated by a large internal ID region that comprises 79% of the residues. This 1,480 amino acid central region was shown by NMR and CD spectroscopy to be disordered.[185] In agreement with these experimental data, 86% of BRCA1 residues are predicted to be disordered. The disordered central region contains binding sites for DNA as well as several protein partners such as p53, retinoblastoma protein, BRCA2; the oncogenes c-Myc and JunB; DNA damage repair proteins such as Rad50 and Rad51; and the Fanconi anemia group A protein.[185] Overall, more than 50 proteins interact with BRCA1, including a variety of DNA damage sensors, DNA repair proteins and signal transducers.[183; 184] The vast majority of these interactions occur in the long central IDR.

Figure 7 represents the results of the PONDR-FIT analysis of these two members of the BRCA family and clearly shows that both, BRCA1 (Figure 7A) and BRCA2 (Figure 7B) are predicted to have significant amount of intrinsic disorder. It is also important to note that the level of predicted intrinsic disorder seeing in BRCA1 agrees well with accumulated experimental data.

3.5.11. *The IgGFc-binding protein*

The IgGFc-binding protein consists of 5,405 amino acids and it is involved in the maintenance of the mucosal structure as a gel-like component of the mucosa.[186] It is mainly expressed in the placenta and in colon epithelium and interacts with the Fc portion of IgG and with Muc2 mucin.[187] FcGF binding protein is believed to play an important role in immune protection and inflammation in the intestines. It is present in higher concentrations in patients with various autoimmune diseases.[188]

Several oncogenic rearrangements and mutations are believed to be responsible for the development of thyroid papillary carcinomas, follicular adenomas, and carcinomas. It is difficult to distinguish between the events that are responsible for each individual carcinoma. Quantitative real-time PCR helped confirm the differential expression. Experiments showed that IgGFc-binding protein is differentially expressed in normal thyroid tissue, thyroid adenomas and thyroid carcinomas.[189] The IgGFc-binding protein gene is constitutively expressed in normal thyroid tissue. However, its expression is significantly increased in follicular thyroid adenomas and significantly decreased in papillary and follicular thyroid carcinomas. As a result, measurement of the expression levels of IgGFc-binding protein in thyroid biopsies helps to make the distinction between a thyroid follicular adenoma and a follicular carcinoma.[189]

Analysis of the IgGFc-binding protein by several disorder predictors revealed that this protein contains up to 30% disordered residues. The large portion of disordered residues is condensed in ~25 long disordered regions, with the longest IDR including 125 residues.

3.5.12. *HC-toxin synthetase*

HC-toxin synthetase 1 (HTS1) is a 5,218 amino acid protein with a molecular mass of 574 kDa. It is found in the organism *Cochliobolus carbonum* (*Bipolaris zeicola*), a filamentous fungal pathogen affecting many plants including corn.[190; 191] This protein is crucial for maintenance of the fungal virulence as it functions to produce peptides that are toxic to the host maize plant, including a cyclic tetrapeptide, HC-toxin.[190] Therefore, HTS1 is a cyclic tetrapeptide synthetase, which is non-ribosomal peptide synthetase able to activate proline and AEO (2-amino-9,10-epoxi-8-oxodecanoic acid), and epimerize L-

Pro.[192] The toxin works by inhibiting the action of host histone deacetylase enzymes.[191] It also contains a D-alanine residue but this is isomerized from L-alanine by a different enzyme in the toxin production pathway and simply incorporated into the toxin by HC-toxin synthetase.[191]

HTS1 contains four amino acid binding domains and it preforms aminoacylation and thioesterification reactions as it synthesizes the toxic peptide.[191] HTS1 also contains a single epimerization motif but this appears to be non-functional as multiple epimerization motifs are necessary to carry out the L to D epimerization reaction.[191] No structure has been determined for this protein but in addition to its recognized amino acid binding domains, it does have some sequence similarity to the ATP-dependent AMP-binding enzyme family.

Disorder predisposition analysis of the HTS1 showed that this protein contains up to 15-20% disordered residues, mostly in a form of relatively short disordered regions, with the longest IDR containing ~70 residues.

Acknowledgements

This work was supported in part by the Programs of the Russian Academy of Sciences for the “Molecular and Cellular Biology” (to V.N.U), the Alberta Innovates Graduate Scholarship in Omics (to Z.P.), and the Natural Sciences and Engineering Research Council (NSERC) Discovery grant (to L.K.).

References

1. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ, Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11**: 161-71, 2000.
2. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**: 635-45, 2004.
3. Uversky VN, The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J Biomed Biotechnol* **2010**: 568068, 2010.
4. Schad E, Tompa P, Hegyi H, The relationship between proteome size, structural disorder and organism complexity. *Genome Biol* **12**: R120, 2011.
5. Xue B, Dunker AK, Uversky VN, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* **30**: 137-49, 2012.
6. Wright PE, Dyson HJ, Linking folding and binding. *Curr Opin Struct Biol* **19**: 31-8, 2009.
7. Uversky VN, Gillespie JR, Fink AL, Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**: 415-27, 2000.
8. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z, Intrinsically disordered protein. *J Mol Graph Model* **19**: 26-59, 2001.
9. Wright PE, Dyson HJ, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**: 321-31, 1999.
10. Dunker AK, Obradovic Z, The protein trinity--linking function and disorder. *Nat Biotechnol* **19**: 805-6, 2001.

11. Dyson HJ, Wright PE, Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* **12**: 54-60, 2002.
12. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z, Intrinsic disorder and protein function. *Biochemistry* **41**: 6573-82, 2002.
13. Dunker AK, Brown CJ, Obradovic Z, Identification and functions of usefully disordered proteins. *Adv Protein Chem* **62**: 25-49, 2002.
14. Tompa P, Intrinsically unstructured proteins. *Trends Biochem Sci* **27**: 527-33, 2002.
15. Uversky VN, Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**: 739-56, 2002.
16. Uversky VN, What does it mean to be natively unfolded? *Eur J Biochem* **269**: 2-12, 2002.
17. Uversky VN, Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell Mol Life Sci* **60**: 1852-71, 2003.
18. Tompa P, Csermely P, The role of structural disorder in the function of RNA and protein chaperones. *Faseb J* **18**: 1169-75, 2004.
19. Daughdrill GW, Pielak GJ, Uversky VN, Cortese MS, Dunker AK, Natively disordered proteins. In *Handbook of Protein Folding* (Buchner, J., Kiefhaber, T., eds.). Wiley-VCH, Verlag GmbH & Co., Weinheim, Germany, Vol., pp. 271-353, 2005.
20. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN, Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS Journal* **272**: 5129-5148, 2005.
21. Dyson HJ, Wright PE, Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**: 197-208, 2005.
22. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK, Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**: 12454-70, 2005.
23. Tompa P, The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* **579**: 3346-54, 2005.
24. Tompa P, Szasz C, Buday L, Structural disorder throws new light on moonlighting. *Trends Biochem Sci* **30**: 484-9, 2005.
25. Uversky VN, Oldfield CJ, Dunker AK, Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* **18**: 343-384, 2005.
26. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK, Intrinsic disorder and functional proteomics. *Biophys J* **92**: 1439-56, 2007.
27. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z, Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* **6**: 1882-98, 2007.
28. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN, 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* **6**: 1899-916, 2007.
29. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN, Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* **6**: 1917-32, 2007.
30. Cortese MS, Uversky VN, Dunker AK, Intrinsic disorder in scaffold proteins: getting more from less. *Prog Biophys Mol Biol* **98**: 85-106, 2008.
31. Dunker AK, Silman I, Uversky VN, Sussman JL, Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* **18**: 756-64, 2008.
32. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN, The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9 Suppl 2**: S1, 2008.

33. Dunker AK, Uversky VN, Signal transduction via unstructured protein conduits. *Nat Chem Biol* **4**: 229-30, 2008.
34. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK, Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9 Suppl 1**: S1, 2008.
35. Russell RB, Gibson TJ, A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett* **582**: 1271-5, 2008.
36. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN, Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* **31**: 328-35, 2009.
37. Tompa P, Fuxreiter M, Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* **33**: 2-8, 2008.
38. Uversky VN, Dunker AK, Biochemistry. Controlled chaos. *Science* **322**: 1340-1, 2008.
39. Uversky VN, Dunker AK, Understanding protein non-folding. *Biochim Biophys Acta* **1804**: 1231-1264, 2010.
40. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK, Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* **103**: 8390-5, 2006.
41. Hsu WL, Oldfield C, Meng J, Huang F, Xue B, Uversky VN, Romero P, Dunker AK, Intrinsic protein disorder and protein-protein interactions. *Pac Symp Biocomput*: 116-27, 2012.
42. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK, Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**: 573-84, 2002.
43. Uversky VN, Roman A, Oldfield CJ, Dunker AK, Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs. *J Proteome Res* **5**: 1829-42, 2006.
44. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN, Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* **45**: 10448-60, 2006.
45. Uversky VN, Oldfield CJ, Dunker AK, Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**: 215-46, 2008.
46. Uversky VN, Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front Biosci* **14**: 5188-238, 2009.
47. Mohan A, Sullivan WJ, Jr., Radivojac P, Dunker AK, Uversky VN, Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol Biosyst* **4**: 328-40, 2008.
48. Xue B, Williams RW, Oldfield CJ, Goh GK, Dunker AK, Uversky VN, Viral disorder or disordered viruses: do viral proteins possess unique features? *Protein Pept Lett* **17**: 932-51, 2010.
49. Goh GK, Dunker AK, Uversky VN, Protein intrinsic disorder and influenza virulence: the 1918 H1N1 and H5N1 viruses. *Virol J* **6**: 69, 2009.
50. Goh GK, Dunker AK, Uversky VN, A comparative analysis of viral matrix proteins using disorder predictors. *Virol J* **5**: 126, 2008.
51. Goh GK, Dunker AK, Uversky VN, Protein intrinsic disorder toolbox for comparative analysis of viral proteins. *BMC Genomics* **9 Suppl 2**: S4, 2008.
52. Xue B, Mizianty MJ, Kurgan L, Uversky VN, Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell Mol Life Sci* **69**: 1211-59, 2012.
53. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE, Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*: 473-84, 1998.

54. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK, Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. *Genome Inform Ser Workshop Genome Inform* **9**: 201-213, 1998.
55. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK, Predicting intrinsic disorder in proteins: an overview. *Cell Res* **19**: 929-49, 2009.
56. Vucetic S, Brown CJ, Dunker AK, Obradovic Z, Flavors of protein disorder. *Proteins* **52**: 573-84, 2003.
57. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK, *IEEE International Conference on Neural Networks, Houston, TX*, 1997.
58. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK, Protein flexibility and intrinsic disorder. *Protein Sci* **13**: 71-80, 2004.
59. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z, Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**: 208, 2006.
60. UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **40**: D71-D75, 2012.
61. 61. Geer LY, Marchler-Bauer A, Geer RC, Han LY, He J, He SQ, Liu CL, Shi WY, Bryant SH, The NCBI BioSystems database. *Nucleic Acids Research* **38**: D492-D496, 2010.
62. Dosztanyi Z, Csizmok V, Tompa P, Simon I, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433-4, 2005.
63. Dosztanyi Z, Csizmok V, Tompa P, Simon I, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology* **347**: 827-839, 2005.
64. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE, ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**: 503-509, 2012.
65. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK, DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* **35**: D786-93, 2007.
66. Peng ZL, Kurgan L, Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr Protein Pept Sci*, 2011.
67. Peng Z, Kurgan L, On the complementarity of the consensus-based disorder prediction. *Pac Symp Biocomput*: 176-87, 2012.
68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Consortium GO, Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**: 25-29, 2000.
69. Anderson TW, Darling DA, Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics* **23**: 193-212, 1952.
70. Wilcoxon F, Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**: 80-83, 1945.
71. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK, Sequence complexity of disordered protein. *Proteins* **42**: 38-48, 2001.
72. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK, Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* **46**: 13468-77, 2007.
73. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN, PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **1804**: 996-1010, 2010.
74. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z, Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* **3**: 35-60, 2005.

75. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL, FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**: 3435-8, 2005.
76. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK, TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett* **15**: 956-63, 2008.
77. Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK, The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput*: 89-100, 2001.
78. Vacic V, Uversky VN, Dunker AK, Lonardi S, Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* **8**: 211, 2007.
79. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank. *Nucleic Acids Res.* **28**: 235-42, 2000.
80. Moh MC, Lee LH, Yang X, Shen S, HEPN1, a novel gene that is frequently down-regulated in hepatocellular carcinoma, suppresses cell growth and induces apoptosis in HepG2 cells. *J Hepatol* **39**: 580-6, 2003.
81. Moh MC, Zhang C, Luo C, Lee LH, Shen S, Structural and functional analyses of a novel ig-like cell adhesion molecule, hepaCAM, in the human breast carcinoma MCF7 cells. *J Biol Chem* **280**: 27366-74, 2005.
82. Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H, Muller-Tidow C, MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**: 8031-41, 2003.
83. Tseng JJ, Hsieh YT, Hsu SL, Chou MM, Metastasis associated lung adenocarcinoma transcript 1 is up-regulated in placenta previa increta/percreta and strongly associated with trophoblast-like cell invasion in vitro. *Mol Hum Reprod* **15**: 725-31, 2009.
84. Ferretti C, Bruni L, Dangles-Marie V, Pecking AP, Bellet D, Molecular circuits shared by placental and cancer cells, and their implications in the proliferative, invasive and migratory capacities of trophoblasts. *Hum Reprod Update* **13**: 121-41, 2007.
85. Garen A, Song X, Regulatory roles of tumor-suppressor proteins and noncoding RNA in cancer and normal cell functions. *Int J Cancer* **122**: 1687-9, 2008.
86. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A, A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**: 39, 2007.
87. Lin R, Maeda S, Liu C, Karin M, Edgington TS, A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* **26**: 851-8, 2007.
88. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, Prasanth KV, The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**: 925-38, 2010.
89. Negro F, Abnormalities of lipid metabolism in hepatitis C virus infection. *Gut* **59**: 1279-87, 2010.
90. Scott BJ, Qutob S, Liu QY, Ng CE, APM2 is a novel mediator of cisplatin resistance in a variety of cancer cell types regardless of p53 or MMR status. *Int J Cancer* **125**: 1193-204, 2009.
91. Kemp TJ, Sadusky TJ, Simon M, Brown R, Eastwood M, Sassoon DA, Coulton GR, Identification of a novel stretch-responsive skeletal muscle gene (Smpx). *Genomics* **72**: 260-71, 2001.

92. Patzak D, Zhuchenko O, Lee CC, Wehnert M, Identification, mapping, and genomic structure of a novel X-chromosomal human gene (SMPX) encoding a small muscular protein. *Hum Genet* **105**: 506-12, 1999.
93. Schraders M, Haas SA, Weegerink NJ, Oostrik J, Hu H, Hoefsloot LH, Kannan S, Huygen PL, Pennings RJ, Admiraal RJ, Kalscheuer VM, Kunst HP, Kremer H, Next-generation sequencing identifies mutations of SMPX, which encodes the small muscle protein, X-linked, as a cause of progressive hearing impairment. *Am J Hum Genet* **88**: 628-34, 2011.
94. Huebner AK, Gandia M, Frommolt P, Maak A, Wicklein EM, Thiele H, Altmuller J, Wagner F, Vinuela A, Aguirre LA, Moreno F, Maier H, Rau I, Giesselmann S, Nurnberg G, Gal A, Nurnberg P, Hubner CA, del Castillo I, Kurth I, Nonsense mutations in SMPX, encoding a protein responsive to physical force, result in X-chromosomal hearing loss. *Am J Hum Genet* **88**: 621-7, 2011.
95. Loret EP, del Valle RM, Mansuelle P, Sampieri F, Rochat H, Positively charged amino acid residues located similarly in sea anemone and scorpion toxins. *J Biol Chem* **269**: 16785-8, 1994.
96. Pallaghy PK, Scanlon MJ, Monks SA, Norton RS, Three-dimensional structure in solution of the polypeptide cardiac stimulant anthopleurin-A. *Biochemistry* **34**: 3782-94, 1995.
97. Norton RS, Structure and structure-function relationships of sea anemone proteins that interact with the sodium channel. *Toxicon* **29**: 1051-84, 1991.
98. Goudet C, Ferrer T, Galan L, Artiles A, Batista CF, Possani LD, Alvarez J, Aneiros A, Tytgat J, Characterization of two Bunodosoma granulifera toxins active on cardiac sodium channels. *Br J Pharmacol* **134**: 1195-206, 2001.
99. Carlsson FH, Snake venom toxins. The primary structures of two novel cytotoxin homologues from the venom of forest cobra (Naja melanoleuca). *Biochem Biophys Res Commun* **59**: 269-76, 1974.
100. Dubovskii PV, Lesovoy DM, Dubinnyi MA, Konshina AG, Utkin YN, Efremov RG, Arseniev AS, Interaction of three-finger toxins with phospholipid membranes: comparison of S- and P-type cytotoxins. *Biochem J* **387**: 807-15, 2005.
101. 101. Dementieva DV, Bocharov EV, Arseniev AS, Two forms of cytotoxin II (cardiotoxin) from Naja naja oxiana in aqueous solution: spatial structures with tightly bound water molecules. *Eur J Biochem* **263**: 152-62, 1999.
102. Grutter MG, Priestle JP, Rahuel J, Grossenbacher H, Bode W, Hofsteenge J, Stone SR, Crystal structure of the thrombin-hirudin complex: a novel mode of serine protease inhibition. *EMBO J* **9**: 2361-5, 1990.
103. Uversky VN, Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. *Chem Soc Rev* **40**: 1623-34, 2011.
104. Gibbs S, Lohman F, Teubel W, van de Putte P, Backendorf C, Characterization of the human spr2 promoter: induction after UV irradiation or TPA treatment and regulation during differentiation of cultured primary keratinocytes. *Nucleic Acids Res* **18**: 4401-7, 1990.
105. Gibbs S, Fijneman R, Wiegant J, van Kessel AG, van De Putte P, Backendorf C, Molecular characterization and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. *Genomics* **16**: 630-7, 1993.
106. Bohren KM, Nadkarni V, Song JH, Gabbay KH, Owerbach D, A M55V polymorphism in a novel SUMO gene (SUMO-4) differentially activates heat shock transcription factors and is associated with susceptibility to type I diabetes mellitus. *J Biol Chem* **279**: 27233-8, 2004.
107. Park WR, Nakamura Y, p53CSV, a novel p53-inducible gene involved in the p53-dependent cell-survival pathway. *Cancer Res* **65**: 1197-206, 2005.
108. Popov S, Rexach M, Zybarth G, Reiling N, Lee MA, Ratner L, Lane CM, Moore MS, Blobel G, Bukrinsky M, Viral protein R regulates nuclear import of the HIV-1 pre-integration complex. *EMBO J* **17**: 909-17, 1998.

109. Jenkins Y, McEntee M, Weis K, Greene WC, Characterization of HIV-1 vpr nuclear import: analysis of signals and pathways. *J Cell Biol* **143**: 875-85, 1998.
110. Piller SC, Ewart GD, Jans DA, Gage PW, Cox GB, The amino-terminal region of Vpr from human immunodeficiency virus type 1 forms ion channels and kills neurons. *J Virol* **73**: 4230-8, 1999.
111. Jacotot E, Ferri KF, El Hamel C, Brenner C, Druillennec S, Hoebeke J, Rustin P, Metivier D, Lenoir C, Geuskens M, Vieira HL, Loeffler M, Belzacq AS, Briand JP, Zamzami N, Edelman L, Xie ZH, Reed JC, Roques BP, Kroemer G, Control of mitochondrial membrane permeabilization by adenine nucleotide translocator interacting with HIV-1 viral protein rR and Bcl-2. *J Exp Med* **193**: 509-19, 2001.
112. Muthumani K, Lambert VM, Shanmugam M, Thieu KP, Choo AY, Chung JC, Satishchandran A, Kim JJ, Weiner DB, Ugen KE, Anti-tumor activity mediated by protein and peptide transduction of HIV viral protein R (Vpr). *Cancer Biol Ther* **8**: 180-7, 2009.
113. Muthumani K, Lambert VM, Sardesai NY, Kim JJ, Heller R, Weiner DB, Ugen KE, Analysis of the potential for HIV-1 Vpr as an anti-cancer agent. *Curr HIV Res* **7**: 144-52, 2009.
114. Emerman M, HIV-1, Vpr and the cell cycle. *Curr Biol* **6**: 1096-103, 1996.
115. Tungaturthi PK, Sawaya BE, Singh SP, Tomkowicz B, Ayyavoo V, Khalili K, Collman RG, Amini S, Srinivasan A, Role of HIV-1 Vpr in AIDS pathogenesis: relevance and implications of intravirion, intracellular and free Vpr. *Biomed Pharmacother* **57**: 20-4, 2003.
116. Majumder B, Venkatachari NJ, Srinivasan A, Ayyavoo V, HIV-1 mediated immune pathogenesis: spotlight on the role of viral protein R (Vpr). *Curr HIV Res* **7**: 169-77, 2009.
117. Andersen JL, Planelles V, The role of Vpr in HIV-1 pathogenesis. *Curr HIV Res* **3**: 43-51, 2005.
118. Sawaya BE, Khalili K, Gordon J, Taube R, Amini S, Cooperative interaction between HIV-1 regulatory proteins Tat and Vpr modulates transcription of the viral genome. *J Biol Chem* **275**: 35209-14, 2000.
119. Chang F, Re F, Sebastian S, Sazer S, Luban J, HIV-1 Vpr induces defects in mitosis, cytokinesis, nuclear structure, and centrosomes. *Mol Biol Cell* **15**: 1793-801, 2004.
120. Rogel ME, Wu LI, Emerman M, The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. *J Virol* **69**: 882-8, 1995.
121. Ramanathan MP, Curley E, 3rd, Su M, Chambers JA, Weiner DB, Carboxyl terminus of hVIP/mov34 is critical for HIV-1-Vpr interaction and glucocorticoid-mediated signaling. *J Biol Chem* **277**: 47854-60, 2002.
122. Jowett JB, Xie YM, Chen IS, The presence of human immunodeficiency virus type 1 Vpr correlates with a decrease in the frequency of mutations in a plasmid shuttle vector. *J Virol* **73**: 7132-7, 1999.
123. Somasundaran M, Sharkey M, Brichacek B, Luzuriaga K, Emerman M, Sullivan JL, Stevenson M, Evidence for a cytopathogenicity determinant in HIV-1 Vpr. *Proc Natl Acad Sci U S A* **99**: 9503-8, 2002.
124. Morellet N, Bouaziz S, Petitjean P, Roques BP, NMR structure of the HIV-1 regulatory protein VPR. *J Mol Biol* **327**: 215-27, 2003.
125. Zhou Y, Lu Y, Ratner L, Arginine residues in the C-terminus of HIV-1 Vpr are important for nuclear localization and cell cycle arrest. *Virology* **242**: 414-24, 1998.
126. Bourbigot S, Beltz H, Denis J, Morellet N, Roques BP, Mely Y, Bouaziz S, The C-terminal domain of the HIV-1 regulatory protein Vpr adopts an antiparallel dimeric structure in solution via its leucine-zipper-like domain. *Biochem J* **387**: 333-41, 2005.
127. Vial S, Lu H, Allen S, Savory P, Thornton D, Sheehan J, Tokatlidis K, Assembly of Tim9 and Tim10 into a functional chaperone. *J Biol Chem* **277**: 36100-8, 2002.

128. Baker MJ, Webb CT, Stroud DA, Palmer CS, Frazier AE, Guiard B, Chacinska A, Gulbis JM, Ryan MT, Structural and functional requirements for activity of the Tim9-Tim10 complex in mitochondrial protein import. *Mol Biol Cell* **20**: 769-79, 2009.
129. Murphy MP, Leuenberger D, Curran SP, Oppliger W, Koehler CM, The essential function of the small Tim proteins in the TIM22 import pathway does not depend on formation of the soluble 70-kilodalton complex. *Mol Cell Biol* **21**: 6132-8, 2001.
130. Wiedemann N, Truscott KN, Pfannschmidt S, Guiard B, Meisinger C, Pfanner N, Biogenesis of the protein import channel Tom40 of the mitochondrial outer membrane: intermembrane space components are involved in an early stage of the assembly pathway. *J Biol Chem* **279**: 18188-94, 2004.
131. Allen S, Lu H, Thornton D, Tokatlidis K, Juxtaposition of the two distal CX3C motifs via intrachain disulfide bonding is essential for the folding of Tim10. *J Biol Chem* **278**: 38505-13, 2003.
132. Vergnolle MA, Baud C, Golovanov AP, Alcock F, Luciano P, Lian LY, Tokatlidis K, Distinct domains of small Tims involved in subunit interaction and substrate recognition. *J Mol Biol* **351**: 839-49, 2005.
133. Chou KC, Energy-optimized structure of antifreeze protein and its binding mechanism. *J Mol Biol* **223**: 509-17, 1992.
134. Hew CL, Wang NC, Yan S, Cai H, Sclater A, Fletcher GL, Biosynthesis of antifreeze polypeptides in the winter flounder. Characterization and seasonal occurrence of precursor polypeptides. *Eur J Biochem* **160**: 267-72, 1986.
135. DeVries AL, Cheng CHC, Antifreeze proteins and organismal freezing avoidance in polar fishes. In *Fish Physiology* (Farrell, A. P., Steffensen, J. F., eds.). Elsevier Academic Press, New York, Vol. 22, pp. 155-201, 2005.
136. Liepinsh E, Otting G, Harding MM, Ward LG, Mackay JP, Haymet AD, Solution structure of a hydrophobic analogue of the winter flounder antifreeze protein. *Eur J Biochem* **269**: 1259-66, 2002.
137. Qi C, Zhu YT, Hu L, Zhu YJ, Identification of Fat4 as a candidate tumor suppressor gene in breast cancers. *Int J Cancer* **124**: 793-8, 2009.
138. Cho E, Feng Y, Rauskolb C, Maitra S, Fehon R, Irvine KD, Delineation of a Fat tumor suppressor pathway. *Nat Genet* **38**: 1142-50, 2006.
139. Harvey K, Tapon N, The Salvador-Warts-Hippo pathway - an emerging tumour-suppressor network. *Nat Rev Cancer* **7**: 182-91, 2007.
140. Yang CH, Axelrod JD, Simon MA, Regulation of Frizzled by fat-like cadherins during planar polarity signaling in the Drosophila compound eye. *Cell* **108**: 675-88, 2002.
141. Saburi S, Hester I, Fischer E, Pontoglio M, Eremina V, Gessler M, Quaggin SE, Harrison R, Mount R, McNeill H, Loss of Fat4 disrupts PCP signaling and oriented cell division and leads to cystic kidney disease. *Nat Genet* **40**: 1010-5, 2008.
142. Yin BW, Dnistrian A, Lloyd KO, Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. *Int J Cancer* **98**: 737-40, 2002.
143. Bast RC, Jr., Feeney M, Lazarus H, Nadler LM, Colvin RB, Knapp RC, Reactivity of a monoclonal antibody with human ovarian carcinoma. *J Clin Invest* **68**: 1331-7, 1981.
144. Bast RC, Jr., Klug TL, St John E, Jenison E, Niloff JM, Lazarus H, Berkowitz RS, Leavitt T, Griffiths CT, Parker L, Zurawski VR, Jr., Knapp RC, A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *N Engl J Med* **309**: 883-7, 1983.
145. Bast RC, Jr., Xu FJ, Yu YH, Barnhill S, Zhang Z, Mills GB, CA 125: the past and the future. *Int J Biol Markers* **13**: 179-87, 1998.
146. Fritsche HA, Bast RC, CA 125 in ovarian cancer: advances and controversy. *Clin Chem* **44**: 1379-80, 1998.

147. Wagner U, Kohler S, Reinartz S, Giffels P, Huober J, Renke K, Schlebusch H, Biersack HJ, Mobus V, Kreienberg R, Bauknecht T, Krebs D, Wallwiener D, Immunological consolidation of ovarian carcinoma recurrences with monoclonal anti-idiotype antibody ACA125: immune responses and survival in palliative treatment. See The biology behind: K. A. Foon and M. Bhattacharya-Chatterjee, Are solid tumor anti-idiotype vaccines ready for prime time? Clin. Cancer Res., 7:1112-1115, 2001. *Clin Cancer Res* **7**: 1154-62, 2001.
148. Rump A, Morikawa Y, Tanaka M, Minami S, Umesaki N, Takeuchi M, Miyajima A, Binding of ovarian cancer antigen CA125/MUC16 to mesothelin mediates cell adhesion. *J Biol Chem* **279**: 9190-8, 2004.
149. Bast RC, Jr., Spriggs DR, More than a biomarker: CA125 may contribute to ovarian cancer pathogenesis. *Gynecol Oncol* **121**: 429-30, 2011.
150. McLemore MR, Aouizerat B, Introducing the MUC16 gene: implications for prevention and early detection in epithelial ovarian cancer. *Biol Res Nurs* **6**: 262-7, 2005.
151. DeSilva U, D'Arcangelo G, Braden VV, Chen J, Miao GG, Curran T, Green ED, The human reelin gene: isolation, sequencing, and mapping on chromosome 7. *Genome Res* **7**: 157-64, 1997.
152. D'Arcangelo G, Miao GG, Chen SC, Soares HD, Morgan JI, Curran T, A protein related to extracellular matrix proteins deleted in the mouse mutant reeler. *Nature* **374**: 719-23, 1995.
153. D'Arcangelo G, Homayouni R, Keshvara L, Rice DS, Sheldon M, Curran T, Reelin is a ligand for lipoprotein receptors. *Neuron* **24**: 471-9, 1999.
154. Forster E, Bock HH, Herz J, Chai X, Frotscher M, Zhao S, Emerging topics in Reelin function. *Eur J Neurosci* **31**: 1511-8, 2010.
155. Hong SE, Shugart YY, Huang DT, Shahwan SA, Grant PE, Hourihane JO, Martin ND, Walsh CA, Autosomal recessive lissencephaly with cerebellar hypoplasia is associated with human RELN mutations. *Nat Genet* **26**: 93-6, 2000.
156. Jossin Y, Goffinet AM, Reelin signals through phosphatidylinositol 3-kinase and Akt to control cortical development and through mTor to regulate dendritic growth. *Mol Cell Biol* **27**: 7113-24, 2007.
157. Weeber EJ, Beffert U, Jones C, Christian JM, Forster E, Sweatt JD, Herz J, Reelin and ApoE receptors cooperate to enhance hippocampal synaptic plasticity and learning. *J Biol Chem* **277**: 39944-52, 2002.
158. Fatemi SH, *Reelin Glycoprotein: Structure, Biology and Roles in Health and Disease*, Springer, New York, 2008.
159. Stone MR, O'Neill A, Catino D, Bloch RJ, Specific interaction of the actin-binding domain of dystrophin with intermediate filaments containing keratin 19. *Mol Biol Cell* **16**: 4280-93, 2005.
160. Bies RD, Phelps SF, Cortez MD, Roberts R, Caskey CT, Chamberlain JS, Human and murine dystrophin mRNA transcripts are differentially expressed during skeletal muscle, heart, and brain development. *Nucleic Acids Res* **20**: 1725-31, 1992.
161. McKinley-Grant LJ, Idler WW, Bernstein IA, Parry DA, Cannizzaro L, Croce CM, Huebner K, Lessin SR, Steinert PM, Characterization of a cDNA clone encoding human filaggrin and localization of the gene to chromosome region 1q21. *Proc Natl Acad Sci U S A* **86**: 4848-52, 1989.
162. Marenholz I, Nickel R, Ruschendorf F, Schulz F, Esparza-Gordillo J, Kerscher T, Gruber C, Lau S, Worm M, Keil T, Kurek M, Zaluga E, Wahn U, Lee YA, Filaggrin loss-of-function mutations predispose to phenotypes involved in the atopic march. *J Allergy Clin Immunol* **118**: 866-71, 2006.
163. Benian GM, Kiff JE, Neckelmann N, Moerman DG, Waterston RH, Sequence of an unusually large protein implicated in regulation of myosin activity in C. elegans. *Nature* **342**: 45-50, 1989.

164. Greene DN, Garcia T, Sutton RB, Gernert KM, Benian GM, Oberhauser AF, Single-molecule force spectroscopy reveals a stepwise unfolding of Caenorhabditis elegans giant protein kinase domains. *Biophys J* **95**: 1360-70, 2008.

165. Pichler A, Gast A, Seeler JS, Dejean A, Melchior F, The nucleoporin RanBP2 has SUMO1 E3 ligase activity. *Cell* **108**: 109-20, 2002.

166. Yokoyama N, Hayashi N, Seki T, Pante N, Ohba T, Nishii K, Kuma K, Hayashida T, Miyata T, Aebi U, et al., A giant nucleopore protein that binds Ran/TC4. *Nature* **376**: 184-8, 1995.

167. Beddow AL, Richards SA, Orem NR, Macara IG, The Ran/TC4 GTPase-binding domain: identification by expression cloning and characterization of a conserved sequence motif. *Proc Natl Acad Sci U S A* **92**: 3328-32, 1995.

168. Kirsh O, Seeler JS, Pichler A, Gast A, Muller S, Miska E, Mathieu M, Harel-Bellan A, Kouzarides T, Melchior F, Dejean A, The SUMO E3 ligase RanBP2 promotes modification of the HDAC4 deacetylase. *EMBO J* **21**: 2682-91, 2002.

169. Yi H, Friedman JL, Ferreira PA, The cyclophilin-like domain of Ran-binding protein-2 modulates selectively the activity of the ubiquitin-proteasome system and protein biogenesis. *J Biol Chem* **282**: 34770-8, 2007.

170. Navarro MS, Bachant J, RanBP2: a tumor suppressor with a new twist on TopoII, SUMO, and centromeres. *Cancer Cell* **13**: 293-5, 2008.

171. Truong K, Su Y, Song J, Chen Y, Entropy-driven mechanism of an E3 ligase. *Biochemistry* **50**: 5757-66, 2011.

172. Bekker-Jensen S, Rendtlew Danielsen J, Fugger K, Gromova I, Nerstedt A, Lukas C, Bartek J, Lukas J, Mailand N, HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes. *Nat Cell Biol* **12**: 80-6; sup pp 1-12, 2010.

173. Kang TH, Lindsey-Boltz LA, Reardon JT, Sancar A, Circadian control of XPA and excision repair of cisplatin-DNA damage by cryptochrome and HERC2 ubiquitin ligase. *Proc Natl Acad Sci U S A* **107**: 4890-5, 2010.

174. Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, Hansen L, Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* **123**: 177-87, 2008.

175. Roy R, Chun J, Powell SN, BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer* **12**: 68-78, 2012.

176. Hussain S, Wilson JB, Medhurst AL, Hejna J, Witt E, Ananth S, Davies A, Masson JY, Moses R, West SC, de Winter JP, Ashworth A, Jones NJ, Mathew CG, Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways. *Hum Mol Genet* **13**: 1241-8, 2004.

177. Ohashi A, Zdzienicka MZ, Chen J, Couch FJ, Fanconi anemia complementation group D2 (FANCD2) functions independently of BRCA2- and RAD51-associated homologous recombination in response to DNA damage. *J Biol Chem* **280**: 14877-83, 2005.

178. Wang X, Andreassen PR, D'Andrea AD, Functional interaction of monoubiquitinated FANCD2 and BRCA2/FANCD1 in chromatin. *Mol Cell Biol* **24**: 5850-62, 2004.

179. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G, Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**: 789-92, 1995.

180. Gutierrez-Enriquez S, de la Hoya M, Martinez-Bouzas C, Sanchez de Abajo A, Ramon y Cajal T, Llort G, Blanco I, Beristain E, Diaz-Rubio E, Alonso C, Tejada MI, Caldes T, Diez O, Screening for large rearrangements of the BRCA2 gene in Spanish families with breast/ovarian cancer. *Breast Cancer Res Treat* **103**: 103-7, 2007.

181. Gonzalez-Hormazabal P, Gutierrez-Enriquez S, Gaete D, Reyes JM, Peralta O, Waugh E, Gomez F, Margarit S, Bravo T, Blanco R, Diez O, Jara L, Spectrum of BRCA1/2 point

mutations and genomic rearrangements in high-risk breast/ovarian cancer Chilean families. *Breast Cancer Res Treat* **126**: 705-16, 2011.

182. Ozcelik H, Schmocker B, Di Nicola N, Shi XH, Langer B, Moore M, Taylor BR, Narod SA, Darlington G, Andrulis IL, Gallinger S, Redston M, Germline BRCA2 6174delT mutations in Ashkenazi Jewish pancreatic cancer patients. *Nat Genet* **16**: 17-8, 1997.
183. Deng CX, BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Res* **34**: 1416-26, 2006.
184. Venkitaraman AR, Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**: 171-82, 2002.
185. Mark WY, Liao JC, Lu Y, Ayed A, Laister R, Szymczyna B, Chakrabartty A, Arrowsmith CH, Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? *J Mol Biol* **345**: 275-87, 2005.
186. Harada N, Iijima S, Kobayashi K, Yoshida T, Brown WR, Hibi T, Oshima A, Morikawa M, Human IgGFc binding protein (FcgammaBP) in colonic epithelial cells exhibits mucin-like structure. *J Biol Chem* **272**: 15232-41, 1997.
187. Johansson ME, Thomsson KA, Hansson GC, Proteomic analyses of the two mucus layers of the colon barrier reveal that their main component, the Muc2 mucin, is strongly bound to the Fcgbp protein. *J Proteome Res* **8**: 3549-57, 2009.
188. Kobayashi K, Yagasaki M, Harada N, Chichibu K, Hibi T, Yoshida T, Brown WR, Morikawa M, Detection of Fcgamma binding protein antigen in human sera and its relation with autoimmune diseases. *Immunol Lett* **79**: 229-35, 2001.
189. O'Donovan N, Fischer A, Abdo EM, Simon F, Peter HJ, Gerber H, Buergi U, Marti U, Differential expression of IgG Fc binding protein (FcgammaBP) in human normal thyroid tissue, thyroid adenomas and thyroid carcinomas. *J Endocrinol* **174**: 517-24, 2002.
190. Panaccione DG, Scott-Craig JS, Pocard JA, Walton JD, A cyclic peptide synthetase gene required for pathogenicity of the fungus Cochliobolus carbonum on maize. *Proc Natl Acad Sci U S A* **89**: 6590-4, 1992.
191. Cheng YQ, Walton JD, A eukaryotic alanine racemase gene involved in cyclic peptide biosynthesis. *J Biol Chem* **275**: 4906-11, 2000.
192. Scott-Craig JS, Panaccione DG, Pocard JA, Walton JD, The cyclic peptide synthetase catalyzing HC-toxin production in the filamentous fungus Cochliobolus carbonum is encoded by a 15.7-kilobase open reading frame. *J Biol Chem* 267: 26044-9, 1992.