

Disease mutations in disordered regions—exception to the rule?†

Vladimir Vacic^a and Lilia M. Iakoucheva^{*b}

DOI: 10.1039/c1mb05251a

Intrinsically disordered proteins (IDPs) have been implicated in a number of human diseases, including cancer, diabetes, neurodegenerative and cardiovascular disorders. Although for some of these conditions molecular mechanisms are now better understood, the big picture connecting distinct structural properties and functional repertoire of IDPs to pathogenesis and disease progression is still incomplete. Recent studies suggest that signaling and regulatory roles carried out by IDPs require them to be tightly regulated, and that altered IDP abundance may lead to disease. Here, we propose another link between IDPs and disease that takes into account disease-associated missense mutations located in the intrinsically disordered regions. We argue that such mutations are more prevalent and have larger functional impact than previously thought. In addition, we demonstrate that deleterious amino acid substitutions that cause disorder-to-order transitions are particularly enriched among disease mutations compared to neutral polymorphisms. Finally, we discuss potential differences in functional outcomes between disease mutations in ordered and disordered regions, and challenge the conventional structure-centric view of missense mutations.

Recent predictions suggest that more than 40% of human proteins have at least one long region (≥ 30 residues) that under physiological conditions does not fold into a fixed three-dimensional structure.¹ These intrinsically unstructured or intrinsically disordered regions (IDRs) mediate important biological functions such as post-translational modification, molecular recognition and assembly, as well as binding to other proteins, DNA and RNA.^{2–6} Intrinsically disordered proteins (IDPs) frequently serve as hubs in protein–protein interaction networks,⁷ and their disordered regions allow binding to multiple partners.^{8,9} In comparison to ordered regions, IDRs generally have lower sequence conservation,¹⁰ with the

exception of IDRs involved in RNA binding and chaperone activity.¹¹ Disordered proteins were shown to be involved in a number of human diseases,^{12,13} and disruption of tight regulation of IDPs could be a contributor to disease pathogenesis.¹⁴ Given the high prevalence of disordered regions in the human proteome^{1,15,16} and their involvement in human diseases, below we explore whether disease-associated mutations could be found in IDRs and what is a possible impact of such mutations on protein disorder.

Historically, disease-associated mutations have been studied from a structural perspective,^{17–22} and much of the attention was focused on understanding how missense mutations influence folding, stability, solubility, activity and other structure-based properties of proteins. Significant progress has been made over the years in classifying potential functional effects of single nucleotide polymorphisms (SNPs), especially in the context of their influence on human health. This is illustrated by the development of numerous

predictors of functional impact of SNPs (ref. 23–25 and others). However, the majority of these methods are structure- and/or conservation-based, which limits their applicability in protein regions with unknown structure or low sequence conservation. In addition, until recently only conserved regions of proteins were considered to be functionally important. As a consequence, existing methods often classify mutations within non-conserved regions as tolerant, not damaging or benign,²⁶ because they are believed to be functionally neutral. For example, the SIFT algorithm tends to incorrectly classify the effect of mutations located in non-conserved,²⁷ solvent accessible or disordered regions of proteins.²⁶ Recent studies demonstrated that prediction accuracy, in particular within disordered regions, can be improved by incorporating prior functional information such as loss or gain of post-translational modification sites or catalytic residues.^{26,28} Here, we focus on missense mutations in IDRs and argue that mutations in generally

^a Department of Computer Science, Columbia University, New York, NY 10027

^b Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093. E-mail: lilyak@ucsd.edu

† Published as part of a Molecular BioSystems themed issue on Intrinsically Disordered Proteins: Guest Editor M. Madan Babu.

non-conserved disordered regions can be highly deleterious because they can produce dramatic changes in disordered structure. Importantly, we propose that properties of mutations in disordered regions need to be taken into account when predicting the effects of missense mutations on protein structure and function. Below, we discuss the differences in the functional impact of mutations in ordered and disordered regions and relate them to different disease mechanisms.

We predicted disorder in the dataset of proteins that carry annotated disease mutations from the UniProt database²⁹ using three different disorder predictors^{30–32} and observed that 20–25% of disease mutations were mapped to predicted disordered regions (Vacic and Iakoucheva, submitted). We believe that this number may be an underestimate because at least some of the mutations in UniProt are annotated as being disease-related because they disrupt important functional sites inferred from known structures. Then, we *in silico* mutated the wild type protein sequences to mimic the annotated disease mutations. We observed that 20% of disease mutations located in disordered regions cause disorder-to-order (D → O) transitions, defined here and throughout this manuscript as a change from predicted disorder (score ≥ 0.5) into order (score < 0.5).³⁰ In the two control datasets, annotated polymorphisms from UniProt and neutral evolutionary substitutions, percentages of mutations that cause D → O transitions were significantly lower (11.5% and 7.3%, Fisher's exact $P = 1.06 \times 10^{-32}$ and 5.47×10^{-105} , respectively). Table 1 shows representative examples of D → O mutations that affect experimentally confirmed disordered regions of proteins from the DisProt database.³³ In total, we have collected over 700 annotated disease mutations from UniProt that cause D → O transitions based on the disorder prediction score. As evident from these examples, disease mutations can also affect disordered regions, and some of them can disrupt disordered conformation via D → O transitions.

Ordered and disordered proteins have distinct functional repertoires: while ordered proteins are mainly involved in metabolism, biosynthesis, catalysis and related cellular processes, disordered proteins carry out regulatory and signaling roles.^{4,5,34}

Disordered regions are believed to be involved in low affinity and high specificity binding of IDPs to their targets.^{35,36} It is therefore likely that the functional impact of disease mutations in these two types of regions would also differ. A plausible hypothesis for the impact of disease mutations in disordered regions is that they primarily disrupt disorder-mediated processes such as protein–protein, protein–DNA, protein–RNA and protein–ligand interactions, post-translational modifications, assembly of macromolecular complexes, and thereby signaling and regulatory networks (Fig. 1).

According to the traditional structure-centric view of disease mutations, a disease may arise from malfunction of a specific protein due to the loss of its stably folded structure or enzymatic activity (Fig. 1). Examples of such disease mechanisms are plentiful in the literature. For instance, in the case of phenylketonuria (OMIM #261600) most of the associated missense mutations impair enzymatic activity of the phenylalanine hydroxylase protein (PHA) by causing its increased instability and aggregation. Furthermore, it was shown that the decrease in PHA stability is the main molecular pathogenic mechanism in phenylketonuria and the determinant of phenotypic outcome in the patients.³⁷ Another example of a metabolic disorder characterized by enzymatic deficiency is homocystinuria (OMIM #236200), which is usually caused by the mutations in the gene that encodes cystathionine beta-synthase.

On the other hand, a new disorder-centric view of missense mutations suggests that a disease may arise from a loss (of wanted) or gain (of unwanted) interactions between a candidate protein and its interaction partners due to mutations that disrupt disordered regions (Fig. 1). Although these two mechanisms are not mutually exclusive—loss of enzyme's activity could in addition eliminate some of its interactions with the corresponding ligands/partners—disruption of signaling and regulatory networks via interaction-specific defects is the most plausible mechanism for diseases that involve mutations in IDRs. This hypothesis agrees with the study by Zhong *et al.*,³⁸ who investigated how disease mutations affect the human protein–protein interaction network. Using a small set of carefully chosen missense mutations, they were able to demonstrate that perturbations of the interactome can

be caused by either a complete loss of gene products (node removal), or by interaction-specific (edgetic) alterations. Mutations leading to node removal were likely to affect buried residues of the protein (comparable to ordered regions), whereas mutations leading to loss or gain of specific interactions were likely to lie on the protein surface³⁸ (comparable to disordered regions). Although both of these mechanisms influence interaction networks, they could have different consequences, especially with regard to disease mechanisms and modes of disease inheritance.³⁸

Role of IDPs as network hubs^{7,8} could further contribute to the network disruption in disease. The 'edgetic' network perturbations that disrupt interactions of hub proteins may result in an imbalanced amount of protein complex subunits. Defective protein complexes may not function properly in the cell, or may be rapidly degraded by the cellular proteolytic machinery. The loss of post-translational modifications (PTMs) could be another potential outcome of the 'edgetic' networks perturbation. Our group and others have previously shown that disorder is required for post-translational modifications such as phosphorylation,² ubiquitination,³ methylation³⁹ and possibly other PTMs. D → O transition mutations could render modification sites less exposed and thus impair the access of modifying enzymes to the PTM sites. As a result, the loss of ubiquitination sites could lead to accumulation of dosage-sensitive IDPs⁴⁰ inside the cell, thereby contributing to disease development. Likewise, access of kinases to phosphorylation sites may be compromised by D → O mutations of the site or its flanking regions, which could influence downstream signaling cascades. D → O mutations could also alter the binding specificity or affinity of IDPs to their partners, thereby leading to either more promiscuous binding or to accumulation of highly stable complexes. Both of these outcomes are undesirable for the finely tuned and dynamic signaling networks, where interactions need to be precise and at the same time easily breakable. This is especially relevant for the fuzzy complexes that rely on dynamic disorder,⁴¹ since dynamically disordered regions could be especially prone to disruption by D → O mutations. In addition, D → O transition mutations could impair regulatory functions of IDPs.

Table 1 Examples of disease mutations located in the experimentally confirmed disordered regions from DisProt proteins

Protein ID	Protein name	Mutation position	Wild type residue	Mutant residue	PONDR disorder score			Disease	DisProt ID	Disordered region(s) position and function
					From	To	Δ			
MECP2	Methyl-CpG-binding protein 2	306	R	C	0.96	0.37	0.59	Rett syndrome (RTT) [MIM:312750]	DP00539	#207-310; #335-486: Molecular recognition effectors, Intraprotein interaction, Protein-DNA binding, Protein-protein binding
		453	R	Q	0.96	0.47	0.49	Rett syndrome (RTT) [MIM:312750]		
BRCA1	Breast cancer type 1 susceptibility protein	227	E	K	0.55	0.15	0.41	Mental retardation syndromic X-linked type 13 [MIM:300055]	DP00238	#170-1649: Molecular recognition effectors, Protein-protein binding
		835	H	Y	0.66	0.25	0.40	Ovarian cancer [MIM:113705]		
		1204	R	I	0.52	0.33	0.19	Familial breast-ovarian cancer type 1 [MIM:604370]		
		1217	S	Y	0.83	0.4	0.43	Breast cancer [MIM:113705, 114480]		
								Breast cancer [MIM:113705, 114480]; Familial breast-ovarian cancer type 1 [MIM:604370]		
TP53	Tumor suppressor p53	17	E	D	0.67	0.47	0.20	Sporadic cancers	DP00086	#1-73: Molecular recognition effectors
		34	P	L	0.56	0.42	0.14	Sporadic cancers		
		35	L	F	0.63	0.47	0.16	Sporadic cancers		
		46	S	F	0.69	0.40	0.29	Sporadic cancers		
		47	P	L	0.69	0.42	0.27	Sporadic cancers		
VHL	von Hippel-Lindau Tumor Suppressor	49	D	Y	0.79	0.34	0.45	Sporadic cancers	DP00287	#1-213
		65	S	W	0.53	0.41	0.12	Von Hippel-Lindau disease [MIM:193300]		
			S	L	0.53	0.48	0.05	Von Hippel-Lindau disease [MIM:193300]		
		66	V	F	0.57	0.39	0.17	Von Hippel-Lindau disease [MIM:193300]		
		167	R	G	0.62	0.38	0.24	Von Hippel-Lindau disease [MIM:193300]		
			R	W	0.62	0.22	0.40	Von Hippel-Lindau disease [MIM:193300]; Pheochromocytoma [MIM:171300]		
		176	R	Q	0.62	0.38	0.24	Von Hippel-Lindau disease [MIM:193300]		
NR3C1	Glucocorticoid receptor	200	R	W	0.51	0.13	0.38	Von Hippel-Lindau disease [MIM:193300]; Erythrocytosis familial type 2 [MIM:263400]	DP00030	#1-500: Molecular recognition effector, Metal binding, Phosphorylation, Protein-protein binding, Protein-DNA binding
		477	R	H	0.70	0.48	0.23	Von Hippel-Lindau disease [MIM:193300]; Erythrocytosis familial type 2 [MIM:263400]		
					0.75	0.4	0.35	Glucocorticoid resistance [MIM:138040]		
TNNI3	Troponin I, cardiac muscle	166	S	F	0.57	0.36	0.2	Cardiomyopathy familial hypertrophic type 7 [MIM:191044]	DP00166	#163-210: Molecular recognition effector, Molecular assembly, Entropic chain, Flexible linkers/spacers, Protein-protein binding
SOD1	Superoxide dismutase	125	D	G	0.52	0.45	0.07	Amyotrophic lateral sclerosis type 1 [MIM:105400]	DP00652	#67-79; #125-141: Metal binding
		126	D	H	0.52	0.41	0.11	Amyotrophic lateral sclerosis type 1 [MIM:105400]		
			D	H	0.58	0.43	0.15	Amyotrophic lateral sclerosis type 1 [MIM:105400]		

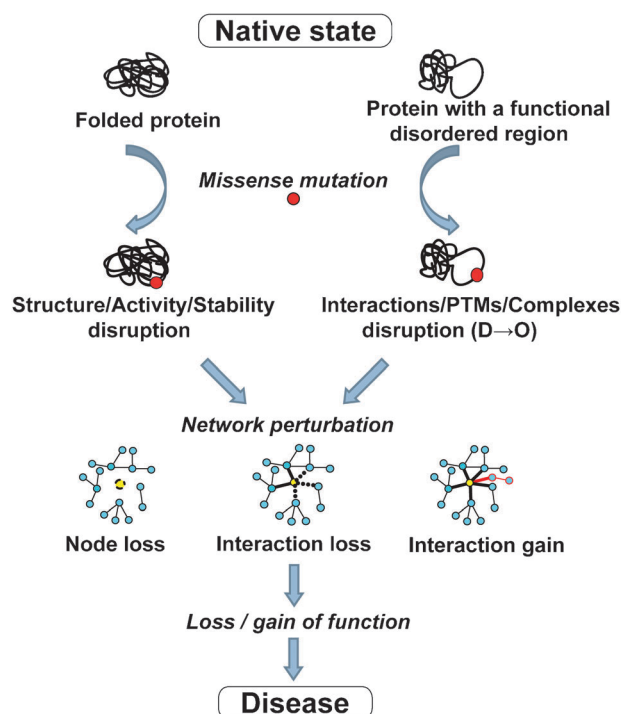


Fig. 1 Schematic diagram of the potential impact of disease mutations in ordered and disordered regions.

As shown previously, IDPs are enriched among transcription and translation regulators, nucleotide-binding proteins and proteins involved in signal transduction.^{12,15,34} By affecting DNA-binding properties of these IDPs, D → O mutations could disrupt transcriptional regulatory networks that control global gene expression. All of these and other ways of network disruptions via D → O mutations in IDPs could trigger disease development.

Another important observation that followed from our analysis of the disease-associated mutations in UniProt is the increased frequency of several specific mutations. When disease mutations were ranked according to their frequency of occurrence in the UniProt database, top five disorder-to-order transition mutations (R → W, R → C, E → K, R → H and R → Q) collectively accounted for 44.0% of all D → O disease mutations (Table 2). Similarly, top five order-to-disorder (O → D) transition mutations (L → P, C → R, G → R, W → R and G → E) collectively accounted for 32.2% of all O → D disease mutations (Table 2). This demonstrates that a limited set of the specific “transition” mutations accounts for a large fraction of D → O and O → D disease mutations. We believe that this observation is important to consider

while developing the classifiers of the functional impact of mutations on protein structure and function, and knowing the preferential “from-to” residue transition could help to better predict which newly discovered mutation is likely to be deleterious. Below, we discuss one example from Table 1, Methyl CpG-binding protein 2 (MeCP2), with three D → O transition mutations that are mapped to its annotated disordered regions from DisProt.

MeCP2 is a methylated DNA-binding protein that mediates transcriptional repression via interaction with the histone deacetylase and is essential for embryonic development. MeCP2 carries a number of missense, nonsense, frame shift and copy number mutations which are associated with various neurodevelopmental disorders such as Rett syndrome, autism spectrum disorders and mental retardation.^{42–44}

The structure and disorder of MeCP2 have been extensively investigated. About 60% of its sequence is intrinsically unstructured, as determined by various experimental methods (CD, NMR, analytical ultracentrifugation and far-UV CD spectroscopy)^{45,46} (Fig. 2). The NMR and X-ray crystal structure of the methyl-CpG binding domain (MBD) of MeCP2 has been solved,^{47,48} and the coordinates of the termini of this domain and several internal residues within MBD could not be assigned, which indicates some amount of disorder even within this structured domain. There are three D → O transition mutations in UniProt that map to the disordered regions of MeCP2 annotated in the DisProt database, R306C, R306H and R453Q (Table 1, Fig. 2). When introduced into the wild type MeCP2 sequence *in silico*, both R306C and R306H mutations result in a dramatic drop of the disorder score in the 207–310 disordered region, which corresponds to a transcriptional repressor domain (TRD) of MeCP2⁴⁹ (Fig. 2). The R453Q mutation causes a drop and a shift in the position of the disordered region 335–486, or C-terminal (CTD-β) domain⁴⁶ (Fig. 2). Both of these domains have been shown to be important for binding to unmethylated DNA, and the synergistic binding to DNA was observed for the TRD-CTD construct, which binds to DNA with 6-fold higher affinity than TRD and 30-fold higher affinity than CTD alone.⁴⁶ Furthermore, it has been demonstrated that CTD-β domain binds to nucleosomes,⁴⁶ most likely to histone H3.⁵⁰ Given the important functional roles of these two domains, the D → O transition mutations could lead to partial or complete impairment of their DNA- and nucleosome-binding properties. Confidently establishing links between specific D → O (and O → D) transition mutations and disruption of domain or protein function using experimental methods is an important step for better understanding the disease mechanisms.

Table 2 Frequencies of the top five D → O and O → D disease mutations from the UniProt database

Substitution	D → O disease mutations (%)	Substitution	O → D disease mutations (%)
R → W	13.1	L → P	11.9
R → C	10.3	C → R	6.6
R → H	7.6	G → R	6.1
E → K	6.7	W → R	4.1
R → Q	6.3	F → S	3.6

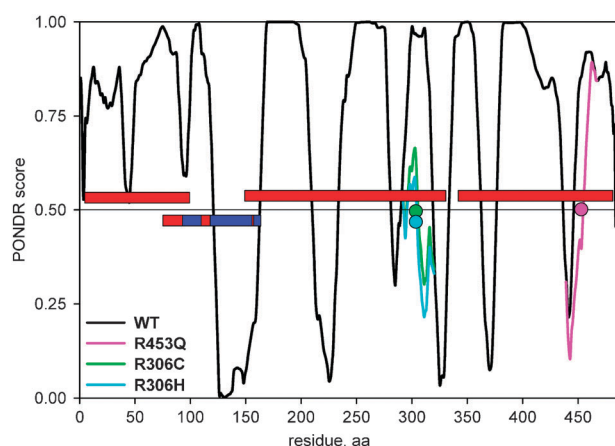


Fig. 2 PONDR VLXT disorder predictions for the wild-type MEC2 and its three D → O mutants that overlap with the experimentally confirmed MEC2 disordered regions from DisProt. The red bars denote the location of the disordered regions, the red/blue bar denotes the MEC2 region for which the structure was determined (red regions within the blue bar indicate approximate location of the disordered regions in the structure), the colored circles denote the location of mutations, the colored lines denote new disorder predictions after introducing mutation into the wild type sequence *in silico*.

Knowing the functional impact of disease mutations in disordered regions has another interesting implication. Recent literature suggests that disordered regions could serve as drug targets for small molecules and short peptides.^{51,52} The potential to target disordered regions carrying disease mutations opens a broad range of possibilities in terms of prioritizing the regions with the most deleterious mutations as drug targets; directing the binding of small molecules towards specific D → O mutations; or even attempting to compensate for the interactions that may be disrupted by such D → O mutations. Since the area of drug development targeting disordered regions, and especially D → O mutations within them, is still largely unexplored, there are many opportunities for future research in this respect.

It is an extremely exciting time for discovery of mutations associated with human diseases. Recent advances in next-generation DNA sequencing technologies⁵³ are bringing a complete catalog of individual genetic variation within reach,⁵⁴ and the decrease in sequencing cost is allowing studies of ever larger disease cohorts.⁵⁵ As the list of mutations associated with human diseases grows, it is likely that some of them will be mapped to protein-coding regions, and a subset of them specifically to disordered regions. However, interpreting disease risk associated with the identified genetic variants still remains a formidable challenge.

Thus, further development of methods to predict functional impact of newly discovered SNPs, especially in disordered regions, is critically needed. This is all the more warranted by the fact that disordered regions have fewer evolutionary constraints compared to ordered regions,⁵⁶ but nevertheless they could carry deleterious mutations, as demonstrated above. We propose that more specialized predictors trained using properties and features of mutations in ordered and disordered regions would be better suited for this purpose than the “one-size-fits-all” models. They are likely to outperform the methods developed to target both ordered and disordered regions without discrimination, because the spectrum of mutations and their functional consequences differ dramatically between these two types of structures. The available domain–domain, protein–protein, and possibly even network-level interaction information should ideally be accounted for while developing such predictors. The first step in this direction has recently been made by incorporating some of the unstructured regions’ properties as training features of the predictor.²⁶

We believe that it is very important that the structure-centric view of mutations changes to account for disease mutations in disordered regions. Although the focus of this opinion was on missense mutations, it is also necessary to recognize that the entire gamut of disease-related mutations including splice-site mutations,

indels, nonsense mutations, and copy number variation could impact disordered regions of proteins in a similar way as they are impacting ordered regions, however with likely varying outcomes. There is still much awaiting to be explored in the area of disease mutations and protein disorder. More rigorous computational and experimental studies integrating genomic, biophysical and biochemical data would contribute to a better understanding of the role of mutations in disordered regions and their relevance to human diseases.

Acknowledgements

This work was supported in whole or in part with the following grants: NSF MCB0444818 (LMI), NIH RO1 HD065288 (LMI), NIH RO1 MH091350 (LMI). We would like to thank the anonymous reviewers for their help in improving the manuscript. Molecular Kinetics Inc is acknowledged for allowing us to use PONDR VLXT predictor.

References

- 1 M. M. Pentony, J. Ward and D. T. Jones, *Methods Mol. Biol.*, 2010, **604**, 369–393.
- 2 L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O’Connor, J. G. Sikes, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res.*, 2004, **32**, 1037–1049.
- 3 P. Radivojac, V. Vasic, C. Haynes, R. R. Cocklin, A. Mohan, J. W. Heyen, M. G. Goebel and L. M. Iakoucheva, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 365–580.
- 4 A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, *Biochemistry*, 2002, **41**, 6573–6582.
- 5 H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 197–208.
- 6 P. Tompa, *Trends Biochem. Sci.*, 2002, **27**.
- 7 A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *FEBS J.*, 2005, **272**, 5129–5148.
- 8 C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal and L. M. Iakoucheva, *PLoS Comput. Biol.*, 2006, **2**, e100.
- 9 D. Ekman, S. Light, A. K. Bjorklund and A. Elofsson, *Genome Biology*, 2006, **7**, R45.
- 10 C. J. Brown, A. K. Johnson, A. K. Dunker and G. W. Daughdrill, *Curr. Opin. Struct. Biol.*, 2011.
- 11 J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers and P. M. Kim, *Genome Biology*, 2011, **12**, R14.
- 12 L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker, *J. Mol. Biol.*, 2002, **323**, 573–584.
- 13 V. N. Uversky, C. J. Oldfield and A. K. Dunker, *Annu. Rev. Biophys.*, 2008, **37**, 215–246.

- 14 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.
- 15 J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, 2004, **337**, 635–645.
- 16 A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Informat.*, 2000, **11**, 161–171.
- 17 R. E. Steward, M. W. MacArthur, R. A. Laskowski and J. M. Thornton, *Trends Genet.*, 2003, **19**, 505–513.
- 18 T. R. Rebbeck, M. Spitz and X. Wu, *Nat. Rev. Genet.*, 2004, **5**, 589–597.
- 19 Z. Wang and J. Moul, *Hum. Mutat.*, 2001, **17**, 263–270.
- 20 D. Chasman and R. M. Adams, *J. Mol. Biol.*, 2001, **307**, 683–706.
- 21 C. Ferrer-Costa, M. Orozco and X. de la Cruz, *J. Mol. Biol.*, 2002, **315**, 771–786.
- 22 C. T. Saunders and D. Baker, *J. Mol. Biol.*, 2002, **322**, 891–901.
- 23 P. C. Ng and S. Henikoff, *Nucleic Acids Res.*, 2003, **31**, 3812–3814.
- 24 S. Sunyaev, V. Ramensky and P. Bork, *Trends Genet.*, 2000, **16**, 198–200.
- 25 P. Yue, E. Melamud and J. Moul, *BMC Bioinformatics*, 2006, **7**, 166.
- 26 M. Mort, U. S. Evani, V. G. Krishnan, K. K. Kamati, P. H. Baenziger, A. Bagchi, B. J. Peters, R. Sathyesh, B. Li, Y. Sun, B. Xue, N. H. Shah, M. G. Kann, D. N. Cooper, P. Radivojac and S. D. Mooney, *Hum. Mutat.*, 2010, **31**, 335–346.
- 27 P. C. Ng and S. Henikoff, *Genome Res.*, 2001, **11**, 863–874.
- 28 S. Li, L. M. Iakoucheva, S. D. Mooney and P. Radivojac, *Pac Symp. Biocomput.*, 2010, 337–347.
- 29 Y. L. Yip, M. Famiglietti, A. Gos, P. D. Duek, F. P. David, A. Gateau and A. Bairoch, *Hum. Mutat.*, 2008, **29**, 361–366.
- 30 P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins: Struct., Funct., Genet.*, 2001, **42**, 38–48.
- 31 Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac and A. K. Dunker, *Proteins*, 2005.
- 32 Z. Dosztanyi, V. Csizmek, P. Tompa and I. Simon, *Bioinformatics*, 2005.
- 33 M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res.*, 2007, **35**, D786–793.
- 34 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome Res.*, 2007, **6**, 1882–1898.
- 35 P. E. Wright and H. J. Dyson, *Curr. Opin. Struct. Biol.*, 2009, **19**, 31–38.
- 36 P. Csermely, R. Palotai and R. Nussinov, *Trends Biochem. Sci.*, 2010, **35**, 539–546.
- 37 A. L. Pey, F. Stricher, L. Serrano and A. Martinez, *Am. J. Hum. Genet.*, 2007, **81**, 1006–1024.
- 38 Q. Zhong, N. Simonis, Q. R. Li, B. Charleatoux, F. Heuze, N. Klitgord, S. Tam, H. Yu, K. Venkatesan, D. Mou, V. Swearingen, M. A. Yildirim, H. Yan, A. Dricot, D. Szeto, C. Lin, T. Hao, C. Fan, S. Milstein, D. Dupuy, R. Brasseur, D. E. Hill, M. E. Cusick and M. Vidal, *Mol. Syst. Biol.*, 2009, **5**, 321.
- 39 K. M. Daily, P. Radivojac and A. K. Dunker, Intrinsic disorder and protein modifications: building an SVM predictor for methylation, *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, CA, 2005.
- 40 J. Gsponer, M. E. Futschik, S. A. Teichmann and M. M. Babu, *Science*, 2008, **322**, 1365–1368.
- 41 P. Tompa and M. Fuxreiter, *Trends Biochem. Sci.*, 2008, **33**, 2–8.
- 42 R. E. Amir, I. B. Van den Veyver, M. Wan, C. Q. Tran, U. Francke and H. Y. Zoghbi, *Nat. Genet.*, 1999, **23**, 185–188.
- 43 M. B. Ramocki, S. U. Peters, Y. J. Tavyev, F. Zhang, C. M. Carvalho, C. P. Schaaf, R. Richman, P. Fang, D. G. Glaze, J. R. Lupski and H. Y. Zoghbi, *Ann. Neurol.*, 2009, **66**, 771–782.
- 44 C. W. Lam, W. L. Yeung, C. H. Ko, P. M. Poon, S. F. Tong, K. Y. Chan, I. F. Lo, L. Y. Chan, J. Hui, V. Wong, C. P. Pang, Y. M. Lo and T. F. Fok, *J. Med. Genet.*, 2000, **37**, E41.
- 45 V. H. Adams, S. J. McBryant, P. A. Wade, C. L. Woodcock and J. C. Hansen, *J. Biol. Chem.*, 2007, **282**, 15057–15064.
- 46 R. P. Ghosh, T. Nikitina, R. A. Horowitz-Scherer, L. M. Gierasch, V. N. Uversky, K. Hite, J. C. Hansen and C. L. Woodcock, *Biochemistry*, 2010, **49**, 4395–4410.
- 47 R. I. Wakefield, B. O. Smith, X. Nan, A. Free, A. Soteriou, D. Uhrin, A. P. Bird and P. N. Barlow, *J. Mol. Biol.*, 1999, **291**, 1055–1065.
- 48 K. L. Ho, I. W. McNae, L. Schmiedebeg, R. J. Klose, A. P. Bird and M. D. Walkinshaw, *Mol. Cell*, 2008, **29**, 525–531.
- 49 X. Nan, F. J. Campoy and A. Bird, *Cell*, 1997, **88**, 471–481.
- 50 T. Nikitina, X. Shi, R. P. Ghosh, R. A. Horowitz-Scherer, J. C. Hansen and C. L. Woodcock, *Mol. Cell. Biol.*, 2007, **27**, 864–877.
- 51 S. J. Metallo, *Curr. Opin. Chem. Biol.*, 2010, **14**, 481–488.
- 52 A. K. Dunker and V. N. Uversky, *Curr. Opin. Pharmacol.*, 2010, **10**, 782–788.
- 53 J. Shendure and H. Ji, *Nat. Biotechnol.*, 2008, **26**, 1135–1145.
- 54 The Thousand Genomes Project Consortium, *Nature*, 2010, **467**, 1061–1073.
- 55 G. M. Cooper and J. Shendure, *Nat. Rev. Genet.*, 2011, **12**, 628–640.
- 56 C. J. Brown, A. K. Johnson and G. W. Daughdrill, *Mol. Biol. Evol.*, 2010, **27**, 609–621.