Structural Capacitance in Protein Evolution and Human Diseases

Chen Li[1,2], Liah V.T. Clark[1], Rory Zhang[1], Benjamin T. Porebski[1,3], Julia M. McCoey[1], Natalie A. Borg[1], Geoffrey I. Webb[4], Itamar Kass[1,5], Malcolm Buckle[6], Jiangning Song[1], Adrian Woolfson[7] and Ashley M. Buckle[1]

*1 - **Department of Biochemistry and Molecular Biology,** Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia*
*2 - **Department of Biology,** Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland*
*3 - **Medical Research Council Laboratory of Molecular Biology,** Francis Crick Avenue, Cambridge, CB2 0QH, UK*
*4 - **Faculty of Information Technology,** Monash University, Clayton, Victoria 3800, Australia*
*5 - **Amai Proteins,** Prof. A. D. Bergman 2B, Suite 212, Rehovot 7670504, Israel*
*6 - **LBPA,** ENS Cachan, CNRS, Université Paris-Saclay, F-94235 Cachan, France*
*7 - **Nouscom,** Baumleingasse, CH-4051 Basel, Switzerland*

***Correspondence to Adrian Woolfson and Ashley M. Buckle:** adrianwoolfson@yahoo.com; ashley.buckle@monash.edu*
https://doi.org/10.1016/j.jmb.2018.06.051
Edited by Dan Tawfik

Abstract

Canonical mechanisms of protein evolution include the duplication and diversification of pre-existing folds through genetic alterations that include point mutations, insertions, deletions, and copy number amplifications, as well as post-translational modifications that modify processes such as folding efficiency and cellular localization. Following a survey of the human mutation database, we have identified an additional mechanism that we term "structural capacitance," which results in the *de novo* generation of microstructure in previously disordered regions. We suggest that the potential for structural capacitance confers select proteins with the capacity to evolve over rapid timescales, facilitating saltatory evolution as opposed to gradualistic canonical Darwinian mechanisms. Our results implicate the elements of protein microstructure generated by this distinct mechanism in the pathogenesis of a wide variety of human diseases. The benefits of rapidly furnishing the potential for evolutionary change conferred by structural capacitance are consequently counterbalanced by this accompanying risk. The phenomenon of structural capacitance has implications ranging from the ancestral diversification of protein folds to the engineering of synthetic proteins with enhanced evolvability.

Introduction

Canonical protein evolution is achieved through the utilization of an array of different genetic mechanisms, including point mutations, recombination, translocations, and duplication. Alterations to the protein-encoding elements of genes are accompanied by modifications to non-coding control elements, epigenetic changes, and post-translational mechanisms that refine the kinetics, spatial localization, synthesis, folding efficiency, and other aspects of protein synthesis and dynamics. The functional and morphological diversity at the protein, cellular, and organismal levels is achieved through the utilization of a combination of such mechanisms. This repertoire of available mechanisms for the implementation of evolutionary change enables adaptations to be furnished at the appropriate level, and allows proteins to efficiently navigate their function spaces to identify optimal phenotypic solutions.

Whereas classic genetic modifications are incremental, more complex mechanisms, such as the phenomenon of genetic capacitance mediated by heat shock proteins like Hsp90, buffer the impact of polymorphisms that may in isolation be maladaptive, allowing their phenotype to remain hidden and offering the possibility for "saltatory" evolutionary change [1]. This enables expansive regions of structural and functional space to be navigated in a single step, releasing phenotypes from local optima and allowing new functions and

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

morphologies to evolve over rapid timescales along routes requiring the simultaneous presence of multiple genetic alterations. The existence of such mechanisms enables evolutionary landscapes [1] to be navigated more efficiently, as searches may be extended beyond the local optima of rugged peaks, so as to reach out across otherwise un-navigable regions of sequence space to identify distant, and potentially more adaptive optima. The capacity for this type of accelerated evolution is especially important in environments characterized by high uncertainty, where a capacity for furnishing rapid and efficient evolutionary change is a prerequisite for survival. Interestingly, there is emerging evidence for profound protein structural changes induced by so-called "hopeful monster" mutations [2–4].

Many proteins, and most notably those involved in cellular regulation and signaling [5,6], contain significant regions of disorder and belong to the class of intrinsically disordered proteins (IDPs) [7]. IDPs can undergo disorder–order transitions, typically upon binding another protein or DNA (coupled folding and binding [8–11]). The energy landscapes of IDPs are typically rugged, featuring a continuum of conformational states that enables interaction with other molecules via both conformational selection and induced fit [8,12,13].

In order to establish whether point mutations within the regions encoding disordered regions may result in microstructuralization and generate nascent microstructural elements that may form substrates for evolution and result in adaptive alterations to protein function, we performed a survey of the human mutation database [14]. Specifically, we performed a bioinformatic analysis to identify mutations predicted to generate localized regions of microstructure in previously disordered regions of target proteins. We report here a new mechanism of protein evolution, which we term "structural capacitance," whereby structural and functional changes at the level of individual proteins may be achieved through the introduction of point mutations influencing key nucleating amino acids located in regions of structural disorder. Once mutated, these residues are predicted to generate new microstructural elements in previously disordered regions that are functionally distinct from the parent fold. These findings have broad implications both for the accelerated non-canonical evolution of protein folds, and for the pathogenesis of human diseases.

Results

Order–disorder transitions associated with mutations in human proteins

We interrogated the human polymorphisms and disease mutations dataset [14], and compiled a dataset of 68,383 unique human mutations (excluding the "unclassified" mutations; see Materials and Methods for more details), comprising 28,662 human disease mutations and 39,721 polymorphisms. We then applied standard algorithms for disorder prediction to every mutation and polymorphism in the dataset. A predictor voting strategy was employed to determine the prediction outcome for each mutation. As there are multiple predictors for protein disorder, the residues were deemed to be located within disordered regions if the number of predictors assigning residues to disordered regions was equal to or larger than the number of predictors assigning the residues to ordered regions. Four types of structural transitions were defined: disorder-to-order (D→O), order-to-disorder (O→D), disorder-to-disorder (D→D), and order-to-order (O→O) (Table S1).

We next interrogated the subset of proteins containing D→O predicted mutations imposing the following selection criteria: (1) mutations located within disordered regions ≥30 amino acids in length (termed "long disordered regions" (LDRs), consistent with other studies [15–19]); (2) LDRs not predicted to be in transmembrane domains; and (3) proteins with LDRs lacking experimentally determined identical or homologous structures. From a BLAST search against the protein databank (PDB), 226 of 1,337 (16.9%) proteins do not currently have experimentally determined structures or homologues. Among 1337 proteins containing predicted D→O mutations, we identified 150 point mutations within LDRs from a total of 131 proteins (Table 1). The workflow is detailed in Fig. S1.

In order to determine whether any of the mutation sites are located in functionally relevant regions, we cross-referenced disease and non-disease mutations in D→O, O→D, D→D, and O→O transitions with the Eukaryotic Linear Motif (ELM) database [20]. ELMs are predominantly functional modules found in intrinsically disordered regions in eukaryotic proteins [21]. All ELMs listed have been experimentally verified (i.e., annotated with experimental evidence showing that the ELM is involved in a functionally relevant interaction). The numbers of mutations found in ELMs were as follows: 13/1,731 (0.75%), 74/13,876 (0.53%), 70/51,317 (0.14%), and 3/1,459 (0.21%) for D→O, D→D, O→O, and O→D mutations, respectively (Tables S2–S5). The number of identified motifs for D→O mutations is much smaller than that for D→D and O→O; however, this most likely reflects the relative scarcity of D→O mutations. For three transitions, more disease mutations than non-disease mutations/polymorphisms were found in ELMs according to the one-tailed Fisher exact test [8 *versus* 5 for D→O disease-associated mutations *versus* polymorphisms (p-value = 0.02); 3 *versus* 0 for O→D disease-associated mutations *versus* polymorphisms (p-value = 0.09); 46 *versus* 28 for D→D disease-

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

associated mutations *versus* polymorphisms (*p*-value = 1.55E−15); 43 *versus* 27 for O→O disease-associated mutations *versus* polymorphisms (*p*-value = 0.02)]. Overall, therefore, we predict that only a relatively small fraction of the identified D→O mutations are part of functionally relevant interactions with other proteins. To further investigate the numbers of mutations that are located in the annotated Pfam domains, we mapped the mutations of four transitions to the Pfam database [22]. The numbers of mutations found in the Pfam domains were as follows: 741/1731 (42.8%; 308 *versus* 433 for disease-associated mutations *versus* polymorphisms), 3538/13,876 (25.5%; 1333 *versus* 2205 for disease-associated mutations *versus* polymorphisms), 800/1459 (54.8%; 474 *versus* 326 for disease-associated mutations *versus* polymorphisms), and 35,521/51,317 (69.2%; 19,500 *versus* 16,021 for disease-associated mutations *versus* polymorphisms) for D→O, D→D, O→D, and O→O, respectively. The mapping results for D→O and O→D transitions are shown in Tables S6 and S7, respectively. On average, approximately half (48.1%) of the mutations were located in Pfam annotated domains.

Identification of structural capacitance elements

We propose that the identified LDRs (Table 1) represent a new class of genetic element, which we have termed "structural capacitance element" (SCE). The D→O mutations identified represent examples of order-inducing substitutions that introduce new microstructure into the parent fold that may confer new functions, or refine existing ones, but which may, in some cases, be of pathogenic significance. There are 21 mutations involving cysteine residues (i.e., X→C and C→X, where X denotes any amino acids) identified within 21 proteins (all involving a single mutation to cysteine; Table 1). None of the identified D→O mutations in LDRs (Table 1) are predicted to be associated with ELMs (Tables S2), indicating that these substitutions are unlikely to interfere with known interactions, resulting instead in the potential for functionality through the generation of new microstructures. The mechanism appears to be initiated by point mutations that change hydrophilic nucleating residues to hydrophobic ones (Fig. 1a). It remains to be seen whether the codons encoding these residues constitute "hotspot" regions within the human genome, and whether there is codon bias at these positions.

The types of mutations in each class (D→O, O→D, D→D, and O→O) appear to be non-random (Table S8). For all documented D→O disease mutations, arginine is the most frequently mutated amino acid (Fig. 1b). The most common classes of disease mutation for D→O and O→D transitions are R→W (59; 11.50%) and L→P (108; 16.62%), respectively (Tables S8A and S8B). For O→O and D→D transitions, the mutation patterns are more evenly distributed (Tables S8C and S8D). For non-disease mutations, the most common type is P→L (153; 12.56%) for D→O and L→P (63; 7.79%) for O→D (Tables S8A and S8B). This is consistent with a recent comparison of mutation frequencies in intrinsically disordered regions of proteins in both disease and non-disease datasets that highlights the previously unappreciated role of mutations in disordered regions [11]. An example of the predicted local increase in ordering due to D→O mutations in both disease and non-disease datasets is shown in Fig. 2a and b.

Given the variability in the predictions among the four predictors tested, which necessitated a majority voting approach, we looked for experimental evidence suggesting that the predicted regions were disordered. Accordingly, we cross-referenced our human disease mutations and polymorphisms dataset against DisProt [23], a database providing experimentally verified disordered regions of proteins. For the resulting matches, we applied four protein disordered region predictors to predict the structural changes following mutation events, using majority voting. Disorder prediction using majority voting predicts that 108 mutations in LDRs result in a D→O structural transition (Table S9). Compared to Table 1, the D→O mutations listed in Table S9 are more reliable, as the LDRs harboring such mutations have been experimentally verified. Examples showing an increase in ordering according to DynaMine based on the DisProt database are shown in Fig. S2.

Discussion

The association between disease phenotypes and the *de novo* formation of protein microstructural elements represents a novel paradigm for understanding the origins of select human diseases, which has previously principally been focused on loss-of-structure and accompanying loss of function. One of the best-studied examples is the tumor suppressor p53, which is inactivated following somatic mutation events in a range of human cancers [24]. Most p53 mutations are loss-of-function mutations impacting the DNA-binding domain through interference with p53–DNA contacts or structural destabilization [25]. It is, however, known that IDPs play a role in cellular regulation and signaling [5,6]. Indeed structural changes in disordered proteins have been implicated in disease processes, with evidence for D→O transitions triggered by disease-causing mutations in functionally important regions (such as regions mediating protein–protein recognition via coupled folding and binding, and DNA binding [10]).

We propose that in addition to canonical loss-of-function mutations, disease-causing mutations may result in gain of function through the binary activation

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

Table 1. Mutations in LDRs of human proteins predicted to produce a D➔O transition[a]

UniProt/dbSNP	Protein	Mutation	Disease	No. disorder predictors[b]	No. order predictors[c]	Average length of LDR[d]	No. disorder predictors in D2P2[e]	No. disorder predictors for LDR[f]
A0JNW5/rs7296162	UHRF1-binding protein 1-like	S1147L	–	4	2[g]	101	6	3
A4D1E1/rs801841	Zinc finger protein 804B	V1195I	–	3[h]	2[g]	37	6	1
A6NJV1/rs2272466	UPF0573 protein C2orf70	Q177L	–	2[h]	4	34	3	1
A7E2F4/rs347880	Golgin subfamily A member 8A	K480N	–	2[h]	2[g]	91	N/A	2
O14645/rs11749	Axonemal dynein light intermediate polypeptide 1	A65V	–	3[h]	3	43	N/A	2
O15078/rs374852145	Centrosomal protein of 290 kDa	R2210C	–	2	3	123	5	1
O15287/rs4986939	Fanconi anemia group G protein	S378L	–	2	2[g]	35	5	1
O43303/rs3751821	Centriolar coiled-coil protein of 110 kDa	P171L	–	4	2[g]	39	7	2
O43734/rs13190932	Adapter protein CIKS	R83W	–	2[h]	3[g]	189	5	2
O60240/rs6496589	Perilipin-1	P194A	–	3[h]	2[g]	32	8	1
O60269/rs4445576	G protein-regulated inducer of neurite outgrowth 2	S328C	–	2[h]	2[g]	31	6	1
O60287/rs762225	Nucleolar pre-ribosomal-associated protein 1	P2071R	–	2	3[g]	53	6	1
O75161/rs17472401	Nephrocystin-4	R848W	–	3[h]	2[g]	45	6	1
O75161/rs571655	Nephrocystin-4	E618K	–	3[h]	2[g]	30	7	1
O75952/rs3786417	Calcium-binding tyrosine phosphorylation-regulated protein	T74M	–	2[h]	2[g]	113	5	2
P07498/rs1048152	Kappa-casein	R110L	–	3[h]	2[g]	35	5	1
P08F94/rs146680689	Fibrocystin	R3957C	Polycystic kidney disease; autosomal recessive (ARPKD) [MIM:263200]	3[h]	4	40	7	1
P10909/rs9331936	Clusterin	N317H	–	2	2[g]	48	4	1
P15502/rs17855988	Elastin	G610R	–	2	2[g]	385	N/A	1
P28290/rs13419020	Sperm-specific antigen 2	R833W	–	3	2[g]	224	5	2
P28290/rs17647806	Sperm-specific antigen 2	P836L	–	4	2[g]	169	7	2
P41440/rs35786590	Folate transporter 1	A558V	–	3	2[g]	51	8	1
P55327/rs35099105	Tumor protein D52	D52Y	–	3[h]	2[g]	55	8	2
P59020/rs13864	Down syndrome critical region protein 9	R76L	–	3[h]	3	31	N/A	1
Q0P670/rs13290	Uncharacterized protein SPEM2	S108A	–	2	2[g]	67	4	1
Q0VG06/rs11552304	Fanconi anemia core complex-associated protein 100	P660L	–	2	2[g]	55	4	1
Q13111/rs35651457	Chromatin assembly factor 1 subunit A	D167V	–	2	3	84	5	2
Q14D04/rs59504298	Ventricular zone-expressed PH domain-containing protein homolog 1	S501L	–	3[h]	2[g]	68	8	2

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

Q15154/rs370429	Pericentriolar material 1 protein	T1543I	–	2	4	41	N/A	1
Q15572/rs4150167	TATA box-binding protein-associated factor RNA polymerase I subunit C	G523R	–	2	2[g]	41	5	1
Q15884/rs35386391	Protein FAM189A2	T233I	–	3[h]	3	82	6	3
Q17RF5/rs2306175	Uncharacterized protein C4orf26	P30L	–	3[h]	3	50	8	2
Q3BBV0/rs11581926	Neuroblastoma breakpoint family member 1	Q850K	–	2[h]	3[g]	131	N/A	1
Q49MG5/rs2305050	Microtubule-associated protein 9	N601D	–	2[h]	2[g]	104	8	1
Q52LG2/rs16986753	Keratin-associated protein 13-2	R26C	–	2	2[g]	88	1	1
Q52M75/rs17366761	Putative uncharacterized protein encoded by LINC01554	R85C	–	2	2[g]	49	5	1
Q53HC0/rs11057401	Coiled-coil domain-containing protein 92	S70C	–	2[h]	3	159	6	1
Q562F6/rs1036533	Shugoshin 2	G9D	–	2	2[g]	55	3	1
Q569K6/rs12167903	Coiled-coil domain-containing protein 157	P191L	–	2[h]	3	41	5	1
Q5JSZ5/rs10736851	Protein PRRC2B	S1630T	–	2[h]	3	183	8	2
Q5SNV9/rs6697244	Uncharacterized protein C1orf167	S848I	–	3	2[f]	30	N/A	1
Q5SQ13/rs11787585	Proline-rich protein 31	L8F	–	2[h]	2[g]	150	6	2
Q5SQN1/rs17851681	Synaptosomal-associated protein 47	R381C	–	4	2[g]	42	5	1
Q5SRN2/rs1003878	Uncharacterized protein C6orf10	P161L	–	2[h]	2[g]	57	5	2
Q5SRN2/rs1033500	Uncharacterized protein C6orf10	P128L	–	2[h]	2[g]	52	6	1
Q5T1N1/rs9440631	Protein AKNAD1	H255Y	–	2[h]	3	193	3	2
Q5THJ4/rs958068	Vacuolar protein sorting-associated protein 13D	S1707F	–	2	2[g]	67	6	1
Q5VWN6/rs56856085	Protein FAM208B	S724Y	–	4	2[g]	144	7	3
Q5VWP3/rs2275769	Muscular LMNA-interacting protein	P376S	–	2[h]	2[g]	163	4	1
Q5VXU9/rs7470491	Uncharacterized protein C9orf84	H416R	–	2	2[g]	86	4	1
Q5VYM1/rs10117097	Uncharacterized protein C9orf131	L285F	–	2[h]	2[g]	256	6	1
Q5W0A0/rs3014939	Glutamate-rich protein 6B	E178K	–	2[h]	2[g]	177	6	1
Q5W0A0/rs749071	Glutamate-rich protein 6B	T427I	–	2[h]	3	45	4	2
Q6L8H2/rs7113784	Keratin-associated protein 5-3	S83C	–	2	2[g]	122	5	1
Q6PJF5/rs11553545	Inactive rhomboid protein 2	D528Y	–	2[h]	4	48	5	1
Q6PJF5/rs387907130	Inactive rhomboid protein 2	P189L	Tylosis with esophageal cancer (TOC) [MIM:148500]	3[h]	3	109	6	2
Q6PJW8/rs12075111	Consortin	R399C	–	3[h]	2[g]	129	7	2
Q6PK04/rs11150805	Coiled-coil domain-containing protein 137	R177W	–	3[h]	3	125	9	3
Q6ZP01/rs13393001	RNA-binding protein 44	D51H	–	2[h]	2[g]	72	N/A	1

(continued on next page)

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

Table 1 (*continued*)

UniProt/dbSNP	Protein	Mutation	Disease	No. disorder predictors[b]	No. order predictors[c]	Average length of LDR[d]	No. disorder predictors in D2P2[e]	No. disorder predictors for LDR[f]
Q6ZR62/rs7474140	Retrotransposon Gag-like protein 4	D162Y	–	2[h]	2[g]	100	4	1
Q6ZU52/rs2236026	Uncharacterized protein KIAA0408	S331L	–	2	2[g]	134	3	1
Q6ZUB1/rs34791830	Spermatogenesis-associated protein 31E1	A736V	–	4	2[g]	133	8	3
Q6ZVD7/rs41278532	Storkhead-box protein 1	N825I	Pre-eclampsia/ eclampsia 4 (PEE4) [MIM:609404]	2	2[g]	83	5	1
Q702N8/rs60540208	Xin actin-binding repeat-containing protein 1	R695C	–	2[h]	3	88	5	1
Q702N8/rs9823779	Xin actin-binding repeat-containing protein 1	R776W	–	3[h]	2[g]	38	6	3
Q711Q0/rs56206226	Uncharacterized protein C10orf71	R320L	–	2[h]	2[g]	440	6	2
Q7Z2D5/rs35285687	Phospholipid phosphatase-related protein type 4	A32V	–	2[h]	2[g]	41	6	1
Q7Z2Z1/rs3743372	Treslin	R1885C	–	2[h]	2[g]	551	6	1
Q7Z3Z2/rs34049451	Protein RD3	R47C	–	2[h]	2[g]	32	6	1
Q7Z402/rs17854512	Transmembrane channel-like protein 7	R59W	–	3[h]	2[g]	49	8	1
Q7Z403/rs34712518	Transmembrane channel-like protein 6	G191D	–	2	2[g]	40	6	1
Q7Z570/rs12105159	Zinc finger protein 804A	G1152R	–	2	2[g]	77	4	1
Q7Z570/rs12476147	Zinc finger protein 804A	Q261L	–	2	2[g]	58	4	1
Q7Z7L8/rs12796667	Uncharacterized protein C11orf96	R144C	–	2[h]	3	155	5	2
Q86TY3/rs3829765	Uncharacterized protein C14orf37	T96I	–	3[h]	2[g]	75	6	3
Q86V40/rs1649292	Metalloprotease TIKI1	P430L	–	2	2[g]	57	6	1
Q8IV16/rs587777636	Glycosylphosphatidylinositol-anchored high density lipoprotein-binding protein 1	G56R	–	2	2[g]	48	6	1
Q8IV16/rs78367243	Glycosylphosphatidylinositol-anchored high density lipoprotein-binding protein 1	S144F	Hyperlipoproteinemia 1D (HLPP1D) [MIM:615947]	3[h]	4	38	8	1
Q8IWZ6/rs119466001	Bardet-Biedl syndrome 7 protein	H323R	Bardet-Biedl syndrome 7 (BBS7) [MIM:615984]	2[h]	2[g]	38	5	1
Q8IXS0/rs10485172	Protein FAM217A	M442V	–	2	2[g]	91	4	1
Q8IYE0/rs1109968	Coiled-coil domain-containing protein 146	N345S	–	2[h]	2[g]	213	6	1
Q8IYE1/rs17853515	Coiled-coil domain-containing protein 13	E375V	–	2[h]	2[g]	92	5	1

Q8IYT3/rs55868409	Coiled-coil domain-containing protein 170	E345K	–	2[h]	2[g]	99	4	1
Q8IYX3/rs861854	Coiled-coil domain-containing protein 116	R96C	–	2	3	136	N/A	2
Q8IZ63/rs3745640	Proline-rich protein 22	P118L	–	2[h]	3	116	5	1
Q8N1H7/rs1033734	Protein SIX6OS1	S309L	–	2[h]	2[g]	198	6	1
Q8N205/rs34818970	Nesprin-4	S224L	–	3[h]	2[g]	74	7	3
Q8N2C7/rs35822936	Protein unc-80 homolog	R131W	–	2[h]	3[g]	48	6	2
Q8N2C7/rs869025316	Protein unc-80 homolog	P1700S	Hypotonia; infantile; with psychomotor retardation and characteristic facies 2 (IHPRF2) [MIM:616801]	2[h]	2[g]	80	3	1
Q8N307/rs11923495	Mucin-20	R666W	–	2	4	163	3	2
Q8N307/rs3762739	Mucin-20	S671C	–	2[h]	4	163	4	2
Q8N387/rs15783	Mucin-15	T202I	–	3[h]	3	37	4	1
Q8N6Y0/rs9676419	Usher syndrome type-1C protein-binding protein 1	M439V	–	2[h]	2[g]	172	5	1
Q8N715/rs17852896	Coiled-coil domain-containing protein 185	R380L	–	2[h]	3	244	5	2
Q8N7R1/rs1689291	POM121-like protein 12	G188E	–	2	2[g]	57	4	1
Q8N8I6/rs2048058	Putative uncharacterized protein encoded by LINC00482	R119C	–	3[h]	3	85	6	2
Q8N9H9/rs1281018	Uncharacterized protein C1orf127	A530V	–	2[h]	3	235	6	2
Q8N9T8/rs34743532	Protein KRI1 homolog	S309L	–	3	2[g]	168	N/A	1
Q8NEV8/rs17108127	Exophilin-5	M512L	–	2[h]	2[g]	256	6	1
Q8NEV8/rs2640785	Exophilin-5	E137V	–	2	2[g]	53	5	1
Q8NEV8/rs3741046	Exophilin-5	R118L	–	2[h]	2[g]	53	5	1
Q8TBE3/rs17054522	Fibronectin type III domain-containing protein 9	P166A	–	3[h]	2[g]	42	8	2
Q8TBZ2/rs9890721	MYCBP-associated protein	R688W	–	4	2[g]	45	N/A	3
Q8TD31/rs130066	Coiled-coil alpha-helical rod protein 1	S164R	–	2	2[g]	166	6	1
Q8TD31/rs130068	Coiled-coil alpha-helical rod protein 1	R417W	–	2[h]	3	45	5	1
Q8TD31/rs130079	Coiled-coil alpha-helical rod protein 1	G575C	–	2	2[g]	151	6	1
Q8TD31/rs2073720	Coiled-coil alpha-helical rod protein 1	K546R	–	2[h]	2[g]	152	6	1
Q8TEV9/rs8080966	Guanine nucleotide exchange protein SMCR8	P524L	–	2[h]	4	166	4	1
Q8TF40/rs12109782	Folliculin-interacting protein 1	V738L	–	2[h]	2[g]	32	5	1

(continued on next page)

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

Table 1 (*continued*)

UniProt/dbSNP	Protein	Mutation	Disease	No. disorder predictors[b]	No. order predictors[c]	Average length of LDR[d]	No. disorder predictors in D2P2[e]	No. disorder predictors for LDR[f]
Q8TF63/rs12520809	Dendritic cell nuclear protein 1	N97D	–	2	2[g]	40	5	2
Q8WVK2/-	U4/U6.U5 small nuclear ribonucleoprotein 27 kDa protein	S114F	–	2[h]	3	135	4	2
Q8WWU5/rs2234045	T-complex protein 11 homolog	G253A	–	2[h]	2[g]	48	6	2
Q8WXD2/rs35664837	Secretogranin-3	M233V	–	2[h]	3	46	5	2
Q8WXH2/rs17853661	Junctophilin-3	P645L	–	4	2[g]	127	6	3
Q92504/rs35690712	Zinc transporter SLC39A7	G124R	–	2	2[g]	90	4	2
Q96FF9/rs34020666	Sororin	S156Y	–	2	2[g]	131	6	1
Q96GE4/rs9910506	Centrosomal protein of 95 kDa	M165I	–	2	2[g]	100	4	1
Q96KD3/rs6949056	Protein FAM71F1	S228L	–	3[h]	2[g]	44	7	1
Q96LP6/rs7484376	Uncharacterized protein C12orf42	P182R	–	2[h]	2[g]	68	5	1
Q96M02/rs11558415	Centrosomal protein C10orf90	M57I	–	2	3	79	5	1
Q99575/rs3824145	Ribonucleases P/MRP protein subunit POP1	S127L	–	2[h]	3[g]	65	5	2
Q9BR39/rs387906898	Junctophilin-2	S165F	Cardiomyopathy; familial hypertrophic 17 (CMH17) [MIM:613873]	4	2[g]	64	7	4
Q9BW71/rs11643314	HIRA-interacting protein 3	G521W	–	4	2[g]	51	9	3
Q9BW71/rs35431046	HIRA-interacting protein 3	A496V	–	2	3	58	4	1
Q9BWW9/rs2076672	Apolipoprotein L5	T323M	–	2[h]	2[g]	114	6	2
Q9H0A9/rs884134	Speriolin-like protein	P113L	–	4	2[g]	72	7	3
Q9H0B3/rs2277921	Uncharacterized protein KIAA1683	P835L	–	2[h]	2[g]	120	6	2
Q9H4K1/rs2142661	RIB43A-like with coiled-coils protein 2	R180C	–	3[h]	3	68	6	1
Q9H799/rs16903518	Protein JBTS17	P2592L	–	3	2[g]	99	5	2
Q9H799/rs377107065	Protein JBTS17	P2750S	–	2	2[g]	50	4	1
Q9HAW4/rs34390044	Claspin	P892T	–	2[h]	3	234	6	1
Q9HBH7/rs1045082	Protein BEX1	M66I	–	2	3	43	4	1
Q9HBH7/rs709036	Protein BEX1	A40V	–	2[h]	2[g]	40	7	2
Q9HCM1/rs3759299	Uncharacterized protein KIAA1551	S1208C	–	3	2[g]	130	7	1

Q9HCM1/rs61353224	Uncharacterized protein KIAA1551	P147S	–	2[n]	2[g]	30	4	1
Q9HCM3/rs2774960	UPF0606 protein KIAA1549	P652L	–	2	2[g]	117	5	2
Q9NSI2/rs3737075	Protein FAM207A	V212L	–	2[h]	2[g]	71	5	1
Q9NWH7/rs1338314	Spermatogenesis-associated protein 6	R333W	–	2	2[g]	154	5	2
Q9NY87/-	Sperm protein associated with the nucleus on the X chromosome C	V59F	–	2[h]	2[g]	95	N/A	2
Q9NZP6/rs36025315	Nuclear pore-associated protein 1	P343A	–	2[h]	2[g]	229	5	1
Q9P2H0/rs7926728	Centrosomal protein of 126 kDa	G238C	–	2[h]	3	147	3	1
Q9UF72/rs35766062	Putative TP73 antisense gene protein 1	P120L	–	2	3[g]	66	N/A	1
Q9UKA4/rs17063163	A-kinase anchor protein 11	H1070R	–	2	2[g]	99	4	1
Q9ULS5/rs17854038	Transmembrane and coiled-coil domain protein 3	P232Q	–	2[h]	3	32	6	1
Q9UPP5/rs7523552	Uncharacterized protein KIAA1107	N715Y	–	3	2[g]	183	N/A	1
Q9Y238/rs9840172	Deleted in lung and esophageal cancer protein 1	N1150D	–	2	2[g]	35	5	1
Q9Y2X0/rs34859566	Mediator of RNA polymerase II transcription subunit 16	L770F	–	2[h]	2[g]	32	5	1
Q9Y334/rs28400001	von Willebrand factor A domain-containing protein 7	R711C	–	2[h]	2[g]	52	5	1
Q9Y448/rs7169404	Small kinetochore-associated protein	R75L	–	2	4	81	4	1
Q9Y6H1/rs864309650	Coiled-coil-helix-coiled-coil-helix domain-containing protein 2	T61I	Parkinson disease 22 (PARK22) [MIM:616710]	2[h]	3	92	7	2

[a] Columns 1 and 2 describe the protein accession numbers in the UniProt database/dbSNP database and protein names, respectively. Column 3 indicates the D➔O mutations, which can be described as *X?Y*, where *X* is the wild-type residue, *Y* is the mutated residue, and *?* is the position. The disease annotations of mutations are shown in column 4. Columns 5 and 6 list the numbers of predictors that agree that the mutations are located in disordered (column 5)/ordered (column 6) regions. The column 7 shows the lengths of predicted LDRs, which are the average length of predicted disordered regions from all the four (VSL2B, IUPred-short, IUPred-long, and DynaMine) predictors. We considered the prediction result to be "ordered" if the score from DynaMine is larger than 0.8 and "disordered" if the score is smaller than 0.69.

[b] Number of predictors that agreed that the wild-type residues are located in disordered region.

[c] Number of predictors that agreed that the mutations are located in ordered region.

[d] Averaged length of disordered regions from different predictors. For Dynamine, <=0.69 was considered disordered, while for VSL2B and IUPred, ≥0.5 was considered disordered.

[e] Number of predictors (out of nine) that agreed that the wild-type residues are located in disordered region according to the prediction results from D2P2 database. "N/A" denotes that the prediction results for current protein could not be found.

[f] Among the predictors agreeing that the wild-type residues are located in disordered region, the number of predictor predicting LDRs (≥30 amino acids). For example, "1" indicates that among the predictors agreeing that the wild-type residues are located in disordered region, only one predictor agreed that it was LDR, suggesting the LDR is potentially biased.

[g] The predicted score of DynaMine for the mutation was between 0.69 and 0.8, which means that the prediction was context dependent. Therefore, this prediction result was not taking into consideration. As such, the total number of predictors for this entry was 3.

[h] The predicted score of DynaMine for the wild-type residue was between 0.69 and 0.8, which means that the prediction was context dependent. Therefore, this prediction result was not taking into consideration. As such, the total number of predictors for this entry was 3.

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

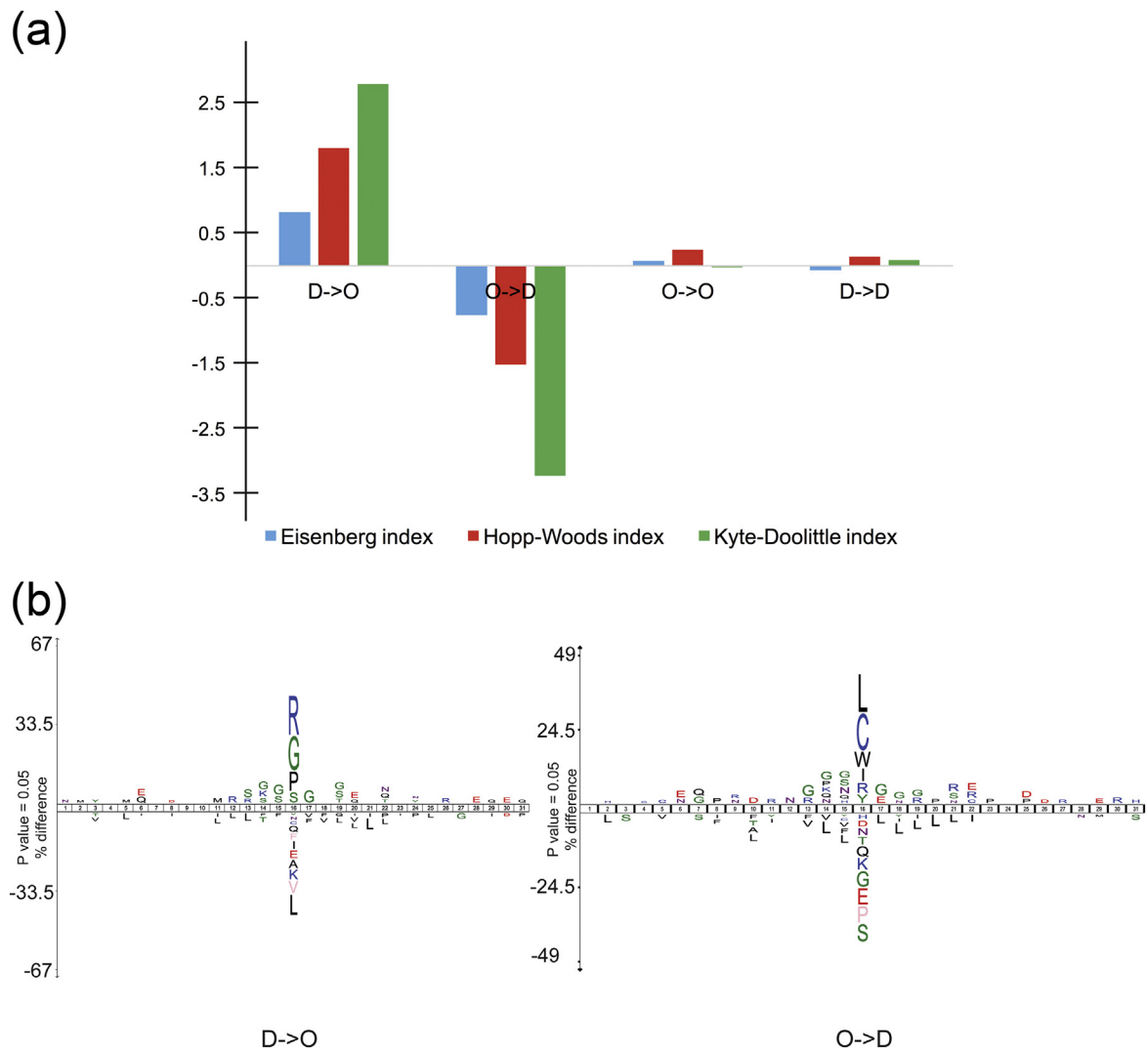

Fig. 1. (a) Mean hydrophobicity changes for disease-causing mutations for four different classes of structure-altering mutations predicted using multiple sequence-based predictors for intrinsically disordered regions (i.e., disorder-to-order is denoted as D➔O, order-to-disorder as O➔D, order-to-order as O➔O, and disorder-to-disorder as D➔D). Bars are shown for the three different hydrophobicity indices used: Eisenberg hydrophobicity index (blue), Hopp–Woods hydrophilicity index (red), and Kyte–Doolittle hydropathicity index (green). (b) iceLogo charts showing the residue conservation around the disease-causing mutation site against a reference set (human Swiss-Prot proteome) for D➔O and O➔D transitions with wild-type residue in the central position. Amino acid residues on top of the *x* axis are significantly conserved, while those underneath it are non-preferred or unfavored according to the reference set.

of cryptic SCEs (Fig. 3). Our findings are in broad agreement with a recent analysis of disease-causing mutations in disordered regions, which demonstrated that a significant number of D➔O mutations are predicted to disrupt protein function [11]. However, here we distinguish mutations that disrupt disorder-based functional properties from those that induce microstructuralization and accompanying gain of function through the phenomenon of structural capacitance. It is possible that some of the D➔O mutations we have identified may induce pathological changes through disrupting known associations with interaction partners, for example, via premature microstructuralization. The lack of evidence for functional interactions and an analysis of ELMs nevertheless suggest that few residues undergoing D➔O mutations form part of an interaction with another protein (Table S2).

In contrast to loss of function through loss of structure in canonical disease-causing mutations [24,25], the complementary phenomenon of gain of structure and function through the introduction of microstructuralization into disordered or unstructured regions of proteins remains undescribed. The characterization of these changes is challenging due to the technical hurdles associated with resolving the

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

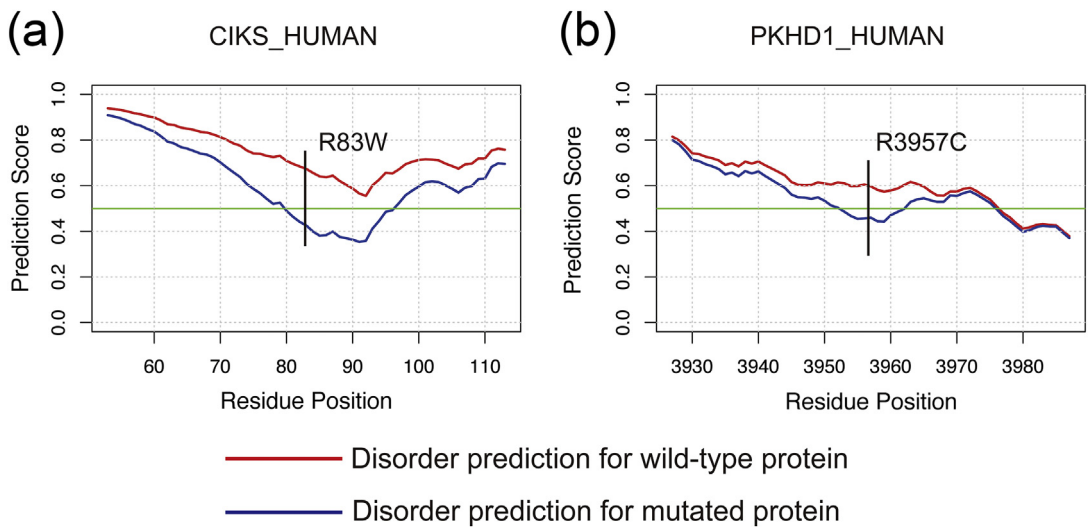

Fig. 2. Predicted disorder score change according to VSL2B for (a) Adapter protein CIKS (UniProt ID: O43734/CIKS_HUMAN; Polymorphism) and (b) Fibrocystin (UniProt ID: P08F94/PKHD1_HUMAN; Polycystic kidney disease; autosomal recessive (ARPKD); [MIM:263200]).

structural properties of a structurally heterogeneous disordered population. Nevertheless, experimental evidence to support the pathogenic relevance of D→O transitions resulting from disease-associated mutations has been described in several proteins [26–28].

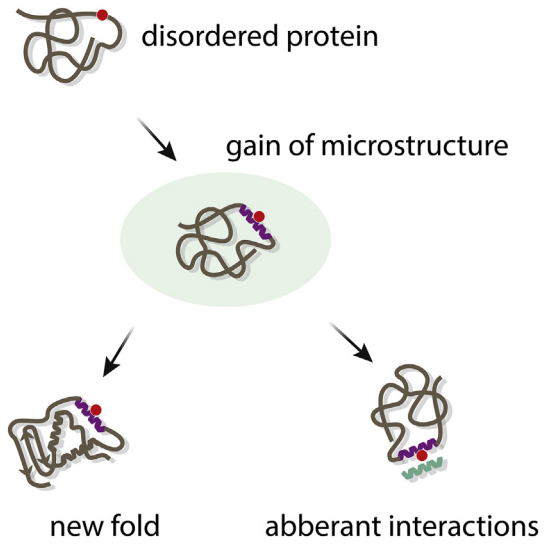

Fig. 3. Disease-causing mutations may result in gain of function through the mechanism of structural capacitance. A D→O mutation (red circle) in a disordered protein results in the generation of local microstructure, or SCE (purple helix). This may be a key nucleating factor in the microevolution of a new adaptive fold, but may also generate inappropriate pathological interactions, through the triggering of inflammatory and autoimmune responses. Aberrant interactions may also promote other pathogenic processes such as aggregate formation, which may result in the formation of pathogenic fibrils.

The phenomenon of structural capacitance has significant implications for protein evolution and for the diversification of organismal form and function over evolutionary time, and may augment other known mechanisms of evolutionary modification such as genetic capacitance [1] and "cryptic" genetic variation [29] that offer the potential for explorations of protein fold and morphological space over compressed timescales. Proteins with high flexibility and dynamics may have a greater intrinsic structural capacitance, increasing their evolvability, and allowing for the more rapid evolution of new folds [30]. The evolution of primordial proteins may have involved either the co-evolution of folds and functions through conformational selection from a repertoire of disordered polypeptides, or the emergence of secondary structure elements followed by the evolution of fully folded proteins [30]. Both scenarios, however, require the prior formation of local structure from an essentially random and disordered population. This raises the issue of how classic Darwinian evolution may proceed in the absence of pre-existing seed structures and functions [31]. We suggest that the microstructuralization necessary for founder events in protein folds may be furnished by structural capacitance (Fig. 3). In this mechanism, SCEs are localized regions of disorder retained within protein structures, and controlled by key capacitance residues that confer the potential to generate new microstructural elements that modify the evolvability of the fold. Structural capacitance generates the potential for micro-structural change, which helps buffer organisms against the vagaries of an uncertain future by furnishing adaptive solutions.

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

In the co-evolutionary model of protein fold and function, microstructuralization of conformationally diverse protein species within a population complements the conformational selection by ligands through the stabilization of functional conformers. The notion that functional selection may occur through the binding of small molecules is compelling, but this process may not inevitably require the prior formation of significant structural scaffolds and could proceed from a relatively small nucleus of structure. Recent work suggests that ligand-binding features arise from the physical and geometric properties of proteins, with structures serving as a feedstock for evolution [32]. This is consistent with findings from directed evolution studies that demonstrate the acquisition of function following only a few prior rounds of selection [33,34]. Structural capacitance may provide the key nucleation event for the formation of a feedstock of molecules with weak functional activity that have the capacity to be fine-tuned and the potential to generate the specificity, high affinity binding and selectivity characteristic of modern enzymes.

In the alternative scenario where the evolution of the fold mimics the folding pathway and the ancestral progenitors of modern folds resemble folding intermediates [30], structural capacitance may introduce reversibility into evolutionary processes in a manner distinct from the ratchet-like and often irreversible mechanism of canonical incremental evolution. The energy landscape of an unstructured or disordered protein may be considered relatively flat with ruggedness depending on hydrophobicity and stereochemistry (Fig. 4a). D➔O mutations with SCEs that induce microstructuralization may induce small impressions, or high-altitude "fissures" into such geographical landscapes (Fig. 4b). Microstructuralization is reversible, allowing rapid conformational transitions and landscape exploration, and minimizes sequestration in dead-end, local minima. The most powerful structural capacitance is predicted to be located in landscapes where D➔O mutations, acting as reversible "binary switches," are expected to have the most significant impact through introducing bias from one lake to another, or transitioning it into a stable "valley." Canonical gradualistic

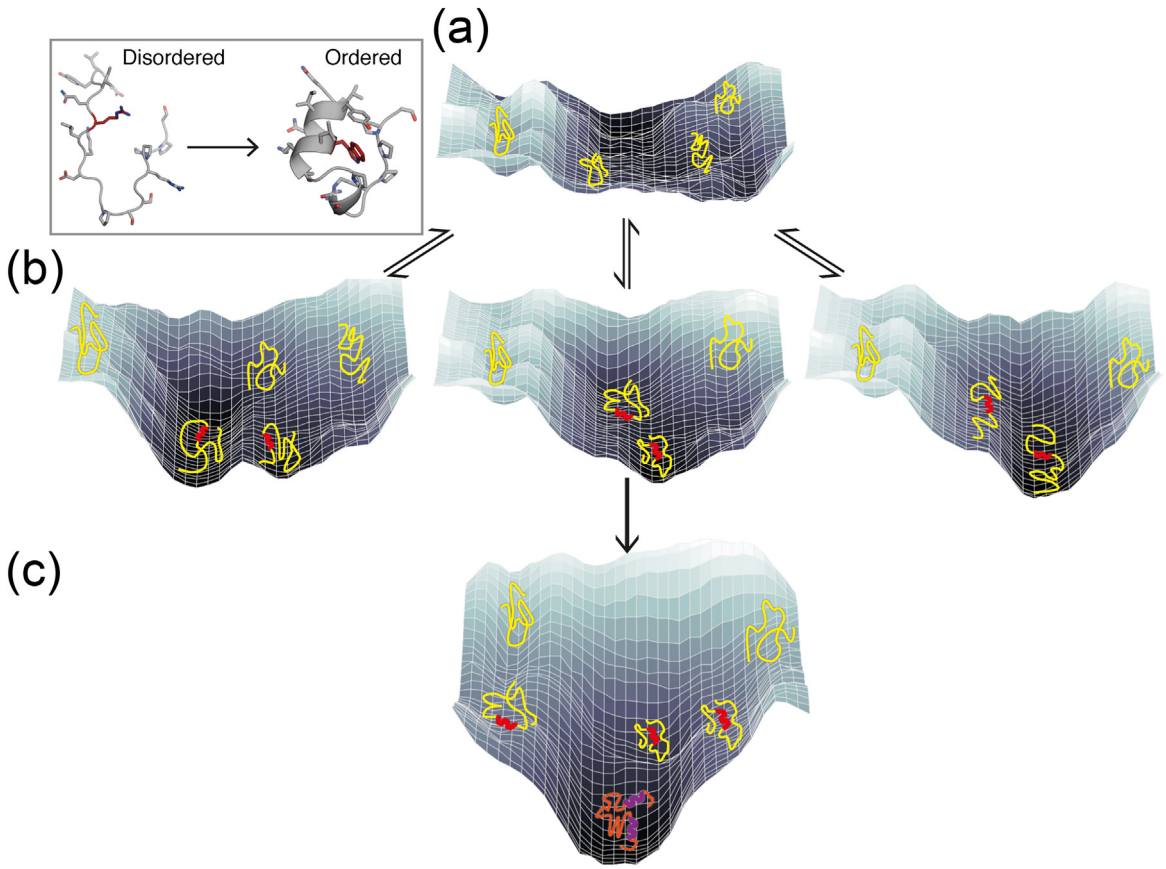

Fig. 4. Structural capacitance and folding energy landscapes. (a) Flat, featureless energy landscape of a disordered protein. (b) D➔O mutations in SCEs induce microstructure and small impressions, allowing conformational transitions and landscape exploration. (c) Canonical incremental evolution optimizes the folding funnel to create new fold. The inset depicts an R➔W mutation (sticks) inducing helix formation and local structuralization, based upon the Trp-cage protein TC5b. Hydrophobic clusters centered around tryptophan are common in several small natural folds.

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

evolution may then optimize the funnel characteristics of the energy landscape (Fig. 4c). These features of structural capacitance would allow for rapid fold generation and may have facilitated some of the major transitions in organismal evolution, complementing and potentiating gradualistic modifications to pre-existing folds. Although there are notable exceptions (serpins, for example, [35,36]), proteins with highly evolved functions are generally situated in deep energy wells at the bottom of the folding landscape, or funnel, preventing major structural changes. Highly optimized active site architectures represent an irreversible evolutionary "ratchet" that may limit the evolutionary adaptability of a fold because non-functional mutants are strongly selected against. Structural capacitance may circumvent such limitations, thereby providing a mechanism for evolvability, and could be exploited for the engineering of artificial proteins with an enhanced capacity for plasticity through micro-evolutionary change [37,38]. Furthermore, such a mechanism might offer some molecular insights into the relationship between cryptic genetic variation [29] and protein evolvability—microstructuralization within SCEs may enable "pre-adapted" phenotypes that confer selective advantages when new selection pressures emerge [39].

Protein folding may nucleate via relatively few key residues, which are typically hydrophobic [40]. Although the energy landscape in which IDPs bind to targets is likely to be complex, experimental evidence supports conformational selection of secondary structure formation, followed by induced fit following binding and completion of global folding [8,41,42]. Our finding that O➔D and D➔O transitions are predominantly associated with mutations to and from proline, respectively (Tables S8A and S8B), is consistent with the demonstrated helix-disrupting properties of proline [43,44]. It is intriguing that 59 (11.5%) of the disease-causing and 80 (6.6%) of non-disease D➔O mutations reported involve a mutation to tryptophan. Hydrophobic clusters centered around tryptophan are common in several small natural folds [45,46] most notably the Trp-cage fold, which is one of the smallest model proteins of just 20-amino-acid residues and folds spontaneously into a stable 3D structure within ~4 μs, featuring a hydrophobic core formed around a central tryptophan residue [47,48]. Along with evidence of residual structure due to hydrophobic collapse around the central tryptophan in the unfolded state of Trp-cage protein TC5b [49], our findings are consistent with tryptophan residues playing an important role in the generation of microstructure from disorder (Fig. 4 inset).

Speculations and hypotheses

We take the opportunity here to discuss further possible implications of our findings that, although speculative and not directly derived from the data, merit mention.

Protein evolution

We propose that the phenomenon of structural capacitance may be extended to a more general

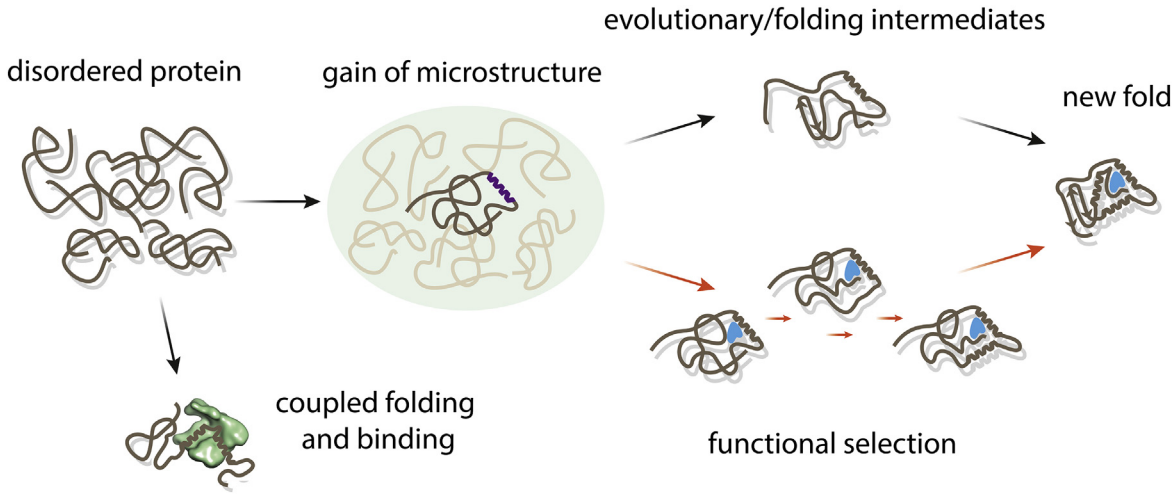

Fig. 5. Structural capacitance incorporating the concepts of classic incremental Darwinian evolution, dynamism, evolvability, and structural diversity provides a potential basis for the generation of microstructure from randomness. The "capacitor" comprises a random ensemble of disordered protein conformations harboring SCEs. These may undergo ordering through either coupled folding-upon-binding of a partner (bottom left) or the fixation of a D➔O mutation that "discharges" the folding information stored within its sequence. A D➔O mutation may bias the random population toward intermittent microstructure, creating a highly evolvable feedstock for subsequent stepwise evolution through random mutation. This process may proceed by selection through either function or folding to produce a new fold.

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

framework encompassing the roles of unstructured proteins (Fig. 5). Approximately 40% of the human proteome is intrinsically disordered [50], but the selective advantage this provides has not been clearly defined. This reservoir of unstructured and highly dynamic protein sequence is highly evolvable for two reasons. First, via the now accepted mechanism of coupled folding and binding (reviewed in Ref. [9]), it may act in concert to engage a broad range of binding partners. This is consistent with the compelling evidence for the role of IDPs in signaling and interaction network hubs [6,24,50]. Second, as described in this work, D→O mutations in SCEs may generate highly evolvable species of conformations with microstructure over rapid timescales that facilitate the evolution of new folds. Both involve a D→O transition, whereby the information for folding is stored in the unstructured protein ensemble. This ensemble is highly evolvable and acts as a "structural capacitor." Release of folding information through microstructuralization is achieved by the binding of a structured physiological partner, as is the case for coupled folding and binding of IDPs, or as the result of a D→O mutation. This concept encapsulates and extends the "dormant foldon" hypothesis [51]. Furthermore, structural capacitance is compatible with the concept of early peptide-world "foldamers" and Dayhoff's hypothesis [52], and more recent hypotheses of early protein evolution driven by oligomerization–duplication–fusion events of short peptides [53,54].

Viral adaptation

The high mutational rate of viruses is a well-characterized phenomenon that allows for adaptation to rapidly changing environments. Recent reports implicate structural disorder in viral adaptation [18] and demonstrate how mutations in disordered regions may promote neostructuralization and accompanying phenotypic divergence [55–57]. Structural capacitance within disordered ensembles of viral proteins may represent a powerful mechanism for enhanced pathogenicity through facilitating the rapid acquisition of microstructure and an accompanying improvement in the ability to interact with host proteins. Molecular recognition elements, found within largely disordered regions, often possess functionally significant residual structure [58] and are key determinants of molecular recognition [59]. Structural capacitance may play a related role in eukaryotic pathogens. Unicellular eukaryotes have considerable variability in their disorder content, which appears to reflect their habitats [18]. The proteomes of parasitic host-changing protozoa, for example, have high levels of disorder, which may represent an adaption to the parasitic lifestyle [60]. Organisms inhabiting environments with high intrinsic frequencies of change, such as microbes, maintain a larger pool of disorder. This is consistent with structural capacitance as constituting a general mechanism for furnishing the capacity to adapt to rapidly changing environments over compressed timescales.

Conclusions

The generation of novel microstructures from conformational "noise" through the mechanism of structural capacitance may have contributed to the functional diversification of the protein repertoire through the origin of new ancestral folds, and in so doing contributed to the origin of life and its subsequent elaboration. Although the reported mutations discussed here are associated with diseases representing a number of different pathogenic types including metabolic, vascular, neoplastic, and congenital, the subset of O→D and D→O mutations appears more likely to have a significant causal role and to be "drivers," than O→O and D→D mutations in which there is no accompanying loss or gain of microstructural change. Given the theoretical nature of this work, we hope that it will prompt experimental validation and further exploration. In summary, the phenomenon of structural capacitance has implications ranging from the ancestral diversification of protein folds to the engineering of synthetic proteins with enhanced evolvability.

Materials and Methods

The overall workflow is shown in Fig. S1.

Data sets

The target dataset was "Human polymorphisms and disease mutations" (http://www.uniprot.org/docs/humsavar) [14]. The release of this dataset, according to the UniProt database, is 2 June 2017. This contains 76,608 human mutations including 29,529 human disease mutations, 39,779 polymorphisms, and 7300 unclassified mutations. Disease mutations were annotated using a basic description of the diseases and their OMIM (http://www.ncbi.nlm.nih.gov/omim/) accession number. Disease mutations are labeled based on literature reports. The UniProt database does not systematically annotate mutations as germline or somatic. For each mutation, this dataset provides detailed information including UniProt (http://www.uniprot.org/) accession number of the original protein, mutated position, wild type and mutated amino acid, and mutation type (i.e., "disease," "polymorphism," and "unclassified"). For the analysis of disease- and non-disease mutations in this study, "unclassified" mutations were removed as the disease annotations for such mutations were ambiguous. Such mutations remain only in Tables S9 for providing

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

a comprehensive and complete D➔O candidate list with experimentally verified disordered regions. After removing the sequences containing uncommon amino acids, the resulting dataset contains 68,383 unique single point mutations (28,662 disease-associated mutations, 39,721 polymorphisms).

Methods

Databases/predictors for disordered region prediction

For both wild-type and mutated proteins, the disorder prediction results were defined using four predictors, namely, VSL2B [61], IUPred (short and long versions) [62,63], and DynaMine [64].

D2P2 database. D2P2 [65] is an online knowledgebase for protein disordered regions prediction results using nine tools for protein disorder prediction: PONDR VLXT [66], PONDR VSL2B [61], IUPred (short and long versions) [62,63], Espritz-D [67], Espritz-X [67], Espritz-N [67], PrDOS [68], and PV2 [69]. In addition, in the updated version of D2P2, MoRF regions (predicted by ANCHOR [70,71]) and post-translational modification sites annotation were used for the investigation of protein binding and function within the disordered regions.

DisProt. DisProt (http://www.disprot.org/index.php [23]; Version: 7 v0.3) harbors experimentally verified intrinsically disorder proteins and disordered regions. DisProt provides detailed function classification, function description and experimental evidence for each entry in this database. The advantage of this database is that the disordered regions harbored in DisProt have been experimentally verified. We used the DisProt database to locate mutations that are located in experimentally validated regions of disorder. We then applied four predictors (VSL2B, IUPred-L, IUPred-S, and DynaMine) to predict disorder–order transitions.

IUPred. IUPred (http://iupred.enzim.hu/ [62,63]) maintains two versions of IUPred including IUPred-S and IUPred-L. Here, "S" and "L" refer to the long LDRs and SDRs, respectively. For the "S" option, the model was trained using a dataset corresponding to missing residues in the protein structures. These residues are absent from the protein structures due to missing electron density in the corresponding X-ray crystal structures. These disordered regions are usually short. Conversely, for the "L" option, the dataset used to train models corresponds to LDRs that are validated by various experimental techniques. In our study, residues with predicted scores equal to or above 0.5 were considered to be located in disordered regions.

PONDR-VSL2B. VSL2B [61] is a widely used sequence-based predictor for intrinsically disordered regions, using Support Vector Machine (SVM). Residues with predicted scores equal to or above 0.5 are considered to be disordered.

DynaMine. DynaMine [64], which is trained with a curated NMR dataset, was used to predict protein disordered regions with only sequence information as the input. Residues with predicted scores below 0.69 are considered to be located in disordered regions, while those with scores above 0.8 are predicted to be in the ordered regions.

Amino acid hydrophobicity indices and sequence motif conservation

Three indices were chosen in our study: Hopp–Woods hydrophilicity index [72], Kyte–Doolittle hydropathicity index [73], and Eisenberg hydrophobicity index [74]. The motif conservation is shown using iceLogo [75] in Fig. 1.

Predictor for protein transmembrane domains

TMHMM [76] employs hidden Markov model for membrane protein topology prediction. Given the fact that the protein transmembrane domains are structurally stable and ordered, TMHMM was used to further validate the predicted disordered regions. Mutations predicted to be in transmembrane regions were discarded.

Protein structure BLAST

In order to ensure that wild-type proteins with predicted disordered regions that lack experimentally determined structures or homologue structures, we performed a BLAST search against the PDB database (http://www.rcsb.org/pdb/software/rest.do) [77] using the protein sequences (*e*-value cutoff = 0.01). Any proteins with predicted disordered regions and BLAST hits against the PDB database were removed.

ELM database mapping

We mapped both disease and non-disease mutations in D➔O, O➔D, D➔D, and O➔O transitions using the ELM (Version: 1.4) (Eukaryotic Linear Motif; http://elm.eu.org/search/ [20]) database. Tables S2–S5 show the mapping results of our mutations of both disease and non-disease for the four structural transitions. All ELMs listed in Tables S2-S5 are experimentally verified (i.e., annotated with experimental evidence showing this ELM to be functional.)

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

Pfam database mapping

In order to characterize the domain context of mutations, we mapped both disease and non-disease mutations to the Pfam (Release: 30) database [22]. The mapping results for disease and non-disease mutations of D➔O and O➔D transitions are shown in Tables S6 and S7.

Acknowledgments

C.L. was supported by China Scholarship Council and Monash University Joint PhD Student Scholarship (2011630031) and by the Bridging Postdoctoral Fellowship of Faculty of Medicine, Nursing and Health Sciences, Monash University (BPF17-0021), and is currently a National Health and Medical Research Council early career fellow (1143366). L.V.T.C. was supported by an Commonwealth Serum Laboratories/ Undergraduate Research Opportunities Program stipend. G.I.W. is an Australian Research Council (ARC) Discovery Outstanding Researcher Award Fellow (DP140100087). N.A.B. is funded by an ARC Future Fellowship (110100223). A.M.B. acknowledges support from the National Health and Medical Research Council (1022688). We thank Dr. Morihiro Hayashida and Prof. Tatsuya Akustu from Kyoto University for assistance with performing the BLAST analysis. We thank Steve Androulakis and the Monash eResearch Centre for assistance with computational tasks, and Mikael Oliveberg, Renwick Dobson, Colin Jackson, and Daniel Christ for helpful comments.

Author Contributions: C.L., L.V.T.C., and R.Z. performed data analysis. B.T.P., J.M., I.K., N.A.B., J.S., and M.B. contributed to writing and conceptual advances. G.I.W. assisted in data analysis. A.M.B., C.L., and A.W. designed the study and drafted the paper.

Competing Financial Interests: The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2018.06.051.

Received 23 February 2018;
Received in revised form 18 June 2018;
Accepted 29 June 2018
Available online xxxx

Keywords:
structural capacitance;
protein evolution;
human diseases;
protein disordered region;

Abbreviations used:
IDPs, intrinsically disordered proteins; LDRs, long disordered regions; ELM, eukaryotic linear motif; SCE, structural capacitance element.

References

[1] D.F. Jarosz, M. Taipale, S. Lindquist, Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms, Annu. Rev. Genet. 44 (2010) 189–216.
[2] T. Arodz, P.M. Plonka, Effects of point mutations on protein structure are nonexponentially distributed, Proteins 80 (2012) 1780–1790.
[3] Y. He, Y. Chen, P.A. Alexander, P.N. Bryan, J. Orban, Mutational tipping points for switching protein folds and functions, Structure 20 (2012) 283–291.
[4] A. Toth-Petroczy, D.S. Tawfik, Hopeful (protein InDel) monsters? Structure 22 (2014) 803–804.
[5] R.B. Berlow, H.J. Dyson, P.E. Wright, Functional advantages of dynamic protein disorder, FEBS Lett. 589 (2015) 2433–2440.
[6] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.
[7] C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins and intrinsically disordered protein regions, Annu. Rev. Biochem. 83 (2014) 553–584.
[8] J. Dogan, S. Gianni, P. Jemth, The binding mechanisms of intrinsically disordered proteins, Phys. Chem. Chem. Phys. 16 (2014) 6323–6331.
[9] T. Kiefhaber, A. Bachmann, K.S. Jensen, Dynamics and mechanisms of coupled protein folding and binding reactions, Curr. Opin. Struct. Biol. 22 (2012) 21–29.
[10] H. Lee, K.H. Mok, R. Muhandiram, K.H. Park, J.E. Suk, D.H. Kim, et al., Local structural elements in the mostly unstructured transcriptional activation domain of human p53, J. Biol. Chem. 275 (2000) 29426–29432.
[11] V. Vacic, P.R. Markwick, C.J. Oldfield, X. Zhao, C. Haynes, V.N. Uversky, et al., Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder, PLoS Comput. Biol. 8 (2012), e1002709.
[12] Y. Chebaro, A.J. Ballard, D. Chakraborty, D.J. Wales, Intrinsically disordered energy landscapes, Sci. Rep. 5 (2015), 10386.
[13] D. Granata, F. Baftizadeh, J. Habchi, C. Galvagnion, A. De Simone, C. Camilloni, et al., The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments, Sci. Rep. 5 (2015), 15449.
[14] Y.L. Yip, M. Famiglietti, A. Gos, P.D. Duek, F.P. David, A. Gateau, et al., Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase, Hum. Mutat. 29 (2008) 361–366.
[15] B. He, K. Wang, Y. Liu, B. Xue, V.N. Uversky, A.K. Dunker, Predicting intrinsic disorder in proteins: an overview, Cell Res. 19 (2009) 929–949.
[16] A. Lobley, M.B. Swindells, C.A. Orengo, D.T. Jones, Inferring function using patterns of native disorder in proteins, PLoS Comput. Biol. 3 (2007) e162.
[17] P. Tompa, Intrinsically disordered proteins: a 10-year recap, Trends Biochem. Sci. 37 (2012) 509–516.
[18] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes

from viruses and the three domains of life, J. Biomol. Struct. Dyn. 30 (2012) 137–149.
[19] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (2004) 635–645.
[20] H. Dinkel, K. Van Roey, S. Michael, N.E. Davey, R.J. Weatheritt, D. Born, et al., The eukaryotic linear motif resource ELM: 10 years and counting, Nucleic Acids Res. 42 (2014) D259–D266.
[21] M. Fuxreiter, P. Tompa, I. Simon, Local structural disorder imparts plasticity on linear motifs, Bioinformatics 23 (2007) 950–956.
[22] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, et al., The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res. 44 (2016) D279–D285.
[23] M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, et al., DisProt: the Database of Disordered Proteins, Nucleic Acids Res. 35 (2007) D786–D793.
[24] D.P. Lane, Cancer. p53, guardian of the genome, Nature 358 (1992) 15–16.
[25] A.C. Joerger, A.R. Fersht, Structure–function–rescue: the diverse nature of common p53 cancer mutants, Oncogene 26 (2007) 2226–2242.
[26] V. Lemma, M. D'Agostino, M.G. Caporaso, M. Mallardo, G. Oliviero, M. Stornaiuolo, et al., A disorder-to-order structural transition in the COOH-tail of Fz4 determines misfolding of the L501fsX533-Fz4 mutant, Sci. Rep. 3 (2013) 2659.
[27] H. Dembinski, K. Wismer, D. Balasubramaniam, H.A. Gonzalez, V. Alverdi, L.M. Iakoucheva, et al., Predicted disorder-to-order transition mutations in IkappaBalpha disrupt function, Phys. Chem. Chem. Phys. 16 (2014) 6480–6485.
[28] J. Mittal, T.H. Yoo, G. Georgiou, T.M. Truskett, Structural ensemble of an intrinsically disordered polypeptide, J. Phys. Chem. B 117 (2013) 118–124.
[29] G. Gibson, I. Dworkin, Uncovering cryptic genetic variation, Nat. Rev. Genet. 5 (2004) 681–690.
[30] N. Tokuriki, D.S. Tawfik, Protein dynamism and evolvability, Science 324 (2009) 203–207.
[31] L.C. James, D.S. Tawfik, Conformational diversity and protein evolution—a 60-year-old hypothesis revisited, Trends Biochem. Sci. 28 (2003) 361–368.
[32] J. Skolnick, M. Gao, Interplay of physics and evolution in the likely origin of protein biochemical function, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 9344–9349.
[33] C. Jurgens, A. Strom, D. Wegener, S. Hettwer, M. Wilmanns, R. Sterner, Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 9925–9930.
[34] G. Song, G.A. Lazar, T. Kortemme, M. Shimaoka, J.R. Desjarlais, D. Baker, et al., Rational design of intercellular adhesion molecule-1 (ICAM-1) variants for antagonizing integrin lymphocyte function-associated antigen-1-dependent adhesion, J. Biol. Chem. 281 (2006) 5042–5049.
[35] J.A. Huntington, R.J. Read, R.W. Carrell, Structure of a serpin–protease complex shows inhibition by deformation, Nature 407 (2000) 923–926.
[36] B.T. Porebski, S. Keleher, J.J. Hollins, A.A. Nickson, E.M. Marijanovic, N.A. Borg, et al., Smoothing a rugged protein folding landscape by sequence-based redesign, Sci. Rep. 6 (2016), 33958.
[37] E.C. Campbell, G.J. Correy, P.D. Mabbitt, A.M. Buckle, N. Tokuriki, C.J. Jackson, Laboratory evolution of protein conformational dynamics, Curr. Opin. Struct. Biol. 50 (2017) 49–57.
[38] E. Campbell, M. Kaltenbach, G.J. Correy, P.D. Carr, B.T. Porebski, E.K. Livingstone, et al., The role of protein dynamics in the evolution of new enzyme function, Nat. Chem. Biol. 12 (2016) 944–950.
[39] E.J. Hayden, E. Ferrada, A. Wagner, Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme, Nature 474 (2011) 92–95.
[40] M. Vendruscolo, E. Paci, C.M. Dobson, M. Karplus, Three key residues form a critical contact network in a protein folding transition state, Nature 409 (2001) 641–645.
[41] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, Preformed structural elements feature in partner recognition by intrinsically unstructured proteins, J. Mol. Biol. 338 (2004) 1015–1026.
[42] C.J. Oldfield, Y. Cheng, M.S. Cortese, P. Romero, V.N. Uversky, A.K. Dunker, Coupled folding and binding with alpha-helix-forming molecular recognition elements, Biochemistry (Mosc) 44 (2005) 12454–12470.
[43] A. Horovitz, J.M. Matthews, A.R. Fersht, Alpha-helix stability in proteins. II. Factors that influence stability at an internal position, J. Mol. Biol. 227 (1992) 560–568.
[44] K.T. O'Neil, DeGrado WF, A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids, Science 250 (1990) 646–651.
[45] X.A. Li, M.J. Sutcliffe, T.W. Schwartz, C.M. Dobson, Sequence-specific 1H NMR assignments and solution structure of bovine pancreatic polypeptide, Biochemistry (Mosc) 31 (1992) 1245–1253.
[46] M. Jager, H. Nguyen, J.C. Crane, J.W. Kelly, M. Gruebele, The folding mechanism of a beta-sheet: the WW domain, J. Mol. Biol. 311 (2001) 373–393.
[47] J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, Designing a 20-residue protein, Nat. Struct. Biol. 9 (2002) 425–430.
[48] L. Qiu, S.A. Pabit, A.E. Roitberg, S.J. Hagen, Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros, J. Am. Chem. Soc. 124 (2002) 12952–12953.
[49] K.H. Mok, L.T. Kuhn, M. Goez, I.J. Day, J.C. Lin, N.H. Andersen, et al., A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein, Nature 447 (2007) 106–109.
[50] V.N. Uversky, A.K. Dunker, Understanding protein non-folding, Biochim. Biophys. Acta 1804 (2010) 1231–1264.
[51] V.N. Uversky, Unusual biophysics of intrinsically disordered proteins, Biochim. Biophys. Acta 1834 (2013) 932–951.
[52] M.L. Romero Romero, A. Rabin, D.S. Tawfik, Functional proteins from short peptides: Dayhoff's hypothesis turns 50, Angew. Chem. Int. Ed. Engl. 55 (2016) 15966–15971.
[53] J. Soding, A.N. Lupas, More than the sum of their parts: on the evolution of proteins from peptides, Bioessays 25 (2003) 837–846.
[54] S. Chessari, R. Thomas, F. Polticelli, P.L. Luisi, The production of de novo folded proteins by a stepwise chain elongation: a model for prebiotic chemical evolution of macromolecular sequences, Chem. Biodivers. 3 (2006) 1202–1210.
[55] J. Siltberg-Liberles, Evolution of structurally disordered proteins promotes neostructuralization, Mol. Biol. Evol. 28 (2011) 59–62.
[56] J.F. Ortiz, M.L. MacDonald, P. Masterson, V.N. Uversky, J. Siltberg-Liberles, Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses, Genome Biol. Evol. 5 (2013) 504–513.

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051

[57] S.S. Borkosky, G. Camporeale, L.B. Chemes, M. Risso, M.G. Noval, I.E. Sanchez, et al., Hidden structural codes in protein intrinsic disorder, Biochemistry (Mosc) 56 (2017) 5560–5569.
[58] A. Mohan, C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, et al., Analysis of molecular recognition features (MoRFs), J. Mol. Biol. 362 (2006) 1043–1059.
[59] P.M. Mishra, V.N. Uversky, R. Giri, Molecular recognition features in Zika virus proteome, J. Mol. Biol. 430 (2018) 2372–2388.
[60] A. Mohan, W.J. Sullivan Jr., P. Radivojac, A.K. Dunker, V.N. Uversky, Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes, Mol. BioSyst. 4 (2008) 328–340.
[61] K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker, Z. Obradovic, Length-dependent prediction of protein intrinsic disorder, BMC Bioinformatics 7 (2006) 208.
[62] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics 21 (2005) 3433–3434.
[63] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, J. Mol. Biol. 347 (2005) 827–839.
[64] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts, W.F. Vranken, From protein sequence to dynamics and disorder with DynaMine, Nat. Commun. 4 (2013) 2741.
[65] M.E. Oates, P. Romero, T. Ishida, M. Ghalwash, M.J. Mizianty, B. Xue, et al., (DP2)-P-2: database of disordered protein predictions, Nucleic Acids Res. 41 (2013) D508-D16.
[66] P. Romero, Z. Obradovic, X.H. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins 42 (2001) 38–48.
[67] I. Walsh, A.J.M. Martin, T. Di Domenico, S.C.E. Tosatto, ESpritz: accurate and fast prediction of protein disorder, Bioinformatics 28 (2012) 503–509.
[68] T. Ishida, K. Kinoshita, PrDOS: prediction of disordered protein regions from amino acid sequence, Nucleic Acids Res. 35 (2007) W460–W464.
[69] M.F. Ghalwash, A.K. Dunker, Z. Obradovic, Uncertainty analysis in protein disorder prediction, Mol. BioSyst. 8 (2012) 381–391.
[70] B. Meszaros, I. Simon, Z. Dosztanyi, Prediction of protein binding regions in disordered proteins, PLoS Comput. Biol. 5 (2009).
[71] Z. Dosztanyi, B. Meszaros, I. Simon, ANCHOR: web server for predicting protein binding regions in disordered proteins, Bioinformatics 25 (2009) 2745–2746.
[72] T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, Proc. Natl. Acad. Sci. U. S. A. 78 (1981) 3824–3828.
[73] J. Kyte, DoolittleR.F. , A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[74] D. Eisenberg, E. Schwarz, M. Komaromy, R. Wall, Analysis of membrane and surface protein sequences with the hydrophobic moment plot, J. Mol. Biol. 179 (1984) 125–142.
[75] N. Colaert, K. Helsens, L. Martens, J. Vandekerckhove, K. Gevaert, Improved visualization of protein consensus sequences by iceLogo, Nat. Methods. 9 (2009) 786–787.
[76] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[77] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

Please cite this article as: C. Li, et al., Structural Capacitance in Protein Evolution and Human Diseases, J. Mol. Biol. (2018), https://doi.org/10.1016/j.jmb.2018.06.051