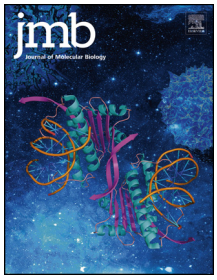

Structural Principles Governing Disease-Causing Germline Mutations

László Dobson[1], Bálint Mészáros[2] and Gábor E. Tusnády[1]

1 - "Momentum" Membrane Protein Bioinformatics Research Group, Institute of Enzymology, RCNS, HAS, PO-Box 7, Budapest H-1518, Hungary
2 - "Momentum" MTA-ELTE Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/c, Budapest H-1117, Hungary

Correspondence to Gábor E. Tusnády: *tusnady.gabor@ttk.mta.hu*
https://doi.org/10.1016/j.jmb.2018.10.005
Edited by Anna Panchenko

Abstract

Advancements in sequencing in the past decades enabled not only the determination of the human proteome but also the identification of a large number of genetic variations in the human population. The phenotypic effects of these mutations range from neutral for polymorphisms to severe for some somatic mutations. Disease-causing germline mutations (DCMs) represent a special and largely understudied class with relatively weak phenotypes. While for somatic mutations their effect on protein structure and regulation has been extensively studied in select cases, for germline mutations, this information is currently largely missing. In this analysis, a large amount of DCMs were analyzed and contrasted to polymorphisms from a structural point of view. Our results delineate the characteristic features of DCMs starting at the global level of partitioning proteins into globular, disordered and transmembrane classes, moving toward smaller structural units describing secondary structure elements and molecular surfaces, reaching down to the smallest structural entity, post-translational modifications. We show how these structural entities influence the emergence and possible phenotypic effects of DCMs.

Introduction

Mutations can impact proteins on several levels, perturbing structural stability, affecting folding, modulating degradation, and can lead to improper trafficking, or the emergence of toxic conformations [1]. The theoretical spectrum of genetic variations ranges from neutral to lethal through increasingly strong phenotypes. Neutral or near-neutral polymorphisms (PMs) are present in a large portion of the human population without any obvious loss of fitness. However, the majority of non-lethal mutations is expected to cause a noticeable phenotypic effect and thus is under significant negative selection. Many of these are disease-associated or disease-causing mutations (DCMs) that form the genetic background of a wide range of pathological conditions. Personalized medicine is a rapidly growing area based on the idea that individual's genetic content can be used to offer a more precise treatment. The early recognition of harmful changes not only helps a more personalized medication but also increases the chance it is delivered prior to irreversible changes.

The identification of a germline mutation cannot usually be done from a single sequencing step without a reference independent of the patient, and the study of familial background is often a must. As a result, early studies considering the phenotypic effects of germline DCMs were restricted to a few proteins only [2,3], and systematic analyses were not possible due to the lack of known variations and structures. In recent years, the number of solved protein structures has grown exponentially, and the amount of available germline DCM data is also constantly increasing, making systematic analyses of DCMs and PMs in proper structural background available The growing number of all-atom structures also helps the development of different prediction methods, helping analyses to be extended to proteins without a solved structure.

One way to describe protein structure is through geometric properties. There are several papers emphasize the importance of surfaces, such as protein–protein interaction (PPI) and buried parts, both large-scale [4–7] and individual studies [8]. Analysis of physico-chemical features adds an additional layer to this information: DCMs often destabilize the structure by disrupting hydrophobic, polar interactions and disulfide and salt bridges [9]. On the other hand, over-stabilization can be also disease causing if the required level of flexibility to maintain a function is damaged [10]. Some of these studies also consider binding affinity [11] or analyzed their effect on enzymatic function [12]. These publications mutually agree that disruption of compact structural entities of rewiring PPI networks carry a lot of disease. Another approach is the consideration of the (lack of) periodic structure elements and domains: several papers investigate the effect of protein disorder. Due to the problematic experimental capture of flexibility, these investigations are usually based on prediction methods, and they conclude that DCMs are depleted in disordered regions and suggest that such variations often stabilize the structure [13–15]. Around 25% of the human proteome is transmembrane protein (TMP): our previous study suggested that non-polar to non-polar amino acid changes and non-polar to charged ones are equally frequent in TMPs carrying DCMs and that mutations to positively charged amino acids in the central of the lipid bilayer often cause disease [16]. Finally, post-translational modifications (PTMs) are the smallest structural/functional units of proteins and DCMs accumulate around these sites compared to expected [5,17].

Although these papers offer prospect to DCMs, to better understand how single-nucleotide polymorphisms perturb function, a more complex point of view is necessary, considering several structural features at the same time rather than individually. DCMs in TMPs, disordered proteins and globular proteins were studied in details individually; however, the presence of DCMs compared between these structural groups was not considered so far. How DCMs are spread in these classes? Do mutational preferences for individual diseases provide a structure-based classification? Protein interactions can be classified based on the interplay between folding and binding [18]. How does such grouping influence our current knowledge regarding the occurrences of DCMs on different surfaces? Predicting topology of TMPs is one of the most reliable parts of structure-based predictions, yet topology-based analysis of TMPs and DCMs is completely missing. Is there any connection between the occurrence of DCMs and biophysical laws aiding the folding of TMPs? How PTMs behave on different surfaces? Phosphomimetics is a common event in cancer. Do germline DCMs also exploit this phenomenon? It is common knowledge that DCMs are rare in disordered regions compared to ordered domain, yet do we see exceptions by utilizing a more complex point of view?

Since the known structural space is constantly growing and new variations are also often deposited into dedicated databases, it is also useful to revisit the topic time to time. Somatic and germline mutations affect protein function and cell survival in different ways. Is this reflected on the structure level of proteins? In the present study, we try to answer these questions using statistical methods as well as by investigation of specific examples.

Results

DCMs preferentially target structured protein regions

To gain a general insight into the connection between various types of mutations and structural regions of proteins, all residues in the human proteome were classified into three distinct structural groups: transmembrane (TM), ordered, and disordered; the occurrences of PMs and DCMs in each class were compared to that expected from random. As expected from their neutral nature, the occurrences of PMs (even if they deviate from the random background) are close to expected values in all structural parts (Fig. 1A, blue triangle). In contrast, DCMs are significantly accumulated on TM and ordered protein regions. This observation is also confirmed by considering domains defined in Pfam, where DCMs are significantly enriched (Supplementary Fig. 1). In contrast, DCMs are less frequent in disordered protein regions, in line with the higher mutation tolerance of protein regions without stable structures [19] (Fig. 1A, red triangle).

While the definitions of ordered and TM protein regions are straightforward (the presence of a structure in aqueous solution or in a lipid bilayer was confirmed for at least in a homolog in all studied cases), the above definition of disordered sequence parts is largely based on prediction. While this enables the analysis on the whole human proteome, the results in theory could depend on the choice of algorithm. In order to eliminate this potential bias [20], we further analyzed disordered residues in three independent ways, employing three different definitions for disordered residues. The first two definitions rely on experimental validation, considering residues with missing coordinates in X-ray PDB structures, or using data from the DisProt database. These two approaches cover two different “flavors” of disorder [21], with the former describing short disordered linkers, loops and terminal regions, while the latter better represents long disordered regions. In a third, more permissive approach, all protein regions falling out of well-characterized structural parts were considered. These regions are also expected to contain a high number of intrinsically

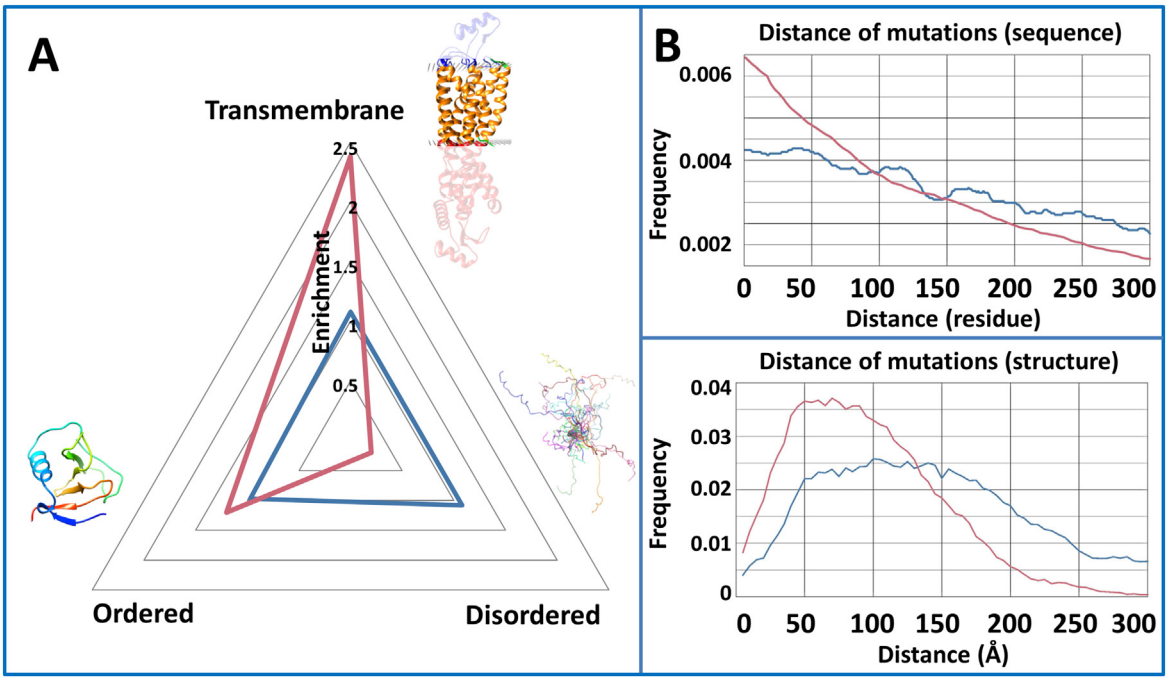

Fig. 1. The space of human protein sequences. A, Enrichments of PMs and DCMs in TM, ordered and disordered residues. *p*-Values are <0.01 in all cases, where the expected number of PMs and DCMs in a certain structural parts was tested (χ^2 test, see Supplementary Material for exact values). B, Distance distributions between PMs and DCMs (top, sequence analysis; bottom, structure analysis). *p*-Values are <0.01 in all cases (two-sample Kolmogorov–Smirnov test; blue, PMs; red, DCMs).

disordered regions/proteins (IDR/IDP), albeit also containing other, possibly structured segments. Using any of the above alternative definitions, the depletion of DCMs in disordered residues is significant (Supplementary Fig. 2).

Proteins are generally modular, and most human proteins contain several structural/functional units (e.g., domains, linkers, TM regions, etc.). The presence of distinct modules is expected to perturb the distribution of the location of mutations with functional importance. In order to test this assumption, we investigated the average distance of various PMs along the protein sequence and contrasted this distribution with that calculated for DCMs. The residual distances of PMs seem to be evenly distributed, in contrast to DCMs, where shorter distances between mutations are more common. (Fig. 1B, top) This suggests that DCMs prefer to cluster close to each other, potentially targeting selected protein modules. This observation is even more pronounced when the spatial distribution of mutations calculated from PDB structures is taken into consideration, showing a peak of DCM distances at 12 Å (Fig. 1B, bottom). This observation hints that certain parts of the protein are more vulnerable to alterations and DCMs tend to specifically target them. In the next sections, we investigate the resilience of specific structural/functional regions to various types of mutations, and how DCMs and PMs perturb various structural units.

Structural features of mutations offer a natural classification of diseases

Most analyzed hereditary diseases are monogenic with all annotated mutations affecting the same protein (2460 out of 2540 diseases—obtained from humsavar, filtered by MIM identifier). However, as the affected proteins can be highly modular including ordered, disordered and TM regions, mutations from a single disease can—in theory—affect more than one type of structural region. While illnesses can be classified according to a wide range of criteria (chromosomal location, severity, inheritance type, etc.), the potential interplay between various types of protein structural regions in the emergence of the phenotype provides a novel, molecular basis for grouping various diseases.

In order to test the validity of a molecular structure-based classification approach, each disease in our data set was represented by a three-element vector that encodes the fraction of associated mutations in ordered, disordered and TM regions. These vectors were used as input to a hierarchical clustering algorithm. The resulting tree depicting linkage of various diseases is shown in Fig. 2A. Apart from hierarchical clustering, k-means clustering was also used, to reveal the optimal number of clusters to partition the data. Inspecting the within-cluster sum of squares as a function of the number of clusters, the most pronounced elbow corresponds to partitioning

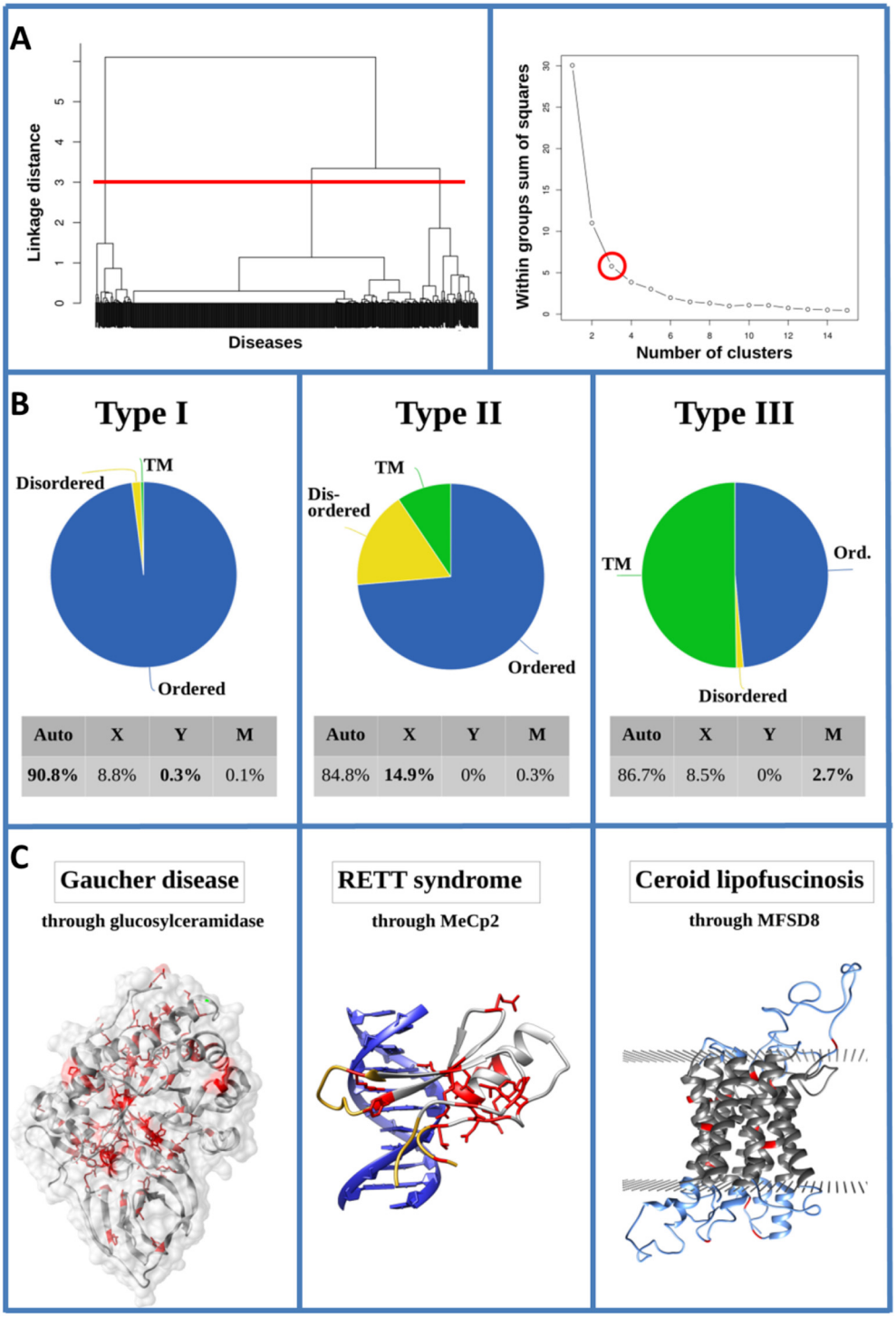

Fig. 2. The structural definition of the three types of diseases. A, the tree resulting from hierarchical clustering of diseases based on structural content (left) and the k-means evaluation of various numbers of clusters. B, The average fraction of mutations falling into the three types of structural regions for diseases in the three classes (top); the fraction of mutations linked to autosomal, X, Y chromosomes, respectively, or the mitochondrial DNA (bottom). Bold values mark higher ratio of diseases (see text for details). C, Examples for the three types of diseases. Positions with known disease-associated germline mutations are marked in red. For MeCp2, regions in gold mark flexible/unstructured regions based on NMR structure (PDB ID: 1qk9). MFSD8 was modeled by Swissmodel [22], and the membrane plane was defined by TMDET [23].

the data into three clusters. In accord, we considered three types of diseases; however, as the k-means analysis shows, further potentially relevant subclasses can be identified by increasing the numbers of final clusters (Fig. 2A).

The three structural composition-based types of diseases show marked differences (Fig. 2B). Type I contains diseases that arise predominantly via ordered mutations. This group is dominated by autosomal mutations but also contains all known hereditary diseases that are linked to Y-chromosomal inheritance. A prototype of this class is Gaucher disease that is caused by mutations in the glucosylceramidase enzyme disrupting lysosomal storage. Glucosylceramidase is a single domain ordered enzyme with a monomeric active form. This domain can be destabilized via a large number of point mutations, most of which (105 out of 113) are buried in the hydrophobic core (Fig. 2C). This destabilization reduces enzyme activity, and the occurring mutations determine the severity and onset age of the disease. This destabilizing effect can be countered by the use of small molecules that can increase domain stability [24].

The second group of diseases exhibit markedly different structural composition. While in this class the majority of mutations also fall into ordered protein regions, there is a significant portion of variations that affect disordered regions as well. This shows that the coordination between order and disorder in disease emergence is a common theme for a wide range of illnesses. Interestingly, this class shows a higher ratio of X-chromosome-linked disease showing a non-trivial connection between preferred protein structural regions and inheritance type. A typical example for these type II illnesses is the X-linked RETT syndrome. This disease is linked to the mutations of the MeCp2 protein, which is a generic transcriptional repressor binding to methylated DNA. The majority of observed mutations (28 out of 46) can be mapped onto the DNA binding domain and disrupt or weaken the DNA binding leading to epigenetic deregulation. While the DNA binding domain has a stable structure in isolation, it contains three discernable disordered/highly flexible regions encompassing the two termini and a middle region between residues 110–120 [25]. The known mutations can be mapped to both the highly structured and the flexible regions, indicating the coordination between the two structural units. Furthermore, the rest of the protein is predicted to be largely unstructured and the rest of the observed germline mutations affect these regions emphasizing the importance of interplay between protein order and disorder.

The third class of diseases marks the realm of TMPs. While ordered mutations still represent a large factor, on average, over half of the mutations linked to diseases of third class affect TM regions. In accord, this class encompasses the vast majority of mitochondrial diseases. A prime example of such disease is ceroid lipofuscinosis that is linked to mutations in the Mfsd8 protein, a lysosomal TMP that is capable of transporting small solutes using chemical gradients. Mfsd8 contains 12 TM regions, half of which carry disease-linked mutations. In addition, three intracellular and one extracellular short helix-connecting loops are also mutated. Half of these mutations cause a change in charge, which can easily modulate the protein's interaction with the surrounding ionic environment. However, the other half of mutations typically affects hydrophobic residues, indicating that the interaction between various TM regions can also be modulated.

Diseases belonging to the three classes of structurally defined groups exhibit different characteristics at the molecular level. While the above examples provide insights into these mechanisms, it is an open question how structure and function are modulated in general via the three classes of structural elements. In the following chapters, we aim to provide an assessment for possible points of modulation in TM, disordered and ordered proteins.

DCMs and the positive-inside rule of TMPs

As TM regions show a strikingly high enrichment in DCMs compared to neutral mutations, the exact location of DCMs in these proteins was analyzed in greater detail. TMPs were partitioned into distinct structural segments: TM and terminal regions, membrane proximal segments, and connecting loops. While the distribution of PMs on all regions was very close to random (see Fig. 3), DCM distributions show targeted enrichment in several region classes. The distribution of DCMs in TM regions of bitopic and polytopic proteins was separately investigated (Fig. 3). Single spanning TM regions are depleted in DCMs, in contrast to TM regions of polytopic proteins. As helix formation is driven by hydrogen bonding between mainchain atoms, single α-helical structures are difficult to disrupt by a change in side chains caused by DCMs. However, in the case of polytopic TM segments, the helix–helix interactions serve as much more accessible targets for structural disruption, as these interactions are typically mediated through sidechain–sidechain interactions. Apart from polytopic TM regions, a distinct enrichment of DCMs can be observed in cytosolic connecting loops (i.e., a five-residue extension from membrane boundary on the cytosolic side).

We further analyzed these cytosolic regions and looked for loops containing positively charged residues that are known to govern the orientation of TMPs in the membrane according to the positive-inside rule [26]. DCMs in these regions are enriched compared to other parts of TMPs. To rule out every other scenario, we have compared the accumulation of DCMs in cytosolic extensions with and without charged residues, as well as extra-cytosolic extensions. The enrichment values calculated separately for these regions show (Fig. 3) that the main enrichment effect can be traced to the presence of inside positive residues.

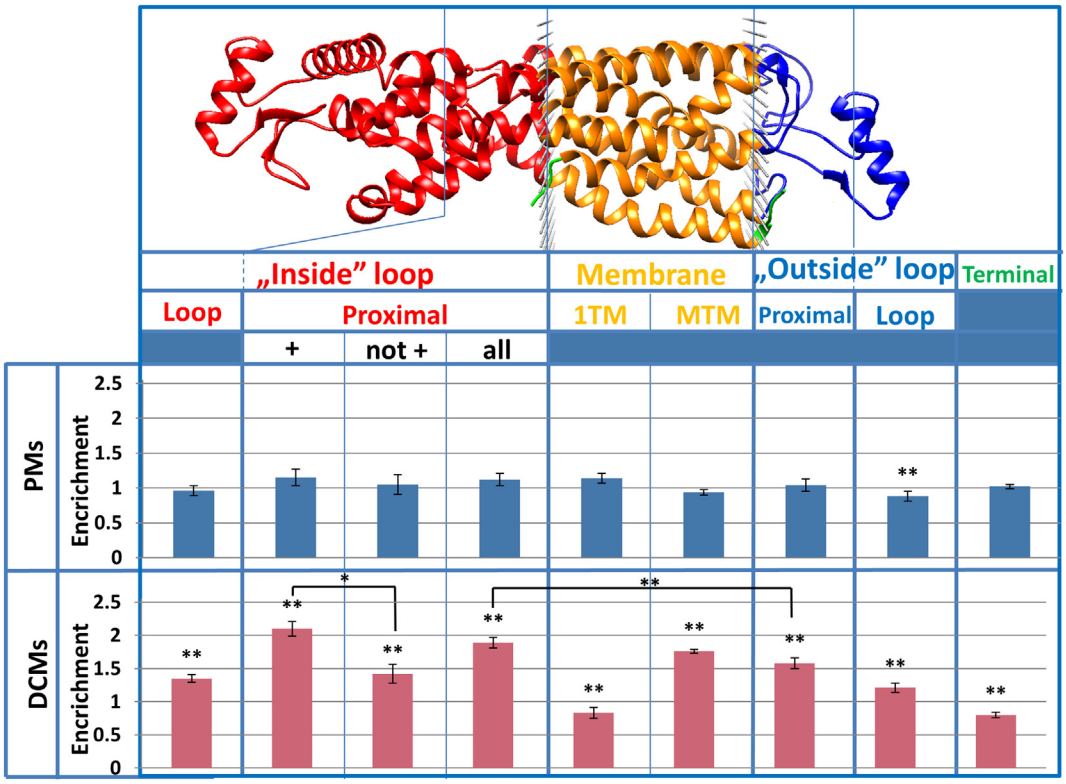

Fig. 3. A, Enrichments of variations on different parts of TMPs. 1TM marks single pass TMPs, and MTM marks polytopic TMPs. ** and * mark significant *p*-values (<0.01 and <0.05, respectively) according to χ^2 test on the observed values (blue, PMs; red, DCMs).

In addition to the positive-inside rule, it was reported that there is a negative-inside depletion/outside enrichment rule [27]. Therefore, we performed an analysis on outside membrane proximal residues. We discriminated regions containing negatively charged residues (D, E). According to this analysis, DCMs are accumulated in membrane proximal extra-cytosolic regions, although negatively charged residues occur in these regions (Supplementary Table). Although a slight increasement is presented in regions containing D, E residues, this result is not significant according to χ^2 test (Supplementary Table).

In contrast, DCMs are depleted in terminal regions, both in polytopic and bitopic proteins (Fig. 3). According to GO overrepresentation analysis, genes mutated near the membrane were exclusively enriched in proteins related to various transport functions, while GO term for genes carrying DCMs in terminal regions included signaling processes as well, for example, "sensory perception of pain" or "localization" (see Supplementary Table).

DCMs in protein–protein interfaces are influenced by how folding occurs

In order to assess how various mutations affect ordered, globular structures and PPIs mediated by them, mutations falling into oligomeric protein complex structures from the PDB were examined (Supplementary Fig. 3). DCMs are enriched on buried residues and protein–protein interfaces; however, exposed residues are depleted in them. In contrast, PMs show a reverse trend being enriched on exposed residues and depleted on buried residues. The difference between DCMs and PMs on buried residues shows that the structural integrity of a folded protein can be more easily disrupted via their core, and hence, these regions are generally less tolerant to variations. A very similar trend is true in the case of residues belonging to interaction interfaces: in these cases, the individual protein structures stay intact, but the complex formation can be often blocked or the interaction strength can be significantly shifted via single-residue changes.

The distribution of DCMs seems selective not only for the spatial location of the residues but also on the secondary structure they belong to. Analyzing protein structures using DSSP for secondary structure assignation and examining the location of variations show that extended residues are enriched in DCMs (especially when mutated to Proline), while irregular residues are depleted in them (Supplementary Fig. 4).

Surface (Supplementary Fig. 5) and secondary structure analyses (Supplementary Fig. 6) were performed on single-subunit proteins as well, and the

results confirm observations made on the oligomer structure set.

While all studied protein complexes have a stable structure, there can be significant differences between individual constituent proteins regarding their monomeric structures. It has been shown that while many IDPs adopt a stable conformation upon interacting with protein partners, their inherent flexibility in their unbound form has a serious impact on the resulting complex and its biophysical properties. We aimed to gain an insight into how the monomeric structural properties of proteins forming interactions reflect in how various mutations target their residues. Regarding the unbound structure of interacting proteins, we distinguish three basic types of protein complexes: autonomous folding and independent binding (i.e., the binding of two or more ordered proteins), coupled folding and binding (where one or more ordered proteins serve as a template to stabilize an IDP partner) and mutual synergistic folding (interactions formed exclusively by IDPs). As known PPIs are dominated by ordered proteins, complexes involving IDPs are scarce in comparison. As a result, the amount of data is very low in some cases; however, several interesting trends can be noticed. DCMs are somewhat enriched on ordered–ordered PPIs; however, this accumulation is not significant, in contrast to complex interfaces where one of the partners is disordered. Moreover, the enrichment of DCMs is even higher for complexes where all partners are disordered (Fig. 4). This trend is unexpected, since DCMs are depleted in disordered residues in general; however, it suggests that disorder to order transitions are prime targets for disease-associated variations. As IDPs undergoing coupled folding and binding almost always donate all their residues into the interaction, these IDPs have virtually no buried residues in their complexed forms. As a result, this category cannot be studied separately. While for IDPs undergoing mutual synergistic folding, a restricted fraction of residues can get buried without being in direct contact with the partner, these residues do not harbor any known DCMs, as are also exempt from further analyses. It is also notable, however, that the data set of complexes exclusively formed by ordered proteins contains dimers only. According to Supplementary Fig. 3, DCMs are enriched on PPIs, when all structures are considered, regardless the number of participating proteins chains. Two proteins were manually discarded from this analysis: transthyretin and superoxide dismutase (UniProt AC: P02766 and P00441, respectively), as they contain extremely high numbers of DCMs that would bias observations. Both proteins represent specialized cases for disease development, having high tendencies to aggregate and form fibrils. These proteins and their corresponding illnesses (familial amyloid polyneuropathy and amyotrophic lateral sclerosis) are heavily studied concerning the structural and functional consequences of mutations.

Stabilizing mutations on exposed residues are harmful when they are in the proximity of protein–protein interfaces

The contribution of a mutation to the stability of a protein or a protein complex can be approximated by the free energy change calculated from introducing the mutation into known structures using FoldX. The calculated $\Delta\Delta G$ energy changes can be categorized as highly stabilizing (<-1.84 kcal/mol), stabilizing (-1.84 to -0.92 kcal/mol), slightly stabilizing (-0.92 to -0.46 kcal/mol), neutral (-0.46 to $+0.46$ kcal/mol), slightly destabilizing ($+0.46$ to $+0.92$ kcal/mol), destabilizing ($+0.92$ to $+1.84$ kcal/mol) and highly destabilizing ($> +1.84$ kcal/mol) [28]. In general, PMs perturb the overall stability to a much lower extent compared to DCMs (Supplementary Fig. 3). This shows that the primary effect of DCMs is the disruption of protein structure by the disruption of PPIs via sterical effects. This idea is supported by the less emphasized difference between $\Delta\Delta G$ values of DCMs and PMs that occur on exposed residues (Fig. 4). Although extreme values slightly shift these averages, PMs are in general neutral on PPIs and exposed residues, and slightly destabilizing on buried residues. In contrast, DCMs are highly destabilizing on PPIs and buried residues, however—similarly to PMs—when they occur on exposed surfaces, their destabilizing effect is much lower (Supplementary Fig. 3). There are several individual cases, when the energy change is negative, yet the variation is harmful. This means that protein structure is not disrupted and the impaired function is caused by other event (e.g., on PPIs, this can be explained that the overall contribution promotes a more compact structure; however, the steric changes prevent the highly specific binding. Another plausible scenario is that interactions required to be dynamic with partners are able to come apart; however, the mutation prevents the dissociation). Over-stabilization can also change catalytic activity or promote aggregation. Considering the three types of complex formation, the same general observations can be made. However, when only one of the partners is disordered, apparently DCMs cause a slightly negative (nearly zero) energy change (on average) on exposed residues.

Disease-associated mutations modulate PTMs at multiple levels

PTMs are regulatory mechanisms that influence the function, localization or the interactions of a protein. Given their importance and their extremely localized nature (being composed of a single residue), PTMs are prime candidate targets for disease-associated mutations, and their alterations are expected to have serious effects. As the most widely occurring PTM is Ser/Thr/Tyr phosphorylation, our analysis of the co-occurrences of PTMs and DCMs/PMs was focused on

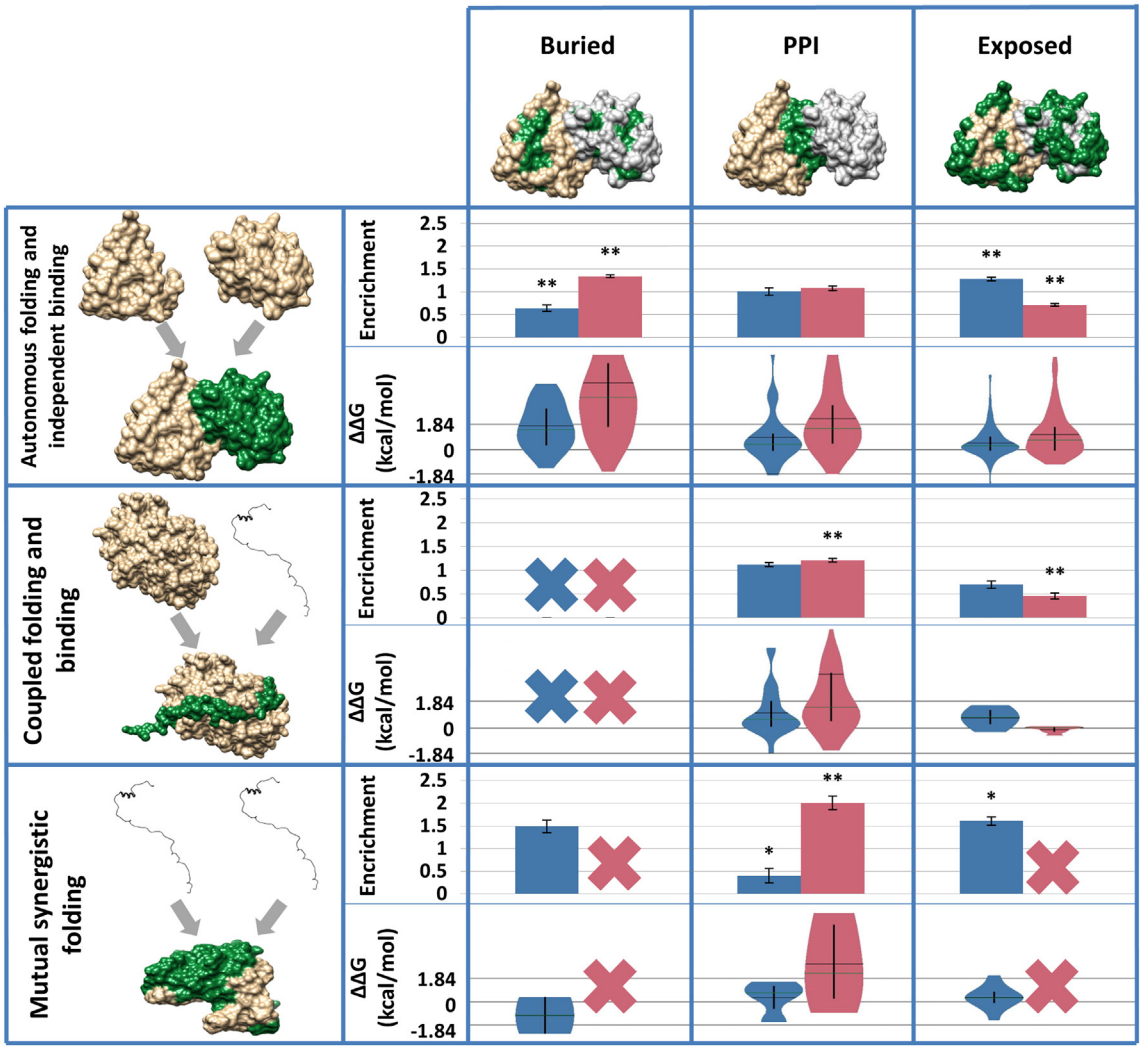

Fig. 4. Variations on protein complex surfaces separated by how folding and binding occurs. Upper rows: Enrichments of variations on different surfaces. ** and * marks significant *p*-value (<0.01 and <0.05, respectively). Lower rows: part of energy change distributions on different surfaces. Vertical black line: interquartile range, horizontal black line: median, horizontal green line: average. The grids help to identify (de)stabilizing ranges. "X" marks surface parts without sufficient number of variations for analysis (blue, PMs; red, DCMs).

phosphosites not considering other types of PTMs, such as methylation or acetylation.

Phosphorylation attaches a phosphate group on one of the three residues Ser, Thr or Tyr. The most fundamental biophysical effect of these modifications is the introduction of a negative charge on an otherwise neutral residue. Phosphorylation events often function as a switch between the non-modified (OFF) state and the modified, charged (ON) state. Mutations affecting phosphosites can rewire these switches in two different ways, with changes to Asp or Glu permanently mimicking the negative charge contribution of the phosphorylation (forced ON state), or changes to any other positively charged or neutral amino acids that abolish the phosphorylation event (forced OFF state).

Table 1 shows the enrichment of DCMs and PMs in phosphorylation sites in disordered regions, and exposed and interface residues of protein complexes. Strikingly, mutations that lead to a forced ON state are virtually non-existent among either DCMs or PMs. While there are two cases where a DCM leads to the induction of a negative charge, both cases entail a Tyr → Asp change. While Asp or Glu can functionally replace a phosphorylated Ser of Thr, both naturally occurring negatively charged amino acids are generally very poor phosphomimetics for Tyr. In contrast, a forced OFF state is achieved in several known cases of phosphosites. The overlap between phosphosites and DCMs/PMs does not differ to a large extent when considering disordered regions. However, in protein complexes, DCMs seem to preferentially target

Table 1. Relative frequencies of mutations occurring on phosphorylation sites in various structural parts (‰).

	Ordered protein–protein complexes				Disordered protein regions	
	Exposed PTMs		Interface PTMs			
Direct modulation (mutated phosphosite)	PMs	DCMs	PMs	DCMs	PMs	DCMs
Forced ON state	0	4	0	0	0	0
Forced OFF state	8	20	14	40	3	1

interface residues when compared to PMs. This preference is easy to explain, as the abolishment of a phosphorylation event can directly modulate the interaction. Interestingly, the difference between PM and DCM enrichment in phosphosites is even larger for exposed residues, outside of direct contact with the partner in the complex structures. These exposed sites might mark recognition sites for other molecular factors absent from the studied complexes. However, the significant enrichment of DCMs underlines the functional importance of these sites with largely unknown functional relevance. This possibly indicates that while the available protein complex structures might give a fair representation of interactions from a structural standpoint, however, a large fraction of functional interactions are currently not represented in the PDB at structural detail.

Apart from the direct mutation of phosphosites, single-residue changes can have a more subtle, indirect effect on phosphorylation-mediated processes. The change in net charge in the spatial vicinity of a phosphosite can alter the interaction mediated by the PTM via electrostatic forces. To test the relevance of this assumption in the case of germline DCMs, protein structures were evaluated with regard to the typical distance between phosphosites and DCMs/PMs that change a neutral amino acid to a charged one. Figure 5 shows that the enrichment of charge-altering various mutations has distinct distributions as a function of the distance measured from the PTM site. Both positive and negative charge-inducing DCMs are enriched in the spatial proximity of PTMs, indicating the relevance of electrostatic modulation. This enrichment vanishes as more distant residues are considered. Negative DCMs seem to have a more long-range effect, with residues even as far as 30–40 Å showing slight enrichments.

Other types of PTMs can be crucially important in functional regulation and signaling processes, with more common ones, such as methylation or acetylation operating through charge as well. These could serve as other possible points of modulation for DCMs, and their analysis would be highly interesting. Unfortunately, current low-throughput data on these modifications is scarce and does not yet enable the systematic analysis of their associations with disease-linked mutations.

Discussion

There are a handful of studies investigating structural aspects of mutations. These papers mostly analyze the presence of disorder, effect of PTMs or the contribution of surfaces to DCMs; however, the different properties are investigated mostly independently. Machine learning is a great and often used [29–34] tool considering interdependence of such features; it rather works like a black-box machine, and therefore, it is hard to place these features into proper biophysical context without a complex analysis. In this manuscript, we attempt to find unrevealed interrelationships between basic

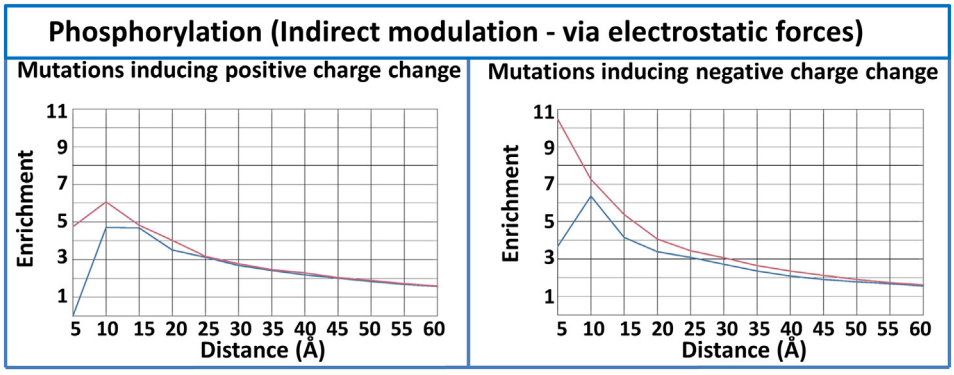

Fig. 5. Enrichment of mutations inducing charge change in the environment of phosphorylated residues (blue, PM; red, DCM).

characteristics. Besides statistical analysis, several pieces of evidence come from detailed structure–function studies of individual proteins.

The amount of germline SNVs and our structural knowledge of the human proteome currently provide the basis for the foundation of large-scale studies concerning the relationship between genetic mutations and protein structure. Our presented work is the first of such works that utilizes the wealth of structural knowledge of human proteins complemented with high-quality TM and disorder predictors available that enable the extension of our analysis to cover the whole human proteome. The usage of state-of-the-art predictors is of utmost importance, as nearly half of human protein-coding genes contain long (>30 amino acid) disordered regions [35], while structure determination of TMPs lags far behind of globular ones, those would largely escape analysis considering experimental data alone.

The power of our approach is apparent in its ability to uncover large-scale systematic trends. These trends clearly show that disease-causing germline mutations (DCMs) are fairly similar to known somatic cancer mutations in many ways. DCMs show a very clear preference for structured protein parts with a pronounced decrease on intrinsically disordered protein regions (IDRs). IDRs are known to be more tolerant against mutations due to the lack of structure that could be disrupted via SNVs. From a phenotypic viewpoint, this would make them ideal candidates for DCMs as mutations in IDRs would potentially lead to weaker phenotype changes, more compatible with the ubiquitous presence of DCMs in all cells. Surprisingly, DCMs seem to not utilize this effect, similarly to cancer mutations [36]. In contrast, DCMs do preferentially target TM regions—a distinct difference compared to somatic cancer mutations. This indicates that while cancer arises primarily as a modulation of regulatory and signaling mechanisms (often manifesting in altered transcription), germline DCMs are more focused on the perturbation of transport mechanisms across membranes.

While the general TM and globular prevalence of DCMs is clear from the above analyses, mutations belonging to given illnesses can show the possible interplay of various structural variations. Considering the structural distribution of DCMs, diseases can be clearly split into various classes. These show that while a large number of conditions have a definite ordered-protein background, in other cases disordered and TM mutations play a heavy role in disease development, and these mutations show interesting coordination patterns. One large class utilizes a fairly high number of disordered mutations, in many cases (such as RETT syndrome; see Fig. 2) complementing variations in ordered domains of the same protein. An alternative molecular approach is presented by diseases mainly focused on the modulation of TM protein regions (such as ceroid lipofuscinosis)—in these cases, the interplay is between TM and ordered domains in the same protein. This shows that illnesses characterized by very different symptoms can share a very similar molecular background, and hence, their therapeutic options may be more similar than expected. In order for structural analyses to have any therapeutic relevance, they must be able to uncover specific molecular ways of alteration that provide a mechanistic understanding of the disease emergence. In accord, we analyzed how DCMs target ordered, disordered and TM regions.

The most germline mutation-sensitive structural regions belong to TM proteins, but this sensitivity is highly dependent of the topology. The highest vulnerability is shown by the membrane-embedded parts of polytopic TMPs. DCMs in this region manifest in the disruption of helix packing either by disrupting weak interactions or by strengthening helix–helix interactions [37]. A prime example of this mechanisms is shown by the thyrotropin receptor, a member of G-protein coupled receptor family that stimulates thyroxine (T4) and triiodothyronine production. Mutation of Val509 in the third TM helix was shown to disrupt the van der Waals packing between the third and fifth TM helices [38,39]. In contrast, the V232D mutation in the middle of the fourth TM segment of cystic fibrosis transmembrane conductance regulator (CFTR) introduces a nonnative hydrogen bond with Q207 of the third TM segment, perturbing the dynamics of the wild-type protein by locking the helices [40]. A second, alternative mechanism is offered by the charge perturbation on membrane proximal parts of polytopic TMPs, especially on the cytosolic side close to positively charged residues. According to the positive-inside rule, positively charged amino acids are more frequent on the cytosolic side near to the membrane [41,42]. It is plausible that variations occurring around this region can prevent the proper folding of TMPs by effectively modulating their orientation via perturbed interaction with the negatively charged glutamate in the cytosolic segment of the translocon, the negatively charged phospholipids on the cytosolic side and the electrochemical membrane potential. These membrane proximal regions are typically highly dynamic and constitute a special class of IDPs [43]. While disordered residues in general are tolerant to variations, these flexible segments are mutation sensitive owing to the fact that they may aid the stabilization of the protein by interacting with lipid head groups.

The vast majority of DCMs in membrane proximal loops alter the transport activity if either positively charged [44,45] or other residues [46,47] are altered, as they are in direct proximity of material transport. A well-characterized such example is TRPV4, a nonselective osmotic and mechano-sensitive cation channel. Different mutations near a functionally important charged residue next to the cytosol-membrane boundary boost protein activity by twisting the backbone and increasing the open probability of the channel [48,49] leading to an increased

intracellular calcium concentration modulating bone development (spondylometaphyseal dysplasia Kozlowski disorder). Other known diseases related to signaling [50,51] or energy production [52] also connected to this region, albeit emerging with a lower frequency.

DCMs targeting ordered and disordered proteins provide molecular mechanisms complementary to the so-far discussed TMP mutations. Similarly to known somatic mutations, germline DCMs also preferentially target either the buried core of domains or—in the case of interacting proteins—PPI interfaces. Buried and interface parts have important role to keep up the structural integrity of proteins or protein complexes. For instance, mutations in ABCC6 are responsible for pseudoxanthoma elasticum, a progressive disorder causing the accumulation of deposits of calcium. It was shown that mutations causing pseudoxanthoma elasticum tend to cluster in the domain–domain interface of the complex, where they can disrupt the proper behavior of the protein [8]. Energetic analysis of such DCMs shows that the main effect of these mutations is the destabilization of ordered monomeric or oligomeric protein structures. Surprisingly, calculated energetic changes are highly destabilizing, with average $\Delta\Delta G$ contributions surpassing 2 kcal/mol for single-residue changes. This shows that the weaker phenotypic changes of germline mutations compared to that of somatic variations are not a reduced $\Delta\Delta G$ change but rather the choice of target proteins. An interesting example is provided by the A → V mutation in the leucine-rich repeat in platelet glycoprotein Ib, responsible for Bernard–Soulier syndrome. On one hand, the mutation results in a conformational change that affects the interaction with the membrane. On the other hand, the same mutation decreases the affinity of binding between the protein and the von Willebrand factor, most likely by altering the orientation of negatively charged residues playing a key role in the interaction [53].

In the case of IDPs, such energetic calculations based on structure are not generally feasible. However, IDPs bound to protein partners generally adopt stable conformations that can lend themselves to analyses similar to that of ordered interactions. The studied IDP complexes show that the two mechanisms available for ordered interacting proteins (i.e., stability changes via buried or interface residues) narrow down to a single option of modulating the interaction strength via DCMs. IDPs binding to ordered partners generally adopt an extended conformation, usually devoid of buried residues. Protein complexes formed exclusively by IDPs, however, have a larger fraction of buried residues, and in theory, this could provide an option for destabilization of the oligomeric structure. Still, our results show that this option is basically never used in known examples, and even IDP-only complexes are targeted solely via interface destabilization. Interestingly, in IDP-mediated interactions, the energetic contributions of germline DCMs surpass the contribution of DCMs in ordered structures, providing a more pronounced destabilizing effect. As IDPs tend to mediate weaker interactions, this comparatively larger energetic effect possibly results in a larger increase in K_d values, promoting a more intensive dissociation, as exemplified by acetylcholinesterase, which is a protein complex formed by various subunits forming a tetramer coiled coil upon oligomerization. Missense mutations in the proline-rich attachment domain are responsible for myasthenic syndrome by preventing the formation of the protein assembly, leading to fatigability and muscle weakness [54].

Such weakened structures and interactions can highlight mutations with serious molecular effect providing insights into how the disease forms. However, while such energetic studies are useful, caution should be exercised when interpreting them, as other cellular processes can modulate structural effects. For example, one of the mutations with the highest destabilizing effect occurs in the Polycomb complex protein BMI-1, a protein playing role in neural stem cell self-renewal [55]. C18Y is a PM that disrupts the RING finger helping zinc coordination [56]; however, the calculated $\Delta\Delta G$ value is ~39 kcal/mol, seemingly incompatible with a neutral phenotypic change. The calculated destabilization indeed does happen; however, the protein unfolds and gets degraded. Mouse studies have shown that as long as the healthy protein gets synthesized from a different allele, the total change in fitness is only marginal compared to healthy ones, and symptoms are negligible compared to Bmi1 –/– mutants [57].

In the mechanisms discussed so far, the affected protein unit was a domain, a TM segment or a disordered binding region, all comprising several residues. Apart from these functional units, a large proportion of proteins feature single-residue functional units as well in the form of PTM sites. According to estimates, the human proteome features over a million PTM sites providing prime targets for disease-associated SNVs [58]. As each type of PTMs requires a specific residue in order to occur (e.g., S/T or Y for phosphorylation), mutations can easily disrupt signaling and regulatory events by changing the residue. This would provide a complementary mechanism of regulatory modulation in addition to the previously discussed structural modulation mechanisms. While PTMs are often modulated via somatic cancer mutations, either as removal of the possibility of a PTM or as the introduction of a phosphomimetic residue [59–61], this mechanisms is largely missing in the case of germline DCMs. Phosphomimetic germline mutations are virtually non-existent, and PTM abolishing mutations are also rare. The disease that shows the highest overlap between known DCMs and PTM sites is the von Hippel Lindau syndrome, which is a cancer-disposition syndrome. In this case, DCMs remove

several phosphorylation sites in the VHL gene; however, the actual emergence of the disease state (renal clear cell carcinoma) requires additional somatic mutations to occur [62]. Furthermore, a large fraction of germline VHL mutations coincide with known somatic mutations, showing that VHL represents the borderline between germline and somatic disease mechanisms. This shows that direct PTM modulation predominantly belongs to the realm of somatic mutations with strong phenotype.

Apart from the direct modulation of PTMs, these sites are readily modifiable via less drastic changes as well. Phosphorylation introduces a negative charge to the protein; therefore, it can be assumed that charge deviance near phosphorylation sites can also have serious consequences. Compared to the direct mutation, a weaker effect can be observed by increasing net charge in proximity to the site. For example, the G121R variation in phosphoglucomutase-1 near to the phosphothreonine at position 115 reduces the catalytic activity of the protein, probably by altering the conformation of the phosphorylation site, as proposed by Lee *et al*. [63].

These results emphasize that, to our current knowledge, the primary effect of germline DCMs is the modulation of structural features of proteins and protein complexes. This means that structural information alone without knowledge of regulatory/pathway information of a given site of a protein should largely be enough to determine the detrimental nature of occurring mutations. Structural considerations in the interpretation of disease-associated mutations with weak phenotypes provide the major clues to the understanding of how these diseases develop. Comprehensive studies based on large-scale data can provide this knowledge and can be further utilized to offer better diagnostics.

Methods

Data sets

Human genetic variation data (PMs and DCMs) were obtained from the UniProt database (version 2017_07) (http://www.uniprot.org/docs/humsavar) [64]. Sequences in the human proteome were filtered to 40% identity with CD-HIT [65], and mutations were kept from only one sequence of each redundancy cluster. After redundancy filtering, 28,499 PMs and 20,608 DCMs remained.

The PDB database [66] was filtered, and only X-ray structures with a resolution better than 3.5 Å were kept. Chimeric proteins were discarded. PDB entries were only kept if nearly all residues have structural information (at least 90% of side-chain atoms are present) or it was completely missing (side-chain atoms are missing, corresponding to disordered/highly flexible residues)—the latter cases provided a subset of disordered proteins. In addition, single-subunit and oligomeric structures were separated into two groups.

Protein complexes formed through autonomous folding and independent binding, coupled folding and binding, and mutual synergistic folding were selected the same way as described by Mészáros *et al*. [18] based on the DIBS [67] and MFIB [68] databases. These data sets already employ a consistent level of similarity filtering, thus no further redundancy filtering was performed.

Both PMs and DCMs were mapped to structures from the PDB. This was done using BLAST search on the redundancy filtered human proteome against sequences in the filtered PDB with a 10^{-10} e-value cutoff. Hits above 40% identity were further processed by a greedy algorithm to select the least structures to which the most mutations can be mapped to. This step was performed for both single-subunit and oligomeric structures. A total of 3916 PMs and 8103 DCMs were mapped onto 1532 oligomeric protein structures, and 2654 PMs and 5217 DCMs were mapped onto 1075 single-subunit proteins. Assigning mutations to proteins where the folding and binding mode could be recognized resulted in 576 structures where 223 PMs and 464 DCMs were identified. On TMPs, 8276 PMs and 7055 DCMs were detected.

Classifying residues as ordered, disordered and TM

Residues containing PMs and DCMs were classified into three classes (TM, disordered or ordered) using two different approaches. In the first one, residues in the membrane regions of TMPs were determined using HTP d1.4 [69,70]. Disordered and ordered residues were defined based on DisProt [71], PFAM [72] and IUPred [73,74] by the following rules: Residues of known Pfam domains were considered to be ordered. Residues outside such domains were considered disordered if they were annotated as such in DisProt or if were predicted to be disordered by IUPred using the standard “long” setting. In the case of TMPs, IUPred was run with the “short” setting, as we found that its accuracy is higher for TMPs [75].

To ensure disorder definition is not biased by the prediction, disordered residues were also assigned using three approaches with different stringency and scope, using structural information as well: (I) using only DisProt annotations, (II) considering only residues with missing side-chain atoms in PDB structures, and (III) considering all proteins for which no structural assignment could be done (non-TMP and no available X-ray structure). While the third category obviously contains non-disordered residues as well, it is expected to be enriched in segments with no stable structure and can be useful when being contrasted with results from strictly structured sets.

Topology from HTP was further refined: TMPs were classified as bitopic and polytopic proteins depending of the number of membrane-spanning segments. Connecting loops between TM helices were divided into "outside membrane proximal" [extra-cytosolic, max 5 amino acid (AA) from membrane boundary]; "outside loop" (extra-cytosolic, at least 6 AA from membrane boundary); "inside membrane proximal, near charged" [cytosolic, max 5 AA from membrane boundary with a positively charged (Arg, Lys) residue within the 5AA]; "inside membrane proximal, far from charged" [cytosolic, max 5 AA from membrane boundary without a positively charged (Arg, Lys) residue within the 5AA]; and "inside loop" (cytosolic, at least 6 AA from membrane boundary).

Hierarchical clustering

Hierarchical clustering was done using the hclust package in R, using Ward's method [76] (ward.D2) and euclidean distances. K-means clustering was performed using the kmeans package in R.

Functional analysis

PANTHER Overrepresentation Test (Release 20,171,205) was performed using PANTHER version 13.0 (release 2017-11-12) using PANTHER GO Slim—biological processes [77].

Residue classifications based on structure

Solvent accessible surface area (SASA) is calculated as described by Lee and Richards [78]. Residues belonging to protein–protein complex interfaces were defined by assessing the SASA values for each subunit in the bound form and by removing all other partner chains in the complex. For interface residues, the SASA values in the two configurations are different. All other residues were classified into buried (up to 9% of maximum accessible surface was calculated in the structure) or exposed (at least 36% of maximum accessible surface was present) [79]. Residues for which side-chain atoms were missing in constitutive 5 amino acids were classified as disordered. DSSP [80] was run on all structures, and residues were classified into four groups: helix ("H," "G" and "I" output), extended ("B" and "E"), turn/bend ("S" and "T") and irregular ("-"). Only standard amino acids were considered.

Energy calculations

Energy calculations were performed using FoldX [81]. Only mutations where the assigned structure and the original protein had 100% sequence identity were considered. $\Delta\Delta G$ calculations were executed on previously optimized structures and were performed five times. All reported $\Delta\Delta G$ values represent the average of these independent runs.

PTMs

Phosphorylation, methylation and acetylation data were collected from PhosphoELM [82], UniProt [64] and PhosphoSitePlus [83] using only low-throughput experiments.

Statistical analysis

χ^2 Tests were performed on all data where indicated. In case of low values (below 5 for any category), Fisher exact test was used. To further eliminate the sporadic error of the data and to estimate the standard deviations, bootstrap analysis was performed by randomly selecting 80% of the data 100 times.

Numbers of occurrences were collected in contingency tables to serve as input for statistical analyses (see Supplementary Material and Supplementary Table 1):

	Mutation (PM or DCM)	No mutation
Not in structural part	x_1	x_2
In structural part	x_3	x_4

The enrichment of one kind of mutation in a certain structural part is the relative frequency of a variation in a given structural part divided by the relative frequency of residues in the given structural part in all residues:

Ratio of relative frequencies (also referred to as enrichment) $= \frac{\frac{x_3}{x_1+x_3}}{\frac{x_3+x_4}{x_1+x_2+x_3+x_4.}}$

	DCM	PM
Not in structural part	d_1	p_1
In structural part	d_2	p_2

Odds ratio was calculated as described Gao *et al.* [12]: $\frac{d_2/p_2}{d_1/p_1}$

Kolmogorov–Smirnov tests were performed to compare samples whether they have the same distribution.

Funding

This work was supported by grants K119287 and 125607 from the Hungarian Scientific Research Fund (OTKA), "Momentum" Program of the Hungarian Academy of Sciences (LP2012/35), and FIEK_16-1-2016-0005 provided by the National Research, Development and Innovation Fund of Hungary.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2018.10.005.

Received 18 September 2018;
Accepted 11 October 2018
Available online 22 October 2018

Keywords:
disease-associated mutations;
germline mutations;
transmembrane proteins;
positive-inside rule;
intrinsically disordered proteins

Abbreviations used:
DCM, disease-causing mutation; IDR, intrinsically disordered region; IDP, intrinsically disordered protein; PM, polymorphism; PPI, protein–protein interaction; PTM, post-translational modification; TM, transmembrane; TMP, transmembrane protein.

References

[1] P.J. Thomas, B.H. Qu, P.L. Pedersen, Defective protein folding as a basis of human disease, Trends Biochem. Sci. 20 (1995) 456–459.
[2] S. Sunyaev, V. Ramensky, P. Bork, Towards a structural basis of human non-synonymous single nucleotide polymorphisms, Trends Genet. 16 (2000) 198–200.
[3] Z. Wang, J. Moult, SNPs, protein structure, and disease, Hum. Mutat. 17 (2001) 263–270, https://doi.org/10.1002/humu.22.
[4] Y. Ye, Z. Li, A. Godzik, Modeling and analyzing three-dimensional structures of human disease proteins, Pac. Symp. Biocomput 2006, pp. 439–450.
[5] M. Mort, U.S. Evani, V.G. Krishnan, K.K. Kamati, P.H. Baenziger, A. Bagchi, et al., In silico functional profiling of human disease-associated and polymorphic amino acid substitutions, Hum. Mutat. 31 (2010) 335–346, https://doi.org/10.1002/humu.21192.
[6] T.A.P. de Beer, R.A. Laskowski, S.L. Parks, B. Sipos, N. Goldman, J.M. Thornton, Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset, PLoS Comput. Biol. 9 (2013), e1003382. https://doi.org/10.1371/journal.pcbi.1003382.
[7] H.C. Jubb, A.P. Pandurangan, M.A. Turner, B. Ochoa-Montaño, T.L. Blundell, D.B. Ascher, Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health, Prog. Biophys. Mol. Biol. 128 (2017) 3–13, https://doi.org/10.1016/j.pbiomolbio.2016.10.002.
[8] K. Fülöp, L. Barna, O. Symmons, P. Závodszky, A. Váradi, Clustering of disease-causing mutations on the domain–domain interfaces of ABCC6, Biochem. Biophys. Res. Commun. 379 (2009) 706–709, https://doi.org/10.1016/j.bbrc.2008.12.142.
[9] P. Yue, Z. Li, J. Moult, Loss of protein structure stability as a major causative factor in monogenic disease, J. Mol. Biol. 353 (2005) 459–473, https://doi.org/10.1016/j.jmb.2005.08.020.
[10] D.P. Ng, B.E. Poulsen, C.M. Deber, Membrane protein misassembly in disease, Biochim. Biophys. Acta (2012) https://doi.org/10.1016/j.bbamem.2011.07.046.
[11] M. Li, M. Petukh, E. Alexov, A.R. Panchenko, Predicting the impact of missense mutations on protein-protein binding affinity, J. Chem. Theory Comput. 10 (2014) 1770–1780, https://doi.org/10.1021/ct401022c.
[12] M. Gao, H. Zhou, J. Skolnick, Insights into disease-associated mutations in the human proteome through protein structural analysis, Structure 23 (2015) 1362–1369, https://doi.org/10.1016/j.str.2015.03.028.
[13] V. Vacic, L.M. Iakoucheva, Disease mutations in disordered regions—exception to the rule? Mol. BioSyst. 8 (2012) 27–32, https://doi.org/10.1039/C1MB05251A.
[14] V. Vacic, P.R.L. 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. https://doi.org/10.1371/journal.pcbi.1002709.
[15] T.A. Peterson, A. Adadey, I. Santana-Cruz, Y. Sun, A. Winder, M.G. Kann, DMDM: domain mapping of disease mutations, Bioinformatics 26 (2010) 2458–2459, https://doi.org/10.1093/bioinformatics/btq447.
[16] J. Molnár, G. Szakács, G.E. Tusnády, Characterization of disease-associated mutations in human transmembrane proteins, PLoS One 11 (2016), e0151760. https://doi.org/10.1371/journal.pone.0151760.
[17] S. Li, L.M. Iakoucheva, S.D. Mooney, P. Radivojac, Loss of post-translational modification sites in disease, Biocomput. 2010, World scientific 2009, pp. 337–347, https://doi.org/10.1142/9789814295291_0036.
[18] B. Mészáros, L. Dobson, E. Fichó, G.E. Tusnády, Z. Dosztányi, I. Simon, Interplay between folding and binding modulates protein sequences, structures, functions and regulation, Structure (2018) , (in preparation) https://www.biorxiv.org/content/early/2017/11/02/211524.
[19] F. Bellido, N. Sowada, P. Mur, C. Lázaro, T. Pons, R. Valdés-Mas, et al., Association between germline mutations in BRF1, a subunit of the RNA polymerase III transcription complex, and hereditary colorectal cancer, Gastroenterology 154 (2018)https://doi.org/10.1053/j.gastro.2017.09.005 (181-194. e20).
[20] F.L. Sirota, H.-S. Ooi, T. Gattermayer, G. Schneider, F. Eisenhaber, S. Maurer-Stroh, Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset, BMC Genomics 11 (Suppl. 1) (2010) S15, https://doi.org/10.1186/1471-2164-11-S1-S15.
[21] S. Vucetic, C.J. Brown, A.K. Dunker, Z. Obradovic, Flavors of protein disorder, Proteins 52 (2003) 573–584, https://doi.org/10.1002/prot.10437.
[22] M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, T. Schwede, Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology, Sci. Rep. 7 (2017) 10480, https://doi.org/10.1038/s41598-017-09654-8.
[23] G.E. Tusnády, Z. Dosztányi, I. Simon, TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates, Bioinformatics 21 (2005) 1276–1277, https://doi.org/10.1093/bioinformatics/bti121.
[24] S.D. Orwig, Y.L. Tan, N.P. Grimster, Z. Yu, E.T. Powers, J.W. Kelly, et al., Binding of 3,4,5,6-tetrahydroxyazepanes to the acid-β-glucosidase active site: implications for pharmacological

chaperone design for Gaucher disease, Biochemistry 50 (2011) 10647–10657, https://doi.org/10.1021/bi201619z.

[25] R.I. Wakefield, B.O. Smith, X. Nan, A. Free, A. Soteriou, D. Uhrin, et al., The solution structure of the domain from MeCP2 that binds to methylated DNA, J. Mol. Biol. 291 (1999) 1055–1065, https://doi.org/10.1006/jmbi.1999.3023.

[26] G. Von Heijne, Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule, J. Mol. Biol. 225 (2) (1992) 487–494.

[27] J.A. Baker, W.-C. Wong, B. Eisenhaber, J. Warwicker, F. Eisenhaber, Charged residues next to transmembrane regions revisited: "positive-inside rule" is complemented by the "negative inside depletion/outside enrichment rule", BMC Biol. 15 (2017) 66, https://doi.org/10.1186/s12915-017-0404-4.

[28] R.A. Studer, P.-A. Christin, M.A. Williams, C.A. Orengo, Stability–activity tradeoffs constrain the adaptive evolution of RubisCO, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 2223–2228, https://doi.org/10.1073/pnas.1310811111.

[29] J. Lugo-Martinez, V. Pejaver, K.A. Pagel, S. Jain, M. Mort, D. N. Cooper, et al., The loss and gain of functional amino acid residues is a common mechanism causing human inherited disease, PLoS Comput. Biol. 12 (2016), e1005091. https://doi.org/10.1371/journal.pcbi.1005091.

[30] C.T. Saunders, D. Baker, Evaluation of structural and evolutionary contributions to deleterious mutation prediction, J. Mol. Biol. 322 (2002) 891–901.

[31] P. Yue, E. Melamud, J. Moult, SNPs3D: candidate gene and SNP selection for association studies, BMC Bioinf. 7 (2006) 166, https://doi.org/10.1186/1471-2105-7-166.

[32] E. Capriotti, R.B. Altman, Improving the prediction of disease-related variants using protein three-dimensional structure, BMC Bioinf. 12 (Suppl. 4) (2011) S3, https://doi.org/10.1186/1471-2105-12-S4-S3.

[33] B. Schuster-Böckler, A. Bateman, Protein interactions in human genetic diseases, Genome Biol. 9 (2008) R9, https://doi.org/10.1186/gb-2008-9-1-r9.

[34] M.W. Gonzalez, M.G. Kann, Chapter 4: protein interactions and disease, PLoS Comput. Biol. 8 (2012), e1002819. https://doi.org/10.1371/journal.pcbi.1002819.

[35] M.E. Oates, P. Romero, T. Ishida, M. Ghalwash, M.J. Mizianty, B. Xue, et al., D2P2: database of disordered protein predictions, Nucleic Acids Res. 41 (2013) D508–D516, https://doi.org/10.1093/nar/gks1226.

[36] M. Pajkos, B. Mészáros, I. Simon, Z. Dosztányi, Is there a biological cost of protein disorder? Analysis of cancer-associated mutations, Mol. BioSyst. 8 (2012) 296–307, https://doi.org/10.1039/c1mb05246b.

[37] D.P. Ng, B.E. Poulsen, C.M. Deber, Membrane protein misassembly in disease, Biochim. Biophys. Acta Biomembr. 1818 (2012) 1115–1122, https://doi.org/10.1016/j.bbamem.2011.07.046.

[38] B. Karges, G. Krause, J. Homoki, K.-M. Debatin, N. de Roux, W. Karges, TSH receptor mutation V509A causes familial hyperthyroidism by release of interhelical constraints between transmembrane helices TMH3 and TMH5, J. Endocrinol. 186 (2005) 377–385, https://doi.org/10.1677/joe.1.06208.

[39] R. Núñez Miguel, J. Sanders, J. Furmaniak, B.R. Smith, Structure and activation of the TSH receptor transmembrane domain, Auto Immun. Highlights 8 (2017) 2, https://doi.org/10.1007/s13317-016-0090-1.

[40] A.G. Therien, F.E.M. Grant, C.M. Deber, Interhelical hydrogen bonds in the CFTR membrane domain, Nat. Struct. Biol. 8 (2001) 597–601, https://doi.org/10.1038/89631.

[41] G. von Heijne, The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology, EMBO J. 5 (1986) 3021–3027.

[42] A. Elofsson, G. von Heijne, Membrane protein structure: prediction versus reality, Annu. Rev. Biochem. (2007) https://doi.org/10.1146/annurev.biochem.76.052705.163539.

[43] I. Stavropoulos, N. Khaldi, N.E. Davey, K. O'Brien, F. Martin, D.C. Shields, Protein disorder and short conserved motifs in disordered regions are enriched near the cytoplasmic side of single-pass transmembrane proteins, PLoS One 7 (2012), e44389. https://doi.org/10.1371/journal.pone.0044389.

[44] G. Conseil, R.G. Deeley, S.P.C. Cole, Functional importance of three basic residues clustered at the cytosolic interface of transmembrane helix 15 in the multidrug and organic anion transporter MRP1 (ABCC1), J. Biol. Chem. 281 (2006) 43–50, https://doi.org/10.1074/jbc.M510143200.

[45] S. Ellard, S.E. Flanagan, C.A. Girard, A.-M. Patch, L.W. Harries, A. Parrish, et al., Permanent neonatal diabetes caused by dominant, recessive, or compound heterozygous SUR1 mutations with opposite functional effects, Am. J. Hum. Genet. 81 (2007) 375–382, https://doi.org/10.1086/519174.

[46] G. Nicolas, C. Pottier, C. Charbonnier, L. Guyant-Marechal, I. Le Ber, J. Pariente, et al., Phenotypic spectrum of probable and genetically-confirmed idiopathic basal ganglia calcification, Brain 136 (2013) 3395–3407, https://doi.org/10.1093/brain/awt255.

[47] P.J. Schwartz, S.G. Priori, R. Dumaine, C. Napolitano, C. Antzelevitch, M. Stramba-Badiale, et al., A molecular link between the sudden infant death syndrome and the long-QT syndrome, N. Engl. J. Med. 343 (2000) 262–267, https://doi.org/10.1056/NEJM200007273430405.

[48] D. Krakow, J. Vriens, N. Camacho, P. Luong, H. Deixler, T.L. Funari, et al., Mutations in the gene encoding the calcium-permeable ion channel TRPV4 produce spondylometaphyseal dysplasia, Kozlowski type and metatropic dysplasia, Am. J. Hum. Genet. 84 (2009) 307–315, https://doi.org/10.1016/j.ajhg.2009.01.021.

[49] J. Teng, S.H. Loukin, A. Anishkin, C. Kung, A competing hydrophobic tug on L596 to the membrane core unlatches S4–S5 linker elbow from TRP helix and allows TRPV4 channel to open, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 11847–11852, https://doi.org/10.1073/pnas.1613523113.

[50] A. Hinney, A. Schmidt, K. Nottebom, O. Heibült, I. Becker, A. Ziegler, et al., Several mutations in the melanocortin-4 receptor gene including a nonsense and a frameshift mutation associated with dominantly inherited obesity in humans, J. Clin. Endocrinol. Metab. 84 (1999) 1483–1486, https://doi.org/10.1210/jcem.84.4.5728.

[51] X. Wang, V. Reid Sutton, J. Omar Peraza-Llanes, Z. Yu, R. Rosetta, Y.-C. Kou, et al., Mutations in X-linked PORCN, a putative regulator of Wnt signaling, cause focal dermal hypoplasia, Nat. Genet. 39 (2007) 836–838, https://doi.org/10.1038/ng2057.

[52] I. Audo, K. Bujakowska, E. Orhan, C.M. Poloschek, S. Defoort-Dhellemmes, I. Drumare, et al., Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness, Am. J. Hum. Genet. 90 (2012) 321–330, https://doi.org/10.1016/j.ajhg.2011.12.007.

[53] J. Ware, S.R. Russell, P. Marchese, M. Murata, M. Mazzucato, L. De Marco, et al., Point mutation in a leucine-rich repeat of platelet glycoprotein Ib alpha resulting in the Bernard–Soulier syndrome, J. Clin. Invest. 92 (1993) 1213–1220, https://doi.org/10.1172/JCI116692.

[54] K. Ohno, A.G. Engel, J.M. Brengman, X.M. Shen, F. Heidenreich, A. Vincent, et al., The spectrum of mutations causing end-plate acetylcholinesterase deficiency, Ann. Neurol. 47 (2000) 162–170.

[55] A.V. Molofsky, R. Pardal, T. Iwashita, I.-K. Park, M.F. Clarke, S.J. Morrison, Bmi-1 dependence distinguishes neural stem cell self-renewal from progenitor proliferation, Nature 425 (2003) 962–967, https://doi.org/10.1038/nature02060.

[56] J. Zhang, K.D. Sarge, Identification of a polymorphism in the RING finger of human Bmi-1 that causes its degradation by the ubiquitin-proteasome system, FEBS Lett. 583 (2009) 960–964, https://doi.org/10.1016/j.febslet.2009.02.023.

[57] I. Park, D. Qian, M. Kiel, M.W. Becker, M. Pihalja, I.L. Weissman, et al., Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells, Nature 423 (2003) 302–305, https://doi.org/10.1038/nature01587.

[58] P. Tompa, N.E. Davey, T.J. Gibson, M.M. Babu, A million peptide motifs for the molecular biologist, Mol. Cell 55 (2014) 161–169, https://doi.org/10.1016/j.molcel.2014.05.032.

[59] M. Krassowski, M. Paczkowska, K. Cullion, T. Huang, I. Dzneladze, B.F.F. Ouellette, et al., ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins, Nucleic Acids Res. 46 (2018) D901–D910, https://doi.org/10.1093/nar/gkx973.

[60] Y. Bu, X. Li, Y. He, C. Huang, Y. Shen, Y. Cao, et al., A phosphomimetic mutant of RelA/p65 at Ser536 induces apoptosis and senescence: an implication for tumor-suppressive role of Ser536 phosphorylation, Int. J. Cancer 138 (2016) 1186–1198, https://doi.org/10.1002/ijc.29852.

[61] H.R. Qin, H.-J. Kim, J.-Y. Kim, E.M. Hurt, G.J. Klarmann, B.T. Kawasaki, et al., Activation of signal transducer and activator of transcription 3 through a phosphomimetic serine 727 promotes prostate tumorigenesis independent of tyrosine 705 phosphorylation, Cancer Res. 68 (2008) 7736–7741, https://doi.org/10.1158/0008-5472.CAN-08-1125.

[62] W.Y. Kim, W.G. Kaelin, Role of VHL gene mutation in human cancer, J. Clin. Oncol. 22 (2004) 4991–5004, https://doi.org/10.1200/JCO.2004.05.061.

[63] Y. Lee, K.M. Stiers, B.N. Kain, L.J. Beamer, Compromised catalysis and potential folding defects in in vitro studies of missense mutants associated with hereditary phosphoglucomutase 1 deficiency, J. Biol. Chem. 289 (2014) 32010–32019, https://doi.org/10.1074/jbc.M114.597914.

[64] S. Pundir, M.J. Martin, C. O'Donovan, UniProt Protein Knowledgebase, Methods Mol. Biol 2017, pp. 41–55, https://doi.org/10.1007/978-1-4939-6783-4_2.

[65] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics (2012) https://doi.org/10.1093/bioinformatics/bts565.

[66] 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, https://doi.org/10.1093/NAR/28.1.235.

[67] E. Schad, E. Fichó, R. Pancsa, I. Simon, Z. Dosztányi, B. Mészáros, DIBS: a repository of disordered binding sites mediating interactions with ordered proteins, Bioinformatics (2017) https://doi.org/10.1093/bioinformatics/btx640.

[68] E. Fichó, I. Reményi, I. Simon, B. Mészáros, MFIB: a repository of protein complexes with mutual folding induced by binding, Bioinformatics (2017) https://doi.org/10.1093/bioinformatics/btx486.

[69] L. Dobson, I. Reményi, G.E. Tusnády, The human transmembrane proteome, Biol. Direct 10 (2015) 31, https://doi.org/10.1186/s13062-015-0061-x.

[70] L. Dobson, I. Reményi, G.E. Tusnády, CCTOP: a Consensus Constrained TOPology prediction web server, Nucleic Acids Res. (2015)https://doi.org/10.1093/nar/gkv451.

[71] D. Piovesan, F. Tabaro, I. Mičetić, M. Necci, F. Quaglia, C.J. Oldfield, et al., DisProt 7.0: a major update of the database of disordered proteins, Nucleic Acids Res. 45 (2016), gkw1279. https://doi.org/10.1093/nar/gkw1279.

[72] 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, https://doi.org/10.1093/nar/gkv1344.

[73] 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, https://doi.org/10.1093/bioinformatics/bti541.

[74] Z. Dosztányi, V. Csizmók, 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, https://doi.org/10.1016/j.jmb.2005.01.071.

[75] G.E. Tusnády, L. Dobson, P. Tompa, Disordered regions in transmembrane proteins, Biochim. Biophys. Acta Biomembr. 1848 (2015) 2839–2848, https://doi.org/10.1016/j.bbamem.2015.08.002.

[76] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244, https://doi.org/10.1080/01621459.1963.10500845.

[77] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, et al., PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements, Nucleic Acids Res. (2016), gkw1138. https://doi.org/10.1093/nar/gkw1138.

[78] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.

[79] B. Rost, C. Sander, Conservation and prediction of solvent accessibility in protein families, Proteins Struct. Funct. Genet. 20 (1994) 216–226, https://doi.org/10.1002/prot.340200303.

[80] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers (1983) https://doi.org/10.1002/bip.360221211.

[81] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: an online force field, Nucleic Acids Res. 33 (2005) W382–W388, https://doi.org/10.1093/nar/gki387.

[82] H. Dinkel, C. Chica, A. Via, C.M. Gould, L.J. Jensen, T.J. Gibson, et al., Phospho.ELM: a database of phosphorylation sites—update 2011, Nucleic Acids Res. 39 (2011) D261–D267, https://doi.org/10.1093/nar/gkq1104.

[83] P.V. Hornbeck, B. Zhang, B. Murray, J.M. Kornhauser, V. Latham, E. Skrzypek, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, Nucleic Acids Res. 43 (2015) D512–D520, https://doi.org/10.1093/nar/gku1267.