Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins

Sina Ghadermarzi[1], *Xingyi Li*[2], *Min Li*[2]* *and Lukasz Kurgan*[1]*

[1] *Department of Computer Science, Virginia Commonwealth University, Richmond, VA, United States,* [2] *School of Computer Science and Engineering, Central South University, Changsha, China*

Recent research shows that majority of the druggable human proteome is yet to be annotated and explored. Accurate identification of these unexplored druggable proteins would facilitate development, screening, repurposing, and repositioning of drugs, as well as prediction of new drug–protein interactions. We contrast the current drug targets against the datasets of non-druggable and possibly druggable proteins to formulate markers that could be used to identify druggable proteins. We focus on the markers that can be extracted from protein sequences or names/identifiers to ensure that they can be applied across the entire human proteome. These markers quantify key features covered in the past works (topological features of PPIs, cellular functions, and subcellular locations) and several novel factors (intrinsic disorder, residue-level conservation, alternative splicing isoforms, domains, and sequence-derived solvent accessibility). We find that the possibly druggable proteins have significantly higher abundance of alternative splicing isoforms, relatively large number of domains, higher degree of centrality in the protein-protein interaction networks, and lower numbers of conserved and surface residues, when compared with the non-druggable proteins. We show that the current drug targets and possibly druggable proteins share involvement in the catalytic and signaling functions. However, unlike the drug targets, the possibly druggable proteins participate in the metabolic and biosynthesis processes, are enriched in the intrinsic disorder, interact with proteins and nucleic acids, and are localized across the cell. To sum up, we formulate several markers that can help with finding novel druggable human proteins and provide interesting insights into the cellular functions and subcellular locations of the current drug targets and potentially druggable proteins.

Keywords: drug targets, druggability, druggable human proteome, drug-protein interactions, protein-protein interactions, intrinsic disorder

OPEN ACCESS

Edited by:
Shandar Ahmad,
Jawaharlal Nehru University,
India

Reviewed by:
Marcelo Adrian Marti,
University of Buenos Aires,
Argentina
Yuanyuan Wang,
St. Jude Children's Research
Hospital, United States

Correspondence:
Min Li
limin@mail.csu.edu.cn
Lukasz Kurgan
lkurgan@vcu.edu

Specialty section:
This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: *30 May 2019*
Accepted: *09 October 2019*
Published: *xx Month 2019*

Citation:
Ghadermarzi S, Li X, Li M and
Kurgan L (2019) Sequence-Derived
Markers of Drug Targets and
Potentially Druggable
Human Proteins.
Front. Genet. 10:1075.
doi: 10.3389/fgene.2019.01075

INTRODUCTION

Knowledge of the drug-target interactions is essential for numerous applications including screening of drug candidates (Schneider, 2010; Núñez et al., 2012; Dalkas et al., 2013; Tseng and Tuszynski, 2015), drug repositioning and repurposing (Chong and Sullivan, 2007; Haupt and Schroeder, 2011; Oprea and Mestres, 2012; Hu and Bajorath, 2013; Li et al., 2016), characterization and mitigation of side-effects of drugs (Lounkine et al., 2012; Wang et al., 2012b; Kuhn et al., 2013; Tarcsay and Keserű, 2013; Hu et al., 2014), and prediction of novel protein-drug interactions (Wang et al., 2016a; Lotfi

Shahreza et al., 2017; Ezzat et al., 2018; Hao et al., 2019; Wang and Kurgan, 2019; Wang and Kurgan, 2018; Wang et al., 2019). Recent analysis reveals that over 95% of the currently known drug targets are proteins and that these proteins facilitate about 93% of known drug-target interactions (Santos et al., 2017). Thus, we focus on the drug-protein interactions and we use the term “drug target” as a synonym for the protein drug target. While earlier works report about 400 drug targets (Hopkins and Groom, 2002; Russ and Lampel, 2005), subsequent studies annotate as many as over 600 drug targets in human (Santos et al., 2017). Furthermore, the druggable human proteome, defined as the full complement of the human drug targets (Hopkins and Groom, 2002; Russ and Lampel, 2005; Rask-Andersen et al., 2014; Cimermancic et al., 2016; Hu et al., 2016), is expected to be much larger. Early estimates place the number of human drug targets at around 3,000 (Hopkins and Groom, 2002; Russ and Lampel, 2005). A more recent analysis approximates this number at 4.5 thousand (Finan et al., 2017), which corresponds to about 22% of the human genome. While the historically typical drug targets include G-protein coupled receptors, nuclear receptors, ion channels, and some of the enzymes (Overington et al., 2006; Imming et al., 2007), recent works suggest that many of the non-enzymes (e.g., scaffolding, regulatory, and structural proteins) and proteins involved in specific protein-protein interactions (PPIs) should be targeted by drugs (Makley and Gestwicki, 2013; Ozdemir et al., 2019), effectively expanding the list of potential drug targets. These observations point to the fact that many of the drug targets remain to be discovered and characterized. The search for these proteins relies on the concept of druggability, which was originally defined based on the presence of structure that favors interactions with drug-like compounds where the corresponding interactions provide desired therapeutic effects (Hopkins and Groom, 2002; Russ and Lampel, 2005; Keller et al., 2006). In a purely structural context, druggability is related to binding of a compound to a given protein target with high affinity (< 1 μM) (Sheridan et al., 2010; Radusky et al., 2014). We focus on the former definition where both the interactions and the therapeutic effects are considered.

One of the key elements in the quest to find druggable proteins is to identify functional and structural characteristics that differentiate drug targets from the non-drug targets (Zheng et al., 2006; Lauss et al., 2007; Bakheet and Doig, 2009; Zhu et al., 2009b; Zhu et al., 2009c; Bull and Doig, Mitsopoulos et al., 2015; 2015; Feng et al., 2017; Kim et al., 2017). In one of the earliest works, Chen *et al.* concentrated on the analysis of structural fold types, target family representation and similarity, pathway associations, tissue distribution, and chromosome location for the drug targets (Zheng et al., 2006). A similar analysis that considered cellular functions, pathway associations, tissue distribution, and subcellular and chromosome location of the drug targets was published soon after by Lauss and colleagues (Lauss et al., 2007). More recent studies have shifted the focus towards characteristic features of the target protein sequence and structure. Bakheet and Doig used a relatively small set of 148 targets to analyze several sequence properties (chain length, hydrophobicity, charge, and isoelectric point), putative secondary structure and transmembrane regions, inclusion of signal peptides, selected set of post-translational modifications (PTMs), as well as the previously studied subcellular location and functions (Bakheet and Doig, 2009). Subsequently, Bull and Doig investigated a similar set of characteristics using a much larger set of 1324 drug targets (Bull and Doig, 2015). They considered a similar set of sequence properties, native secondary structure and signal peptides, selected PTMs, and a few new properties: the number of germline variants, expression levels, and the number of PPIs (Bull and Doig, 2015). The most recent study by Park, Lee, and colleagues expanded the above list of characteristics by inclusion of gene essentiality and tissue specificity (Kim et al., 2017). Moreover, several articles narrowly focused on characteristics that quantify topological features of the underlying PPI networks (Zhu et al., 2009b; Zhu et al., 2009c; Mitsopoulos et al., 2015; Feng et al., 2017). While these studies have considered a broad range of functional and structural features of drug targets, they identified the drug target-specific characteristics by comparing the drug targets against the other human proteins (non-drug targets). However, many of these non-drug targets could be in fact druggable, i.e., as many as 22% according to (Finan et al., 2017). Using the non-drug targets to represent the non-druggable proteins in order to define characteristic features of the druggable targets ultimately creates a bias toward describing the currently known drug targets. Consequently, this reduces our ability to use these characteristics to identify a complete set of druggable proteins.

We address the abovementioned shortcoming of the prior works by comparing sequence-derived characteristics of the drug targets, possibly druggable proteins, and non-druggable proteins using a large and well-curated dataset of human proteins. Our study is novel in four ways. First, we contrast the drug targets (D dataset) not only against all non-drug targets (N dataset), which was also done in prior studies, but also against non-druggable non-drug targets (Nn dataset; the non-drug targets that exclude disease associated proteins) and against possibly druggable non-drug targets (Nd dataset; the non-drug targets that are associated with multiple diseases). The association of the non-drug targets with diseases is necessary for the druggable proteins to exert therapeutic effects. Second, we further compare the D, N, Nd, and Nd proteins against highly promiscuous drug targets that interact with many drugs (Dh dataset) and drug targets that interact with low number of drugs (Dl dataset). This full-spectrum analysis allows us to pinpoint characteristics that differentiate between drug targets, possibly druggable proteins and non-druggable proteins, as well as features that are specific to promiscuous *vs.* non-promiscuous drug targets. Third, we focus on the characteristics that can be quantified directly from the protein sequence or protein name/identifier. This facilitates their use as potential markers for druggability across the entire human proteome. This is in contrast to several related studies that are limited to a relatively small subset of human proteins with solved structures (Hambly et al., 2006; Bull and Doig, 2015; Hu et al., 2016; Wang et al., 2016a; Wang et al., 2019). Fourth, we include several important sequence/protein-derived characteristic that were missed in the past studies including putative intrinsic disorder, residue-level conservation, presence and number of alternative splicing isoforms, inclusion of domains, and solvent accessibility (surface area). Moreover, we cover some of the key characteristics from the prior works, such as the topological features of PPIs, cellular functions, and subcellular locations.

MATERIALS AND METHODS

Datasets

Datasets of Drug Targets (D Dataset), Highly Promiscuous Drug Targets (Dh Dataset), and Low-Interaction Drug Targets (Dl Dataset)

We collect a comprehensive set of drug targets by combining interaction information extracted from several large bioactive compounds-protein interaction databases. We filter these bioactive compounds to include only approved and experimental drugs. Furthermore, we focus on human proteins by excluding protein fragments and proteins from other organisms. We maximize the coverage by first collecting an inclusive set of interactions (including all bioactive compounds and protein chains) and then applying the two filters to obtain a high quality and large set of drugs and proteins.

The data collection protocol follows the work in (Wang and Kurgan, 2019; Wang and Kurgan, 2018). We extract the source data from three large repositories: Drug2gene (Roider et al., 2014), TTD (Zhu et al., 2009a), and GtP (Harding et al., 2017). Drug2gene is one of the most inclusive repositories that aggregates 19 source databases including TTD and GtP and several other major databases like ChEMBL (Gaulton et al., 2016) and DrugBank (Wishart et al., 2017). However, Drug2gene includes older and substantially smaller version of the TTD and GtP resources. Therefore, we integrated the latest versions of these two databases into our dataset. These databases provide a list of drug-protein pairs that use different identifiers and which include other information that could be useful to identify these molecules (like drug structure). The arguably most popular way to identify drugs and proteins are the PubChem CIDs and UniProt accession numbers, respectively. We use these identifiers to map data between the resources. We also merged the drugs with different PubChem CID but identical *simplified molecular-input line-entry system* (SMILES) structures. First, we remove the data collected from TTD and GTP that lacks PubChem CID or UniProt identifiers. Next, we map the proteins in Drug2gene that are represented by Entrez Gene ID into the corresponding UniProt accession numbers. After mapping and combining these datasets and removing duplicates, we obtain 2,490,057 interactions for 591,684 bioactive compounds and 4,128 proteins. Next, we filter this list of compounds using the list of drugs obtained from the DrugBank and ChEMBL. We remove the compounds that do not have the same CID or SMILES structure when compared to the list of DrugBank and ChEMBL drugs. Finally, we remove non-human proteins using a reference human proteome from UniProt. At the end, the set of drug targets (D dataset) includes 33,104 interactions between 4,405 drugs (PubChem CID) and 1,638 protein (UniProt identifiers). We provide the complete D dataset in the supplement. Moreover, we generate an expanded set of human and human-like drug targets that includes proteins in the D dataset plus proteins from other organisms that share high sequence similarity to the human proteins (D+ dataset). More specifically, following recent works (Hu et al., 2014; Wang et al., 2016a; Wang et al., 2019), human proteins that share at least 90% sequence identity quantified using BLAST with default parameters (Altschul et al., 1997) to any of the drug targets were added into the D+ dataset. Consequently, the D+ dataset has 1,762 proteins including 124 proteins that were included based on the high similarity; we list these proteins in the **Supplementary Material**. The number of drug targets in our dataset is slightly higher than the sizes of the datasets used in related studies (in the inverse chronological order): 1604 in (Feng et al., 2017), 1578 in (Kim et al., 2017), 1324 in (Bull and Doig, 2015), and 1,030 in (Rask-Andersen et al., 2014). Compared to popular databases, such as KEGG DRUG and DrugBank, our dataset features a more complete set of interactions (33,104 *vs.* 14,222 and 23,380, respectively (Wang and Kurgan, 2019) while focusing on a smaller and relevant set drugs that specifically target human proteins [4,405 *vs.* 5,045 and 10,562, respectively (Wang and Kurgan, 2019).

Drug targets in our dataset interact with as few as 1 drug and as many as 443 drugs. We investigate whether sequence-derived and functional characteristics of highly promiscuous drug targets are different from the drug targets that interact with a few proteins. To do that we extracted two subsets of the drug targets, the highly promiscuous targets (Dh dataset) that correspond to the top quartile of the targets with the highest interaction counts, and the low-interaction drug targets (Dl dataset) that include the bottom quartile of the drug targets with the lowest numbers of interactions.

Dataset of Non-Drug Targets (N Dataset)

We contrast the sequence-derived and functional characteristics of the proteins in the D, D+, Dh, and Dl datasets against the proteins that are not current drug targets. We collect these non-drug targets (N dataset) by selecting proteins from the UniProt's human proteome that are not in the D dataset. The selection process follows two rules. First, we match the size of the N dataset to the size of the D dataset to ensure robust statistical comparisons between different datasets. Second, when down-sampling the human proteins we ensure that the selected proteins have similar size as the proteins in the D dataset. More specifically, for each protein in the D dataset we pick a human non-drug target at random (without replacement) that has a matching sequence length (with 10% tolerance). We introduce the latter rule since the amount of intrinsic disorder in proteins is dependent on proteins length (HOWELL et al., 2012). The same selection process was used in several related studies (Meng et al., 2015b; Na et al., 2016; Meng et al., 2018) to eliminate protein size bias when studying intrinsic disorder. We provide the list of the 1,638 size-matched proteins that constitute the N dataset in the **Supplementary Material**. Moreover, Section non-druggable and possibly druggable proteins describes how the N dataset is used to derive the dataset of non-druggable non-drug targets (Nn dataset; the non-drug targets that exclude disease associated proteins) and the dataset of possibly druggable non-drug targets (Nd dataset; the non-drug targets that are associated with multiple diseases).

Characterization of Protein Properties

We characterize a broad collection of characteristics of human proteins that include their disease associations, structural properties derived from the sequence (putative intrinsic disorder and surface), sequence properties (domain annotations, alternative splicing, and residue-level conservation), topological properties of the corresponding PPI network (centrality measures and hubs), and functional properties (GO annotations and predicted protein-binding regions). We extract these characteristics directly from the protein sequence or protein names/identifiers. This means that they could be used as potential markers for druggability that cover the entire human proteome.

Disease Associations

The protein-disease association data were collected from DisGeNET (Gutiérrez-Sacristán et al., 2016). DisGeNET integrates several curated databases and offers arguably one of the most complete levels of coverage for human diseases. This database provides association between disease MeSH IDs and Entrez Gene IDs and also provides a mapping between Entrez Gene IDs and UniProt identifiers. We mapped these annotations to our dataset using the UniProt identifiers.

Sequence-Derived Structural Properties

We annotate two relevant structural properties that we can accurately derive from the protein sequences: intrinsic disorder and solvent accessibility. We are unable to directly collect structural data since significant majority of the proteins in the D, D+, and N datasets do not have solved structures.

Intrinsically disordered proteins and protein regions lack a stable tertiary structure in isolation (Dunker et al., 2013; Habchi et al., 2014; Uversky, 2014a). Proteins with disordered regions are crucial for many key cellular functions including molecular recognition and assembly, cell cycle and cell death regulation, signal transduction, transcription, translation, and viral cycle (Dyson and Wright, 2005; Uversky et al., 2005; Liu et al., 2006; Xie et al., 2007; Peng et al., 2012; Xue et al., 2012; Peng et al., 2013; Uversky et al., 2013; Fan et al., 2014; Fuxreiter et al., 2014; Peng et al., 2014b; Xue and Uversky, 2014; Dolan et al., 2015; Meng et al., 2015a; Meng et al., 2015b; Varadi et al., 2015; Babu, 2016; Na et al., 2016; Yan et al., 2016; Wang et al., 2016b; Kjaergaard and Kragelund, 2017). They are also the main contributors to the dark proteome (Hu et al., 2018; Kulkarni and Uversky, 2018). Intrinsic disorder is abundant in the human proteins. Computational studies estimate that about 19% amino acids in eukaryotic proteins are intrinsically disordered (Peng et al., 2015) and over 40% human proteins have at least one long disordered region with 30 or more consecutive residues (Oates et al., 2013). These proteins are particularly relevant to this study since they are associated with several human diseases (Uversky et al., 2008; Babu, 2016; Uversky et al., 2014; Uversky, 2014b) and since they attract recent interest as potent drug targets (Cheng et al., 2006; Uversky, 2012; Dunker and Uversky, 2010; Ambadipudi and Zweckstetter, 2016; Tantos et al., 2015). Intrinsic disorder can be predicted accurately from protein sequence using computational methods (Peng and Kurgan, 2012; Walsh et al., 2015; Lieutaud et al., 2016; Meng et al., 2017a; Meng et al., 2017b). We use one of the leading disorder predictors, IUPred (Dosztányi et al., 2005; Dosztanyi, 2018). This selection is motivated by the fact that IUPred is computationally efficient (i.e., it can be used to process large datasets of proteins, such as the D and N datasets) and since it provides accurate predictions (Peng and Kurgan, 2012; Walsh et al., 2015). We use the IUPred's results to compute the disorder content (fraction of disordered residues in a given protein) and the length of the putative disordered regions.

Solvent accessibility provides a crucial context for the analysis of the residue-level conservation since it allows us to separate conserved residues that are localized on the surface (which include residues that are instrumental for the drug-protein interaction) from those located in the protein core (which are likely responsible for structural stability of the protein). We predict the relative accessible surface area using the ASAquick method (Faraggi et al., 2014). This method predicts relative solvent accessibility from a single sequence (without alignment), and thus it much faster than the other predictors that require calculation of multiple sequence alignment. It also provides accurate prediction, which is why it was recently used in related studies (Zhang et al., 2017; Amirkhani et al., 2018; Meng and Kurgan, 2018). We convert the numeric relative solvent accessibility of residues into a binary annotation (solvent exposed *vs*. buried) using a threshold of 0.15. This value adequately splits the bimodal distribution of solvent accessibility values for the residues in the combined D and N datasets (**Figure S2** in the **Supplementary Material**). We use these results to quantify the fraction of the putative surface residues in a given protein.

We assess quality of these predictions by comparing values of the fraction of the native surface residues that are computed using a limited set of proteins that have structures against the fraction of the predicted surface residues for the same set of proteins. We utilize mapping generated with the SIFTS resource (Velankar et al., 2013) that is available in UniProt to identify structures of the human proteins from the D and N datasets in the PDB database (Berman et al., 2000). We consider structures that cover at least 90% of the corresponding full protein sequences collected from UniProt to ensure that they correspond to a similar set of residues that are covered by the predictions which rely on the full protein chains. We compute the native solvent accessibility from these structures in three steps. First, we remove other molecules (including other protein chains) from the PDB structures. Second, we use DSSP (Kabsch and Sander, 1983; Joosten et al., 2010) to compute solvent accessibility values. Third, we convert the solvent accessibility into the relative solvent accessibility values using the normalization procedure that is described in the ASAquick article (Faraggi et al., 2014). We were able to collect the native solvent accessibility values for 373 drug targets (including 343 proteins from the D dataset, 55 from the Dh dataset, and 103 from the Dl dataset) and 73 proteins non-drug targets (including 39 from the Nd dataset and 12 from the Nn dataset). This corresponds to (373 + 73)/(1762 + 1,638) = 13% structural coverage of the human proteins in our datasets. **Figure S3** compares the distributions of the fractions of the surface residues computed from the protein structures against the fractions that are based on the predicted solvent accessibility for the seven considered datasets. The distributions that rely on the native *vs.* putative solvent accessibility for each of the seven dataset are very similar. The differences are not statistically significant (*p*-values range between 0.17 for the N dataset and 0.88 for the Nd dataset). This results suggests that the solvent accessibility predicted with ASAquick provides an accurate approximation of the native fraction of the surface residues.

Protein Sequence Properties

We use the proteins sequences to annotate the domains, alternative splicing isoforms, and sequence conservation. We collect the domain annotations from Pfam (Calderone et al., 2013) using UniProt identifiers, and we use these annotations to compute the domain boundaries (fraction of the domain-assigned residues) and the number of domains per protein. We obtain the number of alternative splicing isoforms from the UniProt database (UniProt: the universal protein knowledgebase, 2016). We calculate residue-level conservation scores using the relative entropy measure (Wang and Samudrala, 2006) from the PSSMs generated with PSI-BLAST (Altschul et al., 1997). We use a threshold to convert the numeric conservation scores to binary, i.e., a given residue is either conserved (if its conservation score > threshold) or non-conserved (otherwise). We selected the threshold that corresponds to the 80th percentile of the distribution of the conservation scores for the residues in the combined D and N datasets (**Figure S1** in the Supplementary Material). The corresponding threshold value of 0.63 corresponds to an inflection point in the distribution tail where the conserved residues should be located. Using these annotations, we quantify the rate of the conserved residues in the protein sequence and among the residues located on the putative protein surface, given that this is where the drug-protein interaction occurs.

Topological Properties of the Protein-Protein Interaction Network

Motivated by work in (Zhu et al., 2009b; Zhu et al., 2009c; Mitsopoulos et al., 2015; Feng et al., 2017), we quantify the topological characteristics of drug targets and non-drug targets in the human PPI network. We collected the interaction network from the MENTHA resource (Calderone et al., 2013) and directly mapped it to our datasets using UniProt identifiers. MENTHA integrates data coming from several popular databases of PPIs, such as IntAct (Orchard et al., 2014), MINT (Licata et al., 2012), DIP (Salwinski et al., 2004), BioGRID (Oughtred et al., 2019), and MatrixDB (Launay et al., 2015), providing arguably one of the most comprehensive coverage levels. Several different centrality measures can be used to define topological characteristics of proteins in PPI networks (Wang et al., 2013a). We considered a comprehensive set of measures including betweenness centrality (Freeman, 1977), eigenvector centrality (Bonacich, 1987), closeness centrality (Bavelas, 1950), information centrality (Stephenson and Zelen, 1989), degree centrality (Jeong et al., 2001), subgraph centrality (Estrada and Rodriguez-Velazquez, 2005), network centrality (Wang et al., 2012a), and local average connectivity (Li et al., 2011). We reduced this set by removing measures that are redundant (highly correlated). The corresponding subset of four measures (eigenvector, closeness, betweenness and information centrality) has relatively low mutual correlations (< 0.6) while being highly correlated (> 0.8) with at least one of the removed measures. We give the corresponding correlations between these measures on our datasets in Table S1 in the supplement. The eigenvector centrality is an extension of the node degree in which connections to more important nodes have more impact on the score. The nodes that are connected to many highly connected nodes end up having higher score than nodes which are connected to the same number of less-connected nodes (Bonacich, 1987). The closeness centrality measures the average length of the shortest path from the node to other nodes. The nodes with higher closeness centrality on average have smaller distance to the other nodes (Bavelas, 1950). The betweenness centrality quantifies the frequency with which a given node appears in the shortest paths between nodes in the network. Thus, removal of nodes with high betweenness centrality has big impact on the shortest paths between nodes (Freeman, 1977). Finally, information centrality is based on information along the paths from a given node to the other nodes (Stephenson and Zelen, 1989).

Besides quantifying several different topological features, we also annotate hub proteins, defined as proteins that interact with many proteins (Jeong et al., 2001). While early works on hub proteins defined them using a fixed minimal number of (Jeong et al., 2001), more recent studies use a floating threshold defined as a certain percentage of the most connected nodes in a given interactome (Han et al., 2004; Batada et al., 2006; Dosztányi et al., 2006). This results in different cut-offs that define hubs for different interactomes (different organisms) and emphasizes the fact that hubs are a property of the whole interactome system rather than a property of individual proteins. We used the latter definition using the cut-off that corresponds to the 90th percentile of the interaction counts in the complete human PPI network, which is consistent with several recent studies (Han et al., 2004; Batada et al., 2006; Dosztányi et al., 2006). Therefore, we annotate hub proteins as those that have the number of PPIs in the complete interactome collected from MENTHA that is higher than this threshold (i.e., ≥ 77 interactions).

Hub proteins have increased levels of intrinsic disorder (Meng et al., 2015b; Patil et al., 2010) and the disordered regions are often employed to carry out PPIs (Mohan et al., 2006; Vacic et al., 2007; Yan et al., 2016). The disordered protein-binding regions are also linked to certain human diseases (Uversky, 2018). Thus, we also annotate putative disordered protein binding regions. We use ANCHOR (Dosztányi et al., 2009) to predict the disordered protein-binding residues and we aggregate this information to compute the content of disordered protein binding residues for the proteins in our datasets. The selection of this method is motivated by the fact that it is accurate and popular, and provides fast predictions (i.e., is capable of processing our large datasets) (Meng et al., 2017; Katuwawala et al., 2019).

Functional Properties

We annotate cellular functions and subcellular locations of the drug targets and the non-drug targets using the Gene Ontology (GO) terms (Consortium, 2004), which we collect using the PANTHER system (Muruganujan et al., 2018). We annotate and separately analyze the molecular functions, biological processes, and cellular components, where the latter define the subcellular locations.

Statistical and Similarity Analyses

We compare the sequence-derived and functional characteristics between the drug targets, non-drug targets, and possibly druggable proteins using statistical tests of significance of differences. We quantify the significance of the differences using the *t*-test if the underlying measure of the sequence-derived/functional property has normal distribution, and Wilcoxon rank-sum test otherwise. We used the Anderson-Darling test with the *p*-value cutoff of 0.05 to test normality. We use the Fisher's exact test when comparing binary characteristics, including disease associations and presence of hubs.

We annotate the cellular functions and subcellular locations associated with a particular set of proteins using enrichment analysis offered by the PANTHER system (Muruganujan et al., 2018). This system generates a list of annotations that are statistically over-represented when compared with the annotations present in the whole human proteome. PANTHER quantifies the ratios of enrichment and the corresponding *p*-values for each GO term when compared with the reference human proteome. We focus on the GO terms that occur at least 10 times in our datasets (to ensure robustness of statistical analysis), and we annotate a given term as associated with a particular set of proteins if its ratio > 2 (at least two fold increase) and the associated *p*-value (quantified using the False Discovery Rate correction) is < 0.05.

We measure similarity between two sets of proteins by comparing the cellular function and subcellular location GO terms associated with these two protein sets. We calculate this similarity using the GOSemSim package (Li et al., 2010) with default parameters [Wang et al. measure (Wang et al., 2007)] and the reference set to human.

RESULTS AND DISCUSSION

Non-Druggable and Possibly Druggable Proteins

The set of the non-drug targets likely includes a relatively large number of druggable proteins. The ability to characterize properties that differentiate the drug targets and druggable proteins from the non-drug targets hinges on the annotation of the non-druggable and possibly druggable proteins in the set of these non-drug targets. Druggability of proteins requires that they interact with a drug-like compound and that this interaction provides a desired therapeutic effects (Hopkins and Groom, 2002; Russ and Lampel, 2005; Keller et al., 2006). Thus, one way to annotate possibly druggable and non-druggable proteins is to analyze protein-disease associations. **Figure 1** shows the fractions of the proteins associated with different classes of diseases among the drug targets and the non-drug targets. As expected, the

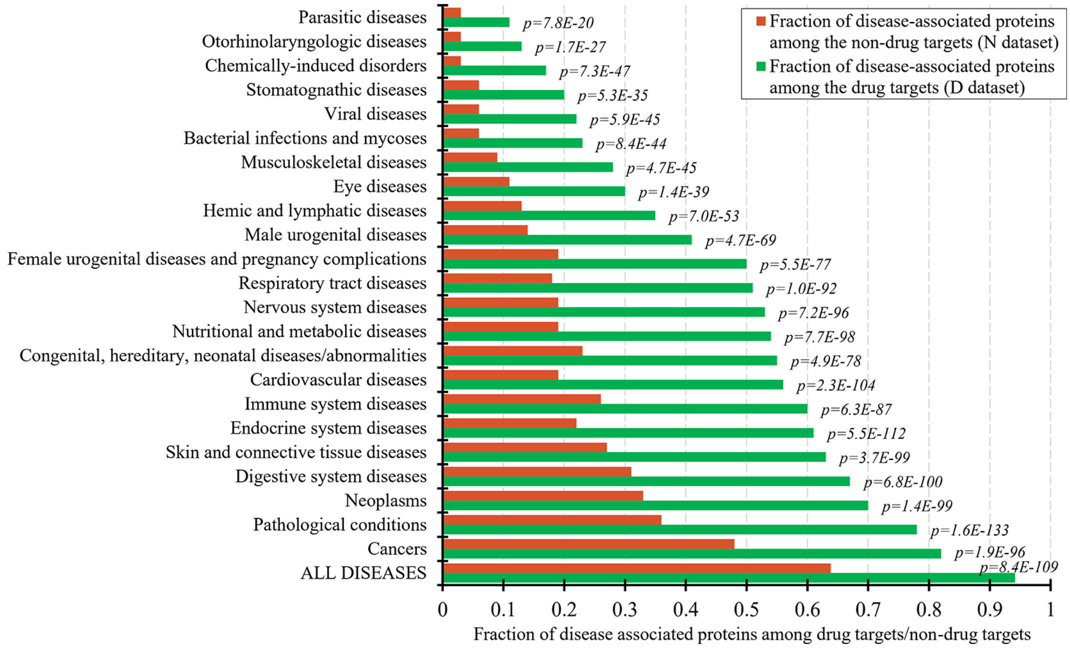

FIGURE 1 | Fraction of drug targets and non-drug targets associated with different classes of diseases. The green and red bars show the fraction of disease associated proteins among the drug targets and non-drug targets for each disease class. The *p*-values quantify the significance of the differences between the two fractions using the Fisher's exact test. The disease classes are sorted by the value of the fraction of the drug targets.

number of the disease associated proteins is significantly higher among the drug targets compared to the non-drug targets. This difference is statistically significant for each of the 23 diseases classes (p-values < 0.0001). About 94% of the drug targets are associated with at least one disease, attesting to the relatively high coverage of these annotations and supporting the fact that the drug targets exert therapeutic effects. The largest fraction of the drug targets (82%) is associated with cancers. To compare, only about 64% of the non-drug targets are disease-associated. The latter suggests that the non-drug targets include both non-druggable proteins (those that lack association with any of the diseases) and possibly druggable proteins (those that are associated with diseases). We note that the use of the diseases associations provides a partial support for their druggability since it does not address the ability of the possibly druggable proteins to interact with drug-like molecules.

Figure 2 analyzes relation between the drug targets, non-drug targets, and disease associations. **Figure 2A** reveals that the disease-associated proteins are likely to be drug targets. About 60% of proteins that are associated with at least one disease are drug targets. The fraction of drug targets increases for the proteins that are associated with more disease. This increase is sharper for a lower number of diseases and plateaus for proteins with about 10 or more disease associations. Therefore, we hypothesize that the non-drug targets with a relatively large number of disease associations can be used as a proxy for possibly druggable proteins. We use the inflection point in **Figure 2A**, which corresponds to proteins with ≥13 disease associations among which 75% are drug targets, to define the set of possibly druggable proteins. **Figure 2B** is a Venn diagram that visualizes overlap between the disease associated proteins (black borders), the drug targets (dataset D; green border), and the non-drug targets (dataset N; red border). We define the set of the non-drug targets that are associated with 13 or more diseases as possibly druggable proteins (Nd dataset; orange area in **Figure 2B**). **Figure 2B** also shows that virtually all drug targets are associated with at least one disease (black border with number of diseases K ≥ 1), while a large portion of the non-drug targets lacks any disease associations (brown area in **Figure 2B**). The latter set of proteins constitutes the set of the non-druggable proteins (Nn dataset).

We test reliability of annotations of the possibly druggable and non-druggable proteins using the 124 human-like drug targets from the D+ dataset that were annotated based on their high sequence similarity to drug targets in other organisms. We found only 4% (5 of the 124) of the human-like drug targets among the 4,869 non-drug targets that are not associated with diseases compared to 67% (83 human-like drug targets) that are among the 4,287 non-drug targets that are associated with 13 or more diseases. The high degree of the latter overlap suggests that the Nd dataset should include a substantial number of druggable proteins. We note that the 4% overlap with the non-drug targets

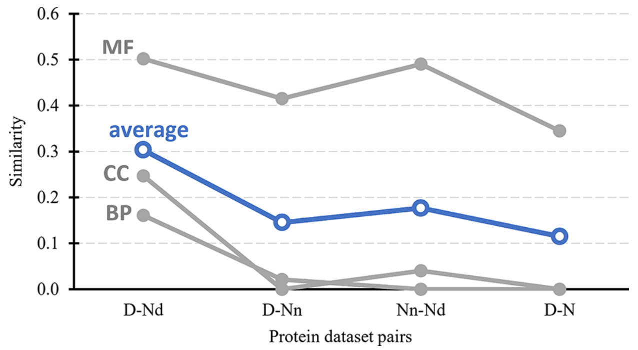

FIGURE 3 | Similarity in cellular processes and subcellular locations between the drug targets (D dataset), possibly druggable proteins (Nd dataset), non-druggable proteins (Nn dataset), and non-drug targets (N dataset). We measure similarity for four pairs of these datasets (D *vs.* Nd, D *vs.* Nn, D *vs.* N, and Nn *vs.* Nd) based on the comparison of the corresponding sets of GO terms associated with these datasets, i.e., GO terms over-represented in a given dataset when compared to the entire human proteome. The GO terms are divided into three categories: MF (molecular functions), BP (biological processes), and CC (cellular components). Similarity was measured with the GOSemSim package (Li et al., 2010). We describe details of these calculations in section Statistical and similarity analyses. The gray markers show the similarity for each GO-term category while the blue markers are the average across the three categories.

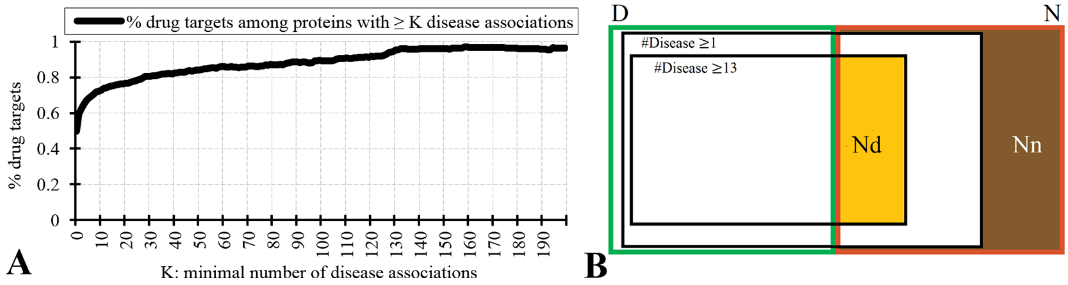

FIGURE 2 | Relation between drug targets, non-drug targets and diseases associations. Panel **A** shows the fraction of the drug targets among proteins associated with a given minimal number of diseases K. Panel **B** is a Venn diagram that visualizes overlap between the disease associated proteins (with K = 1 and K = 13), the drug targets (dataset D; green border), and the non-drug targets (dataset N; red border). Among the non-drug targets we define the Nn dataset of non-druggable proteins (brown area), i.e., the non-drug targets that are not associated with any disease, and the Nd dataset of possibly druggable proteins (orange area), i.e., the non-drug targets that are associated with 13 or more diseases.

that lack diseases associations likely stems from incompleteness of the diseases association data.

Figure 3 further tests the validity of the hypothesis that the Nd and Nn datasets include the possibly druggable and the non-druggable proteins, respectively. It quantifies similarity in the context of cellular functions and subcellular location between the drug targets, possibly druggable proteins, non-druggable proteins, and the non-drug targets. First, we generate a set of GO terms that are associated with each of these datasets, i.e., GO terms over-represented in a given dataset when compared to the human proteome. We perform this analysis separately for each of the three GO terms categories: molecular functions, biological processes, and cellular components; the latter is a proxy for the subcellular location. Next, we calculate similarity between the corresponding sets of dataset-specific GO terms; we describe the details in section Statistical and similarity analyses. The gray lines in **Figure 3** shows the similarity values for each GO term category while the blue lines show the average across the three categories. The left-most set of results reveals that the cellular functions and subcellular location of the drug targets (D dataset) are similar to the possibly druggable proteins (Nd dataset), which aligns with our hypothesis that the Nd dataset in fact includes druggable proteins. The second set of results, which compares the drug targets against the non-druggable proteins (Nn dataset), shows lack of similarity in the biological processes and subcellular locations and modestly reduced levels of similarity in the molecular functions. The corresponding average similarity = 0.145 is lower by a factor of two when compared with the similarity = 0.303 between the drug targets and possibly druggable proteins. The other two sets of results, which compare the possibly druggable against the non-druggable proteins and the drug targets against the non-drug targets, similarly reveal the lack of similarity in the biological processes and subcellular locations, while showing similarity in the molecular functions. The average similarities for these two dataset pairs are low and equal 0.177 and 0.115, respectively, suggesting that the corresponding two pairs of datasets include proteins involved in distinct cellular processes and subcellular locations. To sum up, the above analysis demonstrates that drug targets and the possibly druggable proteins share much higher levels of functional and subcellular location similarity compared to the similarity between possibly druggable proteins, non-druggable proteins, and non-drug targets. This finding, which uses an independent source of information compared to the approach we used to annotate the possibly druggable proteins, supports validity of our annotations of the possibly druggable and the non-druggable proteins.

Comparative Analysis of the Sequence-Derived Structural and Functional Characteristics of the Drug Targets, Possibly Druggable, and Non-Druggable Proteins

Our ability to identify novel druggable proteins relies on the understanding of functional and sequence-derived characteristics that differentiate drug targets from the non-drug targets. We focus specifically on the characteristics that can be quantified from the protein sequence and/or identifier, which allows for a proteome-wide deployment. We compare a broad range of these characteristics between the drug targets, non-drug targets, possibly druggable proteins, and non-druggable proteins. We also investigate differences between the above protein sets and the expanded set of drug targets that includes human and human-like targets (D+ dataset), highly promiscuous drug targets that interact with many drugs (Dh datasets), and drug targets that interact with a low number of drugs (Dl dataset).

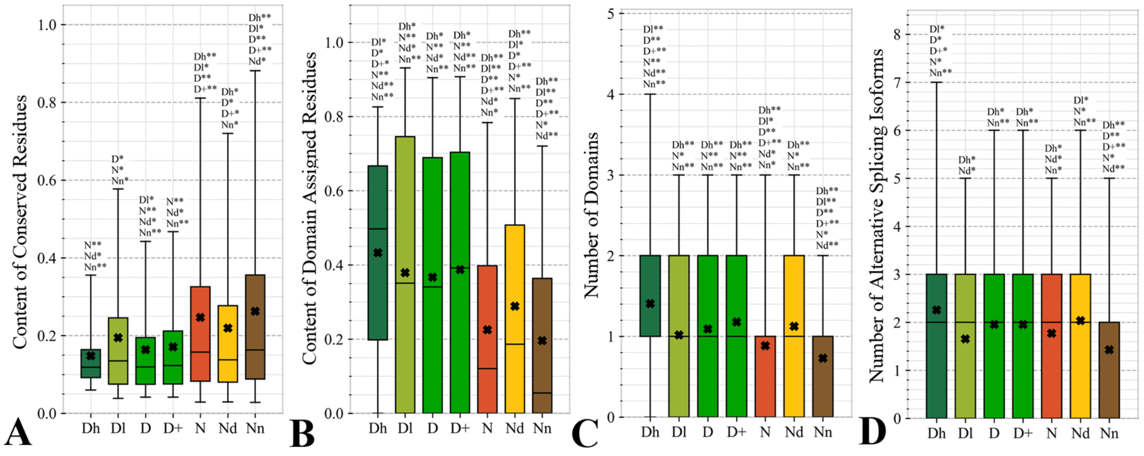

FIGURE 4 | Distributions of the values of the sequence-derived characteristics for the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). (Panels **A**) shows the amount of conserved residues. Panels **B** and **C** focus on the protein domains while Panel **D** quantifies the number of splicing isoforms. The whiskers show the 5 and 95 percentiles, the top and bottom of the box correspond to the first and third quartiles, the middle bar is the median, and the cross marker is the average. The annotation above the whiskers show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section statistical and similarity analyses.

Characteristics Derived From the Protein Sequence

Figure 4 focuses on the characteristics derived directly from the protein sequence, including the residue-level conservation (content of conserved residues in protein chains), number of domains and the content of domain-annotated residues, and the number of the alternative splicing isoforms. **Figure 4A** shows that the drug targets (both D and D+ datasets) have significantly fewer conserved residues than the non-drug targets, possibly druggable proteins and the non-druggable proteins (*p*-value < 0.05). The possibly druggable proteins (orange bars) have significantly lower numbers of conserved residues compared to the non-druggable proteins (brown bars) (*p*-value < 0.05). Moreover, the highly promiscuous drug targets have significantly lower numbers of the conserved amino acids than the non-drug targets and the non-druggable proteins (*p*-value < 0.05), while maintaining similar levels compared to the possibly druggable proteins. Altogether, relatively low numbers of the conserved residues are characteristics for the drug targets and these numbers are also relatively low among the possibly druggable proteins. Interestingly, the residue-level conservation of the residues on the protein surface, where the protein-drug interaction occurs, follows the same pattern (**Figure 5E**). This finding complements prior results that show that drug targets have lower evolutionary rates and higher similarity to orthologous genes (Lv et al., 2016).

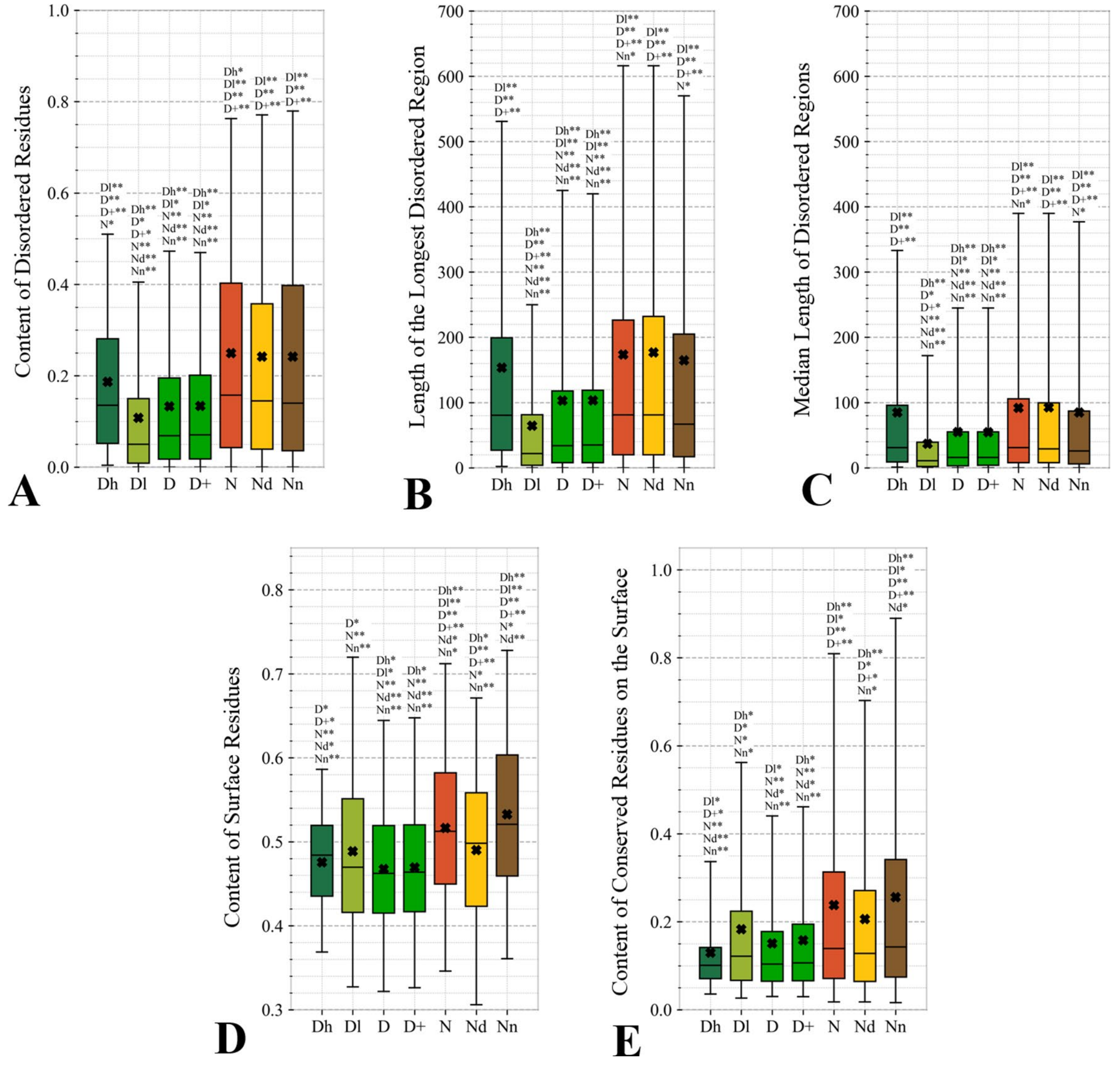

FIGURE 5 | Distributions of the values of the sequence-derived structural characteristics predicted from the protein sequence for the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). Panels **A**, **B**, and **C** quantify the abundance of intrinsic disorder while Panels **D** and **E** quantify the amount of surface and the amount of conserved residues on the surface, respectively. The whiskers show the 5 and 95 percentiles, the top and bottom of the box correspond to the first and third quartiles, the middle bar is the median, and the cross marker is the average. The annotation above the whiskers show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section statistical and similarity analyses.

Figures 4B, C reveal that the drug targets (both D and D+ datasets) have substantially more domains and have larger amounts of domain-annotated residues when compared to the non-druggable proteins (p-value < 0.0001). At the same time, they a similar number domains when contrasted with the possibly druggable proteins. Furthermore, the possibly druggable proteins have significantly higher levels of domain annotations when contrasted against the non-druggable proteins (p-value < 0.0001). The underlying reasons for this enrichment could be two-fold. First, there could be proportionally more multi-domain proteins among the drug targets and the possibly druggable proteins. Consequently, inclusion of a larger number of domains could increase the likelihood that these proteins host at least one druggable domain. However, our result could also mean that these proteins are more studied and understood, and thus their domain annotations are more complete. Moreover, the fact that at least close to half of proteins in all considered datasets have domain annotations, which suggests that they are functionally annotated, suggests that our functional similarity analysis in **Figure 3** should be robust.

The drug targets (both D and D+ datasets) and the possibly druggable proteins have significantly more splicing isoforms compared to the non-druggable proteins (p-value < 0.05) and this increase is even higher for the promiscuous drug targets (p-value < 0.001). This suggests that enrichment in the number of alternative splicing variants could serve as a marker for druggability. The alternative splicing was found to contribute to drug resistance (Siegfried and Karni, 2018; Zhao, 2019), which supports veracity of our result. Interestingly, recent studies suggest that targeting alternative splicing events could lead to therapeutic opportunities (Le et al., 2015; Siegfried and Karni, 2018). Our analysis also reveals that majority of the drug targets and the possibly druggable proteins have multiple isoforms. Thus, gene level analysis of drug targets may not be adequate, considering that these genes would encode multiple proteins.

Overall, we identified three potential sequence-derived markers of druggability. The drug targets and possibly druggable proteins share lower numbers of conserved residues and are more likely to have multiple domains and isoforms when compared to the non-druggable proteins. We also note that the results for the original set of human drug targets (D dataset) are consistent with the results for the expanded set of drug targets (D+ dataset).

Sequence-Derived Structural Properties

This study is the first to analyze two relevant sequence-derived structural characteristics that can be accurately predicted from the protein sequence: intrinsic disorder and solvent accessibility. Proteins with disordered regions are associated with a wide range of human diseases (Uversky et al., 2008; Uversky et al., 2014; Uversky, 2014b; Babu, 2016) while solvent accessibility determines protein surface where the drug-protein interaction happens. We note that while authors in (Kim et al., 2017) computed putative solvent accessibility, they only used it to analyze results concerning enrichment in the PTMs.

Figures 5A–C quantify two key aspects of the disorder: the overall content of disordered residues and the length of disordered regions. Proteins with higher disorder content are functionally distinct from structured proteins while long disordered regions are thought to correspond to disordered protein domains (Tompa et al., 2009; Pentony and Jones, 2010; Peng et al., 2014a). We observe that drug targets (both D and D+ datasets) are significantly less disordered (by a factor of two) and include much shorter disordered regions when compared with the non-drug targets, including both possibly druggable and non-druggable proteins (p-value < 0.001). This is in agreement with a recent study that demonstrates that the current drug targets are biased to exclude disordered proteins (Hu et al., 2016). There are several reasons for this bias. The protein structures are used during the rational drug design process (Gane and Dean, 2000; Lundstrom, 2006; Mavromoustakos et al., 2011; Lounnas et al., 2013) and to gain mechanistic insights into the protein-drug interactions (Pielak et al., 2009; Tan et al., 2013; Christopoulos, 2014) (Altschul et al., 1997; Wang and Samudrala, 2006; Calderone et al., 2013; Orchard et al., 2014; UniProt: the universal protein knowledgebase, 2016). The structures are also indispensable for modeling associated with drug repurposing and repositioning (Moriaud et al., 2011; Ma et al., 2013). This is while proteins with disordered regions are much less likely to have structures (Hu et al., 2018), partly because since they are explicitly avoided in the structural genomics pipeline (Linding et al., 2003; Oldfield et al., 2005; Mizianty et al., 2014). Interestingly, the highly promiscuous drug targets are enriched in disorder when contrasted with the overall set of drug targets and the low promiscuity drug targets (p-value < 0.0001), while their disorder levels are comparable to the possibly druggable proteins. This coincides with the observation that disordered regions are capable of interactions with multiple partners (Oldfield et al., 2008; Hu et al., 2017). Our results suggests that although low disorder amounts are a strong marker for the current drug targets, the set of possibly druggable proteins includes large amounts of disorder. In fact, the disordered proteins may become the key to unlocking a substantial portion of yet to be discovered druggable targets (Uversky, 2012; Hu et al., 2016), especially given their association with numerous human diseases (Uversky et al., 2008; Uversky et al., 2014; Uversky, 2014b; Babu, 2016).

The amount of the putative surface residues for the drug targets (both D and D+ datasets) is significantly smaller that for the non-drug targets, including the possibly druggable and non-druggable proteins (p-value < 0.0001), see **Figure 5D**. This could be driven by the fact that drug targets are often membrane proteins (Yildirim et al., 2007; Rajendran et al., 2010), which means that they have relatively low surface area compared to other proteins. They are also mostly structured proteins (Hu et al., 2016) that are more likely to have globular shape with more buried residues compared to more irregularly shaped/elongated disordered proteins (Peng et al., 2014b; Uversky, 2017). Moreover, presence of disordered regions on the protein surface also leads to an increase of the surface area compared to structured conformations (Wu et al., 2015). Interestingly, the possibly druggable proteins have comparable content of the putative surface residues with the low promiscuity drug targets, which is also significantly smaller when contrasted with the non-druggable proteins (p-value < 0.0001). This again, like in the case of the results in **Figure 4**, shows that the possibly druggable proteins are more similar to drug targets than to the non-druggable proteins. Finally, we observe that the number of conserved residues on the putative surface (**Figure 5E**) maintains the same relation between the different protein sets as the overall number of conserved residues shown in **Figure 4A**, i.e., significantly lower for drug targets (both D and D+ datasets), and lower for the possibly druggable proteins compared to the non-druggable proteins (*p*-value < 0.05).

Topological Features of the Protein-Protein Interaction Networks

Topological features of the PPI networks are among the most studied characteristics of the drug targets (Zhu et al., 2009b; Zhu et al., 2009c; Bull and Doig, 2015; Mitsopoulos et al., 2015; Feng et al., 2017; Kim et al., 2017). A unique aspect of our analysis is that we focus on a set of orthogonal measures, i.e., measures that have low mutual correlations. This offers a more focused and balanced analysis given the high degree of similarity between many of these measures. **Figure 6** reveals that the entire set of four measures of centrality has significantly higher values for the drug targets (both D and D+ datasets) compared to the non-druggable proteins (*p*-value < 0.0001). Our results are in line with several prior studies that correspondingly show that drug targets have more connected and denser local network neighborhoods (Zhu et al., 2009b; Zhu et al., 2009c; Mitsopoulos et al., 2015; Lv et al., 2016). This finding suggests that drug targets are possibly more relevant biologically or are at a higher point of control and thus can better modify physiology, making them better therapeutic targets. The novel element in our study is that we find that all considered network centrality measures for the possibly druggable are even higher than for the drug targets (orange *vs.* green bars in **Figure 6**; *p*-value < 0.05). Consequently, they are also significantly higher than for the non-druggable proteins (orange *vs.* brown bars in **Figure 6**; *p*-value < 0.0001). Thus, our study suggests that these measures can be used as markers of druggability.

Figure 7 analyzes the abundance of the PPI network hubs among the drug targets, possibly druggable and non-druggable proteins. Approximately 17% of the drug targets (for both D and D+ datasets) are hubs and this rate is significantly higher than the 12% rate for the non-drug targets (green *vs.* red bars; *p*-value < 0.0001). Similarly large difference was observed in (Mitsopoulos et al., 2015). Our study reveals additional important details. We observe

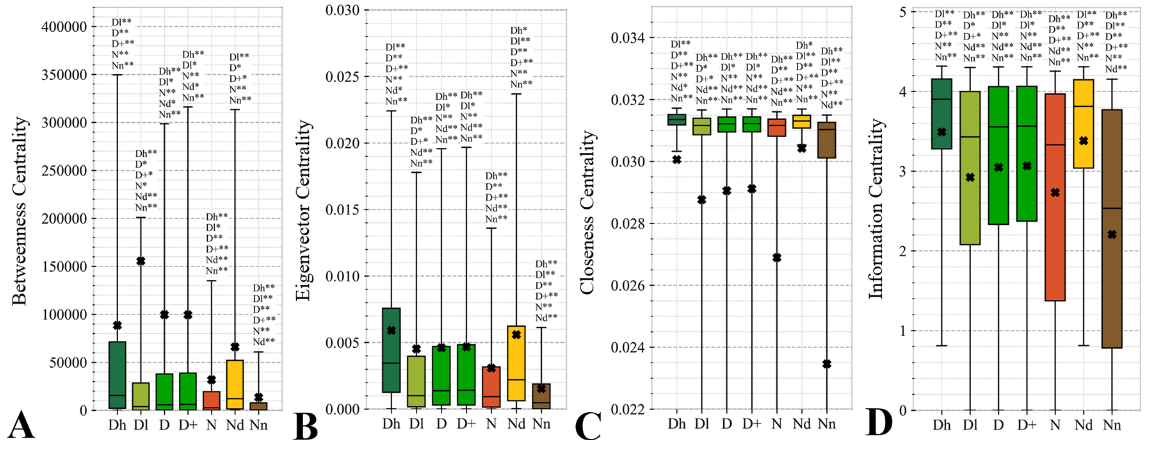

FIGURE 6 | Distributions of the values of the selected orthogonal PPI network properties for the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). Panels A, B, C, and D concern the betweenness centrality, eigenvector centrality, closeness centrality, and information centrality measures, respectively. The whiskers show the 5 and 95 percentiles, the top and bottom of the box correspond to the first and third quartiles, the middle bar is the median, and the cross marker is the average. The annotation above the whiskers show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section statistical and similarity analyses.

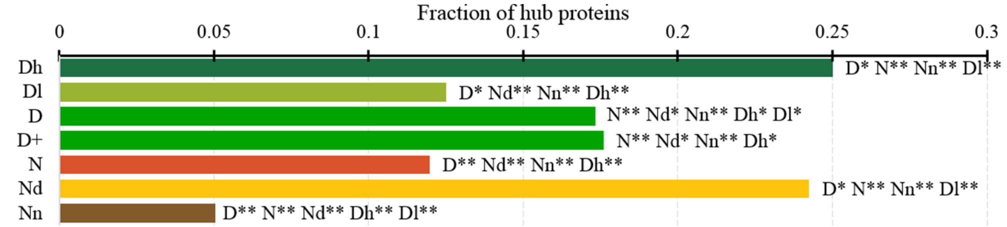

FIGURE 7 | Fraction of hub proteins among the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). The annotation next to the bars show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section statistical and similarity analyses.

that the rate of hubs is very high among the highly promiscuous drug targets (25%) and the possibly druggable proteins (24%), and these rates are significantly higher than the 12% rate for the non-drug targets (p-value < 0.0001) and the 5% rate for the non-druggable proteins (p-value < 0.0001). This suggests that high connectivity in the PPI network is a strong marker for druggability.

Functions and Subcellular Locations of Drug Targets and Possibly Druggable Proteins

Several studies analyzed cellular functions and subcellular locations of the drug targets (Lauss et al., 2007; Bakheet and Doig, 2009; Wang et al., 2013b). The green bars in **Figure 8** provide a list of significantly enriched functions and locations for our set of drug targets. Our results indicate that most of the drug targets are enzymes, including kinases and oxidoreductases, followed by substatial numbers of channels, and in particular ion channels. They are often involved in binding, signalling, regulation, and transport. These finding are in close agreement with the results in (Bakheet and Doig, 2009). **Figure 8** also shows that drug targets are primarily found in membranes, with a large numbers also found in the cytoplasm and the intracellular space. Consistent results are found in (Bakheet and Doig, 2009; Wang et al., 2013b), and these subcellular locations also agree with the observation that membrane proteins are the prime targets for the development of therapeutics (Yildirim et al., 2007; Rajendran et al., 2010).

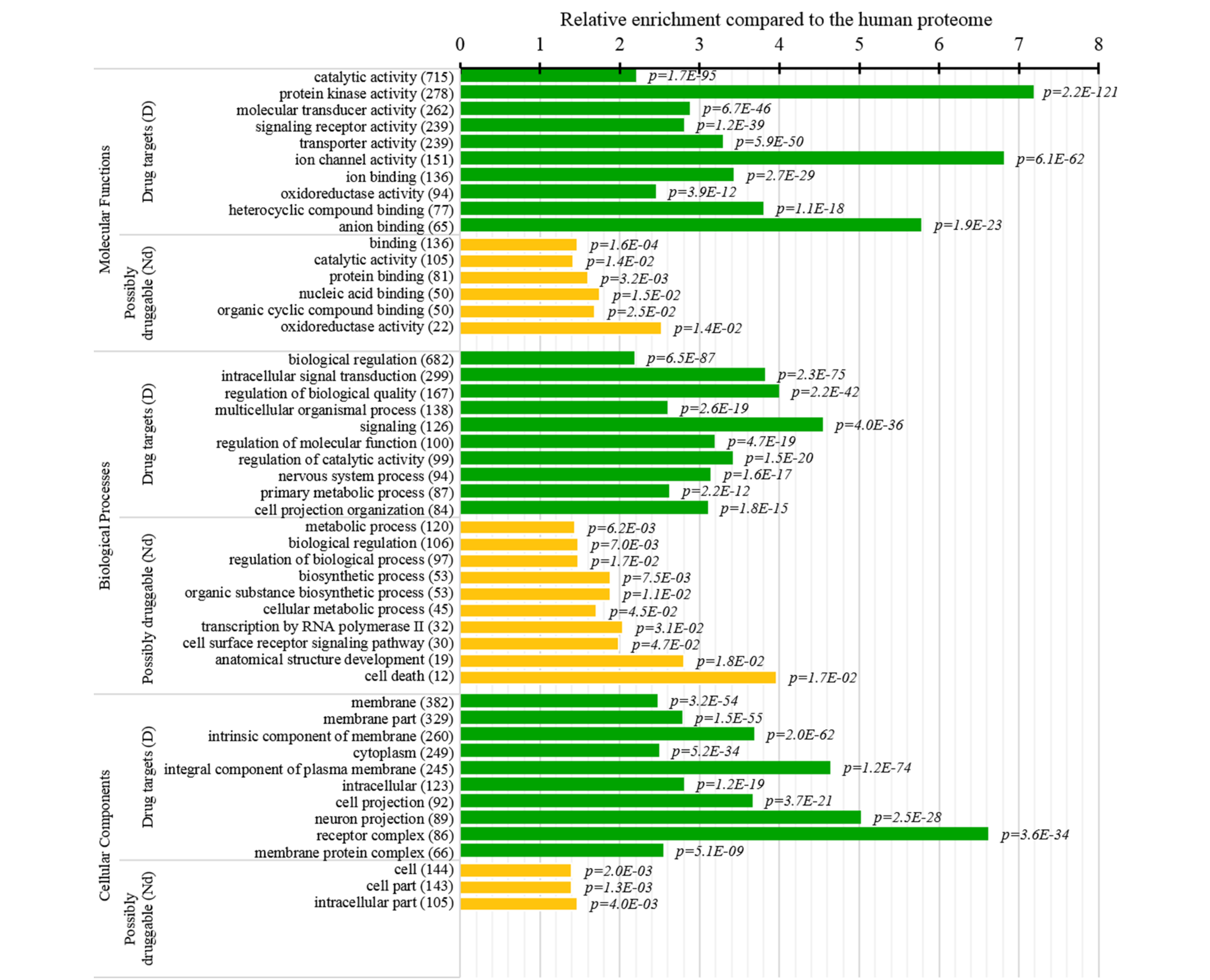

FIGURE 8 | Molecular functions, processes, and subcellular locations that are enriched among the drug targets (D dataset) and the possibly druggable proteins (Nd dataset). We show the top 10 (with the highest counts) over-represented/significantly enriched GO terms for the drug targets (green bars) and the possibly druggable proteins (orange bars). The bars quantify the ratios of enrichment relative to the human proteome and the corresponding p-values are shown on the right. GO terms are identified on the left, including their names and the number of the correspnding proteins in the given dataset. We explain calculation of statistical tests in section statistical and similarity analyses.

This study is the first to perform this type of analysis for the possibly druggable proteins (orange bars in **Figure 8**). Our analysis suggests that the possibly druggable proteins share functional similarities with the drug targets. They are similarly involved in the catalysis, signaling, and binding. However, the possibly druggable proteins tend to bind proteins and nucleic acids, instead of anions and ions which are the main partners for the drug targets. Moreover, the possibly druggable proteins are often involved in the metabolic and biosynthesis processes, and in the cell death cycle. The preference for the protein-protein and protein-nucleic acids binding and the cell death cycle involvement are supported by their significant enrichment in the intrinsic disorder (compared to the drug targets, see **Figures 5A, B**), and the fact that disordered regions are known to facilitate these types of functions (Vuzman and Levy, 2012; Uversky et al., 2013; Fuxreiter et al., 2014; Peng et al., 2015; Basu and Bahadur, 2016; Wang et al., 2016b; Hu et al., 2017; Srivastava et al., 2018). We further investigate this in **Figure 9** that analyzes the differences in the content of the putative disordered protein-protein binding regions. These results confirm the enrichment in the corresponding functional annotations for the possibly druggable proteins. The possibly druggable proteins include a substantial amount of the disordered protein-binding regions, on average about 14% of residues. Moreover, the drug targets (both D and D+ datasets) are significantly depleted in these protein-binding regions (on average only 7% of residues) when compared with the possibly druggable proteins (p-value < 0.0001). Interestingly, **Figure 8** also reveals that the possibly druggable proteins are localized across the cell and they do not have a specifically associated subcellular location, unlike the drug targets that are found mostly in the membranes and cytoplasm. Overall, our empirical analysis provides new insights into the cellular functions and subcellular locations of the druggable proteins.

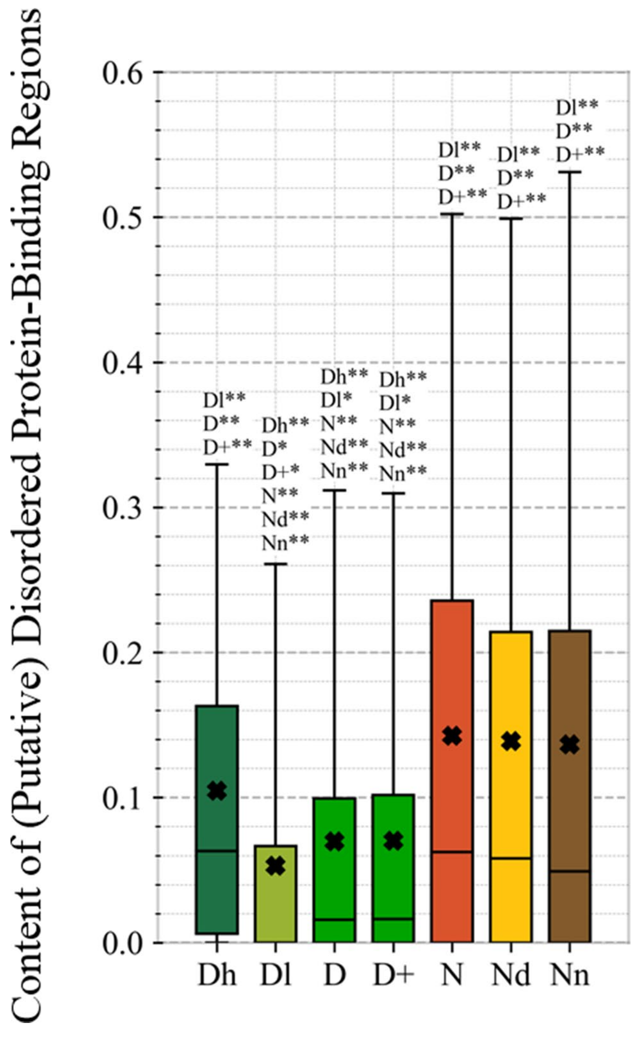

FIGURE 9 | Content of putative protein binding regions in the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). The annotation next to the bars show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section statistical and similarity analyses.

SUMMARY AND CONCLUSIONS

Recent research approximates that the druggable human proteome has about 4,500 proteins (Finan et al., 2017), while there are about 1,600 current drug targets (1,750 drug targets if we include proteins that share high sequence similarity to drug targets that were annotated in other organisms). Annotation of the remaining druggable human proteins would facilitate development and screening of drugs, drug repurposing and repositioning, understanding and mitigation of drug side-effects, and prediction of drug–protein interactions. We contrast the drug targets against the possibly druggable and non-druggable proteins to identify markers that could be used to identify novel druggable proteins. This is in contrast to the prior studies that compare drug targets against non-drug targets (Zheng et al., 2006; Lauss et al., 2007; Bakheet and Doig, 2009; Zhu et al., 2009b; Zhu et al., 2009c; Bull and Doig, 2015; Mitsopoulos et al., 2015; Feng et al., 2017; Kim et al., 2017), thus producing markers that describe current drug target and which implicitly exclude the druggable proteins that are included in the non-drug target set. We annotate the possibly druggable and non-druggable proteins based on the presence and promiscuity of disease associations, and we validate these annotations *via* functional similarity analysis.

We cover a wide range of sequence-derived characteristics to define these markers. These characteristics can be computed across the entire human proteome, allowing for a complete sweep of all potential candidate proteins. We investigate several important characteristic that were missed in the past studies including putative intrinsic disorder, residue-level conservation, presence and number of alternative splicing isoforms, inclusion of domains, and putative solvent accessibility (surface area), as well as the key features from the prior works, such as the topological features of PPIs, cellular functions and subcellular locations. **Figure 10** summarizes the results. It shows the difference in the values of the key markers when comparing the possibly druggable proteins (in orange), the non-druggable proteins (in brown), all non-drug targets (in red), and the expanded set of human and human-like drug targets (in light

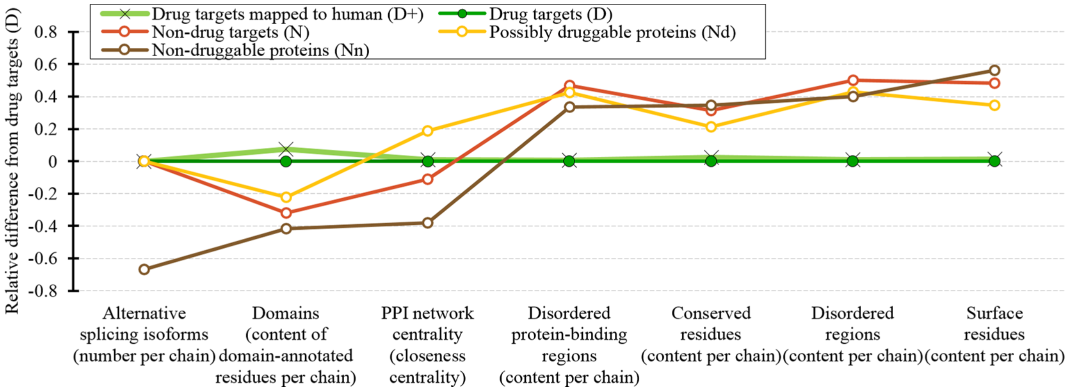

FIGURE 10 | Overview of the sequence-derived markers for the drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). The y-axis quantifies the relative difference of the values of a given protein set X compared to the values of the drug targets (D) set defined as: [median(X)-median(D))/IQR(D), where IQR means the interquartile range. The markers are sorted in the ascending order by the difference for the non-druggable proteins (in brown).

green) against the human drug targets (in dark green). We observe that the possibly druggable proteins are significantly more similar to the drug targets than the non-druggable proteins for majority of the markers. These markers include high abundance of alternative splicing isoforms, relatively large number of domains, higher degree of centrality in the corresponding PPI network (and correspondingly much higher rate of hubs), lower number of conserved residues, and lower number of residues on the putative (sequence-derived) surface. Thus, these factors could serve as high-quality markers for druggability. Results and discussion discusses these findings in the context of the current literature. Moreover, **Figure 10** shows that drug targets (both D and D+ datasets) have significantly depleted levels of intrinsic disorder and intrinsically disordered protein-binding regions when compared with the much higher and comparable levels among the possibly druggable and non-druggable proteins. This suggests that the high levels of disorder combined with the presence of the abovementioned markers should be used together to effectively enlarge the current collection of drug targets. This is in accord with several recent studies that postulate inclusion of the disorder-enriched proteins into the set of druggable proteins (Cuchillo and Michel, 2012; Uversky, 2012; Chen and Tou, 2013; Joshi and Vendruscolo, 2015; Ambadipudi and Zweckstetter, 2016; Hu et al., 2016; Yu et al., 2016).

Our analysis also shows that the possibly druggable proteins are functionally similar to the drug targets, being involved in the catalysis, signaling, and binding. The main difference is that the possibly druggable proteins target interactions with proteins and nucleic acids, unlike the current drug targets that favor interactions with anions and ions. **Figure 10** points to the high amount of the disordered protein-binding regions for the possibly druggable proteins compared to the drug targets, which is in concert with the disordered nature of the druggable proteins. This is in agreement with the literature that shows that disordered regions often facilitate PPIs (Mohan et al., 2006; Vacic et al., 2007; Fuxreiter et al., 2014; Yan et al., 2016; Hu et al., 2017). Finally, we show that the possibly druggable proteins are involved in the metabolic and biosynthesis processes and that they are localized across the cell, without a preference for specific subcellular locations. This is unlike the current drug targets that are located primarily in the membranes.

To sum up, our empirical analysis has led us to formulate several markers that may help with identifying novel druggable human proteins and has produced interesting insights into the cellular functions and subcellular locations of potentially druggable proteins.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

LK conceptualized the study. LK and ML designed the study. SG organized the source databases. SG and XL performed acquisition of data. SG and LK organized and performed statistical analysis. All authors organized, analyzed and interpreted the results. LK and SG wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version, and provided approval for publication of the content.

FUNDING

This research was supported in part by the Robert J. Mattauch Endowment funds to LK, the National Natural Science Foundation of China (No. 61832019), and the 111 Project (No. B18059).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01075/full#supplementary-material

REFERENCES

Altschul, S. F., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi: 10.1093/nar/25.17.3389

Amirkhani, A., et al. (2018). Prediction of DNA-binding residues in local segments of protein sequences with Fuzzy Cognitive Maps. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2018.2890261

Babu, M. M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc Trans.* 44 (5), 1185–1200. doi: 10.1042/BST20160172

Bakheet, T. M., and Doig, A. J. (2009). Properties and identification of human protein drug targets. *Bioinformatics* 25 (4), 451–457. doi: 10.1093/bioinformatics/btp002

Basu, S., and Bahadur, R. P. (2016). A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell Mol. Life Sci.* 73 (21), 4075–4084. doi: 10.1007/s00018-016-2283-1

Batada, N. N., et al. (2006). Stratus not altocumulus: a new view of the yeast protein interaction network. *PloS Biol.* 4, 1720–1731. doi: 10.1371/journal.pbio.0040317

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *J. Acoust. Soc Am.* 22, 725–730. doi: 10.1121/1.1906679

Berman, H. M., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi: 10.1093/nar/28.1.235

Bonacich, P. (1987). Power and centrality: A family of measures. *Am. J. Sociol.* 92 (5), 1170–1182. doi: 10.1086/228631

Bull, S. C., and Doig, A. J. (2015). Properties of protein drug target classes. *PloS One* 10 (3), e0117955. doi: 10.1371/journal.pone.0117955

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). Mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* 10 (8), 690. doi: 10.1038/nmeth.2561

Chen, C. Y., and Tou, W. I. (2013). How to design a drug for the disordered proteins? *Drug Discov. Today* 18 (19-20), 910–915. doi: 10.1016/j.drudis.2013.04.008

Cheng, Y., et al. (2006). Rational drug design *via* intrinsically disordered protein. *Trends Biotechnol.* 24 (10), 435–442. doi: 10.1016/j.tibtech.2006.07.005

Chong, C. R., and Sullivan, D. J. (2007). New uses for old drugs. *Nature* 448 (7154), 645–646. doi: 10.1038/448645a

Christopoulos, A. (2014). Advances in G protein-coupled receptor allostery: from function to structure. *Mol. Pharmacol.* 86 (5), 463–478. doi: 10.1124/mol.114.094342

Cimermancic, P., et al. (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* 428 (4), 709–719. doi: 10.1016/j.jmb.2016.01.029

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32 (suppl_1), D258–D261. doi: 10.1093/nar/gkh036

Cuchillo, R., and Michel, J. (2012). Mechanisms of small-molecule binding to intrinsically disordered proteins. *Biochem. Soc. Trans.* 40 (5), 1004–1008. doi: 10.1042/BST20120086

Dalkas, G. A., et al. (2013). State-of-the-art technology in modern computer-aided drug design. *Briefings Bioinform.* 14 (6), 745–752. doi: 10.1093/bib/bbs063

Dolan, P. T., et al. (2015). Intrinsic disorder mediates hepatitis C virus core-host cell protein interactions. *Protein Sci.* 24 (2), 221–235. doi: 10.1002/pro.2608

Dosztanyi, Z. (2018). Prediction of protein disorder based on IUPred. *Protein Sci.* 27 (1), 331–340. doi: 10.1002/pro.3334

Dosztányi, Z., et al. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21 (16), 3433–3434. doi: 10.1093/bioinformatics/bti541

Dosztányi, Z., et al. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* 5, 2985–2995. doi: 10.1021/pr060171o

Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25 (20), 2745–2746. doi: 10.1093/bioinformatics/btp518

Dunker, A. K., and Uversky, V. N. (2010). Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Curr. Opin. Pharmacol.* 10 (6), 782–788. doi: 10.1016/j.coph.2010.09.005

Dunker, A. K., et al. (2013). What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins* 1 (1), e24157. doi: 10.4161/idp.24157

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6 (3), 197–208. doi: 10.1038/nrm1589

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E.* 71 (5), 056103. doi: 10.1103/PhysRevE.71.056103

Ezzat, A., Wu, M., Li, X. -L., and Kwoh, C. -K. (2018). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics* 20 (4), 1337–1357. doi: 10.1093/bib/bby002

Fan, X., et al. (2014). The intrinsic disorder status of the human hepatitis C virus proteome. *Mol. Biosyst.* 10 (6), 1345–1363. doi: 10.1039/C4MB00027G

Faraggi, E., Zhou, Y., and Kloczkowski, A. (2014). Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins: Struct. Funct. Bioinform.* 82 (11), 3170–3176. doi: 10.1002/prot.24682

Feng, Y., Wang, Q., and Wang, T. (2017). Drug Target Protein-Protein Interaction Networks: A Systematic Perspective. *Biomed. Res. Int.* 2017, 1289259. doi: 10.1155/2017/1289259

Finan, C., Gaulton, A., Kruger, F. A., Lumbers, R. T., Shah, T., Engmann, J., et al. (2017). The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9 (383). doi: 10.1126/scitranslmed.aag1166

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40, 35. doi: 10.2307/3033543

Fuxreiter, M., et al. (2014). Disordered proteinaceous machines. *Chem. Rev.* 114 (13), 6806–6843. doi: 10.1021/cr4007329

Gane, P. J., and Dean, P. M. (2000). Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* 10 (4), 401–404. doi: 10.1016/S0959-440X(00)00105-6

Gaulton, A., et al. (2016). The ChEMBL database in 2017. *Nucleic Acids Res.* 45 (D1), D945–D954. doi: 10.1093/nar/gkw1074

Gutiérrez-Sacristán, A., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45 (D1), D833–D839. doi: 10.1093/nar/gkw943

Habchi, J., et al. (2014). Introducing Protein Intrinsic Disorder. *Chem. Rev.* 114 (13), 6561–6588. doi: 10.1021/cr400514h

Hambly, K., et al. (2006). Interrogating the druggable genome with structural informatics. *Mol. Divers.* 10 (3), 273–281. doi: 10.1007/s11030-006-9035-3

Han, J. D. J., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi: 10.1038/nature02555

Hao, M., Bryant, S. H., and Wang, Y. (2019). Open-source chemogenomic data-driven algorithms for predicting drug–target interactions. *Brief. Bioinform.* 20 (4), 1465–1474. doi: 10.1093/bib/bby010

Harding, S. D., et al. (2017). The iuphar/BPS Guide to pharmacology in 2018: updates and expansion to encompass the new guide to immunopharmacology. *Nucleic Acids Res.* 46 (D1), D1091–D1106. doi: 10.1093/nar/gkx1121

Haupt, V. J., and Schroeder, M. (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings Bioinform.* 12 (4), 312–326. doi: 10.1093/bib/bbr011

Hopkins, A. L., and Groom, C. R. (2002). The druggable genome. *Nat. Rev. Drug Discovery* 1 (9), 727–730. doi: 10.1038/nrd892

HOWELL, M., et al. (2012). Not that rigid midgets and not so flexible giants: on the abundance and roles of intrinsic disorder in short and long proteins. *J. Biol. Syst.* 20 (04), 471–511. doi: 10.1142/S0218339012400086

Hu, Y., and Bajorath, J. (2013). Compound promiscuity: what can we learn from current data? *Drug Discovery Today* 18 (13-14), 644–650. doi: 10.1016/j.drudis.2013.03.002

Hu, G., et al. (2014). Human structural proteome-wide characterization of Cyclosporine A targets. *Bioinformatics* 30 (24), 3561–3566. doi: 10.1093/bioinformatics/btu581

Hu, G., et al. (2016). Untapped Potential of Disordered Proteins in Current Druggable Human Proteome. *Curr. Drug Targets* 17 (10), 1198–1205. doi: 10.2174/1389450116666150722141119

Hu, G., Wu, Z., Uversky, V., and Kurgan, L. (2017). Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int. J. Mol. Sci.* 18 (12). doi: 10.3390/ijms18122761

Hu, G., Wang, K., Song, J., Uversky, V. N., and Kurgan, L. (2018). Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between

Structural Darkness, Intrinsic Disorder, and Crystallization Propensity. Proteomics 18 (21–22), 1800243. doi: 10.1002/pmic.201800243

Imming, P., Sinning, C., and Meyer, A. (2007). Drugs, their targets and the nature and number of drug targets (vol 5, 2006). *Nat. Rev. Drug Discovery* 6 (2), 126–126, pg 821. doi: 10.1038/nrd2132

Jeong, H., Mason, S. P., Barabási, A. -L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411 (6833), 41–42. doi: 10.1038/35075138

Joosten, R. P., Te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., et al. (2010). A series of PDB related databases for everyday needs. *Nucleic Acids Res.* 39 (suppl_1), D411–D419. doi: 10.1093/nar/gkq1105

Joshi, P., and Vendruscolo, M. (2015). Druggability of intrinsically disordered proteins. *Adv. Exp. Med. Biol.* 870, 383–400. doi: 10.1007/978-3-319-20164-1_13

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Orig. Res. Biomol.* 22 (12), 2577–2637. doi: 10.1002/bip.360221211

Katuwawala, A., Peng, Z., Yang, J., and Kurgan, L. (2019). Computational Prediction of MoRFs, short disorder-to-order transitioning protein binding regions. *Comput. Struct. Biotechnol. J.* 17, 454–462. doi: 10.1016/j.csbj.2019.03.013

Keller, T. H., Pichota, A., and Yin, Z. (2006). A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* 10 (4), 357–361. doi: 10.1016/j.cbpa.2006.06.014

Kim, B., Jo, J., Han, J., Park, C., and Lee, H. (2017). In silico re-identification of properties of drug target proteins. *BMC Bioinform.* 18 (Suppl 7), 248. doi: 10.1186/s12859-017-1639-3

Kjaergaard, M., and Kragelund, B. B. (2017). Functions of intrinsic disorder in transmembrane proteins. *Cell. Mol. Life Sci.* 74 (17), 3205–3224. doi: 10.1007/s00018-017-2562-5

Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L. J., Gross, C., Gavin, A. C., et al. (2013). Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.* 9, 663. doi: 10.1038/msb.2013.10

Kulkarni, P., and Uversky, V. N. (2018). Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome. *Proteomics* 18, 21–22. doi: 10.1002/pmic.201800061

Launay, G., et al. (2015). MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Res.* 43 (Database issue), D321–D327. doi: 10.1093/nar/gku1091

Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2007). Characterization of the drugged human genome. *Pharmacogenomics* 8 (8), 1063–1073. doi: 10.2217/14622416.8.8.1063

Lauss, M., Kriegner, A., Vierlinger, K., and Noehammer, C. (2007). Characterization of the drugged human genome. *Pharmacogenomics* 8, 1063–1073.

Le, K. Q., Prabhakar, B. S., Hong, W. J., and Li, L. C. (2015). Alternative splicing as a biomarker and potential target for drug discovery. *Acta Pharmacol. Sin.* 36 (10), 1212–1218. doi: 10.1038/aps.2015.43

Li, F., Yu, G., Wang, S., Bo, X., Wu, Y., and Qin, Y. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26 (7), 976–978. doi: 10.1093/bioinformatics/btq064

Li, M., Wang, J., Chen, X., Wang, H., and Pan, Y. (2011). A local average connectivity-based method for identifying essential proteins from the network level. *Comput. Biol. Chem.* 35 (3), 143–150. doi: 10.1016/j.compbiolchem.2011.04.002

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Brief. Bioinform.* 17 (1), 2–12. doi: 10.1093/bib/bbv020

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., et al. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40 (Database issue), D857–D861. doi: 10.1093/nar/gkr930

Lieutaud, P., Ferron, F., Uversky, A. V., Kurgan, L., Uversky, V. N., and Longhi, S. (2016). How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord. Proteins* 4 (1), e1259708. doi: 10.1080/21690707.2016.1259708

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure* 11 (11), 1453–1459. doi: 10.1016/j.str.2003.10.002

Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006). Intrinsic disorder in transcription factors. *Biochemistry* 45 (22), 6873–6888. doi: 10.1021/bi0602718

Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J., and Green, J.R. (2017). A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics* 19(5), 878-892. doi: 10.1093/bib/bbx017

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486 (7403), 361–367. doi: 10.1038/nature11159

Lounnas, V., Ritschel, T., Kelder, J., Mcguire, R., Bywater, R. P., and Foloppe, N. (2013). Current progress in structure-based rational drug design marks a new mindset in drug discovery. *Comput. Struct. Biotechnol. J.* 5, e201302011. doi: 10.5936/csbj.201302011

Lundstrom, K. (2006). Structural genomics: the ultimate approach for rational drug design. *Mol. Biotechnol.* 34 (2), 205–212. doi: 10.1385/MB:34:2:205

Lv, W., Xu, Y., Guo, Y., Yu, Z., Feng, G., Liu, P., Luan, M., et al. (2016). The drug target genes show higher evolutionary conservation than non-target genes. *Oncotarget* 7 (4), 4961–4971. doi: 10.18632/oncotarget.6755

Ma, D. L., Chan, D. S., and Leung, C. H. (2013). Drug repositioning by structure-based virtual screening. *Chem. Soc. Rev.* 42 (5), 2130–2141. doi: 10.1039/c2cs35357a

Makley, L. N., and Gestwicki, J. E. (2013). Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. *Chem. Biol. Drug Des.* 81 (1), 22–32. doi: 10.1111/cbdd.12066

Mavromoustakos, T., Durdagi, S., Koukoulitsa, C., Simcic, M., Papadopoulos, M. G., Hodoscek, M., et al. (2011). Strategies in the rational drug design. *Curr. Med. Chem.* 18 (17), 2517–2530. doi: 10.2174/092986711795933731

Meng, F., Murray, G. F., Kurgan, L., and Donahue, H. J. (2018). High-throughput prediction of disordered moonlighting regions in protein sequences. *Proteins* 86 (10), 1097–1110. doi: 10.1002/prot.25590

Meng, F., Na, I., Kurgan, L., and Uversky, V. (2015b). Compartmentalization and functionality of nuclear disorder: intrinsic disorder and protein-protein interactions in intra-nuclear compartments. *Int. J. Mol. Sci.* 17 (1). doi: 10.3390/ijms17010024

Meng, F., et al. (2015a). Unstructural biology of the Dengue virus proteins. *FEBS J.* 282 (17), 3368–3394. doi: 10.1111/febs.13349

Meng, F., Uversky, V., and Kurgan, L. (2017a). Computational prediction of intrinsic disorder in proteins. *Curr. Protoc. Protein Sci.* 88, 2 16 1–2 16 14. doi: 10.1002/cpps.28

Meng, F., Uversky, V. N., and Kurgan, L. (2017b). Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell Mol. Life Sci.* 74 (17), 3069–3090. doi: 10.1007/s00018-017-2555-4

Meng, F., Badierah, R. A., Almehdar, H. A., Redwan, E. M., Kurgan, L., and Uversky, V. N. (2018). Functional and structural characterization of osteocytic MLO-Y4 cell proteins encoded by genes differentially expressed in response to mechanical signals in vitro. *Sci. Rep.* 8 (1), 6716. doi: 10.1038/s41598-018-25113-4

Mitsopoulos, C., Schierz, A. C., Workman, P., and Al-Lazikani, B. (2015). Distinctive behaviors of druggable proteins in cellular networks. *PloS Comput. Biol.* 11 (12), e1004597. doi: 10.1371/journal.pcbi.1004597

Mizianty, M. J., Fan, X., Yan, J., Chalmers, E., Woloschuk, C., Joachimiak, A., et al. (2014). Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr. D. Biol. Crystallogr.* 70 (Pt 11), 2781–2793. doi: 10.1107/S1399004714019427

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., et al. (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 362 (5), 1043–1059. doi: 10.1016/j.jmb.2006.07.087

Moriaud, F., Richard, S. B., Adcock, S. A., Chanas-Martin, L., Surgand, J. S., Ben Jelloul, M., et al. (2011). Identify drug repurposing candidates by mining the protein data bank. *Brief Bioinform.* 12 (4), 336–340. doi: 10.1093/bib/bbr017

Muruganujan, A., et al. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 47 (D1), D419–D426. doi: 10.1093/nar/gky1038

Na, I., et al. (2016). Autophagy-related intrinsically disordered proteins in intra-nuclear compartments. *Mol. Biosyst.* 12 (9), 2798–2817. doi: 10.1039/C6MB00069J

Núñez, S., Venhorst, J., and Kruse, C. G. (2012). Target–drug interactions: first principles and their application to drug discovery. *Drug Discovery Today* 17 (1), 10–22. doi: 10.1016/j.drudis.2011.06.013

Oates, M. E., et al. (2013). D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.* 41 (Database issue), D508–D516. doi: 10.1093/nar/gks1226

Oldfield, C. J., et al. (2005). Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins* 59 (3), 444–453. doi: 10.1002/prot.20446

Oldfield, C. J., et al. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9 Suppl 1, S1. doi: 10.1186/1471-2164-9-S1-S1

Oprea, T. I., and Mestres, J. (2012). Drug repurposing: far beyond new targets for old drugs. *AAPS J.* 14 (4), 759–763. doi: 10.1208/s12248-012-9390-1

Orchard, S., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42 (Database issue), D358–D363. doi: 10.1093/nar/gkt1115

Oughtred, R., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47 (D1), D529–D541. doi: 10.1093/nar/gky1079

Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). Opinion - How many drug targets are there? *Nat. Rev. Drug Discovery* 5 (12), 993–996. doi: 10.1038/nrd2199

Ozdemir, E. S., et al. (2019). Methods for Discovering and Targeting Druggable Protein-Protein Interfaces and Their Application to Repurposing. *Methods Mol. Biol.* 1903, 1–21. doi: 10.1007/978-1-4939-8955-3_1

Patil, A., Kinoshita, K., and Nakamura, H. (2010). Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. *Protein Sci.* 19 (8), 1461–1468. doi: 10.1002/pro.425

Peng, Z. L., and Kurgan, L. (2012). Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.* 13 (1), 6–18. doi: 10.2174/138920312799277938

Peng, Z., et al. (2012). More than just tails: intrinsic disorder in histone proteins. *Mol. Biosyst.* 8 (7), 1886–1901. doi: 10.1039/c2mb25102g

Peng, Z., et al. (2013). Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. *Cell Death Differ.* 20 (9), 1257–1267. doi: 10.1038/cdd.2013.65

Peng, Z., et al. (2014b). A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol. Life Sci.* 71 (8), 1477–1504. doi: 10.1007/s00018-013-1446-6

Peng, Z., Mizianty, M. J., and Kurgan, L. (2014a). Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* 82 (1), 145–158. doi: 10.1002/prot.24348

Peng, Z., et al. (2015). Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol. Life Sci.* 72 (1), 137–151. doi: 10.1007/s00018-014-1661-9

Pentony, M. M., and Jones, D. T. (2010). Modularity of intrinsic disorder in the human proteome. *Proteins* 78 (1), 212–221. doi: 10.1002/prot.22504

Pielak, R. M., Schnell, J. R., and Chou, J. J. (2009). Mechanism of drug inhibition and drug resistance of influenza A M2 channel. *Proc. Natl. Acad. Sci. U.S.A.* 106 (18), 7379–7384. doi: 10.1073/pnas.0902548106

Radusky, L., Defelipe, L.A., Lanzarotti, E., Luque, J., Barril, X., Marti, et al. (2014). TuberQ: a Mycobacterium tuberculosis protein druggability database. *Database-the Journal of Biological Databases and Curation*. doi: 10.1093/database/bau035

Rajendran, L., Knolker, H. J., and Simons, K. (2010). Subcellular targeting strategies for drug design and delivery. *Nat. Rev. Drug Discov.* 9 (1), 29–42. doi: 10.1038/nrd2897

Rask-Andersen, M., Masuram, S., and Schioth, H. B. (2014). The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu. Rev. Pharmacol. Toxicol.* 54, 9–26. doi: 10.1146/annurev-pharmtox-011613-135943

Roider, H. G., et al. (2014). Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinform.* 15 (1), 68. doi: 10.1186/1471-2105-15-68

Russ, A. P., and Lampel, S. (2005). The druggable genome: an update. *Drug Discovery Today* 10 (23–24), 1607–1610. doi: 10.1016/S1359-6446(05)03666-4

Salwinski, L., et al. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32 (Database issue), D449–D451. doi: 10.1093/nar/gkh086

Santos, R., et al. (2017). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discovery* 16 (1), 19–34. doi: 10.1038/nrd.2016.230

Schneider, G. (2010). Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* 9 (4), 273–276. doi: 10.1038/nrd3139

Sheridan, R. P., et al. (2010). Drug-like density: a method of quantifying the "Bindability" of a protein target based on a very large set of pockets and drug-like ligands from the protein data bank. *J. Chem. Inf. Model.* 50 (11), 2029–2040. doi: 10.1021/ci100312t

Siegfried, Z., and Karni, R. (2018). The role of alternative splicing in cancer drug resistance. *Curr. Opin. Genet. Dev.* 48, 16–21. doi: 10.1016/j.gde.2017.10.001

Srivastava, A., Ahmad, S., and Gromiha, M. M. (2018). Deciphering RNA-recognition patterns of intrinsically disordered proteins. *Int. J. Mol. Sci.* 19(6), 1595. doi: 10.3390/ijms19061595

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Soc Networks* 11 (1), 1–37. doi: 10.1016/0378-8733(89)90016-6

Tan, Q., et al. (2013). Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* 341 (6152), 1387–1390. doi: 10.1126/science.1241475

Tantos, A., Kalmar, L., and Tompa, P. (2015). The role of structural disorder in cell cycle regulation, related clinical proteomics, disease development and drug targeting. *Expert Rev. Proteomics* 12 (3), 221–233. doi: 10.1586/14789450.2015.1042866

Tarcsay, Á., and Keserű, G. M. (2013). Contributions of Molecular Properties to Drug Promiscuity. *J. Med. Chem.* 56 (5), 1789–1795. doi: 10.1021/jm301514n

Tompa, P., et al. (2009). Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* 31 (3), 328–335. doi: 10.1002/bies.200800151

Tseng, C. Y., and Tuszynski, J. (2015). A unified approach to computational drug discovery. *Drug Discovery Today* 20 (11), 1328–1336. doi: 10.1016/j.drudis.2015.07.004

UniProt: the universal protein knowledgebase (2016). *Nucleic Acids Res.* 45 (D1), D158–D169. doi: 10.1093/nar/gkw1099

Uversky, V. N. (2012). Intrinsically disordered proteins and novel strategies for drug discovery. *Expert Opin. Drug Discovery* 7 (6), 475–488. doi: 10.1517/17460441.2012.686489

Uversky, V. N. (2014a). Introduction to intrinsically disordered proteins (IDPs). *Chem. Rev.* 114 (13), 6557–6560. doi: 10.1021/cr500288y

Uversky, V. N. (2014b). The triple power of D(3): protein intrinsic disorder in degenerative diseases. *Front. Biosci. (Landmark Ed.)* 19, 181–258. doi: 10.2741/4204

Uversky, V. N. (2017). Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* 44, 18–30. doi: 10.1016/j.sbi.2016.10.015

Uversky, V. N. (2018). Intrinsic disorder, protein-protein interactions, and disease. *Adv. Protein Chem. Struct. Biol.* 110, 85–121. doi: 10.1016/bs.apcsb.2017.06.005

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 18 (5), 343–384. doi: 10.1002/jmr.747

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

Uversky, A. V., et al. (2013). On the intrinsic disorder status of the major players in programmed cell death pathways. *F1000Res* 2, 190. doi: 10.12688/f1000research.2-190.v1

Uversky, V. N., et al. (2014). Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.* 114 (13), 6844–6879. doi: 10.1021/cr400713r

Vacic, V., et al. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* 6 (6), 2351–2366. doi: 10.1021/pr0701411

Varadi, M., et al. (2015). Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PloS One* 10 (10), e0139731. doi: 10.1371/journal.pone.0139731

Velankar, S., et al. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* 41 (Database issue), D483–D489. doi: 10.1093/nar/gks1258

Vuzman, D., and Levy, Y. (2012). Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.* 8 (1), 47–57. doi: 10.1039/C1MB05273J

Walsh, I., et al. (2015). Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* 31 (2), 201–208. doi: 10.1093/bioinformatics/btu625

Wang, C., and Kurgan, L. (2018). Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Brief Bioinform.* doi: 10.1093/bib/bby069

Wang, C., and Kurgan, L. (2019). Survey of similarity-based prediction of drug-protein interactions. *Curr. Med. Chem.* 25, 1–1. doi: 10.2174/0929867325666181101115314

Wang, K., and Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinform.* 7, 385. doi: 10.1186/1471-2105-7-385

Wang, J. Z., et al. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23 (10), 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, J., et al. (2012a). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 9 (4), 1070–1080. doi: 10.1109/TCBB.2011.147

Wang, J., et al. (2012b). The relationship between rational drug design and drug side effects. *Brief. Bioinform.* 13 (3), 377–382. doi: 10.1093/bib/bbr061

Wang, J., Peng, W., and Wu, F. X. (2013a). Computational approaches to predicting essential proteins: a survey. *PROTEOMICS–Clin. Appl.* 7 (1-2), 181–192. doi: 10.1002/prca.201200068

Wang, X., et al. (2013b). Evolutionary survey of druggable protein targets with respect to their subcellular localizations. *Genome Biol. Evol.* 5 (7), 1291–1297. doi: 10.1093/gbe/evt092

Wang, C., Hu, G., Wang, K., Brylinski, M., Xie, L., and Kurgan, L. (2016a). PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinformatics* 32, 579–586. doi: 10.1093/bioinformatics/btv597

Wang, C., Uversky, V. N., and Kurgan, L. (2016b). Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* 16 (10), 1486–1498. doi: 10.1002/pmic.201500177

Wang, C., Brylinski, M., and Kurgan, L. (2019). "PDID: Database of Experimental and Putative Drug Targets in Human Proteome," in In Silico Drug Design, ed. K. Roy. Academic Press, London, United Kingdom), 827-847. doi: 10.1016/B978-0-12-816125-8.00028-6

Wishart, D. S., et al. (2017). DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi: 10.1093/nar/gkx1037

Wojcik, S., et al. (2018). Targeting the intrinsically disordered proteome using small-molecule ligands. *Methods Enzymol.* 611, 703–734. doi: 10.1016/bs.mie.2018.09.036

Wu, Z., et al. (2015). In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. *FEBS Lett.* 589 (19 Pt A), 2561–2569. doi: 10.1016/j.febslet.2015.08.014

Xie, H., et al. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* 6 (5), 1882–1898. doi: 10.1021/pr060392u

Xue, B., and Uversky, V. N. (2014). Intrinsic disorder in proteins involved in the innate antiviral immunity: another flexible side of a molecular arms race. *J. Mol. Biol.* 426 (6), 1322–1350. doi: 10.1016/j.jmb.2013.10.030

Xue, B., et al. (2012). Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell Mol. Life Sci.* 69 (8), 1211–1259. doi: 10.1007/s00018-011-0859-3

Yan, J., et al. (2016). Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* 12 (3), 697–710. doi: 10.1039/C5MB00640F

Yildirim, M. A., et al. (2007). Drug-target network. *Nat. Biotechnol.* 25 (10), 1119–1126. doi: 10.1038/nbt1338

Yu, C., et al. (2016). Structure-based inhibitor design for the intrinsically disordered protein c-Myc. *Sci. Rep.* 6, 22298. doi: 10.1038/srep22298

Zhang, J., Ma, Z., and Kurgan, L. (2017). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief. Bioinform.* p, 1–19. doi: 10.1093/bib/bbx168

Zhao, S. (2019). Alternative splicing, RNA-seq and drug discovery. *Drug Discov. Today* 24, 1258–1267. doi: 10.1016/j.drudis.2019.03.030

Zheng, C. J., et al. (2006). Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* 58 (2), 259–279. doi: 10.1124/pr.58.2.4

Zhu, F., et al. (2009a). Update of TTD: therapeutic target database. *Nucleic Acids Res.* 38 (suppl_1), D787–D791. doi: 10.1093/nar/gkp1014

Zhu, M., et al. (2009b). Identifying drug-target proteins based on network features. *Sci. China C. Life Sci.* 52 (4), 398–404. doi: 10.1007/s11427-009-0055-y

Zhu, M., et al. (2009c). The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. *J. Drug Targeting* 17 (7), 524–532. doi: 10.1080/10611860903046610

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.