RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

RSC Advances

ARTICLE

Intrinsically disordered proteins in PubMed: What can the tip of the iceberg tell us about what lies below?

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Shelly DeForte[a] and Vladimir N. Uversky[a,b,c,d,e,*]

Intrinsically disordered proteins (IDPs) have a troubled history in the literature. Historically, a wide variety of terminology has been used to describe these strange proteins that do not adopt a stable three-dimensional structure. We provide here a survey of the current status of both IDPs and IDP terminology in PubMed. We have performed an extensive search of the literature from 1978 through 2014 and compiled a list of 1127 proteins and protein domains and the corresponding citations that refer to these proteins using IDP terminology. We show that papers that use IDP terminology are only the tip of the iceberg in terms of the larger body of literature referring to this group of proteins. Furthermore, our analysis suggests that this is likely due to a lack of perceived relevance rather than a lack of awareness. Finally, we have analyzed the language provided by author keywords, MeSH terms, and abstracts as well as the journals that are currently publishing IDP articles. Our results demonstrate a convergence on a common set of terminology and a rise in the number of papers using this terminology. However, our results also demonstrate that we have not reached the point where IDP terminology is fully accepted and embraced in the literature.

Introduction

Although for a very long time it was believed that the specific functionality of a given protein is predetermined by its unique three-dimensional structure,[1] several important exceptions from the "lock and key" rule were also known. Originally, these functional proteins with flexible structure had been discovered one by one and were considered as some special cases of unique polyfunctional proteins (e.g., serum albumin[2]), or polypeptides with unusual amino acid compositions (e.g., prothymosine α[3-5]), or proteins involved in the binding of large partners (RNA, DNA, proteins, and heme, e.g., histones,[6] ribosomal proteins,[7] myoglobin[8] and cytochrome *c*[9, 10]) or the binding of large quantities of small proteins (e.g., osteocalcin[11]). Furthermore, studies on protein folding pointed out that flexible structure might be of some functional importance. In fact, it has been pointed out that partially structured intermediates accumulated during protein folding (such as molten globule and pre-molten globule), which preserve some main elements of native secondary structure and their crude mutual positions in three-dimensional space, but differ from the rigid globular state by a less tight packing of side chains and by the dramatic increase in the mobility of loops and ends of chain, are almost ideal for some protein functions.[12-14] Therefore, it has been suggested that the molten globule state can exist in a living cell and can be involved in a number of physiological processes.[12-14] The validity of this hypothesis has been confirmed experimentally by showing the involvement of various partially folded intermediates in various biological processes, such as interaction with chaperones,[15] protein insertion into membranes,[16, 17] and interaction with ligands (summarized in refs.[18, 19]). However, in these early studies, even when the functionality has been attributed to the molten globule– or pre-molten globule–like conformations, the major emphasis still was on a concept of rigid three-dimensional structure. It has been hypothesized that the functional partially folded intermediates in a cell represent kinetic folding intermediates trapped by chaperones just after the protein biosynthesis before proteins can completely fold,[12-14] or appear as a result of point mutations preventing polypeptides from complete folding.[14, 20] Some other proteins (such as pore-forming domains of some toxins, or proteins that act as carriers of large hydrophobic ligands) were assumed to originally have a rigid structure but then were forced somehow to denature to carry out their functions.[13, 14]

The situation changed at the turn of century, when it was recognized that such biologically active proteins without unique structures are not merely a set of rare exceptions, but

[a] *Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA.*
[b] *USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA.*
[c] *Department of Biological Science, Faculty of Science, King Abdulaziz University, Jeddah, PO Box 80203, Jeddah 21589, Saudi Arabia.*
[d] *Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia.*
[e] *Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia*
* To whom correspondence should be addressed: E-mail: vuversky@health.usf.edu; Tel: 1-813-974-5816.
Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

instead represent a new and very broad class of proteins.[21-24] Since the publication of the first key studies and reviews describing this new concept, the literature on these proteins has virtually exploded (see Figure 1). The articles returned by intrinsically disordered protein (IDP) search terms each year are increasing at a rate greater than PubMed as a whole. The cumulative distribution of PubMed articles closely follows a parabolic growth curve, while the growth of articles returned by IDP search terms appears to be following a more exponential curve. This creates the illusion that protein intrinsic disorder has become a well-accepted phenomenon. The goal of our study was to validate this hypothesis and to answer an important question: "Are we there yet?"

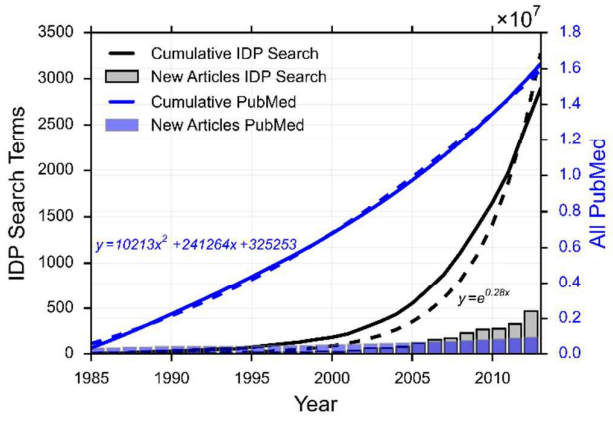


Figure 1. Time-dependent increase in the number of PubMed hits from IDP search terms versus all articles in PubMed. The following term was used to perform this search: "(intrinsically OR natively OR naturally OR inherently) AND (disordered OR unfolded OR unstructured OR denatured OR flexible) AND (protein OR region OR peptide OR domain) AND (1978/1/1:2014/10/15[dp])." A polynomial (for all PubMed) and exponential (for IDP search terms) function were matched to the curves starting with 1985 = 1, and are shown by the dashed lines.

The study of IDPs and intrinsically disordered protein regions (IDPRs) is, in many ways, the flip side of structural biology, with applications as far-reaching and ubiquitous. Initially, there was no consensus as to what these oddly behaving proteins should be called. This led to a number of different terms being used to describe the same phenomenon, such as *natively denatured*, *intrinsically unstructured*, *natively unfolded*, *inherently flexible*, and many others.[25] The term *intrinsically disordered*, however, has emerged as the predominant and agreed-upon term. Therefore, we shall use the terms *IDP* or *IDPR* to refer to these proteins, the phrase *IDP terminology* to refer to the set of terms used to refer to IDPs, and the terms *IDP paper* or *IDP literature* to refer to scientific papers that use IDP terminology.

The validation of protein intrinsic disorder *in vitro* or *in vivo* is challenging, and typically a consensus regarding the presence and nature of intrinsic disorder in a given protein will be developed over many studies. Therefore, it is not surprising that there is a great deal of inconsistency in terms of how and when the language of intrinsic disorder is used in the literature. This has contributed to a bottleneck in the curation of IDPs. While the fields of structural and *un*-structural biology are analogous in some ways, there are also key differences. A three-dimensional structure of a protein will typically be deposited in the Protein Data Bank[26] before papers related to that structure are published. There is no such process for IDPs. The current databases for experimentally verified IDPs, namely DisProt[27, 28] and IDEAL,[28] require the considerable efforts of IDP-focused researchers and can quickly lag behind the expanding literature. Furthermore, IDP-focused proteomics studies, evolutionary studies, disease-related studies, and functional studies require a synthesis of information over many different proteins and many different experiments. We have surpassed the point where it is possible to read all published papers for known, suspected, or recently discovered IDPs, and therefore researchers who specialize in IDPs are in many ways dependent on the presence of appropriate search terms to help them find what they are looking for.

There are many aspects of intrinsic disorder that are of interest to IDP-focused researchers, such as biophysical mechanisms, structural properties, disease-related properties, and structural and functional modifications under mutation, to name just a few. Therefore, the issue of clear language indicating IDPs and IDPRs in the literature is pressing and immediate.

PubMed is a citation aggregator catering primarily to the biomedical field. One advantage of PubMed is that it includes biomedical literature from MEDLINE, which is the U.S. National Library of Medicine's bibliographic database. A key feature of MEDLINE is that records are indexed with Medical Search Headings (MeSH) that connect related terms in a hierarchical structure, allowing for more targeted searching, even when precise keywords are not used. The PubMed search engine has become increasingly sophisticated with the ongoing expansion of MeSH terms and the official addition of author keywords in 2013.[29]

The field of IDPs is slightly behind the curve, however, as the term "intrinsically disordered proteins" was not added to MEDLINE's MeSH terms until 2014. This represents a potential boon for the organization and connection of IDP literature going forward. However, as PubMed does not retroactively index entries, the bulk of the IDP literature must still be referenced using a variety of keywords that search the abstract and title, which will often result in an incomplete picture.

At this significant juncture, it is our intention to present a survey of the use of IDP terminology (the tip of the iceberg) and present a picture of IDPs in the literature outside of this identifying terminology (what lies below), in hopes that this will encourage the community of researchers working with IDPs and IDPRs to contribute to a better connected body of research going forward.

Results and discussion

The incidence of intrinsic disorder in Swiss-Prot

Swiss-Prot is a database of manually curated protein sequences for over five hundred thousand proteins.[30] It is a

subset of the much larger UniProt Knowledgebase (KB), which contains an additional 50 million proteins that have been automatically annotated. Inspection of the composition of Swiss-Prot by organism quickly reveals that it is not a representative set of proteins across all proteomes. Bacteria are highly represented and compose approximately 61% of Swiss-Prot. Eukaryotes represent the second-largest group, at 32%, while archaea and viruses represent approximately 3% each. The large number of bacteria from similar proteomes results in many identical sequences, and because of this, only 84% of the Swiss-Prot sequence space is unique when identical sequences are clustered (calculated using the search term "uniprot: (reviewed: yes) AND identity: 1.0").

We used the regression-based, fast disorder predictor RAPID[31] to evaluate the percent predicted disorder over the entire Swiss-Prot database. Table I provides the numbers of proteins for each organism in each predicted disorder interval. As expected, proteins from bacteria and archaea were predicted to be the most structured, while eukaryotic proteins were predicted to be the least structured. We found that 20% of eukaryotic proteins in Swiss-Prot were predicted to be more than 30% disordered. However, it should be noted that the UniProt consortium places a high priority on the annotation of enzymes,[30] which is likely to skew the sequence space into one that is more highly structured, so this number should not be taken as representative for all eukaryotes.

Table I. The number of proteins in each disorder fraction interval separated by organism. Each protein is given a score that represents the predicted percent disordered. This table groups the proteins by the amount of predicted disorder (0–10%, 10–20%, etc.), and separates by organism.

The Number of UniProt IDs Associated with Percent Predicted Disorder				
Disorder	Archaea	Bacteria	Eukaryote	Virus
0–10%	9997 (54.0%)	172146 (52.0%)	69921 (40.0%)	7517 (47.0%)
10–20%	5086 (28.0%)	97867 (30.0%)	45295 (26.0%)	4295 (27.0%)
20–30%	1673 (9.1%)	30352 (9.2%)	24579 (14.0%)	2056 (13.0%)
30–40%	824 (4.5%)	13919 (4.2%)	14947 (8.6%)	937 (5.8%)
40–50%	346 (1.9%)	7391 (2.2%)	7759 (4.5%)	567 (3.5%)
50–60%	213 (1.2%)	3710 (1.1%)	4011 (2.3%)	342 (2.1%)
60–70%	102 (0.56%)	1795 (0.54%)	2129 (1.2%)	177 (1.1%)
70–80%	43 (0.23%)	1063 (0.32%)	1303 (0.75%)	127 (0.79%)
80–90%	23 (0.13%)	649 (0.2%)	1085 (0.63%)	40 (0.25%)
90–100%	41 (0.22%)	2598 (0.78%)	2338 (1.3%)	72 (0.45%)

Intrinsic disorder prediction and literature citations in Swiss-Prot

We then looked at the number of associated citations by disorder prediction interval. The citations in UniProt are not exhaustive, but rather cover those articles that provide the evidence needed for UniProt annotations. Therefore, the number of citations is an indication of the breadth of coverage in the literature on topics such as post-translational modifications, protein-protein interactions, subcellular locations, functions, and sequence-specific information.

In Figure 2, the relative number of literature citations is displayed for each predicted disorder interval. For all disorder intervals, the majority (60–64%) of bacteria have a single citation or no citations (25–30%), with little variance. We expect that in most cases, the single citation is a proteome-level study that produced the sequence in question. Eukaryotic proteins, however, tell a different story. The number of proteins that have greater than five citations actually peaks to 20% of the set within the disorder interval between 40 and 50%. The number of citations then falls sharply to a majority with one citation (59–60%) between 80 and 100% predicted disorder.

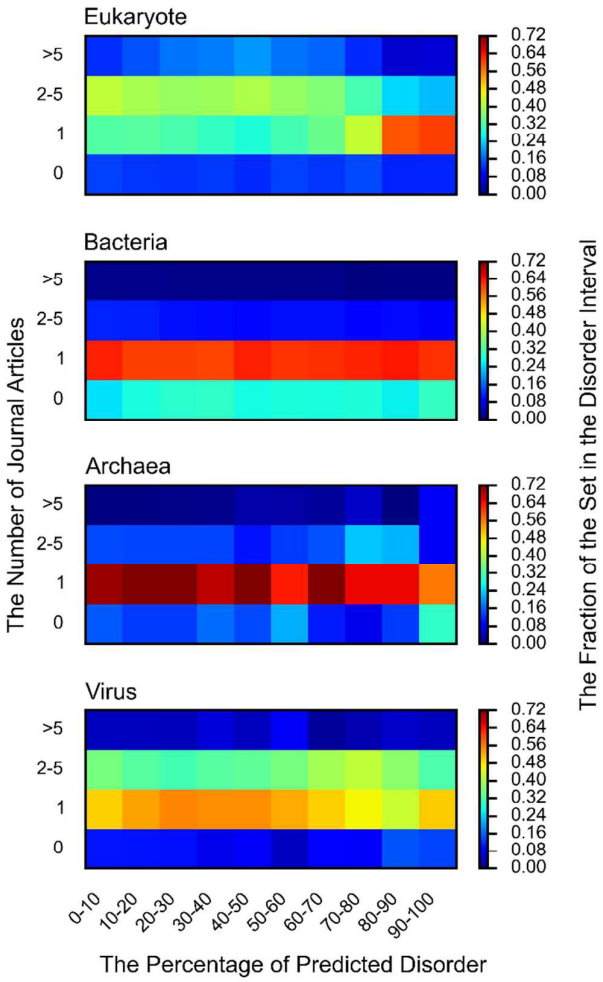


Figure 2. The number of PubMed articles linked to by Swiss-Prot, sorted by disorder fraction interval. The X axis groups the proteins by fraction of predicted disorder (0–10%, 10–20%, etc.). The Y axis is the number of PubMed articles linked to by each protein entry (0, 1, 2–5, >5). The heat map corresponds to the relative fraction of the set in a disorder interval column.

This demonstrates that a eukaryotic protein with a medium to high amount of predicted disorder is more likely to receive a large number of citations in Swiss-Prot than a completely structured or completely disordered protein. Surprisingly, it is the highly disordered proteins in archaea and viruses that are

more likely to receive a higher number of literature citations. Literature citations numbering between two and five jump up from 15% to 23% in the intervals between 70 and 90% predicted disorder in archaea. A similar occurrence happens in viruses, at the interval between 60 and 90%, where the fraction of proteins receiving between two and five citations jumps from 36% to 42%. Raw scores are available in the supplementary materials.

IDP terminology in PubMed

In order to investigate the use of IDP terminology in PubMed, we created 1127 search terms for proteins that have been referred to using IDP terminology in the literature, representing a total of 2278 papers.

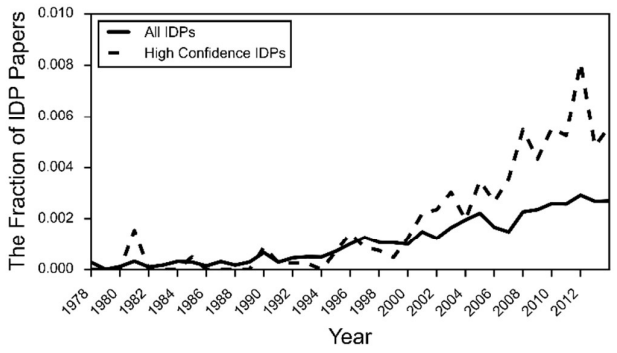


Figure 3. The fraction of PubMed IDs using IDP terminology by year. The fraction for each year is calculated by the number of PubMed IDs that use IDP terminology for specific proteins, divided by the total number of PubMed IDs obtained by a general search for those proteins. "High Confidence IDPs" are those defined as having greater than three associated papers using IDP terminology and greater than 30% disorder prediction by RAPID.

These were obtained through an extensive manual literature search (see Methods for search terms and criteria, and supplementary materials for the list of terms and associated PubMed IDs). 630 of our proteins could be linked to a DisProt ID, but 497 could not, demonstrating a significant expansion in the literature since the last DisProt update in 2013. Furthermore, this represents only those proteins that are referred to using IDP terminology, and therefore should not be considered to represent the entire set of IDPs.

Our objective was to compare the total number of papers that used IDP terminology to describe specific proteins to the total body of literature for those proteins. For each search term, we recorded the number of returned articles from PubMed, and calculated the fraction of all papers using IDP terminology during a given year, over the total amount of related literature for that year. For the entire set of IDP search terms, there is slow growth in the usage of IDP terminology to about 0.003 of the total set. However, because our criteria were simply that a protein be referred to as an IDP, there are likely to be some search terms in our set representing proteins that have been incorrectly assigned as IDPs. Therefore, we also created a set of "high confidence IDPs," which we defined as those that had more than three associated articles using IDP terminology and a greater than 30 RAPID score. These proteins have a somewhat more dramatic rise and peak, at just below 0.008. It should be noted that there is a certain amount of expected noise in the set due to irrelevant results from the search terms, or incorrectly assigned proteins. However, even given this expected noise, this fraction is still surprisingly low. We think it is unlikely that this fraction of papers that use IDP terminology represents the only papers that are relevant to IDP researchers.

Authors of IDP literature

We hypothesize that the small fraction of papers that use IDP terminology and slow pace of increase shown in Figure 3 could have several explanations:

1. The researchers are unaware the protein is intrinsically disordered.
2. The researchers do not think intrinsic disorder is relevant, or do not believe the protein is intrinsically disordered.
3. The researchers are using non-standard language to describe the structural properties of the protein.

In order to investigate the general awareness of intrinsic disorder, we examined the number of new authors publishing papers that use IDP terminology. Amongst our IDP papers there were 8425 unique authors. 6548 of those authors appeared on only 1 paper, 1151 appeared on 2 papers, 361 appeared on 3 papers, and the remaining 365 authors appeared on anywhere between 4 and 58 papers. While we expect some noise due to variations in spelling or the presence of identical names, these numbers seem to indicate that a wide variety of researchers are contributing to the IDP literature. Figure 4 shows the number of new researchers contributing to IDP papers per year. This shows both a growth of the number of authors per IDP paper and also a steady increase in the number of new contributing authors. It is not clear whether these are researchers who will continue to contribute to the IDP literature in the future, however.

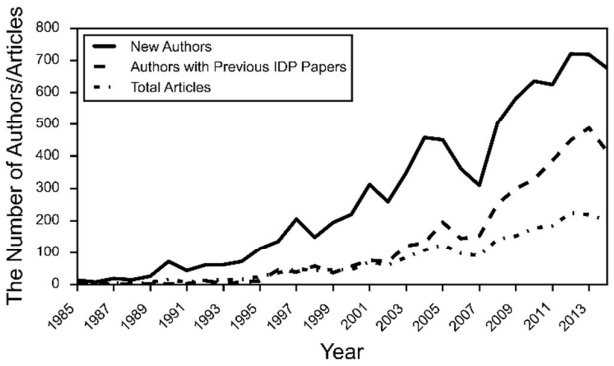


Figure 4. The number of new authors and authors with existing IDP papers per year. An author was counted as "new" at each year if they were not within the set of authors publishing in previous years; however, once an author was counted as "new," subsequent publications in the same year were counted in the set of authors with previous IDP papers.

We found it useful to conceptually separate authors into one of the following groups:

1. Authors who primarily study IDPs and contribute to papers on a variety of proteins. We would expect these authors to have a low number of papers for a specific protein, but that a high fraction of those papers would use IDP terminology.
2. Authors who primarily study one or more specific IDP proteins, but do not focus on the disordered properties of the protein. We would expect these authors to have a higher number of papers in the subject area but a low fraction of IDP papers.
3. Authors who primarily study one or more specific proteins and also focus on the intrinsically disordered properties of those proteins. We would expect these authors to have a high number of papers in the field and that a high fraction of those papers would use IDP terminology.

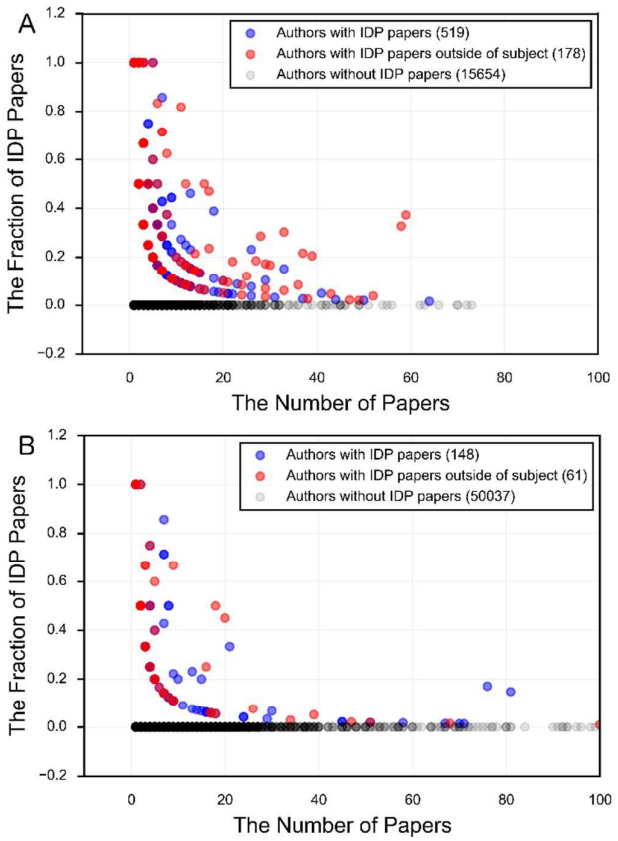


Figure 5. The number of papers per author for the search term in PubMed, plotted against the fraction of those papers that use IDP terminology. Each point represents an author on one or more papers associated with the given search term. The darker the dot, the larger the concentration of authors at that point. Blue dots are authors who have an IDP paper in the field in question (alpha-synuclein or tau, in this case), while the red dots are authors who have an IDP paper in the field in question and also have an IDP paper in a different field. The fraction of IDP papers is the number of papers by that author that use IDP terminology divided by all papers for that author and search term. The following search terms were used: A) (Top) "alpha synuclein" B) (Bottom) "tau AND (protein OR Alzheimer's OR tauopathies OR neuronal)."

Starting with this premise, we looked at two well-known IDPs: tau and alpha-synuclein. The first surprising result is that even with these well-known IDPs, the fraction of papers using IDP terminology is still very low. Alpha-synuclein peaks at a fraction of 0.05 IDP papers, while tau peaks at 0.008. Figure 5 shows, for each author, the number of papers published in the field versus the fraction of those papers that use IDP terminology for alpha-synuclein (Figure 5A) and tau (Figure 5B). Generally, it appears that the authors publishing papers using IDP terminology are publishing few papers in the field (group 1), and those who publish a large quantity of papers are, generally speaking, not using IDP terminology (group 2). Furthermore, a large number of papers using IDP terminology are published by authors who have published a research study on a different protein that is also an IDP, thus increasing the likelihood that it is IDP researchers (group 1) who are using IDP terminology.

The group of researchers who focus on the IDP properties of specific proteins (group 3), who would appear in the center and upper right portion of the graph, is fairly small, at least for alpha-synuclein and tau. While it is possible that the authors in the field are unaware of the IDP properties of these proteins, we feel this is unlikely for well-known IDPs such as alpha-synuclein and tau. Instead, it seems more likely that many authors are not using IDP terminology because they either do not believe the protein is intrinsically disordered, or they do not think the intrinsically disordered nature of the protein is relevant to the study in question.

The use of IDP terminology and predicted disorder

We then examined if the predicted amount of disorder was correlated with the use of IDP terminology in the literature. Figure 6 shows each of 1127 search terms plotted with the percent predicted disorder versus the fraction of IDP papers for the search term. The fraction of total proteins is highest for the 10–20% predicted disorder category, suggesting a high emphasis on smaller disordered regions in the literature. The mean fraction of papers that use IDP terminology is the highest as the protein becomes more disordered. This is not surprising, because it is in these proteins that intrinsic disorder is most likely to be persistently relevant. When weighting all search terms equally, the mean fraction of IDP papers reaches a fairly high percentage: nearly 20% in the case of highly disordered proteins. However, it is important to bear in mind that those proteins with a fraction of 1.0 have only one paper in most cases.

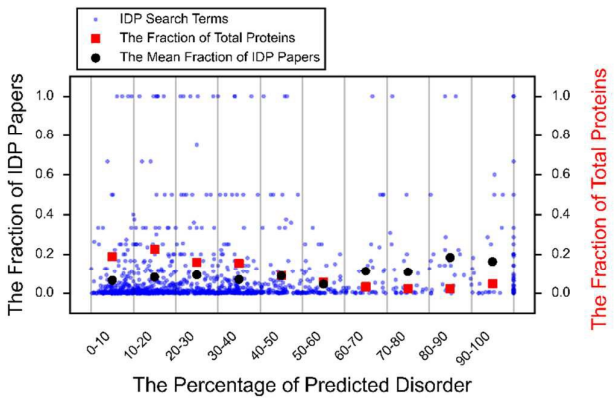


Figure 6. The fraction of predicted disorder versus the fraction of PubMed IDs that use IDP terminology. Each blue dot represents a protein search term. The percent predicted disorder for the protein is plotted against the number of PubMed IDs that use IDP terminology divided by all PubMed IDs associated with that protein search term. For each predicted disorder interval (0–10%, 10–20%, etc.), the fraction of the total proteins in that interval is plotted in red. The mean of the fraction of PubMed IDs that use IDP terminology is plotted for each fraction of disorder interval in black. This is the mean over the column, delineated by a grey line.

IDP terminology in abstracts, keywords, and MeSH terms

From our set of 2278 IDP papers, we examined the use of IDP terminology in the abstracts, the use of keywords, and the journals publishing IDP papers. Figure 7 clearly demonstrates the increasing consensus over the use of the term "intrinsically disordered" to describe the phenomenon in abstracts. However, until approximately 2007, the four most common IDP terms were in nearly equal use. However, in 41% of IDP abstracts, these four common terms did not appear, and we found that certain MeSH terms were disproportionately represented in this set. Not surprisingly, these terms tended to be oriented towards three-dimensional structure and enzyme catalysis, such as "crystallization" (85% of abstracts using this MeSH term did not use common IDP terminology), "catalysis" (100%), and "hydrogen bonding" (76%). Only one IDP-associated term was encountered among these MeSH terms, which was "pliability." This term had a low occurrence overall (15 appearances), but it appears in some cases that it was used before the IDP-specific MeSH term was introduced.

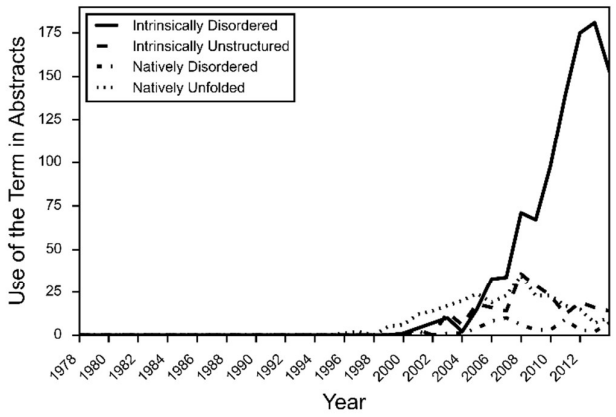


Figure 7. The usage of specific IDP terminology in PubMed abstracts. For each year, the abstracts for the PubMed IDs in the IDP set were searched for each of the terms listed, and the number was counted. These are the top 4 terms over all abstracts out of 20 total search terms (supplementary materials).

Surprisingly, 2110 of the 2278 IDP papers in the set did not have keywords available. However, this makes sense in light of the fact that PubMed did not officially add author keywords until 2013. Table II shows the occurrences of keywords for those entries with the author keyword field available. There were 67 appearances of either "intrinsically disordered protein" or "intrinsically disordered proteins." Not surprisingly, method-related terms such as "nuclear magnetic resonance/nmr," "molecular dynamics," and "circular dichroism" were common as well. Figure 8 shows a "Wordle" for the keywords in our set, with common words emphasized through an increase in size.

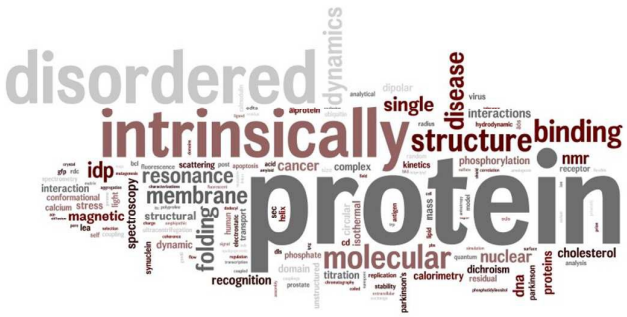


Figure 8. A Wordle for the keywords in IDP papers. The size of each word is increased in proportion to its number of occurrences.

MeSH terms appeared in 2198 of the articles; however, only 66 of the articles had the MeSH term "intrinsically disordered proteins." This is explained by the fact that the MeSH terms are not back indexed, and the term "intrinsically disordered proteins" was not introduced until 2014. The top 10 MeSH terms can be seen in Table III, and not surprisingly, emphasize protein structure, binding, and sequence. The full list of keywords and MeSH terms associated with IDP papers can be found in the supplementary materials.

Table II. Keywords and relative usage in IDP papers. The top 10 most common author supplied keywords in IDP papers.

The Top 10 Keyword Phrases	
Keyword Phrase	Appearances
intrinsically disordered protein(s)	67
nuclear magnetic resonance / NMR	14
alpha-synuclein	10
protein folding	9
protein structure	9
IDP	9
circular dichroism	8
protein-protein interactions	7
molecular dynamics	6
phosphorylation	6

The dearth of author-supplied keywords and IDP-specific MeSH terms in the literature means that either the title or

abstract must contain IDP terminology in order for the majority of papers to be retrieved in an IDP-specific search. It is very possible that this has significantly contributed to the low percentage of papers that are searchable by IDP terminology.

Overall, chemistry-focused journals were highly represented, with medical and biological journals following close behind, and a smaller number of structural biology and computationally-focused journals. It should be noted, however, that our set did not include reviews, proteomic studies, or protocol papers, and these are the top journals for primarily experimental studies on individual proteins. The full list of journals publishing IDP papers can be found in the supplementary materials.

Table III. MeSH terms and relative usage in IDP papers. The top 10 most common MeSH terms in IDP papers.

The Top 10 MeSH Terms	
MeSH Term	Appearances
humans	1010
amino acid sequence	923
molecular sequence data	897
protein structure, tertiary	688
models, molecular	673
animals	668
protein binding	634
protein conformation	634
protein structure, secondary	541
binding sites	441

We also looked at which journals are publishing IDP papers. Table IV shows the top 10 journals publishing IDP papers.

Table IV. Journals publishing IDP papers and the number of articles. This does not include reviews, proteomic studies, or protocol papers.

The Top 10 Journals Publishing IDP Papers	
Journal	Papers
J Biol Chem	256
Biochemistry	239
J Mol Biol	171
Proc Natl Acad Sci U S A	108
PLoS One	75
Protein Sci	65
J Am Chem Soc	51
Biophys J	51
Biochim Biophys Acta	50
Proteins	47

Experimental

PubMed coverage of predicted IDPs in Swiss-Prot

The Swiss-Prot database was downloaded from the UniProt KB on May 11, 2015. Fragments were removed, but no other filtering was applied. Duplicate sequences from multiple organisms were left in the set, as it was our goal to survey the entire database. The number of citations for each entry was obtained using UniProt filtering and selecting the "PubMed ID" column. The organism designations are provided by UniProt at http://www.uniprot.org/docs/speclist.

Disorder prediction for Swiss-Prot was obtained using the fast regression-based disorder predictor RAPID at http://biomine-ws.ece.ualberta.ca/RAPID/index.php.[31] We felt this predictor was the best choice because we needed to process a large number of sequences (all Swiss-Prot sequences), and RAPID provides high speed with high-quality predictions. RAPID was compared with 21 disorder predictors[31] and performed as well or better than any publically available predictor that performs at the speed we needed for such a large dataset. Furthermore, we did not need the detail provided by individual residue prediction and consensus methods, and instead only needed to place proteins within a disorder prediction bin (0–10%, 10–20%, etc.), and therefore we felt that a single predictor was sufficient. All parsing of the raw data files was done through custom Python scripts.[32]

IDP terminology in PubMed

IDPs in the literature can be referred to by a number of different terms. In order to try to maximize coverage while minimizing irrelevant results, we used the search term "(intrinsically OR natively OR naturally OR inherently) AND (disordered OR unfolded OR unstructured OR denatured OR flexible) AND (protein OR region OR peptide OR domain) AND (1978/1/1:2014/10/15[dp])" in PubMed at http://www.ncbi.nlm.nih.gov/pubmed/. This search covers the date ranges from January 1, 1978 through October 15, 2014. This search yielded 3343 results.

From the initial 3343 results, we manually examined each paper to try to ascertain which proteins were referred to as an IDP or indicated to have an IDPR. We recorded these names using the same language used in the corresponding literature. We discarded review, theory, proteomic, and method papers, as well as irrelevant results. This filtering resulted in 2278 PubMed articles attached to 1127 search terms, each corresponding to a protein or protein domain.

Our emphasis was primarily on the language used in the literature, and therefore we did not evaluate the experimental evidence. Because curation was not the primary objective of this project and naming conventions vary, there may be some duplicates and incorrect assignments, but we attempted to minimize this as much as possible. For each of the 1127 identified proteins and protein domains, we created a search term and attempted to maximize relevant results by adding qualifiers as necessary. For instance, the search term we created for tau was "tau AND (protein OR Alzheimer's OR tauopathies OR neuronal)," because a search for "tau" alone would return many irrelevant results. Similarly, the search term for p53 was "p53 AND (CTD or C-terminal or C-terminus)," because we wanted to specifically target our search towards the region that had been identified as intrinsically disordered. We attached DisProt and UniProt IDs to each protein search term; however, in many cases, this required an educated guess due to variations in naming conventions. In some cases, more than one UniProt and/or

DisProt ID was attached when multiple organisms were referred to in the paper(s). In cases where only a domain was mentioned, a UniProt ID was not assigned. There were 630 proteins in our set that could be attached to DisProt IDs. For each UniProt ID assigned to a protein, a disorder prediction was obtained by RAPID.

In order to get the number of both IDP and non-IDP papers per year, PubMed was automatically queried for each PubMed ID using the Biopython suite of tools[33] and custom Python scripts. The fraction of IDP papers is calculated as the number of IDP papers divided by the entire set of papers for that protein search query. The set of 2278 PubMed IDs formed the basis for the IDP-specific author, keyword, abstract, and journal data. This data was extracted from the corresponding MEDLINE entries using Biopython.

Conclusions

The field of protein intrinsic disorder has suffered from a lack of clear and consistent language describing the phenomenon in the literature. Many combinations of the terms *intrinsically*, *natively*, *naturally,* and *inherently* have been used with *disordered*, *unfolded*, *unstructured*, *denatured,* and *flexible* to describe proteins without a unique three-dimensional structure. However, we have shown here that a consensus in the literature has converged on the term *intrinsically disordered*, and accompanying this convergence has been a significant growth in the use of IDP terminology overall. However, this growth must be understood in the larger context of the body of literature referring to IDPs that do not use IDP terminology. For those who primarily study IDPs, it can seem as though there has been an explosion in the IDP literature. However, we have shown here that the number of papers *using IDP terminology* is still only a very small fraction of the greater body of literature referring to IDPs. In fact, the number of papers that use IDP terminology is only the tip of the iceberg in terms of the research that is happening in the field.

The curation of IDPs and the synthesis of literature that goes into building IDP theory requires painstaking manual literature searches that are further hindered by an inconsistent use of IDP terminology. It appears that the majority use of IDP terminology is by researchers who primarily study IDPs and not researchers who primarily study a specific protein or proteins. We would argue that consistent usage of IDP terminology will not increase significantly until more researchers see the value in using IDP terminology to describe the proteins they study. Therefore, the challenge is one of both increasing awareness and also expanding the perceived *relevance* of intrinsic disorder. The introduction of the MeSH term “intrinsically disordered proteins” and the addition of author keywords to PubMed allow for better indexing of IDPs in PubMed, and we highly recommend that studies involving IDPs recommend this MeSH term upon submission. Furthermore, we recommend the inclusion of the term “intrinsically disordered protein(s)” in the author keyword list.

Finally, we have provided here a list of 1127 proteins and protein domains that have been referred to using IDP terminology, along with their associated PubMed IDs (supplementary material), which includes 497 proteins not currently in DisProt. We hope this will provide a useful starting point for the further curation of recently recognized IDPs.

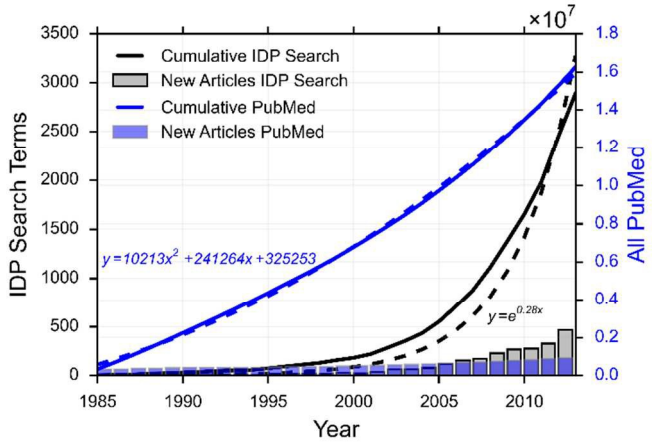
Acknowledgements

This work was supported in part by a grant from Russian Science Foundation RSCF № 14-24-00131.

Notes and references

1. E. Fischer, *Ber. Dt. Chem. Ges.*, 1894, **27**, 2985-2993.
2. F. Karush, *J. Am. Chem. Soc.*, 1950, **72**, 2705-2713.
3. J. D. Watts, P. D. Cary, P. Sautiere and C. Crane-Robinson, *Eur J Biochem*, 1990, **192**, 643-651.
4. K. Gast, H. Damaschun, K. Eckert, K. Schulze-Forster, H. R. Maurer, M. Muller-Frohne, D. Zirwer, J. Czarnecki and G. Damaschun, *Biochemistry*, 1995, **34**, 13211-13218.
5. V. N. Uversky, J. R. Gillespie, I. S. Millett, A. V. Khodyakova, A. M. Vasiliev, T. V. Chernovskaya, R. N. Vasilenko, G. D. Kozlovskaya, D. A. Dolgikh, A. L. Fink, S. Doniach and V. M. Abramov, *Biochemistry*, 1999, **38**, 15009-15016.
6. M. Boublik, E. M. Bradbury, C. Crane-Robinson and E. W. Johns, *Eur J Biochem*, 1970, **17**, 151-159.
7. S. Y. Venyaminov, A. T. Gudkov, Z. V. Gogia and L. G. Tumanova, *Absorption and circular dichroism spectra of individual proteins from Escherichia coli ribosomes*, Pushchino, Russia, 1981.
8. E. Breslow, S. Beychok, K. D. Hardman and F. R. Gurd, *J Biol Chem*, 1965, **240**, 304-309.
9. E. Stellwagen, R. Rysavy and G. Babul, *J Biol Chem*, 1972, **247**, 8074-8077.
10. W. R. Fisher, H. Taniuchi and C. B. Anfinsen, *J Biol Chem*, 1973, **248**, 3188-3195.
11. D. T. Isbell, S. Du, A. G. Schroering, G. Colombo and J. G. Shelling, *Biochemistry*, 1993, **32**, 11352-11362.
12. V. E. Bychkova, R. H. Pain and O. B. Ptitsyn, *FEBS Lett*, 1988, **238**, 231-234.
13. V. E. Bychkova and O. B. Ptitsyn, *Chemtracts Biochem. Molec. Biol.*, 1993, **4**, 133-163.
14. O. B. Ptitsyn, V. E. Bychkova and V. N. Uversky, *Philos Trans R Soc Lond B Biol Sci*, 1995, **348**, 35-41.
15. J. Martin, T. Langer, R. Boteva, A. Schramel, A. L. Horwich and F. U. Hartl, *Nature*, 1991, **352**, 36-42.
16. F. G. van der Goot, J. M. Gonzalez-Manas, J. H. Lakey and F. Pattus, *Nature*, 1991, **354**, 408-410.
17. F. G. van der Goot, J. H. Lakey and F. Pattus, *Trends Cell Biol*, 1992, **2**, 343-348.
18. V. N. Uversky and N. V. Narizhneva, *Biochemistry (Mosc)*, 1998, **63**, 420-433.
19. V. N. Uversky, in *Recent Research Developments in Biophysics & Biochemistry*, ed. S. G. Pandalai, Transworld Research Network, Kerala, India, 2003, vol. 3, pp. 711-745.
20. V. E. Bychkova and O. B. Ptitsyn, *FEBS Lett*, 1995, **359**, 6-8.
21. P. E. Wright and H. J. Dyson, *J Mol Biol*, 1999, **293**, 321-331.
22. V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins*, 2000, **41**, 415-427.

23. A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J Mol Graph Model*, 2001, **19**, 26-59.
24. P. Tompa, *Trends Biochem Sci*, 2002, **27**, 527-533.
25. A. K. Dunker, M. M. Babu, E. Barbar, M. Blackledge, S. E. Bondos, Z. Dosztányi, H. J. Dyson, J. Forman-Kay, M. Fuxreiter, J. Gsponer, K.-H. Han, D. T. Jones, S. Longhi, S. J. Metallo, K. Nishikawa, R. Nussinov, Z. Obradovic, R. Pappu, B. Rost, P. Selenko, V. Subramaniam, J. L. Sussman, P. Tompa and V. N. Uversky, *Intrinsically Disordered Proteins*, 2013, **1**.
26. W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten and G. Vriend, *Nucleic Acids Res*, 2015, **43**, D364-368.
27. M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res*, 2007, **35**, D786-793.
28. S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, S. D. Murakami, R. Koike, H. Hiroaki and M. Ota, *Nucleic Acids Res*, 2014, **42**, D320-325.
29. S. Torre, *Author Keywords in PubMed*, NLM Tech Bull., 2013 Jan-Feb.
30. U. Consortium, *Nucleic Acids Res*, 2015, **43**, D204-212.
31. J. Yan, M. J. Mizianty, P. L. Filipow, V. N. Uversky and L. Kurgan, *Biochim Biophys Acta*, 2013, **1834**, 1671-1680.
32. W. McKinney, *Journal*, 2010, 51-56.
33. P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. de Hoon, *Bioinformatics*, 2009, **25**, 1422-1423.



Papers that use IDP terminology represent only the tip of the iceberg of the larger body of literature on this subject.