Molecular
BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 3262–3273

www.rsc.org/molecularbiosystems

PAPER

Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification†

Harry Amri Moesa, Shunichi Wakabayashi, Kenta Nakai and Ashwini Patil*

Received 25th May 2012, Accepted 25th September 2012
DOI: 10.1039/c2mb25202c

Intrinsically disordered regions in proteins are known to evolve rapidly while maintaining their function. However, given their lack of structure and sequence conservation, the means through which they stay functional is not clear. Poor sequence conservation also hampers the classification of these regions into functional groups. We studied the sequence conservation of a large number of predicted and experimentally determined intrinsically disordered regions from the human proteome in 7 other eukaryotes. We determined the chemical composition of disordered regions by calculating the fraction of positive, negative, polar, hydrophobic and special (Pro, Gly) residues, and studied its maintenance in orthologous proteins. A significant number of disordered regions with low sequence conservation showed considerable similarity in their chemical composition between orthologs. Clustering disordered regions based on their chemical composition resulted in functionally distinct groups. Finally, disordered regions showed location preference within the proteins that was dependent on their chemical composition. We conclude that preserving the overall chemical composition is one of the ways through which intrinsically disordered regions maintain their flexibility and function through evolution. We propose that the chemical composition of disordered regions can be used to classify them into functional groups and, together with conservation and location, may be used to define a general classification scheme.

Introduction

The abundance of intrinsically disordered regions (IDRs) in proteins and their importance in protein function have been extensively studied over the past decade.[1–4] IDRs play an important role in the functions of proteins either as flexible linkers,[5] regions that undergo a disorder-to-order transition upon binding,[6] or in dynamic complexes.[7] It has been observed that IDRs evolve more rapidly than ordered regions within proteins.[8,9] While the amino acid sequence is often not conserved in IDRs, it is still important for maintaining the structural flexibility and function. For instance, the IDR in the DNA binding protein RPA70 maintains its disordered state despite poor sequence conservation across three taxa.[10] The poor sequence conservation in IDRs raises two important issues.

Firstly, it is not clear how the IDRs stay functional in the absence of a specific structure or a conserved sequence, both of which are associated with functional domains. While recent studies have looked at the sequence conservation of IDRs, they do not explain this phenomenon.[11] The amino acid sequence of an IDR is not random,[12] presumably because it plays an important role in defining the structure (or the lack thereof) and function. Recent computational studies have shown that IDRs with similar amino acid content co-occur with specific structural domains indicating a possibility of shared function.[12] Indeed, the use of amino acid content similarity alone can help identify IDRs with similar functions at statistically significant levels.[13] Thus, the amino acid content and the resultant overall chemical composition of the IDR are clearly important. For instance, it has been observed that the amino acid composition, specifically the fraction of Gly residues, is similar in the N-terminal domains of core histones H2A and H4 in spite of poor sequence similarity.[14] In the case of several DNA-binding proteins, the net charge of the disordered tail affects their ability to efficiently search DNA for binding regions.[15] The net charge has also been observed to affect the overall dimensions of the protein containing the IDR.[16] This raises the possibility that IDRs evolve to maintain not only their sequence but also their chemical composition.

A second related problem, as a consequence of the lack of sequence conservation, is the difficulty in defining a classification system for IDRs that is similar to that of conserved domains, as in Pfam,[17] in order to separate them into functional groups. A unified classification system is important not only in improving our understanding of IDRs, but would also facilitate the

Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. E-mail: ashwini@hgc.jp; Fax: +81-3-5449-5133; Tel: +81-3-5449-5131

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25202c

annotation of uncharacterized proteins with large IDRs and no identifiable sequence homologs. The current domain classification techniques are based either on structure[18] or on sequence conservation.[17] Due to their lack of structure and sequence conservation, a large number of IDRs are not amenable to these classification techniques. Though several attempts have been made to classify IDRs,[19,20] there is no generally accepted means of classifying them into functional groups based on their sequence.

In this study, we addressed the aforementioned two issues. We assigned amino acids to groups based on their chemical and structural properties and used these residue types to (1) test the maintenance of the fraction of charged, polar, hydrophobic and special (Pro, Gly) residues across orthologous IDRs and, (2) classify IDRs into functionally distinct groups based on their chemical composition as given by their residue type content. We found that a large proportion of IDRs with low sequence conservation show significant conservation of their chemical composition. Additionally, IDRs can be classified into five broad functional groups based on their chemical composition and show a composition dependent location preference within proteins. We propose that poorly conserved IDRs maintain their function by preserving their chemical composition and this property, along with conservation and location, can be used in classifying them into functional groups.

Results

We studied the conservation of intrinsically disordered regions (IDRs) from human proteins in 7 other eukaryotes (chimp, dog, rat, mouse, fly, worm and yeast) and attempted to classify the IDRs based on their conservation, chemical composition and location into functionally distinct groups. Human proteins with orthologs in at least 4 of the 7 species were selected. Orthologous proteins were aligned using ClustalW.[21] IDRs longer than 30 residues were predicted using DisoPred2[22] and IUPred.[23] Experimentally determined IDRs were obtained from DisProt[24] by aligning all IDRs in DisProt to human proteins with 4 or more orthologs and selecting those with 90% or greater sequence identity with the human proteins.

Conservation

We studied the conservation of IDRs at three levels (Fig. 1):

Table 1 Residue types assigned to each amino acid based on their side-chain properties

Type	Amino acid
Positive	Arg, Lys
Negative	Asp, Glu
Polar	Cys, Gln, His, Ser, Thr, Tyr, Asn
Hydrophobic	Ala, Phe, Ile, Leu, Met, Val, Trp
Special	Pro, Gly

(1) *Residue conservation* – The residue conservation score of an IDR indicates the level of sequence conservation of the IDR within the orthologous proteins (see Materials and methods). It was calculated for each IDR across 8 species using a scoring scheme similar to that proposed by Bellay *et al.*[11] We chose this scoring scheme because its utility in differentiating IDRs based on conservation has already been demonstrated.[11]

(2) *Residue type conservation* – In order to determine if the residue type was more frequently conserved in IDRs compared to the amino acid residue itself, we assigned one of five types to each amino acid depending on the nature of its side chain (Table 1). We chose these five categories to highlight the differences in the chemical and structural properties of amino acid residues. Hence, Pro and Gly were assigned to a separate category due to their special structural properties.[25] We then calculated the type conservation score of each IDR similar to the residue conservation score (see Materials and methods) to determine how often a residue type was conserved within the aligned regions in orthologous proteins.

(3) *Type content conservation* – Type content of an IDR is the fraction of each residue type in the IDR (described in Table 1). It indicates the overall chemical composition of the IDR. Type content conservation is the maintenance of the fraction of residue types in regions within orthologous proteins that are aligned to the reference IDR. It was calculated as the average Euclidean distance between the type content of an IDR in a human protein and that in aligned regions within orthologous proteins (see Materials and methods). A smaller distance between the human IDR and its orthologs indicates a greater similarity in content and hence higher type content conservation and similar in chemical composition.

Fig. 2A shows the distribution of the residue and residue type conservation scores in all predicted IDRs. 98% IDRs show a greater level of residue type conservation than simple

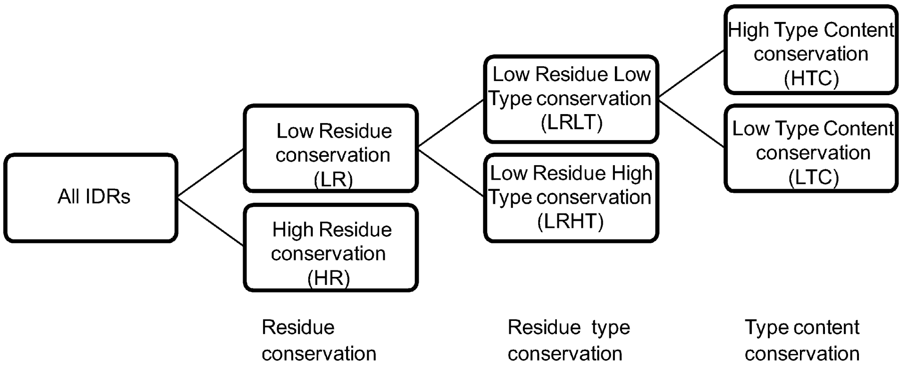

All IDRs
Low Residue conservation (LR)
High Residue conservation (HR)
Low Residue Low Type conservation (LRLT)
Low Residue High Type conservation (LRHT)
High Type Content conservation (HTC)
Low Type Content conservation (LTC)
Residue conservation
Residue type conservation
Type content conservation

Fig. 1 IDR groups based on conservation. Groups of IDRs based on their residue, residue type and type content conservation. Residue conservation: conservation of amino acid sequence; residue type conservation: conservation of each residue type in the amino acid sequence; type content conservation: conservation of the chemical composition (position independent).

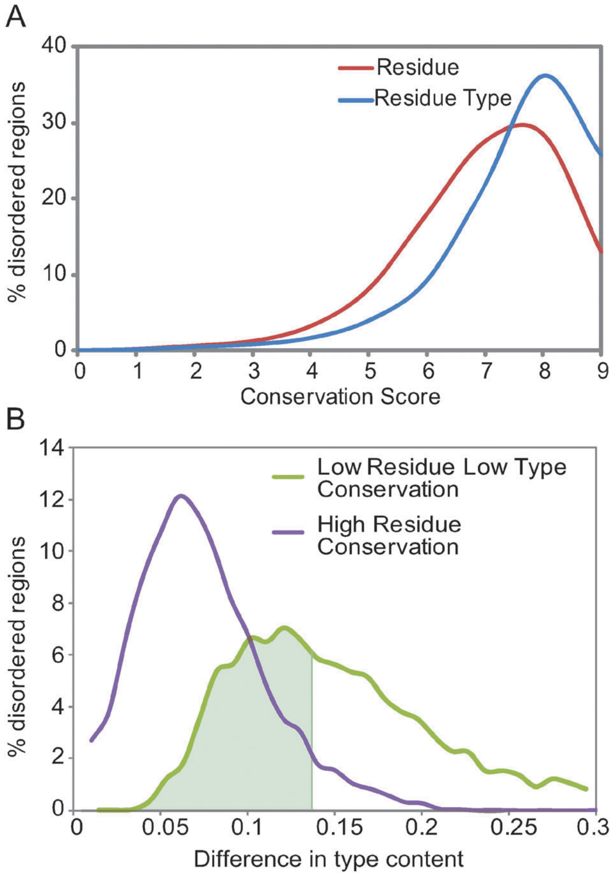


Fig. 2 Residue and type content conservation in IDRs. (A) Distribution of the residue and residue type (positive, negative, polar, hydrophobic and special) conservation of human intrinsically disordered regions. A greater proportion of disordered regions show residue type conservation. (B) Distribution of the difference in residue type content (denoted by Euclidean distance) of IDRs in orthologous proteins having high residue conservation (purple) and low residue, low residue type conservation (green). Almost half the IDRs with low residue type conservation have conserved residue type content similar to that in IDRs with high residue conservation (shaded area).

residue conservation ($p \ll 0.01$), with only 2% showing no difference between the two conservation scores. A similar tendency was also observed in experimentally determined IDRs from DisProt, which have a greater average residue type conservation score compared to the average residue conservation score (Table S1 (ESI†), $p \ll 0.01$).

We then focused on two groups of IDRs in order to study the differences between those with and without position dependent sequence conservation (Fig. 1):

(1) IDRs with high residue conservation (HR) – IDRs with a residue conservation score greater than average. These IDRs have high sequence conservation across orthologous proteins

(2) IDRs with low residue and low type conservation (LRLT) – IDRs with both, residue conservation and residue type conservation scores, less than average. These IDRs not only show poor amino acid conservation but also poor residue type conservation and thus lack significant position dependent sequence conservation.

We chose the average values of conservation scores as cutoffs in defining different IDR groups because the threshold suggested by Bellay *et al.*[11] was unsuitable for our distribution of conservation scores.

We compared the conservation of chemical composition in IDRs within the two groups defined above through their type content similarity (Fig. 2B). As expected, IDRs with high residue conservation scores, *i.e.* high sequence conservation, have greater type content similarity with their orthologs. This is indicated by a lower average Euclidean distance between the type content of the IDRs and their aligned orthologous regions. However, 52% of IDRs with low residue conservation (LRLT) have type content similarity with their orthologs that is as high as that of IDRs with high residue (HR) conservation (within 2 standard deviations of the average Euclidean distance). This result indicates that these IDRs show similar chemical composition in orthologous proteins despite having poor sequence conservation.

This result was confirmed in IDRs predicted using IUPred in order to eliminate bias caused by using a particular disorder predictor (Fig. S1 and Tables S1, S2, ESI†). Further confirmation was obtained in experimentally determined IDRs from DisProt, where 51% of the IDRs with low residue conservation show type content conservation similar to that found in highly conserved IDRs (Table S2, ESI†). Additionally, all conservation scores based on ClustalW alignments of orthologous regions were significantly different from those obtained using random alignments (see Materials and methods for details) confirming the reliability of the alignments used (Fig. S2 (ESI†), $p \ll 0.01$).

Based on these results, we conclude that a significant number of IDRs with poor sequence conservation maintain their type content and thus their chemical composition through evolution. We propose that this preservation of chemical composition is the means through which IDRs retain their function and flexibility in spite of a lack of sequence conservation.

Conservation dependent characteristics

In order to study the differences in IDRs based on the conservation of their type content, we further separated the LRLT IDRs into two groups (Fig. 1, and Table S2, ESI†):

(1) IDRs with high type content conservation (HTC) – LRLT IDRs with type content difference between orthologous regions within two standard deviations of that in HR IDRs. These IDRs show high conservation of chemical composition across orthologs.

(2) IDRs with low type content conservation (LTC) – LRLT IDRs with type content difference within orthologous regions greater than two standard deviations of that in HR IDRs. These IDRs show poor conservation of chemical composition across orthologs.

We compared the properties of IDRs in these two groups with HR IDRs. Fig. 3A shows the relationship between the type content difference between orthologous IDRs (given by their average Euclidean distance from the human IDR) and their residue conservation score. While the HR IDRs show a high negative correlation between the type content difference and the residue conservation score among orthologs ($r = -0.65$, $p < 0.01$), the HTC IDRs show a relatively poor

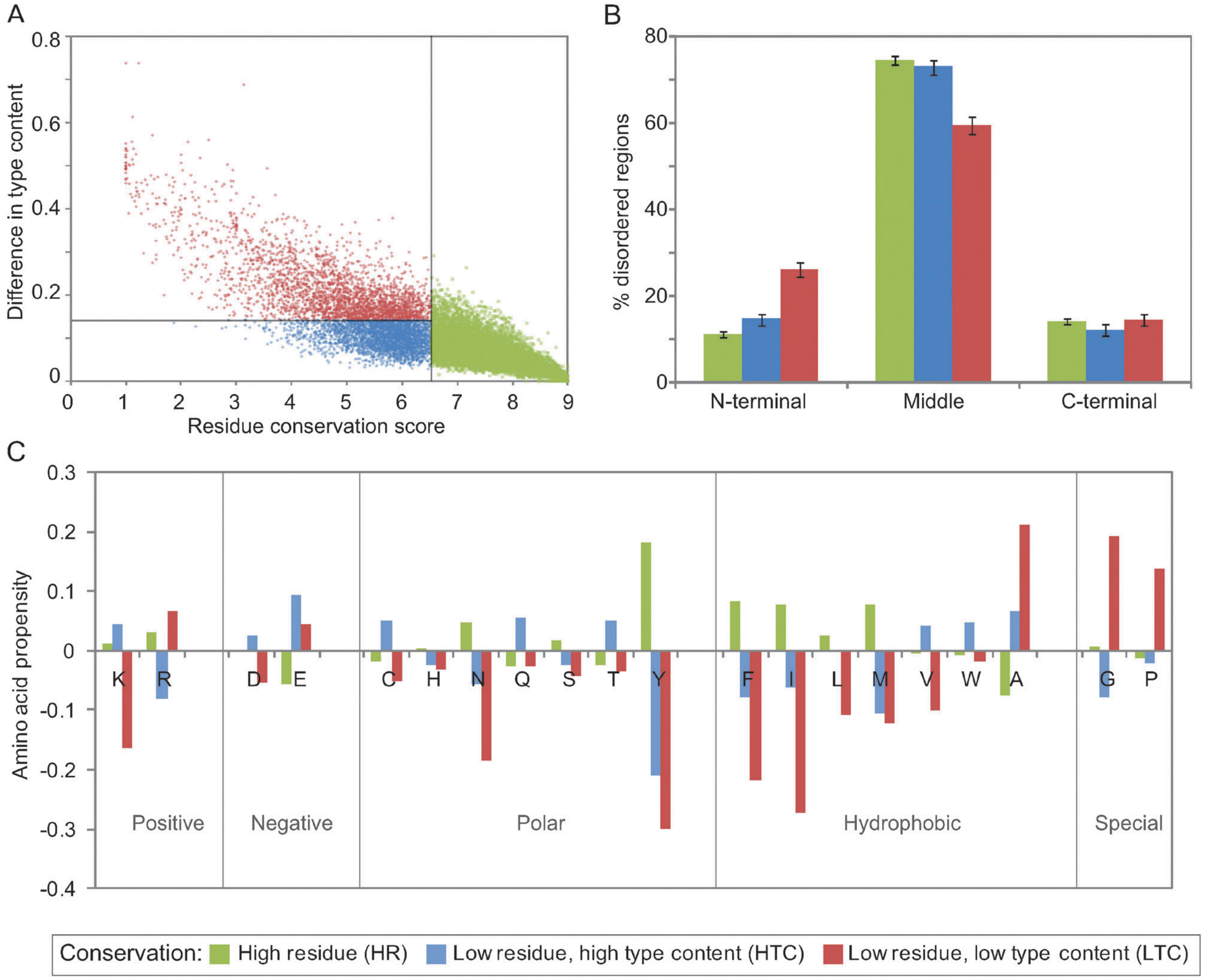


Fig. 3 Characteristics of IDRs based on conservation. (A) Relationship between residue conservation and the difference in type content within orthologs. The graph highlights the three conservation-based groups of IDRs studied. (B) Location preferences, and (C) amino acid propensity of IDRs with high residue conservation (green), low residue conservation with high type content similarity (blue) and low residue conservation with low type content similarity (red). Error bars indicate 95% confidence intervals. IDRs show distinct location preference and amino acid propensity depending on their level and type of conservation.

correlation ($r = -0.22$, $p < 0.01$). The LTC IDRs also show a high negative correlation between the type content difference and the residue conservation score primarily because of the poor conservation of both residue and type content in these IDRs.

We defined IDRs within 40 residues of either end of the protein as terminal IDRs[26] to compare the location preference of IDRs in the three groups within proteins. While all IDRs are more likely to be away from the termini, LTC IDRs are enriched at the N-terminal region of proteins and depleted in the middle (Fig. 3B, $p \ll 0.01$). HTC IDRs show location preference that is similar to that of HR IDRs. All types of IDRs are equally likely at the C-terminal region.

The amino acid propensity provides further evidence of the differences in IDRs in these three groups (Fig. 3C). Polar residues, especially Tyr is abundant in highly conserved IDRs possibly indicating conserved phosphorylation sites or binding sites. Several hydrophobic residues, with the exception of Trp, are also enriched and could indicate binding sites. On the other hand, HTC IDRs show an abundance of charged (Glu, Asp, Lys) and polar residues (Gln, Cys, Thr), along with a few hydrophobic residues (Val, Ala), especially Trp. The charged and polar residues play an important role in defining the overall chemical composition of the IDR and, possibly, its flexibility. The enrichment of certain hydrophobic residues might indicate the presence of small linear motifs within these IDRs. Finally, the LTC IDRs are abundant in not just Pro and Gly, but also Ala, and to some extent Arg and Glu. These IDRs show a specific amino acid composition, but lack conservation of relative proportions of these residues across orthologous regions distinguishing them from HTC IDRs. It has been previously shown that the combined Pro and Gly content can be a distinguishing factor in proteins showing elastomeric properties or a tendency towards amyloid formation.[27] This raises the possibility that the combined content of Pro and Gly residues might show greater conservation than other residue

Table 2 Gene ontology molecular function terms enriched in proteins containing IDRs with high residue conservation (HR), IDRs with low residue conservation but high type content conservation (HTC), and IDRs with low residue and type content conservation (LTC) ($p \leq 0.01$, FDR ≤ 1)

High residue (HR)	High type content (HTC)	Low type content (LTC)
Transcription regulator activity	Adenyl ribonucleotide binding	Cation binding
Sequence-specific DNA binding	Adenyl nucleotide binding	Metal ion binding
Transcription factor activity	Purine nucleoside binding	Ion binding
Transcription activator activity	Nucleoside binding	
DNA binding	ATP binding	
Transcription repressor activity	Nuclease activity	
Transcription factor binding	ATPase activity	
Transcription cofactor activity	Purine nucleotide binding	
Chromatin binding	Ribonucleotide binding	
Transcription coactivator activity	Purine ribonucleotide binding	
RNA polymerase II transcription factor activity		
Ras guanyl-nucleotide exchange factor activity		
Guanyl-nucleotide exchange factor activity		
GTPase regulator activity		

types in LTC IDRs. We found that while the combined fraction of Pro and Gly residues is better conserved than hydrophobic and polar residues, they are not as strongly conserved as charged residues (Fig. S3, ESI,† $p \ll 0.01$). Further, the combined conservation of Pro and Gly residues is inversely proportional to their content in the IDR, with fewer Pro and Gly residues being better conserved ($r = 0.71$ between Pro and Gly content and difference between orthologous IDRs, $p \ll 0.01$). These results indicate that greater conservation of the Pro and Gly residues is more prevalent when they are present in smaller amounts in LTC IDRs.

We performed Gene Ontology term enrichment analysis of proteins containing IDRs in HR, HTC and LTC IDRs to identify functional associations of IDRs based on their levels of conservation (Table 2). IDRs with high residue conservation (HR) are enriched in proteins involved in transcription regulation and DNA binding. Some of the proteins containing HR IDRs are transcription factors with Homeobox domains or zinc finger domains that bind to DNA. On the other hand, HTC IDRs with low residue and high type content conservation are enriched in proteins showing ATPase activity and nuclease activity, and are involved in the biological processes of DNA replication and repair among others. These IDRs are often found in RNA helicases and kinases. Finally, the LTC IDRs which show neither sequence nor type conservation are abundant in proteins enriched in ion binding functionality. These IDRs are often found in peptidases.

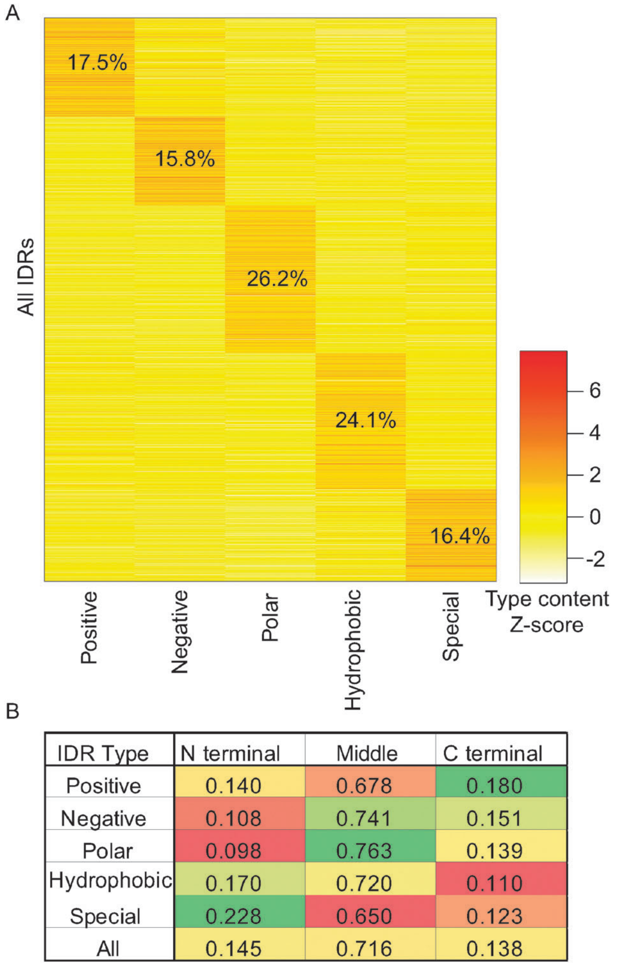
These results demonstrate that IDRs with differing levels of conservation have distinct location preference, amino acid composition and functional properties.

Chemical composition based classification

We next attempted to classify all predicted IDRs into functionally distinct groups based on their chemical composition. All IDRs were clustered into 5 groups (positive, negative, polar, hydrophobic, special) based on their residue type content. Each cluster contained IDRs that were enriched in one type of residue over all others. For instance, the positive cluster contained IDRs enriched in positively charged residues and depleted in other residue types. Similarly, IDRs in the hydrophobic cluster show the greatest enrichment in hydrophobic residues above the expected values. As seen in Fig. 4A, the IDRs separate into distinct clusters based on their residue type content. IDRs with an over-representation of polar residues form the largest cluster.

We evaluated the location preference of IDRs in these clusters within proteins to determine any composition dependence. Regions in all clusters are most abundant away from the N- and C-termini in the protein (Fig. 4B, and Fig. S4, ESI†). However, IDRs enriched in hydrophobic and special residues were located at the N-terminal region more frequently than expected ($p \ll 0.01$). On the other hand, IDRs enriched in positive residues were more likely to be located at the C-terminal region ($p \ll 0.01$). Further, IDRs enriched in polar and negative residues were under-represented at the N-terminal region ($p < 0.01$), while IDRs in the special and positive clusters were significantly under-represented away from the termini ($p < 0.02$). These results demonstrate the association between the location of IDRs in proteins and their chemical composition.

Fig. 5 shows the broad biological process categories associated with IDRs in different clusters (see Table S3 (ESI†) for over-represented molecular function and cellular component terms). IDRs in the positive cluster are related to RNA processing and chromatin binding. This is not unexpected, since RNA and DNA are negatively charged. Based on location preference results (Fig. 4B), these IDRs often appear to be C-terminal tails and are included in several proteins from the DEAD box helicase family. These proteins have positively charged C-terminal tails that enable their non-specific binding to RNA.[28] Negative IDRs are involved in protein folding and ion transport. These include heat shock proteins and ion exchangers. Polar IDRs are enriched in signal transduction and are generally a part of receptors and kinases. IDRs in the hydrophobic cluster (having hydrophobic content above average) are present in proteins acting in development and the regulation of neurogenesis. Finally, proteins with IDRs abundant in Pro and Gly are involved in epidermis and ectoderm development, and macromolecule metabolic process. Many of the IDRs in the special cluster form coiled coils and are associated with fibrous proteins like keratin and collagen. Thus, we show that IDRs in each of the five groups are associated with distinct functions.



IDR Type	N terminal	Middle	C terminal
Positive	0.140	0.678	0.180
Negative	0.108	0.741	0.151
Polar	0.098	0.763	0.139
Hydrophobic	0.170	0.720	0.110
Special	0.228	0.650	0.123
All	0.145	0.716	0.138

Fig. 4 Classification of IDRs based on chemical composition. (A) Clustering of IDRs into 5 groups based on the residue type content. IDRs in each cluster are enriched for a single residue type. Each row in the heatmap corresponds to an IDR and each value indicates the *z*-score for the fraction of five types of residues in one IDR. (B) Location preference of the IDRs in type based clusters. Columns are colored by fraction of IDRs in each cluster at a specific location. Green indicates the highest propensity at a specific location while red indicates a low propensity with yellow denoting an average value. IDRs show distinct location preference based on their chemical composition.

Additionally, when IDRs in each cluster were further separated based on their location within the protein, they showed enrichment of distinct GO terms based on their location and chemical composition (Fig. 5 and Table S4, ESI†). For instance, C-terminal IDRs in the positive cluster are associated with mRNA processing, as previously discussed. On the other hand, those in the middle are associated with chromatin organization. Similarly, N-terminal IDRs in the negative clusters are enriched in GO terms related with amino acid transport, those in the middle are associated with regulation of kinase activity, and those at the C-terminus are associated with protein folding and vesicle-medicated transport. Thus, IDRs show a better functional separation using a combination of chemical composition and location within the protein.

We also separately clustered HR, HTC and LTC IDRs by type content to determine the impact of conservation on IDR type and function (Fig. S5, ESI†). While all three types of IDRs can be clustered into five groups based on their chemical composition, the fraction of IDRs in each cluster varies depending on their level and type of conservation, and is representative of the overall chemical composition of each group (Fig. 6). While HTC IDRs are more likely to be present in charged clusters, HR IDRs are significantly more likely to be in hydrophobic clusters. LTC IDRs are most abundant in the special cluster and under-represented in the charged clusters. IDRs in each group also show essentially the same location preferences as those observed for all IDRs (Fig. S5, ESI†). GO term enrichments show results similar to those of all IDRs but with slight differences in the functional associations of IDRs in the negative, hydrophobic and special clusters (Tables S5–S7, ESI†). IDRs in positive clusters for all three groups are enriched in RNA processing functions. Similarly, those in the polar cluster in HR and HTC groups are enriched in signal transduction related functions. The hydrophobic IDRs in the HTC group are localized in the mitochondria. Special IDRs in the HR group are enriched within the collagen and the extracellular matrix.

We studied HTC IDRs further to determine the role of residue distribution and the presence of short functional regions in long IDRs. We studied the extent of residue clustering in HTC IDRs and their potential effect on function. 18% of the IDRs had 10-residue clusters containing at least 70% residues of the same type, while only 2.2% IDRs contained such clusters of size 20. More than half the clusters consisted of polar residues only, though these are unlikely to be functionally significant given the general abundance of polar residues in IDRs. Apart from these, clusters of negatively charged residues were most abundant while those of positively charged residues were depleted (Table S8, ESI†). Of the 3042 HTC IDRs, only 6 had clusters of size 20 containing 90% residues of one type, while 43 IDRs with clusters of size 10 had this property. The IDRs containing such residue clusters were weakly associated with functions. IDRs containing clustered hydrophobic residues were associated with sequence-specific DNA binding ($p = 0.01$), while IDRs with clustered special residues were associated with the extracellular matrix and collagen ($p = 0.01$). These results confirm the existence of residue clusters in IDRs, though their role in function is not clear.

For 142 HTC IDRs greater than 200 residues in length, we attempted to identify shorter regions having a distinct chemical composition and hence function. We hypothesized that such short sub-regions would show greater content conservation compared to the entire IDR and identified sub-regions longer than 30 residues showing greater type content conservation than the entire IDR (see Materials and methods for details). 26% of the IDRs contained at least 2 such regions. 63% of these regions showed an enrichment of charged residues but only 16 IDRs showed content that was significantly different from their parent IDR. However, all the short sub-regions showed greater residue and a residue type conservation score than their parent IDR indicating the presence of short conserved regions within long IDRs that are potentially important for their function.

Thus, the location and functions of IDRs are dependent on their chemical composition as represented by type content and this property can be used to classify them into functionally distinct groups. These results also indicate that in IDRs with

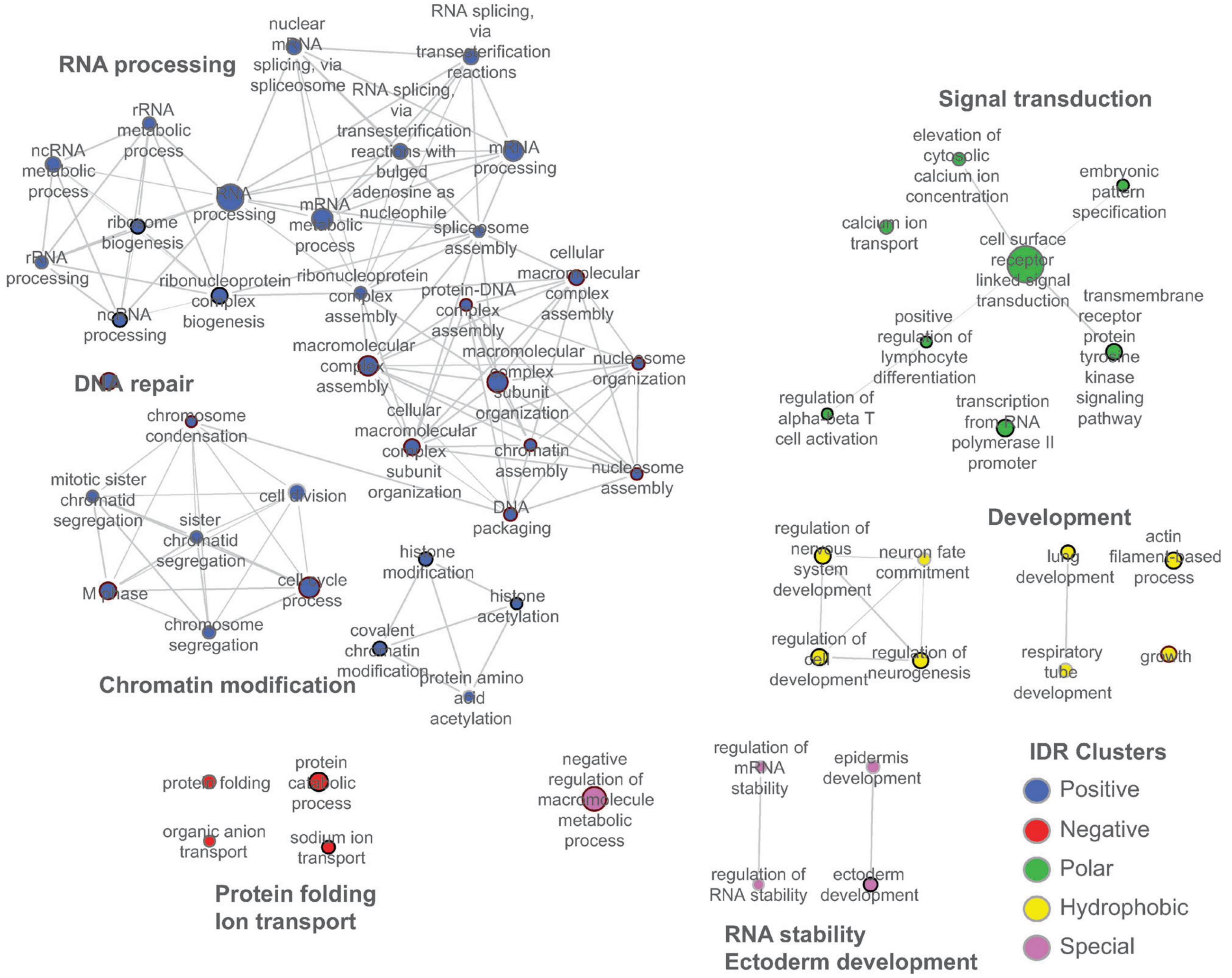

RNA processing
nuclear mRNA splicing, via spliceosome
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
rRNA metabolic process
ncRNA metabolic process
mRNA processing
RNA processing
mRNA metabolic process
ribosome biogenesis
spliceosome assembly
rRNA processing
ribonucleoprotein complex biogenesis
ribonucleoprotein complex assembly
cellular macromolecular complex assembly
ncRNA processing
protein-DNA complex assembly
macromolecular complex assembly
macromolecular complex subunit organization
nucleosome organization
DNA repair
chromosome condensation
cellular macromolecular complex subunit organization
chromatin assembly
nucleosome assembly
mitotic sister chromatid segregation
cell division
DNA packaging
sister chromatid segregation
histone modification
cell cycle process
M phase
histone acetylation
chromosome segregation
covalent chromatin modification
protein amino acid acetylation
Chromatin modification
Signal transduction
elevation of cytosolic calcium ion concentration
embryonic pattern specification
calcium ion transport
cell surface receptor linked signal transduction
transmembrane receptor protein tyrosine kinase signaling pathway
positive regulation of lymphocyte differentiation
regulation of alpha-beta T cell activation
transcription from RNA polymerase II promoter
Development
regulation of nervous system development
neuron fate commitment
lung development
actin filament-based process
regulation of cell development
regulation of neurogenesis
respiratory tube development
growth
protein folding
protein catabolic process
organic anion transport
sodium ion transport
Protein folding
Ion transport
negative regulation of macromolecule metabolic process
regulation of mRNA stability
epidermis development
regulation of RNA stability
ectoderm development
RNA stability
Ectoderm development
IDR Clusters
Positive
Negative
Polar
Hydrophobic
Special

Fig. 5 Functional groups of IDR clusters based on chemical composition. Gene Ontology Biological Process terms enriched in proteins containing IDRs in 5 types of clusters. Each IDR cluster shows a distinct function. Blue: positive, red: negative, green: polar, yellow: hydrophobic and pink: special. Nodes correspond to IDR-containing proteins with the specified GO terms. Node size corresponds to the number of proteins with the GO term. Edges indicate the presence of proteins sharing the connected GO terms. Node borders indicate location specific GO terms, black: C terminal, brown: N terminal, dark gray: middle, light gray: term not found in location specific clusters. Additional GO terms shown in Tables S3 and S4 (ESI†).

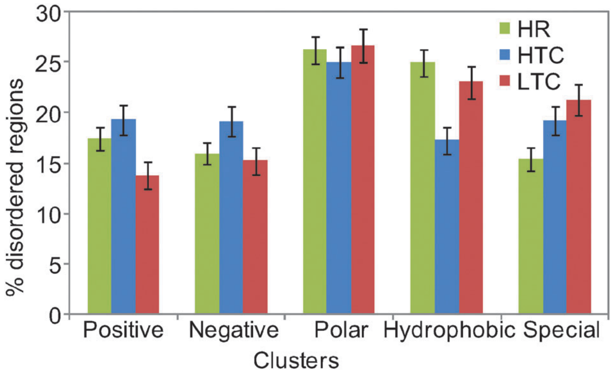


Fig. 6 Distribution of differentially conserved IDRs in composition-based clusters. Fraction of IDRs in each cluster based on chemical composition for IDRs with high residue conservation (HR) in green, IDRs with low residue conservation but high type content conservation (HTC) in blue, and IDRs with low residue and type content conservation (LTC) in red. Error bars indicate 95% confidence intervals.

low sequence conservation, a characteristic chemical composition is important for its function. Further, IDRs with differing levels of conservation and distinct locations show some differences in their composition based functional associations indicating that conservation, chemical composition and location may be used to obtain functional groups of IDRs.

Discussion

It is known that IDRs are poorly conserved compared to ordered regions in proteins.[8] However, the mechanism through which they maintain their function despite rapid changes in their sequence during evolution is not known. In this study, we propose a possible explanation. Our finding that IDRs with low sequence conservation show similar chemical composition within orthologous proteins suggests that preservation of the overall chemical composition may be one of the ways through which IDRs stay functional. The maintenance of chemical composition in IDRs has been suspected.[14,29] However, the extent of its prevalence has not been previously studied. The results of this study indicate that this is a frequent phenomenon in IDRs. Our results also show that a reduced amino acid alphabet can be used to describe disordered regions as has been previously observed.[30] In order to eliminate any bias as a result of disorder prediction, we confirmed our results in experimentally identified IDRs and those predicted by another disorder predictor.

Our results show the presence of three types of IDRs based on conservation:

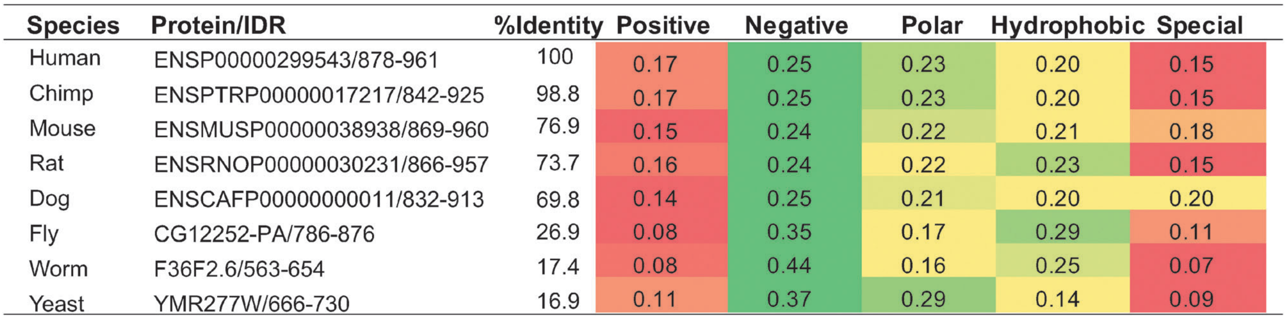
(1) IDRs with high sequence conservation (including those with high residue type conservation) whose function is most likely a result of folding on binding since a conserved sequence may be a requisite for a specific structure on folding. Fig. S6 (ESI†) shows the multiple sequence alignment of such an IDR in the human tubulin beta-4 chain. This IDR is highly conserved among orthologous proteins, is highly acidic and possibly binds cations.[24]

(2) IDRs with low sequence conservation between orthologs but that maintain their chemical composition and hence their function. While these IDRs are expected to have a function mainly as a result of their characteristic chemical composition, it is likely that they contain small conserved regions in the form of linear motifs. This could explain the weak negative correlation between the residue conservation of these IDRs and their type content difference with orthologous regions. Additionally, in order for the motifs to be functional, it might be necessary to maintain the type content in the disordered region immediately surrounding the motif. This group of IDRs is similar to the "flexible disorder" proposed by Bellay *et al.* that functions in signal transduction.[11] Indeed, their study found an abundance of small liner motifs in flexible IDRs. Small clusters of residues seen in some IDRs may also play an important role in their function. An IDR at the C-terminal tail of the TFIIF-associating CTD phosphatase, CTDP1, falls in this category. A few charged and hydrophobic residues at the very end of the C-terminal region are conserved across all species considered here and are important for binding to the TFIIF, RAP74 (Fig. 7 and Fig. S6, ESI†).[31] Though this IDR shows significant conservation among mammals, it is poorly conserved in lower eukaryotes. However, as indicated by the type content of the IDR in orthologous proteins, despite poor sequence conservation, the overall chemical composition of the IDR is preserved with a high abundance of negatively charged residues, depletion of positively charged and special residues and average amounts of hydrophobic and polar residues.

(3) IDRs with neither sequence conservation nor maintenance of chemical composition. These IDRs are sequences abundant in Pro, Gly and Ala with interspersed charged residues. Based on the co-occurrence of Ala with Pro and Gly residues in these IDRs, Ala appears to be more closely related to Pro and Gly than other hydrophobic residues, at least within disordered regions, and should be classified as a special residue in future studies. It is not clear how these IDRs maintain their function. It is possible that they do so by maintaining the abundance of Pro and Gly residues which maintain the disorderliness. The high conservation of charged residues demonstrates their importance in the function of LTC IDRs. In a separate study, we observed that repetitive and non-repetitive disordered regions in proteins encoded by nucleotide repeats are abundant in Pro, Gly and Ala residues that are poorly conserved (unpublished results). This raises the possibility that LTC IDRs may be encoded by highly mutable repetitive regions in DNA. However, these IDRs need to be further studied.

The frequent residue type conservation in IDRs over residue conservation alone is similar to the behavior of ordered domains. While we have focused on the properties of poorly conserved IDRs, it will be interesting to see if chemical composition is similarly maintained in poorly conserved ordered regions in disordered as well as ordered proteins, in future studies.

Apart from the classification of IDRs based on conservation, we also propose a broad classification scheme based on their type content. The separation of IDRs into functionally distinct groups based on their type content emphasizes the importance of their chemical composition in their function. However, this classification does not overlap with the flavors of disordered regions,[19] or that based on charge and hydropathy alone,[20] that have been previously proposed. The results here present the possibility of classifying IDRs into functional



Species	Protein/IDR	%Identity	Positive	Negative	Polar	Hydrophobic	Special
Human	ENSP00000299543/878-961	100	0.17	0.25	0.23	0.20	0.15
Chimp	ENSPTRP00000017217/842-925	98.8	0.17	0.25	0.23	0.20	0.15
Mouse	ENSMUSP00000038938/869-960	76.9	0.15	0.24	0.22	0.21	0.18
Rat	ENSRNOP00000030231/866-957	73.7	0.16	0.24	0.22	0.23	0.15
Dog	ENSCAFP00000000011/832-913	69.8	0.14	0.25	0.21	0.20	0.20
Fly	CG12252-PA/786-876	26.9	0.08	0.35	0.17	0.29	0.11
Worm	F36F2.6/563-654	17.4	0.08	0.44	0.16	0.25	0.07
Yeast	YMR277W/666-730	16.9	0.11	0.37	0.29	0.14	0.09

Fig. 7 Example of an HTC IDR – CTDP1. Percent identity and type content of the C-terminal intrinsically disordered region in the TFIIF-associating CTD phosphatase, CTDP1 (ENSP0000029954, Disprot id: DP00177) in human and 7 other species. The fractions of residues are colored with red denoting a low propensity, green denoting a high propensity and yellow indicating an average value in rows according to their abundance. The IDR shows conserved chemical composition with relative amounts of different residue types (high negative, low positive and special, medium polar and hydrophobic content) being maintained between orthologs despite poor sequence identity. Complete sequence alignment is shown in Fig. S6 (ESI†).

groups based on chemical composition. Including location information into the classification scheme provides better functional separation of IDRs. While the classification shown here is simplistic, it demonstrates the utility of chemical composition as a means of classification. Undoubtedly, there are more than five types of IDRs with varying combinations of residue type content, though they are not represented in this study. For instance, in both HR and HTC IDRs, but to a greater extent in HTC IDRs, negative residues are frequently enriched in IDRs within the positive cluster (Fig. S5, ESI†). Introducing a cluster specifically for such IDRs will help in better identifying their functions. Thus, adding more complexity into the classification with combinations of residue types will provide better groups of IDRs. The location preference of IDRs suggests this as an additional feature that may be used for classification of IDRs.

However, several issues need to be addressed. The alignment of disordered regions is inherently difficult due to poor sequence conservation and low sequence complexity. We have tried to overcome this problem by performing multiple alignments of whole proteins (ordered and disordered regions) instead of the IDRs alone hypothesizing that the alignments of disordered regions would be improved by those of proximal ordered and conserved regions. We also show that the scores based on alignments used in this study are significantly better than those obtained from random alignments. Additionally, our finding that the chemical composition within a section of IDRs with poor sequence conservation (HTC IDRs) is maintained within orthologs indicates that these alignments are reliable. This could be the result of a few conserved residues within the disordered region, as demonstrated in the multiple sequence alignment of CTDP1 and its orthologs (Fig. S6B, ESI†). However, there is no way to judge the accuracy of the alignments in the LTC IDRs which do not show conservation of sequence or chemical composition. It is possible that these IDRs are so poorly conserved that they cannot be accurately aligned to their orthologous regions using sequence-based alignment algorithms. This in turn would hinder the identification of common sequence or composition patterns within these IDRs. Another possibility is that a multiple sequence alignment program other than CLUSTALW[21] might perform better with disordered regions. However, a comparison of CLUSTALW with MUSCLE[32] in the alignment of disordered regions does not show considerable differences in the conservation scores (unpublished results). Thus, the conclusions of this study regarding the preservation of chemical composition may be applicable only to IDRs that can be successfully aligned using the current methods. Future studies on IDRs will need to develop special alignment tools, perhaps based on chemical composition rather than sequence, to address this issue. An additional concern is the choice of an appropriate cutoff to separate IDRs with high and low residue or type content conservation. There is no consensus on a value of sequence identity that may be considered as a cutoff. However, comparing the differences in the distribution of residue conservation scores between our study and that of Bellay *et al.*, it appears to depend on the species chosen for testing.[11] While the use of average scores is sufficient to separate the IDRs into distinct groups in this study, this issue needs to be addressed in greater detail in the future. Lastly, the classification system presented here is tested on predicted IDRs due to the lack of experimentally determined information and will need to be reconfirmed as the availability of these data increases.[33]

In spite of these concerns, the results of our study suggest an explanation for the maintenance of disorderliness and function in rapidly evolving IDRs. They also suggest a means of improving the function prediction of proteins with large IDRs and few or no known annotated sequence homologs. The current function prediction methods rely heavily on sequence homology at some level[34] and hence are unsuitable for proteins with large IDRs having poor sequence conservation. We have previously shown that amino acid content similarity, instead of sequence similarity, can be used to predict a function associated with an IDR.[13] While the method works better than random, the results here suggest several avenues for improvement, such as location preference. Indeed, location of IDRs has been previously used in the function prediction of proteins with long disordered regions.[35]

Conclusion

We investigated the conservation of amino acid residues and chemical composition in intrinsically disordered regions (IDRs) from human proteins in 7 other eukaryotes. We found that IDRs with poor sequence conservation maintain their chemical composition. This suggests that the overall chemical composition of the IDR is important for its function and is one of the ways through which IDRs maintain their function or disorderliness. Additionally IDRs with different levels of conservation also have differing location preferences and functional enrichments. We also show that IDRs can be classified into five broad groups that are functionally distinct based on their chemical composition. IDRs show specific location preferences based on their conservation and their chemical composition. Finally, the findings of this study demonstrate that conservation, chemical composition and location can be used to distinguish between functionally distinct intrinsically disordered regions and provide a means of improving functional prediction and annotation of these regions.

Materials and methods

Dataset

Two types of datasets were used in this study for the calculation of conservation scores and type content:

(1) *Dataset of predicted disordered regions*: All human proteins were taken from Ensembl.[36] Orthologs of human proteins in the species *P. troglodytes*, *M. musculus*, *R. norvegicus*, *C. familiaris*, *D. melanogaster*, *C. elegans* and *S. cerevisiae* were identified with InParanoid.[37] Unique orthologs within each species were chosen based on the highest InParanoid score or the greatest length, in the event of multiple hits having the same score. Human proteins with at least 4 orthologs were chosen. Orthologous proteins were aligned using ClustalW[21] with the human proteins as a reference. Predicted disordered regions longer than 30 residues were downloaded from Disodb[38] (which contains disorder predictions made by Disopred2[22] at a false positive rate of 5%). This resulted in a set of 6751 human

proteins with 14 612 disordered regions. Disordered regions were also predicted using IUPred[23] which predicted 9464 disordered regions longer than 30 residues with a score greater than 5 in 4287 proteins.

(2) *Dataset of experimentally determined disordered regions*: all disordered regions longer than 30 residues were obtained from DisProt.[24] These were aligned to human proteins with orthologs in 4 or more species and disordered regions were assigned where the sequence similarity was greater than 90%. The resulting 102 disordered regions were then used for the calculation of conservation scores and type content.

Score calculation

(1) *Residue conservation score*: The residue conservation score was calculated using the scoring scheme similar to that proposed by Bellay *et al.*[11] Briefly, residue conservation ($\mathrm{RC_i}$) for each position in the multiple sequence alignment was defined as follows:

$$\mathrm{RC_i} = f\left(\frac{N_\mathrm{i}}{N_\mathrm{ortho}}\right) \quad (1)$$

where,

N_i : number of orthologous proteins with the same residue at position i,

N_ortho : number of orthologous proteins,

$f(x)$: function that returns a value based on x as follows:

X	Return value
$x < 0.1$	1
$0.1 \le x < 0.2$	2
$0.2 \le x < 0.3$	3
$0.3 \le x < 0.4$	4
$0.4 \le x < 0.5$	5
$0.5 \le x < 0.6$	6
$0.6 \le x < 0.7$	7
$0.7 \le x < 0.8$	8
$0.8 \le x$	9

Average residue conservation of each disordered region was calculated as follows:

$$\mathrm{RC} = \frac{\sum \mathrm{RC_i}}{L} \quad (2)$$

where,

RC : average residue conservation of the disordered region,

$\mathrm{RC_i}$: residue conservation at position i in the disordered region,

L : length of the disordered region.

Gap positions in the reference (human) disordered region were not included in the final score of the disordered region. Gap positions in the aligned orthologous regions were assigned a 0 score and did not contribute to the final conservation score at that position.

(2) *Residue type conservation score*: Residues in the disordered regions and the aligned regions in orthologous proteins were replaced with a residue type as described in Table 1. The residue type conservation score was calculated in the manner similar to eqn (1) as follows:

$$\mathrm{RTC_i} = f\left(\frac{N_\mathrm{i}}{N_\mathrm{ortho}}\right) \quad (3)$$

where,

N_i : number of residues with the same type at position i in orthologous proteins,

N_ortho : number of orthologous proteins,

$f(x)$: function that returns a value based on x as described in eqn (1).

The average residue type conservation score of a disordered region of length L was calculated as follows:

$$\mathrm{RTC} = \frac{\sum \mathrm{RTC_i}}{L} \quad (4)$$

where,

RTC : average residue type conservation of a disordered region,

$\mathrm{RTC_i}$: residue type conservation at position i in the disordered region,

L : length of the disordered region.

Gaps in alignments were handled as previously described.

(3) *Type content and conservation*: For each type as described in Table 1, the proportion of residue type T in a disordered region was defined as follows:

$$P_\mathrm{T} = \frac{N_\mathrm{T}}{L} \quad (5)$$

where,

P_T : proportion of residue type T in a disordered region,

N_T : number of occurrences of residue type T,

L : length of the disordered region.

Type content (TC) of a disordered region was defined as:

$$\mathrm{TC} = (P_L, P_\mathrm{P}, \cdots, P_\mathrm{S}) \quad (6)$$

where,

P_T : proportion of residue type T based on Table 1.

Type content conservation was determined as the average Euclidean distance between the type content of orthologous IDRs. The Euclidean distance of the type content of a disordered region in protein 1 and orthologous protein 2 was defined as:

$$\mathrm{d}(\mathrm{TC_1}, \mathrm{TC_2}) = \sqrt{\sum_\mathrm{T} ({P_\mathrm{T}}^1 - {P_\mathrm{T}}^2)^2} \quad (7)$$

where,

${P_\mathrm{T}}^1$: proportion of residue type T in protein 1,

${P_\mathrm{T}}^2$: proportion of residue type T in an orthologous disordered region in protein 2.

The residue type content score of a disordered region was defined as the average of Euclidean distance of residue type content between the reference human protein and all orthologous proteins having an aligned disordered region:

$$\mathrm{TCD} = \frac{\sum_{N_\mathrm{ortho}} \mathrm{d}(\mathrm{TC_R}, \mathrm{TC_i})}{N_\mathrm{ortho}} \quad (8)$$

where,

$\mathrm{TC_R}$: residue type content of reference (human) protein,

$\mathrm{TC_i}$: residue type content of aligned ortholog i,

N_{ortho} : number of orthologous proteins.

Gaps were not included in this scoring scheme.

Random alignments were made by selecting 1000 IDRs from human and mouse, aligning their termini and randomly inserting gaps in one or both IDRs. Residue, residue type and type content conservation scores were then calculated for these random alignments according to eqn (1)–(8).

(4) *Grouping of disordered regions*: Disordered regions were divided into the following categories:

(i) High residue conservation (HR): IDRs with a residue conservation score greater than average.

(ii) Low residue conservation (LR): IDRs with a residue conservation score less than average.

(iii) Low residue high type conservation (LRHT): IDRs with a residue conservation score less than average and a residue type conservation score greater than average. These IDRs were not studied separately due to their similarity to HR IDRs.

(iv) Low residue low type conservation (LRLT): IDRs with a residue conservation score less than average and a residue type conservation score less than average.

(v) High type content conservation (HTC): LRLT IDRs with type content distance less than the average plus two standard deviations of type content score of HR IDRs. A smaller distance between orthologous IDRs indicates a greater similarity in type content.

(vi) Low type content conservation (LTC): LRLT IDRs with type content distance greater than the average plus two standard deviations of type content score of HR IDRs. A greater distance between orthologous IDRs indicates a lower similarity in type content.

Tables S1 and S2 (ESI†) provide the average values used and the number of IDRs in each group. Statistical significance of the differences in the groups was calculated using the Wilcoxon rank sum test.

(5) *Location preference*: IDRs within 40 residues of the terminal regions were assigned to the N- or C-terminal. The remaining IDRs were denoted as middle or non-terminal. Statistical significance for the differences in location preference was calculated using the Hypergeometric distribution.

(6) *Amino acid propensity*: The amino acid propensity of a group of IDRs was calculated as follows.

The proportion of amino acid X in i IDRs within group G was calculated as:

$$P_{\text{GX}} = \frac{\sum_{\text{G}i} N_{\text{X}i}}{\sum_{\text{G}i} L_i} \tag{9}$$

where,

P_{GX} : proportion of amino acid X in all disordered regions in group G,

$N_{\text{X}i}$: number of occurrences of amino acid X in the ith disordered region in group G,

L_i : length of the ith disordered region in group G.

The proportion of amino acid X in all disordered regions was calculated as:

$$P_{\text{AX}} = \frac{\sum_i N_{\text{X}i}}{L_i} \tag{10}$$

where,

P_{AX} : proportion of amino acid X in all disordered regions,

$N_{\text{X}i}$: number of occurrences of amino acid X in the ith disordered region,

L_i : length of the ith disordered region.

Hence, the propensity of an amino acid X was calculated as:

$$\text{Prop}_{\text{X}} = \frac{P_{\text{GX}} - P_{\text{AX}}}{P_{\text{AX}}} \tag{11}$$

(7) *Clustering*: For each disordered region i, a Z-score was calculated for its type content in category T (as defined in Table 1) as follows:

$$Z_{\text{T}i} = \frac{P_{\text{T}i} - P_{\text{Tavg}}}{P_{\text{Tstd}}} \tag{12}$$

$Z_{\text{T}i}$: Z-score for content of type T in disordered region i based on Table 1,

$P_{\text{T}i}$: proportion of residues of type T in disordered region i,

P_{Tavg} : average proportion of residues of type T in all disordered regions considered,

P_{Tstd} : standard deviation for the proportion of residues of type T in all disordered regions considered.

All IDRs were then clustered using the z-scores for the content of each residue type with the k-means algorithm in R. Since we were specifically interested in IDRs with an over-representation of one of the 5 types (positive, negative, polar, hydrophobic and special), a cluster count of 5 was assigned during clustering. The default distance metric (Euclidean distance) was used for the clustering. Chemical composition based clusters were further divided into 3 groups based on the location of the IDRs (N-terminal, middle and C-terminal).

(8) *Gene ontology term enrichment*: Proteins containing IDRs within each group were analyzed using DAVID[39] for GO term enrichment. Similarly, for IDRs within type based clusters, proteins were identified and GO term enrichment analysis was performed using DAVID. GO terms unique to each group/cluster were selected for the figures and tables. Gene ontology enrichment maps were drawn in Cytoscape[40] using the Enrichment Map tool.[41]

(9) *Identification of short conserved regions in long IDRs*: HTC IDRs longer than 200 residues were selected for the identification of short functional sub-regions since they have a greater probability of having such short regions given that 90% of IDRs were less than 200 residues in length. For each residue in an IDR, we calculated the type content conservation as the average Euclidean distance between the residue content of 50 residues upstream and downstream of the selected residue and that of the aligned region in orthologous proteins. We then identified highly conserved regions as those with 30 or more residues having a content conservation score <0.143 (the lowest type content distance threshold for HTC IDRs). To identify highly conserved sub-regions, we selected those sub-regions that had a score lesser than the parent IDR. The chemical composition for each sub-region was calculated and compared to that of the parent IDR to determine differences in overall content.

Abbreviations

IDR	Intrinsically disordered region
GO	Gene ontology

Author contributions

AP conceived of the project and designed experiments. HAM and SW collected data, performed alignments and calculated conservation scores. AP, HAM and KN analyzed results. AP and HAM wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Prof. Haruki Nakamura, Dr Daron Standley, Dr Shunsuke Teraguchi and Dr Kuo-ching Liang for useful discussions, and Dr Nikolay Dokholyan and the two anonymous reviewers for their helpful suggestions. The supercomputing resource was provided by the Human Genome Center, The Institute of Medical Science, The University of Tokyo.

References

1 L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker, *J. Mol. Biol.*, 2002, **323**, 573–584.

2 A. Patil and H. Nakamura, *FEBS Lett.*, 2006, **580**, 2041–2045.

3 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome. Res.*, 2007, **6**, 1882–1898.

4 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.

5 A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *FEBS J.*, 2005, **272**, 5129–5148.

6 P. E. Wright and H. J. Dyson, *Curr. Opin. Struct. Biol.*, 2009, **19**, 31–38.

7 T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, M. Tyers and J. D. Forman-Kay, *Structure*, 2010, **18**, 494–506.

8 C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams and A. K. Dunker, *J. Mol. Evol.*, 2002, **55**, 104–110.

9 C. J. Brown, A. K. Johnson, A. K. Dunker and G. W. Daughdrill, *Curr. Opin. Struct. Biol.*, 2011, **21**, 441–446.

10 G. W. Daughdrill, P. Narayanaswami, S. H. Gilmore, A. Belczyk and C. J. Brown, *J. Mol. Evol.*, 2007, **65**, 277–288.

11 J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers and P. M. Kim, *Genome Biol.*, 2011, **12**, R14.

12 S. Teraguchi, A. Patil and D. M. Standley, *BMC Bioinf.*, 2010, **11**(Suppl 7), S7.

13 A. Patil, S. Teraguchi, H. Dinh, K. Nakai and D. M. Standley, *Pac. Symp. Biocomput.*, 2012, **17**, 164–175.

14 J. C. Hansen, X. Lu, E. D. Ross and R. W. Woody, *J. Biol. Chem.*, 2006, **281**, 1853–1856.

15 D. Vuzman and Y. Levy, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 21004–21009.

16 J. A. Marsh and J. D. Forman-Kay, *Biophys. J.*, 2010, **98**, 2383–2390.

17 M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, *Nucleic Acids Res.*, 2012, **40**, D290–D301.

18 A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin, *Nucleic Acids Res.*, 2008, **36**, D419–D425.

19 S. Vucetic, C. J. Brown, A. K. Dunker and Z. Obradovic, *Proteins*, 2003, **52**, 573–584.

20 F. Huang, C. Oldfield, J. Meng, W. L. Hsu, B. Xue, V. N. Uversky, P. Romero and A. K. Dunker, *Pac. Symp. Biocomput.*, 2012, 128–139.

21 M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins, *Bioinformatics*, 2007, **23**, 2947–2948.

22 J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, 2004, **337**, 635–645.

23 Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon, *J. Mol. Biol.*, 2005, **347**, 827–839.

24 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–D793.

25 N. V. Dokholyan and E. I. Shakhnovich, *J. Mol. Biol.*, 2001, **312**, 289–307.

26 M. Y. Lobanov, E. I. Furletova, N. S. Bogatyreva, M. A. Roytberg and O. V. Galzitskaya, *PLoS Comput. Biol.*, 2010, **6**, e1000958.

27 S. Rauscher, S. Baud, M. Miao, F. W. Keeley and R. Pomes, *Structure*, 2006, **14**, 1667–1676.

28 J. Banroques, O. Cordin, M. Doere, P. Linder and N. K. Tanner, *J. Mol. Biol.*, 2011, **413**, 451–472.

29 P. Tompa and M. Fuxreiter, *Trends Biochem. Sci.*, 2008, **33**, 2–8.

30 E. A. Weathers, M. E. Paulaitis, T. B. Woolf and J. H. Hoh, *FEBS Lett.*, 2004, **576**, 348–352.

31 B. D. Nguyen, K. L. Abbott, K. Potempa, M. S. Kobor, J. Archambault, J. Greenblatt, P. Legault and J. G. Omichinski, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 5688–5693.

32 R. C. Edgar, *Nucleic Acids Res.*, 2004, **32**, 1792–1797.

33 S. Fukuchi, S. Sakamoto, Y. Nobe, S. D. Murakami, T. Amemiya, K. Hosoda, R. Koike, H. Hiroaki and M. Ota, *Nucleic Acids Res.*, 2012, **40**, D507–D511.

34 T. Hawkins, S. Luban and D. Kihara, *Protein Sci.*, 2006, **15**, 1550–1556.

35 A. Lobley, M. B. Swindells, C. A. Orengo and D. T. Jones, *PLoS Comput. Biol.*, 2007, **3**, e162.

36 P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kahari, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Harrow, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa and S. M. Searle, *Nucleic Acids Res.*, 2012, **40**, D84–D90.

37 G. Ostlund, T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings and E. L. Sonnhammer, *Nucleic Acids Res.*, 2010, **38**, D196–D203.

38 M. M. Pentony and D. T. Jones, *Proteins*, 2010, **78**, 212–221.

39 W. Huang da, B. T. Sherman and R. A. Lempicki, *Nat. Protoc.*, 2009, **4**, 44–57.

40 M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang and T. Ideker, *Bioinformatics*, 2011, **27**, 431–432.

41 D. Merico, R. Isserlin, O. Stueker, A. Emili and G. D. Bader, *PLoS One*, 2011, **5**, e13984.