博士論文

The Conservation and Classification of Intrinsically Disordered Regions in Proteins
（タンパク質の天然変性領域の保存と分類）

ハリー・アムリ・ムサ

HARRY AMRI MOESA

博士論文

The Conservation and Classification of Intrinsically Disordered Regions in Proteins

タンパク質の天然変性領域の保存と分類

Table Of Contents

1. Introduction

Until around a decade ago, it was dominantly believed that a specific function of a protein is determined by its unique three-dimensional (3D) structure. The idea of natively unfolded proteins had challenged the then popular structure-centric viewpoint. However, in recent years, intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are considered as established concepts.

The abundance of IDRs in proteins and their importance in protein function have been extensively studied in the past decade. Intrinsically unstructured proteins are involved frequently in numerous processes in the cell such as in transcriptional activation, cell-cycle regulation, membrane transport, molecular recognition and signaling. Intrinsic disorder might be responsible for the binding diversity of the proteins involved in the broad cascade of protein-protein interactions(1). Moreover, IDRs are important for hub protein function. Previous study showed that hubs have more disordered residues than non-hubs (2) by providing hubs with the ability to bind multiple structurally diverse targets, thus enabling them to regulate multiple networks (3). IDPs with long IDRs are also associated with biological processes through bioinformatics approaches (4).

The complex sets of protein interactions could be understood by incorporating intrinsic disorder in the binding regions of proteins. Intrinsic disorders can serve as the structural basis for hub protein promiscuity, can bind to structured hub proteins, and can provide flexible links between functional domains with the linkers enabling mechanism that facilitate binding diversity (5).

It was also suggested that IDRs also play important role in disorder to order transition upon bindings (6), and contain transient structures to facilitate dynamic complexes (7).

Molecular evolution of disordered proteins differ from ordered proteins, in which they have a different pattern of accepted point mutations, exhibit higher rates of insertions and deletions, and evolve more rapidly than ordered proteins, and often evolve at a fast rate compared to ordered proteins. Despite of that, the conservation of their functions indicates that these proteins have important roles in organisms (8),(9).

The absence of a specific structure or a conserved sequence does not prevent IDRs to stay functional. Previous studies have looked at sequence conversations of IDRs, and divided disordered regions into regions where the underlying amino acid sequence is also conserved (constrained disorder), regions where there appears to be selection on the structural property of disorder itself rather than the specific sequence (flexible disorder), and disordered residues that were not conserved (non-conserved disorder) (10). However, this classification is not sufficient to explain structural conservation phenomenon of disordered regions despite of the absence of sequence conservation.

Previously, our group has shown that using amino acid content similarity can help identify IDRs with similar functions through a thorough statistical scoring of amino acid sequences of mouse proteins. The more similar the content, the more similar the functions, in which it was shown in analysis result of Pfam domain and GO terms (11). This result indicated that similarity of content is one of important parameters in IDRs analysis, thus, as well as overall chemical composition of the IDR.

Evidence is mounting that in many cases there is a direct correlation between amino acid composition, intrinsic disorder, and protein function. For instance, a study on intrinsically disordered tails in DNA-binding proteins indicated that the net charge of disordered tail can both increase the affinity of the globular region to the DNA and facilitate transfer from one DNA molecule to another, thus results in a more efficient DNA search (12). A further study using computational simulation showed that the net charge of intrinsically disordered tail indeed affects the ability to efficiently search DNA for binding regions (13).

These evidences raise the possibility that IDRs evolve to maintain not only their sequence but also their chemical composition.

On the other hand, defining a classification system for IDRs is proved a challenge, as a consequence of the lack of sequence conservation. The current domain classification techniques are based either on structure, such as SCOP (14) or sequence conservation, such as Pfam (15). IDRs are not amenable to these classification techniques, due to their lack of structure and structure conservation. It is necessary to separate them into functional groups, similar to that of conserved domains. A unified classification system is important to improve our understanding of IDRs, and also to facilitate the annotation of uncharacterized proteins with large IDRs and no identifiable sequence homologs.

Several studies have attempted to classify IDRs. A study classified disordered regions in 3 flavors by their compositions, sequence locations, and biological function (16). Another implemented predictor parameters to classify IDRs, based on charge, hydrophobicity, and function distribution (17). However, there is no generally accepted means of classifying IDRs into functional groups based on their sequence.

Our study addressed the two issues:

1) To investigate how the IDRs stay functional in the absence of a specific structure or a conserved sequence, both of which are associated with functional domains.

2) To define a classification system for IDRs that is similar to that of conserved domains.

2. Chemical composition conservation in intrinsically disordered regions

We studied the IDRs using proteins of human and the orthologs from other 7 eukaryotes (chimp, dog, rat, mouse, fly, worm and yeast). For information of disordered regions, we used the information from regions predicted by DisoPred2(18) and IUPred(19), and the information of experimentally determined disordered regions from DisProt(20). Using CLUSTALW(21), we aligned the human proteins with proteins of other 7 eukaryotes and selected orthologous proteins that have more than 4 orthologs, then from those proteins we selected IDRs whose length is longer than 30 residues.

2.1. Conservation in IDRs

We ran the study of conservation in IDRs at three levels (Figure 1).

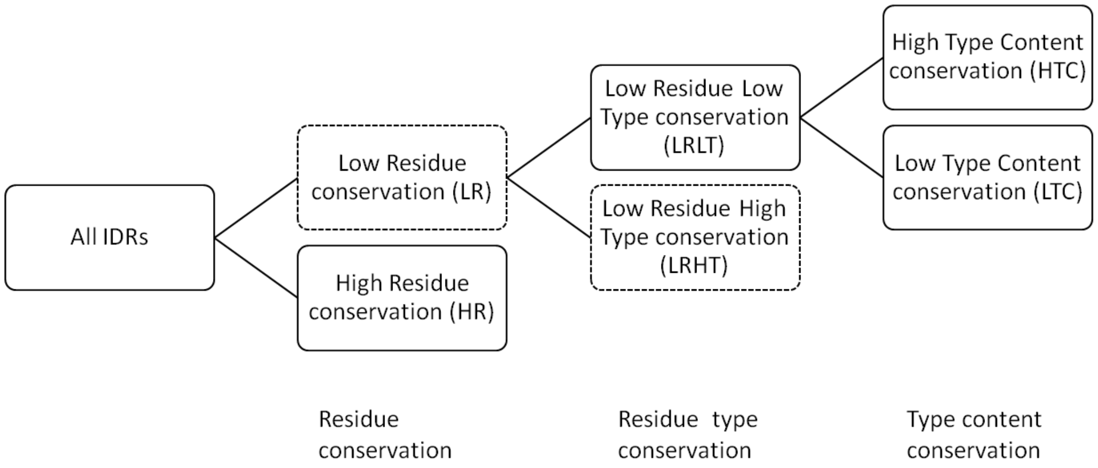

All IDRs
Low Residue conservation (LR)
High Residue conservation (HR)
Low Residue Low Type conservation (LRLT)
Low Residue High Type conservation (LRHT)
High Type Content conservation (HTC)
Low Type Content conservation (LTC)
Residue conservation
Residue type conservation
Type content conservation

Figure 1. Flowchart groups of IDRs based on their residue, residue type and type content conservation. Groups marked in solid lines were specifically studied in this work.

1. *Residue conservation* - The residue conservation score for each IDR within the 8 species was calculated using the scoring scheme proposed by Bellay et al(10). The residue conservation score of an IDR indicates the level of sequence conservation of the IDR within the orthologous proteins (see Methods).

2. *Residue type conservation* - In order to determine if the residue type was more frequently conserved in IDRs compared to the residue itself, we assigned one of five types to each amino acid depending on the nature of its side chain (Table 1). Pro and Gly were assigned to a special category due the nature of their side chains(15). We then calculated the type conservation score of each IDR similar to the residue

conservation score (see Methods) to determine how often a residue type was conserved within the aligned regions in orthologous proteins.

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

Table 1. Residue Types assigned to each amino acid.

3. *Type content conservation* - Type content of an IDR is the fraction of each residue type in the IDR. Type content conservation is the maintenance of the fraction of residue types (described in Table 1) in regions within orthologous proteins that are aligned to the IDRs irrespective of the sequence conservation. 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 Methods). A smaller distance between the human IDR and its orthologs indicates a greater similarity in content and hence a higher type content conservation.

Figure 2A shows the distribution of the residue and residue type conservation scores in all IDRs. IDRs show a greater level of residue type conservation than simple residue conservation ($p << 0.01$). A similar tendency was also identified in experimentally determined IDRs from DisProt, where the average residue type conservation score was greater than the average residue conservation score (Supp. Table S1, $p << 0.01$)

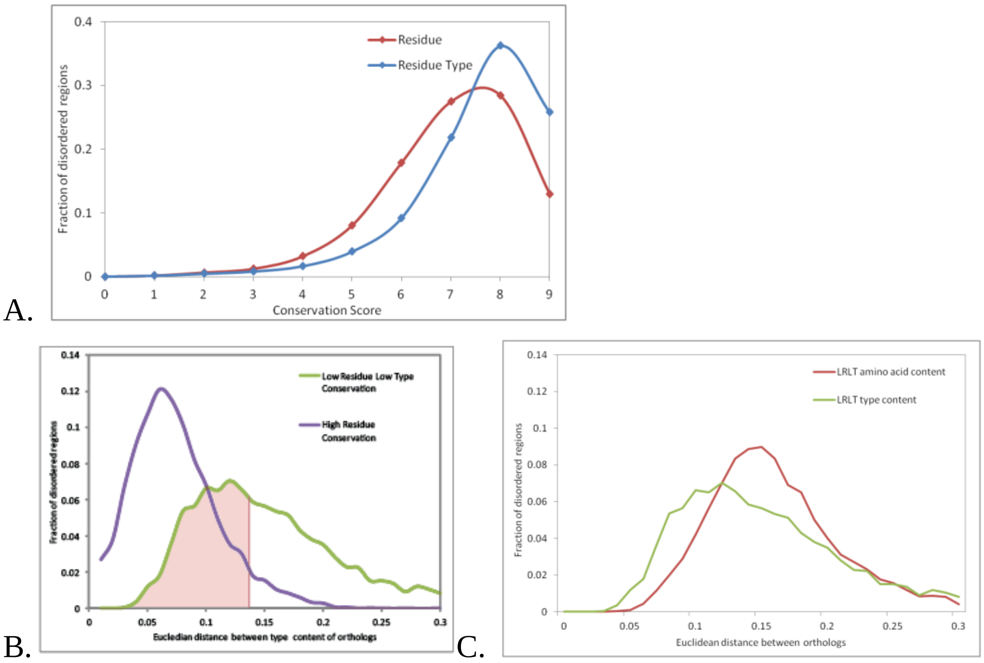


Figure 2. A) Distribution of the residue and residue type (positive, negative, polar, hydrophobic, special) conservation of human intrinsically disordered regions. A greater proportion of disordered regions show residue type conservation. B) Distribution of the residue type content similarity using Euclidean distance for intrinsically disordered regions with high residue conservation (purple) and low residue with low residue type conservation (green). Almost half the IDRs with low residue type conservation have conserved residue type content similar to those with high residue conservation (shaded area). C) Distribution of the average Euclidean distance between the amino acid content and residue type content of orthologous LRLT IDRs.

We then divided the IDRs into two groups (Figure 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 sequence conservation but also poor residue type conservation and thus lack any kind of position dependent conservation.

We compared the type content conservation in IDRs within these two groups (Figure 2B). As expected, IDRs with high residue conservation score i.e. high sequence conservation have greater type content similarity with their orthologs as indicated by the lower average Euclidean distance. 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 from the average Euclidean distance), despite having very low sequence conservation. This indicates that several IDRs with low residue conservation have similar type content within their orthologous regions in other species. This result was confirmed in IDRs from DisProt where 51% of the IDRs with low residue conservation show type content conservation similar to that found in highly conserved IDRs (Supp. Table S2).

Additionally, the type content conservation of LRLT IDRs was better than the amino acid content conservation as indicated by a lower average Euclidean distance between the type content of human IDRs with their respective orthologous regions as compared to that of the amino acid content (Figure 2C, p<<0.01). Thus, some IDRs with low sequence similarity show high residue type content similarity across species at levels better than the amino acid content similarity between the orthologous IDRs. We conclude that IDRs with poor sequence conservation maintain their function and flexibility by preserving their type content and thus their overall nature (charge, hydropathy, polarity and the fraction of Pro and Gly).

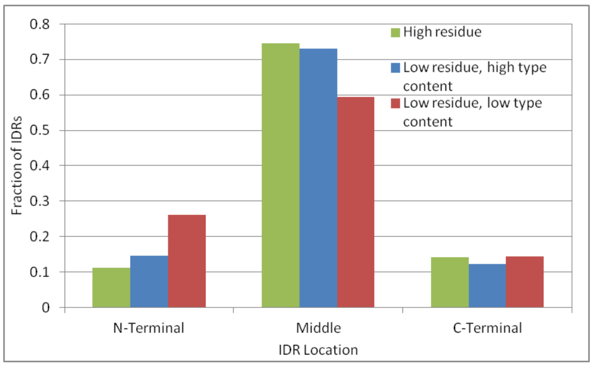
2.2. Characteristics of intrinsically disordered regions with respect to conservation

In order to study the differences in IDRs based on their type content conservation, we further separated the LRLT IDRs into two groups (Figure 1):

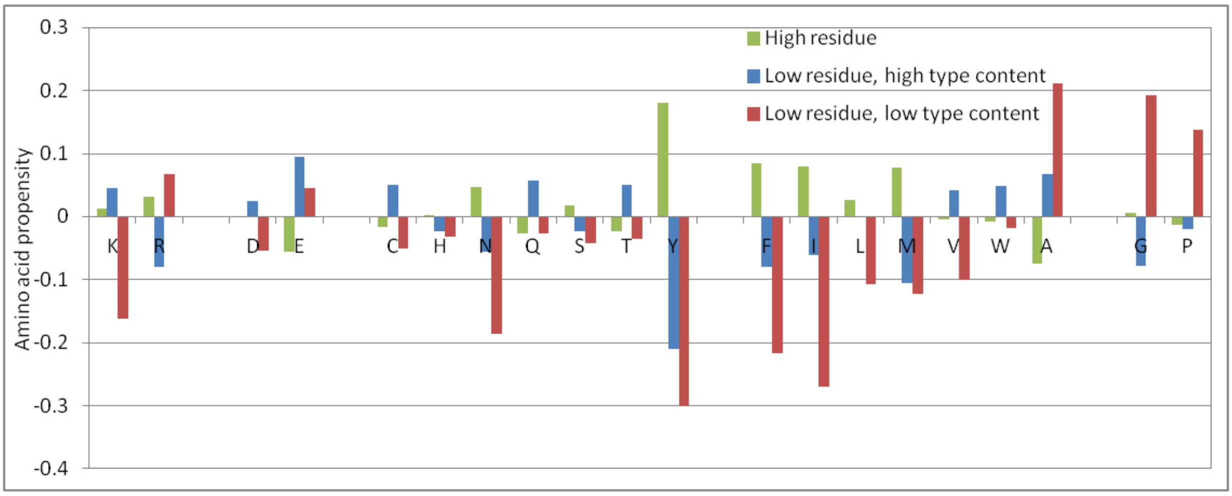
1) IDRs with high type content conservation (HTC) - IDRs with type content conservation that is within two standard deviations of that in HR IDRs.

2) IDRs with low type content conservation (LTC) – IDRs with type content conservation that is lower than that within two standard deviations of HR IDRs.

We compared the preferential location, overall amino acid propensity and functional enrichment of IDRs within these two groups with high residue conservation (HR) IDRs.



A.



B.

Figure 3 A) Location preferences, and B) 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).

IDRs within 40 residues of either terminal of the protein were considered as terminal IDRs. While all IDRs are more likely to be in the center than at either terminal, LTC IDRs are enriched at the N-terminal and depleted in the center of the protein (Figure 3A, p << 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.

The amino acid propensity provides further information about the differences in IDRs in these three groups (Figure 3B). Polar residues, especially Tyr and Ser are abundant in highly conserved IDRs possibly indicating conserved phosphorylation sites. Several hydrophobic residues are also enriched indicating the presence of binding regions. However, Trp is conspicuously depleted in HR IDRs. On the other hand, HTC IDRs show an abundance of charged residues (Glu, Lys) as well as polar residues (Gln) and hydrophobic residues, especially Trp. The charged and polar residues play an important role in defining the overall nature of the IDR and, possibly, its flexibility. The enrichment of certain hydrophobic residues might indicate the presence of small linear motifs within these IDRs. The LTC IDRs show an abundance of not just Pro and Gly, but also Ala and to some extent Arg and Glu. Based on this result, Ala bears a greater similarity to Pro and Gly, at least in disordered regions, than to other hydrophobic residues.

Table 2 shows the results of Gene Ontology term enrichment analysis. IDRs with high residue conservation are enriched in proteins involved in transcription regulation and DNA binding. The abundance of these IDRs in transcription regulation indicates the need for specific sequences for binding other proteins and DNA which are conserved through evolution. On the other hand, HTC IDRs with low residue conservation and high type content are enriched in proteins showing ATP activity and nuclease activity and involved in the regulation of Ras protein signal transduction and DNA replication and repair among others. Finally, the LTC IDRs which show the neither sequence nor type conservation are abundant in proteins enriched in ion binding functionality. The IDRs are often found in peptidases.

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		

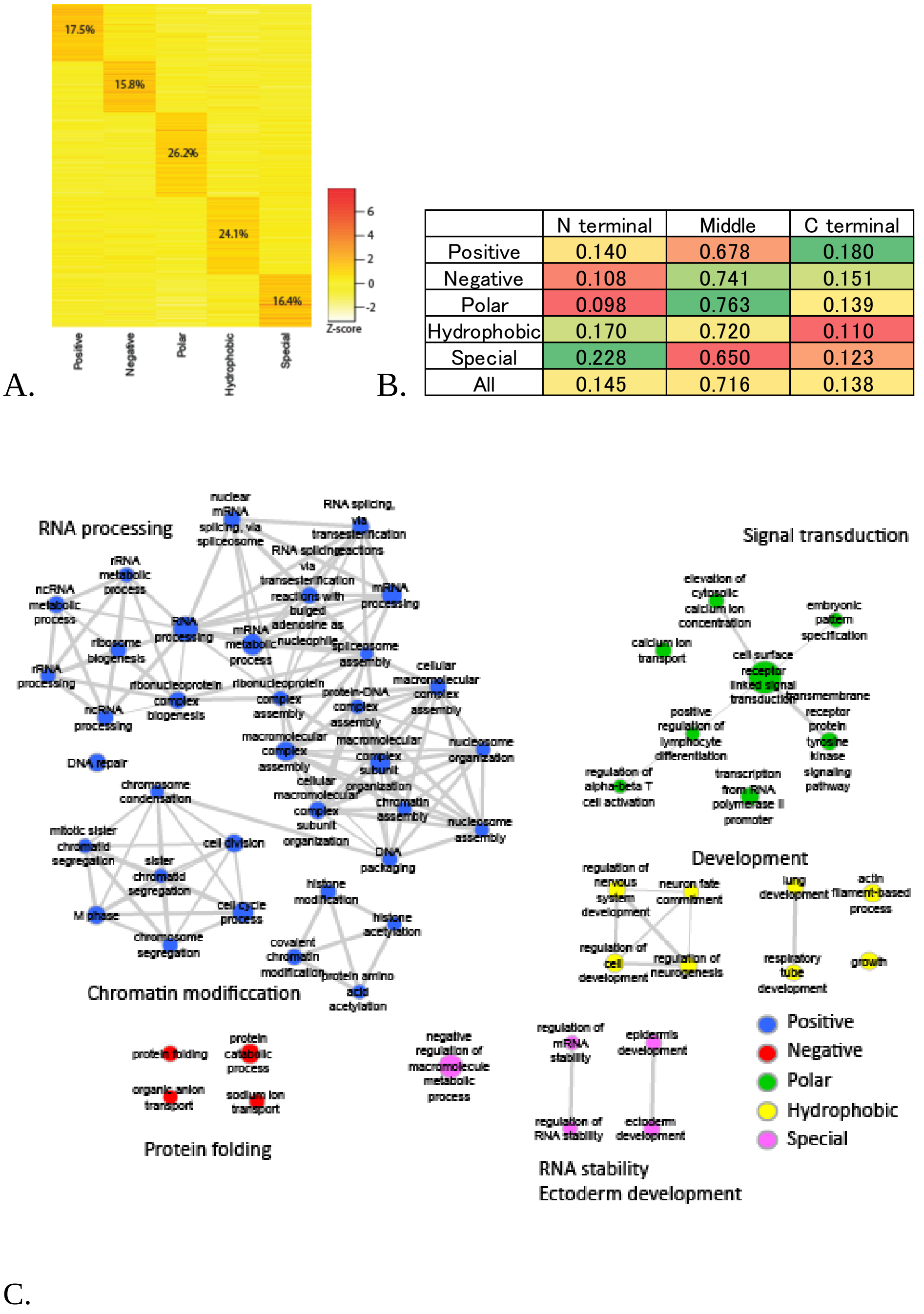
Table 2. Gene Ontology Molecular Function terms enriched in proteins with HR, HTC and LTC IDRs (p <= 0.01, FDR <=1).

2.3. Classification of intrinsically disordered regions based on type content

We attempted to classify IDRs into functionally distinct groups based on type content. We clustered all IDRs 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 above all others. For instance, the positive cluster contained IDRs with the largest fraction of positive residues. As seen in Figure 4A, the IDRs separate into clear clusters based on their residue type content. IDRs with an over-representation of polar residues form the largest cluster.

We checked the location preference of these clusters within the protein. Regions in all clusters are most abundant away from the N- and C-terminal regions in the protein (Figure 4B). However, IDRs in hydrophobic and special clusters are located at the N-terminal more frequently than expected ($p << 0.01$). On the other hand, IDRs in the positive cluster are more likely to be located at the C-terminal ($p << 0.01$). Further, polar and negative regions are under-represented at the terminal regions ($p < 0.01$).

Functionally, IDRs separate very clearly based on their type content with an over-representation of a particular residue type associating with a specific function (Figure 4C). IDRs in the positive cluster are related to RNA and chromatin binding which often appear to be C-terminal tails, based on location preference results (Figure 4B). These include several proteins from the DEAD box helicase family.



	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

Figure 4. Clustering of IDRs into 5 clusters based on the residue type content for A) high residue conservation IDRs. Each value in the heatmap indicates a z-score. Clustering is done using k-means. B) Location preference of the IDRs in type based clusters. C) GO Biological Process terms enriched in proteins containing IDRs in 5 types of clusters. Blue: positive, red: negative, green: polar, yellow: hydrophobic, pink: special.

These proteins have positively charged C-terminal tails that enable their non-specific binding to RNA18. Negative IDRs are involved in protein folding and ion transport. These include heat shock proteins and ion exchanger proteins. Polar IDRs are enriched in signal transduction and are generally a part of receptors and kinases. Hydrophobic IDRs are present in proteins acting in 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 are associated with fibrous proteins like keratin and collagen.

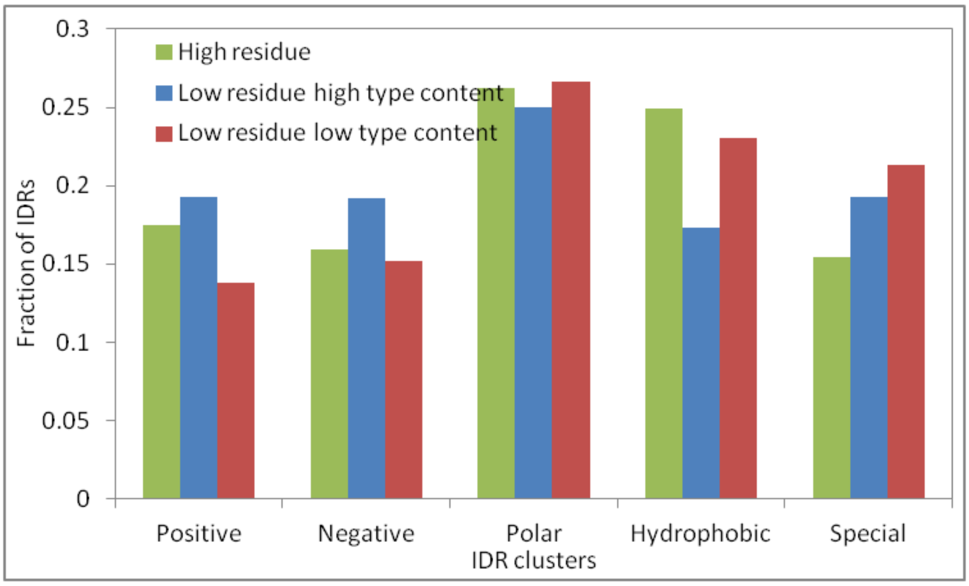


Figure 5. Fraction of IDRs in each type of cluster for HR IDRs (green), HTC IDRs (blue) and LTC IDRs (red). Cluster of each IDRs type and location preferences are shown in Figures S2 and S3.

We also separately clustered HR, HTC and LTC IDRs by type content to determine the impact of conservation on IDR type and function. While all the types of IDRs can be clustered, the fraction of IDRs in each cluster is representative of their overall nature (Figure 5). While HTC IDRs are more likely to be in charged and special clusters, HR IDRs are significantly more in likely to be hydrophobic clusters and slightly more in polar clusters. LTC IDRs are most abundant in the special cluster and under-represented in the charged clusters. The prevalence of IDRs having more than one residue type enriched is also common, especially within the HTC IDRs. IDRs in each group also show essentially the same location preferences with positive IDRs enriched at the C-terminal, hydrophobic and special IDRs at the N-terminal, and polar and negative IDRs away from the terminal regions. GO term enrichments however highlight the differences within these groups (Tables S3-5).

Though all clusters are associated with distinct functions, they are slightly different for each type of IDRs. IDRs in positive clusters for all three groups are enriched in RNA processing functions. Those in the polar cluster in HR and HTC groups are enriched in signal transduction related functions. The IDRs in negative, hydrophobic and special clusters show considerable differences based on conservation. 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 conclude that type content can be used to classify IDRs into functionally distinct groups. However, a classification scheme allowing combination of various residue types is more suitable.

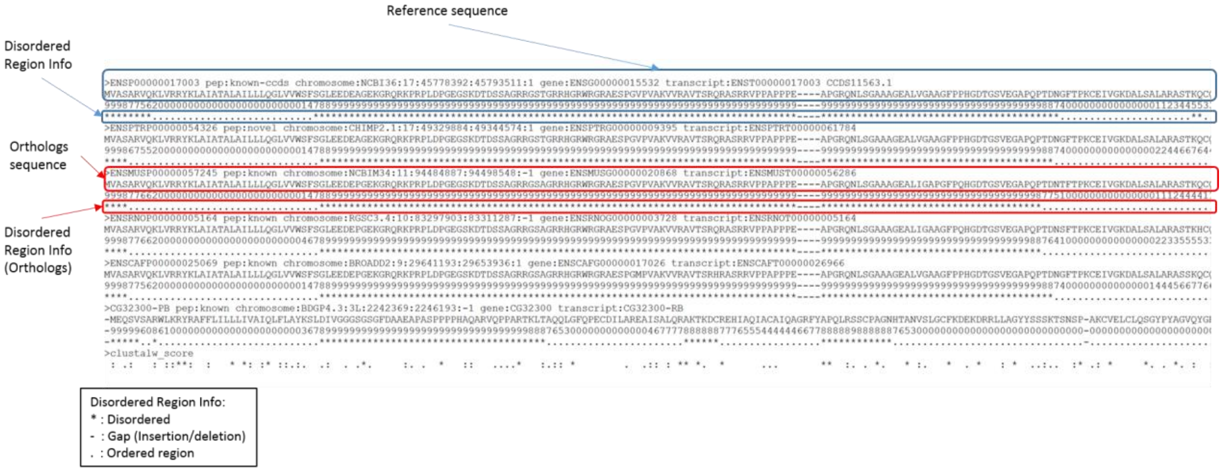
2.4. Material and Methods

2.4.1. Dataset

We used two datasets in this study for the calculation of conservation scores and type content:

1. Dataset of predicted disordered regions: Proteins with predicted disordered regions were downloaded from Disodb(26) for H. sapiens, P. troglodytes, M. musculus, R. norvegicus, C. familiaris, D. melanogaster, C. elegans and S. cerevisiae. Disodb contains disordered regions predicted by Disopred2(18) with a false positive rate of 5%. Groups of orthologous proteins within the 8 species were determined with InParanoid(27) using the human proteins as reference. 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. Orthologous proteins were aligned using ClustalW(21) with the human proteins as reference. And then, we mapped the disordered regions retrieved from Disodb to the result of the alignment. We used human proteins with at least 4 orthologs and one or more disordered regions longer than 30 residues. This resulted in a set of 6751 human proteins with 14612 disordered regions longer than 30 residues.

Figure 6 shows the fraction of the result of alignment along with the disordered regions information.



```
>ENSP00000017003 pep:known-ccds chromosome:NCBI36:17:45778392:45793511:1 gene:ENSG00000015532 transcript:ENST00000017003 CCDS11563.1
MVASARVQKLVRRYKLAIATALAILLLQGLVVWSFSGLEEDEAGEKGRQRKPRPLDPGEGSKDTDSSAGRRGSTGRRHGRWGRAESPGVPVAKVVRAVTSRQRASRRVPPAPPPE----APGRQNLSGAAAGEALVGAAGFPPHGDTGSVEGAPQPTDNGFTPKCEIVGKDALSALARASTKQC(
99987756200000000000000000000000014788999999999999999999999999999999999999999999999999999999999999999999999999999----99999999999999999999999999999999999998874000000000000000011234455 3
********...........................********************************************************************************----****************************************.......................**.
>ENSPTRP00000054326 pep:novel chromosome:CHIMP2.1:17:49329884:49344574:1 gene:ENSPTRG00000009395 transcript:ENSPTRT00000061784
MVASARVQKLVRRYKLAIATALAILLLQGLVVWSFSGLEEDEAGEKGRQRKPRPLDPGEGSKDTDSSAGRRGSTGRRHGRWGRAESPGVPVAKVVRAVTSRQRASRRVPPAPPPE----APGRQNLSGAAAGEALVGAAGFPPHGDTGSVEGAPQPTDNGFTPKCEIVGKDALSALARASTKQC(
99986755200000000000000000000000014788999999999999999999999999999999999999999999999999999999999999999999999999999----9999999999999999999999999999999999999887400000000000000002244676 4
****...............................********************************************************************************----***************************************...........................
>ENSMUSP00000057245 pep:known chromosome:NCBIM34:11:94484887:94498548:-1 gene:ENSMUSG00000020868 transcript:ENSMUST00000056286
MVASARVQKLVRRYKLAIATALAILLLQGLVVWSFSGLEEDEPGEKGRQRKPRPLDPGEGSKDTDSSAGRRGSAGRRHGRWGRAESPGVPVAKVVRAVTSRQRASRRVPPAPPPE----APGRQNLSGAAAGEALIGAPGFPQHGDTGSVEGAPQPTDNTFTPKCEIVGKDALSALARASTKQC(
99987766200000000000000000000000014788999999999999999999999999999999999999999999999999999999999999999999999999999----999999999999999999999999999999999999877510000000000000001112444441
****...............................********************************************************************************----*************************************.............................
>ENSRNOP00000005164 pep:known chromosome:RGSC3.4:10:83297903:83311287:-1 gene:ENSRNOG00000003728 transcript:ENSRNOT00000005164
MVASARVQKLVRRYKLAIATALAILLLQGLVVWSFSGLEEDEPGEKGRQRKPRPLDPGEGSKDTDSSAGRRGSAGRRHGRWGRAESPGVPVAKVVRAVTSRQRASRRVPPAPPPE----APGRQNLSGAAAGEALIGAAGFPQHGDTGSVEGAPQPTDNSFTPKCEIVGKDALSALARASTKHC(
99987766200000000000000000000000046789999999999999999999999999999999999999999999999999999999999999999999999999999----99999999999999999999999999999999999988764100000000000000022335555 3
****...............................********************************************************************************----****************************************..........................
>ENSCAFP00000025069 pep:known chromosome:BROADD2:9:29641193:29653936:1 gene:ENSCAFG00000017026 transcript:ENSCAFT00000026966
MVASARVQKLVRRYKLAIATALAILLLQGLVVWSFSGLEEDEPGEKGRQRKPRPLDPGEGSKDTDSSAGRRGSAGRRHGRWGRAESPGMPVAKVVRAVTSRHRASRRVPPAPPPE----APGRQNLSGAAAGEALVGAAGFPPHGDTGSVEGAPQPTDNGFTPKCEIVGKDALSALARASSKQC(
99987756200000000000000000000000014788999999999999999999999999999999999999999999999999999999999999999999999999999----999999999999999999999999999999999999988740000000000000001444566776
****...............................********************************************************************************----******************************************........................
>CG32300-PB pep:known chromosome:BDGP4.3:3L:2242369:2246193:-1 gene:CG32300 transcript:CG32300-RB
-MEQSVSARWLKRYRAFFLILLLIVAIQLFLAYKSLDIVGGGSGSGFDAAEAPASPPPPHAQARVQPPARTKLTAQQLGFQPECDILAREAISALQRAKTKDCREHIAQIACAIQAGRFYAPQLRSSCPAGNHTANVSLGCFKDEKDRRLLAGYYSSSKTSNSP-AKCVELCLQSGYPYAGVQYGI
-99999608610000000000000000000000367899999999999999999999999999999999988876530000000000000046777788888877765554444446677888889888888765300000000000000000000000000000-000000000000000000000
-*****..*..........................****************************************......................******.................************................................-.....................
>clustalw_score
 :  .:    :  ::**:   :     *  *: :*  ::.:.   .:   .  .*.       :. .   *    ::    ....*    :.:: *         .  .:: :  ** *.   *        ...         **   :. . *.    * :.    *   . *    : *   ..:..   :* .: *       *. . *. :
```

Fig 6. The fraction of aligment result used in calculation of scores for IDRs grouping.

2. Dataset of experimentally determined disordered regions: All disordered regions longer than 30 residues were obtained from DisProt(20). These were aligned to the 6751 human proteins in the above dataset and disordered regions were assigned where the sequence similarity was greater than 90%. These disordered regions were then used for the calculation of conservation scores and type content.

2.4.2. Score calculation

1. Residue conservation score: The residue conservation score was calculated using the scoring scheme proposed by Bellay et al(10). Briefly, residue conservation (RCi) for each position was defined as follows:

$$RC_i = f(\frac{N_i}{N_{ortho}})$$

(1)

where,

Ni : Number of orthologous proteins with the same residue at position i

Northo : Number of orthologous proteins

f(x) : Function that returns value based on x as follows:

x	Return value
x ‹ 0.1	1
0.1 ≤ x ‹ 0.2	2
0.2 ≤ x ‹ 0.3	3
0.3 ≤ x ‹ 0.4	4
0.4 ≤ x ‹ 0.5	5
0.5 ≤ x ‹ 0.6	6
0.6 ≤ x ‹ 0.7	7
0.7 ≤ x ‹ 0.8	8
0.8 ≤ x	9

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

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

where,

RC : Average residue conservation of disordered region

RCi : Residue conservation at position i in the disordered region

L : Length of the disordered region

Regarding gaps in orthologs, we counted them as residues different to residues in reference sequence.

2. Residue type conservation score: Residue in the disordered region 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 Equation 1 as follows:

$$\mathrm{TC_i} = \mathrm{f}(\frac{\mathrm{N_i}}{\mathrm{N_{ortho}}})$$

(3)

where,

Ni : Number of residues with the same type in position i on orthologous proteins

Northo : Number of orthologous proteins

f(x) : Function that returns value based on x as described in Equation 1

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

$$\mathrm{TC} = \frac{\sum \mathrm{TC_i}}{\mathrm{L}} \qquad (4)$$

where,

TC : Average residue type conservation of a disordered region

TCi : Residue type conservation at position i in the disordered region

L : Length of the disordered region

Regarding gaps in orthologs, we counted them as residues different to residues in reference sequence.

3. Amino acid content and conservation: Amino acid content (AAC) of a disordered region was defined as:

$$\mathrm{AAC} = (\mathrm{P_A}, \mathrm{P_D}, \cdots, \mathrm{P_V}) \quad (5)$$

where,

$\mathrm{P_X}$: Proportion of amino acid X

Euclidean distance of amino acid content of a disordered region in protein 1 and its ortholog, protein 2 was calculated as:

$$\mathrm{d}(\mathrm{AAC_1}, \mathrm{AAC_2}) = \sqrt{\sum\nolimits_{\mathrm{X}} (\mathrm{P_X^1} - \mathrm{P_X^2})^2} \quad (6)$$

where,

$\mathrm{P_X^1}$: Proportion of amino acid X in disordered region in protein 1

$\mathrm{P_X^2}$: Proportion of amino acid X in orthologous disordered region in protein 2

Amino acid content score (ACD) of a disordered region was defined as the average of Euclidean distance of amino acid content between the reference human protein and all orthologous proteins with regions aligned to the disordered region:

$$\mathrm{ACD} = \frac{\sum_{\mathrm{N_{ortho}}} \mathrm{d}(\mathrm{AAC_R}, \mathrm{AAC_i})}{\mathrm{N_{ortho}}}$$

(7)

where,

$\mathrm{AAC_R}$: Amino acid content of IDR in reference (human) protein

$\mathrm{AAC_i}$: Amino acid content of aligned region ortholog i

$\mathrm{N_{ortho}}$: Number of orthologous proteins

4. 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:

$$\mathrm{P_T} = \frac{\mathrm{N_T}}{\mathrm{L}}$$

(8)

where,

$\mathrm{P_T}$: Proportion of residue type T in a disordered region

$\mathrm{N_T}$: Number of occurrences of residue type T

L : Length of disordered region

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

$$\mathrm{TC} = (\mathrm{P_L}, \mathrm{P_P}, \cdots, \mathrm{P_S}) \qquad (9)$$

where,

$\mathrm{P_T}$: Proportion of residue type T based on Table 1

Euclidean distance in the type content of a disordered region in ortholog 1 and ortholog 2 was defined as:

$$\mathrm{d}(\mathrm{TC}_1, \mathrm{TC}_2) = \sqrt{\sum_{\mathrm{T}} (\mathrm{P}_{\mathrm{T}}^1 - \mathrm{P}_{\mathrm{T}}^2)^2}$$

(10)

where,

$\mathrm{P}_{\mathrm{T}}^1$: Proportion of residue type T in protein 1

$\mathrm{P}_{\mathrm{T}}^2$: Proportion of residue type T in orthologous disordered region in protein 2

Residue type content score of a disordered region is defined as the average of Euclidean distance of amino acid type content between the reference human protein and all orthologous proteins having an aligned disordered region:

$$\mathrm{TCD} = \frac{\sum_{\mathrm{N_{ortho}}} \mathrm{d}(\mathrm{TC_R}, \mathrm{TC_i})}{\mathrm{N_{ortho}}} \qquad (11)$$

where,

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

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

$\mathrm{N_{ortho}}$: Number of orthologous proteins

5. Grouping of disordered regions: Disordered regions were divided into the following categories:

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

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

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

iv) 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.

v) 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.

The following table shows the fraction of IDRs based on their groups of dataset with disordered regions retrieved from Disodb.

HR			8075	55%
LR 6537	LRHT		773	5%
	LRLT 5764	HTC	2722	19%
		LTC	3042	21%
Total			14612	100%

Tables S1 and S2 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.

6. 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.

7. Clustering: For each disordered region i, a Z-score was calculated for its type content in category T (as defined in Table1) as follows:

$$Z_{\mathrm{Ti}} = \frac{P_{\mathrm{Ti}} - P_{\mathrm{Tavg}}}{P_{\mathrm{Tstd}}} \qquad (12)$$

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

P_{Ti} : 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

The IDRs in HR and HTC groups were then clustered using the z-scores for type content with the k-means algorithm in R. Since IDRs with an over-representation of one of the 5 types (positive, negative, polar, hydrophobic, special) were desired, a cluster count of 5 was defined during the k-means. The default distance metric (Euclidean distance) was used for the clustering.

8. Gene Ontology term enrichment: Proteins containing IDRs within each group were analyzed using DAVID(28) for GO term enrichment. Similarly, for IDRs with type based clusters, proteins were identified and GO term enrichment 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(29) using the Enrichment Map tool(30).

3. Discussion and Conclusion

3.1. Discussion

Intrinsically disordered regions are poorly conserved compared to ordered regions in proteins. However, the mechanism through which they maintain their function despite rapid evolution in their sequence is not known.

We propose a possible explanation in this thesis. We found that to stay functional, IDRs probably preserved the overall chemical composition. This is indicated by our discovery that IDRs with low sequence conservation show similar chemical composition within orthologous proteins.

Previously, chemical composition maintenance in IDRs has been suspected (22). However, the degree of its commonness across eukaryotes has not been previously investigated. Our study indicates that chemical composition maintenance is a frequent phenomenon in IDRs, and chemical composition can be used to describe disordered regions, as has been previously observed (23)

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

(1) IDRs with high sequence conservation.

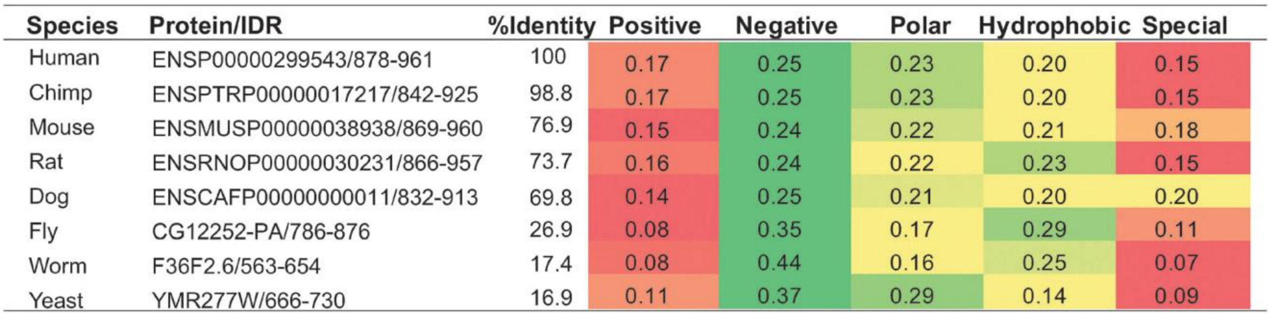
These IDRs includes those with high residue type conservation, whose function is probably resulted from folding on binding. A conserved sequence is necessary for a specific structure on folding. One of the examples of IDR within this type is the human tubulin beta-4 chain. This IDR is highly conserved among orthologous proteins. Fig S1 shows the multiple sequence alignment of such IDR. In our dataset, these IDRs gave the largest proportion compared with the other groups of IDRs with the percentage of 55%.

(2) *IDRs with low sequence conservation between orthologs but maintain their function through preservation of their chemical composition.*

These IDRs are expected to have a function as a result of their characteristic chemical composition. The weak negative correlation between the residue conservation of these IDRs and their type content difference with orthologous regions could be explained if we assume that they contain small conserved regions in the form of linear motifs.

Previously, Bellay et.al proposed a group of IDRs, namely “flexible disorder” IDRs, that functions in signal transduction (10). This type of IDR is similar with the proposed group, since it might be required to maintain the type content in the disordered region immediately surrounding the motif, in order for the motifs to be functional, in which Bellay group found a large number of small linear motifs in flexible disorder IDRs.

Small clusters of residues found in some IDRs may also play an important role in the IDRs functions. For instance, CTDP1, and IDR at the C-terminal tail of the TFIIF-associating CTD phosphatase, falls in this category. A few charged and hydrophobic residues at the very end of the C-terminal region are conserved across all species studied here. Those residues are important for binding to the TFIIF, RAP74 (Figure 7) (24). This IDR shows significant conservation among mammals, but it is poorly conserved in lower eukaryotes.



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

Figure 7. Example of an HTC IDR – CTDP1. Percent identity and type content of the C-terminal intrinsically disordered region in the TFIIFassociating 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.

Our results indicated that despite of the poor sequence conservation, according to the type content of the IDR in orthologous protein, the overall chemical composition of the IDR is preserved. Moreover, it has 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 neither conservation nor maintenance of chemical composition.*

These IDRs are abundant in Pro, Gly and ALa with interspersed charged residues. It is not clear how these IDRs maintain their function. However, based in the co-occurrence of Ala with Pro and Gly residues, Ala appears to be more closely related to Pro and Gly compared to other hydrophobic residues. Hence, Ala could be classified as a special residue in future investigations. It is also possible that these IDRs maintain their function 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 Low Type Content (LTC) IDRs.

In future studies, it will be interesting to see if chemical composition is similarly maintained in poorly conserved ordered regions in disordered as well as ordered proteins, considering the frequent residue type conservation in IDRs over residue conservation itself is similar to the behavior of ordered domains.

Classification of IDRs based on conservation has been attempted before ((10),(16),(17)), including our study. However, aside from that, we also propose a broad classification scheme based on their type content. To emphasize the importance of the IDRs' chemical composition in their function, we separated IDRs into functionally distinct groups based on their type content.

However, our classification result does not overlap with the flavors of disordered regions (16) or that based on charge and hydropathy, that have been previously proposed (17). We show that it is possible to classify IDRs based on their chemical composition into functional groups. By adding location information into the classification scheme, the result offers better functional separation of IDRs. It demonstrates the utility of chemical composition as a means of classification, despite of the simplistic classification.

In future studies, in order to provide better groups of IDRs, adding more complexity into the classification with combinations of residue types will help in better identifying the functions of IDRs. It is clear that there are more than five types of IDRs with varying combinations of residue type content. For instance, in both high residue and high type content IDRs, but to a greater extent in high type content IDRs, negative residues are frequently enriched in IDRs within the positive cluster. Introducing cluster for such IDRs might help in identifying their functions. The location preference of IDRs suggests that it is also an additional feature to classify IDRs.

However, we also have issues that need to be addressed in this study. Disordered region is inherently difficult in alignment, due to poor sequence conservation and low sequence complexity. To overcome this problem, we tried to perform multiple alignments of whole proteins (ordered and disordered regions), instead of the IDRs alone. We hypothesized that alignments of proximal ordered and conserved regions would improve those of disordered regions. We show that the scores based on alignments in this study are better than those obtained from random alignments. Indications show that these alignments are reliable, from our finding that the chemical compositions within a section of IDRs with poor sequence conservation is maintained within orthologs. There is also a possibility that this is the result of a few conserved residues within the disordered regions.

On the other hand, in the case of low type content IDRs, it is difficult to judge the accuracy of the alignments, which do not show conservation of sequence or chemical composition. Sequence-based alignment algorithms cannot accurately align these IDRs due to poor conservation. This problem naturally hinders the identification of common sequence or composition patterns within these IDRs.

To study the preservation of chemical composition further, a more accurate multiple sequence alignment method is required. Hence, to address this issue, future studies on IDRs will need to develop special alignment tools, which perhaps based on chemical composition rather than sequence.

Another concern is the choice of the appropriate cutoff to separate IDRs with high and low residue or type content conservation, due to lack of consensus on a value of sequence identity that may be considered as a cutoff. Apparently it depends in the species chosen for testing, based on comparison of our study and that of Bellay, et al., (10). This study uses average scores to separate IDRs into distinct groups, which is proven sufficient. However, more detailed parameters are needed in the future.
Another issue is that, due to the lack of experimentally determined information, the classification system presented in our study is tested on predicted IDRs. The classification system needs to be reconfirmed as the availability of experimentally

determined information increases, especially the information regarding the orthologous proteins.

The results of our study propose an explanation of the maintenance of disorderliness and function in rapidly evolving IDRs. The current function prediction methods rely heavily on sequence homology, and hence are unsuitable for proteins with large intrinsically disordered regions.

In that regard, our results also suggests a means of improving the function prediction of proteins with large IDRs and few or no known annotated sequence homologs, while suggesting several avenues for improvement.

One of the avenues is location preference. A previous study used location preference in the function prediction of proteins with long disordered regions (25). Our group has previously shown that amino acid content similarity, instead of sequence similarity, can be used to predict a function associated with an IDR. The result of our current study could improve the performance of the prediction method.

3.2. Conclusion

We ran investigation on the conservation of amino acid residues and type content in intrinsically disordered regions (IDRs) in human proteins within 7 other eukaryotes. Instead of the poor sequence conservation, we found that IDRs show type content conservation similar to that of highly conserved IDRs. This suggests the possibility that the overall nature of the IDR is more important for its function and is one of ways through which IDRs maintain their function or disorderliness despite rapid sequence evolution. As addition to it, IDRs with different levels of conservation also have differing location preferences and functional enrichments. Based on their type content, we also show that IDRs can be classified into 5 broad groups that are functionally distinct. IDRs show specific location preferences based on their conservation and their type content with positive IDRs located at the C-terminus, polar and negative IDRs in the middle and the hydrophobic and special IDRs at the N-terminus of proteins. Finally, the findings of this study demonstrate that conservation, residue type content and location can be used to distinguish between functionally distinct intrinsically disordered regions and provides a means of improve functional annotation and prediction of these regions.

4. Acknowledgement

I would like to express my sincere thanks to Professor Kenta Nakai, who kindly supervised and supported me to complete my research project. Because of him, I could have learned various bioinformatics skills and trends, which broadened my scientific vision.

Also, I would like to express my gratitude to Assisant Professor Ashwini Patil, for all of her support for this doctoral research project. Her advices and inputs really helped me in finishing this project.

Finally, I would like to express the most special thanks to my lovely wife, Prieka Khusnul Khatima, who was the strongest motivation for pursuing this research. Her devotion to me was the greatest power that accomplished this dissertation.

This dissertation was co-authored by Assistant Professor Ashwini Patil, Shunichi Wakabayashi and Professor Kenta Nakai. Dissertation is based on previously published work; Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification., Moesa HA, Wakabayashi S, Nakai K, Patil A., *Mol Biosyst.* 2012 Oct 30;8(12):3262-73

5. References

1. *Intrinsic disorder in cell-signaling and cancer-associated proteins.* **Iakoucheva, L M, et al.** 3, 2002, J Mol Biol, Vol. 323, pp. 573-84.

2. *Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks.* **Patil, Ashwini and Nakamura, Haruki.** 8, 2006, FEBS Letters, Vol. 580, pp. 2041-45.

3. *Intrinsically disordered proteins: regulation and disease.* **Babu, M M, et al.** 2011, Current Opinion in Structural Biology, Vol. 21, pp. 432-40.

4. *Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.* **Xie, H, et al.** 5, 2007, J Proteome Res., Vol. 6, pp. 1917-32.

5. *Flexible nets. The roles of intrinsic disorder in protein interaction networks.* **Dunker, A K, et al.** 2005, FEBS J, Vol. 272, pp. 5129-48.

6. *Linking folding and binding.* **Wright, P E and Dyson, H J.** 2009, Curr Opin Struct Biol, Vol. 19, pp. 31-8.

7. *Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase.* **Mittag, T, et al.** 2010, Structure, Vol. 18, pp. 494-506.

8. *Evolution and disorder.* **Brown, C J, et al.** 2011, Curr Opin Struct Biol, Vol. 21, pp. 441-6.

9. *Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation.* **Daughdrill, G W, et al.** 2007, J Mol Evol, Vol. 65, pp. 277-88.

10. *Bringing order to protein disorder through comparative genomics and genetic interactions.* **Bellay, J, et al.** 2011, Genome Biol, Vol. 12, p. R14.

11. *Functional Annotation of Intrinsically Disordered Domains by Their Amino Acid Content Using IDD Navigator.* **Patil, A, et al.** 2012, Pac Symp Biocomput, Vol. 17, pp. 164-75.

12. *Searching DNA via a "Monkey Bar" mechanism: the significance of disordered tails.* **Vuzman, D, Azia, A and Levy, Y.** 3, 2010, J Mol Biol., Vol. 396, pp. 674-84.

13. *DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail.* **Vuzman, D and Levy, Y.** 2010, Proc Natl Acad Sci U. S. A, Vol. 107, pp. 21004-9.

14. *Data growth and its impact on the SCOP database: new developments.* **Andreeva, A, et al.** 2008, Nucleic Acids Res, Vol. 36, pp. D419-25.

15. *The Pfam protein families database.* **Punta, M, et al.** 2012, Nucleic Acids Res., Vol. 40, pp. D290-301.

16. *Flavors of protein disorder.* **Vucetic, S, et al.** 2003, Proteins, Vol. 52, pp. 573-84.

17. *Subclassifying disordered proteins by the ch-cdf plot method.* **Huang, F, et al.** 2012, Pac Symp Biocomput, pp. 128-39.

18. *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* **Ward, J J, et al.** 3, 2004, J Mol Biol, Vol. 337, pp. 635-45.

19. *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* **Dosztanyi, Z, et al.** 4, 2005, J. Mol. Biol., Vol. 347, pp. 827-39.

20. *DisProt: the Database of Disordered Proteins.* **Sickmeier, M, et al.** 2007, Nucleic Acids Res, Vol. 35, pp. D786-93.

21. *Clustal W and Clustal X version 2.0.* **Larkin, M A, et al.** 2007, Bioinformatics, Vol. 23, pp. 2947-8.

22. *Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains.* **Hansen, J C, et al.** 2006, J. Biol. Chem., Vol. 281, pp. 1853-56.

23. *Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein.* **Weathers, E A, et al.** 2004, FEBS Lett., Vol. 576, pp. 348-52.

24. *NMR structure of a complex containing the TFIIF subunit RAP74 and the RNA polymerase II carboxyl-terminal domain phosphatase FCP1.* **Nguyen, B D, et al.** 10, 2003, Proc. Natl. Acad. Sci. U. S. A., Vol. 100, pp. 5688-93.

25. *Inferring function using patterns of native disorder in proteins.* **Lobley, A, et al.** 2007, PLoS Comput Biol, Vol. 3, p. e162.

26. *Modularity of intrinsic disorder in the human proteome.* **Pentony, M M and Jones, D T.** 2010, Proteins, Vol. 78, pp. 212-21.

27. *InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.* **Ostlund, G, et al.** 2010, Nucleic Acids Res, Vol. 38, pp. D196-203.

28. *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* **Huang da, W, Sherman, B T and Lempicki, R A.** 2009, Nat. Protoc., Vol. 1, pp. 44-57.

29. *Cytoscape 2.8: new features for data integration and network visualization.* **Smoot, M E, et al.** 3, 2011, Bioinformatics, Vol. 27, pp. 431-2.

30. *Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.* **Merico, D, et al.** 2011, PLoS One, Vol. 5, p. e13984.

6. Appendix

6.1. Tables

Table S1. Average values of residue, residue type and type content conservation used as cutoffs for the definition of IDR groups.

Average Value	Predicted	DisProt
Residue conservation score	6.537	7.465
Residue type conservation score	7.123	7.919
Type content conservation score*	0.105	0.059
HR Type conservation score + 2 standard deviations †	0.143	0.077

* Type content conservation score is the Euclidean distance between the type content of an IDR in a human protein with that in a corresponding aligned sequence in other orthologous proteins.

† The value used as a cutoff for type content conservation is the average value + 2 standard deviations of HR IDRs.

Table S2. Number of IDRs within each sub group as determined by the levels of residue, type and type content conservation in all predicted IDRs and those identified in DisProt.

Intrinsically disordered regions	**Predicted**	**DisProt**
Total	14612	102
High residue conservation (HR)	8075	58
Low residue conservation (LR)	6537	44
Low residue and low type conservation (LRLT)	5764	39
Low residue, low type and high type content conservation (HTC)	3042	20
Low residue, low type and low type content conservation (LTC)	2722	19

Term	Positive	Negative	Polar	Hydrophobic	Special
GOBP	RNA processing DNA packaging	Protein folding	Cell surface receptor linked signal transduction	Actin filament-based process	Ectoderm development
GOMF	RNA binding Helicase activity	Unfolded protein binding	Cytokine binding Peptide receptor activity	Lipid binding	Extracellular matrix structural constituent
GOCC	Protein-DNA complex Ribonucleoprotein complex		Intermediate filament	Ion channel complex	Collagen Proteinaceous extracellular matrix

Table S3. Representative GO terms enriched in clusters of all IDRs (p<0.01). For a complete list refer to additional files.

Table S4. GO terms enriched in HR clusters by type content ($p < 0.01$)

Term	Positive	Negative	Polar	Hydrophobic	Special
GOBP	-RNA processing -ribonucleoprotein complex biogenesis -histone acetylation -spliceosome assembly	-protein modification by small protein conjugation or removal	-morphogenesis of an epithelium -neural tube formation	-cognition -exocytosis	-ectoderm development -epidermis development
GOMF	-histone methyltransferase activity	-ubiquitin-protein ligase activity		-small GTPase regulator activity -lipid binding	-extracellular matrix structural constituent -ligand-dependent nuclear receptor transcription coactivator activity
GOCC	-histone acetyltransferase complex		-apicolateral plasma membrane -apical junction complex	-postsynaptic membrane -perinuclear region of cytoplasm -clathrin adaptor complex	-collagen -extracellular matrix

Table S5. GO terms enriched in HTC clusters by type content ($p < 0.01$)

Term	Positive	Negative	Polar	Hydrophobic	Special
GOBP	-ncRNA metabolic process -RNA processing -ribosome biogenesis -cell division -telomere organization -chromatin assembly or disassembly	-maintenance of fidelity during DNA-dependent DNA replication	-regulation of ARF protein signal transduction		-asymmetric protein localization
GOMF	-ATP-dependent helicase activity -purine NTP-dependent helicase activity -RNA binding -damaged DNA binding -cyclin-dependent protein kinase activity	-DNA secondary structure binding	-hyaluronic acid binding		
GOCC	-nucleoplasm -nuclear speck	-mismatch repair complex -endoplasmic reticulum part		-mitochondrial matrix -mitochondrial lumen	

Term	Positive	Negative	Polar	Hydrophobic	Special
GOBP	-RNA processing -ncRNA metabolic process	-microtubule-based process	-DNA repair -response to inorganic substance	-coenzyme metabolic process -monovalent inorganic cation transport -coenzyme biosynthetic process	-response to insulin stimulus -response to peptide hormone stimulus
GOMF	-metallopeptidase activity -peptidase activity, acting on L-amino acid peptides	-DNA bending activity -sulfate transmembrane transporter activity	-histone methyltransferase activity -calcium ion binding -copper ion transmembrane transporter activity		-sodium channel activity -enzyme binding -diacylglycerol binding -kinase activator activity -hydrolase activity
GOCC	-intracellular organelle lumen -organelle lumen -membrane-enclosed lumen				-Golgi cisterna membrane

Table S6. GO terms enriched in LTC clusters by type content ($p < 0.01$)

6.2. Figures

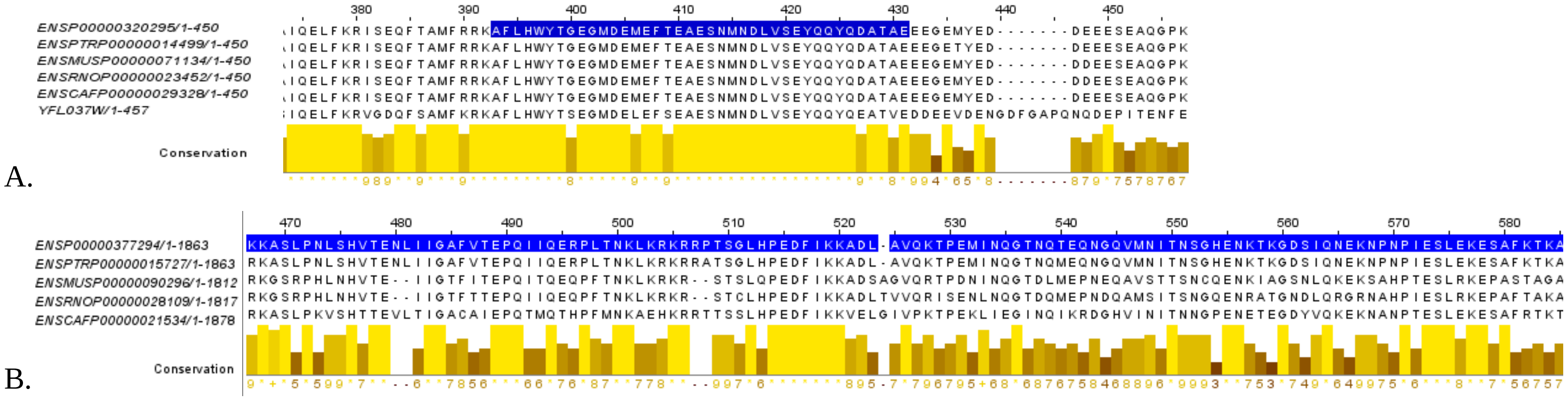


Figure S1. A) Alignment and conservation of a highly conserved (HR) IDR in the human Tubulin beta-4 chain (ENSP00000320295). This region is highly acidic and may be used o bind cations. B) Alignment of part of the Ser rich BRCT_assoc domain in the BRCA1 human protein. This IDR extends over the entire central portion of the protein and has poor conservation but high type content conservation across species.

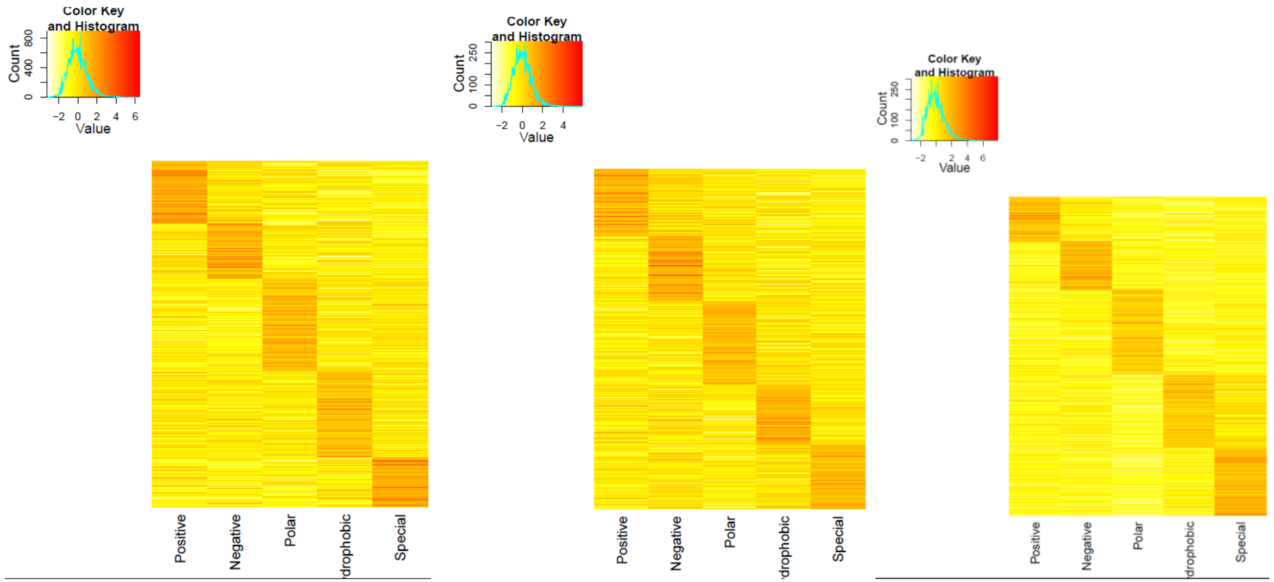
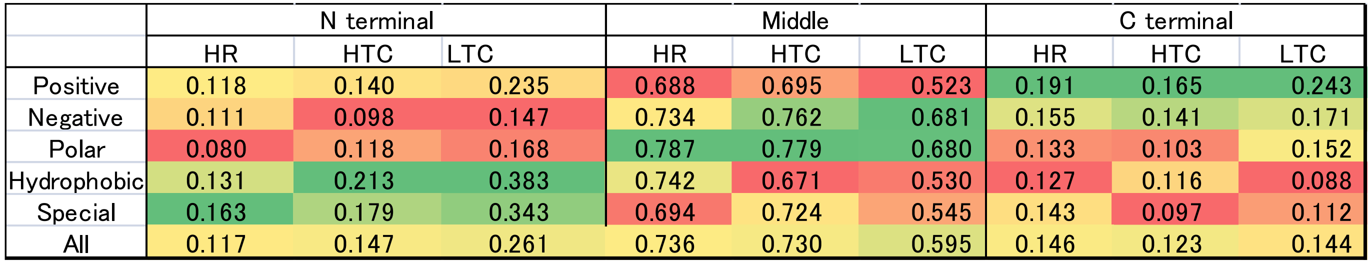


Fig S2. HR, HTC and LTC IDR clusters based on type content.



	N terminal			Middle			C terminal		
	HR	HTC	LTC	HR	HTC	LTC	HR	HTC	LTC
Positive	0.118	0.140	0.235	0.688	0.695	0.523	0.191	0.165	0.243
Negative	0.111	0.098	0.147	0.734	0.762	0.681	0.155	0.141	0.171
Polar	0.080	0.118	0.168	0.787	0.779	0.680	0.133	0.103	0.152
Hydrophobic	0.131	0.213	0.383	0.742	0.671	0.530	0.127	0.116	0.088
Special	0.163	0.179	0.343	0.694	0.724	0.545	0.143	0.097	0.112
All	0.117	0.147	0.261	0.736	0.730	0.595	0.146	0.123	0.144

Fig S3. Location preferences of the 5 type clusters in HR, HTC and LTC IDRs.

博士論文

The Conservation and Classification of Intrinsically Disordered Regions in Proteins
(タンパク質の天然変性領域の保存と分類)

ハリー・アムリ・ムサ

HARRY AMRI MOESA