Systematic analysis of short evolutionarily invariant motifs in intrinsically disordered regions

A dissertation submitted for the partial fulfillment of BS-MS dual degree in Science

by

Nitish Tayal

MS08040

Indian Institute of Science Education and Research Mohali

Knowledge City, Sector 81, SAS Nagar, Manauli, PO 140306

May 2013

Certificate of Examination

This is to certify that the dissertation titled "**Systematic analysis of short evolutionarily invariant motifs in intrinsically disordered regions"** submitted by Mr. Nitish Tayal (Reg. No. MS08040) for the partial fulfillment of BS-MS dual degree program of IISER Mohali has been examined by the thesis committee duly appointed by the institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Samrat Mukhopadhyay   Dr. Shashi Bhushan Pandit   Dr. Kuljeet Singh Sandhu

(Supervisor)

Dated:

Declaration of Authorship

The work presented in this dissertation has been carried out by me under the guidance of Dr. Kuljeet Singh Sandhu at the Indian Institute of Science, Education and Research, Mohali.

This work has not been submitted in part or in full for a degree, a diploma or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Nitish Tayal

(Candidate)

Date:

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Kuljeet Singh Sandhu

(Supervisor)

Acknowledgement

I would like to express my deepest gratitude to Dr. Kuljeet Singh Sandhu for guiding me in the final year of my thesis. He provided me with the freedom to think and encouraged me to come up with new ideas. I was especially motivated by the discussions that we had in the mess or moving around in the campus. Coming in contact with him, I could really appreciate what Science is all about. During this period, he also showed me the freedom that comes while doing computational biology. As he said, there is a lot of data out there but fewer people to analyze it. Under his guidance, I realized that if one can frame a good question, then there is nothing to stop him in Computational Biology. The discussions are like gems that I will be taking with me to start a new journey. Apart from Science, we also talked about life in general, which will surely help me in becoming a responsible citizen and then a successful scientist. Wherever I go, whatever I do, his presence in the form of his thoughts will always be with me.

I would also like to thank Dr. Jagdeep Singh for giving me the opportunity to work with *Caenorhabditis elegans* which is an elegant system to study gene expression. He encouraged me to come up with my own ideas given the constraints of the lab. He taught me to apply my thinking practically and encouraged me to change the guide when I was not finding the work interesting. I would like to thank Shruti and Avinash for teaching me to handle *Caenorhabditis elegans*. They allow me to question everything and encourage me to find the answer out. I was deeply moved by the clarity of their concepts and hope that they will surely become good researchers. It was an enjoying experience to be with them.

I would like to acknowledge the suggestions given by Dr. Shashi Bhushan Pandit at different phases of the project. It was always encouraging to have him especially because he also taught me the BIO-455 course i.e. Computational Structural Biology which helped me a lot during my project. Also, we are using his expertise in molecular dynamics to analyze the conformation of the CoRE peptides which is a future prospective. I would like to thank Miss Preeti for taking up this project. She has really devoted a lot of time in understanding and doing the molecular dynamics simulations. I would also like to thank the members of Bioinformatics Lab especially Priya and Ashutosh for their helping nature.

I would like to give my respects to the thesis committee consisting of Dr. Samrat Mukhopadhyay, Dr. Shashi Bhushan Pandit and Dr. Kuljeet Singh Sandhu (Supervisor) for taking out the time to read my thesis and giving valuable suggestions. I would like them to guide me further in my life too if time permits. I would like to thank Prof. P. Guptasharma for allowing me to shift my lab in the middle of my final year. I would also like to thank IISER Mohali for giving me this wonderful opportunity, first to study in a research oriented environment and then allowing me to do the final year research. The scientific discussions with the faculties and students over these years have opened my mind to the existence of different possibilities in the ever changing field of Science. Especially the discussions with Dr. Lolitika Mandal and Dr. Manash K. Paul have shaped my life for which I am grateful to them. I would also like to thank facilities provided by the IISER Mohali (library, sports etc.) which allow us to do healthy research.

At last, I would like to thank my friends, Karishma and Anuj for being always with me even in the tough situation when I decided to change my lab. It was due to their constant support that I did not feel lost at any time. I would like to acknowledge the respect that I got from junior batches and the love given by the seniors and PhD students. My time at IISER would be unforgettable due to the presence of many friends notably, Aastha Sindhwani, Abhilasha Joshi, Arashdeep Singh, Archana, Bhupinder, Gagan Preet Singh, Hitesh Gakhar, Kapil Dave, Manish Pareek, Mayank Chugh, Mayank Mishra, Mehreen Khaleel, Rahul Chhajwa, Reema, Ritu, Rivi Verma and Vinit Kumar. I have done projects with some of them whereas others have been involved in discussions about Science or life in general. I am taking a lot from here and hope that I will be able to do justice to the support and guidance provided by IISER Mohali.

Contents

List of Abbreviations

φ	phi angle in Ramachandran Plot
ψ	psi angle in Ramachandran Plot
ΔG	Free energy
ΔH	Change in the enthalpy
ΔS	Change in entropy
20S	20 Svedberg Unit
AA	Amino Acid
Ala	Alanine
Arg	Arginine
Asn	Asparagine
CD	Circular dichroism
CoRE	Conserved Recognition Elements
DNA	Deoxyribonucleic Acid
DSSP	Define Secondary Structure of Proteins
FDR	False Detection Rate
Gly	Glycine
H	Shannon entropy
<H>	Net hydrophobicity
IDP	Intrinsically disordered Proteins
Ile	Isoleucine
IUPred	Prediction of Intrinsically Unstructured Proteins
Leu	Leucine
Lys	Lysine
Met	Methionine
MoREs	Molecular recognition Elements

NMR	Nuclear Magnetic Resonance
P_{ij}	Amino acid propensity of i amino acid in j state
PDB	Protein Data Bank
Phe	Phenylalanine
Pro	Proline
<R>	Net charge
RCSB	Research Collaboratory for Structural Bioinformatics
R_H	Relative Shannon entropy
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
SAXS	Small Angle X-ray Scattering
Ser	Serine
SLiMs	Short Linear Motifs
sm-FRET	Single Molecule Förster Resonance Energy Transfer
T	Temperature
Thr	Threonine
Tyr	Tyrosine
UV	Ultra Violet
Val	Valine
VMD	Visual Molecular Dynamics

List of Figures

Abstract

Intrinsically disordered regions in proteins are known to mediate macromolecular interactions. Despite recent advances, it is yet not very clear how these regions recognize the specific targets without any defined three dimensional conformation. Here, we report a repertoire of evolutionary conserved *de novo* peptides in long disordered regions, which have strong tendency to retain certain preferred conformations. The peptides exhibit distinct amino acid propensities, particularly the enrichment of Gly, Ala, Arg, Lys, when compared to other known short linear motifs. The peptides show significantly conserved Ramachandran conformations, secondary structures and the three dimensional folds when compared to neighboring regions. Significant enrichment of DNA-binding and extracellular matrix binding functions of proteins with these peptides further explains the abundance of positively charged amino acids like Arginine and Lysine. DNA-binding function of motifs was further confirmed through prediction of DNA binding residues. Importantly, the non-synonymous single nucleotide mutations in the peptides are predicted to be highly intolerant for the protein function when compared to neighbouring regions, hinting at their indispensable function in proteins. Preferred left handed bridge-conformation of enriched Glycines in CoREs suggest that Gly-to-nonGly mutations within CoRE can alter the backbone conformation and consequently the function, a hypothesis that we reconciled using mutation data. Overall, our observations uncover an evolutionary strategy wherein certain set of peptides, which have strong tendency to retain their conformations are recognized and utilized in disordered regions for molecular recognition. We therefore, propose that these peptides could serve as anchors for initial recognition, followed by other binding and folding events, during macromolecular interactions of disordered proteins. Structured nature of CoREs suggests possibilities to inhibit the molecular interactions using small molecules mimicking CoRE conformations.

Keywords Intrinsic disorder, molecular interaction, DNA-binding, conformation, Ramachandran plot.

Chapter 1. Introduction

Protein Structure Function paradigm

The paradigm claimed that a particular sequence of amino acids in the protein gives rise to a particular 3D structure which is responsible for its particular function. [1] This paradigm was originally inspired by Emil Fisher's work on glycolytic enzymes who coined the lock and key model for enzyme activity. [2] Consequently, it was shown that proteins can lose and gain back its native activity in different environment. [3] In the presence of denaturant like urea or Triton X, the protein will lose its native structure along with its activity. [3] On the removal of the denaturant the structure as well as the function is revived without the help of any other protein. [3] This means that information about the structure is present within the sequence of the protein. Also, the loss of structure results in loss of function making it necessary to have an ordered 3D-structure. [2] However, with the realization of presence of flexible regions in proteins and the existence of intrinsically disorder proteins, the paradigm started fading its truth. The initial realization came when the missing residues in X-ray structures were analyzed by Uversky et al and through other biophysical techniques. [1] He proposed that missing residues are not due to systematic errors but is the intrinsic flexibility provided by the disorder regions. As these regions are flexible, the different ensemble would have different orientations and hence the resolution of the structure is poor. Meanwhile, we also came across intrinsically disordered proteins which do not have any ordered native structure in vitro. Keeping all these observation in mind, the protein trinity paradigm was established which states that protein can exist either in ordered state, the molten globule state or the random coil state and any of these can be the native form of protein capable of function. [2] The ordered state has a fixed tertiary structure whereas the molten globule states have some native like secondary structure but dynamic tertiary structure. [4] The random coil on the other hand has no secondary structure and is highly flexible. [4] With this background, the protein structure function space is divided into two sub spaces. [5] First one contains globular proteins which have binding pockets, active sites and interaction surfaces. [5] The other one consists of sorting signals, post translational modification sites and macromolecular interaction sites. [5] This other group is made up of disorder regions or intrinsic disordered proteins.

Intrinsic structural disorder

Before going further we must define disorder regions in proteins. As stated earlier, the missing residues in X-ray crystallography which are important for function contributed to disorder regions due to flexibility. Apart from this, loops or coils as defined by DSSP are the ones containing disorder. [5] However, every loop is not necessarily disorder. [5] The definition of disorder contains that the loop must be highly mobile. [5] As propensity scales can be derived from B-factors, one can use them to define disorder proteins by X-ray structures. [5] Also, we should be careful in differentiating flexibility and disorder. [6] Disorder means a lack of constraints on many degrees of freedom of the polypeptide chain and no permanent structure. [6] On the other hand, flexibility implies concerted changes that affect a few degrees of freedom modifying the overall structure without destroying it. [6] So, apart from flexibility, the absence of secondary structure defines disorder. [6] The folding of the protein is entropy driven process especially in aqueous solution. [7] This entropy is largely increased by the hydrophobic collapse in case of the proteins having ordered structure. [7] But for this to happen, the protein must contain minimum fraction of hydrophobic residues. [7] If the hydrophobic residues are smaller than this fraction, then blobs of hydrophobic residue surrounded by hydrophilic residues are formed but the overall structure contains flexible random coils. [7] These flexible domains in the proteins are composed of charged residues such that the value of (2.785 <H> - <R> - 1.151) is negative. Here <H> is the net hydrophobicity whereas <R> is the net charge of the residues. [8] This comes from the Uversky plot where the higher net charge corresponds to the disorder region whereas higher hydrophobicity corresponds to ordered structure. [8]

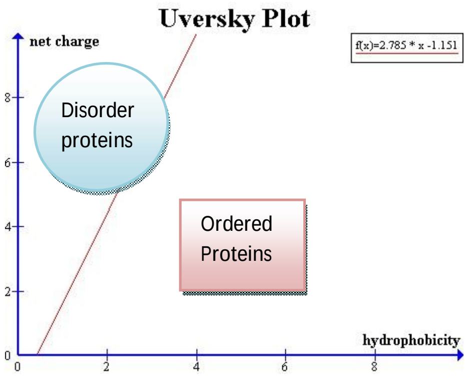

Fig. 1 Uversky plot of net charge against net hydrophobicity. Disorder proteins contain higher net charge whereas ordered proteins are enriched in hydrophobic residues.

The existence of net charge ensures that electrostatic repulsions forbid that part to fold owing to large positive enthalpy change for the residues to come together and the low hydrophobicity also ensure that change in entropy is very less. Therefore, the free energy of folding (ΔG),

$$\Delta G = \Delta H - T\Delta S$$

at normal temperature would be positive and hence, the folding would not happen. This can happen at the global level where the whole protein contains very less hydrophobic residues (as in intrinsically disordered proteins) or locally where some part of the protein is disordered due to presence of high net charge. [8] Due to this reason, the disorder can be predicted from the sequence of the proteins. This makes the proteome wide prediction of disorder easier. Also, the Uversky plot can be considered as the definition to disorder in case, there is any confusion. However, if the disordered region of protein comes in the contact with other macromolecules which nullify the effect of charge then the region can fold. [6] This is known as coupled binding and folding. [9] The functional sites present in these disorder regions came to be known as linear motifs. This allows the proteins to interact with multiple partners. However, the presence of charged residues in disorder regions allows non specific interactions which results in promiscuous interactions. [10]

Toxicity and in vivo existence

There is still debate on the in vivo existence of disorder in intrinsically disorder proteins. The disorder within the protein may not interact with other proteins but these intrinsically disordered proteins have huge entropy and can form even non-specific electrostatic interactions. These non-specific interactions would lead to the toxicity inside the cells due to aggregation or interfering with other protein's functions. Due to this, some studies reveal that these proteins would always be present in bound state. [6] On the contrary, there are proteins whose function is absolutely dependent on their disorder regions both in vitro and in vivo. For eg: in cell NMR studies revealed that microtubule binding region of tau protein is ordered whereas its long projection domain remains largely disordered. [11] Also, IDPs are sensitive to ubiquitin independent degradation by 20S proteasome both in vitro and in vivo. This pathway requires the protein to be present in disordered state. [11] Also, the expression level of these disordered proteins is less preventing the toxicity. [6] This implies that disorder exists even in vivo. [11] Hence, the debate is still open and demands further observations and analysis.

Function of disorder

The disorder region can function as the flexible region which provides the protein, the ability to cover more volume. This is important if the interacting partner is macromolecule whose concentration is low especially due to low mobility of macromolecules owing to their size and crowding inside cell. [12] Due to this, enzymes which function on small molecules usually do not contain disorder regions whereas the proteins involved in macromolecular processes (like transcription factors) contain disorder regions. This ability to cover more volume due to the presence of flexible arm is known as fly casting mechanism. [12] However, recent studies indicate that fly casting mechanism do not increase the capture efficiency but increases the binding efficiency because it requires fewer encounter times. [9] This happens because ordered structures have higher capture efficiency but do not evolve to bound state that often. [9] On the other hand the encounter times required for disorder proteins undergoing coupled binding and folding is very less giving them kinetic advantage. [9] However, this study needs to be further assesed and tested on different disorder interactions. [9] Apart from covering more volume, these regions are also implicated in multiple molecular recognition pathways where they undergo coupled binding and folding. The existence of partner induces the protein to fold in such a manner that it exposes the residues suitable for binding. [6] Study of coupled folding and binding leads to the concept of molecular recognition elements (MoREs) of 10-70 residues in length. [13, 14] Also, small functional units in linear motifs consisting of 3-10 residues called SLiMs came into light. The existence of these elements allow adaptability in binding, high functional diversity, transient and specific binding and frequent regulation by post translational modifications. [13] In general, the disorder present in terminals are implicated in molecular recognition particularly, DNA binding, whereas the middle disorder region is implicated to impart flexibility and individuality to domains so that they can have different functions without interference from other domains. These regions are also targetted for post translational modifications and are susceptible to protein degradation. The disorder is also implicated in various neurodegenerative diseases like Alzheimer’s or Parkinson’s disease where intrinsically disordered proteins in extracellular matrix aggregates causing the synaptic communication to shut down.

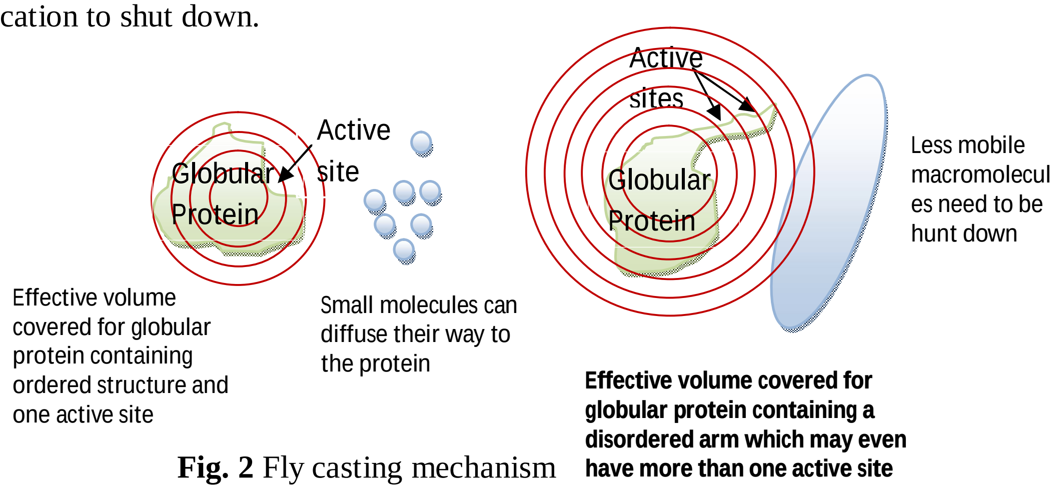
Globular Protein
Active site
Effective volume covered for globular protein containing ordered structure and one active site
Small molecules can diffuse their way to the protein
Active sites
Globular Protein
Less mobile macromolecul es need to be hunt down
Effective volume covered for globular protein containing a disordered arm which may even have more than one active site

Fig. 2 Fly casting mechanism

Network hubs

Inside the cell, the proteins interact with each other in order to carry out biological processes. The networks of these interactions are largely scale free. This happens due to the presence of two major types of hubs. The globular proteins which are essential for cell like ribosome proteins, forms one kind of hubs called party hub because they were conserved in evolution whereas the proteins interacting with them evolved as evident from their ancestry. [15] Hence, a number of proteins have stable interactions with them. On the other hand lays disorder proteins which allow transient interactions with different proteins. Therefore, while making protein-protein interaction networks, they form the hubs to which many proteins are connected. The specificity comes from the hydrogen bonding and hydrophobic effect due to the presence of a few specific residues in the disordered arm. The arm itself can move around covering a huge volume as explained earlier. Due to the presence of transient interactions, these hubs are called date hubs. [15] This ability of disordered proteins makes them well suited to act as signaling molecules or transcription factors. Both of them have different partners at different times inside the cell and need to switch depending upon the conditions. The localization and low rate of synthesis of these proteins may help them to avoid non-specific interactions. [6]

Experimental methods

The most extensively used method to study protein disorder is multidimensional NMR complimented by small angle X-ray scattering combined with molecular dynamics simulations. [16] NMR parameters like chemical shift, N^{15}-H^{1} heteronuclear Overhauser effect, relaxation and residual dipolar coupling values are sensitive to local structures. Also, long range structural constraints can be determined through paramagnetic relaxation enhancement NMR measurements and SAXS experiments. [16] These conformers were then molecularly simulated to provide first quantitative structural description of function of IDPs. [16] Also, CD analysis showed the disorder to order transition due to characteristic signals from different secondary structures. Moreover Far UV CD spectra can distinguish molten globule and globular proteins from random coil. [2] Apart from these, single molecule studies are also possible consisting of single molecule fluorescence resonance energy transfer (sm-FRET) measurements. [16] Atomic Force microscopy allows the visualization of conformational changes or the study of energetic and dynamics of structural ensemble of IDPs. [16]

Evolution of disorder

The disorder regions are present extensively in all eukaryotes. The possible mechanism for their evolution can be repeat expansion where a part of genome is copied multiple times. Another method could be horizontal transfer through viruses as they are enriched in disorder regions like eukaryotes. The disorder regions do not have secondary structure so the random insertions, substitutions and deletions of few residues would not affect the function very often. [17] However, if the mutation is in the functional region which is important for flexibility or the molecular recognition, then these mutations may be deleterious. The mutations in disorder regions leading to deleterious effect are therefore, scale free distributed. In contrast, the mutations in globular proteins (highly ordered) have less freedom because they are also involved in making the specific structure required for the function. This also implies that the conserved sequences present in the disorder regions are essential for the function of the region. This along with an observation of AT-hook leads to the hypothesis on which this final year thesis is based.

Hypothesis

The conserved sequences present in the disorder regions of the proteins have functional roles and hence, have a specific local structure enabling it to execute that function. The conservation of structure is just in these conserved sequences and not in the neighboring residues making them absolutely necessary for function.

Motivation

The AT-hook is a conserved motif present in DNA binding proteins (like transcription factors). The essential motif is RGRP (Arg-Gly-Arg-Pro) where each is the single letter code for amino acids. It was observed that this motif is present in two different conformations. The bound conformation has larger end to end distance whereas the unbound form is relatively compact. Also, the superposition of these residues from the PDB shows that the structure is invariable in each state irrespective of the protein they are present in.

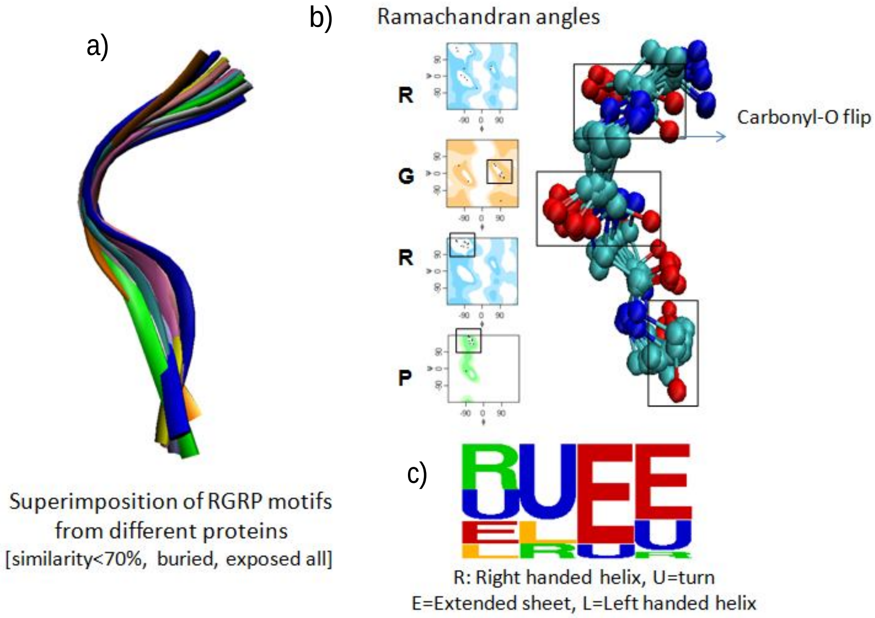
a)
b)
Ramachandran angles
R
G
R
P
-90 0 90
Carbonyl-O flip
c)
Superimposition of RGRP motifs
from different proteins
[similarity<70%, buried, exposed all]
R: Right handed helix, U=turn
E=Extended sheet, L=Left handed helix

Fig. 3 Conformation analysis of AT-hook (unpublished data). (a) Superposition of AT-hook from proteins with <70% similarity. (b) Ramachandran Plot for the 4 residues that are conserved across different proteins. (c) Conformational strings for the proportion of residue present in specific region in Ramachandran Plot (E: extended beta sheet, R: right handed alpha helix, U: turn)

This motivated us to look for the existence of other such motifs which are evolutionary conserved. The motifs like MoREs and SLiMs were already present but the way they are projected as linear motifs did not approve us. We thought that instead of being linear, these motifs should make some anchor like structures which are able to fish out macromolecules. As we had already observed such type of structure in AT-hook analysis, therefore, such existence was plausible. Moreover, the SLiMs and MoREs present in literature are more general in nature. We intend to look for specific sequences without allowing any mutation in those sequences. As the sequences are conserved in different species, they are thought to do similar function in them. Therefore, structure should remain same. The applicative motivation was their use in synthetic polypeptide formation where specific fold is required and targeting them for curing diseases. With this motivation in mind, we started the project.

Chapter 2. Materials and Methods

Perl was used as the programming language for text based search. R and Microsoft Excel were used for the statistical analysis. The programs written for different analysis are present in the supplementary section.

Sequence analysis

Uniprot database (uniprot-sprot FASTA file downloaded in Oct, 2012) was used as the starting point where all the manually curated, non-redundant protein sequences are maintained. [18]

Removal of redundancy

The sequences downloaded from Uniprot database were clustered based on homology greater than 70% and largest sequence among them was taken as the representative sequence to remove the identical sequences from whole database. CD-HIT was downloaded as C-program application from net and used for this purpose.

Disorder Prediction

IUPRED was used to predict disorder regions present in the proteins by estimating their pair wise inter-residue interaction energy based on the presumption that IUP do not fold due to their inability to form sufficient stabilizing interactions. [19] Apart from it, it can also predict short and long disorder regions present between the ordered structures in protein. It uses a 20X20 energy predictor matrix parameterized by statistical method to approach pairwise energies of globular proteins of known structures. [19] As no training on disordered proteins is involved, this distinction underlines that the lack of a well-defined three-dimensional structure is an intrinsic property of certain evolved proteins. [19] This approach was turned into a position specific method to predict protein disorder by considering only the local sequential environment of residues within 2–100 residues in either direction. [19] The score is then smoothed over a window-size of 21. [19] The C-program for IUPRED was downloaded from the web. It takes a single fasta sequence and converts the disorder region in capitals and order in small letters. It was called by a Perl program (present in supplementary data) to do the prediction for all the sequences making a temporary file containing single fasta sequence. To validate this assignment, an unsupervised method was also used called Fold Index. It uses Uversky definition for distinction between disorder and order regions. [20] It is based solely on the average hydrophobicity of the amino acids and their net charge. [20] The mean net charge, <R> is the absolute value of the difference between the numbers of positively and negatively charged residues at pH 7.0, divided by the total residue number, and the mean hydrophobicity, <H> is the sum of all residue hydrophobicity, divided by the total number of residues, using the Kyte/Doolittle scale rescaled to 0 to 1. [20] An index (I) was defined as

$I = 2.785 \langle H \rangle - \langle R \rangle - 1.151$ [20]

The positive value for this index means ordered structure whereas a negative value means disorder. [20] The perl program for it (given in supplementary data) uses a web interface and calls the output for each sequence from the web. The predicted disorder is capitalized and checked for consistency.

False Detection Rate (FDR)

Peptides of size ranging from 3 residues to 20 residues were created from the long disorder regions present in the proteins using a sliding window method. To test the statistical significance of occurrence of these residues, FDR analysis was done. It was found out that peptides of size 6 or more residues occur with considerably low false detection rate (FDR). So, a size cutoff of 6 residues was taken for further analysis.

Complexity analysis (Shannon entropy)

Shannon entropy is the measure of uncertainty in a random variable. The low complexity peptides occur in different species more frequently but contain less information. Higher the complexity of the sequence, higher will be the significance if it is conserved in different species. The complexity of the peptides were calculated by the Shannon entropy given by

$$H = -\sum_{i=1}^{k}\left(\frac{n_i}{l}\right)\log_2\left(\frac{n_i}{l}\right)$$

Where, n_i is the count of amino-acid *i. k* is the total number of amino-acids possible (i.e. 20) and *l* is the length of the peptide analyzed.

The peptides having Shannon entropy of greater than 2.2 and species conservation of atleast 6 different species were selected for further analysis. Conjugated peptides were concatenated if they are also conserved in more than 6 species giving longer peptides. The final set of these peptides were called CoRE (Conserved recognition elements). Further analysis was done on these peptides.

Amino acid propensity

The occurrence of different amino acids was calculated for the CoRE as well as the disorder and order regions. This was normalized to the occurrence in the uniprot database giving the propensity.

$$P_{ij} = \frac{n_{ij} \big/ \sum n_{ij}}{N_i \big/ \sum N_i} \quad [21]$$

Where, n_{ij} is the count of amino acid in state *j* (ordered, disordered, SLiM or CoRE) and N_i is the count of amino acid *i* in the UniProt database. Propensities were than normalized between 0 to 1 using following equation: $P_{ij}^{'} = \frac{P_{ij} - \min(P_{ij})}{\max(P_{ij}) - \min(P_{ij})}$

AA propensities in CoREs were subtracted from that of disordered regions and sorted based on their significance of over-representation in figure 1C.

Structural Analysis

PDB seqres

PDB seqres records were downloaded on 3rd Dec. 2012. The CoRE peptides obtained from the sequence analysis were searched into this seqres record for getting the pdb ids. The pdb ids containing the CoRE peptides were downloaded from RCSB after applying a cutoff of 70% similarity and resolution of < 3Å. This ensures that the overall sequence of the proteins is atleast 30% different and there is no redundancy in the sequence. Also, the resolution limits excludes the low quality pdb ids. Therefore, whatever conformation comes out is not the result of the whole protein being the same or errors in crystallization.

Secondary structure assignment (DSSP)

Define Secondary Structure of Proteins (DSSP) provide the secondary structure information from the pdb records given the high resolution while obtaining the structure information. The pdb records of the ids were downloaded from RCSB PDB. Then DSSP was run to assign secondary structure information from the coordinates given in pdb. The surface buried string was provided by assigning surface (S) if the exposed surface area is greater than 25% of the area exposed in the tripeptide Gly-X-Gly.

Structure Conservation

The coordinates of the CoRE peptides were compared against the pre and post peptide region by calculating RMSD of the superimposed structures and calculating end to end distances. The structures were superimposed using VMD and RMSD was calculated from them.

$\mathrm{RMSD} = \sqrt{((\Sigma\delta^2)/N)}$ [22]

Where δ is the distance between the respective residues in the aligned structure of peptides and N is the total number of residues present. The summation is done over all the residues present. The smaller value of RMSD (towards 0) indicates the conservation of structure.

Apart from RMSD, end to end distances provide the information about the space available inside the CoRE and their conservation further indicates the conservation of structure as well as specificity for some macromolecule whose size is comparable.

End-distance = $\sqrt{}\ ((x_1\text{-}x_2)^2 + (y_1\text{-}y_2)^2 + (z_1\text{-}z_2)^2)$

Where, (x_1, y_1, z_1) are the coordinates of N-terminal peptide whereas the (x_2, y_2, z_2) are the coordinates of C-terminal peptides.

Conformational strings

Conformation of each amino acid was designated as right-handed helical (R), left handed helical (L) and extended beta sheet (E) conformation as per the standard phi-psi ranges given in the figure S2A. Conformation outside the phi-psi ranges of R, L and E was designated as unassigned (U). Secondary structure letter H(alpha-helix), B(isolated beta bridge), E(extended strand), G(3/10 helix), I(pi helix), T(H-bond turn), S(bend) for residues were obtained from Definition of Secondary Structure of Proteins (DSSP) program [23]. Surface accessibility of each residue was designated at Surface (S) if the accessible surface area of an amino acid is at-least 25% of total surface area of that amino acid, else it was assigned Buried (B) notation. Each peptide was, then, assigned conformational string of R, E, L and U letters, a secondary structure string of H, B, E, G, I, T and S letters, and surface accessibility string of S and B letters. Relative Shannon’s entropy for conservation of multiple conformational strings, secondary structure strings and surface accessibility strings was calculated using following equation

$$R_H = \log_2(k) - \sum_{i=1}^{k} \left(\frac{n_i}{N}\right) \log_2\left(\frac{n_i}{N}\right)$$

Where, n_i/N is the fraction of positions conserved in the multiple sequence alignment and k is the total number of possible amino-acids. The cosine of the angle between C_i=O and C_{i-}

$_1$=O bonds was calculated using DSSP program [23]. Control analyses were performed by scrambling CoRE sequences and searching for its multiple occurrences in PDB.

Bridge regions in Ramachandran plot

The Glycine showed higher propensity in the CoRE peptides so, the Ramachandran plot of the Glycine was made after calculating phi and psi angles from the DSSP files generated from coordinates given in pdb files. This study was also important because it is reported that glycine act as surrogate D-amino acid as they act as structural probes and provide conformational stability. [24] The b_r region (towards right handed α-helix) is defined as -120 <= phi <= -60 and -30 <= psi <= 30 and b_l region (towards left handed α-helix) was defined as 60 <= phi <= 120 and -30 <= psi <= 30. These ranges were taken from literature [25] and shown in the figure S4A.

Functional analysis

Gene ontology

The Uniprot ids containing the CoRE region and present in *Homo sapiens* were submitted to babelomics 4.3 to check for gene ontology against disorder and order ids present in same species. Specifically, the molecular function, cellular component, biological process and activity were checked for these sequences.

DNA binding ability (BindN)

As the DNA binding comes at the top with significant adjusted p-value, we checked for the potential of the CoRE to bind with DNA using BindN data for Homo sapiens. BindN is the DNA binding data provided by ensembl which predicts the binding potential of each residue while binding to the DNA. The human BindN data was downloaded and the CoRE peptides were mapped to this. No binding potential (indicated by - sign) was assigned a score of 0 whereas binding (indicated by + sign) was assigned a score of 1. Then position wise average was taken for the CoRE peptides and 10 residues pre and post peptide each.

Mutation analysis

To find out the importance of these residues, mutation analysis was done. The CoRE peptides were mapped to the disorder regions in the Ensembl Human database after doing the same sequence analysis on the Human proteins provided by Ensembl. These Ensembl ids were fished out from the Sift and Polyphen database and the scores for pre and post peptides along with the CoRE peptide were averaged position wise from each database separately. The score below 0.05 for any amino acid change means that the position is intolerant to this mutation whereas a higher score towards 1 means that it is tolerant to the mutation.

Chapter 3. Results

Evolutionarily conserved elements in long disordered regions

A total of 27781 long disordered regions (LDRs) mapping to 18993 proteins in non-redundant (<70% sequence identity) UniProt (Magrane and Consortium 2011) sequences were identified using the IUPred [26], a large fraction (73%) of which was also confirmed by FoldIndex algorithm [20] (figure S1A-B, Materials and Methods). LDRs were then subjected to a brute force search to calculate the occurrence of identical peptides of different lengths across different species represented in UniProt [18] database. As shown in the fig. 4A, the occurrence of disordered peptides of length ≥6 amino acids is rarer among species, having median value of 1, in the database while the peptides shorter than 6 amino acids were relatively common among species. To extract the peptides that were significantly enriched across multiple species, we calculated false detection rate (FDR) for the occurrence of hexamer peptides in different species and selected the cut-off of 6 species corresponding to FDR of ≤0.01 (Materials and methods). Furthermore, the disordered regions are enriched in low complexity sequences, particularly the tandem repeats of hydrophilic and charged amino-acids [27-29], which have high probability to occur in more than one species by chance. To remove these sequences, we calculated the sequence complexity using Shannon's entropy approach (Materials & Methods). As apparent from the fig. 4B, the distribution plot of Shannon's entropy shows distinct peaks corresponding to low and high complexity peptides. We, therefore, took the peptides of high sequence complexity for further analysis. This led to a final dataset of 879 peptides of length ≥6 residues. 66.2% of CoREs mapped to vertebrates, of which 57.7% were shared by invertebrates, unraveling their wide-spread conservation, in contrast to earlier reports wherein short linear motifs were claimed to be conserved primarily within vertebrates (figure S1D). Interestingly a large fraction (88%) of total CoREs found in viruses were also shared by metazoans, suggesting that viruses might acquire CoREs from host through horizontal gene transfer, possibly reflecting a strategy to hijack the host cell regulation (figure S1D).

Interestingly, these peptides, though of similar length as Short Linear Motifs (SLiMs), did not overlap with SLiMs present in UniProt. There were only 50 (5.7%) peptides out of 879, which overlapped with SLiMs. Moreover, the peptides were significantly different from SLiMs in amino-acid composition. While SLiMs show significant enrichment of Asn, Met, Ile, Val, Leu, Phe, Tyr, Pro amino acids over disordered regions, our peptides were enriched in Gly, Ala, Arg, Lys, Leu and Pro amino acids (fig. 4C). We therefore distinguished our peptides from SLiMs using a distinct term: Conserved Recognition Elements (CoREs). The over-representation of ancient amino-acids like Gly and Ala further reinforces the evolutionary conserved nature of COREs. Strong under-representation of hydrophilic amino acids like Ser, Thr and positive enrichment of a few hydrophobic amino acids like Ala and Leu might ascribe the tendency to be ordered when compared to neighboring regions, a property that we explored in detail in the further analysis.

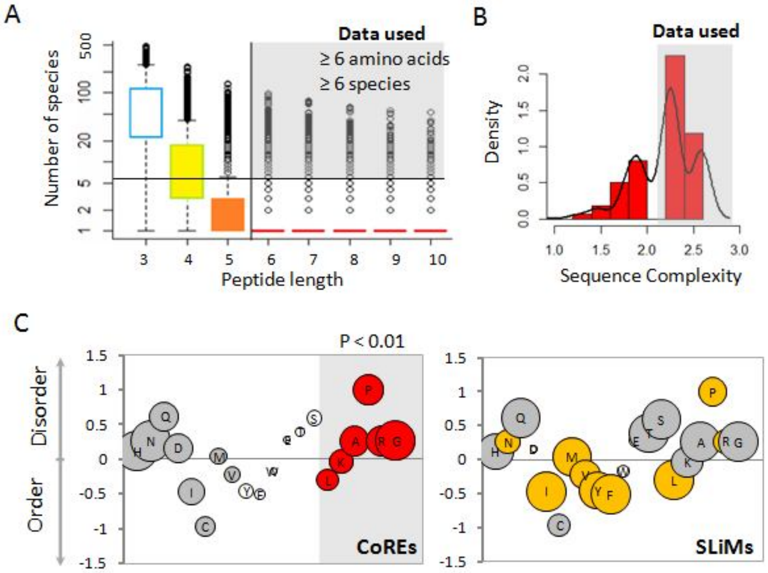

Fig. 4 Identification of Conserved Recognition Elements (CoREs). (A) Boxplot for the distribution of number of species for the occurrence of peptides of different lengths. Peptides of length 6 occurring at-least in 6 different species (FDR<0.01, highlighted in gray), were selected for analysis. (B) Complexity analysis of peptide sequences. Peptides with low sequence complexity (Shannon's entropy ≤ 2) were removed from the analysis. (C) left panel: Bubble plot representing amino acid propensities of CoRE motifs. Bubbles are scaled as per 1-log2(p) values (Fisher's exact test) for the difference in the amino acid propensities in CoREs and LDRs in the background of disorder (positive y-axis) and order (negative y-axis) promoting amino acid propensities. Bubbles are ordered as per their significance of enrichment. Gray color of the bubbles signifies under-representation, white is no enrichment and red represents over-representation. Right panel: Same analysis as in left panel, but for the known SLiM motifs. Gray color represents underrepresentation, white as no enrichment and Orange as over-representation of amino acids. The order of bubbles is the same in left panel.

CoREs exhibit conserved three dimensional conformations

Evolutionary invariance might suggest preferred three dimensional orientations of CoREs. To test this, we performed following three analyses:

1. CoREs were searched in Protein Data Bank in order to fetch their naturally occurring conformations. We obtained multiple occurrences of 51 peptides from non-redundant (<70% sequence identity) crystal structures of high resolution (<3 Å). The dihedral angles *φ and ψ* were calculated for each amino acid in the CoRE and upto 10 amino acids on either side of the CORE. The amino acids were categorized in right handed helical (R), extended beta sheet (E) and left handed helical (L) conformations based on their respective *φ-ψ* values (figure S2A). Conformational conservation across multiple occurrences of a CoRE was assessed using Shannon's Entropy. The analysis suggested significant preference of CoRE motifs to retain particular conformation, when compared to neighboring regions (fig. 5A-i, 6B). Similarly, the *secondary structure of COREs also* exhibited significant conservation as compared to neighboring regions (fig. 5A-ii). Interestingly, however, the CoREs were variable in their surface accessibility (figure S2B) suggesting that the conformations of CoREs are conserved despite their differential structural environments. Furthermore, the orientation of carbonyl oxygen with respect to its previous carbonyl oxygen was also significantly conserved in the CoREs. This was assessed by calculating the cosine of the angle of between C_i=O and C_{i-1}=O bonds (fig. 5A-iii).
2. To analyze the overall 3D orientation of CoREs, we measured the end-to-end euclidian distances of CoREs and their neighboring peptides, of same length as CoRE, across different conformations. The end-to-end distances were significantly smaller for the CoRE regions, when compared to neighboring regions, suggesting folded nature of CoREs (fig. 5A-iv). Moreover, the standard deviations of end-to-end distances of COREs were significantly lower than that of neighboring regions (fig. 5A-v). Similarly, multiple structural alignments of the CoREs suggested significant retention of 3D fold when compared to neighboring regions (median RMSD of 0.37 vs. 2.16 Å respectively, fig. 5A-vi). As a control, we performed the same analyses on scrambled CoREs (Material & Methods). The scrambled peptides did not exhibit significantly conserved 3-dimensional conformations (fig. 5B, figure S2C-E).

3. Since the above analysis was performed regardless of whether the CORE occurred in the disordered region, in principle, it remained unaddressed whether the COREs exhibit any tendency to mould into preferred conformation in the disordered state. To address this, we searched our peptides in the NMR structures deposited in PDB and extracted the ones, which occurred in the disordered regions (≥20 aa). In adherence to our earlier observations, the analyses of end-to-end distances and the Root Mean Squared Deviation (RMSD) of structural alignments of multiple NMR models of the same protein revealed a greater conservation of 3-dimensional fold of CORE motifs (median RMSD of 0.34 vs. 0.48 Å respectively, fig. 5C-E). This is also noteworthy that 78% (11/14) these examples did not involve any direct interaction with other macromolecule and hence represent native unbound state of CoREs. Limited number of NMR examples (7 peptides across 14 PDBs) does not allow for statistical power in the analysis, which might explain insignificant p-values for certain comparisons.

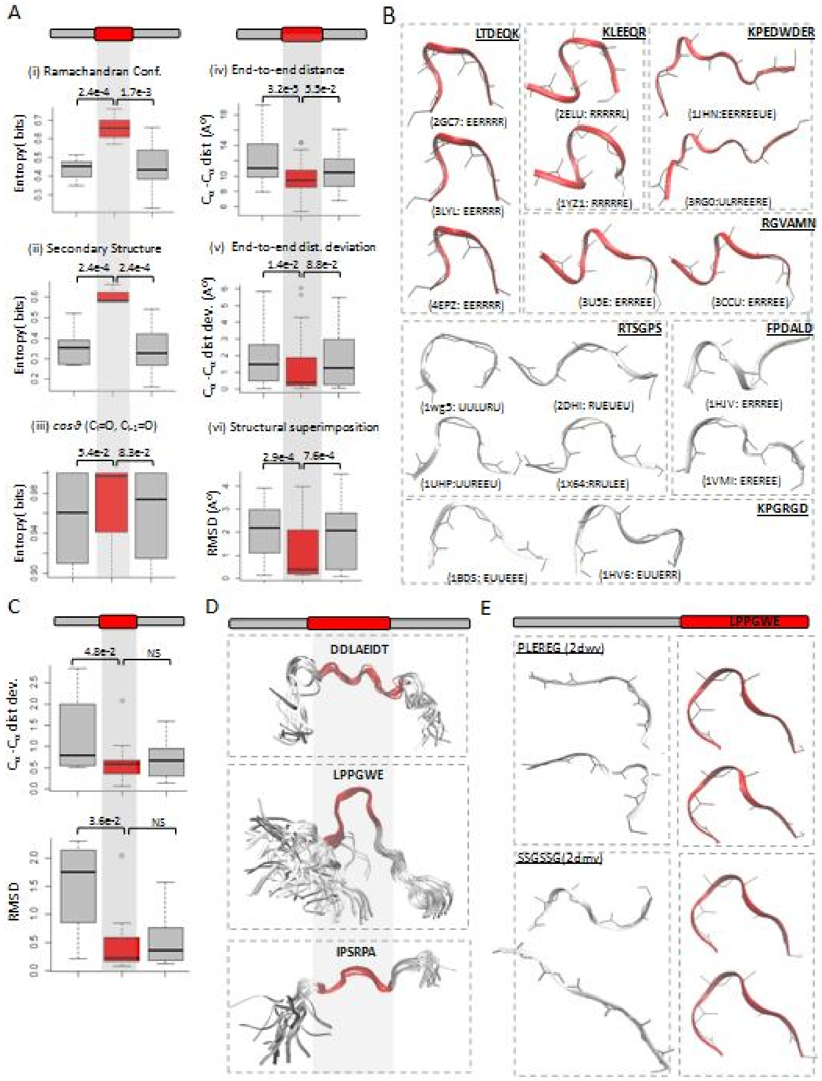

Fig. 5 Conformational analysis of CoREs. (A) From top-left to bottom-right corner: Conservation analysis of Ramachandran conformational strings, secondary structure strings, carbonyl oxygen orientation, end to end distances, deviation in end-to end distances and root mean squared deviation of superimpositions of CoRE structures found in Protein Data Bank. Shannon's entropy was calculated for I, ii & iii as elaborated in Materials and Methods. CoREs (red) were compared to their immediate upstream and downstream neighboring peptides for each analysis. P-values were calculated using Mann-Whiteny *U* test. (B) Examples of CoRE (Red) and the control (gray, scrambled CoREs) peptides. Below each peptide, respective PDB ID and conformational string is mentioned. (C) End-to-end distance deviation (top panel) and structural superimposition (bottom panel) of multiple NMR models of CoREs and their neighboring regions. Mann-Whitney U test was use to calculate p-values.

Contd. Legend (B) Examples of multiple structural alignments of NMR models of CoREs (red) and their neighboring regions (gray). (C) An example showing folds of LPPGWE and its preceding peptide peptides across different solution structures available in PDB. Only two models were picked (randomly) from each NMR PDB for representation purpose.

Evidence that CoREs are functional

Enrichment of positively charged amino-acids hinted at CoRE's interactions with negatively charged macro-molecules like DNA. To assess this possibility, we performed Gene Ontology (GO) analysis of CoRE-containing and LDR-containing proteins in Homo sapiens. The analysis revealed significant enrichment of DNA-binding functions in CoRE-containing proteins when compared to LDR-containing proteins (fig. 6A). Similarly, DNA-binding domains were over-represented in CoRE contacting proteins, when compared to LDR contacting proteins (figure S3A). The functional enrichment analyses, therefore, suggest the prominent roles of CoREs in DNA binding. To test if CoREs themselves are involved in the protein-DNA interactions, we predicted the DNA-binding residues in the CoREs and the neighboring regions using BindN algorithm. [30] BindN predicts the DNA binding potential of amino-acids through supervised learning of known DNA-protein complexes available in PDB. The analyses revealed greater potential of DNA-binding in the CoRE regions when compared to neighboring regions (fig. 6B, left panel). Similar results were obtained for *Saccharomyces cerviseae* and *Drosophila melanogaster*, suggesting an evolutionary conserved DNA-binding function of CoREs (figure S3B)

Evolutionarily invariance, conformational constrained nature and DNA-binding function of CoRE suggest their functionally important role in the proteins. To test their functional indispensability, we performed SIFT and PolyPhen predictions on the CoREs and the adjacent regions. The comprehensive algorithmic frameworks of SIFT [31] and PolyPhen [32] algorithms predict the effect of single residue mutations in the proteins and have been successfully applied to many systems. SIFT and PolyPhen predictions for the effect of single amino acid mutations in all the human proteins are available publicly (Materials and Methods). We obtained the predictions for CoREs and the adjacent regions. A lower SIFT score represent more intolerant consequence of mutation and a score of 0.05 has been recommended as cut-off for tolerant and intolerant mutation. CoRE regions showed significantly lower average SIFT scores when compared to neighboring regions (inverse of SIFT score was plotted in fig. 6B, right panel). Moreover, out of 19 possible mutations of a residue, a significant number showed deleterious effect (score < 0.05) in CoREs when compared to neighboring regions (figure S3C), strongly suggesting an indispensable nature CoRE peptides. Similar results were obtained by PolyPhen predictions, supporting the functional relevance of CoREs (figure S3C). Searching through disease associated mutations in UniProt revealed that single amino acid changes in RSDELTRH motif leads to decreased DNA-binding of BRG2 transcription factor to its target promoter and consequently leads to demyleniating neuropathies (fig. 6C), suggesting the important role of CoREs in transcriptional regulation through their DNA binding function.

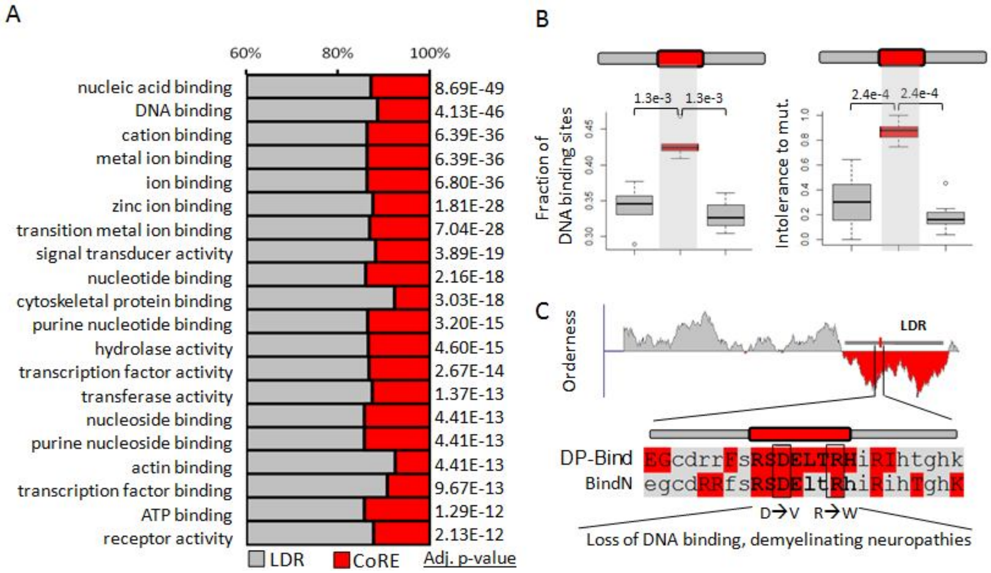

Fig. 6 Functional anatomy of CoRE motifs. (A) Enrichment of Gene Ontology terms in proteins containing CoREs over proteins containing LDRs. P-values are adjusted using Benjamin Hoechberg method of multiple corrections. (C) DNA binding potential (top panel) predicted using BindN algorithm, intolerance to single amino acid substitutions predicted using SIFT algorithm. P-values for the differences between distributions were calculated using Mann-Whitney U test. (C) An example of a CoRE, for which experimental mutation data was available. Upper panel: Predicted intrinsic disorder (red: disordered, grey: ordered). Lower panel: Magnified CoRE region, DNA binding residues (red) predicted using DP-Bind/BindN algorithms and known single amino acid substitutions, which leads to loss of DNA biding and demyelinating neuropathies.

Aberration in symmetry of Glycyl conformation in CoREs

We further focussed on the Glycyl residues which were over-represented in COREs. Glycine is special amino acid, which has greater phi-psi torsional freedom, when compared to L-amino acids, due to lack of C_β atom. The di-peptide units of glycyl occupy symmetrical energy contour around origin (phi=0. Psi=0) in the Ramachandran plot. However, this symmetry is not generally observed in the known three-dimensional (tertiary) protein structures [33]. It is clear that Glycyl occupies preferred foci in the bridge regions, BR (-

120<=phi<=-60, -30 <=psi<=30) and BL (120>=phi<=60. -30<=psi<=30) of the Ramachandran plot, which were not the minimal energy region in the standard Ramachandran plot for a di-peptide of glycyls. [34] Moreover, significant bias was observed for the clustering of phi-psi values in the BL region when compared to BR region. It is possible that Glycine could serve as surrogate for the conformation of a D-amino acid since only Glycine can opt such conformation and, therefore, could have been evolutionarily selected to stabilize certain conformations in proteins [24] particularly turns, which are not allowed for L-amino acids. If this hypothesis is true, the evolutionarily invariant Glycyl residues in CoREs might tend to exhibit greater asymmetry around origin and would show greater skew towards BL than BR conformation. To test this, we performed the following analysis:

1) We plotted the phi-psi values for the glycyl residues present in the COREs. Glycines present in the neighboring regions were taken as control. There was significant difference in the fraction of phi-psi values occupying the regions which are disallowed for the L-amino acids, for the glycines in the COREs and in the controls. Glycines in the COREs were significantly more common in the L-disallowed region, when compared to control glycines from the neighboring regions (fig. 7A).

2) We defined the phi-psi boundaries for BR and BL regions in the Ramachandran plot (Supplementary fig 4A, Materials and Methods) and calculated BL/BR ratio for the glycines in the CORE and control regions. The Glycines in the COREs exhibit significantly greater BL/BR ratio when compared to control regions (fig. 7B). The above two observations are interesting since any mutation of Glycyl to non-glycyl residue would alter the backbone conformation significantly if the Glycine had occurred in the L-disallowed region. The conformational alternation might, consequently, alter the function, if any, of the CORE motif. To support the hypothesis, we analyzed the conformations of all the Glycines which have undergone disease associated Gly-to-nonGly mutations in all human proteins present in UniProt. The analysis revealed a high BL/BR ratio when compared to the neighboring Gly

residues. It has earlier been proposed that such skew might relate to solvent interactions. To demarcate our observation from solvent effect, we analyzed the conformation of Glycines, which were on the surface. Once again, the glycines which undergo disease causing mutations showed significantly greater BL/BR ratio than the control Glycines (figure S4B). The analyses consistently support our hypothesis that the mutation of Glycines in BL conformation could alter the function by altering the local backbone conformation.

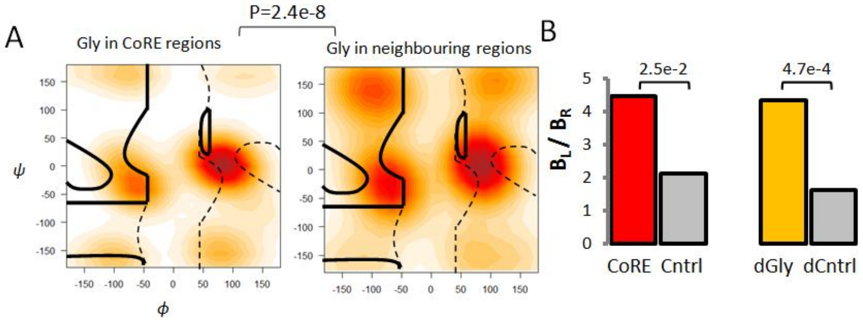

Fig. **7** Role of Glycyl residues in CoREs. **(A)** Ramachandran plots of glycyl residues in CoREs (left) and neighboring regions (right). Solid line represents the L-allowed region and dotted line signifies the allowed region for glycyl residues. P-value for the difference in number of phi-psi values in L-allowed and L-disallowed region was calculated using Fisher's exact test **(B)** Ratio of left handed bridge region to right handed bridge region for the glycyls in CoREs, neighboring control regions (Cntrl), disease associated sites (dGly) and the glycyls adjacent to disease sites(dCntrl).

Chapter 4. Discussion

Intrinsically disordered regions are generally viewed as flexible chains having diverse interactions of high specificity and low affinity. While low affinity is well explained through thermodynamic arguments, high specificity of these regions is hard to be explained considering lack of three dimensional folds. It has been proposed that specificity is brought through short linear motifs, though the structural and mechanistic aspects of these motifs remain elusive, partly due to difficulty in delineating the functional motif from that of false positives and mainly because they fall in disordered regions. Is there any structural perquisite for molecular recognition through short motifs? Our stringent approach of identifying evolutionarily invariant peptides with <1% false discovery rate provides us a good starting point to attempt for conformational analysis of short recognition motifs. Our observations suggested that the CoREs mould into crescent like folded conformations, which were significantly conserved regardless of where they were found. The observations were further supported through analysis of multiple NMR models of CoREs present in disordered regions. We, therefore, propose that long disordered regions could contain several CoRE like short and highly ordered anchors for molecular recognition. The proposal could fill the conceptual gap between the intrinsic structural disorder and the structural basis of molecular recognition.

Short motifs have been shown to mediate several different functions, like phosphorylation, sub-cellular localization, protein-protein interactions, cleavage/degradation, ligand binding etc. However, their role in DNA-binding is not well explored. Our analysis reports short motifs, which are primarily present in DNA binding protein and exhibit greater potential to bind to DNA consistently across genera. We propose that CoREs might serve as a bait for primary DNA recognition by long disordered regions, which could facilitate the search for DNA via a fly-casting like or monkey-bar mechanisms. We argue that CoREs might endow high specificity for DNA-protein interactions on following grounds: (i) COREs are pre-primed to retain their conformations, which might help them dock to specific sites on DNA. (ii) Several known short DNA binding motifs like Dm-Antp, Dm-NK2, Ce-SKN-1 and

AT-hook, which are enriched in R, K and G amino acids, specifically bind to minor groove of DNA through non-electrostatic component of Gibbs energy of binding, in contrast to non-specific electrostatic binding of charged residues of disordered regions. (iii) Transcription factors, in the absence of their long disordered regions, lose their ability to recognize their specific target DNA sites, suggesting that disordered regions need disordered regions for specificity, possibly through short motifs like CoREs. (iv) Non-specific binding of CoREs would fail to explain strong evolutionary conservation of CoREs. (v) Non-trivial role of CoREs was also supported by SIFT and PolyPhen predictions. Interestingly, single amino acid changes (D→ V or R→W) in RSDELTRH CoRE sequence in BRG2 transcription factor lead to loss of DNA-binding and strongly associate with demyelinating neuropathies like Charcot-Marie-Tooth, supporting potential functional roles for CoREs. Therefore, CoREs could play important roles in transcriptional regulatory dynamics in the nucleus. Being short in nature, CoREs can rapidly evolve from non-CoRE regions through substitution mutations, as opposed to slower evolution through duplication events, and henceforth, could efficiently rewire transcriptional regulatory network during evolutionary time-scale.

Nonetheless, we do not rule out the possibility that CoREs are not required for initial recognition of DNA as hypothesized above. Other possibilities could include: (i) CoREs might be required as accessory DNA binding motifs, which stabilize DNA-protein complex. (ii) CoREs could play important role in the process of coupled folding and binding by serving as rigid hinges to endow suitable orientation of disordered regions to facilitate couple folding and binding. (iii) CoREs might not mediate DNA recognition at all, but are required for the appropriate assembly of transcriptional protein complex, wherein CoRE mediates inter-domain or protein-protein communications (ref). (iv) CoREs might bind to a particular DNA-binding domain of the same protein or the interacting protein to inhibit the DNA-binding and thus modulate the protein-DNA interactions via competitive inhibition mechanism (ref). (v) Conservation of CoREs might also relate to post-translation modifications of Arg or Lysine residues, which plays critical role in chromatin remodelling (ref) and DNA binding [35].

Short motifs identified in this study are de novo. Neither they overlapped with SLiM motifs nor they are similar to SLiMs in composition. Glycine is the most enriched amino acid in the CoREs when compared to disordered regions or SLiMs. Their preferred left handed bridge conformations further distinguish CoREs from their adjacent regions. Since left handed bridge conformation cannot be occupied by any non-glycine L-amino acid, mutations in BL glycines to a non-glycine amino acid would tend to shift the BL conformation to L-allowed range of Ramachandran plot. Conserved glycines in CoREs, therefore, might have been evolutionarily selected to acquire certain conformations, which are not allowed for other amino-acids. If this hypothesis is true, the glycines located at disease associated mutation sites might show a skew towards BL conformation. This, indeed, was observed when we analyzed the conformation of all the disease associated glycine substitutions in Human proteins. Distinct neighborhood has been suggested for the glycines having negative and positive phi values, though we did not observe any such preference in our data (figure S4C) suggesting that BL conformation of Glycines in CoREs is independent of its immediate neighboring residues.

To gain further insights to the role of glycyl residues in CoREs, we focussed on the evolutionarily invariant core "RGRP" of AT-Hook DNA binding motif, which is also present in our dataset. PDB search for this motif revealed 24 instances in non-redundant entries. The conformational analysis of the motif suggested two different preferred conformations; one structurally alignable to energy minimized unbound conformation (C-shape) of AT-hook and other alignable to DNA-bound conformation (extended) of AT hook (figure S4D). Phi-Psi values clearly showed a bias towards BL region for the unbound conformation and shifted to non-BL for the DNA bound conformation (figure S4E), suggesting that the BL conformation of Glycyl residue might play a role in ascribing proper crescent like fold for DNA recognition and that the Glycine dependent conformational change might further facilitate the induced fit of AT-hook to the DNA groove. On similar lines, role of Glycine in specific DNA-binding has also been postulated by others too [35-37].

Structured nature of the CoREs also hints at an interesting therapeutic strategy, wherein molecular recognition can be blocked or downplayed using: 1) Small molecules complementary to CoRE, or 2) Small molecule mimicking the conformation of a particular CoRE. Indeed, Nutilin an analogue of MDM2 motif has been successfully applied to inhibit the interaction of MDM2 to P53, thereby stabilizing the P53 and consequently provoking senescence of cancer cells [38]. Similarly integrin inhibitor Cilngitide, an analogue of Arg-Gly-Asp motif, induces cellular detachment and apoptosis of epithelial and glioma cells [39]. Given that disordered regions can implicate in diseases through promiscuous interactions [40], constrained conformation of CoREs could help designing antagonistic molecules to inhibit certain macro-molecular interactions.

Chapter 5. Conclusion

We conclude that that long disordered regions can contain short motifs, which are invariantly conserved across vertebrates and invertebrate proteomes. These motifs are pre-disposed to retain their conformations, as against the prevailing view of multi-conformational dynamics of disordered regions, and might serve to make primary contact during recognition of macromolecules, DNA in particular, in the cell. Mutations in these motifs are predicted to be functionally intolerant, suggesting their possible implications in diseases. We further hypothesize that distinct backbone conformation of Glycines, might have been evolutionarily selected and could play important role in ascribing suitable conformation for macromolecular recognition.

Chapter 6. Future prospective

To consolidate the conformationally constrained nature of CoREs, Molecular Dynamics will be performed in near future. Presence of evolutionarily conserved and conformationally ordered recognition elements might explain the specific interactions by intrinsically disordered regions. It is exp[ected that in the absence of specific recognition elements intrinsically disordered proteins can have promiscuous interactions in the cell. Do ordered recognition elements limit the promiscuous interactions of IDPs? A thorough analysis of protein-protein iteractions and protein-DNA interaction networks might shed some light on this. Moreover, a single amino-acid change in a non-CoRE region might lead to sudden appearance of CoRE during evolution. It would be interesting to study how appearance and disappearance of CoREs in the evolution rewires transcriptional regulatory networks.

Chapter 7. Caveats

One may have used TrEMBL database rather than Uniprot which contain more proteins to increase the number of CoREs. We deliberately avoid that because Uniprot contains manually curetted proteins. Due to this we avoided the inclusion of peptides that are wrongly stated.

The peptides are present in the disorder regions so, we were not able to get sufficient PDB hits for all peptides. We included the NMR analysis to account for these disorder regions. However, given more solved PDB structure the statistics could have been improved.

The mutation data used is the prediction from sift and polyphen data generated from human genome based on conservation and occurrence of SNPs. This may not be the actual mutation effect. The mutation analysis needs to be done experimentally if possible.

Bibliography

1. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding.* Biochimica et biophysica acta, 2010. 1804(6): p. 1231-64.
2. Dunker, A.K., et al., *Intrinsically disordered protein.* Journal of molecular graphics & modelling, 2001. 19(1): p. 26-59.
3. Chen, J.W., et al., *Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions.* Journal of proteome research, 2006. 5(4): p. 879-87.
4. Uversky, V.N., *Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?* Cellular and molecular life sciences : CMLS, 2003. 60(9): p. 1852-71.
5. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics.* Structure, 2003. 11(11): p. 1453-9.
6. Janin, J. and M.J. Sternberg, *Protein flexibility, not disorder, is intrinsic to molecular recognition.* F1000 biology reports, 2013. 5: p. 2.
7. Miao, J., J. Klein-Seetharaman, and H. Meirovitch, *The optimal fraction of hydrophobic residues required to ensure protein collapse.* Journal of molecular biology, 2004. 344(3): p. 797-811.
8. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. 41(3): p. 415-27.
9. Huang, Y. and Z. Liu, *Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism.* Journal of molecular biology, 2009. 393(5): p. 1143-59.
10. Vavouri, T., et al., *Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity.* Cell, 2009. 138(1): p. 198-208.
11. Tompa, P., *Unstructural biology coming of age.* Current opinion in structural biology, 2011. 21(3): p. 419-25.

12. Shoemaker, B.A., J.J. Portman, and P.G. Wolynes, *Speeding molecular recognition by using the folding funnel: the fly-casting mechanism.* Proceedings of the National Academy of Sciences of the United States of America, 2000. 97(16): p. 8868-73.

13. Fuxreiter, M., P. Tompa, and I. Simon, *Local structural disorder imparts plasticity on linear motifs.* Bioinformatics, 2007. 23(8): p. 950-6.

14. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* Journal of molecular biology, 2006. 362(5): p. 1043-59.

15. Ekman, D., et al., *What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?* Genome biology, 2006. 7(6): p. R45.

16. Tompa, P., *Intrinsically disordered proteins: a 10-year recap.* Trends in biochemical sciences, 2012. 37(12): p. 509-16.

17. Brown, C.J., et al., *Evolution and disorder.* Current opinion in structural biology, 2011. 21(3): p. 441-6.

18. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database (Oxford), 2011. 2011: p. bar009.

19. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. 21(16): p. 3433-4.

20. Prilusky, J., et al., *FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded.* Bioinformatics, 2005. 21(16): p. 3435-8.

21. Reddy, B.V., S. Datta, and S. Tiwari, *Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability.* Protein engineering, 1998. 11(12): p. 1137-45.

22. Lindorff-Larsen, K. and J. Ferkinghoff-Borg, *Similarity measures for protein ensembles.* PloS one, 2009. 4(1): p. e4203.

23. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. 22(12): p. 2577-637.

24. Horng, J.C., F.W. Kotch, and R.T. Raines, *Is glycine a surrogate for a D-amino acid in the collagen triple helix?* Protein science : a publication of the Protein Society, 2007. 16(2): p. 208-15.

25. Eswar, N., et al., *Influence of solvent molecules on the stereochemical code of glycyl residues in proteins.* Proteins, 2002. 49(3): p. 326-34.

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

27. Sandhu, K.S., *Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins.* J Mol Recognit, 2009. 22(1): p. 1-8.

28. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins, 2001. 42(1): p. 38-48.

29. Tompa, P., *Intrinsically unstructured proteins evolve by repeat expansion.* Bioessays, 2003. 25(9): p. 847-55.

30. Wang, L. and S.J. Brown, *BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences.* Nucleic Acids Res, 2006. 34(Web Server issue): p. W243-8.

31. Ng, P.C. and S. Henikoff, *Predicting deleterious amino acid substitutions.* Genome Res, 2001. 11(5): p. 863-74.

32. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. 7(4): p. 248-9.

33. Ramakrishnan, C., N. Srinivasan, and D. Prashanth, *Conformation of glycyl residues in globular proteins.* Int J Pept Protein Res, 1987. 29(5): p. 629-37.

34. Karplus, P.A., *Experimentally observed conformation-dependent geometry and hidden strain in proteins.* Protein Sci, 1996. 5(7): p. 1406-20.

35. Dery, U., et al., *A glycine-arginine domain in control of the human MRE11 DNA repair protein.* Mol Cell Biol, 2008. 28(9): p. 3058-69.

36. Dongre, M., et al., *Evidence on how a conserved glycine in the hinge region of HapR regulates its DNA binding ability: lessons from a natural variant.* J Biol Chem, 2011. 286(17): p. 15043-9.

37. Qin, F., et al., *Specific recognition between intrinsically disordered LEF and DNA.* Phys Chem Chem Phys, 2012. 14(2): p. 538-45.

38. Vassilev, L.T., et al., *In vivo activation of the p53 pathway by small-molecule antagonists of MDM2.* Science, 2004. 303(5659): p. 844-8.

39. Oliveira-Ferrer, L., et al., *Cilengitide induces cellular detachment and apoptosis in endothelial and glioma cells mediated by inhibition of FAK/src/AKT pathway.* J Exp Clin Cancer Res, 2008. 27: p. 86.

40. Babu, M.M., et al., *Intrinsically disordered proteins: regulation and disease.* Curr Opin Struct Biol, 2011. 21(3): p. 432-40.

Appendix

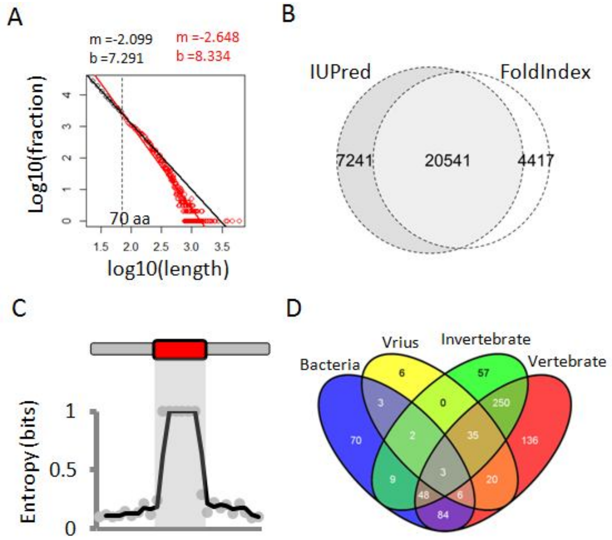

Figure S1. (A) log-log plot for the distribution of LDR lengths. At 70 aa, the difference between the two slopes (black and red corresponding to "< cutoff" and "≥ cut-off" values respectively) was maximum (=0.545) (B) Overlap of number of LDRs predicted by IUPred and FoldIndex algorithms. (C) Sequence conservation in the CoREs and the neighboring regions. (D) Venn diagram for the number of CoREs (FDR ≤ 0.01) found in major life kingdoms represented in UniProt.

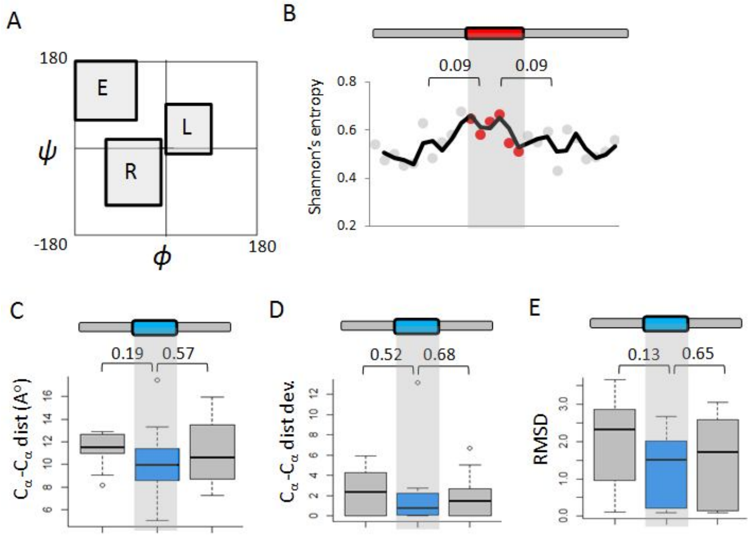

Figure S2. (A) Ramachandran plot showing the phi-psi ranges for extended sheet (E), right handed helical (R) and left handed (L) conformations. **(B)** Conservation of surface accessibility of CoREs and the adjacent regions**. (C-E)**Structural analyses of control peptides (scrambled CoREs): **(C)** End to end distances, **(D)** deviation in end-to end distances and, **(E)** root mean squared deviation of superimpositions of scrambled CoREs found in Protein Data Bank. P-values were calculated using Mann-Whitney U test. Examples of these control peptides are shown in figure 2B (in grey color).

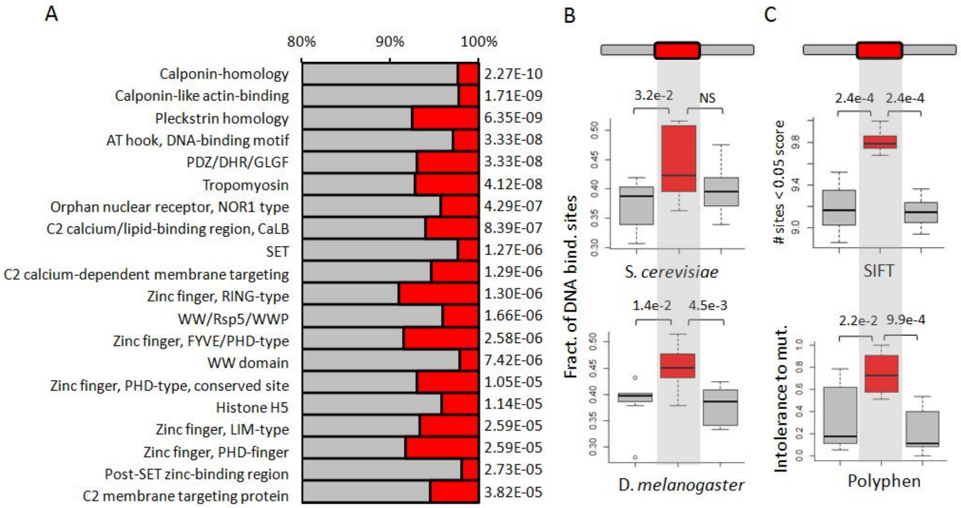

Figure S3. (A) Enrichment of protein domains in CoRE containing human proteins, when compared to LDR containing proteins. The analysis was done using FatiGO. **(B)** Predicted DNA Binding potential of CoREs in yeast and fly. **(C)** Predictions for the effect of mutation using SIFT (left panel, number of amino acid changes showing the SIFT score < 0.05) and PolyPhen (right panel) algorithms. P-values were calculated using Mann-Whitney U test.

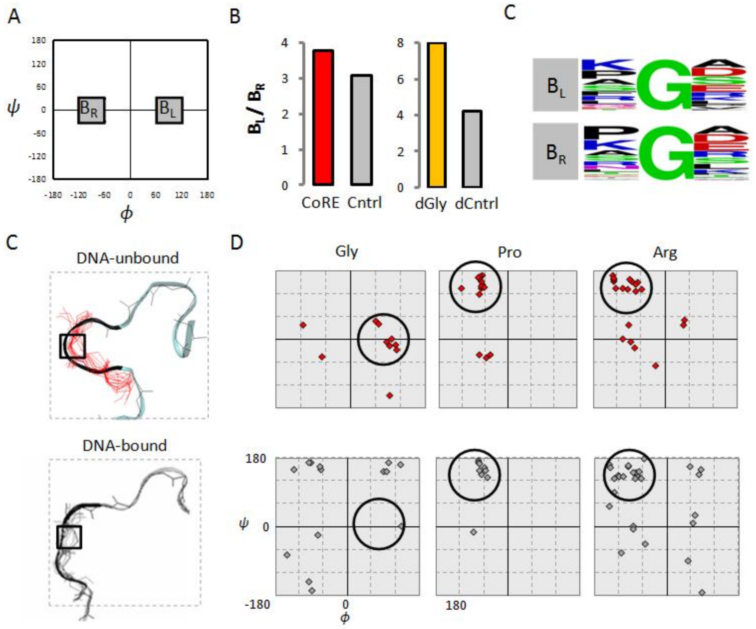

Figure S4. (A) Ramachandran plot showing phi-psi ranges for right handed bridge (B_R) and left handed bridge (B_L) conformations. **(B)** B_R/ B_L ratio for the surface-exposed glycines in CoREs, adjacent regions (Cntrl), disease associated sites (dGly) and corresponding control sites (dCntrl). **(C)** Sequence neighborhood of glycines in the B_L and B_R conformations. **(D)** Conformations of AT-hook core motif (RGRP) alignable to DNA unbound (upper panel) and bound conformations (lower panel). **(E)** Ramachandran plots of Glycine, Proline and Arginine residues occurring in RGRP motifs in DNA unbound (upper panel) and DNA-bound (lower panel) conformation in high resolution structures of RGRP motifs available in PDB.