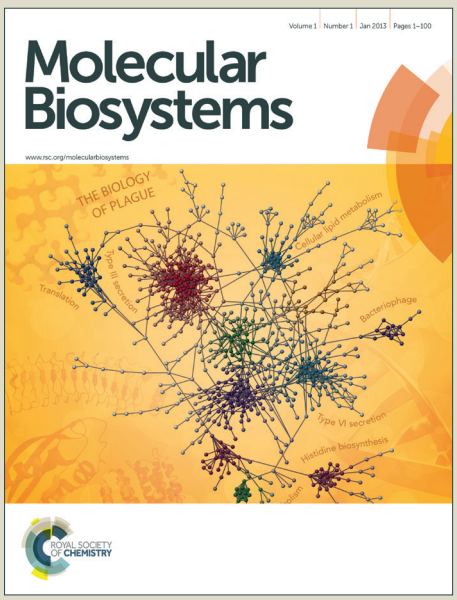
Molecular BioSystems

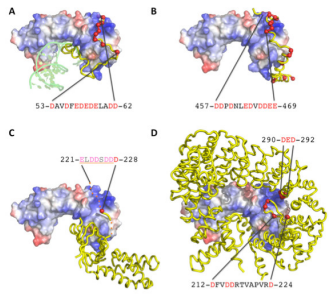
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Table of contents entry



D/E-rich proteins might be involved in DNA mimicry, mRNA processing and regulation of the transcription complex.

Molecular BioSystems

Paper

Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry

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

DOI: 10.1039/x0xx00000x

www.rsc.org/

Chia-Cheng Chou[ab] and Andrew H.-J. Wang[ab*]

Aspartic acid and glutamic acid repeats in proteins exhibit strong negative charge distribution and they may play special biological roles. From 39,684 unique structural data in the RCSB Protein Data Bank (PDB), 173 structures were found to contain ordered D/E-rich repeat structures, and 57 of them were related to DNA/RNA functions. The frequency of occurrence of glutamic acid (36.90%) was higher than that of aspartic acid (27.02%). Glycine (2.38%), alanine (2.68%), valine (3.54%), leucine (5.57%), and isoleucine (3.34%), but not methionine (0.91%), were the most abundant hydrophobic residues. The available complex structures suggested that D/E-rich proteins might be involved in DNA mimicry, mRNA processing and regulation of the transcription complex. The region surrounding the D/E-rich repeat sequences play important roles in the binding specificity toward the target proteins. The numbers and composition of aspartic acid and glutamic acid might also affect binding properties. Aspartic acid and glutamic acid are disorder-promoting residues in the intrinsically disorder proteins. Our findings suggest the D/E-rich repeats are unique components of intrinsically disordered proteins which are involved in the gene regulation and could serve as potential druggable fragments or drug targets.

Introduction

Aspartic acid and glutamic acid repeats in proteins are highly important owing to their negative charges, and many studies have revealed the properties underlying their interaction with metal ions. For instance, the muscular protein aspolin in zebrafish is an extreme case in that most of its sequence comprises aspartic acid (179 Asp among 186 total amino acids). Aspolin has since been identified as a paralog of a histidine-rich calcium binding protein and regulates the calcium concentration in the sarcoplasmic reticulum of striated muscle.[1] Aspartic acid-rich proteins are also major components of the soluble organic matrix of mollusk shells,[2] and at least 10 different proteins have been identified.[3] Additionally, GARPs (glutamic acid-rich proteins) in the outer segments of rod photoreceptors have been found to have low affinity but high capacity for Ca^{2+} binding.[4] Furthermore, the overexpression of SH3BGR (SH3 domain binding glutamic acid-rich protein) might alter specific functions of muscle tissue and therefore take part in the pathophysiology of muscular hypotonia in Down syndrome.[5] SMAPs (small acidic proteins), which contain additional Phe residues in the C-terminal aspartic acid-rich domain, are responsive for auxin signaling in the root of *Arabidopsis thaliana*.[6] Glutamic acid-rich proteins are also markers for the diagnosis of parasitic malaria[7] and babesia.[8]

Some reports further indicate that D/E-rich proteins are involved in the regulation of gene expression. Two acidic regions of the chromatin insulator Cp190 mediate the association and dissociation of chromosomes in *Drosophila melanogaster*.[9] The truncated C-terminal nucleolin cleaved by MMP7 (matrix metalloproteinase 7) at D255 in the D/E-rich sequence induces MMP9 expression to promote tumor malignancy.[10] Loss of function of mouse Pagr1a (Pax-interacting protein 1-associated glutamate rich protein 1a) reduces the expression of BMP2 (bone morphogenetic protein 2), the regulator of extraembryonic development.[11] In another aspect, our previous studies of acidic DNA mimic proteins, including UGI/SAUGI,[12,13] ICP11,[14] DMP19,[15] DMP12[16] and ARN,[17] have already shown that surface Asp and Glu residues can mimic the phosphates of the DNA backbone.[18,19] These residues do not have a repeated primary sequence but are sequentially distributed and form a pattern that mimics the phosphate backbone in DNA on the protein surface. Some translation initiation factors have also been found to resemble the shape of tRNA.[20] It is intriguing to speculate whether concentrated D/E-rich repeat sequences could represent similar structural behaviors involved in DNA/RNA-related functions, in addition to their metal ion binding properties.

Most D/E-rich repeats are predicted to be unstructured and therefore their functions are not easily investigated. To determine the possible functions of D/E-rich repeats, we first searched through 14,296 reviewed records from the yeast genome in the UniProt database (http://www.uniprot.org),[21] and 109 protein candidates were filtered with the search criteria of >70% D/E in a 30-aa fragment (Supplementary Table 1). More than half (68, 62.39%) of the candidates were DNA/RNA-related proteins. Five that were annotated as uncharacterized proteins were applied to search in protein-protein interaction databases. It was surprising that all had the potential to bind DNA/RNA-related proteins (Supplementary Table 2), which implies that D/E-rich repeats might have specific biological functions. Accompanying the growth of RCSB Protein Data Bank (PDB, http://www.rcsb.org/pdb/)[22] and also the contribution of structural genomics projects in the past 10 years, comprehensive structural information for various protein structures should enable molecular-level elucidation of structure-function relationships. From the 39,684 unique pieces of data in the PDB, 173 structures were found to contain ordered D/E-rich repeat structures, and 32.94% were related to DNA/RNA functions. The available

[a] Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan.
[b] Core Facilities for Protein Structural Analysis, Academia Sinica, Taipei, Taiwan.
Email: ahjwang@gate.sinica.edu.tw
† Electronic Supplementary Information (ESI) available. See
DOI: 10.1039/x0xx00000x

complex structures suggest that they might be involved in DNA mimicry, mRNA processing and the regulation of the transcription complex. We therefore hypothesize that D/E-rich repeat sequences play important roles in the regulation of gene expression, and detailed composition and sequence specificity are discussed.

Materials and methods

Yeast protein sequence data from the UniProt database

Protein sequences in the yeast genome were downloaded from the UniProt database; 14,296 reviewed datasets were selected in total.

Interacting proteins from protein-protein interaction databases

Information regarding interacting proteins was obtained from four online protein-protein interaction databases, specifically BioGRID (http://thebiogrid.org),[23] IntAct (http://www.ebi.ac.uk/intact/),[24] DIP (http://dip.doe-mbi.ucla.edu/dip/Main.cgi)[25] and STRING (http://string-db.org),[26] which each contain comprehensive information from experimental datasets and computational prediction strategies.

Structural data from the PDB

To understand the structure and function of D/E-rich sequences, 39,684 non-redundant structures released until 2014/07/31 were downloaded from the PDB databank. Only the sequences with 3D coordinates were analyzed. The amino acid sequences were then extracted from the C_α atom in protein coordinates using the modified Python script in pdb-tools (http://code.google.com/p/pdb-tools/). The first model in the NMR data and the first alternative conformer in the X-ray structures were considered to simplify the calculation. Secondary structure information was calculated by the DSSP program.[27] DSSP classified the secondary structures according to 7 types using hydrogen-bonding patterns, but they were merged into α-helix, β-strand and non-structured only in this study.

The definition of "D/E-rich repeat"

Amino acid repeats (AARs) are abundant in protein sequences. Luo and Nijveen classified AARs into three categories depending on the characteristics of the repeat units.[28] The first approach is to classify AARs according to the similarity among the repeat units. The second is based on the distance between adjacent units. The third takes the complexity of the sequence pattern of the repeat units into consideration. The simple repeats in the third type are often called simple sequences (SSs) or low complexity regions (LCRs), which are composed of limited sets of amino acids including repeats of one or more residues.[29]

Our study aimed to identify the SSs/LCRs composed of aspartic acid or glutamic acid. To avoid short tandem repeat motifs such as WD, DEAD, DExxD, metal binding motifs and others, the search criterion focused on longer sequences and defined to find 10 amino acids of fragments with more than seven D/E residues. Then, the overlapped fragments were merged together. The filtered sequences were checked carefully, and redundant chains and duplicated proteins were excluded. Proteins containing the metal ions were also analyzed. For the classification of secondary structures, the sequence lengths for helices or strands were limited to more than two residues.

Results and discussions

The distribution of D/E-rich proteins in PDB

The search results returned 173 proteins as hits (141 from X-ray, 27 from NMR and 5 from cryo-EM), and detailed descriptions are listed in Supplementary Table 3. The number of residue of these proteins varied from 23 (4CAY_C and 2XZE_Q) to 1476 (3H0G_M) and the pI values from 3.13 (4CAY_C) to 10.06 (2GD5_D). The statistics of the distributions of the lengths, secondary structures and amino acid compositions are shown in Fig. 1A. According to our strategy, the D/E-rich repeat lengths ranged from 10 to 18, and most of them (82 fragments) contained 10 amino acids. *Neurospora crassa* plasma membrane ATPase (1MHS_B) was the only D/E-rich repeat with two fragments. In total, 50.57% of the fragments were composed of a helical structure, 39.08% were unstructured, and 9.77% were β-strands (Fig. 1B). Only one fragment contained both a helix and a β-strand. Evaluation of the frequency of occurrence of individual residues in the fragments indicated that glutamic acid occurred more often (36.90%) than aspartic acid (27.02%) (Fig. 2A). Glycine (2.38%), alanine (2.68%), valine (3.54%), leucine (5.57%), and isoleucine (3.34%), but not methionine (0.91%), were the most abundant hydrophobic residues (Fig. 2B). The histidine residue, which is common in metal binding motifs such as zinc fingers, was rare (0.25%) among all the fragments. In contrast with the observed frequency of occurrence in vertebrates, the ratio of small (glycine, alanine and proline), polar uncharged (serine, threonine, asparagine, glutamine) and positively charged (histidine, lysine and arginine) residues are significantly lower (Fig. 2C).

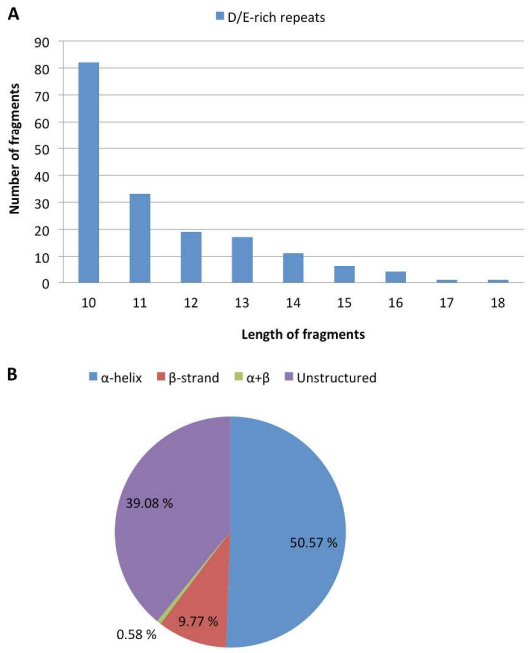


Fig. 1 (A) Distribution of the length/number and (B) secondary structures of the D/E-rich repeats identified from the PDB.

Interestingly, all of the fragments had been exposed to solvent, which implicated possible functions involving interaction with other molecules rather than the maintenance of intra-molecular contacts. Among the 173 protein candidates, 58 (33.52%) proteins contained metal ions, of which the top three ions were Mg^{2+} (20, 11.56%), Ca^{2+} (19, 10.98%) and Zn^{2+} (11, 6.35%). However, most did not interact with D/E-rich repeats. The Calx Na^{+}/Ca^{2+} exchanger (2DPK_A)

bound four Ca^{2+} with two aspartic acids, two glutamic acids, two main-chain oxygens in the D/E-rich repeats and three water molecules. The D/E-rich 892-DQDDDDDPDTE-902 in the BK channel (3NAF_A) only coordinated one Ca^{2+} through two aspartic acids and two main-chain oxygens. The human thrombospondin-2 (1YO8_A) bound 30 Ca^{2+} but only at specific repeat motifs comprising mainly aspartic acids. Metal binding capacity appeared to be conferred by a type of tandem repeat, not simple D/E repeat sequences. Wang *et al*. previously indicated that S is better than G in acidic peptides to inhibit the formation of calcium oxalate monohydrate crystals, which is the primary mineral of kidney stones.[30] In our results, the frequency of S was less than G (1.21% and 2.37%, respectively), which agreed with the observation that the structural D/E-rich repeat does not primarily function in metal binding.

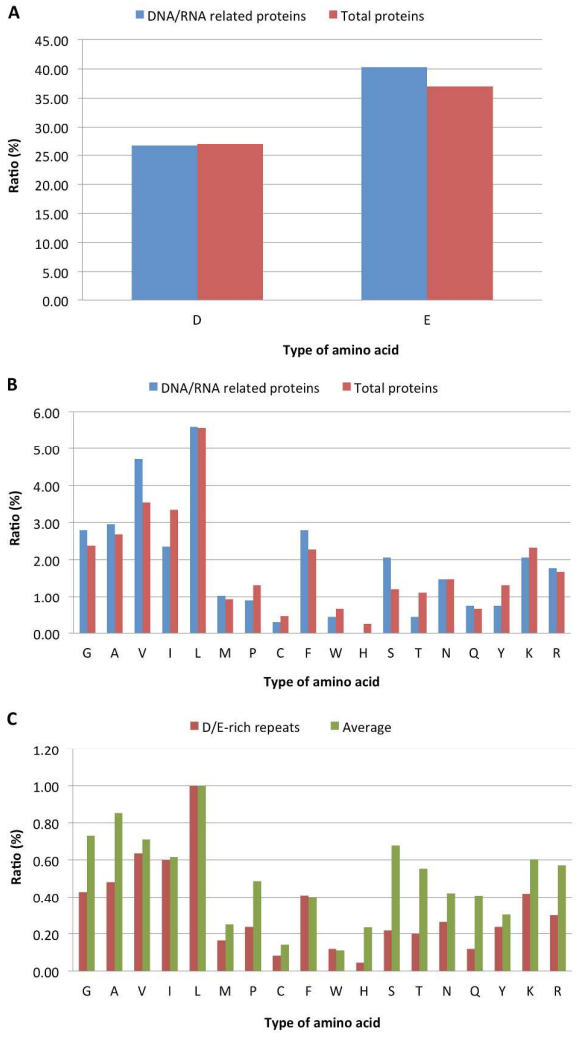


Fig. 2 Frequency of occurrence of individual residues in D/E-rich repeats. (A) D and E only; (B) the remaining 18 amino acids. (C) Comparison with the observed frequency in vertebrates and normalized by leucine.

Biological significance of D/E-rich repeat-containing proteins

Many proteins have more than one LCR. Toll-Riera *et al*. suggested that low-complexity sequences contribute to the formation of novel protein coding sequences.[31] Ekman and co-workers noted that highly connected 'hub' proteins contain an increased fraction of LCRs compared to non-hub proteins.[32] Coletta *et al*. suggested that LCR-containing proteins tend to have more interactions in PPI (protein-protein interaction) databases.[33] In their report, LCRs were found to have position-dependent roles. Centrally-located LCRs are enriched for transcription-related GO terms, whereas terminal LCRs are enriched for translation and stress response-related terms.

In our results, the functions of 173 proteins were spread over varied cellular activities, including histone chaperones, nucleosome and chromatin packing, transcription factors, DNA methylation, DNA helicases, ribosomal proteins, RNA polymerase subunits, translation factors, RNA cleavage, RNA splicing, Ca^{2+} regulation, transporters, protein kinases/phosphatases, oxygenases/reductases, hydrolases/dehydrolases, proteases/peptidases, synthases, transferases, ubiquitin binding proteins, protein assembly, chaperones, nuclear import/export, protein-protein interaction and others. Compared with previous studies investigating the relationship between the position and function of LCRs, only 24 (13.87%) were located at the N- or C-terminus. This might be due to

Table 1 Functions of 173 D/E-rich repeat-containing proteins in the PDB

Function	Number
DNA/RNA related proteins	
Histone, nucleosome, and chromatin protein	11
Transcription factor	7
DNA methylation	2
DNA helicase and recombinase	4
ssDNA binding protein	1
RNA polymerase subunit	3
Ribosomal protein	7
Translation factor	6
RNA cleavage	4
RNA splicing	4
mRNA triphosphatase	1
RNA binding	3
Signal recognition particle	4
Enzymes	
Protein kinase and phosphatase	6
Oxygenase and reductase	7
Hydrolase and dehydrolase	5
Isomerase	2
Protease and peptidase	4
Synthase	8
Transferase	3
Kinase and phosphorylase	3
Mutase	1
Enolase	1
Glucosidase	1
Cystathionase	1
Penicillin binding protein	2
Lactamase	1
Ubiquitin binding protein	8
Chaperone	6
Nuclear import and export	4
Protein assembly	10
Ca^{2+} regulation	5
Transporter	10
Sensor	2
Lipid transfer	1
Signaling	2
Electron transfer	5
Prion inhibition	2
Others	9
Unknown function	7

the flexibility of terminal structures and therefore their invisibility in crystal structures. Interestingly, DNA/RNA-related proteins comprised 32.94% of total proteins (57 of 173, Table1). The frequency of occurrence of glutamic acid in these 57 proteins was higher (40.21%) than aspartic acid (26.66%) compared with the total data, although there was no further evidence to illustrate the structural significance. From these proteins, certain complex structures were revealed to be important for the function of D/E-rich repeats, and they are summarized below.

1. DNA mimicry

a. DNA methyltransfrase-1 (3SWR_A and 3PT6_B)

Maintenance of genomic methylation patterns is mediated primarily by DNA methyltransferase-1 (DNMT1).[34] Unmethylated DNA is excluded from the active site of DNMT1 by the binding of the CXXC domain, whereas the presence of an acidic autoinhibitory CXXC-BAH1 linker (701-EADDDEEADDD-711) positioned directly between the DNA and the active site prevents the entrance of DNA into the catalytic pocket. Comparison of the mouse DNMT1-DNA and M.HhaI-DNA complexes indicated that the autoinhibitory linker could mimic the phosphate backbone and push the DNA substrate away (Fig. 3A).

b. Histone chaperones Chz1 and ANP32E-ZID (4CAY_C and 2JSS_B)

Fig. 3B-C shows the structures of two chaperone proteins, Chz1 (chaperone for Htz1-H2B) and ANP32E-ZID (acidic nuclear phosphoprotein 32 kilodalton e - Z interacting domain), which bind H2A.Z-H2B. Chz1 forms a long irregular chain capped by two short helices and uses both positively and negatively charged residues to stabilize the histone dimer.[35] ANP32E regulates H2A.Z deposition at promoters and strikingly preserves enhancers and insulator sites free of H2A.Z nucleosomes.[36] The model for the nucleosome suggests that the two chaperones not only block interactions with histones but also prevent binding to DNA, which suggests that the D/E-rich repeats (3-EDSESDMDD-11 and 21-EGEEEEDDLAEID-33 in Chz1, and 224-EEIQDEEDDDDYVE-237 in ANP32E) could mimic the behavior of DNA. Another chaperone, FACT (facilitates chromatin transcription), contains an E790D/K in the D/E-rich repeat that suppresses the effects of H3-L61W, which disturbs the structure of nucleosomes, presumably by destabilizing the interface between H3 and H4.[37]

c. BRCA2-interacting protein DSS1 (1MJE_B)

DSS1 (deleted in split-hand/split foot syndrome) binds BRCA2 (breast cancer susceptibility gene 2) in an extended conformation involving interaction with the helical domains OB1 (oligonucleotide/oligosaccharide binding domain 1) and OB2. The binding is characterized by hydrophobic interactions and also a large number of acidic DSS1 residues (14-EEDDEFEEFPAE-25 and a few D/E residues in the region from 40-61) interacting with the basic groove on BRCA2. The binding residues on both proteins are conserved, and one BRCA2 residue, R2580, is mutated in cancer.[38] DSS1 might potentially mimic the regulation of the accessibility of a subset of the putative DNA binding sites on the helical and OB1 domains (Fig. 3D).

2. mRNA processing

a. Antitoxin RelB (1WMI_D)

Prokaryotic chromosomes contain toxin-antitoxin loci, which are composed of two genes organized in an operon that encodes a stable toxin and a labile cognate antitoxin, respectively.[39] In the archaeal

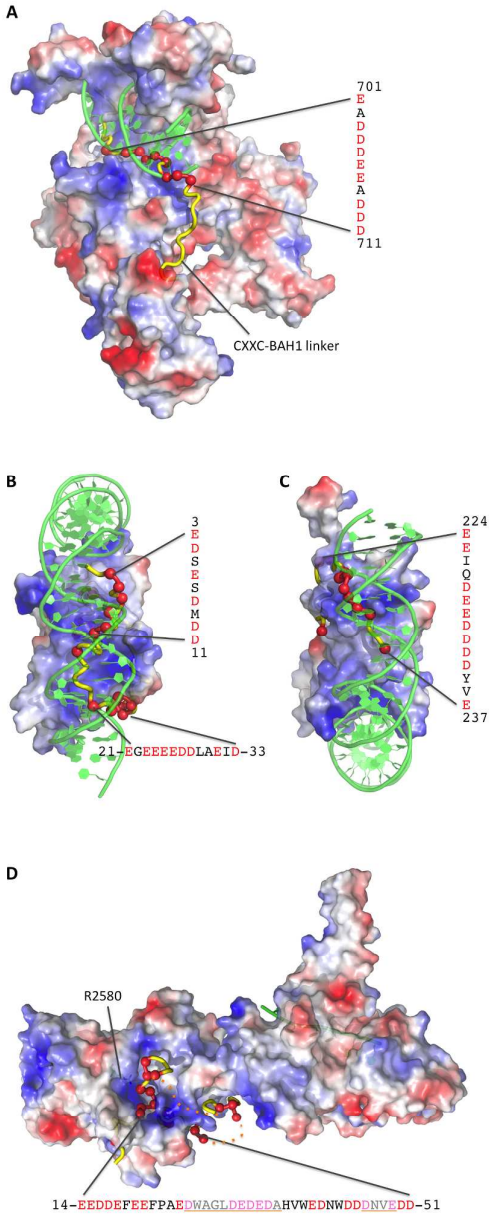


Fig. 3 DNA mimic proteins. Complex models of (A) human DNMT1 and dsDNA substrate, (B) Chz1 chaperone, H2A/H2B, and the chromosome fragment, (C) ANP32E-ZID, H2A/H2B, and the chromosome fragment, and (D) the crystal structure of the BRCA2, DSS1, and ssDNA fragment complex. The D/E-rich containing proteins are represented as yellow loops and the Asp/Glu residues are represented as red spheres. For DNMT1, only the autoinhibitory linker is presented as the yellow loop, and the remaining structure is present as the electrostatic surface potential (and also the binding partners in each complex). The conserved R2580 in DNM1 is shown as a ball-and-stick. The dsDNA substrate, chromosome and ssDNA fragments are in transparent green. The unmodelled residues are highlighted with orange dots. The Asp/Glu residues in the sequence label are in red. The missing sequence is in gray and marked with an orange underline, and the Asp/Glu residues are in pink.

*rel*BE system, RelE assists in the degradation of mRNA in a codon-specific manner positioned at the ribosome A-site.[40] The interface of RelB and RelE represents a high degree of charge complementary. The negatively charged E31, D33, D35 and E40 in the D/E-rich repeat (30-EERDEDITEEE-40) of RelB interact with K47, R58, R65 and K81 in RelE. Mutagenesis experiments have also shown that R40, L48, R58, R65 and R85 in RelE appear to be essential residues for functional activity. A model with mRNA (from 4V7J) indicates that RelB inhibits the binding of mRNA and might behave as an RNA mimic (Fig. 4A).

b. Splicing factor U2AF65 (4FXW_C)

The essential splicing factors U2AF65 and SF1 cooperatively bind consensus sequences at the 3' ends of introns. Phosphorylation of SF1 in a highly conserved 'SPSP' motif enhances its interaction with U2AF65 and pre-mRNA.[41] The W22 of ULM (U2AF ligand motif) near the SPSP motif in SF1 is buried within a hydrophobic pocket, and positively charged residues interact with the negatively charged residues E394, E397, D401, and E405 of the UHM (U2AF homology motif), which are found in the identified D/E-rich repeats (387-EELLDDEEYEEIVEDVRDE-405) of U2AF65 (Fig. 4B). The acidic surface of the D/E-rich repeats might also be used for the regulation of pre-mRNA binding.

3. Regulation of transcription complex

a. TBP-TAF1/Brf1 (4B0A_A, 1NGM_N, 3OC3_A and 1NH2_B)

The transcription factor complex TFIID is composed of a TBP (TATA-box binding protein) and 13 TAFs (TBP-associated factors) and provides a regulatory platform for transcription initiation. The yeast TAND2 region (the D/E-rich repeat 53-DAVDFEDEDELADD-62) of TAF1 independently exerts an inhibitory effect on transcription and competes with TFIIA in binding TBP (Fig. 5A and 5C).[42] The TBP surface groove is also critical for interactions with the preinitiation complexes (PICs) of PolI, PolII and PolIII. Notably, the D/E-rich repeat (457-DDPDNLEDVDDEE-469) of Brf1 (RNA polymerase III transcription initiation factor 90 kDa subunit) in the PolIII complex[43] binds the same groove on TBP but with an inversed orientation (Fig. 5B). Another regulator, Mot1, also competes for the same surface but uses two discrete fragments (Fig. 5D).

b. SEM1/DSS1 (3T5V_F and 3T5X_B)

SEM1 (suppressor of exocyst mutants), the DSS1 homolog, has different properties to assist with the assembly of the TREX-2 transcription-export complex. SEM1 stabilizes the Thp1 cofactor, promotes interaction with Sac3 and provides a platform that mediates nucleic acid binding.[44] Comparison of DSS1 and PCID2 (Thp1 homolog) indicated that their C-terminal portions (61-EENWDDVEVDDD-72 in SEM1 and 40-EDNWDDDNVEDD-51 in DSS1) both bind the same positively charged groove on Thp1/PCID2 (Fig. 6A-B). The other N-terminal D/E-rich repeat, 30-EEDDEFED-37, is also visible. Co-expression of SEM1/DSS1 is essential to obtain soluble Thp1/PCID2. SEM1/DSS1, which is also associated with a wide range of conserved complexes, including the 19S proteasome lid and the CSN, eIF3 and integrator complexes.[45]

Conclusions

Many types of AARs have previously been identified. For example, leucine-rich repeats are important for protein-protein interactions, and the repeat fragment in zinc finger transcription factors binds to the *cis*-elements of DNA promoters. PolyQ repeats are found in the Forkhead box protein, the androgen receptor, and result in several

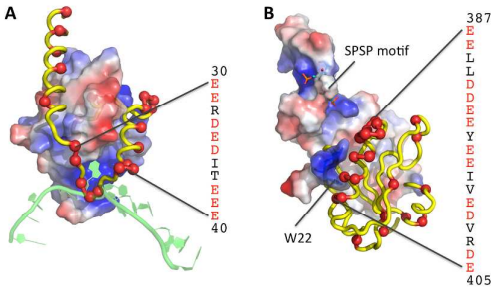


Fig. 4 mRNA processing proteins. The complex model of (A) RelB, RelE and mRNA, and the crystal structure of (B) U2AF65 and SF1. The representation diagram is the same as in Fig. 3. The mRNA fragment is in green. The SPSP motif and the W22 in U2AF65 are represented as ball-and-stick.

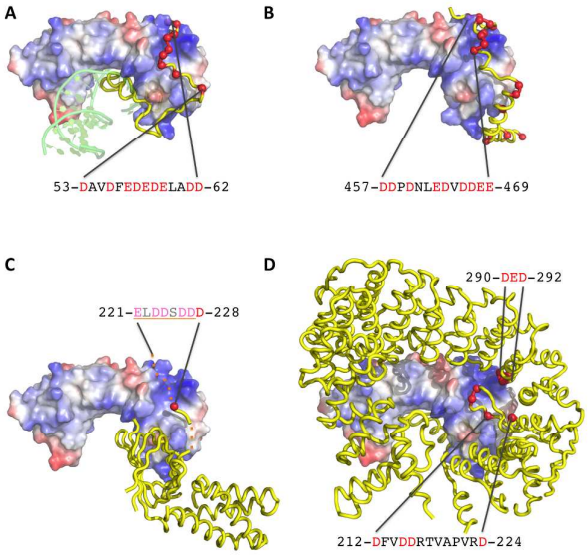


Fig. 5 TAFIID transcription initiation complexes. The crystal structures of TBP with (A) TAF1, (B) Brf1, (C) TFIIA, and (D) Mot1. The representation diagram is the same as in Fig. 3. The dsDNA fragment is in green.

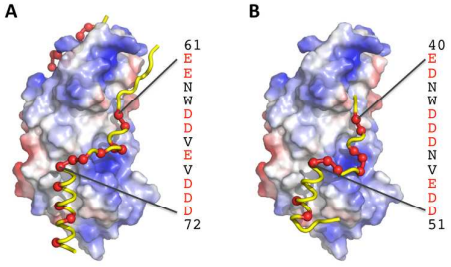


Fig. 6 TREX-2 transcription export complexes. The crystal structures of (A) SEM1, Thp1 and Sac3, and the complex model of (B) DSS1, Thp1 and Sac3. The representation diagram is the same as in Fig. 3.

neurological disorders such as mental retardation, Huntington's disease (HD), inherited ataxias and muscular dystrophy.[46] Other single AARs, including polyL, polyA and polyH, can also be found in many other proteins.[22,23] Zhu and Karlin also pointed out that the percentage of proteins with at least one significant linear charge cluster is about 20 ~ 25% in most eukaryotic species, about 35% in Drosophila, and about 6 ~ 8% in *E.coli* among protein primary sequences.[47]

A report from Huntley and Golding in 2002 indicated that only 15 of 4442 (0.3%) possible eukaryotic PDB protein structures existed with complete structural information for the LCR sequences.[48] LCRs are normal but are not present at the same high frequency within structural databases, and therefore LCRs are largely under-represented in PDB. The lack of structures indicates this type of simple sequence might produce disordered structures. Wootton[29] and Saqi[49] indicated that when a LCR was present in a protein sequence, it formed well-ordered structures that were often helical. Karlin and Burge found that this type of homopolymer is primarily composed of the amino acid residues Q, N, S, T, P, H, G, A, D and E, and the length is generally less than 20 residues long.[50] Polyalanine is often used as a standard model. Under aqueous conditions, short stretches of polyalanine will form helical structures; however, this can be nucleation dependent, and with an appropriate sequence they can form a stable beta-plated sheet.[51,52]

From our current analysis of PDB, the major functions of structural D/E-rich repeats appear to be DNA/RNA-related. The complex structures also provide some structural evidence. Although many do not have complex structures, the current results suggest it is possible that D/E-rich repeats participate in the regulation of protein complexes for gene expression and might mimic the shape of DNA or RNA, which is similar to the findings in our previous studies of many DNA-mimic proteins. The surrounding domain can maintain the binding specificity. For example, nucleolin is cleaved by MMP7 at D255 of 252-DDEDDDDEDDE-262. The remaining D/E-rich sequence might result in a loss of control and lead the truncated nucleolin to bind the 3'UTR of tumor-promoting mRNA and induce tumor malignancy. It is also possible that the D/E-rich repeat around the cutting site might increase the catalytic activity of MM7 due to its capacity for Ca^{2+} and Zn^{2+} binding. In Cp190, the N-terminal BTB/POZ (bric-abrac/poxvirus and zinc finger) domain and D-rich repeat (263-286, 11 aspartic acids and 1 glutamic acid) mediate the association, and the C-terminal E-rich (592-1096, 53 aspartic acids and 92 glutamic acids) domain is required for dissociation. The number and composition of aspartic acids and glutamic acids might also affect the binding properties.

Transcription factors are highly attractive but difficult drug targets due to the intrinsically disordered nature of their binding sites. These regions might bind their partners with a relatively short fragment that has high curvature upon binding.[53] Nutlin-2, a *cis*-imidazoline analog, has been shown to mimic the crucial residues of p53 fragment when bound to Mdm2, with two bromophenyl groups fitting into Mdm2 in the same pockets as p53's Trp23 and Leu26, and with an ethyl-ether side chain filling the spot normal taken by Phe19.[54] c-Myc-Max transcriptional complex can also be inhibited through the disruption of binding domains which are both disordered without dimer formation.[55] In our studies, D/E-rich repeats appear to be key regions to modulate the activity of TFIID complexes, particularly for TBP binding. Another transcription factor, TFIIE, was also found to share the same binding region on TFIIH with p53 through its acidic domain.[56] Radivojac *et al.* analyzed the intrinsically disordered proteins (IDPs) and grouped the amino acid residues into order-promoting (C, W, Y, I, F, V, L, H, T and N), disorder-promoting (D, M, K, R, S, Q, P and E) and neutral (A and G). H, T, N and D could also be considered neutral due to the lower difference criterion.[57] Homma *et al*. also proposed that the intrinsically disordered regions in nuclear proteins tend to be negatively charged and are involved in protein-protein interactions.[58] The non-D/E residues in the D/E-rich repeats could be important factors due to the transition of the frequency of occurrence. The incorporation of unnatural amino acids in the non-D/E region could be a way to increase the binding specificity, which has been applied on the generation of therapeutic antibody conjugates.[59] These findings suggest the D/E-rich repeats are unique components of intrinsically disordered proteins which are involved in the gene regulation and could serves as potential druggable fragments or drug targets.

Acknowledgements

This work was supported in part by Academia Sinica and a grant from the National Science Council (NSC100-2325-B-001-029) to AHJW for Core Facilities for Protein Structural Analysis at Academia Sinica, Taiwan. We also thank the National Research Program for Biopharmaceuticals (supported by grant no. NSC 10102325-B-492-001) as well as the National Center for High Performance Computing at the National Applied Research Laboratories of Taiwan for providing computing resources. DSSP and PyMol are components of software packages provided by SBGrid Consortium.[59]

References

1 S. Kinoshita, E. Katsumi, H. Yamamoto, K. Takeuchi and S. Watabe, *Mar. Biotechnol.*, 2011, **13**, 517-526.
2 S. Weiner, *Calcif. Tissue Int.*, 1979, **29**, 163-167.
3 B. A. Gotliv, N. Kessler, J.L. Sumerel, D. E. Morse, N. Tuross, L. Addadi and S. Weiner, *Chembiochem.*, 2005, **6**, 304-314.
4 S. Haber-Pohlmeier, K. Abarca-Heidemann, H. G. Körschen, H. K. Dhiman, J. Heberle, H. Schwalbe, J. Klein-Seetharaman, U. B. Kaupp and A. Pohlmeier, *Biophys J.*, 2007, **92**, 3207-3214.
5 P. Scartezzini, A. Egeo, S. Colella, P. Fumagalli, P. Arrigo, D. Nizetic, R. Taramelli and A. Rasore-Quartino, *Hum. Genet.*, 1997, **99**, 387-392.
6 A. Nakasone, M. Kawai-Yamada, T. Kiyosue, I. Narumi, H. Uchimiya and Y. Oono, *J. Plant Physiol.*, 2009, **166**, 1307-1313.
7 J. H. Kattenberg, I. Versteeg, S. J. Migchelsen, I. J. González, M. D. Perkins, P. F. Mens and H. D. Schallig, *MAbs*., 2012, **4**, 120-126.
8 A. A. Mousa, S. Cao, G. O. Aboge, M. A. Terkawi, A. El Kirdasy, A. Salama, M. Attia, M. Aboulaila, M. Zhou, K. Kamyingkird, P. F. Moumouni, T. Masatani, S. A. El Aziz, W. M. Moussa, B. Chahan, S. Fukumoto, Y. Nishikawa, S. S. El Ballal and X. Xuan, *Exp. Parasitol*., 2013, **135**, 414-420.
9 D. Oliver, B. Sheehan, H. South, O. Akbari and C. Y. Pai, *BMC Cell Biol*., 2010, **11**, 101-116.
10 T. I. Hsu, S. C. Lin, P. S. Lu, W. C. Chang, C. Y. Hung, Y. M. Yeh, W. C. Su, P. C. Liao and J. J. Hung, *Oncogene*, 2015, **34**, 826-837.
11 A. Kumar, M. Lualdi, J. Loncarek, Y. W. Cho, J. E. Lee, K. Ge and M. R. Kuehn, *Dev. Dyn*., 2014, **243**, 937-947.
12 C. D. Mol, A. S. Arvai, R. J. Sanderson, G. Slupphaug, B. Kavli, H. E. Krokan, D. W. Mosbaugh and J. A. Tainer, *Cell*, 1995, **82**, 701-708.
13 H. C. Wang, K. C. Hsu, J. M. Yang, M. L Wu, T. P. Ko, S. R. Lin and A. H. J. Wang, *Nucleic Acids Res.*, 2014, **42**, 1354-1364.

14 H. C. Wang, H. C. Wang, T. P. Ko, Y. M. Lee, J. H. Leu, C. H. Ho, W. P. Huang, C. F. Lo and A. H. J. Wang, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 20758-20763.
15 H. C. Wang, T. P. Ko, M. L Wu, C. H. Ho, S. C. Ku, H. J Wu and A. H. J. Wang, *Nucleic Acids Res.*, 2012, **40**, 5718-5730.
16 H. C. Wang, M. L Wu, T. P. Ko and A. H. J. Wang, *Nucleic Acids Res.*, 2013, **41**, 5127-5138.
17 C. H. Ho, H. C. Wang, T. P. Ko, Y. C. Chang and A. H. J. Wang, *J. Biol. Chem.*, 2014, **289**, 27046-27054.
18 C. D. Putnam and J. A. Tainer, *DNA repair*, 2005, **4**, 1410-1420.
19 H. C. Wang, C. H. Ho, K. C. Hsu, J. M. Yang and A. H. J. Wang, *Biochemistry*, 2014, **53**, 2865−2874.
20 D. E. Brodersen and V. Ramakrishnan, *Nat. Struct. Mol. Biol.*, 2003, **10**, 78-80.
21 UniProt Consortium, *Nucleic Acids Res*., 2014, **42**, D191-198.
22 F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Jr. Meyer, M. D. Brice, J. R. Rodgers, O, Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol*., 1977, **112**, 535-542.
23 A. Chatr-Aryamontri, B. J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. M. Rust, M. S. Livstone, R. Oughtred, K. Dolinski and M. Tyers, *Nucleic Acids Res*., 2012, **41**, D816-823.
24 S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard and H. Hermjakob, *Nucleic Acids Res.*, 2011, **40**, D841-846.
25 L. Salwinsk, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, *Nucleic Acids Res.*, 2004, **32**, D449-451.
26 A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen, *Nucleic Acids Res.*, 2013, **41**, D808-815.
27 R. P. Joosten, T. A. te Beek, E. Krieger, M. L. Hekkelman, R. W. Hooft, R. Schneider, C. Sander and G. Vriend, *Nucleic Acids Res*., 2011, **39**, D411-419.
28 H. Luo and H. Nijveen, *Brief Bioinform.*, 2013, **15**, 582-591.
29 J. Wootton, *Curr. Opin. Struct. Biol*., 1994, **4**, 413-421.
30 L. Wang, S. R. Qiu, W. Zachowicz, X. Guan, J. J. DeYoreo, G. H. Nancollas and J. R. Hoyer, *Langmuir*, 2006, **22**, 7279-7285.
31 M. Toll-Riera, N. Radó-Trilla, F. Martys and M. M. Albá, *Mol. Biol. Evol*., 2011, **29**, 883-886.
32 D. Ekman, S. Light, A. K. Björklund and A. Elofsson, *Genome Biol*., 2006, **7**, R45.
33 A. Coletta, J. W. Pinney, D. Y. W. Solís, J. Marsh, S. R. Pettifer and T. K. Attwood, *BMC Syst. Biol.*, 2010, **4**, 43-55.
34 J. Song, O. Rechkoblit, T. H. Bestor and D. J. Patel, *Science*, 2011, **331**, 1036-1040.
35 Z. Zhou, H. Feng, D. F. Hansen, H. Kato, E. Luk, D. I. Freedberg, L. E. Kay, C. Wu and Y. Bai, *Nat. Struct. Mol. Biol.*, 2008, **15**, 868-869.
36 A. Obri, K. Ouararhni, C. Papin, M. L. Diebold, K. Padmanabhan, M. Marek, I. Stoll, L. Roy, P. T. Reilly, T. W. Mak, S. Dimitrov, C. Romier and A. Hamiche, *Nature*, 2014, **505**, 648-653.
37 D. J. Kemble, F. G. Whitby, H. Robinson, L. L. McCullough, T. Formosa and C. P. Hill, *J. Biol. Chem*., 2013, **288**, 10188-10194.
38 H. Yang, P. D. Jeffrey, J. Miller, E. Kinnucan, Y. Sun, N. H. Thoma, N. Zheng, P. L. Chen, W. H. Lee and N. P. Pavletich, *Science*, 2002, **297**, 1837-1848.
39 H. Takagi, Y. Kakuta, T. Okada, M. Yao, I. Tanaka and M. Kimura, *Nat. Struct. Mol. Biol.*, 2005, **12**, 327-331.
40 K. Pedersen, A. V. Zavialov, M. Y. Pavlov, J. Elf, K. Gerdes, and M. Ehrenberg, *Cell*, 2003, **112**, 131–140.
41 W. Wang, A. Maucuer, A. Gupta, V. Manceau, K. R. Thickman, W. J. Bauer, S. D. Kennedy, J. E. Wedekind, M. R. Green and C. L. Kielkopf, *Structure*, 2013, **21**, 197-208.
42 M. Anandapadamanaban, C. Andresen, S. Helander, Y. Ohyama, M. I. Siponen, P. Lundström, T. Kokubo, M. Ikura, M. Moche and M. Sunnerhagen, *Nat. Struct. Mol. Biol.*, 2013, **20**, 1008-1014.
43 Z. S. Juo, G. A. Kassavetis, J. Wang, E. P. Geiduschek and P. B. Sigler, *Nature*, 2003, **422**, 534–539.
44 A. M. Ellisdon, L. Dimitrova, E. Hurt and M. Stewart, *Nat. Struct. Mol. Biol*., 2012, **19**, 328-336.
45 E. Pick, K. Hofmann and M. H. Glickman, *Mol. Cell*, 2009 **35**, 260–264.
46 T. Shimohata, O. Onodera and S. Tsuji, *Neuropathology*, 2000, **20**, 326-333.
47 Z. Y. Zhu and S. Karlin, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 8350-8355.
48 M. A. Huntley and G. B. Golding, *Proteins*, 2002, **48**, 134-140.
49 M. Saqi, *Protein. Eng*, 1995 **8**, 1069 –1073.
50 S. Karlin and C. Burge, *Proc Natl Acad Sci USA*, 1996, **93**, 1560-1565.
51 C. A. Rohl, W. Fiori and R. L. Baldwin, *Proc. Natl. Acad. Sci. USA*, 1999, 96, 3682-3687.
52 S. E. Blondelle, B. Forood, R. A. Houghten and E. Perez-Paya, *Biochemistry*, 1997, **36**, 8393-8400.
53 A. K. Dunker and V. N. Uversky, *Curr. Opin. Pharmacol*., 2010, **10**, 782-788.
54 L. T. Vassilev, *Cell Cycle*, 2004, **3**, 419-421.
55 S. J. Metallo, *Curr. Opin. Chem. Biol.*, 2010, **14**, 481-488.
56 M. Okuda, A. Tanaka, M. Satoh, S. Mizuta, M. Takazawa, Y. Ohkuma and Y. Nishimura, *EMBO. J*., 2008, **27**, 1161-1171.
57 P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky and A. K. Dunker, *Biophys. J.* 2007, **92**, 1439-1456.
58 K. Homma, S. Fukuchi, K. Nishikawa, S. Sakamoto and H. Sugawara, *Mol Biosyst.*, 2012, **8**, 247-255.
59 K. Wals and H. Ovaa, *Front. Chem.*, 2014, **2**, 15.
60 A. Morin, B. Eisenbraun, J. Key, P. C. Sanschagrin, M. A. Timony, M. Ottaviano and P. Sliz, *eLife*, 2013, **2**, e01456.