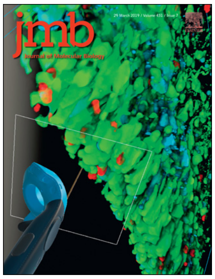

Dynamic Studies on Intrinsically Disordered Regions of Two Paralogous Transcription Factors Reveal Rigid Segments with Important Biological Functions

Snigdha Maiti [1,†], Bidisha Acharya [1,†], Veda Sheersh Boorla [2], Bharat Manna [3], Amit Ghosh [3] and Soumya De [1]

*1 - **School of Bioscience,** Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India*
*2 - **Department of Chemical Engineering,** Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India*
*3 - **School of Energy Science and Engineering,** Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India*

***Correspondence to Soumya De:** School of Bioscience, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India. somde@iitkgp.ac.in*
https://doi.org/10.1016/j.jmb.2019.02.021
Edited by Richard W. Kriwacki

Abstract

Long stretches of intrinsically disordered regions (IDRs) are abundantly present in eukaryotic transcription factors. Although their biological significance is well appreciated, the underlying structural and dynamic mechanisms of their function are still not clear. Using solution NMR spectroscopy, we have studied the structural and dynamic features of two paralogous HOX transcription factors, SCR and DFD, from *Drosophila*. Both proteins have a conserved DNA-binding homeodomain and a long stretch of functionally important IDR. Using NMR dynamics, we determined flexibility of each residue in these proteins. The flexibility of the residues in the disordered region is not uniform. In both proteins, the IDRs have short stretches of consecutive residues with relatively less flexibility, that is, higher rigidity. We show that one such rigid segment is specifically recognized by another co-transcription factor, thus highlighting the importance of these rigid segments in IDR-mediated protein–protein interactions. Using molecular dynamics simulation, we further show that the rigid segments sample less conformations compared to the rest of the residues in the disordered region. The restrained conformational sampling of these rigid residues should lower the loss in conformational entropy during their interactions with binding partners resulting in sequence specific binding. This work provides experimental evidence of a “rigid-segment” model of IDRs, where functionally important rigid segments are connected by highly flexible linkers. Furthermore, a comparative study of IDRs in paralogous proteins reveals that in spite of low-sequence conservation, the rigid and flexible segments are sequentially maintained to preserve related functions and regulations of these proteins.

Introduction

An intrinsically disordered region (IDR) in a protein is a stretch of amino acids that remains unfolded even in the presence of other folded domains in the complete protein. In eukaryotes, at least one-third of all proteins contain long IDRs [1]. Recent studies have shown that proteins containing IDRs play important roles in cell signaling events and regulation of various biological processes [2]. IDR functions may arise from a specific disordered conformation, from inter-conversion between several disordered conformations, and transitions between disordered and ordered states [3]. IDRs are involved in multiple interactions with diverse partners and are known to play key roles in protein–protein interaction networks [4]. Large multi-protein complexes employ these IDRs to assemble the component proteins [5]. One such assembly where IDRs play important roles is the recruitment of cofactors by transcription factors during regulation of gene expression.

Transcription factors are DNA-binding proteins that control gene expression. These proteins bind to specific promoter or enhancer sequences in the DNA and subsequently recruit the transcriptional machinery. Due to their central role in many biological processes, transcription factors are under tight regulation, disruption of which may lead to several diseases ranging from developmental disorders to cancer. Transcription factors in eukaryotes have long stretches of disordered regions that are necessary for their function [6]. For example, IDRs of 30 residues or more are found in 92% and 96% of transcription factors in *Homo sapiens* and *Drosophila melanogaster*, respectively [7]. Typically, proteins in a transcription factor family have a conserved DNA-binding domain that recognizes very similar DNA sequences [8]. The altered DNA-binding specificity, leading to paralog-specific biological functions, is provided by the appended regions that may contain other folded domains and IDRs [9]. These IDRs regulate transcription factor function via post-translational modifications such as phosphorylation, ubiquitination, and sumoylation [10]; interaction with other transcription factors and proteins [11]; and auto-inhibition [8]. All of these processes require specific sequences on IDRs to interact with enzymes and proteins. The underlying mechanism by which IDRs achieve such site-specific interactions in the absence of any folded conformation is still not clear and is the focus of the present study.

In this study, using solution NMR spectroscopy, we have characterized the structure and dynamics of two transcription factors from *Drosophila*, Sex Combs Reduced (SCR) and Deformed (DFD). Both have long stretches of disordered regions. These proteins belong to the family of HOX transcription factors that play key roles in the morphological development of all bilateral animals. In *Drosophila*, SCR controls the development of the adjacent segments namely labial and prothorax [12], while DFD is required for the development of the maxillary and mandibular segments [13]. SCR and DFD have 417 and 586 amino acids, respectively, and contain a conserved 60-residue DNA-binding homeodomain. The remaining residues, based on sequence, are predicted to be intrinsically disordered. For SCR, it has been shown that the homeodomain and its preceding ~30 residues are sufficient to carry out most of its *in vivo* functions such as homeotic transformations, transcriptional regulation, and protein–protein interactions [14]. Moreover, these N-terminal 30 residues have been shown to regulate the functional specificities in both SCR and DFD [15]. Hence, SCR$^{K298–K384}$ and DFD$^{T337–K426}$ constructs, which include the 60-residue homeodomain and ~30 residues preceding it, have been used in the present study. From a comparative study of these paralogous proteins, we aim to elucidate the important structural and dynamic features of their IDRs that are required for the functional regulation of these transcription factors.

We have completely assigned the ^{1}H, ^{13}C, and ^{15}N chemical shifts of the backbone atoms of SCR$^{K298–K384}$ and DFD$^{T337–K426}$ by solution NMR spectroscopy. Chemical shift-based secondary structure prediction and ^{15}N dynamic studies show that ~30 residues in the N-terminus are completely disordered, while the DNA-binding homeodomain is properly folded in both proteins. Reduced spectral density analysis of the ^{15}N dynamics data revealed variations in flexibility in the disordered region. A closer inspection showed that the disordered region is composed of segments of rigid and flexible residues. Interestingly, one of these segments of rigid residues is conserved in both SCR and DFD, and using NMR titration experiments, we show that this rigid segment is specifically recognized by a co-transcription factor Extradenticle (EXD). Based on our findings in these two paralogous proteins, we propose that IDRs have segments of rigid residues that are functionally important and suggest a straightforward method of their identification from backbone NMR relaxation experiments. Furthermore, using molecular dynamics (MD) studies, we show that residues in the rigid segments sample fewer conformations compared to other residues in the disordered region. Thus, the loss in conformational entropy should be less for these rigid segments enabling them to interact specifically with partner molecules. Overall in this study, we show that IDRs in HOX transcription factors have interspersed rigid segments that are functionally important and report a method to identify them by solution NMR spectroscopy.

Results

Backbone assignments reveal that SCR and DFD have disordered N-terminal region followed by a well-folded homeodomain

The ^{15}N–^{1}H heteronuclear single quantum coherence (HSQC) spectra of SCR$^{K298–K384}$ and DFD$^{T337–K426}$ showed well-dispersed peaks indicating the presence of a folded domain. These ^{15}N–^{1}H HSQC peaks were assigned by standard ^{1}H/^{13}C/^{15}N heteronuclear NMR experiments (Fig. 1). The backbone chemical shifts of ^{1}H$^{\alpha}$, ^{1}H^{N}, ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, and ^{13}CO, nuclei were used to predict the secondary structures of SCR$^{K298–K384}$ and DFD$^{T337–K426}$ using the program MICS (Fig. 2a and e) [16]. In both proteins, the secondary structure prediction shows the presence of three alpha-helices that form the DNA-binding homeodomain. These secondary structures agree well with the crystal structure of SCR in complex with DNA (PDB: 2R5Y). The N-terminal residues of both proteins (SCR: K298 to Y331; DFD: T337 to T374) show lack of secondary structure, highlighting their highly flexible nature. Thus, SCR$^{K298–K384}$ and DFD$^{T337–K426}$ have ~30 disordered residues followed by a properly folded 60-residue homeodomain.

Backbone ^{15}N relaxation studies of SCR and DFD

Amide ^{15}N relaxation data (R_1, R_2, and NOE) were collected for both SCR$^{K298–K384}$ and DFD$^{T337–K426}$

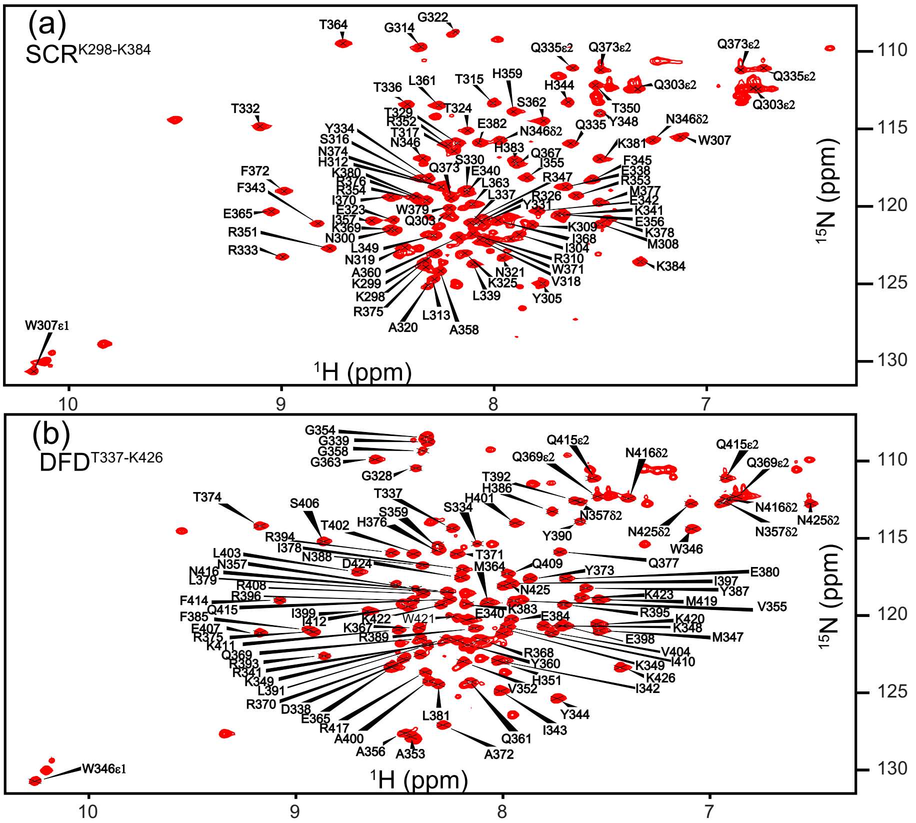

Fig. 1. Assigned ^{15}N–^{1}H HSQC spectra of SCR$^{K298–K384}$ and DFD$^{T337–K426}$. The ^{15}N–^{1}H HSQC spectra of (a) SCR$^{K298–K384}$ and (b) DFD$^{T337–K426}$ were assigned by triple-resonance experiments. The peaks of the homeodomain are well dispersed, while those of the N-terminal residues appear in the 7.5- to 8.5-ppm range in the proton dimension, which is typical for disordered regions.

at 25 °C. The heteronuclear {^{1}H}–^{15}N NOE values are in the range of 0.7 to 1.0 for the homeodomain residues (Fig. 2b and f). This is again consistent with a well-folded domain. For the N-terminal regions in SCR$^{K298–K384}$ (Fig. 2b) and DFD$^{T337–K426}$ (Fig. 2f), the hetero-NOE values progressively decrease as residues get farther from the homeodomain, indicating loss of ordered structure and increased flexibility. Interestingly, a small conserved sequence "YPWMK(R/K)," which is situated ~20 residues away from the folded homeodomain, has positive or very low negative hetero-NOE values in both proteins. The "YPWMK(R/K)" region is flanked with residues with negative hetero-NOE values on either side. This indicates that this sequence is relatively more ordered than the neighboring residues within the disordered N-terminal region.

In both proteins, the transverse relaxation rate constants (R_2) have an average value of ~8 s^{-1} for the homeodomain which decreases to ~2.5 s^{-1} for the N-terminal region (Fig. 2c and g). This further indicates the lack of order in the N-terminal residues. Again the "YPWMK(R/K)" residues have relatively higher R_2 values compared to the neighboring residues, further indicating the presence of a relatively more ordered stretch of residues than the neighboring residues within the disordered N-terminal region. The longitudinal relaxation rate constants (R_1) show much less variation (Fig. 2d and h). The R_1 values for the N-terminal residues are marginally smaller compared to the homeodomain. Collectively, these data indicate that in both proteins the homeodomain is properly folded and has a stable three-dimensional structure, whereas the N-terminal residues are disordered except the conserved "YPWMK(R/K)" sequence that displays relatively higher order compared to the remaining residues in the disordered region.

Reduced spectral density mapping reveals varying degree of flexibility in the disordered regions

Reduced spectral density mapping was done for SCR$^{K298–K384}$ and DFD$^{T337–K426}$ using the corresponding ^{15}N relaxation data sets (Fig. 3 and S1). The extent of motions of the N–H bond vectors for each residue can be estimated from these spectral density functions. For rigid regions, the spectral

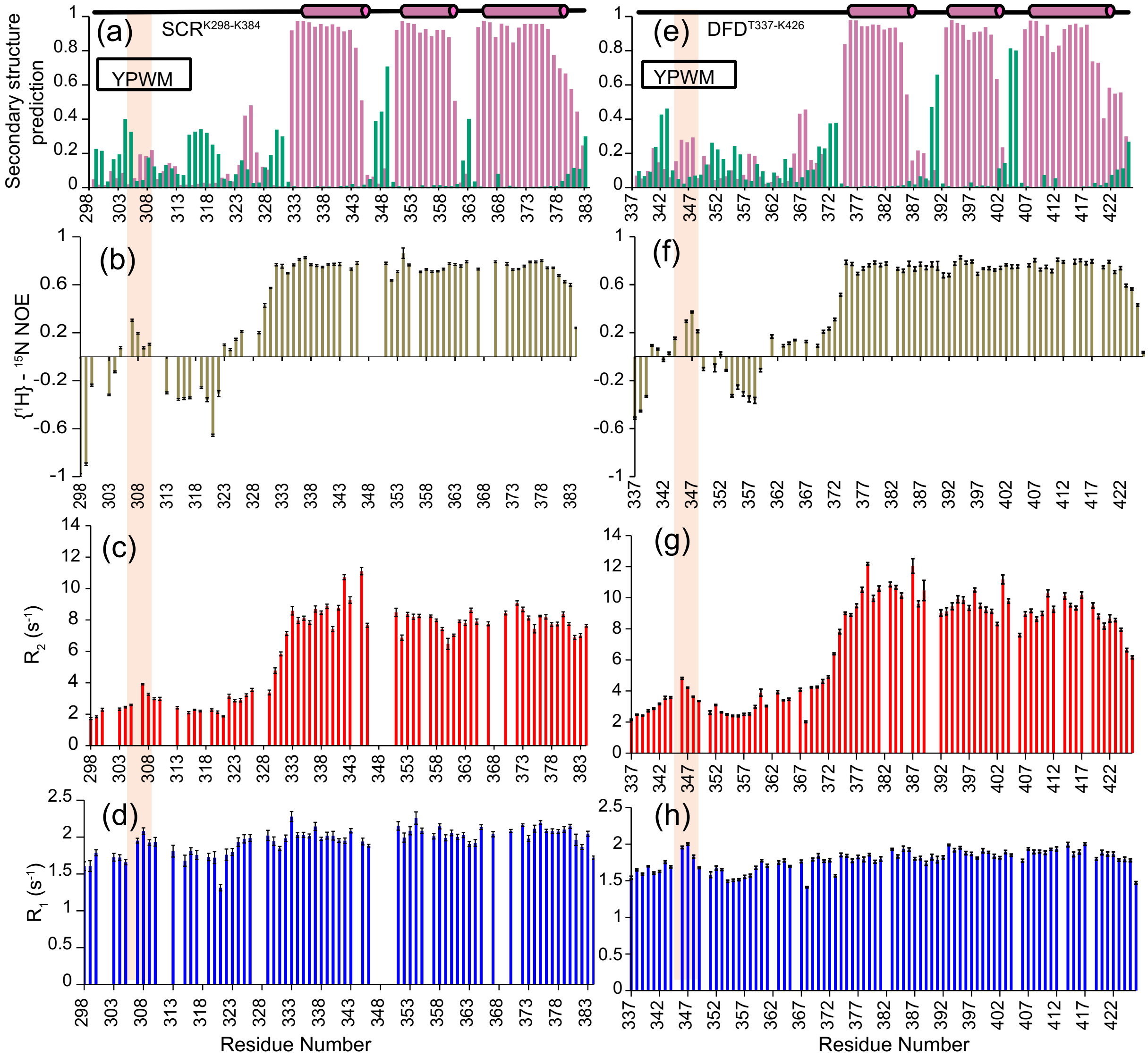

Fig. 2. SCR$^{K298–K384}$ and DFD$^{T337–K426}$ have disordered N-terminus followed by folded homeodomain. Secondary structure predictions of (a) SCR$^{K298–K384}$ and (e) DFD$^{T337–K426}$ based on backbone chemical shifts indicate disordered N-terminal region and folded homeodomain in both proteins. The homeodomain contains three alpha helices (shown as cylinders) that align well with the DNA-bound structure of SCR (PDB: 2R5Y). Heteronuclear ^{15}N-NOE of (b) SCR$^{K298–K384}$ and (f) DFD$^{T337–K426}$ have values greater than 0.7 for most of the homeodomain residues and rapidly decrease for the N-terminal residues. The "YPWM" residues are highlighted and show positive ^{15}N-NOE. Transverse relaxation rate constants (R_2) for (c) SCR$^{K298–K384}$ and (g) DFD$^{T337–K426}$ have an average value of 8/s for the homeodomain that decreases to 2.5/s for the N-terminal residues. Again the "YPWM" and their neighboring residues have significantly higher R_2 values. Longitudinal relaxation rate constants (R_1) for (d) SCR$^{K298–K384}$ and (h) DFD$^{T337–K426}$ are relatively featureless. The disordered region has marginally lower R_1 values compared to the homeodomain.

density function is dominated by the low-frequency components, that is, $J(0)$ and $J(\omega_N)$, while for flexible regions, the higher-frequency component, that is, $J(0.87\omega_H)$ also makes significant contribution [17]. The R_1, R_2, and hetero-NOE measurements were done at 14.1-T (600-MHz) field strength, and thus, the spectral densities were measured at 0-, 60-, and 522-MHz frequencies. In both proteins, $J(0)$ has an average value of ~2.3 ns/rad for the homeodomain, which decreases to ~0.55 ns/rad for the N-terminal region (Fig. S1). On average, the $J(0)$ and $J(\omega_N)$ terms of the N-terminal residues decrease by factors of 0.24 and 0.79 to those of the homeodomains, respectively. In contrast, the $J(0.87\omega_H)$ terms of the N-terminal residues are approximately 3.43 times larger than those of the homeodomain. Thus, the higher-frequency component $J(0.87\omega_H)$ has increased contribution in the N-terminal region. This indicates

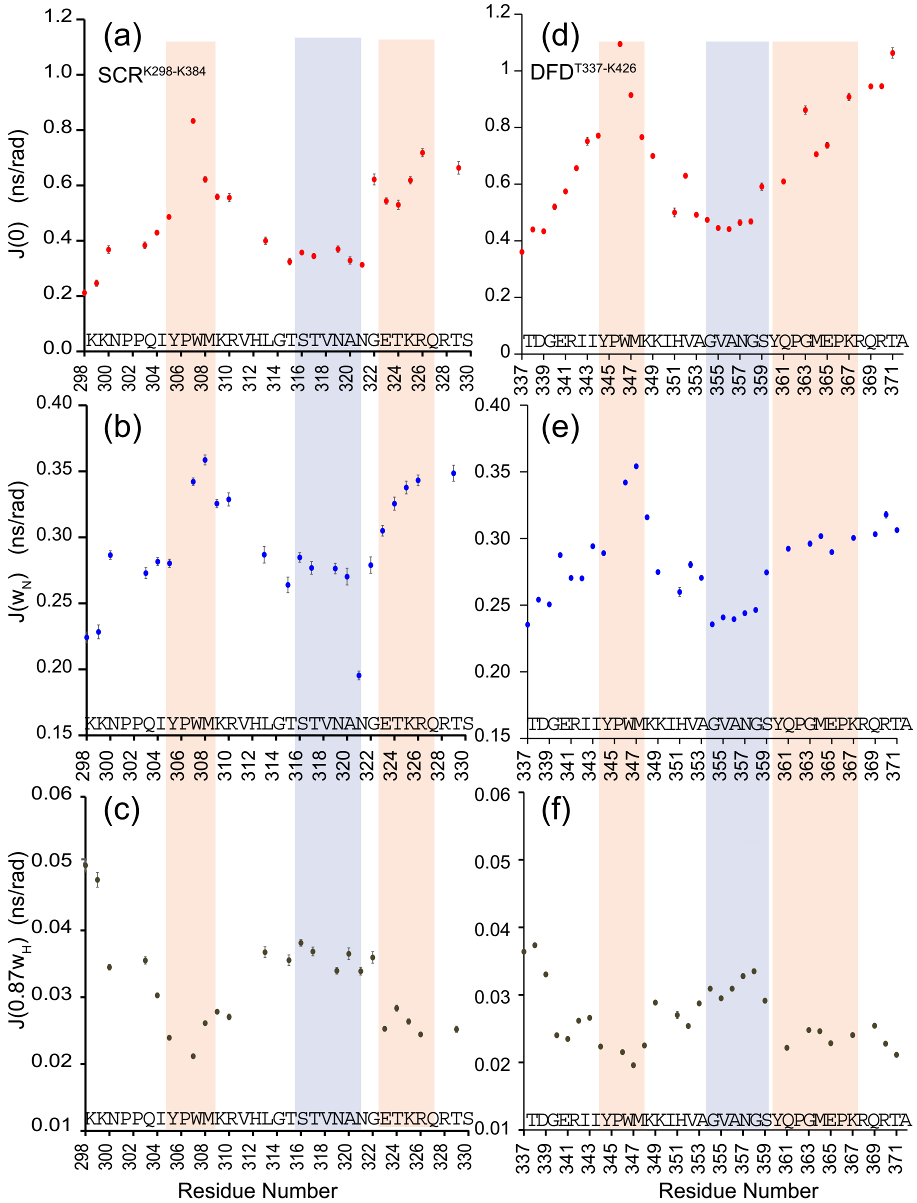

Fig. 3. Reduced spectral density mapping reveal variation in flexibility in the disordered N-terminus. Reduced spectral densities $J(0)$, $J(\omega_N)$, and $J(0.87\omega_H)$ for SCR$^{K298–K384}$ (a–c) and DFD$^{T337–K426}$ (d–f) were determined using Eqs. (1a) to (1c). The rigid segments are highlighted in pink, while the flexible segment is highlighted in blue. The rigid segments show increased $J(0)$ and $J(\omega_N)$ and decreased $J(0.87\omega_H)$, while it is reverse for the flexible segments in these proteins.

that the N-terminal residues are more mobile due to lack of ordered structure as compared to the homeodomain residues in both proteins.

It is important to note that in both proteins the residues in the N-terminal region are not equally flexible. To probe the variation of flexibility in this region, we compared the three spectral density values of each residue to the average value in this N-terminal region of ~30 residues. Consecutive residues with significantly higher $J(0)$ and $J(\omega_N)$ and significantly lower $J(0.87\omega_H)$ constitute a relatively rigid region. On the other hand, consecutive residues with significantly lower $J(0)$ and $J(\omega_N)$ and significantly higher $J(0.87\omega_H)$ constitute a relatively flexible region. This analysis revealed three distinct segments with varying degrees of flexibility (Fig. 3) in both proteins. These segments in SCR$^{K298–K384}$ are as follows: ^{305}YPWMKR^{310} (rigid), ^{316}STVNAN^{321} (flexible), and ^{322}GETKR^{326} (rigid). Similar regions in DFD$^{T337–K426}$ are as follows: ^{343}IYPWMKK^{349} (rigid), ^{354}GVANGS^{359} (flexible), and ^{360}YQPGMEPK^{367} (rigid). Interestingly, these rigid and flexible segments are sequentially equivalent regions in these two proteins (Fig. 4a).

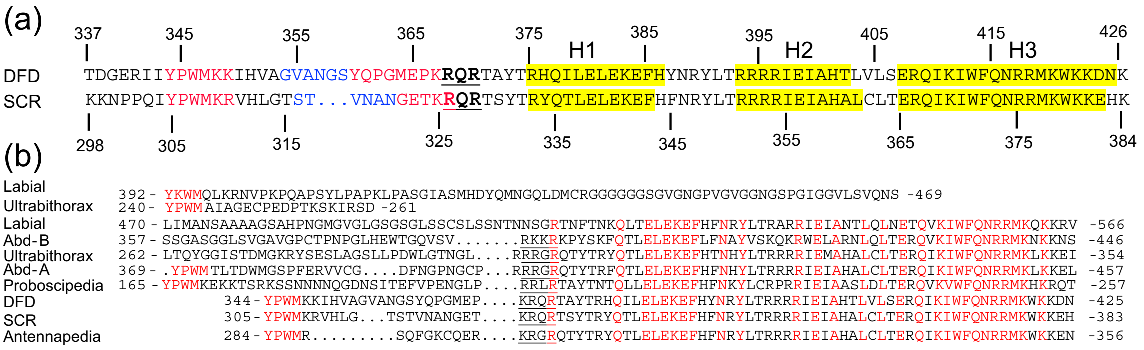
(a)
337
345
355
365
375
385
395
405
415
426
H1
H2
H3
DFD TDGERIIYPWMKKIHVAGVANGSYQPGMEPKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDNK
SCR KKNPPQIYPWMKRVHLGTST...VNANGETKRQRTSYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEHK
298
305
315
325
335
345
355
365
375
384
(b)
Labial 392 - YKWMQLKRNVPKPQAPSYLPAPKLPASGIASMHDYQMNGQLDMCRGGGGGGSGVGNGPVGVGGNGSPGIGGVLSVQNS -469
Ultrabithorax 240- YPWMAIAGECPEDPTKSKIRSD -261
Labial 470 - LIMANSAAAAGSAHPNGMGVGLGSGSGLSSCSLSSNTNNSGRTNFTNKQLTELEKEFHFNRYLTRARRIEIANTLQLNETQVKIWFQNRRMKQKKRV -566
Abd-B 357 - SSGASGGLSVGAVGPCTPNPGLHEWTGQVSV.......RKKRKPYSKFQTLELEKEFLFNAYVSKQKRWELARNLQLTERQVKIWFQNRRMKNKKNS -446
Ultrabithorax 262 - LTQYGGISTDMGKRYSESLAGSLLPDWLGTNGL....RRRGRQTYTRYQTLELEKEFHTNHYLTRRRRIEMAHALCLTERQIKIWFQNRRMKLKKEI -354
Abd-A 369 - .YPWMTLTDWMGSPFERVVCG....DFNGPNGCP...RRRGRQTYTRFQTLELEKEFHFNHYLTRRRRIEIAHALCLTERQIKIWFQNRRMKLKKEL -457
Proboscipedia 165 - YPWMKEKKTSRKSSNNNNQGDNSITEFVPENGLP....RRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQT -257
DFD 344 - YPWMKKIHVAGVANGSYQPGMEP....KRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIKIWFQNRRMKWKKDN -425
SCR 305 - YPWMKRVHLG...TSTVNANGET....KRQRTSYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEH -383
Antennapedia 284 - YPWMR.........SQFGKCQER....KRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN -356

Fig. 4. Sequence alignment of *Drosophila* HOX transcription factors. (a) Sequences of the constructs $\mathrm{SCR}^{\mathrm{K298–K384}}$ and $\mathrm{DFD}^{\mathrm{T337–K426}}$ are aligned. The three helices in the homeodomain as determined from backbone chemical shifts are highlighted in yellow. The folded and the disordered regions have 85% and 38% sequence identity, respectively. In the disordered region, the rigid and flexible segment residues are colored pink and blue, respectively. The residues RQR (bold and underlined) have been shown to interaction with the DNA minor groove and are postulated to determine DNA sequence specificity. (b) Sequences of eight HOX transcription factors in *Drosophila* are aligned. The homeodomain is highly conserved especially the DNA recognition helix H3. Except Abd-B, other seven HOX factors have the "YPWM" motif albeit at varying distances from the homeodomain. The "RQR" motif is underlined to show its relative conservation. The intervening residues between the "YPWM" motif and the homeodomain help determine DNA binding specificity of the HOX–EXD complex resulting in paralog-specific functions.

Residue-wise flexibility in the disordered region is readily obtained from $R_1R_2/(1 - \mathrm{NOE})$

Although reduced spectral density analysis highlighted varied degree of flexibility in the disordered regions of $\mathrm{SCR}^{\mathrm{K298–K384}}$ and $\mathrm{DFD}^{\mathrm{T337–K426}}$, it would be beneficial for studies on IDRs to identify rigid and flexible segments directly from the raw data, that is, R_1, R_2, and NOE values. Our data reveals that for residues in the rigid regions, spectral density function is dominated by low-frequency components $J(0)$ and $J(\omega_N)$, while for residues in flexible regions, the higher-frequency component $J(0.87\omega_H)$ also becomes significant. Hence, for each residue in the IDR the product $J(0)$ * $J(\omega_N)/J(0.87\omega_H)$ gives a measure of relative rigidity where rigid and flexible regions have higher and lower than average values, respectively (Figs. 5a and c and S2). In these plots, the aforementioned rigid and flexible segments can be clearly identified. It has been noted by several groups that spectral density functions $J(0)$, $J(\omega_N)$, and $J(0.87\omega_H)$ are dominated by R_2, R_1, and (1 − NOE) data, respectively [17–20]. This observation is validated by the corresponding correlation plots between these parameters for both $\mathrm{SCR}^{\mathrm{K298–K384}}$ and $\mathrm{DFD}^{\mathrm{T337–K426}}$ (Fig. 5). Thus, replacing the spectral density function at these three frequencies by the equivalent relaxation data set, residue-wise rigidity may also be obtained from $R_1R_2/(1 - \mathrm{NOE})$ (Fig. 5b and d). The correlation between $J(0)$ * $J(\omega_N)/J(0.87\omega_H)$, and $R_1R_2/(1 - \mathrm{NOE})$ is 0.99 for both proteins (Fig. 5h and l). Thus, the product $R_1R_2/(1 - \mathrm{NOE})$ gives a measure of rigidity for each residue in an IDR. A segment of rigid residues can be identified as consecutive residues with $R_1R_2/(1 - \mathrm{NOE})$ values more than average of the disordered region. This provides a simple analysis based on well-established NMR relaxation methods to identify rigid segments in IDRs of proteins.

SCR and DFD interact with partner protein EXD through identified rigid segment

In order to test the functional significance of the identified rigid segments, we studied the interaction of the HOX proteins SCR and DFD with their co-transcription factor EXD, which also has a DNA-binding homeodomain [21]. Several promoters recognized by HOX transcription factors have composite binding sites for both HOX and EXD [22]. It has been shown that when SCR and EXD homeodomains are bound to DNA, the disordered region of SCR also interacts with EXD through the "YPWM" motif [23], which has been identified as a rigid segment in free HOX protein in our study. The DNA most likely acts as a scaffold and facilitates the interaction of HOX and EXD. We wanted to test whether the HOX–EXD interaction also occurs in the absence of DNA. EXD gene corresponding to residues A238–I300 was cloned into pET28 (a+) vector, expressed, and purified. A ^{15}N–^{1}H HSQC spectrum was collected for ^{15}N-labeled sample and showed well-dispersed peaks indicating a properly folded homeodomain (Fig. S3). Unlabeled homeodomain of $\mathrm{EXD}^{\mathrm{A238–I300}}$ (residues A238 to I300) was titrated into ^{15}N-labeled $\mathrm{SCR}^{\mathrm{K298–K384}}$ or $\mathrm{DFD}^{\mathrm{T337–K426}}$, and each titration point was monitored by ^{15}N–^{1}H HSQC experiment (Figs. 6 and S4). Consecutive residues Y305 to R310 (P306 lacks backbone NH) in the rigid segment of $\mathrm{SCR}^{\mathrm{K298–K384}}$

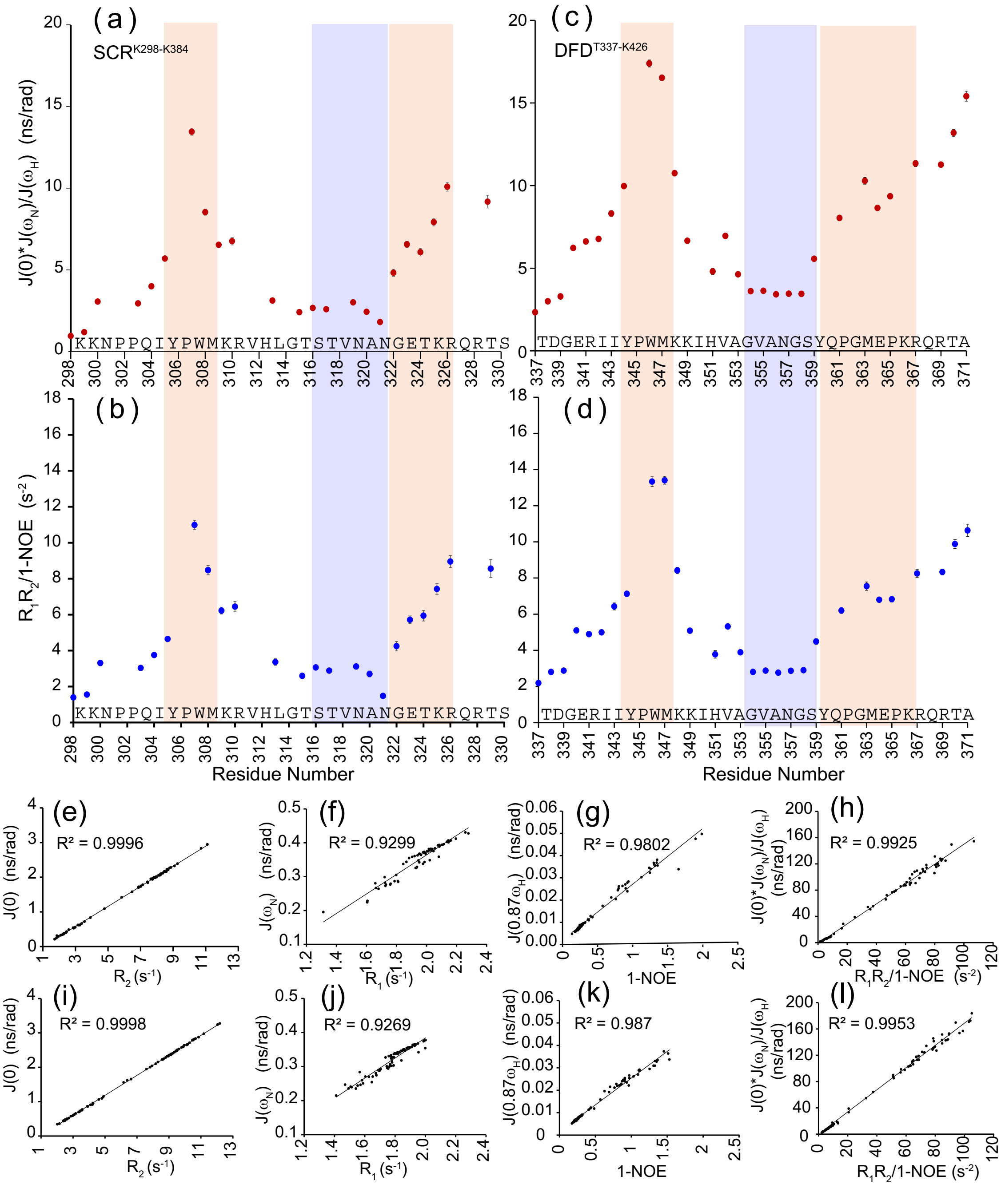

Fig. 5. Identifying the rigid and flexible segments in IDRs from the ^{15}N relaxation data. The rigid and flexible segments in the IDRs are readily identified from the residue-wise plot of $J(0) * J(\omega_N)/J(0.87\omega_H)$ for the disordered region of (a) SCR$^{K298–K384}$ and (c) DFD$^{T337–K426}$. The same segments can also be identified directly from the plot of $R_1*R_2/(1 - \mathrm{NOE})$ for (b) SCR$^{K298–K384}$ and (d) DFD$^{T337–K426}$. The rigid and flexible segments are highlighted in pink and blue, respectively. The correlation plots between the relaxation data and spectral density functions for all residues are shown for (e–h) SCR$^{K298–K384}$ and (i–l) DFD$^{T337–K426}$.

showed significant chemical shift changes (Fig. 6a). Similarly, the stretch of residues R341 to K349 (P345 lacks backbone NH) in the rigid segment of DFD$^{T337–K426}$ showed significant chemical shift changes (Fig. 6b). The remaining residues of both SCR$^{K298–K384}$ and DFD$^{T337–K426}$ have no significant chemical shift perturbation, indicating a very specific interaction between the rigid segment of the HOX proteins and EXD. Also no other residues in the flexible region and the homeodomain showed significant

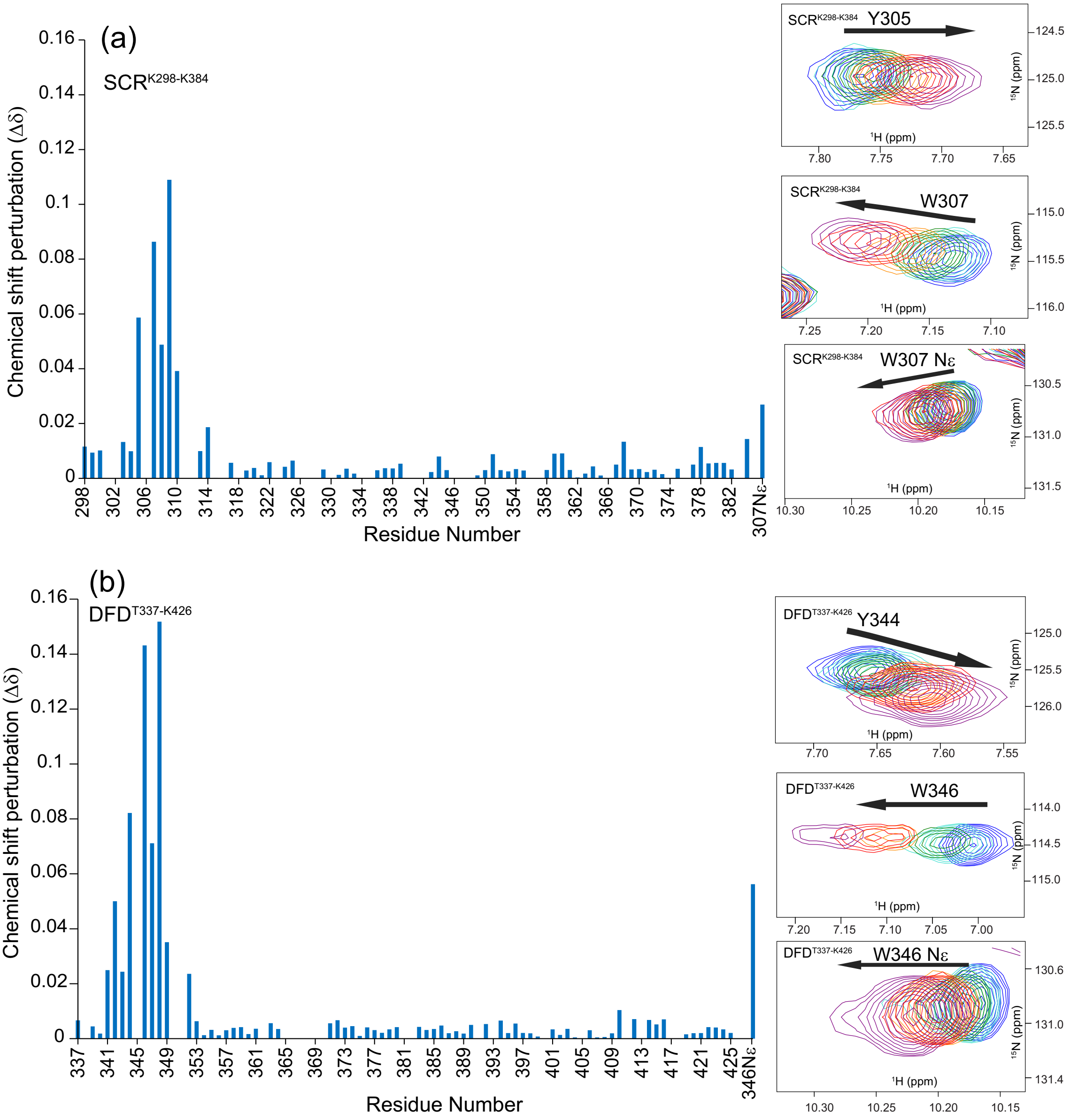

Fig. 6. Titration of ^{15}N-labeled HOX with EXD. The chemical shift perturbation of free and bound SCR$^{K298–K384}$ (a) and DFD$^{T337–K426}$ (b) is shown for each residue. In the inset, the change in peak position is shown for the Tyr and Trp residues in the "YPWM" motif for both proteins. HOX-to-EXD molar ratio in each titration is as follows: blue, 1:0; turquoise, 1:0.5; green, 1:1; orange, 1:2; red, 1:3; purple, 1: 4.5 for SCR and 1:5 for DFD.

chemical shift changes. Thus, the HOX–EXD interaction is specifically mediated by the "YPWM" rigid segment even in the absence of DNA.

In the crystal structures of SCR and EXD bound to DNA (PDB: 2R5Y and 2R5Z), the residues Y305 to R310 of SCR interact with EXD (Fig. 7). These are the exact same residues of SCR$^{K298–K384}$ that show significant chemical shift changes in our titration experiment. Thus, SCR and EXD interact in a very similar manner through the six ^{305}YPWMKR^{310} residues both in the presence or absence of DNA. DFD$^{T337–K426}$, similarly, showed significant chemical shift changes for residues R341 to K349 and thus interacts with EXD through nine residues in the disordered region. To check the effect of this relatively larger binding interface of DFD (nine residues) with respect to SCR (six residues), we determined the HOX–EXD dissociation constants from the chemical shift changes. The dissociation constants (K_D) were found to be 2.6 ± 0.4 and 1.3 ± 0.4 mM for the binding of EXD$^{A238–I300}$ to SCR$^{K298–K384}$ and DFD$^{T337–K426}$, respectively. Thus, DFD$^{T337–K426}$ binds to EXD$^{A238–I300}$ relatively more tightly than SCR$^{K298–K384}$ in the absence of DNA.

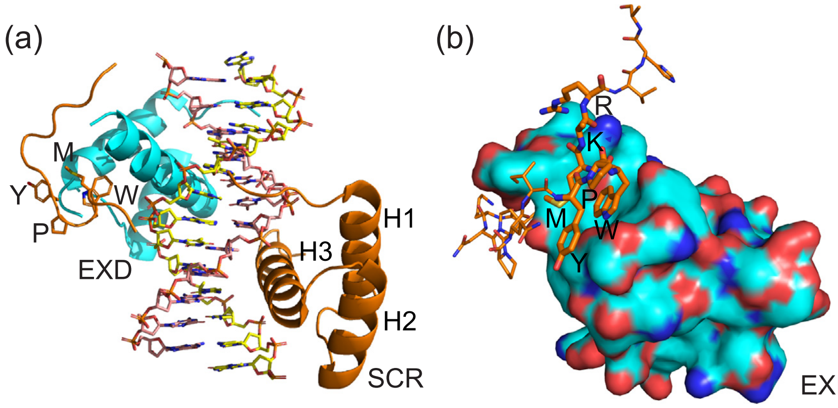

Fig. 7. SCR-EXD bound to DNA. (a) Crystal structure (PDB: 2R5Y) of SCR (orange) and EXD (cyan) bound to DNA. The conserved DNA recognition helix H3 inserts into the DNA major groove, while the N-terminal disordered region loops over the DNA and binds EXD. The flexible region also contacts the DNA. (b) The ^{305}YPWMKR^{310} rigid segment (stick representation) in the N-terminal disordered region is shown bound to EXD (surface representation). The W307 residue binds a hydrophobic pocket in EXD and results in proper orientation of the transcription factors, SCR and EXD, for cooperative DNA binding. Here, carbon atoms are colored orange (SCR), cyan (EXD), pink (DNA), and yellow (DNA); oxygen atoms are colored red; nitrogen atoms are colored blue; phosphorous atoms in DNA are colored orange; and sulfur atom in EXD is colored yellow.

MD simulation reveals restricted conformation sampling by the rigid segments

Our dynamic studies by solution NMR spectroscopy revealed that the disordered N-terminal regions in both SCR$^{K298–K384}$ and DFD$^{T337–K426}$ have stretches of rigid and flexible residues. This experimental observation is reminiscent of the “rigid-segment model” of folded proteins proposed by Fitzkee and Rose [24]. This model postulates that “…protein structures are partitioned alternately into rigid segments linked by individual flexible residues” [24]. In case of IDRs, we observe by NMR dynamic experiments that rigid segments are separated by stretches of flexible residues. In order to further test whether a disordered region of a protein indeed follows the “rigid-segment model,” MD simulations were performed on a 91-residue model of SCR$^{K298–K384}$. Two starting structures with different conformations of the flexible region were used to perform two 100-ns MD simulations.

For folded proteins, flexibility is measured from MD simulations by aligning each structure in the trajectory to a reference structure and computing root mean square fluctuation (RMSF). For IDRs, such an approach does not give a meaningful result as the RMSF keeps on increasing for residues farther from the aligned structured region (Fig. 8a). To get a meaningful measure of residue-wise flexibility, an algorithm was devised, which is loosely based on the method used by Fitzkee and Rose [24] to identify flexible residues that connect the rigid segments. Since torsion angle dynamics also occurs in the ps–ns timescale [25], we reasoned that two 100-ns simulations should capture the dynamics of the disordered region. For each residue, the Φ and Ψ torsion angles were determined from each structure in the two 100-ns MD trajectories. The corresponding Ramachandran plot for each residue consisting of 200,000 torsion angles was divided into 2° by 2° boxes. All the boxes that were occupied at least twice were counted and normalized by the total number of boxes, that is, 32,400 resulting in a flexibility index for each residue (Fig. 8b).

In the homeodomain, the helices and the loops exhibit low and high flexibility indices, respectively (Fig. 8b). This is expected as helices are restrained by backbone hydrogen bonds and have very restricted backbone torsion angles, while loops are less restrained. The N-terminal disordered region also shows high flexibility index. Inspecting the rigid and flexible segments, which were identified by NMR studies, shows that overall the rigid segments have lower flexibility index compared to the flexible segments. This is indeed consistent with our postulate that the rigid segments in these IDRs sample relatively less conformations. This is clearly seen in the Ramachandran plots for the residues Y305 and R326 in the rigid, and R310 and N321 in the flexible segments (Fig. 8c). Thus, the flexibility index described here is a convenient way of measuring residue-wise dynamics from MD simulations and is especially useful for IDRs.

Discussion

Traditional structural biology relies on the structure–function paradigm that elucidates the function of a protein from its well-defined three-dimensional

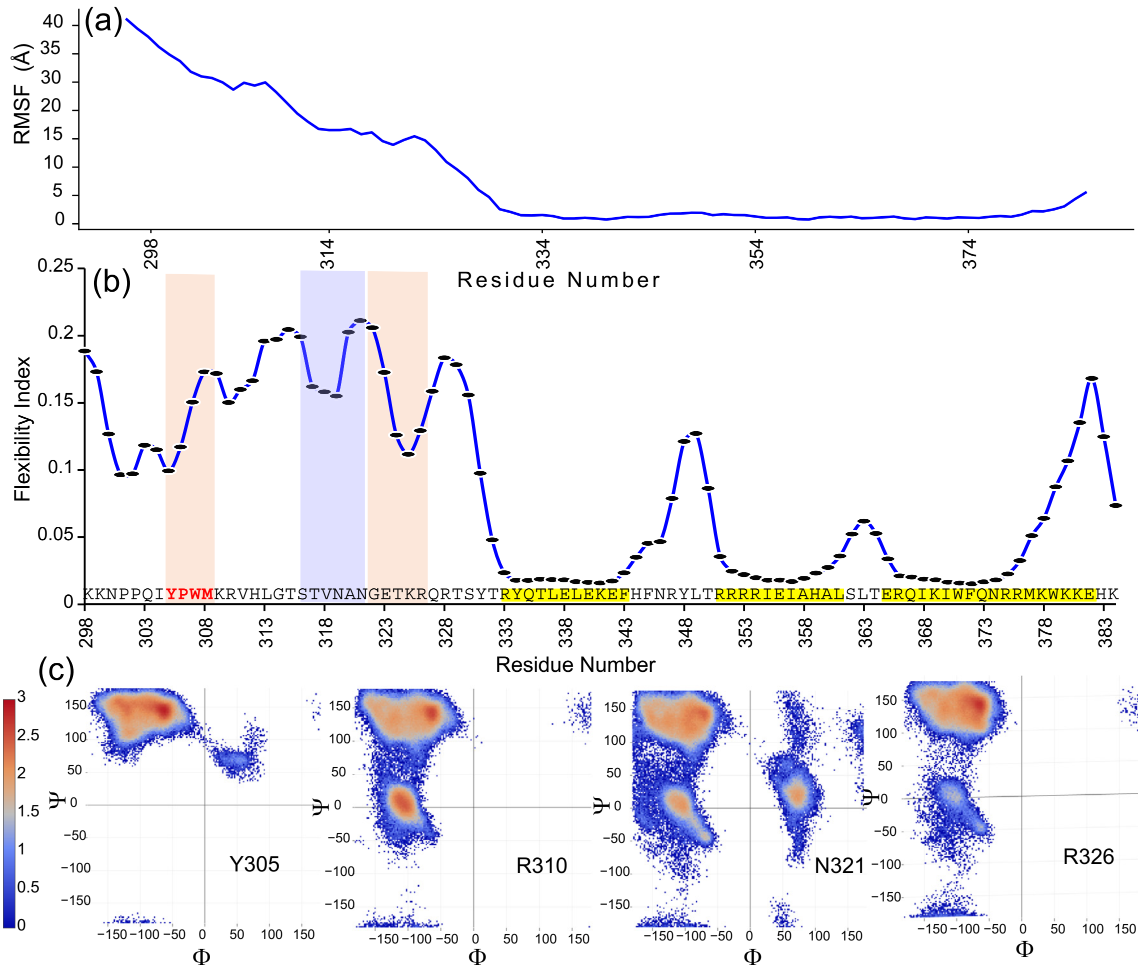

Fig. 8. MD simulation reveals less conformational sampling for rigid segments. (a) RMSF for each residue was calculated from 10,000 structures extracted at 10-ps interval from one 100-ns MD simulation. RMSF in the disordered region keeps on increasing as residues get farther from the folded domain. (b) Residue-wise flexibility index was calculated from 200,000 structures extracted from two 100-ns MD simulations. It shows rigid helices and flexible loops in the homeodomain. The N-terminal disordered region shows variable flexibility. The rigid and flexible segments identified by NMR studies are highlighted in pink and blue, respectively. (c) Ramachandran plot for four residues from 200,000 structures. Residues Y305 and N321 are from rigid and flexible segments, respectively. The flexible residue N321 occupies more Φ–Ψ space and hence samples more conformations compared to the rigid residue Y305. Two arginines, R310 and R326, are also shown for comparison. Population in each box is shown in log scale, 10^0 is blue and 10^3 is red.

structure. This paradigm works well for folded proteins and has led to many important discoveries [26], but it is not adequate to study IDRs in proteins. IDRs are devoid of any three-dimensional fold and consist of dynamically exchanging conformations [27]. Hence, it is not trivial to describe the structural behavior of such dynamic proteins and infer biological roles from them [28]. Several approaches using solution NMR spectroscopy have been described to structurally characterize IDRs. NMR chemical shifts [29–31], hydrogen exchange rates [32], ^{15}N-relaxation measurements [33,34], J-coupling [35], and residual dipolar couplings [36] have been used to determine the existence of local transient secondary structures, while paramagnetic relaxation enhancement has been used to study transient long-range contacts [37,38]. Although these experiments provide important insights into IDRs, the total restraints obtained are much less than the overall conformations sampled by IDRs [39]. Thus, the structural characterization of IDRs is an underdetermined problem. Hence, instead of attempting an ensemble description, we have focused on one structural characteristic of IDRs, that is, residue-wise rigidity, which we postulate is highly significant for their biological function. To investigate this, we chose two paralogous HOX transcription factors, SCR and DFD, from *Drosophila* that have well-characterized biological roles of their IDRs [15]. Although the DNA-binding homeodomains are highly conserved, the disordered regions have little sequence identity (Fig. 4a). Hence, the method to identify rigid segments, described here, should be applicable to other IDRs.

Identified rigid regions are functionally important

Analysis of ^{15}N-relaxation measurements revealed two distinct stretches of relatively rigid residues in the disordered region of both SCR$^{K298–K384}$ and DFD$^{T337–K426}$. One stretch of residues "YPWMK(R/K)" contains a conserved motif "YPWM" in HOX proteins (Fig. 4b). The "YPWM" motif is recognized by HOX cofactors named TALE (three amino acid loop extension), a family of transcription factors that cooperatively bind DNA along with HOX proteins (Fig. 8). In *D. melanogaster*, EXD is the TALE cofactor. Using NMR titration experiments, we show that the EXD$^{A238–I300}$ homeodomain specifically binds only to the "YPWM" motif in the disordered regions of both SCR$^{K298–K384}$ and DFD$^{T337–K426}$ even in the absence of DNA. Thus, this rigid segment is indeed involved in recruiting the co-transcription factor EXD resulting in cooperative DNA binding by HOX and EXD.

This cooperative HOX–EXD binding is very significant for HOX function as it modulates the DNA-binding specificity of the HOX factors. The isolated DNA binding homeodomain of most HOX proteins recognizes very similar AT-rich sequences such as $^{5'}$-TAAT[T/G]A$^{-3'}$ [40,41]. Since most HOX paralogs in an organism recognize very similar sequences, the DNA-binding by their homeodomains is not sufficient to explain their distinct and very specific biological functions. Moreover, the *Drosophila* genome has more than 150,000 copies of these sequences, which is much more than the number of annotated protein-coding genes in this organism [42], indicating that HOX factors should recognize a longer stretch of DNA sequence. An expanded DNA sequence, such as $^{5'}$-AGATTTATGG$^{-3'}$, is recognized by the HOX–EXD combination, where one half-site is bound by EXD and the other half-site by HOX proteins (Fig. 8a). In combination with EXD, the HOX proteins can recognize DNA sequences different from the consensus DNA sequence [43]. It has been proposed that the cooperative DNA binding by EXD through the "YPWM" motif elicits a latent DNA-binding specificity in HOX transcription factors. Thus, the "YPWM" motif plays an important role in specific promoter site recognition by the HOX–EXD combination in *Drosophila* and HOX–TALE combinations in other vertebrates.

The other identified rigid region in the disordered N-terminus of SCR and DFD is not conserved. Interestingly, the residues $^{325-}$KRQRT^{-329} (SCR numbering) that immediately follow this rigid region are conserved in many HOX factors (Fig. 4b). The arginine residues in this segment have been shown to contact the negatively charged backbone of the DNA and thus also help in determining the HOX–EXD specificity [23]. Overall, the intervening residues between the "YPWM" motif and the DNA-binding homeodomain in both SCR and DFD have been shown to play a critical role in determining paralog-specific DNA-binding specificity. A chimeric DFD with the SCR intervening sequences has been shown to exhibit SCR-like functions *in vivo* [15]. Thus, both the rigid segments in the disordered regions of SCR and DFD, identified in our study, play very important roles in determining the specific biological functions of these proteins.

Importance of rigidity in IDRs and its origin

Biological functions of IDRs rely on their ability to act as interaction modules to multitude of partners [44–46]. Binding to a partner molecule requires the selection of a few binding-competent conformations from the vast number of conformations explored by IDRs. This large loss in conformational entropy prohibits binding interactions. Based on a survey of structures of IDRs bound to partner proteins, it has been proposed that short regions in IDRs may exhibit limited conformations, which correspond to the bound form [47]. Our ^{15}N-relaxation measurements on free proteins indeed show the existence of dynamically rigid stretches of amino acids separated by highly flexible residues. Rigid segments in IDRs ensure that these stretches of amino acids sample significantly less number of conformations thus, minimizing the loss in conformational entropy and increasing their binding affinity to partner molecules.

We performed two 100-ns MD simulations, resulting in 200,000 structures, to determine the conformational space sampled by the residues in SCR$^{K298–K384}$. For each residue, the Φ–Ψ torsion angles from the 200,000 structures were mapped in the Ramachandran plot, which gives a measure of the conformations sampled by the residue. Our MD simulations on SCR$^{K298–K384}$ indeed show that the amino acids in the rigid segments sample less Φ–Ψ space as compared to those in the flexible segments (Fig. 8c). Using NMR-monitored titration experiments, we further show that the rigid "YPWM" segment in both SCR$^{K298–K384}$ and DFD$^{T337–K426}$ is specifically bound by the EXD transcription cofactor. Moreover, using chemical shift perturbations, we show that the exact same residues, that is, ^{305}YPWMKR^{310} of SCR$^{K298–K384}$ interact with EXD both in the absence (Fig. 6) or presence of DNA (Fig. 7).

In order to compare the structures of the ^{305}YPWMKR^{310} residues in the bound and free SCR$^{K298–K384}$, we mapped their torsional angles in the crystal structure and our MD simulations (Fig. S5). The Φ–Ψ torsion angles of each residue in the ^{305}YPWMKR^{310} segment from the MD simulations occupied certain area in the Ramachandran map, signifying the conformations sampled by these residues in the free state. The corresponding Φ–Ψ torsion angles of each residue in the ^{305}YPWMKR^{310} segment in the crystal structure were found to be within this area in the Ramachandran map. This indicates that although the rigid segment samples relatively less number of conformations, it does sample the binding competent conformation (Fig. S5). Inspection of the torsion angles of the bound ^{305}YPWMKR^{310} segment shows that residues ^{306}PWM^{308} have helical torsion angles, while the other three residues have torsion angles corresponding to an extended conformation.

It is important to note that in the first rigid segment containing the "YPWM motif," residues R341 to K349 in DFD$^{T337–K426}$ have higher R_1R_2/(1 – NOE) values compared to the residues I304 to R310 in SCR$^{K298–K384}$ (Fig. 5), indicating a more relative rigidity of this segment in DFD$^{T337–K426}$ compared to SCR$^{K298–K384}$. Thus, upon binding, this rigid segment in DFD$^{T337–K426}$ should experience relatively less decrease in conformational entropy compared to SCR$^{K298–K384}$, resulting in tighter binding of the "YPWM" motif to EXD. This is indeed observed in our titration experiments. Thus, the interspersed rigid segments in IDRs enable them to bind various partner molecules and perform their biological functions.

Since long-range contacts in IDRs are almost negligible, the rigid segments are stabilized by local interactions within a few residues. Inspection of the sequences of the rigid segments in SCR$^{K298–K384}$ and DFD$^{T337–K426}$ reveals two types of interactions: hydrophobic interactions and electrostatic interactions. Conformations of the "YPWM" motif are most likely restricted by hydrophobic interactions between the bulky hydrophobic side chains. The other rigid segment has more charged residues. The fraction of charged residues and their distribution in IDRs can restrain their conformation sampling [48]. On the other hand, the identified flexible region in SCR$^{K298–K384}$ and DFD$^{T337–K426}$ consists of mostly neutral and polar residues such as Ser, Thr, Asn, Gly, and Ala, which are incapable of both hydrophobic and electrostatic interactions. Thus, in the absence of any restraints, these residues can sample conformations extensively resulting in a flexible region. These regions act as flexible linkers that join the functionally important rigid segments. As they are sterically more malleable, these linkers can also increase the interaction efficiency of IDRs with various partner molecules, especially in large multi-protein complexes. Moreover, due to their high conformational entropy, the linker regions themselves interact weakly with other molecules, thus, minimizing non-specific interactions of IDRs.

The importance of small motifs in IDRs in facilitating protein–protein interactions is getting significant attention in the past decade. The eukaryotic linear motif resource lists over 3000 short linear motifs (SLiMs), which are within 3 to 15 amino acids in length [49]. It is estimated that the human proteome may contain ~ 10^6 such motifs. The eukaryotic linear motif database is curated manually and uses the defined motifs to detect SLiMs in query sequences [50]. Another approach utilizes knowledge based on PDB structures of short peptides bound to partner proteins to identify molecular recognition features (MoRFs) within IDRs [51,52]. In this study, we describe an experimental method to identify short rigid sequences in IDRs in free proteins that play important role in protein–protein or protein–DNA interactions. These experimentally derived rigid segments are equivalent to the SLiMs or MoRFs identified by bioinformatics approach. We envision that identification of similar rigid segments in other IDRs can further enhance the determination of SLiMs and MoRFs.

Conclusions

Although several models have been proposed, a complete structural description of IDRs, similar to the folded proteins, remains elusive due to the lack of sufficient experimental restraints. In this study, we show that an important characteristic of IDRs, that is, residue-wise rigidity, can be experimentally determined from well-established ^{15}N-relaxation measurements. We show that the IDRs are composed of short segments of rigid residues with limited backbone motions that are linked by stretches of flexible amino acids. In the context of HOX transcription factors, we further demonstrate that one of the identified rigid segments, which is conserved in the family, specifically interacts with a co-transcription factor. Thus, the rigid segments in the IDRs can fine tune protein function through specific interactions with other molecules.

Experimental Methods

Protein expression and purification

SCR, DFD, and EXD genes were cloned into pET28a(+) expression vector that results in N-terminal His$_6$-tagged protein. The SCR$^{K298–K384}$, DFD$^{T337–K426}$, and EXD$^{A238–I300}$ constructs were generated by PCR. The sole cysteine in SCR$^{K298–K384}$ (Cys362) was mutated to serine by site-directed mutagenesis. These proteins were expressed in *Escherichia coli* BL21(λDE3) cells at 37 °C. For ^{15}N/^{13}C labeling, M9 minimal media was used. For unlabeled proteins, LB media was used. Overnight culture (9–10 h) in 10 ml LB at 37 °C was used to inoculate 250 ml of LB and grown at 37 °C till the cell density reaches OD$_{600}$ of 3.0 [53]. The cells were harvested and resuspended into 250 ml of M9 minimal media supplemented with 0.25 g ^{15}NH$_4$Cl and 1 g $^{13}C_6$-glucose as the sole nitrogen and carbon source, respectively. After induction with 1 mM IPTG for 4 h at 37 °C, cells were harvested and lysed by sonication in the presence of 4 M guanidine HCl in the lysis buffer (100 mM Tris–Cl, 200 mM NaCl, 10 mM imidazole) at pH 8.2. The cell lysate was centrifuged at 16,639*g* for 40 min at room temperature. The target proteins were purified using Ni-NTA column that binds the His_6 tag. The affinity tag was removed by the Thrombin CleanCleave Kit from Sigma-Aldrich. The proteins were exchanged into the final buffer [20 mM sodium phosphate, 50 mM NaCl (pH 5.5)]. For titration experiments, the final buffer had 150 mM NaCl. Protein concentrations were determined by UV absorption using predicted molar absorptivity (ε280) [54]. For heteronuclear NMR experiments, ^{15}N-labeled and ^{15}N–^{13}C double-labeled protein samples were used. The proteins were 0.2–0.6 mM with 7% D_2O for spin lock. For long-term stability of the proteins, 0.04% NaN_3 and 0.4 mM PMSF were also added to the final sample.

Backbone assignment of SCR$^{K298–K384}$ and the DFD$^{T337–K426}$

NMR experiments were performed on Bruker 600 MHz spectrometer at 25 °C. For sequential backbone assignment, ^{15}N,^{13}C double-labeled protein samples were used and triple-resonance experiments such as HNCACB, CBCA(CO)NH, HNCO, HN(CA)CO, and (H)CC(CO)NH-TOCSY were collected [55]. In addition, HNN spectrum was also collected to assign DFD$^{T337–K426}$ [56]. These spectra were processed and analyzed using NMRPipe [57] and Sparky [58], respectively. Secondary structure was predicted from backbone chemical shifts (^{1}HN, ^{15}N, ^{13}Cα, ^{13}Cβ, and ^{13}CO) using the program MICS [16].

Backbone amide ^{15}N relaxation measurements

Amide ^{15}N R_1, R_2, and steady-state heteronuclear NOE experiments were collected at 25 °C for both proteins. Spectra for R_1 (50, 100, 150, 200, 400, 600, 900, and 1200 ms) and R_2 (50, 100, 150, 200, 300, 400, 500, 600, and 700 ms) time series were collected in random order to minimize any systematic error. A delay of 3 s between scans was used. Relaxation rate constants R_1 and R_2 were determined by fitting the peak intensities to single exponential decay ($I_t = I_0$ * exp.($-t$ * R_i)), where 'I_t' is the peak intensity, 't' is the relaxation delay, I_0 is the initial intensity, and R_i is either R_1 or R_2 [59,60]. Uncertainties in the rate constants were estimated by Monte Carlo simulation. The heteronuclear {^{1}H}–^{15}N NOE values were determined from the ratio of the peak heights acquired with and without 3 s of ^{1}H saturation and a total recycle delay of 5 s. Uncertainties in hetero-NOE were estimated by propagation of error using the spectral noise.

Reduced spectral density analysis of amide ^{15}N relaxation data

Reduced spectral density mapping was done for the two constructs, SCR$^{K298–K384}$ and DFD$^{T337–K426}$, by fitting the relaxation data to the Eqs. (1a)–(1c) [61]. Uncertainties in $J(0)$, $J(\omega_N)$, and $J(0.87\omega_H)$ were determined by Monte Carlo simulation.

$$J(0.87\omega_H) = \frac{4}{5}\frac{1}{d^2}\frac{\gamma_N}{\gamma_H} R_1(\mathrm{NOE}-1) \quad \mathrm{srad}^{-1} \tag{1a}$$

$$J(\omega_N) = \frac{1}{3d^2 + 4c^2}\left[4R_1 - 7d^2 J(0.87\omega_H)\right] \quad \mathrm{srad}^{-1} \tag{1b}$$

$$J(0) = \frac{R_2 - \left(\frac{3}{8}d^2 + \frac{1}{2}c^2\right)J(\omega_N) - \frac{13}{8}d^2 J(0.87\omega_H)}{\frac{d^2}{2} + \frac{2}{3}c^2} \tag{1c}$$

where

$$d = \frac{\mu_0 h \gamma_H \gamma_N}{8\pi^2}\left\langle \frac{1}{r_{NH}{}^3} \right\rangle$$

$$c = \frac{\omega_N}{\sqrt{3}}(\sigma_\parallel - \sigma_\perp)$$

$$r_{NH} = 1.02\ \text{Å}$$

$$\sigma_\parallel - \sigma_\perp = -160 \text{ ppm for backbone NH;}$$
$$-89.6 \text{ ppm for Trp residue NH}$$

NMR titration experiments

Unlabeled EXD$^{A238–I300}$ (0.65 mM stock solution) was titrated into ^{15}N-labeled SCR$^{K298–K384}$ or DFD$^{T337–K426}$, and ^{15}N-HSQC spectra were collected by cryoprobe-equipped 600 MHz Avance III Bruker spectrometer. The initial concentrations of SCR$^{K298–K384}$ and DFD$^{T337–K426}$ were 0.25 and 0.28 mM, respectively. The molar ratios of EXD$^{A238–I300}$ to SCR$^{K298–K384}$ in the titration sets were 0, 0.5, 1, 1.5, 2, 3, 4, and 4.5, whereas for DFD$^{T337–K426}$, the titration sets were 0, 0.5, 1, 1.5, 2, 3, 4, and 5. Binding of HOX proteins to EXD$^{A238–I300}$ occurred in the fast exchange limit; thus, amide $^1H^N$ and ^{15}N assignments were obtained by tracking shifts relative to the initial free HOX protein. Combined amide chemical shifts were obtained as

$\Delta\delta_{obs} = \{(\Delta\delta_H^2 + (0.154 * \Delta\delta_N)^2\}^{½}$, where $\Delta\delta_H$ and $\Delta\delta_N$ are the observed shifts from the free state in the proton and nitrogen dimension, respectively. Dissociation constants (K_D) were determined using the following equation

$$\Delta\partial_{obs} = \frac{\Delta\partial_{max}\left\{([P]_t + [L]_t + K_D) - \left(([P]_t + [L]_t + K_D)^2 - 4[P]_t[L]_t\right)^{1/2}\right\}}{2[P]_t} \quad (2)$$

where $\Delta\delta_{max}$ is the maximum shift change upon saturation, $[P]_t$ is the total concentration of HOX protein and $[L]_t$ is the total concentration of EXD$^{A238–I300}$. Uncertainties in K_D were determined from the standard deviation of K_D obtained from the residues Y305, W307 backbone amides, and W307 side-chain amide peaks in SCR$^{K298–K384}$, as these peaks showed no spectral overlap. Corresponding residues in DFD$^{T337–K426}$ were used to determine the average and standard deviation of K_D.

MD simulation

A model of SCR$^{K298–K384}$ was built based on the SCR–DNA structure (PDB: 2R5Y) using the program I-TASSER [62]. Starting from this model, an ensemble of structures was generated by sampling backbone conformations of the disordered region using Backrub application in Rosetta [63]. Two structures with significantly different conformations of the disordered region were used as the starting structures for two independent MD trajectories of 100 ns each. A common protocol, as described below, was used for both trajectories.

All-atom CHARMM36m [64], a modified version of CHARMM36 [65] force field, which is specifically designed for modeling intrinsically disordered proteins, was implemented. The protein was solvated in a cubic box using TIP3P [66] explicit water model, with a padding distance of 25 Å. The system retained a total charge of +15 after solvation, which was neutralized with 15 chloride (Cl^-) ions. In addition, sodium chloride (NaCl) was added to adjust the salt concentration to 150 mM. The system was subjected to minimization for 10 ps followed by annealing to raise the temperature to room temperature (300 K). Furthermore, equilibration was performed for 1 ns using constant-temperature, constant-pressure (NpT) ensemble for relaxing the system. The temperature was controlled by using Langevin dynamics, and the pressure was kept constant by Nosé–Hoover Langevin piston [67]. Particle-mesh Ewald method was used to treat the long-range electrostatic interactions [68]. The equation of motion was integrated using a time step of 2 fs with the help of r-RESPA [69] multiple-time step scheme. The cutoff for non-bonded interaction was set to 12 Å. MD simulation was performed for 100 ns for both simulations in NpT ensemble using NAMD [69]. The trajectory data were saved at an interval of 1 ps for analysis.

RMSFs were calculated for 10,000 structures, where each structure was extracted from one 100-ns MD trajectory at 10-ps interval. The homeodomains of these structures were aligned, and RMSF between Cα atoms was calculated for each residue.

Determination of flexibility index from MD trajectories

From both 10-ns MD trajectories, one structure at every 1 ps was extracted resulting in a set of 200,000 structures. For each residue, the Φ and Ψ torsion angles were measured from each of these 200,000 structures. The Ramachandran plot (Φ *versus* Ψ) of each residue was divided into 32,400 grid boxes (2° × 2°), and 200,000 (Φ, Ψ) pairs of torsion angles of the residue were placed in their corresponding boxes in the plot. For each residue, the flexibility index was determined by counting the number of boxes occupied at least twice and dividing it by the total number of boxes, that is, 32,400. The grid boxes that were occupied at least twice were considered to eliminate the outliers. The final flexibility index was calculated by a running average of 3 residues that was assigned to the middle residue.

Accession numbers

The chemical shifts for DFD and SCR are submitted to BMRB under the accession numbers 27621 and 27622, respectively.

Acknowledgments

We acknowledge Aditya Jyoti Basak for helpful discussions. This work was supported by the Science and Engineering Research Board (SERB), India; IIT Kharagpur, India Early Career Research (ECR) award (ECR/2016/000847 dated 07/03/2017), and ISIRD grant from IIT-Kharagpur (IIT/SRIC/ISIRD/2015-2016 dated 14/12/2015) to S.D.

Author Contributions: S.M. and B.A. contributed equally to this work. S.M., B.A., and S.D. conceptualized the idea of the work. S.M. and B.A. performed the sample preparations and NMR experiments. S. M., B.A., and S.D. processed and analyzed the data. B.M. and A.G. performed MD simulation and analyzed the data. V.S.B. and S.D. designed the algorithm for the calculation of flexibility index. All authors contributed in writing the manuscript and have given approval to the final version of the manuscript.

Conflict of Interest: The authors declare that they have no conflicts of interest with the contents of this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.02.021.

Received 9 November 2018;
Received in revised form 31 January 2019;
Accepted 15 February 2019
Available online 22 February 2019

Keywords:
NMR spectroscopy;
HOX;
backbone dynamics;
molecular dynamics;
flexibility index

†S.M. and B.A. contributed equally to this work.

Abbreviations used:
HOX, SCR, Sex Combs Reduced; DFD, Deformed; HSQC, heteronuclear single quantum coherence; MD, molecular dynamics.

References

[1] N.S. Bogatyreva, A.V. Finkelstein, O.V. Galzitskaya, Trend of amino acid composition of proteins of different taxa, J. Bioinforma. Comput. Biol. (2006), https://doi.org/10.1142/S0219720006002016.

[2] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D^2 concept, Annu. Rev. Biophys. 37 (2008) 215–246, https://doi.org/10.1146/annurev.biophys.37.032807.125924.

[3] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C.H. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59, https://doi.org/10.1016/S1093-3263(00)00138-8.

[4] V.N. Uversky, Unusual biophysics of intrinsically disordered proteins, Biochim. Biophys. Acta, Proteins Proteomics 1834 (2013) 932–951, https://doi.org/10.1016/j.bbapap.2012.12.008.

[5] P. Tompa, The interplay between structure and function in intrinsically unstructured proteins, FEBS Lett. 579 (2005) 3346–3354, https://doi.org/10.1016/j.febslet.2005.03.072.

[6] Y. Minezaki, K. Homma, A.R. Kinjo, K. Nishikawa, Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation, J. Mol. Biol. 359 (2006) 1137–1149, https://doi.org/10.1016/j.jmb.2006.04.016.

[7] J. Liu, N.B. Perumal, C.J. Oldfield, E.W. Su, V.N. Uversky, A.K. Dunker, Intrinsic disorder in transcription factors, Biochemistry. (2006), https://doi.org/10.1021/bi0602718.

[8] S. De, A.C.K. Chan, H.J. Coyne, N. Bhachech, U. Hermsdorf, M. Okon, M.E.P. Murphy, B.J. Graves, L.P. McIntosh, Steric mechanism of auto-inhibitory regulation of specific and non-specific DNA binding by the ETS transcriptional repressor ETV6, J. Mol. Biol. 426 (2014) 1390–1406, https://doi.org/10.1016/j.jmb.2013.11.031.

[9] P.C. Hollenhorst, L.P. McIntosh, B.J. Graves, Genomic and biochemical insights into the specificity of ETS transcription factors, Annu. Rev. Biochem. 80 (2011) 437–471, https://doi.org/10.1146/annurev.biochem.79.081507.103945.

[10] A. Bah, J.D. Forman-Kay, Modulation of intrinsically disordered protein function by post-translational modifications, J. Biol. Chem. 291 (2016) 6696–6705, https://doi.org/10.1074/jbc.R115.695056.

[11] S. De, M. Okon, B.J. Graves, L.P. McIntosh, Autoinhibition of ETV6 DNA binding is established by the stability of its inhibitory helix, J. Mol. Biol. 428 (2016) 1515–1530, https://doi.org/10.1016/j.jmb.2016.02.020.

[12] M.J. Gorman, T.C. Kaufman, Genetic analysis of embryonic cis-acting regulatory elements of the *Drosophila* homeotic gene sex combs reduced, Genetics 140 (1995) 557–572, http://www.genetics.org/content/140/2/557.abstract.

[13] W. McGinnis, R. Krumlauf, Homeobox genes and axial patterning, Cell. 68 (1992) 283–302, https://doi.org/10.1016/0092-8674(92)90471-N.

[14] D.K. Papadopoulos, V. Vukojević, Y. Adachi, L. Terenius, R. Rigler, W.J. Gehring, D.K. Papadopoulos, V. Vukojevicb, Y. Adachic, L. Tereniusb, R. Rigler, W.J. Gehring, Linked references are available on JSTOR for this article : function and specificity of synthetic Hox transcription factors, In vivo 107 (2010) 4087–4092.

[15] R. Joshi, L. Sun, R. Mann, Dissecting the functional specificities of two Hox proteins, Genes Dev. 24 (2010) 1533–1545, https://doi.org/10.1101/gad.1936910.

[16] Y. Shen, A. Bax, Identification of helix capping and β-turn motifs from NMR chemical shifts, J. Biomol. NMR 52 (2012) 211–232, https://doi.org/10.1007/s10858-012-9602-0.

[17] M. Buck, H. Schwalbe, C.M. Dobson, Main-chain dynamics of a partially folded protein: 15N NMR relaxation measurements of hen egg white lysozyme denatured in trifluoroethanol, J. Mol. Biol. 257 (1996) 669–683, https://doi.org/10.1006/jmbi.1996.0193.

[18] R. Ishima, K. Nagayama, Protein backbone dynamics revealed by quasi spectral density function analysis of amide N-15 nuclei, Biochemistry. 34 (1995) 3162–3171, https://doi.org/10.1021/bi00010a005.

[19] N.A. Farrow, O. Zhang, J.D. Forman-Kay, L.E. Kay, Comparison of the backbone dynamics of a folded and an unfolded SH3 domain existing in equilibrium in aqueous buffer, Biochemistry. 34 (1995) 868–878, https://doi.org/10.1021/bi00003a021.

[20] J. Wirmer, P. Wolfgang, H. Schwalbe, Motional properties of unfolded ubiquitin: a model for a random coil protein, J. Biomol. NMR (2006), https://doi.org/10.1007/s10858-006-9026-9.

[21] R.S. Mann, M. Affolter, Hox proteins meet more partners, Curr. Opin. Genet. Dev. 8 (1998) 423–429, https://doi.org/10.1016/S0959-437X(98)80113-5.

[22] H.D. Ryoo, R.S. Mann, The control of trunk Hox specificity and activity by extradenticle, Genes Dev. 13 (1999) 1704–1716, https://doi.org/10.1101/gad.13.13.1704.

[23] R. Joshi, J.M. Passner, R. Rohs, R. Jain, A. Sosinsky, M.A. Crickmore, V. Jacob, A.K. Aggarwal, B. Honig, R.S. Mann, Functional specificity of a Hox protein mediated by the recognition of minor groove structure, Cell. 131 (2007) 530–543, https://doi.org/10.1016/j.cell.2007.09.024.

[24] N.C. Fitzkee, G.D. Rose, Reassessing random-coil statistics in unfolded proteins, Proc. Natl. Acad. Sci. (2004), https://doi.org/10.1073/pnas.0404236101.

[25] C. Narayanan, K. Bafna, L.D. Roux, P.K. Agarwal, N. Doucet, Applications of NMR and computational methodologies to study protein dynamics, Arch. Biochem. Biophys. (2017) https://doi.org/10.1016/j.abb.2017.05.002.

[26] E. Mastrangelo, M. Nardini, M. Bolognesi, One hundred years of X-ray diffraction, 50 years of structural biology, Rend. Lincei 24 (2013) 93–99, https://doi.org/10.1007/s12210-012-0214-0.

[27] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm, J. Mol. Biol. 293 (1999) 321–331, https://doi.org/10.1006/jmbi.1999.3110.

[28] T. Chouard, Structural biology: breaking the protein rules, Nature. 471 (2011) 151–153, https://doi.org/10.1038/471151a.

[29] W.Y. Choy, J.D. Forman-Kay, Calculation of ensembles of structures representing the unfolded state of an SH3 domain, J. Mol. Biol. 308 (2001) 1011–1032, https://doi.org/10.1006/jmbi.2001.4750.

[30] C. Fisher, A. Huang, C. Stultz, Modeling intrinsically disordered proteins with Bayesian statistics, J. Am. Chem. (2010) 14919–14927.

[31] V. Ozenne, F. Bauer, L. Salmon, J.R. Huang, M.R. Jensen, S. Segard, P. Bernadó, C. Charavay, M. Blackledge, Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables, Bioinformatics. 28 (2012) 1463–1470, https://doi.org/10.1093/bioinformatics/bts172.

[32] K.A. Crowhurst, M. Tollinger, J.D. Forman-Kay, Cooperative interactions and a non-native buried Trp in the unfolded state of an SH3 domain, J. Mol. Biol. 322 (2002) 163–178, https://doi.org/10.1016/S0022-2836(02)00741-6.

[33] H. Schwalbe, K.M. Fiebig, M. Buck, J.A. Jones, S.B. Grimshaw, A. Spencer, S.J. Glaser, L.J. Smith, C.M. Dobson, Structural and dynamical properties of a denatured protein. Heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea, Biochemistry. (1997) https://doi.org/10.1021/bi970049q.

[34] N.A. Farrow, O. Zhang, J.D. Forman-Kay, L.E. Kay, Characterization of the backbone dynamics of folded and denatured states of an SH3 domain, Biochemistry. 36 (1997) 2390–2402, https://doi.org/10.1021/bi962548h.

[35] A.B. Mantsyzov, Y. Shen, J.H. Lee, G. Hummer, A. Bax, MERA: a webserver for evaluating backbone torsion angle distributions in dynamic and disordered proteins from NMR data, J. Biomol. NMR 63 (2015) 85–95, https://doi.org/10.1007/s10858-015-9971-2.

[36] M.R. Jensen, L. Salmon, G. Nodet, M. Blackledge, Defining conformational ensembles of intrinsically disordered and partially folded proteins from chemical shifts, J. Am. Chem. Soc. 132 (2010) 1270–1272.

[37] J.R. Allison, P. Varnai, C.M. Dobson, M. Vendruscolo, Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements, J. Am. Chem. Soc. 131 (2009) 18314–18326, https://doi.org/10.1021/ja904716h.

[38] S. Bibow, V. Ozenne, J. Biernat, M. Blackledge, E. Mandelkow, M. Zweckstetter, Structural impact of proline-directed pseudophosphorylation at AT8, AT100, and PHF1 epitopes on 441-residue tau, J. Am. Chem. Soc. 133 (2011) 15842–15845, https://doi.org/10.1021/ja205836j.

[39] T. Mittag, J.D. Forman-Kay, Atomic-level characterization of disordered protein ensembles, Curr. Opin. Struct. Biol. 17 (2007) 3–14, https://doi.org/10.1016/j.sbi.2007.01.009.

[40] M.F. Berger, G. Badis, A.R. Gehrke, S. Talukder, A.A. Philippakis, L. Peña-Castillo, T.M. Alleyne, S. Mnaimneh, O.B. Botvinnik, E.T. Chan, F. Khalid, W. Zhang, D. Newburger, S.A. Jaeger, Q.D. Morris, M.L. Bulyk, T.R. Hughes, Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences, Cell. 133 (2008) 1266–1276, https://doi.org/10.1016/j.cell.2008.05.024.

[41] M.B. Noyes, R.G. Christensen, A. Wakabayashi, G.D. Stormo, M.H. Brodsky, S.A. Wolfe, Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites, Cell. 133 (2008) 1277–1289, https://doi.org/10.1016/j.cell.2008.05.023.

[42] R.S. Mann, K.M. Lelli, R. Joshi, Chapter 3 Hox specificity. Unique roles for cofactors and collaborators, Curr. Top. Dev. Biol. (2009), https://doi.org/10.1016/S0070-2153(09)88003-4.

[43] M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H.J. Bussemaker, R.S. Mann, Cofactor binding evokes latent differences in DNA binding specificity between hox proteins, Cell. 147 (2011) 1270–1282, https://doi.org/10.1016/j.cell.2011.10.053.

[44] A.K. Dunker, M.S. Cortese, P. Romero, L.M. Iakoucheva, V.N. Uversky, Flexible nets: the roles of intrinsic disorder in protein interaction networks, FEBS J. 272 (2005) 5129–5148, https://doi.org/10.1111/j.1742-4658.2005.04948.x.

[45] T. Mittag, L.E. Kay, J.D. Forman-Kaya, Protein dynamics and conformational disorder in molecular recognition, J. Mol. Recognit. 23 (2010) 105–116, https://doi.org/10.1002/jmr.961.

[46] Z. Dosztányi, J. Chen, A.K. Dunker, I. Simon, P. Tompa, Disorder and sequence repeats in hub proteins and their implications for network evolution, J. Proteome Res. 5 (2006) 2985–2995, https://doi.org/10.1021/pr060171o.

[47] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, Preformed structural elements feature in partner recognition by intrinsically unstructured proteins, J. Mol. Biol. 338 (2004) 1015–1026, https://doi.org/10.1016/j.jmb.2004.03.017.

[48] R.K. Das, R.V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues, Pnas. 110 (2013) 13392–13397, https://doi.org/10.1073/pnas.1304749110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1304749110.

[49] M. Gouw, S. Michael, S. Hugo, M. Kumar, B. Lang, B. Bely, B. Chemes, N.E. Davey, Z. Deng, A. Huber, S. Kleinsorg, F. Diella, G. Clara-marie, K.V. Roey, B. Altenberg, A. Rem, S. Schlegel, T.J. Gibson, The eukaryotic linear motif resource—2018 update, 46, 2018 428–434, https://doi.org/10.1093/nar/gkx1077.

[50] H. Dinkel, K. Van Roey, S. Michael, N.E. Davey, J. Weatheritt, D. Born, T. Speck, D. Kru, G. Grebnev, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, B. Chemes, J. Glavina, I.E. Sa, M. Seiler, F. Diella, T.J. Gibson, The eukaryotic linear motif resource ELM: 10 years and counting, 42, 2014, https://doi.org/10.1093/nar/gkt1047.

[51] A. Mohan, C.J. Oldfield, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, V.N. Uversky, Analysis of Molecular Recognition Features (MoRFs), 2006 1043–1059, https://doi.org/10.1016/j.jmb.2006.07.087.

[52] J. Yan, A.K. Dunker, V.N. Uversky, L. Kurgan, Molecular recognition features (MoRFs) in three domains of life, Mol. BioSyst. 12 (2016) 697–710, https://doi.org/10.1039/C5MB00640F.

[53] A. Sivashanmugam, V. Murray, C. Cui, Y. Zhang, J. Wang, Q. Li, Practical protocols for production of very high yields of recombinant proteins using *Escherichia coli*, Protein Sci. (2009), https://doi.org/10.1002/pro.102.

[54] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, Protein identification and analysis tools on the ExPASy server, in: J.M. Walker (Ed.), Proteomics Protoc. Handb, Humana Press, Totowa, NJ 2005, pp. 571–607, https://doi.org/10.1385/1-59259-890-0:571.

[55] M. Sattler, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients, Prog. Nucl. Magn. Reson. Spectrosc. 34 (1999) 93–158, https://doi.org/10.1016/S0079-6565(98)00025-9.

[56] S.C. Panchal, N.S. Bhavesh, R.V. Hosur, Improved 3D triple resonance experiments , HNN and HN(C)N , for HN and ^{15}N sequential correlations in (13C, 15N) labeled proteins: application to unfolded proteins, Signals. (2001) 135–147.

[57] F. Delaglio, S. Grzesiek, G.W. Vuister, G. Zhu, J. Pfeifer, A. Bax, NMRPipe: a multidimensional spectral processing system based on UNIX pipes, J. Biomol. NMR 6 (1995) 277–293, https://doi.org/10.1007/BF00197809.

[58] W. Lee, W.M. Westler, A. Bahrami, H.R. Eghbalnia, J.I. Markley, PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy, Bioinformatics. 25 (2009) 2085–2087, https://doi.org/10.1093/bioinformatics/btp345.

[59] M.J. Stone, P.E. Wright, W.J. Fairbrother, A.G. Palmer, J. Reizer, M.H. Saier, Backbone dynamics of the Bacillus subtilis glucose permease IIA domain determined from 15N NMR relaxation measurements, Biochemistry. 31 (1992) 4394–4406, https://doi.org/10.1021/bi00133a003.

[60] M.J. Stone, K. Chandrasekhar, P.E. Wright, H.J. Dyson, A. Holmgren, Comparison of backbone and tryptophan side-chain dynamics of reduced and oxidized *Escherichia coli* thioredoxin using 15N NMR relaxation measurements, Biochemistry. 32 (1993) 426–435, https://doi.org/10.1021/bi00053a007.

[61] N. Farrow, O. Zhang, A. Szabo, D. Torchia, Spectral density function mapping using 15N relaxation data exclusively, J. Biomol. NMR 6 (1995).

[62] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725–738, https://doi.org/10.1038/nprot.2010.5.

[63] C.A. Smith, T. Kortemme, Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction, J. Mol. Biol. 380 (2008) 742–756, https://doi.org/10.1016/j.jmb.2008.05.023.

[64] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B.L. De Groot, H. Grubmüller, A.D. MacKerell, CHARMM36m: an improved force field for folded and intrinsically disordered proteins, Nat. Methods 14 (2016) 71–73, https://doi.org/10.1038/nmeth.4067.

[65] R.B. Best, X. Zhu, J. Shim, P.E.M. Lopes, J. Mittal, M. Feig, A.D. MacKerell, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles, J. Chem. Theory Comput. 8 (2012) 3257–3273, https://doi.org/10.1021/ct300400x.

[66] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926–935, https://doi.org/10.1063/1.445869.

[67] S.E. Feller, Y. Zhang, R.W. Pastor, B.R. Brooks, Constant pressure molecular dynamics simulation: the Langevin piston method, J. Chem. Phys. 103 (1995) 4613–4621, https://doi.org/10.1063/1.470648.

[68] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: an N•log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092, https://doi.org/10.1063/1.464397.

[69] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD, J. Comput. Chem. 26 (2005) 1781–1802, https://doi.org/10.1002/jcc.20289.