CHAPTER 3

Identification of specific evolutionarily conserved regions on DDX3X
CHAPTER 3: Identification of Specific Evolutionarily Conserved Functional Residues on DDX3X

CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 3: Identification of specific evolutionarily conserved functional residues on DDX3X</td>
<td>68-110</td>
</tr>
<tr>
<td>3.A. Abstract</td>
<td>66</td>
</tr>
<tr>
<td>3.B. Introduction</td>
<td>67-68</td>
</tr>
<tr>
<td>3.C. Rationale and Hypothesis</td>
<td>69</td>
</tr>
<tr>
<td>3.D. Objectives</td>
<td>69</td>
</tr>
<tr>
<td>3.E. Material and Methods</td>
<td>70-77</td>
</tr>
<tr>
<td>3.E.1. Collection of Sequence Data Set</td>
<td>71</td>
</tr>
<tr>
<td>3.E.2. Alignment and Phylogenetic Ranking</td>
<td>71-72</td>
</tr>
<tr>
<td>3.E.4. Protein Structures and Refinement</td>
<td>74</td>
</tr>
<tr>
<td>3.E.5. Estimating Constraints</td>
<td>74-75</td>
</tr>
<tr>
<td>3.E.6. Pharmacophore Modeling</td>
<td>75</td>
</tr>
<tr>
<td>3.E.7. Flexibility Analyses</td>
<td>75-76</td>
</tr>
<tr>
<td>3.E.8. Docking Simulations</td>
<td>76</td>
</tr>
<tr>
<td>3.E.9. Molecular Dynamics Simulations</td>
<td>77</td>
</tr>
<tr>
<td>3.E.10. Statistical Analyses</td>
<td>77</td>
</tr>
<tr>
<td>3.F. Results</td>
<td>78-107</td>
</tr>
<tr>
<td>3.F.1 Identification of unbiased pool of DEAD box helicase sequences for selection analyses</td>
<td>79-82</td>
</tr>
<tr>
<td>3.F.2 Selection Analyses and Identification of Functionally Important Residues</td>
<td>82-84</td>
</tr>
<tr>
<td>3.F.3</td>
<td>Mapping of Functionally Important Residues on DDX3X Surface</td>
</tr>
<tr>
<td>3.F.4</td>
<td>Role of Functional Residues in Generating Constraints at ATP Binding Site of DDX3X</td>
</tr>
<tr>
<td>3.F.5</td>
<td>Role of Functional Residues in Regulating RNA Unwinding Function of DDX3X</td>
</tr>
<tr>
<td>3.F.6</td>
<td>DDX3X-CRM-1 Docking and Functional Effects of Constraints on DDX3X-CRM-1 Interaction</td>
</tr>
<tr>
<td>3.F.7</td>
<td>Selected Functional Residues on DDX3X</td>
</tr>
<tr>
<td>3.F.8</td>
<td>Identifying Specificity of Functional Clusters Towards DDX3X</td>
</tr>
<tr>
<td>3.F.9</td>
<td>Role of DDX3X Functional Clusters in Regulating Conformational Transition</td>
</tr>
<tr>
<td>3.F.10</td>
<td>Pharmacophore Design Based on Functional Regions</td>
</tr>
<tr>
<td>3.G.</td>
<td>Discussion</td>
</tr>
</tbody>
</table>
3.A. ABSTRACT

Differential host-pathogen interactions direct viral replication in infected cells. In HIV-1 infected cells, nuclear export of viral RNA transcripts into cellular cytoplasm is a critical event for efficient translation of HIV-1 proteins and in turn, for production of infectious virus particles. This step requires association of > 50 host factors with a 16kDa protein of HIV-1 aptly named Rev (Regulator of expression of Virion particles). HIV-1 Rev, a viral accessory protein primarily regulates efficient RNA translocation by elegantly achieving optimal post transcriptional regulation of nascent HIV-1 RNA molecule. To achieve this, Rev closely associates with DDX3X, a DEAD box RNA helicase.

DDX3X is ubiquitously present across divergent species with multiple inferred functions, attributable to its functional motifs. However, the identity of DDX3X regions specifically exploited by HIV-1 is still a topic of research. Based on the assumption that study of protein evolution can potentially identify functional residues on its surface, in this part of the study, we analyzed the co-evolution pattern of HIV-1 Rev with DDX3X and identified novel functionally important residues on the surface of DDX3X. Further, employing computational approaches, we show that these functional residues, depending on their location, are capable of regulating ATPase and RNA helicase functions of DDX3X. Additionally, using stepwise docking simulations, putative DDX3X-CRM-1 interface was also marginalized. Structural comparisons with different DEAD box helicases aided in detection of inherent DDX3X specificity for the identified functional clusters. Finally, the potential of these residues in designing better blockers against HIV-1 replication was assessed.
3.B. INTRODUCTION

HIV-1 gene expression is regulated by complex interplay of cellular and viral factors. One such important step is post-transcriptional regulation of HIV-1 gene expression by HIV-1 Rev-host factors interaction. Rev binds to a highly structured element called Rev Response Element (RRE) present in all viral transcripts (Malim, Hauber et al. 1989). With the aid of CRM-1/Exportin-1, Rev takes the bound viral RNA transcripts out of the nucleus (Askjaer, Jensen et al. 1998). In doing so, Rev uses the CRM-1 dependent nuclear-cytoplasmic transport pathway normally associated for export of various proteins in the cell (Neville, Stutz et al. 1997). This pathway employs plethora of host factors. The relation of these host factors in Rev-CRM-1 mediated export of nascent HIV-1 transcripts is not clear. One of the recent findings has shown direct physical interaction of an energy dependent DEAD box RNA helicase DDX3X with expotin-1 (CRM-1) with implications in HIV-1 Rev function (Yedavalli, Neuveut et al. 2004). Also, analyses of HIV-1 Rev protein interaction network described DDX3X as component of highly active sub network of HIV-1 host protein interactions. Generally, DEAD box RNA helicases are most ubiquitously present in the cellular milieu and perform diverse functions related with RNA metabolism. As described in Review of Literature, these helicases are characterized by strong sequence homology with nine highly conserved motifs encompassing the protein sequence. One of these motifs is the DEAD box motif and thereby referred to as DEAD box helicases (Luking, Stahl et al. 1998).

Protein-protein interaction studies revealed that DDX3X acts as an effector rather than cargo of CRM1/Exportin-1 (Yedavalli, Neuveut et al. 2004). The actual effector function carried out by this protein is not clear yet, but it seems to regulate HIV-1 Rev mediated RNA export out of the nucleus. Analysis of similar helicases reveals that DDX3X could participate in the terminal step of RNA export by removing proteins that are bound to RNA through nucleopore complex (Rocak and Linder 2004). Recently, DDX3X has also been related to innate antiviral immune mechanisms which highlighted its importance as a molecule being at the crossroads
of viral replication and antiviral activity (Soulat, Burckstummer et al. 2008). DDX3X had emerged as a validated target for development of antiviral therapeutics against HIV-1 (Schroder 2010). This focus on DDX3X can be attributed to 1) the identification of DDX3X as an important interactor of HIV-1 Rev protein in most of the proteomic screenings (Naji, Ambrus et al. 2012) along with 2) high mutability of HIV-1, rendering the virus resistant to drugs.

A series of ring expanded nucleoside analogues (REN) were identified to be capable of blocking HIV-1 replication by interfering with DDX3X function (Yedavalli, Neuveut et al. 2004, Garbelli, Beermann et al. 2011). Pharmacophore modeling and combinatorial synthesis of DDX3X helicase region was also carried out to shortlist novel intervention sites (Garbelli et al., 2011b). Additionally, point mutations of DDX3X ATPase motif (Motif I) or DDX3X helicase motif (Motif III) have been shown to abolish the unwinding activity of DDX3X along with reduction of HIV-1 replication (Yedavalli, Neuveut et al. 2004, Garbelli, Beermann et al. 2011, Garbelli, Radi et al. 2011, Schroder 2011). Deletion mutants of a number of DEAD box helicases have been described to affect HIV-1 replication by, either hampering ATP hydrolysis or by uncoupling ATP hydrolysis from RNA unwinding (Sengoku, Nureki et al. 2006). However, preferential utilization of DDX3X by HIV-1, despite presence of paraphernalia of >45 DEAD box helicases suggest involvement of certain specific functional surfaces of DDX3X in HIV-1 life cycle.

In this study, we attempted to identify these DDX3X specific functional surfaces with the help of evolutionary history of this protein. We combined a series of evolution based models to study DEAD box helicase family with an aim to infer inherent selection pattern representative of DDX3X. We have also exploited the knowledge about structural architectures of different DEAD box proteins to finally pin point critical functional clusters on DDX3X.
3.C. RATIONALE AND HYPOTHESIS

From previous studies, it has become evident that DDX3X is a critical contributor of HIV-1 nuclear export event. The specificity of DDX3X for HIV-1 suggests existence of potentially important functional clusters on this protein. Further, the observed dispensability of DDX3X for cellular metabolism makes it an ideal intervention target against HIV-1 replication. In this study, we employed an approach that involved identification of functional clusters on DDX3X and combined it with detailed structural analyses of DEAD box helicases.

We hypothesize that evolution dependent selection pressure on DDX3X can potentially generate certain residue clusters that are highly specific for DDX3X activity. These clusters, once identified, can directly be targeted to impact HIV-1 replication.

3.D. OBJECTIVES

1. To identify novel evolutionarily conserved functional residues on DDX3X

2. To infer the specificity of identified functional regions towards DDX3X.
3.E.

Material

And Methods
3.E.1 Collection of Sequence Data Set

In order to identify selection pattern on DDX3X, primate cluster of sequences were selected from UniRef 90 database (Suzek, Huang et al. 2007) (for bias towards human sequences) for each DEAD box family member. The UniRef databases provide clustered sets of sequences from UniProt Knowledgebase (including splice variants and isoforms) and selected UniParc records, in order to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences (but not their descriptions) from view. The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues (from any organism) into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries, and links to the corresponding UniProtKB and UniParc records. UniRef90 and UniRef50 are built by clustering UniRef100 sequences with 11 or more residues using the CD-HIT algorithm (Li, Jaroszewski et al. 2001) such that each cluster is composed of sequences that have at least 90% or 50% sequence identity, respectively, to the longest sequence (UniRef seed sequence). Using these clusters, a total of 178 complete nucleotide coding sequences (CDS) belonging to 45 DEAD box family members from 6 different species were retrieved from NCBI Reference Sequence (RefSeq) collection (Pruitt, Tatusova et al. 2005).

3.E.2 Alignment and Phylogenetic Ranking

The selected 178 DEAD box helicase sequences were aligned using a method described by Gonnet, Cohen and Brenner (1992) (Gonnet, Cohen et al. 1992) utilizing 3 iterations and divergent cut off delay of 30% (Clustal X) (Thompson, Gibson et al. 1997). Classical distance measures were used to estimate alignment of proteins. The data obtained was used to generate a new distance matrix. The latter was used to refine the alignment, estimate a new distance matrix. The process was repeated 3 times. This method gave better alignments compared to normalizations using widely used PAM 250 matrix (Point Accepted Mutation corresponding to 250 mutations per 100 amino acids of sequence). Briefly, initial alignment was carried using Gonnet matrix and subsequent refinements were done using PAM 250 matrix.
Aligned sequences were then manually checked in MEGA 4 with in-built Clustal W (Tamura, Peterson et al. 2011) based on two criteria:

1) Average identity between sequences was greater than 50% at the amino acid level and

2) Percentage of gaps was less than 25%.

Based on the above mentioned criteria, 57 members were shortlisted for evolutionary analyses. DEAD box helicases with highest pair wise alignment score with respect to DDX3X were finally chosen. Evolutionary history of the final 19 refined and shortlisted coding sequences (corresponding to amino acid position 167 to 520) was inferred by Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 8.299 was shown. The tree was drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura, Nei et al. 2004) and were in the units of the number of base substitutions per site. Codon positions included were 1st + 2nd + 3rd + Noncoding. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). There were a total of 789 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 (Kumar, Tamura et al. 2004, Tamura, Peterson et al. 2011).

3.E.3 Selection Analysis

Recombination in the dataset was tested by using a Genetic Algorithm for Recombination Detection (GARD) (Pond and Frost 2005). Identified breakpoints were assessed by Kishino-Hasegava test implemented in DataMonkey server (Pond and Frost 2005). To test for diversifying selection and to infer codons under positive selection the $\omega$ ratio was calculated with the computer program Codeml from PAML package (Yang 2007). The relative fit of codon substitution models was evaluated with likelihood ratio tests (LRT), which were assumed to be $\chi^2$ distributed with
degrees of freedom equal to the difference in number of parameters between models. LRTs for positive selection compare a model in which there is class of sites with $\omega > 1$ against a model that does not allow for this class. LRTs were tested with M8 (beta and $\omega$) and M8a (beta and $\omega > 1$) models. In order to examine the robustness of the positive selections identified by PAML, datasets were also analyzed using two additional methods using HYPHY package (available from DataMonkey facility) (Pond and Frost 2005). These models were single likelihood ancestor counting (SLAC) and fixed effects likelihood (FEL) (Pond and Frost 2005).

For identification of lineage and site specific selection patterns CodeML implemented in PAML program was used. CodeML needs three inputs:

- A multiple sequence alignment file (MSA), under phylip format.
- A phylogenetic tree, under nexus format.
- A control file where all option are provided.

Some of the terms employed in this analyses are summarized below:

$dS = \text{Synonymous mutation substitution rate}$

$dN = \text{Non-synonymous mutation substitution rate}$

$\omega = \frac{dN}{dS}$

$kappa = \frac{ts}{tv}$ (transition/transversion rate)

$dN/dS < 1 \Rightarrow \text{purifying selection}$

$dN/dS = 1 \Rightarrow \text{neutral evolution}$

$dN/dS > 1 \Rightarrow \text{positive Darwinian selection}$

For the phylogeny based identification of functionally important residues on DDX3X surface, conservation scores were obtained by likelihood based approach using Jules-Thornton-Taylor (JTT) matrix and using Consurf server (PSI BLAST E-value cut-off = 0.001; Iterations = 3) (Landau, Mayrose et al. 2005). The topology based functional residue classification was obtained by structural alignment of identical proteins using Dali facility (Holm, Kaariainen et al. 2008) and subsequent measure of root mean square deviations (RMSD).
3.E.4  Protein Structures and Refinement:

Protein structure coordinates for DDX20 (2OXC), DDX10 (2PL3), DDX41 (2P6N), DDX3X (2I4I), DDX19 (3EWS and 3G0H), DDX47 (3BER), DDX5 (3FE2), DDX18 (3LY5), DDX52 (3DKP), DDX20 (3B7G), DDX53 (3IUY), DDX25 (2RB4), DDX4 (2DB3), eIF4 (2G9N) and Exportin-1 (CRM-1) were obtained from RCSB PDB (Berman, Battistuz et al. 2002). The structures were energy minimized (steps = 100, step size = 0.2 Å) using rotamers derived from Dunbrack library (Dunbrack 2002) and the charges were computed using Antechamber method (Wang, Wang et al. 2006). All the structural modeling was carried out in UCSF Chimera extensible molecular modeling environment (Pettersen, Goddard et al. 2004).

3.E.5  Estimating Constraints

For the generation of all-atom contact map, DDX3X structure was refined by REFMAC, hydrogen atoms were added using REDUCE and contact dots were calculated with PROBE. All these tools were accessed through MolProbity (Davis, Murray et al. 2004). Docking of ATP molecule to DDX3X was carried using ArgusLab software, version 4.0 (Joy, Nair et al. 2006). Protein Ligand docking efficiencies had been successfully tested by various groups for ArgusLab (Joy, Nair et al. 2006). ATP mol2 file was obtained from ChemBank (Seiler, George et al. 2008). Hydrogen atoms were added to ligand coordinate file before docking with ArgusLab. Mutant DDX3X structures were generated in Chimera and each structure was minimized. The docking between each mutant DDX3X receptor and ATP was performed by defining a binding site using bound AMP molecule (cubic box = 80×80×80 cubic angstrom). Docking simulations were carried out using “ArgusDock” as docking engine, “Dock” was chosen as the calculation type and AScore was used as the scoring function. All mutant DDX3X-ATP energy results were compared with control docking of wild type DDX3X and ATP. Charge distribution was computed by mapping electrostatic potential (ESP) onto electron
density (ArgusLab). This calculation helped to show regions that might act as critical electron acceptor or electron donor. Knowledge based scoring function for protein-ligand interactions implemented in DrugScore was used to analyze individual atomic contributions of functional residues towards ATP binding (Gohlke, Hendlich et al. 2000).

Relative difference in binding energy between mutant and wild type DDX3X and ATP molecule was obtained from ArgusLab AScore. All the structural alignments were carried out in MatchMaker tool of Chimera and Root Mean Square Deviations (RMSD) were obtained. RNA-DDX3X docking was performed as described above (ArgusLab). DDX3X RNA binding was identified by structural alignment between DDX3X and DDX4 using Chimera. Alanine scanning mutagenesis was performed using Concoord/PBSA (Benedix, Becker et al. 2009). This process mutates structure using WHATIF and a structure ensemble is generated using Concoord. Interaction energies were evaluated by Lennard-Jones potential and Coulomb's Law (parameters from GROMOS96 53a6 force field).

3.E.6 Pharmacophore Modeling

For ligand based pharmacophore design, multiple flexible alignments of ligands was employed using PharmaGist (Schneidman-Duhovny, Dror et al. 2008). The ligand mol2 coordinate files were generated in ArgusLab. Receptor based pharmacophore was defined in FlexX using interaction constraints at ATP binding site of DDX3X (Rarey, Kramer et al. 1996). Binding site was identified using bound AMP as an anchor (distance = 10 Å).

3.E.7 Flexibility Analysis

We performed a Normal Mode Analysis (NMA) on the DDX3X protein structure. We used the approximate NMA developed by Hinsen (Hinsen 1998), which represents very well low-frequency domain motions at negligible computational cost. Briefly, a normal mode analysis consists of the diagonalization of
the matrix of the second derivatives of the energy with respect to the displacements of the atoms, in mass-weighted coordinates (Hessian matrix). The eigenvectors of the Hessian matrix are the normal modes, and its Eigen values are the squares of the associated frequencies. The low-frequency normal modes are believed to be the ones functionally important. The NMA tools are implemented in the Molecular Modeling Toolkit (MMTK) (*Hinsen 1998*). All modes were calculated, i.e., three times the number of Ca atoms (3×418 = 1254). The lowest frequency modes are selected.

A vector field representation was calculated as described by Thomas *et al* (*Thomas, Hinsen et al. 1999*). The vector field was calculated over cubic regions with an edge length of 3 Å, containing on average 1.3 Ca atoms. The vector field defined on a regular lattice at the center of each cube is the mass-weighted average of the displacements of the atoms in the cube.

### 3.E.8 Docking Simulations

We carried out a series of conformation based local and global docking simulations to identify DDX3X-CRM-1 interaction interface. Initial orientations for global docking simulations were carried out in Hex (*Ritchie and Kemp 2000*) which uses Spherical Polar Fourier (SPF) correlations to accelerate the calculations. Final local docking steps were done in FireDock (*Andrusier, Nussinov et al. 2007, Mashiach, SchnieDman-DuhoVny et al. 2008*) which predicts the structure of protein complexes given the structures of the individual components and an approximate binding orientation. Both of these programs were tested in CAPRI (Critical Assessment of Predicted Interactions) blind trials and were shown to give accurate results within 1 angstrom of RMSD (*Vajda and Kozakov 2009*). Contacts between amino acids were deduced using molprobity. Hydrogen atoms were added using REDUCE and all atom contact map generated by the program PROBE of MOLPROBITY. Amino acid side chain volume was obtained by Voronoi Polyhedra (*Vajda and Kozakov 2009*).
3.E.9 Molecular Dynamics Simulations

All molecular dynamics simulations were carried out in NAMD (Phillips, Braun et al. 2005). Six independent molecular dynamics simulations corresponding to “open” and “closed” states of DDX3X were carried out for 300ps with an equilibration phase of 50ps. The molecular trajectories for the two systems generated by molecular dynamics simulations were analyzed in VegaZZ (Pedretti, Villa et al. 2002, Pedretti, Villa et al. 2004). Basically, the same MD simulation protocol was applied for all the simulations. We used NPT ensemble with explicit solvent and periodic boundary conditions (Berendsen and Hayward 2000). Integration of equation of motion was carried out by Leap-Frog algorithm. All bonds containing hydrogen atoms were constrained to their equilibrium values by means of SHAKE algorithm. The velocities were reassigned after every 0.5ps according to Maxwell-Boltzmann distribution. The complexes were immersed in a cubic box (72x72x72Å) and the solvated system was neutralized by steepest descent and conjugate gradient minimization. Simulation was basically carried out using the protocol described by Moustafa et al. and was composed of following steps: 1) Short energy minimization (water and ions only), 2) Short constraint dynamics (only water and ions were allowed to move under NPT ensemble or constant pressure and temperature) and 3) Energy minimization (protein side chains + water ions) 4) System heating and all atom dynamics. MD trajectories were analyzed in VMD, VegaZZ and Chimera (Moustafa, Shen et al. 2011).

3.E.10 Statistical Analysis

All the statistical analyses were carried out in Microsoft Excel software and using standard statistical tables. The alignments were represented using WebLogo (Crooks, Hon et al. 2004). Sets of data were analyzed using Student's t-test. Statistical significance was assessed using P-value: P<0.05, P<0.01 and P<0.001. Standard errors were obtained for all the observations.
3.F.

RESULTS
3.F.1 Identification of unbiased pool of DEAD box helicase sequences for selection analyses.

Initially, 1486 DEAD box helicase sequences belonging to Eutherian cluster were retrieved from Uniprot (Figure 3.1A). The sequences were compared with the RNA helicase database (Jankowsky, Guenther et al. 2011) to ensure all the DEAD box helicase members are selected. The sequences encompassed a total of 45 DEAD box helicase members from seven different species. Coarse alignment by a combination of PAM120 and PAM250 matrices (after giving different scores to gaps and indels or insertion deletions) resulted in 648 refined sequences sharing >45% identity to DDX3X (Figure 3.1B). Subsequently, local alignment corresponding to the highly conserved DEAD box helicase regions was carried out using BLOSUM 62 and BLOSUM 45 matrices. The sequences with alignment score values < 28 with respect to DDX3X were selected (Figure 3.1B).

Figure 3.1: Identification of DEAD box helicase sequence cluster: (A) Alignment scores for DEAD box helicase sequences compared to DDX3X. Only representative DEAD box members are shown for simplicity. (B) Schematic representation of iterative selection of sequence cluster with respect to DDX3X.
PAM matrices globally align more divergent sequences, while BLOSUM matrices can be used to align local highly conserved regions in a sequence. Gonnet is a modification of PAM matrices and is employed for refinement of aligned sequences in an iterative manner.

Finally, we are left with 178 nucleotide coding sequences (CDS) corresponding to 45 DEAD box family members (primate cluster). Initial alignment revealed that genes were still too divergent for analysis of positive selection. So, the sequences with high number of In-Dels (Insertion-Deletion) were removed. In addition to that, regions with high diversity were also deleted from all the sequences. DEAD box amino acid region 167 to 520 was finally obtained from remaining 57 CDS. Out of these 57 sequences, those with highest pair wise alignment score with DDX3X coding sequence were kept for further study and rest were discarded (Figure 3.2A). We started by analyzing the lineage specific differences in selection patterns between DEAD box helicase sequences from different primate species. A Neighbor-Joining tree corresponding to 57 primate sequences with bootstrap values and branch lengths was constructed and is shown in Figure 3.2A. We started by estimating global dN/dS ratio for each of the branches (for details see Material and Methods and Appendix A.2). Finally, likelihood ratio test was carried out for significance of model depicting positive selection compared to null model. This analysis did not suggested evidence for predominant positive selection pattern in one primate lineage compared to another. This could be attributed to relatively similar branch length values for different primate lineages (Figure 3.2A). Due to absence of lineage specific selection pattern, neighbor joining tree generated using homologous sequences for a given DEAD box helicase resulted in bias towards more similar sequences. For instance, DDX3X coding sequence depicted an alignment score of >50 for most of the homologues, resulting in high level of background synonymous variations (Figure 3.2B). This might interfere with identification of important residues under purifying selection. To reduce this similarity based bias, finally 19 DEAD box sequences belonging to humans were finalized out of 57 members. These members were represented in a (Figure 3.2B)
Figure 3.2: Shortlisted DEAD box helicases after alignment based iterations. (A) Neighbor Joining tree for DEAD box helicase members from Eutherian cluster (B) Conserved DEAD box motifs common to all DEAD box helicases including DDX3X with inter-motif distances. (C) Neighbor joining tree for selected DEAD box members. The tree had sum of branch lengths = 8.2995. Cut off for bootstrap test (1000 replicates) = .50%. Branch lengths were shown at the bottom of each branch.
phylogram generated by Neighbor Joining method using amino acids from translated coding sequences (Figure 3.2C). Each individual branch lengths were utilized for subsequent selection analyses.

### 3.F.2 Selection Analysis and Identification of Functionally Important Residues

We aimed to study site specific selection pattern in the selected 167–520 amino acid region of DEAD box members. However, due to the effects of recombination, there is always a probability of obtaining false positive results in any selection study. In order to negate this effect, we first analyzed the recombination acting in these sequences. Recombination analysis revealed presence of three breakpoints in sequences which could alter the outcome of selection studies (Figure 3.3A). These breakpoints were present at nucleotide positions 186, 769 and 1122 respectively. We used these breakpoints to generate four separate Neighbor-Joining (NJ) trees. Using these trees as an input, four separate selection analysis were carried (Figure 3.3B).

![Figure 3.3: Recombination analyses of DDX3X:](image)

(A) Major breakpoints shown with respect to DDX3X coding sequence. Nucleotide position of each breakpoint is highlighted. Separate Neighbor-Joining trees corresponding to each of the DDX3X fragment are shown in (B).
Maximum Likelihood based pair wise comparison of null and alternate hypothesis using a pair of evolutionary models (M0 vs M1 and M8 vs M8a) (Figure 3.4A and Figure 3.4B) revealed a log likelihood ratio difference in favor of the alternate hypothesis ($2\Delta L = 10.4$ at 0.005 level of significance). Briefly, certain amino acid sites depicted elevated level of substitution ($\omega>1$) indicating positive selection. Based on likelihood-ratio tests, both selection models were a better fit to the data, either in comparison to a nearly neutral model (Model 0 vs. 1a, $p<0.05$) or to a model with a Beta distribution (Model 8 vs. Model 8a, $p<0.01$) (Table 3.1). A naïve empirical Bayes (NEB) analysis indicated that positive selection is evident only at a small fraction of sites and the sites are depicted in Table 3.1 (for details of the terms, please refer to Appendix).

**Figure 3.4:** Schematic representation of the site specific codon substitution models employed for the selection analyses: (A) The models are based on analyses of pattern of synonymous/non synonymous changes ($K_s/K_a$) (B) Based on $K_s/K_a$ parameters, different models have been devised. The models are compared pair wise and significance of one model over another is estimated (representing positive and negative selection at a residue position).

To assess the robustness of positive selection, SLAC and FEL analyses were also carried out. Only the sites which were common in all the four methods were selected as being positively selected (Total 7). In addition to positive selection, 158 sites were identified to be under purifying or negative selection with 54 sites depicting strong Bayes factor scores (>30). This analysis described that DDX3X is under strong purifying selection (as inferred from negative selection) with few regions showing evidence for positive selection.
Conserved sites under negative selection are indicative of active sites or protein-protein interaction epitopes while highly variable sites may represent sites subjected to positive Darwinian selection. Such positively selected sites may be interpreted as being a consequence of molecular adaptation, which confers an evolutionary advantage to the protein.

Table 3.1: Parameters for detection of positive selection with positively selected sites shown

<table>
<thead>
<tr>
<th>MODEL</th>
<th>LognL</th>
<th>Parameter estimates</th>
<th>2ΔL</th>
<th>Positively Selected Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M8 (Beta and ω)</strong></td>
<td>19819.6</td>
<td>κ: 1.27707; α: 0.63411; β: 1.27467; additional omega category: 1.5 prob (additional omega category): 0.0172209</td>
<td>10.4 (0.005)</td>
<td>T198, S244, P247, A253, F402, G406, S410, F430, T438, H485, S520</td>
</tr>
<tr>
<td><strong>M8a (Null)</strong></td>
<td>19824.6</td>
<td>κ: 1.23607; α: 0.710644; β: 1.63009; additional omega category: 1 prob (additional omega category): 0.0699265</td>
<td>NOT ALLOWED</td>
<td></td>
</tr>
<tr>
<td><strong>SLAC</strong></td>
<td>--------</td>
<td>--------</td>
<td>N/A (0.2)</td>
<td>T198, P247, E256, R294, P297, I310, C317, K335, S382, G406, F430, T438, A467, H485</td>
</tr>
<tr>
<td><strong>FEL</strong></td>
<td>--------</td>
<td>--------</td>
<td>N/A (0.2)</td>
<td></td>
</tr>
</tbody>
</table>

84
3.F.3 Mapping of Functionally Important Residues on DDX3X

Surface

Next, we used the selection information (from Figure 3.4 and Table 3.1) to highlight functionally relevant residue clusters on DDX3X protein. By the term functionally important residues, we mean the residues required by the protein to perform its molecular function or biological role in such a way that they cannot be changed without affecting the function of protein. We identified functional residues based on two criteria:

1. Residues under positive selection and present on surface
2. Surface residues under negative selection with inferred functional importance

For identifying functional residues with positive selection, we simply mapped the positively selected sites on DDX3X (from Table 3.1) and related them with Accessible surface area (ASA) (Figure 3.5A and Figure 3.5B). Sites with accessible area greater than 30% were identified as functional.

Out of the identified positively selected sites, Gly$^{406}$, Phe$^{430}$, and His$^{485}$ were identified as functional residues, solely based on hydrophobicity or surface accessibility profiles. As shown in Figure 3.5C, Gly$^{406}$ constituted a part of β-turn connecting α helix of motif V (in N terminal domain) with the variable β sheet at the C terminal domain. Owing to the inter domain location, Gly$^{406}$ appears to participate in DDX3X conformational transition events. Phe$^{430}$ and His$^{485}$ encompassed DDX3X exposed site with a cavity of ~540Å$^2$, adjacent to ATP binding site. Thr$^{198}$ harbors nucleotide binding contact shell with direct van der wall’s contact with bound adenosine moiety. Based on structural comparisons, Pro$^{247}$ and Glu$^{256}$ appear to participate in RNA-DDX3X binding by regulating conformational transition of DDX3X (Figure 3.5D). Finally Phe$^{430}$ and Thr$^{438}$ complete the group of positive selected sites with a predominant exposed, non ATP binding and non RNA binding function of Pro$^{247}$ and Glu$^{256}$. Thus, each of the positively selected site can be ascribed a functional cluster on DDX3X (Figure 3.5E).
Figure 3.5: Mapping of positively selected sites on DDX3X surface: (A) Hydrophobicity plot of DDX3X (B) Plot of % Accessible Surface Area (ASA) versus positively selected residues. Structural features of positively selected sites shown for positions 406,430,438 (C) and 247, 256 (D) respectively. (E) Significant positively selected clusters mapped on DDX3X surface (stereo view).

For short listing functional importance in negatively selected amino acids, we employed a phylogenetic sequence based method profiling of DEAD box helicases. The alignment of selected DEAD box helicases was used to generate a residue wise conservation score using maximum likelihood approach (Jules-Thornton-Taylor matrix).
The conservation score corresponds to evolutionary rate of that position. Positions were color coded according to these scores and are depicted in Figure 3.6A. Only those residues were selected as functional which show high conservation scores along with surface accessibility of greater than 30%. These functional residues identified by sequence based phylogenetic approach were further validated using an independent approach based on structure comparisons. Pair wise structural alignment between DDX3X and other members of DEAD box helicase family revealed that the Root Mean Square Deviations (RMSD) at the functional residue sites identified by sequence based approach, was least while the RMSD at the non-functional sites was maximum (Figure 3.6B).

**Figure 3.6: Functional residue identification on DDX3X:** (A) Residues under negative selection (yellow line) grouped into functional amino acids based on conservation scores. The numbering corresponds to amino acid positions in DDX3X. f: functional residues; s: structural residues; e: exposed; b: buried C) Graph between Root Mean Square Deviation (RMSD) and residue position represents structural alignment of DEAD box members. D) DDX3X structurally aligned with DEAD box helicases.
3.F.4 Role of Functional Residues in Generating Constraints at ATP Binding Site of DDX3X

ATP binding to DDX3X was utilized as a potential target where a block was shown to result in inhibition of HIV-1 Rev activity (Yedavalli, Zhang et al. 2008). ATP analogs, which bind to ATP binding site of DDX3X, were shown to inhibit HIV-1 replication (Garbelli, Beermann et al. 2011, Garbelli, Radi et al. 2011). We thought that identification and analysis of critical and specific functional residues at this site would aid in explaining the specificity of DDX3X-ATP binding site and designing of selective and efficient inhibitors that could result in inhibition of HIV-1 replication. So, we studied the constraints acting at this site of DDX3X with reference to functional residues and analyzed their selectivity towards DDX3X. We started by short listing the functional residues that formed DDX3X-ATP binding site. We observed that ATP binds in a groove formed due to the bi-lobed nature of DDX3X protein structure (Figure 3.7A).

Using DDX3X crystal structure (PDB ID: 2I4I), an all atom contact map was created and AMP was selected as a marker point (AMP was present in the bound form in DDX3X crystal structure). Amino acids in the radius of 7 Å of AMP (corresponding to distance up to which inter atomic van der Wall forces are felt) were taken as members of its binding site. This selection revealed residue positions 200–211 and 225–231 in the proximity of AMP with Tyr\textsuperscript{200}, Pro\textsuperscript{203}, Trp\textsuperscript{204}, Gln\textsuperscript{207}, Gln\textsuperscript{225}, Thr\textsuperscript{226}, Gly\textsuperscript{227}, Ser\textsuperscript{228}, Gly\textsuperscript{229}, Lys\textsuperscript{230} and Thr\textsuperscript{231} being the functional residues (based on conservation score). ATP molecule was docked to DDX3X structure and the structure was energy minimized. We utilized inter-atomic interaction information for the identified functional residues to study constraints acting on them. Electrostatic potential-mapped electron density surface with inter-atomic contacts revealed that residues 227–231 formed close and direct inter-atomic associations (inter-atomic distance between 1–3 Å) with phosphate and ribose moieties of bound ATP molecule (ATP tail) with a preponderance of net positive charge (Positive ESP or Electrostatic...
Potential) (Figure 3.7B). Also, these residues were found to be highly conserved among DEAD box members (Figure 3.7C).

Figure 3.7: Functional residues at the ATP binding site of DDX3X: A) DDX3X structure (wireframe) showing the bound AMP (sphere) at the groove formed by bilobed DDX3X. B) Electrostatic potential map showing charge distribution near essential functional residues. Red are regions of negative ESP (Electrostatic potential). White represents positive ESP C) Alignment showing entropy at each functional residue position with most conserved residues having highest bit score. D) Atomic contributions for ATP binding for functional residues. Atoms contributing maximum are shown with spheres of large radius. The inter-atomic distance are shown alongside critical ATP-DDX3X association and is given in the units of Angstroms. E) Binding energy change values on substitutions at various functional residue positions. Substitutions: red – with positive charged residues; green – with negative charged residues; blue – with polar residues; yellow – with aromatic residues and purple – with non polar residues. F) Atomic displacements on ATP-AMP conversion. Thr226, Lys230 and Thr231 occupy critical positions in contact with residues of motifs II and IV. G) Q207 and P204 occupied crucial positions in the ATP binding site of DDX3X. Q207 forms the base of this cavity while P204 helps in alpha helix torsion angle. (For direct online access: doi:10.1371/journal.pone.0009613.g003)

Gly227 and Gly229 atomic contributions were found to be maximum towards binding ATP tail, with their Cα and N atoms in close contact with ATP H-5 and O-1 atoms (1.3245 Å and 1.2867 Å) respectively (Figure 3.7D). Substitutions with positively
charged and large polar aromatic residues at this site were found to substantially reduce ATP-DDX3X binding energy (Figure 3.7E). On the other hand, Lys\textsuperscript{230} hydrogen atoms were in close proximity to phosphate group of ATP (ATP tail) (with inter-atomic distances = 2.0965 Å–3.6543). These three functional residues thus carry greater share of ATP-DDX3X binding energy with strong positive charge responsible for stabilizing ATP tail region (Figure 3.7B and 3.7D). Also, in association with Thr\textsuperscript{226} and Thr\textsuperscript{231}, Lys\textsuperscript{230} makes important inter-domain contacts with motifs which participate in RNA unwinding process of DDX3X (Motifs II and IV) (Figure 3.7F). Probably, because of this direct role in ATPase activity and indirect role in RNA unwinding mechanism, Thr\textsuperscript{226}, Thr\textsuperscript{231} and Lys\textsuperscript{230} were found to be less tolerant towards any change (average ΔBE >2 kcal/mol). Among other functional residues, Tyr\textsuperscript{200} near the purine ring of bound ATP molecule was quite tolerant towards substitutions, which could be attributed to the ability of AMP purine ring to adopt different poses on binding DDX3X (Figure 3.7D). Pro\textsuperscript{203} was another interesting functional residue found to maintain DDX3X-ATP binding site (probably because of its function of extension of otherwise α helical segment of the ATP binding cleft) with significant effect on ATP binding energies (Figure 3.7E). Thr\textsuperscript{204} and Gln\textsuperscript{207} occupied interface of positive and negative charge distribution of the ATP binding cleft (Figure 3.7B) and were responsible for forming its (ATP binding cleft) base (Figure 3.7G). They were highly conserved among DEAD box members and most of the size changing variations here were damaging for ATP binding (because of changes in ATP binding cavity area).

So, taken together, we found that depending on the contribution towards ATP binding and depending on the role in maintaining the structural integrity of the ATP binding site, different functional residues at the DDX3X-ATP binding site were under various constraints which maintain their sequence conservation. In addition to the functional residues described above, we performed a DDX3X specific functional residue identification and found Thr\textsuperscript{198}, Arg\textsuperscript{199}, Arg\textsuperscript{202} and Gln\textsuperscript{225} to be specific to DDX3X.
Figure 3.8: Comparison of ATP binding site of DDX3X with other helicases: (A) Structural alignments of DDX3X with DDX53, DDX47 and EIF4A showing variations at residue positions 198–201 (B) Conformations of the DDX3X specific residues at position 198 to 201. Corresponding residues in other DEAD box helicases mainly yield two kinds of surfaces as shown. DDX3X surface has larger binding cavity at this site which was not found in any other helicase.

Next, we examined whether the functional residue constraints were responsible for generating similar architecture of ATP binding site for all DEAD box members or protein specific variations exist. We used crystal structures of available DEAD box helicases and compared their ATP binding regions, encompassing the functional residues shown in Figure 3.7B, with DDX3X ATP binding site. As shown in Figure 3.8A, the root mean square deviations (RMSD) of structural alignment between these helicases were less than 1Å, indicating a predominantly similar arrangement of functional residues. However, the positions Thr198, Arg199, Tyr200 and Thr201 showed variations among all the studied structural alignments. In addition, these residues showed specific structural arrangement, exclusive to DDX3X, which was not seen in any other helicase. Surface volume measurements (Figure 3.8B, mesh surface) at these residue positions showed the depth of DDX3X-ATP binding cavity to be more compared to other helicases (average Δvol = 58 Å).
3.F.5 Role of Functional Residues in Regulating RNA Unwinding Function of DDX3X

Apart from ATP binding site, another important DDX3X surface is the region where RNA binds. Since RNA binding region was not clearly understood for DDX3X, we carried out a structure based alignment using DDX3X structure as query (in CATH database) and identified DDX4 as a DEAD box helicase which shares >70% similarity with DDX3X. This helicase was recently co-crystallized along with bound synthetic RNA molecule (Sengoku, Nureki et al. 2006). Since the RMSD (Root Mean Square Deviation) of the RNA binding region of DDX4 and the corresponding region of DDX3X was less than 0.5 Å (Figure 3.9A), we decided to utilize the RNA binding site of DDX4 as a backbone to identify RNA binding region in DDX3X.

By homology, we identified DDX3X amino acid region spanning amino acids 300 to 370 as potential RNA binding site (Figure 3.9B). Docking of the synthetic RNA (from DDX4 structure) to this region of DDX3X using distance constraints (ArgusLab) revealed that DDX3X RNA binding region was mainly composed of an alpha helices emanating from multiple motifs (Figure 3.9B). The functional residues within 7 Å of bound RNA formed this helix and were described in Figure 3.9C. We carried out computational alanine scanning mutations at these regions and identified the binding “hot spots” among the functional residues by obtaining the binding energy change values (ΔΔG). Functional residues that showed ΔΔG >2kcal/mol after alanine substitutions were considered important for RNA-DDX3X interaction and are shown in Figure 3.9C. These residues were part of the conserved DEAD box motifs II, III, IV and V (Figure 3.2B).

In addition to binding strength between RNA and DDX3X, this region also requires certain level of conformational rearrangement and flexibility in order to accomplish the process of RNA unwinding. We decided to analyze the role of functional residues in providing flexibility to this RNA binding region of DDX3X. For this, we started by estimating the flexibility at this region, as a whole, and then
used this information to highlight important functional residues showing maximum atomic displacements.

We used NMA (Normal Mode Analyses) of DDX3X which is an approach based on the hypothesis that vibrational normal modes of a protein exhibiting the lowest frequencies (also named soft modes) describe the largest movements in a protein and are the ones functionally relevant (for details see Materials and Methods). Each normal mode represented a combination of conformational displacements achieved by amino acid residues in a protein. Normal modes were generated by MMTK package using Cα atoms of each residue as representative of the masses of whole amino acid. Out of 222 calculated DDX3X normal modes (each mode represents protein motion), we selected mode numbers 7-9, because these

**Figure 3.9: Functional residues at the RNA binding site**: A) Structural alignment of C alpha traces of DDX3X and DDX4. B) RNA binding region of DDX3X structurally alignment with DDX4 (RMSD = 0.4). C) Functionally relevant residues at DDX3X RNA binding site highlighted and represented in terms of Flexible (Yellow) and Rigid (Green) residues.
modes represented the lowest deformation energy values (low deformation energy means a mode with large rigid regions and a good chance for describing individual domain motions). To analyze displacements associated with these three modes, we calculated normalized square atomic displacement for each mode, which is the square of the displacement of each $C\alpha$ atom, normalized, so that the sum over all the DDX3X $C\alpha$ atoms equals one. Graphs for all three modes showed peaks for normalized atomic displacements corresponding to RNA binding region of DDX3X thus, indicating that this region was flexible (Figure 3.10A). Residue wise flexibility was also mapped on the DDX3X structure and was shown in the form of vector field to highlight the direction and amplitude of displacements (Figure 3.10B).
We identified the functional residues corresponding to these peaks and plotted the normalized atomic displacements of each functional residue with conservation score values (Figure 3.10C). As shown in Figure 3.9C, the binding “hot spot” functional residues were mostly rigid. As previously discussed, these residues were members of conserved DEAD box motifs II, III, IV and V. On the other hand, the functional flexible residues were found to be separate from the conserved motifs of DEAD box helicases, thus implicating presence of different sets of functional residues for providing DDX3X-RNA binding and for providing flexibility for RNA unwinding respectively.

3.6 DDX3X-CRM-1 Docking and Functional Effects of Constraints on DDX3X-CRM-1 Interaction

Another aspect of DDX3X function is its interaction with Exportin-1 (CRM-1) to accomplish efficient transport of newly formed HIV-1 RNA with the help of HIV-1 Rev protein. We wondered whether the functional residues that we found in our study could play any role in DDX3X-CRM-1 interaction (and subsequently in Rev Activity). Since the exact binding region for DDX3X-CRM-1 is not known, we started by docking DDX3X onto CRM-1 using a combination of algorithms based on conformation based matching. Before performing this analysis we checked the accuracy of these docking algorithms using known complexes (data not shown). We realized that our approach would find the interface of a known complex within 1.5 Å
of fluctuation (RMSD between experimental and docked complex was <1.5 Å). Using a combination of Geometric Hashing, Pose-Clustering matching and spherical polar Fourier (SPF) techniques (Ritchie and Kemp 2000), we generated initial starting positions for proteins compared to each other. For this, Exportin-1 PDB structure (ID: 3GB8) and DDX3X PDB structure (ID: 2I4I) were used (obtained from RCSB PDB). Pre-docking structure edit was done to add hydrogen residues, remove surface water molecules and to correct deviant bond angles and atoms (Andrusier, Nussinov et al. 2007). Molecular centroids of two protein structures were adjusted so that they fall within 30 angstrom radius wall. We used initial steric scan of 20 followed by a final search at N = 30 (other parameters were Distance range = 40, Scan Step = 0.75, Sub steps = 2) (Ritchie and Kemp 2000). All but top 25000 entries were discarded. Spatially similar orientations were grouped together and clusters were made starting from the solution with highest energy value. Out of these, top 10 clustered orientations are given in Table 3.2 and Table 3.3 in order to give an indication of the parameters considered for this analysis.

Table 3.2: Global and Local docking simulation parameters for DDX3X-CRM-1 interaction

<table>
<thead>
<tr>
<th>Solution No</th>
<th>Score</th>
<th>Area</th>
<th>ACE</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17206</td>
<td>2686.00</td>
<td>288.52</td>
<td>-1.20 1.09 2.56 4.95 160.55 -6.24</td>
</tr>
<tr>
<td>2</td>
<td>16856</td>
<td>4283.50</td>
<td>126.91</td>
<td>-235-1.19 1.07 -10.55 73.54 44.37</td>
</tr>
<tr>
<td>3</td>
<td>16300</td>
<td>2770.70</td>
<td>359.85</td>
<td>0.03 -0.93 -0.15 -77.93 132.05 57.65</td>
</tr>
<tr>
<td>4</td>
<td>16014</td>
<td>3230.80</td>
<td>144.95</td>
<td>-0.51 -1.10 -2.09 26.21 101.0 47.46</td>
</tr>
<tr>
<td>5</td>
<td>15726</td>
<td>2386.20</td>
<td>309.05</td>
<td>-0.63 0.18 1.50 39.91 121.60 5.27</td>
</tr>
<tr>
<td>6</td>
<td>15422</td>
<td>2923.00</td>
<td>440.90</td>
<td>1.46 0.10 0.08 -25.47 103.53 57.70</td>
</tr>
<tr>
<td>7</td>
<td>15130</td>
<td>2334.90</td>
<td>-22.58</td>
<td>2.53 -0.43 -1.07 -30.91 61.84 60.60</td>
</tr>
<tr>
<td>8</td>
<td>15024</td>
<td>2002.10</td>
<td>425.51</td>
<td>-0.82 0.593.0613 6177.95 -15.96</td>
</tr>
<tr>
<td>9</td>
<td>14698</td>
<td>1949.60</td>
<td>268.94</td>
<td>-2.24 -1.11 1.52 53.64 134.97 49.15</td>
</tr>
<tr>
<td>10</td>
<td>14532</td>
<td>2193.50</td>
<td>309.15</td>
<td>115 0.10 -1.72 81.04 133.72 51.83</td>
</tr>
<tr>
<td>11</td>
<td>14492</td>
<td>2248.90</td>
<td>-75.41</td>
<td>1.35 1.36 0.72 -78.72 78.75 23.03</td>
</tr>
</tbody>
</table>
We sought to refine these initial positions generated by rigid body docking algorithms by Monte Carlo minimization of the binding score function (Ritchie and Kemp 2000). The refined candidates were ranked by the binding score. This score includes Atomic Contact Energy, softened van der Waals interactions, partial electrostatics and additional estimations of the binding free energy (Vajda and Kozakov 2009). To do this, we utilized top 500 transformation files obtained by global docking approach. Refined candidates were ranked according to their energy function as discussed before and as shown in Table 3.2. Final refinement was again carried out on the best 60 structures obtained following initial local refinement to reduce the incidence of clashes (using full interface Side-chain optimization). Structures obtained following final local docking were arranged on the basis of their global energy parameters. Out of these complexes, five structures having highest global energy parameters were selected (Figure 3.11).

Structure parameters were employed to dock DDX3X to CRM-1. To select one potential complex out of five structures, we utilized an additional phylogenetic approach based on maximum likelihood analysis to highlight clusters of interaction.
“hot spots” on both DDX3X and CRM-1 surface. Then, we selected the DDX3X-CRM-1 complex (out of final five complexes) that had an interface with maximum concentration of hot spots (Figure 3.11B). Based on this, we could select structure four, as shown in Figure 7B, to be the possible DDX3X-CRM-1 complex. DDX3X region encompassing amino acid residues 420–560 formed contact with CRM-1 region with amino acids 500–580 (Figure 3.11C). Most of the potential functional residues under positive selection were present in this region (G406, F430, and H485). Next, we used the residue wise conservation scores (from Figure 3.5) and selected the functional residues in this region. These functional residues were evaluated for their role in DDX3X-CRM-1 binding by running a carbon probe of 1.4 Å (a carbon probe represents a peptide binding region) on the DDX3X part of the DDX3X-CRM-1 interface. In addition, alanine scan of the functional residues at this interface was also carried out. Finally, functional residues contributing highly towards DDX3X-CRM-1 binding (found by carbon probe-DDX3X functional residue binding energy values) and showing maximum change in binding energy (ΔG as found by alanine scanning) were identified as residues crucial for DDX3X.

Figure 3.11: Docking of DDX3X-CRM-1: (A) Top 5 docked complexes (B) Phylogenetic identification of clusters with binding hot spots. (C) Interface of selected complex with identified functional residues. (D) Binding energy values are shown in Red.
3.F.7 Selected functional residues on DDX3X

Based on overall evolution and structure based analyses described above, we could shortlist three novel functional regions on DDX3X. These regions encompassed amino acid positions 200-210, 330-340 and 408-430 respectively. The structural location of each of the functional cluster is described in Figure 3.12.

Amino acid position 200-210 on DDX3X is located in proximity to the bound nucleoside in inter-domain cleft of surface area 1650Å² (Figure 3.12A). Position 330-340 encompassed RNA binding region (Figure 3.12B), while functional region 408-430 encompassed residues distributed on both N and C terminal domains of DDX3X (Figure 3.12C).

**Figure 3.12: DDX3X functional clusters:** (A) DDX3X 200-210 (yellow surface): with ATP binding site shown in electrostatic surface representation. Bound adenosine moiety shown as stick representation. (B) DDX3X 330-340 (Red ribbon): with RNA binding site shown as green ribbon representation. Bound U6 RNA highlighted as stick representation. (C) DDX3X-CRM-1 interface (Red ribbons): Two distinct DDX3X conformations are shown (yellow ribbons for “open” and cyan ribbons for “closed” state) with respect to predicted DDX3X-CRM-1 interface
3.F.8 Identifying specificity of functional clusters towards DDX3X

Mechanistically, DEAD box helicases possess two forms: a RNA and ATP bound conformation (“closed”) in which two RecA like domains are at an average distance of 10 Å from each other. Following ATP hydrolysis, two domains move apart from each other with conformational changes in various regions representing conserved motifs (Figure 3.13A). This conformational change allows RNA unwinding and release of RNA molecule. DDX3X “open” conformation structure coordinates are available at a resolution of 2.20Å (Hogbom, Collins et al. 2007), but its RNA bound closed state is not known. In order to study the relative positions of evolutionary conserved functional regions in DDX3X “open” and “closed” state structures, we utilized the known “closed” state structure of DDX19 and DDX4 to homology model the “closed” state of DDX3X. Validation and short refinement was applied to correct small inaccuracies in the template and final model was optimized by WHAT_CHECK. The erroneous residues (i.e. B factor <0 or >100, unusual bond lengths >4σ from average) were further refined by restricted molecular dynamics (MD) techniques according to Berendsen et al (Berendsen and Hayward 2000). To test the validity of this method, DDX19 closed conformation was used as template to model known closed state of DDX4 and vice versa. Structural alignment indicated similar architecture for modeled and crystal structures for DDX4 and DDX19 with root mean squared deviation (RMSD) as low as 0.7Å (Figure 3.13B and 3.13C).
Figure 3.13: DDX3X “open” and “closed” state structures: (A) Schematic representation of conformational states of DEAD box helicases (B) Structural alignment of “closed” and “open” state structures of DEAD box helicases (C) DDX3X “closed” and “open” state structures.

In order to structurally assess the specificity of identified DDX3X functional regions, we carried out comparison of DDX3X functional regions with 32 representative DEAD box helicase members belonging to six established groups of DEAD box helicases. These members are represented in Table 3.4.

Table 3.4: Classification of DEAD box helicases

<table>
<thead>
<tr>
<th>Group</th>
<th>Member</th>
<th>Group</th>
<th>Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>eIF4A</td>
<td>DDX2A</td>
<td>Has1</td>
<td>DDX18</td>
</tr>
<tr>
<td></td>
<td>DDX48</td>
<td></td>
<td>DDX10</td>
</tr>
<tr>
<td></td>
<td>DDX6</td>
<td></td>
<td>DDX55</td>
</tr>
<tr>
<td></td>
<td>DDX19</td>
<td></td>
<td>DDX31</td>
</tr>
<tr>
<td></td>
<td>DDX25</td>
<td></td>
<td>DDX47</td>
</tr>
<tr>
<td></td>
<td>DDX39</td>
<td></td>
<td>DDX49</td>
</tr>
<tr>
<td></td>
<td>UAP56</td>
<td></td>
<td>DDX27</td>
</tr>
<tr>
<td></td>
<td>DDX20</td>
<td>Rrp3</td>
<td>DDX54</td>
</tr>
<tr>
<td>DDX3X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DDX56</td>
</tr>
</tbody>
</table>
For structural comparison, crystal coordinates of DDX20 (2OXC), DDX10 (2PL3), DDX41 (2P6N), DDX3X (2I4I), DDX19 (3EWS and 3G0H), DDX47 (3BER), DDX5 (3FE2), DDX18 (3LY5), DDX52 (3DKP), DDX20 (3B7G), DDX53 (3IUY), DDX25 (2RB4), DDX4 (2DB3) and eIF4 (2G9N) were obtained after refinement, while other remaining structures were generated by homology modeling in SWISS MODEL workspace (Schwede, Kopp et al. 2003). As shown in Figure 3.14A, each of the identified functionally conserved DDX3X regions (viz. 200-210, 330-340 and 408-430) showed high sequence specificity when compared against representative sequences of six DEAD box helicase group members (Figure 3.14A). The process was validated by benchmarking against known crystal structures as described in Methods.

DDX3X RNA unbound state (Figure 3.13C) was obtained from crystal coordinates of DDX3X structure with PDB ID 2I4I, while RNA bound state was homology modeled using DDX4 and DDX19 RNA bound states as templates (Figure 3.13C). Finally, a total of 64 structures (both homology modeled and crystal coordinates) representing 32 RNA bound and 32 RNA unbound states were compared with DDX3X RNA bound and unbound states and the Root Mean Squared Deviations (RMSD) of each of the three functional regions (DDX3X 200-210, DDX3X 330-340 and DDX3X 408-430) was computed.
Figure 3.14: Specificity of DDX3X functional regions: (A) Sequence alignment of DDX3X with 6 representative sequences of DEAD box helicases belonging to distinct groups. More hydrophobic sequences are shown in green. Hydrophilic region are blue. Bars represent (from top to bottom) sequence conservation, quality of sequence information and consensus sequence for each of the functional cluster respectively. (B) Atomic details of each of the DDX3X functional region is shown with potential function described in the context of structural arrangement and interactions with adenosine (for DDX3X 200-210) or RNA (for DDX3X 330-340). Graphs represent RMSD values of comparison of each of the DDX3X functional region with similar region of 26 other DEAD box helicases (representing 6 groups of DEAD box helicases family). Structures for DEAD box helicase “open” and “closed” forms were obtained from RCSB PDB or were generated by homology modeling.

Structural alignment of DDX3X with each of the representative member revealed existence of marked deviation of DDX3X functional regions. On an average, DDX3X (200-210) depicted an RMSD (Root Mean Squared Deviation) of ~0.9-1.1Å with most of the other DEAD box helicases (Figure 3.14B). Expectedly, intra group comparison of DDX3X (among members belonging to p68 group, of which DDX3X is also a member) revealed least deviation pattern with RMSD clustered at around 0.75Å (Figure 3.14B). On the other hand, DDX3X(330-340) showed an average RMSD of ~1.65Å. Interestingly both DDX3X (200-210) and DDX3X(330-340) yielded 30-35% higher deviations when DDX3X RNA unbound structures were compared with RNA unbound structures of other 28 DEAD box helicases. DDX3X RNA bound state revealed an average RMSD = 1.78Å for DDX3X (408-430) region. However, similar region showed significant structural uniformity among other DEAD box helicases (RMSD = 1.234Å), thus establishing existence of specific architecture.
of this region in DDX3X. Overall, this analysis described specificity of DDX3X functional regions at amino acid positions 200-210, 330-340 and 408-430.

3.F.9 Role of DDX3X Functional Clusters in Regulating Conformational Transition

In order to understand the role of functional clusters in defining “open” state of DDX3X, we homology modeled DDX3X Δ(200-210) and DDX3X Δ(330-340) based on DDX3X crystal structure coordinates (obtained following conjugate gradient minimization of DDX3X PDB: 2I4I structures) (Hogbom, Collins et al. 2007). A 300ps molecular dynamics simulation for “open” RNA unbound states of DDX3X WT, DDX3X Δ(200-210) or DDX3X Δ(330-340) revealed that the MD average structures were significantly similar to the original structure with all the secondary structure elements of original structure remaining intact indicative of minimum effect of deletions on the “open” conformation of DDX3X (1.046Å > RMSD > 0.7653Å) (Figure 3.15A). The coordination pattern of bound nucleotide is shown in Figure 3.15B and is in good agreement with the X-ray structure, validating our modeling procedure. RMSD for last 50 frames of 300ps molecular dynamics simulation showed similar pattern for both wild type as well as DDX3X mutants. The pattern of deviation of Cα atoms was similar for wild type as well as mutant structures with mutants depicting ~60% greater structural flexibility.

Knowing that DDX3X “open” state remained relatively unchanged following deletions of functional clusters, we performed three additional 300ps MD simulations corresponding to “closed” state of DDX3X/DDX3X mutants, so as to comprehend the outcome of reduced protein flexibility in DDX3X Δ(200-210) and DDX3X Δ(330-340). Thermodynamic properties such as temperature and energies converged to stable values in all the simulations following an equilibration of 50ps. DDX3Xwt was relatively stable during the simulation with average RMSD hovering around 0.4Å compared to 2-2.2Å for mutant DDX3X simulations. Secondary structure elements were subsequently analyzed for last 50ps of MD simulations. As shown in Figure
3.15C, DDX3XΔ(200-210) lacked the conserved motif I α helix for the last 50ps of MD simulations, indicating a destabilizing effect of deleting ATPase functional cluster. On the other hand, the ATP binding motif I of wild type DDX3X reflected intact nucleotide binding shell reminiscent of “open” state of DDX3X (Figure 3.15C). Superposition of MD simulated ATPase region from wild type DDX3X with available crystal coordinate of DDX3X “open” state measured an average RMSD of 0.56Å compared to 0.986Å<RMSD <1.542Å for MD simulated DDX3X Δ(200-210). Similarly, MD simulation for DDX3X Δ(330-340) showed loss of association between motifs Ia and motif IV. Overall, RNA binding functional cluster depicted abnormal residue displacements with >1.8Å difference between wild type DDX3X and DDX3XΔ(330-340). Individual Root Mean Squared Fluctuations (RMSF) between stable N-3 of adenosine moiety and Ca of DDX3XΔ(200-210) and between 2'-OH of bound U6 RNA and DDX3XΔ(330-340) echoed the observed differences in secondary structure elements and depicted higher fluctuations for mutant DDX3X variants, thus explaining loss of original structural arrangements.
Figure 3.15: Comparison between wild type and mutant DDX3X “open” and “closed” states: (A) Superposition of MD simulated final structures for DDX3X WT (Red), DDX3X ΔATPase (Green) and DDX3X Δhelicase (Cyan) in the “open” state showing relatively similar arrangement of secondary structure elements. (B) Coordination pattern of bound adenosine moiety representative of DDX3X WT and mutants in DDX3X “open” state, indicative of validity of simulation procedure. (C) 50ps energy minimization followed by 300ps MD simulation for WT, ATPase and helicase mutants of DDX3X. Left: Step I represents 50 pico seconds energy minimization of DDX3X WT and DDX3X ΔATPase. Green ribbon: DDX3X ATPase region at stage 0 of energy minimization; Red ribbon: DDX3X ATPase at 50ps of energy minimization. Cyan ribbons represent arrangement at the DDX3X ATPase site after 300ps of MD simulation with difference between WT and ΔATPase clearly observable. Right: Similar energy minimization and MD simulation for DDX3X Δhelicase region and DDX3X WT. All these structures and procedures were validated by standard methods and are described in text.

3.F.10 Pharmacophore Design Based on Functional Regions

The AMP analogs were shown by various groups as efficient blockers of DDX3X ATPase activity (and HIV-1 replication) (Yedavalli, Zhang et al. 2008). We carried out a multiple flexible alignment of all the AMP analogs in order to analyze their stability and selectivity, with respect to the constraints that we identified in Figure 3.6 and Figure 3.7. The aligned analogs were employed for the generation of a pharmacophore (i.e. corresponding binding site on DDX3X). The pharmacophore, as shown in Figure 3.16A revealed that these analogs mainly utilize the functional residues which are under structural constraints in all the DEAD box helicases. Because of this, these analogs can bind to most of the helicases and, thus possess minimum selectivity. In fact, we docked the pivot molecule (AMMPNP which is a nonhydrolyzable analog) to different DEAD box helicases and found that, with small variations in ligand-DDX3X binding energy values, the ligand was capable of binding all helicases. We thought that to improve the DDX3X specific selectivity of analogs, residue positions with high atomic contributions towards ATP binding (Figure 3.7 and 3.8) can be combined with DDX3X specific functional residues that were shown in Figure 3.3. We used residue positions 227 to 231 and residue positions 198 to 201 to design a novel pharmacophore specific to DDX3X (Figure 3.16B). The residues 227-231 could serve to provide ample binding strength, because of their maximum contributions towards ATP-DDX3X binding while residues 198 to 201 could provide selectivity towards DDX3X.
Comparing the DDX3X RNA binding site from different DEAD box helicases (using approach similar to Figure 3.16B) showed that this region exhibits better member specific selectivity than the ATP binding region. We thought of utilizing the flexible and rigid residues (from Figure 3.9 and 3.10) to identify a potential site which can be used as a target for a blocker molecule. To do that, the RNA binding region was examined for the presence of cavities on the DDX3X surface. A cavity encompassing regions 330 to 345 of RNA binding region was found to include both flexible and rigid flexible residues. This cavity was selected and the nature of residues in this cavity was analyzed by constructing a map of charge distribution, calculated using ArgusLab as shown in Figure 3.16C. Important electron donors and acceptors belonging to crucial functional residues were highlighted.

**Figure 3.16: Identifying DDX3X specific regions as a pharmacophore:** (A) Binding surface of AMP analogs. Pharmacophore for ATP binding site (B) and RNA binding site (C) based on the evolutionary conserved functional residues.
3.G. DISCUSSION

In this chapter, we identified critical functional residues regulating ATPase and RNA unwinding function of DDX3X (Figures 3.3, 3.4, 3.5, 3.6) and predicted DDX3X-CRM-1 interaction region (Figure 3.7). We believe this information could provide a new and detailed glimpse of the mechanism by which complex interplay of DDX3X-HIV-1 Rev-CRM-1 takes place and leads to efficient viral replication. Using the functional residue information at the DDX3X ATPase site, we showed that most of these residues are under strong functional constraints and that validates their conservation pattern (Figure 3.6). Using computation based biochemical approaches; we then identified the cause for these constraints as well (Figure 3.7). This information helped us in comparing the ATP binding sites of different DEAD box helicases (Figure 3.8). Most of the functional residues were common to all the sites with little structural variations (Figure 3.16). But, certain ATP proximal functional residues showed remarkable specificity for DDX3X. We believe that these regions are responsible for providing different levels of ATPase activities to different DEAD box helicases. The binding energy for ATP binding regions of different DEAD box helicases, when compared, showed substantial variations, mainly attributable to amino acid positions 198 to 201 (the DDX3X specific region). Using ligand based pharmacophore modeling; we showed that most of the available AMP analogs bind residues which were common to all DEAD box helicases (Figure 3.16). In fact, docking of these analogs to different DEAD box helicases revealed that they were capable of binding to all of these members with minor variations in binding energy values. We utilized the functional residue information to select region on DDX3X that could be employed for generation of DDX3X specific analog. Surface topology of the selected region showed that this surface had specific conformation specific to DDX3X. Comparing this surface for other helicases revealed a predominant two kind of topologies (Figure 3.16) which were different for DDX3X. So, combining surface and electron donor acceptor profile of the selected region at ATP binding site, we identified a pharmacophore that can be potentially employed to design DDX3X specific ATP analog (Figure 3.16).
As found by docking simulations, the binding of DDX3X to CRM-1 positions DDX3X in such a way that its helicase domain was found to come in proximity of HIV-1 Rev, which binds CRM-1 (Figure 3.10). It indicated that HIV-1 Rev transfers the viral RNA to DDX3X for unwinding during the process of nuclear transport. Also the flexible and rigid functional residues at the RNA binding region indicate their role in RNA unwinding process (Figure 3.9). The identification of functional residues at this site showed that the flexible and rigid functional residues constitute separate clusters and the highly conserved motifs of the DDX3X encompassed a predominance of rigid residues. This indicates that residues required for maintaining DDX3X dynamics are different from residues required for efficient RNA-DDX3X binding. This information can help in understanding the mechanism of HIV-1 RNA unwinding by DDX3X in more detail. Further, we showed that RNA binding region showed better selectivity and specificity towards DDX3 as compared to the ATP binding region by comparing the structural topology of different DEAD box helicases. We believe that this region can be utilized to generate better selective inhibitors for potential block in HIV-1 replication (Figure 3.16). Analysis of functional residues is crucial for characterization of domains present in a protein. Functional residue information of DDX3X (Figure 3.4) would provide more information about the mechanism by which this helicase acts and about the selectivity that differentiates DDX3X from rest of the helicases of the family (Figure D1). Direct role of DDX3X in HIV-1 Rev mediated nuclear export of unspliced viral RNA further makes it important to understand the functional sites of DDX3X.
Development of potential ligand molecules that block specific DDX3X function could help in achieving the aim of designing ideal intervention against HIV-1 replication.