Chapter 3

Shape, Flexibility and Packing of Proteins and Nucleic Acids in Complexes

3.1 Introduction

Protein-protein complexes and protein-nucleic acid complexes form biological assemblies, characterized by highly specific shape complementarity. Changes in shape due to complex formation may modulate various molecular interactions which underpin many biophysical processes like cell cycle regulation, signal transduction, gene expression and molecular recognition [92–99]. Different studies reveal that complex formation of proteins with other proteins and/or nucleic acids is accompanied by changes in the internal configuration of atoms in a molecule or large scale conformational changes [100–102]. These induced structural changes may affect the internal dynamics of a protein by exposing a catalytic site in an enzyme or a binding site in a regulatory protein [103,104]. The variations in shape reflects the extent to which the different macromolecular interactions at the molecular surface of the complexes affect
3.1 Introduction

the structure, packing and flexibility of proteins with spherical or aspherical shapes, which consequently reflects their specific biological functions.

The shape of any object have been determined earlier from the ensemble averages of the various combinations of different Cartesian components of the radius of gyration tensor ($R_G$), where the values of the upper and lower bounds correspond to a highly symmetric object \([47, 48, 105]\). A spherically symmetric object has all components of $R_G$ equal. For objects with random shapes, the values of $R_G$ are computed for each member of the set and then averaged over all the members of the set, ensuring that the orientational averaging preserves their average intrinsic spatial anisotropy. Uncomplexed globular proteins have densely packed structures inspite of numerous voids and cavities present in protein interiors and the shapes of these maximally compact structures are expected to be spherical. However, a calculation of the shape parameters reveals a considerable degree of spatial anisotropy in globular proteins and RNA molecules with compact folded states \([57, 88]\). The shape analysis of semiflexible molecules like nonglobular proteins and DNA shows that most of them have aspherical prolate structures; the short DNA molecules behave as stiff elastic rods, while the longer DNA molecules and nonglobular proteins have typically extended structures resembling a real polymer with excluded volume interactions and obeys Flory’s scaling law for a self-avoiding walk \([106]\).

The geometry and composition of interfaces in the complexes of proteins and nucleic acids have been intensively studied \([107, 108]\) by comparing various physico-chemical properties, such as solvation, accessible surface area, shape and occurrence of different types of contacts resulting from different interactions \([109, 110]\). Protein-protein complexes are primarily characterized by hydrophobic \([111]\) and electrostatic interactions \([112]\) widely used in protein-protein docking for exploring the possible binding modes between two proteins \([56]\). The shape of proteins/nucleic acids often provide a fundamental insight in the properties of their respective structure and association \([113]\). Many computational docking approaches utilize the binding-related anisotropic shape of the proteins and nucleic acids in their complexed states as a guide.
to search the position of binding sites in proteins and predict the structures of these complexes. These studies clearly demonstrate the existence of a prominent binding-related anisotropic shape characteristic of larger proteins which facilitates protein-protein recognition. However, there has been no quantitative assessment of the shape complementarity in terms of the global shape determinants when proteins/nucleic acids undergo conformational changes from the unbound to the bound complexed structures.

The compatibility between the atomic interactions between the amino acids of a protein and the nucleotides of DNA plays a major role in dictating specific recognition. This compatibility may be determined by the extent of change in the physical and topological properties like the shape, flexibility and packing density of the protein-nucleic acid complex relative to that of the protein or nucleic acid. Recognition of a specific base sequence by a DNA binding protein involves the unwinding of the DNA to form hydrogen bonds with specific amino acids, mainly in the major grooves [114]. In many complexes, DNA may bend optimizing the protein-DNA interface [115], while in others, DNA undergoes a variety of conformational changes as in the opening of the minor groove of protein-DNA complex [116]. The sequence dependence of the minor groove width of DNA induce variations in the electrostatic potential that creates specific binding sites for positively charged amino acids within the minor groove [117]. Other studies of protein-DNA complexes highlights the significance of DNA base-amino acid type correlation in terms of electrostatic complementarity and indicates that the variations in the shapes of DNA observed in protein-DNA complexes is strongly correlated with the conformational preferences of DNA [118]. The quantitative characterization of the shape complementarity of the protein-DNA contacts may offer valuable insights in understanding how proteins recognize DNA sequences and lead to the development of improved prediction algorithms.

This work aims to investigate the variation in shapes of proteins and nucleic acids in different hetero-complexes of protein-protein, protein-DNA and protein-RNA by evaluating the different shape descriptors. The rapidly expanding database of the
3.1 Introduction

Solved crystal structures of the protein-protein and protein-nucleic acid complexes provide a unique resource for studying the structural and functional characteristics of the interactions of proteins and nucleic acids. The three-dimensional coordinates of all proteins, DNA and RNA are obtained from the July 2009 list of the Protein Data Bank (PDB) and April 2009 list of the Nucleic Acid Data Bank (NDB). The intrinsic flexibility of these complexes, which influences the nature of the interactions among proteins, DNA and RNA are measured in terms of the persistence length ($l_p$). The packing density of proteins is calculated by using Voronoia program [119], based on an improved Voronoi cell algorithm using hyperboloid interfaces to construct atomic volume.

Analysis of the number and type of two-body contacts determines the packing density of the protein and the hydrogen bonding between proteins in different complexes, thus imparting a definite shape to the protein-protein complex. Rather than investigating amino acid and base specific interactions in individual or highly related protein-nucleic acid complexes, we formulate simple geometric criteria for a generic recognition code that adequately explains observations for all proteins and nucleic acids. In this study we select well-defined three-dimensional X-ray crystallographic structures to calculate the different geometric properties of proteins and nucleic acids compared to those in complexes. Relative to the globular proteins, the intrinsically disordered proteins lack precisely defined structure and exist instead as dynamic ensembles of interconverting structures. Investigation of nucleic acid induced/assisted protein folding is beyond the scope of the present chapter.

The chapter is organized as follows: Section 3.2 describes the methodology for calculating the size, shape parameters, persistence length and the packing density. Section 3.3 presents and discusses the results of our calculations. Section 3.4 briefly summarizes the conclusions of our study.
3.2 Method

The shape of protein-protein and protein-nucleic acid (protein-DNA and protein-RNA) complexes may be determined from the atomic coordinates of the currently available three dimensional structures analyzed by NMR, X-ray crystallography, fiber diffraction or cryo-electron microscopy [102,120–125].

The non-redundant database of the protein-protein and protein-nucleic acid complexes consists of 2396 protein-protein complexes, 1466 protein-DNA complexes and 477 protein-RNA complexes. This database comprises of the protein-protein complexes and their corresponding uncomplexed counterparts from the protein docking benchmark sets [126,127], which is a subset of the selected database of proteins compiled from the Protein Data Bank (PDB) and Nucleic Acid Data Bank (NDB). The selection criteria required all X-ray crystallographic structures of protein-protein and protein-nucleic acid complexes, to have a resolution less than 3.0 Å and R factor less than 0.35. The data set of protein-protein complexes contains 738 single-single-chain and 1658 multi-chain protein complexes, comprising a total of 1476 single-chain proteins and 4978 multi-chain proteins of size, \( N > 30 \) residues. The protein-DNA database comprises of 1013 single-chain proteins, 1206 multi-chain proteins of size \( N > 30 \) and 3299 DNA with chain length greater than 5 nucleotides. The data set of 477 protein-RNA complexes consists of 1225 single-chain proteins, 385 multi-chain proteins and 590 RNA with chain length greater than 10 nucleotides. The characteristic shape of single-single-chain proteins in complexes are expected to be rather different from the multi-chain ones due to the presence of interchain interactions which affects the overall structure of the complexed protein.

The size of proteins, DNAs and RNAs are usually measured by mean radius of gyration \( \langle R_G \rangle \), which may be determined from the coordinates of their respective three dimensional structures available from either PDB or NDB. Small angle neutron scattering or X-ray scattering experiments may be used to determine \( R_G^2 \), which may be calculated from the equation,
\[ \langle R_G^2 \rangle = \frac{1}{2M^2} \sum_{i}^{N} \sum_{j}^{N} m_i m_j (r_i - r_j)^2 \]  

(3.1)

where \( N \) is the number of residues, \( m_i \) and \( r_i \) are the mass and position of the \( i^{th} \) residue and \( M \) is the total mass of protein or nucleic acid. This equation is obtained from the definition of \( R_G \), by substituting the center of mass coordinates in terms of the residue distance coordinates. The shape of any given conformation may be calculated from the different components of the radius of gyration, \( R_G \). Among a number of shape descriptors based on \( R_G \), we use two rotationally invariant quantities, the asphericity parameter (\( \Delta \)) and shape parameter (\( S \)) which may be evaluated from the inertial tensor [57,64,65]

\[ T_{(\alpha,\beta)} = \frac{1}{2M^2} \sum_{i}^{N} \sum_{j}^{N} m_i m_j (r_i\alpha - r_j\alpha) (r_i\beta - r_j\beta) \]  

(3.2)

where \( r_{i\alpha} \) is the \( \alpha^{th} \) component of the position of the \( i^{th} \) residue and \( \alpha, \beta = x, y, z \). The squares of the three principal radii of gyration, \( R_G \) are calculated from the trace of the matrix \( T_{(\alpha,\beta)} \), which amounts to the sum of its eigenvalues, \( \lambda_1, \lambda_2 \) and \( \lambda_3 \).

\[ R_G^2 = trT = \sum_{i=1}^{3} \lambda_i \]  

(3.3)

These eigenvalues are used to calculate the asphericity parameter, \( \Delta \), which measures the average deviation of the chain conformation from spherical symmetry. In three dimensions, the extent of asphericity is expressed as

\[ \Delta = \frac{3}{2} \frac{\sum_{i=1}^{3} (\lambda_i - \lambda)^2}{(trT)^2} \]  

(3.4)

where

\[ \lambda = \frac{\sum_{i=1}^{3} \lambda_i}{3} \]  

(3.5)

The lowest value of \( \Delta \) is zero (the value for a sphere) and the highest value is one (the value for a rod).
3.2 Method

The shape parameter ($S$) measures the ellipticity of the chain conformation by characterizing it as either prolate or oblate. In three dimensions, $S$ may be calculated from

$$S = 27 \prod_{i=1}^{3} (\lambda_i - \lambda) (\text{tr}T)^3$$

(3.6)

The value of $S$ is bounded, following the inequality $-1/4 \leq S \leq 2$. $S = 0$ implies that the molecule is completely symmetric i.e. spherical. Positive value of $S$ correspond to prolate ellipsoids, i.e. tending towards the rod shape and negative values indicate oblate ellipsoids, i.e. tending towards disk shapes. $\Delta$ and $S$ provide a quantitative estimate of shapes using the atomic coordinates of the three dimensional structures of protein-protein and protein-nucleic acid complexes. There are no tacit assumptions about the extent of packing and steric overlaps which introduce a finite degree of arbitrariness [5].

Persistence length ($l_p$) is a direct measure of the intrinsic elasticity/flexibility of the polypeptide or nucleic acids, which modulates the interactions between different proteins or proteins and nucleic acids. It is the typical length scale over which thermal fluctuations are correlated. Persistence length may be measured by various experimental techniques like electron microscopy, light scattering, linear flow birefringence and dichroism [18, 66]. The worm-like chain model [16] (WLC) is used for describing the elastic and statistical properties of proteins and nucleic acids theoretically, especially the double stranded DNA. The theory requires two input parameters: i) contour length ($L$), ii) persistence length ($l_p$). The persistence length ($l_p$) may be estimated either from i) the end-to-end distance probability distribution function or ii) the inter-residue distance distribution in the limit $r >> R_G$. The mean-field approximated distribution function for the end-to-end distance for WLC model is given by [67]

$$P (R_e) = \frac{1}{4\pi R_e^2 N_c} \exp \left( \frac{-3L}{4l_p \left( 1 - (R_e/L)^2 \right)} \right)$$

(3.7)

where $N_c$ is the normalization constant, $L$ and $l_p$ are the contour length and
3.2 Method

The persistence length of the chain respectively. $P(R_e)$ may be experimentally determined from single molecule fluorescence resonance energy transfer (FRET) measurements. For small $l_p$, $P(R_e)$ is Gaussian, whereas for large $l_p$, $P(R_e)$ approaches the rigid rod limit and exhibits a non-Gaussian behavior. For molecules of large contour length ($L$), the end-to-end distance may be approximated by $\langle R_e^2 \rangle \approx \langle R_G^2 \rangle \approx Ll_p$. The inter-monomer distance distribution function $P(r)$ for $r >> R_G$ follows an exponential decay given by

$$P(r) = \beta \exp \left( -\frac{1}{1-x^2} \right)$$  \hspace{1cm} (3.8)

where $\beta$ is an arbitrary constant and $x = l_pr/R_G^2$. The persistence length, $l_p$ may be calculated by fitting $P(r)$ to the above equation.

For maximally extended chain conformations, $\langle R_e^2 \rangle \approx L^2$, the inter-residue distance distribution function $P(r)$ is different from the end-to-end distribution function $P(R_e)$. The contour length ($L$) may be calculated from the maximum inter-residue distance, $r_{max}$ [42, 68]. In this case, the persistence length $l_p$ may be evaluated from the unperturbed radius of gyration for the worm-like chain (WLC), according to Benoit and Doty [46]. Radius of gyration is given by

$$\frac{\langle R_G^2 \rangle}{l_p^2} = \frac{L}{3l_p} - 1 + 2l_p \frac{L}{L} - 2 \left( \frac{l_p}{L} \right)^2 \left( 1 - \exp \left( -\frac{L}{l_p} \right) \right)$$ \hspace{1cm} (3.9)

This expression holds true for large values of $l_p$ or when the $L/l_p$ ratio is small. $l_p$ is calculated by expanding the exponential up to the sixth order term. Simplification of this expression yields

$$\left[ \frac{1}{12L} - \frac{\langle R_G^2 \rangle}{L^3} \right] l_p^2 - \frac{l_p}{60} + \frac{L}{360} = 0$$ \hspace{1cm} (3.10)

which yields two values of $l_p$, equal in magnitude but opposite in sign. The negative value of $l_p$ is ignored as it represents an unphysical value.

Packing of protein structures is an important determinant of their stability, function and flexibility. The packing of the proteins in the different complexes is calculated by using Voronoia, a database for the analysis of atomic-scale packing density,
which is based on an improved Voronoi Cell algorithm [119]. It uses hyperboloid interfaces instead of the planar ones, to access the atomic volumes and resolves the solvent-accessible and inaccessible regions of the residues in the protein. The extent of packing in proteins in complexed states may be evaluated from the number and nature of two body contacts. The amino acids are classified into nine groups based on their charge and van der Waals radius [4]. The heavy atoms of side-chain of each amino acid are considered for calculating the $s - s$ contacts. side-chain-side-chain contact occurs when at least a pair of heavy atoms from the two residues in each chain are separated by a distance $d \leq 5.2$ Å. Backbone-backbone contacts ($b - b$) are facilitated when the distance between two backbones is $\leq 6.5$ Å, where the backbones are represented by C-α atoms. For $s - b$ contacts, the minimum distance between the backbone and side-chain residues should be $\leq 5.5$ Å [57]. The $s - b$ contacts optimize interactions due to side-chain packing and hence stabilizes the protein conformations [69].

The difference in interactions between the protein-protein complexes and the protein-nucleic acid complexes is due to the hydrogen bonding in the complexation sites [128]. The hydrogen bonds are formed between nitrogen and oxygen where nitrogen mostly acts as the donor and rarely as an acceptor. For calculating the hydrogen bonds involved in the interaction sites, the minimum and maximum distance between the donor and the acceptor is chosen between 2.4 Å to 3.5 Å while the distance between the hydrogen and the acceptor lie between 1.5 Å and 2.5 Å [129]. The position of hydrogen is not reported in most crystal structures. The missing $H$ atoms are incorporated in all protein structure files by using the AMBER software [130]. We calculated all inter-residue contacts in protein complexes with distance less than 3.5 Å. The hydrophobicity of the interacting residues in complexes is also taken in account to classify the difference in interactions. Finally, we also calculate the contact density of the proteins to measure the extent of change compared to the uncomplexed proteins. The contact density is simply the number of non-hydrogen atoms surrounding the C-α atom within the mean cutoff radius $R_c$. For proteins $R_c$ is chosen to be
3.3 Results and Discussions

3.3.1 Number Distribution

The number distributions of biomolecules in the different types of complexes are illustrated in Fig. 3.1. The protein-protein complexes are subdivided into the single-single-chain protein complexes and multi-chain protein complexes. There are 1476

![Figure 3.1 Number distribution of proteins/nucleic acids as a function of chain length $N$ for (a) proteins ($a_1$ and $a_2$ for single-chain proteins and multi-chain proteins) in protein-protein complexes, (b) proteins and DNA ($b_1$, $b_2$, and $b_3$ for single-chain proteins, multi-chain proteins and DNA respectively) in protein-DNA complexes and (c) protein and RNA ($c_1$, $c_2$ and $c_3$ for single-chain proteins, multi-chain proteins and RNA respectively) in protein-RNA complexes.](image)

7.3 Å as it marginally deflects from the mean value [131]. To exclude the covalent neighbours the lower cutoff of 3.5 Å is introduced. For uncomplexed proteins, the mean contact density is 67.5.
single-chain proteins shown in Fig. 3.1(a), where the chain length varies from 30 to 1476 residues and approximately 83% of the proteins range between 30 to 300 residues. Fig. 3.1(a2) depicts the chain length distribution for 4978 multi-chain proteins with the chain length ranging from 30 to 7860 residues, of which the chain length of 61.7% proteins lie between 30 to 500 residues and 24.6% proteins range between 500 to 1000 residues.

For the protein-DNA complexes, the number distributions for 1013 single-chain proteins, 1206 multi-chain proteins and 3299 DNA are indicated in Fig. 3.1(b1), Fig. 3.1(b2), Fig. 3.1(b3) respectively. For the single-chain proteins, the chain length varies from 30 to 2103 residues. The number of amino acid residues of 53.6% proteins vary between 30 to 300 and 40.1% proteins are in the range of 300 to 600 residues. For the multi-chain proteins, $N$ varies from 30 to 4480 out of which 62.02% proteins lie in the range between 30 to 500 residues. For DNA, the chain length varies from 6 to 147 residues with 80.3% DNA ranging from 6 to 20 residues.

Protein-RNA complexes comprise of 1225 single-chain proteins, 385 multi-chain proteins and 590 RNA. The number distributions are shown in Fig. 3.1(c1), Fig. 3.1(c2), Fig. 3.1(c3) respectively. The chain length for single-chain proteins varies from 32 to 1454 with 56.8% proteins lying between 30 to 200 residues and 25.3% vary from 300 to 500 residues. For the multi-chain proteins, the number distribution is relatively broad ranging from 30 to 3024 residues, with 85.12% proteins ranging between 30 to 1000 residues. For RNA the width of the distribution is narrow varying from 10 to 362 residues, 81.2% of which are of 10 to 80 residues.

### 3.3.2 Radius of Gyration

The radius of gyration describes the overall size of the proteins and nucleic acids in their respective complexes. The value of $\langle R_G^2 \rangle$ may be either measured by using small angle X-ray or neutron scattering or calculated from Eqn. 3.1. The size of the protein molecules in protein-protein and protein-DNA complexes follows Flory’s scaling law given by $\langle R_G^2 \rangle^{1/2} = (3.07 \pm 0.03)N^{(0.33 \pm 0.006)}$ and $\langle R_G^2 \rangle^{1/2} = (3.33 \pm 0.03)N^{(0.33 \pm 0.006)}$
3.3 Results and Discussions

Figure 3.2 Radius of gyration as a function of $N$. The straight line (red) is the best fit line for the data set of single-chain proteins in (a) protein-protein, (b) protein-DNA, (c) protein-RNA complexes, (d) DNA, (e) RNA in complexes and (f) uncomplexed proteins with the correlation-coefficients of 0.81, 0.87, 0.89, 0.91, 0.83 and 0.79 respectively.

as depicted in Fig. 3.2(a) and Fig. 3.2(b) respectively. The numerical value of the Flory’s exponent $\nu = 0.33$ implies that the proteins are compact spherical molecules, exhibiting the collapsed walk behavior similar to a polymer chain in bad solvent. This is rather similar to the scaling relation in single-chain uncomplexed proteins, shown in Fig. 3.2(f), where $\langle R_G^2 \rangle^{1/2} = (2.9 \pm 0.01)N^{(0.34 \pm 0.02)}$. The results obtained from the analysis of shape parameters and packing density also validates the above observation.

The radius of gyration $R_G$ of proteins in protein-RNA complexes varies as $\langle R_G^2 \rangle^{1/2} = (2.69 \pm 0.03)N^{(0.37 \pm 0.005)}$ as shown in Fig. 3.2(c). For RNA molecules in protein-RNA complexes the size scales according to $\langle R_G^2 \rangle^{1/2} = (4.25 \pm 0.04)N^{(0.39 \pm 0.01)}$ as indicated in Fig. 3.2(e). The Flory’s exponents for proteins and RNA, $\nu = 0.37$ and $\nu = 0.39$, closely resemble the value $\nu = 2/5$ for proteins in physiological condition
representing an intermediate scaling regime [132]. This clearly shows that proteins and RNA in the protein-RNA complexes are more compact compared to polymers in the good solvent but not as much as polymers in a poor solvent. Proteins are usually less compact than collapsed polymer chains in bad solvents as they are not very well-packed because of the cavities which arise as a consequence of regular secondary structures in folded proteins [133], even though the core of the protein may be compact because of the stabilizing hydrophobic interactions [134]. Uncomplexed single-chain proteins and RNA are compact molecules following the scaling relation $\langle R_G^2 \rangle^{1/2} \sim N^{0.33}$. Complexation is usually marked by an increase in size for both proteins and RNA. In protein-DNA complexes, DNA molecules are maximally expanded and represent typically non-compact structures. The size varies as a self-avoiding walk (SAW) according to $\langle R_G^2 \rangle^{1/2} = (2.81 \pm 0.01)N^{(0.63 \pm 0.005)}$ as shown in Fig. 3.2(d).

The plots of $R_G$ vs $N$ as depicted in Fig. 3.2 for the protein-protein and protein-nucleic acid complexes reveal that the compactness of proteins remains the same in protein-protein and protein-DNA complexes when compared to single-chain uncomplexed globular proteins. For RNA and proteins in protein-RNA complexes, the compactness decreases due to complexation as compared to the uncomplexed RNA or single-chain proteins, which closely resembles the natively unfolded proteins. Smaller DNA are less compact than their uncomplexed form (rod like behavior) while the longer ones show similar compactness.

### 3.3.3 Shape Parameters

**Asphericity**

The distribution of the asphericity parameter for proteins in protein-protein, protein-DNA and protein-RNA complexes is illustrated in Fig. 3.3. Fig. 3.3(a) shows that the $\Delta$ value varies from 0 to 0.1 for 75.8% proteins while it ranges between 0.1 and 0.3 for 18% proteins in protein-protein complexes with an extended tail in the regime $\Delta < 0.3$. Values of $\Delta$ for 80.56% of uncomplexed single-chain proteins lie between 0 to 0.1 and that of 16.69% proteins between 0.1 and 0.3. Similar behavior is exhibited
3.3 Results and Discussions

Figure 3.3 Distribution of asphericity parameter ($\Delta$) for single-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes respectively. Thin bars (red) represent the distribution of uncomplexed proteins.

in Fig. 3.3(b) and Fig. 3.3(c) which represents the protein distribution plot for the protein-DNA and protein-RNA complexes respectively. The value of $\Delta$ varies from 0 to 0.1 for 75.8\% proteins and from 0.1 to 0.3 for 21\% proteins in protein-DNA complexes. Fig. 3.4(a), Fig. 3.4(b) and Fig. 3.4(c) display $P(\Delta)$ as a function of $\Delta$ for the multi-chain proteins in different complexes. The distribution of $\Delta$ clearly portrays that the oligomeric proteins have a broad distribution and are less spherically symmetric compared to the uncomplexed single-chain proteins. Fig. 3.5 confirms the same trend in depicting the distribution of $P(R_\Delta)$ vs $R_\Delta$.

$$R_\Delta = \frac{\Delta_o}{\Delta_i}$$  \hspace{1cm} (3.11)

where $\Delta_o$ and $\Delta_i$ are the respective asphericity parameters of the oligomer and
3.3 Results and Discussions

The individual chains in the oligomer. Fig. 3.3 clearly depicts that the distribution of the asphericity parameter for proteins in the protein-protein and protein-DNA complexes does not differ much from the single-chain uncomplexed proteins while the distribution is comparatively broader for proteins in the protein-RNA complexes.

Typically DNA molecules are double helical with a diameter of \( \approx 20 \, \text{Å} \). Fig. 3.6(a) shows the plot of \( P(\Delta) \) vs \( \Delta \) for DNA molecules where \( \Delta \) values range from 0 to 0.8. These values predict that even though 22% DNA have \( \Delta > 0.5 \), 53.5% of the DNA are spherical with \( \Delta < 0.3 \). The \( \Delta \) values of 24.5% proteins vary between 0.3 to 0.5.

This sphericity may arise due to the small size of majority of DNA for which the values of \( R_G \) may be equal to the chain radius. Among proteins in the protein-RNA complexes, 69.61% proteins have \( \Delta \) values between 0 to 0.1 and 24.89% range between 0.1 to 0.3. Fig. 3.6(b) displays \( P(\Delta) \) of RNA, 61% of which have \( \Delta < 0.2 \) and 20%
3.3 Results and Discussions

Figure 3.5 Distribution of $R_\Delta$ for the multi-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes respectively.

Figure 3.6 Distribution of asphericity parameter ($\Delta$) for (a) DNA and (b) RNA in their respective complexes.
3.3 Results and Discussions

Figure 3.7 Distribution of shape parameter ($S$) for single-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes respectively. Thin bars (red) represent the distribution of uncomplexed proteins.

have $\Delta$ values between $0.2 - 0.3$. This analysis proves that majority of the complexed proteins and DNA/RNA are mostly spherically symmetrical.

Shape

The distribution of the shape parameter ($S$) for complexed single-single-chain proteins is displayed in Fig. 3.7(a), Fig. 3.7(b) and Fig. 3.7(c) respectively. The plots show that 63.3%, 68.2% and 56.4% of proteins in the respective protein-protein, protein-DNA and protein-RNA complexes have $S \approx 0$ and are spherically symmetrical. Only 6.75%, 4.43% and 7.12% of proteins in protein-protein, protein-DNA and protein-RNA complexes have $S < 0$ and are oblate-shaped. For uncomplexed single-chain proteins, 67.82% have $S \approx 0$ and 5.16% proteins are oblate with $S$ from $-0.01$ to $-0.20$. The shape parameter distribution for multi-chain proteins in protein-protein,
3.3 Results and Discussions

Figure 3.8 Distribution of shape parameter \((S)\) for multi-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes respectively.

proteins in protein-protein, (b) protein-DNA and (c) protein-RNA complexes respectively indicates that approximately 50% of proteins are prolate and 30% of them are almost spherical with \(S \approx 0\) in all the complexes. Approximately 16%, 11% and 17% proteins in protein-protein, protein-DNA and protein-RNA complexes are oblate. These results validate that the intrachain interactions in multi-chain proteins significantly change their overall shape.

The values of \(S\) for DNA as depicted in Fig. 3.9(a) reveals that 70% of DNA are prolate, 14% are spherical with \(S \approx 0\) and 16% are oblate-shaped. Among the prolate-shaped DNA, 25% have \(S < 0.5\). Majority of uncomplexed DNA are prolate. Thus the shape of DNA in complexed state exhibits a crossover from the prolate to more spherical/oblate structures in the complexed state. Shape parameter analysis from Fig. 3.9(b) reveals that 35.2% of RNA are spherical in complexed state 40.93%
3.3 Results and Discussions

Figure 3.9 Distribution of shape parameter ($S$) for (a) DNA and (b) RNA in their respective complexes.

of RNA have $S < 0.3$ and 12% are oblate in shape. This shows that in the complexed form RNA are less prolate compared to the uncomplexed molecules where the $S$ value is centered around $S = 0.3$ [88]. There is no explicit dependence of $\Delta$ or $S$ on $N$ as is evident from the correlation of the plot of $\Delta$ or $S$ vs $N$.

3.3.4 Persistence Length

The theoretical and experimental studies for proteins [135], DNA [136] and RNA [137] demonstrate that they can be suitably described by the WLC model. The inter-residue distance distribution function $P(r)$ may be calculated from the three dimensional coordinates of the center of mass of residues in the chain. The distribution $P(x)$ for complexed proteins and RNA depicted in Fig. 3.10 shows that irrespective of the size, sequence or the nature of interactions which stabilize the native state, the tail region of $P(r)$ (for $r > R_G$) for all proteins coincide. Hence it may be inferred that the distance distribution function of RNA and proteins structures in complexes are well described by the worm-like chain model and the value of the persistence length is obtained by fitting $P(r)$ to Eqn. 3.8 in the limit of $r > R_G$ [88]. The inter-residue
### 3.3 Results and Discussions

#### Figure 3.10

(i) Distance distributions $P(r)$ as a function of $r$ for single-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes and (d) for RNA in protein-RNA complexes. (ii) The distribution of $P(x)/\beta$ vs $x$ determines the dependence of $P(r)$ on the dimensionless variable $x = r l_p / R_G^2$ for proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes and for (d) RNA in protein-RNA complexes.

The distance distribution function for DNA is similar to the distance distribution function for prolate objects as determined from small-angle scattering experiments [81,82]. The persistence length of DNA may be evaluated from the Benoit-Doty expression for the unperturbed radius of gyration of WLC given by Eqn. 3.10 [106].

Fig. 3.11 depicts the persistence length as a function of the chain length for the proteins, DNA and RNA in different complexes. Fig. 3.11(a) sketches the persistence length of proteins in the single-single-chain protein complexes, where $l_p$ is found to vary from 2.14 to 10.54 Å increasing sub-linearly with the chain length $N$ according to the relation $l_p = (1.05 \pm 0.03)N^{(0.32\pm0.006)}$. The linear fit of the plot of $l_p$ vs $N$...
3.3 Results and Discussions

Figure 3.11 Persistence length as a function of $N$. The straight line (red) is the best fit line for the data set of single-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes, (d) DNA, (e) RNA in complexes and (f) for the uncomplexed proteins.

![Figure 3.11 Persistence length as a function of N.](image)

gives a correlation-coefficient of 0.802. Similarly from Fig. 3.11(b), $l_p$ for proteins in protein-DNA complexes scale as $l_p = (1.12 \pm 0.03)N^{(0.33\pm0.005)}$. The plot of $l_p$ vs $N$ is linear which yields a correlation of 0.88 with $l_p$ ranging from 2.84 to 15.36 Å. For proteins in protein-RNA complexes, $l_p$ for proteins ranges from 3.1 to 13.4 Å while Fig. 3.11(c) shows that $l_p$ varies with $N$ and follows $l_p = (0.90 \pm 0.03)N^{(0.37\pm0.005)}$ with a correlation-coefficient of 0.88. The scaling exponent $\nu$ is related to different types of interactions between the proteins and protein-nucleic acid which stabilizes the complex at length scales comparable to $l_p$ and the compactness of the structure varies from a maximally compact collapsed polymer to that of proteins in intermediate solvent condition ($\nu \sim 2/5$). Persistence length for proteins in protein-protein and protein-DNA complexes are similar to that of the uncomplexed proteins, while $l_p$ for proteins in protein-RNA complexes resembles that of the natively unstructured proteins [132,138] (protein molecules in intermediate solution condition).
Fig. 3.11(d) displays a plot of \( l_p \) vs chain length \( (N) \) for DNA molecules. The value of \( l_p \) ranges from 2.8 to 24.6 Å for DNA molecules of size, 6 − 147 base pairs. The plot shows that \( l_p \) scales with \( N \) according to \( l_p = (0.97 \pm 0.02)N^{(0.72\pm0.01)} \) with a correlation-coefficient of 0.79. \( l_p \) increases linearly with the contour length \( L \) and \( L/l_p = 5.95 \). The flexibility of DNA increases in the complexed state and they behave close to semiflexible polymers rather than rigid rods in the limit of small chain lengths. Fig. 3.11(e) depicts the variation of \( l_p \) with \( N \) for RNA. The calculated \( l_p \) varies from 2.7 to 23.7 Å scaling sub-linearly with respect to the chain length as \( l_p = (1.4 \pm 0.05)N^{(0.39\pm0.01)} \) with a correlation-coefficient of 0.78. The persistence length of RNA is dictated by the effective balance of various competing interactions like stacking, hydrogen bonding, hydrophobic forces and repulsion between the phosphate groups. In uncomplexed RNA, due to the presence of these interactions, \( l_p \) scales as the range of these interactions i.e. \( l_p = N^{0.33} \). The higher scaling exponent of the complexed RNA arises due to the presence of non-local interactions which increases with \( N \) competing with the requirement that these complexes should adopt a unique native fold minimizing entropic fluctuations. The flexibility of RNA decreases due to complexation and they scale as proteins in an intermediate solvent. Persistence length of RNA in complexed form is approximately same as that of the proteins in protein-RNA complexes. This finding suggests that RNA are as compact and densely packed as proteins and these molecules are more intermingled in complexed form.

### 3.3.5  \( \chi^2 \) Analysis

To evaluate the accuracy and statistical significance of the respective scaling relations for \( R_G \) and \( l_p \), \( \chi^2 \) values are calculated from the difference of the observed and expected values. Values of \( \chi^2 \) for the power law fit is evaluated by the combination of errors.

\[
sA^2 = \left( \frac{\delta A}{\delta m} \right)^2 \times sm^2 + \left( \frac{\delta A}{\delta c} \right)^2 \times sc^2
\] (3.12)

where \( A \) is the expected value given by \( A = 10^{(m \times \log(B) + c)} \), \( sA \) is the standard error on \( A \), \( m \) is the gradient, \( sm \) is standard error on the gradient, \( c \) is y-intercept and \( sc \)
standard error on the y-intercept. The correctness of the fit is verified by calculating the $\chi^2_{\text{CRIT}}$ values from the number of degrees of freedom of the system. If the $\chi^2$ values are significantly smaller than the $\chi^2_{\text{CRIT}}$ values, the null hypothesis is accepted and the accuracy of the fit is verified [90]. The $\chi^2$ values for the $R_G$ vs $N$ plot are 134.67, 125.0, 130.22 for proteins in different complexes, which are significantly lower than the corresponding $\chi^2_{\text{CRIT}}$ values of 1494.34, 1091.26, 1303.4. The $\chi^2$ values for DNA and RNA in complexes are 26.86 and 117.17 while for uncomplexed proteins it is 1097.55. These values are significantly lower than the corresponding $\chi^2_{\text{CRIT}}$ values of 2219.0, 875.54 and 4576.79 respectively. Similarly, the $\chi^2$ values for the $l_p$ vs $N$ plot are 43.82, 39.63, 41.41 for complexed proteins, 73.19, 57.43 for DNA and RNA and 64.82 for uncomplexed proteins respectively. All values of $\chi^2$ are found to be significantly smaller than the corresponding $\chi^2_{\text{CRIT}}$ values. Thus the accuracy of both fits are verified.

### 3.3.6 Two Body Contact Analysis

Proteins form densely packed structures resembling the characteristics of high-density liquids and glasses. The number and type of two body contacts determine the packing efficiency in proteins, which dictate their overall shape. We have analyzed the number and type of two body contacts in the chosen data sets of complexes for single-single-chain proteins with the cutoffs $D_{ss} = 5.2$ Å, $D_{sb} = 5.5$ Å, $D_{bb} = 6.5$ Å using the nine classes of amino acids [4]. The two body contacts are calculated between backbones and side-chains, between backbones and between side-chains.

In Fig. 3.12, the distributions, $P(Q_1)$ of $Q_1 = N_{bb}/N_{ss}$ and $P(Q_2)$ of $Q_2 = N_{sb}/N_{ss}$ where $N_{bb}$, $N_{sb}$, $N_{ss}$ are the numbers of $b-b$, $s-b$, $s-s$ contacts respectively, reveal distinct characteristic features for the proteins in three different types of complexes. For proteins in protein-protein and protein-DNA complexes, the distribution $P(Q_1)$ is peaked around 1 and $P(Q_2)$ around 2, which implies that in the native state $s-s$ contacts are as frequent as the $b-b$ contacts. The larger peak value for $P(Q_2)$ suggest the predominance of the $s-b$ contacts which implies that the molecules have maximum
3.3 Results and Discussions

Figure 3.12 Distributions of $Q_1 = N_{bb}/N_{ss}$ and $Q_2 = N_{sb}/N_{ss}$, where $N_{bb}$ is the number of $b-b$ contacts, $N_{sb}$ is the number of $s-b$ contacts and $N_{ss}$ is the number of $s-s$ contacts in single-chain proteins in (a) protein-protein, (b) protein-DNA and (c) protein-RNA complexes respectively and for the uncomplexed proteins.

short range interactions [91] while the diminished $s-s$ contact frequency value indicates that the molecule is less densely packed. Analysis of the nature of amino acids that participate in $s-s$ contact of proteins in protein-protein complexes highlights that the largest contribution is due to hydrophobic residues i.e. 45.91%, of which 14.08% is between small hydrophobes, 4.26% between small and large hydrophobes and only 1.45% between large hydrophobes. The polar residues constitute 23.04% and the charged residues contribute 20.72%. Analysis of the nature of amino acids that form $s-s$ contacts in protein-DNA complexes is displayed in Fig. 3.12, where the number of hydrophobic residues is found to be 35.35%, polar residues constitute 29.34% and 24.12% are charged residues.

For the proteins in protein-RNA complexes $P(Q_1)$ and $P(Q_2)$ are peaked around 1 and 2.5 respectively, implying that two body contacts are largely dominated by the $s-b$ contacts, while the $s-s$ contacts have the same frequency as that of the $b-b$ contacts. Similar to the protein-protein and protein-DNA complexes, only
the short range interactions are predominant in the protein-RNA complexes. The density is intermediate between the compact and elongated proteins as the respective $s - s$ contacts are neither very dominant nor the least frequent. These proteins are stabilised mainly because of the large percentage of the $s - b$ contacts which is predominant in acquiring the functionally active state. Analysis of the nature of their amino acids side-chains taking part in $s - s$ contact shows that about 25.66% are hydrophobic, 27.28% are polar and 43.39% are charged in nature.

For the uncomplexed proteins $s - s$ contact analysis of the residues shows that 41.81% are hydrophobic, 21.69% are polar and 25.02% are charged in nature. $P(Q_1)$ distribution for all the complexes and uncomplexed proteins is similar implying that on complex formation there is no appreciable change in the $b - b$ contacts. The $P(Q_2)$ distribution for uncomplexed proteins and protein-DNA complexes are similar, implying that the $s - b$ contacts remain almost same in the protein-DNA complexes compared to the uncomplexed counterpart. Considerable deviations in $P(Q_2)$ distribution are noted for the protein-protein and protein-RNA complexes. The distributions are comparatively broader with an increase in the average value of $P(Q_2)$ indicating that the number of $s - b$ contacts increase and it is the maximum for protein-RNA complexes.

### 3.3.7 Packing Density

The average packing density of the protein molecules is an important physical property for evaluating the protein structure [20, 139]. For uncomplexed small proteins, the average packing density of protein interiors is similar to that of a crystalline solid [133], i.e. 0.74. In Fig. 3.13, average packing density of proteins in complexed form and uncomplexed form is plotted with respect to the chain length $N$ which shows that the packing densities decreases with the increase in chain length [140]. Small protein molecules in all the three complexes ($N = 30 – 300$) exhibit a broad range of packing densities varying from the 0.67 to 0.87 while for large proteins it ranges from 0.69 to 0.74. The average packing density of single-chain proteins in protein-protein,
3.3 Results and Discussions

![Figure 3.13](image)

**Figure 3.13** Packing density as a function of \( N \) for (a) the uncomplexed proteins, (b) single-chain proteins in protein-protein complex, (c) single-chain proteins in protein-DNA complex and (d) single-chain proteins in protein-RNA complexes respectively.

protein-DNA and protein-RNA complexes are illustrated in Fig. 3.13(b), Fig. 3.13(c) and Fig. 3.13(d) respectively. The average packing density for 61.03%, 62.17% and 75.04% proteins in the protein-protein and protein-DNA and protein-RNA complexes respectively lie in the range 0.67 – 0.72, while only a minority of proteins in their respective complexes i.e. 11.1%, 5.16% and 6.67% exhibit an average packing density > 0.78. This shows that most proteins in the respective complexed states are less packed compared to the uncomplexed ones. Fig. 3.13(a) depicts that 56% of proteins have an average packing density ranging from 0.74 – 0.72 and for 34% it lies between 0.72 – 0.69. Proteins with a higher packing density have smaller chains lengths without any complex tertiary structures.

The contact density of complexed and uncomplexed proteins provide a direct measure of the differences in their packing density. The C-\( \alpha \) contact density distribution,
3.3 Results and Discussions

\[ P(n_k) \]

is calculated with a cutoff \( R_c = 7.3 \text{ Å} \) for the chosen data set of protein-protein and protein-nucleic acid complexes. From the contact density plot, Fig. 3.14, it may be noted that though the distribution pattern is similar in all cases, the cumulative distribution for proteins in protein-protein complexes is slightly lower than the distribution for the proteins in protein-nucleic acid complexes. The distribution of contact density is broad spanning the entire range of \( 10 \) – \( 100 \). For small values of the contact density \( n_k \), the contribution of \( P(n_k) \) is mainly due to the C-\( \alpha \) atoms near the surface of the protein. The higher flank of \( P(n_k) \) drops off sharply at the close packing limit of \( n_k \approx 80 \). The mean contact densities for the proteins in protein-protein, protein-DNA and protein-RNA complexes are 55, 54 and 56 respectively, which are significantly smaller compared to the value 67.5 [131] corresponding to the uncomplexed proteins.

The differences among the interfaces of the protein-protein and protein-nucleic acid complexes may be determined by calculating the percentage of hydrogen bonding. Fig. 3.15 illustrates the distribution of hydrogen bonds with respect to the chain length of the protein for all three types of complexes. In protein-protein complexes,
the average number of hydrogen bonds are less compared to that in protein-DNA and protein-RNA complexes. The ratio of hydrogen bonds to the total number of $s-s$ contacts in interfacial area is equal to 0.5 for the protein-nucleic acid complexes while for protein-protein complexes the value of the same ratio is 0.2. Various types of interactions exist between the amino acid residues that are in contact with the nucleic acids. Most contact interactions are positively charged constituting 42%, while 21.75% are hydrophilic and 21% contacts are hydrophobic. Unlike uncomplexed proteins the polar/charged residues play a dominant role in the intrachain and interchain interactions for protein-RNA and protein-DNA complexes while for protein-protein complexes, the intrachain interactions are predominantly hydrophobic i.e. 56.22%, the polar and charged residue contributions are 23.03% and 20.72% respectively. The interchain (hydrogen bonding) interactions are composed of 46.54% hydrophobic, 24.24% polar and 24.12% charged residue contributions, of which 14.08% are between small hydrophobes 4.26% are between large hydrophobes and 1.45% are bei-
3.4 Conclusions

A separate analysis for the protein docking benchmark set comprising of 140 non-redundant protein-protein complexes and their corresponding uncomplexed counterparts, reveals that the results for this data subset are almost similar to the selected database of 4339 protein-protein and protein-nucleic acids complexes. For the uncomplexed proteins, the value of the Flory’s exponent for the scaling of $R_G$ is calculated to be 0.33 for the docked protein set, which is very close to the value of 0.34 found in the single-single-chain protein complexes. The flexibility of the uncomplexed proteins in the benchmark set of docked proteins is also found to be similar to that of the main non-redundant database of complexed proteins. The persistence length, $l_p$ varies as $l_p = 1.02N^{0.33}$ for the docked protein set and $l_p = 1.00N^{0.34}$ for the database of complexed proteins compiled from PDB. Comparisons of the distribution of asphericity, shape parameter and packing density is found to follow a similar trend for both sets. The above study thus highlights the point that for high resolution protein-protein and protein-nucleic acid complexes, the effect of X-ray, NMR or other crystallographic techniques are not responsible for the differences in their physical properties compared to the uncomplexed proteins or nucleic acids.

3.4 Conclusions

This study deals with the analysis of shape, flexibility and packing of biomolecules which reveals that there are considerable changes in the shape and topological properties of proteins, DNA and RNA in their respective complexed states. The structural features of various complexes of proteins and nucleic acids are studied using the currently available three dimensional structures in PDB and NDB. The size of proteins and nucleic acids, as measured by $R_G$, follows Flory’s scaling law with different exponents. Proteins in protein-protein and protein-DNA complexes are compact following the collapsed walk behavior with the size exponent 0.33 while DNA molecules are maximally expanded representing a self-avoiding walk with a scaling exponent
of 0.63. Proteins and RNA in protein-RNA complexes have intermediate compact structures and resembles the scaling behavior of proteins in physiological condition representing an intermediate scaling regime between a collapsed and a random walk. Although the overall scaling law for the size of RNA is almost similar to that of the proteins, there are considerable differences in their shapes. The distribution of the asphericity, \( \Delta \) and shape parameter, \( S \) show that complexed DNA and RNA are largely prolate and are considerably more aspherical compared to the proteins which are more spherically symmetric with less prolate and oblate shapes. The shape of DNA in complexed state switches from a prolate to more compact spherical/oblate structures in the complexed state. This variation in shape reflects the conformational preferences of DNA subsequent to the binding of the protein, which may be coded in specific nucleotide sequences [141].

The persistence length, \( l_p \) increases consistently with the chain length, \( N \) for all proteins and nucleic acids present in protein-protein and protein-nucleic acid complexes implying an increase of the non-local interactions which stabilize the respective folded conformations. For small DNA molecules (6 – 147 base pairs) \( l_p \) is larger compared to RNA and proteins in protein-protein and protein-RNA complexes. The flexibility of DNA increases, while RNA decrease on complexation as opposed to the proteins where it remains almost unchanged, reflecting a clear separation in the energy scales stabilizing the secondary and tertiary structures of nucleic acids. The increase in flexibility of DNA may be due to the compression or expansion of the major/minor grooves or bending of DNA in such a way so as to optimize the protein-DNA interface, resulting in large conformational changes. This conformational and directional complementarity of protein-DNA contacts assists in protein-DNA recognition [118].

The change in packing density of proteins may be assessed from the analysis of two body contacts in proteins, hydrogen bonding in interface area and the contact density of the C-\( \alpha \) atoms. The \( s-b \) contacts are found to be more prevalent compared to the \( s-s \) and \( b-b \) contacts. Least \( s-s \) contacts are observed for protein in protein-RNA complexes while for proteins in protein-protein and protein-DNA complexes \( s-s \)
and $b - b$ contacts are equally frequent. The cumulative distribution for proteins in protein-protein complexes is slightly lower than the distribution for the proteins in protein-nucleic acid complexes. The mean contact densities for the proteins in protein-protein, protein-DNA and protein-RNA complexes are significantly less compared to that of the uncomplexed proteins.

Hydrogen bonds are the most predominant stabilizing interactions, contributing approximately 50% in protein-nucleic acid interfaces, while the average number of hydrogen bonds are much less in the interface region of protein-protein complexes. The polar/charged residues play a dominant role in the intrachain and interchain interactions for protein-RNA and protein-DNA complexes and may be involved in the long distance pre-orientation of the peptide chains. For protein-protein complexes, the intrachain and interchain interactions are predominantly hydrophobic. The degree of asymmetry in the complexed proteins may be due to dispersions in the energetic interactions involving side-chains and the peptide backbone.