Computational Structural Bioinformatics II: Molecular Models

Chandrajit Bajaj

November 17, 2010

Contents

A Molecules **5**
A.1 Internal Coordinates . . . 5
A.2 LEG (Labelled Embedded Graph) Representations . . . 7
A.3 FCC (Flexible Chain Complex) Representations . . . 13
A.3.1 Hierarchical Representation . . . 14
A.3.2 Flexibility representation . . . 15
A.3.3 Denavit-Hartenberg scheme . . . 15
A.3.4 Flexibility analysis in molecules - creation of flexible models . . . 16
A.4 Flexibility in RNA . . . 18
A.4.1 Reduced conformation space . . . 18
A.4.2 Classification of RNA using clustering . . . 20
A.4.3 Division of RNA backbone by *suites* . . . 22

Appendix A

Molecules

A.1 Internal Coordinates

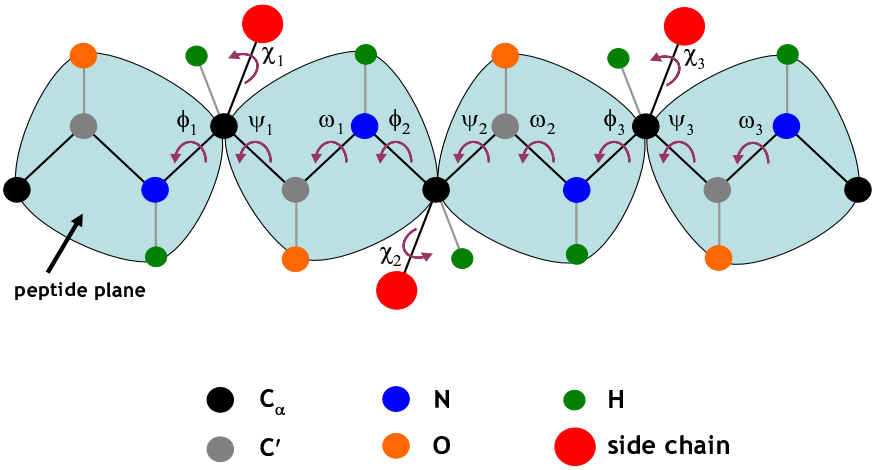


Figure A.1: A polypeptide chain with backbone dihedrals (ψ, ϕ, ω) and side-chain dihedrals (χ) shown.

Proteins have a naturally occurring backbone consisting of $-NH-C(H)R-CO-$ sequences, where R is some functional group defined for 20 different amino acids. These functional groups appear as side-chains connected to the backbone. As in all organic molecules, each type of bond formed in a protein conforms to the characteristic bond length and bond angles for that type. Hence, the conformation of a protein can be approximately defined by a set of *dihedral angles* (or *torsional angles*) that determine the orientation of different chemical groups along and around the backbone.

The following three dihedral angles determine the conformation of the backbone (see Figure A.1).

ϕ_i. This is the angle between the planes $C_{i-1}-N_i-C_{\alpha_i}$ and $N_i-C_{\alpha_i}-C'_i$, i.e., the angle of rotation ($-180^\circ \leq \phi_i < +180^\circ$) around the $N_i-C_{\alpha_i}$ bond. A positive change in the ϕ_i value occurs by counter-clockwise rotation of the $C_{i-1}-N_i-C_{\alpha_i}$ plane around the $N_i-C_{\alpha_i}$ bond.

ψ_i. This is the angle of rotation ($-180^\circ \leq \psi_i < +180^\circ$) around the $C_{\alpha_i}-C'i$ bond, and is determined by the angle between the $N_i-C_{\alpha_i}-C'i$ and $C_{\alpha_i}-C'i-N_{i+1}$ planes.

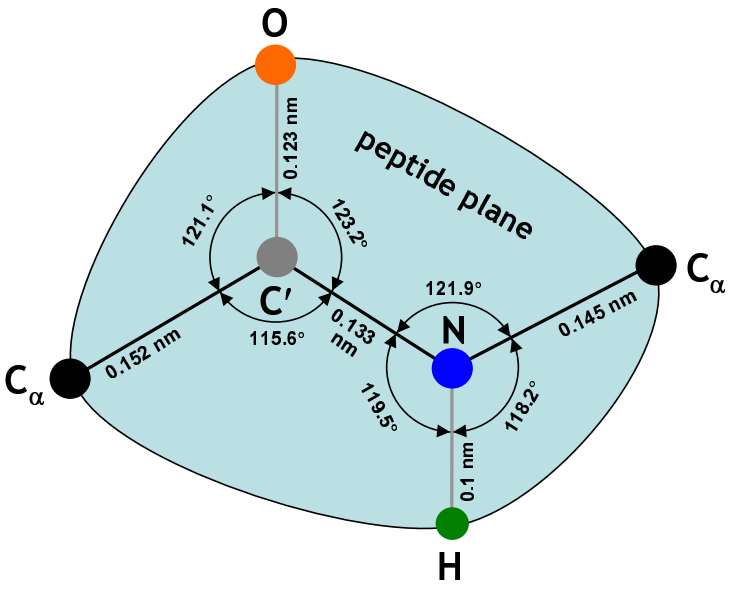


Figure A.2: A peptide plane with all bond lengths and bond angles shown [9].

ω_i. This is the angle of rotation around the peptide bond ($C'_{i-1} - N_i$), and is given by the dihedral angle between the $C_{\alpha_{i-1}} - C'_{i-1} - N_i$ and $C'_{i-1} - N_i - C_{\alpha_i}$ planes. The partial (40 %) double-bond character of the peptide bond and the steric interactions between adjacent side-chains causes the amide group (N_i, C_{α_i}, H_i, C'_{i-1}, O_{i-1} and $C_{\alpha_{i-1}}$) to be almost planar with the distance between $C_{\alpha_{i-1}}$ and C_{α_i} as large as possible (see Figure A.1 for bond lengths and bond angles on this plane). Therefore, almost always $\omega_i \approx 180^\circ$ (for *trans*-peptides), or $\omega_i \approx 0^\circ$ (for *cis*-peptides).

More than than 99.9% of all residues (except proline) are *trans*-peptides, and hence have $\omega_i \approx 180^\circ$. Approximately 5% of all proline peptide bonds have $\omega_i \approx 0^\circ$.

The side chains change conformation through torsional changes in the χ_i angles.

χ_i. Depending on the amino acid type of the side chain there can be up to 4 such successive angles per side chain: $\chi_{i,1}$, $\chi_{i,2}$, $\chi_{i,3}$ and $\chi_{i,4}$. However, for *Glycine* side chain which consists of only one hydrogen atom, and *Alanine* whose side chain is only a single methyl group, these angles are undefined. For all other side chains $\chi_{i,1}$ is defined as the dihedral angle between the planes $N - C_\alpha - C_\beta$ and $C_\alpha - C_\beta - X$, where X is either C_γ, or C_{γ_1} (Val, Ile), O_γ (Ser), O_{γ_1} (Thr), or S_γ (Cys). All side chain dihedrals have values clustered near three conformers known as *gauche*$^+$ or g^+ ($+60^\circ$), *trans* or *t* (180°), and *gauche*$^-$ or g^- (-60°).

Figure A.3 shows the side-chain dihedrals of all amino acids except Glycine and Alanine. Table A.1 shows that about 90% of all side-chains in proteins can be completely described with three dihedral angles (i.e., $\chi_{1,1}$, $\chi_{1,2}$ and $\chi_{1,3}$), and only two dihedral angles (i.e., $\chi_{1,1}$ and $\chi_{1,2}$) are necessary to completely specify more than two-thirds of them.

number of dihedrals (d)	frequency (%)
$d \leq 4$	100.00
$d \leq 3$	89.48
$d \leq 2$	70.64
$d \leq 1$	23.46

Table A.1: Amino acid frequencies in proteins based on the number of (side-chain) dihedrals they have (based on data in [27]).

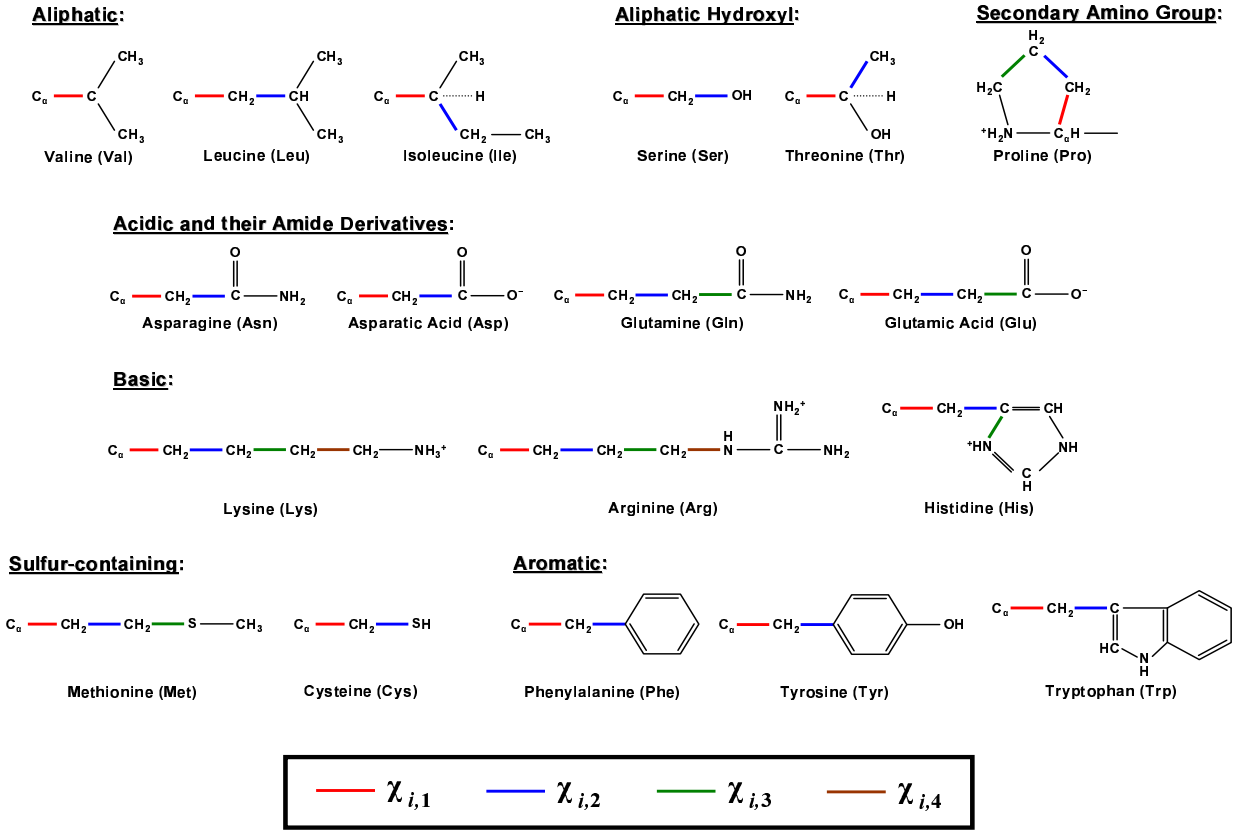


Figure A.3: Side-chain dihedrals ($\chi_{i,1}$, $\chi_{i,2}$, $\chi_{i,3}$, $\chi_{i,4}$) are shown for 18 of the 20 amino acids. The remaining two, i.e., Glycine (Gly) and Alanine (Ala), do not have any side-chain dihedrals. Adapted from [30].

A.2 LEG (Labelled Embedded Graph) Representations

The **LEG** representation of a molecule is simply an annotated graph representation of the chemical structure of the molecule, in which each node represents an atom and each edge a chemical bond. Each atom may be annotated by its symbol and the **vdW** radius, each edge may be annotated by the length of the corresponding chemical bond and possibly a dihedral angle, and each pair of consecutive edges by a bond angle.

In Figure A.3 we show the chemical structures of various amino acids, and in Tables A.3, A.2, A.4 and A.5 we list all possible **vdW** radii, bond lengths and bond angles, respectively, that appear in these chemical structures. Using these information, it is straight-forward to construct the required **LEG** representations of the amino acids.

Since secondary structures (e.g., α-helices and β-sheets) are composed of primary structures (i.e., amino acids), the **LEG** representation of secondary structures can also be constructed from the information in Figure A.3 and Tables A.2, A.4 and A.5. However, the (ϕ, ψ) dihedral angles of the residues in α-helices and β-sheets lie in fairly restricted ranges: $(-45°, -60°)$ for α-helices, $(-120°, 115°)$ for parallel β-sheets, and about $(-140°, 135°)$ for anti-parallel β-sheets. The bond lengths and bond angles may also change slightly.

We can use geometric properties of α-helices and β-sheets in order to extract them from the **LEG** representation L of the given protein P.

Extracting α-helices from L. We traverse L along the peptide backbone of P, and using the internal coordinates (i.e., bond lengths, bond angles, dihedral angles, etc.), bond types and atom types specified in L, we detect and output all maximal contiguous segments of this backbone (along with side chains) that satisfy the following properties of α-helices.

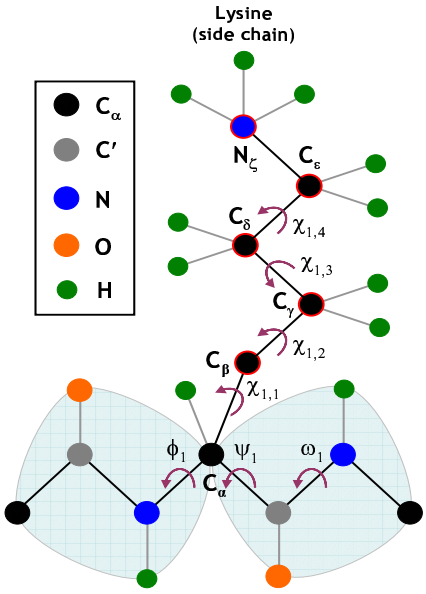


Figure A.4: A Lysine side-chain with side-chain dihedrals ($\chi_{1,1}$, $\chi_{1,2}$, $\chi_{1,3}$, $\chi_{1,4}$).

Atom or Group	Symbol	R_{vdW} (Å)	Notes
>CHR	CA	1.90	Main-chain α-carbon (excluding α-carbon of Gly)
>C=O	C	1.75	Main-chain carbonyl carbon
>CH—	CH	2.01	Side-chain aliphatic carbon with one hydrogen (C^β of Ile, C^γ of Leu, C^β of Thr, C^β of Val)
$>CH_2$	CH2	1.92	Side-chain aliphatic carbon with two hydrogens, except those at β-position and those next to a charged group (C^γ of Arg, $C^{\gamma 1}$ of Ile, C^γ and C^δ of Lys, C^γ of Met, C^γ and C^δ of Pro)
$>CH_2^\beta$	CH2b	1.91	Side-chain aliphatic carbon with two hydrogens at β-position (C^β of Arg, Asn, Asp, Cys, Gln, Glu, His, Leu, Lys, Met, Phe, Pro, Ser, Trp, Tyr)
$>CH_2^{ch}$	CH2ch	1.88	Side-chain aliphatic carbon next to a charged group (C^δ of Arg, C^γ of Glu, C^ϵ of Lys)
$-CH_3$	CH3	1.92	Side-chain aliphatic carbon with three hydrogens (C^β of Ala, $C^{\gamma 2}$ and $C^{\delta 1}$ of Ile, $C^{\delta 1}$ and $C^{\delta 2}$ of Leu, $C^{\gamma 2}$ of Thr, $C^{\gamma 1}$ and $C^{\gamma 2}$ of Val)
−CH=	CHar	1.82	Aromatic carbon with one hydrogen (carbon atoms on the rings of Phe, Trp and Tyr)
>C=	Car	1.74	Aromatic carbon with no hydrogen (C^γ of Phe, C^γ and $C^{\epsilon 2}$ of Trp, C^γ of Tyr)
−CH=	CHim	1.74	C^δ and C^ϵ on the imidazole side-chain of His
>C=O	Cco	1.81	Side-chain carbonyl carbon (C^γ of Asn, C^δ of Gln)
$-COO^-$	Ccoo	1.76	Side-chain carboxyl carbon (C^γ of Asp, C^δ of Glu)
−SH	SH	1.88	S on Cys
−S—	S	1.94	S on Met
>NH	N	1.70	Main-chain amide nitrogen
>NH	NH	1.66	Side-chain nitrogen with one hydrogen ($N^{\epsilon 1}$ of Trp)
$>NHn^+$	NH+	1.65	$N^{\delta 2}$ and $N^{\epsilon 1}$ of His (n = 0 or 1; may be partially charged)
$-NH_2$	NH2	1.62	Side-chain neutral nitrogen with two hydrogen ($N^{\delta 2}$ of Asn, $N^{\epsilon 2}$ of Gln)
$-NH_2^+$	NH2+	1.67	Side-chain partially charged nitrogen on Arg
$-NH_3^+$	NH3+	1.67	Side-chain nitrogen on Lys
>C=O	O	1.49	Main-chain carbonyl oxygen
>C=O	Oco	1.52	Side-chain carbonyl oxygen ($O^{\delta 1}$ of Asn, $O^{\epsilon 1}$ of Gln)
$-COO^-$	Ocoo	1.49	Side-chain carboxyl oxygen ($O^{\delta 1}$ and $O^{\delta 2}$ of Asp, $O^{\epsilon 1}$ and $O^{\epsilon 2}$ of Glu)
−OH	OH	1.54	Side-chain hydroxyl oxygen (O^γ of Ser, $O^{\gamma 2}$ of Thr, O^η of Tyr)
H_2O	H2O	1.68	Water oxygen

Table A.2: List of van der Waals radii for 25 protein atoms [22].

Atom type	Description
C	Carbonyl C atom of the peptide backbone
C5W	Tryptophan C^{γ}
CW	Tryptophan $C^{\delta 2}$, $C^{\epsilon 2}$
CF	Phenylalanine C^{γ}
CY	Tyrosine C^{γ}
CY2	Tyrosine C^{ζ}
C5	Histidine C^{γ}
CN	Neutral carboxylic acid group C atom
CH1E	Tetrahedral C atom with one H atom
CH2E	Tetrahedral C atom with two H atoms (except CH2P, CH2G)
CH2P	Proline C^{γ}, C^{δ}
CH2G	Glycine C^{α}
CH3E	Tetrahedral C atom with three H atoms
CR1E	Aromatic ring C atom with one H atom (except CR1W, CRH, CRHH, CR1H)
CR1W	Tryptophan $C^{\zeta 2}$, $C^{\eta 2}$
CRH	Neutral histidine $C^{\epsilon 1}$
CRHH	Charged histidine $C^{\epsilon 1}$
CR1H	Charged histidine $C^{\delta 2}$
N	Peptide N atom of proline
NR	Unprotonated N atom in histidine
NP	Pyrrole N atom
NH1	Singly protonated N atom (His, Trp, peptide)
NH2	Doubly protonated N atom
NH3	Triply protonated N atom
NC2	Arginine $N^{\eta 1}$, $N^{\eta 2}$
O	Carbonyl O atom
OC	Carboxyl O atom
OH1	Hydroxyl O atom
S	S atom
SM	Methionine S atom
SH1E	Singly protonated S atom

Table A.3: List of atom types [7].

- The amino acids in an α-helix are arranged in a right-handed helical structure with each amino acid corresponding to a 100° turn in the helix and a 1.5 Å translation along the helical axis. Thus there are 13 atoms and 3.6 amino acid residues per turn, and each turn is 5.4 Å wide (see Figure A.5).

Bond type	Bond length (Å)	Bond type	Bond length (Å)
C5W–CW	1·433	CH1E–CH1E	1·540
CW–CW	1·409	CH1E–CH2E	1·530
C–CH1E	1·525	CH1E–CH3E	1·521
C5–CH2E	1·497	CH1E–N	1·466
C5W–CH2E	1·498	CH1E–NH1	1·458
CF–CH2E	1·502	CH1E–NH3	1·491
CY–CH2E	1·512	CH1E–OH1	1·433
C–CH2E	1·516	CH2E–CH2E	1·520
CN–CH2E	1·503	CH2P–CH2E	1·492
C–CH2G	1·516	CH2P–CH2P	1·503
C5W–CR1E	1·365	CH2E–CH3E	1·513
CW–CR1E	1·398	CH2P–N	1·473
CW–CR1W	1·394	CH2G–NH1	1·451
CF–CR1E	1·384	CH2E–NH1	1·460
CY–CR1E	1·389	CH3E–NH1	1·460
CY2–CR1E	1·378	CH2E–NH3	1·489
C5–CR1H	1·354	CH2E–OH1	1·417
C5–CR1E	1·356	CH2E–S	1·822
C–N	1·341	CH2E–SM	1·803
C–NC2	1·326	CH2E–SH1E	1·808
C5–NH1	1·378	CH3E–SM	1·791
CW–NH1	1·370	CR1E–CR1E	1·382
C–NH1	1·329	CR1E–CR1W	1·400
C–NH2	1·328	CR1W–CR1W	1·368
C5–NR	1·371	CR1E–NH1	1·374
C–O	1·231	CRH–NH1	1·345
CN–O	1·208	CRHH–NH1	1·321
C–OC	1·249	CR1H–NH1	1·374
CY2–OH1	1·376	CRH–NR	1·319
C–OH1	1·304		

Table A.4: Bond lengths in proteins [7].

- The *C*=*O* group of residue i forms a hydrogen bond with the *N*-*H* group of residue $i+4$.
- Amino acid residues in an α-helix typically have dihedral angles $\phi \approx -45^\circ$ and $\psi \approx -60^\circ$.

Extracting β-sheets from L. We scan the peptide backbone of P given in L, and detect and output all maximal contiguous segments of this backbone (along with side chains) that satisfy the following properties of β-sheets.

- Each β-strand can be viewed as a helical structure with two residues per turn. The distance between two such consecutive

Angle type	Angle (°)	Angle type	Angle (°)
C5W–CW–CW	107·2	CH3E–CH1E–CH3E	110·8
CW–C5W–CH2E	126·8	CH3E–CH1E–NH1	110·4
C5W–CW–CR1E	133·9	CH3E–CH1E–OH1	109·3
CW–CW–CR1E	118·8	C–CH2E–CH1E	112·6
CW–CW–CR1W	122·4	C5–CH2E–CH1E	113·8
CW–C5W–CR1E	106·3	CF–CH2E–CH1E	113·8
CW–CW–NH1	107·4	C5W–CH2E–CH1E	113·6
CH1E–C–N	116·9	CY–CH2E–CH1E	113·9
CH1E–C–NH1	116·2	C–CH2E–CH2E	112·6
CH1E–C–O	120·8	C–CH2G–NH1	112·5
CH1E–C–OC	117·0	C–CH2G–NH3	112·5
CH2E–C5–CR1E	129·1	CH1E–CH2E–CH1E	116·3
CH2E–C5–CR1H	131·2	CH1E–CH2E–CH2P	104·5
CH2E–CF–CR1E	120·7	CH1E–CH2E–CH2E	114·1
CH2E–C5W–CR1E	126·9	CH1E–CH2E–CH3E	113·8
CH2E–CY–CR1E	120·8	CH1E–CH2E–OH1	111·1
CH2E–C–N	118·2	CH1E–CH2E–S	114·4
CH2G–C–N	118·2	CH1E–CH2E–SH1E	114·4
CH2E–C5–NH1	122·7	CH2E–CH2E–CH2E	111·3
CH2E–C–NH1	116·5	CH2E–CH2P–CH2P	106·1
CH2G–C–NH1	116·4	CH2P–CH2P–N	103·2
CH2E–C–NH2	116·4	CH2E–CH2E–NH1	112·0
CH2E–C5–NR	121·6	CH2E–CH2E–NH3	111·9
CH2E–C–O	120·8	CH2E–CH2E–SM	112·7
CH2G–C–O	120·8	CY2–CR1E–CR1E	119·6
CH2E–C–OC	118·4	CW–CR1E–CR1E	118·6
CH2G–C–OC	118·4	CW–CR1W–CR1W	117·5
CR1E–CY2–CR1E	120·3	CF–CR1E–CR1E	120·7
CR1E–CY–CR1E	118·1	CY–CR1E–CR1E	121·2
CR1E–CF–CR1E	118·6	C5–CR1E–NH1	106·5
CR1W–CW–NH1	130·1	C5–CR1H–NH1	107·2
CR1E–C5–NH1	105·2	C5W–CR1E–NH1	110·2
CR1H–C5–NH1	106·1	C5–CR1E–NR	109·5
CR1E–CY2–OH1	119·9	CR1E–CR1E–CR1W	121·1
N–C–O	122·0	CR1W–CR1W–CR1E	121·5
NC2–C–NC2	119·7	CR1E–CR1E–CR1E	120·0
NC2–C–NH1	120·0	NH1–CRHH–NH1	108·4
NH1–C–O	123·0	NH1–CR1E–NR	111·7
NH2–C–O	122·6	C–N–CH1E	122·6
OC–C–OC	122·9	C–N–CH2P	125·0
C–CH1E–CH1E	109·1	CH1E–N–CH2P	112·0
C–CH1E–CH2E	110·1	C–NH1–CH1E	121·7
C–CH1E–CH3E	110·5	C–NH1–CH2G	120·6
C–CH1E–N	111·8	C–NH1–CH2E	124·2
C–CH1E–NH1	111·2	C–NH1–CH3E	120·6
C–CH1E–NH3	111·2	C5–NH1–CRHH	109·3
CH1E–CH1E–CH2E	110·4	C5–NH1–CRH	109·0
CH1E–CH1E–CH3E	110·5	CW–NH1–CR1E	108·9
CH1E–CH1E–NH1	111·5	CRHH–NH1–CR1H	109·0
CH1E–CH1E–OH1	109·6	CRH–NH1–CR1E	106·9
CH2E–CH1E–CH3E	110·7	C5–NR–CR1E	105·6
CH2E–CH1E–N	103·0	CR1E–NR–CR1E	107·0
CH2E–CH1E–NH1	110·5	CH2E–SM–CH3E	100·9
CH2E–CH1E–NH3	110·5	CH2E–S–S	103·8

Table A.5: Bond angles in proteins [7].

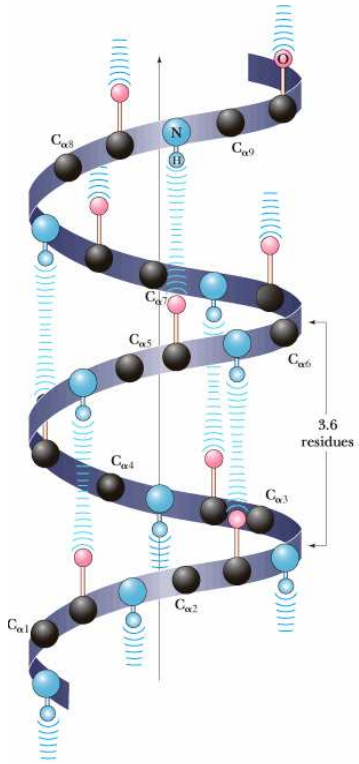


Figure A.5: Geometric structure of an α-helix [9].

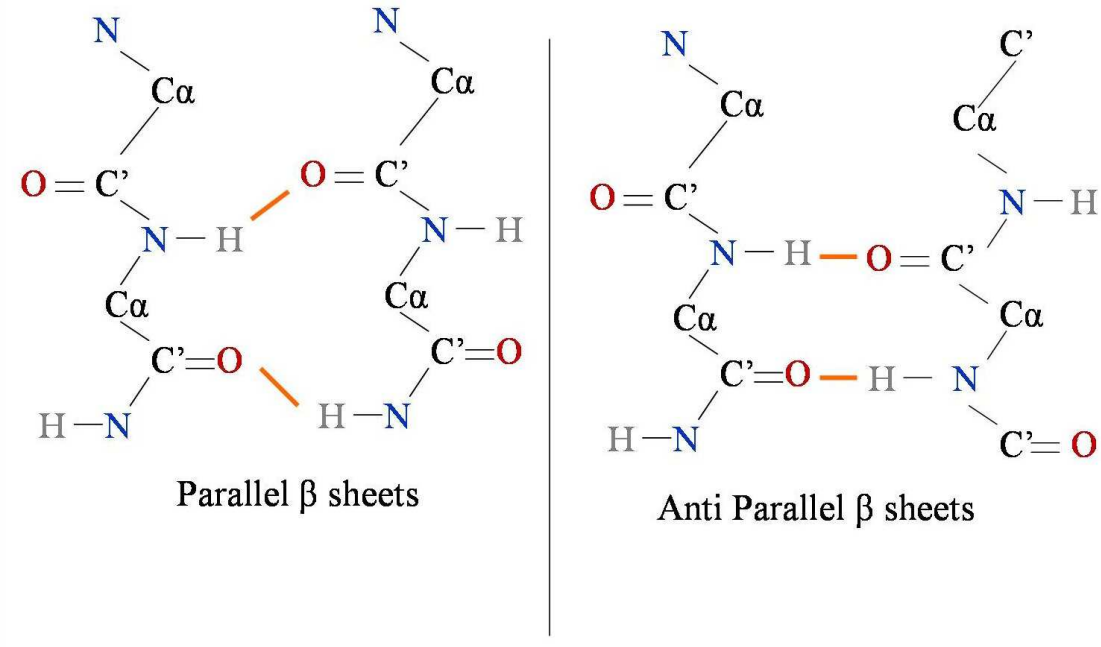


Figure A.6: Geometric structure of a β-sheet [9].

residues is 3.47 Å in anti-parallel β-sheets and 3.25 Å in parallel β-sheets.

- Unlike α-helices the C=O groups in the backbone of a β-strand form hydrogen bonds with the N-H groups in the backbone of adjacent strands.
  - In parallel β-sheets all N-termini of adjacent strands are oriented in the same direction (see Figure A.7(b)). If the C_α atoms of residues i and j of two different strands are adjacent, they do not hydrogen bond to each other, rather rasidue i may form hydrogen bonds to residues $j-1$ or $j+1$ of the other strand.
  - In anti-parallel β-sheets the N-terminus of one strand is adjacent to the C-terminus of the next strand (see Figure A.7(a)). If a pair of C_α atoms from two successive β-strands are adjacent, then unlike in parallel β-sheets they form hydrogen bonds to each other's flanking peptide groups.
- The (ϕ, ψ) dihedrals are about $(-120°, 115°)$ in parallel β-sheets, and about $(-140°, 135°)$ in anti-parallel β-sheets.
- Unlike in α-helices, peptide carbonyl groups in successive residues point in alternating directions.

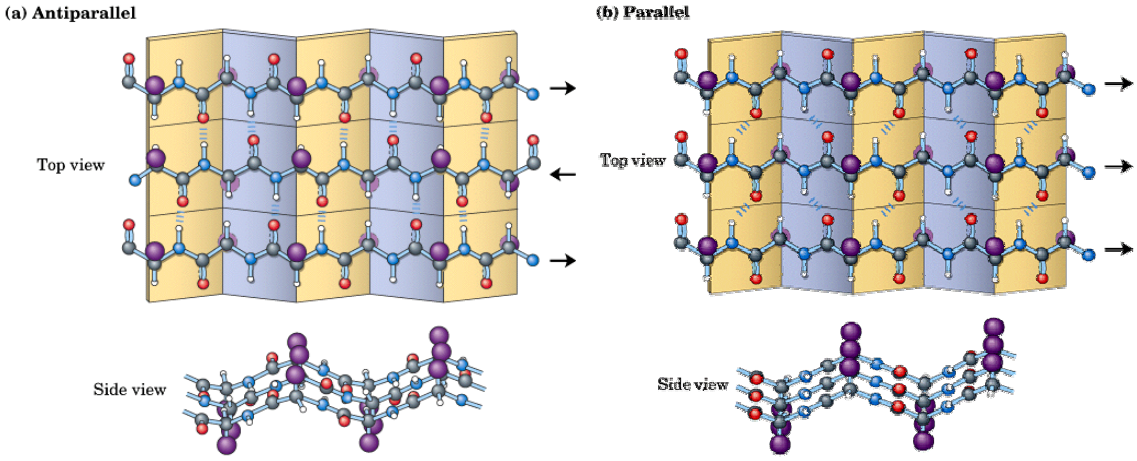


Figure A.7: Two types of β-sheets: (a) anti-parallel, and (b) parallel [13].

A.3 FCC (Flexible Chain Complex) Representations

Complex biomolecules have a naturally occurring backbone, forming chains which flex through their torsion angles. This *nerve* is biochemically well defined, and described by a labeled complex. Structural (shape) and functional properties of a biomolecule can be described as a labeled *sheath* around the central *nerve*. This combined representation (Flexible Chain Complex, or FCC) of a *nerve* and a *sheath* describe a flexible biomolecule.

The nerve of the FCC. The chain complex consists of the following elements.

- *Vertices*: Atom or pseudo atom positions. Atom positions are obtained typically from the PDB files. For pseudo atoms, we use the centers of a set of enclosing spheres which represent the finer level using some error norm like the Hausdorff error.
- *Edges*: Bonds or pseudo bonds. This is again from the PDB or from the hierarchical complex formed by clustering the finer resolutions to a DAG.
- *Faces*: Residues, bases or pseudo structures.

These elements are labeled with the following attributes.

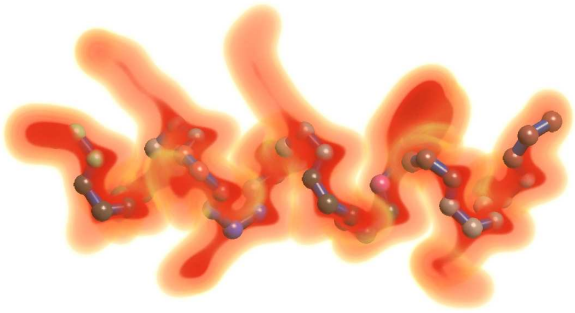


Figure A.8: Flexible Chain Complex: Combined volume (through hardware accelerated 3D texture mapping based volume rendering) and imposter rendering, showing the chain together with the high density volumetric regions formed by the functional groups protruding outwards from the chain.

- Position, length, areas.
- Ranges for flexible angles, lengths.
- Sub structural markers.
- Field attributes.

We allow the molecules to flex around their torsion angles as it is widely accepted that bond angles and bond lengths do not have much flexibility. In protein chains, the ϕ and ψ angle variations are obtained and stored in the complex attributes. For RNA, we have 8 different torsion angles along the backbone. The ranges for these atoms are obtained either from molecular dynamics simulations or from NMR analysis for certain structures.

The sheath of the FCC. The surrounding volume, sub volumes and surfaces of a biomolecule are used to represent shape, volumetric properties (like electrostatics, hydrophobicity) and surface properties (like curvatures). These representations enjoy a dual implicit and explicit representation.

- *Implicit volumetric representation* In this representation, we have a vector containing of (a). A set of centers of expansion points, (b) A parameter referred to as the blobbiness parameter which is useful to represent the van der Waals forces in a continuous and hierarchical fashion, and (c), a set of radii. These parameters are necessary and sufficient to define the electron density function of a molecule. For functions like hydrophobicity and electrostatics, charges at each center of expansion is required.
- *Explicit volumetric representation* There are three representations which can be used for explicitly describing a volumetric function.
  - *Simplicial representation*: The data is described over a simplex like a surface grid at the vertices.
  - *Tensor product*: An explicit grid is used to represent the functions. The size of such a representation can be very large. Hence it is useful to develop compression based algorithms to represent and visualize such a representation.
  - *Multipole summations*: Since our data set consists of a set of vertices and functions which are summations of functions defined over this limited set, Multi-Pole type summations can be used efficiently to represent the data sets.

A.3.1 Hierarchical Representation

Both the skeletal and the volumetric features are represented in a hierarchical fashion. We have a biochemical based static hierarchy of the molecules, with atoms at the finest resolution. Groups of atoms are collapsed to form residues and residues form secondary structures. Chains consist of a set of these secondary structures. A dynamic hierarchy, which could be more useful for interactive dynamic level of detail rendering and manipulation is also performed as outlined in [1].

(a) Atom level

(b) Residue level with a low $G_{dropoff}$ factor

(c) Residue level with a high $G_{dropoff}$ factor

Figure A.9: LOD volume rendering of a large ribosomal subunit (1JJ2.pdb). The parameter $G_{dropoff}$ controls the spread of the density around a pseudo atom when blurring the chain complex

Once a flexible chain complex hierarchy is rebuilt due to dynamic changes in the molecule, the implicitly defined volumetric and surface properties can be quickly updated. Explicit volumes can also be extracted in a hierarchical fashion.

When we have a hierarchical representation of a FCC skeleton, we implicitly have a hierarchical representation of the surrounding differentiable sheath. In figure A.9, we show the large ribosomal subunit at three different levels of a hierarchy.

A.3.2 Flexibility representation

The paper [40] describes how to store the flexibility information in a structure. More specifically, they describe existing and new methods to obtain new atom positions when rotations are performed. Three schemes for storing and manipulating rotation matrices are given below.

Simple rotations scheme. A tree is constructed from the molecule by taking any atom as the root, and bonds in the molecule as bonds in the tree. Rings in a protein are simply taken as a single atom. When a torsional angle changes at a node, then all the nodes below it are rotated to new positions. This rotation update involves a matrix multiplication. The update has to be from the node to the leaves and numerical errors can occur due to manipulating positions of atoms down a chain for each rotation.

Consider a bond b_i rotated by angle θ_i. Let $\mathbf{v}$ be a vector along the bond and T be the translation matrix formed by the i^{th} atoms position. Then the update matrix is

$$T \begin{pmatrix} v_x^2 + (1-v_x^2)cos\theta_i & v_xv_y(1-cos\theta_i) + v_zsin\theta_i & v_zv_x(1-cos\theta_i) + v_ysin\theta_i & 0 \\ v_xv_y(1-cos\theta_i) + v_zsin\theta_i & v_y^2 + (1-v_y^2)cos\theta_i & v_yv_z(1-cos\theta_i) - v_xsin\theta_i & 0 \\ v_zv_x(1-cos\theta_i) - v_ysin\theta_i & v_yv_z(1-cos\theta_i) + v_xsin\theta_i & v_z^2 + (1-v_z^2)cos\theta_i & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} T^{-1} \tag{3.1}$$

A.3.3 Denavit-Hartenberg scheme

In this scheme, we again maintain a tree, with matrices and update from a root to the leaf. But now, the matrices no longer need the information on the current position of the atom, but only the rotations it underwent as a single matrix. Hence this is numerically stable.

To construct the matrix, we first define a local frame at each node. The origin and the vectors are the node position and

- **w** the bond from the node to its parent
- **u** a vector perpendicular to the previous vector and the bond containing this atom and a child. This means that a frame is to be defined for each child.
- **v** a vector perpendicular to the above two.

The matrix which takes a point from one frame defined at a node to the frame of the parent of that node is defined as

$$\begin{pmatrix} cos\theta_i & -sin\theta_i & 0 & 0 \\ sin\theta_i cos\phi_{i-1} & cos\theta_i cos\phi_{i-1} & -sin\phi_{i-1} & -l_i sin\phi_{i-1} \\ sin\theta_i sin\phi_{i-1} & cos\theta_i sin\phi_{i-1} & cos\phi_{i-1} & -l_i cos\phi_{i-1} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.2}$$

θ_i is the torsional angle of bond b_i
ϕ_{i-1} is the bond angle between bonds b_{i-1} and b_i

Atomgroup scheme. This scheme eliminates the requirement for multiple frames and frames where the bond does not rotate. It simply aggregates the tree into a new tree where sets of vertices (atoms) which do not have rotatable bonds are collapsed into a new vertex. Here, we define the local frame as the atomgroup origin and the vectors

- $\mathbf{w}_i$ as a vector along the bond to atomgroup $i-1$
- $\mathbf{u}_i$ as any vector perpendicular to $\mathbf{w}_i$
- $\mathbf{v}_i$ as any vector perpendicular to the above two.

Let the frames after and before rotation be $[\mathbf{x}_i, \mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i]$ and $[\mathbf{x}'_i, \mathbf{u}'_i, \mathbf{v}'_i, \mathbf{w}'_i]$. In this case the transformation matrix, which takes a point in frame i to local frame at $i-1$ (rotated by θ around the connecting bond) is defined as the product

$$\begin{pmatrix} \mathbf{u}_{i-1}.\mathbf{u}'_i & \mathbf{u}_{i-1}.\mathbf{v}'_i & \mathbf{u}_{i-1}.\mathbf{w}_i & \mathbf{u}_{i-1}.(\mathbf{x}_i - \mathbf{x}_{i-1} \\ \mathbf{v}_{i-1}.\mathbf{u}'_i & \mathbf{v}_{i-1}.\mathbf{v}'_i & \mathbf{v}_{i-1}.\mathbf{w}_i & \mathbf{v}_{i-1}.(\mathbf{x}_i - \mathbf{x}_{i-1} \\ \mathbf{w}_{i-1}.\mathbf{u}'_i & \mathbf{w}_{i-1}.\mathbf{v}'_i & \mathbf{w}_{i-1}.\mathbf{w}_i & \mathbf{w}_{i-1}.(\mathbf{x}_i - \mathbf{x}_{i-1} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cos\theta_i & -sin\theta_i & 0 & 0 \\ sin\theta_i & cos\theta_i & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.3}$$

The concatenation of such matrices until the root gives the global position of the atomgroup.

A.3.4 Flexibility analysis in molecules - creation of flexible models

One classification of flexibility analysis methods in the biomolecular area is given by [14] as

- Molecular dynamics

  Molecular dynamics involves simulation of the protein in a solvent environment and saving the conformation state at regular time intervals. Since this simulation is often at very small time scales, (pico or nano seconds), large conformational changes (which occur over micro or milli seconds) will not be recorded. Hence obtaining flexibility analysis through molecular dynamics is limited. An adaptive solver is given in [17]. By allowing users to interact with the system, conformational changes can be forced and observed [21], [34]. A multiple grid method for solving the electrostatics efficiently [33]. Compact structural domains were computed in [12] using simple force calculations in a protein structure.

- Xray Crystallography and Nuclear Magnetic Resonance (NMR)

  Xray Crystallography is used to obtain high resolution images of proteins, upto the atomic level. Most structure in the PDB are generated using this method.

  NMR techniques have been used to obtain dynamic conformations of proteins. The basic idea behind NMR is that atoms have an intrinsic property spin, which determines its behavior when exposed to magnetic fields. Different atoms are seen to emit different frequencies of light, providing an image of the underlying protein as a signature. NMR imaging yields lower resolution results than xray crystallography.

  Given the large number of states which could be obtained from molecular dynamics, NMR and xray crystallography, the following methods generate certain important conformal states by reducing the number of degrees of freedom in the protein.

- Comparison of conformal states

  Protein dynamics give rise to a large number of conformations. Analyzing these conformations for any problem, including flexible protein docking is not computationally feasible. Hence many methods are used to reduce these conformations to a new basis, where the principal basis gave the large fluctuations efficiently. Many authors [37], have shown that the main conformational changes of a protein is mostly captured by using only a few bases and projection vectors, [36]. Normal mode analysis and principal component analysis are two methods to reduce the dimensionality of the problem.

  Singular Value Decomposition (SVD) is commonly used to find basis vectors to reduce the dimensionality of a set of vectors. An equivalent formulation using Principal Component Analysis (PCA) is also done. Consider the column vectors of a matrix A as the zero mean weighted atomic displacement positions. Usually, this vector is also aligned with a given conformation, so that the displacements are relative. The SVD of a matrix is

  $$SVD(A) = U \sum V^T \tag{3.4}$$

  U, V are orthonormal matrices
  $\sum$ is a diagonal matrix
  The diagonal matrix has entries are all non negative and decreasing, called the singular values.

  In this decomposition, the set of left column vectors of U are the basis set for A, and the vectors in V^T are the projections along these basis vectors with magnitudes given by the singular values. Hence, we have an ordering on the influence of the basis vectors for the matrix.

  To apply the PCA algorithm, a matrix A is defined with elements a_{ij} as follows

  $$a_{ij} = ((x_i - x_{i,avg})(x_j - x_{j,avg})) \tag{3.5}$$

  The eigenvector problem $AW = W\zeta$ is solved to get the axis vectors and the corresponding fluctuations in the eigenvectors and eigenvalues [19].

  In [10], a theorem relating the atom displacements to the frequencies of vibrations is presented. In this paper, the authors prove that if a large molecule only flexes around a certain minimal energy state, approximated by a multidimensional parabola, then the average displacements of the atom positions is the sum of the contributions from each normal mode, which is proportional to the inverse square of the frequency [19]. For Normal Mode Analysis (NMA), the moment matrix diagonalized is

  $$A = k_B T F^{-1} \tag{3.6}$$

  k_B is the Boltzmann constant,
  T is the absolute temperature,
  F is a matrix of the second derivatives of the potential energy at a minimum point.

  Successful modeling of the Chaperonin GroEL was performed using NMA in [23]. To avoid the computations on a large matrix, [35] compute a blocked version of NMA by grouping residues.

  Gaussian Network Models (GNM) are used in [18]. In this model, the correlation matrix is formed as

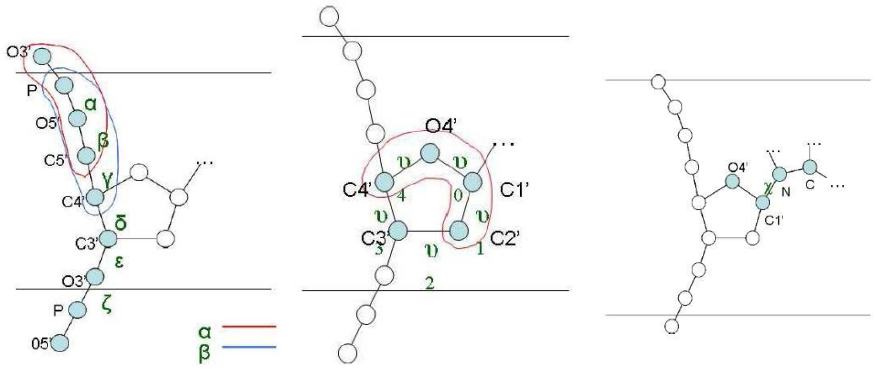
  $$(3kT/\gamma[T^{-1}]_{ij}) \tag{3.7}$$

  k is the boltzmann constant,
  T is the absolute temperature,
  γ is a harmonic potential,
  T is a nearness matrix, called Kirchoff matrix

  The kirchoff matrix inverse can only be approximated since its determinant is 0.

- Deriving flexibility through a single structure.

  Non polar regions in protein tend to lie in the interior and this hydrophobic effect folds the protein. In [39], the authors describe how to capture this information into rigid domains of the protein. Their assumption is that rigid domains folded



(a) The six backbone torsion angles for a RNA

(b) The torsion angles defined along the sugar ring

(c) Nucleotides can rotate about the χ torsion angle

Figure A.10: The torsion angles around which a RNA can flex.

by the hydrophobic effect behave as a *compact unit* during conformational changes. To quantify this, they hierarchically grouped residues in a protein to form a tree, using a coefficient of compactness Z given by

$$Z = \frac{\text{accessible surface area of segment}}{\text{surface area of sphere of equal volume}} \tag{3.8}$$

Static core or the backbone of molecules and their associated rigid domains were computed in [3] using two different conformations of a given protein. α helices, β strands and loops were segmented. Similar pairs of segments were clustered in a tree-like fashion using a rmsd calculation. Domains or compact units of a protein were also computed by [32]. The heuristic they used was that the amount of internal contact a domain had was larger than the amount of contact it had with the rest of the protein. Hence by choosing suitable split planes along the sequence, they form compact sequences. Extending this idea, a Monte Carlo sampling in internal coordinates using relevant torsion angles was performed in [24]. They obtained a set of low energy conformations for any given protein structure as a representation of its flexibility. Using graph theoretical algorithms, [14] obtained flexible and rigid domains in a protein.

A.4 Flexibility in RNA

Flexibility in RNA is given by three sets of angles

- The backbone torsion angles.
- The angles on the sugar ring, also defined by amplitude and a phase.
- An angle about which the residue can flex.

The angles are shown in figure A.10. Due to the large number of angles, people have studied and proposed various means to reduce the conformational space.

A.4.1 Reduced conformation space

Due to the large number of angles defining the flexibility of nucleotides, it is useful to find fewer pseudo torsion angles to represent the other angles.

Reduction to two angles. Duarte et al. have reduced the number of torsion angles necessary to describe an RNA molecule to two, η and θ [5], [4]. Figure A.11 gives the relative positions of these angles and the specific atoms of the backbone involved.

η is the torsion angle resulting from $C4'_{i-1} - P_i - C4'_i - P_{i+1}$. The atoms connected $P_i - C4'_i - P_{i+1} - C4'_{i+1}$ create θ [5]. In their most recent publication, Duarte et al. combined the η and θ data with position information to describe the overall

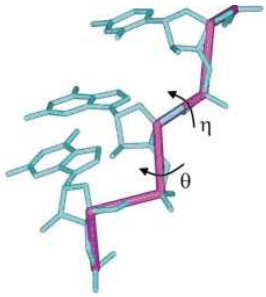


Figure A.11: The backbone torsion angles are represented by just two pseudo rotation angles η and θ.

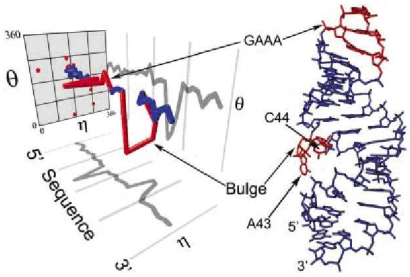


Figure A.12: Example of 3D representation of RNA structure. Plotting the RNA chains using only two angles per residue in a 3D plot shows similar structures along the *worms*

structure of the RNA molecule. Using PRIMOS [6] to create an "RNA worm" - a sequential description of the angle data - allows for analysis of the structure on a nucleotide by nucleotide basis.

After all η and θ angles have been calculated from the PDB [8], NDB [2], and RNABASE [38] data, PRIMOS creates an RNA worm file which gets deposited into a database. The two angles are plotted and a 3^{rd} dimension, sequence, is added to the graph to form a 3D representation of structure. See Figure A.12.

In this plot, A-helices (the most common form of RNA; represented in blue) travel in relatively straight lines, whereas the motifs/other features of the RNA show large deviations from the straight line (shown in red).

To compare RNA worm representations, and thus conformational variations between molecules, it is necessary to find the difference between the η and θ values in the two molecules. Simply put:

$$\Delta(\eta,\theta)_i = \sqrt{(\eta^A - \eta_i^B)^2 + (\theta^A - \theta_i^B)^2} \tag{4.9}$$

The larger the value of $\Delta(\eta,\theta)_i$ the more extreme the disparity between the two RNA fragments, chains, or molecules.

Further, Duarte et al. use this method to compare ribosomal complexes, search for existing motifs, identify new motifs, and characterize two different types of the same motifs. To compare ribosomal complexes, Duarte et al use PRIMOS to calculate differences in h and q when the ribosome is in different conformational states. For example, the conformational state of the ribosome is altered during antibiotic binding or during different stages of translation. The same method can be used to compare conformational states of ribosomes from different species.

To find existing motifs in RNA structures, they used PRIMOS to create another RNA worm database. From this database, a fragment of RNA that contained the motif of interest was selected and compared to every other fragment of the same size within the database and given a score according to equation 4.10.

$$\overline{\Delta(\eta,\theta)} = \frac{\sum_{i=1}^{n} \Delta(\eta,\theta)_i}{n} \tag{4.10}$$

The scores were sorted in increasing order. The smaller scores indicate a closer match.

Reduction to four angles and binning. Hershkovitz et al. [11] suggest a more complex, yet complementary, method to that of Duarte [5]. This method involves calculating four torsion angles, α, γ, δ and ζ, and binning these angles into allowable ranges.

Angle	Bin 1	Bin 2	Bin 3	Bin 4
α	40 - 90	135 - 190	260 - 330	other
γ	35 - 75	150 - 200	260 - 320	other
δ	68 - 93	130 - 165	other	
ζ	255 - 325	other		

Table A.6: Classification of angles:discrete ranges of angles or "bins" as defined by Hershkovitz.

"Binning" is a term used to describe the technique used by Hershkovitz to classify various RNA configurations into discrete bins. For example, nucleotides in the A-form helix, the most common conformation of RNA, have a bin number of 3111 where each number represents which "bin", or range, the torsion angles belong to (i.e. α is in bin 3, or 260° - 320° and γ, δ and ζ and are in bin 1, or 35° - 75°). See Table A.6.

The bin number combination 3111 is then assigned an ASCII character, "a". All combinations of bin numbers are assigned a unique ASCII character, enabling the entire RNA chain to be described by a sequence of letters that represent the structure of the molecule. Their goals were to recognize and catalogue all the RNA conformational states, eliminate any unnecessary angle information, and to assess the validity of their binning model by comparing it to a torsion-matching model. The torsion-matching method for RNA motif searching is a brute force method. So while it is highly accurate, it is computationally expensive as it involves calculating all backbone angles including a ribopseudorotation phase angle, P, for each residue and comparing each set of angles to all other sets of angles in the molecule.

After using the binning method for all RNA fragments and molecules in their database, Hershkovitz et al found 37 distinct conformational states of RNA. Table A.13 lists the assigned bin numbers, the corresponding ASCII symbols, and the observed frequency of these 37 conformational states.

Because this method allows the three dimensional structure of an RNA molecule to be displayed as a sequence of characters, it facilitates motif searching. Without computational aides, one could see that a string of repeating letters (other than "a") represents a possible motif.

Hershkovitz et al suggest an alternative to the Ramachandran plots traditionally used for representing angle distributions. The tree diagram in figure A.14 is a natural progression from the four integer code, or bin. Here the widths of the line correspond to the log of the number of residues in each bin.

A.4.2 Classification of RNA using clustering

Nucleotides from the large ribosomal subunit (1JJ2.pdb) were clustered into commonly occurring structures by Schneider et al. [31]. They classified the non A-type nucleotides separately (830 of them). Eighteen distinct non A-type conformations and fourteen A-type conformations were reported. They report that a large number of the RNA were very close (in a RMSE sense) to the clusters. The authors also say that their results agree with those from Murray et al. [25].

The steps used in obtaining the conformations were as follows.

- Separate the A-type from the non A-type nucleotides.
- Plot the histogram for the backbone ($\alpha, \beta, \gamma, \delta, \varepsilon, \zeta$) and the base ($\chi$) angles.
  - α and γ were seen to have tri-modal distributions.
  - β has a wide gaussian with 180 as its center.
  - ε has values greater than 180 due to the ring, and lacked a gaussian shape.
  - *delta* also was constrained by the ring, and had a sharp bimodal distribution due to the C3'-endo and C2'-endo ribose puckers.
  - The base χ angle was largely bimodal, due to the two main configuration of bases (anti and syn).
  - There was a wide distribution of ζ.
- Plot 2D scatter plots for the following angle pairs : [α, ζ], [β, ζ], [ε, ζ], [γ, α], [χ, ζ] and [χ, δ].
  - The reason for choosing the above sets were not given.

Ascii letter[a]	Bin number	Frequency
a	3111	1709
e	3112	169
r	3122	124
i	2211	103
o	2111	58
t	4111	48
n	1111	37
s	2122	34
l	1211	31
c	3121	30
u	4211	28
d	1121	26
p	4122	21
m	1122	21
h	3411	18
g	1322	18
b	1112	14
f	3211	14
y	4112	13
w	2212	11
k	4121	11
v	3212	10
x	3222	10
z	1331	9
j	4222	9
q	3321	8
1	1212	8
2	3422	8
3	4311	8
4	4411	8
5	2121	7
6	3322	7
7	2222	7
8	2411	7
9	1311	7
0	1221	7
+	3311	6

[a]The assignment of characters to configuration classes was made by frequency of observation. The choice of letter assignment was taken from http://www.askoxford.com/asktheexperts/faq/aboutwords/frequency. All bins with less than five residues are denoted by * and are omitted from this table.

Figure A.13: Classification into 37 clusters through binning

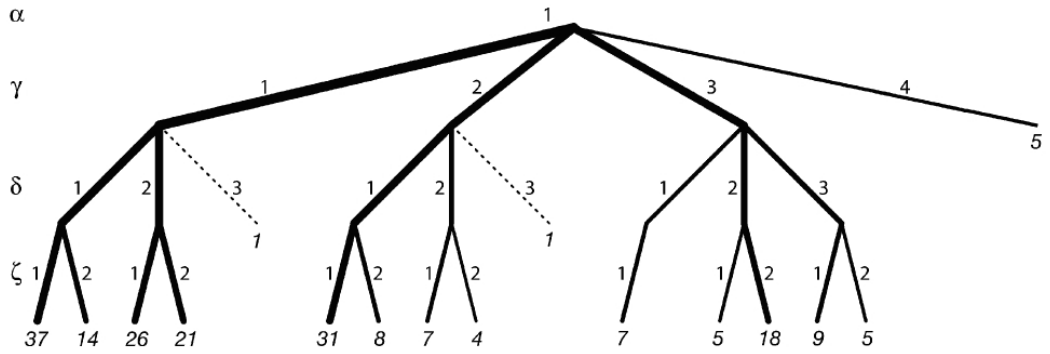


Figure A.14: This tree represents the case where α is 1. There are three others; one for each possible value of α.

  - Clusters were found in the pairs [ζ, α], [α, γ] and [χ, δ].
  - The lack of clusters in other plots led to clustering of 3 tuples of angles.

- From the features and distributions seen in the 1D and 2D plots, the authors choose six 3D plots to base their clusters on to classify the structure of nucleotides.
  - The following six 3tuples were chosen for clustering: [ζ_i, α_{i+1}, δ_i], [ζ_i, α_{i+1}, γ_{i+1}], [α_i, γ_i, δ_i], [ζ_i, α_{i+1}, χ_i], [ζ_i, α_{i+1}, ε_i] and [ζ_i, δ_i, χ_i].
  - The clusters in the 3D plots were assigned peaks and labeled.
  - Each nucleotide was assigned the corresponding label from each plot, if any, or simply a '-'.
- Each nucleotides 6 letter classification was clustered using lexicographic clustering. The authors do not mention why this method was used.
- From this clustering, eighteen distinct non A-type conformations and fourteen A-type conformations were reported.

A.4.3 Division of RNA backbone by *suites*

Murray et al. [25] identify several problems associated with the methods of Murthy [26], Hershkovitz [11], and Duarte [5]. While these methods are excellent at finding and comparing RNA motifs in a large nucleic acid sample, they oversimplify the problem of determining RNA backbone structure. As a result, Murray et al. propose to analyze the folding structure of RNA molecules on a more detailed level, correct the artifacts created in the data structures (sometimes caused by NMR or X-ray crystallography), produce *low-noise data distributions*, and create a list of the resulting, distinct RNA backbone conformers.

The traditional nine angles of the RNA backbone and its bases (i.e. $\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\chi$, and the 2 puckering angles of the sugars) were reduced to six. χ was not included in the model. The two puckering angles were combined and represented as δ, where δ was bimodal - either C3' endo or C2' endo. This allowed the six remaining angles two be divided into 2 sets of 3D distributions, α,β,γ and δ,ε,ζ. Dividing the RNA backbone into *heminucleotides*, a term coined by Malathi and Yathinda [28], in this manner provided some advantage to the traditional phosphate - phosphate division in that it reduced the dimension of the problem and made visualization more feasible. In other words, two 3D plots can be created using α,β,γ data and δ,ε,ζ data respectively. See figure A.15

The methods of Murray et.al were fairly straightforward. They obtained the sequence and structure data samples from the Protein Database and/or the Nucleic Acid Database. From these samples they calculated all the dihedral angles and added hydrogens with REDUCE [15]. The backbone steric hindrances were calculated with PROBE and CLASHLIST [16]. A clash was noted when the overlap between two atoms was greater than 0.4Å. The angles, quality, resolution, base id, highest crystallographic B factor, and d-e-z values were entered into excel. Images were created using the software PREKIN and MAGE from the same authors. For each of the seven peaks created in the δ,ε,ζ distributions, the α,β,γ set was plotted. Finally, a quality filter was applied to rule out nucleotides with greater than 2.4Å resolution.

210 potential RNA conformers were determined from which 146 had an acceptable (low) amount of steric hindrance. 42 conformers had actual cluster points from the data.

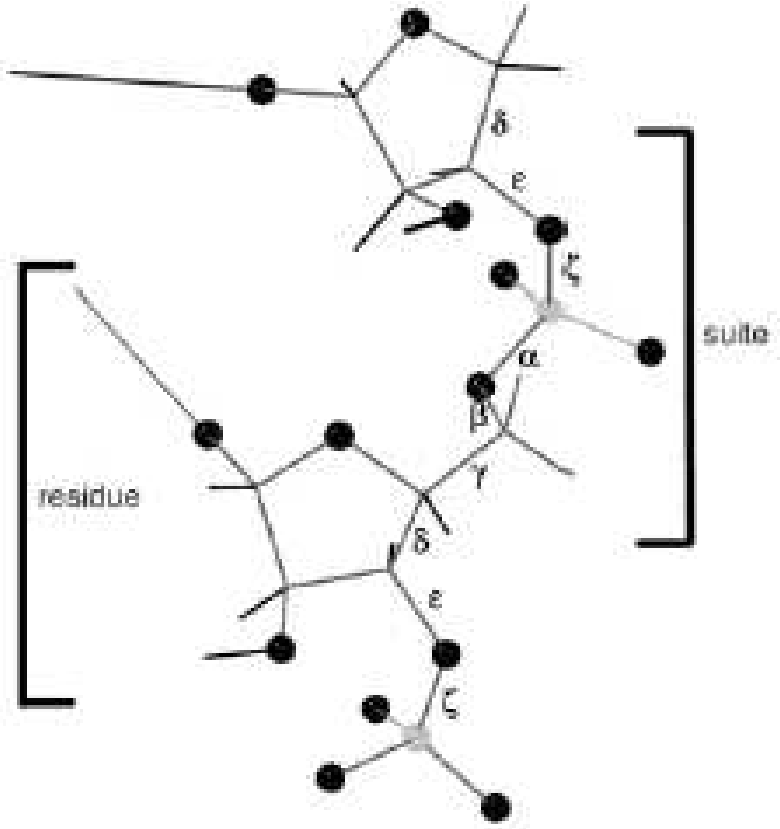


Figure A.15: Division of angles into residue and "suite" data

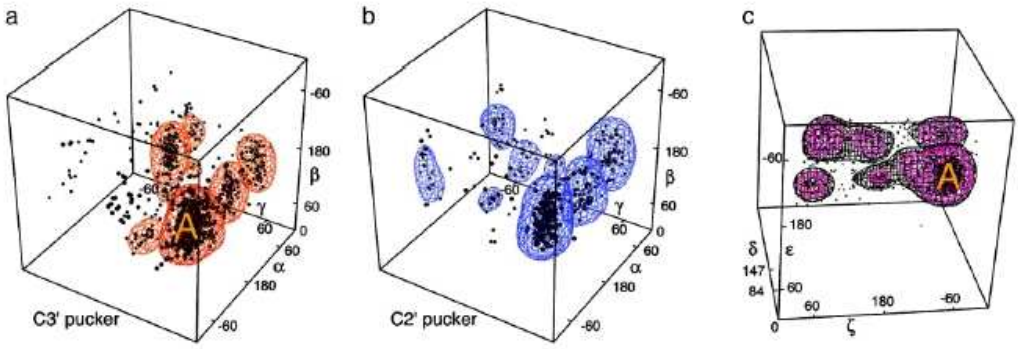


Figure A.16: 3D visualization of clusters

Bibliography

[1] C. Bajaj, V. Pascucci, A. Shamir, R. Holt, and A. Netravali. Dynamic maintenance and visualization of molecular surfaces. *Dis. App. Math.*, 127(1):23–51, 2003.

[2] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. R. Srinivasan, and B. Schneider. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, 63:751–759, 1992.

[3] N. Boutonnet, M. Rooman, and S. Wodak. Automatic analysis of protein conformational changes of by multiple linkage clustering. *Journal of Molecular Biology*, 253(4):633–647, 1995.

[4] D. CM and P. AM. Stepping through an rna structure: a novel approach to conformational analysis. *Journal of Molecular Biology*, 284:1465–1478, 1998.

[5] D. CM, W. LM, and P. AM. Rna structure comparison, motif search and discovery using a reduced representation of rna conformational space. *Nucleic Acids Research*, 31:4755–4761, 2003.

[6] C. Duartes. Primos. software:www.pylelab.org.

[7] R. A. Engh and R. Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallographica Section A*, 47(4):392–400, July 1991.

[8] F.C.Bernstein, T.F.Koetzle, G.J.B.Williams, E. Jr, M.D.Brice, J.R.Rodgers, O.Kennard, T.Shimanouchi, and M.Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[9] R. Garrett and C. Grisham. *Biochemistry*. Saunders Collge Publishing, New York, 2nd edition, 1999.

[10] N. Go. A theorem on amplitudes of thermal atomic fluctuations in large molecules assuming specific conformations calculated by normal mode analysis. *Biophysical Chemistry*, 35(1):105–112, January 1990.

[11] E. Hershkovitz, E. Tannenbaum, S. B. Howerton, A. Sheth, A. Tannenbaum, and L. D. Williams. Automated identification of rna conformational motifs: theory and application to the hm lsu 23s rrna. *Nucleic Acids Res.*, 31(21):6249–6257, November 2003.

[12] L. Holm and C. Sander. Parser for protein folding units. *Proteins*, 19(3):256–268, July 1994.

[13] H. R. Horton, L. A. Moran, R. S. Ochs, D. J. Rawn, and K. G. Scrimgeour. *Principles of Biochemistry*. Prentice Hall, 3rd edition, July 2002.

[14] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins: Structure, Function, and Genetics*, 44:150–165, 2001.

[15] W. JM, L. SC, R. JS, and R. DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol.*, 284(4):1735–1747, January 1999.

[16] W. JM, L. SC, L. TH, T. HC, Z. ME, P. BK, R. JS, and R. DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol.*, 285(4):1711–1733, January 1999.

[17] L. KalÃ©, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. Namd2: greater scalability for parallel molecular dynamics. *J. Comput. Phys.*, 151(1):283–312, 1999.

[18] O. Keskin, R. L. Jernigan, and I. Bahar. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophysical Journal*, 78(4):2093–2106, April 2000.

[19] A. Kitao and N. Go. Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology*, 9(2):164–169, 1999.

[20] B. Lee and F. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, 55(3):379–400, February 1971.

[21] J. Leech, J. Prins, and J. Hermans. Smd: Visual steering of molecular dynamics for protein design. *IEEE Computational Science and Engineering*, 3(4):38–45, 1996.

[22] A.-J. Li and R. Nussinov. A set of van der waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins: Structure, Function, and Genetics*, 32(1):111–127, 1998.

[23] J. Ma and M. Karplus. The allosteric mechanism of the chaperonin groel: A dynamic analysis. *Proceedings of the National Academy of Sciences USA.*, 95(15):8502–8507, July 1998.

[24] V. N. Maiorov and R. A. Abagyan. A new method for modeling large-scale rearrangements of protein domains. *Proteins*, 27:410–424, 1997.

[25] L. J. Murray, W. B. A. III, D. C. Richardson, and J. S. Richardson. Rna backbone is rotameric. *Proceedings of the National Academy of Sciences U S A*, 100(24):13904–13909, November 2003.

[26] V. L. Murthy, R. Srinivasan, D. E. Draper, and G. D. Rose. A complete conformational map for rna. *Journal of Molecular Biology*, 291(2):313–327, August 1999.

[27] J. Park, S. Teichmann, T. Hubbard, and C. Chothia. Intermediate sequences increase the detection of homology between sequences. *Journal of Molecular Biology*, 273(6):349–354, 1997.

[28] M. R and Y. N. Backbone conformation in nucleic acids: an analysis of local helicity through heminucleotide scheme and a proposal for a unified conformational plot. *Journal Biomolecular Structural Dynamics*, 3(1):127–144, August 1985.

[29] F. Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6:151–176, June 1977.

[30] T. Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002.

[31] B. Schneider, Z. MorÃ¡vek1, and H. M. Berman. Rna conformational classes. *Nucleic Acids Research*, 32(5):1666–1677, 2004.

[32] A. S. Siddiqui and G. J. Barton. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, 4(5):872–884, May 1995.

[33] R. D. Skeel, I. Tezcan, and D. J. Hardy. Multiple grid methods for classical molecular dynamics. *Journal of Computatioanl Chemistry*, 23(6):673–684, 2002.

[34] J. E. Stone, J. Gullingsrud, and K. Schulten. A system for interactive molecular dynamics simulation. In *Proceedings of the 2001 symposium on Interactive 3D graphics*, pages 191–194. ACM Press, 2001.

[35] F. Tama, F. X. Gadea, O. Marques, and Y. H. Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41(1):1–7, October 2000.

[36] M. Teodoro, J. G. N. Phillips, and L. E. Kavraki. Molecular docking: A problem with thousands of degrees of freedom. In *Proc. of the 2001 IEEE International Conference on Robotics and Automation (ICRA 2001)*, pages 960–966, Seoul, Korea, May 2001. IEEE press.

[37] M. L. Teodoro, G. N. Phillips, Jr., and L. E. Kavraki. A dimensionality reduction approach to modeling protein flexibility. In *Proceedings of the sixth annual international conference on Computational biology*, pages 299–308. ACM Press, 2002.

[38] M. VL and R. GD. Rnabase: an annotated database of rna structures. *Nucleic Acids Research*, 31(1):502–504, 2003.

[39] M. H. Zehfus and G. D. Rose. Compact units in proteins. *Biochemistry*, 25(19):5759–5765, September 1986.

[40] M. Zhang and L. E. Kavraki. A new method for fast and accurate derivation of molecular conformations. *Journal of Chemical Information and Computer Sciences*, 42(1):64–70, 2002.