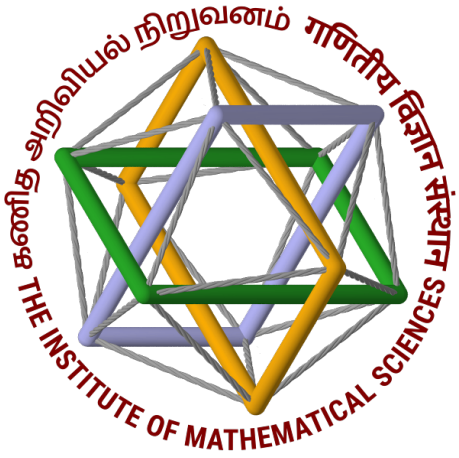
Finding Distance and Orientation Dependent Statistical Potential for Protein–Nucleic Acid Interactions

A Report Submitted in Fulfilment of the Requirement for the Summer Internship

By
Arkoparno Das



Under the supervision of
Dr. Debayan Chakraborty
Professor, Theoretical Physics
The Institute of Mathematical Sciences (IMSc), Chennai
IV Cross Road, CIT Campus, Taramani, Chennai 600 113, Tamil Nadu, India

FROM
Bidhannagar College, West Bengal State University
City Center, 3rd Ave, EB-2, Sector 1, Bidhannagar, Kolkata, West Bengal 700064

AND
INDIAN INSTITUTE OF TECHNOLOGY – MADRAS (IIT–Madras)
Indian Institute Of Technology, Chennai, Tamil Nadu 600036

Contents

Abstract

Protein–nucleic acid interactions play an important role in processes such as gene regulation, DNA repair, and RNA processing. Understanding the geometry and energetics of these interactions can help improve computational models for docking, binding site prediction, and functional annotation. In this project, we focus on developing a distance- and orientation-dependent statistical potential for protein–DNA and protein–RNA complexes. To ensure unbiased statistics, a non-redundant dataset of high-resolution structures was curated from the Protein Data Bank, with redundancy reduction performed using sequence clustering tools. For each protein residue and nucleotide, centers of mass were calculated for key structural components—protein backbone and side chain, and nucleotide phosphate, sugar, and base. Using these centroids, pairwise distances and orientation angles (θ, ϕ) were computed for all residue–nucleotide contacts within a defined cutoff. The resulting distributions were normalized and transformed into statistical potentials through Boltzmann inversion, generating energy landscapes that capture both spatial and angular preferences of interaction. While this report presents the methodology and preliminary visualizations, the approach lays the foundation for a more comprehensive potential that could be applied in molecular modeling and bioinformatics pipelines.

Introduction

Protein–nucleic acid complexes (e.g., protein–DNA/RNA assemblies) are central to gene regulation, replication, transcription, and repair (5). Computational modeling of these complexes remains a challenge because it must account for both the chemical specificity and the precise geometrical complementarity of the interface (3). In practice, empirical (knowledge-based) statistical potentials derived from databases of known structures provide effective scoring functions for biomolecular binding and folding (3). These potentials are based on the observed frequency of structural features that follow a Boltzmann-like distribution, such that the potential energy U relates observed contact probabilities to effective free energies (3). In this way, millions of protein and protein–nucleic acid complexes in the Protein Data Bank (PDB) can be mined to extract pseudo-potentials that correlate with binding affinity and structural quality (3).

However, most conventional potentials depend only on pairwise distance. Protein–nucleic acid interfaces are inherently directional: precise alignment of hydrogen-bond donors/acceptors and aromatic stacking angles is critical for specificity. Incorporating angular orientation into the potential can therefore capture interaction preferences that distance alone misses. Indeed, recent orientation-sensitive models have outperformed distance-only scores. For instance, Zhou and Skolnick (4) introduced a generalized orientation-dependent all-atom potential and showed it far outperforms its distance-only counterpart (DFIRE) in recognizing native protein folds. Likewise, Takeda et al. (2) developed a residue-level orientation-dependent potential for transcription factor–DNA docking, reporting significantly higher docking accuracy when (θ, ϕ) angles are included. These results suggest that a statistical potential $U(r, \theta, \phi)$, depending on both intermolecular distance r and orientation angles θ, ϕ, can more faithfully model the geometry of protein–DNA/RNA binding.

In this work, we will build a large, non-redundant dataset of protein–DNA/RNA complexes, compute the centers of mass (centroids) of protein residues (backbone/side chain) and nucleotide fragments (phosphate, sugar, base), and measure all pairwise distances r and angles (θ, ϕ) between centroids across the interface. From these statistics, we will

compute distribution functions $g(r)$, $g(\theta)$, $g(\phi)$, and their joint distribution $g(r, \theta, \phi)$, and apply Boltzmann inversion to obtain a three-dimensional statistical potential. The workflow follows established methods for knowledge-based potentials, with the novel emphasis on orientation. Throughout, we will filter our dataset by sequence identity (e.g., using MMseqs2) to ensure non-redundancy.

Data Collection

Protein–nucleic acid complex structures were obtained from the Protein Data Bank (PDB) (https://www.rcsb.org), which serves as the primary repository of experimentally determined 3D biomolecular structures. Initially, a broad dataset of more than 3000 protein–DNA and protein–RNA interaction structures was downloaded. In addition, reference structures of all 20 nucleosomes were downloaded from the PDB for residue geometry checks and centroid calculations. Only X-ray crystallographic structures with resolution < 2.0 Å were retained, ensuring high coordinate accuracy.

For the proof-of-concept phase, a smaller and more homogeneous subset was used: approximately 20 nucleosome core particle structures (histone–DNA complexes) were selected as the representative dataset. These complexes were chosen because they are well-characterized, high-resolution examples of stable protein–DNA interactions, and they provide a consistent framework for initial analysis.

Redundancy Filtering

To minimize statistical bias arising from overrepresented protein and nucleic acid families, a redundancy reduction step was implemented. MMseqs2 (Many-against-Many Sequence Searching) was selected for clustering due to its high computational efficiency and scalability, offering significant performance advantages over traditional tools such as CD-HIT and UCLUST.

All protein and nucleic acid sequences were first retrieved in FASTA format from the PDB using the RCSB API and Biopython scripts. Clustering was then performed separately for protein chains and nucleic acid chains to ensure that both components of the complexes were independently non-redundant.

An iterative clustering approach was adopted, running MMseqs2 over identity thresholds ranging from 1% to 100% to systematically evaluate the effect of different cutoffs on dataset size.

Following the complete 1%–100% clustering sweep, a 30% sequence identity cutoff was selected for the final dataset. The selection was based on the following rationale:

- **Avoiding the "twilight zone" bias:** Below 25% sequence identity, structural similarity may still persist, but sequence similarity becomes unreliable for accurate grouping. The 30% threshold is widely accepted in structural bioinformatics as a safe limit for generating non-redundant datasets.
- **Maximizing structural diversity:** At higher cutoffs (e.g., 50% or 70%), the dataset was dominated by similar histone variants, leading to reduced diversity. The 30% cutoff retained a broader range of unique structures.

After redundancy filtering at 30% identity, the dataset size was reduced accordingly for downstream analysis.

Computing Residue Centroids

To enable a simplified geometric representation of protein–nucleic acid complexes, each residue or nucleotide fragment was reduced to one or more centroid points representing its major functional groups. This abstraction facilitates downstream analyses such as contact detection, interaction geometry, and statistical modeling, while reducing computational complexity.

Protein Residues

Two centroids were computed for each protein residue:

- **Backbone centroid (bb)**: Average coordinates of backbone atoms N, Cα, C, and O.
- **Side-chain centroid (sc)**: Average coordinates of all remaining heavy atoms in the residue (excluding backbone atoms). For glycine, which lacks a side chain beyond H, a null or placeholder value was assigned for the side-chain centroid.

Nucleotide Residues

Three centroids were computed for each nucleotide residue, following functional partitioning commonly used in DNA/RNA binding studies:

- **Phosphate (P)**: Includes the phosphorus atom and its directly bonded oxygen atoms.
- **Sugar (S)**: Includes all carbon and oxygen atoms forming the sugar ring.
- **Base (B)**: Includes all nitrogen and carbon atoms forming the aromatic base (purine or pyrimidine).

Methodology

1. Each PDB file was parsed using Biopython to extract atomic coordinates and residue-level metadata (chain ID, residue name, residue ID).
2. Residues were classified as amino acids or nucleotides based on standard three-letter or one-letter codes.
3. Coordinates for each group were averaged to compute the center of mass (COM).
4. All calculations were performed using NumPy for vectorized efficiency.

All COM data were stored in `pdb_protein_COM.csv` and `pdb_dna_COM.csv` for downstream analysis.

Defining Interaction Pairs

After computing centroids, all potential protein–nucleotide interaction pairs were enumerated. Each interaction consisted of:

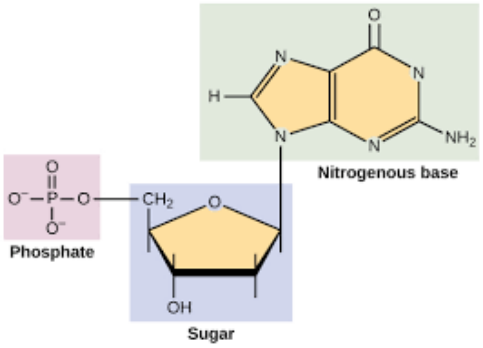
- **Protein residue:** identified by chain ID, residue number, and centroid type (backbone "bb" or side-chain "sc").
- **Nucleotide:** identified by chain ID, residue number, and centroid type (phosphate "P", sugar "S", or base "B").

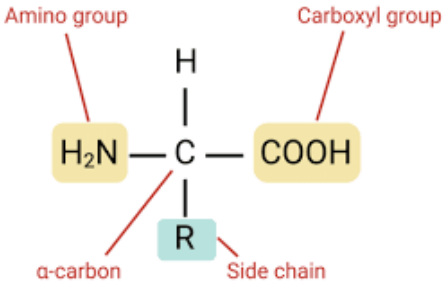
For each complex, every protein residue centroid was paired with every nucleotide centroid within the same structure. The interaction type was assigned by centroid categories (e.g., `bb_to_P`, `sc_to_B`).

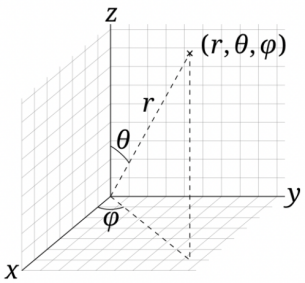
Unlike approaches such as GOAP or ANDIS, which define atom-wise local reference frames for each residue or base, this study employed a global reference frame based on the vector from the protein centroid to the nucleotide centroid. This simplification maintains biologically meaningful spatial relationships while reducing computational complexity.

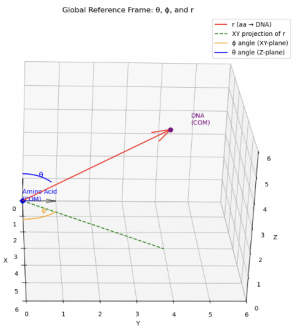
All selected pairs were recorded in `interactions.csv` with the following columns:

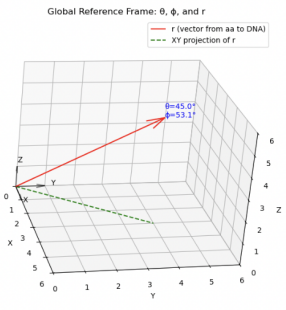
```
pdb_id, aa_type, aa_resid, aa_chain, dna_type, dna_resid, dna_chain,
interaction_type
```











Computing Distance and Orientations

Data Preparation

The list of interaction pairs (`interactions.csv`) was loaded alongside the precomputed centroid tables (`pdb_protein_COM.csv` and `pdb_dna_COM.csv`) using Pandas.

Vector and Distance Calculation

For each interaction:

1. Retrieve the centroid coordinates based on the interaction type (e.g., `bb_x, bb_y, bb_z` for backbone vs. `P_x, P_y, P_z` for phosphate).
2. Compute the vector:
$$\vec{v} = DNA_{COM} - Protein_{COM}$$
3. Compute the Euclidean distance:
$$r = \sqrt{v_x^2 + v_y^2 + v_z^2}$$
4. Discard pairs with $r = 0$ or invalid (NaN) coordinates.

Orientation Angles

Two orientation angles were calculated for each valid pair:

- **Polar angle** (θ) — inclination from the z-axis:
$$\theta = \arccos\left(\frac{v_z}{r}\right)$$
Values were clipped to $[-1, 1]$ before applying arccos, then converted to degrees (0°–180°).
- **Azimuthal angle** (ϕ) — rotation in the xy-plane:
$$\phi = \text{atan2}(v_y, v_x) \mod 2\pi$$
Converted to degrees (0°–360°).

Output

The resulting dataset (`interaction_orientations.csv`) contained the following fields:

`pdb_id, aa_type, aa_resid, aa_chain, dna_type, dna_resid, dna_chain, interaction_type, r,` θ, ϕ

Results: Histogram and Potential Calculation

To quantify the spatial preferences of protein–nucleotide interactions, all recorded interaction pairs were aggregated, and distribution functions along with corresponding potentials of mean force (PMFs) were computed. Each amino acid–nucleotide–centroid interaction type (e.g., `Ala--A, bb_to_P`) was treated independently.

Binning Strategy

- **Radial distance** (r): Range 0–150 Å, divided into 151 bins ($\Delta r \approx 1$ Å).
- **Polar angle** (θ): Range 0°–180°, divided into 36 bins ($\Delta\theta = 5°$).
- **Azimuthal angle** (ϕ): Range 0°–360°, divided into 72 bins ($\Delta\phi = 5°$).

The coordinate system is global, with the vector

$$\vec{v} = \mathrm{DNA}_{\mathrm{COM}} - \mathrm{Protein}_{\mathrm{COM}}$$

defined for each interaction pair.

Counting and Normalization

Histogram counts $n(r)$, $n(\theta)$, and $n(\phi)$ were computed for each interaction subset using `numpy.histogram`.

Radial distribution $g(r)$:

$$g(r) = \frac{n(r)}{4\pi r^2 \, \Delta r \, \rho}$$

where ρ is the average particle density for the given interaction set.

Polar distribution $g(\theta)$: Normalized by the Jacobian factor $\sin\theta$ to account for the solid angle:

$$g(\theta) = \frac{n(\theta)}{2\pi \, \sin\theta \, \Delta\theta}$$

Azimuthal distribution $g(\phi)$: Uniform normalization over $[0, 2\pi)$:

$$g(\phi) = \frac{n(\phi)}{\Delta\phi}$$

Boltzmann Inversion

Potentials of mean force were computed via Boltzmann inversion:

$$U(r) = -k_B T \ln g(r), \quad U(\theta) = -k_B T \ln g(\theta), \quad U(\phi) = -k_B T \ln g(\phi)$$

where:

$$k_B T \approx 0.593 \text{ kcal/mol} \quad at \quad T = 300 \text{ K}.$$

Empty bins ($g = 0$) were clipped or shifted to avoid $-\infty$ values.

Combined Potential

Assuming additivity of radial, polar, and azimuthal contributions:

$$U_{\text{total}}(r, \theta, \phi) = U(r) + U(\theta) + U(\phi)$$

This potential was stored as a 3D grid and later flattened for visualization.

Implementation

- **Data source:** `interaction_orientations.csv` containing all r, θ, and ϕ values.
- **Processing:** Python scripts (`NumPy`, `Pandas`) for histogramming and PMF calculation.
- **Output:** Potential tables for each interaction type, saved for downstream modeling.

Visualization

To interpret the computed potentials of mean force, visual plots were generated for every amino acid–nucleotide–centroid interaction type. For each interaction set, the following visualizations were produced:

Plot Types

1. 2D Heatmaps (three per interaction type):

- $U(r, \theta)$ at fixed ϕ — reveals preferred distance–polar angle combinations.
- $U(r, \phi)$ at fixed θ — highlights azimuthal orientation preferences.
- $U(\theta, \phi)$ at fixed r — visualizes directional tendencies in angular space.

All heatmaps used a logarithmic color scale (`LogNorm` in `Matplotlib`) to enhance contrast in low-energy regions.

Axes:

- r: radial distance in Å.
- θ: polar angle in degrees (0°–180°).
- ϕ: azimuthal angle in degrees (0°–360°).
- Energy: kcal/mol.

2. 3D Scatter Plots:

- Each point represents a triplet (r, θ, ϕ) for a given interaction type.
- Points are colored by the total potential U_{total} to illustrate the full spatial energy landscape.
- These plots are useful for spotting clusters of favorable geometries.

Scale of Analysis

We analyzed:

$$20\ aminoacids \times 4\ nucleotides \times 6\ interactioncategories \times 4\ plottypes$$

$$= 1{,}920\ plots\ in\ total.$$

This exhaustive coverage ensures that all interaction types are represented in multiple geometric views.

Repository and Data

All code, scripts, and processed CSV outputs will be made publicly available in a GitHub repository for transparency and reproducibility.

Repository (placeholder): https://github.com/Arkoparno/Protein-nucleic-acid-potentials.git

References

Huang, S.-Y., and Zou, X. A knowledge-based scoring function for protein–RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Research*, 42(7):e55, 2014. 10.1093/nar/gku077.

Takeda, T., Corona, R. I., and Guo, J. A knowledge-based orientation potential for transcription factor–DNA docking. *Bioinformatics*, 29(3):322–330, 2013. 10.1093/bioinformatics/bts699.

Yu, Z., Yao, Y., Deng, H., and Yi, M. ANDIS: An atomic angle- and distance-dependent statistical potential for protein structure quality assessment. *BMC Bioinformatics*, 20(1):299, 2019. 10.1186/s12859-019-2898-y.

Zhou, H., and Skolnick, J. GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical Journal*, 101(8):2043–2052, 2011. 10.1016/j.bpj.2011.09.012.

Zhu, X., Liu, L., He, J., Fang, T., Xiong, Y., and Mitchell, J. C. iPNHOT: A knowledge-based approach for identifying protein–nucleic acid interaction hot spots. *BMC Bioinformatics*, 21(1):289, 2020. 10.1186/s12859-020-03636-w.