Pocket Seeker

Binding Site Predictor

Supporting Information Report

Marc Arreaza

Marta Pérez

A report submitted in partial fulfilment of the requirements of
the University of Pompeu Fabra for the master's degree of
Bioinformatics in Health Sciences in *Structural Bioinformatics and Introduction to Python courses*

April 18, 2025

Chapter 1

Dataset Description and Benchmarking

PocketSeeker is a machine learning model based on a random forest algorithm, trained on a curated dataset of 103 protein-DNA complex structures derived from the sc-PDB database. This model leverages structural features of proteins to predict potential ligand-binding sites.

The dataset used for training and testing PocketSeeker consists of 103 protein-DNA complexes, as listed in Table 1.1. It is divided into two subsets: 75% of the data is used for training the model, while the remaining 25% serves as an independent test set. The PDB IDs of the protein structures in the dataset were obtained from the Protein Data Bank (PDB).

The training set is used to build the random forest model, while the test set is employed to evaluate the model's performance on unseen data. This split ensures that the model's ability to generalize and accurately predict binding pockets in novel structures is properly assessed.

Table 1.1: The dataset of 103 protein-DNA complexes with PDB ID listed. PDB ID is from the Protein Data Bank (PDB).

Training dataset			
1AAY	1AIS	1AZ0	1B3T
1BF5	1BHM	1CKQ	1CKT
1ECR	1EWQ	1FOS	1G9Z
1J5N	1JMC	1MHT	1QBJ
1QRV	1QZG	1RUN	1TRO
1TW8	2A0I	2BPF	2CCZ
2ERE	2FQZ	2MXF	2NTZ
2VLA	2VS7	2VYE	2XRO
3BS1	3GPX	3IV5	3KDE
3KTQ	3NH0	3OD8	3ODC
3OSG	3PVV	3QMG	3RN2
3SPD	3UFD	3WPC	3WTS
4A15	4ATK	4B5F	4BNC
4GZN	4HF1	4HN5	4KMF
4LLN	4M9E	4MHG	4QJU
4QTJ	4R56	4RDU	4RKG
4S04	4WCG	4XQ8	4ZSF
5CO8	5DFF	5DWB	5EXH
5EYO	5F55	5KUB	5SZX
5T5C	1A73	1BPX	1CMA
1GXP	1N6Q	1PNR	1TN9
2G1P	2UZK	2VWJ	3IMB
3NCI	3OOL	3QFQ	3SM4
3UPU	4CH1	4HQB	4L5S
4LJR	4TMU	4Y5W	5AWH
5FD3			
Independent test dataset			
5HRI	5TWP		

Chapter 2

Features Description

To effectively characterize the structural and physicochemical properties of ligand-binding sites, PocketSeeker utilizes a comprehensive set of 48 features derived from each residue. These features are selected to capture relevant biological and geometric information necessary for accurate prediction. They are grouped into eight categories based on their biological role and computational derivation: Coevolution, Secondary Structure, Accessible Surface Area (ASA), Physicochemical Properties, Solvent Exposure, PSSM (Position Specific Scoring Matrix), Concavity, and Distance to Core.

Table 2.1: The total 131 features used in this study, grouped by category.

Feature Symbol	Description	Category
Na	Number of atoms	Physicochemical
Nec	Number of electrostatic charges	Physicochemical
pI	Isoelectric point	Physicochemical
Mass	Molecular mass	Physicochemical
Enc	Expected Number of Contacts	Physicochemical
Hidrofobicity	Hydrophobicity	Physicochemical
Polarity	Polarity	Physicochemical
Center_of_mass_x	Center of mass (x-axis)	Physicochemical
Center_of_mass_y	Center of mass (y-axis)	Physicochemical
Center_of_mass_z	Center of mass (z-axis)	Physicochemical
A	Probability of Alanine	PSSM
R	Probability of Arginine	PSSM
N	Probability of Asparagine	PSSM
D	Probability of Aspartic acid	PSSM
C	Probability of Cysteine	PSSM
Q	Probability of Glutamine	PSSM
E	Probability of Glutamic acid	PSSM
G	Probability of Glycine	PSSM
H	Probability of Histidine	PSSM
I	Probability of Isoleucine	PSSM
L	Probability of Leucine	PSSM
K	Probability of Lysine	PSSM
M	Probability of Methionine	PSSM
F	Probability of Phenylalanine	PSSM
P	Probability of Proline	PSSM
S	Probability of Serine	PSSM
T	Probability of Threonine	PSSM
W	Probability of Tryptophan	PSSM
Y	Probability of Tyrosine	PSSM
V	Probability of Valine	PSSM
MI	Mutual information score	Coevolution
DI	Direct information score	Coevolution
SASA	Solvent-accessible surface area (Shrake-Rupley method)	ASA
Total_SASA	Total SASA of all atoms in the residue	ASA
Main-chain_SASA	SASA contributed by backbone atoms (N, C_α, C, O)	ASA
Side-chain_SASA	SASA from side chain atoms unique to each amino acid	ASA
Polar_SASA	SASA from polar atoms (e.g., O, N)	ASA
ASA	Relative solvent accessibility (DSSP-based method)	ASA
Non-polar_SASA	SASA from non-polar atoms (mainly carbon)	ASA
SS	Secondary structure (e.g., helix, sheet, coil)	Secondary Structure
Phi	Phi torsion angle	Secondary Structure
Psi	Psi torsion angle	Secondary Structure
Theta(i-1⇒i+1)	Local backbone angle	Secondary Structure
Tau(i-2⇒i+2)	Extended backbone angle	Secondary Structure
Residue_Depth	Residue depth from surface	Solvent exposure
CN	Contact number	Solvent exposure
Distance_to_Core	Distance to Core	Distance to Core
Concavity	Concavity	Concavity