# *Protein fold recognition*

## Introduction

Fold recognition, one assigns a probe amino acid sequence of unknown structure to one of a library of target 3D structures. Correct assignment depends on effective scoring of the probe sequence for its compatibility with each of the target structures. Here we show that, amino acid sequence of the probe, sequence-derived properties of the probe sequence (such as the predicted secondary structure) are useful in fold assignment. The additional measure of compatibility between probe and target is the level of agreement between the predicted secondary structure of the probe and the known secondary structure of the target fold.

## What is protein fold recognition?

Protein threading, also known as fold recognition, is a method of protein modeling which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure. It differs from the homology modeling method of structure prediction as it (protein threading) is used for proteins which do not have their homologous protein structures deposited in the Protein Data Bank (PDB), whereas homology modeling is used for those proteins

## Why it is important ?

- Disease discovery
- Respiratory drug discovery

## How do computer scientists predict protein by fold recognition as mentioned in the different research papers?

- The construction of a structure template database
- The design of the scoring function
- Threading alignment
- Threading prediction

**Methodology:**

| Papers | Datasets | Feature extraction techniques | Algorithms | Evaluation | Tools | summary |
|---|---|---|---|---|---|---|
| **Improving protein fold recognition and template-based modeling by employing probabilistic-based matching**<br><br>**https://academic.oup.com/bioinformatics/article/27/15/2076/404317**<br><br>**ID:011123020** | SALIGN Benchmark<br>For alignment accuracy<br><br>Lindahl and Scop for fold recognition | HHPRED and Boosthreader. | Smith walterman | Domingues,F.S. *et al.* (2000) Structure-based evaluation<br><br>performance of SPARKS-X is consistently better than that of HHPRED | SPARK-X | We found that consensus prediction from multiple fold recognition servers, the use of multiple templates and model refinement. |
| **Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition**<br><br>**http://www.mdpi.com/1422-0067/17/12/2118/htm**<br><br>**ID: 011123020** | **PSI BLAST** | Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs.<br><br>A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition | Ensemble classifier.<br><br>Single classifier. | we evaluate and compare the recognition performance of existing methods used in the last 10 years on a benchmark dataset | PFPA method is developed and freely available at http://server.malab.cn/PFPA/index.html | First, they demonstrate accurate, robust, and reliable performance. Second, constructing informative and effective prediction engines remains a great challenge . |

| Papers | Datasets | Feature extraction technique | Algorithms | Evaluation | Tools | Summary |
|---|---|---|---|---|---|---|
| **Protein fold recognition using geometric kernel data fusion**<br><br>https://academic.oup.com/bioinformatics/article/30/13/1850/2422171<br><br>**Id:011121090** | SCOP PDB-40D benchmark Dataset<br><br>Ding and Dubchak (DD) (Ding and Dubchak, 2001) | A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition<br><br>. | The approximate AGH mean algorithm | Improving the accuracy of psi-blast protein database searches with composition-based statistics<br><br>Improving taxonomy-based protein fold recognition by using global and local features | geometric kernel fusion frameworks are publicly available at http://people.cs.kuleuven.be/*raf.vandebril/homepage/software/geomean.php?menu¼5/ | Understanding the relationship between primary and tertiary structure in proteins is one of the main objectives of protein sequence analysis. |
| **Protein folds and protein folding**<br><br>**Id:011121090** | NMR (Kazmirski et al., 2001). Early work on CI2 established the benchmarks | similar functions and structural features suggest a common origin; and the fold, wherein families and superfamilies with conserved core topological arrangement are grouped.<br><br>SCOP folds are also grouped by secondary structure class: all-a, all-b, a/b | Domain partitioning algorithm<br><br>The structural comparison algorithm used in the initial releases of CATH, | partitioned by expert duration have been compared both with each other and with domain boundaries generated by the crystallographers for the structure | PDB:1AU7 | protein folding must move beyond model systems. By coupling the broad-based knowledge of protein domain partitioning and fold classification with high throughput MD simulations, |

| Papers | Datasets | Feature extraction techniques | Algorithms | Evaluation | Tools | summary |
|---|---|---|---|---|---|---|
| **Performance of Machine Learning Techniques in Protein Fold Recognition Problem**<br><br>**http://ieeexplore.ieee.org/abstract/document/5480307/**<br><br>**Id: rumi** | SCOP dataset | The classification of proteins in SCOP database has been carried out by visual inspection and comparison of structures. | SVM and MLP(on 10 fold-cross validation) | MLP gives better classification accuracy for individual fold of the protein feature than Support Vector Machine (SVM) | | -Compared the classification performance of Support Vector Machine (SVM) and Multilayer Perceptron (MLP) on the SCOP dataset;<br>-Unique learning capability of the Neural Nets;<br>-In future, intend to investigate the classification model of multilayer perceptron, support vector machine and other techniques |
| **ımportance of Dimensionality Reduction in Protein Fold Recognition**<br><br>**http://ieeexplore.ieee.org/abstract/document/7476132/**<br><br>**ID: rumi** | DD, EDD,TG | DRLDA, FNLDA, EFR and MLDA to reduce the dimension of n-grams. | kNN(k=5) | The interaction of amino acids does contain useful information provided it is utilized in an appropriate way. However, thechallenge is to develop a suitable DRT and/or a classifier that can efficiently estimate parameters for recognition . | | - interaction between amino acids<br>- high dimensional features using quad-grams<br>- dimensionality reduction techniques to show usefulness of extracted features. |

| Papers | Datasets | Feature extraction techniques | Algorithms | Evaluation | Tools | summary |
|---|---|---|---|---|---|---|
| **A Segmentation-Based Method to Extract Structural and Evolutionary Features for Protein Fold Recognition**<br><br>**ID : 011131057** | Four Data Sets,<br><br>1. TG<br>2. EDD<br>3. F92<br>4. F110 | SegmentationBased Feature Extraction Technique<br><br>1. Embedded in position specific scoring matrix **(PSSM)** to provide local evolutionary information<br><br>2. embedded in the predicted secondary structure of proteins using **SPINE-X** to provide structural information | 1. Support Vector Machine (SVM)<br><br>2. Grid Search Algorithm (for proposed feature vector to calculate SVM parameters) | 10-Fold Cross Validation Evaluation Method | | Increasing the number of folds the complexity of the problem is increasing and therefore, more discriminatory information is required to tackle this problem. |
| **Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models**<br><br>**ID : 011131057** | Main Three Data Sets,<br><br>1. DD (Ding and Dubchak)<br><br>2. EDD (Extended Ding and Dubchak)<br><br>3. TG (Taguchi and Gromiha)<br>Additional,<br>1. Lindahl | Extract evolutionary information (extract monogram, bigram, and trigram feature groups) using,<br><br>1. Position Specific Scoring Matrices **(PSSM)**<br>2. Remote Homology Detection Technique<br>a. **HHblits** to produce a **HMM profile**<br>b. **HMMFold** method to extract n-gram feature groups from the **HMM profile** | 1. Naive Bayes<br>2. Bayes network<br>3. K-nearest neighbor (KNN)<br>4. Support Vector Machine (SVM-using linear kernel)<br>5. Random Forest (RF)<br><br>Best result achieved from,<br><br>**Support Vector Machine (SVM-using linear kernel)** | 2-Fold Cross Validation Evaluation Method (for Lindahl)<br><br>10-Fold Cross Validation Evaluation Method (for others) | | using the HMM profile is more effective than the PSSM profile for the protein fold recognition problem and should be considered the main resource to extract sequence profiles and evolutionary information for this task. |

| Papers | Datasets | Feature extraction techniques | Algorithms | Evaluation | Tools | summary |
|---|---|---|---|---|---|---|
| **Coarse grained contact potential helps improve fold recognition sensitivity in template based protein structure modeling**<br><br>**http://ieeexplore.ieee.org/document/7723749/**<br><br>**ID: 011 141 068** | | sequence profiles; predicted secondary structures; depth-dependent structure profiles; solvent accessibility; backbone dihedral torsion angles; hydrophobic scoring matrix | Sequence Template alignment;<br><br>Pairwise Sequence Alignment;<br><br>ICOSA score;<br><br>MUSTER score;<br><br>Ranking algorithm; | Global Distance Test;<br><br>Total Score (GDT-TS) compared to the native structure | **MUSTER** https://zhanglab.ccmb.med.umich.edu/MUSTER/<br><br>**BLAST** | -Modeling protein structure<br>-Experiment determined structure<br>-most likely to resemble target sequence<br>-ICOSA Score<br>-MUSTER Score<br>-Enhanced template sequence technique<br>-Combining ICOSA and MUSTER Score<br>-Improves fold recognition accuracy and sensitivity in template based structure modeling |
| **SWISS MODEL: Modeling protein tertiary and quaternary structure using evolutionary information**<br><br>**https://academic.oup.com/nar/article/42/W1/W252/2435313**<br><br>**ID: 011 141 068** | PDB | Annotation of ligands, Oligomeric structure prediction, Modeling of ligands | Sensitive HMM | CAMEO project, GMQE, ITASSER library | **SWISS model** https://swissmodel.expasy.org<br><br>**HHblits**<br>**BLAST** | -Modeling protein tertiary and quaternary structure using evolutionary information<br>-Homologous modeling<br>-user friendly webserver<br>-automated system, 3D modeling for amino acid and sequence<br>-model quality estimation, sensitivity of predictions |

## Conclusion:

The recent growth of the number of experimentally determined protein structures in Protein Data Banks (PDB) has provided hundreds of thousands of structural templates to support reliable template-based protein structure modeling. Taking advantage of large-scale data processing platforms, sequence alignments with a large number of potential structural templates can be completed rapidly. In these research papers, various methods are tested with benchmark datasets, to improve and increase evolution of bio science. And will help further experiments.