Bioinformatics III
Structural Bioinformatics and Genome Analysis

BIOINF

Chapter 5 Homology 3D Structure Prediction

- 5.1 Introduction
- 5.2 Comparative Modeling
  - **Sequence-Sequence Comparison**
- 5.3 Threading
  - **Sequence-Structure Alignment**

Chapter 6 Ab Initio Prediction and Molecular Dynamics

- 6.1 Introduction
- 6.2 Ab Initio Methods

5. Homology 3D Structure Prediction
5.1 Introduction

- Homology search
  - Prediction of proteins 3D structure based on their primary sequence
  - The new sequence has an homolog with the same solved structure
  - Prediction of new structures

- Process of folding from amino acid sequence into a protein is poorly understood : many local effects dependent

  - Quantum mechanics: to find a minimum energy state of the amino acid sequence
  - Molecular Dynamics

5. Homology 3D Structure Prediction
5.1 Introduction

- Fold recognition/ Structure prediction
  - Sequence comparison: No 3D but databases as NR (sequence-sequence, sequence-profile, profile-profile alignments)
  - Secondary structure prediction
  - Sequence-Structure alignments / Structures comparison: **Threading** or the use of a solved 3D protein structure to search for compatibilities of sequences with known 3D folds
- Proteins have limited variety of shapes: most folds are known
  → **Comparative Modeling success**

Sequence-Sequence Comparison

To find homologies

- For high sequence similarities: Pairwise alignment methods (Waterman-Smith, FASTA, BLAST, PSI-BLAST)

- For remote homologous similarities: Alignment-based Methods and discriminative Methods (only positive examples)

  - PSI-BLAST: More than one iteration through NR, profile generated and used as template for comparing unknown structures, folds and folds classes
  - FPS: Family Pairwise Search based on BLAST (comparisons of new sequence)

Sequence-Sequence Comparison (cont.)

For remote homologous similarities

- SVMs based protein homology: relay on a kernel specially designed for protein sequences
  - Fisher kernel: HMMs and alignments
  - Mismatch kernel: sequence identities
  - SVM-Mismatch kernel applied to profiles (PSI-BLAST and NR)
  - SVM- pairwise method: SW score as the feature vector
  - SVM using the SW kernel: SW pairwise score as kernel matrix
  - SVM using Local Alignment kernel: gap penalties and BLOSUM matrices
  - SVM with LA and SW- kernels applied to profiles
  - SVM using oligomer based distances: construction of a feature space of indicative patterns (PROSITE and BLOCKS)
  - SVM-HMMSTR: profile construction from SwissProt data base

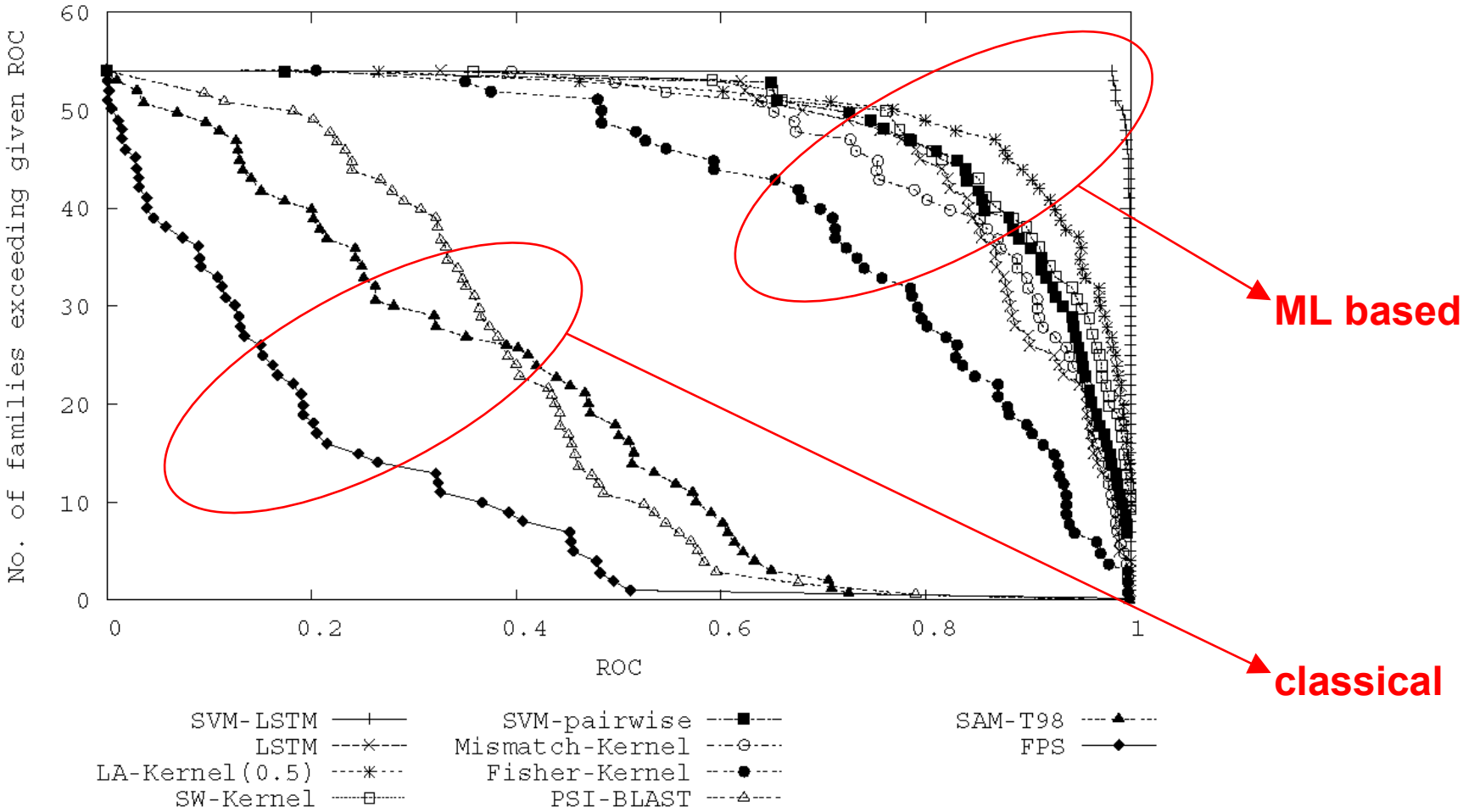
- LSTM recurrent Network

Sequence-Sequence Comparison (cont.)

- **Overview: example from the remote homology detection benchmark:** http://www.cs.columbia.edu/compbio/svm-pairwise
  - Data set: 54 superfamily tasks from SCOP with one family holdig Positive and Negative examples (in and out of belonging family)
  - Goal: Detection of examples from outside the fold
  - Quality by the area under ROC curves: values from 0.5 (random guessing) to 1.0 (perfect prediction)
  - Quality by the area under ROC50 curves: up to 50 false positives

Sequence-Sequence Comparison (cont.)

Result on benchmark data (Sensitivity Vs Specificity)



method	m	p	S	ROC	ROC50	time
(a) PSI-BLAST	+	-	-	0.693	0.264	5.5s
(b) FPS	-	-	-	0.596	–	6800s
(c) SAM-T98	+	-	-	0.674	0.374	200s
(d) Fisher	-	-	+	0.773	0.250	>2000s
(e) Mismatch	-	-	+	0.872	0.400	68 h
(f) Pairwise	-	-	+	0.896	0.464	>194h
(g) SW	-	-	+	0.916	0.585	>129h
(h) LA 1	-	-	+	0.923	0.661	550h
(h) LA 2	-	-	+	0.925	0.649	550h
(i) Oligomer	-	-	+	0.919	0.508	2000s
(j) HMMSTR	-	+	+	–	0.640	>500h
(j) Mismatch 1	-	+	+	0.974	0.756	>500h
(j) Mismatch 2	-	+	+	0.980	0.794	>500h
(j) AF-GSM	-	+	+	0.926	0.549	>620h
(j) BF-GSM	-	+	+	0.934	0.669	>620h
(j) BV-GSM	-	+	+	0.930	0.666	>620h
(j) SW-GSM	-	+	+	0.948	0.711	>620h
(j) AF-PSSM	-	+	+	0.978	0.816	>620h
(j) BF-PSSM	-	+	+	0.980	0.854	>620h
(j) BV-PSSM	-	+	+	0.973	0.855	>620h
(j) SW-PSSM	-	+	+	0.982	0.904	>620h
(k) LSTM	+	-	-	0.932	0.652	20s

Sequence-Sequence Comparison (cont.)

Profile-profile alignment performs better for homology remote detection than sequence-sequence or sequence-profile alignments

PSSM: Position Scoring Specific Matrix

GSM: Global Scoring Matrix

AF, BF and BV: All Fixed-width, Best Fixed-width and Best variable-width ω-mer

Sequence-Structure Alignments

Structure prediction from sequence or fold recognition

"..also known as fold recognition, is a method of computational protein structure prediction used for protein sequences which have the same fold as proteins of known structures but do not have homologous proteins with known structure. Protein threading predicts protein structures by using statistical knowledge of the relationship between the structure and the sequence" Wikipedia

In PDB Ratio sequence to structure 7/1 and structures submitted in the past three years have similar structural folds

Number of folds is small: Similar structures or folds do not have similar sequences

Proteins with different sequences but do fold into similar structures

5. Homology 3D Structure Prediction
5.3 Threading

BIOINF

Sequence-Structure Alignments

Dictionary of solved structures are available DSSP

- **Number of folds is limited (High chance to detect the structure of new sequence in the dictionary)**
- **Evaluate the fitness of the query sequence for each of the possible structures (SSEs matching, residue environment matching)**
- **Post-processing of the results need due to the low accuracy (50%) finding the correct fold (filtering by other predictions or known experimental data)**

Goal

From native fold approximation of the energy or part of it and comparison with the energy of the new sequence squeezed into this fold to determine if it is a suited fold for the sequence or not

"The prediction is made by "threading" each amino acid contained in the target sequence to a position in the template structure, and evaluating how well the target fits the template" **Wikipedia**

Sequence-Structure Alignments

Folds as cores or SSEs BUT not loops or turns (high variation)

Decoys generation and evaluation to fix the range of energy values for a native fold and for sequences not fitting in the fold

Decoys energy values computation to separate the native fold from similar ones :

"Energy of native fold with original sequence should be less than the energy of a random sequence"

Conformation of non-native Decoys: Parameter-Independent Decoys in which conformation pairs of torsional angles from native decoys are perturbed by

$-30^\circ \leq \Phi \leq 30^\circ$

Sequence-Structure Alignments

Computational limitations due to empirical physical energy function (water Vs molecules simulation energies)

Concepts

Energy as values based on potentials: Cβ-Cβ distances from 3 Å to 13 Å

Unknown structure: Problem sequence Target

Known structure: Template sequence

BIOINF

Threading Method design

A. Structure template database : Size and quality of the cores in the template dictionary (as high the number higher probabilities to find an existing one)

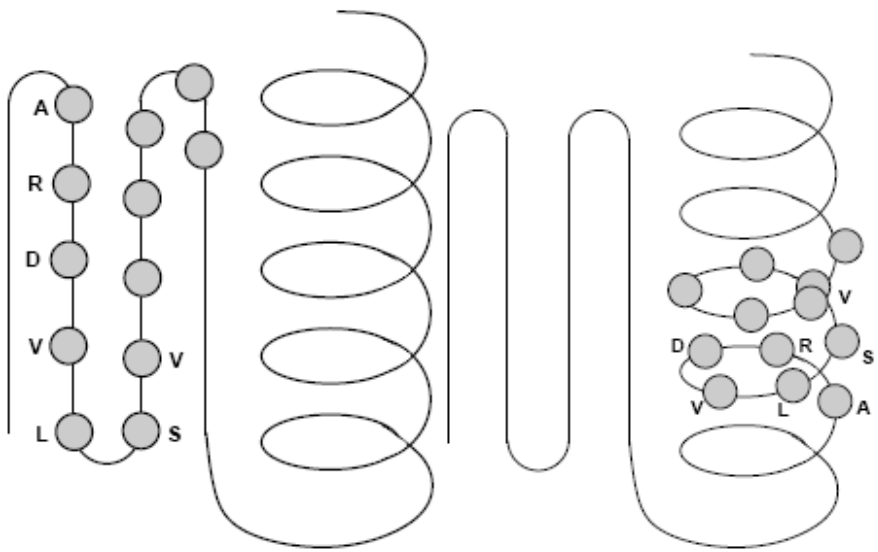
- Domains by CATH or SCOP
- Bias introduced by 3D potential function deductions
- NMR and x-ray crystallography

B. Scoring Function: Potential and energy function and how it is optimized to evaluate target fitness into the folds template

- Description of core elements: hydrophobic and hydrophilic residues, neighbor relation, number and types of contacts, environment
- Contact potentials: knowledge-based potentials and potential of mean forces
- Potentials and configuration of the query sequence to compute the energy (normalization to obtain the energy)

C. Optimization procedure to find the best fold the sequence has in the known structure

Goal is the energy function
Difficulties due to gaps (loops and turns length variability)



For pairwise contact potentials, procedure as a NP-hard:

DDP: iteratively a residue is placed in another position and all other residues are optimized for the new position

Frozen approximation: template residues are kept and new query residue is inserted

Sampling and searching methods: Gibs sampling, Monte Carlo.,

Mean field approaches and branch and bound algorithms

For singleton
procedure as sequence- sequence alignment:
alignment of new sequences to the new positions

5. Homology 3D Structure Prediction
5.3 Threading

D. Final selection of the template once the optimal energy on each structure/fold is computed

"..construct a structure model by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template" W.

By Decoys construction
Deviation of the native fold by perturbation in torsional angles of $30^\circ \leq \Phi \leq 30^\circ$
Minimizing the energy of native fold with respect the current potential function

By Z-score to measure how the energy value obtained deviates σ from the mean value μ
Mean μ and variance σ^2 should be computed
μ and σ estimated: sequences of other folds are threaded through the fold
A Gaussian distribution CAN NOT be assumed!!!!!

Energy parameter optimization

$$E = \sum_i s(a_i, p(a_i)) + \sum_i \sum_{j:i<j} S_{ij}\ c(a_i, a_j)$$

a_i, a_j amino acids positions

S_{ij} contact matrix

C_{ij} contact potential

For a single pairwise contact potential

$$E = \sum_i \sum_{j:i<j} S_{ij}\ c(a_i, a_j) \qquad E_0 < E_p$$

Z-score

$$Z = \frac{E_0 - \mu}{\sigma}$$

Decoys generated: only the μ and covariance of the contact maps have to be computed

$$\mu = \langle E \rangle = \sum_i \sum_{j:i<j} \langle S_{ij} \rangle\ c(a_i, a_j) \qquad \sigma^2 = \sum_i \sum_{j:i<j} \sum_k \sum_{l:k<l} \mathrm{cov}(S_{ij}, S_{kl})\ c(a_i, a_j)\ c(a_k, a_l)$$

Substituting E_o, μ and σ

$$Z = \frac{\sum_i \sum_{j:i<j} c(a_i, a_j) \ (S_{ij}^0 - \langle S_{ij} \rangle)}{\sqrt{\sum_i \sum_{j:i<j} \sum_k \sum_{l:k<l} \mathrm{cov}(S_{ij}, S_{kl}) \ c(a_i, a_j) \ c(a_k, a_l)}}$$

In vector notation

$$Z = \frac{\boldsymbol{c}^T \ (\boldsymbol{s}^0 - \langle \boldsymbol{s} \rangle)}{\boldsymbol{c}^T \boldsymbol{S} \boldsymbol{c}}$$

c vector with components c (a*i*, a*j*)

s vector with components S_{ij} (analog for s^0)

S covariance matrix of s

P-SVM: z-score as a classification problem with native fold as the only member of the positive class

$$Z = \frac{\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{c}}{\boldsymbol{c}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{c}}$$

Maximize Z by

$$\frac{1}{2} \boldsymbol{c}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{c} - \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{c}$$

Energy

$$E = c^T s$$

The goal is

$$E_0 - E = c^T (s^0 - s) > 0$$

can be learned by Perceptron learning rule or one-class SVM

When different sequences are used

$$E = \sum_{i=1}^{20} \sum_{j=i}^{20} S_{ij} \, c_{ij}$$

c (a*i*, a*j*) replaced by c_{ij}

S_{ij} a_i in contact with a_j

$$\mu = \langle E \rangle = \sum_i \sum_{j:i<j} \langle S_{ij} \rangle \, c_{ij}$$

$$\sigma^2 = \sum_i \sum_{j:i<j} \sum_k \sum_{l:k<l} \mathrm{cov}(S_{ij}, S_{kl}) \, c_{ij} c_{kl}$$

Chapter 6 Ab Initio Prediction and Molecular Dynamics
6.1 Introduction

Ab initio and molecular dynamics : insights into protein folding and stability

Ab Initio

Use of amino acids sequence as the ONLY input for 3D prediction
Experimental data can be included (Rosetta method)
Novel structure to be determined with no homolog known structure (no threading methods): Prediction of new structures

Molecular dynamics

Force fields not always modeled correctly
Computation of many sums over all atoms or sets of atoms
Simulation of water and its interaction with many molecules
Downscale macroscopic parameters: dielectric constant.,
No simulation of the context in the cell: chaperones not considered
Simulation in femtoseconds: gaps of 10^{12}
Computing time of 10^{12} CPU-years

BIOINF

Rosetta Method: the way to a fold protein

Local folds

- Constructed based on small fragments
- Library of 3 and 9 residues from which folds are generated
- Sequence and profile-profile method extracts the appropriate fold by sampling possible conformation by Monte Carlo approach

Scoring function

- Hydrophobic burial
- Pairwise interaction (electrostatic and disulfide bonds)
- α helix and β strand and spherical packing
- β strand packing

Improvement by

- filtering out non-plausible folds as poorly formed β strand, low contact order or packed interior
- Information from homologous sequence

- **Rigid body models:** Secondary structures are predicted and represented as rigid models where the torsion angles are only changeable at the junctions of those bodies
- **Lattice representations:** Residues are restricted to points on a regular 3D lattice
- **Potential functions:** Molecular mechanics and force fields are used but computationally expensive because water must be also modeled
- **Optimization techniques and search methods:** Energy landscape of the current conformation must be sampled (torsion angles variation, direct movements of the atoms or fragments insertions). Monte Carlo simulation, evolutionary or genetic algorithms and simulated annealing can be used. The candidate solutions are filtered and checked for plausibility. As fewer candidates to be considered more detailed the model

Threading, Ab initio

- **Performance:** Threading methods perform better being comparable methods Rosseta and Ab initio

- **Threading programs:**
  - PROSPECT [Xu and Xu, 2000]
  - Tasser
  - FAMS
  - Zhang (threading + clustering)