

Protein Threading Algorithms Used in Protein Structure Prediction

Yuting Xiao

Division of Science and Mathematics
University of Minnesota, Morris

April 15, 2017

Outline

Background

Introduction

Sequence Profile-Profile Alignment(PPAs)

Profile-Hidden Markov Models(HMMs)

Summary

Background

Amino Acids

- ▶ **Amino acids**, also called **residues**.
- ▶ 20 different amino acids
- ▶ Unique single letter
- ▶ **Primary Structure**, linear combination of amino acids
- ▶ **Secondary Structure**, natural folds
- ▶ **Tertiary Structure**, 3-D structure

Background

Protein Structure

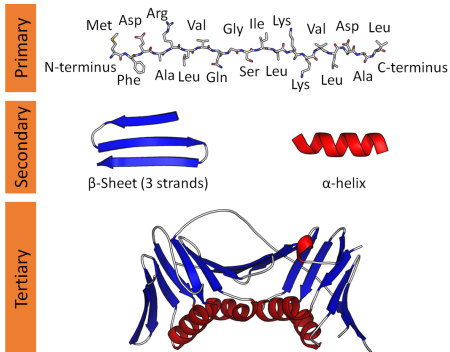


Figure: Protein Structure

Introduction

- ▶ Why predicting protein structure?
- ▶ Basic Strategy
- ▶ Template-Based Modeling

Introduction

Why predicting protein structure?

- ▶ One important topic
- ▶ Functionality is closely related to structures
- ▶ Discovering novel drugs for diseases

Introduction

Basic Strategy

- ▶ Unknown protein's primary structure (**target**)
- ▶ Currently known protein structures (**templates**).
- ▶ Constructing target's structure based on templates' structures

Introduction

Basic Strategy

- ▶ Protein Data Bank (PDB):
 - ▶ Templates
 - ▶ Coordinate Files
 - ▶ Atoms in each protein, and their 3D location in space
- ▶ Modeling Method
 - ▶ **Template-Based Modeling**
 - ▶ Free Modeling

Introduction

Template-Based Modeling

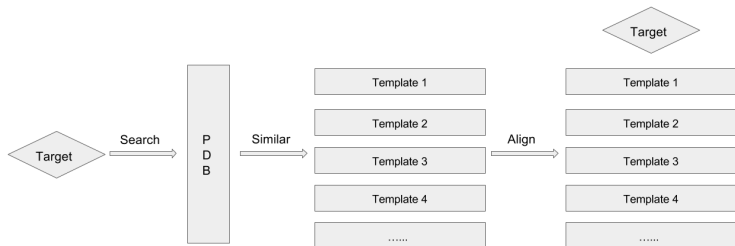


Figure: Protein Threading

Introduction

Template-Based Modeling

- ▶ Aligned Regions
- ▶ Unaligned Regions

Sequence Profile-Profile Alignment(PPAs)

- ▶ Sequence
- ▶ Pairwise Sequence Alignment
- ▶ Multiple Sequence Alignment and Profile
- ▶ PPA Program
- ▶ Improvement

Sequence Profile-Profile Alignment(PPAs)

Sequence

Sequence 1 : L E V K
Sequence 2 : L D I R
Sequence 3 : L E I K
Sequence 4 : L D V E

L --- Leucine
E --- Glutamic Acid
D --- Aspartic Acid
V --- Valine
I --- Isoleucine
K --- Lysine
R --- Arginine

Sequence Profile-Profile Alignment (PPAs)

Pairwise Sequence Alignment

- ▶ There are **many ways** to align two protein sequences, and for each amino acid pair, we can find either a **match** (blue), a **mismatch** (red) or an **insertion or deletion** ("-") represents a gap)

Index : 0 1 2 3 4

Sequence 1: L E V - K

Sequence 2: L D - I K

Figure: Pairwise Sequence Alignment

Sequence Profile-Profile Alignment(PPAs)

Pairwise Sequence Alignment

- ▶ If we adopt a scoring method for each possible alignment, the best alignment is therefore the one with the highest score.

Index :	0	1	2	3	4	
Sequence 1:	L	E	V	-	K	
Sequence 2:	L	D	-	I	K	
	+2	+0	-1	-1	+2	=2

Match +2
Mismatch 0
Gap -1

Figure: Pairwise Sequence Alignment

Sequence Profile-Profile Alignment(PPAs)

Multiple Sequence Alignment

- ▶ A **profile** is a 20 by L table of frequencies for a multiple sequence alignment with length L . Each entry $p_{i,j}$ represents the **probability** of amino acid type i occur in the j th column.
- ▶ Profile is a better representation for multiple sequence alignment.

Sequence Profile-Profile Alignment(PPAs)

Profile

Sequence 1 : L E V K
Sequence 2 : L D I R
Sequence 3 : L E I K
Sequence 4 : L D V E

L --- Leucine
E --- Glutamic Acid
D --- Aspartic Acid
V --- Valine
I --- Isoleucine
K --- Lysine
R --- Arginine

Figure: Protein Sequence Examples

Sequence Profile-Profile Alignment(PPAs)

Profile

	Index 0	Index 1	Index 2	Index 3
D	-	0.5	-	-
E	-	0.5	-	0.25
L	1	-	-	-
I	-	-	0.5	-
V	-	-	0.5	-
R	-	-	-	0.25
K	-	-	-	0.5
...

Sequence Profile-Profile Alignment(PPAs)

PPA program

- ▶ **I-TASSER**
- ▶ PPA program reduces multiple sequence alignments to pairwise alignment between profiles

Sequence Profile-Profile Alignment(PPAs)

PPA program

- ▶ Use target sequence as input, and search through PDB using PSI-BLAST

```
>NP_002583.1 proliferating cell nuclear antigen [Homo sapiens]  
MFEARLVQGSILKKVLEALKDLNEACWDISSSGVNLQSMDSHVSLVQL  
TLRSEGFDTYRCDRNLAMGVNLTSMKILKCAGNEDIITLRAEDNADTLA  
LVFEAPNQEKVSDYEMKLMDL DVEQLGIPEQEYSCVVKMPSGEFARICRD  
LSHIGDAVVISCAKDGVKFSASGELGNGNIKLSQTSNV DKEEEEAVTIEMN  
EPVQLTFALRYLNFFTKATPLSSTVTLSMSADVPLVVEYKIADMGHLKYLLA  
PKIEDEEGS
```

Sequence Profile-Profile Alignment (PPAs)

PPA program



Sequence Profile-Profile Alignment(PPAs)

PPA program

- ▶ Construct target profiles
- ▶ Align target profiles against all pre-calculated profiles in database, where each profile represents **a specific set of protein families**

Sequence Profile-Profile Alignment(PPAs)

PPA program Scoring function

- ▶ Use *dynamic programming* to find the overall best alignment

$$S(i, j) = \sum_{k=1}^{20} F_q(i, k) L_t(j, k) + c_1 \sigma[s_q(i), s_t(j)] + \text{shift}$$

Frequency of k th amino acid at the i th position of the target profiles

Frequency of k th amino acid at the j th position of the template profiles

Return 1 if target and template have the same secondary structure; Return 0 otherwise

Frequency likelihood

Secondary structure fitness

Sequence Profile-Profile Alignment(PPAs)

New Improvement in I-TASSER Suite

- ▶ Added structural environment fitness score, $E(j, AA_q(i))$
 - ▶ torsion angle
 - ▶ solvent accessibility
 - ▶ secondary structure

$$S_{\text{Env-PPA}}(i, j) = S(i, j) + c_2 E(j, AA_q(i))$$

Profile-Hidden Markov Models(HMMs)

- ▶ Structures
- ▶ Coin Toss Example
- ▶ Profile-HMM
- ▶ Pairwise Profile-HMM Alignment
- ▶ Scoring function
- ▶ Improvements

Profile-Hidden Markov Models(HMMs)

Structures

- ▶ Two layers structure:
 - ▶ Visible layer
 - ▶ Invisible layer
- ▶ **Markov chain**

Profile-Hidden Markov Models(HMMs)

Coin Toss Example

- ▶ Given two coins that has different probability of heads and tails:

	Coin A	Coin B
Head (H)	0.5	0.3
Tail (T)	0.5	0.7

- ▶ Suppose we are given an observation sequence of HHTHTH
- ▶ Without knowing which coin was used for each toss
- ▶ What would the best explanation for an observations of such sequence?

Profile-Hidden Markov Models(HMMs)

Coin Toss Example

- ▶ **Transition probabilities** are given as below:

	Coin A	Coin B
Coin A	0.9	0.2
Coin B	0.1	0.8

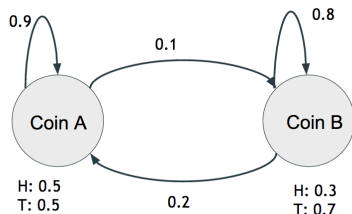
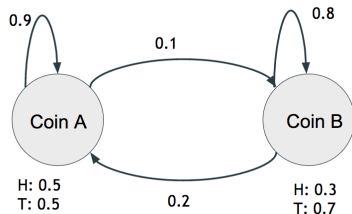


Figure: Hidden Markov Model for Coin Toss Example

Profile-Hidden Markov Models(HMMs)

Coin Toss Example



- ▶ If the coin sequence is AABAAB
- ▶ The probability for observations HHTHTH is:

$$P = \mathbf{0.5} * 0.9 * \mathbf{0.5} * 0.1 * \mathbf{0.7} * 0.2 * \mathbf{0.5} * 0.9 * \mathbf{0.5} * 0.1 * \mathbf{0.3}$$
$$= 2.12625 * 10^{-5}$$

Profile-Hidden Markov Models(HMMs)

Profile-HMM

► HHpred

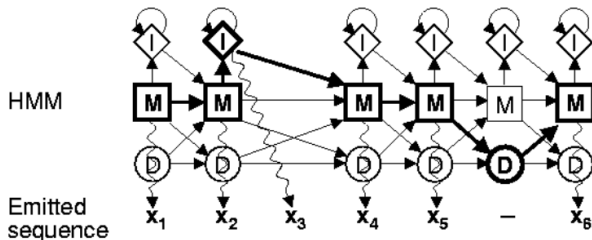


Figure: Example of a Profile-Hidden Markov Model

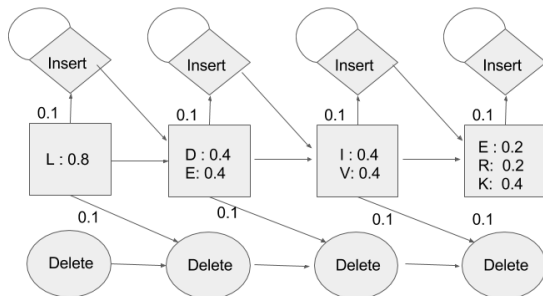
Profile-Hidden Markov Models(HMMs)

Profile

	Index 0	Index 1	Index 2	Index 3
D	-	0.4	-	-
E	-	0.4	-	0.2
L	0.8	-	-	-
I	-	-	0.4	-
V	-	-	0.4	-
R	-	-	-	0.2
K	-	-	-	0.4
Insert	0.1	0.1	0.1	0.1
Delete	0.1	0.1	0.1	0.1

Profile-Hidden Markov Models(HMMs)

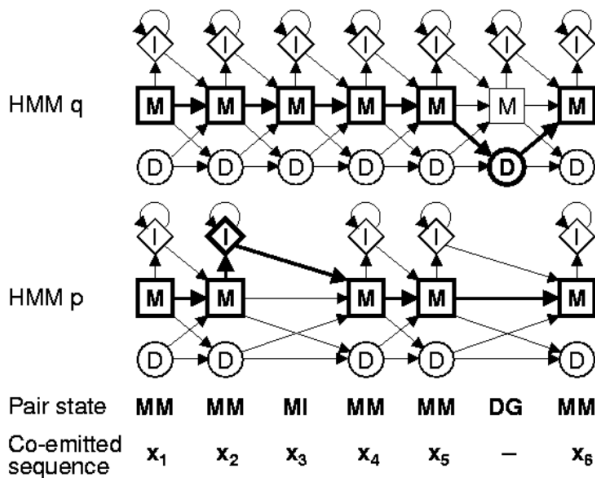
Profile-HMM



Profile-Hidden Markov Models(HMMs)

Pairwise Profile-HMM Alignment

- ▶ Example of a pairwise profile-HMM alignment:



Profile-Hidden Markov Models(HMMs)

Profile-HMMs Scoring Function

- ▶ Five possible pair states can co-emit amino acids or gaps: MM, MI, IM, DG and GD
- ▶ Log-sum-of-odds Score:

$$S_{LSO} = \log \sum_{x_1, \dots, x_L} \frac{P(x_1, \dots, x_L | \text{co-emission on path})}{P(x_1, \dots, x_L | \text{Null})}$$

Profile-Hidden Markov Models(HMMs)

Profile-HMMs Scoring Function

- ▶ Log-sum-of-odds Score:

$$S_{LSO} = \sum_{k: X_k Y_k = MM} S_{aa}(q_{i(k)}, p_{j(k)}) + \log \mathcal{P}_{tr}$$

- ▶ Column Score:

$$S_{aa}(q_i, p_j) = \log \sum_{a=1}^{20} \frac{q_i(a)p_j(a)}{f(a)}$$

- ▶ Also use dynamic programming, with a dynamic matrix for each co-emit state pair, to determine the best alignment

Profile-Hidden Markov Models(HMMs)

New Improvement

- ▶ Reaserchers Xin Deng and Jianlin Cheng from University of Missouri-Columbiacan
- ▶ Additional structural information
 - ▶ protein solvent accessibility
 - ▶ torsion angles
- ▶ Improved alignment accuracy





Summary

- ▶ *Critical Assessment of protein Structure Prediction (CASP)*
- ▶ The I-TASSER server (zhang-server) — top 3 places
- ▶ The HHpred server — top 10 places





Summary

- ▶ We looked at two different and popular approaches used in protein threading process
- ▶ Many different improvements have been proposed for both methods
- ▶ However, there is no single method outperforms all others on every target yet, which leaves room for improvement

References I

-  S. F. Altschul, J. C. Wootton, E. Zaslavsky, and Y.-K. Yu.
The construction and use of log-odds substitution scores for multiple sequence alignment.
Computational Biology, 2010.
-  X. Deng and J. Cheng.
Enhancing hmm-based protein profile-profile alignment with structural features and evolutionary coupling information.
BMC Bioinformatics, 2014.
-  A. Roy, A. Kucukural, and Y. Zhang.
I-tasser: a unified platform for automated protein structure and function prediction.
Nature protocols, 5(4):725738, 2010.
-  J. Söding.
Protein homology detection by HMMHMM comparison.
Bioinformatics, 21(7):951960, 2005.

References II

-  J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang.
The I-TASSER suite: protein structure and function prediction.
Nature methods,12(1):78, 2015.
-  J. Yang and Y. Zhang.
Protein structure and function prediction using I-TASSER.
Current protocols in bioinformatics,pages 58, 2015.
-  B.-J. Yoon.
Hidden markov models and their applications in biological sequence analysis.
Current genomics,10(6):402415, 2009.
-  Y. Zhang.
Progress and challenges in protein structure prediction.
Current opinion in structural biology,18(3):342348, 2008.

Questions?

Thank You!