

Biological Sequence Analysis (Ch 1. Introduction)

Biointelligence Laboratory
School of Computer Sci. & Eng.
Seoul National University
Seoul 151-742, Korea

This slide file is available online at

<http://bi.snu.ac.kr/>

Motivation of computational biology

- ▶ Organize, classify and parse the sequence data
- ▶ Behind the string of bases or amino acids is the whole complexity of molecular biology
- ▶ Capturing some of this complexity
- ▶ Direct experimentation is the most reliable way to determine a biological molecule's structure or function
- ▶ Obtaining DNA sequence of the gene for an RNA or protein is far easier than direct experimentation
- ▶ Biological information can be inferred from sequence alone.
- ▶ HMM is used for statistical analysis of sequence

Sequence similarity, homology, and alignment

- Two sequences are homologous
 - ▶ New sequences are adapted from pre-existing sequences rather than invented de novo.
 - ▶ Nature is a tinkerer and not an inventor [Jacob 1977]
 - ▶ We are transforming information by homology.
- Alignment between two strings
 - ▶ Evolving sequences accumulate insertions, deletions, substitutions
 - ▶ Scoring of alignment can be very complex.
 - ▶ Probabilistic modeling methods

Overview of the book

- Pairwise alignment
 - ▶ Pairwise alignment (2),
 - ▶ Markov chains and hidden Markov models (3),
 - ▶ Pairwise alignment using HMMs (4)
- Multiple alignment
 - ▶ Profile HMMs for sequence families (5)
 - ▶ Multiple sequence alignment methods (6)
- Phylogenetic trees
 - ▶ Building phylogenetic trees (7)
 - ▶ A probabilistic approach to phylogeny (8)
- RNA structure
 - ▶ Transformational grammars (9)
 - ▶ RNS structure analysis (10)

Probabilities and probabilistic models (1)

- Model: a system that simulates the objects (or sequences) under consideration
- Probabilistic model: one that produces different outcomes with different probabilities
 - ▶ (Example 1) A model of a roll of a dice have six parameters
 - ▶ (Example 2) A model of a sequence of three consecutive rolls [1, 6, 3]
- Biological sequences: a strings from a finite alphabet (4 nucleotides or 20 amino acids)

Probabilities and probabilistic models (2)

- Maximum likelihood estimation : maximize $P(D | \theta)$
- Overfitting: from a limited amount of data the model is well adapted to the training data but not generalize well to a new data (Ex) $[TTT] : p(H)=0, p(T)=1$

- Conditional, joint and marginal probabilities

- ▶ Probability of rolling i given dice 1 : $p(i | D_1)$
- ▶ Probability of picking dice j and rolling an i :

$$p(i, D_j) = p(D_j)p(i | D_j)$$

- ▶ Probability of rolling i : $p(i) = \sum_j p(D_j)p(i | D_j)$

Probabilities and probabilistic models (3)

- Bayes theorem and model comparison

- ▶ Testing of fair dice : Pick a dice at random and roll it for 3 times,

$$p(\text{unfair dice} | (5,5,5)) = \frac{p((5,5,5) | \text{unfair dice}) p(\text{unfair dice})}{p((5,5,5))}$$

- ▶ Protein sequence: extracellular proteins have a slightly different amino acid composition than intracellular proteins.

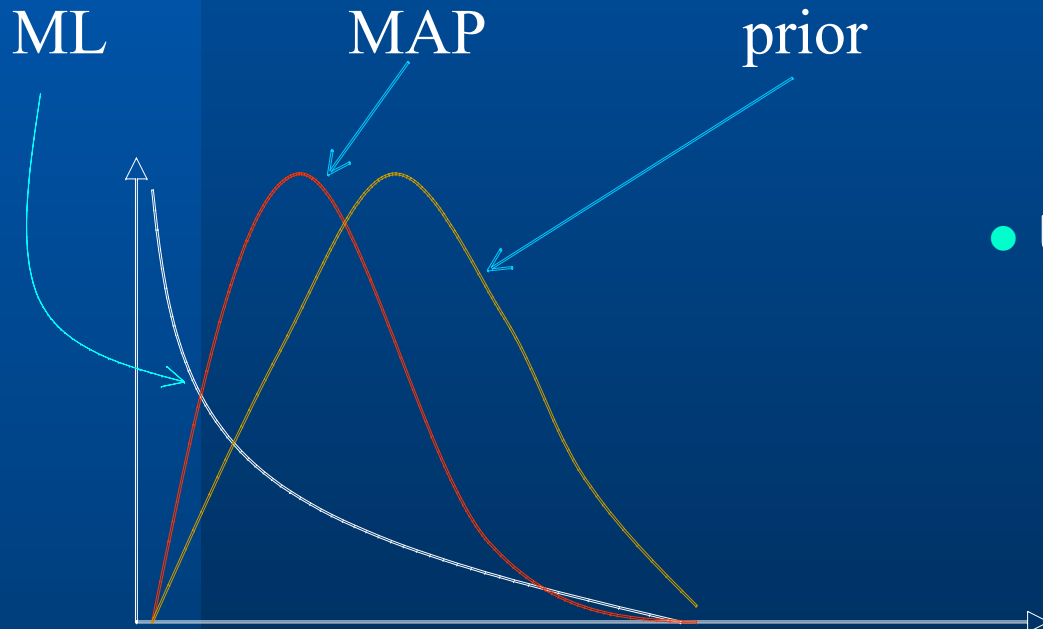
Test whether a new sequence is extracellular or not:

$$p(\text{ext} | x) = \frac{p(x | \text{ext}) p(\text{ext})}{p(x)} = \frac{p^{\text{ext}} \prod_i q_{x_i}^{\text{ext}}}{p(x)}$$

Probabilities and probabilistic models (4)

- Bayes theorem and parameter estimation (MAP)

$$p(\theta | D) = \frac{p(\theta)p(D | \theta)}{\int_{\theta'} p(\theta')p(D | \theta')}$$



(E.g.) Roll a dice to have
(1,3,4,2,4,6,2,1,2,2):

MLE (P5)=0

- Use of pseudo-count for the insufficient data