11. Multiple Sequence Alignment

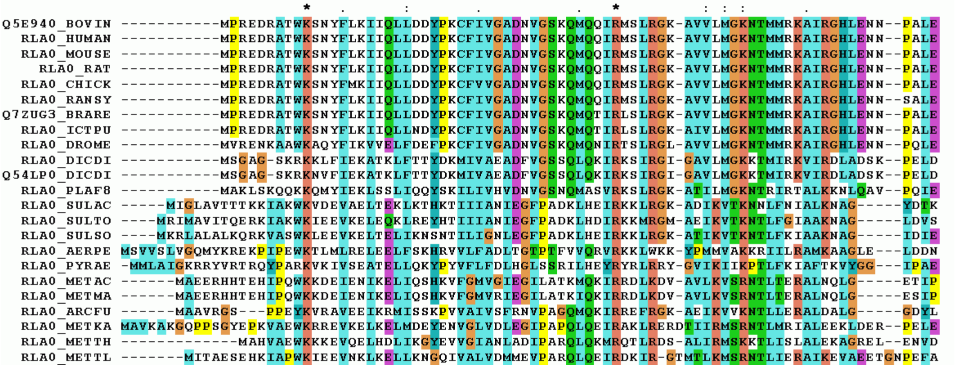
Outline

Introduction

Probabilistic models of MSA

Multiple sequence alignment

- multiple sequence alignment (MSA) is sequence alignment of three or more biological sequences such as DNA, RNA, or protein
- an example protein MSA



```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
```

Multiple sequence alignment

- each row of the MSA corresponds to the sequence of a specific protein
- each column of the MSA corresponds to a position in the sequence
- dash symbol means the sequence does not have an amino acid aligned at that position
- protein sequences are in the same MSA are evolutionarily related: they are homologous
- homologous sequences are derived from a common ancestor, so they are similar in sequence, structure, and function

Multiple sequence alignment

- MSA of a protein contains more information than a single sequence
- can be used to identify conserved regions in the protein
- conserved regions are often important for the protein's function
- used to infer the evolutionary relationships between the sequences
- used to search for homologous sequences in a database
- used to predict the structure and function of a protein

Multiple sequence alignment

- multiple algorithms exist for constructing MSAs
- most algorithms require a query sequence and a database of sequences
- they iteratively search for homologous sequences in the database and align them
- example algorithms: Clustal Omega, MUSCLE

Protein family

- a protein family is a group of proteins that share a common evolutionary origin
- members of a protein family are homologous and have similar sequences, structures, and functions
- sequences of a protein family are aligned to create a multiple sequence alignment
- the Pfam database is a collection of protein families

Outline

Introduction

Probabilistic models of MSA

Probabilistic models in general

- data: $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, where $x^{(i)}$ is a data sample and could be a scale or a vector
- a probabilistic model of the data defines a probability distribution $P(x; \theta)$
- θ is a set of parameters that define the model
- assumes that the observed data are generated by the model, i.e., the data are samples from the distribution $P(x; \theta^*)$
- θ^* is the true parameter value of the model
- learning the model means estimating the parameters θ from the data

A simple example of probabilistic model

- observed data: $0.43, 2.49, -1.91, 0.29, -2.1, 0.44$
- model:

$$p(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $\theta = (\mu, \sigma^2)$ is the set of parameters
- how to estimate θ from the data?

Maximum likelihood estimation

- a general approach to estimate the parameters of a probabilistic model based on the observed data

- estimates the parameters θ by maximizing the likelihood function

$$L(\theta) = P(x^{(1)}, x^{(2)}, \ldots, x^{(N)}; \theta) = \prod_{i=1}^{N} P(x^{(i)}; \theta)$$

- the estimate $\hat{\theta} = \arg\max_{\theta} L(\theta)$

- it is often easier to maximize the log-likelihood function

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \theta)$$

The probability distribution $P(x;\theta)$

- an assumption about the data and an approximation of the true distribution
- several factors influence the choice of the distribution
  - the nature of the data
  - the complexity of the model
  - the computational cost of estimating the parameters
  - the interpretability of the model
  - the need of sampling from the distribution or computing the likelihood
- by choosing a distribution with inherent structures, we could infer the structures from data

Example $P(x, \theta)$ with varying complexity and structrues

- a Gaussian distribution
- a mixture of Gaussians
- a Gaussian process
- a hidden Markov model
- the Boltzmann machine
- large language models
- autoregressive probabilistic models
- variational autoencoders
- restricted Boltzmann machines
- the Ising model
- the Potts model
- large language models of proteins

Probabilistic models of MSA

- a MSA is a collection of sequences: $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, where $x^{(i)}$ is a sequence of amino acids
- a probabilistic model of MSA defines a probability distribution $P(x; \theta)$
- $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ are assumed to be samples from the distribution $P(x; \theta^*)$
- two examples of probabilistic models of MSA
  - MSA profile (position independent model)
  - Potts model (directed coupling analysis)

MSA profile

- assumes that amino acids at each position are independent
- the probability of a sequence is the product of the probabilities of each amino acid at each position

$$P(x; \theta) = \prod_{k=1}^{L} P(x_k; \theta_k)$$

- L is the length of the sequence and θ_k is the set of parameters for the k-th position
- $P(x_k; \theta_k)$ is the probability distribution of amino acid types at the k-th position

MSA profile

- assume there are no gaps in the MSA, then x_k has 20 possible values (20 amino acids)
- the probability distribution $P(x_k; \theta_k)$ is a multinomial distribution

$$P(x_k = i; \theta_k) = \theta_{i,k}$$

- $\theta_{i,k}$ is the probability of the k-th position being the i-th amino acid and $\sum_{i=1}^{20} \theta_{i,k} = 1$
- estimate $\theta_{i,k}$ with MLE and is equal to the frequency of each amino acid at each position

$$\hat{\theta}_{i,k} = \frac{N_{i,k}}{N}$$

- $N_{i,k}$ is the number of times the i-th amino acid appears at the k-th position in the MSA

MSA profile

- is a matrix of size $20 \times L$

$$\begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \dots & \theta_{1,L} \\ \theta_{2,1} & \theta_{2,2} & \dots & \theta_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{20,1} & \theta_{20,2} & \dots & \theta_{20,L} \end{pmatrix}$$

- used by many ML methods as input features
- captures more information about a protein family than a single sequence
- easy to sample sequences from the distribution and compute the likelihood of a sequence
- ignores dependency between positions

Potts model

- a more complex model that captures the dependency between positions
- assumes that the probability of a sequence is given by a Boltzmann distribution

$$P(x;\theta) = \frac{1}{Z(\theta)} e^{-E(x;\theta)}$$

- $E(x;\theta)$ is the “energy” of the sequence and $Z(\theta)$ is the partition function

$$Z(\theta) = \sum_{x} e^{-E(x;\theta)}$$

- the sum in the partition function is over all possible sequences
- how many possible sequences are there for a given length L?

Potts model

- the energy function is given by

$$E(x;\theta) = \sum_{k=1}^{L} h_k(x_k) + \frac{1}{2}\sum_{k=1}^{L}\sum_{l=1}^{L} J_{kl}(x_k, x_l)$$

- $h_k(x_k)$ is the “field” at position k
- $h_k(x_k)$ captures preferences of the amino acid types at position k
- $J_{kl}(x_k, x_l)$ is the “coupling” between positions k and l
- $J_{kl}(x_k, x_l)$ captures the dependency between the amino acid types at positions k and l

The field term

- assume there are no gaps in the MSA, then x_k has 20 possible values (20 amino acids)
- to specify the field term, we need to define $h_k(x_k)$ for each amino acid type
- let $h_k(x_k = i) = h_{i,k}$, the field term at the k-th position is given by

$$h_k(x_k) = \sum_{i=1}^{20} h_{i,k} \cdot \mathbb{1}\{x_k = i\}$$

- the total field term is given by

$$E_f(x;\theta) = \sum_{k=1}^{L} h_k(x_k) = \sum_{k=1}^{L} \sum_{i=1}^{20} h_{i,k} \cdot \mathbb{1}\{x_k = i\}$$

The coupling term

- to specify it, we need to define $J_{kl}(x_k, x_l)$ for each pair of amino acid types
- let $J_{kl}(x_k = i, x_l = j) = J_{i,j}^{k,l}$, the coupling term at positions k and l is given by

$$J_{kl}(x_k, x_l) = \sum_{i=1}^{20} \sum_{j=1}^{20} J_{i,j}^{k,l} \cdot \mathbb{1}\{x_k = i\} \cdot \mathbb{1}\{x_l = j\}$$

- the total coupling term is given by

$$\frac{1}{2} \sum_{k=1}^{L} \sum_{l=1}^{L} J_{kl}(x_k, x_l) = \frac{1}{2} \sum_{k=1}^{L} \sum_{l=1}^{L} \sum_{i=1}^{20} \sum_{j=1}^{20} J_{i,j}^{k,l} \cdot \mathbb{1}\{x_k = i\} \cdot \mathbb{1}\{x_l = j\}$$

- the factor of $\frac{1}{2}$ is to avoid double counting