

## 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  -----M*PREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0_HUMAN   -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0_MOUSE   -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0_RAT      -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0_CHICK    -----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0_RANSY    -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE  -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0 ICTPU    -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
RLA0_DROME    -----MVRENKAAWKAQYFIKVVLFDEFPPKCFIVGADNVGSKMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PQLE
RLA0_DICDI    -----MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSQLOKIRKSIRGI-GAVLMGKNTMIRKVVIRDLADSK--PELD
Q54LP0_DICDI  -----MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSQLOKIRKSIRGI-GAVLMGKNTMIRKVVIRDLADSK--PELD
RLA0_PLAF8    -----MAKLSKQKKQMYIEKLSLILQQYSKILIVHDNVGSGNOMASVRKSLRGK-ATILMGKNTIRRTALKKNLQAV--PQIE
RLA0_SULAC    -----MIGLAVTTT*KKIAKKVDEVAELTEKLKTHKTIIANIEGFPADKLHEIRKKLRGK-ADIKVTNNLNFNIALKNAG----YDTK
RLA0_SULTO    -----MRIMAVITQERKIAKKVIEEVKELEKLEHYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTNTLFGIAAKNAG----LDVS
RLA0_SULSO    -----MKRLALALKQKRVASWKKLEEVKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTNTLFLKIAAKNAG----IDIE
RLA0_AERPE    MSVYVSLVQMYKREK*IP*EWKTLMLRELELFSKHRVVLADLTGTP*FVVRVRKKLWKK-YPMVAKKRILLAMKAAGLE--LDDN
RLA0_PYRAE    MMLAIGKRRYVTRQY*PARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRY-GVIKIIP*TLFKIAFTKVYGG--IPAE
RLA0_METAC    -----MAERHHTTEHIPQWKKDEIENIKELIQSHKVF*GMVRIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG----ETIP
RLA0_METMA    -----MAERHHTTEHIPQWKKDEIENIKELIQSHKVF*GMVRIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG----ESIP
RLA0_ARCFU    -----MAAVRGS-----PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMKIRREFRGK-AEIKVVKNTLLERALDNLG--GDYL
RLA0_METKA    MAVKAKGQPPSGYE*PKVAEWKRRREVKKLELMDEYENVGLVDLEGIPAPOLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
RLA0_METTH    -----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPAROLQKMRQTLRDS-ALIRMSKKTLLISLAEKAGREL--ENVD
RLA0_METTL    -----MITAESEHKIAPWKIEEVNKKLELLKNGQIVALVDMMEV*PAROLQEIRDKIR-GTMTLKMSRNTLLERAIKEVAEETGNPEFA
  
```

## 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 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