

Biological sequence analysis in drug discovery

Jitao David Zhang

Jul 31, 2019

The Central Dogma of Molecular Biology

DNA/RNA sequence analysis

How does BLAST work

Probabilistic modelling of biological sequences

Conclusions

The Central Dogma of Molecular Biology

The central dogma

See the notes

Examples of drugs that target DNA or RNA

Examples of drugs that target DNA or RNA

More and more drugs target DNA/RNA of the CD graph

Sequence analysis allows designing potent and specific ASOs

Sequence analysis is critical for phylogenetic analysis

Sequence analysis is critical for phylogenetic analysis

Sequence analysis helps us understanding evolution of sequences

DNA/RNA sequence analysis

DNA is a string of characters and often form double helixes

RNA is a string of characters and is often single-stranded

Levenshtein distance

- The minimum number of operations required to transform string a to string b with following operations:

Levenshtein distance

- The minimum number of operations required to transform string a to string b with following operations:
 - Insertion, for instance sit **bat** \rightarrow **bait**

Levenshtein distance

- The minimum number of operations required to transform string a to string b with following operations:
 - Insertion, for instance sit **bat** \rightarrow **bait**
 - Deletion, e.g. **boat** \rightarrow **bot**

Levenshtein distance

- The minimum number of operations required to transform string a to string b with following operations:
 - Insertion, for instance sit bat \rightarrow bait
 - Deletion, e.g. boat \rightarrow bot
 - Substitution, e.g. pig \rightarrow big

Levenshtein distance

- The minimum number of operations required to transform string a to string b with following operations:
 - Insertion, for instance `sit bat` \rightarrow `bait`
 - Deletion, e.g. `boat` \rightarrow `bot`
 - Substitution, e.g. `pig` \rightarrow `big`
- The Levenshtein distance between two strings a, b of length $|a|$ and $|b|$ respectively is given by $\text{lev}_{a,b}(|a|, |b|)$ where

$$\text{lev}_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $\text{lev}_{a,b}(i, j)$ is the distance between the first i characters of a and the first j characters of b .

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Question

What is the Levenshtein distance between **AATGCTGCTT** and **AATGCATT**?

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Question

What is the Levenshtein distance between **AATGCTGCTT** and **AATGCATT**?

Answer

3

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Question

What is the Levenshtein distance between **AATGCTGCTT** and **AATGCATT**?

Answer

3

- Beyond bioinformatics, the Levenshtein distance is often used in computational linguistics and natural language processing. For instance, check out *How to Write a Spelling Corrector* by Peter Norvig.

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Question

What is the Levenshtein distance between **AATGCTGCTT** and **AATGCATT**?

Answer

3

- Beyond bioinformatics, the Levenshtein distance is often used in computational linguistics and natural language processing. For instance, check out *How to Write a Spelling Corrector* by Peter Norvig.

Levenshtein distance (continued)

- The Levenshtein distance is often referred to as the *edit distance*.

Question

What is the Levenshtein distance between **AATGCTGCTT** and **AATGCATT**?

Answer

3

- Beyond bioinformatics, the Levenshtein distance is often used in computational linguistics and natural language processing. For instance, check out *How to Write a Spelling Corrector* by Peter Norvig.

Levenshtein distance applies to biological sequences

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):
 - ...**DTGAQTT**... (yeast)

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):
 - ...DTGAQTT... (yeast)
 - ...D**S**GAQTT... (fruit fly)

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):
 - ...DTGAQTT... (yeast)
 - ...DSGAQTT... (fruit fly)
 - ...DSGAQMT... (human)

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):
 - ...DTGAQTT... (yeast)
 - ...D**S**GAQTT... (fruit fly)
 - ...D**S**GAQ**M**T... (human)

- The simplest substitution matrix: the Identity matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):
 - ...DTGAQTT... (yeast)
 - ...D**S**GAQTT... (fruit fly)
 - ...D**S**GAQ**M**T... (human)

- The simplest substitution matrix: the Identity matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

Substitution matrix

- Substitution matrix describes the rate at which one character changes to other character states over time.
- Example of DNA-damage-inducible 1/2 protein (Silva *et al.*, Nature Reports, 2016):
 - ...DTGAQTT... (yeast)
 - ...DSGAQTT... (fruit fly)
 - ...DSGAQMT... (human)

- The simplest substitution matrix: the Identity matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

Questions

(1) What are the advantage and disadvantage of using the identity matrix? (2) What other alternatives can you imagine?

- Log-odds matrices: we can express the probabilities of substitution with log-odds scores. The scores matrix S is defined as

$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}}$$

where $M_{i,j}$ is the probability that character i transforms into character j , and p_i , p_j are the frequencies of character i and j .

- Log-odds matrices: we can express the probabilities of substitution with log-odds scores. The scores matrix S is defined as

$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}}$$

where $M_{i,j}$ is the probability that character i transforms into character j , and p_i , p_j are the frequencies of character i and j .

- Commonly used log-odds substitution matrices

- Log-odds matrices: we can express the probabilities of substitution with log-odds scores. The scores matrix S is defined as

$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}}$$

where $M_{i,j}$ is the probability that character i transforms into character j , and p_i , p_j are the frequencies of character i and j .

- Commonly used log-odds substitution matrices
 - PAM (Point Accepted Mutation) matrix, developed by Margaret Dayhoff in the 1970s, works well to compare closely related species, e.g. rat and mouse.

Substitution matrix continued

- Log-odds matrices: we can express the probabilities of substitution with log-odds scores. The scores matrix S is defined as

$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}}$$

where $M_{i,j}$ is the probability that character i transforms into character j , and p_i , p_j are the frequencies of character i and j .

- Commonly used log-odds substitution matrices
 - PAM (Point Accepted Mutation) matrix, developed by Margaret Dayhoff in the 1970s, works well to compare closely related species, e.g. rat and mouse.
 - BLOSUM (BLOCK SUBstitution Matrix), developed by Steven and Jorja G. Henikoff in early 1990s, works well for evolutionarily divergent sequences, say zebrafish and human.

The Needleman-Wunsch algorithm

The Needleman-Wunsch algorithm

Dynamic programming underlies the Needleman-Wunsch algorithm

The Smith-Waterman algorithm

The Smith-Waterman algorithm

SM reaches local alignment whereas NW reaches global alignment

How does BLAST work

Sequence query is frequently used in drug discovery

Anyone can query similar sequences using the BLAST tool

How BLAST in principle works

How BLAST in principle works

BLAST is a heuristic method

Statistical models are important components of bioinformatics

Probabilistic modelling of biological sequences

Model DNA/RNA sequences using Markov chains

Model DNA/RNA sequences using Markov chains

Markov chain is a probabilistic model of biological sequences

HMMs allow sequence discrimination

Conclusions

- Phylogenetic analysis of drug targets

- Phylogenetic analysis of drug targets
 - Is my target present in mouse/rat/rabbit...?

Other sequence-analysis tasks in drug discovery

- Phylogenetic analysis of drug targets
 - Is my target present in mouse/rat/rabbit...?
 - Is the function of my target conserved in animals?

Other sequence-analysis tasks in drug discovery

- Phylogenetic analysis of drug targets
 - Is my target present in mouse/rat/rabbit...?
 - Is the function of my target conserved in animals?
- Prediction of RNA secondary structure

Other sequence-analysis tasks in drug discovery

- Phylogenetic analysis of drug targets
 - Is my target present in mouse/rat/rabbit...?
 - Is the function of my target conserved in animals?
- Prediction of RNA secondary structure
- Protein sequence and structure analysis

Other sequence-analysis tasks in drug discovery

- Phylogenetic analysis of drug targets
 - Is my target present in mouse/rat/rabbit...?
 - Is the function of my target conserved in animals?
- Prediction of RNA secondary structure
- Protein sequence and structure analysis
 - Discussed in the follow-up sessions

- Rosalind

- Rosalind
- Teaching RNA algorithms by Backofen Lab at U Freiburg, with source code available on Github.

- Rosalind
- Teaching RNA algorithms by Backofen Lab at U Freiburg, with source code available on Github.
- An Introduction to Applied Bioinformatics