

14 Sequence patterns

14.1 Sequence patterns

A sequence pattern can be used to find protein family members. The concept of finding patterns is similar to creating regular expressions.

The PROSITE language

- x: An arbitrary amino acid
- -: Separating elements
- []: A list of amino acids
- {}: A list of not accepted amino acids
- (): A range of an element

Example of PROSITE

Find all matched sequences for the patterns. Assume the alphabet $M = \{A, G, C, T\}$.

Pattern 1: A - [GC] - {AGC}

AGT
ACT

Pattern 2: A - x(1,3) - G

AxG
AxxG
AxxxG

Exercise 14.1

Find all matched sequences for the pattern. Assume the alphabet $M = \{A, G, C, T\}$.

Pattern: [AC] - {GCT} - x(1, 2) - T

14.2 Pattern comparison

Information theory is used to score a pattern. Calculated scores indicate the specificity of patterns.

Information theory for amino acid distribution

$I(a)$ represents the information content of amino acid a when the probability of a is p_a .

$$I(a) = -\log p_a$$

$H(M)$ represents the entropy of all amino acids.

$$H(M) = -\sum_{a \in M} p_a \log p_a$$

Example of entropy calculation

Calculate the information content of an amino acid A when $p(A) = 0.25$. Use base 2.

$$I(a) = -\log p_a = -1 \times -2 = 2$$

Calculate the entropy of M that contains three pseudo amino acids A, B, C with the probabilities $p(A) = 1/2$, $p(B) = 1/4$, and $p(C) = 1/4$.

$$\begin{aligned} H(M) &= -p(A) \cdot \log p(A) - p(B) \cdot \log p(B) - p(C) \cdot \log p(C) \\ &= (-0.5) \times (-1) + (-0.25) \times (-2) + (-0.25) \times (-2) = 1.5 \end{aligned}$$

Scores of patterns

Patterns can be scored for their specificity.

p'_{ai} represents an adjusted probability of a at position i .

$$p'_{ai} = \frac{p_a}{\sum_{b \in K_i} p_b}$$

I'_{K_i} represents the information content of a set of amino acids K_i at position i .

$$I'(K_i) = H(M) - \left(-\sum_{a_i \in K_i} p'_{a_i} \log p'_{a_i} \right)$$

The information content of a pattern P can be the sum of I'_{K_i} for all i .

$$I(P) = \sum_i I'(K_i) - c \sum_k (j_k - i_k)$$

where c is a constant value for a wildcard region $x(j_k - i_k)$.

Example of pattern specificity scores

Calculate the information content of the following patterns when the probabilities of the pseudo amino acids are $p(A) = 1/2$, $p(B) = 1/4$, and $p(C) = 1/4$. Use $c = 0.1$.

$\mathbf{P}_1 : A - B$

$$I'(K_1) = 1.5 - (-1 \log 1) = 1.5$$

$$I'(K_2) = 1.5 - (-1 \log 1) = 1.5$$

$$I(P_1) = 1.5 + 1.5 = 3$$

$\mathbf{P}_2 : [AC] - B$

$$p'_{A1} = \frac{1/2}{(1/2 + 1/4)} = \frac{1/2}{3/4} = \frac{2}{3}$$

$$p'_{C1} = \frac{1/4}{(1/2 + 1/4)} = \frac{1/4}{3/4} = \frac{1}{3}$$

$$\begin{aligned} I'(K_1) &= 1.5 - \left(-\frac{2}{3} \log \frac{2}{3} \right) - \left(-\frac{1}{3} \log \frac{1}{3} \right) \\ &= 1.5 + \frac{2}{3}(1 - \log 3) - \frac{2}{3} \log 3 \\ &= \frac{13}{6} - \log 3 \end{aligned}$$

$$I'(K_2) = 1.5 - (-1 \log 1) = 1.5$$

$$I(P_2) = 1.5 + \frac{13}{6} - \log 3 = \frac{22}{6} - \log 3 = 2.817$$

$\mathbf{P}_3 : A - B - x(1, 2)$

$$I(P_3) = 3 - 0.1 \times (2 - 1) = 2.9$$

14.3 Pattern discovery

Sequence patterns can be created from MSAs.

Pattern discovery methods

- Comparison-based methods
- Pattern-driven methods

Pivot-based methods

The pivot-based method is one of the comparison-based methods. One sequence is used as pivot, and all the rest of the sequences are compared with this pivot to find similar segments.

Example of pivot-based methods

Find similar segments of size 3 with one edit distance without insertion/deletion for the following sequences. The first sequence should be used as pivot.

Seq1 GATC
Seq2 GGACCG
Seq3 GAG
Seq4 GGGT

Select segments for GAT.

Seq2 GAC
Seq3 GAG
Seq4 GGT

Construct a pattern for the segment of GAT.

G - [AG] - [CGT]

Select segments for ATC.

Seq2 ACC

Construct a pattern for the segment of ATC.

A - [TC] - C

Pratt

Pratt is one of the pattern-driven methods.

$$C_1 - x(i_1, j_1) - C_2 - x(i_2, j_2) - \dots - C_{p-1} - x(i_{p-1}, j_{p-1}) - C_p$$

Example of Pratt

Find a matched pattern for the following three sequences.

Seq1 ACTG
Seq2 ACGT
Seq3 ACT

$$A - x(0, 0) - C - x(1, 2)$$

Pratt procedure

Pratt uses the following three main steps to find the best matched pattern.

- Preprocessing
- Searching
- Specialization