cours 5:

recherche de motifs: Brute force

Identifier les motifs dans les régions régulatrices

- La découverte de motifs similaires dans les régions régulatrices (promoteur) de plusieurs gènes suggère une relation de régulation entre ces gènes.

- Problèmes
  - Nous ne connaissons pas la séquence de motifs à l'avance.
  - Nous ne savons pas où se situe le motif par rapport au début des gènes.
  - Un motif peut différer légèrement d'un gène à l'autre.

Identifier les motifs dans les régions régulatrices

- La découverte de motifs similaires dans les régions régulatrices de plusieurs gènes suggère une relation de régulation entre ces gènes.

- Question

- Comment identifier les vrais motif lié à de factor de transcription de motifs «aléatoires» qui ne représentent pas une corrélation réelle entre les gènes?

Trouver un motif: Générer de donnée artificiel

- Étant donné un échantillon aléatoire de séquences d'ADN

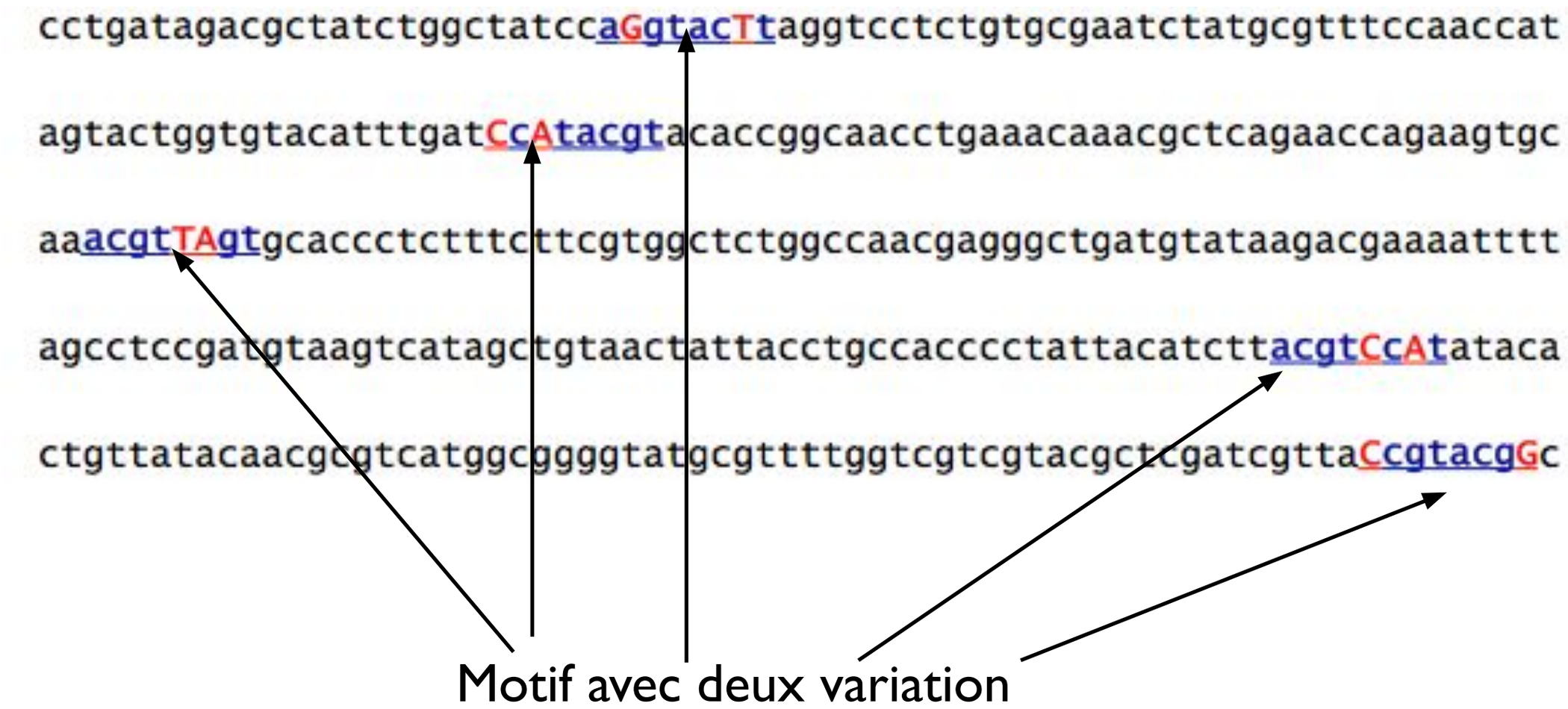
```
cctgatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcgtttccaaccat
agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc
aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt
agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatcttacgtacgtataca
ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgttaacgtacgtc
```

- Trouver le motif qui a été implanté dans chacune des séquences individuelles

- Par exemple: la séquence implanté est de longueur 8.

Trouver un motif: Générer de donnée artificiel

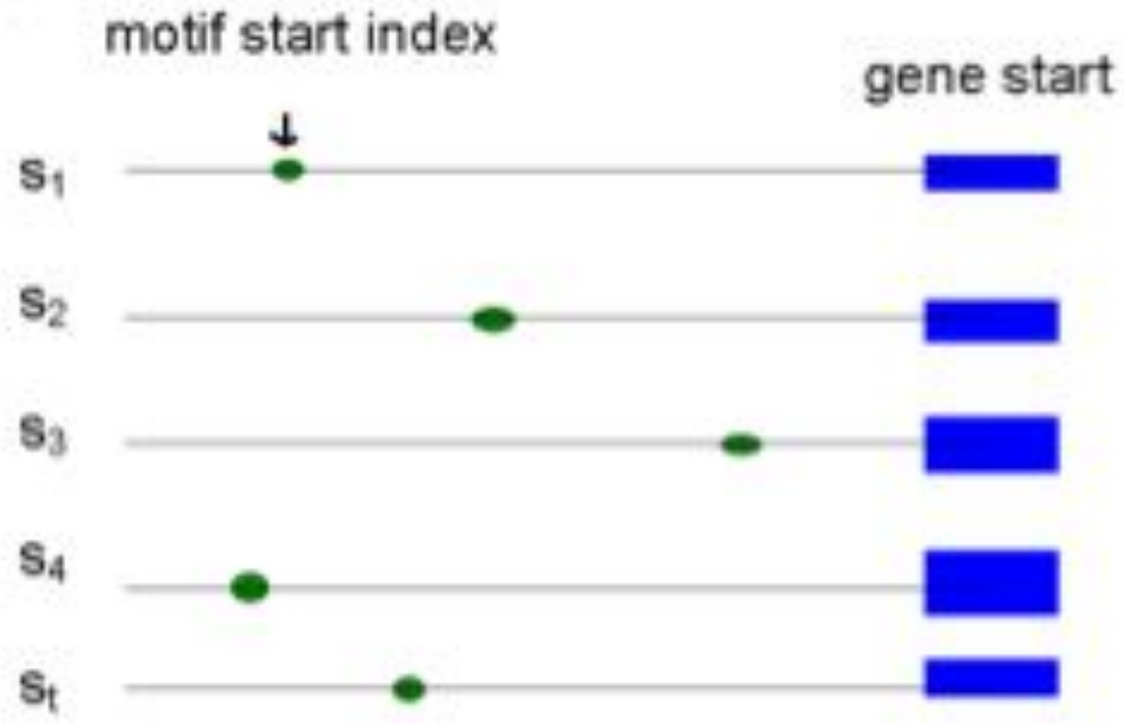
- Le motif n'est pas nécessairement le même dans les sequence, car des mutations aléatoires (substitutions) peuvent se produire dans les séquences



Trouver un motif: The Motif Finding Problem

- Comment trouver des motifs variables?
  - Disons que nous savons où le motif commence dans chaque séquence.

    - Les positions de début de motif dans leurs séquences peuvent être représentées comme $s = (s_1, s_2, \ldots, s_t)$



Trouver un motif: The Motif Finding Problem

- Comment trouver des motifs variables?
  - Extraire et aligner les modèles à partir de leurs index de départ $s = (s_1, s_2, \ldots, s_t)$

Alignment

```
a G g t a c T t
C c A t a c g t
a c g t T A g t
a c g t C c A t
C c g t a c g G
```

  - Construire une matrice de profil avec les fréquences de chaque nucléotide dans les colonnes

Profile

A	3	0	1	0	3	1	1	0
C	2	4	0	0	1	4	0	0
G	0	1	4	0	0	0	3	1
T	0	0	0	5	1	0	1	4

  - Le nucléotide consensus dans chaque position a le plus haut score dans la colonne

Consensus A C G T A C G T

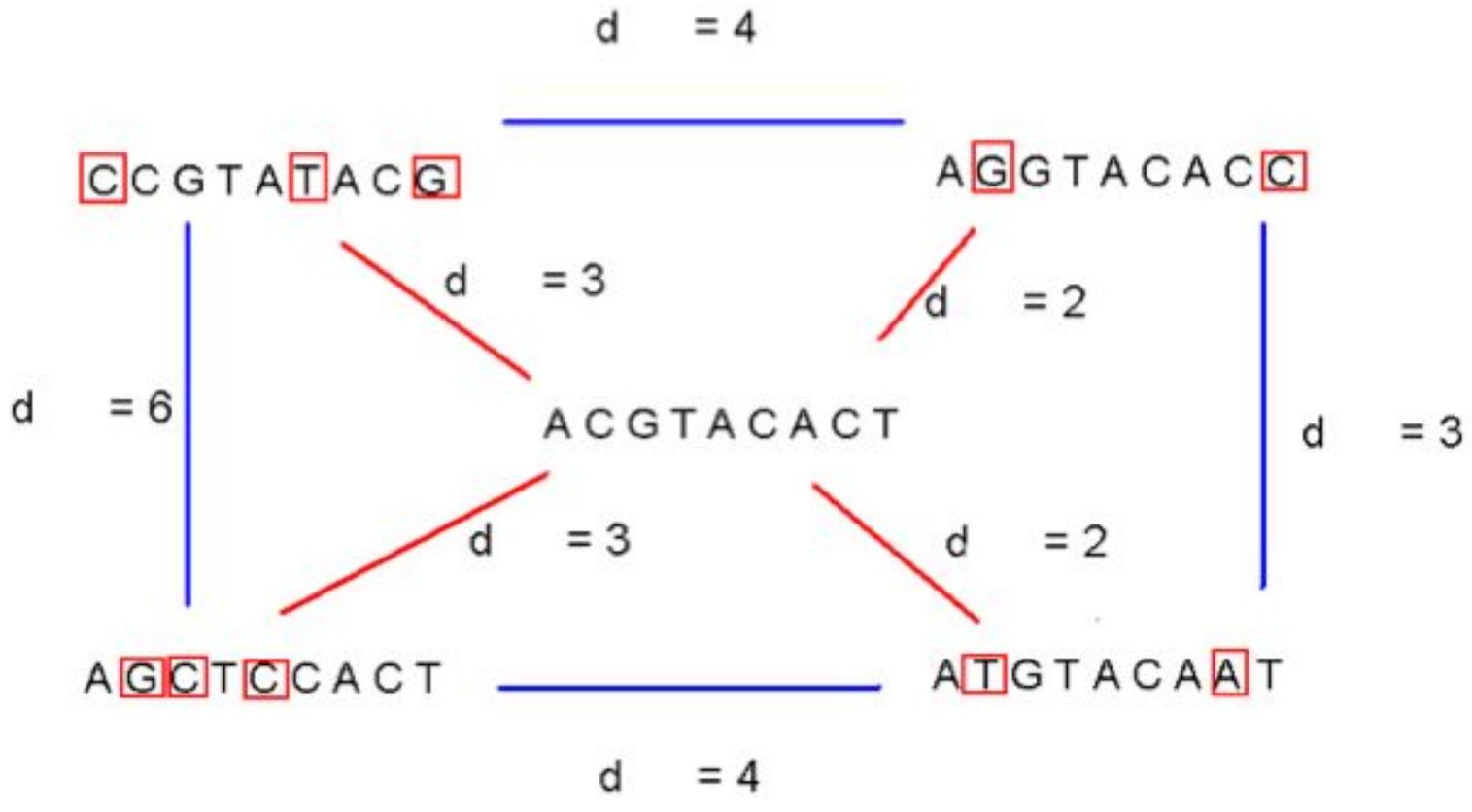
  - **L'idée principale est : nous voulons choisir les positions de départ que nous donner la meilleure séquence consensus**

Trouver un motif: The Motif Finding Problem

- Sequence consensus

  - Pensez à la séquence consensus comme un motif «ancêtre», à partir duquel des motifs mutés ont émergé

  - La distance entre un motif réel et la séquence consensus est généralement inférieure à la distance entre deux motifs réels



Trouver un motif: The Motif Finding Problem

- Evaluer les motifs

  - Nous avons une estimation de la séquence consensus, mais à quel point ce consensus est-il «bon»?

  - Nous pouvons définir une fonction d'évaluation pour comparer différentes séquences consensus.

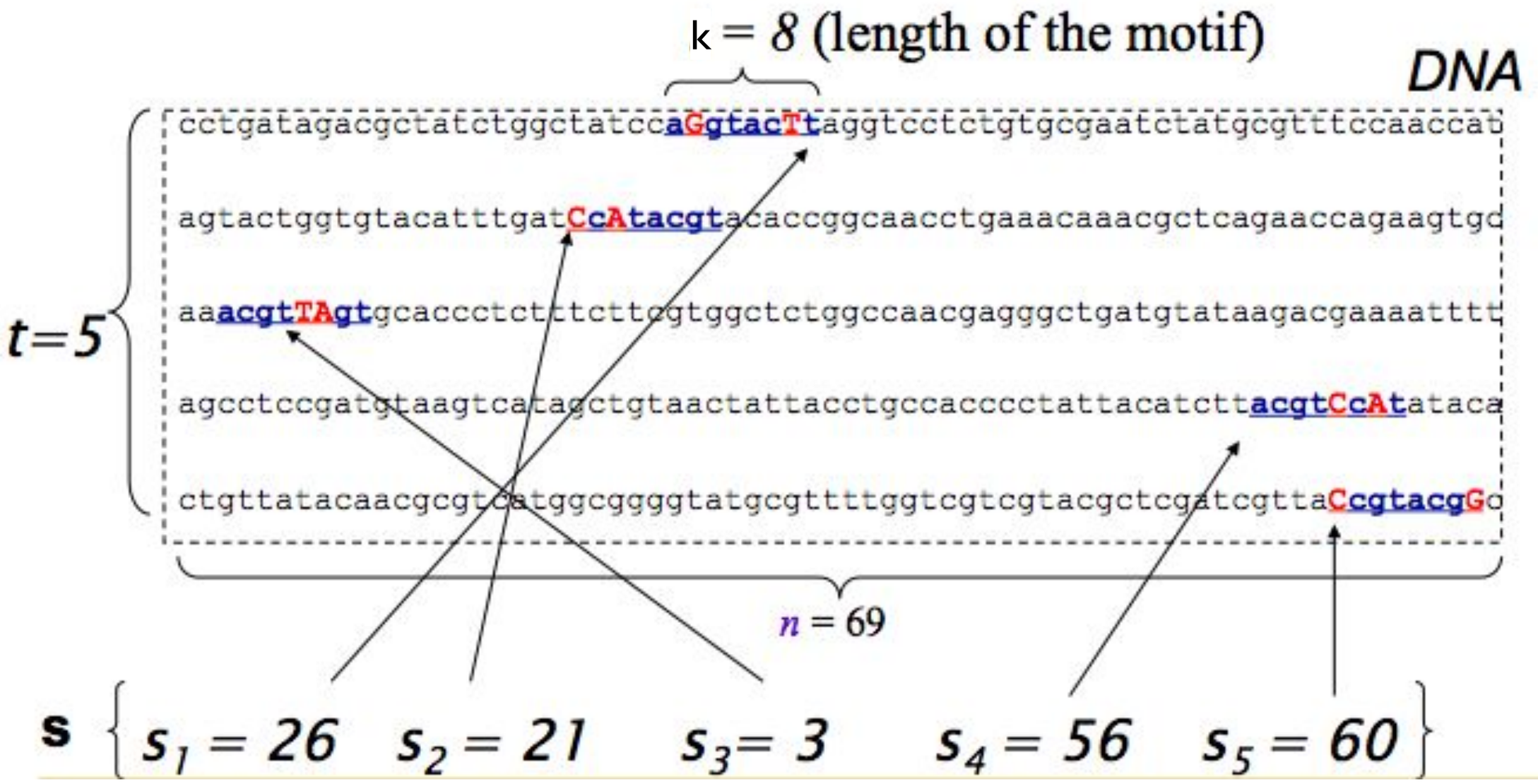
  - **L'idée principale est : nous voulons choisir les positions de départ que nous donner la meilleure séquence consensus**

Trouver un motif: The Motif Finding Problem

- <u>Paramètre du algorithme</u>
  - t - nombre de séquences d'ADN
  - n - longueur de chaque séquence d'ADN
  - ADN - échantillon de séquences d'ADN (stocké sous la forme d'un tableau t x n) •
- k - longueur du motif (k-mer)
- s_i - position de départ d'un k-mer dans la séquence i
- $s = (s_1, s_2, \ldots, s_t)$ positions de départ du motif

Trouver un motif: The Motif Finding Problem

- Paramètre du algorithme



Trouver un motif: The Motif Finding Problem

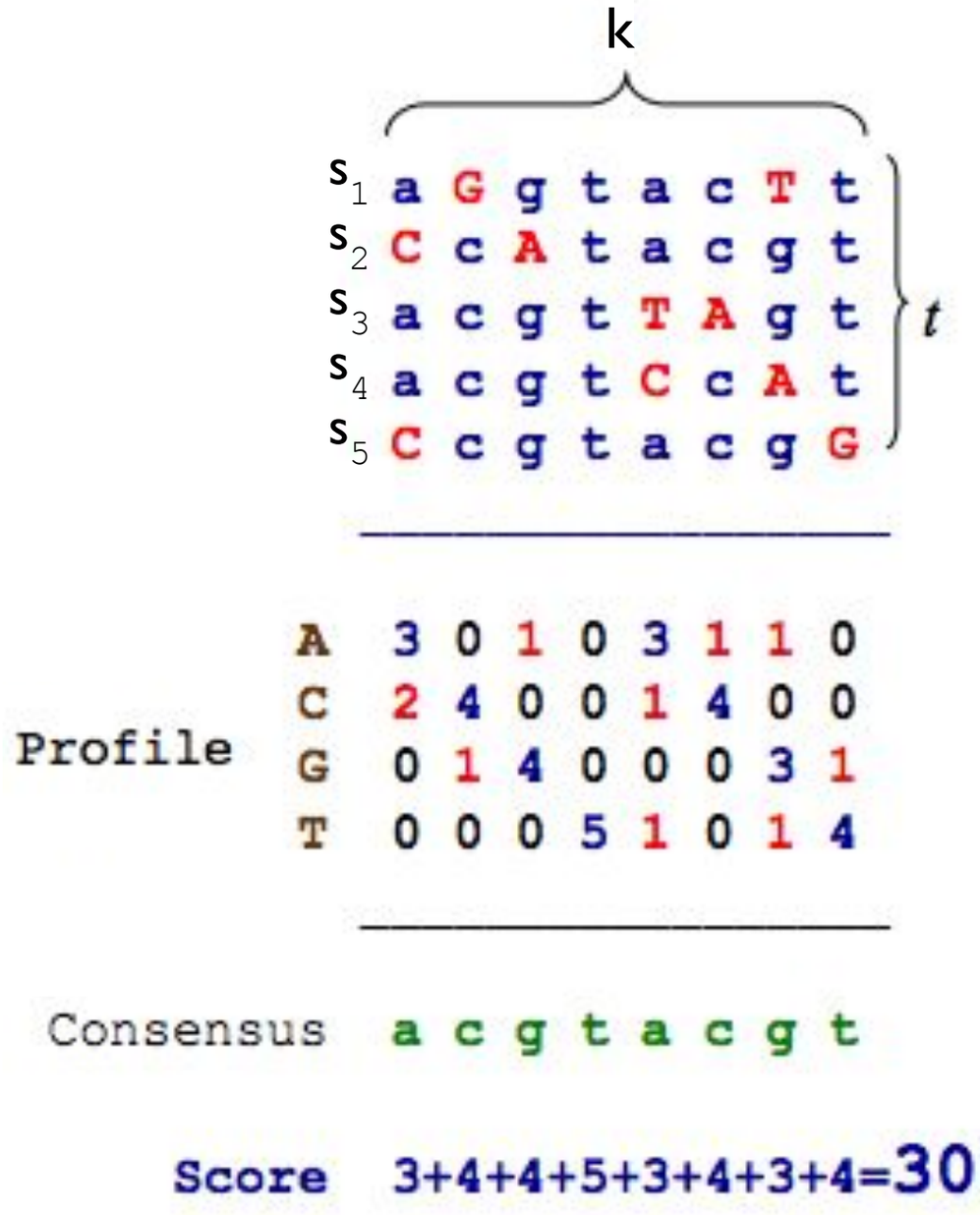
- <u>Evaluer les motifs</u>
  - Étant donné les positions de départ s = (s_1, ... s_t) et les sequences d'ADN:

$$\text{Score}(s, \text{DNA}) = \sum_{i=1}^{k} \max_{v \text{ in } \{A, C, G, T\}} (\text{count}(v, i))$$

  - count (v, i) représente le fréquence du nucléotide v dans le motif à partir de s_i

Trouver un motif: The Motif Finding Problem

- Evaluer les motifs : exemple



Trouver un motif: The Motif Finding Problem

- <u>Definition formelle</u>

  - But: À partir d'un ensemble de séquences d'ADN, trouver un ensemble de k-mers, un de chaque séquence, qui maximise le score consensuel

  - Entrée: une matrice t x n d'ADN, et k, la longueur du motif à trouver

  - Sortie: un vecteur de t positions de départ $s = (s_1, s_2, \ldots, s_t)$ maximisant le score (s, ADN)

Brute Force Method

- Calculer les scores pour chaque combinaison possible de positions de départ s.

- Le meilleur score déterminera le meilleur profil et la meilleure séquence consensus.

- Le but est de maximiser Score (s, ADN) en faisant varier les positions de départ s_i, où:

$$s_i = [1, ..., n-k+1]$$

$$i=[1,..., t]$$

Brute Force Method

```
1.  BruteForceMotifSearch(DNA, t, n, k)
2.  bestScore ← 0
3.  for each s=(s1,s2 , . . ., st) from (1,1 . . . 1)
                                    to (n - k +1, …, n - k +1)
4.      if (Score(s,DNA) > bestScore)
5.          bestScore ← score(s, DNA)
6.          bestMotif ← (s1,s2 , . . . , st)
7.  return bestMotif
```

Brute Force Method : complexité

- Variant (`n - k + 1`) positions dans chacune des t séquences, nous regardons (`n - k + 1`)^t ensembles s.

- Pour chaque s, la fonction d'évaluation fait k opérations, donc la complexité de l'algorithme est

- k (`n - k + 1`)^t = O (kn^t)

- Cela signifie que pour t = 8, n = 1000, k = 10, nous devons effectuer environ 10*1000^8 = 10^{25} calculs - l'algorithme prendra des milliards d'années pour se terminer sur une telle instance de problème