Cours en Analyse de données séquentielles

Chapitre 4: Fouille des motifs séquentiels pour les données biologiques

Dr D. AKROUR
2ème année Master
Systèmes d'Information, Optimisation et Décision

2024-2025

Plan

- Bioinformatiques
- Les séquences biologiques
- Évolution des séquences biologiques
  - Mutation
  - Séquences homologues
- Alignement des Séquences Biologiques
- Méthodes d’évaluation de l’alignement
- Algorithme d’alignement
  - L'algorithme Needleman-Wunsch
- Analyse des Séquences Biologiques
  - Chaîne de Markov
  - Modèle de Markov Cachées (HMM)

Bioinformatique

- **Discipline:** combinant **biologie**, **informatique** et **mathématiques**.
- **Objectif :** Analyser des séquences biologiques (ADN, ARN, protéines).
- **Défis:**
  - Séquences longues et complexes,
  - Contiennent des informations cachées d'une grande signification biologique.
  - Méthodes traditionnelles inefficaces pour extraire des motifs colossaux.

Bioinformatique

- **Utilité** de l'analyse séquentielle en bioinformatique:
  - comprendre les **processus biologiques** et l'**évolution des séquences**.
  - **traiter** efficacement ces **grandes quantités** de données et **découvrir** des **séquences homologues**, révélant des relations évolutives entre les espèces.
- **Algorithmes:**
  - la **programmation dynamique**
  - les **modèles de Markov cachés** (HMM),

Types de séquences biologiques

- **ADN** (Acide Désoxyribonucléique):
  - Bases : Adénine (A), Thymine (T), Cytosine (C), Guanine (G).
  - Contient les instructions génétiques nécessaires au développement et au fonctionnement des organismes.
- **Protéines**:
  - Constituées d'acides aminés (20).
  - Exemples : Alanine (A), Glycine (G), Leucine (L).

Remarque: un **résidu** est une **base** dans le cas d'un **nucléotide** ou un acide aminé dans le cas d'une protéine.

Évolution des séquences biologiques

Mutation

- **Concept:**
    - Les mutations entraînent des changement génétiques au cours de l'évolution,
    - Modifiant l’apparence et la fonction des séquences, contribuant ainsi à la diversité génétique et influençant l'évolution des espèces.

Évolution des séquences biologiques

Mutation

- **Type:**
  - **Substitution** : Remplacement d'un résidu par un autre.
    - Exemple: si dans une séquence d'ADN < $ATCG$ >, le C est remplacé par G, la séquence devient < $ATGG$ >.
  - **Insertion** : Ajout de résidus.
    - Exemple: en ajoutant un A dans la séquence < $ATCG$ >, on obtient < $ATCAG$ >
  - **Délétion** : Suppression de résidus.
    - Exemple: en supprimant T de < $ATCG$ >, on obtient < ACG >.

Évolution des séquences biologiques

Séquences homologues

- **Concept:**
    - Des séquences qui partagent une origine évolutive commune.
    - Elles proviennent d'un ancêtre commun et ont souvent des similitudes significatives dans leur structure ou fonction, bien qu'elles aient pu évoluer et subir des mutations au fil du temps.

Évolution des séquences biologiques

Séquences homologues

- **Exemple:**
  - La séquence ancestrale < $ATCGTACG$ >.
  - Au fil du temps, évolue et subi des mutations, donnant naissance à deux séquences homologues chez deux espèces différentes :
    - < $ATCGTTCG$ >
      - une substitution a remplacé A par T en sixième position
    - < $ATCGACG$ >.
      - une délétion a été introduite après le G en quatrième position

Alignement des Séquences Biologiques

- **Principe:** Identifier les similitudes entre séquences
- **Objectif:**
  - Détecter des séquences homologues et découvrir des mutations,
  - Construire des arbres phylogénétiques et reconstruire des séquences ancestrales, fournissant ainsi des informations clés pour retracer l'histoire évolutive des séquences et étudier leur divergence entre espèces.

Alignement des Séquences Biologiques

- **Méthode:**
  - Positionner les séquences de manière à maximiser leur similarité, permettant ainsi de repérer les mutations, telles que les substitutions, insertions et délétions, et de localiser précisément où elles se sont produites.
  - On utilise pour cela des **Gaps** (représentés par des tirets)
  - Les Gaps peuvent indiquer :
    - soit une perte de résidus par délétion dans la séquence comportant le gap,
    - soit un gain de résidus par insertion dans l'autre séquence.

Alignement des Séquences Biologiques

- **Exemple:**
  - X et Y sont deux séquences d'ADN homologues que nous voulons aligner

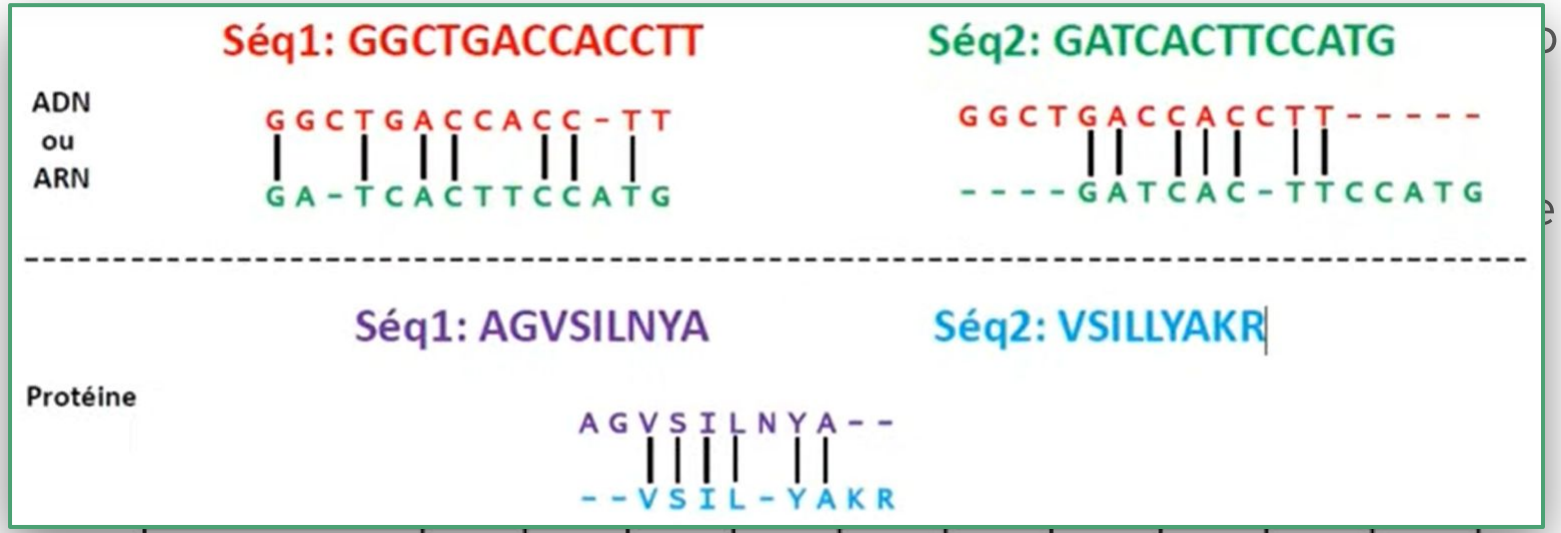
X=<ACCTGATCCG> Y=<ACTGATCAG>

  - On introduit un gap en position 3 de la deuxième séquence pour maximiser les similarités.

Position	1	2	3	4	5	6	7	8	9	10
X	A	C	C	T	G	A	T	C	C	G
Y	A	C	-	T	G	A	T	C	A	G

Alignement des Séquences Biologiques

- **Exemple:**



Y	A	C	-	T	G	A	T	C	A	G

Types d’Alignement des Séquences Biologiques

- **Alignement par paire**
  - Compare deux séquences à la fois.
  - L’objectif est de comprendre les relations entre deux séquences spécifiques.
- **Alignement multiple**
  - Alignement simultané de trois séquences ou plus
  - Identifier les régions conservées sur plusieurs séquences.
  - Ce type d’alignement est important pour l'analyse phylogénétique et l'étude des familles de protéines

Types d’Alignement des Séquences Biologiques

- **Alignement global**
  - Proposée par **Needleman-Wunsch**,
  - Comparer les séquences sur toutes leurs longueurs du début jusqu’à la fin,
  - Cela convient particulièrement lorsque les séquences entières présentent un intérêt.
- **Alignement local**
  - Proposée par **Smith-Waterman**,
  - Comparer les sous-séquences
  - Elle est utile pour les séquences qui peuvent avoir des motifs ou des domaines conservés mais qui diffèrent considérablement dans d'autres régions.

Méthodes d’évaluation de l’alignement

- Pour comparer les alignements de séquences, il est essentiel de déterminer un score qui estime la qualité et la robustesse de chaque alignement.
- Le score de similarité est la somme des scores des comparaisons individuelles entre les résidus alignés des deux séquences.
- L'objectif est d'atteindre le score maximal parmi tous les alignements possibles

Méthodes d’évaluation de l’alignement

- **Scores :**
  - Score d'**identité** (match) : pour les résidus identiques alignés,
  - Pénalité de **substitution** (mismatch) : pour les résidus différents alignés,
  - Pénalité de **gap** (indel) : pour l'alignement d'un résidu avec un gap.

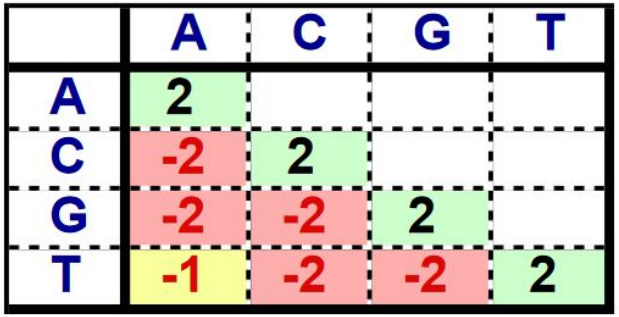
Méthodes d’évaluation de l’alignement

- **Exemple:**
  - un score d’identité de +1,
  - une pénalité de -3 pour une substitution,
  - une pénalité de -4 pour un alignement avec un gap.

Deux alignements avec des scores différents. L'alignement de gauche est meilleur que l'alignement de droite car son score global est plus élevé.

Méthodes d’évaluation de l’alignement

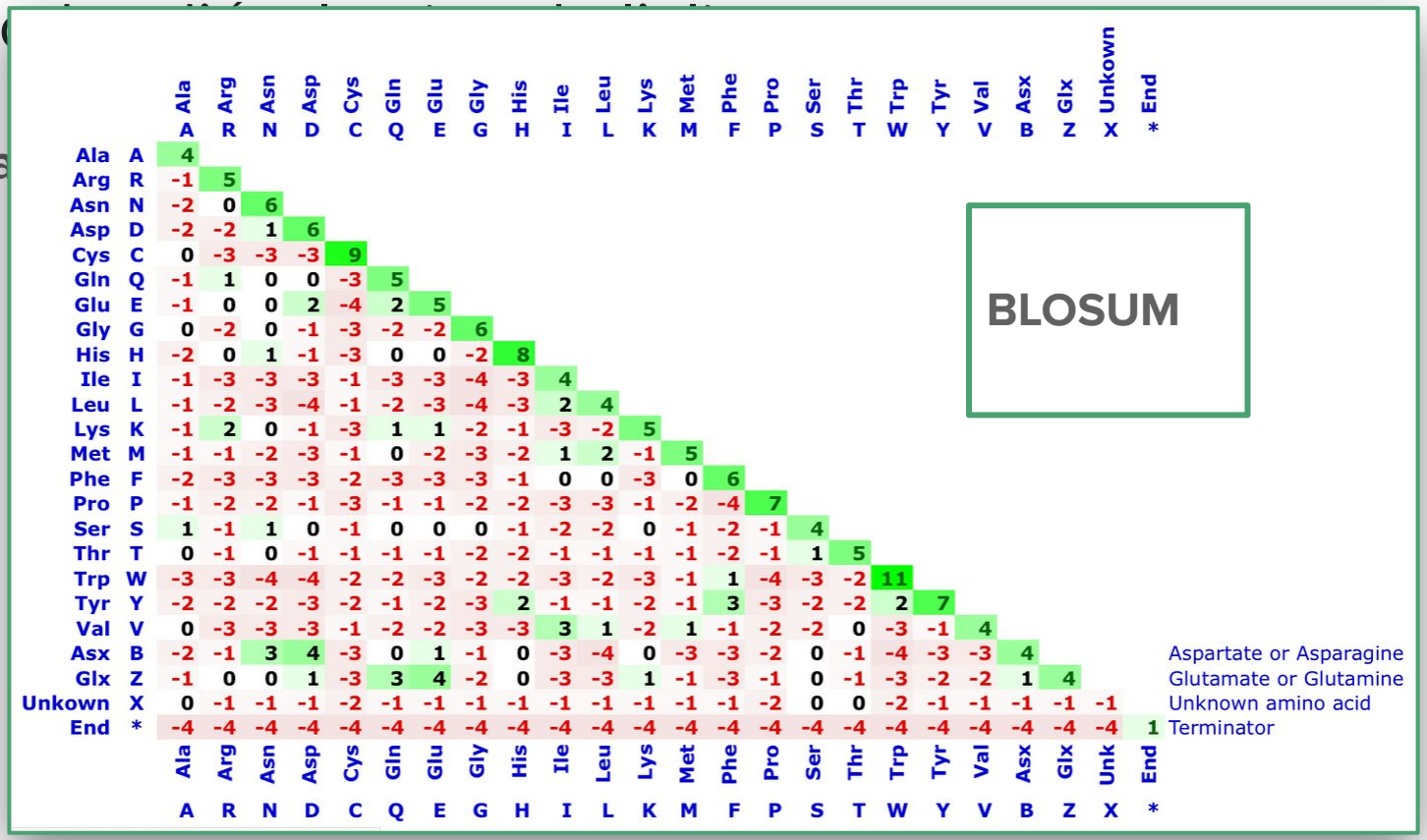
- **Matrices de substitution :**
  - Residue categories (Phylip)
  - PAM (Dayhoff, 1979).
  - BLOSUM (Henikoff & Henikoff, 1992).
  - ...



	A	C	G	T
A	2			
C	-2	2		
G	-2	-2	2	
T	-1	-2	-2	2

Méth

- Ma



		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	Unkown	End
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
Asx	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
Glx	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
Unkown	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Méth

- Ma
  -
  -
  -
  -

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	Asx B	Glx Z	Unkown X	End *
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1												
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3												
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
Asx	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			Aspartate or Asparagine
Glx	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		Glutamate or Glutamine
Unkown	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	Unknown amino acid
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1 Terminator

- Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acides aminés).
- La diagonale correspond aux identités.
- Le triangle inférieur correspond à des substitutions.
- Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.

Méthodes d’évaluation de l’alignement

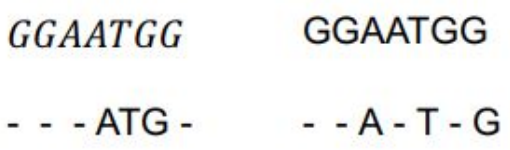
- **Matrices de substitution:**
  - attribuent des valeurs spécifiques pour les identités et substitutions entre résidus, en fonction de leurs propriétés.
  - Ces valeurs sont basées sur des facteurs biologiques.
  - Par exemple, l'alignement de deux cystéines peut recevoir un score plus élevé en raison de leur forte conservation et de leurs contraintes structurelles.

Méthodes d’évaluation de l’alignement

- **Coût d’alignement avec les gaps:**
  - **Coût de gap linéaire**
    - Chaque position d’un gap est facturée de façon identique.
    - $Coût\ total\ du\ gap = n \times coût\ du\ gap\ par\ site$ (où n est la longueur du gap)
    - Par exemple, un gap de trois positions coûtera trois fois celui d'un gap d'une seule position
  - **Coût de gap affiné**
    - Ce coût distingue l'ouverture et l'extension d'un gap, avec une pénalité plus élevée pour ouvrir un gap (O) et une pénalité moindre pour chaque extension (E).
    - $Coût\ total\ du\ gap = O + n \times E$
    - Ce modèle reflète mieux les insertions et suppressions en biologie

Méthodes d’évaluation de l’alignement

- **Coût d’alignement avec les gaps:**
  - **Exemple:** on a 2 alignement



- Un coût de gap linéaire évalue ces deux alignements de manière similaire.
- Un coût de gap affiné pénaliserait davantage le deuxième alignement, car il comporte trois gaps distincts (de longueurs 2, 1, et 1), contre deux gaps dans le premier (de longueurs 3 et 1).
- La réduction du coût pour les gaps terminaux favoriserait également le premier alignement, puisque le deuxième inclut un gap interne, alors que le premier n'a que des gaps terminaux.

Algorithme d’alignement

- L'alignement optimal de deux séquences est très coûteux (un problème NP-difficile),
- c'est pourquoi des méthodes heuristiques et de programmation dynamique sont souvent utilisées.
- Ces méthodes visent soit à maximiser les scores d'identité, soit à minimiser les pénalités en fonction des coûts définis.

Algorithme d’alignement (L'algorithme Needleman-Wunsch)

- Il s’agit d’un algorithme de **programmation dynamique** pour l’alignement **global** et optimal entre **deux** séquences.
- **Trois étapes** à suivre:
  - Remplir une matrice des scores
  - Retour arrière (Backtracing)
  - Génération de l’alignement

Algorithme d’alignement (L'algorithme Needleman-Wunsch)

1. **Initialisation de la matrice:**
   - Créer une matrice $M[n + 1][m + 1]$ où n et m sont les longueurs des séquences A et B.
   - Initialiser la première ligne et la première colonne de la matrice avec des valeurs multiples de la pénalité de gap.
2. **Remplissage de la matrice:**
   - Pour chaque cellule $M[i][j]$ calculer le score comme suit:

$$M[i][j] = max \begin{cases} M[i-1][j] + pénalité\ de\ gap \\ M[i-1][j-1] + score\ (A[i], B[j]) \\ M[i][j-1] + pénalité\ de\ gap \end{cases}$$

Algorithme d’alignement (L'algorithme Needleman-Wunsch)

1. **Traceback:**
    - Une fois la matrice remplie, effectuer un traceback à partir de la cellule $M[n][m]$ pour construire l'alignement optimal.
    - Suivre les flèches qui indiquent le chemin optimal jusqu'à atteindre le début des séquences.

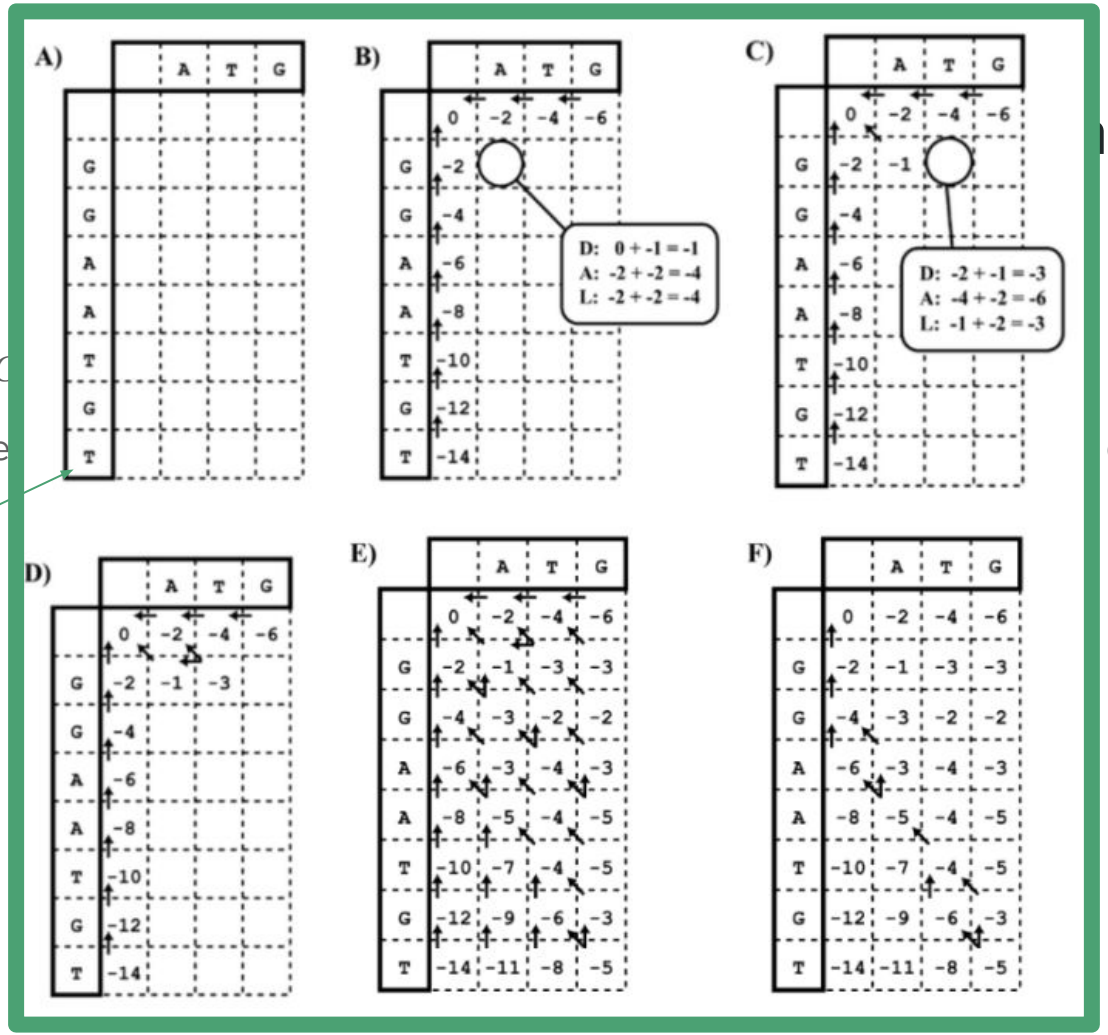
Algorithme d’alignement (L'algorithme Needleman-Wunsch)

- **Exemple:**
  - $A =< ATG >$ et $B =< GGAATGG >$,
  - un score de match de +1, une pénalité de mismatch de -1 et une pénalité de gap (indel) de -2

Algorithme n-Wunsch)

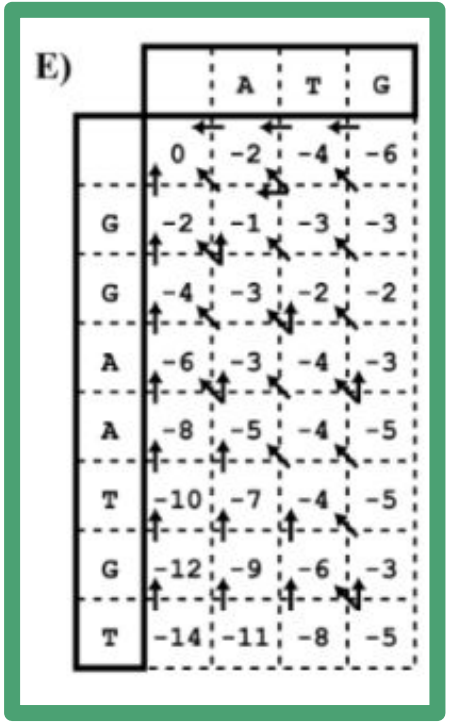
- **Exemple:**
  - $A =< ATC$
  - un score gap (indel) de -2

G



Algorithme d’alignement (L'algorithme Needleman-Wunsch)

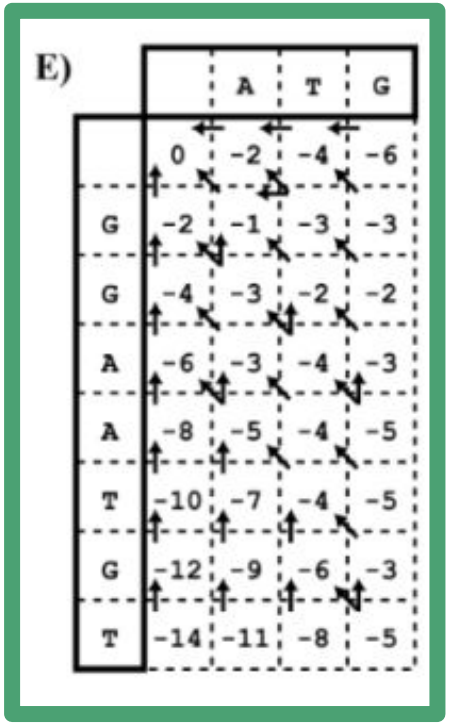
- **Exemple:**
  - $A =< ATG >$ et $B =< GGAATGG >$,
  - **Flèche diagonale :** Les résidus correspondants à la ligne et à la colonne de la matrice doivent être alignés.
  - **Flèche verticale :** Le résidu de la séquence sur l'axe vertical doit être aligné avec un gap dans la séquence sur l'axe horizontal.
  - **Flèche horizontale :** Le résidu de la séquence sur l'axe horizontal doit être aligné avec un gap dans la séquence sur l'axe vertical.



Algorithme d’alignement (L'algorithme Needleman-Wunsch)

- **Exemple:**
  - $A =< ATG >$ et $B =< GGAATGG >$,
  - Si plusieurs chemins mènent au sommet de la matrice, cela signifie qu'il existe plusieurs alignements qui donnent le même score, et tous ces alignements sont considérés comme optimaux:

GGAATGG	GGAATGG	GGAATGG	GGAATGG
- - - ATG -	- - - AT - G	- - A - TG -	- - A - T - G



Analyse des Séquences Biologiques

- **Définition :** Modélisation probabiliste pour analyser des séquences pour en découvrir les significations et les fonctions.
- **Algorithme:**
  - Les chaînes de Markov,
  - Les modèles de Markov cachés (HMM)
- **Particularité:** Ces modèles sont efficaces car ils supposent que la probabilité d'un état donné dépend uniquement de l'état précédent, ce qui les rend particulièrement adaptés aux données séquentielles biologiques.

Analyse des Séquences Biologiques (Les chaînes de Markov)

- **Définition :**
  - Les chaînes de Markov sont des systèmes mathématiques qui passent d'un état à un autre au sein d'un ensemble fini d'états.
  - L'état futur dépend uniquement de l'état actuel, et non de la séquence d'événements qui le précède
- **Application:** Modélisation de séquences biologiques pour la prédiction.

Analyse des Séquences Biologiques (Les chaînes de Markov)

- **Méthode:**
  - Définir les états de la chaîne de Markov.
  - Créer une matrice de transition avec les probabilités de passage d'un résidu à un autre.
  - Utiliser la matrice pour prédire le prochain résidu.
- **Exemple:** <AGCTAGCAGT>
  - Nous pouvons ainsi prédire le prochain nucléotide en fonction du nucléotide actuel.
  - Par exemple, si le nucléotide actuel est G, le prochain nucléotide sera probablement G avec une probabilité de ⅔ ou T avec une probabilité de ⅓.

	A	C	G	T
A	0	0	1	0
C	½	0	0	½
G	0	⅔	0	⅓
T	1	0	0	0

Analyse des Séquences Biologiques (Modèles de Markov Cachés (HMM))

- **Définition:**
  - des modèles probabilistes qui génèrent des séquences via des transitions entre des états de Markov, incluant des états cachés non observables directement.
- **Applications :**
  - Identification de régions codantes (exons) et non-codantes (introns).
  - Ce processus s’appelle en biologie l'annotation des gènes
  - Algorithmes associés : Forward, Viterbi, Baum-Welch.

Analyse des Séquences Biologiques (Modèles de Markov Cachés (HMM))

- **États cachés :** (on ne sait pas a priori quelles parties de la séquence sont codantes ou non.)
  - **Codant :** correspond aux régions de l'ADN qui produisent des protéines.
  - **Non-codant :** correspond aux régions qui ne produisent pas de protéines.
- **Symboles observables :**
  - Les bases d'ADN : A, T, C, G.
  - Celles-ci sont visibles dans la séquence et peuvent être observées directement.
- **Transitions entre états :**
  - La séquence peut alterner entre les régions codantes et les régions non-codantes. Par exemple, une région codante peut être suivie d'une région non-codante et vice versa.

Analyse des Séquences Biologiques (Modèles de Markov Cachés (HMM))

- **Probabilités d'émission :**
  - Pour chaque état, il existe une probabilité spécifique d'observer chaque base.
  - Par exemple :
    - Dans un état "codant", certaines bases peuvent être plus fréquentes en raison des préférences génétiques (par exemple, certains triplets de bases sont souvent observés dans des régions codantes).
    - Dans un état "non-codant", la distribution des bases peut être plus aléatoire.

Analyse des Séquences Biologiques (Modèles de Markov Cachés (HMM))

- **Utilisation de l’HMM pour l’identification:**
  - En appliquant l'HMM sur la séquence d'ADN, on peut estimer la séquence des états cachés (codant ou non-codant) pour chaque segment de la séquence d'ADN observée.
  - L'HMM utilise les probabilités de transition (passage de codant à non-codant) et les probabilités d'émission (chances de voir A, T, C, G dans chaque état) pour "deviner" où se trouvent les gènes.
  - Grâce à des algorithmes comme Viterbi, on peut calculer la séquence d'états la plus probable pour chaque base et ainsi identifier les zones codantes.