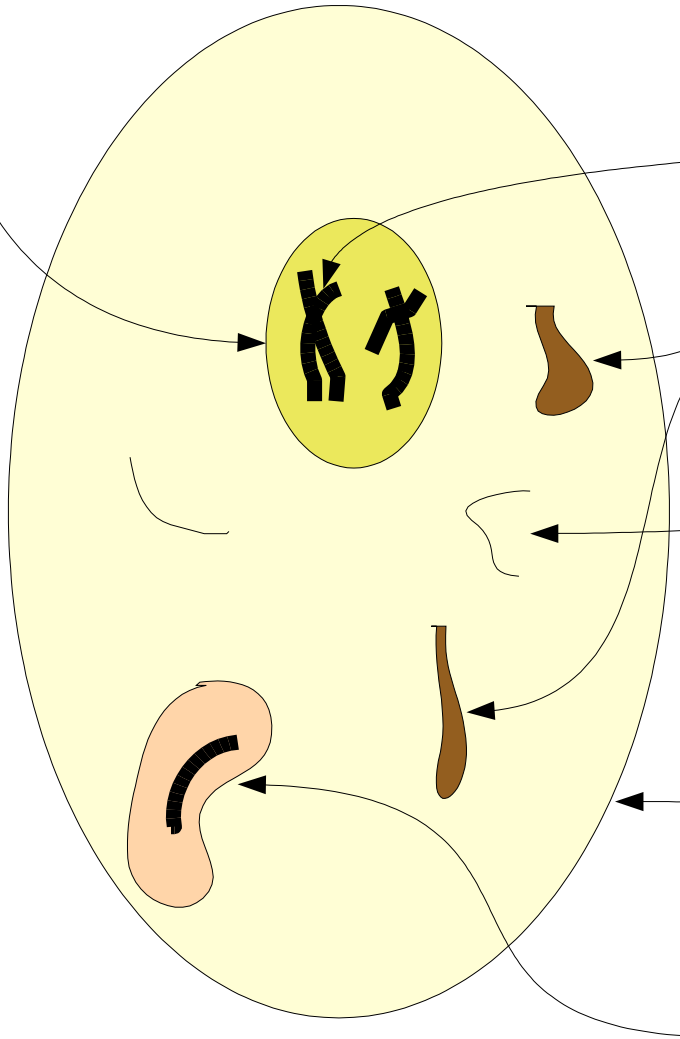
Quelques approches mathématiques en génomique

Notions de bases en biologie

Analyses de séquences

Analyses de données de transcriptome

Une cellule



Noyau

ADN
nucleotides {A, C, G, T}
de 10^4 à 10^{11}

Ribosomes

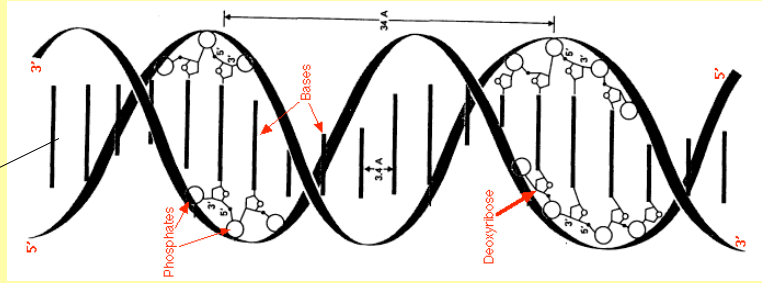
ARN
nucléotides {A, C, G, U}
jusqu'à 10^6

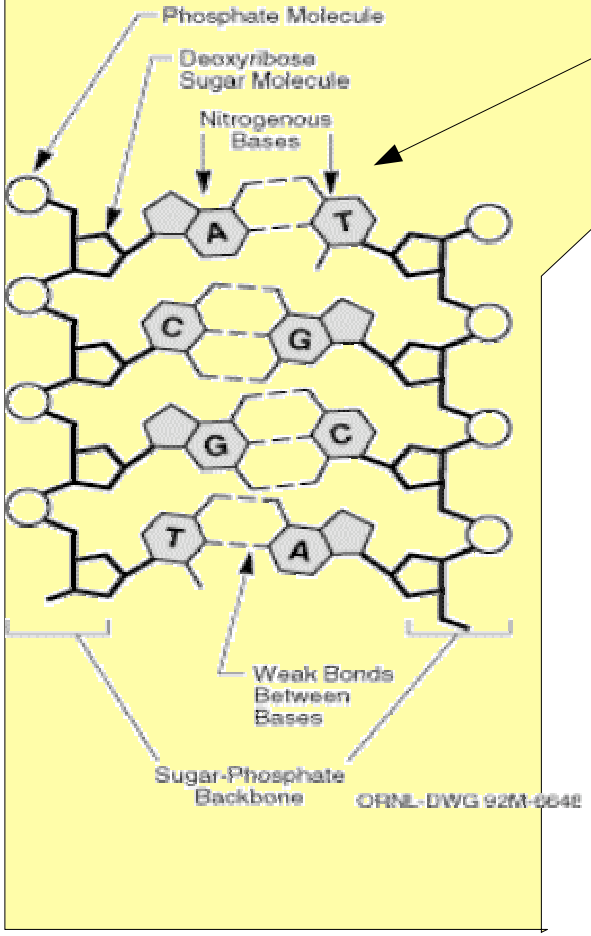
Protéines
20 acides aminés
jusqu'à 10^3

Mitochondries, Chloroplastes

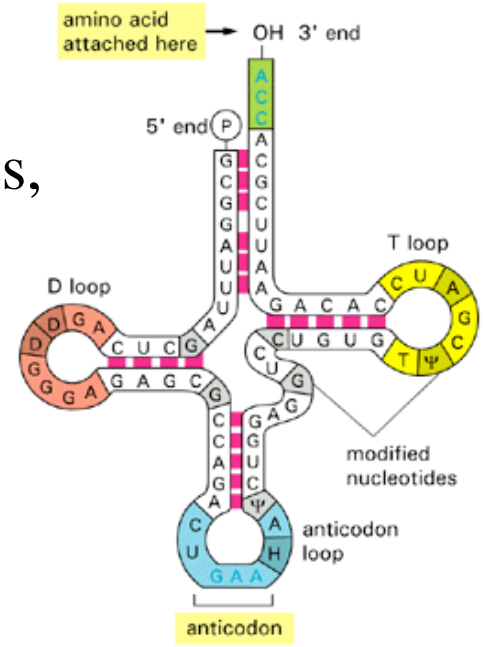
Structures

ADN: double brin (la sequence et sa complementaire reverse), double hélice

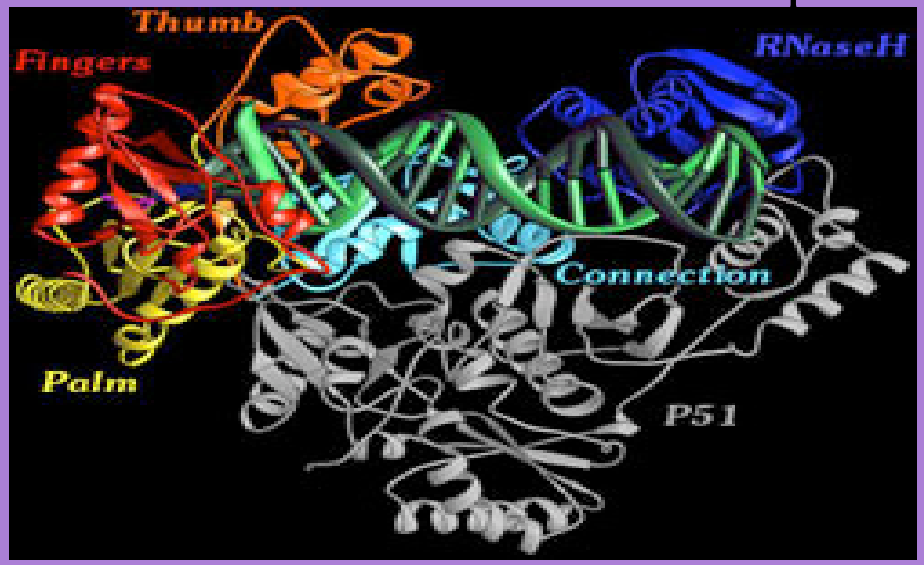




ARN: simple brin, structures parfois complexes (ribosomes, tARN)

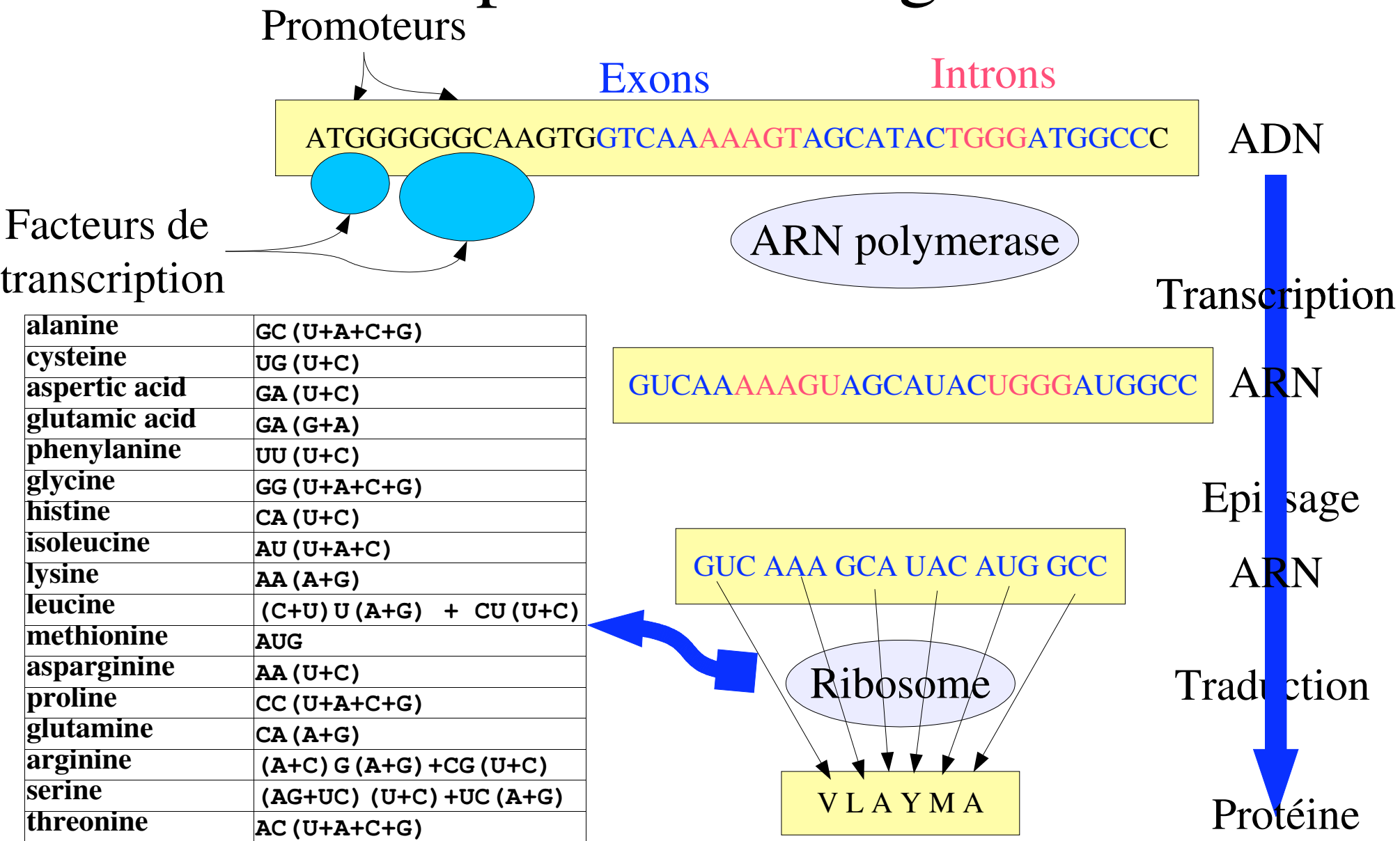


Proteines : structures très complexes



La structure est essentielle à la fonction, pas toujours déterminée par la séquence.

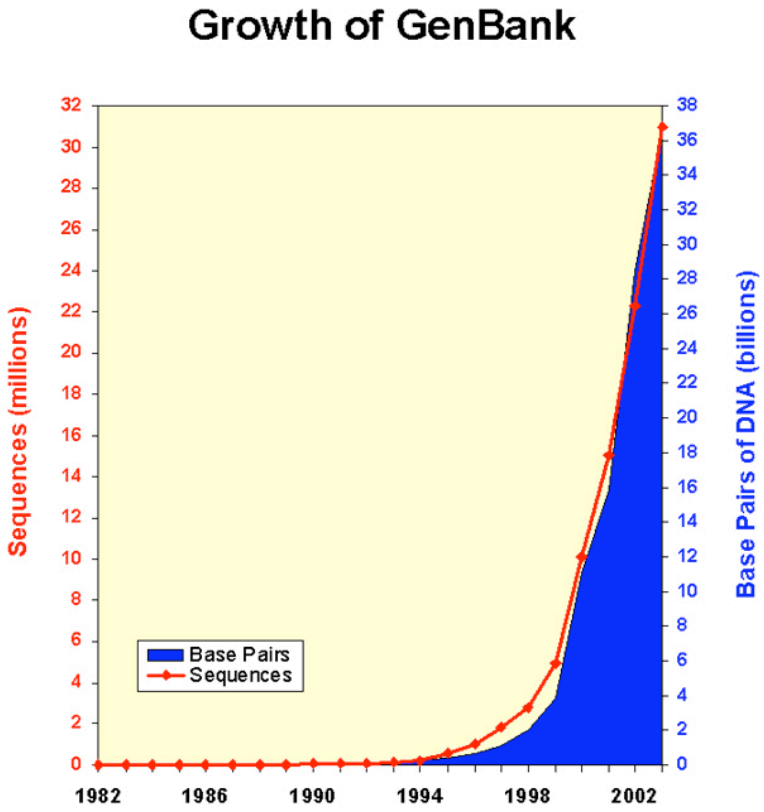
Expression des gènes



alanine	GC (U+A+C+G)
cysteine	UG (U+C)
aspertic acid	GA (U+C)
glutamic acid	GA (G+A)
phenylanine	UU (U+C)
glycine	GG (U+A+C+G)
histine	CA (U+C)
isoleucine	AU (U+A+C)
lysine	AA (A+G)
leucine	(C+U) U (A+G) + CU (U+C)
methionine	AUG
asparginine	AA (U+C)
proline	CC (U+A+C+G)
glutamine	CA (A+G)
arginine	(A+C) G (A+G) +CG (U+C)
serine	(AG+UC) (U+C) +UC (A+G)
threonine	AC (U+A+C+G)
valine	GU (U+A+C+G)
tryptophan	UGG
tyrosine	UA (U+C)

Données disponibles

Les séquences d'ADN, ARN et protéines, collectées depuis les années 70, grandes bases de données (EMBL, GenBank, SwissProt etc.)



Les quantités d'ARNm et de protéines (transcriptome, protéome), collectées depuis les années 90, quelques bases de données (Stanford Microarrays Database)

Analyse de séquences

Modélisation/Annotation

Comparaison

Evolution

Smodélisation de séquences

Les séquences ne sont pas homogènes

- À petite échelle, à cause de leur fonction: parties codantes ou pas, télomères etc.
- À grande échelle : isochores (Bernardi, Karlin)

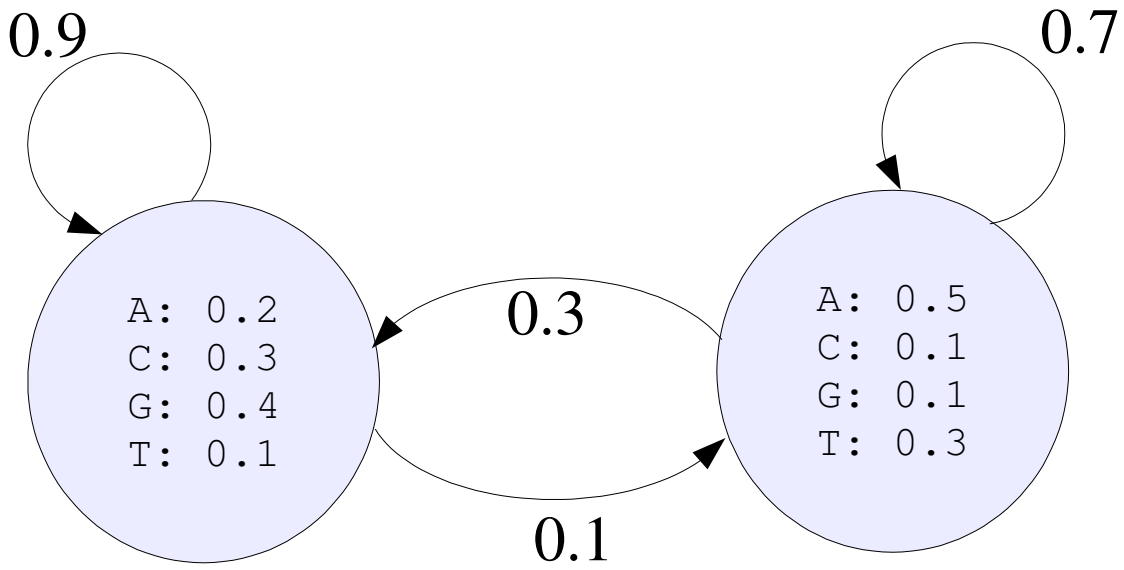
Modèles nécessaires:

- Pour évaluer la significativité de diverses mesures
  exemple: trouver des mots avec des fréquences inattendues (Prum)
- Pour reconnaître les sous-parties des séquences – annotation

Markov et Markov à état cachés

Markov à états cachés

Exemple:



Markov à états cachés

Question 1 – Décoder: Identifier les états à partir d'un ensemble d'observation (d'émissions)

Maximum de vraisemblance : algorithme de Viterbi

Question 2 - Apprentissage: Estimer les paramètres (de transitions et d'émissions) à partir d'un ensemble d'observation (d'émissions)

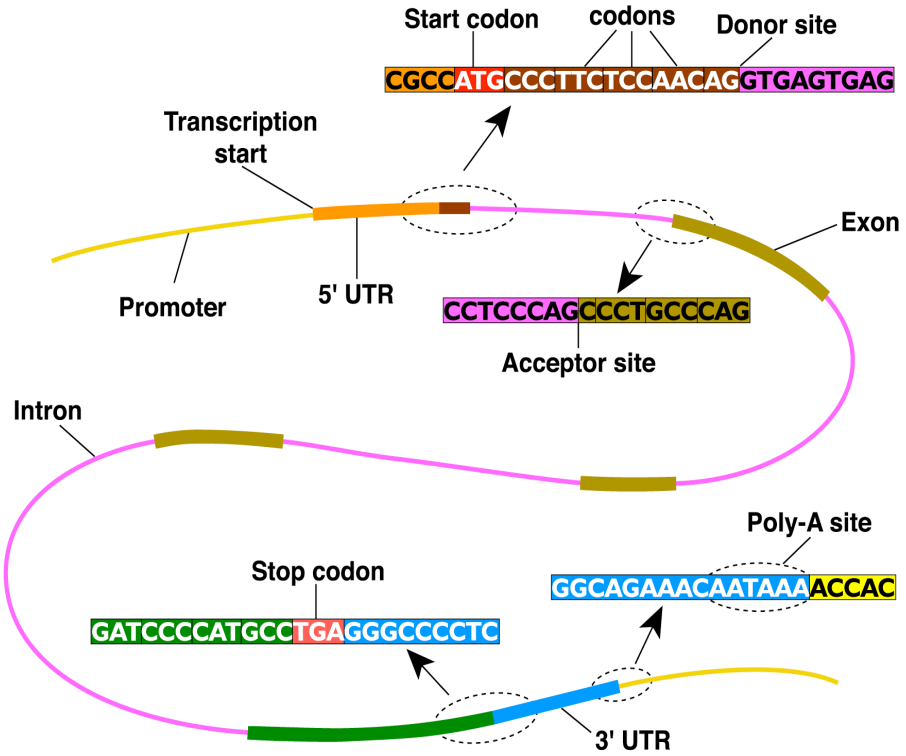
Maximum de vraisemblance: Baum and Welsh

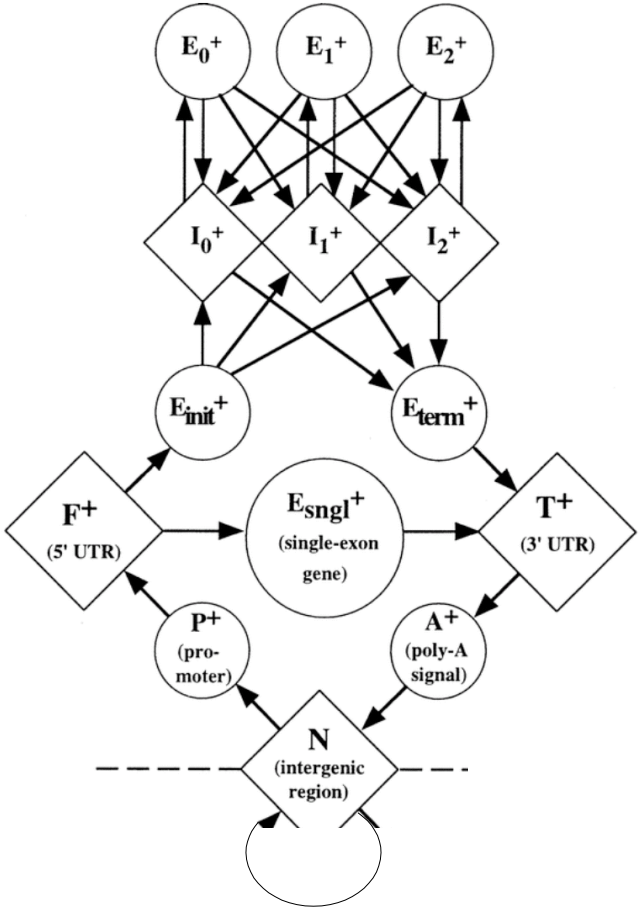
Consistance de l'estimateur de maximum de vraisemblance pour les HMM (Leroux)

Annotation de séquences

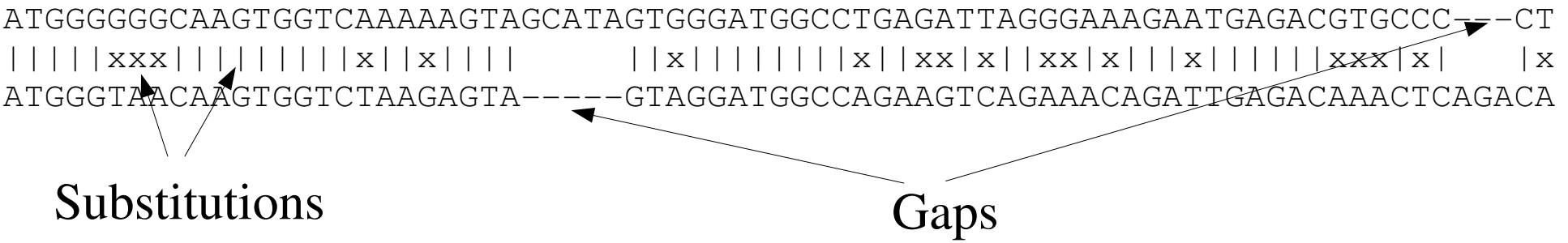
Problème: Identifier les parties d'une séquence brute.

Trouver les gènes HMM (GenScan – Burge, K





Comparaison de séquences – Alignement



Score:

Matrice de substitution

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

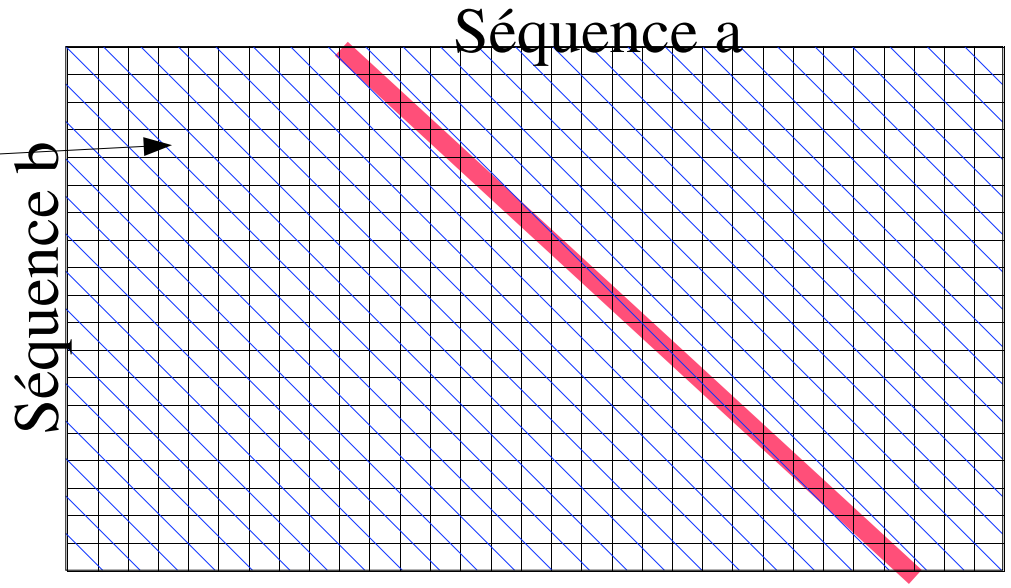
Penalité de gap fonction lineaire or affine de la longueur du gap

Score = somme score de subsititution individuels – pénalités de gap

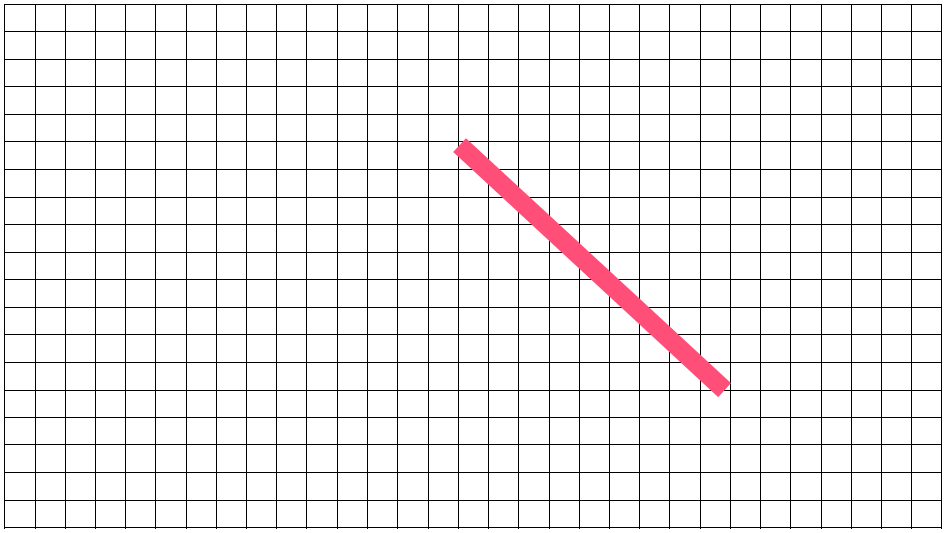
Alignement global et local

Cas des alignements sans gap

(i,j) = score de substitution entre la lettre en position i de la sequence a et celle en position j de la séquence b

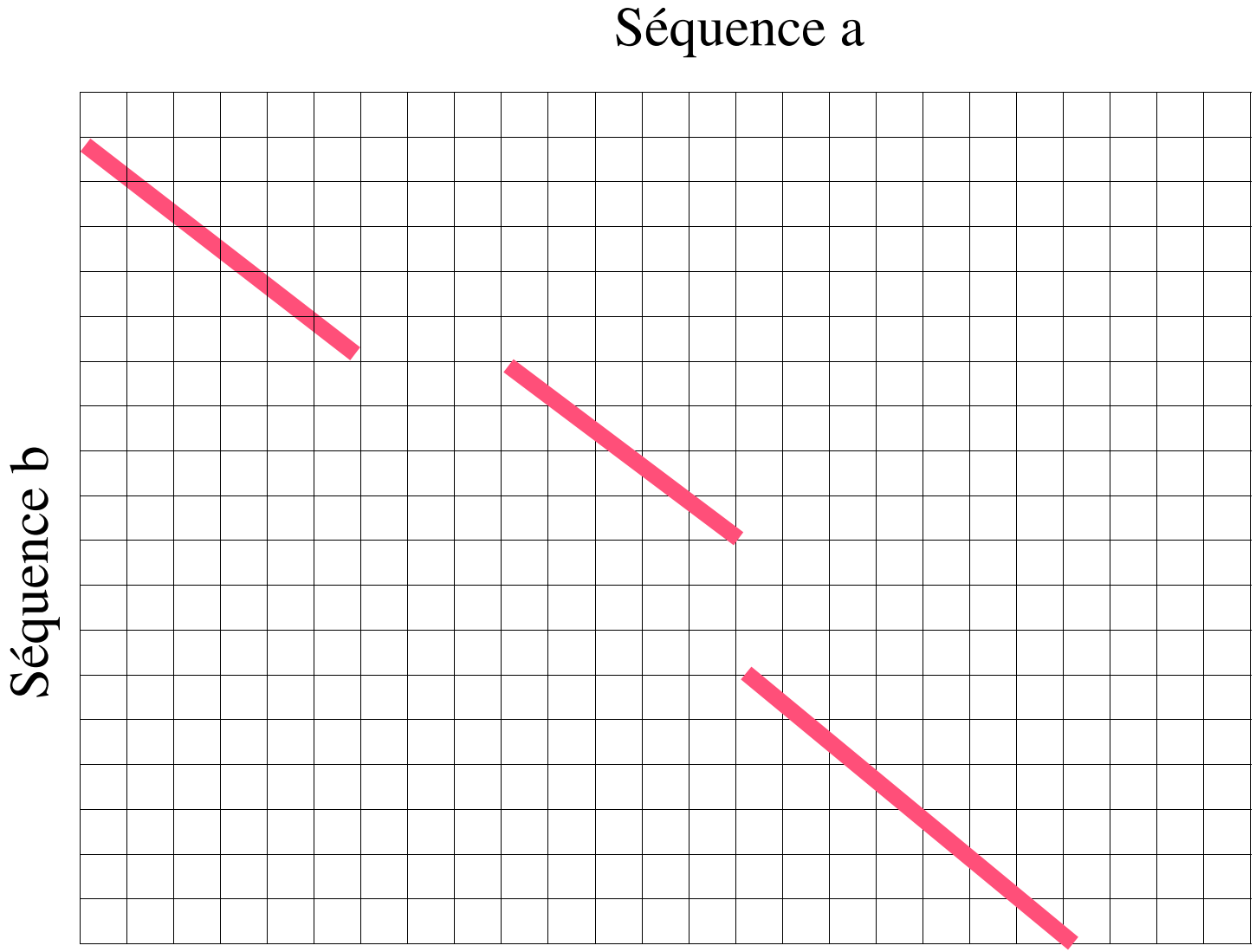


Alignement global :
On aligne l'ensemble des sequences



Alignement local :
Trouve deux sous-séquence avec un score maximum

Alignment avec gaps



Transition de phase pour les alignements locaux

L'espérance du score maximum d'un alignement local de deux séquences iid de longueur N croît en O(N) ou en O(log(N)) selon les paramètres du score.

Dans le cas d'un alignement local sans gap sur un alphabet de c lettres avec comme matrice de substitution (a_{ij}) $a_{ii}=1$ $a_{ij}=-m$ if i différent de j,

Si $m<1/(c-1)$, l'espérance et linéaire sinon logarithmique.

(Arratia, Waterman, Dembo, Karlin).

Significativité d'un aligment

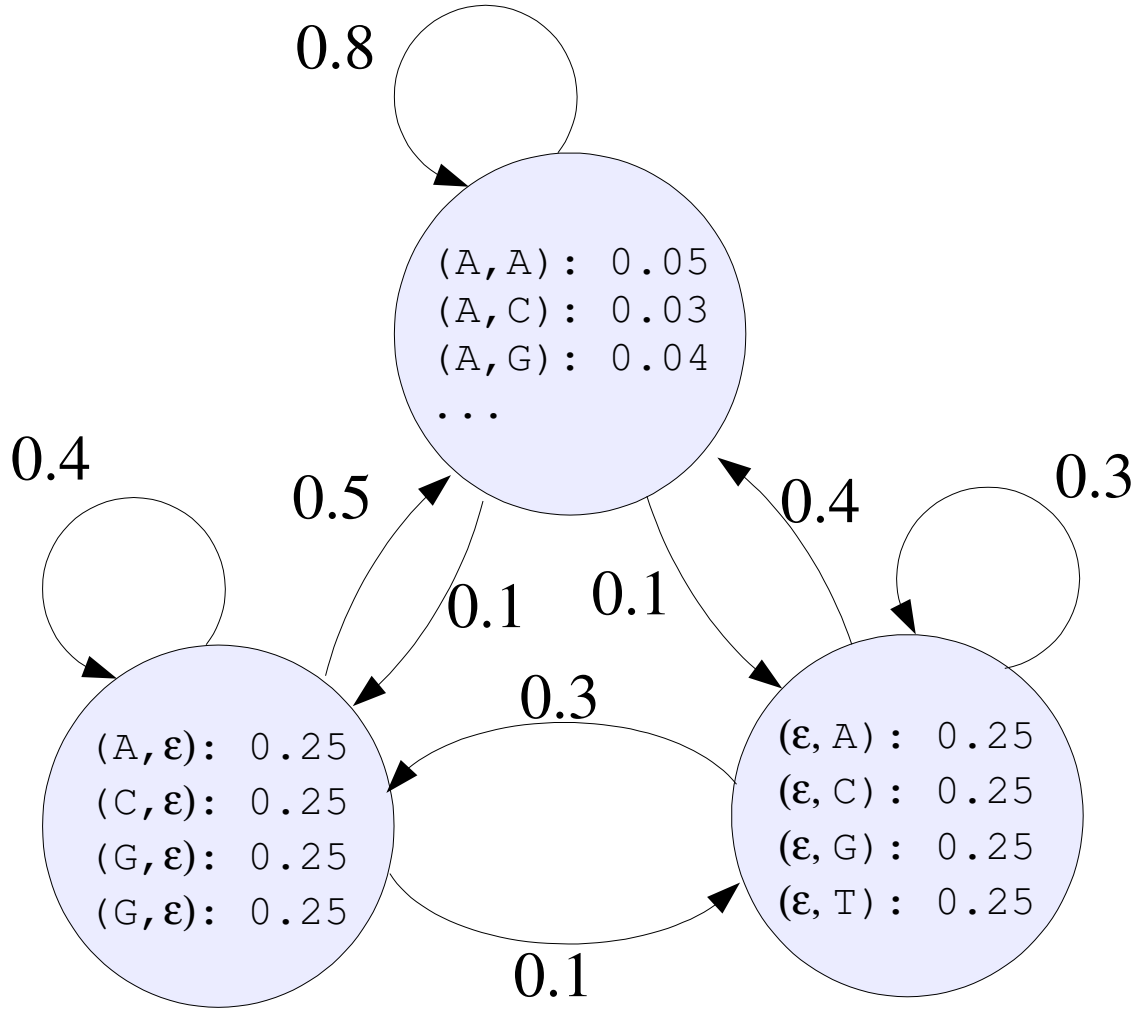
Problème : sous des hypothèses raisonnables sur les séquences (iid, Markov), calculer la probabilité d'obtenir un score supérieur à une valeur donnée

Alignement global : la loi du score est inconnue...
quelques résultats pour la plus longue sous séquence commune
i. e. score de substitution = identité pas de pénalité de gap
(Chvàtal, Sankoff, Dancik, Martinez)

Alignement local : sous différentes restrictions sur les paramètres, la loi du score maximum (p-value) suit une distribution de valeur extrème:
$P(M>s) = 1-\exp(-K.m.n.\exp(-\lambda s))$
L'outil de recherche BLAST renvoie aussi une E-value (espérance du nombre de sous-séquences dans la base avec un score plus grand que s)
(Altschul, Karlin, Dembo, Waterman,Vingron, Mott, Hwa, Bundschul, Siegmund,Yakir...)

Alignement – Pairing et multiple HMM

HMM dont les symboles d'émission sont {A, C, G, T, ε}x{A, C, G, T, ε}



Durbin software: Hmmer, SAM

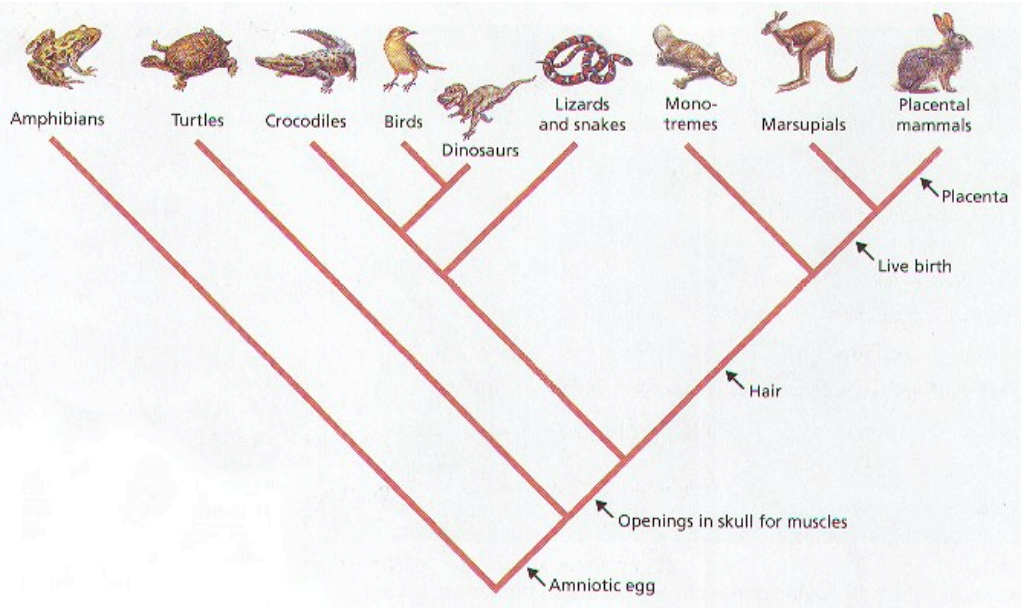
Alignment multiple

Très haute complexité algorithmique

Méthode heuristiques souvent basées sur les alignement par paires des séquences

Phylogenie

Question: reconstruire la généalogie d'un ensemble de séquence



Les données sont en général un alignement multiple de séquences

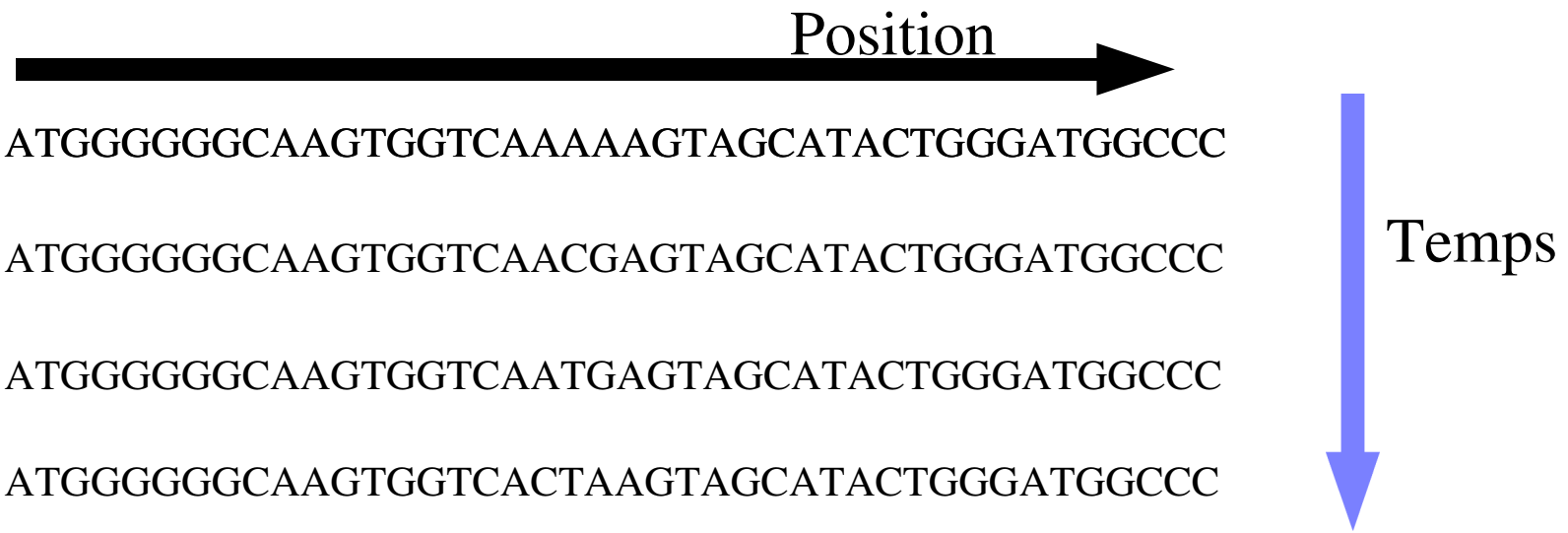
Trois grandes classes de méthodes :

- Distance (clustering hiérarchique, distance d'arbre)
- Parsimonie (cout de transformation)
- Maximum de vraisemblance (modèle d'évolution)

Evolution de séquences

Hypothèse habituelles:
Le processus d'évolution :

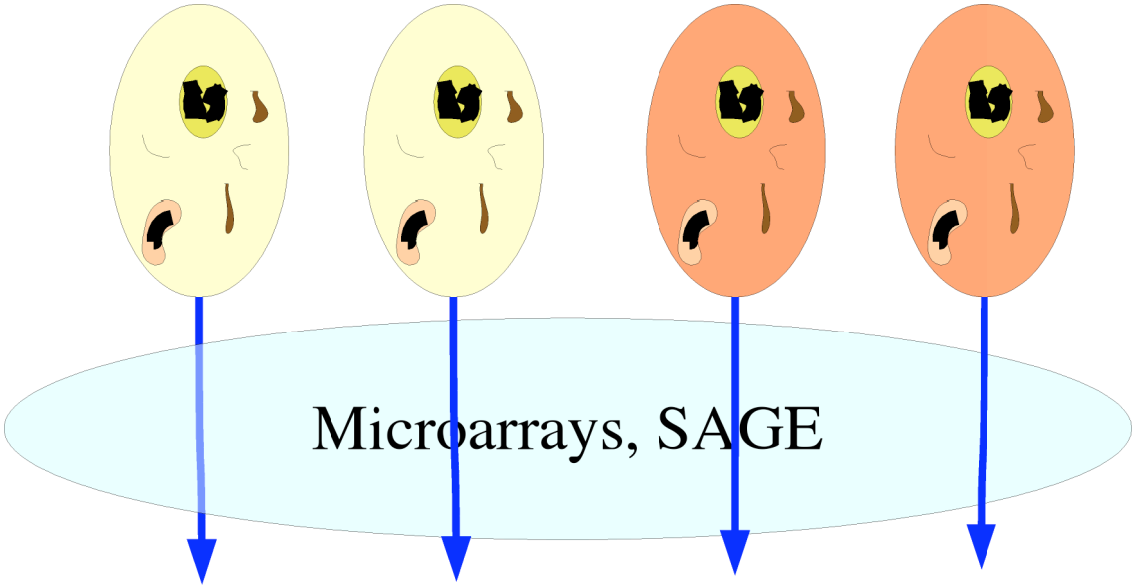
- N'a pas de mémoire (Markovien)
- Est homogène en temps et position



Matrice PAM (Dayhoff)

matrice BLOSUM (Henikoff)

Analyses de données de transcriptome



Gène	Condition A	Condition A	Condition B	Condition B
ytvB	593,24	642,07	412,52	412,34
yesX	1303,86	1406,24	961,97	996,28
lgt	769,97	837,97	755,45	776,41
yqaE	827,72	930,24	547,62	590,38
ykkD	1959,14	2136,97	1513,69	1493,24
ycbJ	598,41	495,9	436,52	412,52
yjbQ	1729,69	1803,21	1287,31	1319,48
ppsC	871,69	877,83	570,72	590,41

Jusqu'à de l'ordre de 10^4 mesures simultanées.
Tous les gène d'un organisme

Niveau d'expression d'un gène = quantité de son ARNm

Questions

Question1: Identifier les gènes différentiellement exprimés entre deux conditions ou plus

Test de student, de Wilcoxon, ANOVA

Question 1bis: Reconnaître des type de cellules d'après l'expression des gènes

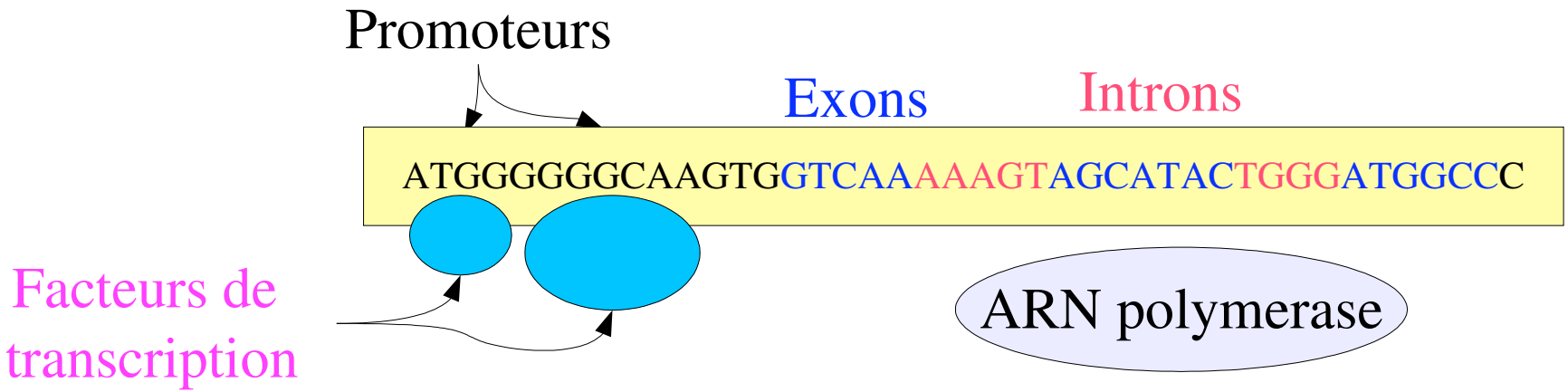
Analyse discriminante, arbre de décision, SVM

Question 2: Trouver des groupes de gène co-exprimés

Distance entre « gènes »: Spearman, corrélation...
Méthode de classification : UPGMA, SOM

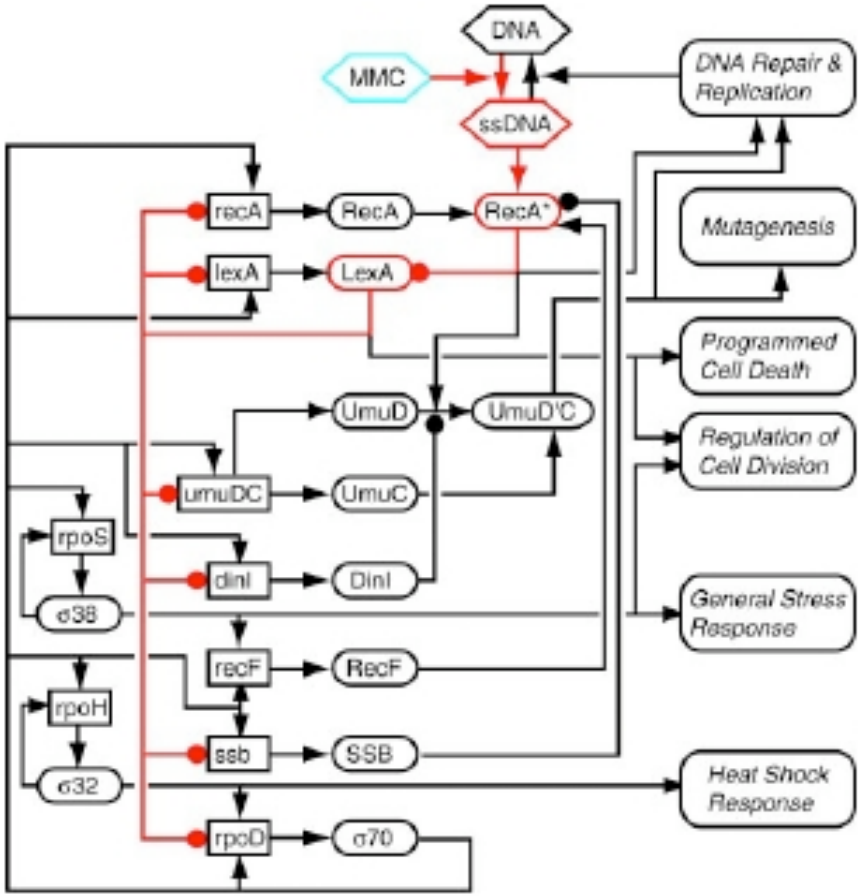
(Eisen, Churchill)

Réseaux de régulation



Problème : Etudier les interaction entre les gènes

Réseaux de régulation



Réponse d' E. coli après un domage sur l'ADN, Gardner

Question 1: Etudier la dynamique d'un réseau donné

Question2: Inférer le réseau de régulation à partir de données expérimentales (données d'expression)

Réseaux Bayesiens, Booléens, de Petri

Organism	Taille du génome	Nombre de gènes
Human (*Homo sapiens*)	3 billion	30,000
Laboratory mouse (*M. musculus*)	2.6 billion	30,000
Mustard weed (*A. thaliana*)	100 million	25,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	12.1 million	6,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9700	9