Comparaisons locales et matrices de score

Équipe Bonsai

http://www.lifl.fr/bonsai

année 2013

Contenu du cours

- Matrices de scores
- Recherches locales : BLAST et FastA

Exemples pour l'ADN

Identité (similarité)

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

BLAST (similarité)

	A	C	G	T
A	1	-3	-3	-3
C	-3	1	-3	-3
G	-3	-3	1	-3
T	-3	-3	-3	1

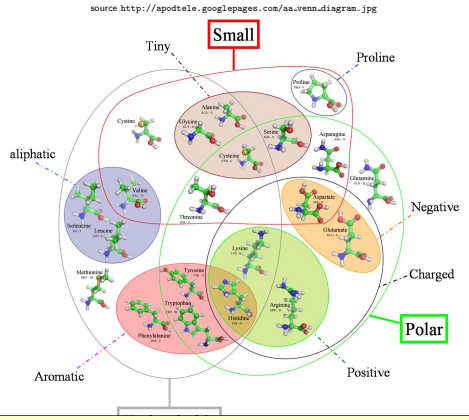
Transition/Transversion

	A	C	G	T
A	5	-4	-3	-4
C	-4	5	-4	-3
G	-3	-4	5	-4
T	-4	-3	-4	5

Matrice BLOSUM-62 pour les protéines

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =  0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```


Matrice BLOSUM-62 pour les protéines



De l'importance des matrices de scores

- impliquées dans toutes les analyses par comparaison de séquences
- résultats fortement dépendants de la matrice
- représentent implicitement une théorie de l'évolution (matrices protéiques)
- la compréhension d'une matrice ⇒ un bon choix

Similarité vs. distance

- un élément de la matrice représente :
  - le coût du remplacement d'une base par une autre (distance)
  - la mesure de la similarité du remplacement
- association entre
  - distance → phylogénie
  - similarité → recherche dans des bases de données
- même principe de recherche :
  - **maximiser un score ≡ minimiser une distance**
  - matrice de distance et de similarité peuvent être déduites l'une de l'autre

Comment obtenir une telle matrice ?

pour l'ADN

⇒

souvent données de manière *ad hoc*

Comment obtenir une telle matrice ?

pour les protéines

- matrices log odds ratio

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- exprime le ratio entre :
  la probabilité que deux résidus i et j soient alignés par descendance
  et la probabilité que ceux-ci soient alignés par chance
- explication :
  - q_{ij} = la fréquence que l'alignement de i et j soit observé dans des séquences homologues.
  - p_i = la fréquence d'occurrence de i
  - un score est > 0 si la proba d'un match significatif est $>$ à la proba d'un match aléatoire

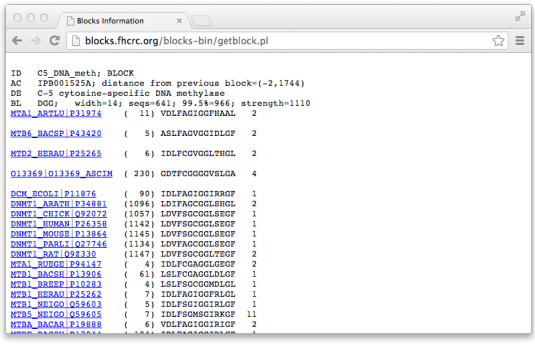
⇒ matrices PAM et BLOSUM

Matrice BLOSUM

BLocks SUbstitutions Matrices
Henikoff & Henikoff, 1992

- fréquence de changements entre deux acides aminés avec conservation de structure
  échantillon : BLOCKS (alignements multiples sans gaps)
- BLOSUM-N : seuil de similarité, N = % de similarité
- convient bien pour la recherche de similarités locales
- la plus courante : BLOSUM-62 (matrice par défaut de BLAST)

Matrice BLOSUM - Matériel



Matrice BLOSUM - Schéma de Construction

1. éliminer les alignements d'identité $< n\%$ ⇒ **BLOSUM-n**.
2. compter le nombre $\mathbf{f_{ij}}$ de paires d'aa i et j.
3. calcul de la fréquence $\mathbf{q_{ij}}$ des paires d'aa i et j :

$$q_{ij} = \frac{f_{ij}}{\#\text{ total de paires}}$$

4. calcul de la fréquence *marginale* $\mathbf{p_i}$ des aa i :

$$p_i = \sum_j q_{ij}$$

(plus exactement $p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2$)

5. calcul de la matrice log odds $\mathbf{S_{ij}}$:

$$S_{ij} = c \times \log \frac{q_{ij}}{p_i p_j}$$

(plus exactement $S_{i \neq j} = c \times \log \frac{q_{ij}}{2 p_i p_j}$ et $S_{ii} = c \times \log \frac{q_{ii}}{p_i p_i}$)

Matrice PAM

Accepted Point Mutation / Percent Accepted Mutation
Dayhoff, 1979

- fréquence de changements entre acides aminés
  Reconstitution de l'évolution avec la construction d'arbres phylogénétiques pour 71 familles de protéines
- convient bien pour les séquences avec un ancêtre commun
- possibilité de choix d'une matrice en fonction de l'évolution supposée
  PAM-*N* : *N* mutations acceptées par 100 acides aminés
- si la distance mutationnelle n'est pas connue, faire plusieurs essais
  PAM 40, PAM 120, PAM 250, par exemple.

Matrice PAM - Construction

1. alignements globaux de familles de protéines (identité > 85%)
2. reconstruction d'une phylogénie et des ancêtres (71 familles)
3. compter le nombre de fois $\mathbf{A_{ij}}$ où un aa i est remplacé par un aa j dans toutes les comparaisons 2 à 2
4. estimation de la mutabilité $\mathbf{m_j}$ d'un aa j
5. calcul de la matrice de probabilité de mutations

$$M_{ij} = \lambda \frac{m_j A_{ij}}{\sum_i A_{ij}} \text{ et } M_{jj} = 1 - \lambda m_j$$

m_j : probabilité pour j de muter

$\frac{A_{ij}}{\sum_i A_{ij}}$: probabilité conditionnelle pour j, s'il mute, de muter en i

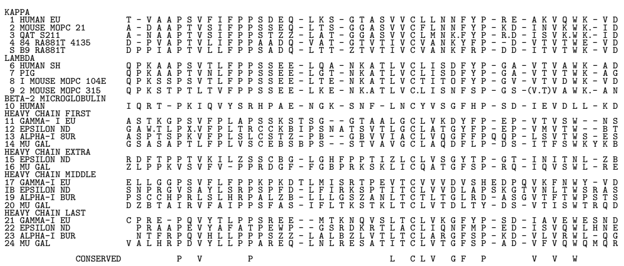
6. calcul de la matrice log odds

$$S_{ij} = \log \frac{M_{ij}}{p_i}$$

$S_{ij} = \log \frac{p_j \times M_{ij}}{p_i p_j}$: comme p_j est la proba de j dans une séq., et M_{ij} la prob, pour un occ. j, de muter en i durant un laps de temps donné, le

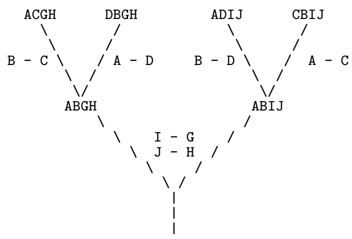
Matrice PAM - Construction (1/6)

alignements de 71 groupes de l'*Atlas of protein sequence*



Matrice PAM - Construction (2/6)

phylogenie et reconstruction des ancetres pour chacun des 71 groupes



Matrice PAM - Construction (3/6)

nombre d'*accepted point mutations*

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala																				
arg	30																			
asn	109	17																		
asp	154	0	532																	
cys	33	10	0	0																
gln	93	120	50	76	0															
glu	266	0	94	831	0	422														
gly	579	10	156	162	10	30	112													
his	21	103	226	43	10	243	23	10												
ile	66	30	36	13	17	8	35	0	3											
leu	95	17	37	0	0	75	15	17	40	253										
lys	57	477	322	85	0	147	104	60	23	43	39									
met	29	17	0	0	0	20	7	7	0	57	207	90								
phe	20	7	7	0	0	0	0	17	20	90	167	0	17							
pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Matrice PAM - Construction (4/6)

Exemple de calcul :

```
Alignement                A D A
                          A D B

Acides aminés             A    B    D
Chgt. observés            1    1    0
Fréquence d'occ.          3    1    2
Mutabilité relative     .33    1    0
```

Mutabilités relatives de tous les aa :

Ser	149	Ala	100	Gln	98
Met	122			Asp	90
Asn	111			Thr	90
Ile	110			Gap	84
Glu	102			Val	80
				Lys	57
				Pro	56
				His	50
				Gly	48
				Phe	45
				Arg	44
				Leu	38
				Tyr	34
				Cys	27
				Trp	22

Matrice PAM - Construction (5/6)

probabilités de mutation 1-PAM (une mutation pour 100 aa),
(ajustement obtenu grâce à λ)

probabilités *à l'échelle* $\times 10000$

	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
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	0	1	0	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Matrice PAM

probabilites de mutation PAM-250 = (1-PAM)^{250}

probabilités *à l'échelle* $\times 100$

	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
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	1	0	1	0	1	0	0	1	0	1	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Matrice PAM - Construction (6/6)

Fréquences et log odds matrice PAM-250 :

Ala	0.096	Cys C	12																				
Gly	0.090	Ser S	0	2																			
Lys	0.085	Thr T	-2	1	3																		
Leu	0.085	Pro P	-3	1	0	6																	
Val	0.078	Ala A	-2	1	1	1	2																
Thr	0.062	Gly G	-3	1	0	-1	1	5															
Ser	0.057	Asn N	-4	1	0	-1	0	0	2														
Asp	0.053	Asp D	-5	0	0	-1	0	1	2	4													
Glu	0.053	Glu E	-5	0	0	-1	0	0	1	3	4												
Phe	0.045	Gln Q	-5	-1	-1	0	0	-1	1	2	2	4											
Asn	0.042	His H	-3	-1	-1	0	-1	-2	2	1	1	3	6										
Pro	0.041	Arg R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	8									
Ile	0.035	Lys K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								
His	0.034	Met M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							
Arg	0.034	Ile I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	-2	5						
Gln	0.032	Leu L	-8	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	8					
Tyr	0.030	Val V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				
Cys	0.025	Phe F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-5	-2	-4	-5	0	1	2	-1	9		
Met	0.012	Tyr Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		
Trp	0.012	Trp W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	
			C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp	

Matrice PAM - Propriétés

- la probabilité de mutation d'un aa est fixé à 1% (1-PAM)
- probabilités de mutation PAM-1 $\Rightarrow$ définition d'une unité de changement évolutif
- applications successives $\Rightarrow$ 2,3,4,... changements $\Rightarrow \neq$ modèles
- les opérations suivantes sont équivalentes :
  - appliquer n fois les probabilites de mutation PAM-1
  - appliquer 1 fois les probabilites de mutation PAM-1^n
  - modifier les élts de PAM-1 par une constante multiplicative λ si celle ci est proche de 1.
    ($M_{ij} = \lambda m_j A_{ij} / \Sigma_i A_{ij}$ et $M_{jj} = 1 - \lambda m_j$)
- PAM contient de l'information sur la composition :
  - PAM-0 : $M_{ij} = 0$ et $M_{ii} = 1$
  - PAM-∞ : approche asymptotique de la composition en aa

PAM vs. BLOSUM

	% $\neq$ observé	dist. évolutive PAM
BLOSUM-80	1	1
PAM-1	5	5
faible divergence	10	11
	15	17
	20	23
	25	30
	30	38
BLOSUM-62	34	47
PAM-120	40	56
	45	67
	50	80
	55	94
	60	112
	65	133
BLOSUM-45	70	159
PAM-250	75	195
forte divergence	80	246
	85	328

Recherche dans les banques de données

- on se donne :
  - une séquence requête q
  - une banque de séquences $T = \{t_1, ..., t_n\}$
- on veut :

  trouver des alignements significatifs entre q et les t_i
- les algorithmes classiques ne fonctionnent pas : prennent trop de temps, il faut trouver des parades

FASTA

Pearson et Lipman, 1988

- alignement global avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :
  1. trouve tous les mots identiques de longueur $\geq l$ communs à q et t_i
  2. sélectionne ceux de score suffisament élevé (score PAM par exemple)
  3. sélectionne une diagonale d (du dotplot) contenant le maximum de mots identiques de longueur $\geq l$
  4. procède à un alignement global "classique" dans une bande de largeur $2k$ autour de la diagonale d
- deux paramètres : k et l, l généralement de longueur 6 pour l'ADN et 2 pour les protéines

Blast

- naît en 1990 : trouve des matchs significatifs sans gaps
- évolution vers une version 2, avec gaps
  - NCBI-Blast
  - WU-Blast : très similaire à NCBI-Blast (mix entre Blast1 et FASTA pour la dernière étape)
- évolution vers des versions avec raffinement des résultats

Blast 1

- Trouve les mots **similaires** de taille fixe entre q et t_i.

  (taille par defaut : ADN $\rightarrow$ 11 , Protéines $\rightarrow$ 3)
- Ne considère que les couples de mots . . .

(Protéines) **similaires** $\rightarrow$ score des mots alignés $\geq$ seuil T

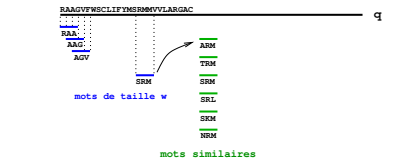
(à l'origine $T = 13$, actuellement $T = 11$ sur BLOSUM-62)

(ADN) **identiques** $\rightarrow$ pas de seuil T donc moins sensible.

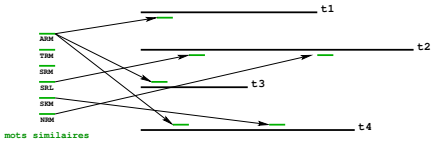
- Chaque couple de mots entre q et un t_i forme un *hit*
- Chaque hit est étendu à gauche et à droite : l'extension est stoppée lorsque le score du hit décroît de plus de X (*X-drop*)

Blast 1 - Schématiquement

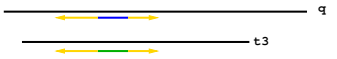
1 -



2 -



3 -

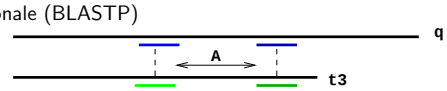


Blast 1

- un hit est un mot "commun" de taille fixée w (et de score supérieur à un seuil T dans le cas de BLAST-P) sur les deux séquences q et t_i
- chaque hit **étendu** (X-drop) forme un LMSP : *Localy Maximal scoring Segment Pair*
- ne conserve que les LMSP de score supérieur à un score seuil donné, les HSP : *High scoring Segment Pair*
- significativé évaluée (pour chaque t_i) sur le meilleur HSP trouvé nommé MSP : *Maximun scoring Segment Pair*

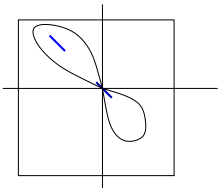
NCBI - Blast 2 (Gapped-blast)

idée 1 : se baser sur 2 hits distants au maximum de A sur la même diagonale (BLASTP)



$\rightarrow$ baisser le seuil de score de chaque hit $T = 13 \rightarrow 11$ pour conserver une bonne sensibilité

idée 2 : étendre les hits comme dans Blast 1 (X-drop) mais en **autorisant les gaps**



MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)
- à réserver à des requêtes du style : trouver la séquence dans la banque
- **évolution** : Discontiguous MegaBLAST
  - principe : utiliser une *graine espacée* plutôt qu'un *mot exact* (graine contiguë)
  - **exemple** : graine espacée 1001011001011001011101 plutôt que graine contiguë 11111111111
  - peut se réveler meilleur que BLAST (en particulier avec *graines espacées multiples*).

Définition des graines contiguës vs espacées

graine contiguë : 111111

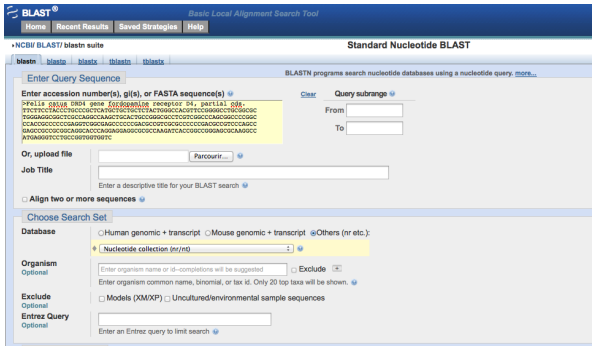
```
      111111
ATCAGTGCAAATGCGCAAGA
||||||:|||||||||.||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

```
       11101011
ATCAGTGCGAATGCGCAAGA
||||||:||:||||||.||||
ATCAGCGCAAATGCTCAAGA
```

- Les graines espacées peuvent être *bien choisies* pour mieux détecter les alignements (*Keich Li Ma Tromp DAM 2004*)
- Il est possible d'utiliser plusieurs graines espacées de formes différentes pour améliorer la sensibilité de la recherche

Exemple : *Basic BLAST → nucleotide blast → nr*



Exemple : Résultats

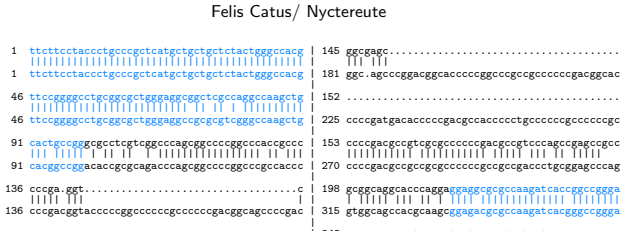
```
Query=  Felis catus DRD4 gene fordopamine receptor D4
        (276 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences
          1,174,453 sequences; 5,001,591,585 total letters


                                                      Score    E
Sequences producing significant alignments:          (bits) Value

gi|AB069665  Felis catus DRD4 gene f...               210   5e-52
gi|AB069662  Nyctereutes procyonoide...               157   7e-36
gi|AB069661  Canis lupus DRD4 gene f...               157   7e-36
gi|AB069666  Bos taurus DRD4 gene fo...               143   1e-31
gi|291947|   Homo sapiens Dopamine D4 recep...        135   2e-29
```


Exemple : Résultats

```
>gi|18143632|dbj|AB069662.1|AB069662 Nyctereutes procyonoides
 DRD4 gene  fordopamine receptor D4. Length = 393

 Score =  157 bits (79), Expect = 7e-36
 Identities = 94/99 (94%)
 Strand = Plus / Plus

Query 1 ttcttcctaccctgcccgctcatgctgctgctctactgggccacgttcc 48
        ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct 1 ttcttcctaccctgcccgctcatgctgctgctctactgggccacgttcc 48

Query 49 ggggcctgcgggcgctgggaggccggctcgccaggccaagctgcactgccgg 99
         |||||||||||||||||||| || || | ||||||||||||| |||||
Sbjct 49 ggggcctgcgggcgctgggaggccgcgcgtcgggccaagctgcacggccgg 99


 Score =  107 bits (54), Expect = 5e-21
 Identities = 60/62 (96%)
 Strand = Plus / Plus

Query 215 ggaggcgcgccaagatcaccggccgggagcgcaaggccatgagggtcct 252
          |||| ||||||||||||||| ||||||||||||||||||||||||||
Sbjct 332 ggagacgcgccaagatcacgggccgggagcgcaaggccatgagggtcct 379

Query 253 tgccggtggtggtc 276
          |||||||||||||
Sbjct 380 tgccggtggtggtc 393
```

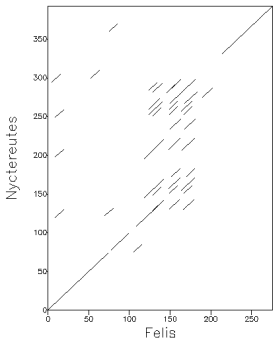

BLAST : ... vs Alignement Réel

Felis Catus/ Nyctereute



BLAST : ... vs Alignement Réel

Felis Catus/ Nyctereute



BLAST : Significativité des MSPs

- deux séquences peuvent toujours être alignées
- il existe toujours un (au moins) alignement de meilleur score S entre deux séquences (un MSP)

question : ce score est-il suffisamment élevé pour prouver une homologie ?

problème : peut-on trouver un MSP de meilleur score dans deux séquences aléatoires ?

BLAST : Mesures de significativité

- la **p-valeur (p-value)**

  mesure la *Probabilité* que 2 séquences aléatoires de même longueur et de même composition possèdent un MSP de score $\geq S$

- la **E-valeur (E-value)**

  mesure l'*Esperance* E du nombre n de MSPs de score $\geq S$ dans 2 séquences aléatoires de même longueur et de même composition

$$E = \sum_n p(n) \times n$$

Calcul de la E-value

- soient deux séquences a et b aléatoires suivant une distribution de probabilité connue
- on suppose que les MSPs sont données par les diagonales du dotplot
- plutôt que de décrire un alignement par des paires de lettres tirées aléatoirement, on peut le décrire par une suite de scores tirés aléatoirement
- on veut calculer l'esperance du nombre de MSPs de score $\geq$ à S

Calcul de la E-value

- Selon *Karlin et Altschul 1991* :

$$\textbf{E-value} = Kmne^{-\lambda S} \qquad \textbf{p-value} = 1 - e^{-\text{e-value}}$$

avec m la taille de la séquence requête, n la taille de la banque de données, S le score du hit (K et λ dépendent de la matrice de score, K peut être ajusté en fonction du coût des gaps)

Calcul du bit-score

- si S est le score d'un hit
- le bit-score (score normalisé) est :

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- l'expression de la E-value devient :

$$E(S) = mn2^{-S'}$$

Variation de la E-value

- si la taille de la séquence query augmente : la E-value . . . . . .
- si la taille de la banque est divisée par deux : la E-value . . . . . .
- si le score augmente : la E-value . . . . . .
- quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ? . . . . . .
- si on passe la E-value à 0.01, quel sera le bit-score ? . . . . . .

Variations de la E-value

```
Mus musculus chromosome 5, clone RP23-301L9, complete sequence
          Length = 212246

 Score = 32.2 bits (16), Expect = 2.1
 Identities = 19/20 (95%)
 Strand = Plus / Plus

Query: 2      ttcattatgaagcacgagga 21
               ||||||||||  |||||||||
Sbjct: 136843 ttcattatgatgcacgagga 136862


Mus musculus BAC clone RP23-13L19 from chromosome 9, complete sequence
          Length = 224108

 Score = 30.2 bits (15), Expect = 8.1
 Identities = 15/15 (100%)
 Strand = Plus / Plus

Query: 6      ttatgaagcacgagg 20
               |||||||||||||||
Sbjct: 93798 ttatgaagcacgagg 93812


Lambda     K      H
    1.37    0.711     1.31
Number of Hits to DB: 99,084,306
Number of Sequences: 2130505
Number of extensions: 2852
Number of successful extensions: 19
Number of sequences better than 10.0: 0
Number of HSP's better than 10.0 without gapping: 0
Number of HSP's successfully gapped in prelim test: 16
length of query: 21
length of database: 10,249,863,584
```


Variations de la E-value

```
Mus musculus chromosome 5, clone RP23-301L9, complete sequence
          Length = 212246

 Score = 32.2 bits (16), Expect = 2.1
 Identities = 19/20 (95%)
 Strand = Plus / Plus

Query: 2      ttcattatgaagcacgagga 21
               ||||||||||  |||||||||
Sbjct: 136843 ttcattatgatgcacgagga 136862

------------------------------------------------------------------

Mus musculus, clone RP23-277C24, complete sequence
          Length = 199946

 Score = 32.2 bits (16), Expect = 2.1
 Identities = 16/16 (100%)
 Strand = Plus / Minus


Query: 1      attcattatgaagcac 16
               ||||||||||||||||
Sbjct: 69080 attcattatgaagcac 69065
```


Variations de la E-value

```
Query length : 21
Mus musculus, clone RP23-277C24, complete sequence
          Length = 199946

 Score = 32.2 bits (16), Expect = 7.6
 Identities = 16/16 (100%)
 Strand = Plus / Minus


Query: 1      attcattatgaagcac 16
               ||||||||||||||||
Sbjct: 69080 attcattatgaagcac 69065

------------------------------------------------------------------

Query length : 20
Mus musculus, clone RP23-277C24, complete sequence
          Length = 199946

 Score = 32.2 bits (16), Expect = 5.1
 Identities = 16/16 (100%)
 Strand = Plus / Minus


Query: 1      attcattatgaagcac 16
               ||||||||||||||||
Sbjct: 69080 attcattatgaagcac 69065
```


Les différents programmes BLAST

Query \ Database	nucléique	protéique	nucléique traduit
nucléique	blastn	x	x
protéique	x	blastp	tblastn
nucléique traduit	x	blastx	tblastx

Le bon programme pour la bonne requête

extrait de "BLAST Program Selection Guide"

- MEGABLAST is the tool of choice to identify a nucleotide sequence
- Discontiguous MEGABLAST is better at finding nucleotide sequences similar, but not identical, to your nucleotide query
- les pages "Search for short nearly exact matches"
  - nucleotide : useful for primer or short nucleotide searches
  - proteins : optimized to find matches to a short peptide
  - principales différences :
    - taille de mots plus petite
    - suppression des filtres
    - relâchement de la E-value
    - matrice de score PAM30 (au lieu de BLOSUM62) pour les protéines

Evolutions de Blast : PSI-Blast

PSI-Blast is designed for more sensitive protein-protein similarity searches

Position Specific Iterated BLAST

1. recherche initiale avec BLASTp
2. construction d'un alignement multiple, puis d'un profil
   - à partir d'un alignement multiple des meilleurs hits
   - construit une matrice position-spécifique :
     - chaque colonne représente un AA
     - chaque ligne une position dans l'alignement
3. nouvelle recherche avec le profil et modification

réitère le processus un certain nombre de fois ou juqu'à convergence

Profil - exemple

```
                    # Pure Frequency Matrix
Alignement multiple:  # Columns are amino acid counts A->Z
                    # Rows are alignment positions 1->n
                    Simple
T-VAAPSVFIFPPSDEQ   Name            mymatrix
A-DAAPTVSIFPPSSEQ   Length          17
A-NAAPTVSIFPPSTZZ   Maximum score   60
D-PVAPTVLIFPPAADQ   Thresh          75
DPPIAPTVLLFPPSADQ   Consensus       APPAAPTVLIFPPSADQ
                    200200000000000000010000000
                    000000000000000100000000000
                    000100000000010200000100000
                    300000001000000000000100000
                    500000000000000000000000000
                    000000000000000500000000000
                    000000000000000000140000000
                    000000000000000000500000000
                    000001000002000000200000000
                    000000000400100000000000000
                    000005000000000000000000000
                    000000000000000500000000000
                    000000000000000500000000000
                    100000000000000000400000000
                    200100000000000000110000000
                    000220000000000000000000010
                    000000000000000040000000010
```


Evolution de Blast : PHI-Blast

PHI-BLAST can do a restricted protein pattern search

Pattern Hit Initiated BLAST

- pour les séquences protéiques
- entrée : une séquence et un motif (expression régulière à la Prosite)
- restriction de la banque aux séquences pour lesquelles le motif est retrouvé
- puis application de BLAST
- couplage possible avec PSI-Blast

Sensibilité et spécificité

On peut classer les résultats d'une méthode en 4 catégories :

VP les vrais positifs *(classé positif* **et bien** *positif)*

FP les faux positifs *(classé positif* **mais réellement** *négatif)*

ex. alignements parasites, prédiction pour une fonction que la séquence n'a pas en réalité...

FN les faux négatifs *(classé négatif* **mais réellement** *positif)*

ex. alignements perdus, prédiction négative pour une fonction que la séquence a en réalité...

VN les vrais négatifs *(classé négatif* **et bien** *négatif)*

$$\text{sensibilité} = \frac{VP}{VP+FN} \qquad \text{spécificité} = \frac{VN}{VN+FP}$$

- sensibilité : capacité de la méthode à ne pas "louper" de Positifs
- spécificité : capacité de la méthode à ne pas "ramener" de Négatifs