

H el ene TOUZET

touzet@lifl.fr

Table de hachage

- ▷ **k-mers** : mots de longueur k
- ▷ Fonction de hachage (pour l'ADN):

$$e : \{A, C, G, T\} \rightarrow \{0, 1, 2, 3\}$$

$$\mathcal{H}(i) = \sum_{j=0}^{k-1} e(a_{i+j})4^{k-j-1}$$

$$\mathcal{H}(i+1) = 4 \times \mathcal{H}(i) + e(a_{i+k}) \pmod{4^k}$$

- ▷ **Exemple** : 5-mers de *AGTACCGAA*

A G T A C C G A A

.....177.....

.... 709

..... 790

..... 88

.... 352

Application : BLAST

Basic Local Alignment Search Tool

Altschul et al. - 1997

- ▷ Programme pour la recherche de similarités dans de grandes banques de données

EMBL, Swissprot, ...

- ▷ Utilise un algorithme heuristique linéaire pour l'alignement local :
 - Ne s'intéresse qu'aux séquences avec un fort taux de similarité

BLAST ne construit pas un alignement avec toutes les séquences de la banque.

- Tire parti du caractère biologique des séquences

Des séquences similaires ont des segments communs de taille k quasi-identiques.

Par défaut : ADN $k=11$ ou 13 , protéines $k=3$

Étape 0:

Pré-traitement de la banque de données : indexation de tous les k -mers

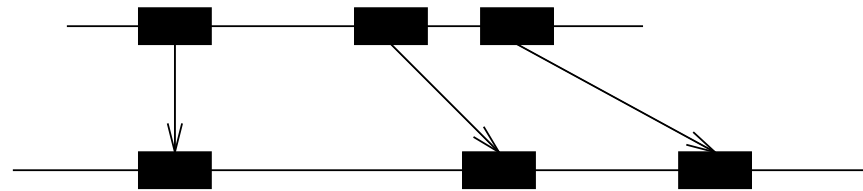
$4^{11} = 4\,194\,304 \ll$ taille de la banque de données

Étape 1:

Construction d'une table de hachage recensant tous les mots de longueurs k apparaissant dans la séquence requête avec un score $> T$.

Étape 2:

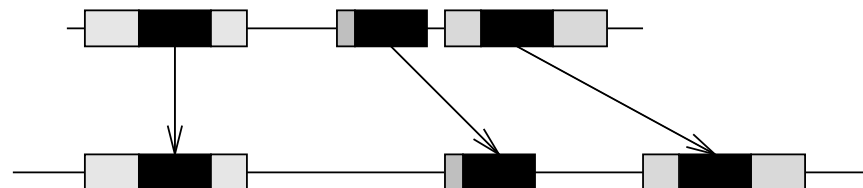
Localisation des k -mers dans la banque de données



→ HSP : *High Scoring Pairs*

Étape 3 :

Extension de ces points d'ancrage de proche en proche, pour avoir un score **significatif**.



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

Sequences producing significant alignments:		Score	E
		(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

ALIGNMENTS

>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 ggggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgg 99
|||||
Sbjct 49 ggggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgg 99

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

Query 215 ggaggcgcgccaagatcaccggccgggagcgcgaaggccatgagggtcct 252
||||
Sbjct 332 ggagacgcgccaagatcacgggcccgggagcgcgaaggccatgagggtcct 379

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

>gi|AB032908 Hylobates pileatus gene for dopamine receptor D4,
partial cds, drd4, 7-repeat allele. Length = 507

Score = 42.1 bits (21), Expect = 0.27

Identities = 45/53 (84%)

Strand = Plus / Plus

```
52  ggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgggcgcc 104
    ||||| ||||| ||||| ||| || | ||||| ||||| |||| |
4   ggcctgcagcgctgggagggtggcacgtcgcgccaagctgcacggccgcgcgcc 56
```

Alignement Felis Catus/ Nyctereute

```

1      ttcttcctaccctgcccgctcatgctgctgctctactgggccacgttccggggcctgcgg      60
      |||
1      ttcttcctaccctgcccgctcatgctgctgctctactgggccacgttccggggcctgcgg      60

61     cgctgggaggcggctcgccaggccaagctgcaactgccggggcgcctcgtcggcccagcggc    120
      |||
46     cgctgggaggccgcgcgctcggggccaagctgcacggccggacaccgcgacaccagcggc    120

121    cccggcccaccgcccccca.ggt.....c      144
      |||
121    cccggcccgccacccccgacgggtacccccggccccccgccccccgacggcagccccgac    180

145    ggcgagc.....      151
      |||
181    ggc.agcccggacggcacccccggcccgcgccccccgacggcacccccgatgacacccc    239

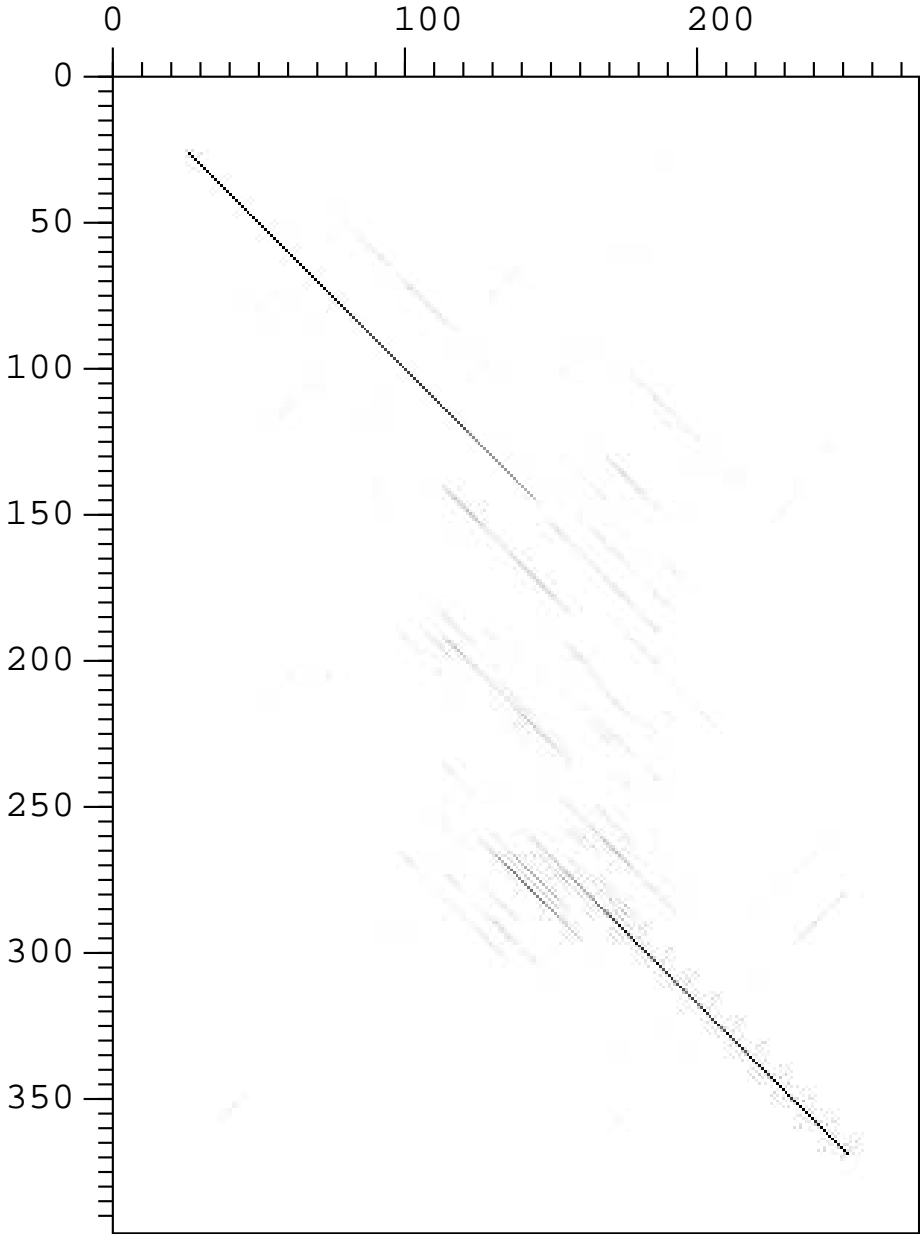
152    .....ccccgacgccgtcgcgccccccgacgccgt    192
      |||
240    cgacgccaccccctgccccccgccccccgccccccgacgccgccgccccccgcccga    299

193    cccagccgagccgcccggcaggcaccaggaggaggcgcgccaagatcacggccggga    242
      |||
300    ccctgcgagccccctggcagccacgcaagcggagacgcgccaagatcacgggcccggga    359

243    gcgcaaggccatgagggtcctgccggtggtggtc      276
      |||
360    gcgcaaggccatgagggtcctgccggtggtggtc      393

```

AB069665. (horizontal) vs. AB069662. (vertical)



Query= actgagcatagctgga (16 letters)

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

Sequences producing significant alignments:	Score (bits)	E Value
gi AC021642.14 Mus musculus chromosome 10 ...	32	1.0
gi AC079858.8 Homo sapiens BAC clone...	30	4.0
gi AC090032.2 Canis familiaris clone...	30	4.0
gi AF289076 Homo sapiens chromosome 8...	30	4.0
...		

ALIGNMENTS

>gi|AC021642.14| Mus musculus chromosome 10 clone.
Length = 203839

Query: 1 actgagcatagctgga 16
 |||||
Sbjct: 195114 actgagcatagctgga 195129

>gi|16973779|gb|AC079858.8| Homo sapiens BAC clone
Length = 82719

Query: 1 actgagcatagctgg 15
 |||||
Sbjct: 48150 actgagcatagctgg 48164

Query= actgagcatagctggac (17 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences

1,174,453 sequences; 5,001,591,585 total letters

Sequences producing significant alignments:	Score	E
	(bits)	Value
gi AC021642.14 Mus musculus chromosome ...	34	0.25
gi AL121894.26 Human DNA sequence fro...	32	1.0
gi AC079858.8 Homo sapiens BAC clone ...	30	4.0

ALIGNMENTS

>gi|AC021642.14| Mus musculus chromosome 10 clone

Query: 1 actgagcatagctggac 17

|||||

Sbjct: 195114 actgagcatagctggac 195130

>gi|9944239| Human DNA sequence

Query: 2 ctgagcatagctggac 17

|||||

Sbjct: 21064 ctgagcatagctggac 21079

>gi|AC079858.8| Homo sapiens BAC clone

Query: 1 actgagcatagctgg 15

|||||

Sbjct: 21064 ctgagcatagctggac 21079

Query= actgagcatagctggat (17 letters)

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

Sequences producing significant alignments:	Score (bits)	E Value
gi AC021642.14 Mus musculus chromosome 10 c...	32	1.0
gi AC079858.8 Homo sapiens BAC clone ...	30	4.0
gi AF07784 1 Sulfolobus solfataricus ...	30	4.0
...		

ALIGNMENTS

>gi|AC021642.14| Mus musculus chromosome 10 clone

```
Query: 1      actgagcatagctgga 16
          |||
Sbjct: 195114 actgagcatagctgga 195129
```

>gi|AC079858.8| Homo sapiens BAC clone

```
Query: 1      actgagcatagctgg 15
          |||
Sbjct: 48150  actgagcatagctgg 48164
```

Query= actgagcatag (11 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences

1,174,453 sequences; 5,001,591,585 total letters

No significant similarity found.

E-value \leq 1000

Query= actgagcatag (11 letters)

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

Sequences producing significant alignments:	Score (bits)	E Value
gi AC092378.3 Homo sapiens chromosome 16 clo...	22	967
gi NM_131197.1 Danio rerio endothelin recept...	22	967
gi AC084013.5 Homo sapiens BAC clone ...	22	967
gi AC092203.15 Mus musculus clone rp23-422n18...	22	967
gi AP000003 Pyrococcus horikoshii OT...	22	967
gi 10727456 Drosophila melanogaster ...	22	967
...		

ALIGNMENTS

>gi|AC092378.3| Homo sapiens chromosome 16 clone
Length = 199869

Query: 1 actgagcatag 11
|||
Sbjct: 78821 actgagcatag 78811

Génome de la drosophile

Query= actgagcatag (11 letters)

Database: D. melanogaster genomic nucleotide sequences
1170 sequences; 122,655,632 total letters

Sequences producing significant alignments:	Score (bits)	E Value
gi AE002770 Drosophila melanogaster g...	22	24
gi AE003609 Drosophila melanogaster g...	22	24
gi AE003450 Drosophila melanogaster g...	22	24
gi AE003426 Drosophila melanogaster g...	22	24
gi AE003484 Drosophila melanogaster g...	22	24
...		

ALIGNMENTS

>gi|7289299|gb|AE002770.1|AE002770 Drosophila melanogaster
genomic scaffold 142000013385552

Query: 1 actgagcatag 11
 |||||
Sbjct: 17834 actgagcatag 17844

Les différentes versions de BLAST

BLASTN : séquences nucléiques

BLASTP : séquences protéiques

BLASTX : une séquence nucléique comparée à une base de données protéique. (*Traduction suivant les 6 cadres de lecture.*)

TBLASTX : une séquence protéique comparée à une base de données nucléique

TBLASTN : une séquence nucléique comparée à une base de données nucléiques, chacune suivant tous les cadres de lecture. (*Cela revient à faire 36 fois BLASTP.*)

Psi-BLAST : itération de BLAST

Phi-BLAST : motifs

MEGABLAST : forte similarité

"Advanced BLAST"

Accès aux paramètres du programme

- ▷ Longueur des mots d'ancrage
- ▷ Seuil de la E-value
- ▷ Matrices de similarités
- ▷ Pénalités de gaps
- ▷ Code génétique
- ▷ Filtre des régions de faible complexité

Etc.

Les régions de faible complexité

Option "low complexity"

▷ Régions avec un fort biais de composition

```
PPCDPPPPKDKKKKDDGPP  
AAATAAAAAAAAAATAAAAT
```

Queue Poly-A, séquence Alu, répétitions CA, région riche en Proline . . .

▷ Peu de pertinence biologique

▷ Par défaut, BLAST masque ces régions de faible complexité

N → nucléotides

X → acides aminés

Psi-BLAST (Position Specific Iterated BLAST)

1. Recherche initiale avec BLAST
2. Construction d'un alignement multiple, puis d'un **profil** à partir des séquences trouvées

L'alignement utilise la séquence requête comme modèle

- 3 Nouvelle recherche sur la banque de données à l'aide du profil

L'algorithme est une adaptation de BLAST. Le critère de significativité est la E-value

Retour à **2.** jusqu'à ce qu'il y ait convergence, ou après un nombre déterminé d'itérations

Profil ?

Exemple : hormone pancréatique (PP)

NEUY_CARAU/29-64	AEE..LAKYYSALRHYINLITRQRY
PYY_HUMAN/29-64	PEE..LNRYYASLRHYLNLVTRQRY
PMY_PETMA/1-36	PEE..LSKYMLAVRNYINLITRQRY
PPY_LOPAM/1-36	PED..WASYQAAVRHYVNLITRQRY
PAHO_BOVIN/30-65	PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61	VED..LIRFYNDLQQYLNVVTRHRY
PAHO_ANSAN/1-36	VED..LRFYYDNLQQYRLNVFRHRY
NPF_HELAS/4-39	PNE..LRQYLKELNEYA AIMGRTRF
NPF_MONEX/1-39	DNKAALRDYLRQINEYFAIIGRPRF

Score de substitutions : PSSM (Position Specific Score Matrix)

Columns are amino acid counts A->Z, Rows are alignment positions 1->n

```
1 0 0 1 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 2 0 0 0 0 0
0 0 0 0 7 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 3 4 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 7 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
3 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 3 1 0 0 0 0 0 0 0 0 0
0 0 0 1 0 1 0 0 0 0 2 0 0 0 0 0 2 2 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0
1 0 0 0 0 0 0 0 0 0 0 2 1 0 0 0 1 0 0 0 0 0 0 0 4 0 0 0
3 0 0 1 0 0 0 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0
3 0 0 1 2 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 6 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 2 5 0 0 0 0 0 0 0 0 0
0 0 0 0 2 0 0 3 0 0 0 0 0 1 0 0 2 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 0
0 0 0 0 0 1 0 0 3 0 0 2 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0
2 0 0 0 0 0 0 0 0 0 0 1 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 2 0 0 4 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 4 0 0 1 1 0 0 0 0 0 0 0 0 3 0 0 0 0 0
0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 2 4 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0
```

Gaps :
valeur par défaut
de BLAST

PHI-Blast (Pattern-Hit Initiated BLAST)

Requête :

- une séquence protéique
- un motif modélisé par une expression régulière

Exemple : C2H2 pour le doigt de zinc : $C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H$

Problème : Quelles sont les séquences contenant le motif et présentant une similarité locale autour du motif ?

Algorithme en deux temps :

1. sélection des séquences contenant le motif
2. application de BLAST à ces séquences

Exemple : la séquence

```
>sp|093602|ATF2_CHICK  
MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFGPARNDSVIVADQTPTPTRFLKN  
CEEVGLFNELASPFENEFFKASEDDIKKMPLDLSPLATPIIRNKIEEPSVVETTHQDSPL  
PHPESTTNDKEVSLQQTAQPTSTIVRPASLQVPNVLLTSSDSSVIIQQAIPSPTSSTVI  
TQAPSSNRPIVPVPGPFLLLHLPNGQTMPVAIPASITNSNVHVPAAVPLVRPVTMVPSI  
PGIPGPSSPQPVQSEAKLRLKAALTQQHPQVTNGDTAKGHPSGLVRTQSEPRPQSLQQP  
ATSTTETPASPAQPTQQTPNTGRRRRRAANEDPDEKRRKFLERNRAAASRCRQKRKVVWVQ  
SLEKKAEDLSSLNGQLQNEVTLLRNEVAQLKQLLLAHKDCPVTAMQKKSGYHTADKDDSS  
EDISVPSSPHTEAIQHSSVSTSNQVSSTSKAEAVATSVLTQLADQSSEPGLPQVGVVPPS  
QAQPSGS
```

avec un motif **C2H2** (position 9 sur la séquence)

PHI-BLAST permet de filtrer les hits.

```
>gi|18401576|ref|NP_566582.1| bZIP family transcription factor Length = 135
```

```
Score = 49.7 bits (117), Expect = 9e-05
```

```
Identities = 27/76 (35%), Positives = 47/76 (61%)
```

```
320 NTGGRRRRAANEDPDEKRRKFLERNRAAASRCRQKRKVVVQSLEKKAEDLSSLNGQLQNE  
    +T RRR      D + + K L RNR +A + R+++KV+V LE +A +L + N QL+ +  
50  STAKRRRGRNPVDKEYRSLKRLLRNRVSAQQARERKKVYVSDLESRANELQNNNDQLEEK
```

```
380 VTLLRNEVAQLKQLLL 395  
    ++ L NE   L+++L+  
110 ISTLTNENTMLRKMLI 125
```

Trouvé par BLAST, mais pas par PHI-BLAST (et ne contenant pas de doigt de zinc)

PHI-BLAST permet de récupérer des hits plus fins.

>NP_694513.1 metal-regulatory transcription factor 1 Length = 593

Score = 23.6 bits (69), Expect = 0.002

Identities=70/295 (23%), Positives=114/295 (38%), Gaps=52/295 (17%)

```
4   DKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKF-----GPARNDSVIVADQT--PTP
      *****
      +KPF C + GC + F+++ L H H+ F N S+ ++D + T
297 EKPFPCPSDGCEKTFSSQYSLKSHIRGHDKGPSFTVSGHPLSEDANHSLCLSDLSLISTD

55  TRFLKNCEEVGLFNELASP-----FENEFKKASEDDIK-----KMPLDLSPLATPI
      + +N GL +P F++ SEDD K L+ SP + P
357 SELQENHNSQGLDLNSVTPIRIFELMFQSPENSVSEDDPKPTESLAESFGLEPSPQSAP-

101 IRNKIEEPSVETTHQDSPLPHPESTTNDKEVSLQQTAAQPTSTIVRPASLQVPNVLLTS
      +TH P P P + ++ S+ AQ T P + Q P ++S
416 -----ADASTHPAFPQPPSTCSS---SCSITAPAQDAQT--PPTTQQAPPPAVSS

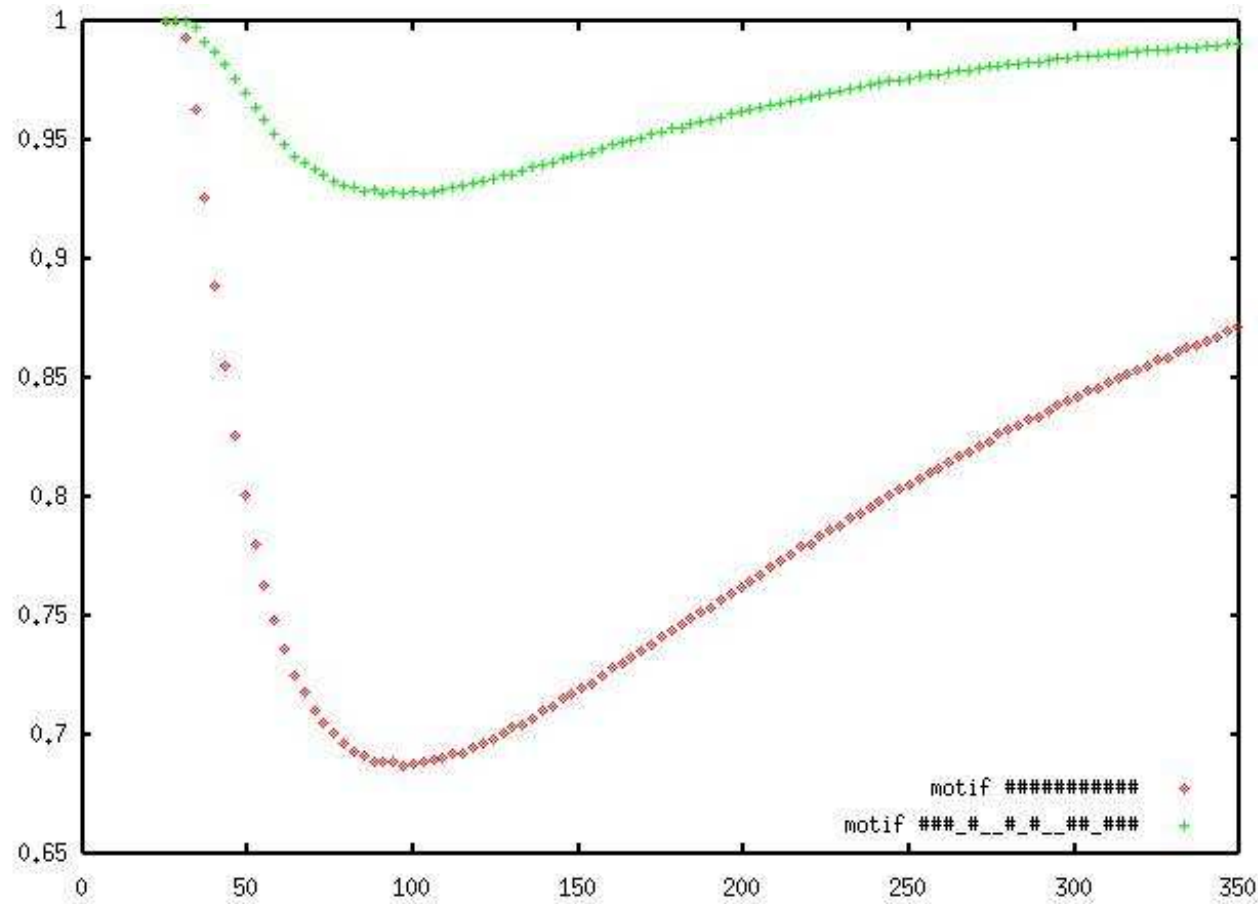
161 SDSSVIIQQAIPSPSTSVIT--QAPSSNRPIVPVPGPFLLLHLPNGQTMPVAIPASIT
      S + A PS + ++ APS+ + + + P+ ++ ++PA
462 SSQTSSFPSAPPSSSQPAEVSSPSAPSATQHYYMAQS-----VSSPSAASVS-SVPAGTA

219 NSNVHVPAAVPLVRPVTMVPISIPGIPGPSSPQPQVQSEAKLR---LKAALTQQHPQ 270
      V VPL P T+ SI G P V S+ L+ AA QQ+P+
516 EVTAAVTHTVPLAAPPTI--SIAPTLG-LQPSLVMSDQNLQWILSSAASAQQNPE 567
```

Trouvé par PHI-BLAST, mais pas par BLAST (et contenant bien un doigt de zinc)

PatternHunter

Utiliser des “graines espacées” au lieu de motifs contigus



Tous les alignements de score 25, identité 1, mismatch -3. Source: L. Noé, G. Kucherov - LORIA