

Méthodes de filtrage

Hélène Touzet

Équipe Bioinfo — LIFL — USTL

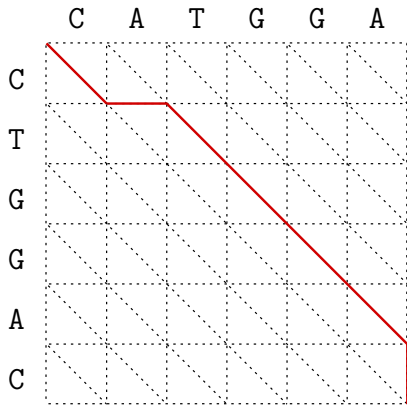
Master recherche informatique

www.lifl.fr/~touzet/masterrecherche.html

Le rêve d'un biologiste

- Avec les algorithmes vus au cours précédent, je sais qu'on est capable de comparer deux séquences, et de regarder si elles ont une fonction commune.
- Que faire si on ne dispose que d'une séquence ?
 - ▶ La comparer avec toutes les autres séquences existantes sur terre, disponibles dans les banques de données
 - ▶ Besoin d'algorithmes d'alignements locaux extrêmement efficaces

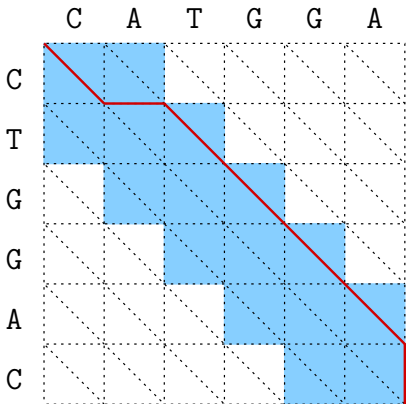
Rappel : alignement global par programmation dynamique



C	A	T	G	G	A	-
C	-	T	G	G	A	C

- Cas général : $O(n^2)$

Rappel : alignement global par programmation dynamique



C	A	T	G	G	A	-
C	-	T	G	G	A	C

- Cas général : $O(n^2)$

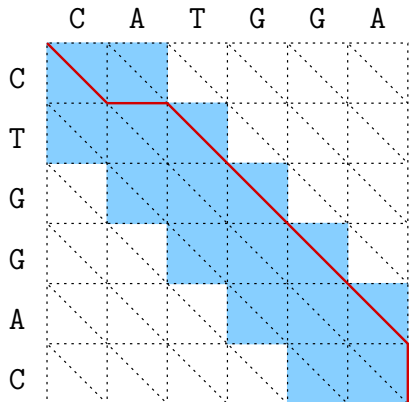
- Séquences similaires :

(au plus k indels)

k -strip = $\{(x, y); |x - y| \leq k\}$

Complexité : $O(kn)$

Rappel : alignement global par programmation dynamique



C	A	T	G	G	A	-
C	-	T	G	G	A	C

– Cas général : $O(n^2)$

– Séquences similaires :
(au plus k indels)

k -strip = $\{(x, y); |x - y| \leq k\}$

Complexité : $O(kn)$

Cette optimisation ne marche pas pour l'alignement local

Filtrage: accélération de l'alignement local



Ne pas construire d'alignement avec toutes les séquences de la banque de données, mais seulement avec un petit nombre de séquences candidates qui peuvent être sélectionnées de manière très efficace

- ▶ Filtre : couple de positions entre la requête et la banque de données susceptible de produire une similarité
- ▶ Filtrage avec perte: pas de garantie de trouver l'intégralité des alignements pertinents
- ▶ 3 exemples : BLAST, BLAT et les heuristiques à base de graines espacées

Comment évaluer la qualité d'un filtre ?

- ▶ Temps de calcul
- ▶ Qualité du résultat
 - ▶ Vrais positifs (VP) : bons alignements détectés
 - ▶ Faux positifs (FP) : mauvais alignements retenus par le filtre
 - ▶ Vrais négatifs (VN) : mauvais alignements effectivement non-détectés
 - ▶ Faux négatifs (FN) : bons alignements non détectés

$$\frac{VP}{VP+FN}$$

Sensibilité

$$\frac{FP}{VP+FP}$$

Sélectivité

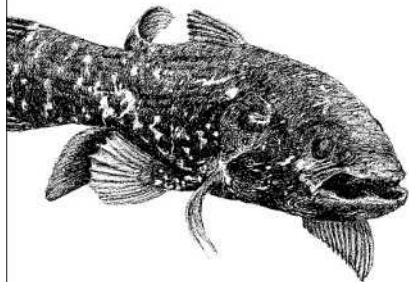
BLAST

Basic Local Alignment Search Tool - Altschul *et al.* - 1997

- ▶ Recherche de similarités dans de grandes banques de données
Genbank, EMBL, Swissprot, ...
- ▶ Utilise un algorithme heuristique linéaire pour l'alignement local
- ▶ **Filtre** : mot exact commun entre la séquence requête et la banque
- ▶ *k*-mer : mot de longueur *k*
Par défaut : *ADN* $k=11$ ou 13 , *protéines* $k=3$

An Essential Guide to the Basic Local Alignment Search Tool

BLAST



O'REILLY®

Jan Korf, Mark Yandell & Joseph Bedell

Plus de 100 000 citations
scientifiques

Plus de 150 000 requêtes
quotidiennes (NCBI)

Étape 1: pré-traitement de la banque de données

- ▶ Indexation de tous les k -mers
Pour $k = 11$, $4^{11} = 4\,194\,304 \ll$ taille de la banque de données
- ▶ Stockage dans une table de hachage (sans collision)
- ▶ Fonction de hachage (pour l'ADN):

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

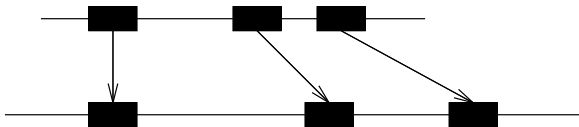
$$\mathcal{H}(i) = \sum_{j=0}^{k-1} e(a_{i+j})4^{k-j-1}$$

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

- ▶ **Exemple** : 5-mers de AGTACCGAA

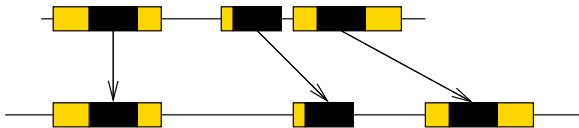
```
A G T A C C G A A
.....177.....
.....709.....
.....790.....
.....88.....
.....352.....
```

- ▶ **Étape 2:** Filtrage - localisation des k -mers de la séquence requête 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 (bits)	E 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 ttcttctaccctgcccgctcatgctgctgctctactgggccacgttc 48
|
Sbjct 1 ttcttctaccctgcccgctcatgctgctgctctactgggccacgttc 48

Query 49 ggggcctcggcgctgggaggcgctcgcaggccaagctgcaactgccgg 99
|
Sbjct 49 ggggcctcggcgctgggaggcgctcgcaggccaagctgcaactgccgg 99

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

Query 215 ggaggcgcgccaagatcacggccgggagcgcaaggccatgagggtcct 252
|
Sbjct 332 ggagacgcgccaagatcacggccgggagcgcaaggccatgagggtcct 379

Query 253 tgccggtggtggtc 276
|
Sbjct 380 tgccggtggtggtc 393

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.*)

BLAT : accélération de BLAST

BLAST-Like Alignment Tool - JW Kent - 2002

- ▶ Privilégie la rapidité, au détriment de la sensibilité, en réduisant la taille de l'index
- ▶ Index : 11-mers non chevauchants
Tient en mémoire RAM : moins de 1 Gb
- ▶ Adapté à la recherche de séquence avec plus de 95% d'identité et de longueur supérieure à 40 bases
Application: localiser un gène dans un génome

Filtres à base de graines espacées



Utiliser des graines avec des espaces - des trous -
à la place des graines contigues de BLAST

Exemple: les alignements de longueur 8 avec au plus une substitution

ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG
ACTGACTG	TCTGACTG	AATGACTG	ACGGACTG	ACTAACTG

ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG
ACTGTCTG	ACTGAATG	ACTGACCG	ACTGACTA

Graine contigue de poids 5: #####

Filtres à base de graines espacées



Utiliser des graines avec des espaces - des trous -
à la place des graines contigues de BLAST

Exemple: les alignements de longueur 8 avec au plus une substitution

ACTGACTG
|||||||
ACTGACTG
détecté

ACTGACTG
|||||||
TCTGACTG
détecté

ACTGACTG
| |||||
AATGACTG
détecté

ACTGACTG
|| |||||
ACGGACTG
détecté

ACTGACTG
||| ||||
ACTAACTG
non détecté

ACTGACTG
|||| |||
ACTGTCTG
non détecté

ACTGACTG
||||| ||
ACTGAATG
détecté

ACTGACTG
|||||| |
ACTGACCG
détecté

ACTGACTG
|||||||
ACTGACTA
détecté

Graine contigue de poids 5: #####

Sensibilité : 7/9

Filtres à base de graines espacées



Utiliser des graines avec des espaces - des trous - à la place des graines contigues de BLAST

Exemple: les alignements de longueur 8 avec au plus une substitution

ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG
ACTGACTG	TCTGACTG	AATGACTG	ACGGACTG	ACTAACTG

ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG
ACTGTCTG	ACTGAATG	ACTGACCG	ACTGACTA

Graine espacée de poids 5: ##_ ###

Filtres à base de graines espacées



Utiliser des graines avec des espaces - des trous - à la place des graines contigues de BLAST

Exemple: les alignements de longueur 8 avec au plus une substitution

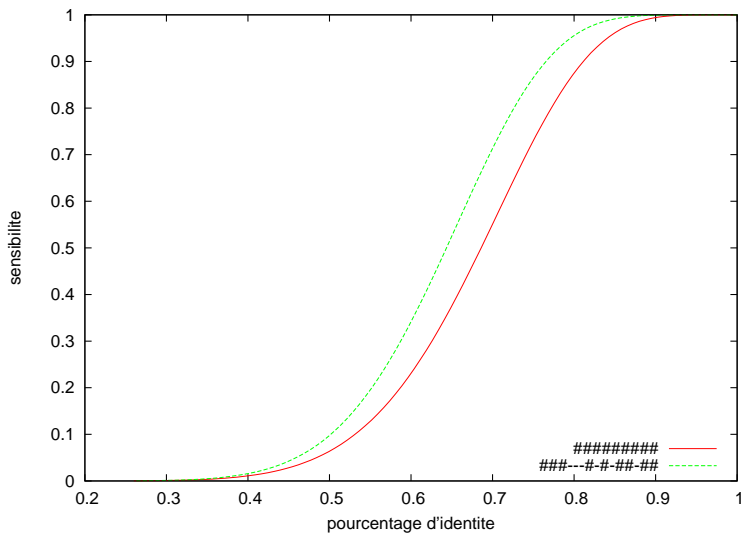
ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG
ACTGACTG	TCTGACTG	AATGACTG	ACGGACTG	ACTAACTG
déTECTÉ	déTECTÉ	déTECTÉ	déTECTÉ	déTECTÉ

ACTGACTG	ACTGACTG	ACTGACTG	ACTGACTG
ACTGTCTG	ACTGAATG	ACTGACCG	ACTGACTA
déTECTÉ	non déTECTÉ	déTECTÉ	déTECTÉ

Graine espacée de poids 5: ##_ ###

Sensibilité : 8/9

Sensibilité: graine espacée vs graine contigue



Comment calculer la sensibilité d'une graine ?

Données

- ▶ une graine espacée π
 - ▶ w , son poids (nombre de #)
 - ▶ m , son étendue (nombre de # et nombre de -)

- ▶ un modèle pour les alignements
 - ▶ p , le pourcentage d'identité ($0 \leq p \leq 1$)
 - ▶ n , la longueur

Question

Quelle est la proportion des alignements qui sont détectés par la graine π ?

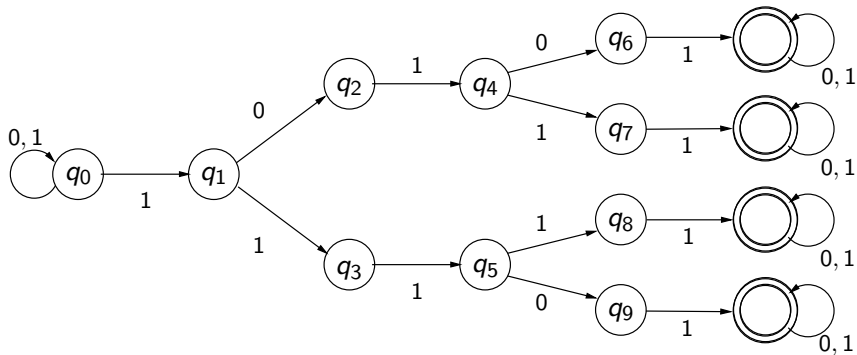
Algorithme de Buhler (2003)

- ▶ Les alignements sont modélisés par des mots sur l'alphabet $\{0, 1\}$
 - ▶ 1 \rightarrow match, identité
 - ▶ 0 \rightarrow mismatch, substitution

ACTGACTG	ACTGACTG
X	X
TCTGACTG	AATGACTG
01111111	10111111

- ▶ $\mathcal{L}(\pi)$: ensemble des mots $\{0, 1\}^*$ détectés par la graine π
- ▶ **Etape 1** : construction d'un automate fini déterministe qui reconnaît $\mathcal{L}(\pi)$
- ▶ **Etape 2**: calcul à partir de l'automate de la probabilité qu'un mot de longueur n appartienne à $\mathcal{L}(\pi)$

Exemple: graine $\pi = \#-\#-\#$ (Laurent Noé)



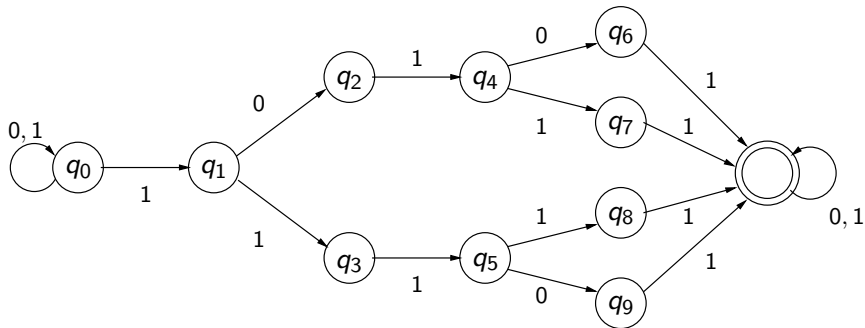
Automate non-déterministe pour $\mathcal{L}(\pi)$

Longueur: s

Nombre d'états acceptants: 2^{s-w}

Nombre d'états $\leq (w + 2)2^{s-w}$

Exemple: graine $\pi = \#-\#-\#$ (Laurent Noé)



Automate non-déterministe pour $\mathcal{L}(\pi)$

On peut fusionner les états correspondant à des $\#$ en partant de la droite

Construction efficace de l'automate déterministe

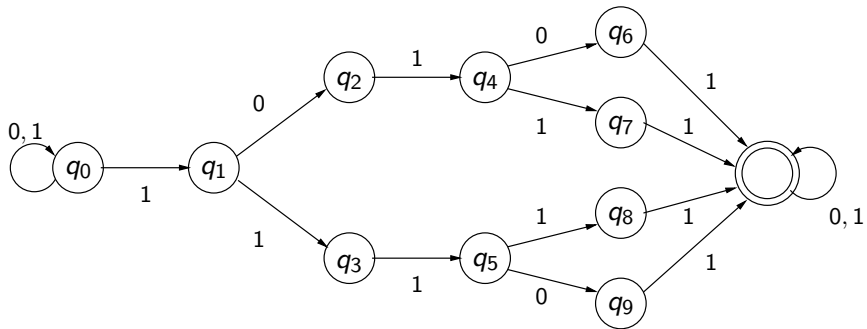
- ▶ Les états sont les mêmes que pour l'automate non-déterministe
- ▶ Modification de l'état initial q_0
- ▶ Ajout de transitions étiquetées par 0 à partir des états non-complets (qui correspondent à des $\#$)

Création d'un **tableau de bord**:

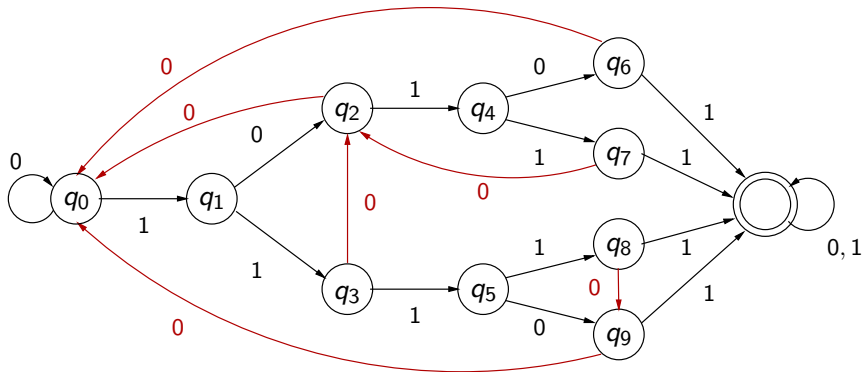
- ▶ soit q un état non-complet de l'automate
- ▶ soit u le mot menant de q_0 à q
- ▶ soit w le plus long suffixe propre de $u0$ qui soit également un préfixe d'un mot de $\mathcal{L}(\pi)$.
- ▶ soit q' l'état atteint à partir de q_0 en lisant w

$$\text{Bord}(q) = q'$$

Suite de l'exemple: graine #_#_#



Suite de l'exemple: graine #-#-#



état	q_2	q_3	q_6	q_7	q_8	q_9
Bord	q_0	q_2	q_0	q_2	q_9	q_0

Calcul de la sensibilité

- ▶ α : alignement, mot sur $\{0, 1\}^*$
- ▶ p : probabilité de lire 1 à une position donnée de α
- ▶ $\Phi_x(q)$: ensemble des états étiquetés par la lettre x qui mènent à l'état q
- ▶ $P(i, q, x)$: probabilité d'atteindre l'état q avec le préfixe $\alpha[1..i]$ en ayant lu x en dernière lettre

$$P(i, q, 0) = (1 - p) \sum_{q' \in \Phi_0(q)} P(i - 1, q', 0) + P(i - 1, q', 1)$$

$$P(i, q, 1) = p \sum_{q' \in \Phi_1(q)} P(i - 1, q', 0) + P(i - 1, q', 1)$$

- ▶ Mise en œuvre : programmation dynamique
- ▶ Complexité : $O(nw2^{s-w})$

Calcul de la sensibilité

- ▶ α : alignement, mot sur $\{0, 1\}^*$
- ▶ p : probabilité de lire 1 à une position donnée de α
- ▶ $\Phi_x(q)$: ensemble des états étiquetés par la lettre x qui mènent à l'état q
- ▶ $P(i, q, x)$: probabilité d'atteindre l'état q avec le préfixe $\alpha[1..i]$ en ayant lu x en dernière lettre

$$P(i, q, 0) = (1 - p) \sum_{q' \in \Phi_0(q)} P(i - 1, q', 0) + P(i - 1, q', 1)$$

$$P(i, q, 1) = p \sum_{q' \in \Phi_1(q)} P(i - 1, q', 0) + P(i - 1, q', 1)$$

- ▶ Mise en œuvre : programmation dynamique
- ▶ Complexité : $O(nw2^{s-w})$

Le problème de la sensibilité d'une graine espacée est NP-complet (Li-Ma 2004)

Performances comparées

Comparaison exhaustive entre le génome humain (3 milliards de bases) et le génome de la souris (3 milliards de bases)

Temps CPU estimés pour un Pentium III 700MH, 1GB :

- ▶ Alignement local exact (Smith&Waterman) : 100 siècles
- ▶ BLAST : 19 années
- ▶ Graines espacées (PatternHunter) : 20 jours

Source : Ming Li - CPM 2005

Comment interpréter les résultats

- ▶ Problème commun à BLAST, BLAT, aux heuristiques à base de graines espacées
- ▶ Mesure de significativité (rareté) de l'alignement : E-valeur
 $E(S,n,m)$: Nombre moyen d'alignements ayant un score supérieur ou égal à S quand on cherche dans une banque de taille m avec une séquence requête de longueur n .
- ▶ Décrit le bruit aléatoire qui existe lorsque on aligne des séquences
- ▶ Croit de manière proportionnelle en fonction de n et de m
- ▶ Décroit de manière exponentielle en fonction du score S
- ▶ Plus la E-valeur est proche de 0, plus la similarité est significative

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

		Score	E
Sequences producing significant alignments:		(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 ≤ 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

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