La gestion de données dans le cadre d'une application de recherche d'alignement de séquence : BLAST.

Gaël Le Mahec

L’algorithme BLAST.

- **B**asic **L**ocal **A**lignment **S**earch **T**ool est un algorithme de recherche d’homologies entre une séquence et une banque de séquences.

- Il emploie une méthode heuristique limitant la recherche à des séquences plus susceptibles de présenter un bon alignement avec la séquence requête.
  - L’algorithme est basé sur l’hypothèse que deux séquences présentant de fortes similarités doivent contenir des sous-séquences identiques ou très proches.
  - Il repère ces sous-séquences et les étend tant que possible.
  - Il attribue un “score” de similarité qui déterminera si la séquence doit être retenue.

Exemple.

BLAST trouve une région de similarité avec un alignement de taille 10 (paramètre) qu'il va ensuite étendre.

```
A C A G T A A T T T C A G T A C T T
| |   -   -   | |   -   -   | | | | | | | | | -   -
A C T T C A A C C T C A G T A C C G

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

A C G A A A G G T A C T A G G A T C
| | | | | -   | |   -   | | | | | | | | -
A C G A A T G G T C C T A G G A T T
```

Sensibilité.

Les paramètres, notamment la taille de l'alignement de départ, influent largement sur la sensibilité (capacité à ne pas rater de bon alignement).
Ainsi, une valeur trop grande, si elle accélère grandement le temps d'éxécution, risque de rater de nombreux alignements.
Par exemple :

```
A C G A A A G G T A C T A G G A T C
|   |   |   |   |   -   |   |   |   -   |   |   |   |   |   |   |   -
A C G A A T G G T C C T A G G A T T
```

On a 88% de similitude mais pas d'alignement exact de taille 10, l'alignement est perdu.

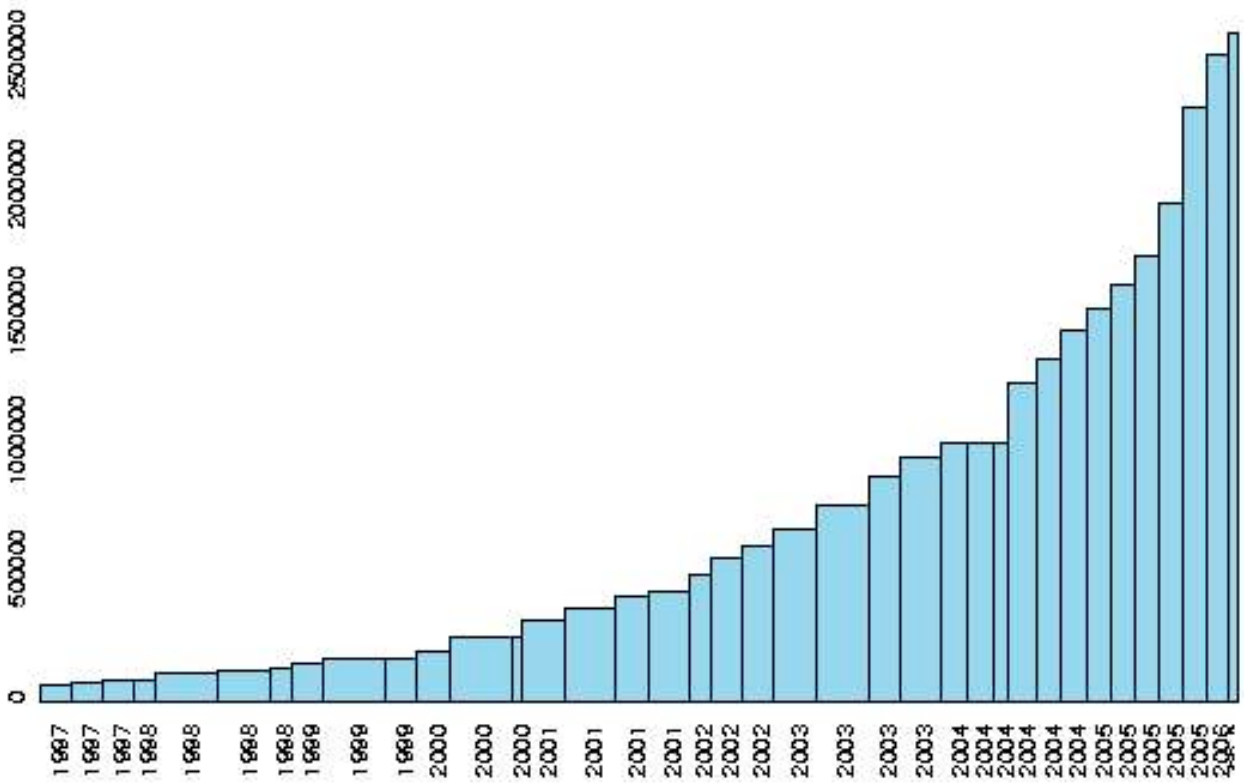
Les bases de données biologiques.

Parmi les bases les plus utilisées pour la recherche d'alignement on notera les suivantes :

- NCBI (**N**ational **C**enter for **B**iotechnology **I**nformation).
  - La base GenBank d'environ 32,500,000 séquences ($\sim$ 37,900,000,000 nucléotides). Une nouvelle version disponible environ tous les 2 mois.
  - La base de protéines orthologues COGs. $\sim$ 5000 protéines et 67 génomes complets.
- EMBL (The **E**uropean **M**olecular **B**iology **L**aboratory)
  - La base TrEMBL (**Tr**anslated **EMBL**) d'environ 2,6 millions de séquences, soit environ 840 millions d'acides aminés. Mise à jour en moyenne tous les deux mois.
- ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem)
  - La base Swiss-Prot d'environ 200,000 séquences soit environ 75,000,000 acides aminés. Mise à jour hebdomadaire et une nouvelle version tous les 6 mois environ.

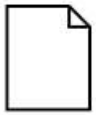
Croissance des bases.

La croissance actuelle de la taille de ces bases est exponentielle et ne semble pas devoir s'arrêter dans un avenir proche. Voici par exemple la croissance de la base TrEMBL depuis 1997 :



Exécution de BLAST.

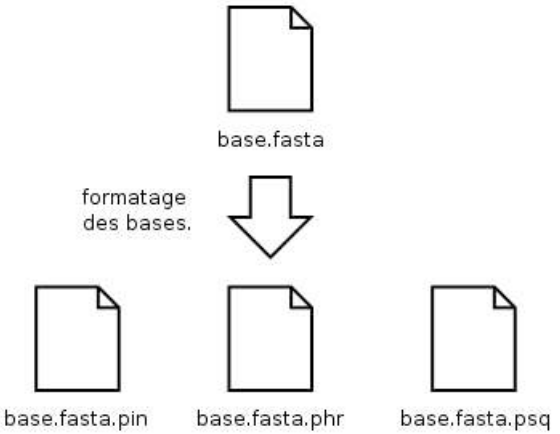
Les bases de données biologiques sont des bases à faible structuration, il s'agit bien souvent de simples fichiers plats et parfois de fichiers dans une arborescence de répertoires. Il faut formater ces bases avant leur utilisation par BLAST. Ainsi, le déroulement normal dans une exécution de BLAST est le suivant :



base.fasta

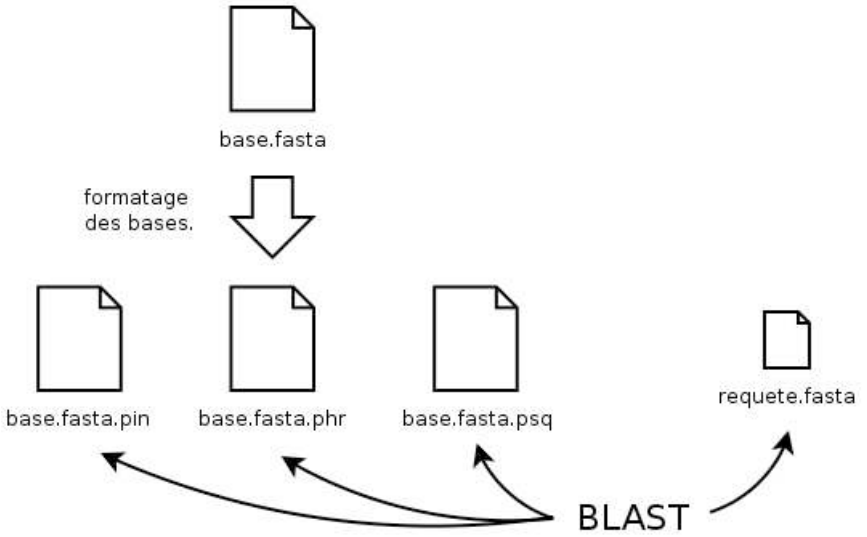
Exécution de BLAST.

Les bases de données biologiques sont des bases à faible structuration, il s'agit bien souvent de simples fichiers plats et parfois de fichiers dans une arborescence de répertoires. Il faut formater ces bases avant leur utilisation par BLAST. Ainsi, le déroulement normal dans une exécution de BLAST est le suivant :



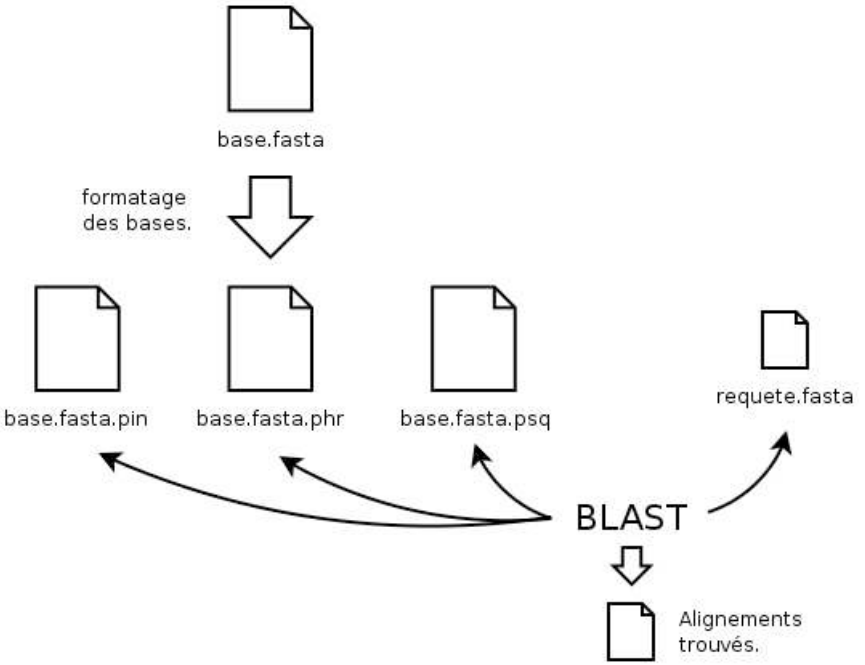
Exécution de BLAST.

Les bases de données biologiques sont des bases à faible structuration, il s'agit bien souvent de simples fichiers plats et parfois de fichiers dans une arborescence de répertoires. Il faut formater ces bases avant leur utilisation par BLAST. Ainsi, le déroulement normal dans une exécution de BLAST est le suivant :



Exécution de BLAST.

Les bases de données biologiques sont des bases à faible structuration, il s'agit bien souvent de simples fichiers plats et parfois de fichiers dans une arborescence de répertoires. Il faut formater ces bases avant leur utilisation par BLAST. Ainsi, le déroulement normal dans une exécution de BLAST est le suivant :



Mise à jour.

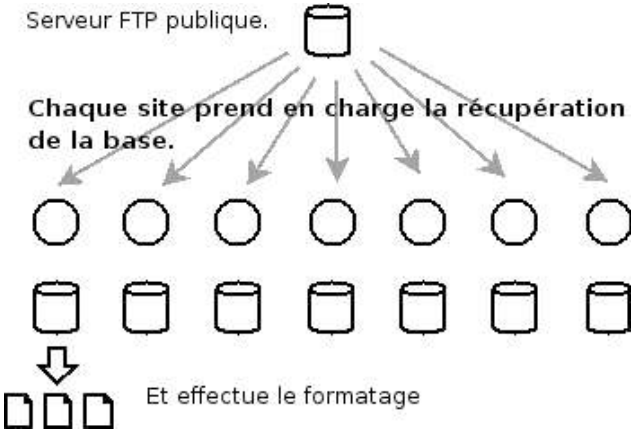
Le formatage des bases nécessite d' établir la manière dont on effectuera leurs mises à jour.

- Mise à jour parallèle de chacun des replicas. Chaque site récupère une nouvelle version et effectue le formatage.

Mise à jour.

Le formatage des bases nécessite d' établir la manière dont on effectuera leurs mises à jour.

- Mise à jour parallèle de chacun des replicas. Chaque site récupère une nouvelle version et effectue le formatage.



Mise à jour.

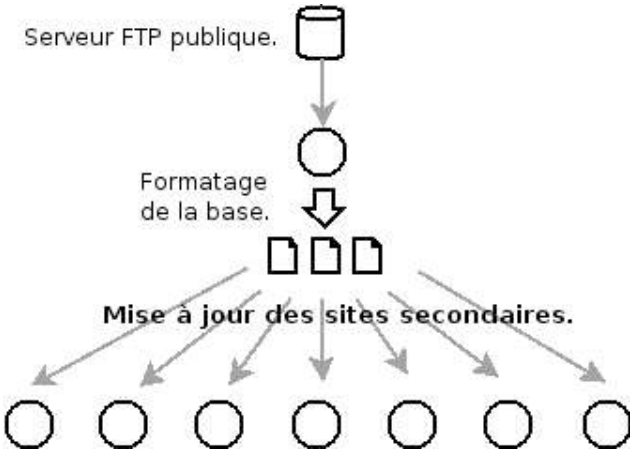
Le formatage des bases nécessite d' établir la manière dont on effectuera leurs mises à jour.

- Mise à jour parallèle de chacun des replicas. Chaque site récupère une nouvelle version et effectue le formatage.
- La base est récupérée sur un site qui effectue le formatage puis chaque replica est mis à jour depuis cette version de référence.

Mise à jour.

Le formatage des bases nécessite d' établir la manière dont on effectuera leurs mises à jour.

- Mise à jour parallèle de chacun des replicas. Chaque site récupère une nouvelle version et effectue le formatage.
- La base est récupérée sur un site qui effectue le formatage puis chaque replica est mis à jour depuis cette version de référence.



Mise à jour.

Le formatage des bases nécessite d' établir la manière dont on effectuera leurs mises à jour.

- Mise à jour parallèle de chacun des replicas. Chaque site récupère une nouvelle version et effectue le formatage.
- La base est récupérée sur un site qui effectue le formatage puis chaque replica est mis à jour depuis cette version de référence.

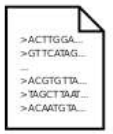
Suivant la taille de la base et les possibilités de connexions des sites, les mises à jour des bases peuvent s'avérer être une source de bloquage pour l'utilisation de BLAST sur la grille. En effet, une cohérence stricte des données est exigée pour obtenir des résultats significatifs. Les données n'étant plus utilisables tant que dure l'opération, il convient de s'assurer que celle-ci prendra le moins de temps possible.

Répartition des requêtes.

Une requête peut consister en la recherche d'alignement de plusieurs dizaines de milliers de séquences contre une seule base. On peut ainsi répartir facilement la charge sur autant de noeuds disposant de celle-ci.
La durée d'éxécution de telles requêtes peut rendre envisageable de placer temporairement la base sur les noeuds disponibles.

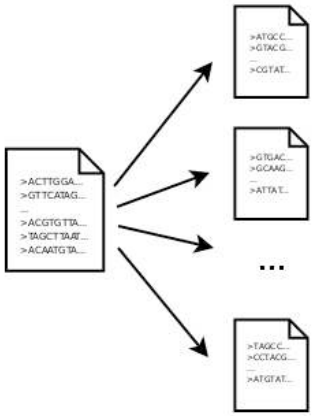
Répartition des requêtes.

Une requête peut consister en la recherche d'alignement de plusieurs dizaines de milliers de séquences contre une seule base. On peut ainsi répartir facilement la charge sur autant de noeuds disposant de celle-ci.
La durée d'éxécution de telles requêtes peut rendre envisageable de placer temporairement la base sur les noeuds disponibles.



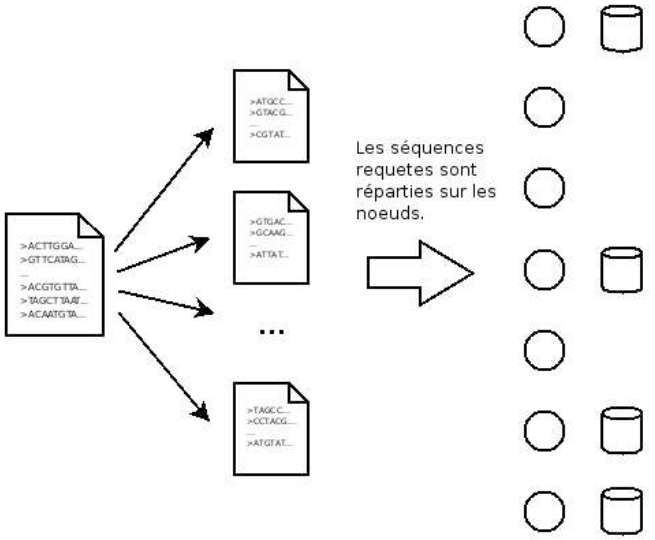
Répartition des requêtes.

Une requête peut consister en la recherche d'alignement de plusieurs dizaines de milliers de séquences contre une seule base. On peut ainsi répartir facilement la charge sur autant de noeuds disposant de celle-ci.
La durée d'éxécution de telles requêtes peut rendre envisageable de placer temporairement la base sur les noeuds disponibles.



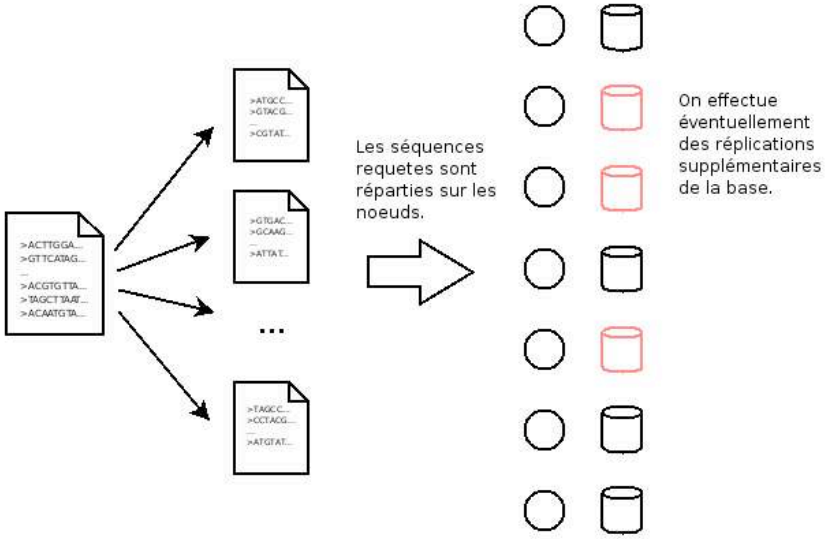
Répartition des requêtes.

Une requête peut consister en la recherche d'alignement de plusieurs dizaines de milliers de séquences contre une seule base. On peut ainsi répartir facilement la charge sur autant de noeuds disposant de celle-ci.
La durée d'éxécution de telles requêtes peut rendre envisageable de placer temporairement la base sur les noeuds disponibles.



Répartition des requêtes.

Une requête peut consister en la recherche d'alignement de plusieurs dizaines de milliers de séquences contre une seule base. On peut ainsi répartir facilement la charge sur autant de noeuds disposant de celle-ci.
La durée d'éxécution de telles requêtes peut rendre envisageable de placer temporairement la base sur les noeuds disponibles.



Utilisation des résultats.

Il est courant qu'on effectue une première utilisation de BLAST comme filtrage des séquences à utiliser. Les résultats obtenus sont alors utilisés comme paramètre d'entrée d'autres requêtes.
Suivant le contexte, il peut être utile de conserver ces données sur les noeuds de calcul en vue de leur réutilisation ultèrieure.

Conclusion

BLAST est un outil extrêmement populaire chez les biologistes et nécessite un temps de calcul qui grandit en même temps que la taille des bases sur lesquelles il s'applique. L'utilisation de la grille comme support d'exécution est un bon moyen d'accélérer les recherches d'alignements mais pose des problèmes de gestion de données non triviaux.

- Mettre à jour les bases très régulièrement tout en assurant la cohérence des données.
- Répartir les bases sur les noeuds afin de permettre une utilisation optimale de celles-ci. Cette répartition doit pouvoir être effectuée dynamiquement.
- Conserver et/ou transmettre les résultats en vue d'une réutilisation.

- L'algorithme BLAST.
- *Exemple.*
- *Sensibilité.*
- Les bases de données biologiques.
- *Croissance des bases.*
- Exécution de BLAST.
- Mise à jour.
- Répartition des requêtes.
- Utilisation des résultats.
- Conclusion