Université d'Alexandrie

# Initiation à la Bioinformatique

```
>gb|ACN22993.1|  cytochrome oxidase subunit I [Metacarcinus edwardsii]
 gb|ACN22994.1|  cytochrome oxidase subunit I [Metacarcinus edwardsii]
 gb|ACN22995.1|  cytochrome oxidase subunit I [Metacarcinus edwardsii]
  ▶7 more sequence titles
 Length=185

 Score =  360 bits (923),  Expect = 4e-98, Method: Compositional matrix adjust.
 Identities = 185/185 (100%), Positives = 185/185 (100%), Gaps = 0/185 (0%)

Query  1    TSLSLIIRAELGQPGTLISNDQIYNVVVTAHAFVMIFFMVMPIMIGGFGNWLVPLMLGAP  60
            TSLSLIIRAELGQPGTLISNDQIYNVVVTAHAFVMIFFMVMPIMIGGFGNWLVPLMLGAP
Sbjct  1    TSLSLIIRAELGQPGTLISNDQIYNVVVTAHAFVMIFFMVMPIMIGGFGNWLVPLMLGAP  60

Query  61   DMAFPRMNNMSFWLLPPSLTLLLMSGMVESGVGTGWTVYPPLAGAIAHAGASVDMGIFSL  120
            DMAFPRMNNMSFWLLPPSLTLLLMSGMVESGVGTGWTVYPPLAGAIAHAGASVDMGIFSL
Sbjct  61   DMAFPRMNNMSFWLLPPSLTLLLMSGMVESGVGTGWTVYPPLAGAIAHAGASVDMGIFSL  120

Query  121  HLAGVSSILGAVNFMTTVINMRSFGMTLDQMPLFVWAVFITAILLLLSLPVLAGAITMLL  180
            HLAGVSSILGAVNFMTTVINMRSFGMTLDQMPLFVWAVFITAILLLLSLPVLAGAITMLL
Sbjct  121  HLAGVSSILGAVNFMTTVINMRSFGMTLDQMPLFVWAVFITAILLLLSLPVLAGAITMLL  180

Query  181  TDRNL  185
            TDRNL
Sbjct  181  TDRNL  185
```

## Basic Local Alignment Search Tool
## BLAST

Auteur(s): Mohamed GAD

2009/2010

Auteur(s) :

Mohamed GAD

Professeur à l'institut des études supérieures et de la recherche

El Shatby, Alexandrie, EGYPTE

Mèl : esmailgadmoh@yahoo.fr

2009/2010

## Introduction

BLAST est l'abréviation de « Basic Local Alignment Search Tool » ou, en français, L'outil de recherche basique d'alignement local. Il ressemble à google dans le fonctionnement (figure 1). On utilise google pour chercher les bases de données d'internet sur les informations d'un mot clé (cancer, par exemple). On cherche les bases de données d'internet sur des sujets qui ressemblent ou qui contiennent le mot clé. BLAST, quand à lui, cherche les bases de données des protéines et ADNs pour des séquences (sujets) qui ressemblent à notre séquence (requête) utilisée comme mot clé.


**Figure 1:** Le lien fonctionnel entre google et BLAST

## Comment ça marche

Admettons que nous comparons l'être humain avec les bases de données (figure 2)

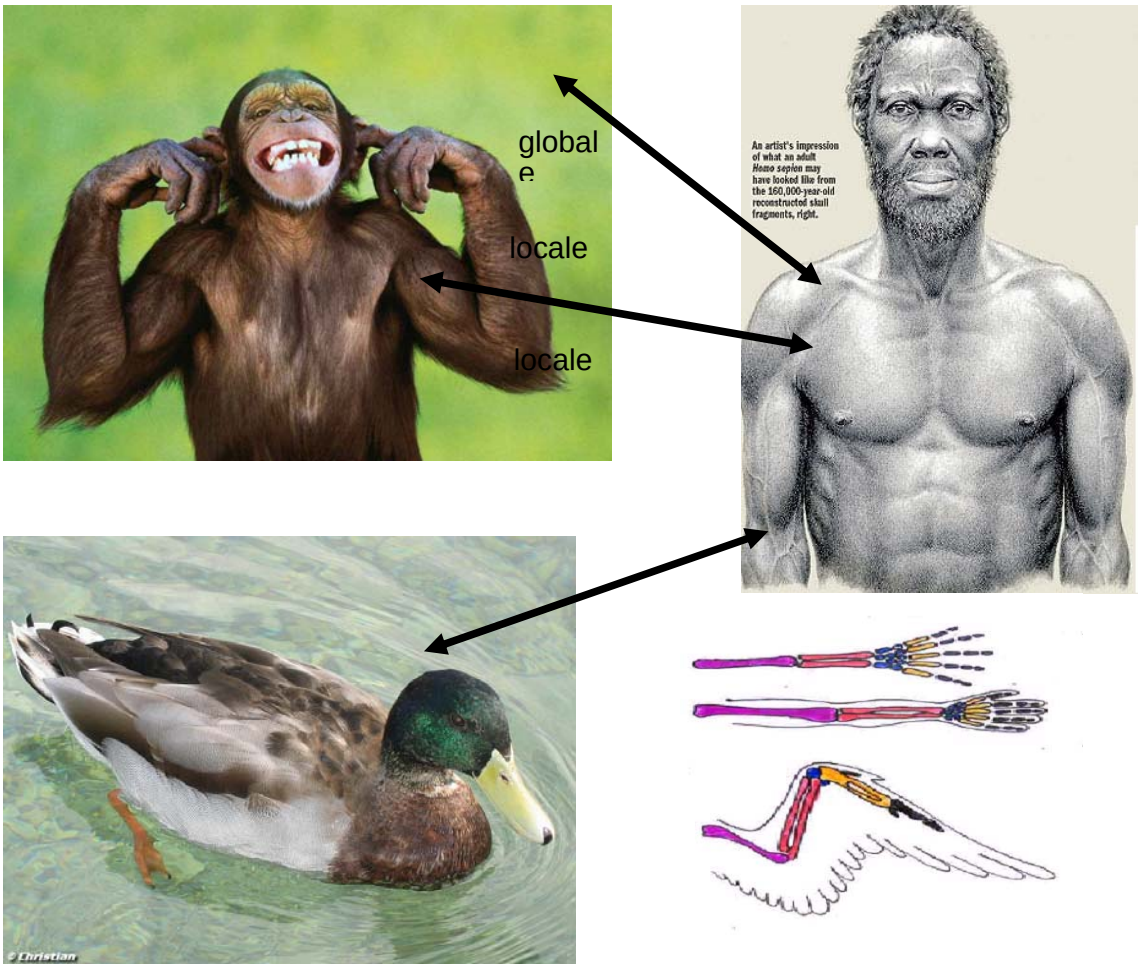


**Figure 2:** Le principe de BLAST

Globalement et morphologiquement on va trouver que le champanzé est l'animal qui ressemble plus à l'être humain. En revanche, on ne peut pas faire la même chose avec un canard, car globalement il n'y a pas de ressemblance entre un humain et un canard. Par contre, on peut comparer localement (une partie) : l'aile d'un canard avec le bras d'un humaine. Là, on va trouver une ressemblance entre l'homme et le canard. En ce qui concerne BLAST, il utilise l'alignement local pour comparer les séquences. Il divise la séquence en

question « requête » en morceaux composées de trois acides aminés (en cas des protéines) ou 11 nucléotides (en cas d'ADNs). Ces morceaux sont nommés mots. En cherchant les bases de données de séquences avec ces mots on trouvera plusieurs mots (mots voisins) qui ressemble à ceux de la requête (figure 3). Les mots voisins appartiennent à un ou plusieurs séquences sujets.

| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 12 | | | | | | | | | | | | | | | | | | | | C |
| S | 0 | 2 | | | | | | | | | | | | | | | | | | | S |
| T | −2 | 1 | 3 | | | | | | | | | | | | | | | | | | T |
| P | −3 | 1 | 0 | 6 | | | | | | | | | | | | | | | | | P |
| A | −2 | 1 | 1 | 1 | 2 | | | | | | | | | | | | | | | | A |
| G | −3 | 1 | 0 | −1 | 1 | 5 | | | | | | | | | | | | | | | G |
| N | −4 | 1 | 0 | −1 | 0 | 0 | 2 | | | | | | | | | | | | | | N |
| D | −5 | 0 | **0** | −1 | 0 | 1 | 2 | 4 | | | | | | | | | | | | | D |
| E | −5 | **0** | **0** | −1 | 0 | 0 | 1 | 3 | 4 | | | | | | | | | | | | E |
| Q | −5 | −1 | −1 | **0** | 0 | −1 | 1 | 2 | 2 | 4 | | | | | | | | | | | Q |
| H | −3 | −1 | −1 | 0 | −1 | −2 | 2 | 1 | 1 | 3 | 6 | | | | | | | | | | H |
| R | −4 | 0 | −1 | 0 | −2 | −3 | 0 | −1 | −1 | 1 | 2 | 6 | | | | | | | | | R |
| K | −5 | 0 | 0 | −1 | −1 | −2 | 1 | 0 | 0 | 1 | 0 | 3 | 5 | | | | | | | | K |
| M | −5 | −2 | −1 | −2 | −1 | −3 | −2 | −3 | −2 | −1 | −2 | 0 | 0 | 6 | | | | | | | M |
| I | −2 | −1 | 0 | −2 | −1 | −3 | −2 | −2 | −2 | −2 | −2 | −2 | −2 | 2 | 5 | | | | | | I |
| L | −6 | −3 | −2 | −3 | −2 | −4 | −3 | −4 | −3 | −2 | −2 | −3 | −3 | 4 | 2 | 6 | | | | | L |
| V | −2 | −1 | 0 | −1 | 0 | −1 | −2 | −2 | −2 | −2 | −2 | −2 | −2 | 2 | 4 | 2 | 4 | | | | V |
| F | −4 | −3 | −3 | −5 | −4 | −5 | −4 | −6 | −5 | −5 | −2 | −4 | −5 | 0 | 1 | 2 | −1 | 9 | | | F |
| Y | 0 | −3 | −3 | −5 | −3 | −5 | −2 | −4 | −4 | −4 | 0 | −4 | −4 | −2 | −1 | −1 | −2 | 7 | 10 | | Y |
| W | −8 | −2 | −5 | −6 | −6 | −7 | −4 | −7 | −7 | −5 | −3 | 2 | −3 | −4 | −5 | −2 | −6 | 0 | 0 | 17 | W |
| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |

**Figure 3:** Un exemple d'une matrice de score

Maintenant, BLAST va établir un seuil pour diminuer le nombre de mots obtenus. Le seuil est établi par la matrice de score (scoring matrix, figure 3)

**Matrices de score « substituion »**

C'est une méthode algorithmique qui calcule la probabilité qu'une lettre dans un mot requête (acide aminé ou nucléotide) est remplacé par un autre dans un mot sujet. Ces probabilités se trouvent dans un tableau (matrice). Il y a plusieurs matrices de scores (PAM, BLOSUM)

BLAST utilise ces matrices de score pour déterminer un score de seuil. Ensuite, un score global va être calculé pour tous les mots retenus de la même séquence sujet et BLAST va afficher les séquences avec les scores les plus élevés (High Scoring Segment Pair, HSP). Entre les deux séquences (figure 4) on trouvera une ligne qui montre les lettres (acides aminés ou nucléotides) identiques dans les deux séquences, les lettres qui ont été remplacées par d'autre (positives) et les régions dans les séquences qui n'étaient pas pris en considération (vide). Les positives, ce sont les différences entre les deux séquences que le BLAST considère comme acceptable basé sur la nature de la lettre et la matrice de score utilisée. C'est

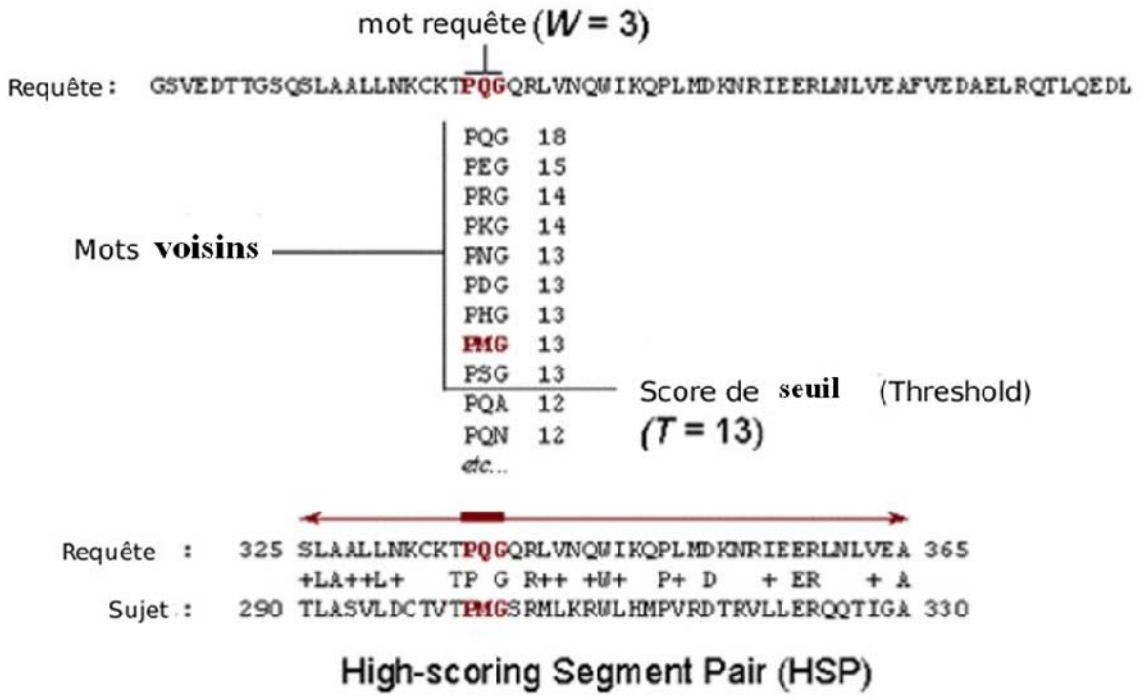


**Figure 4:** Le principe d'action de BLAST

Comme le cas du remplacement d'une leucine (L) par une isoleucine (I). Les régions vides présentent les parties de séquences qui n'étaient pas prises dans l'alignement local.

**BLAST jargon**

Query: séquence requête
subject: séquence sujet de la banque de donnée
HSP: High Scoring Pairs
E-Value: Expect value (fréquence d'occurrence par hasard)
nr: non redundant databases
Score: valuer calculée par BLAST pour déterminer le degré de ressemblance entre deux séquences. Le plus le score augmente le plus les deux séquences se ressemblent
Gap: saut, une ou plusieurs distance ajoutées dans la séquence requête et/ou sujet pour optimiser l'alignement

E-value: on l'appelle aussi « Expect » c'est la probabilité que les deux séquences se ressemblent par hasard. Le moins la valuer E est le plus la ressemblance des deux séquences. Fréquence d'occurrence aléatoire

Le BLAST se trouve sur le site de NCBI à l'adresse suivante:
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome

Figure 5 montres la page d'accueil de BLAST. Elle est composée de trois sections :

1. La section « BLAST genome »; qui contient les séquences intégrales de quelques génomes. En utilisant cette section on compare un ADN avec un des génomes de cette section

2. La section « Basic BLAST » qui contient les outils pour comparer une molécule (ADN ou protéine) avec les bases de données

    - Des protéines
        - Protein BLAST (BLASTp); cherche dans les bases de données de protéines en utilisant une protéine requête
        - BLASTx ; cherche dans les bases de données de protéines en utilisant la traduction d'un ADN
    - Des ADNs
        - nucleotides BLAST (BLASTn); cherche dans les bases de données de ADN en utilisant une ADN requête
        - tBLASTn; cherche dans la base de données de l'ADN traduit en utilisant une protéine requête
        - tBLASTx; cherche dans la base de donnée de l'ADN traduit en utilisant une ADN requête traduite

3. La section « Specialized BLAST » qui contient des outils qui comparent une molécule avec les bases de données pour :
    - Primer BLAST ; Fabriquer des amorces
    - Trace archives ; chercher l'archive de BLAST pour des activités faites précédemment sur le site
    - Conserved domain ; chercher dans la base de données des domaines conservées
    - SNPs (single nucleotide polymorphism); chercher dans la base de données du polymorphisme des nucléotides
    - Align ; comparer de séquences d'ADNs ou de protéines

**Pourquoi utilisons-nous BLAST?**

1. Pour connaître la fonction d'un gène « BLASTn ». En comparant ce gène avec les molécules des bases de données nous pouvons déterminer à quelle famille ce gène appartient-il.
2. Pour connaître la fonction d'une protéine « BLASTp » en le comparant avec les bases de données.
3. Pour connaître l'emplacement et les effets de SNPs « Align »sur une protéine en comparant la protéine mutée avec la normale.
4. Pour sélectionner des patrons « templates » qui, ensuite, vont être utilisés pour prédire la structure 2D et 3D des protéines « psi-BLAST »
5. Pour fabriquer des amorces qui vont être utilisés pour le PCR « Primer BLAST »
6. Pour déterminer dans une séquence les régions susceptibles de participer dans des site actifs d'une protéine

**BLASTp : comment ça marche**

La page (figure 6) est divisé en deux sections; basique et avancée « algorithm parameters ».
La section basique est divisée en trois parties:

1. Enter Query Sequence.
2. Choose Search Set
3. Program Selection

Dans cette section on soumet la séquence, on sélectionne les bases de données et le programme BLAST. En général on sélectionne BLASTp

La section avancée est divisé en trois parties:

1. General Parameters
2. Scoring Parameters
3. Filters and Masking

Cette section est pour les personnes expérimentées. En général on sélectionne les régions de basse complexité

Le « complexity region » pour éliminer les régions de la séquences qui influenceraient le résultat de BLAST comme les séquences répétées eucaryotiques « repeats ». Ces régions seront remplacées par des X ou des N et elles n'interviendront plus dans le score

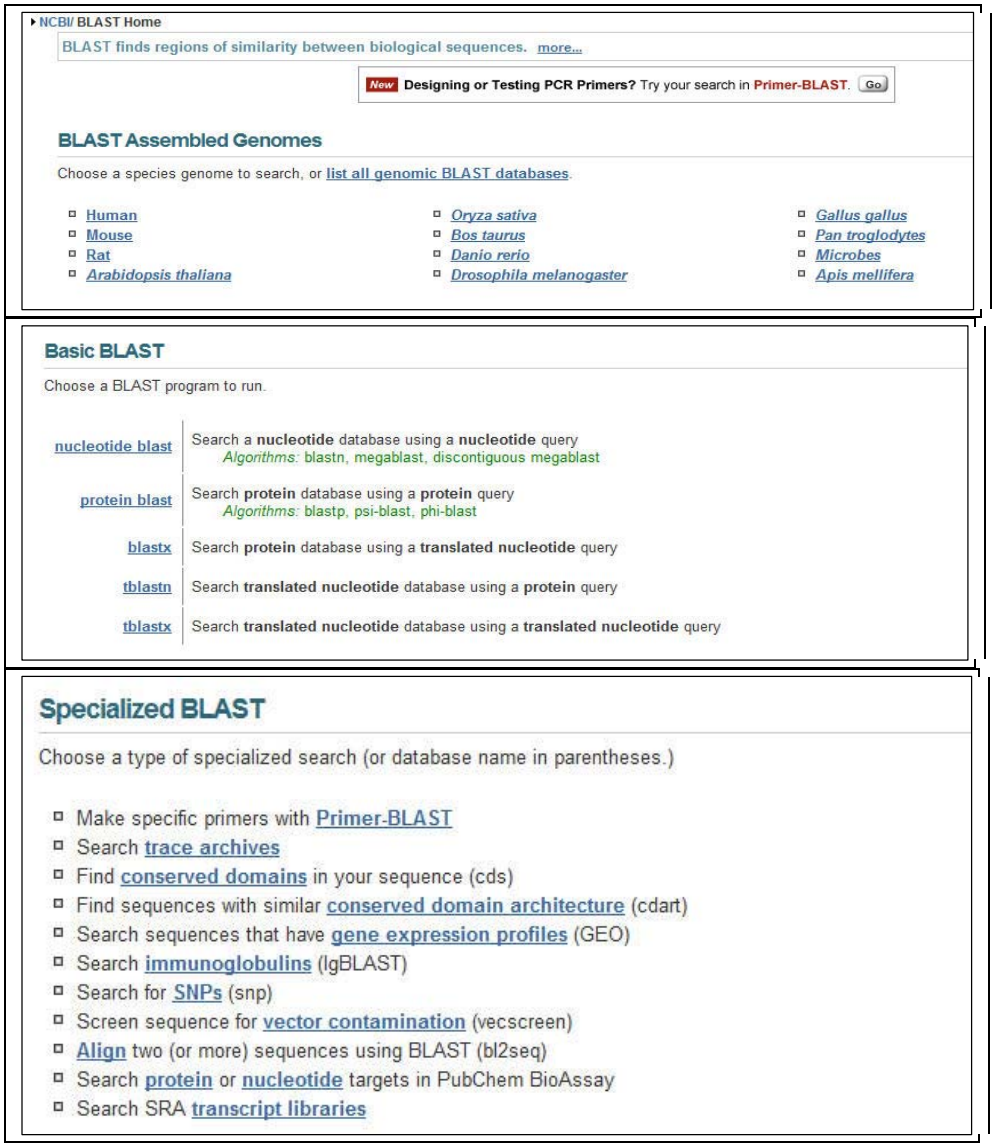


**Figure 5:** Page d'accueil de BLAST.

**Comparons-nous les protéines ou les ADNs ?**

La meilleure façon de détecter des similitudes entre séquences est généralement la comparaison au niveau protéique.

1. Il existe 20 acides aminés contre 4 nucléotides. La probabilité de trouver une "lettre" donnée par hasard est donc plus importante avec les nucléotides.
2. Plusieurs codons produisent le même acide aminé (tableau 1). 134 / 549 substitutions de bases sont synonymes. Les séquences protéiques sont plus informatives.
3. Les outils de comparaison est plus puissants pour les acides aminées dû à l'utilisation des propriétés physicochimiques ou des substitutions observées dans l'évolution.

Par conséquent, lorsque les acides aminés sont différents, on est capable de retrouver des similitudes. On en est tout à fait incapable au niveau des nucléotides.

Il faut prendre en compte que les comparaisons avec les séquences protéiques ne permettent de détecter que les régions codantes. Évidemment, on utilisera toujours la séquence ADN ou ARN pour analyser ce qui n'est pas traduit.

**Tableau 1:** Les codes des acides aminés

| | T | C | A | G |
|---|---|---|---|---|
| T | TTT Phe (F)<br>TTC Phe (F)<br>TTA Leu (L)<br>TTG Leu (L) | TCT Ser (S)<br>TCC Ser (S)<br>TCA Ser (S)<br>TCG Ser (S) | TAT Tyr (Y)<br>TAC Tyr (Y)<br>TAA Stop<br>TAG Stop | TGT Cys (C)<br>TGC Cys (C)<br>TGA Stop<br>TGG Trp (W) |
| C | CTT Leu (L)<br>CTC Leu (L)<br>CTA Leu (L)<br>CTG Leu (L) | CCT Pro (P)<br>CCC Pro (P)<br>CCA Pro (P)<br>CCG Pro (P) | CAT His (H)<br>CAC His (H)<br>CAA Gln (Q)<br>CAG Gln (Q) | CGT Arg (R)<br>CGC Arg (R)<br>CGA Arg (R)<br>CGG Arg (R) |
| A | ATT Ile (I)<br>ATC Ile (I)<br>ATA Ile (I)<br>ATG Met (M) | ACT Thr (T)<br>ACC Thr (T)<br>ACA Thr (T)<br>ACG Thr (T) | AAT Asn (N)<br>AAC Asn (N)<br>AAA Lys (K)<br>AAG Lys (K) | AGT Ser (S)<br>AGC Ser (S)<br>AGA Arg (R)<br>AGG Arg (R) |
| G | GTT Val (V)<br>GTC Val (V)<br>GTA Val (V)<br>GTG Val (V) | GCT Ala (A)<br>GCC Ala (A)<br>GCA Ala (A)<br>GCG Ala (A) | GAT Asp (D)<br>GAC Asp (D)<br>GAA Glu (E)<br>GAG Glu (E) | GGT Gly (G)<br>GGC Gly (G)<br>GGA Gly (G)<br>GGG Gly (G) |

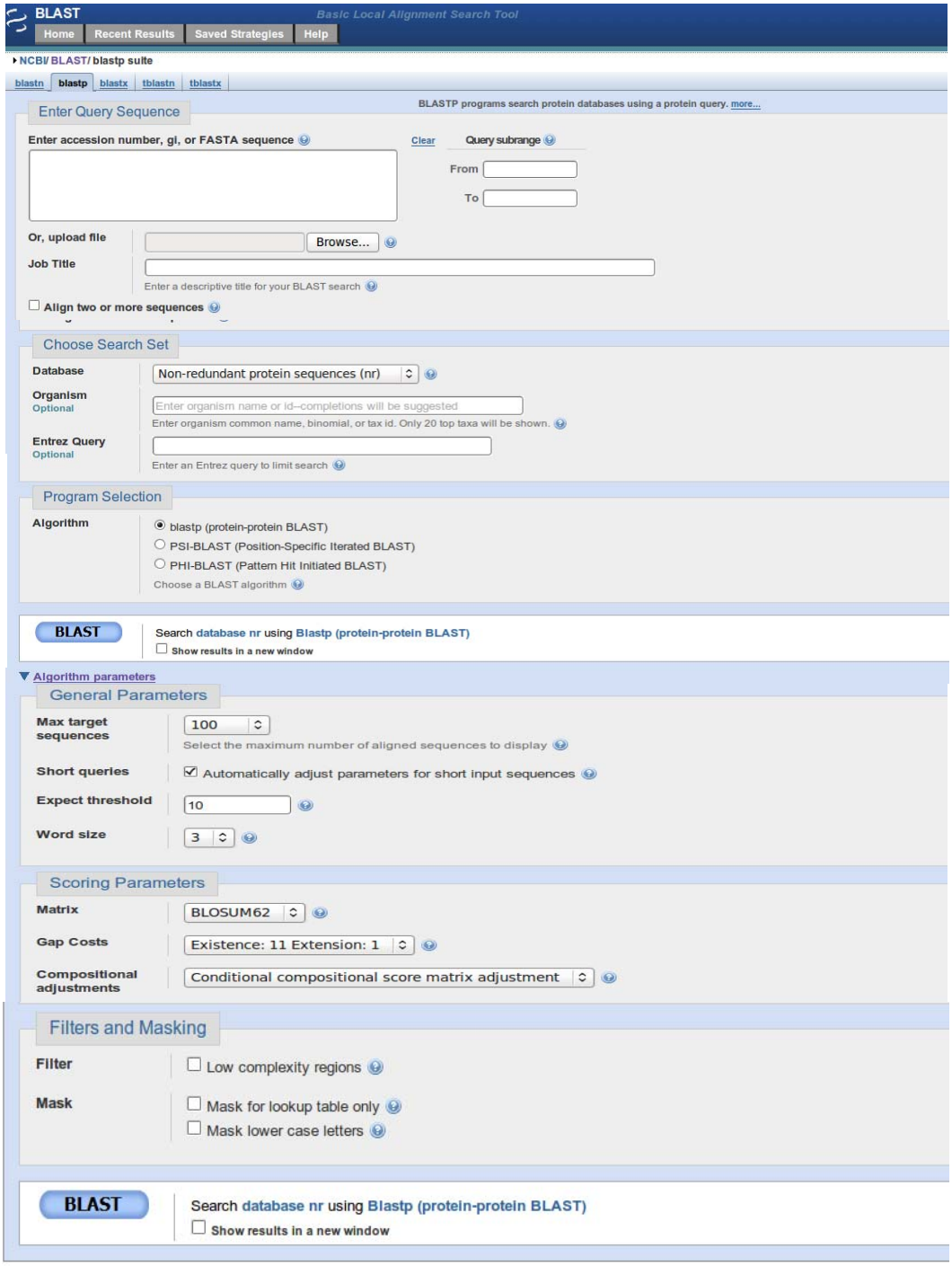


**Figure 6**: La page d'accueil de BLASTp