Recherche de séquences par similarité

Recherche de similarités dans les bases de données de séquences

- Les diapos précédentes illustraient des alignements entre une paire de séquences choisies a priori.
- Dans certains cas on dispose de la séquence d'un gène ou d'une protéine d'intérêt, et on désire l'aligner avec ***toutes*** les séquences connues.
- Ceci permet par exemple de détecter des ressemblances entre une séquence inconnue (par exemple une séquence codante qu'on vient d'identifier dans un génome nouvellement séquencé) et des séquences déjà présentes dans une base de données, et dont la fonction est connue.
- La recherche par similarité est aujourd'hui l'approche principale pour assigner des fonctions aux gènes (alignement d'ADN) et aux protéines (alignement de séquences peptidiques).

Homologie et analogie

- La similarité entre deux traits (organes, séquences) peut s'interpréter par deux hypothèses alternatives: homologie et analogie.
- **Homologie**
  - La similarité s'explique par le fait que les deux séquences divergent d'un ancêtre commun.
  - Les différences entre les deux caractères homologues résultent de l'accumulation de mutations à partir de l'ancêtre commun. Il s'agit donc d'une évolution par ***divergence évolutive***.
- **Analogie**
  - Ressemblance entre deux traits (organes, séquence) qui ne résulte pas d'une origine ancestrale commune (par opposition à l'homologie).
  - Les traits similaires sont apparus de façon ***indépendante***. Leur ressemblance peut éventuellement manifester l'effet d'une pression évolutive qui a sélectionné les mêmes propriétés.
  - Dans ce cas, on parle de ***convergence évolutive***.

Matrices de substitutions

- Une ***matrice de substitution*** associe un score à chaque paire de résidus qu'on peut trouver dans un alignement.
  - Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acide aminés).
  - La diagonale correspond aux identités.
  - Le triangle inférieur correspond à des substitutions.
  - Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.
- Les ***scores négatifs*** sont considérés comme des pénalités associées à certaines substitutions qu'on n'observe que rarement dans les alignements. Les algorithmes d'alignements tenteront donc d'éviter ces substitutions.
- Les ***scores positifs*** correspondent à des substitutions qu'on observe plus souvent que prévu, dans les alignements d'un grand nombre de séquences. Ceci suggère que ces substitutions particulières sont moins dommageable que d'autres, et on les qualifie donc de « ***substitutions conservatives*** » ou encore de « ***mutations ponctuelles acceptées*** » (***PAM***).
- Au sein d'un alignement, le terme ***similarité*** désigne les positions où se superposent des résidus ayant un score positif dans la matrice de substitution (identité ou substitution conservative).

Matrice de substitutions entre nucléotides

	A	C	G	T
A	2			
C	-2	2		
G	-2	-2	2	
T	-1	-2	-2	2

Matrice de substitutions entre acides aminés

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i *de 1 à* L) , des scores des paires de résidus ($r_{1,I}$ et $r_{2,I}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    - Valeurs typiques: entre -10 et -15
  - Pénalité d'extension de gap (***ge***)
    - Valeurs typiques: entre -0.5 et -2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	R	L	A	S	V	E	T	D	M	P	-	-	-	-	-	L	T	L	R	Q	H
	T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H

S

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement ($i \; de \; 1 \; à \; L$) , des scores des paires de résidus ($r_{1,I}$ et $r_{2,I}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    - Valeurs typiques: entre -10 et -15
  - Pénalité d'extension de gap (***ge***)
    - Valeurs typiques: entre -0.5 et -2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
	R	L	A	S	V	E	T	D	M	P	-	-	-	-	-	L	T	L	R	Q	H	
	.	|	.	|	:	:	|	.	:	.	go	ge	ge	ge	ge	.	.	|	.	.	|	
	T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H	
S	-1	+4	+0	+4	+1	+2	+5	-1	+2	-1	-10	-1	-1	-1	-1	-1	-2	+4	-2	-1	+8	= 7

Résultat de BLAST – Requête peptidique vs DB de peptides

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
          (N-terminal); homoserine dehydrogenase I (C-terminal)
          [Escherichia coli K12]
         Length = 820

 Score =  344 bits (882), Expect = 2e-95
 Identities = 247/821 (30%)  Positives = 410/821 (49%)  Gaps = 44/821 (5%)

Query: 16   KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
            KFGG+S+A+ + +LRVA I+   ++    +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5    KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75   QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
                R +  +L++GL   A+     L  +  FV          +  GI+       D++ A ++
Sbjct: 65   SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127  GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
              GE   S  +M+ VL  +G     +D  E L A       +   + E         ++   H
Sbjct: 124  CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184  PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
               +++ GF + N  GE V+LGRNGSDYSA  + A         IW+DV GVY+ DPR+V
Sbjct: 184  ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244  KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
             DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241  PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299  ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
            E  L      + +++ +++ +     P  +         +  + RA++  + +   +
Sbjct: 301  EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356
```

- La ligne entre les séquences “Query” et “Sbjct” indique les correspondances entre acides aminés.
- Identités
- Substitutions “conservatives”: paires de résidus distincts mais dont la substitution est *généralement* moins délétère que pour d’autres paires de résidus.
- Substitutions non conservatives
- Positives: identités + substitutions conservatives.
- Gaps: lacunes insérées dans une séquence afin d’optimiser l’alignement des fragments avoisinants.

Résultat de BLAST – Requête peptidique vs DB de peptides

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
          (N-terminal); homoserine dehydrogenase I (C-terminal)
          [Escherichia coli K12]
         Length = 820

 Score =  344 bits (882), Expect = 2e-95
 Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
           KFGG+S+A+ + +LRVA I+    ++    +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5   KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
               R +  +L++GL  A+    L  +  FV          +  GI+      D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
             GE  S  +M+ VL  +G     +D  E L A      +   + E         ++   H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
              +++ GF + N  GE V+LGRNGSDYSA  + A         IW+DV GVY+ DPR+V
Sbjct: 184 ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
           DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
           E  L      + +++ +++ +     P  +         +   + RA++  + +   +
Sbjct: 301 EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356
```

- Exemple de résultat de recherche par similarité de séquences.
  - Requête (query): metA
  - Protéine identifiée dans la base de données: (subject): thrA.
- Le premier critère d'évaluation d'un résultat de BLAST:
  - La e-valeur (**expect**) indique le nombre de faux-positifs attendus au hasard, si l'on plaçait le seuil au niveau du score observé (344 bits dans ce cas-ci).
  - Plus la e-valeur est faible, plus le résultat est fiable.
  - Si la e-valeur est >= 1, le résultat est peu fiable (on aurait facilement obtenu un « aussi bon » alignement avec des séquences aléatoires.

Comment interpréter la similarité entre séquences de HIV et de SARS-CoV-2 ?

Alignement de séquences de SARS-CoV-2 sur le génome du HIV

Haut: fragment le plus significatif de l'alignement de la séquence du gène S sur le génome du VIH. Noter le score Expect = 7.5. Ce score n'est significatif que s'il est nettement inférieur à 1.

Bas: fragment le plus significatif de l'alignement d'une séquence aléatoire sur le génome du VIH. Noter le score Expect = 2.1, supérieur à 1 et donc non-significatif (comme on s'y attend, puisque la séquence est aléatoire).

Conclusion: l'alignement sur lequel s'appuient Perez et Luc Montagnier correspond à ce qu'on s'attend à trouver par hasard en alignant des séquences de cette taille.

HIV-1 isolate 19828.PPH11 from Netherlands envelope glycoprotein (env) gene, partial cds

Sequence ID: HQ644953.1	Length: 1143	Number of Matches: 1	Range 1: 967 to 994

Score	Expect	Identities	Gaps	Strand
38.3 bits(41)	7.5	25/28(89%)	0/28(0%)	Plus/Plus

```
Query  86   AATGGTACTAAGAGGTTTGATAACCCTG  113
           ||||||||||| ||||| |||||| |||
Sbjct  967  AATGGTACTAAAAGGTTAGATAACACTG  994
```

HIV-1 isolate patient B clone 16.3 from Netherlands envelope glycoprotein (env) gene, complete cds

Sequence ID: HQ386166.1	Length: 2580	Number of Matches: 1	Range 1: 2493 to 2523

Score	Expect	Identities	Gaps	Strand
39.2 bits(42)	2.1	27/31(87%)	0/31(0%)	Plus/Minus

```
Query  351   CCTAAAAGTTCTTTGTAATAACTGTATTATT  381
             ||||||||||||||||||| | ||| |||
Sbjct  2523  CCTAAAAGTTCTTTGTAATATTTCTATAATT  2493
```


Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.

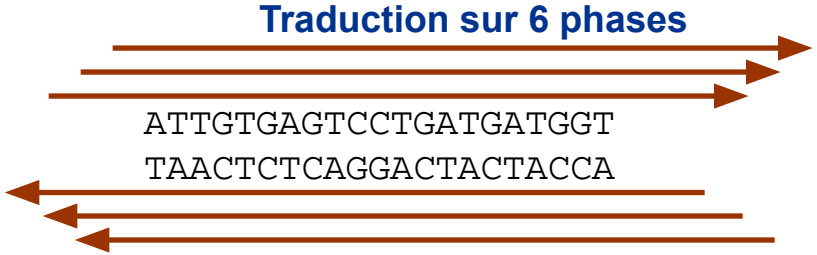
Modalités de BLAST

DNA versus protein searches

- When the query is a coding DNA sequence, it is recommended to apply searches with the translated rather than raw DNA sequences
  - This allows to introduce a substitution matrix (PAM, BLOSUM, ...), which better reflects the evolutionary changes.
  - It has been shown that some distant relationships can be detected with translated searches, but escape detection with the DNA search.
  - It is easier to filter out low complexity regions from proteins than from DNA sequences.

Traduction d'une séquence nucléique dans les 6 phases

- Si l'on dispose d'une séquence nucléique, on peut facilement déduire la séquence de la protéine qui pourrait être produite par sa traduction, sur chacun des 6 brins.
- Si cette séquence n'est pas codante, on s'attend à trouver des codons stop assez fréquemment (3 codons sur 64).
- Cependant, rien n'empêche d'aligner les 6 séquences ainsi produites avec d'autres séquences peptidiques.



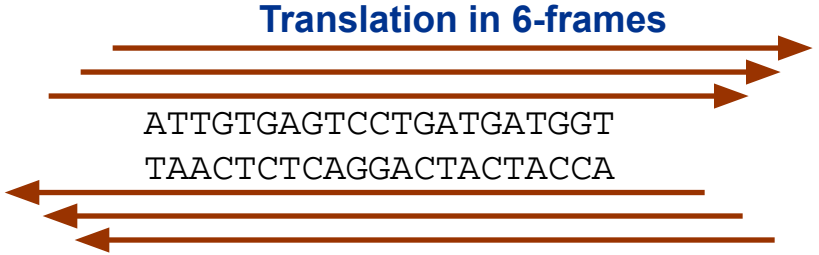
Résultat

```
F1       I   V   S   P   D   D   G
F2        L   *   V   L   M   M   V
F3         C   E   S   *   *   W   X
       1 ATTGTGAGTCCTGATGATGGT 21
         ----:----|----:----|-
       1 TAACACTCAGGACTACTACCA 21
F6        X   T   L   G   S   S   P
F5       X   Q   S   D   Q   H   H
F4         N   H   T   R   I   I   T
```

- Outil: http://www.ebi.ac.uk/Tools/st/emboss_sixpack/

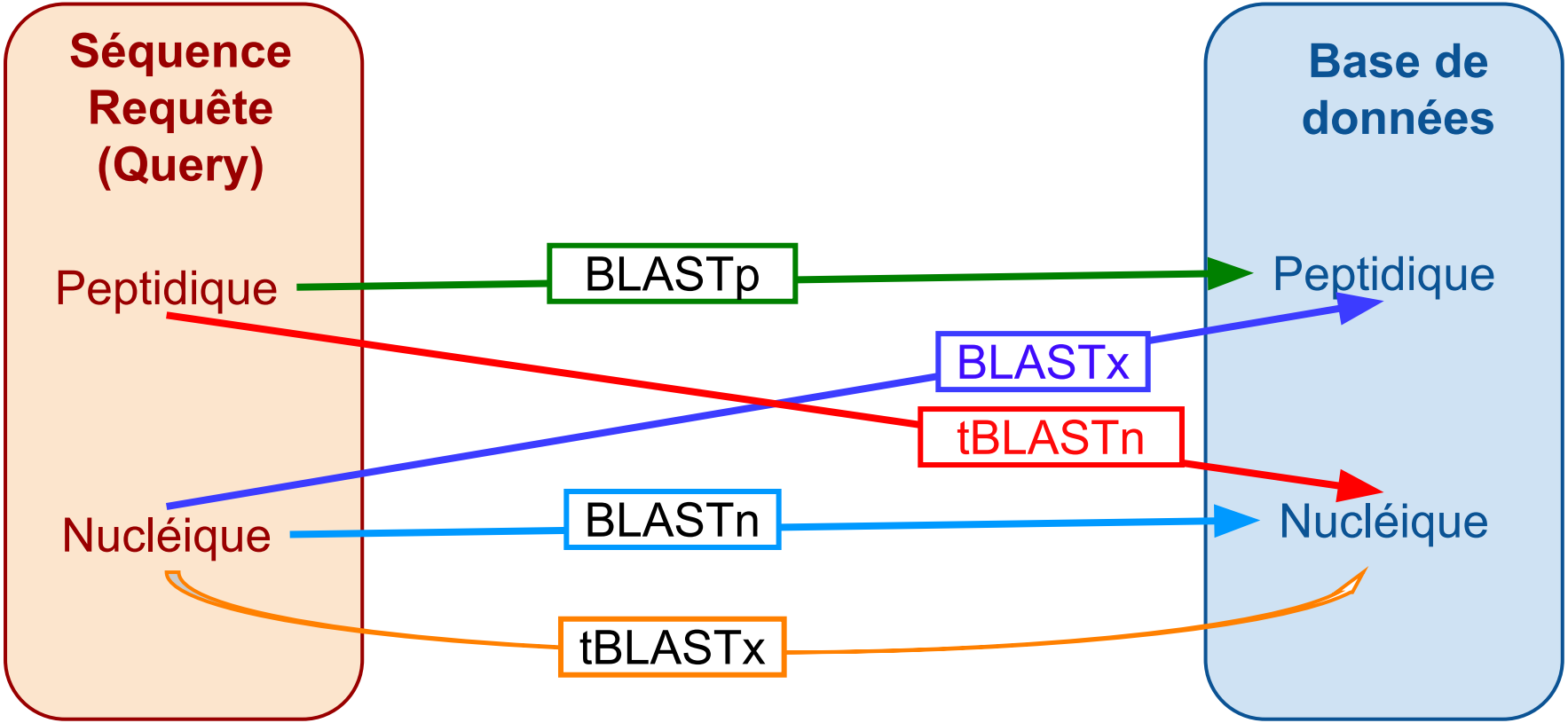
BLAST - a family of purpose-specific programs

- Different program names exist, depending on the type (protein or nucleic acid) of query and database sequences.
- For comparison between nucleic acids and proteins, the nucleic acid is translated in the 6 frames (3 frames per strand)



Query	Database	Program	Application examples	Study cases
protein	protein	blastp	Starting from a protein of known function detect putative homologs in the whole Uniprot database.	Collect sequences similar to the blue-sensitive opsin in all human proteins.
nucleic acid	nucleic acid	blastn	Match RNAi against a genome. Match mRNA (or EST) against a genome.	
nucleic acid (translated)	protein	blastx	After having sequenced a piece of DNA, search potentially coding fragments + their putative homologs without any prior knowledge of gene positions in the query sequence.	
protein	nucleic acid (translated)	tblastn	- Identify a genomic region likely to code for an homolog of a protein of interest. - Identify pseudo-genes (defective genes, with many stop codons) for a protein of interest in a genome.	Do cats see colors ? Get Human blue-sensitive opsin protein, connect UCSC genome browser, use BLAT to find similarities in Cat genome
nucleic acid (translated)	nucleic acid (translated)	tblastx		

Les modalités de BLAST



Source: Emese Meglecz