Biologie :
Théorie de l’évolution
→ mutations aléatoires
+ pression de sélection

Biologie :
Théorie de l'évolution
→ mutations aléatoires
+ pression de sélection

adaptation au milieu / à l'environnement
+ capacité à se reproduire et transmettre les mutations «bénéfiques»

Biologie :
Théorie de l'évolution
→ mutations aléatoires
+ pression de sélection

→ adaptation au milieu / à l'environnement
+ capacité à se reproduire et transmettre les mutations «bénéfiques»

Histoire de l'évolution d'une famille de séquences

→ Séquences apparentées / conservées
= pas identiques mais similaires

= dérivant d'une même séquence ancêtre
→ séquences homologues

Biologie :
Théorie de l'évolution
→ mutations aléatoires
+ pression de sélection

→ adaptation au milieu / à l'environnement
+ capacité à se reproduire et transmettre les mutations «bénéfiques»

Histoire de l'évolution d'une famille de séquences

→ Séquences apparentées / conservées
= pas identiques mais similaires

= dérivant d'une même séquence ancêtre
→ séquences homologues

Bioinformatique :
- méthodes
  de traitement ou d'analyse
- données

Biologie :
Théorie de l'évolution
→ mutations aléatoires
+ pression de sélection

→ adaptation au milieu / à l'environnement
+ capacité à se reproduire et transmettre les mutations «bénéfiques»

Histoire de l'évolution d'une famille de séquences

→ Séquences apparentées / conservées
= pas identiques mais similaires

= dérivant d'une même séquence ancêtre
→ séquences homologues

Bioinformatique :
- méthodes de traitement ou d'analyse
- données

→ alignement de 2 séquences
alignement multiple

→ banques de données

GenBank, UniProt, Pfam, PubMed

Biologie :
Théorie de l'évolution
→ mutations aléatoires
+ pression de sélection

→ adaptation au milieu / à l'environnement
+ capacité à se reproduire et transmettre les mutations «bénéfiques»

Histoire de l'évolution d'une famille de séquences

→ Séquences apparentées / conservées
= pas identiques mais similaires

= dérivant d'une même séquence ancêtre
→ séquences homologues

Recherche de séquences homologues dans une banque

BLAST

Bioinformatique :
- méthodes
de traitement ou d'analyse
- données

→ alignement de 2 séquences
alignement multiple

→ banques de données

GenBank, UniProt, Pfam, PubMed

Biologie :
Théorie de l'évolution
→ mutations aléatoires
+ pression de sélection

→ adaptation au milieu / à l'environnement
+ capacité à se reproduire et transmettre les mutations «bénéfiques»

Histoire de l'évolution d'une famille de séquences

→ Séquences apparentées / conservées
= pas identiques mais similaires

= dérivant d'une même séquence ancêtre
→ séquences homologues

Recherche de séquences homologues dans une banque

BLAST

Recherche des domaines présents sur une séquence

InterProScan

Bioinformatique :
- méthodes
de traitement ou d'analyse
- données

→ alignement de 2 séquences
alignement multiple

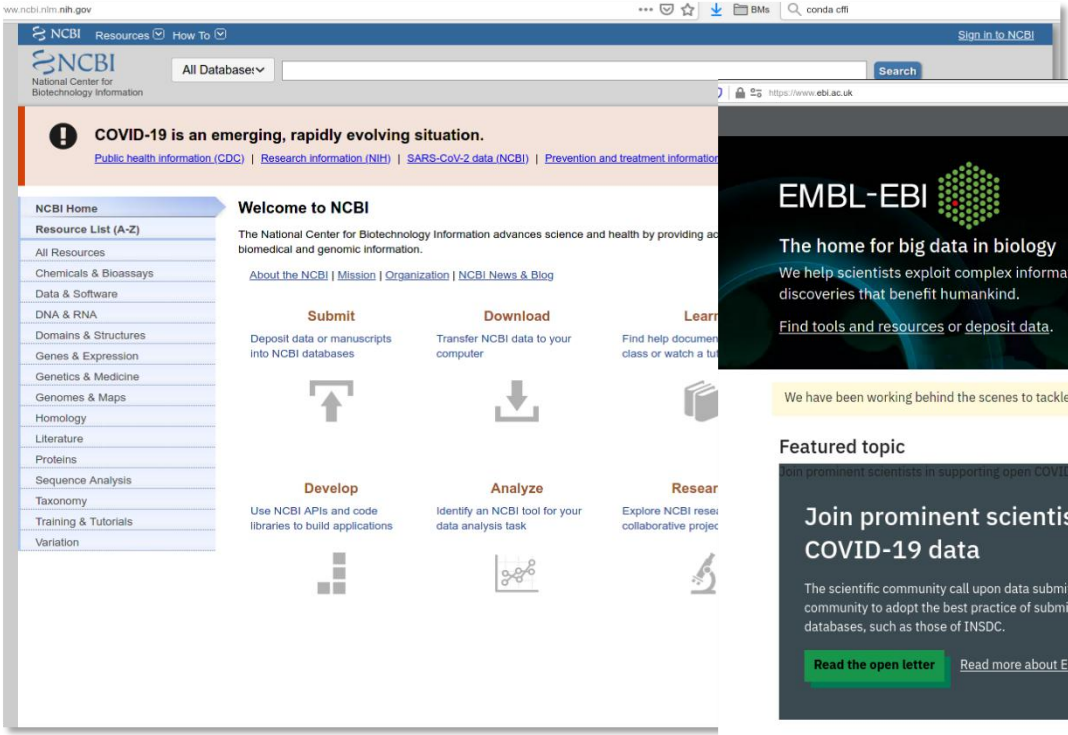
→ Famille de séquences
→ motifs ou modèle mathématique

→ banques de données

GenBank, UniProt, Pfam, PubMed

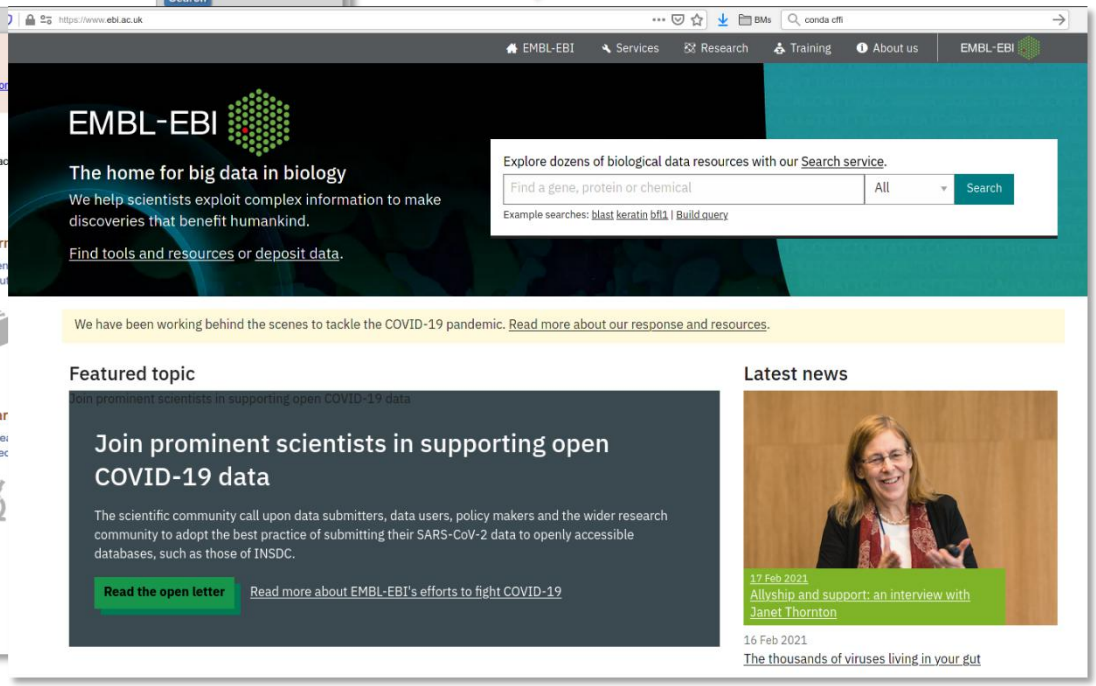
Instituts, banques de données et services

- Instituts ou centres, ex: NCBI, EBI, ...
  - Banques de données, ex: UniProt, Pfam, PubMed, ...
  - Services, ex: BLAST, InterProScan, ...



NCBI
National Center for Biotechnology Information

EBI
European Bioinformatics Institute

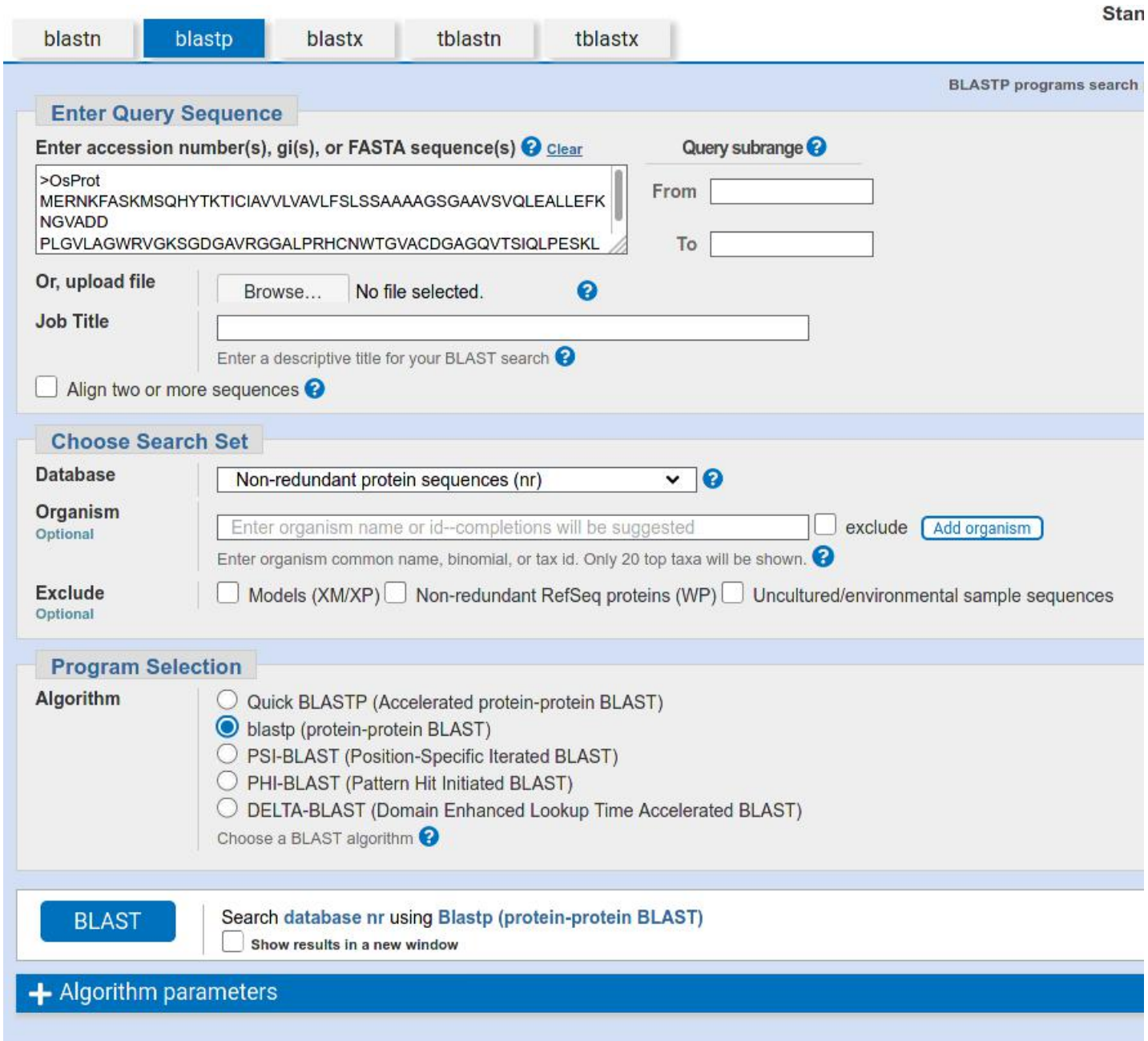


Banques de données

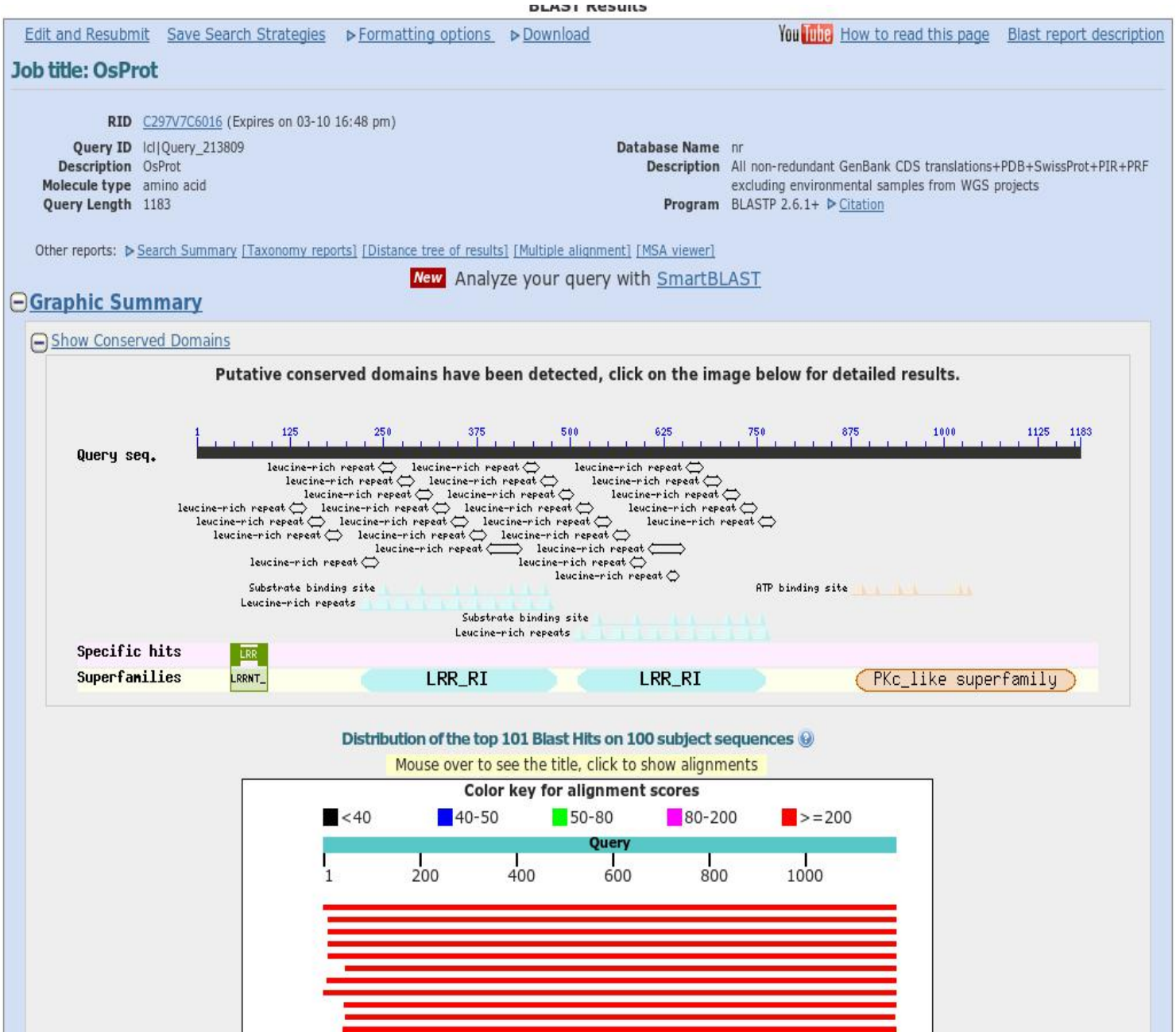
- GenBank, EMBL, ...
  - séquences nucléiques
- UniProt (séquences protéiques) fusion de :
  - trEMBL : traduction automatique des séq. nuc.)
  - SwissProt : Séquences protéiques expertisées (**Reviewed**)
- PubMed, PubMedCentral, ... : articles de la litératures bio-médicale
- GEO : Gene Expression Omnibus
  - dépôt des données d'expression des gènes (microarray + RNAseq)
- SRA : dépôt pour les données de séquençage NGS
- OMIM : maladies génétiques humaines
- SNP : Single Nuclotide Polymorphism
- dbVar : Variant en nombre de copies CNV (Copy Number Variation)
- Taxonomy
- ...

- Basic Local Alignment Search Tool
- Recherche dans une banque de séquences (Séquences «Subject») des séquences similaires à une séquence requête «Query»
- Paramètres (en entrée du programme) :
  - séquence Query
  - banque dans laquelle faire la recherche
  - type de recherche
    - blastp : séquence Query protéique → banque de séquences protéiques
    - bastn : séquence Query nucléique → banque de séq. nuc.
    - blastx : nuc. traduite en séq. prot. → banque prot.
    - tblastn : prot. → banque nuc. traduites en protéines
    - tblastx : nuc. traduite en prot. → banque nuc. traduites en prot.
  - système de score pour les substitutions (ex : A/T ou I/L) et les insertion/délétions

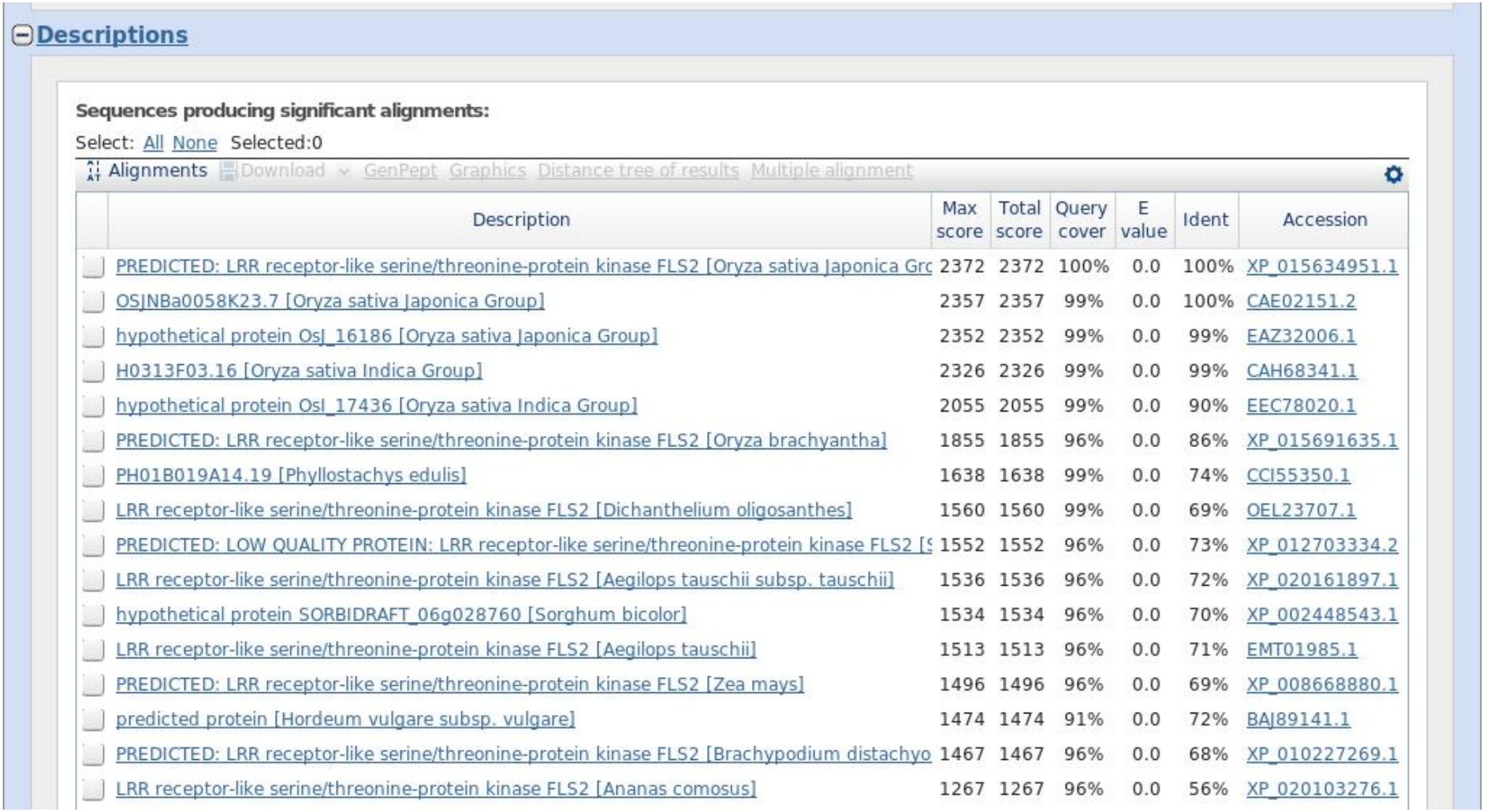
BLAST


Stan
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
>OsProt
MERNKFASKMSQHYTKTICIAVVLVAVLFSLSSAAAAGSGAAVSVQLEALLEFK
NGVADD
PLGVLAGWRVGKSGDGAVRGGALPRHCNWTGVACDGAGQVTSIQLPESKL
From
To
Or, upload file
Browse...
No file selected.
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Non-redundant RefSeq proteins (WP)
Uncultured/environmental sample sequences
Program Selection
Algorithm
Quick BLASTP (Accelerated protein-protein BLAST)
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters

BLAST Résultats (1)



BLAST Résultats (2) : liste



Descriptions

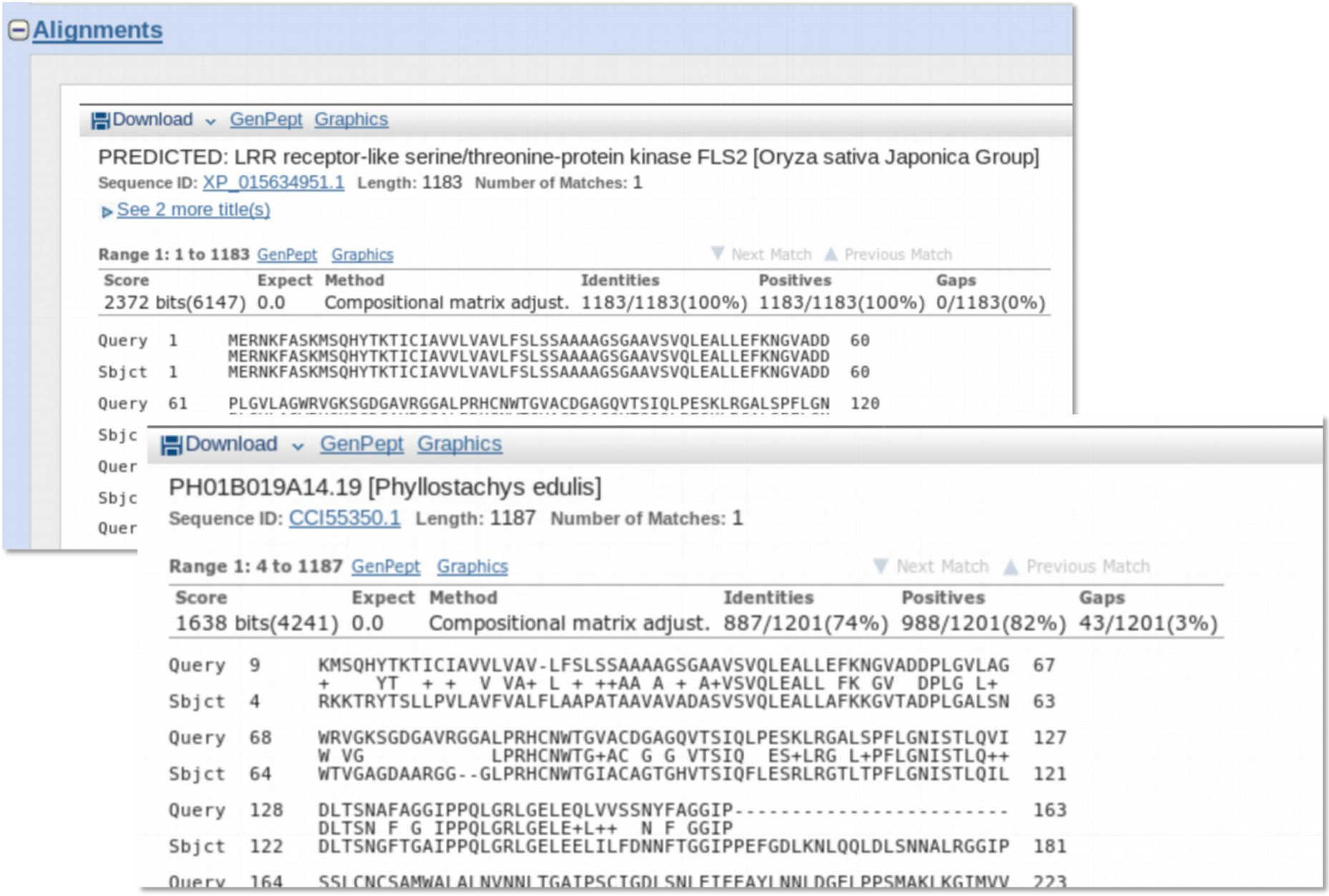
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza sativa Japonica Grc	2372	2372	100%	0.0	100%	XP_015634951.1
OSJNBa0058K23.7 [Oryza sativa Japonica Group]	2357	2357	99%	0.0	100%	CAE02151.2
hypothetical protein OsJ_16186 [Oryza sativa Japonica Group]	2352	2352	99%	0.0	99%	EAZ32006.1
H0313F03.16 [Oryza sativa Indica Group]	2326	2326	99%	0.0	99%	CAH68341.1
hypothetical protein OsI_17436 [Oryza sativa Indica Group]	2055	2055	99%	0.0	90%	EEC78020.1
PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza brachyantha]	1855	1855	96%	0.0	86%	XP_015691635.1
PH01B019A14.19 [Phyllostachys edulis]	1638	1638	99%	0.0	74%	CCI55350.1
LRR receptor-like serine/threonine-protein kinase FLS2 [Dichanthelium oligosanthes]	1560	1560	99%	0.0	69%	OEL23707.1
PREDICTED: LOW QUALITY PROTEIN: LRR receptor-like serine/threonine-protein kinase FLS2 [S	1552	1552	96%	0.0	73%	XP_012703334.2
LRR receptor-like serine/threonine-protein kinase FLS2 [Aegilops tauschii subsp. tauschii]	1536	1536	96%	0.0	72%	XP_020161897.1
hypothetical protein SORBIDRAFT_06g028760 [Sorghum bicolor]	1534	1534	96%	0.0	70%	XP_002448543.1
LRR receptor-like serine/threonine-protein kinase FLS2 [Aegilops tauschii]	1513	1513	96%	0.0	71%	EMT01985.1
PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Zea mays]	1496	1496	96%	0.0	69%	XP_008668880.1
predicted protein [Hordeum vulgare subsp. vulgare]	1474	1474	91%	0.0	72%	BAJ89141.1
PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Brachypodium distachyo	1467	1467	96%	0.0	68%	XP_010227269.1
LRR receptor-like serine/threonine-protein kinase FLS2 [Ananas comosus]	1267	1267	96%	0.0	56%	XP_020103276.1

BLAST Résultats (3) : alignements



Alignments

Download GenPept Graphics

PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza sativa Japonica Group]

Sequence ID: XP_015634951.1 Length: 1183 Number of Matches: 1

See 2 more title(s)

Range 1: 1 to 1183 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
2372 bits(6147)	0.0	Compositional matrix adjust.	1183/1183(100%)	1183/1183(100%)	0/1183(0%)

```
Query  1     MERNKFASKMSQHYTKTICIAVVLVAVLFSLSSAAAAGSGAAVSVQLEALLEFKNGVADD  60
             MERNKFASKMSQHYTKTICIAVVLVAVLFSLSSAAAAGSGAAVSVQLEALLEFKNGVADD
Sbjct  1     MERNKFASKMSQHYTKTICIAVVLVAVLFSLSSAAAAGSGAAVSVQLEALLEFKNGVADD  60

Query  61    PLGVLAGWRVGKSGDGAVRGGALPRHCNWTGVACDGAGQVTSIQLPESKLRGALSPFLGN  120
```

Download GenPept Graphics

PH01B019A14.19 [Phyllostachys edulis]

Sequence ID: CCI55350.1 Length: 1187 Number of Matches: 1

Range 1: 4 to 1187 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1638 bits(4241)	0.0	Compositional matrix adjust.	887/1201(74%)	988/1201(82%)	43/1201(3%)

```
Query  9     KMSQHYTKTICIAVVLVAV-LFSLSSAAAAGSGAAVSVQLEALLEFKNGVADDPLGVLAG  67
             +    YT  + +  V VA+ L + ++AA A + A+VSVQLEALL FK GV  DPLG L+
Sbjct  4     RKKTRYTSLLPVLAVFVALFLAAPATAAVAVADASVSVQLEALLAFKKGVTADPLGALSN  63

Query  68    WRVGKSGDGAVRGGALPRHCNWTGVACDGAGQVTSIQLPESKLRGALSPFLGNISTLQVI  127
             W VG           LPRHCNWTG+AC G G VTSIQ  ES+LRG L+PFLGNISTLQ++
Sbjct  64    WTVGAGDAARGG--GLPRHCNWTGIACAGTGHVTSIQFLESRLRGTLTPFLGNISTLQIL  121

Query  128   DLTSNAFAGGIPPQLGRLGELEQLVVSSNYFAGGIP----------------------  163
             DLTSN F G IPPQLGRLGELE+L++  N F GGIP
Sbjct  122   DLTSNGFTGAIPPQLGRLGELEELILFDNNFTGGIPPEFGDLKNLQQLDLSNNALRGGIP  181

Query  164   SSLCNCSAMWALALNVNNLTGAIPSCIGDLSNLETFFAYLNNLDGELPPSMAKLKGIMVV  223
```

Alignement de deux séquences

Alignement :

```
AACT--GGTAACCG
AGCTACGGT--CCG
```

➡ Calcul d'un score

Le score de l'alignement doit prendre en compte toutes les positions alignées : identités, substitutions et indels. Chacun de ces événements va recevoir un poids, appelé score élémentaire s_e. Le score de l'alignement correspondra à la somme des scores élémentaires correspondant aux positions alignées.

$$S = \sum_{i=1}^{l} Se(i)$$

Où l est le nombre de positions alignées

exemple: l = 14

s_e identité = +2
s_e substitution = -1
s_e indels = -2

➡ S = 9

Algorithme de programmation dynamique

Etant donné un système de score, garantit l'obtention de l'alignement optimal

Hypothèse : l'évolution est parcimonieuse
Signification: pour trouver l'alignement optimal, l'algorithme va rechercher le chemin permettant de passer d'une séquence à l'autre avec le minimum de changements

Deux types de score en fonction des algorithmes :

- score d'homologie: la valeur du score diminue avec le nombre de différences observées entre les deux séquences

- score de distance: la valeur du score augmente avec le nombre de différences observées entre les deux séquences

Exemples de systèmes de scores

	Score d'homologie	Score de distance
identité	+1	0
mismatch	-1	+1
indel	-2	+2

Algorithme de programmation dynamique

Trois types d'algorithmes d'alignement de deux séquences:

- alignement global (proposé en premier par Needleman and Wunsch). Les séquences vont être alignées sur toutes leurs longueurs (du premier au dernier résidus). Utilisé quand les séquences ont à peu près la même longueur

- alignement semi-global (pas de pénalités des gaps aux extrémités). Utilisé quand une séquence est plus courte que l'autre ou quand on recherche des chevauchements aux extrémités.

- alignement local (connu comme l'algorithme de Smith and Waterman). L'algorithme recherche les deux sous-régions les plus conservées entre les deux séquences. Seulement ces deux régions seront alignées.

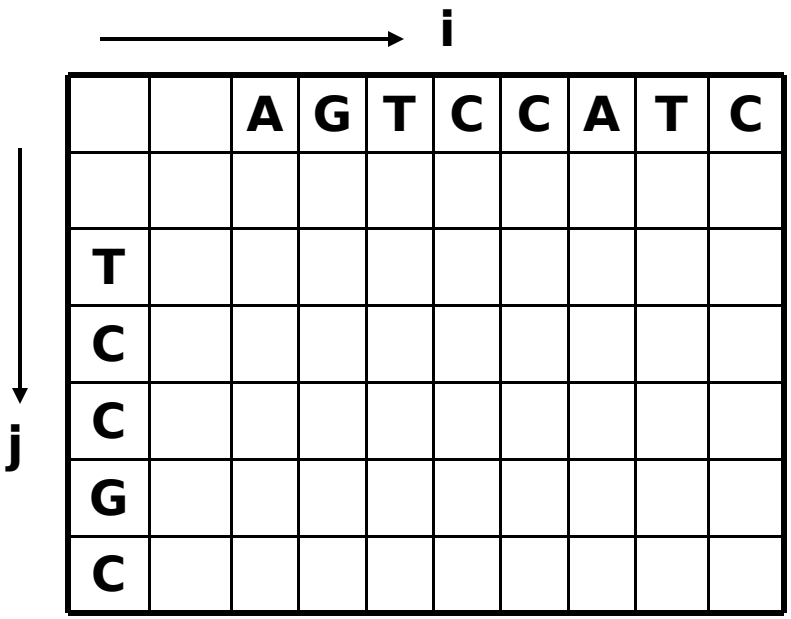
Algorithme de programmation dynamique

Comment ça marche ?

Prenons comme exemple deux séquences X et Y de longueur M et N :

X = AGTCCATC M=8
Y = TCCGC N=5

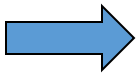
Matrice de programmation dynamique :



Le score optimal sera calculé récursivement. Le score calculé pour la cellule (i,j) correspondra au meilleur alignement des résidus
$x_1 \ldots\ldots x_i$ avec les résidus $y_1 \ldots\ldots y_j$

Alignement de deux séquences protéiques

Les acides aminés composant une protéine peuvent avoir des propriétés physico-chimiques similaires.



La structure 3D dépend de ces caractéristiques

Une similitude au niveau de ces propriétés sera suffisante pour permettre la substitution d'un acide aminé en un autre sans perturber la fonction de la protéine (par exemple, échange de l'acide aminé hydrophobe valine en leucine).

Lors de la comparaison de deux séquences protéiques, nous devons prendre en compte ces similitudes et pas seulement les identités.

Comment quantifier la similitude entre deux acides aminés ?

- calculer une distance entre acides aminés basée sur leurs caractéristiques
- estimer la fréquence de substitution de l'acide aminé X en Y au cours de l'évolution

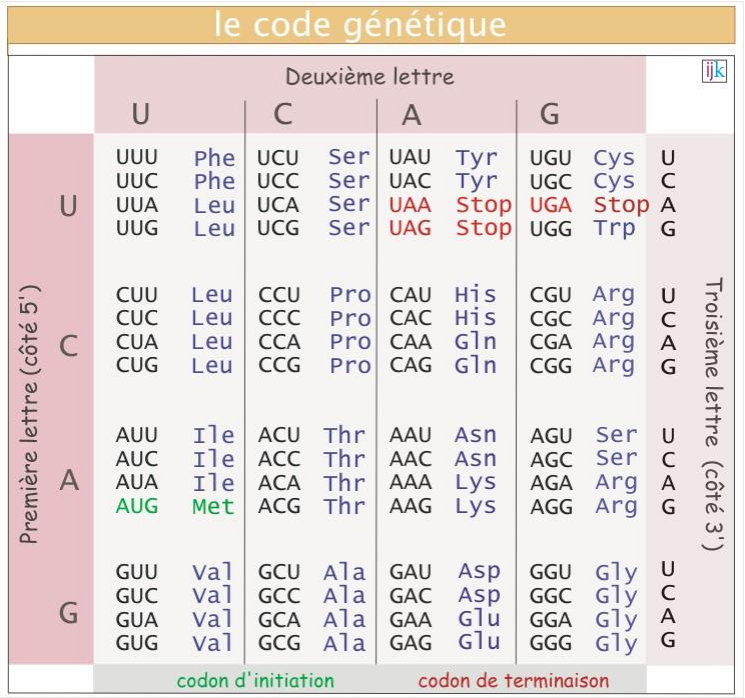
Les deux approches donnent une matrice (20,20) symétrique par rapport à la diagonale. Cependant, les matrices les plus utilisées ont été obtenues par la seconde approche et sont appelées « matrices de substitution »

Approches basée sur les caractéristiques des a.a.

Basée sur le code génétique : une substitution d'un a.a. en un autre se produit d'autant plus rarement que cela nécessite un plus grand nombre de mutations au niveau ADN.

→ Matrice génétique (Fitch, 1966)

Identité : +3
1 mutation ADN = 2 nt identiques : +2
2 mutations ADN = 1 nt identique : +1
3 mutations ADN = 0 nt identique : 0



Basée sur les propriétés physico-chimiques des a.a. :

- composition, polarité, volume moléculaire (Grantham, 1974)
- volume et polarité (Miyata *et al.*, 1979)
- paramêtres de Chou et Fasman (structures secondaires), polarité et hydrophobicité (Rao, 1987)

Approches basée sur les fréquences de substitutions des a.a. au cours de l'évolution

Principe :

• les séquences homologues ont conservées des fonctions similaires
• deux a.a. se ressembleront d'autant plus que la fréquence de substitution observée est grande puisque ces substitutions n'auront pas modifié la fonction de la protéine
• il est possible d'estimer la fréquence avec laquelle un a.a. est remplacé par un autre au cours de l'évolution à partir de séquences homologues alignées

Principales approches :

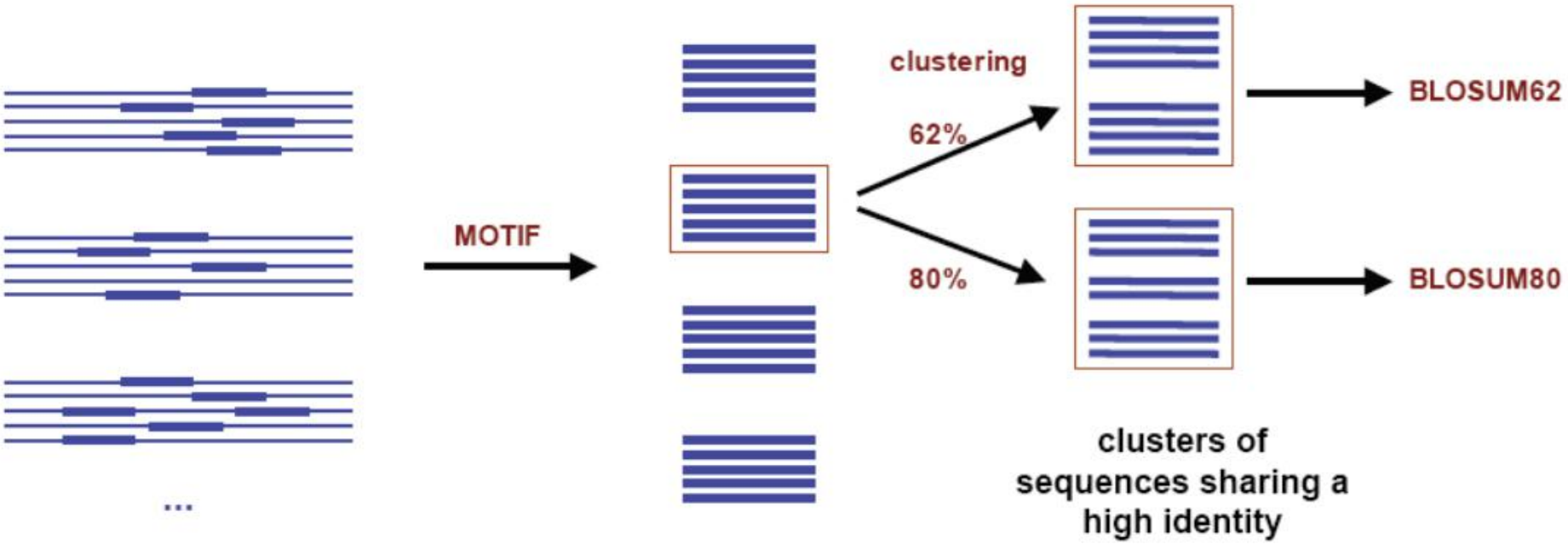
• Comparaison directe des séquences (alignement global) : matrices PAM (Dayhoff, 1978)
• Comparaison des domaines protéiques (régions les plus conservées) : matrices **BLOSUM** (Henikoff et Henikoff, 1992)
• Alignement des séquences en comparant leur structure secondaire ou tertiaire

Matrices BLOSUM (Henikoff et Henikoff, 1992)

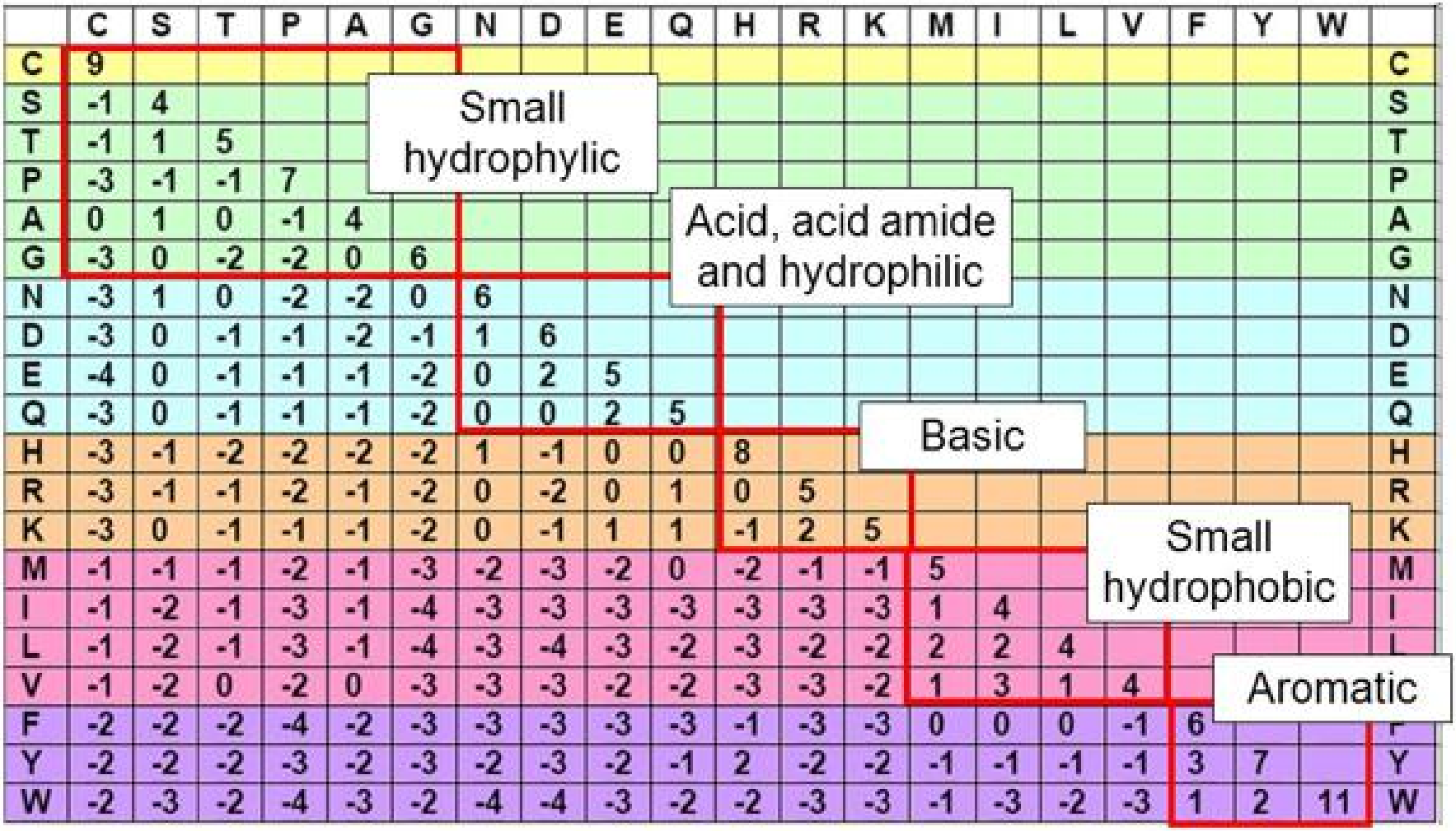
BLOSUM : BLOcks SUbstitution Matrix

Principe :
• Obtention à partir de blocs de séquences alignées (alignement multiple sans brèche)
• Pour une paire d'a.a. : log (fréquence observée / fréquence attendue)

Avantages par rapport aux matrices PAM :
• contrairement aux matrices PAM, les matrices BLOSUM pour différentes distances évolutives sont obtenues directement avec des séquences plus ou moins divergentes
• l'utilisation de blocs plutôt que de séquences complètes : modélise les contraintes uniquement sur les régions conservées
• obtenues à partir d'un plus grand jeu de données (>2000 blocks, > 500 familles)



La matrice BLOSUM62



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

Alignement de deux séquences protéiques : Matrices de substitution

Famille de matrices correspondant à différentes distances évolutives entre les séquences :

PAM120 et BLOSUM80 : estimation des fréquences de substitution entre acides aminés pour des séquences proches dans l'évolution (courtes distances)

PAM250 et BLOSUM45 : estimation des fréquences de substitution entre acides aminés pour des séquences distantes dans l'évolution (longues distances)

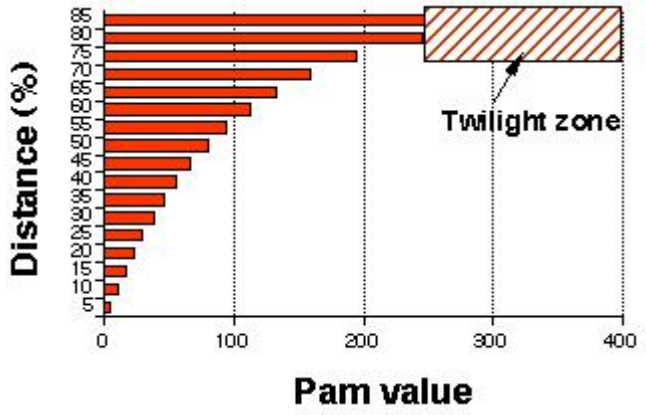
PAM160 et BLOSUM62 : estimation des fréquences de substitution entre acides aminés pour des séquences ayant des distances évolutives intermédiaires.

Source figure : ebi.ac.uk

longueur	matrice	ouverture de gap	extension de gap
300+	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
300+	PAM250	-10	-2
85-300	PAM120	-16	-4

Recommandations (à adapter)

distance %	PAM
1	1
25	30
50	80
80	246



Source figure : Infobiogen.fr

Banque de domaines

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Family: *DnaJ* (PF00226)

1471 architectures 68327 sequences 12 interactions 7853 species 87 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Summary: DnaJ domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Chaperone DnaJ | Pfam | InterPro

This is the Wikipedia entry entitled "Chaperone DnaJ". **More...**

Chaperone DnaJ

Edit Wikipedia article

In molecular biology, **chaperone DnaJ**, also known as **Hsp40** (heat shock protein 40 kD), is a molecular chaperone protein. It is expressed in a wide variety of organisms from bacteria to humans.[1][2]

Contents[hide]

Function

Molecular chaperones are a diverse family of proteins that function to protect proteins from irreversible aggregation during synthesis and in times of cellular stress. The bacterial molecular chaperone DnaK is an enzyme that couples cycles of ATP binding, hydrolysis, and ADP release by an N-terminal ATP-hydrolyzing domain to cycles of sequestration and release of unfolded proteins by a C-terminal substrate binding domain. Dimeric GrpE is the co-chaperone for DnaK, and acts as a nucleotide exchange factor, stimulating the rate of ADP release 5000-fold.[3] DnaK is itself a weak ATPase; ATP hydrolysis by DnaK is stimulated by its interaction with another co-chaperone, DnaJ. Thus the co-chaperones DnaJ and GrpE are capable of tightly regulating the nucleotide-bound and substrate-bound state of DnaK in ways that are necessary for the normal housekeeping functions and stress-related functions of the DnaK molecular chaperone cycle.

DnaJ domain

PDB rendering based on 1hdj.

Identifiers	
Symbol	DnaJ
Pfam	PF00226

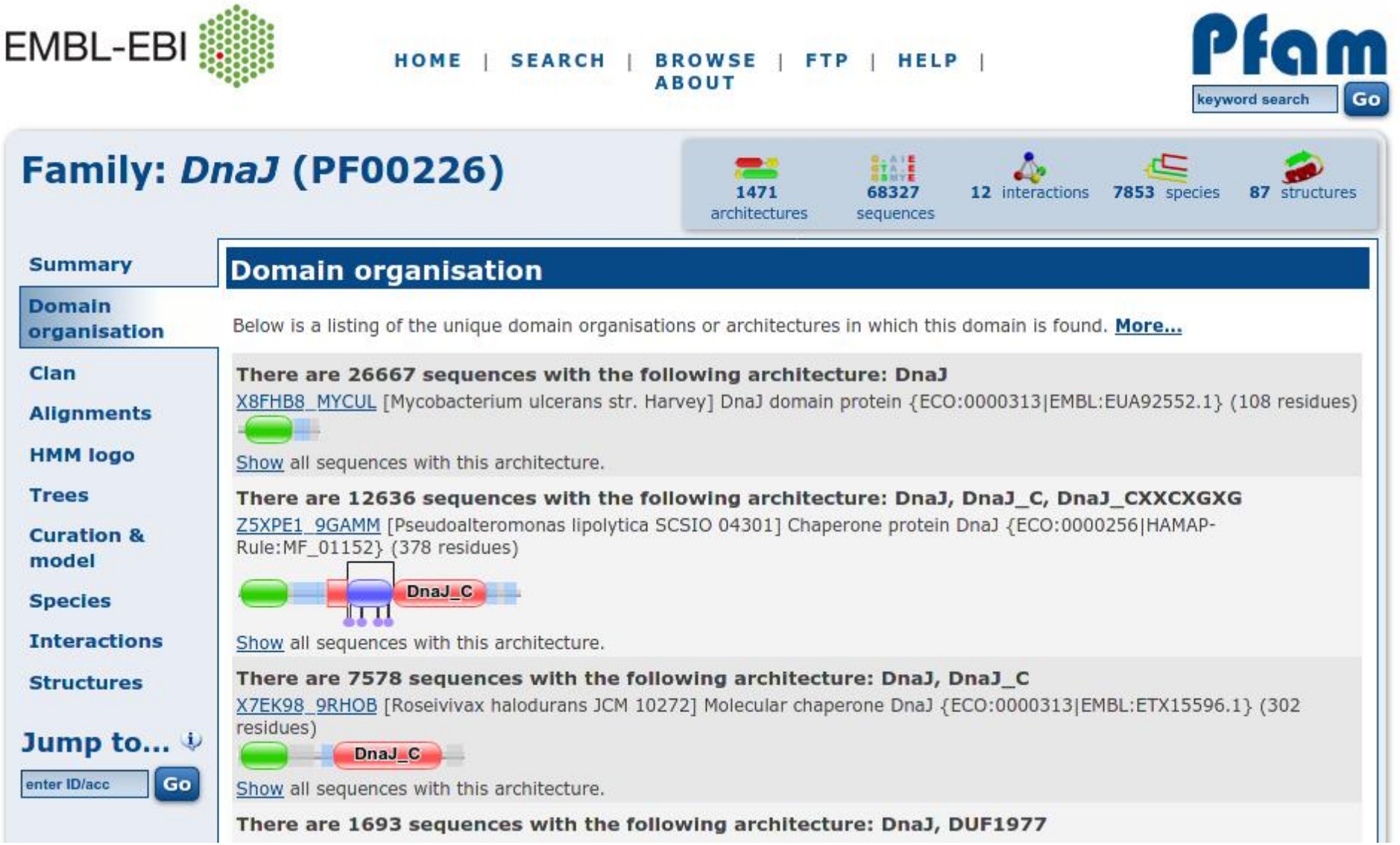
Domaine protéique : séquence conservée

EMBL-EBI

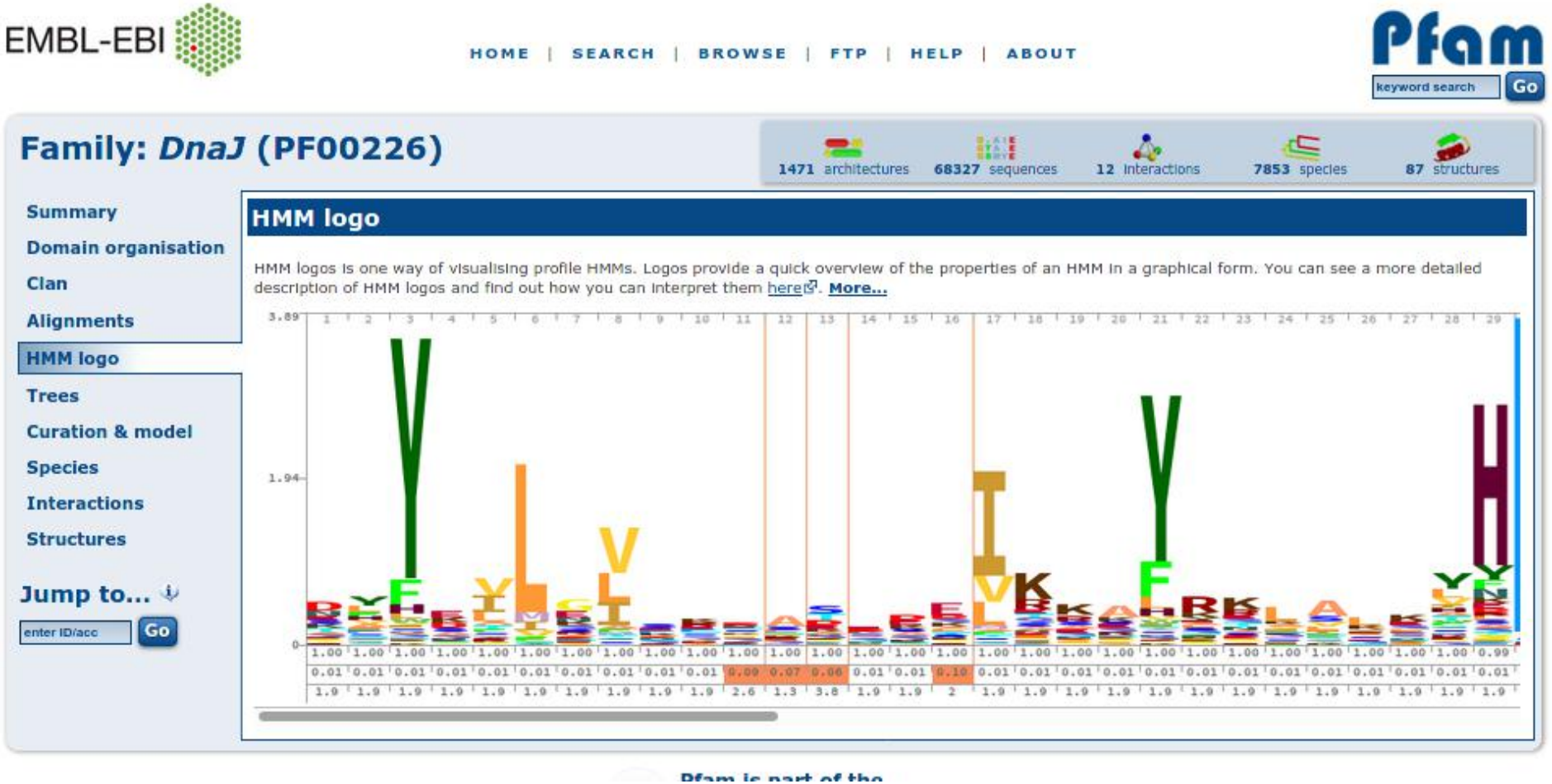
Seed sequence alignment for PF00226

```
Y6693_DROME/15-79          DVYKLMELARG....A.G.....EKE..VKKAYHKLSLLVHPD..RV.....PEEQK..AESTE....KFK.VLSKLYQV
DNJ13_ARATH/15-80          ELYALLNLSPE....A.S.....DEE..IRKAYRQWAQVYHPD..KI....QSPQMK..EVATE....NFQ.RICEAYEI
YCJD_SCHPO/4-68            DYYAILNITPK....A.S.....AEE..IKYAYKKAALETHPD..RV....SPSAR...ARATE....QFQ.LVNEAYYV
Q9SJI1_ARATH/76-138        DYYAVLGLLPD....A.T.....QEE..IKKAYYNCMKSCHPD..LS.....GNDP....ETTN.....FCMFINDIYEI
JJJ2_YEAST/13-74           TYYSILGLTSN....A.T.....SSE..VHKSYLKLARLLHPD..KT.....KSDK.....SEE....LFK.AVVHAHSI
YCJ3_SCHPO/8-72            ELYLALGLPKD....A.T.....SDQ..IKESYYRLSRLFHPD..RH.....TADQK..AAAEE....KFQ.IIQHAYEV
JJJ1_YEAST/4-67            CYYELLGVETH....A.S.....DLE..LKKAYRKKALQYHPD..KN.....PDNV...EEATQ....KFA.VIRAAYEV
O62360_CAEEL/3-66          CHYEVLEVERD....A.D.....DDK..IKKNYRKLALKWHPD..KN.....PDRI...EECTQ....QFR.LLQAAYDV
Q9W0X8_DROME/3-66          CYYEELELQRN....A.N.....DGD..IKSAYRKMALRWHPD..KN.....PDRL...AEAKE....RFQ.LIQQAYEV
DJP1_YEAST/6-68            EYYDLLGVSTT....A.S.....SIE..IKKAYRKKSIQEHPD..KN.....PNDP....TATE....RFQ.AISEAYQV
CAJ1_YEAST/6-68            EYYDILGIKPE....A.T.....PTE..IKKAYRRKAMETHPD..KH.....PDDP....DAQA....KFQ.AVGEAYQV
Q9XEM8_ORYSI/285-347       AYYDTLGVSVD....A.S.....PAE..IKKAYYLKAKQVHPD..KN.....PGNP....DAAQ....KFQ.ELGEAYQV
Q9SJS8_ARATH/6-68          EYYEILGVKTD....A.S.....DAE..IKKAYYLKARKVHPD..KN.....PGDP....QAAK....NFQ.VLGEAYQV
O49288_ARATH/6-67          VYYDVLGVTPS....A.S.....EEE..IRKAYYIKARQVHPD..KN.....QGDP...LAAEK.....Q.VLGEAYQV
DNJ10_ARATH/6-68           EYYDVLGVSPT....A.T.....ESE..IKKAYYIKARQVHPD..KN.....PNDP....QAAH....NFQ.VLGEAYQV
YAY1_SCHPO/8-71            EYYDLLGISTD....A.T.....AVD..IKKAYRKLAVKYHPD..KN.....PDDP...QGASE....KFQ.KISEAYQV
YHXB_SCHPO/9-72            DYYDILNISVD....A.D.....GDT..IKKSYRRLAILYHPD..KN.....RENP...EAARE....KFQ.KLAEAYQV
XDJ1_SCHPO/6-70            KLYDILEVHFE....A.S.....AEE..IKKSYKRLALLHHPD..KA.....PIHEK..EEAAE....RFR.GVQEAYDI
Q9VGR0_DROME/10-72         NLYDLLGISLE....S.D.....QNE..IRKAYRKKALECHPD..KN.....PDNP....KAVE....RFH.ELSKALEI
CWC23_YEAST/15-84          NLYDVLELPTP....L.DVHTIYDDLPQIKRKYRTLALKYHPD..KH.....PDNP....SIIH....KFH.LLSTATNI
CWC23_YEAST/15-84 (SS)     -HHHHTT----------HHHHHHHHHHHHHHH-TT..T-------------------HHH.HHHHHHHH
HLJ1_YEAST/21-82           EFYEILKVDRK....A.T.....DSE..IKKAYRKLAIKLHPD..KN......SHP....KAGE....AFK.VINRAFEV
DNJ1_CAEEL/137-198         DYYEILKIDKK....A.S.....DDD..IRKEYRKLALKLHPD..KC......RAP....HATE....AFK.ALGNAYAV
Q9VFP0_DROME/106-167       DYYEVLGVSKT....A.T.....DSE..IKKAYKKLALQLHPD..KN......KAP....GAVE....AFK.ALGNAAGV
YNF5_SCHPO/113-174         QYYEILDLKKT....C.T.....DTE..IKKSYKKLALQLHPD..KN......HAP....SADE....AFK.MVSKAFQV
DJB12_MOUSE/111-172        DYYEILGVSRS....A.S.....DED..LKKAYRKLALKFHPD..KN......HAP....GATE....AFK.AIGTAYAV
O43978_BABBO/14-72         KFYKVLGLSRD....C.S.....ESE..IKKAYRKLAIKHHPD..KG.............GDSE....MFK.EITRAYEV
O24074_MEDSA/14-72         KYYDILGVSKS....A.S.....EDE..IKKAYRKAAMKNHPD..KG......GDP.......E....KFK.ELGQAYEV
DNJH2_ALLPO/13-71          KYYEVLGVSKN....A.T.....PED..LKKAYRKAAIKNHPD..KG......GDP.......E....KFK.EIGQAYEV
O96455_DICDI/5-66          KFYDILGVARD....A.S.....ETD..IKKAYRKLAIKYHPD..KN......PDP....AAVE....KFK.ELTVAYEV
Q9U062_GIAIN/6-65          EFYDLLGVSPS....A.D.....PQT..IKKRTTKLARKYHPD..KP............TGDEE....LFN.KIGRAYEV
```

Architecture en domaines



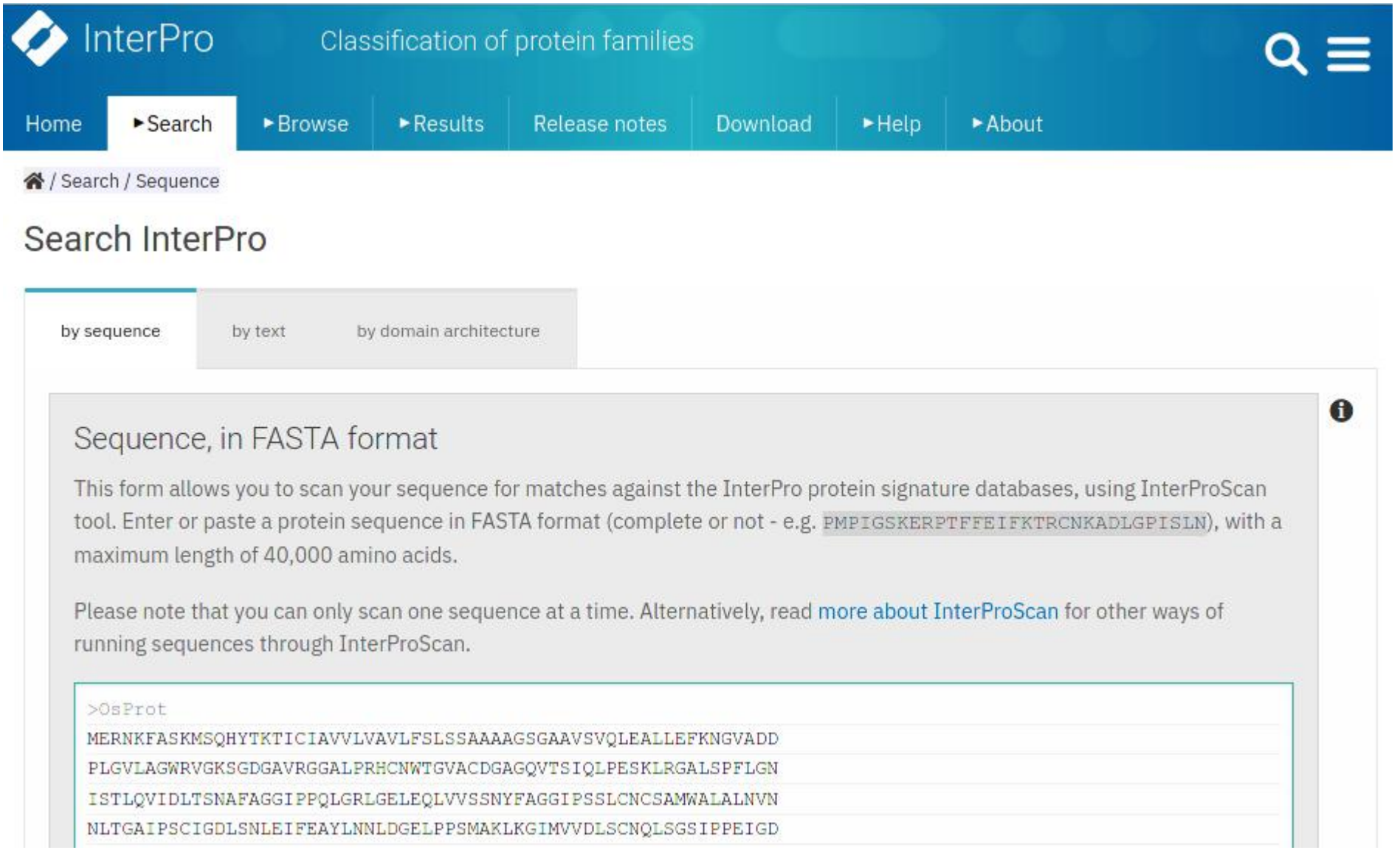
- Pfam : modèle mathématique (HMM)



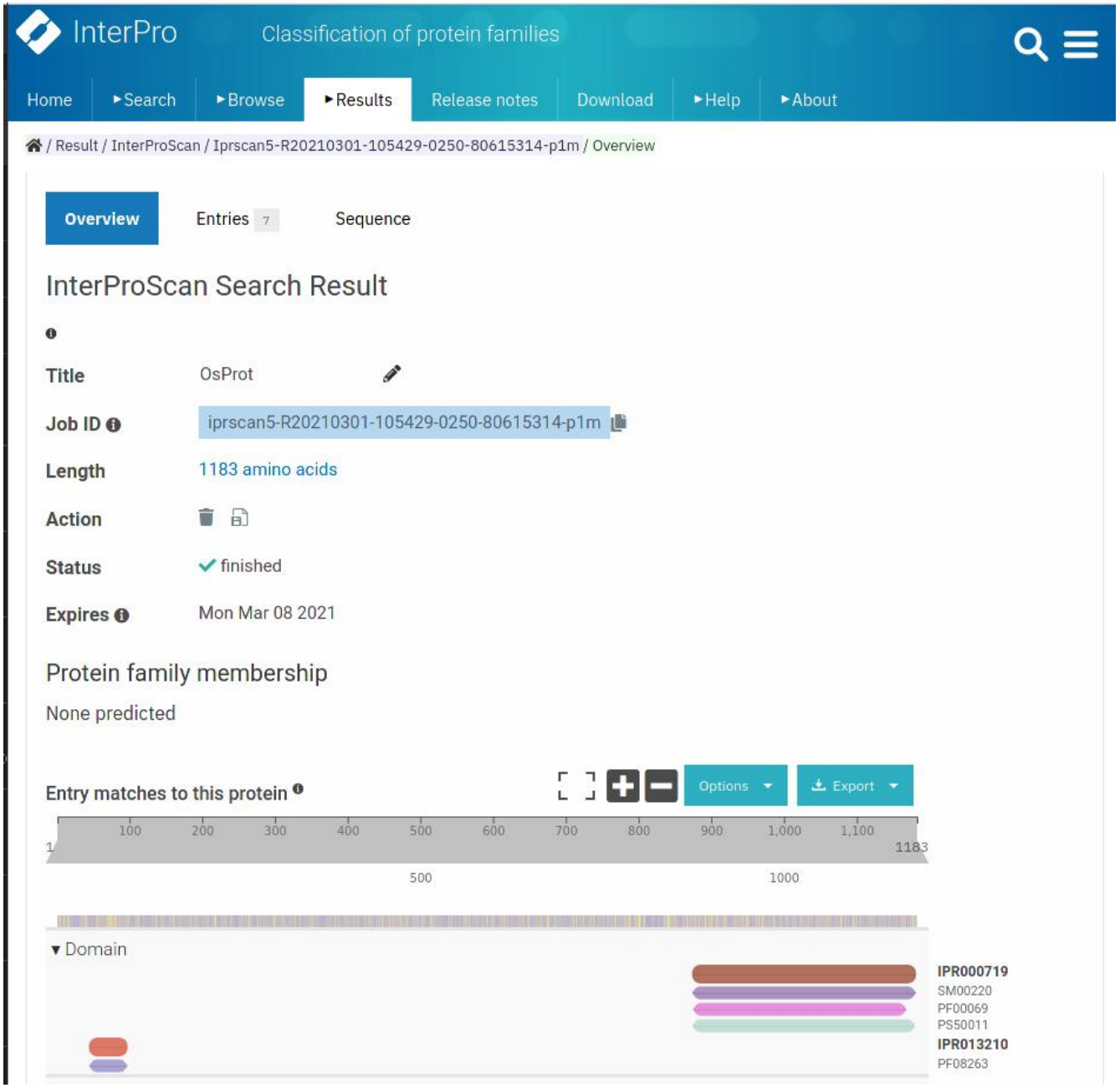
- 2 principales utilisations :
  - chercher tous les domaines présents sur une séquence donnée
  - chercher toutes les séquences arborant un domaine donné

- InterPro : méta-banque de données regroupant différentes banques de domaines
- InterProScan : exemple de recherche de domaines présents sur une séquence données

InterProScan



InterProScan résultats (1)



InterProScan résultats (2)