

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

Bioinformatique :

- méthodes  
de traitement ou d'analyse  
- données

alignement de 2 séquences  
alignement multiple

BLAST

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

Recherche des domaines  
présents sur une séquence

Bioinformatique :

- méthodes  
de traitement ou d'analyse  
- données

alignement de 2 séquences  
alignement multiple

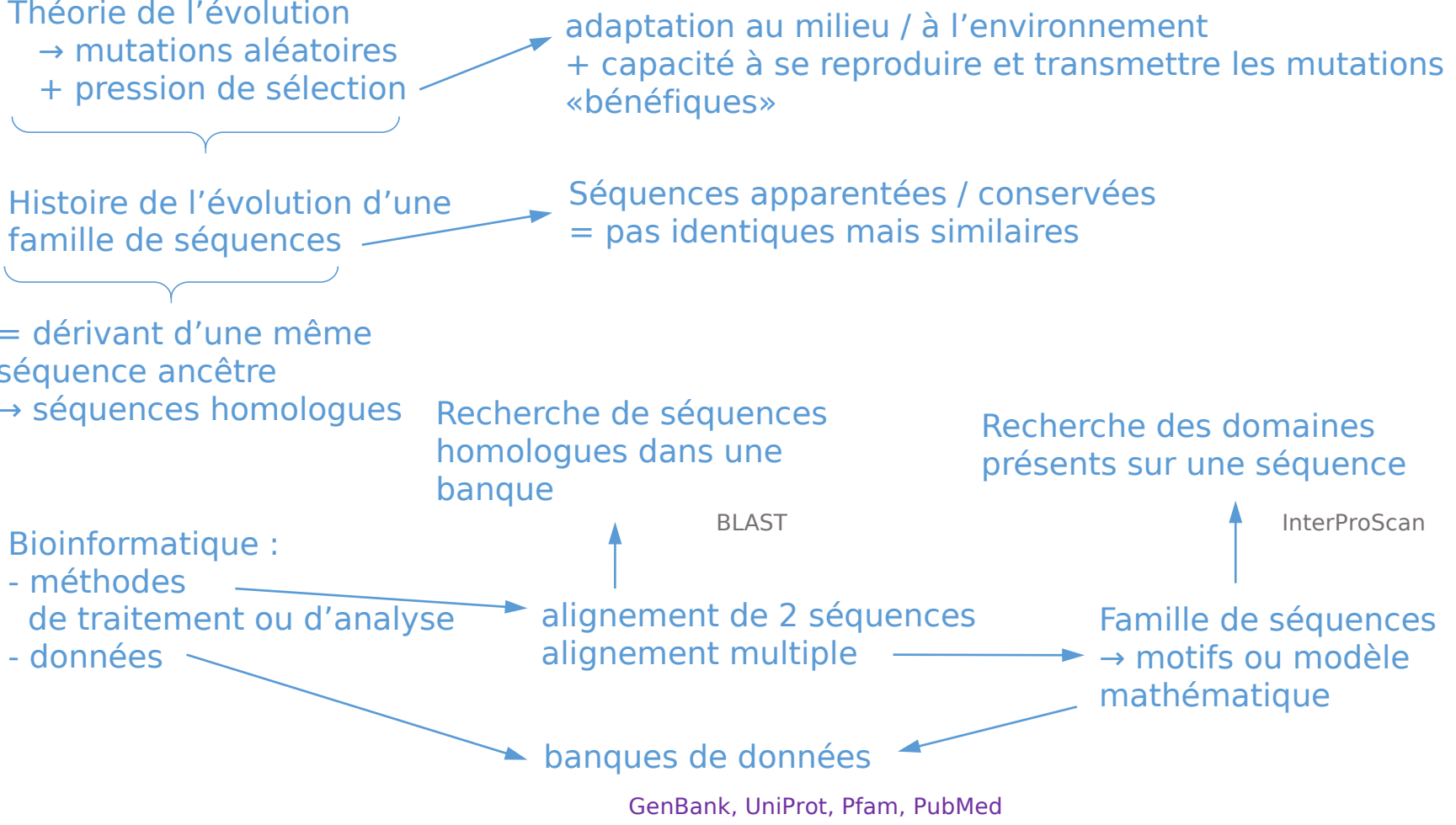
Famille de séquences  
→ motifs ou modèle  
mathématique

banques de données

GenBank, UniProt, Pfam, PubMed

BLAST

InterProScan



# Instituts, banques de données et services

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

The image displays two web browser screenshots side-by-side. The left screenshot shows the NCBI (National Center for Biotechnology Information) homepage. It features a navigation menu on the left with categories such as 'All Resources', 'Data & Software', 'Genetics & Medicine', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area includes a 'Welcome to NCBI' message and several service tiles: 'Submit' (Deposit data or manuscripts into NCBI databases), 'Download' (Transfer NCBI data to your computer), 'Learn' (Find help documents, class or watch a tutorial), 'Develop' (Use NCBI APIs and code libraries to build applications), 'Analyze' (Identify an NCBI tool for your data analysis task), and 'Research' (Explore NCBI research collaborative projects). A prominent banner at the top left of the NCBI page reads 'COVID-19 is an emerging, rapidly evolving situation.' with links to public health information, research information, SARS-CoV-2 data, and prevention and treatment information.

The right screenshot shows the EMBL-EBI (European Bioinformatics Institute) homepage. It features a search bar at the top with the text 'Explore dozens of biological data resources with our Search service.' and a search input field containing 'Find a gene, protein or chemical'. Below the search bar is a 'Featured topic' section titled 'Join prominent scientists in supporting open COVID-19 data', which includes a sub-heading 'Join prominent scientists in supporting open COVID-19 data' and a paragraph about the scientific community's call to action. A 'Read the open letter' button is visible. To the right of the featured topic is a 'Latest news' section with a photo of Janet Thornton and a caption: '17 Feb 2021 Allyship and support: an interview with Janet Thornton'. Below this is another news item dated '16 Feb 2021' titled 'The thousands of viruses living in your gut'.

NCBI  
National Center for Biotechnology Information



## 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 littérature 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 Nucleotide 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
      - blastn : 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

- blastn
- blastp**
- blastx
- tblastn
- tblastx

BLASTP programs search

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

```
>OsProt
MERNKFASKMSQHYTKTICIAVVLVAVLFLSSAAAAGSGAAVSVQLEALLEFK
NGVADD
PLGVLAGWRVKGSGDGAVRGGALPRHCNWTGVACDGAGQVTSIQLPESKL
```

Query subrange [?](#)

From

To

Or, upload file

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

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 (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
<input type="checkbox"/>	<a href="#">PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza sativa Japonica Gr</a>	2372	2372	100%	0.0	100%	<a href="#">XP_015634951.1</a>
<input type="checkbox"/>	<a href="#">OSJNBa0058K23.7 [Oryza sativa Japonica Group]</a>	2357	2357	99%	0.0	100%	<a href="#">CAE02151.2</a>
<input type="checkbox"/>	<a href="#">hypothetical protein OsJ_16186 [Oryza sativa Japonica Group]</a>	2352	2352	99%	0.0	99%	<a href="#">EAZ32006.1</a>
<input type="checkbox"/>	<a href="#">H0313F03.16 [Oryza sativa Indica Group]</a>	2326	2326	99%	0.0	99%	<a href="#">CAH68341.1</a>
<input type="checkbox"/>	<a href="#">hypothetical protein Osl_17436 [Oryza sativa Indica Group]</a>	2055	2055	99%	0.0	90%	<a href="#">EEC78020.1</a>
<input type="checkbox"/>	<a href="#">PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Oryza brachyantha]</a>	1855	1855	96%	0.0	86%	<a href="#">XP_015691635.1</a>
<input type="checkbox"/>	<a href="#">PH01B019A14.19 [Phyllostachys edulis]</a>	1638	1638	99%	0.0	74%	<a href="#">CCI55350.1</a>
<input type="checkbox"/>	<a href="#">LRR receptor-like serine/threonine-protein kinase FLS2 [Dichantheium oligosanthos]</a>	1560	1560	99%	0.0	69%	<a href="#">OEL23707.1</a>
<input type="checkbox"/>	<a href="#">PREDICTED: LOW QUALITY PROTEIN: LRR receptor-like serine/threonine-protein kinase FLS2 [S</a>	1552	1552	96%	0.0	73%	<a href="#">XP_012703334.2</a>
<input type="checkbox"/>	<a href="#">LRR receptor-like serine/threonine-protein kinase FLS2 [Aegilops tauschii subsp. tauschii]</a>	1536	1536	96%	0.0	72%	<a href="#">XP_020161897.1</a>
<input type="checkbox"/>	<a href="#">hypothetical protein SORBIDRAFT_06g028760 [Sorghum bicolor]</a>	1534	1534	96%	0.0	70%	<a href="#">XP_002448543.1</a>
<input type="checkbox"/>	<a href="#">LRR receptor-like serine/threonine-protein kinase FLS2 [Aegilops tauschii]</a>	1513	1513	96%	0.0	71%	<a href="#">EMT01985.1</a>
<input type="checkbox"/>	<a href="#">PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Zea mays]</a>	1496	1496	96%	0.0	69%	<a href="#">XP_008668880.1</a>
<input type="checkbox"/>	<a href="#">predicted protein [Hordeum vulgare subsp. vulgare]</a>	1474	1474	91%	0.0	72%	<a href="#">BAJ89141.1</a>
<input type="checkbox"/>	<a href="#">PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2 [Brachypodium distachyo</a>	1467	1467	96%	0.0	68%	<a href="#">XP_010227269.1</a>
<input type="checkbox"/>	<a href="#">LRR receptor-like serine/threonine-protein kinase FLS2 [Ananas comosus]</a>	1267	1267	96%	0.0	56%	<a href="#">XP_020103276.1</a>



# 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	MERNKFASKMSQHYTKTICIAVVLVAVLFSLSAAAAGSGAAVSVQLEALLEFKNGVADD	60
Sbjct	1	MERNKFASKMSQHYTKTICIAVVLVAVLFSLSAAAAGSGAAVSVQLEALLEFKNGVADD	60
Query	61	PLGVLAGWRVKGSGDGA VRGGALPRHCNWTGVACDGAGQVTSIQLPESKLRGALSPFLGN	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 - LFSLSAAAAGSGAAVSVQLEALLEFKNGVADDPLGVLAG	67
Sbjct	4	+ YT + + V VA+ L + ++AA A + A+VSVQLEALL FK GV DPLG L+ RKKTRYTSLLPVLAVFVALFLAAPATAAVAVADASVSVQLEALLAFKKGVTADPLGALS	63
Query	68	WRVGKSGDGA VRGGALPRHCNWTGVACDGAGQVTSIQLPESKLRGALSPFLGNISTLQVI	127
Sbjct	64	W VG LPRHCNWTG+AC G G VTSIQ ES+LRG L+PFLGNISTLQ++ WTVGAGDAARGG - -GLPRHCNWTGIACAGTGHVTSIQFLESRLRGTLPFLGNISTLQIL	121
Query	128	DLTSNAFAGGIPPQLGRLGELEQLVSSNYFAGGIP-----	163
Sbjct	122	DLTSN F G IPPQLGRLGELE+L++ N F GGIP DLTSNGFTGAIPPQLGRLGELEELILFDNNFTGGIPPEFGDLKNLQQLDLSNNALRGGIP	181
Query	164	SSI CNCSAMWAI AI NVNNI TGATPSC TGD I SNI E TFFAYI NNI DGEI PPSMAKI KGTMVV	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.

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

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



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.



ou



- 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 :

		→ i							
		A	G	T	C	C	A	T	C
↓ j	T								
	C								
	C								
	G								
	C								

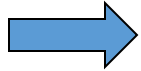
Le score optimal sera calculé récursivement. Le score calculé pour la cellule (i,j) correspondra au meilleur alignement des résidus

$x_1 \dots x_i$  avec les résidus  $y_1 \dots y_j$

# 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)
- paramètres de Chou et Fasman (structures secondaires), polarité et hydrophobicité (Rao, 1987)

le code génétique									
	Deuxième lettre								
	U	C	A	G					
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

codon d'initiation      codon de terminaison

## 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)

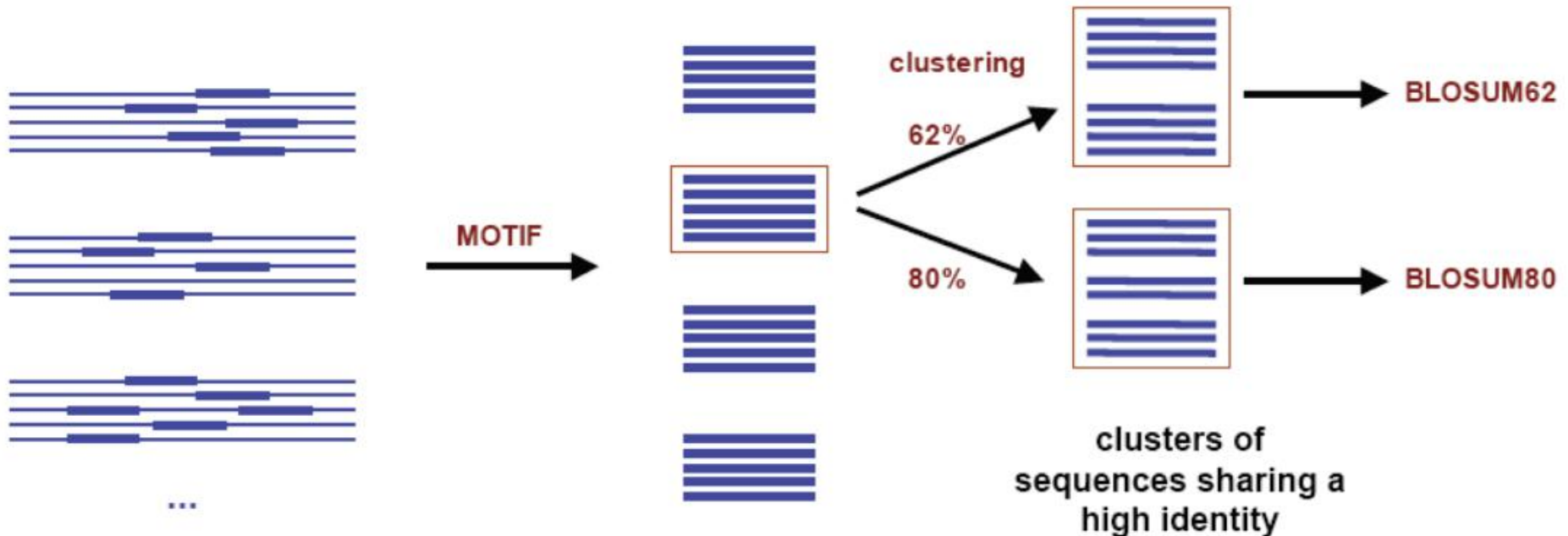
BLOSUM : BLOcks SUBstitution Matrix

Principe :

- Obtention à partir de blocs de séquences alignées (alignement multiple sans brèche)
- Pour une paire d'a.a. :  $\log(\text{fréquence observée} / \text{fréquence attendue})$

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)



# Alignement de deux séquences protéiques : Matrices de substitution

## 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

Small hydrophylic

Acid, acid amide and hydrophilic

Basic

Small hydrophobic

Aromatic

# 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)

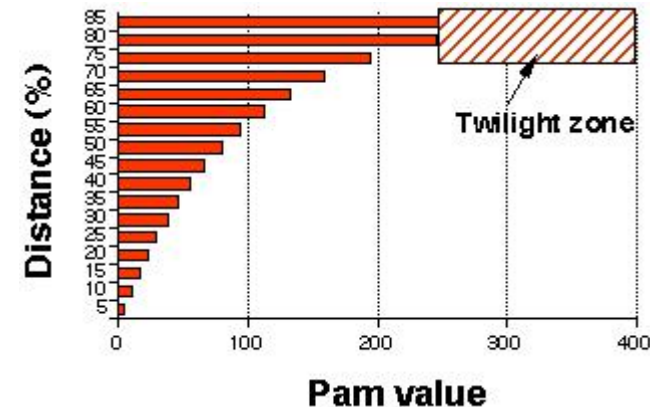
PAM160 et BLOSUM62 : estimation des fréquences de substitution entre acides aminés pour des séquences ayant des distances évolutives intermédiaires.



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

distance %	PAM
1	1
25	30
50	80
80	246



Recommandations (à adapter)

Source figure : Infobiogen.fr



## 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...



## 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

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.<sup>[1][2]</sup>

### Contents [\[hide\]](#)

- Function
- Domain architecture
- Proteins containing a DnaJ domain
- References

## 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.<sup>[3]</sup> 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

<b>Symbol</b>	DnaJ
<b>Pfam</b>	PF00226 <a href="#">↗</a>



## Seed sequence alignment for PF00226

```

Y6693_DROME/15-79      DVYKLMELARG...A.G.....EKE..VKKAYHKLSSLVHPD..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..AAAE...KFK.IIQHAYEV
JJJ1_YEAST/4-67       CYYELLGVETH...A.S.....DLE..LKKAYRKKALQYHPD..KN.....PDNV...EETQ....KFA.VIRAAYEV
O62360_CAEEL/3-66     CHYEVLEVERD...A.D.....DDK..IKKNYRKLALKWHPD..KN.....PDRI...EECTQ...QFR.LLQAAYDV
Q9W0X8_DROME/3-66     CYYEELQLRN...A.N.....DGD..IKSAYRKMALRWHPD..KN.....PDRL...AEAKE...RFQ.LIQQAYEV
DJP1_YEAST/6-68       EYYDLLGVSTT...A.S.....SIE..IKKAYRKKSIQHPD..KN.....PNDP....TATE...RFQ.AISEAYQV
CAJ1_YEAST/6-68       EYYDILGIKPE...A.T.....PTE..IKKAYRRKAMETHPD..KH....PDDP...DAQA...KFK.AVGEAYQV
Q9XEM8_ORYSI/285-347  AYYDTLGVSV...A.S.....PAE..IKKAYYLKAKQVHPD..KN.....PGNP....DAAQ...KFK.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.....QGD...LAAEK...Q.VLGEAYQV
DNJ10_ARATH/6-68      EYYDVLGVSP...A.T.....ESE..IKKAYYIKARQVHPD..KN.....PNDP....QAAH...NFQ.VLGEAYQV
YAY1_SCHPO/8-71       EYYDLLGISTD...A.T.....AVD..IKKAYRKLAVKYHPD..KN.....PDDP...QBASE...KFK.KISEAYQV
YHXB_SCHPO/9-72       DYYDILNISVD...A.D.....GDT..IKKSYRRLAILYHPD..KN....RENP...EAARE...KFK.KLAEAYQV
XDJ1_SCHPO/6-70       KLYDILEVHFE...A.S.....AEE..IKKSYRRLALLHHPD..KA.....PIHEK..EEAAE...RFR.GVQAYDI
Q9VGR0_DROME/10-72   NLYDLLGISLE...S.D.....QNE..IRKAYRKLALQHPD..KN.....PDNP...KAVE...RFH.ELSKALEI
CWC23_YEAST/15-84     NLYDVLELPTP...L.DVHTIYDDLQIKRKYRTLALKYHPD..KH.....PDNP...SIH...KFK.LLSTATNI
CWC23_YEAST/15-84 (SS) -HHHHH-TT-----HHHHHHHHHHHHHHHHH-TT-T-----HHH.HHHHHHHH
HLJ1_YEAST/21-82      EFYEILKVDK...A.T.....DSE..IKKAYRKLAIKHPD..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..IKKAYKLLALQLHPD..KN.....KAP...GAVE...AFK.ALGNAAGV
YNF5_SCHPO/113-174   QYYEILDLLKKT...C.T.....DTE..IKKSYKLLALQLHPD..KN.....HAP...SADE...AFK.MVSKAFQV
DJB12_MOUSE/111-172  DYYEILGVSR...A.S.....DED..LKKAYRKLALKFHPD..KN.....HAP...GATE...AFK.AIGTAYAV
O43978_BABB0/14-72   KFYKVLGLSRD...C.S.....ESE..IKKAYRKLAIKHPD..KG.....GDP...GDP...MFK.EITRAYEV
O24874_MEDSA/14-72   KYDILGVSKS...A.S.....EDE..IKKAYRKAAMKNHPD..KG.....GDP...GDP...E...KFK.ELGQAYEV
DNJH2_ALLPO/13-71    KYEVLGVSKN...A.T.....PED..LKKAYRKAAMKNHPD..KG.....GDP...GDP...E...KFK.EIGQAYEV
O96455_DICDI/5-66    KFYDILGVARD...A.S.....ETD..IKKAYRKLAIKYHPD..KN.....PDP...AAVE...KFK.ELTVAYEV
Q9U062_GIAIN/6-65    EFYDILGVSPS...A.D.....PQT..IKKRTTKLARKYHPD..KP.....TDEE...TDEE...LFN.KIGRAYEV
  
```



## 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...



## Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

**There are 26667 sequences with the following architecture: DnaJ**

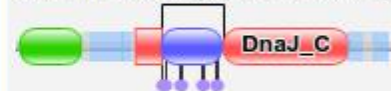
[X8FHB8\\_MYCUL](#) [Mycobacterium ulcerans str. Harvey] DnaJ domain protein {ECO:0000313|EMBL:EUA92552.1} (108 residues)



[Show](#) all sequences with this architecture.

**There are 12636 sequences with the following architecture: DnaJ, DnaJ\_C, DnaJ\_CXXCXGXG**

[Z5XPE1\\_9GAMM](#) [Pseudoalteromonas lipolytica SCSIO 04301] Chaperone protein DnaJ {ECO:0000256|HAMAP-Rule:MF\_01152} (378 residues)



[Show](#) all sequences with this architecture.

**There are 7578 sequences with the following architecture: DnaJ, DnaJ\_C**

[X7EK98\\_9RHOB](#) [Roseivivax halodurans JCM 10272] Molecular chaperone DnaJ {ECO:0000313|EMBL:ETX15596.1} (302 residues)



[Show](#) all sequences with this architecture.

**There are 1693 sequences with the following architecture: DnaJ, DUF1977**

- Pfam : modèle mathématique (HMM)



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT



## 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

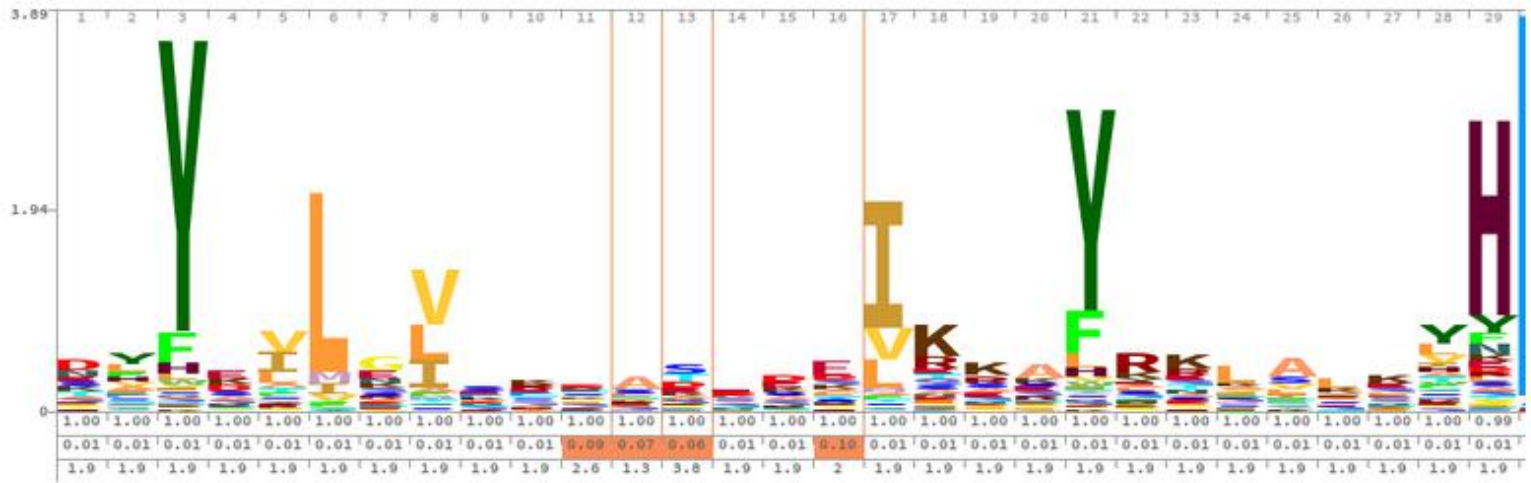
Jump to... ↓

enter ID/accession

Go

### HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them [here](#). [More...](#)



Pfam is part of the

- 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



## Search InterPro

by sequence

by text

by domain architecture


### Sequence, in FASTA format

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool. Enter or paste a protein sequence in FASTA format (complete or not - e.g. `PMPIGSKERPTFFEIFKTRCNKADLGPISLN`), with a maximum length of 40,000 amino acids.

Please note that you can only scan one sequence at a time. Alternatively, read [more about InterProScan](#) for other ways of running sequences through InterProScan.

```
>OsProt
MERNKFASKMSQHYTKTICIAVVLVAVLFSLSAAAAGSGAAVSVQLEALLEFKNGVADD
PLGVLAGWRVKGSGDGA VRGGALPRHCNWTGVACDGAGQVTSIQLPESKLRGALSPFLGN
ISTLQVIDLTSNAFAGGI PPQLGRLGELEQLVVSSNYFAGGIPSSLCNCSAMWALALNVN
NLTGAIPSCIGDLSNLEIFEAYLNNLDGELPPSMAKLGIMVVDLSCNQLSGSIPPEIGD
```

## InterProScan résultats (1)

 **InterPro** Classification of protein families 🔍 ☰


Home ▶ Search ▶ Browse ▶ **Results** Release notes Download ▶ Help ▶ About


🏠 / Result / InterProScan / Iprscan5-R20210301-105429-0250-80615314-p1m / Overview

**Overview** Entries 7 Sequence



### InterProScan Search Result

📘

**Title** OsProt 

**Job ID** iprscan5-R20210301-105429-0250-80615314-p1m 

**Length** 1183 amino acids

**Action**  

**Status** ✓ finished

**Expires** 📘 Mon Mar 08 2021

### Protein family membership

None predicted


**Entry matches to this protein** 📏 + - Options Export

1 1183

100 200 300 400 500 600 700 800 900 1,000 1,100

500 1000

▼ Domain



**IPR000719**  
SM00220  
PF00069  
PS50011

**IPR013210**  
PF08263



## InterProScan résultats (2)

InterPro Classification of protein families

Home Search Browse Results Release notes Download Help About

Entry matches to this protein ⓘ [ ] + - Options Export

1 100 200 300 400 500 600 700 800 900 1,000 1,100 1183

500 1000

▼ Domain

▼ Homologous Superfamily

▼ Repeat

▼ Active Site

▼ Unintegrated

IPR000719: Prot\_kinase\_dom  
SM00220: serkin\_6  
PF00069: Pkinase  
PS50011: PROTEIN\_KINASE\_DOM  
IPR013210: LRR\_N\_plant-typ  
PF08263: LRRNT\_2

IPR011009: Kinase-like\_dom\_sf  
SSF56112: Protein kinase-like (PK-like)  
IPR032675: LRR\_dom\_sf  
G3DSA:3.80.10.10: Ribonuclease Inhibitor

IPR001611: Leu-rich\_rpt  
PF13855: LRR\_8  
IPR003591: Leu-rich\_rpt\_typical-subtyp  
SM00369: LRR\_typ\_2

IPR008271: Ser/Thr\_kinase\_AS  
PS00108: PROTEIN\_KINASE\_ST

G3DSA:1.10.510.10: Transferase(Phosphotrans  
SSF52047: RNI-like  
SSF52058: L domain-like

PF00069  
Pkinase  
Pfam  
877 - 1167