



Faculté des Sciences – Rabat



Laboratoire de Microbiologie et Biologie Moléculaire

Université Mohamed V – Agdal • Faculté des Sciences B.P. 1014 –
Rabat – MAROC



Institut de recherche
pour le développement

TD Biologie Moléculaire SVI – S6

Introduction à la bioinformatique

Dr. Laila Sbabou

Mai 2010

Introduction

Qu'est ce que la bioinformatique?

Historique de la Bioinformatique

Acquisition, organisation et stockage des données

Banques de séquences nucléiques

Banques de séquences protéiques

Autres types de banques de données

Analyse de données:

Recherche de similarités dans les bases de données

Recherche des phases de lectures ouvertes (ORF)

Introduction

Qu'est ce que la bioinformatique?

Une approche *in silico* de la biologie qui consiste en une analyse informatisée des données biologiques en utilisant un ensemble de moyens (concepts, méthodes, implémentations, logiciels, etc...).

Discipline complémentaire aux approches classiques de la biologie:

In vivo (tests au sein des organismes vivants)

In situ (tests dans les milieux naturels)

In vitro (tests dans des tubes)

Introduction

Historique de la Bioinformatique

1965: Première compilation de protéines (M. Dayhoff *et al.*) : 50 entrées

1967: Article : "*Construction of Phylogenetic Trees*" (Fitch & Margoliash)

1970: Programme d'alignement global de séquences (algorithme de Needleman & Wunsch).

1972: Premier microprocesseur Intel 8008

1973: Génie Génétique (Cohen *et al.*)

1977: Micro-ordinateurs

Séquençage d'ADN : F. Sanger / Maxam & Gilbert

Première suite logicielle bioinformatique (*Staden*)

1980: Premières méthodes de prédiction et d'alignement.

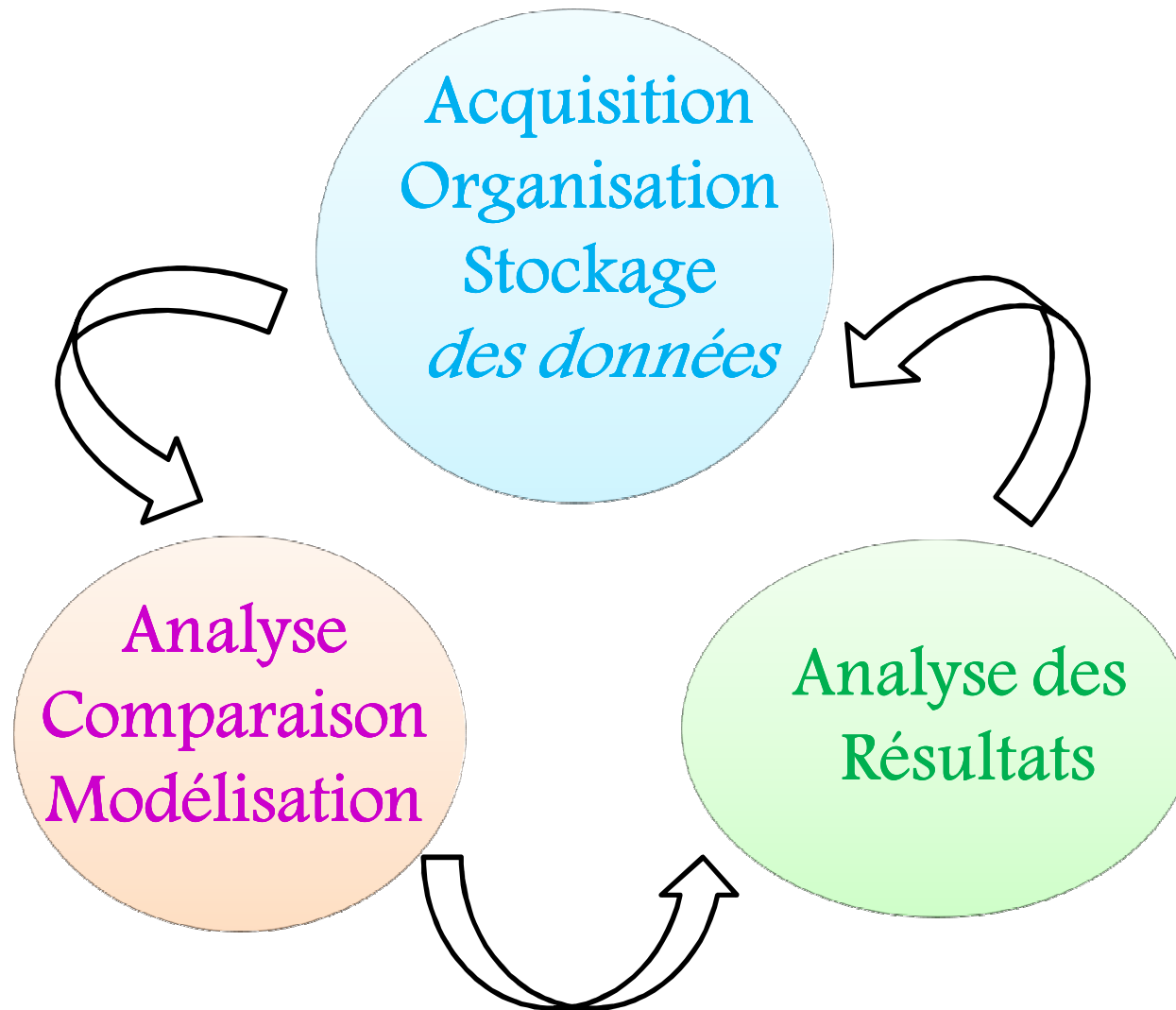
Première bases de données *EMBL*, *GenBank* et *PIR*

1981: *GenBank* : 270 séquences

Programme d'alignement local de séquences (Smith & Waterman)

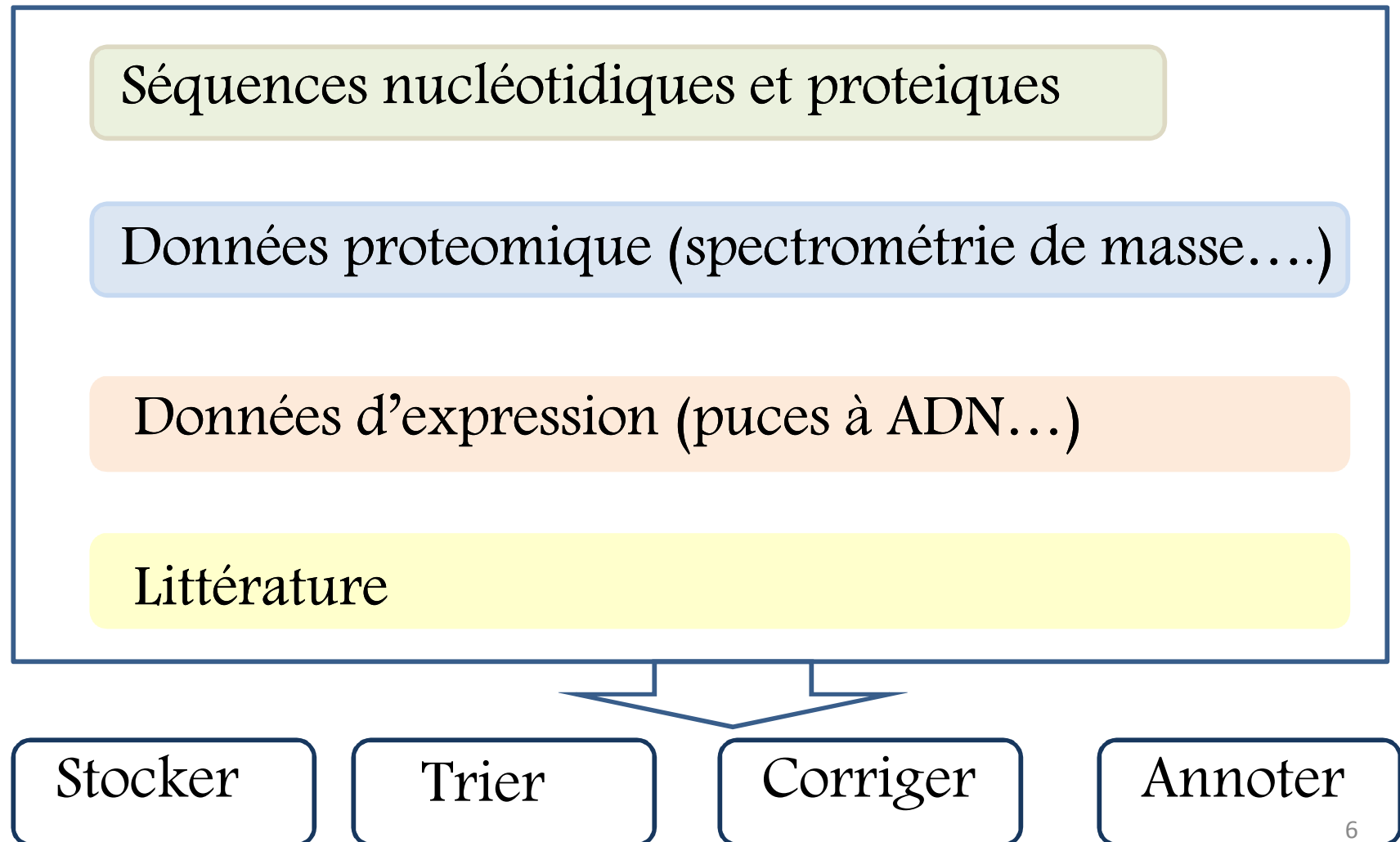
1985: Programme d'alignement local de séquences "FASTA" (Pearson & Lipman)

1990: Programme d'alignement local de séquences "BLAST" (Altschul *et al.*)



Acquisition, organisation et stockage des données:

Types de données:



Acquisition, organisation et stockage des données:

Banques de séquences nucléiques:

Divers Banques de Données:

GenBank:

La banque américaine maintenue au NCBI (National Center for Biotechnology Information) à Bethesda (USA)

EMBL :

La banque européenne maintenue à l'EBI (European Bioinformatics Institut) à Cambridge (UK)

DDBJ :

La banque japonaise (DNA Data Bank of Japan)

Acquisition, organisation et stockage des données:

Banques de séquences proteiques:

Swiss Prot:

- Niveau élevé d'Annotation (manuelle):
- Description de la fonction des proteines, structure des domaines et modification post-traductionnelle.... etc

TrEMBL :

Données générées par traduction automatique des informations génétiques de la banque de données EMBL (d'où TrEMBL =Traduction EMBL)

Annotation automatique

Prosite:

Base de données de familles et domaines de protéines

GenePept:

Traduction automatique des CDS de GenBank

PIR (Protein Information Resource):

Groupe établi par le National Biomedical Research Foundation (NBRF)

Identification et interpretation de l'information des séquences proteiques

Expazy:

Base de données protéomique

En 2002: consortium UniProt (Universal Protein Resource) formé par le groupe SwissProt-TrEMBL et le groupe PIR

Acquisition, organisation et stockage des données:

Autres types de banques de données:

Banques de Structure:

Ex: la Protein Database PDB dédiée aux structures protéiques déterminées expérimentalement

Banques dédiées à un organisme particulier:

Ex: *Arabidopsis thaliana* (TAIR, ABRC....)

Colibri (*E. coli*)

Subtilis (*Bacillus subtilis*)

Flybase (Drosophile)

Banques dédiées à un type de séquences particulier:

- Analyse des promoteurs (PromScan, Virtual Footprint, Prokaryotic Promoter Prediction 'PPP', Scope...)
- Analyse des terminateurs de la transcription (FindTerm, RibEx....)

Analyse des données:

Par comparaison de séquences afin de trouver des similarités (Voir si ma séquence ressemble à d'autres déjà connues), définir la structure et Identifier les domaines et motifs connus.

Acides nucléiques:

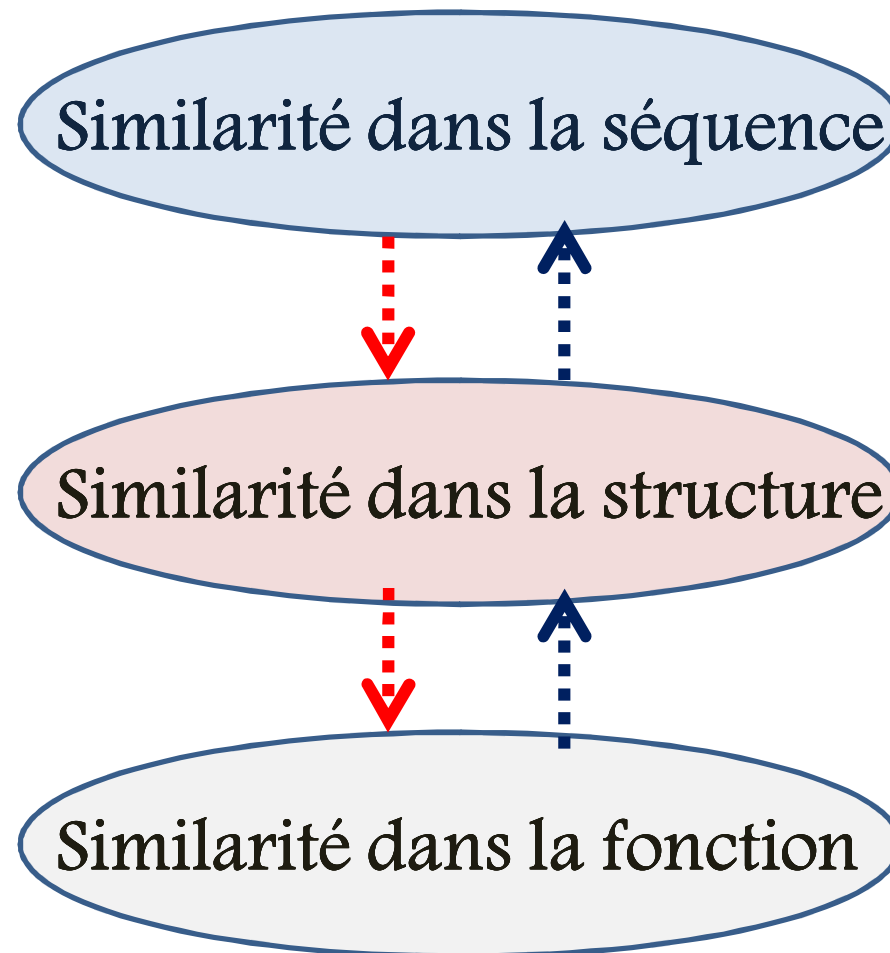
- Recherche de phase de lecture ouverte (ORF) d'un gène
- Dédurre la structure Intron/Exon d'un gène
- Etablir l'arbre phylogénique
- Recherche d'Etiquettes EST et du profil d'expression des gènes (profiling)
- Analyse de génomes entiers

Protéines:

- Dédurre la séquence de la protéine à partir d'une séquence d'ADN (traduction in silico)
- Identifier la famille, sites actifs et domaines fonctionnelles
- prédiction des modifications post-traductionnelles et structures secondaires
- Etablir un arbre Phylogénique

Analyse des données:

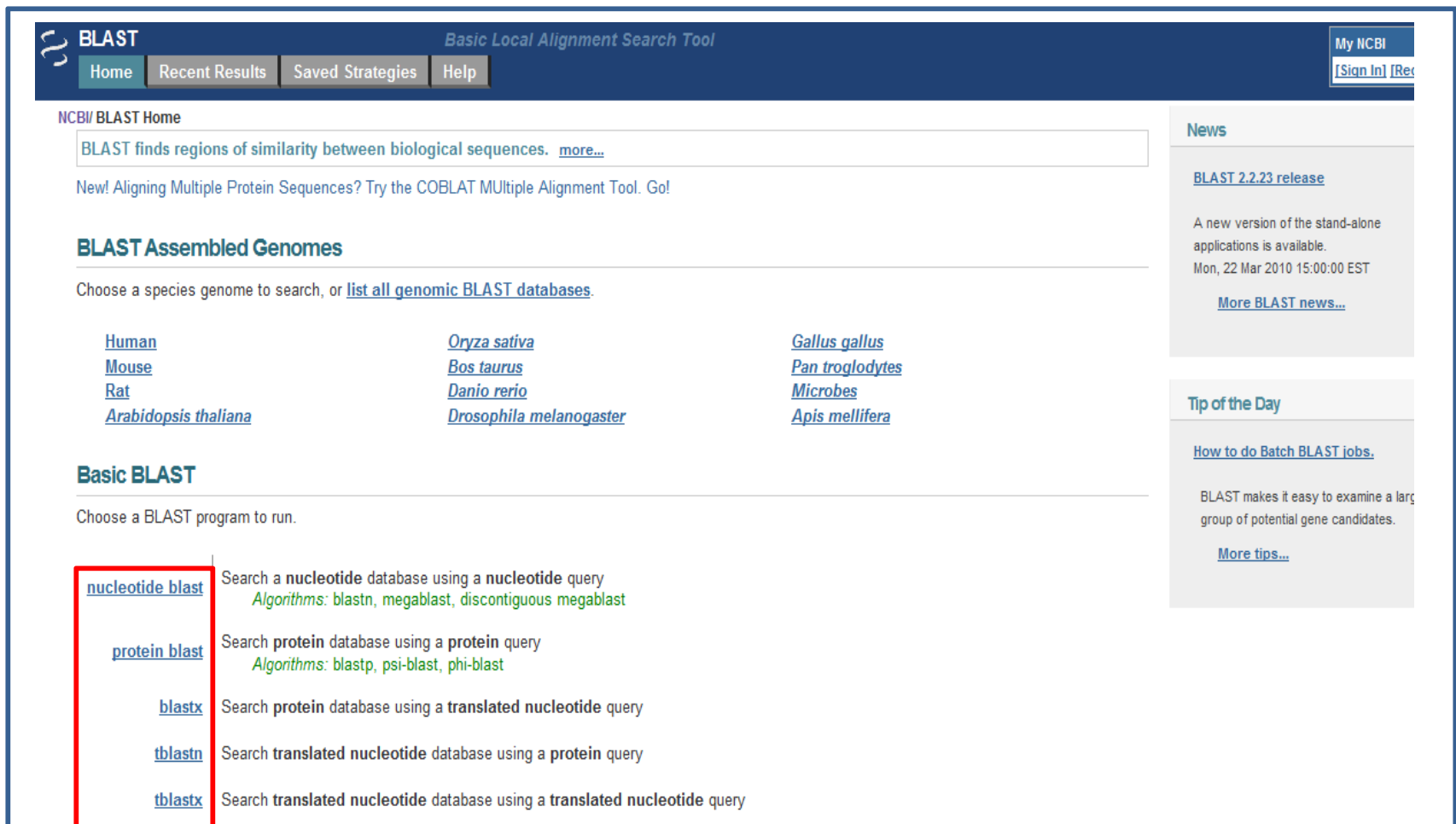
Analyser ma séquence pour ?



Analyse des données:

Recherche de similarités dans les banques de données:

Le Blast: Basic Local Alignment Search Tool (Ex. Blast sur NCBI)



The screenshot shows the NCBI BLAST homepage. At the top, there's a navigation bar with 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. Below this, the main content area is divided into sections. The 'BLAST Assembled Genomes' section lists various species for genome search, including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. The 'Basic BLAST' section provides a list of search options, with 'nucleotide blast' highlighted by a red box. The right sidebar contains 'News' and 'Tip of the Day' sections.

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New! Aligning Multiple Protein Sequences? Try the COBLAT Multiple Alignment Tool. Go!

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

Human	Oryza sativa	Gallus gallus
Mouse	Bos taurus	Pan troglodytes
Rat	Danio rerio	Microbes
Arabidopsis thaliana	Drosophila melanogaster	Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

News

[BLAST 2.2.23 release](#)

A new version of the stand-alone applications is available.
Mon, 22 Mar 2010 15:00:00 EST

[More BLAST news...](#)

Tip of the Day

[How to do Batch BLAST jobs.](#)

BLAST makes it easy to examine a large group of potential gene candidates.

[More tips...](#)

Analyse des données:

Programme	Séquence requête	Base de données
Blast N	Nucléique	Nucléique
Blast P	Protéique	Protéique
Blast X	Nucléique	Protéique
t blast N	Protéique	Nucléique
t blast X	Nucléique	Nucléique

Résultat du Blast

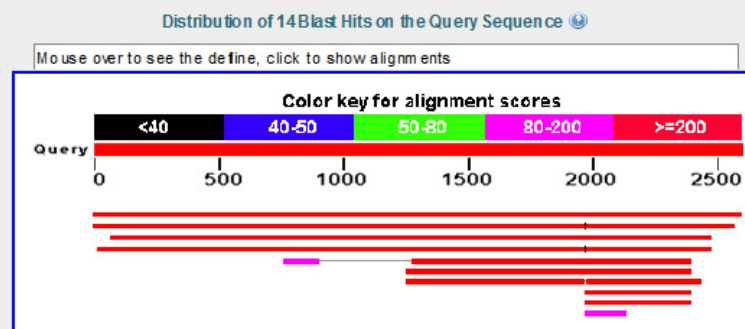
[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [Help](#)

gi|229615779|gb|FJ236985.1| Lupinus albus...

Query ID	Id 41729	Database Name	nr
Description	gi 229615779 gb FJ236985.1 Lupinus albus scarecrow 1 (SCR1) mRNA, complete cds	Description	All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
Molecule type	nucleic acid	Program	BLASTN 2.2.22+ Citation
Query Length	2581		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#)

▼ Graphic Summary



▼ Descriptions

Legend for links to other resources: [Un](#) [E](#) [G](#) [S](#) [M](#)

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
FJ236985.1	Lupinus albus scarecrow 1 (SCR1) mRNA, complete cds	4767	4767	100%	0.0	100%	
FJ236987.1	Lupinus albus scarecrow 1 (SCR1) gene, complete cds	3616	4729	99%	0.0	100%	
FJ236986.1	Lupinus albus scarecrow 2 (SCR2) mRNA, complete cds	2819	2819	92%	0.0	88%	
FJ236988.1	Lupinus albus scarecrow 2 (SCR2) gene, complete cds	2185	2864	94%	0.0	91%	
AB048713.1	Pisum sativum PsSCR mRNA for SCARECROW, complete cds	883	1054	48%	0.0	88%	
XM_002323076.1	Populus trichocarpa GRAS family transcription factor (GRAS1), mRNA	695	695	43%	0.0	78%	G
AP010542.1	Lotus japonicus genomic DNA, chromosome 1, clone: LjB20B09, BM1478 complete sequence	582	1060	45%	3e-162	86%	
AC124215.21	Medicago truncatula clone mth2-32m6, complete sequence	436	436	16%	2e-118	85%	
AC155890.2	Medicago truncatula chromosome 7 clone mth2-49p3, complete sequence	436	436	16%	2e-118	85%	
AB048714.1	Pisum sativum PsSCR gene for SCARECROW, partial cds	189	189	6%	7e-44	86%	

Analyse des données:

Recherche des phases de lectures ouvertes (ORF) (Utiliser l'ORF finder de ncbi)

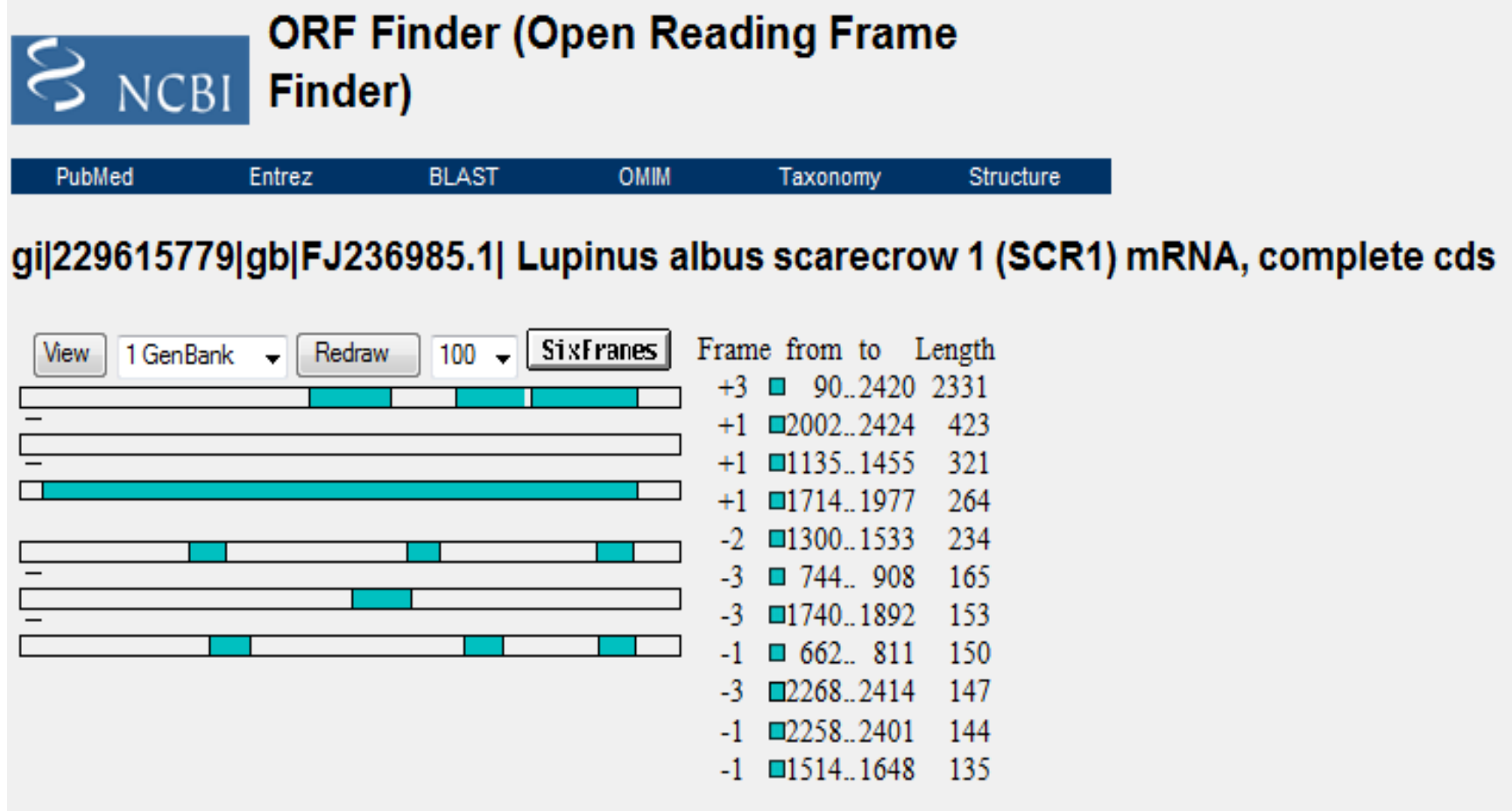
NCBI / Resources / Sequence analysis / Tools / ORF finder

The screenshot shows the NCBI ORF Finder web interface. The top navigation bar includes links to PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The left sidebar contains links to NCBI, Tools for data mining, GenBank sequence submission support and software, and FTP site download data and software. The main content area is titled "ORF Finder (Open Reading Frame Finder)" and contains a description of the tool. Below the description, there is a form with the following fields:

- Enter GI or ACCESSION**: A text input field.
- or sequence in FASTA format**: A text area containing a FASTA sequence for *Lupinus albus* scarecrow 1 (SCR1) mRNA.
- FROM:**: Two empty text input fields.
- Genetic codes**: A dropdown menu set to "1 Standard".

The "OrFind" button is highlighted with a blue border.

Analyse des données:



A utiliser la séquence de cette protéine dans
une nouvelle recherche

BlastP et **tBlastN**

Analyse des données:

Comparaison des séquences par alignement

Utiliser ClustalW (<http://www.ebi.ac.uk/Tools/clustalw2/index.html>)

The screenshot displays the EBI ClustalW2 web interface. The top navigation bar includes links for Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. A search bar is present with a 'Go' button. The left sidebar contains a 'Help Index' and 'Similar Applications' like Align, Kalign, MAFFT, MUSCLE, and T-Coffee. The main content area is titled 'ClustalW2' and provides a brief description of the tool. Below this is a large configuration panel with several sections of dropdown menus:

- YOUR EMAIL:** A text input field.
- ALIGNMENT TITLE:** A dropdown menu with 'Sequence' selected.
- RESULTS:** A dropdown menu with 'interactive' selected.
- ALIGNMENT:** A dropdown menu with 'full' selected.
- SCORE TYPE:** A dropdown menu with 'percent' selected.
- NO END GAPS:** A dropdown menu with 'yes' selected.
- TOPDIAG:** A dropdown menu with 'def' selected.
- PAIRGAP:** A dropdown menu with 'def' selected.
- GAP EXTENSION:** A dropdown menu with 'def' selected.
- GAP DISTANCES:** A dropdown menu with 'def' selected.
- ITERATION:** A dropdown menu with 'none' selected.
- NUMBER:** A dropdown menu with '1' selected.
- OUTPUT FORMAT:** A dropdown menu with 'aln w/numbers' selected.
- OUTPUT ORDER:** A dropdown menu with 'aligned' selected.
- TREE TYPE:** A dropdown menu with 'none' selected.
- PHYLOGENETIC TREE:** A dropdown menu with 'ofl' selected.
- CORRECT DIST.:** A dropdown menu with 'ofl' selected.
- IGNORE GAPS:** A dropdown menu with 'ofl' selected.
- CLUSTERING:** A dropdown menu with 'NJ' selected.

Below the configuration panel is a text area for entering sequences, with a 'Help' button. The text area contains the following sequence data:

```
>gi|229615779|gb|FJ236985.1| Lupinus albus scarecrow 1 (SCR1)
mRNA, complete cds
ATTCAAAACAATTITACGTATCACTATCAACTATCACTCCCTATCTATCTCACTCTTGGCTTCTCCAT
AA
ATTGGATGCTTCTGATGCTATGCTCAGCGGTGCCAGCAATAATAACTGAAAGATGGCAATAACAATA
AT
GGTGGTAGTTCCTTTGACTAGTGCCTCTAATAACTCTAGTAAATTTAGCAGTGAAGAGCACACTCACAG
TC
AGAAAGCTACAGCATTCCCATTCGGGACAAAGGAAAAATGGTGAGAAAGAGGATGGCTTCTGAGATGGAA
CC
```

At the bottom, there is an 'Upload a file:' section with a 'Browse...' button, and 'Run' and 'Reset' buttons.

Exemple d'un Multi-alignement

Séquence 1 →	LaSCR1	1	MLSGASNNNT	EDGNNNN---	GGSPLOTSASN	NSSKFSSEEH	THSQKLQSH	SGQRKMVRKR	57
Séquence 2 →	LaSCR2	1	****G*****	***S***NIN	*****	***NI*****	**T*Q*****	*GK*****	60
Séquence 3 →	AtSCR	1	-----MA	*S*DFNG---	*QP*PH*PLR	TT*SG**S--	SNNRGPPPPP	PPPLV*****	47
	LaSCR1	58	MASEMEPTHT	MLPHGAVVGG	RFPRR--CND	DNMSLLLNCS	SLPPAVEKTT	SFNNNDYHYK	115
	LaSCR2	61	*****	*****IDS-	*****SSS*N	N***SV*I**	*****	*****-Y*****	118
	AtSCR	48	L****S----	-----	-----	-----	-----	--SNPD*NNS	61
	LaSCR1	116	TSSSSKVDNN	VAVVPNPTTP	NYSTM LLPSS	SCSTTINPNY	HISQRQQQDQ	NQLTSPAVCG	175
	LaSCR2	119	*****R****	AH*****	****L*****	-----	N*I*T*****	*****	176
	AtSCR	62	SRPFR*SHL	LDSNY*TV**	QQPPS*TAAA	TV*SQPN*--	-----	----PLS***	105
	LaSCR1	176	FSGLPLFPAS	QQRNHHHNSS	SSSSTGANVE	VAASPSMEDN	NNN--SAATA	WIDGILKDLI	233
	LaSCR2	177	*****	*****--	-*T****T**	*****	***NN****D	*****	233
	AtSCR	106	*****V**SD	RG-----	-----*R**M	MSVQ*-*DQD	SSSSSASPTV	*V*A*IR***	151
	LaSCR1	234	HSSNSVSIPQ	LINNVREIIY	PCNP NLAVVL	EYRLRLLTSH	DNTSAAPNND	-----SPNSS	288
	LaSCR2	234	*****	**S*****	*****	*****	****T****N	NNNNN*S*PA	293
	AtSCR	152	***T*****	**Q***D**F	*****GAL*	*****S*MLL	DPS*SS----	-----D*SP-	201
	LaSCR1	289	AVGKNNTTEV	GVVLNQNHPR	LPSTTTTVNV	IPDNFPDPDS	-GAAPLVMNQ	MLSNWVVLPI	347
	LaSCR2	294	S*R*****	*EG****Q**	***GAN--**	MH*****S*	S****V*****	*****G*****	351
	AtSCR	202	-----Q*F*P	LYQISN*PS*	PQQQQQHQQQ	QQQHK**P*P	-----	-----IQQ	239

* = Identité

- = Gap (introduits pour optimiser l'alignement entre les deux séquences)

