- 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)

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 seins des 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.

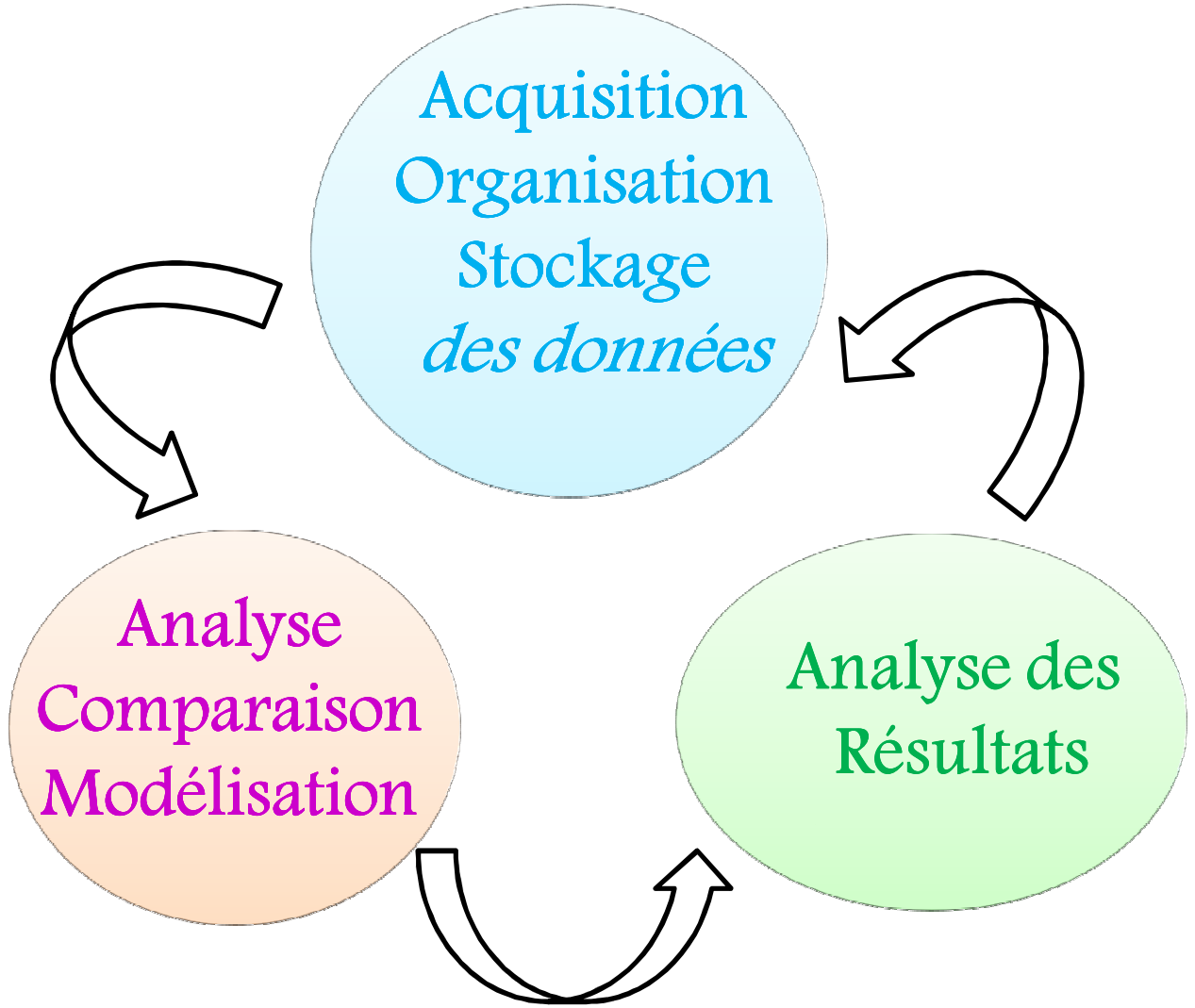
Première bases de données *EMBL*, *GenBank* et *PIR*

1981: GenBank : 270 séquences

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

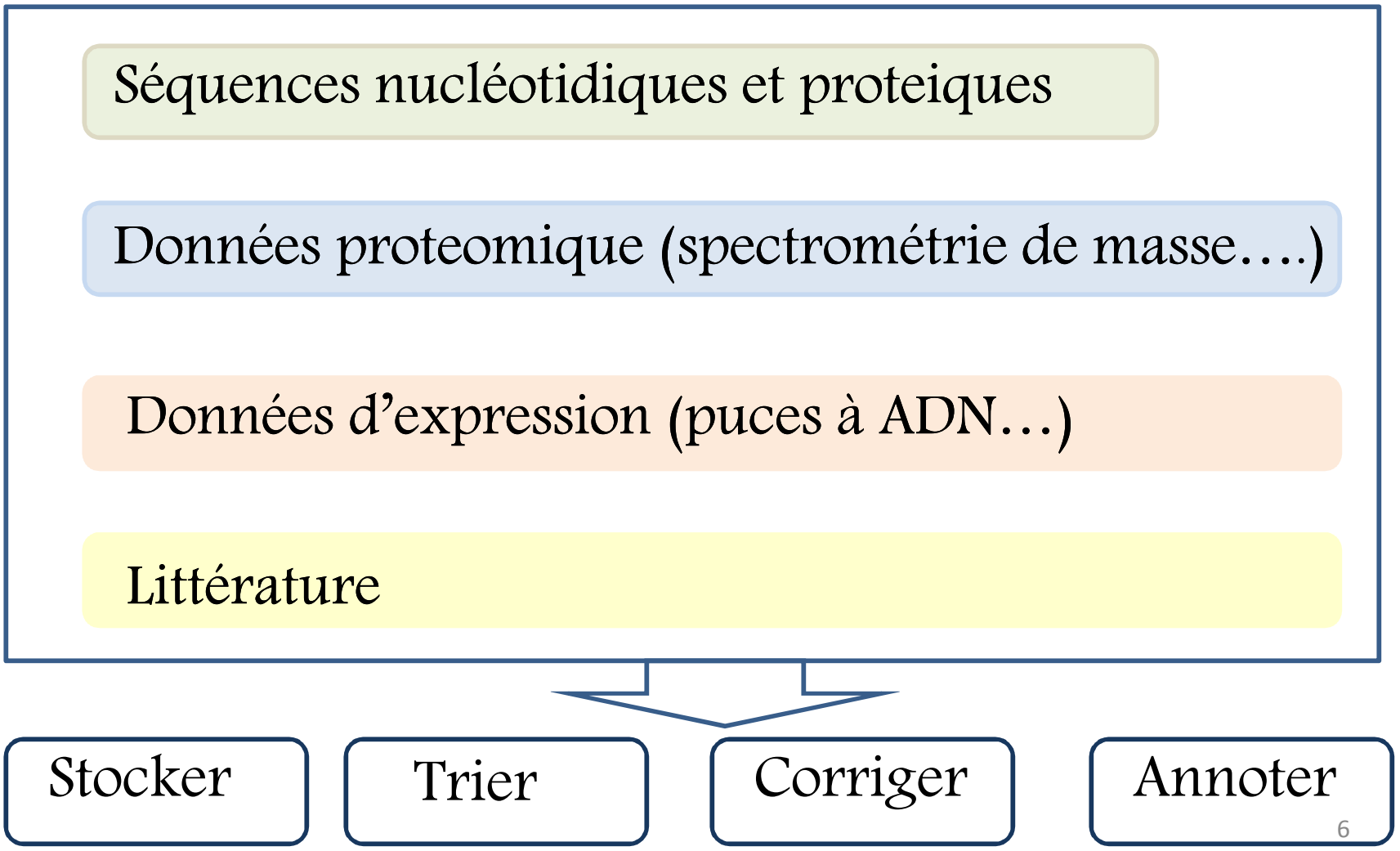
1985: Programme d'alignement local de séquences "FASTA" (Pearson & Lipman)

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


Acquisition
Organisation
Stockage
des données
Analyse des
Résultats
Analyse
Comparaison
Modélisation

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 Ressource):

Groupe établit 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 au structures proteiques détérminées experimentalement

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 transcriptio (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:

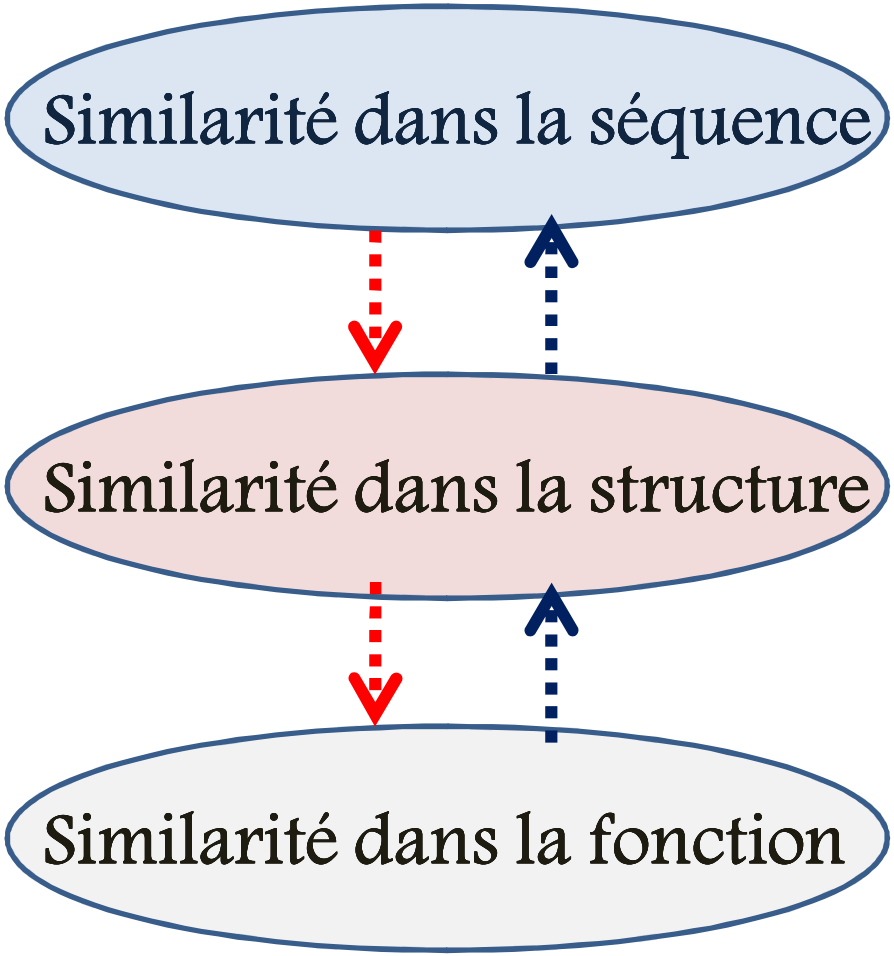
- Recherche de phase de lecture ouverte (ORF) d'un gène
- Déduire la structure Intron/Exon d'un gène
- Etablir l'arbre phylogènique
- Recherche d'Etiquettes EST et du profil d'expression des gènes (profiling)
- Analyse de génomes entiers

Protéines:

- Déduire 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:

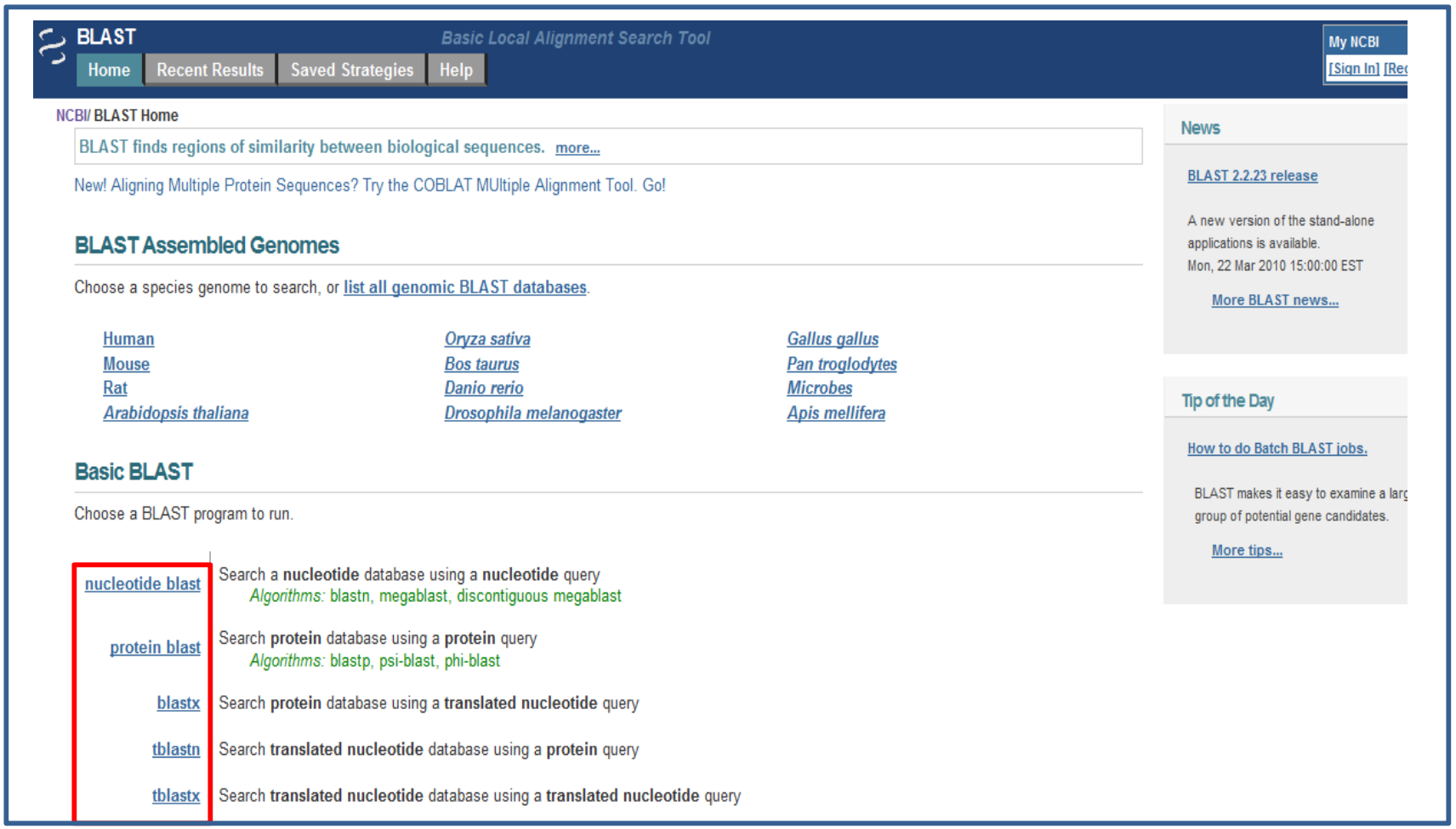
Analyser ma séquence pour ?



Analyse des données:

Recherche de similarités dans les banques de données:

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



Analyse des données:

Programme	Séquence requête	Base de données
BlastN	Nucléique	Nucléique
BlastP	Protéique	Protéique
BlastX	Nucléique	Protéique
tblastN	Protéique	Nucléique
tblastX	Nucléique	Nucléique

Résultat du Blast

Edit and Resubmit Save Search Strategies Formatting options Download

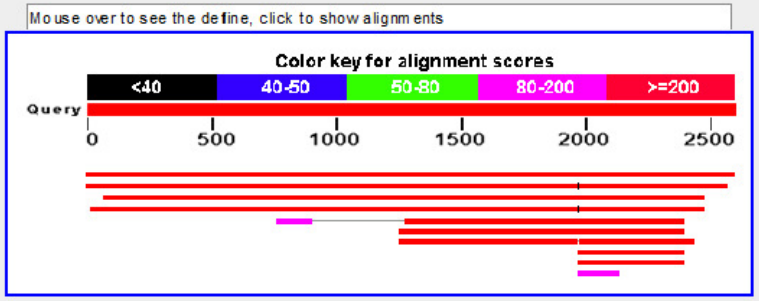
gi|229615779|gb|FJ236985.1| Lupinus albus...

Query ID	lcl|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

Distribution of 14 Blast Hits on the Query Sequence



Descriptions

Legend for links to other resources: UniGene GEO Profile Structure MapViewer M

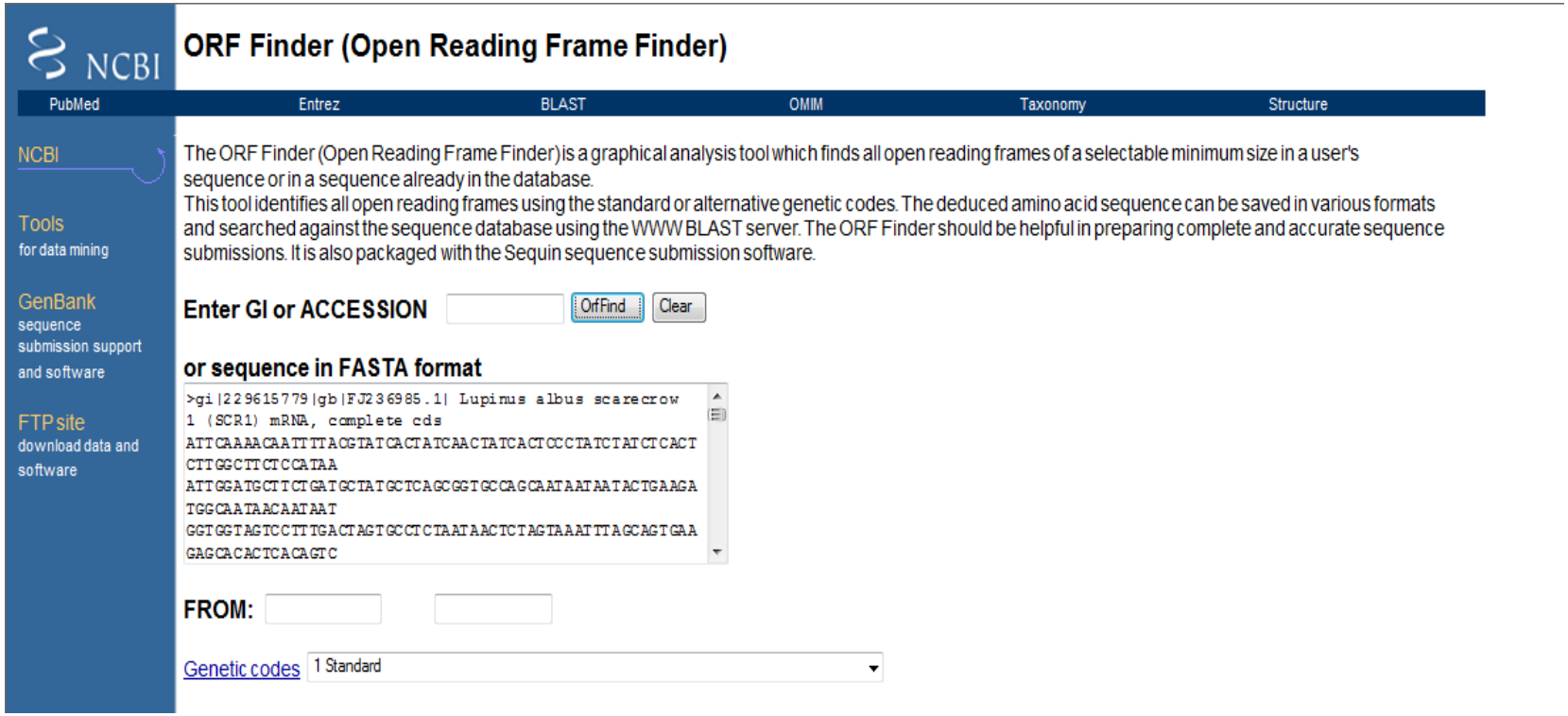
Sequences producing significant alignments:
(Click headers to sort columns)

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%	

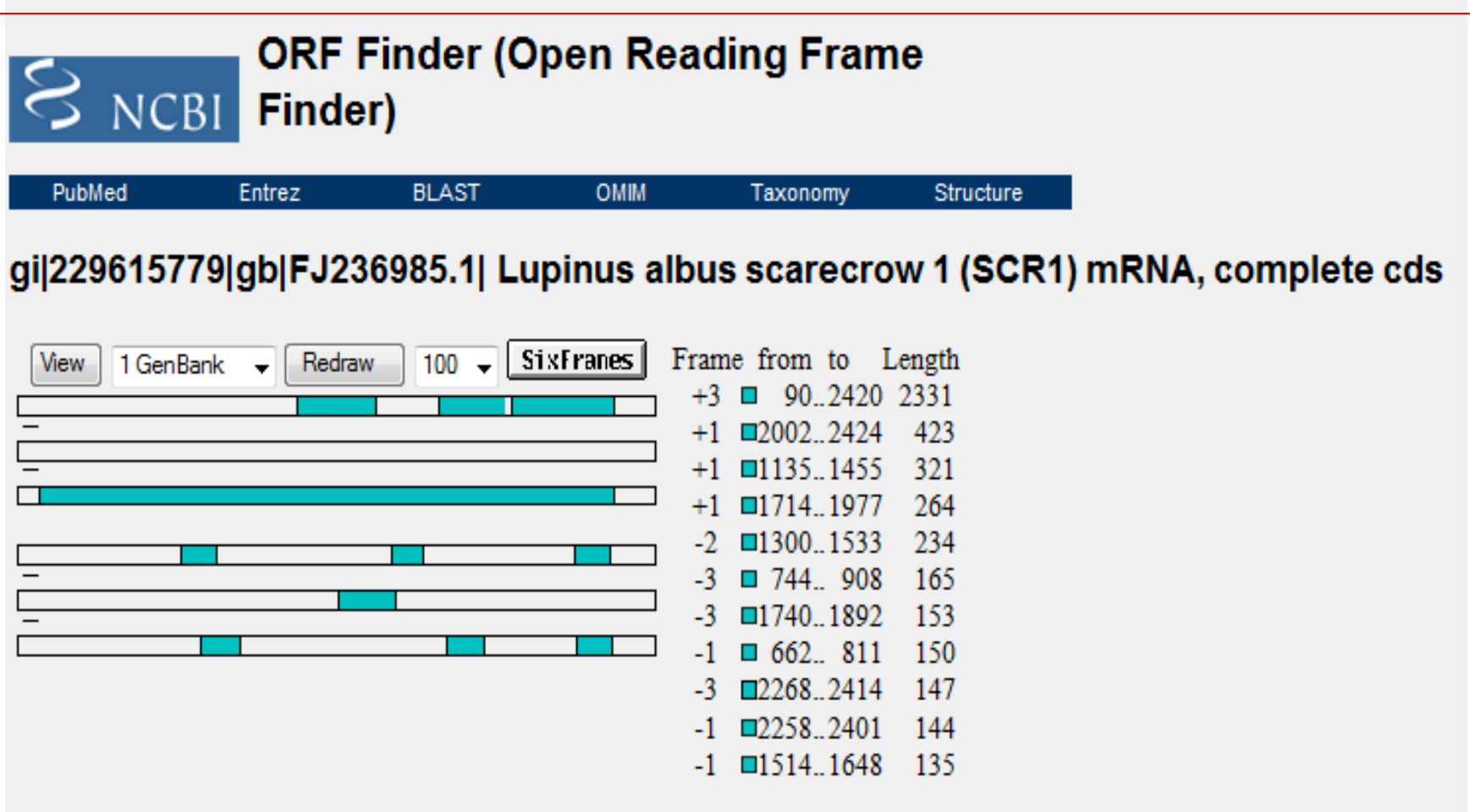
Analyse des données:

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

NCBI / Resources/ Sequence analysis / Tools/ ORF finder



Analyse des données:



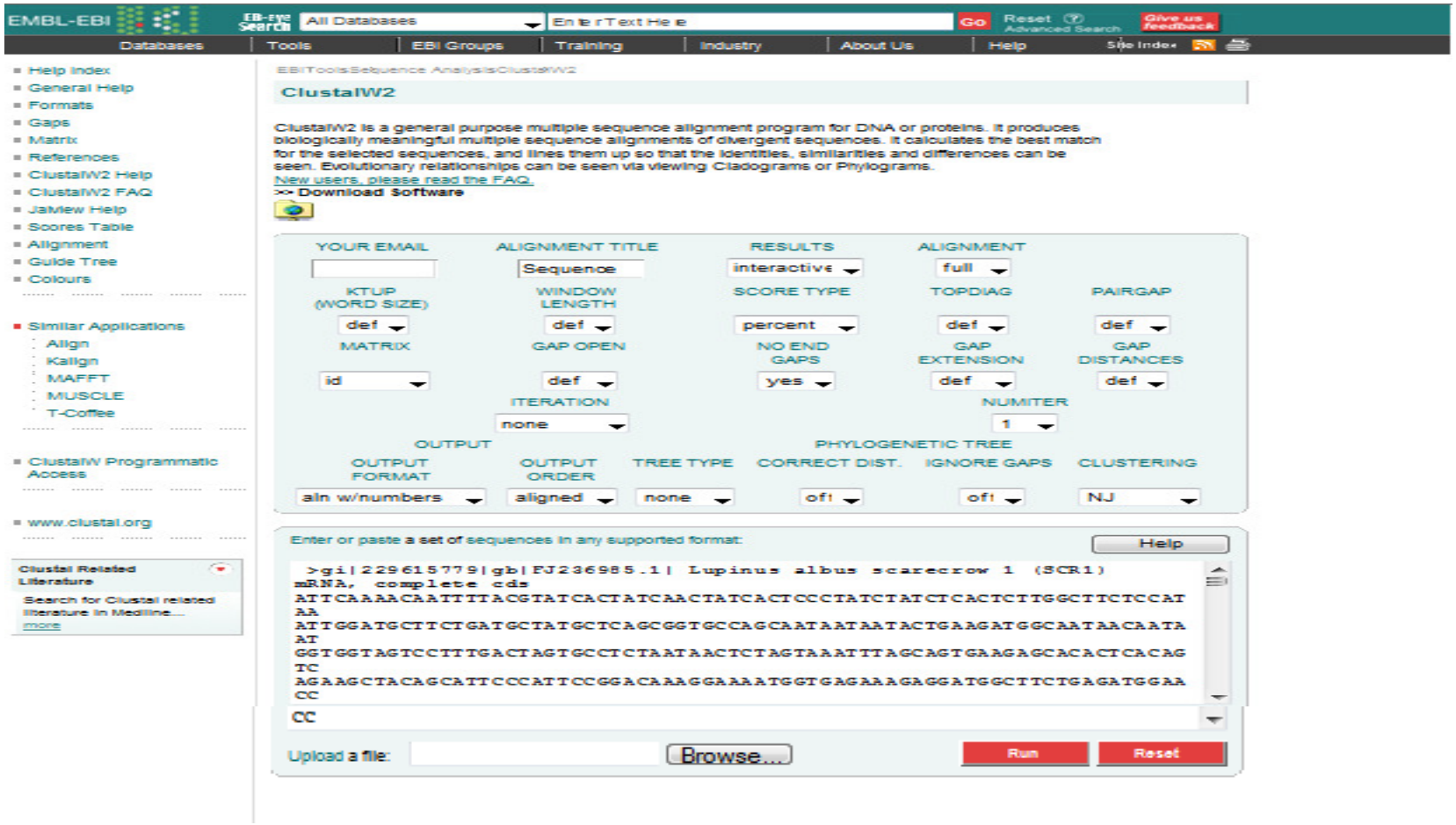
A utiliser la séquence de cette protéine dans une nouvelle recherche

BlastP et tBlastN

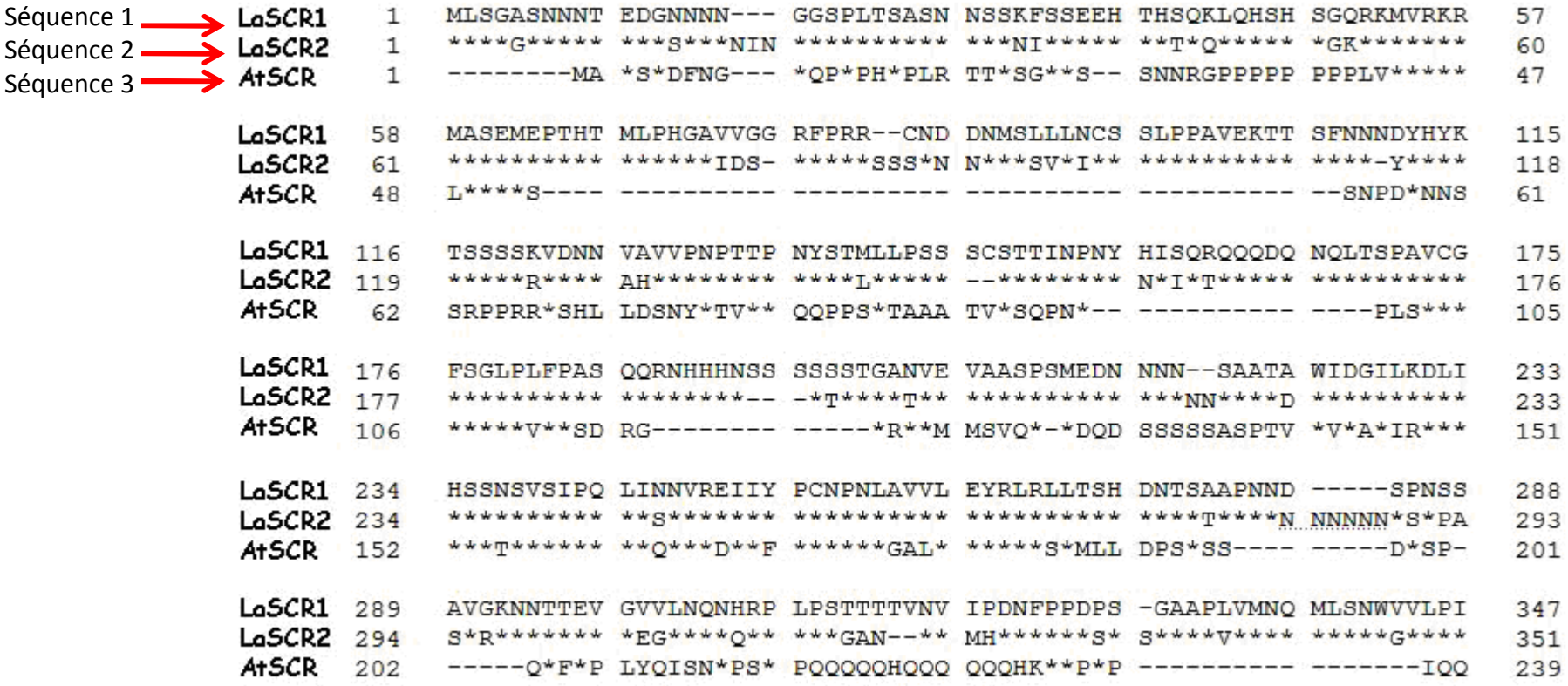
Analyse des données:

Comparaison des séquences par alignement

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



Exemple d`un Multi-alignement



* = Identité

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