

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 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.

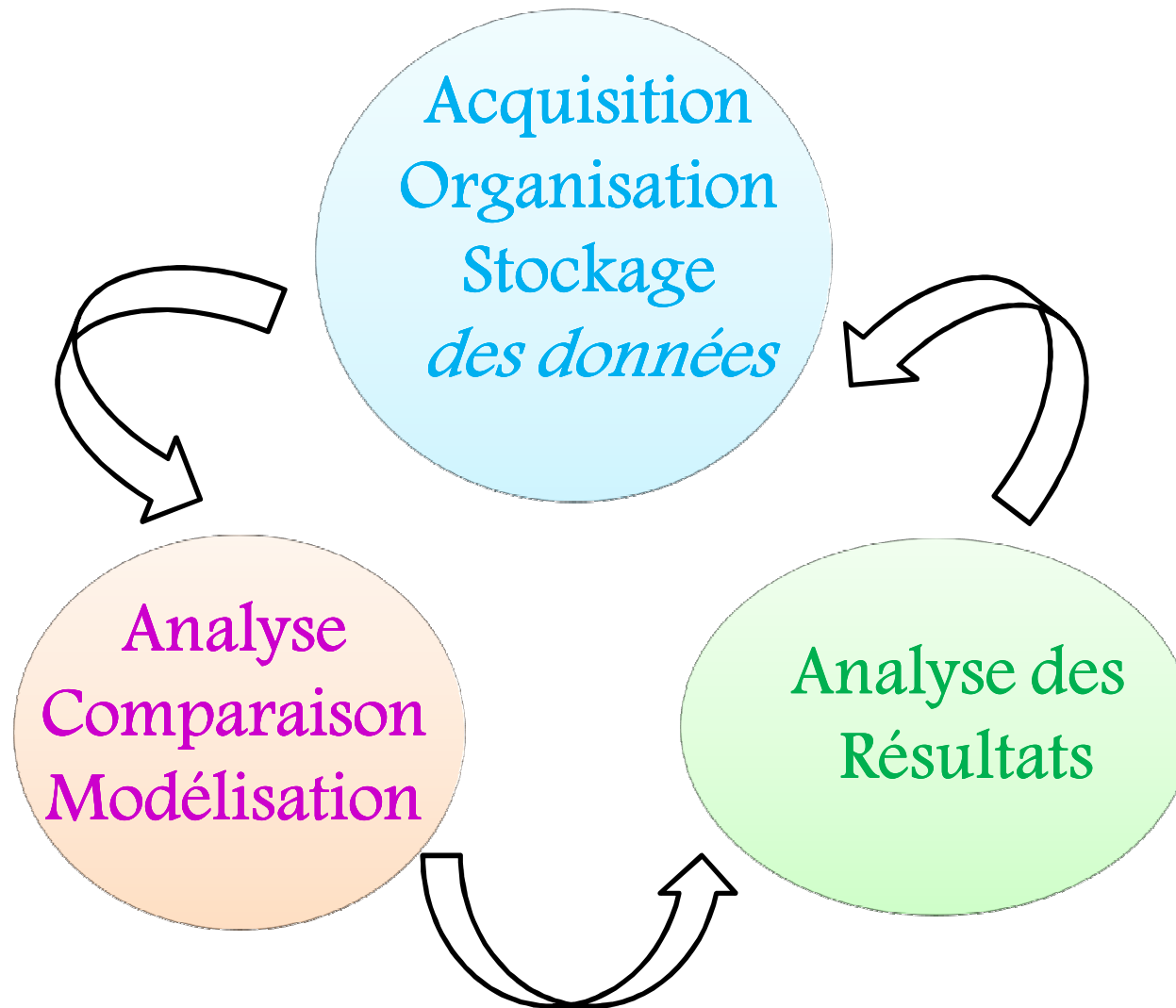
 Première bases de données *EMBL*, *GenBank* et *PIR*

1981: *GenBank* : 270 séquences

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

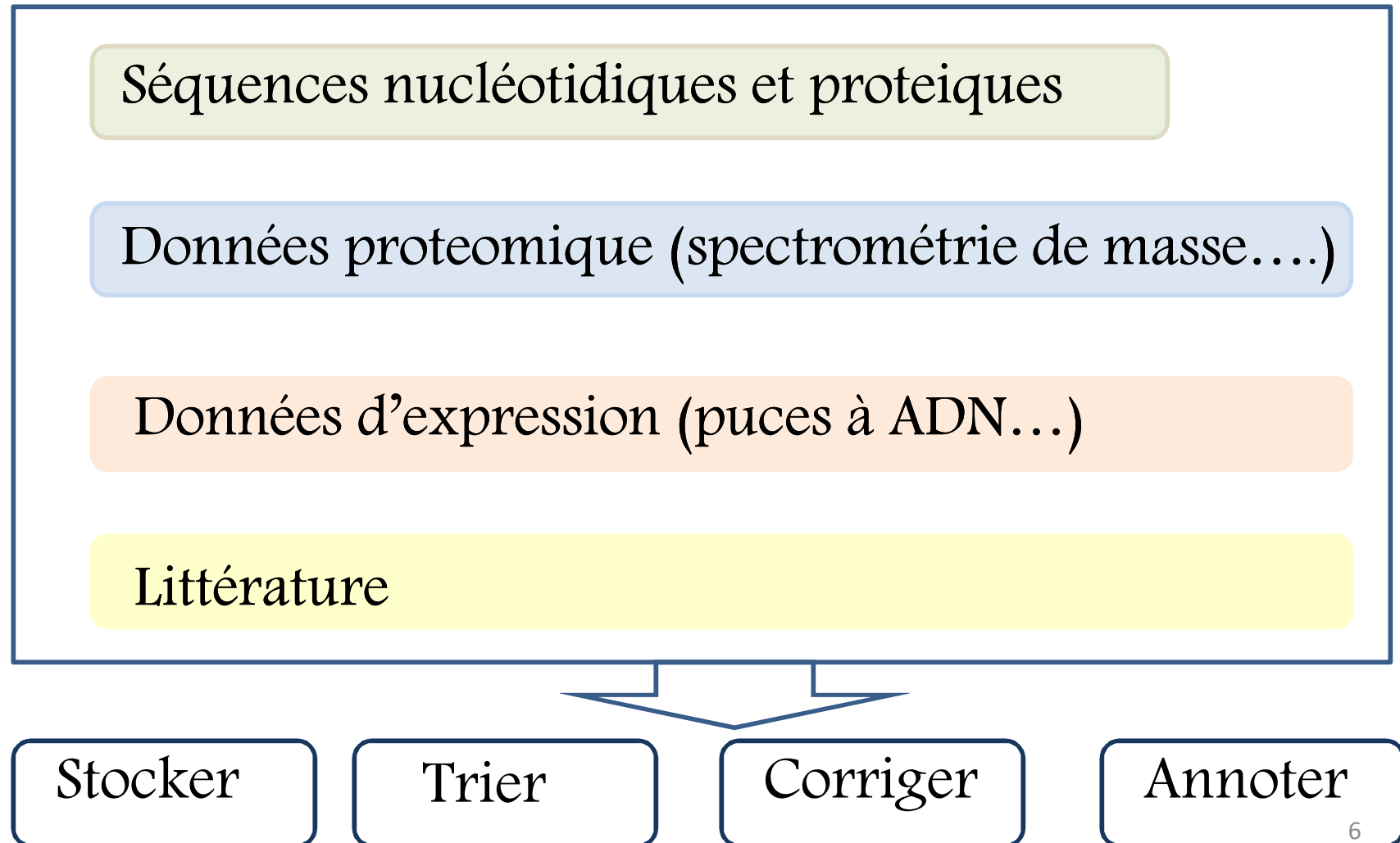
1985: Programme d'alignement local de séquences "FASTA" (Pearson & Lipman)

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:

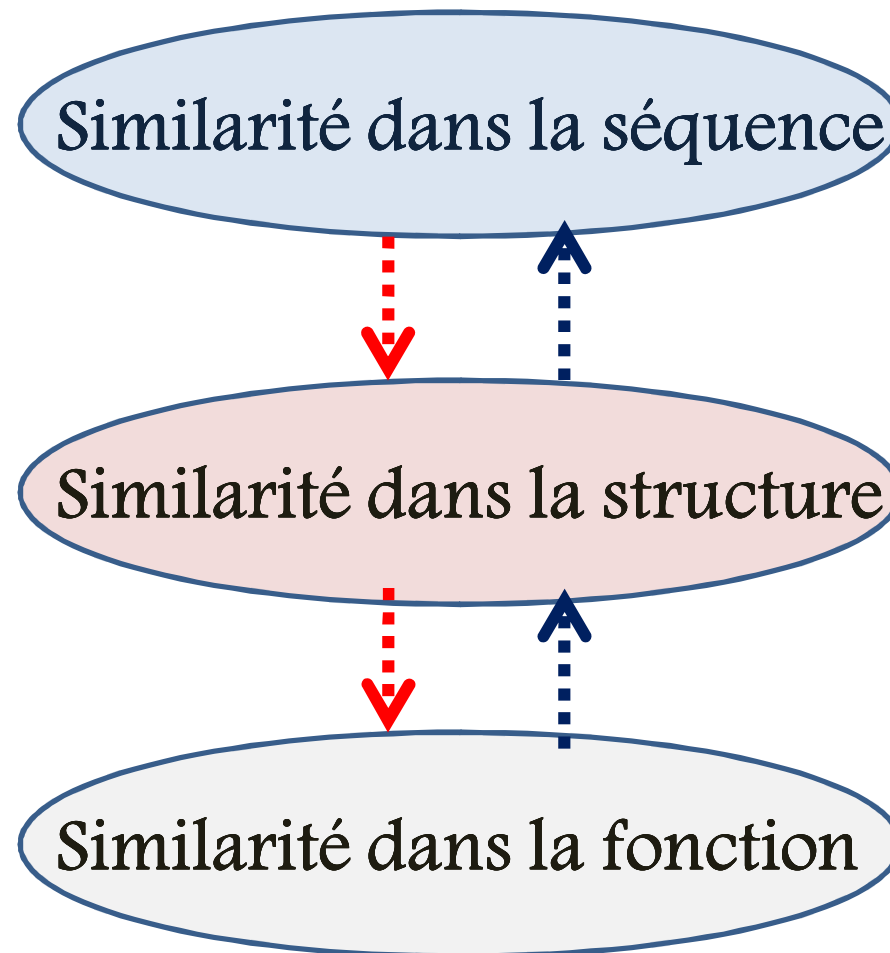
- Recherche de phase de lecture ouverte (ORF) d'un gène
- Déduire la structure Intron/Exon d'un gène
- Etablir l'arbre phylogénique
- Recherche d'Etiquettes EST et du profil d'expression des gènes (profiling)
- Analyse de génomes entiers

Protéines:

- 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 Alignment Search Tool (Ex: Blast sur NCBI)

The screenshot shows the NCBI BLAST website. At the top, there is a navigation bar with the BLAST logo, the text "Basic Local Alignment Search Tool", and buttons for "Home", "Recent Results", "Saved Strategies", and "Help". On the right side of the navigation bar, there are links for "My NCBI", "Sign In", and "Register".

The main content area is divided into several sections:

- NCBI/ BLAST Home:** A introductory text box stating "BLAST finds regions of similarity between biological sequences. [more...](#)" and a link to "New! Aligning Multiple Protein Sequences? Try the COBLAT Multiple Alignment Tool. Go!".
- BLAST Assembled Genomes:** A section where users can choose a species genome to search, or [list all genomic BLAST databases](#). It lists various species: Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera.
- Basic BLAST:** A section where users can choose a BLAST program to run. It lists several options, each with a description and the algorithms used:
 - [nucleotide blast](#): Search a nucleotide database using a nucleotide query. Algorithms: blastn, megablast, discontinuous megablast.
 - [protein blast](#): Search protein database using a protein query. Algorithms: blastp, psi-blast, phi-blast.
 - [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.

On the right side of the page, there are two sidebar sections:

- News:** A section titled "BLAST 2.2.23 release" with a sub-header "A new version of the stand-alone applications is available." and a date "Mon, 22 Mar 2010 15:00:00 EST". It includes a link for "More BLAST news...".
- Tip of the Day:** A section titled "How to do Batch BLAST jobs." with a sub-header "BLAST makes it easy to examine a large group of potential gene candidates." and a link for "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](#)

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

Query ID	lc 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

Mouse over to see the define, click to show alignments

▼ Descriptions

Legend for links to other resources: [U](#) Un [G](#) GEO [E](#) E [S](#) Str [G](#) Gene [M](#) Map [S](#) S [M](#) Sewer [M](#) 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	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)

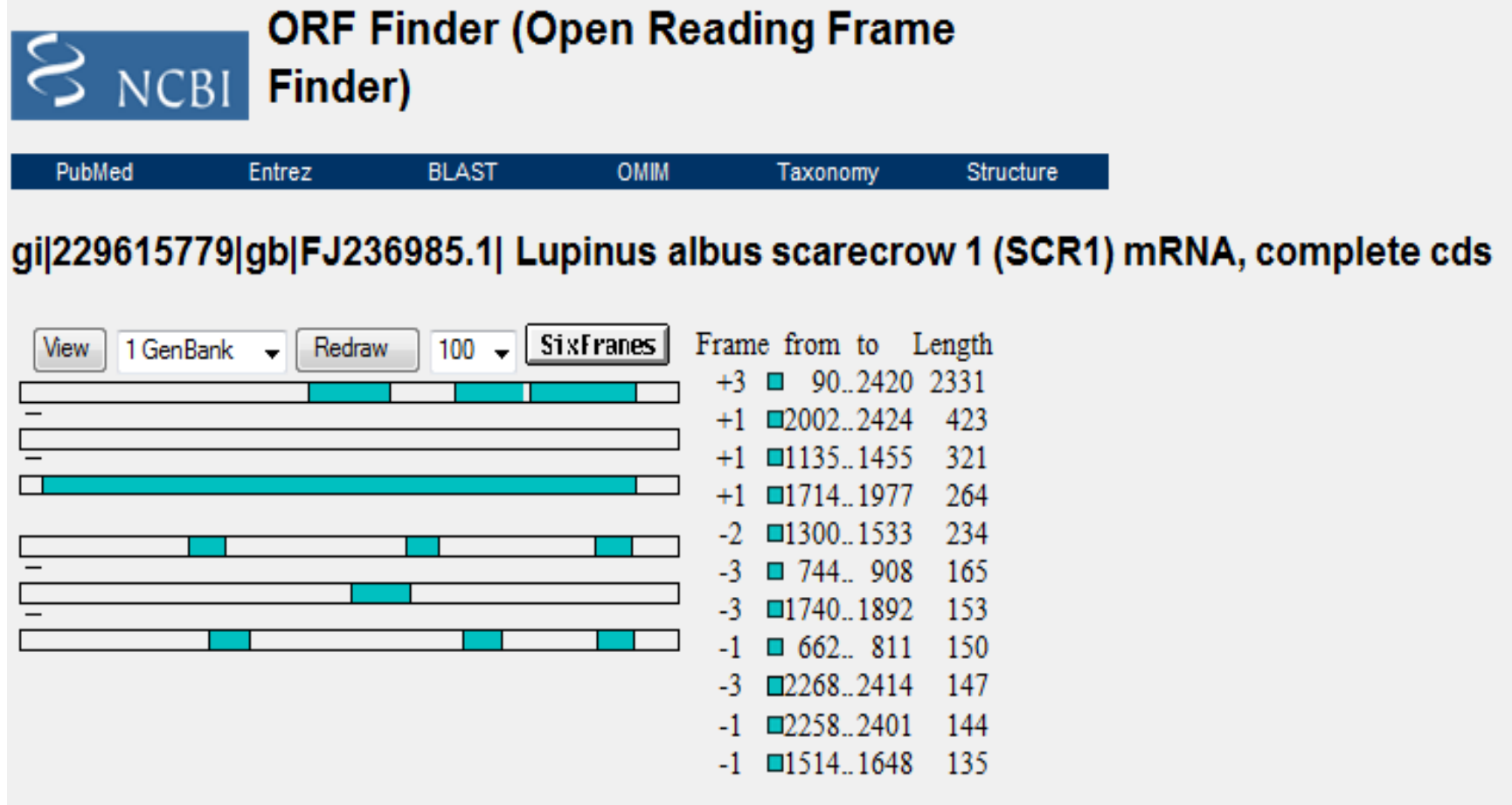
NCBI / Resources / Sequence analysis / Tools / ORF finder

The screenshot shows the NCBI ORF Finder web interface. The page title is "ORF Finder (Open Reading Frame Finder)". The navigation bar includes links for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The left sidebar contains links for NCBI, Tools for data mining, GenBank sequence submission support and software, and FTP site for downloading data and software. The main content area provides a description of the tool and a form for inputting a sequence. The description states: "The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software." The input form includes a text box for "Enter GI or ACCESSION" with an "OrFind" button and a "Clear" button. Below this is a text area for "or sequence in FASTA format" containing the following sequence:

```
>gi|229615779|gb|FJ236985.1|Lupinus albus scarecrow  
1 (SCR1) mRNA, complete cds  
ATTCAAAACAATTITACGTATCACTATCAACTATCACTGCCATCTATCTCACT  
CTTGGCTTCTCCATAA  
ATTGATGCTTCTGATGCTATGCTCAGCGTGCAGCAATAAATACTGAAGA  
TGGCAATAACAATAAT  
GGTGGTAGTCCITTGACTAGTGCCTCTAATAACTCTAGTAAATTTAGCAGTGAA  
GAGCACACTCACAGTC
```

 Below the text area are input fields for "FROM:" and a dropdown menu for "Genetic codes" set to "1 Standard".

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 ClustalW2 web interface. At the top, there is a navigation bar with 'EMBL-EBI' and 'EBI-EMBL Search' logos, a search box containing 'All Databases', and a 'Go' button. Below the navigation bar, there are several menu items: Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. A sidebar on the left contains a 'Help Index' with links to 'General Help', 'Formats', 'Gaps', 'Matrix', 'References', 'ClustalW2 Help', 'ClustalW2 FAQ', 'Jalview Help', 'Scores Table', 'Alignment', 'Guide Tree', and 'Colours'. Below the sidebar, there are sections for 'Similar Applications' (Align, Kalign, MAFFT, MUSCLE, T-Coffee), 'ClustalW Programmatic Access', and 'www.clustal.org'. The main content area is titled 'ClustalW2' and contains a description of the program. Below the description, there is a 'Download Software' link. The main configuration area is divided into several sections: 'YOUR EMAIL' (input field), 'ALIGNMENT TITLE' (input field with 'Sequence'), 'RESULTS' (dropdown menu with 'interactive'), and 'ALIGNMENT' (dropdown menu with 'full'). Below these are several other options: 'KTUP (WORD SIZE)' (dropdown with 'def'), 'WINDOW LENGTH' (dropdown with 'def'), 'SCORE TYPE' (dropdown with 'percent'), 'TOPDIAG' (dropdown with 'def'), 'PAIRGAP' (dropdown with 'def'), 'MATRIX' (dropdown with 'id'), 'GAP OPEN' (dropdown with 'def'), 'NO END GAPS' (dropdown with 'yes'), 'GAP EXTENSION' (dropdown with 'def'), 'GAP DISTANCES' (dropdown with 'def'), 'ITERATION' (dropdown with 'none'), and 'NUMBER' (dropdown with '1'). Below these are 'OUTPUT' options: 'OUTPUT FORMAT' (dropdown with 'aln w/numbers'), 'OUTPUT ORDER' (dropdown with 'aligned'), 'TREE TYPE' (dropdown with 'none'), 'PHYLOGENETIC TREE' options: 'CORRECT DIST.' (dropdown with 'of1'), 'IGNORE GAPS' (dropdown with 'of1'), and 'CLUSTERING' (dropdown with 'NJ'). At the bottom, there is a text input field for 'Enter or paste a set of sequences in any supported format:' with a 'Help' button. Below the input field, there is a 'Browse...' button and 'Run' and 'Reset' buttons. The input field contains the following text: '>gi|229615779|gb|FJ236985.1| Lupinus albus scarecrow 1 (SCR1) mRNA, complete cds ATTCAAAA CAATT TACGTATCACTATCAACTATCACTCCCTATCTATCTCACTCTTGGCTTCTCCAT AA ATTGATGCTTCTGATGCTATGCTCAGCGGTGCCAGCAATAATAACTGAAAGATGGCAATAACAATA AT GGTGGTAGTCCTTTGACTAGTGCCTCTAATAACTCTAGTAAATTTAGCAGTGAAGAGCACACTCACAG TC AGAAGCTACAGCATTCCCATTCGGACAAAGGAAAAATGGTGAGAAAGAGGATGGCTTCTGAGATGGAA CC'.

Exemple d'un Multi-alignement

Séquence 1	→	LaSCR1	1	MLSGASNNNT	EDGNNNN---	GGSPLOTSASN	NSSKFSSEEH	THSQKLQHS	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	DNMSLLLNC	SLPPAVEKTT	SFNNNDYHYK	115
		LaSCR2	61	*****	*****IDS-	*****SSS*N	N***SV*I**	*****	****-Y****	118
		AtSCR	48	L****S----	-----	-----	-----	-----	--SNPD*NNS	61
		LaSCR1	116	TSSSSKVDNN	VAVVPNPTTP	NYSTMLLPSS	SCSTTINPNY	HISQRQQDQ	NQLTSPAVCG	175
		LaSCR2	119	*****R****	AH*****	****L*****	-----	N*I*T*****	*****	176
		AtSCR	62	SRPPRR*SHL	LDSNY*TV**	QQPPS*TAAA	TV*SQPN*--	-----	----PLS***	105
		LaSCR1	176	FSGLPLFPAS	QQRNHHHNS	SSSSTGANVE	VAASPSMEDN	NNN--SAATA	WIDGILKDLI	233
		LaSCR2	177	*****	*****--	-*T****T**	*****	***NN****D	*****	233
		AtSCR	106	*****V**SD	RG-----	-----*R**M	MSVQ*-*DQD	SSSSASPTV	*V*A*IR***	151
		LaSCR1	234	HSSNSVSIPO	LINNVREIIY	PCNPNLAVVL	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	IPDNFPPDPS	-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)

