
Projet L2D -Introduction à la bioinformatique

Manuel d'installation

Projet de Programmation

L2D1



Illustration 1: Image de présentation

Projet L2D -Introduction à la bioinformatique

Les informations d'identification du document :

Référence du document :
Version du document : 2
Date du document : 01/05/2021
Auteurs : Mehdi Hamiche Manal Boutajar Adelin Bodnar

Les éléments de vérification du document :

Validé par :
Validé le :
Soumis le :
Type de diffusion :
Confidentialité :

Les éléments d'authentification :

Maître d'ouvrage :	Chef de projet :
Date / Signature :	Date / Signature :

Projet L2D -Introduction à la bioinformatique

Sommaire

<u>1. Introduction</u>	5
1.1. Objectifs et méthodes	5
1.2. Documents de référence	6
<u>2. Guide de lecture</u>	7
2.1. Maîtrise d'œuvre	7
<i>2.1.1. Responsable</i>	7
<i>2.1.2. Personnel administratif</i>	7
<i>2.1.3. Personnel technique</i>	7
2.2. Maîtrise d'ouvrage	8
<i>2.2.1. Responsable</i>	8
<i>2.2.2. Personnel administratif</i>	8
<i>2.2.3. Personnel technique</i>	8
<u>3. Concepts de base</u>	9
<u>4. Installation du matériel</u>	11
<u>5. Paramétrage du système</u>	12
<u>6. Installation du logiciel</u>	13

Projet L2D -Introduction à la bioinformatique

<u>7. Installation des données</u>	14
<u>8. Autres informations</u>	18
<u>9. Annexes</u>	19
<u>8. Glossaire</u>	20
<u>9. Références</u>	22
<u>10. Index</u>	23

1. Introduction

Le manuel d'installation est un document rassemblant l'ensemble des procédures nécessaires à la mise en place de notre application dans son environnement de production (conditions réelles d'utilisation). Il permet à un administrateur système d'installer et de configurer l'application sur des systèmes informatiques. Les rubriques du manuel d'installation sont les suivantes :

- **Installation du matériel** (préciser le matériel à installer et les opérations nécessaires à sa mise en fonction)
- **Paramétrage du système** (lister les opérations nécessaires pour paramétrier convenablement le système)
- **Installation du logiciel** (lister les opérations nécessaires pour installer le logiciel : copie de fichiers, etc.)
- **Installation des données** (lister les opérations nécessaires à la mise en place des données de l'application : copie de fichiers, création de base de données, etc.)
- **Autres informations** (lister d'autres opérations utiles : informer les utilisateurs, mettre hors service une application le temps de l'installation, procédure de mise à jour, etc.)

1.1. Objectifs et méthodes

Permettre à notre programme d'être plus simple à utiliser que ce soit dans la visibilité ou dans la façon d'utiliser.

- Alignement de séquences :
 - protéiques,
 - nucléotidiques

Projet L2D - Introduction à la bioinformatique

- Global
- Multiple
- Interface graphique permettant de :
 - voir les alignements globaux et multiples,
 - générer le logo

Modules

1. *Alignement global* - 2 séquences alignées via l'algorithme de Needleman - Wunsch
2. *Alignement multiple* - amélioration de l'alignement global (consensus) - alignement de trois séquences au moins
3. *Logo* - représentation graphique de l'alignement multiple
4. *Interface graphique* - relie les 3 modules et les présente

1.2. Documents de référence

Les documents du projet servant à l'élaboration du présent document :

- Le cahier des charges,
- Le cahier de recettes,
- La conception générale

2. Guide de lecture

2.1. Maîtrise d'œuvre

La maîtrise d'œuvre présente l'équipe du développement chargé du bon suivi du manuel d'installation et des besoins dont le maître d'ouvrage fait commande.

Elle représente l'équipe du développement :

- Adelin Bodnar
- Manal Boutajar
- Mehdi Hamiche

Cette équipe veillera au bon suivi du manuel d'installation coordonnées avec la conception générale représentant les besoins des enseignants encadrants.

2.1.1. Responsable

Il est conseillé pour le responsable de la maîtrise d'œuvre de lire le document dans sa totalité afin de prendre conscience de l'ensemble des éléments.

2.1.2. Personnel administratif

Il est conseillé pour le personnel administratif de lire la présentation du produit, paramétrage du système.

2.1.3. Personnel technique

Il est conseillé pour le personnel technique de prendre en compte la partie sur le concept de base, installation du matériel, installation du logiciel et installation des données.

2.2. Maîtrise d'ouvrage

Projet L2D - Introduction à la bioinformatique

La maîtrise d'ouvrage représente dans notre cas le client du projet, c'est-à-dire les personnes dont les besoins permettent la conception du projet.

La maîtrise d'ouvrage est assistée par l'équipe de la maîtrise d'œuvre et donc ce rôle sera assuré par les enseignants encadrants Dragutin Jastrebic et Koviljka Lukic Jastrebic.

2.2.1. Responsable

Il est conseillé pour le responsable de la maîtrise d'ouvrage de lire le document dans toute sa totalité afin de prendre conscience de l'ensemble des documents.

2.2.2. Personnel administratif

Il est conseillé pour le personnel administratif de lire la présentation du produit, paramétrage du système.

2.2.3. Personnel technique

Il est conseillé pour le personnel technique de prendre en compte la partie sur le concept de base, installation du matériel, installation du logiciel et installation des données.

3. Concepts de base

Pour bien comprendre ce document, lire le manuel d'utilisation pour avoir les notions sur les différents modules abordés et quelques définitions pour ne pas être perdu :

• Bio-informatique	➔ Interdiscipline
• Modules	➔ Alignement : <ul style="list-style-type: none">• Global (backtracking),• Multiple (séquences consensus) ➔ Weblogo ➔ Interface graphique
• Needleman-Wunsch	➔ Algorithme
• Fasta	➔ Format
• Programmation orienté objet	➔ Java <ul style="list-style-type: none">• Connaissance;• Compréhension du code (Eclipse : installation Windows);• Exécution :<ul style="list-style-type: none">• classes,• packages,• méthodes,• variables etc...
• Programme	➔ Application alignant les séquences

Projet L2D - Introduction à la bioinformatique

	protéiques et nucléotidiques (archive .zip)
--	---

4. Installation du matériel

Pour pouvoir exécuter le programme, il faut installer un IDE comportant un compilateur pour le langage JAVA comme Eclipse, Netbeans, ect :

Installation Eclipse, Netbeans

Après l'installation d'éclipse et récupération des classes java du projet, on exécute la classe **FramePrincipale** pour avoir la page d'accueil où on trouve une introduction sur le projet et comment utiliser l'application web.

5. Paramétrage du système

On doit avoir à côté d'Eclipse et des classes nécessaire pour le bon fonctionnement de l'interface, des fichiers .fasta et .aln-fa (ou .txt) pour parcourir ensuite sur l'interface afin de réaliser des alignements multiples et des logo des séquences protéiques ou nucléotidiques.

6. Installation du logiciel

Le logiciel sera organisé sous forme de projet Java qu'il sera possible d'ouvrir avec un IDE.

Le dit projet doit être téléchargé à un emplacement connu qui sera nécessaire celui-ci avec l'IDE.

Dans le dossier devront obligatoirement se trouver le fichier “dna.jpg” ainsi que le dossier “src” contenant les fichiers :

- “**FramePricipale.java**”
- “**FrameAG.java**”
- “**FrameAM.java**”
- “**ResAlignment.java**”
- “**Sequence.java**”
- “**Fasta1.java**”
- “**AlignementG.java**”
- “**Alignement_Multiple.java**”
- “**Matrice.java**”
- “**Main.java**”
- “**TestWeblogo.java**”

Les fichiers peuvent être placés aux côtés du fichier “dna.jpg” pour saisir seulement leur nom au format “nom.extension” si vous ne souhaitez pas utiliser la fonction parcourir.

7. Installation des données

Installation de différents logiciels pour mieux comprendre notre application à travers d'autres langages :

PYTHON (Version 3.8)

Installation biotite - pip install biotite

Installation biopython – pip install biopython

Installation urllib – pip install urllib

Installation weblogo – pip install weblogo

Voici différents exemples :

```
Python 3.8 (64-bit)
>>> from Bio import pairwise2
>>> ## load the module
>>> ## globalxx
>>> alignments = pairwise2.align.globalxx("ACTG", "CTTG")
>>> for alignment in alignments:
...     ## Each alignment is a tuple consisting of the two aligned sequences,
...     ## the score, the start and the end positions of the alignment
...     ## (in global alignments the start is always 0 and the end the length of the alignment).
...     print(alignment)
...
...     ## print the alignment in a nicer format
...     from Bio.pairwise2 import format_alignment
...     print(format_alignment(*alignment))
...     print(repr(alignment))
...
Alignment(seqA='ACT-G', seqB='-CTTG', score=3.0, start=0, end=5)
ACT-G
|| |
-CTTG
Score=3

Alignment(seqA='ACT-G', seqB='-CTTG', score=3.0, start=0, end=5)
Alignment(seqA='AC-TG', seqB='-CTTG', score=3.0, start=0, end=5)
AC-TG
| |
-CTTG
Score=3

Alignment(seqA='AC-TG', seqB='-CTTG', score=3.0, start=0, end=5)
>>>
```

Illustration 2: PSA Python

Projet L2D - Introduction à la bioinformatique

Illustration 3: NCBI - NUCLEOTIDE - ACCÉDER AUX BDDs BIOLOGIQUES

```
(base) C:\Users\hamichi>python
Python 3.8.5 (default, Sep  3 2020, 21:29:08) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> from Bio import Entrez
>>> from Bio import SeqIO
>>> Entrez.email = "hami.chemehdi@gmail.com"
>>> handle = Entrez.efetch(db="protein", rettype="fasta", retmode="text",
...     id="P01317")
>>> for seq_record in SeqIO.parse(handle, "fasta"):
...     print (seq_record.id)
...
P01317.1|INS_BOVIN
>>> from Bio import Entrez
>>> from Bio import SeqIO
>>> Entrez.email = "hami.chemehdi@gmail.com"
>>> handle = Entrez.efetch(db="protein", rettype="fasta", retmode="text",
...     id="P01317")
>>> file = open("P01317.fasta", "w")
>>> file.write(""><stdin>, line 2
...
...     id="P01317")
SyntaxError: invalid syntax
>>> handle = Entrez.efetch(db="protein", rettype="fasta", retmode="text",
...     id="P01317")
>>> for seq_record in SeqIO.parse(handle, "fasta"):
...     print (seq_record.seq)
...
MALWTRRLPQLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPQVQGALEAGGGAGGLEGPQPKRGIVEQCCASVCSLYQLENYCN
>>> from Bio import Entrez
>>> from Bio import SeqIO
>>> Entrez.email = "hami.chemehdi@gmail.com"
>>> handle = Entrez.efetch(db="protein", rettype="fasta", retmode="text",
...     id="P01317")
>>> for seq_record in SeqIO.parse(handle, "fasta"):
...     print (seq_record.id,seq_record.seq)
...
P01317.2|INS_BOVIN MALWTRRLPQLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPQVQGALEAGGGAGGLEGPQPKRGIVEQCCASVCSLYQLENYCN
>>>
```

Illustration 4: NCBI - PROTEINE - ACCEDER AUX BDDs BIOLOGIQUES

```
C:\Python27\python.exe
]
>>> import urllib
>>>
>>> f = file
>>>
>>> file = urllib.urlopen("http://tugows.org/entry/pdbj-pdb/6m17/keywords.json")
>>> jsondata = file.read()
>>> f.close()
>>>
>>>
>>> print(jsondata)
[
    [
        "2019-NCOV RBD",
        "ACE2-B0AT1 COMPLEX",
        "MEMBRANE PROTEIN",
        "MEMBRAN",
        "PROTEIN-VIRAL PROTEIN COMPLEX"
    ]
]
>>>
```

Illustration 5: DDBJ Exemple RECUPERATION DE DONNÉES AU FORMAT .FASTA

Illustration 6: BLASTN BIOPYTHON

Projet L2D - Introduction à la bioinformatique

R – Rstudio (version 1.4.1106)

Installation msaR (Library) (voir forge)

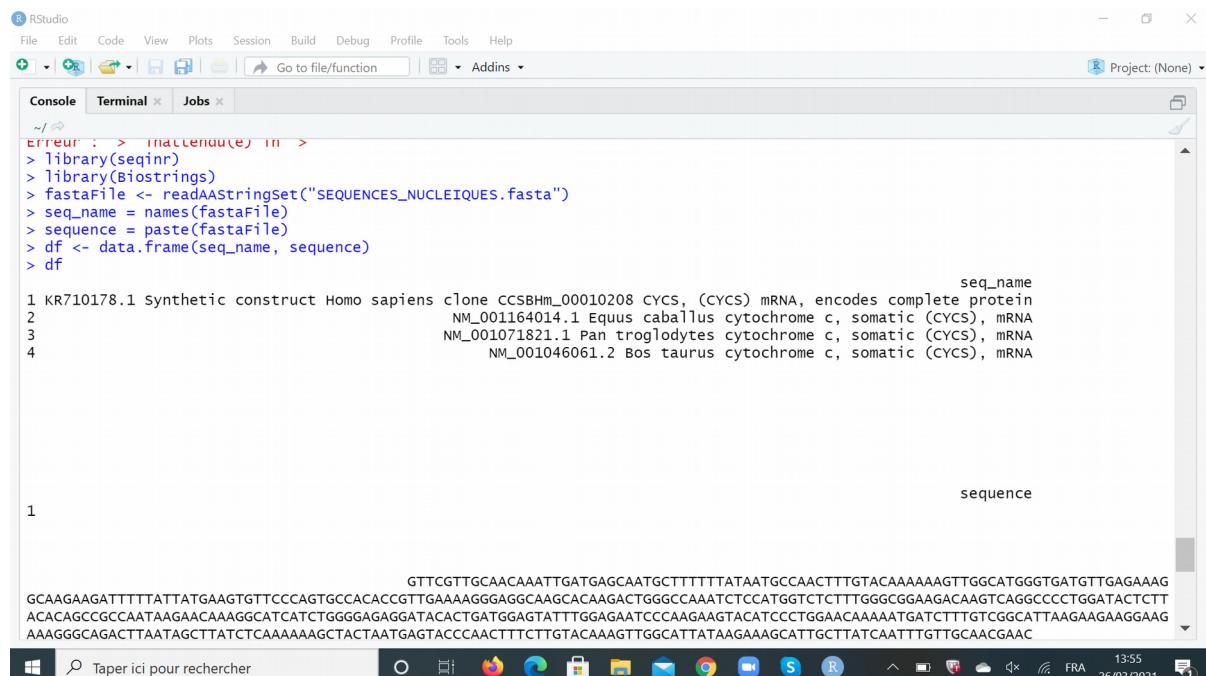
Voici quelques exemples :

```
> library(msa)
> CYTOCHROME_C <- readAStringset("c:/Users/hamic/Desktop/.fasta/CYTOCHROME_C.fasta")
> cytochrome_CMsaAlignment <- msa(CYTOCHROME_C)
use default substitution matrix
> cytochrome_CMsaAlignment
CLUSTAL 2.1

Call:
msa(CYTOCHROME_C)

MsaAMultipleAlignment with 21 rows and 105 columns
  names
  aln
[1] MGDIKEGKKIFVQKCSQCHTVEKGKHKTGPNLNGLF...ENPKKYIPGTMIFAGIKKKSERADLIAYLKDATSK sp|P00018.2|CYC_D...
[2] MGDIKEGKKIFVQKCSQCHTVEKGKHKTGPNLNGLF...ENPKKYIPGTMIFAGIKKKSERADLIAYLKDATSK sp|P00019.2|CYC_S...
[3] MGDIKEGKKIFVQKCSQCHTVEKGKHKTGPNLNGLF...ENPKKYIPGTMIFAGIKKKSERADLIAYLKDATSK sp|P00017.2|CYC_A...
[4] MGDIKEGKKIFVQKCSQCHTVEKGKHKTGPNLNGLF...ENPKKYIPGTMIFAGIKKKSERVDIAYLKDATSK sp|P67881.2|CYC_C...
[5] MGDAEAGKKIFVQKCAQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFAGIKKKSERADLIAYLKDATAK sp|P00020.2|CYC_A...
[6] MGDAEAGKKIFVQKCAQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFAGIKKKSEREDLIYKLQQATSS sp|P00015.3|CYC_2...
[7] MGDAEAGKKIFVQKCAQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFAGIKKKTERERDLIAYLKKATNE sp|P00004.2|CYC_H...
[8] MGDAEAGKKIFVQKCAQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFAGIKKKTERERDLIAYLKKATNE sp|P68096.2|CYC_E...
[9] MGDAEAGKKIFVQKCAQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFAGIKKKGEREDLIAYLKKATNE sp|P62894.2|CYC_B...
...
[14] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P99999.2|CYC_H...
[15] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P99998.2|CYC_P...
[16] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P00002.2|CYC_M...
[17] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P00003.2|CYC_A...
[18] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P00014.2|CYC_M...
[19] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P00015.3|CYC_2...
[20] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P00004.2|CYC_H...
[21] MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE sp|P00008.2|CYC_R...
Con MGDAEAGKKIFIMKSQCHTVEKGKHKTGPNLWGLF...ENPKKYIPGTMIFVGIKKKEERADLIAYLKKATNE Consensus
> |
```

Illustration 7: msa R - alignement multiple avec séquences cytochrome c



The screenshot shows the RStudio interface with the console tab active. The console window displays the following R code and its output:

```
Erreur : > inattendue(e) in >
> library(seqinr)
> library(biostrings)
> fastafile <- readAStringset("SEQUENCES_NUCLEIQUES.fasta")
> seq_name = names(fastafile)
> sequence = paste(fastafile)
> df <- data.frame(seq_name, sequence)
> df
```

The output shows the names and sequences of several cytochrome c genes from different species:

```
seq_name
1 KR710178.1 synthetic construct Homo sapiens clone CCSBhM_00010208 (CYCS) mRNA, encodes complete protein
  NM_001164014.1 Equus caballus cytochrome c, somatic (CYCS), mRNA
  NM_001071821.1 Pan troglodytes cytochrome c, somatic (CYCS), mRNA
  NM_001046061.2 Bos taurus cytochrome c, somatic (CYCS), mRNA
```

Below the console, the sequence viewer shows the sequence for KR710178.1. The sequence is a long string of nucleotides starting with GTCGTTGCAACAAATTGATGAGCAATGCTTTATAATGCCAATTGTACAAAAAGTTGGCATGGGTGATGTTGAGAAAAG.

Illustration 8: Test Rstudio Covid visualisation séquence

8. Autres informations

Pour information, lors du choix de réaliser des logos de séquences protéiques ou nucléotidiques sur l'interface, il sera obligatoire de fermer l'application après les résultat du logo et puis exécuter de nouveau la **FramePrincipale** pour pouvoir faire un autre logo d'une autre séquence, contrairement aux alignements globaux et multiples qui ne nécessite pas fermer le programme, il suffit juste d'appuyer sur le bouton « **Retour** » en haut à gauche de l'interface.

9. Annexes

- Documentation java ;
- Logiciel MAFFT - EMBOSS - MULTALIN
 - tester les alignements
- Forge : nos différents tests

8. Glossaire

- ★ **BIO-INFORMATIQUE** : discipline permettant d'analyser de l'information biologique qu'il y a dans les séquences nucléotidiques et protéiques
- ★ **PROGRAMMATION ORIENTÉE OBJET** : met en œuvre différents objets (instances de classes). Chaque objet associe des données et des méthodes (fonctions) agissant exclusivement sur les données
- ★ **ALGORITHME NEEDLEMAN - WUNSCH** : effectue un alignement global maximal de deux chaînes de caractères

- ★ **ALIGNEMENT GLOBAL** : recouvre les séquences alignées (par exemple 2 séquences) sur l'ensemble de leur longueur
- ★ **ALIGNEMENT MULTIPLE** : aligner un ensemble de séquences homologues, comme des séquences protéiques ou nucléiques qui assure des fonctions similaires dans différentes espèces vivantes
- ★ **WEBLOGO** : application web conçue pour rendre la génération de logos de séquence aussi simple que possible
- ★ **INTERFACE GRAPHIQUE** : relie les 3 modules et les présente de manière simple et efficace

- ★ **GAP, MATCH, MISMATCH** : valeurs pour alignement global
- ★ **FASTA** : format de fichier texte dans lesquels les séquences sont affichées par une suite de lettres

- ★ **NCBI** : National Centre for Biotechnology Information
 - **PROTEIN** : Base de données pour la récupération des séquences protéiques en format .fasta
 - **GENE** : Base de données pour récupération des séquences nucléiques en format .fasta

Projet L2D - Introduction à la bioinformatique

- **NUCLEOTIDE** : Base de données pour la récupération des séquences nucléiques en format .fasta
- **UNIPROT** : Base de données pour la récupération des séquences protéiques en format .fasta
- **DDBJ** : Base de données pour la récupération des séquences nucléiques en format .fasta
- **DNA** : Data Bank Japan
- **EMBL** : European Molecular Biology Laboratory
- **EMBOSS** : outil bioinformatique pour l'alignement global – NEEDLE
- **BLAST** : outil bioinformatique pour l'alignement local
- **MULTALIN** : outil bioinformatique pour l'alignement multiple

9. Références

[Alignement de séquences](#)

[BNLEARN](#)

[DDBJ](#)

[National Centre for Biotechnology Information](#)

[Réseaux Bayésiens](#)

[UniProt](#)

[Alignement global](#)

[Alignement multiple](#)

[JAVA - Main \(site 1\)](#)

[JAVA - Main \(site 2\)](#)

[Chaînes de caractères](#)

[Tableau](#)

[Needleman et Wunsch – Java](#)

[Needleman et Wunsch – Python](#)

[EMBOSS](#)

[MAFFT](#)

10. Index

Index des figures

<i>Illustration 1: Image de présentation</i>	1
<i>Illustration 2: PSA Python</i>	14
<i>Illustration 3: NCBI - NUCLEOTIDE - ACCÉDER AUX BDDs BIOLOGIQUES</i>	15
<i>Illustration 4: NCBI - PROTEINE - ACCÉDER AUX BDDs BIOLOGIQUES</i>	15
<i>Illustration 5: DDBJ Exemple RECUPERATION DE DONNÉES AU FORMAT .FASTA</i>	16
<i>Illustration 6: BLASTN BIOPYTHON</i>	16
<i>Illustration 7: msa R - alignement multiple avec séquences cytochrome c</i>	17
<i>Illustration 8: Test Rstudio Covid visualisation séquence</i>	17

Index lexical

alignement global	6, 20
Alignement global	6, 22
ALIGNEMENT GLOBAL	20
alignement multiple	6
Alignement multiple	6, 22
ALIGNEMENT MULTIPLE	20

Projet L2D - Introduction à la bioinformatique

Alignement_Multiple	13
AlignementG	13
aln-fa	12
application	20
BIO-INFORMATIQUE	20
Eclipse	9, 11 sv
fasta	12
Fasta	9, 13
FASTA	20
Fasta1	13
FrameAG	13
FrameAM	13
FramePricipale	13
FramePrincipale	11
GAP	20
Interface graphique	6, 9, 20
INTERFACE GRAPHIQUE	20
java	11, 13, 19

Projet L2D - Introduction à la bioinformatique

Java	9, 13, 22
JAVA	11, 22
logo	6, 9, 12, 20
Logo	6
LOGO	20
Main	13
manuel d	9
manuel d'installation	5, 7
Manuel d'installation	1
MATCH	20
Matrice	13
MISMATCH	20
NEEDLEMAN - WUNSCH	20
Netbeans	11
nucléiques	20
nucléotidiques	5, 10, 12, 20
protéiques	5, 10, 12, 20
ResAlignement	13

Projet L2D - Introduction à la bioinformatique

Sequence	13
WEBLOGO	20