

Introduction à la Bioinformatique

I. Motivation et objectifs

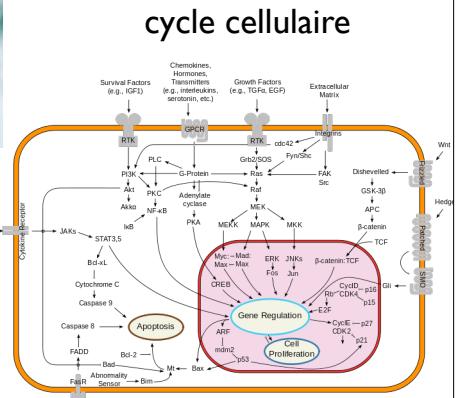
1

Propriétés et modèles?



Cancer
l'autisme
alzheimer

cycle cellulaire



3-1

La bio-quoi???

Définition simple:

La bioinformatique est l'analyse des données biologiques pour :

1. l'identification des propriétés intéressantes et
2. la construction des modèles biologiques en utilisant ces données.

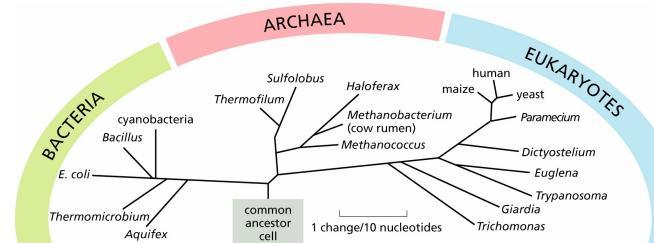
2

Propriétés et modèles?



3-2

Les données



Tous les organismes connus utilisent 2 types de données :

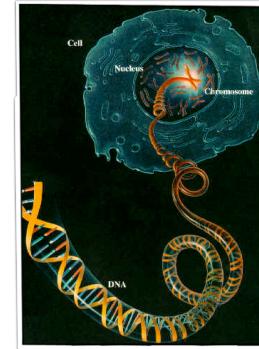
1. séquences de nucléotides
2. séquences d'acides aminés

Et leurs structures 3D

4

La structure des données

ADN



strings

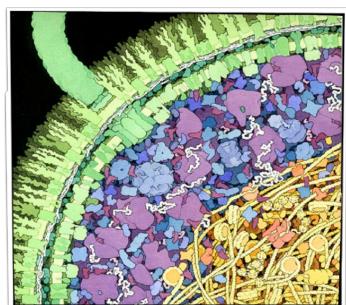
```
... TTGTACATCTCTATCTACTTATCGTCTAGCAGCAGC  
TACTCGTCTGACTGTCGATGCTACGTCTAGTCATGCTAC  
TATCGATGCACTGCGATCGTAATCGGGCTACTAGGGCGG  
GTGTCATATACTAGCCTCTAGCGCTAGCAGCTGATGATC  
TAGTCGTTCATGTCGATGCACTGAGCTAGTCTAGCTATCTA  
TACTAGCAGCGATGCTAGCGTACGCTAGCTATATAGCTAC  
TCTGATATACTGCCGCTACTGACCTACTGCAAGCAGCTGAC  
TGCTGACTGCTGACTGACGTAGCTGACATTGATGCTAGC  
TAGCTTACATGCCGATGCTAGCTAGCGATGCTACGTA  
GCCTACGGTACTTGGGATGCTAGCTGCTTAGTCGATT  
GTGCGATGACTCTGTCGAGTCACTGAGCTGATGACTG  
ACTGACGTCGACTGATGCACTGACTGACTGACTGACTGC  
ATGTCGTCGACTGACTGACGCTGCACTGACTGCACTGAC  
GTGCACTGACTGACTGACTGCGCTGACTGACTGACTG  
CTGACTGACTGTCAGTGA  
CTGACTGACTGACTGACTGACTGACTGACG...
```

séquences d'acides nucléiques

5

La structure des données 2

Protéines



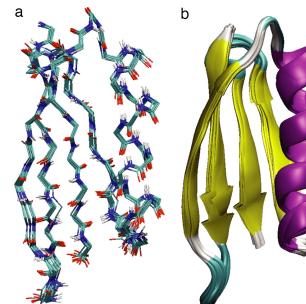
strings

130	140	150	160	170	180
EARSLLTGET	GYIPSNVAF	VDSIQAEEWY	FCKLGRKDAE	RQLLSFGNPR	GTFPLIRESQ
190	200	210	220	230	240
TKGAYSLSIR	DWDDMKGDHV	KHYKIRKLON	GGYYITTRAQ	FETIQLQVQH	YSERAAGLCC
250	260	270	280	290	300
RLVVPCHKGM	PRLTDLSVKT	KDVWVEIPRES	LQLIKRLGLNG	QFGEVWLQTV	NGNTKVIAKLT

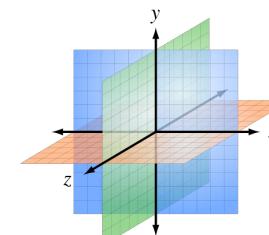
séquences d'acides aminés

6

Protéines



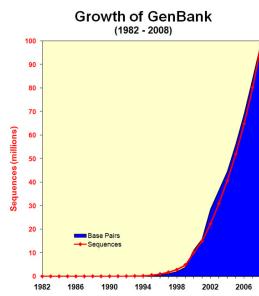
l'ADT Atome (x,y,z)



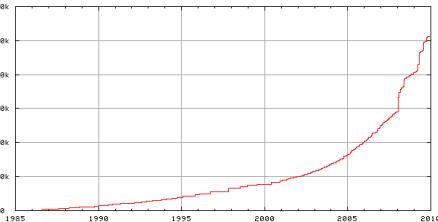
séquences de coordonnées d'atomes

7

Explosion de données



Uniprot/swissprot statistiques



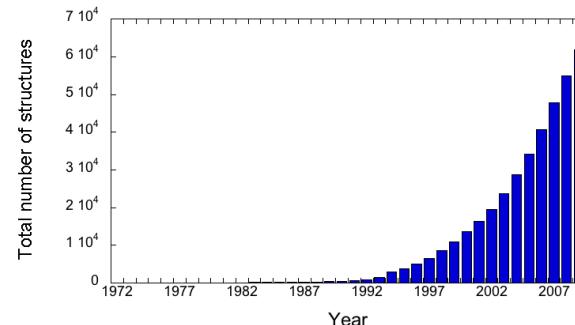
Release 174.0, oct 2009,
99 116 431 942 paires de
base et 98 868 465
séquences

Release 57.11, Novembre 2009
512 994 séquences

8

Explosion de données 2

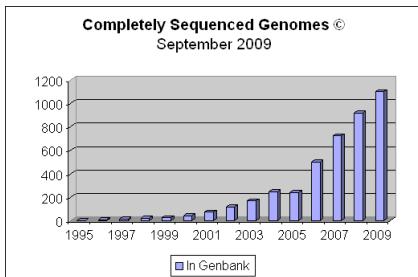
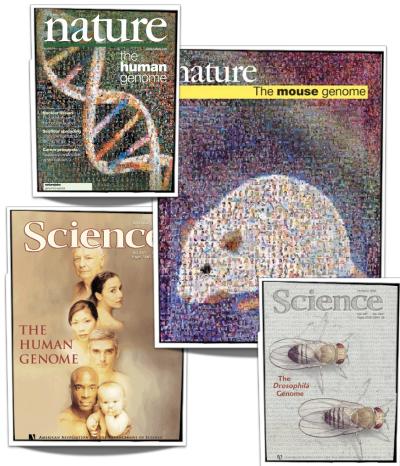
Nombre des structures de protéines



RCSB base de données de protéines version du 3 décembre 2009

9

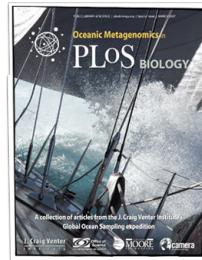
Explosion de données 3



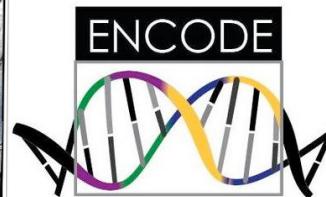
10

Explosion de données 4

ça ne termine pas ...

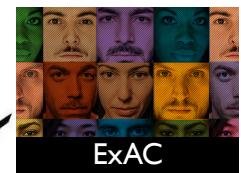


Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* 5(3): e77.



Source: <http://www.genome.gov/10005107>

ENCODE Project Consortium. "The ENCODE (ENCyclopedia of DNA elements) project." *Science* 306.5696 (2004): 636-640.



Source: <http://exac.broadinstitute.org>

Lek M, et al. Analysis of protein-coding genetic variation in 60,708 humans. *Nature*. August 18, 2016. DOI:10.1101/030338.

11

L1 motivation et objectives - 13 September 2018

On a les données ...

```
... TTGTACATCTCATCTACTTATCGCTAGCAGCAGC  
TACTGATCGTAGTCCTCGTAGTCATTCATCGTAC  
TATCGATCGAGTCGATCGTAAATCGGGTAGTAGGCCGG  
GTGTCAATATAGCCTCTAGGGCTAGCAGCTGATCGATC  
TAGTCGTTCATGTCGATCGAGCTAGTCAGTCGTATCTA  
TACTAGCGACGATCTAGCGTACCGTAGCTATACTAC  
TCTGATATACTGCCGCTAGTAGCTACTGCAGCTGAC  
TGCTGACTGCTGACTGACGTAGCTGACATTGCTAGC  
TAGCTTACATCGCGATCGTAGCTAGCGATCGTACCGTAGC  
GCCTAGCGGACTCTGCGATCGTAGCTGCTGTAGTCGATT  
GTGCGATAGTCAGTCAGTCAGTCAGTCAGTCAGTC  
ACTGACGTCGACTGATCGACTGACTGACTGACTG  
ATGTCGTCGACTGACTGACCGCTGACTGACTG  
GTCGACTGATGACTGACTGCGCGTCAGTCAGTCAGTC  
CTGACTGACTGTCAGTCAGTCAGTCAGTCAGTCAGC...
```

12

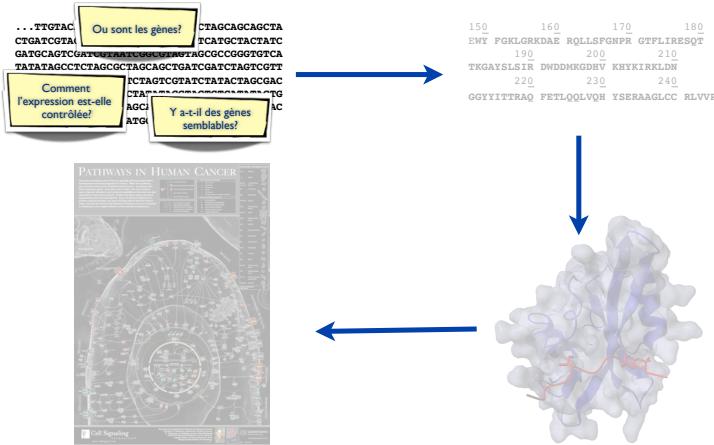
... et maintenant?

```
... TTGTACATCTCATCTACTTATCGCTAGCAGCAGC  
TACTGATCGTAGTCCTCGTAGTCATTCATCGTAC  
TATCGATCGAGTCGATCGTAAATCGGGTAGTAGGCCGG  
GTGTCAATATAGCCTCTAGCGCTAGCAGCTGATCGATC  
TAGTCGTTCATGTCGATCGAGCTAGTCAGTCGTATCTA  
TACTAGCGACGATCTAGCGTACCGTAGCTATACTAC  
TCTGATATACTGCCGCTAGTAGCTACTGCAGCTGAC  
TGCTGACTGCTGACTGACGTAGCTGACATTGCTAGC  
TAGCTTACATCGCGATCGTAGCTAGCGATCGTACCGTAGC  
GCCTAGCGGACTCTGCGATCGTAGCTGCTGTAGTCGATT  
GTGCGATAGTCAGTCAGTCAGTCAGTCAGTCAGTC  
ACTGACGTCGACTGATCGACTGACTGACTGACTG  
ATGTCGTCGACTGACTGACCGCTGACTGACTG  
GTCGACTGATGACTGACTGCGCGTCAGTCAGTCAGTC  
CTGACTGACTGTCAGTCAGTCAGTCAGTCAGTCAGC...
```



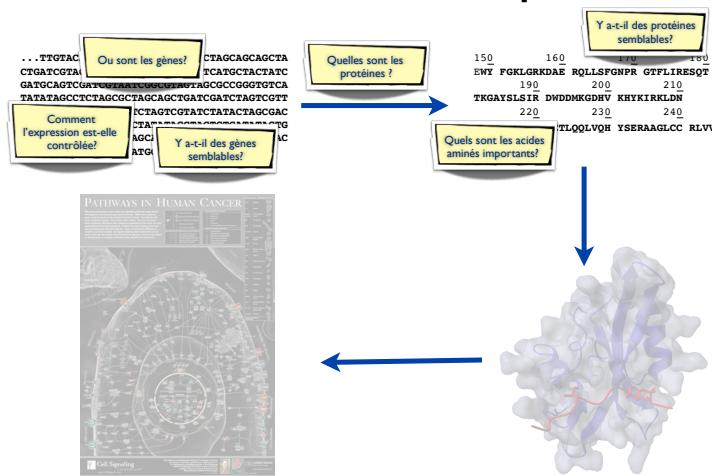
13

Le rôle joué par l'informatique



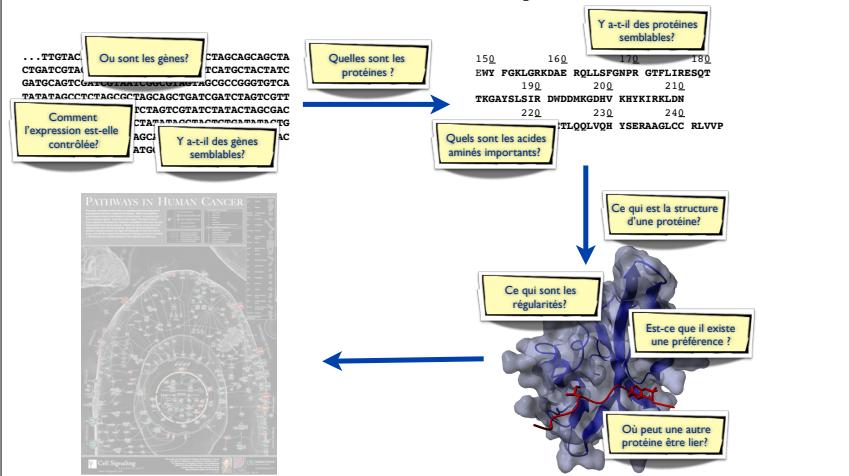
14

Le rôle joué par l'informatique



15

Le rôle joué par l'informatique



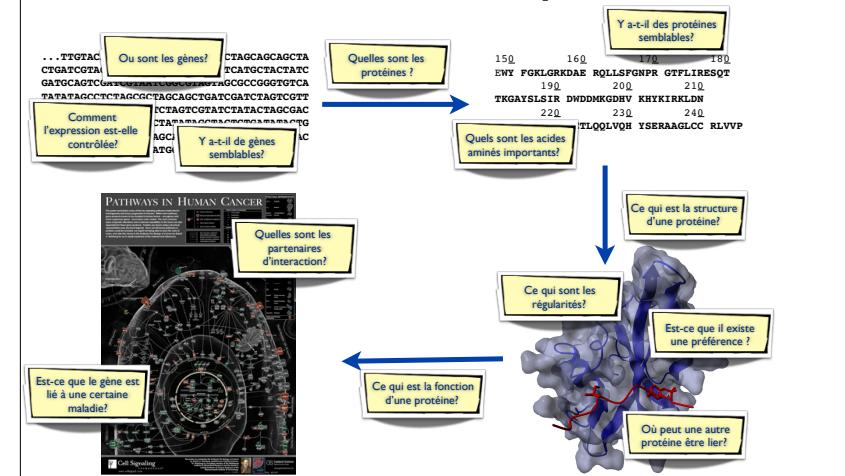
16

Objectifs du cours

- Comprendre le vocabulaire biologique et bioinformatique
- Comprendre les questions moléculaires et les techniques bioinformatiques qui ont résolu ces questions.
- Être capable d'expliquer et reproduire certains algorithmes
- Être capable de choisir entre les algorithmes connus pour résoudre un problème moléculaire
- Être capable de créer un nouveau algorithme qui peut résoudre une certaine question biomoléculaire
- Comprendre les publications dans le domaine.
- **Augmentez votre intérêt pour des autres sciences et la recherche interdisciplinaire**

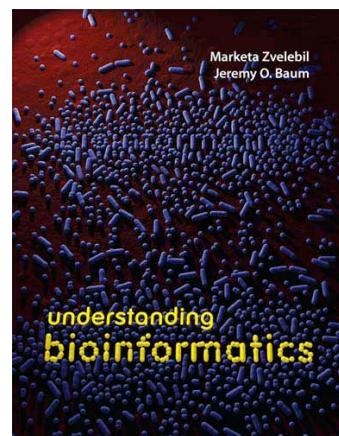
18

Le rôle joué par l'informatique



17

Référence

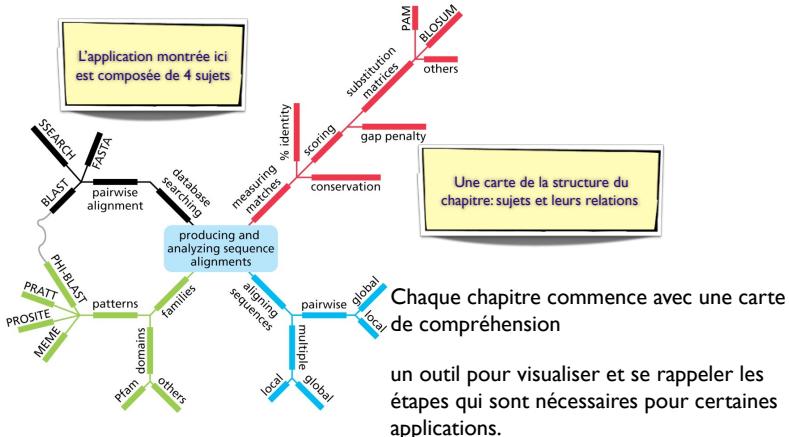


Site web: <http://www.garlandscience.com/textbooks/0815340249.asp>
Marketa Zvelebil, Jeremy O. Baum. Understanding Bioinformatics August 2007 Paperback: 978-0-8153-4024-9

19

- Part 1: Background Basics**
1. The Nucleic Acid World
 2. Protein Structure
 3. Dealing with Databases
- Part 2: Sequence Alignments**
4. Producing and Analyzing Sequence Alignments
 5. Pairwise Sequence Alignment and Database Searching
 6. Patterns, Profiles, and Multiple Alignments
- Part 3: Evolutionary Processes**
7. Recreating Evolutionary History
 8. Building Phylogenetic Trees
- Part 4: Genome Characteristics**
9. Revealing Genome Features
 10. Gene Detection and Genome Annotation
- Part 5: Secondary Structures**
11. Obtaining Secondary Structure from Sequence
 12. Predicting Secondary Structures
- Part 6: Tertiary Structures**
13. Modeling Protein Structure
 14. Analyzing Structure-Function Relationships
- Part 7: Cells and Organisms**
15. Proteome and Gene Expression Analysis
 16. Clustering Methods and Statistics
 17. Systems Biology
- Appendices: Background Theory
Appendix A. Probability, Information, and Bayesian Analysis
Appendix B. Molecular Energy Functions
Appendix C. Function Optimization

Cartes de compréhension



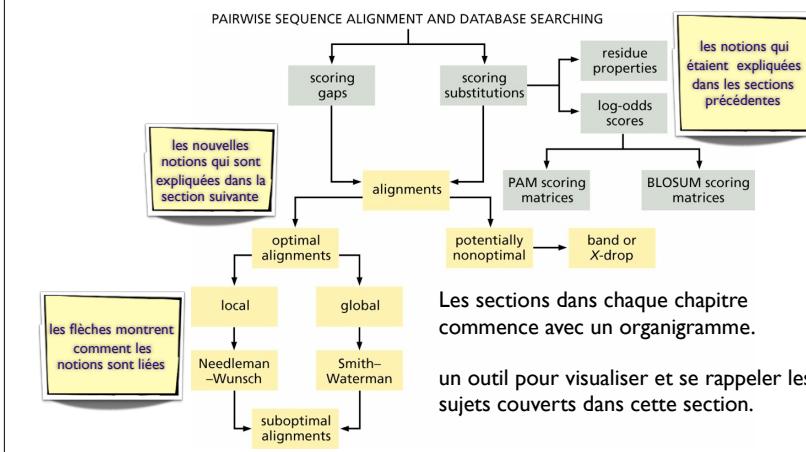
20

Programme (provisoire)

Date	Théorie (Lundi)	TP (vendredi)	Livre
17 septembre 2018	Introduction et Fondations moléculaire	No TP	5-44
24 septembre 2018	Alignment des séquences	P1	71-161
1 octobre 2018	Comment rédiger le rapport, Intro Jupyter et questions liées au projet 1	P1	
8 octobre 2018		DL P1	
15 octobre 2018	Alignment de groupes de séquences et Profiles	P2	165-219
22 octobre 2018	Modèles Markoviens Cachés et Profiles	P2	165-219
29 octobre 2018	Semaine Tampon		
5 novembre 2018	Questions liées au projet 2	DL P2	
12 novembre 2018	Trouver des séquences homologues	No TP	71-161
19 novembre 2018	Structures secondaires	P3	
26 novembre 2018	Arbres phylogénétiques	P3	223-312
3 décembre 2018	Questions liées au projet 3	DL P3	
10 décembre 2018	Session finale + explication examen	No TP	
17 décembre 2018	Semaine Tampon		

22

Organigrammes



21

Organisation

- Chaque **Lundi** entre 10h00 et 12h00
- Local Forum H
- **Réunions pour le portfolio:**
 - **Vendredi** de 10h à 12h dans 2 NO4.008
 - Mrs. Charlotte Nachtegaal
 - Assistance pour les trois mini projets

23

Portfolio

- Chaque étudiant construira pendant l'année un portfolio composé de 3 mini projets

Un **portfolio** ou **portefolio** est un dossier personnel dans lequel les acquis de formation et les acquis de l'expérience d'une personne sont définis et démontrés en vue d'une reconnaissance par un établissement d'enseignement ou un employeur.

- Utilisant Jupyter notebooks (python)
- Ecrire un rapport !!
- 50% de la note finale

24

Portfolios

- Date limite projet 1; **12 octobre 2017**
- Date limite projet 2; **9 Novembre 2017**
- Date limite projet 3; **7 Décembre 2017**

26

Portfolio

Le portfolio sera composé de 3 mini projets

- 1.Une implémentation de l'algorithme Needleman-Wunsch et l'algorithme de Smith Waterman qui sera comparée avec le logiciel LALIGN
- 2.Un algorithme qui construit des PSSM pour des ensembles de séquences alignées et qui sont utilisé dans l'alignement.
- 3.Une implémentation de l'algorithme GOR IV pour la prédiction de la structure secondaire des séquences.

25

IMPORTANT

- pour chaque projet créez un Jupyter notebook (python)
- Ecrivez un rapport scientifique (introduction, méthodes, résultats, conclusion)
- Utilisez des exemples pour illustrer votre code
- Ajoutez
 - des explications: le but du projet, le traitement des données
 - des figures
 -

27

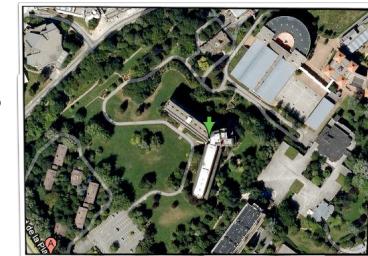
L'examen

- Examen écrit **janvier 2018 (3h)**
 - la théorie (transparents et livre)
 - Un exemple est en ligne sur le site web
 - 50% des points
- Continuation orale après l'examen écrit
 - +2 ou -2 !

28

Mes coordonnées

- Tom Lenaerts
- Bureau : 8ième étage,
2 NO 8.117
- téléphone ULB :
02/650 60 04
- courrier électronique: tlenaert@ulb.ac.be
- <http://www.ulb.ac.be/di/map/tlenaert/>



29