

Introduction

Francisco M. Couto

Data Processing

2019/20

<https://fenix.ciencias.ulisboa.pt/courses/pd-2254879305238291>


Página Inicial · Processamento de Dados · Faculdade de Ciências da Universidade de Lisboa - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Página Inicial - Process... x +

← → ↻ 🏠 🔒 https://fenix.ciencias.ulisboa.pt/courses/pd-2254879305238291 ... 🌟 📄 🗄️ ☰

FenixEdu Login



Página Inicial

Processamento de Dados PT / EN

Últimos anúncios


Bibliografia


3 Setembro 2019, 15:57 · Francisco José Moreira Couto

Introdução à Bioinformática Via Linha de Comando
<https://www.gradiva.pt/catalogo/46792/introducao-a-bioinformatica-via-linha-de-comando>

Data and Text Processing for Health and Life Sciences
<http://labs.rd.ciencias.ulisboa.pt/book/>

Corpo Docente

Francisco José Moreira Couto Responsável

Docente a definir

Página Inicial

Grupos

Avaliação

Bibliografia

Horário

Métodos de Ensino e Avaliação

Objetivos

Planeamento

Programa


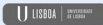
Turnos

Anúncios

Sumários

Appointments

- Francisco Couto
Tuesdays, 3.30pm, 6.3.08
- Lívio Rodrigues
Wednesdays, 2pm, 1.3.19



Faculdade de Ciências

Generic Plan

#T	Theory	Theory and Practical	#TP
1	Introduction to Data Processing		
2	Resources; Caffeine Example	Caffeine Example	1
3	Unix Shell; Web Identifiers	Unix Shell; Web Identifiers	2
4	Data Retrieval and Extraction	Data Retrieval and Extraction	3
5	Task Repetition	Task Repetition	4
6	XML Processing; Text Retrieval	XML Processing; Text Retrieval	5
7	Pattern Matching; Regular Expressions	Pattern Matching; Regular Expressions	6
8	Position; Tokenization; Relation Extraction	Position; Tokenization; Relation Extraction	7
9	Classes; URIs and Labels	Classes; URIs and Labels	8
10	Parent Classes; Ancestors	Parent Classes; Ancestors	9
11	My Lexicon; Generic Lexicon; Case insensitive	My Lexicon; Generic Lexicon; Case insensitive	10
12	Entity Linking; Large Lexicons	Entity Linking; Large Lexicons	11
13	Revisions	Revisions	12

TP13,TP11, TP12 - Francisco Couto

TP14,TP17, TP15, TP16 - Lívio Rodrigues

TPs

	Tuesdays	Fridays	Mondays	Tuesdays	Wednesdays	Thursdays	Fridays
Week	T11 Tue	T12 Fri	TP14,19	TP11,13	TP16	TP12	TP15,17,18
16/9/2007	1	1	-	-	-	-	-
23/9/2007	2	2	-	-	-	-	1
30/9/2007	3	3	1	1	1	1	2
7/10/2007	4	4	2	2	2	2	3
14/10/2007	5	5	3	3	3	3	4
21/10/2007	6	6	4	4	4	4	5
28/10/2007	7	Holiday	5	5	5	5	Holiday
4/11/2007	8	7	6	6	6	6	6
11/11/2007	9	8	7	7	7	7	7
18/11/2007	10	9	8	8	8	8	8
25/11/2007	11	10	9	9	9	9	9
2/12/2007	12	11	10	10	10	10	10
9/12/2007	13	12	11	11	11	11	11
16/12/2007	-	13	12	12	12	12	12

$$\text{Final Grade} = ((E + T) / (20 + T)) * 20$$

- **T = TPs classes**
 - Between 0 and 4
 - Individual for each topic
 - 11 topics available
 - Maximum 0.5 per topic
 - i.e. 8 topics gives the maximum grade
 - Periodical Evaluation
not redoable in the **special period** of exams
- **E = Exam**
 - Between 0 and 20
 - Written Exam
 - Estimated dates:
 - 10-01-2019 - 1pm
 - 03-02-2019 - 4.30pm

Examples

- E = 7.4 (minimal grade) and T = 4 (maximum grade)
 - $((7.4+4)/(20+4))*20 = 9.5$
 - Approved with 10
- E = 14 and T = 2
 - $((14+2)/(20+2))*20=14.54$
 - Approved with 15
- E = 17 and T = 4
 - $((17+4)/(20+4))*20=17.5$
 - Approved with 18
- E = 17 and T = 0
 - $((17+0)/(20+0))*20=17$
 - Approved with 17

Theoretical Classes Quiz

- Exam has 5 multiple choice questions
 - with a penalty for each wrong answer
- Quiz in the end of theoretical class
 - for each 3 correct quiz answers one penalty is removed
 - 11 classes, maximum 3 penalties removed

Bibliography

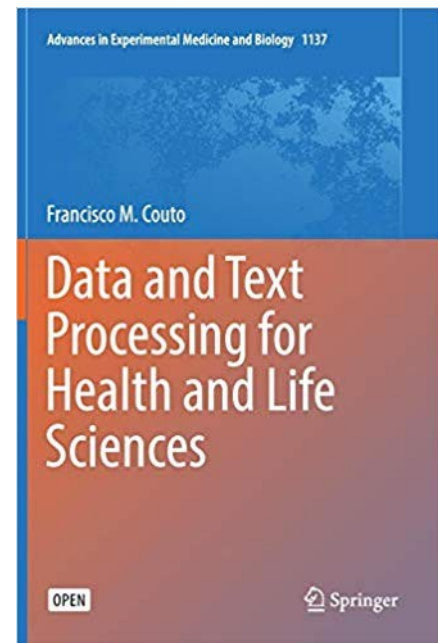
<http://labs.rd.ciencias.ulisboa.pt/book/>

- eBook
- Second Edition Draft
- Portuguese Version

Biblioteca de Biologia no C2

- Slides
- Workbook
- File Archive
- Test Script
- Video Tutorials

*Lançamento
23 Set 2019*



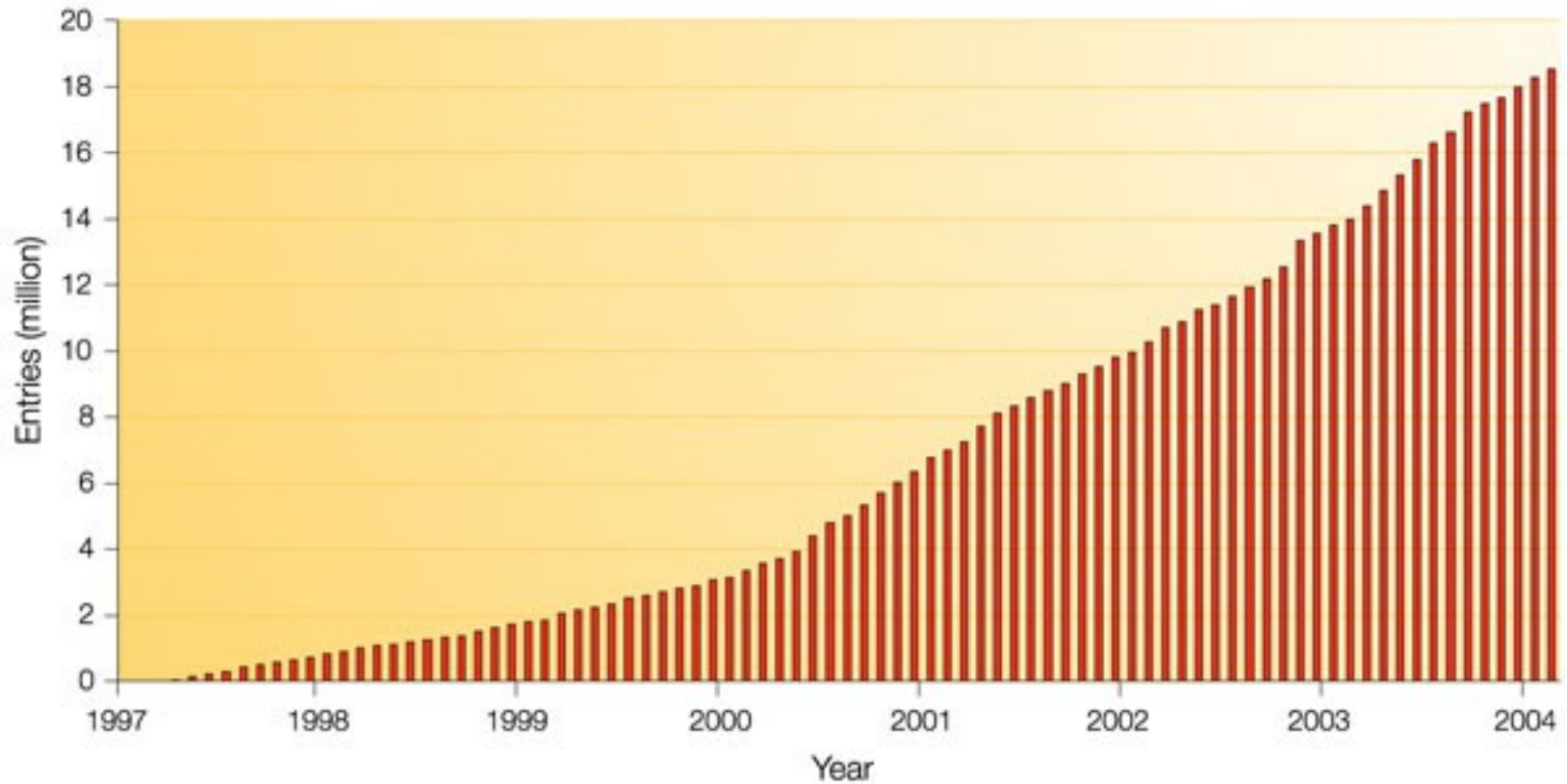
Software

- Spreadsheet application
 - LibreOffice Calc or Microsoft Excel
- Text Editor
 - notepad++ (Windows), TextEdit (macOS) or gedit (Linux)
- Terminal for shell scripting
 - Default in Linux or macOS
 - Windows:
 - Windows 10: Windows Subsystem for Linux
 - MobaXterm (Available at the labs)
 - Cygwin

WHY DATA PROCESSING?

Big Data

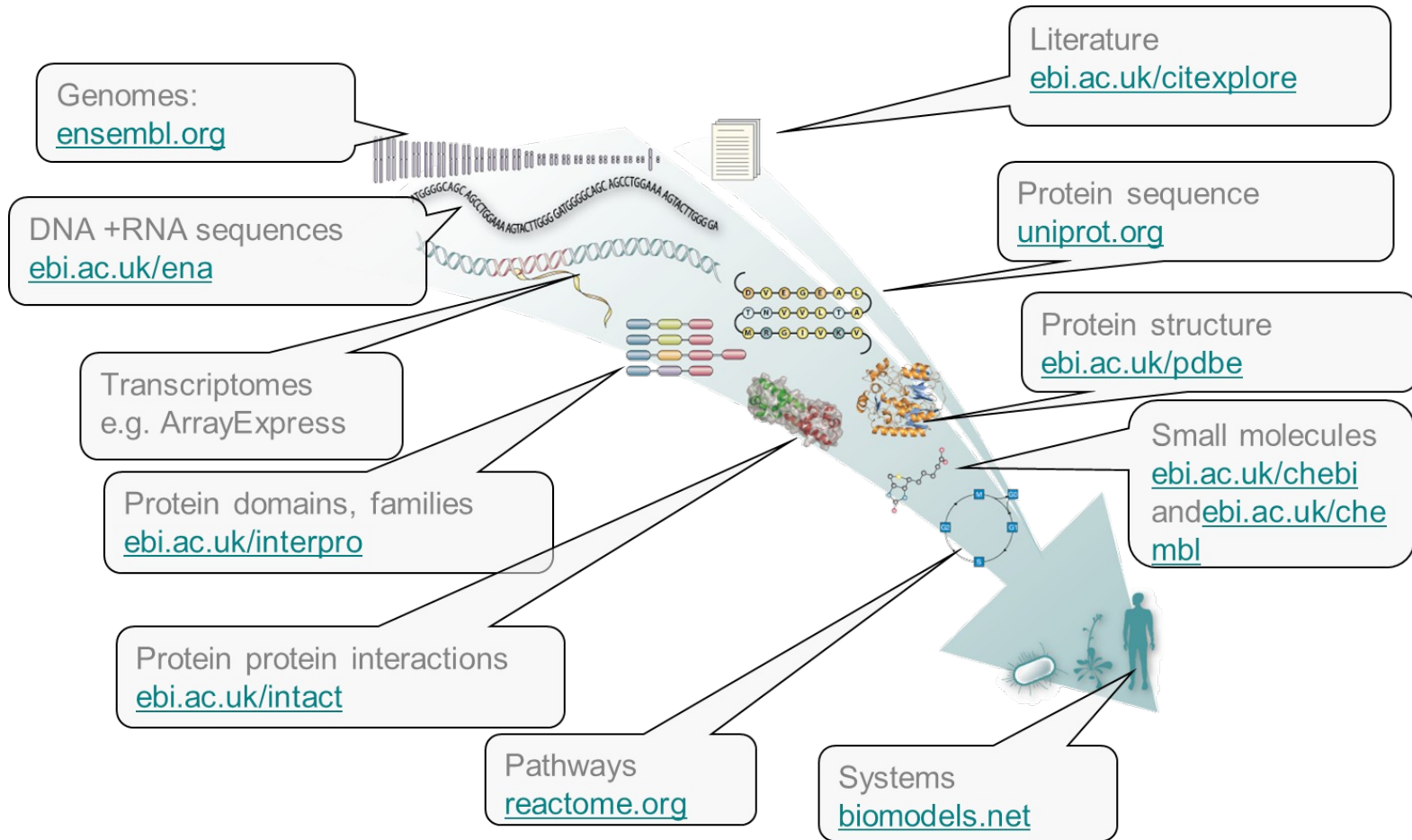
Growth of entries in DNA-sequence databases



Nature Reviews | **Genetics**

Heterogeneity

EBI: a data hub for bioinformatics in Europe



Main Goal

Learn computational techniques to:

Automate and **Replicate** the:

Web retrieval

process data and text files

semantic resources exploration

”the collection and manipulation of items of data to produce meaningful information” – [Wikipedia](#)

Master in Bioinformatics and Computational Biology

Since 2002 at FCUL

<http://bbc.fc.ul.pt>

Formação, séc. XXI Bioinformática - Exploração da Informação

28

centroatlantico.pt magazine • Outubro 2003

FORMAÇÃO, SÉC. XXI

Por **Francisco M. Couto**, Prof. Assistente na Faculdade de Ciências da Universidade de Lisboa, e **Mário J. Silva**, Prof. Associado na Faculdade de Ciências da Universidade de Lisboa

Bioinformática

- Exploração da Informação

A bioinformática é uma disciplina científica recente, cujo principal objectivo é a produção de conhecimento de interesse para a biotecnologia.

A biotecnologia tem como objectivo a produção e transformação industrial de materiais de natureza biológica. Os exemplos mais conhecidos de aplicações deste tipo de tecnologia são: o conhecimento do genoma de cada indivíduo para previsão de doenças e eventual tratamento pelos medicamentos mais apropriados; a manipulação genética de sementes permitindo obter plantas de maior rendimento; a substituição de materiais poluentes como os plásticos, combustíveis e antibióticos por materiais de origem biológica com um nível de poluição muito inferior.

A bioinformática estuda técnicas inovadoras de manipulação, gestão, e análise de grandes quantidades de informação biológica, permitindo aos cientistas extrair conhecimento a partir dessa informação. As fronteiras que limitam a variedade das aplicações da bioinformática são difíceis de identificar, pois esta integra conhecimentos de diversas áreas da ciência, como a biologia, a bioquímica, a estatística, a matemática e, naturalmente, a informática. O factor comum de todas as suas aplicações é o uso de sistemas computacionais no tratamento de informação biológica para a obtenção eficaz de importantes resultados científicos.

O grande interesse pela bioinformática nos últimos anos deve-se sobretudo à explosão da informação disponível proveniente dos esforços de sequenciação dos genomas de diferentes organismos. Esta informação permitiu o estudo de processos biológicos relacionados com o genoma, o que gerou ainda mais informação. Para a gerir têm sido criadas diversas bases de dados de grande dimensão (Tabela

Nome	Principais Características
Oracle	Muito utilizado pela indústria; Grande capacidade de dados; Sistema comercial.
PostgreSQL	Tipo de dados flexíveis; Vasto conjunto de funcionalidades; "Open Source".
MySQL	Facilidade na instalação e no uso; Rapidez na execução das operações "Open Source".

Tabela 1. Principais sistemas de gestão de bases de dados (SGDBs) utilizados em Bioinformática.

1). Por exemplo, a base de dados GenBank (<http://www.ncbi.nih.gov/GenBank>) disponibilizava através da Internet em Julho de 2003 cerca de 20GB só em sequências, resultante de um crescimento exponencial desde a sua criação. Este valor não conta com a informação descritiva de cada sequência, que é ainda de maior dimensão e de enorme importância.

A gestão destas bases de dados afigurou-se desde cedo como um processo complexo. A ausência de recursos para caracterização das entidades armazenadas foi infelizmente acompanhada pela utilização de métodos simplistas de anotação, causas da maioria das incongruências encontradas presentemente nas bases de dados. A integração de diversas fontes de informação é uma forma viável de completar e corrigir o conhecimento sobre as entidades biológicas, mas o objectivo, a estrutura, a nomenclatura e o tipo de informação variam nas diferentes bases de dados, tornando assim pouco viável a sua integração. Contudo, todo esse conhecimento biológico está presente na literatura, pois esta tem sido o meio preferi-

Career Opportunities

- An Explosion Of Bioinformatics Careers
 - in Science of June 13, 2014 DOI
<http://dx.doi.org/10.1126/science.opms.r1400143>
- Global Bioinformatics Market Will reach USD **12,542.4 million** in 2020
 - in Finances, December 31, 2014
<http://www.finances.com/analyses-and-opinions/analysis-opinions/49771-global-bioinformatics-market-will-reach-usd-12542-4-million-2020.htm>

Testimonials

- Experts agree that
 - the most successful bioinformaticians (and the ones who land the jobs) are those who have a **multitude of skills**
- At Roche,
 - “we offer continuous training in various areas and encourage our staff to attend conferences, publish, or pursue **higher degrees**”