ANALISYS OF BIOLOGICAL DATABASES

Bioinformatics

What Is Bioinformatics?

Information technology applied to analysis and management of biological data.

Implications in several fields:

Artificial intelligence

Robotics

Genome analysis

DNA sequences or proteins

tridimensional structures

Bioinformatics Definitions

- Fredj Tekaia at the Institute Pasteur offers this definition of bioinformatics:

  "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."

- According to Damian Counsell from bioinformatics.org

  “use *computers* to *store*, *retrieve*, *analyze* or *predict* the *composition* or the *structure* of biomolecules, including your genetic material (*nucleic acids*) and the products of your genes: *proteins*. These are the concerns of *classical* bioinformatics, dealing primarily with sequence analysis.”

“New” Bioinformatics

- **comparative genomics -** look for differences and similarities between all the genes of multiple species
- **functional genomics -** identifying gene functions and associations
- **proteomics -** catalogue the activities and characterize interactions between all gene products (in humans)
- **structural genomics -** crystallize and/or predict the structures of all proteins (in humans)

Computer Scientist/Mathematician Perspective

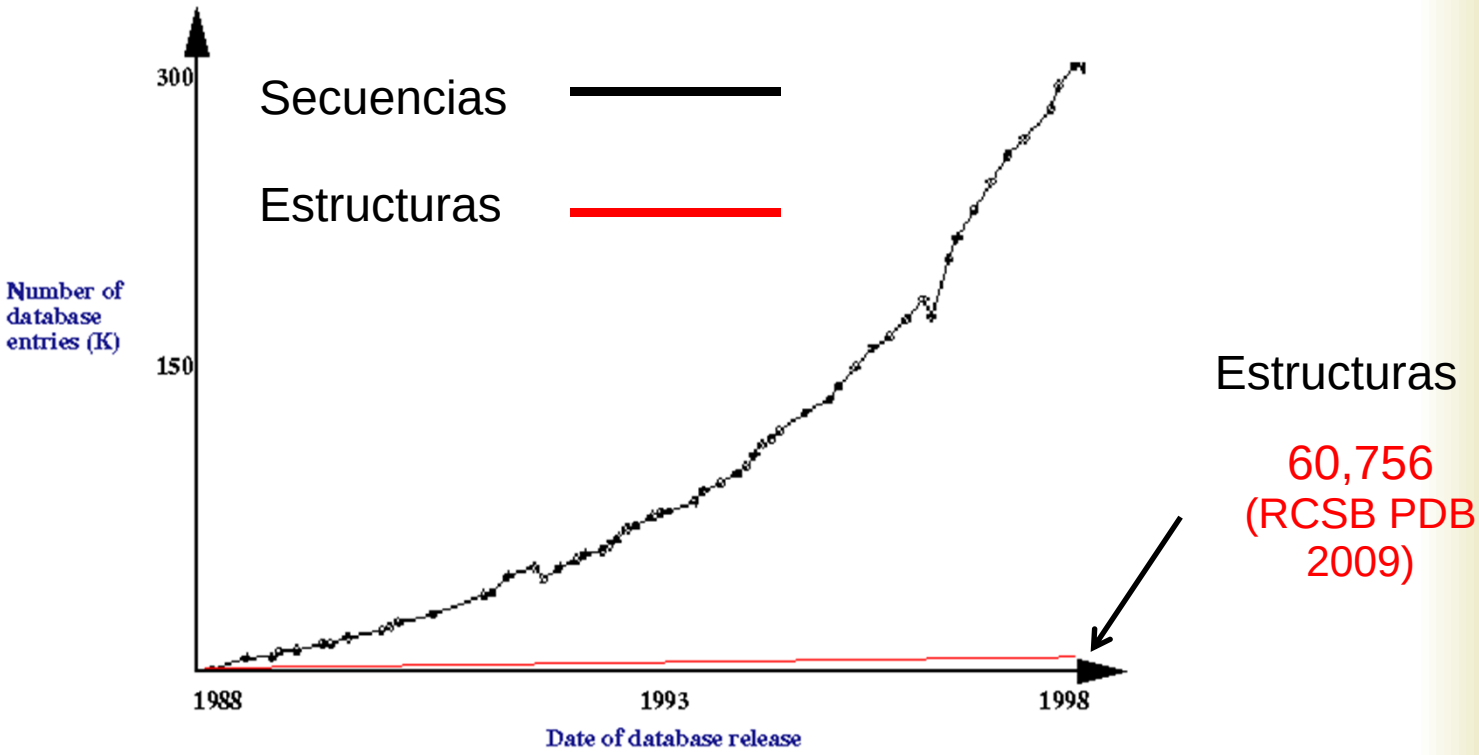
- Develop algorithms for sequence analysis
- Write programs using Perl/Python scripting languages
- Understand statistics used to align sequences

Chemist/Biologist Perspective

- Understand how to use existing bioinformatics software to
  - Retrieve gene sequence information (from **Genbank**)
  - Conduct similarity searches (using **BLAST**)
  - Align multiple sequences (with **Clustal W**)
  - Predict protein functional motifs (with **BLOCKS** and **ProDom**)
  - Display and compare 3-D protein structures (with **Spdbv**, **Pymol** and **Yasara**)

The legacy of the genome projects

SEQUENCE-STRUCTURE DEFICIT



Non-redundant growth of sequences during 1988-2002 (—) & the corresponding growth in the number of structures (—).

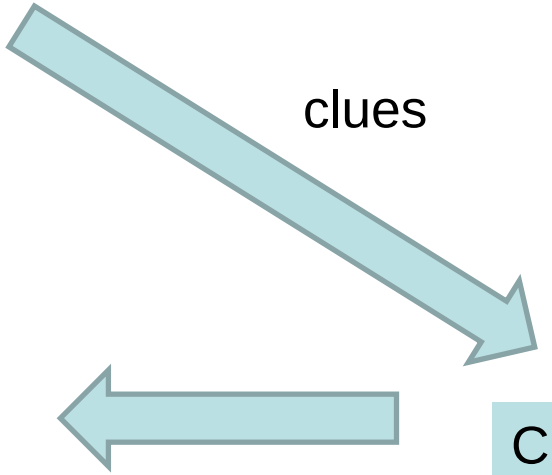
Importance of Bioinformatics

Due to sequence/structure deficit:

Rationalization of sequence information for:

Storage

Analysis tools



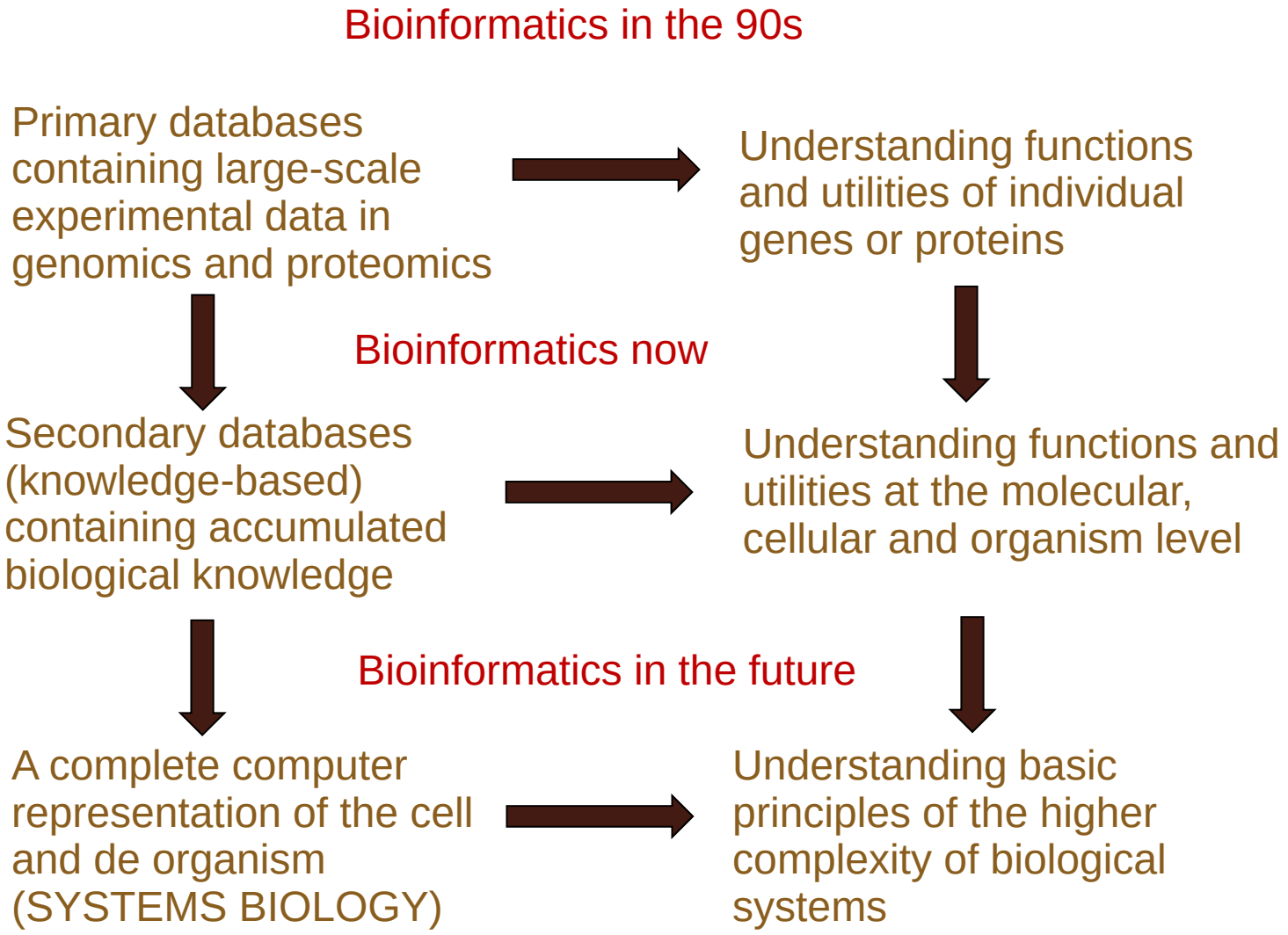
Biophysics and biochemical knowledge

Characteristics:

- Structural
- Functional
- Evolutive

Recognition and prediction

Evolution of Bioinformatics



Bioinformatics for us

The design, construction and use of software tools to generate, store, annotate, access and analyse data and information relating to Molecular Biology

Here we consider the use of Bioinformatics tools rather than their design and construction

Here we consider mainly the access and analysis of data and information items, but also their generation, storage or annotation (ADAN database)

Is Bioinformatics the Solution?

Bioinformatics **does not provide** definitive answers

Computational methods only provide "clues"

Results from Bioinformatics are generating questions that can be tested experimentally

The results obtained can serve as a laboratory guide to design experiments.

CHALLENGE: Design an analysis strategy which would be able to efficiently gather the biological knowledge in databases

Tight interaction between the “dry” and “wet” approaches is the best way to move forward

Course Outline

Bioinformatics and Internet
Genomes and sequence analysis
Proteomes and sequence analysis
Structural Bioinformatics

Course Objectives

Summarize the most important bioinformatics tools available that can be done easily with present computing programs and databases

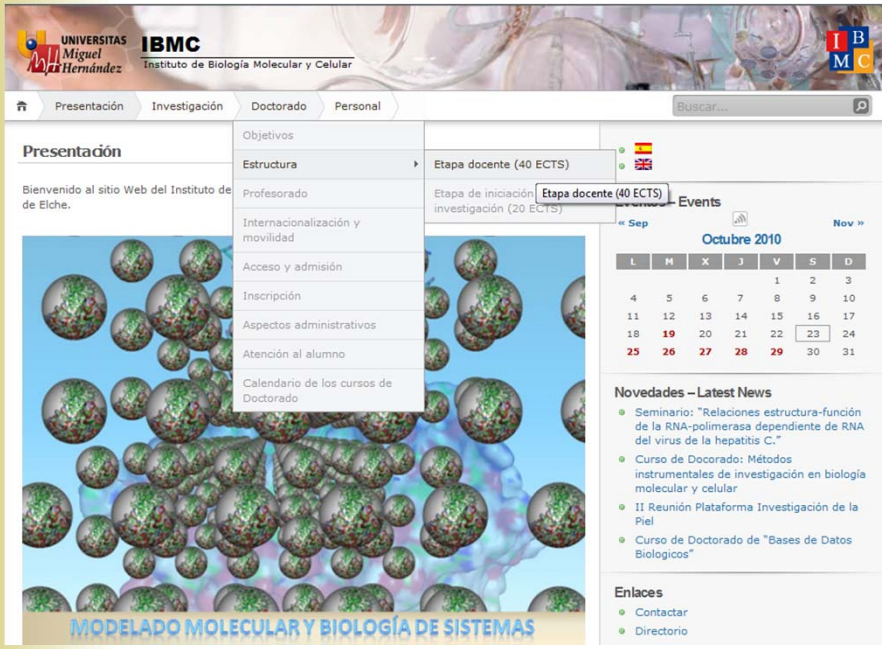
Analyze nucleotide and amino acid sequences

Analyze macromolecular structures

Infer their function in an adequate biological context

Course Material

http://shaker.umh.es/doctoradoBD/basesdatos_8786_1081/private/default.htm



Login: alumno
Passw: doctorado20102011

How To Pass the Course

1. Full attendance to the Course (25 Oct - 5 Nov 2010)

2. A well organized directory tree clearly containing
   a) Complete report (*MY_RESULTS.doc* in MS-Word) with the answers and explanations required in the exercises (in the root).

   b) The required downloaded files and the Word, Excel, html, etc. files asked in the exercises (distributed in the directories).

3. Final test with a real lab situation: Friday, 5 November 2010

4. It is also acknowledged to
   a) Have a positive attitude
   b) Try to improve yourself everyday
   c) Do as much as you can
   d) Try to do your best

Data Storage

ATTENTION: Your computer is FROZEN. This means that you can NOT store data permanently in C:\.

You can save data or install programs, BUT all changes will be lost upon computer reboot.

STORE YOUR DATA IN:

Remote Server as in “proteo.ibmc.umh.es”. Use **SSH** (sFTP) protocols to store data remotely

Remote Server from other sources

Locally: Memory sticks (you will need a big one) or drive T:\ in your local machine (thawed but dangerous!)

Doctorate Students 2010

AYALA TORRES, JOSE LEONARDO jatorresx@yahoo.com
BELLO GIL, DANIEL dabe_gil@yahoo.es
CALERO MARTÍNEZ, ALEX alexcaleromartinez@yahoo.com
DOMENECH MATA, ROSA MARIA rdomenech@umh.es
GARCIA VALTANEN, PABLO valtanen4@hotmail.com
GOMEZ-HURTADO CUBILLANA, ISABEL gomezhurtado_isa@gva.es
GUTIERREZ CASTRO, NOELIA AMALIA noelia@quimicasvinalopo.com
LOPEZ CORDOBA, AINARA ainara.lopezc@umh.es
LOPEZ PEREZ, MIRIAM miriam.lopez@umh.es
MARTINEZ LOPEZ, ALICIA alicia.martinez@umh.es
MARTIN MARCOS, RAQUEL r.martin@umh.es
MATHIVANAN, SAKTHIKUMAR sakthikumarmathivanan@gmail.com
MEDINA GALI, REGLA MARÍA reglita2000@yahoo.com
MELGAREJO JUAN, PATRICIA fashion69victim@hotmail.com
MONTOYA DIAZ, ESTEFANIA emontoya@umh.es
NEMESIO DE CASTRO, HENRIQUE henriquenemesio@gmail.com
PEREZ SANCHEZ, ALMUDENA hada_almu@hotmail.com
TOMAS MENOR, LAURA laura_tms20@hotmail.com
VEGARA GOMEZ, SALUD vegara84s@hotmail.com

LOGINS for proteo.ibmc.umh.es

E-mail: gregorio@umh.es

Login: your_nick (i.e.: gregorio)

Passwd: your_nick (i.e.: gregorio)

PASSWORD CHANGE REQUIRED

Computer Tools

PLATFORMS

Unix / Linux

PC

INTERNET

TCP/IP

FTP, **SSH**

WWW

HTTP, **HTML**, URL

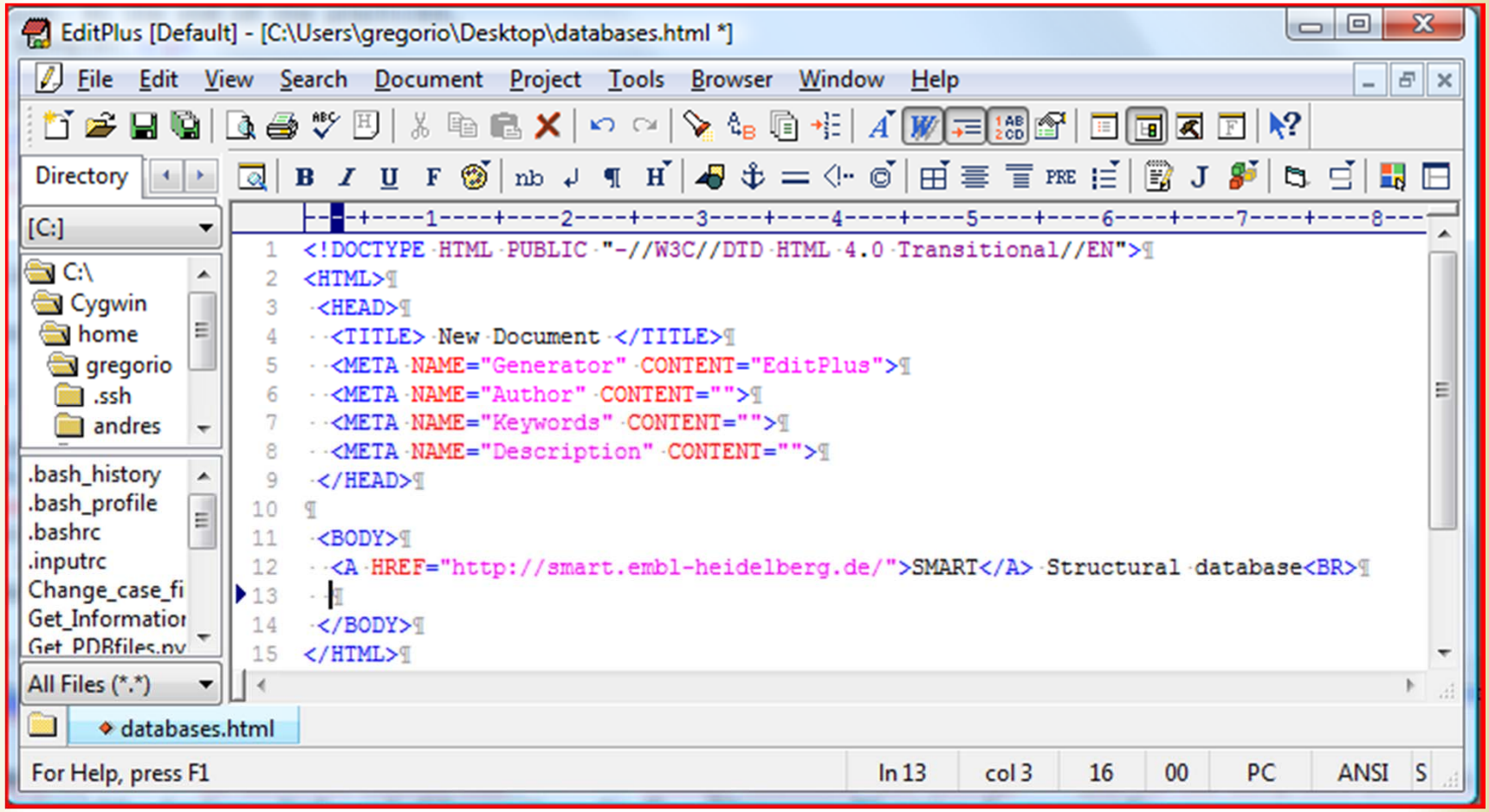
LOCALS

Local servers (programs and calculations)

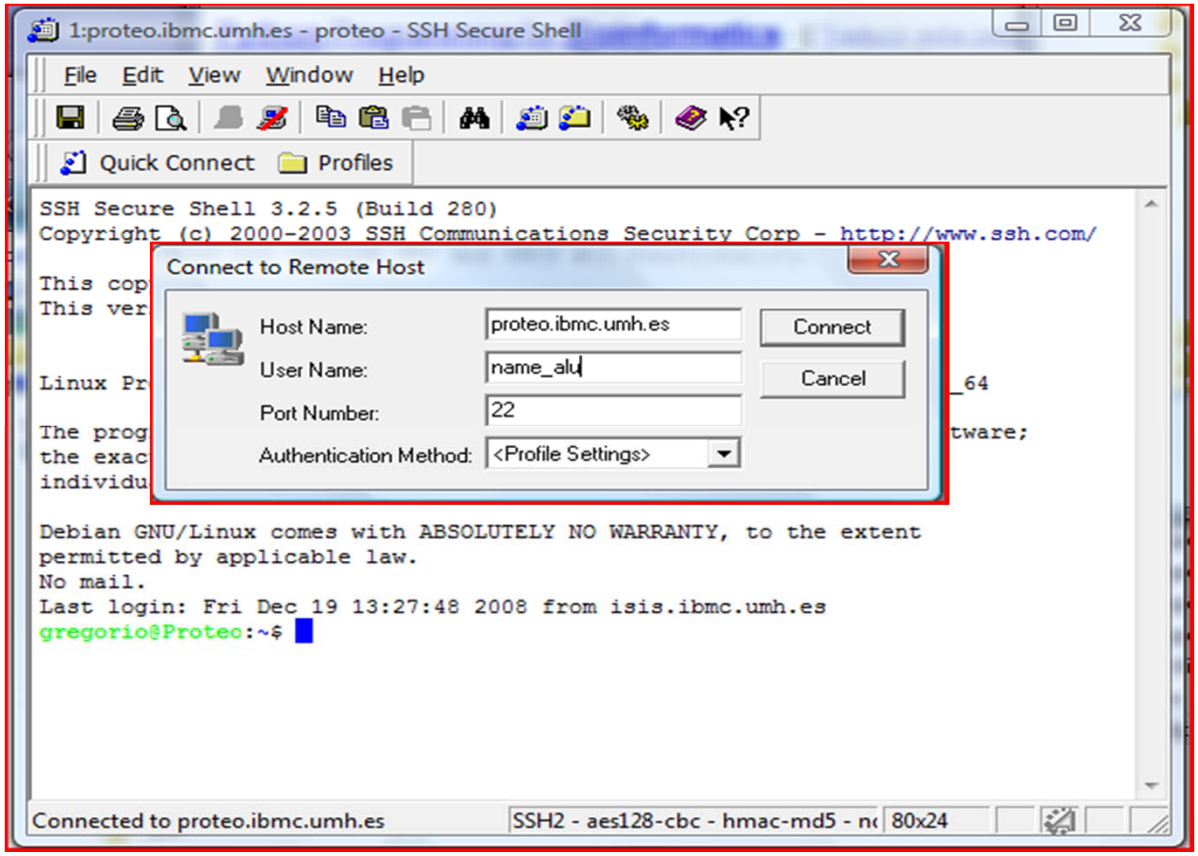
Remote control: X-Win32, Terminal Server

“Scripting” languages: **python**, perl

HTML Language



SSH: Execute Program



Changing password with command: *passwd*
USAGE=> your_nick@Proteo:~$passwd (and press *enter)*

General Comments

- use `pwd` and `ls` often

  If you are unfamiliar with the Unix command line and/or navigating around a hierarchical file system, use the `pwd` (print name of current/working directory) and `ls` (list directory contents) shell commands as much as you need to stay oriented in the file system. It's always helpful to draw a quick picture.

- use `man` often

  Use the `man` command copiously. Unix has a very useful on-line manual that you can read with the `man` command. For example, if you can't remember how to use the `ls` command to list a directory contents with file modification dates, type `man ls`. This will display the on-line manual page which describes the `ls` command and allow you quickly learn how do it. `man -k` is a useful option when you can't remember the name of the command you want to read about. (Look up `man -k` in the on-line manual by typing `man man`).

Unix Shell Commands Used

- ls
- pwd
- cd
- mkdir
- ..
- man
- setenv
- echo
- foreach
- >
- |
- cp
- ln -s

- cat
- more
- head
- tail
- wc
- grep
- sort
- awk
- sed
- vi or emacs

Making, changing and deleting folders

$mkdir folder_name (enter)
Creates a folder

$ls or ls –al (enter)
List the files w or w/o details

$cd folder_name (enter)
Changes the prompt to the desired folder

$pwd (enter)
Gives you the location in the tree (hierarchically)

$cd .. (enter)
Return to the parent folder

$rm file_name (enter)
Delete the especified file/s (DANGEROUS)

$rm –R folder_name (enter)
Delete the folder and files/folder inside (VERY DANGEROUS)

$whoami
Remind you who are you if you don´t remember well

Make a tree of folders

Mandatory folders:

- EXE => To execute programs and generate files or folders. Once you have had the results, you must move the data to an appropriate location.

(move implies copy to other location and delete the source)

- RESULTS=> Stores all the results obtained for the different exercises. This folder will contain the file: “MY_RESULTS.doc” and a collection of folders name P1 to P13 (the index, see main web page). Each of this folders will contain sub-folders named P1.1, P1.2, etc. (the exercises).

- EXAM=> Will contain the result of the exam the last day of the course. The answers and the correct folder structure will also be evaluated.

Optional folders:

- PERSONAL=> Store temporary results, data downloaded files, etc. It can be unstructured and contain one or more folder and sub-folders.

Some useful advice for file names and directories

Never use white spaces in sequence names. Use instead: “_” or “-”

Never use special symbols. Stick to plain letters, numbers and the underscore sign to replace the spaces.

Never use written accent used in other languages (Spanish á, é, í, etc.)

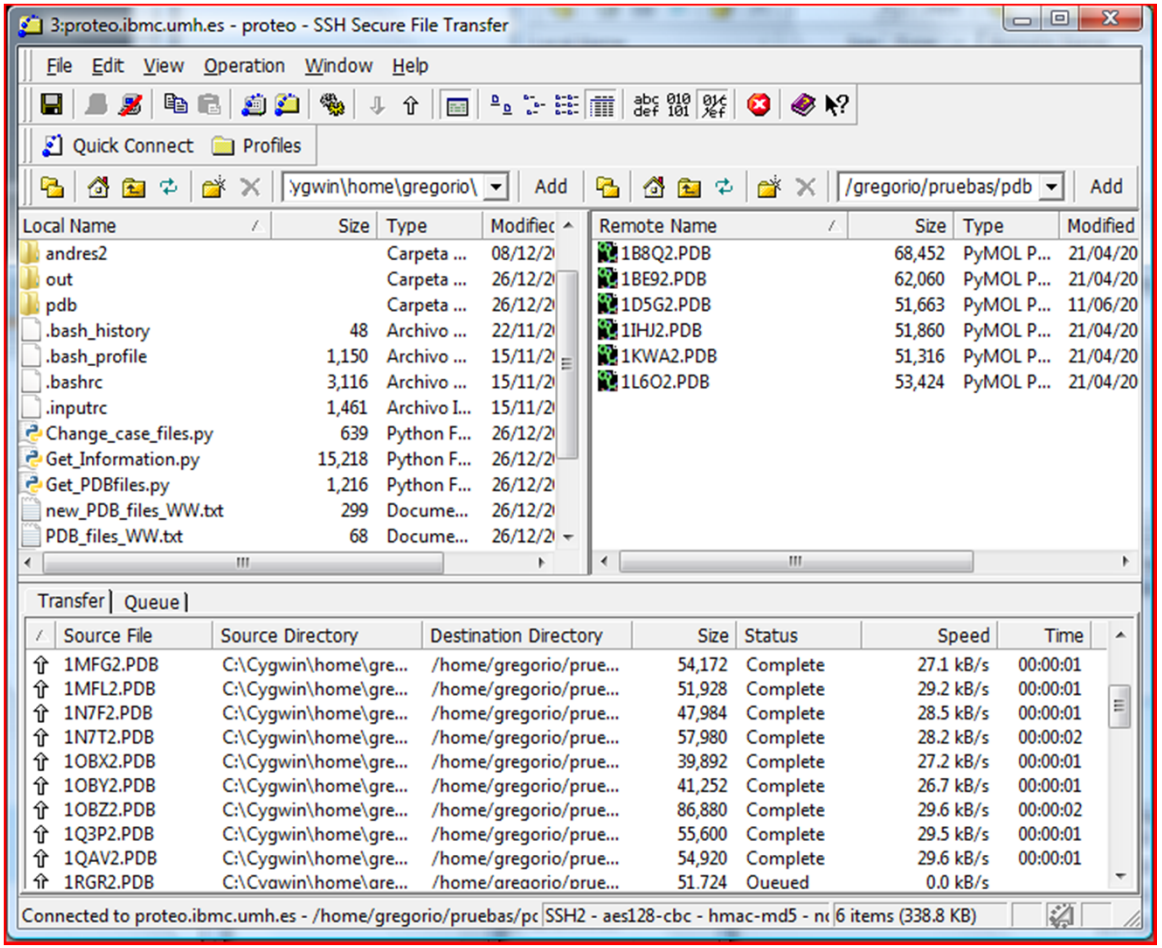
Avoid all other signs, specially the most tempting ones: @ # | * < >….

Never use names longer than 15 characters (rule of fifteen)

Never give the same name to 2 different sequences in your set. Some programs accept it, some others don’t

SSH: transfer a file

LOCAL
(your PC)



REMOTE
(server Proteo)

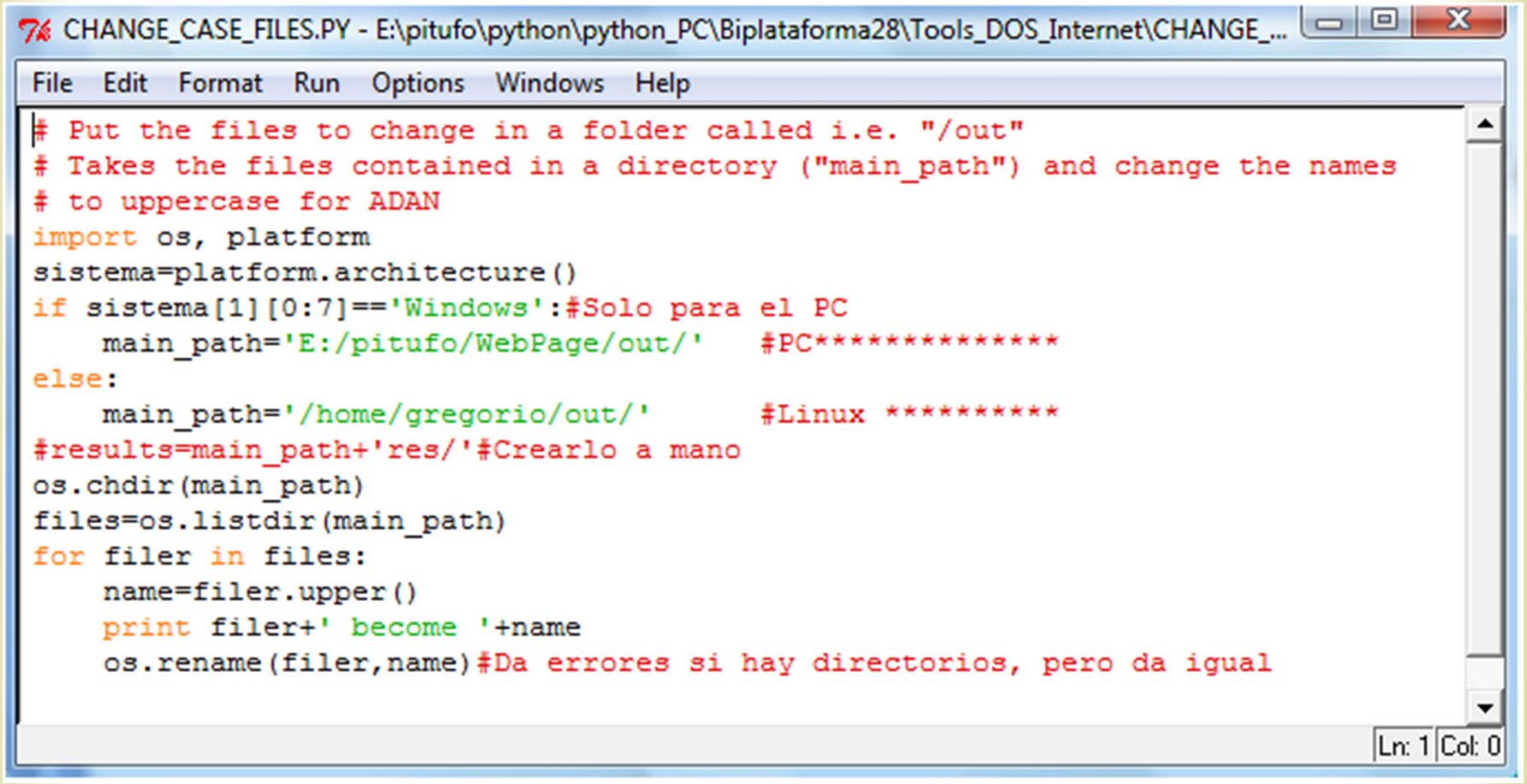
File can be tranferred from PC Windows (local) to Linux (remote), or between Linux folders (remote to remote)

What is Python?

- A portable, interpretive, object-oriented programming language
- Elegant syntax
- Powerful high-level built-in data types
  - Numbers, strings, lists, dictionaries
- Full set of string operations

- Previously used C++
- Scripting languages useful for bioinformatics
- Perl is “bioinformatics standard”
- Python is more “robust” for larger software projects

Python IDLE example



Useful Tutorials

- DNA from the Beginning
  - http://www.dnaftb.org/dnaftb/
- Python Tutorial
  - http://www.python.org/doc/current/tut/tut.html

Python Development Open-Source Software

- Python interpreter - will run on windows
- Allows you to Edit and prepare the script for use. It is an integrated environment for python called pythonwin.

FASTA Format

A very common format for sequence data is derived from conventions of FASTA, a program for FAST Alignment by *W.R. Pearson*.
Many programs use FASTA format for reading sequences, or for reporting results.

A sequence in FASTA format:

- Begins with a single-line description. A **>** must appear in the first column. The rest of the title line is arbitrary but should be informative.
- Subsequent lines contain the sequence, one character per residue.
- Use one-letter codes for nucleotides or amino acids specified by the International Union of Pure and Applied Chemistry (IUB/IUPAC).

>sp|P00674|RNP_Horse ribonuclease pancreatic

```
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHE
PLADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQ
TSQKERHIIVACEGNPYVPVHFDASVEVST
```