DBW – Databases and Web development

Josep Ll. Gelpí. gelpi@ub.edu

Aims

- Review a number of technologies to handle bioinformatics data:
  - Computer communication, design of web applications, basic database design and optimization.
  - This is NOT a programming course, it is about designing and building applications in an heterogenous scenario

- The final objective is to built a **fully operative application** using the appropriate combination of the techniques reviewed.

Bioinformatics & Internet

- Bioinformatics Tools and data should be available through web

- Ex. Nucleic Acid Research reviews:
  - Database Issue (January) 1170 DBs
  - Web Server Issue (July) 1200 Servers

NAR Database issue recommendations for authors

- “The pre-submission enquiry must present a working web accessible database “
- “The quality, quantity and originality of data as well as the quality of the web interface are the most important. Good data with a poor interface or vice versa are never sufficient for consideration. “
- “Do get a domain name for your website. URLs to specific IP addresses/ports are unlikely to stand the test of time.”
- (…)

Nucleic Acids Research, Volume 35, Issue suppl_1, 1 January 2007, Pages D1–D2

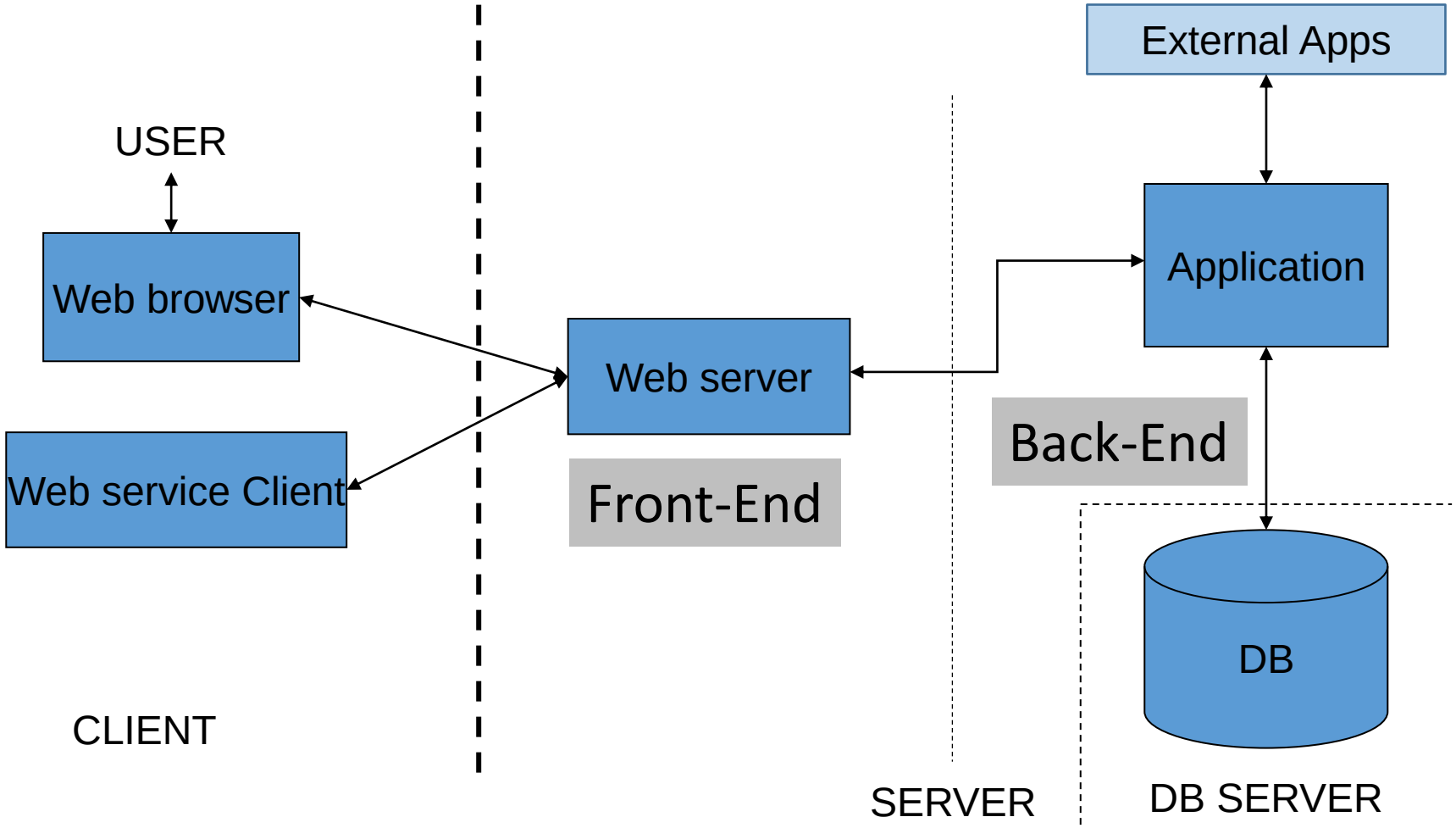
https://academic.oup.com/nar/article/35/suppl_1/D1/1088333

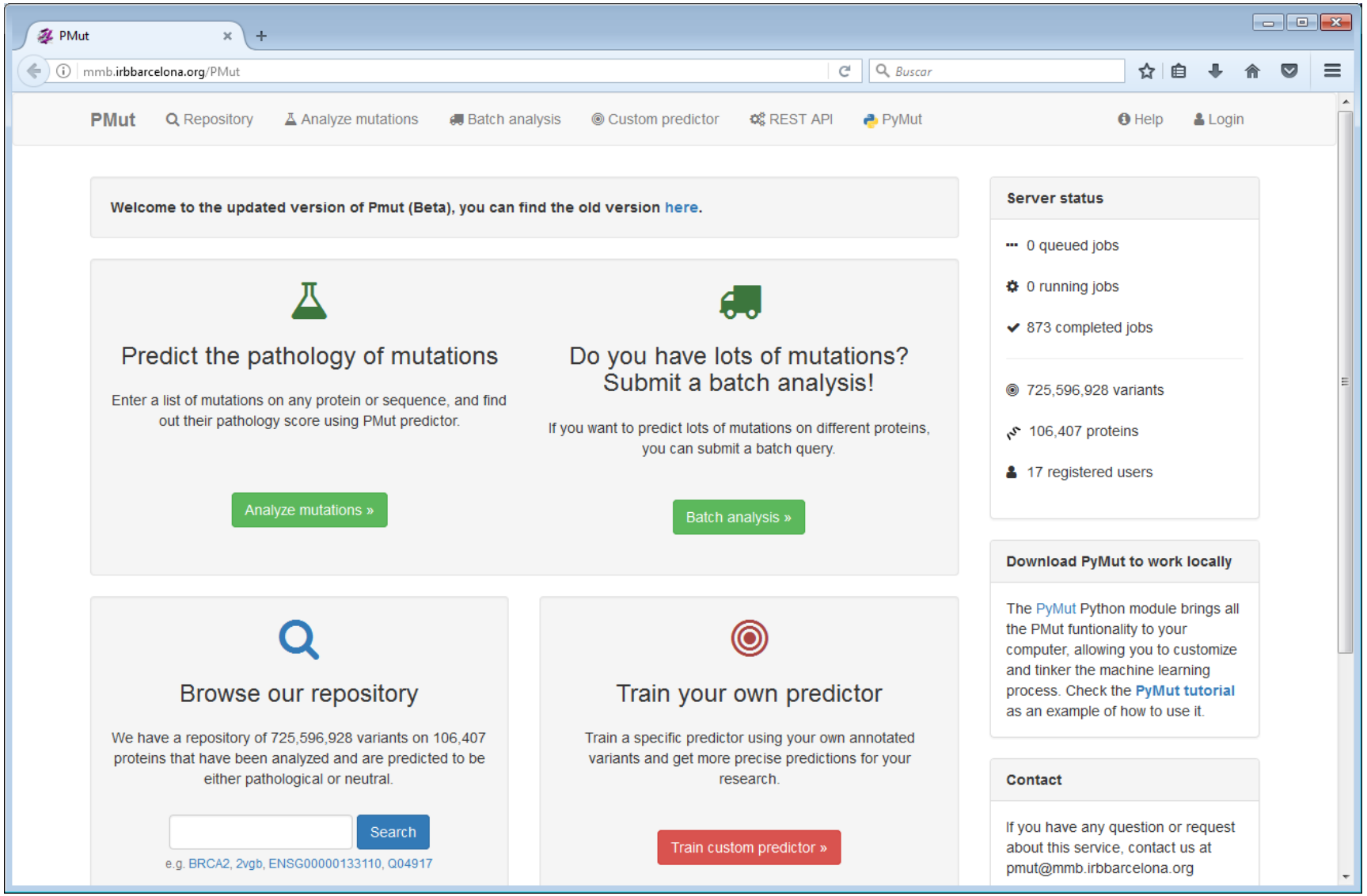
Web applications by access type

- Web interfaces
  - Provide a user friendly interface (web based) to “human” users
    - Users known how to use the interface
    - There is no need to install software
    - Single operations (no large scale)
    - Must adapt to navigation uses (low latency, synchronous answers,…)
- Web services & APIs
  - Provide a programmatic interface (using Web protocols)
  - Intented to interact with software, not with humans
    - Well-defined data formats required.
    - Adequated for large scale operations
- Modern applications will normally offer both
  - Web frontend is normally just another client of the API’s

- Application styles
  - Access to data
    - Friendly interface to data repositories
  - Web Interfaces to stand-alone software
    - Collect input parameters and redirect output
  - Workbenches (e.g. Galaxy)
  - On-purpose applications & DBs

Web application anatomy

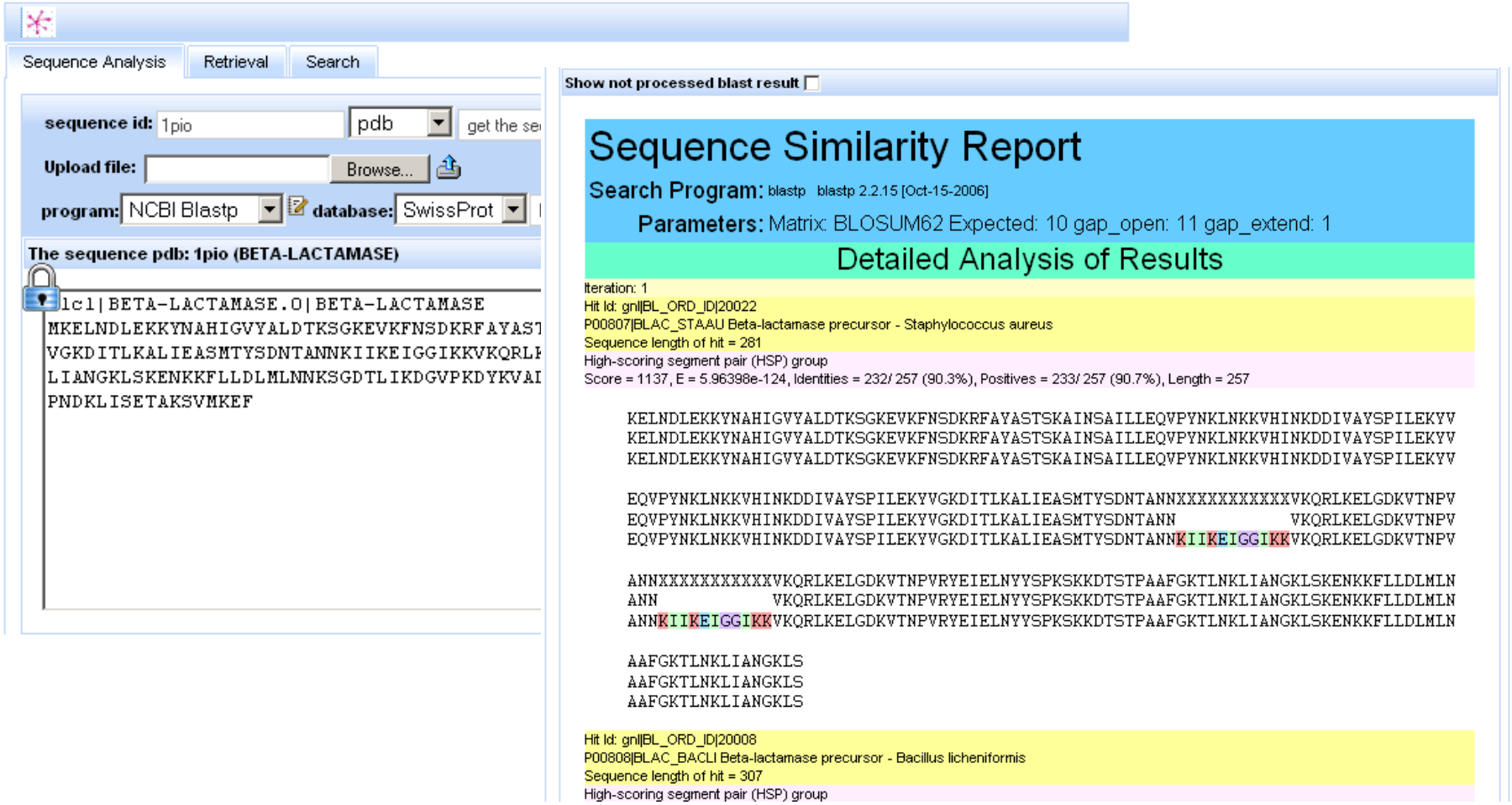


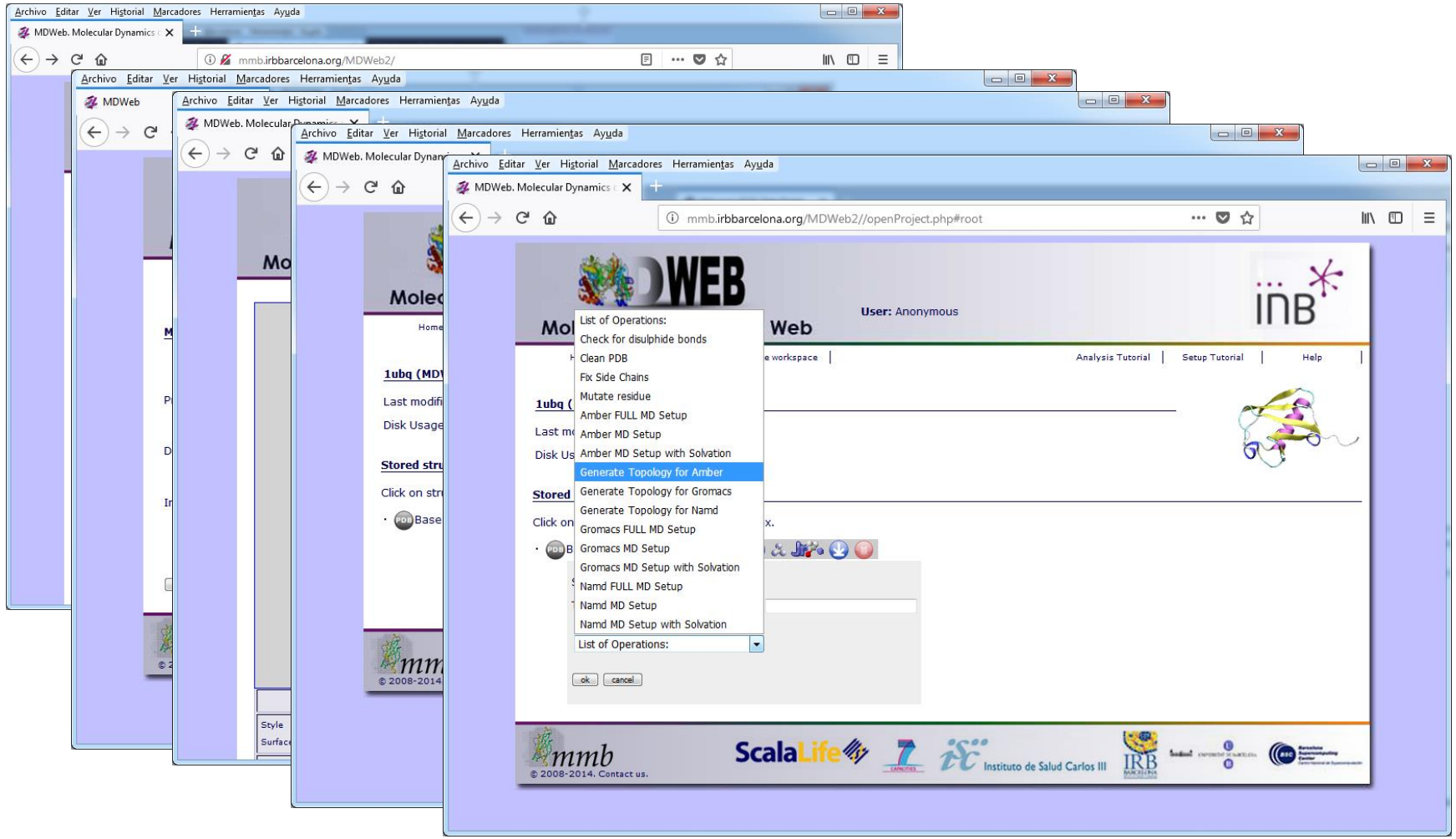


López-Ferrando *et al.* **Nucl. Ac. Res** 2017


Welcome | ICGC Data Portal
https://dcc.icgc.org
Buscar
Microsoft Office Home
Documentos de Google
Hojas de Cálculo de G...
EndNote
Aul@-ESCI: Entrar al si...
Calendario y horarios
Campus Virtual de la ...
MICINN - Sede Electró...
Login
ICGC Data Portal
Cancer Projects
Advanced Search
Data Analysis
DCC Data Releases
Data Repositories
e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049
About Us
The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.
To access ICGC controlled tier data, please read these instructions.
New features will be regularly added by the DCC development team. Feedback is welcome.
PCAWG
PanCancer Analysis of WHOLE GENOMES
The Pancancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from
https://dcc.icgc.org/icgc-in-the-cloud
Consortium.
Data Release 26
Dec 7th, 2017
Donor Distribution by Primary Site
Cancer projects 76
Cancer primary sites 21
Donors with molecular data in DCC 17,440
Total Donors 20,383
Tutorial
EXAMPLE QUERIES
1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available
ICGC International Cancer Genome Consortium
in the cloud
ICGC data is now available on commercial and
Go to ICGC in the Cloud Home

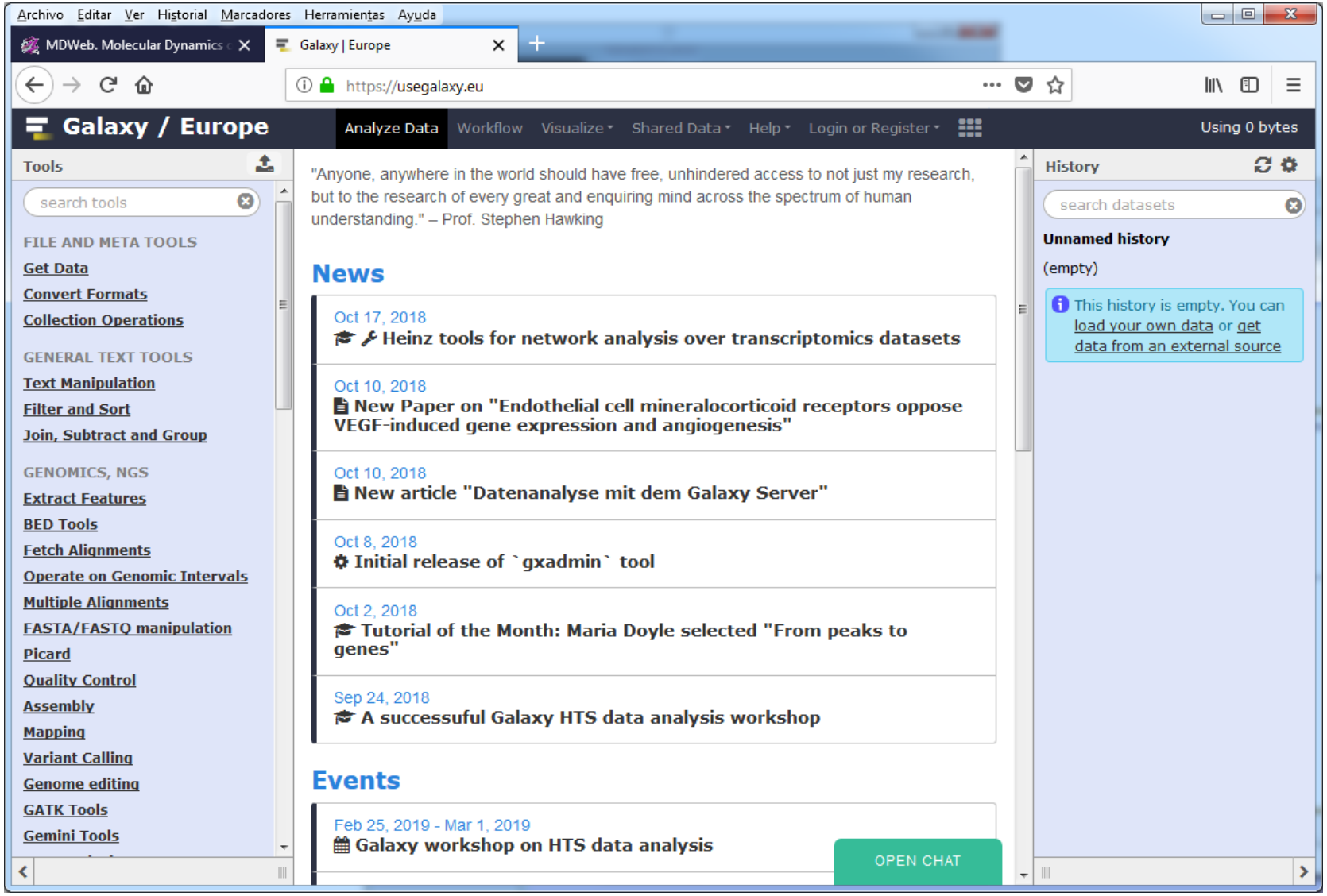
Web interfaces to apps.





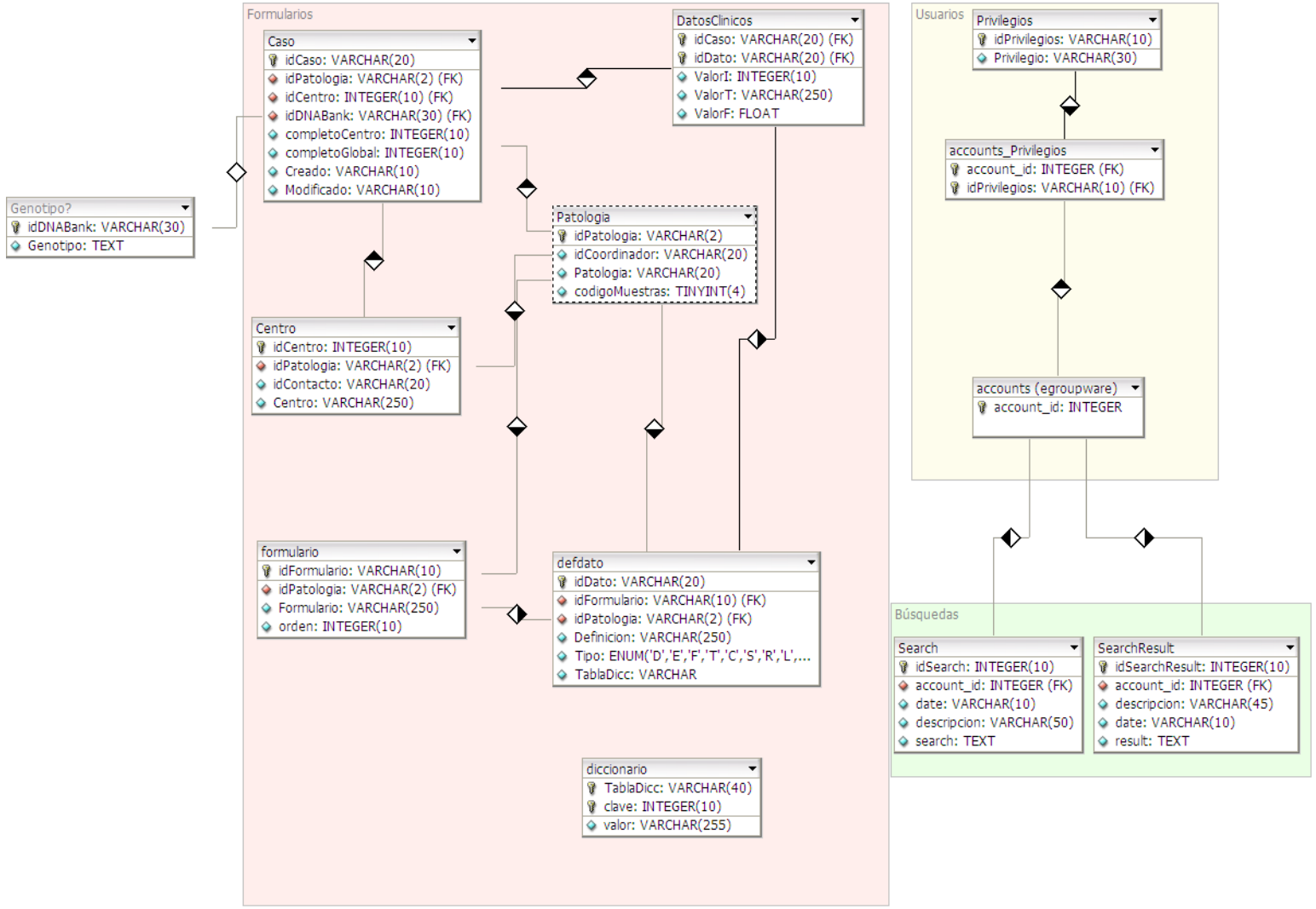
http://mmb.irbbarcelona.org/MDWeb2

Bioinformatics. 2012 28(9):1278-9.
doi: 10.1093/bioinformatics/bts139



https://usegalaxy.eu/

Special purpose applications & DBs



IMID-KIT [imidclinica] - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos

Dirección http://inb.bsc.es/imid_kit/egroupware/imidclinica/index.php

Google Marcadores Pop-ups permitidos Corrector ortográfico Configuración

imidclinica

Referencia:
Especialidad:
Centro:

Form
Codigo Banco AD
Datos Centro Co
Datos Completos

Cuestionario cli

A. Criterios de In
B. Criterios de Ex
C. Antecedentes
D. Variables Clíni
E. Variables Clíni
F. Enfermedades Asociadas (Diagnosticadas y Documentadas) (0/36)
G. Tratamiento (0/155)
G1. Tratamiento Etanercept y Adalimumab (0/48)
G2. Tratamiento Rituximab y Anakinra (0/49)
H. Actividad de la Enfermedad en el Momento de la Extracción (1/75)

Cuestionario epidemiologico

Extracción (0/0)
Actividades (0/0)
Hábitos (0/0)
Cuestionario Dem
Estudios (0/0)
Cuestionario Gene

imidclinica

Busqueda de Casos

- Ref. Caso contiene
- Ref. Banco DNA contiene
- Datos completos centro
- Datos completos
- Centro(s)
  - Servicio de Dermatología, Hospital Universitario Gregorio Marañón (M
  - Servicio de Dermatología, Hospital General Universitario de Valencia
  - Servicio de Dermatología, Hospital Universitario 12 de Octubre (Madr
  - Servicio de Dermatología, Complejo Hospitalario Juan Canalejo (A Co
  - Servei de Dermatologia, Hospital de la Santa Creu i Sant Pau (Barcel

Cuestionarios clinicos

Selecciona los campos a incluir en la busqueda

[Expandir Todos] [Colapsar Todos] [Seleccio

Selecciona los campos a incluir en la busqueda

[Expandir Todos] [Colapsar Todos] [Seleccionar] [Limpiar]

- Reumatología
  - A. Criterios de Inclusión
  - B. Criterios de Exclusión
  - C. Antecedentes Familiares
    - Psoriasis — si no (Cualquiera)
    - EII — si no (Cualquiera)
      - Crohn
        - Parentesco: padre, madre, hermanas/hermanos (Cualquiera)
      - Colitis Ulcerosa
    - AIC — si no (Cualquiera)
    - Otras Enfermedades — si no (Cualquiera)
  - D. Variables Clínicas y Biológicas Articulares
  - E. Variables Clínicas Extra-Articulares
  - F. Enfermedades Asociadas (Diagnosticadas y Documentadas)
  - G. Tratamiento
  - G1. Tratamiento Etanercept y Adalimumab
  - G2. Tratamiento Rituximab y Anakinra

imidclinica

Busqueda de casos

#1	Reumatología > C. Antecedentes Familiares > Psoriasis: si	3	Resultado	Grabar set
#2	Reumatología > C. Antecedentes Familiares > EII > Crohn: SI	0		
#3	Reumatología > C. Antecedentes Famil si			
	Operadores posibles: O, Y, NO			

[Grabar búsqueda actual] [Nueva Busqueda]

imidclinica

#4. #1 Y #3:

Num Casos: 2

ID.	Especialidad	Centro	Datos Cuest.	Id DNA Bank	Datos Epid.	Compl. Centro	Completo
30112345	Reumatología	Unitat de Recerca de Reumatologia, Hospital Universitari Vall d'Hebron (Barcelona)	22/492		0/0		
3012345	Reumatología	Unitat de Recerca de Reumatologia, Hospital Universitari Vall d'Hebron (Barcelona)	25/492		0/0		

Volver a la búsqueda

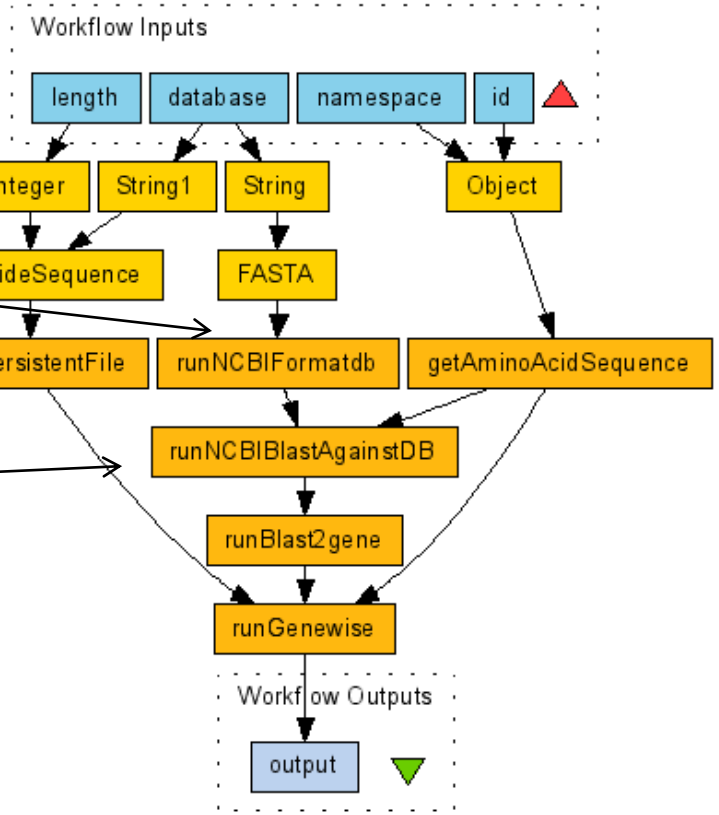
Bioinformatics web-services and workflows

Input: Protein Id and DNA genomic sequence

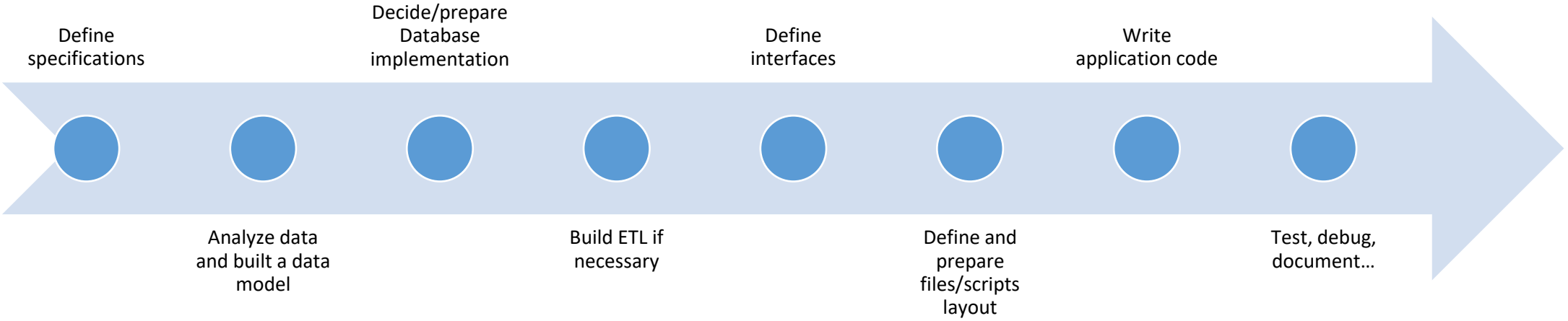
Building of Blast Database from DNA seq.

BLAST Search

Run GeneWise to detect gene structure



Building a (web) application, usual steps



Course logistics

- Web site(s)
  - Course materials:
    - http://mmb.pcb.ub.es/formacio/
  - Personal sites:
    - http://mmb.pcb.ub.es/formacio/~uXXXXXX

- Server
  - SSH Access
    - ssh mmb.pcb.ub.es –p 22122 –l uXXXXXX
    - Password dbw_uXXXXXX
  - SCP
    - scp –P 22122 uXXXXXX@mmb.pcb.ub.es …
  - MySQL/MongoDB Access
    - Localhost only
    - DBs on demand

Software to install

- Ideally Linux (may need root privileges)
- From Linux distributions
  - A Web server
    - Apache (with PHP 7.x)
    - Nginx (better for Python apps)
  - MYSQL server
  - MYSQL Workbench or phpMyAdmin
- Your preferred software code editor
- MongoDB (optional)
  - Install drivers for PHP/python if needed

Evaluation

- Exercices, in-class projects (20%)
- Personal web site (20%)
- Web application project (60%)
  - Progress presentations
  - Fully operative web application using DBs

Evaluation

- Web application project
  - 3-4 people / group
  - Free subject (bioinformatics preferred)
  - **Should include DB management, web interface, users' management** (Mysql or MongoDB)
  - May use **fake data if necessary**
  - Available at the personal web sites of all team
  - Preferred languages: PHP, Python
  - Source code at github or equivalent

Evaluation

- Web application project
  - Steps (Deadlines):
    - Initial specification (Presentation 21st Jan)
    - Data analysis & Database design (Presentation 31st Jan)
    - Project prototype Demo (Presentation 9th Feb)
    - Mid development review (meeting around 22th Feb)
    - Final application (End of Term)
- Fully Installed and functional on course server
  - PHP projects will use Apache
  - Python projects will use uwsgi/nginx (Flask temp server not acceptable)

Basic computer communication protocols

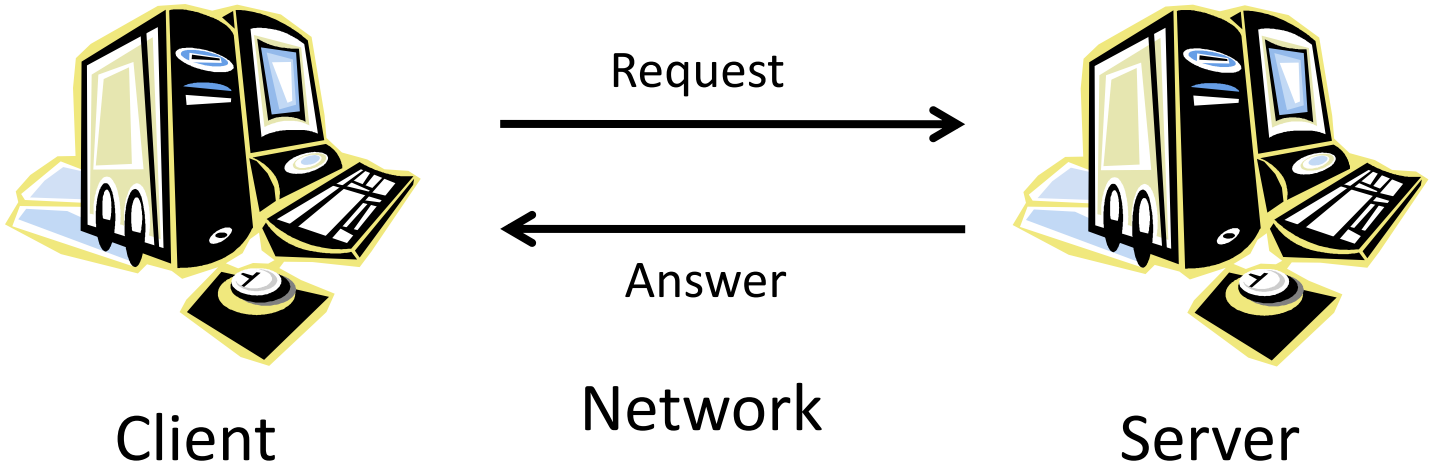
Aim & Outline

- Understand the basic components of computer communication protocols
  - Concepts of client and server
  - Addressing servers and data
    - Computer addresses (MAC Address, IP Address, DNS)
    - Ports
    - Resource identification: URL/URI concepts
  - Client/server transactions
    - HTTP protocol

Present internet

- Huge network of computers using common communication protocols (TCP/IP, HTTP)
- Distributed, no central servers
  - (Well, not really true in bioinformatics)
- Common languages: HTML/CSS/JS (XML, JSON)
- Content originally static, but dynamic behaviour is possible through web applications

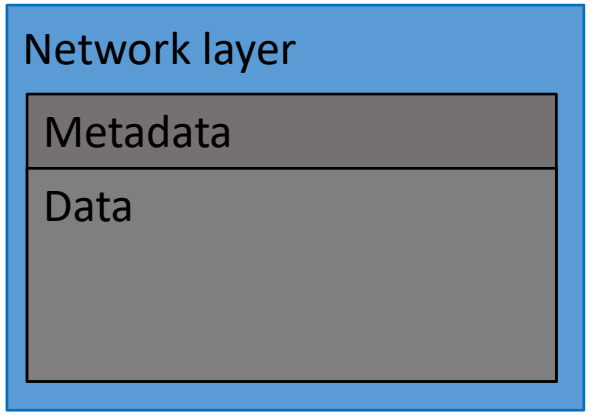
Components



- Client and Server logic and physical addresses
- Data
- Data meta-information
    - Nature of data
    - Request (what to do)
    - Applications involved (email, web, etc.)

How it works: TCP/IP

- Packet switching
  - Packet switching breaks the signal in small fragments (“packets”) each of them containing the complete information about source and destination
    - Packets can share a single communication line
  - Users have the idea of a dedicated line but, in fact, it is not. Of course, the bandwidth is limited.
- Computers connected to internet should have addresses/ports
  - MAC Address: Address of the physical interface
  - IP Address: Unique address of the computer
  - Unique Host name
  - Ports to point to specific applications



IP addresses & Host names (DNS services)

- Allow to find destination irrespective of the nature of the network media.
- Each device has a **“unique” IP address**
- IPv4: 32 bits (4 x 1 byte (0-255) numbers)
  - Max: 2^{32} : aprox 4.3×10^9
  - P. ex. 84.88.74.180 (mmb.pcb.ub.es)
  - The 4 levels are hierarchical
- Some addresses are reserved, and some networks are “local”
- IPv6: 128 bits (16 bytes). Max: 2^{128} (3.4×10^{38})

- IP addresses are not easy. Most hosts have also a “name”:

  f. ex. www.ncbi.nlm.nih.edu
- Host names have a structure similar to IP addresses:
  - Top domains (.es, .edu, correspond to full class domains and subnets are indicated by prefixes).
  - ub.edu (161.116.x.x)
  - bq.ub.edu (161.116.72.x)
  - www.bq.ub.edu (161.116.72.181)

How to address applications in a server: **Ports**

- Each host has **one IP address** but has **several ports** for known services

- Ports are 2-Byte numbers.
  - 0-1023 are "**Well known ports**" (Telnet: 23, FTP: 21, HTTP: 80, ...).
  - 1024-49151 are "**Registered ports**", usually managed by applications (MySQL: 3306)
  - 49,152-65,535 are "**Dynamic and/or private ports**" freely usable.

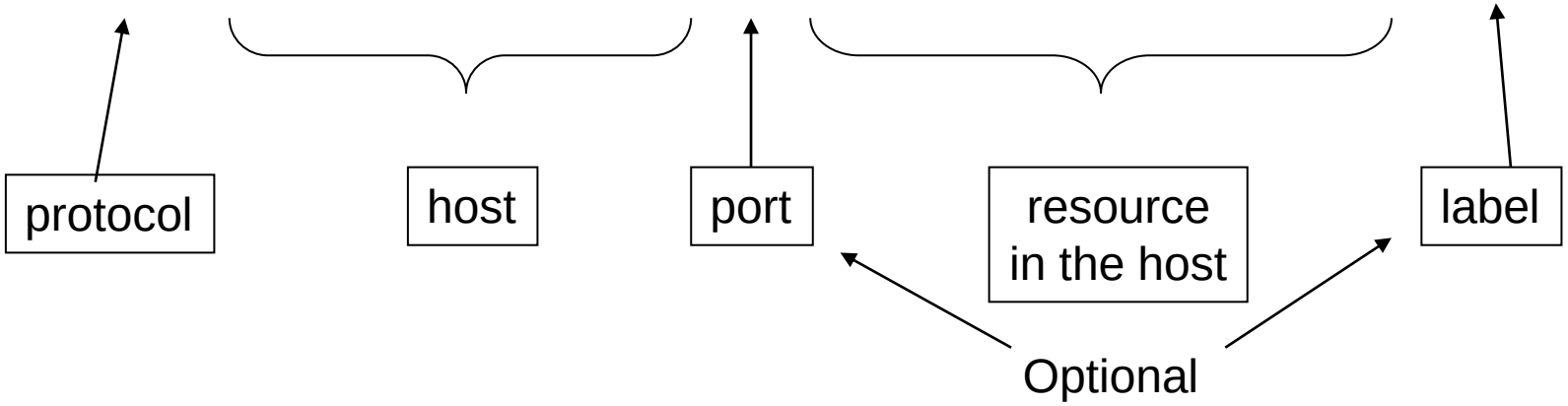
- **Communication to ports triggers the specific application** to deal with the data

- However, different ports from official ones can be used to:
  - Hide applications
  - Have more then one server in the same IP address
  - Hide servers in internal networks.

URI/URLs

- Resources must be identified in a way that includes all the necessary details:

http://mmb.pcb.ub.es:80/courses/master.htm#top



Missing parts of the URL are added by the browser by default!!

Client – server communication (HTTP)

- Most Web Applications use HTTP (hypertext transfer protocol), although sometimes FTP, SMTP

- HTTP is a client-server communication protocol
  - Link between client and server is dynamic
  - Usually **limited to a single transaction**
  - Requests composed by a query operation and a variable set of headers (Metadata)
  - Answers: headers + data

- Relevant Operations: GET, POST
  - GET: Simple retrieval, all information/parameters included in the URL
    - Simple queries, static information
    - Required to be used as hypertext links
  - POST: Query defines the resource, but input data follows
    - Input data can be of any type (including binaries, whole files) or size (within limits)
  - PUT: Similar to POST. Used in APIs

- Relevant HTTP headers
  - Content-type (POST): input data format
  - Content-type (Answer): Data MIME type (text/html, image/jpg, …)
  - Location: Redirects browser
  - Set-cookie: Set a “cookie” on users’ software.