DBW – Databases and Web development

MSc Bioinformatics
for Health Sciences

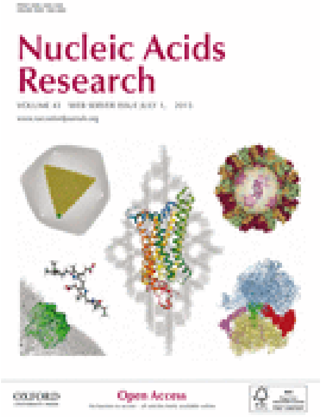
Aims

- 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

- 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

- “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.”
- (…)

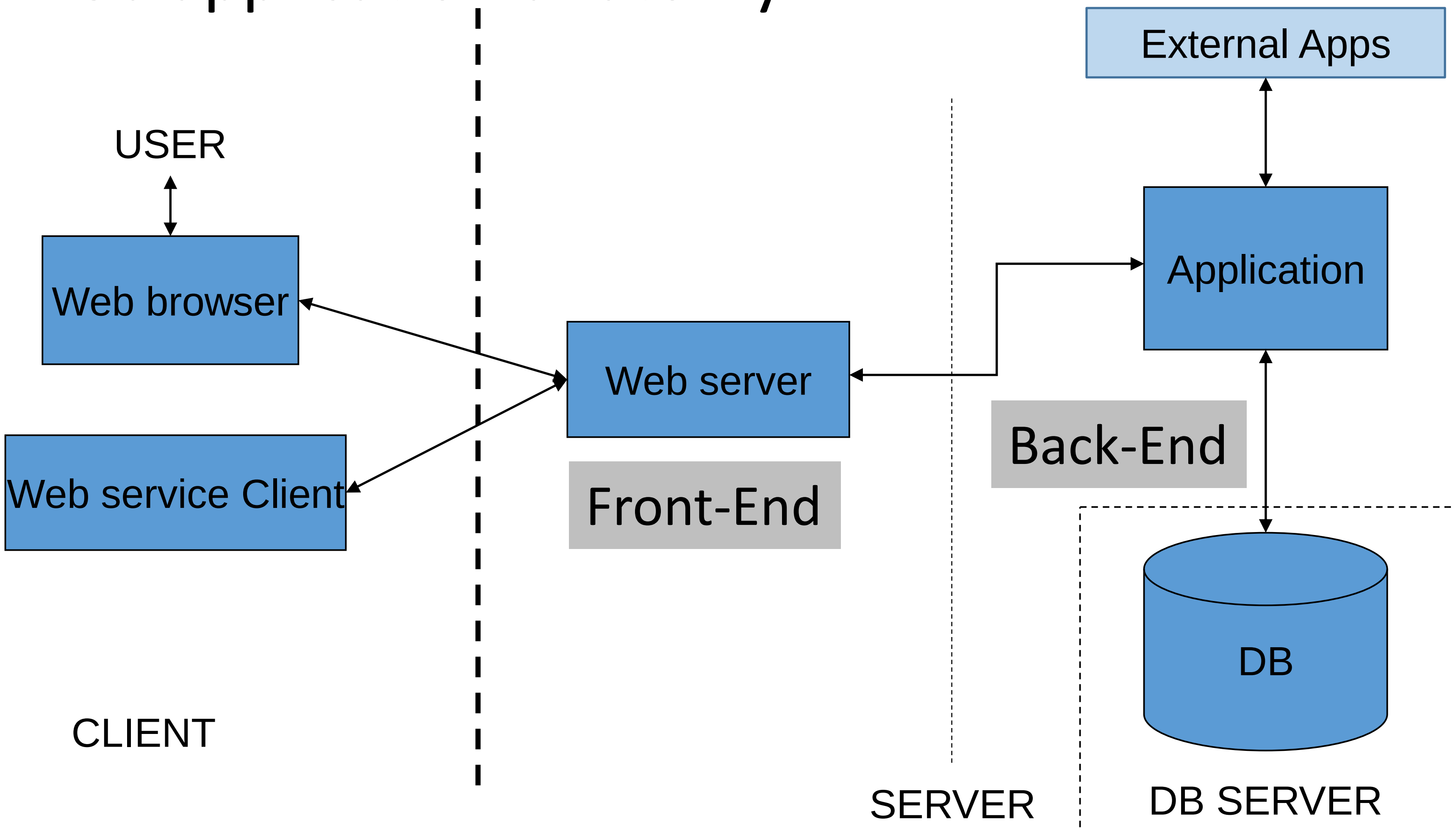
Nucleic Acids Research, Volume 35, Issue suppl_1, 1 January 2007, Pages D1–D2

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

Web applications by access type

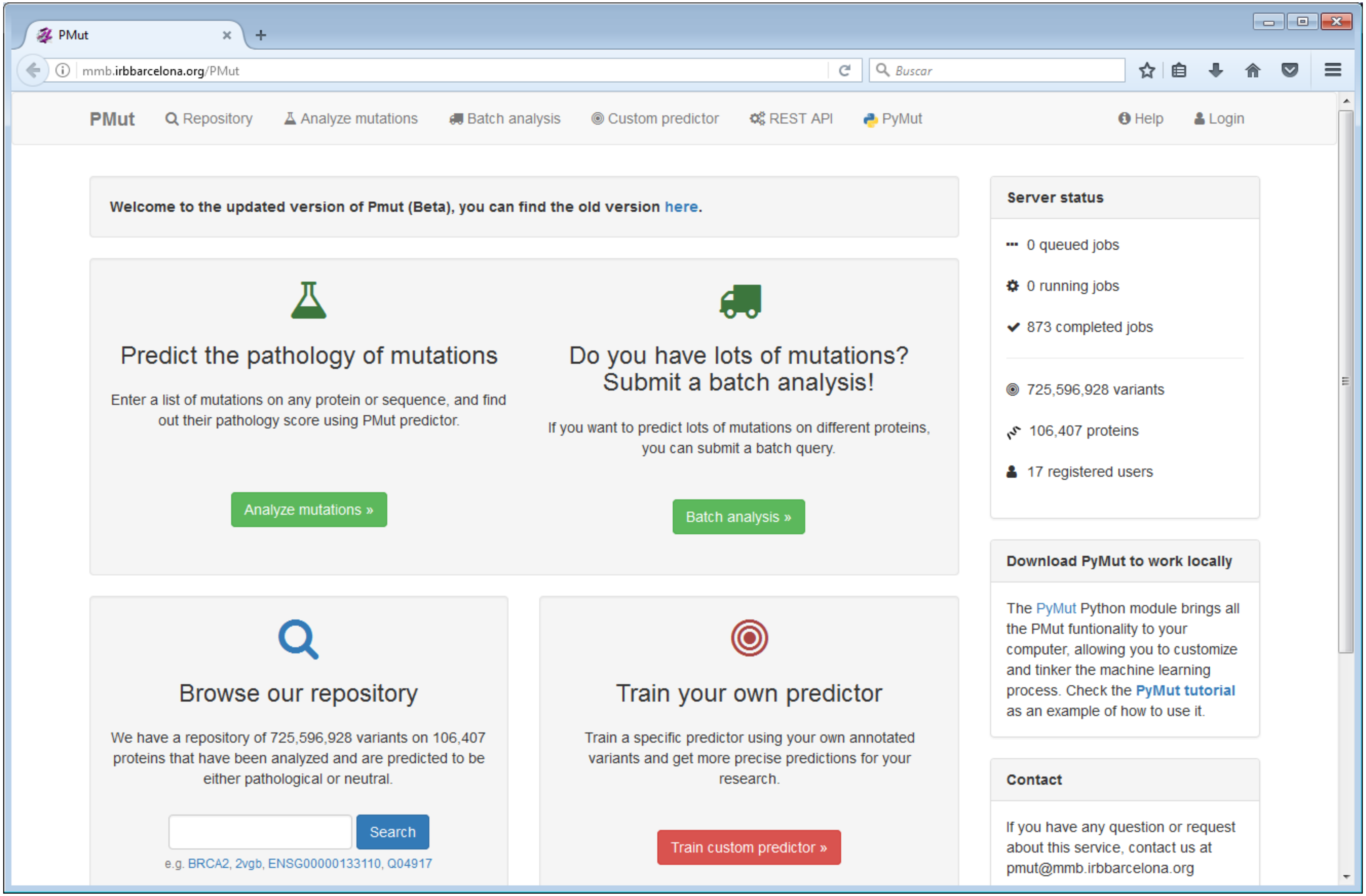
- Web interfaces
  - Provide a user friendly interface (web based) to “human” users
    - Users known how to use the interface
    - There is no need to install software
    - Single operations (no large scale)
    - Must adapt to navigation uses (low latency, synchronous answers,…)
- Web services
 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

Web application anatomy



Web 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 services (programmatic access)



Welcome to the updated version of Pmut (Beta), you can find the old version here.

Predict the pathology of mutations

Enter a list of mutations on any protein or sequence, and find out their pathology score using PMut predictor.

Do you have lots of mutations? Submit a batch analysis!

If you want to predict lots of mutations on different proteins, you can submit a batch query.

Browse our repository

We have a repository of 725,596,928 variants on 106,407 proteins that have been analyzed and are predicted to be either pathological or neutral.

e.g. BRCA2, 2vgb, ENSG00000133110, Q04917

Train your own predictor

Train a specific predictor using your own annotated variants and get more precise predictions for your research.

Server status

- 0 queued jobs
- 0 running jobs
- ✓ 873 completed jobs
- 725,596,928 variants
- 106,407 proteins
- 17 registered users

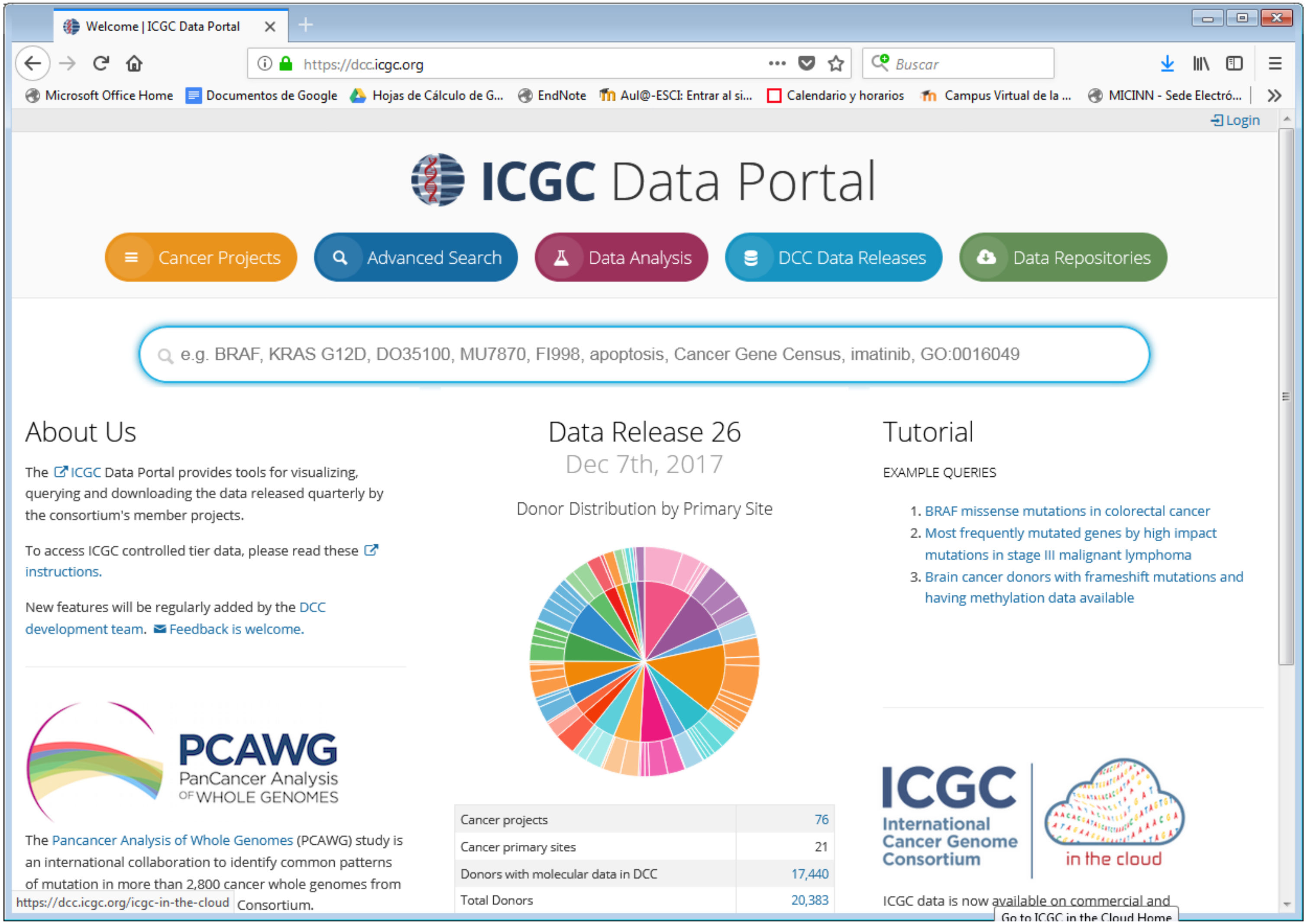
Download PyMut to work locally

The PyMut Python module brings all the PMut funtionality to your computer, allowing you to customize and tinker the machine learning process. Check the **PyMut tutorial** as an example of how to use it.

Contact

If you have any question or request about this service, contact us at pmut@mmb.irbbarcelona.org

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



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.

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

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

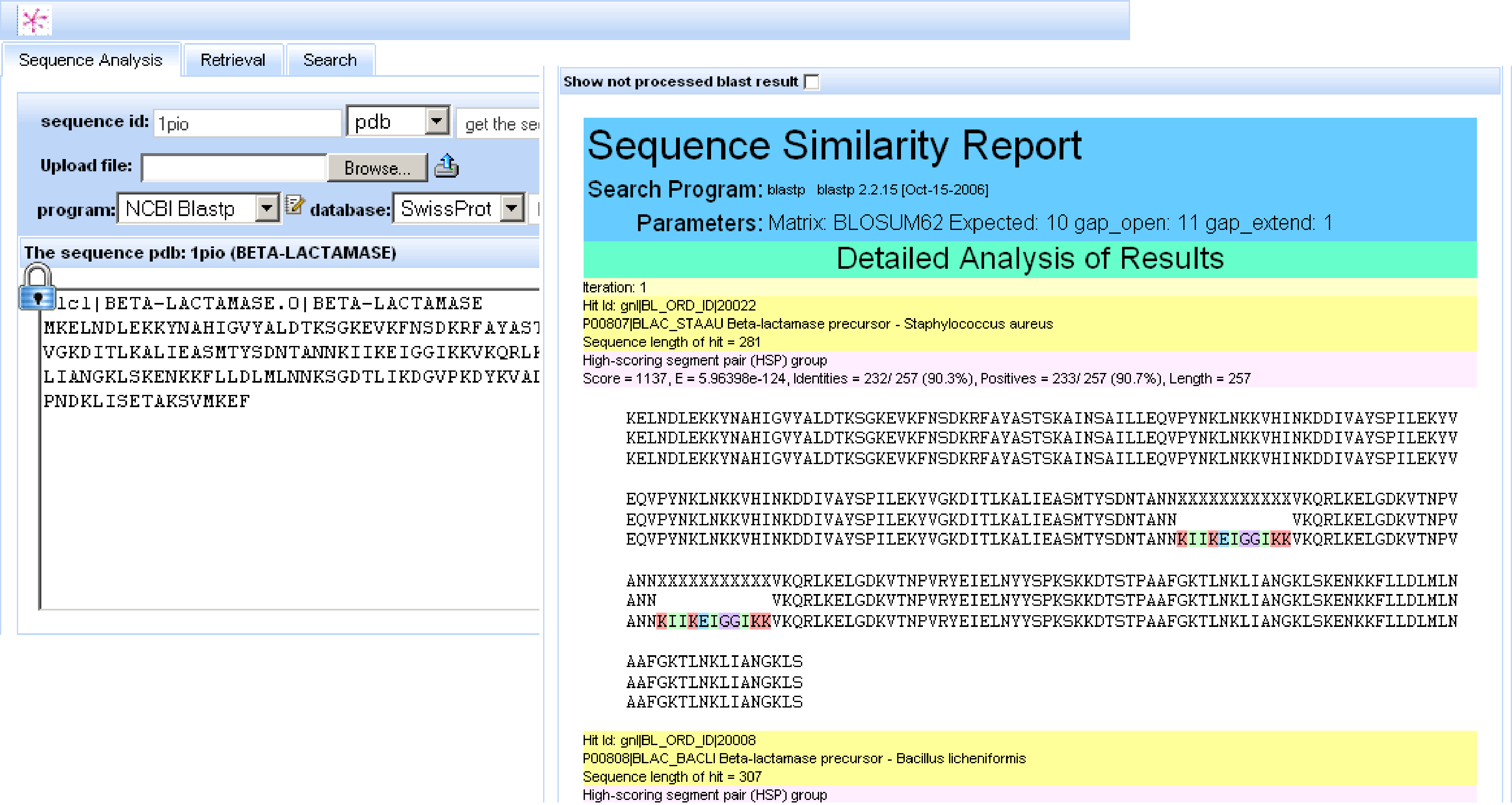
Tutorial

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 data is now available on commercial and

Web interfaces to apps.



sequence id: 1pio | pdb | get the se

Upload file: Browse...

program: NCBI Blastp **database:** SwissProt

The sequence pdb: 1pio (BETA-LACTAMASE)

```
lcl|BETA-LACTAMASE.0|BETA-LACTAMASE
MKELNDLEKKYNAHIGVYALDTKSGKEVKFNSDKRFAYAST
VGKDITLKALIEASMTYSDNTANNKIIKEIGGIKKVKQRL
LIANGKLSKENKKFLLDLMLNNKSGDTLIKDGVPKDYKVA
PNDKLISETAKSVMKEF
```

Show not processed blast result □

Sequence Similarity Report

Search Program: blastp blastp 2.2.15 [Oct-15-2006]

Parameters: Matrix: BLOSUM62 Expected: 10 gap_open: 11 gap_extend: 1

Detailed Analysis of Results

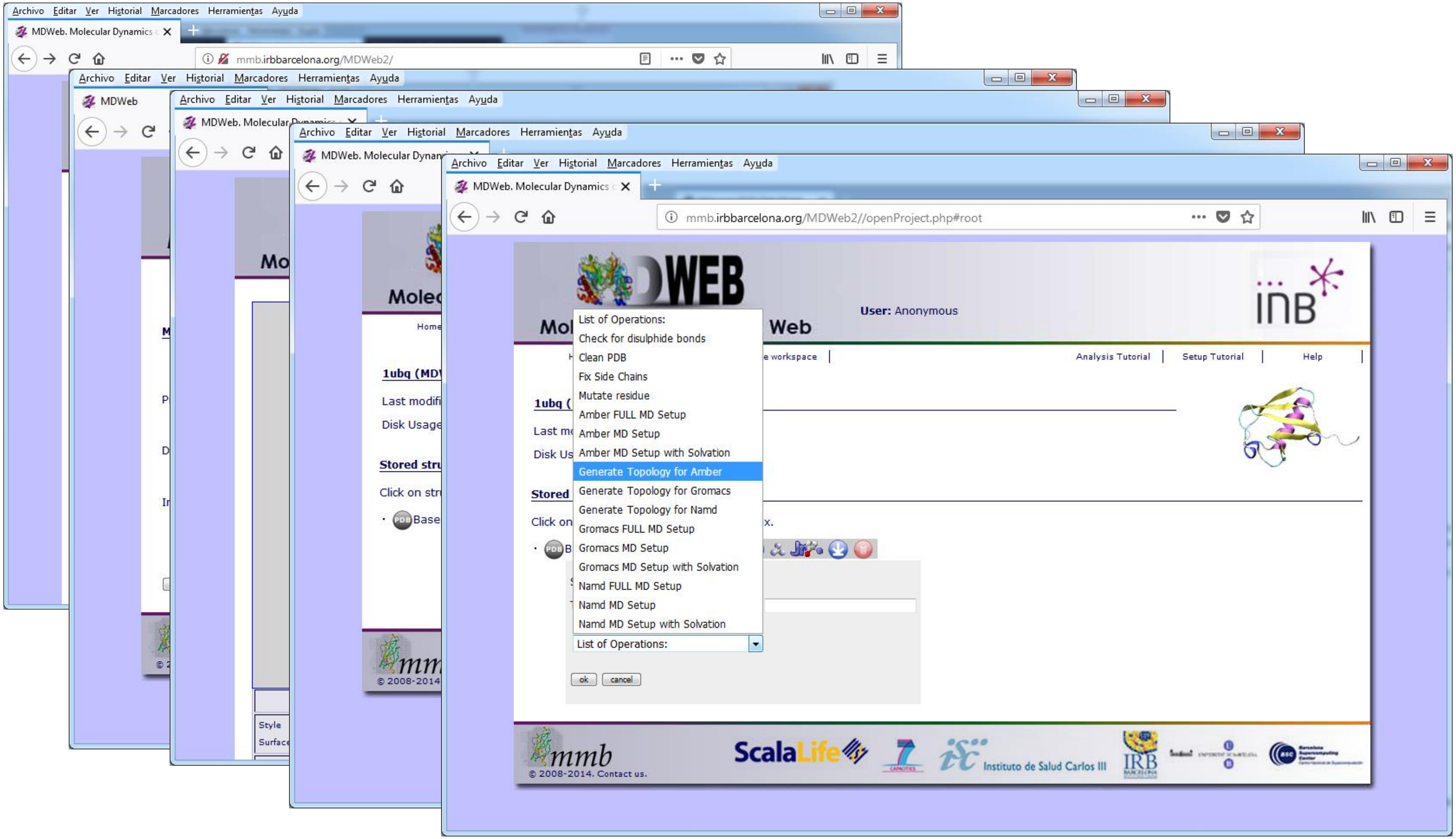
Iteration: 1

Hit Id: gnl|BL_ORD_ID|20022
P00807|BLAC_STAAU Beta-lactamase precursor - Staphylococcus aureus
Sequence length of hit = 281

High-scoring segment pair (HSP) group
Score = 1137, E = 5.96398e-124, Identities = 232/ 257 (90.3%), Positives = 233/ 257 (90.7%), Length = 257

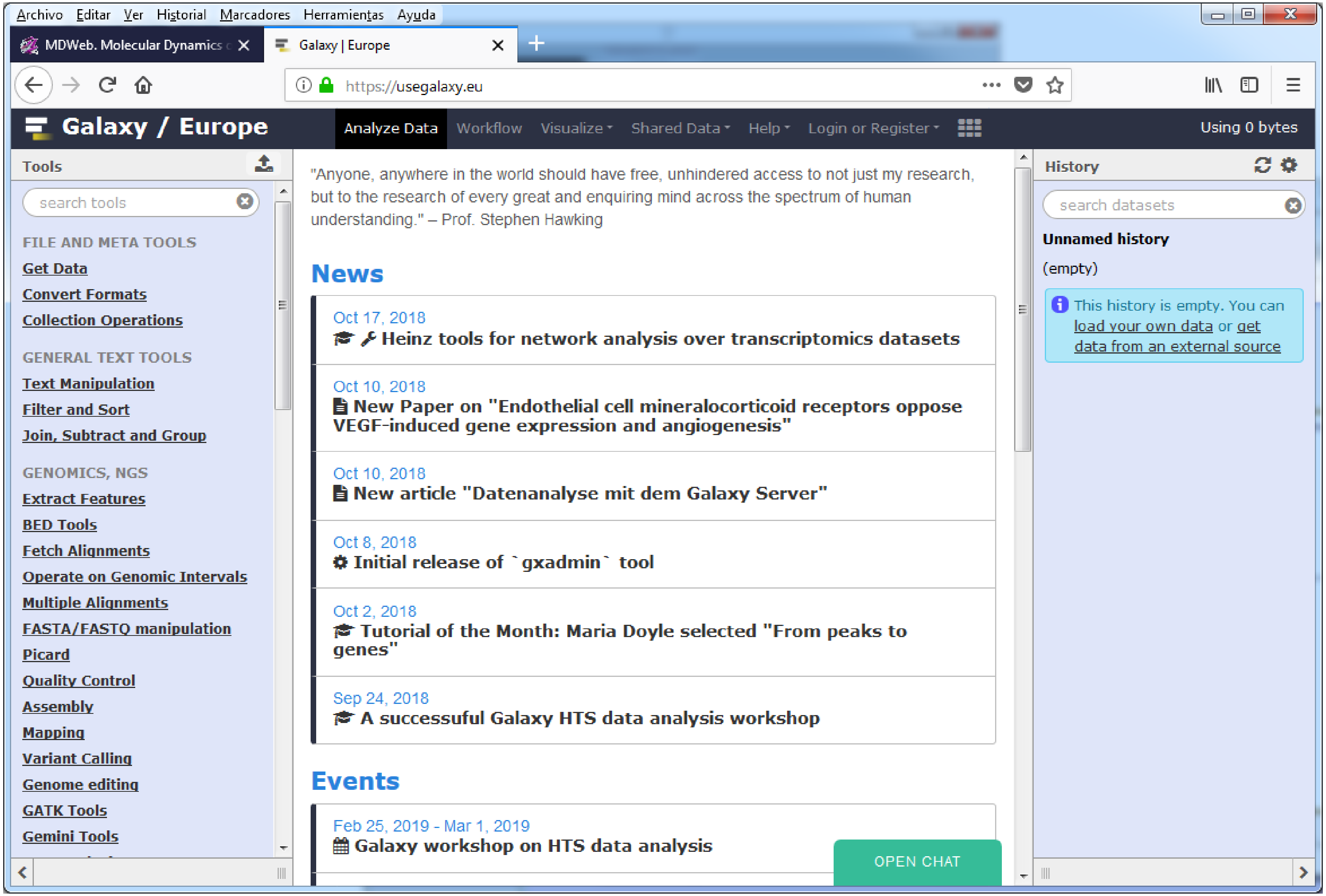
```
KELNDLEKKYNAHIGVYALDTKSGKEVKFNSDKRFAYASTSKAINSAILLEQVPYNKLNKKVHINKDDIVAYSPILEKYV
KELNDLEKKYNAHIGVYALDTKSGKEVKFNSDKRFAYASTSKAINSAILLEQVPYNKLNKKVHINKDDIVAYSPILEKYV
KELNDLEKKYNAHIGVYALDTKSGKEVKFNSDKRFAYASTSKAINSAILLEQVPYNKLNKKVHINKDDIVAYSPILEKYV

EQVPYNKLNKKVHINKDDIVAYSPILEKYVGKDITLKALIEASMTYSDNTANNXXXXXXXXXXXVKQRLKELGDKVTNPV
EQVPYNKLNKKVHINKDDIVAYSPILEKYVGKDITLKALIEASMTYSDNTANN           VKQRLKELGDKVTNPV
EQVPYNKLNKKVHINKDDIVAYSPILEKYVGKDITLKALIEASMTYSDNTANNKIIKEIGGIKKVKQRLKELGDKVTNPV

ANNXXXXXXXXXXXVKQRLKELGDKVTNPVRYEIELNYYSPKSKKDTSTPAAFGKTLNKLIANGKLSKENKKFLLDLMLN
ANN           VKQRLKELGDKVTNPVRYEIELNYYSPKSKKDTSTPAAFGKTLNKLIANGKLSKENKKFLLDLMLN
ANNKIIKEIGGIKKVKQRLKELGDKVTNPVRYEIELNYYSPKSKKDTSTPAAFGKTLNKLIANGKLSKENKKFLLDLMLN

AAFGKTLNKLIANGKLS
AAFGKTLNKLIANGKLS
AAFGKTLNKLIANGKLS
```

Hit Id: gnl|BL_ORD_ID|20008
P00808|BLAC_BACLI Beta-lactamase precursor - Bacillus licheniformis
Sequence length of hit = 307
High-scoring segment pair (HSP) group


Archivo Editar Ver Historial Marcadores Herramientas Ayuda
MDWeb. Molecular Dynamics
mmb.irbbarcelona.org/MDWeb2//openProject.php#root
MDWEB
Molecular Dynamics on Web
User: Anonymous
INB
workspace
Analysis Tutorial
Setup Tutorial
Help
1ubq
Last modified
Disk Usage
Stored structures
Click on
List of Operations:
Check for disulphide bonds
Clean PDB
Fix Side Chains
Mutate residue
Amber FULL MD Setup
Amber MD Setup
Amber MD Setup with Solvation
Generate Topology for Amber
Generate Topology for Gromacs
Generate Topology for Namd
Gromacs FULL MD Setup
Gromacs MD Setup
Gromacs MD Setup with Solvation
Namd FULL MD Setup
Namd MD Setup
Namd MD Setup with Solvation
List of Operations:
ok
cancel
mmb
© 2008-2014. Contact us.
ScalaLife
Instituto de Salud Carlos III
IRB Barcelona

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

http://mmb.irbbarcelona.org/MDWeb2



"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

Oct 17, 2018
Heinz tools for network analysis over transcriptomics datasets

Oct 10, 2018
New Paper on "Endothelial cell mineralocorticoid receptors oppose VEGF-induced gene expression and angiogenesis"

Oct 10, 2018
New article "Datenanalyse mit dem Galaxy Server"

Oct 8, 2018
Initial release of `gxadmin` tool

Oct 2, 2018
Tutorial of the Month: Maria Doyle selected "From peaks to genes"

Sep 24, 2018
A successuful Galaxy HTS data analysis workshop

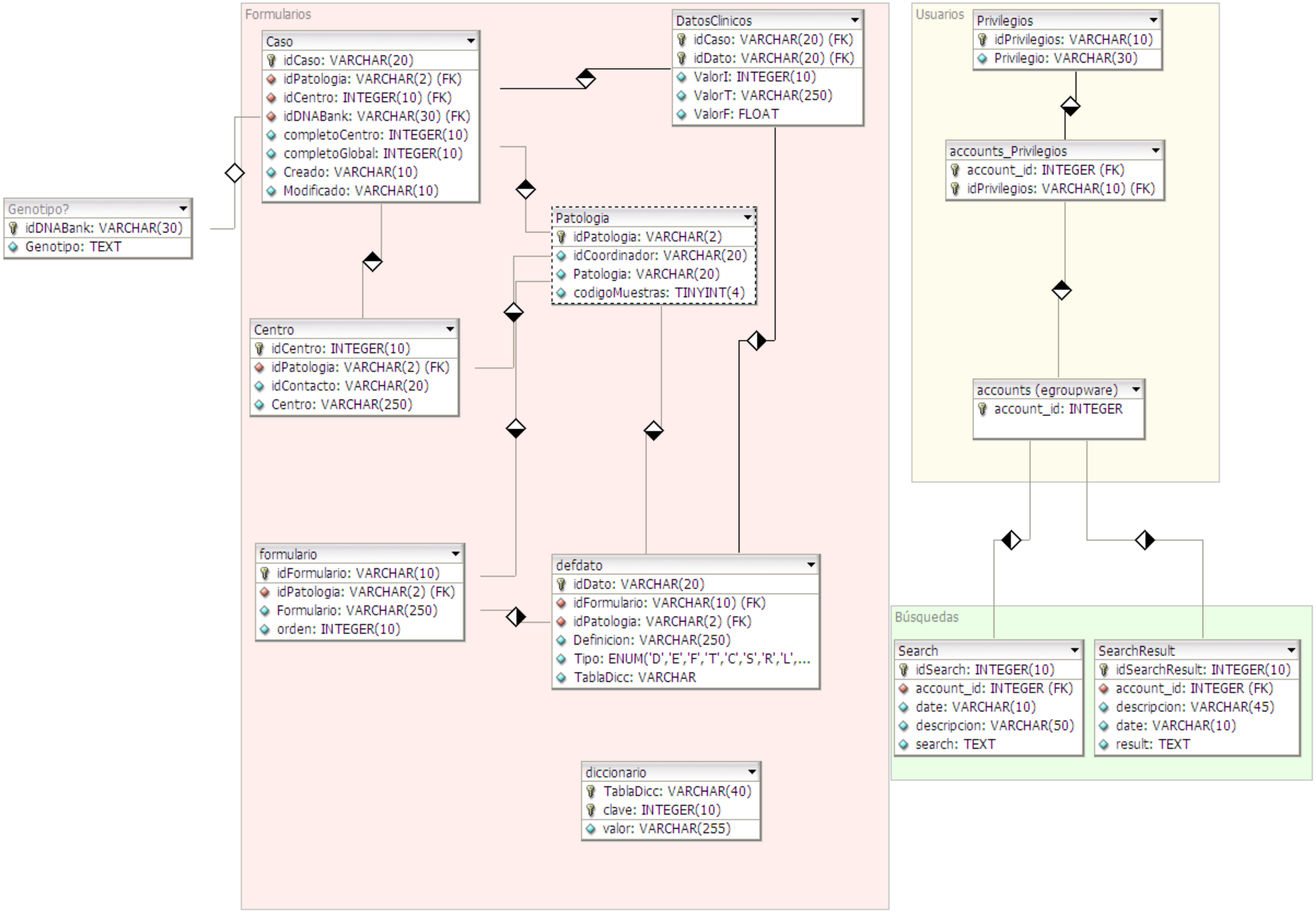
Events

Feb 25, 2019 - Mar 1, 2019
Galaxy workshop on HTS data analysis

This history is empty. You can load your own data or get data from an external source

https://usegalaxy.eu/

Special purpose applications & DBs



Formularios

Caso
- idCaso: VARCHAR(20)
- idPatologia: VARCHAR(2) (FK)
- idCentro: INTEGER(10) (FK)
- idDNABank: VARCHAR(30) (FK)
- completoCentro: INTEGER(10)
- completoGlobal: INTEGER(10)
- Creado: VARCHAR(10)
- Modificado: VARCHAR(10)

DatosClinicos
- idCaso: VARCHAR(20) (FK)
- idDato: VARCHAR(20) (FK)
- ValorI: INTEGER(10)
- ValorT: VARCHAR(250)
- ValorF: FLOAT

Genotipo?
- idDNABank: VARCHAR(30)
- Genotipo: TEXT

Patologia
- idPatologia: VARCHAR(2)
- idCoordinador: VARCHAR(20)
- Patologia: VARCHAR(20)
- codigoMuestras: TINYINT(4)

Centro
- idCentro: INTEGER(10)
- idPatologia: VARCHAR(2) (FK)
- idContacto: VARCHAR(20)
- Centro: VARCHAR(250)

formulario
- idFormulario: VARCHAR(10)
- idPatologia: VARCHAR(2) (FK)
- Formulario: VARCHAR(250)
- orden: INTEGER(10)

defdato
- idDato: VARCHAR(20)
- idFormulario: VARCHAR(10) (FK)
- idPatologia: VARCHAR(2) (FK)
- Definicion: VARCHAR(250)
- Tipo: ENUM('D','E','F','T','C','S','R','L',...
- TablaDicc: VARCHAR

diccionario
- TablaDicc: VARCHAR(40)
- clave: INTEGER(10)
- valor: VARCHAR(255)

Usuarios

Privilegios
- idPrivilegios: VARCHAR(10)
- Privilegio: VARCHAR(30)

accounts_Privilegios
- account_id: INTEGER (FK)
- idPrivilegios: VARCHAR(10) (FK)

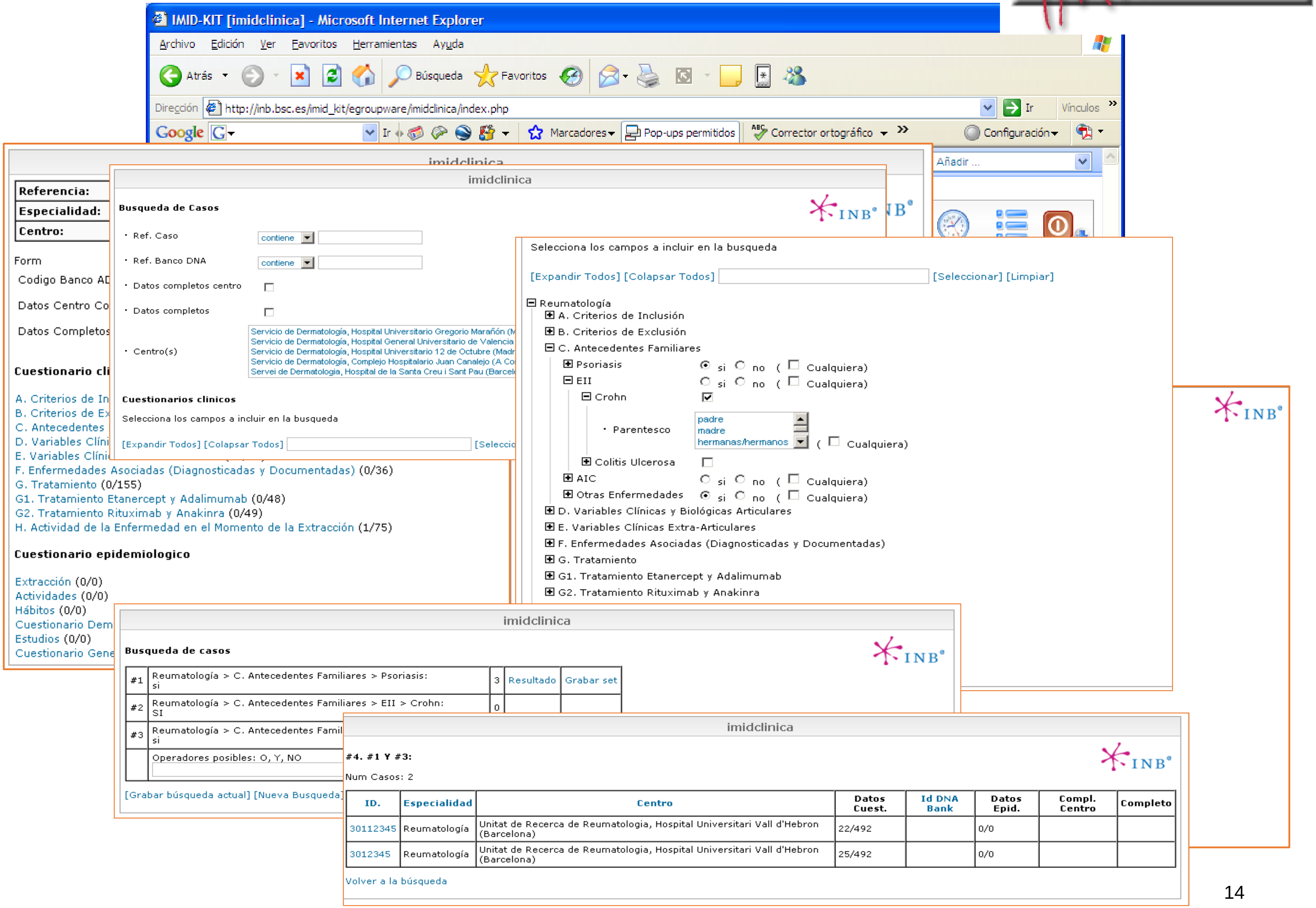
accounts (egroupware)
- account_id: INTEGER

Búsquedas

Search
- idSearch: INTEGER(10)
- account_id: INTEGER (FK)
- date: VARCHAR(10)
- descripcion: VARCHAR(50)
- search: TEXT

SearchResult
- idSearchResult: INTEGER(10)
- account_id: INTEGER (FK)
- descripcion: VARCHAR(45)
- date: VARCHAR(10)
- result: TEXT

IMID-Kit



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

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

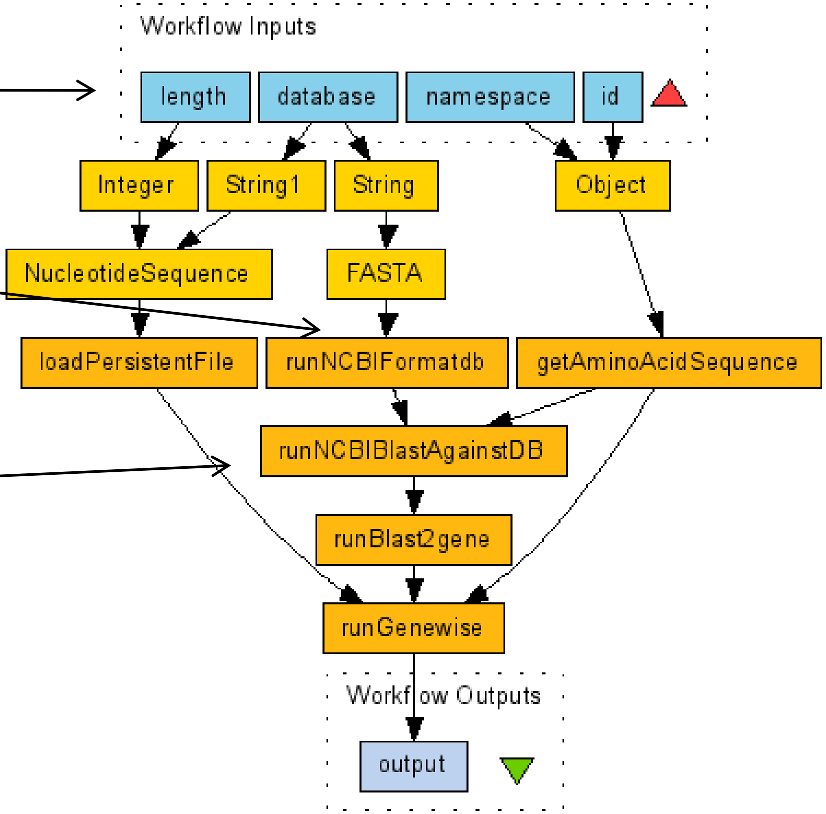
Bioinformatics web-services and workflows

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

1. Define specifications
2. Analyze data and built a data model
3. Decide/prepare Database implementation
4. Build ETL if necessary
5. Define interfaces
6. Define and prepare files/scripts layout
7. Write application code
8. Test, debug, document…

Course logistics

- Web site(s)
  - Course materials:
    - http://mmb.pcb.ub.es/formacio/
  - Personal sites:
    - http://mmb.pcb.ub.es/formacio/~dbwXX
  - SSH Access
    - ssh mmb.pcb.ub.es –p 22021 –l dbwXX
  - MySQL Access
    - Localhost only

Software to install

- Ideally Linux (may need root privileges)

- From Linux distribution
  - Apache Web Server (v. 2.x)
    - With PHP 5.x and mysql support
  - MYSQL server (v. 5.x)
  - MYSQL Workbench or phpMyAdmin

- Netbeans (PHP module) (optional)

- MongoDB (optional)

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
  - May use fake data if necessary
  - Available at the personal web site
  - Preferred languages: PHP, Perl, Mysql
  - Source code at github

Evaluation

- Web application project
  - Steps:
    - Initial specification
    - Data analysis & Database design
    - Project Demo
    - Final application
- Installed on server
  - mmb.pcb.ub.es/formacio/~dbwXX
    - Account dbwXX

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)
    - Ports
    - Resource identification
      - DNS
      - URL/URI concept
  - Client/server transactions
    - HTTP protocol

Present internet

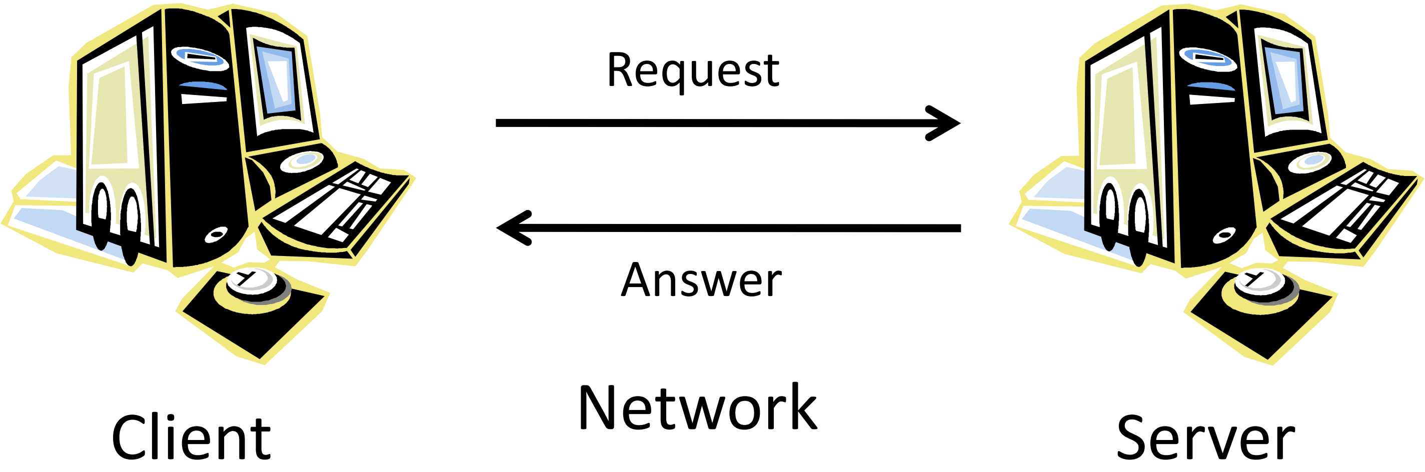
- Huge network of computers using common communication protocols (TCP/IP, HTTP)

- Distributed, no central servers
  - (Well, not really true in bioinformatics)

- Common language: HTML/CSS (XML)

- Content mostly static, but dynamic behaviour is possible through web applications

Components



Request

Answer

Client

Network

Server

- 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
  - MAC Address: Address of the physical interface
  - IP Address: Address of the computer

IP addresses

- 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. 161.116.222.59 (mmb.pcb.ub.es)
  - 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})

Names *vs* addresses (Domain Name System)

- 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.es (161.116.x.x)
  – bq.ub.es (161.116.154.x)
  – www.bq.ub.es (161.116.154.18)

How to address applications in a server: Port ids.

- 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.

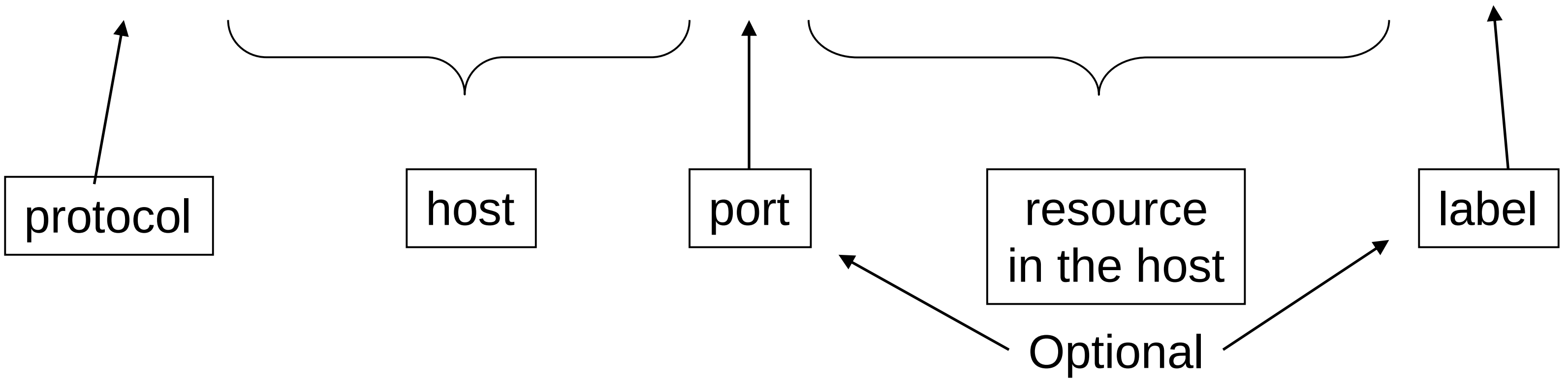
- Communication to ports triggers the specific application to deal with the data

- 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



protocol

host

port

resource in the host

label

Optional

Missing parts of the URL are filled by default!!

Client – server communication

- 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.
  - Answers: headers + data

Client – server communication

- 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)

- 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.

Cookies

- Small information tags sent as HTTP headers and stored in the browser side
  - Are associated with a URL, and are sent back to the server whenever that URL is visited within a expiration date

```
Set-Cookie:
  PHPSESSID=bb56ee648aeac6923e3360a7b8284a6f;
  path=/
```

- Useful to “remember” clients, but some people disables them!