

DBW – Databases and Web development



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 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

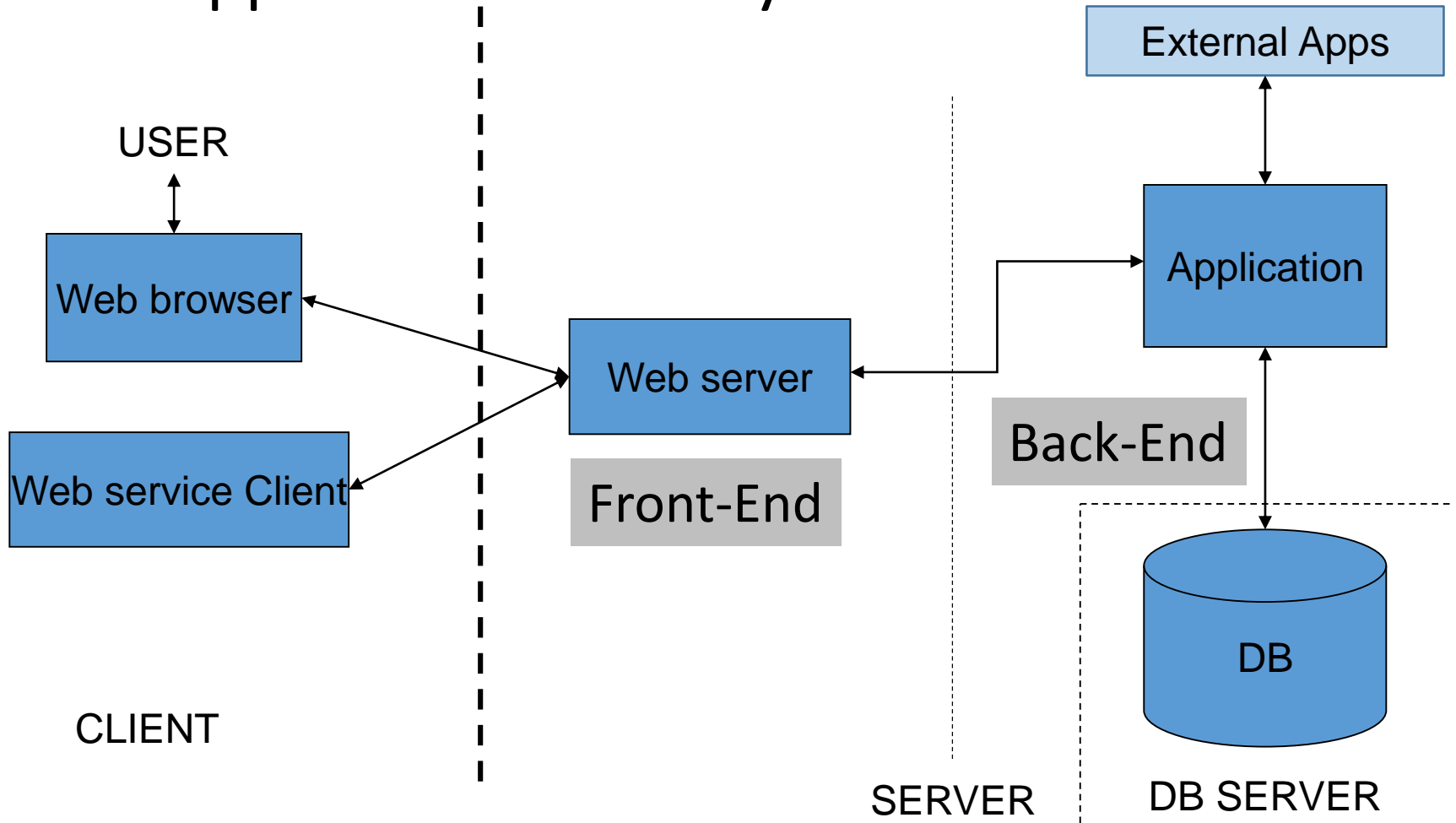
- Tools and data should be available through web
- Ex. Nucleic Acid Research reviews:
 - Database Issue (January) 1170 DBs
 - Web Server Issue (July) 1200 Servers



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)
 - Intended 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)

MolecularModelingandBioinformatics

[New search](#)[Protein Data Bank](#)

PDB local mirror at MMB: 117976 total entries (Proteins: 106948, Nuc. Acids: 2908, Prot-Nuc Complexes 6009, Carboh. 18)

PDB Id

Uniprot Acc.

Sequence
searchUpload File: [Examinar...](#) No se ha seleccionado

Exact match

☒ Protein ☐ Nucleic acid

Advanced Search:

Resolution

From to Compound
type☒ Any ☐ carb ☐ nuc ☐ other ☐ prot ☐ prot-

Exp. type

☒ Any ☐ ELECTRON_CRYSTALLOGRAPHY
NEUTRON_DIFRACCTON ☐ NMR ☐ SOLID-

Simulation

☐ MoDEL ☐ BigNASim

MolecularModelingandBioinformatics

[New search](#)[Protein Data Bank](#)

Entry: 2KI5

Classification

Transferase

Type

Prot

Deposition Date

02/12/99

Title

Herpes simplex type-1 thymidine kinase in complex with the drug aciclovir at 1.9a resolution

Source

Human herpesvirus 1

Authors list

Batuwangala, T., Bennett, M. S., Champness, J. N., Rutherford, T., Sanderson, M. R., Summers, W. C., Sun, H., Wien, F., Wright, G.

Resolution

1.9

Exp. type

X-RAY DIFFRACTION

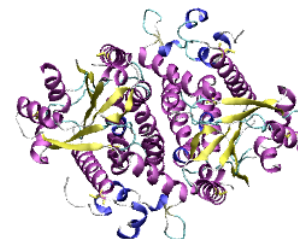
Structure:

Asymmetric Unit

[PDB File](#)[3D View](#)

BioUnit(s)

1

[PDB File](#)[3D View](#)View entry at [\[PDB\]](#)

Web interfaces to apps.

Sequence Analysis

Retrieval

Search

sequence id: 1pio

pdb

get the se

Upload file:

Browse...

program: NCBI Blastp

database: SwissProt

The sequence pdb: 1pio (BETA-LACTAMASE)

1c1 | BETA-LACTAMASE.O | BETA-LACTAMASE

MKELNDLEKKYNAHIGVYALDTKSGKEVKFNSDKRFAYAST

VGKDITLKALIEASMTYSNTANNKIIKEIGGIKKVKQRLP

LIANGKLSKENKKFLDLMLNKSGLTIKDGVPKDYKVAI

PNDKLISSETAKSVMKEF

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: gn|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

EQVPYNKLNKKVHINKDDIVAYSPILEKYVGKDITLKALIEASMTYSNTANNXXXXXXXXXXXXVKQRLKELGDKVTNPV

EQVPYNKLNKKVHINKDDIVAYSPILEKYVGKDITLKALIEASMTYSNTANN VKQRLKELGDKVTNPV

EQVPYNKLNKKVHINKDDIVAYSPILEKYVGKDITLKALIEASMTYSNTANNKIIKEIGGIKKVKQRLKELGDKVTNPV

ANNXXXXXXXXXXXXVKQRLKELGDKVTNPVRYEIELNYYSPKSKKDTSTPAAFGKTLNKLIANGKLSKENKKFLDLMLN

ANN VKQRLKELGDKVTNPVRYEIELNYYSPKSKKDTSTPAAFGKTLNKLIANGKLSKENKKFLDLMLN

ANNKIIKEIGGIKKVKQRLKELGDKVTNPVRYEIELNYYSPKSKKDTSTPAAFGKTLNKLIANGKLSKENKKFLDLMLN

AAFGKTLNKLIANGKLS

AAFGKTLNKLIANGKLS

AAFGKTLNKLIANGKLS

Hit id: gn|BL_ORD_ID|20008

P00808|BLAC_BACLI Beta-lactamase precursor - Bacillus licheniformis

Sequence length of hit = 307

High-scoring segment pair (HSP) group

Tools

search tools

[Get Data](#)[Lift-Over](#)[Text Manipulation](#)[Datamash](#)[Convert Formats](#)[Filter and Sort](#)[Join, Subtract and Group](#)[Fetch Alignments/Sequences](#)[NGS: QC and manipulation](#)[NGS: Mapping](#)[NGS: RNA Analysis](#)[NGS: SAMtools](#)[NGS: BamTools](#)[NGS: Picard](#)[NGS: VCF Manipulation](#)[NGS: Peak Calling](#)[NGS: Variant Analysis](#)[NGS: RNA Structure](#)[NGS: DuplexNovo](#)[Operate on Genomic Intervals](#)[Statistics](#)[Graph/Display Data](#)[CloudMap](#)[Phenotype Association](#)[BEDTools](#)[Genome Diversity](#)

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).

Galaxy 101

Start small

The very first tutorial you need



Tweets

[Follow](#)**Galaxy Project** @galaxyproject

20h

In 2 weeks: NGS Data Analysis in Galaxy (@CTMMTraIT course)

bit.ly/1P06TzI #usegalaxy

Expand

**Galaxy Project** @galaxyproject

21h

History

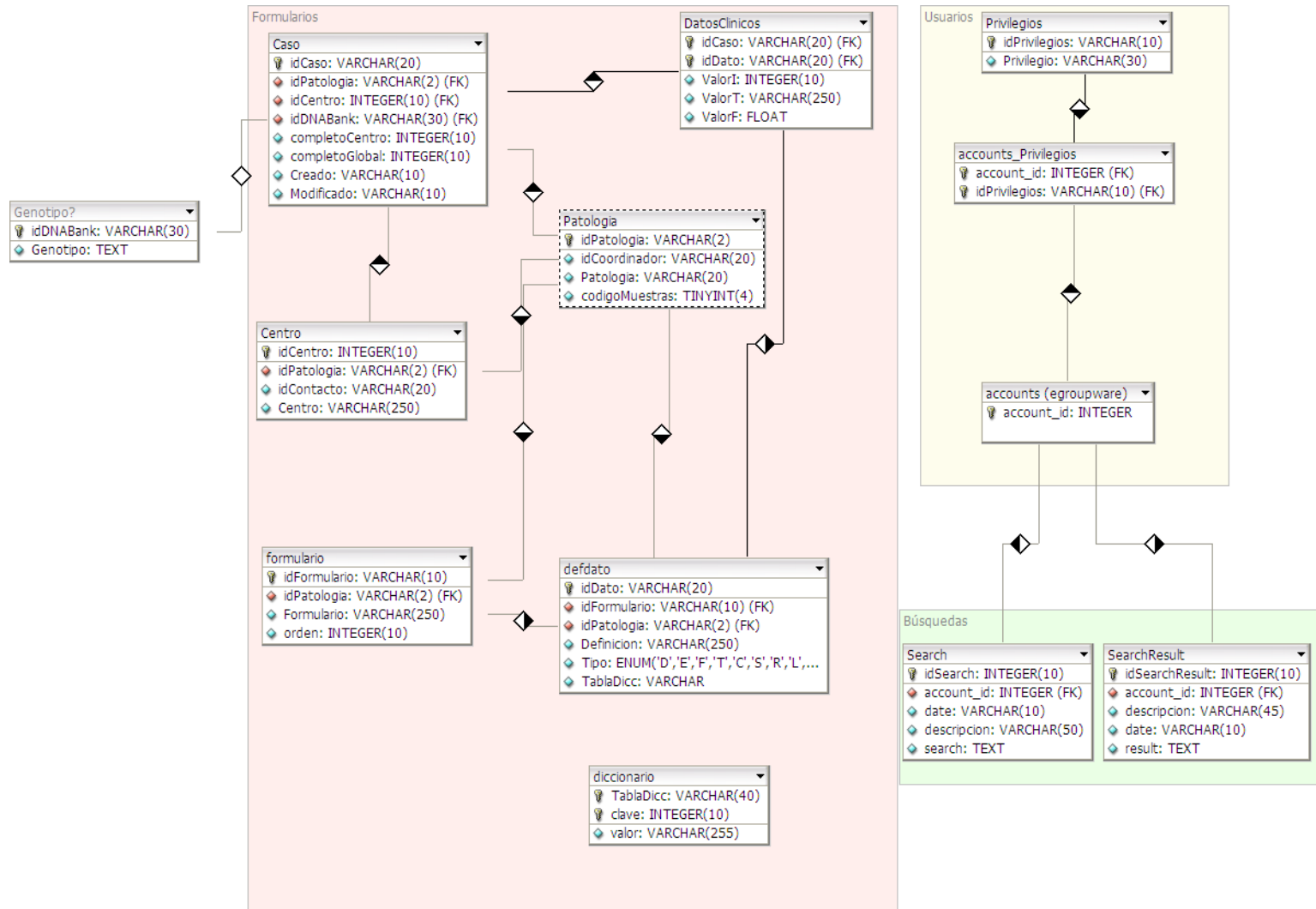
search datasets

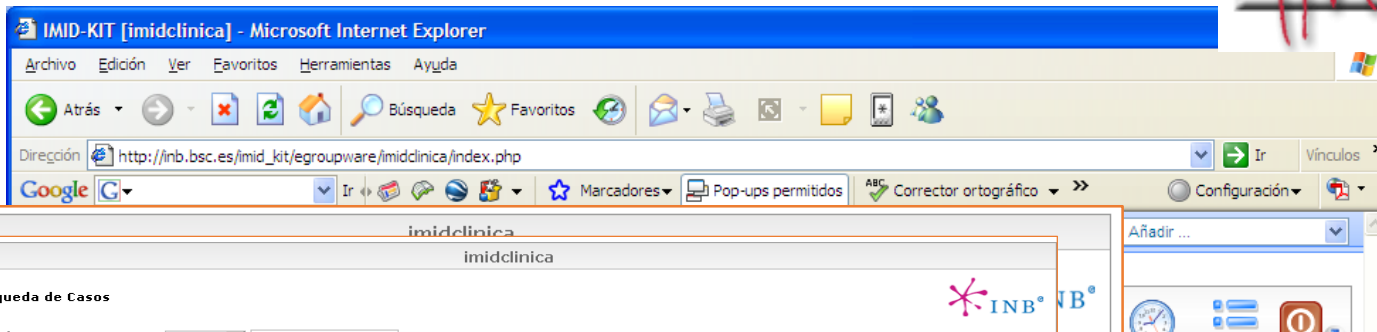
Unnamed history

0 b

i This history is empty. You can [load your own data](#) or [get data from an external source](#)

Special purpose applications & DBs





Referencia:

Especialidad:

Centro:

Form

Código Banco AD

Datos Centro Co

Datos Completos

Cuestionario clínico

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 epidemiológico

Extracción (0/0)

Actividades (0/0)

Hábitos (0/0)

Cuestionario Dem

Estudios (0/0)

Cuestionario Gene

imidclinica

Busqueda de casos

#	Reumatología > C. Antecedentes Familiares > Psoriasis:		Resultado	Grabar set
#1	si	3		
#2	Reumatología > C. Antecedentes Familiares > EII > Crohn:	0		
#3	Reumatología > C. Antecedentes Famil			

Operadores posibles: O, Y, NO

#4. #1 Y #3:

Num Casos: 2

[Grabar búsqueda actual] [Nueva Búsqueda]

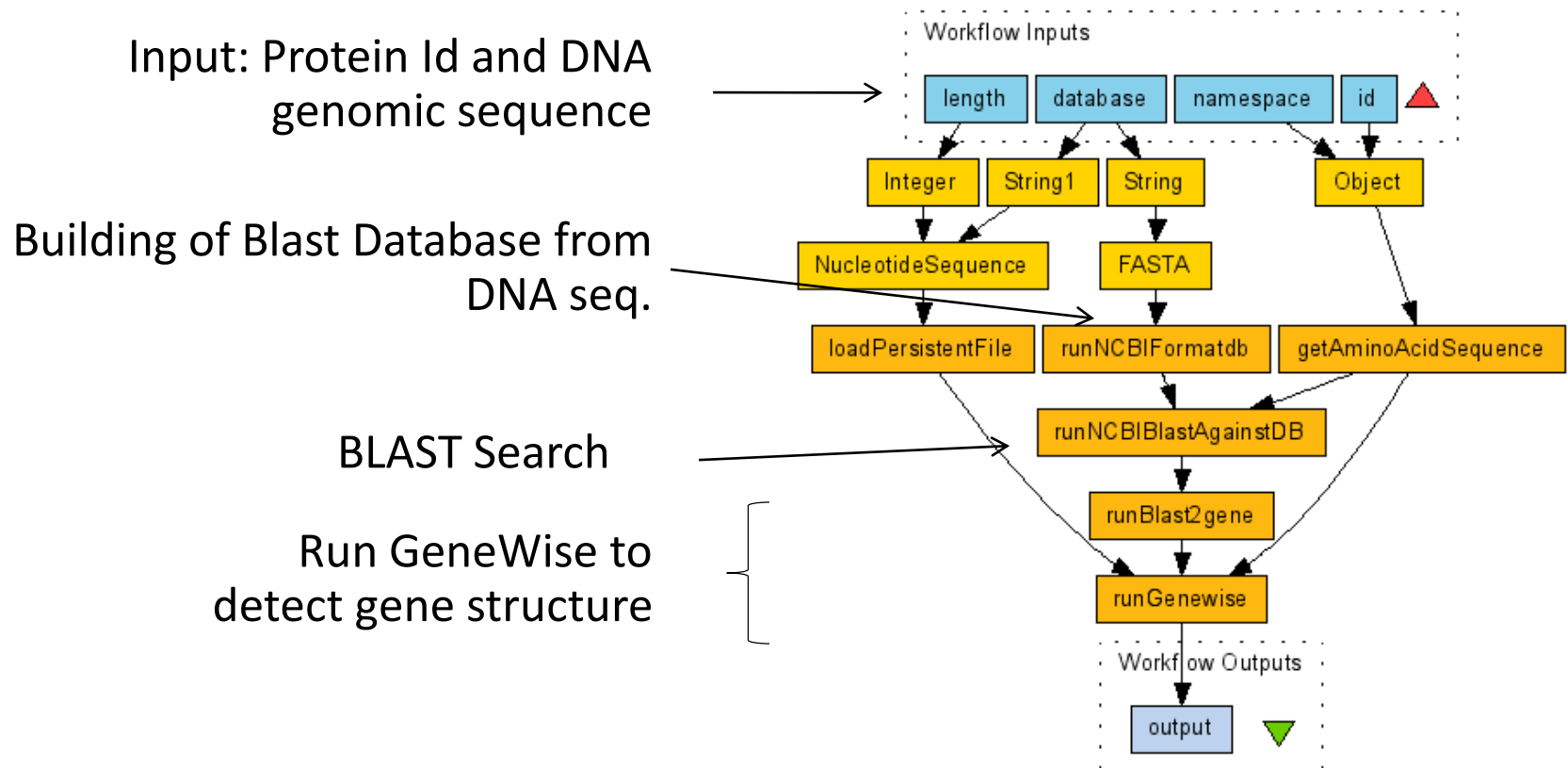
imidclinica

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



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

Evaluation

- Web application project
 - Steps:
 - Initial specification (15 Jan)
 - Data analysis & Database design (25 Jan)
 - Project Demo (6 Feb)
 - Final application (~6 Mar)
- Installed on server
 - mmb.pcb.ub.es/formacio/~dbwXX
 - Account dbwXX

Basic computer communication protocols

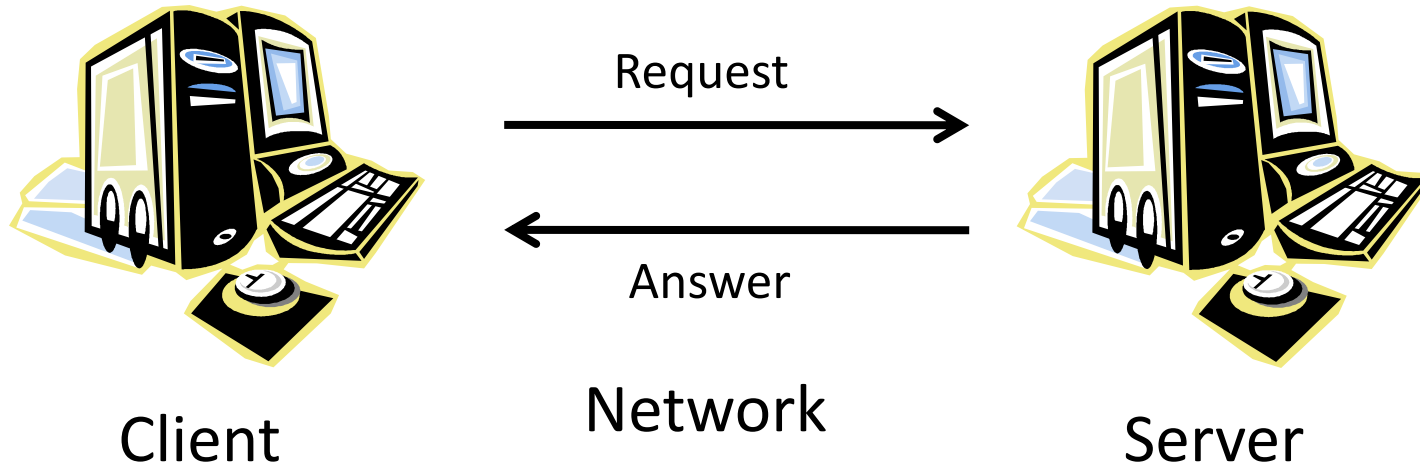
Aim & Outline

- 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



- 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)

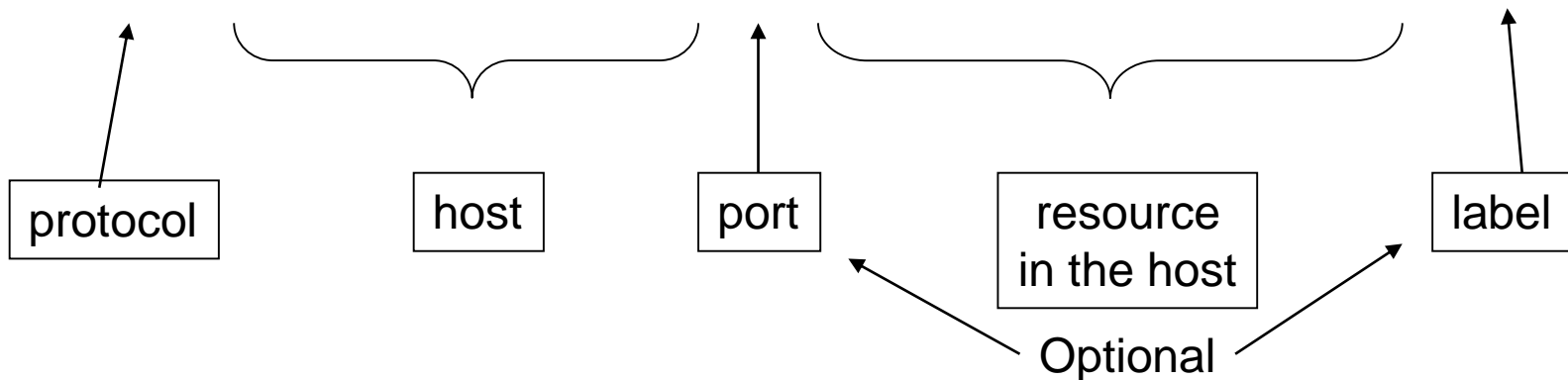
How to address applications in a server: Ports.

- 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 than 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 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!