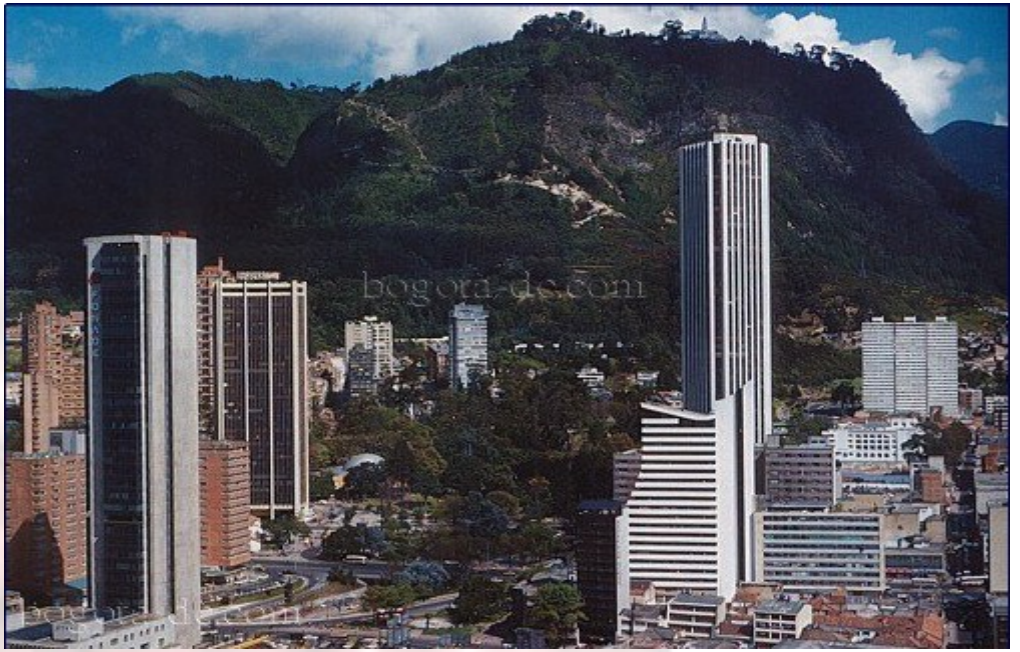
Introducción a la Bioinformática

Centro de Bioinformática
Instituto de Biotecnología
Universidad Nacional de Colombia

Andrés M. Pinzón
cPhD - Universidad de los Andes

7'000.000 de habitantes

Capital Mundial del Libro 2007



2600m sobre el nivel del mar

SEMANA CULTURA
2006

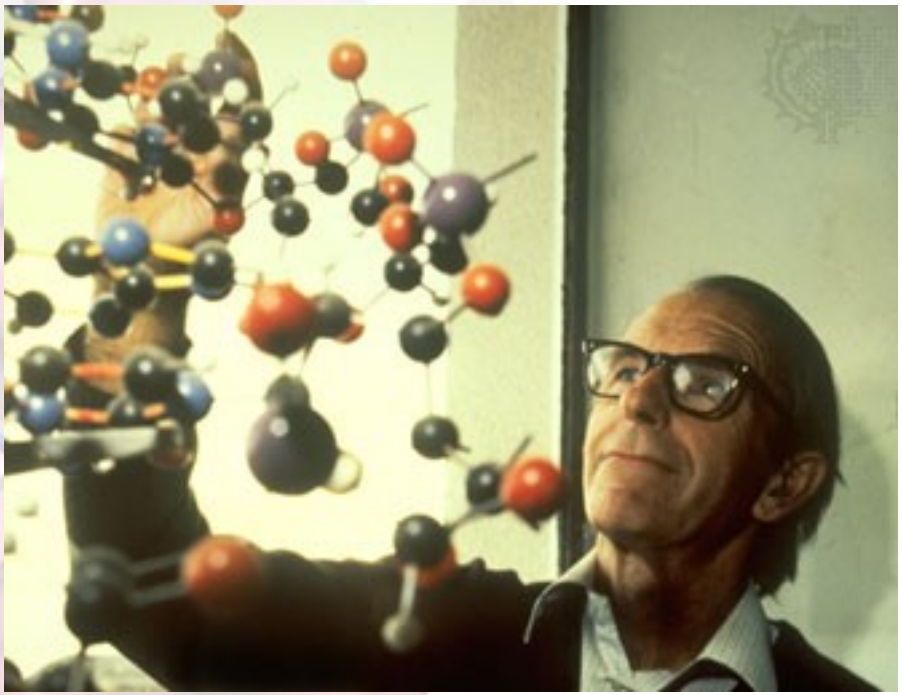
LA UN ES
PUBLICA
NO SE VENDE. NO!
NI SE PRIVATIZA

LA UN ES PUBLICA
NO SE VENDE NI
SE PRIVATIZA

“Actualmente en Biología el número de datos es mucho mayor a la capacidad de análisis del investigador”.

1958 Primera secuencia de una proteína (insulina. Sanger, Premio Nobel)[1]

1975 Primera secuencia de ADN (fago PHI-X174 Sanger, Nobel 1980)



1986 desarrollo de PCR

… inicio de la era genómica!!

1 http://www.genetics.org/cgi/content/full/162/2/527

BD para las ciencias de la vida

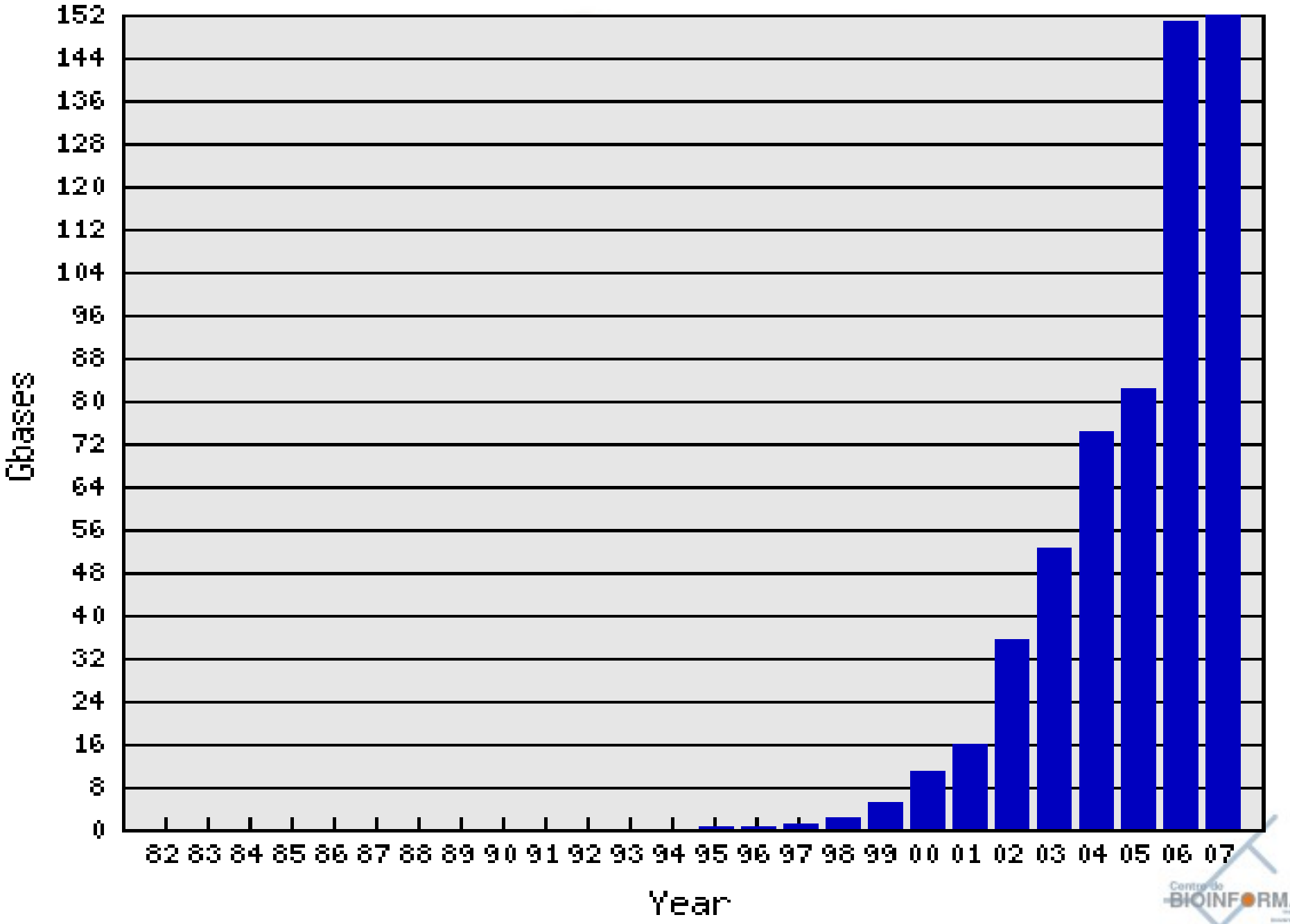
- DNA
- Proteínas
- Genómicas
- Dominios/familias protéicas
- Mutación/polimorfismo
- Proteómica (2D gel, MS)
- 3D estructura
- Metabolismo
- Bibliografía
- SNPs, ESTs, Microsatélites, etc..
-

Algunas estadísticas

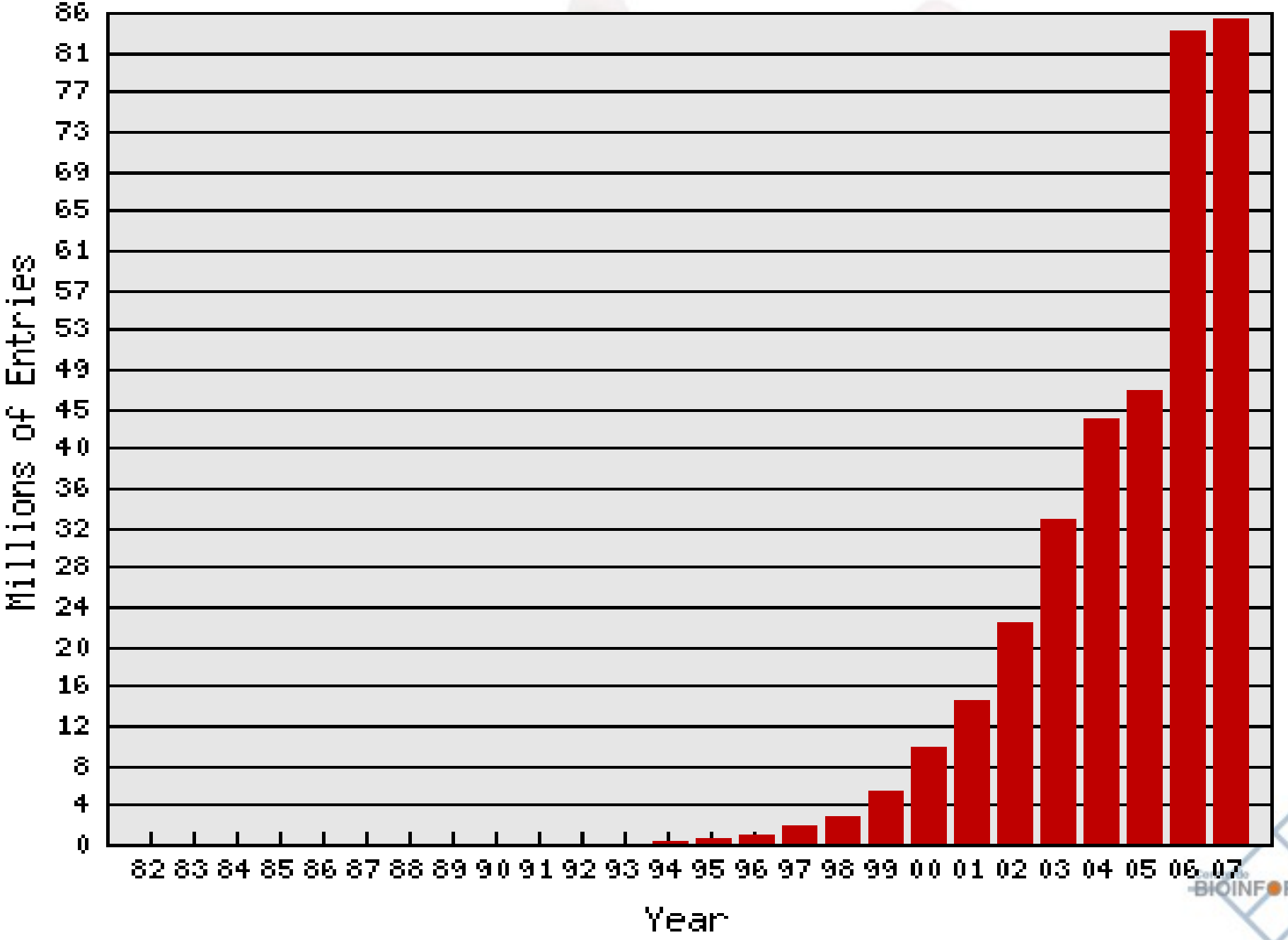
- Más de 1000 bases de datos
- Generalmente accesibles vía WEB
  - Biohunt: http://www.expasy.org/BioHunt/
  - Amos' links: www.expasy.ch/alinks.html

- Tamaño variable: 100Kb a 100Gb
  - DNA (EMBL): > 68 Gb (comprimida!)
  - 80,591,891 entradas.
  - 146,595,277,574 nucleótidos.

Crecimiento de la base datos EMBL en gigabases.



Crecimiento de la base datos EMBL en millones de entradas.



Sin embargo...

Tenemos las secuencias pero...

¿Como podemos saber qué partes de ese DNA controlan los diversos procesos químicos de la vida?

Conocemos la función y estructura de algunas proteínas pero...

Cómo determinamos nuevas funciones?
Como predecir la conformación espacial de una proteína, basados únicamente en su secuencia?

Entendemos el código genético pero...

Como encontrar nuevas palabras significativas que podamos añadir al diccionario del DNA?

BIOINFORMÁTICA

Biología Computacional
Aplicación de técnicas analíticas y cuantitativas para el modelamiento de sistemas biológicos.

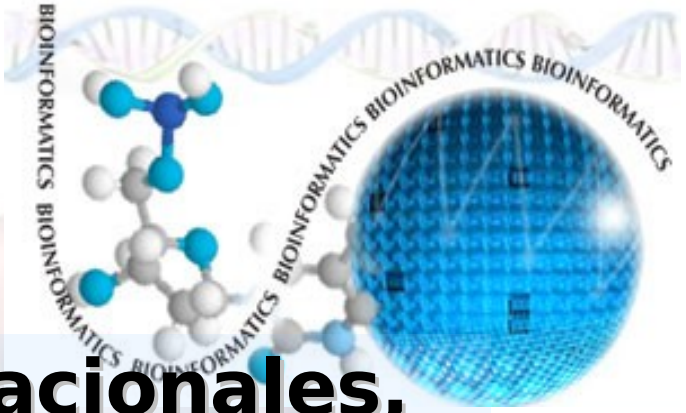
"La bioinformática comprende los métodos matemáticos, estadísticos y computacionales que pretenden solucionar problemas biológicos usando secuencias de ADN y aminoácidos e información relacionada".

Fredj Tekaia - Instituto Pasteur

“La bioinformática es el estudio de la información biológica desde su almacenamiento en el genoma hasta la obtención de los productos génicos en la célula” esto involucra la creación y desarrollo de tecnologías informáticas y computacionales para la resolución de problemas en biología molecular”.

Stanford Center for Profesional Development, 2002

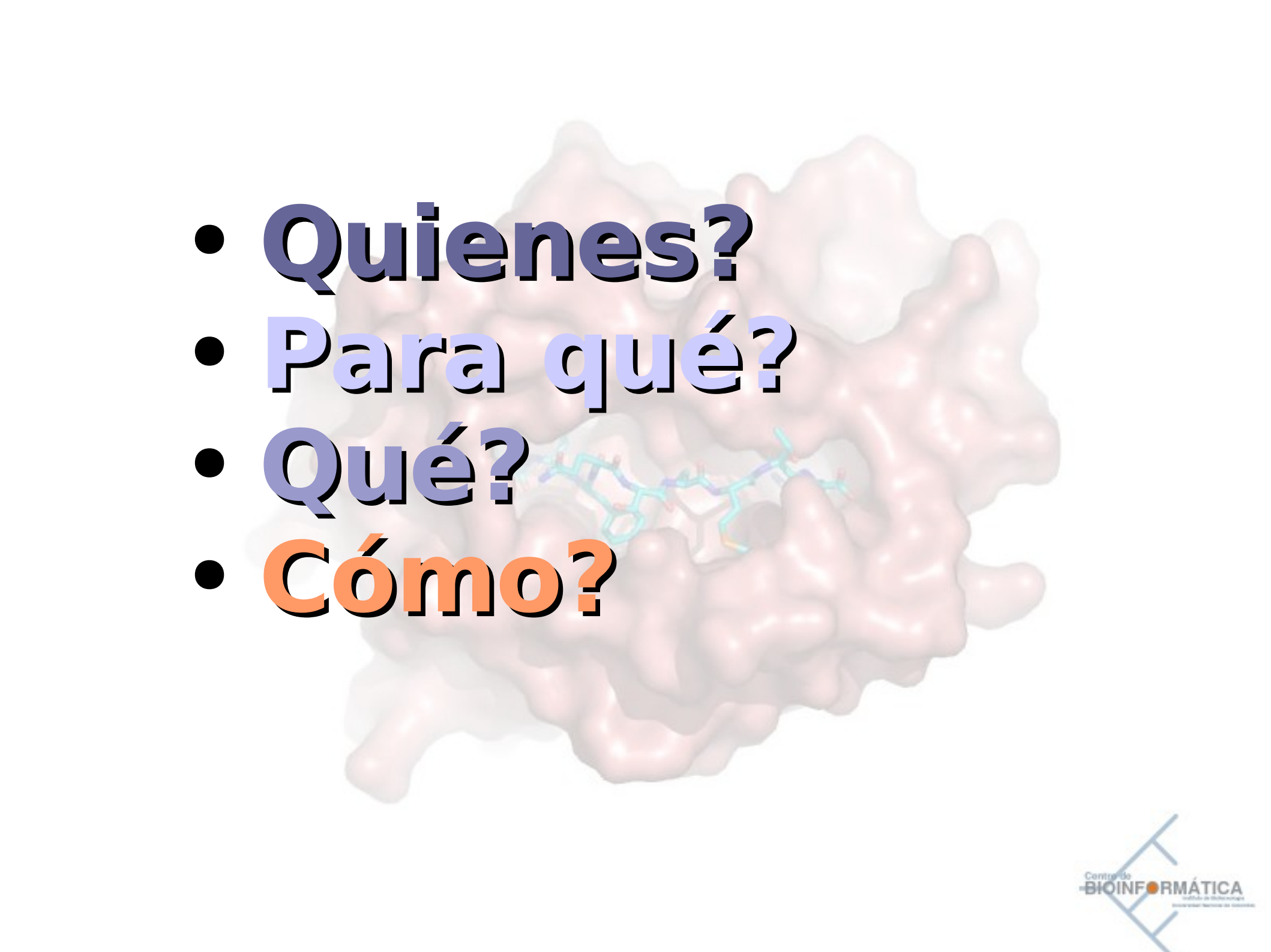
Nuestra definición...



Uso de técnicas computacionales, matemáticas y estadísticas para el análisis, interpretación y generación de datos biológicos.

CARACTERISTICAS:

- INTERDISCIPLINA Y COLABORACIÓN ENTRE GRUPOS.
- INTEROPERATIVIDAD E INTERDEPENDENCIA DE LOS DATOS.
- FORMACIÓN DE REDES.



- Quienes?
- Para qué?
- Qué?
- Cómo?

Centro de BIOINFORMÁTICA

¿Qué personas pueden hacer Bioinformática?

Necesario:

- Conocimiento y entendimiento del Dogma Central de la Biología molecular.
- Conocimiento en Biología Molecular (bioquímica, biología molecular, biofísica molecular).

Muuuuuuuuuuy recomendado:

- Conocimiento en el manejo de sistemas de cómputo.

Recomendado:

- Manejo básico de linea de comandos en ambientes UNIX (GNU/Linux).

Muy deseable:

- Experiencia con algún lenguaje de programación.

¿Qué se busca con el uso de la Bioinformática?

“Profundizar en nuestro entendimiento acerca de los organismos vivos y sus relaciones, partiendo desde el genoma que les codifica”.

Los análisis potenciales en el campo de la Biología molecular son tan diversos como la vida misma.

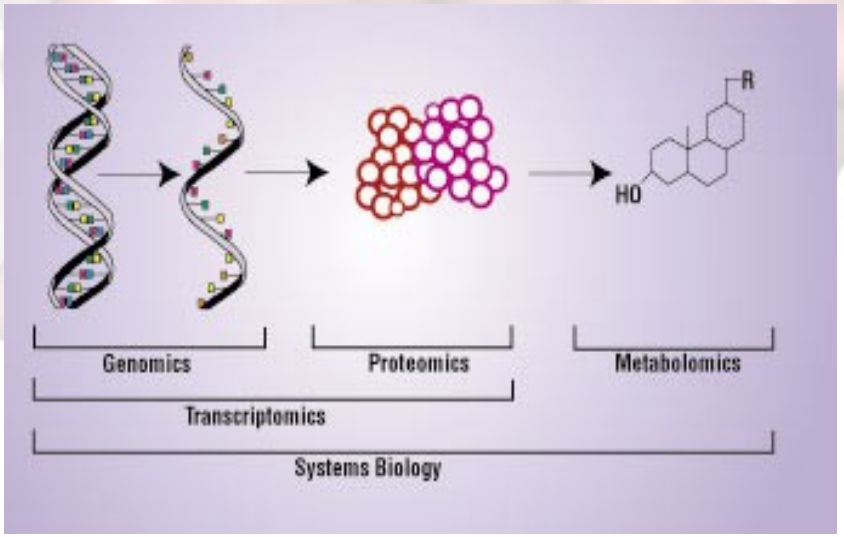
- Genómica comparativa.
- Análisis de DNA (ORFs, Contenidos GC, etc).
- Recuperación de secuencias.
- Ensamblaje de secuencias.
- Predicción de estructuras protéicas.
- Visualización de estructuras protéicas.
- Microarreglos.
- PCR.
- Filogenia.
- Educación.

La bioinformática provee algoritmos, bases de datos, interfaces y herramientas estadísticas para resolver nuestras preguntas!

Centro de BIOINFORMÁTICA

Bioinformática y "ómicas"

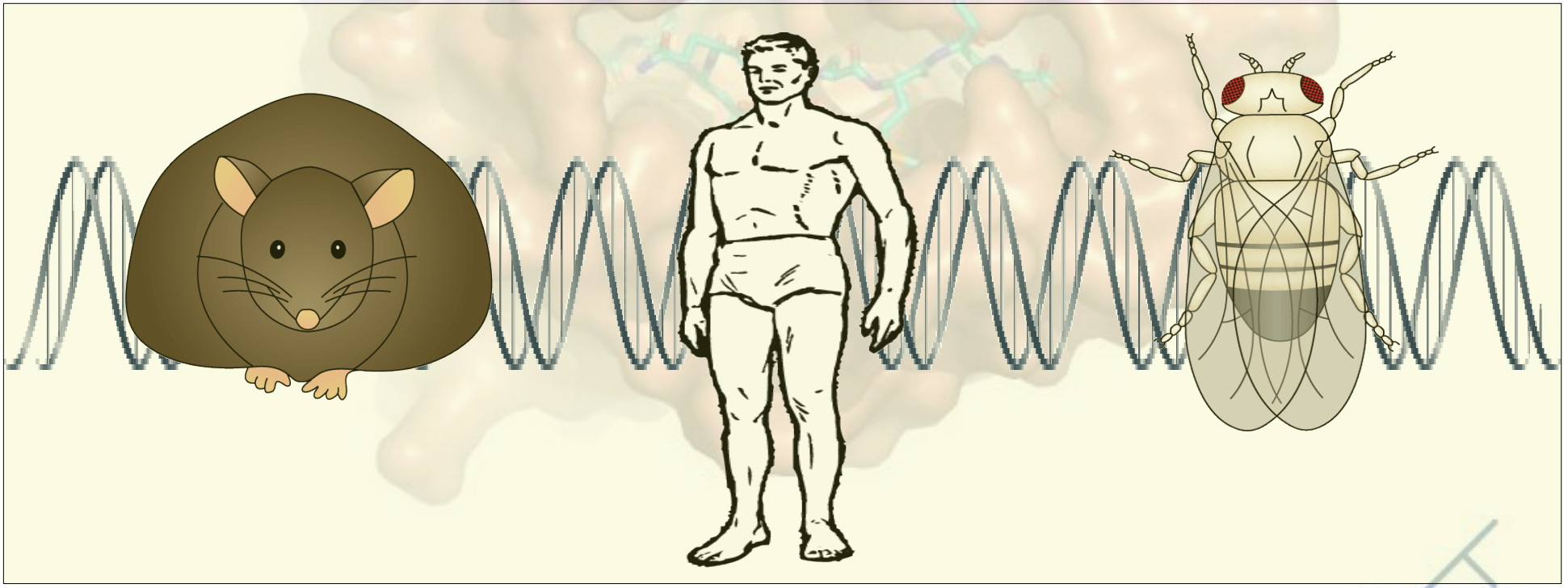
- **Genómica:** Estudia el genoma de los organimos. Uso sistemático de la información genómica y su asociación con otros datos.
- **Transcriptómica:** transcriptoma es el conjunto de todos los mRNAs (o transcriptos), dado un conjunto de condiciones externas. El genoma es fijo (ok, existe polimorfismo genético!) el transcriptoma varia dependiendo del contexto.
- **Proteómica:** es el estudio a gran escala de las proteínas, particularmente su estructura y función.
- **Metabolómica:** estudia el conjunto completo de metabolitos (intermediarios metabólicos, hormonas, metabolitos secundarios etc.) encontrados en una muestra biológica, por ejemplo un organismo.



Genómica comparativa

Alineamiento de genes de diferentes especies buscando secuencias en común.

- Relaciones evolutivas.
- Clasificación de familias de genes.
- Tiempos de divergencia.



Genómica comparativa

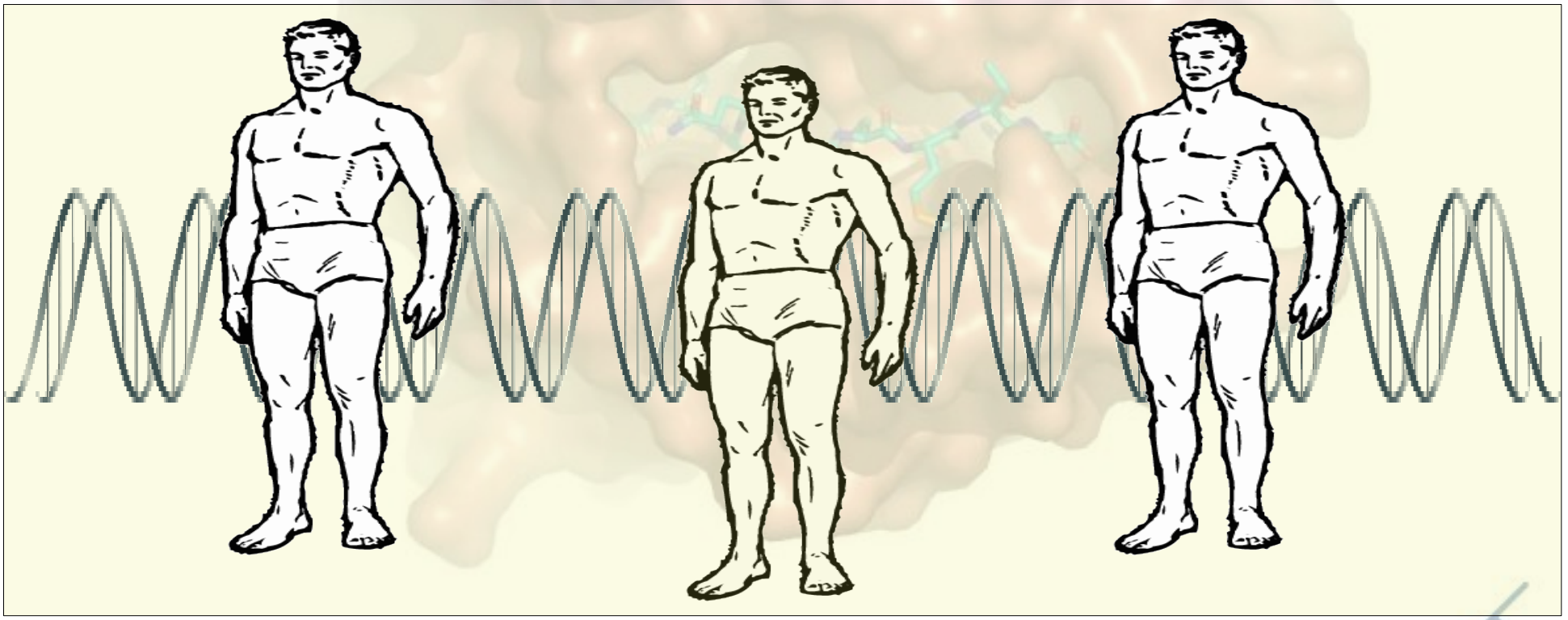
Species	Similarity	Source
Human	99.9%	quoted by U.S.A. President Clinton, Jan 2000, State of the Union address; also Human Genome Project
	100%	identical twins
Chimpanzee	98.4%	sources: Americans for Medical Progress; Jon Entine in the San Francisco Examiner
	98.7%	Richard Mural of Celera Genomics, quoted on MSNBC
Bonobo		equal to chimpanzee
Gorilla	98.38%	based on study of intergenic nonrepetitive DNA in *Am J Hum Genet*. (2001) Feb;**682**:444-56
Mouse	98%	source: Americans for Medical Progress
	85%	comparing all protein coding sequences, NHGRI
Dog	95%	Jon Entine in the *San Francisco Examiner*
C. elegans	74%	Jon Entine in the San Francisco Examiner
Banana	50%	source: Americans for Medical Progress
Daffodil	35%	Steven Rose in *The Guardian* 22 January 2004

Centro de BIOINFORMÁTICA

Genómica comparativa

Alineamiento de genes de iguales especies buscando regiones en común.

- Identificación de polimorfismos.
- Medicamentos personalizados (farmacogenética/farmacogenómica).
- Tratamiento de enfermedades a nivel genético.



Mismas herramientas ... preguntas diferentes!

Ensembl

Search all Ensembl: Anything Go

Ensembl v39 - Jun 2006 Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server

What's New in Ensembl 39

- Honeybee dropped from Ensembl (*Apis mellifera*)
- New Opossum assembly and genebuild (*Monodelphis domestica*)
- New Mouse assembly and genebuild (*Mus musculus*)
- New Ciona savignyi assembly and genebuild (*Ciona savignyi*)
- New Zebrafish assembly and genebuild (*Danio rerio*)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other sites using the Ensembl system

- EBI Genome Reviews database - mainly archaea and bacteria.
- VEGA - Vertebrate Genome Annotation

More...

Mammalian genomes

Homo sapiens
NCBI 36 | Vega

Pan troglodytes
PanTro 1.0

Macaca mulatta
MMUL 0.1 | *pre!*

Mus musculus
UPDATED! NCBI m36 | Vega

Rattus norvegicus
RGSC 3.4

Pre! *Oryctolagus cuniculus*
RABBIT

Canis familiaris
CanFam 1.0 | Vega | *pre!*

Bos taurus
Btau 2.0

Pre! *Dasypus novemcinctus*
ARMA

Pre! *Loxodonta africana*
BROAD E1

Pre! *Echinops telfairi*
TENREC

Monodelphis domestica
UPDATED! MonDom 4

Other species

Gallus gallus
WASHUC 1

Xenopus tropicalis
JGI 4.1

Danio rerio
UPDATED! Zv 6 | Vega

Takifugu rubripes
FUGU 4.0

Tetraodon nigroviridis
TETRAODON 7

Pre! *Gasterosteus aculeatus*
BROAD S1

Ciona intestinalis
JGI2

Ciona savignyi
NEW! CSAV 2.0

Drosophila melanogaster
BDGP 4

Anopheles gambiae
AgamP3

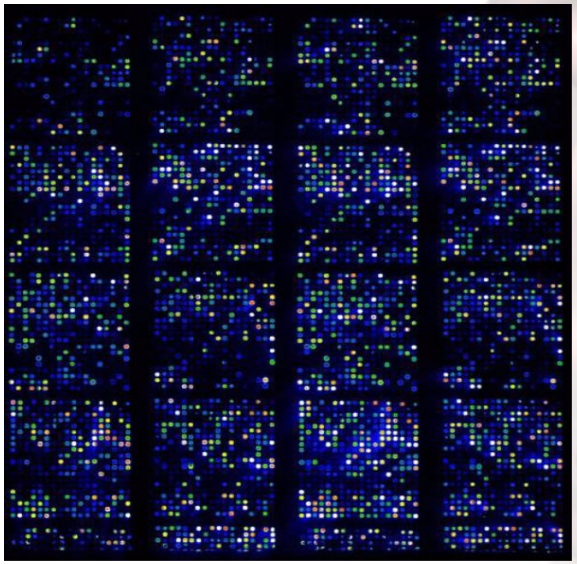
Pre! *Aedes aegypti*
AEDES 1

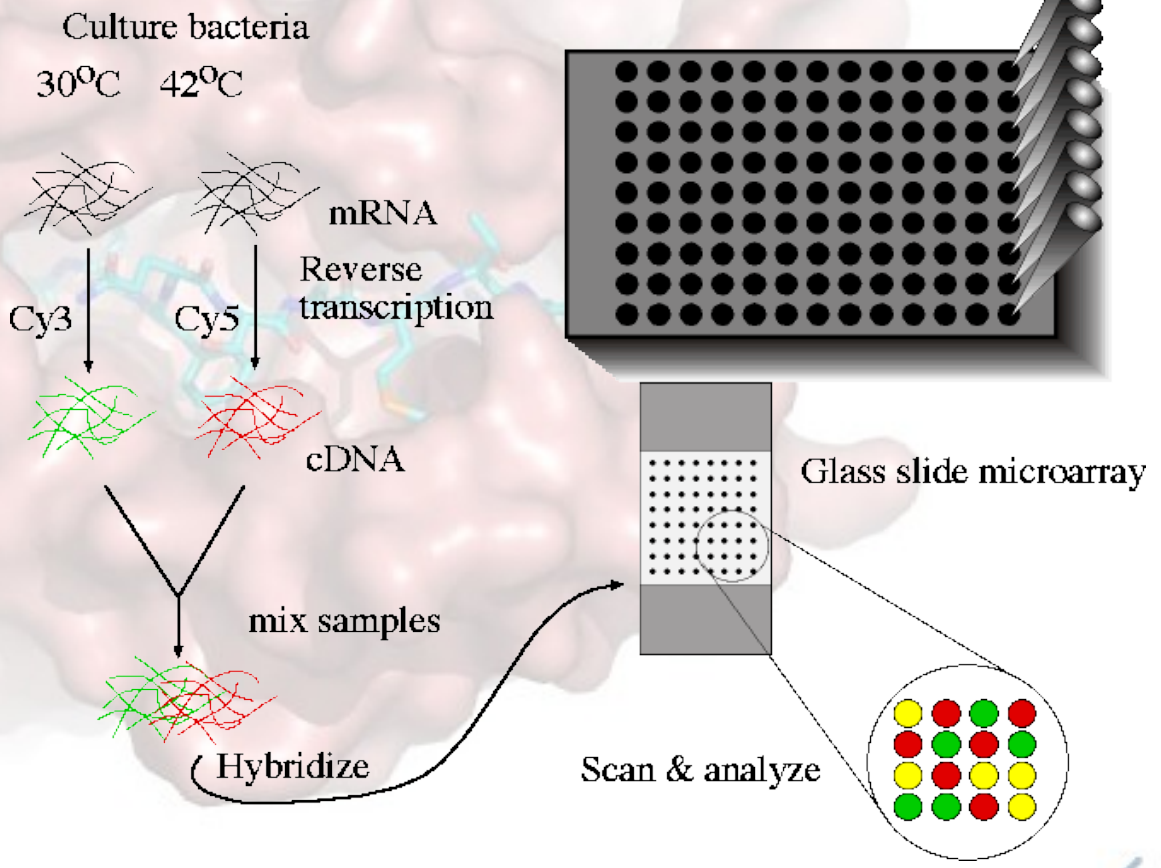
Caenorhabditis elegans
WS 150

Saccharomyces cerevisiae
SGD 1

Genómica funcional

¿Qué genes se expresan, dónde, cuando, función?



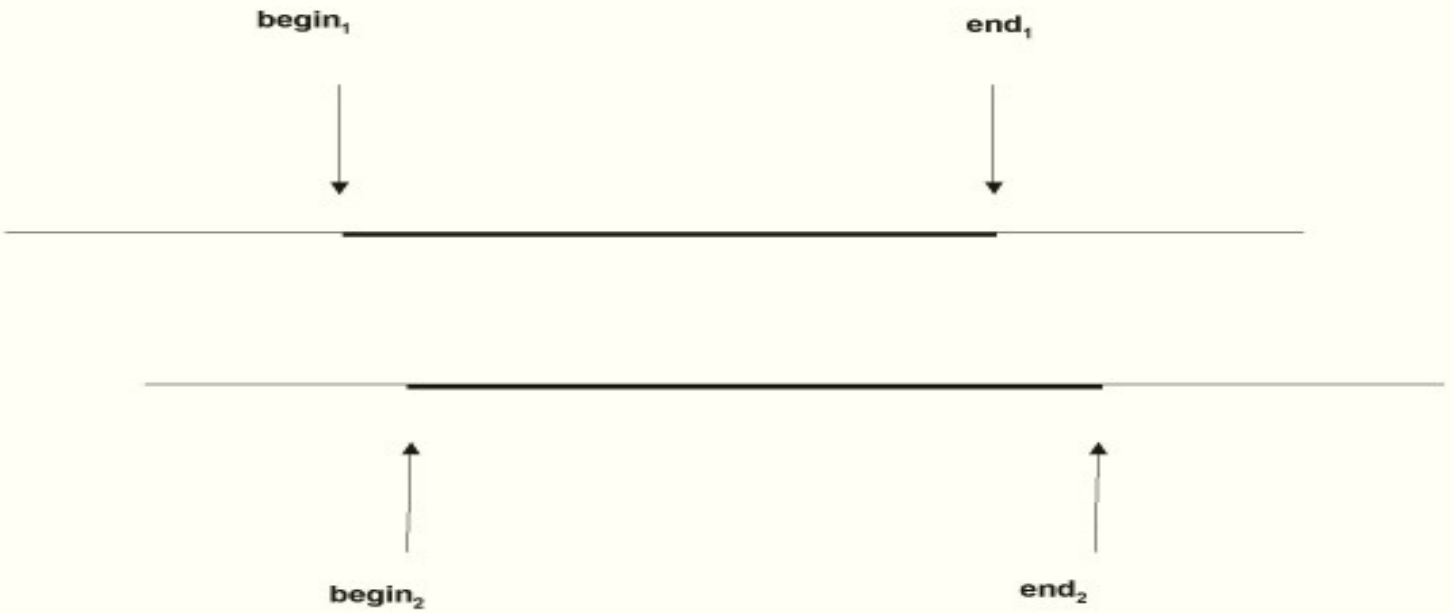


Microarreglos

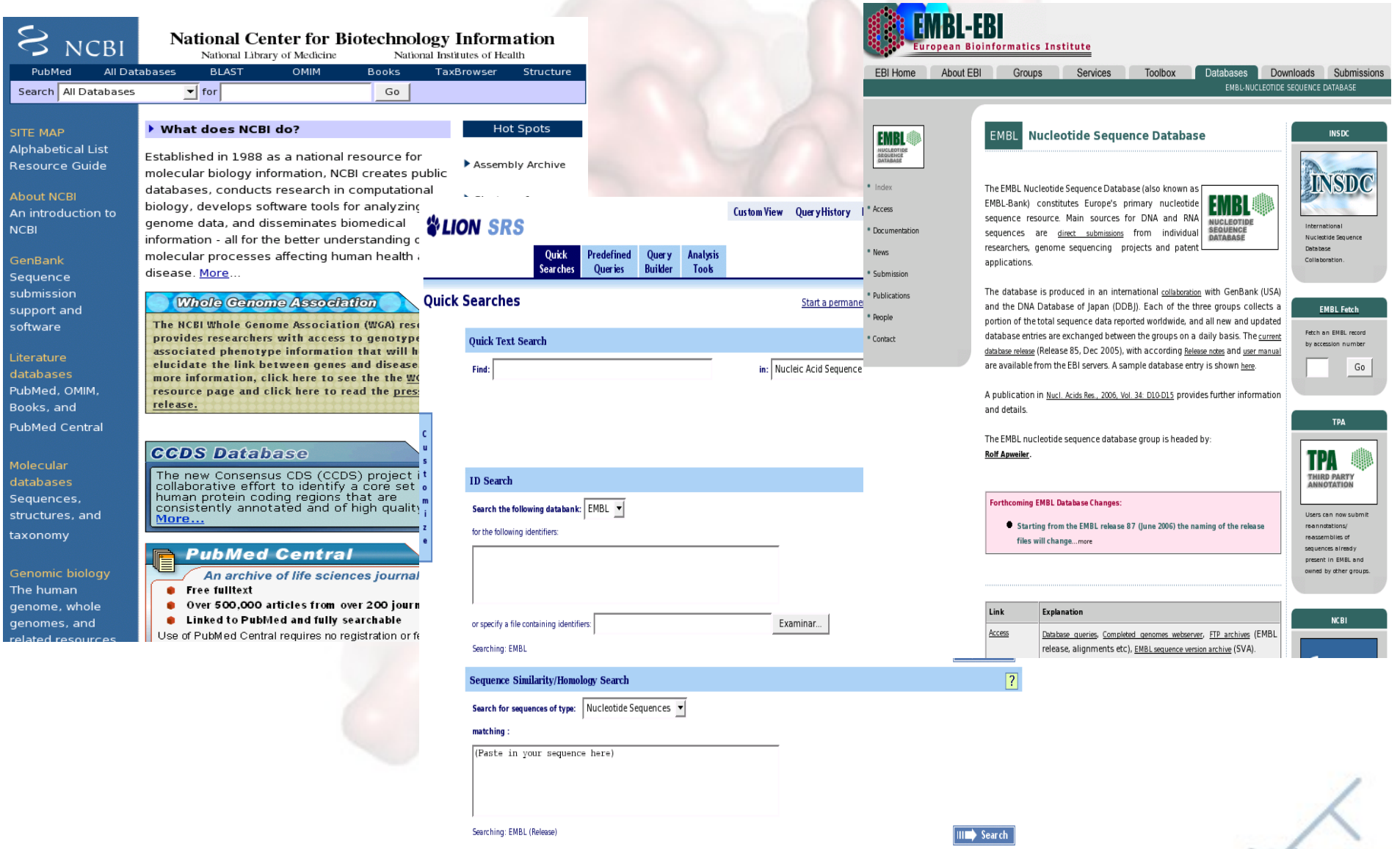
Alineamiento de secuencias

Sec. 1 MALVEDNNAVAVSFLRNSFSEEQILVKYGVDG

Sec. 2 HKSFEDNNAVAVSSEFALLEQILVKDG



Recuperación de secuencias



http://www.ncbi.nlm.nih.gov/

Centro de BIOINFORMÁTICA

- Index
- Access
- Documentation
- News
- Submission
- Publications
- People
- Contact

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 85, Dec 2005), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucl. Acids Res., 2006, Vol. 34: D10-D15 provides further information and details.

The EMBL nucleotide sequence database group is headed by:
Rolf Apweiler.

Forthcoming EMBL Database Changes:

- **Starting from the EMBL release 87 (June 2006) the naming of the release files will change**...more

Link	Explanation
Access	Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA).

INSDC

International Nucleotide Sequence Database Collaboration.

EMBL Fetch

Fetch an EMBL record by accession number

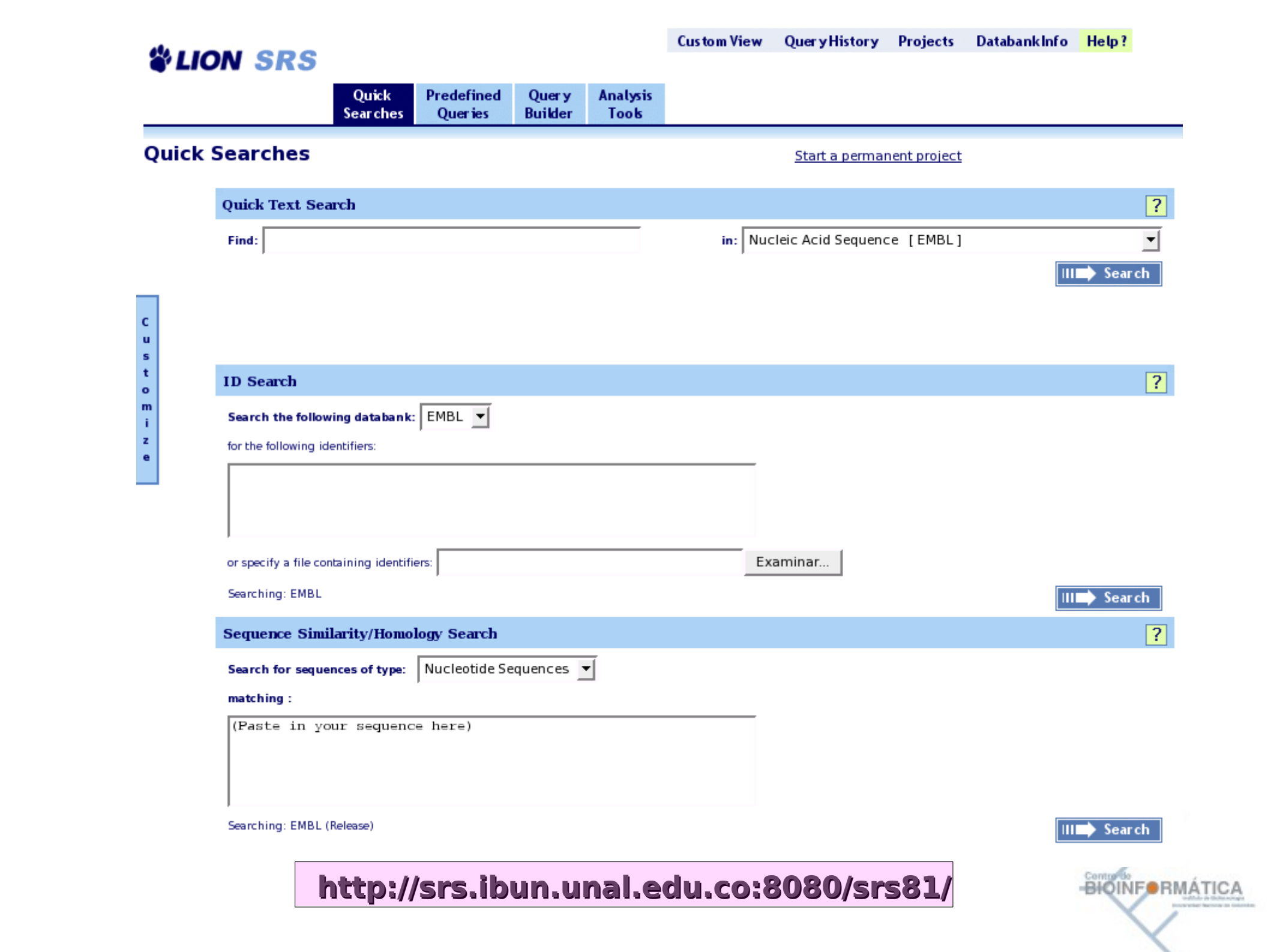
Go

TPA

Users can now submit re-annotations/ re-assemblies of sequences already present in EMBL and owned by other groups.

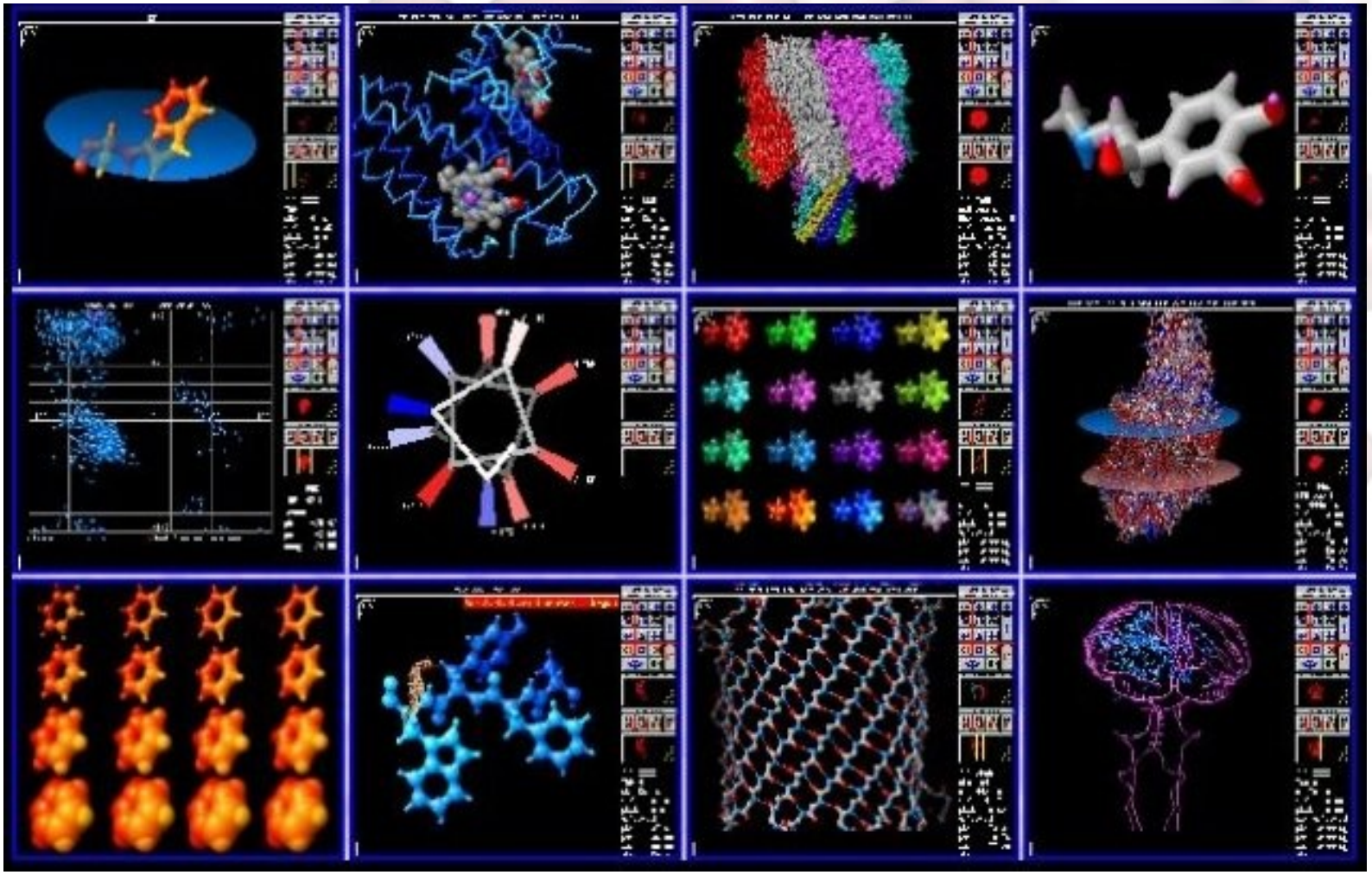
NCBI

http://www.ebi.ac.uk/embl/


LION SRS
Custom View
Query History
Projects
DatabankInfo
Help?
Quick Searches
Predefined Queries
Query Builder
Analysis Tools
Quick Searches
Start a permanent project
Quick Text Search
Find:
in:
Nucleic Acid Sequence [EMBL]
Search
Customize
ID Search
Search the following databank:
EMBL
for the following identifiers:
or specify a file containing identifiers:
Examinar...
Searching: EMBL
Search
Sequence Similarity/Homology Search
Search for sequences of type:
Nucleotide Sequences
matching :
(Paste in your sequence here)
Searching: EMBL (Release)
Search
http://srs.ibun.unal.edu.co:8080/srs81/
Centro de BIOINFORMÁTICA

Visualización de estructuras protéicas

¿Visualización o predicción?

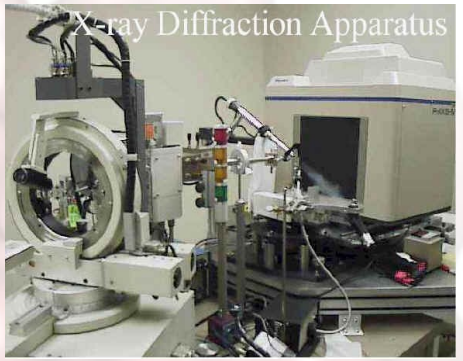


GARLIC

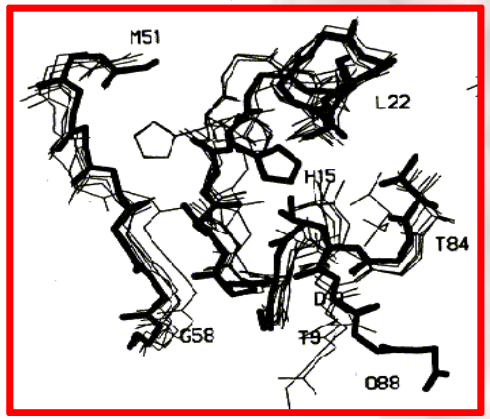
Centro de BIOINFORMÁTICA

¿Predicción o determinación?

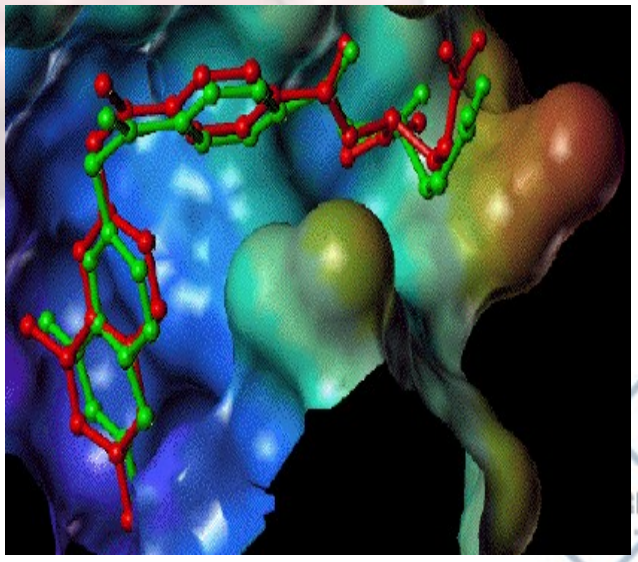
MLDTNMKTQLKAYLEKLTKPVELIATLDDSAKSAEIKELL
TKPVELIATLDDSAEIKELLSAIAMLKSAEIKELLAEIKELL
LKAYLEKLTKPMLDTNMKTQLKKSAEIKELLKSAEIKELL



Cristalografía de Rayos X



Predicción 2D

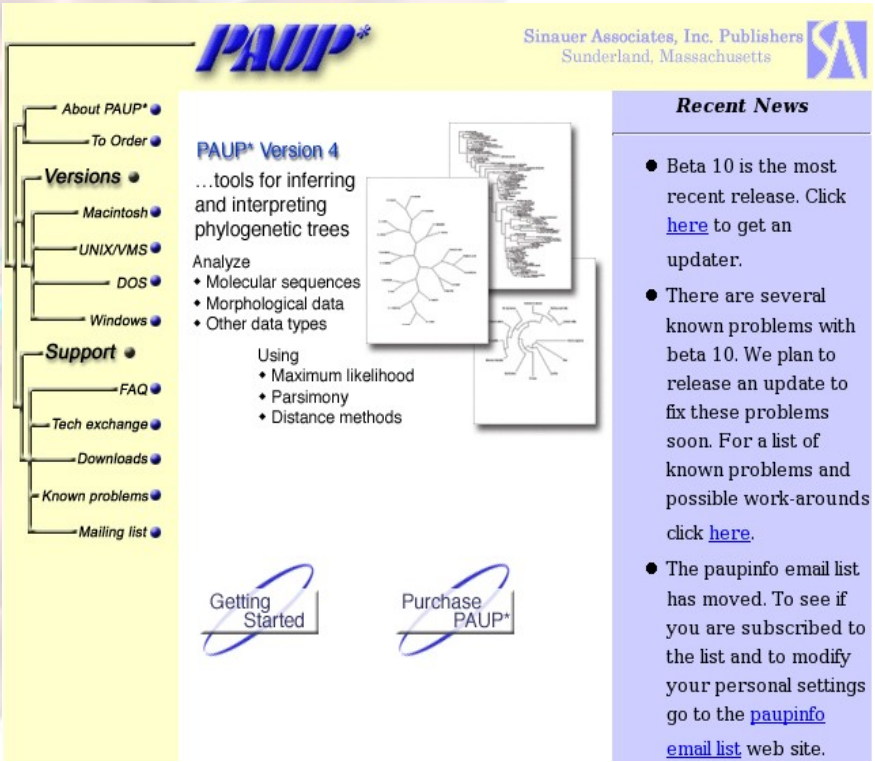


- Modelamiento homólogo (modelamiento comparativo).
- Métodos Ab initio

Análisis Filogenéticos

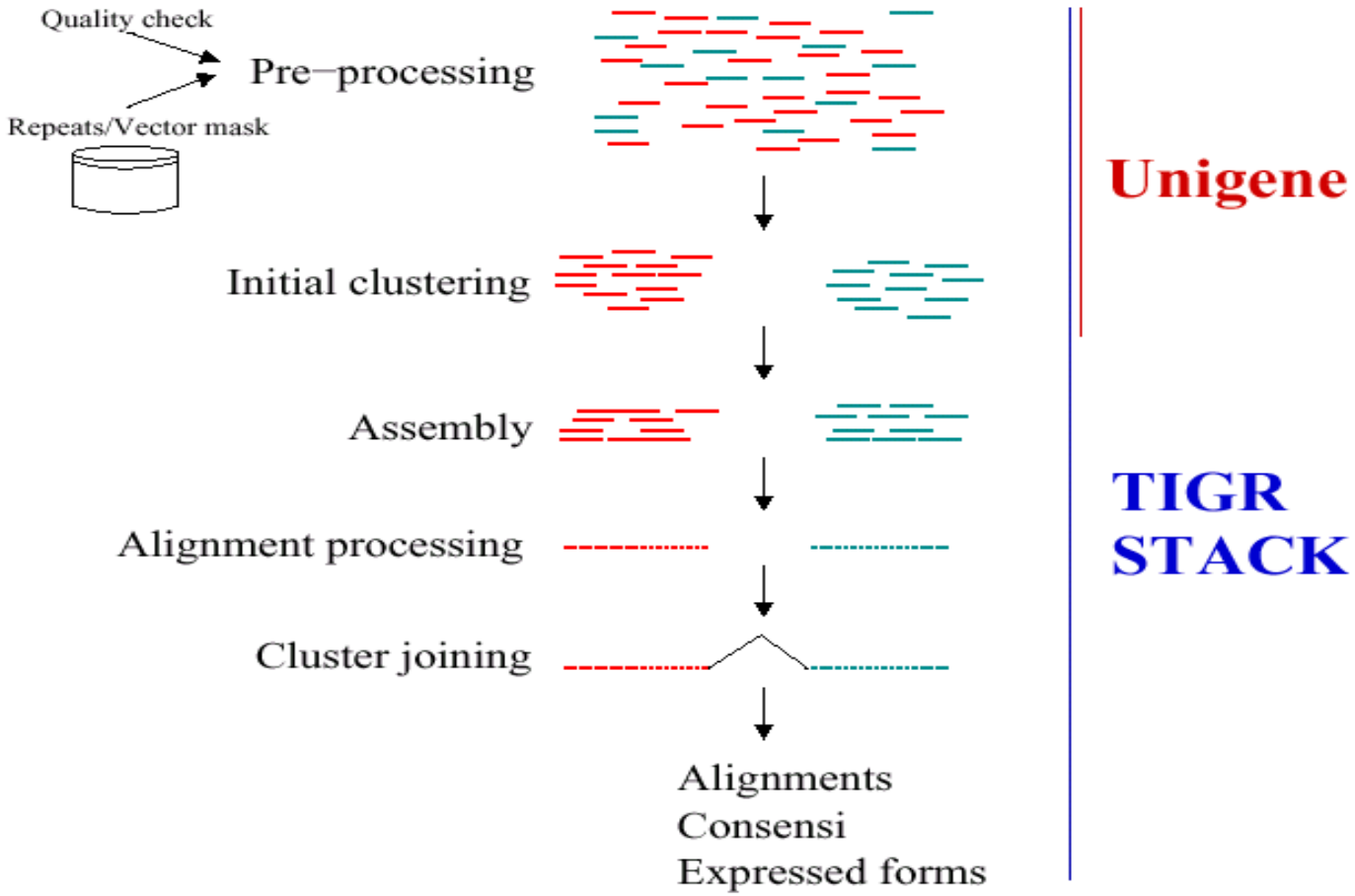
- **Evolución molecular de familias de proteínas**
- **Creación de árboles taxonómicos**
- **Reconstrucción evolutiva de rutas metabólicas.**

- QuickTree: Reconstrucción de árboles filogenéticos.
- Phylip (Phylogenie Inference Package).
-Mavric (Python): Manipulación y visualización de árboles.
- PAUP.



Ensamblaje de Secuencias

EST clustering pipeline



Nuestra Bioinformática

http://bioinf.ibun.unal.edu.co

EMBnet Colombia - Centro de Bioinformática del Instituto de Biotecnología - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://bioinf.ibun.unal.edu.co/

Ir paup

Cross Cron BEL SEF Manuales-LuCAS BLAST program ... Dictionary Gmail The Artemis Man... Msatfinder manual ISCB

Gmail - Inbox (11) | PHYLIP Home Page | PAUP* 4.0 | (Sin Título) | Developing Bioinformatics ... | EMBnet Colombia - Cen...

EMBnet Colombia

UNIVERSIDAD NACIONAL DE COLOMBIA

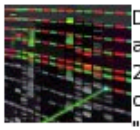
Herramientas

| EMBOSS | BLAST | BLEE | SMS SEWER | BACTNAME | SINCO | DOTLET

Completa suite con más de 100 herramientas para análisis moleculares.

Puede consultar nuestra página de ayuda para mayor información acerca de estas herramientas.

Curso Microarreglos



Del 28 de Noviembre al 2 de Diciembre de 2005 tendra lugar el curso Intrnacional: **"Introduccion al Analisis de Microarreglos"**, organizado por el CBIB en colaboracion con varios grupos e institutos a nivel mundial. Mayor informacion

Curso Bioinformatica

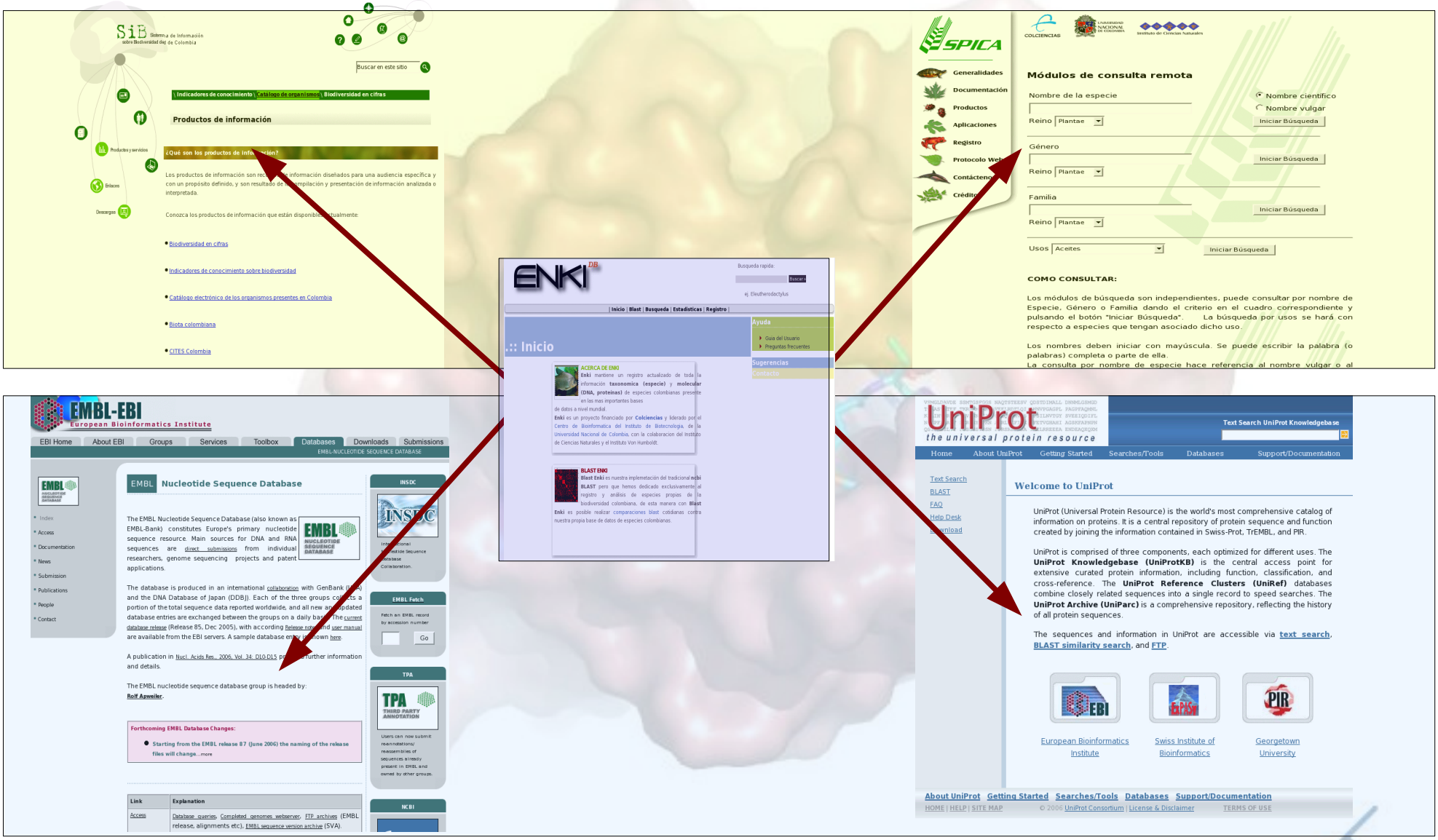
EMBnet Colombia

EMBnet Colombia es el nodo Colombiano de la **Red Europea de Biología Molecular** (Embnet) , una red mundial de centros en bioinformática con nodos en los 5 continentes. EMBnet Colombia se encuentra a cargo del **Centro de Bioinformática del Instituto de Biotecnología** (CBIB) de la Universidad Nacional de Colombia.

Centro de Bioinformática - Instituto de Biotecnología 2005
Universidad Nacional de Colombia

http://bioinf.ibun.unal.edu.co/wEMBOSS

Biodiversidad y Bioinformática



Centro de BIOINFORMÁTICA

Biodiversidad y Bioinformática

ENKI DB

Busqueda rapida:

Buscar »

ej. Eleutherodactylus

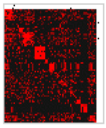
| Inicio | Blast | Busqueda | Estadisticas | Registro |

.:: Inicio

ACERCA DE ENKI

Enki mantiene un registro actualizado de toda la información **taxonomica (especie)** y **molecular (DNA, proteínas)** de especies colombianas presente en las mas importantes bases de datos a nivel mundial.

Enki es un proyecto financiado por **Colciencias** y liderado por el Centro de Bioinformatica del Instituto de Biotecnologia, de la Universidad Nacional de Colombia, con la colaboracion del Instituto de Ciencias Naturales y el Instituto Von Humboldt.



BLAST ENKI

Blast Enki es nuestra implemetación del tradicional **ncbi BLAST** pero que hemos dedicado exclusivamente al registro y análisis de especies propias de la biodiversidad colombiana, de esta manera con **Blast Enki** es posible realizar comparaciones blast cotidianas contra nuestra propia base de datos de especies colombianas.

Ayuda

- Guia del Usuario
- Preguntas frecuentes

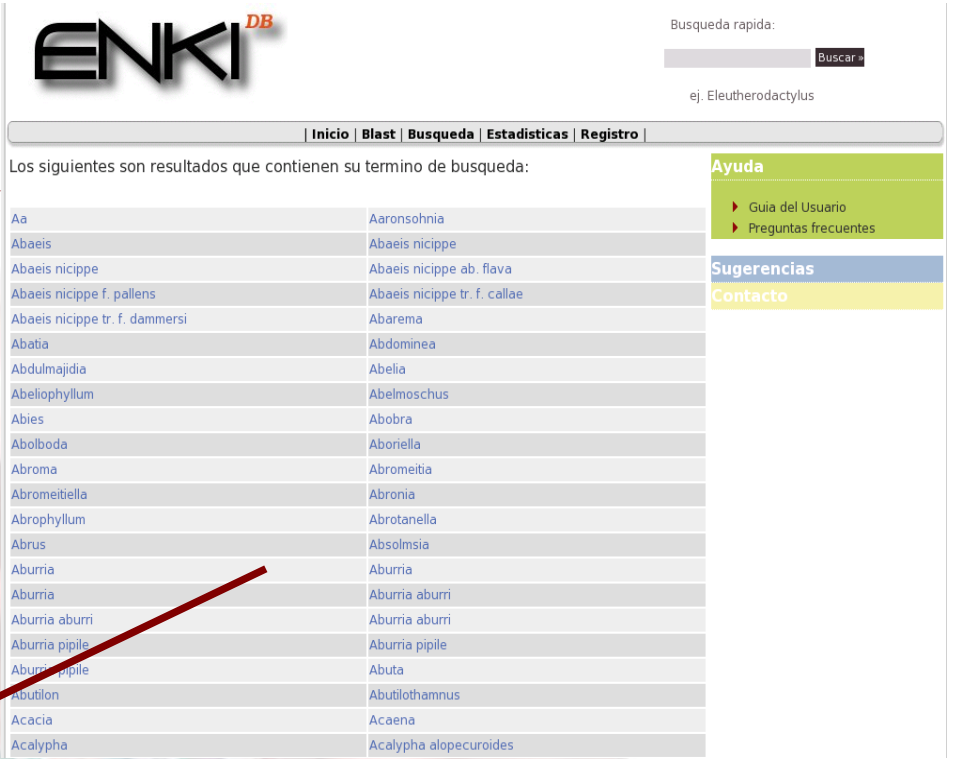
Sugerencias

Contacto

33308 especies registradas.
413983 registros en Uniprot.
Corresponden a: 11779 (35.4%).

11212 registros en EMBL.

Centro de BIOINFORMÁTICA



Centro de BIOINFORMÁTICA

Algunos de nuestros proyectos...

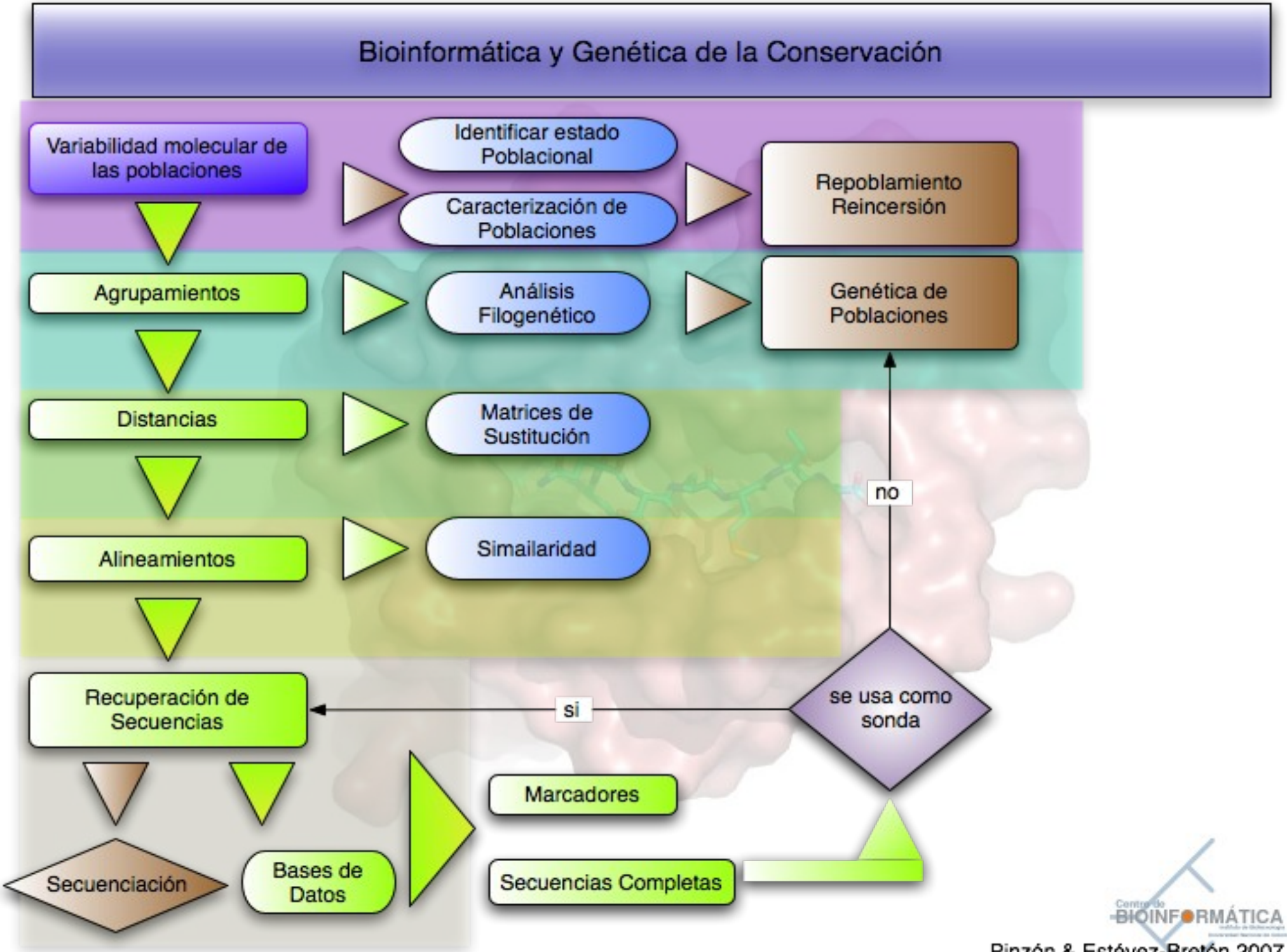
Microsatellites in Phytophtora ESTs: Survey, transferability and association with pathogenesis related genes

Búsqueda e identificación de nuevos candidatos a vacuna contra la Malaria producida por *Plasmodium vivax*.

Búsqueda de dominios específicos para la clase Hexápoda (Phylum Artrópoda).

Modelo para la identificación de genes para Betalactamasas de espectro extendido.

Sequence Retrieve System Federation.


Bioinformática y Genética de la Conservación
Variabilidad molecular de las poblaciones
Identificar estado Poblacional
Caracterización de Poblaciones
Repoblamiento Reincersión
Agrupamientos
Análisis Filogenético
Genética de Poblaciones
Distancias
Matrices de Sustitución
Alineamientos
Simailaridad
no
Recuperación de Secuencias
si
se usa como sonda
Marcadores
Secuenciación
Bases de Datos
Secuencias Completas
Centro de BIOINFORMÁTICA
Pinzón & Estévez-Bretón 2007

...Preguntas?