



# Computación Científica para las Ciencias de la Vida y la Salud

*José R. Valverde*



# Resumen

- Introducción
- Los 90
- El cambio de milenio
- Los 10
- El futuro

# Introducción



Perdón por el inglés

# EMBnet

- En 1990 era obvio que el EMBL no podía atender a toda la comunidad.
- Se estableció una Red de Nodos de Excelencia para proporcionar a nivel nacional
  - herramientas de computación
  - formación
  - desarrollo
  - soporte

Los 90



# Formación

- Análisis de Secuencias
- Evolución
- Cursos generales en cada institución
- Apoyo personalizado para problemas específicos

# Servicios en los 90

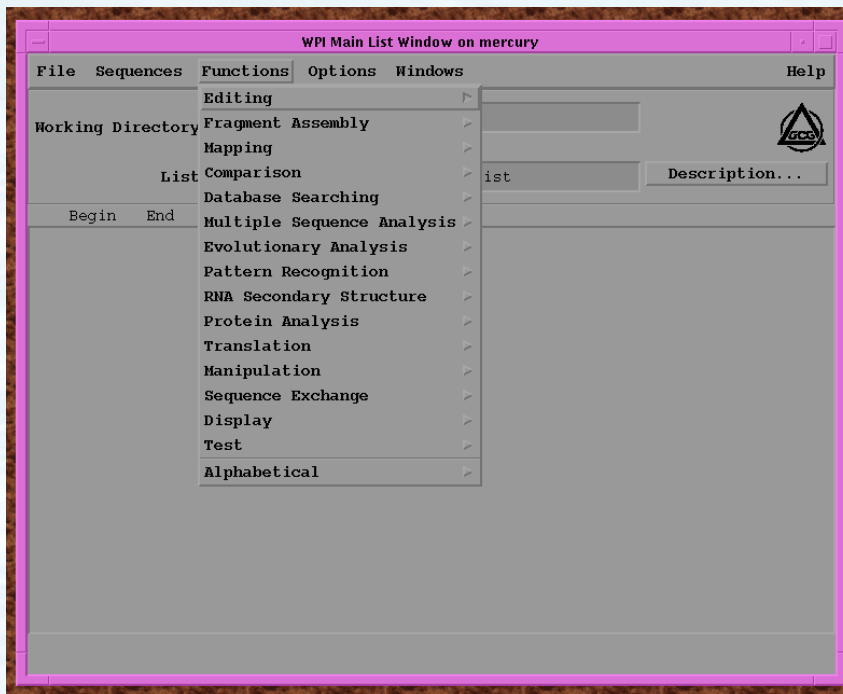
- Ámbito institucional, regional o nacional
- Análisis de Secuencias: GCG, otros
- Filogenia: PHYLIP, GCG-PAUP
- Acceso a Bases de Datos
- Análisis de Imagen
- Análisis de estructura



# GCG

- Exhaustivo

- Secuenciación
- Mapeo
- Comparación
- Evolución
- Estructura 1 y 2D
- Reconocimiento de patrones
- Caro (24-48 K€)



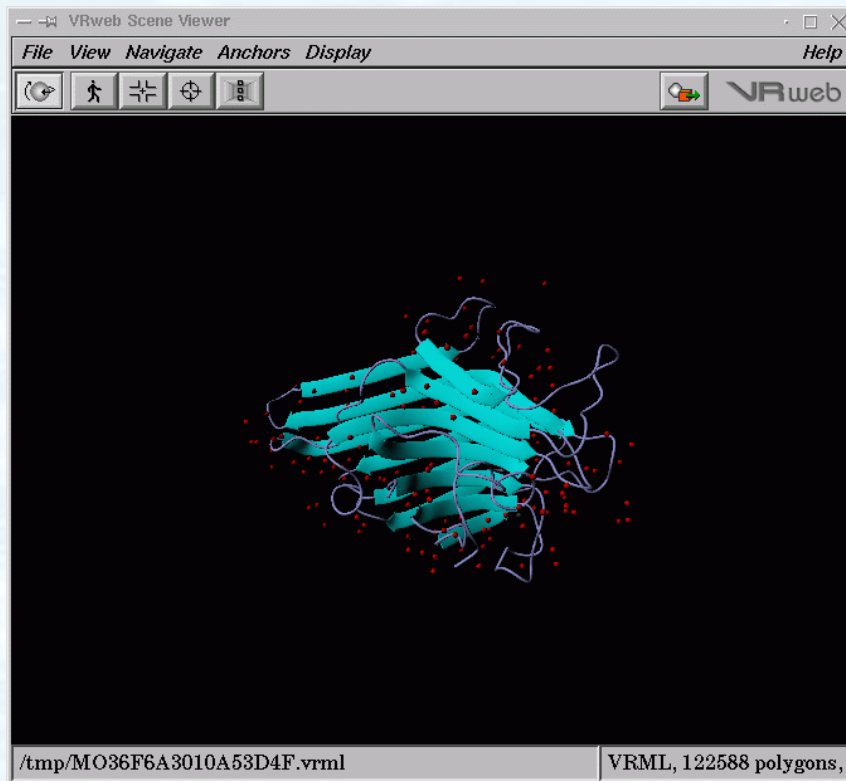
1995 GCG

# Bases de datos

- Copia local
  - SRS: Sequence Retrieval System
    - modelo mixto
    - Muy caro
    - Gratis (acad.)
  - Herramientas académicas
- 1996 SRS

The screenshot shows the SRS web interface with a navigation bar at the top containing buttons for 'Top Page', 'Query Form', 'Query Manager', 'View Manager', 'Databanks', and 'Help'. Below the navigation bar, there is a search section with a 'Search' label and a link to 'SWISSPROT SWISSNEW'. A red arrow points to a 'Do Query' button, followed by a 'Reset' button and a 'Combine searches with' dropdown menu set to 'AND'. There is also a checkbox for 'Append wildcard \*\* to words'. Below this, there are four rows of input fields, each with an 'Info' button and an 'AllText' dropdown menu. At the bottom, there is a section for 'Include fields in output' with a list of fields: ID, AccNumber, Description, GeneName, Keywords, Date, and Organism. There are also options for 'Entry List in chunks of' (set to 5), 'Sequence Format' (set to '\* default \*'), 'Use view' (set to 'SequenceSimple'), and 'Retrieve set of' (set to 'entry').

# Structura Molecular



1999 Virtual Reality

- Servicio especializado
- Alto coste computacional
- Visualización
  - Rasmol/MolMol
  - Virtual Reality
- Resolución
- Dinámica Molecular



# Preparando el milenio

- EMBnet
  - Federación de servicios distribuídos (SRS, blast)
  - Preparar la revolución de las \*ómicas
- Extensión de coordinación internacional
  - Europa se “extiende” a Asia, Australia, África y América
- Mejoras de infraestructura de comunicaciones

# Interfaces Web y EMBOSS

- GCG muy caro
- EMBnet desarrolla EMBOSS

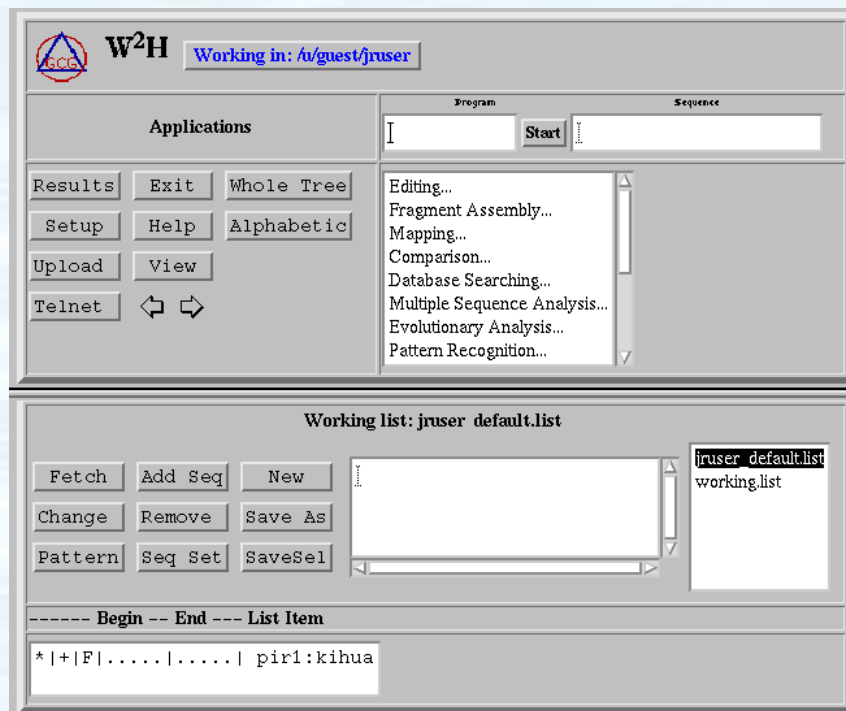
– Exhaustivo

- EMBnet desarrolla interfaces web

– w2h

– www2gcg

– wemboss



1998 w2h

El cambio de milenio



# FLOSS

- Free and Libre Open Source Software
- El software libre reduce costes significativamente
- El ecosistema natural del FLOSS es el público
  - bien común: financiación común
- ¿Cómo puede sobrevivir la *res privada* frente a la *res pública*?

# Formación

- Sigue habiendo alta demanda de formación general en Bioinformática
- Formación especializada en campos específicos
  - Filogenia
  - Estadística
  - Modelización
- Se inicia la autoformación electrónica

# Colaboración en EMBnet (2002)

- 41 Centros de excelencia
  - 31 nodos nacionales
  - 10 nodos especialistas
  - Más de 30.000 usuarios registrados
- Colaboración en
  - Conocimiento
  - Formación
  - Desarrollo
  - Servicios

AR AU AT BE BR CA CH CL  
CN CO CU DK ES FI FR DE  
GR HU IE IN IL IT MX NL  
NO PO PT RU SK SE UK ZA  
EBI ETI EU Roche ICGEB  
Lion MIPS Pharmacia  
Sanger UMBER

**La provisión de  
servicios  
requiere  
coordinación  
para estar al día**



# Nuevos servicios

- Un solo servicio es insuficiente para todas las necesidades de una comunidad
- Nueva especialización requiere nuevos servicios
  - Análisis de Secuencias
  - Proteómica
  - Genómica
  - Modelización
  - Evolución, etc...

# Infraestructura: Hardware (2002)

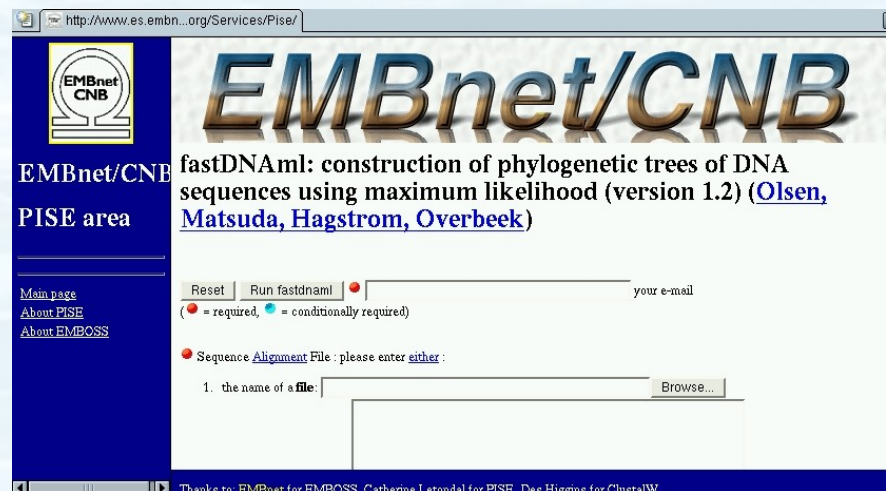
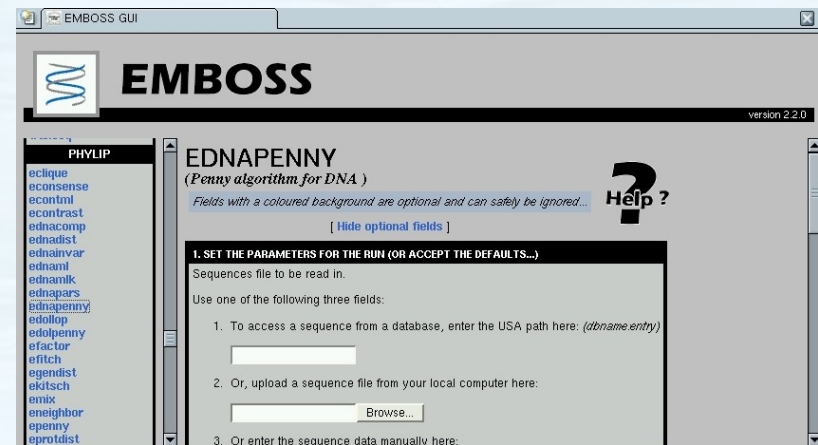
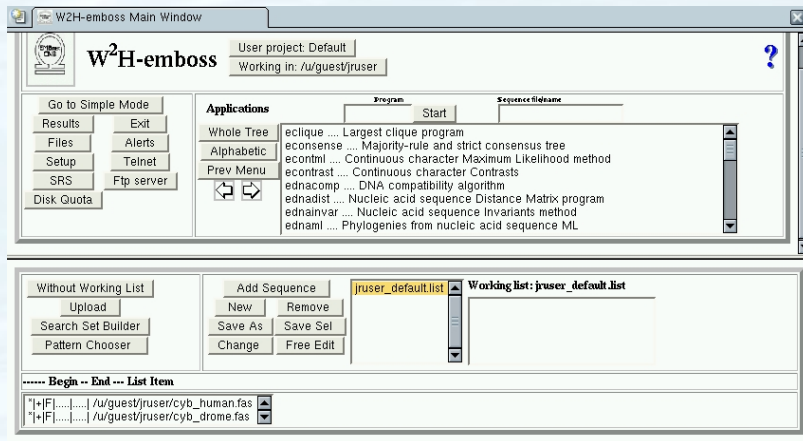
- Los servicios comunes requieren una infraestructura que pueda satisfacer a todos los usuarios. P. ej. EMBnet/CNB
  - 2 SGI PowerChallenge
  - 2 IBM Power
  - ~1TB hard disk storage
  - Cluster Linux IA32 (16 CPU, 16 GB RAM)
  - Estaciones de trabajo
  - Sistema centralizado de backup

# Análisis de Secuencias

- EMBOSS/wEMBOSS
  - EMBnet desarrolla EMBOSS y varios interfaces web (se puede elegir)
  - Inicia una **tendencia a reemplazar software comercial por FLOSS**
- Secuenciación:
  - FLOSS: Staden package, ACeDB, Phred, Phrap, Linkage Analysis...
- **Tendencia al uso vía Web**



# Interfaces web



# Bases de datos

- SRS se mantiene en el mundo académico
- Dificultades para mantenerlas al día
  - Comienza el abandono de las copias locales y regreso a las centrales vía Web
- Inicio de la **aparición de bases de datos especializadas locales**
- MySQL, PostgreSQL van ganando terreno



# Filogenia

- ~~PAUP (GCG)~~
- ClustalX, T-Coffee/Mocca, MAFFT, MUSCLE...
- PHYLIP
- Se popularizan las interfaces WWW para métodos clásicos (UPGMA, ML, MP)
- **Mr.Bayes (nuevas tendencias y especialización)**



# WebPHYLP

WebPhylip

1. Parsimony  
[Help](#) [Run](#)

2. Parsimony +  
Branch&Bound  
[Help](#) [Run](#)

3. Compatibility  
[Help](#) [Run](#)

4. Max. Likeli.  
[Help](#) [Run](#)

5. Max. Likeli.  
with mol. clock  
[Help](#) [Run](#)

---

[Do consensus,  
tree editor](#)

---

[Draw trees](#)

---

[Back to Menu](#)

**DNAPENNY - Branch and bound to find  
all most parsimonious trees  
for nucleic acid sequence parsimony criteria**

DNAPENNY is a program that will find all of the most parsimonious trees implied by your data when the nucleic acid sequence parsimony criterion is employed. It does so not by examining all possible trees, but by using the more sophisticated "branch and bound" algorithm, a standard computer science search strategy first applied to phylogenetic inference by Hendy and Penny (1982). (J. S. Farris [personal communication, 1975] had also suggested that this strategy, which is well-known in computer science, might be applied to phylogenies, but he did not publish this suggestion).

There is, however, a price to be paid for the certainty that one has found all members of the set of most parsimonious trees. The problem of finding these has been shown (Graham and Foulds, 1982; Day, 1983) to be NP-complete, which is equivalent to saying that there is no fast algorithm that is guaranteed to solve the problem in all cases (for a discussion of NP-completeness, see the Scientific American article by Lewis and Papadimitriou, 1978). The result is that this program, despite its algorithmic sophistication, is VERY SLOW.

Use ordinary parsimony ☒ or [threshold](#) parsimony? ☐

Input threshold value (larger than 1):

---

[Simple](#) branch and bound?

How many groups of 100 trees?  After  trees, report progress of run.

---

Input sequence [type](#)?  Number of data sets:

Use previous data set?  if no, type data set below.

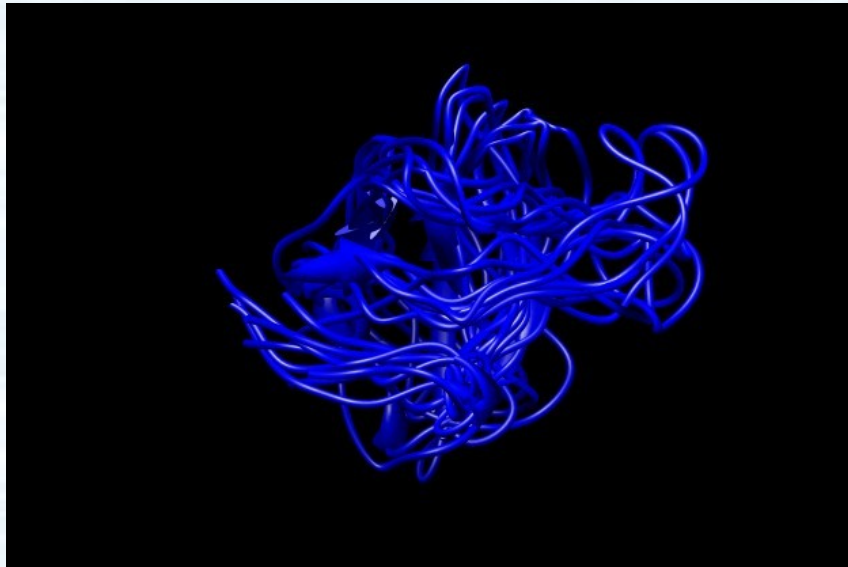
[Formatted](#) Input Sequences: an [example](#)

---

# Modelización Molecular

- Predicción de estructura (web)
- Análisis de estructura (web)
- **Docking & QSAR**
- **Dinámica Molecular**
- **Química cuántica y modelos mixtos**

# Modelización por homología

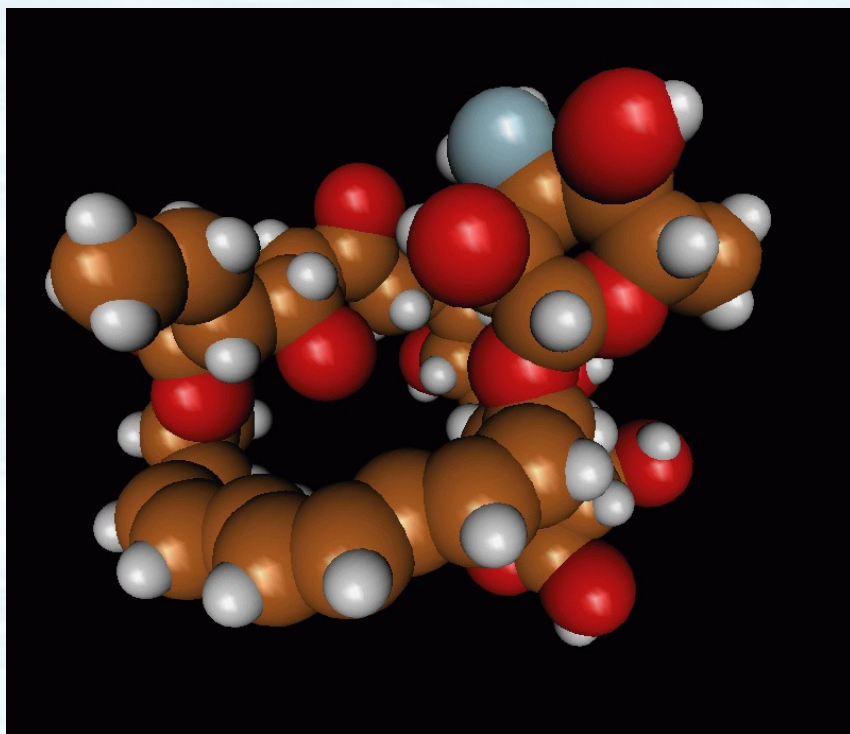


- Modelos de SHA3 superpuestos (2002)
- Generados en Web
  - PSSM
  - Modeller
  - SwissPDB
- **Refinados por MD**



# Modelado Cuántico

- 2001 modelo de antibiótico macrólido



- estructura inicial definida con JME
- Generado en Corina
- **Refinado con GAMESS-US y MOPAC**

# Desarrollo web

- Interno: la mayoría de servicios amplían sustancialmente su oferta de servicios vía Web:
  - Simplificar el uso de herramientas complejas
  - Captar usuarios externos
  - Visibilidad
- Se reduce la necesidad de soporte local



# El desarrollo del Web

**Moldy**

**Ye olde Modelling Wizard**

**EMBNet/CNB**

**Obtaining the initial model**

To build an initial model submit your sequence to one of the following modelling services:

**Homology modelling servers**

- SwissModel
- The CD4 models server
- SDSC Protein Structure Homology Modeling Server
- 3D-JIGSAW

**Threading servers**

- 3D-psp
- The Sausage Machine
- FUGUE
- LOOP
- Superfamily

**Compare the different servers**

- EVA

Once you receive your initial model by e-mail you may proceed to

© EMBnet/CNB José R. Valverde

**YaMI**

**Yet another Modeller Interface**

PIR Sequence File

Template PDB Codes

Number of models

Refinement: Very Fast MD

Include hetero atoms: Yes No

Include waters: Yes No

Include hydrogens: Yes No

Alignment: Yes No

**WebGramm (LR-HR)**

**Web interface for High or Low-Resolution docking with Gramm.**

Matching mode: Generic Helix

High or low resolution docking: Low High

Docking type resolution: All Hydrophobic

Receptor molecule in PDB format

Ligand molecule in PDB format

Run Gramm

Page loaded.

**Applications**

The following TINKER applications are supported under this WWW interface:

Application	Description	Help	Do it!
ANALYZE	Provides information about a specific molecular structure.	?	?
ANNEAL	Molecular dynamics simulated annealing computation.	?	?
DYNAMIC	Molecular dynamics (MD) or stochastic dynamics (SD) computation.	?	?
GDA	Straub's Gaussian Density Annealing algorithm	?	?
MINIMIZE	Nonlinear conjugate gradient minimization	?	?
NEWTON	Truncated Newton minimization	?	?
OPTIMIZE	Variable metric energy minimization	?	?
PCS	Potential smoothing and search algorithm for a "global" optimization	?	?



# Herramientas comunes

- Herramientas científicas de uso general
  - LIMS
- Groupware (2003-2004)
  - Wiki, blogs, gestores de contenido
  - Gestión de Proyectos
  - Gestión de Clientes
  - Moodle (e-Learning, hoy estándar)
- Profesionalización de los servicios


# Profesionalización

- Servicios públicos
  - Seguimiento de errores y listas de usuarios
  - Sourceforge, Savannah...
- Servicios personalizados
  - Gestión de proyectos
  - Gestión de clientes
  - Formación
  - etc...



# PredictProtein (2001)

### The PredictProtein server



[PP mirrors](#) : JavaScript links (for Netscape 3+): [Europe](#) - [Australia](#) - [Asia](#) - [America](#)

**It is** PredictProtein is a service for sequence analysis, and structure prediction.

**It does** You submit any protein sequence. PredictProtein retrieves similar sequences in the database and predicts aspects of protein structure ([brief introduction](#)).

**You can**

- [submit](#) a protein sequence for prediction ([default](#), [advanced](#), [expert](#))
- scan through the PredictProtein [help documents](#)
- get a brief [listing](#) of what PredictProtein does
- get more information about [where to go from here and options](#) for PredictProtein
- use post-processing [tools](#) (e.g. alignment display)

**META-PP can**

- [submit](#) requests to [other servers](#) (through ONE [META](#) interface!).

**Methods used** [PHDsec](#) - [PHDacc](#) - [PHDhm](#) - [GLOBE](#) - [BLAST](#) - [MaxHom](#) - [TOPTIS](#) - [COILS](#) - [CYSPPRED](#) - [ProSite](#) - [SEG](#) - [ProDom](#) - [EvalSec](#) -

**Slow line?** [Download](#) all files from the PP home directory.

**From here** [NEWS](#) [WHAT](#) [WAIT](#) [HINT](#)

**Contact** [predict\\_help@columbia.edu](#) Copyright: Bu

**Version** Last update of WWW pages: Mar2, 2000

Search PP pages for:

Note: search only for PP in New York

PP mirrors in Europe

1. Germany: [EMBL Heidelberg](#)
2. England: [EBI Hinxton \(UK\)](#)
3. Italy: [Univ 'Tor Vergata', Rome](#)
4. Netherlands: [CAOS Nijmegen](#)
5. Russia: [RAS Fushchino](#)
6. Spain: [CNB Madrid](#)
7. Switzerland: [GlaxoWellcome Geneva](#)

File Edit View Go Bookmarks Tools Tabs Help

Back Forward Home Bookmarks

http://sting.cnb.csic.es/SMS/index\_s.html

Embrapa Informática Agropecuária

Bioinformatics LABORATORY

Home About Us Contact Us Help

Supported By: Java Protein Dossier STING 1.1 Build Molec Description References Modules Rate

TRAJE GALLERY

Java Protein Dossier

See information on what you need to run properly Java Protein Dossier and BLUE STAR STING components

Launch JAVA PROTEIN DOSSIER with a public PDB file

Choose the size of the BLUE STAR STING window:

☒ 640x480 ☐ 800x640 ☐ 1024x768 ☐ 1280x1024

Enter 4 letter code of the PDB format file (Example: type in 1cho). Then STING IT!

PDB Code:  STING IT!

Launch PROTEIN DOSSIER with your local file in TGZ format

Browse to TGZ format file at any position on your local computer. Then STING IT!

Choose the size of the BLUE STAR STING window:

x1024

Mozilla: Choose Computational Engine

Archivo Modificar Ver Ir Comunicador Ayuda

Anterior Siguiente Recargar Inicio Buscar Mozilla Imprimir S

Marcadores Dirección: http://www.es.embnet.org/c Sitios parecidos

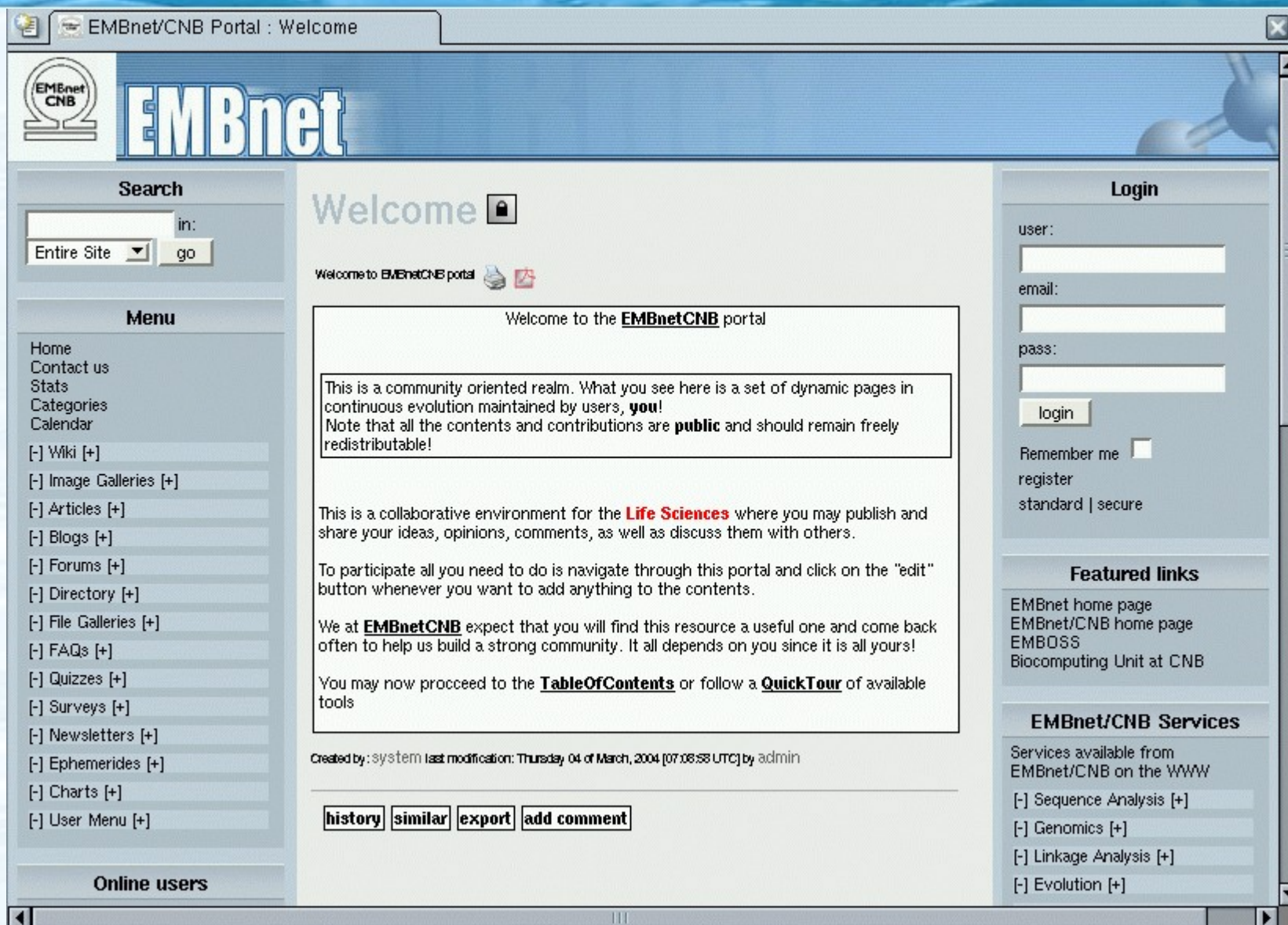
### Choose Computational Engine

Engine	Description
<input checked="" type="radio"/> Gamess	Ab initio and semi-empirical calculations
<input type="radio"/> Mopac	Semi-empirical calculations

Select Server

Job Manager Build Molecule Choose Engine Job Options Submit Job

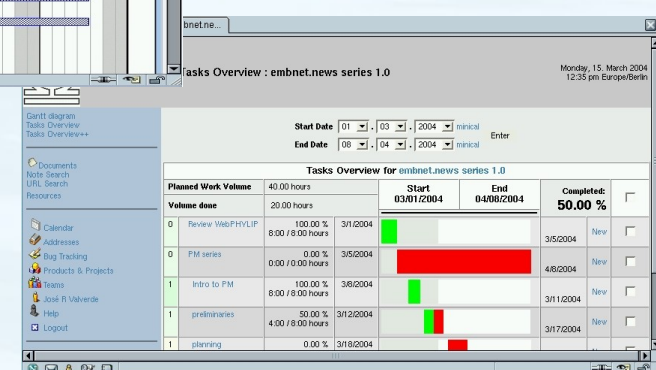
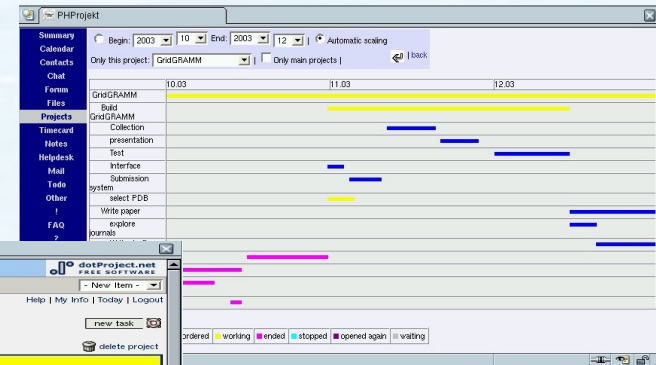
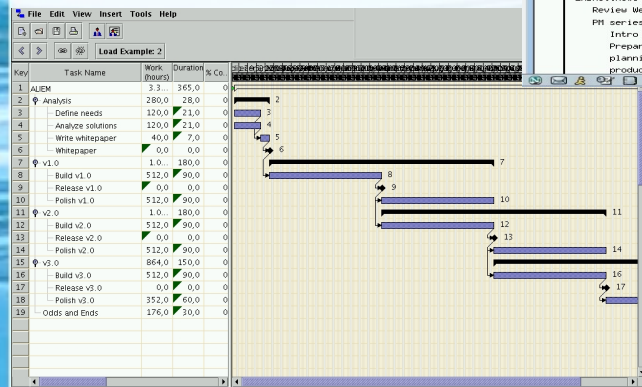
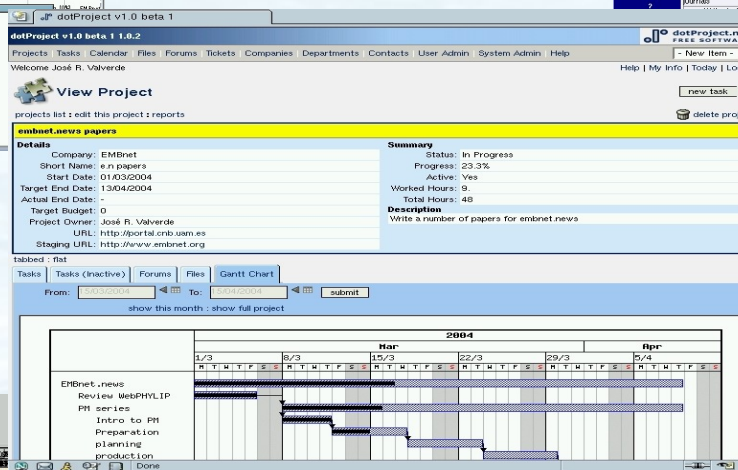
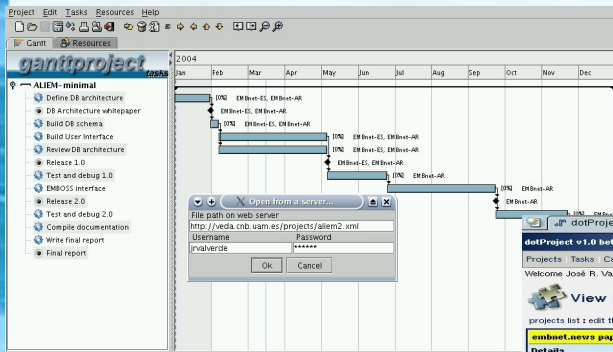




The Bioportal at CNB, Nov. 2003



# Project Management Tools (2004)



# Coordinación

- EMBnet
  - Desarrollo de portal e-learning común
  - Coordinación con ISCB, Organizaciones médicas, etc..
- Red Iberoamericana de Bioinformática
- La palabra mágica es **sinergia interdisciplinar**.
  - Biología - Medicina - Farmacología - Química - Física...



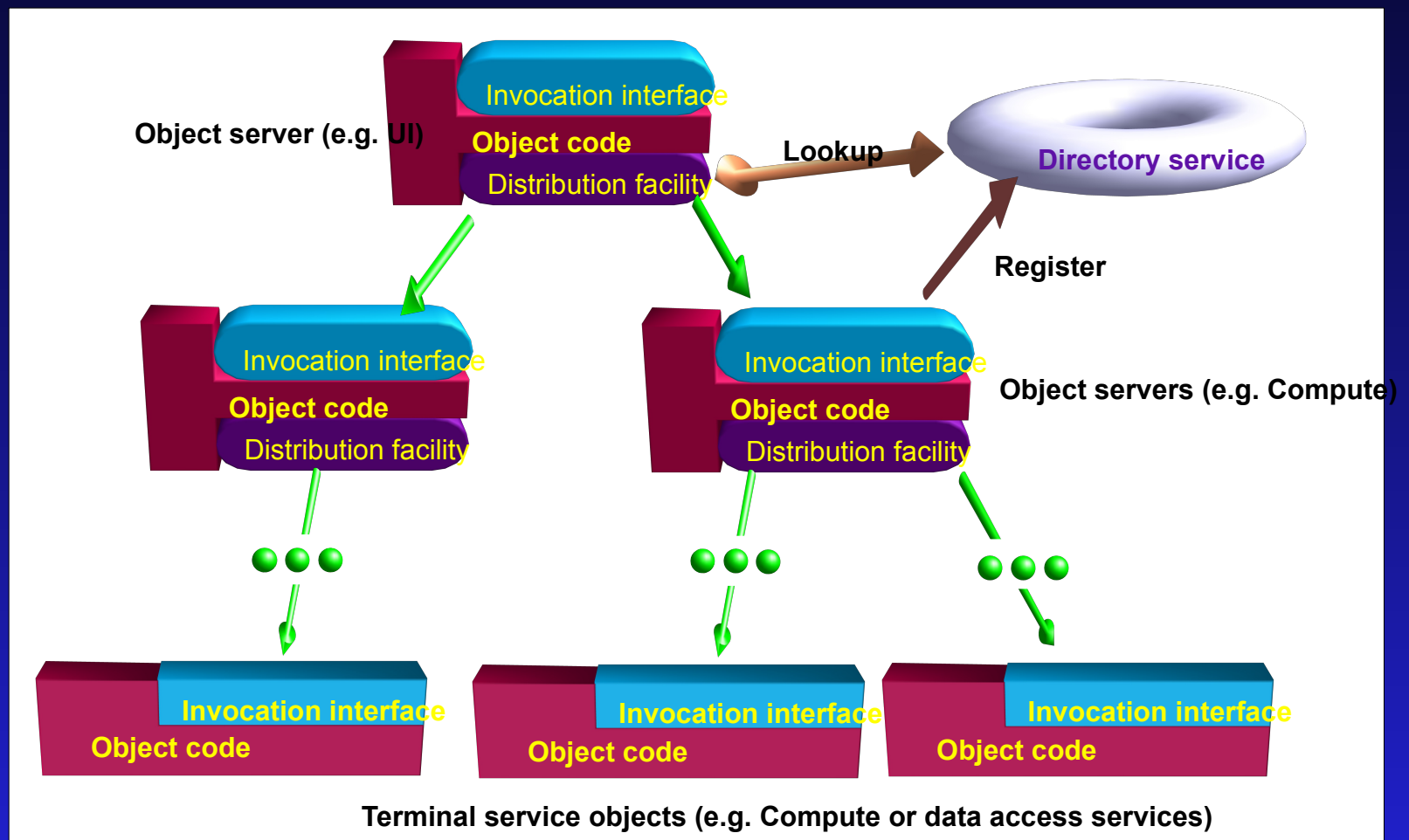
# Desarrollo Grid

- Necesidad detectada en 1998 y presentada a la ESF en 1999
- Arquitectura diseñada para 2001
- Lanzamiento de EGEE (CERN, 2004)
  - Análisis de secuencias
  - Interacciones Proteína-Ligando/Proteína
  - Genética de poblaciones
  - Filogenia...

# Arquitectura EMBgrid

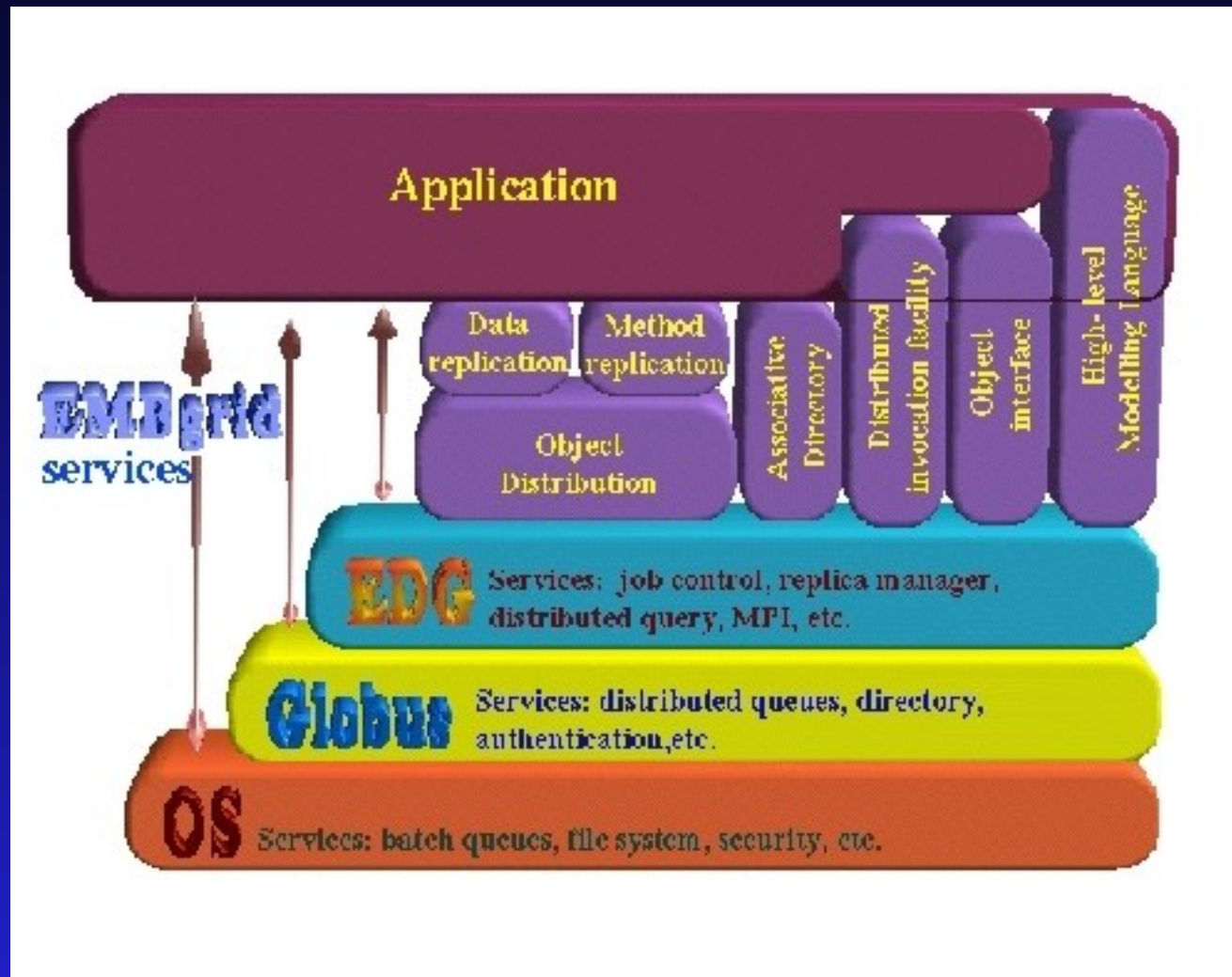
- Propuesta en 2001 por EMBnet tras dos años de deliberación interna
- Diseñado para copar con el **atabang!**  
**NGS**
- Precedía las tecnologías Grid y Cloud
- **Es mejor resolver los problemas con tiempo**

# An object distribution architecture



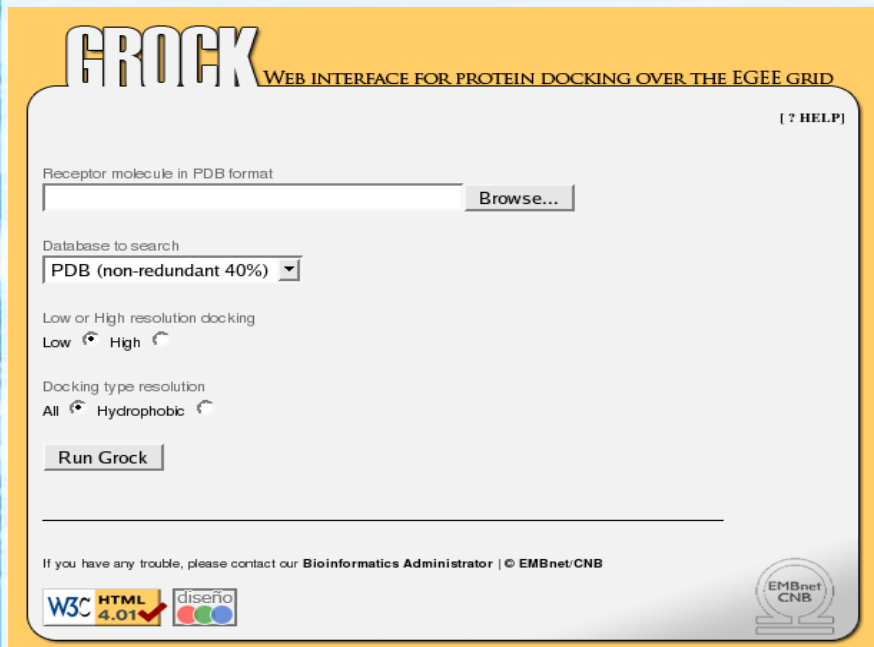


# A logical architecture



# GROCK (2005)

- Evolución de GridGRAMM
- Docking masivo en la Grid
- 10.000 docks de estructura completa en 2-3 días
  - interacciones
  - nuevos efectos



The screenshot shows the GROCK web interface. At the top, the logo 'GROCK' is displayed in a stylized font, followed by the text 'WEB INTERFACE FOR PROTEIN DOCKING OVER THE EGEE GRID'. A '[ ? HELP ]' link is in the top right corner. The main form area contains the following elements:

- A text input field labeled 'Receptor molecule in PDB format' with a 'Browse...' button to its right.
- A dropdown menu labeled 'Database to search' with 'PDB (non-redundant 40%)' selected.
- Radio buttons for 'Low or High resolution docking', with 'Low' selected.
- Radio buttons for 'Docking type resolution', with 'All' selected.
- A 'Run Grock' button.

At the bottom of the interface, there is a footer with the text 'If you have any trouble, please contact our Bioinformatics Administrator | © EMBnet/CNB'. To the left of the footer are logos for 'W3C', 'HTML 4.01', and 'diseño'. To the right is the 'EMBnet CNB' logo.

# Los servicios al final de los 00

- Servicios públicos
  - ámbito mundial, sencillos, utilidad general, requieren infraestructura sólida
- Servicios localizados
  - ámbito local, resolución de problemas especializados, infraestructura sólida
- Aparece el “bioinformático de cabecera”
  - asociado a un grupo, aplicado a un proyecto, equipamiento *ad hoc*



# Colapso general

- Uso en 2007: 398 computadores del CNB, 655.324 de todo el mundo
- Visitas en 2007: 7.515.008 computadores de todo el mundo
- Para 2007 la demanda privada de expertos en Grid/Cloud drenó nuestro servicio.
- La financiación escasea y los ingresos se vuelven más difíciles de justificar

# Tendencias

- Todos los campos nuevos necesitan un soporte especializado
- A medida que se popularizan, pasan a ofrecerse vía Web
- A medida que se ofrecen en Web decrece la necesidad de soporte local
- Al desaparecer el soporte local tienden a desaparecer los servicios Web

# Lecciones

- La previsión y el éxito puede pagarse muy caro si no se aprecian
- La (Bio)informática no es gratuita aunque lo parezca
  - FLOSS ayuda a reducir costes
  - Requiere desarrollo **y mantenimiento**
  - Abandonado, el software se pudre
  - El coste a pagar es una pérdida de competitividad y conocimiento



# Computación Científica en los 10

# Formación

- Persiste la necesidad de formación general
  - Aún escasa en la Universidad española :(
- Formación especializada
  - Usuario: Interpretación de resultados
  - Especialista: Manejo de herramientas
- Necesidad de
  - Formación general en la carrera
  - Harmonización de formación especializada

# Coordinación

- Aumenta la presión hacia los servicios “de cabecera”
  - La crisis conduce a visiones “miopes” y falta de coordinación
  - Falta de preparación para el futuro
- Nuevas iniciativas
  - Justificación práctica detallada
  - SEQAHEAD: competitividad
  - FreeBIT: ahorro



# Infraestructura

- Servicios autónomos
  - Dependen de ayudas financieras
  - Pueden abordar grandes problemas
- Servicios satélite
  - Dependientes de recursos compartidos
  - Pueden abordar problemas generales
- Servicios aislados
  - Problemas de andar por casa

# Servicios generales

- Problemas de propósito general en Web
- Permiten abordar grandes problemas
- Transparentes al usuario
  - Necesidad de justificar su existencia
  - Muy sensibles a cambios sociopolíticos, económicos y modas
- Necesarios para la actividad de fondo y para grandes problemas

# Servicios especializados

- Atención especializada
- Evolución natural hacia servicios web
  - Erosión de los servicios generales
- Coordinación con otros servicios especializados
  - Visión reducida / Sinergias
- Redes de Excelencia



# Servicios de cabecera

- Atención específica de proyecto
  - Erosión de los servicios especializados
- Periodo vital reducido
  - Escasa experiencia
- Coordinación limitada
  - Habitualmente con servicios especializados locales
- Línea borrosa servicio/investigador



Oportunidades

# NGS

- La secuenciación llevada a su máxima potencia
- Está revolucionando la práctica
- Aún no ha empezado a desarrollar todo su potencial



# NGS hoy

- Producción de grandes cantidades de secuencias:
  - genomas completos
  - metagenomas
- El problema actual
  - Cómo analizar un genoma
  - Cómo comparar un genoma con otro
  - Medir cambios cuantitativos/cualitativos

# NGS mañana

- ¿Qué haremos cuando...
  - cada uno de nosotros tenga su genoma secuenciado?
  - queramos personalizar la medicina?
  - estudiemos poblaciones enteras?
- Sobre todo sabiendo todo lo que ignoramos
  - Genes y sus productos
  - Posibles mutaciones y combinaciones

# Propuesta EMBnet (2009)

- Construir un sistema que aúne todos los genomas de todos los ciudadanos
  - Integración con historias clínicas
  - Anotación sobre la marcha por especialistas (científicos, médicos)
  - Evaluación y toma de decisiones sobre la marcha
  - Sistemas de alerta al ciudadano
  - Sistemas de protección de la información



# Análisis de macromoléculas

- Entornos de visualización avanzados
  - Chimera, PyMol, SPBV, VMD...
- Análisis avanzado
- Predicción funcional a partir de la secuencia, evolución y estructura 3D
- Análisis masivo a partir de datos de alto rendimiento

# Predicción de estructura

- Sistemas avanzados
  - Modelado *ab initio*
  - Modelado cuántico (miles de átomos) con MOPAC, SIESTA, GAMESS, NW-CHEM, MPQC, ABINIT, OCTOPUS...
- Requiere un conocimiento detallado del problema
  - Actualmente exige verificación minuciosa
    - experimento, teoría



# Ab initio prediction



Comparison of *ab initio*  
predicted and known  
structures (human and  
mouse)

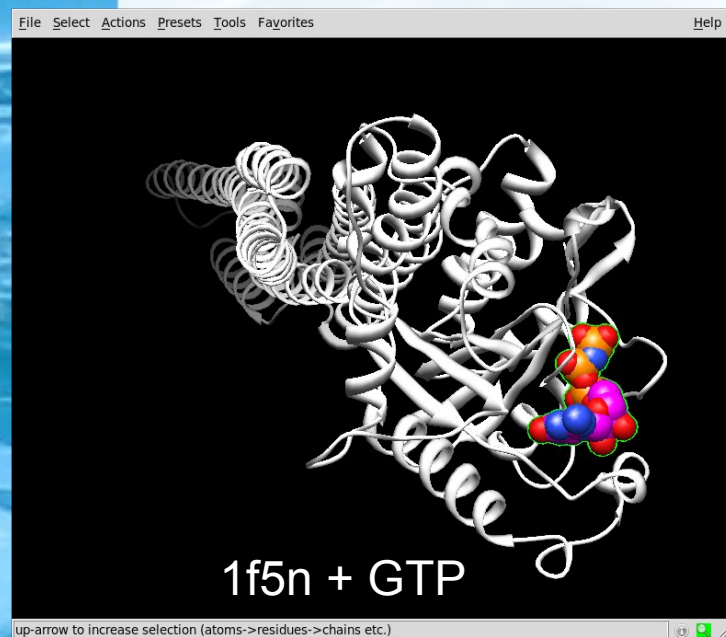




# Estructura de complejos

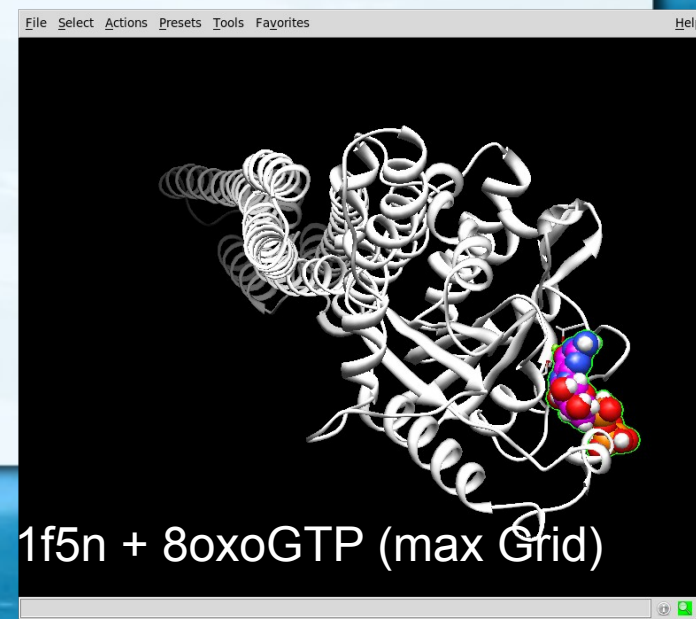
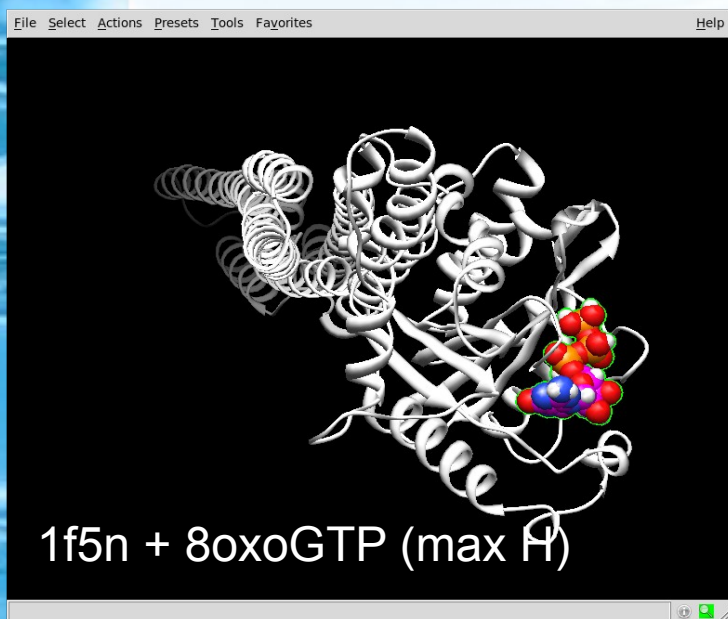
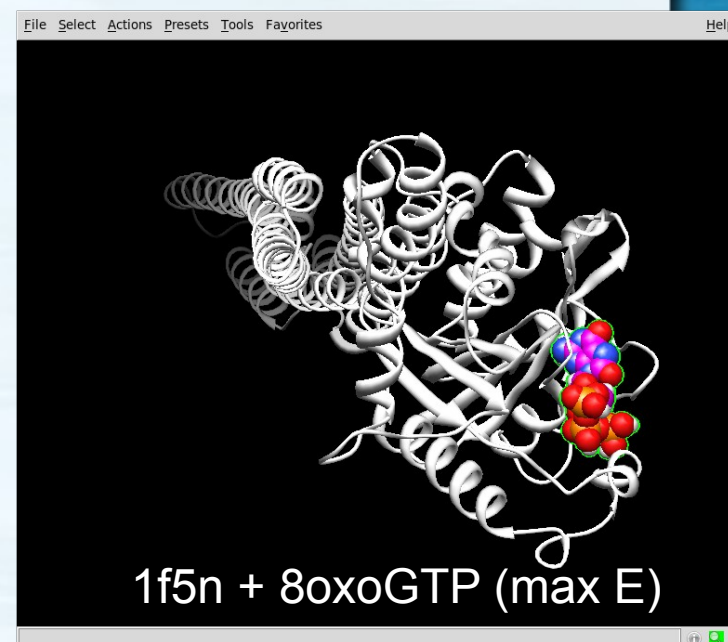
- Complejos de interacción (Farma, función)
  - Proteína-Proteína
  - Proteína-Substrato
  - Medicina personalizada
- El conocimiento del problema es fundamental también
  - Oportunidad para servicios especializados

# Docking 8oxoGTP



Over 200 docking experiments, 100 models each. Which are the correct ones?

Autodock  
Dock6  
3D-Dock  
Gramm  
Etc..





# Simulación Molecular

- Dinámica molecular (de fs a  $\mu$ s)
  - in vacuo
  - en solución
  - en membranas
  - multi-molecular
- Permite entender el comportamiento molecular
- Requiere un conocimiento detallado del problema



# Biología Cuántica

- Quantum Mechanics/Dynamics and TD-DFT
- Modelización de reacciones enzimáticas
  - Medicina personalizada
  - Predicción (dis)funcional
- Requiere conocimientos avanzados y análisis en profundidad

El futuro

# Está por escribir

- FLOSS reduce costes y maximiza compartir el conocimiento
  - Lo que es bueno para tí lo es para mí
  - ~~Lo que es bueno para tí es malo para mí~~
- Requiere desarrollo y mantenimiento
  - Los costes -como el software- pueden compartirse
  - Pero hay que asumirlos en cualquier caso
  - La importancia de las RRPP



# Pero no es tan malo...

- Hay espacio de sobra para los servicios, formación y soporte especializados
  - Sólido modelo de negocio
  - No tiene pinta de ir a desaparecer...
  - ..siempre que lleguen nuevos servicios
    - Compartir conocimiento acelera el desarrollo
    - Frenar el desarrollo facilita el aprendizaje
    - Cuando un servicio es común, pasa al web.

# Requerimientos

- Culturales
  - cultura de colaboración competitiva
  - adaptación al cambio
- Apoyo institucional a servicios generales
  - si nadie los mantiene desaparecerán
- Inversión en grandes problemas
  - puede dejarse para otros perdiendo competitividad

# Gracias

- A todos por venir
- A CYTED
- A EARTH
- 
- **Por hacer de éste un evento especial.**