

The mission of the EBI

Jean-Jack Riethoven
EMBL-EBI

EMBL Outstation
European Bioinformatics Institute



EBI has its roots in the EMBL Data Library

- **1980 EMBL established the "Nucleotide Sequence Data Library"**
- **The original goal was to build a database of DNA sequences rather than putting them on the pages of journals**

EMBL Outstation
European Bioinformatics Institute



The EMBL Data Library

- **World's first nucleotide sequence database**
- **Support of journals in data submission**
- **Network access 1987**
- **Nucleotide and Protein Sequences 1987**
- **Value added specialist databases — e.g., EPD 1988**
- **EMBnet 1988**
- **CD-ROM 1989**
- **Pointers between databases 1990**
- **Genome projects — C. elegans, Yeast**
- **Patent Data 1992**
- **CD-ROM indices & software 1992**
- **Parallel architectures — Blitz 1992**

EMBL Outstation
European Bioinformatics Institute



European Bioinformatics Institute

- **Long term viability**
- **Scale of task**
- **Fragile grant support**
- **Some organisational requirements for service activities different from research**
- **The need for R&D**
- **Collaborate as peers with global partners**
- **Better support for industry**
- **Genomics context**

EMBL Outstation
European Bioinformatics Institute



Establishing the EBI

- **EMBL Council voted in December 1992 to establish the EBI as an EMBL Outstation**
- **March 1993 the decision accept the offer to locate it in Hinxton**
- **August 1993 first person moves to Hinxton**
- **May 1994 MRC made temporary facilities available to relocate Data Library**
- **September 1994 we started to offer our services from Hinxton**
- **September 1995 — moved into permanent facilities**

EMBL Outstation
European Bioinformatics Institute



The Arguments for Relocation

- **The vision:**
Wellcome Trust Genome Campus
- **The Cambridge academic context**
- **The facilities offered**
- **Resources to support the transition**

EMBL Outstation
European Bioinformatics Institute



Molecular biology becomes information intensive

- For more than two decades 3D protein structures have been collected in an electronic database
- Protein and DNA sequences collected in central repositories
- Genome projects and automated sequencing cause a huge surge in the scale of the task

EMBL Outstation
European Bioinformatics Institute



Macromolecular Information

- Nucleotide Sequence
- Protein Sequence
- Protein Structure
- Protein Function

EMBL Outstation
European Bioinformatics Institute



DNA

```
ccagaagaatccaga  
aggggcttggaaag  
cttttgctataag  
cgggtggcaagtg  
cctcaaaacgtagta  
cgcctcgcctcct
```

EMBL Outstation
European Bioinformatics Institute



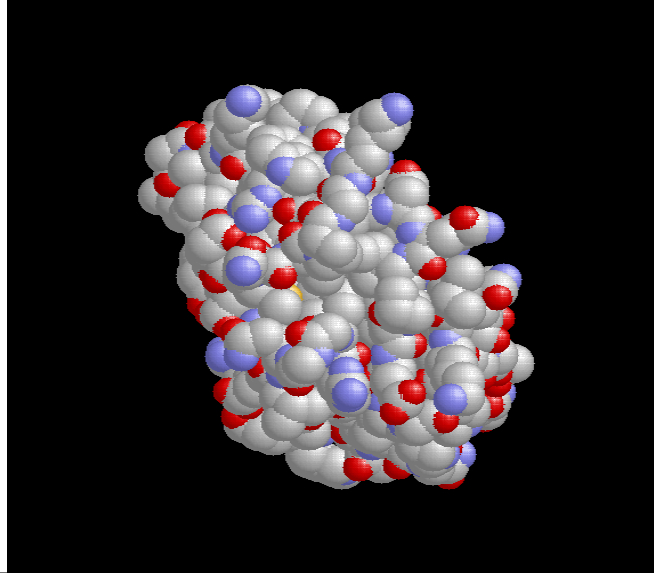
Protein

```
MFREDLAFLOGI  
EFSSEQTRANSE  
RELQVWGENNS  
EAGADRQGTVSE  
PQITLWQRPLVT  
LGGQLKEALLDT  
DDVYLFEMNLRD
```

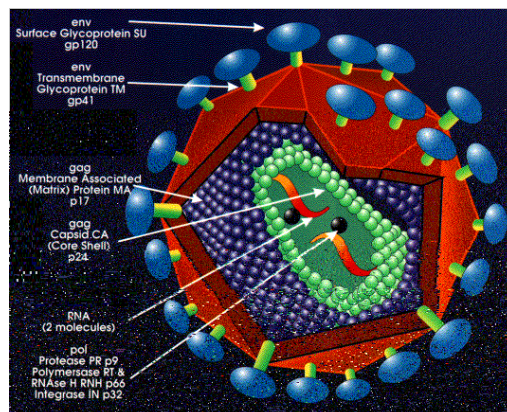
EMBL Outstation
European Bioinformatics Institute



Protein Structure



Protein Function



EMBL Outstation
European Bioinformatics Institute



Nucleotide Sequence Database

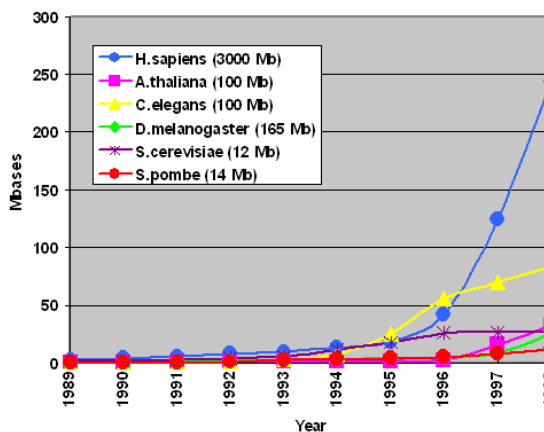
- Over 900 000 000 base pairs
- Collaboration with GenBank® and DDBJ
- Data from genome projects
- Partial sequencing — ESTs, STSs
- Patent data
- A new sequence every minute...

EMBL Outstation
European Bioinformatics Institute



Species Sequenced

Progress of Major Genome Sequencing Projects



EMBL Outstation
European Bioinformatics Institute



Predicted completion dates

- **Predicted Completion dates**
- **Homo sapiens: 2006**
- **Arabidopsis thaliana: 2002**
- **Caenorhabditis elegans: 1998 (Done)**
- **Drosophila melanogaster: 2005**

EMBL Outstation
European Bioinformatics Institute



EBI and the Web

- **Move towards web-based services**
- **'Older' services of EBI (by snailmail or e-mail) are phased out**
- **In this talk:**
 - **Quick intro on databases hosted by EBI**
 - **Services interacting with those database (e.g. new interactive FASTA & BLAST, SRS, S&W)**
 - **Other services (e.g. BiowURLd, BioInformer Events)**

EMBL Outstation
European Bioinformatics Institute



Databases - an overview

- EMBL Nucleotide Sequence Database
- SWISS-PROT Protein Sequence Data Bank
- IMGT Immunogenetics Database
- Radiation Hybrid Database
- BioCatalogue of Software
- MSD/3DB/PDB
- Around seventy additional specialised molecular biology databases

EMBL Outstation
European Bioinformatics Institute



EMBL Nucl. Sequence Database

- is a curated nucleotide sequence database established in 1982 by the EMBL Data Library
- doubles in size every 12 months
- contains currently (release 53) more than 1.9 million nucleotide sequence entries comprising over 1,281 million nucleotides
- data collection is done in collaboration with GenBank and DDBJ
- http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html

EMBL Outstation
European Bioinformatics Institute



SWISS-PROT - an overview

- is a curated protein sequence database established in 1986 by Amos Bairoch in Geneva and maintained collaboratively with EMBL since 1987
- contains currently (release 35) nearly 70 000 protein sequence entries (25 million amino acids) with:
 - a high level of annotations
 - a minimal level of redundancy
 - a high level of integration with other databases
- http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html

EMBL Outstation
European Bioinformatics Institute



Databases at EBI - an overview

- **IMGT:** <http://www.ebi.ac.uk/imgt/>
integrated specialised database containing nucleotide sequence information of genes important in the function of the immune system (>24000 entries)
- **RHdb:** <http://www.ebi.ac.uk/RHdb/>
is a database of raw data used in constructing radiation hybrid maps. This includes STS data, scores, experimental conditions, and extensive cross references (>65000 entries)
- **Biocat:** <http://www.ebi.ac.uk/Biocat/>
database containing entries on software related to bioinformatics, molecular biology and genetics (release 5.6)
- **More at:** <http://www.ebi.ac.uk/dbases/topdata.html>

EMBL Outstation
European Bioinformatics Institute

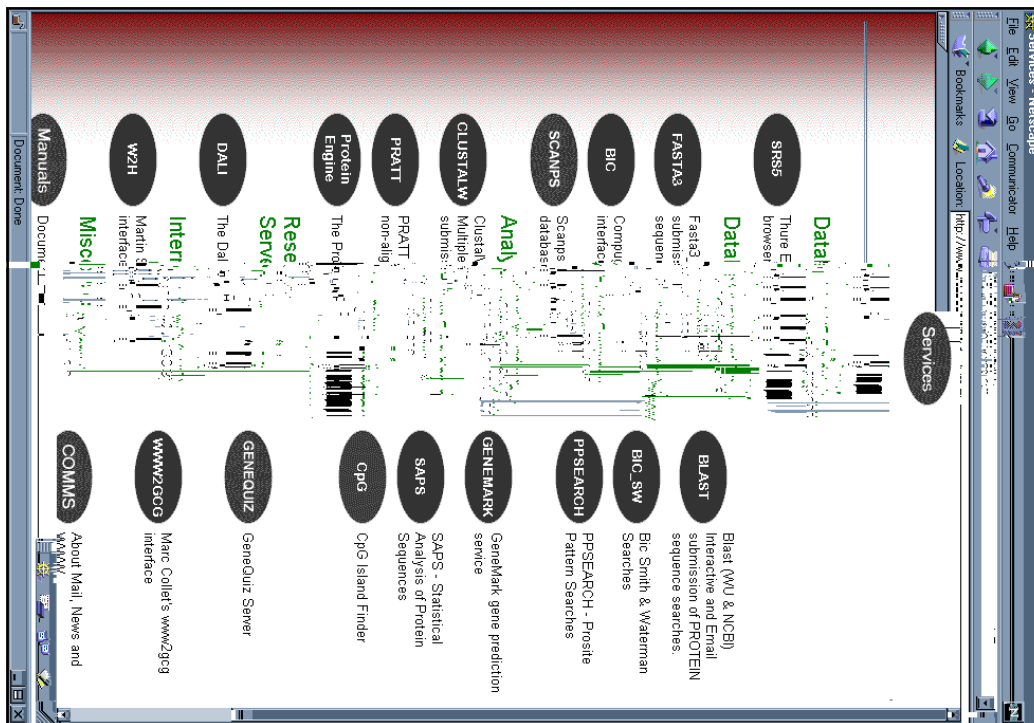


Webin - submission tool

- **New web-based tool to submit nucleotide sequences + related biological information to EMBL/Genbank/DDBJ**
- **Preferred tool for submission**
- **<http://www.ebi.ac.uk/submission/webin.html>**

EMBL Outstation
European Bioinformatics Institute





Three new interactive search services

- **Interactive FASTA3, BLAST, S&W**
- **Typical wait one to two minutes**
- **Running on 3 SGI Challenge servers, 2 DEC 8400's. S&W on Compugen's Bioccelerator-2**
- **<http://www2.ebi.ac.uk/fasta3/>**
- **<http://www2.ebi.ac.uk/blast2/>**
- **http://www2.ebi.ac.uk/bic_sw/**

EMBL Outstation
European Bioinformatics Institute



SRS - details

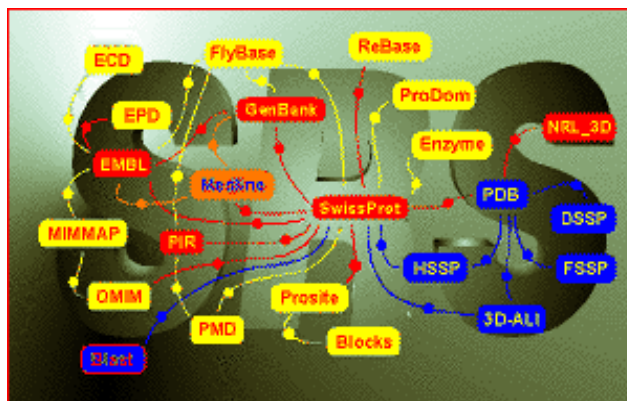
- Handles flat-file biological databases
- Indexing of fields for rapid searches
- Web-based front-end providing query interface

EMBL Outstation
European Bioinformatics Institute



SRS - details

- summary table of query hits
- HTML-enhanced presentation of flat file entries
- processing of cross-references to related databases



EBI Web Applications

- **ClustalW (multiple sequence alignment)**
<http://www2.ebi.ac.uk/clustalw/>
- **GeneMark gene prediction server (interactive in future)**
<http://www2.ebi.ac.uk/genemark/>
- **ProteinMachine / ProteinColourer**
<http://www2.ebi.ac.uk/translate/>
- **SAPS Statistical Analysis of Protein Sequences**
<http://www2.ebi.ac.uk/SAPS/>

EMBL Outstation
European Bioinformatics Institute

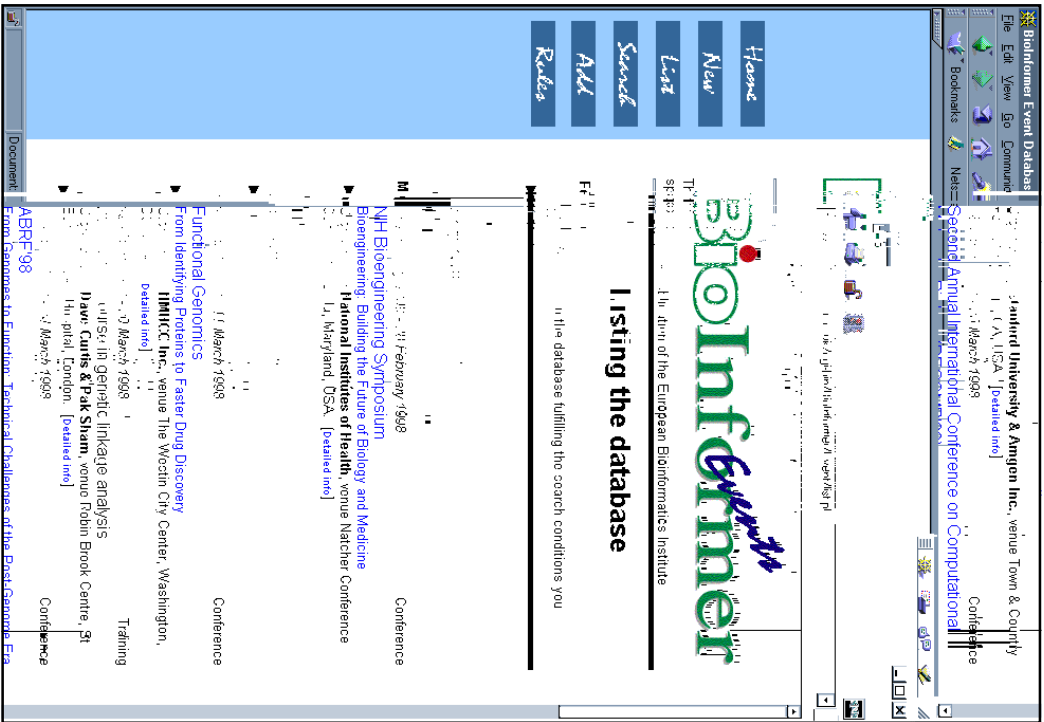


Other Services

- **BiowURLd: user-maintained searchable database of bioinformatics, biochemistry & molecular biology oriented sites on the web (near 1000 entries)**
 - <http://www.ebi.ac.uk/htbin/bwurld.pl>
- **BioInformer Event Database: user-maintained searchable database of bioinformatics related events (conferences, workshops, trainings)**
 - <http://bioinformer.ebi.ac.uk/Events/>

EMBL Outstation
European Bioinformatics Institute





Questions and comments



EMBL Outstation
European Bioinformatics Institute

