MASTER IN TECNOLOGIE BIOINFORMATICHE APPLICATE ALLA MEDICINA PERSONALIZZATA

Protein Sequence Analysis

Luis Sanchez (Protein Design Group, CNB-CSIC, Madrid)
Ana Rojas (Structural Bioinformatics Group, CNIO, Madrid)
Florencio Pazos (Protein Design Group, CNB-CSIC, Madrid)

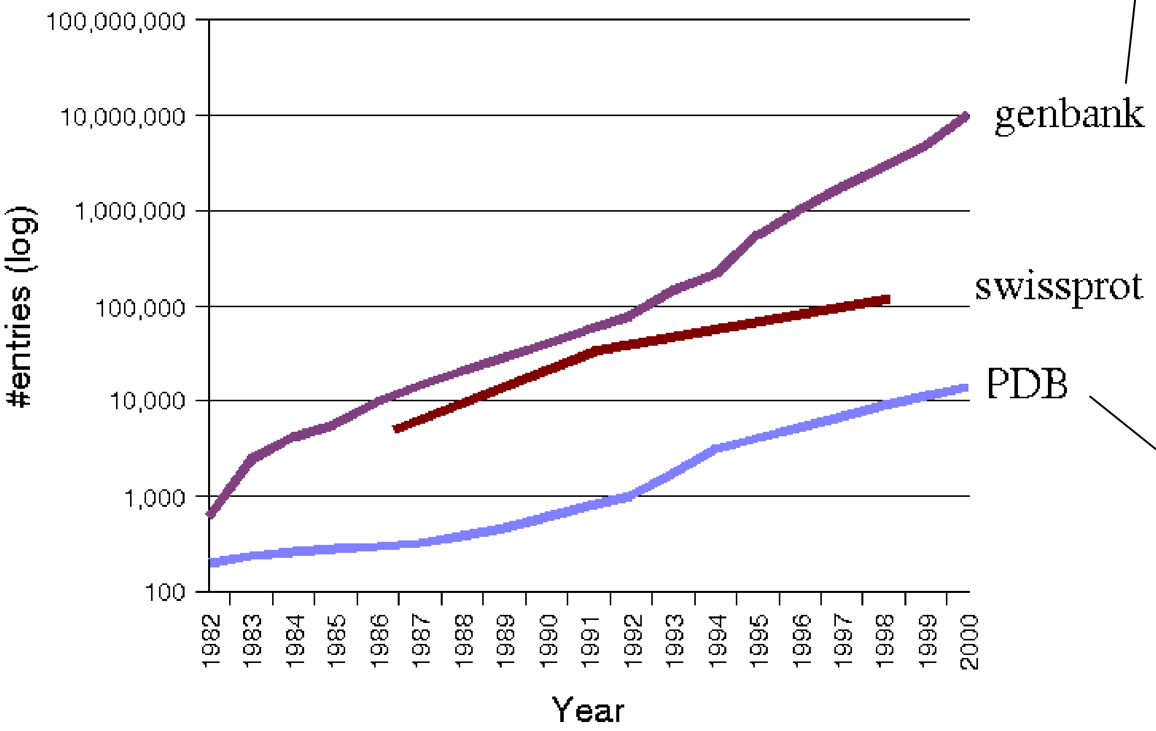
Protein Sequence Analysis

Introduction

Florencio Pazos (CNB-CSIC)

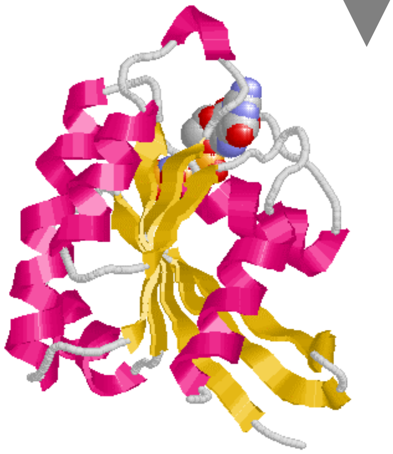
Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es

Level of Knowledge on Protein Sequences, Structures and Functions

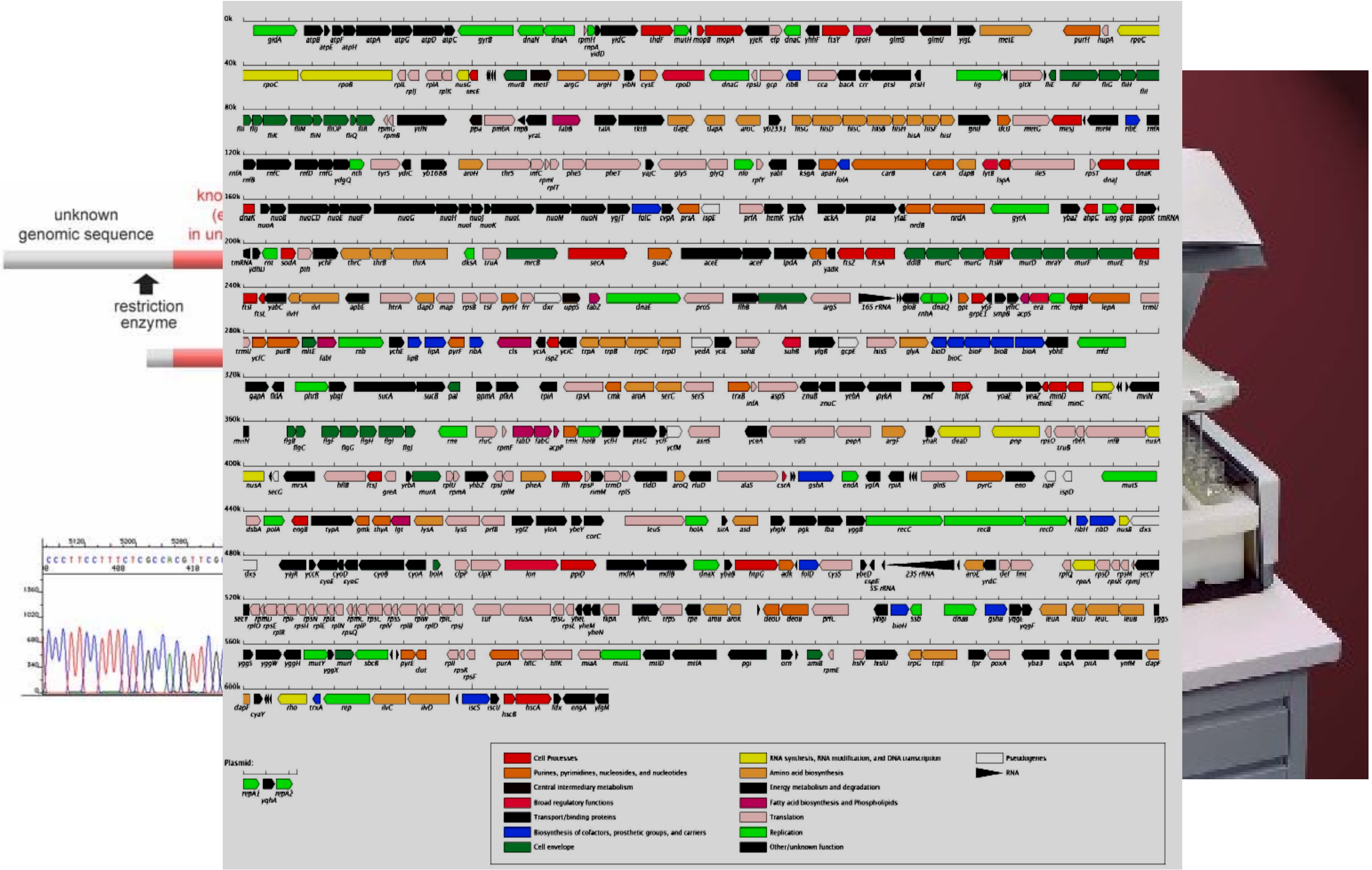


MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

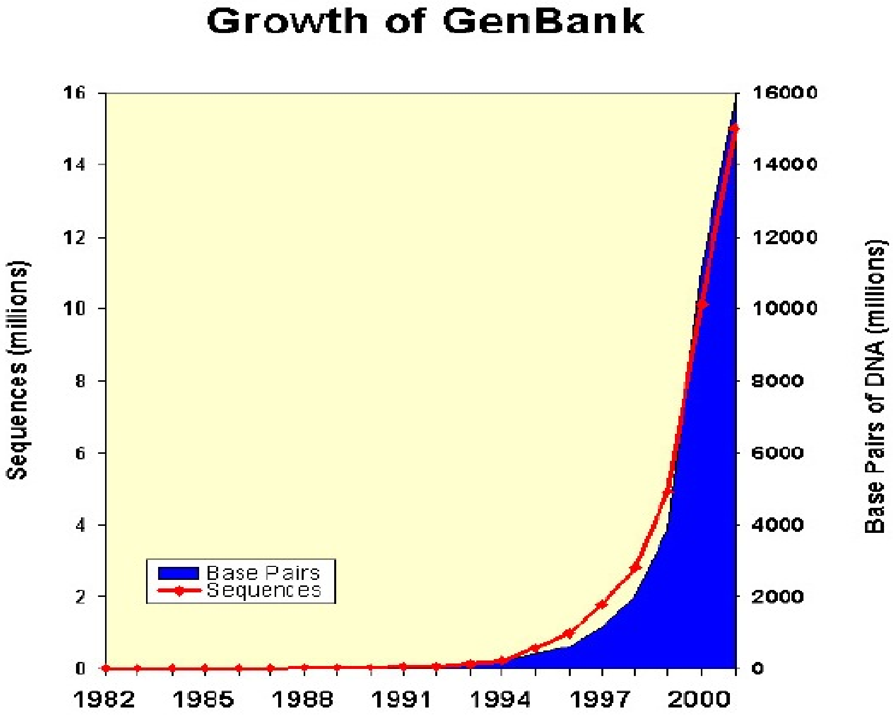
MLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMIL
VGNKCDLEDERV

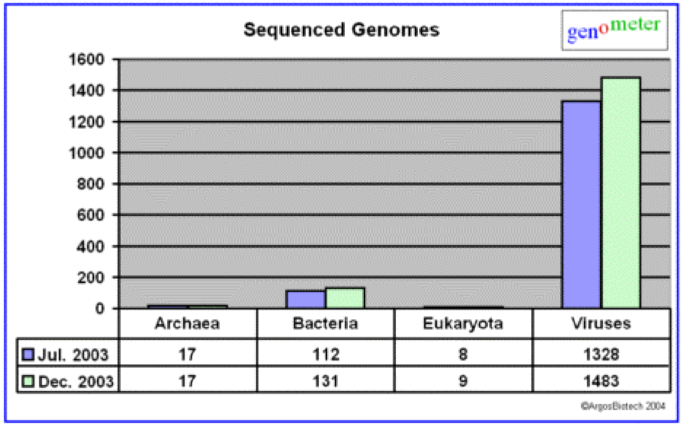


Determining protein sequences
DNA sequencing

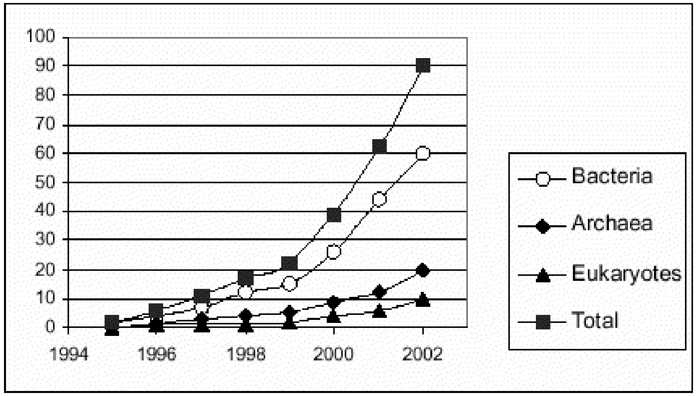

unknown genomic sequence
kno
(e
in un
restriction enzyme
0k
40k
80k
120k
160k
200k
240k
280k
320k
360k
400k
440k
480k
520k
560k
600k
Plasmid:
Cell Processes
Purines, pyrimidines, nucleosides, and nucleotides
Central intermediary metabolism
Broad regulatory functions
Transport/binding proteins
Biosynthesis of cofactors, prosthetic groups, and carriers
Cell envelope
RNA synthesis, RNA modification, and DNA transcription
Amino acid biosynthesis
Energy metabolism and degradation
Fatty acid biosynthesis and Phospholipids
Translation
Replication
Other/unknown function
Pseudogenes
RNA

Determining protein sequences
Genome sequencing



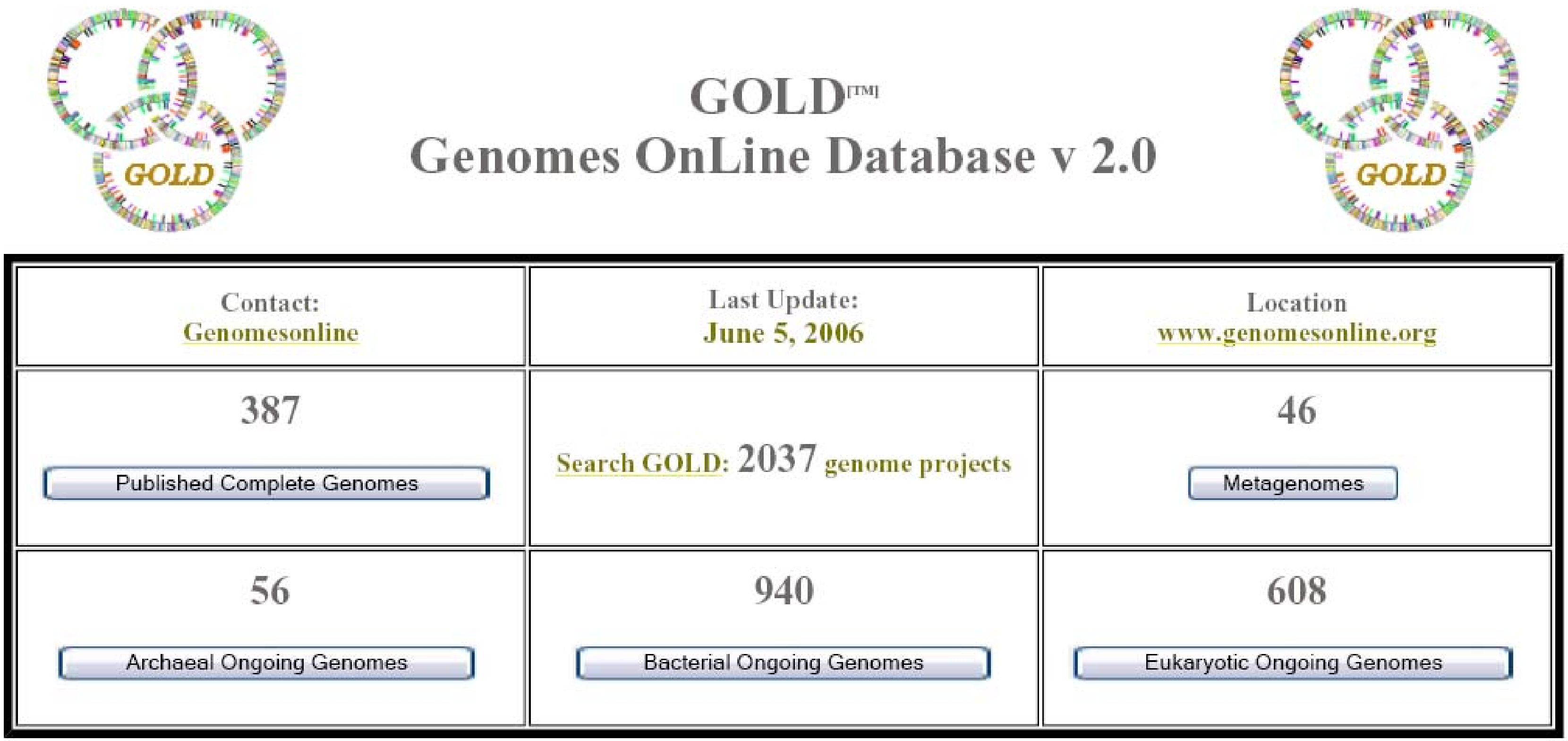


	Archaea	Bacteria	Eukaryota	Viruses
Jul. 2003	17	112	8	1328
Dec. 2003	17	131	9	1483



• Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. (2003) A vision for the future of genomic research. *Nature*, **422**, 835-847.

Finished genomes & sequencing projects

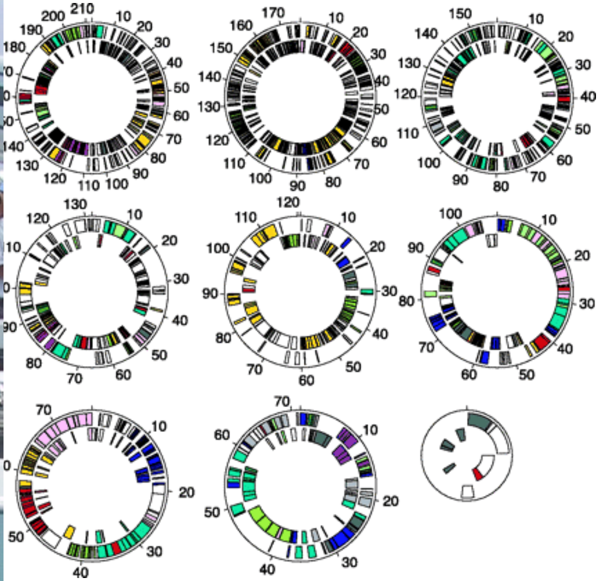


GOLD[TM]
Genomes OnLine Database v 2.0

Contact: Genomesonline	Last Update: June 5, 2006	Location www.genomesonline.org
387 Published Complete Genomes	Search GOLD: 2037 genome projects	46 Metagenomes
56 Archaeal Ongoing Genomes	940 Bacterial Ongoing Genomes	608 Eukaryotic Ongoing Genomes

http://genomesonline.org/

Determining protein sequences
“Environmental sequencing”
Metagenomes



- Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. (2003) A vision for the future of genomic research. *Nature*, **422**, 835-847.
- Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. (2004). *Science* **304,** 66-74.

```
TCCAAACCCAGGCTCTCTCCCAAACCAGTTTGCGGCAGATGGCCAGTGGAACCTCACTCTCCTCATCAGTAAAAAGGGGGCAGAGTGAGGGTCCTGAGAGCTAGTACAGGGACTGTGTGAAGTAGACAATG
CCCAGTGTTTAGCGTAAGAATCAGGGTCCAGCTGGTGCTCCCTAAACAGCAGCTGCTGTTCACTGTTGAAAGGCGCTCTGGAAGGCCAGGCGCGGTGGCTCATGCTTGTAATCCCAGCACTGTGGGAGGCC
GAGGTGGGCGGATCACCTGAGGTAGGGAGTTCGAGACCAGCCTGACCAACGTGGAGAAACCCCATCTCTCCTAAAAATACAAAATTAGCCAGGCGTGGTAGCACATACCTGTAATCCCAGCGACTCGGGAG
GCTGAGGCAAGAGAATTGCTTGAAACCAGCAGGGGAGGTTGTGGTGAGCCAAGATCGAGCCATTGCACTCCAGCCAGGGCAACAAGAGGCAAAATGGCGAAACTCCATCTCCGAGAAAAAAAAAAAAAAAG
AATACTTTCTGAAAGTATTTATTCATACAAATAAAGACTTGACCCATAAGGTAGGAACGCAAATGGGCCACGGAATCACTCATTCCACAGTATACACCGAGTGCCCTTGAAGTGCTGGGCACTGCTCCAGG
ATTGGGGGCATATTGGTGAAAAGAGAAGCAAGCCTGCCTGCTCAGATGGCAGGGAATGGGGAAAAACAGGGAGACAGTTTCCTGTTTGAGATGTTGGGAGTCTGCTTCGAGTAGTATATTTACTGGAAATA
GACCACTAACTTGGATGTCCCTTTTTGGAAATGTGCCTGCGTCCAGGGCTGGGTTGGGGCCCCAATGAACTTTGGCTCTGACATAGCTGTTGCCACACTCAGTGGAACTGAATCCATGTTTGCCTTCACCC
GGCATCCTTCACCCCAACTCTCCCCGCCACAACATACATCCCATGCCAGCCTGGGGACCCTCAAAGGTGCTTCATCATTAGGTTTGTGGCTGGGTCCTACTGAAGTAAGTCTTGGCACTCAGAGGGATAGG
AATTGAATGAAGACATGAGATTCCTCTGCGGGAGGCCTCTCTAGGAAATCTGTGGACTCACACGTTTACTAATGTTGCTGCAGCCCCGCACCCACCTTGGCCTTGGGCAGCCATACTCTAGGGCTTTTGTA
ACCTCTCCATGTGAGGAACTCAAATTAGACCTGGGTTTGGAGGCGGTGCTCCGAGCTGGCCTTTGGGGGAGGTTTTGTGCGAGGCATTTCCCAAGTGCTGGCAGGATTGTGTCACAGACACAGAGTAAACT
TTTGCTGGGCTCCAAGTGACCGCCCATAGTTTATTATAAAGGTGACTGCACCCTGCAGCCACCAGCACTGCCTGGCTCCACGTGCCTCCTGGTCTCAGTATGGCGCTGTCCTGGGTTCTTACAGTCCTGAG
CCTCCTACCTCTGCTGGAAGCCCAGATCCCATTGTGTGCCAACCTAGTACCGGTGCCCATCACCAACGCCACCCTGGACCGGGTGAGTGCCTGGGCTAGCCCTGTCCTGAGCACATGGGCAGCTGCCTCCC
TTCTCTGGGCTTCCCTTTACCTGCTGGCTGTGGTCGCACCCCCACTCCCAGCTCTGCCTTTTTCTCTTCTGGGTCCCCAGGGTGAAATTCTCACCAGCCCAGGGGACTCTGGAGGCACCCCCTGCCTCCAA
ACACAGAAGCCTCACTGCAGAGTCCTTCAGGCAGGATTGTGTCACAGACACAGAGTAAACTTTTGCTGGGCTCCAAGTGACCGCCCATAGTTTATTATAAAGGTGACTGCACCCTGCAGCCACCAGCACTG
CCTGGCTCCACGTGCCTCCTGGTCTCAGTATGGCGCTGTCCTGGGTTCTTACAGTCCTGAGCCTCCTACCTCTGCTGGAAGCCCAGATCCCATTGTGTGCCAACCTAGTACCGGTGCCCATCACCAACGCC
ACCCTGGACCGGGTGAGTGCCTGGGCTAGCCCTGTCCTGAGCACATGGGCAGCTGCCTCCCTTCTCTGGGCTTCCCTTTACCTGCTGGCTGTGGTCGCACCCCCACTCCCAGCTCTGCCTTTTTCTCTTCT
GGGTCCCCAGGGTGAAATTCTCACCAGCCCAGGGGACTCTGGAGGCACCCCCTGCCTCCAAACACAGAAGCCTCACTGCAGAGTCCTTCCGGAGGACGGTTCTGTGCTGGGCCTGGAGGGGCTGCCTGGGG
GGCAATGACTGATCCTCAGGGTGAGCTCCTGCATGCGCACTGCCCACCAGGGGCCTCATCTCCCCATCTGCAAAATCAGGGAGAGATCTGCCTGAGTCTCCTCCCAGCTGACAGTCAAAGATTCAGCATCA
AGCCCCCATCACCAGCTCCCCCCTTCTCCCCAGATCACTGGCAAGTGGTTTTATATCGCATCGGCCTTTCGAAACGAGGAGTACAATAAGTCGGTTCAGGAGATCCAAGCAACCTTCTTTTACTTTACCCC
CAACAAGACAGAGGACACGATCTTTCTCAGAGAGTACCAGACCCGGTGAGAGCCCCCATTCCAATGCACCCCCGATCTCAGCTGTCTGGCCAGAAGACCTGAGCAAGTCCCTCCTTCTTCCTGGCCTTGGC
CTTCCCATGGGTGGAACCGGGAGGGTTGGCTTTAATCTCCACCAGAACTCTTGCCCCGGGACTGTGATGGGCGATTGGCCACTTCTCCTCGATAACATTACTGTTTTTCTTCCGCCTTCTGGTTGACTTTA
GCCAGAACCAGTGCTTCTATAACTCCAGTTACCTGAATGTCCAGCGGGAGAATGGGACCGTCTCCAGATACGGTGAGGGCCAGCCCTCAGGCAGGAGGGTTCACCGTGGGAACAGGGCAGGCCAGCATAAG
GTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTGGGCACAGAGAAATAACCACTAACATTTTTGAGCTCTTACCACGTGCTCAGAAAAAATCCCTAAGAAGACACTGAGAGAATTAGATGAGGAAACATAAG
AACAGAGACCTCAAATAGTTTCCCCAAGGTCACACAGCTTATAATTAGAACTAGAATTGGAACTCCAGGCTGGCTTCAGATCTGCCTCTCTCTCACGCCCTCTTTAAGATCCTTTGCAAACCAATGGTAGA
AGCCTGTATGTTGGAGAGGTGGTACCTTCAACTATGTCCCCCATCACCGCAGAGGTGGCACATGGCAGGGATCTGATGGAGCTGAACTGACATCATTTAGCATCCCGAGCCTCCTCTCTGGGCCTCATTTT
CCTCCTCTGTAAAACGGGGAGAAAGGCCCTGACAGCCACAGTCTGTGTGAGGCTCCTGAGATCTCATGTACAGAAAGTGCTTGGCGTGGAGCTGGGCACGCAGCAGGGGCTGGGCACACGGTGGCCCAAAG
GAGACCCGGGCCTTCACTGATGGGCTTTGTGGCCCCGGACACATTTCTCTTCCAGAGGGAGGCCGAGAACATGTTGCTCACCTGCGTTCCTTAGGGACACCCCTAGGACTCCTCACCTGTAAGACAGGCAC
CATTGTGCCATCCCATGTTCTCACCCAGAGGCTCTTAAGACCTTGATGTTTGGTTCCTACCTGGACGATGAGAAGAACTGGGGGCTGTCTTTCTATGGTAGGCATGCTTAGCAGCCCCAAACTCATGCCCC
TCTCAGGCCTCACCCCCCATTCACCCACCCCTGGGCTGGCCCCTAGAACCCCAGCCCTCCCTGGCCTCCGCCGGGCCCCACCATGTCCCCAGTCAGTCTCCTTGCTCCCCCTGCAGCTGACAAGCCAGAGA
CGACCAAGGAGCAACTGGGAGAGTTCTACGAAGCTCTCGACTGCTTGTGCATTCCCAGGTCAGATGTCATGTACACCGACTGGAAAAAGGTAAACGCAAGGGATTGGACATTGCCCACCTTGTCCATGGCC
CAACTTGGGCAGCCCCAGAGGCCCAGAGCAGGAAAGCTGCCAGGCAAGGCTGCACAGCTAGGCAGATCTTCTGCTTTTAGGCACCTGCCTCACTGTAGGGACAGCTGAGCTCTACAGAGGCCCAGGGGTGG
TGGATGAGAGCCCAGGAGGGAGAAGTCCCTGTGAAACCAGGGAGGACCTGAAAGCTAACAGGAGGGAACAGCGTGAGCCACGGGGTTGGGGGATTGGCAATTGGAGGGGACGTAATGCGGGGAGTTACCAC
CTACAGACGCGTCCCAAACCCCAGGCTTTCACCCCAACCTCCACTCCCCGCTCATTTTTAATACCCGTGCAGTGGGGAATTGATACTGTGGTTTTCAATGTCACCCACACTGCAGCACGGCCACAGTCACC
ATCCCGATTTTTGCTACAAATGAAAATTACTGTATAATGAGCTCCTTAACACTTTTCTTTAAACCTGTGTTTGGAAGACTTGTGTTGGTGTGGCCCTGTGCCCTAATACCTGTGAAATCACAGCACCGATG
AGCTGGTTCCAATTTTTAAAATATATACATGCAGTACTTCCATGACTATTCAAAGAAAAACAATTCCTTCCATTTGCCACCTGAGATGACCACCAGGGATGTGAACTACCTCCTGCCCCATCCCCAGCCCC
AGGATCCTGGGACAGGGCTTATGAACGCAACCACTGTAGTCAGCTCACTTGATCCACAGCCTGGCACCTCCACTGTCTGGCTAGGGAGCCTCGAATGGGTCCCAAGGCCACCCTGCTCCTCAGTTACATCA
TCTGCATAGTAGTGGTGGTTGTGAGGAATTCAGGAGCTGCAGCATAAGGGCCCTGCAGGTACTATGTGCTCAGTAAATGCCAGTGGTTCTTAAGGGTCTGAGCTCCCATTGTAGAGGCAAGTAAGCTGAGG
TTCAGAGAAGAAAATGACTTGCCCAAGATCACCCAGCTGGGAAGTGACAGTGCCAGGGTTGGAGCCCTGGTTGAGCTGGTTCCACAGGCCAGAGCTCATTCTGCCCTCTCCCCGGAAGACCTCCCACCCTG
TCCCCATGCCTCTGCTTCTCCCTCACCCCAATTCCCCGCTGCCTTCTAGGATAAGTGTGAGCCACTGGAGAAGCAGCACGAGAAGGAGAGGAAACAGGAGGAGGGGGAATCCTAGCAGGACACAGCCTTGG
ATCAGGACAGAGACTTGGGGGCCATCCTGCCCCTCCAACCCGACATGTGTACCTCAGCTTTTTCCCTCACTTGCATCAATAAAGCTTCGCATCGGCCTTTCGAAACGAGGAGTACAATAAGTCGGTTCAGG
AGCCCTCAGGCAGGAGGGTTCACCGTGGGAACAGGGCAGGCCAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTGGGCACAGAGGCCACCCTGGACCGGGTGAGTGCCTGGGCTAGCCCTGTC
CTGAGCACATGGGCAGCTGCCTCCCTTCTCTGGGCTTCCCTTTACCTGCTGGCTGTGGTCGCACCCCCACTCCCAGCCCCCAACTCTCCCCGCCACAACATACATCCCATGCCCAGGAGGGTTCACCGTGG
                GAACAGGGCAGGCCAGCATAAGGTGGGGGCTGGATGTAGAGCCCTGGAGGCTTTGGGCACAGAGGCCACCCTGGACCGGGTGAGTGCCTGGGCTAGCCC
```

Interpreting protein sequences in functional terms

Where are we know????

Thisisanexampleofwhatwehaveachievedinthelasttwentyye
arsandwhatthechallengesaretomakesensefromtheKnowndat
asetwhatwearegeenratinginahighthroughputscale

WHAT WE WANT TO ACHIEVE?

Sequence Analysis

This is an example of what we have achieved in the last twenty years and what the challenges are: to make sense from the known data set that we are generating in a high Throughput scale.

....TO MAKE SENSE OUT OF IT

Taken from G. van Omen

Reductionism

- Reductionism has been very successful in Biology (Molecular Biology). *"The ultimate aim of the modern movement in biology is to explain* ***<u>all</u>*** *biology in terms of physics and chemistry"*. *F. Crick (1966)*

- Biological systems: prototype of complex systems. => Many biological phenomena could never be explained as a simple combination of the properties of the components ("the whole is more than the sum of the parts").

- Genome sequencing: Neither the number nor the characteristics of genes and proteins account for many characteristics of the organisms:
  - Similar number of genes in *Drosophila* y *C. elegans*.
  - High sequence similarity between human and mouse.
  - ...

•Van Regenmortel, M.H. (2004) Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. *EMBO Rep*, **5**, 1016-1020.

The “-omics” Paradigm of Biology

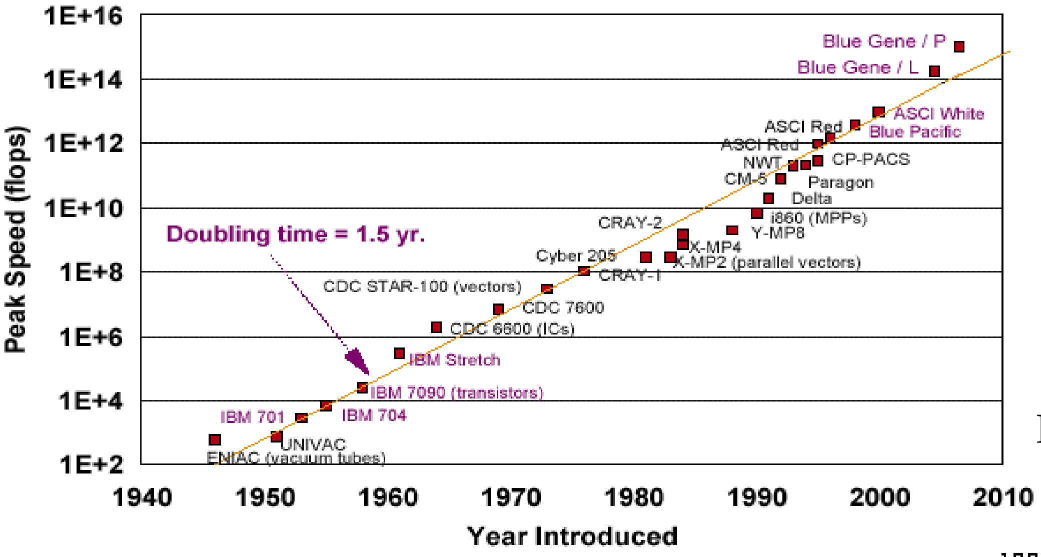
Genomic era → “Post-genomic” era

(massive production of biological data –sequences, …-)

(analysis and interpretation)

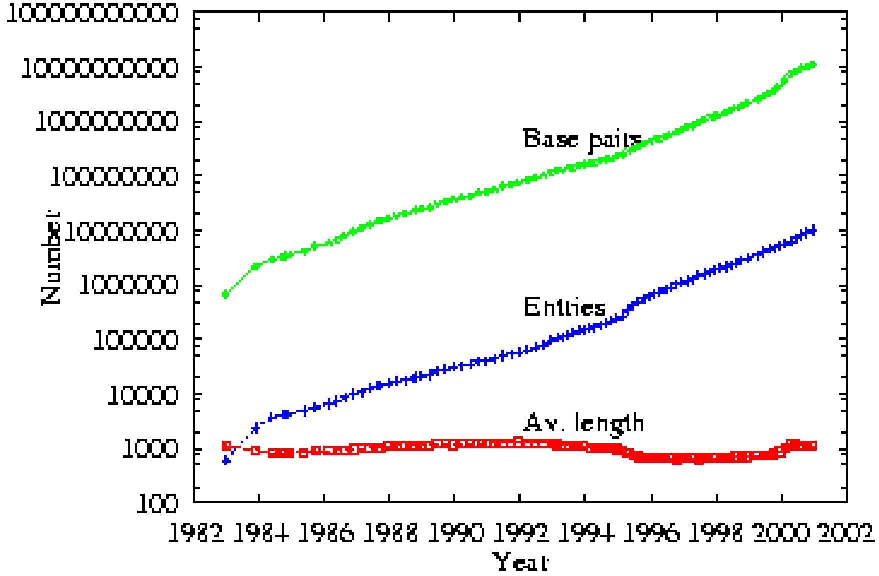
“Pre-genomic” era: the data itself contain the interpretation.
No data processing needed for obtaining the biological knowledge.
I.e. gel.

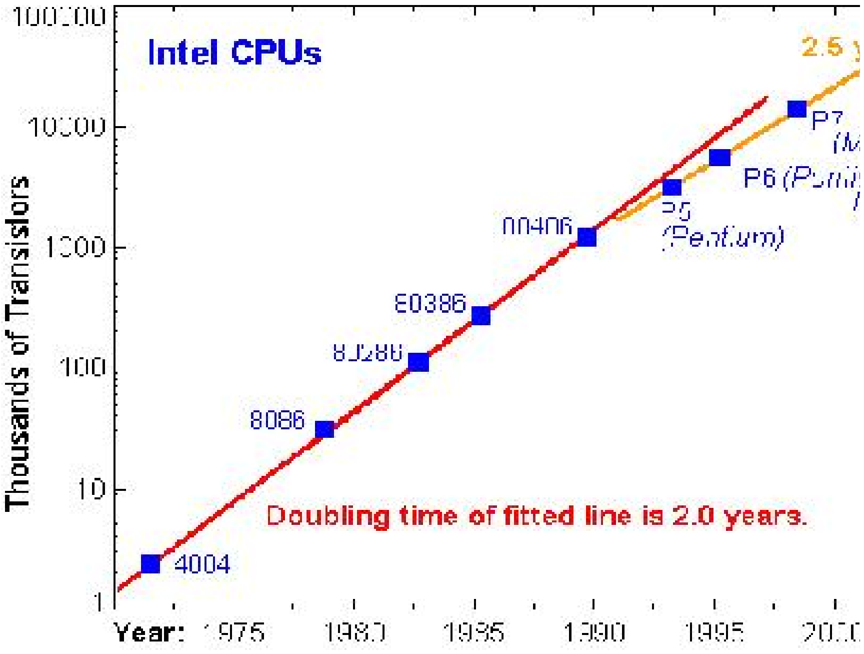
Can computers deal with that?



Moore’s law

Growth of the GenBank Database

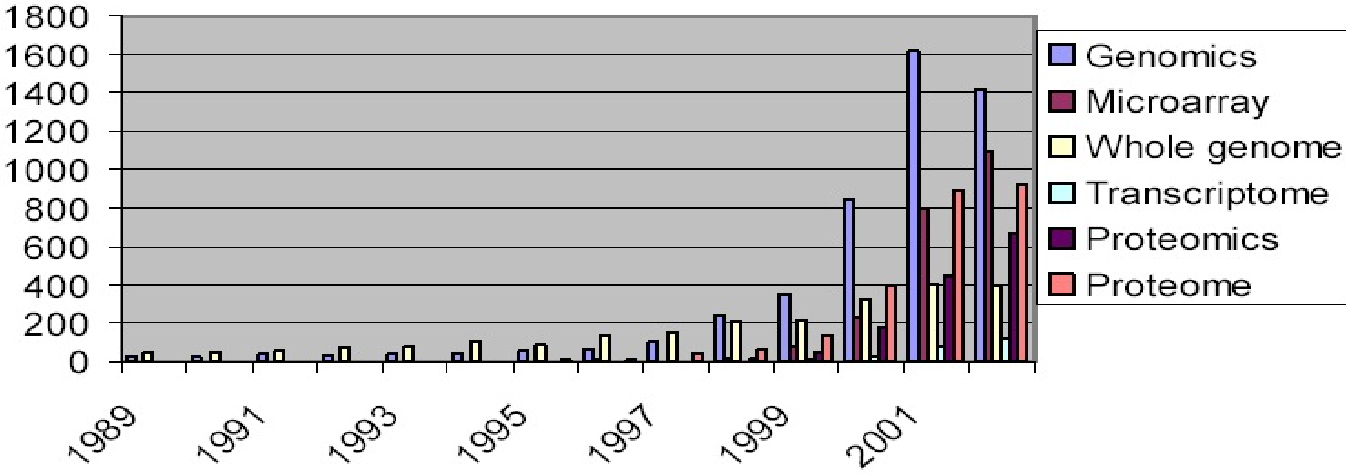


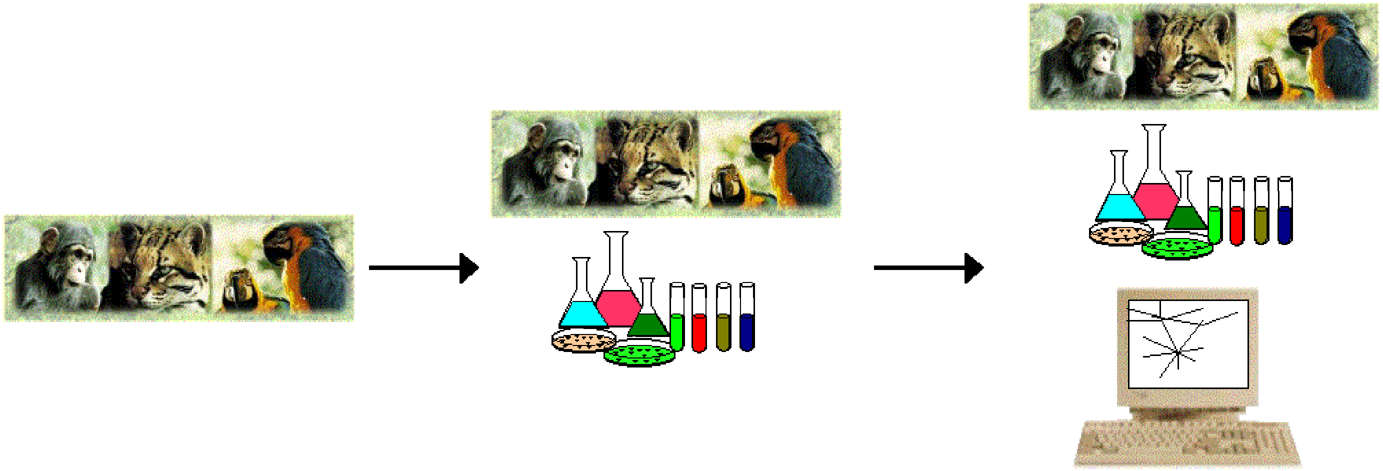


Center for Biological Sequence Analysis
The Technical University of Denmark

“-omics”

Publications (through Sept. 2002)





In vivo

In vivo + in vitro

In vivo + in vitro + in silico

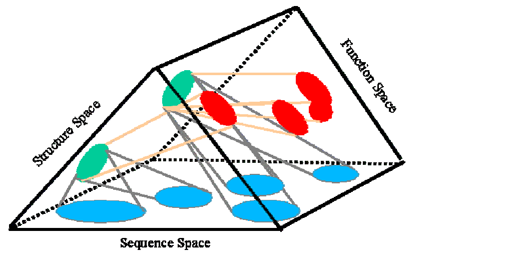
Protein Sequence Analysis

Module Overview

Florencio Pazos (CNB-CSIC)

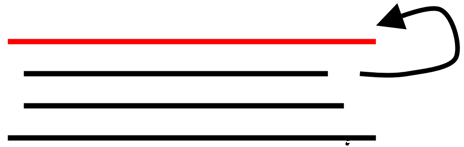
Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es

Protein Sequence Analysis
Module Overview – Week 1



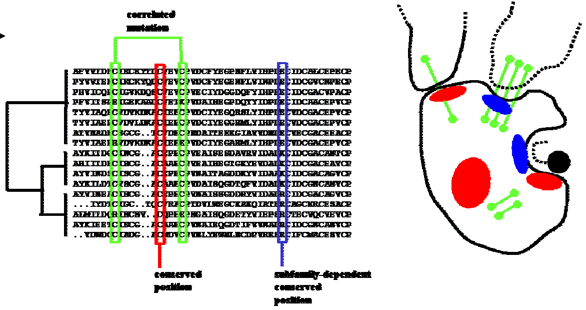
Monday 19th

- Introduction – Module overview
- Characteristics of the sequence space and relationships with structure and function spaces
- Sequence relationships for function prediction



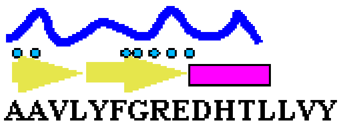
Tuesday 20th

- Extraction of functional features from sequence alignments
- Practical “Extraction of functional features from sequence alignments”



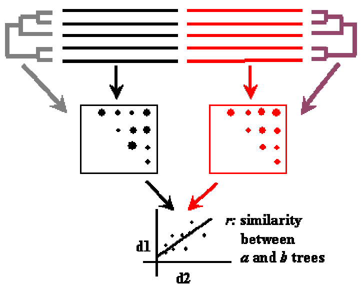
Wednesday 21st

- Extraction of structural features from sequence alignments
- Practical “Extraction of structural features from sequence alignments”



Thursday 22nd

- Sequence alignments for the prediction of protein-protein interactions
- Practical “Sequence alignments for the prediction of protein-protein interactions”



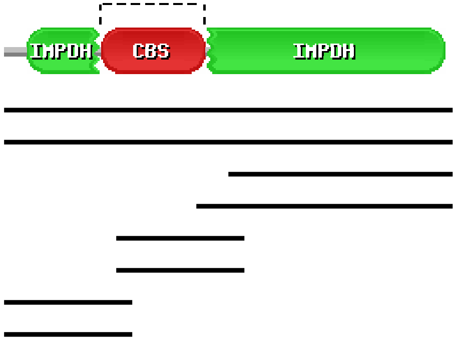
Friday 23rd

- Practical Work

Protein Sequence Analysis
Module Overview – Week 2

Monday 26th

- Sequence alignments and phylogeny
- Practical “Phylogeny”.

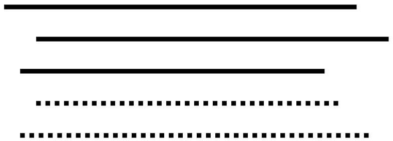


Tuesday 27th

- Protein domains
- Practical “Protein domains – PFAM”

Wednesday 28th

- Remote homology
- Practical “Remote Homology – PsiBLAST, HMMER”



Thursday 29th

- Practical Work / Group Presentations

Friday 30th

- Seminar

Master TecBio

Protein Sequence Analysis

Luis Sanchez (Protein Design Group, CNB-CSIC)
Ana Rojas (Structural Bioinformatics Group, CNIO)
Florencio Pazos (Protein Design Group, CNB-CSIC)

http://pdg.cnb.uam.es/cursos/Sardinia06/

sanchez@cnb.uam.es
arojas@cnio.es
pazos@cnb.uam.es