

Bioinformatics in view of
Protein Sequence
Protein Structure
Protein Function (only a glimpse)

Bioinformatics in view of

1. Protein Structure - Sequence:

- classification
- rules

2. Protein Space -

basis for protein families

ProtoNet - global classification

3. Structural Genomics -Integration

4. In practice

expression, MS,

2D-technology, PTM (if enough time)

Bioinformatics in Brief

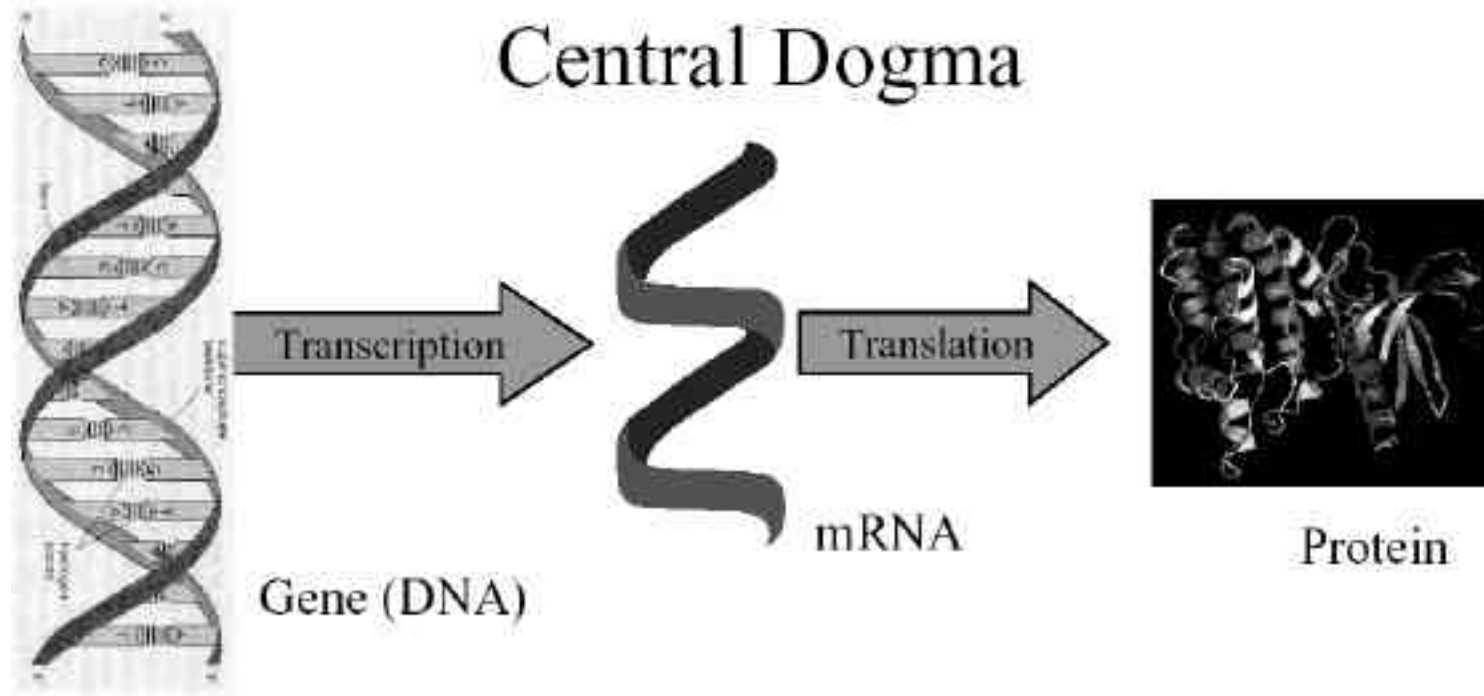
This week:

Bioinformatics -what is it, what for

DB for structures

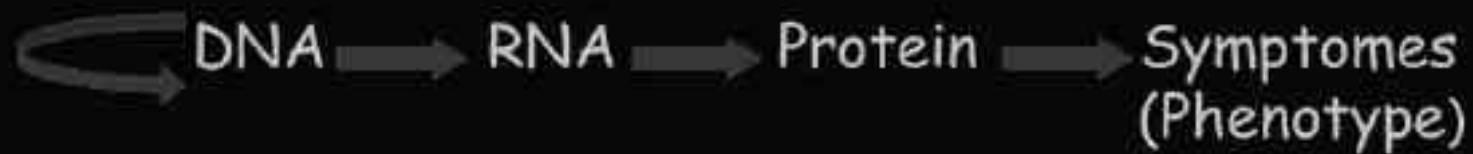
Structure Classification

Structure-Function link



Cells express different subset of the genes in different tissues and under different conditions

Central Paradigm of Molecular Biology



Hidden level of information

Fragile sites
DNA expansion
Structure shift

Compaction/accessibility

Modification

Amplification

Localization

Modification

Protein-protein interaction

Ligand interaction

Dynamic information

Central Paradigm of Bioinformatics

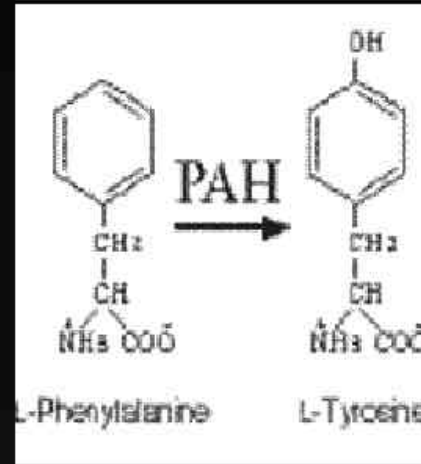
Genetic
Information

Molecular
Structure

Biochemical
Function

Symptoms

```
SRAAINKHIVA  
VSYQTVSRVUN  
VSTATVSRALA  
GWTTTVMHVIN  
SGVSAVSAILN  
GVSEKTRRDLN  
TAYATIMVWVE  
GSQPTVSRRLA  
MSIATITRGEN  
ISRETVGRILK  
FDISRLSHLFR  
LRPSRLAHLFR  
MTWETISRLLG  
TLEFHLRLEK
```

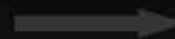
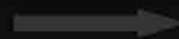


Phenylketonuria PKU

Biological Revolution Necessitates Bioinformatics

- New bio-technologies (automatic sequencing, DNA chips, protein identification, mass specs., etc.) produce large quantities of biological data.
- It is impossible to analyze data by manual inspection.
- Bioinformatics: Development of algorithms that enable the analysis of the data (from experiments or from databases).

Data produced by
biologists and
stored in database



New information
for biological
and medical use

Bioinformatics
Algorithms and Tools

A Big Goal

“The greatest challenge, however, is analytical. ...
Deeper biological insight is likely to emerge from
examining datasets with scores of samples.”

Eric Lander, “array of hope” *Nat. Gen.* 1999.

BIOINFORMATICS:

Provide methodologies for
elucidating biological knowledge
from biological data.



What is BIOINFORMATICS ?

A field of science in which Biology, Computer Science and Information Technology merge into a single discipline.

Goal: To enable the discovery of new biological insights and create a global perspective for biologists.

Disciplines:

- Development of new algorithms and statistics to assess relationships among members of large data sets.
- Analysis and interpretation of various types of data.
- Development and implementation of tools to efficiently access and manage different types of information.

Why use BIOINFORMATICS ?

- An explosive growth in the amount of biological information necessitates the use of computers for cataloging and retrieval.
- A more global perspective in experimental design (from "one scientist = one gene/protein/disease" paradigm to whole organism consideration).
- Data mining - functional/structural information is important for studying the molecular basis of diseases (and evolutionary patterns).

Why is it Hard to Elucidate from Sequence?

- Genetic information is redundant
 - Genetic code
 - Accepted amino acid replacements
 - Intron-Exon variation
 - Strain variation
- Structural information is redundant
 - Conformational changes
 - Different structures may result in similar functions
 - Different sequences result in the same structure
- Single genes have multiple functions.
 - May act as an metabolic enzyme and as a regulator.
 - Genes are 1-dimensional but function depends on 3-dimensional structure.

Bioinformatics

Genomics

Genomics

- Genomics includes the genetic mapping, physical mapping and sequencing of entire genomes
- Sequenced genomes include Human, Mouse, Fruit-Fly, Yeast, Arabidopsis, rice, *E. coli*, *M. tuberculosis*



Mycobacterium tuberculosis

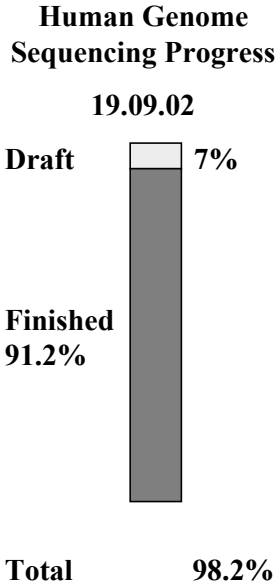
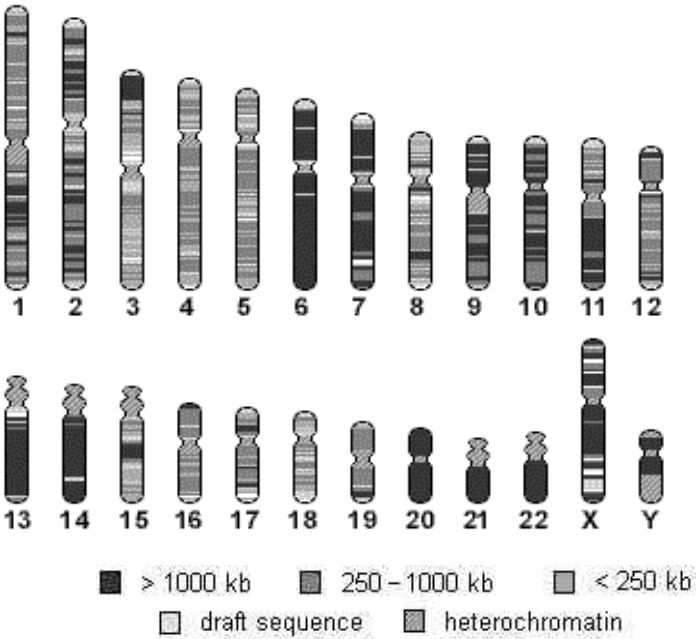


Proteins Class1&2

M. Linial

'02-'03

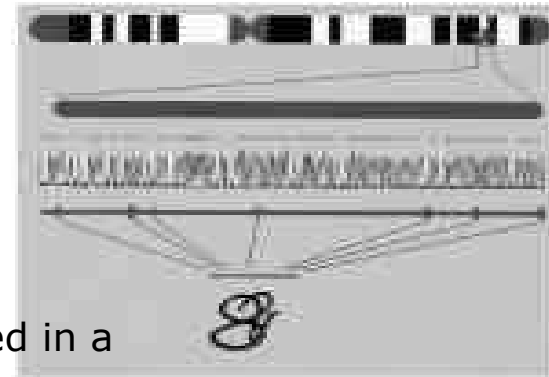
The Genomic Human data



And from a Practical View

Huge amount of data:
Mostly discrete (nt, aa, snp..)
Mostly accurate & reproducible
Mostly comparable

DB - A large collection of 'structured' data stored in a computer system.



So, one can apply:

Data mining
Statistics and validation
Database search
Computational tricks, old and new

Searching database for studying proteins

- ◆ Start with a genomic information (one of the new malaria/Anthrax genes..)
- ◆ Define the coding region (not so easy in many cases...)
- ◆ Finding clues
 - ◆ translate all 6 frame
 - ◆ test codon preference
 - ◆ nucleotide/ AA composition
 - ◆ intron/exon boundaries
 - ◆ double check the ends..
 - ◆ Biological support - EST

- ◆In the end - yes, we have a deduced AA sequence.

Searching database for studying proteins

- ◆ **Medical orientation:**

- ◆ Find Information on diseases or mutation related to that gene

- ◆ **Biochemical orientation:**

- ◆ Function ??
- ◆ Partners ??
- ◆ Localization??
- ◆ Homologues ??
- ◆ Modifications ??
- ◆ Biochemical pathway ??

- ◆ **Combined orientation:**

- ◆ Structural model (or even better solving the structure)
- ◆ Developing and designing drugs, inhibitors, antibodies...

For studying individual (or small set) protein - why we need all of this ?

- ◆ How general is the protein: Genomes – Human, Mouse, Yeast, E.coli...
- ◆ How diverged is the (putative) function?
- ◆ Where are the most 'important parts?' (evolution conservation)
- ◆ Are potential modification sites conserved?
- ◆ What are the regulatory elements that control its expression?
- ◆ Level of expression in various condition
 - ◆ Disease related mutation?
 - ◆ SNP variations in disease
- ◆ protein expression in different tissues, conditions, organisms..
- ◆ Family members.
- ◆ 3D Structures and models
- ◆

The Biggest trap...

◆ Lacking of a dynamic information

- ◆ The 'coding' of dynamic information is yet unresolved
- ◆ (probably the most important aspect of function...)

◆ The outcome:

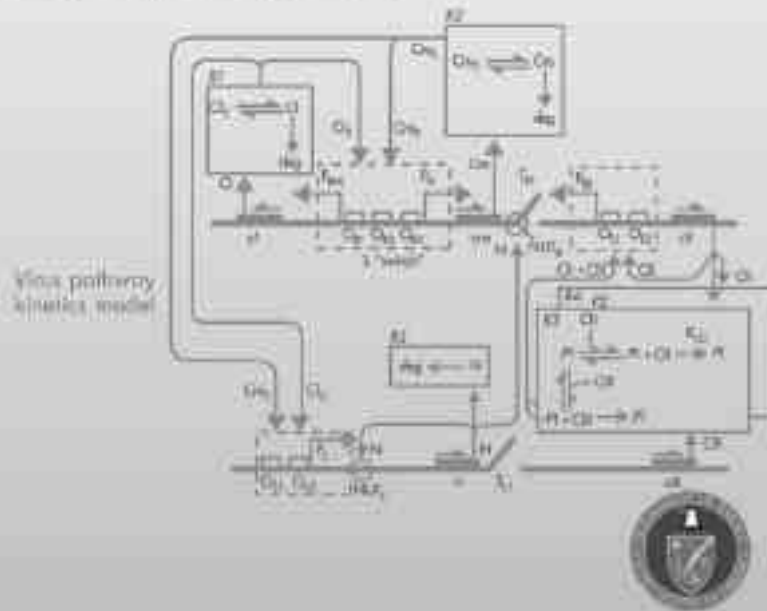
- ◆ Each protein is >10 different variance (each carry an alternative function)

◆ Why?

COMPUTATIONAL CAPABILITIES TO UNDERSTAND
AND PREDICT COMPLEX BIOLOGICAL SYSTEMS



IBM SP supercomputer
at Oak Ridge
National Laboratory



This trap applies to most current ‘static’ bioinformatics

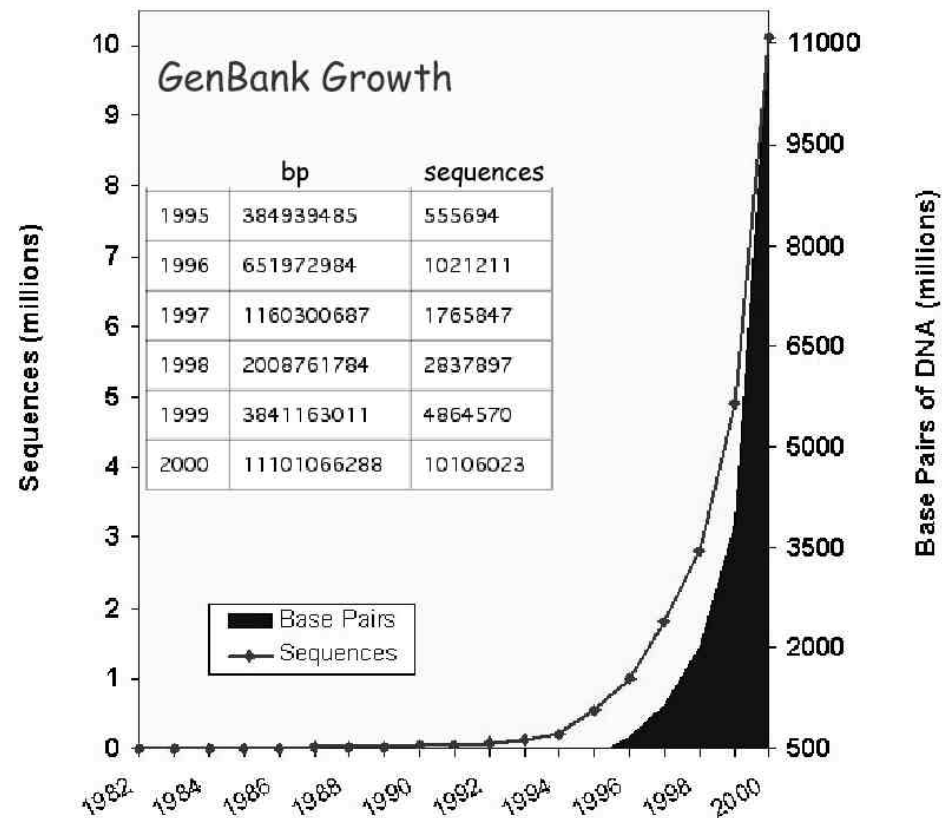
Ex: virus life cycle

System Biology

The growth in biological data

Expected:
Million
proteins
(shortly)

Now:
~800,000
nr



NCBI
National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed Entrez BLAST OMM TracEMBL Structure

Search for

SITE MAP

- About NCBI: general and contact information
- GenBank: sequence submissions, support and software
- Molecular databases: sequence, structure and taxonomy
- Literature databases: PubMed, OMM and Pubmed Central **ICM**
- Genomic biology: the human genome, whole genomes and related resources
- Tools: for data mining
- Research at NCBI

What does NCBI do?

Established in 1998 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

Draft Human Genome
Explore [human genome resources](#) or browse the human genome sequence using the [Map Viewer](#).

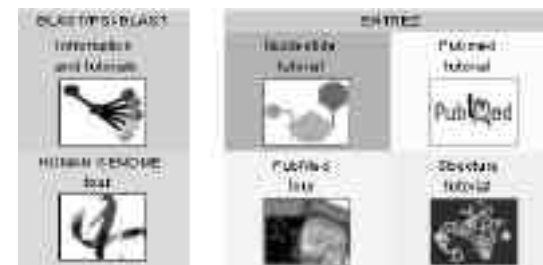
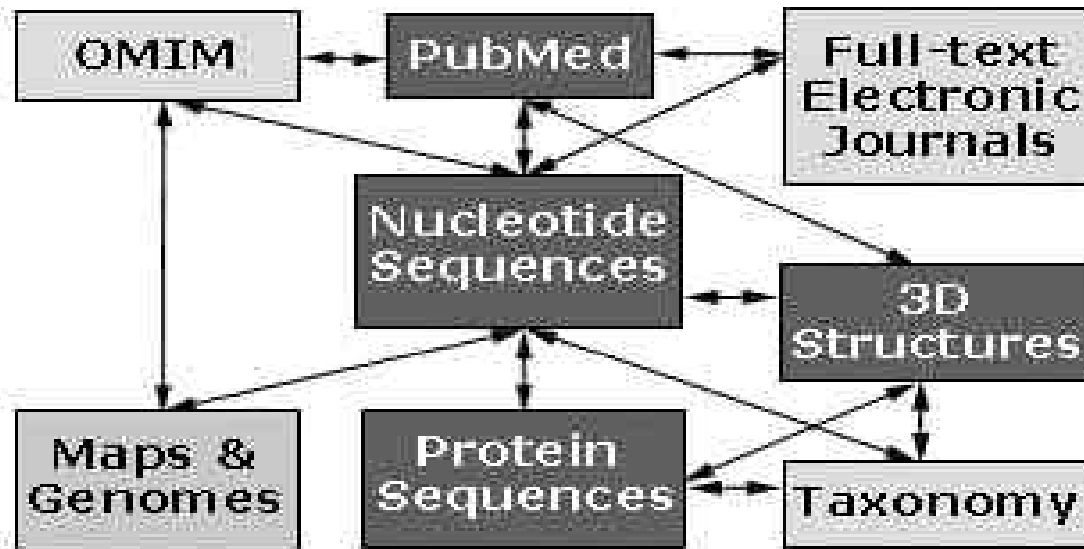
DART: A new tool

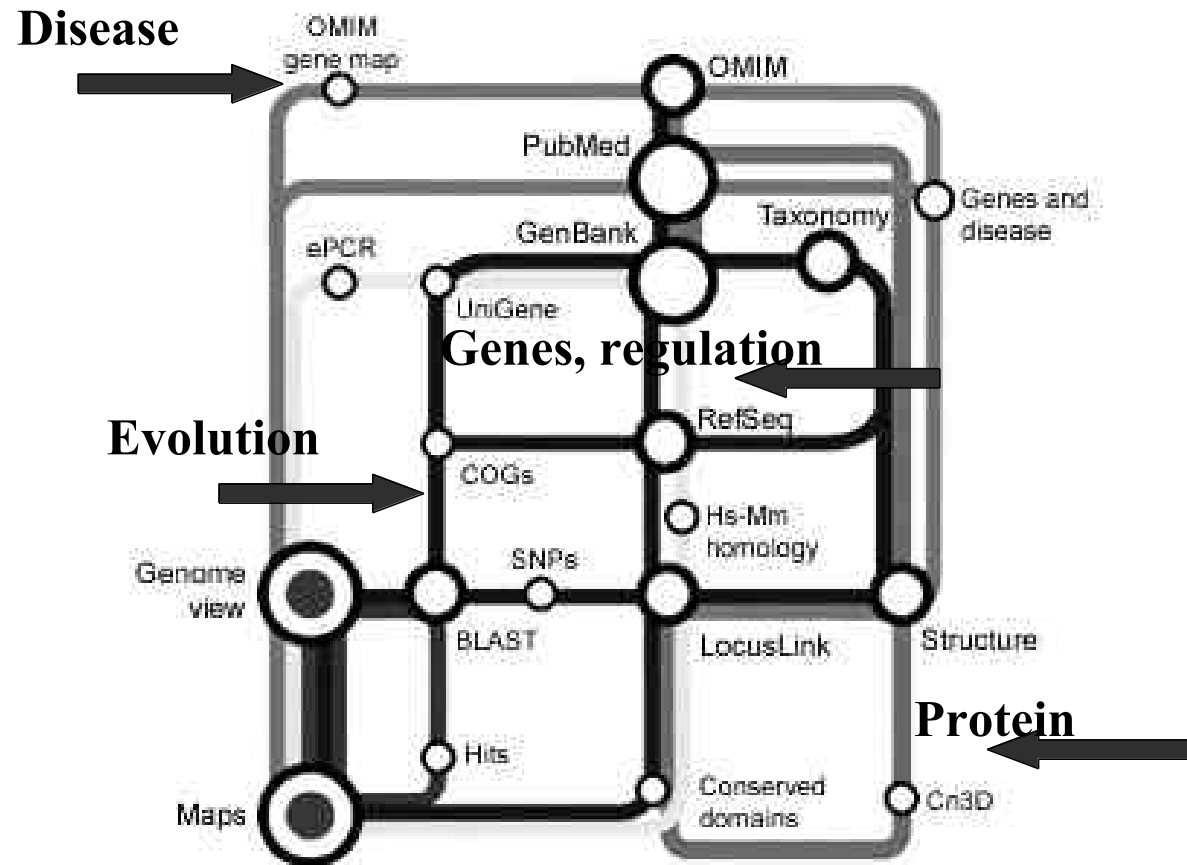
Want to locate protein neighbors by domain architecture? Learn about NCBI's new Domain Architecture Retrieval Tool.
<http://www.ncbi.nlm.nih.gov/>

Hot Spots

- ▶ Cancer genome anatomy project
- ▶ Clusters of orthologous groups
- ▶ Coffee Break
- ▶ Electronic PCR
- ▶ Gene expression omnibus
- ▶ Genes and disease
- ▶ Human genome resources
- ▶ Human map viewer
- ▶ Human/mouse homology maps

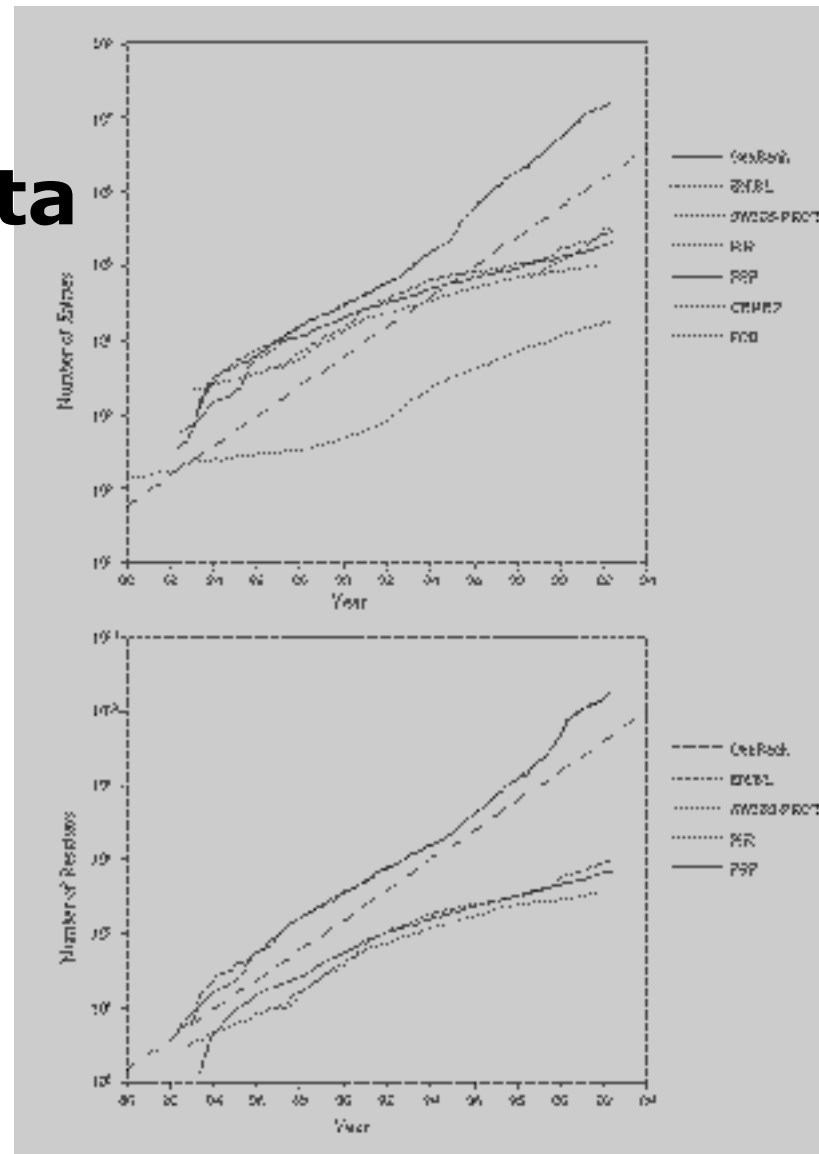
Linking Databases – Integration














The growth in biological data

Protein structures lagging behind



Primary Sequence Databases

- In the early 1980s several primary database projects evolved in different parts of the world

Nucleic Acids	Protein
EMBL 	PIR 
GenBank 	MIPS 
DDBJ 	Swiss-Prot 
	TrEMBL 
	NRL_3D 
	GenPept 

Inherited Problems of Databases 20 Years Later

- During the early 1980s no one envisaged that databases would become so huge
- Many databases are regulated by users rather than centrally, except for Swiss-Prot
 - Only the owner of the sequence data can name it
 - Dependency on annotation of submitter
 - Sequences are not up to date
 - Large degree of redundancy in databases

Swiss-Prot

- Established in 1986 and maintained collaboratively by SIB (Swiss Institute of Bioinformatics) and EBI/EMBL
- Provides high-level annotations, including description of protein function, structure of protein domains, post-translational modifications, variants, etc
- Aims to be minimally redundant
- Linked to many other resources -**Consider the best**



TrEMBL

- Translated EMBL was created in 1996 as a computer annotated supplement to Swiss-Prot.
- Contains translations of all coding sequences in EMBL
- SP-TrEMBL contains entries that will be incorporated into Swiss-Prot
- REM-TrEMBL contains entries that are not destined to be included in Swiss-Prot (Ig, T-cell receptors, patented sequences) (no accession #)

