Bioinformatics: Knowledge-representation in molecular biology

Sándor Pongor
Protein Structure and Bioinformatics, ICGEB, Trieste
Representation of biological knowledge

Source: NCBI
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1962</td>
<td>Pauling’s theory of molecular evolution</td>
</tr>
<tr>
<td>1967</td>
<td>Margaret Dayhoff’s Atlas of Protein Sequences</td>
</tr>
<tr>
<td>1970</td>
<td>Needleman-Wunsch algorithm</td>
</tr>
<tr>
<td>1977</td>
<td>DNA sequencing and software to analyze it (Staden)</td>
</tr>
<tr>
<td>1981</td>
<td>The concept of a sequence motif (Doolittle)</td>
</tr>
<tr>
<td>1982</td>
<td>Phage lambda genome</td>
</tr>
<tr>
<td>1983</td>
<td>Database search (Wilbur-Lipman)</td>
</tr>
<tr>
<td>1985</td>
<td>FASTP/FASTN: fast sequence similarity searching</td>
</tr>
<tr>
<td>1987</td>
<td>Sequence profiles</td>
</tr>
<tr>
<td>1987</td>
<td>EMBL, Genbank, Swiss-Prot databases</td>
</tr>
<tr>
<td>Year</td>
<td>Event</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1988</td>
<td>National Center for Biotechnology Information (US)</td>
</tr>
<tr>
<td>1988</td>
<td>EMBnet network for database distribution</td>
</tr>
<tr>
<td>1990</td>
<td>BLAST: fast sequence similarity searching</td>
</tr>
<tr>
<td>1991</td>
<td>EST: expressed sequence tag sequencing</td>
</tr>
<tr>
<td>1993</td>
<td>Sanger Centre, Hinxton, UK</td>
</tr>
<tr>
<td>1994</td>
<td>EMBL European Bioinformatics Institute, Hinxton, UK</td>
</tr>
<tr>
<td>1995</td>
<td>First bacterial genomes</td>
</tr>
<tr>
<td>1996</td>
<td>Yeast genome</td>
</tr>
<tr>
<td>1997</td>
<td>PSI-BLAST</td>
</tr>
<tr>
<td>1998</td>
<td>Worm (multicellular) genome</td>
</tr>
<tr>
<td>2000+</td>
<td>The rice and human genomes.</td>
</tr>
<tr>
<td></td>
<td>Microarrays, high throughput methods, new generation sequencing…</td>
</tr>
</tbody>
</table>
The ingredients

- Data collection techniques (DNA sequencing, protein sequencing, microarrays)
- Theoretical milestones (concepts of DNA structure, protein structure, evolution)
- Algorithms and programs (BLAST, FASTA)
- Databases
- Institutions
- Complex genomic and high throughput data
The evolution of bioinformatics as seen in the 90’s

- **bioinformatics in the 90s**
  - primary databases containing large-scale experimental data in genomics and proteomics
  - understanding functions and utilities of individual genes or proteins

- **bioinformatics now**
  - secondary databases (knowledge bases) containing accumulated biological knowledge
  - understanding functions and utilities at the molecular, cellular and organism levels

- **bioinformatics in the future**
  - a complete computer representation of the cell and the organism
  - understanding basic principles of the higher complexity of biological systems
Bioinformatics is an approach to biology...
Model, description and visualization
Why is bioinformatics important?

- “A paradigm shift in biology: from data collection to data processing”

- “Biotechnology is the industrial use of biological information”
  Lee Hood, in *The Economist*, 1997
Current trends

- Massive data processing
- Massive generation of data: sequences (genomics) functions (functional genomics) and structures (structural genomics)
- Interpretation of data: data mining, data warehousing techniques

Informatics strike back...
Structural genomics

- Classify proteins (Database of protein motifs)
- Choose and express representative proteins from all families
- Determine structure by X-ray or NMR
- Predict the rest by homology modelling

Bioinformatics
Functional genomics

- Sequence complete genome
- Identify protein coding regions
- Identify unique genes
- Gene knockout
- Functional analysis (phenotype, detailed functional characterization..)
- Structural studies, drug development

Bioinformatics
New data representations

Data (property) visualization
What bioinformatics is…

- Data Analysis
- Interpretation
- Processing of raw sequence data & instrument output
- Database maintenance
- Biocomputing, biomathematics

Infrastructure

Research

USER
Bioinformatics: managing information for the life sciences

- For: Biomedicine, Agriculture
- In: Academic and Industrial Research and Development, Medical Practice
- Scientific Infrastructure (service)
- Advanced Informatics (research)
- Education (biologists vs. informaticians)

An independent field of study but a general approach to biology
Current challenges – the three gaps

- Understanding (“annotating”) new data: “annotational gap”

- Translating data to practice: personalized medicine, epidemics… “translational gap”

- Making users (biologists, medical doctors) aware of what is there and how to use it… (“communication gap”)
Protein Structure and Bioinformatics

- Established as a resource group for protein chemistry and protein engineering
- In charge of bioinformatics services since 1991
- Research projects on bioinformatics, structural biology, systems modeling
- Currently includes 12 researchers and students
ICGEB bioinformatics

- Biological computing service for 800 users from 47 countries
- 2-3 training courses per year, 1400+ students in 18 years...
- Methods development (classification, machine learning)
- WWW-services: DNA-tools, protein domain identification
EMBnet: A world wide network of bioinformatics

- 32 national nodes, 35,000 registered users.
- 12 specialist nodes including all major European database producers.
- Includes China, India, Australia
- Education: High level courses organised in member countries. WWW-tutorials.
- A coordinated network of bioinformatics services, a global technical and educational resource

www.embnet.org
Bioinformatics summer course

-One of the oldest continuous teaching traditions in Europe, over 1300 students since 1991
- Introduction to theory and practice of bioinformatics
- Contact with the centers of (NCBI, EBI, SwissProt, KEGG)
A take-home message

- Bioinformatics is a general approach (paradigm) in biology today.
- A bioinformatician has to understand:
  - The biological question, the biological model
  - The data-collection technology, the data model
  - The mathematics/statistics of data-evaluation
- This is why we have this course
Thank you for your attention…
Bioinformatics:
Knowledge-representation in molecular biology

Sándor Pongor
Protein Structure and Bioinformatics, ICGEB, Trieste
An overview of bioinformatics

- History and development
- Model, description and visualization
  - Sequences
  - 3D structures
  - Networks
  - Text (abstracts)
- Similarity and classification:
  - similarity measures (structured, unstructured)
  - database search
  - consensus descriptions
- Integrated resources
The subjects: Molecular structures

Sequences

Extended sequences (e.g. disulphide-topologies)

Domain-cartoons (sec. str. cartoons)

3D structures

Diagrams (hydrophobicity plots, helical circles)

3D cartoons
A structural model

Substructures

Relationships

Structure

Susbstructures, relations, rules = onthology

Entity-relationship model
Structures As Database Records

Identification
Name of protein
Organism
Function
Cross-references
...
Domain structure
Sec. structure
Disulphides
...

Sequence (structure)
quinetdttvivtvtpprarivgyltvgllseeg
depqylldlpstatsvnipltprgrkytvnyeises
egeqnliilstsqttadappdptvqvdvdsiv
wsrprapitgyrvyspsvgssteltnlpelatnsv
tldqpvqyvnityyaveengestpvfiquertt
gyrvdipvvnlpghqrlpvrntfatevg
vtyhfkv

Database record, fields

ANNOTATIONS
SEQUENCE OR STRUCTURE
CIPKWNRCGPKMDGVCCEPYTCTSDYYYYNC
The subjects of bioinformatics

- Stored data = descriptions for computers
- Visualization, text = simplified descriptions for humans
- Models (knowledge)
SEQUENCES
Model: Chemical structure

Description: Series of characters

Simplified and/or extended visualization

IFPPVPGP

domain1

Binding site
Sequences as language

Character strings, computer-languages, Chomsky et al, etc.
3D STRUCTURES
Chimie dans l’espace

Van t’Hoff
1852-1911

1898
Some molecules are more equal then others…

…”This figure is purely diagrammatic. The two ribbons symbolize the phosphate-sugar chains, and the horizontal rods the pairs of the bases holding the chains together. The vertical line marks the fibre axis”
Protein models
3D structures

- Model: 3D chemical structures
- Description: 3D coordinates
- Simplified and/or extended visualization

\[(x_i, y_i, z_i)_n\]
NETWORKS
Small molecules – classical graphs

Loschmidt, 1861
Kekulé, 1865
Crum Brown, 1861
Cayley, 1872
Van’t Hoff, 1898
The transcription regulatory networks

E. coli

S. cerevisiae
TEXTS (article abstracts)
The language of bibliographies

Structures As Database Records

Identification
Name of protein
Organism
Function
Cross-references
...
Domain structure
Sec. structure
Disulphides
...

Keyword-collections, ontologies, etc.
Texts (abstracts)

- Model: ??
- Description: structured files (records, fields), standardized language
- Simplified and/or extended visualization
Models

SEQUENCES

3-D

NETWORKS

TEXT
An overview of bioinformatics

- History and development
- Model, description and visualization
  - Sequences
  - 3D structures
  - Networks
  - Text (abstracts)
- Similarity and classification:
  - similarity measures (structured, unstructured)
  - database search
  - consensus descriptions
- Integrated resources
The concept of similarity I

Shared parts

Shared context

...easier if modular
The concept of similarity II

...Easy for humans, hard for computers
Multiple Objects

Similarity groups or neighborhoods

Multiple alignments

Evolutionary trees

Structural similarity

Metabolic pathways

Subunit structures, ligands

Genomes

Trajectories

Context (function)
Similarity of molecules

Shared parts

Shared relations

Shared context
Quantitative comparison

Unstructured models
- Typical form: numbers, lists, vectors (x1, x2, ... x3)
- Similarity score
- Clustering, classification etc.

Structured models
- Typical form: sequences, networks etc.
- Alignment (matching)
- Similarity score
- Clustering, classification etc.
Quantification of sequence similarity: sequence alignment and its scoring

Score = sum contributions of matches subtract penalties for mismatches
"The similarity of objects can be best described as partial identities of components and relationships.

Erich Goldmeier, The similarity of perceived forms, 1936
Using similarity: Comparing one sequence with a group (database)

**Ranked list of best similarities**

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWISSALL:IAAI</td>
<td>457.36</td>
<td>ALPHA-AMYLASE INHIBITOR AAI. 2/9</td>
</tr>
<tr>
<td>SWISSALL:O426</td>
<td>152.82</td>
<td>CELLULOSE BINDING PROTEIN</td>
</tr>
<tr>
<td>SWISSALL:GUX</td>
<td>145.77</td>
<td>EXOGLUCANASE I PRECURSOR</td>
</tr>
<tr>
<td>SWISSALL:Q126</td>
<td>145.66</td>
<td>CELLULASE (EC 3.2.1.91)</td>
</tr>
</tbody>
</table>

Similarities??

**EXPECTation Threshold (E parameter)**

```
<table>
<thead>
<tr>
<th>V</th>
<th>Observed Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>6336 1688</td>
</tr>
<tr>
<td>6310</td>
<td>4648 1618</td>
</tr>
<tr>
<td>3980</td>
<td>3030 886</td>
</tr>
<tr>
<td>2510</td>
<td>2144 706</td>
</tr>
<tr>
<td>1580</td>
<td>1438 438</td>
</tr>
<tr>
<td>1000</td>
<td>1000 272</td>
</tr>
<tr>
<td>631</td>
<td>728 185</td>
</tr>
<tr>
<td>398</td>
<td>543 141</td>
</tr>
<tr>
<td>251</td>
<td>402 103</td>
</tr>
<tr>
<td>158</td>
<td>299 63</td>
</tr>
<tr>
<td>100</td>
<td>236 43</td>
</tr>
<tr>
<td>63.1</td>
<td>134 3</td>
</tr>
<tr>
<td>39.8</td>
<td>131 2</td>
</tr>
<tr>
<td>25.1</td>
<td>129 2</td>
</tr>
<tr>
<td>15.8</td>
<td>127 0</td>
</tr>
<tr>
<td>10.0</td>
<td>127 1</td>
</tr>
<tr>
<td>6.31</td>
<td>126 0</td>
</tr>
<tr>
<td>3.98</td>
<td>126 4</td>
</tr>
<tr>
<td>2.51</td>
<td>122 0</td>
</tr>
<tr>
<td>1.58</td>
<td>122 0</td>
</tr>
<tr>
<td>1.00</td>
<td>122 0</td>
</tr>
<tr>
<td>0.63</td>
<td>122 0</td>
</tr>
<tr>
<td>0.40</td>
<td>122 4</td>
</tr>
<tr>
<td>0.25</td>
<td>122 0</td>
</tr>
<tr>
<td>0.16</td>
<td>122 0</td>
</tr>
<tr>
<td>0.10</td>
<td>122 0</td>
</tr>
<tr>
<td>0.063</td>
<td>122 0</td>
</tr>
<tr>
<td>0.040</td>
<td>122 0</td>
</tr>
<tr>
<td>0.025</td>
<td>122 0</td>
</tr>
<tr>
<td>0.016</td>
<td>122 1</td>
</tr>
<tr>
<td>0.010</td>
<td>121 0</td>
</tr>
<tr>
<td>0.0063</td>
<td>120 0</td>
</tr>
<tr>
<td>0.0040</td>
<td>120 0</td>
</tr>
<tr>
<td>0.0025</td>
<td>120 1</td>
</tr>
<tr>
<td>0.0010</td>
<td>120 0</td>
</tr>
</tbody>
</table>

BLAST program
Using similarity: comparing a group with itself

Similarity group or neighbourhood

Multiple alignment

Mathematical consensus for database search

Regular expressions
Consensus sequence
Frequency matrix
Markov chains
Neural networks
etc.

CLUSTAL program

Publish

CGPK-MDGVPCCEPY
CGGQNWGPTCCASG
CSPTSYN---CCR--
CSRLMY---DCCT--
CIPYYL---DCCEPL
### Similarities: a practical overview

<table>
<thead>
<tr>
<th>SEQUENCES</th>
<th>3D</th>
<th>NETWORKS</th>
<th>PAPERS</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Bulk</th>
<th>“Glycine-rich”</th>
<th>“α-helical”</th>
<th>“scale-free”</th>
<th>“genomics”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure-alignment</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td>(metabolic pathways)</td>
<td>same author, common references</td>
</tr>
<tr>
<td>Motifs</td>
<td>G-RR</td>
<td><img src="image3.png" alt="Image" /></td>
<td></td>
<td>“Joe Doe, folding”</td>
</tr>
</tbody>
</table>
An overview of bioinformatics

- History and development
- Model, description and visualization
  - Sequences
  - 3D structures
  - Networks
  - Text (abstracts)
- Similarity and classification:
  - similarity measures (structured, unstructured)
  - database search
  - consensus descriptions
- Integrated resources
Biological knowledge as a network of data

The world according to a PC...

Source: NCBI
Search on a preprocessed, integrated database: the importance of a good neighbourhood.
Models are human constructs...

THIS IS NOT A PIPE!
Models are human constructs...

THIS IS NOT A MOLECULE
The central dogma:

DNA $\rightarrow$ RNA $\rightarrow$ Protein

Dogma, paradigm, mythology
New central dogma:
Self-assembly, catalysis, replication, networks

Polymers: Initiate, elongate, terminate, fold, modify, localize, degrade

Evolution + Self assembly, Systems biology
Summary of topics discussed

- History and development
- Model, description and visualization
  - Sequences
  - 3D structures
  - Networks
  - Text (abstracts)
- Similarity and classification:
  - similarity measures (structured, unstructured)
  - database search
  - consensus descriptions
- Integrated resources
Summary of the introduction

- Bioinformatics is the science of biological information or rather a computer-based approach to biological problems.
- All kinds of biological data are *structures defined with entities and relationships* (metabolites, genes, networks).
- Typical tasks: Similarity search, categorization and clustering
- Simultaneous handling of many, complex data-types
On-line help to this lecture

- **Bioinformatics tutorials on-line**
  
  http://www.ebi.ac.uk/2can/home.html

- **ICGEBnet**
  
  http://www.icgeb.org/~netsrv/

- **The Trieste bioinformatics course**
  
  http://www.icgeb.org/~netsrv/netcourse.html
Reading about bioinformatics

- In depth introduction
- Genomics research problems
- Evolutionary principles
- Math principles
Theoretical overview: Sándor Pongor

Sequence database searching, theory and practice (Dave Judge and Jack Leunissen)

Nucleic acid databases, Medline, Pubmed (David Landsman)

Functional genomics databases, KEGG (Minoru Kanehisa)

EBI Services (Jim Watson)

Protein databases, Swissprot, Prosite (Marie-Claude Blatter)

Genome analysis (Martin Bishop)