Bioinformatics:
problems, algorithms and perspectives.

Jaroslaw Meller

Biomedical Informatics, Children’s Hospital Research Foundation, University of Cincinnati

Dept. of Informatics, Nicholas Copernicus University

Quiz:

codon

transcription

PCR

microarray

float

compiler

NP-hard

sequence alignment

Bioinformatics: putting A,T,C,G's into computer ...

In order to make sense out of it and facilitate further experiments by inference, modeling and computer simulations.

What is bioinformatics?

“Roughly, bioinformatics describes any use of computers to handle biological information. In practice the definition used by most people is narrower; bioinformatics to them is a synonym for "computational molecular biology“ -- the use of computers to characterize the molecular components of living things.”

“The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.”

More: Bioinformatics_Org Bioinformatics FAQ.htm

Bioinformatics vs. computational biology:

"Computational biology is not a "field", but an "approach" involving the use of computers to study biological processes and hence it is an area as diverse as biology itself."

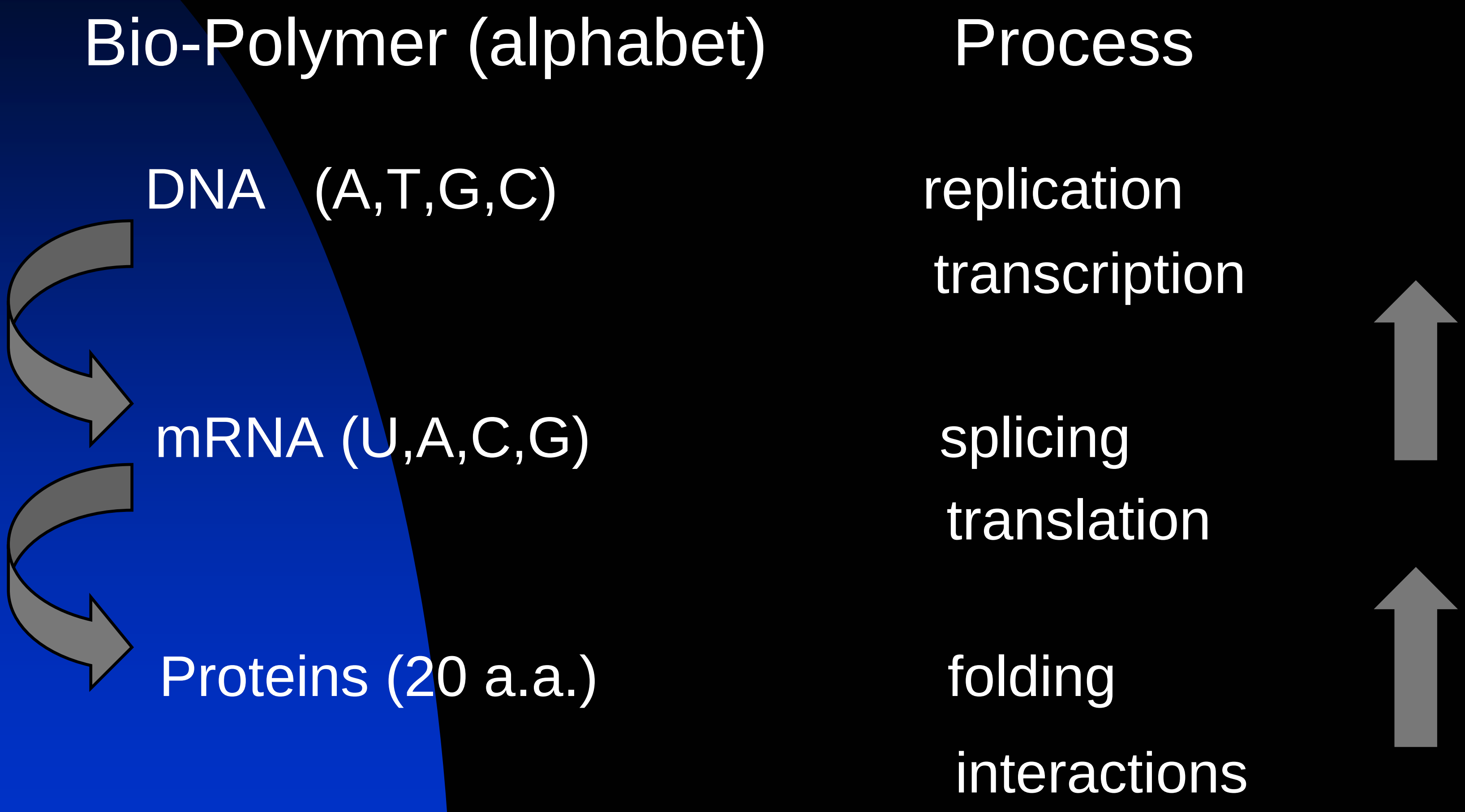
Richard Durbin, Head of Informatics at the Wellcome Trust Sanger Institute:

"I do not think all biological computing is bioinformatics, e.g. mathematical modeling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information."

Related fields:

- Genomics (functional, structural)
- Proteomics
- Cheminformatics
- Pharmacogenomics
- Medical Informatics

Bioinformatics: genes, proteins and computers ...



Bio-Polymer (alphabet)	Process
DNA (A,T,G,C)	replication
	transcription
mRNA (U,A,C,G)	splicing
	translation
Proteins (20 a.a.)	folding
	interactions

Lipids, polysaccharides, membranes and signal transduction, environmental signals etc.

Problems and methods:

Problem → Algorithms → Programs

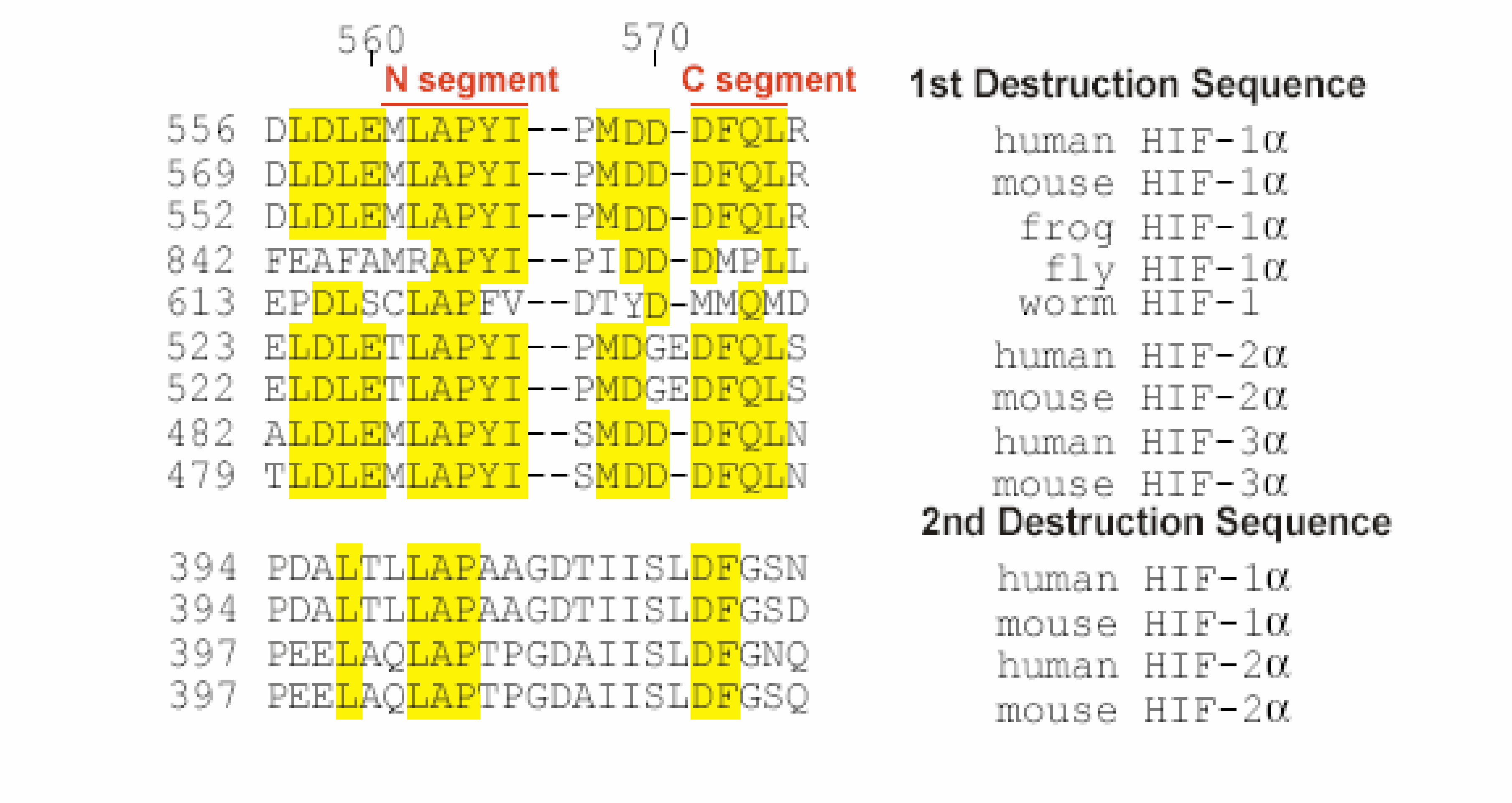
Sequencing → Fragment assembly problem → The Shortest Superstring Problem → Phrap (Green, 1994)

Gene finding → Hidden Markov Models, pattern recognition methods → GenScan (Burge & Karlin, 1997)

Sequence comparison → pairwise and multiple sequence alignments → dynamic algorithm, heuristic methods → BLAST (Altschul et. al., 1990)

Searching for binding motif:

LxxLAP motifs found on human RNA Pol II (C-terminus of Rpb1).



```
         560          570
      N segment    C segment
556 DLDLEMLAPYI--PMDD-DFQLR
569 DLDLEMLAPYI--PMDD-DFQLR
552 DLDLEMLAPYI--PMDD-DFQLR
842 FEAFAMRAPYI--PIDD-DMPLL
613 EPDLSCLAPFV--DTYD-MMQMD
523 ELDLETLAPYI--PMDGEDFQLS
522 ELDLETLAPYI--PMDGEDFQLS
482 ALDLEMLAPYI--SMDD-DFQLN
479 TLDLEMLAPYI--SMDD-DFQLN

394 PDALTLLAPAAGDTIISLDFGSN
394 PDALTLLAPAAGDTIISLDFGSD
397 PEELAQLAPTPGDAIISLDFGNQ
397 PEELAQLAPTPGDAIISLDFGSQ
```

1st Destruction Sequence

```
human HIF-1α
mouse HIF-1α
 frog HIF-1α
  fly HIF-1α
 worm HIF-1
human HIF-2α
mouse HIF-2α
human HIF-3α
mouse HIF-3α
```

2nd Destruction Sequence

```
human HIF-1α
mouse HIF-1α
human HIF-2α
mouse HIF-2α
```

Trying out the bioinformatist’s routine: BLAST searches.

- Let us BLAST some sequences …

  NCBI HomePage.htm

- Why homology is so useful?

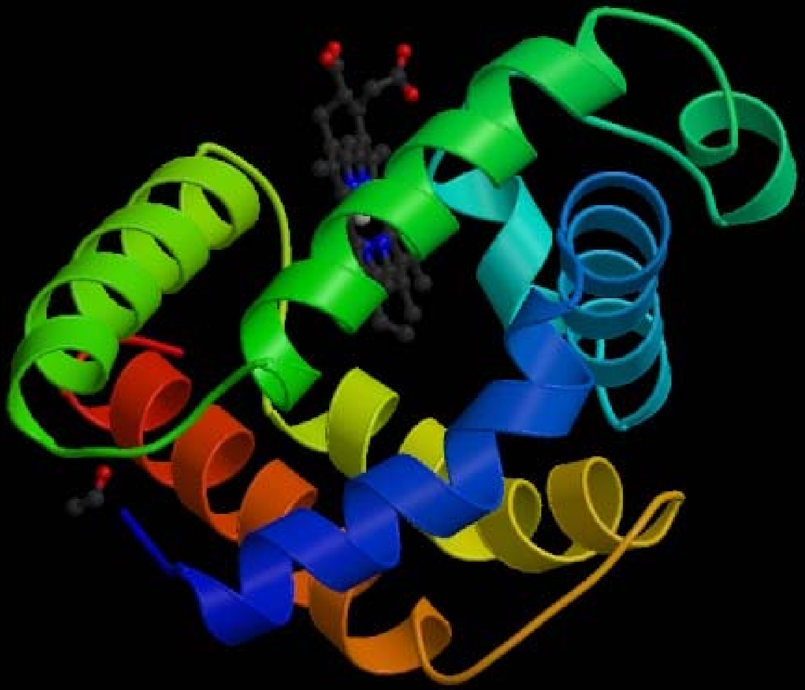
From genes to drugs: protein machinery of life

- Genes determine protein sequences
- Proteins are crucial agents in living organisms
- Understanding genes = understanding proteins with their structure and function

Significance of protein folding problem

V
L
S
E
G
E
W
Q
L
V
L
V
.
.
.

→



⇄ O_2

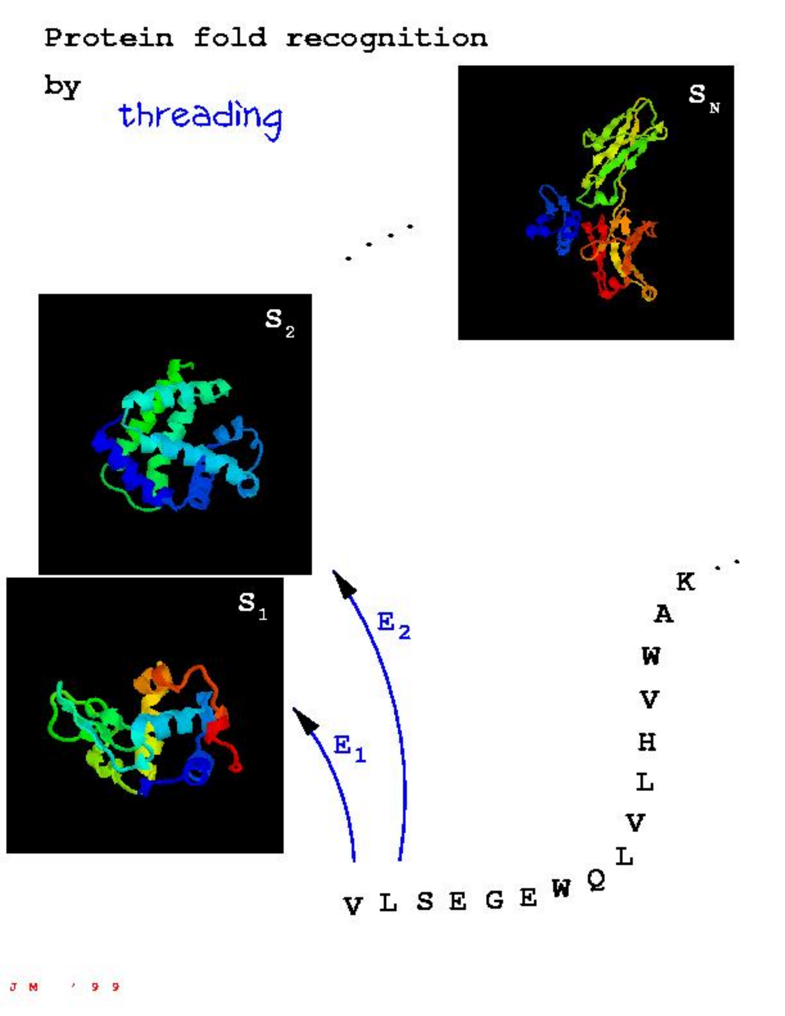
Sequence *folds into a 3D* structure *to perform a* function

Deciphering protein structure and function

- Experiment (X-ray, NMR): months

- Atomistic (physical principles based) simulations: weeks
- Homology based modeling: hours
- Sequence similarity based annotations: seconds

Assigning fold and function utilizing similarity to experimentally characterized proteins:



- Sequence similarity: BLAST and others
- Beyond sequence similarity: matching sequences and shapes (threading)

Importance of bioinformatics databases:

- DNA, mRNA, EST’s sequences, genes: GenBank → NCBI HomePage.htm
- Protein and nucleic acid structures: Protein Data Bank (PDB) → www.google.com
- Protein motifs: PROSITE
- Protein families: PFAM

Examples of further problems and methods:

- Microarray differential gene expression analysis → various clustering, pattern recognition and data mining algorithms → GeneSpring, J-express etc.
- Structural genomics and protein folding → global optimization methods
- Pattern searches → finite automata parsing, suffix trees → grep etc.
- Interactome and functional pathways analysis and prediction → chemical kinetics, graph theory etc.
- SNP’s, haplotypes and individual variation → statistical inference, correlations with disease states

Let us check out some recent papers …

- "Bioinformatics" is one of the major journals in the field:

  Bioinformatics -- Table of Contents (18 [4]).htm

And some links:

Bioinformatics Links.htm

Assignments:

- Devise an algorithm to align optimally two sequences over a random alphabet (e.g. A,T,G,C) with a pairwise score +1 for a match, 0 for a mismatch and –1 for a gap.
- Devise and algorithm to find efficiently if a given string contains a specified substring.
- What is the meaning of AUG and UGA codons? Devise a simple method for gene prediction in prokaryotic genomes.