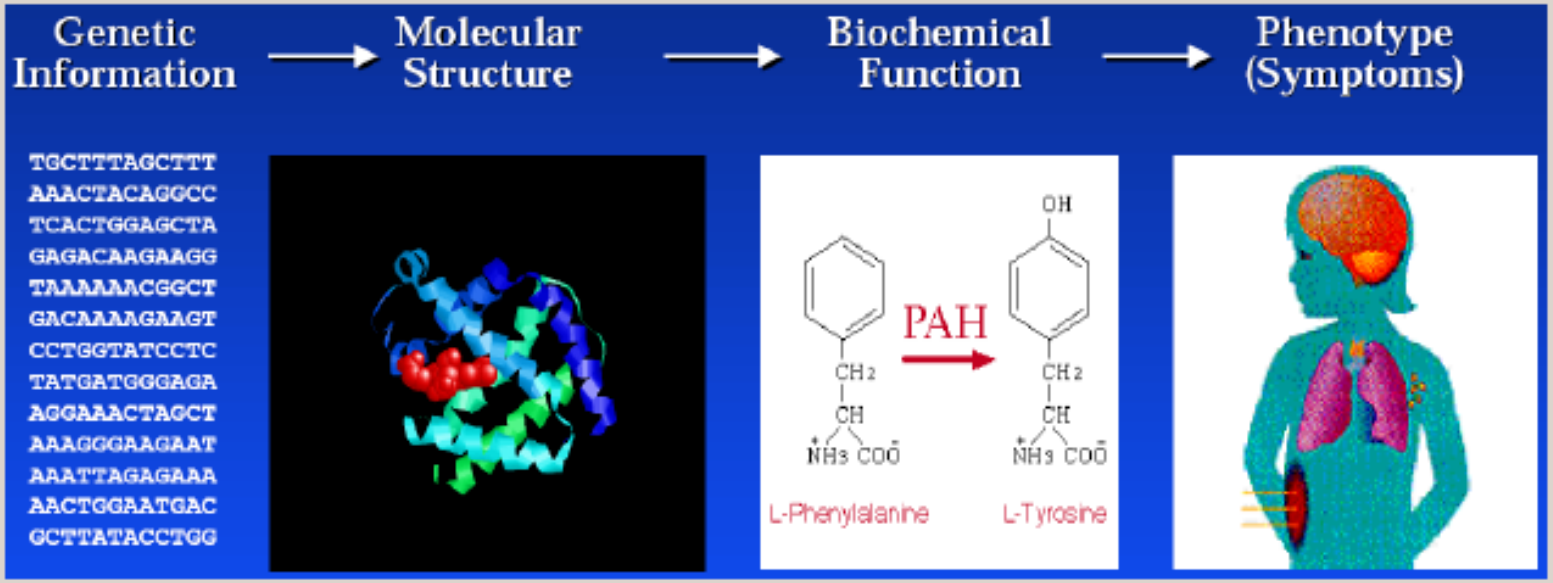
Why Bioinformatics?

Hammad Naveed
hammad.naveed@nu.edu.pk

Motivation

"Biology easily has 500 years of exciting problems to work on."
Donald Knuth (Stanford Professor & famous computer scientist)

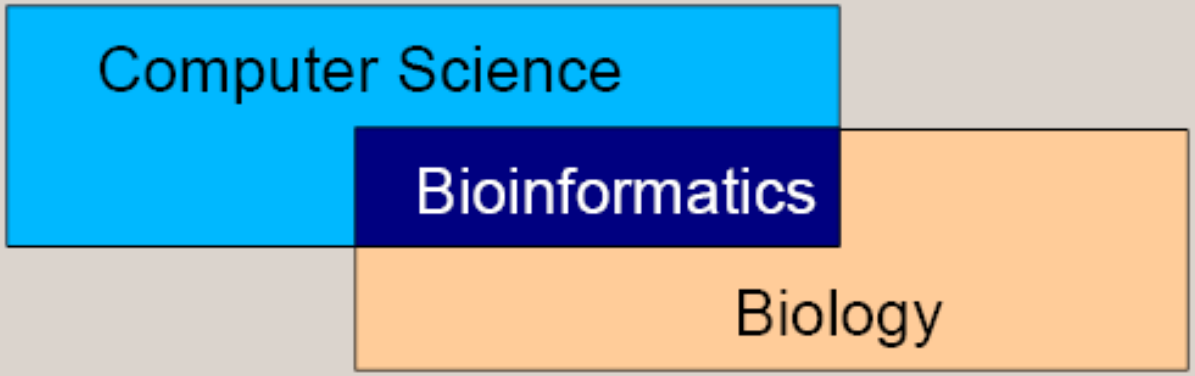


By developing techniques for analyzing sequence data and related structures, we can attempt to understand molecular basis of life.

http://cmgm.stanford.edu/biochem218/

Bioinformatics

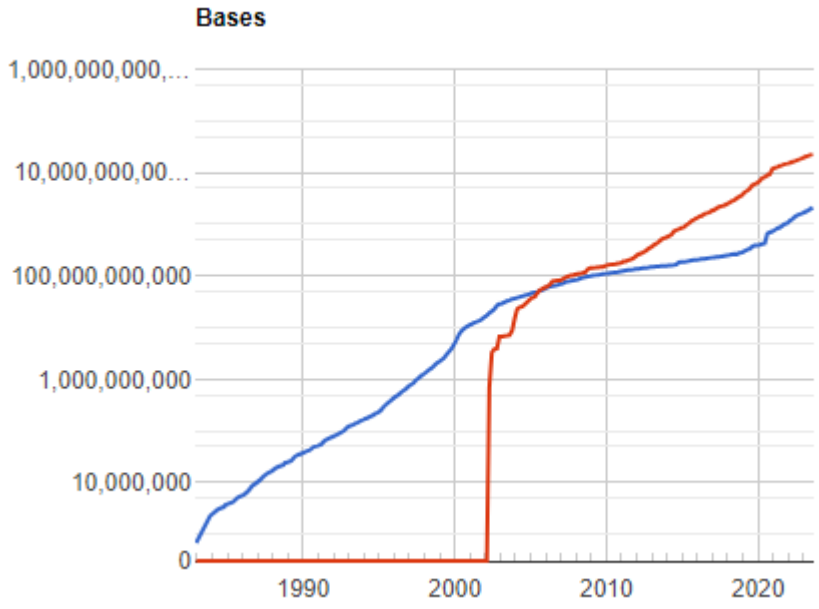
What is bioinformatics? *Application of techniques from computer science to problems from biology.*

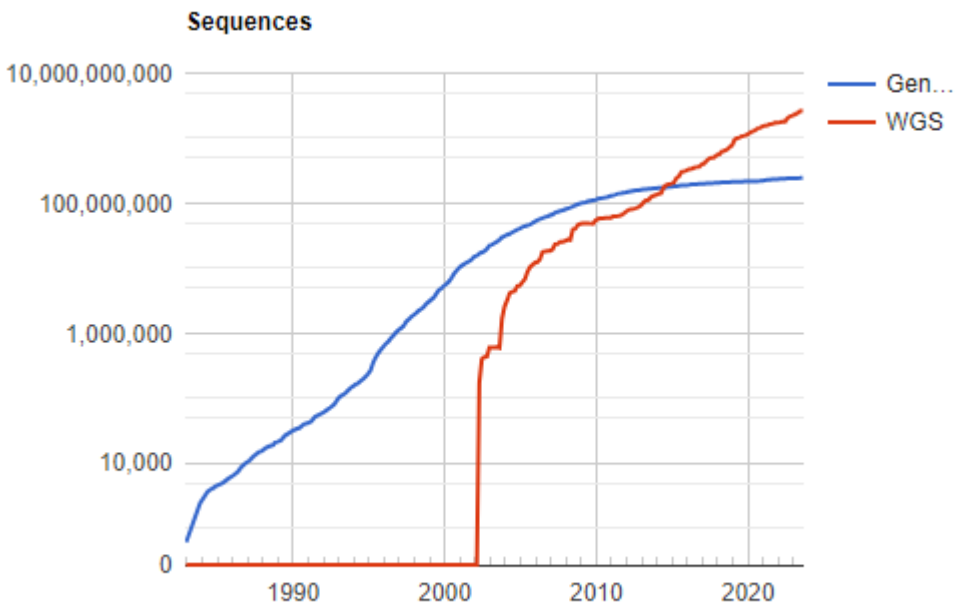


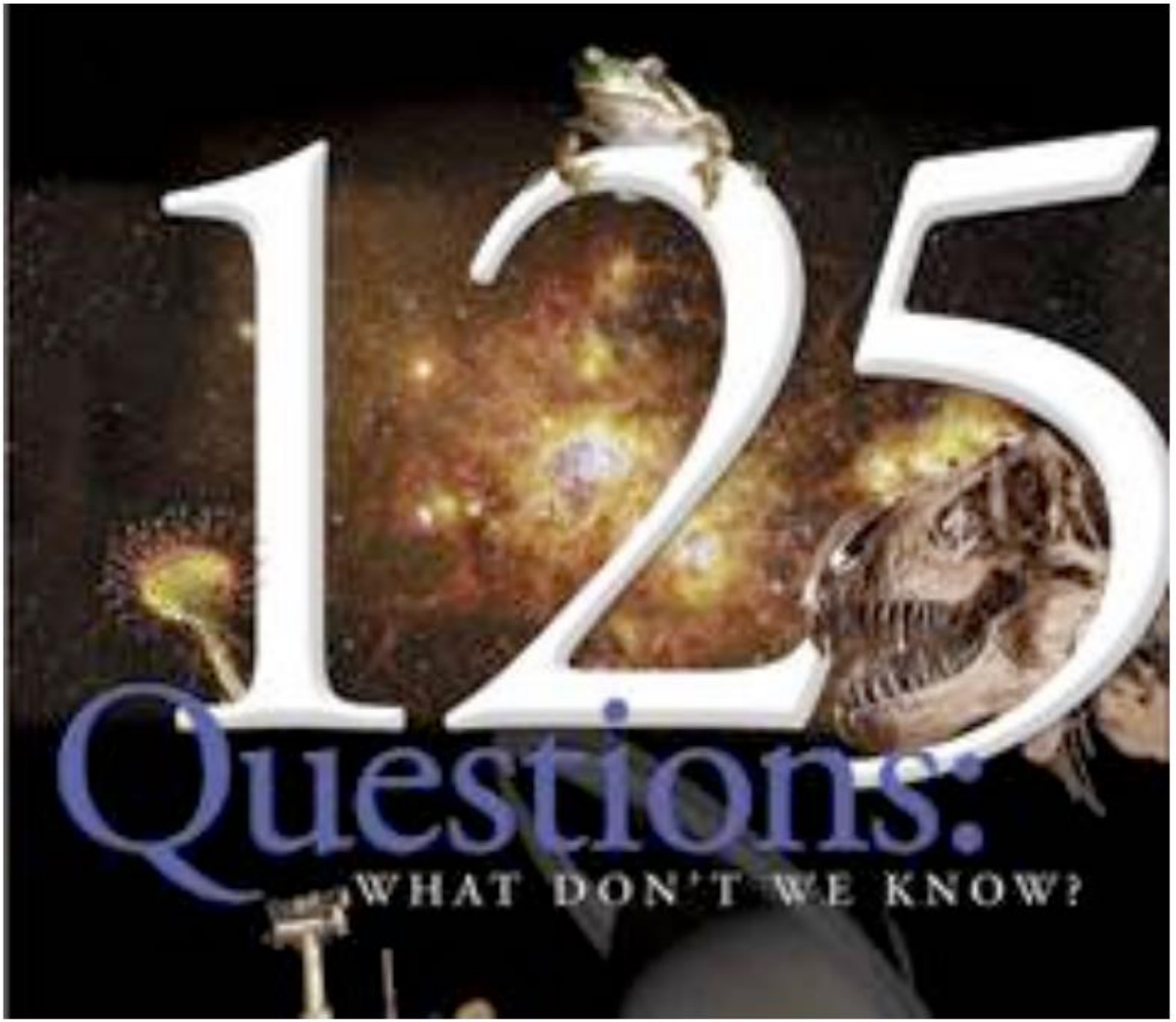
Why is it interesting?

- Important problems.
- Massive quantities of data.
- Desperate need for efficient solutions.
- Success is rewarded.

Data Explosion







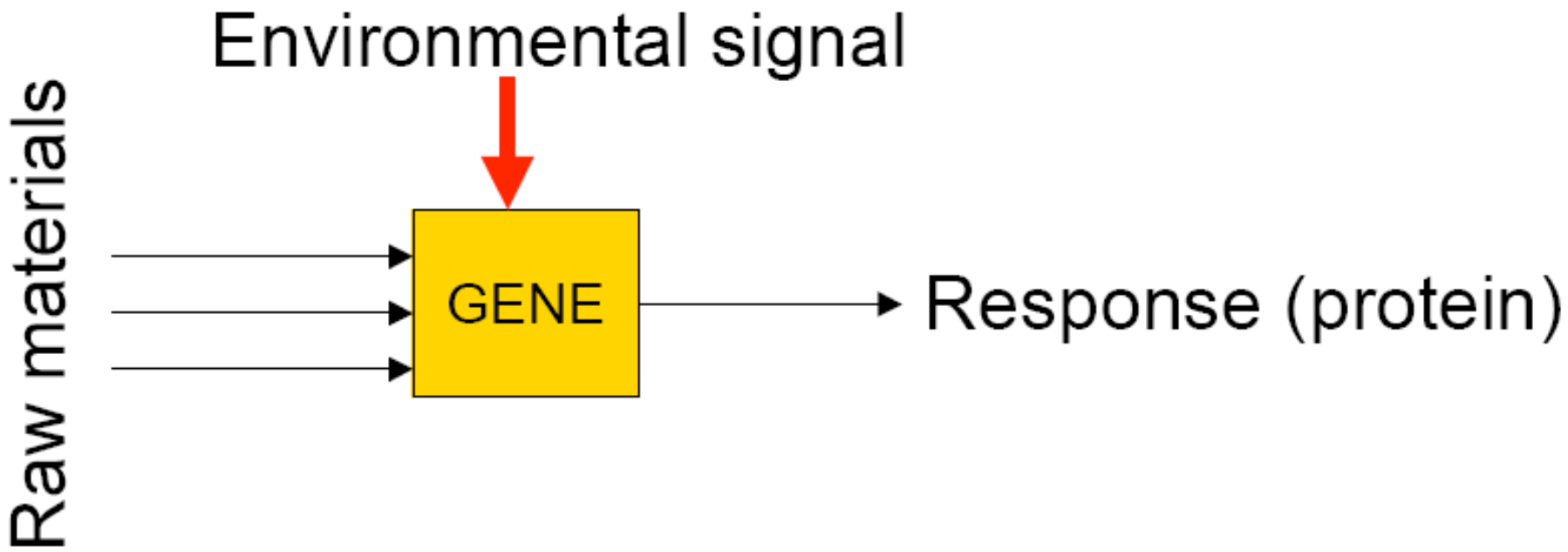
Special issue of journal *Science,* July 1, 2005.

>What Is the Universe Made Of?>What is the Biological Basis of Consciousness?>Why Do Humans Have So Few Genes?>To What Extent Are Genetic Variation and Personal Health Linked?>Can the Laws of Physics Be Unified?>How Much Can Human Life Span Be Extended?>What Controls Organ Regeneration?>How Can a Skin Cell Become a Nerve Cell?>How Does a Single Somatic Cell Become a Whole Plant?>How Does Earth's Interior Work?>Are We Alone in the Universe?>How and Where Did Life on Earth Arise?>What Determines Species Diversity?>What Genetic Changes Made Us Uniquely Human?>How Are Memories Stored and Retrieved?>How Did Cooperative Behavior Evolve?>How Will Big Pictures Emerge from a Sea of Biological Data?>How Far Can We Push Chemical Self-Assembly?>What Are the Limits of Conventional Computing?>Can We Selectively Shut Off Immune Responses?>Do Deeper Principles Underlie Quantum Uncertainty and Nonlocality?>Is an Effective HIV Vaccine Feasible?>How Hot Will the Greenhouse World Be?>What Can Replace Cheap Oil -- and When?>Will Malthus Continue to Be Wrong?

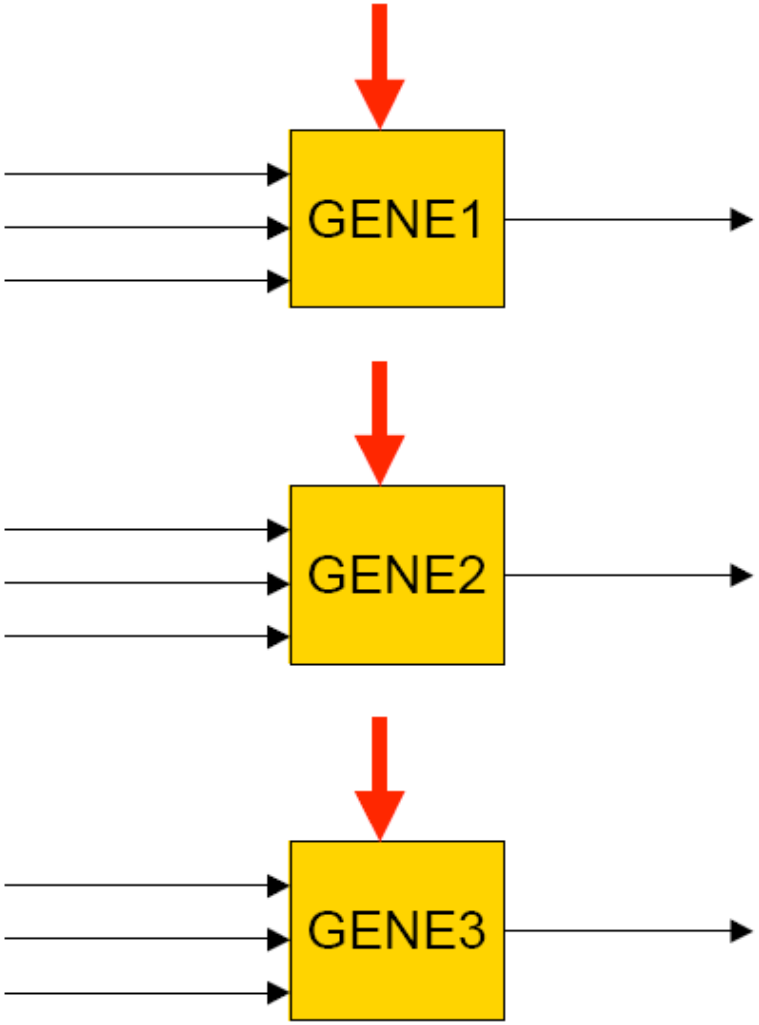
- Many of the most profound scientific questions of today are within the realm of bioinformatics research

- “Why do humans have so few genes ?”

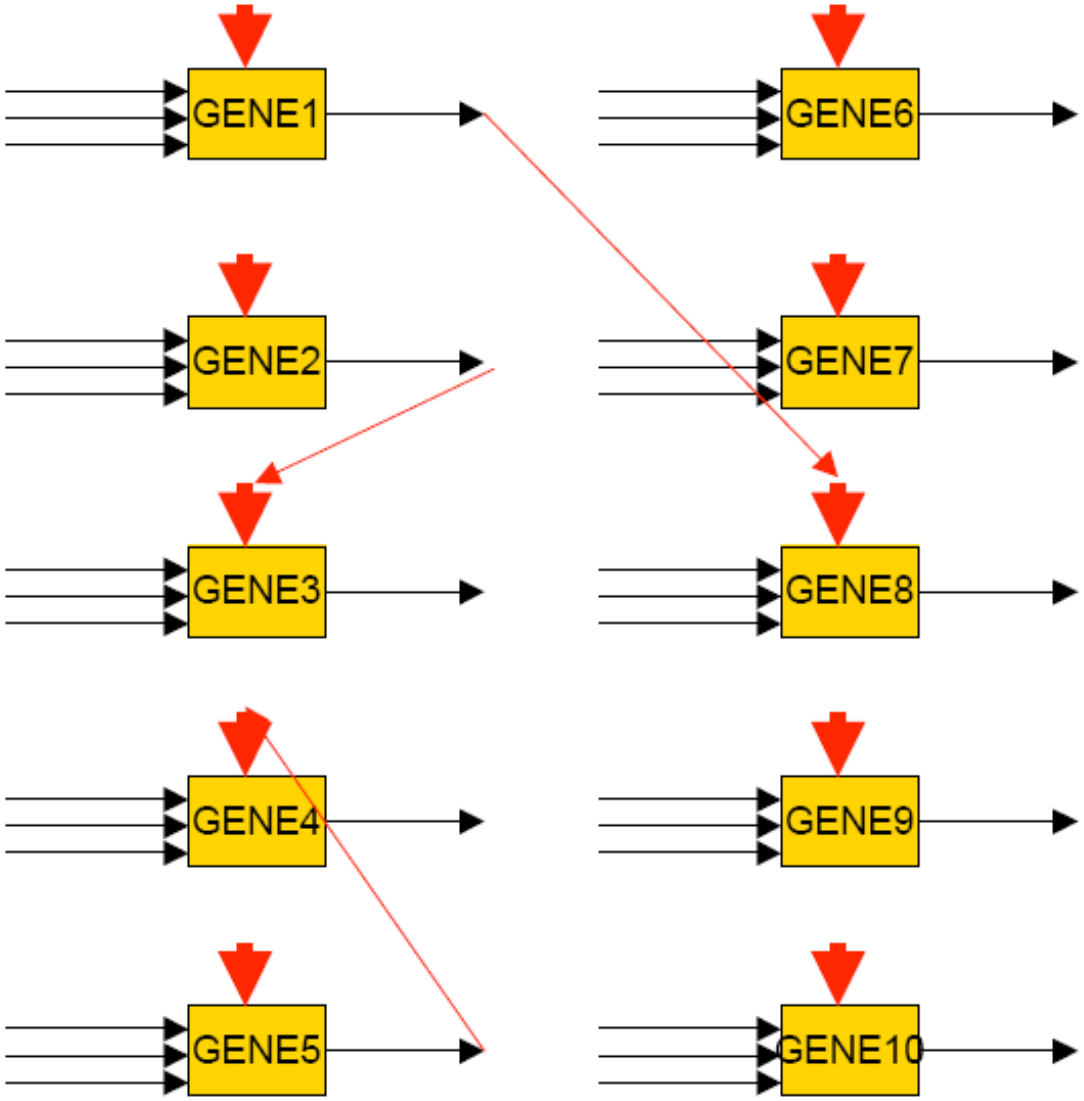
A simple organism



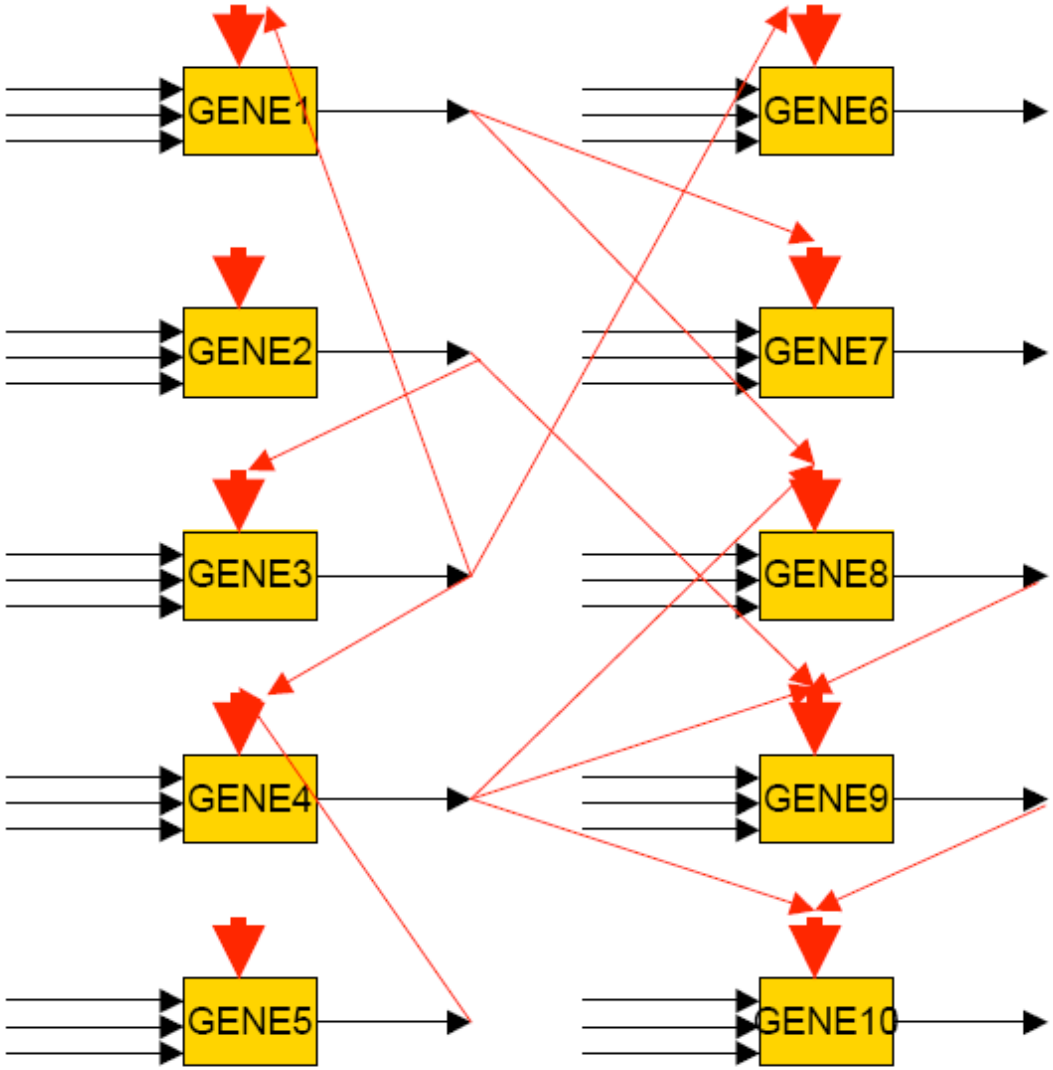
A simple organism



A simple organism



A complex organism



Regulatory networks

- This may be the reason why humans have so few genes (the circuit, not the number of switches, carries the complexity)

- Bioinformatics can unravel such networks, given the genome (DNA sequence) and gene activity information

Decoding the regulatory network

- Find patterns (“motifs”) in DNA sequence that occur more often than expected by chance
- An example computational problem:
  - Given a string of length 10,000 over the alphabet {A,C,G,T}
  - Count the number of occurrences N_w of every 6 letter word w

Decoding the regulatory network

- Are there specific words that occur more frequently than expected by chance?
- What is expected by chance?
- What is “more frequently”?
- Interesting mathematical questions

Comparing DNA

- Humans are about 99.9% identical to each other, DNA-wise.
- How do we know that ?
- Compare the genome of two individuals.
- The computational problem: Are two sequences similar?

Sequence alignment

- Why is this a problem?
- The two sequences will differ by “substitutions”, “insertions” and “deletions” accumulated during evolution
- The comparison algorithm has to be robust to such possibilities.
  - A special technique called “dynamic programming” does all this, and is “efficient”

Sequence alignment

- Why should we care?
- Compare human genome with fish. You'll see some portions that are highly similar.
- These "conserved" portions are often genes…
- … or regulatory sequences! The regulatory network again.

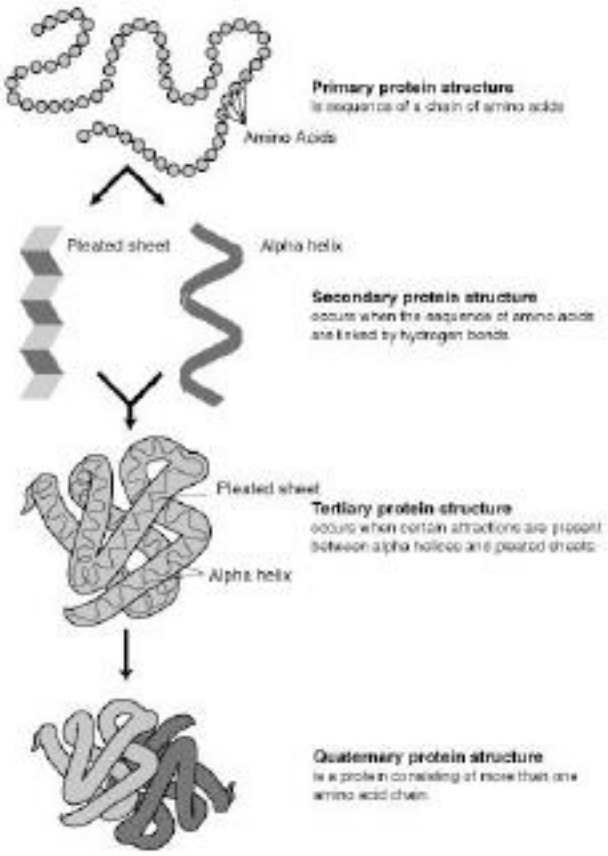
On counting genes

- The original question was “Why do humans have so few genes?”
- How do we know how many genes there are in the human genome ? (And where they are in the genome)
- Experiments can be designed, but bioinformatics plays a major role

Gene prediction

- The task of predicting the locations of genes in a new genome (“annotation”)
- Gene prediction software
- The more sophisticated ones use “Hidden Markov models” (HMM) and multiple species comparison

Protein structure prediction



Protein structure prediction

- Can we predict the 3D structure of a protein from its amino acid sequence?

- Why ?
  - One good reason: structure gives clues about function. If we can tell the structure, we can perhaps tell the function
  - We can design amino acid sequences that will fold into proteins that do what we want them to do. Drug design !!

Many more challenges

- New types of data come due to technological breakthroughs in biology
- High throughput data carries unprecedented amount of information
- Too much noise
- Bioinformatics removes the noise and reveals the truth

Bioinformatics

- Is not about one problem (e.g., designing better computer chips, better compilers, better graphics, better networks, better operating systems, etc.)
- Is about a family of very different problems, all related to biology, all related to each other
- How can computers help solve any of this family of problems ?

Bioinformatics and You

- You can learn and develop new tools of bioinformatics
- These tools owe their origin to computer science, information theory, probability theory, statistics, etc.
- You can learn the language of biology, enough to understand what the problems are
- Surprising biological problems convert to computer science problems naturally