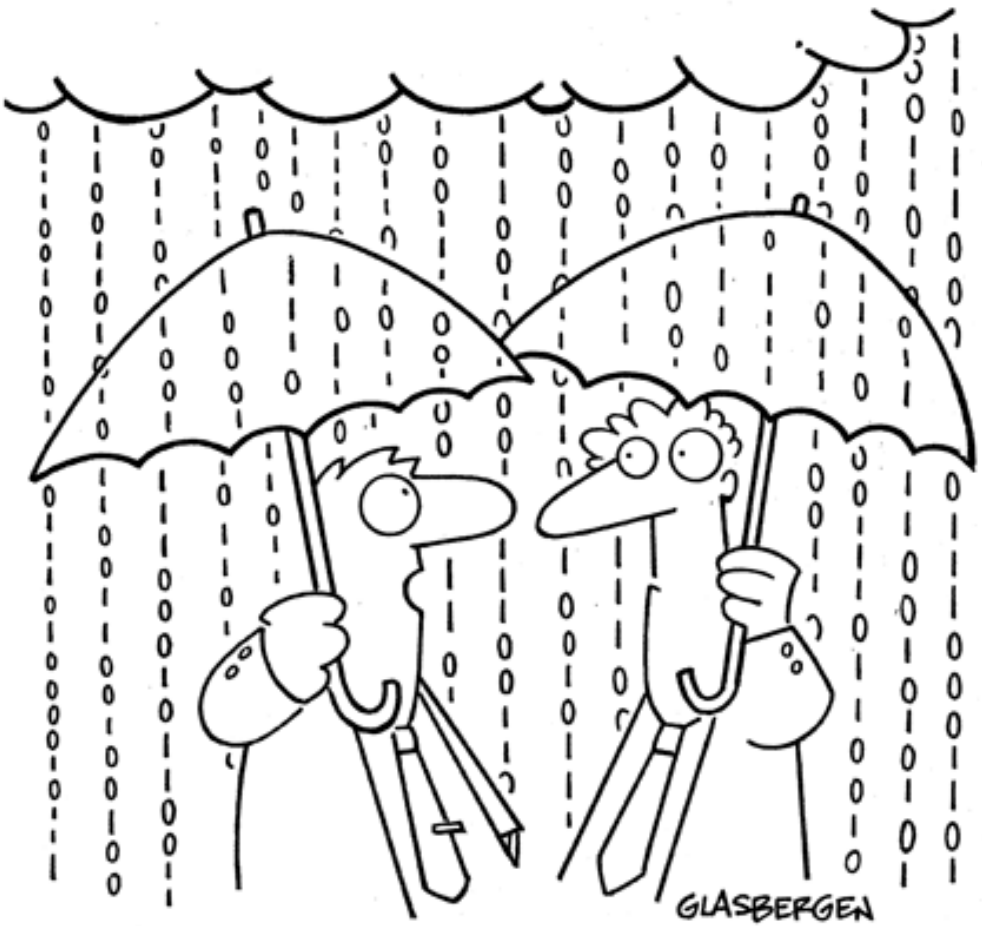
Biology & Big Data

Debasis Mitra

Professor, Computer Science, FIT

Cloud?



"I don't know much about cloud computing, but I think it might be responsible for the strange weather we're having."

Data as Service

- **Transparent to user**
- **Multiple locations**
- **Robustness**
- **Cost-based speed**

Software as Service

- **Software location transparent**
- **Pay-as-you-use**
- **Software running platform transparent**

Underlying middle-layer makes everything transparent

Platform as Service

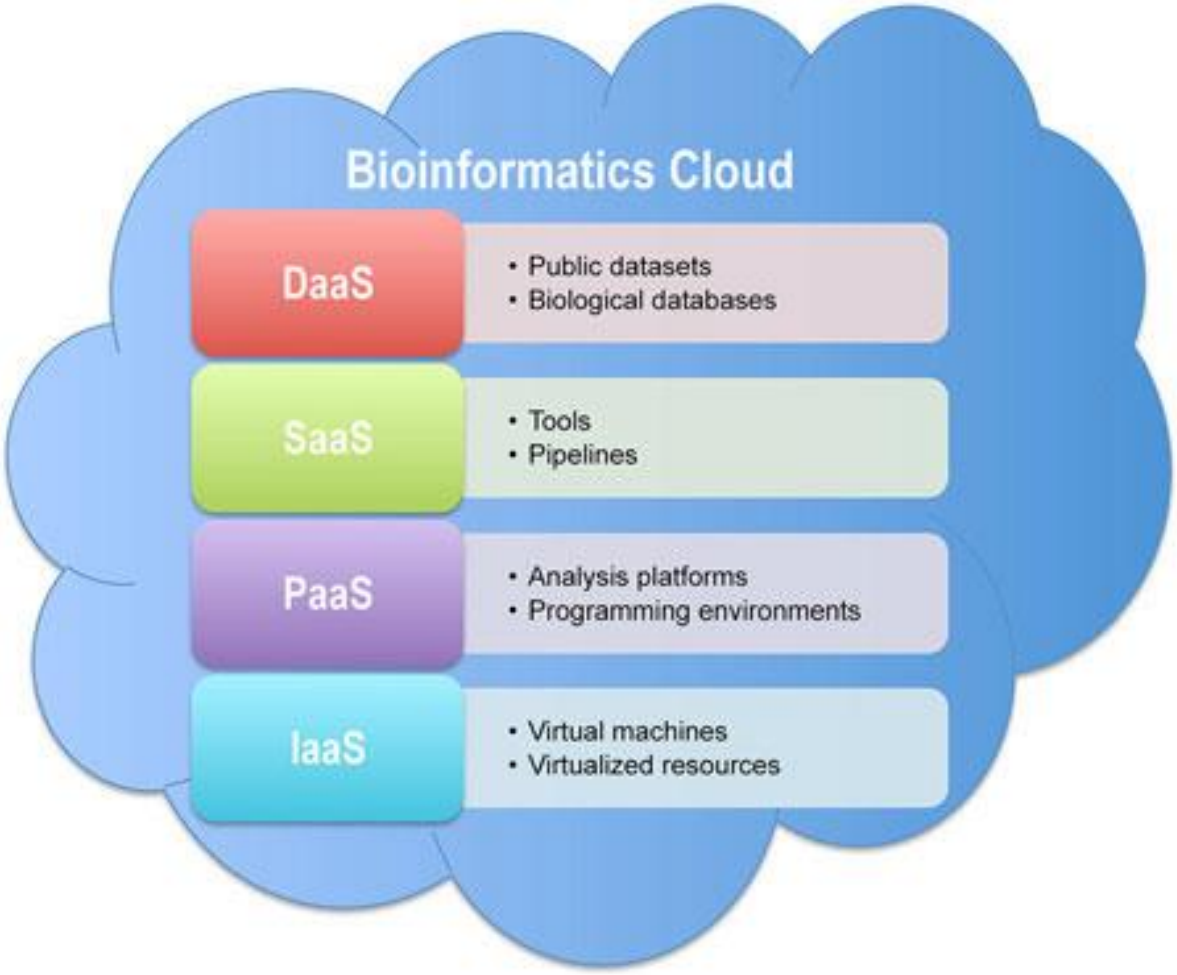
- **Middle-layer +**
- **Software environment +**
- **User interface**

Infrastructure as Service

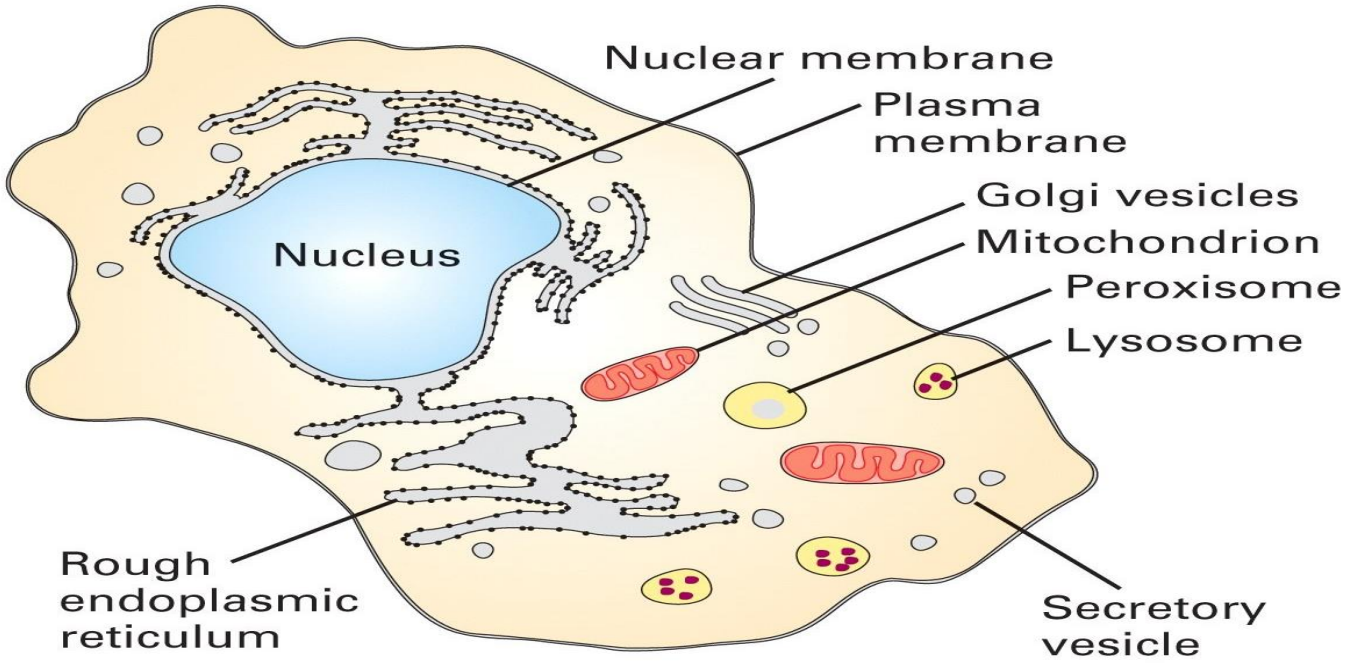
- **Hardware +**
- **Internet +**
- **Virtual Machine**

Source: Dai et al. Biology Direct 2012, 7:43 http://www.biology-direct.com/content/7/1/43

Bio-informatics Cloud / Grid

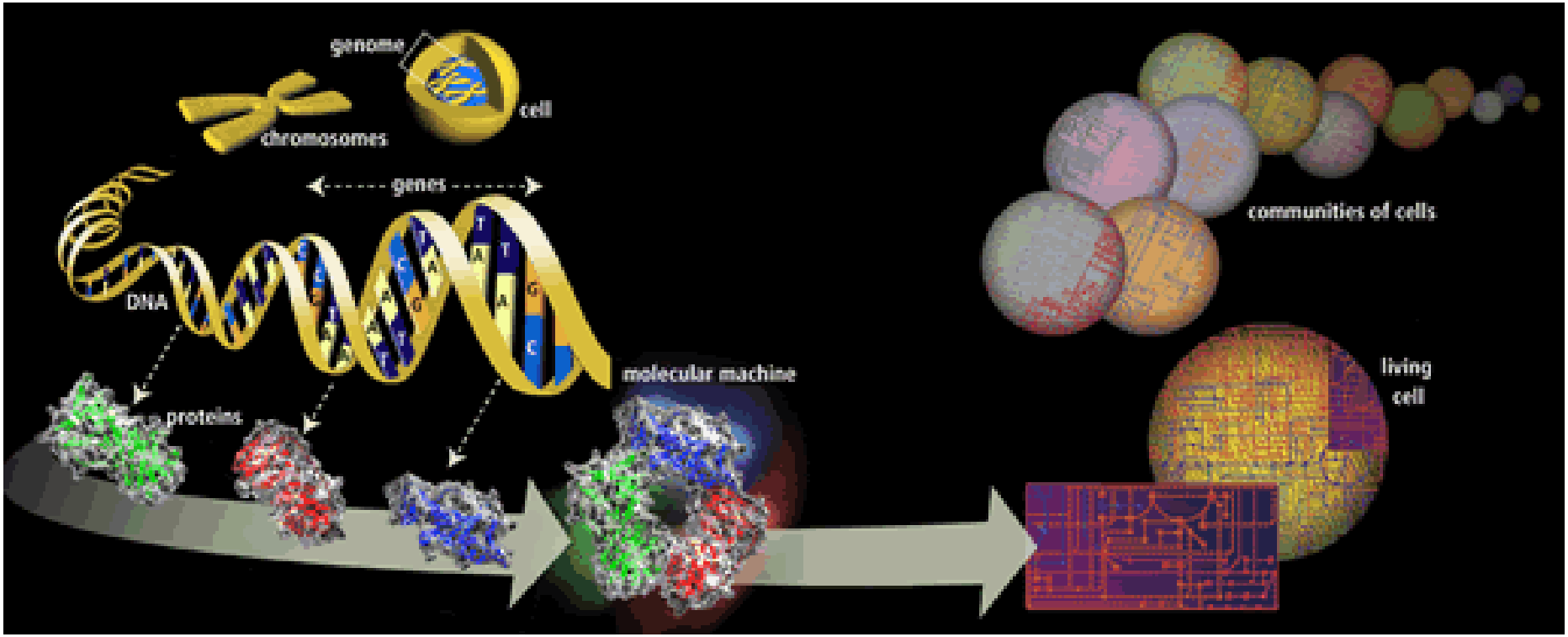


Life begins with Cell



- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

Molecular Biology

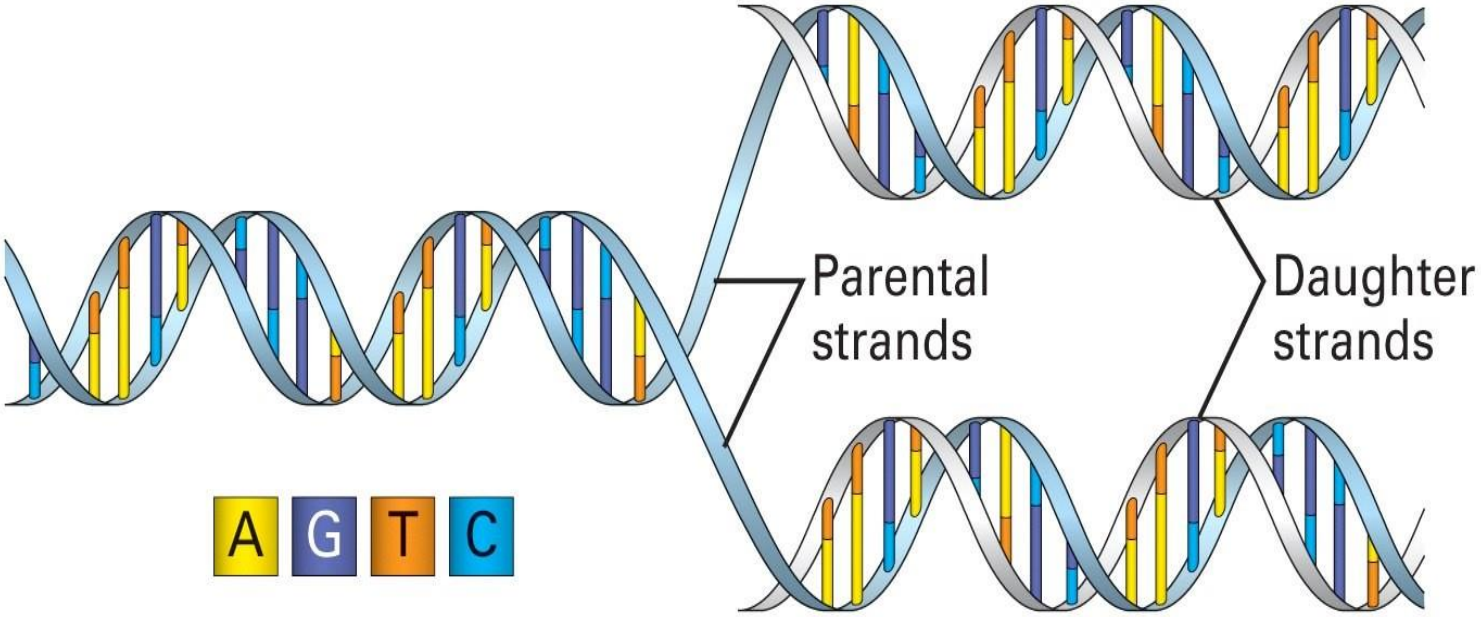


An Introduction to Bioinformatics Algorithms (Computational Molecular Biology)
Neil C. Jones*, *Pavel A. Pevzner

Organizations of life

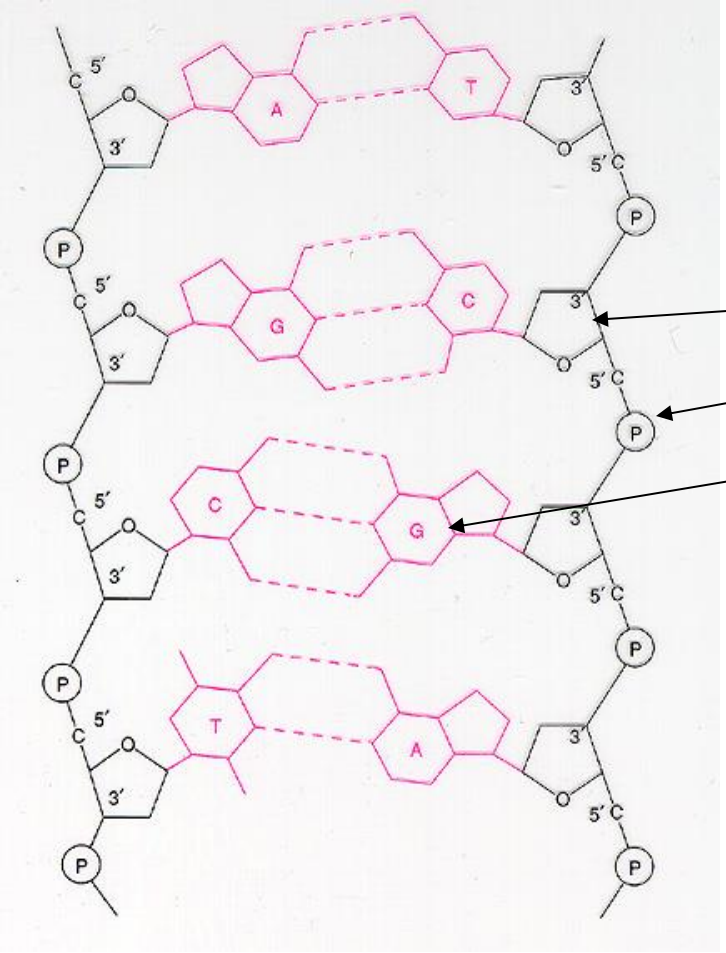
- **Nucleus** = **library**
- **Chromosomes** = **bookshelves**
- **Genes** = **books**
- Same libraries and the same sets of books for every cell in an organism
- Books represent all the information to carry out its various functions.

DNA: The Code of Life



- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

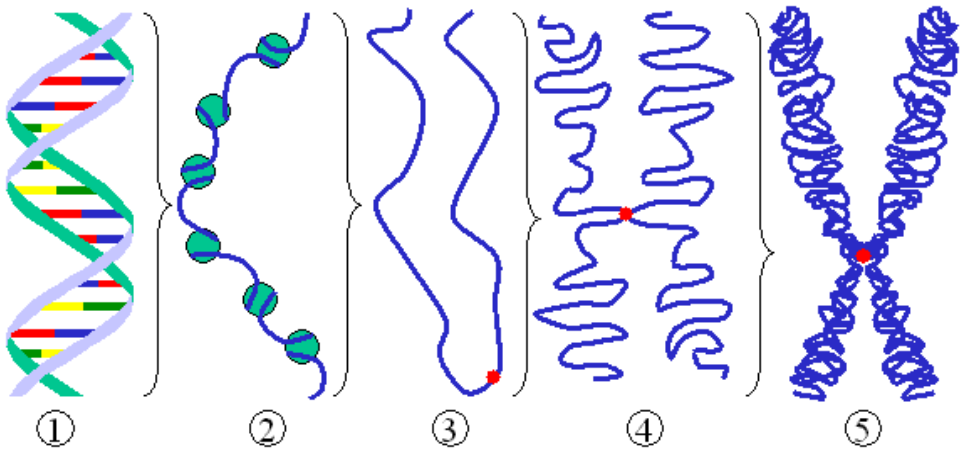
DNA, continued



- DNA has a double helix structure which is composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

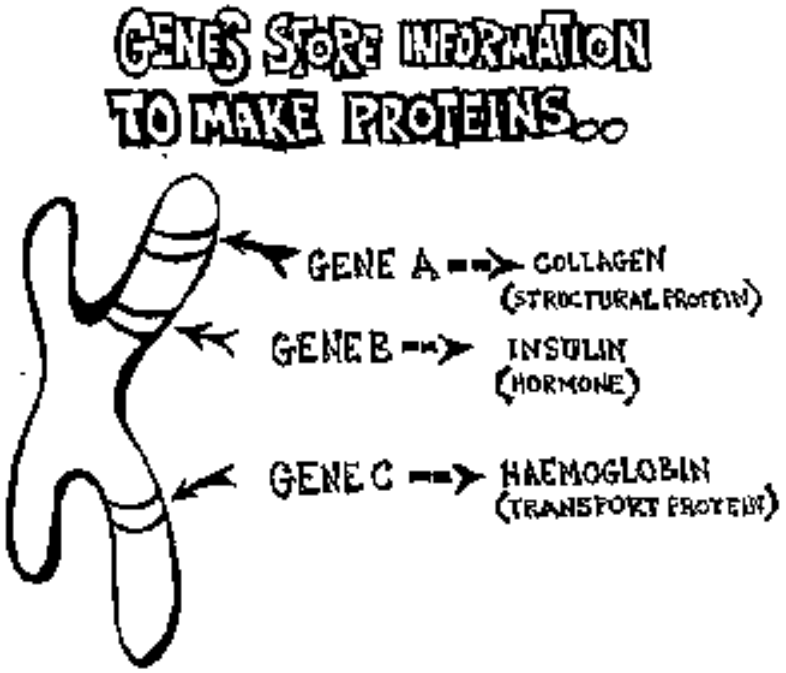
5’ ATTTAGGCC 3’
3’ TAAATCCGG 5’

Genetic Information: Chromosomes



Genes Make Proteins

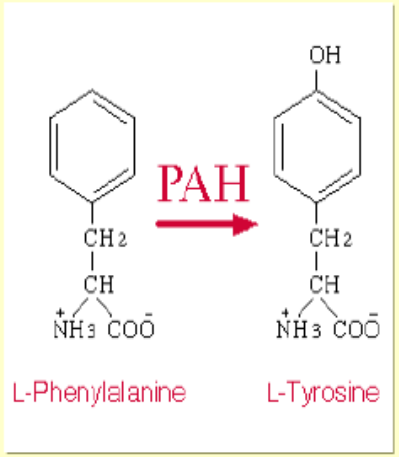
- genome-> genes ->protein(forms cellular structural & life functional)->pathways & physiology

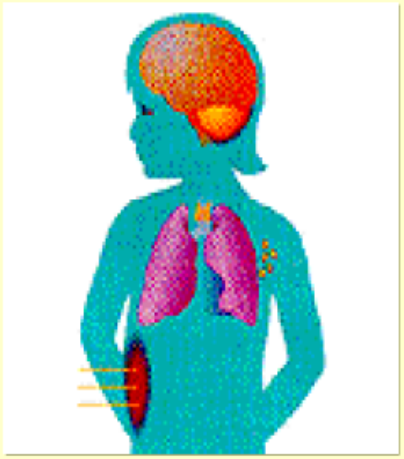


Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHLD
NLKGTFSQLS
ELHCDKLHVD
PENFRLLGNV
LVCVLARNFG
KEFTPQMQAA
YQKVVAGVAN
ALAHKYH





Some Terminology

- **Genome**: **an organism's genetic material**
  - ***Human genome is about 3,000,000,000 base-pair long***

- **Gene**: **a discrete units of hereditary information located on the chromosomes and consisting of DNA.**

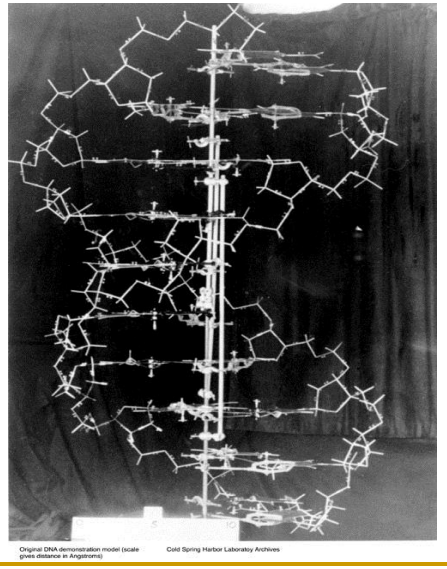
- **Nucleic acid**: **Biological molecules(RNA and DNA) that allow organisms to reproduce: A, T (U), C, G**

Major events in the history of Molecular Biology 1952 - 1960

- **1952-1953** James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA with four molecules A, T, C, G

James Watson and Francis Crick

- 1 Biologist
- + 1 Physics student
- + 900 words
- ----------
- = Nobel Prize



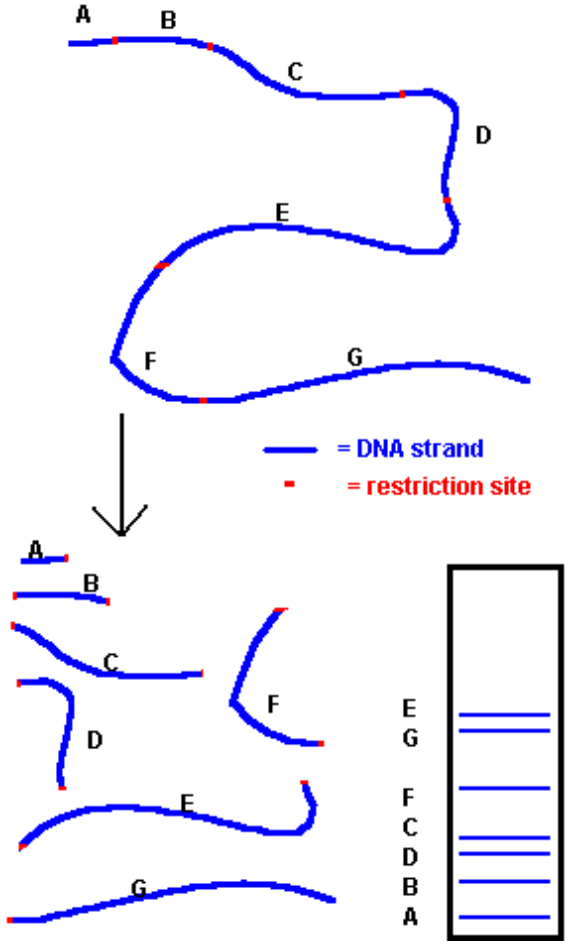
Original DNA demonstration model (scale gives distance in Angstroms) Cold Spring Harbor Laboratory Archives

Major events in the history of Molecular Biology 1970

- **1970** Howard Temin and David Baltimore independently isolate the first restriction enzyme

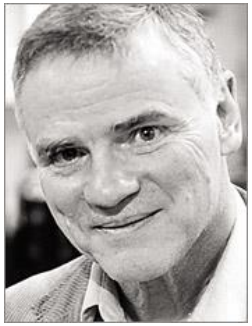
- DNA can be cut into reproducible pieces with restriction enzymes;

  (gene cloning or recombinant DNA technology)

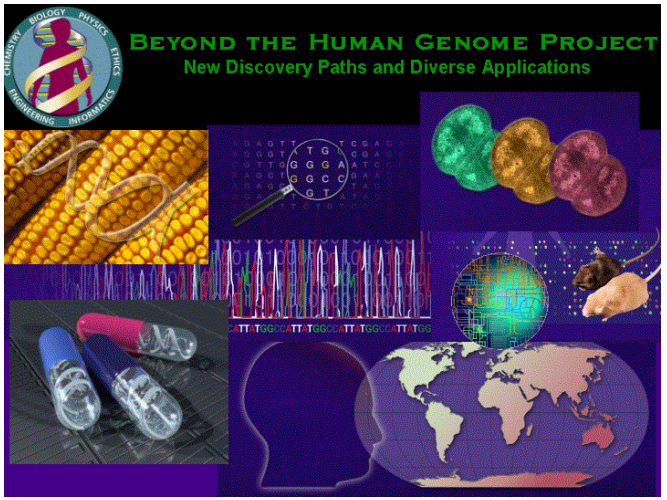


Major events in the history of Molecular Biology 1986 - 1995

- **1986** Leroy Hood: Developed automated sequencing mechanism
- **1986** Human Genome Initiative announced
- **1990** The 15 year Human Genome project is launched by congress
- **1995** Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published (These maps provide the locations of “markers” on each chromosome to make locating genes easier)



Leroy Hood



Major events in the history of Molecular Biology 1995-1996

- **1995** John Craig Venter: First bacterial genomes sequenced

John Craig Venter

- Challenged the genome sequencing project

  by developing '*shotgun*'

  approach –
- approach depends on assembling the sequences by **computer**

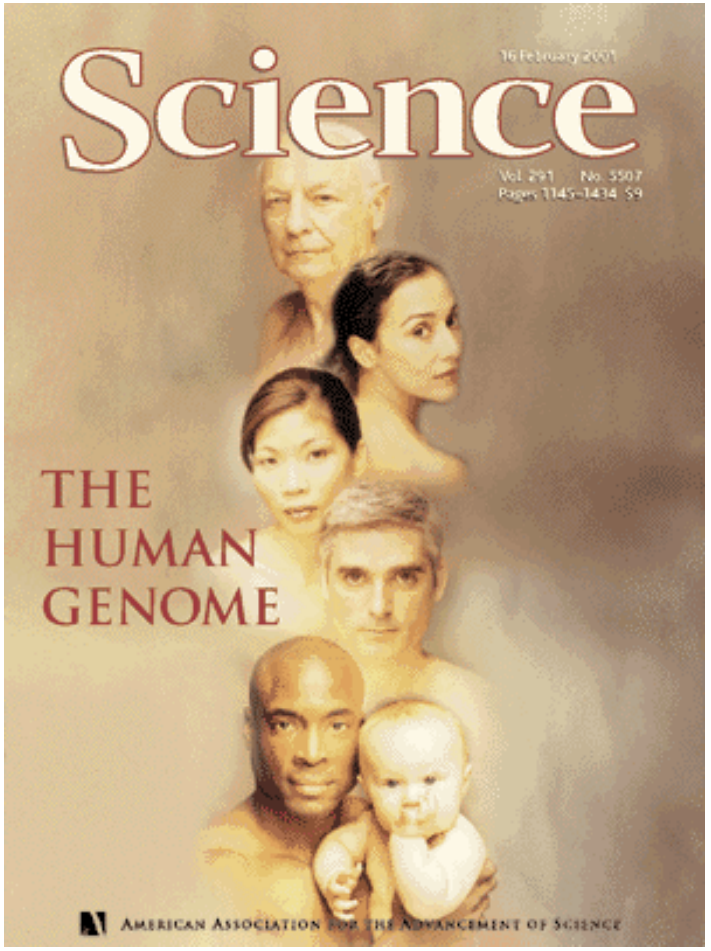
Major events in the history of Molecular Biology 1997 - 1999

- **1997** E. Coli sequenced
- **1998** PerkinsElmer, Inc.. Developed 96-capillary sequencer
- **1998** Complete sequence of the Caenorhabditis elegans genome
- **1999** First human chromosome (number 22) sequenced

Major events in the history of Molecular Biology 2000-2001

- **2000** Complete sequence of the euchromatic portion of the Drosophila melanogaster genome

- **2001** International Human Genome Sequencing:first draft of the sequence of the human genome published



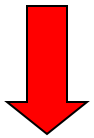
Major events in the history of Molecular Biology 2003- Present

- **April 2003** Human Genome Project Completed. Mouse genome is sequenced.
- **Jan 15, 2014, *Illumina*** Your genome may be sequenced < $1,000
- **IBM Challenge**: <$100

Why sequence?
MUtAsHONS

- What happens to genes when the DNA is mutated?

Normal DNA sequence: ATCTAG



Mutated DNA sequence: ATCGAG

The Good, the Bad, and the Silent

- Mutations can serve the organism in three ways:

A mutation can cause a trait that enhances the organism's function:

- The Good : Mutation in the sickle cell gene provides resistance to malaria.

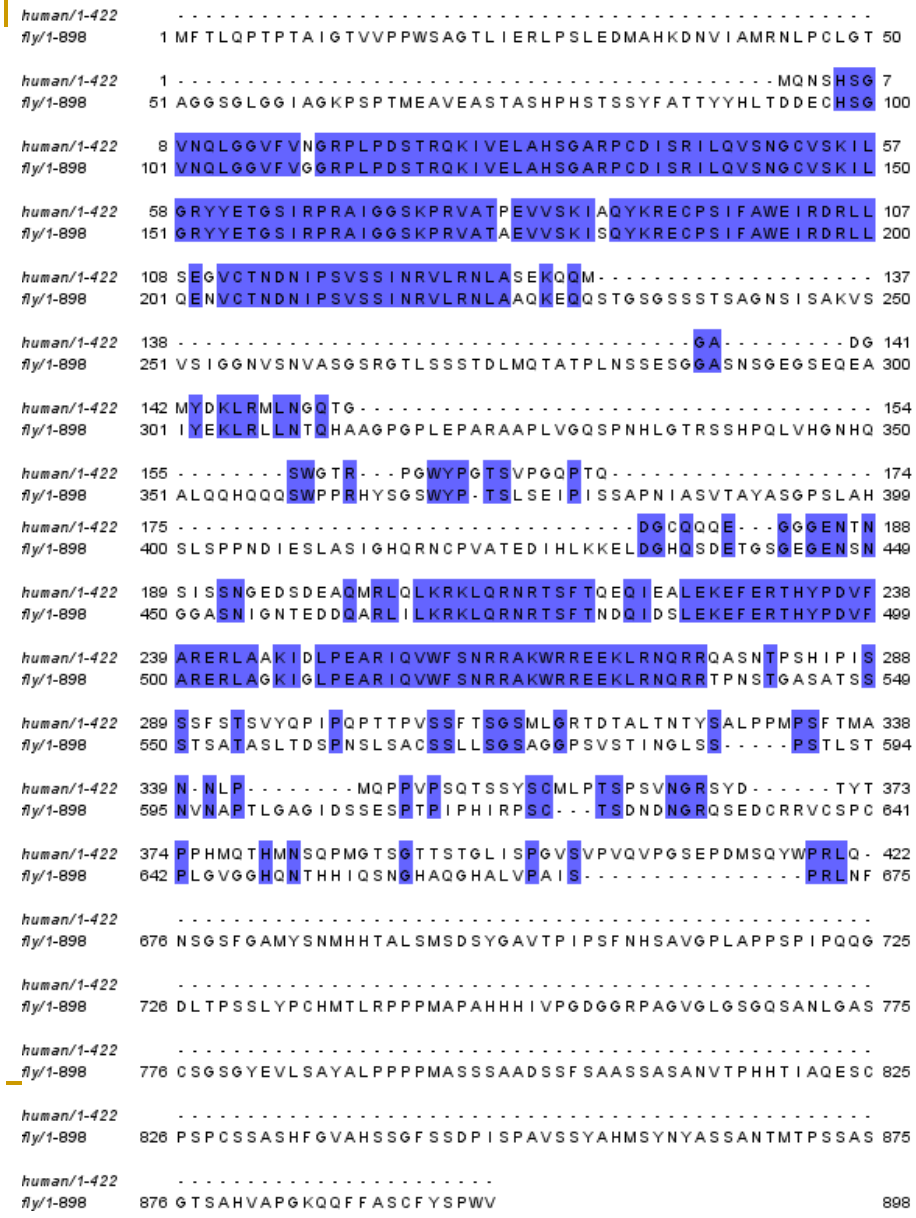
- The Bad : A mutation can cause a trait that is harmful, sometimes fatal to the organism:

Huntington's disease, a symptom of a gene mutation, is a degenerative disease of the nervous system.

- The Silent: A mutation can simply cause no difference in the function of the organism.

Campbell, Biology, 5th edition, p. 255

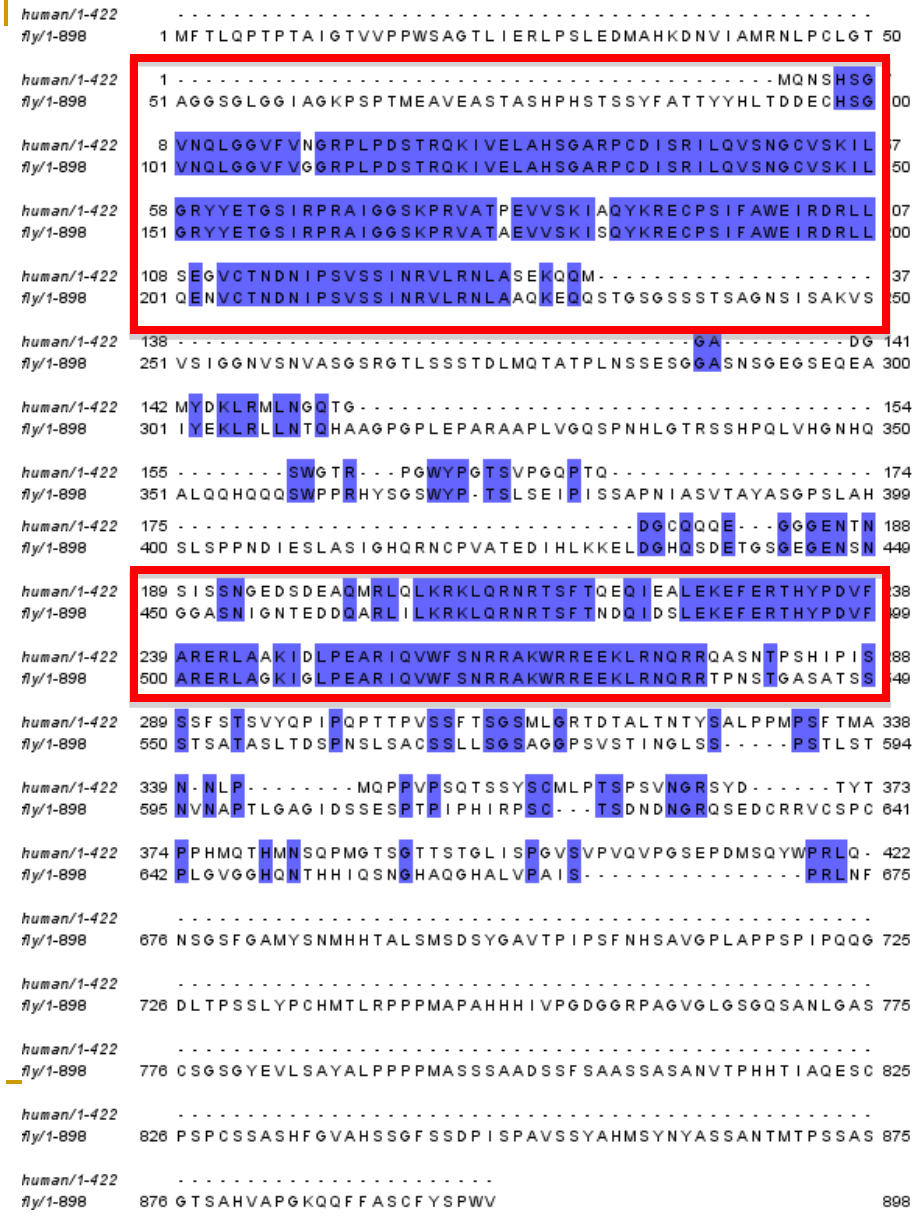
Real data: human & fruitfly eyeless



- **This is a global alignment of human & fruitfly Eyeless-gene**

Next few slides are from
Dr Avril Coghlan, Sanger Institute, UK

Real data: human & fruitfly eyeless

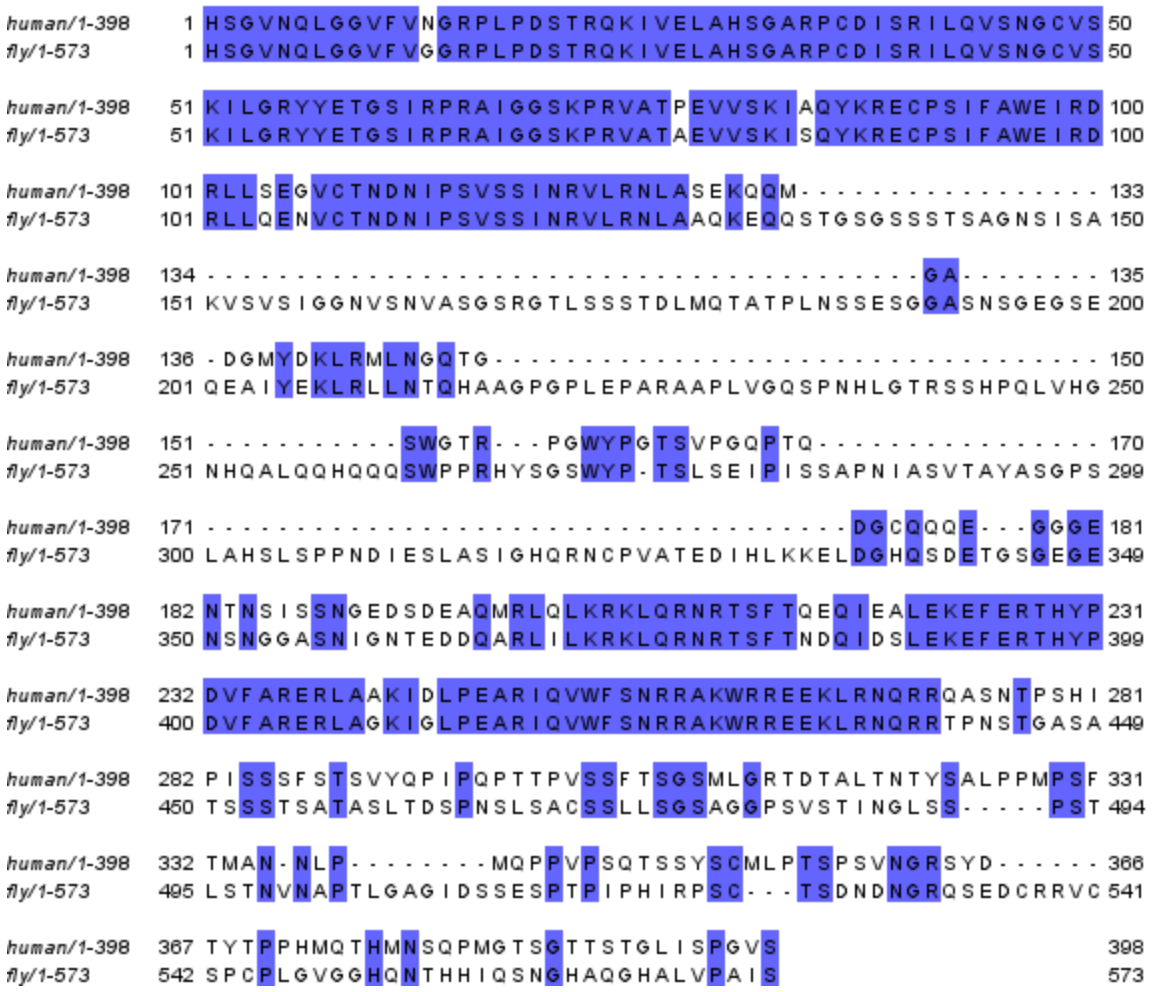


There are 2 short regions of high similarity

Outside those regions, there are many mismatches and gaps

It might be more sensible to make local alignments of one or both of the regions of high similarity

Real data: human & fruitfly



This is a local alignment of human & fruitfly

What parts of the sequences were used in the local alignment?

The Smith-Waterman algorithm

- **local alignment** of 2 sequences

The **alignment of all possible subsequences (parts) of sequences S_1 and S_2**

The 0^{th} row and 0^{th} column of *T* are first filled with zeroes

The recurrence relation used to fill table *T* is:

$$T(i, j) = \max \begin{cases} T(i\text{-}1, j\text{-}1) + \sigma(S_1(i), S_2(j)) \\ T(i\text{-}1, j) + \text{gap penalty} \\ T(i, j\text{-}1) + \text{gap penalty} \\ 0 \end{cases}$$

The traceback starts at the **highest scoring cell** in the matrix *T*, and travels up/left **while the score is still positive**

- eg., to find the best local alignment of sequences “ACCTAAGG” and “GGCTCAATCA”, using +2 for a match, -1 for a mismatch, and -2 for a gap:

We first make matrix *T* (as in N-W):

The 0th row and 0th column of *T* are filled with zeroes

The recurrence relation is then used to fill the matrix *T*

		G	G	C	T	C	A	A	T	C	A
	0	0	0	0	0	0	0	0	0	0	0
A	0										
C	0										
C	0										
T	0										
A	0										
A	0										
G	0										
G	0										

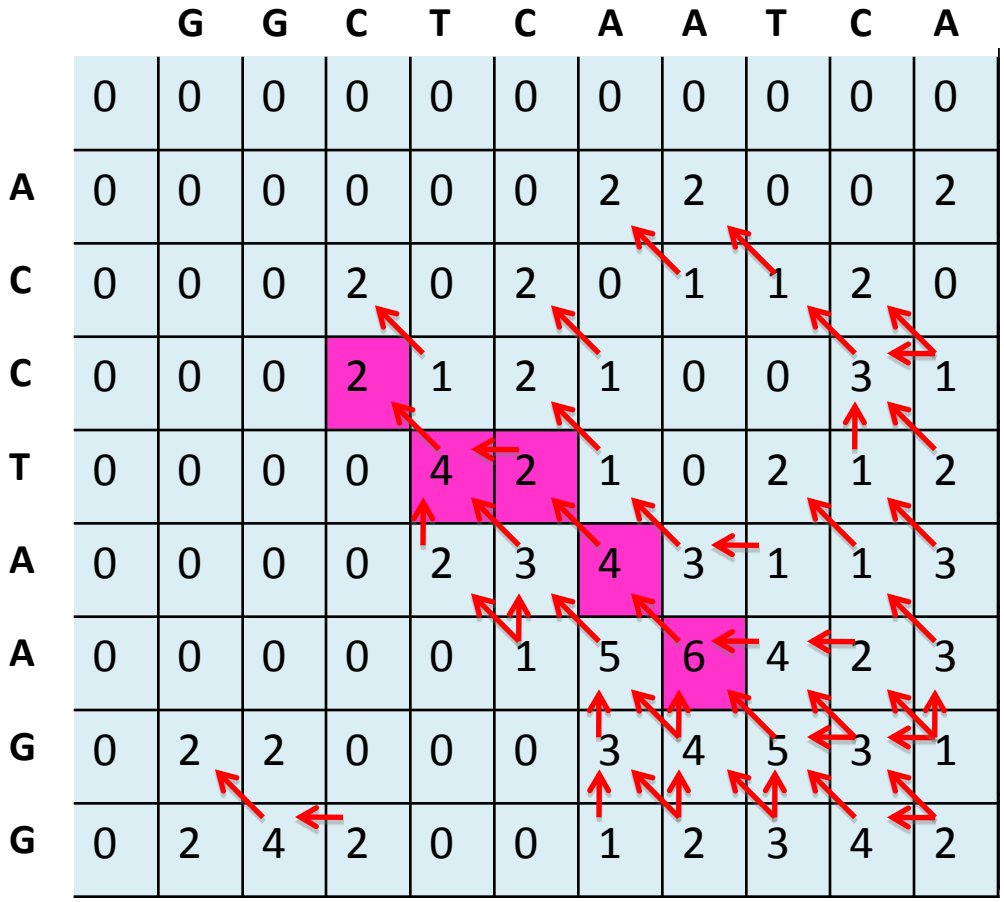
We first calculate $T(1,1)$ using the recurrence relation:

$$T(i,j) = \max \begin{cases} T(i\text{-}1, j\text{-}1) + \sigma(S_1(i), S_2(j)) = \mathbf{0 - 1 = -1} \\ T(i\text{-}1, j) + \text{gap penalty} = \mathbf{0 - 2 = -2} \\ T(i, j\text{-}1) + \text{gap penalty} = \mathbf{0 - 2 = -2} \\ \mathbf{0} \end{cases}$$

The maximum value is 0, so we set $T(1,1)$ to 0

		G	G	C	T	C	A	A	T	C	A
	0	0	0	0	0	0	0	0	0	0	0
A	0	0	?								
C	0										
C	0										
T	0										
A	0										
A	0										
G	0										
G	0										

You fill in the whole of *T*, recording the previous cell (if any) used to calculate the value of each *T*(*i*, *j*):



		G	G	C	T	C	A	A	T	C	A
	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	2	2	0	0	2
C	0	0	0	2	0	2	0	1	1	2	0
C	0	0	0	2	1	2	1	0	0	3	1
T	0	0	0	0	4	2	1	0	2	1	2
A	0	0	0	0	2	3	4	3	1	1	3
A	0	0	0	0	0	1	5	6	4	2	3
G	0	2	2	0	0	0	3	4	5	3	1
G	0	2	4	2	0	0	1	2	3	4	2

The traceback starts at the **highest scoring cell** in the matrix *T*, and travels up/left **while the score is still positive**

		G	G	C	T	C	A	A	T	C	A
	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	2	2	0	0	2
C	0	0	0	2	0	2	0	1	1	2	0
C	0	0	0	2	1	2	1	0	0	3	1
T	0	0	0	0	4	2	1	0	2	1	2
A	0	0	0	0	2	3	4	3	1	1	3
A	0	0	0	0	0	1	5	6	4	2	3
G	0	2	2	0	0	0	3	4	5	3	1
G	0	2	4	2	0	0	1	2	3	4	2

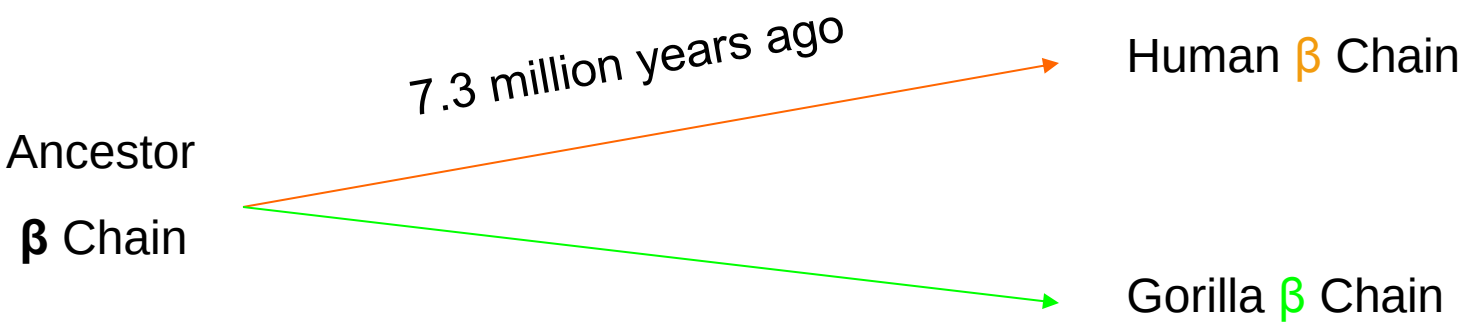
Work out the best local alignment from the traceback:

```
C T C A A
| |   | |
C T - A A
```

Score of the alignment is in the **bottom right cell of the traceback** (6 = 4 × (score of 2 per match) + 1 × (-2 per gap))

Molecular Clock

- Linus and Pauling found that α-chains of human and gorilla differ by 2 residues, and β-chains by 1 residues.
- They then calculated the time of divergence between human and gorilla using evolutionary molecular clock.
- Gorilla and human β chain were found to diverge about 7.3 years ago.



Beta globins:

- Beta globin chains of closely related species are highly similar:
- Observe simple alignments below:

Human β chain: MVHLT**PE**EK**S**AV**TA**LWGKV N**V**D**E**VGGEALGRLL

Mouse β chain: MVHLT**DA**EK**A**AV**NG**LWGKVN**P**D**D**VGGEALGRLL

Human β chain: VVYPWTQR**F**F**E**SFGDLS**TPD**A**V**MGNPKVKAHGKKV**LG**

Mouse β chain: VVYPWTQR**Y**F**D**SFGDLS**SAS**A**I**MGNPKVKAHGKK V**IN**

Human β chain: AF**S**DGL**A**HLDNLKGTFA**T**LSELHCDKLHVDPENFRLLGN

Mouse β chain: AF**N**DGL**K**HLDNLKGTFA**H**LSELHCDKLHVDPENFRLLGN

Human β chain: **VL**V**C**VL**A**HH**F**GKEFTP**PV**QAA**Y**QKVVAGVA**N**ALAHKYH

Mouse β chain: **MI** V**I** VL**G**HH**L**GKEFTP**CA**QAA**F**QKVVAGVA**S**ALAHKYH

There are a total of **27** mismatches, or (147 – 27) / 147 = **81.7 %** identical

Beta globins: Cont.

Human β chain: MVH L TPEEKSAVTALWGKVNVDEVGGEALGRLL

Chicken β chain: MVHWTAEEKQL I TGLWGKVNVAECGAEALARLL

Human β chain: VVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG

Chicken β chain: IVYPWTQRFF ASFGNLSSPTA I LGNPMVRAHGKKVLT

Human β chain: AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGN

Chicken β chain: SFGDAVKNLDNIK NTFSQLSELHCDKLHVDPENFRLLGD

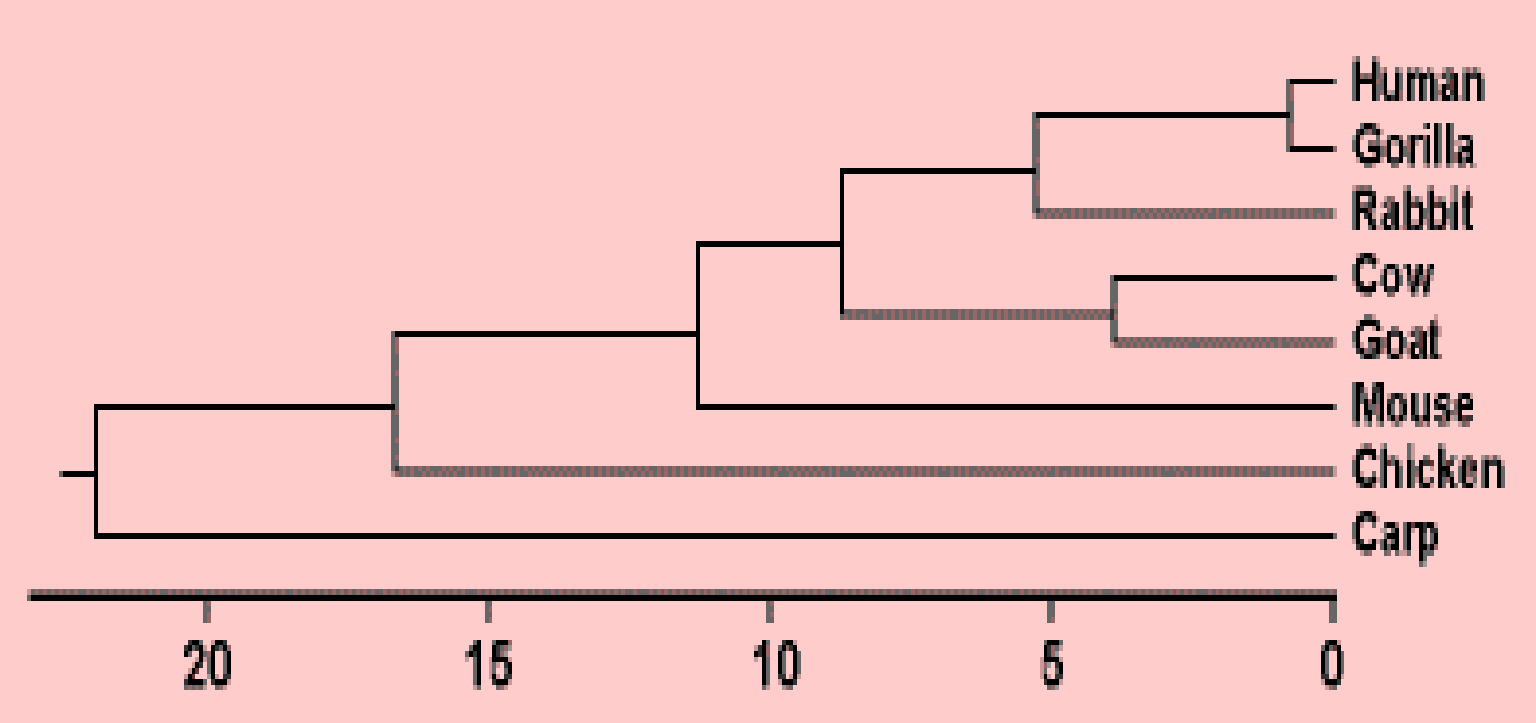
Human β chain: VLVCVLAHHFGKEFTPPVQAAY QKVVAGVANALAHKYH

Mouse β chain: I L I I VLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH

-There are a total of **44** mismatches, or (147 – 44) / 147 = **70.1 %** identical

- As expected, mouse β chain is '*closer*' to that of human than chicken's.

Molecular evolution can be visualized with phylogenetic tree.



Phylogenetic tree of Beta globin (Aligned using Clustal, PAM250)

Mouse and Human overview

- Mouse has 2.1 x10^9 base pairs versus 2.9 x 10^9 in human.
- About 95% of genetic material is shared.
- 99% of genes shared of about 30,000 total.
- The 300 genes that have no homologue in either species deal largely with immunity, detoxification, smell and sex*

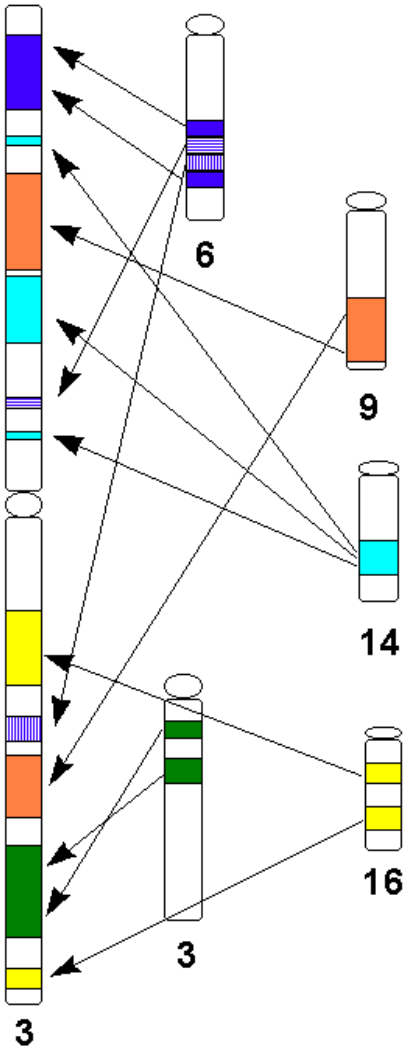
*Scientific American Dec. 5, 2002

Human and Mouse

Significant chromosomal rearranging occurred between the diverging point of humans and mice.

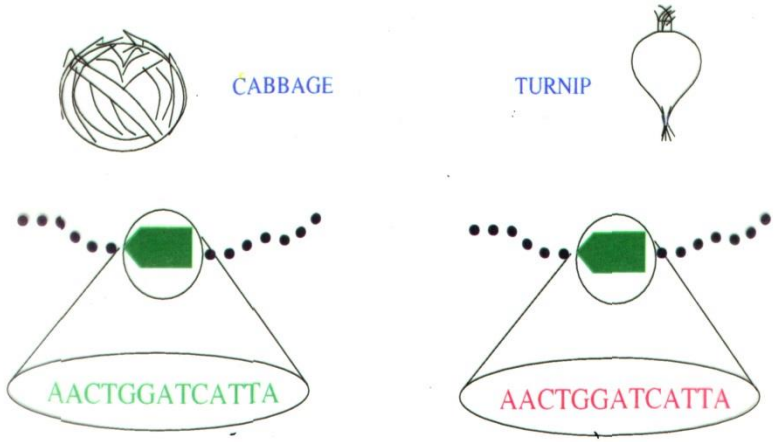
Here is a mapping of human chromosome 3.

It contains homologous sequences to at least 5 mouse chromosomes.



Important discovery

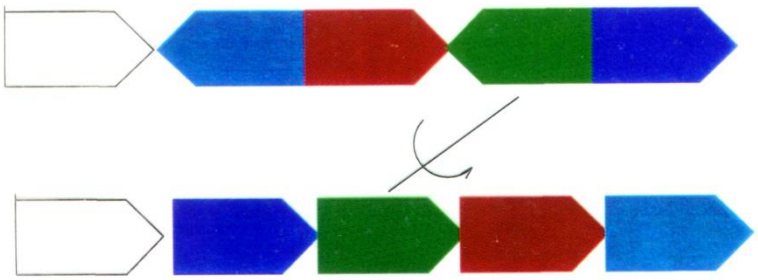
GENE SEQUENCE COMPARISON



AACTGGATCATTA
AACTGGATCATTA

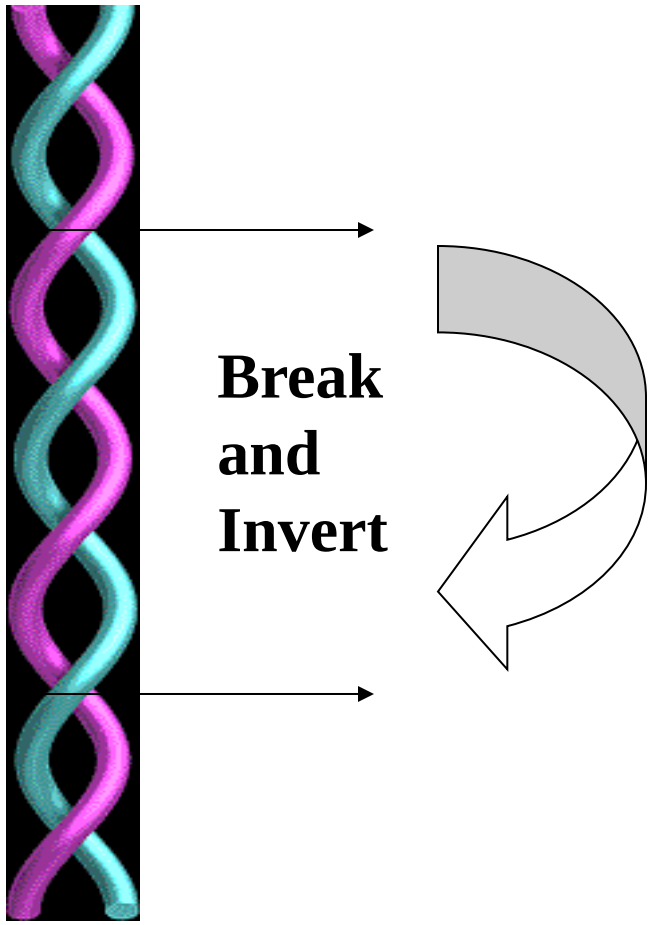
Comparing gene sequences yields no evolutionary information

GENE ORDER COMPARISON



Evolution is manifested as the divergence in Gene Order

DNA Reversal



5’ A T G C C T G T A C T A 3’

3’ T A C G G A C A T G A T 5’

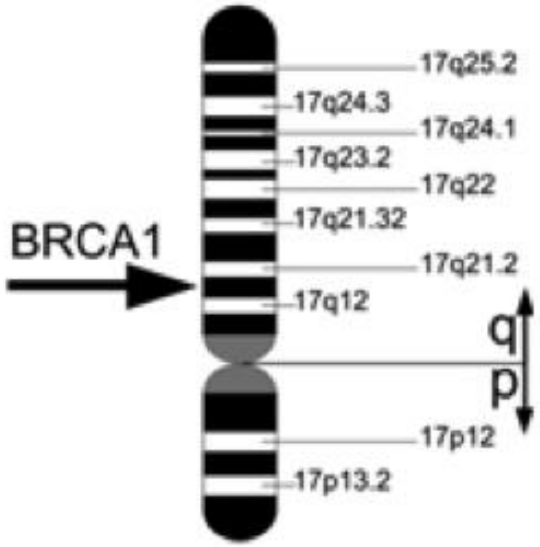
5’ A T G T A C A G G C T A 3’

3’ T A C A T G T C C G A T 5’

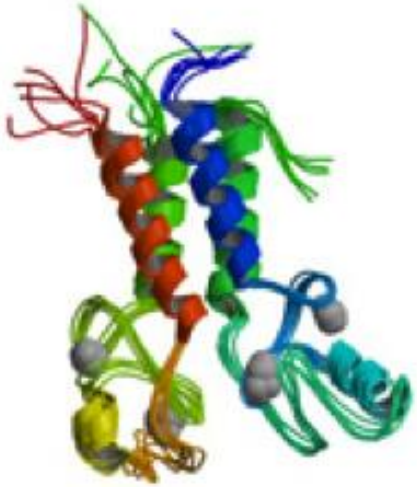
Waardenburg's syndrome

- Genetic disorder
- Characterized by loss of hearing and pigmentary dysphasia
- Found on human chromosome 2

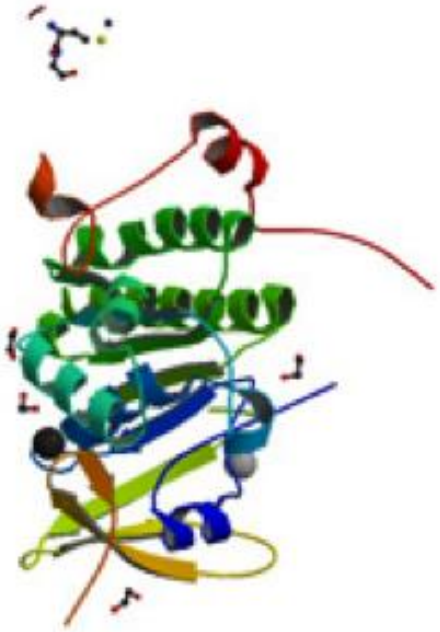
Chromosome 17



BRCA1



BRCA2



It is Sequenced, What's Next?

- Tracing Phylogeny
  - Finding family relationships between species by tracking similarities between species.
- Gene Annotation (cooperative genomics)
  - Comparison of similar species.
- Proteomics
  - From DNA sequence to a folded protein.
- Determining Regulatory Networks
  - How the genes react to certain stimuli?
  - How cancer progresses?

Sequence Driven Problems in Bioinformatics

- Genomics
  - Fragment assembly of the DNA sequence.
    - Not possible to read entire sequence.
    - Cut up into small fragments using restriction enzymes.
    - Then need to do fragment assembly. Overlapping similarities to matching fragments.
    - N-P complete problem.
  - Finding Genes
    - Identify open reading frames
      - Exons are spliced out.
      - Junk in between genes

More Computing problems…

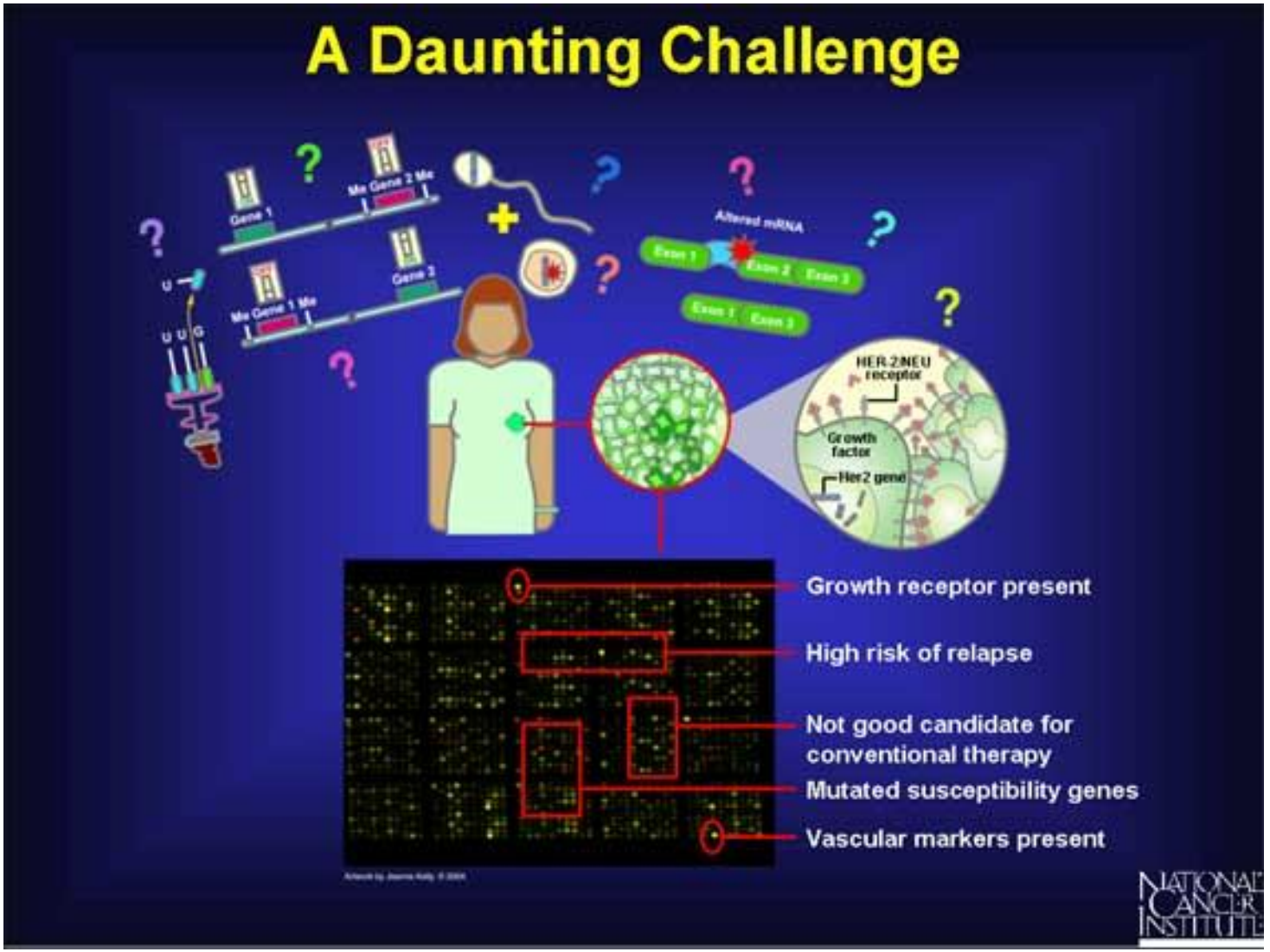
- Proteomics
  - Protein Folding
    - 1D Sequence → 3D Structure
    - What drives this process?
  - Identification of functional domains in protein's sequence
    - Determining functional pieces in proteins.
- *Most problems need mathematical solutions & For Large data they need software*

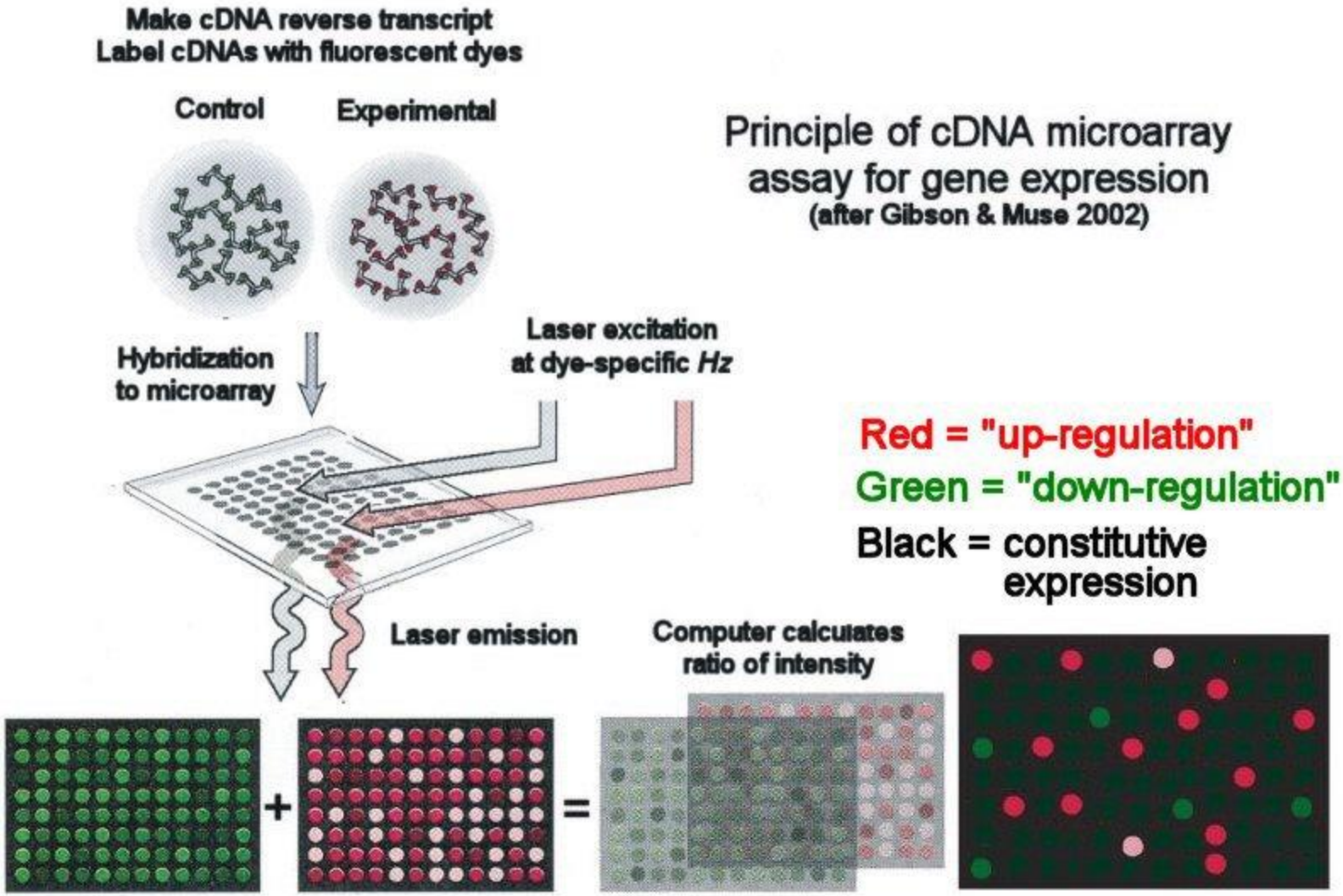
Biological Databases

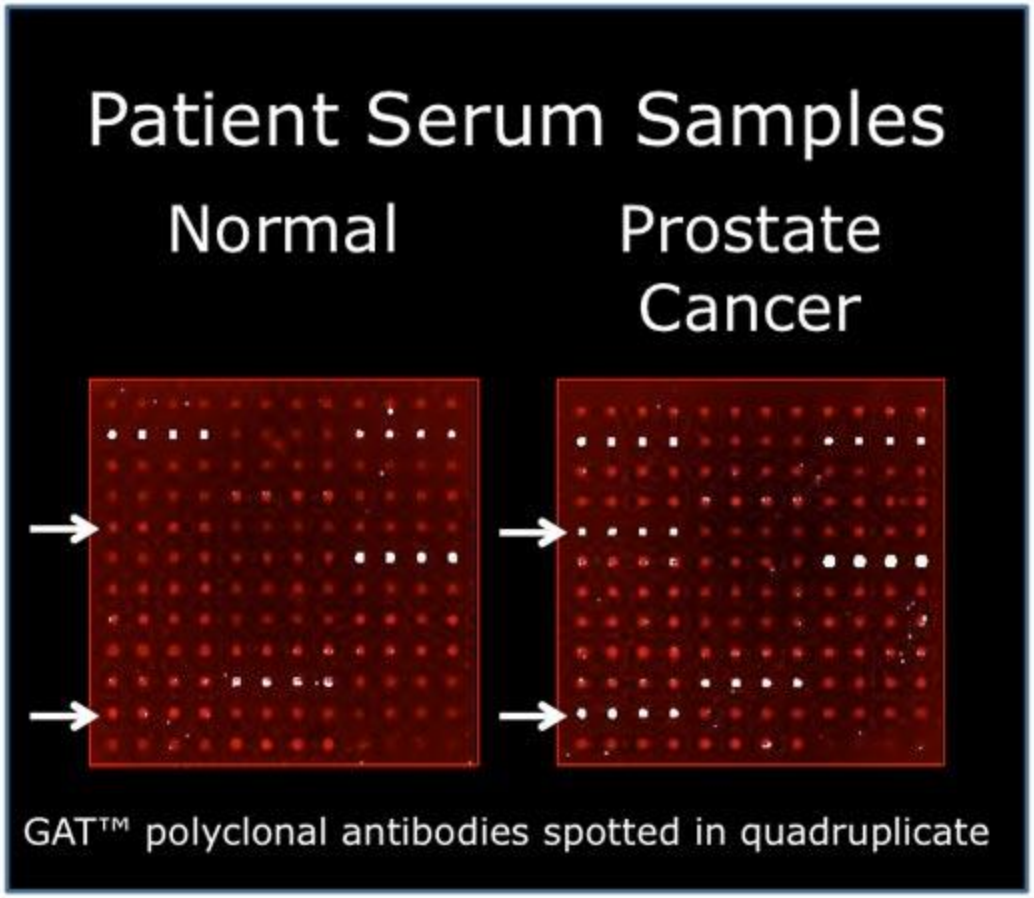
- Online databases to archive, search, and run programs on – massive amount of data

- NCBI GeneBank http://ncbi.nih.gov
  Huge collection of databases, the most prominent being the nucleotide sequence database
- Protein Data Bank http://www.pdb.org
  Database of protein 3D-structures
- SWISSPROT http://www.expasy.org/sprot/
- Database of annotated protein sequences
- PROSITE http://kr.expasy.org/prosite
  Database of protein active site motifs

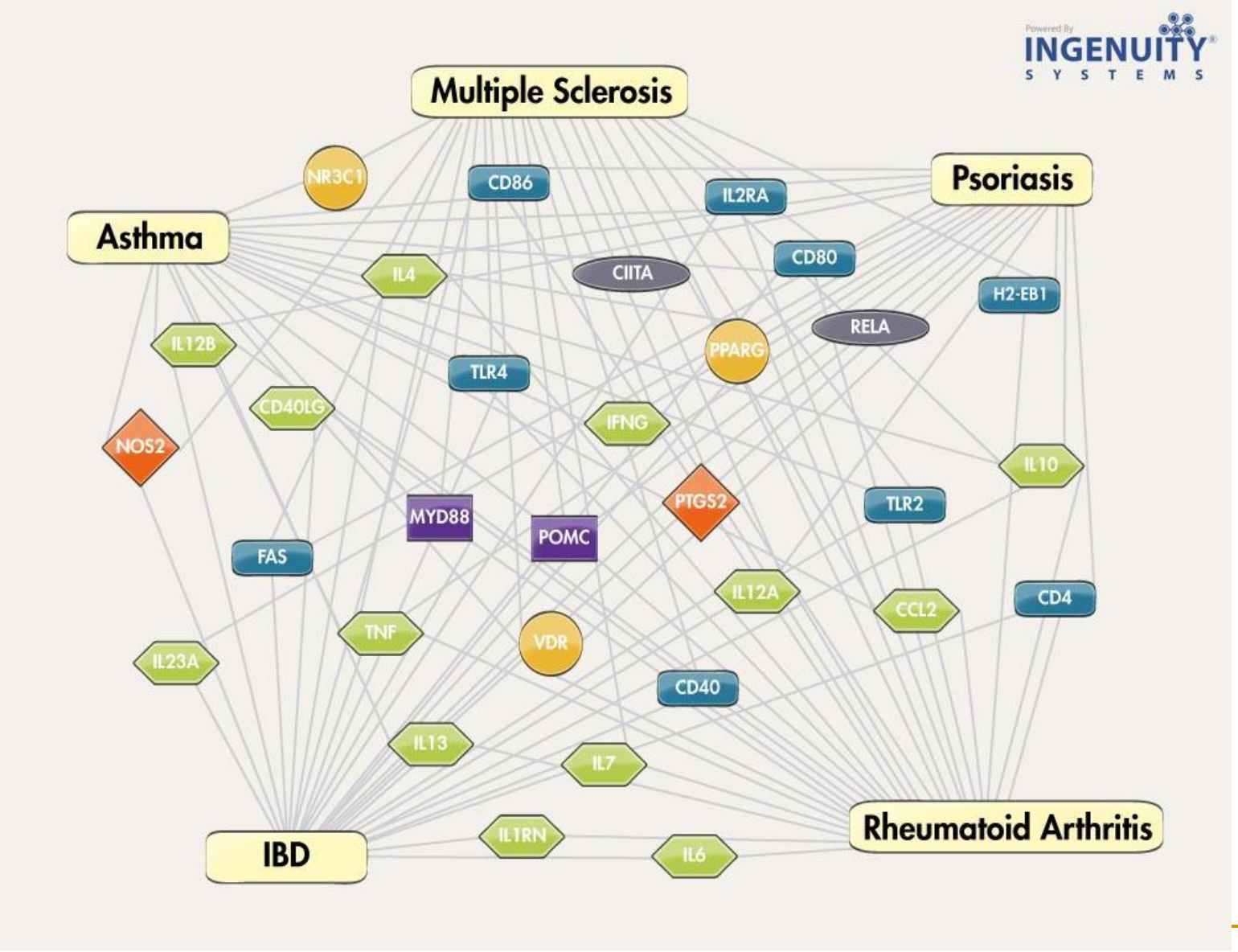
How gene expressions translate to Prognosis or Diagnosis: A Cancer Question




Make cDNA reverse transcript
Label cDNAs with fluorescent dyes
Control
Experimental
Principle of cDNA microarray assay for gene expression
(after Gibson & Muse 2002)
Hybridization to microarray
Laser excitation at dye-specific Hz
Red = "up-regulation"
Green = "down-regulation"
Black = constitutive expression
Laser emission
Computer calculates ratio of intensity


Patient Serum Samples
Normal
Prostate Cancer
GAT™ polyclonal antibodies spotted in quadruplicate

Microarray Experiments lead to understanding Gene Regulatory Networks



Big Data?

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Current Thoughts on Big-data & Biology – a Conference

http://www.triconference.com/Bioinformatics-Genome/

"Bioinformatics for Big Data: How Applications of Big Data will Drive Research Forward"

February 10-12, 2014 | Moscone North Convention Center | San Francisco, CA

Session: TECHNOLOGIES GENERATING BIG BIOMEDICAL DATA

11:00 Cancer Genomics

David Haussler, Ph.D., Distinguished Professor and Director,
Center for Biomolecular Science & Engineering, Univ. of California Santa Cruz

UCSC has built the Cancer Genomics Hub (CGHub) for the US National Cancer Institute, designed to hold up to 5 petabytes of research genomics data (up to 50,000 whole genomes), including data for all major NCI projects. To date it has served more than 8.3 petabytes of data to more than 300 research labs. Cancer is exceedingly complex, with thousands of subtypes involving an immense number of different combinations of mutations. The only way we will understand it is to gather together DNA data from many thousands of cancer genomes so that we have the statistical power to distinguish between recurring combinations of mutations that drive cancer progression and "passenger" mutations that occur by random chance. Currently, with the exception of a few projects such as ICGC and TCGA, most cancer genomics research is taking place in research silos, with little opportunity for data sharing. If this trend continues, we lose an incredible opportunity. Soon cancer genome sequencing will be widespread in clinical practice, making it possible in principle to study as many as a million cancer genomes. For these data to also have impact on understanding cancer, we must begin soon to move data into a global cloud storage and computing system, and design mechanisms t hat allow clinical data to be used in research with appropriate patient consent. A global alliance for sharing genomic and clinical data is emerging to address this problem. This is an opportunity we cannot turn away from, but involves both social and technical challenges.

Translation:
Huge database of mutated gene's

Session: DATA STORAGE AND MAINTENANCE

2:20 Implementing Big Data Analysis and Archival Solutions for NGS Data

*Zhiyan Fu, Ph.D., Chief Scientific Computing Officer, Genome Institute of Singapore (A*STAR)*

This presentation shows the latest development in big data analysis, compression and storage management. It provides a practical case to implement the big data technologies to a mid-size genome center. Attendees will understand the challenges of big data life-cycle management in a genome center and see how the latest big data technologies are implemented, and the pros and cons of some of the techniques, including Hadoop, HDF5, and different NGS compression algorithms evaluated by GIS.

Translation:

Experience in managing Next Gen. Sequencing Data storage

Session: HOW BIG DATA WILL DRIVE RESEARCH FORWARD

1:45 Harnessing Big Data to Accelerate Drug Development

Vinod Kumar, Ph.D., Senior Investigator, Computational Biology, GlaxoSmithKline Pharmaceuticals

With the rapid development of high-throughput technologies and ever-increasing accumulation of whole genome-level datasets, an increasing number of diseases and drugs can be comprehensively characterized by the changes they induce in gene expression, protein, metabolites and phenotypes. Integrating and querying such large volumes of data, often spanning domains and residing in diverse sources, constitutes a significant obstacle. This talk presents two distinct approaches that utilize these data types to systematically evaluate and suggest new disease indications for new and existing drugs.

Translation:
Predict new disease indications from Data for Drug Modeling, first by software

Session: BIG DATA DRIVING PERSONALIZED MEDICINE

5:05 It's Not Just About Big Data...
Big Analytics for Identifying What Works and for Whom in Healthcare
Iya Khalil, Ph.D., Executive Vice President and Co-Founder, GNS Healthcare

We are living in the era of big data in healthcare, with unprecedented ability to collect data at multiple levels (genomic/'omic', phenotypic, health records, mobile health, etc.) and at scale. The key will be leveraging advanced analytics and appropriate feedback loops to identify what works on an individual patient level.

Translation:
Smart Algorithms needed for connecting multi-omics Big Data

Thank you!

Debasis Mitra

Professor, Computer Science, FIT

dmitra@cs.fit.edu