How Computers Can Cure Cancer

COS 116: 5/1/07 guest lecture:

Matthew Hibbs

Lewis-Sigler Institute for Integrative Genomics
Department of Computer Science

Many slides from Olga Troyanskaya

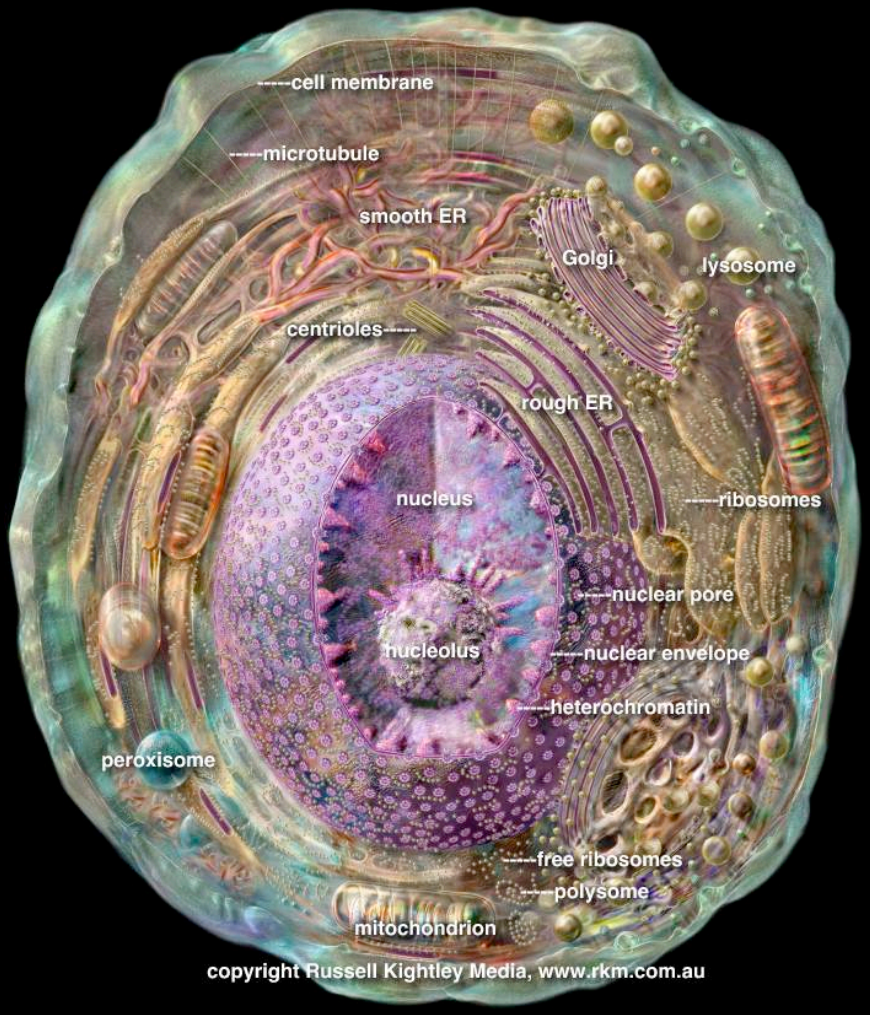
Big Challenges

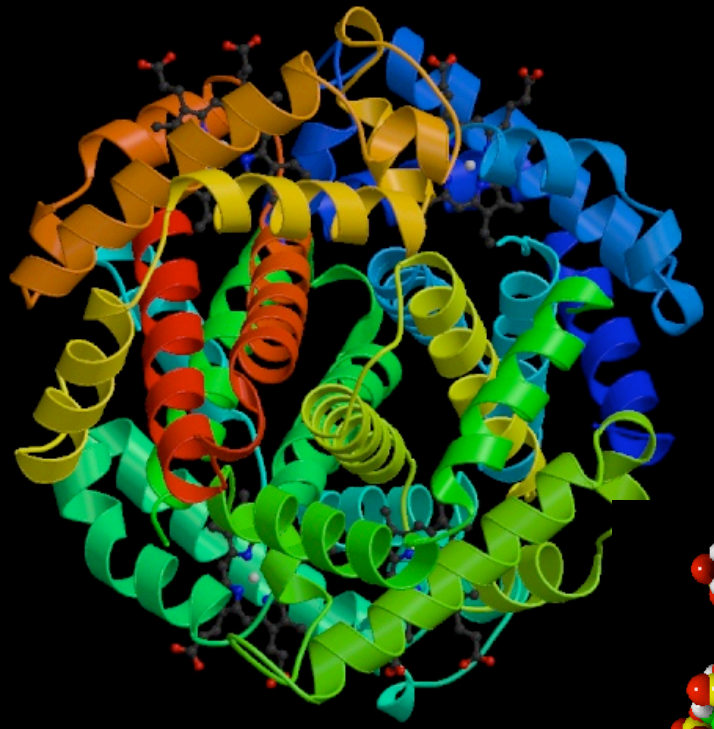
- Understand biology at multiple levels
  - genomic, cellular, organismal, etc.
- Diagnose and treat disease
  - cancer, malaria, etc.
- CS Can Help!
  - Post-genomic era
  - Huge amounts of data generated
  - Algorithms and methods for analysis

Bio 101

Or, Why do we care?

Cells, Proteins, DNA





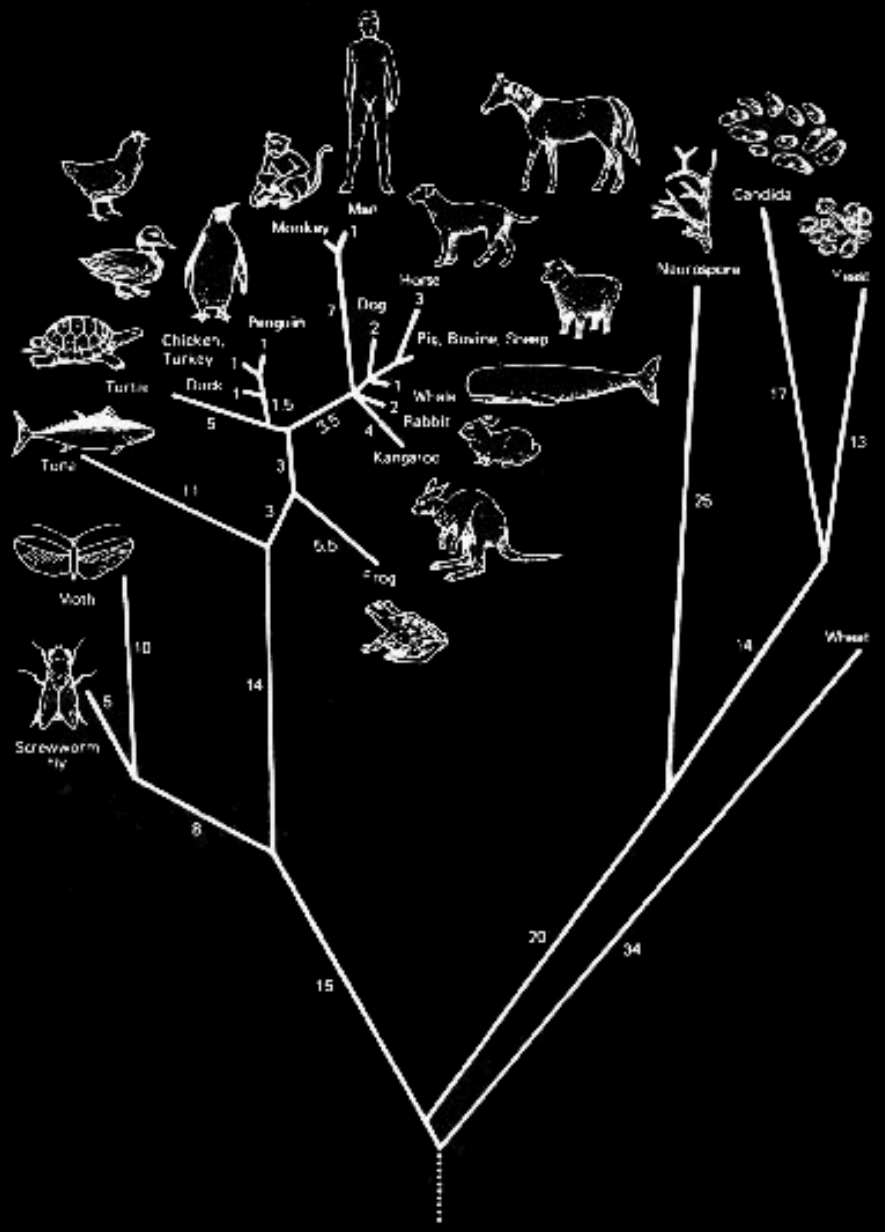
Evolution is Key

Common descent of organisms implies that they will share many "basic technologies."

Development of new adaptations in response to environmental pressure can lead to "specialized technologies."

More recent divergence implies more shared technologies between species.

All of biology is about two things: understanding shared or unshared features.



Model Organisms

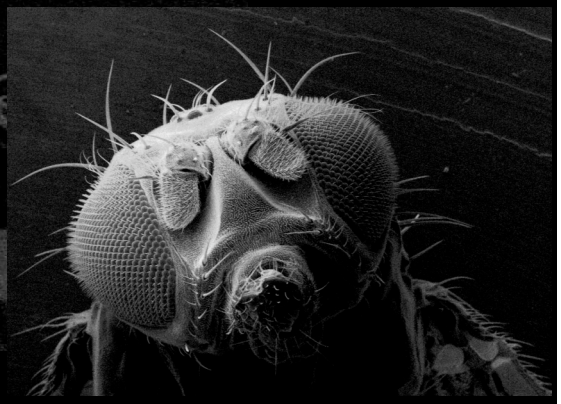
E. coli

cerevisiae

M. musculus

C. elegans

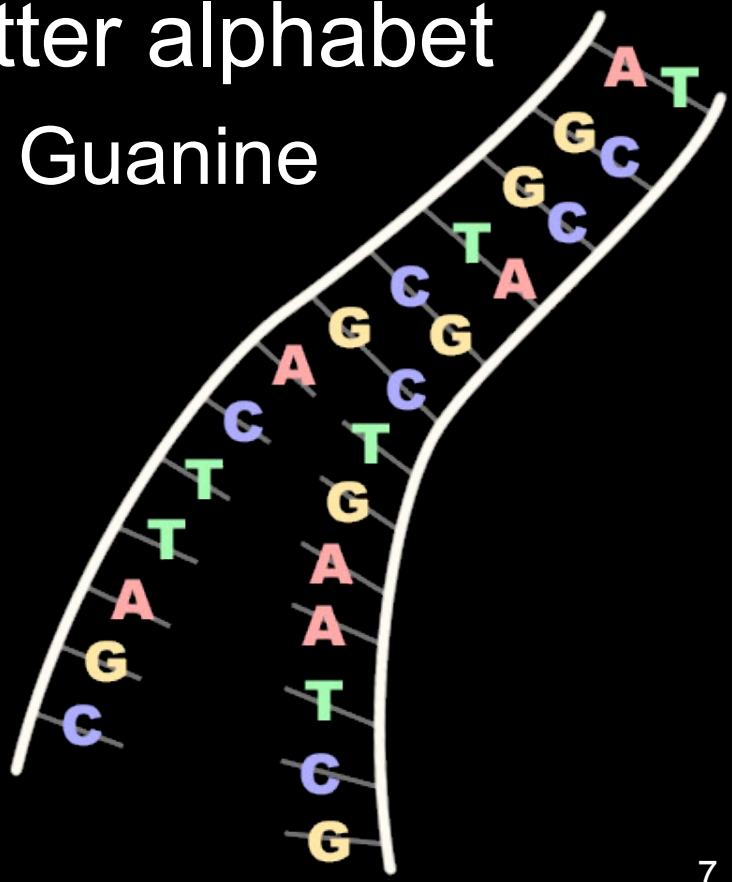
D. melanogaster



D. melanogaster

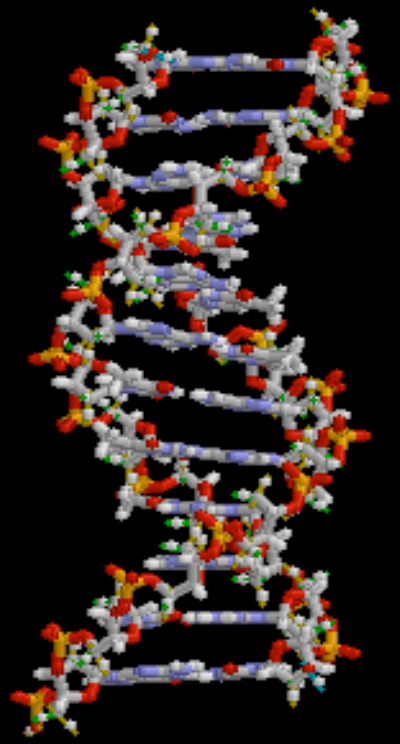
What is DNA?

- Deoxyribonucleic Acid
- Instructions, written in a 4-letter alphabet
  - Adenine, Thiamine, Cytosine, Guanine



DNA Structure

James Watson



Francis Crick

DNA Content

E. coli - 4 million bp (~1.36 mm) - 3000 genes

Baker's yeast - 13.5 million bp (~4.6 cm) - 6000 genes

Human - 3 billion bp (~1 m) - ~25,000 genes

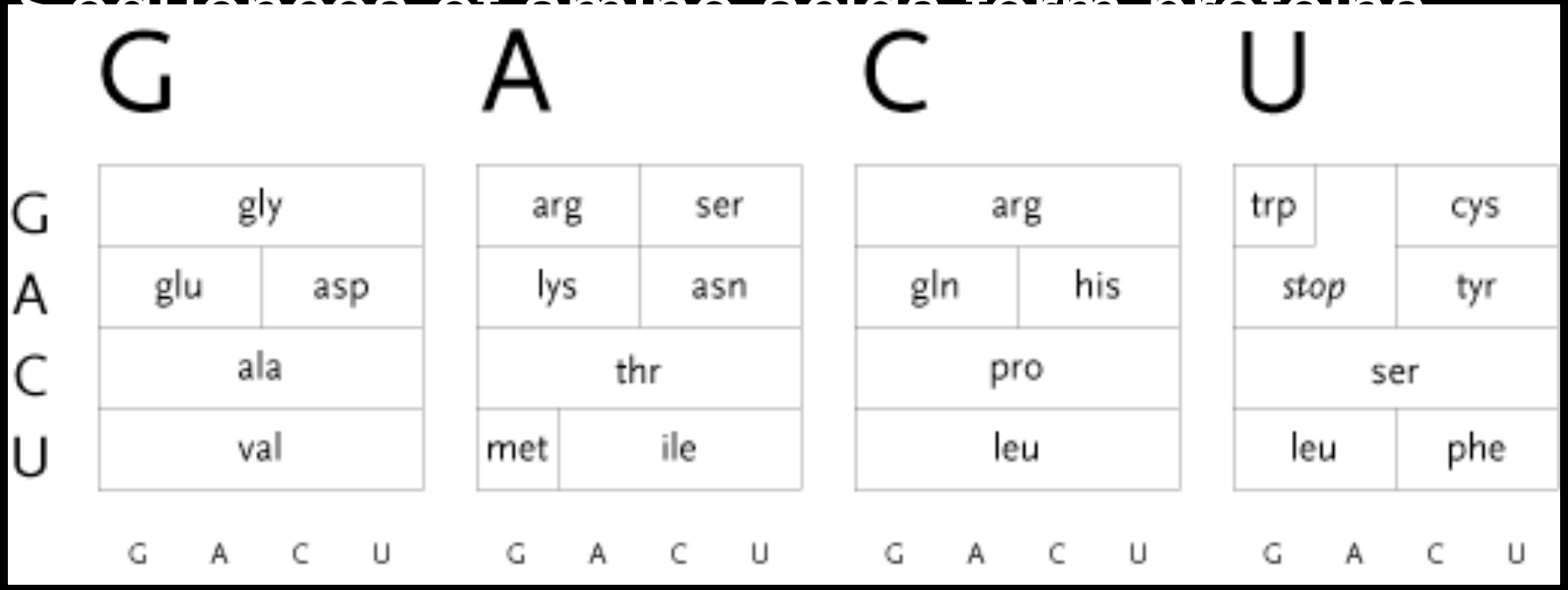
Waterlily

Salamander

Inch plant

What does DNA do?

- Each base pair triplet (codon) encodes an amino acid
- Sequences of amino acids form proteins
-



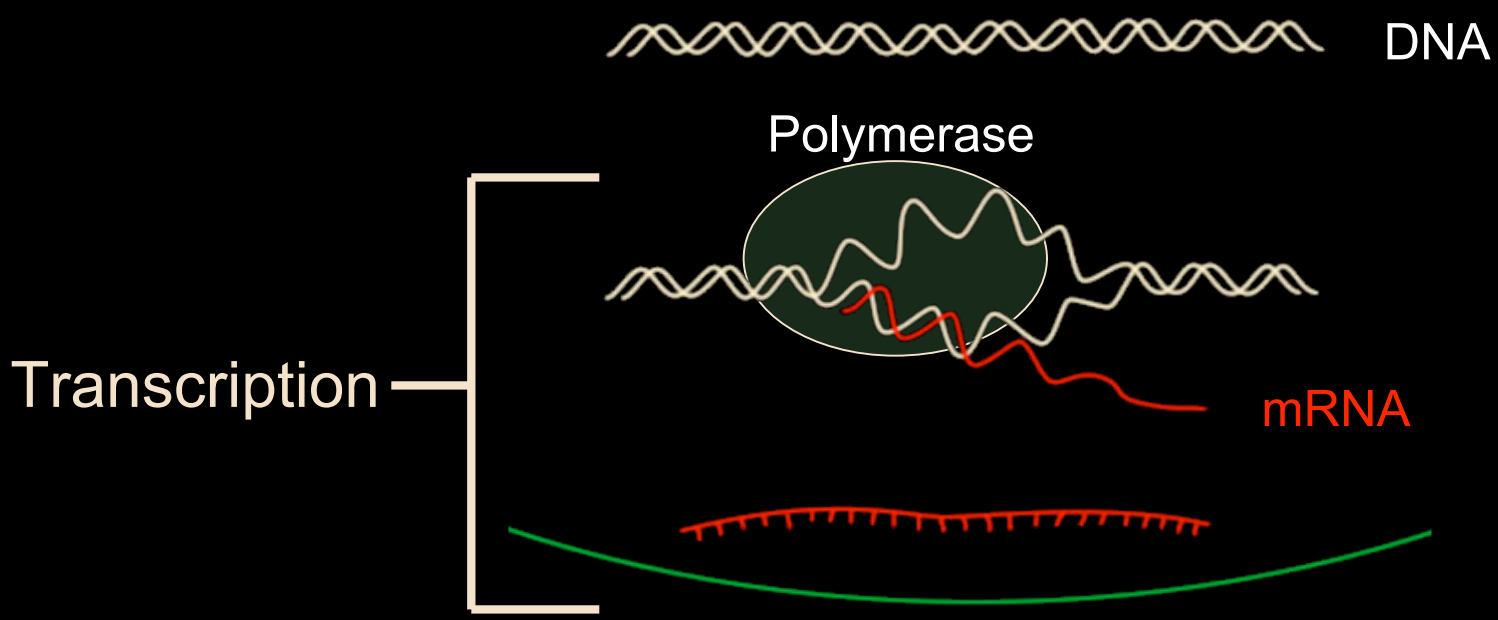
	G	A	C	U
G	gly	arg / ser	arg	trp / cys
A	glu / asp	lys / asn	gln / his	*stop* / tyr
C	ala	thr	pro	ser
U	val	met / ile	leu	leu / phe

Table by Ben Fry

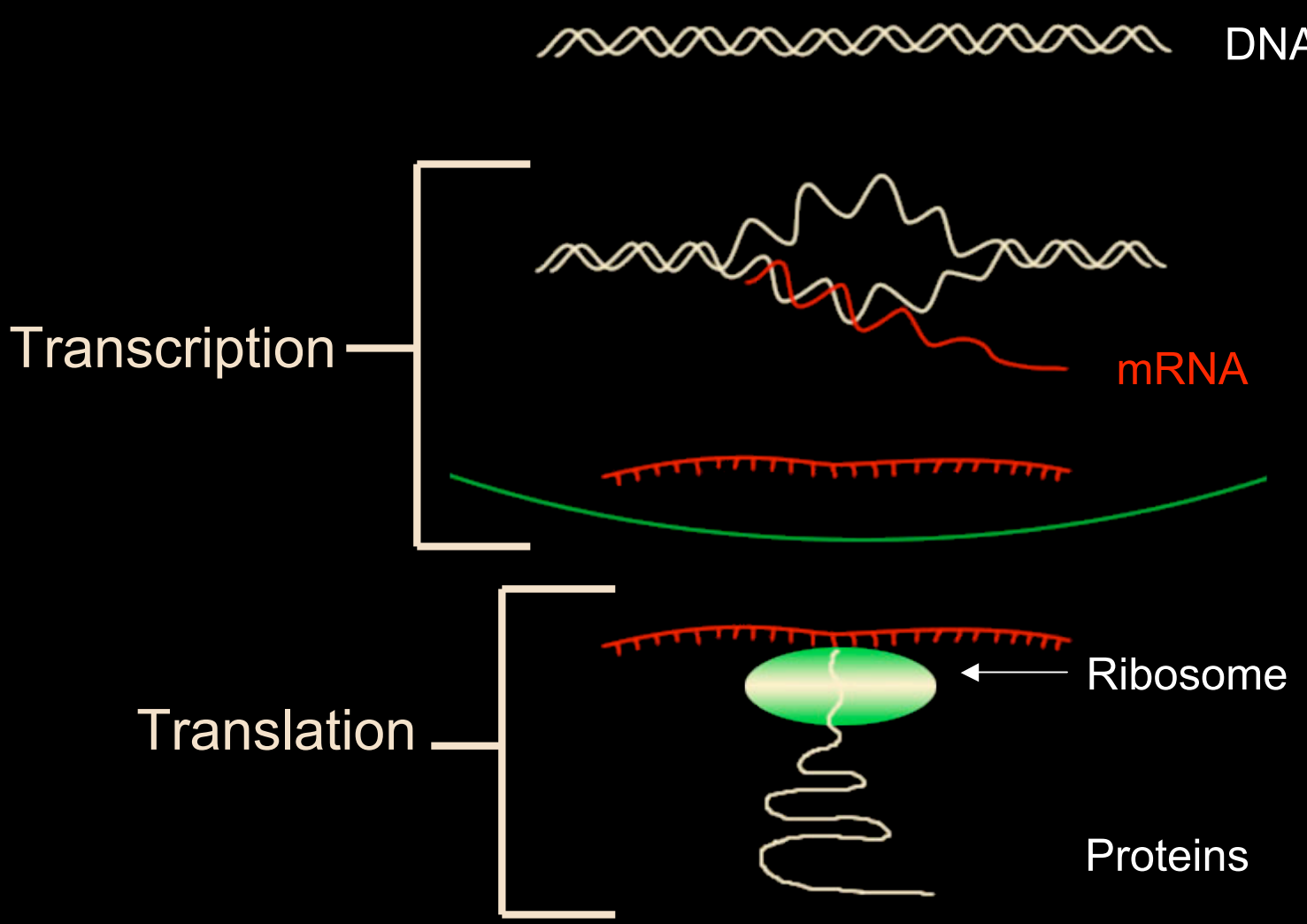
Central Dogma

DNA

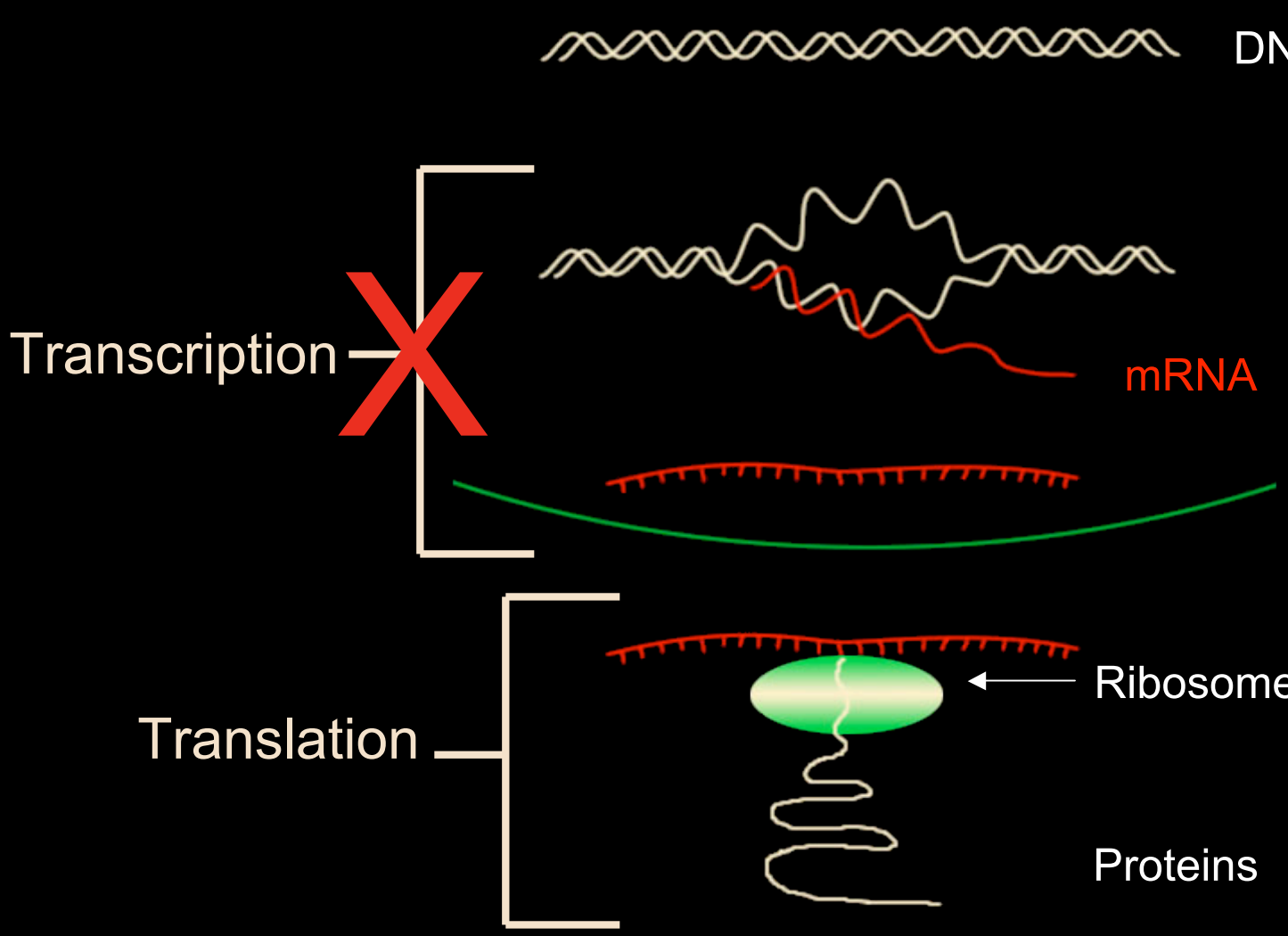
Central Dogma



Central Dogma

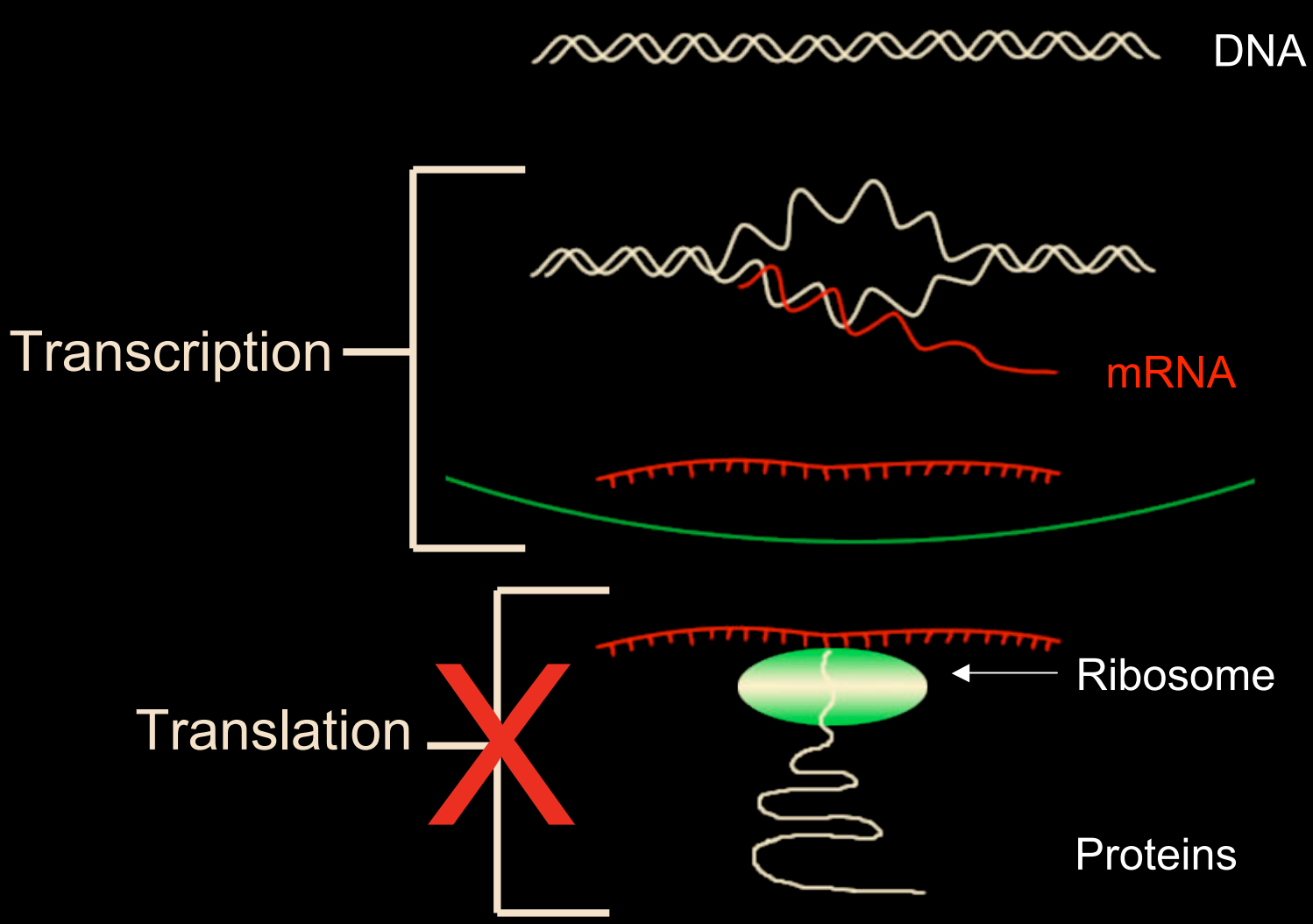


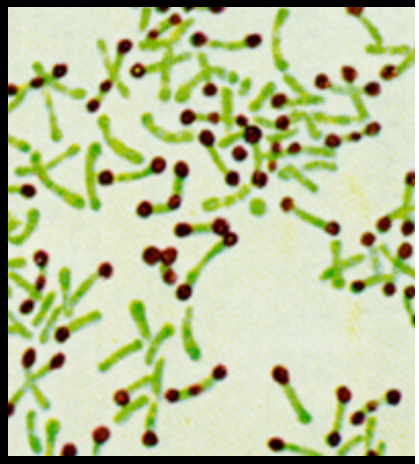
Central Dogma



Death Cap

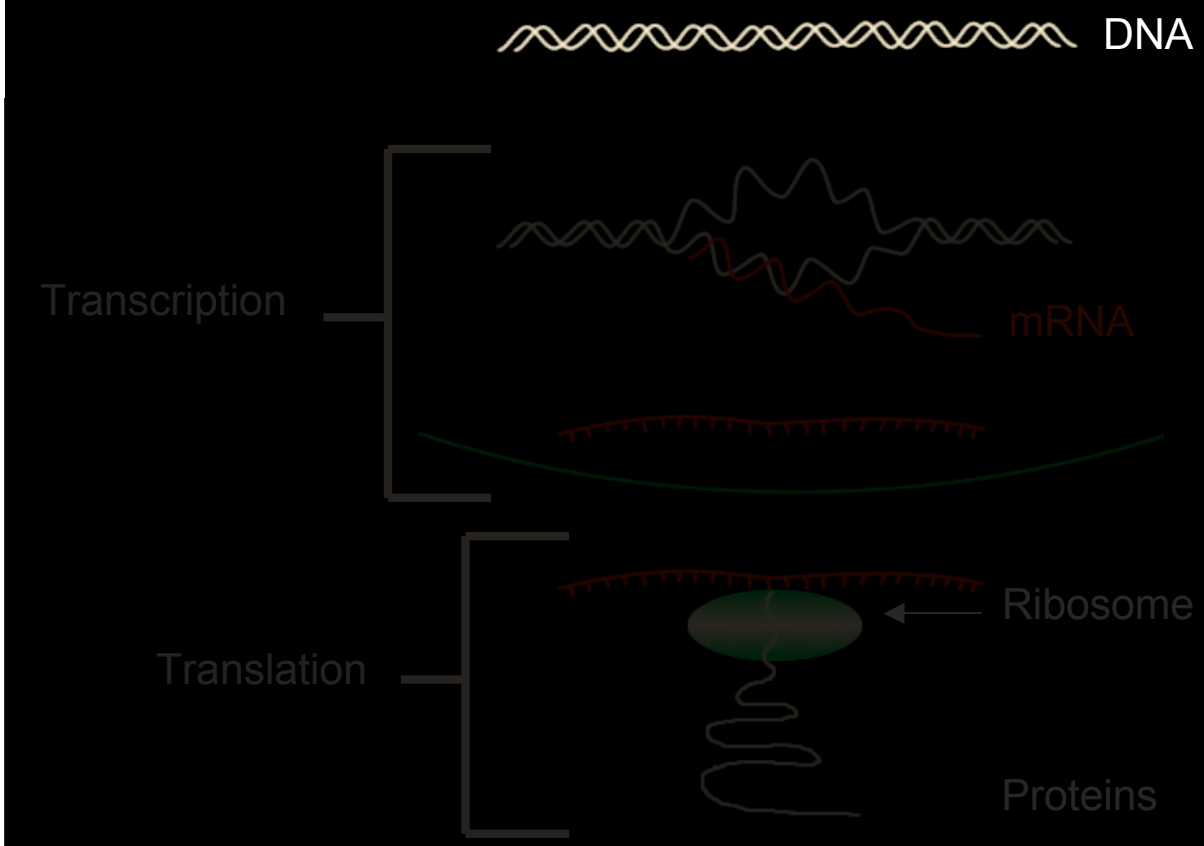
Central Dogma





Diphtheria

How Can CS Help?



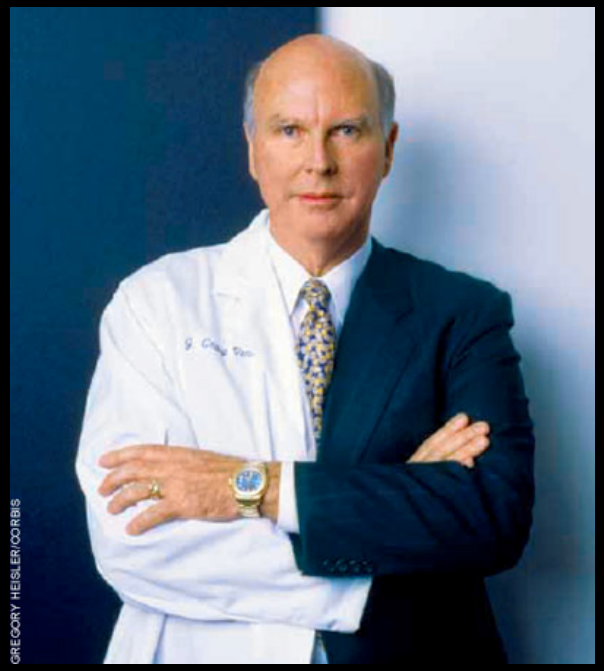
The Sequence

Sequence Analysis

- Knowledge of the sequence is the foundation of modern biology/genetics research
- The sequence is huge, computers and algorithms were vital
- Given the sequence for many organisms, we open up the doors for current and future research

The Great Sequence Race

- End of the 20th century, 2 major human sequencing projects
- Gov't run Human Genome Project vs. Celera
- Craig Venter, president of Celera developed the "winning" technique
  - used computers over lab techniques



“Shotgun” sequencing

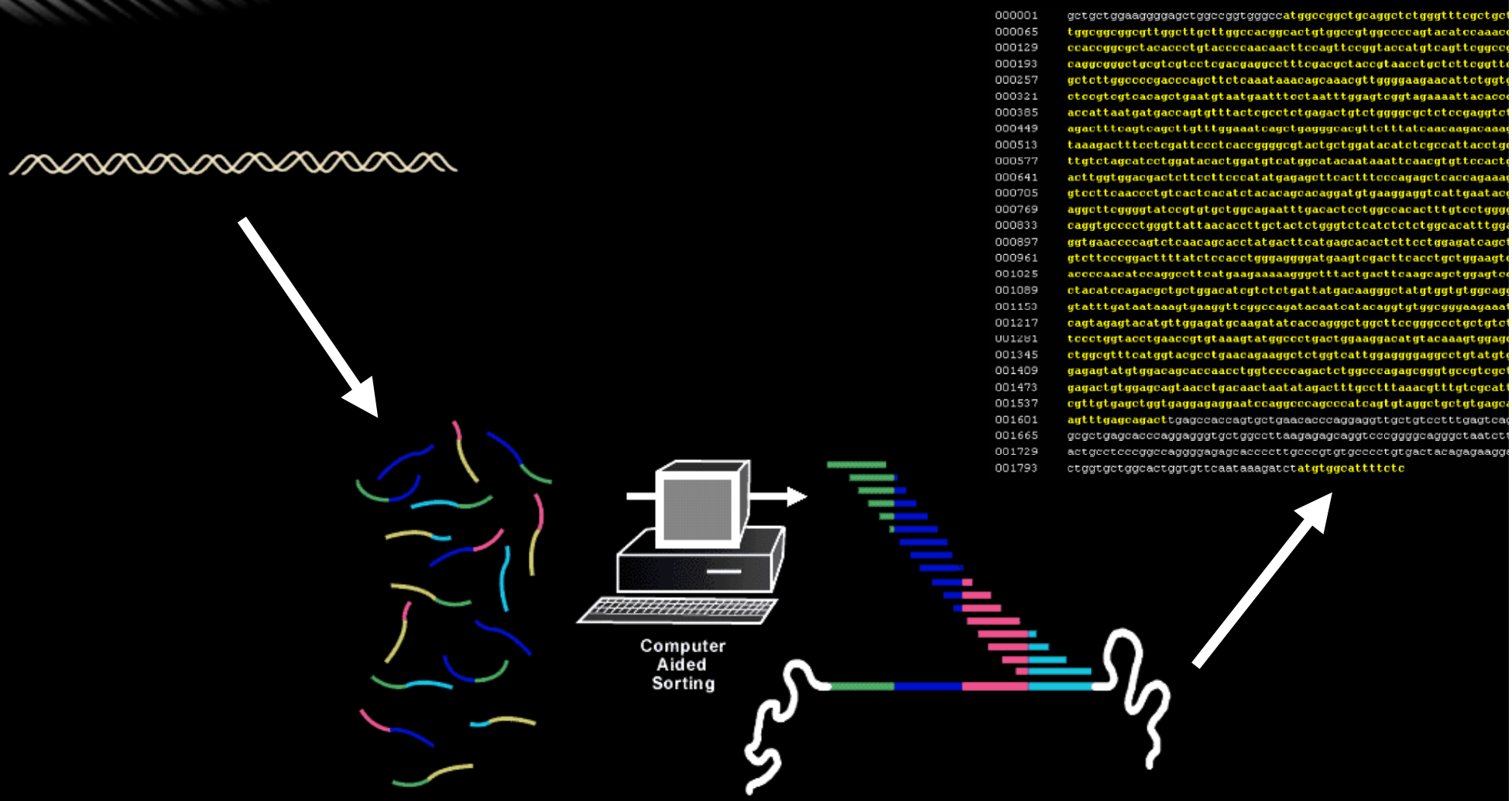
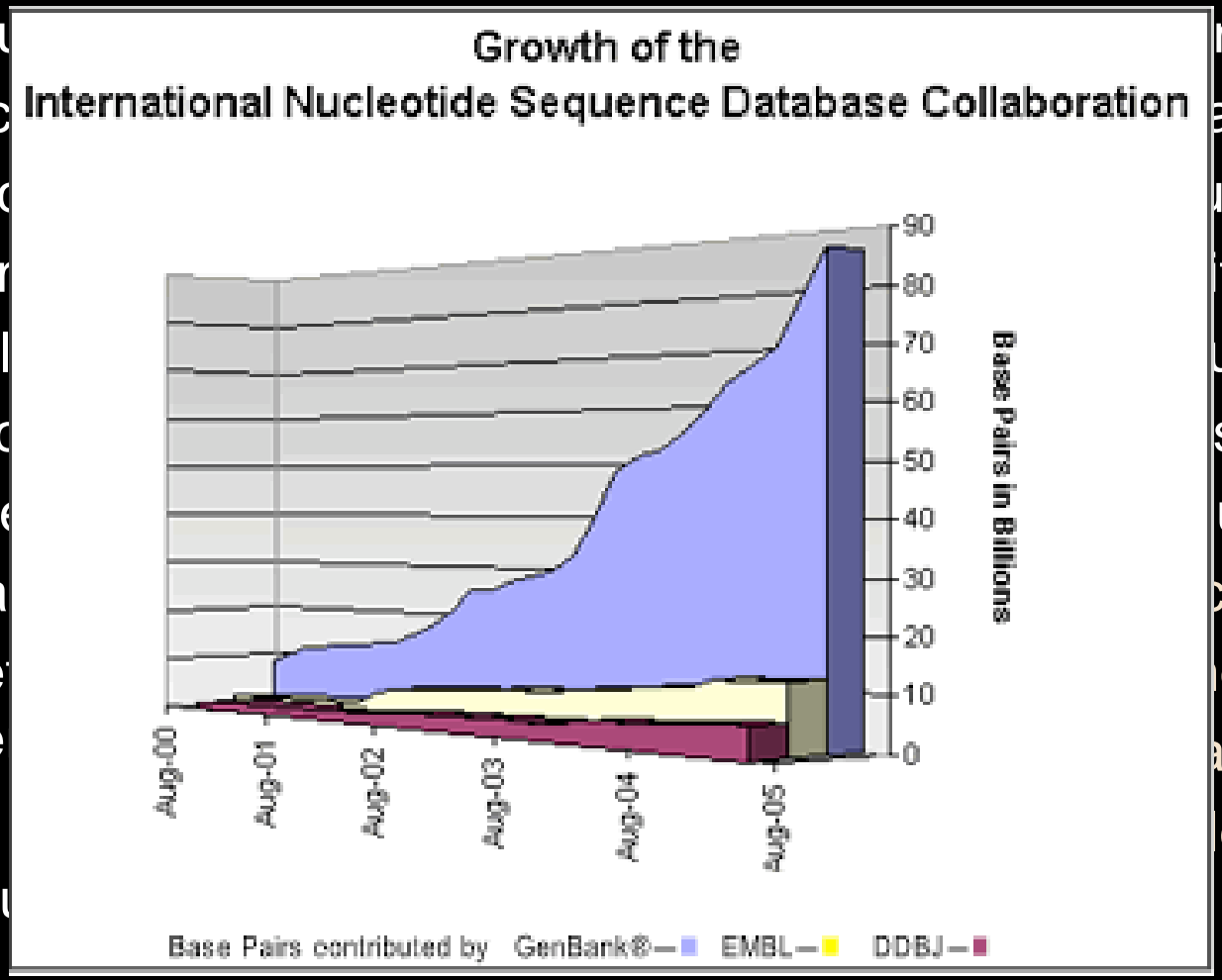


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

1000s of Sequenced Genomes

- Aqu … ri
- Arc … annaschii
- Bac … uberculosis
- Bor … talium
- Chl … umoniae
- Esc … shii
- Hae … um
- Pla … erevisiae
- Me … ogaster
  the … ana
- Cau … egans



BLAST

- *S. cerevisiae* gene msh2 (YOL090W)
- Human mutator gene (hMSH2). GenBank: U03911

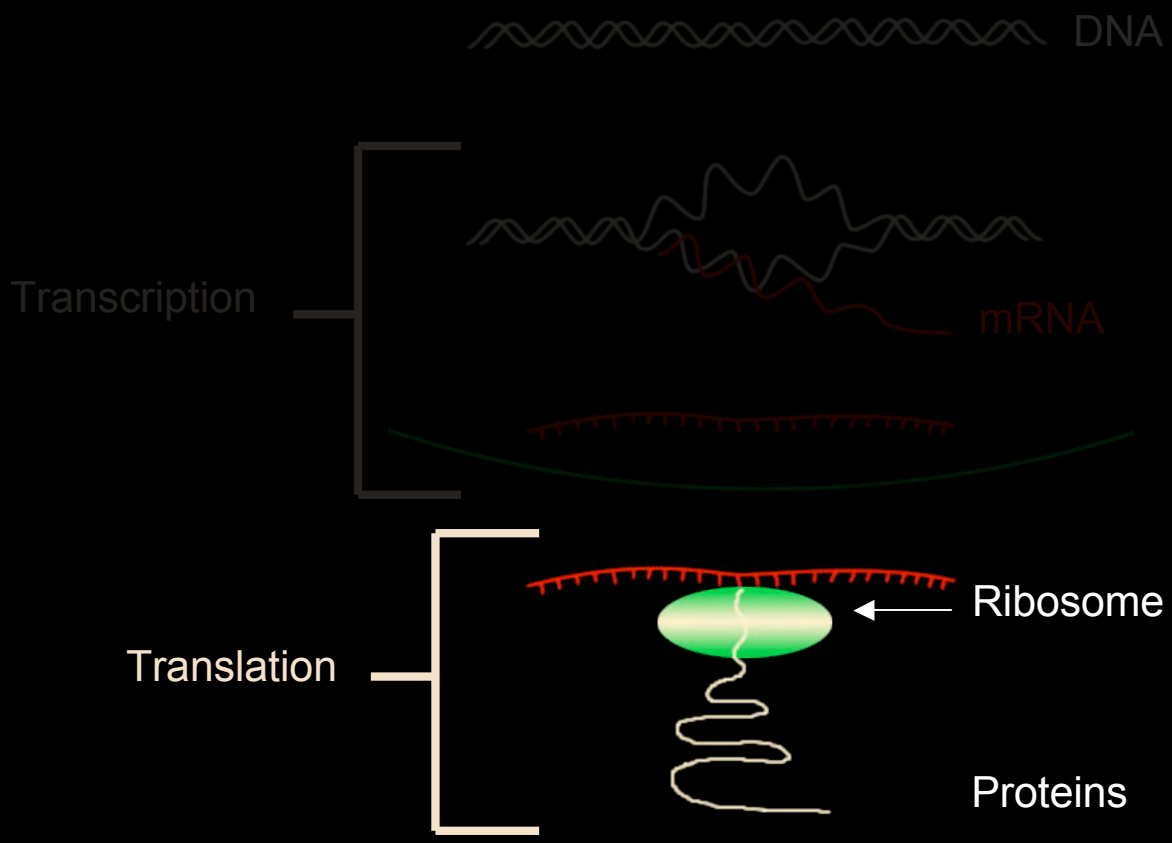
(Yeast sequence on top).

BLAST P-value: 3.8e-255
Percent Similarity (|+|): 64.7 Percent Identity (|): 43.0

```
  1 MSSTRPELKFSDVSEERNFYKKYTGLPKKPLKTIRLVDKGDYYTVIGSDA 50
    |.   ..|    | ..| .| | | |||.|| .||||.|||||||. |.||
  1 MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYTAHGEDA 50

 51 IFVADSVYHTQSVLKNCQLDPVTAKNFHEPTKYVTVSLQVLATLLKLCLL 100
    ||.| .||.||||||   |||..|||||.    |.||   |..|||  ||
 51 LLAAREVFKTQGVIKY..MGPAGAKNLQS....VVLSKMNFESFVKDLLL 94

 01 DLGYKVEIY.........DKGWKLIKSASPGNIEQVNELMNMNIDSSII 140
      .||||||           |.|| |  .||||||.|.||||  | | |
 95 VRQYRVEVYKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMSAS 144

 41 IASLKVQWNSQDGNCIIGVAFIDTTAYKVGMLDIVDNEVYSNLESFLIQL 190
    || | |. .. ||.  |||||||... |||| ||.||| |||||.|||||
 45 IGVVGVKMSAVDGQRQVGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQI 194

 91 GVKECLVQDLTSNSNSNAEMQKVINVIDRCGCVVTLLKNSEFSEKDVELD 240
    |.||||.|    ||..|||.|| .||||.|.|||  |..|||.|||  |
 95 GPKECVLPG....GETAGDMGKLRQIIQRGGILITERKKADFSTKDIYQD 240

_41 LTKLL....GDDL.ALSLPQKYSKLSMGACNALIGYLQLLSEQDQVGKYE 285
    |.|||    |||| .  |||. ..|.||. .||| ||||||||...|.||
241 LNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKFLELLSDDSNFGQFE 290
```

Human hereditary colon cancer gene was found by looking for a gene similar to MSH2 gene in yeast (these genes are 65% similar)!

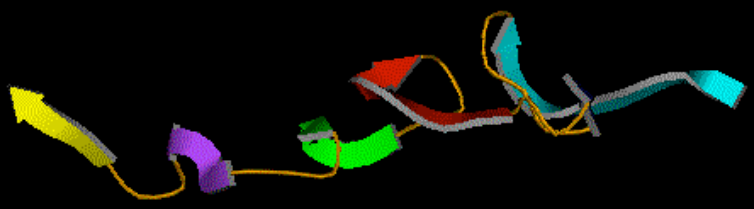
How Can CS Help?



The Proteins

Protein Structure

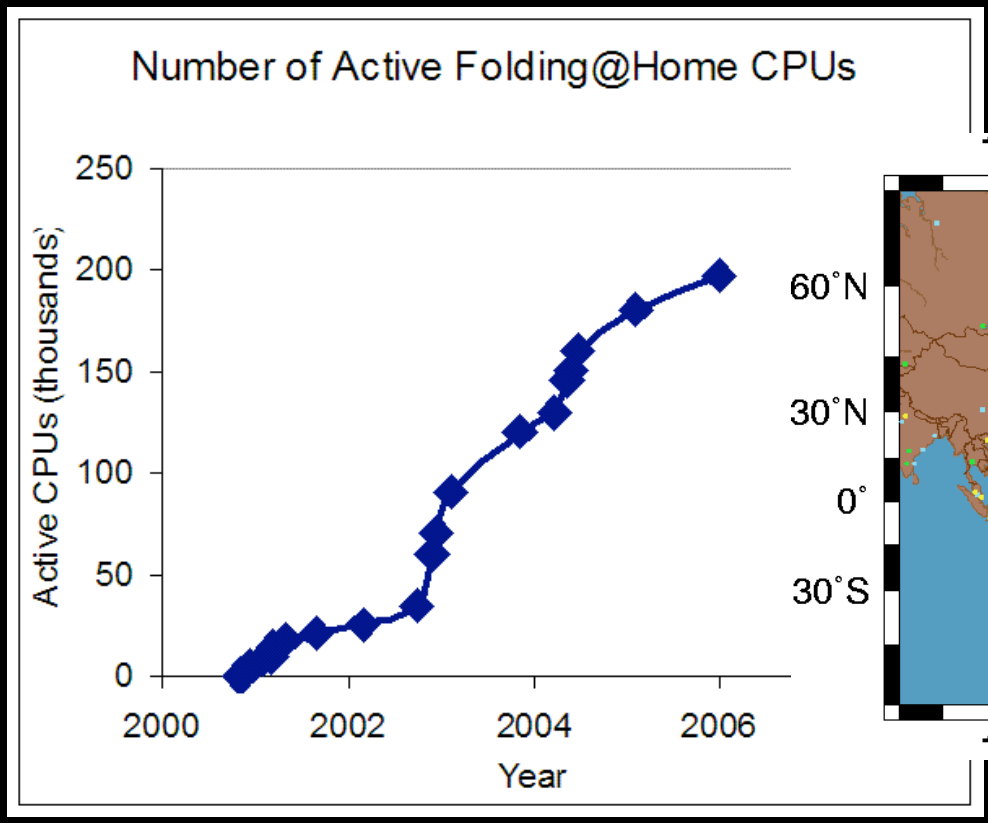
- Proteins fold into complex 3D structures
- Only functional if properly folded

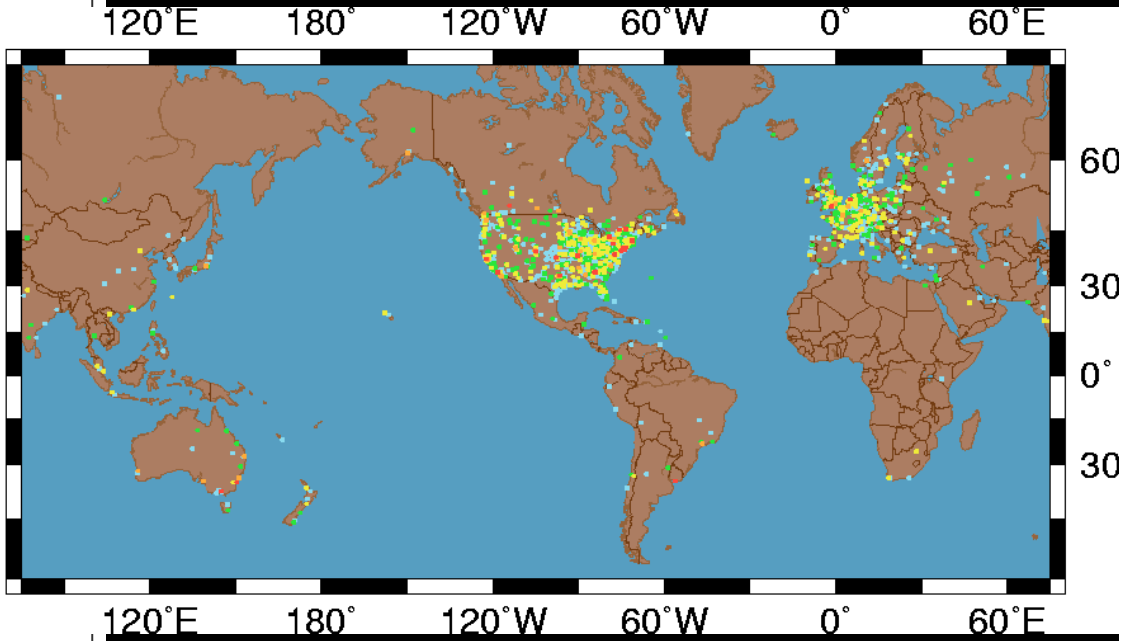


Courtesy of the Zhou Laboratory, The State University of New York at Buffalo

Folding@home distributed computing

- Project at Stanford to use computer downtime to process folds

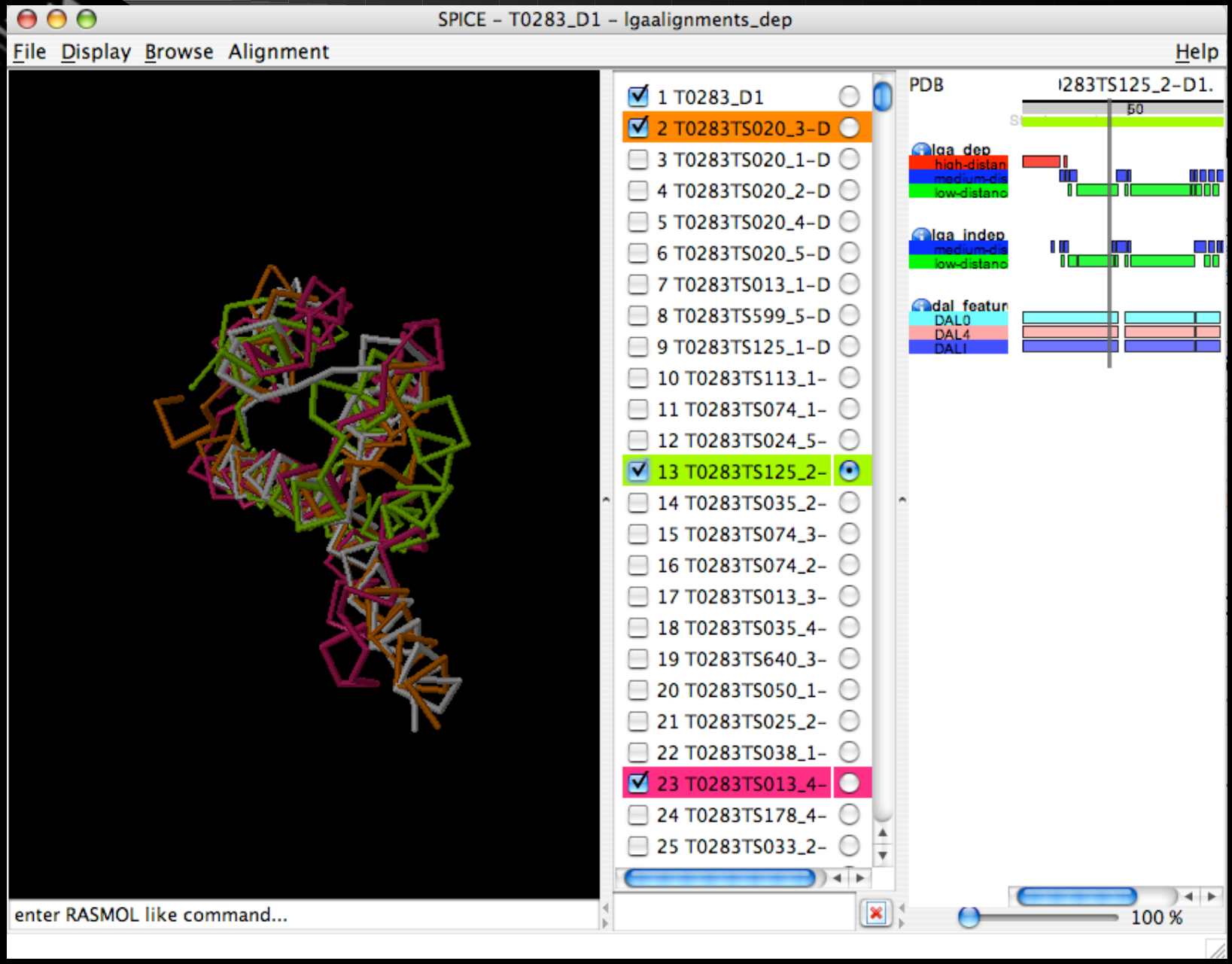




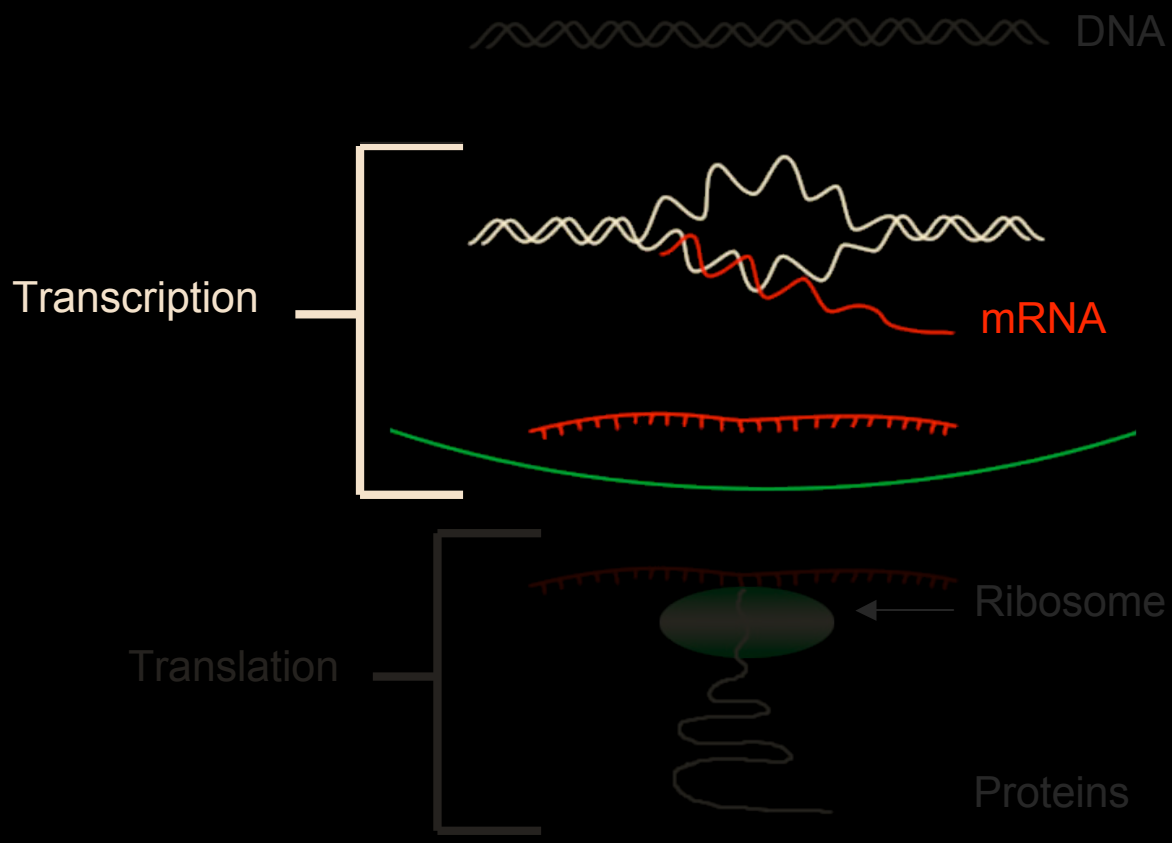
CASP

- Annual competition to predict protein structures
- A few crystallized proteins are “held back” from publication until after the contest
- Given just the sequence, groups try to predict the real structure
  - All groups use CS to accomplish this
  - Machine learning, data mining, phylogeny, etc.

CASP

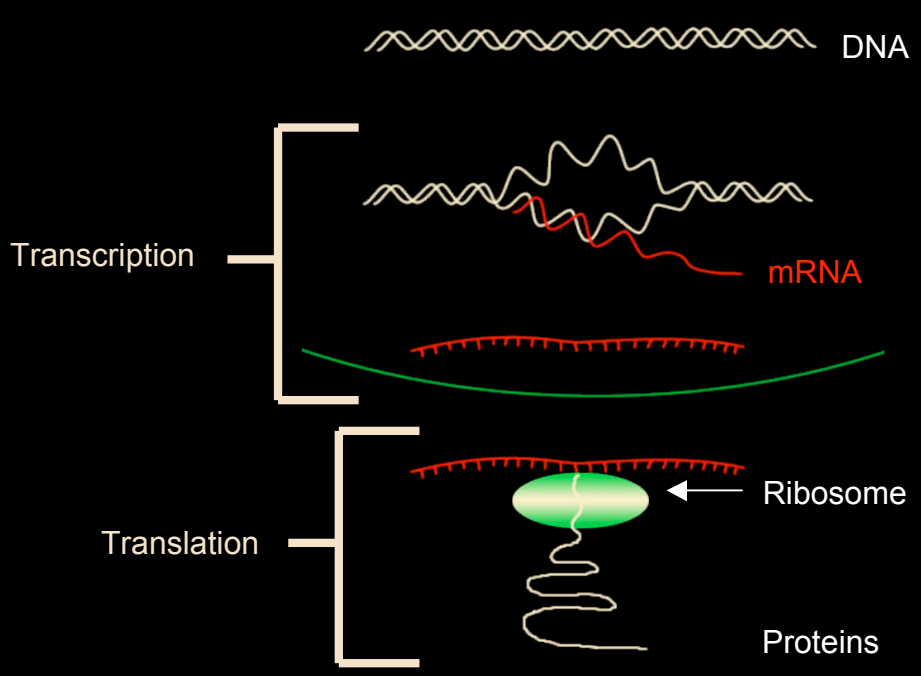

SPICE – T0283_D1 – lgaalignments_dep
File Display Browse Alignment
Help
1 T0283_D1
2 T0283TS020_3-D
3 T0283TS020_1-D
4 T0283TS020_2-D
5 T0283TS020_4-D
6 T0283TS020_5-D
7 T0283TS013_1-D
8 T0283TS599_5-D
9 T0283TS125_1-D
10 T0283TS113_1-
11 T0283TS074_1-
12 T0283TS024_5-
13 T0283TS125_2-
14 T0283TS035_2-
15 T0283TS074_3-
16 T0283TS074_2-
17 T0283TS013_3-
18 T0283TS035_4-
19 T0283TS640_3-
20 T0283TS050_1-
21 T0283TS025_2-
22 T0283TS038_1-
23 T0283TS013_4-
24 T0283TS178_4-
25 T0283TS033_2-
PDB
283TS125_2-D1.
50
lga_dep
high-distan
low-distanc
lga_indep
low-distanc
dal_featur
DAL0
DAL4
DAL1
enter RASMOL like command...
100 %

How Can CS Help?



The Transcriptome

Central Dogma



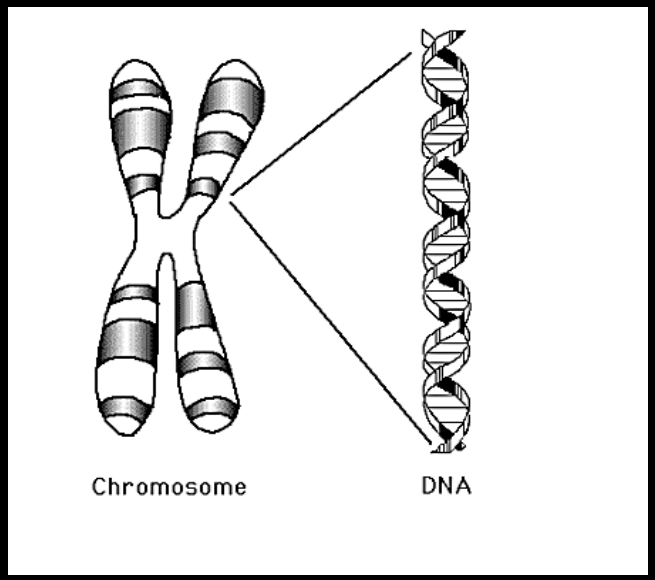
(Almost) all organisms have this basic process in common

Mammals have 99% of their genomes in common

So what makes organisms different from each other?

Small DNA changes → Big phenotype changes

Proteins



DNA

Gene Expression

Proteins

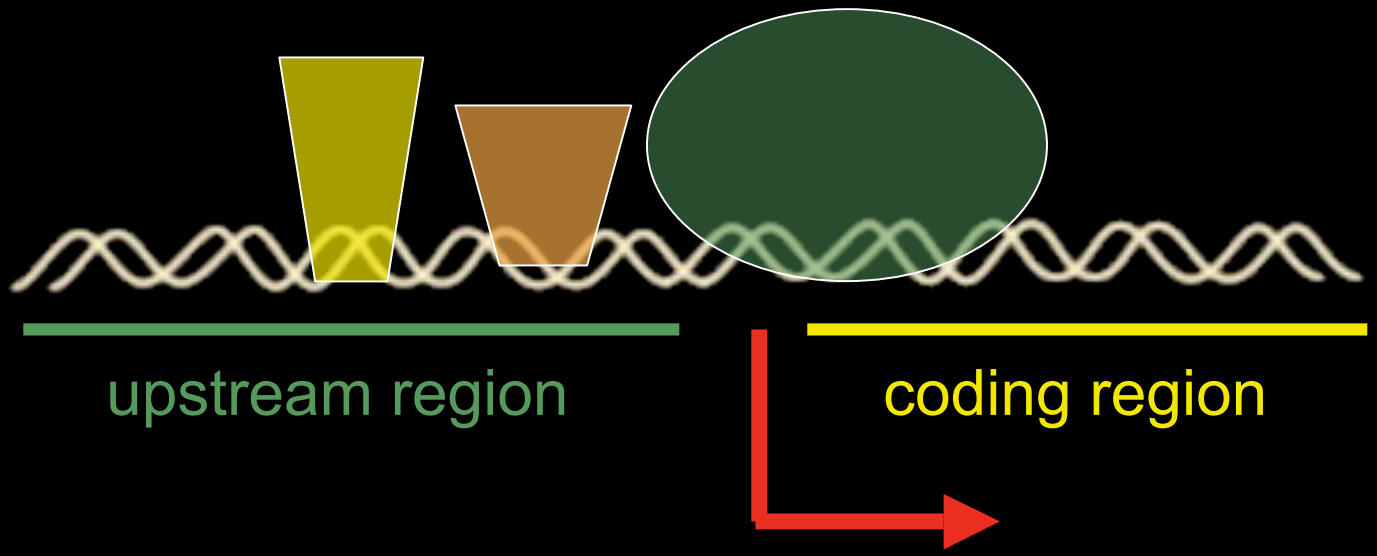
Transcription -- it's important

- Not all genes/proteins are made at all times
- Amounts made change over time

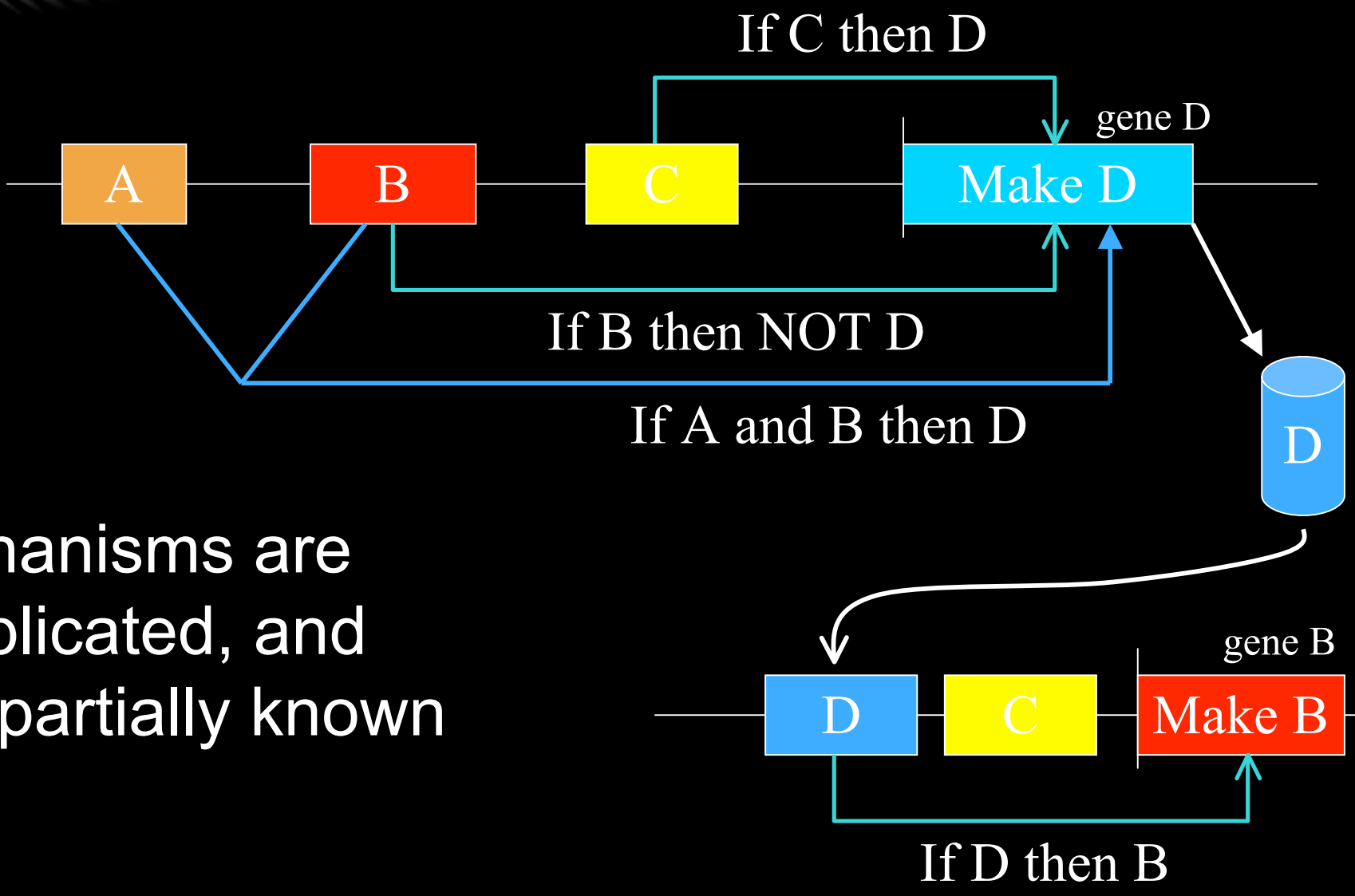
- Understanding regulation of what makes proteins is a big, key problem

Transcription

- Housekeeping vs. specialized genes
- Polymerase is the protein that makes mRNA
- Transcription factors recruit/repress polymerase



“Gene Regulatory Circuits”



Mechanisms are complicated, and only partially known

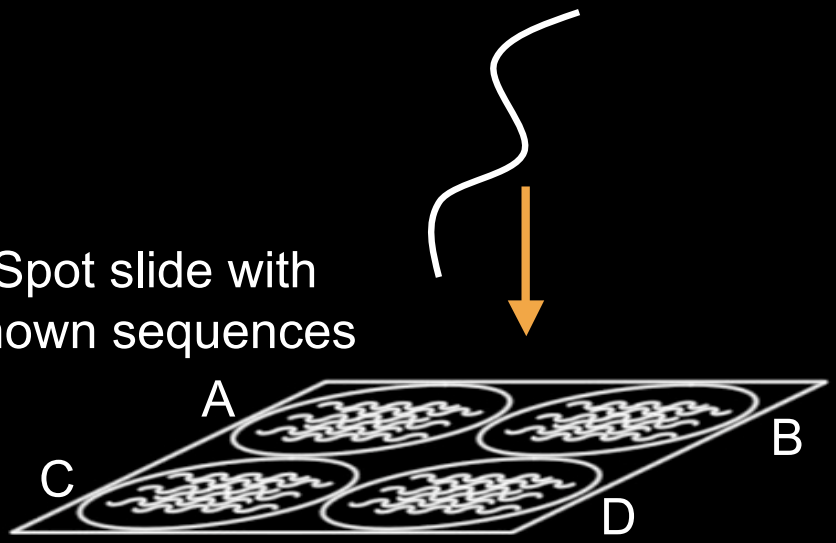
Measuring Transcription

- Differences in transcription are important
  - Organismal differences
  - Disease, immune response, etc.
- How can we measure transcriptional changes?
  - Count the number of mRNA created for each gene
  - Microarrays are a method to do this

Microarray Methodology

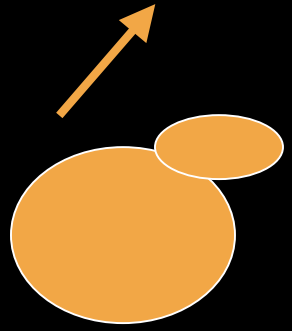
Microarray Methodology

Spot slide with known sequences



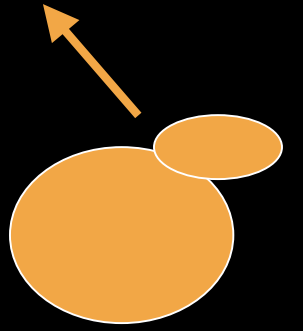
Microarray Methodology

reference mRNA



Reference sample

test mRNA



Test cells

Spot slide with
known sequences

Microarray Methodology

reference mRNA

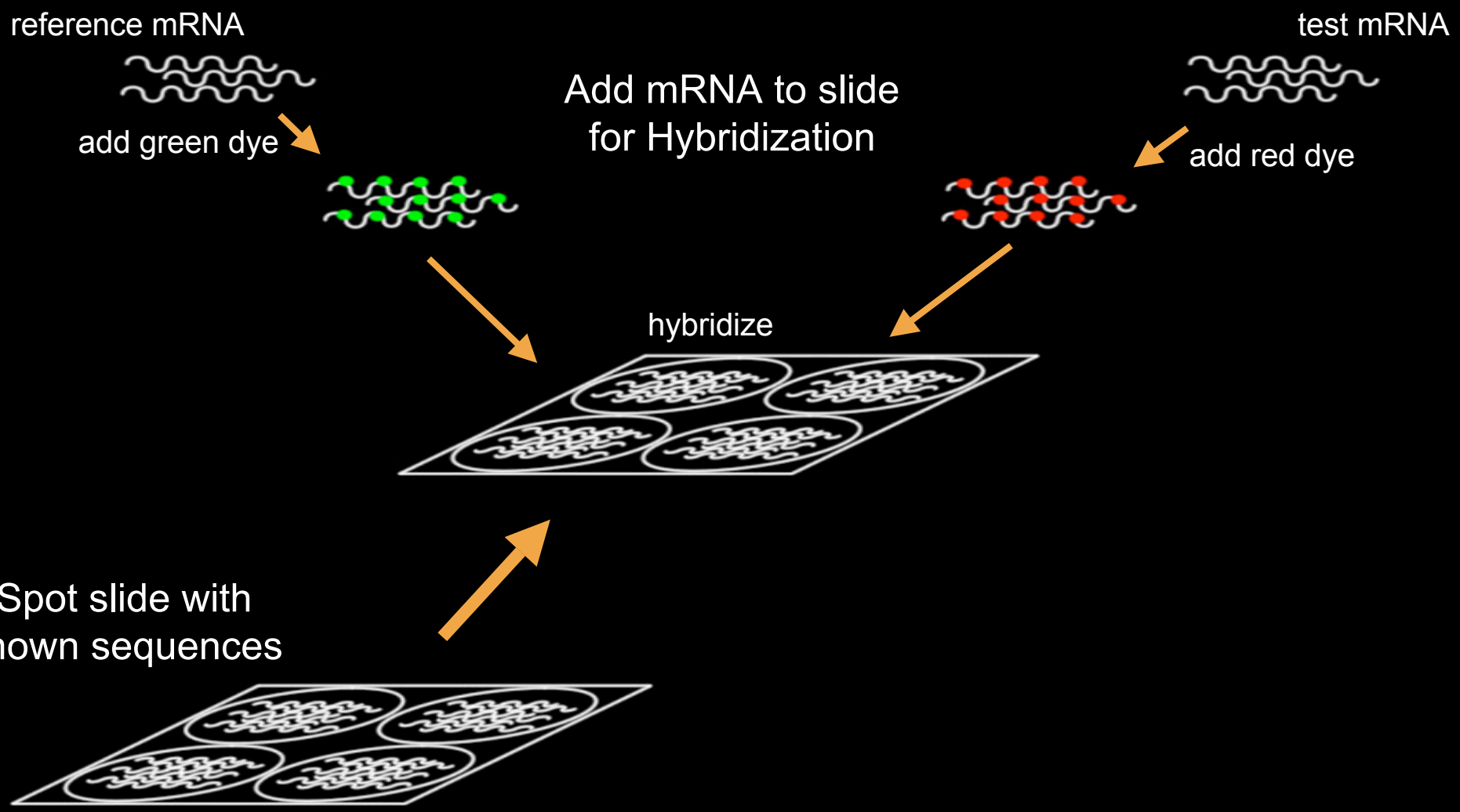
add green dye

test mRNA

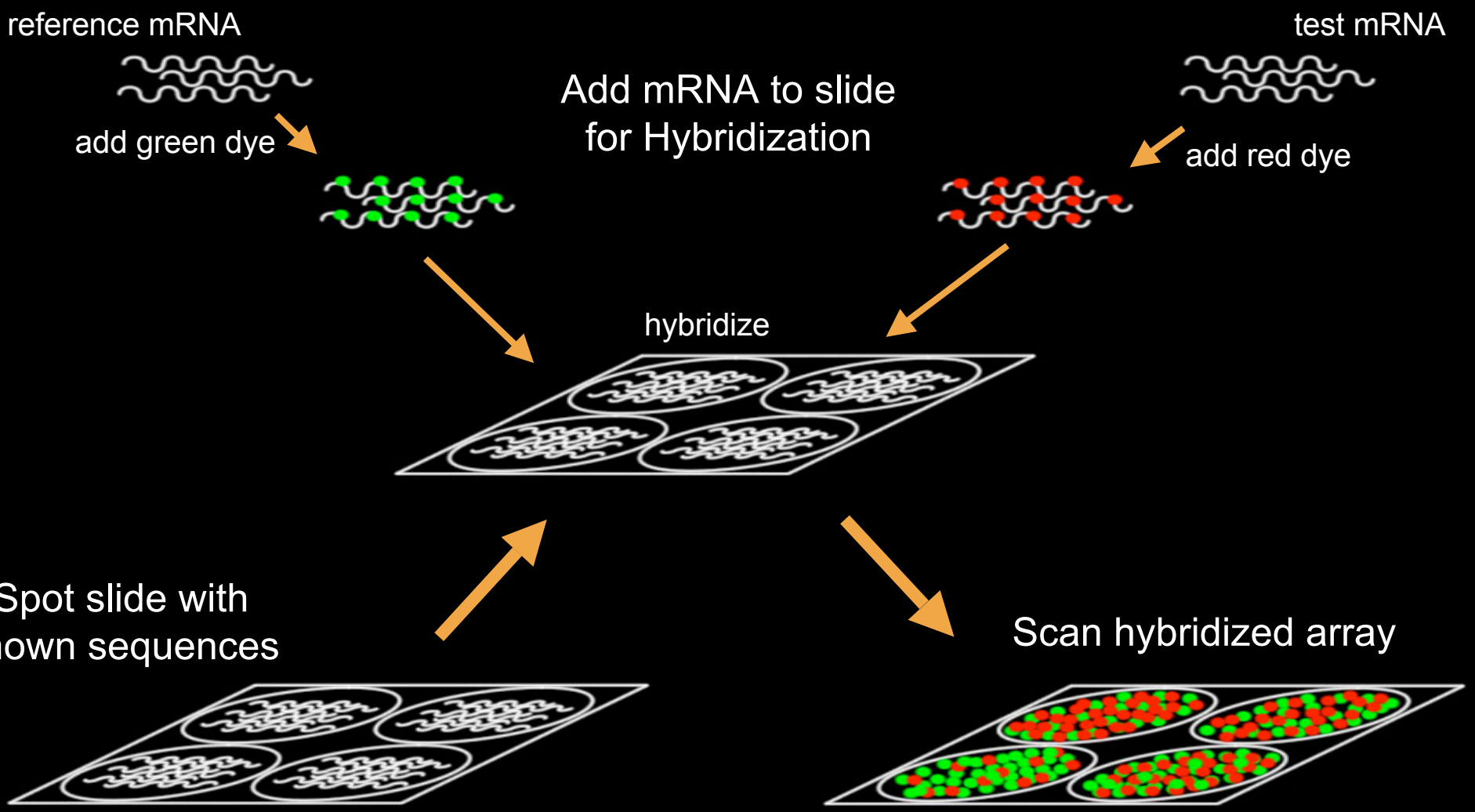
add red dye

Spot slide with known sequences

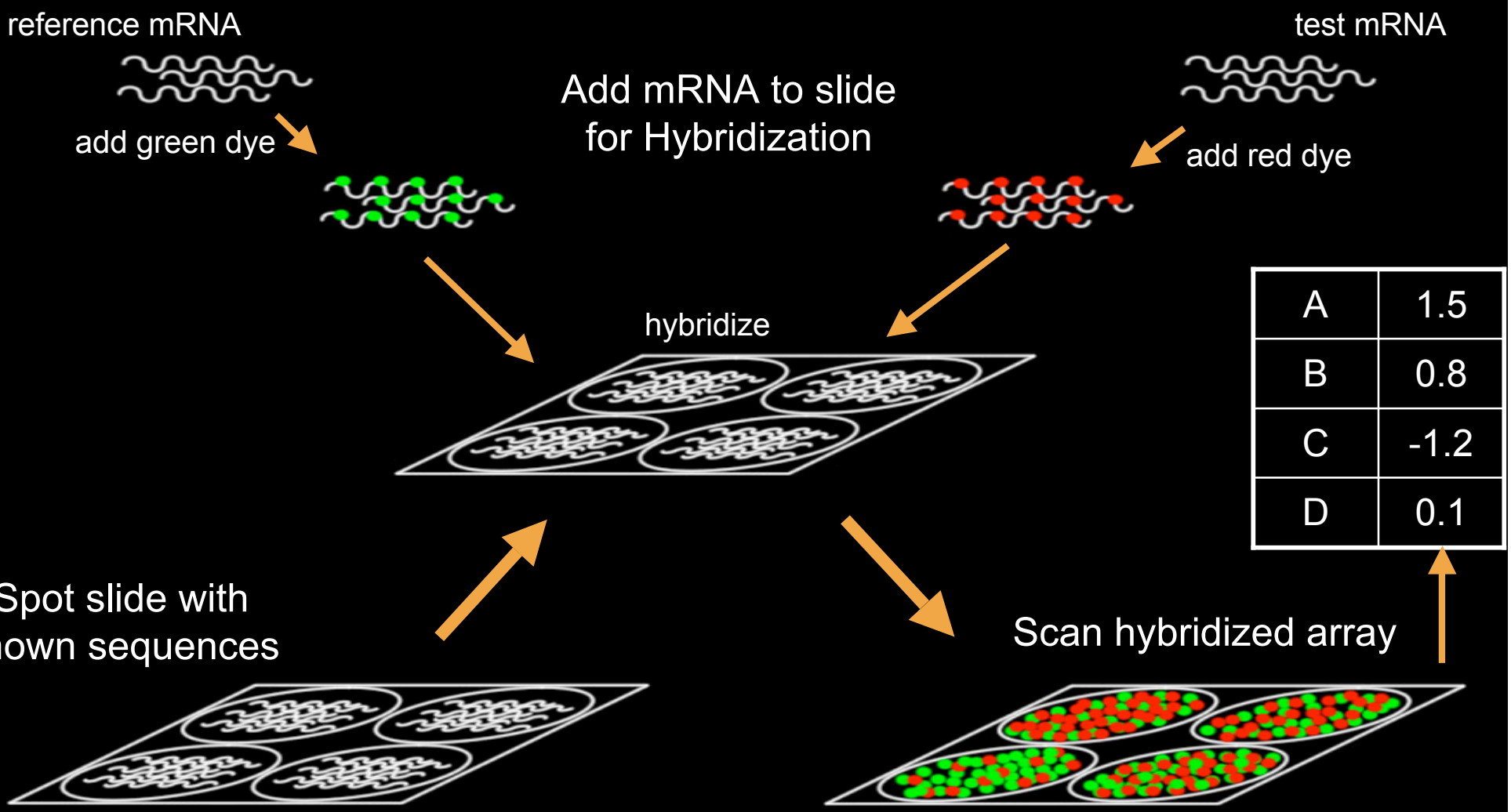
Microarray Methodology



Microarray Methodology

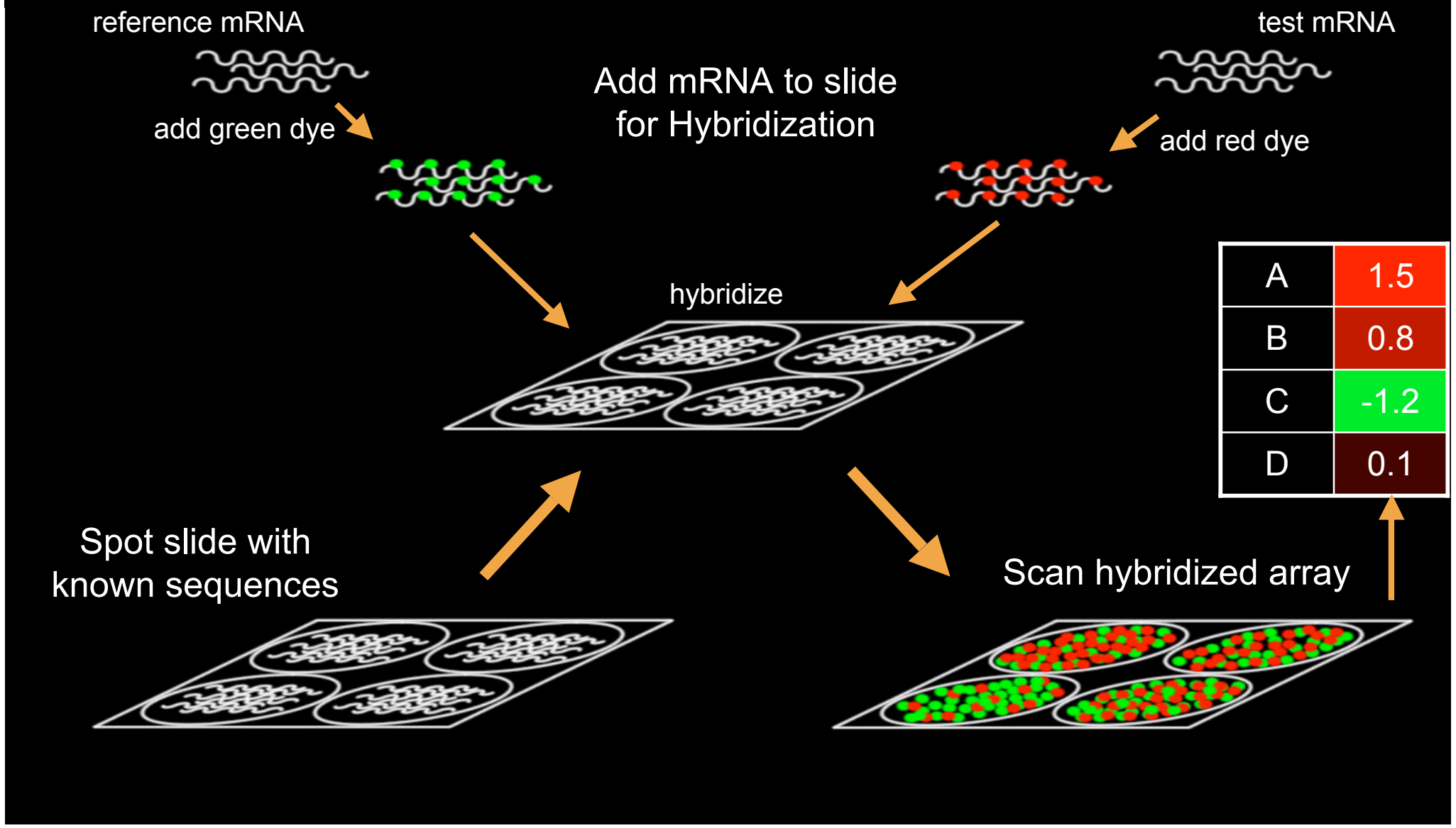


Microarray Methodology

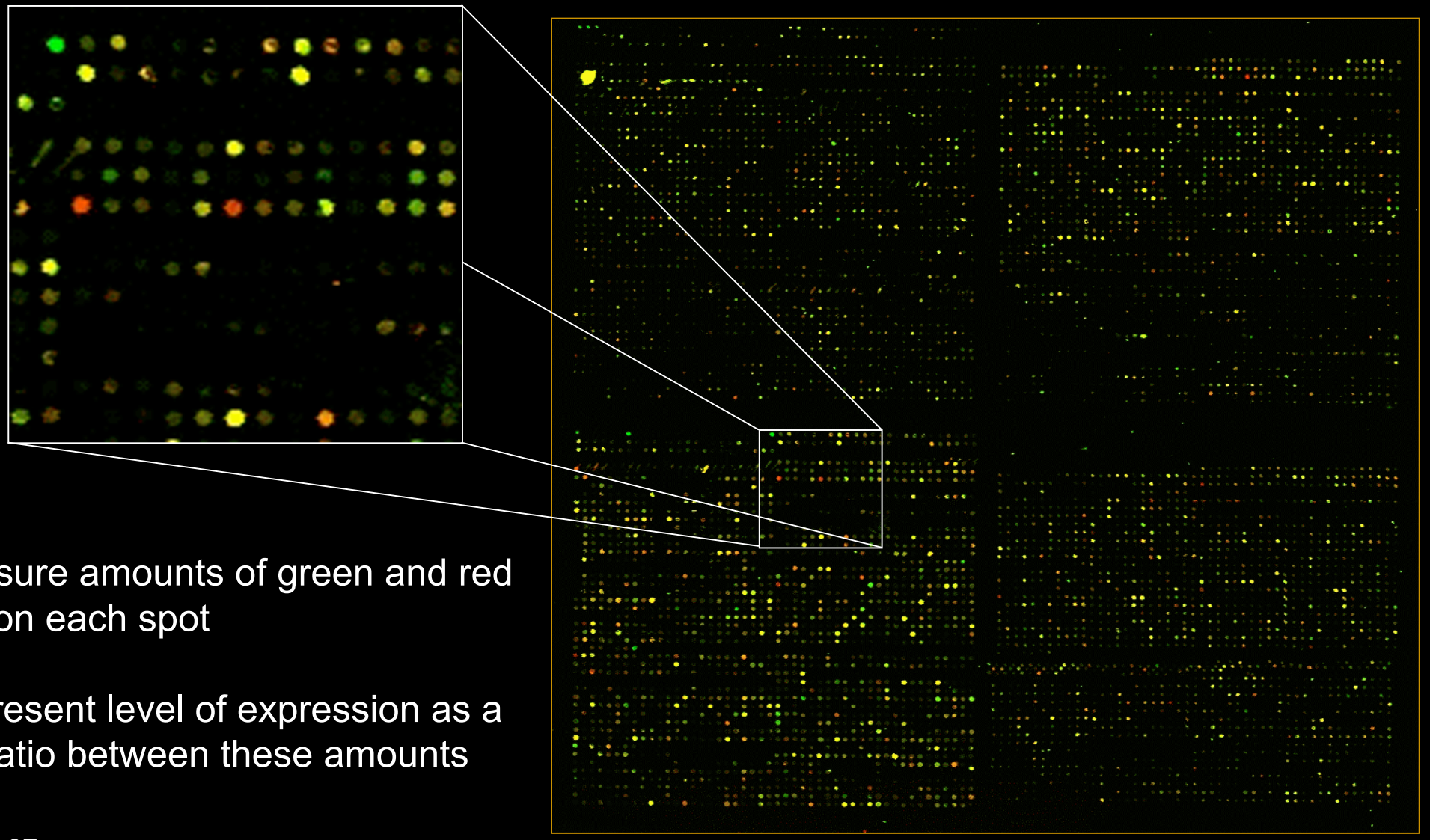


A	1.5
B	0.8
C	-1.2
D	0.1

Microarray Methodology



Microarray Outputs



Measure amounts of green and red dye on each spot

Represent level of expression as a log ratio between these amounts

Raw Image from Spellman et al., 98

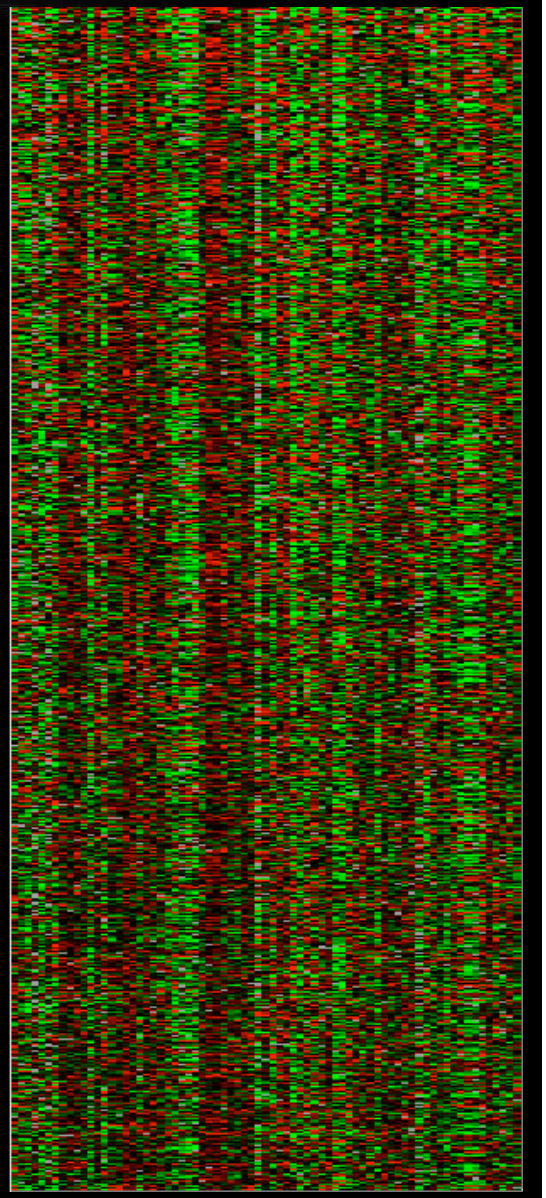
Microarray Outputs

Experiments →

Genes ↓

	alpha0	alpha7	alpha14	alpha21
YAL001C	-0.15	-0.15	-0.21	0.17
YAL002W	-0.11	0.1	0.45	1.52
YAL003W	-0.14	-0.71	0.1	-0.32
YAL004W	-0.02	-0.48	-0.11	0.12
YAL005C	-0.05		-0.47	-0.06
YAL007C	-0.6	-0.45	-0.13	0.35
YAL008W	-0.28	-0.22	-0.06	
YAL009W	-0.03	-0.27	0.17	-0.12
YAL010C	-0.05	0.13	0.13	-0.21
YAL011W	-0.31	-0.43	-0.3	-0.23
YAL012W	0.02	-0.33	-0.49	-0.3
YAL013W	-0.36	-0.19		-0.32
YAL014C	-0.1	-0.15	-0.01	-0.25
YAL015C		-0.01	0.12	-0.23
YAL016W	0.06	0.01	0.17	-0.14
YAL017W	-0.4	-0.22	0.19	-0.2
YAL018C	0.46	0.28	0.16	-1.72
YAL019W	-0.24	-0.95	-0.23	0.12

→



Data Analysis

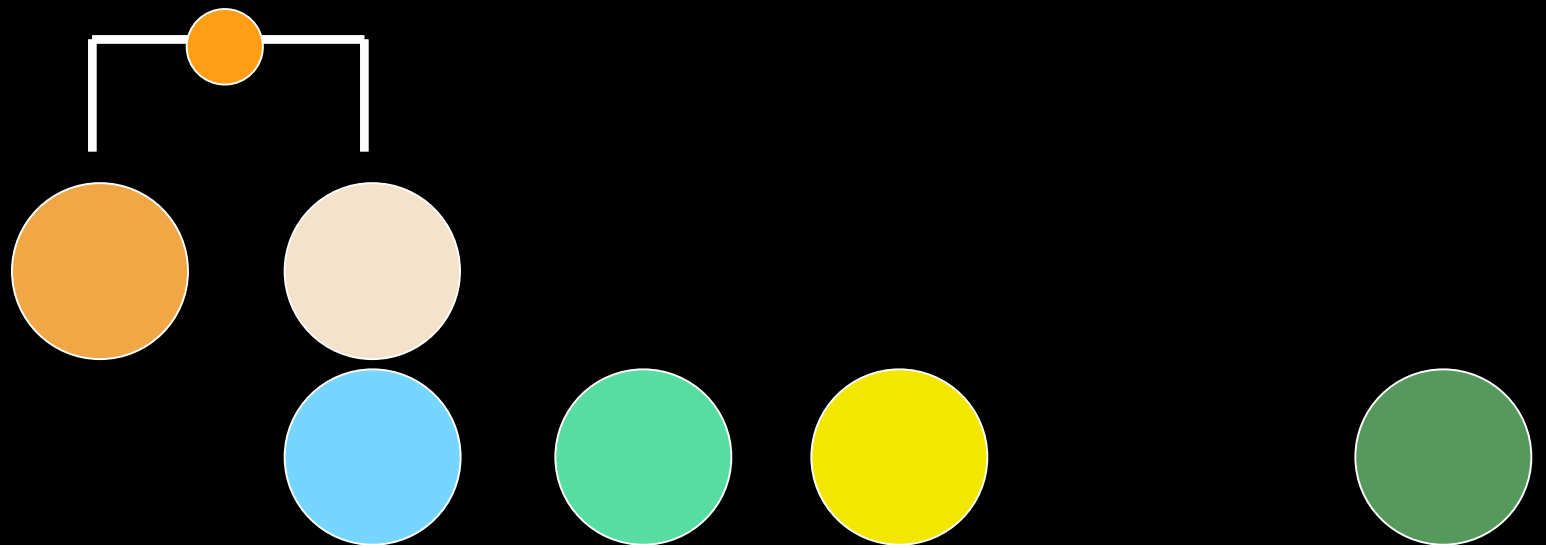
- Lots and lots of data
  - Thousands of genes x hundreds of conditions
- Traditional biology is on the scale of tens of genes in a handful of conditions
- Computer Science regularly deals with large amounts of data
  - Internet, digital animation, cryptography

Running Times

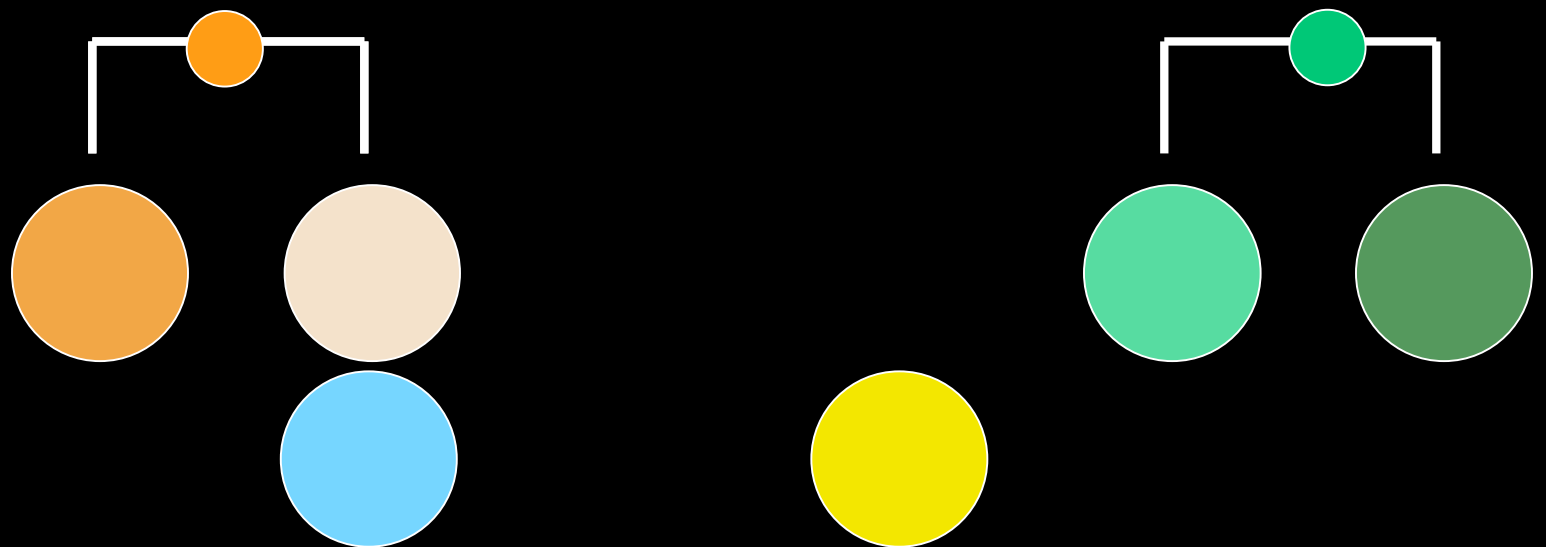
n	10	100	1000	10000
n log n	10	200	3000	40000
n^2	100	10000	1e6	1e8
n^3	1000	1e6	1e9	1e12
2^n	1024	~1e30	~1e301	$\sim \infty$

Hierarchical clustering

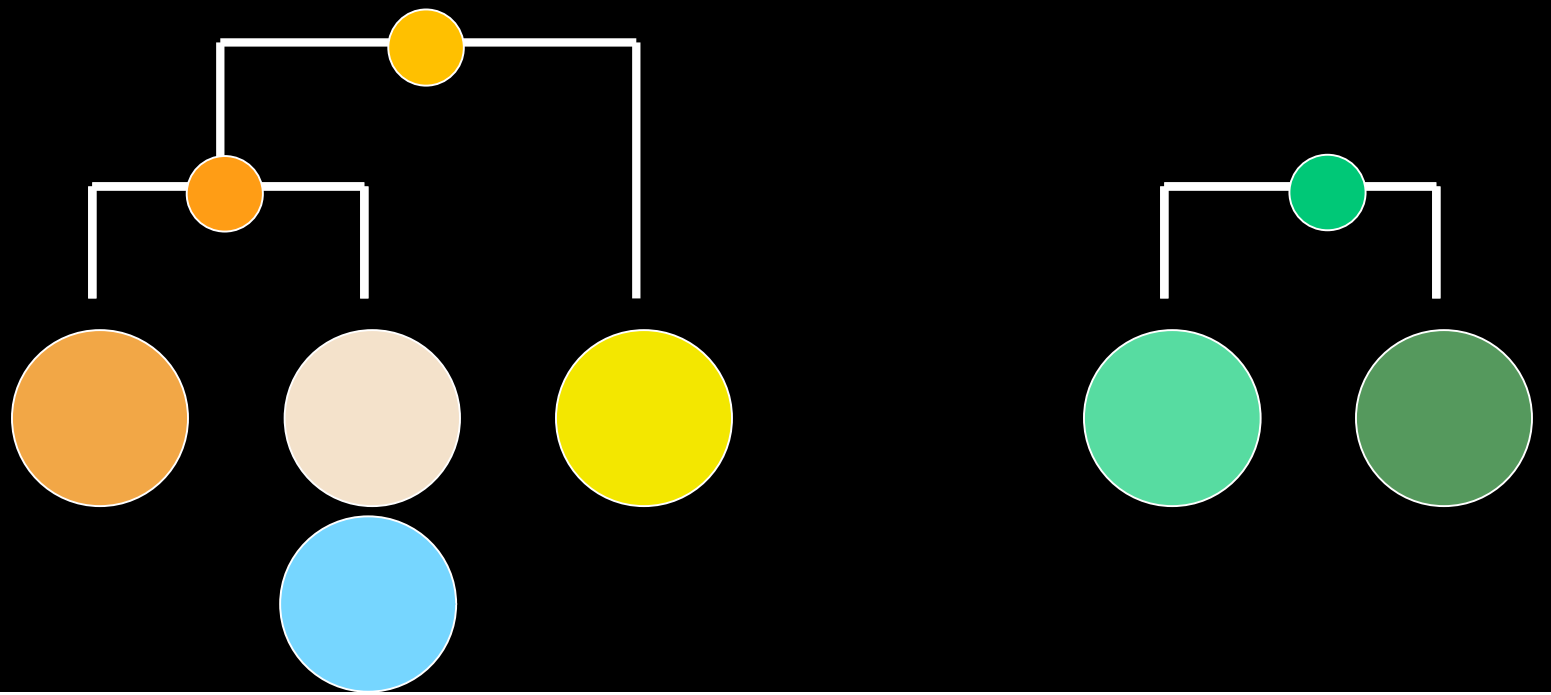
Hierarchical clustering



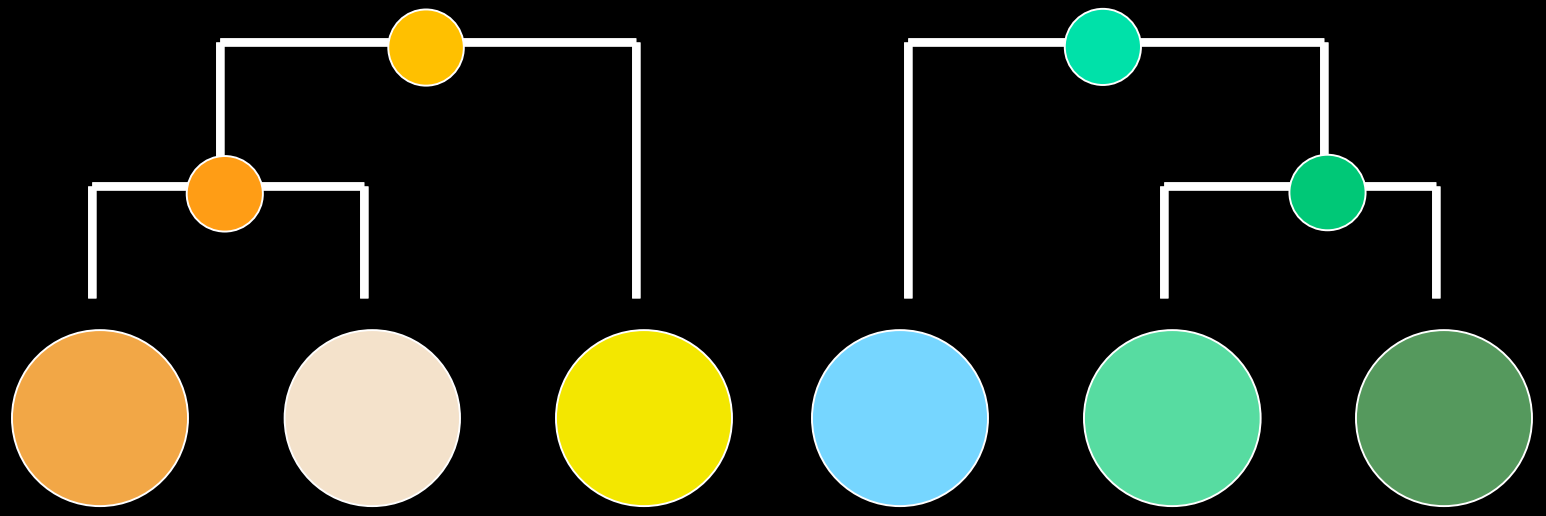
Hierarchical clustering



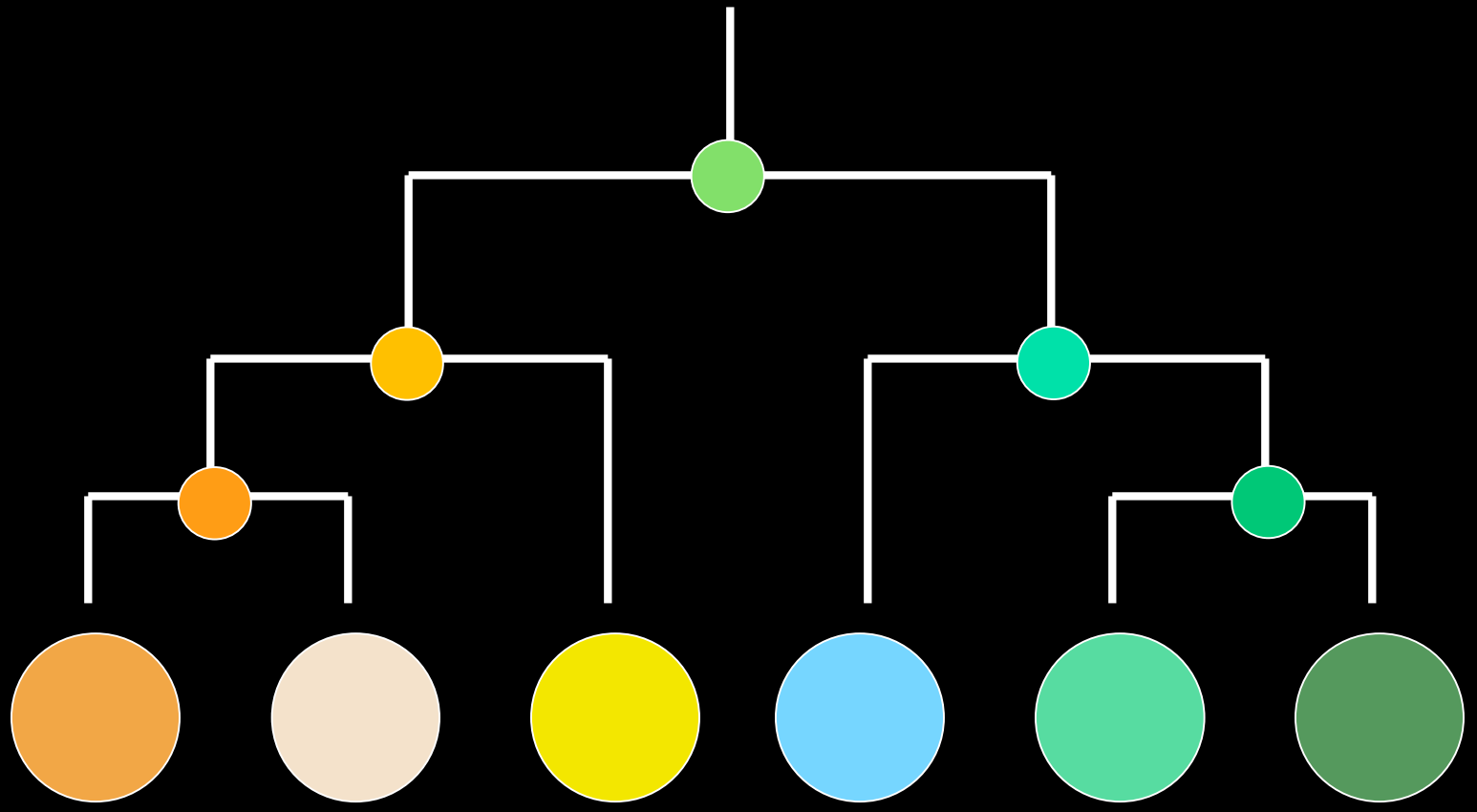
Hierarchical clustering



Hierarchical clustering

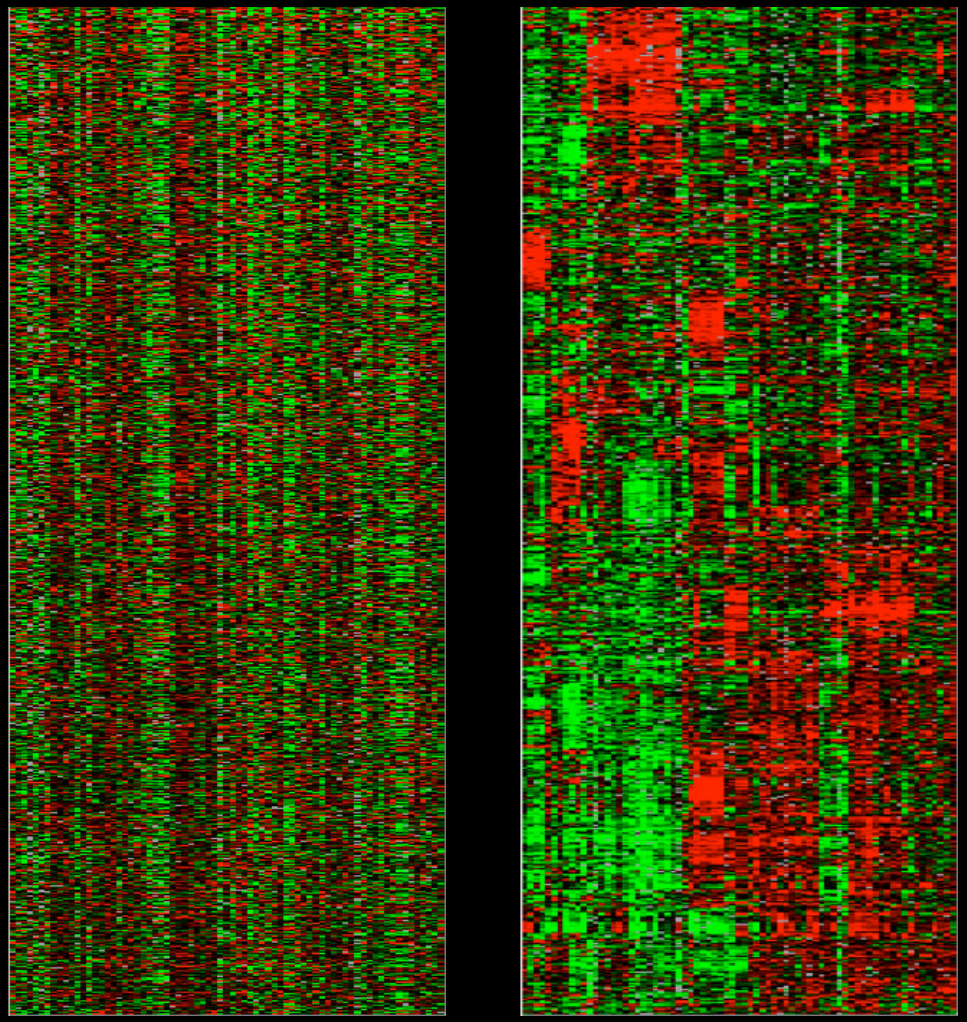


Hierarchical clustering



Clustering Analysis

- Distance metrics
  - Euclidean
  - Pearson
  - Spearman
  - …
- Algorithms
  - Hierarchical
  - K-means
  - SOM
  - …

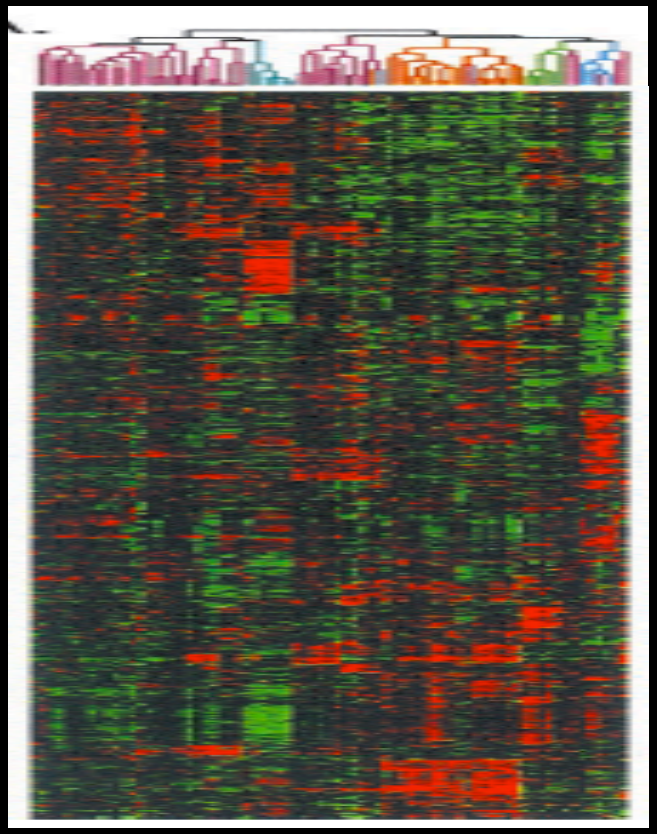


Cancer treatment

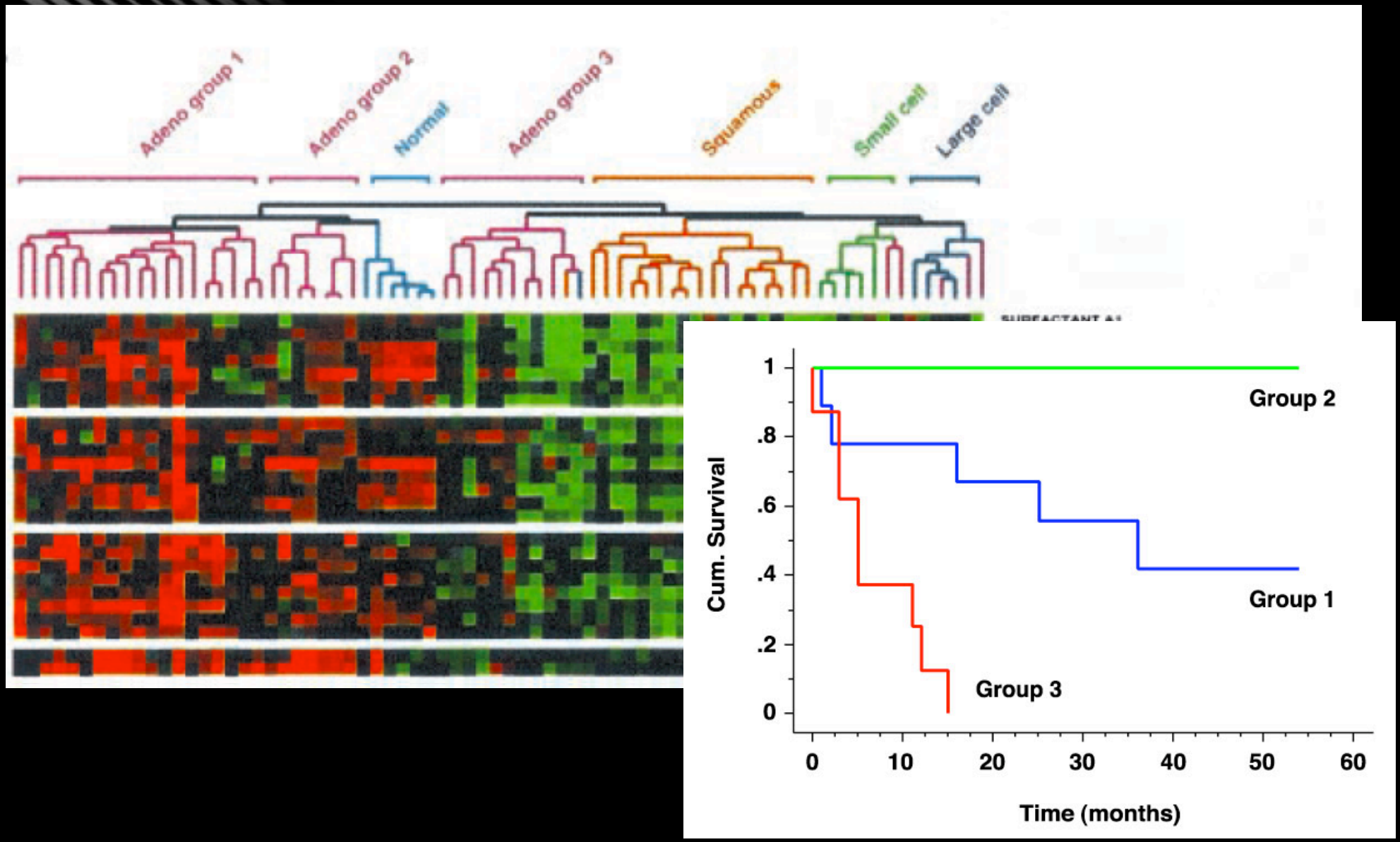
- Oncologist examines biopsies under microscope
  - Different cancer types look different
  - Treatment differs by type
- Some cancers that look the same have very different clinical outcomes
- Is there a difference we can't see?

Cancer microarrays

- Each condition is a patient biopsy
- Hierarchically cluster together genes and patients



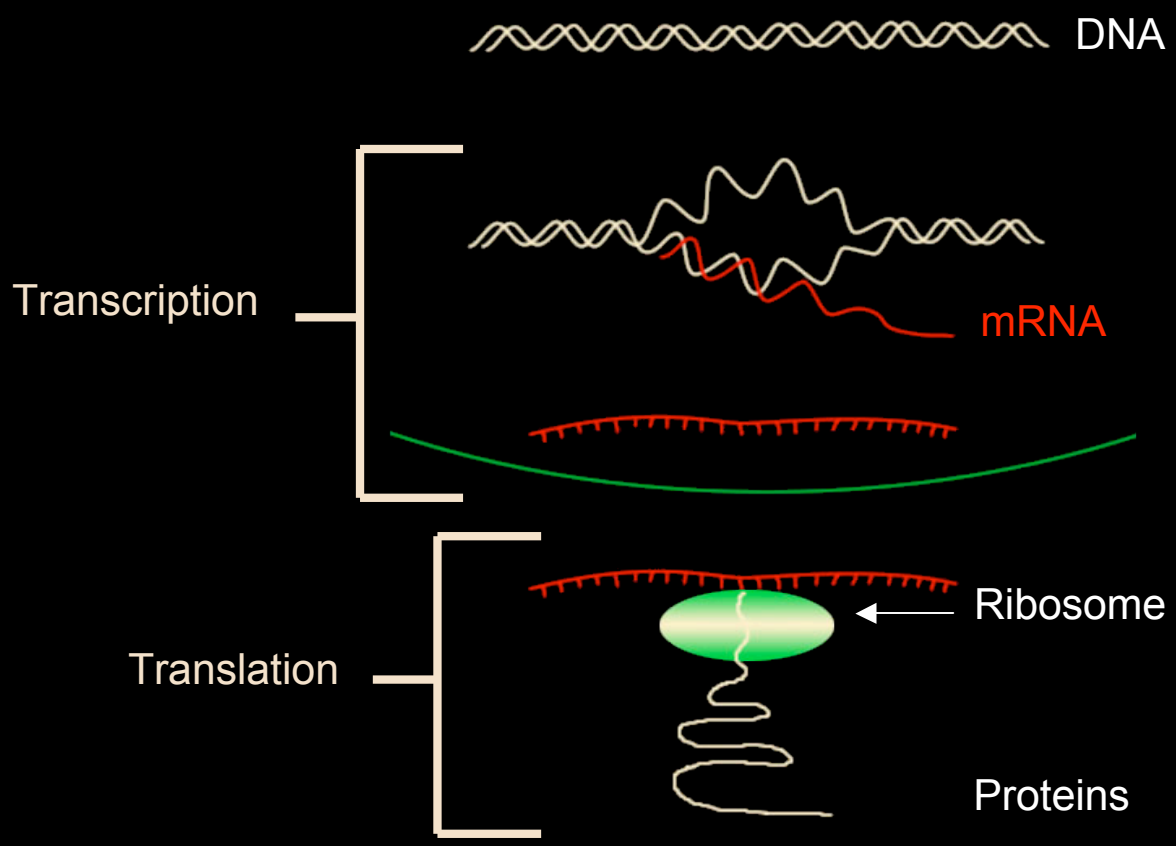
Cancer clusters



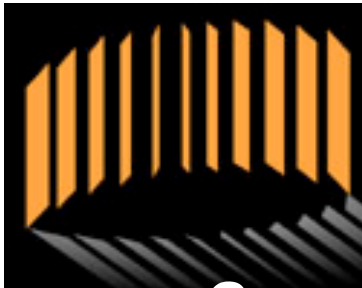
Personalized medicine

- Microarrays show more specific information
- Whole-genome level of analysis
- Individual arrays can show a doctor how a specific patient's cells behave

How Can CS Help?



Systems Biology



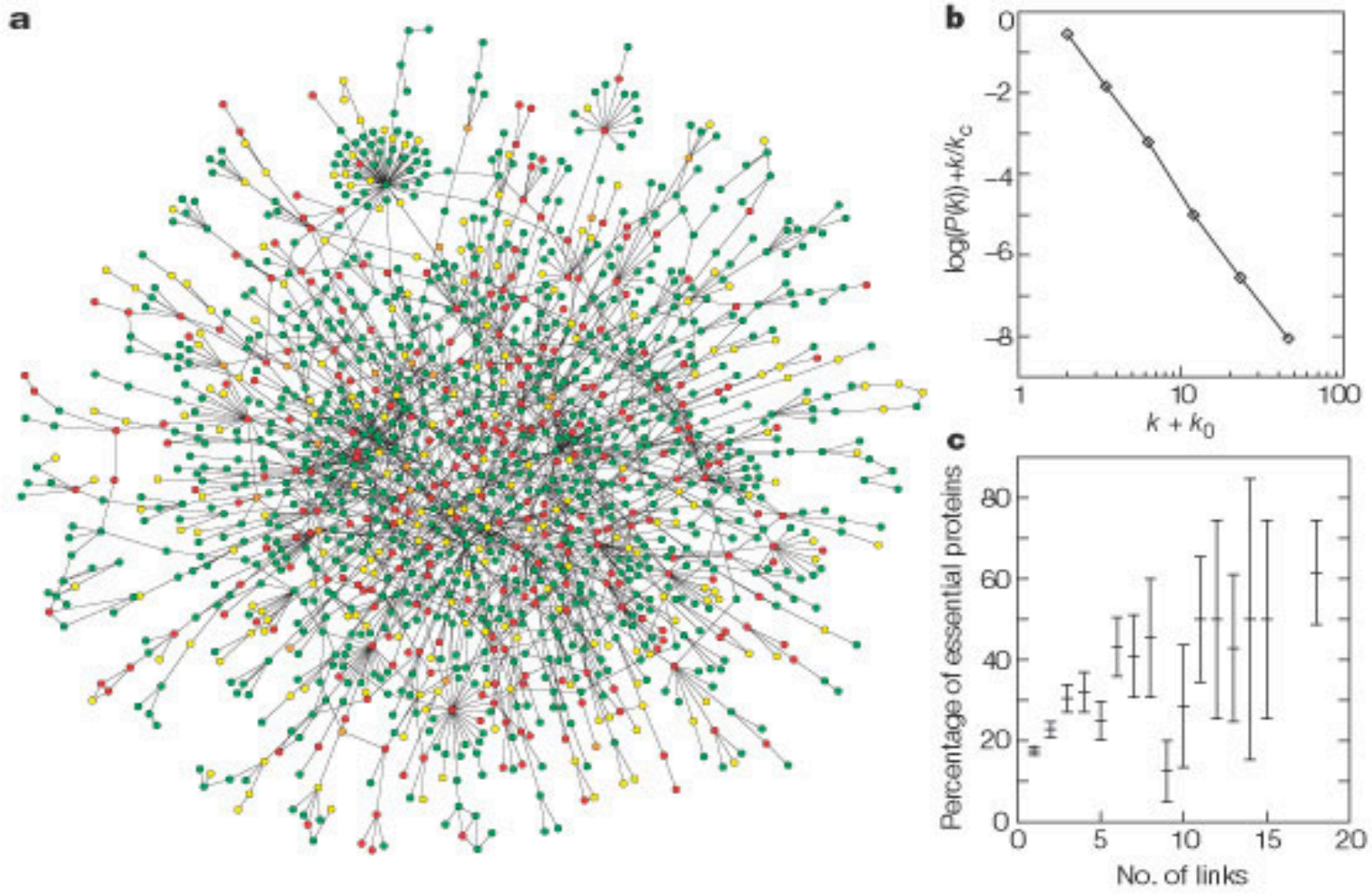
Systems-level challenges

- **Gene function annotation – what does a gene do**
  - ~30,000 genes in the human genome => systems-level approaches necessary
  - A modern human microarray experiment produces ~500,000 data points => computational analysis & visualization necessary
  - Many high-throughput functional technologies => computational methods necessary to integrate the data
- **Biological networks – how do proteins interact**
  - Large amounts of high-throughput data => computation necessary to store and analyze it
  - Data has variable specificity => computational approaches necessary to separate reliable conclusions from random coincidences
- **Comparative genomics – comparing data between organisms**
  - Need to map concepts across organisms on a large scale => practically impossible to do by hand
  - High amount of variable quality data => computational methods needed for integration, visualization, and analysis
  - Data often distributed in databases across the globe, with variable schemas etc => data storage and consolidation methods needed

Biological networks

- Interaction maps (no directions)
- Pathway models (dynamic or static)
- Metabolic networks
- Genetic regulatory networks

Yeast Interaction Network



Hawoong Jeong et al. Oltvai Centrality and lethality of protein networks. Nature 411, 41-42 (2001)

What are functions of genes?

- Signal transduction: sensing a physical signal and turning into a chemical signal
- Structural support: creating the shape and pliability of a cell or set of cells
- Enzymatic catalysis: accelerating chemical transformations otherwise too slow.
- Transport: getting things into and out of separated compartments

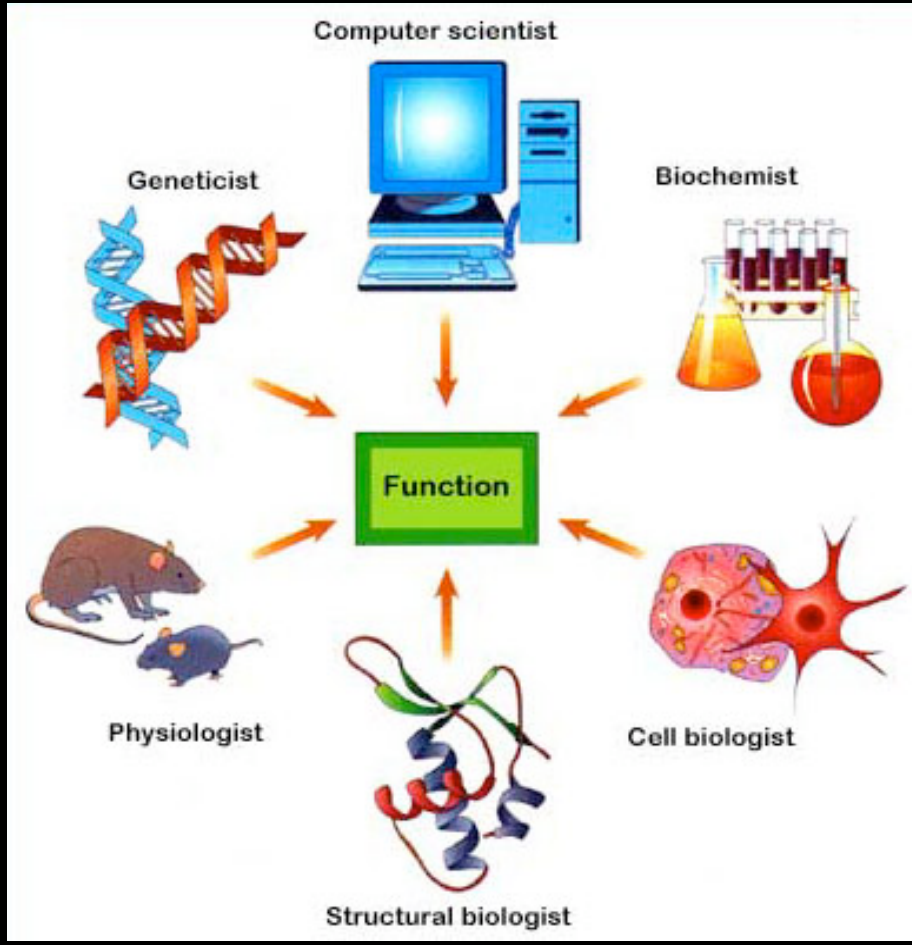
What are functions of genes?

- Movement: contracting in order to pull things together or push things apart.
- Transcription control: deciding when other genes should be turned ON/OFF
- Trafficking: affecting where different elements end up inside the cell

Function

- To study WHAT proteins DO, HOW they INTERACT, and HOW they are REGULATED, need data beyond genomic sequence

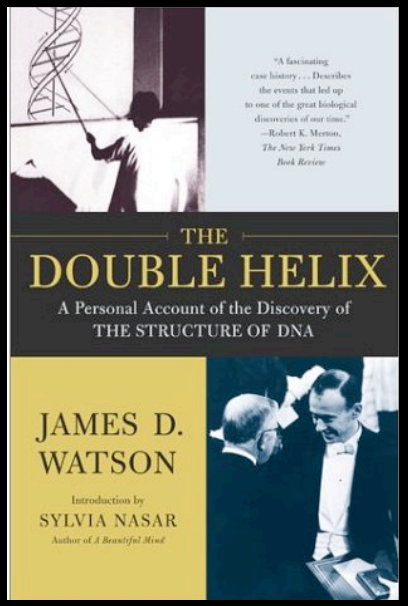
- Genomics/Bioinformatics is fundamentally a COLLABORATIVE and MULTIDISCIPLINARY effort

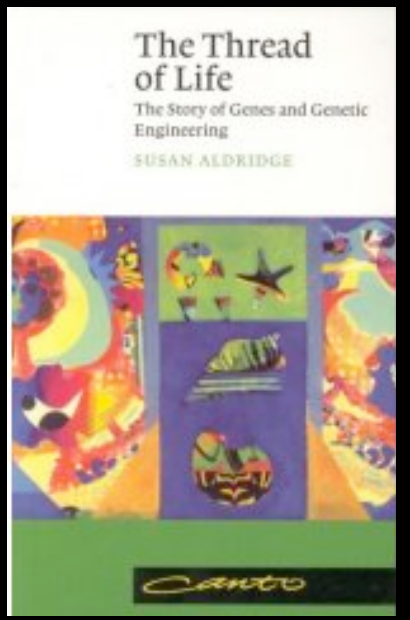


BioInformatics

- Many, many problems in modern biology benefit from computer science
- Very young field, just beginning to see practical applications and results
- CS already a vital part of any biology laboratory
- Much, much more to be done

Questions?


"A fascinating case history . . . Describes the events that led up to one of the great biological discoveries of our time."
—Robert K. Merton, The New York Times Book Review
THE
DOUBLE HELIX
A Personal Account of the Discovery of THE STRUCTURE OF DNA
JAMES D. WATSON
Introduction by
SYLVIA NASAR
Author of A Beautiful Mind


The Thread of Life
The Story of Genes and Genetic Engineering
SUSAN ALDRIDGE
Canto