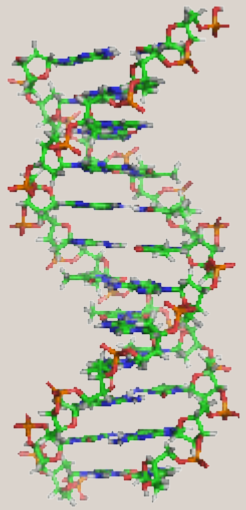
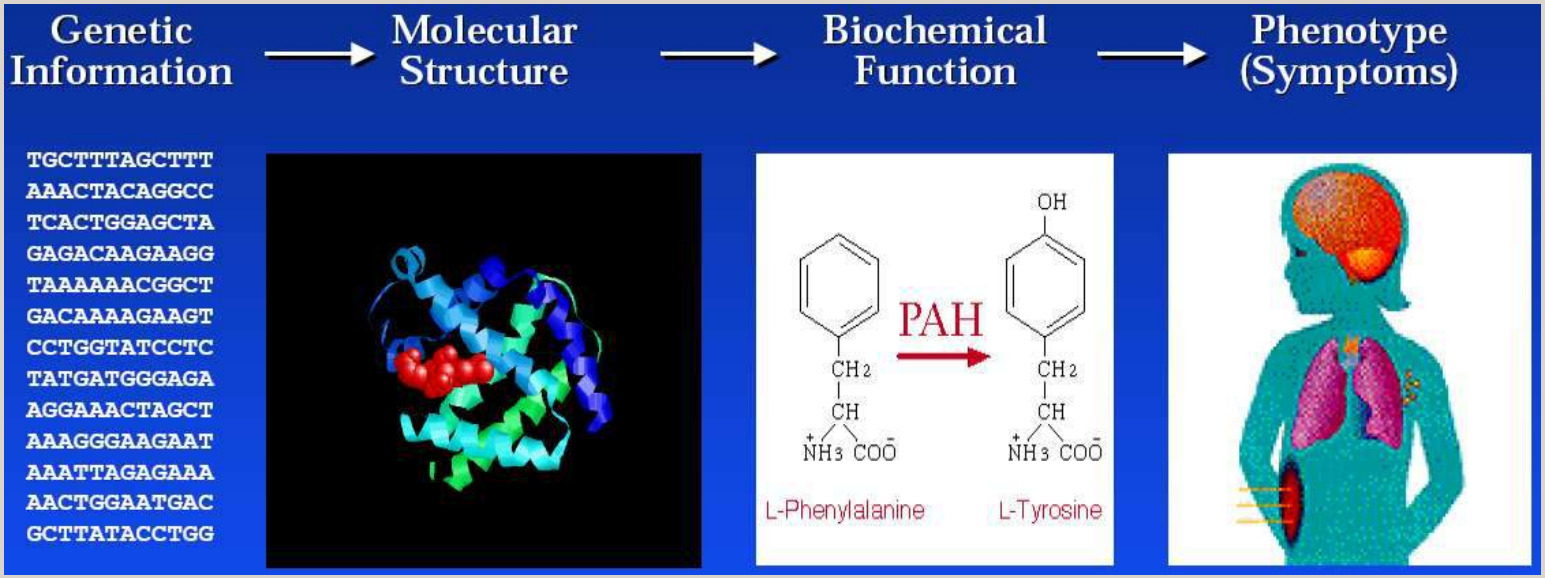
Introduction to Bioinformatics

Dan Lopresti
Associate Professor
Office PL 404B
dal9@lehigh.edu



Motivation

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

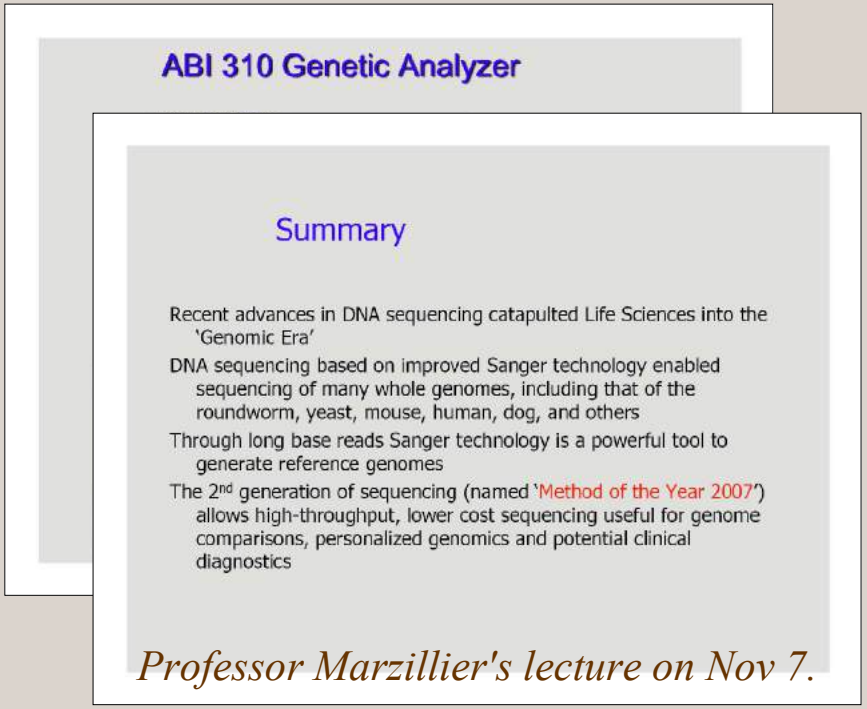


By developing techniques for analyzing sequence data and related structures, we can attempt to understand molecular basis of life.

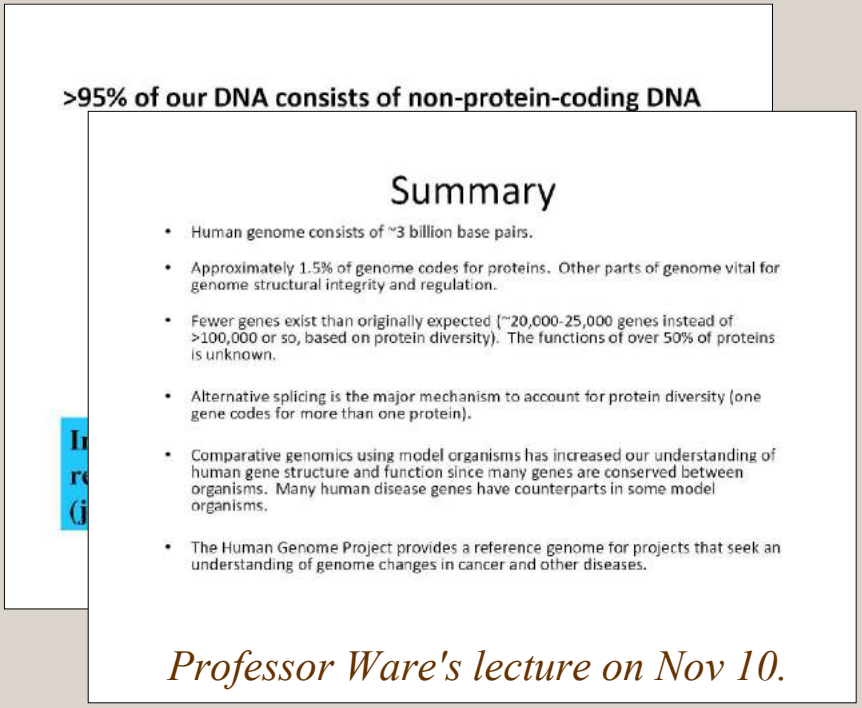
http://cmgm.stanford.edu/biochem218/

Before We Get Going

Recall your recent lectures by Professors Marzillier and Ware who presented biological background:



Professor Marzillier's lecture on Nov 7.

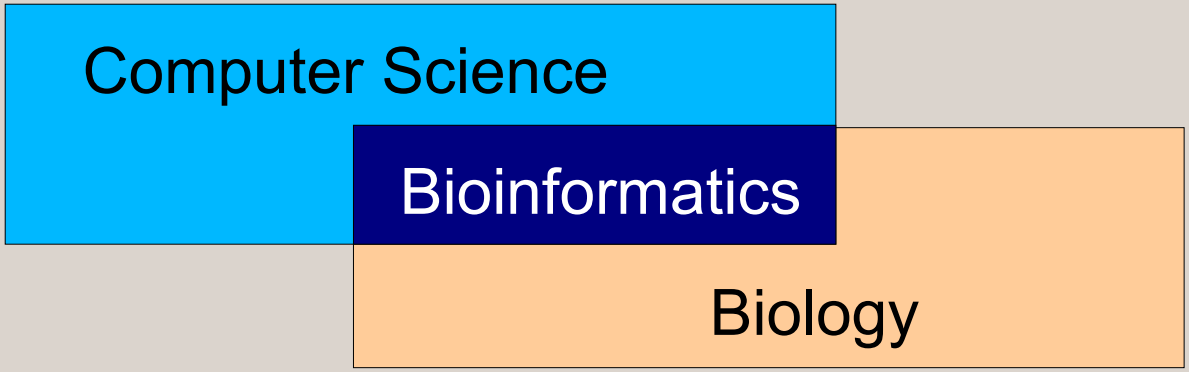


Professor Ware's lecture on Nov 10.

Today I'll focus on the related computational questions.

Bioinformatics

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



Why is it interesting?

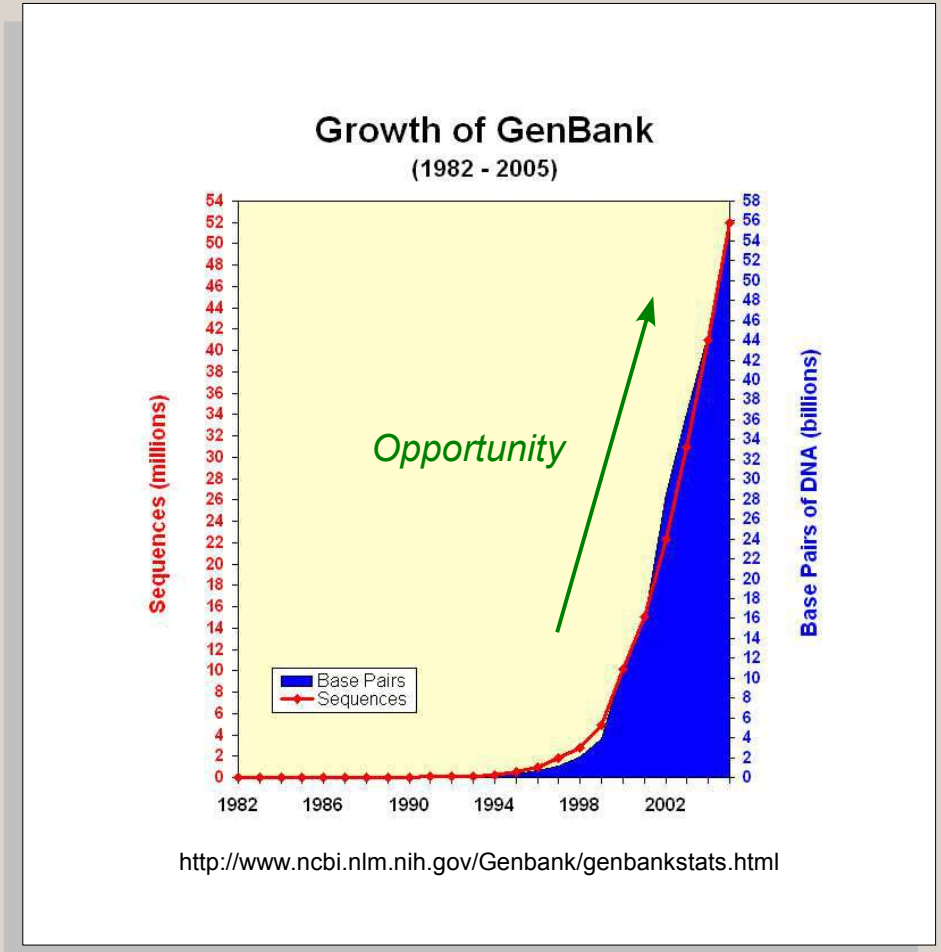
- Important problems.
- Massive quantities of data.
- Desperate need for efficient solutions.
- Success is rewarded.

Data Explosion

Our genetic identity is encoded in long molecules made up of four basic units, the nucleic acids:

(1) *Adenine*,
(2) *Cytosine*,
(3) *Guanine*,
(4) *Thymine*.

To first approximation, DNA is a language over a four character alphabet, $\{A, C, G, T\}$.



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Genomes

Complete set of chromosomes that determines an organism is known as its *genome*.

Us

Mus musculus

GenBank Release 121.0 — December 15, 2000

Species	Haploid genome size	Bases	Entries
Homo sapiens	3,400,000,000	6,702,881,570	3,918,724
Mus musculus	3,454,200,000	1,291,602,139	2,456,194
Drosophila melanogaster	180,000,000	487,561,384	166,554
Arabidopsis thaliana	100,000,000	242,674,129	181,388
Caenorhabditis elegans	100,000,000	203,544,197	114,553
Tetraodon nigroviridis	350,000,000	165,539,271	188,993
Oryza sativa			1,411
Rattus norvegicus			8,598
Bos taurus			9,473
Glycine max			1,802
Medicago truncatula			4,535
Trypanosoma brucei			1,334
Lycopersicon esculentum			7,112
Giardia intestinalis			4,328
Strongylocentrotus purpuratus			7,532
Entamoeba histolytica			9,938
Hordeum vulgare	—	44,489,692	57,779
Danio rerio	1,900,000,000	40,906,902	83,726
Zea mays	5,000,000,000	36,885,212	77,506
Saccharomyces cerevisiae	12,067,280	32,779,082	18,361

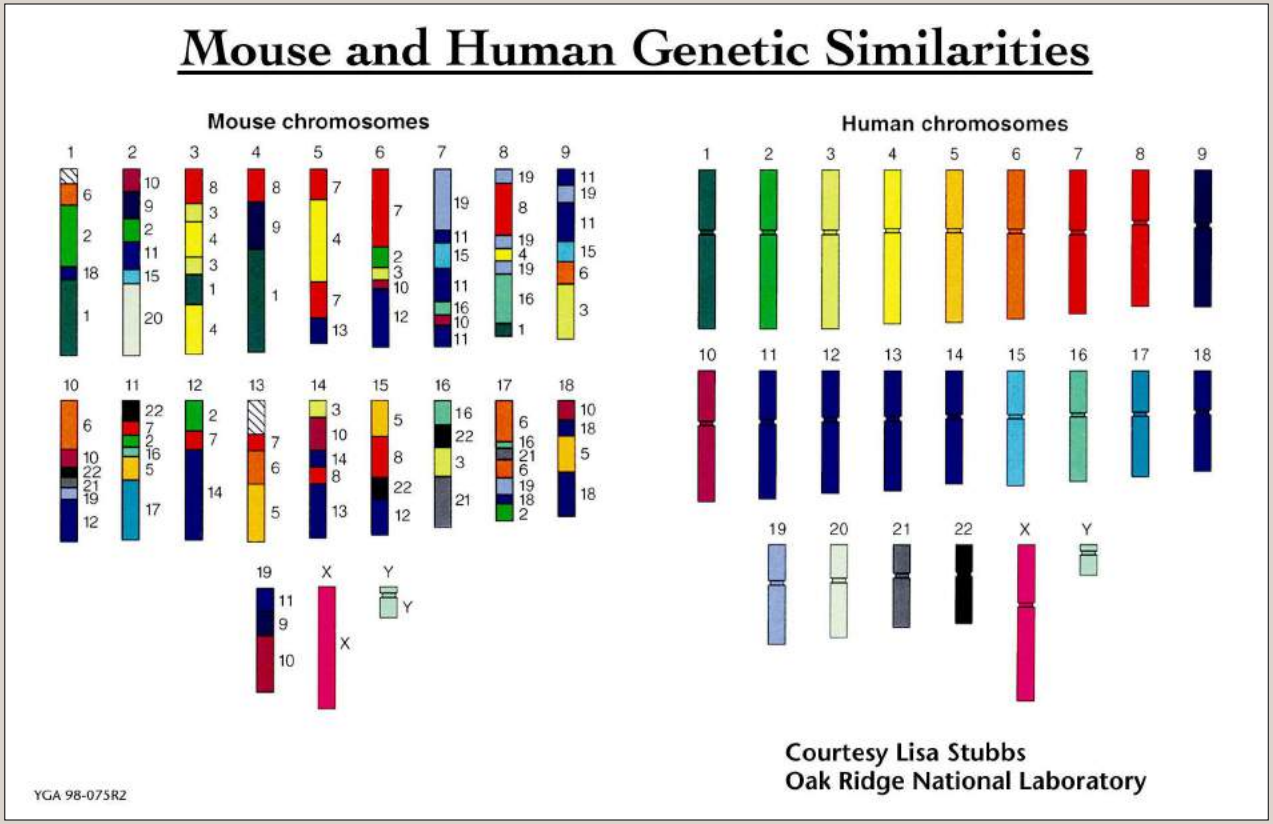
Conclusion: size does not matter!

(But you already knew this. ☺)

http://www.cbs.dtu.dk/databases/DOGS/
http://www.nsrl.ttu.edu/tmot1/mus_musc.htm
http://www.oardc.ohio-state.edu/seedid/single.asp?strID=324

Comparative Genomics

Recall this amazing diagram from Professor Ware's lecture:



How did we decipher these relationships?

http://www.ornl.gov/sci/techresources/Human_Genome/graphics/slides/ttmousehuman.shtml

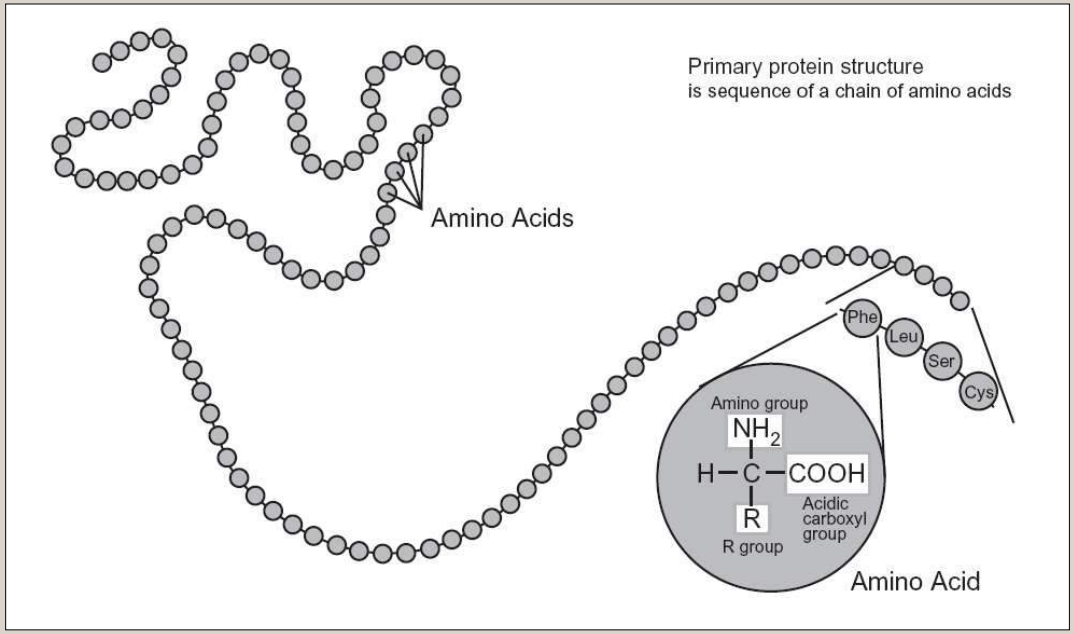
Algorithms are Central

An *algorithm* is a precisely-specified series of steps to solve a particular problem of interest.

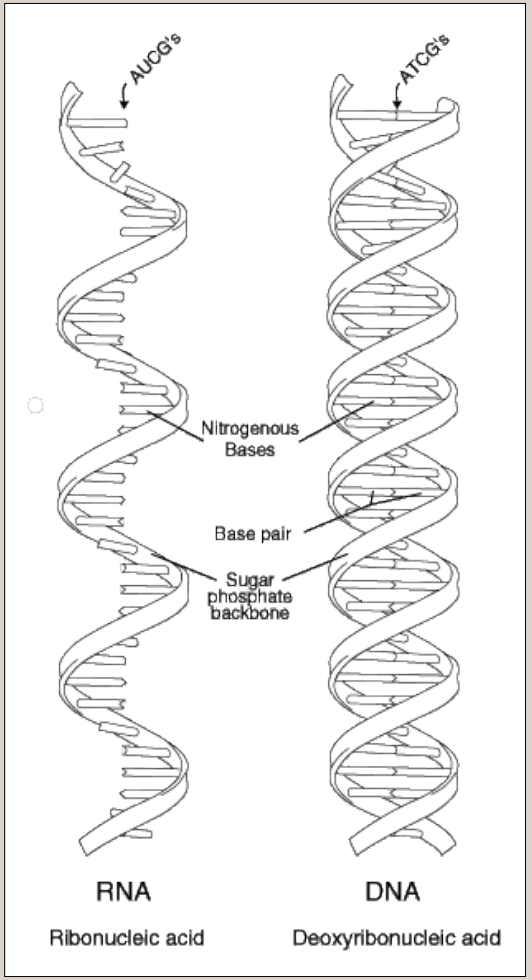
- Develop model(s) for task at hand.
- Study inherent computational complexity:
  - Can task be phrased as an optimization problem?
  - If so, can it be solved efficiently? Speed, memory, etc.
  - If we can't find a good algorithm, can we prove task is "hard"?
  - If known to be hard, is there approximation algorithm (one that works at least some of the time or comes close to optimal)?
- Conduct experimental evaluations (perhaps iterate above steps).

Sequence Nature of Biology

Macromolecules are chains of simpler molecules.



http://www.accessexcellence.org/AB/GG/aminoAcid.html



http://www.accessexcellence.org/AB/GG/rna.html

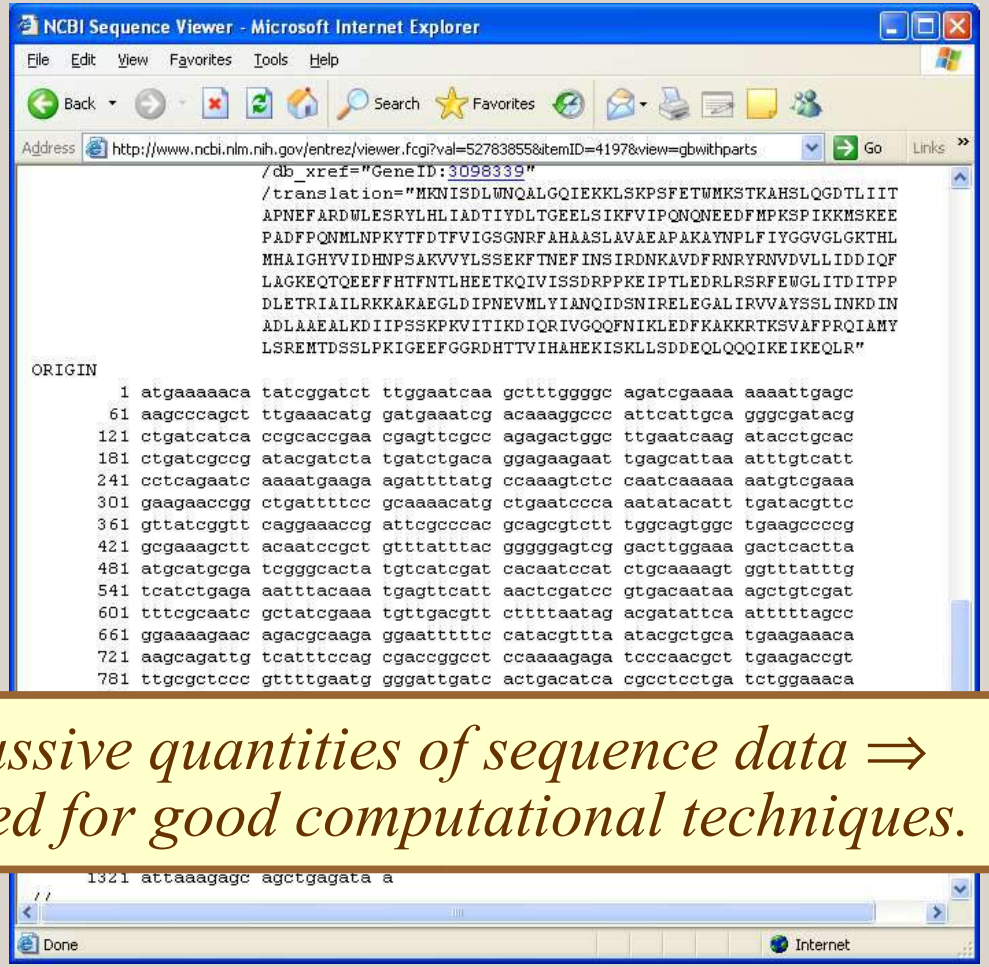
In the case of proteins, these basic building blocks are *amino acids*.

In DNA and RNA, they are *nucleotides*.

NCBI GenBank

National Center for Biotechnology Information (NCBI), which is branch of National Library of Medicine (NLM), which is branch of National Institutes of Health (NIH), maintains *GenBank*, a worldwide repository of genetic sequence data (all publicly available DNA sequences).

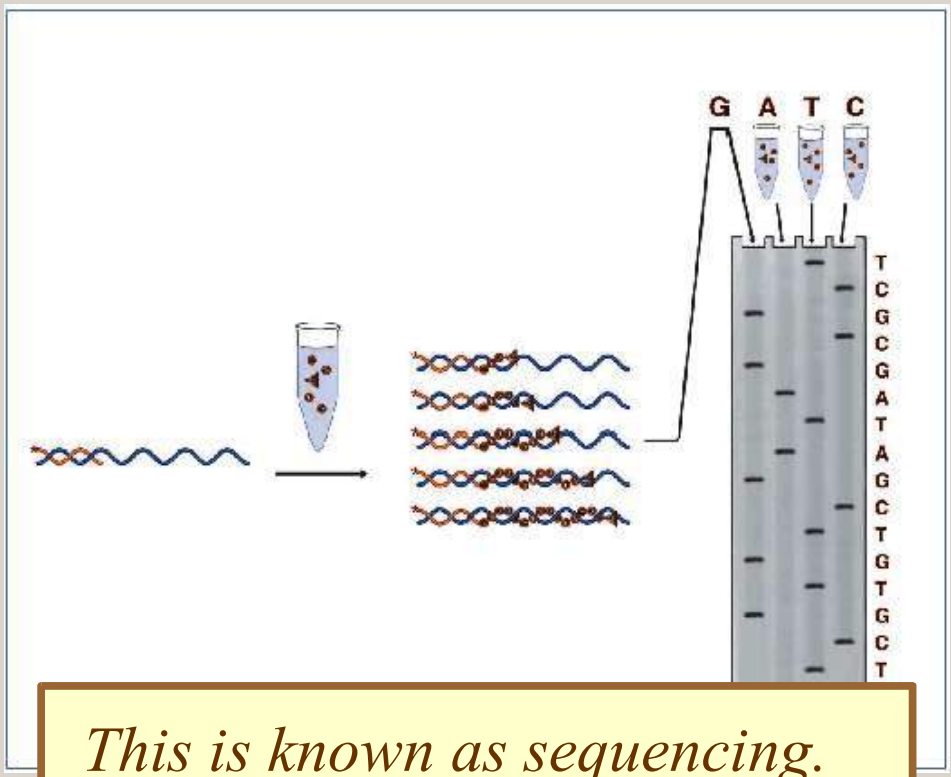
http://www.ncbi.nlm.nih.gov/



Massive quantities of sequence data ⇒ need for good computational techniques.

Reading DNA

Recall Professor Marzillier's lecture:



This is known as sequencing.

http://www.apelex.fr/anglais/applications/sommaire2/sanger.htm
http://www.iupui.edu/~wellsctr/MMIA/htm/animations.htm

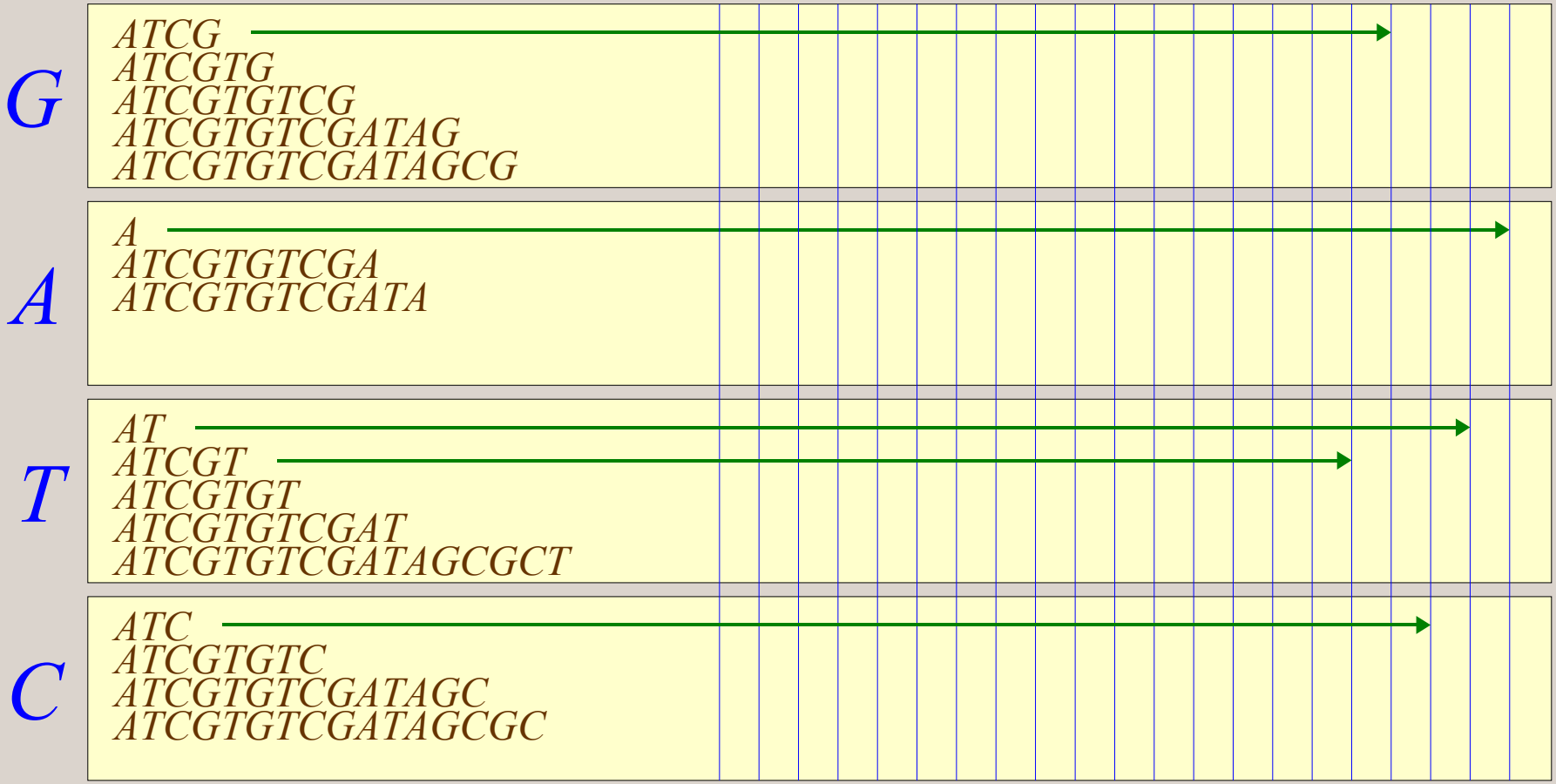
Gel electrophoresis is process of separating a mixture of molecules in a gel media by application of an electric field.

In general, DNA molecules with similar lengths will migrate same distance.

Make DNA fragments that end at each base: *A*, *C*, *G*, *T*. Then run gel and read off sequence: *ATCGTG* ...

Reading DNA

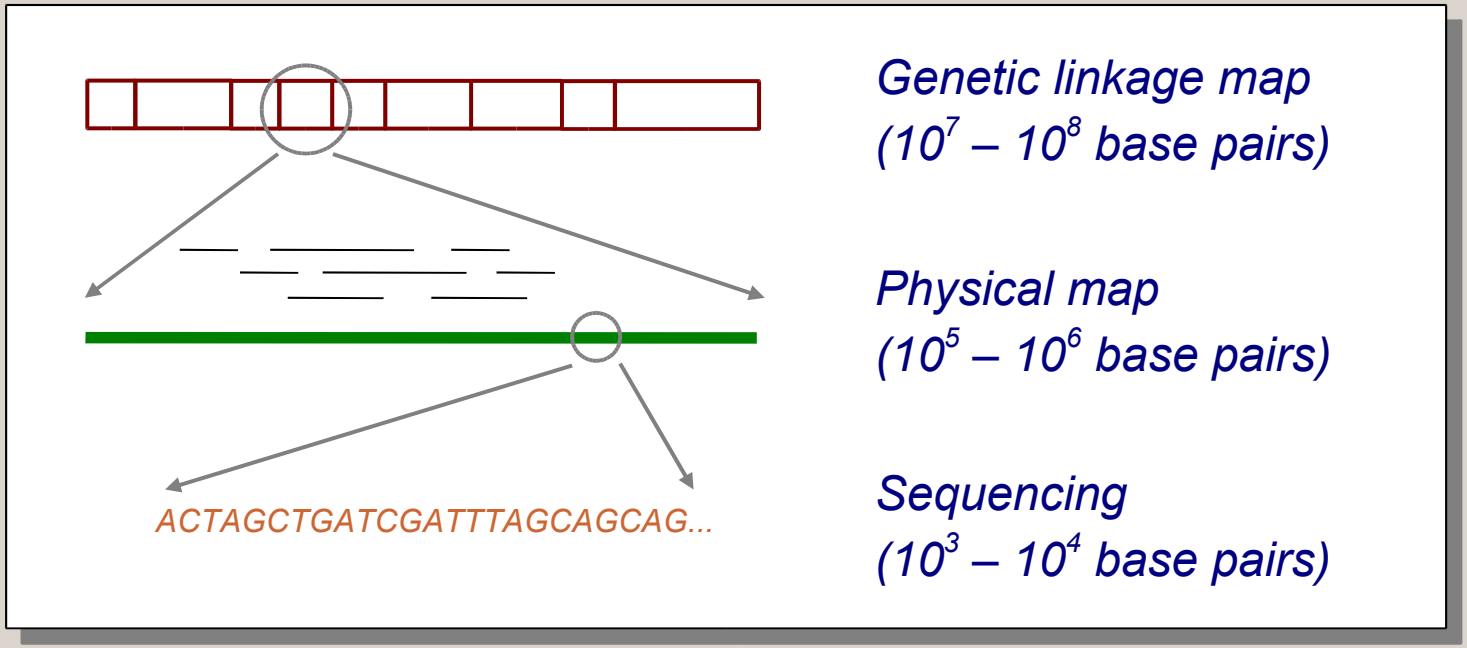
Original sequence: *ATCGTGTCGATAGCGCT*



Sequencing a Genome

Most genomes are enormous (e.g., 10^{10} base pairs in case of human). Current sequencing technology, on the other hand, only allows biologists to determine ~10^3 base pairs at a time.

This leads to some very interesting problems in bioinformatics ...

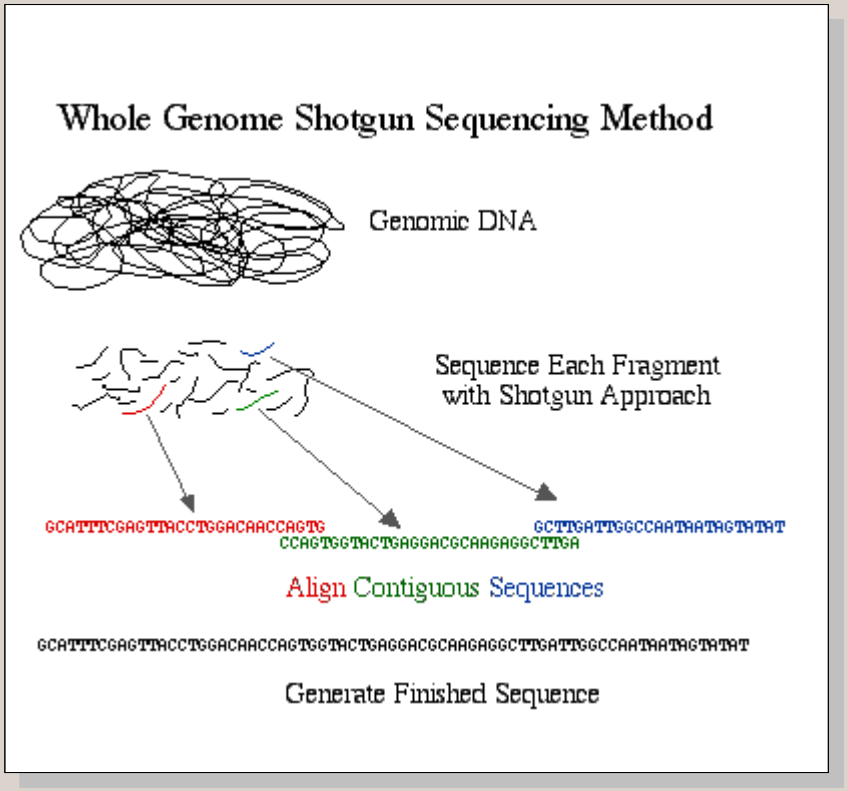


Sequencing a Genome

Genomes can also be determined using a technique known as *shotgun sequencing*.

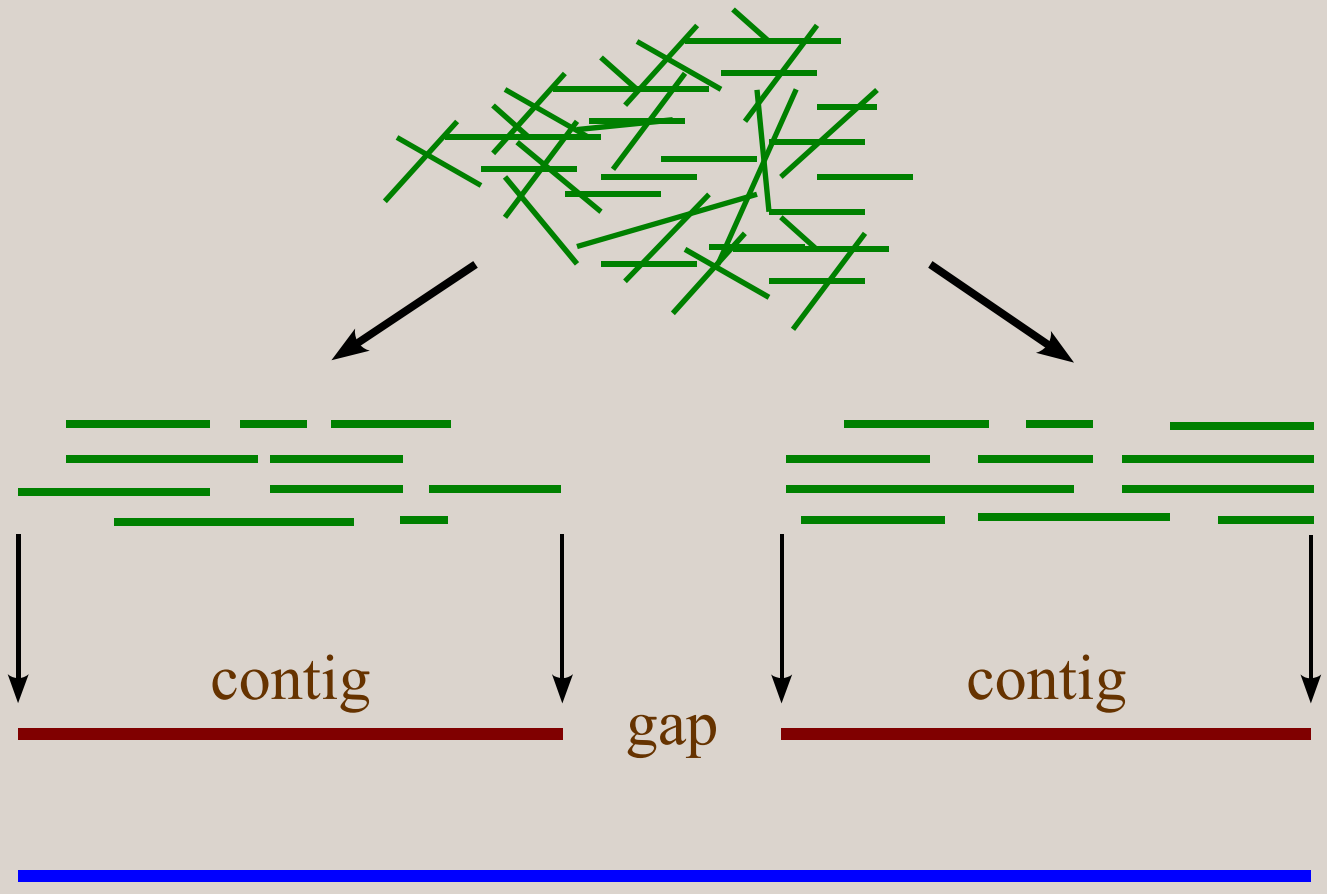
Computer scientists have played an important role in developing algorithms for assembling such data.

It's kind of like putting together a jigsaw puzzle with millions of pieces (a lot of which are "blue sky").



http://occawlonline.pearsoned.com/bookbind/pubbooks/bc_mcampbell_genomics_1/medialib/method/shotgun.html

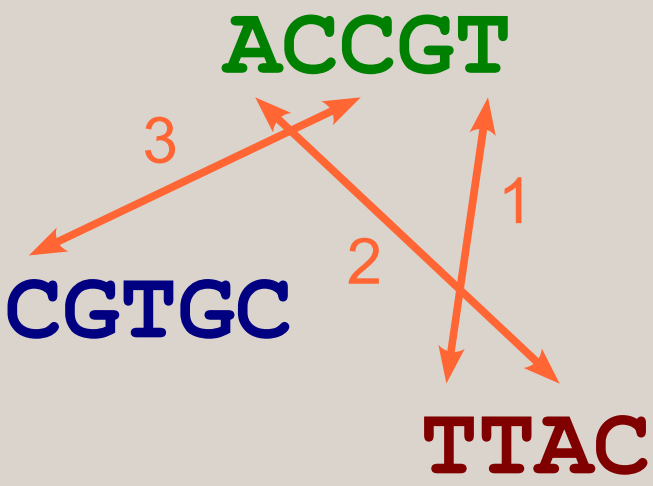
Sequence Assembly



Sequence Assembly

A simple model of DNA assembly is the *Shortest Supersequence Problem*: given a set of sequences, find the shortest sequence *S* such that each of original sequences appears as subsequence of *S*.

Look for overlap between *prefix* of one sequence and *suffix* of another:



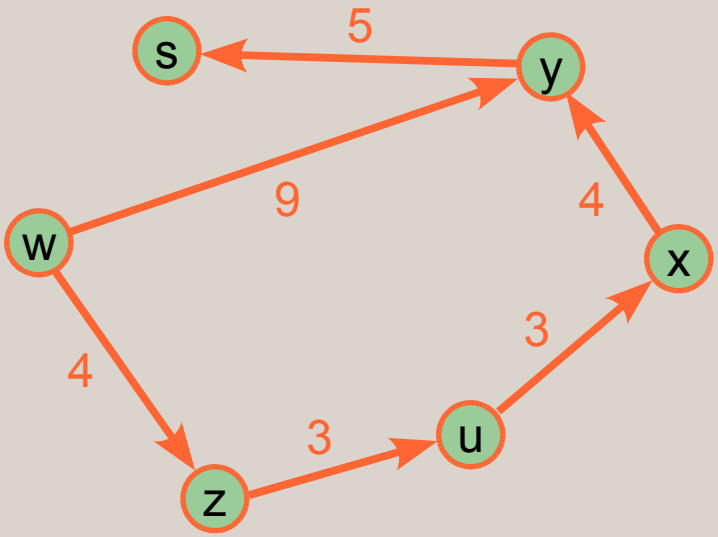

```
--ACCGT--
-----CGTGC
TTAC-----
----------
TTACCGTGC
```

Sequence Assembly

Sketch of algorithm:

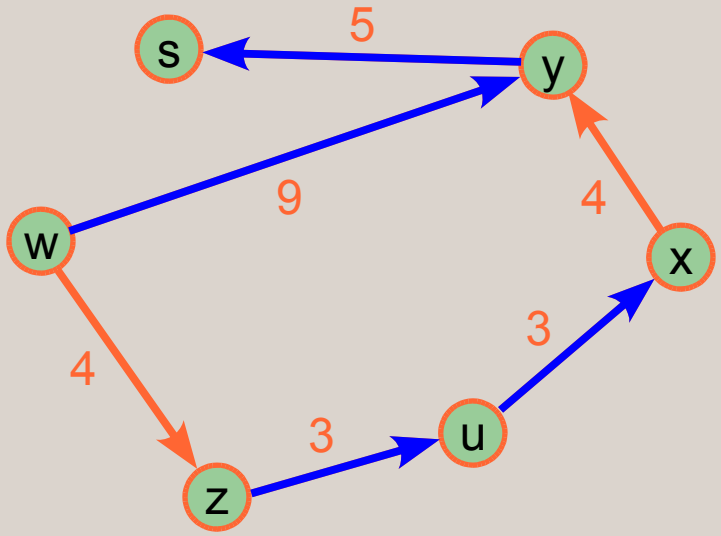
- Create an *overlap graph* in which every node represents a fragment and edges indicate overlap.
- Determine which overlaps will be used in the final assembly: find an *optimal spanning forest* in overlap graph.

```
W = AGTATTGGCAATC
Z = AATCGATG
U = ATGCAAACCT
X = CCTTTTGG
Y = TTGGCAATCA
S = AATCAGG
```



Sequence Assembly

- Look for paths of maximum weight: use greedy algorithm to select edge with highest weight at every step.
- Selected edge must connect nodes with in- and out-degrees <= 1.
- May end up with set of paths: each corresponds to a contig.



W→Y→S	AGTATTGGCAATC TTGGCAATCA AATCAGG AGTATTGGCAATCAGG
Z→U→X	AATCGATG ATGCAAACCT CCTTTTGG AATCGATGCAAACCTTTTGG

Sequence Comparison

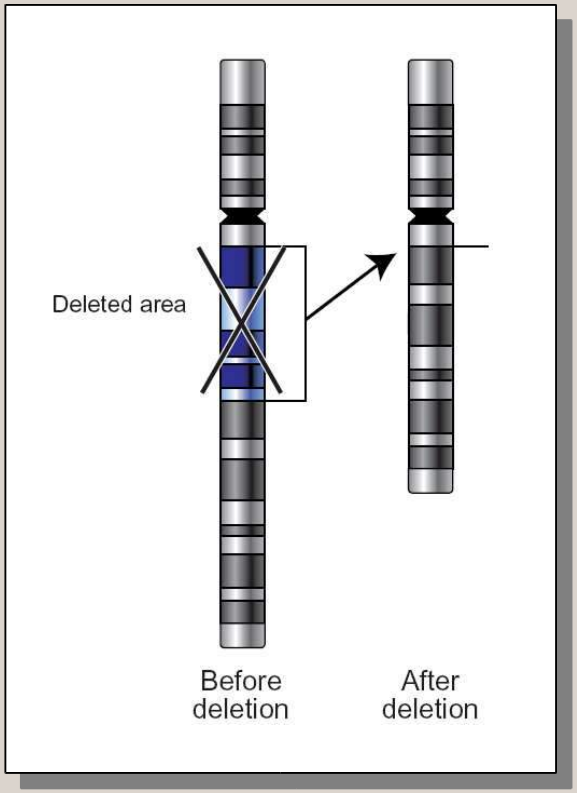
What's the problem? Google for biologists ...

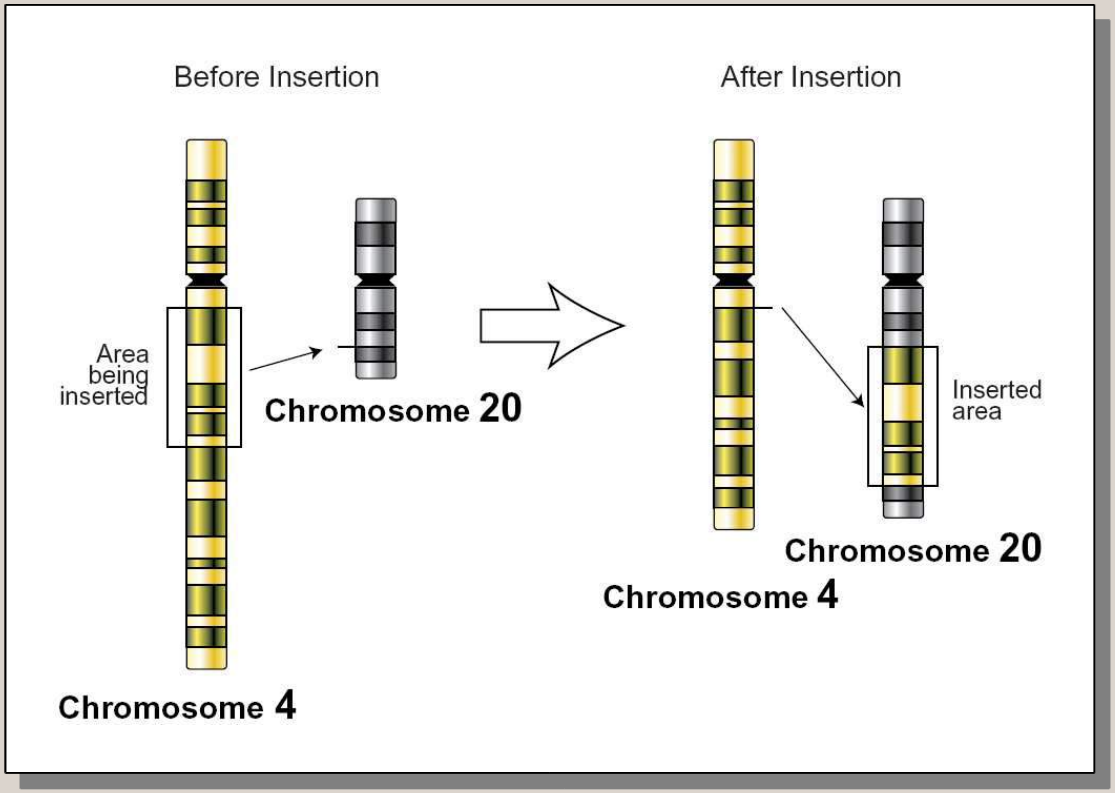
- Given new DNA or protein sequence, biologist will want to search databases of known sequences to look for anything similar.
- Sequence similarity can provide clues about function and evolutionary relationships.
- Databases such as GenBank are far too large to search manually. To search them efficiently, we need an algorithm.

Shouldn't expect exact matches (so it's not really like google):

- Genomes aren't static: mutations, insertions, deletions.
- Human (and machine) error in reading sequencing gels.

Genomes Aren't Static



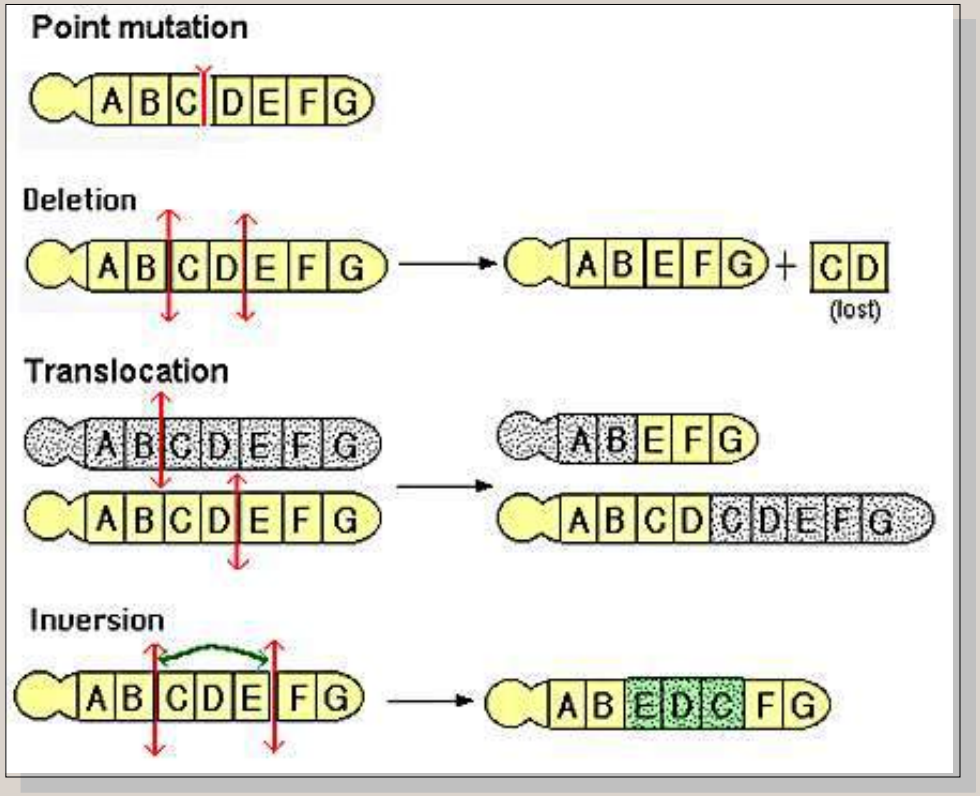


Sequence comparison must account for such effects.

http://www.accessexcellence.org/AB/GG/nhgri_PDFs/deletion.pdf http://www.accessexcellence.org/AB/GG/nhgri_PDFs/insertion.pdf

Genomes Aren't Static

Different kinds of mutations can arise during DNA replication:



http://www.accessexcellence.org/AB/GG/mutation.htm

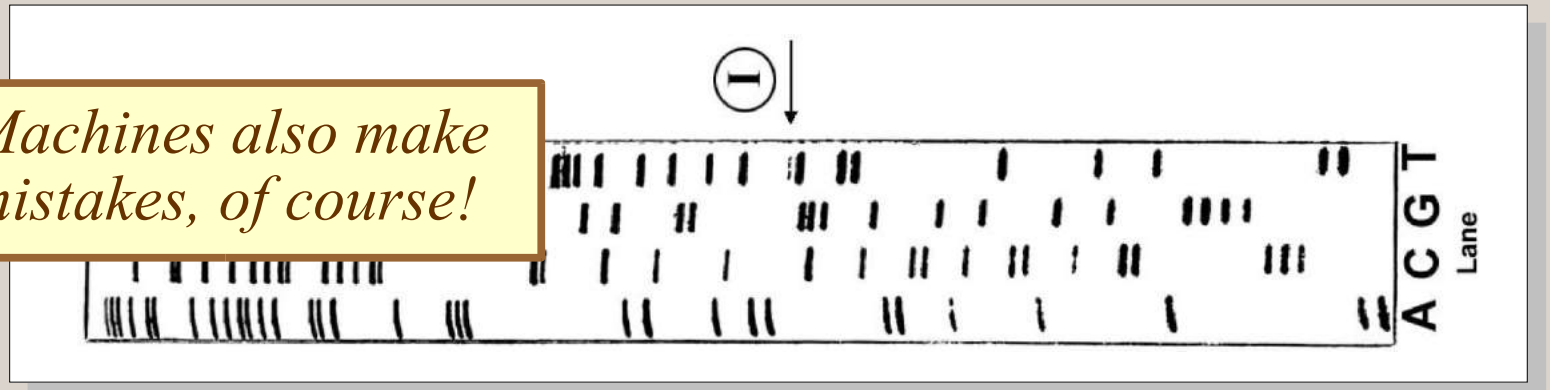
The Human Factor

In addition, errors can arise during the sequencing process:

> "...the error rate is generally less than 1% over the first 650 bases and then rises significantly over the remaining sequence."
>
> http://genome.med.harvard.edu/dnaseq.html

A hard-to-read gel (arrow marks location where bands of similar intensity appear in two different lanes):

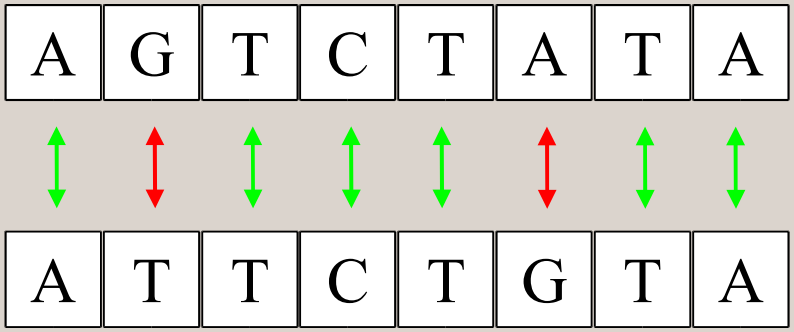
Machines also make mistakes, of course!



http://hshgp.genome.washington.edu/teacher_resources/99-studentDNASequencingModule.pdf

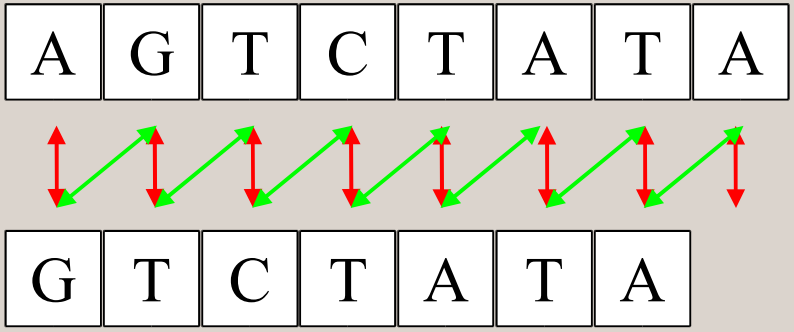
Sequence Comparison

Why not just line up sequences and count matches?



Difference = 2

Doesn't work well in case of deletions or insertions:



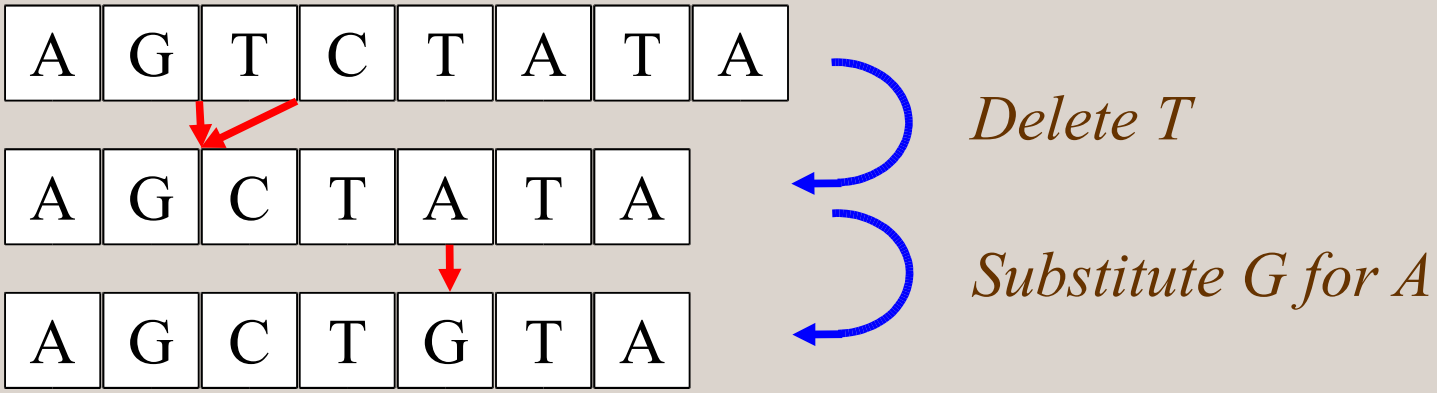
Difference = 8

One missing symbol at start of sequence leads to large difference!

Sequence Comparison

Instead, we'll use a technique known as *dynamic programming*.

- Model allows three basic operations: delete a single symbol, insert a single symbol, substitute one symbol for another.
- Goal: given two sequences, find the shortest series of operations needed to transform one into the other.

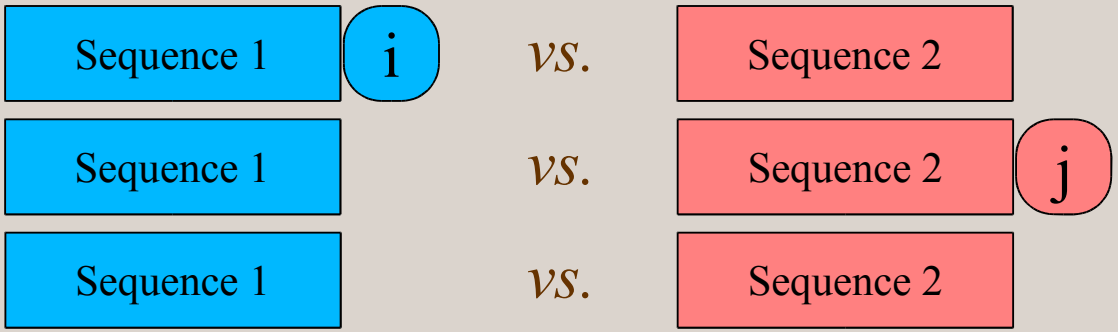


Sequence Comparison

How can we determine optimal series of operations?

- Approach is to build up longer solutions from previously computed shorter solutions.
- Say we want to compute solution at index *i* in first sequence and index *j* in second sequence:

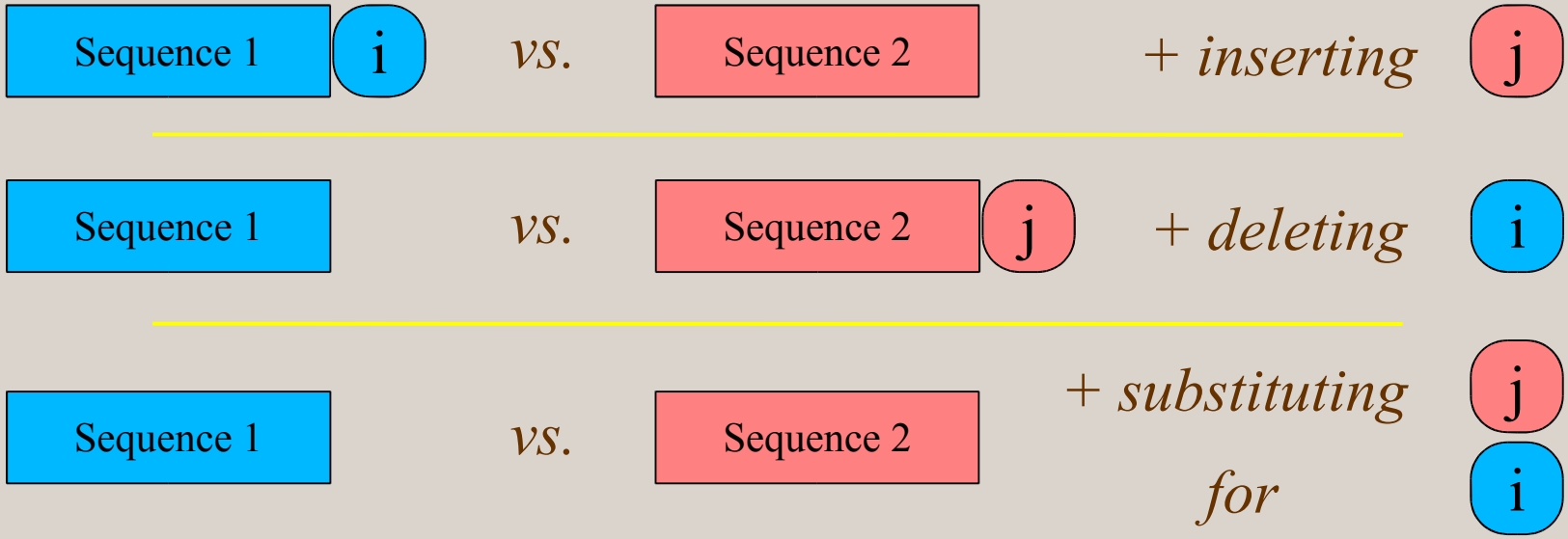
Assume that we already know the best way to compare:



Sequence Comparison

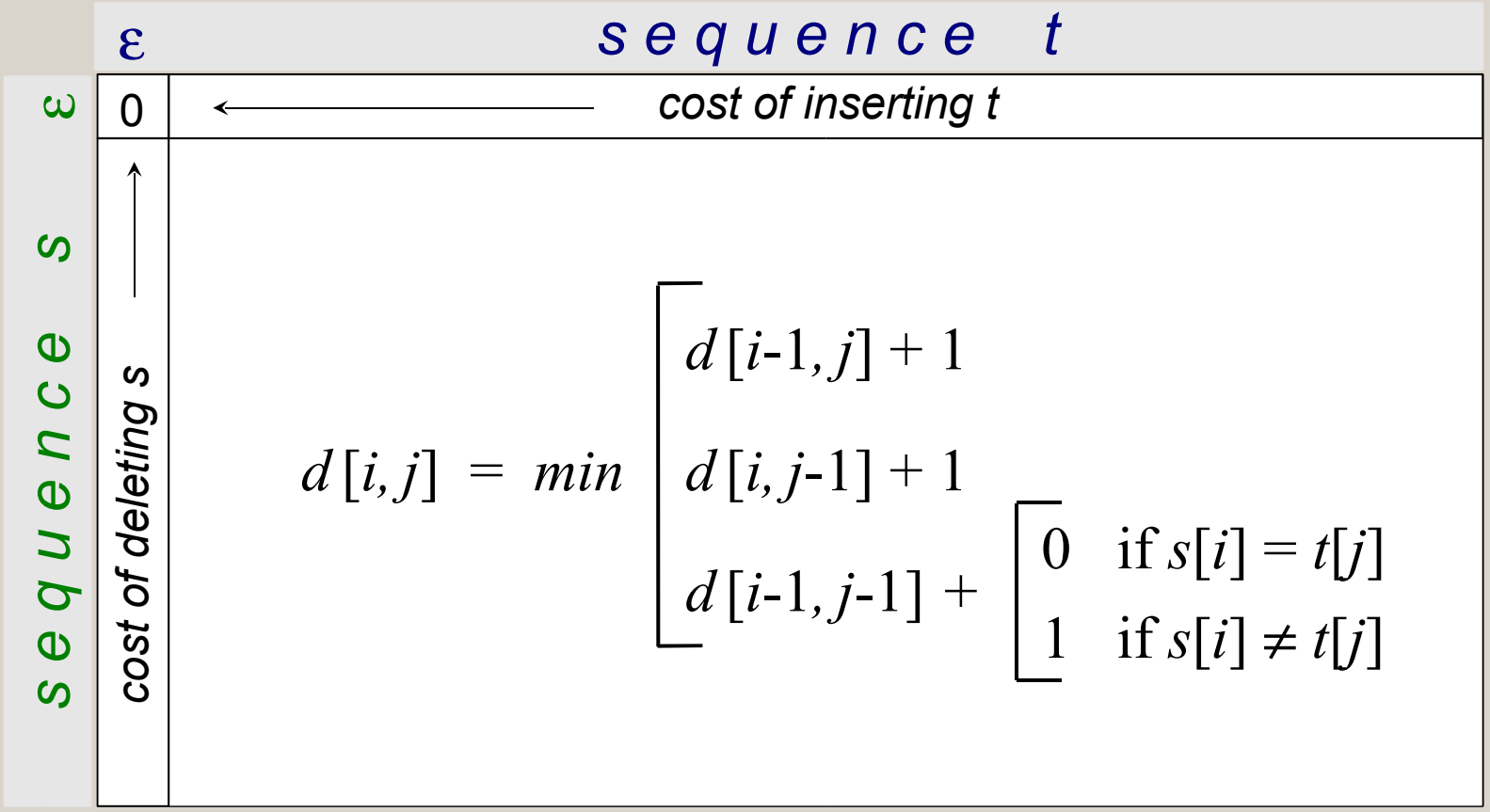
So, best way to do this comparison:

Is best choice from following three cases:



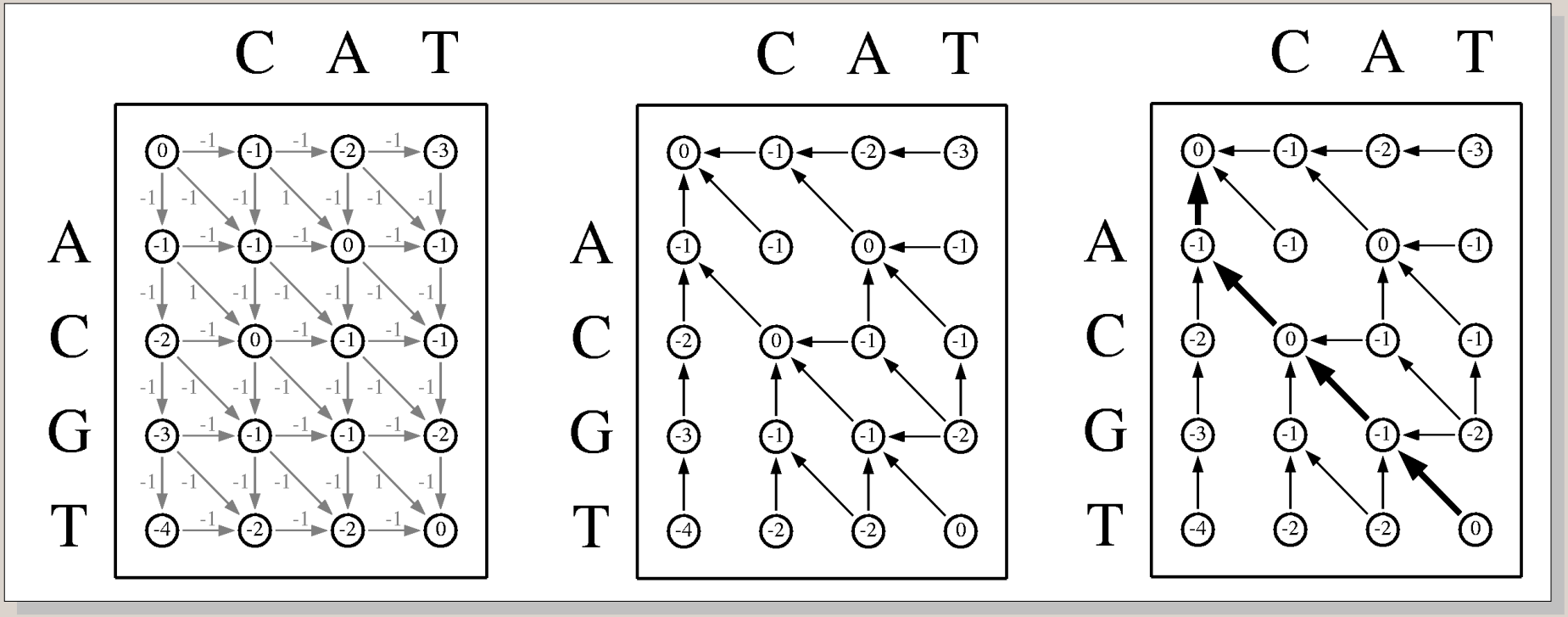
Sequence Comparison

Normally, this computation builds a table of distance values:



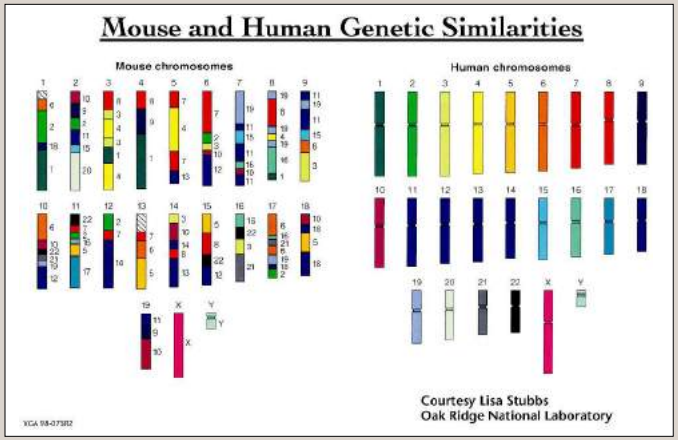
Sequence Comparison

By keeping track of optimal decision, we can determine operations:



Genome Rearrangements

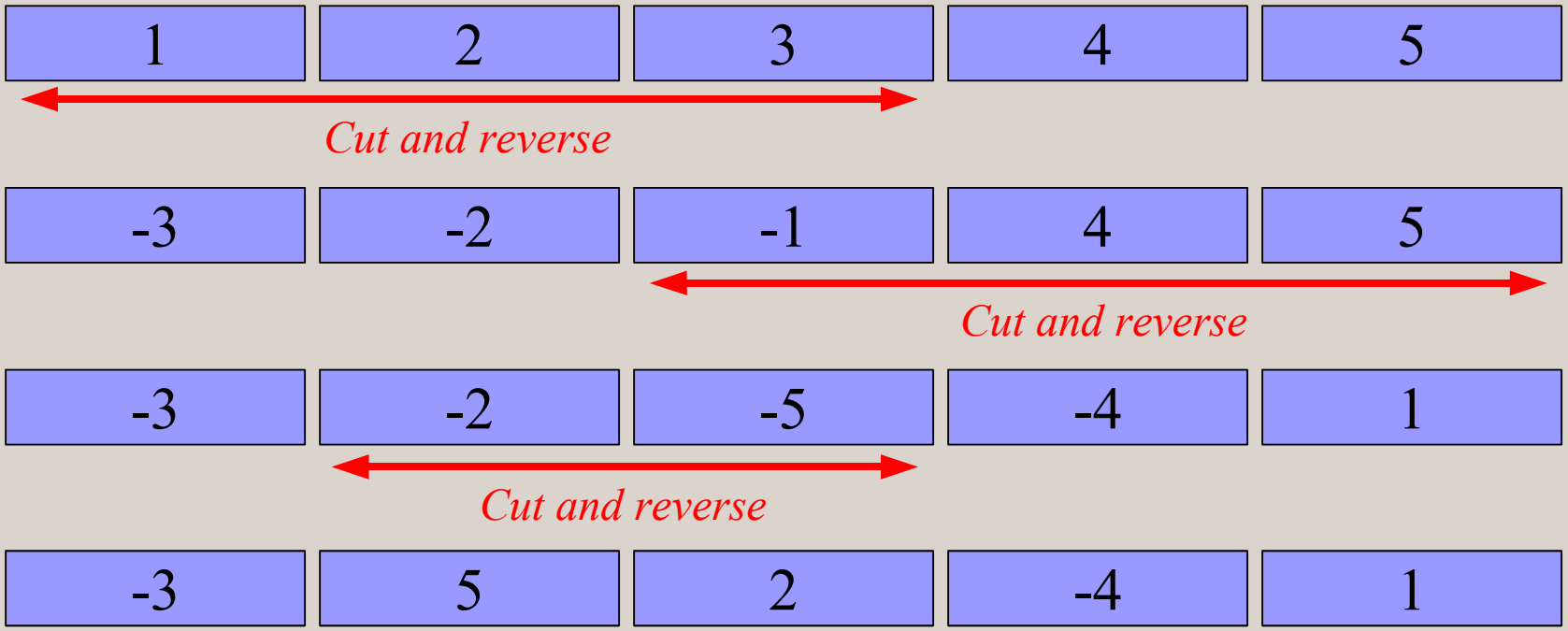
Recall what we saw earlier:



- 99% of mouse genes have homologues in human genome.
- 96% of mouse genes are in same relative location to one another.
- Mouse genome can be broken up into 300 *synteny blocks* which, when rearranged, yield human genome.
- Provides a way to think about evolutionary relationships.

Reversal Distance

Human Chromosome X

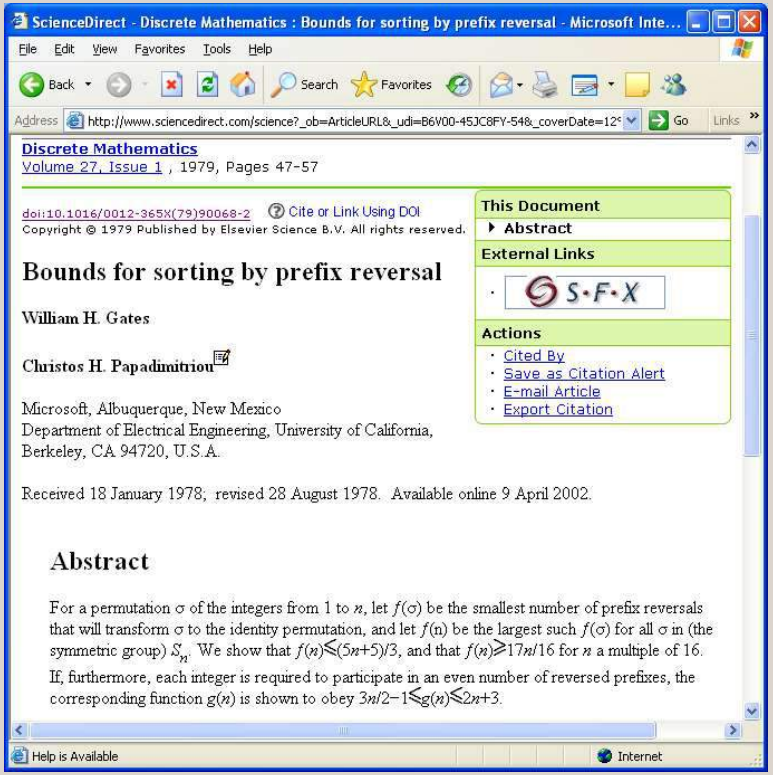


Mouse Chromosome X

Reversal distance is the minimum number of such steps needed.

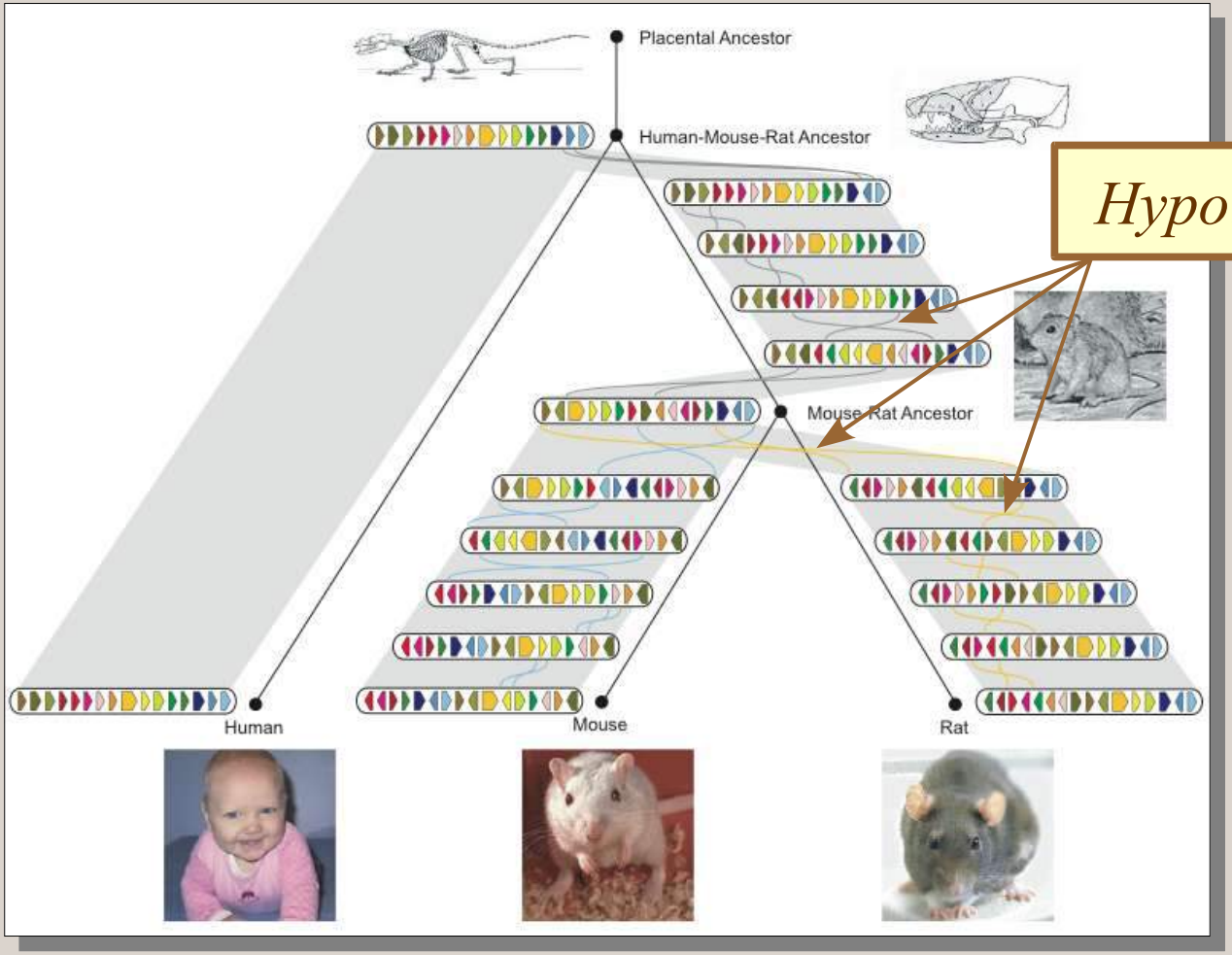
Interesting Sidenote

Early work on a related problem, sorting by prefix reversals, was performed in 1970's by Christos Papadimitriou, a famous computer scientist now at UC Berkeley, and one "William H. Gates" ...



Yes, that Bill Gates ...

History of Chromosome X



Rat Consortium, Nature, 2004

Waardenburg's Syndrome

Mouse provides insight into human genetic disorder:

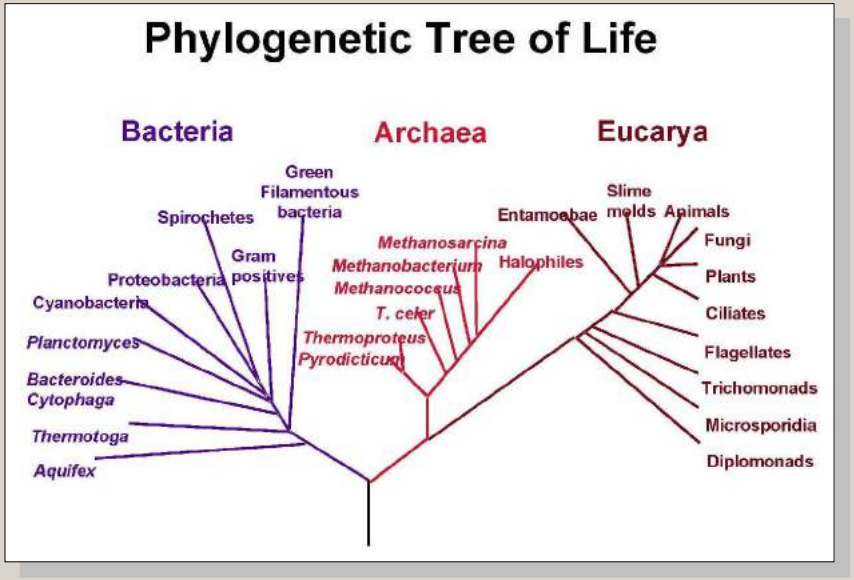
- Waardenburg's syndrome is characterized by pigmentary dysphasia.
- Disease gene linked to Chromosome 2, but not clear where it was located.

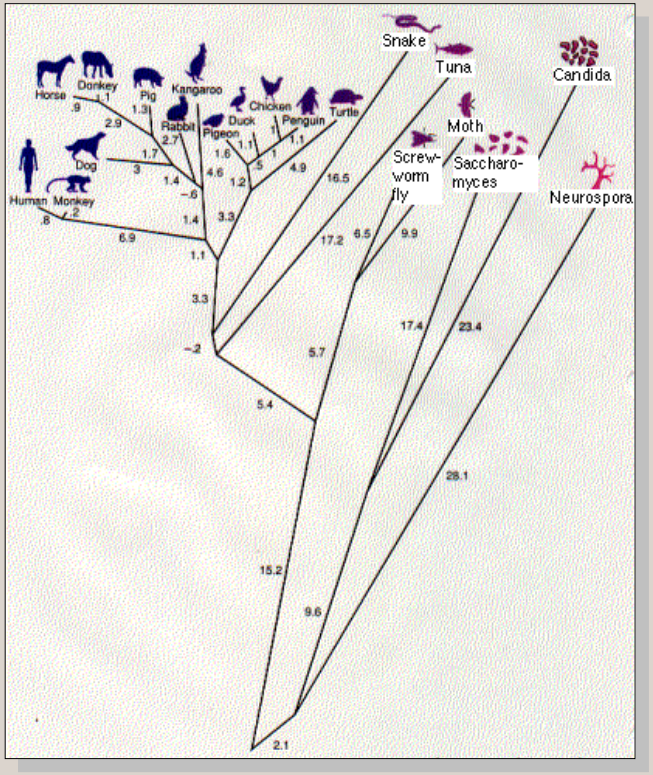
"Splotch" mice:

- A breed of mice (with splotch gene) had similar symptoms.
- Scientists succeeded in identifying location of gene in mice.
- This gave clues as to where same gene is located in humans.

Building the "Tree of Life"

Scientists build phylogenetic trees in an attempt to understand evolutionary relationships. Reversal distance is often used here.



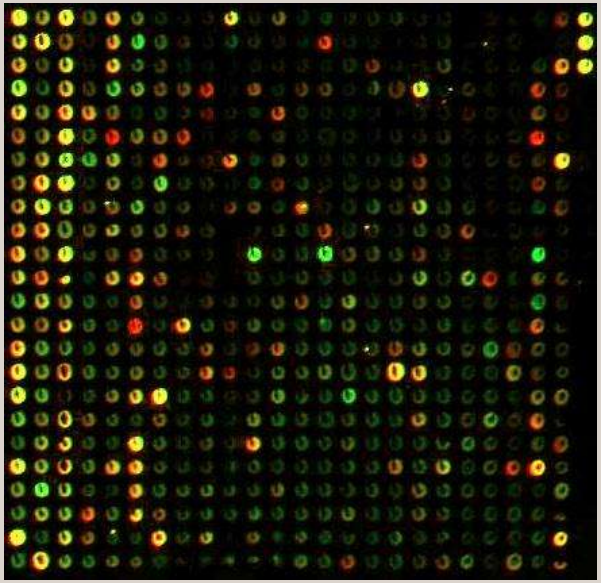


http://en.wikipedia.org/wiki/Phylogenetic_tree
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/Taxonomy.html

Note: these trees are "best guesses" and certainly contain some errors!

DNA Microarrays

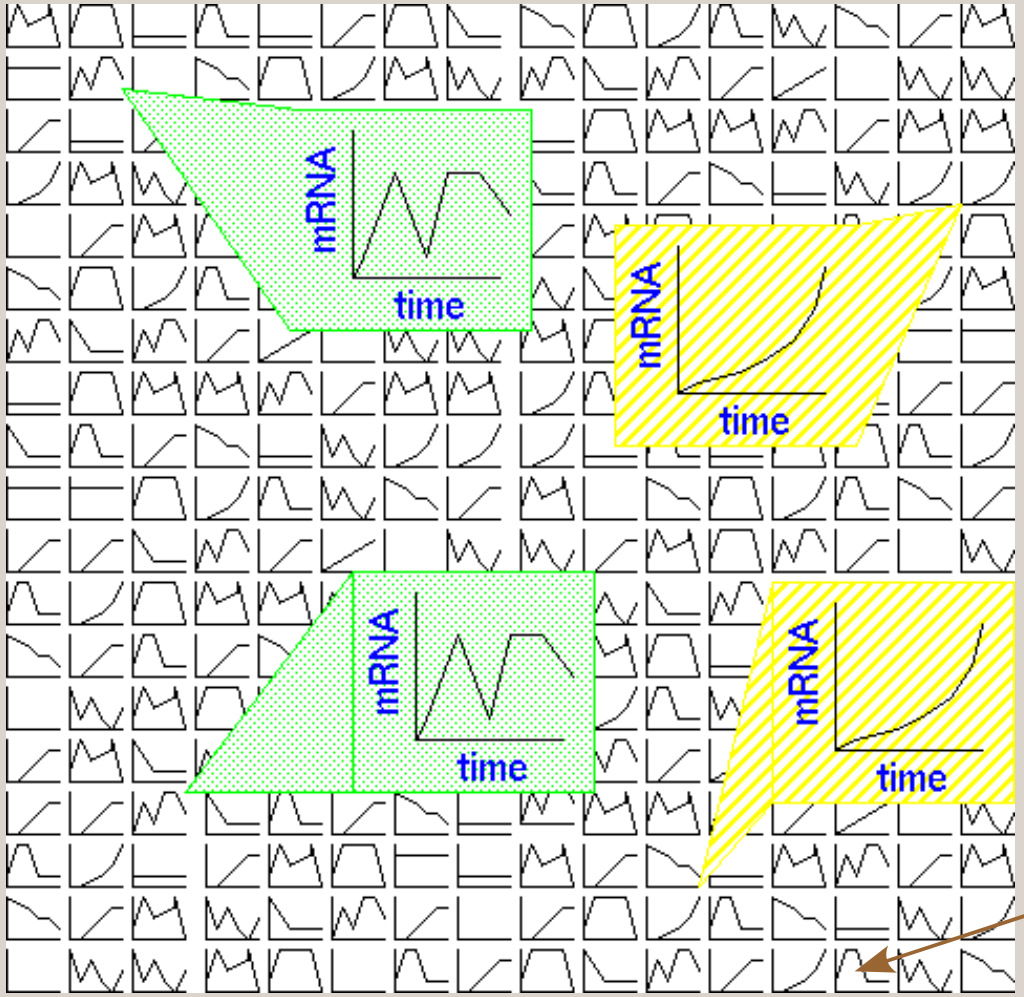
- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression).
- Differential expression, changes over time.
- Single microarray can test ~10k genes.
- Data obtained faster than can be processed.
- Want to find genes that behave similarly.
- A pattern discovery problem.



green = repressed

red = induced

Using DNA Microarrays



- Track sample over a period of time to see gene expression over time.
- Track two different samples under same conditions to see difference in gene expressions.

Each box represents one gene's expression over time

http://www.bioalgorithms.info/presentations/Ch10_Clustering.ppt

DNA Microarrays

K-means clustering is one way to organize this data:

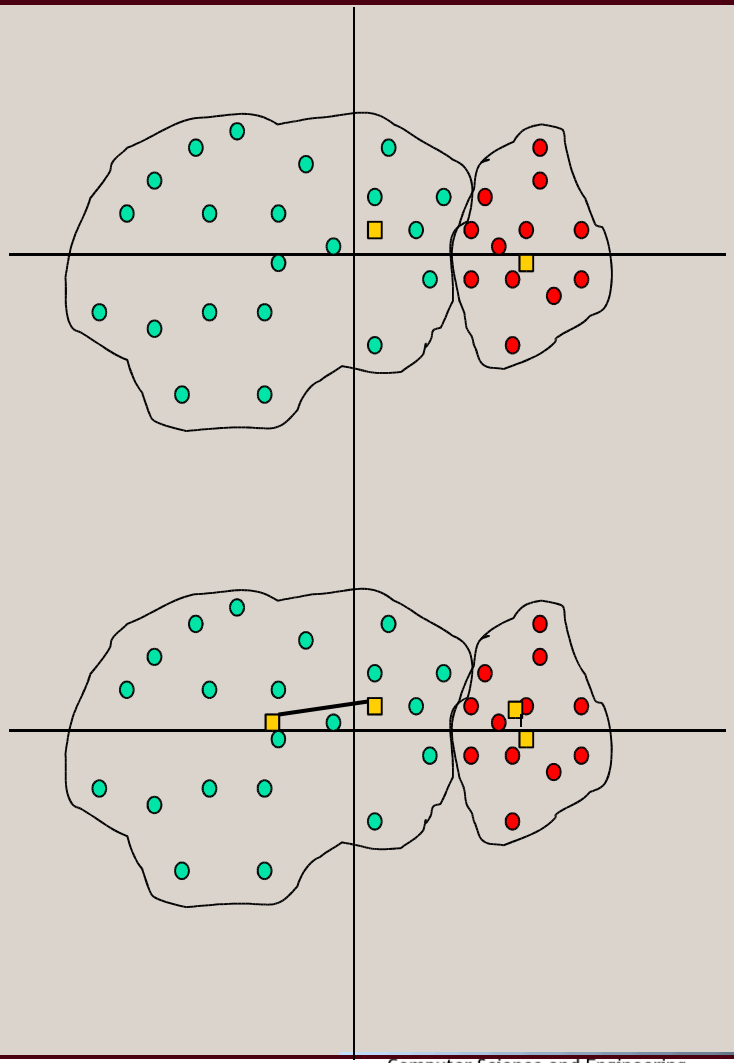
- Given set of *n* data points and an integer *k*.
- We want to find set of *k* points that minimizes the mean-squared distance from each data point to its nearest cluster center.

Sketch of algorithm:

- Choose *k* initial center points randomly and cluster data.
- Calculate new centers for each cluster using points in cluster.
- Re-cluster all data using new center points.
- Repeat second two steps until no data points are moved from one cluster to another or some other convergence criterion is met.

Clustering Microarray Data

- Pick $k = 2$ centers at random.
- Cluster data around these center points.

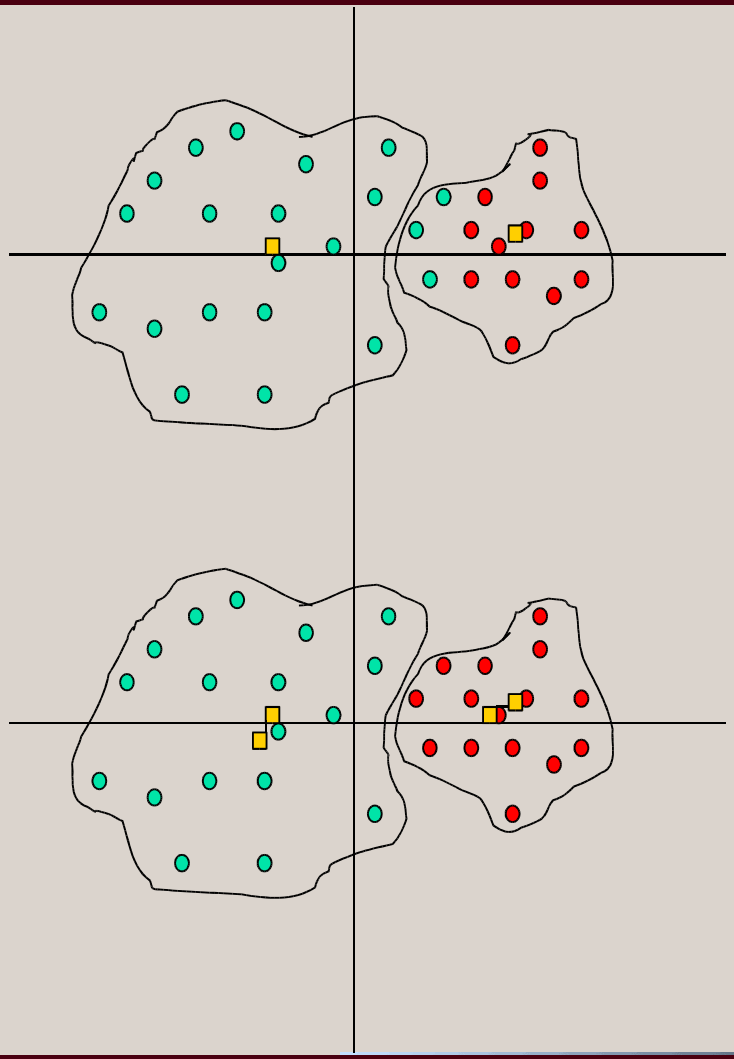


- Re-calculate centers based on current clusters.

From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.

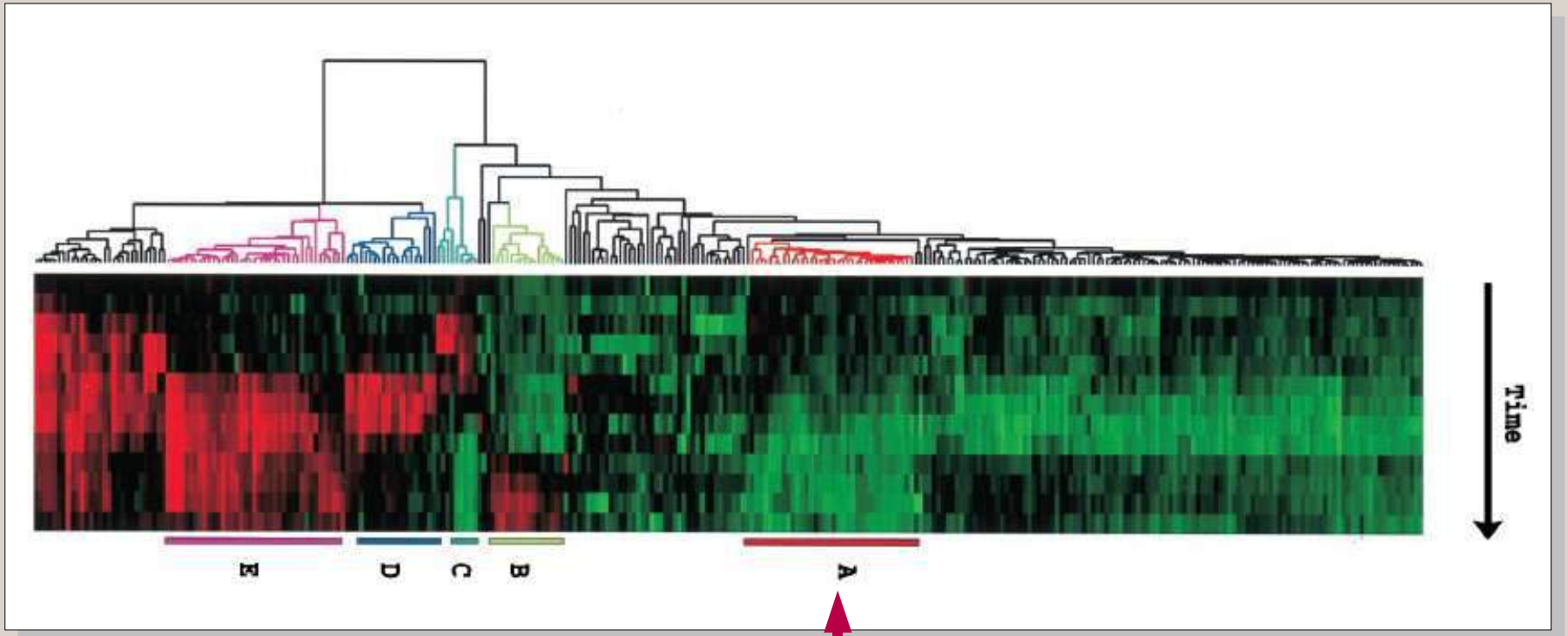
Clustering Microarray Data

- Re-cluster data around new center points.
- Repeat last two steps until no more data points are moved into a different cluster.



From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.

Example of Hierarchical Clustering



Different genes that express similarly

From "Cluster analysis and display of genome-wide expression patterns" by Eisen, Spellman, Brown, and Botstein, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 14863–14868, December 1998

Why Study Bioinformatics?

- Still many urgent open problems $\Rightarrow$ lots of opportunities to make fundamental contributions (and become rich and famous).
- Stretch your creativity and problem-solving skills to the limit.
- Join a cross-disciplinary team – work with interesting people.
- Participate in unlocking the mysteries of life itself.
- Make the world a better place.

CSE Course in Bioinformatics

In CSE 308, we cover:

- Intro to molecular biology & algorithms,
- Basic programming for bioinformatics,
- Genetic sequence comparison & alignment,
- Physical mapping, sequencing, and assembly of DNA,
- Standard formats and sources for genomic data,
- Advanced topics: DNA microarrays, genome rearrangements, RNA and protein structure prediction, etc.

CSE 308 is <u>not</u> a programming course!

Materials @ http://www.cse.lehigh.edu/~lopresti/courses.html
Questions: dal9@lehigh.edu

Thank you!