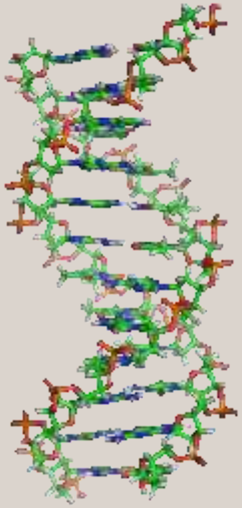
Introduction to Bioinformatics

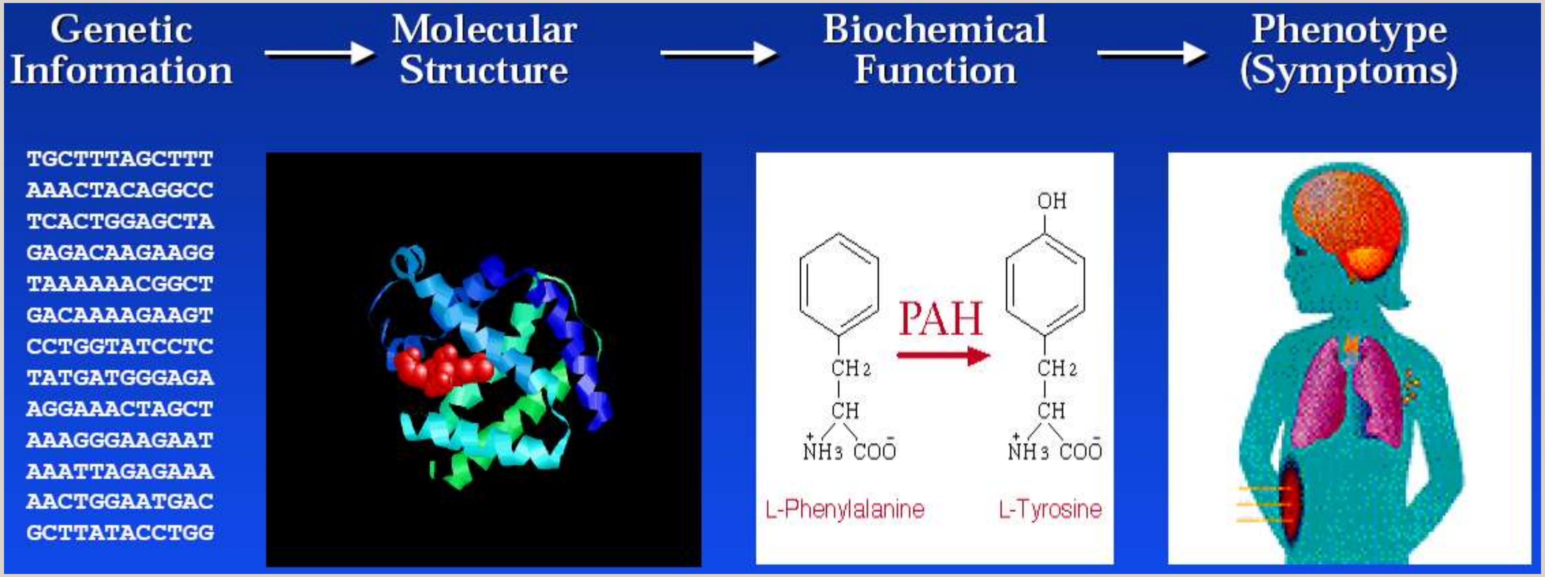
Dan Lopresti
Computer Science and Engineering
Office Building C 337
dal9@lehigh.edu



In 2017 when I gave this talk …

Motivation

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



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

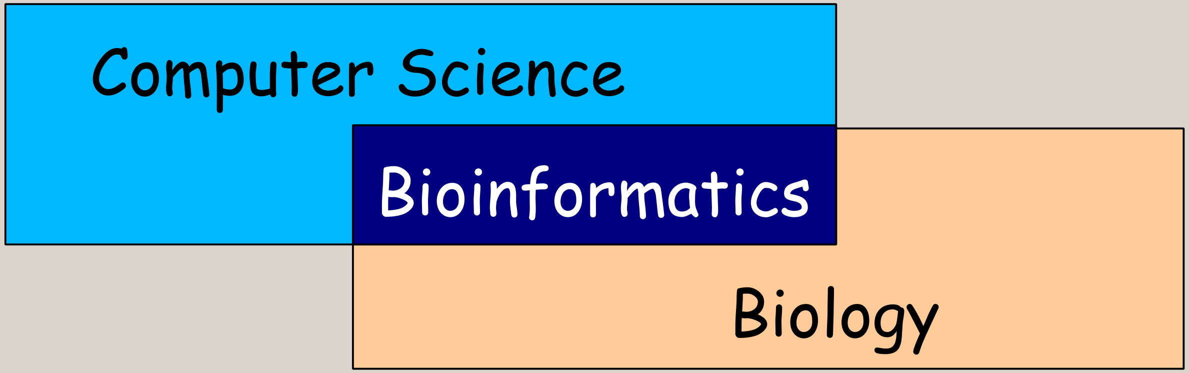
TGCTTTAGCTTT
AAACTACAGGCC
TCACTGGAGCTA
GAGACAAGAAGG
TAAAAAACGGCT
GACAAAAGAAGT
CCTGGTATCCTC
TATGATGGGAGA
AGGAAACTAGCT
AAAGGGAAGAAT
AAATTAGAGAAA
AACTGGAATGAC
GCTTATACCTGG

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

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

Bioinformatics

What is bioinformatics? Application of methods from computer science to biology.



Computer Science

Bioinformatics

Biology

Why is it interesting?

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

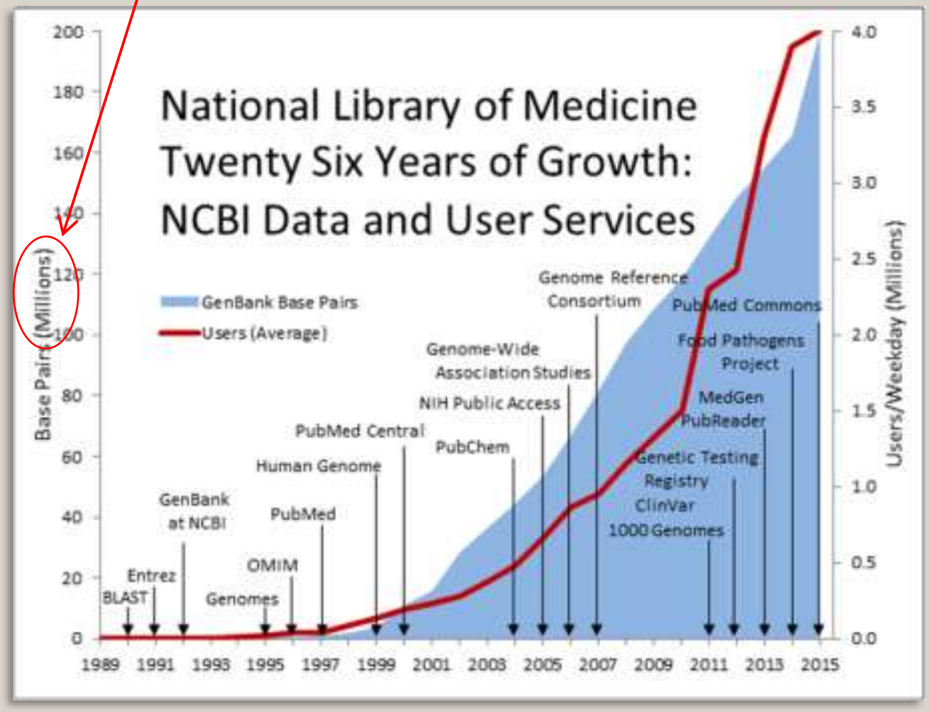
Data Explosion

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* }.

NLM / NIH seems to have made a mistake: this should be billions, not millions!

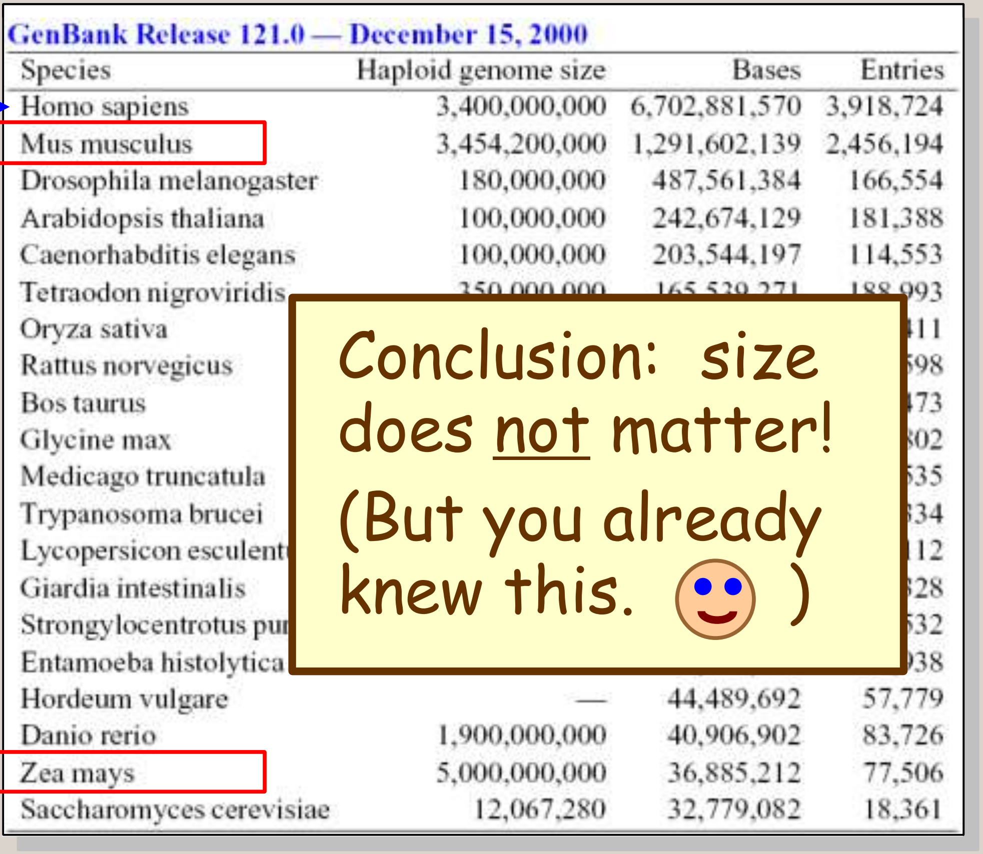


https://www.nlm.nih.gov/about/2017CJ.html

Genomes

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

Us



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	[illegible]	[illegible]	[illegible]
Rattus norvegicus	[illegible]	[illegible]	[illegible]
Bos taurus	[illegible]	[illegible]	[illegible]
Glycine max	[illegible]	[illegible]	[illegible]
Medicago truncatula	[illegible]	[illegible]	[illegible]
Trypanosoma brucei	[illegible]	[illegible]	[illegible]
Lycopersicon esculent	[illegible]	[illegible]	[illegible]
Giardia intestinalis	[illegible]	[illegible]	[illegible]
Strongylocentrotus pur	[illegible]	[illegible]	[illegible]
Entamoeba histolytica	[illegible]	[illegible]	[illegible]
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

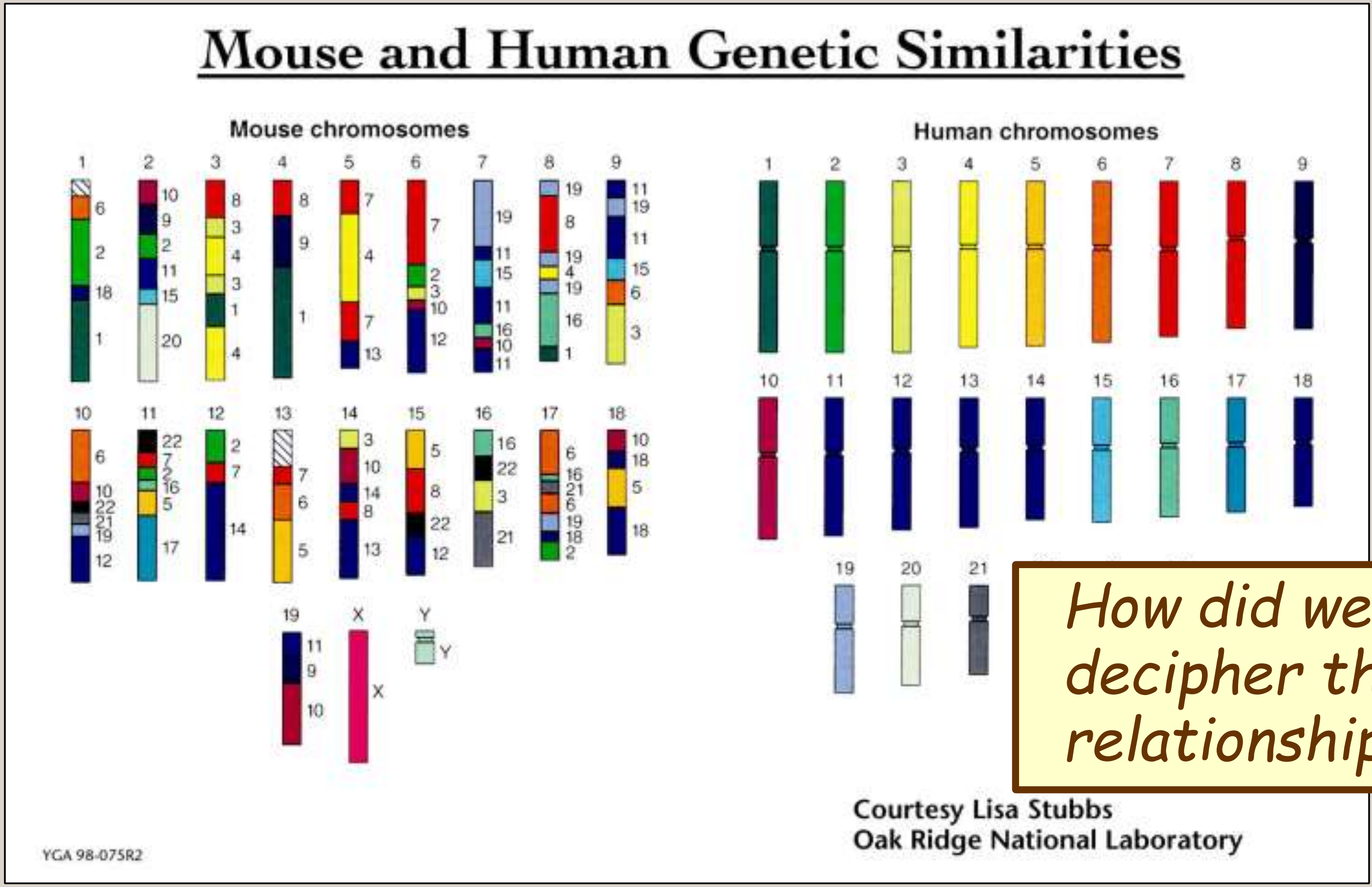
Conclusion: size does <u>not</u> matter!

(But you already knew this. ☺)

Mus musculus

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



Mouse and Human Genetic Similarities

Courtesy Lisa Stubbs
Oak Ridge National Laboratory

How did we decipher these relationships?

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

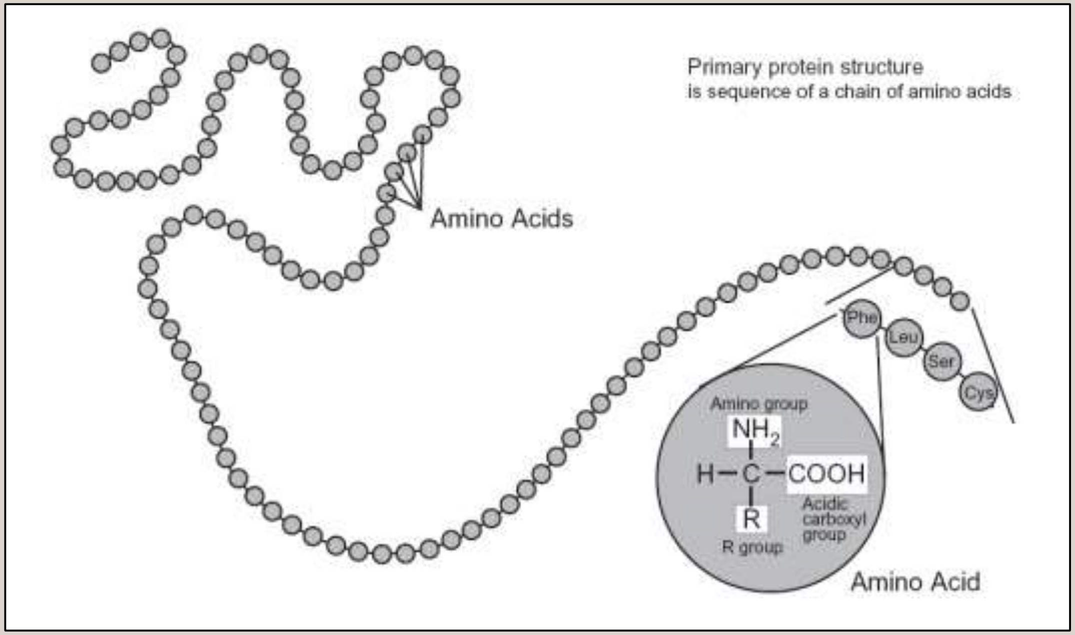
Algorithms are Central

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?
  - Can it be solved efficiently? Speed, memory, etc.
  - If we can't find good algorithm, can we prove task hard?
  - If known to be hard, is there approximation algorithm (works some of the time or comes close to optimal)?
- Conduct experimental evaluations (iterate above steps).

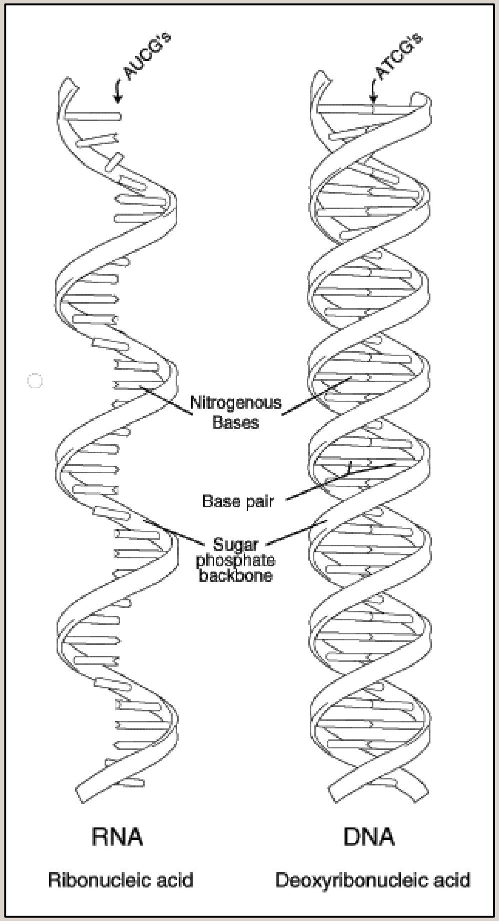
Sequence Nature of Biology

Macromolecules are chains of simpler molecules.



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

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

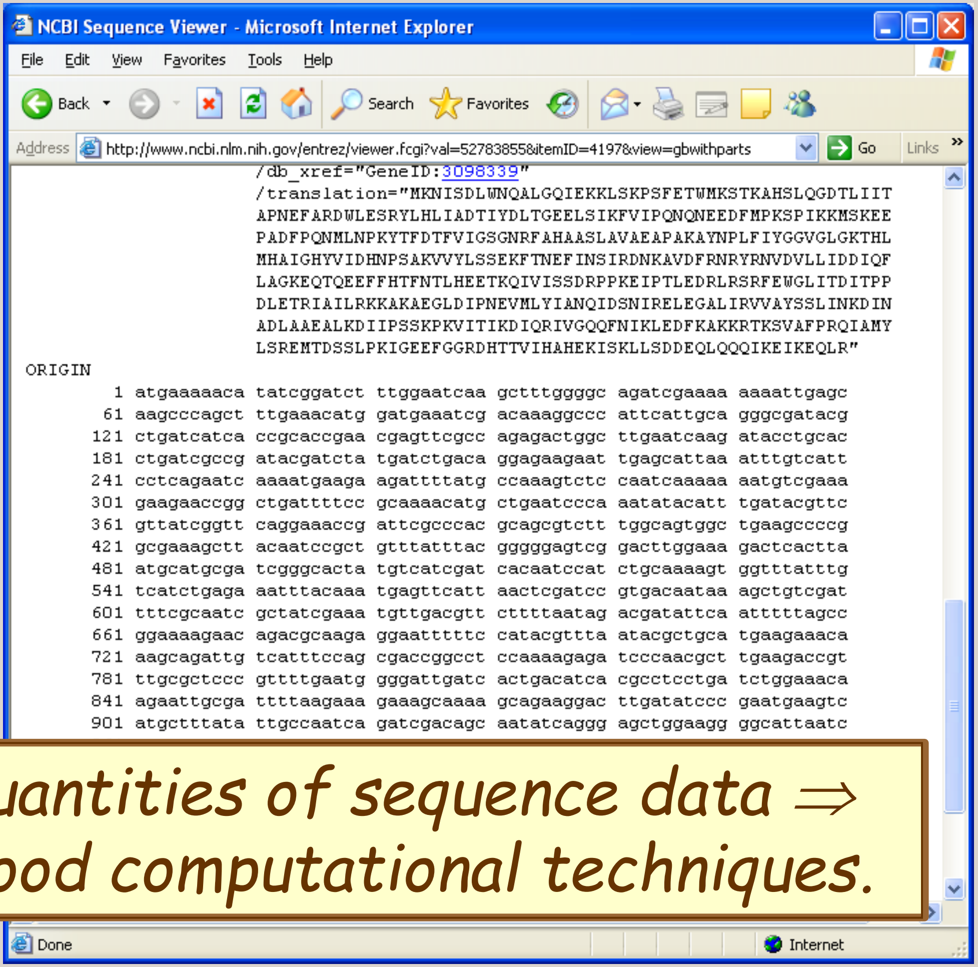


In case of proteins, these basic building blocks are *amino acids*.

In DNA and RNA, they are *nucleotides*.

NCBI GenBank

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


NCBI Sequence Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
Done
Internet

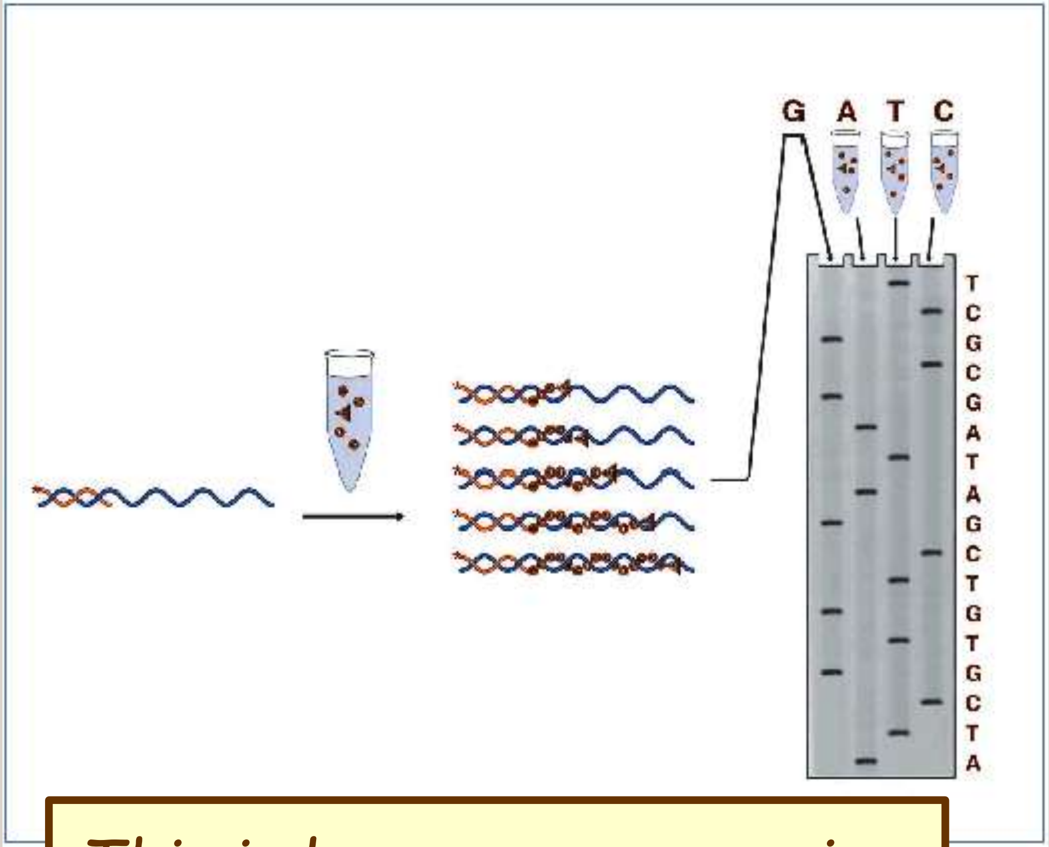
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=52783855&itemID=4197&view=gbwithparts

```
                     /db_xref="GeneID:3098339"
                     /translation="MKNISDLWNQALGQIEKKLSKPSFETWMKSTKAHSLQGDTLIIT
                     APNEFARDWLESRYLHLIADTIYDLTGEELSIKFVIPQNQNEEDFMPKSPIKKMSKEE
                     PADFPQNMLNPKYTFDTFVIGSGNRFAHAASLAVAEAPAKAYNPLFIYGGVGLGKTHL
                     MHAIGHYVIDHNPSAKVVYLSSEKFTNEFINSIRDNKAVDFRNRYRNVDVLLIDDIQF
                     LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPPKEIPTLEDRLRSRFEWGLITDITPP
                     DLETRIAILRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKDIN
                     ADLAAEALKDIIPSSKPKVITIKDIQRIVGQQFNIKLEDFKAKKRTKSVAFPRQIAMY
                     LSREMTDSSLPKIGEEFGGRDHTTVIHAHEKISKLLSDDEQLQQQIKEIKEQLR"
ORIGIN
        1 atgaaaaaca tatcggatct ttggaatcaa gctttggggc agatcgaaaa aaaattgagc
       61 aagcccagct ttgaaacatg gatgaaatcg acaaaggccc attcattgca gggcgatacg
      121 ctgatcatca ccgcaccgaa cgagttcgcc agagactggc ttgaatcaag atacctgcac
      181 ctgatcgccg atacgatcta tgatctgaca ggagaagaat tgagcattaa atttgtcatt
      241 cctcagaatc aaaatgaaga agattttatg ccaaagtctc caatcaaaaa aatgtcgaaa
      301 gaagaaccgg ctgattttcc gcaaaacatg ctgaatccca aatatacatt tgatacgttc
      361 gttatcggtt caggaaaccg attcgcccac gcagcgtctt tggcagtggc tgaagccccg
      421 gcgaaagctt acaatccgct gtttatttac gggggagtcg gacttggaaa gactcactta
      481 atgcatgcga tcgggcacta tgtcatcgat cacaatccat ctgcaaaagt ggtttatttg
      541 tcatctgaga aatttacaaa tgagttcatt aactcgatcc gtgacaataa agctgtcgat
      601 tttcgcaatc gctatcgaaa tgttgacgtt cttttaatag acgatattca atttttagcc
      661 ggaaaagaac agacgcaaga ggaatttttc catacgttta atacgctgca tgaagaaaca
      721 aagcagattg tcatttccag cgaccggcct ccaaaagaga tcccaacgct tgaagaccgt
      781 ttgcgctccc gttttgaatg gggattgatc actgacatca cgcctcctga tctggaaaca
      841 agaattgcga ttttaagaaa gaaagcaaaa gcagaaggac ttgatatccc gaatgaagtc
      901 atgctttata ttgccaatca gatcgacagc aatatcaggg agctggaagg ggcattaatc
```

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

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

Reading DNA



This is known as sequencing.

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

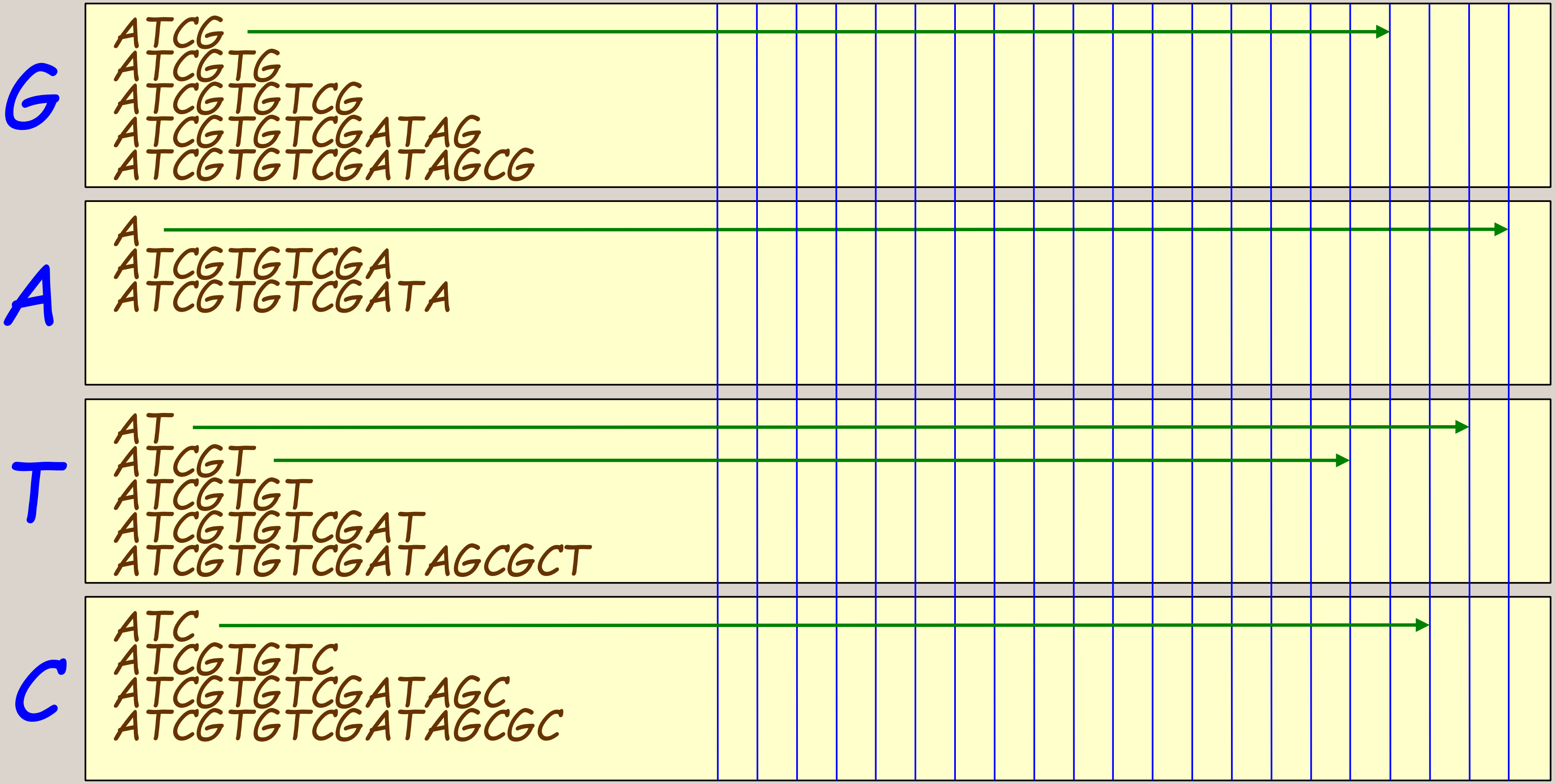
Gel electrophoresis separates mixture of molecules in a gel media by application of an electric field.

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

Make DNA fragments that end at each base. Then run gel and read off sequence: ATCGTG ...

Reading DNA

Original sequence: *ATCGTGTCGATAGCGCT*



G

ATCG
ATCGTG
ATCGTGTCG
ATCGTGTCGATAG
ATCGTGTCGATAGCG

A

A
ATCGTGTCGA
ATCGTGTCGATA

T

AT
ATCGT
ATCGTGT
ATCGTGTCGAT
ATCGTGTCGATAGCGCT

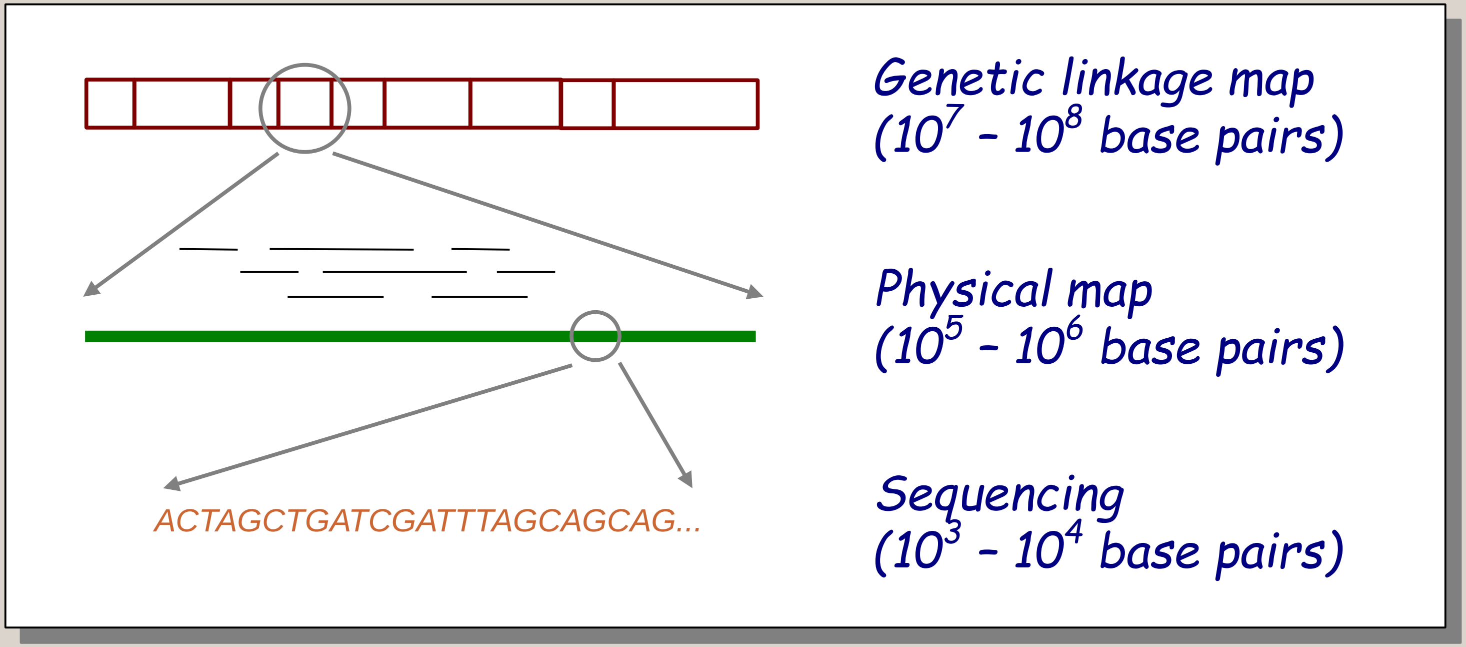
C

ATC
ATCGTGTC
ATCGTGTCGATAGC
ATCGTGTCGATAGCGC

Sequencing a Genome

Most genomes are enormous (e.g., 10^{10} base pairs for human). But current sequencing technology only allows biologists to determine ~10^{3} base pairs at a time.

Leads to some very interesting problems in bioinformatics!



Genetic linkage map
(10^{7} - 10^{8} base pairs)

Physical map
(10^{5} - 10^{6} base pairs)

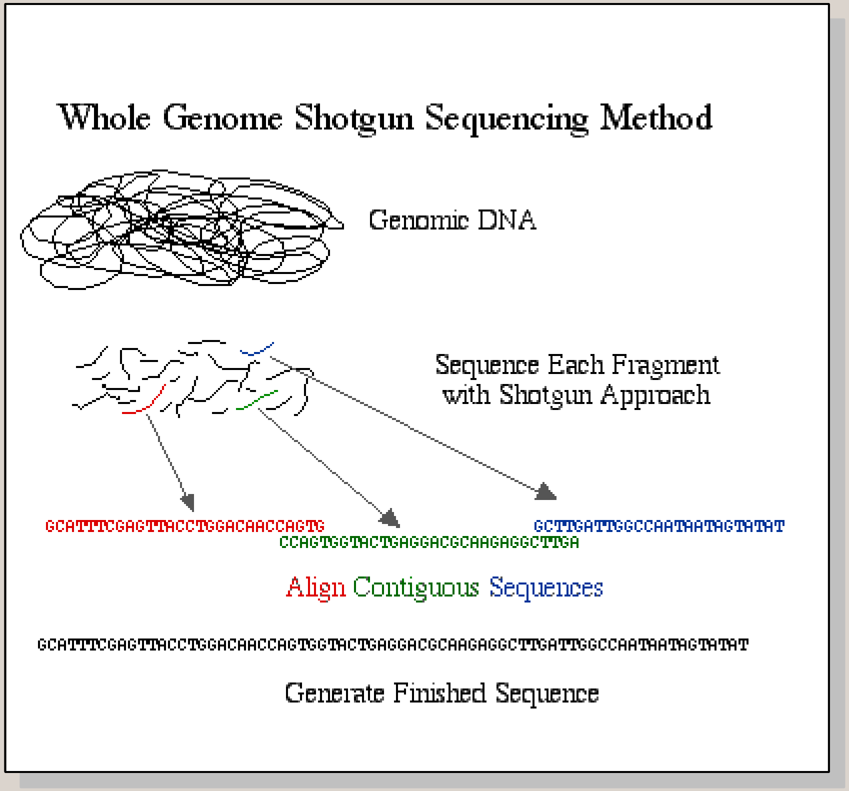
Sequencing
(10^{3} - 10^{4} base pairs)

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



Whole Genome Shotgun Sequencing Method

Genomic DNA

Sequence Each Fragment
with Shotgun Approach

```
GCATTTCGAGTTACCTGGACAACCAGTG
                            CCAGTGGTACTGAGGACGCAAGAGGCTTGA
                                                        GCTTGATTGGCCAATAATAGTATAT
```

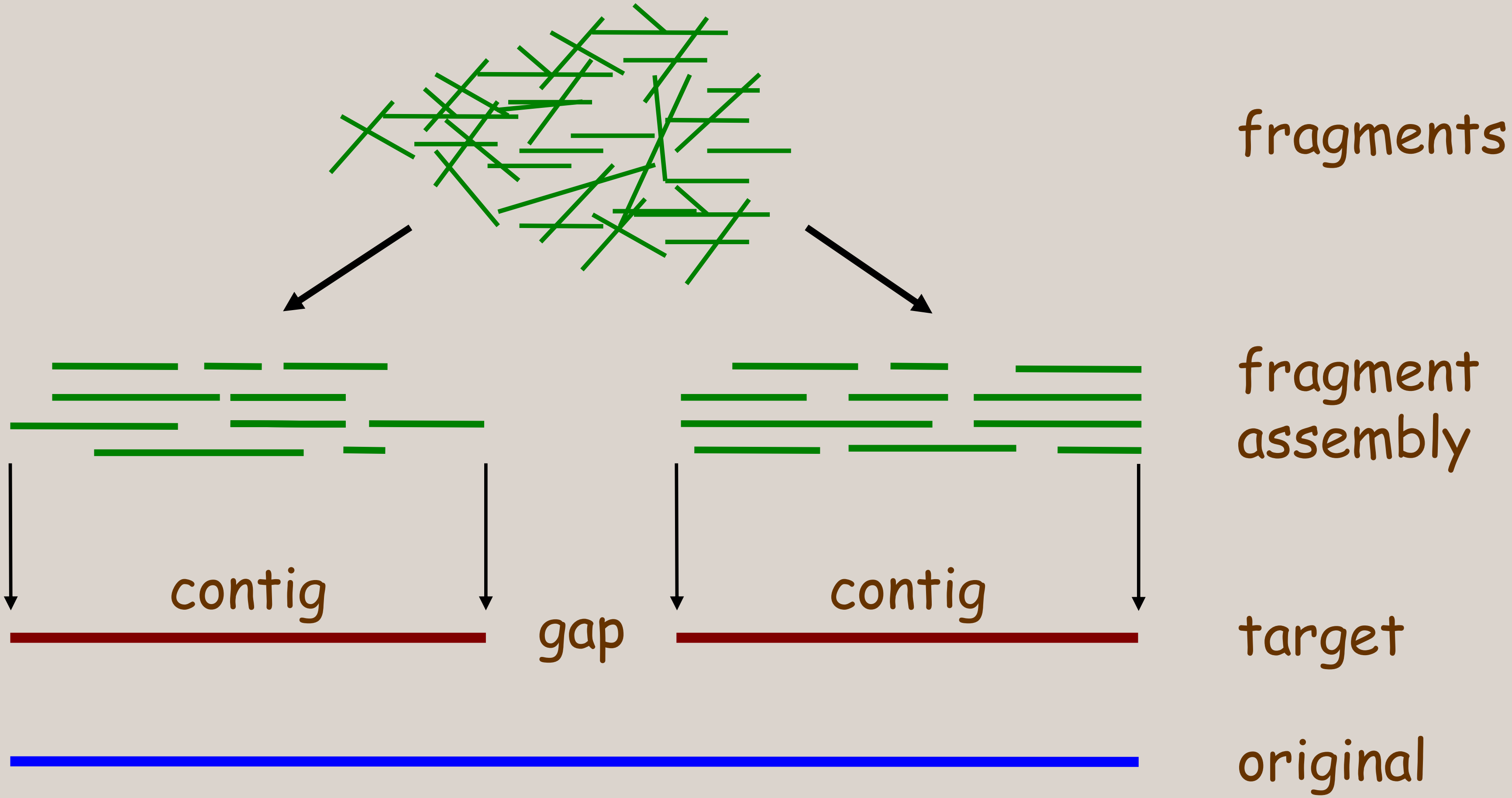
Align Contiguous Sequences

```
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
```

Generate Finished Sequence

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

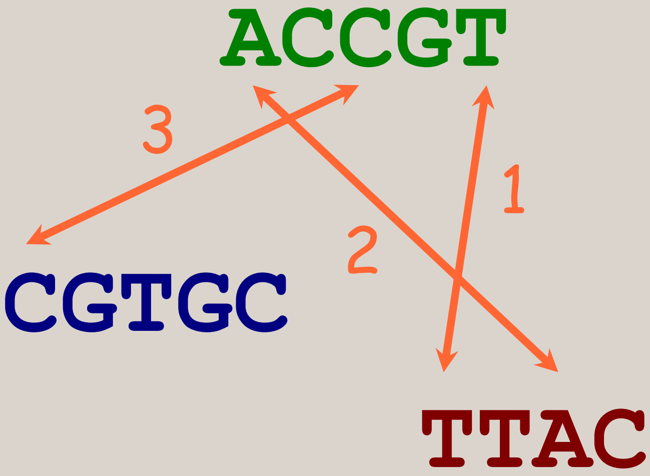
Sequence Assembly



Sequence Assembly

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

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



ACCGT

3

1

2

CGTGC

TTAC

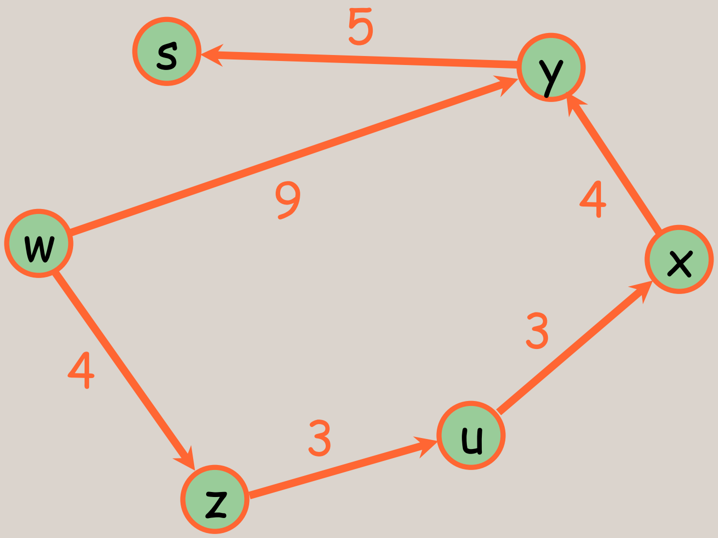
```
--ACCGT--
----CGTGC
TTAC-----
_________
TTACCGTGC
```

Sequence Assembly

Sketch of algorithm:

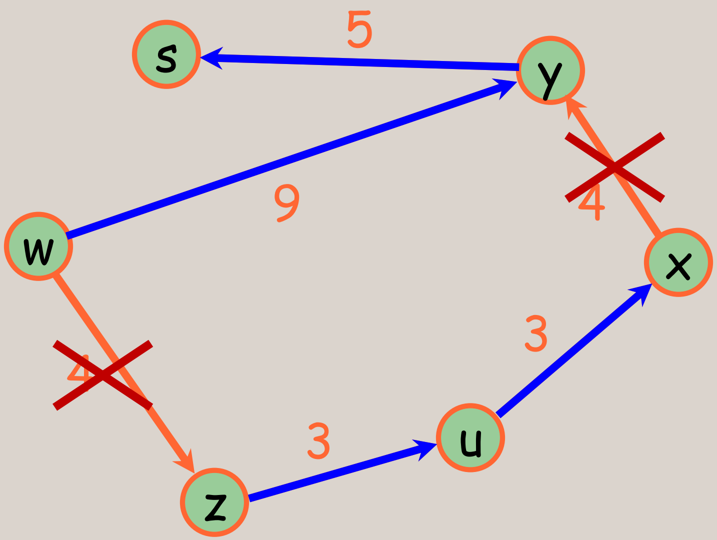
- Create an *overlap graph* in which every node represents a fragment and edges indicate overlap.
- Determine which overlaps will be used in final assembly: find an *optimal spanning forest* in overlap graph.

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 each step.
- Edge must connect nodes with in- and out-degrees <= 1.
- May end up with set of paths: each yields a contig.



W→Y→S

```
AGTATTGGCAATC
    TTGGCAATCA
         AATCAGG
----------------
AGTATTGGCAATCAGG
```

Z→U→X

```
AATCGATG
     ATGCAAACCT
            CCTTTTGG
--------------------
AATCGATGCAAACCTTTTGG
```

Sequence Comparison

What's the problem? Kind of like google for biologists ...

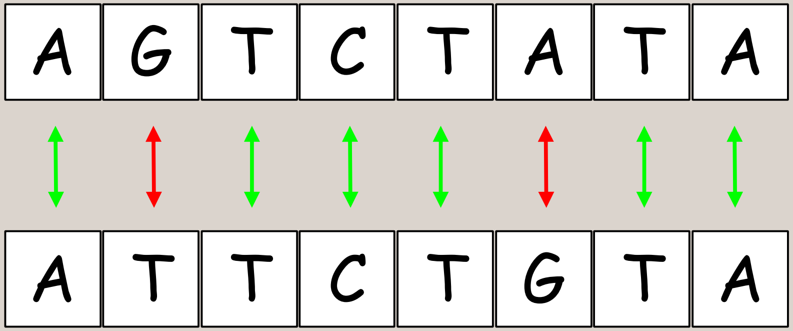
- Given new DNA or protein sequence, biologist will want to search databases of known sequences for similarities.
- Sequence similarity can provide clues about function and evolutionary relationships.
- Databases such as GenBank are too big for manual search. To search them efficiently, we need an algorithm.

Can't expect exact matches (i.e., not really like google):

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

Sequence Comparison

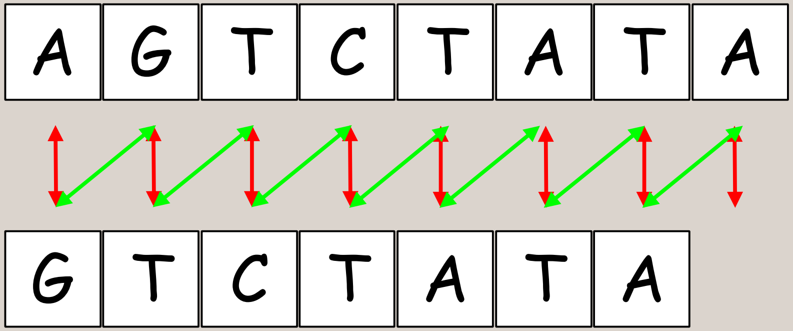
Why not just line up sequences and count matches?



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

→ Difference = 2

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



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

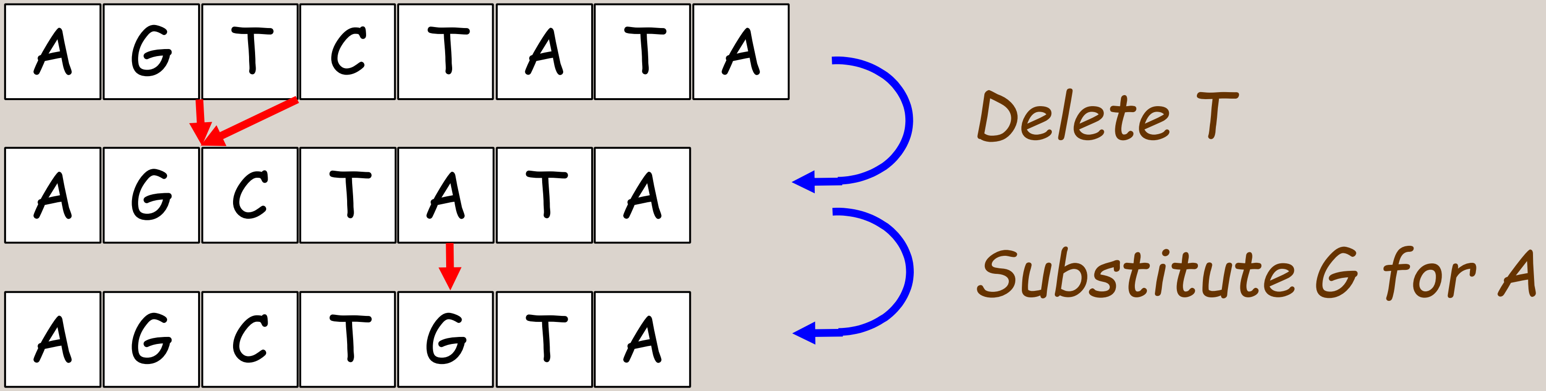
→ Difference = 8

One missing symbol at start leads to large difference!

Sequence Comparison

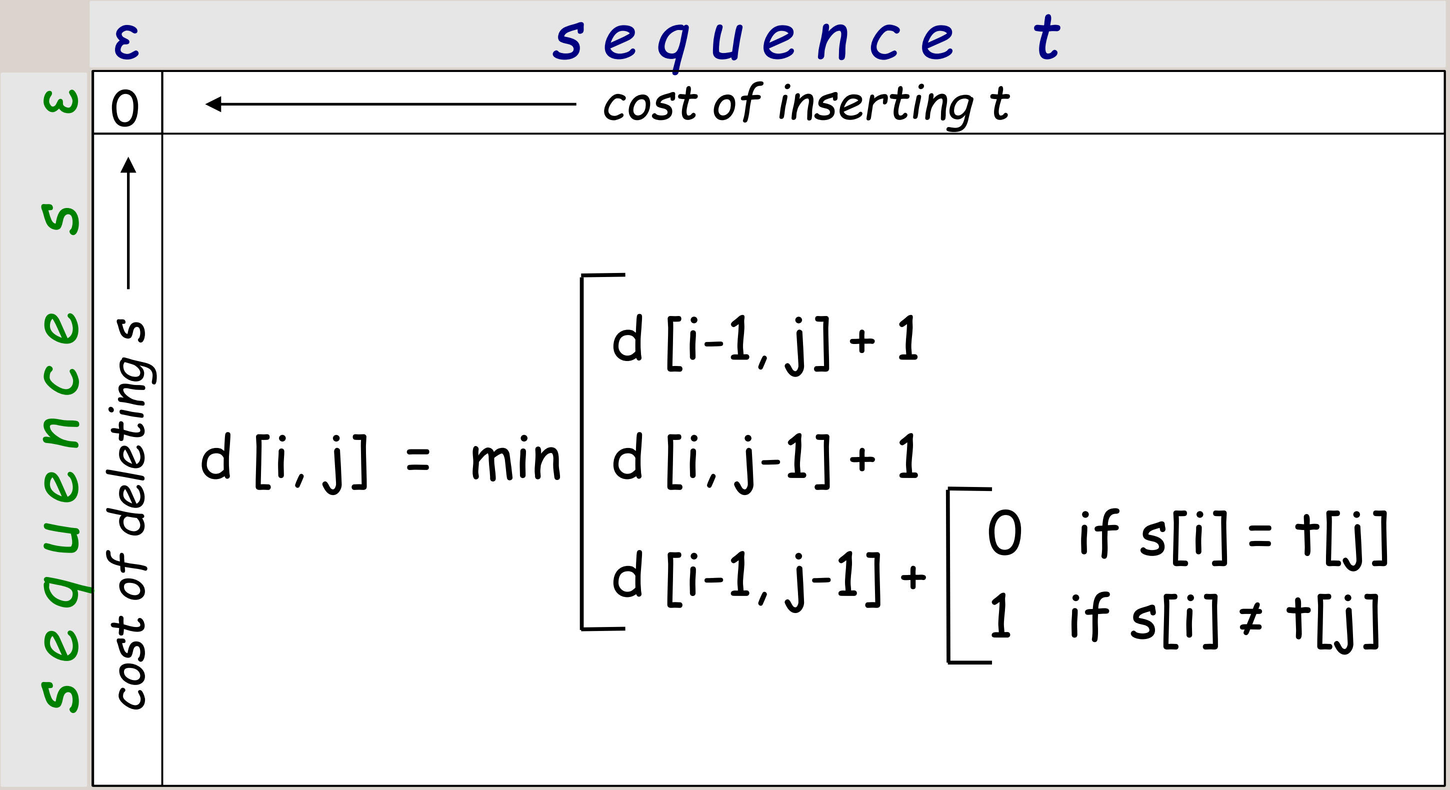
Instead, we'll use technique known as *dynamic programming.*

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



Sequence Comparison

Elegant optimization algorithm builds table of values, working from shorter prefixes to longer prefixes:



ε sequence t

ε sequence s

0

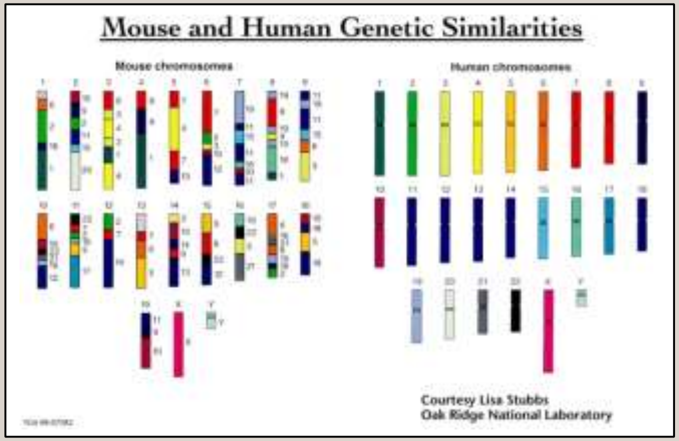
cost of inserting t

cost of deleting s

$$d[i, j] = \min \begin{cases} d[i-1, j] + 1 \\ d[i, j-1] + 1 \\ d[i-1, j-1] + \begin{cases} 0 & \text{if } s[i] = t[j] \\ 1 & \text{if } s[i] \neq t[j] \end{cases} \end{cases}$$

Genome Rearrangements

Recall what we saw earlier:



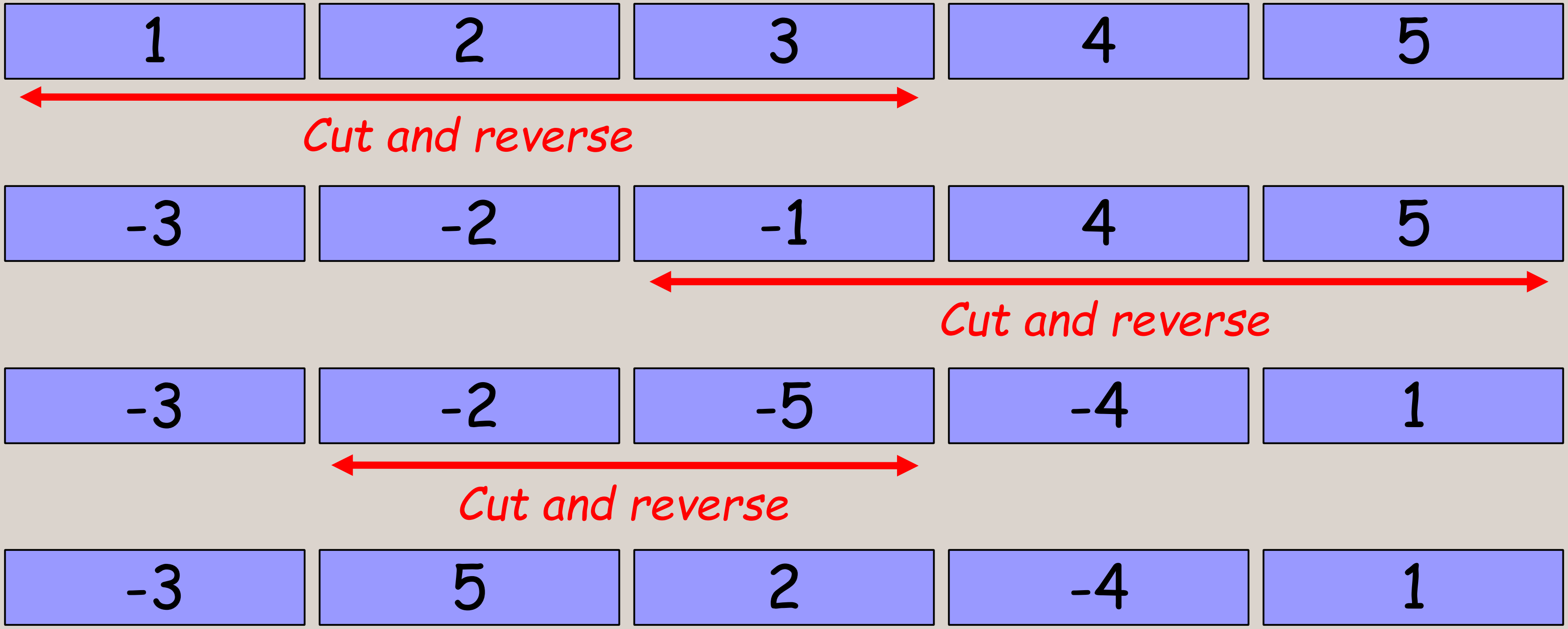
Mouse and Human Genetic Similarities

Courtesy Lisa Stubbs
Oak Ridge National Laboratory

- 99% of mouse genes have homologues in human genome.
- 96% of mouse genes are in same relative location.
- 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



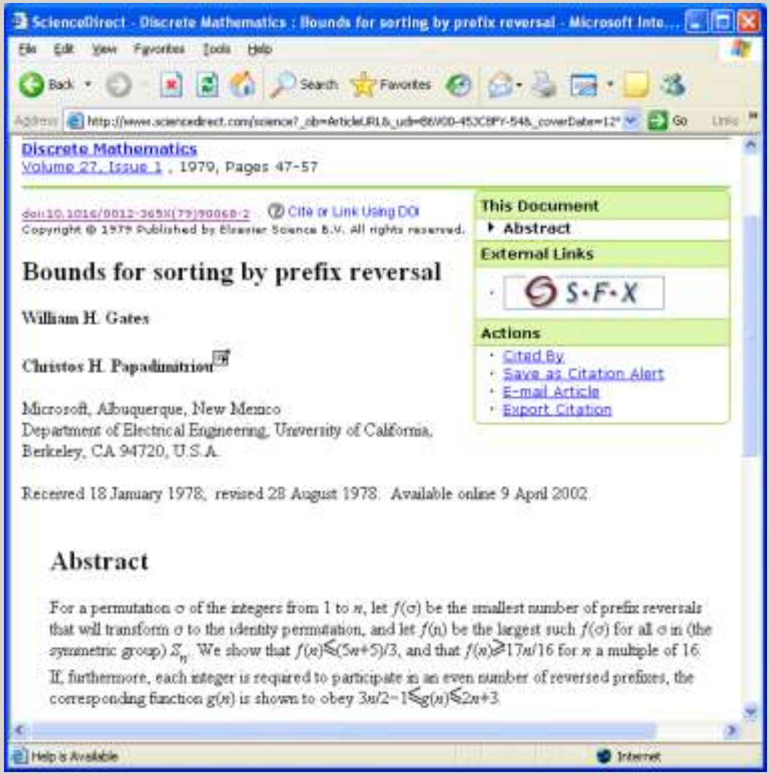
1	2	3	4	5
-3	-2	-1	4	5
-3	-2	-5	-4	1
-3	5	2	-4	1

Mouse Chromosome X

Reversal distance is minimum number of steps needed.

Interesting Sidenote

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



ScienceDirect - Discrete Mathematics : Bounds for sorting by prefix reversal - Microsoft Inte...
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V00-453C8FY-548_coverDate=12" Go Links

Discrete Mathematics
Volume 27, Issue 1, 1979, Pages 47-57

doi:10.1016/0012-365X(79)90068-2 Cite or Link Using DOI
Copyright © 1979 Published by Elsevier Science B.V. All rights reserved.

This Document
Abstract
External Links
S•F•X
Actions
Cited By
Save as Citation Alert
E-mail Article
Export Citation

Bounds for sorting by prefix reversal

William H. Gates

Christos H. Papadimitriou

Microsoft, Albuquerque, New Mexico
Department of Electrical Engineering, University of California, Berkeley, CA 94720, U.S.A.

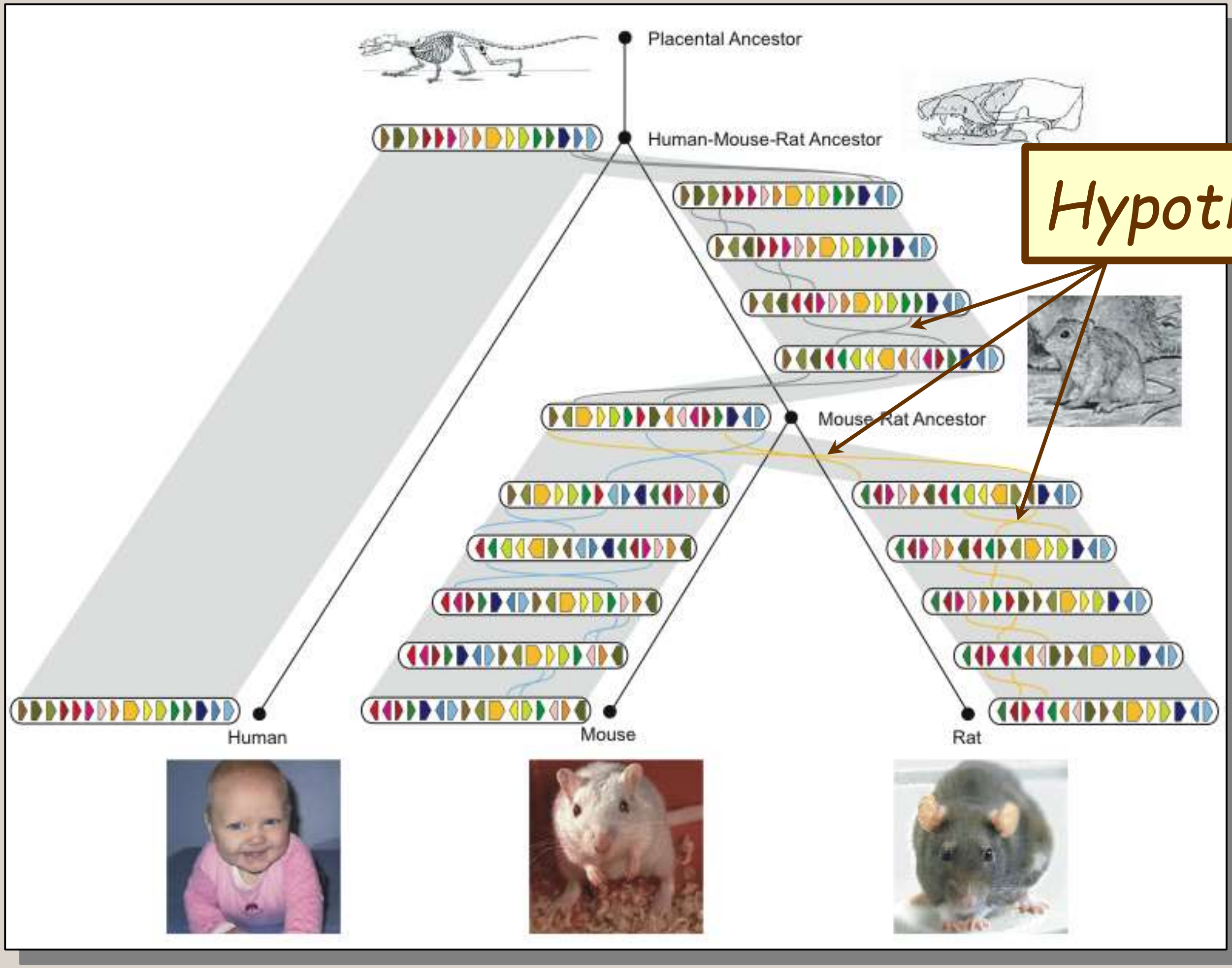
Received 18 January 1978; revised 28 August 1978. Available online 9 April 2002.

Abstract

For a permutation σ of the integers from 1 to n, let $f(\sigma)$ be the smallest number of prefix reversals that will transform σ to the identity permutation, and let $f(n)$ be the largest such $f(\sigma)$ for all σ in (the symmetric group) S_n. We show that $f(n) \leqslant (5n+5)/3$, and that $f(n) \geqslant 17n/16$ for n a multiple of 16. If, furthermore, each integer is required to participate in an even number of reversed prefixes, the corresponding function $g(n)$ is shown to obey $3n/2-1 \leqslant g(n) \leqslant 2n+3$.

Yes, that Bill Gates ...

History of Chromosome X

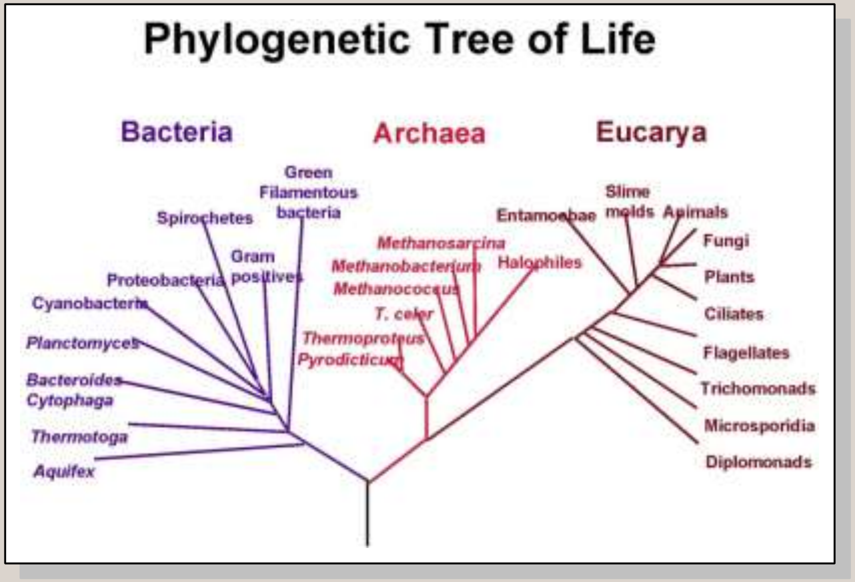


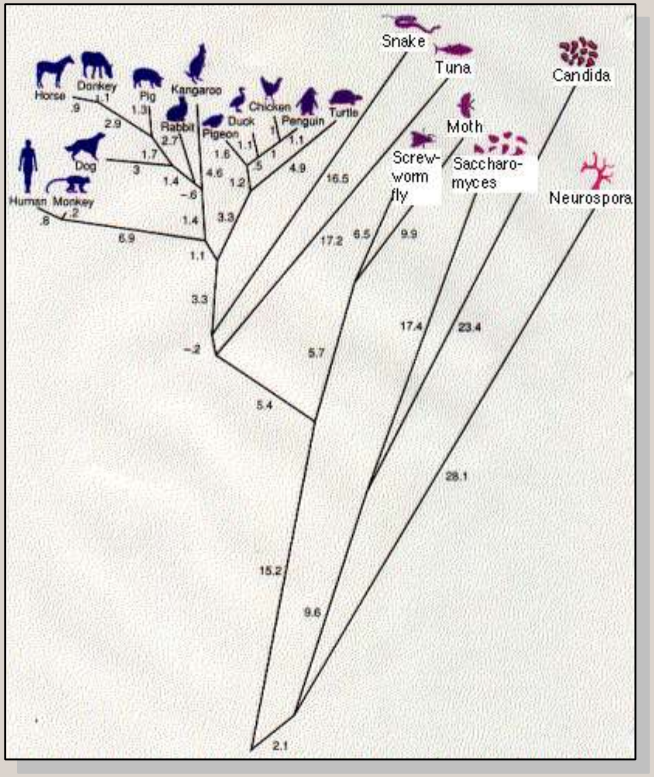
Hypothesized reversals

Rat Consortium, Nature, 2004

Building the "Tree of Life"

Scientists build phylogenetic trees to help understand evolutionary relationships. Reversal distance often used.





http://en.wikipedia.org/wiki/Phylogenetic_tree

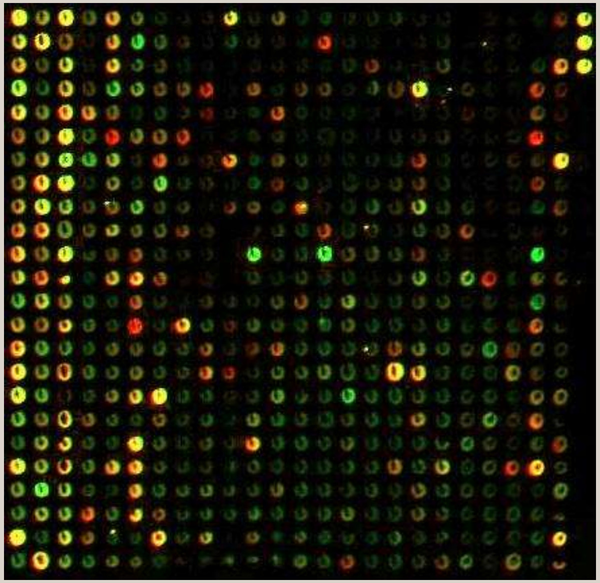
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/Taxonomy.html

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

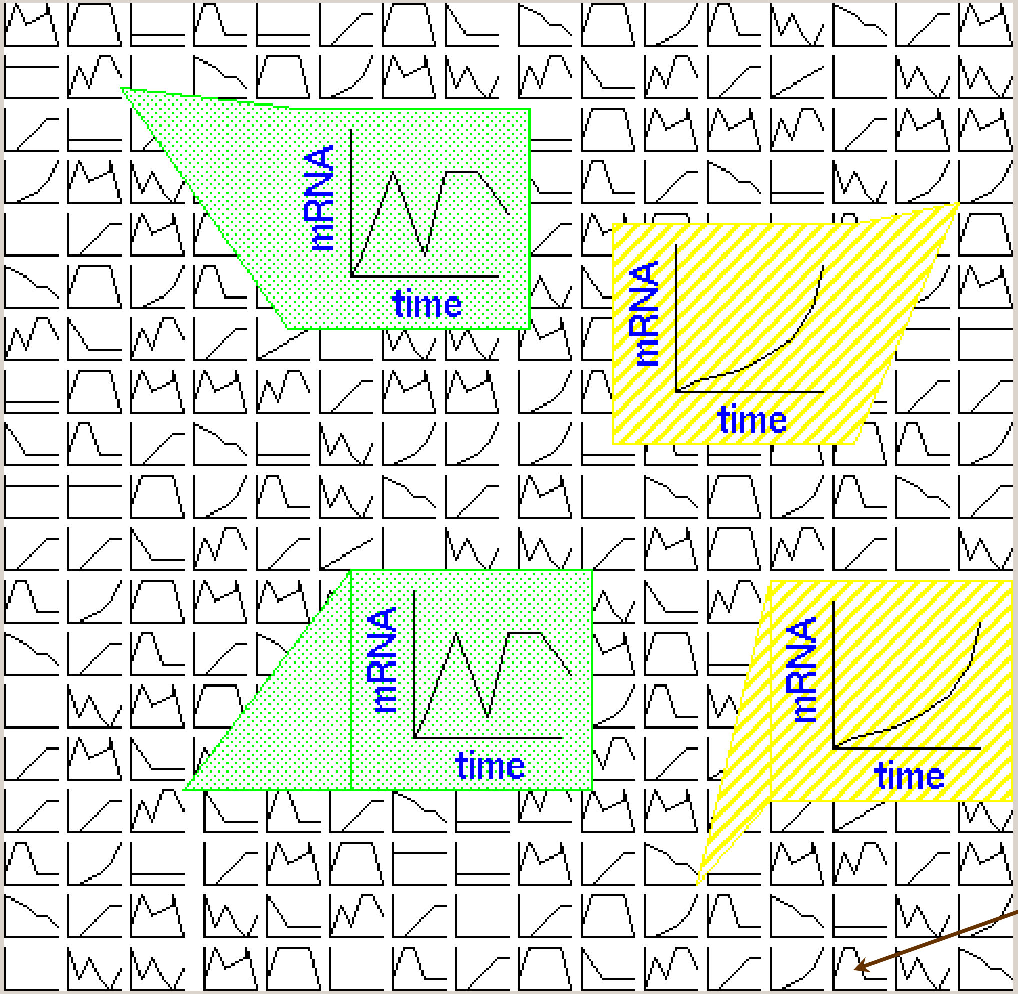
DNA Microarrays

- Allows simultaneous measurement of transcription level for every gene in a genome (gene expression).
- Differential expression, want to find genes that behave similarly over time.
- One microarray can test ~10k genes.
- Data obtained much faster than we can process it!
- Must find ways to uncover patterns.

green = repressed
red = induced



Using DNA Microarrays



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

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

Each cell represents one gene's expression over time

DNA Microarrays

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

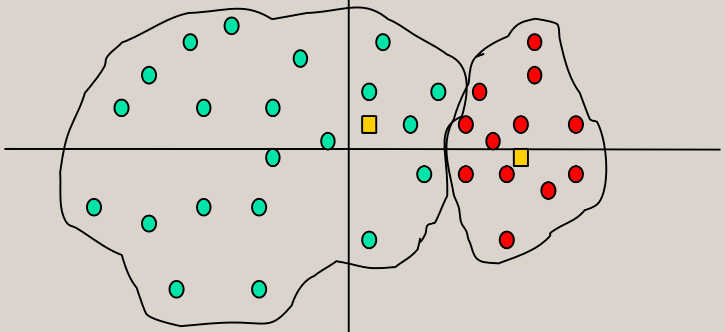
- Given set of *n* data points and an integer *k*.
- We want to find set of *k* points that minimizes mean-squared distance from each data point to nearest center.

Sketch of algorithm:

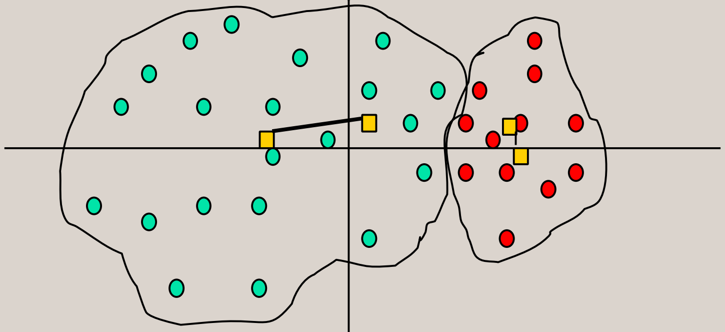
- Choose *k* initial center points randomly and cluster data.
- Calculate new centers for clusters using points in cluster.
- Re-cluster all data using new center points.
- Repeat second two steps until no data points change clusters, or some other convergence criterion is met.

Clustering Microarray Data

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



- Re-calculate centers based on current clusters.

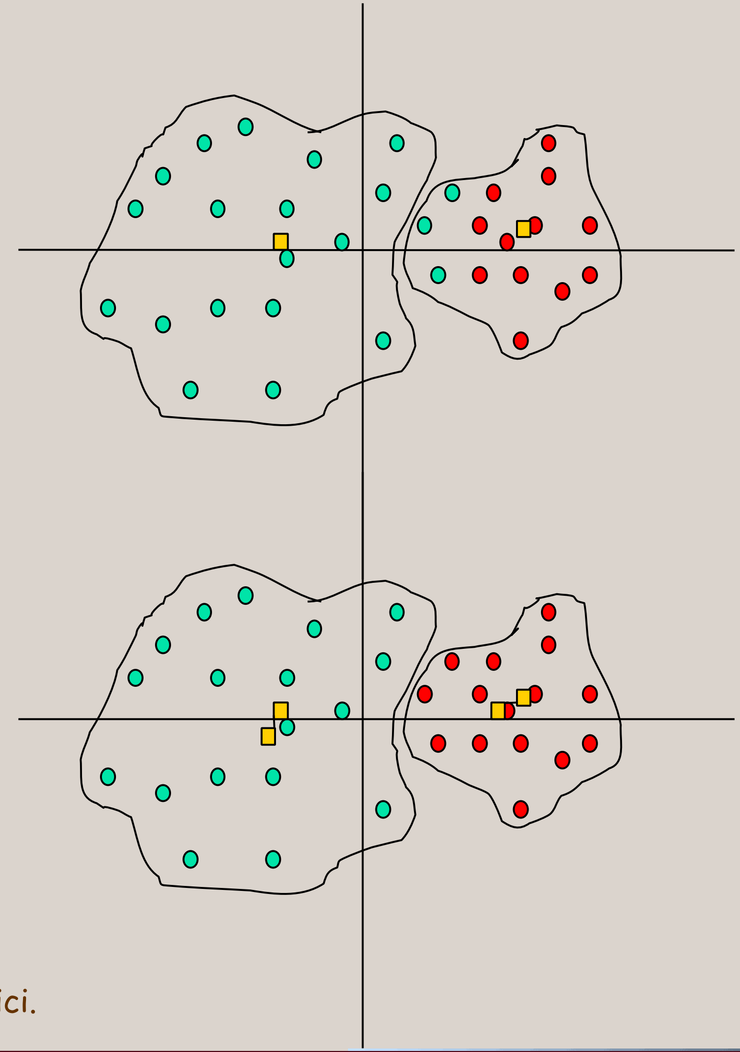


From "Data Analysis Tools for DNA Microarrays" by Sorin Draghici.

Clustering Microarray Data

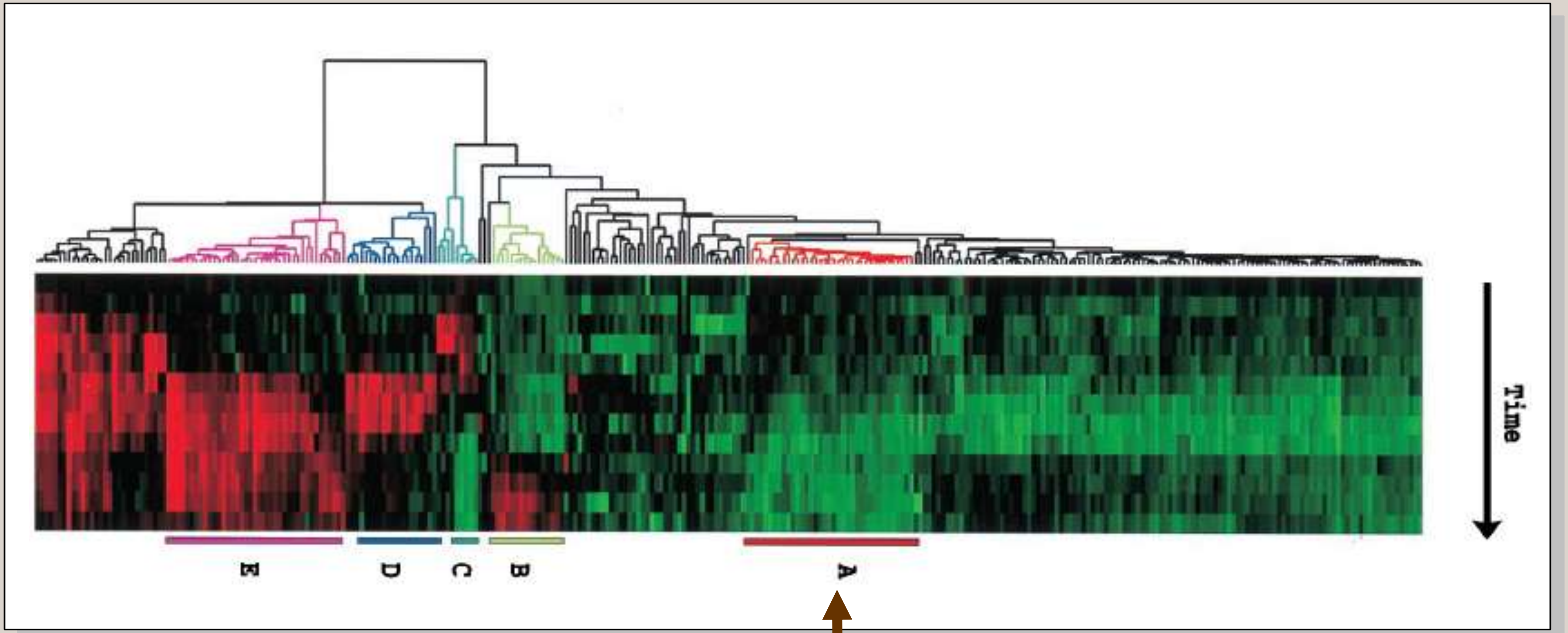
- Re-cluster data around new center points.

- Repeat last two steps until no data points change clusters.



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?

- Many unanswered questions ⇒ opportunities to make fundamental contributions (+ become rich and famous).
- Stretch your creativity and problem-solving skills.
- Cross-disciplinary teams: work with interesting people.
- Participate in unlocking the mysteries of life itself.
- Make the world a better place.

Intro to Bioinformatics

Prof. Brian Chen

CSE 308 / BioE 308 covers:

- Intro to molecular biology & algorithms,
- Genetic sequence comparison & alignment,
- Sequencing & assembly of DNA,
- DNA microarrays,
- Gene regulatory networks,
- Genome annotation,
- Transcription factor binding site prediction,
- Standard formats and sources for genomic data, etc.

CSE 308 is not a programming course! It's for BioS, BioE, CSE, and Math students.

Questions: chen@cse.lehigh.edu

Structural Bioinformatics

Prof. Brian Chen

CSE 307 / BioE 307 covers:

- Geometric modeling for proteins,
- Structure alignment & protein folding,
- Protein surfaces, cavities, electrostatics,
- Protein-protein and protein-DNA
- Interfaces and interactions,
- Protein structure prediction, simulation, docking,
- Structural bioinformatics in pharmaceutical discovery,
- Function annotation, active site prediction, etc.

For seniors in BioS, BioE, CSE, and Math.

Questions: chen@cse.lehigh.edu

HHMI
Howard Hughes Medical Institute

LEHIGH UNIVERSITY.

Computer Science and Engineering

Thank you!

Beat Lafayette!