

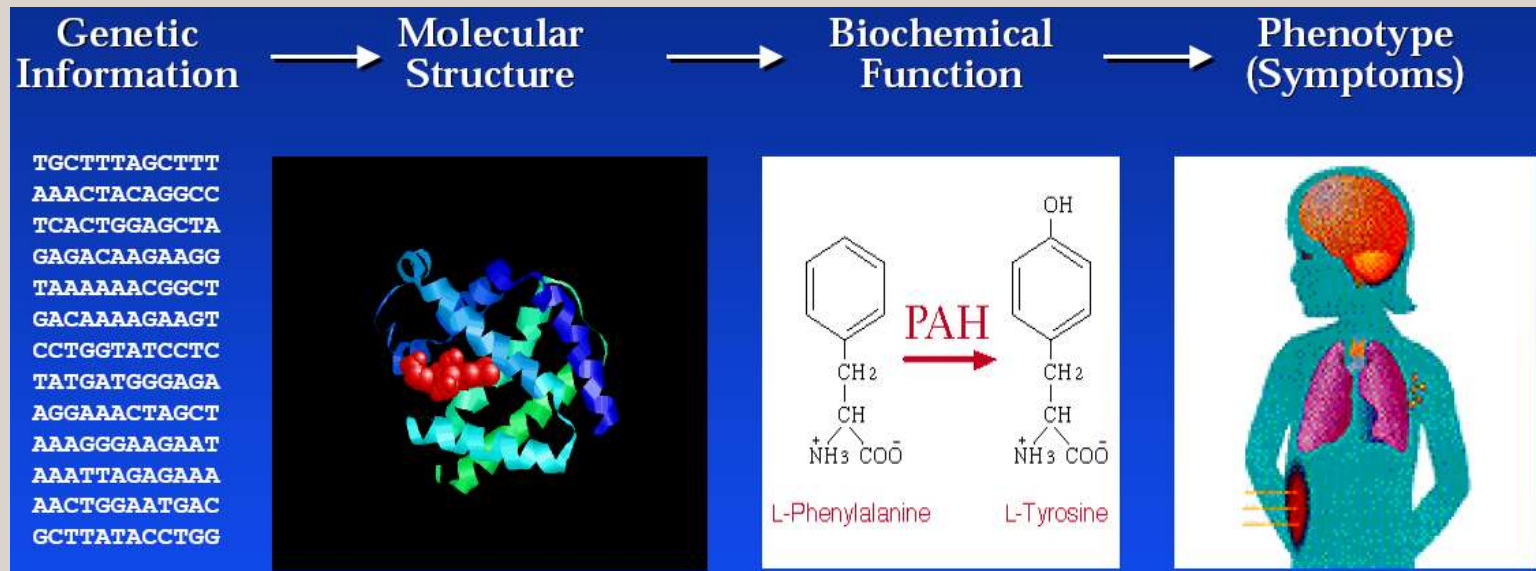
A Brief Introduction to Bioinformatics



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)

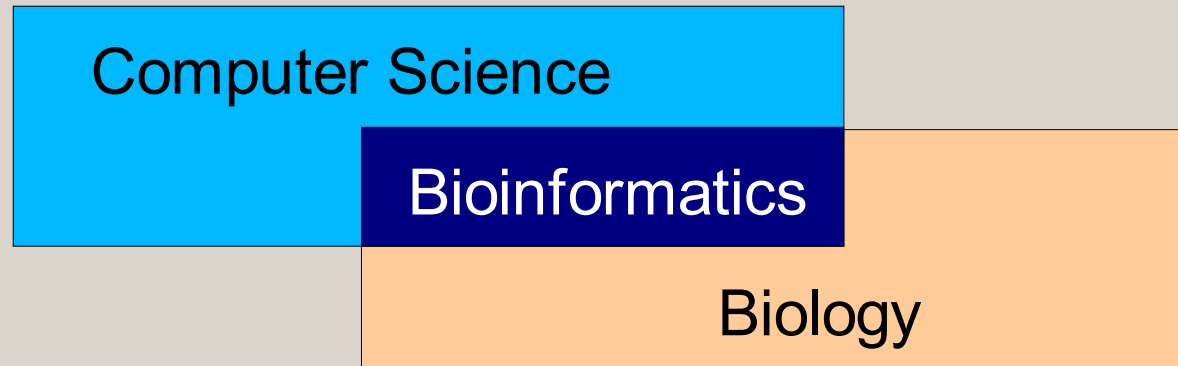


By developing techniques for analyzing sequence data and related structures, we can attempt to understand genetic nature of diseases.

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

Bioinformatics

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



Why is it interesting?

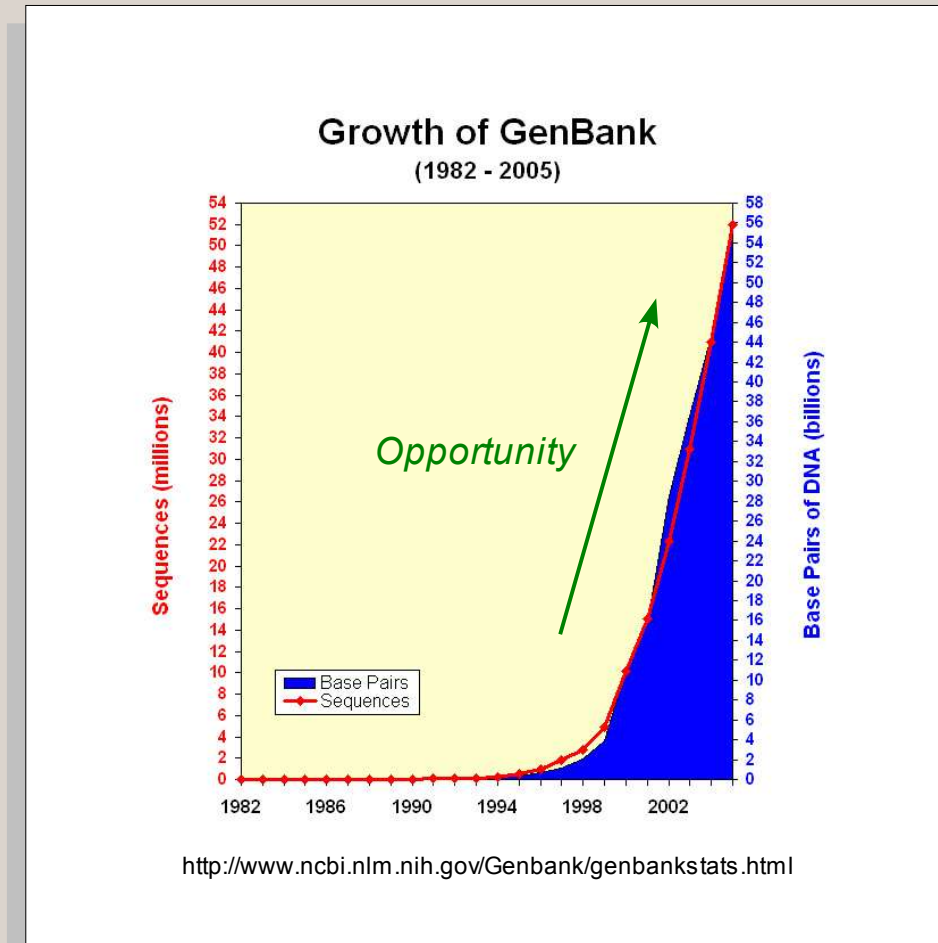
- Important problems.
- Massive quantities of data.
- Desperate need for efficient solutions.
- Success is rewarded.

Data Explosion

Genetic identity of most organisms 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 language over 4 character alphabet, $\{A, C, G, T\}$.



Genomes

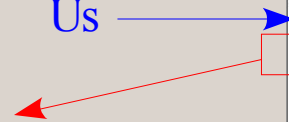
Complete set of chromosomes that determines an organism is known as its *genome*.



Mus musculus



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	400,000,000	125,948,974	151,411
Rattus norvegicus	2,900,000,000	106,344,366	218,598
Bos taurus	3,651,500,000	71,215,626	159,473
Glycine max	1,115,000,000	62,817,102	141,802
Medicago truncatula	400,000,000	50,991,920	104,535
Trypanosoma brucei	35,000,000	49,855,996	91,334
Lycopersicon esculentum	655,000,000	49,415,566	97,112
Giardia intestinalis	12,000,000	47,639,714	54,328
Strongylocentrotus purpur	900,000,000	47,590,936	77,532
Entamoeba histolytica	—	44,522,016	49,938
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

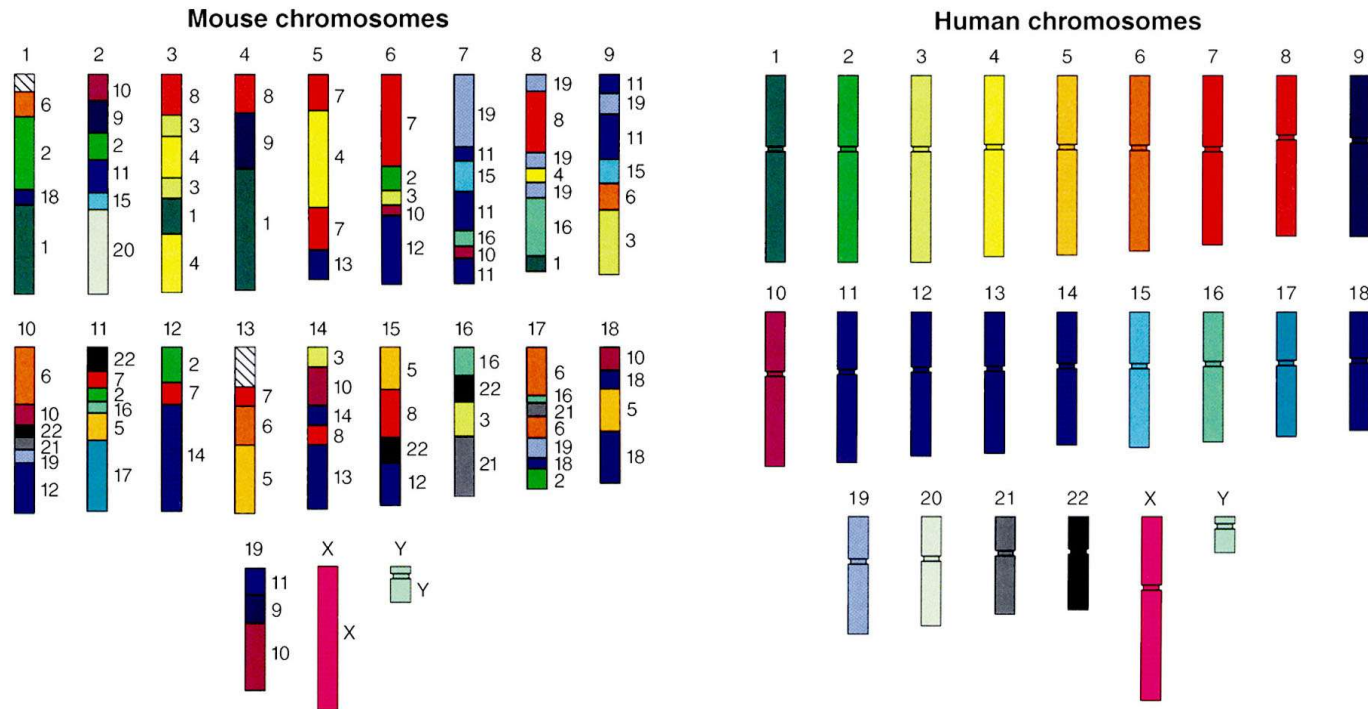
<http://www.cbs.dtu.dk/databases/DOGS/>

http://www.nsrll.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

YGA 98-075R2

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

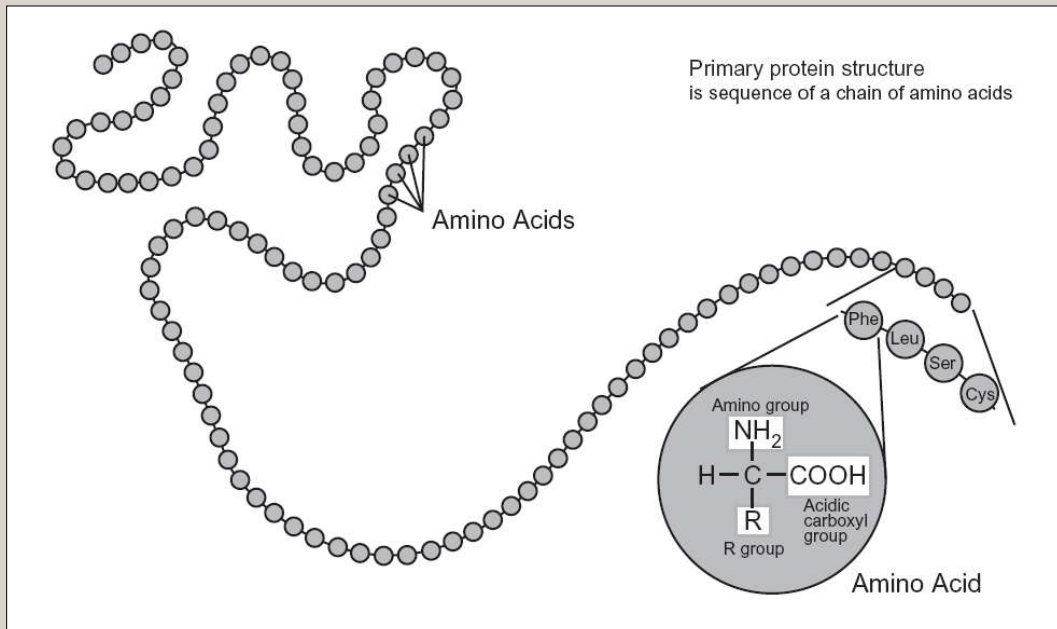
Algorithms are Central

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).

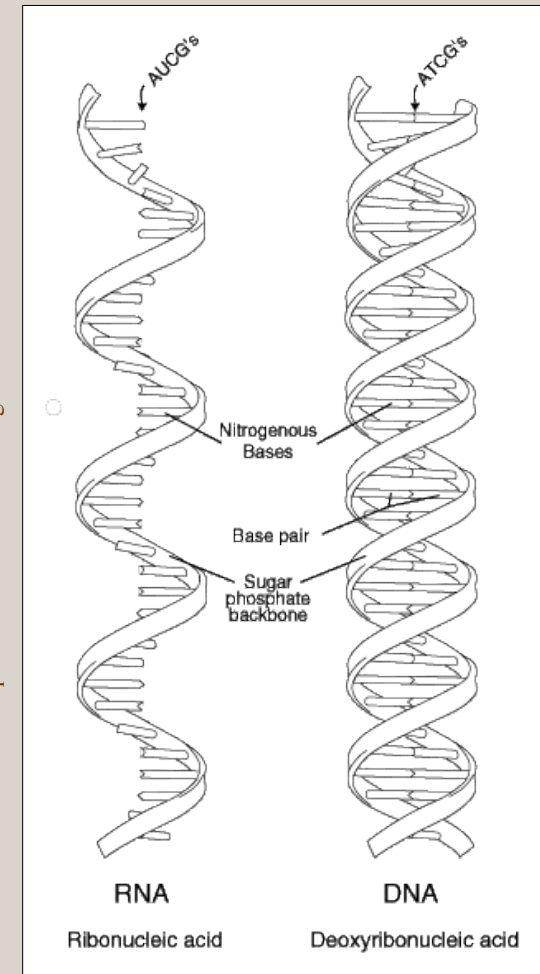
Sequence Nature of Biology

Macromolecules are chains of simpler molecules.



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

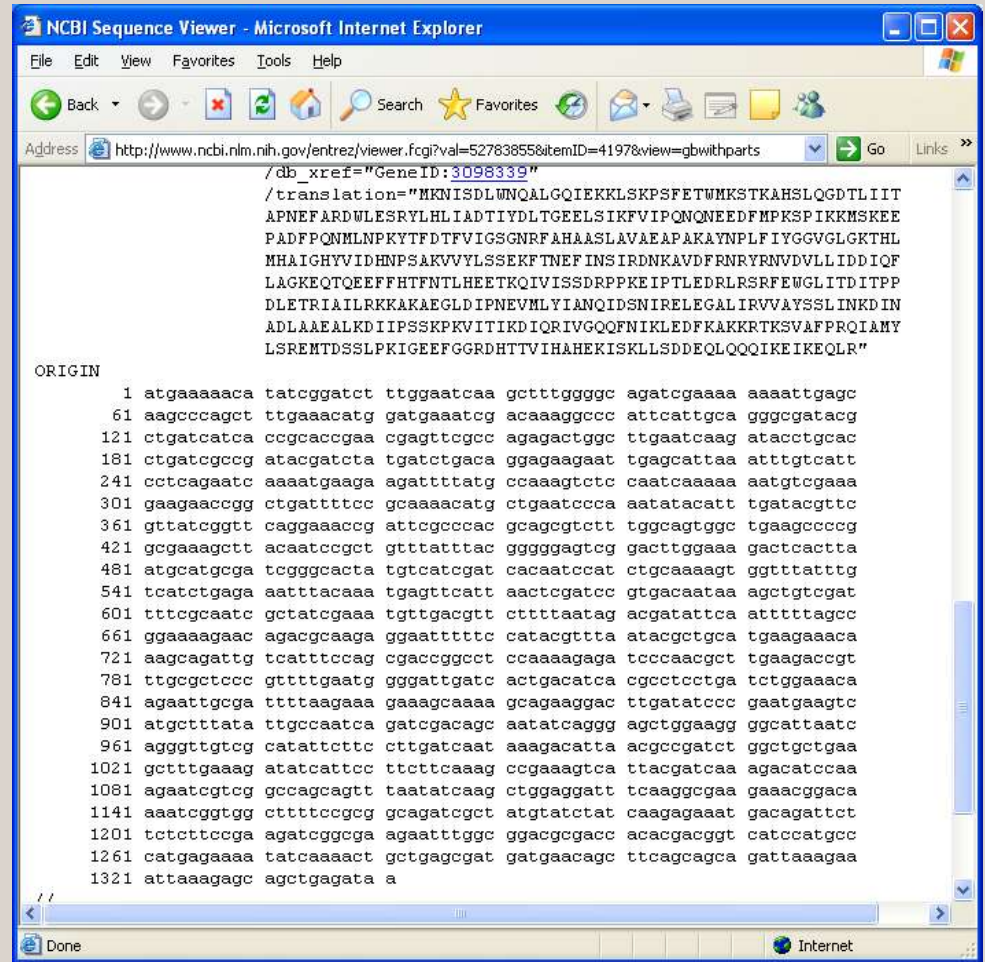
In DNA and RNA, they are *nucleotides*.



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/>



The screenshot shows a web browser window titled "NCBI Sequence Viewer - Microsoft Internet Explorer". The address bar contains the URL: <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=52783855&itemID=4197&view=gbwithparts>. The main content area displays the following information:

```
/db_xref="GeneID:3098339"
/translation="MKNISDLUNQALGQIEKLSKPSFETWMSKTKAHSLOQDTLIIT
APNEFARDWLESRYLHLIADTIYDLTGEELS IKFVIPQNQNEEDFMPKSP IKKMSKEE
PADFPQNMLNPKYTFDTFVIGSGNRF AHAASLAVAEAPAKAYNPLF IYGGVGLGKTHL
MHAIGHYVIDHNPSAKVYVLSSEKFTNEFINS IRDNKAVDFRNRVYRVVDVLLIDDIQF
LAGKEQTQEEFFHTFNTLHEETKQIVISSDRPKEIPTLEDRLRSRFEMGLITDITPP
DLETRIALLRKKAKAEGLDIPNEVMLYIANQIDSNIRELEGALIRVVAYSSLINKDIN
ADLAAELKDIIPSSPKVITIKDIQRIVGQQFNKLEDFKAKKRTKSVAFPRQIAMY
LSREMTDSSLPKIGIEFGGRDHTTVIHAHERISKLLSDDEQLQQQIKEIKEQLR"
```

ORIGIN

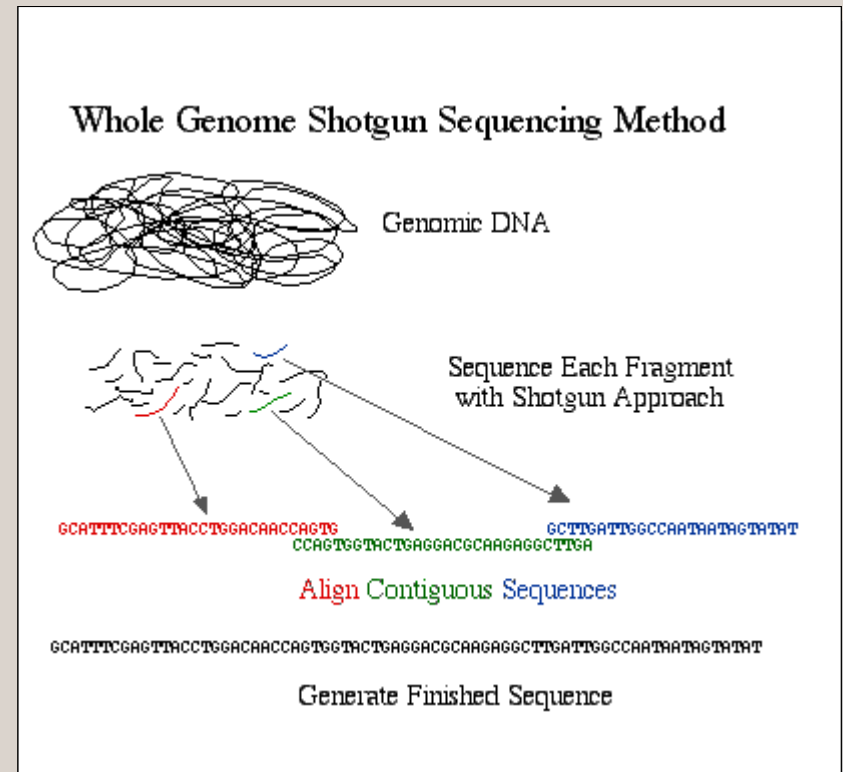
```
1 atgaaaaaca tatcggtatc ttggaatcaa gctttggggc agatcgaaaa aaaattgagc
61 aagcccagct ttgaaacatg gatgaaatcg acaaaggccc attcattgca gggcgcatac
121 ctgatcatca ccgcaccgaa cgagttcgcc agagactggc ttgaatcaag atacctgac
181 ctgatcgccc atacgatcta tgatctgaca ggagaagaat tgaccataaa atttgtcatt
241 cctcagaatc aaaaatgaaga agattttatg ccaaagtctc caatcaaaaa aatgtcgaaa
301 gaagaaccgg ctgattttcc gcaaaacatg ctgaatccca aatatacat tgatcagttc
361 gttatcggtt caggaaaccc attcgcccac gcagcgtctt tggcagtgcc tgaagccccg
421 gcgaaagcct acaatccgct gtttatttac gggggagtcg gacttgaaaa gactcactta
481 atgcatcgca tcgggcacta tgcctcatcg cacaatccat ctgcaaaaag ggtttatttg
541 tcactctgaga aatttacaaa tgagttcatt aactcgatcc gtgacaataa agctgtcgat
601 tttcgcaatc gctatcgaaa tgttgacgtt cttttaatag acgatattca atttttagcc
661 ggaaaagaac agacgcaaga ggaatttttc catacgttta atacgttca tgaagaaaca
721 aagcagatgg tcatttccag cgaccggcct ccaaaagaga tcccacgct tgaagaccgt
781 ttgcgctccc gttttgaatg gggattgac actgacatca cgcctcctga tctggaaa
841 agaattgcga ttttaagaaa gaaagcaaaa gcagaaggac ttgatatccc gaatgaagtc
901 atgctttata ttgccaatca gatcgacagc aatatacagg agctggaagg ggcattaatc
961 aggttgtgct catattcttc cttgatcaat aaagacatta acgccgatct ggctgctgaa
1021 gctttgaaag atatcattcc tcttccaag cggaaagtea ttacgatcaa agacatccaa
1081 agaatcgctg gccagcagtt taatatcaag ctggaggatt tcaaggcgaa gaaacggaca
1141 aaatcggtgg cttttccgcg gcagatcgct atgtatctat caagagaaat gacagattct
1201 tctcttccga agatcggcga agaatttggc ggacgggacc acacagcgtt catccatgcc
1261 catgagaaaa tatcaaaact gctgagcgat gatgaacagc ttcacagcga gattaagaa
1321 attaaaagac agctgagata a
```

Sequencing a Genome

Genomes are determined using a technique known as *shotgun sequencing*.

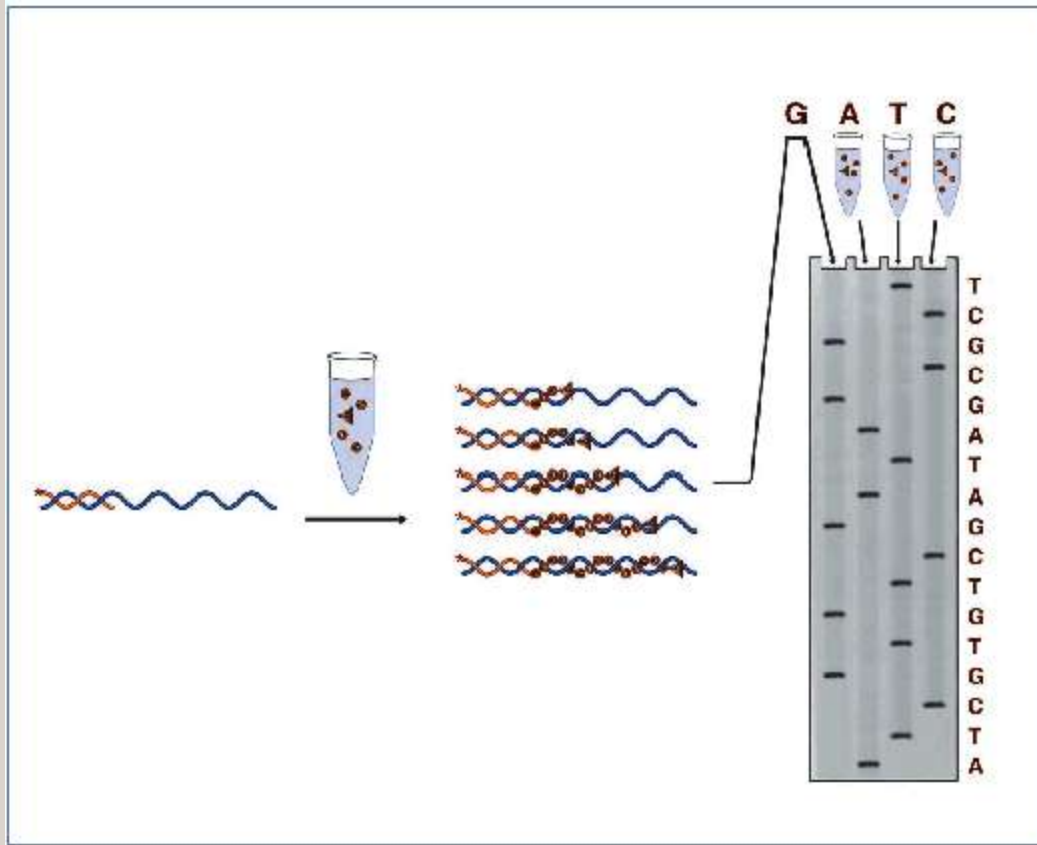
Computer scientists have played an important role in developing algorithms for assembling such data.

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



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

Reading DNA



This is known as *Sanger sequencing*.

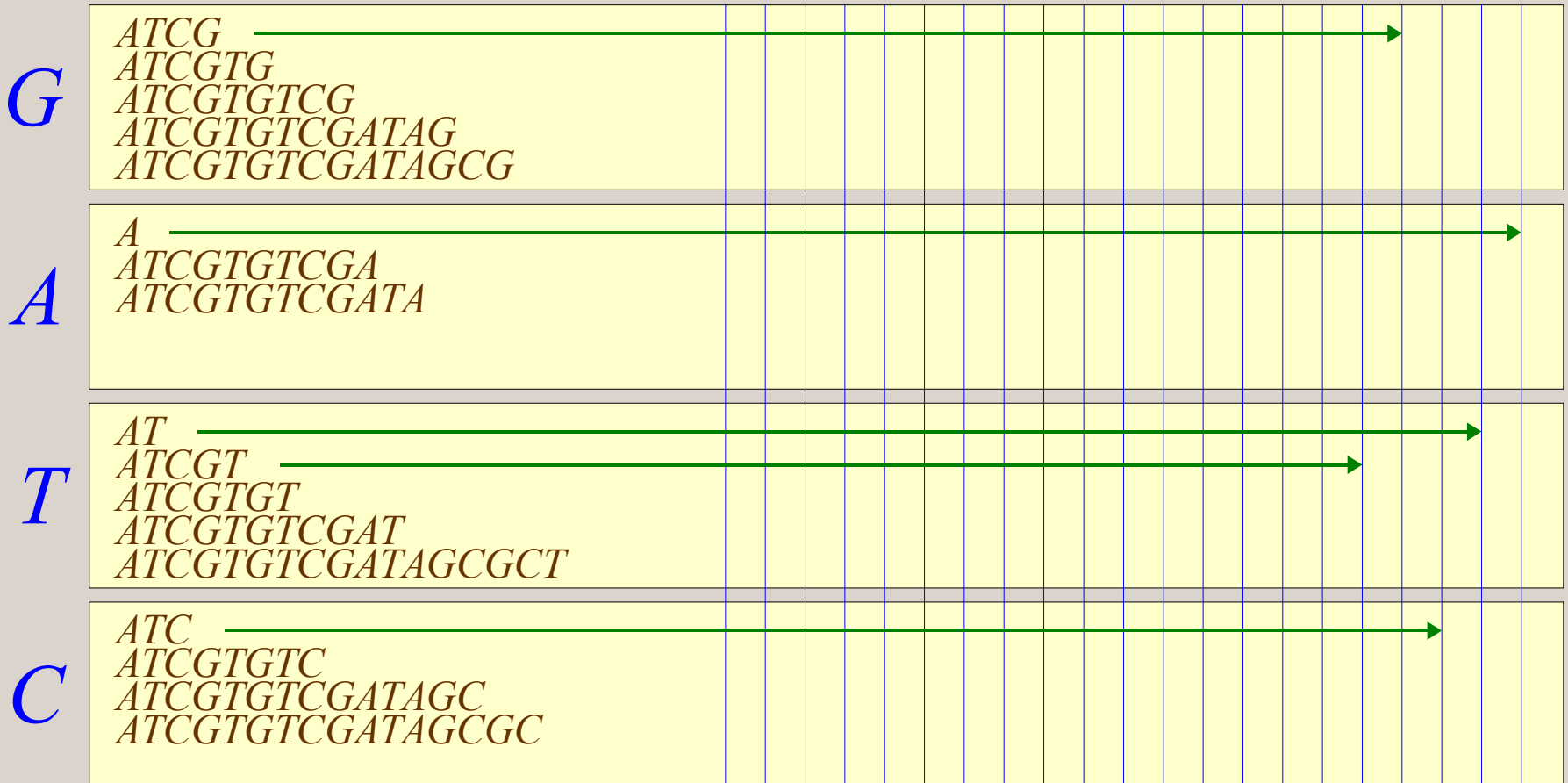
<http://www.apelx.fr/anglais/applications/sommaire2/sanger.htm>
<http://www.iupui.edu/~wellsctr/MMIA/html/animations.htm>

Gel electrophoresis is a 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.

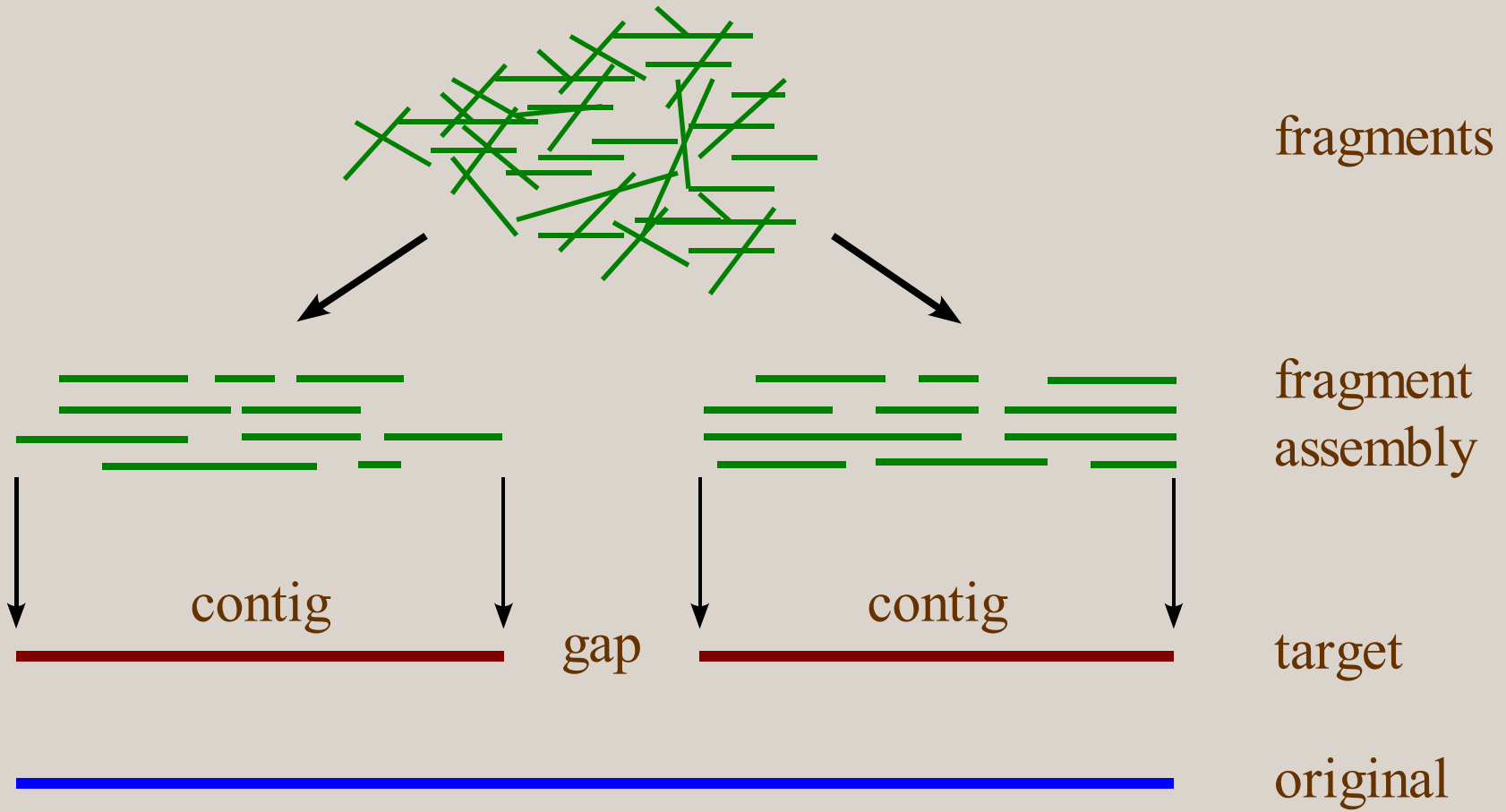
First "starve" DNA copy reaction at each base: *A, C, G, T*. Then run gel and read off sequence: *ATCGTG ...*

Reading DNA

Original sequence: *ATCGTGTCGATAGCGCT*



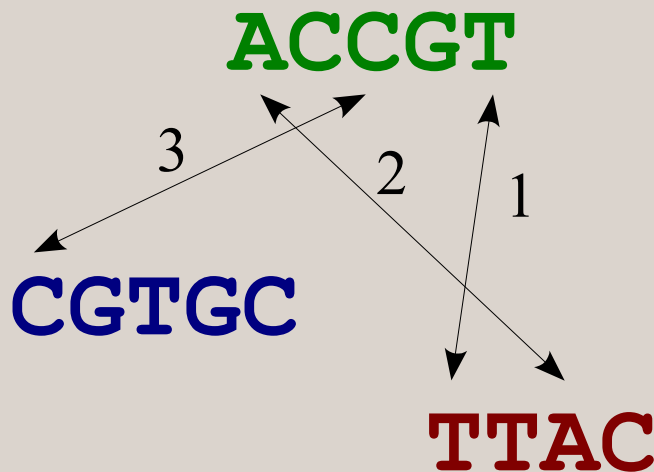
Sequence Assembly



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

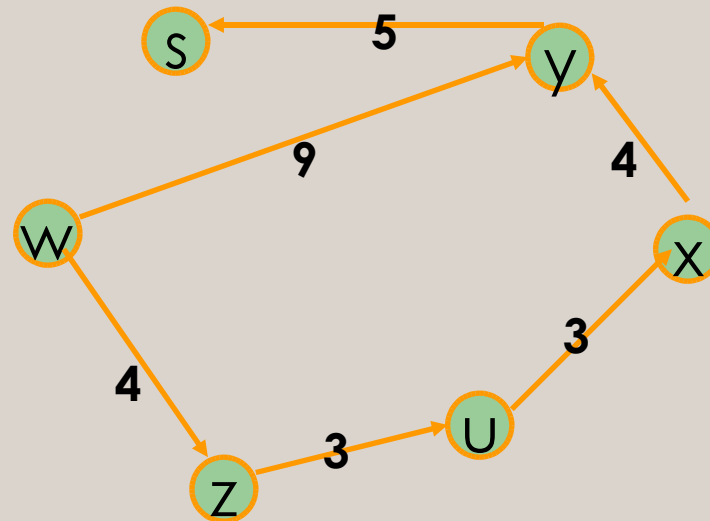
Z = AATCGATG

U = ATGCAAACCT

X = CCTTTTGG

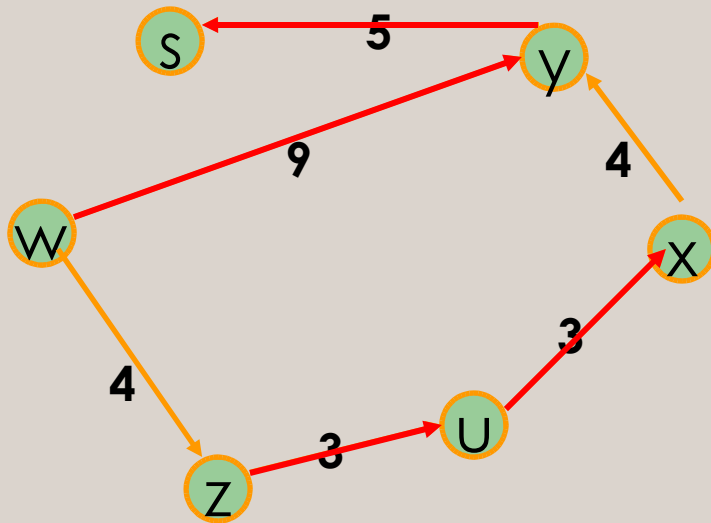
Y = TTGGCAATCA

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

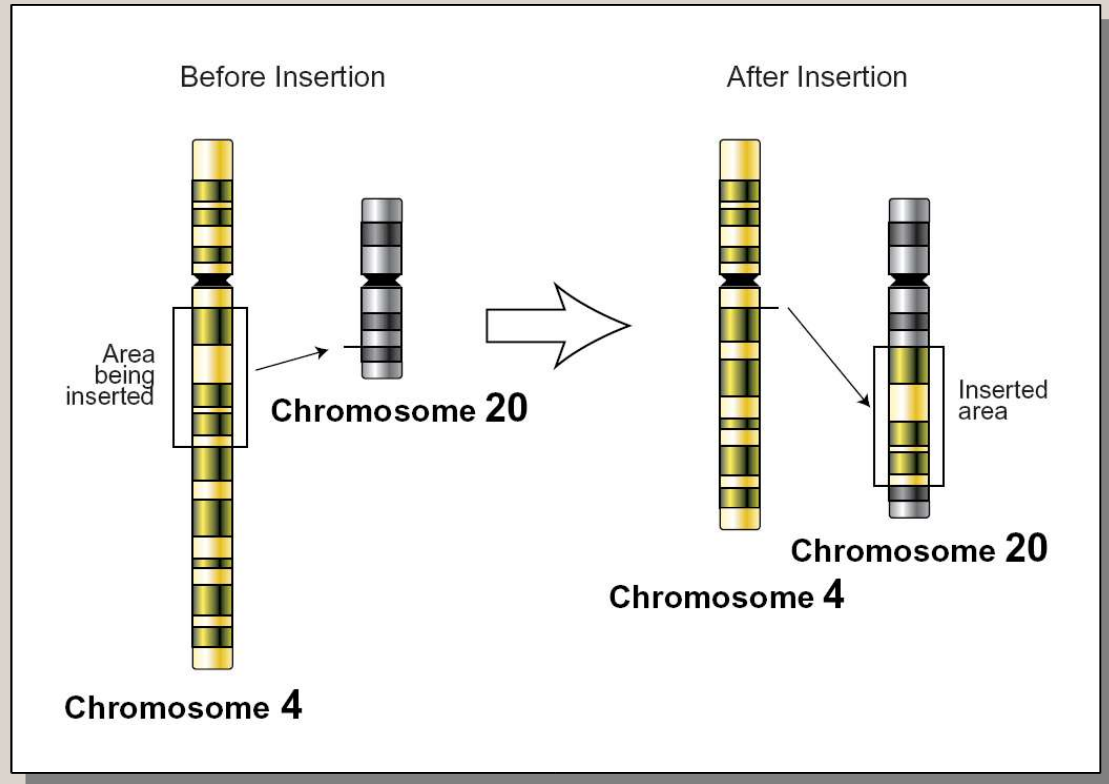
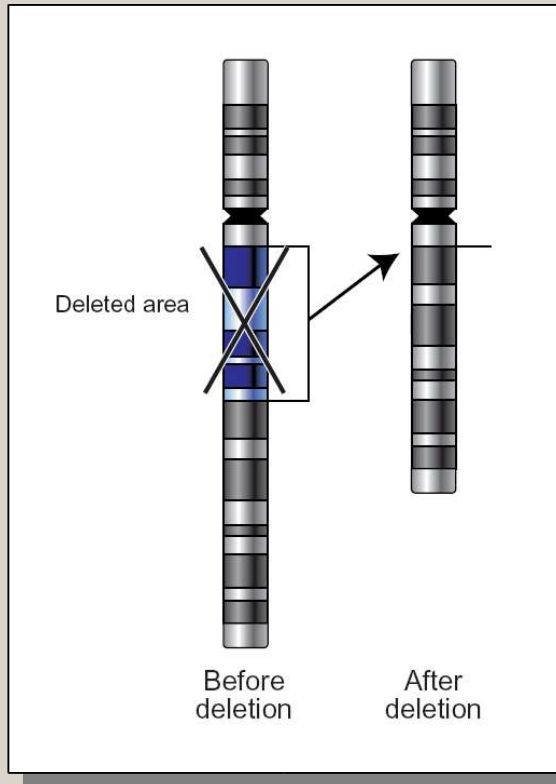
What's the problem? Google for biologists ...

- 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



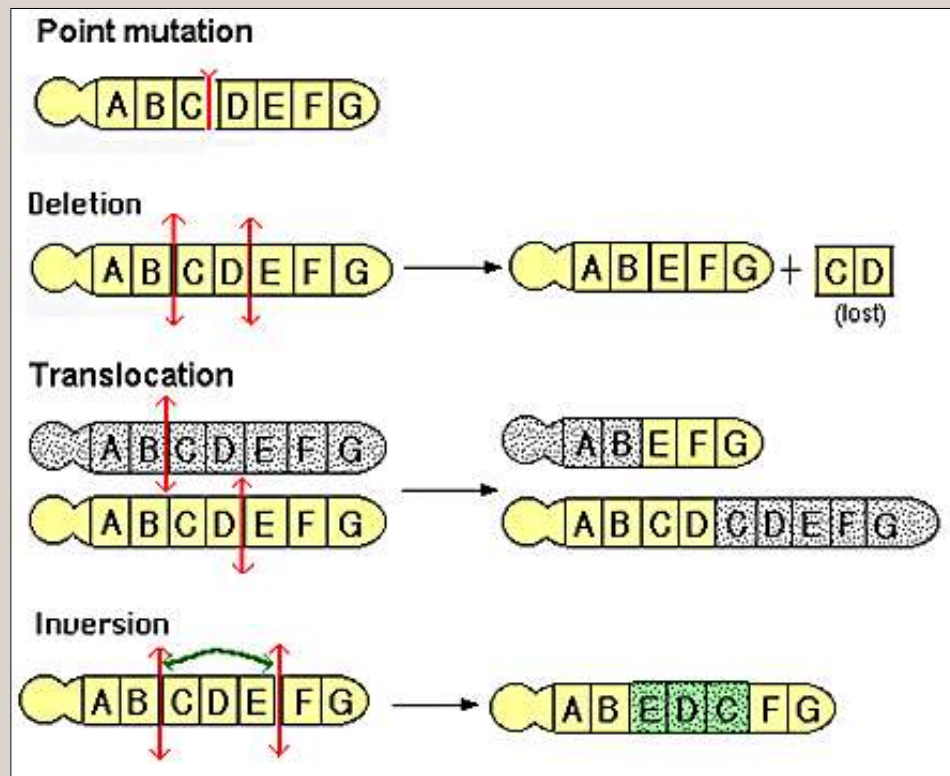
Sequence comparison must account for such effects.

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>

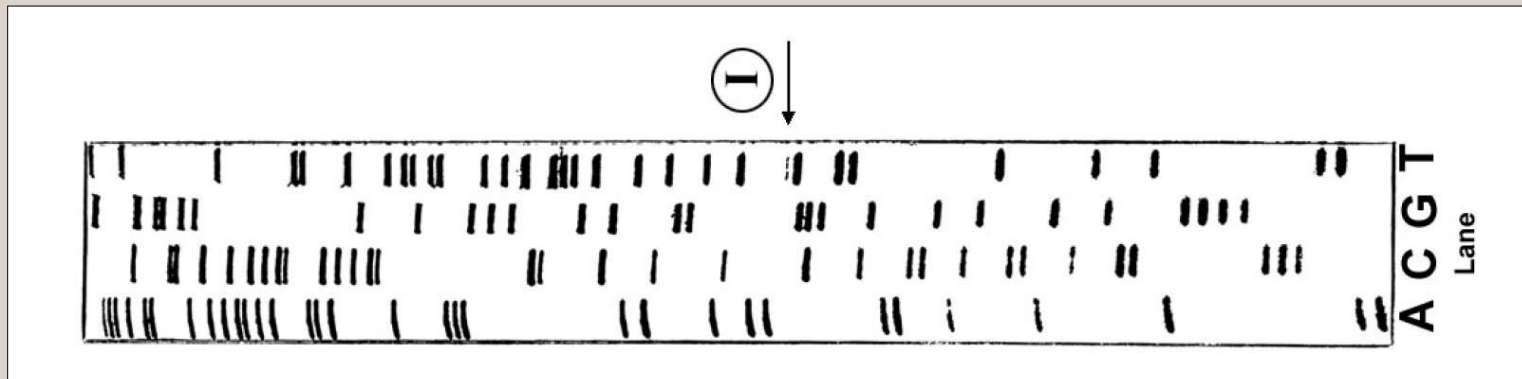
The Human Factor

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>

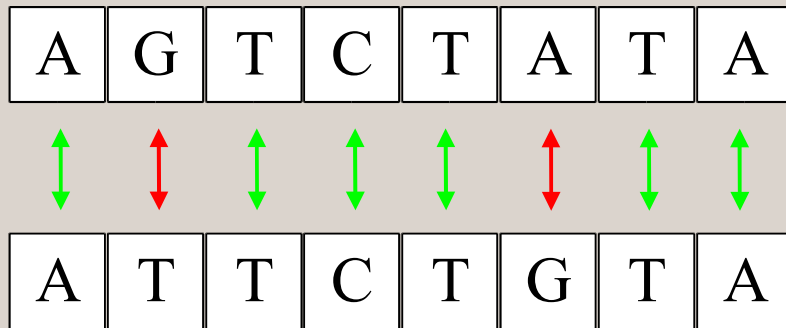
A hard-to-read gel (arrow marks location where bands of similar intensity appear in two different lanes):



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

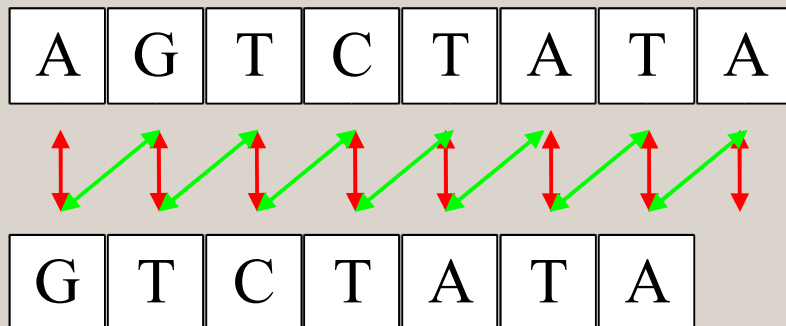
Sequence Comparison

Why not just line up sequences and count matches?



→ *Difference = 2*

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



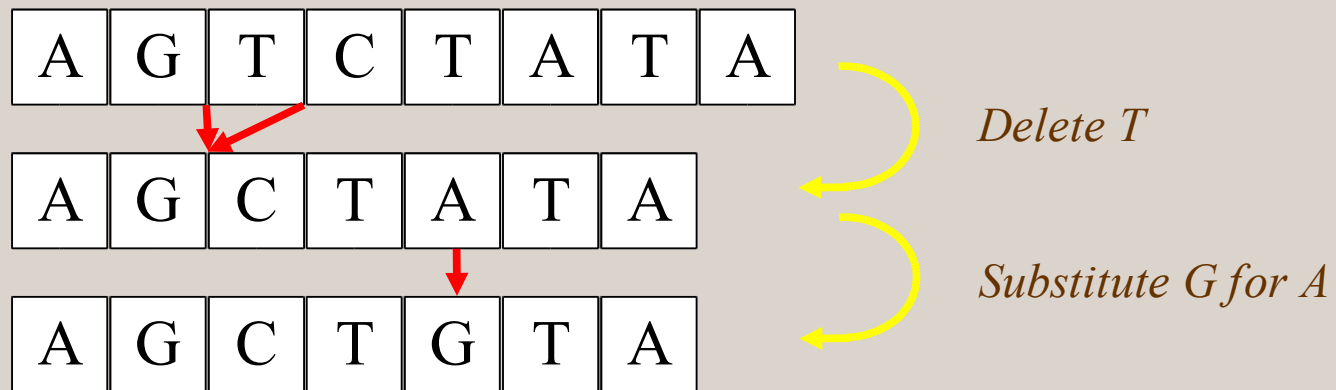
→ *Difference = 8*

One missing symbol at start of sequence leads to large distance.

Sequence Comparison

Instead, we'll use a basic 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.



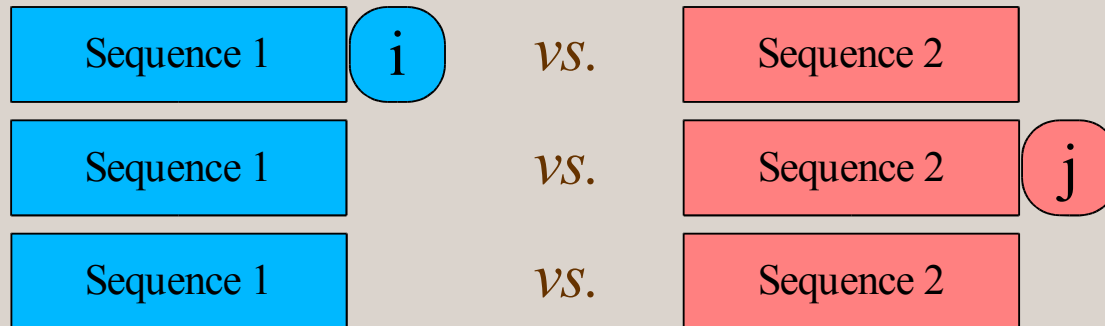
Sequence Comparison

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:



Note that we already know the best way to compare:

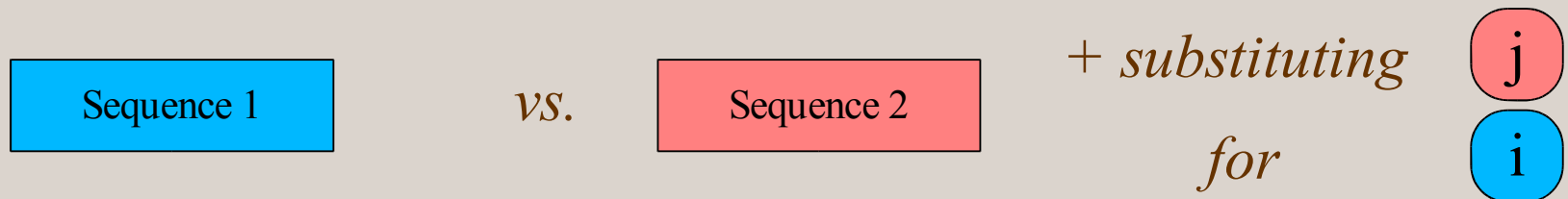
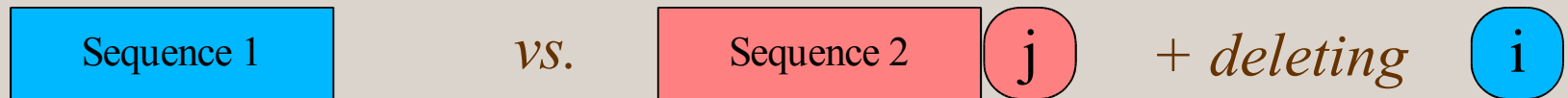
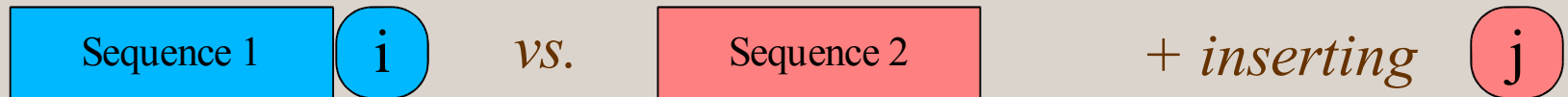


Sequence Comparison

So, best way to do this comparison:



Is best choice from following three cases:



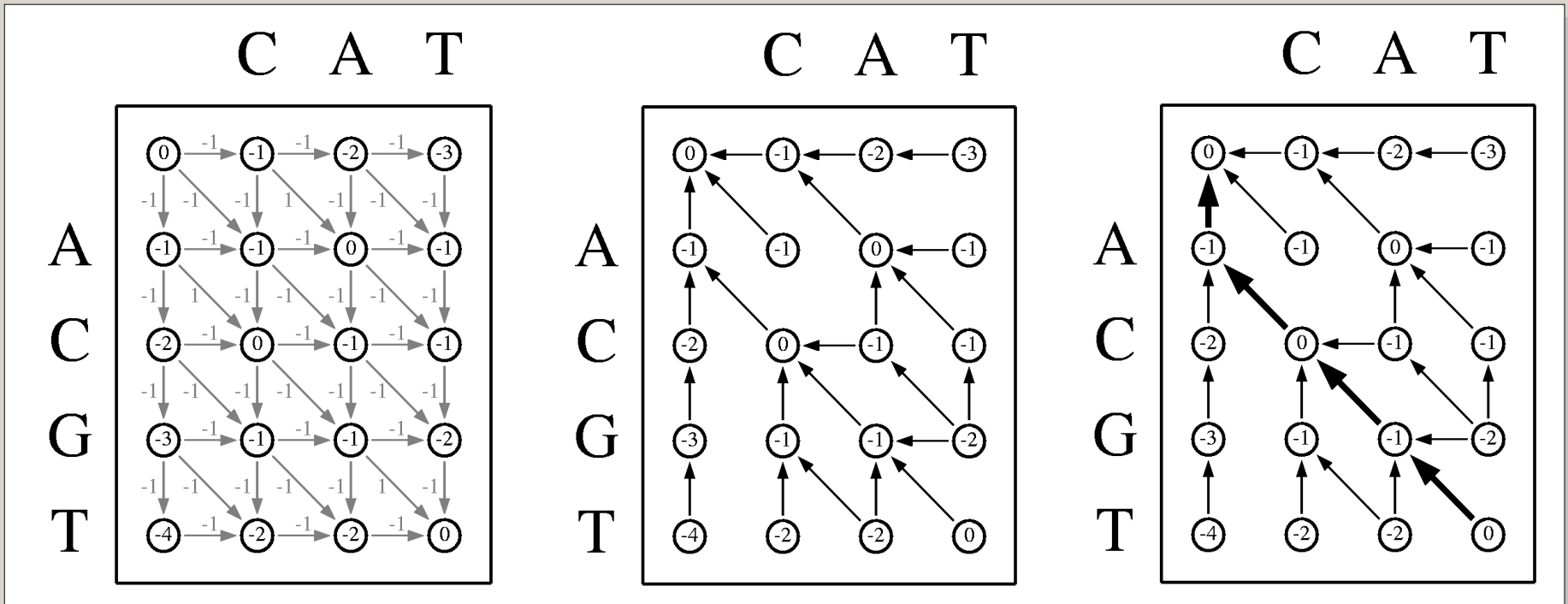
Sequence Comparison

Normally, this computation builds a table of distance values:

		ϵ	<i>sequence t</i>	
<i>sequence s</i>	ϵ	0	← <i>cost of inserting t</i>	
	↑ <i>cost of deleting s</i>		$d[i,j] = \min \begin{cases} d[i-1,j] + 1 \\ d[i,j-1] + 1 \\ d[i-1,j-1] + \begin{cases} 0 & \text{if } s[i] = t[j] \\ 1 & \text{if } s[i] \neq t[j] \end{cases} \end{cases}$	

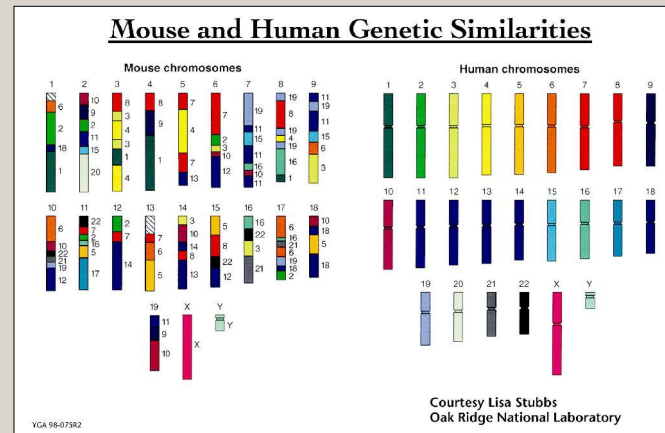
Sequence Comparison

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



Genome Rearrangements

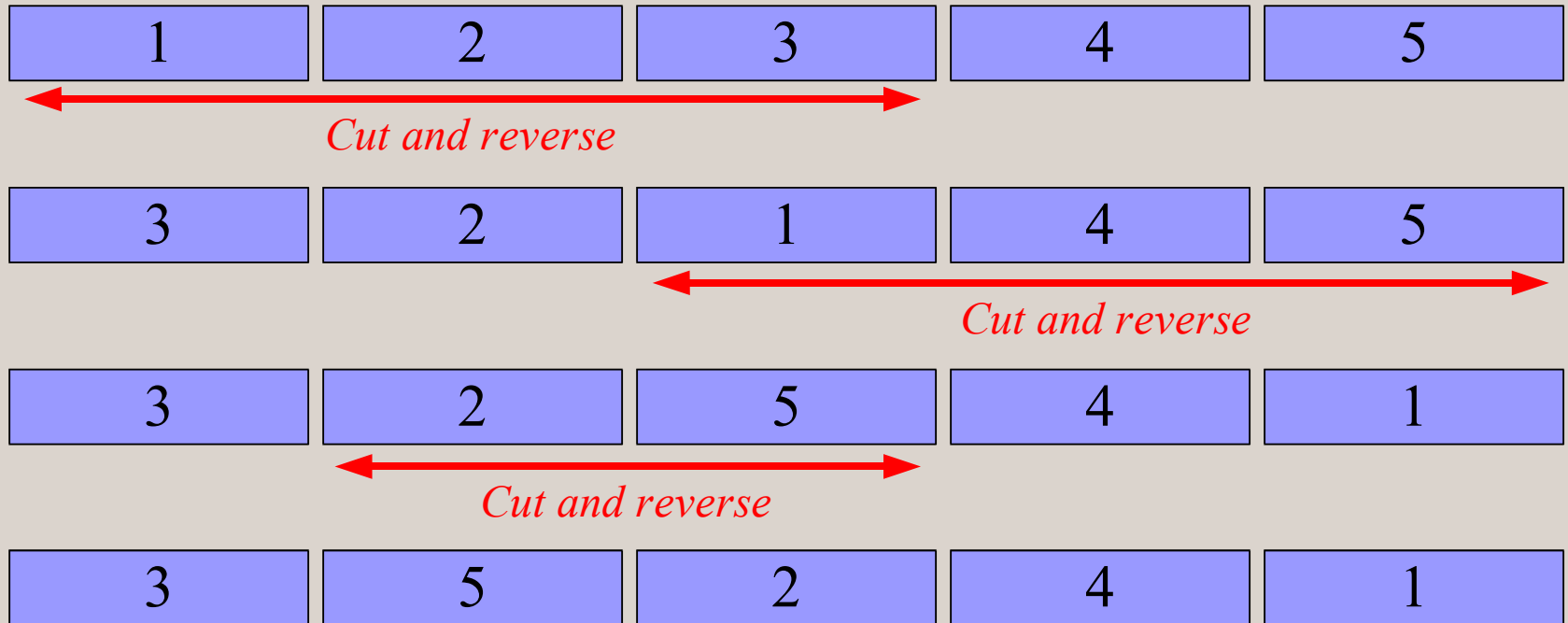
Recall what we saw earlier:



- 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

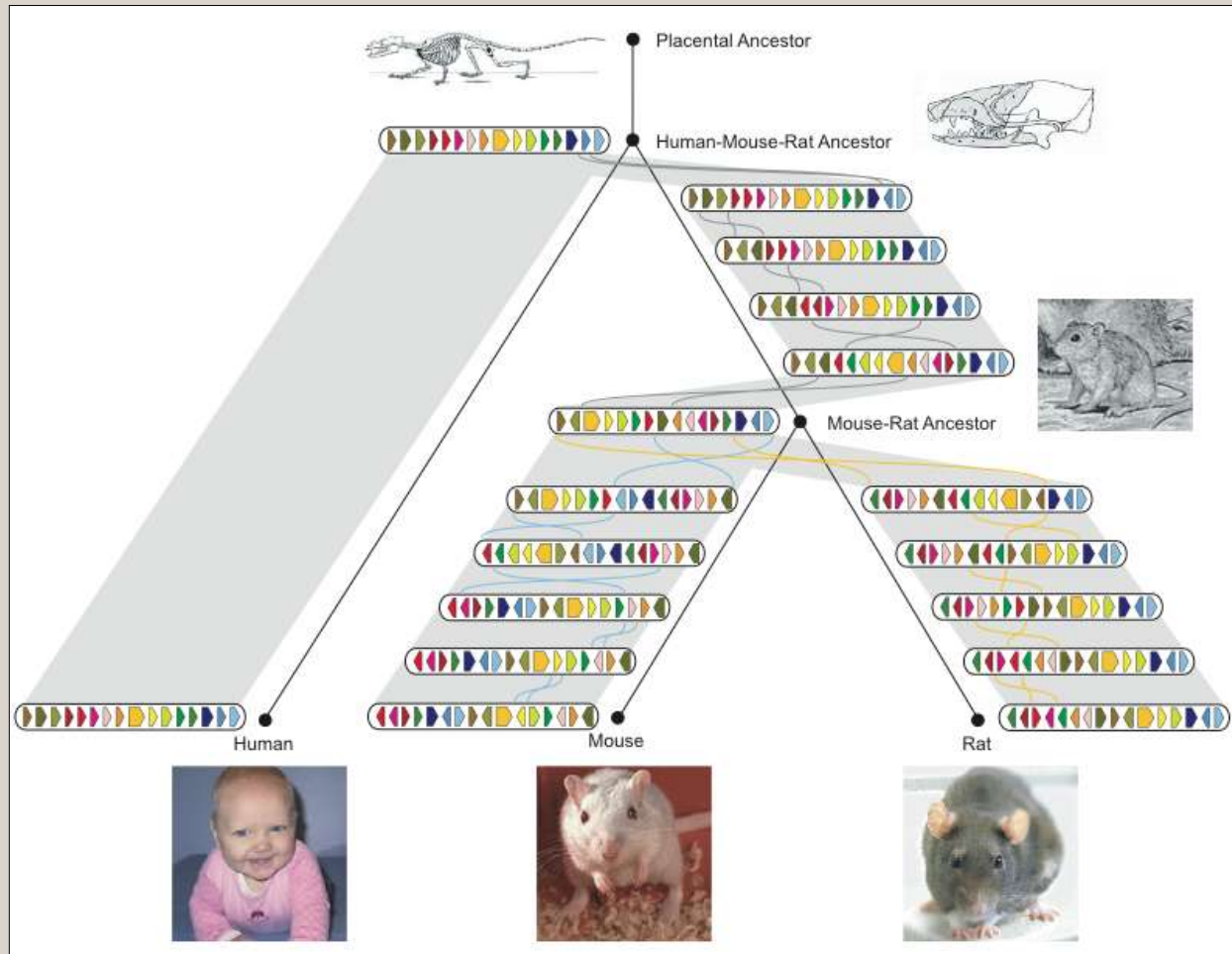
Human Chromosome X



Mouse Chromosome X

Reversal distance is the minimum number of such steps needed.

History of Chromosome X



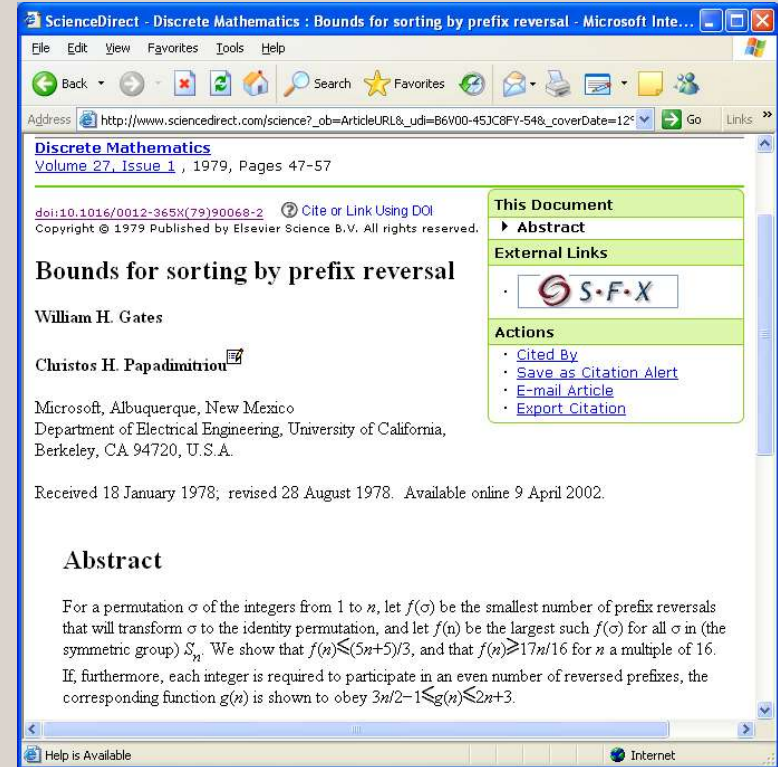
Rat Consortium, Nature, 2004

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 ...



ScienceDirect - Discrete Mathematics : Bounds for sorting by prefix reversal - Microsoft Inte...

File Edit View Favorites Tools Help


Address http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V00-45JC8FY-54&_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.

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.

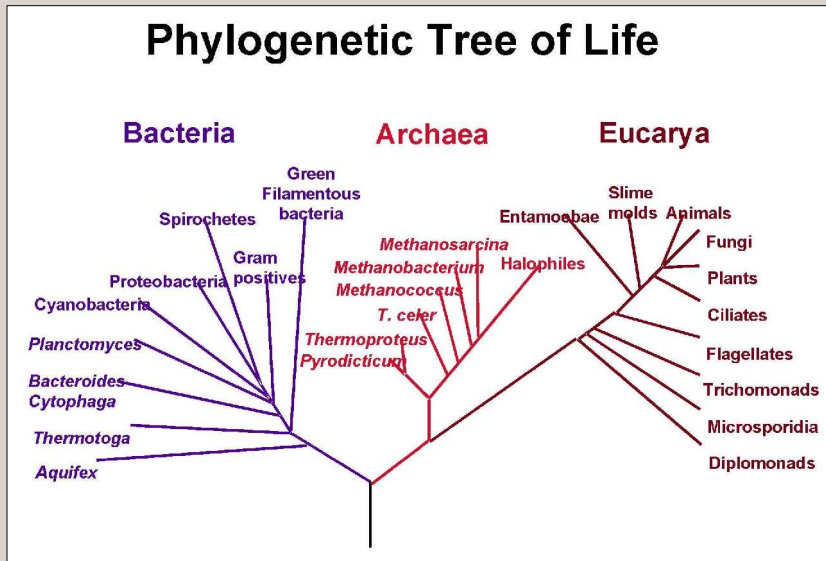
Abstract

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 $f(n) \leq (5n+5)/3$, and that $f(n) \geq 17n/16$ 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 $3n/2 - 1 \leq g(n) \leq 2n + 3$.

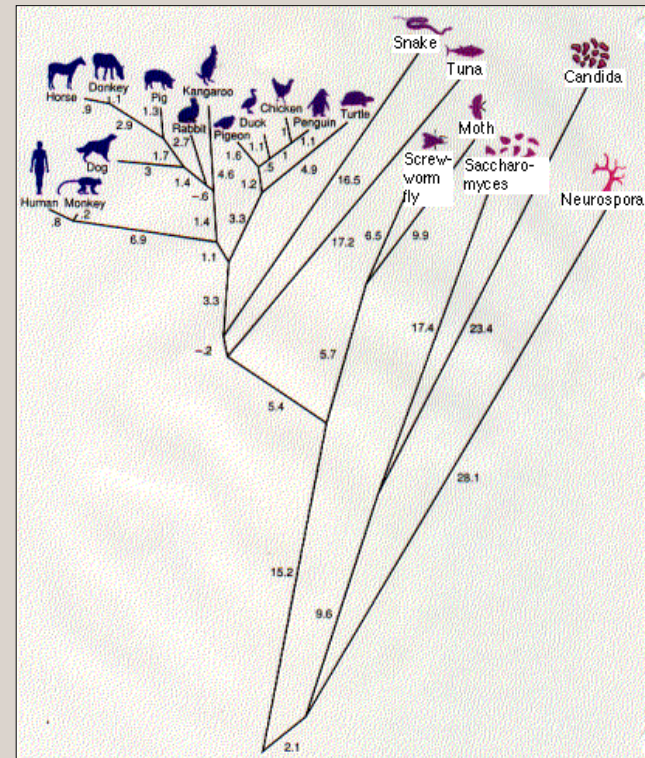
Help is Available Internet

Building the “Tree of Life”

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



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

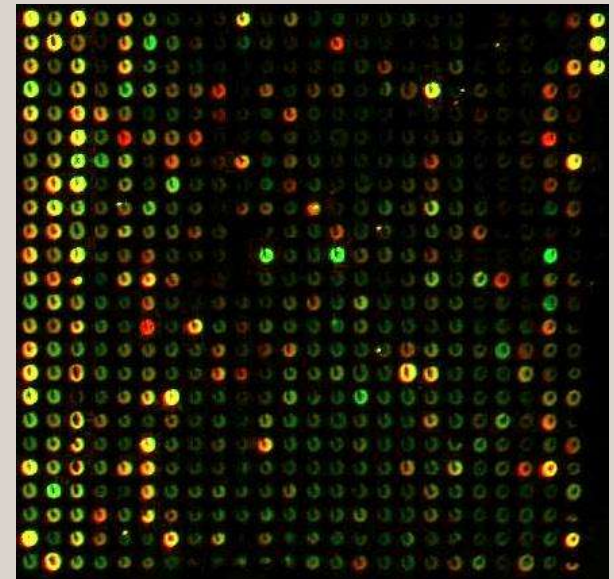


http://en.wikipedia.org/wiki/Phylogenetic_tree
<http://users.rcn.com/jkimball.ma.ultranet/Biology/Pages/T/Taxonomy.html>

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



DNA Microarrays

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

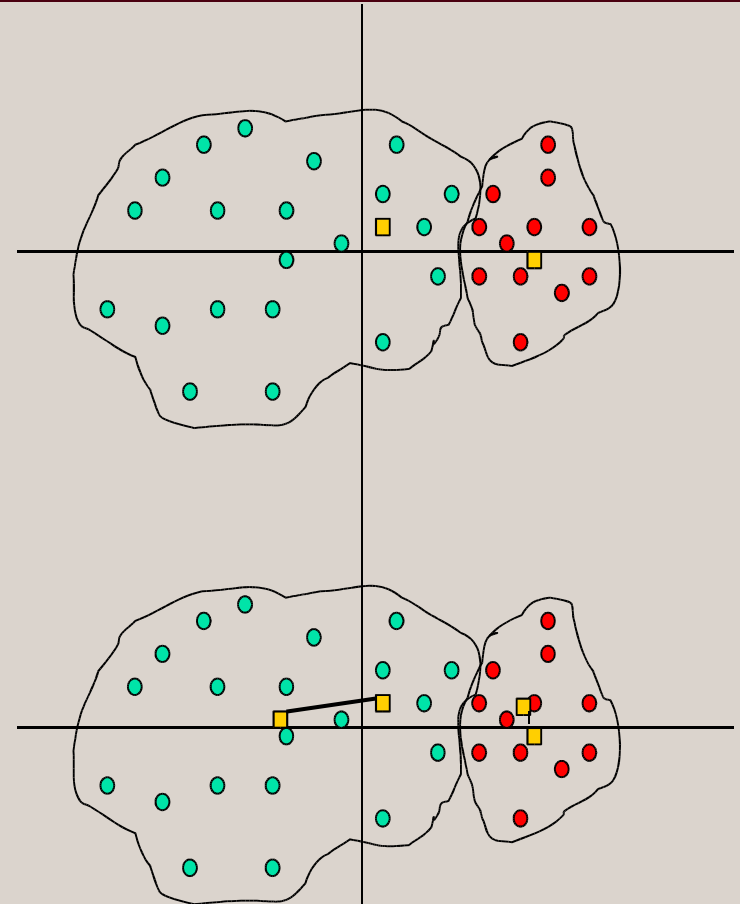
- Given set of n data points and an integer k .
- We want to find set of k points that minimizes the mean-squared distance from each data point to its nearest cluster center.

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

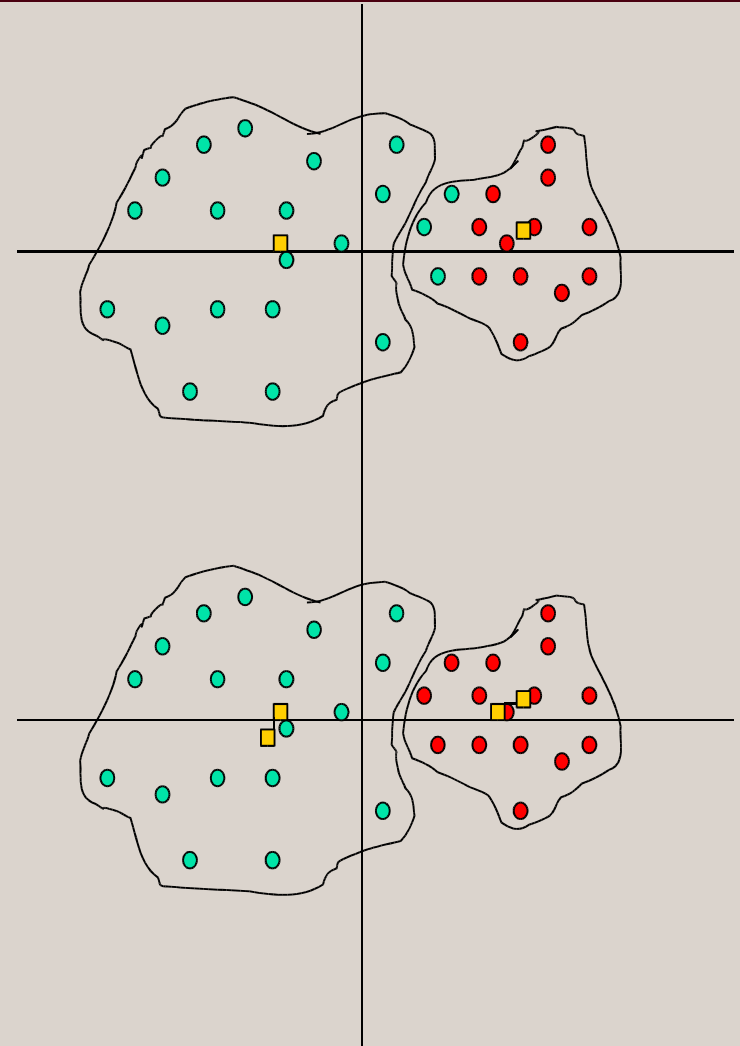
- Pick $k = 2$ centers at random.
- Cluster data around these center points.
- Re-calculate centers based on current clusters.



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

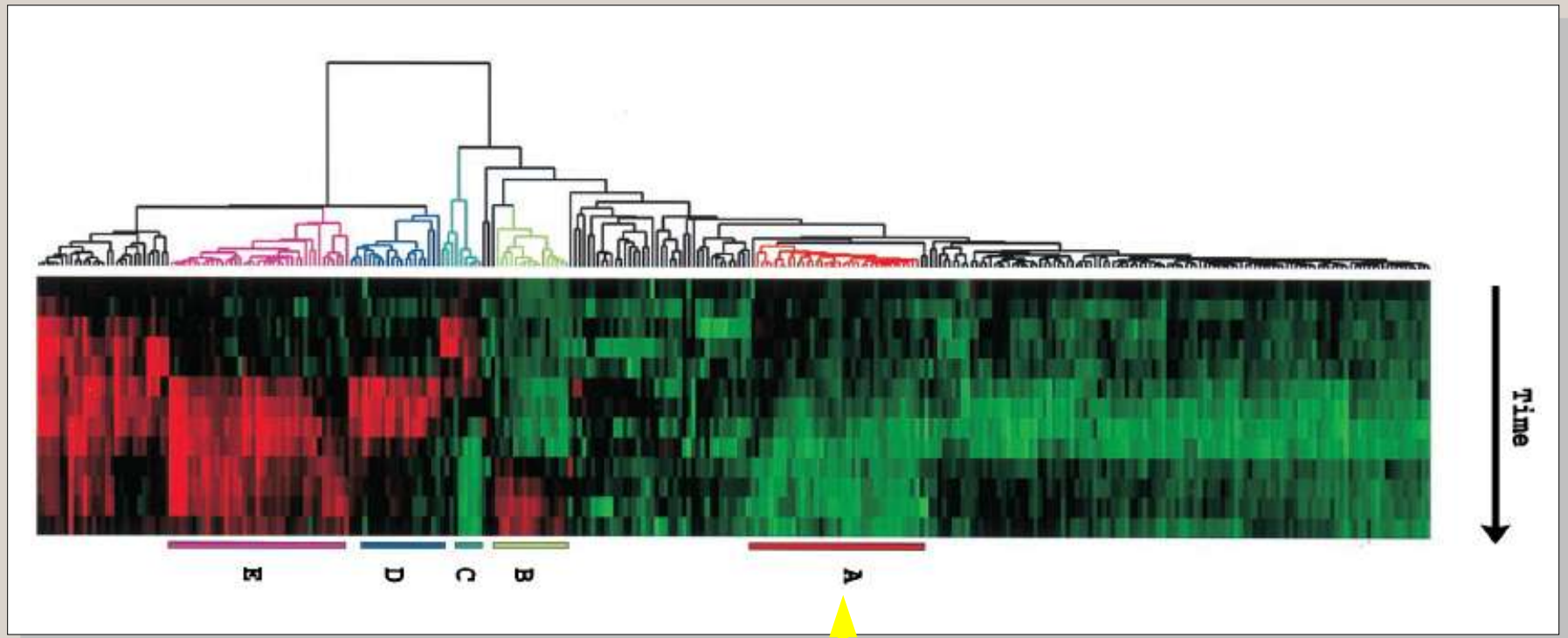
Clustering Microarray Data

- 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 308/408, we study algorithms for:

- Sequence comparison & alignment (pairwise & multiple).
- Sequence assembly (shotgun sequencing).
- Physical mapping of DNA.
- Constructing phylogenetic (evolutionary) trees.
- Computing genome rearrangements.
- DNA microarray analysis.
- DNA computing.

Materials @ <http://www.cse.lehigh.edu/~lopresti/courses.html>

Questions: dal9@lehigh.edu

Thank you!

Backup Slides

Bioinformatics and Computer Science

Recall that bioinformatics is the application of techniques from computer science to problems from biology.

Particularly relevant subfields:

- pattern recognition (classifying unknown inputs),
- image processing (machine vision),
- databases (efficient storage and retrieval),
- data mining (extracting knowledge from vast datasets),
- graphics & visualization (assists in data analysis),
- robotics (automating DNA microarray experiments).

Sequence Comparison

Simple example:

		q	u	i	c	k
	0	1	2	3	4	5
h	1	1	2	3	4	5
a	2	2	2	3	4	5
c	3	3	3	3	3	4
k	4	4	4	4	4	3

So the distance between “hack” and “quick” is 3.

This corresponds to one insertion and two substitutions.

If one sequence has length M and other has length N , then table has $M*N$ entries. We say the time complexity is $O(MN)$.