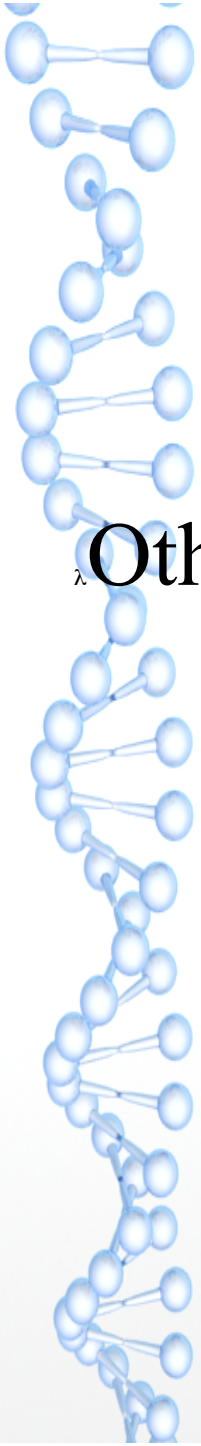
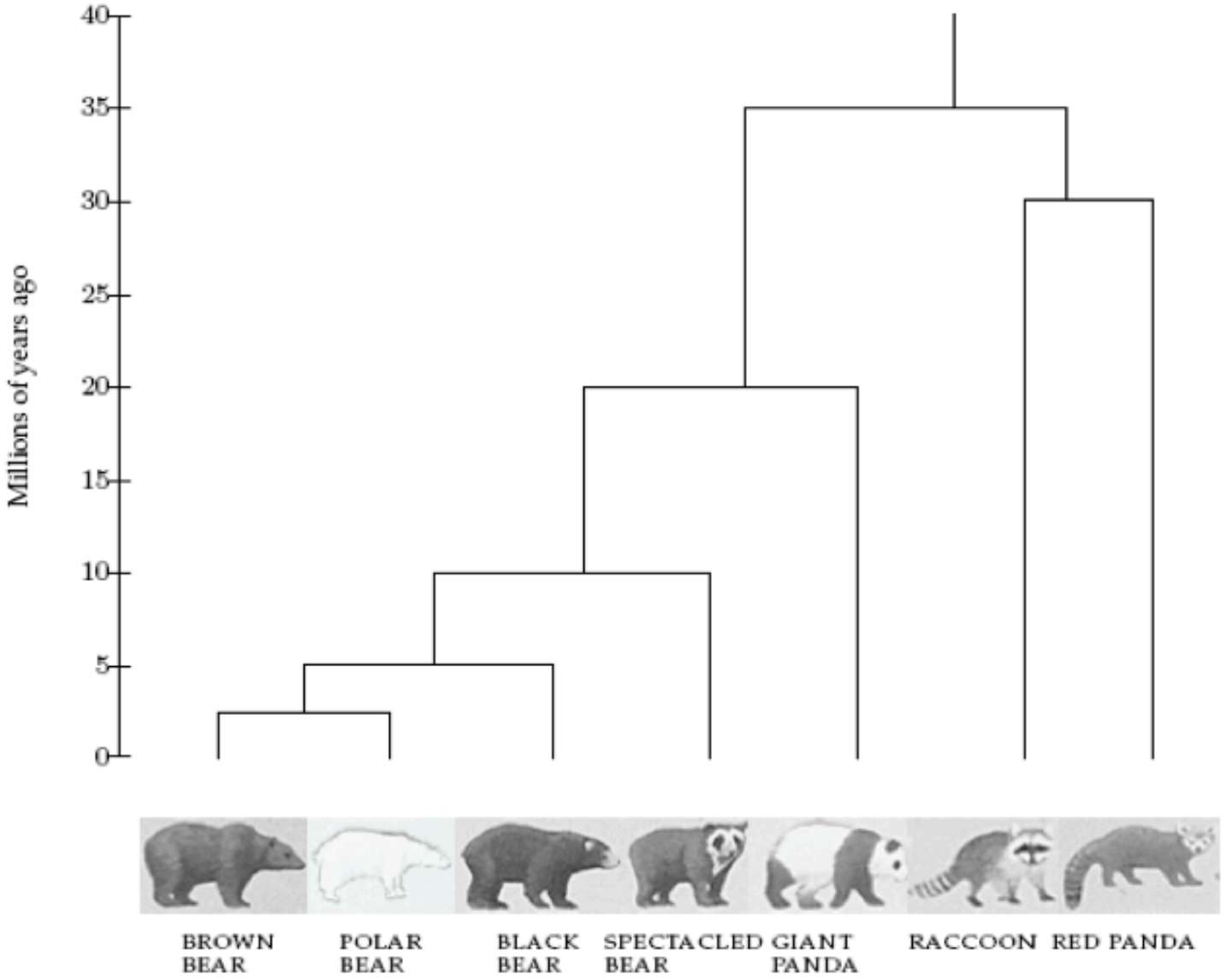


Other approaches to inferring phylogenetic trees


Millions of years ago
40
35
30
25
20
15
10
5
0
BROWN BEAR
POLAR BEAR
BLACK BEAR
SPECTACLED BEAR
GIANT PANDA
RACCOON
RED PANDA

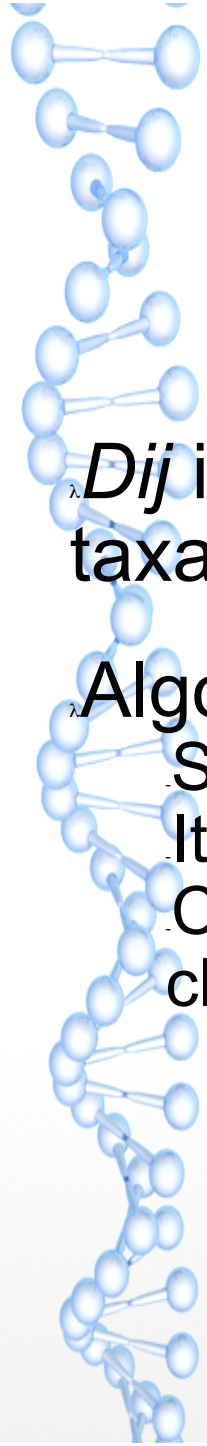
Review

- UPGMA
  - This method requires ultrametric property: for three distances, two are equal and third is <= the first two
  - Relies on a molecular clock
- Neighbor joining
  - This method requires additive property: distances between two nodes is sum of their edges
  - Often produces a good tree even with non-additive data

UPGMA Method

- Unweighted Pair Group Method using Arithmetic Averages
- Distance is defined between two clusters *Ci* and *Cj* such that:

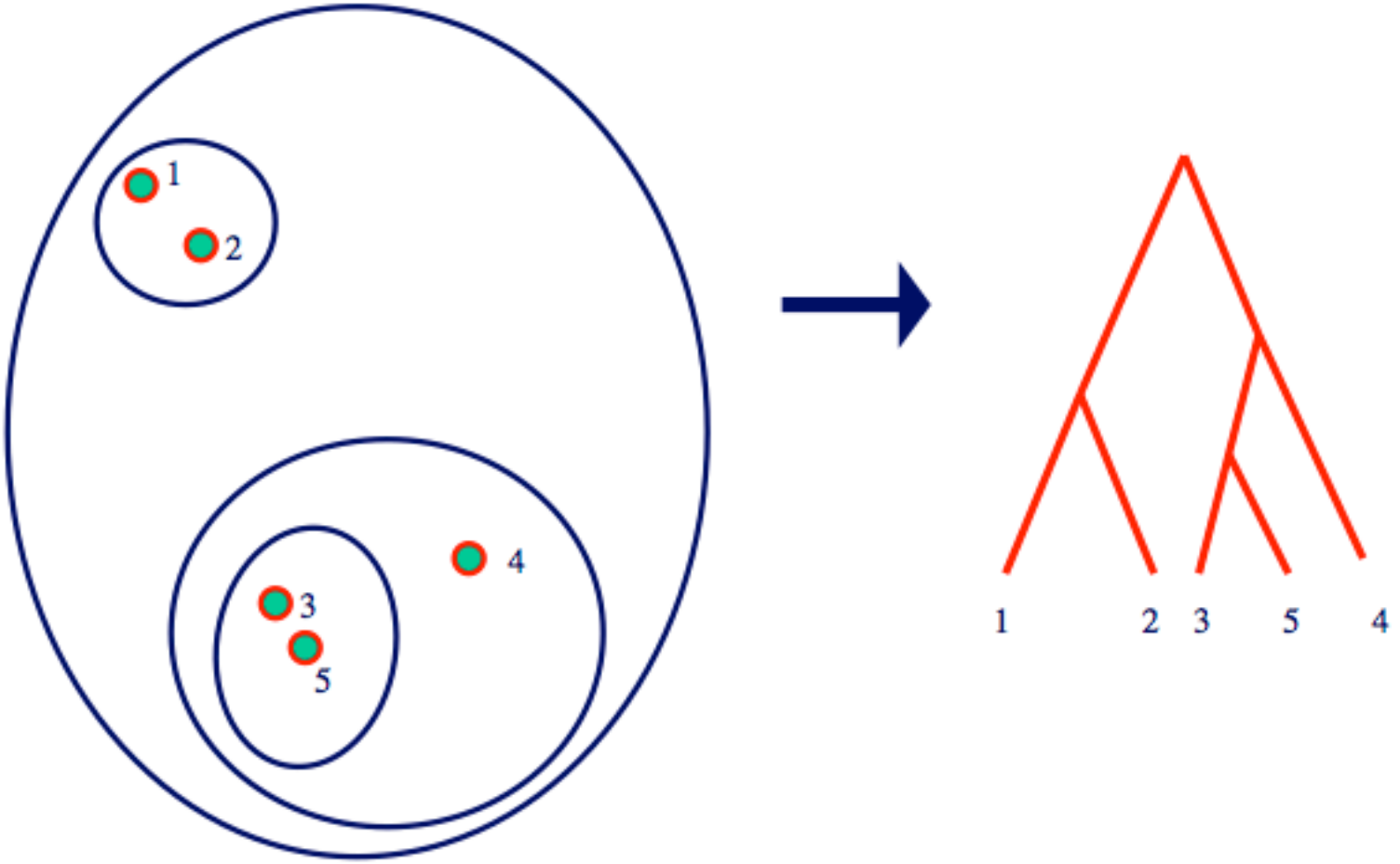
$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$$



Basic idea

- *Dij* is the average distance between pairs of taxa from each cluster
- Algorithm:
  - Start with one taxa per cluster
  - Iteratively pick two clusters and merge
  - Create a new node in the tree for the merged cluster

UPGMA visually

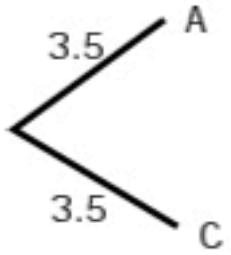


More specifics

- Place each taxon at height 0 in the tree
- While more than two clusters:
  - Determine clusters with smallest *dij*
  - Merge clusters into a new one *Ck*
  - Make a new node *k* at height *dij*/2
  - Replace *Ci* and *Cj* with *Ck*
  - Recompute distance of C*k* to other clusters
- Hook in the two remaining clusters to the root with height calculated as above.

Step 1: pick smallest and update distances

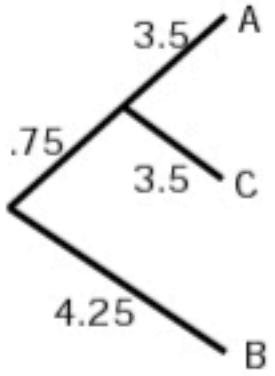
	A	B	C	D
A	0			
B	8	0		
C	7	9	0	
D	12	14	11	0



$$M_{B(AC)} = (M_{BA} + M_{BC})/2 = (8+9)/2=8.5$$
$$M_{D(AC)} = (M_{DA} + M_{DC})/2= (12+11)/2=11.5$$

Step 2

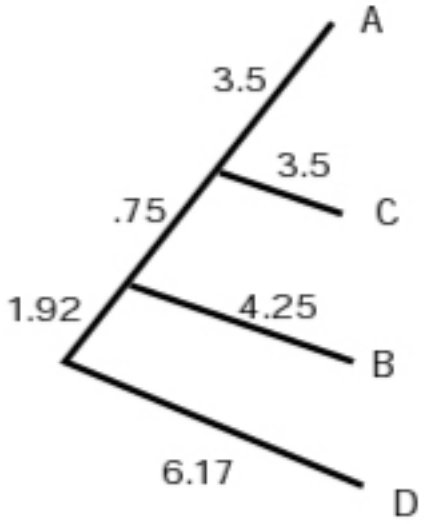
	AC	B	D
AC	0		
B	8.5	0	
D	11.5	14	0



$$M_{(ABC)D} = (M_{AD} + M_{BD} + M_{CD})/3 = (12+14+11)/3$$

Step 3

	ABC	D
ABC	0	
D	12.33	0



Molecular clocks

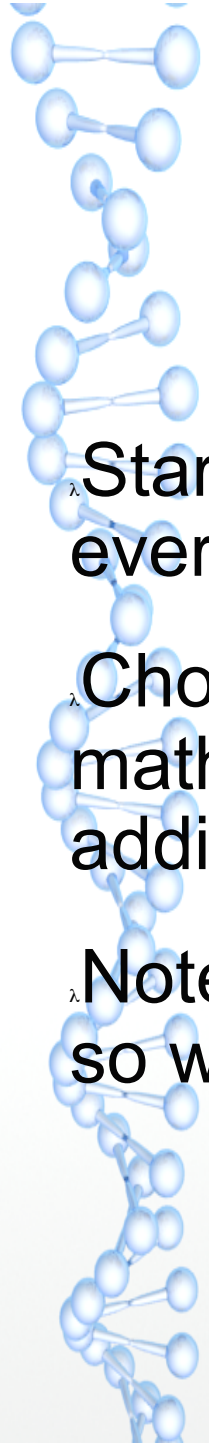
- A molecular clock assumption is divergence is uniform and equal across all branches of the tree
- Seldom (never?) true in practice
- If it is true, these data are called *ultrametric*.

Neighbor joining

- Does not assume a molecular clock, but does assume additively
  - Distance between a pair of leaves is sum of edges between them
- Constructs an unrooted tree iteratively, just like UPGMA
- Two differences:
  - How subtrees selected
  - How distances are updated
- Root can be added via inclusion of an “outgroup”

Basics

- NJ is a greedy algorithm that starts with a center star tree (all taxa connected to a single root)
- Criterion for merging is key: identifies topological neighbors using math that is correct for all additive distance matrices.
- Once merged, the two taxa are treated as a single taxon.



Idea behind NJ algorithm

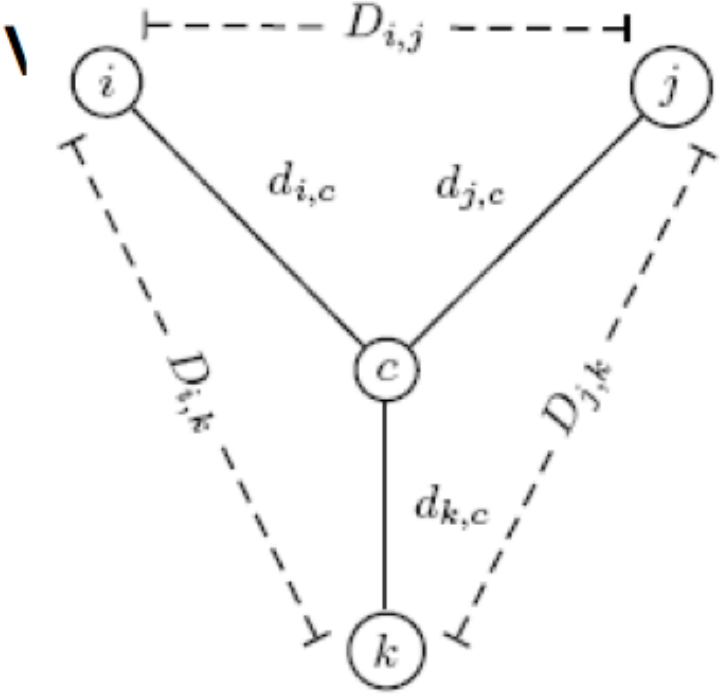
- Start with a center star phylogeny where everyone is connected to a single node.
- Choose two sequences to merge based on the mathematically optimal topology under an additive scoring scheme.
- Note this tree starts out unrooted and remains so without an outgroup.

Caveats

- Matrix is updated iteratively after merge.
- Produces unrooted trees; need an outgroup for a rooted version
- Always gives the true tree if distances are additive (may not with noise)

Reconstructing a 3 leaved tree

- Tree reconstruction for any 3x3 matrix is straightforward
- We have 3 leaves *i*, *j*, *k* and a center

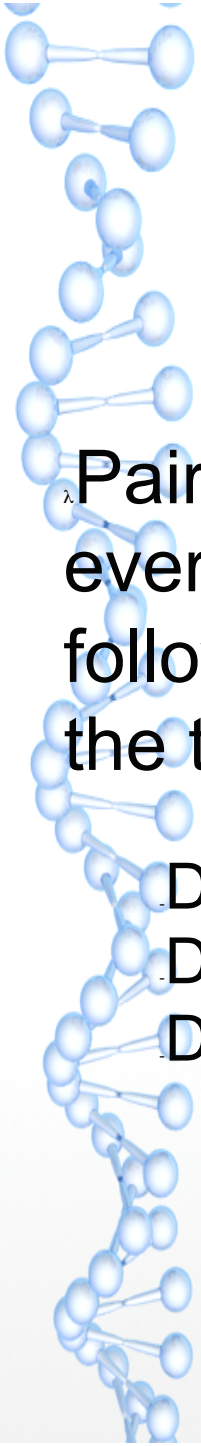


Observe:

$$d_{ic} + d_{jc} = D_{ij}$$

$$d_{ic} + d_{kc} = D_{ik}$$

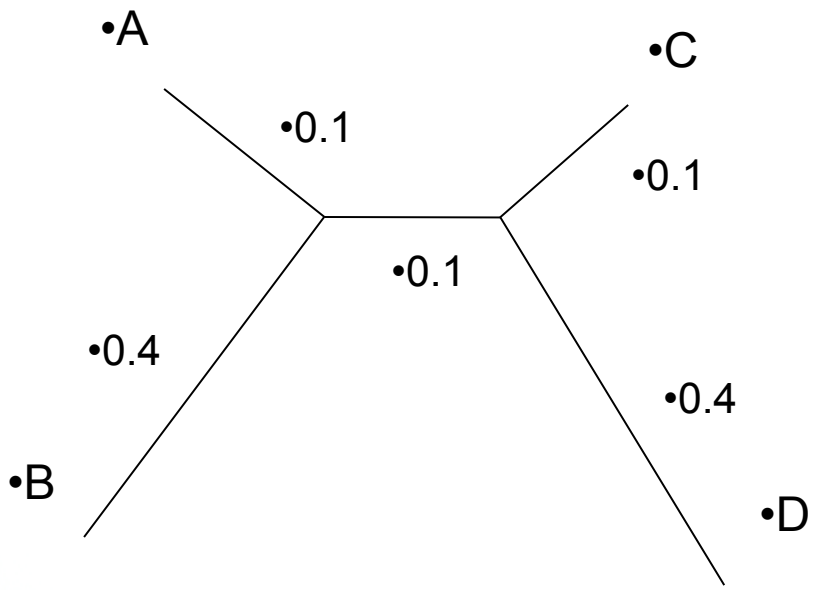
$$d_{jc} + d_{kc} = D_{jk}$$



Four-point condition

- Pairwise distances are additive if and only if for every set of four leaves *i,j,k,l*, two of the following three sums are equal and larger than the third:
  - Dij + Dkl
  - Dik + Djl
  - Dil + Djk

Example



- Neighbor-joining will find the correct tree here

One more thing

- There are an assortment of formats for trees as there is with DNA sequence data.

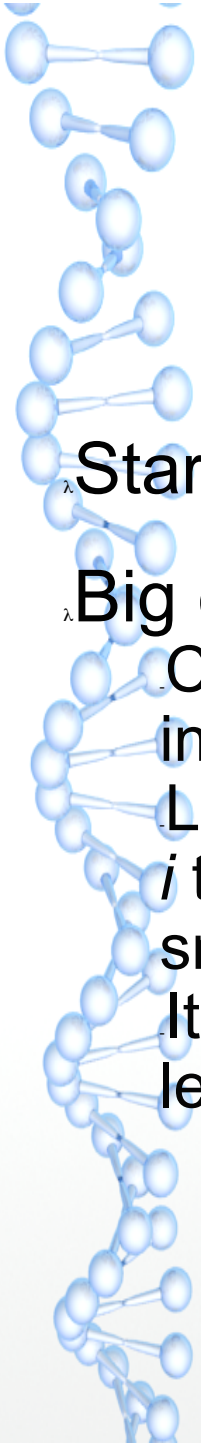
- Newick format indicated in the text is one of the more common ones (like FASTA is for sequences)
  - Ex: ((A,B),(C,D))

Calculating distances

- Saitou and Nei (1987)

$$D_{ij} = d_{ij} - (r_i + r_j)$$

$$r_i = \frac{1}{|L| - 2} \sum_{k \in L} d_{ik}$$



NJ algorithm

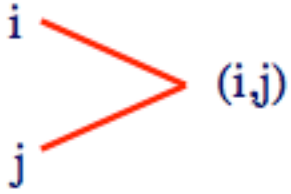
- Start with a center star tree and break off pairs
- Big caveat:
  - Compute (almost) “average” distance to other nodes including those in the tree
  - Look for nodes that its distance (Mij) – distance from *i* to everything else – distance *j* to everything else is smallest
  - It turns out the above equates to minimizing branch lengths in the complete tree

Why compute more distances?

- The most crucial piece of NJ is computing a new matrix using the previously mentioned equations.
- This allows us to choose the smallest one greedily, in something called the “4-point condition.”
- UPGMA is a simpler form of NJ that is correct when distance between all taxa and the root is the same.
  - This seems true but rarely holds up in observed DNA sequence data

Main idea

Step 3: Define a new cluster (i, j), with a corresponding node in the tree



Distance from i and j to node (i,j):

$$d_{i,(i,j)} = 0.5(M_{ij} + u_i - u_j)$$

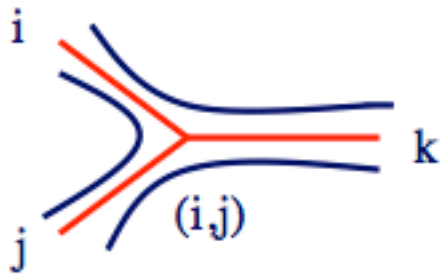
$$d_{j,(i,j)} = 0.5(M_{ij} + u_j - u_i)$$

Default: split distance but if on average one is further away, make it longer

Finishing up

Step 4: Compute distance between new cluster and all other clusters:

$$M_{(ij)k} = \frac{M_{ik} + M_{jk} - M_{ij}}{2}$$



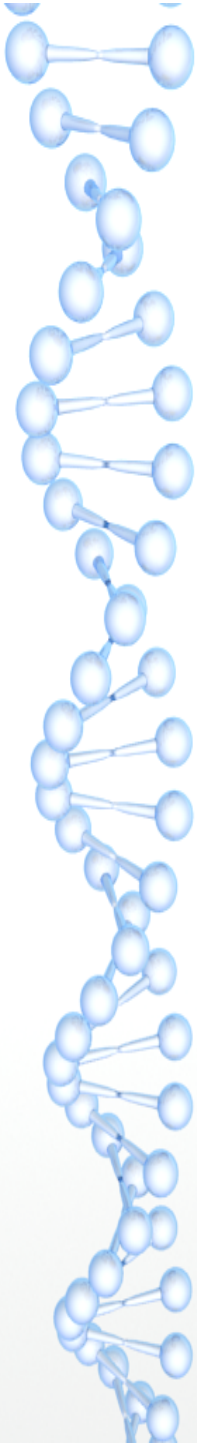
Step 5: Delete *i* and *j* from matrix and replace by *(i, j)*

Step 6: Continue until only 2 leaves remain

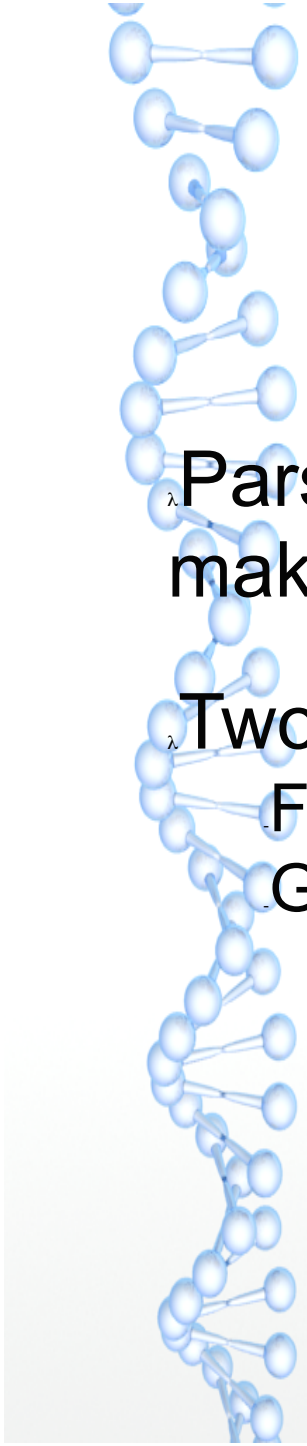
Finishing up NJ

- Build a NJ tree for the matrix earlier:

	•A	•B	•C	•D	•E
•A	•0	•5	•3	•8	•10
•B	•5	•0	•5	•8	•10
•C	•3	•5	•0	•8	•10
•D	•8	•8	•8	•0	•1
•E	•10	•10	•10	•1	•0

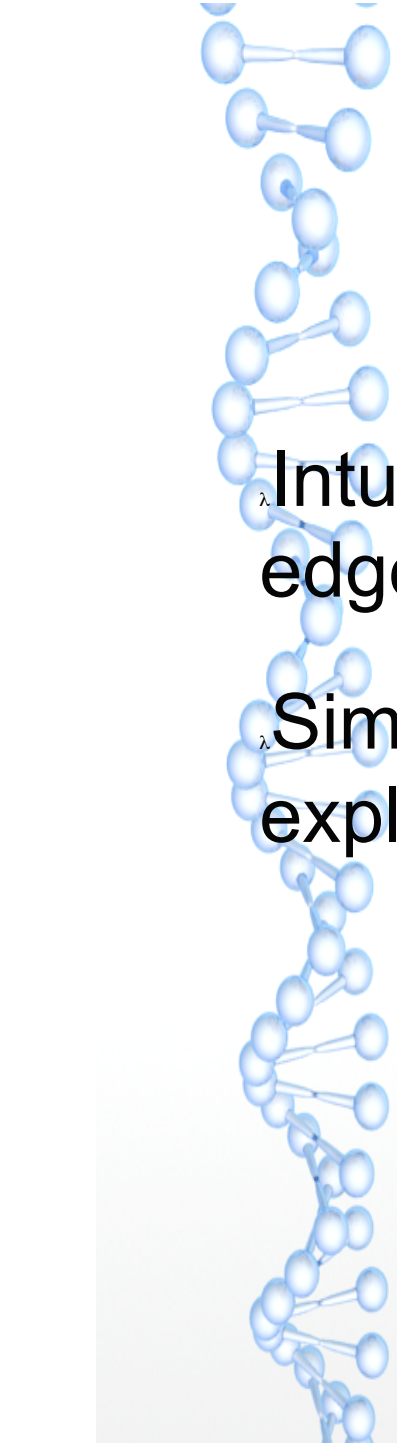


λ Parsimony



Intro

- Parsimony is a simple and fast approach, making it popular
- Two distinct subproblems:
  - Find the history of mutations (easy)
  - Given an alignment, infer tree (hard)



Parsimony

- Intuitively, we want to measure changes along edges of trees.
- Similar to Okham's razor: find simplest explanation that works

par·si·mo·ny

/ˈpärsəˌmōnē/

noun

1. extreme unwillingness to spend money or use resources.
   "a great tradition of public design has been shattered by government parsimony"
   synonyms: cheapness, miserliness, meanness, parsimoniousness, niggardliness, close-fistedness, closeness, penny-pinching; More

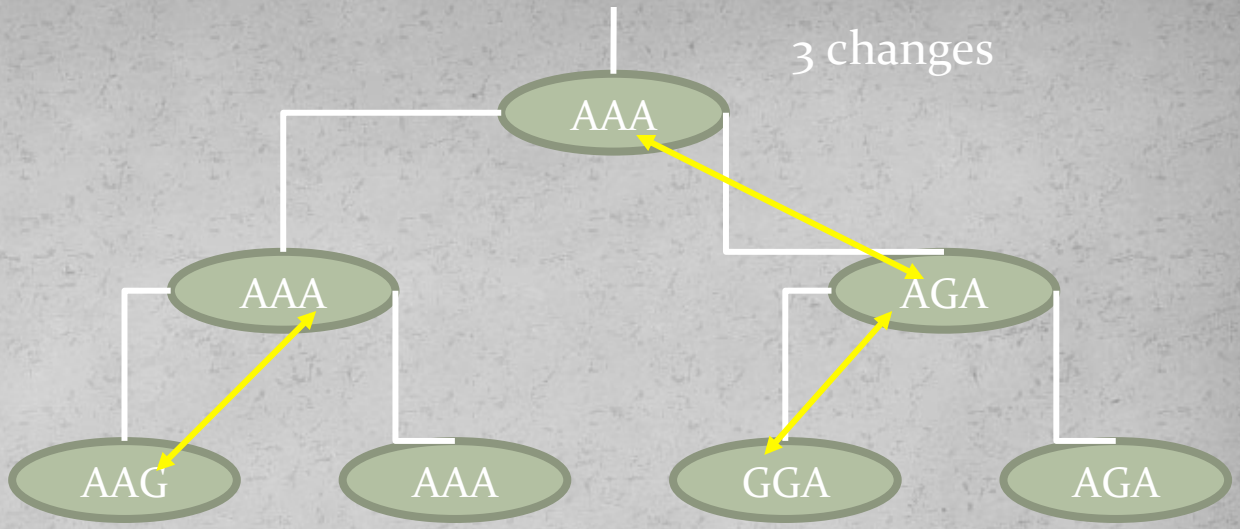
Example

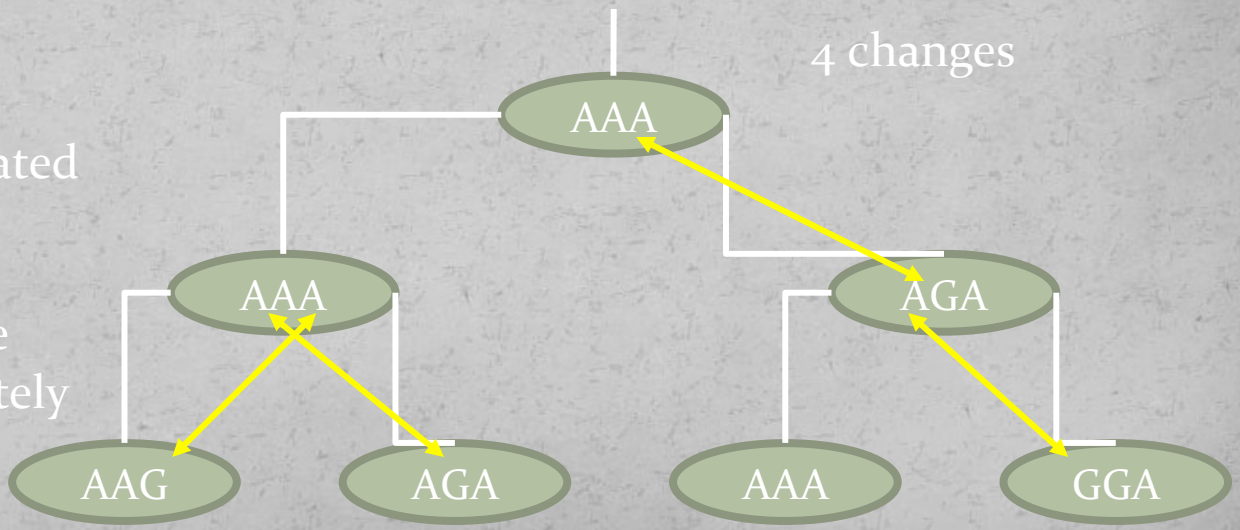
AAG

AAA

GGA

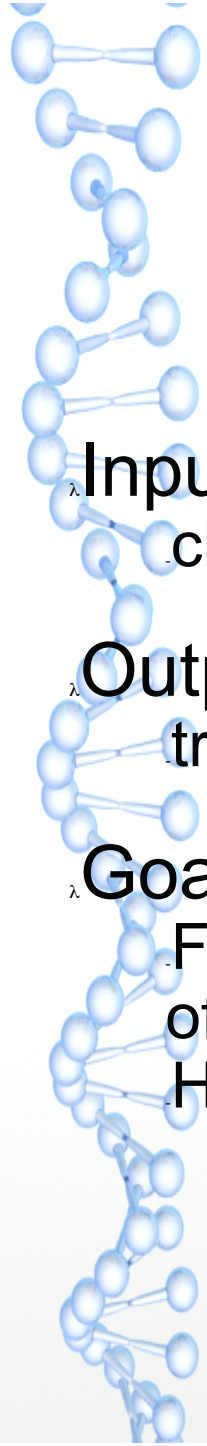
AGA





Notice: Sites are treated independently.

We can estimate the cost for each separately and sum over them.

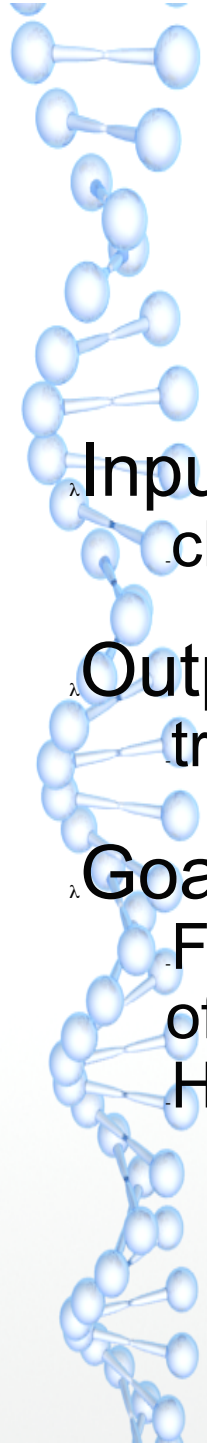


Overview

- Input:
  - character-based data such as an alignment
- Output:
  - tree that requires minimal number of changes
- Goal:
  - Find right tree topology (how things branch) instead of the actual lengths of edges
  - Hard part is finding optimal topology

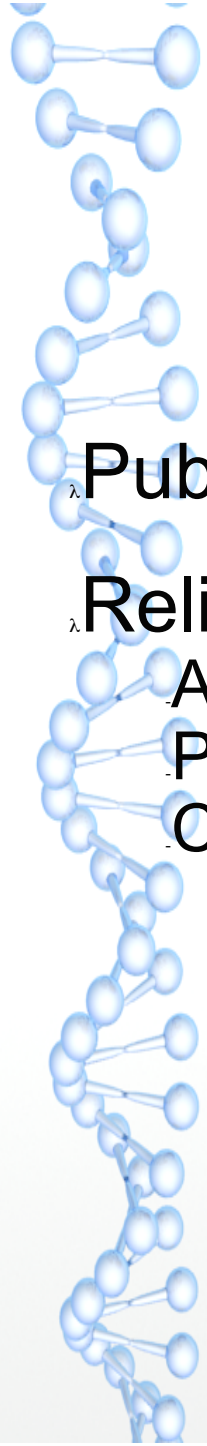
In-class example

	•1	•2	•3	•4	•5	•6
•Cat	•A	•T	•A	•C	•A	•G
•Dog	•A	•T	•A	•G	•T	•G
•Chimp	•A	•C	•A	•C	•A	•G
•Cow	•T	•C	•G	•G	•T	•A
•Bat	•T	•C	•G	•C	•A	•G



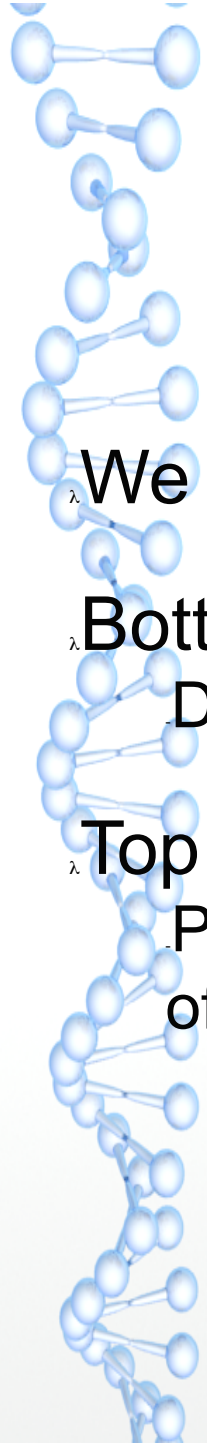
Overview

- Input:
  - character-based data such as an alignment
- Output:
  - tree that requires minimal number of changes
- Goal:
  - Find right tree topology (how things branch) instead of the actual lengths of edges
  - Hard part is finding optimal topology



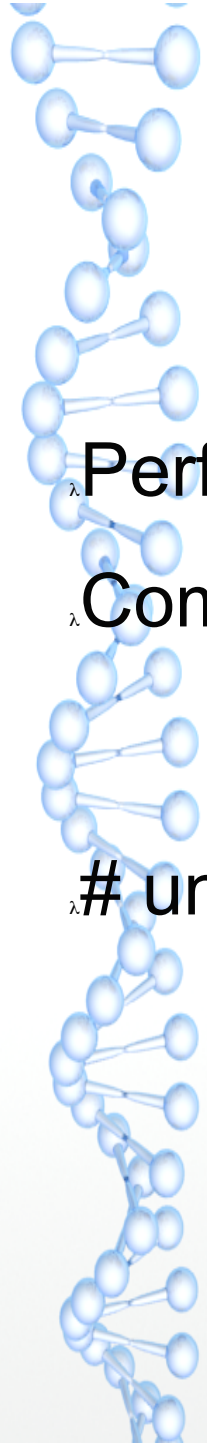
Fitch's algorithm

- Published in 1971
- Relies on the following assumptions:
  - Any state can convert to any other state
  - Positions in an input are independent
  - Cost is uniform, i.e., A -> T = G -> T



Fitch's algorithm

- We will do two traversals of the tree
- Bottom - up (leaves to root)
  - Determine set of possible states for each node
- Top - down (root to leaves)
  - Pick the ancestral state for each node from the set of possibilities

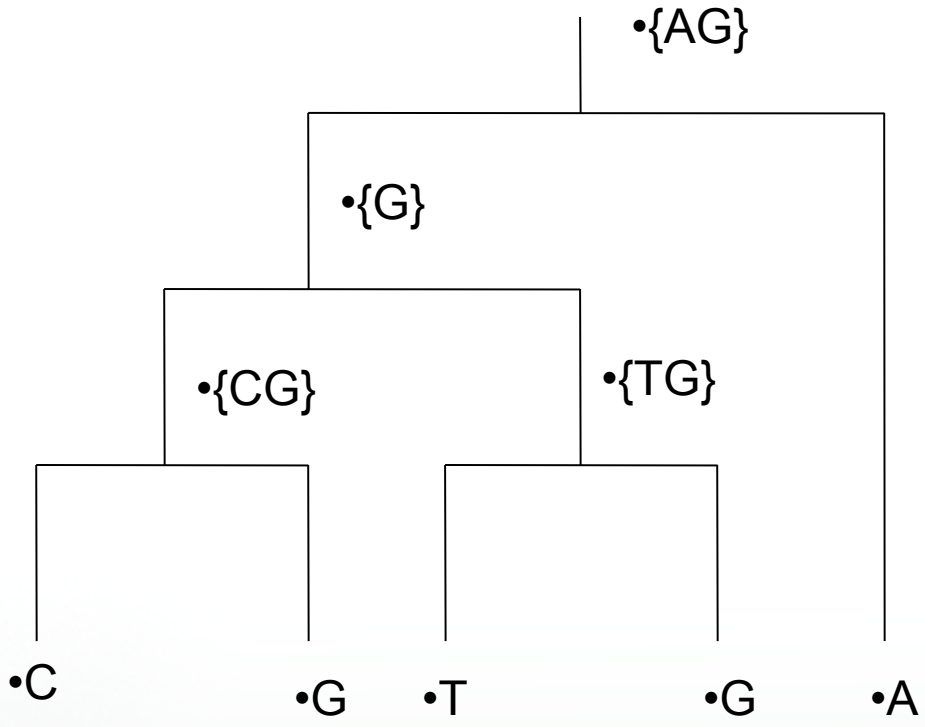


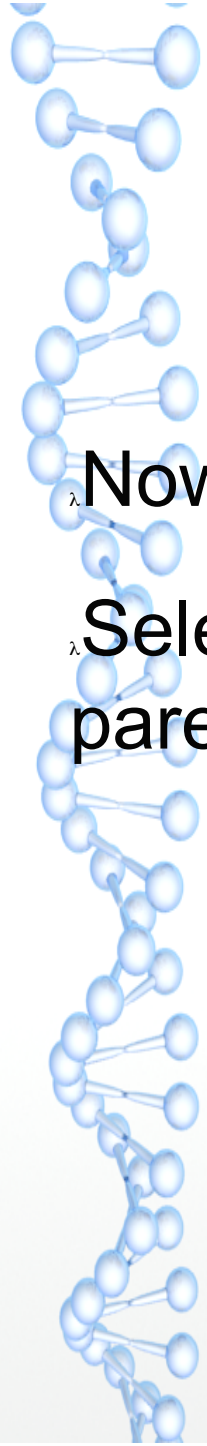
Step 1

- Perform a post-order traversal of the tree
- Compute:

- # union operations = # of changes

Example

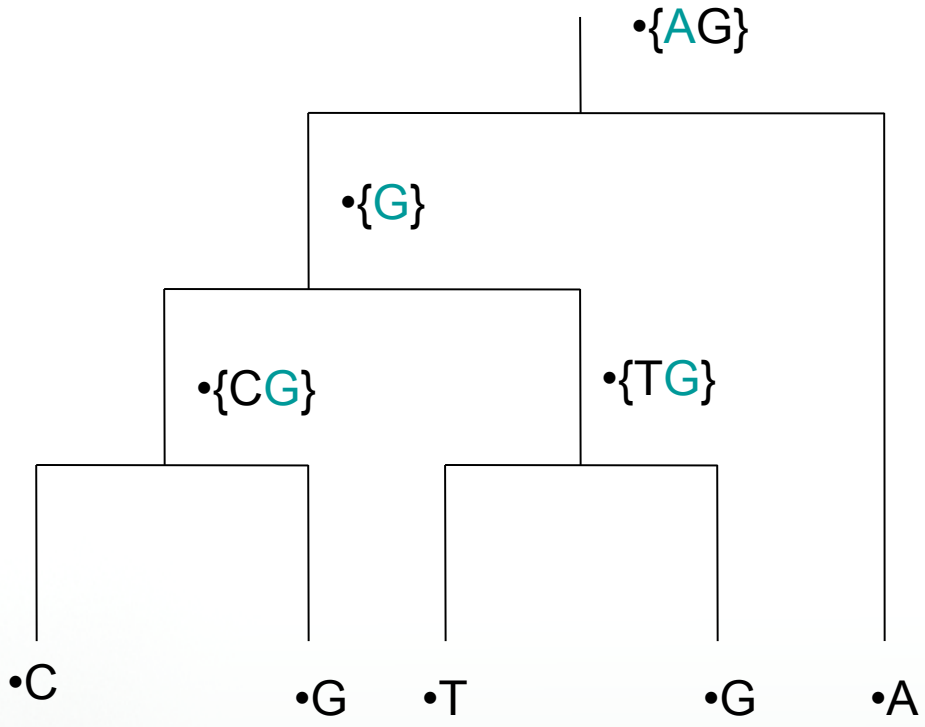


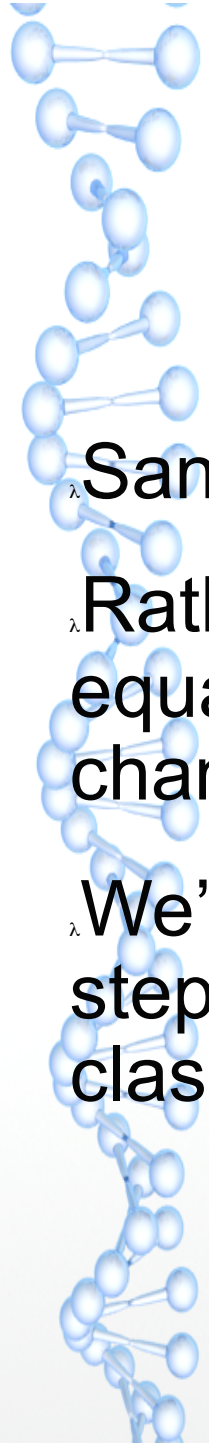


Step 2

- Now do a preorder traversal of tree
- Select state *rj* for an internal node *j* with a parent node *i* as follows:

Example



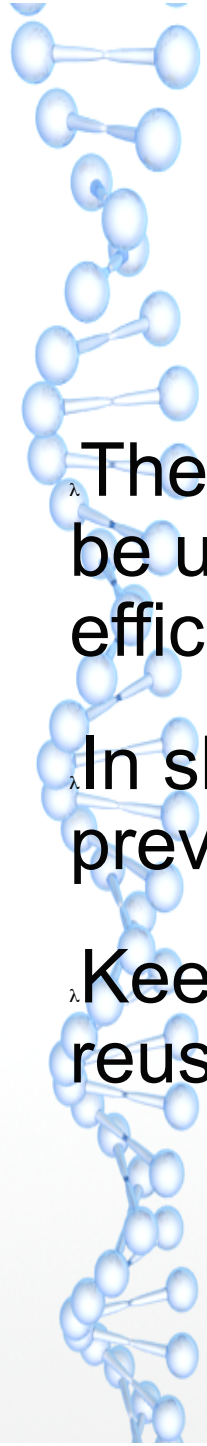


Weighted Parsimony

- Sankoff & Cedergren (1983)
- Rather than assume all state changes are equally likely, use different costs for different changes
- We'll need to propagate costs up during first step of approach, but will not cover this in class

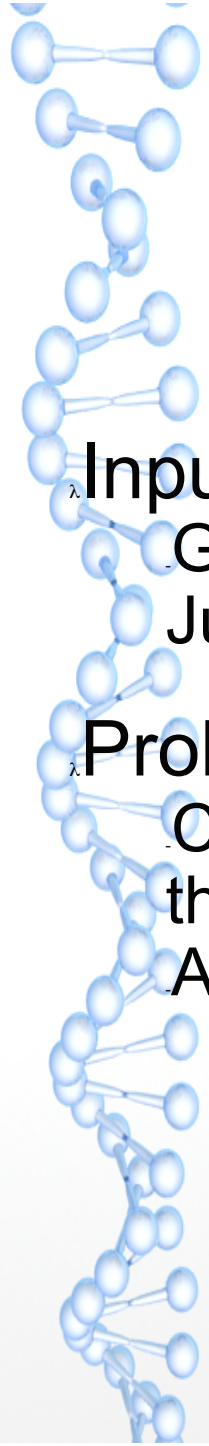
Methods for exploring tree space

- Consider any single internal edge in a tree
- There are 3 ways the four subtrees can be grouped:
  - AB - CD; AC - BD; AD - BC
- Using this rule to look at trees is called nearest neighbor interchange.



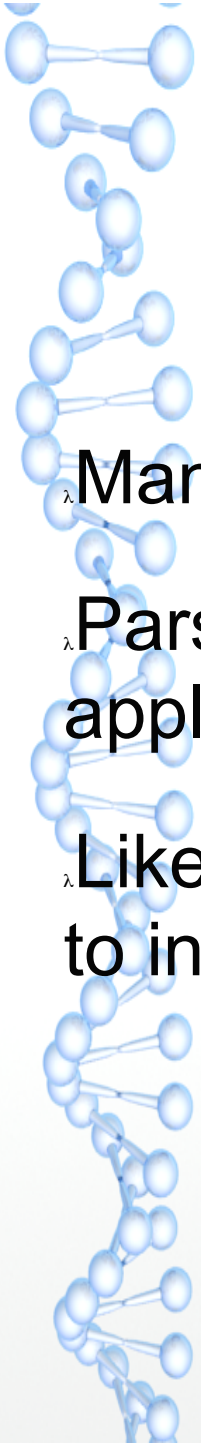
Exact method

- There is a branch and bound approach that can be used to calculate the best tree more efficiently.
- In short, if a tree you build is worse than a previously discovered tree, you stop
- Keep track of all partial trees such that you can reuse information as much as possible



Probabilistic methods

- Input:
  - Given an alignment and a mutation model (e.g., Jukes-Cantor)
- Problem:
  - Compute the likelihood of a tree as a product of the individual likelihoods from the alignment
  - Assumes columns are independent



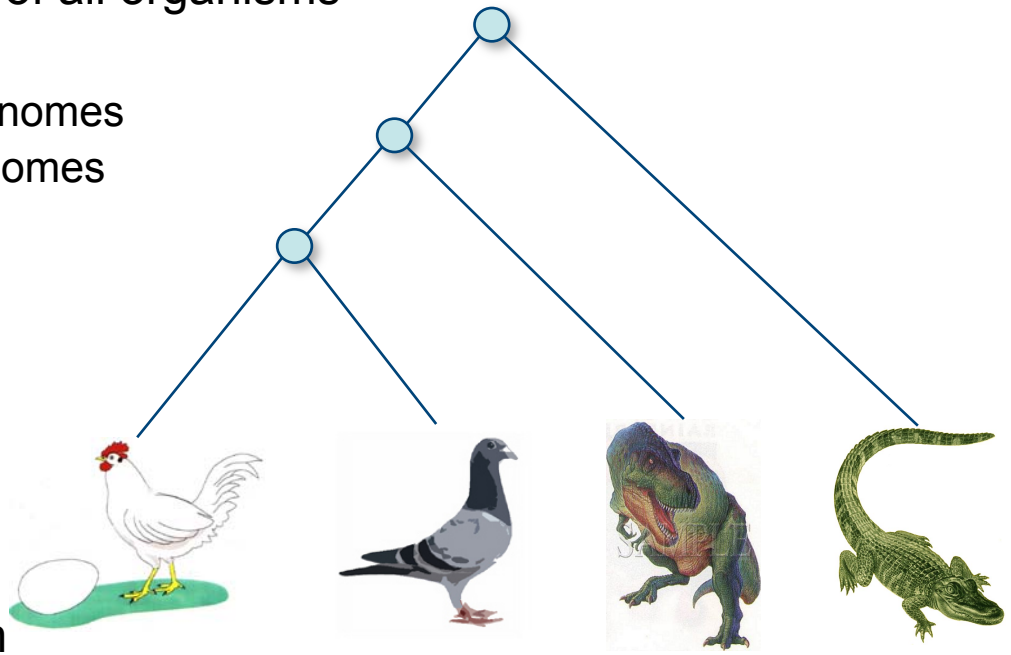
Conclusions

- Many algorithms exist for these problems
- Parsimony generally does pretty well for most applications
- Likelihood, however, is becoming popular due to increased computational power.

Some future directions for sequencing

Organism sequencing

- Sequence a large fraction of all organisms
- Deduce ancestors
  - Reconstruct ancestral genomes
  - *Synthesize* ancestral genomes
  - Clone—Jurassic park!



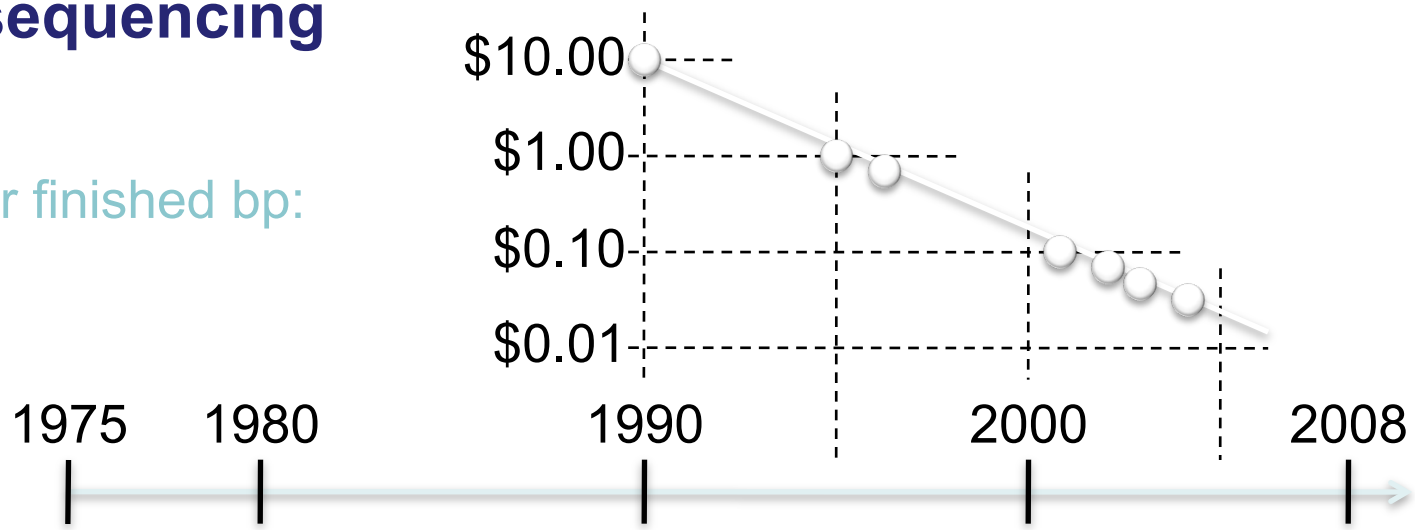
- Study evolution of function
  - Find functional elements within a genome
  - How those evolved in different organisms
  - Find how modules/machines composed of many genes evolved

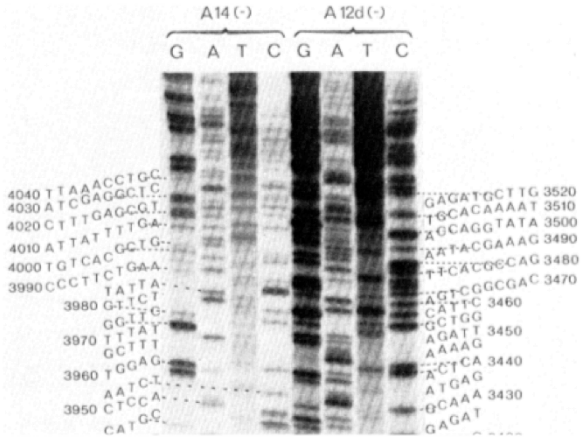
Sequencing technology

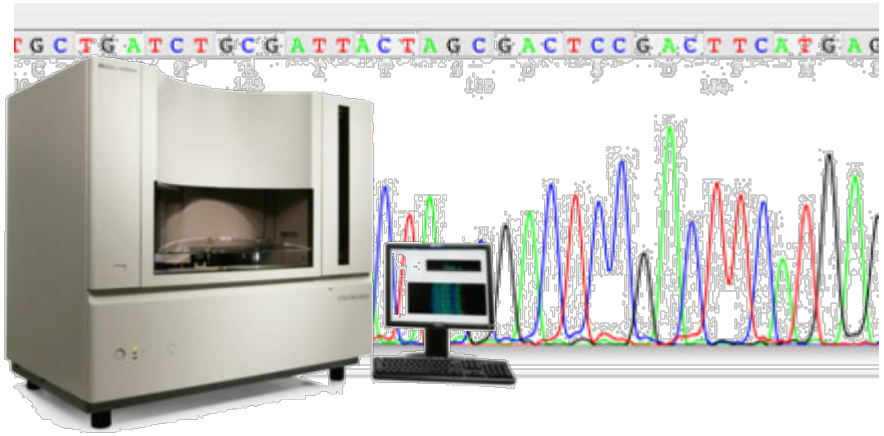
Source bioinformatics.org

Sanger sequencing

Cost per finished bp:







Read length: 15 – 200 bp → 500 – 1,000 bp

Throughput: “grad-student years” → $2 \cdot 10^6$ bp/day

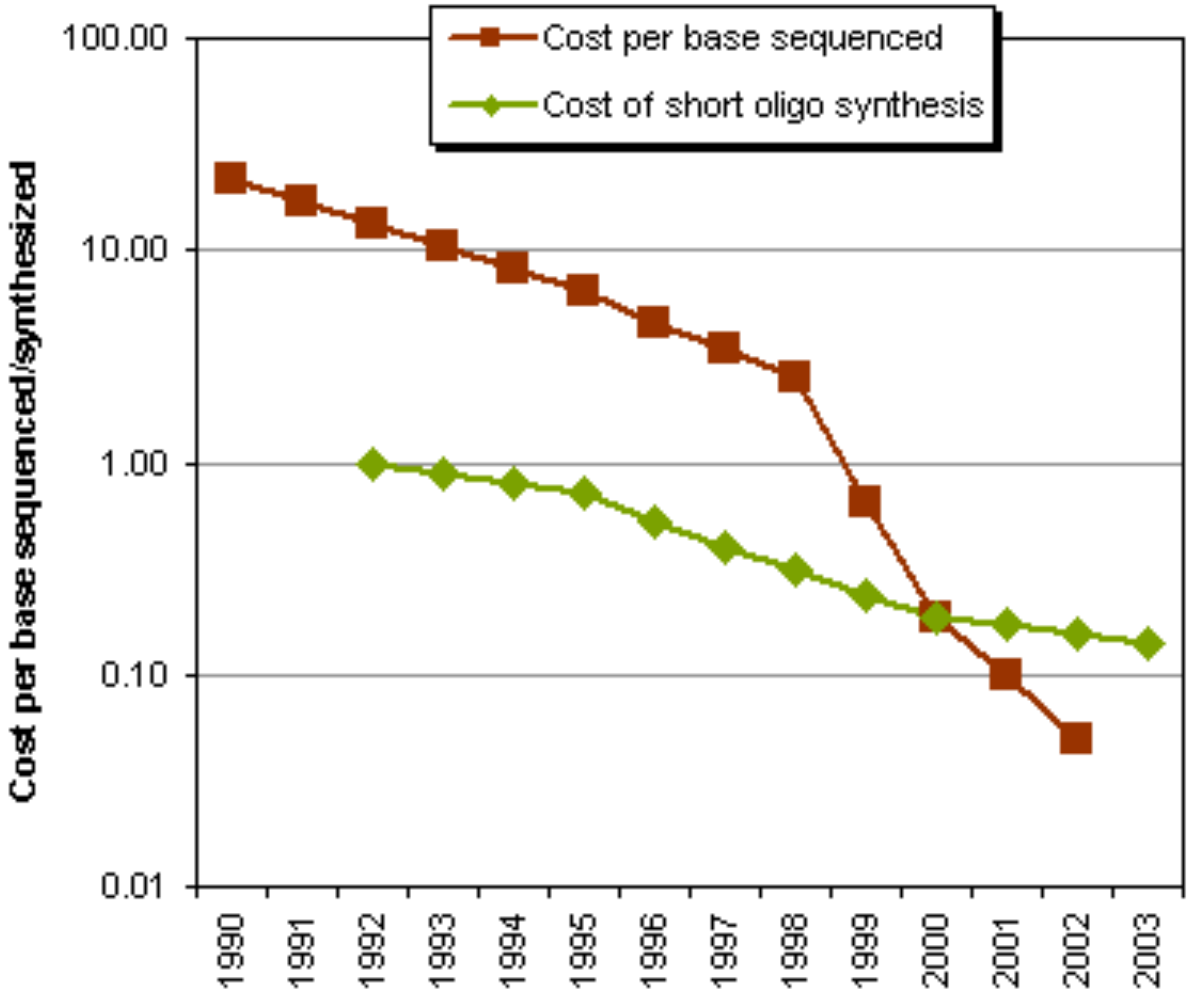
$985 deCODEme
(November 2007)

$399 Personal Genome Service
(November 2007)

$2,500 Health Compass service
(April 2008)

Genetic Information Nondiscrimination Act
(May 2008)

$~1,000 Whole-genome sequencing
(2013)


Cost per base sequenced
Cost of short oligo synthesis
Cost per base sequenced/synthesized
100.00
10.00
1.00
0.10
0.01
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003

Sequencing technology

Next-generation sequencing

Genome Sequencer / FLX

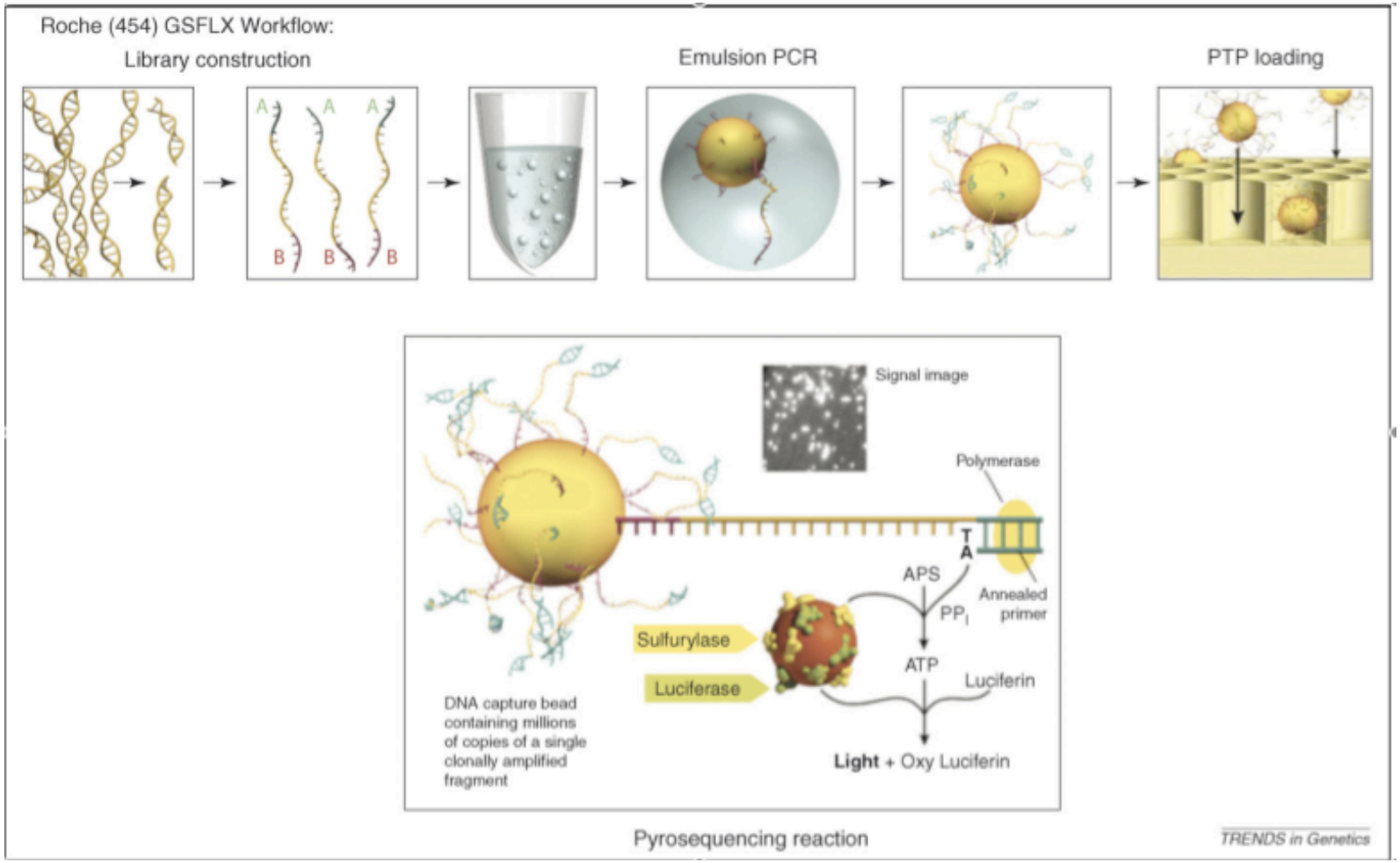
Read length: 450 bp "short reads"

Throughput: 600-800 Mb/day

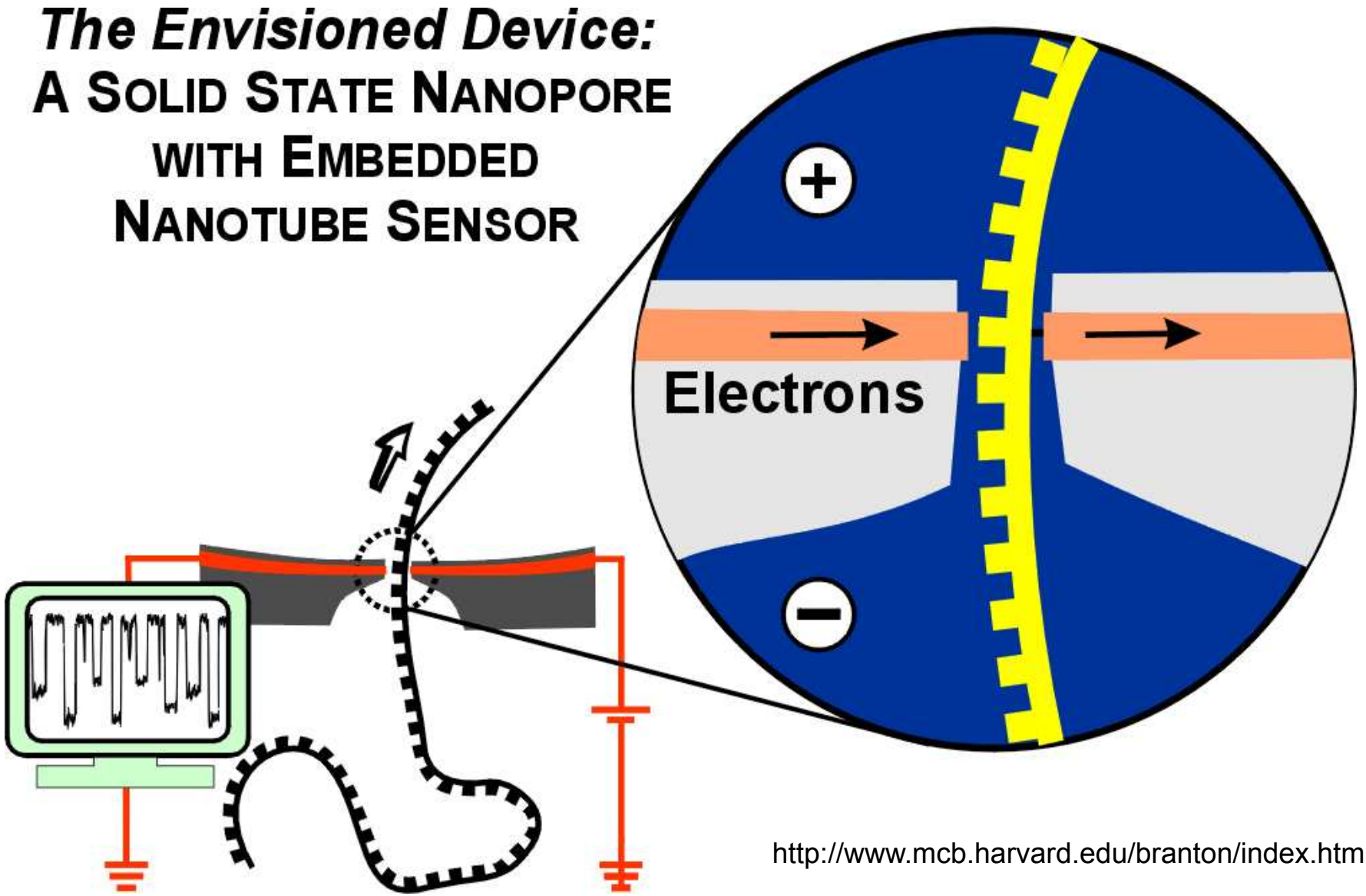
Cost: ~ 20,000 bp/$

De novo: yes

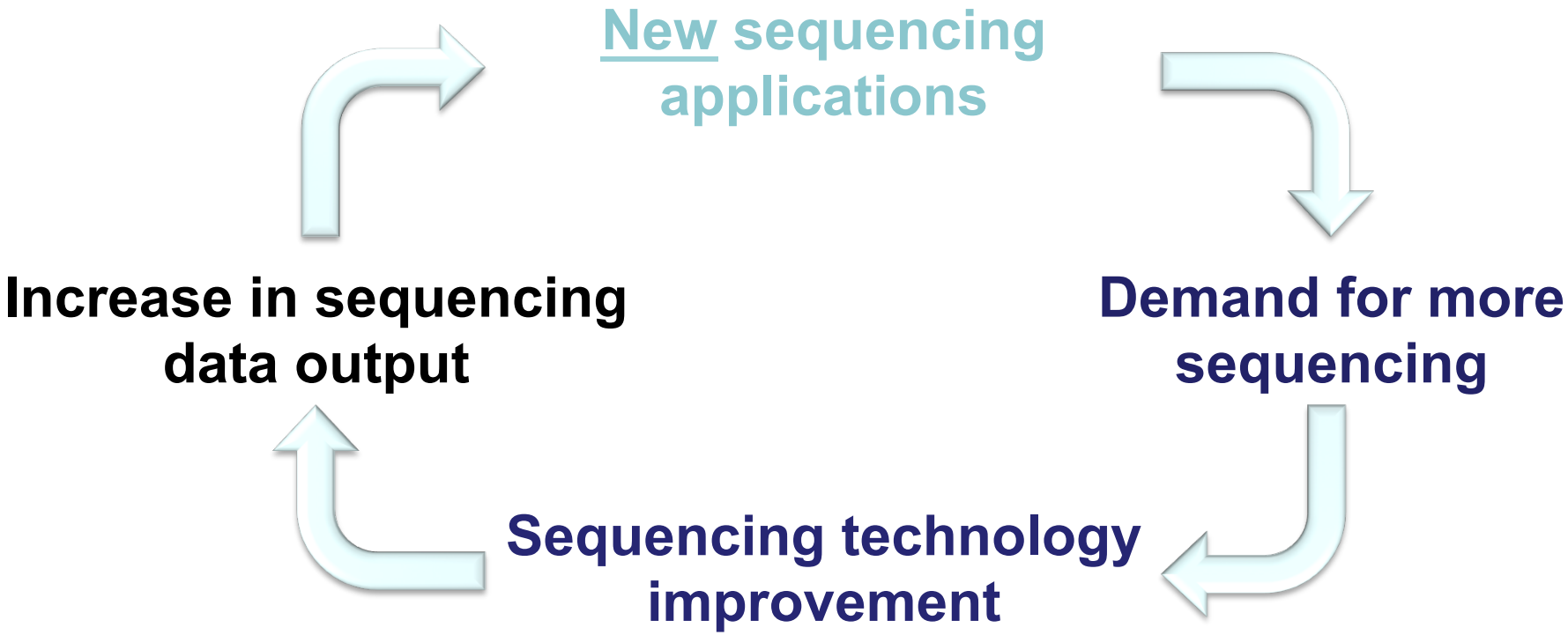
454 Sequencing



Nanopore Sequencing



http://www.mcb.harvard.edu/branton/index.htm

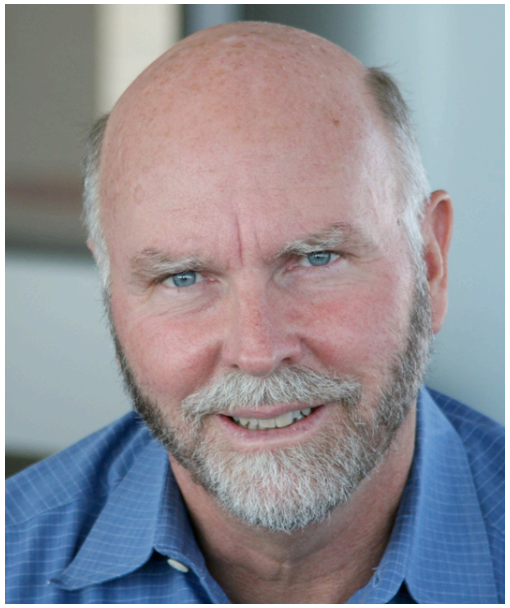

New sequencing applications
Demand for more sequencing
Sequencing technology improvement
Increase in sequencing data output

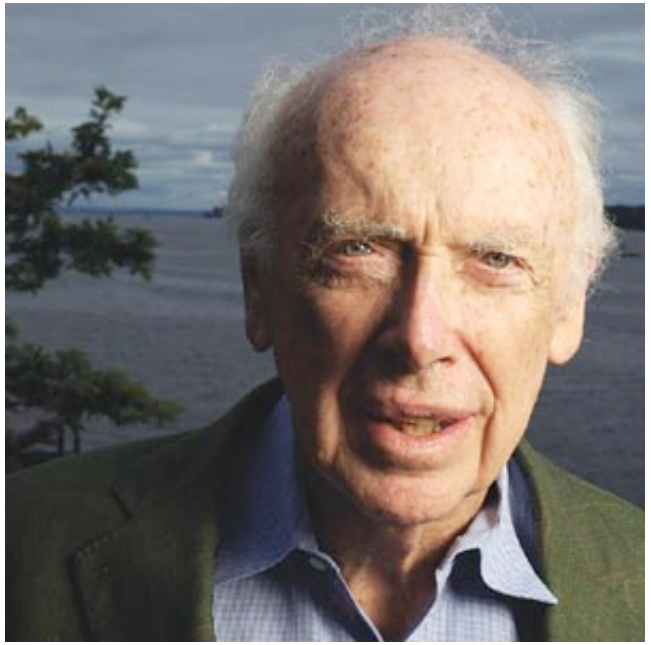
Applications

Whole-genome sequencing
Comparative genomics
Genome resequencing
Structural variation analysis
Polymorphism discovery
Metagenomics
Environmental sequencing
Gene expression profiling
Genotyping
Population genetics
Migration studies
Ancestry inference
Relationship inference
Genetic screening
Drug targeting
Forensics

Goal

- Sequencing a human genome





Technology	Read length (bp)	Pairing	bp / $	*de novo*
Sanger	1,000	longish	1,000	yes
454	250	shortish	10,000	yes
Solexa/ABI	30	shortish	100,000	maybe

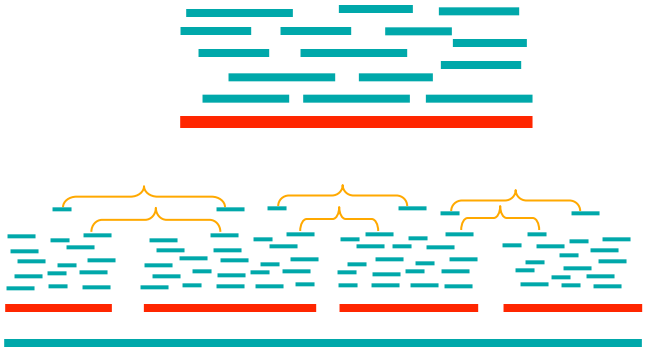
Application	Sanger	454	Solexa/ABI
Bacterial sequencing	yes	yes	maybe
Mammalian sequencing	yes	sort of	probably no
Mammalian *re*sequencing	Lots of $$	Lots of $$	yes

Overlap-Layout-Consensus

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs



Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

Triazzle: A Fun Example

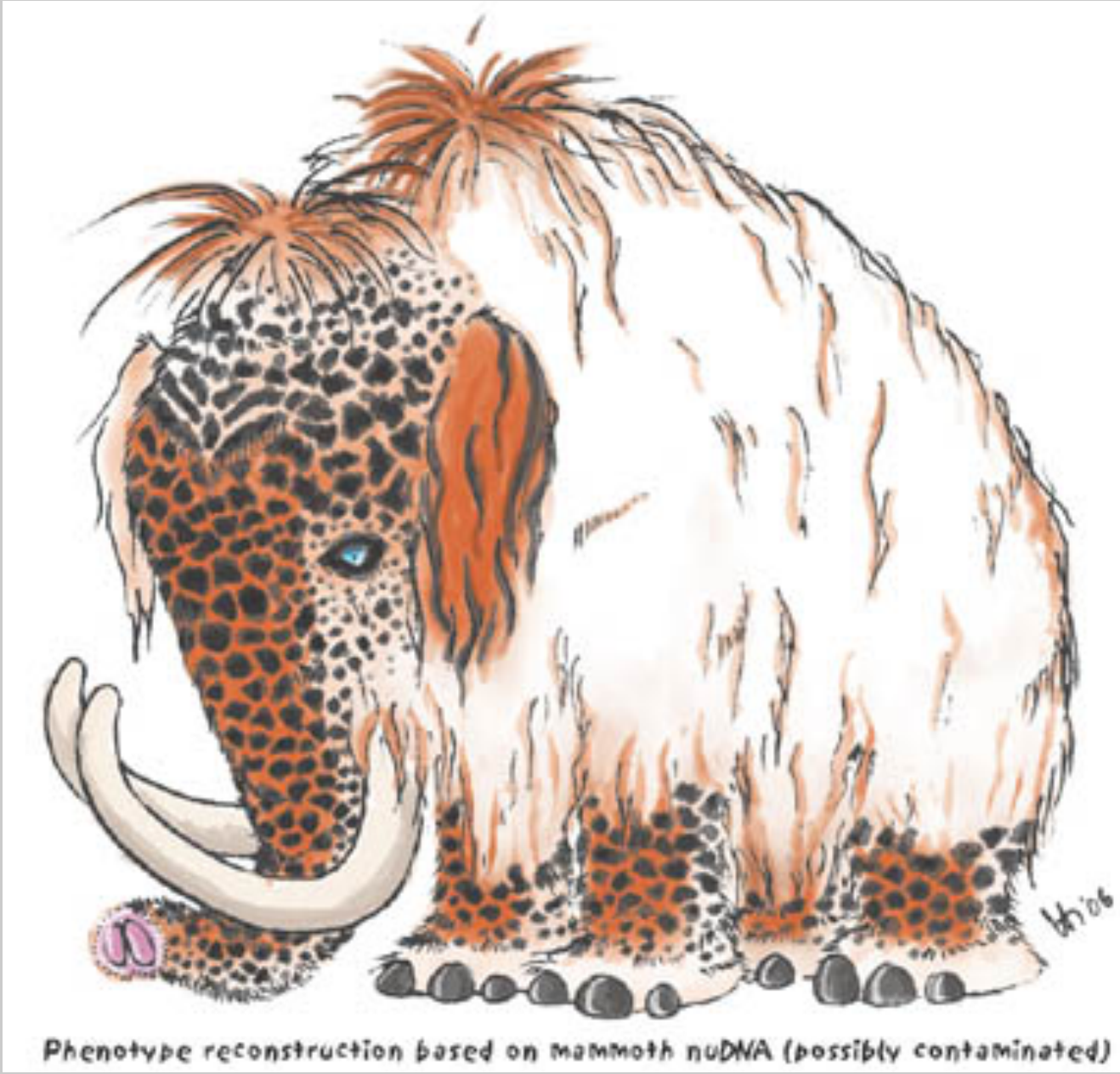
The puzzle looks simple

BUT there are repeats!!!

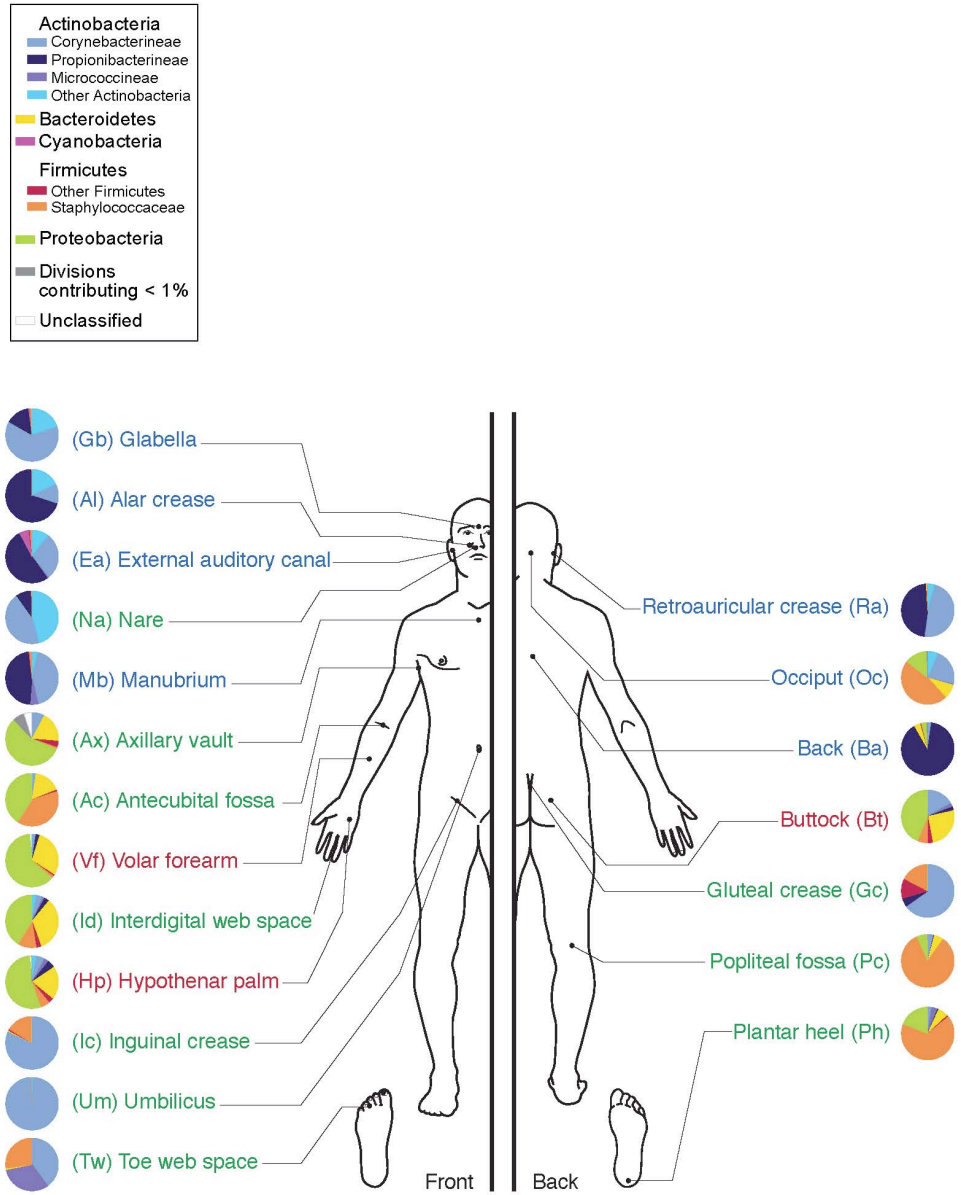
The repeats make it very difficult.

Try it – $10.99 at www.triazzle.com

iPhone version too!



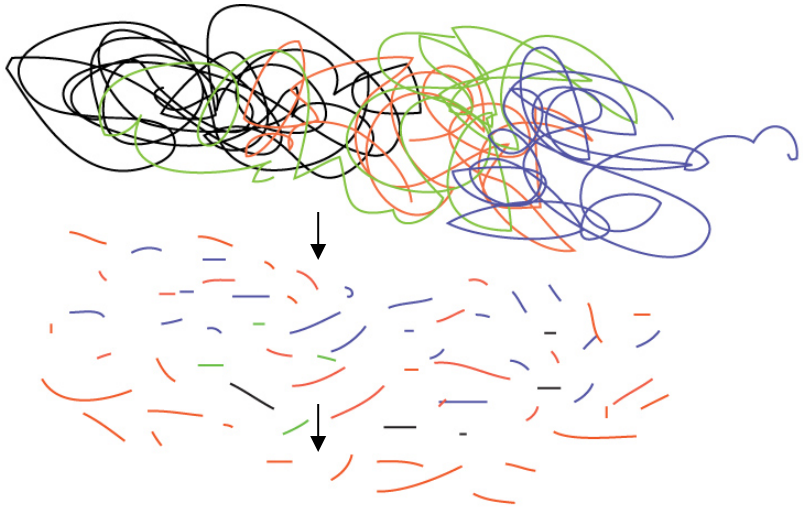
EMBO reports 7, 2, 136–139 (2006)


Actinobacteria
Corynebacterineae
Propionibacterineae
Micrococcineae
Other Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Other Firmicutes
Staphylococcaceae
Proteobacteria
Divisions contributing < 1%
Unclassified
(Gb) Glabella
(Al) Alar crease
(Ea) External auditory canal
(Na) Nare
(Mb) Manubrium
(Ax) Axillary vault
(Ac) Antecubital fossa
(Vf) Volar forearm
(Id) Interdigital web space
(Hp) Hypothenar palm
(Ic) Inguinal crease
(Um) Umbilicus
(Tw) Toe web space
Retroauricular crease (Ra)
Occiput (Oc)
Back (Ba)
Buttock (Bt)
Gluteal crease (Gc)
Popliteal fossa (Pc)
Plantar heel (Ph)
Front
Back


PHASE TWO: INTERPRETATION
SHERMAN Inquirer-Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

Whole Gene Shotgun Sequencing for Metagenomics

Multiple genomes



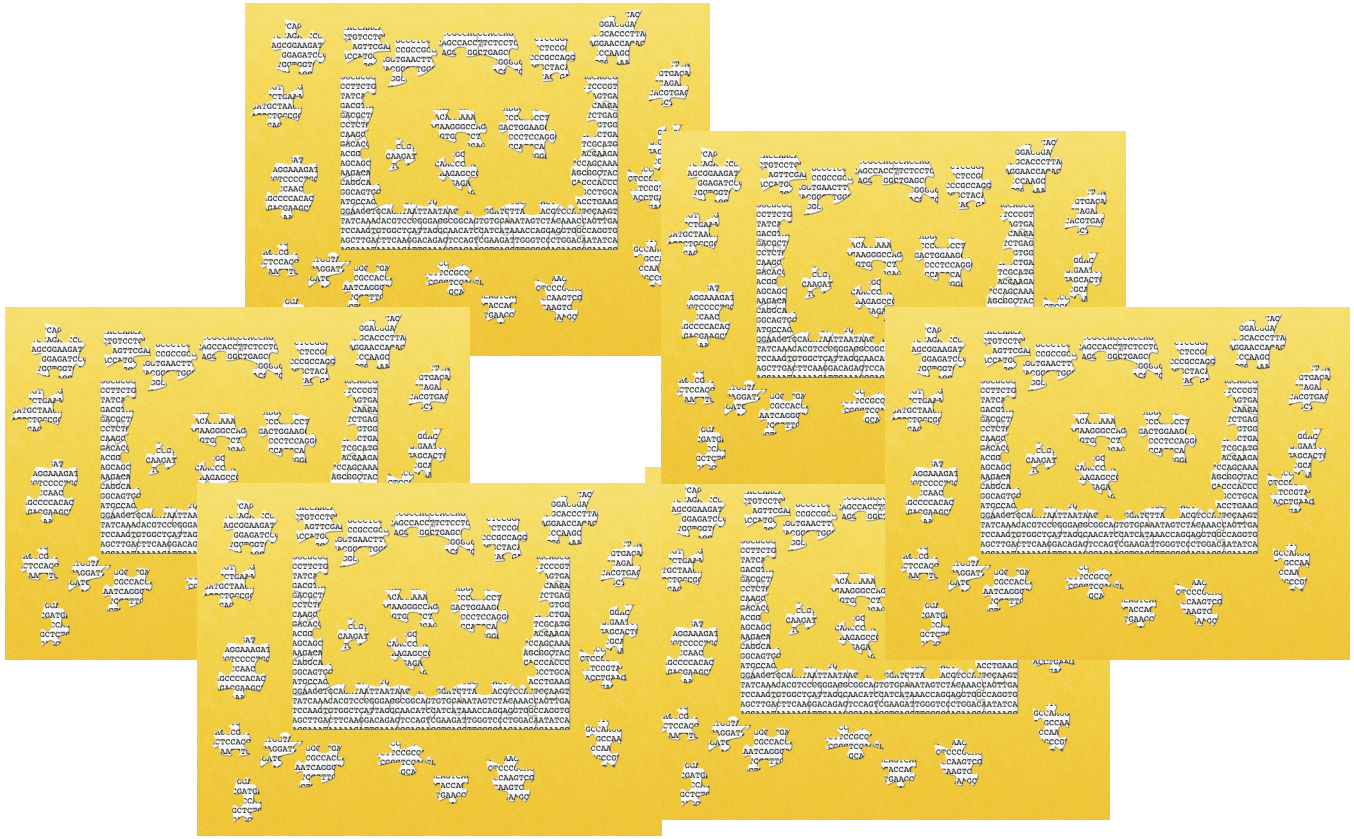
Random genomes fragmentation

...ACACATATACATACAGAGATAGCCCAGATG
AGCCCAGATGGCGCTGCTGCTGGGCGCG.....

...ACATACATACAGAGATAAAAGATG
AAAAGATGCCAGATGGCGCTGCTGCTGGGCGCG.....

...ACACCATACAGAGATAGGGGTGGG
GGGTGTGGAGCCCAGATGGCGCTGCTGCTGG.....

Genomes assembly using overlaps



“Working with thousands of jigsaw puzzles"

What is Metagenomics?

- **Metagenomics** (**Environmental Genomics** or **Community Genomics**) is the study of genomes recovered from environmental samples
- Pro: No need to culture (grow in lab) them
- Con: Heavily uses bioinformatics tools to facilitate insight

Why is Metagenomics Important?

- Some reasons include:
  - Organisms can be studied directly in their environments
  - There are significant advantages for viral metagenomics, because of difficulties cultivating the appropriate host
  - Genomic information has advanced research in a diverse fields such as forensic science

Many projects, many fragments

- Examples:
  - Prokaryote:
    - Sargasso Sea (Venter et al 2004) : **1.6 billion** base pairs generated estimated to come from **1800** genomic species
  - Viral:
    - Marine water (Breitbart et al 2002) Mission Bay and Scripps Pier. 873 sequences for the Mission Bay and **1061** for Scripps Pier with respectively more than **65%** and **73%** of unknown

Many projects, many fragments

- Three years after the Marine Water project, most of sequences are still unique. Despite the fact that GenBank has more than doubled in size.

- All of the Metagenome projects have generated enormous amounts of data that still cannot be assembled or annotated.