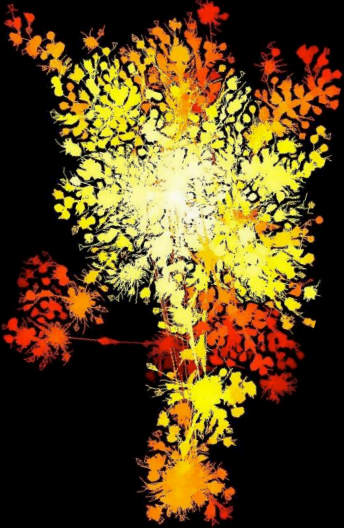
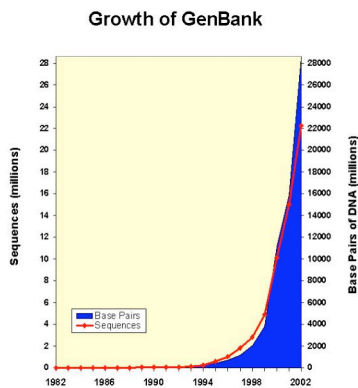


What is bioinformatics?

- Computers and biology?
- the **organization** and the **analysis** of biological data
- The new biology of the 21st century?

A network diagram with yellow and orange nodes and red connections, resembling a tree or branching structure.

What is bioinformatics?

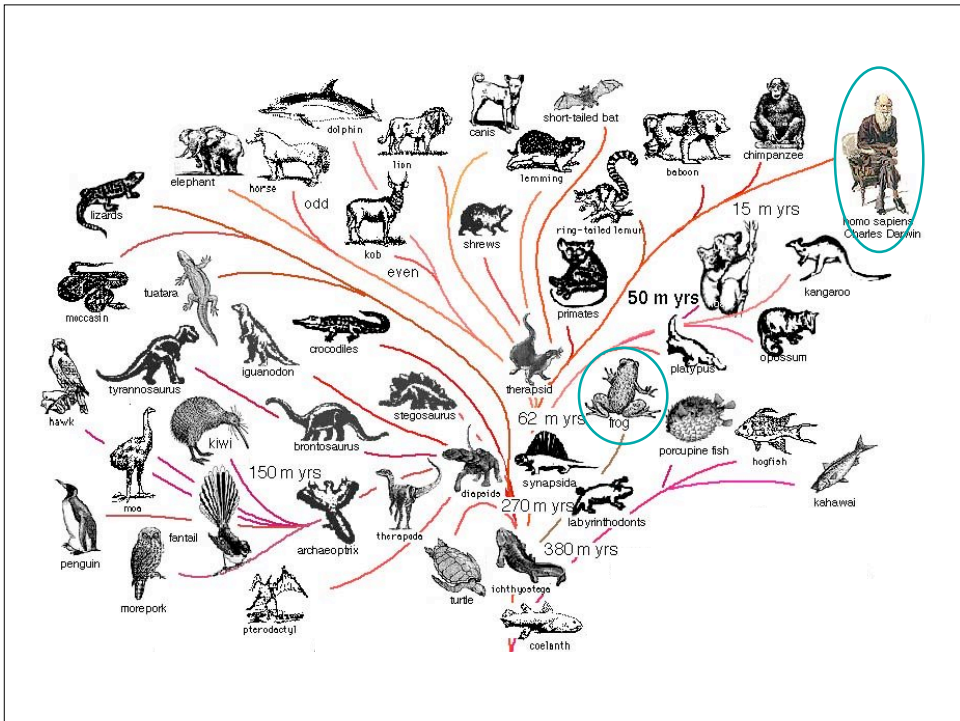
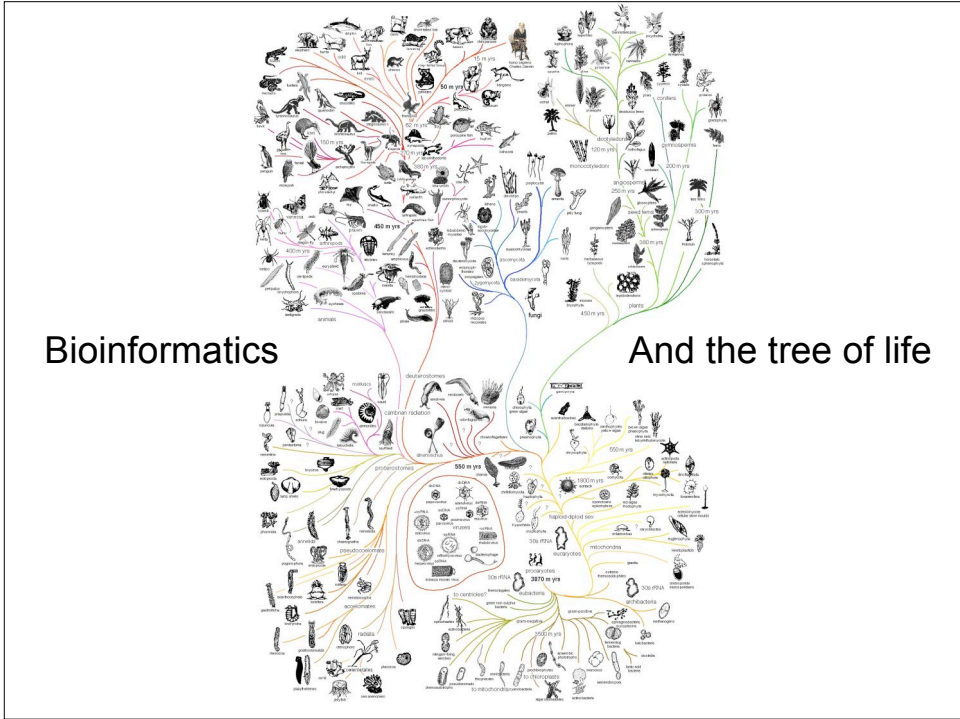


- Rapidly growing biological databases contain information about
 - Cellular and molecular biology
 - Ecology and Evolutionary biology
 - Microbiology
 - Genomics
 - Proteomics

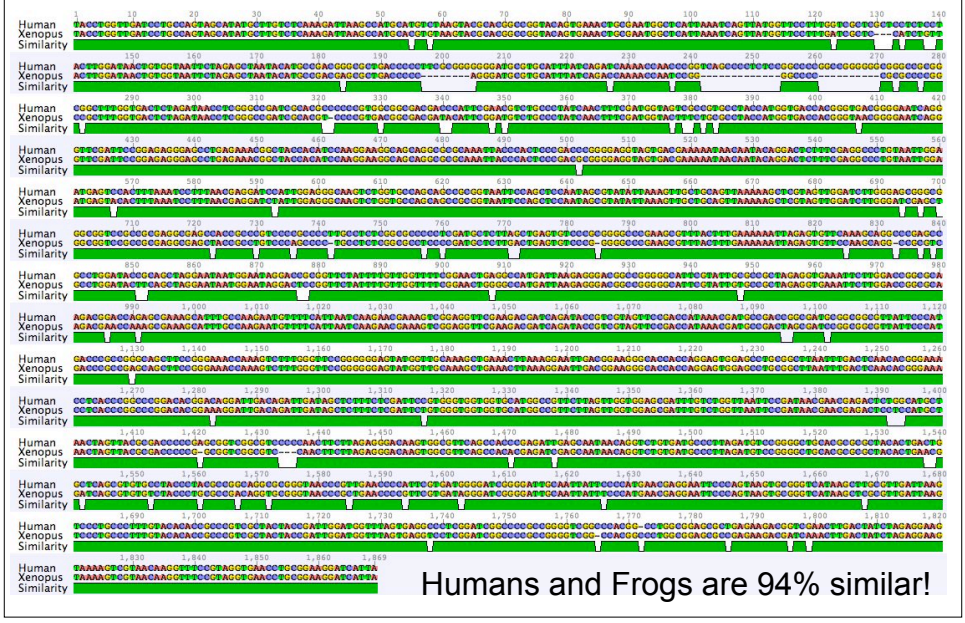
What is bioinformatics?

- Bioinformatics combines the tools and techniques of
 - mathematics,
 - statistics,
 - computer science and
 - biology
- in order to understand biological data (i.e. to understand biology!)

<i>E. coli</i>	KSTCTGVEMFRKLLDEGRAGENVGVLLRGIKREEIERGQVLA-----KPGTIKPHTKFESEVY
<i>P. woesei</i>	KS----IEMHHEPLEEALPGDNIGFNVRGVSKNDIKRGDVAG-HTNPPTVVRTKDTFKAQII
<i>H. salinarum</i>	KT----IEMHHEVPNAEPGDNVGFNVRGIGKDDIRRGDVCG-PADPPPSV---ADTFQAQVV
<i>H. sapiens</i>	KS----VEMHREALSEALPGDNVGFNVKNVSVKDVRRGNVAGDSKNDPPME---AAGFTAQVI
<i>S. acidocaldarius</i>	RS----IETHRTKIDKAEPGDNIGFNVRGVVEKKDVKRGDVAG-SVQNPPTV---ADEFTAQVI



Pairwise alignment of human and frog



Pairwise sequence alignment

Sequences

$$x = a c g g t s c a$$
$$y = a w g c c t t c a$$

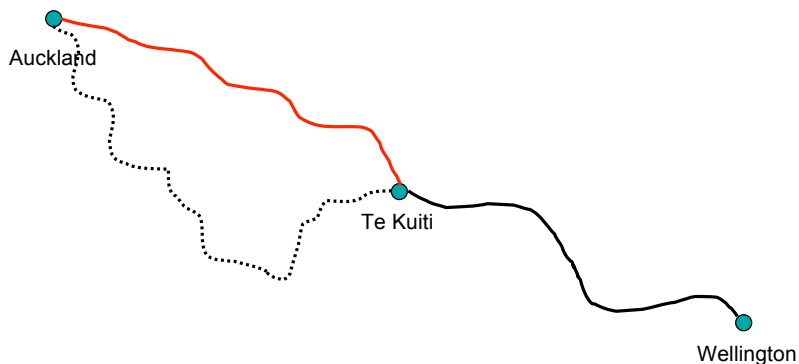
Alignment

$$x' = a - c g g - t s c a$$
$$y' = a w - g c c t t c a$$

Dynamic programming algorithm

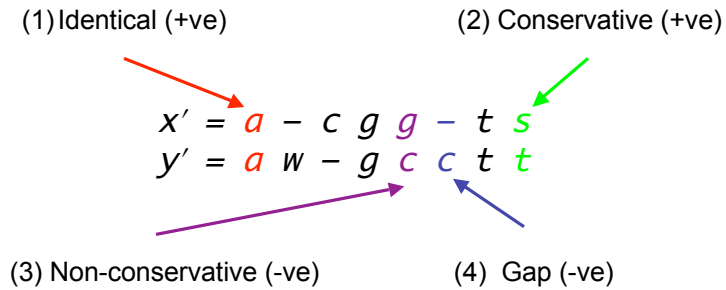
- computation is carried out bottom-up
- store solutions to subproblems in a table
- all possible subproblems solved once each, beginning with smallest subproblems
- work up to original problem instance
- only optimal solutions to subproblems are used to compute solution to problem at next level
- DO NOT carry out computation in recursive, top-down manner
- same subproblems would be solved many times

Principle of Optimality



Scoring

- Numeric score associated with each column
- Total score = sum of column scores
- Column types:

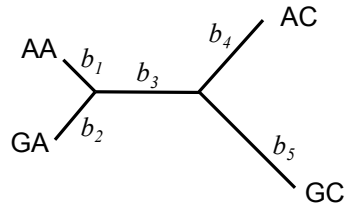
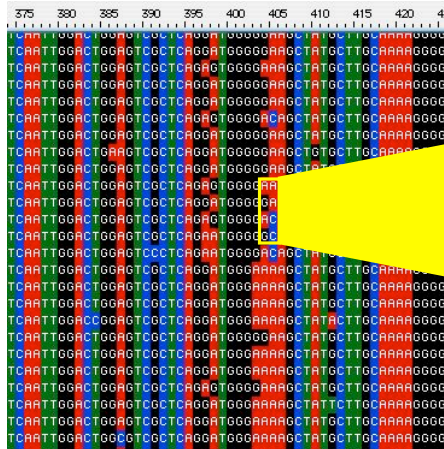


Comparing many species



This is called a “multiple sequence alignment”

Reconstructing the evolutionary tree

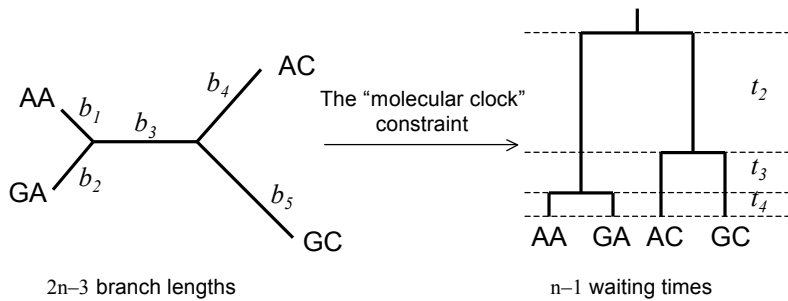


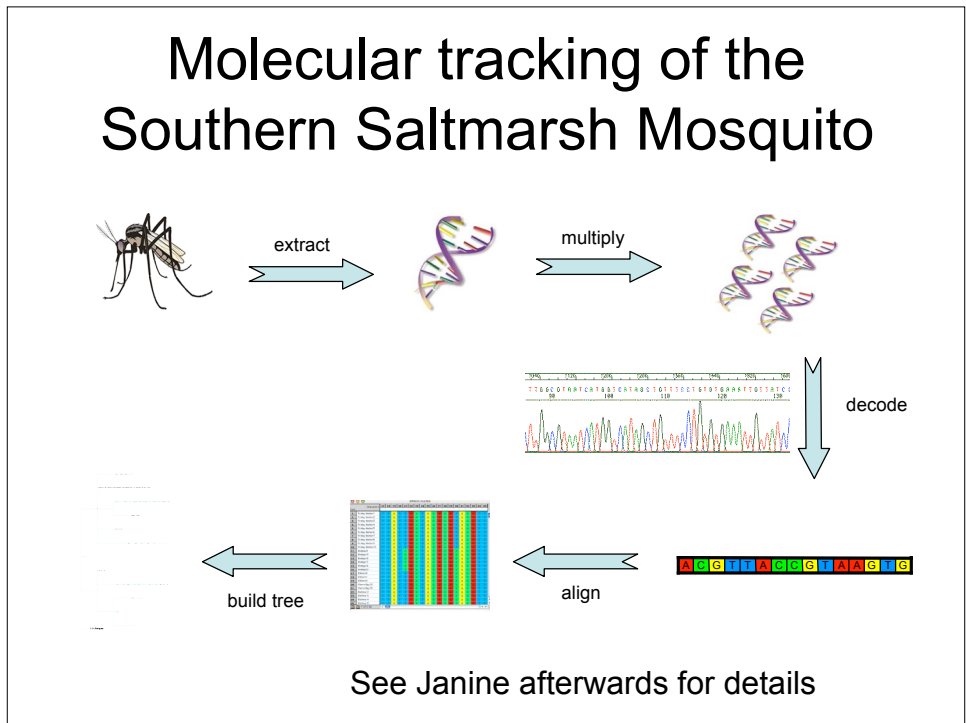
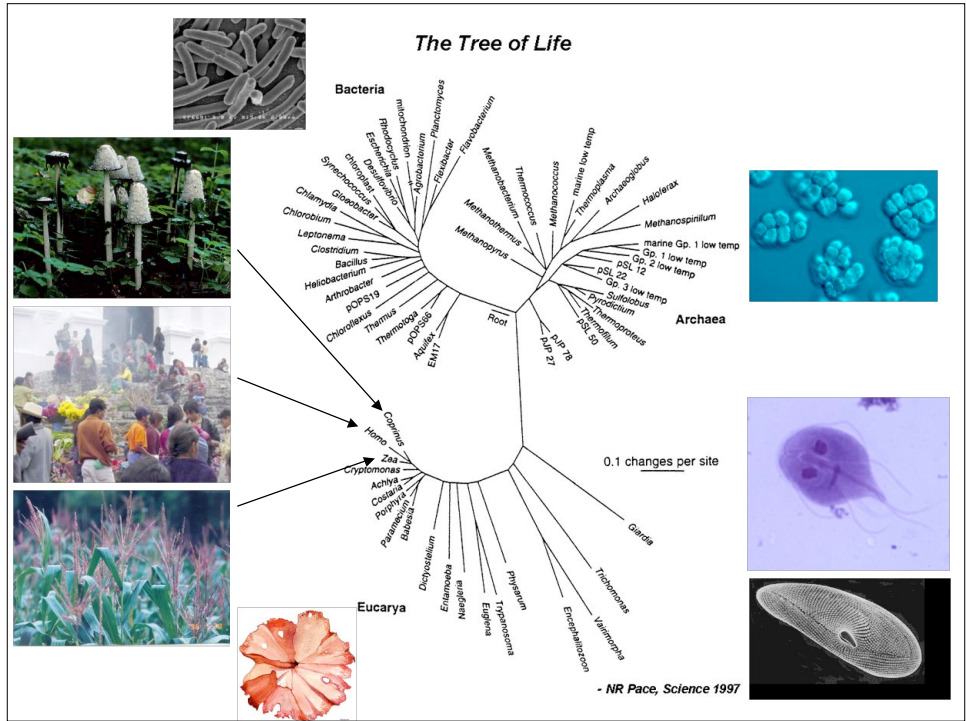
The probability of the sequence alignment,

$$\Pr\{D | T, Q\}$$

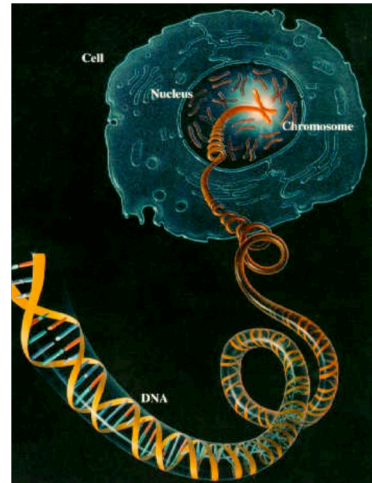
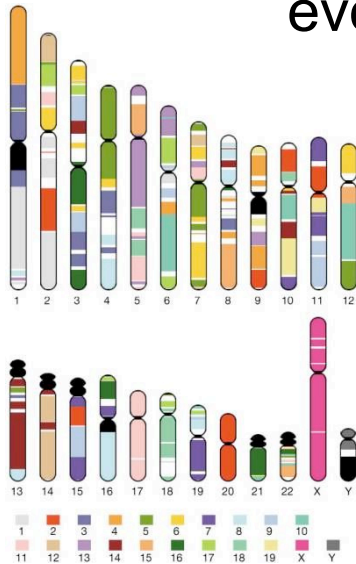
can be efficiently calculated given a tree and branch lengths (T), and a probabilistic model of mutation represented by an instantaneous rate matrix (Q). **In phylogenetics (the study of evolutionary trees), branch lengths are traditionally unconstrained.**

The molecular clock assumption



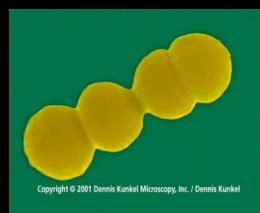
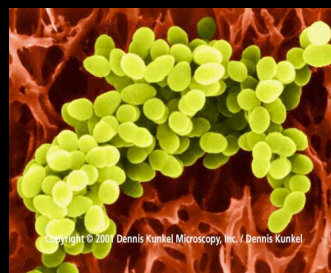


Bioinformatics and genome evolution



Bioinformatics and Disease

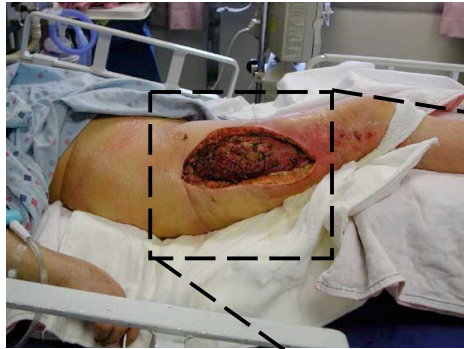
- Bacterial infection can cause serious illnesses, e.g.
 - Necrotising Fasciitis (“Flesh eating disease”)
 - Toxic Shock Syndrome (TSS)
 - (e.g. the illness Lana Cockroft got from the cut in her foot during “Celebrity Treasure Island”).



T: *Staphylococcus aureus*

B: *Streptococcus pyogenes*,

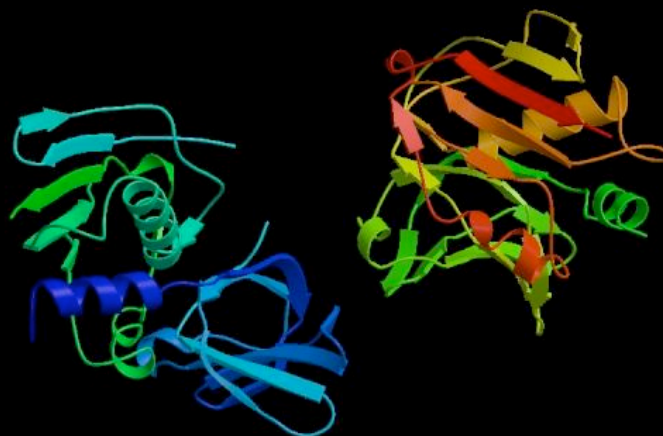
Bioinformatics and Disease



Necrotizing fasciitis



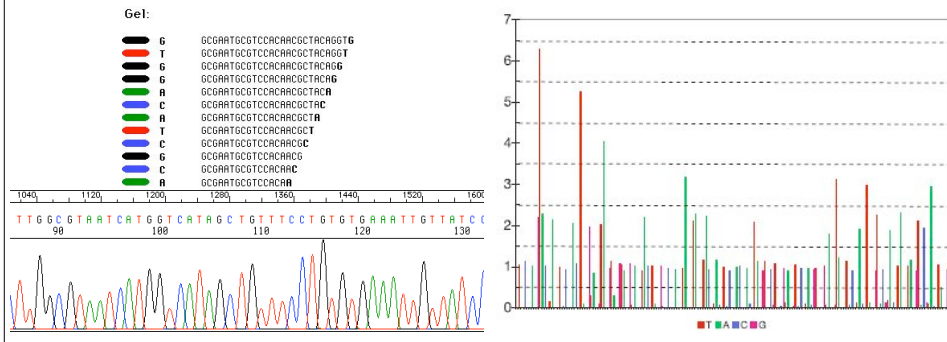
Bioinformatics and Disease



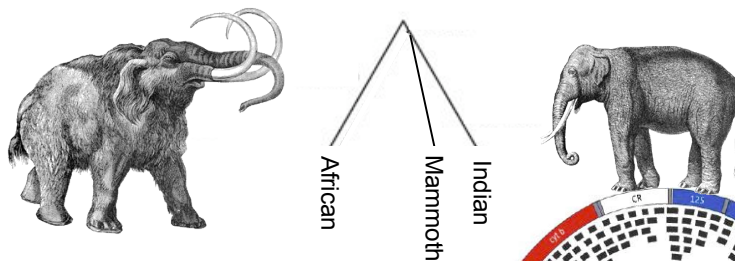
Identify protein structure of superantigens
- so that potential drug targets can be uncovered.

Bioinformatics and sequencing technology

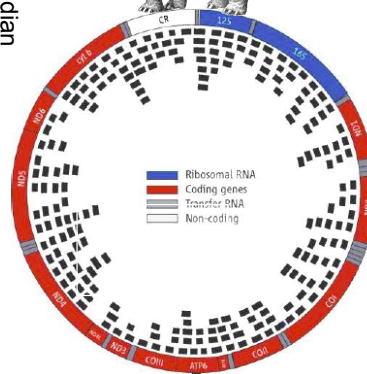
- “Old” Technology
 - 1977
 - Small Scale
 - Slow
 - Length: long \approx 700bp
- New Technology: 454
 - 2003
 - Large scale: Whole genome
 - Fast
 - Length: short \approx 100bp
 - <http://www.454.com>



Bring the mammoth back

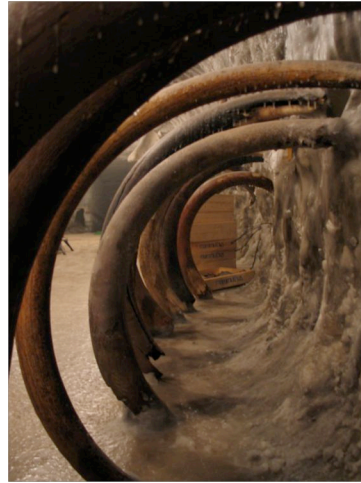


Using this new sequencing technology and bioinformatics to compare the sequences to the African elephant genome, scientists sequenced 28 million nucleotides of the mammoth genome from frozen bones!



Where does the DNA come from?

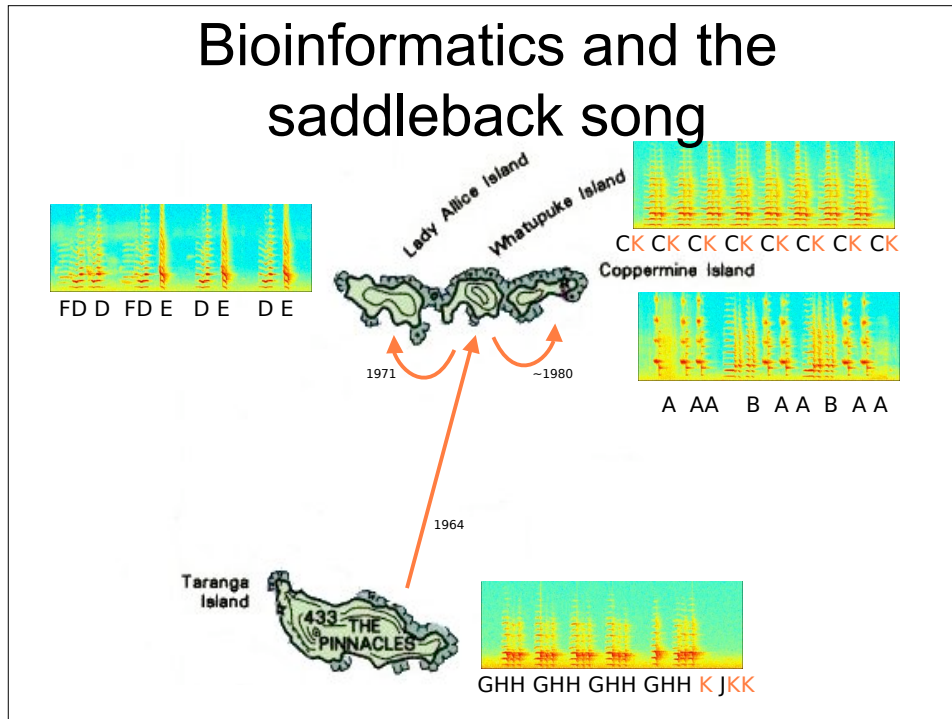
- Scientists Sequence DNA of Woolly Mammoth (extinct 10,000 years ago)
 - <http://www.science.psu.edu/alert/Schuster12-2005.htm>



Bioinformatics and the saddleback song



Bioinformatics and the saddleback song



Conclusion

- Bioinformatics is fast becoming a central part of the biological sciences
 - Ecology (saddlebacks)
 - Evolutionary biology (tree of life)
 - Health and disease (flesh-eating bacteria)
 - Ancient DNA (mammoth)
 - Molecular biology and cell biology
- Bioinformatics draws skills not just from biology, but also from the 'hard' sciences like mathematics, computer science and statistics - this type of knowledge is becoming increasingly important in biology.