Computational Biology in South Africa

Part I

- Some classic problems in molecular evolution
- Extension to an outline of some current research

Part II

- Computational Biology in South Africa
  - Summary of research in South Africa
  - Recent developments
- General perspectives

Bioinformatics

Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology

The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Example – The Phylogeny Problem

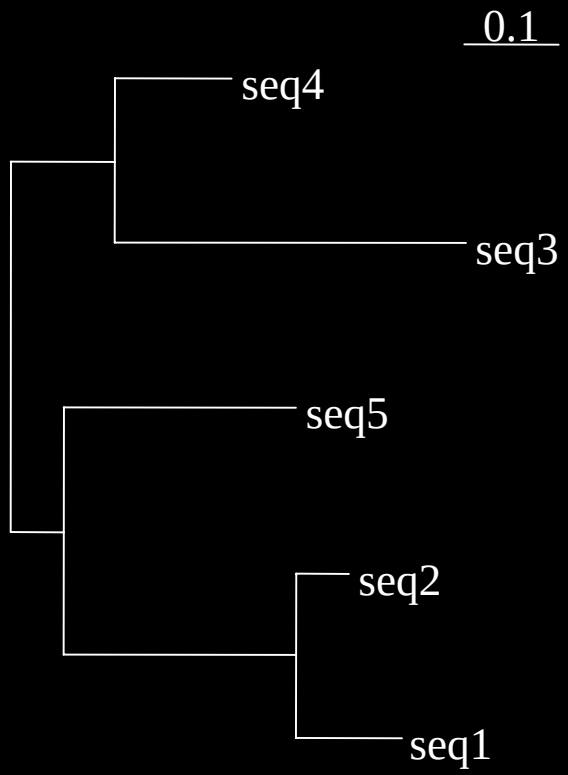
Simplified problem statement: Given a set of aligned DNA sequences that are all derived from a single ancestral sequence through evolutionary processes infer the ‘phylogenetic tree’ that represents the relationships between the sequences.

DNA sequence – a sequence of characters from the four letter alphabet (A, C, G, T)

Aligned sequences

seq1: AGCTAA
seq2: AGCTAG
seq3: GGAGTT
seq4: GCATAT
seq5: ACATGG

Tree



The problem of inferring the phylogenetic tree consists of two parts

I Deciding the score of a tree (given the data)

II Searching among trees for the tree that gives the highest score

Maximum likelihood method for inferring the phylogenetic tree

Tree score = likelihood of the tree (i.e. probability of generating the observed sequences given the tree)

We will discuss

a. how the tree score is calculated

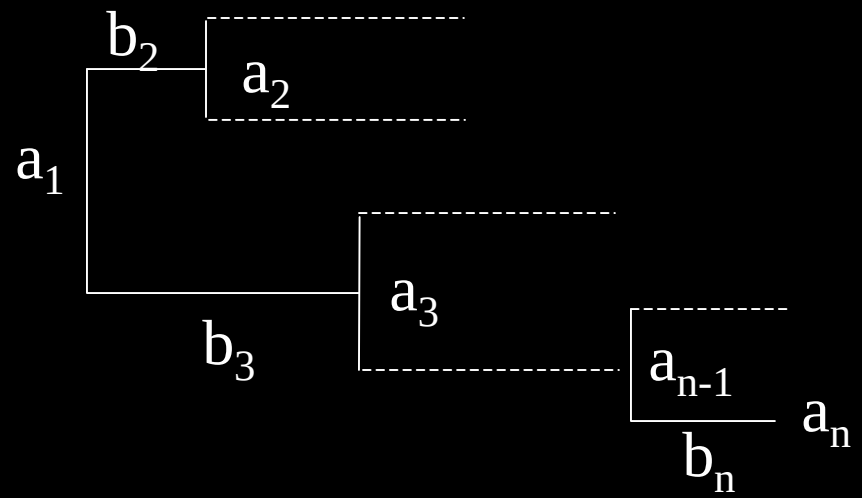
b. how to search among trees for the highest scoring tree

Given a branch with a particular length we can come up with a value for the probability of a given substitution event (replacement of one nucleotide with another) between the two ends of the branch

We can use this to calculate the likelihood of a tree as a product of terms that look like this:

$$L(T) = \sum_{a_1}\sum_{a_2}\sum_{a_3}\cdots\sum_{a_n} P(a_1)P(a_2 \mid b_2, a_1)P(a_3 \mid b_3, a_1)\ldots P(a_{n-1} \mid b_n, a_n)$$

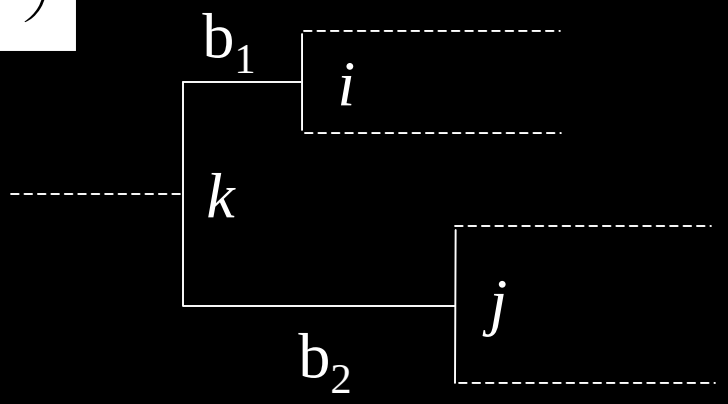
NB: simplification – treating all sites the same!



A standard dynamic programming algorithm can be used to avoid doing all the summations

Let L_s^i be the likelihood of the *subtree* decended from node *i*, given that the character at node *i*, is *s*, then for the situation shown

$$L_r^k = \left(\sum_s L_s^i P(s \mid r, b_1) \right) \left(\sum_s L_s^j P(s \mid r, b_2) \right)$$



Searching for the best tree

There are a great many possible trees for realistic numbers of sequences

$$\frac{(2n-5)!}{2^{n-3}(n-1)!}$$

Branch and bound algorithms developed

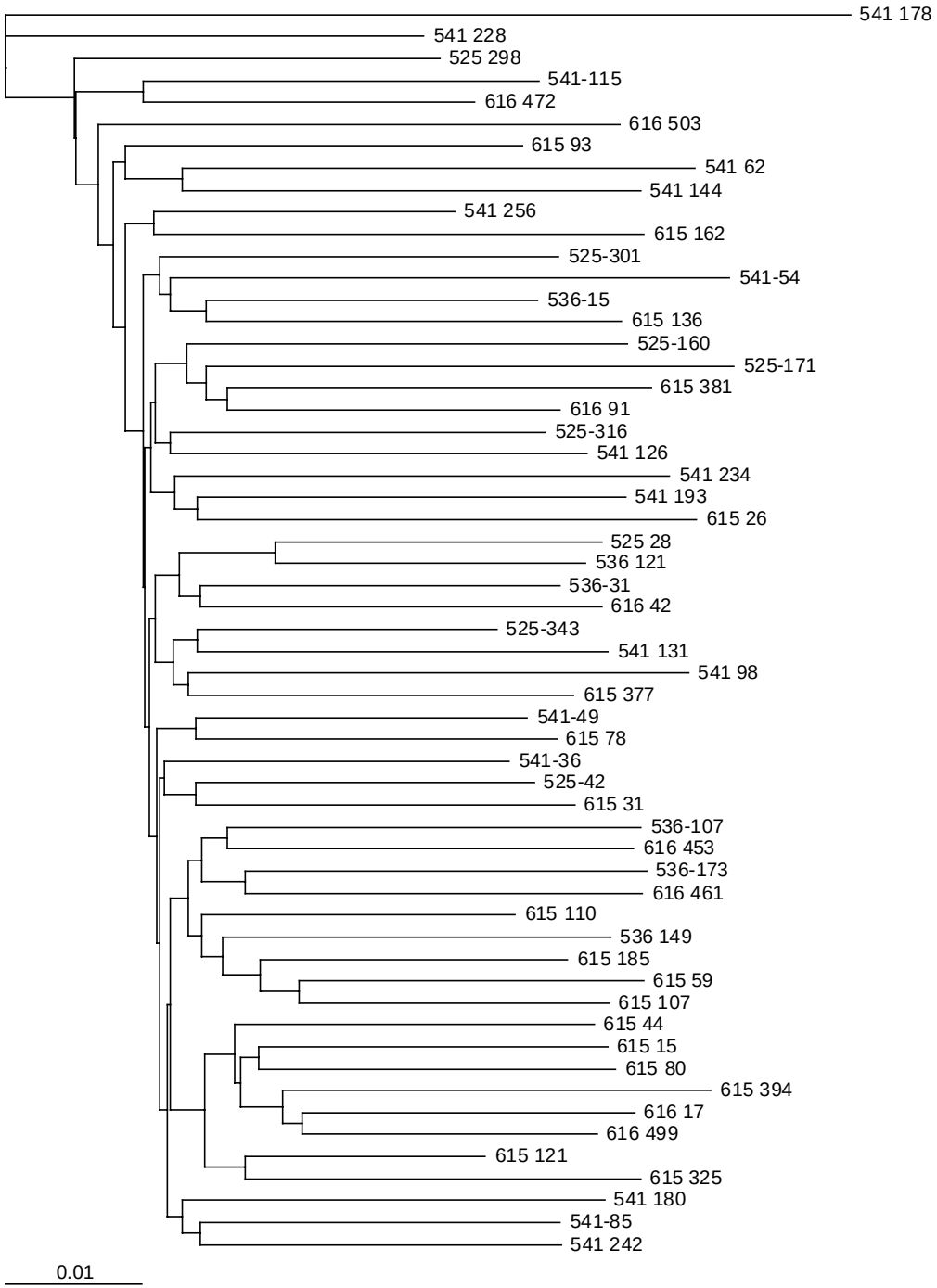
Heuristic algorithms normally applied (greedy algorithms that search in tree neighbourhoods by breaking and reconnecting trees)

An area of active research

Applications –

e.g. molecular

epidemiology



Digression

SACEMA: South African Centre for Epidemiological Modelling and Analysis

Mathematical modeling of

spread of disease (on the population level)

host-viral interactions (cell level)

interaction between diseases (TB-HIV)

etc…

Caprisa: Collaborative AIDS programme of research in South Africa

Nelson Mandela School of Medicine, UKZN

A Bayesian approach to the phylogeny problem

$$P(T \mid D) = \frac{P(D \mid T)P(T)}{P(D)}$$

P(D) would require integration over all possible trees – not possible – but can be estimated using Markov Chain Monte Carlo

Normally flat (uninformative) prior probabilities are used for the trees

Not all sequence sites are the same…

The selection problem:

Given a set of sequences and a tree

a. determine whether there is evidence for a subset of sites evolving under ‘positive selection’

b. if so determine which sites are likely to be evolving in this way

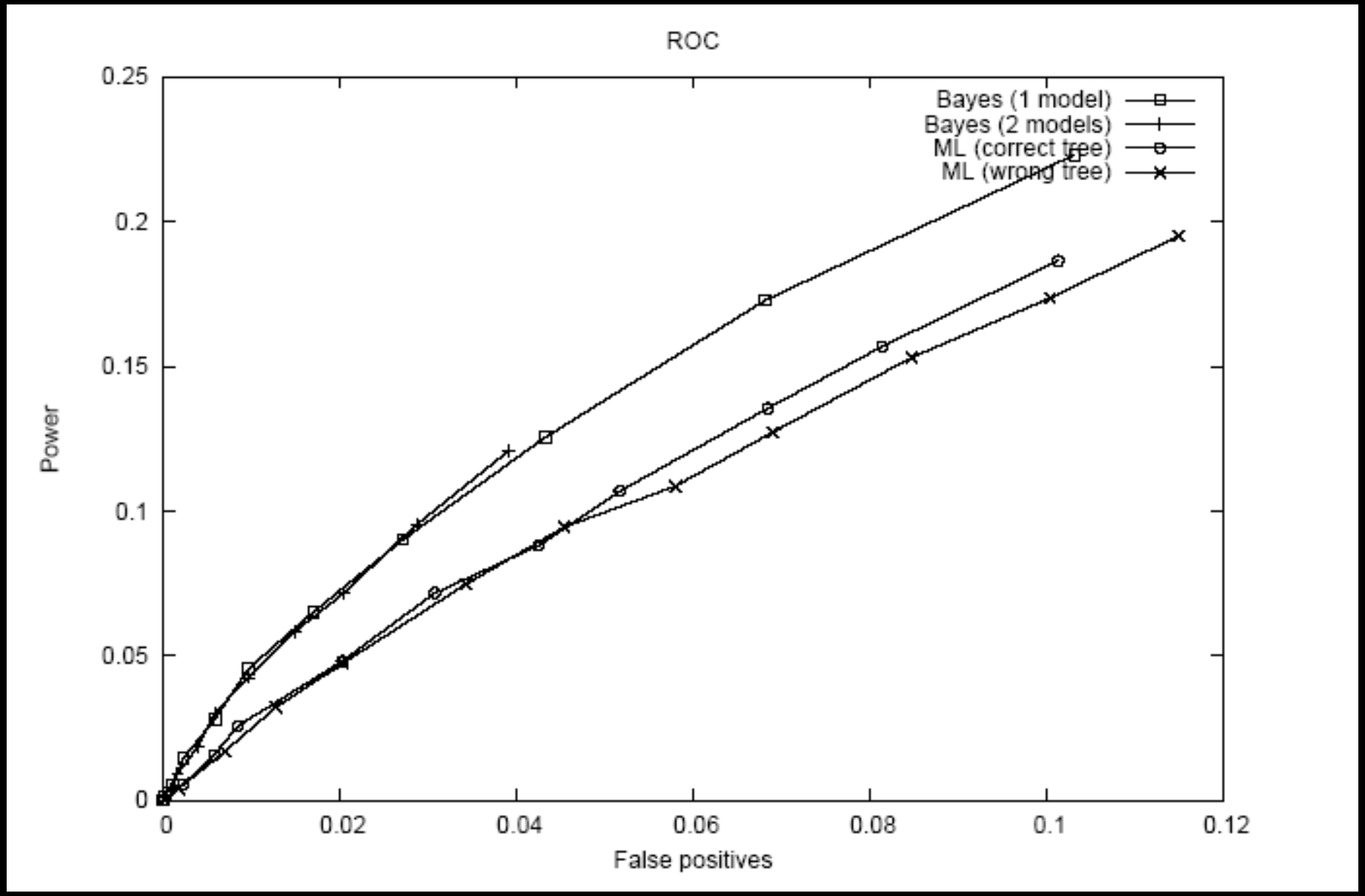
Over the last few years the selection problem has been tackled using model comparison in a likelihood framework (likelihood ratio tests)

$$\frac{P(D|\text{ML values of constrained model}, T)}{P(D|\text{ML values of unconstrained model}, T)}$$

The unconstrained model is the same as the constrained model but with the addition of a class of very rapidly evolving sites

We are currently developing a method to infer selection, taking account of uncertainty in tree topology and other parameters using Bayesian Model Comparison

- Sample from trees and all model parameters according to posterior probabilities of trees/parameters
- Compare evidence in favour of the unconstrained model before and after examining the data


ROC
Bayes (1 model)
Bayes (2 models)
ML (correct tree)
ML (wrong tree)
Power
False positives
0.25
0.2
0.15
0.1
0.05
0
0
0.02
0.04
0.06
0.08
0.1
0.12

More problems derived from molecular sequences

Clustering problem: Given a large number of molecular sequences perform pair-wise comparisons and cluster ‘similar sequences’ – SANBI, University of the Western Cape

Recombination problem: Given a set of sequences infer recombined sequences (i.e. mozaics) – Recombination Detection Program (RDP) – University of Cape Town

Beyond sequence

Expression bioinformatics: The when and where of a gene's function

Structural Bioinformatics: Gene sequences are one dimensional but proteins are not

Systems Biology: Proteins actually function (and are regulated) in very complex networks

Ontologies and text mining: create systems to enable integration and intelligent mining of very large amounts of data

And more…

Expression bioinformatics

The microarray analysis problem: find patterns in large and extremely noisy datasets of gene expression information – University of Cape Town, University of Pretoria/CSIR

Structural Bioinformatics (some examples)

University of Pretoria: Model interactions between drugs and target molecules (especially Malaria proteins)

University of Cape Town & University of Western Cape: MSc in Structural Biology; structure determination; molecular dynamics to estimate constraints on structure given low-resolution protein structure data from NMR.

Systems Biology (examples)

- The Virtual Cell
- Network motifs
- Control analysis (University of Stellenbosch – Triple J Group)

Ontologies

- The Gene Ontology (GO)
- The Expression Vocabulary (eVOC) – SANBI, University of the Western Cape

Mining PubMed

Recent national developments relevant to Bioinformatics/Computational Biology in South Africa

The National Bioinformatics Network

7 nodes – UWC, UCT, Wits, Stellenbosch, UP, Rhodes, UKZN

Note: SANBI is a node of NBN

Perspectives

How do cells work?

modeling, virtual cell etc.

What kinds of interventions are possible?

medicine, biotechnology

How do cells process information?

e.g. stomatal opening, ecological modeling

How do individuals within a population differ?

individualized medicine, molecular archaeology, conservation genetics, complex traits

What makes organisms different?

comparative genomics etc.

How did organisms come to be the way they are?

molecular evolution, selection

How do organisms interact?

systems biology of several organisms, immunoinformatics, ecological modeling

What causes an organism to function incorrectly

e.g. mapping genetic disorders

Data

- Complete genomes (*many!*)
- Expression information (*especially microarray*)
- Proteomics
- Protein-protein interactions
- Metabolic networks
- Regulatory relationships (*transcription factors and binding sites*)
- Quantitative data for modeling (*e.g. HIV CD4 counts & viral loads*)
- Polymorphism data (*dbSNP*, *HapMap*)

Acknowledgements

Konrad Scheffler

NBN