Molecular Evolution, Genomic Analysis and FreeBSD

Joseph Mingrone

Department of Mathematics and Statistics
Dalhousie University

About the Bielawski Group

- Small group (5 - 10 members)
- Molecular Evolution and Genomic Analysis
  - We analyze DNA sequences to make inferences
  - The Research is purely computational
- Multidisciplinary

Outline

- Hardware
- Research tracks
  - Modelling evolution at the molecular level
  - MircroBiome / Metagenomics
- Software, Design decisions, Observations

Hardware

Computing Cluster: Awarnach

- Purchased from Sun in 2006
- Sun Fire V40z master node (two dual core Opteron 870, 16 GB ECC RAM)
- Twenty compute nodes (X4100), two dual core Opteron 270, 4 GB ECC RAM, two 73 GB disks and four gigabit ethernet ports
- 48-port gigabit SMC switch

Hardware

Storage

- Asus RS300-E7-PS4 1U Server
- E3-1230V2 Xeon CPU
- Four Intel 60GB SSD (520 Series)
- LSI 9205-8e SAS controller
- Supermicro SC847E16-RJB0D1 JBOD with 10 WD30EFRX 2TB Hard Drives (ZFS raidz)

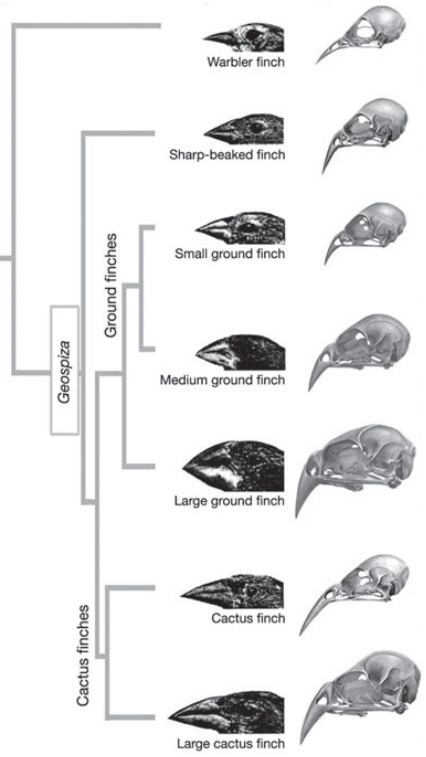
Hardware

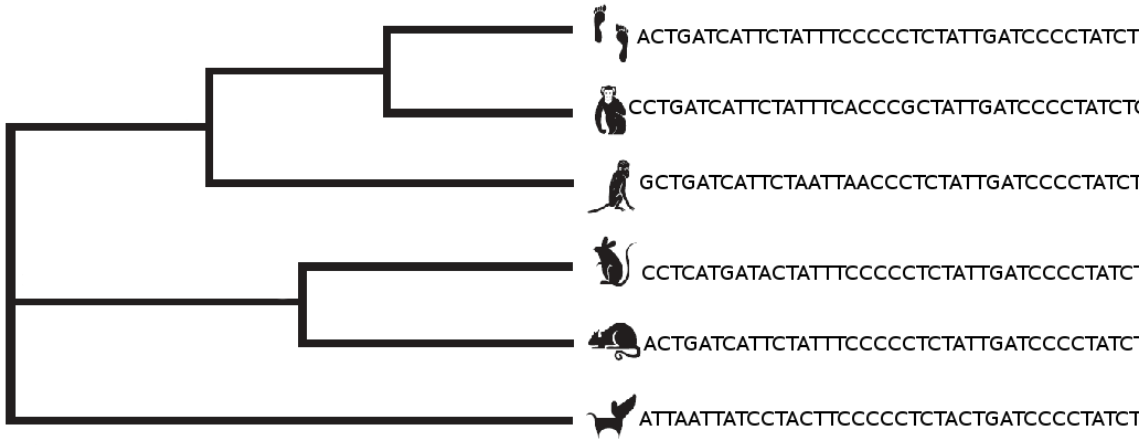
New Compute Node

- New Compute Node
- Four 12-core 6348 CPUs
- 256 GB RAM

Track 1: Background

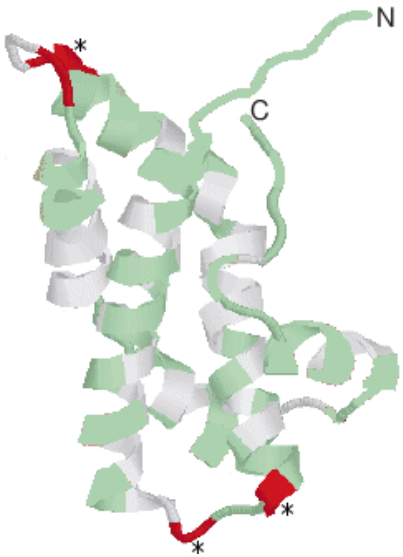
- Studying evolution is studying the past
  - Clues in present day to infer past processes
  - Morphology from fossil record
  - Studies with some organisms
  - Genetic material contains many markers of events in evolutionary history





Track 1: Molecular Evolution Modelling

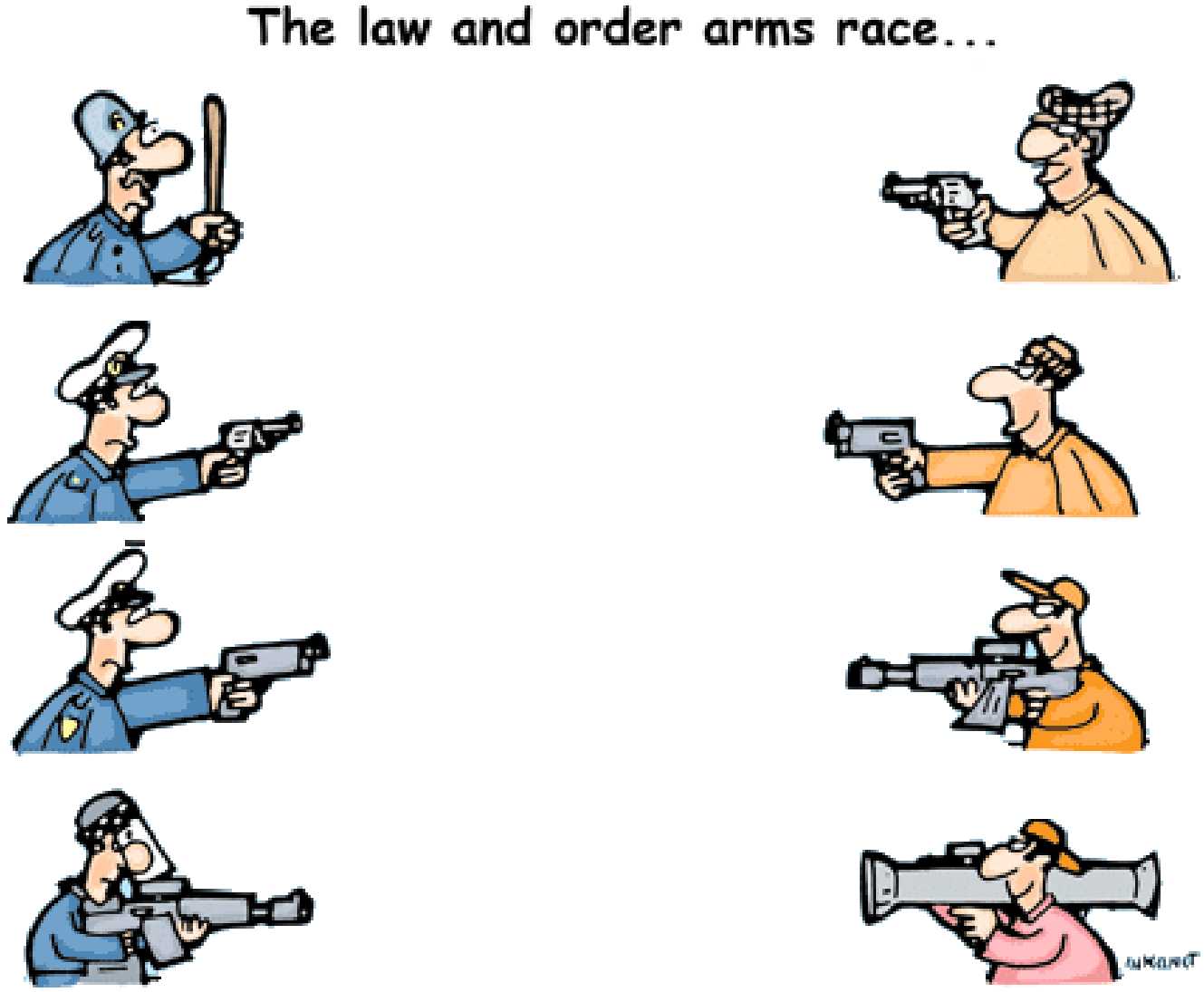
- Classify selection pressure at the protein and amino acid level



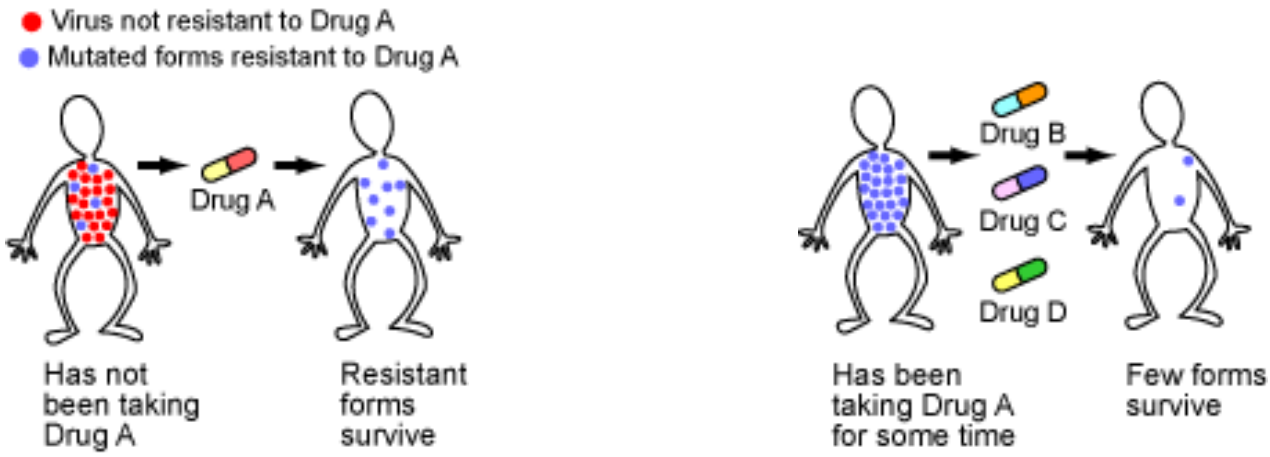
  - Purifying Selection
  - Neutral Evolution
  - Positive Selection

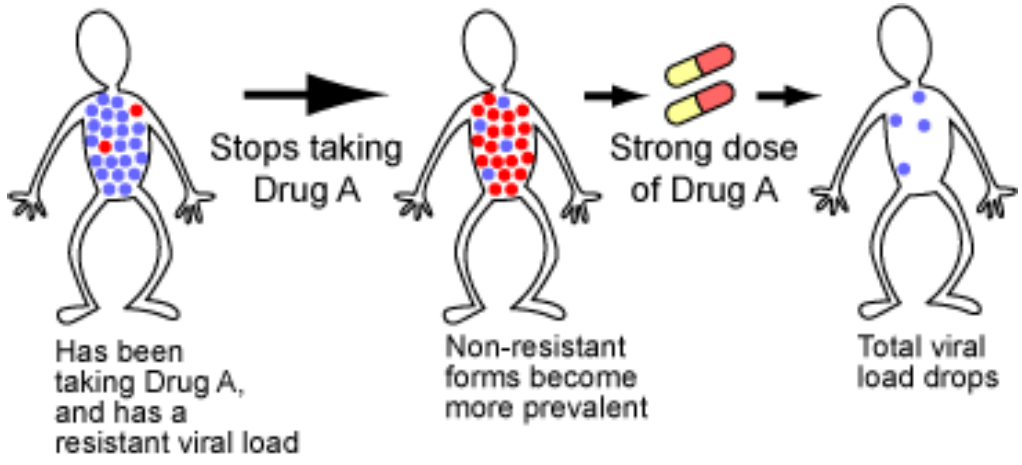
- Usually most interested in detecting sites under positive selection

Track 1: Arms Race

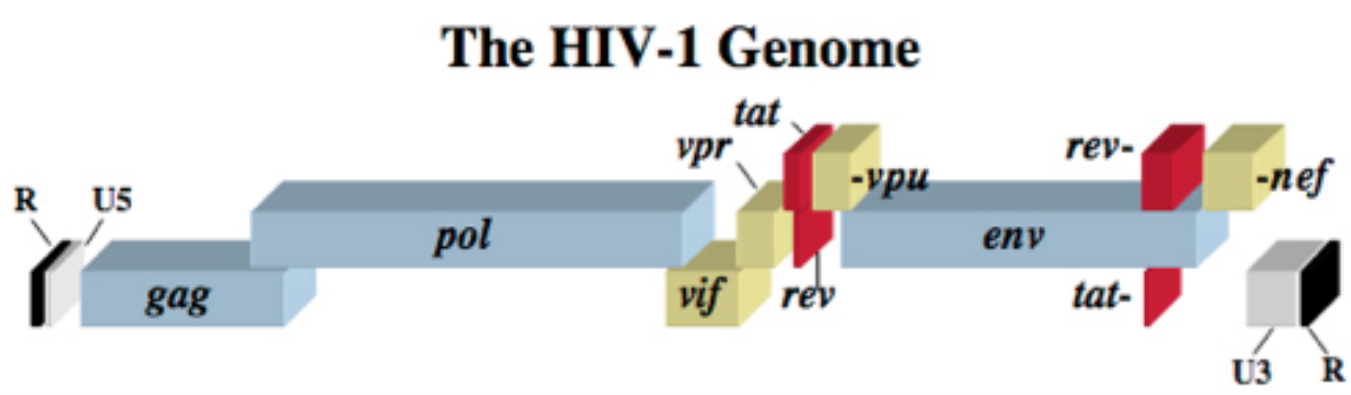


Track 1: Arms Race





Track 1: Analysis of HIV Genes

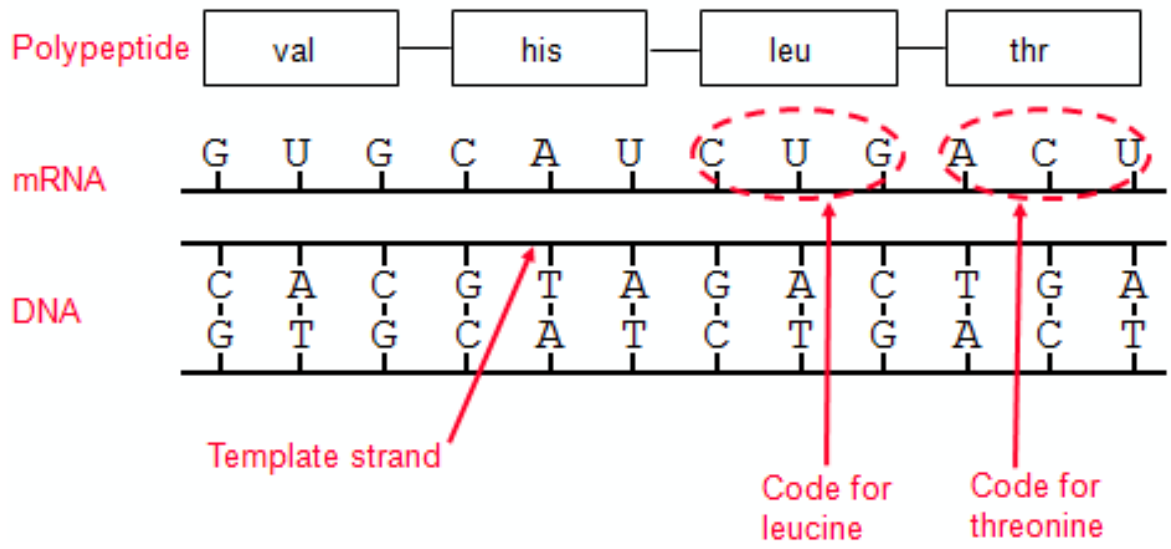


- Viral Envelop (Env) 3/91
- DNA polymerase (Pol) 11/947
- Viral infectivity factor (Vif) 10/192

Track 1: Background

The Central Dogma of Biology

- DNA is made up of four nucleotides with four different bases: Adenine (A), Cytosine (C), Guanine (G), or Thymine (T)
- Codons specify amino acids, the building blocks of protein



Track 1: Redundant Code

	T	C	A	G
T	TTT phe	TCT ser	TAT tyr	TGT cys
	TTC phe	TCC ser	TAC tyr	TGC cys
	TTA leu	TCA ser	TAA stop	TGA stop
	TTG leu	TCG ser	TAG stop	TGG trp
C	CTT leu	CCT pro	CAT his	CGT arg
	CTC leu	CCC pro	CAC his	CGC arg
	CTA leu	CCA pro	CAA gln	CGA arg
	CTG leu	CCG pro	CAG gln	CGG arg
A	ATT ile	ACT thr	AAT asn	AGT ser
	ATC ile	ACC thr	AAC asn	AGC ser
	ATA ile	ACA thr	AAA lys	AGA arg
	ATG met	ACG thr	AAG lys	AGG arg
G	GTT val	GCT ala	GAT asp	GGT gly
	GTC val	GCC ala	GAC asp	GGC gly
	GTA val	GCA ala	GAA glu	GGA gly
	GTG val	GCG ala	GAG glu	GGG gly

- $4^3 = 64$ possible codons (61 sense), 20 amino acids
- Code is redundant
- Nucleotide substitution may or may not mean change in amino acid

Track 1: A Measure of Selection Pressure

d_S = rate of synonymous substitutions

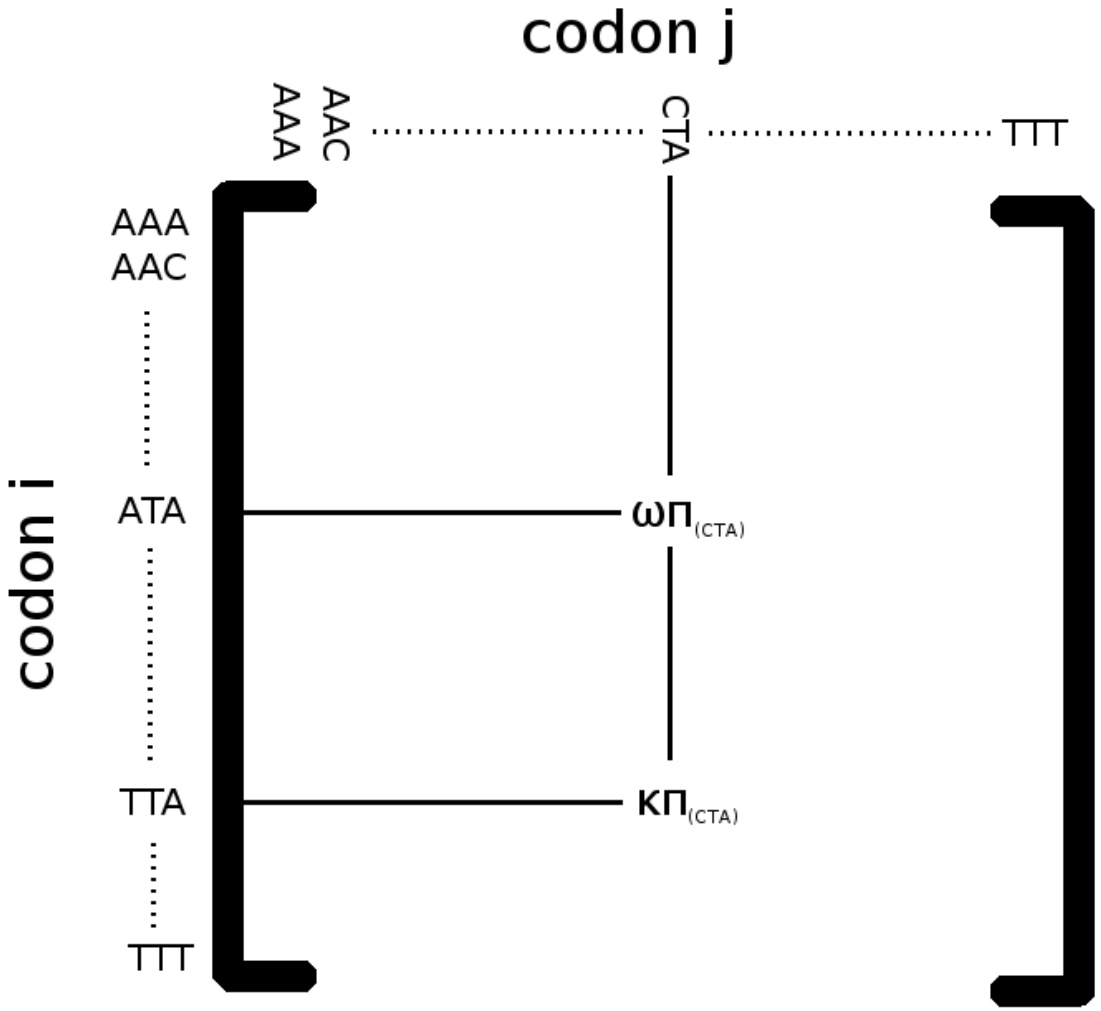
d_N = rate of nonsynonymous substitutions

$d_N/d_S = 1$	Neutral Evolution
$d_N/d_S < 1$	Purifying Selection
$d_N/d_S > 1$	Positive Selection

$$d_N/d_S = \omega$$

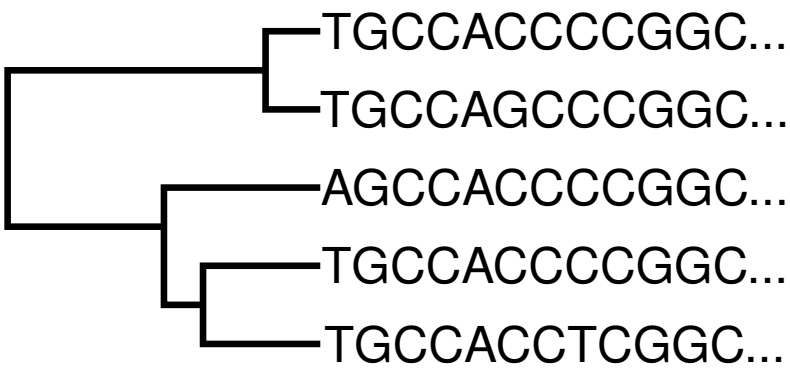
A measure of the strength and direction of selection pressure

Track 1: Markov Process



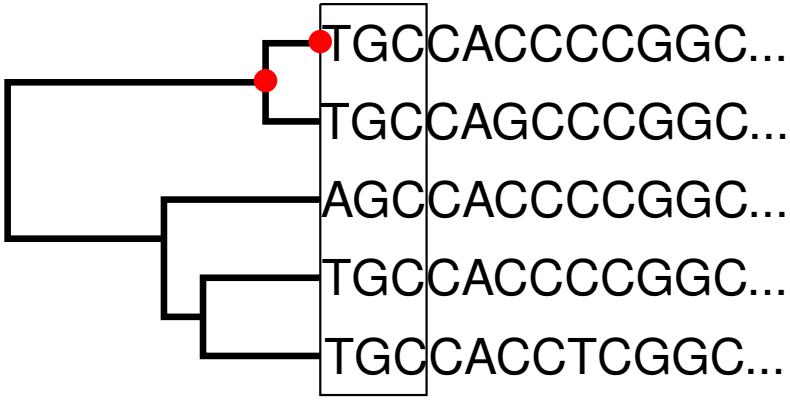
$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

Track 1: Estimating Model Parameters is Computationally Intense



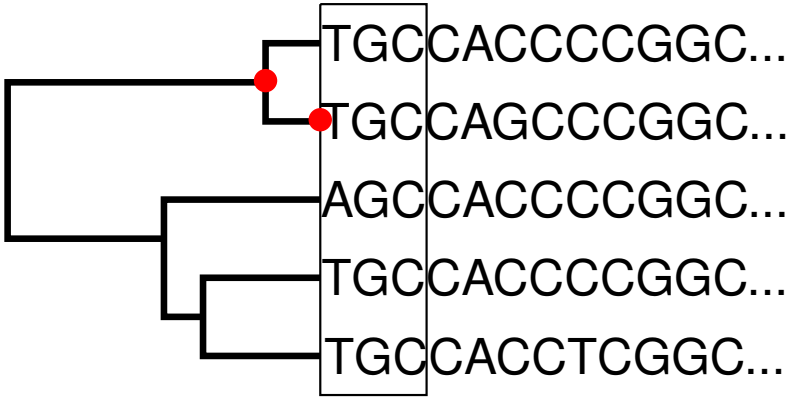
$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

Track 1: Estimating Model Parameters is Computationally Intense



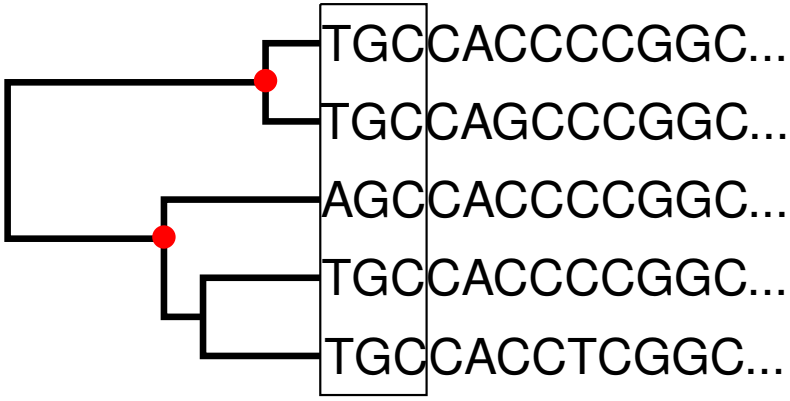
$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

Track 1: Estimating Model Parameters is Computationally Intense



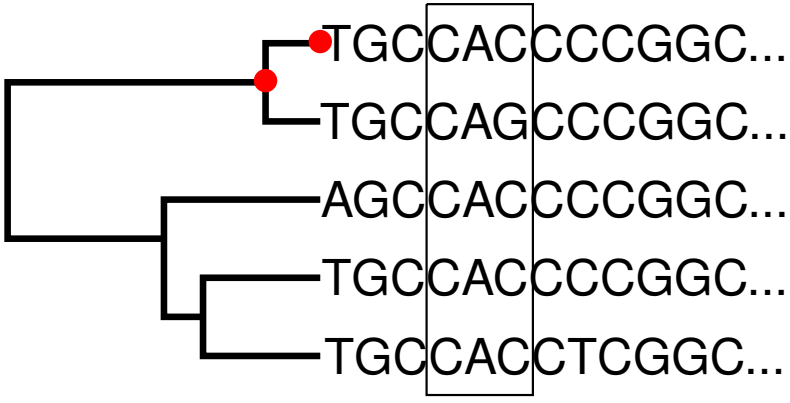
$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

Track 1: Estimating Model Parameters is Computationally Intense



$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

Track 1: Estimating Model Parameters is Computationally Intense



$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

Track 2: MircroBiome / Metagenomics

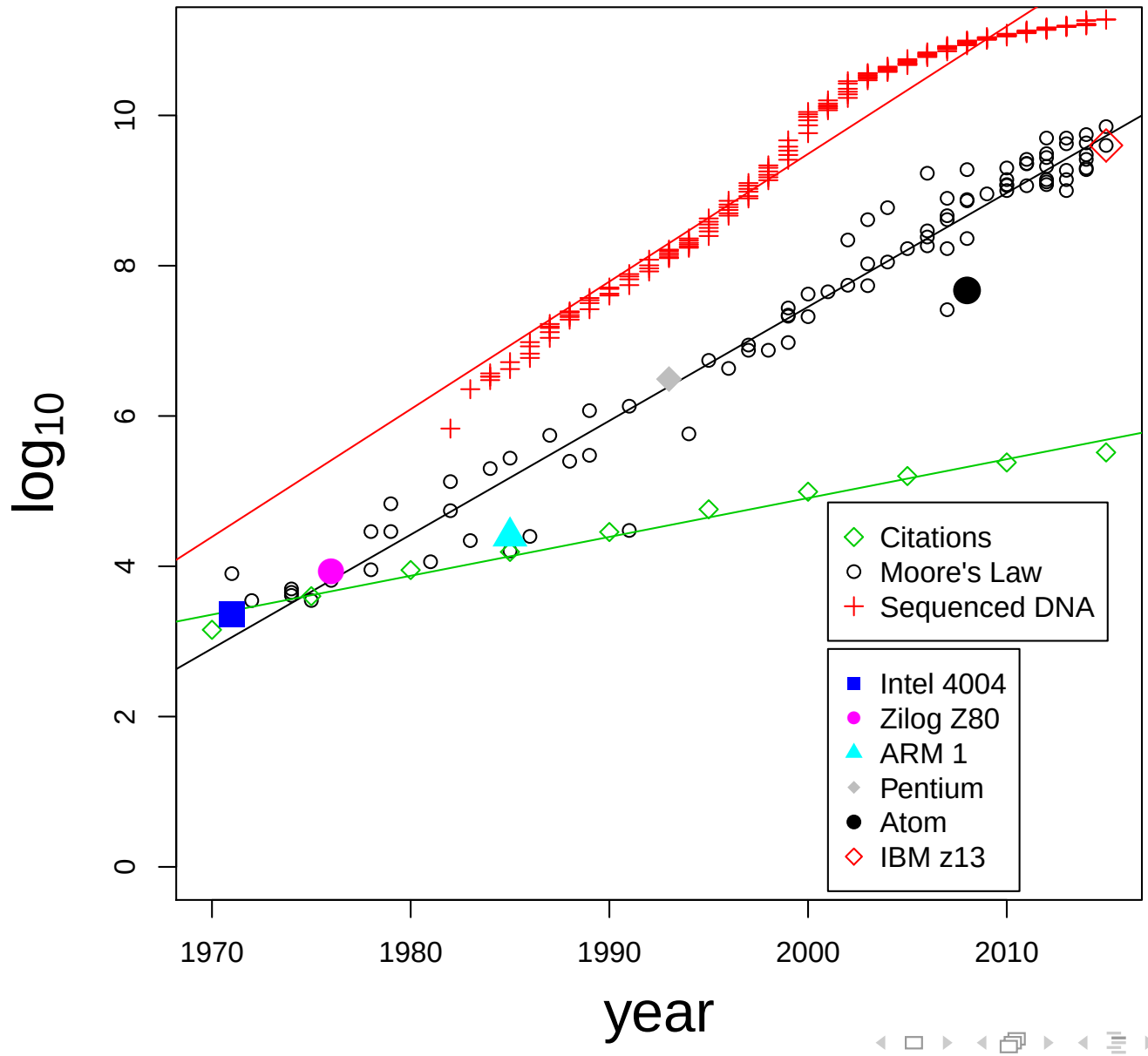
- Culturing diverse microbial communities in the lab can be difficult
- Sequence DNA as it exists in the natural environment (metagenomics)
- Understand associations between changes in microbiomes and changes in complex systems
  - BiomeNET BioMiCo: group microbial communities according to properties

El-Swais, Heba, et al. “Seasonal assemblages and short-lived blooms in coastal north-west Atlantic Ocean bacterioplankton.” Environmental microbiology (2014).

Shafiei, Mahdi, et al. “BioMiCo: a supervised Bayesian model for inference of microbial community structure.” Microbiome 3.1 (2015): 1-15.

Shafiei, Mahdi, et al. “BiomeNet: A Bayesian model for inference of metabolic divergence among microbial communities.” (2014): e1003918.

Research and Computing Technology are Coupled



Software

- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node

Software

- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node
- Poudriere, VirtualBox (MatLab)

Software

- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node
- Poudriere, VirtualBox (MatLab)
- ZFS (master node and storage server), NFS

Software

- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node
- Poudriere, VirtualBox (MatLab)
- ZFS (master node and storage server), NFS
- math/R, lang/perl5.20

Software

- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node
- Poudriere, VirtualBox (MatLab)
- ZFS (master node and storage server), NFS
- math/R, lang/perl5.20
- print/texlive-full

Software

- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node
- Poudriere, VirtualBox (MatLab)
- ZFS (master node and storage server), NFS
- math/R, lang/perl5.20
- print/texlive-full
- biology/paml / Proteus

Software

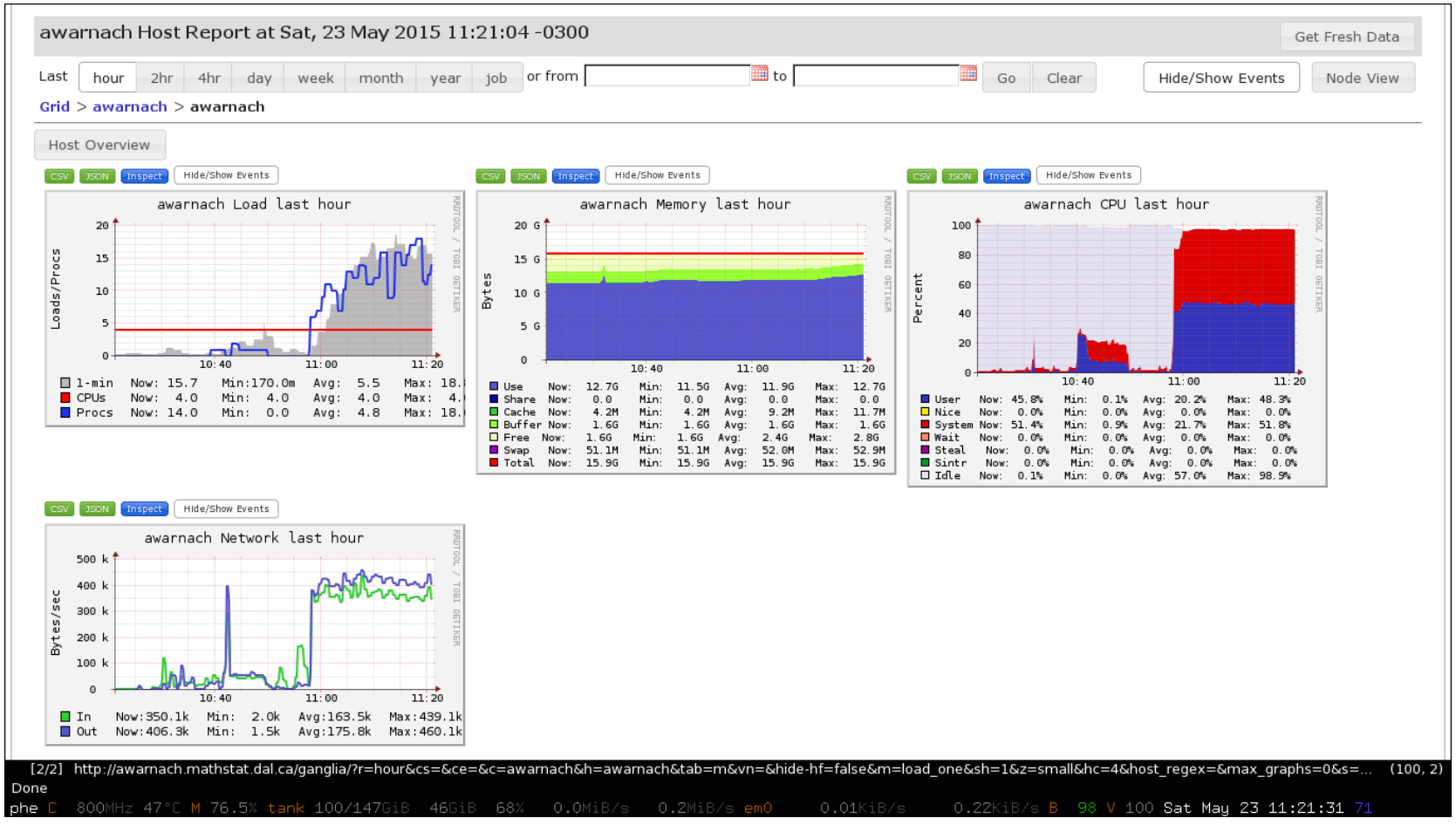
- Originally Solaris, since 2009 FreeBSD release (7.x - FreeBSD 10.1), STABLE on storage / new compute node
- Poudriere, VirtualBox (MatLab)
- ZFS (master node and storage server), NFS
- math/R, lang/perl5.20
- print/texlive-full
- biology/paml / Proteus
- biology/ncbi-blast+, Diamond, Humann, Qiime, Metaphlan, Biomenet, BioMico, BMTagger, PEAR

Software

security/tmux-cssh

Software

sysutils/ganglia-monitor-core sysutils/ganglia-webfrontend



Software

sysutils/ganglia-monitor-core sysutils/ganglia-webfrontend

Software

Basic Unix Tools

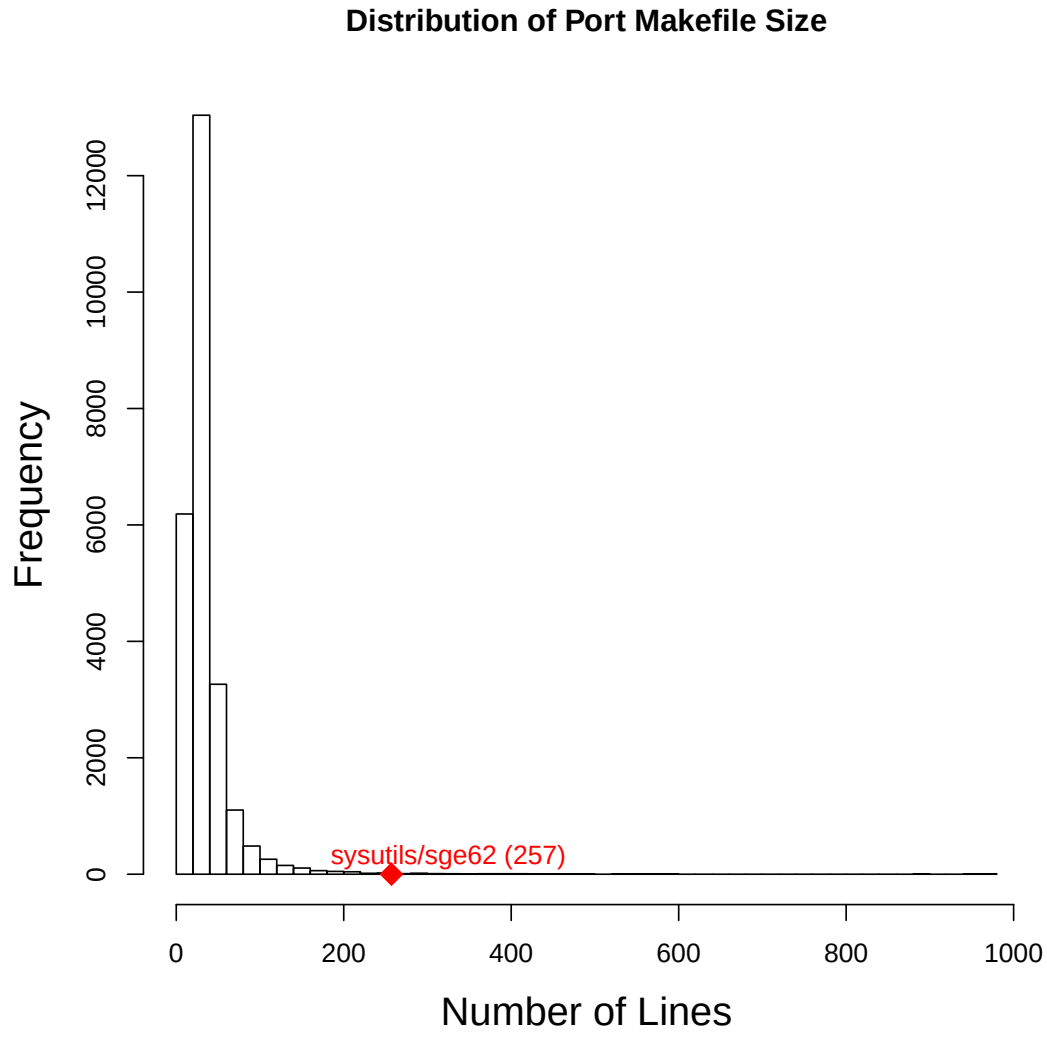
```
#!/bin/sh

fping=/usr/local/sbin/fping
nodes="aw1 aw2 aw3 aw4 aw5 aw6 aw7 aw8 aw9 aw10 aw11 aw12 aw13 aw14 aw15 aw16 aw18 aw19 aw20"
rdo_user=root

for host in `${fping} -a ${nodes}`; do
        printf "*** ${host} ***\n"
        ssh ${rdo_user}@${host} $*
done
```

Software

Batch Submission, Resource Management with Sun Grid Engine

Software

Batch Submission, Resource Management with Sun Grid Engine



Software

sysutils/slurm-hpc (slurm-wlm)

- Despite the name, Simple Linux Utility for Resource Management (Slurm): “Portability: Written in C with a GNU autoconf configuration engine. While initially written for Linux, Slurm has been ported to a diverse assortment of systems.”

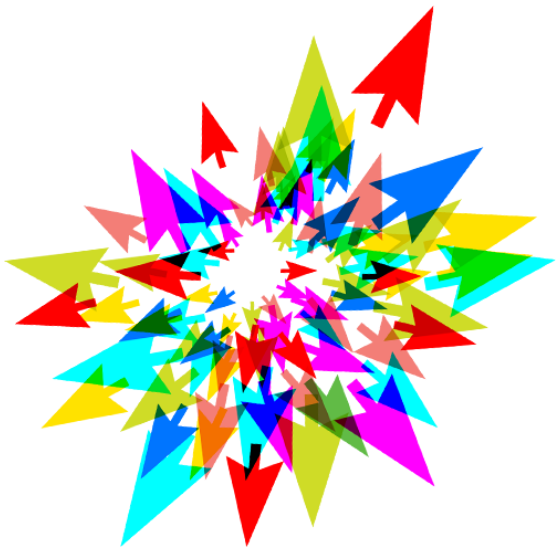
- “Sequoia, an IBM BlueGene/Q system at Lawrence Livermore National Laboratory with 1.6 petabytes of memory, 96 racks, 98,304 compute nodes, and 1.6 million cores, with a peak performance of over 17.17 Petaflops.”

Software

Porting Software Written by Biologists

Other Cluster Options

compute canada | calcul canada



ACENET

accelerate discovery

Thank You

Questions?

Image credits:
Arms race: www.inkcintc.com.au

HIV images: http://evolution.berkeley.edu/evolibrary/article/medicine_04