Phylogenetics

Tanmoy Bhattacharya

Los Alamos National Laboratory

4^{th} qBio Summer School

July 28, 2010

outline

1. Inference in historical sciences.
2. State versus Process.
3. Importance of Time Scales.
4. Maximum Likelihood and Bayesian.
5. Information Criteria.
6. Example from HIV:
   - Getting the tree.
   - Using the tree: timing.
   - Using the tree: population history.
   - Using the tree: counting.
   - Practical uses.
7. Conclusions.

Science and repeatability

These steps must be repeatable in order to dependably predict any future results. (Wikipaedia)

The Book of Optics by Ibn al-Haytham (1021): conscious reliance upon *repeated* observations to infer *regularities*.

Repeated observations can come from:

- Performing controlled experiments, or
- Selecting from a stream of data.

Small correlation structure leads to *independent* observations.

History may have long correlations: not independent.

- Cosmic Variance: the problem of one universe.
- Large planets have lower density: rule?
- Widespread language similarities: cognitive structure?
- Four limbs: adaptation?
- Similarity across religions: common truth?

Galton's Problem

Ought we ... to begin by discussing each separate species—man, lion, ox, and the like—taking each kind in hand independently of the rest ... (De partibus Animalium)

In 1889 Sir Edward Tylor presented a paper on correlations between marital systems and societal complexity.

Sir Francis Galton pointed out confounding by borrowing and common descent.

General problem of dealing with autocorrelation called Galton's problem by Raoul Naroll in 1961.

If the cause of the correlation is known, one can reduce it by various methods: data selection, multiple regression, or lagging.

How do we know what is independent?

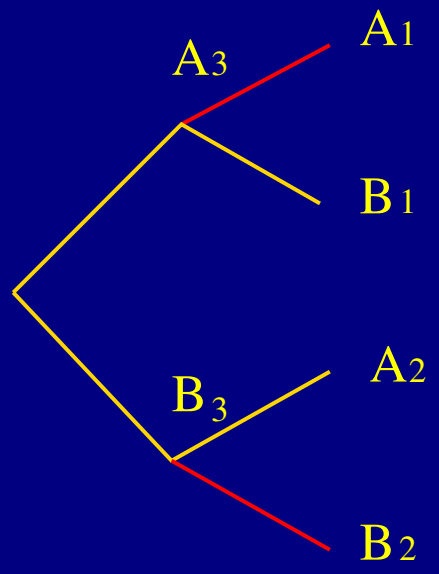
Synchronic vs. Diachronic

Problem is one of modeling: does synchronic data have enough information about

- Past situations?
- Process rules?

Often, causal structure restricted to a simply connected directed network (tree).

The independence structure of the tree allows us to look at multiple independent realizations.



When tree can be reconstructed, it allows us to deduce diachronic processes from synchronic data alone up to an unknown time scale.

Highly constrained prblem: $n(n-1)/2$ distances among n taxa has only $2n-2$ independent parameters.

State v Prcess

‘State’ is

- a statistic of the past;
- sufficient to predict the future.

Changes an integro-differential equation into differential equations:
Future State = Change Rule (Past State)
Change Rule is process.
Time scale separation: State changes faster than Process.
State variables few compared to history:

- Newton’s laws: only position, velocity, and environment.
- State of society: Institutions, norms, knowledge, and myth.
- Linguistics: Language as spoken.

Not always obvious:

- Temperature of floating bodies for motion.
- Remembered poetry on language.
- Family traditions in societal change.

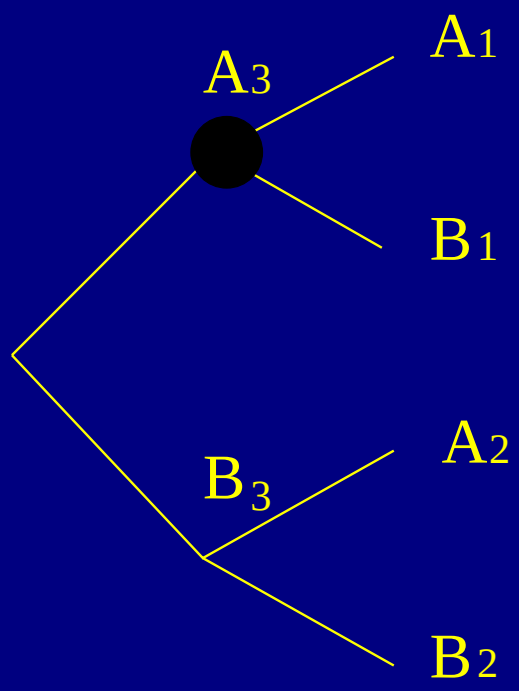
Branching Processes

Branching processes have simple causal structure.

Past and present conditionally independent given the state.

State at a branch point ('node') splits history into three conditionally independent sectors.

Almost stationary process: the probabilistic rules of change are constant.
Determine the process by fitting to the data.

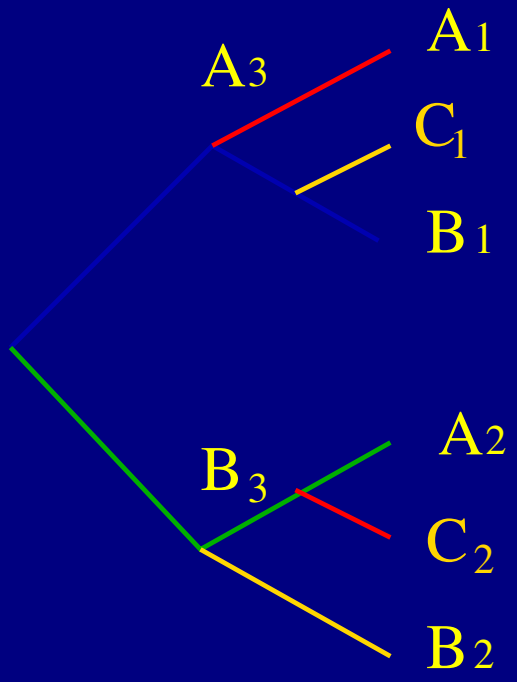


Timescales

Example:

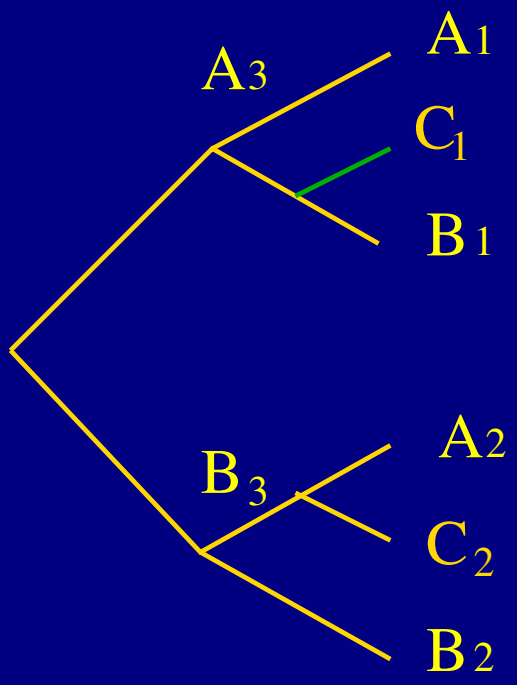
- Large number of almost independent traits.
- Varying at different rates.
- Rates constant, though different for different traits.

Fast Traits



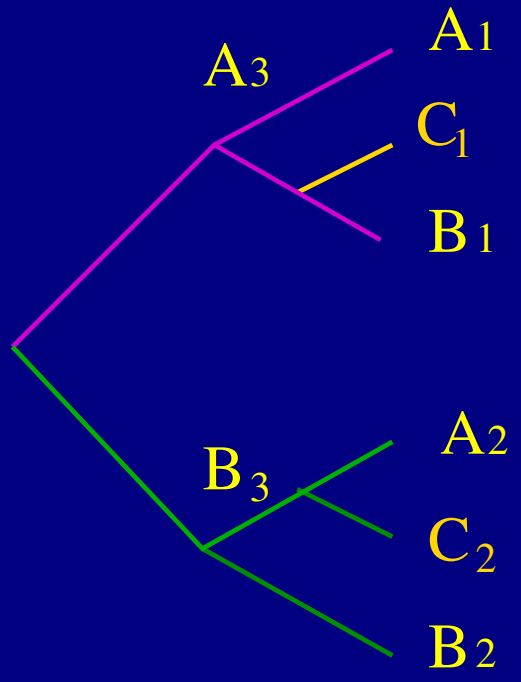
Almost random patterns. No correlations.

Slow Traits



Essentially constant. Replicated initial conditions.

Informative Traits



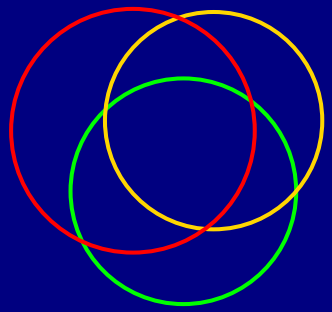
Intermediate: Traits that change about once on the timescale

- partition the data
- consistent with a tree
- no correlations except tree concordance

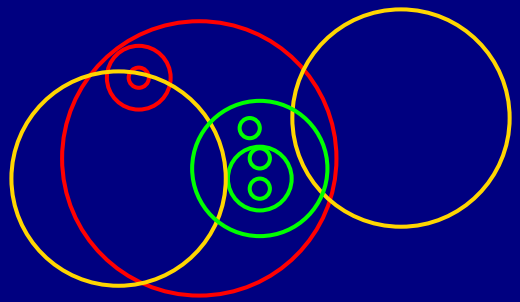
Excess of traits partitioned consistent with the same tree.

Hierarchical structuring of non-coextensive traits.

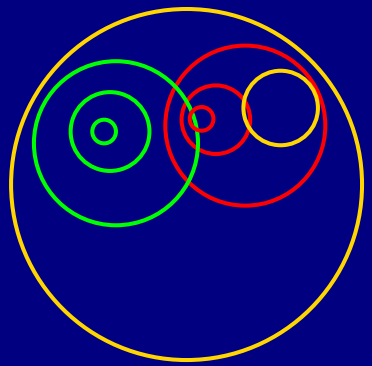
Hierarchical Structure



Fast Random Traits



Hierarchical Structure



Implicational scaling

Stochastic Processes

A stochastic process described by a probability law on states. Usually taken to be infinitely divisible

$$\mathrm{p}(S_0 \to S_t; t) = \sum_{S_{t'}} \mathrm{p}(S_0 \to S_{t'}; t')\ p(S_{t'} \to S_t; t - t')$$

Leads to a differential formulation:

$$\frac{\mathrm{dp}(S;t)}{\mathrm{d}t} = \sum_{S'} T(S, S'; t)\ \mathrm{p}(S'; t)$$

which can be formally solved:

$$\mathrm{p}(S;t) = \sum_{S'} [\mathcal{T} \exp \int_{t_0}^{t} T(S, S'; t)\mathrm{d}t]\, p(S'; t_0)$$

This allows us to calculate the probability of any set of observations if we are given a model T and the tree that the data was generated under.

Forward process; Backward inference

Model is constant! How do we determine it?

- Choose model parameters.
- Forward: Either simulate or calculate expected observations.
- Evaluate how expected the actual observations are.
- Backward: Vary model and choose best.

How does this perform with large amount of independent data?
True Model M: Independent observations with prbability $\{p_i\}$.
Observation i seen with frequency $f_i \approx p_i$.

Let model M_j assign it probablities p_{ij}. Log Total probability

$$\ln \mathcal{L}(M_j) \equiv \ln p(\text{Data} \mid M_j) = \sum_i f_i \ln p_{ij} \; .$$

Maximize over j subject to $\sum_i p_{ij} = 1$: $p_{ij} = f_i$.

The model chosen by Maximum Likelihood is consistent.

Maximum Likelihood Method

Maximum Likelihood Estimate is efficient.

Cramér-Rao inequality: sample-to-sample fluctuation of an estimate b for parameter β bounded by:

$$\sigma_b^2 \geq \left(\frac{\partial^2 \ln \mathcal{L}}{\partial \beta^2} \right)^{-1} \left\langle \frac{\partial b}{\partial \beta} \right\rangle^2 .$$

Sample fluctuation of an estimate cannot be smaller than product of

- Sensitivity of likelihood to parameter.
- Dependence of estimate on parameter.

Equality often reached by Maximum Likelihood estimate.

Example

Toss a loaded coin many times: how do we determine probability of heads?

As the probability of heads increases, the fraction of heads increases. So, does the number of runs of heads.

Can use different features of the data, *e.g.,*

- Fraction of heads.
- Average length of runs of heads.

Maximum likelihood chooses fraction of heads, because it is most informative.

Similarly, no need to sort out non-informative fast and slow traits: Maximum Likelihood method weights each trait at its proper time depth.

Model misspecification

Likelihood based methods are ideal if

- There is enough data.
- Model in the specified class.

When model does not allow features of the data: one can get bizarre results.

Model: Bengali derived from the Vedic language of India.
We want to find how long back Vedic was spoken.
If we do not recognize that many modern words are actually from Persian, English, Portuguese, Dravidian, Austrasiatic, etc., we will get a very wrong answer.
But, if we allow a probability for random new word: we start getting reasonable results.
Rare forgotten processes can sometimes be replaced by uncorrelated random processes.

Likelihood Ratio

For Normal distribution, $-2\ln\mathcal{L} \equiv \chi^2 + \sum \ln(2\pi\sigma^2)$.

- Maximum Likelihood is a generalization of minimum χ^2.
- Provides confidence intervals.

Adding parameters gives better fit even at random.
Best fit models with δ more parameters: $2\Delta\ln\mathcal{L} \sim \chi^2_\delta$.

- Quantifies overfitting.

This can be generalized to Bayesian posterior

$$\mathrm{p}(M_i \mid \text{Data}) \propto \mathrm{p}(M_i)\mathcal{L}(M_i)\,.$$

- Can incorporate prior knowledge.
- Penalizes extra parameters more.
- Can be evaluated by a Markov Chain Monte Carlo.

Markov Chain Monte Carlo

How to sample a random distribution?
Replace ensemble average by time average. Choices:

1. Design a deterministic ergodic process.
2. Use Markov processes.

If there is a random Markov *process*, $p(X \to Y)$, such that

- $\frac{p(X \to Y)}{p(Y \to X)} = \frac{\pi(Y)}{\pi(X)}$,
- Every state is reachable in one or more moves,
- The process is not periodic, and
- The expected number of moves to return is finite,

Then, this ergodic process samples X in proportion to $\pi(X)$.
Much easier problem because $p(X \to Y)$ is *local*.

Example:

- Choose a small neighbourhood $\{X_j\}$ for each X_i.
- Choose $p(X_i \to X_j) = 1$ if $\pi(X_j) > \pi(X_i)$.
- Choose $p(X_i \to X_j) = \pi(X_i)/\pi(X_j)$ otherwise.
- Check the criteria.

Information Criteria

How do we decide when a richer model should be used?

Three problems with increasing parameters:

- More parameters make estimation more noisy.
- More parameter models more sensitive to noisy estimates.
- Many more multiparameter models to choose from: fairness?

Akaike Information Criterion: Choose number of parameters to maximize predictability.
Bayesian Information Criterion: Use a prior on number of parameters; minimize unassumed coincidences.

Akaike Information Criterion

Two parts:

- Parameter fits non-reproducible noise.
- Parameter mispredicts future observations.

Let,

- θ be the true model,
- X be some observations,
- Y be similar future observations.

Let θ_X be the model estimated from X: *i.e.,* the model that maximizes the likelihood $\mathcal{L}(\theta_X|X)$. But, what we really want is a model that assigns high probability $p(Y)$ to future observations.

Asymptotically, for a k-parameter model,

$$\mathrm{E}_X \log \mathcal{L}(\theta_X|X) = \mathrm{E}_X \log \mathcal{L}(\theta|X) + \frac{k}{2}\,,$$

because it fits some noise.

But, the noise is different on almost every dataset, so θ_X is worse than θ except on X.

In particular, asymptotically, on average

$$\mathrm{E}_Y \log \mathcal{L}(\theta|Y) = \mathrm{E}_Y \mathrm{E}_X \log \mathcal{L}(\theta_X|Y) + \frac{k}{2}\,.$$

Thus, on average,

$$\mathrm{E}_{X,Y} \log p(Y|\theta_X) = \mathrm{E}_X \log \mathcal{L}(\theta_X|X) - k\,.$$

For best predictability, maximize this, *i.e.,* minimize $\mathrm{AIC} \equiv -2 \log \mathcal{L} + 2k$.

Bayesian Information Criterion

Data can't prove a hypothesis: it can rule it out.

Bayesian prior: how much support from data do we need to rule out the hypothesis?

Uniform prior: Each hypothesis has equal *a priori* probability $= 1/\text{Number of hypothesis}$.

If lower parameter model a point in higher parameter space: zero *a priori* probability; needs *infinite* data to override!

Use distributions (Dirac δ functions): total probability of lower parameter model equal to total probability of higher parameter model.

This is equivalent to letting data resolution decide the ‘size’ of the lower dimensional point.

If data sample is of size n, error bars are size $1/\sqrt{n}$. So, a k lower-dimensional point has relative volume $\exp -\frac{k}{2} \log n$.

Equal probability means, lower dimensional model has correspondingly larger *a priori* weight compared to every point in the higher dimension: so, choosing a higher parameter model requires that much more evidence.

So, to maximize Bayesian posterior, minimize $\mathrm{BIC} \equiv -2 \log \mathcal{L} + k \log n$.

Since sample size arbitrary: being just able to rule out lower dimensional model is surprising. BIC requires that the fit be better in proportion to data.

Example: phylogeny in biology

Can one use these methods to infer history of life? Is history of life tree-like?

But history of what?

Traits are inherited

- From parents.
- From peers.
- From physical environment.
- From previous changes in environment . . . .

Look for a large co-inherited bundle of traits. Define this as 'vertical transmission.' Other inheritences referred to this baseline.

One such large bundle: genetic traits.

Genotype

Most of life has a strong genotype-phenotype separation.

- Genotype encodes heredity: phenotype is selectable.
- Genotype to phenotype program not easily invertible.
- Genotype changes mainly dictated by chemistry: almost stationary process.

Genotype changes randomly, weakly filtered by selection. Vast amount of almost independent traits, almost stationary process.

- Most of life close to fitness maximum.
- Robustness: Few changes fatal, most neutral.
- High mutation rates harmful.
  (Eigen's law: no more than 1 change/unit/generation.)
- Mutation rate maximized for adaptability.

Genetic Change

Most of the changes are ‘point mutations’:

GTAAGACAGTATGATCAGATACTCATAGAAATCTGTGGA $\rightarrow$
GTAAAACAATATGATCAGGTATCTATAGAAATTTGTGGA .

Some regions are prone to insertions or deletions.

AGTAATACTACTAGTAAT $\leftrightarrow$
ACT ATACTA AAT .

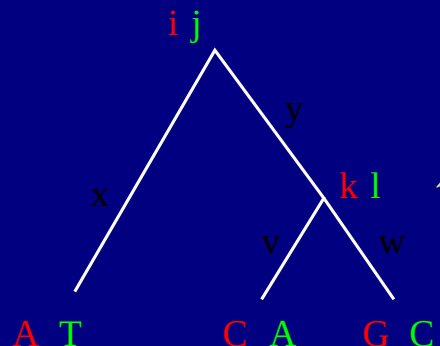
Daughter may form from parts of different parents

...AGGATGGAC...
...TTTATGCTG... $\rightarrow$ ...AGGATGCTG...

Stationary Independent Sites Model

Consider only point mutations and assume

- many sites change at different rates r_i,
- rates are almost constant over time,
- relative probabilities of base substitutions T the same, and
- sites in the genome change almost independently.



'Feynmann Diagram':

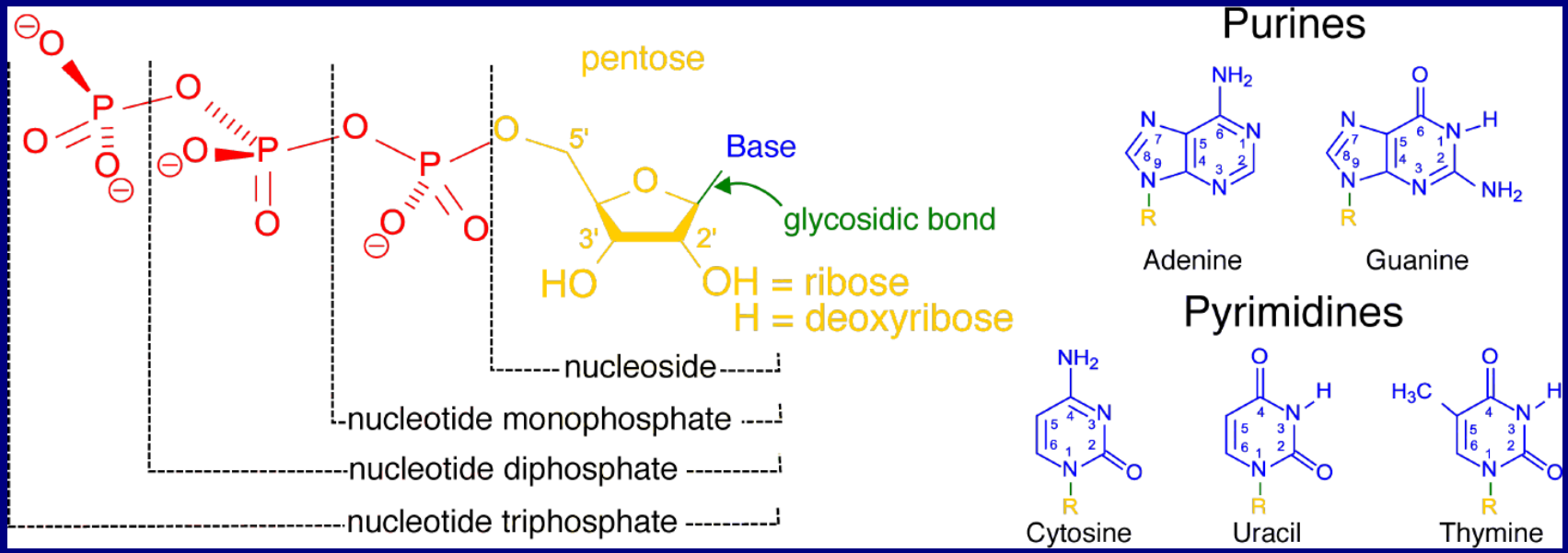
$$\mathcal{L} = \Big(\sum_{ik} \overline{p}_i (e^{r_1 x T})_{Ai} (e^{r_1 y T})_{ki} (e^{r_1 v T})_{Ck} (e^{r_1 w T})_{Gk}\Big) \times \Big(\sum_{jl} \overline{p}_j (e^{r_2 x T})_{Tj} (e^{r_2 y T})_{lj} (e^{r_2 v T})_{Al} (e^{r_2 w T})_{Cl}\Big),$$

where $\overline{p}$ are the initial probabilities.

- Propose reconstructions
- Evaluate reconstructions
- Find history and process that give best reconstructions.

Base substitution models

Base substitution often due to chemical ‘error’.
Some bases more similar than others:



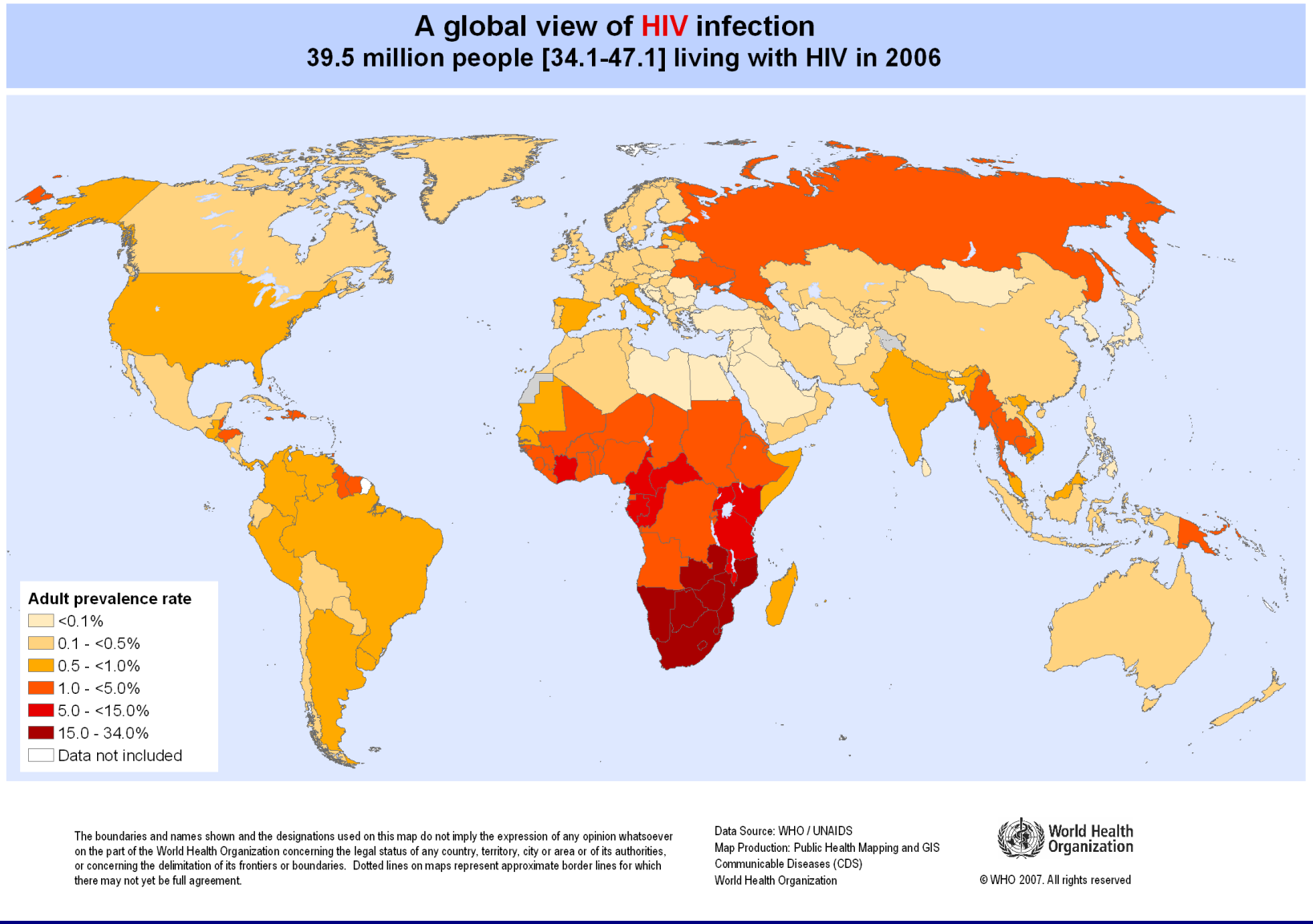
Wikimedia

Usually, $C \leftrightarrow T$ and $G \leftrightarrow A$ faster than the rest.

The calculation simplifies further if

- $\overline{p}$ is stationary distribution: $T\overline{p} = 0$, and
- $T_{ij}\overline{p}_j = T_{ji}\overline{p}_i \qquad \forall i, j.$

HIV: A worldwide pandemic



HIV: A social problem

Affects the sexually active (productive) age group.

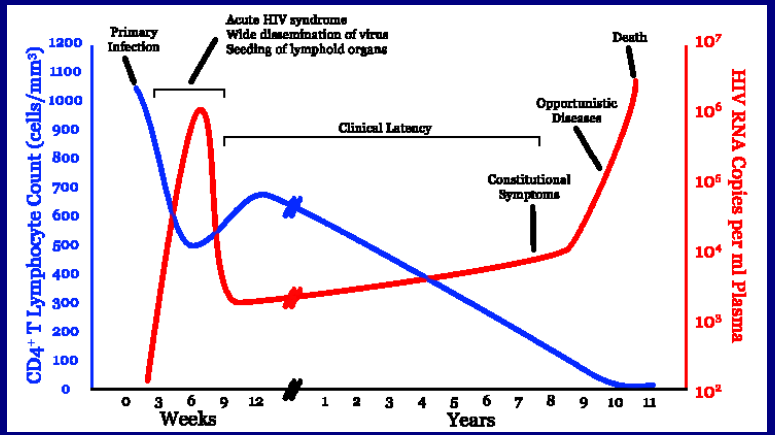
Destroys families: orphans and the aged.

Asymptomatic phase before AIDS: 9 years untreated.

Escapes single drugs within a couple of days. Escapes double combinations in years.

Prevention difficult

- lifestyle changes (*e.g.*, monogamy, condoms, circumcision)
- screening blood
- sterilizing needles
- expensive drugs

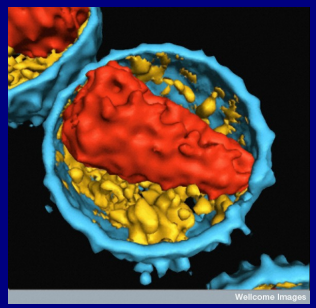


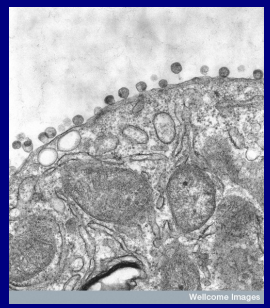
Medipedia

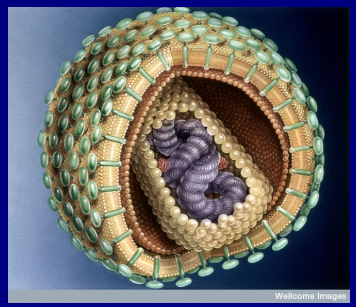
HIV: The Virus

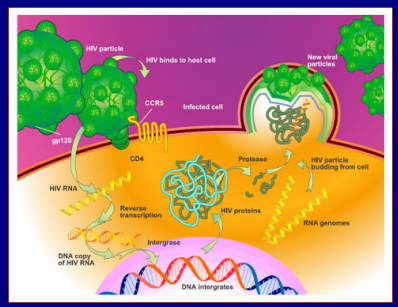
We know

the structure, the function, and the infection dynamics.





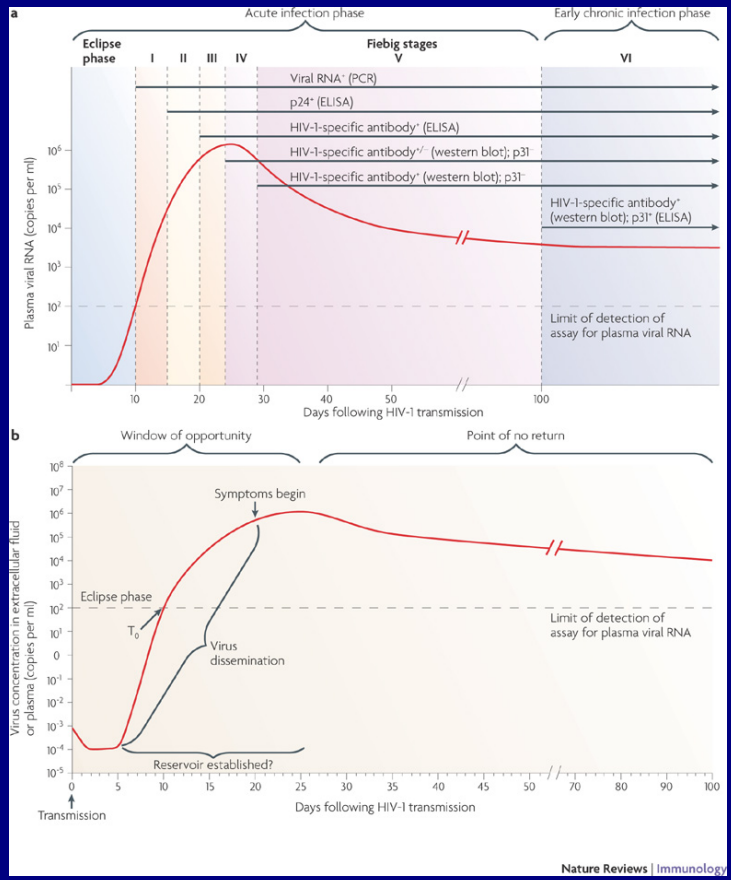




Adcock Ingram

the genetics,

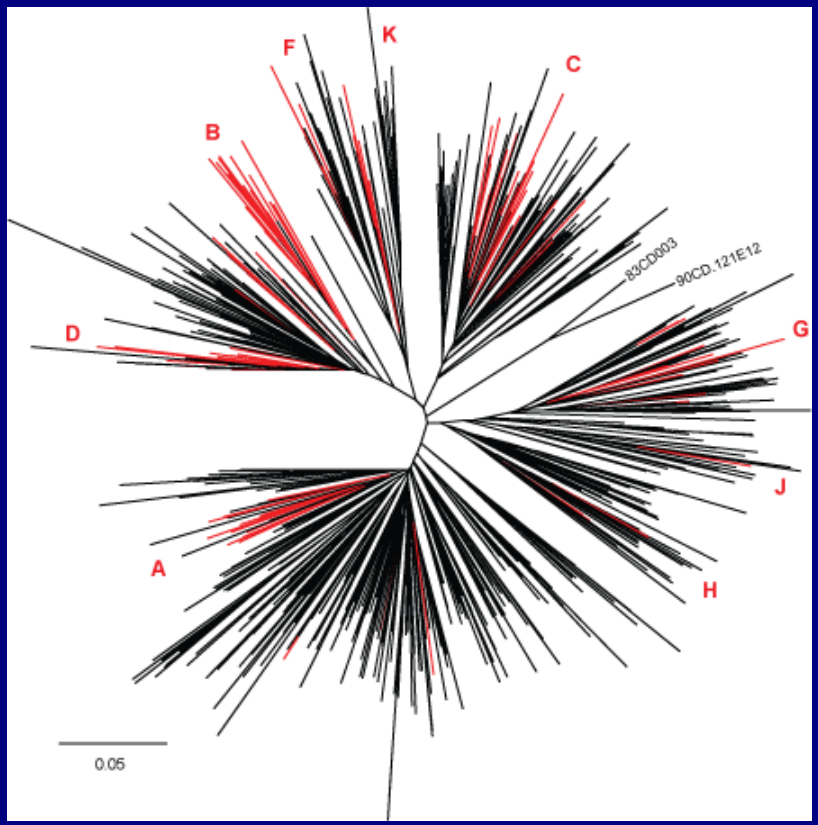
HIV Database



McMichael *et al.*,
Nat. Rev. Immun. 2010 Jan; 10(1):11–23.

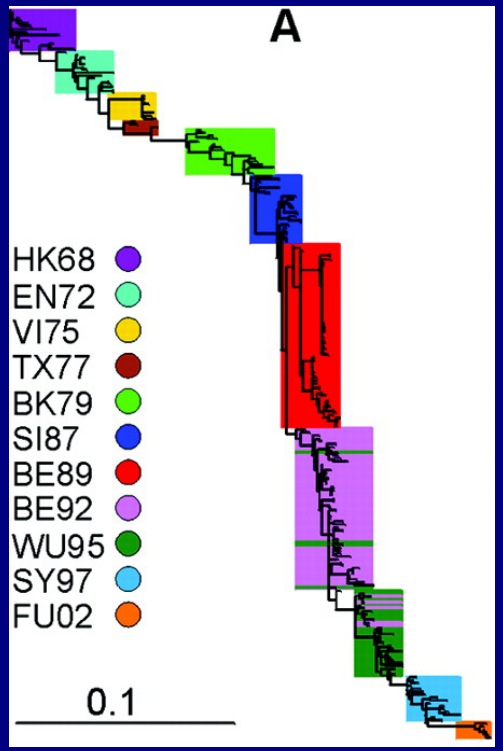
HIV: Extreme Diversity

No effective vaccine yet!



HIV

Archer and Robertson,
AIDS 2007 Aug 20; 21(13):1693–1700



Influenza

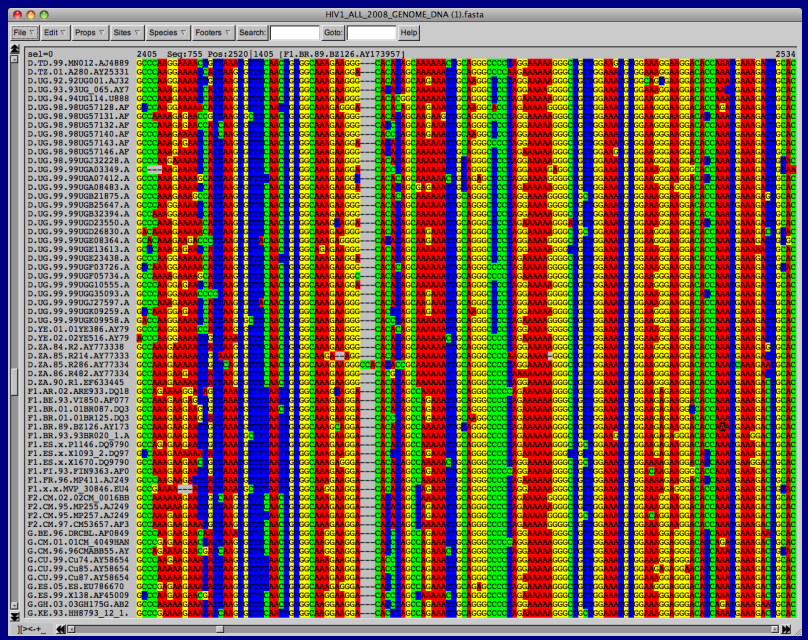
Smith *et al.*,
Science 2004 Jul 16: 305(5682):371–376

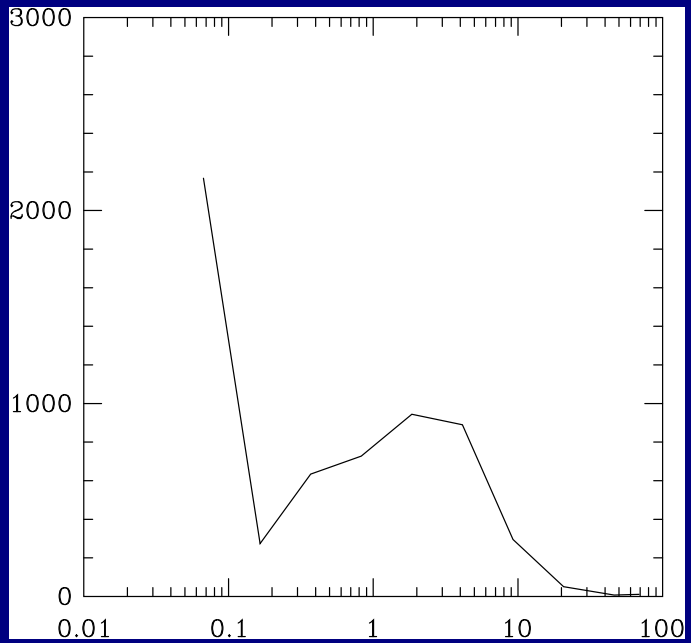
Need to *understand* biology and evolution.

HIV

HIV is a virus about 9719 bases long. More than 1200 almost complete sequences known.

The rates vary from site to site.

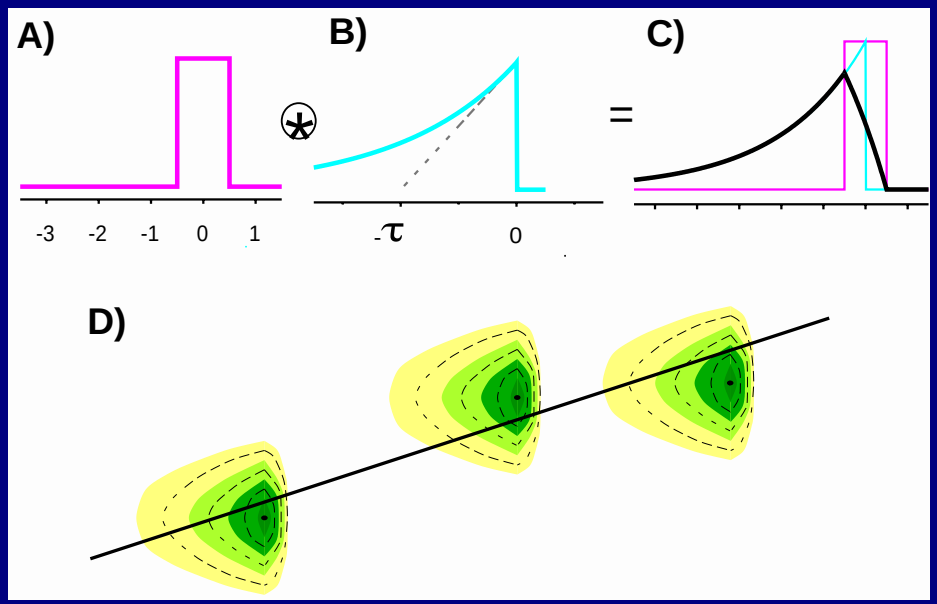




Different kinds of changes have different probabilities:
Transversion:Transition CT:Transition GA :: 1:3:4

Error Model

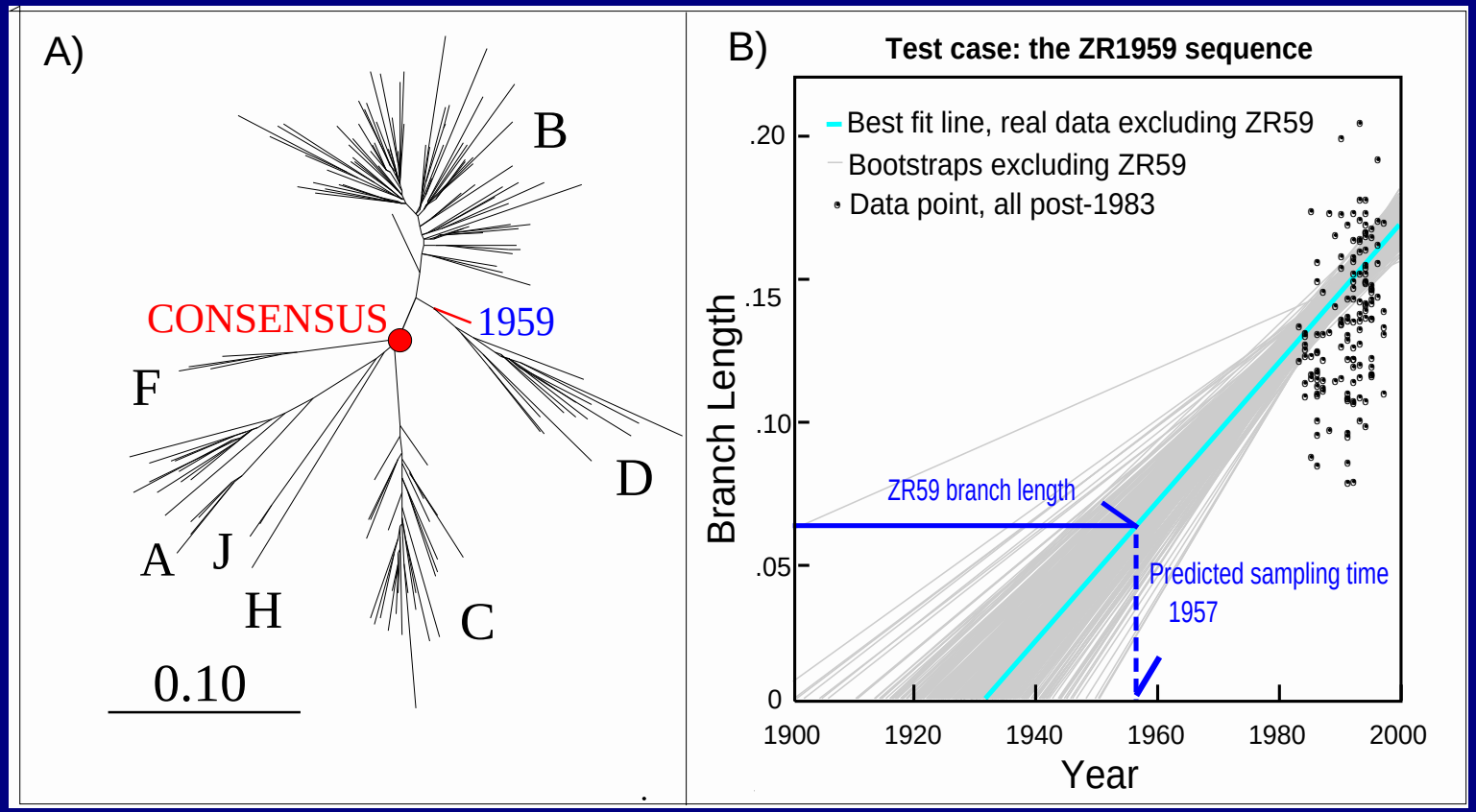
- Random substitution processes imply Poisson error on the branch lengths.
- Sampling time known only to one year.
- Virus integrates into host DNA, stops evolving and is expressed much later. Assume an exponential latency, $e^{-t/\tau}$, with unknown parameter τ.



- Use Maximum Likelihood to estimate τ and the best fit line.

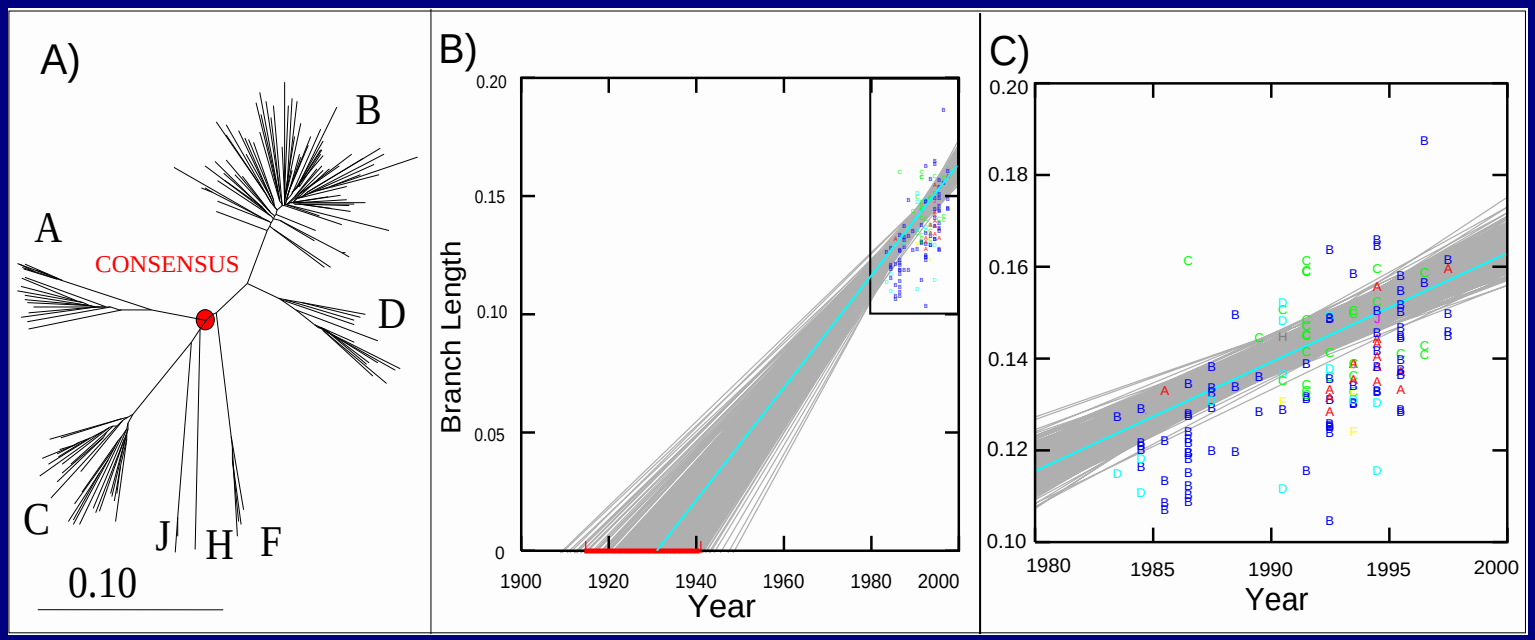
Model verification

This model correctly predicts the time of the earliest HIV sequence:



Origin of HIV

- Tested two genes:
  - env (gp160): 141 sequences of 2038 bases each.
  - gag: 64 sequences of 1363 bases each.
- Results consistent:
  - env: estimate 1931; 95% CI 1915 – 1941
  - gag: estimate 1934; 95% CI 1869 – 1950

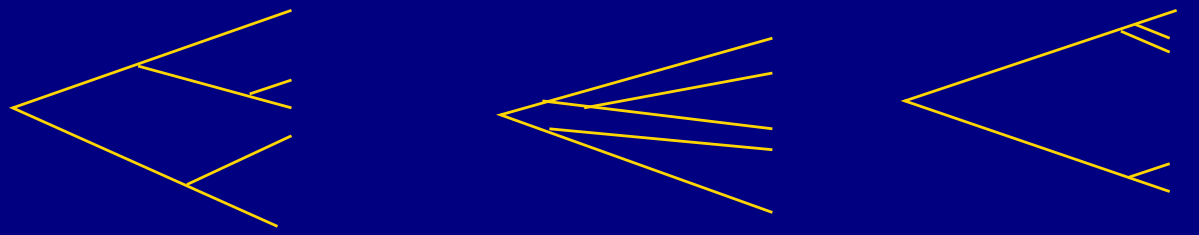


Current results at the earlier end.

Uses of phylogeny: Coalescent Theory

In a population of size N, two randomly selected strains had a common parent with probability $1/N$.

Phylogeny provides estimates of relative times of common ancestors.



Constant Growing Contracting

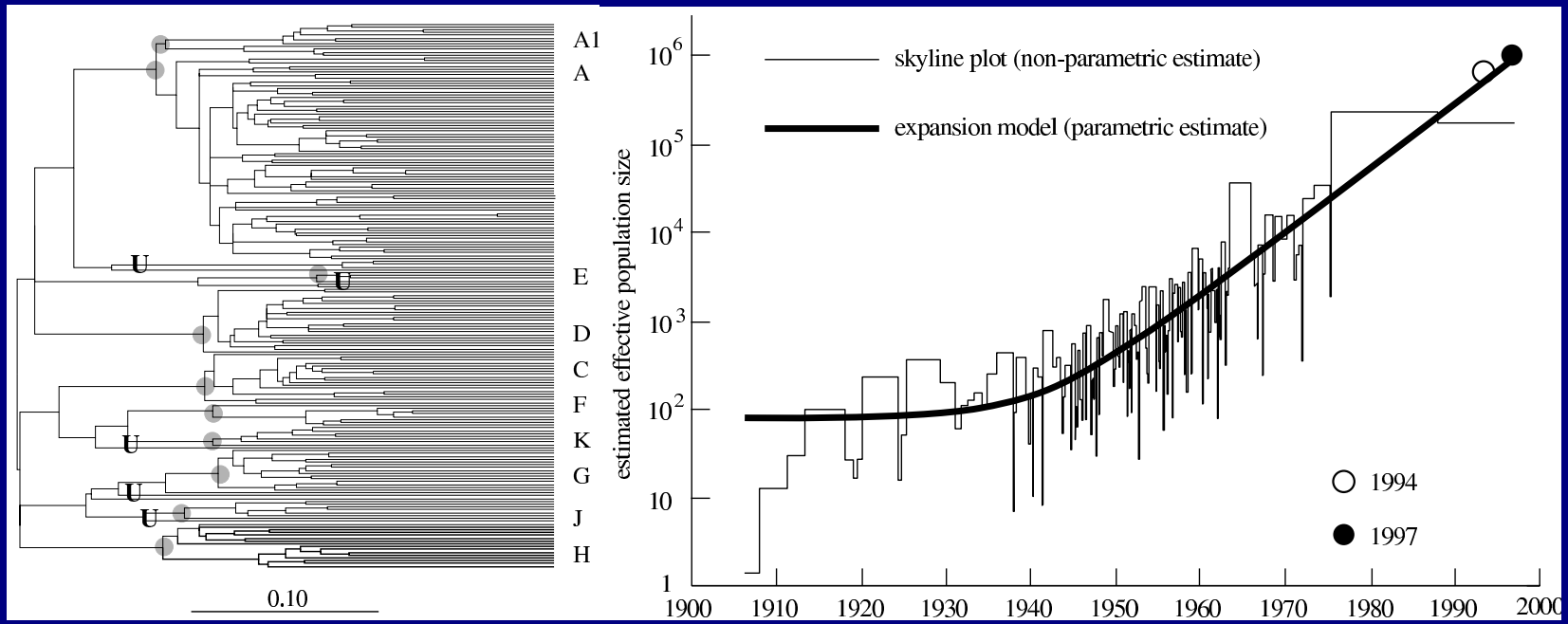
More ‘coalescenses’ when population size small.

In randomly mixed situations, can estimate populations in the past!

Example: Growth of HIV infections

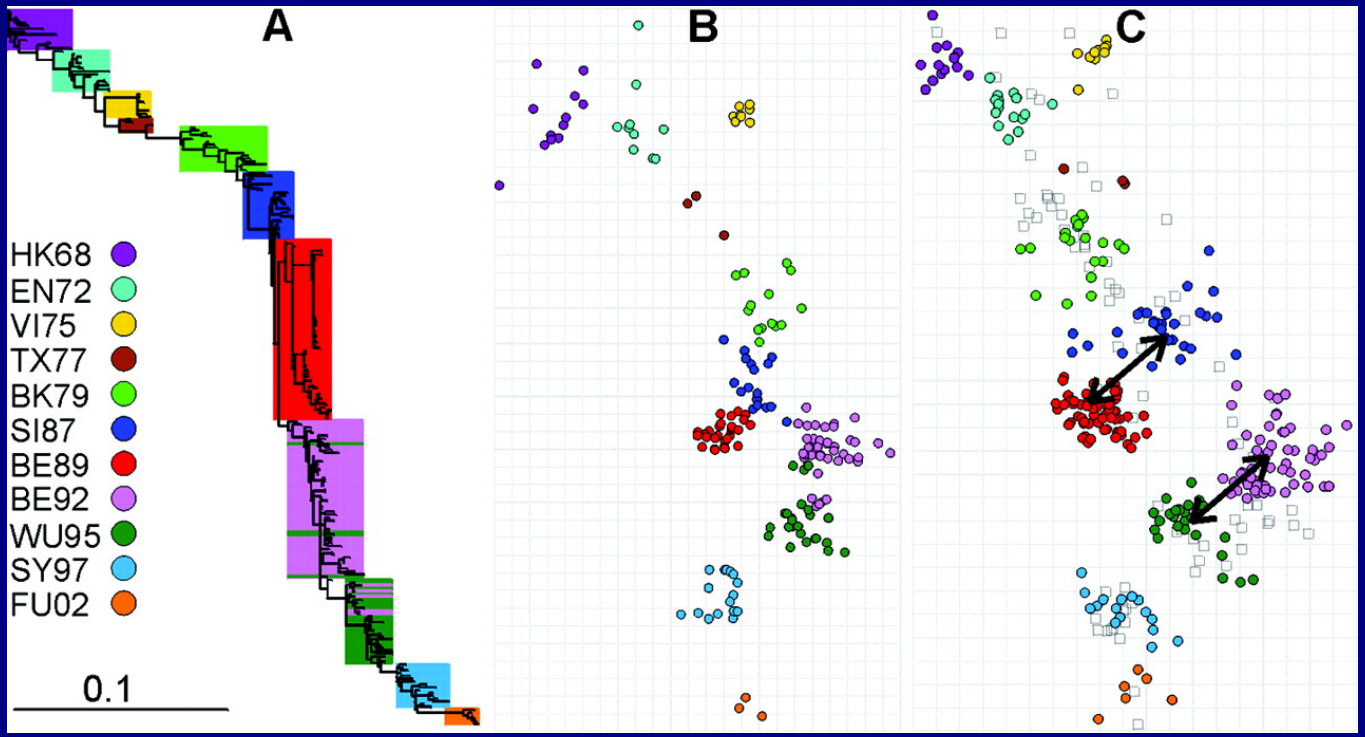
Study HIV within Democratic Replubic of Congo.
Fits the model

$$N = N_0(\alpha + (1-\alpha)e^{rt})$$



Uses of phylogeny: centralized sequences

Influenza phylogenetic and immunological distance correlated



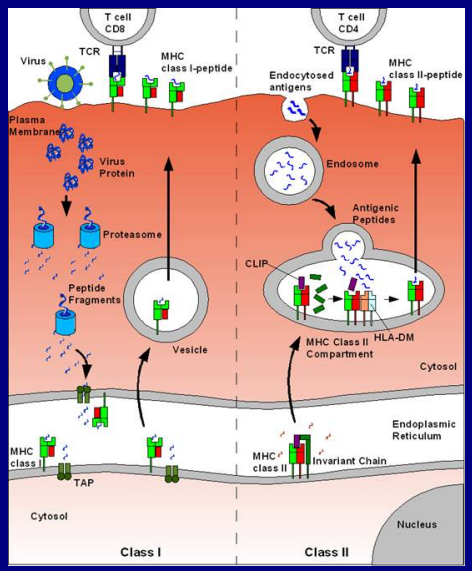
Smith *et al.*
Science 2004 Jul 16: 305(5682):371–376

Ancestral HIV half the distance to current circulating strains.
Reconstruct ancestor as vaccine reagent.

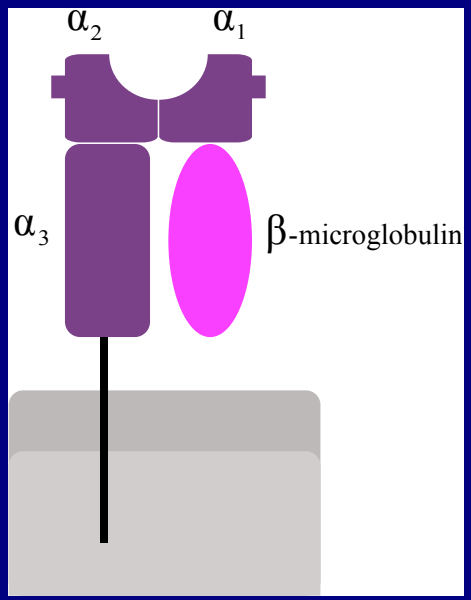
Uses of phylogeny: A Problem in HIV immunology

Cytotoxic T Lymphocytes (CTL) recognize and kill virally infected cells.

Small bits of virus presented for recognition by class I Human Leukocyte Antigen (HLA).



Halling-Brown, Ph.D. Thesis



en.wikipaedia:atropos235

Which bits of the virus (epitopes) are recognized by CTLs?

Solution

Direct Solution

Find the HLA type of the individual.
Find sequence positions where change is selected over time.
Construct overlapping stretches of small peptides.
Study binding.

Statistical Solution

Individuals differ in HLA.
If an epitope recognized, it may escape by changing.
Study a population and HIV extracted from them.
If an HIV sequence position *correlates* with host HLA, it is likely to be in an epitope.

Incorrect Results!!!

89/202 (37–51%) positions in the protein Pol correlated with host HLA.

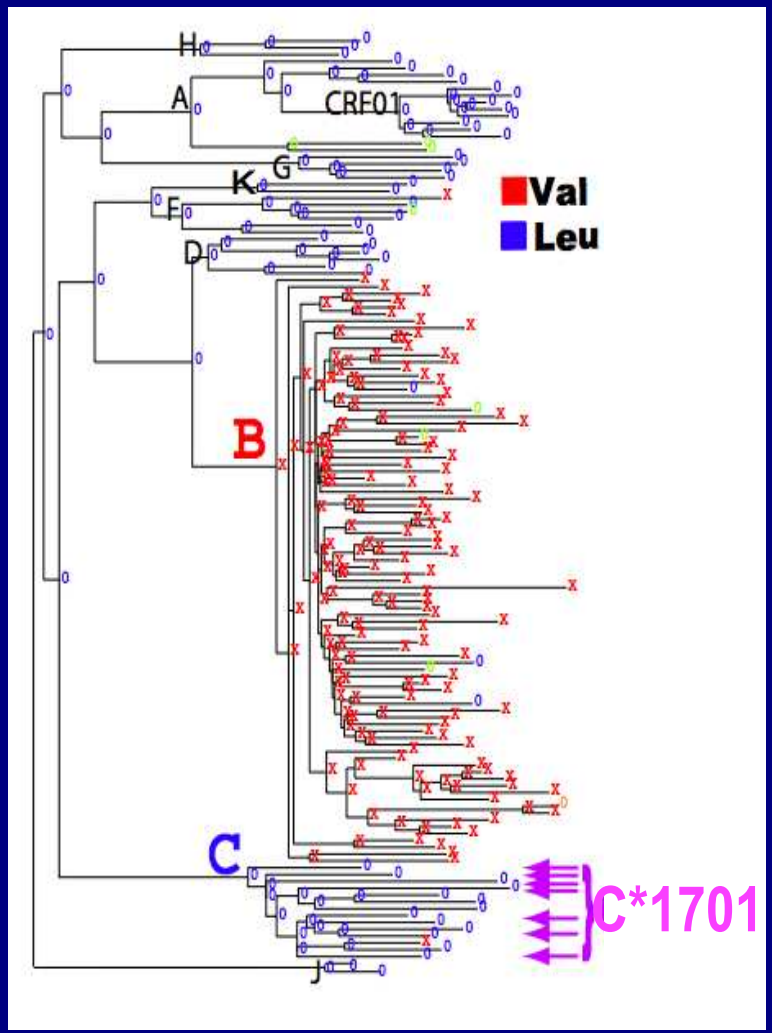
11 (3–10%) significant even after correction for multiple tests.

Moore *et al.,* Science 2002 296:1439–1443.

A separate study: 346/624 (51–59%) positions correlated.

80 (10–16%) significant with a cutoff on false positive rate of 20%. Kiepiela *et al.,* Nature 2004 432:769–775.

Conclusion:
CTL escape significant factor in shaping HIV evolution.



Dataset from Perth dominated by B subtype.

All C*1701+ people in the dataset are C subtype infected. C*1701 common in south Africa, south African epidemic dominated by C subtype.

All clades except B are dominated by Leucine.

Valine is not *changing to Leucine in C*1701+ people.*

Correlation better explained by common descent and migration.

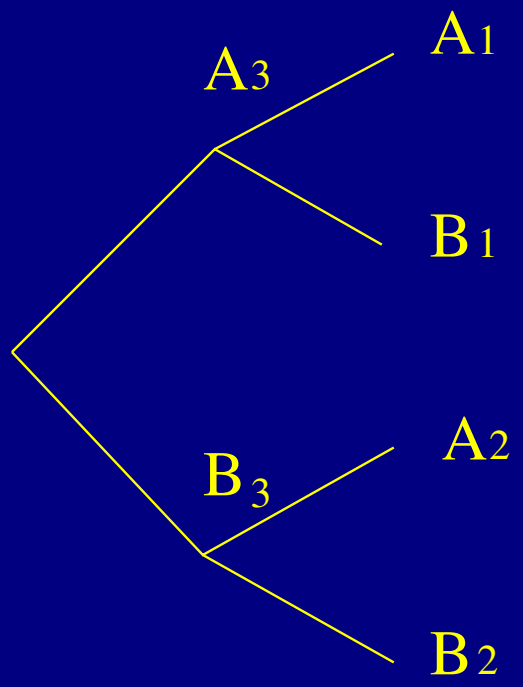
Phylogeny provides a Model of Covariance

In 1985 Felsenstein proposed phylogenetically independent contrasts.

Consider a trait diffusing on a phylogenetic tree:

- The changes on each branch are independent variables, with variance given by branch length.
- The ancestral state can be estimated from the descendants by a mean, weighted by inverse branch lengths.
- A contrast is normalized difference between two daughter nodes:

$$\frac{A/\sigma_A^2 - B/\sigma_B^2}{1/\sigma_A^2 + 1/\sigma_B^2}$$



Markovian Processes

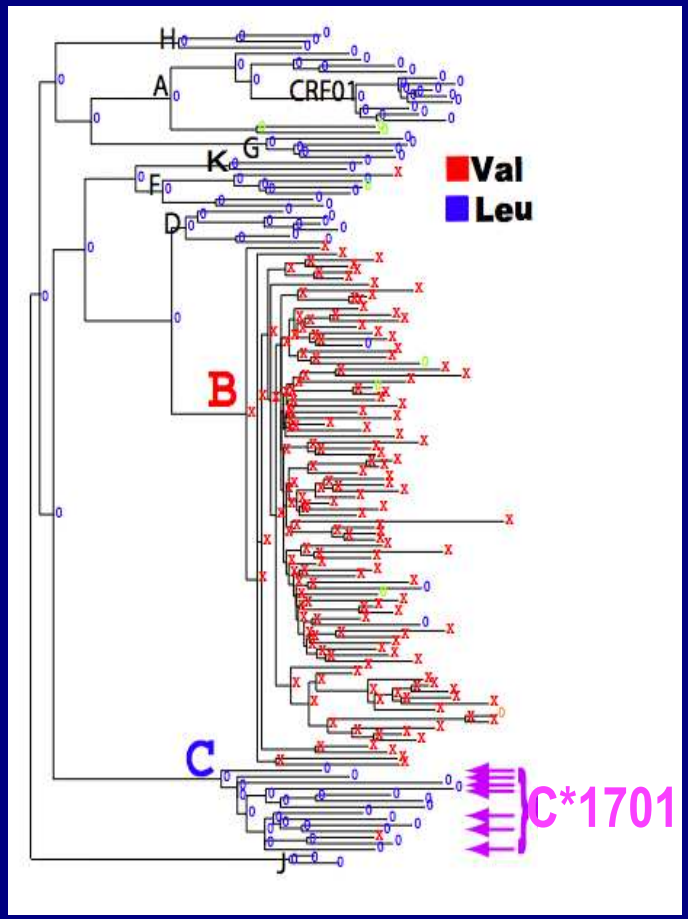
In a Markov process, the *changes* at various instants, conditioned by the state at that instant, are still independent: so instead of looking at the state, one can look at the change.

This formalizes the intuitive problem noticed before: Valine was not changing to Leucine in C*1701+ people.

So, we devise the following method:

- Calculate the ancestor of sequences
- Select cases with common ancestor.
- Correlate *change* or *not change* with feature.

Count



In other words instead of looking at a table like:

	V	Not-V
C*1701+	0	7
Not C*1701+	115	45

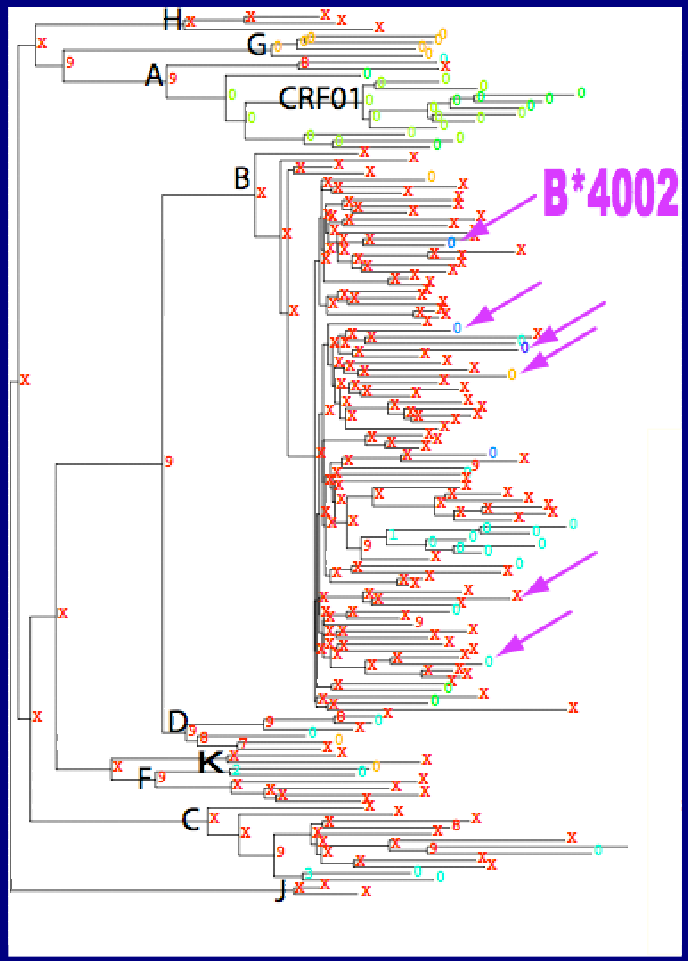
p = 0.0002

we should look at tables like:

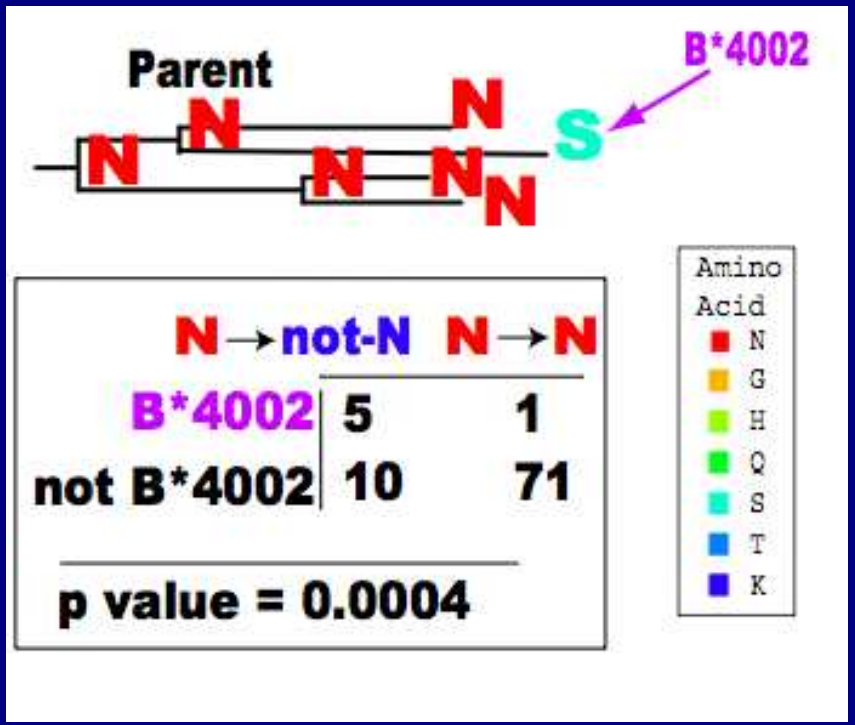
	V→ V	V→ Not-V
C*1701+	0	0
Not C*1701+	75	7

p = 1

True correlaton



Restricted to the cases when the parent is an Aspargine, the daughter changes when the patient is B*4002+



Sensitivity and Specificity

The new method does correct for the phylogenetic artifact.

Without phylogenetic correction, *silent* mutations correlate at the same rate with host HLA.
Silent: 10/153 (3–12%)
Non-silent: 138/1732 (7–9%)
Some silents are very significant: $p < 0.00002$.

With phylogenetic correction 62/80 significant cases due to clade association, and only 7/80 strongly supported.

4/6 were known epitopes, and 2 more were later experimentally found to be true.

Also performs well on synthetic data.

False Positives

Problem in this case was because

- HIV transmitted through social (sexual) contact.
- Human populations are not panmictic.
- HIV transmission clusters correlate with human genetics.

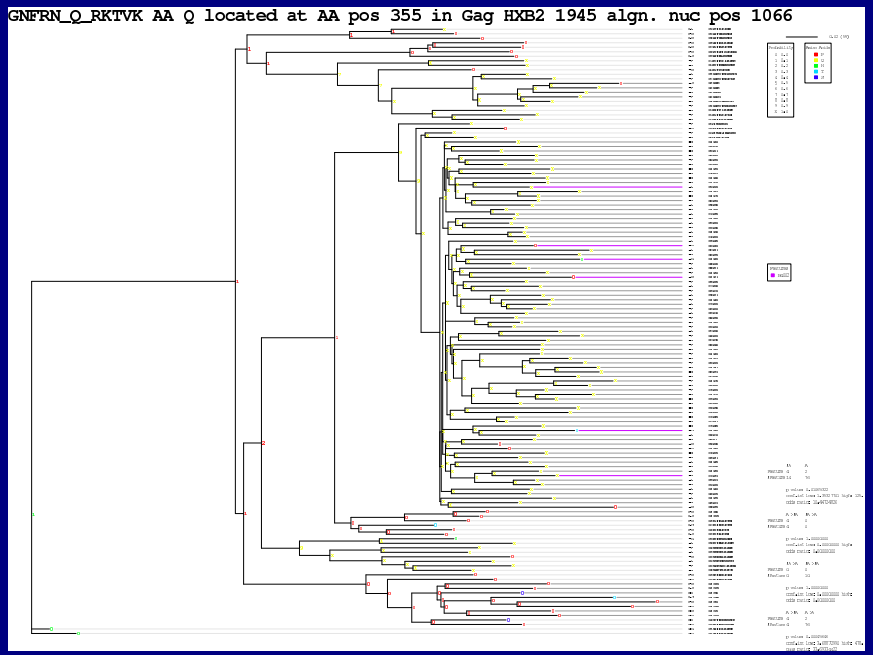
Time scales:

- Humans are only 10^4 generations old.
- Mutation rate 2.5×10^{-8} per base per generation.
- Human genetic decorrelation time: $> 10^8$ generations.
- HIV only 2×10^4 generations old.
- Mutation rate 2.3×10^{-5} per base per generation.
- Most clusters at least a factor of 2 younger.
- HIV decorrelation time $> 10^5$ generations.

Phylogenetics remain important.

Missing true positives

But more generally, it is a question of statistical independence: closely related pieces of evidence overcounted.
This can lead to false negatives as well as false positives.

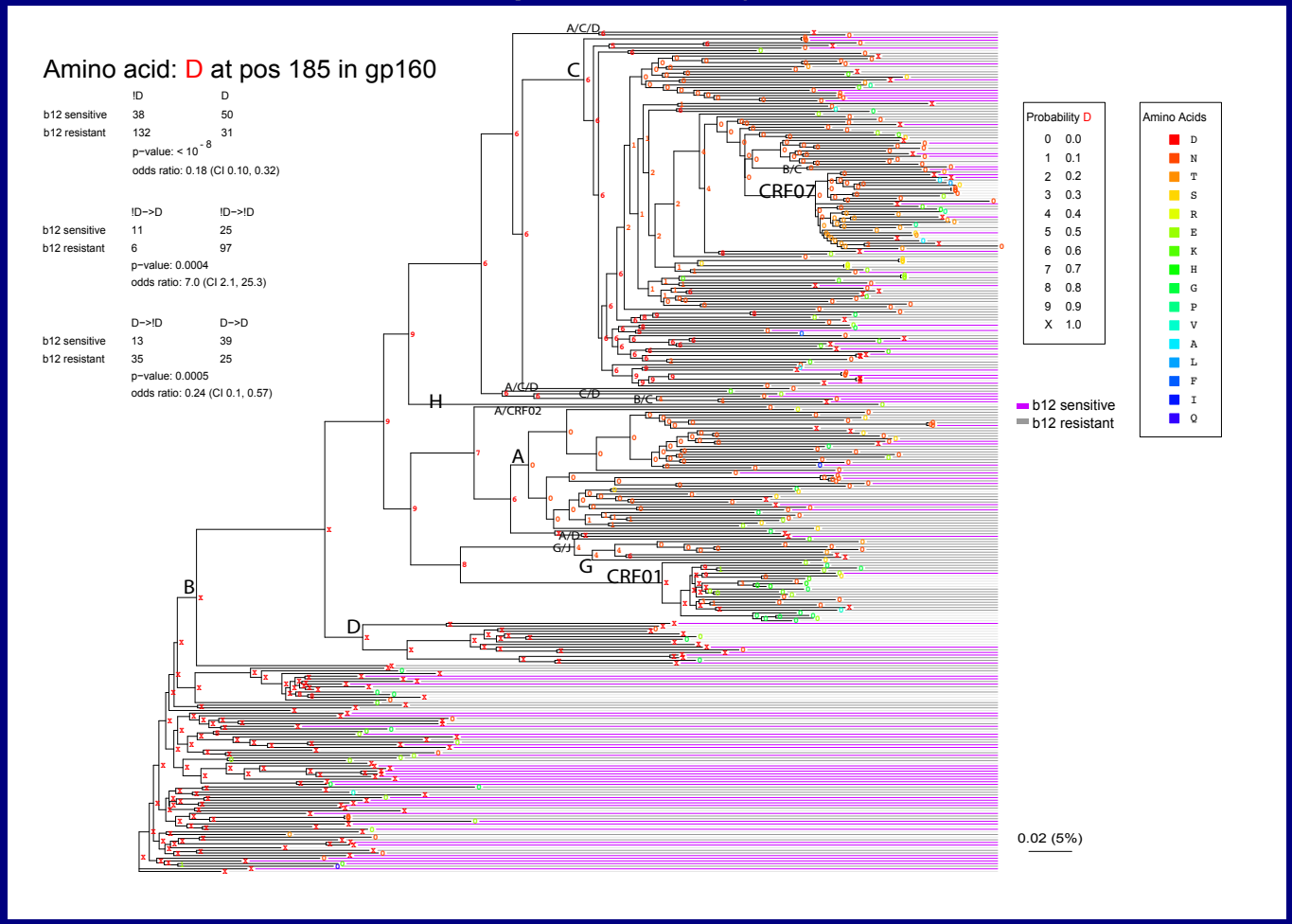


Non-B subtype is predominantly P, not Q, and has no B*4002+ve.

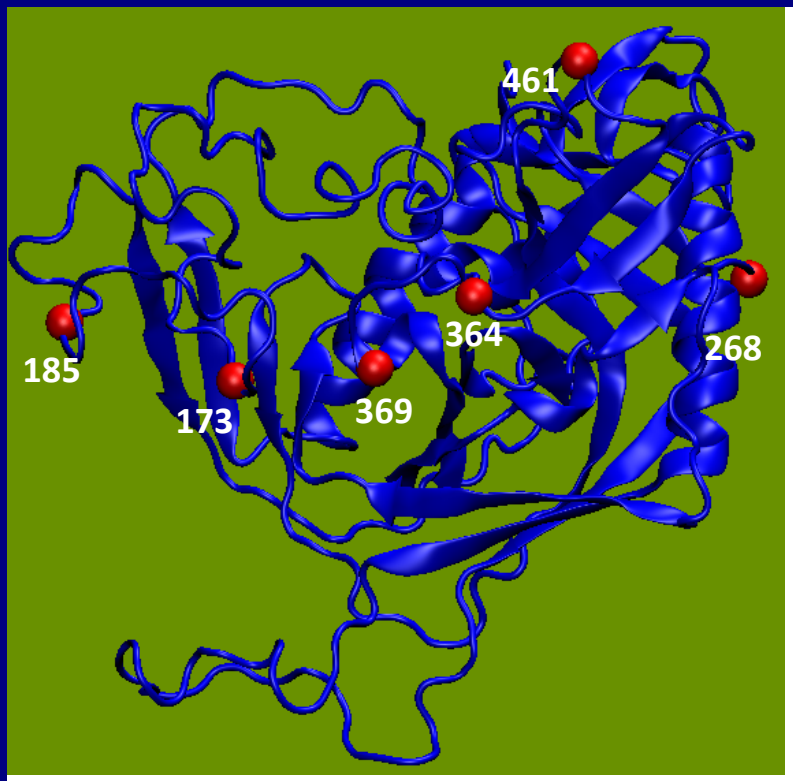
			Q→	
	not Q	Q	not Q	Q
B*4002+	4	2	4	2
B*4002-	14	76	4	76
		p = 0.01		p = 0.00004

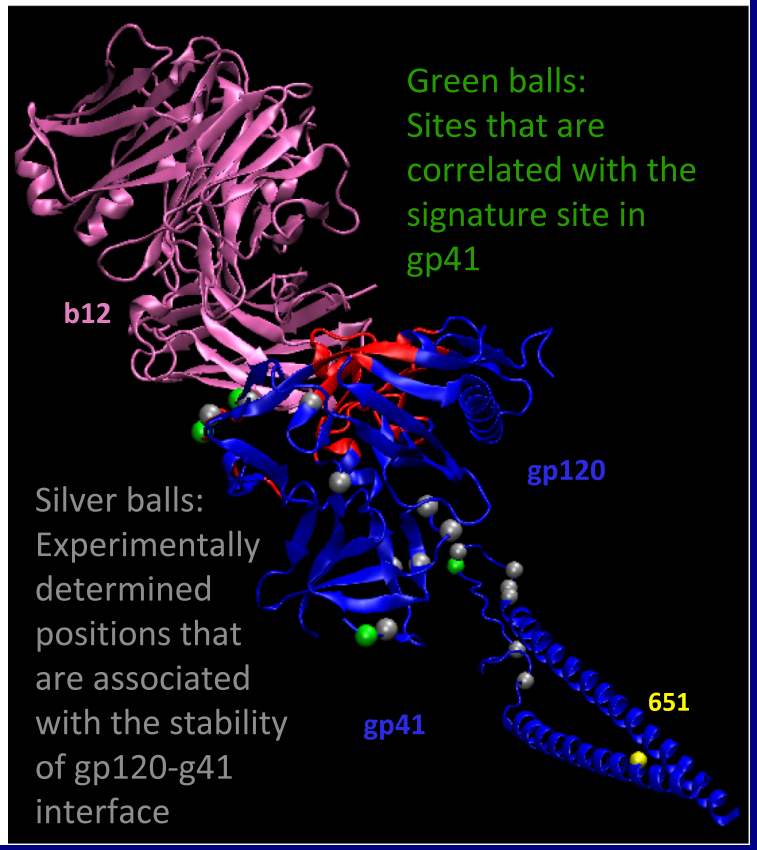
Structure from function

HIV escapes antibodies binding to protein.
Patterns of escape points to binding regions.
Conclusions about quaternary structure.



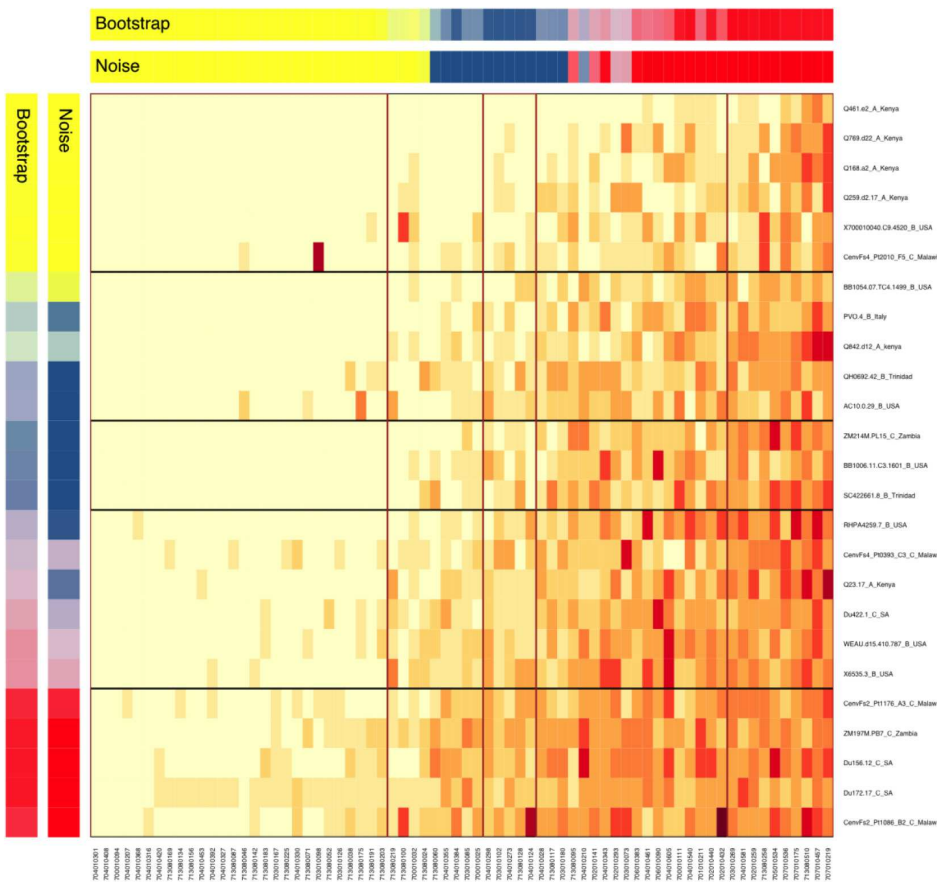
Results


461
185
364
268
173
369


Green balls:
Sites that are correlated with the signature site in gp41
b12
gp120
Silver balls:
Experimentally determined positions that are associated with the stability of gp120-g41 interface
gp41
651

Immunogenicity

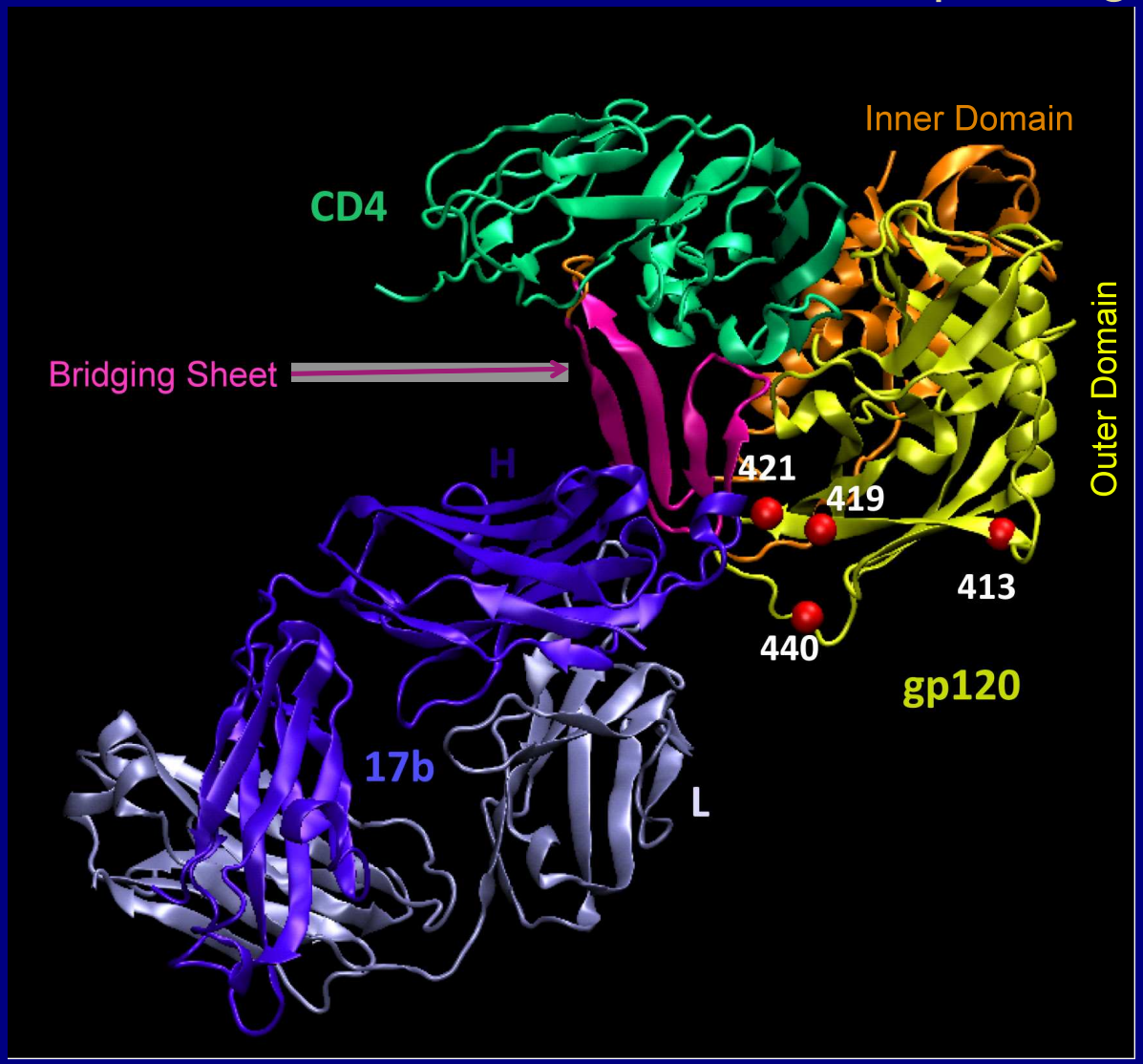
Some antibodies neutralize many viruses. Rarely produced.
Some people make good antibodies.



Host genetics or viral genetics?

Results

All viral sites cluster in CCR5 coreceptor region.



Beyond Biology

Can these methods be applied to other historical sciences?

- Is trait inheritance more important than trait genesis?
- Are there individuals: stable collections of traits?
- Are states easy to define: a closed system?
- Is there vertical transmission: coinheritance bundles?
- Do the coinherited traits show hierarchical structure?
- Are their traits of differring rates?
- Are the distances explained by a tree?
- Is the change process stationary?

These conditions are probably satisfied in many fields, and underlying laws may be discoverable.
Incorrect counting probably common: incorrect deduction of laws.
Important to study historical processes quantitatively.