# INFERENCE OF DIVERGENCE TIMES AS A STATISTICAL INVERSE PROBLEM

*Steven N. Evans, Don Ringe, and Tandy Warnow*

The Program for Evolutionary Dynamics at Harvard University

March 21, 2005

"The past is a foreign country; they do things differently there." Lesley P. Hartley, *The Go-Between* (1953).

## YES OR NO?

Forster & Toth (2003):

> Phylogenetic time estimates ... are statistically feasible once the language tree has been correctly reconstructed, by uncovering any recurrent changes of the items.

# INVERSE PROBLEMS

inverse problem
= reverse engineering a complex system from its outputs

- partial information, often in the form of low-dimensional projections of high-dimensional systems

- issues of non-uniqueness and lack of identifiability

"What song the sirens sang, or what name Achilles assumed when he hid himself among women, though puzzling questions, are not beyond all conjecture." Sir Thomas Browne, *Religio Medici* (1643).

## WHY DO WE NEED MODELS?

- models circumscribe the universe of possible explanations

- any model is a trade-off of

*richness* and *plausibility*
**versus**
*inferential feasibility*

## WHY DO WE NEED MODELS? (cont.)

- more mathematically sophisticated $\not\Rightarrow$ more plausible

- randomness is often used as a *proxy* for unknowable complexity

- models and inferences from them should be consistent, both mathematically and with what is already understood empirically about the system under consideration

- bad models aren't just useless, they can be positively misleading

# THE SIREN SONG...

## Berk and Freedman 2003

> If the random-sampling assumptions do not apply, or the parameters are not clearly defined, or the inferences are to a population that is only vaguely defined, the calibration of uncertainty offered by contemporary statistical technique is in turn rather questionable. Thus, investigators who use conventional statistical technique turn out to be making, explicitly or implicitly, quite restrictive behavioral assumptions about their data collection process. By using apparently familiar arithmetic, they have made substantial empirical commitments; the research enterprise may be distorted by statistical technique, not helped.

# ...AND THE SIRENS KEEP SINGING...

Berk and Freedman 2003

> ... researchers may find themselves assuming that their sample is a random sample from an imaginary population. Such a population has no empirical existence, but is defined in an essentially circular way – as that population from which the sample may be assumed to be randomly drawn. At the risk of the obvious, inferences to imaginary populations are also imaginary.

# MODELS IN LINGUISTIC PHYLOGENY

- same (topological) rooted tree for all characters, one leaf for each language
- state space for each character (= universe of possible values)
- probability distribution at the root for each character
- character-specific *substitution* mechanisms for each edge (conditional probabilities)
- Markov random field on vertices
- only observe character states at leaves (ancestral states hidden)
- different characters usually assumed independent

## HOW DO WE USE MODELS TO MAKE PHYLOGENETIC INFERENCES?

- maximum likelihood = find the choice of parameters (tree + "numerical" parameters) that make the data most likely

- could also use Bayesian and other methods

# SOME ISSUES TO CONSIDER

- What is a reasonable state space? (Cf. biology, where we naturally have 4 bases or 20 amino acids.)

- Shouldn't look at the data first to decide what the state space is - interpretation of probabilities becomes very problematic.

- What are reasonable substitution mechanisms – do we understand the *geometry* of lexical, phonological or morphological "space"? (Cf. biology where transition mechanisms are informed by biochemistry and analysis of independent data - Dayhoff, PAM.)

- Need understanding of linguistic processes - not black boxes.

- Many-to-one coding can destroy the Markov structure.

# MODELS, BOXES, AND TICKETS

- The usual probabilistic models for linguistic evolution are meant to describe *generic* characters from some population that is being sampled, and inferences are only justified to the extent that this is a reasonable assumption.

- Essentially, we have the following *Gedanken-experiment*: there is a population of possible states for a character that are akin to tickets in a box, some states appear on more tickets than others in proportion to how *likely* the state is to be exhibited by the character, and we imagine that nature has somehow shaken up the box and chosen a ticket at random to give us the observed state of the character.

## MISSING DATA, ASCERTAINMENT, AND WHAT'S OUR STATE SPACE?

- If someone was allowed to rummage through the box and discard tickets before the drawing took place or we are able to look at a ticket after the drawing and can accept or reject it, then the proportions of tickets originally in the box no longer describe the experiment and we need to consider another, perhaps substantially more complex, box that somehow incorporates this *a priori* or *a posteriori* winnowing.

# WHAT CAN WE DO WITH SUCH MODELS?

Recent theory has established:

- The tree topology (= shape) can be recovered with a high degree of accuracy with not too many characters.

- That is, can do *cladistics* = order divergences along lineages.

- Data can be reasonably heterogeneous if we don't care about estimating substitution mechanisms.

- To *date* divergence times we need something more.

# MODELS THAT PERMIT DATING

- the tree has edge lengths (= time durations between divergences)

- the substitution mechanism on an edge is tied somehow to length

- typically, substitution comes from running a Markov chain for the time specified by the edge length

- there is control on heterogeneity of mechanisms between characters - i.e., we need "replication"

# SOME ISSUES TO CONSIDER ABOUT MODELS FOR DATING

- Do we believe the scenario implicit in the Markov chain assumption?

- How much can Markov chain mechanisms vary from edge to edge and character to character?

- Arbitrary heterogeneity leads to unidentifiability - different edge lengths give the same probability distribution for the data.

- Same mechanism for all characters and edges not tenable.

- What to do?

# A FIX FOR DATING MODELS?

• Assume that on each edge for each character we have the *same* Markov chain run at different rates.

• Rate for (character, edge) pair is of the form

(character-specific rate) $\times$ (edge-specific rate)

• Each character behaves like any other, just "scaled" up or down. Is this tenable?

• Still too many parameters! Unidentifiability. What to do?

## A FIX FOR EDGE-SPECIFIC RATES?

- Don't allow edge-specific rates to vary too much from "mother" edge to "daughter" edges.

- More precisely, subtract an extra *ad hoc*, apples-and-oranges "roughness penalty" from the likelihood and maximize the resulting quantity (idea due to Mike Sanderson in biology, based on non-parametric regression methods, used by Gray and Atkinson).

- Quite arbitrary, introduces bias.

- Still have too many parameters. What to do?

# A FIX FOR CHARACTER-SPECIFIC RATES?

- Rather than assume character-specific rates are arbitrary, assume they are *independent random draws* from an unknown probability distribution in some *low-dimensional parametric family*, typically the gamma distributions.

- No scientific rationale for gammas - just flexible and leads to nice formulae.

- Random rates more than a methodological convenience: definite empirical commitment. All characters really the same.

- Depends on believing the gamma assumption - gammas can be embedded in a larger parametric family where two rate distributions from the family give the same data distributions for TWO DIFFERENT TREES! (E. + Warnow)

# EVALUATING MODELS

- There are no *gold standards* of sufficient external verification.

- There are necessary minimal standards of internal consistency.

  - Models should make sense mathematically.

  - Methods based on restrictive models chosen for mathematical convenience should still work on simulated "data" from richer, more plausible models.

  - Do we have appropriate richer models for such testing?

# CONCLUSION

- Current dating models in linguistics *and* biology are very restrictive in order to overcome problems of unidentifiability.

- Current linguistic models are black boxes expropriated from biology, without much attempt to validate empirically whether biological change and linguistic change are similar processes.

- We need appropriate models of linguistic evolution.

- We need standards for evaluating models.

# EPILOGUE

The universe is not only stranger than we imagine,
it is stranger than we can imagine.

– J.B.S. Haldane

As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don't know
We don't know.

– Donald Rumsfeld, U.S. Secretary of Defense