Foundations of the Age-Area Hypothesis

Matt Baker

Background

- The economic basis for indigenous institutions:
    - Baker (2003, 2008), Baker and Miceli (2005), Baker and Jacobsen (2007, 2008).
- Question: How environment, technology, and institutions co-evolve

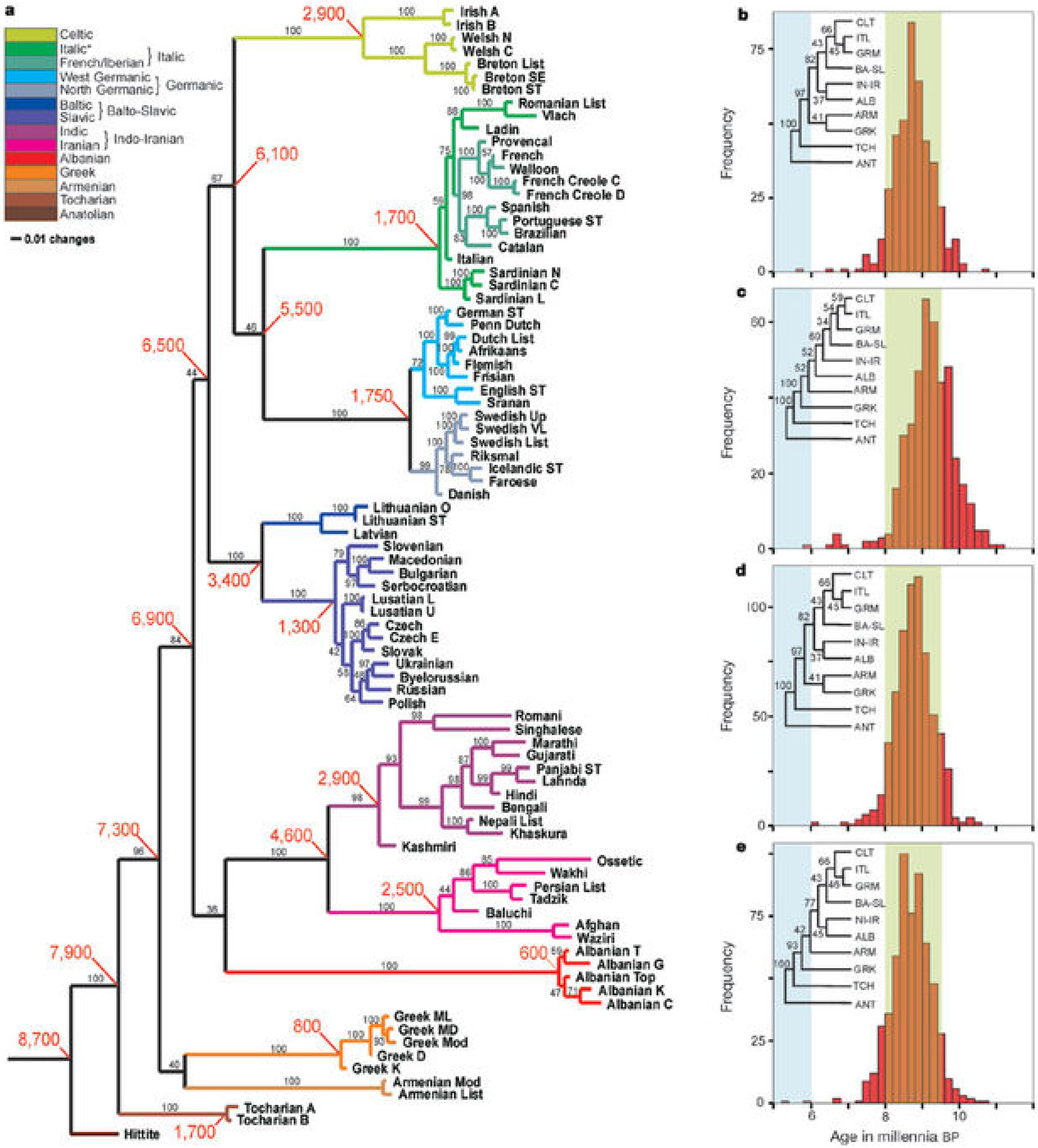
Recently

- Applications in economic growth:
  - Alesina et. al. (2005), Spolaore and Wacziarg (2013), Michalopoulus (2012), Fenske (2012)
- Computational linguistics and Phylogenetic approaches to analyzing cultural diversity (Mace, 2006)
  - computing power!
- Incorporation of geographical data into analyses.

Question: How did ethnic and geographic diversity that we observe today come about?

Cultural similarities

- Related to genetic similarities
- Computational linguistics - treat aspects of language like a genetic code with drift.
- Build *Phylogenies* of related cultures; epitome Mace (2006).
- Atkinson and Gray (2006) example: Indo-European Tree.
- Fairly sophisticated machinery for doing this! Great way to build "Path dependency" and "drift" into the analysis



a
Celtic
Italic*
French/Iberian
Italic
West Germanic
North Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
0.01 changes
Irish A
Irish B
Welsh N
Welsh C
Breton List
Breton SE
Breton ST
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
Italian
Sardinian N
Sardinian C
Sardinian L
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Sranan
Swedish Up
Swedish VL
Swedish List
Riksmal
Icelandic ST
Faroese
Danish
Lithuanian O
Lithuanian ST
Latvian
Slovenian
Macedonian
Bulgarian
Serbocroatian
Lusatian L
Lusatian U
Czech
Czech E
Slovak
Ukrainian
Byelorussian
Russian
Polish
Romani
Singhalese
Marathi
Gujarati
Panjabi ST
Lahnda
Hindi
Bengali
Nepali List
Khaskura
Kashmiri
Ossetic
Wakhi
Persian List
Tadzik
Baluchi
Afghan
Waziri
Albanian T
Albanian G
Albanian Top
Albanian K
Albanian C
Greek ML
Greek MD
Greek Mod
Greek D
Greek K
Armenian Mod
Armenian List
Tocharian A
Tocharian B
Hittite
2,900
6,100
1,700
5,500
6,500
1,750
3,400
1,300
6,900
2,900
4,600
7,300
2,500
7,900
600
8,700
800
1,700
b
c
d
e
CLT
ITL
GRM
BA-SL
IN-IR
NI-IR
ALB
ARM
GRK
TCH
ANT
Frequency
Age in millennia BP

Practical Questions:

- Where did this tree originate?
- How did the peoples of the tree come to be where they are?
- Which related cultures have been in close proximity, and for how long?

Questions of geography, cultural/lingustic drift, and time.

The Age-Area Hypothesis (AAH)

- Sapir (1916) - *the root of the Phylogenetic tree is the most likely geographical point of origin.*
- Recursive application - migratory routes
- Used to resolve historical debates, but also could be important in creating new theories

Old applications and continuing debates

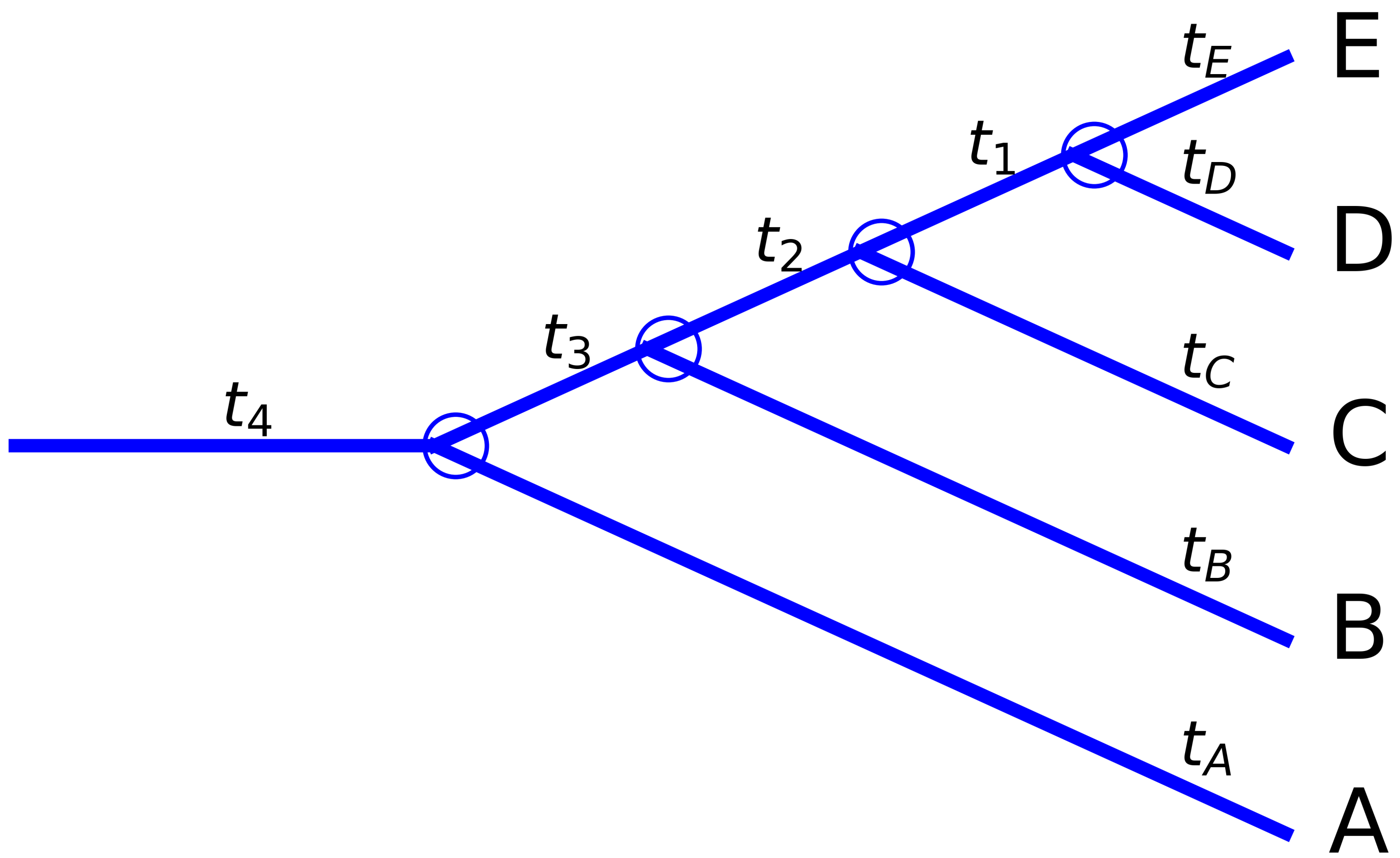
- Origins of Athabaskan/Na-Dene speakers
- Indo-European origins
- Afro-Asiatic origins
- Spread of Bantu peoples
- Native American population dispersal

On the need or theory...

Greenhill and Gray (2005) write: "many expansion scenarios are little more than plausible narratives. A common feature of these narratives is the assertion that a particular lineof evidence (archaeological, linguistic, or genetic) is 'consistent with' the scenario. 'Consistent with' covers a multitude of sins.

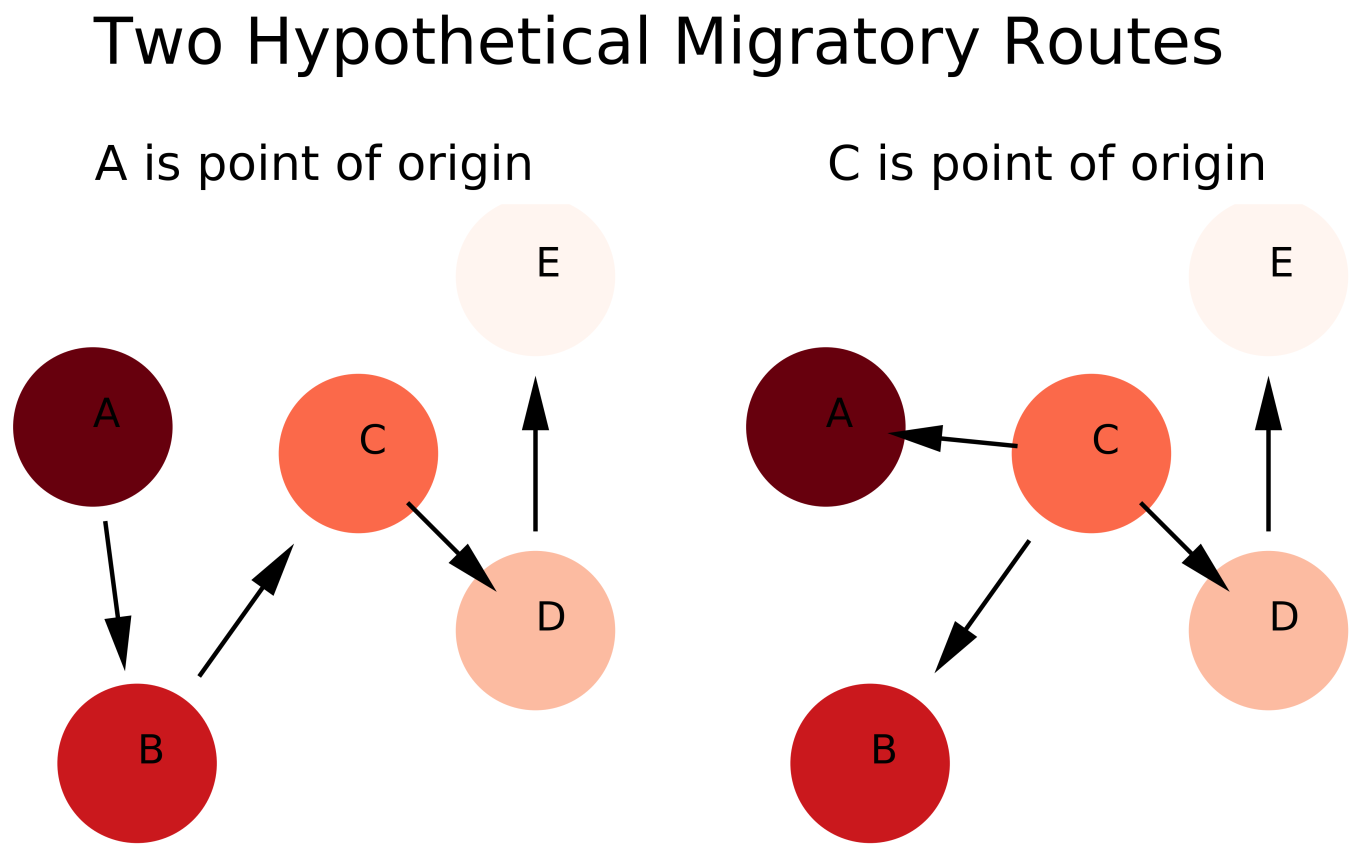
So why believe the AAH (or not)?

- Occam's Razor?
- Minimum effort or # of moves?
- Dyer (1956, p. 613) hits upon the idea of conserving moves of a particular sort: "...the probabilities of different reconstructed migrations are in inverse relation to the number of language movements required."



t_E
E
t_1
t_D
D
t_2
t_3
t_C
C
t_4
t_B
B
t_A
A

Problem Preview



Candidate Migratory Histories:

- A is point of origin - A to B to C to D to E
- C is point of origin - C to A, C to B, C to D to E
- Both are consistent with observed phylogenetic difference or drift. **The tree tells us which migrations happened first!**
- Note "minimum moves" doesn't get us very far. Both have four moves!

Basic Model:

- **Assume a full, rooted binary tree**
  - Tree with z terminal nodes will have $z - 1$ internal nodes, which are the minimal number of geographic moves needed to span the tree.
- **Current locations coincide with historic locations**
- **All constituents of the tree observed**

Definitions

Migratory Event

A location jump from one location to a new, unoccupied one.

Migratory Chain

A sequence of "forward moving," jumps in space migratory events that end at a terminal node/taxa/culture.

Migratory History

A collection of chains spanning the whole tree, with a "deepest chain" starting at a given location.

The Basic Idea - assumptions

1. A migratory chain occupies one location at a time.
2. When a chain moves from its location, a new chain starts in its place.
3. Migratory chains move to new locations at random times, according to an Exponential/Poisson density.
4. Each migratory chain is unique in that it has its own parameters.

Chain One:

- requires a chain from A to B to C to D to E (or D to E)
- By the previous rules, new chains start at A, B, C, and D. Let T denote the length of the tree.
- Likelihood:

$$L_A = \frac{(\lambda_1 T)^4 e^{-\lambda_1 T}}{4!} \times \frac{(\lambda_A t_A)^0 e^{-\lambda_A t_A}}{0!} \frac{(\lambda_B t_B)^0 e^{-\lambda_3 t_B}}{0!} \frac{(\lambda_C t_C)^0 e^{-\lambda_C t_C}}{0!} \frac{(\lambda_D t_D)^0 e^{-\lambda_D t_D}}{0!}$$

- Seems like overkill, but notice the dead branches

Log-Likelihood:

$$
\begin{aligned}
\ln L_A = 4\ln(\lambda_1 T) - 4\lambda_1 T - \ln(4!) \\
- \lambda_A t_A - \lambda_B t_B - \lambda_C t_C - \lambda_D t_D
\end{aligned}
$$

Optimized with $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0$, and then:

$$\lambda_1 = \frac{4}{T}$$

Substituting this all back into the original likelihood gives "Profile" or "Concentrated" likelihood:

$$L_A = \frac{4^4 e^{-4}}{4!}$$

Chain Two:

Log-Likelihood

$$L_C = \frac{(\lambda_1(t_4 + t_A)^1 e^{-\lambda_1(t_4+t_A)}}{1!} \frac{(\lambda_2(t_3 + t_B))^1 e^{-\lambda_B(t_3+t_B)}}{1!}$$
$$\times \frac{(\lambda_3(t_2 + t_1 + t_E))^2 e^{-\lambda_3(t_2+t_1+t_E)}}{2!}$$
$$\times \frac{(\lambda_C t_C)^0 e^{-\lambda_C t_C}}{0!} \frac{(\lambda_D t_D)^0 e^{-\lambda_D t_D}}{0!}$$

Highlight: fewer degenerate chains, and more active chains!

Chain two profile/concentrated-likelihood:

$$L_C = \frac{1^1 e^{-}}{1!} \frac{1^1 e^{-1}}{1!} \frac{2^2 e^{-2}}{2!} = \frac{2^2 e^{-4}}{2!}$$

Comparison of L_A and L_C is a race between $\frac{4^4}{4!}$ and $\frac{2^2}{2!}$.

Relative likelihood: $P(A|A \text{ or } C) = \frac{\frac{4^4}{4!}}{\frac{4^4}{4!} + \frac{2^2}{2!}} = .84$

Key Feature:

$$h(n) = \frac{n^n}{n!}$$

is a convex function. Breaking it up into smaller chunks, or spreading n around is a losing proposition. So:

$$h(n) > h(n-k)h(k)$$

Questions:

- How can these ideas be related to a notion of distance or divergence?
- How can divergence and probability be tied together, as the AAH supposes?

A Start:

- With each location k, there are a family of possible migratory histories $\mathcal{H}_k$ that explain the phylogeny.
- For $H_k \in \mathcal{H}_k$, define $N(H_k)$ as a count of the migratory chains in the history.
- Define $n(C)$ as a count of the number of events in a migratory chain, and then define:
- $n^*_{H_k} = \max_{C_{ik} \in H_k}[n(C_{1k}), n(C_{2k}), ..., n(C_{N(H_k)k})]$ The maximum node count for a chain in H_k.

Definition: Dyen Divergence

Start with a function $D_{H_k} = m(n^*_{H_k}, N(H_k))$, where m is increasing in its first argument, and decreasing in the second. Define now the *Dyen Divergence* as

$$D_k = \max[D_{H_{1k}}, D_{H_{2k}}, ..., D_{H_{Ik}}]$$

A family of divergence measures. Examples:

- $D^1_k = n^*_{H_k} - N(H_k)$
- $D^2_k = \frac{n^*_{H_k}}{N(H_k)}$

Age-Area Theorem

Suppose model assumptions hold, and define a Dyen Divergence measure. Then:

$$D_k \geq D_j \implies L_k \geq L_j$$

Further

$$\begin{aligned} k &= \arg\max \left[D_1, D_2, D_3, ..., D_n\right] \\ \implies k &= \arg\max \left[L_1, L_2, L_3, ..., L_n\right] \end{aligned}$$

Proof

Basic idea is to note likelihood obeys

$$L_k \propto \prod_{j=1}^{N(H_k^*)} h(n_j), \quad \sum n_j = I$$

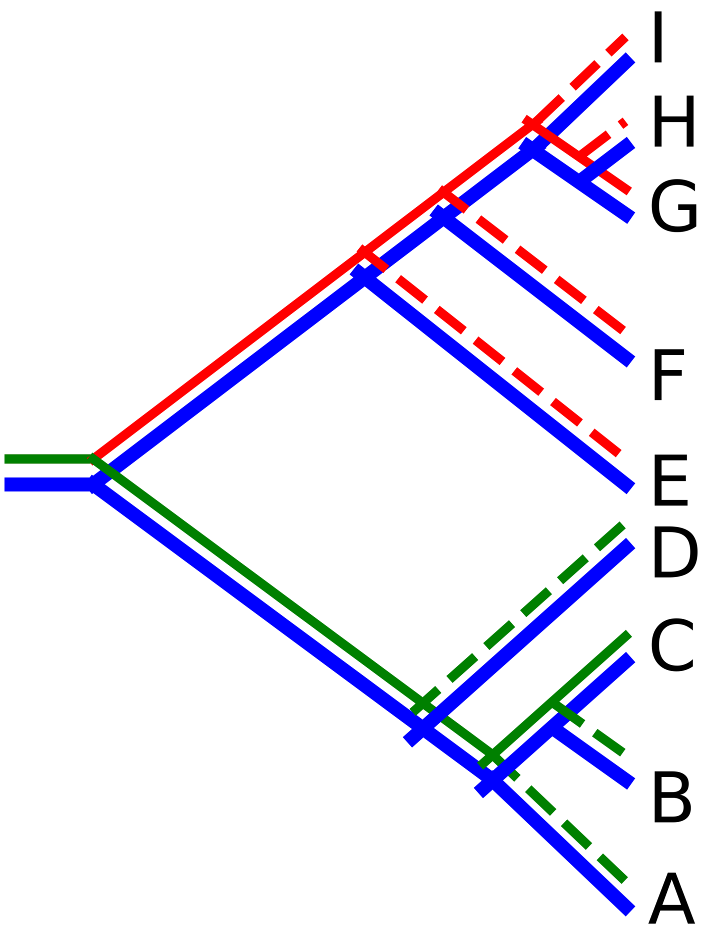
Because of convexity of $h(n)$, pile up as many n's in as few chains as possible. Analogy: a risk-loving investor with fixed assets and a bunch of investment choices.

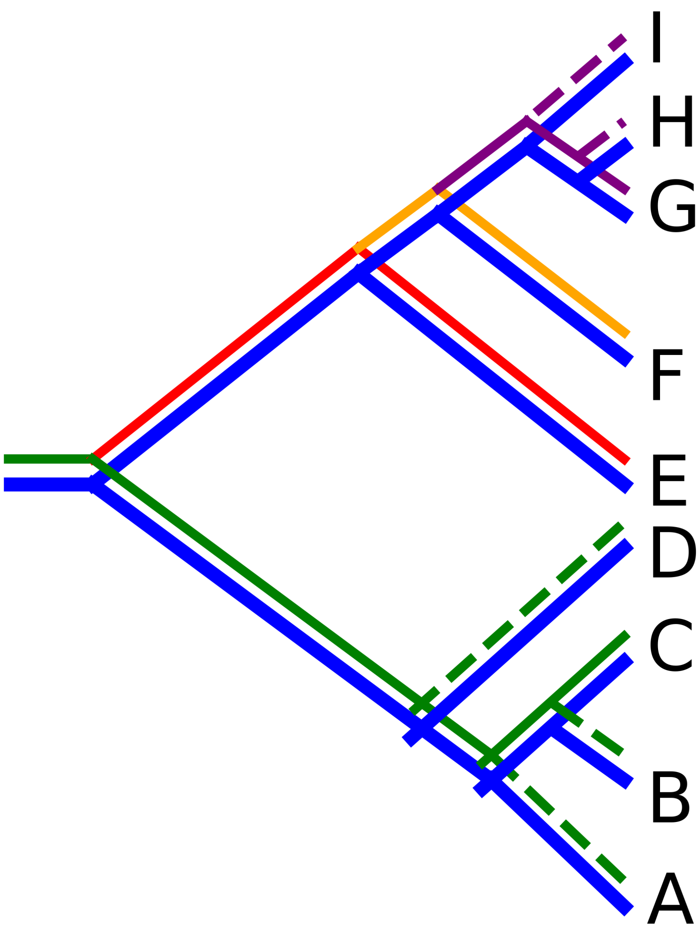
Possible Dyen Divergences:

$$D_i^1 = \frac{n^*_{H_k^*}}{N(H_k^*)}$$

$$D_i^2 = n^*_{H_k^*} - N(H_k^*)$$

Additional Example: G versus I





Comparison of divergence measures

- If E if the point of origin:
  - chain from E to D to A to B to C
  - chain from E to F to I to G to H
  - Dyen Divergence - 2 chains, 4 events each. $D^1 = 2$, $D^2 = 2$.
- If D is the point of origin:
  - chain from I to D to A to B to C
  - chain from I to E, chain from I to F
  - chain from I to H to G.
  - Dyen Divergence - 4 chains, with 4, 1, 1, and 2 events. $D^1 = 0$, $D^2 = 1$.

Likelihoods

- For E, calculating it out gives relative likelihood as .84.
- A better contender to E, however, is D. Two chains, one with 5 events, and one with 3.
    - Dyen Divergence - $D^1 = \mathbf{3}$, $D^2 = \mathbf{2.5}$
- Seems like there is plenty of room for stuff like this in Phylogenetic analysis.

Bells and Whistles

- Known branch lengths - doesn't change much - Exponential becomes Poisson.
- Algorithm for calculation - one can traverse the tree backwards, using dynamic programming to pick out the most likely path
- Include other information in the decision (for example, physical distance).

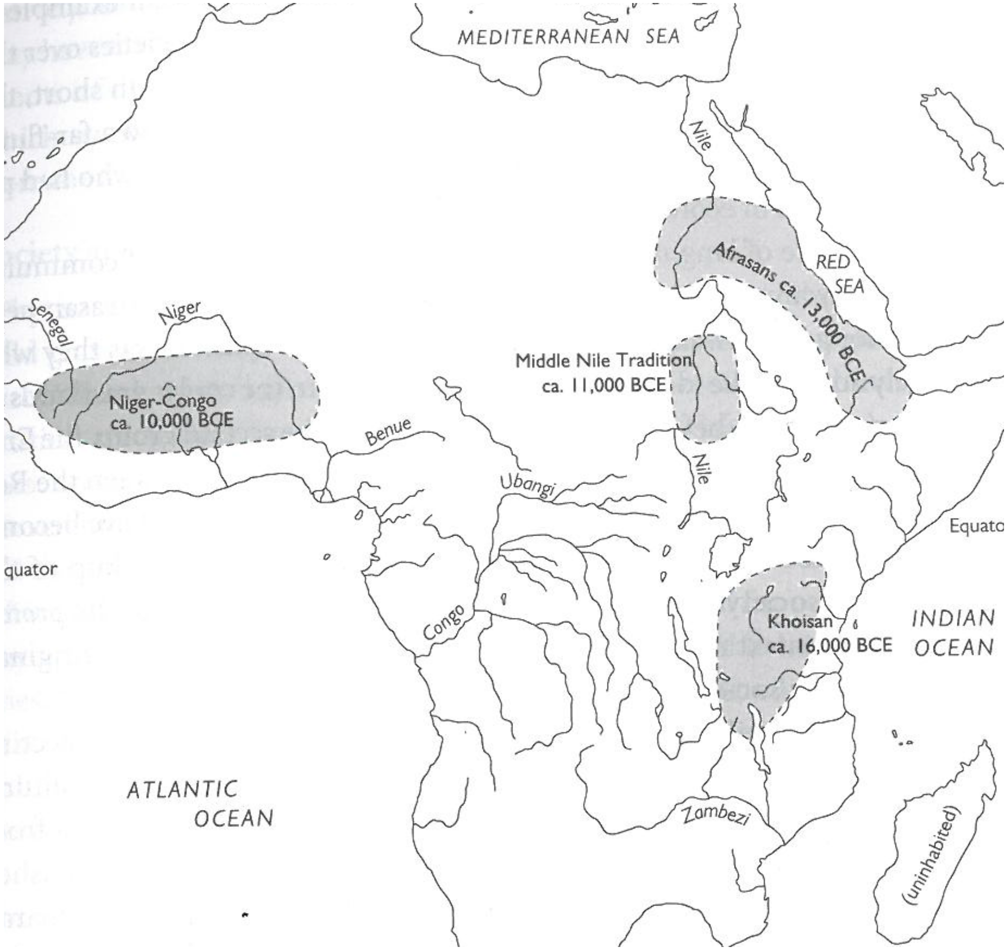
Micro foundations

- Why would one believe the exponential/Poisson arrival rate story?
- Idea: stochastic population growth model
  - Some development creates a superabundance of resources.
  - Once a barrier is achieved, the superabundance dissipates.
  - Too many people at this point in time, so some segment of the population moves on
  - Once the population has moved on, a new superabundance parameter is drawn.

Applications

- Afrasan (Afroasiatic in older sources) and its point of origin. Arabic and Semitic languages, Ancient Egyptian, and Ethiopiac languages as well. Where did it all begin?
- NaDene phylogeny and its point of origin. Simulating *spatial and temporal points of origin*.

Ehret Figure - origins of Afrasans



Probabilities of Origin Points

- Link to Afrasan Map
- Link to Na Dene Map
- Link to Khoisan Map

Sampling NaDene History

- Idea: blend known branch lengths, migratory distances, and estimation of a linguistic phylogeny using standard methods.
- Create a probability distribution over histories.
- Idea follows Baxter and Ramer (1993), Mace and Pagel(1993) - use the type of first letter for Swadesh lists.
- Lexicostatistics/Glottochronology

```
cwittl-cw-ctssscwlilccltttyiNycciiccsts-s  HAIDA
citllyyccltttctclNcliccssicsycccccticti    CHIRICAHUA
sNNcltllsccsttcsssssssssiiiiissttcttttcs  NAVAJO
sNtlNtllyct-tctstmc-clc-Nicc-sytsct-lm-i  JICARILLA
sNNlNNllyc--sctcNiwscl--Nict-sstsct-lp-y  JICARILLA_APACHE
sNNlNclNyct-cctcNNcsclc-Nict-cctcct-cp-c  SAN_CARLOS
wNNlNcllyit-cctcNNc-clc-Nicc-sytsct-cm-i  LIPAN
sNNlitllyii-tctcNcccclc-Nicc-cstcci-cc-i  BEAVER
sNNlNtllytt-tttcNNctcNt-titt-sttctt-cp-i  CARRIER
sNNiNtllyti-iiiiiiiiilt-iicp-sptccN-ti-i  C_CARRIER
sNNiNtllsti-ttiiiiiicNt-citt-sstccN-ti-i  KUTCHIN
iiiiNttl-tiiiiiiiiiiiiiii-tt-sstcctcit-p  HARE
sNciNtllyct-tttcNNciclt-Nicc-ssttct-cp-i  CHIPEWYAN
sNNlitllytt-tctctNwc-N--ticc-sstcct-c--y  SARCEE
sNNlNtlNyct-tctcNicsclp-licc-ysccci-lp-i  HUPA_2
sNNlNNllyct-tctcNccsclp-licc-yctcci-lp-i  MATTOLE
stttNilt-t--tityttwsclpstc-c--sc-cictc-s  KATO
sNNlNtllyNt-tctcNccsclp-licc-yctcci-lp-i  GALICE
sNNlNtllyciitciiiiiiiiiiiyttysstccittiii  TANACROSS
sNNlNNlcyct-tctcNicsclp-licc-cstcci-lp-i  EYAK
ittssNcc-ccccsNcccctcsttNccciccccccctsscc  TLINGIT
```

Words: I, you, we, one, two, person, fish, dog, louse, tree, leaf, skin, blood, bone, horn, ear, eye, nose, tooth, tongue, knee, hand, breast, liver, drink, see, hear, die, come, sun, star, water, stone, fire, path, mountain, night, full, new, name.

Estimation using MCMC methods

- Density is $P(H|T)P(T)$, coupled with prior on certain split dates and on tree structure.
- Simulation from distribution estimated using linguistics

Conclusions:

- Recent literature on diversity is getting more sophisticated and multidisciplinary
- Doesn't mean it should sacrifice rigor.
- Current paper: formalize and operationalize some of this
- In the future: formal models of borrowing, interaction, and cultural evolution.