James Fogleman
Silvana Gaudieri
Phillip Harris
Eleanor Steinberg
Anne Yoder

mail

- mail/room
- evo15
- passwords are mole94

computers

- talk to Brenden
- don't use Swofford
- Vax a ... evo 3
- evo 5 ... vax .. normal
- evo 1 ... vax .. normal
- evo 4
- two Indigos

two disks ... VMS?
unix?

MACS

- WORKSHOP1-25
- SERVER
  - HAS MISC. UTILITIES
  - MOUNTED BY RESTARTED

- TRANSFER FILES
  - CHOOSER --- APPLESHARE
  - WORKSHOP COMPUTERS - name = Workshop_#
  - PASSWORD = WORK#

EVOL1

- doubleclick
- go to open to get to other computers

- dir

MAIL

- mail enter
  - send ... enter
  - to? ... "name"
  - subject --

  quit when done
  exit

} INSIDE DOMAIN

OUTSIDE {
- send enter
- to type 2 on keypad get SMTP% "address"

SUBDIRECTORIES

LOGOFF

FTP

TELNET PROGRAM

① make sure FTP is enabled
② set transfer directory
-unselect drive

mput
mget

③ type FTP ... go to network ... send IP #

④ put "file name"

⑤

PUTTING TO VAX

① set directory
② enable

} PUNT VIEW SEQ FILES

FETCH

host
user-ID
passwd

automatically does binaries

MANUALS - BLACK BOOKS

WORKSHOP

SEQUENCE
- load master (for master.seq)
- C
- move cursor to right

- font
  BKG Cycle --- color

- ctrl Z exits

W80
W132

Vtwidth

sugin

ERROR IN MANUAL
PF1 4 ... GOLD BOTTOM

GOLD PF1 GOLD	Help PF2 Help	Find PF3 Find Next	undel line PF4 del line
command 7	8	replace 9 append	—
top 4 advance	bottom 5 backup	paste 6 cut	undel. char 7 del char
change case 1 move 10 char	2 end of line	3 mvt char	substitute ENTER
open New Line 0 Beginning of Line		Reset select	enter

SEQUENCE NOTES

STARTING
① TYPE SEQUENCE
② ENTER # SEQS, LENGTH

COMMANDS

CONSENSUS
BASE_COMPOSITION
NWS
INVERT
DOT_PLOT

LOAD filename ... opens file
WRITE filename ... save file
PRINT filename ... sams file in print format
— / .. executes VMS command
@ — .. executes SEQ commands from file
CREATE .. creates new seq. line
DIRECTORY .. lists directory
EXIT ... leaves + save
QUIT ... leaves w/o save
CHANGE .. enter edit mode

GOLD Append puts it in other things

MOVING
ARROWS ... 1 character
WORD ... 10 characters
PAGE ... 50-100
BEG_LINE
END_LINE
GOLD # ARROW = repeat # arrow
GOLD FIND GAATTC RETURN = move to GAATTC
ADVANCE ... set direction R
BACKUP ... " " L
ctrl F ... move to next residue
GOLD ... 2nd FUNCTION

GOLD command pack=compact

FONTS
BRGCuck ... color

data: weighting mask

GOLD COMMAND WRITE -- saves

GOLD COMMAND QUIT .... QUIT # NO SAVE
GOLD COMMAND EXIT -- SAVES
GOLD COMMAND SIMILARITY/WEIGH=SEQ# [#..#]
/OUTPUT=NAME /free

CTRLZ TO COMMAND

WRITE/PHYLIP/SEQUENCES=[#..#]
-remove F

WRITE/PAUP/SEQ=[#..#]

WRITE/PAUP/WEI=2/SEQ=[2..7]

Phylip Notes

PROGRAMS

PROTPARS
DNAPARS
DNAMOVE ... reconst. ancestral
DNAPENNY ... branch & bound
DNACOMP ... compatability
DNAINVAR ...
DNAML ... max. lik
DNAMLK ... assumes mol. clock
DNADIST
PROTDIST ... makes dist. matrix
RESTML ... restriction sites
SEQBOOT ... bootstrapping
COALUKE
FITCH ... takes dist. matrix & makes tree; no clock
KITSCH ... " " " ; clock
NEIGHBOR ... UPGMA
CONTML ... gene freqs → tree
GENDIST ... gene freqs → dist.
CONTRAST ...
MIX ...
MOVE ...
PENNY ...
DOLLOP ...
DOLMOVE ...
DOLPENNY ...
CLIQUE ...
FACTOR ...
DRAWGRAM ... plots rooted trees
DRAWTREE ... plots unrooted trees
CONSENSE ... consensus tree
RETREE ... reads in tree

8/8/94

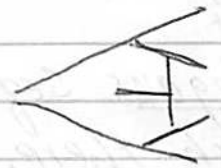
W. Fitch

Definitions

① Tree ... in math is a type of graph
... graphs have points & edges

... a tree is a connected graph ... minimally

② Network ... e.g. recombination trees



← not a minimal graph

③ Rooting - most recent ancestor of all in group

... parsimony ... uniformly weighted trees
can be rooted anywhere

... many ways to root tree
① distance from tip
② outgroup
- must be useful outgroup
.. too distant ≅ random
.. random roots the longest branch
... ∴ want one that is very close
... but not w/in group

Bifurcating trees

Trifurcating trees
- usually used to indicate ambiguity about deep branches

Data
① homologous alignment

② similarity ≠ homology

③ gene tree ≠ species tree

ⓐ orthologous (= exact) sequences are those for which there is an exact correspondence betw. gene & organism tree.

ⓑ paralogous ... not same as organism tree
- e.g. duplications (e.g. lysozyme)

Methods

1. Distances
   a. hybridization
   b. immunology
   c. sequence
      - but problems
        a. back/2ary mutations

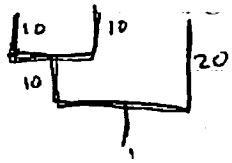
Jukes-Cantor = ~~for DNA~~

DNA: $d = -3/4 \ln(1 - 4/3\lambda)$ λ = % sites that differ
- only works if all sites variable
- all sites can accept all changes

AA: $d = -19/20 \ln(1 - 20/19\lambda)$
- but assumes all 20 aa are possible

a) UPGMA — unweighted pair group method using averages

	A	B	C
A		20	38
B			42
C			

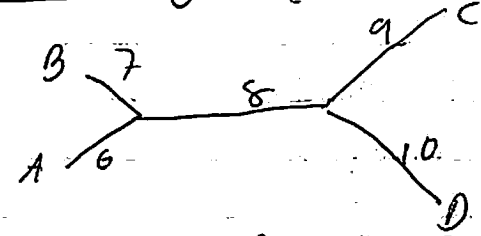


Makes all branches from a node equal (averages distances)

b) Fitch-Margoliash
- similar but allows variable length from node

c) Farris ... Wagner distance
- length = l_{min} = Σ branch lengths
- so minimizes branch lengths as long as none are smaller than distance

d) Neighbor-Joining (Fitch & ??) (Nei)



true #s = distance

	A	B	C	D
A		13	23	24
B			24	25
C				19
D				

$d_{AB} + d_{CD} < d_{AC} + d_{BD} = d_{AD} + d_{BC}$

$13+19 < 23+25 = 24+24$ = SHOULD ALWAYS HOLD

METRICS

$d(AA) = 0$

$d(AB) = d(BA)$

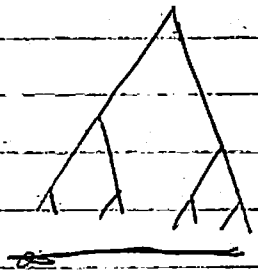
$d(AB) + d(BC) \geq d(AC)$ triangle inequality

why doesn't it always hold

- homoplasy = multiple events

2) Sequence methods

a) Maximum likelihood Felsenstein et al

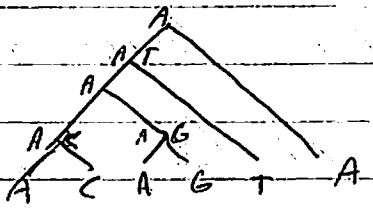


transformation matrix = prob. of nucleotide changes

- pick tree that gives maximum likelihood of getting that data

b) Maximum parsimony Hennig et al

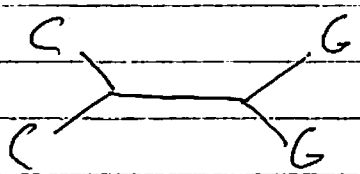
- designed for morphology



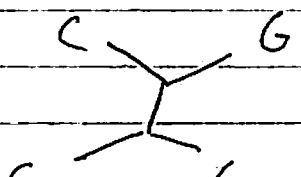
- what are minimal # changes
  - pick intersections
  - if no consensus put both
  - do again

- can get an estimate of ancestral sequence

c) Evolutionary parsimony J Lake



parsimony would link this way

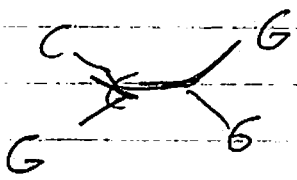


what if this is true?

@only uses transversions

- ASSUMES EACH TRANSVERSION IS EQUALLY LIKELY

- SUBTRACT OUT LIKELIHOOD OF GETTING



WHEN IT IS WRONG

- EXPECTED VALUE FOR ~~DOUBLE~~ WRONG TREES IS ZERO

- ∴ CALLEO INVARIANCE

- But need lots of data

CONFIDENCE

1) Bootstrapping or jack-knifing

jack-knifing - sample w/o replacement

bootstrapping - sample w/ replacement (one site can be done twice)
- 37% (e^{-1}) are NOT sampled
- allows scoring of ~~branch lengths~~ / clades reliability

CLADE = monophyletic group

- appears to be conservative
- prob of being correct > bootstrap value if bootstrap > 75%

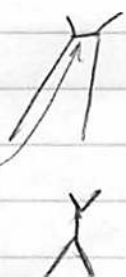
2) Likelihood
- can find value for diff. trees but NOT groups w/in tree

FAILURES IN TREES

① alignment error - most are progressive ... pw comps then grouping
- progressivity is BAD for trees
- LAKE showed order biases tree

② homoplasy - multiple events in one position

③ unequal rates.
- kills maximum parsimony if tree is because two long branches become alike especially if deep branch short



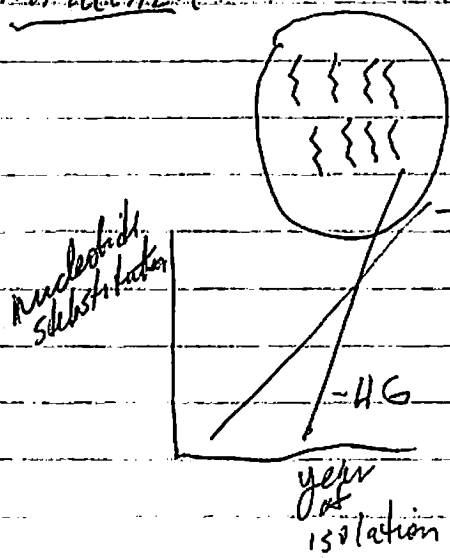
Tree Failures

4) Composition

- e.g. GC content - changes v. rapidly

Fitch - Applications

Influenza



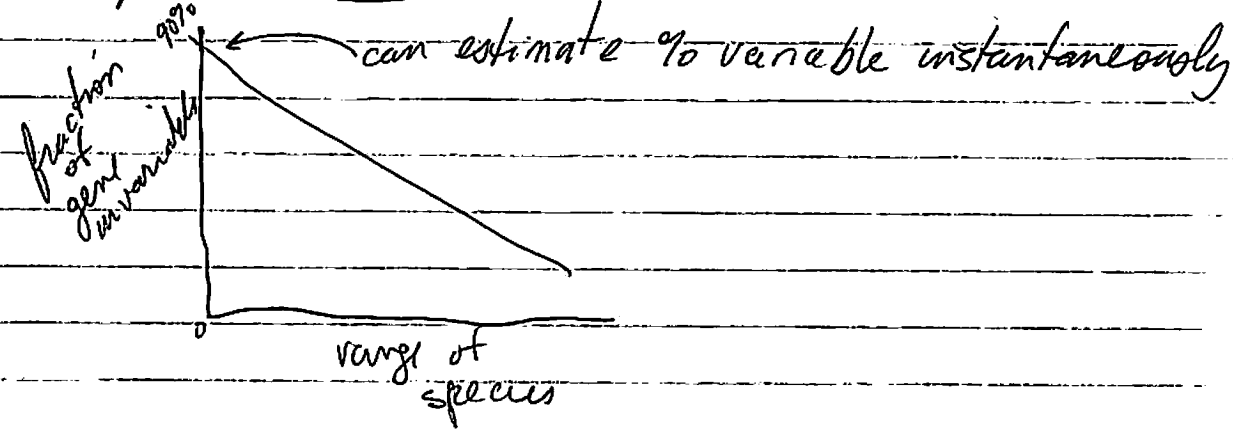
NUCLEOTIDE

ANTIGENIC SITE vs NON-ANTIGENIC SITE

nucleotide rate high in all
aa rate high only in mammals

Cytochrome C Rates

- can estimate # positions that should have NO change using Poisson
- Compare to real



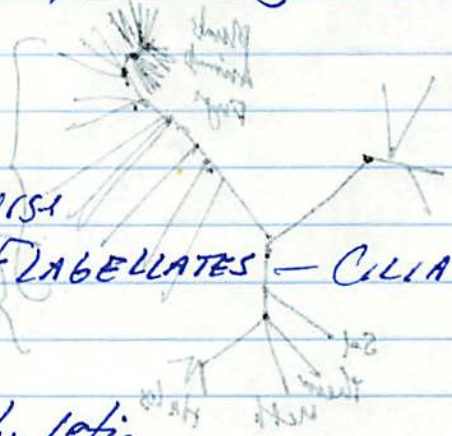
- Fixation Rate Per Covarion (sites free to vary) is relatively constant

- SOD

- ISs in E. coli

Mitch Sogin –

① Plants vs. Animals
but -- Euglena
② Protists vs Plants vs Animals
③ 5 Kingdoms



PROTISTS – V Diverse
AMOEBOIDS – FLAGELLATES – CILIATES –

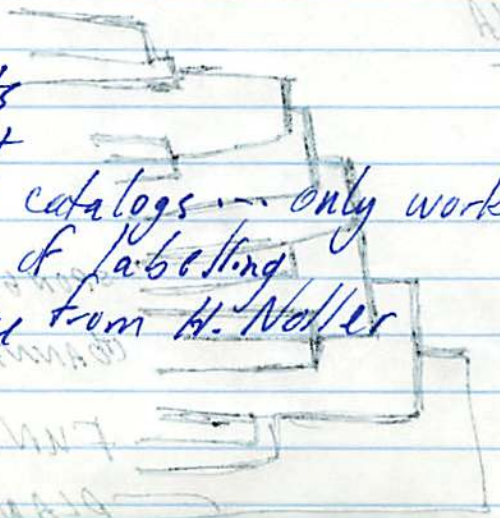
Amoebas
-not monophyletic

Chlorophytes ...
Dinoflagellates ... diff. chlorophyll that chlorophytes

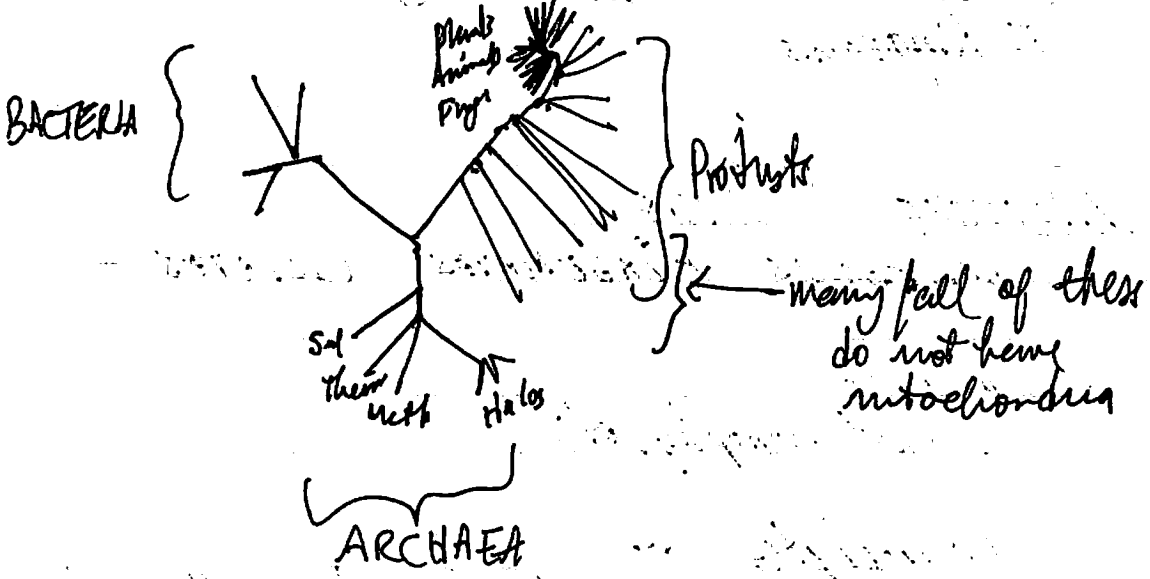
Flagellates
-Euglena
-Trypanosomes

Ciliated

TREES
-rDNA
-Carl Woese ...
-20 Fingerprints
-55 ... too short
-Oligonucleotide catalogs -- only worked for bacteria because of labeling
-1st sequence from H. Noller



© Clone
© Alignment ... does by eye
... cannot align hypervariable regions
... so he doesn't use them

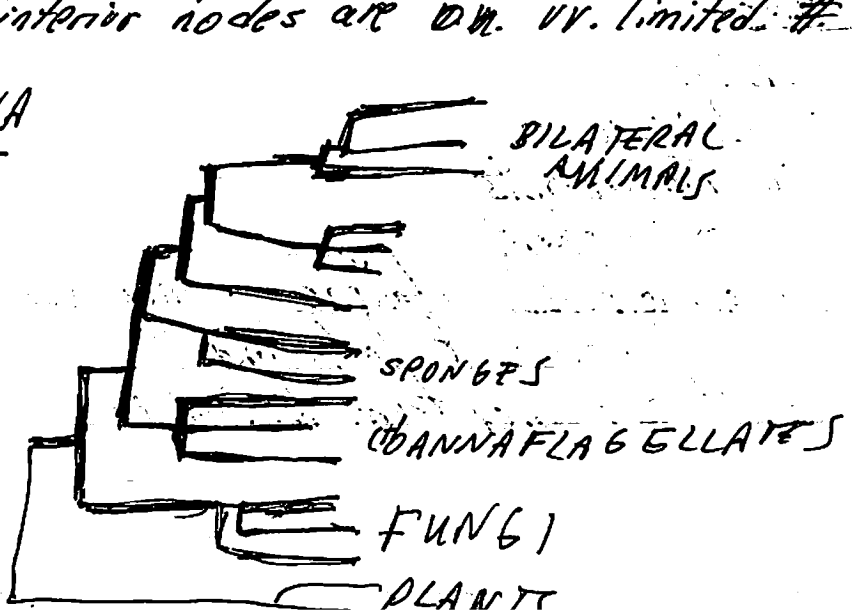


CROWN OF TREE
Animals / Plants / Fungi

SOGIN 2 Field et al - Molecular Phylogeny of the Animal Kingdom

© Animals in two clades
© But Mitch says rt sequencing was wrong & interior nodes are on. vr. limited # of positions

NEW rRNA

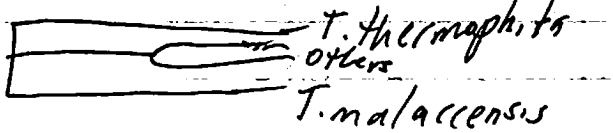


- Hasegawa also did this w/ EF.
- Baldauf & Palmer similar

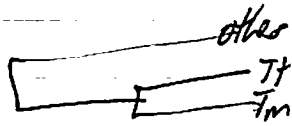
Phylogenetic w/in closely related groups
- use variable regions
- Fernandez et al. ?

Tetrahymena
- 13 species rDNA sequences
- rDNA RFLP's
- 2D cytoskeletal proteins (even more resolution)
- micronucleus has ≈ 1 rRNA gene
- similar branching w/ mt DNA

INTRON TREE



rDNA



Conclusion
- lateral transfer of introns

Introns
- type I introns commonly in most conserved domains if anywhere
- variable regions pick up long new stretches

Other
- length not correlated w/ phylogeny
- organization not correlated either

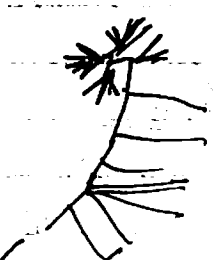
Plasmodium berghei:
- two diff. Px. units
- A gene is colinear w/ rRNA
- C gene is not colinear w/ rRNA

- Suggests C gene is pseudogene
- but not random variation
  - variation maps to variable regions
  - so is it a pseudogene?
  - discovered expression in diff hosts.

ATTINE ANTS
- leaf harvesters fungal gardens
- ~200 diff ant species
- QUESTION

J. WETTERER
- ANT & FUNGI PHYLOGENY ARE PARALLEL

DEPTH OF BRANCHING

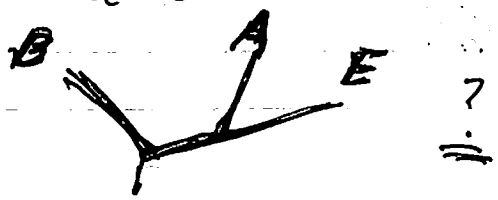


One explanation
- eukaryotes v.v. old
- but why not much biochem. diversity?
  - maybe because it hasn't been looked for

- but why not Euk fossils for v.v. long?
  - suggests definition of Euk. fossil may be bad (>1 u)
  - suggests may not form fossils

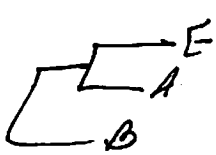
Sogin contd

Where is the root?

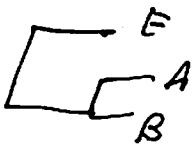


Why this tree?
- © proteins (RNA pols) more similar
- © duplicated genes as outgroup - Gogarten, Iwabe

Mitch -- doesn't like this because rRNA sequences in eukaryotes would have had to incr. at beginning of euk. evolution & then slow



prot. similarity



rRNA similarity

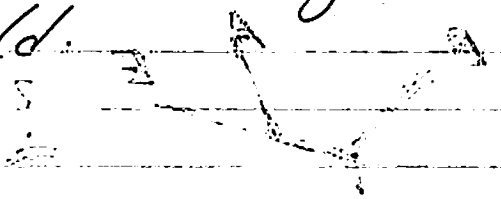
How explain?
- © Woese says progenote
  - micells
  - small genes
  - RNA based system

The progenote isolated itself from micelle world

splicing / cytoskeleton — eukaryotes — suggests cytoskeleton allowed for endosymbiosis and this then took in Archaeal genomes for its nucleus.

cell wall — Archaea, Bacteria

Thus should/if find cytoskeleton homologs in the bacterial world.



... - everything hollow

VMS

DCL = language to talk to VMS

Getting out of commands

LEAST SEVERE - ctrl-Z sends exit signal
- ctrl-C sends "cancel" "interrupt"
MOST - ctrl-Y sends interrupt

TYPE ... view files
DIR -- list
MAKE
CASE ... insensitive
CREATE
OPTIONS ... use /

TYPE
/page 1 line at a time

RECALL ... shows last command

SHOW
DEF ... shows current location [DIR.DIR.FILE]

— ... '-'

SET
DEF ... moves to new location [ ]

CREATE/DIR [.name] ... makes dir. below where you are
COPY ...
RENAME ... [

edit
*change - shows whole screen
*exit - saves

#.exe.

.com
.command files
- all lines begin w/ $!
- "!" marks comment line

eve - extended vax editor
- ctrl z - immediately out
- use DO key - commands - either says DO or TR keypad
- HELP

/out= creates file

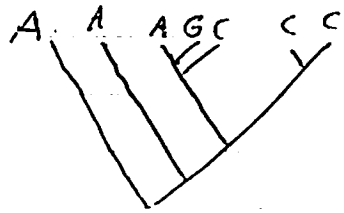
Dave's Swafford & Maddison

D. Maddison &

(1) Known phylogeny uses
  (a) infer ancestral state
  (b) types of changes / process
    - Markovian model of state change
      = prob. of change betw. nodes or betw. generations
      - depend only on state

A C T G
A
C
T
G

(2) example - MOST PARSIMONIOUS RECONSTRUCTION



MPR = most parsimonious reconstruction of ancestral state given a tree & data

- PROBLEM - parsimony underestimates #s of changes & homogenizes regions

- TESTS - HOLMQUIST & TUTENO
  HILLIS ET AL w/ VIRUS
  97% ACCURACY w/ PARSIMONY

(3) W. Maddison method
- imagines a fully symmetrical tree w/ 64 taxa, binary character
- Markovian model

	0	1
0	0.9	0.1
1	0.1	0.9

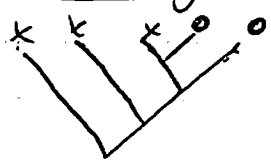
- prob. of accurate reconstruction is worse in deeper branches

How does one find MPR?
Farris & Fitch
- Sankoff et al

Objection function - some value that is optimized
- eg # of steps

Variation in MPR's
- # of changes along branch
- evolution of a character

How choose among MPR's



① Deltran vs. Acctran

Extreme much of the time but not always

Deltran = delayed transformation
= changes farthest from root

Acctran = opposite

② Cost Matrices
- change weight of diff changes
- influences MPRs

TREELENGTH - count #s of steps for MPRs

INFERRING MATRICES
① use MLE to compare matrices but hard to choose matrix
② reconstruct ancestral states & make matrix of freq of Δs

BUT PARSIMONY MAY HOMOGENIZE

SWAFFORD

CLASSIFICATION OF METHODS

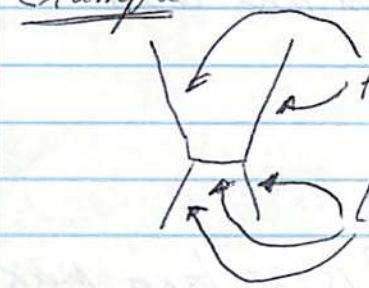
ALGORITHM
- Average-linkage
- 4-Pt. Metric
- NJ

CRITERION
- Max-parsimony
- Additive tree distance
- MLE
- Spectral analysis

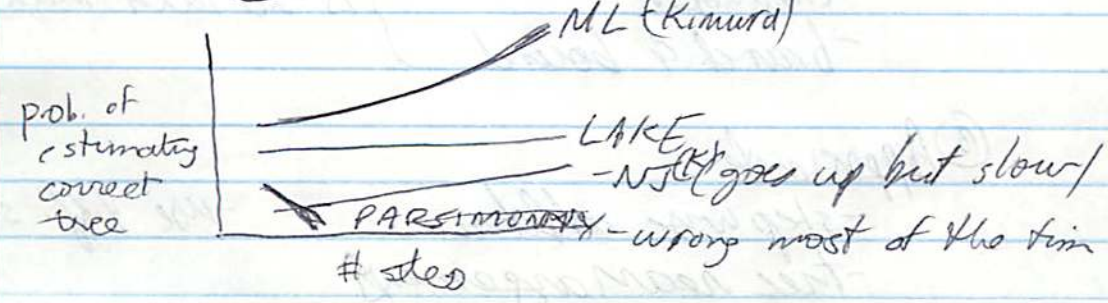
QUALIFICATION OF METHODS

- Consistency = dependent on model
- Robustness - how is it affected by violations of assumptions
- Efficiency - how much data needed to get X% accuracy
- ~~Predictive Ability~~
- Practicality

Example

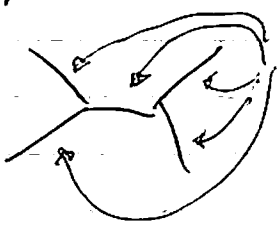


+10/1 TRANS/TRANSV.

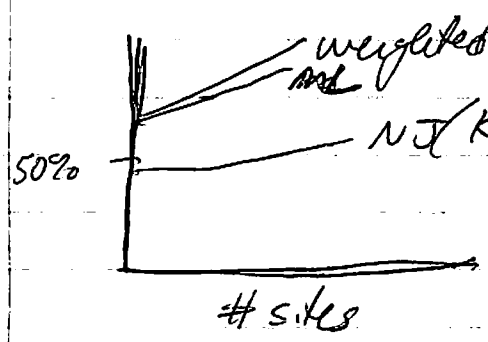


So methods are consistent but needs lots of data

Example 2



All very high rates of change + transition bias



CONCLUSION

- Diff conditions --- diff methods best

SEARCHING FOR TREES

① hillclimbing - go up when can

but - bad w/ multiple optima

- plateaus - stays in one area
- flat -

② others

Ⓐ Exact

- exhaustive
- branch & bound

} 15-20 taxa max

Ⓑ Approximate

- stepwise addition
- tree rearrangements

- use diff. starts

Exhaustive

} - get nice distribution pattern
- some information

Branch & Bound

- cuts off those searches when get too high a tree length

Stepwise Addition

- only take one path at each step
- but can get stuck in local optima

Nearest-Neighbor exchange

- too limited

Tree Bisection Reconnection

- still get stuck on islands
- solution
  - use diff. starting points

Example - MtDNA

Minimum evolution method
① construct NJ tree
② get other trees
③ evaluate trees w/ ME score

Swafford / Maddison

MacClade ⟵ sets up files for PAUP

Tree

Character
- good alignments
- black selection

- Quadra
- PowerPC

NEXUS
TEXT
PHYLIP
NBRF

Character status
- coding position
- translate

Type Edit --- matrix

PAUP

3.1.1 ...
- SINNAUER IS NEXT DISTRIBUTOR
- ~3 MONTHS
- C COMPILER VERSION

Editor
- #NEXUS
- [INFO
- BEGIN DATA

AMBIGUITY CODES = EQUATE MACROS
R ==> (A, G)

Execute
can do from edit file or directly

Search
- Branch & bound ...
- then ~~slooow~~ examine

ROOTING
- doesn't affect tree length
- default chooses 1st taxa
- select define outgroup

- can do midpt. rooting

SHOW TREE
EXAMINE TREE
DESCRIBE TREE cladogram phylogram
tree length
consistency index

PRINTING - multiple trees/page

WHY? CHANGING CHARACTER TYPES (MARICES)

- using step matrices tree slows down

GAPS
- using ?
① program assumes each char. is independent
② can code gap as a separate character

UNINFORMATIVE CHARS

① use for bootstrapping

INCLUDE/EXCLUDE CHARS
SET CHAR WEIGHTS

PAUPER PAUPEI DL

PAUPRAT

Paup & McClade II

- MacClade 3,04 #75
SINNAUER
publish.sinauer.com

- Save tree in PAUP

- Open tree in MacClade

ROOTING ... tool
CUTTING ... (option +) = cut all below

= remove branch

= rotate

= search for trees

collapse without

FOOTNOTES - ANY CELL IN MATRIX

paintbrush = force ancestors to have certain chars.

TRACE ALL CHANGES = COLORIZES BY # OF CHANGES ON BRANCH

LIST TOOL -- SHOWS Changes on a Branch

CHART
- can show just transversions Distance

PAUPDOC
PAUPDOM

PAUP

Changes

D=distance

Compare trees

L=likelihood
M=max likelihood

Save tree for PAUP

- Create new tree file
- Save as many trees as you want

Loading constraint trees

Load constraints
① backbone - only some of the taxa
② monophyly constraints

Compute consensus
- strict -- shows only those that are shared

Search
- include constraints (are compat)
- or converse constraints (are NOT compat)

must use ~~random~~ addition

- Decay Index - # of steps it takes to decay a group

~~Bootstrap~~
Exhaustive searches ~~delete~~
- shape of distribution (G1 statistic)

Bootstrap

- # replicates
- weights

1 2 3 4 5 6 7 8

10 * * * * * 10 means take 1, 2, 3 one set
& 7-8 on another in 10 of
bootstraps

Lavaan Addition Sequence

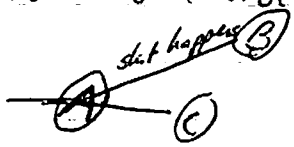
- choose stepwise addition panel
- select random
- close search --- load info

7:30pm-Lab

GARY OLSEN

GARY'S WORLD VIEW

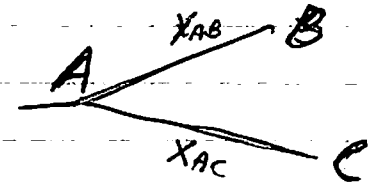
① there was a tree / history



- A is most recent common ancestor
- B & C diverge ; possibly one more than other

EVOLUTIONARY DISTANCE

- number of fixed genetic events separating two genes



$$X_{BC} = X_{AB} + X_{AC}$$

MOLECULAR PHYLOGENY

- history of a portion of the genome
- not always the organism tree
- what region to choose to get organism tree?
- should do with multiple genes

WHY?

① organism phylogeny / evolution
② gene " " "
③ process of sequence change
④ similarity searches
⑤ function

Homology I
- common ancestry of gene - similarity in excess to that expected by chance

Homology II
- common ancestry of residues
(what about when there is a deletion then reappearance)

Alignment
- expressing hypothesis about homology
① Terminal length variation

④ AUCGCG
AUCCGG
↓ - don't know what to do with this residue
- so leave it out

② Internal length variation
There no succession to the hypen

- indels - insertion in one / deletion in other
(hypothesis of a genetic event)

- multiple possibilities for some ~~residues~~ indels

- where does a gap go?

- if you don't know ... don't use that region

③ Aligning to 2ary structure
- e.g. tRNA
- may have no identical 1° structure

Why use 2ary structure

① most likely correspond to phylog. positions
② assumes unlikely to have non-pt. mutations
③ helps alignment

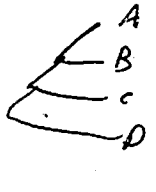
METHODS

① BY HAND
② PAIRWISE
③ MULTIPLE

- STAR -- ALIGN ALL TO REFERENCE
- BLAST -- CONSERVED REGIONS
- GLOBAL -- BEGINNING TO END

All to alignment as a tree

- PROGRESSIVE (PILEUP)
- CLUSTAL
- TREEALIGN



ⓐ generate tree from PW similarity
ⓑ align most similar
ⓒ parsimoniously create consensus
ⓓ align w/ next

- but the alignment will reproduce the history generated in step 1

- treealign can search alternative trees

CAVEATS

① optimal doesn't mean correct
② multiple alignments often assume a specific tree
③ if a region of an alignment is wrong it contributes nonsense to the tree

omit regions of uncertain alignment

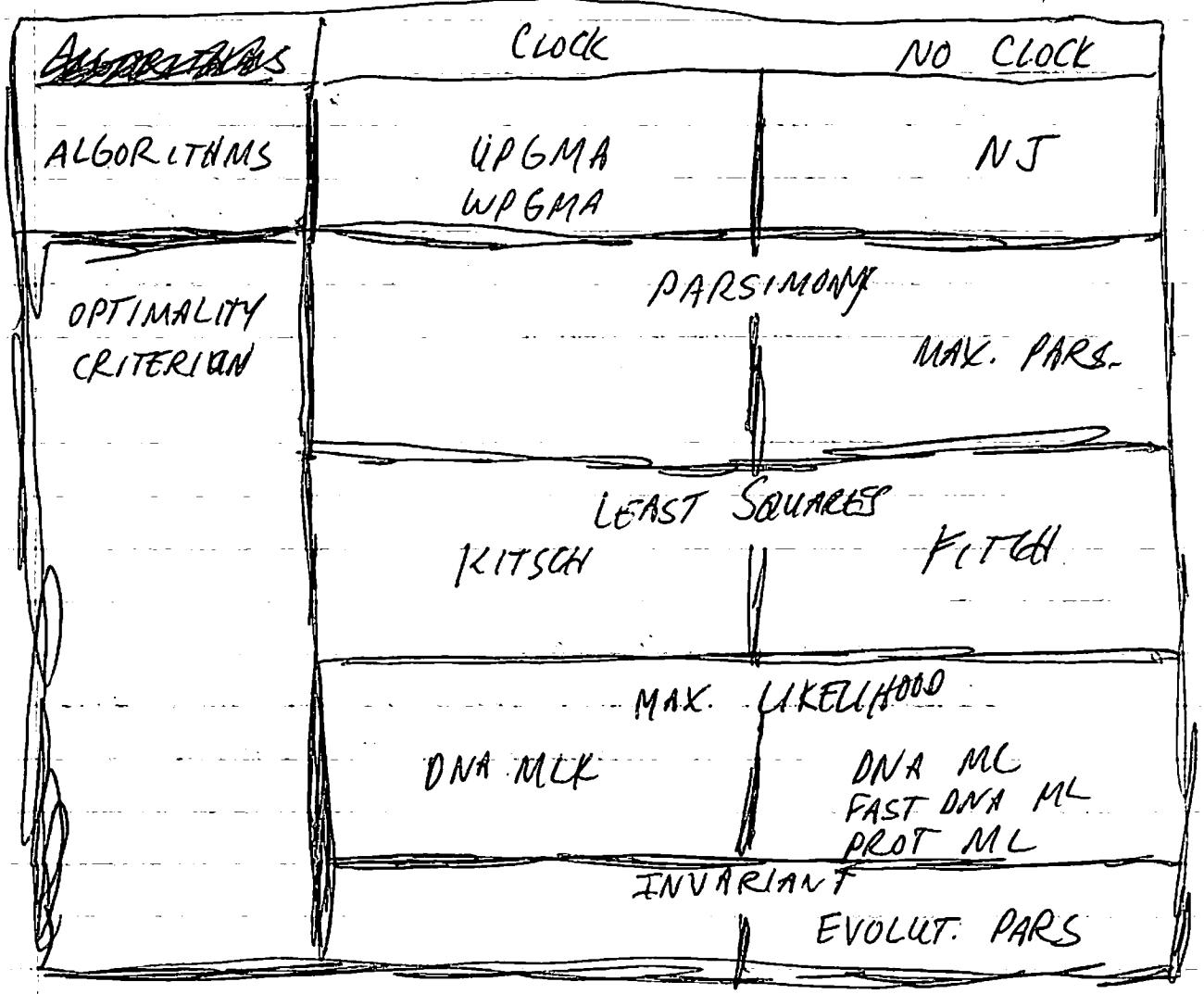
① all alignments are the same then OK

Gary Olsen II

ALIGNMENT REVIEW
- M McClure ... MBE??

METHOD
① align
② mask
③ trees

TREES & METHODS
① algorithms - just make tree
② optimality - make tree most consistent w/ data



Distance
~~don't know them~~
only observe differences

∴ ~~must make models about relationship betw observed diff & # events~~
for distance & ML

MODELS	JC	K2	Fels.	Parsimo
Homologous positions	+	+	+	+
Tree branches indep changes along	+	+	+	+
Positions indep.	?	?	?	?
Indels?	no	no	no	+/-
All sites same rate	+	+	-	-
Same base freq	+	+	-	-
Trans = Transver	+	-	-	-
Transitions all =	+	+	+	-
Transversi all =	+	+	+	-
Changes all rare	-	-	-	+

- use to segs

which the expected #
e observed

probability of giving
nces observed today

a

① groups most similar
② averages similarity to all others
③ takes next group

v. fast
- doesn't allow vary change

Non-equal distance
- still additive

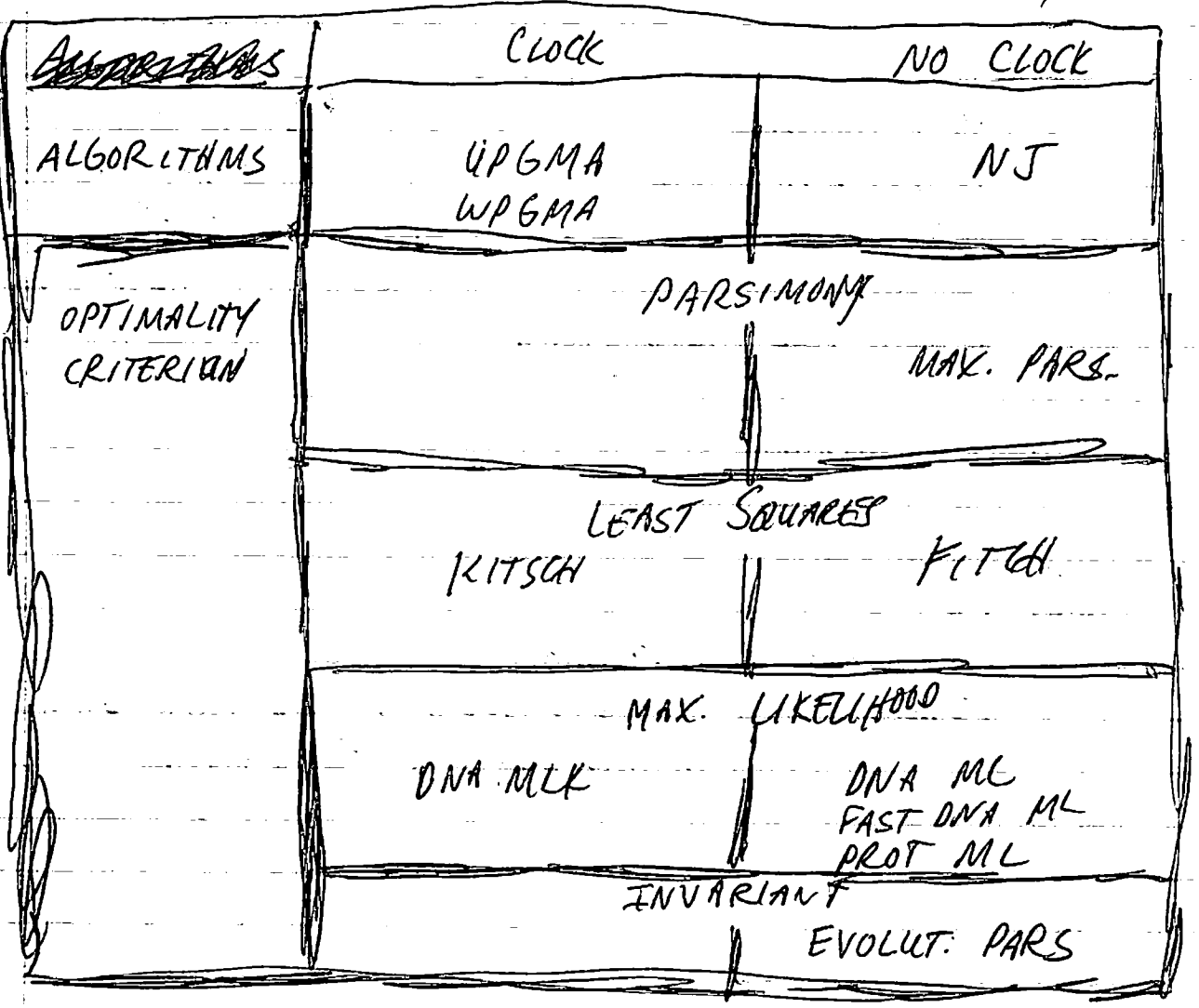
Distance
- only observe differences
- ∴ must make models about relationships betw observed differences & # of events

Gary Olsen II

ALIGNMENT REVIEW
- M McClure ... MBE??

METHOD
① align
② mask
③ trees

TREES & METHODS
① algorithms - just make tree
② optimality - make tree most consistent w/ data



Parsimony - fewest events to give rise to seqs

Least sq. dist - want the history for which the expected # of differences best fits the observed

ML - the history w/ the highest probability of giving rise to exactly the sequences observed today

TERMS

① similarity = DNA vs. prot. difference
② dissimilarity = 1-S
③ distance = 0 ~ Identical
higher = increasing

UPGMA

- cluster analysis
- groups by similarity
- tree vs. clock-like
- v. fast
- doesn't allow vary change

① groups most similar
② averages similarity to all others
③ takes next group

Non-equal distance

- still additive

Distance

- only observe differences
- must make models about relationships betw observed differences & # of events

Distance

~~① don't know them~~

② only observe differences.

∴ ~~must make models about~~ relationship betw. observed ~~diff. & # events~~

→ for distance & ML

MODELS	JC	K2	Fels.	Parsimony
Homologous positions ?	+	+	+	+
Tree branches indep.	+	+	+	+
Positions indep.	?	?	?	?
Indels?	no	no	no	+/−
All sites same rate	+	+	−*	−
Same base freq	+	+	−*	−
Trans = Transv	+	−	−	−
Transitions all =	+	+	+	−
Transversi all =	+	+	+	−
Changes all rare	−	−	−	+

TERMS

① similarity : DNA vs prot. difference

② dissimilarity: 1-S

③ distance : 0=∞

higher = increase

UPGMA

- cluster analysis
- groups by similarity
- tree is clock-like
- v. fast
- doesn't allow unequal change

① groups most similar

② averages similarity to all other

③ takes next group

Non-equal distance

- still additive

MAXIMUM LIKELIHOOD

- same models as distance
- traces evolution of tree

WHAT'S IN AN ARCHAEA GENOME

- RANDOM DNA SEQS
  - Tc cells banding protein
    - like EUKS
    - branches deeply
    - internal duplications
    - using Iwabe method

Tc cells branches most deeply

PROT-ML

- Hasegawa program
  - v. slow
  - doesn't do any branch swaps

LEAST SQUARES

$$\text{ERROR} = \sum_{\alpha = \text{seq. pairs}} \left(X_{E\alpha} - X_{O\alpha} \right)^2$$

$X_{E\alpha}$ expected, $X_{O\alpha}$ observed

WEIGHTED LS

$$\text{ERROR} = \sum_{\alpha} W_\alpha \left(X_{E\alpha} - X_{O\alpha} \right)^2$$

1. Fitch-Margoliash $= \frac{(X_E - X_O)^2}{X_O X}$
2. Cavalli & Edwards $=$ none
3. Inverse uncertainty of the differences $= \frac{(X_E - X_O)^2}{\sigma_O^2}$

DISTANCE IS FROM THE BIOLOGY

- compartmented form tree
- esp. useful for non-char distances

Tc. celer

- prot sequence has changed less than EUKS
- distance betw. 1st & 2nd half w/in species
- Tc. celer is lowest suggest it is the least randomized

Gary Olsen - Seqedit

dir data:

Sequence

#

#

- load/locus=* "file" all loci in genbank format
- name of sequence set = default file name

- seq = sequence file
- write = save
-

8/10/94

Joe Felsenstein

LIKELIHOOD & PHYLOGENY

- Ideally one should have weighted parsimony that
  - (a) takes into account non-parsimonious as well as parsimonious reconstruction
  - (b) weighted differently in branches of diff. length
  - (c) weight diff. kinds of events

He Says -- maximum likelihood does this

Maximum Likelihood - Invented by Fisher (one of the most unpleasant people that ever lived)

- Disadvantages
  - need model for change
  - computationally slower than parsimony
    - probabilities
    - Joe wrote it

- Justification

H_1 hypothesis 1 ... that the branch length is 0.1
H_2 hypothesis 2 ... that the branch length is 0
D data
$|$ symbol for given

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1 \& D)/P(D)}{P(H_2 \& D)/P(D)} = \frac{\text{Prob } D|H_1}{\text{Prob } D|H_2} \frac{P(H_1)}{P(H_2)}$$

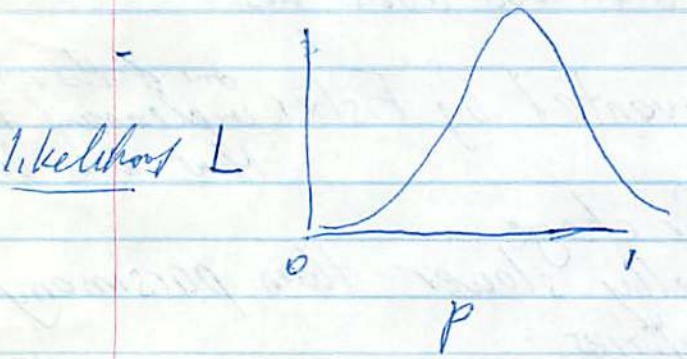
posterior odds ratio ← likelihood ratio ← prior odds ratio

- prior prob. rarely known
- but if enough data likelihood ratio outweighs it

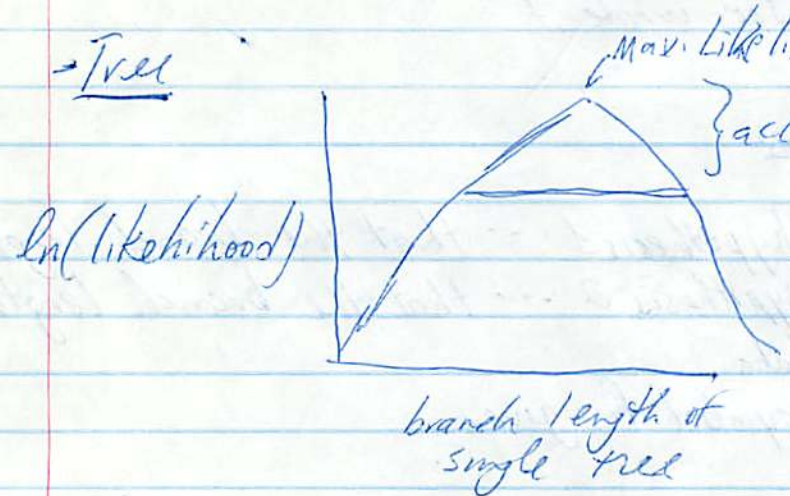
Coin tossing

= H H T H T T T H T T H p = prob heads
1-p = prob. of tail

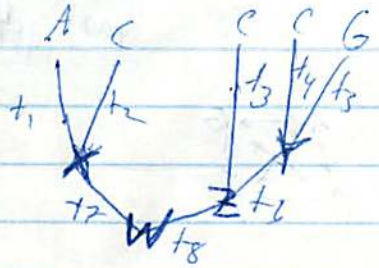
- prob of this data is $p^5(1-p)^6$



- Tree



Likelihood for Trees



① Need
- alignment
- probability model for how bases change
- trees w/ branch lengths to evaluate

② Assume
- independent ~~alignment~~ evolution of sites

Likelihood = P(D) site 1 × P(D) site 2

Probability models
- Markov models (past state not important)

③ compute prob. of data for each possible path of state changes $W \to X \to A$
$W \to X \to C$

$L_1 = P(W) \cdot P(X/W \cdot t_7) \cdot P(A/X \cdot t_1) \cdots$
(X given W)

$L = \Sigma$ for all possible X, Y, W, Z
because want likelihood for all that true for all unknowns

Origin of numerical taxonomy 1964

MODELS OF DNA CHANGE

① Jukes-Cantor 1969
- all equal
- P=change $= \frac{1}{4}(1-e^{-4at})$

② Kimura 2 Parameter 1980
- transitions diff. from transversions

③ Felsenstein
- allow unequal base freqs
- transitions vs. transversion

④ Hasegawa et al 1984

~~All models of~~
Branch length
- are product of rates and time

but not all sites change at same rate

So he has a method
- suppose there are three rates
- can do MLE for tree

but what about changes over tree

$$P = \left(\sum_{\text{paths}} \text{prob. data given tree} \times \text{path}\right)(\text{Prob path})$$

transition/transversion ratio

to use the likelihood machine and crank the handle and something will drop out

Evolution w/in species

① Likelihood based method

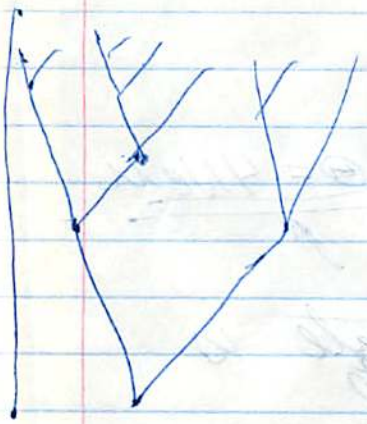
② random mating method (w/ selfing allowed)

- coalescence
- trees of genes
- depth depends on effective population size

③ what do we want to know - effective pop. size

④ trees of genes

⑤ Kingman coalescent
- randomly branching tree
- from present → past lineages come together one after another



t_k = time to coalescent is drawn from an exponential distribution

$= 4N_e / k(k-1)$ N_e = effective pop. size, k = # of branches

- time to coalescence increases as you go back because there are fewer taxa

How estimate N_e?

① Two levels of variability

② randomness of coalescence of lineages
means you ~~cannot~~ don't get similar trees

- can improve this by looking at more loci & more gene copies

③ randomness of substitutions
" " " mutations

- can use more sites if all part of same tree

Likelihood = L = Prob (Data / ⊕)

$= \sum_{\text{genealogies } G}$ Prob (Tree / ⊕) & Prob (Data / Tree) ⊕ = $4N_e\mu$

↑

∴ all they can estimate is product of size and rate

$= \sum$ Prob (Tree / ⊕) · Prob (Data / Tree)

↑ Known prior ↑ likelihood of tree

many to do

How do this?

① sum of all trees can be reduced by subsampling trees

- subsample trees by Hastings-Metropolis method
  - method samples preferentially those of interest by using likelihood

SIMULATED ANNEALING
- method gradually lowers the maximum so that it ends up on highest peak

Recombination

PROGRAM

LAMARC -- Likelihood Analysis Metropolis & COALESCE

C program --- available where phylip is

How do this?

① sum of all trees can be reduced by subsampling trees

- subsample trees by Hastings-Metropolis method
  - method samples preferentially those of interest by using likelihood

SIMULATED ANNEALING
- method gradually lowers the maximum so that it ends up on highest peak

Recombination

PROGRAM

LAMARC -- Likelihood Analysis Metropolis & COALESCE

C program --- available where phylip is

Computers

New Data Set
- 21 segs
- two tricks
- trim down to do MLE

BBS
David Swofford

PAUP ON EVOLY
- evol4

paup - need to be in directory w/ data files
help (lists commands) command name
EXEcute - (runs file) file name
set pause - will pause every screen full
set nopause
showtrees -
showtrees all

- TAXA block
  - some commands only need taxa info (eg Consensus)
- delete 7-. (delete 7-end)
- criterion likelihood

Paup

Mania – Multiple alignment never is automated

PHYLIP
SUN

infile
outfile
treefile

eg dnapars type the letter to toggle options

INTERLEAVED - like an alignment

- DNADIST - calculates distance matrix

- FITCH - fitch tree
- DRAWTREE set to tech tronics emulator

Problem

① choose masks...

DNA
PROT
} SAME MASK... ALL GAPPED POSITIONS PLUS 80

② copy files to all accounts

③ PHYLIP

- RUN PHYLIP
- SAVE EDIT

AGCT
0111

← fastest

MOLE(7)(8) - α's

EVOL3 - v. fast

WEDPM - PCR

Peter Waddell -

Calculation of probabilities of sequence patterns for a specified weighted tree.

edge weight = branch length

① Take Tree

eg

② Break up into path sets

- path set - set of non intersecting distances among even # taxa

d_{14} ⌐ = #

d_{24} ⌐ = #

d_{1234} X = path set

matrix eg purines/pyrimidines

state	1	2	3	4
1	0	0	0	0
2	0	0	0	1
3	0	1	0	1
4	0	0	1	1

p trees' Paths

step (1) | break down to more accurate dists

r smaller paths

step (2) |

s(t) pattern probabilities

STEP 1

after expressing tree as vector calc. all path lengths

STEP 2

shrinking tree generalized distances into observed distances

STEP 3

using observed mismatch counts to solve for pattern probabilities

---

LOG DET

- Lockhart et al 1994

- sites are evolving identically w/ reasonable independence } assumes all site evolve at same rate

- M is any 4×4 rate matrix

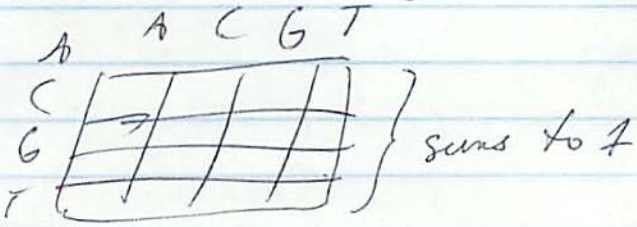
- Good at dealing w/ GC content

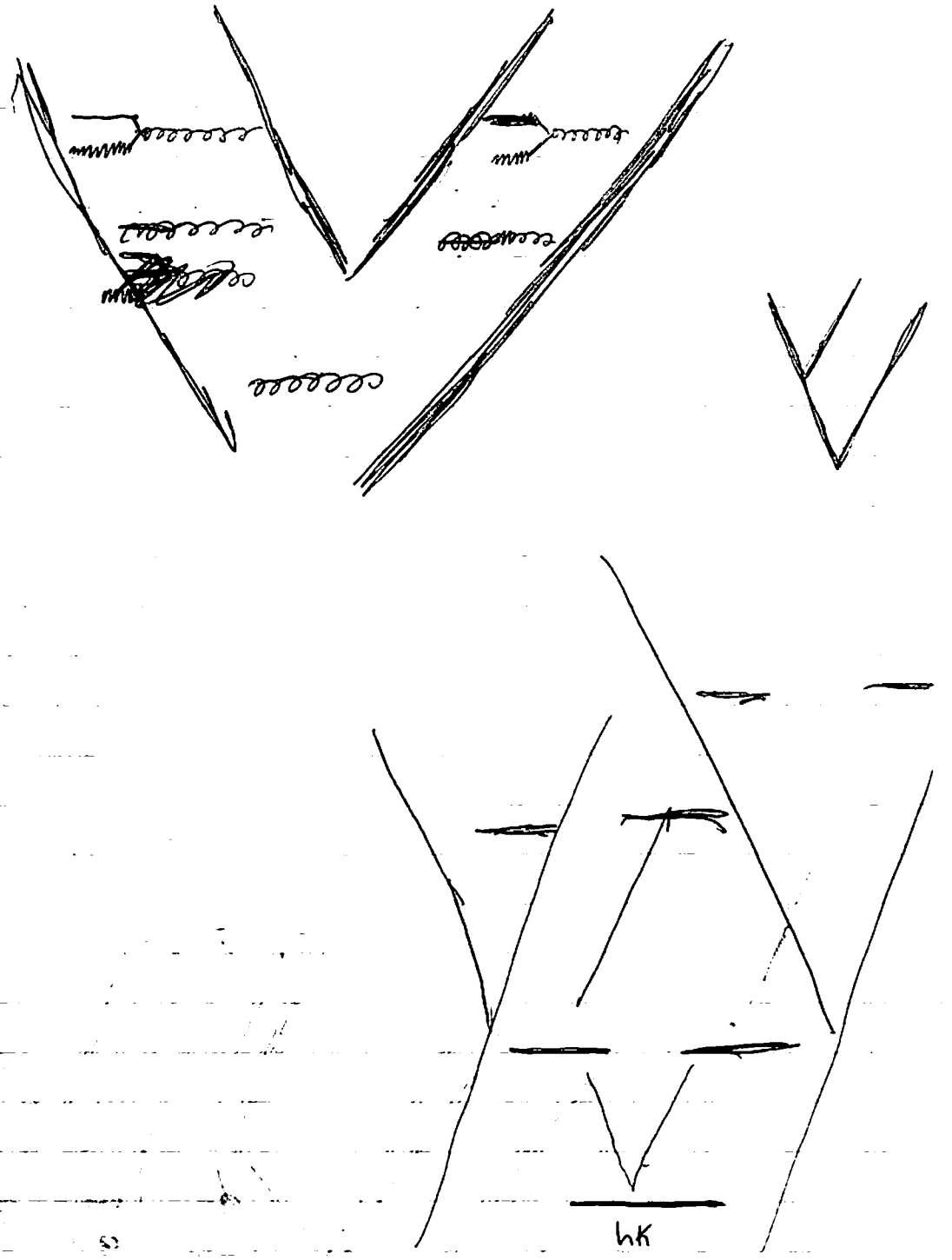
- $-\ln(\det(F_{ij}))$ = corrected dist

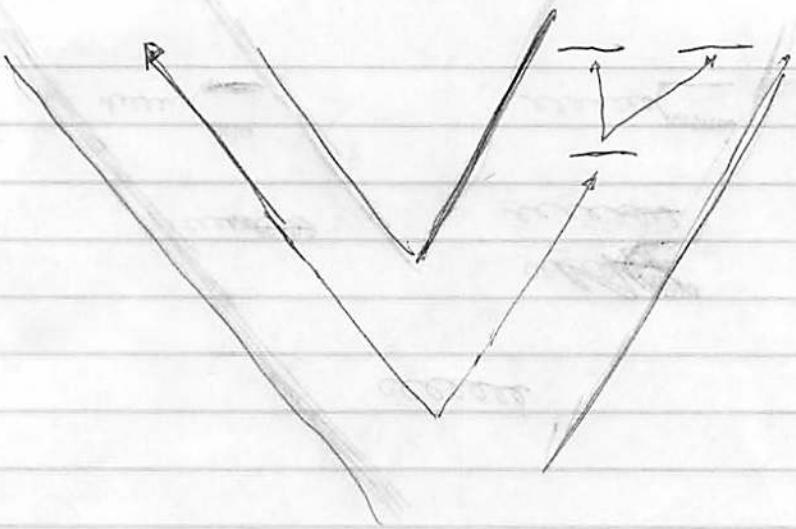
= matrix of observed divergences

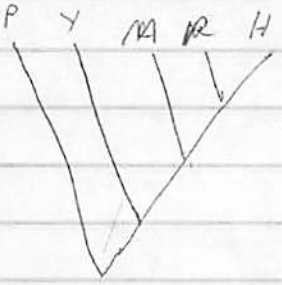
+ might have high variance

Fig =

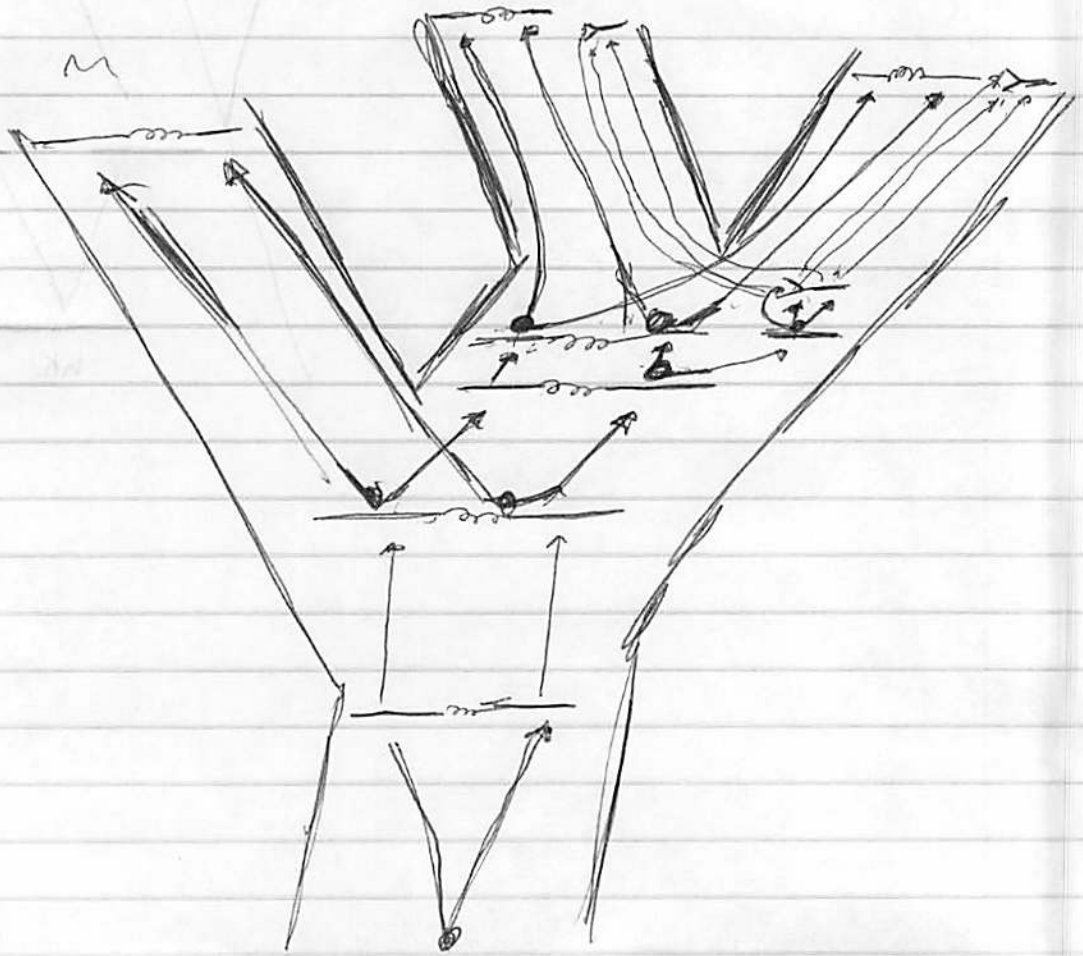



hK



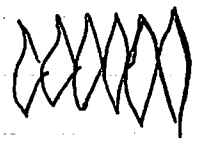


overlay L-term
overlay N-term



8/12/94

Bruce Walsh



-Coalescence vs. Drift in genealogies

- coalescence is dependent on drift and population size
- it can be decoupled from mutation

- Mutation & alleles
  - ~~①~~ infinite alleles } work for DNA but not great
  - ② infinite sites
  - ③ finite sites
  - ④ ~~Single event~~ stepwise

Drift & Mutation & Selection

① Selection
- too many possibilities to model in the real world

Selection vs. Pop. Size

e.g. Codon bias ... how happens in bacteria
- ① Eukes - strong selection
- ② Bacts - large pop. size

Substitution Rate

$K = 2Nu\,U(2N)^{-1}$ = subst rate

neutral mutation rate is dependent on population size but pop. size determines # a mutation is neutral

"Large organisms have slow generations & small pop. size"

∴ clocks that are correlated w/ time and not generations ~~both~~ generations may be so because those w/ longer generations have smaller pop. size.

Clock problems

① can say that due to pop. size diff
② can say that due to physiology

8/12/94

Steve O'Brien

Conservation Genetics

Cats -

- (like butterflies) lots of field observation
- started w/ cheetah breeding problem
- cheetahs
  - 60-70 mph ~100 yards
  - v. popular historical
  - but never really able to breed them
- ~~dad~~ - approached O'Brien
  - infant mortality v. high

- cheetahs in SA
  - abnormal spermatozoa
  - tons of defects
  - 7/90 ~~#~~ morphological abnormal

- isozymes
  - cheetahs v. v. low heterozygosity (0 variation)

- other genetic characters same
  - mtDNA
  - microsatellites - none in some
  - skin grafting
    - from self - OK
    - from cat - OK
    - from other cheetah - reject
  - MHC

- o

No more prizes for predicting rain
Prizes only for building arks

VI

Separate problem
- homogenization of immune system leads to great susceptibility to virulent pathogens

- Coryza
- virus that normally kills v. few cats killed 50-70% of the cheetahs

Hypothesizes bottleneck

When?
① hunting
② ~10,000 years ago many mammals died
- cheetahs in NA died

Rapidly evolving genes/loci
- multilocus RFLPs mini-satellites
- using human time estimate

Variation
High –
Medium –
Low
None –

Lions
① Serengeti ~ 3000 lions
② Ngoro goro crater – 100 lions
- known bottleneck
③ Lions in India - known bottleneck

Demographic collapse probability e.g. log growth problem

Same pattern for - sperm morphology
- testosterone levels

Observations were only congenital defects in

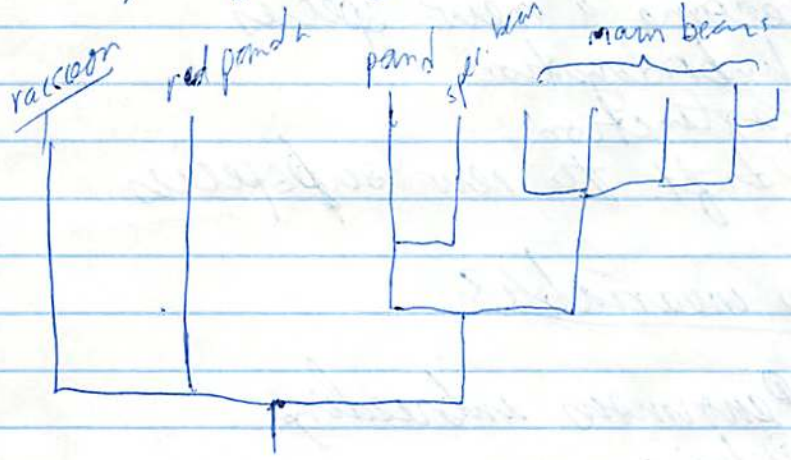
Florida Panther / Puma
- same story -- v. little variation
- Everglades -- different character
Hybrids are fine
Teams from Everglades park

Steve O'Brien

Phylogenetics for Users

① e.g. Panda

(a) morphologically ~~appears~~ somewhat like red pandas



(b) 12S mtDNA

(c) jackknifing - removing the outgroups one by one

② Cats

- mtDNA - RFLPs were too big
  - a piece of mt like DNA is in nucleus
  - chromo 2
  - integrated mtDNA segment
- Variation
  - can compare rates in mt & nucleus

USFWS

- hybrid policy
- does an endangered species that has crossed w/ another species deserve protection

Hybrid Policy
- based on biological species concept
- but should/n't be applied to subspecies

Potential fates of Sub-species
① become a new species
② hybridization
③ extinction
④ drift to new subspecies

Why so invariable?

① anonymous inbreeding
② drive

inbreeding

8/13/94

Steve O'Brien

Retrovirus

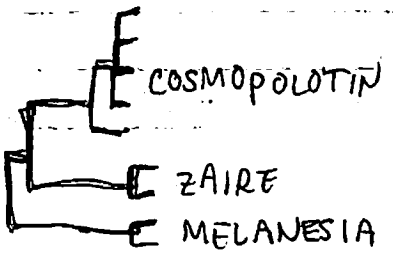
- HTLV I
- HIV
- RSU
- HepB

HTLV I

- sequence variation in env
- occurs in many parts of world
- found in many simian species

ENV TREE

- 600 bp - humans



- simians
  - viruses broken up - not specific for host
  - asian viruses deep branching
  - when include humans get interleaved branches

LENTIVIRUS

HIV, SIV

- SIV doesn't kill African Macaques
- SIV does " Asian " "
- FIV - not in Asia
  - evolves v. rapidly
  - looked at pol

CODOV

MHC.
HLA
- in MHC region
- synonymous vs. non-synonymous subst.

SUMMARY - uses of phylogenetics

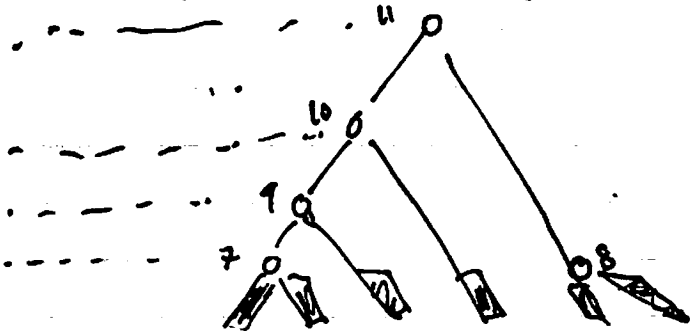
1. taxonomic recognition
2. rank endangerment
3. evaluating fly adaptation
4. hybridization
5. subspecies vs species
6. virus divergence
7. genomic vestiges

Oxf. Surv. Evol. Biol 7:1-44. '90, 8/15/94

Dick Hudson - Coalescence

Inferring Population Genetic Parameters

pop. genet models → gene trees → genetic variation



- what is topology?
- what are branch lengths?
- what is sum of branch lengths?
- what is total time?

$T(\#)$ ~ have statistical distribution
= in generations

- E(# diff. betw. pairs of seqs)
- " E(# polymorphic sites in a sample of seqs)
- frequency spectrum
  n_1 - # of polym. sites in which 1 seq has derived state
  n_2 - # " " " " " 2 " " " "

Assumptions

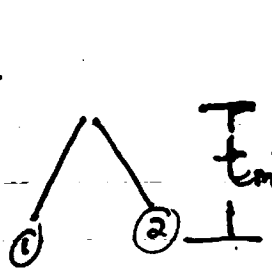
① Neutral mutation process ... assumes all variation is due to neutral mutations -- no selection on this variation

② Offspring copies of the gene differ from parents by a Poisson distributed # of neutral mutations with mean (μ) - mutation for whole region

③ Mutations in diff. indiv. & at diff. times are independent

④ 'u' is constant
⑤ ∞-site (no back mutations, multiple hits)

① Example



-GIVEN TREE

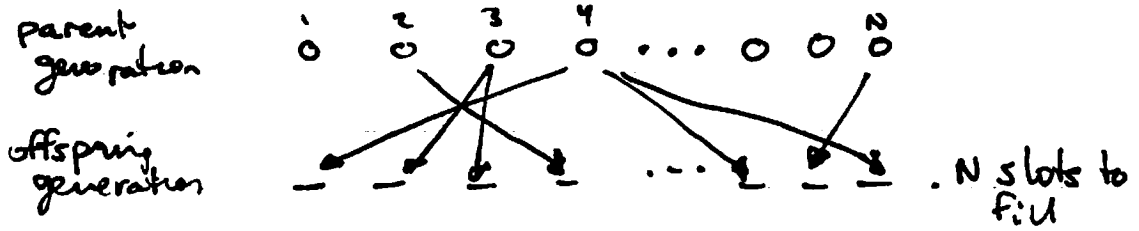
diff$_{1-2}$ = poisson dist. w/ mean 2 ut

-CAN DO WITH ANY GIVEN TREE

② Pop. Genetics

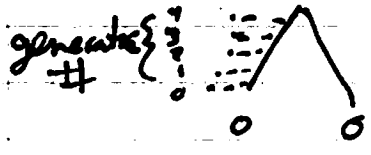
⑤ Wright-Fisher sampling (to produce succeeding generations)

-Haploid
-For diploids change everything to 2N.



- Fill N slots
- sampling randomly w/ replacement (distrib. of offspring is Poisson w/ mean 1 if N is v. large)
- many generations

⑥ Sampling



$P(t_{mrca}=1) = \frac{1}{N}$

$P(t_{mrca}>1) = 1 - \frac{1}{N}$

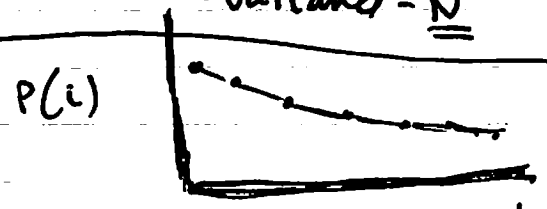
SQ

-if N is large $P(t_{mrca}=1)$ is low

$P(t_{mrca} > t) = \left(1 - \frac{1}{N}\right)^t$

$P(t_{mrca} = i) = \left(1 - \frac{1}{N}\right)^{i-1} \frac{1}{N}$: (prob. going back to $i-1$)(prob. of in $i-1$ of coalescing at i)

t_{mrca} = exponentially distrib.
: mean $\cong N$ generations
: variance $= N^2$



$E(\text{\#diff btw. 2 seqs}) = 2u\,E(t_{mrca})$

$= 2uN$

$= \Theta$

SEQUENCE DIFF

(A) $P(\text{2 seqs have no diff}) = e^{-2ut}$ (given t)

$E(e^{-2ut_{mrca}}) = \frac{1}{1+\Theta}$

(B) $P(\text{2 seqs have no diff})$

- follow lineages back and count likelihood of mutation ≠ common ancestor

- estimate of likelihood of common ancestor or mutation $\cong \frac{1}{N} + 2u$

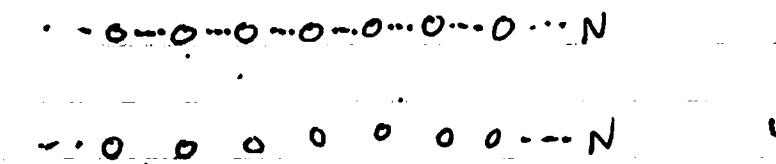
- estimation of likelihood that that was a mutation is $\frac{2u}{\frac{1}{N} + 2u}$

- estimation of likelihood that 1st event is coalescence is $\frac{\frac{1}{N}}{\frac{1}{N} + 2u} = \frac{1}{1+\Theta}$

$$\text{prob. } j \text{ diff} = \left(\frac{2u}{\frac{1}{N}+2u}\right)^{j}\left(\frac{\frac{1}{N}}{\frac{1}{N}+2u}\right)$$

$$= \left(\frac{\theta}{1+\theta}\right)\left(\frac{1}{1+\theta}\right)$$

SAMPLE SIZES BIGGER THAN 2



what is prob. if we have N indiv. that they all have distinct parents one generation back?

N = 4

= (Prob that 1 distinct from 2)(Prob. that 3 dist. from 1,2)
(Prob that 4 dist. from 1,2,3)

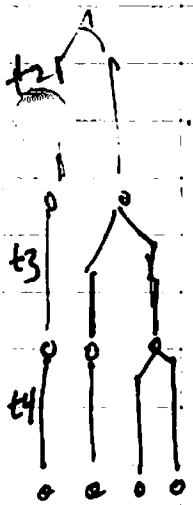
$$= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\left(1-\frac{3}{N}\right) = 1-\left(\frac{1}{N}+\frac{2}{N}+\frac{3}{N}\right)+\left(\frac{a}{N_2}+\frac{b}{N_2}+\frac{c}{N_2}\right)$$

$$= 1-\frac{\binom{N_2}{2}}{N} \quad \text{N choose 2}$$

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

suggests you can ignore this most of the time

$$= 1-\left(\frac{N!}{2!(N-2)!}\right)$$



$t(4) \cong$ exponentially variable w/ mean $= \frac{N}{\binom{4}{2}} = \frac{N}{6}$

$t(3) \cong$ " " " " $= \frac{N}{\frac{3}{2}} = \frac{N}{3}$

$t(2) \cong$ " " " " $= N$

$t_{mrca} = t_4 + t_3 + t_2$

$E(t_{mrca}) = 2N\left(1-\frac{1}{n}\right)$ n = sample size

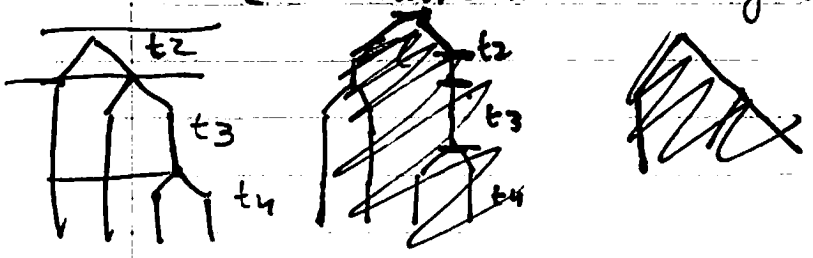
Offspring #

- Wright Fisher assumes 1 thing
- can repeat w/ $N_e = \frac{N}{\sigma^2}$ = $\frac{\#}{\text{variance in offspring \#}}$

Mutations

- E # of mutations on tree given tree = $u \cdot \sum_{\text{all}}^{\text{branches}}$ distances

$= u \cdot E(\sum \text{distances})$



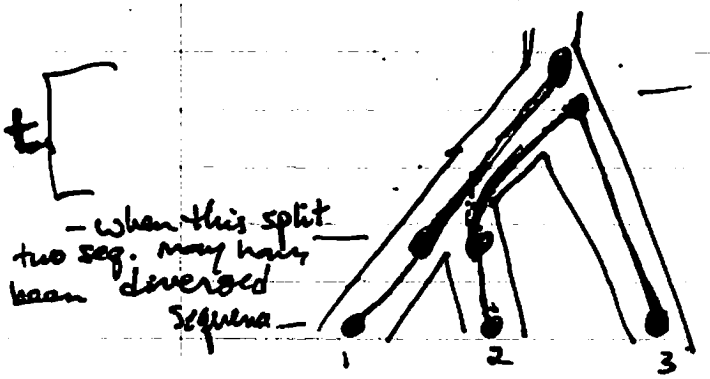
$= u \cdot 4E(t_4) + 3E(t_3) + 2E(t_2)$

total distance at $t_2 = 2t_2$

" " " $t_3 = 3t_3$

$E(\#\text{ polymorphic sites}) = \theta \sum_{j=1}^{n-1} \frac{1}{j}$

BETWEEN SPECIES



- possible that 1 & 2 may not coalesce until 2 & 3 coalesce

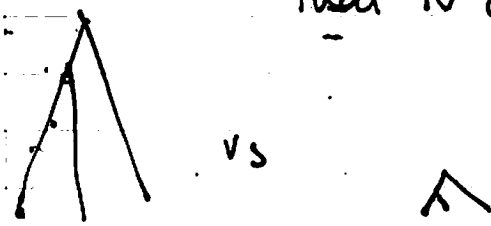
- prob. that (2,3) coalesces = prob. that 1 & 2 remain distinct & that then 2 & 3 coalesce

- prob. non-concordance of gene & species tree $= \frac{2}{3} e^{-t/N}$ - this is affected by bottlenecks, selection

can you use this for lat. transfer?

Make Tree
- uniform random variable
- $-\ln(\text{uniform}) \sim$ exponential

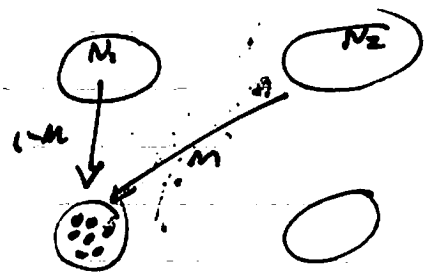
Variable Population Size
- ~~change~~ time intervals by fraction that N changes



vs

= leads to lots of rare variants because none like a star tree

Migration
- essentially use WF model
- but prob. that indiv. comes from other population is m



prob of migration = $\frac{2m}{\frac{1}{N}+2m}$

mean t of migration event = $\frac{1}{2m+\frac{1}{N}}$

$E(t_s) = 2N$ - variance not constant w/ m

Peg Riley : My Hero

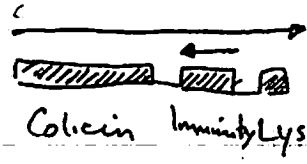
Citations

Riley 1993 MBE 10:1048-1059 } Colicin
Riley 1993 MBE 10:1380-1395 }
Sharp et al PNAS 89:9836-9840 - Restr. Modif.
Ohta & Bishop MolPhyEvol 1:87-90 } Serine Proteases
Ohta MBE 1994 in press }

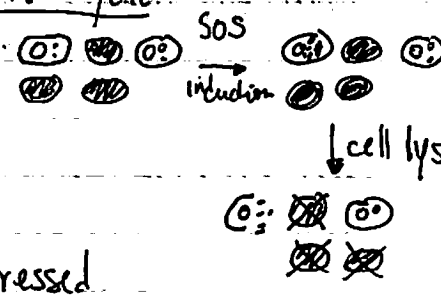
pop. gen. process —predict→ gene trees —predict→ genetic variation

inference
selection? migration? recombination?

Colicin gene cluster



Life cycle



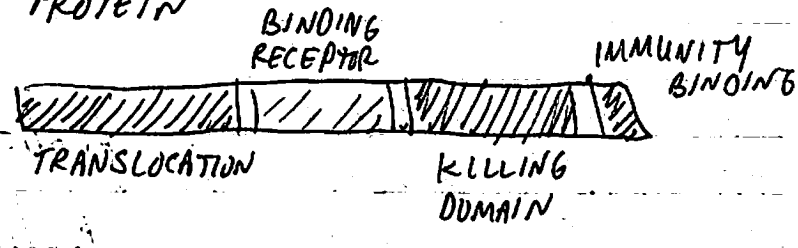
CELLS ARE RESISTANT TO THE COLICINS THEY CARRY

① Always on plasmids
② Immunity protein usually constit. expressed
- Col & Lys SOS inducible
③ Colicin recognizes cell surface receptor
- IF NO IMMUNITY PROTEIN -
either
ⓐ forms ion channel
ⓑ enters cyto. → degrades rRNA
ⓒ enters cyto → degrades DNA
ⓓ considers them analogous to viruses

Phenotypic screen

- lawn of E. coli sensitive to all colicins
- toothpick onto lawn
- grow O/N → plaques form

COLICIN PROTEIN



5 Classes
- betw. classes no a.a seq. similarity

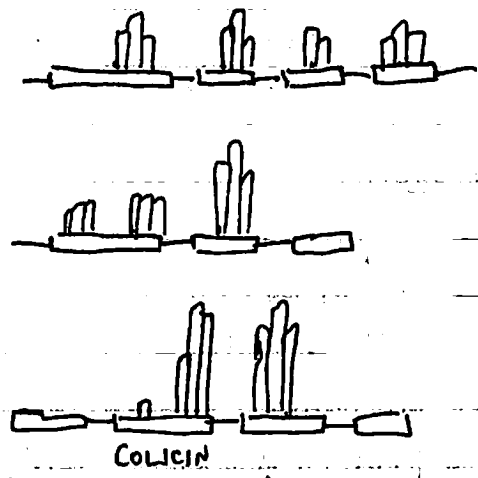
Pattern of substitution

3 Pairs of Closely Related Colicins

DNA 95% ident { ColE3 / ColE6

DNA 90% ident { ColE2 / ColE9

DNA 55% ID { ColIa / ColIb



substitutions

- suggest clustering of substitutions
- G test for homogeneity suggests clustering is real

- mutation
- 2ary structure

Why variation?

① functional constraint
- but should not have variation in synony. subst.

② recombinational diversification
- suggests NOT likely bec.
  - ⓐ too many events
  - ⓑ need 2nd coli in popln.
  - ⓒ need 2 in individ
  - ⓓ need recomb. too

- what about

- is there any info on mutation rate?

- what about the induction?

③ mutator
- genome wide (alter trans: transversion)
- locus specific
- seq. specific

④ diversification - selection
- like F. tetic virus stuff

- what is arg gst mut
- what is info on mutation
- what variation in vall would be needed to get this pattern?

RESTRICTION-MODIFICATION

TYPE I)

EcoB, EcoK, EcoD, StySB, StySP } K Family
- chromosomally encoded
- same location as A

EcoA, EcoE, CfrA } A Family
- chromosomal
- same location as C

EcoR124, EcoDXX1 } R Family
- plasmid

3 Polypeptides
- R = endodeoxyribo
- M = methylation
- S = sequence recognition

3 possibilities
- if dimethyl - fall off
- if hemimethy - methyl
- if no methy - cut

Sharp

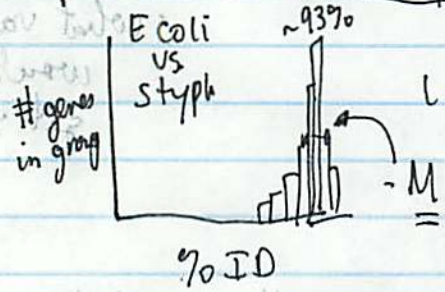
- M locus comparison
  - Intraspecific w/in family - v. highly similar
  - Intraspecific betw. " - not v. highly similar
  - Interspecific b/w " - v. highly similar

Assumes

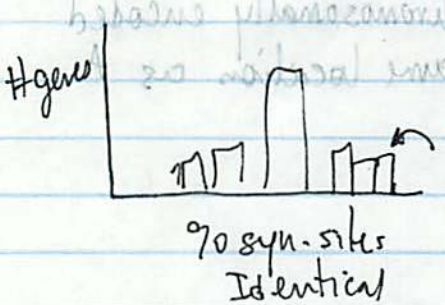
- ~25-30% ID represents homology

Interspecific vs. Intraspecific

Interspecific



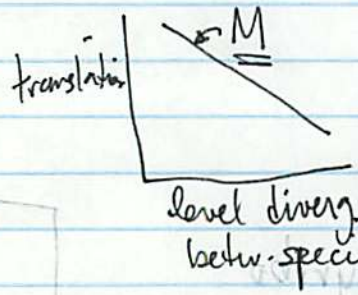
M in EC & Sty falls in this range



M comp. is here

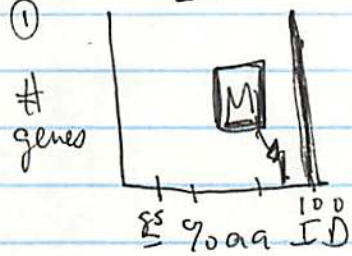
Codon bias

- highly tx genes ⟹ have high codon bias ⇒ low divergence betw. species

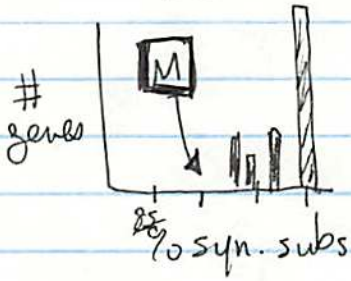


Intraspecific

① Ec.

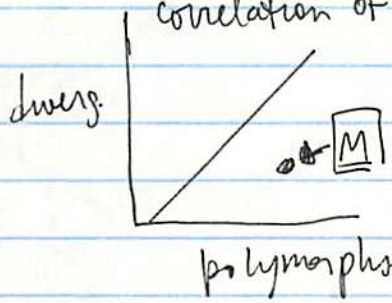


Ec



Neutral theory

⑥ correlation of divergence w/ polymorphism



② Explanations

ⓐ recombination generates variation
ⓑ selection at linked site
ⓒ selection at M-locus

} either of these could be due to selection for new RM systems

Are there sites of recognition sites?

what about variation w/in RM system?

would you expect most bias mean selection?

8/16/94

Roger Milkman Evolution of the E.coli chromosome

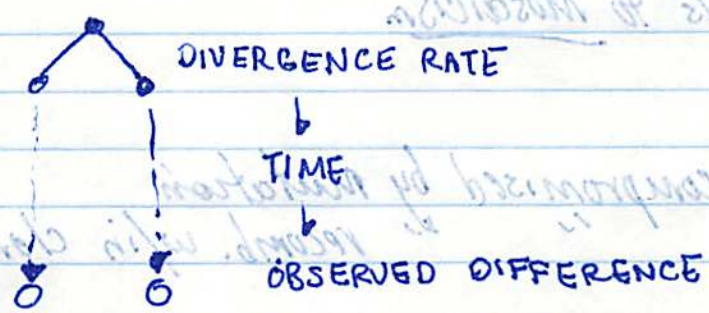
Mosacism

} suggests recombination

POPULATION GENETICS/EVOLUTION

- "gene" in gene tree is essentially a recombination unit

E. coli



Periodic Selection

- 1951 Atwood & Breyer?

① Every once in a while a rare favorable mutation arises
② Begins to spread
③ Can spread all over world - (w/o recomb. this is 1 genome)
④ Could replace all existing forms (essentially a bottleneck)

actually doesn't eliminate all competitors

Confirmed in batch transfers & chemostats

Mutation

- $\sim 3 \times 10^{-10}$/nucl.-generation
- in population of 10^9 get most mutations in short time
- ∴ rare events would be
  - double substitutions } esp. if each alone are
  - triple substitutions } harmful

Thermococcus
Sulfolobus acidocaldarius

Selection vs. Drift will affect likelihood of spread

Recombination - can counted genome wide pdic selection

- in bacteria -
  - conjugal
  - phage mediated ~50-100 kbp

- rare but lots of time

- leads to mosaicism

Clones

- not compromised by mutation
- " " " recomb. w/in clone
- can get replaced
- is compromised by recomb. from out of group
- can be nested

- can be applied to parts of the genome

Trp Operon

- multiple clones in wild pop. detected by sequence comparison
- calculates divergence times from

$$A \rightleftarrows B$$

- reversible first order reaction

SAYS E.COLI IS 100 MYO

Recombination

- suggest transformation is likely to not occur in E. coli in wild
- conjugation - does happen
- restriction enzymes - recombinogenic

8/16/94

Robin Gutell - rRNA structure

Started w/ H. Noller
- 1st 16s rRNA sequence

Methodology
A) © dot plot against self w/ highlighting potential 2ary structure

© do same thing but do comparison w/ other species

B) Folding algorithms — Woese et al 1980 NAR
- searching for stable energies
- look for compensating changes in other species

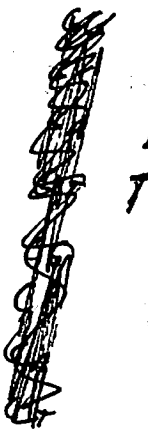
C) Red-Dot Green-Dot
- transitions get red dots
- transversions get green dots
- look for inverse patterns
- continue w/ new sequences
- see if compensatory changes have occurred multiple times

D) Covariation in columns in alignment — Gutell et al 1985
- transform nucl. pattern into ## pattern
- search for residues w/ sim. pattern
- ignores wc base-pairs

- COVARIATION MAY BE DUE TO PAIRING BUT MAY BE DUE TO SELECTION FOR ANTIPAIRING

Triplet Variation
- looking for tricorrellation

- take 2 nucleotides
and search for singles that vary
with this G G C / G C T T T T A A A A
G C A T G C A T G C A T G C A T



Variation in 16S rRNA structure

# seqs	16S	23S
Arch	120	17
Bact	~~130~~	72
Euks	~~700~~	42
Organ ~~1111~~	110	89
CPST	40	32

- w/in GROUPS ONLY A LITTLE VARIATION IN LENGTH EXCEPT IN mt & nuclear

- some w/ huge inserts in mature RNA

- Cren. hard. Hr mt DNA
- fragmented

- Physarum - edited

8/17/94

NORM PACE: USE OF PHYLOGENY - WHO'S OUT THERE?

Why know so little about natural world?

① most is non-cultivatable
- normally can culture $1/10^4$ of organisms

Environments

① open ocean
- seem to be sterile because unculturable
- but lots of stuff < 2 uM

Molecular Phylogeny

① can provide "definition" of organism by placement on the tree
② can infer physiology / biochemistry...
③ identify cultivatable model

Methods

① shotgun cloning
- water → biomass → DNA → library → rDNA
- 50% of marine picoplankton are in cyanobacteria and chloroplast group

② PCR

② Epulopsium
- giant bacteria in surgeon fish stomach
- no nuclear membrane
- selected organism
- PCR... groups w/ Gram +'s (close to Clostridium)
- daughter cells w/in cell - no binary fission and then are released

④ physical separation

④ Fluorescent probes
- phylogenetic stains
e.g - Nasonia wasp SON-KILLER
- closest to Proteus vulgaris

⑤ Survey biodiversity

High Temperature
① hydrothermal vents
② yellowstone

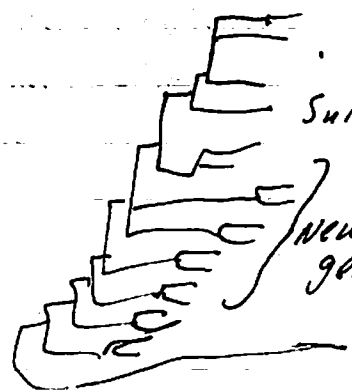
Octopus Springs
① 92°C
② lots of biomass
③ three dominant species
H₂ oxidizer → - related to Hg. pyrophilus
- related to Th. martiana

most deeply diverging bacterial group

Sue Barns

Black Pool
① 80°C - 93°C
② lots of organisms
③ single dd Nucleotide tracts
④ sequence small regions
⑤ sequence completes of uniques (>85% diff)
⑥ many new ones are very deeply branching
⑦ new sequences basis
Lakes tree

- Deep branches are unreliable
- (can support w/ signature segs



Hydrothermal vents

① Believes there is a "girdle of slime" circling the earth below the crust

② Vents
- some v.v. hot (black smokers)
- ~~most~~ 26-40°C

High Temperatures

① Contact slides ...
- if they grow there might get microcolonies

Juan de Fuca vent caps
- looks like some colonies of Pyrodictium relatives

RNAse P

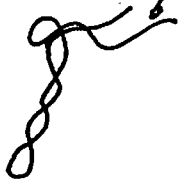
-120 aa protein P-Protein
-400 nt RNA P-RNA -catalytic

① believe the role of the protein is electrostatic because RNA can fold on own in high ionic solution

Phylogeny to Predict Structure

RNA secondary structure ... two or more continuous canonical base pairs to give rise to A-form RNA

RNA 3° structure ... single bp



RNase P

① secondary structure

② too many possible 2ary structures
-(e.g. Dot Plot vs COMPLEMENT)

SAYS YOU WOULDN'T PUBLISH A CONJECTURAL SEQUENCE

③ so ... which are correct

① predict w/ computer (free energy)

② chemical / enzyme assays
-good for 3ary structure
-but can't get inside

③ phylogenetic comparisons

"In after the essence of P"

Phylogenetic comparison
① look for covariation that preserves complementarity

RNase P
① B. subt vs. E. coli very diff to align
- ∴ 1° seq. evolves more rapidly than rRNA
- solution use more close relatives

② Solved structure somewhat -
ⓐ ... a lot of variation in structure
ⓑ ... core of about 270 molecules

Consensus
① problem ... what do you do in variable regions?
- use sequence from those w/o the region

② commonalities should point to f(x) regions
- good for mutagenesis
- similar to hammerhead
- can use as RNA probes

'ONING
① heterologous hybridization
② consensus probes

New Sequences ...
- want them for 3ary structure
- so using PCR to pull out RNase P from natural populations
- probably related to stacking
- photoaffinity crosslinking
- use RT elongation
- length variation useful to mutate

George Fox

Sequence Space

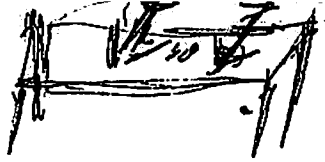
suggests the evolutionary process maintains structural constraints that are f(x) necessary

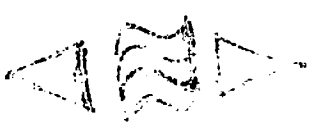
Structure Space
- a subset of seq. space consisting of only those seqs. consistent w/ biological f(x)
- can determine # of seqs that could have a particle structure

Evolutionary Rates
a) rate of events changing sequence
b) prob. that event accepted

Structural evolution
- a seq. change that leads to a change in structure space
- two types
  - equilibrium - a sequence can belong (almost) to two structures.
    - single change may favor either
  - direct change

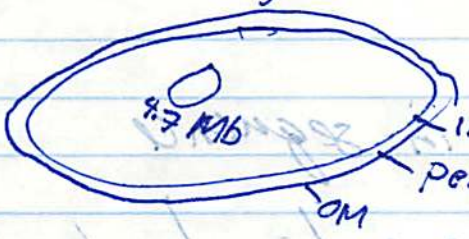
SAYS THAT "RARE" EVENTS AREN'T REALLY "RARE" JUST THAT THERE ARE LOTS OF POSSIBLE RARE EVENTS





EXPLORING SEQUENCE SPACE

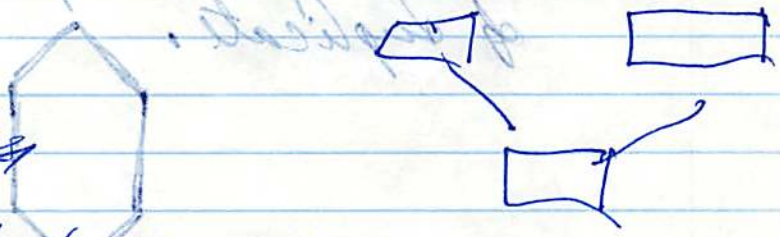
Monica Riley



Dan
Andy
Jim
Mark

chemoautotroph
CHOPKINS - negd

ECOCYC
- knowledge base



Number of ancestral sequences
- How many
- How many also in Archaes / Eukes

OPW comp for all E.coli prots (~2000)

- 899 have at least 1 partner using PAM250

- w/ 20% >> aa ID

- 97% of prot. pairs have similar f(x)

	# prots	# sets
singles	967	967
doubles	252	126
triples	93	31
4-6	384	77
>6	170	14

Doubled Enzymes

~⅔ are similar in sequence

Is it one big shbang? one big duplication?

No relationship betw. position & age of duplicate.

Laura Landweber

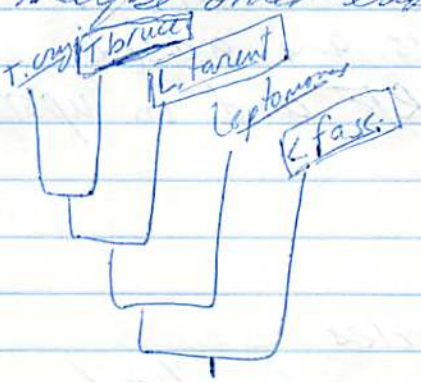
Genetic Diversity
- Early eukaryotes -
- have interesting & unique nuclear metabolism

~~Genetics~~
- Static data (DNA sequence) vs. Processed Data

RNA editing
Definition - any form of sequence modification after transcription

History
Fagan et al 1988
- Cryptic gene modified in huge amount
- Cyt c oxIII - edited

Organisms
- T. Brucei mt have editing
- maybe other euks do



Models
① editing only w/in group
② editing ancient - may reflect fossil of early mt

Why?

Distribution of RNA editing

① Herpetomonas

- Kinetoplast DNA - contains 1000's of minicircles & dozens of maxicircles

- cox III - same location but diff. amts of editing

~~Amplified~~ ~~PCR cDNA~~ w/ flanking genes

- lots of size variation

② amplify cDNA

- little size variation

③ v. v. low levels of G T in DNA
v. low C in RNA

- encoded U's are deleted
- other U's are added
- lots in ORF and in upstream

Mechanism

- guide RNAs

- small minicircles
- complementary to edited genes

- synony. subs. limited bec. still have to pair w/ RNA not just make prot.

8/18/94

what is mutation rate of guide RNAs?

- what about tRNA selection?

UG$_C^U$

AA subst

C

G
UUA L
CUX

I AUU
C
A

V

GUU

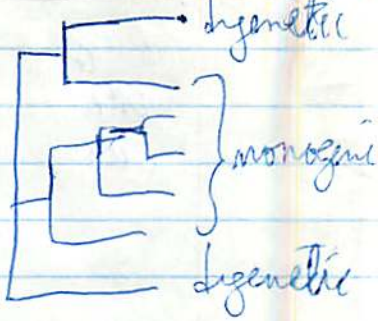
- in kinetoplasts L → V changes occur more freq. by UUA → GUU (not CUU → GUU) by a frameshift

DNA
- highly conserved
- almost all A's & G's
- Cs could pair w/ G's & confuse editing

AA
- relatively highly conserved
- variable regions correlate w/ frameshift regions
- similarity among species is low in species w/ editing

Phylogeny of RNA editing

Nuclear 16S like
Mt 9 & 12S



How would RNA editing been lost?
- perhaps reversal fix of edited gene can recombine using conserved template

Mechanism of RNA editing
- e.g. MURF4
① edited at 5' end
- 1st guide RNA corresponds w/ never edited 3' end
② next guide is better paired bec. ∴ wobble

Why edit?
① allows translational control
② can control AUG - like in humans
- usu. v. 1st site

PCR assay

Other editing
① Plant editing
prob. by deamination → C→U in land mt
U→C at some sites in fern mt

prob. by deamination A→I - brain glut. receptor

② Physarum mt 16S-like
C insertions & A insertions

Are these bizarre systems present simply bec. enzymes that could do the forward reaction were present and of course could do reverse

DNA Editing

- Actin I gene in Ciliated protozoa
  - two types of nuclei
  - micronuclei = germ line
  - macronuclei = contained cut DNA

- extensive rearrangements needed
- including reordering

Mechanism

- seems to ~~be~~ be guided by flanking regions

Paup email publish@sinnauer.com

FTP - copy ftpservers.txt (evol4.mbl.edu)
/home/mb4/dbd/Internet

DNA Patterns

Comparisons w/o alignment

- 4 base window
- count occurrence of each 4 combinations
  - normal distribution
  - compare distribution w/ other organisms
    - w/ correlation
    - compare catalog strand (Gb coding strand)
      - correlation of strands v. high
      - not due to palindromes
      - selected non-coding sequences & similar

- Correlation
  - T7 v. strange

- ~~Yeast~~ 18 organisms
  - ~~fly~~
  - ~~worm~~
  - ~~human~~
  - ~~rabbit~~
  - ~~Ecoli~~

Mt -- two strands don't look alike
yeast ets ok