Markovian Models for Genome Rearrangement Evolution

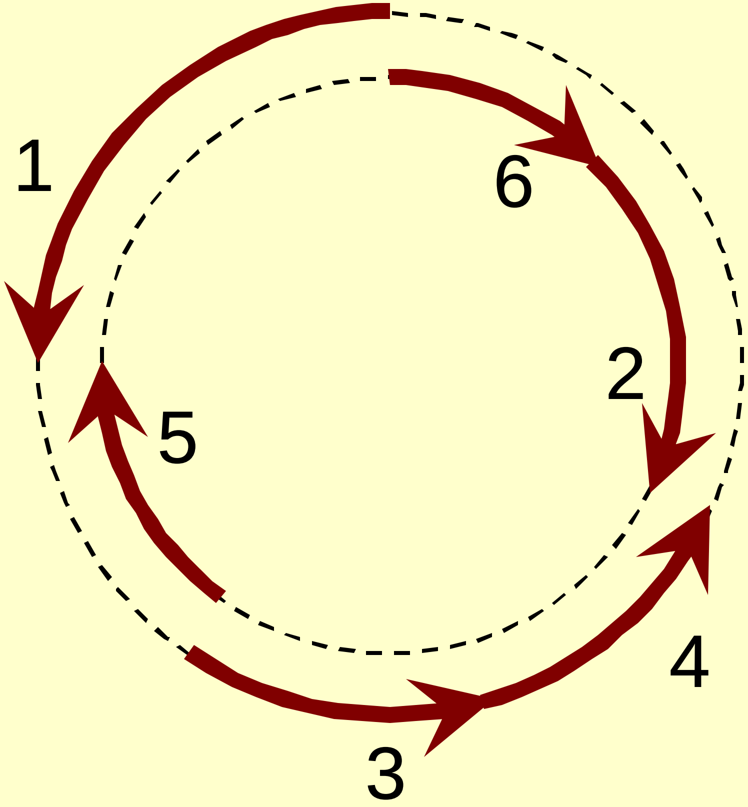
Li-San Wang

Department of Computer Sciences
University of Texas at Austin

Outline

- Genome Rearrangement Evolution
  - The GNT Model
- Distribution of evolutionary distances
  - Breakpoint distance
  - Inversion distance
- Simulation study: accuracy of tree reconstruction
- Future work

Genomes As Signed Permutations



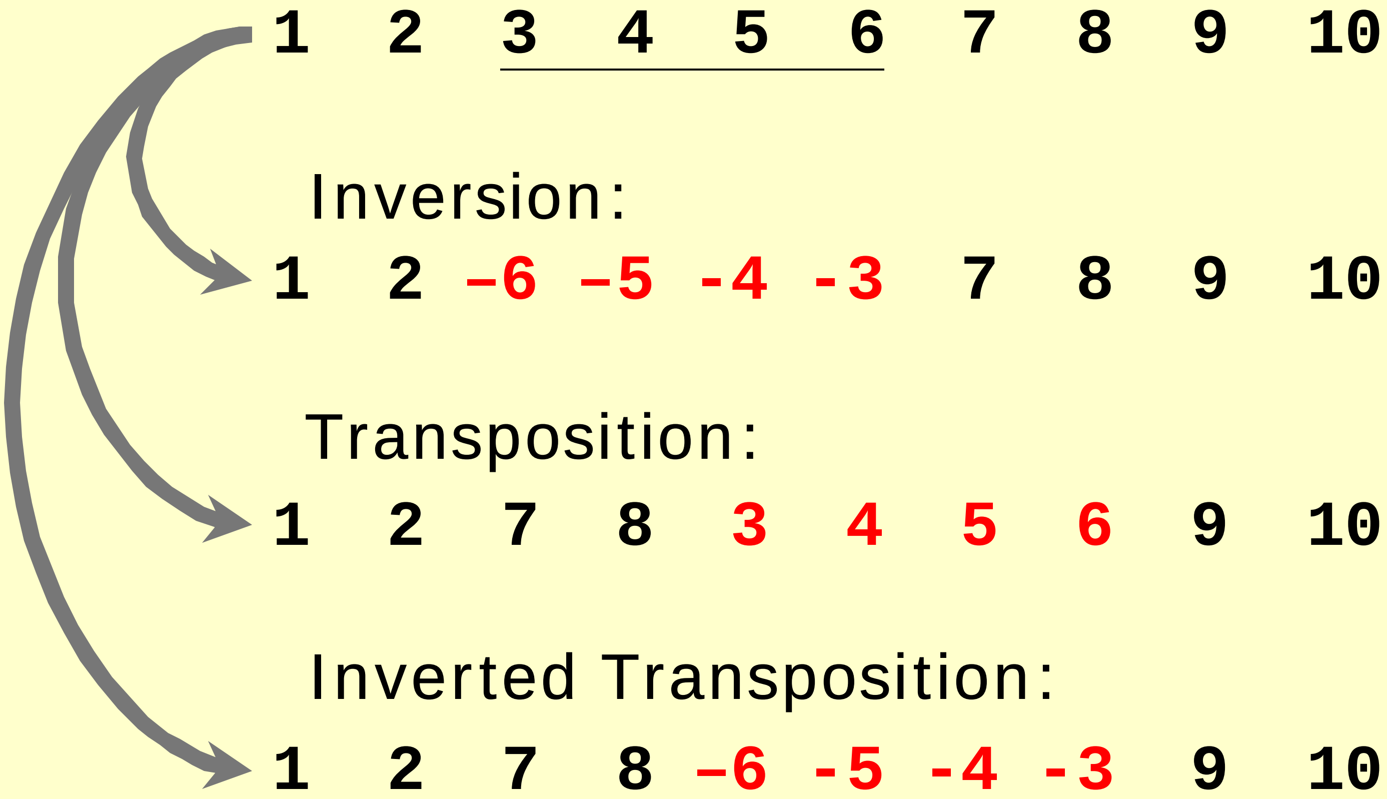
1 –5 3 4 -2 -6

or

5 –1 6 2 -4 -3

etc.

Genomes Evolve by Rearrangements



1 2 3 4 5 6 7 8 9 10

Inversion:

1 2 –6 –5 -4 -3 7 8 9 10

Transposition:

1 2 7 8 3 4 5 6 9 10

Inverted Transposition:

1 2 7 8 –6 -5 -4 -3 9 10

Our Model: the Generalized Nadeau-Taylor Model *[STOC'01]*

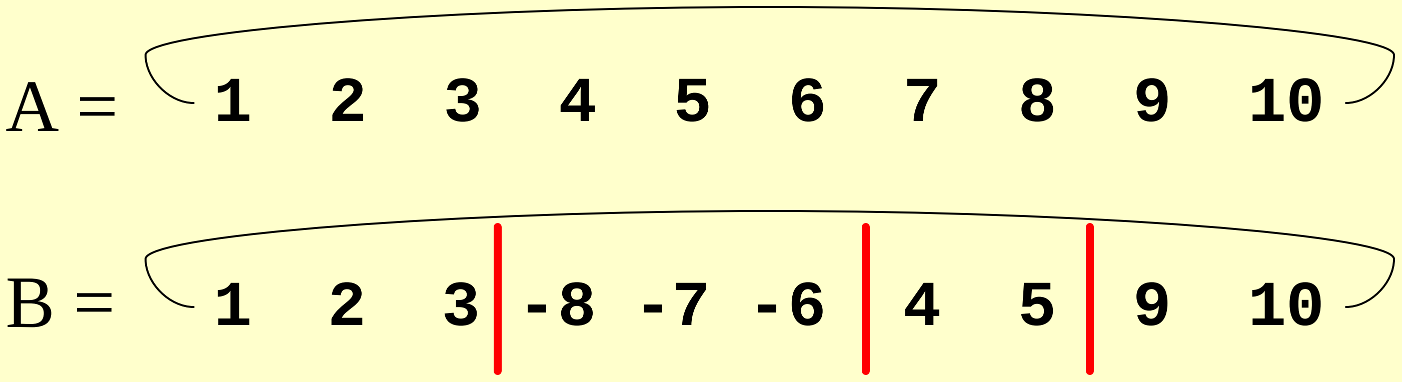
- Three types of events:
  - Inversions (INV)
  - Transpositions (TRP)
  - Inverted Transpositions (ITP)
- Events of the same type are equiprobable
- Probabilities of the three types have fixed ratio

$$\begin{aligned} & \Pr(r \in INV) : \Pr(r \in TRP) : \Pr(r \in ITP) \\ = \quad & (1-\alpha-\beta) : \alpha : \beta \end{aligned}$$

- We focus on signed circular genomes in this talk.

Edit Distances Between Genomes

- (**INV**) Inversion distance *[Hannenhalli & Pevzner 1995]*
  - Computable in linear time *[Moret et al 2001]*
- (**BP**) Breakpoint distance *[Watterson et al. 1982]*
  - Computable in linear time
  - NJ(BP): *[Blanchette, Kunisawa, Sankoff, 1999]*

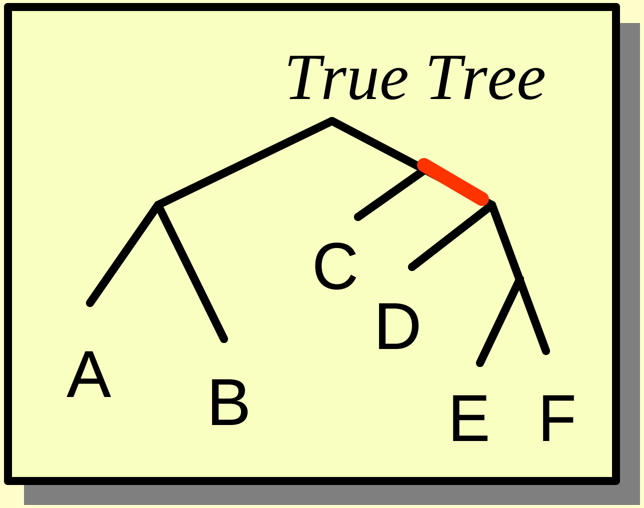


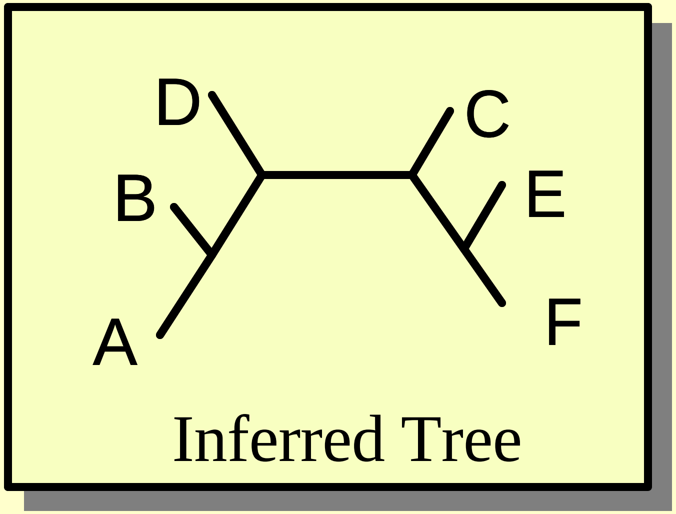
A = 1 2 3 4 5 6 7 8 9 10

B = 1 2 3 | -8 -7 -6 | 4 5 | 9 10

BP(A,B)=3

Quantifying Error

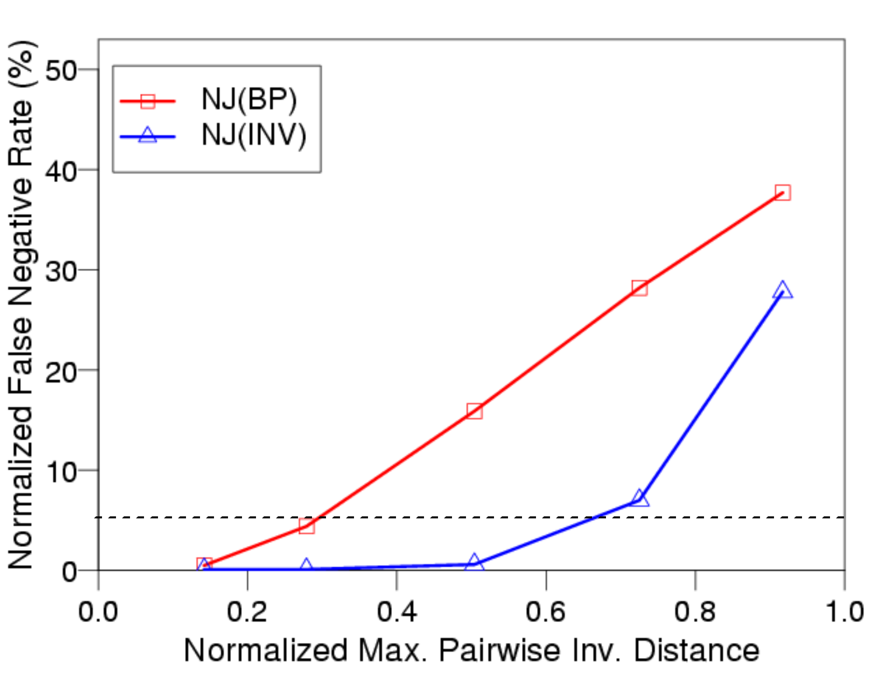




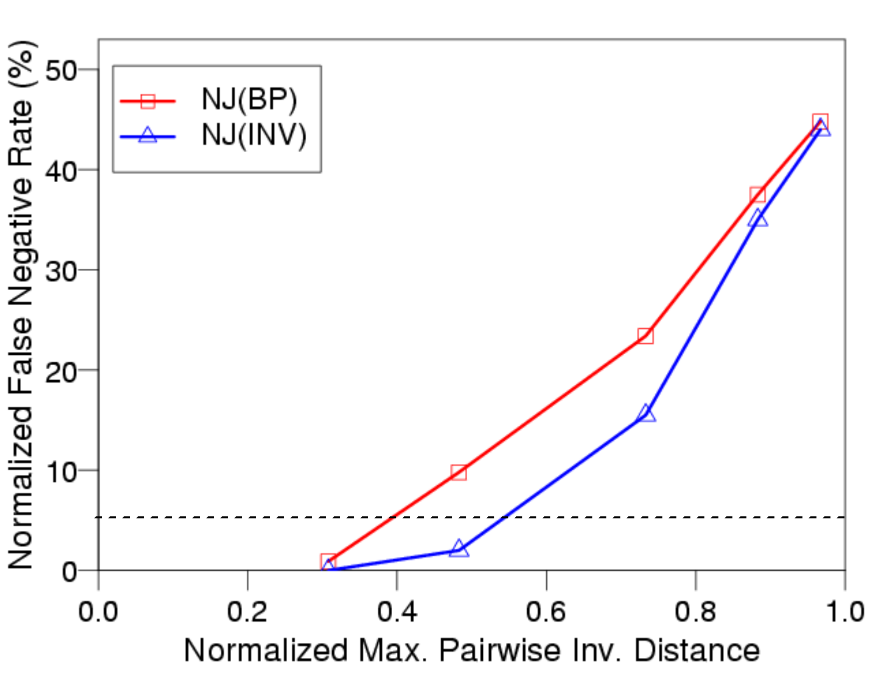
FN: false negative (missing edge)

1/3=33.3% error rate

NJ(BP) and NJ(INV)



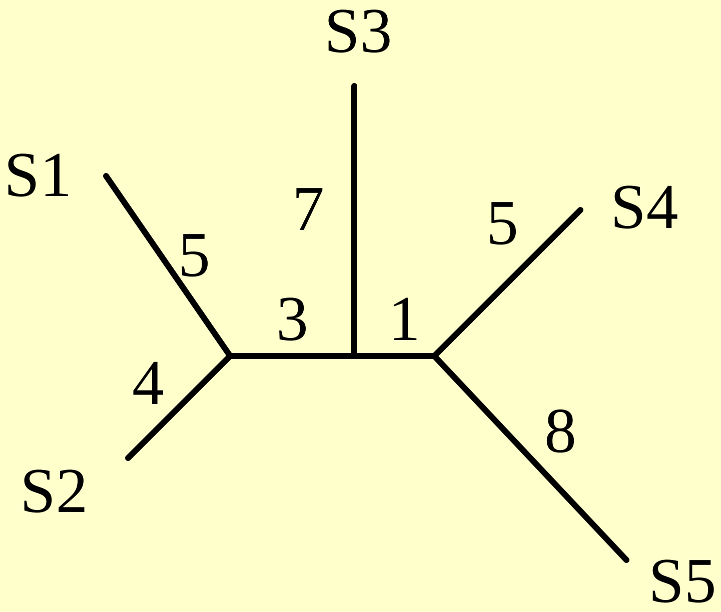
Inversion only



Transpositions/
inverted transpositions only

120 genes, 160 leaves
Uniformly Random Trees

Additive Distance Matrix and True Evolutionary Distance (T.E.D.)



	S1	S2	S3	S4	S5
S1	0	9	15	14	17
S2		0	14	13	16
S3			0	13	16
S4				0	13
S5					0

Theorem [Waterman *et al.* 1977] Given an $m \times m$ additive distance matrix, we can reconstruct a tree realizing the distance in O(m^2) time.

Error Tolerance of Neighbor Joining

Theorem *[Atteson 1999]*

Let $\{D_{ij}\}$ be the true evolutionary distances, and $\{d_{ij}\}$ be the estimated distances for T.

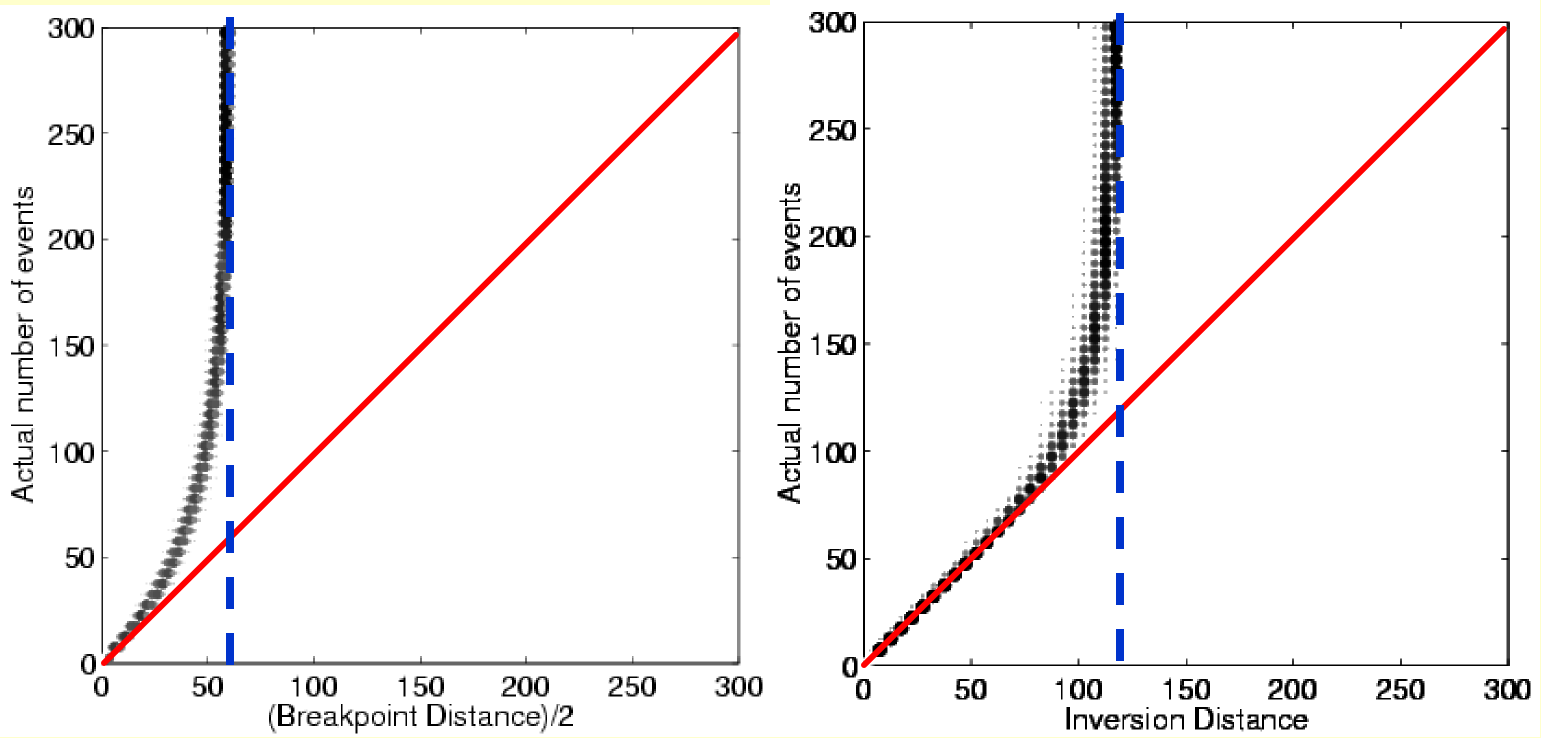
Let e be the length of the shortest edge in T.

If for all taxa i,j, we have

$$|D_{ij} - d_{ij}| < \frac{1}{2}e$$

then neighbor joining returns T.

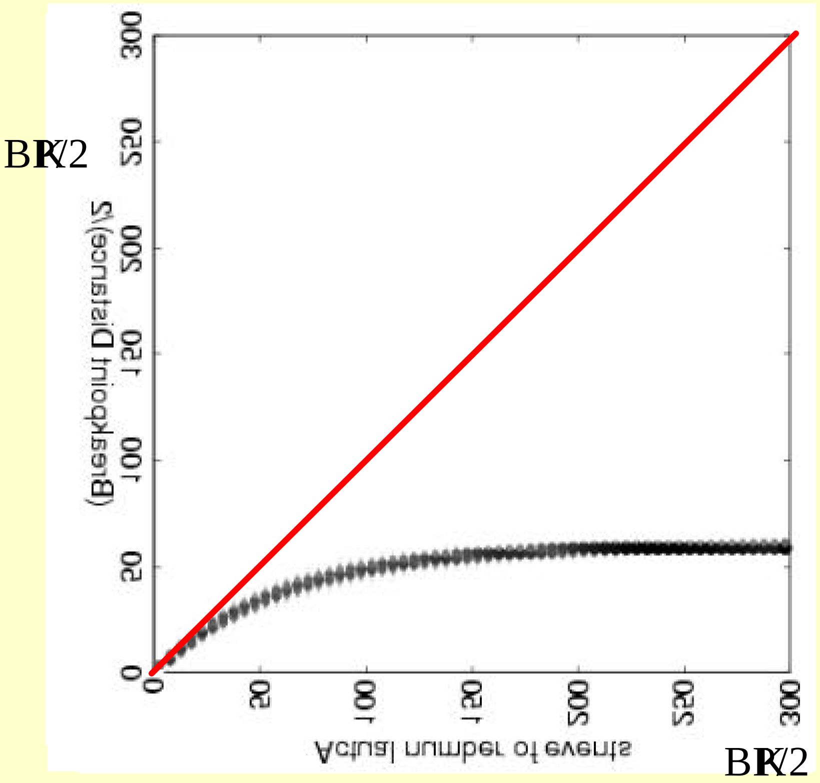
BP and INV



BP/2 vs K (120 genes) INV vs K

(K: Actual number of inversions) (Inversion-only evolution)

Estimate True Evolutionary Distances Using BP



To use the scatter plot to estimate the actual number of events (K):

1. Compute BP/2
2. From the curve, look up the corresponding value of K

BP/2 vs K (120 genes)

(K: Actual number of inversions) (Inversion-only evolution)

Using Breakpoints to Estimate T.E.D.

- Compute $f_n(k) = \mathrm{E}[\mathrm{BP}(G_0, G_k)]$
(i.e. the expected number of breakpoints after *k* random events; n is the number of genes)

- Given two genomes *G* and *G'*:
  - Compute breakpoint distance $d = \mathrm{BP}(G, G')$
  - Find *k* so that $f_n\ (k)$ is closest to *d*

- Challenge: finding $f_n\ (k)$

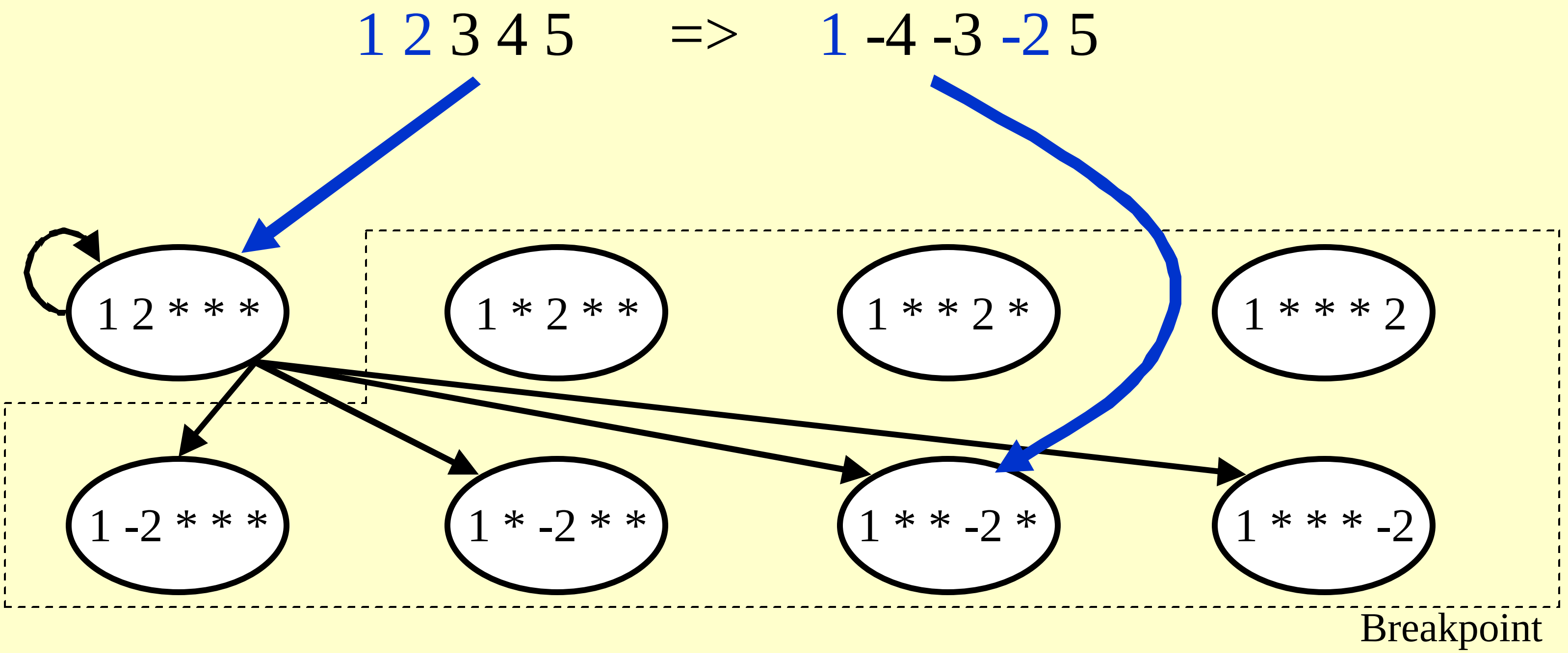
True Evolutionary Distance (t.e.d.) Estimators for Gene Order Data

T.E.D. Estimator	**Exact-IEBP** *[WABI'01]*	**Approx-IEBP** *[STOC'01]*	**EDE** *[ISMB'01]*
Based on the Expectation of	Breakpoint distance (Exact)	Breakpoint distance (Approx.)	Inversion distance (Approx.)
Derivation	Analytical	Analytical	Empirical
Model knowledge	Required	Required	Inversion-only

IEBP: Inverting the Expected BreakPoint distance
EDE: Empirically Derived Estimator

Exact-IEBP *[WABI'01]*

- Breakpoints are identically distributed: use linearity



1 2 3 4 5 => 1 -4 -3 -2 5

1 2 * * *

1 * 2 * *

1 * * 2 *

1 * * * 2

1 -2 * * *

1 * -2 * *

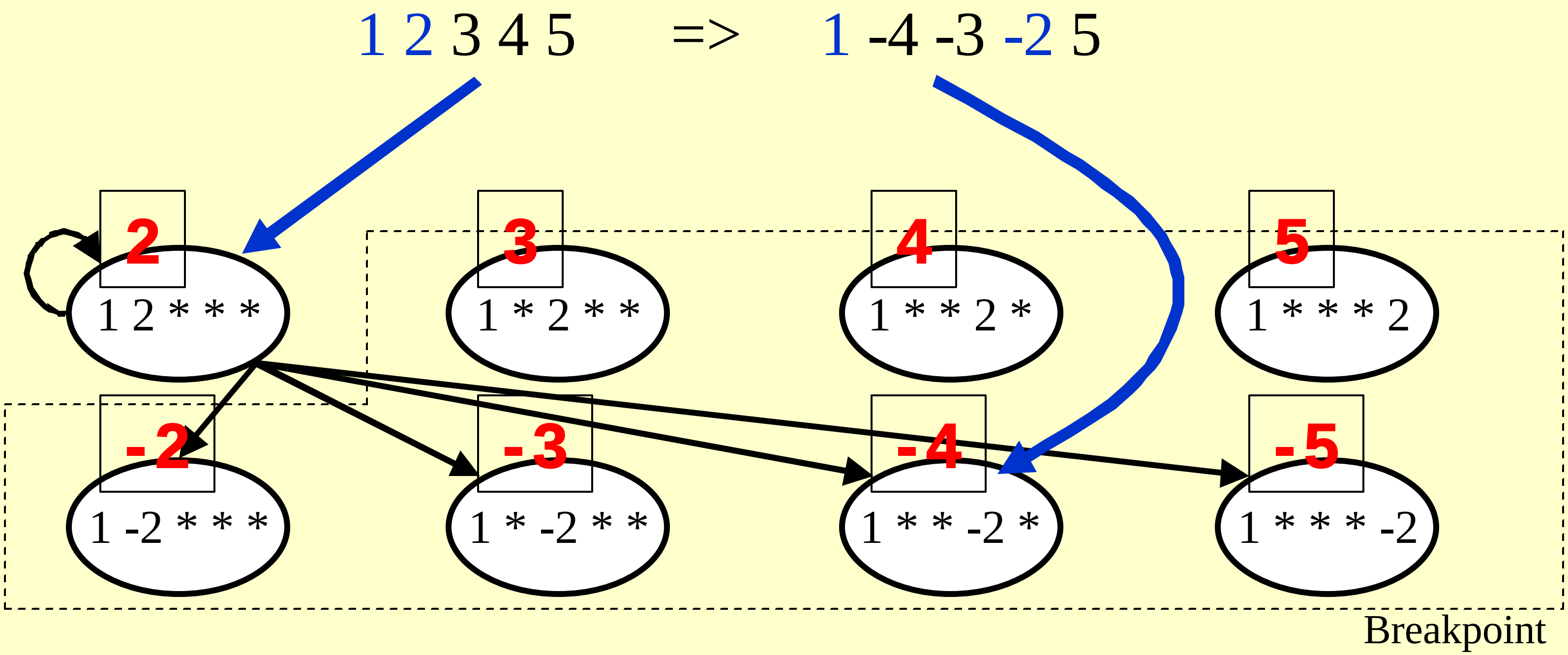
1 * * -2 *

1 * * * -2

Breakpoint

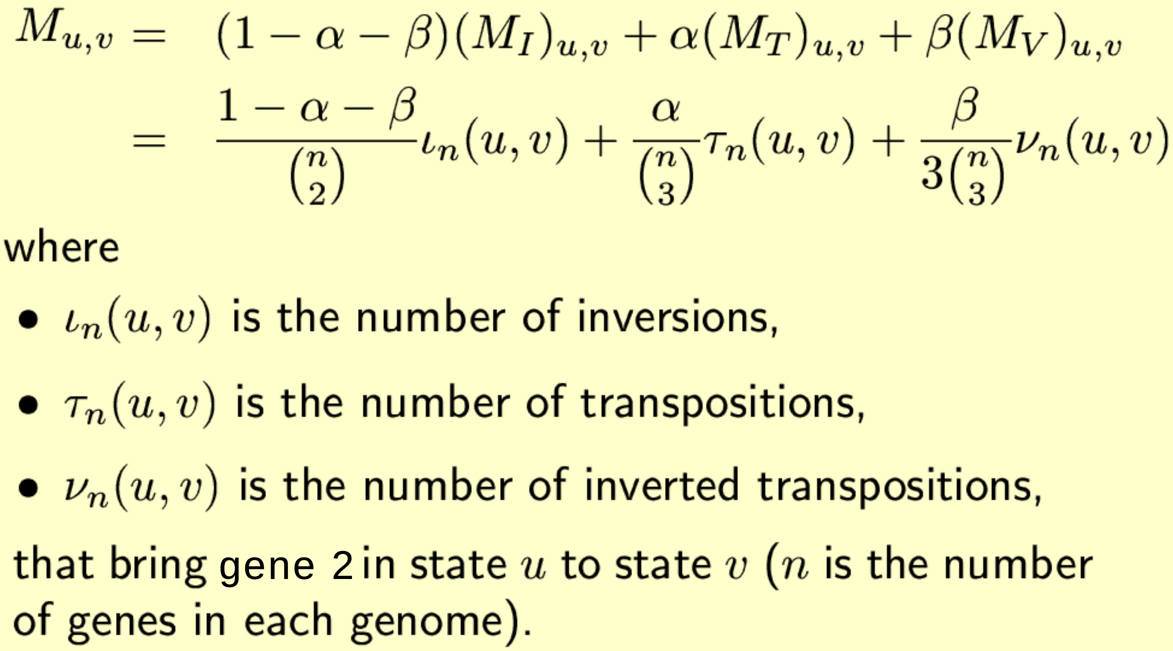
State Notation

- The sign and position of gene 2 with respect to gene 1 (at pos 1) is {-*n*, -(*n*-1), …, -2, 2, 3, …, *n*}.



Markov Chain for a Breakpoint

- Let n be the number of genes
- Each breakpoint (in particular, bp between genes 1 and 2) is a Markov process with 2(n-1) states
- We have



$$\begin{aligned} M_{u,v} &= (1-\alpha-\beta)(M_I)_{u,v} + \alpha(M_T)_{u,v} + \beta(M_V)_{u,v} \\ &= \frac{1-\alpha-\beta}{\binom{n}{2}}\iota_n(u,v) + \frac{\alpha}{\binom{n}{3}}\tau_n(u,v) + \frac{\beta}{3\binom{n}{3}}\nu_n(u,v) \end{aligned}$$

where

- $\iota_n(u,v)$ is the number of inversions,
- $\tau_n(u,v)$ is the number of transpositions,
- $\nu_n(u,v)$ is the number of inverted transpositions,

that bring gene 2 in state u to state v (n is the number of genes in each genome).

- The probability trasitional matrix is easily obtained:

$$\iota_n(u,v) \quad = \quad \begin{cases} \min\{|u|-1, |v|-1, n+1-|u|, n+1-|v|\} & (\textit{if} \quad uv < 0) \\ 0 & (\textit{if} \quad u \neq v, uv > 0) \\ \binom{|u|-1}{2} + \binom{n+1-|u|}{2} & (\textit{if} \quad u = v) \end{cases}$$

$$\tau_n(u,v) \quad = \quad \begin{cases} 0 & (\textit{if} \quad uv < 0) \\ (\min\{|u|, |v|\} - 1)(n+1-\max\{|u|, |v|\}) & (\textit{if} \quad u \neq v, uv > 0) \\ \binom{n+1-|u|}{3} + \binom{|u|-1}{3} & (\textit{if} \quad u = v) \end{cases}$$

$$\nu_n(u,v) \quad = \quad \begin{cases} (n-2)\iota_n(u,v) & (\textit{if} \quad uv < 0) \\ \tau_n(u,v) & (\textit{if} \quad u \neq v, uv > 0) \\ 3\tau_n(u,v) & (\textit{if} \quad u = v) \end{cases}$$

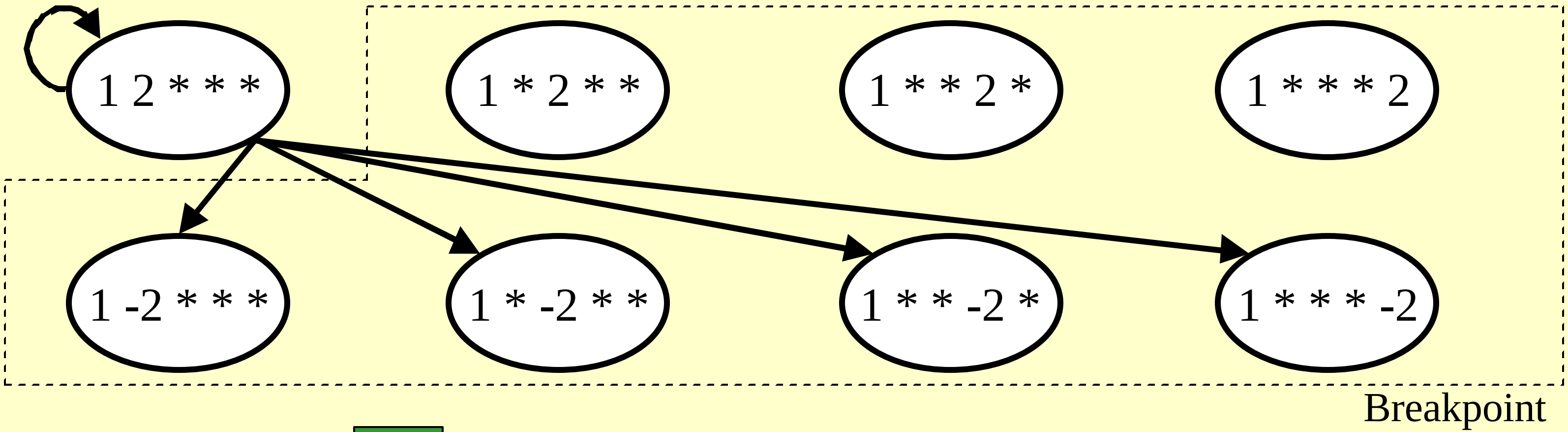
$$M_I = \frac{1}{\binom{10}{2}} \begin{array}{r|ccccccccc|ccccccccc|}
 & -10 & -9 & -8 & -7 & -6 & -5 & -4 & -3 & -2 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
-10 & 36 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-9 & 0 & 29 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 1 \\
-8 & 0 & 0 & 24 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & 3 & 3 & 3 & 2 & 1 \\
-7 & 0 & 0 & 0 & 21 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 4 & 4 & 3 & 2 & 1 \\
-6 & 0 & 0 & 0 & 0 & 20 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 5 & 4 & 3 & 2 & 1 \\
-5 & 0 & 0 & 0 & 0 & 0 & 21 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 4 & 4 & 3 & 2 & 1 \\
-4 & 0 & 0 & 0 & 0 & 0 & 0 & 24 & 0 & 0 & 1 & 2 & 3 & 3 & 3 & 3 & 3 & 2 & 1 \\
-3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 29 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 1 \\
-2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 36 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hline
2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 36 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 1 & 0 & 29 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 2 & 3 & 3 & 3 & 3 & 3 & 2 & 1 & 0 & 0 & 24 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 1 & 2 & 3 & 4 & 4 & 4 & 3 & 2 & 1 & 0 & 0 & 0 & 21 & 0 & 0 & 0 & 0 & 0 \\
6 & 1 & 2 & 3 & 4 & 5 & 4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 20 & 0 & 0 & 0 & 0 \\
7 & 1 & 2 & 3 & 4 & 4 & 4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 21 & 0 & 0 & 0 \\
8 & 1 & 2 & 3 & 3 & 3 & 3 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 24 & 0 & 0 \\
9 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 29 & 0 \\
10 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 36
\end{array}$$

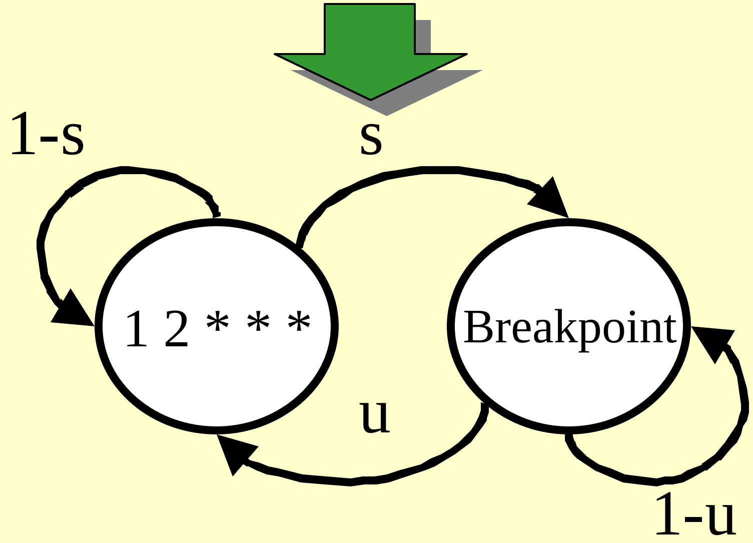
(n=10)

Exact-IEBP

- There are $2(n-1)$ states.
- The transitional matrix has dimension $2(n-1) \times 2(n-1)$.
- To compute $E[BP(G_0,G_k)]$ for k up to $2n$ takes $O(n^3)$-time. ($2n$ matrix-vector multiplications)

Reducing the State Space



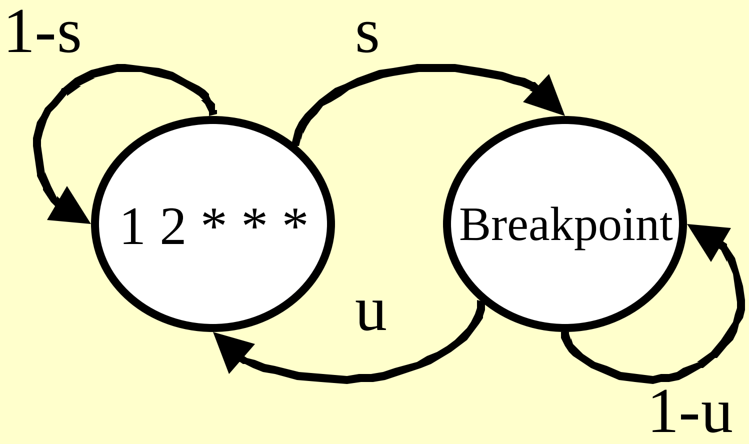


Approx-IEBP [STOC'01]:

- 2 states
- Not a Markov process
- Simple closed-form formula with provable error bound

Lower and Upper Bounds

- Under the GNT model, *s* is constant
- *u* is not constant, but has good lower and upper bounds: u_{max} and u_{min}
- Parameter *u* is small with respect to *s*

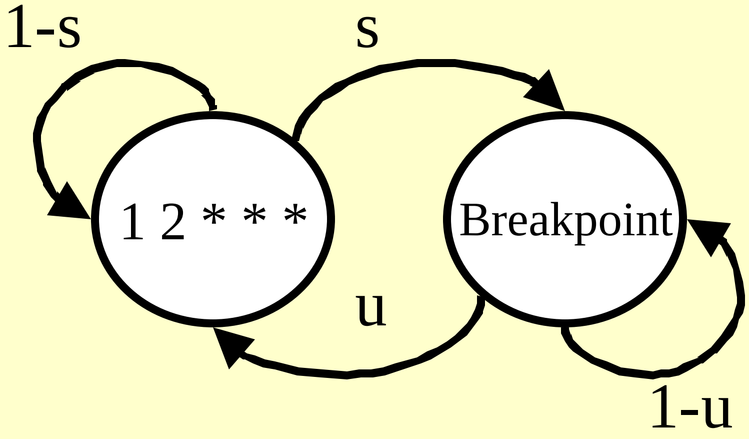


Inversion-Only Evolution

- Unsigned genome: $u_{min}=u_{max}$ -> Markov Process *[Caprara & Lancia, 2000]*
- Signed genome:

$$s = \frac{n-1}{\binom{n}{2}} = \frac{2}{n}$$

$$u_{\min} = 0,\ u_{\max} = \frac{1}{\binom{n}{2}}$$



- The two Markov chains (s,u_{min}) and (s,u_{max}) give lower and upper bounds to the expectation of breakpoint distance.

GNT Model

- $s = (1 - a - b)s_I + a\, s_T + b\, s_{IT}$
  $u_{\min} = (1 - a - b)u_{I,\min} + a\, u_{T,\min} + b\, u_{IT,\min}$
  $u_{\max} = (1 - a - b)u_{I,\max} + a\, u_{T,\max} + b\, u_{IT,\max}$

- $P_k^L \le \Pr(B_1(G_k \mid G_0) = 1) \le P_k^H,$ where

  $$P_k^L = s\frac{1-(1-s-u_{\max})^k}{1-(1-s-u_{\max})} \qquad P_k^H = s\frac{1-(1-s-u_{\min})^k}{1-(1-s-u_{\min})}$$

- $\mathcal{F}_k = \frac{n}{2}(P_k^L + P_k^H) \sim E[BP(G_k, G_0)]$

Approx-IEBP

[Wang & Warnow, STOC'01]

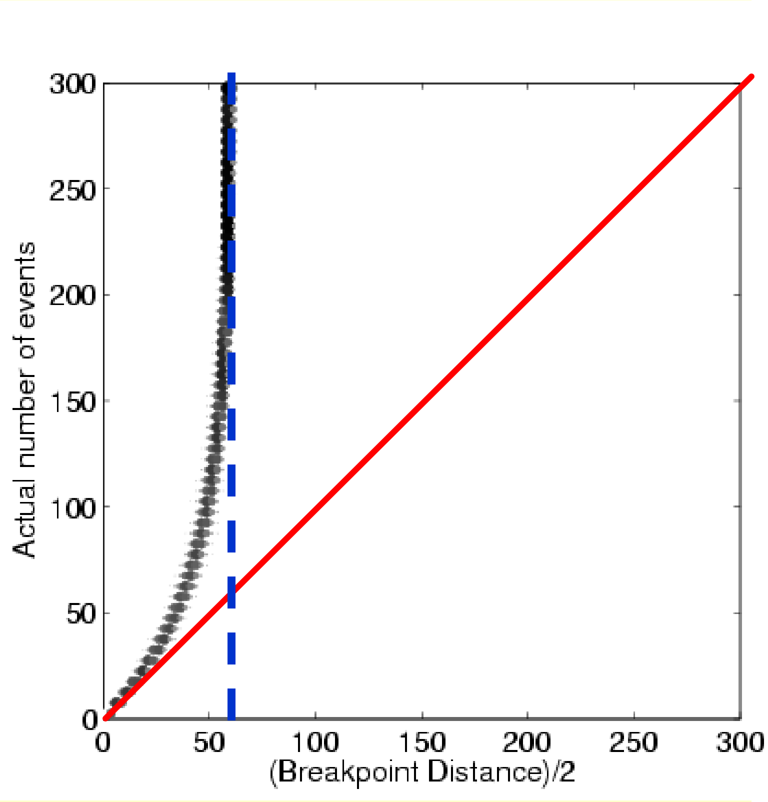
Theorem *Let G_k be the genome obtained after applying k random rearrangement events to genome G_0 according to the GNT model with parameters α and β. Let $\mathcal{F}_k$ be the estimate to $E[BP(G_k, G_0)]$ in the Approx-IEBP distance.*
For all $k > 0$,

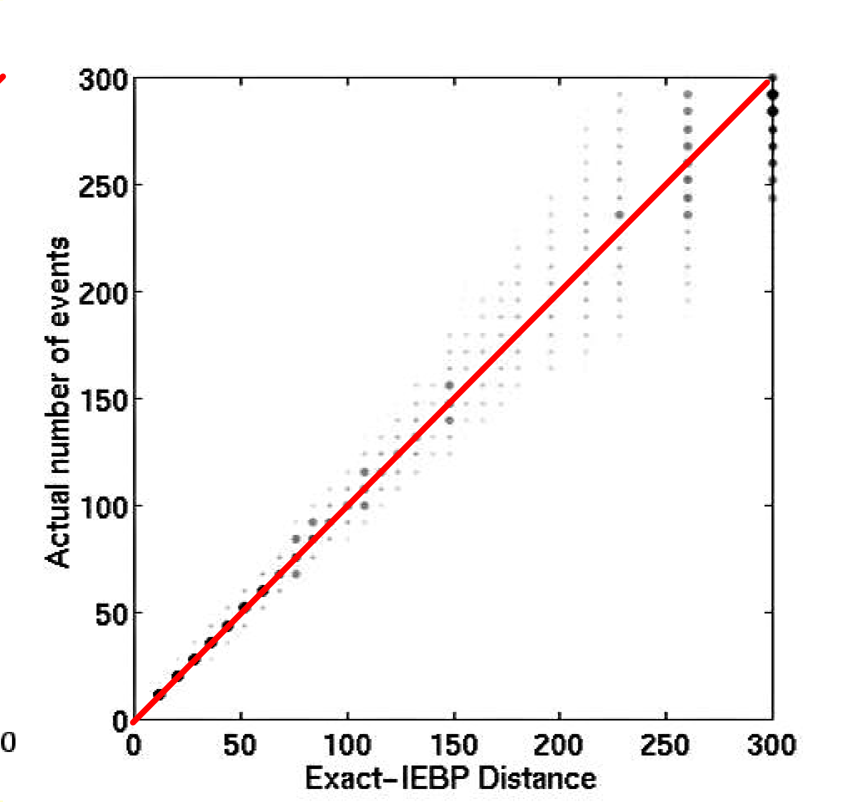
$$|\mathcal{F}_k - E[BP(G_k, G_0)]| \leq 1 + \frac{1}{n-1}, \ \textit{and}$$

$$\phi^{-1} \leq \frac{\mathcal{F}_k}{E[BP(G_k, G_0)]} \leq \phi$$

where $\phi = 1 + \frac{2+4\alpha+2\beta}{2+\alpha+\beta} n^{-1} + O(n^{-2})$.

True Evolutionary Distance Estimators





BP vs K (120 genes) Exact-IEBP vs K

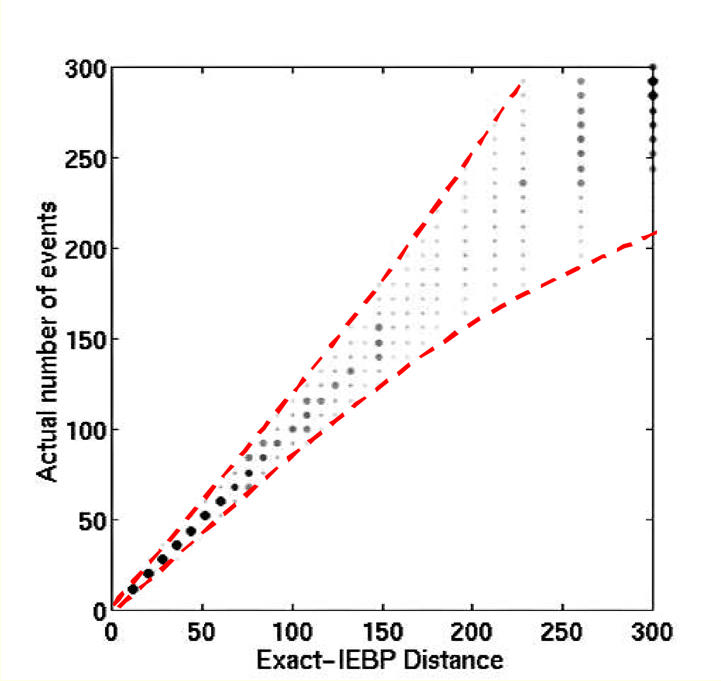
(K: Actual number of inversions) (Inversion-only evolution)

Variance of True Evolutionary Distance Estimators

- There are new distance-based phylogeny reconstruction methods (though designed for DNA sequences)
  - **Weighbor** *[Bruno et al. 2000]*

    uses the variance of good *t.e.d.*s, and yield more accurate trees than NJ.

- Variance estimates for the *t.e.d.*s *[Wang WABI'02]*
  - Weighbor(IEBP), Weighbor(EDE)



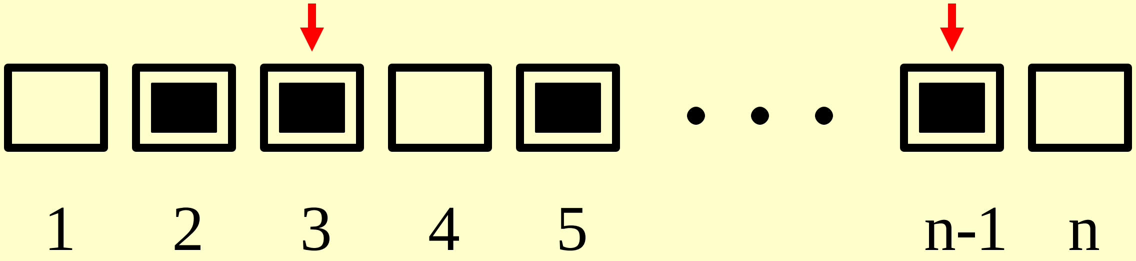
K vs Exact-IEBP (120 genes)

Deriving Var(BP)

- Difficulties in deriving Var(BP):
  - Even E(BP) is only in the form of unsimplified sums *[RECOMB '99, WABI '01].*
  - Breakpoints are not independent.
- We will use an approximating model to examine all breakpoints simultaneously
  - Idea: once two adjacent genes are separated, it is hard to bring the two genes back again (especially when there are many genes).

Approximating Model

- Approximating box model: boxes correspond to breakpoints.
- An approximation (using *n* boxes) can be obtained in the following way:
  - Every inversion chooses two boxes and put a ball in them if they are empty.
  - The BP distance is approximated by the number of nonempty boxes.



Approximating Model

- **Notations:**
  - Let B_i=1 if box i is not empty, 0 if it is.
  - We use inversion-only model to illustrate; let i and j be the two breakpoints corresponding to the two endpoints of the inversion being applied.
  - Let the number of breakpoints be b.
  - Let n be the number of genes.

Why the Approximation Works

- Case analysis: *[Hannenhalli and Pevzner 1995]*

Case	?BP	Condition	# inversions	
1	+2	$B_i=B_j=0$	$\binom{n-b}{2}$	
2	+1	$B_i=0$, $B_j=1$ or $B_i=1$, $B_j=0$	$b(n-b)$	
3a	0	$B_i=B_j=1$		Total $\binom{b}{2}$
3b	-1	$B_i=B_j=1$, one/both of $(g_{i-1}, -g_j)$, $(-g_i, g_j)$ adjacencies are in $G_{0.}$	$\leq b$	
3c	-2			

- When b is small, probability of case 3 out of cases 1, 2, and 3 is small (when n is large)
- When b is large, probability of 3b/3c out of case 3 is small
- As a result we can ignore cases 3b/3c
-> As a breakpoint is asserted, it does not disappear

Derivation of the Variance

- Fix k. Let $S=\left(\frac{1}{\binom{n}{2}}(x_1x_2+x_1x_3+\ldots+x_{n-1}x_n)\right)^k$
  - Each term in the expansion of S is a way of applying k inversions

    E.g. $x_1^3x_2x_3^2$: box 1 three times, 2 once, 3 twice
  - The coefficient of the term is the probabilities of such k inversions
  - If transpositions and inverted transpositions are present:

$$S=\left(\frac{1-\alpha-\beta}{\binom{n}{2}}\sum_{1\leq i<j\leq n}x_ix_j+\frac{\alpha+\beta}{\binom{n}{3}}\sum_{1\leq i<j<l\leq n}x_ix_jx_l\right)^k$$

- Let $S(a_1,a_2,\ldots,a_n)$ be the value of S when we let $x_i=a_i$ for all i.
- Let $S_j=S(\underbrace{1,1,1,\ldots,1}_{j\ 1's},0,\ldots,0)$

Derivation of Var(BP)

- Let u_i be the sum of coefficients of all terms in the expansion of S in the following form:

$$x_1^{a_1} x_2^{a_2} \cdots x_i^{a_i} \quad (a_1, a_2, ..., a_i > 0)$$

Then $\binom{n}{i} u_i$ is the probability of having i nonempty boxes after k events.

- We want to compute

$$Z_a = \sum_{i=0}^{n} i(i-1)\cdots(i-a+1)\binom{n}{i} u_i = n(n-1)\cdots(n-a+1)\sum_{i=a}^{n}\binom{n-a}{i-a} u_i$$

In particular,

$$z_1 = \sum_{i=1}^{n} i \binom{n}{i} u_i = E[b \mid k] \approx E[BP(G_0, G_k)]$$

$$z_2 = \sum_{i=1}^{n} i(i-1) \binom{n}{i} u_i = E[b^2 - b \mid k] \approx E[BP^2(G_0, G_k) - BP(G_0, G_k)]$$

$$\begin{aligned} S &= \left(\frac{1}{\binom{n}{2}} \left(\sum_{1 \le i < j \le n} x_i x_j \right) \right)^k \\ &= \sum_{1 \le i \le n} \sum_{\{t_1, t_2, \ldots, t_i\} \subseteq \{1, 2, \ldots, n\}} \sum_{\substack{a_1, a_2, \ldots, a_i \ge 1 \\ a_1 + a_2 + \ldots + a_i = 2k}} c(t_1, t_2, \ldots, t_i, a_1, a_2, \ldots, a_i) x_{t_1}^{a_1} x_{t_2}^{a_2} \cdots x_{t_i}^{a_i} \end{aligned}$$

$$\begin{aligned} S_j &= \sum_{1 \le i \le j} \sum_{\{t_1, t_2, \ldots, t_i\} \subseteq \{1, 2, \ldots, j\}} \sum_{\substack{a_1, a_2, \ldots, a_i \ge 1 \\ a_1 + a_2 + \ldots + a_i = 2k}} c(t_1, t_2, \ldots, t_i, a_1, a_2, \ldots, a_i) \\ &= \sum_{1 \le i \le j} \sum_{\{t_1, t_2, \ldots, t_i\} \subseteq \{1, 2, \ldots, j\}} u_i = \sum_{1 \le i \le j} \binom{j}{i} u_i \end{aligned}$$

Lemma Let a be some given integer such that $1 \leq a \leq n$. Let us be given $\{u_1, u_2, \ldots, u_n\}$ such that

$$\sum_{i=0}^{j} \binom{j}{i} u_i = \sum_{i=0}^{n} \binom{j}{i} u_i = S_j$$

for all j, $1 \leq j \leq n$. We have

$$\sum_{i=n-a}^{n} (-1)^{n-i} \binom{a}{n-i} S_i = \sum_{i=0}^{n} \binom{n-a}{i-a} u_i$$

Expectation and Variance *[WABI'02]*

- Let b$_k$ be the number of nonempty boxes after k (box choosing) iterations in the approximation model. Let *a* + *ß* = *?*. We have

$$S_{n-1} = (1 - \frac{2+\gamma}{n})^k, S_{n-2} = \left(\frac{(n-3)(n-2-2\gamma)}{n(n-1)}\right)^k.$$

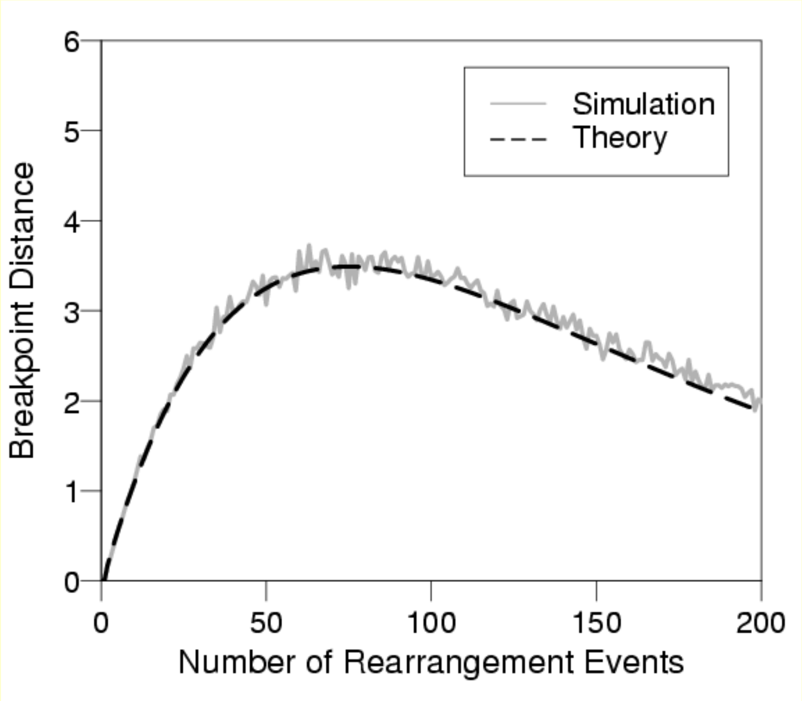
$$Eb_k = n(1 - S_{n-1})$$

$$Varb_k = nS_{n-1} - n^2 S_{n-1}^2 + n(n-1)S_{n-2}^2$$

- We use the delta method to obtain the variance of IEBP:

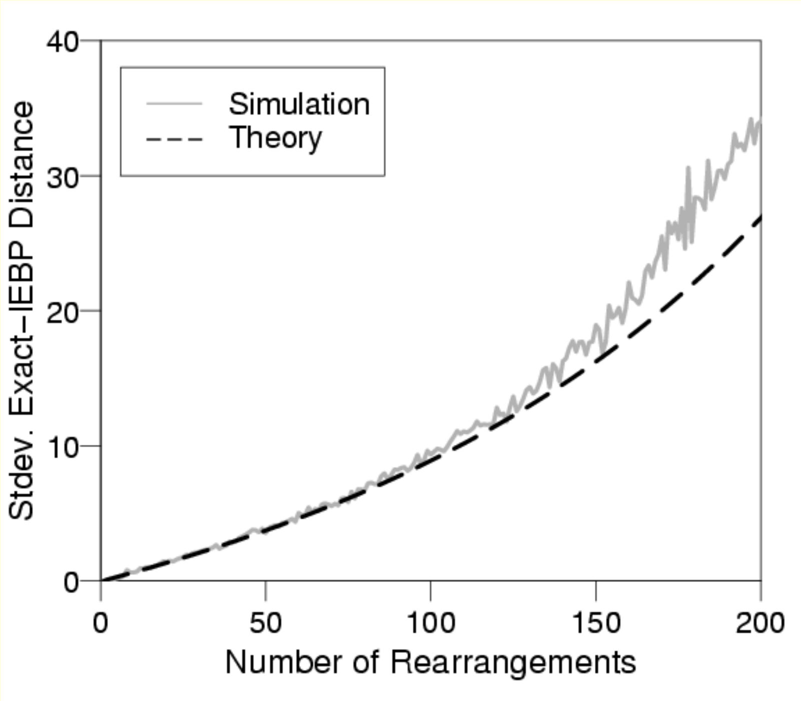
$$\text{Var } \widehat{k}(b_k) \simeq (\frac{d}{dk}Eb_k)^{-2}\text{Var } b_k = \frac{\left(1 - nS_{n-1} + (n-1)(\frac{S_{n-2}}{S_{n-1}})\right)}{nS_{n-1}(\ln(1 - \frac{2+\gamma}{n}))^2}.$$

Simulation Results



$\mathsf{Var}(BP_k)$

Variance of BP distance after k events



$\mathsf{Var}\ \widehat{k}(b_k)$

Variance of IEBP

(120 genes, inversion only)

Regression Formula for E(INV) and Var(INV)

- Let *n* be the number of genes, *x* be the normalized number of inversions (*k/n*), and *f(x)* be the normalized expectation of the inversion distance (*f(x)* seems to be roughly independent of *n*)
- We use nonlinear regression to obtain easily computable formulas for E(INV) and Var(INV):

$$f(x) = \min\{\frac{x^2+bx}{x^2+cx+b}, x\} \quad (x = \frac{k}{n})$$

1. $f(0)=0$ 2. $f'(0)=1$

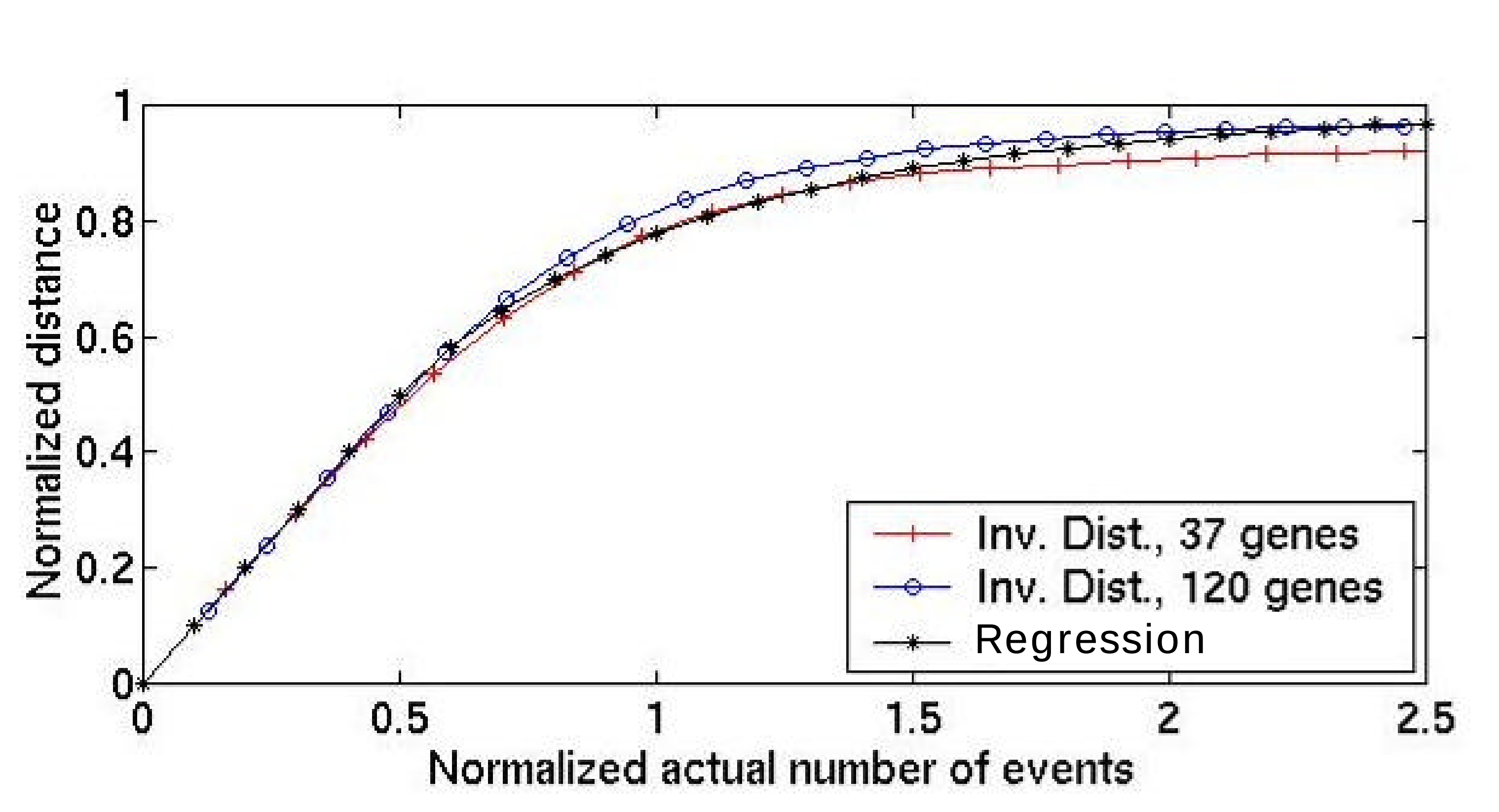
3. $0 \le f(x) \le x$

4. $f^{-1}(y)$ exists for all $y: 0 \le y \le 1$

-> b=0.5956, c=0.4577

EDE

[Moret, Wang, Warnow, & Wyman, ISMB’01]



Formula for Var(INV) and Var(EDE)

- Let *n* be the number of genes, *x* be the normalized number of inversions (*k/n*), and $g_n(x)$ be the standard deviation of the inversion distance.
- The regression of $g_n(x)$: we use the following form

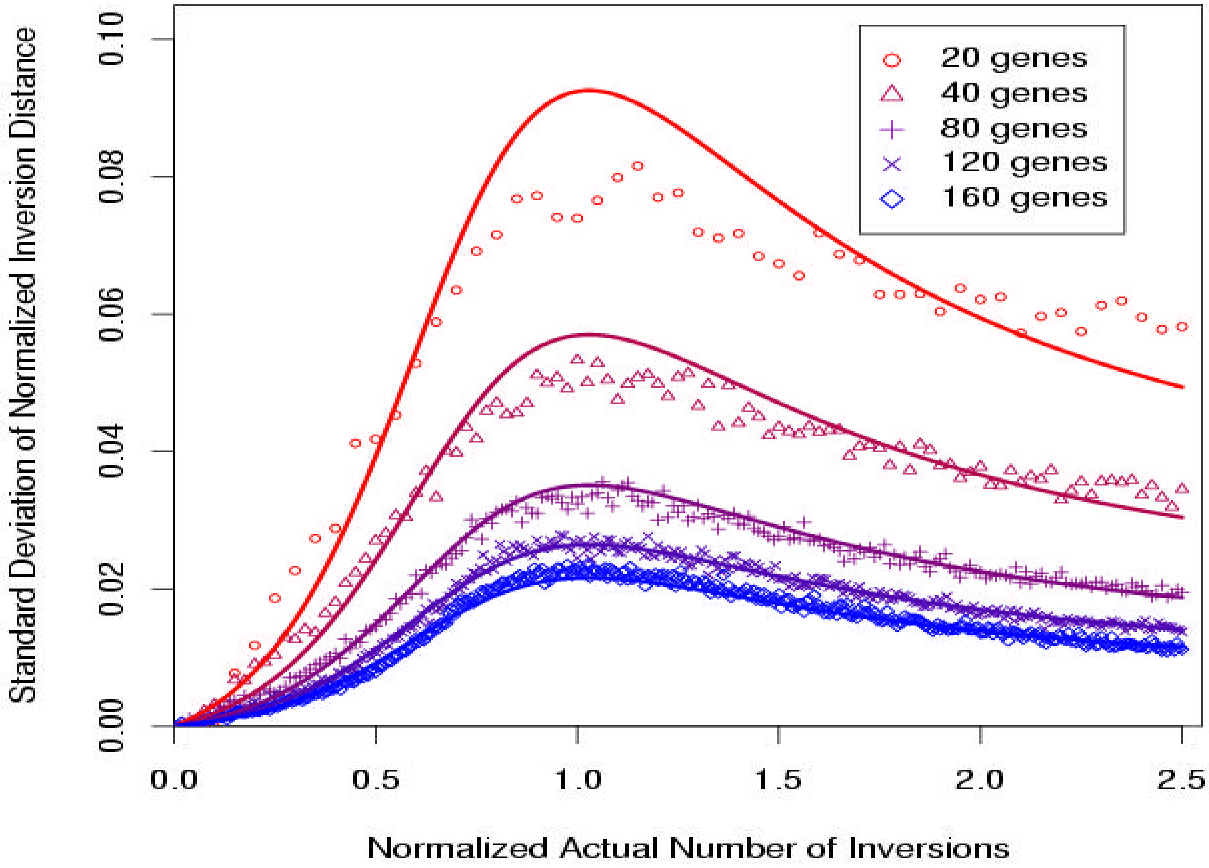
$$g_n(x) = n^q \frac{ux^2 + vx}{x^2 + wx + t}$$

  q=-0.6998, u=0.1684, v=0.1573, w=-1.3893, and t=0.8224.

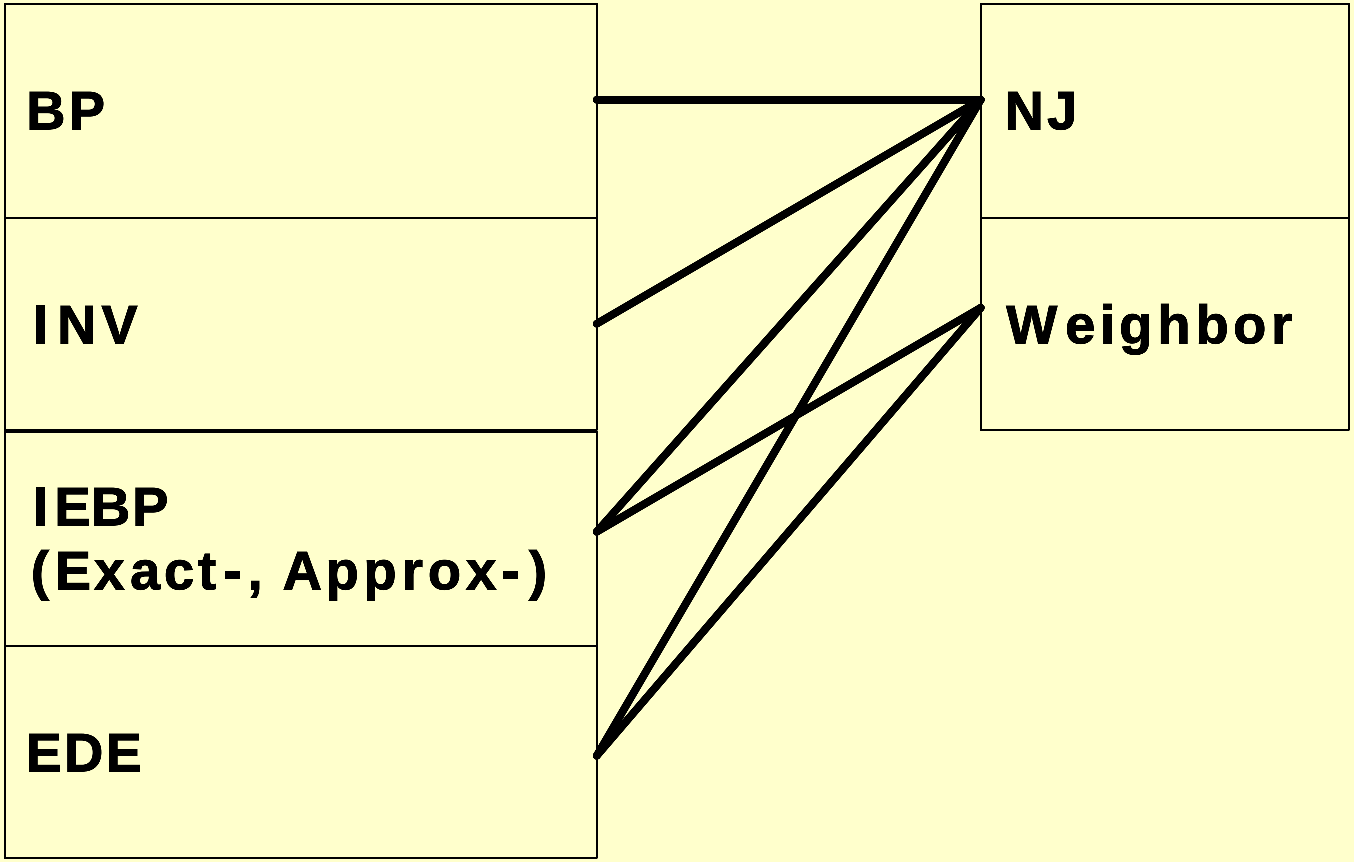
- Var(EDE) can be obtained using the delta method on Var(INV).

Regression for Var(INV)

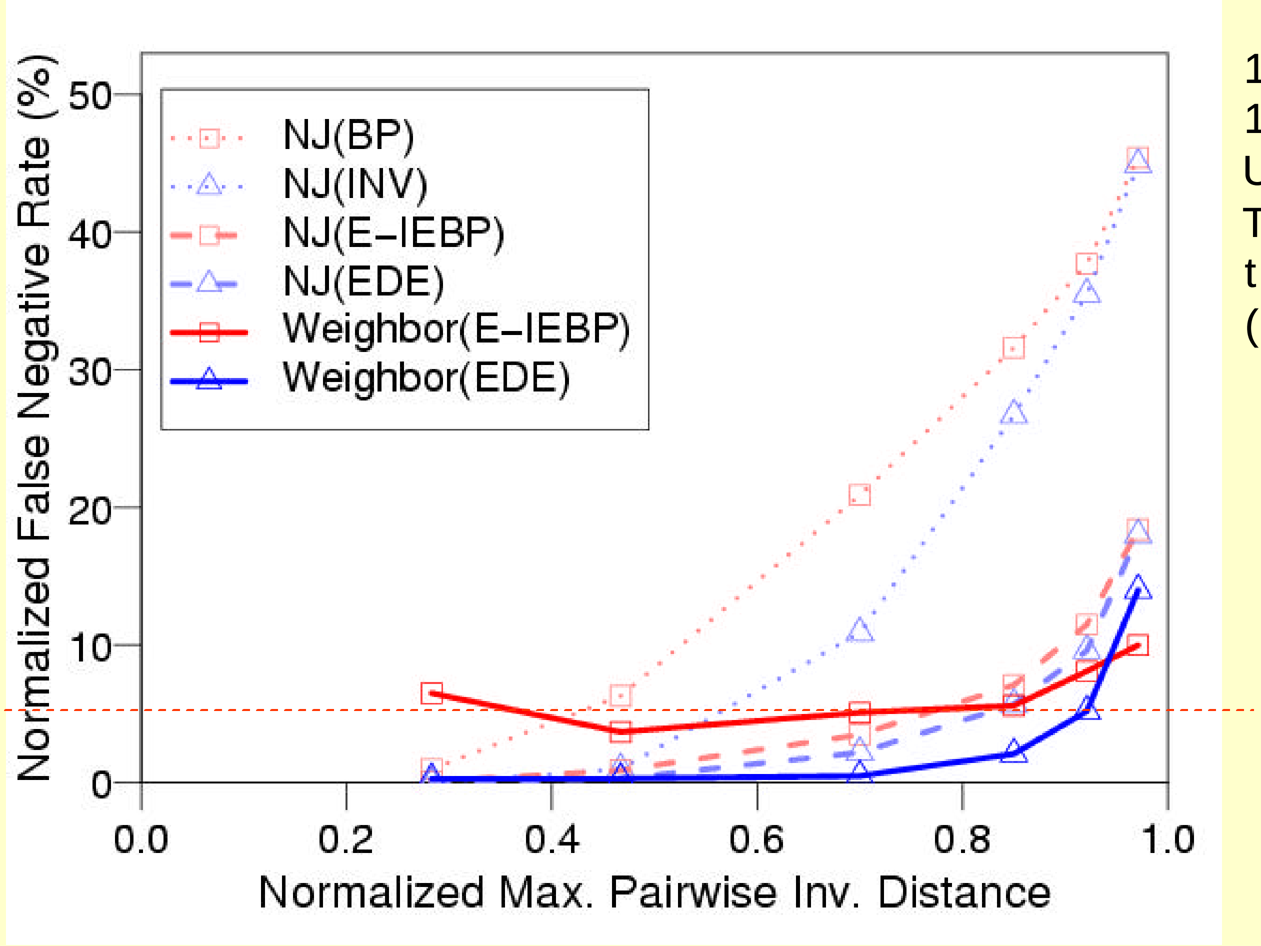
Regression: solid lines, Simulation: dots



Distance-Based Methods



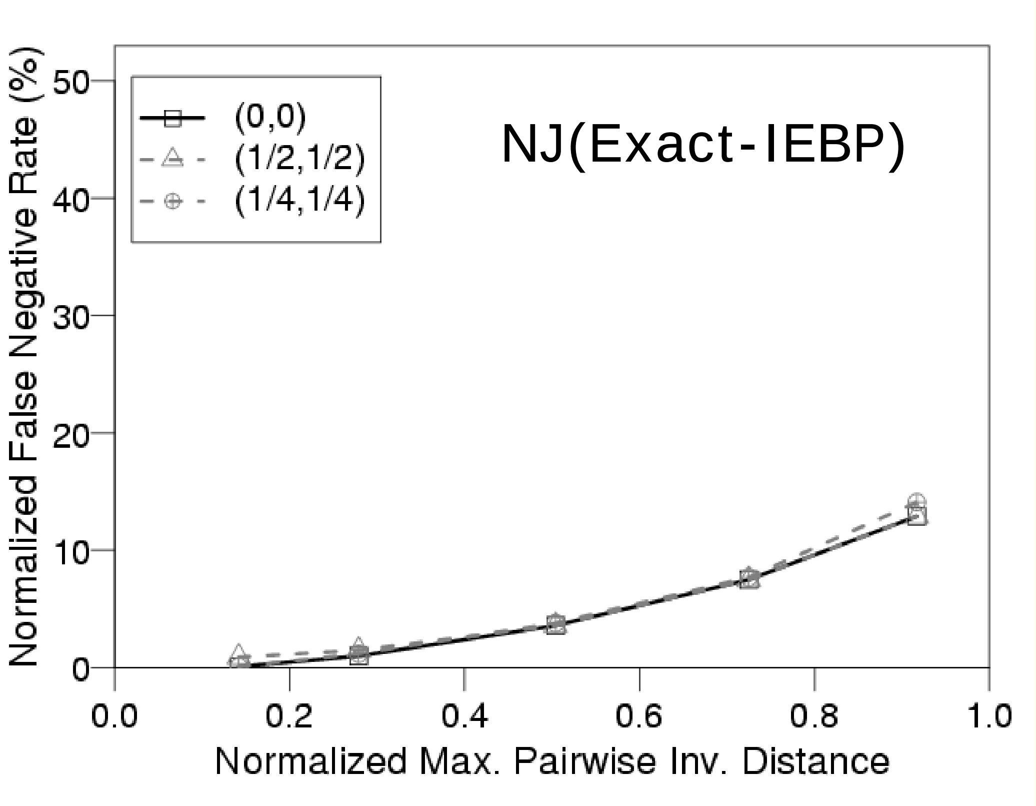
Using T.E.D. Helps



5%

120 genes
160 taxa
Uniformly random trees
Transpositions/inverted transpositions only
(180 runs per figure)

IEBP is Robust to Model Violations



120 genes, 160 taxa
Uniformly Random Trees
(alpha,beta)=(0,0) (inversion only)

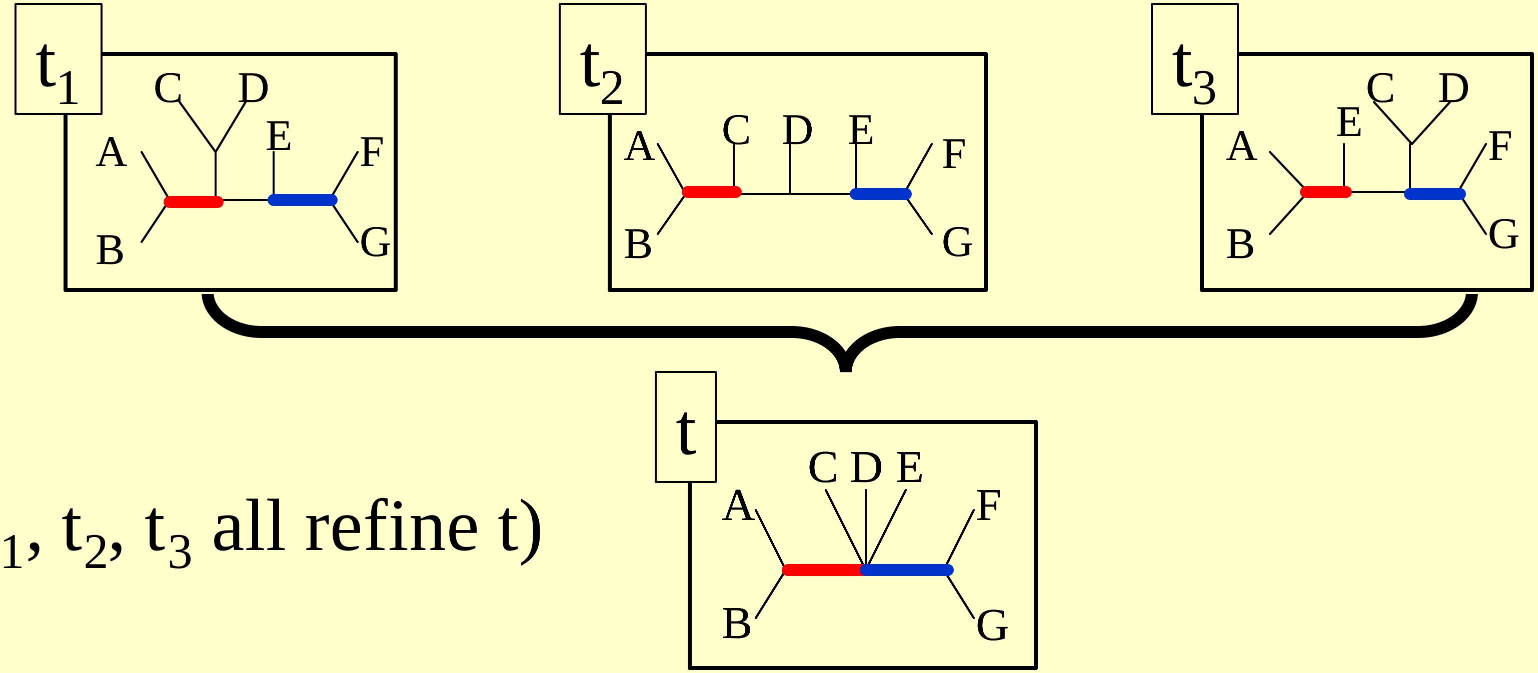
Maximum Parsimony Returns Thousands of Trees

- Example:
  - The complete *Caesalpinia* dataset:
    7095 trees on 82 taxa.
  - The *Astericeae* dataset:
    34,560 trees on 288 taxa.

- Consensus methods are necessary so we can summarize so many trees.
- Current approaches are limited to the strict consensus and majority consensus trees, and lose information

Postprocessing: Traditional Approaches

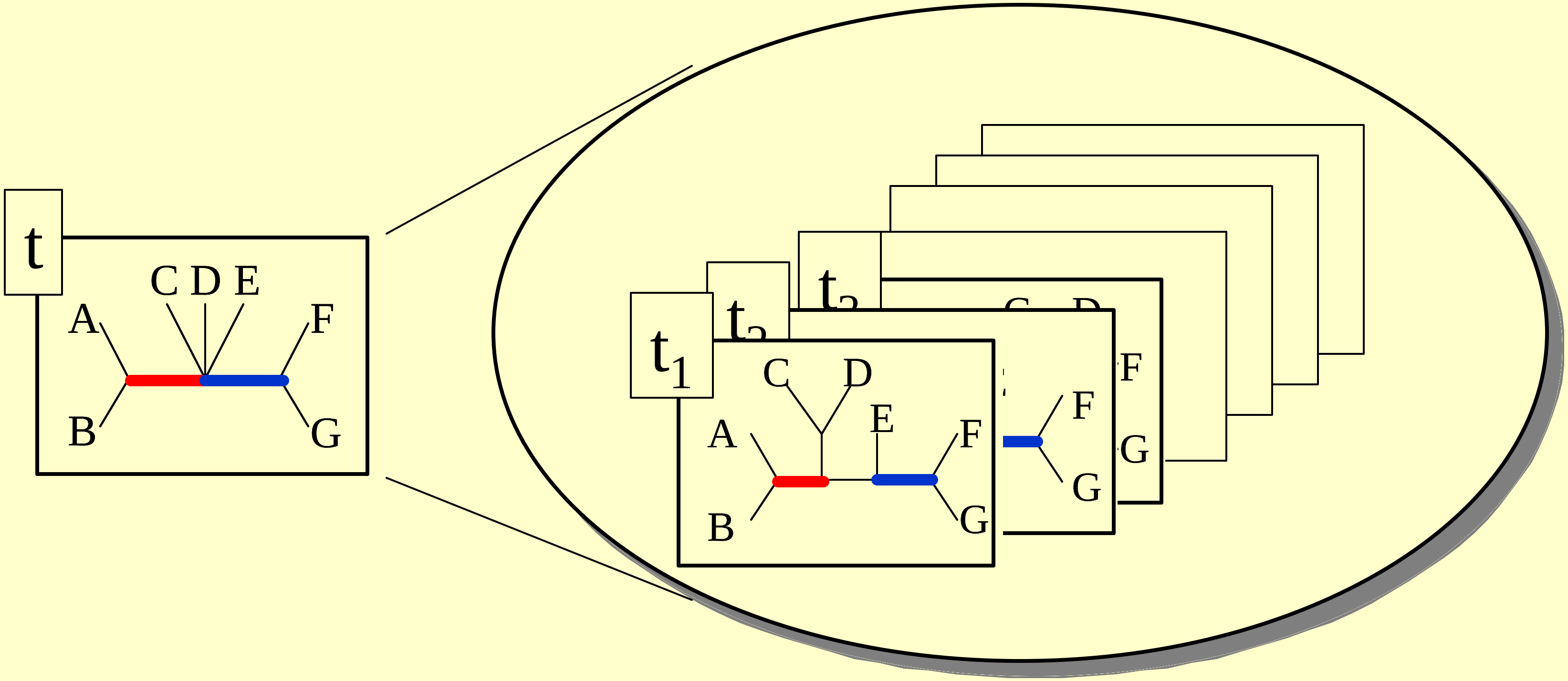
- Single-tree consensus
  Example: strict consensus



(t_1, t_2, t_3 all refine t)

How Do We Interpret the Consensus Tree

- Given a nonbinary consensus tree *t*, every binary tree that refines *t* is equally probable to be the true tree:

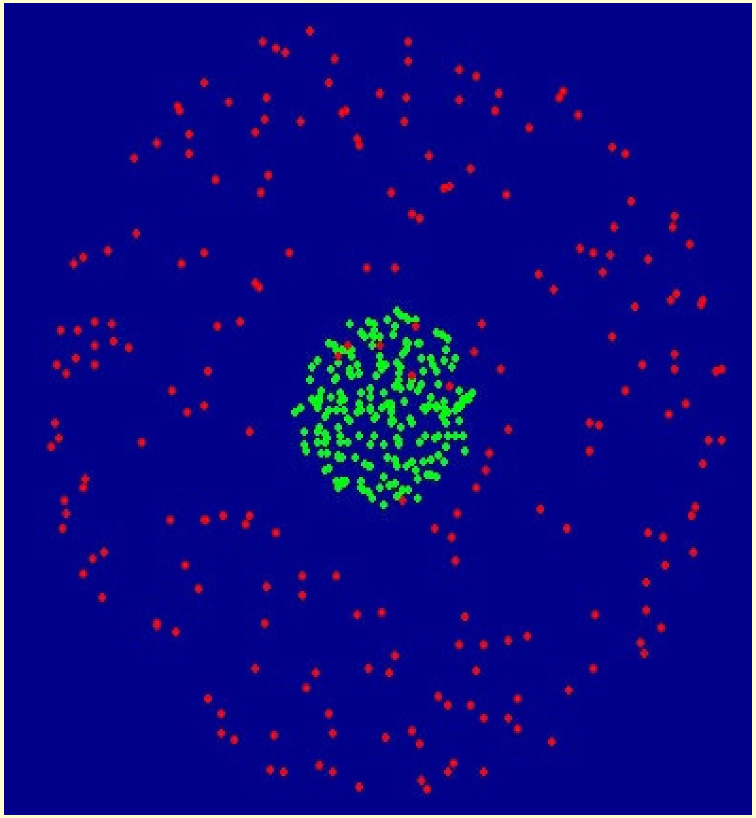


(15 refinement trees)

Disadvantages of Single-Tree Consensus

- Loses a lot of information
- Sensitive to outlier trees
- Sensitive to small perturbations in the dataset

Sometimes A Cluster is Enough (Campanulaceae)

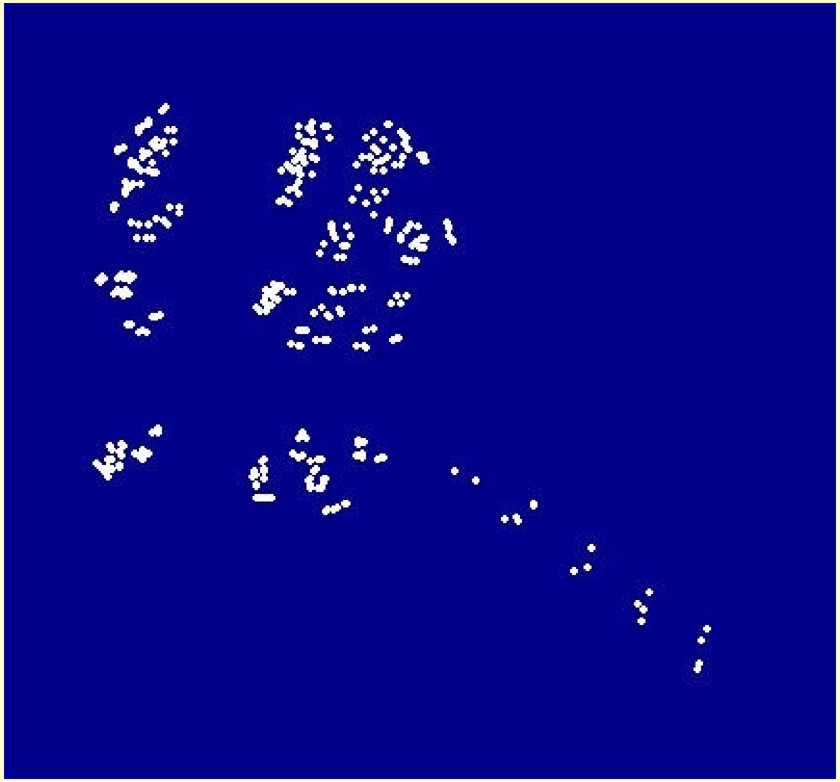


The *Campanulaceae* Gene-Order Dataset

1. 13 taxa (outgroup Tobacco)
2. 216 trees

(Courtesy Nina Amenta and Jeff Klingner)

Complex Structure in the Inferred Set of Trees



The *Caesalpinia* cpDNA Dataset

1. 51 taxa

2. 342 trees

(Courtesy Nina Amenta and Jeff Klingner)

Why We Want to Cluster Trees

- Dividing trees into clusters, and use the consensus trees from each cluster to represent “conflicting hypotheses” for the true phylogeny.

- Merits:
  - Represent the input set of trees better
  - Identify outliers
  - Restrict perturbations to a small number of clusters

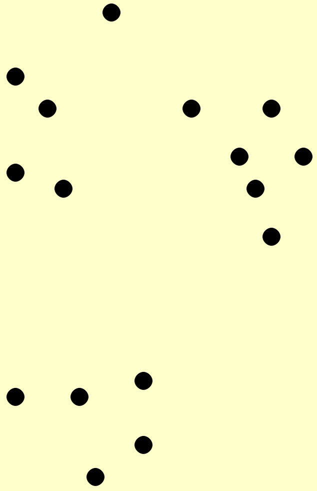
Biological Criteria

- Number of clusters
- Number of edges of the consensus
- Diameter of a cluster
- Density of clusters
- Etc.

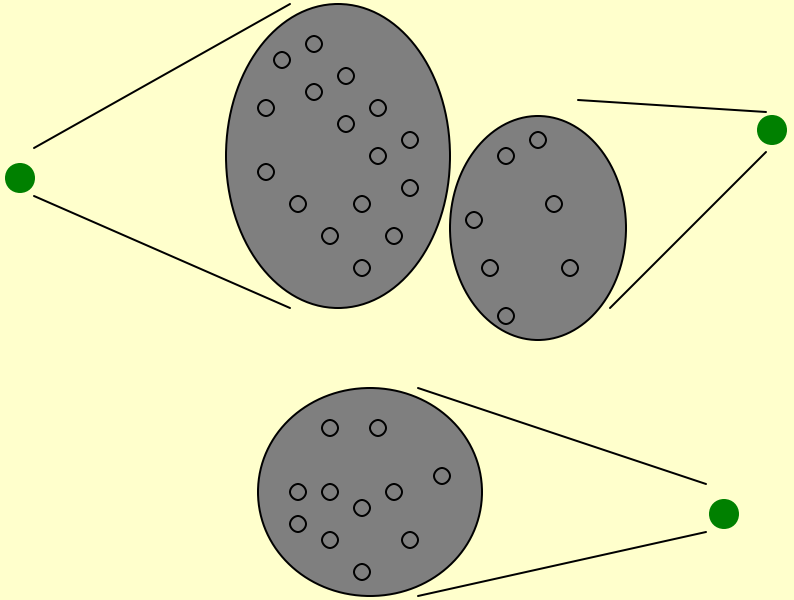
Information Loss:
How We Interpret the Clustering

- We can define distributions for both the original set of trees and the clustering.

Input set of tree *T:*
All trees are equally probable.



Clustering $\{C_1, C_2, \ldots, C_k\}$:
All trees refining any of SC(C_i) are equally probable.



Distributions

- Input set of tree *T*:

$$f_T(t) = \begin{cases} \frac{1}{|T|} & \text{if } t \in T \\ 0 & \text{othewise} \end{cases}$$

- Clustering $\{C_1, C_2, \dots, C_k\}$: let

$$B = \bigcup_{i=1}^{k} B(C_i)$$

$$f_C(t) = \begin{cases} \frac{1}{|B|} & \text{if } t \in B \\ 0 & \text{otherwise} \end{cases}$$

(Here B(C) is the set of binary trees that refine the strict consensus of C)

Information Loss (KL)

- The distance between the two distributions is the loss of information due to clustering.
  - L_1 distance
  - L_2 distance
  - L_∞ distance

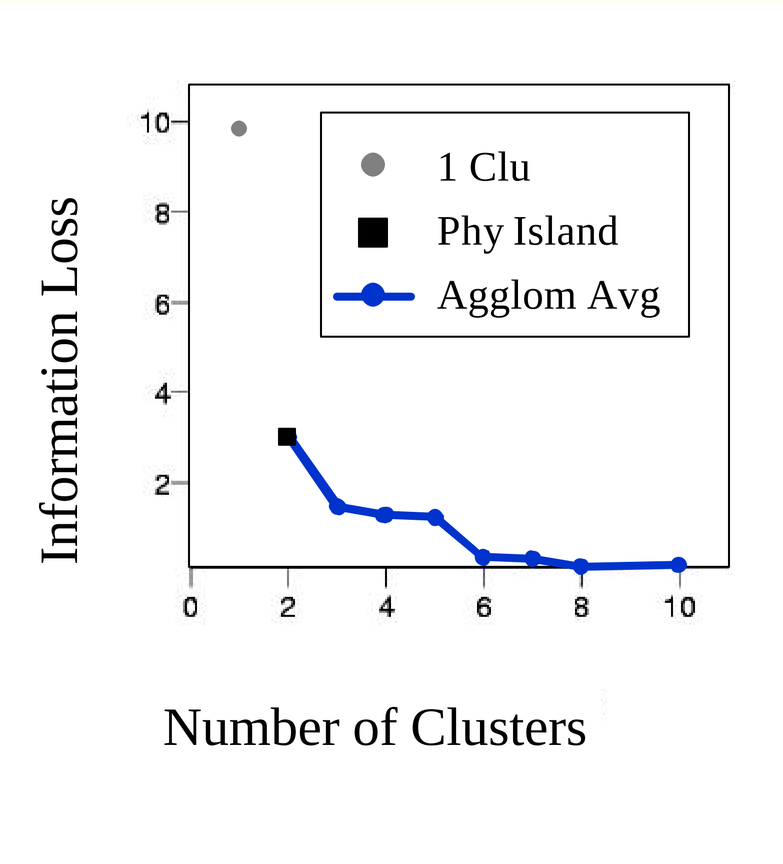
$$\left.\begin{array}{l}\end{array}\right\} L_x(T,C) = \sum_t \| f_T(t) - f_C(t) \|_x$$

  - Kullback-Leibler distance (relative entropy):

$$KL(T,C) = \sum_t f_T(t) \ln \frac{f_T(t)}{f_C(t)}$$

Postprocessing of Phylogenetic Analysis Using Clustering *[ISMB'02]*

- The first framework using clustering algorithms in the postprocessing of phylogenetic analyses.
  - Improves upon the traditional single-consensus approach in terms of information loss
- Identifies outliers in the *Caesalpinia* dataset
  - Improves the resolution of the strict consensus by 36%
  - Only loses 4% of the trees



Caesalpinia (51 taxa, 450 trees)

Clu No.	No. of Trees	% Edges lost
1clu	450	22.9%
1	108	10.4%
2	324	12.5%
3	18	10.4%
1+2	432	14.6%

KL(Agg-complete, 3clu) = 1.449269
KL(1clu) = 9.790346

Improvement: (22.9-14.6)/22.9 = 36%
% trees dropped: 18/450=4%

Acknowledgements

- University of Texas
  Tandy Warnow (Advisor)
  Robert K. Jansen Stacia Wyman
- University of New Mexico
  Bernard M.E. Moret David Bader
  Jijun Tang Mi Yan
- Central Washington University
  Linda Raubeson
- University of Ottawa
  David Sankoff
- University of Canterbury
  Mike Steel
- LIRMM
  Olivier Gascuel

- **Genome rearrangement phylogeny**

1. **[STOC' 01]** Li-San Wang and Tandy Warnow,
"Estimating true evolutionary distances between genomes,"
Proceedings of the Thirty-Third Annual ACM Symposium on the Theory of Computing (STOC'01), pp. 637-646, Crete, Greece (2001).
2. **[ISMB' 01]** Bernard M.E. Moret, Li-San Wang, Tandy Warnow, and Stacia Wyman,
"New approaches for reconstructing phylogenies based on gene order," Proceedings of S. Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB-2001), pp.165-173, (2001).
3. **[WABI' 01]** Li-San Wang,
"Exact-IEBP: A New Technique For Estimating Evolutionary Distances Between Whole Genomes," L*ecture Notes for Computer Sciences No. 2149: Proceedings of the First Workshop on Algorithms in BioInformatics (WABI'01),* pp. 175-188, 2001.
4. **[PSB' 02]** Li-San Wang, Robert Jansen, Bernard Moret, Linda Raubeson, and Tandy Warnow,
"Fast Phylogenetic Methods For Genome Rearrangement Evolution: Empirical Study,"
Proceedings of Fifth Pacific Symp. of Biocomputing (PSB'02), pp. 524-535*, Hawaii, USA 2002.*
5. **[WABI' 02]** Li-San Wang, *"Distance-Based Genome Rearrangement Phylogeny Using Weighbor," Lecture Notes for Computer Sciences No. 2452: Proceedings of the Second Workshop on Algorithms in BioInformatics (WABI'02),* pp. 112-125, 2002.

- **Postprocessing by clustering**

1. **[ISMB' 02]** Cara Stockham, Li-San Wang, and Tandy Warnow,
"Statistically Based Postprocessing of Phylogenetic Analysis by Clustering,"
Bioinformatics: supplemental issue, Proceedings of the 10th International Conference on Intelligent Systems and Molecular Biology (ISMB 2002), pp. 285-293, August 2002.

http://www.cs.utexas.edu/users/lisan/