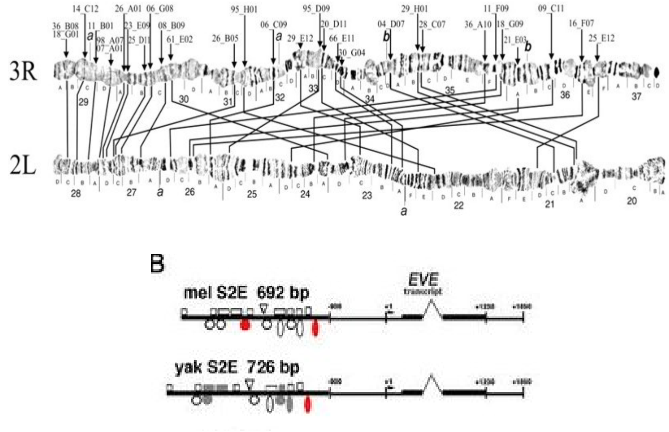
# Two Postcards from the Edge

Rick Durrett



# Genome Rearrangement

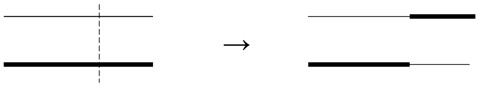
Joint work with

Nathanael
Berestycki

## Genome Rearrangement

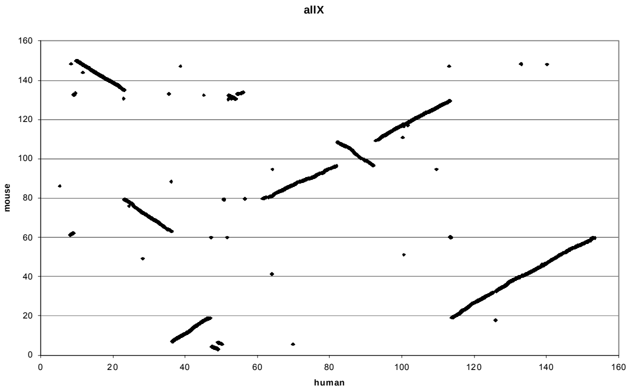
Genomes evolve by **inversions** that reverse the order of segments of chromosomes

**Translocations** between chromosomes



**Fissions** and **fusions** that change chromosome number. Today we will restrict our attention to inversions.

## Human vs Mouse X chromsome



## Human vs. Mouse X chromosome

The relationship may be described by a signed permutation

1 –7 6 –10 9 –8 2 –11 –3 5 4

**Parsimony Approach:** What is the minimum number of inversions needed to transform this arrangement back to the identity?

**Hannenhalli and Pevzner (1995)** developed a polynomial algorithm for the inversion distance

## Distance = 7

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | –7 | 6 | –10 | 9 | –8 | 2 | –11 | –3 | 5 | 4 |
| 1 | –7 | 6 | –10 | –9 | –8 | 2 | –11 | –3 | 5 | 4 |
| 1 | –7 | 6 | –10 | –9 | –8 | 2 | –4 | –5 | 3 | 11 |
| 1 | –7 | 4 | –2 | 8 | 9 | 10 | –6 | –5 | 3 | 11 |
| 1 | –7 | 4 | 5 | 6 | –10 | –9 | –8 | 2 | 3 | 11 |
| 1 | –7 | 4 | 5 | 6 | –3 | –2 | 8 | 9 | 10 | 11 |
| 1 | 2 | 3 | –6 | –5 | –4 | 7 | 8 | 9 | 10 | 11 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

## D. repleta 2 vs. D. melanogaster 3R

unsigned comparison, parsimony distance ≤ 54

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 36 | 37 | 17 | 40 | 16 | 15 | 14 | 63 | 10 | 9 |
| 55 | 28 | 13 | 51 | 22 | 79 | 39 | 70 | 66 | 5 |
| 6 | 7 | 35 | 64 | 33 | 32 | 60 | 61 | 18 | 65 |
| 62 | 12 | 1 | 11 | 23 | 20 | 4 | 52 | 68 | 29 |
| 48 | 3 | 21 | 53 | 8 | 43 | 72 | 58 | 57 | 56 |
| 19 | 49 | 34 | 59 | 30 | 77 | 31 | 67 | 44 | 2 |
| 27 | 38 | 50 | 26 | 25 | 76 | 69 | 41 | 24 | 75 |
| 71 | 78 | 73 | 47 | 54 | 45 | 42 | 46 | | |

## Durrett (2003) J. Theoretical. Prob.

Let φ = –2 + # of conserved adjacencies

If there are n markers, φ is an eigenfunction of the Markov chain with eigenvalue (n-1)/(n+1)

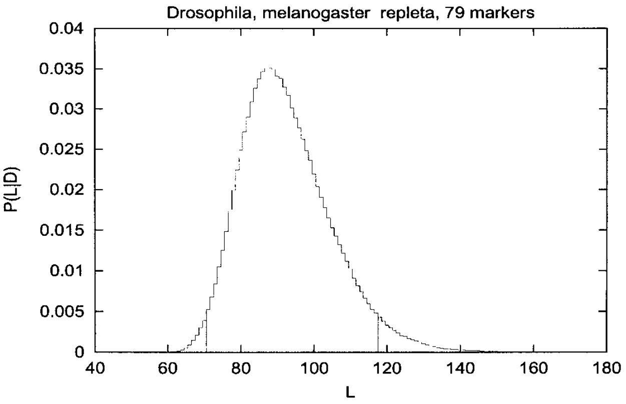
Conserved adjacencies = 11, n = 79

$$\text{Set } 78\left(\frac{78}{80}\right)^m = 9 \text{ and solve}$$

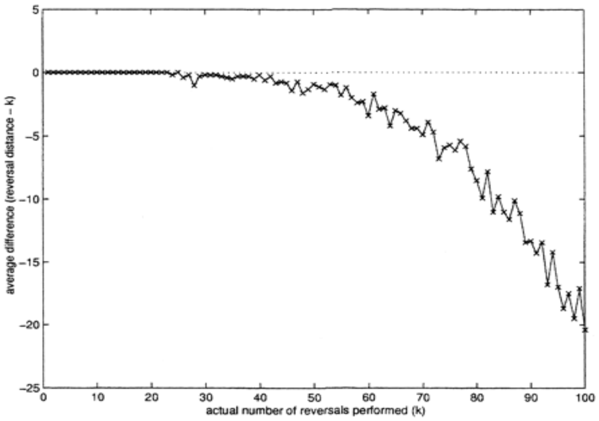
$$m = \frac{\log(9/78)}{\log(78/80)} = 85.3 \text{ [pars. 54]}$$

## Bayesian Estimation

parsimony 54, moment est. 85.3



## When is the parsimony estimate reliable?



## Random Transpositions

For simplicity consider random transpositions instead of inversions

(1 7 4) (2) (3 12) (5 13 9 11 6) (8 10 14)

This permutation has five cycles.

Distance from identity = n – # of cycles
= 14 – 5 = 9

## Coagulation Fragmentation

(1 7 4) (2) (3 12) (5 13 9 11 6) (8 10 14)

If we transpose two markers in different cycles they merge, e.g., 7 and 9

(1 9 11 6 5 13 7 4) (2) (3 12) (8 10 14)

If we pick two in the same cycle, e.g., 13 and 11) it breaks into two

(1 7 4) (2) (3 12) (5 11 6) (9 13) (8 10 14)

## Connections with random graphs

When we transpose i and j connect them with an edge. As long as we can ignore fragmentation, cycles in permutation = components in graph

When # of edges is cn [out of n(n-1)/2], is ≈ an Erdös-Rényi random graph, p = 2c/n.

When c < ½ all components small and fragmentation can be ignored

## Phase transition, cn inversions

When c < ½ distance is roughly the number of transpositions

When c > ½ the behavior of large cycles becomes complicated but (a) there are at most $n^{1/2}$ cycles of size > $n^{1/2}$ and (b) fragmentation can be ignored for smaller cycles. Number of cycles in permutation is ≈ number of components in random graph

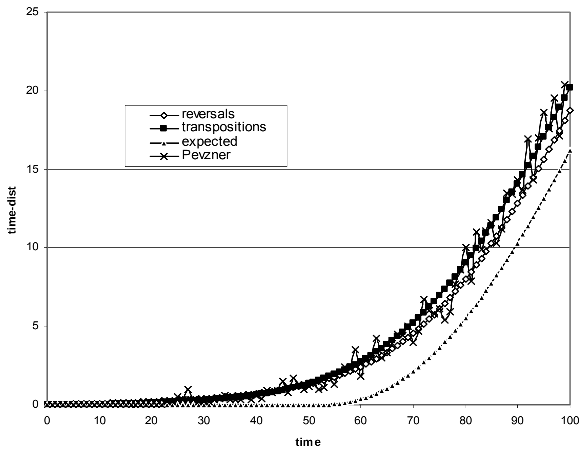
## The answer

$$u(c) = 1 - \sum_{k=1}^{\infty} \frac{1}{c} \frac{k^{k-2}}{k!} (ce^{-c})^k$$

**Theorem. The distance from the identity at time cn/2 is ~ u(c)n.**

When c < 1, u(c) = c/2, sublinear for c > 1

kth term is fraction of vertices in components of size k in Erdös Rényi random graph



## Regulatory Sequence Evolution

Joint work with

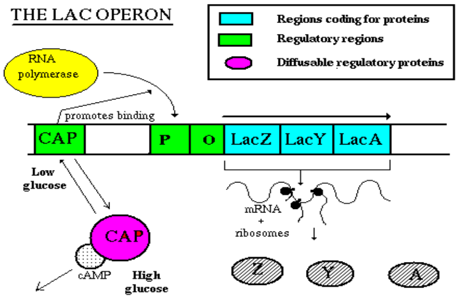
Deena Schmidt

Graduating May 2007

## Human and chimpanzee DNA is 98.7% identical

## But there are significant phenotypic differences

## Differences can come from gene regulation Is 6 million years enough ?



## Main Question

Regulatory sequences are often 6-9 nucleotides long and appear within 1kb (1000 nucleotides) of the start of a gene.

**Q. How long does it take for a specified word to appear in a region this size in some individual in the population?**

We suppose the mutation is advantageous and then sweeps to fixation.

## Stone and Wray (2001)

Six letter words in a 2kb region

| | |
|---|---|
| Humans | 5950 years |
| Mice | 80 years |
| Drosophila | 24 years |
| C. elegans | 4 years |
| Yeast | 73 days ! |

## Stone and Wray's argument

Simulation for 2kb region in one individual: mean 952 mutations for six letter word $= 4.76 \circ 10^8$ generations (they take $\mu = 10^{-9}$)

Assume individuals independent! Divide by 2 DNA strands $\circ\ 10^6$ individuals $= 238$ generations

Multiply by 25 years per generation = 5950 years

## What’s wrong with this?

Individuals are not independent!

Two humans differ at 0.1% of their DNA

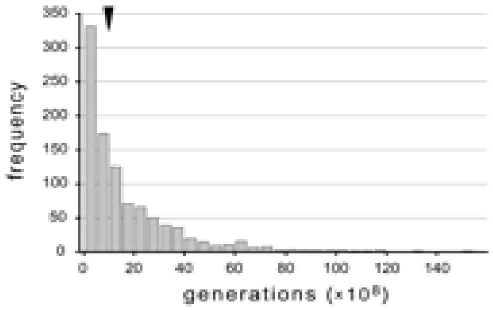
Human effective population size is ≈ $10^4$ not $10^6$

Polymorphism $\dfrac{2\mu}{1/2N+2\mu} = \dfrac{4N\mu}{1+4N\mu}$

If μ = 2.5 × $10^{-8}$ this is 0.001 when N = $10^4$

MacArthur and Brookfield (2004) Mol. Biol. Evol.

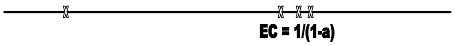
## Stone Wray simulations



## Outline

- W nucleotides in one DNA sequence
- L nucleotides in one DNA sequence
- W nucleotides in N diploids
- L nucleotides in N diploids

## W letters in one DNA sequence



Kac $E_W T_W = 4^W$. Let $a = P_{W-1}(T_W < T_0)$.

Poisson clumping heuristic $E_\pi T_W \approx 4^W/(1-a)$

Aldous-Fill. Proposition 23, Chapter 3

$$|P_\pi(T_W > t) - \exp(-t/E_\pi T_W)| \le \tau_2/E_\pi T_W$$

## 1024 nucleotides in one DNA sequence

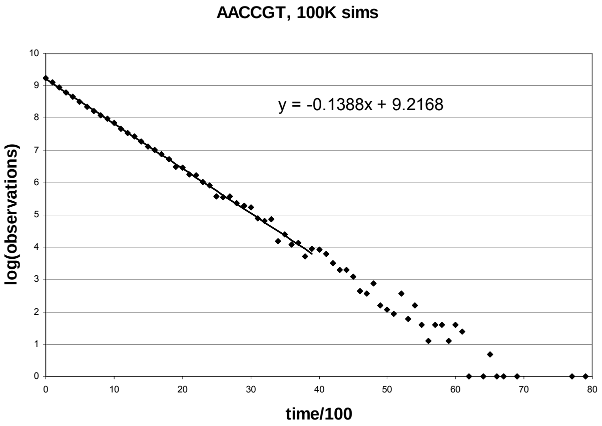
| W | P( wait = 0) |
|---|---|
| 6 | 0.2211 |
| 8 | 0.015504 |

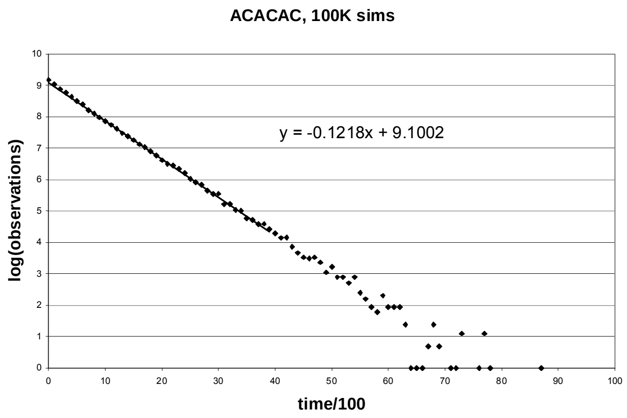
Using Arratia, Goldstein, and Gordon (1989)

and Poisson clumping ideas under $P_\pi$

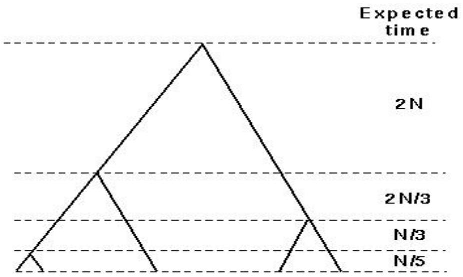
$$T_W \approx p\,\delta_0 + (1-p)\exp(\mu)$$

$$\mu = (4^W / WL)\, EC$$





## The Coalescent



When there are k lineages coalescence occurs at rate $C_{k,2}$ /2N

## W nucleotides in N diploids

Expected total time in genealogical tree

$$2N\sum_{k=2}^{2N} k\cdot\frac{1}{C_{k,2}} = 4N\sum_{j=1}^{2N-1}\frac{1}{j} \sim 4N\log(2N)$$

$\mu = 10^{-8}$   $N=10^4$   $W=8$   P( mutation ) = 0.0316

96.84% of the time no variation in population

## Fixation chain

$F = \{ t : X_t(i) = X_t(1) \text{ for all } i \}$

$T(n+1) = \inf\{ t > T(n) : t \in F, X_t(1) \neq X_{T(n)}(1) \}$

$Y_n = X_{T(n)}(1)$   $L_n$ = # of letters matching target

$\tau_k = \inf \{ n : L_n = k \}$

Mutations occur at rate 2NWμ and go to fixation with probability 1/2N, so target word is reached soon after $\tau_{W-1}$

## Killed fixation chain

$$\rho = \frac{2\mu N/9W}{1/2N + 2\mu N/9W} = \frac{4\mu N^2/9W}{1+4\mu N^2/9W}$$

Kill the fixation chain with probability 1 in state W-1 and with probability ρ in state W-2 and let S be the death time.

The expected time to find the target word in a population of size $10^4$ is ≈ $E_\pi S/(W\mu)$

## L nucleotides in N diploids

$M_i$ = number of words in segment of length L=1024 with i mismatches compared to target word

| W | 6 | 8 |
|---|---|---|
| $EM_1$ | 4.5 | 0.375 |
| $EM_2$ | 33.75 | 3.94 |

W=6: wait for a mutation in one of the $10^4$ individuals to give you what you want

$$375,000 = \frac{1}{2 \times 10^{-4} \times (1/3)} \cdot 25$$

W=6. Poisson mean 4.5 number of matches – 1, so waiting time has mean 100,000 years

W=8. With probability 1-exp(-3/8) = 0.3127, we have a match – 1.

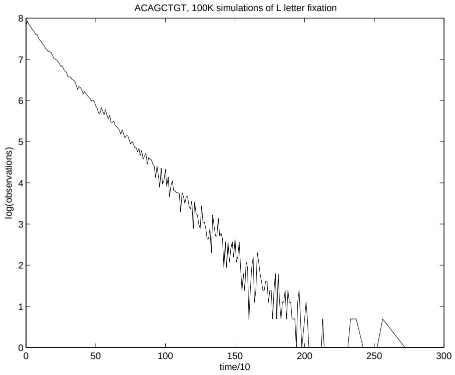
If no match – 1, we have to run killed L letter fixation chain

## Simulation of killed fixation chain

| | P(S=0) | ES |
|---|---|---|
| ACAGCTGT | .3185 | 253.77 |
| ACAGACAG | .3162 | 279.09 |
| AAAACAAA | .3123 | 295.94 |
| ACACACAC | .2817 | 327.91 |

Fixation happens at rate Lμ = $10^{-5}$ so 250 corresponds to 25 million generations or 625 million years (5 × $10^{11}$ events)

## Mystery: why is S approx. exp.?

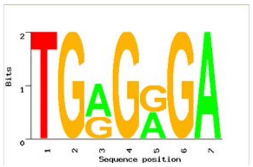


## Moral of the story

Words of length 6 in a one kb region can evolve in 100,000 years

If we want an exact match of an 8 nucleotide sequence then unless there is a match minus 1 in the initial condition this will take an average of 650,000,000 years

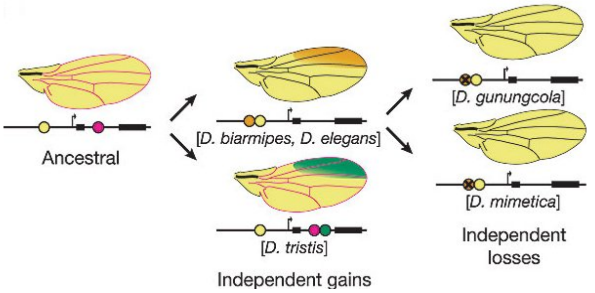
## Imperfect matches save the day



However gene regulation does not require an exact match to the target word. If 7 out of 8 is good enough, there are 3.94 match -2's in 1kb, so about 60,000 years is enough (an intelligent design)

## Future Work

Our analysis requires $N^3\mu^2$ to be small so it is not valid for Drosophila N = $10^6$ μ = $10^{-8}$



## Thanks

## www.math.cornell.edu/~durrett/

If that went by too fast, a PDF version of the talk can be found on my web page along with copies of all of the papers