Wald Lecture 1:
Philosophy and Anecdotes

Rick Durrett, Cornell U

PDF's of talks (6 slides per page) and papers:
www.math.cornell.edu/~durrett/

Recent Wald Lectures in Probability

(2005) S. Varadhan
(1999) Charles Newman
(1996) Tom Liggett
(1993) David Aldous
(1991) E.B. Dynkin
(1987) Persi Diaconis
(1986) Harry Kesten
(1979) Frank Spitzer
(1978) Mark Kac

What is good applied probability?

What is good applied probability?

Answer : see my publication list!

Wrong Answer: The Viagra Standard

If it's not hard it's not good.

In my experience papers submitted to applied probability journals are judged primarily on the difficulty and novelty of the mathematics involved.

Is the purpose of reality to inspire the creation of new probability?

OR

Is the purpose of probability to develop models and results to help us understand the world around us?

"Mathematicians are more inclined to build fire stations than to put out fires."

Kai Lai Chung in *Markov Chains*

Is there a place in applied probability journals for papers that tackle real applied problems but use mathematical arguments that are:

"well-known" to mathematicians (at least 10 people in the world understand them).

too difficult to be published in biology journals (they use more than calculus and undergraduate probability).

Think about this question as I describe two papers that are "not hard" but tackle real biological problems

Does subfunctionalization explain the persistence of gene duplicates?

Joint work with
Lea Popovic (Concordia)

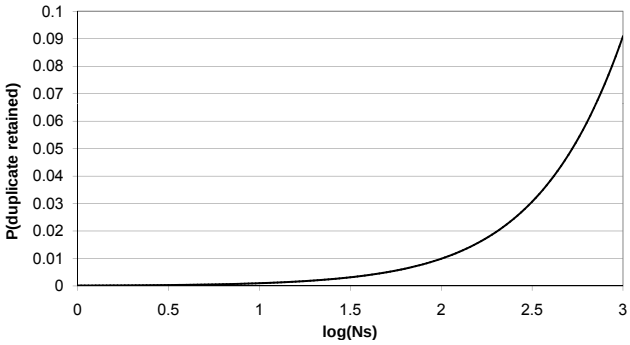
Paradox of gene duplication

About 15% of genes in the human genome are duplicates.

Gene duplications are traditionally considered to be a major evolutionary source of new protein functions. (Ohno 1970)

However, most computations predict that very few gene duplicates will have beneficial mutations that lead to new gene functions.

Walsh (1995)



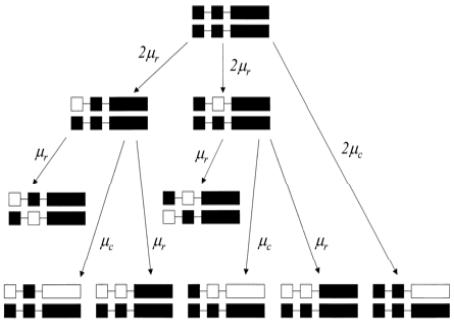
Subfunctionalization

Consider a gene regulated by two transcription factors.

μ_r = rate of regulatory region loss

μ_c = rate mutation destroys gene

Lynch and Force (2000)



P(subfunctionalization)

On N=1 chromosome, Markov chain

If $\mu_r = \mu_c$ then p = 2/9.

If $\mu_r = \mu_c$ /30 then p ≈ 1/1000.

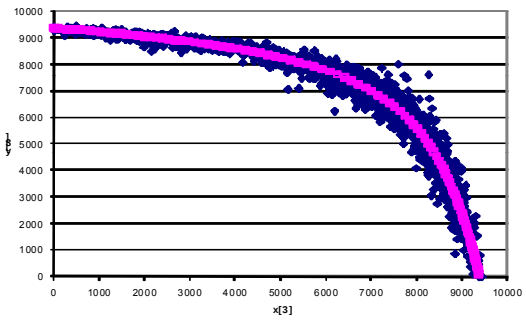
General N – use diffusion approximation

Linked (0 recomb) – tandem duplication

Unlinked (∞ recomb) – genome duplication

Ward and Durrett (2004)

simulation of diffusion in d=6 for unlinked case



Durrett and Popovic (2008)

In the limit as the population size N → ∞, the diffusion stays near a one dimensional curve of equilibria for the ODE (5 eq., 6 unk.).

The subfunctionalization outcomes are far from this curve, so by large deviations results of Wentzell-Freidlin their probability decays exponentially fast as N gets large.

Why is this trivial?

Katzenberger (1991) proved a general result for SDE's forced onto a manifold by a large drift (which we cite and is his 4th citation).

Therefore "the maths is not new and the paper should not be published in Stochastic Processes and their Applications."

Why is the referee wrong?

To prove our result we need to

1. Compute the curve of equilibria
2. Linearize around the curve and show the real parts of all five eigenvalues are negative.

Once this is done we don't need K's result.

All we need to do is to apply Ito's formula to Φ the function that projects points in the state space onto their ODE limits on the curve.

In Biology there is no QED.

In applications to biology a theorem is not the end of the story. The two math biologists who refereed (and ultimately accepted) the paper for Annals of Applied Probability had different objections.

Our result shows that in a large population **of constant size** then **in the absence of positive selection** subfunctionalization is unlikely.

To explain the objections

1. Population bottlenecks or subdivision may create small populations in which subfunctionalization can occur.
2. Fixation may be driven by positive selection (but earlier work shows this is unlikely).

Regulatory sequence evolution

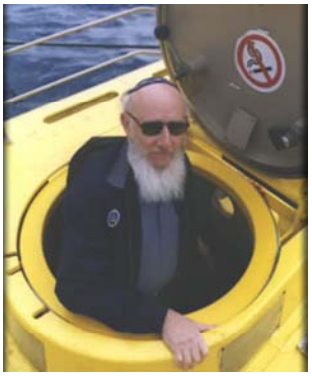
Joint with

Deena

Schmidt

IMA→MBI

Ann. Appl. Prob. 17 (2007) 1-32



We would like to thank Robert Adler for accepting it

Human and chimpanzee DNA is 98.7% identical

But there are significant phenotypic differences

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 6-9 letter word from a 4 letter alphabet to appear in a 1kb region in some individual in the population?

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 !

Eric Siggia's question to us.

Aren't these guys off by a factor of a million (for Drosophila)?

Stone and Wray's argument

Simulation for six letter window: mean 952 mutations to reach target. Take $\mu = 10^{-9}$ and divide by 2Kb gives $4.76 \circ 10^{8}$ generations

Assume individuals independent!

Divide by 2 DNA strands $\circ 10^{6}$ individuals = 238 generations

Multiply by 25 years per generation = 5950 years

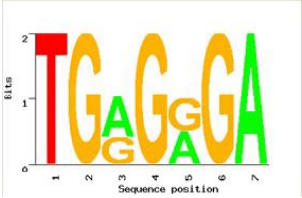
Results of our calculations (humans)

Words of length 6 in a 1 kb region evolve in exponential mean 100,000 years

If we want an exact match of an 8 nucleotide sequence then unless there is a 7 out of 8 match in the population consensus sequence this will take an average of **650,000,000 years**

Imperfect matches save the day.

Gene regulation does not require an exact match to the target word. If 7 out of 8 is good enough, then 60,000 years is enough (an intelligent design!)



Why is this trivial?

Uses mostly standard results:

Arratia-Goldstein-Gordon (1989) version of the Chen-Stein method

Aldous' Poisson clumping ideas (one success leads to others later and in nearby positions)

Calculations for the coalescent and for the Moran model

It only took us two years to work out the details.

What makes this difficult?

We are interested in words of length W = 6 to 9. It is not sensible to prove limit theorems in which W and L tend to infinity.

We want a number for the mean waiting time, not some unspecified constant.

We would like some assessment of the error involved in the approximation.

Waiting for k mutations

With Deena Schmidt
and Jason Schweinsberg

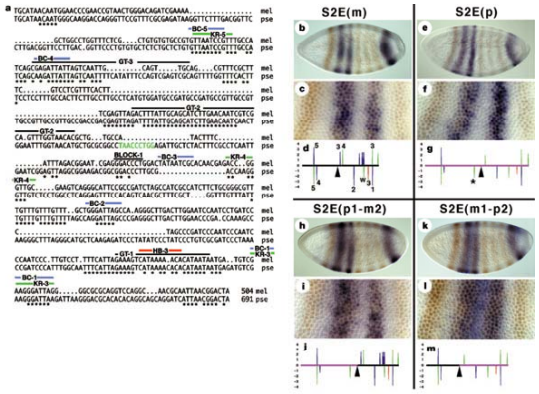
Details in Lecture 2

Jason Schweinsberg

By day mild mannered associate professor at UCSD, but on the weekend…

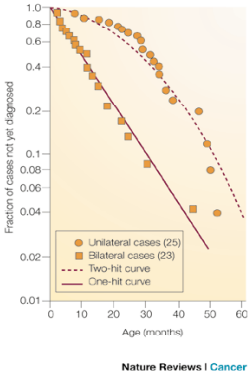
Jason talks at 15:10 on Wednesday
on Loop erased random walk
IS17 Random Processes with Interactions

Wait for 2: Eve2 in Drosophila



Retinoblastoma

Two hits = cancer



Limits to Darwinism

In the last 50 years the malaria parasite *Plasmodium falciparum* has evolved resistance to chloroquine. This is due to two amino acid altering substituions in the gene *PfCRT*.

Michael Behe in his book The Edge of Evolution calls such an event a *chloroquine complexity cluster,* or CCC.

Michael Behe concludes

There are 5000 species of modern mammals. If each species had an average of a million members and if a new generation appeared every year, and if this went on for two hundred million years, the likelihood of a single CCC appearing in the whole bunch over that entire time would only be 1 in a hundred.

This shows there are limits to evolution.

Behe is very wrong

Theorem. Suppose $Nu_1 \to 0$ and $Nu_2^{1/2} \to \infty$.
$P(T_2 > t/ Nu_1 u_2^{1/2}) \to \exp(-t)$

$N = 10^6$, $u_1 = u_2 = 10^{-9}$, the waiting time is exp mean 31.6 million generations for one prespecified pair of mutations in one species.

Not 1/100 in 5000 species in 200 million years.

Invited Session 14
Monday 14-15:45 in LT 28
Probability Problems from Genetics

Lea Popovic
Deena Schmidt
Tom Kurtz

Wald Lecture 2:
My work in Genetics with Jason Schweinsberg

1. Waiting for k mutations
2. Λ coalescents: theory and applications

If you like the movie, you'll love the book

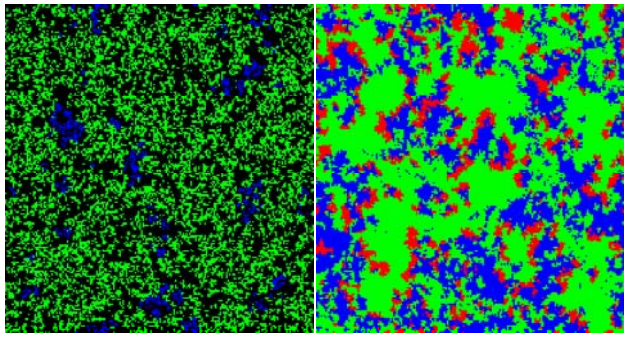
Probability Models for DNA Sequence Evolution

2nd Edition
Springer (Probability and its Applications)
432 pages, $84.95

Wald Lecture 3:
Coexistence in stochastic spatial models

Twenty years of results,
eight $1000 open problems,
and lots of pictures

Coexistence?: #1. no, #5. yes



Related Sessions

IS30 Models with Spatial Effects M 16:15-18

Ed Perkins (joint with Cox, Durrett, Merle)

Jeremy Quastel, Balint Toth

C73 Stochastic Processes 6 Fri 14:00-15:45

Daniel Remenik 15:20

Goldschmidt, Rolski, Vares, Sidoravicius