Bphys/Biol-E 101 = HST 508 = GEN224

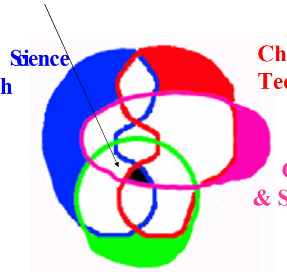
Instructor: George Church
Teaching fellows: Lan Zhang (head), Chih Liu, Mike Jones, J. Singh, Faisal Reza, Tom Patterson, Woodie Zhao, Xiaoxia Lin, Griffin Weber

Lectures Tue 12:00 to 2:00 PM Cannon Room (Boston)
Tue 5:30 to 7:30 PM Science Center A (Cambridge)
Your grade is based on five problem sets and a course project, with emphasis on collaboration across disciplines.

Open to: upper level undergraduates, and all graduate students. The prerequisites are basic knowledge of molecular biology, statistics, & computing.

Please hand in your questionnaire after this class.
First problem set is due Tue Sep 30 before lecture via email or paper depending on your section TF.

Intersection (not union) of:



Computer Science
& Math

Chemistry &
Technology

Genomics
& Systems

Biology, Ecology, Society, & Evolution

Bio 101: Genomics & Computational Biology

Tue Sep 16 *Intro 1:* **Minimal "Systems", Statistics, Computing**
Tue Sep 23 *Intro 2:* Biology, comparative genomics, models & evidence, applications
Tue Sep 30 *DNA 1:* Polymorphisms, populations, statistics, pharmacogenomics, databases
Tue Oct 06 *DNA 2:* Dynamic programming, Blast, multi-alignment, **H**idden**M**arkov**M**odels
Tue Oct 14 *RNA 1:* 3D-structure, microarrays, library sequencing & quantitation concepts
Tue Oct 21 *RNA 2:* Clustering by gene or condition, DNA/RNA motifs.
Tue Oct 28 *Protein 1:* 3D structural genomics, homology, dynamics, function & drug design
Tue Nov 04 *Protein 2:* Mass spectrometry, modifications, quantitation of interactions
Tue Nov 11? *Network 1:* Metabolic kinetic & flux balance optimization methods
Tue Nov 18 *Network 2:* Molecular computing, self-assembly, genetic algorithms, neural-nets
Tue Nov 25 *Network 3:* Cellular, developmental, social, ecological & commercial models
Tue Dec 02 Project presentations
Tue Dec 09 Project Presentations
Tue Dec 16 Project Presentations

Intro 1: Today's story, logic & goals

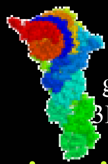
Life & computers : Self-assembly required
 Discrete & continuous models
 Minimal life & programs
Catalysis & Replication
 Differential equations
 Directed graphs & pedigrees
Mutation & the Single Molecules models
 Bell curve statistics
Selection & optimality

acgt
1
0
1
1 1
0 0
1 1
1
0
1

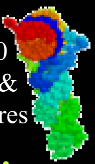
1
0
1
1
0
1

00=a
01=c
10=g
11=t

gggatttagctcagtt
gggagagcgccagact
gaa

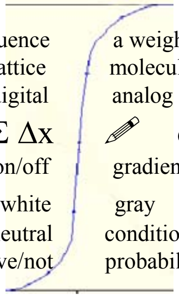


Post- 300
genomes &
3D structures



gat
ttg gag
gtcctgtgttcgatcc
acagaattcgcacca

Discrete Continuous



Discrete	Continuous
a sequence	a weight matrix of sequences
lattice	molecular coordinates
digital	analog (16 bit A2D converters)
$\Sigma\, \Delta x$	✎ dx
neural/regulatory on/off	gradients & graded responses
sum of black & white	gray
essential/neutral	conditional mutation
alive/not	probability of replication

Bits (discrete)

bit = binary digit
1 base >= 2 bits
1 byte = 8 bits

+ Kilo	Mega	Giga	Tera	Peta	Exa	Zetta	Yotta +
3	6	9	12	15	18	21	24
- milli	micro	nano	pico	femto	atto	zepto	yocto -

	Kibi	Mebi	Gibi	Tebi	Pebi	Exbi
1024 =	2^{10}	2^{20}	2^{30}	2^{40}	2^{50}	2^{60}

http://physics.nist.gov/cuu/Units/prefixes.html

Defined quantitative measures

Seven basic (Système International) SI units:
s, m, kg, mol, K, cd, A

(some measures at precision of 14 significant figures)

Quantal: Planck time, length: 10^{-43} seconds, 10^{-35} meters, mol=6.0225 10^{23} entities.

casa.colorado.edu/~ajsh/sr/postulate.html
physics.nist.gov/cuu/Uncertainty/
scienceworld.wolfram.com/physics/SI.html

Quantitative definition of life?

Historical/Terrestrial Biology vs "General Biology"

Probability of replication … of complexity from simplicity
(in a specific environment)

Robustness/Evolvability
(in a variety of environments)

Examples: mules, fires, nucleating crystals, mold-replicas, pollinated flowers, viruses, predators, geological layers, molecular ligation, factories, self-assembling machines.

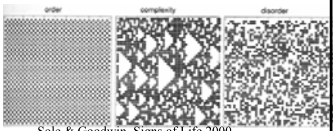
Complexity definitions

1. Computational Complexity = speed/memory scaling P, NP

2. Algorithmic Randomness (Chaitin-Kolmogorov)

3. Entropy/information

4. Physical complexity
(Bernoulli-Turing Machine)



Sole & Goodwin, Signs of Life 2000

Crutchfield & Young in Complexity, Entropy, & the Physics of Information 1990 pp.223-269
www.santafe.edu/~jpc/JPCPapers.html

Why Model?

- To understand biological/chemical data.
 (& **design** useful modifications)

- To **share** data we need to be able to **search, merge, & check** data via models.

- Integrating diverse data types can reduce random & systematic errors.

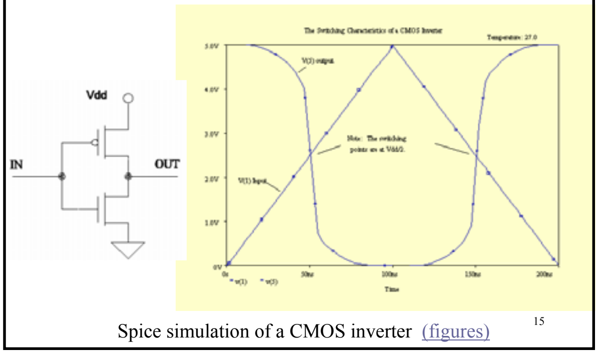
Which models will we search, merge & check in this course?

- Sequence: Dynamic programming, assembly, translation & trees.
- 3D structure: motifs, catalysis, complementary surfaces – energy and kinetic optima
- Functional genomics: clustering
- Systems: qualitative & boolean networks
- Systems: differential equations & stochastic
- Network optimization: Linear programming

Intro 1: Today's story, logic & goals

Life & computers : Self-assembly required
Discrete & continuous models
Minimal life & programs
Catalysis & Replication
Differential equations
Directed graphs & pedigrees
Mutation & the Single Molecules models
Bell curve statistics
Selection & optimality

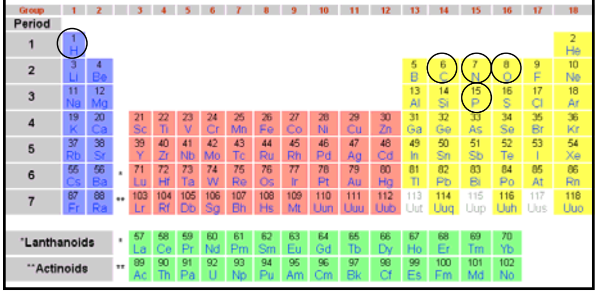
Transistors > inverters > registers > binary adders > **compilers** > application programs



Spice simulation of a CMOS inverter (figures)

Elements of RNA-based life: C,H,N,O,P

Useful for many species:
Na, K, Fe, Cl, Ca, Mg, Mo, Mn, S, Se, Cu, Ni, Co, B, **Si**



Group	1	2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period																			
1	1 H																		2 He
2	3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca		21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr		39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	*	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Uun	111 Uuu	112 Uub	113 Uut	114 Uuq	115 Uup	116 Uuh	117 Uus	118 Uuo
*Lanthanoids			*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb		
**Actinoids			**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No		

Minimal self-replicating units

Minimal theoretical composition: 5 elements: C,H,N,O,P
Environment = water, NH_4^+, 4 NTP$^-$s, lipids

Johnston et al. Science 2001 292:1319-1325 RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension.

Minimal programs

perl -e "print exp(1);"   2.71828182845905
excel: = EXP(1)   2.71828182845905000000000
f77: print*, exp(1.q0)   2.71828182845904523536028747135266
Mathematica: N[Exp[1],100]   2.71828182845904523536028747135266249775
7247093699959574966967627724076630353547594571382178525166427

- Underlying these are algorithms for arctangent and hardware for RAM and printing.
- Beware of approximations & boundaries.
- Time & memory limitations. E.g. first two above 64 bit floating point: 52 bits for mantissa (= 15 decimal digits), 10 for exponent, 1 for +/- signs.

Self-replication of complementary nucleotide-based oligomers

```
5'ccg + ccg     =>    5'ccgccg
          5'CGGCGG

CGG   +  CGG   =>    CGGCGG
          ccgccg
```

Sievers & Kiedrowski 1994 Nature 369:221
Zielinski & Orgel 1987 Nature 327:347

Why Perl & Excel?

In the hierarchy of languages, **Perl** is a "high level" language, optimized for easy coding of string searching & string manipulation. It is well suited to web applications and is "open source" (so that it is inexpensive and easily extended). It has a very easy learning curve relative to C/C++ but is similar in a few way to C in syntax.

Excel is widely used with intuitive stepwise addition of columns and graphics.

Facts of Life 101

Where do parasites come from?

(computer & biological viral codes)

AIDS - HIV-1

26 M dead (worse than black plague & 1918 Flu)

www.apheda.org.au/campaigns/images/hiv_stats.pdf

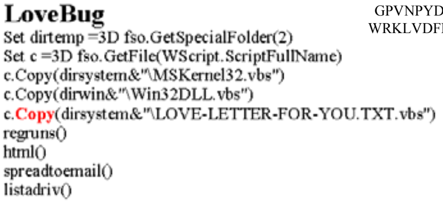
www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=11676

Polymerase drug resistance mutations

M41L, D67N, T69D, L210W, T215Y, H208Y
PISPIETVPVKLKPGMDGPK VKQWPLTEEK
IKALIEICAE **L**EKDGKISKI
GPVNPYDTPV FAIKKK**N**S**D**K
WRKLVDFREL NKRTQDFCEV

Computer viruses & hacks :
over $3 trillion/year

www.ecommercetimes.com/perl/story/4460.htm



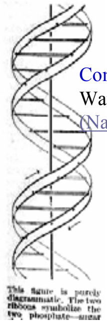
LoveBug

```
Set dirtemp =3D fso.GetSpecialFolder(2)
Set c =3D fso.GetFile(WScript.ScriptFullName)
c.Copy(dirsystem&"\MSKernel32.vbs")
c.Copy(dirwin&"\Win32DLL.vbs")
c.Copy(dirsystem&"\LOVE-LETTER-FOR-YOU.TXT.vbs")
regruns()
html()
spreadtoemail()
listadriv()
```

Conceptual connections

Concept	Computers	Organisms
Instructions	Program	Genome
Bits	0,1	a,c,g,t
Stable memory	Disk,tape	DNA
Active memory	RAM	RNA
Environment	Sockets,people	Water,salts
I/O	AD/DA	proteins
Monomer	Minerals	Nucleotide
Polymer	chip	DNA,RNA,protein
Replication	Factories	1e-15 liter cell sap
Sensor/In	Keys,scanner	Chem/photo receptor
Actuator/Out	Printer,motor	Actomyosin
Communicate	Internet,IR	Pheromones, song

Self-compiling & self-assembling

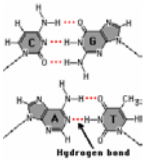


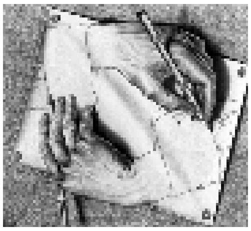
This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar

Complementary surfaces

Watson-Crick base pair

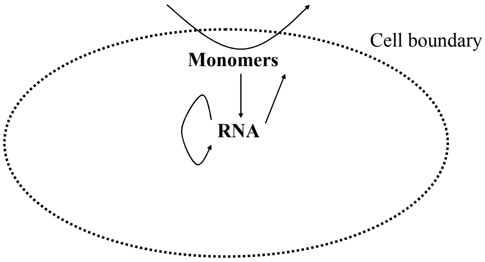
(Nature April 25, 1953)





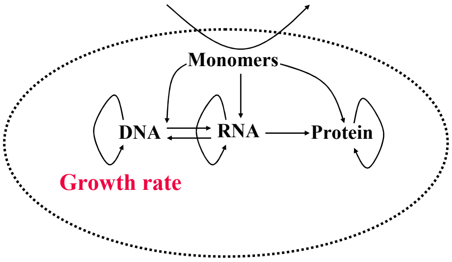
Minimal Life:

Self-assembly, Catalysis, Replication, Mutation, Selection

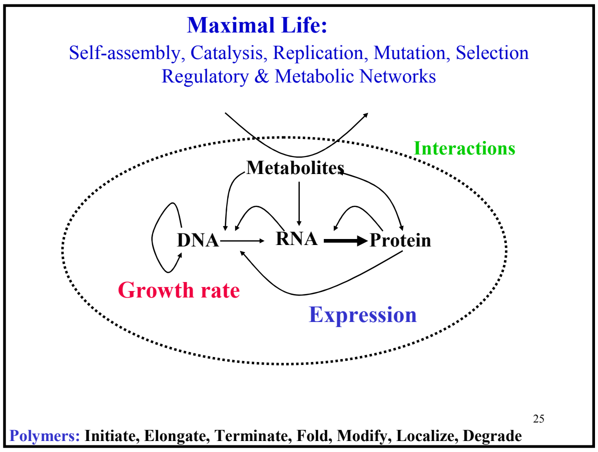


Replicator diversity

Self-assembly, Catalysis, Replication, Mutation, Selection
Polymerization & folding (Revised Central Dogma)



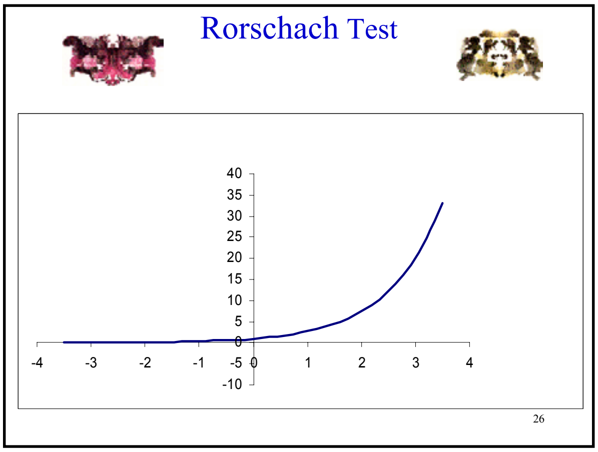
Polymers: Initiate, Elongate, Terminate, Fold, Modify, Localize, Degrade



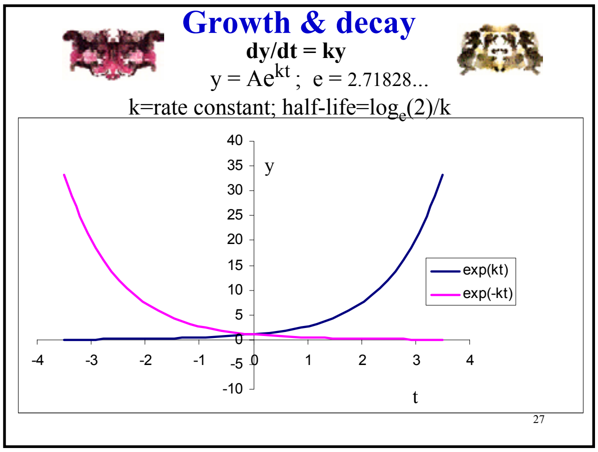
Maximal Life:

Self-assembly, Catalysis, Replication, Mutation, Selection
Regulatory & Metabolic Networks

Polymers: Initiate, Elongate, Terminate, Fold, Modify, Localize, Degrade



Rorschach Test

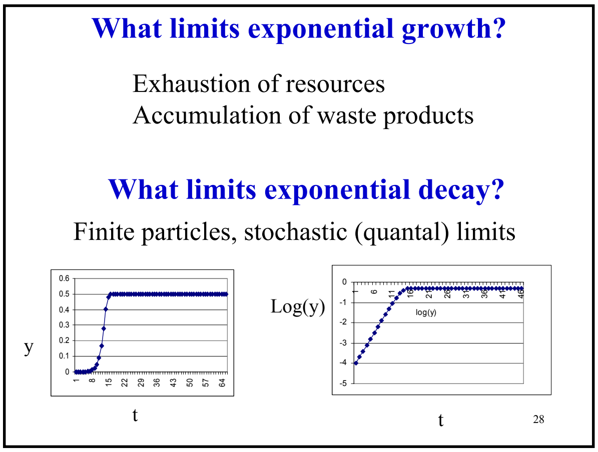


Growth & decay

dy/dt = ky

$y = Ae^{kt}$; e = 2.71828...

k=rate constant; half-life=$\log_e(2)/k$

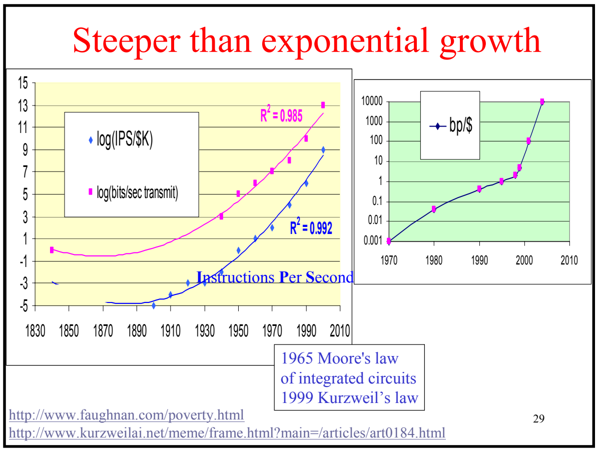


What limits exponential growth?

Exhaustion of resources
Accumulation of waste products

What limits exponential decay?

Finite particles, stochastic (quantal) limits



Steeper than exponential growth

1965 Moore's law
of integrated circuits
1999 Kurzweil's law

http://www.faughnan.com/poverty.html
http://www.kurzweilai.net/meme/frame.html?main=/articles/art0184.html

Computational power of neural systems

1,000 MIPS (million instructions per second) needed to derive edge or motion detections from video "ten times per second to match the retina … The 1,500 cubic centimeter human brain is about 100,000 times as large as the retina, suggesting that matching overall human behavior will take about 100 million MIPS of computer power … The most powerful experimental supercomputers in 1998, costing tens of millions of dollars, can do a few million MIPS."

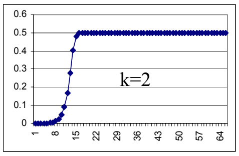
"The ratio of memory to speed has remained constant during computing history [at Mbyte/MIPS] … [the human] 100 trillion synapse brain would hold the equivalent 100 million megabytes."
--Hans Moravec http://www.frc.ri.cmu.edu/~hpm/book97/ch3/retina.comment.html

2002: the ESC is 35 Tflops & 10Tbytes. http://www.top500.org/

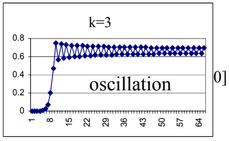
Post-exponential growth & chaos

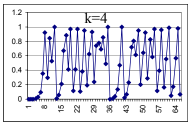
Excel:
A3=k*A2*(1-A2)
A4=k*A3*(1-A3)
…

k = growth rate
A= population size (min=0, max=1)



Smooth approach to plateau





chaos

Intro 1: Today's story, logic & goals

Life & computers : Self-assembly required

Discrete & continuous models

Minimal life & programs

Catalysis & Replication

Differential equations

Directed graphs & pedigrees

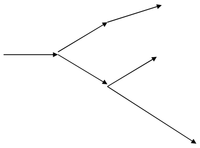
Mutation & the Single Molecules models

Bell curve statistics

Selection & optimality

Inherited Mutations & Graphs

Directed Acyclic Graph (DAG)
Example: a mutation pedigree
Nodes = an organism, edges = replication with mutation



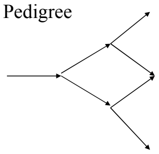
time →

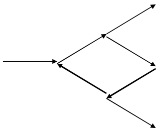
hissa.nist.gov/dads/HTML/directAcycGraph.html

Directed Graphs

Directed Acyclic Graph:
Biopolymer backbone
Phylogeny
Pedigree

Cyclic:
Polymer contact maps
Metabolic &
Regulatory Nets





Time →

Time independent or implicit
↔

System models	Feature attractions
E. coli chemotaxis	Adaptive, spatial effects
Red blood cell metabolism	Enzyme kinetics
Cell division cycle	Checkpoints
Circadian rhythm	Long time delays
Plasmid DNA replication	Single molecule precision
Phage λ switch	Stochastic expression

also, all have large genetic & kinetic datsets.

Intro 1: Today's story, logic & goals

Life & computers : Self-assembly required

Discrete & continuous models

Minimal life & programs

Catalysis & Replication

Differential equations

Directed graphs & pedigrees

Mutation & the Single Molecules models

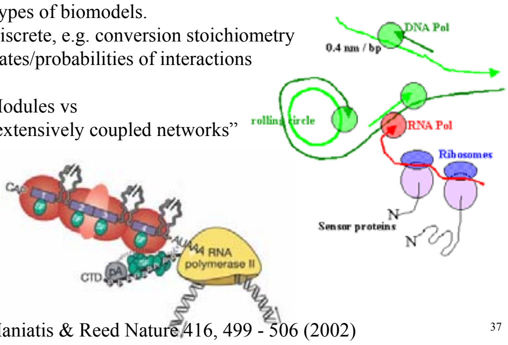
Bell curve statistics

Selection & optimality

Bionano-machines

Types of biomodels.
Discrete, e.g. conversion stoichiometry
Rates/probabilities of interactions

Modules vs
“extensively coupled networks”

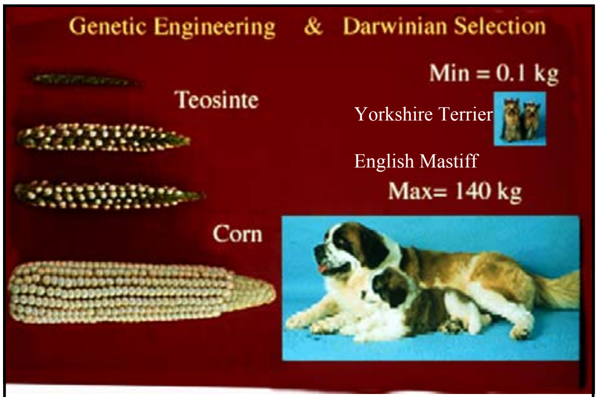


Maniatis & Reed Nature 416, 499 - 506 (2002)

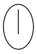
Types of Systems Interaction Models

Quantum Electrodynamics	subatomic	
Quantum mechanics	electron clouds	
Molecular mechanics	spherical atoms	**nm-fs**
Master equations	stochastic single molecules	
Fokker-Planck approx.	stochastic	
Macroscopic rates ODE	**Concentration & time (C,t)**	
Flux Balance Optima	**dC_{ik}/dt optimal steady state**	
Thermodynamic models	$dC_{ik}/dt = 0$ k reversible reactions	
Steady State	$\Sigma dC_{ik}/dt = 0$ (sum k reactions)	
Metabolic Control Analysis	$d(dC_{ik}/dt)/dC_j$ (i = chem.species)	
Spatially inhomogenous	dCi/dx	
Population dynamics	as above	**km-yr**

Increasing scope, decreasing resolution



How to do single DNA molecule manipulations?

One DNA molecule per cell 

Replicate to two DNAs.
Now segregate to two daughter cells
If totally random, **half** of the cells will have too many or too few.
What about human cells with 46 chromosomes (DNA molecules)?

Dosage & loss of heterozygosity & major sources of mutation in human populations and cancer.

For example, trisomy 21, a 1.5-fold dosage with enormous impact.

Mean, variance, & linear correlation coefficient

Expectation E (rth moment) of random variables X for any distribution f(X)

First moment= Mean μ ; variance σ^2 and standard deviation σ

$E(X^r) = \Sigma X^r f(X)$ $\mu = E(X)$ $\sigma^2 = E[(X-\mu)^2]$

Pearson correlation coefficient $C = cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]/(\sigma_X \sigma_Y)$

Independent X,Y implies C = 0,
but C =0 does not imply independent X,Y. (e.g. $Y=X^2$)

$P = TDIST(C*sqrt((N-2)/(1-C^2))$ with dof= N-2 and two tails.

where N is the sample size.

www.stat.unipg.it/IASC/Misc-stat-soft.html

Binomial frequency distribution as a function of $X \in \{int\ 0 \ldots n\}$

p and q $0 \le p \le q \le 1$ $q = 1 - p$ two types of object or event.

Factorials $0! = 1$ $n! = n(n-1)!$

Combinatorics (C= # subsets of size X are possible from a set of total size of n)

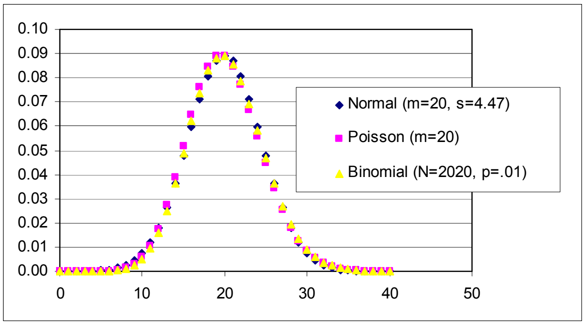
$$\frac{n!}{X!(n-X)!} = C(n,X)$$

$$B(X) = C(n, X)\, p^X q^{n-X} \quad \mu = np \quad \sigma^2 = npq$$

$$(p+q)^n = \Sigma\, B(X) = 1$$

$B(X: 350, n: 700, p: 0.1) = 1.53148\times10^{-157}$
=PDF[BinomialDistribution[700, 0.1], 350] Mathematica
~= 0.00 =BINOMDIST(350,700,0.1,0) Excel

Mutations happen



Poisson frequency distribution as a function of $X \in \{\text{int } 0 \ \ldots\infty\}$

$P(X) = P(X\text{-}1)\ \mu/X \quad = \quad \mu^x e^{-\mu}/ X! \quad \sigma^2 = \mu$

n large & p small → $P(X) \cong B(X) \quad \mu = np$

For example, estimating the expected number of positives in a given sized library of cDNAs, genomic clones, combinatorial chemistry, etc. X= # of hits.

Zero hit term = $e^{-\mu}$

Normal frequency distribution as a function of $X \in \{-\infty\ldots\ \infty\}$

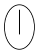
$Z= (X-\mu)/\sigma$

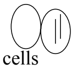
Normalized (standardized) variables

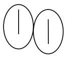
$N(X) = \exp(-Z^2/2) / (2\pi\sigma)^{1/2}$
probability density function

npq large → $N(X) \cong B(X)$

One DNA molecule per cell







Replicate to two DNAs.
Now segregate to two daughter cells
If totally random, **half** of the cells will have too many or too few.
What about human cells with 46 chromosomes (DNA molecules)?

Exactly 46 chromosomes (but any 46):
$B(X) = C(n,x)\ p^x q^{n-x}$
n=46*2; x=46; p=0.5
B(X)= 0.083

$P(X) = \quad \mu^x e^{-\mu}/ X!$
μ=X=np=46, P(X)=0.058

But what about exactly the correct 46?
$0.5^{46} = 1.4 \times 10^{-14}$

Might this select for non random segregation?

What are random numbers good for?

- Simulations.
- Permutation statistics.

Where do random numbers come from?

$X \in \{0,1\}$

```
perl -e "print rand(1);"                  0.116790771484375
0.8798828125   0.692291259765625     0.1729736328125

excel:   = RAND()   0.4854394999892640  0.6391685278993980
                    0.1009497853098360

f77:  write(*,'(f29.15)') rand(1)  0.513854980468750
                    0.175720214843750    0.308624267578125

Mathematica: Random[Real, {0,1}]          0.7474293274369694
                   0.5081794113149011  0.02423389638451016
```

Where do random numbers come from really?

Monte Carlo.

Uniformly distributed random variates X_i = remainder(aX_{i-1} / m)

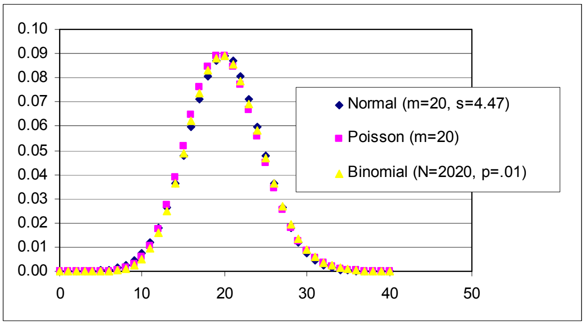
For example, a= 7^5 m= 2^{31} -1

Given two X_j X_k such uniform random variates,

Normally distributed random variates can be made

(with $\mu_X = 0$ $\sigma_X = 1$)
X_i = sqrt(-2log(X_j)) cos($2\pi X_k$) (NR, Press et al. p. 279-89)

Mutations happen



Intro 1: Summary

Life & computers : Self-assembly required
Discrete & continuous models
Minimal life & programs
Catalysis & Replication
Differential equations
Directed graphs & pedigrees
Mutation & the Single Molecules models
Bell curve statistics
Selection & optimality