Your grade is based on six 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 before Lecture 3 starts via email or paper depending on your section TF.
Bio 101: Genomics & Computational Biology

**Week#1**  *Intro 1: Computing, Statistics, Perl, Mathematica*
**Week#2**  *Intro 2: Biology, comparative genomics, models & evidence, applications*
**Week#3**  *DNA 1: Polymorphisms, populations, statistics, pharmacogenomics, databases*
**Week#4**  *DNA 2: Dynamic programming, Blast, multi-alignment, HiddenMarkovModels*
**Week#5**  *RNA 1: 3D-structure, microarrays, library sequencing & quantitation concepts*
**Week#6**  *RNA 2: Clustering by gene or condition, DNA/RNA motifs.*
**Week#7**  *Protein 1: 3D structural genomics, homology, dynamics, function & drug design*
**Week#8**  *Protein 2: Mass spectrometry, modifications, quantitation of interactions*
**Week#9**  *Network 1: Metabolic kinetic & flux balance optimization methods*
**Week#10**  *Network 2: Molecular computing, self-assembly, genetic algorithms, neural-nets*
**Week#11**  *Network 3: Cellular, developmental, social, ecological & commercial models*
**Week#12**  *Project presentations*
**Week#13**  *Project Presentations*
**Week#14**  *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 0 1 1
00=a
1 0 1 1 0 1 1
01=c
1 0 1 1 0 1 1
10=g
1 0 1 1 0 1 1
11=t
Post- 300 genomes & 3D structures
Discrete

- a sequence
- lattice
- digital
- neural/regulatory on/off
- sum of black & white
- essential/neutral
- alive/not

Continuous

- a weight matrix of sequences
- molecular coordinates
- analog (16 bit A2D converters)
- \( \Sigma \Delta x \)
- gradients & graded responses
- gray
- conditional mutation
- 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}

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=$6.0225 \times 10^{23}$ 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, pollinated flowers, viruses, predators, 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)

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

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
**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, Si

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1H</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>87</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Lanthanoids*  
**Actinoids**
Minimal self-replicating units

Minimal theoretical composition: 5 elements: C,H,N,O,P
Environment = water, NH₄⁺, 4 NTP⁻s, lipids


Minimal programs

perl -e "print exp(1);" 2.71828182845905
excel: = EXP(1) 2.718281828459050000000000
f77: print*, exp(1.q0) 2.71828182845904523536028747135266
Mathematica: N[ Exp[1],100] 2.718281828459045235360287471352662497757247093699959574966967627724076630353547594571382178525166427

• 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

\[ 5'\text{ccg} + \text{ccg} \Rightarrow 5'\text{ccgccg} \]

\[ 5'\text{CGGCGG} \]

\[ \text{CGG} + \text{CGG} \Rightarrow \text{CGGC} \text{CGG} \]

\[ \text{ccgccg} \]

Sievers & Kiedrowski 1994 Nature 369:221
Zielinski & Orgel 1987 Nature 327:347
Why Perl & Mathematica?

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.

Mathematica is intrinsically stronger on math (symbolic & numeric) & graphics.
Where do parasites come from?
(computer & biological viral codes)

Over $12 billion/year
on computer viruses (ref)
(http://virus.idg.net/crd_virus_126660.html)

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

AIDS - HIV-1 (download)

Polymerase drug resistance mutations
M41L, D67N, T69D, L210W, T215Y, H208Y
PISPIETVPVKLPGMGDGPK
VKQWPLTEEK
IKALIEICAE LEKDGKISKI
GPVNYPDTPV FAIKKKNSDK
WRKLVDFREL NKRTQDFCEV
## Conceptual connections

<table>
<thead>
<tr>
<th>Concept</th>
<th>Computers</th>
<th>Organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructions</td>
<td>Program</td>
<td>Genome</td>
</tr>
<tr>
<td>Bits</td>
<td>0,1</td>
<td>a,c,g,t</td>
</tr>
<tr>
<td>Stable memory</td>
<td>Disk, tape</td>
<td>DNA</td>
</tr>
<tr>
<td>Active memory</td>
<td>RAM</td>
<td>RNA</td>
</tr>
<tr>
<td>Environment</td>
<td>Sockets, people</td>
<td>Water, salts</td>
</tr>
<tr>
<td>I/O</td>
<td>AD/DA</td>
<td>proteins</td>
</tr>
<tr>
<td>Monomer</td>
<td>Minerals</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>Polymer</td>
<td>chip</td>
<td>DNA, RNA, protein</td>
</tr>
<tr>
<td>Replication</td>
<td>Factories</td>
<td>1e-15 liter cell sap</td>
</tr>
<tr>
<td>Sensor/In</td>
<td>Keys, scanner</td>
<td>Chem/photo receptor</td>
</tr>
<tr>
<td>Actuator/Out</td>
<td>Printer, motor</td>
<td>Actomyosin</td>
</tr>
<tr>
<td>Communicate</td>
<td>Internet, IR</td>
<td>Pheromones, song</td>
</tr>
</tbody>
</table>
Transistors > inverters > registers > binary adders > **compilers** > application programs

Spice simulation of a CMOS inverter ([figures](http://et.nmsu.edu/~etti/spring97/electronics/cmos/cmostran.html))
Self-compiling & self-assembling

Complementary surfaces
Watson-Crick base pair
(Nature April 25, 1953)
(http://www.sil.si.edu/Exhibitions/Science-and-the-Artists-Book/bioc.htm#27)
Minimal Life:
Self-assembly, Catalysis, Replication, Mutation, Selection
Replicator diversity
Self-assembly, Catalysis, Replication, Mutation, Selection
Polymerization & folding (Revised Central Dogma)

Monomers
DNA → RNA → Protein
Polymers: Initiate, Elongate, Terminate, Fold, Modify, Localize, Degrade
Maximal Life:
Self-assembly, Catalysis, Replication, Mutation, Selection
Regulatory & Metabolic Networks

Metabolites
DNA → RNA → Protein

Growth rate
Expression

Interactions

Polymers: Initiate, Elongate, Terminate, Fold, Modify, Localize, Degrade
Growth & decay

\[ \frac{dy}{dt} = ky \]

\[ y = Ae^{kt} ; \quad e = 2.71828... \]

\[ k = \text{rate constant}; \quad \text{half-life} = \frac{\log_e(2)}{k} \]
What limits exponential growth?

Exhaustion of resources
Accumulation of waste products

What limits exponential decay?
Finite particles, stochastic (quantal) limits
Solving differential equations

Mathematica: **Analytical (formal, symbolic)**

\[ \text{In[2]} := \text{DSolve\left\{y'[t] == y[t], y[0] == 1\right\}, y[t], t} \]
\[ \text{Out[2]} = \left\{\left\{y[t] = e^t\right\}\right\} \]

**Numerical (&graphical)**

\[ \text{NDSolve}\left\{y'[t] == y[t], y[0] == 1\right\}, y, \{t, 0, 3\}\]
\[ \text{Plot[Evaluate[y[t] /. %], \{t, 0, 3\}]} \]
(Hyper)exponential growth

See http://www.faughnan.com/poverty.html
See 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."


Post-exponential growth & chaos

Pop[k_][y_] := k y (1 - y);
ListPlot[NestList[Pop[1.01], 0.0001, 3000], PlotJoined->True];

\[ k = \text{growth rate}\]
\[ y = \text{population size}\]

http://library.wolfram.com/examples/iteration/iterate.nb
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

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

- *E. coli* chemotaxis
- Red blood cell metabolism
- Cell division cycle
- Circadian rhythm
- Plasmid DNA replication
- Phage $\lambda$ switch

## Feature attractions

- Adaptive, spatial effects
- Enzyme kinetics
- Checkpoints
- Long time delays
- Single molecule precision
- Stochastic expression

Also, all have large genetic & kinetic datasets.
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”

Types of Systems Interaction Models

Quantum Electrodynamics
Quantum mechanics
Molecular mechanics
Master equations
Fokker-Planck approx.
Macroscopic rates ODE
Flux Balance Optima
Thermodynamic models
Steady State
Metabolic Control Analysis
Spatially inhomogenous Population dynamics

subatomic
electron clouds
spherical atoms
stochastic single molecules
stochastic
Concentration & time (C,t)
dC_{ik}/dt optimal steady state
dC_{ik}/dt = 0 k reversible reactions
ΣdC_{ik}/dt = 0 (sum k reactions)
d(dC_{ik}/dt)/dC_j (i = chem.species)
dCi/dx
as above

Increasing scope, decreasing resolution

nm-fs
km-yr
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.
Most RNAs < 1 molecule per cell.

See Yeast RNA 25-mer array in
Wodicka, Lockhart, et al. (1997)
Nature Biotech 15:1359-67

Mean, variance, & linear correlation coefficient

Expectation $E (rth \text{ moment})$ of random variables $X$ for any distribution $f(X)$

First moment $= \text{Mean } \mu$; variance $\sigma^2$ and standard deviation $\sigma$

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

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

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

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

where $N$ is the sample size.

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

- Binomial
- Normal ($m=20, s=4.47$)
- Poisson ($m=20$)
- Binomial ($N=2020, p=0.01$)
Binomial frequency distribution as a function of $X \in \{\text{int } 0 \ldots n\}$

$p$ and $q$ \quad \quad 0 \leq p \leq q \leq 1 \quad \quad q = 1 - p \quad \quad \text{two types of object or event.}$

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

Combinatorics \quad (C = \# \text{ subsets of size } X \text{ are possible from a set of total size of } n)$

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

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

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

$B(X: 350, n: 700, p: 0.1) = 1.53148 \times 10^{-157}$

$=\text{PDF[ BinomialDistribution[700, 0.1], 350]} \quad \text{Mathematica}$

$\sim 0.00 = \text{BINOMDIST(350,700,0.1,0)} \quad \text{Excel}$
Poisson frequency distribution as a function of $X \in \{\text{int} \ 0 \ \to \infty\}$

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

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

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 $X \in \{-\infty... \infty\}$. 

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

Normalized (standardized) variables 

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

probability density function 

npq large $\rightarrow$ $N(X) \cong B(X)$
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

\[ B(X) = 0.083 \]

\[ P(X) = \frac{\mu^x e^{-\mu}}{X!} \]

\[ \mu = X = np = 46, \ P(X) = 0.058 \]

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 $X_i = \text{remainder}(aX_{i-1} / m)$

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 $\mu_X = 0$ $\sigma_X = 1$)

$$X_i = \sqrt{-2\log(X_j)} \cos(2\pi X_k) \quad (\text{NR, Press et al. p. 279-89})$$

Mutations happen.

- Binomial ($i$, $p=0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10$)
- Normal ($m=20$, $s=4.47$)
- Poisson ($m=20$)
- Binomial ($N=2020$, $p=0.01$)
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
Computation and Biology share a common obsession with strings of letters, which are translated into complex 3D and 4D structures. Evolution (biological, technical, and cultural) will probably continue to act via manipulation of symbols (A, C, G, T, 0 & 1, A-Z) plus "selection" at the highest "systems" levels. The power of these systems lies in complexity.
Simple representations of them (fractals, surgery, and drugs) may not be as fruitful as detailed programming of the symbols aided by hierarchical models and highly-parallel testing. Local decisions no longer stay local. Examples are the Internet, computer viruses, genetically modified organisms (GMOs), replicating nanotechnology, bioterrorism, global warming, and biological species transport. Information (& education) is becoming increasingly easy to spread (and hard to control). We are on the verge of begin able to collect data on almost any system at costs of terabytes-per-dollar.

The world is manipulating increasingly complex systems, many at steeper-than-exponential rates. Much of this is happening without much modeling. Some people predict a "singularity" in our lifetime or at least the creation of systems more intelligent (and/or more proliferative) than we are (possibly as little as 100 Teraflops/terabytes). We need to not only teach our students how to cope with this, but start thinking about how to teach these "intelligent" systems as if they were students. As integrated circuits reach their limit soon, the next generation of computers may be based on quantum computing and/or biologically inspired. We need to be able to teach our students about this revolution, and via the Internet teach anyone else listening.