Sequencing & understanding biosphere omes, 3D molecular design
30-Jan-2018  Biophysics 242.  George Church

Reading --------------- Writing ----------- ArithmELSI --------------
Biophysics 242: Transformative Biotechnologies (R-W-A) Refactored

Jan 23  (1) Exponential, Logistic, Prioritizing global challenges
Jan 30  (2) Sequencing & understanding biosphere omes, 3D molecular design
Feb  6  (3) Genome Edit/Write (beyond CRISPR), Codon Recoding
Feb 13  (4) In situ Sequencing: Mammalian Cell Atlas & BRAIN Initiative
Feb 20  (5) Epigenetic programming, signaling pathways, SynEvoDevo
Feb 27  (6) Microbiomes: therapeutics, diagnostics, nanopores
Mar  6  (7) Synthetic Organs for VUS & Transplantation
Mar 20  (8) Aging Reversal
Mar 27  (9) Global Warming
Apr  3  (10) Germline editing & *H. sapiens* 2.0
What should we do with technologies?

- Global warming
- Infectious diseases (poverty)
- Psychiatric & cognitive
- Aging reversal
- Backup planet colonies

Climate change skeptics: wattsupwiththat.com
Problem #1A: Please start your shared notes

openwetware.org/wiki/Harvard:Biophysics_101/2018

Each student will participate in projects/problems via web page or wiki.

No prerequisites. Each of you brings some expertise to be integrated with the goals and talents of other team members.
Global warming solutions brainstorm (started in class)
CO₂ to sugar via biophotosyn or electricity
Wind, photovoltaic, nuclear, geothermal
Bioelectricity
Population control
Electric cars
Syn ozone

Post-brainstorm: deeper dive & quantitative considerations
Optimize Mkm²: 13 arctic, 19 deserts, 50 agricultural, 40 forests
Carbon Gton: 1,400 tundra, 850 Air, 375 Tropical forests
Proxima Centauri b
4.2 light years, 20% light speed
Meter-scale sails
Gigawatt hours per launch
Cameras, thrusters, power, navig/commun.
Gram-scale ‘nanocraft’
A Roadmap to Interstellar Flight, Philip Lubin 2016 Arxiv

What about ng-scale? Deceleration? Build communication on arrival?
What will we take with us?
Space Organ & Ecosystem \(\rightarrow\) Omes

Gravity (0% ISS, 38% Mars)
Radiation
Limited Biome
Neuro-behavioral
## Biopolymers → Omes

<table>
<thead>
<tr>
<th>Name</th>
<th>(Examples)</th>
<th>Max length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynucleotides (DNA, RNA…)</td>
<td></td>
<td>300M</td>
</tr>
<tr>
<td>Polypeptides (collagen, vancomycin…)</td>
<td></td>
<td>100K</td>
</tr>
<tr>
<td>Polyketides (fats, tetracycline)</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Polysaccharides (cellulose, starch…)</td>
<td></td>
<td>∞</td>
</tr>
<tr>
<td>Polyterpenes (cholesterol, rubber)</td>
<td></td>
<td>16K</td>
</tr>
<tr>
<td>Polyaminoacids (lignin, polyalkaloids…)</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>Polypyrroles (heme, vitamin B12 …)</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Polyesters (PHA, PHV, P3HB, P4HB…)</td>
<td></td>
<td>30K</td>
</tr>
</tbody>
</table>
Self-compiling & self-assembling

Complementary surfaces
Watson-Crick base pair
(Nature April 25, 1953)

M.C. Escher
Side chains: **AEGIS**, Romesberg, Aptamers

Backbones: **GNA**, **HNA**, **LNA**, **PNA**, **RNA**, **TNA**, **XNA**...
A semi-synthetic organism that stores & retrieves increased genetic information. Zhang ... Romesberg 2017 Nature.
Non-Watson-Crick base pairs & their associated isostericity matrices. Leontis ... Westhof 2002 NAR
Nucleotides That Lack Tautomeric, Protonated, or Deprotonated Versions Complementary to Natural Nucleotides.

Winiger CB ... Benner S (2017) ACS Synth Biol.
**Backbones:** diversity of artificial genetic polymers. Anosova ... Egli. NAR 2015

CeNA: Cyclohexene  
HNA: 2,3-dideoxy-1,5-anhydro-Darabino-hexitol nucleic acid,  
hNA: 2,3-dideoxy-β-D-glucopyranose  
TNA: α-L-threofuranosyl-(3→2)
22 Proteinogenic Amino Acids
Non-standard Amino Acids

<table>
<thead>
<tr>
<th>AA</th>
<th>#atoms(non-H)</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ketone</td>
<td>9</td>
<td>Orthogonal chem</td>
</tr>
<tr>
<td>Azobenzyl</td>
<td>14</td>
<td>Photoisomers</td>
</tr>
<tr>
<td>Coumarinyl</td>
<td>13</td>
<td>Redox</td>
</tr>
<tr>
<td>Cytosinyl</td>
<td>8</td>
<td>Peptide-NA</td>
</tr>
<tr>
<td>Dansyl</td>
<td>17</td>
<td>Fluorescent</td>
</tr>
<tr>
<td>Biotinyl-Lys</td>
<td>19</td>
<td>Binds Avidin</td>
</tr>
</tbody>
</table>
Chirality

19 L-amino acids:
H toward you; CO R N clockwise.
Mirror world: מירור העולם
B-DNA & α-helical protein chirality
Left (mirror) Right (standard)

0.34 nm

CH₂OH

CH₂OH

OH

OH

COOH

COOH

NH₂

NH₂
Problem #2A: accessing GenBank

github.com/LennonLab/GenomeDorm/blob/master/scripts/fetch-genomes.py

Y chromosome: PCDH11Y gene

ncbi.nlm.nih.gov/gene/83259

ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi?REQUEST=CCDS&DATA=CCDS14776.1

&rettype=gb&from=5000044&to=5742228

Code to find codons?
Problem #2B: How to understand Omes?

How to apply such understanding?

- Phylogenetic alignments
- Molecular simulations
- Crosslink sequencing (ENCODEx) vs in situ imaging
- Genome-wide Association (GWAS) vs Causality tests
Holley et al. (1965)
Science 147: 1462-1465

Kim...Rich (1974)
PNAS 71:4970-4974

Fig. 2. Schematic representation of three conformations of the alanine RNA with short, double-stranded regions.
Molecular simulations
(Energy minimization, trajectories, approximations)

Quantum Electrodynamics (QED) Schwinger
Born-Oppenheimer Approximation

Quantum Engines Molecular Orbital Methods
Semiempirical Hartree-Fock methods
Modified Intermediate Neglect of Differential Overlap (MINDO)
Modified Neglect of Diatomic Overlap (MNDO) - AMPAC, MOPAC
SemiChem Austin Model 1 (SAM1) - Explicitly treats d-orbitals.
ab initio Hartree-Fock programs: GAMESS, Gaussian

Semiempirical Engines (Molecular Mechanics) from above & spectroscopy
AMBER, Discover, SYBYL, CHARMM, MM2, MM3, ECEPP.
(Chemistry at HARvard Molecular Mechanics)

Exact Stochastic Simulation of Coupled Chemical Reactions.
Molecular mechanics

\[ F = m \ a \]
\[-dE/dr_i = F_i = m_i \ d^2r_i/dt^2 \quad r = \text{position (radius)}\]
\[ dt \sim 1 \ fs (1e-15 \ sec) \]
\[ v_i(t+dt/2) = v_i(t-dt/2) + a_i(t) \ dt \quad \text{update velocity & } r \]
\[ r_i(t+dt) = r_i(t) = v(t+dt/2)dt \]

\[ E = E_b + E_\theta + E_\omega + E_{vdw} + E_{\text{electrostatic}} \]

\[ E_b = 0.5 \ k_b (r-r_0)^2 \]
\[ E_\theta = 0.5 \ k_\theta (\theta - \theta_0)^2 \]
\[ E_\omega = k_\omega \left[ 1 + \cos(n \omega - l) \right] \]
\[ E_{vdw} = A(r/r_{v0})^{-12} - B(r/r_{v0})^{-6} \]
\[ E_{\text{electrostatic}} = \frac{q_i q_j}{e r} \]

(Ref)
Molecular modeling examples

- Ribosome
- Aminoacyl-tRNA synthetases
- Essential proteins NSAA
- AAV
- Pore-polymerase fusions
- Recombination machinery
- Allosteric DNA binding

World’s only fully open access human Genomic, Environmental, Trait data & cells

NIST + FDA + PGP = Genomeinabottle.org
Ancestrally diverse trios.

NIH ENCYclopedia Of DNA Elements
11 isogenic cell types.

Critical Assessment of Genome Interpretation

US, Canada, UK, Austria, Korea, China
World’s only fully open access human Genomic, Environmental, Trait data & cells

Fibroblasts, B-cells, PSCs, excitatory/inhibitory neurons, pyramidal neurons, oligodendrocytes, endothelial cells, myocytes, cardiomyocytes, stromal cells.

NIH ENCYclopedia Of DNA Elements 11 isogenic cell types.

Epigenetic tools: Hi-C, Chip-Seq, RNA-Seq, RAMPAGE, smallRNA, DNaseSeq
GWAS: Multigenics (e.g. height): too many genetic + environmental components?

N=253,288  9500 SNPs : 29% of phenotypic variance
... but some variants have large effects → causality tests

Growth hormone therapies & non-GH alleles

- Turner syndrome
- Chronic renal failure
- Prader–Willi syndrome
- Intrauterine growth retardation
- idiopathic short stature
- AIDS Muscle wasting

FDA Nov-2015: "AquAdvantage salmon is as safe to eat as any non-genetically engineered (GE) Atlantic salmon, and also as nutritious."
Rorschach

Normal (m=20, s=4.47)
Poisson (m=20)
Binomial (N=2020, p=.01)
Bell-curves

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

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

Factorials  \[ 0! = 1 \quad n! = n(n-1)! \]

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

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

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

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

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

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

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

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

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

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 $= e^{-\mu}$
Normalized (standardized) variables

\[ Z = \frac{(X - \mu)}{\sigma} \]

probability density function

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

npq large \( \rightarrow \) \( N(X) \approx B(X) \)

Normal frequency distribution as a function of \( X \in \{-\infty... \infty\} \)
Utility of random numbers?

- Simulations
- Permutation statistics
- DBPCRCT

![Graph showing utility of random numbers](image)
Where do random numbers come from?

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

**Perl**

```
perl -e "print rand(1);"
```

0.116790771484375 0.8798828125 0.692291259765625

**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?

Hardware: Monte Carlo. Classical & Quantum
Software: PRNGs

Uniformly distributed \( X_i = \text{Mod}(aX_{i-1}, m) \) e.g., \( a = 7^5 \quad m = 2^{31} - 1 \)

Normally distributed random variates (with \( \mu_X = 0 \quad s_X = 1 \))
\( X_n = \sqrt{-2\log(X_j)} \cos(2\pi X_k) \)

xoroshiro128+ (XOR/rotate/shift/rotate)
fastest full-period generator passing BigCrush without systematic failures. Longer period: xorshift1024*

Numerical Recipes Press et al. p. 279-89
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 are the odds of getting correct segregation of all human DNA molecules (chromosomes) during mitosis if movement is random?**

Significance: 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.