Sequence Analysis 2015 — Lecture 24

Single Nucleotide Polymorphisms + Systems Biology
What are polymorphisms?

- Genetic differences between individuals in a population.
- Changes related to alleles
  - **Single nucleotide polymorphisms** (one base substitution)
    - Noncoding
    - Coding
      - synonymous -- same amino acid, different codon
      - non-synonymous
        - missense -- change in amino acid
        - nonsense -- stop codon
  - **Frame-shifts**
    - One or more base Insertion/deletion

SNPs = single nucleotide polymorphisms
How To View all SNPs associated with a gene

Starting with...

a gene name

1. Search the Gene database with the gene name. If you know the gene symbol and species, enter them as follows: tpo[gene] AND human[orgn]
2. Click on the Gene ID of the desired gene.
3. In the list of Links on the right, click "SNP:GeneView". If the link is not present, no SNPs are currently linked to this gene.

a nucleotide or protein accession number (e.g. NM_001126)

1. Search the Nucleotide or Protein database with the accession number.
2. In the Links menu in the upper right, click on "GeneView in dbSNP". If the link is not present, click on the "Gene" link in the same menu and continue at step 3 above under "a gene name".

a nucleotide sequence

1. Go to the BLAST home page and click "nucleotide blast" (blastn) under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. Continue at step 2 under "a nucleotide or protein accession number" above.

a protein sequence

1. Go to the BLAST home page and click "protein blast" (blastp) under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. Continue at step 2 under "a nucleotide or protein accession number" above.
Search Gene database

### Search results
**Items:** 2

<table>
<thead>
<tr>
<th>Name/Gene ID</th>
<th>Description</th>
<th>Location</th>
<th>Aliases</th>
<th>MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPO</td>
<td>thyroid peroxidase</td>
<td>Chromosome 2, NC_000002.12</td>
<td>MSA, TDH2A, TPX</td>
<td>606765</td>
</tr>
<tr>
<td></td>
<td>[Homo sapiens (human)]</td>
<td>(1413461..1542727)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THPO</td>
<td>thrombopoietin</td>
<td>Chromosome 3, NC_000003.12</td>
<td>MGDF, MKCSF, ML, M PLLG, THCYT1, TPO</td>
<td>600044</td>
</tr>
<tr>
<td></td>
<td>[Homo sapiens (human)]</td>
<td>(184371935..184379688, complement)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Gene ID**

Contains basic description, gene location in chromosome coordinates
Click on Gene ID
exons
swipe on ruler.
select
zoom on range

Genomic regions, tran
Genomic Sequence: NC_000002.12: 1.4M..1.6M

NC_000002.12: 1.4M..1.6M
Genes, NCBI Homo sapien
NM_000547.5
NM_175719.3
NM_01206744.1
NM_01206745.1
XM_01150300.1
NM_175721.3
XM_01150381.1
NM_175722.3
XM_01150379.1
XM_01150382.1
LOC105373362
XR_922719.1

LOC105373362
XR_922719.1

Go to reference sequence detail
Graphics FASTA GenBan
Tools Tracks

Range ToolTip
Range: 1492355 .. 1498015

Zoom On Range
Zoom To Sequence
Modify Range
Set New Marker For Selection
BLAST Search (Selection)
Primer BLAST (Selection)
Download FASTA (Selection)
SNPs
again, swipe and zoom...
again...
This one is “missense” Asp —> Glu mutation.
Go back to GenBank entry. On the right, find GeneView in the menu.
# Same SNP, in GeneView

<table>
<thead>
<tr>
<th>contig reference</th>
<th>G</th>
<th>Asp [D]</th>
<th>1</th>
<th>707</th>
</tr>
</thead>
<tbody>
<tr>
<td>1496109 2218</td>
<td>rs199561216</td>
<td>0.000</td>
<td>0.0002</td>
<td>Yes</td>
</tr>
<tr>
<td>1496115 2224</td>
<td>rs377151900</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1496127 2236</td>
<td>rs732608</td>
<td>0.493</td>
<td>0.4389</td>
<td>Yes</td>
</tr>
<tr>
<td>1496128 2237</td>
<td>rs138891531</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1496133 2242</td>
<td>rs775227003</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1496134 2243</td>
<td>rs744314413</td>
<td>0.000</td>
<td>0.0002</td>
<td>Yes</td>
</tr>
<tr>
<td>1496137 2246</td>
<td>rs201782331</td>
<td>0.000</td>
<td>0.0002</td>
<td>Yes</td>
</tr>
<tr>
<td>1496144 2253</td>
<td>rs761527805</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1496152 2261</td>
<td>rs764757383</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[back to SNP summary page]
[back to the chromosome browser]
High LD* → Low/no Recombination
\(r^2 \approx 1\) SNP1 “tags” SNP2

Low LD* → More Recombination
Many possibilities. \(r^2 << 1\)

*Linkage disequilibrium
Population descriptors: YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe)
Linkage Disequilibrium

Triangle plot shows LD values using $r^2$ or D'/LOD scores in one or more HapMap populations.

Phased haplotype track shows all 120 chromosomes with alleles colored yellow and blue.
Tagging SNPs, tSNPs

• SNPs that are highly correlated are redundant information

• tSNPs are selected as the minimal non-redundant set of SNPs in a population, such that the genotypes can be reconstructed from the tSNPs.

• tSNPs allow genotyping with fewer steps
  – PCR amplification experiments determine which base is present.

• Block based tagging

Block based tagging requires that haplotype "blocks" first be inferred. In the majority of cases when you are investigating association within a candidate gene you are likely to start off with a large number of potential SNPs to choose from, and using various measures of linkage disequilibrium and inferred haplotypes it is possible to define 'haplotype blocks' of markers that are in strong LD with each other, but not with those in other blocks. The exact definition of a haplotype block is open to interpretation, and there are a number of different methods for choosing your haplotype blocks (Gabriel et al 2002, )
tSNPs (tag SNPs, Tagging SNPs) are SNPs that correlate strongly with other SNP, therefore they give information on the haplotype without having to sequence the whole genome.

Highly correlated SNP have high LD. SVD finds repeating, nearly identical rows in the haplotype matrix.
Uses of SNPs

• Personalized medicine
  – SNP-associated Sensitivity to
    – diseases
    – drugs
    – chemicals
    – pathogens
    – vaccines

• Livestock breeding

• Human migrations
Human migration out of Africa

Each individual has a set of SNPs and a family history, marking geographic location. Tree is built ignoring the geographic location, to model migration paths.

mtDNA: mitochondrial DNA
NRY: nonrecombining region of the Y chromosome

systems biology models

- Stocks and flows
- Finite Difference simulation
- Evolving model
- Training the model
- Systems engineering
**Stocks and flows**

**Stock**: represents a physical quantity, such as the concentration of ATP. Includes a description and a quantity.

**Flow**: represents conversion of one or more stock(s) into other stocks. Includes a rate equation and associated variables, which can themselves be stocks.

![Diagram showing ATP, ADP, Pi, and ATPase with hydrolysis and ATPase reactions](image-url)
Stocks are compounds, which have concentrations. Flows are reactions, which depend on the concentration (expression levels) of enzyme stocks, and the concentration of cofactor stocks (not shown in this graphic).
Simple finite difference simulation

1. For each flow, multiply by time to get flux.
2. Subtract flux from source, add flux \((x\) stoicheometry) to sink.
3. Repeat for all flows.

Assumes no feedback, inhibition, cooperativity…

\[
\frac{T}{\Delta t} = \text{number of time steps}
\]

matrix of flows from stock \(i\) to stock \(j\)

\[
t=0 \quad t=T
\]

stocks, initial & final

Fecal metagenomics was applied to subjects given live-culture yogurt or acidified milk product. The abundance of certain species explained the degree of butyric acid produced.
Example of systems model for protein folding: GeoFold

A protein structure is treated as a geometric chain, with pivot points and hinges. Each pivot or hinge produces two new smaller “intermediate” structures.
GeoFold: Directed acyclic graph of unfolding pathways

fully folded state (1).

flows (pivot)

stocks (folding intermediates)

Two products for each flow

small pieces.
GetCuts sets up the DAG. UnFoldSim does the finite difference simulation.

GetCuts

UnFoldSim

// ----------------- Variables -----------------
// f, u1, u2 = {coordinate sets}. (nodes)
// cut = {f, u1, u2, cuttype}. (edges)
// DAG = directed acyclic graph. {nodes, edges}.
// cuttype = (break|pivot|hinge|melt).
// kf(cut), ku(cut) = rate constants for folding, unfolding.
// C(f), dc(f) = concentration of f, rate of change in concentration of f.
// t, dt, f0 = time, time step, protein concentration

// ---------------- Functions -----------------
// GetBreak, GetPivot, GetHinge find new geometrically possible cuts,
// returning False when there are no more.
// Split(f, cuttype) returns two subsets of f using cuttype.
// Split(f, melt) returns {∅, ∅}
// Exists(f, DAG) returns True if f is already in DAG.

GeoFOLD:
1. f = entire protein
2. DAG = ∅
3. GetCuts(f, DAG)
4. UnFoldSim(DAG)

GetCuts(f, DAG):
5. If (f == ∅) Return
6. If Exists(f, DAG) Return
7. While (cuttype != melt) {
8. If (GetBreak(f)) {cuttype = break}
9. ElseIf GetPivot(f)) {cuttype = pivot}
10. ElseIf GetHinge(f)) {cuttype = hinge}
11. Else {cuttype = melt}
12. {u1, u2} = Split(f, cuttype)
13. DAG = DAG u {f, u1, u2, cuttype}
14. GetCuts(u1, DAG)
15. GetCuts(u2, DAG)
16. }

UnFoldSim(DAG):
17. C(0) = f0
18. For All cut ∈ DAG {calculate kf(cut) and ku(cut)}
19. While Not converged {
20. t += dt
21. For All f { dc(f) = 0 }
22. For All cut ∈ DAG {
23. dc(f) = dc(f) - C(f)*ku(cut) + C(u1)*C(u2)*kf(cut)
24. dc(u1) = dc(u1) + C(f)*ku(cut) - C(u1)*C(u2)*kf(cut)
25. dc(u2) = dc(u2) + C(f)*ku(cut) - C(u1)*C(u2)*kf(cut)
26. }
27. For All f { C(f) = C(f) + dc(f)*dt }
28. Plot t, C
29. If (All |dc| < VerySmall) Then converged
30. }
Results of finite difference simulation starting with 100% folded state.

\[ k_u = \frac{\ln(2)}{t_{1/2}} \]
GeoFold reproduces experimental folding data in disulfide mutants: DHFR

ECOME - a systems model with an evolving structure

Stocks

- Species are measured in units of “biomass”
• Species can be autotrophs (green) or heterotrophs (red)
Flows — carbon metabolism

- Plants catalyze $\text{CO}_2 \rightarrow \text{CH}$

Plants grow proportional to biomass.
Constraints

• The sun's maximum total input to the food web is fixed.

All plants stop growing when the sum total biomass $\geq$ sun limit
• Primary consumers (herbivores) get biomass from plants
Secondary consumers (carnivores) get biomass from other animals.

etc.
• All species also catalyze CH $\rightarrow$ CO$_2$.
i.e. All species lose biomass to respiration and natural death, at a constant rate.
• Predator species collapse when prey is scarce.

\[
\text{new mass needed to survive} = \text{current mass} \times k.
\]

Plant mass that is left
...along with the prey species.
• Fed fraction grows
• Unfed fraction dies.
• ...becomes CO$_2$.

...unfed fraction dies

...collapse follows on the next cycle.
Variables.

- **Holling response functions** modify predator/prey relationship.

Predators starve when prey is scarce.

**Canadians versus Cod**

Too small. Predator can't find it.
Variables.

- Speciation. Species split to create new species with selective advantage(s).

http://www.bioinfo.rpi.edu/bystrc/ecome/local.mpeg
ECOME: boom/bust cycles
ECOME: Introducing hyper-evolution
InsightMaker is a free systems modeling environment. Users can create stocks, flows, and variables, set their values and run the model. Model values can be optimized to fit real data (e.g. population measurements).
The World Model

The Limits to Growth, Meadows et al, 1972
Systems engineering

Systems biology

1. Omics data
   - Design of experiment and hypothesis-testing

9. Experimental validation
   - Design of experiment and hypothesis-testing
   - Non-validated details

8. Genome-scale metabolic model
   - Model construction
   - Pathway gaps

7. Pathway databases and literature

Bioinformatics

2. Poorly and unannotated genes/proteins
   - Data

6. Annotation database
   - Pathway gaps
   - Storage of annotations
   - Annotations for modeling and gap-filling

5. Annotated genes/proteins
   - Synthetic pathways
   - Genetic bottlenecks

3. High-throughput annotation
   - Poorly and incorrect annotations

4. Second-generation annotation
   - General annotations

10. Construct design

11. Improved genome

Metabolic engineering/synthetic biology
1. What is a SNP?
2. How do we locate SNPs?
3. What are Tagging SNPs?
4. What is a haplotype?
5. What is linkage disequilibrium?
6. What is systems biology?
7. What are the components of a system model?
8. How does a finite difference simulation work?
9. How can metagenomics feed into systems biology?
10. What are stocks? What are flows?