Announcements

  - Lecture slides, problem sets, etc.
- Course communications via Piazza
  - Auditors please sign up too

- HW2 will be posted later today.

Class Topics

1. Genome context:
   - cells, DNA, central dogma
2. Genome content / genome function:
   - genes, gene regulation, epigenetics, repeats, SARS-CoV-2
3. Genome sequencing:
   - technologies, assembly/analysis, technology dependence
4. Genome evolution:
   - evolution = mutation + selection, main forces of evolution:
     Neutral evolution, Negative selection, Positive selection
5. Population genomics:
   - Human migration, paternity testing, forensics, cryptogenomics
6. Genomics of human disease:
   - personal genomics, GxE disease types, deep dive monogenics
7. Comparative Genomics:
   - Genomics of amazing animal adaptations, ultraconservation
"non coding" RNAs (ncRNA)

Central Dogma of Biology:

DNA → Transcription (Polymerases) → mRNA → Translation (Ribosome, tRNA) → Protein

Active forms of “non coding” RNA

DNA → Transcription (Polymerases) → mRNAs, miRNA, snoRNA, snRNA → Reverse transcription (of eg retrogenes) → RNA → Translation (Ribosome, tRNA) → Protein
What is ncRNA?

- Non-coding RNA (ncRNA) is an RNA that functions without being translated to a protein.
- Known roles for ncRNAs:
  - RNA catalyzes excision/ligation in introns.
  - RNA catalyzes the maturation of tRNA.
  - RNA catalyzes peptide bond formation.
  - RNA is a required subunit in telomerase.
  - RNA plays roles in immunity and development (RNAi).
  - RNA plays a role in dosage compensation.
  - RNA plays a role in carbon storage.
  - RNA is a major subunit in the SRP, which is important in protein trafficking.
  - RNA guides RNA modification.
  - RNA can do so many different functions, it is thought in the beginning there was an RNA World, where RNA was both the information carrier and active molecule.

"non coding" RNAs (ncRNA) subtype:

Small structural RNAs (ssRNA)

Remember with protein coding genes?

Gene (DNA) sequence determines protein (AA) sequence, which determines protein (3D) structure, which determines protein’s function.
It's the same with small structural RNAs!

For example, tRNA Structure...

...Facilitates tRNA Function

RNP = RNA-protein Complexes

Proteins and ncRNA co-operate. For example the ribosome is a complex of proteins+RNAs.

Predicting ssRNA Structure

To a first approximation: Structural RNAs fold into their most energetically stable structure. An RNA structure stores its energy in the form of bonds between complementary nucleotides.

RNA structure prediction challenge: Given a ssRNA sequence, find its most energetically favorable possible structure, or in other words, the structure which maximizes the energy stored in its bond — this is the predicted structure of this ssRNA.

ssRNA structure rules

- Canonical basepairs:
  - Watson-Crick basepairs:
    - G = C
    - A = U
  - Wobble basepair:
    - G = U
- Stacks: continuous nested basepairs (energetically favorable)
- Non-basepaired loops:
  - Hairpin loop
  - Bulge
  - Internal loop
  - Multiloop
- Pseudo-knots
Ab initio RNA structure prediction: Recursive formulation

- Objective: Maximizing # of base pairings (Nussinov et al., 1978)

Simple model: $w(i, j) = 1$ iff GC(AU/GU) pair

Fancier model: GC > AU > GU

Dynamic programming

- If you naively try to evaluate $S(0, n)$ using recursion – it will work, but you will be re-evaluating the same terms over and over.

- Dynamic programming is a clever way of organizing term calculations, so that each is called & computed exactly once.
  - Compares a sequence against itself in a dynamic programming matrix

- Three steps

1. Initialization

   Example: GGGAAAUCC

   \[ S(i, i) = 0 \quad \forall \quad 1 \leq i \leq L \rightarrow \text{the main diagonal} \]
   \[ S(i, i - 1) = 0 \quad \forall \quad 2 \leq i \leq L \rightarrow \text{the diagonal below} \]

   $L$: the length of input sequence
2. Recursion

Fill up the table (DP matrix) — diagonal by diagonal

\[
S(i, j) = \begin{cases} 
S(i+1, j-1) + w(i, j) & \text{if } i, j \text{ are complementary} \\
S(i+1, j) & \text{otherwise}
\end{cases}
\]

3. Traceback

The structure is:

What are the other “optimal” structures?

Pseudoknots drastically increase computational complexity
Objective: Minimize Secondary Structure

Free Energy at 37 °C:

Instead of \((i, j)\), measure and sum energies:

\[
AG_{\text{Hairpin}} = AG_{\text{CGCG}} + 2AG_{\text{GUA}} + 2AG_{\text{CGU}} = -2.0 \text{ kcal/mol} + 2(0.9) \text{ kcal/mol} + 2(-0.9) \text{ kcal/mol} = -7.7 \text{ kcal/mol}
\]

\[
AG_{\text{Helix}} = AG_{\text{GCGCG}} + 4AG_{\text{AT}} = 5.0 \text{ kcal/mol} + 4(-1.6) \text{ kcal/mol} = 3.4 \text{ kcal/mol}
\]

\[
AG_{\text{Total}} = AG_{\text{Helix}} + AG_{\text{Hairpin}} = 3.4 \text{ kcal/mol} - 7.7 \text{ kcal/mol} = -4.3 \text{ kcal/mol}
\]


---

Zuker’s algorithm MFOLD: computing loop dependent energies

\[
W_{ij} = \min \left\{ W_{i+1,j}, \right. \\
\left. W_{i,j-1}, \right. \\
\left. \frac{V_{i,j}}{W_{K,j}}, \right. \\
\left. \frac{V_{i,j}}{W_{M,j}} \right\}
\]

\[
V_{ij} = \min \left\{ W_{ij}, \right. \\
S_{ij}, \\
V_{Bij}, \\
V_{Mij} \right\}
\]

\[
V_{Bij} = \min \left\{ (I_{ij} + j) + V_{ij} \right\}
\]

\[
V_{Mij} = \min \left\{ W_{ij} + W_{i+1,j-1} \right\}
\]

---

RNA structure: other representations

- Base-pairing defines a secondary structure. The base-pairing is usually non-crossing.
- In this example the pseudo-knot makes for an exception.
Stochastic context-free grammar

1. A CFG
   - \( S \rightarrow gSu \)
   - \( S \rightarrow dSu \)
   - \( S \rightarrow dSg \)
   - \( S \rightarrow gSc \)
   - \( S \rightarrow aSa \)
   - \( S \rightarrow aSc \)
   - \( S \rightarrow gSa \)
   - \( S \rightarrow gSc \)
   - \( S \rightarrow SuS \)

2. A derivation of \( \text{"accuggacccuuagaggu"} \)

3. Corresponding structure

ssRNA transcription

- ssRNAs like tRNAs are usually encoded by short "non coding" genes, that transcribe independently.
- Found in both the UCSC “known genes” track, and as a subtrack of the RepeatMasker track

"non coding" RNAs (ncRNA) subtype:

- microRNAs (miRNA/miR)
MicroRNA (miR)

- miR match to target mRNA is quite loose.

⇒ a single miR can regulate the expression of hundreds of genes.

MicroRNA Transcription

MicroRNA Transcription

MicroRNA Transcription
MicroRNA (miR)

- mRNA
- ~70nt
- ~22nt
- miR match to target mRNA is quite loose.
- A single miR can regulate the expression of hundreds of genes.

Computational challenges:
- Predict miRs.
- Predict miR targets.

MicroRNA Therapeutics

Idea: bolster/inhibit miR production to broadly modulate protein production
Hope: “right” the good guys and/or “Wrong” the bad guys
Challenge: and not vice versa.

Other Non-Coding Transcripts
IncRNAs (long non coding RNAs)

Don’t seem to fold into clear structures (or only a sub-region does). Diverse roles only now starting to be understood. Example: bind splice site, cause intron retention.

→ Challenges: 1) detect and 2) predict function computationally

X chromosome inactivation in mammals

Dosage compensation

Xist – X inactive-specific transcript

lincRNAs & mRNAs often have structural motifs

SARS-CoV-2 Too...

Same bases both code for protein and participate in RNA structures

RNA genome!

System output measurements

We can measure non/coding gene expression! (just remember – it is context dependent)
1. First generation mRNA (cDNA) and EST sequencing:

In UCSC Browser:
2. Gene Expression Microarrays ("chips")

3. RNA-seq

"Next" (2nd) generation sequencing.

Fun computational challenges abound
Still Plenty “weird” transcripts out there

Common denominators still worked out..

Central Dogma had just miRNAs..
The million dollar question

Human Genome
Transcribed (Tx)
Tx from both strands
Leaky tx?
Functional?

In HW2 you’ll measure the sizes of some of these entities yourselves!

* True size of set unknown

Gene Finding II: technology dependence

Challenge:
“Find the genes, the whole genes, and nothing but the genes”

We started out trying to predict genes directly from the genome.
When you measure gene expression, the challenge changes:
Now you want to build gene models from your observations.
These are both technology dependent challenges.
The hybrid: what we measure is a tiny fraction of the space-time state space for cells in our body. We want to generalize from measured states and improve our predictions for the full compendium of states.

Circle of Life

To a first approximation this is all your genome cares about:
Making the right genes, at the right amount, at the right time, In the right cells.
One genome, ten trillion cells, hundred years of life.
Genes

- Gene production is conceptually simple
  - Contiguous stretches of DNA transcribe (1 to 1) into RNA
  - Some (coding or non-coding) RNAs are further spliced
  - Some (mi)RNAs are then translated into protein (4 to 20+1)
  - Other (nc)RNA stretches just go off to do their thing as RNA
- The devil is in the details, but by and large — this is it.

(none/coding) Gene finding - classical computational challenge:
1. Obtain experimental data (from limited cell types/conditions)
2. Find features in the data (e.g., genetic code, splice sites)
3. Generalize from features (e.g., predict genes yet unseen)
4. Predict contexts and functions (e.g., via family members)

Cell content is constantly refreshing

The cell is constantly making new proteins and ncRNAs. These perform their function for a while, and are then degraded. Newly made coding and non-coding gene products take their place. The picture within a cell is constantly "refreshing".

Cell differentiation

That is exactly what happens when cells differentiate during development from stem cells to their different final fates.
Genes usually work in groups

Biochemical pathways, signaling pathways, etc.
We’d like to catalog the functions of every gene.

Keyword lists are not enough

- Sheer number of terms too much to remember and sort
  - Need standardized, stable, carefully defined terms
  - Need to describe different levels of detail
  - So defined terms need to be related in a hierarchy

- With structured vocabularies/hierarchies
  - Parent/child relationships exist between terms
  - Increased depth = Increased resolvability
  - Can annotate data at appropriate level
  - May query at appropriate level

- Effectively a directed acyclic graph (DAG)
  - Node levels can be somewhat deceptive, depending on heterogeneity of curation efforts in different portions of the DAG

Anatomy Hierarchy

Anatomy keywords

Organ system
Cardiovascular system
Heart

Annotate genes to most specific terms

molecular function

- nucleic acid binding
- enzyme

DNA binding
- transcription factor
- DNA helicase
- DNA polymerase
- DNA methyltransferase
- DNA ligase
- RNA helicase

ATP-dependent helicase

- ATP-dependent helicase

DNA-dependent helicase

- DNA-dependent helicase

RNA helicase

- RNA helicase

- DNA-dependent helicase

- ATP-dependent helicase

- DNA-dependent helicase

- RNA helicase

- DNA-dependent helicase

- RNA helicase
1. Annotate at appropriate level, query at appropriate level
2. Queries for higher level terms include annotations to lower level terms

Genes usually work in groups
When a cell changes its behavior to take on a new activity, it stops producing some groups of genes and starts producing other groups of genes, to stop and start the relevant biological processes, respectively.

Let's compare which genes are active
Cluster all genes for differential expression

Determine cut-offs, examine individual genes

Ask about whole gene sets
Simplest way to ask: Hypergeometric

- Genes measured: N = 20,000
- Total genes in set 3: K = 11
- We picked the top n = 100 diff. expressed genes.
- Of them k = 8 belong to gene set 3.
- Under a null of randomly distributed genes, how surprising is it?

\[ P\text{-value} = \text{Pr}(k \geq 8 | N, K, n) \]

A low p-value, as here, suggests gene set 3 is highly enriched among the diff. expressed genes. Now see what (pathway/process) gene set 3 represents, and build a novel testable model around your observations.

GSEA (Gene Set Enrichment Analysis)

- Dataset distribution
- Gene set 3 distribution
- Gene set 1 distribution

Multiple Testing Correction

Note that statistically you cannot just run individual tests on 1,000 different gene sets. You have to apply further statistical corrections, to account for the fact that even in 1,000 random experiments a handful may come out good by chance.

(eg experiment = Throw a coin 10 times. Ask if it is biased. If you repeat it 1,000 times, you will eventually get an all heads series from a fair coin. Mustn’t deduce that the coin is biased)
RNA-seq

“Next” (2nd) generation sequencing.


64

What will you test?

Also note that this is a very general approach to test gene lists. Instead of a microarray experiment you can do RNA-seq. Advantage: RNA-seq measures all genes (up to your ability to correctly reconstruct them). Microarrays only measure the probes you can fit on them. (Some genes, or indeed entire pathways, may be missing from some microarray designs).


65

Single gene in situ hybridization

Recall a human is $10^{13}$ cells living for $10^9$ seconds. Which cells will you assay? When?


66