Statistical modeling of biopolymer sequences

Modeling biological sequences

- Kinds of questions we want to ask
  - Is this sequence a motif (e.g., binding site, splice site)?
  - Is this sequence part of the coding region of a gene?
  - Are these two sequences evolutionarily related?
  - ...

- What we will not address (covered last semester)
  - How two (or more) sequences can be optimally aligned
  - How sequencing results of a clone library can be assembled
  - What is the most parsimonious phylogeny of a set of sequences

- Machine learning: extracting useful information from a corpus of data $D$ by building good (predictive, evaluative or decision) models
Modeling biological sequences, ctd

• Computational analysis only generate hypothesis, which must be tested by experiments
  – Site-directed mutagenesis (to alter the sequence content)
  – Knockouts/insertions of genes/sites (deletion/addition of elements)
  – Functional perturbations (pathway inhibitors, drugs, ...)

• How to choose experimental models?
  – bacteria, yeast, C. Elegans, Drosophila, mouse, human(?) ...

• From one-way learning to close-loop learning:
  – Active learning: can a machine design smart experiments?

Probabilistic models for sequences

• We will use probabilistic models of sequences -- not the only approach, but usually the most powerful, because

  – sequences are the product of an evolutionary process which is itself stochastic in nature,
  – want to detect biological "signal" against "random noise" of background mutations,
  – data may be missing due to experimental reasons or intrinsically unobservable, and
  – we want to integrate multiple (heterogeneous) data and incorporate prior knowledge in a flexible and principled way,
  – ....
Hierarchical structure of the genome

Gene structure in prokaryotes
Gene structure in prokaryotes

• A protein-coding gene consists of the following, in 5’ to 3’ order
  – An upstream regulatory region, generally < 50 bp, which turns transcription on and off.
  – A transcription start site where RNA polymerase incorporates 1st nucleotide into nascent mRNA.
  – A 5’ untranslated region, generally < 30bp, that is transcribed into mRNA but not translated.
  – The coding region of the gene (typically=1000bp), consisting of a sequence of codons.
  – The translation stop site marking the end of coding region. Consists of a stop codon, which causes the release of the polypeptide at conclusion of translation.
  – A 3’ untranslated region, transcribed into RNA but not translated.
  – The transcription stop site marking where the RNA polymerase concludes transcription.

The bacterial genome

The E. coli chromosome
Gene structure in eukaryotes

A typical gene consist of the following, in 5' to 3' order:

- An upstream regulatory region, often larger and more complex than in prokaryotes, parts of which may be several thousand bases or more upstream of transcription start site.
- A transcription start site.
- A 5' untranslated region, often larger than in prokaryotes, and which may include sequences playing a role in translation regulation.
- The coding sequence, which unlike the case with prokaryotes, may be interrupted by non-coding regions called introns. These are spliced out of the transcript to form the mature mRNA (and sometimes the splicing can occur in more than one way).
- The translation stop site.
- A 3' untranslated region, which may contain sequences involved in translational regulation.
- A polyadenylation (polyA) signal, which indicates to the cell's RNA processing machinery that the RNA transcript is to be cleaved and a poly-adenine sequence (AAAAA...) tail appended to it.
- The transcription stop site.
Alternative splicing

The human genome
Basic Probability Theory Concepts

- A **sample space** \( S \) is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (\( S \) can be finite or infinite.)
  - E.g., \( S \) may be the set of all possible nucleotides of a DNA site

- A **random variable** is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)
  - E.g., seeing an "A" at a site \( \Rightarrow X=1 \), o/w \( X=0 \).
  - This describes the true or false outcome a random event.
  - Can we describe richer outcomes in the same way? (i.e., \( X=1, 2, 3, 4 \), for being A, C, G, T) --- think about what would happen if we take expectation of \( X \).

- Random vector
  - \( X=[X_A, X_T, X_G, X_C]^T \), \( X=[0,0,1,0]^T \Rightarrow \) seeing a "G" at site /

Basic Prob. Theory Concepts, ctd

- **(In the discrete case)**, a probability distribution \( P \) on \( S \) (and hence on the domain of \( X \)) is an assignment of a non-negative real number \( P(s) \) to each \( s \in S \) (or each valid value of \( x \)) such that \( \Sigma_{s \in S} P(s)=1 \). (0\( \leq P(s) \leq 1 \))
  - intuitively, \( P(s) \) corresponds to the frequency (or the likelihood) of getting \( s \) in the experiments, if repeated many times
  - call \( \theta_s = P(s) \) the **parameters** in a discrete probability distribution

- A probability distribution for a sample space is sometimes called a probability model, in particular if several different distributions are under consideration
  - write models as \( M_1, M_2 \), probabilities as \( P(X|M_1), P(X|M_2) \).
  - E.g., \( M_1 \) may be prob. dist. appropriate if \( X \) is from splice site, \( M_2 \) is for the "background".
  - \( M \) is usually a two-tuple of \{dist. family, dist. parameters\}
Basic Prob. Theory Concepts, ctd

• For events $E$ (i.e., $X=x$) and $H$ (say, $Y=y$), the conditional probability of $E$ given $H$, written as $P(E|H)$, is

$$P(E \text{ and } H)/P(H)$$

(= the probability of both $E$ and $H$ are true, given $H$ is true)

• $E$ and $H$ are (statistically) independent if

$$P(E) = P(E|H)$$

(i.e., prob. $E$ is true doesn’t depend on whether $H$ is true); or equivalently

$$P(E \text{ and } H) = P(E)P(H).$$

• $E$ and $F$ are conditionally independent given $H$ if

$$P(E,F|H) = P(E|H)$$

or equivalently

$$P(E,F|H) = P(E|H)P(F|H).$$

Basic Prob. Theory Concepts, ctd

• Joint probability dist. on multiple variables:

$$P(X_1,X_2,X_3,X_4,X_5,X_6) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)P(X_5|X_4)P(X_6|X_5)$$

• If $X_i$’s are independent: ($P(X_i) = P(X_i)$)

$$P(X_1,X_2,X_3,X_4,X_5,X_6) = P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)P(X_6) = P(X_i)$$

• If $X_i$’s are conditionally independent, the joint can be factored to simpler products, e.g.,

$$P(X_1,X_2,X_3,X_4,X_5,X_6) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)P(X_5|X_4)P(X_6|X_5)$$

• The **Graphical Model** representation
Basic Prob. Theory Concepts, ctd

- The Bayesian Theory: (e.g., for data $D$ and model $M$)

\[
P(M|D) = \frac{P(D|M)P(M)}{P(D)}
\]

- the posterior equals to the likelihood times the prior, up to a constant.

- This allows us to capture uncertainty about the model in a principled way

Probabilities on sequences

- Let $S$ be the space of DNA or protein sequences of a given length $n$. Some simple assumptions for assigning probabilities to sequences:

  - **Equal frequency assumption**: All residues are equally probable at any position; i.e., $P(X_i = r) = P(X_i = q)$ for any two residues $r$ and $q$, for all $i$.
    - this implies that $P(X_i = r) = 1/|A|$, where $A$ is the residue alphabet ($1/20$ for proteins, $1/4$ for DNA)

  - **Independence assumption**: whether or not a residue occurs at a position is independent of what residues are present at other positions.
    - probability of a sequence
      
      \[
P(X_1, X_2, ..., X_n) = \theta_r \cdot \theta_q \cdot ... \cdot \theta_s = \theta^N
      \]
Failure of Equal Frequency Assumption for (real) DNA

- For most organisms, the nucleotides composition is significantly different from 0.25 for each nucleotide, e.g.,
  - *H. influenza* 0.31 A, 0.19 C, 0.19 G, 0.31 T
  - *P. aeruginosa* 0.17 A, 0.33 C, 0.33 G, 0.17 T
  - *M. janaschii* 0.34 A, 0.16 C, 0.16 G, 0.34 T
  - *S. cerevisiae* 0.31 A, 0.19 C, 0.19 G, 0.31 T
  - *C. elegans* 0.32 A, 0.18 C, 0.18 G, 0.32 T
  - *H. sapiens* 0.30 A, 0.20 C, 0.20 G, 0.30 T

- Note symmetry: $A \cong T$, $C \cong G$, even though we are counting nucleotides on just one strand. Explanation:

General Hypothesis Regarding Unequal Frequency

- Neutralist hypothesis: mutation bias (e.g., due to nucleotide pool composition)

- Selectionist hypothesis: selection
The multinomial model for sequence

- For a site $i$, define its residue identity to be a random vector:
  \[
  X_i = \begin{pmatrix} X_{i,A} \\ X_{i,C} \\ X_{i,G} \\ X_{i,T} \end{pmatrix}, \quad \text{where } X_{ij} \in [0,1], \quad \text{and } \sum_{j \in \{A,C,G,T\}} X_{ij} = 1
  \]
- $X_{ij} = 1$ w.p. $\theta_j$, $\sum_{k \in \{A,C,G,T\}} \theta_k = 1$.
- The probability of an observation $s = C$ (i.e., $x_{i,C} = 1$) at site $i$:
  \[
P(X_{i,C}) = P(X_{i,j} = 1, \text{ where } j \text{ index then observed at } i))
  = \theta_j = \theta_A^{\text{val}} \times \theta_C^{\text{val}} \times \theta_G^{\text{val}} \times \theta_T^{\text{val}} = \theta_j^{\text{val}}.
  \]
- The probability of a sequence $(x_1, x_2, ..., x_N)$:
  \[
P(x_1, x_2, ..., x_N) = \prod_{i=1}^N P(x_i) = \prod_{i=1}^N \theta_j^{\text{val}} = \prod_{i=1}^N \theta_j^{\text{val}}
  \]

Parameter estimation

- Maximum likelihood estimation: $\theta = \arg \max_\theta P(D | \theta)$
  - multinomial parameters:
    \[
    \{\theta_1, \theta_2, ..., \theta_k\} = \arg \max_\theta \theta_k^{\text{val}}, \quad \text{s.t. } \theta_j = 1
    \]
    It can be shown that: $\theta_k^{\text{ML}} = \frac{n_k}{N}$
- Bayesian estimation:
  - Dirichlet distribution: $P(\theta) = \frac{\Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \prod_k \theta_k^{\alpha_k-1} = C(\alpha) \prod_k \theta_k^{\alpha_k-1}$
    - Posterior distribution of $\theta$ under the Dirichlet prior:
      \[
P(\theta | x_1, ..., x_N) \propto \theta_k^{\alpha_k-1} \theta_j^{\text{val}} = \theta_k^{\alpha_k-1+\alpha_k}
      \]
    - Posterior mean estimation:
      \[
      \theta_k = \frac{\alpha_k + n_k}{N + \sum_k \alpha_k}
      \]
Models for homogeneous sequence entities

• Probabilities models for long "homogeneous" sequence entities, such as:
  – exons (ORFs)
  – introns
  – inter-genetic background
  – protein coiled-coil (other other structural) regions

• Assumptions:
  – no consensus, no recurring string patterns
  – have distinct but uniform residue-composition
  – every site in the entity are iid samples from the same model

• The model:
  – a single multinomial: $X \sim \text{Mul}(\theta)$

Models for homogeneous sequence entities, ctd

• Limitations
  – non-uniform residue composition (e.g., CG rich regions)
  – non-coding structural regions (MAR, centromere, telomere)
  – di- or tri- nucleotide couplings
  – estimation bias
  – evolutionary constrains
Site models

• Probabilities models for short sequences, such as:
  – splice sites
  – translation start sites
  – promoter elements
  – protein "motifs"

• Assumptions:
  – different examples of sites can be aligned without indels (insertions/deletions) such that tend to have similar residues in same positions
  – drop equal frequency assumption; instead have position-specific frequencies
  – retain independence assumption (for now)

Site models ctd.

• Applies to short segments (<30 residues) where precise residue spacing is structurally or functionally important, and certain positions are highly conserved

  – DNA/RNA sequence binding sites for a single protein or RNA molecule
  – Protein internal regions structurally constrained due to folding requirements; or surface regions functionally constrained because bind certain ligands
Nucleotide Counts for 8192 C. elegans 3' Splice sites

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3276</td>
</tr>
<tr>
<td>C</td>
<td>570</td>
</tr>
<tr>
<td>G</td>
<td>598</td>
</tr>
<tr>
<td>T</td>
<td>5553</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.000</td>
</tr>
<tr>
<td>C</td>
<td>0.000</td>
</tr>
<tr>
<td>G</td>
<td>0.000</td>
</tr>
<tr>
<td>T</td>
<td>0.000</td>
</tr>
</tbody>
</table>

3' Splice site - C. elegans

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3276</td>
</tr>
<tr>
<td>C</td>
<td>570</td>
</tr>
<tr>
<td>G</td>
<td>598</td>
</tr>
<tr>
<td>T</td>
<td>5553</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.000</td>
</tr>
<tr>
<td>C</td>
<td>0.000</td>
</tr>
<tr>
<td>G</td>
<td>0.000</td>
</tr>
<tr>
<td>T</td>
<td>0.000</td>
</tr>
</tbody>
</table>
5' Splice sites - C. elegans

Limitation of Site Models

- Failure to allow indels means variably spaced subelements are "smeared", e.g.:
  - branch site, for 3' splice sites;
  - coding sequences, for both 3' and 5' sites

- Independence assumption
  - usually OK for protein sequences (after correcting for evolutionary relatedness)
  - often fails for nucleotide sequences; examples:
    - 5' sites (Burge-Karlin observation);
    - background (dinucleotide correlation).
Why correlation?

- Splicing involves pairing of a small RNA with the transcription at the 5' splice site.
- The RNA is complementary to the 5' sr consensus sequence.
- A mismatch at position -1 tends to destabilize the pairing, and makes it more important for other positions to be correctly paired.
- Analogy can be easily drawn for other DNA and protein motifs.

Comparing alternative probability models

- We will want to consider more than one model at a time, in the following situations:
  - To differentiate between two or more hypothesis about a sequence
  - To generate increasingly refined probability models that are progressively more accurate
Comparing alternative probability models, ctd.

• First situation arises in testing biological assertion, e.g., “is this a coding sequence?” Would compare two models:

1. one associated with a hypothesis \(H_{\text{coding}}\) which attaches to a sequence the probability of observing it under experiment of drawing a random sequence from the genome

2. one associate with a hypothesis \(H_{\text{noncoding}}\) which attaches to a sequence the probability of observing it under experiment of drawing a random non-coding sequence from the genome.

Likelihood Ratio Test

• The posterior probability of a model given data is:

\[ P(M|D) = \frac{P(D|M)P(M)}{P(D)} \]

• Given that all models are equally probable \textit{a priori}, the posterior probability ratio of two models given the same data reduce to a \textit{likelihood ratio}:

\[ LR(M_a, M_0 | D) = \frac{P(D | M_a)}{P(D | M_0)} \]

– the numerator and the denominator may both be very small!

• The log likelihood ratio (LLR) is the logarithm of the likelihood ratio:

\[ LLR(M_a, M_0 | D) = \log P(D | M_a) - \log P(D | M_0) \]