Lecture 3:
Probability Models for Sequences

- Probability models
  - – Equal frequency & independence assumptions
- ‘Background’ models
  - – Failure of equal frequency assumption
    - Neutralist vs selectionist interpretations
  - – Failure of independence assumption
    - Markov models
- Assessing significance of sequence patterns
  - – Simulations

Probability Models of Sequences

- Sample questions when interpreting genomes:
  - Is this sequence a splice site?
  - Is this sequence part of the coding region of a gene?
  - Are these two sequences evolutionarily related?
  - Does this sequence show evidence of selection?
- Computational analysis can’t answer:
  - only generates *hypotheses*

    which must ultimately be tested by experiment.
- *But* hypotheses should
  - have some reasonable chance of being correct, and
  - carry indication of reliability.

- We use *probability models* of sequences to address such questions.
- Not the only approach, but usually the most powerful, because
  - seqs are products of evolutionary process which is *itself* probabilistic
  - want to detect biological “signal” against “noise” of background sequence or mutations

Models: simplicity vs complexity

- “*All models are wrong; some models are useful.*” – George Box
- “*What is simple is always wrong. What is not is unusable.*” – Paul Valery
- “*Everything should be made as simple as possible, but not simpler*.” – Albert Einstein (?)
- Some disadvantages of complexity:
  - computational challenge
  - (lack of) interpretability
  - overfitting

Basic Probability Theory Concepts

- A *sample space* S is set of all possible outcomes of a conceptual, repeatable experiment.
  - $|S| < \infty$ in most of our examples.
  - e.g. S = all possible sequences of a given length.
- Elements of S are called *sample points*.
  - e.g. a particular seq = outcome of "experiment" of extracting seq of specified type from a genome.
- A *probability distribution* P on S assigns non-neg real number $P(s)$ to each $s \in S$, such that

$$\Sigma_{s \in S} P(s) = 1$$

(So $0 \leq P(s) \leq 1 \ \ \forall s$)

  - Intuitively, $P(s)$ = fraction of times one would get s as result of the expt, if repeated many times.

- A *probability space* (S,P) is a sample space S with a prob dist'n P on S.
- Prob dist'n on S is sometimes called a *probability model* for S, particularly if several dist'ns are being considered.
  - Write models as M_1, M_2 , probabilities as $P(s \mid M_1)$, $P(s \mid M_2)$.
  - e.g.
    - M_1 = prob dist'n for splice site seqs,
    - M_2 = prob dist'n for "background" (arbitrary genomic) seqs.

- An *event* E is a criterion that is true or false for each $s \in S$.
  - defines a subset of S (sometimes also denoted E).
  - $P(E)$ is defined to be $\Sigma_{s|E \text{ is true}} P(s)$.
- Events $E_1, E_2, \ldots, E_n$ are *mutually exclusive* if no two of them are true for the same point;
  - then $P(E_1 \text{ or } E_2 \text{ or } \ldots \text{ or } E_n) = \Sigma_{1 \le i \le n} P(E_i)$.
- If $E_1, E_2, \ldots, E_n$ are also *exhaustive*, i.e. every s in S satisfies E_i for some i, then $\Sigma_{1 \le i \le n} P(E_i) = 1$.

- For events E and H, the *conditional probability* of E given H, is

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

(= prob that both E and H are true, given H is true)

  - – undefined if $P(H) = 0$.

- E and H are (*statistically*) *independent* if

$$P(E) = P(E \mid H)$$

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

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

Probabilities on Sequences

- Let S = space of DNA or protein sequences of length n. Possible assumptions for assigning probabilities to S:
  - *Equal frequency assumption*: All residues are equally probable at any position;
    - $P(E_r^{(i)}) = P(E_q^{(i)})$ for any two residues r and q,
      - where $E_r^{(i)}$ means residue r occurs at position i, then
    - Since for fixed i the $E_r^{(i)}$ are mutually exclusive and exhaustive,

$$P(E_r^{(i)}) = 1 / |A|$$

where A = residue alphabet

$$P(E_r^{(i)}) = 1/20 \text{ for proteins, } 1/4 \text{ for DNA).}$$

  - *Independence assumption*: whether or not a residue occurs at a given position is independent of residues at other positions.

- Given above assumptions, the probability of the sequence

  s = ACGCG

  (in the space *S* of all length 5 sequences) is calculated by considering 5 events:

  - Event 1 is that first nuc is A. Probability = .25.
  - Event 2 is that 2d nuc is C. Probability = .25.
  - Event 3 is that 3d nuc is G. Probability = .25.
  - Event 4 is that 4th nuc is C. Probability = .25.
  - Event 5 is that 5th nuc is G. Probability = .25.

  By independence assumption, prob of all 5 events occurring is the product $(.25)^5 = 1/1024$.

  Since *s* is the only sequence satisfying all 5 conditions, $P(s) = 1/1024$.

- More generally, under equal freq and indep assumptions,

  prob of nuc sequence of length n = $.25^n$,

  prob of protein sequence of length n = $.05^n$

  in the space S of length n sequences.

‘Background’ models

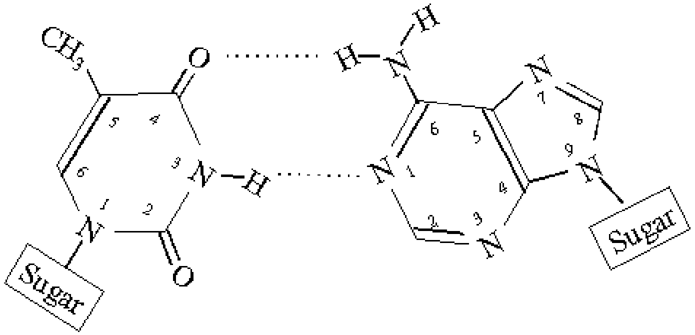
- ‘Average’ model for genome; contrasted with ‘foreground’ models (for sites & other regions of interest)

- *Whole genome* vs *non-site*

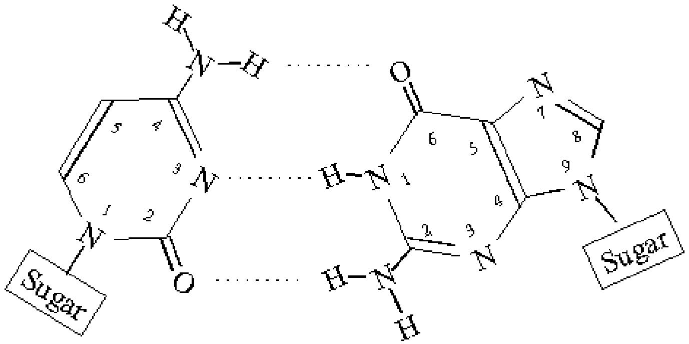
Genome background models: Failure of equal frequency assumption

- For most organisms, the genomic nucleotide composition is significantly different from .25 for each nucleotide, e.g.:
  - *H. influenza* .31 A, .19 C, .19 G, .31 T
  - *P. aeruginosa* .17 A, .33 C, .33 G, .17 T
  - *M. janaschii* .34 A, .16 C, .16 G, .34 T
  - *S. cerevisiae* .31 A, .19 C, .19 G, .31 T
  - *C. elegans* .32 A, .18 C, .18 G, .32 T
  - *H. sapiens* .29 A, .21 C, .21 G, .29 T

- Note approximate symmetry: A $\cong$ T, C $\cong$ G,
  - even though we're counting nucs on just one strand.
  - Expect *exact* equality when counting both strands
- Explanation:
  - Although individual biological features may have non-symmetric composition (local *asymmetry*),
  - usually features are distributed approx *randomly* w.r.t. strand,
  - so local asymmetries *cancel*, yielding overall symmetry.



Thymine (T) Adenine (A)



Cytosine (C) Guanine (G)

General Hypotheses Regarding Unequal Frequency

- Neutralist hypothesis: *mutation bias*
  - e.g. due to nucleotide pool composition
- Selectionist hypothesis: *selection*
  - selection on (many) particular nucleotides
  - selection on mutational bias mechanisms
  - …

Genome background models: Failure of independence assumption

Nucleotide Freqs (*C. elegans* chr. 1):
A 4575132 (.321) ; C 2559048 (.179) ; G 2555862 (.179); T 4582688 (.321)

dinucleotide frequencies (5′ nuc to left, 3′ nuc at top – e.g. obs freq of *ApC* is .047): (Note "symmetry"!)

Observed

	A	C	G	T
A	0.135	0.047	0.051	0.088
C	0.061	0.035	0.033	0.051
G	0.063	0.034	0.034	0.047
T	0.061	0.064	0.061	0.135

Expected (*under independence*)

	A	C	G	T
A	0.103	0.057	0.057	0.103
C	0.057	0.032	0.032	0.058
G	0.057	0.032	0.032	0.057
T	0.103	0.058	0.057	0.103

Observed / Expected

	A	C	G	T
A	1.314	0.818	0.885	0.853
C	1.055	1.075	1.031	0.886
G	1.106	1.062	1.074	0.818
T	0.597	1.105	1.056	1.313

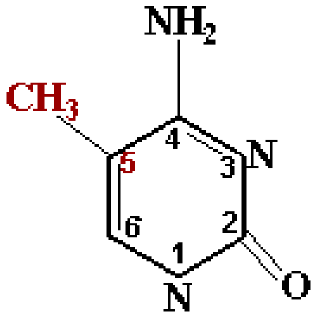
Dinucleotide frequencies

- Underrepresentation of *T*p*A*: found in nearly all genomes;
  - reason unknown:
    - neutral (mutation patterns)?
    - selection?
- Overrepresentation of *A*p*A*, *T*p*T*, *C*p*C*, *G*p*G* – also frequently observed in other organisms.
- Unlike mammalian genomes, no underrepresentation of *C*p*G* in *C. elegans*
  - *C*p*G* not methylated in *C. elegans* (or most other non-vertebrates).

Dinucleotide Freqs – *H. sapiens* Chr.21

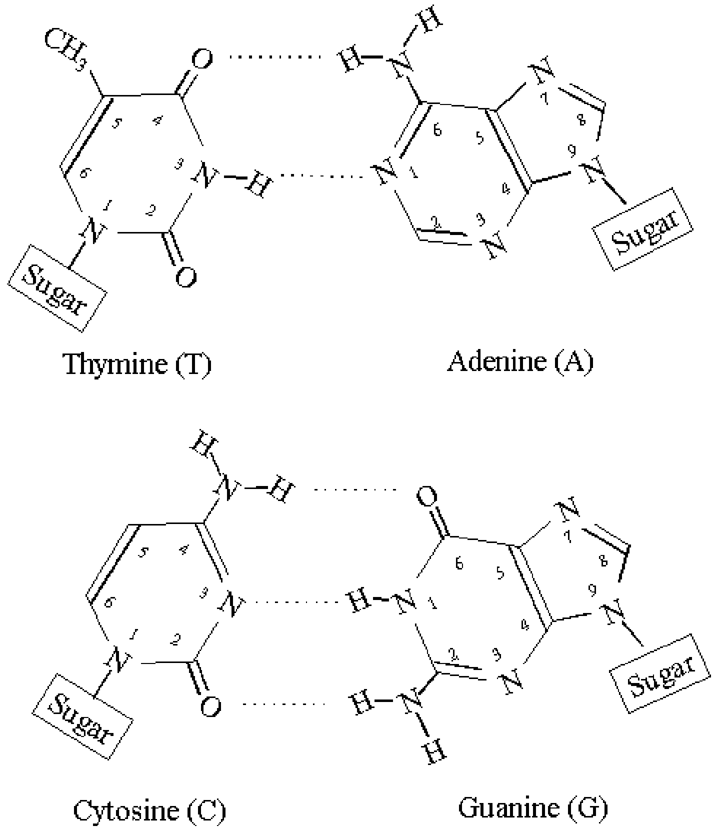
```
 Nucleotide Freqs:
    A 10032226  0.297; T  9962530  0.295
    G  6908202  0.204; C  6921020  0.205
  Entropy: 1.976 bits

    Observed Dinuc Freqs          Expected (under independence)
     A     C     G     T               A     C     G     T
A   0.099 0.051 0.069 0.078          0.088 0.061 0.061 0.087
C   0.073 0.052 0.011 0.069          0.061 0.042 0.042 0.060
G   0.059 0.043 0.052 0.050          0.061 0.042 0.042 0.060
T   0.066 0.059 0.072 0.098          0.087 0.060 0.060 0.087

                     Observed / Expected
                     A     C     G     T
                A   1.124 0.839 1.139 0.891
                C   1.204 1.243 0.260 1.139
                G   0.974 1.025 1.245 0.839
                T   0.752 0.976 1.204 1.125
```

5-methylcytosine (^{m}C): the ‘5th base’



- Comprises ~1-6% of mammalian & plant genomes

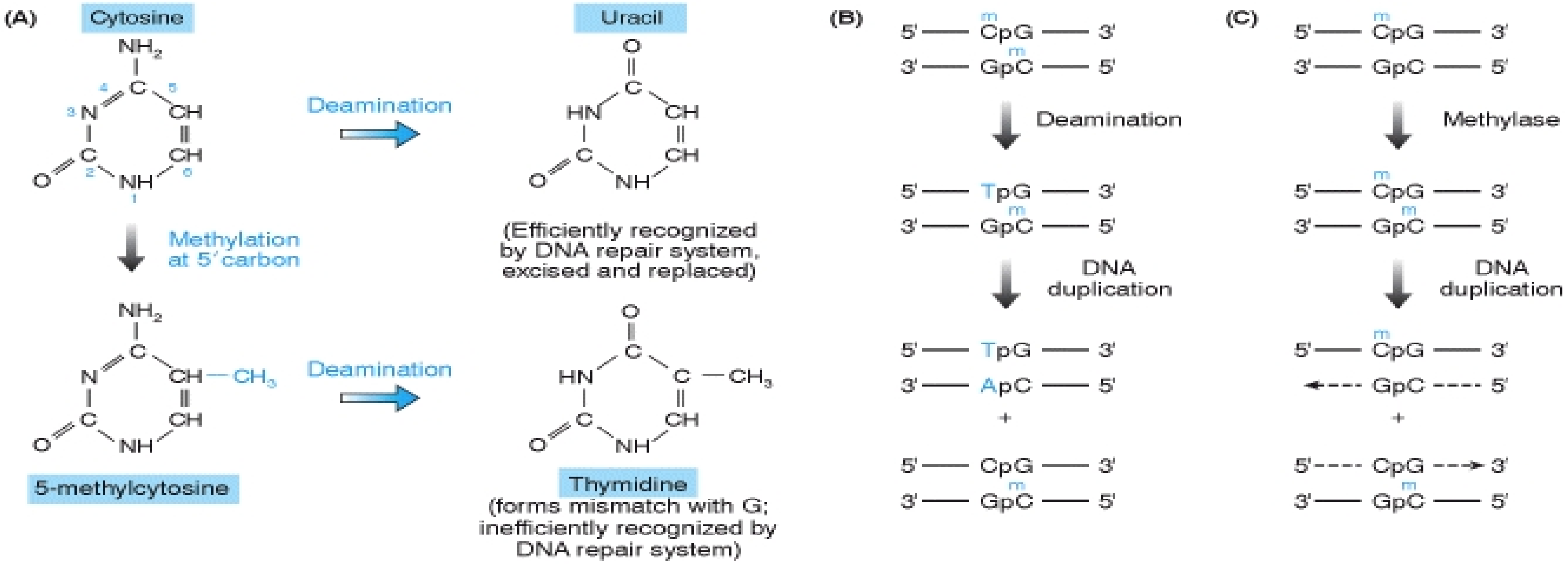
- Methylation does *not* affect base-pairing:



- But it *does* affect
  - protein binding, e.g. Sp1, EGR1, CTCF

    ⇒ effects on gene expression, development, cellular differentiation, transposon suppression, embryogenesis, imprinting, X-inactivation, chromatin structure, tumorigenesis

    mouse methyltransferase knockouts are embryonic lethal
  - mutation rate: ^{m}C is a mutation ‘hotspot’:



http://www.ncbi.nlm.nih.gov

- In mammals methylated C’s (nearly) always occur as part of a CpG dinucleotide:

5′ ^{m}C G 3′
3′ G ^{m}C 5′

- But some Cs *not* in CpGs are methylated, in some cell types

- as many as 20-30% of all new single-base mutations in mammalian genomes may be at CpGs, judging from
  - analysis of disease-causing mutations,
  - comparison of closely related species
  - polymorphism data
- As a result, CpGs are substantially underrepresented in mammalian DNA:
  - expected frequency .21 .21 = .044 (in mammalian genomes, G+C freq is about .42, A+T about .58)
  - only see about 1/5 that many.
- Conversely, TpGs and CpAs are overrepresented

Dinucleotide Freqs – *H. sapiens* Chr.22

```
Nucleotide Freqs:
   A  8745910  0.261; T  8720493  0.261
   G  7999585  0.239; C  7997931  0.239
 Entropy: 1.999 bits

   Observed Dinuc Freqs         Expected (under independence)
     A     C     G     T             A     C     G     T
A  0.077 0.051 0.075 0.058         0.068 0.062 0.062 0.068
C  0.077 0.071 0.016 0.075         0.062 0.057 0.057 0.062
G  0.061 0.057 0.071 0.051         0.062 0.057 0.057 0.062
T  0.047 0.061 0.077 0.076         0.068 0.062 0.062 0.068

                 Observed / Expected
                   A     C     G     T
                A  1.125 0.817 1.205 0.855
                C  1.233 1.236 0.285 1.206
                G  0.975 0.989 1.237 0.818
                T  0.684 0.977 1.233 1.124
```

Genome background models: Failure of independence assumption

Nucleotide Freqs (*C. elegans* chr. 1):
A 4575132 (.321) ; C 2559048 (.179) ; G 2555862 (.179); T 4582688 (.321)

dinucleotide frequencies (5′ nuc to left, 3′ nuc at top – e.g. obs freq of *ApC* is .047): (Note "symmetry"!)

	Observed				**Expected (*under independence*)**			
	A	**C**	**G**	**T**	**A**	**C**	**G**	**T**
A	0.135	0.047	0.051	0.088	0.103	0.057	0.057	0.103
C	0.061	0.035	0.033	0.051	0.057	0.032	0.032	0.058
G	0.063	0.034	0.034	0.047	0.057	0.032	0.032	0.057
T	0.061	0.064	0.061	0.135	0.103	0.058	0.057	0.103

Observed / Expected

	A	**C**	**G**	**T**
A	1.314	0.818	0.885	0.853
C	1.055	1.075	1.031	0.886
G	1.106	1.062	1.074	0.818
T	0.597	1.105	1.056	1.313

Conditional probability (in *C. elegans*) of a given nucleotide (top) occurring, given the preceding nucleotide (left)

	A	**C**	**G**	**T**
A	**0.421**	**0.147**	**0.159**	**0.274**
C	**0.338**	**0.193**	**0.185**	**0.284**
G	**0.355**	**0.190**	**0.192**	**0.263**
T	**0.191**	**0.198**	**0.189**	**0.421**

Markov models

- Such conditional probabilities can be used to define a ***first-order Markov model*** (or ***Markov chain model***) for background sequence probabilities:

$$P(s_1\, s_2\, s_3 \cdots s_n)$$

$$\equiv P(s_1)\, P(s_2 \mid s_1)\, P(s_3 \mid s_2) \cdots P(s_n \mid s_{n\text{ -}1})$$

- Similarly, one can define an a ***order-k Markov model*** in which the probability of s_i is conditional on $s_{i\text{-}k} \cdots s_{i\text{-}2}\, s_{i\text{-}1}$

(i.e. the k preceding residues)

- Note that the required number of parameters is exponential in k
- ***independence model = order-0 Markov model***

Assessing significance of sequence patterns

- Problem: Is a particular sequence pattern, e.g.
  - a match between genomes, or
  - a region of a particular composition (e.g. GC-rich)

likely to be "biologically significant", e.g. indicating

  - an evolutionary relationship, or
  - a functional feature

Assessing significance of sequence patterns

- Idea:
  - specify a scoring system for patterns of the given type
  - find the score *distribution* in *negative controls*
    - i.e. sequences not expected to contain the biological feature
  - Scores occurring in real sequence, but not in negative controls, *may* have biological significance
- Caveats:
  - Control may be inadequate in quantity / quality
  - ‘Biologically significant’ ≠ interpretable
    - can’t infer function!!

‘Negative control’ sequences

1. real biological ‘background’ sequences known not to have the feature in question
   - ideal if available – but usually hard to find!

2. simulated sequences
   - requires probability model retaining *some* features of real sequences
   - Quantity: In general, want multiple such sequences
   - Quality: is the model complex enough?

Theoretical score distributions

- For simple probability models, can sometimes avoid simulations by finding a ***theoretical*** probability distribution
  - approximate, e.g. Karlin-Altschul for BLAST hits
  - or exact

  for the scores.
- Alternatively, can fit a theoretical distribution to the observed scores for simulated data
  - Avoids need for large number of simulations

Homework 2

- Purpose: Assess significance of HW 1 genomic matches
- Simulate negative controls using two different background sequence models:
  - Order 0 Markov
  - Order 1 Markov
- Then find matches (using HW 1 suffix array method) between real sequence and these control sequences
  - Ideally should do lots of simulations!!