CpG Islands and Hidden Markov Models

CSE 527

Notes, Lecture12, 11/7/05

Tyler Powell powellt@u

Independence – Assumption that next base in DNA sequence is independent of previous bases

- Assumed in most models
- Not realistic
    - Physical, electrostatic fit
    - Bad fits in some location must be compensated for.

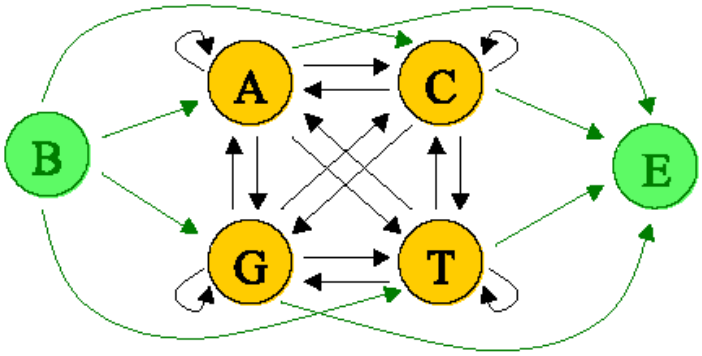
CpG Islands – "XpY" is a common notation for a pair of consecutive nucleotides on one DNA strand; the "p" is mnemonic for the DNA phosphorate backbone joining them, to distinguish it from, e.g., an X-Y Watson-Crick base pair. "A CpG island" is a region of the genome in which CpG dinucleotides occur higher frequency than would be expected based on the overall frequency of C's and G' in the vicinity. There is a story behind how they arise and why they're interesting:

- The C of a CpG dinucleotide is often methylated (in Eukaryotes)
    - Methylation prevents protein binding (shuts off gene)
    - Can still be copied (copy is unmethylated)
    - Passes down information on "type" of cell
- CpG is less common than randomly expected
    - This is due to the fact that methyl-C mutates to T easily
        - Therefore, in methylated regions, an unmutated C is uncommon
    - EXCEPT in promoter regions
        - Promoter regions often not methylated, so mutation not as prevalent
        - Therefore promoter regions contain more CpG occurrences, creating CpG Islands
- In these CpG Islands, weight matrix model is less accurate because it assumes independence. A better model is the Markov chain.

Problem:
CpG islands have more CpG's than elsewhere. Can we find these islands?

Markov Chains – A seqeunce of random variables is a k-th order Markov chain if for all I: $P(x_i|x_1,x_{2,}\ldots,x_{i-1}) = P(x_i|x_{i-k},x_{i-k+1},\ldots x_{i-1})$. That is, the probability of x_i is only dependent on the previous k values of x.

First order (k=1) – Probability of next value only dependent on current value:



This model shows all possible transitions from one base to another in a DNA sequence. You can imagine that each arrow in the figure has a different probability associated with it are the probability the base will transition along the arrow. The beginning and end states can be used to simulate more likely start/end bases and to control length. So, in this case:

States:	A,C,G,T
Emissions:	Corresponding letter
Transitions:	$a_{st} = P(x_i=t \mid x_{i-1}=s)$

Probability of a given sequence:

$$p(x) = p(x_1)\prod_{i=1}^{n-1} a_{x_i,x_{i+1}}$$ (Multiplication of probability of next state from current)

Values of a are trained from data sets. So values of a must be found for bases in islands and not in islands. After a values have been found, the model must score the data set. Log likelihood ratio is convenient for this:

$$S(x) = \log\frac{P(x \mid +)}{P(x \mid -)} = \sum_{i=1}^{L} \log\frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

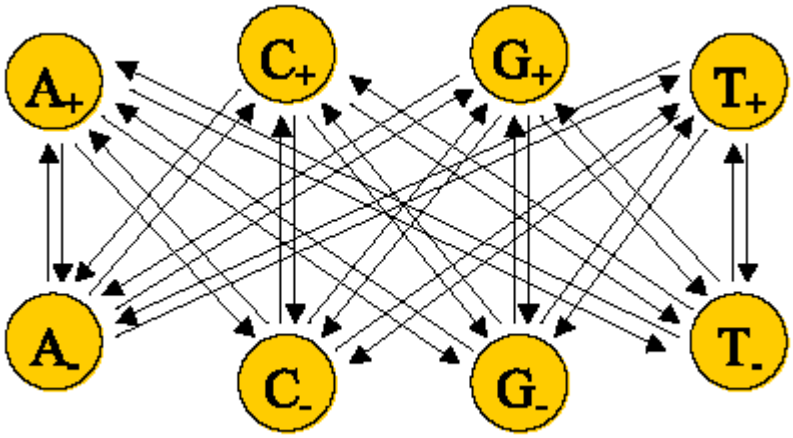
Where the beta values can be found from the a values.

Log likelihood uses ratios of sequence probabilities from CpG islands and background. Therefore, if it's positive, it's more likely to be a CpG island. This model works well on training set.

Problem update:

- We now can determine if a given sequence is likely a CpG island
- However, we still can't find the islands imbedded in a long
    - Approach 1:
        - Score 100 bp windows
            - Simple
            - Inflexible (might not find shorter sequences)
    - Approach 2:
        - Combined model
            - Same as before, but add transitions from island to background and back

- Probabilities found from data with transition probabilities based on average length of sequences.
- Current letter does not tell you if you're in island state or background state (*Hidden* Markov model)



Visual representation of hidden Markov model. Now the transitions between background and CpG islands have been added.

Observed data: emission sequence
Hidden data: state/transition sequence

The hidden data is what we are interested in (the transitions from CpG islands to background).

How to uncover this hidden data and how to use this model will be covered in subsequent lectures.