CSE 527 **Justin Shaffer**
11.1.06

Substring parsimony problem
- Goal is to determine sequence similarity across species using phylogenic trees
- Want to find all possible sets of k-mers so that parsimony score is at most a threshold value

Small Example
- Find motif of length four between five species
- One mutation needed

Plant example (rcbs)
- Sequences are aligned at start codon, but motifs are not aligned
- Proteins regulate when these genes are expressed or not
- 50% of protein mass of green leaves – key player in photosynthesis

An Exact Algorithm
- generalized from Sankoff and Rousseau
- need to keep track of adjacent bases to run algorithm
- at each leaf, build a table that says what are the scores for the motifs
- need a table of size 4^k – 4 bases to pick from
- ACGT can become ACGG through one mutation – total cost of two for two leaves

Recurrence
- $W_u[s]$ → score of W of string s at node u
- Running time → $O(k * 4^{2k})$ if two children per node
- If sequence is length 5, running time is $4^{2*5} = 2^{20}$ = order of 1 million; probably can be done on desktop computer

Improvements
- Are only looking for small parsimony scores
    - Can limit the number of trees that you need to look it → cuts running time
    - Allows algorithm to be practical
    - Webserver apps exist for this algorithm → few minutes to find result

B-actin example
- Motif of length 13 matched in all 10 organisms
- Can it happen by chance? → $4^{13} = 2^{26}$ chance that it occurs randomly
    - Most likely this is a conserved sequence that occurs for a reason

Motifs absent from some species
- Motifs can be lost from say fish compared to mammals over time
- Parsimony score would then be bad when comparing fish to mammals just because they are not there
- 2 Objective functions → smallest parsimony score and largest part of tree

    - how to accommodate both?
        - Ask for tree sizes based on number of mutations needed
    - As get farther away in phylogeny, larger parsimony score

Conclusions
- Well motivated problem definition
- Exact algorithm for solving it
- Linear in # species, exponential in parsimony score
- Able to discover highly conserved regions, both known and not yet known
    - Experiments can determine TF binding sites as well

New topic
Markov Models and Hidden Markov Models

DNA Methylation
- Recognizing CpG islands – adjacent C and G on same strand
- C of CpG can be methylated in eukaryotes (~70-80% in mammals)
- Plants seem to have evolved CpG methylation independently
- One function of methylation → silences transcription; could limit TF binding
    - Also could use to delineate stem cell derivatives (stem cell → kidney or liver?)
- If both strands methylated, upon cell division, daughter strands not methylated, but parent strands are, so DNA is said to be hemi-methylated
- DNA methyltransferases can then methylate the daughter strands, thus forming fully-methylated DNA and turning off that gene
- Is a key obstacle to cloning → need to de-methylate DNA to differentiate cells into other kinds of cells (allow for selective gene expression)
- X-inactivation → for females (two X chromosomes) large parts of one chromosome are methylated to disallow double expression of some proteins
- Housekeeping genes are not methylated

CpG Islands
- Methyl-C can mutate to T → pretty easy to do (amino → carboxyl and H → CH_3)
- The net result is that CpG is less common than expected
    - Frequency of CpG < (Frequency of C) * (Frequency of G)
- Because housekeeping genes are not methylated, they retain their CpG character
    - The CpG to TpG transition is less likely here, so there are large densities of CpG in promoter regions
- Typical length → few 100 to few 1000 bp
- Questions
    - If short sequence (200 bp), is that a CpG island?
    - If long sequence (1000-10000 bp), can you find CpG islands?

Markov and Hidden Markov Models (HMM)
- Textbook and tutorial article are good references

- Can handle dependency on adjacent positions (CpG)

Markov Chains
- Zeroth order means that the system is independent of its neighbors
- 1st order depends on previous value (ACGT)

1st order Markov Model
- States: A,C,G,T
- Have probabilities of transitioning from A→G, T→C, etc, etc…
- Emissions → corresponding letter
- Can add special Begin and End states
  - Proteins always start with Met
- Probabilities are dependent only on the previous state → $P(x_n \mid x_{n-1})$

Is a short sequence a CpG island?
- Gather statistics from larger sequences and CpG islands
  - Probability of CpG is 27%
  - For non-CpG island data → 8%
- Discrimination/classification
  - Look at the log-likelihood ratio of CpG model vs background (non-CpG) model
  - Take logs of ratios of tables for CpG and non-CpG data to create a table for the beta data
    - CpG is $2^{1.8}$ more likely in CpG model vs background
  - Add up beta values for all subsequent pairs in data, if score is positive it most likely is a CpG island
- Since Cs are more likely in CpG islands, could also have an independent model looking at the number of Cs in data
- Do not need to fully understand mechanism

1st order Weight Matrix Model (WMM)
- Column n is dependent on column (n-1)
- Need to have enough example data to cover number of parameters
- WMM are usually 0th order because of this
- Number of parameters grows exponentially with order of model

Given a long sequence where are the CpG islands in it? (if any?)
- Approach 1 → score subsequences (windows) of the model
  - Simple, but arbitrary, is fixed length a good idea?, inflexible
- Approach 2 → combine the +/- models

Combined model
- Incorporates CpG+ and CpG- models
- E.g. How often is a CpG- A followed by a CpG+ G ??

    - Involves crossing between models
- Defines probability distributions of sequences
- Is it more likely a stretch of NT comes from the CpG+ or the CpG-?
    - Every NT is ambiguous in terms of where it came from
    - The state sequence is hidden → only emission sequence is detected
        - A *Hidden* Markov Model (HMM)
    - Can try to label states (NT) to determine which part of model NT came from
        - Impossible though, because there are non-zero probabilities that either part of model can emit entire sequence
    - What is the most probable path through the model that emits this sequence? Where did it come from?