Scoring Matrices
CS795

Huzefa Rangwala
CS795 (Spring 2010)
rangwala AT cs DOT gmu DOT edu
Why are scoring matrices needed?

- Excercise!!
Why are scoring matrices needed?

- Play the most critical role in determining how two sequences align with each other.
  - After all, the alignment algorithm tries to find the alignment that optimizes the score.
- Different scoring matrices can lead to dramatically different alignments.
- Should capture common function/role/property between pairs.
- The score should encode the biological likelihood of ‘a’ to be evolved into ‘b’
Types of Scoring Matrices

- Identity Matrices (1 - identical pairs and 0 - non-identical pairs)
- DNA Scoring Matrix - definitions are based on whether there is an inversion, transversion or transition.
- Chemical similarity scores (for proteins) - based on which chemical properties they share like hydrophobocity, size, weight, buried/not
- Observed matrices - constructed by analyzing substitution frequencies of matrices in alignments of known families of proteins. Most widely used, aimed at capturing evolutionary signals.
Constructing Amino Acid Scoring Matrices

Let $p_a, p_b$ be the occurrence probability of amino acids $a$ and $b$ (i.e., their normalized frequency).
Let $p_{a,b}$ be the joint probability of $a$ aligned against $b$ in the correct alignment.
Then:

$$S(a, b) = \log \frac{p_{a,b}}{p_a \ p_b}.$$
Constructing Amino Acid Scoring Matrices

- Thus, in order to determine/construct a scoring matrix we need to determine the various independent and joint probabilities.

- General Idea
  - Determine these experimentally by looking at confirmed alignments.

- Potential Problems:
  - How to ensure coverage of all possible cases?
  - Different sequences are at different evolutionary distances from each other
    - $p_{a,b}$ can be too pessimistic if the estimates are based on related or it can be too optimistic if it is based on distant proteins.
PAM Matrices

- PAM matrices were one of the first widely-used amino acid scoring matrices.
  - Developed Dayhoff and coworkers in 1979.
- They are derived by considering how the sequence of related proteins changes as a function of their evolutionary divergence.
- They are widely used even today.
PAM Distance

- PAM stands for *point accepted mutation* or *percent accepted mutation*.
- “Point accepted mutation”
  - Single amino acid change that was incorporated into the protein and passed on its progeny
    - that is, it was not lethal!
- Definition:
  - Two sequence $S_1$ and $S_2$ are one PAM unit diverged if a series of point accepted mutations (and no insertions or deletions) has converted $S_1$ to $S_2$ with an average of one point mutation per 100 amino acids.
- It is a measure of evolutionary divergence.
  - If two sequences are at a distance of k-PAM units this does not mean that they have a k% sequence difference
    - a position can mutate back to its original amino acid
    - In general, 200 PAM units is about 20% sequence identity
Ideal Way of Building PAM Matrices

- Using the PAM distance measure we can build a sequence of scoring matrices, which are suited for scoring sequences at different levels of evolutionary divergence
  - PAM-k matrices
- How?
  - Collect a large set of sequence pairs that are k-PAM units diverged
  - Use these alignments to compute the required single and joint probability distributions
  - Build the matrix using the log-odds-ratio of the various probabilities.
- What is the problem with this approach?
Practical Approach for Building PAM Matrices

- Approach for small number of PAM units
  - Take a set of sequence pairs that are highly similar (less than 15% divergent)
    - Sequences will be of the same length
  - Use these alignments to compute the required probabilities and construct PAM-k matrices for small values of k.
Practical Approach for Building PAM Matrices

- **Approach for large number of PAM units**
  - Let $p_{i,j}^1$ be the probability of amino acid $i$ being mutated to amino acid $j$ in sequences that are one-PAM unit apart (i.e., the joint probability).
  - Let $P^1$ be the $20 \times 20$ matrix of these probabilities, and let $P^k = (P^1)^k$ be the matrix obtained by multiplying $P^1$ with itself $k$ times.
  - $p_{i,j}^k$ is treated as the probability of amino acid $i$ being mutated to amino acid $j$ in sequences that are $k$-PAM units apart.
  - Then:
    \[
    S^k(i, j) = \log \frac{f_i p_{i,j}^k}{f_i f_j} = \log \frac{p_{i,j}^k}{f_j},
    \]
    where $f_i$ and $f_j$ is the frequency of amino acids $i$ and $j$. Note that $f_i p_{i,j}^k$ is treated as the “observed” frequency of $i$ aligned against $j$. 

BLOSUM

- BLOSUM matrices were constructed by analyzing *blocks* in multiple sequence alignment of related proteins.
  - A block is a highly conserved region of a MSA and corresponds to a motif.
  - There is a database of such blocks, called the BLOCKS database.
- Initial set of BLOSUM matrices were constructed by looking around 3000 blocks from 800 proteins.
- Overall approach:
  - For each block, look at all pairwise induced alignments (restricted within these blocks) and compute the required background and joint probability distributions for the standard log-odds-ratio scoring method.
BLOSUM and Evolutionary Distance

- Evolutionary distance-sensitive matrices are constructed by eliminating certain sequences from each block.
- BLOSUM-X matrices:
  - For each block, eliminate each sequence that has at least X% sequence identity with another sequence in the block (w.r.t. the block)
    - i.e., find a maximal independent set of sequences less than X% sequence identity.
  - Perform the probability estimation exercise using these reduced blocks.
- BLOSUM62 is a widely used matrix
  - Default choice for blastp.
PAM vs BLOSUM

- PAM - usually for global alignments
- BLOSUM - usually for local alignments

There are many problem-specific scoring matrices, and in general constructing one for your own problem will improve your results!
Position Specific Scoring Matrices

- Scoring matrices used to model a particular MSA and are derived directly from it.
  - Closely related to sequence-against-profile alignment.

**Definition 1 (Position Specific Scoring Matrix (PSSM))** Given a multiple sequence alignment $\mathcal{M}$ of length $n$, the PSSM is a $20 \times n$ matrix $S$, such that $S(i, k)$ is the score of aligning amino acid $i$ against column $k$ of the MSA.

Specifically,

$$S(i, k) = \log \frac{Q_i}{P_i},$$

where $Q_i$ is the estimated probability of observing amino acid $i$ in column $k$, and $P_i$ is the estimated probability of observing amino acid $i$ in the MSA.
Position Specific Scoring Matrices

How to estimate $Q_i$?

- From the raw frequencies of the particular column. That is, $Q_i = f_i$, where $f_i$ is the observed frequency of the $i$th amino acid.
  - It may under-sample certain amino acids due to the finite number of sequences.
- Use \textit{pseudo-counts} by taking advantage of existing scoring matrices. In this approach,

\[
Q_i = \frac{\alpha f_i + \beta \gamma_i}{\alpha + \beta},
\]

where $\gamma_i = \sum_j \frac{f_j}{P_j} q_{j,i}$, and $q_{j,i}$ is the probability of amino acid $j$ being mutated to amino acid $i$ and is derived from the scoring matrix.
PSI-BLAST

- Readings

- External Links for Reference
  - http://www.techfak.uni-bielefeld.de/bcd/Curric/PrwAli/nodeD.html