6.874/... Recitation 2

Courtesy of an MIT Teaching Assistant.
Reminders

• Pset 1 posted – due Feb 20\textsuperscript{th} (no extra problem)
• Pset 2 posted – due Mar 13\textsuperscript{th}
• Project teams due – Feb 25\textsuperscript{th}
  • Interests and background directory has been posted
• Lecture videos will be posted on MITx soon – next week?
Today

• Clustering (6.874 topic)
• Biology review
• Alignment
Clustering – K-means

• Group points together based on how ‘close’ they are to each other
• Dataset of unlabelled points: \( X = \{x_1, x_2, ..., x_N\}, x_n \in \mathbb{R}^d \)
• Assume K clusters – each is defined by a centroid \( \mu_k \)
• \( r_{nk} = 1 \) if \( x_n \) belongs to cluster \( k \)
• Find unknowns \( \mu_k \) and \( r_{nk} \)

\[
J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| x_n - \mu_k \|^2
\]
Algorithm 10.1 $K$-Means Clustering

1. Randomly assign a number, from 1 to $K$, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the $K$ clusters, compute the cluster centroid. The $k$th cluster centroid is the vector of the $p$ feature means for the observations in the $k$th cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).
### Table 1 Gene expression similarity measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan distance (city-block distance, L1 norm)</td>
<td>$d_{fg} = \sum_c</td>
<td>e_{fc} - e_{gc}</td>
</tr>
<tr>
<td>Euclidean distance (L2 norm)</td>
<td>$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$</td>
<td></td>
</tr>
<tr>
<td>Mahalanobis distance</td>
<td>$d_{lg} = (e_f - e_g')\Sigma^{-1}(e_f - e_g)$, where $\Sigma$ is the (full or within-cluster) covariance matrix of the data</td>
<td></td>
</tr>
<tr>
<td>Pearson correlation (centered correlation)</td>
<td>$d_{fg} = 1 - r_{fg}$ with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}<em>f)(e</em>{gc} - \bar{e}<em>g)}{\sqrt{\sum_c (e</em>{fc} - \bar{e}<em>f)^2\sum_c (e</em>{gc} - \bar{e}_g)^2}}$</td>
<td></td>
</tr>
<tr>
<td>Uncentered correlation (angular separation, cosine angle)</td>
<td>$d_{fg} = 1 - r_{fg}$ with $r_{fg} = \frac{\sum_c e_{fc}e_{gc}}{\sqrt{\sum_c e_{fc}^2\sum_c e_{gc}^2}}$</td>
<td></td>
</tr>
<tr>
<td>Spellman rank correlation</td>
<td>As Pearson correlation, but replace $e_{gc}$ with the rank of $e_{gc}$ within the expression values of gene $g$ across all conditions $c = 1 \ldots C$</td>
<td></td>
</tr>
<tr>
<td>Absolute or squared correlation</td>
<td>$d_{fg} = 1 -</td>
<td>r_{fg}</td>
</tr>
</tbody>
</table>

$d_{fg}$, distance between expression patterns for genes $f$ and $g$. $e_{gc}$, expression level of gene $g$ under condition $c$.  

Courtesy of Macmillan Publishers Limited. Used with permission.  
Source: D’haeseleer, Patrik. "How Does Gene Expression Clustering Work?."  
Hierarchical clustering

• Organize data in a tree
  • Leaves are individual genes/species
  • Path lengths between leaves are distances
  • Similar points should lie in same lower subtrees

• Used to reveal evolutionary history of sequences
Hierarchical clustering

• What dissimilarity measure should be used?
  • To compare individual points
• What type of linkage should be used?
  • To compare clusters with each other
• Where do we cut dendrogram to obtain clusters?
Algorithm 10.2  *Hierarchical Clustering*

1. Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:
   
   (a) Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.
<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>$d(A, B) = \max_{x \in A, y \in B} d(x, y)$</td>
<td>Maximal intercluster dissimilarity</td>
</tr>
<tr>
<td>Single</td>
<td>$d(A, B) = \min_{x \in A, y \in B} d(x, y)$</td>
<td>Minimal intercluster dissimilarity</td>
</tr>
<tr>
<td>Average</td>
<td>$d(A, B) = \sum_{x \in A, y \in B} \frac{d(x, y)}{</td>
<td>A</td>
</tr>
<tr>
<td>Centroid</td>
<td>$d(A, B) = d\left(\sum_{x \in A} \frac{x}{</td>
<td>A</td>
</tr>
</tbody>
</table>
Model selection

• Hierarchical clustering – cutoff tree at certain point
• K-means – how to choose K?
  • Balance number of clusters (# of parameters – K=n is uninformative) and the variance of the clusters
  • BIC Score – general model selection criterion
  • $\text{BIC} = -2 \times \log\text{likelihood} + d \times \log(N)$
  • Can use to decide whether to split a cluster
  • Computer BIC score of cluster and two potential child clusters – if BIC score is lower after split, do not accept split
Biclustering

• Simultaneous clustering of rows and columns of a matrix
• Bicluster – subset of rows which exhibit similar behavior across a subset of columns, or vice versa
Fig. 4. Bicluster structure. (a) Single bicluster, (b) exclusive row and column biclusters, (c) checkerboard structure, (d) exclusive rows biclusters, (e) exclusive columns biclusters, (f) nonoverlapping biclusters with tree structure, (g) nonoverlapping nonexclusive biclusters, (h) overlapping biclusters with hierarchical structure, and (i) arbitrarily positioned overlapping biclusters.
Biology Review
Selection

• Negative selection (purifying/natural selection) – removal of deleterious traits
• Positive selection – increases prevalence of adaptive traits

• Thinking about selection happening at different levels
  • Protein level: Sequence -> Structure -> Function
  • RNA level: splicing, degradation/processing (NMD)
  • DNA level: DNA-protein binding sites
Synonymous/Non-synonymous mutations

• Redundancy built into the genetic code
• Synonymous – one base changes for another in an exon, but the resulting amino acid sequence is unchanged
• Non-synonymous – new AA
• Can affect splicing, mRNA processing - so may not be silent

![Image by MIT OpenCourseWare.](https://ocw.mit.edu)
Side-chain biochemistry

• Amino acids classified by properties of side chains
  • Grouped by general properties
• Substitutions of amino acid with another of similar chemical properties may conserve protein function
Side chain size (Trp – W)

D. Amino Acids with Hydrophobic Side Chain

A A A A A A A A A

Valine (Val) Valine (Val) Valine (Val) Valine (Val) Valine (Val) Valine (Val) Valine (Val) Valine (Val) Valine (Val)
V V V V V V V V V

Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile) Isoleucine (Ile)
I I I I I I I I I

Leucine (Leu) Leucine (Leu) Leucine (Leu) Leucine (Leu) Leucine (Leu) Leucine (Leu) Leucine (Leu) Leucine (Leu) Leucine (Leu)
L L L L L L L L L

Methionine (Met) Methionine (Met) Methionine (Met) Methionine (Met) Methionine (Met) Methionine (Met) Methionine (Met) Methionine (Met) Methionine (Met)
M M M M M M M M M

Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe) Phenylalanine (Phe)
F F F F F F F F F

Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr) Tyrosine (Tyr)
Y Y Y Y Y Y Y Y Y

Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp) Tryptophan (Trp)
W W W W W W W W W

© unknown source. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [http://ocw.mit.edu/help/faq-fair-use/](http://ocw.mit.edu/help/faq-fair-use/).
Disulfide bond

• Important in protein folding – holds two distant portions of protein together
• Occurs between Cys residues
Next-generation sequencing

• Sequencing is always of DNA
  • Need to convert RNA to DNA by reverse transcription (RT)

• Illumina is current leader in the field
  • 8 lanes on a flow cell
  • Each lane can sequence 200 million 100bp reads – 20 Gbps!
  • Can sequence multiple samples per lane by barcoding
  • Requires (heterogeneous) population of cells to get enough DNA for sample

• Single cell sequencing applications are becoming more common (RNAseq)

• Single molecule technologies are still being developed – PacBio
Alignment
Alignment

\[ F(i, j) = \max \begin{cases} 
F(i-1, j-1) + s(x_i, y_j), \\
F(i-1, j) - d, \\
F(i, j-1) - d. 
\end{cases} \]

\[ F(i, j) = \max \begin{cases} 
0, \\
F(i-1, j-1) + s(x_i, y_j), \\
F(i-1, j) - d, \\
F(i, j-1) - d. 
\end{cases} \]
Local alignment example

Do a local alignment between these using PAM250 and gap penalty -2:

### AWEK

### FWEF
Local alignment solution

<table>
<thead>
<tr>
<th></th>
<th>Gap</th>
<th>A</th>
<th>W</th>
<th>E</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-8</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>19</td>
<td>17</td>
</tr>
</tbody>
</table>

alignment: W E  
W E
Global alignment solution

<table>
<thead>
<tr>
<th></th>
<th>Gap</th>
<th>A</th>
<th>W</th>
<th>E</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap</td>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
<td>-8</td>
</tr>
<tr>
<td>F</td>
<td>-2</td>
<td>-4</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
</tr>
<tr>
<td>W</td>
<td>-4</td>
<td>-6</td>
<td>13</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>-6</td>
<td>-4</td>
<td>-6</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>F</td>
<td>-8</td>
<td>-6</td>
<td>-4</td>
<td>15</td>
<td>12</td>
</tr>
</tbody>
</table>

alignment: A  W  E  K  
            F  W  E  F
<table>
<thead>
<tr>
<th></th>
<th>Global</th>
<th>Semiglobal</th>
<th>Local (gapped)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalties at edges?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Reset to 0 instead of including negative entries?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>End of alignment</td>
<td>Bottom right entry</td>
<td>Highest score entry in bottom row or rightmost column</td>
<td>Highest score entry in matrix</td>
</tr>
</tbody>
</table>
Reminders

- Pset 1 posted – due Feb 20th (no extra problem)
- Pset 2 posted – Due Mar 13th
- Project teams due – Feb 25th
  - Interests and background directory has been posted
- Lecture videos will be posted on MITx soon – next week?