Whole genome alignments

http://faculty.washington.edu/jht/GS559_2013/

Genome 559: Introduction to Statistical and Computational Genomics
Prof. James H. Thomas
Extreme value distribution

\[ P \quad S \geq x = 1 - e^{-e^{-x}} \]

S is data score, x is test score

\[ P \quad S \geq x = 1 - e^{-e^{-\lambda(x-\mu)}} \]

S is data score, x is test score, \( \mu \) is mode, \( \lambda \) is width
Summary score significance

- A **distribution** plots the frequencies of types of observation.
- The area under the distribution curve is 1.
- Most statistical tests compare observed data to the expected result according to a **null hypothesis**.
- Sequence similarity scores follow an **extreme value distribution**, which is characterized by a long tail.
- The **p-value** associated with a score is the area under the curve to the right of that score.
- Selecting a **significance threshold** requires evaluating the cost of making a mistake.
- **Bonferroni correction**: Divide the desired p-value threshold by the number of statistical tests performed.
- The **E-value** is the expected number of times that a given score would appear in a randomized database.
Whole genome alignments

Why?

• genome-wide alignment data (efficient)

• inference of shared (orthologous) genes across species

• genome evolution
known gap in assembly
alignment discontinuity (e.g. translocation break point)
questionable alignment segment
= sequence present but unalignable

individual genome alignments, darker = higher scoring

averaged conservation for 17 genomes

UCSC Browser track
How are genome-wide alignments made?

- mouse and human genomes are each about $3 \times 10^9$ nucleotides.

- how many calculations would a dynamic programming alignment have to make?

- at a minimum - 3 integer additions and 3 inequality tests for each DP matrix position

- DP matrix size is $3 \times 10^9$ by $3 \times 10^9$

- about $6 \times (3 \times 3 \times 10^{18}) = 5.4 \times 10^{19}$ calculations!

  Age of the universe is about $4.3 \times 10^{17}$ seconds

(by the way, there are other problems too, including assuming colinearity)
Making large searches faster

- Most common method is the BLAST search (Basic Local Alignment Search Tool). Only the initial step is different from dynamic programming alignment.

- Search sequence broken into small words (usually 3 residues long for proteins). $20 \times 20 \times 20 = 8,000$ protein words. These act as seeds for searches.

- The target dataset is pre-indexed for all positions that match each search word above some score threshold (using a score matrix such as BLOSUM62).
BLAST searches (cont.)

- For example, the search sequence word “WVH” might score above threshold with these indexed sequences:

<table>
<thead>
<tr>
<th>Indexed word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WVH</td>
<td>23</td>
</tr>
<tr>
<td>WIH</td>
<td>22</td>
</tr>
<tr>
<td>WVY</td>
<td>17</td>
</tr>
<tr>
<td>WIY</td>
<td>16</td>
</tr>
</tbody>
</table>

- Target sequences around each indexed word hit are retrieved and the initial match is extended in both directions:

```plaintext
...VFEWVHLLP...  your sequence
doctorase    database (many sites)
```
Result - instead of aligning these 3 amino acids to *everything*, they are aligned only with the tiny fraction of sequence regions that are good candidates for a valid alignment.

(note- blast actually looks for **two** such matches close to each other)
Extension and scoring

...QSVFEWVHLLPPGA...
..WIY..

Match Score: 16
Total Score: 16

...QSVFEWVHLLPPGA...
..WIYQ..

Match Score: -3
Total Score: 13

...QSVFEWVHLLPPGA...
..WIYQK..

Match Score: -2
Total Score: 11

...QSVFEWVHLLPPGA...
..WIYQKA..

Match Score: -1
Total Score: 10

[mention gap variant]
Extension termination and Reporting

- Extension is continued until the alignment score drops below some threshold (usually 0, like local alignments).

- Extensions whose maximal cumulative score is above some threshold are kept for reporting to user.

- For web interfaces, various formatting, links, and overviews are added.

- It is also easy to set up blast on your local computer; useful for custom databases and automation.
Key to speed: word matching and prior indexing

• Though gapped blast local alignment is slow, only a very small part of total search space is analyzed.

• Because word matches are indexed prior to the search, the relevant parts of search space are reached quickly.

• Tradeoff is in sensitivity – occasionally matches will be missed (e.g. when they are distant enough and dispersed enough that no local word pairs match well enough).
BLAST whole genome against another

- Runtime (my desktop) for mouse vs. human, about 24 hours*.
- Extract best match segments, reverse blast
- Keep reciprocal best match regions as anchors
- Schematic of part of results:

* megablastn with repeat-masked human genome
Dynamic programming after BLAST matching

Anchored DP alignment: if two reciprocal best blast matches are nearby and in the same orientation, DP align everything between them.

M x N manageable