PAM and BLAST

Ulf Leser
This Lecture

• **Substitution Matrices**
  - PAM distance
  - PAM matrices

• **Scaling up Local Alignments**
  - BLAST
**Substitution Matrices**

- **Recall**
  - A *scoring function* is a function $s: \Sigma \times \Sigma' \rightarrow \text{Integer}$
  - We also call $s$ a substitution matrix
  - Direct similarity
    $$sim'(A, B) = \sum_{i=1}^{n} s(A[i], B[i])$$

- **DNA**: symmetric, simple matrices

- **Protein sequences** are different
  - Very heterogeneous properties
  - Very different impact on folding
  - Substitutions *may change the 3D structure completely* or not at all
Amino Acids

**Small**
- **Glycine** (Gly, G) MW: 72.06
- **Alanine** (Ala, A) MW: 89.09
- **Serine** (Ser, S) MW: 97.08, $pK_a = 16$
- **Threonine** (Thr, T) MW: 101.11, $pK_a = 16$
- **Cysteine** (Cys, C) MW: 121.15, $pK_a = 8.36$

**Hydrophobic**
- **Valine** (Val, V) MW: 99.14
- **Leucine** (Leu, L) MW: 113.16
- **Isoleucine** (Ile, I) MW: 113.16
- **Methionine** (Met, M) MW: 131.19
- **Proline** (Pro, P) MW: 87.12

**Aromatic**
- **Phenylalanine** (Phe, F) MW: 147.18
- **Tyrosine** (Tyr, Y) MW: 163.18
- **Tryptophan** (Trp, W) MW: 186.21

**Acidic**
- **Aspartic Acid** (Asp, D) MW: 115.09, $pK_a = 3.9$
- **Glutamic Acid** (Glu, E) MW: 129.12, $pK_a = 4.87$

**Amide**
- **Asparagine** (Asn, N) MW: 114.11
- **Glutamine** (Gln, Q) MW: 128.14
- **Histidine** (His, H) MW: 137.14, $pK_a = 6.04$
- **Lysine** (Lys, K) MW: 128.17, $pK_a = 10.79$
- **Arginine** (Arg, R) MW: 156.10, $pK_a = 12.48$
Example

|   | A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  | B  | Z  |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4  | -1 | -2 | -2 | 0  | -1 | -1 | 0  | -2 | -1 | -1 | -1 | -1 | -1 | -2 | -1 | 1  | 0  | -3 | -2 | 0  | -2 | -1 |
| R | -1 | 5  | 0  | -2 | -3 | 1  | 0  | -2 | 0  | -3 | -2 | 2  | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0  |    |
| N | -2 | 0  | 6  | 1  | -3 | 0  | 0  | 0  | 1  | -3 | -3 | 0  | -2 | -3 | -2 | 1  | 0  | -4 | -2 | -3 | 3  | 0  |    |
| D | -2 | -2 | 1  | 6  | -3 | 0  | 2  | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0  | -1 | -4 | -3 | -3 | 4  | 1  |    |
| C | 0  | -3 | -3 | -3 | 9  | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 |    |
| Q | -1 | 1  | 0  | 0  | -3 | 5  | 2  | -2 | 0  | -3 | -2 | 1  | 0  | -3 | -1 | 0  | -1 | -2 | -1 | -2 | 0  | 3  |    |
| E | -1 | 0  | 0  | 2  | -4 | 2  | 5  | -2 | 0  | -3 | -3 | 1  | -2 | -3 | -1 | 0  | -1 | -3 | -2 | -2 | 1  | 4  |    |
| G | -2 | 0  | 0  | -1 | -3 | -2 | -2 | 6  | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0  | -2 | -2 | -3 | -3 | -1 | -2 |    |
| H | -2 | 0  | 1  | -1 | -3 | 0  | 0  | -2 | 8  | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2  | -3 | 0  | 0  |    |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4  | 2  | -3 | 1  | 0  | -3 | -2 | -1 | -3 | -1 | 3  | -3 | -3 |    |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2  | 4  | -2 | 2  | 0  | -3 | -2 | -1 | -2 | -1 | 1  | -4 | -3 |    |
| K | -1 | 2  | 0  | -1 | -3 | 1  | 1  | -2 | -1 | -3 | -2 | 5  | -1 | -3 | -1 | 0  | -1 | -3 | -2 | -2 | 0  | 1  |    |
| M | -1 | -1 | -2 | -3 | -1 | 0  | -2 | -3 | -2 | 1  | 2  | -1 | 5  | 0  | -2 | -1 | -1 | -1 | -1 | 1  | -3 | -1 |    |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0  | 0  | -3 | 0  | 6  | -4 | -2 | -2 | 1  | 3  | -1 | -3 | -3 |    |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7  | -1 | -1 | -4 | -3 | -2 | -2 | -1 |    |
| S | 1  | -1 | 1  | 0  | -1 | 0  | 0  | 0  | -1 | -2 | -2 | 0  | -1 | -2 | -1 | 4  | 1  | -3 | -2 | -2 | 0  | 0  |    |
| T | 0  | -1 | 0  | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -1 | -2 | -1 | 1  | 5  | -2 | -2 | 0  | -1 | -1 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -3 | -2 | -3 | -1 | 1  | -4 | -3 | -2 | 11 | 2  | -3 | -4 | -3 |    |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2  | -1 | -1 | -2 | -1 | 3  | -3 | -2 | -2 | 2  | -7 | -1 | -3 | -2 |    |
| V | 0  | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3  | 1  | -2 | 1  | -1 | -2 | -2 | 0  | -3 | -1 | 4  | -3 | -2 |    |
| B | -2 | -1 | 3  | 4  | -3 | 0  | 1  | -1 | 0  | -3 | -4 | 0  | -3 | -3 | -2 | 0  | -1 | -4 | -3 | -3 | 4  | 1  |    |
| Z | -1 | 0  | 0  | 1  | -3 | 3  | 4  | -2 | 0  | -3 | -3 | 1  | -1 | -3 | -1 | 0  | -1 | -3 | -2 | -2 | 1  | 4  |    |

Where do all these numbers come from?
Is it Really Necessary?

- Let’s count how often a particular AA was replaced by any other AA
  - Using “sure” sequence alignments
- Replacement rate of Alanin (A) := 100%
- Obviously no equal distribution
- Even if we assume that mutations happen more or the less at the same rate
- ... they obviously don’t survive with the same probability
  - Mutations are suppressed to different degrees
  - A-W (Tryptophan): Strong suppression
  - A-S (Serin): Little suppression

<table>
<thead>
<tr>
<th>Code</th>
<th>Häufigkeit</th>
<th>Mutierbarkeit</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>0.091</td>
<td>54</td>
</tr>
<tr>
<td>A</td>
<td>0.077</td>
<td>100</td>
</tr>
<tr>
<td>G</td>
<td>0.074</td>
<td>50</td>
</tr>
<tr>
<td>S</td>
<td>0.069</td>
<td>117</td>
</tr>
<tr>
<td>V</td>
<td>0.066</td>
<td>98</td>
</tr>
<tr>
<td>E</td>
<td>0.062</td>
<td>77</td>
</tr>
<tr>
<td>K</td>
<td>0.059</td>
<td>72</td>
</tr>
<tr>
<td>T</td>
<td>0.059</td>
<td>107</td>
</tr>
<tr>
<td>I</td>
<td>0.053</td>
<td>103</td>
</tr>
<tr>
<td>D</td>
<td>0.052</td>
<td>86</td>
</tr>
<tr>
<td>P</td>
<td>0.051</td>
<td>58</td>
</tr>
<tr>
<td>R</td>
<td>0.051</td>
<td>83</td>
</tr>
<tr>
<td>N</td>
<td>0.043</td>
<td>104</td>
</tr>
<tr>
<td>Q</td>
<td>0.041</td>
<td>84</td>
</tr>
<tr>
<td>F</td>
<td>0.040</td>
<td>51</td>
</tr>
<tr>
<td>Y</td>
<td>0.032</td>
<td>50</td>
</tr>
<tr>
<td>M</td>
<td>0.024</td>
<td>93</td>
</tr>
<tr>
<td>H</td>
<td>0.023</td>
<td>91</td>
</tr>
<tr>
<td>C</td>
<td>0.020</td>
<td>44</td>
</tr>
<tr>
<td>W</td>
<td>0.014</td>
<td>25</td>
</tr>
</tbody>
</table>
Filling a Substitution Matrix

• We need **app. 400 values**
  – Matrix need not be symmetric

• **Possibility 1: Analytical**
  – Capture weight, polarity, size, ...
  – Find a scoring scheme to measure the difference between two AA
  – Needs to produce a single value per AA pair
  – Every scheme would be very hard to justify
  – Not used in practice

• **Possibility 2: Empirical**
  – See and count which substitutions survived at which frequency
  – Therefore, we need examples: *Pairs of homologues and aligned protein sequences*
  – Problem: Discern short-term from long-term effects
Margaret O. Dayhoff

- “Deduce evolutionary relationships of the biological kingdoms, phyla, and other taxa from sequence evidence”

- Collection of all known protein sequences
  - First edition: 65 proteins
  - Several releases followed
  - Resulted in the Protein Information Resource (PIR)

Thanks to Antje Krause
PAM: Point-Accepted Mutations


- PAM has two meanings
  - 1 PAM – Unit for measuring the similarity of two AA sequences
  - PAM-X matrix – Substitution matrix to use when aligning two sequences that are X PAM distant
PAM as Distance Measure

• Definition
Let $S_1$, $S_2$ be two protein sequences with $|S_1|=|S_2|$. We say $S_1$ and $S_2$ are $x$ PAM distant, iff. $S_1$ most probably was produced from $S_2$ with $x$ single mutations per 100 AAs

• Remarks
− PAM is motivated by evolution
− Assumptions: Mutations happen with the same rate at every position of a sequence
− If mutation rate is high, mutations will occur again and again at the same position
− PAM ≠ %-sequence-identity
PAM as Distance Measure

- PAM does not deal with INSDELS – only replacements
- The PAM distance of two sequences can be estimated easily from their \%-sequence-similarity
  - Jukes-Cantor model
  - Method skipped
- Sequences with a PAM distance of 250 and above (\%-sequence-identity < 20\%) are probably not homologues
  - Twilight zone
  - Which \%-sequence-identity will two random protein sequences have?
Generating a PAM Matrix

- The PAM-X matrix contains a measure for the probability that a given AA was replaced by another given AA in two sequences that are x PAM distant.

- Estimated from data
  - Let \((S_{1,1}, S_{2,1}), \ldots, (S_{1,n}, S_{2,n})\) by \(n\) pairs of aligned sequences, each \(x\) PAM distant
  - Compute \(f(i)\), the relative frequency of AA \(A_i\) in all pairs
  - Compute \(f(i,j)\), the relative replacement frequency from \(A_i\) to \(A_j\)
    - Number of positions \(k\) in the alignments with \(S_{1,z}[k]=A_i\) and \(S_{2,z}[k]=A_j\) or vice versa
  - Then

\[
M_x(i, j) = \log \left( \frac{f(i, j)}{f(i) \times f(j)} \right)
\]
Some Explanations

• Again: A log-likelihood ratio, combining
  – chances to generate this mutation by chance given the relative frequencies of the two involved AAs
  – observed frequency of this mutation

\[
M_x(i, j) = \log \left( \frac{f(i, j)}{f(i) \cdot f(j)} \right)
\]

• Meaning
  – M(i,j) = 0: No selection
  – M(i,j) < 0: Negative selection, suppression of mutation
  – M(i,j) > 0: Positive selection, mutation is favored
Example

\begin{align*}
S_{1,1}: & \quad \text{ACGGTGAC} \\
S_{2,1}: & \quad \text{AGG\_TGCC} \\
S_{1,2}: & \quad \text{GTT\_AGCTA} \\
S_{2,2}: & \quad \text{TTTCAG\_TA} \\
S_{1,3}: & \quad \text{GGTCAA} \\
S_{2,3}: & \quad \text{AGTC\_A}
\end{align*}

Relative frequencies

\begin{align*}
A: & \quad 11/42 \\
C: & \quad 8/42 \\
G: & \quad 12/42 \\
T: & \quad 11/42
\end{align*}

Mutation rates

\begin{array}{|c|c|c|c|}
\hline
 & A & C & G \\
\hline
A & 4/19 & 1/19 & 0/19 \\
C & 2/19 & 1/19 & 0/19 \\
G & 4/19 & 1/19 & \\
T & 5/19 & & \\
\hline
\end{array}

Matrix

\begin{array}{|c|c|c|c|}
\hline
 & A & C & G \\
\hline
A & 0,48 & 0,02 & -0,15 \\
C & 0,46 & -0,01 & - \\
G & 0,41 & -0,15 & \\
T & 0,58 & & \\
\hline
\end{array}
Problems

• Depends on predefined alignments
  – Hard to define uniquely

• For larger $x$ values, we need a substitution matrix to find optimal alignments
  – A hen-egg problem

• Several assumptions
  – Mutations are equally likely at all positions in a sequence
  – Mutation probabilities are independent from sequence neighbors

• Requires large $n$ for each $x$ to adequately capture rare mutations
  – Impossible
Real Substitution Matrices

• One more assumption: **Molecular clock**
  – Assume that mutations appear with equal rate over time
  – PAM-x matrices can be computed by iterated application of PAM-1

• Complete procedure
  – Choose set of \( n \) pairs with small distance and align manually
  – Use these alignments to compute \( M_1 \)
  – Compute \( M_x = (M_1)^x \)

• **BLOSUM: BLOcks SUbstitution Matrix**
  – Henikoff and Henikoff, 1993
  – Removes assumption of equal mutation rates across each sequence
    by only considering **conserved blocks**
  – Direct estimation for different PAM distances instead of error-
    propagating self multiplication
This Lecture

- Substitution Matrices
  - PAM distance
  - PAM matrices
- Scaling up Local Alignments
  - BLAST
Scaling Up Local Alignment

- Searching similar sequences (with a high local alignment score) is a fundamental operation in Bioinformatics
- Sequence databases grow exponentially
- Naïve method does not scale at all
- We need faster algorithms, even if they sometimes fail
Similarity Search Problems and their Accuracy

• Task: Given a sequence $s$ and a database $D$, find all sequences $T$ in $D$ that are sufficiently similar to $s$
  – Often, exactly computing $T$ is not feasible and not necessary (think of the WWW)
  – Assume a method that finds a set $X$ of answers for $s$

• How good is this method?
  – Some sequences will be in $X$ and $T$ – true positives
  – Some will be in $X$ but not $T$ – false positives
  – Some will be in $T$ but not $X$ – false negatives
  – Some will be neither in $X$ nor $T$ – true negatives

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reality</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>TruePositive (TP)</td>
<td>FalsePositive (FP)</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>FalseNegative (FN)</td>
<td>TrueNegative (TN)</td>
</tr>
</tbody>
</table>
Precision and Recall

- **Precision** = \( TP/(TP+FP) \)
  - What is the fraction of correct answers in \( X \)?
  - Related to specificity

- **Recall** = \( TP/(TP+FN) \)
  - Which fraction of correct answers from \( T \) are also in \( X \)?
  - Also called sensitivity

- **Trade-Offs**
  - Usual methods compute a score per element of \( D \)
  - All sequences with a score above a threshold \( t \) are returned as \( X \)
  - Increasing \( t \): higher precision, lower recall
  - Lowering \( t \): lower precision, higher recall
  - ... if the score correlates with correctness ...

\begin{table}
\begin{tabular}{|c|c|c|}
\hline
Reality & Prediction & Reality \\
\hline
+ & TruePositive (TP) & FalsePositive (FP) \\
- & FalseNegative (FN) & TrueNegative (TN) \\
\hline
\end{tabular}
\end{table}
Example

• Let $|DB| = 1000$, $|X| = 15$, $|T| = 20$, $|X \cap T| = 9$

<table>
<thead>
<tr>
<th></th>
<th>Real: Positive</th>
<th>Real: Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg: Positive</td>
<td>TP = 9</td>
<td>FP = 6</td>
</tr>
<tr>
<td>Alg: Negative</td>
<td>FN = 11</td>
<td>TN = 974</td>
</tr>
</tbody>
</table>

- Precision = $\frac{TP}{TP+FP} = \frac{9}{15} = 60\%$
- Recall = $\frac{TP}{TP+FN} = \frac{9}{20} = 45\%$

• Assume we increase $t$: $|X| = 10$, $|X \cap T| = 7$

<table>
<thead>
<tr>
<th></th>
<th>Real: Positive</th>
<th>Real: Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg: Positive</td>
<td>TP = 7</td>
<td>FP = 3</td>
</tr>
<tr>
<td>Alg: Negative</td>
<td>FN = 13</td>
<td></td>
</tr>
</tbody>
</table>

- Precision: 70%, recall = 35%
BLAST

• Altschul, Gish, Miller, Myers, Lipman: „Basic Local Alignment Search Tool“, J Mol Bio, 1990
  – A heuristic algorithm for sequence similarity search
  – Much faster than the naïve way
  – High recall, but not perfect
  – Very successful: You “blast” a sequence
  – NCBI runs thousands of BLAST searches every day

• A family of tools
  – Gapped-BLAST, PSI-BLAST, MegaBlast, BLAST-ALL, PATHBLAST, Name-BLAST, ...
  – BLAST for DNA, protein, DNA-protein, protein-DNA, ...
  – We only look at the simple DNA-DNA version
  – We skip several small and dirty tricks
Fundamental Idea

- **Seeds**: If two sequences have a good local alignment, then this area contains, with very high probability, a smaller area where the match is even better (or even exact)

```
TTGACTCGATTATAGTCGCGGATATACTATCG
CCTATCACAAGAATATAGTCCCTGATCCAGC

TTGACTC  GATTTATAGTCGCGGAT  ATACTATCG
CCTATCAACA  GAATATAGTCCCTGAT  CCAGC

TTGACTC  GATTTATAGTCGCGGAT  ATACTATCG
CCTATCAACA  GAATATAGTCCCTGAT  CCAGC
```
Algorithm

• Given \( s \) and \( DB = \{d_i\} \)
• 1. Compute all substrings \( s_i \) of \( s \) of length \( w \)
   – Also called \( w \)-grams
   – How many?
• 2. Find all approximate occurrences of all \( s_i \) in all \( d_j \)
   – Gapfree alignment using a matrix; score must be above threshold \( t \)
   – Hits are called \textit{seeds} – approx. occurrences of some \( s_i \) in some \( d_j \)
• 3. Extend \textit{seeds} to left and right in \( s_i \) and \( d_j \) until
   – Constantly update the similarity score
   – ... the score drops too sharply
   – ... \( s_i \) or \( s_j \) ends
   – ... the score gets too bad compared to other hits found earlier
Example

\[w=5, \ t=3, \ Matrix: \ M=+1, \ R=-1\]
\[s=ACGTGATA\]
\[d=GATTGACGTGACTGCTAGTGATACTATAT\]
\[s_1=ACGTG\]
\[s_2=CGTGA\]
\[s_3=GTGAT\]
\[s_4=TGATA\]
Visualization
Properties

- **Finding seeds efficiently** requires more work
  - Pre-compute all w-grams of all D
  - Group by w-gram
  - Called a **hash-index** (should be kept in main memory)
  - Lookup: Given w, find all matching w-grams, then all seeds

- **Exclusion method**
  - Vast majority of all sequences are never looked at because they do not contain a seed
  - This “seed” idea is used in essentially all fast alignment methods

- **Where it fails**
  - **Sensitive to t**: Too high – missing hits; too low – too slow
  - Does not consider gaps
BLAST Screenshots
BLAST-2


- Faster
  - BLAST: 90% of time spend in extensions
  - BLAST2: Two seeds in short distance
    - Needs a decrease in $t$

- Higher sensitivity
  - BLAST didn’t even consider gaps in the extension phase
  - BLAST2: Full local alignment starting from seeds
    - Allows an increase of $t$
Further Reading

- Substitution matrixes: Krane & Raymer, Chapter 3
- BLAST, BLAST2: Merkl & Waack, Chapter 12