Alignment: The process of lining up two or more sequences to achieve maximal levels of **identity** (and **conservation**, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

What are the important parameters in establishing a homology between a query sequence and a sequence of the database?

- The scoring scheme (PAM/BLOSUM)
- The score (size) of the actual alignment
- The size of the database
A Dayhof matrix a (biomatch /random) probability model

The scoring scheme (PAM/BLOSUM)
## Major Differences between PAM and BLOSUM

<table>
<thead>
<tr>
<th>PAM</th>
<th>BLOSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built from global alignments</td>
<td>Built from local alignments</td>
</tr>
<tr>
<td>Built from small amount of Data</td>
<td>Built from vast amount of Data</td>
</tr>
<tr>
<td>Counting is based on minimum replacement or maximum parsimony</td>
<td>Counting based on groups of related sequences counted as one</td>
</tr>
<tr>
<td>Perform better for finding global alignments and remote homologs</td>
<td>Better for finding local alignments</td>
</tr>
<tr>
<td>Higher PAM series means more divergence</td>
<td>Lower BLOSUM series means more divergence</td>
</tr>
</tbody>
</table>
Alignment scoring

We want to assign a score to an alignment that gives a measure of the relative likelihood that the two sequences are related as opposed to being unrelated.

. . . . S T S A . .
. . . . S T T L . .

We can score the ratio of two probabilities match (M) vs. random (R)

(additive scoring) $\log \left( \frac{P(x,y|M)}{P(x,y|R)} \right) + \log \left( \frac{P(x,y|M)}{P(x,y|R)} \right) + \log \text{ etc}$

$= 2 + 3 + 1 - 2 + \text{ etc}$
BIT scores and E values
Meaning of Bit Score

- **Bit Score**
  - Raw scores have no meaning without knowledge about the scoring scheme used, it's like citing a distance without specifying feet, meters, miles, light years etc.
  - Raw scores are normalized to get Bit scores by incorporating information about the scoring scheme used
  - On average you get 1-2 bits per match
  - Bit score represents Scores in sequence comparisons that is independent of the size of the database
  - So what about E values?
Expect values an example with dice

Suppose you have rolled a die 6 times and the scores are kept in a database.

What is the probability that there is a 3 in this database?

The probability of at least one 3 in 6 rolls is:

$$1 - (5/6)^6 = 1 - 0.3349 = 0.6651$$

Expect value: when you do 60 rolls and you have fair die you expect to roll 3 about 10 times or once per 6 rolls.

The probability $P$ of at least one 3 in 60 rolls is:

$$1 - (1/e^{10}) = 0.99$$
E values an example with DNA sequences

consider the sequence

GATC

We expect to find this sequence randomly once per 256 bases

The odds are $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$

or $1$ out of $2^8$
BIT scores

consider the sequence

GATC

E = 1 per $2^8$

Suppose we compare this sequence (GATC) with a database of one sequence of 4099 bp. We make a sliding window of 4

GGATTATTCGATCGGAATTT....... . .

GATC

GATC

GATC

GATC
With word length of 4 (κ-tuple of 4)
It is possible to make
4099-κ+1
4096 or $2^{12}$ pairwise comparisons
GGATTATTCTGATCGGAATTT...........
GATC
GATC
GATC
GATC
If we give a value of
2 bits to a matching pair
this means that for the sequence GATC we CAN NOW SUM the SCORES
a maximum score of $2+2+2+2 = 8$ bits
( a normalised score of 2 bits per match)
The size (search space) of this database is 12 bits ($2^{12}$ pairwise comparisons)
We expect to find this $2(12-8)=2^4=16$ times
or $4096/256=16$
The probability P of finding GATC once in this database is however almost 1
Does this annotation work (in an intuitive sense)?

What if the search space (sequence length) is doubled?
What if there are two equal sequences in the search space?

\[
(seq1 = 4099-4+1) + (seq2 = 4099-4+1) = \ln(4096 + 4096) = 13 \text{ bit}\]

\[
\ln 2
\]

13-8=5 \text{ E value=32}

The chance of finding GATC once in this search space is again almost 1

But you see and end of sequence effect
What if we search for GAATTCGATCAAGCTT

The word length or κ-tuple = 16
The chance of finding this sequence randomly is \((1/4)^{16}\)
or in bits \(2 \times 16 = 32\)

The search space in this example now is (with a window of 16)
\(\ln((4099-16+1) = \text{slightly less than 12 bit}\)

The E (expect) value is now
is \(2^{(12-32)} = 2^{-18} = 0,0000038\)

The probability that we find this motif in a database of this size is
\(P = 1-\left(\frac{1}{e^{0,0000038}}\right) = 0,0000038\)
The Statistics of Sequence Similarity Scores

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone.

In other words the E-value gives a measure of the amount of expected “background noise” in the database. Small E-values are similar to P-values. Larger alignments result in exponentially smaller E-values.
What is officially the Expect (E) value?

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. It decreases exponentially with the Score (S) that is assigned to a match between two sequences. Essentially, the E value describes the random background noise that exists for matches between sequences.
Meaning of E value in Database Searches in simple terms

- **E- Value** =

- The number of equal or higher scores expected at random for a given High Scoring Pair (HSP).

- E value of 10 for a match means, in a database of current size, one might expect to see 10 matches with a similar or better score, simply by chance alone.

- E- value is the **most commonly used threshold** in database searches. Only those hits with E-values **smaller** than the set threshold will be **reported in the output**

- Increasing E value enable you to see biologically significant, but statistically insignificant matches

- E - value ranges between 0 to any value
• P-Value (Probability value)

  • The probability of achieving an equal or higher score at random for a HSP
  
  • \( P = 1-e^{-E} \)

  • A P value of 0.03 means that, there is a 3% chance of obtaining an equal or higher score at random for a particular HSP

  • When \( E < 0.01 \), P-values and E-values are nearly identical

  • P values range from 0 to 1
In the limit of sufficiently large sequence lengths $m$ and $n$, the statistics of HSP scores are characterised by two parameters, $K$ and $\lambda$.

Most simply, the expected number of HSPs high scoring pairs with score at least $S$ is given by the formula

$$E = Kmn e^{-\lambda S}$$

We call this the $E$-value for the score $S$. 
BIT scores
Lambda:
Lambda is an experimentally determined parameter describes the increase in information of added amino acids pair function: to normalise bit scores: a different matrix requires a different lambda

Gapped blast-scores: lambda changes

<table>
<thead>
<tr>
<th>Matrix</th>
<th>GOP</th>
<th>GEP</th>
<th>lambda ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOSUM62</td>
<td>11</td>
<td>1</td>
<td>0.85</td>
</tr>
<tr>
<td>BLOSUM62</td>
<td>7</td>
<td>2</td>
<td>0.77</td>
</tr>
<tr>
<td>BLOSUM80</td>
<td>11</td>
<td>1</td>
<td>0.92</td>
</tr>
</tbody>
</table>
INTERPRETING OUTPUT

Scores

For the list of sequences, the High Score is the score of the highest-scoring segment pair for the pairwise comparison of that database sequence and your query.

For the segment pair display, the Score is the sum of the scoring matrix values in the segment pair being displayed.
There is a probability, for instance $2.7e-160$ in the example, associated with each pairwise comparison in the list and with each segment pair alignment. For the list, this number is the probability that by chance you would observe a score or group of scores as high as the observed high score (or scores) when you do a search of the same size.
Bit Score

Each aligned segment pair has a normalised score expressed in bits, the bit score.

It determines the size of the search space you would have to look through to find an **High Scoring Pair** (HSP or part of a sequence) score as good as this one by chance.

If the bit score is 30, you have to look at about 1 billion independent segment pairs to find a score this good by chance alone.

Each additional bit doubles the size of the search space.
Bit scores represent a probability level for sequence comparisons that is independent of the size of the search.

The size of the search space is proportional to the product of the query sequence length (e.g. 235) x the sum of the lengths of the sequences in the database. Swiss Prot = 12,496,420

This product is multiplied by a coefficient $K$ to get the size of the search space.

When searching protein databases with protein queries, $K$ is about 0.13. This product for the search is for example $x 12,496,420 x 0.13$ or about 0.38 billion,

so a bit score of 30 corresponding to a space of 1 billion could easily have occurred by chance alone.
In the limit of sufficiently large sequence lengths $m$ and $n$, the statistics of HSP scores are characterised by two parameters, $K$ and $\lambda$.

Most simply, the expected number of high scoring pairs (HSPs) which score at least $S$ is given by the formula

$$E = Kmn e^{-\lambda S}$$

We call this the $E$-value for the score $S$. 
The End
<table>
<thead>
<tr>
<th>PAM distance</th>
<th>Percent conserved</th>
<th>Bits/Unit score</th>
<th>Average information per position (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>99.7</td>
<td>1.992</td>
<td>1.97</td>
</tr>
<tr>
<td>5.3</td>
<td>94.9</td>
<td>0.968</td>
<td>1.63</td>
</tr>
<tr>
<td>16.0</td>
<td>85.6</td>
<td>0.595</td>
<td>1.18</td>
</tr>
<tr>
<td>30.2</td>
<td>75.0</td>
<td>0.396</td>
<td>0.79</td>
</tr>
<tr>
<td>47.0</td>
<td>65.1</td>
<td>0.275</td>
<td>0.51</td>
</tr>
<tr>
<td>65.0</td>
<td>56.5</td>
<td>0.196</td>
<td>0.32</td>
</tr>
<tr>
<td>86.0</td>
<td>48.8</td>
<td>0.138</td>
<td>0.19</td>
</tr>
<tr>
<td>109.0</td>
<td>42.5</td>
<td>0.096</td>
<td>0.11</td>
</tr>
</tbody>
</table>
BLASTP, Protein Query Searching a Protein Database
Each database sequence is compared to the query in a separate protein-protein pairwise comparison.

BLASTX, Nucleotide Query Searching a Protein Database
The query is translated, and each of the six products is compared to each database sequence in a separate protein-protein pairwise comparison.

BLASTN, Nucleotide Query Searching a Nucleotide Database
Each database sequence is compared to the query in a separate nucleotide-nucleotide pairwise comparison.
**TBLASTN, Protein Query Searching a Nucleotide Database**

Each nucleotide database sequence is translated, and each of the six products is compared to the query in a separate protein-protein pairwise comparison.

**TBLASTX, Nucleotide Query Searching a Nucleotide Database**

The query and each database sequence are both translated in six frames, and each of the 12 products is compared in 36 different pairwise comparisons. This program is compute-intensive and is therefore limited to searches of Alu, STS, and EST sequences.