3.3 Database searching for similar sequences

3.3.1 Importance of database searches for similar sequences

“The massive quantities of data generated by genomic research ... We are swimming in a rapidly rising sea of data...how do we keep from drowning?”

Simian sarcoma virus (SSV) oncogene, v-sis

- platelet-derived growth factor (PDG)
- 31AA (26 similar) / 39AA (35 similar)
- It appeared likely that the retrovirus had acquired the gene from the host cell as some kind of genetic exchange event and then had produced a mutant form of the protein that could compromise the function of the normal protein when the virus infected another animal
Biological question and context of the search

- Query sequence

- DB of sequences

Comparison

- Program Parameters (scoring system)

Observed Similarities (alignments, scores)

- Biological

Interpretation

- Statistical

Biological Inferences

(structural, functional, evolutionary features)
Uses of Similarity Searches

- Find orthologous proteins sequences in different organisms (ortho = same function)
- Find paralogous sequences in same organism (para = originate from gene duplication)
- Align genomic sequences and expressed gene sequences (cDNAs, ESTs) to locate gene and ORFs in gene.
Difficulties in Similarity Searching and Solutions

- Slowness of dynamic programming alignment algorithm - use a word search, heuristic method (nearly always works but is not a guaranteed algorithm).
- Low complexity regions in sequences give false high scores - filter out these regions in the query sequence.
- How good is the match? - Produce an optimal local alignment.
- How significant is the alignment score? - Provide an expect score (E value) for finding a match with an unrelated sequence in searching a sequence database of the same size.

Mount D, 2002
3.3.2 Word or $k$-tuple methods

- Used by the FASTA and BLAST programs
- They align two sequences very quickly, by first searching for identical short stretches of sequences (called words or $k$-tuples) and by then joining these words into an alignment by the dynamic programming method.
- They are heuristic methods, reliable in a statistical sense, and usually provide a reliable alignment.
- Heuristic methods: an empirical method of computer programming in which rules of thumb are used to find solutions and feedback is used to improve performance.
Heuristic method for sequence alignment is used to try to find a reasonable alignment. An example of a heuristic alignment method is starting with a search for similar sequence patterns and then building a longer alignment based on these patterns.

Rather than comparing individual residues in the two sequences, FASTA and BLAST search for matching sequence patterns or words (Wilbur and Lipman, 1983). The programs then attempts to build a local alignment based on these word matches.

Methods of Similarity Searching

• FASTA (fast alignment) and BLAST (basic local alignment search tool) algorithms
• PSI-BLAST (position-specific-iterated) and PHI-BLAST(pattern-hit-iterated)
• MAST and MetaMEME on MEME server: search with PSSMs or HMMs representing domains
FASTA & BLAST story

1985 : FASTP (D. Lipman and W. Pearson)
   Global gapped alignments

1988 : FASTA (W. Pearson and D. Lipman)
   Local gapped alignments

1990 : BLAST1
   (S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman)
   Local ungapped alignments

Gapped BLASTs :

1996: WU–BLAST2 (W. Gish)

1997: NCBI–BLAST2 (and PSI–BLAST)
   (S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang,
   W. Miller and D. Lipman)
FASTA Algorithm

$k$-tuples 1-2 for amino acid and 4-6 for nucleotide
BLAST

- BLAST (Basic Local Alignment Search Tool) (Altschul et al 1990), a heuristic method designed to detect alignments that optimize a measure of local similarity known as the maximal segment pair (MSP) score.
- Word length: 11 (DNA sequences) and 3 (protein)
- Word index
- High-scoring segment pairs (HSP)
First step:

For each position $p$ of the query, find the list of words of length $w$ scoring more than $T$ when paired with the word starting at $p$:

List of words of length $w$, scoring more than $T$ with the $p$-word.
Second step:

For each words list, identify all exact matches with DB sequences:

\[\text{p-word} \rightarrow \text{words list} \rightarrow \text{DB sequences}\]
Third step:
For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.

HSP = High Scoring Segment Pair
MSP = Maximal Segment Pair

Reports all HSPs having score S above a threshold, or equivalently, having E-value below a threshold.

E-value = the number of HSPs having score S (or higher) expected to occur only by chance.

Apply sum-statistics to evaluate the significance of a combination of HSPs involving the same DB sequence.
Ungapped extension of hits
Ungapped and gapped extensions

Ungapped extension

Gapped extension by full DP

Gapped extension by «banded DP»

Gapped extension by «score-limited DP»
BLASTall

Query Sequence

Amino acid Sequence

BLASTp  tBLASTn

BLASTp

Protein Database

DNA Sequence

BLASTn  BLASTx  tBLASTx

DNA Sequence

Nucleotide Database

Translated

Translated

megasBLAST

......
Web-based BLAST Input
Web-based BLAST Output

BLASTN 2.2.4 [Aug-26-2002]

Reference:

RID: 1031981000-015047-25897

Query= cluster8190_1
(1447 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)
1,367,736 sequences; 6,442,224,816 total letters
Web-based BLAST Output

ii. Graphical overview of database matches
### Web-based BLAST Output

iii. Descriptions of each hit (alignment)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>21314556</td>
<td>gb</td>
<td>AF513016.1</td>
</tr>
<tr>
<td>gi</td>
<td>21261723</td>
<td>emb</td>
<td>AJ487671.1</td>
</tr>
<tr>
<td>gi</td>
<td>17226389</td>
<td>gb</td>
<td>AF440218.1</td>
</tr>
<tr>
<td>gi</td>
<td>22061915</td>
<td>ref</td>
<td>XM_027060.4</td>
</tr>
<tr>
<td>gi</td>
<td>21411328</td>
<td>gb</td>
<td>BC031006.1</td>
</tr>
<tr>
<td>gi</td>
<td>21410232</td>
<td>gb</td>
<td>BC031008.1</td>
</tr>
<tr>
<td>gi</td>
<td>16198354</td>
<td>gb</td>
<td>BC015821.1</td>
</tr>
<tr>
<td>gi</td>
<td>516580</td>
<td>gb</td>
<td>S69022.1</td>
</tr>
<tr>
<td>gi</td>
<td>4557774</td>
<td>ref</td>
<td>NM_000432.1</td>
</tr>
<tr>
<td>gi</td>
<td>34845</td>
<td>emb</td>
<td>X66141.1</td>
</tr>
<tr>
<td>gi</td>
<td>2460246</td>
<td>gb</td>
<td>AF020768.1</td>
</tr>
<tr>
<td>gi</td>
<td>1220300</td>
<td>gb</td>
<td>M22815.1</td>
</tr>
<tr>
<td>gi</td>
<td>34686</td>
<td>emb</td>
<td>X14332.1</td>
</tr>
<tr>
<td>gi</td>
<td>56682</td>
<td>emb</td>
<td>X07314.1</td>
</tr>
<tr>
<td>gi</td>
<td>20841890</td>
<td>ref</td>
<td>XM_132324.1</td>
</tr>
<tr>
<td>gi</td>
<td>6754777</td>
<td>ref</td>
<td>NM_010861.1</td>
</tr>
<tr>
<td>gi</td>
<td>12832295</td>
<td>dbj</td>
<td>AK002367.1</td>
</tr>
<tr>
<td>gi</td>
<td>188984</td>
<td>gb</td>
<td>M91602.1</td>
</tr>
</tbody>
</table>
Web-based BLAST Output

iv. Pairwise alignments
### Web-based BLAST Output

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)

- **Posted date:** Sep 4, 2002 12:27 AM
- **Number of letters in database:** 2,147,257,520
- **Number of sequences in database:** 1,367,736

<table>
<thead>
<tr>
<th>Lambda</th>
<th>K</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.37</td>
<td>0.711</td>
<td>1.31</td>
</tr>
</tbody>
</table>

- **Matrix:** blastn matrix: 1 -3
- **Gap Penalties:** Existence: 5, Extension: 2
- **Number of Hits to DB:** 3,320,183
- **Number of Sequences:** 1367736
- **Number of extensions:** 3320183
- **Number of successful extensions:** 18698
- **Number of sequences better than 10.0:** 207
- **Length of query:** 1447
- **Length of database:** 6,442,224,816
- **Effective HSP length:** 22
- **Effective length of query:** 1425
- **Effective length of database:** 6,412,134,624
- **Effective search space:** 9137291839200
- **Effective search space used:** 9137291839200

- **T:** 0
- **A:** 30
- **X1:** 6 (11.9 bits)
- **X2:** 15 (29.7 bits)
- **S1:** 12 (24.3 bits)
- **S2:** 20 (40.1 bits)

v. Statistical details of the search
Local BLAST

On Windows PC:

(1) Download from NCBI and install

(2) Configuration: using the ncbi.ini file

[NCBI]

Data="C:\path\data\"

(3) Create or format your databases
Difference between FASTA and BLAST

- Whereas FASTA searches for all possible words of the same length, BLAST confines the search to the words that are the most significant.
- For the BLAST algorithm, the word length is fixed at 3 (formerly 4) for proteins and 11 for nucleic acids (3 if the sequences are translated in all six reading frames). This length is the minimum needed to achieve a word score that is high enough to be significant but not so long as to miss short but significant patterns.
- FASTA theoretically provides a more sensitive search of DNA sequence databases because a shorter word length may be used.
Beyond sequences: database searching using scoring matrices, profiles and PSSMs

Previously used a single sequence to search a sequence database to locate similar sequences

Now interested in searching a database with a representation of a local multiple sequence alignment — a PSSM (no gaps), profile (gaps) or HMM (gaps)

Can also search a sequence for matches to a database of models in the form of PSSMs, profiles or HMMs (e.g. the PFAM database)

- PSSM: position-specific scoring matrix; HMM: Hidden Markov Models
How do PSSMs etc. find additional matches than searching with single sequences?

If search with a set of single query sequences, then matches must align with query with Blosum62 substitutions at each position.

If using a PSSM, profile of HMM then can also match combinations of amino acids.

<table>
<thead>
<tr>
<th>Blast search</th>
<th>PSSM or HMM search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query</strong></td>
<td><strong>Query</strong></td>
</tr>
<tr>
<td>..GYD...</td>
<td>..GYD...</td>
</tr>
<tr>
<td><strong>Database sequences</strong></td>
<td><strong>Database sequences</strong></td>
</tr>
<tr>
<td>..GYD...</td>
<td>..GYD...</td>
</tr>
<tr>
<td>..GWD...</td>
<td>..SYE...</td>
</tr>
<tr>
<td>matched</td>
<td>matched</td>
</tr>
<tr>
<td>..GWE...</td>
<td>..SWE...</td>
</tr>
<tr>
<td>(Blosum62 substitutions)</td>
<td>(PSSM or HMM substitutions)</td>
</tr>
</tbody>
</table>
Searching a database with a PSSM or HMM

- Recall that the PSSM is a matrix of scores with the columns representing positions in a local multiple sequence alignment and the rows the 20 amino acids.

- Similarly, an HMM is a probabilistic model of a multiple sequence alignment.
Other tools: PSI-BLAST

- PSI–Blast is position-specific-iterated-blast
- Generates a PSSM from the matched sequences in a BLAST search – it uses the entire sequence

Query sequence

Blast database

Matched sequences

Add more sequences to PSSM

PSSM

Caution! Can switch domains
PHI-BLAST: regular expressions

Regular expressions provide a wide range of pattern finding options e.g. [LIVMF]-G-E-x (5,11)-[IV] implemented in the prosite catalog and used in phi-blast. Prosite entry for ATP/GTP binding site motif A (P loop) is [AG] – x(4)-G-K-[ST].

In phi-blast, a regular expression is provided – database is searched for sequences with pattern, and then psi-blast used to expand search.
3.3.4 Assessing the significance of sequence alignments

Global alignments:
the distribution is not known

Local alignments without gaps:
theoretical work: Karlin–Altschul statistics
  ➡️ Extreme–value distribution

Local alignments with gaps:
empirical studies
  ➡️ Extreme–value distribution
The Gumbel Extreme Value Distribution: distribution of scores expected by aligning two random sequences of the same length and composition follow a distribution called the extreme value distribution (Gumbel EJ, 1962).
## A. Score distribution

<table>
<thead>
<tr>
<th>z'</th>
<th>opt</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>176</td>
<td>0:***</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0:=</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>0:=</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>2:*</td>
</tr>
<tr>
<td>28</td>
<td>21</td>
<td>17:*</td>
</tr>
<tr>
<td>30</td>
<td>109</td>
<td>102:*</td>
</tr>
<tr>
<td>32</td>
<td>355</td>
<td>393:***</td>
</tr>
<tr>
<td>34</td>
<td>1053</td>
<td>1067:======*</td>
</tr>
<tr>
<td>36</td>
<td>2214</td>
<td>2191:=============*</td>
</tr>
<tr>
<td>38</td>
<td>4004</td>
<td>3620:=============*==</td>
</tr>
<tr>
<td>40</td>
<td>5285</td>
<td>5050:=============*==</td>
</tr>
<tr>
<td>42</td>
<td>6350</td>
<td>6173:=============*==</td>
</tr>
<tr>
<td>44</td>
<td>6716</td>
<td>6809:=============*==</td>
</tr>
<tr>
<td>46</td>
<td>6847</td>
<td>6935:=============*==</td>
</tr>
<tr>
<td>48</td>
<td>6661</td>
<td>6640:=============*==</td>
</tr>
<tr>
<td>50</td>
<td>5777</td>
<td>6059:=============*</td>
</tr>
<tr>
<td>52</td>
<td>5015</td>
<td>5327:=============*</td>
</tr>
<tr>
<td>54</td>
<td>4344</td>
<td>4550:=============*</td>
</tr>
<tr>
<td>56</td>
<td>3772</td>
<td>3801:=============*</td>
</tr>
<tr>
<td>58</td>
<td>3025</td>
<td>3120:=============*</td>
</tr>
<tr>
<td>60</td>
<td>2475</td>
<td>2528:=============*</td>
</tr>
<tr>
<td>62</td>
<td>1968</td>
<td>2026:=============*</td>
</tr>
<tr>
<td>64</td>
<td>1607</td>
<td>1612:=============*</td>
</tr>
<tr>
<td>66</td>
<td>1362</td>
<td>1274:=============*</td>
</tr>
<tr>
<td>68</td>
<td>983</td>
<td>1002:=============*</td>
</tr>
<tr>
<td>70</td>
<td>823</td>
<td>785:=====*</td>
</tr>
<tr>
<td>72</td>
<td>597</td>
<td>614:=====*</td>
</tr>
<tr>
<td>74</td>
<td>584</td>
<td>478:=====*</td>
</tr>
<tr>
<td>76</td>
<td>456</td>
<td>372:====*</td>
</tr>
<tr>
<td>78</td>
<td>306</td>
<td>289:====*</td>
</tr>
<tr>
<td>80</td>
<td>238</td>
<td>225:====*</td>
</tr>
<tr>
<td>82</td>
<td>230</td>
<td>172:====*</td>
</tr>
<tr>
<td>84</td>
<td>141</td>
<td>136:====*</td>
</tr>
<tr>
<td>86</td>
<td>131</td>
<td>105:====*</td>
</tr>
<tr>
<td>88</td>
<td>93</td>
<td>82:====*</td>
</tr>
<tr>
<td>90</td>
<td>52</td>
<td>63:==</td>
</tr>
<tr>
<td>92</td>
<td>55</td>
<td>49:==</td>
</tr>
<tr>
<td>94</td>
<td>41</td>
<td>38:==</td>
</tr>
<tr>
<td>96</td>
<td>37</td>
<td>29:==</td>
</tr>
<tr>
<td>98</td>
<td>20</td>
<td>23:==</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>17:==</td>
</tr>
<tr>
<td>102</td>
<td>16</td>
<td>14:==</td>
</tr>
<tr>
<td>104</td>
<td>7</td>
<td>10:==</td>
</tr>
<tr>
<td>106</td>
<td>9</td>
<td>8:==</td>
</tr>
<tr>
<td>108</td>
<td>7</td>
<td>6:==</td>
</tr>
<tr>
<td>110</td>
<td>3</td>
<td>5:==</td>
</tr>
<tr>
<td>112</td>
<td>6</td>
<td>4:==</td>
</tr>
<tr>
<td>114</td>
<td>2</td>
<td>3:==</td>
</tr>
<tr>
<td>116</td>
<td>4</td>
<td>2:==</td>
</tr>
<tr>
<td>118</td>
<td></td>
<td>1:==</td>
</tr>
</tbody>
</table>

Inset = represents 2 library sequences
Figure 3.17. Probability values for the extreme value distribution (A) and the normal distribution (B). The area under each curve is 1.
The extreme value distribution

\[ Y_{ev} = \exp \left[ -x - e^{-x} \right] \]

The normal distribution

\[ Y_n = \frac{1}{\sqrt{2\pi}} \exp \left[ -x^2 / 2 \right] \]
The probability that score \( S \) will be less than value \( x \), 
\[ P(S < x) \]
, is obtained by calculating the area under curve from \( -\infty \) to \( x \)

\[
P(S < x) = \exp[-e^{-x}]
\]

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

**P-value:** probability that there is at least one random MSP having score \( S \) or greater
E-value: the number of HSPs having score $S$ (or higher) expected to occur only by chance

$E$-value is given by the formula:

$$E = K m n e^{-\lambda x}$$

$$P(S' \geq x) = 1 - \exp E$$

$$P(S' \geq x) \approx E$$

$$P(S \geq x) = 1 - \exp[-e^{-x}] \rightarrow e^{-x}$$
The Gumbel extreme value distribution

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

\[ \lambda = \pi(\sigma \sqrt{6}) \]

\[ \mu = \frac{x - \gamma}{\pi} = x - 0.450\sigma \]

\[ = (\ln Kmn)/\lambda \]

\[ P(s \geq x) \approx Kmn e^{-\lambda s} \]

\[ e^{y} \approx e^{-x} \]

\[ P(S \geq x) = 1 - \exp[-e^{-x}] \rightarrow e^{-x} \]
\[ P(S \geq x) = 1 - \exp[-e^{-x}] \rightarrow e^{-x} \]

\[ P(S \geq x) \approx e^{-x} \]

\[ P(S' \geq x) \approx Kmne^{-\lambda x} \]

\[ P(S' \geq x) \approx E \]
By normalizing a raw score using the formula

\[ S' = \lambda S - \ln Kmn \]
Need of empirical simulations of the random distribution in order to calculate its parameters

- **Blast2**: artificial random sequences
- **Fasta**: uses results from the search: real unrelated sequences
BLASTP 2.0.5 [May-5-1998]
Query= human XP-F repair gene  (905 letters)

Database: Non-redundant SwissProt sequences 74,596 sequences; 26,848,718 total letters

B.

Color Key for Alignment Scores

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>Black</td>
</tr>
<tr>
<td>40-50</td>
<td>Blue</td>
</tr>
<tr>
<td>50-80</td>
<td>Green</td>
</tr>
<tr>
<td>80-200</td>
<td>Red</td>
</tr>
<tr>
<td>&gt;=200</td>
<td>Pink</td>
</tr>
</tbody>
</table>

Distribution of 11 BLAST Hits on the Query Sequence

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp</td>
<td>Q92889</td>
<td>XPF_HUMAN DNA-REPAIR PROTEIN COMPLEMENTING XP-F CELL ...</td>
<td>1659</td>
</tr>
<tr>
<td>sp</td>
<td>P36617</td>
<td>RA16_SCHPO DNA REPAIR PROTEIN RAD16</td>
<td>485</td>
</tr>
<tr>
<td>sp</td>
<td>P06777</td>
<td>RAD1_YEAST DNA REPAIR PROTEIN RAD1</td>
<td>231</td>
</tr>
<tr>
<td>sp</td>
<td>P40562</td>
<td>YIS2_YEAST PUTATIVE ATP-DEPENDENT RNA HELICASE YIR002C</td>
<td>37</td>
</tr>
<tr>
<td>sp</td>
<td>Q10202</td>
<td>YAXB_SCHPO PUTATIVE ATP-DEPENDENT RNA HELICASE C13F4.11C</td>
<td>36</td>
</tr>
</tbody>
</table>
Database: Non-redundant SwissProt sequences
Number of letters in database: 26,848,718
Number of sequences in database: 74,596

Lambda K H
0.320 0.136 0.394

Gapped
Lambda K H
0.270 0.0470 0.230

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1

Number of Hits to DB: 42777291
Number of Sequences: 74596
Number of extensions: 1706128
Number of successful extensions: 4638
Number of sequences better than 10.0: 12
Number of HSP's better than 10.0 without gapping: 4
Number of HSP's successfully gapped in prelim test: 8
Number of HSP's that attempted gapping in prelim test: 4616
Number of HSP's gapped (non-prelim): 16

length of query: 905
length of database: 26848718
effective HSP length: 55
effective length of query: 850
effective length of database: 22745938
effective search space: 19334047300
effective search space used: 19334047300

T: 11
A: 40
X1: 16 (7.4 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 41 (21.8 bits)
S2: 68 (30.9 bits)
FASTA (3.14 April, 1998) function (optimized, BL50 matrix) ktup: 2
  join: 39, opt: 27, gap-pen: -12/ -2, width: 16 reg.-scaled

<table>
<thead>
<tr>
<th>The best scores are:</th>
<th>initn</th>
<th>init1</th>
<th>opt</th>
<th>z-sc</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPF_HUMAN</td>
<td>11/97</td>
<td>( 905)</td>
<td>5893</td>
<td>5893</td>
<td>6529.7</td>
</tr>
<tr>
<td>RA16_SCHPO</td>
<td>11/97</td>
<td>( 892)</td>
<td>1569</td>
<td>519</td>
<td>749</td>
</tr>
<tr>
<td>RAD1_YEAST</td>
<td>11/97</td>
<td>(1100)</td>
<td>975</td>
<td>362</td>
<td>619</td>
</tr>
<tr>
<td>YIS2_YEAST</td>
<td>11/95</td>
<td>( 993)</td>
<td>37</td>
<td>37</td>
<td>161</td>
</tr>
<tr>
<td>YAXB_SCHPO</td>
<td>10/96</td>
<td>( 578)</td>
<td>91</td>
<td>91</td>
<td>133</td>
</tr>
</tbody>
</table>