Bioinformatics

Sequence Analysis: Part II. Database Searching

Fran Lewitter, Ph.D.
Head, Biocomputing
Whitehead Institute
Course Syllabus

Jan 7  Sequence Analysis I. Pairwise alignments, database searching including BLAST (FL) [1, 2, 3]
Jan 14 Sequence Analysis II. Database searching (continued), Pattern searching (FL)[7]
Jan 21 No Class - Martin Luther King Holiday
Jan 28 Sequence Analysis III. Hidden Markov models, gene finding algorithms (FL)[8]

Feb 4  Computational Methods I. Genomic Resources and Unix (GB)
Feb 11 Computational Methods II. Sequence analysis with Perl. (GB)
Feb 18 No Class - President's Birthday
Feb 25 Computational Methods III. Sequence analysis with Perl and BioPerl (GB)

Mar 4  Proteins I. Multiple sequence alignments, phylogenetic trees (RL) [4, 6]
Mar 11 Proteins II. Profile searches of databases, revealing protein motifs (RL) [9]
Mar 18 Proteins III. Structural Genomics: structural comparisons and predictions (RL)
Mar 25 Microarrays: designing chips, clustering methods (FL)
Topics to Cover

• Introduction
• Scoring alignments
• Alignment methods
  – Dot matrix analysis
  – Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  – Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST)
• Significance of alignments
• Database searching methods
• Demo
Questions

• Why do a database search?
• What database should be searched?
• What alignment algorithm to use?
• What do the results mean?
Issues affecting DB Search

• Substitution matrices
• Statistical significance
• Filtering
• Database choices
BLASTP Results

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>gi</th>
<th>ref</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi 17647947</td>
<td>ref NP_523565.1</td>
<td>(NM_078841) spellcheckerl [Dro...</td>
<td>613</td>
</tr>
<tr>
<td>gi 17508445</td>
<td>ref NP 491202.1</td>
<td>(NM_058801) DNA mismatch repair...</td>
<td>392</td>
</tr>
<tr>
<td>gi 17462805</td>
<td>ref XP_003992.5</td>
<td>(XM_003992) similar to mutS (E...</td>
<td>230</td>
</tr>
<tr>
<td>gi 17545870</td>
<td>ref NP 519272.1</td>
<td>(NC_003295) PROBABLE DNA MISMA...</td>
<td>223</td>
</tr>
<tr>
<td>gi 17934261</td>
<td>ref NP 531051.1</td>
<td>(NC_003304) DNA mismatch repair...</td>
<td>218</td>
</tr>
<tr>
<td>gi 17988084</td>
<td>ref NP 540718.1</td>
<td>(NC_003317) DNA MISMATCH REPAIR...</td>
<td>211</td>
</tr>
<tr>
<td>gi 17508447</td>
<td>ref NP 491163.1</td>
<td>(NM_058762) DNA mismatch repair...</td>
<td>189</td>
</tr>
<tr>
<td>gi 17534743</td>
<td>ref NP 495451.1</td>
<td>(NM_063050) recombination prote...</td>
<td>159</td>
</tr>
<tr>
<td>gi 17539736</td>
<td>ref NP 502531.1</td>
<td>(NM_070130) Yeast DNA mismatch...</td>
<td>152</td>
</tr>
<tr>
<td>gi 13182948</td>
<td>gb</td>
<td>AAK15001.1</td>
<td>AF233343_4</td>
</tr>
<tr>
<td>gi 17535283</td>
<td>ref NP 495458.1</td>
<td>(NM_063057) T02G5.6.p [Caenorr...</td>
<td>70</td>
</tr>
<tr>
<td>gi 17485036</td>
<td>ref XP_037759.2</td>
<td>(XM_037759) KIAA0376 protein [...</td>
<td>36</td>
</tr>
<tr>
<td>gi 17933948</td>
<td>ref NP 530738.1</td>
<td>(NC_003304) ATP-dependant DNA ...</td>
<td>32</td>
</tr>
</tbody>
</table>

Alignments

>gi|17647947|ref|NP_523565.1| (NM_078841) spellcheckerl [Drosophila melanogaster]
   gi|17298155|gb|AAF53392.1| (AE003643) spell gene product [Drosophila melanogaster]

Length = 854

Score = 613 bits (1582), Expect = e-175
Identities = 352/854 (41%), Positives = 508/854 (59%), Gaps = 86/854 (10%)
Topics to Cover

• Introduction
• Scoring alignments
• Alignment methods
  – Dot matrix analysis
  – Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  – Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST)
• Significance of alignments
• Database searching methods
• Demo
Significance of Alignment

How strong can an alignment be expected by chance alone?

• Real but non-homologous sequences
• Real sequences that are shuffled to preserve compositional properties
• Sequences that are generated randomly based upon a DNA or protein sequence model
Extreme Value Distribution

- When 2 sequences have been aligned optimally, the significance of a local alignment score can be tested on the basis of the distribution of scores expected by aligning two random sequences of the same length and composition as the two test sequences.
Statistical Significance

- Raw Scores - score of an alignment equal to the sum of substitution and gap scores.

- Bit scores - scaled version of an alignment’s raw score that accounts for the statistical properties of the scoring system used.

- E-value - expected number of distinct alignments that would achieve a given score by chance. Lower E-value => more significant.
Some formulas

\[ E = Kmn e^{-S} \]

This is the Expected number of high-scoring segment pairs (HSPs) with score at least \( S \) for sequences of length \( m \) and \( n \).

This is the E value for the score \( S \).
Topics to Cover

• Introduction
• Scoring alignments
• Alignment methods
• Significance of alignments
• Database searching methods
  – BLAST - ungapped and gapped
  – BLAST vs. FASTA
  – PSI-BLAST
  – PHI-BLAST
  – Pattern searching
• Demo
Low Complexity Regions

- Local regions of biased composition
- Common in real sequences
- Generate false positives on BLAST search

- DUST for BLASTN (n’s in sequence)
- SEG for other programs (x’s in sequence)

Filtering is only applied to the query sequence (or its translation products), not to database sequences.
Filtered Sequence

> HUMAN MSH2

NEEYTKNKTEYEE

TALTTEETLTLT

WIBR Bioinformatics Course. © Whitehead Institute, 2002
Example Alignment w/o filtering

Score = 29.6 bits (65), Expect = 1.8
Identities = 22/70 (31%), Positives = 32/70 (45%), Gaps = 12/70 (17%)

Query: 31   PPPTQGAPRTSSFTPTTLT-------------NGTSHSPTALNGAPSPPNGFS 71
            PPP+   Q   R   S   +   T   T           NG+S   S   ++   +   S   +   S
Sbjct: 1221 PPPSVQNQQRWGSSSVITTTTCQQRQSQVSPHSNGSSSSSSSSSSSSSSSSSSSSST 1273

Query: 72   NGPSSSSSSSLANQQLP 88
            +   SSSS+SS    Q   P
Sbjct: 1274 SNCSSSSASSCQYFQSP 1290
Example BLAST w/ filtering

Score = 36.6 bits (83), Expect = 0.67
Identities = 21/58 (36%), Positives = 25/58 (42%), Gaps = 1/58 (1%)

Query: 471 AEDALAVINQQEDSSESNCGRKASETCSGCNTARYCGSFCQH KDWE-K HHHICGQT 527
       A D  V  Q  +  +  C  CG  A  TCS  C  A  YC  Q  DW  H  C  Q+
Sbjct: 61 ASDTECVCLQLKSGAHLCRVCGCLAPMTCSRCKQAHYCSKEHQTLDWQLGHKQACTQS 118

Score = 37.0 bits (84), Expect = 0.55
Identities = 18/55 (32%), Positives = 22/55 (39%)

Query: 483 DSSESCWNCGRKASETCSGCNTARYCGSFCQH KDWKHHHICGQTLQAOQQQGDT P 537
       D  C  CG  A++  C+  C  ARYC  Q  DW  H  C  +  D  P
Sbjct: 75 DGPGLCRICGCSAAKKCAKCQVARYCSQAHQVIDWPAHKLECAKAATDGSITDEP 129
WU-BLAST vs NCBI BLAST

- WU-BLAST first for gapped alignments
- Use different scoring system for gaps
- Report different statistics
- WU-BLAST does not filter low-complexity by default
- WU-BLAST looks for and reports multiple regions of similarity
- Results will be different
BLAT

- Developed by Jim Kent at UCSC
- BLAT is not BLAST
- For DNA it is designed to quickly find sequences of $\geq 95\%$ similarity of length 40 bases or more.
- For proteins it finds sequences of $\geq 80\%$ similarity of length 20 amino acids or more.
- DNA BLAT works by keeping an index of the entire genome in memory - non-overlapping 11-mers ($< 1$ GB of RAM)
- Protein BLAT uses 4-mers ($\sim 2$ GB)
FASTA

- Index "words" and locate identities
- Rescore best 10 regions
- Find optimal subset of initial regions that can be joined to form single alignment
- Align highest scoring sequences using Smith-Waterman
PSI-BLAST

- Position specific iterative BLAST uses a profile (or position specific scoring matrix, PSSM) that is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search.
- The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero.
- The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" is used to refine the profile. This iterative searching strategy results in increased sensitivity.
Start with a BLASTP search

Sequences with E-value BETTER than threshold

<table>
<thead>
<tr>
<th>Score</th>
<th>E (bits)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>236</td>
<td>5e-62</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>4e-16</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>2e-13</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>2e-13</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>2e-13</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>2e-11</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>2e-11</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>3e-11</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>8e-11</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>9e-11</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>1e-10</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>3e-10</td>
<td></td>
</tr>
</tbody>
</table>
PSI-BLAST - Iteration 1

- gi|16078034|ref|NP_388850.1| (NC_000964) similar to hypothetical... 108 2e-23
- gi|16329299|ref|NP_440027.1| (NC_000911) unknown protein [Synech... 107 4e-23
- gi|15228790|ref|NP_191814.1| (NC_003074) putative protein [Arabi... 106 8e-23
- gi|15801753|ref|NP_287771.1| (NC_002655) ynaF gene product [Esch... 106 8e-23
- gi|16129337|ref|NP_415894.1| (NC_000913) putative filament protein... 106 9e-23
- gi|16760235|ref|NP_455852.1| (NC_003198) conserved hypothetical ... 106 1e-22
- gi|15888243|ref|NP_353924.1| (NC_003062) AGR_C_1650p [Agrobacter... 106 1e-22
- gi|15674218|ref|NP_268393.1| (NC_002662) conserved hypothetical ... 106 1e-22
- gi|15807354|ref|NP_296084.1| (NC_001263) conserved hypothetical ... 105 2e-22
- gi|17231624|ref|NP_488172.1| (NC_003272) hypothetical protein [N... 104 2e-22
- gi|2507516|sp|P37903|UP003_ECOLI UNKNOWN PROTEIN 2D_000B3L FROM 2... 103 6e-22
- gi|16081690|ref|NP_394062.1| (NC_002578) conserved hypothetical ... 103 6e-22
### PSSM from PSI-BLAST

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

- N

**Amino acids**
Pattern Hit Initiated (PHI)-BLAST

> HUMAN MSH2

MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTTVRLFDRGDFYTAHGEDALLAAREVFKTQGVIMYMPGA
GAKNLQSVVLSKMNFSFVKDLLLRQYRVEYKRNAGNKASKENDWYLAYKASPGNLSQFEDILFNGND
MSASIGVGVKMSAVDGQROVQVGVYGVYDSIQRKLGLCEFIPDNDQFNSNLEALLIQGKPKECVLPGGETAGDM
GKLRQIIQRGGILITERKKADFSKTDIYQDLNRLKKGKGEQMSAVLPEMENQVAVSSLSAVIKFLELL
SDDSNGFGQFELTTFDSPQYMKLAVAATRNLQFSERQGVEDTTGTSQSLAALLNCKTPQGQLRNQWIKQPL
MDKNIREERNLVFEAVDEAELRQTLQEDDLRRFDPDLNRLAKKFQRAANLQDCYRLQGINQLPNIVQA
LEKHEGKHKQLLLAVFVTPLTDLRDSDFSKFQEMIETTLMDDQVENHEFLVKSDFPNLSELREIMNDLEK
KMQLSTLSAAARDLGLDPGQIKLDSAQFQYFVTCKEEKVLRNKSTVDIQKNGVKNTSNKLTSNL
EEYTKNKTEYEAAQDAIVKEIVNISSGYVEMQTLNDVLAQLDAVVSFAHSNGAPVPYVRPALIEKKGQG
RIILKASRHACEVQDEIAFIPNDVYFEKDKQMFIHTGNPMGKSTYIRQTGVIVLMAQICFCVPCESA
EVSIADVCLARVGAGDSQLKGVSTFMAEMLETASLRSATKDSLIIIIDFGRGTSTYDGFLAWAISEYI
ATKIGAFCMFATHFHELTALONQIPTVNNLTVAILTTELTMLVYDGGVCQDSFGIHVATLANFPKHV
IEACKPKELEEFQYGESQQYDIMEPAAKCYLERFQGEKIIQEFLSKVQMPFTEMSEEINITIKLKQ
LKAEVIAKNNSFVNEIIISRKVT

DNA mismatch repair proteins mutS family signature
>gi|4099512|gb|AAD00647.1| (U87911) MutS homolog 2 [Arabidopsis thaliana]
Length = 117

Score = 136 bits (364), Expect = 1e-40
Identities = 88/117 (75%), Positives = 98/117 (83%)

Query: 668 TGPNMGKSTYRQTGVVLMAQIGCFVPCESAEVSVDCILARVGAGDSQLKVSTFMA 727
       TGPNMGKST+IRG GVIVLMAQ+G FVPC+ A +SI DCI ARVGAGD QL+GVSTFM
Sbjct: 1 TGPNMGKSTFIROVGVVLMAQVGSFVPCDKASISIRDCIFARVGAGDCQLRGVSTFMQ 60

Query: 728 EMLETASILRAATKSLIIIDELGRGRTSTYDGFLAIAISEYIATKIGAFCMFATHF 784
pattern 743 ****************
       EMLETASIL+ AT SLIIIDELGRGRTSTYDGFLAIA E++ A +FATH+
Sbjct: 61 EMLETASILRAATKSLIIIDELGRGRTSTYDGFLAIAICEHLQVVKRAPTLFATHY 117
Basic Searching Strategies

- Search early and often
- Use specialized databases
- Use multiple matrices
- Use filters
- Consider Biology
Demo

- Readseq
- Entrez
- NCBI
  - BLAST2.0
  - PSI-BLAST
  - PHI-BLAST
- WU-BLAST2
- FASTA
- Smith-Waterman
BLAST and FASTA Citations

- Nature Genetics, 1994, 6: 119-129.

http://www.ebi.ac.uk/blast2/
http://www2.ebi.ac.uk/fasta33/
http://www2.ebi.ac.uk/bic_sw/