Heuristic alignment / statistical analysis (2)

Pål Sætrom

Topics
- Introduction
- Heuristic alignment
  - Dot matrices
  - Keyword based searches
  - BLAST
- Statistical Analysis
  - Hypothesis testing
  - Model-based significance
  - Comparisons

Summary – heuristic alignment
- DP-based homology search is too slow for practical applications
- l-mer filtering improves homology search performance

Assessing search result significance
- How to determine if two sequences are homologs?
  - Local alignment score $x$
  - Homologous sequences?
- What is the probability $p$ of having a local alignment score $\geq x$ if the sequences are not homologs?
  - Null hypothesis ($H_0$)
    - Reject $H_0$ if $p$ is low
    - sequences are homologs ($H_1$)
- What is the probability of score $\geq x$ if sequences are random?

Randomization-based hypothesis testing
$P(LA(Q, rand(D)) \geq LA(Q, D))$?
RandomizationTest($Q, D$):
1. $x = LA(Q, D)$
2. $count = 0$
3. $its = 10000$
4. for $i$ in range($its$):
   1. if $LA(Q, rand(D)) \geq x$:
      1. $count += 1$
5. return ($count + 1$) / ($its + 1$)
Distribution-based hypothesis testing
- Randomization requires multiple LA evaluations
- Instead: Use distribution that estimates score distribution
- BLAST (with suitable similarity matrix):
  - Two random sequences
  - Number of matches with scores at least $S'$ is Poisson distributed with parameter $\nu = Kmn e^{-\mu}$
  - $K$ constant, $\mu$ positive root of $1 = \exp(-E(S'))$

BLAST p-values
- Probability of finding at least one segment pair with score greater or equal to $S'$:
  $P(S') = P(Z_S \geq 1) = 1 - e^{-\nu} = 1 - \exp(-Kmn e^{-\mu S})$
- Power series expansion gives
  - $P(S') = E(S')$
  - For small $E$ (large $S'$)

Significance of BLAST in practice
1. Find best local alignment (score $S'$)
2. Assess statistical significance
   1. Calculate $K$ and $\mu$
   2. Find probability of $S'$ occurring by chance (when aligning random sequences)
      $P(S') = E(S') = Kmne^{-\mu S}$
      - Require $E(S') << 0$

Assessing search performance
- Which parameter setting is best?
- Is DP-based LA better than BLAST?
- Is my NewAndCleverHeuristic better than BLAST?

Search results as Venn diagram

Assessing search performance
- Sensitivity (true positive rate) $Se = TP/(TP+FN)$
- Specificity (1-false positive rate) $Sp = TN/(TN+FP)$
- Positive predictive value $PPV = TP/(TP+FP)$
- Correlation coefficient $cc = (TP \cdot FN - FP \cdot TN)/(\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FP)})$
Different thresholds give different performances

Search methods return prioritized lists:

ROC-curve show all Se/Sp tradeoffs

Different thresholds give different TP, FP, TN, and FN

Quality measure: Receiver operating characteristics (ROC) score calculated as area under curve

Relative True Positives

Random

Perfect

ROC\textsubscript{n}: Area under ROC curve up to \( n \) FP

\[ \text{ROC}\textsubscript{n} = \frac{1}{n P} \sum_{i=1}^{n} TP_i \]

ROC\textsubscript{n} example

- 20 positives (homologs)
- Two different methods, P1 and P2
- ROC\textsubscript{5}
  - P1: P, P, P, N, N, P, N, P, N, P, P, N, ...
  - P2: N, P, P, N, P, P, P, N, N, P, P, N, ...

  \[ \text{ROC}_{n}(P1) = \frac{1}{5 \times 20} \times (3+3+4+5+7) = 22/100 \]
  \[ \text{ROC}_{n}(P2) = \frac{1}{5 \times 20} \times (0+2+5+6+7) = 19/100 \]

Summary (2)

- Answering whether two sequences with LA score \( x \) are homologs
  - ...means answering
    - What is the probability \( p \) of having a local alignment score \( \geq x \) if the sequences are not homologs?
- Randomization
- Distribution-based estimates