Entropy

David Bernick
A case

Consider a set of related DNA polymerases. These are directly from our research and may be useful as future enzymes.

We are about to invest significant time and money in these predictions.

• Are they worth our investment?
Candidate viral DNA polymerase genes

- Aligned sequences are shown emphasizing their similarity to a true positive – phi29 DNA polymerase.
- Consider the meaning of gaps, identical columns (green), identical to phi29 (yellow) or similar (aqua).
Catalytic residues

DNA pol B proteins have been extensively studied, esp. phi29

– Asp12, Glu14 and Asp66 are conserved catalytic residues of the Exonuclease domain

– Asp249 and Asp458 are conserved catalytic residues of the Palm domain – the catalytic nucleotidyl transferase
Conservation in a column

• Conservation vs. Entropy
• We can compute maximum column Entropy as:

\[ S_{\text{max}} = - \sum_i P(x_i) \log P(x_i) \]

\[ = - \sum_i \frac{1}{N} \log \frac{1}{N} \]

\[ = \log N \]

and observed column Entropy as

\[ S_{\text{obs}} = - \sum_{n=1}^{N} P_n \log_2 P_n \]
Compare column Entropy to a Uniform Null Model

• and compare the entropies of the 2 models

\[ R = S_{\text{max}} - S_{\text{obs}} = \log N - \left( -\sum_{n=1}^{N} P_n \log_2 P_n \right) \]

WebLogo: A Sequence Logo Generator. Crooks GE, et al. 2004 Genome Research
R is Relative Entropy

\[ R = S_{\text{max}} - S_{\text{obs}} = \log N - \left( -\sum_{n=1}^{N} P_n \log_2 P_n \right) \]

\[ Q_x = \frac{1}{N} \]

\[ H(P \parallel Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \]

\[ = \sum_i P(x_i) \left( \log P(x_i) - \log \frac{1}{N} \right) \]

\[ = \sum_i P(x_i) \log P(x_i) + P(x_i) \log N \]

\[ \sum_i P(x_i) \log P(x_i) + P(x_i) \log N = \sum_i P(x_i) \log P(x_i) + \sum_i P(x_i) \log N \]

\[ = \log N + \sum_i P(x_i) \log P(x_i) \]

\[ H(P \parallel Q) = R \]
Sequence Logo
with small sample correction
Sensitivity and Specificity

Sensitivity measures the ability to recover Positives from a dataset
Specificity measures the ability to reject Negatives from a dataset

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \\
\text{Specificity} = \frac{TN}{TN + FP}
\]

FP -- Type I error
FN -- Type II error
Receiver Operating Characteristic

ROC Chart

ROC Space

Perfect Classification

Random guess

TPR or sensitivity

FPR or (1 - specificity)

Better

Worse

A

B

C