StatsM254 Statistical Methods in Computational Biology Lecture 12b - 05/22/2014

Basics of Information Theory

Lecturer: Jingyi Jessica Li *Scribe: Austin Quach*

This lecture is based on the tutorial paper titled *Information theory in molecular biology* [1].

1 Shannon's entropy (an uncertainty measure)

Suppose X is a discrete random variable that takes N different values, then **Shannon's entropy** $H(X)$ for X is defined as:

$$H(X) = -\sum_{i=1}^{N} P_i \log P_i \tag{1}$$

Entropy itself only describes the uncertainty in this random variable or an event but it never means information, *e.g.* if an observer is interested in a sequence of L nucleotides, then the number of possible sequences is 4^L. Entropy describes the observer's uncertainty about the true sequence. At Position 1, if the observer has no prior knowledge then he/she would guess: $P_A = P_T = P_C = P_G = \frac{1}{4}$. Therefore, the entropy at Position 1 is:

$$H(X_1) = -\sum_{i \in \{A,T,G,C\}} \frac{1}{4} \log \frac{1}{4} \overset{\text{base } 2}{=} 2 \quad (\log_2 \text{ relates } H \text{ to the unit of information known as the } \textit{bit}) \tag{2}$$

On the other hand, if the observer knows A is more likely to be at Position 1, then the entropy will be less than 2.

There is also a book titled *Elements of information theory* [2] for those interested in learning more. This book describes the entropy function as corresponding to the smallest number of yes/no questions necessary on average to identify the state of random variable X (we need a minimum of 2 questions to determine the nucleotide at Position 1). In other words, entropy can be viewed as the "[...] length of the shortest description of X [...]".

2 What is Information?

Information is defined as a relative quantity which measures the difference in entropy (uncertainty). For example, we can consider the case where X and Y are two discrete random variables. We introduce the **joint entropy** $H(XY)$ which measures the uncertainty about the joint distribution of X and Y, *i.e.* if X takes M values and Y takes N values, then XY can take MN values. If X and Y are independent, then we

can show that the joint entropy is simply a sum of the individual entropies of X and Y:

$$\begin{aligned} H(XY) &= -\sum_{i=1}^{M}\sum_{j=1}^{N} P_{ij}\log P_{ij} \\ &= -\sum_{i=1}^{M}\sum_{j=1}^{N} p_i q_j \log\left(p_i q_j\right) \\ &= -(\sum_{i=1}^{M}\sum_{j=1}^{N} p_i q_j \log p_i + \sum_{i=1}^{M}\sum_{j=1}^{N} p_i q_j \log q_j) \\ &= -\sum_{i=1}^{M} p_i \log p_i - \sum_{j=1}^{N} q_j \log q_j \\ &= H(X) + H(Y) \end{aligned} \tag{3}$$

Because of this the **mutual information** I is defined as:

$$I(X,Y) = H(X) + H(Y) - H(XY) \tag{4}$$

In the case that that they are completely dependent, $X = Y$ $(M = N)$:

$$H(XY) = \sum_{i=1}^{M} P_i \log P_i = H(X) \quad \text{(because } P_{ij} = P_i \text{ if } j = 1 \text{ or } P_{ij} = 0 \text{ otherwise)} \tag{5}$$

We also have another definition called **conditional entropy** $H(X|Y)$ which measures the extent of reduction in the uncertainty of X conditioned on Y:

$$H(X|Y) = H(XY) - H(Y) \tag{6}$$

Substituting, we can express the mutual information as:

$$\Rightarrow I(X,Y) = H(X) - H(X|Y) \tag{7}$$

We can apply this concept in the context of biological sequences. The highest entropy we can have for DNA at one position is 2. If we have prior knowledge that the sequence in a specific region is GC-rich, then the entropy will be much less. In the study RNA secondary structure, researchers use the reduction in entropy to guide them in finding the sequences that should bind together. The base entropy for protein sequences is higher because there are 20 possible amino acids, *i.e.* $\log_2 20 = 4.32$ bits.

Also to recap on the **mutual information coefficient (MIC)**, the MIC method estimates the mutual information by drawing an $m \times n$ grid on a scatterplot of X and Y to transform X and Y into discrete variables. Using these discretized variables and a normalization factor we can find:

$$MIC = \frac{I(X,Y)}{\log \min\left(m,n\right)} \tag{8}$$

There is some dispute of the utility of MIC as the normalization term in the denominator is thought to be a source of bias and it has been shown to be no better than the mutual information itself. Additionally the method uses heuristic methods to discretize the variables which also imposes biases.

References

[1] Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, 1(1), 3-22.

[2] Cover, T. M., Thomas, J. A. (1991). *Elements of information theory*. John Wiley and Sons.