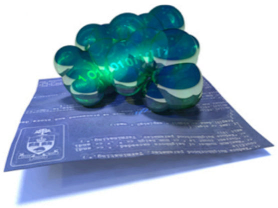
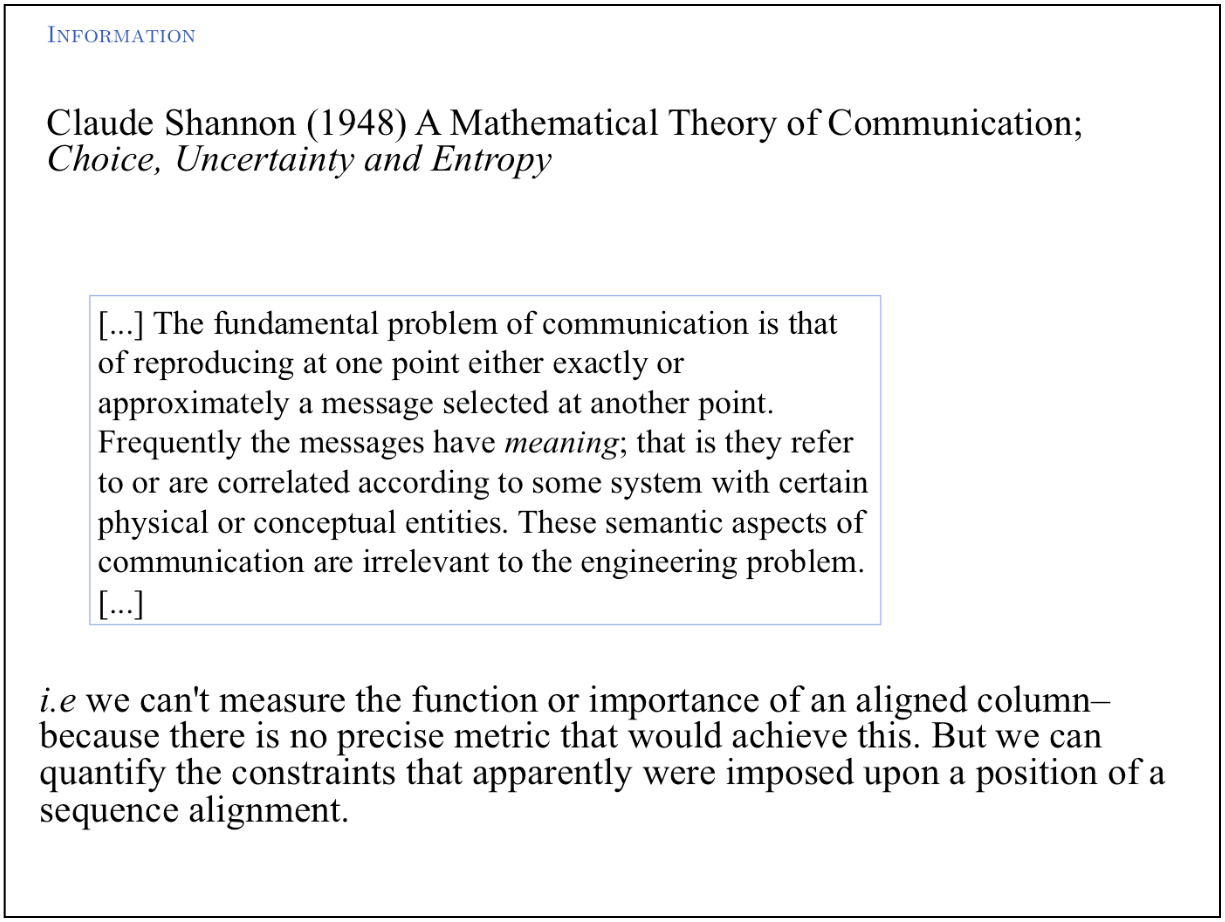
A
Bioinformatics
Course

Information Theory



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



Information

Claude Shannon (1948) A Mathematical Theory of Communication;
Choice, Uncertainty and Entropy

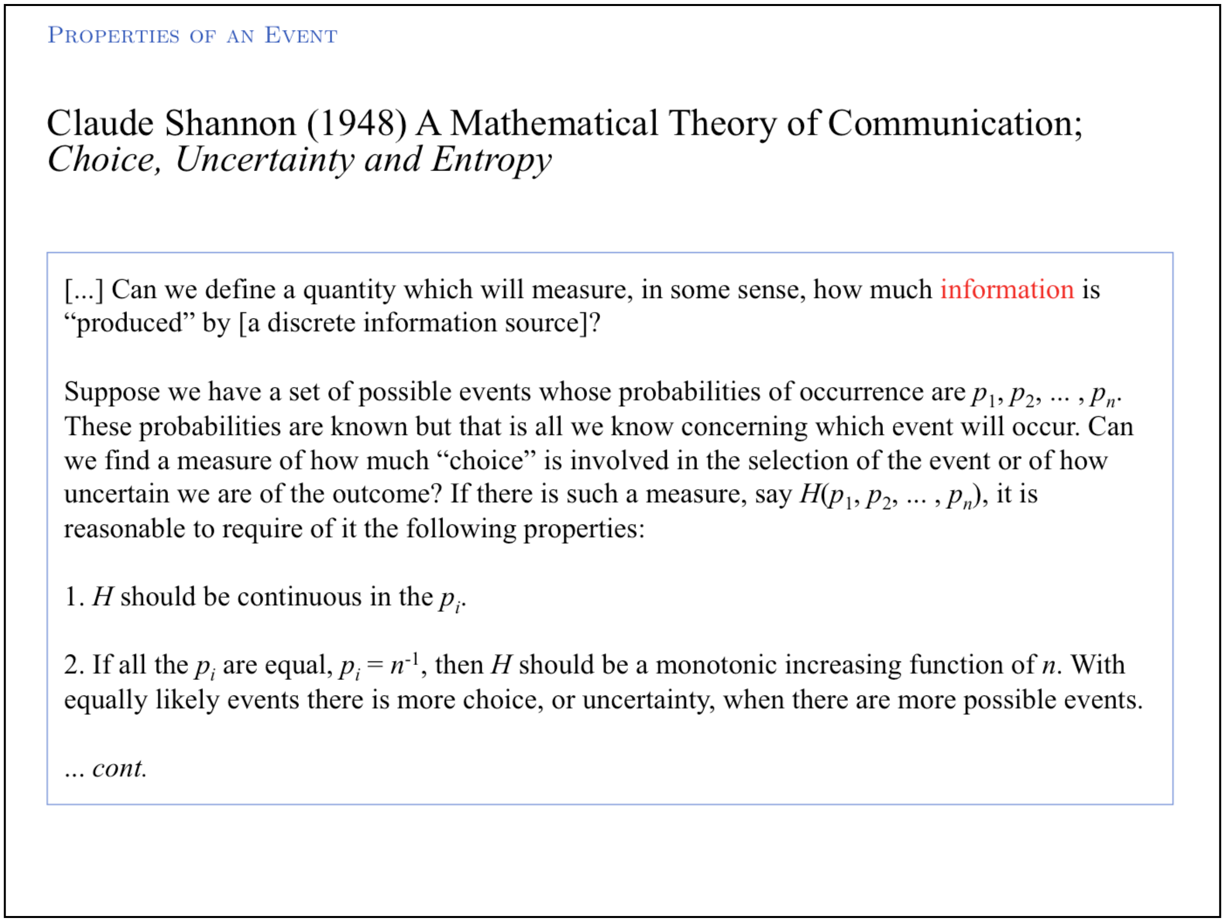
> [...] The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. [...]

i.e we can't measure the function or importance of an aligned column– because there is no precise metric that would achieve this. But we can quantify the constraints that apparently were imposed upon a position of a sequence alignment.

A quantitative theory of information was formulated by Claude Shannon in 1948. It applies in many ways to biological sequences (and biological systems in general), because it quantifies how different an **observed** distribution of states (e.g. amino acids, or nucleotides) is from an expected distribution, e.g. produced by a stochastic process, or merely reflecting general database trends. Observing a difference between observation and expectation ...

... implies that some selective process was operating on the sequence, which means ...

... there is some functional significance to it.

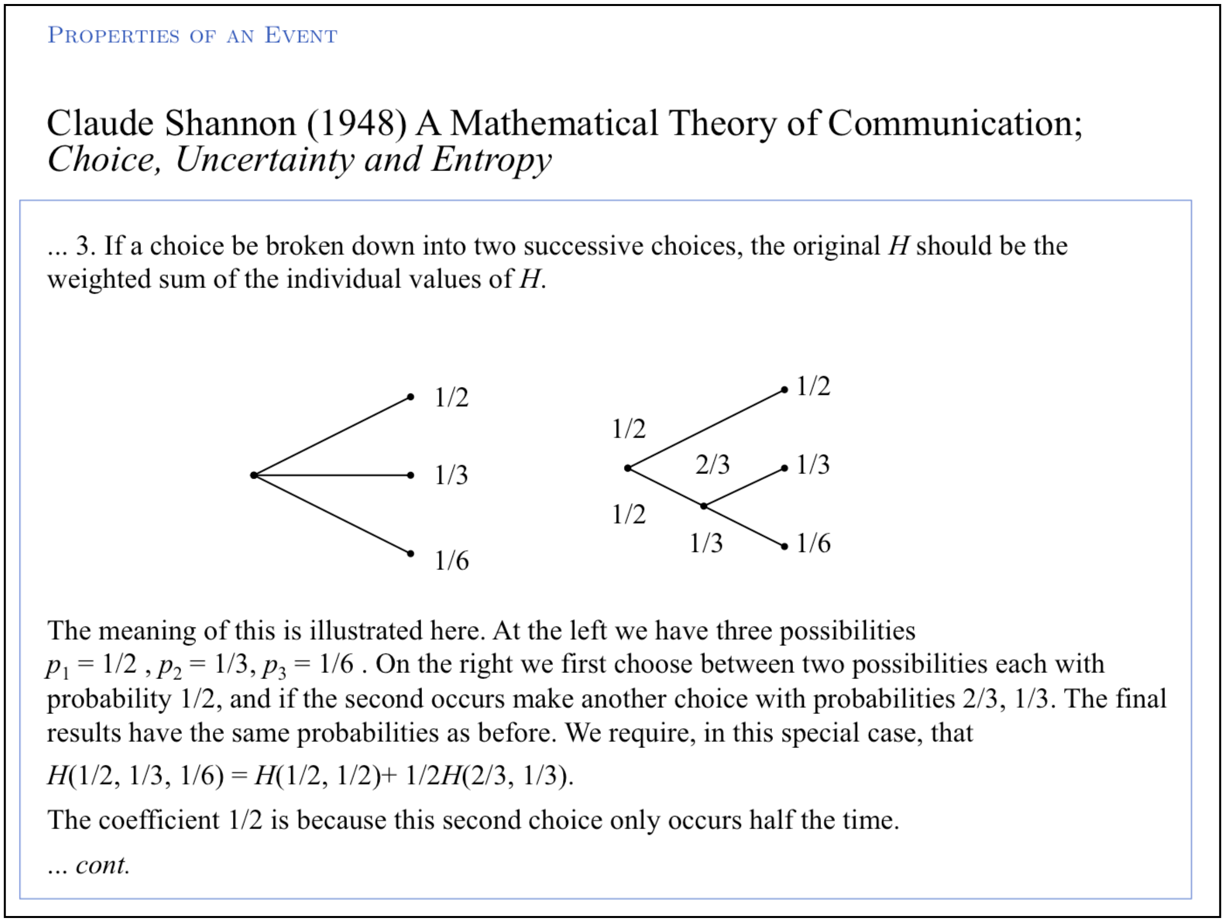


Properties of an Event

Claude Shannon (1948) A Mathematical Theory of Communication; *Choice, Uncertainty and Entropy*

> [...] Can we define a quantity which will measure, in some sense, how much information is "produced" by [a discrete information source]?
>
> Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \dots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome? If there is such a measure, say $H(p_1, p_2, \dots, p_n)$, it is reasonable to require of it the following properties:
>
> 1. H should be continuous in the p_i.
>
> 2. If all the p_i are equal, $p_i = n^{-1}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
>
> ... *cont.*

Shannon figured out the formula that measures the difference between observation and expectation from a simple constraint. He sought out how to define a quantity, H, that satisfied a number of intuitive properties that a measure of information has: ...



Properties of an Event

Claude Shannon (1948) A Mathematical Theory of Communication; *Choice, Uncertainty and Entropy*

... 3. If a choice be broken down into two successive choices, the original *H* should be the weighted sum of the individual values of *H*.

1/2
1/3
1/6
1/2
1/2
2/3
1/3
1/2
1/3
1/6

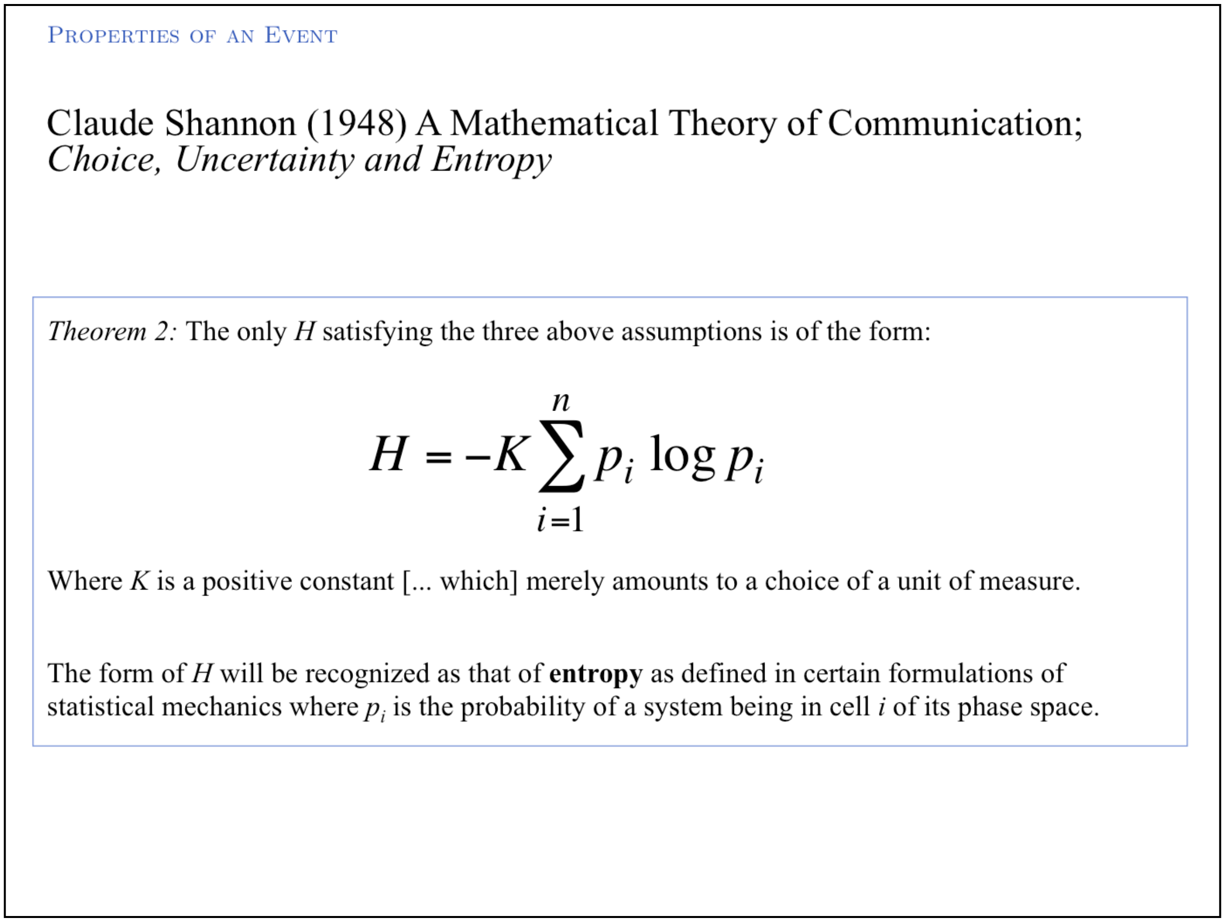
The meaning of this is illustrated here. At the left we have three possibilities $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/6$. On the right we first choose between two possibilities each with probability 1/2, and if the second occurs make another choice with probabilities 2/3, 1/3. The final results have the same probabilities as before. We require, in this special case, that

$H(1/2, 1/3, 1/6) = H(1/2, 1/2) + 1/2H(2/3, 1/3)$.

The coefficient 1/2 is because this second choice only occurs half the time.

... *cont.*

... any valid measure must be independent of the sequence of choices we make to observe a particular distribution.



Properties of an Event

Claude Shannon (1948) A Mathematical Theory of Communication;
Choice, Uncertainty and Entropy

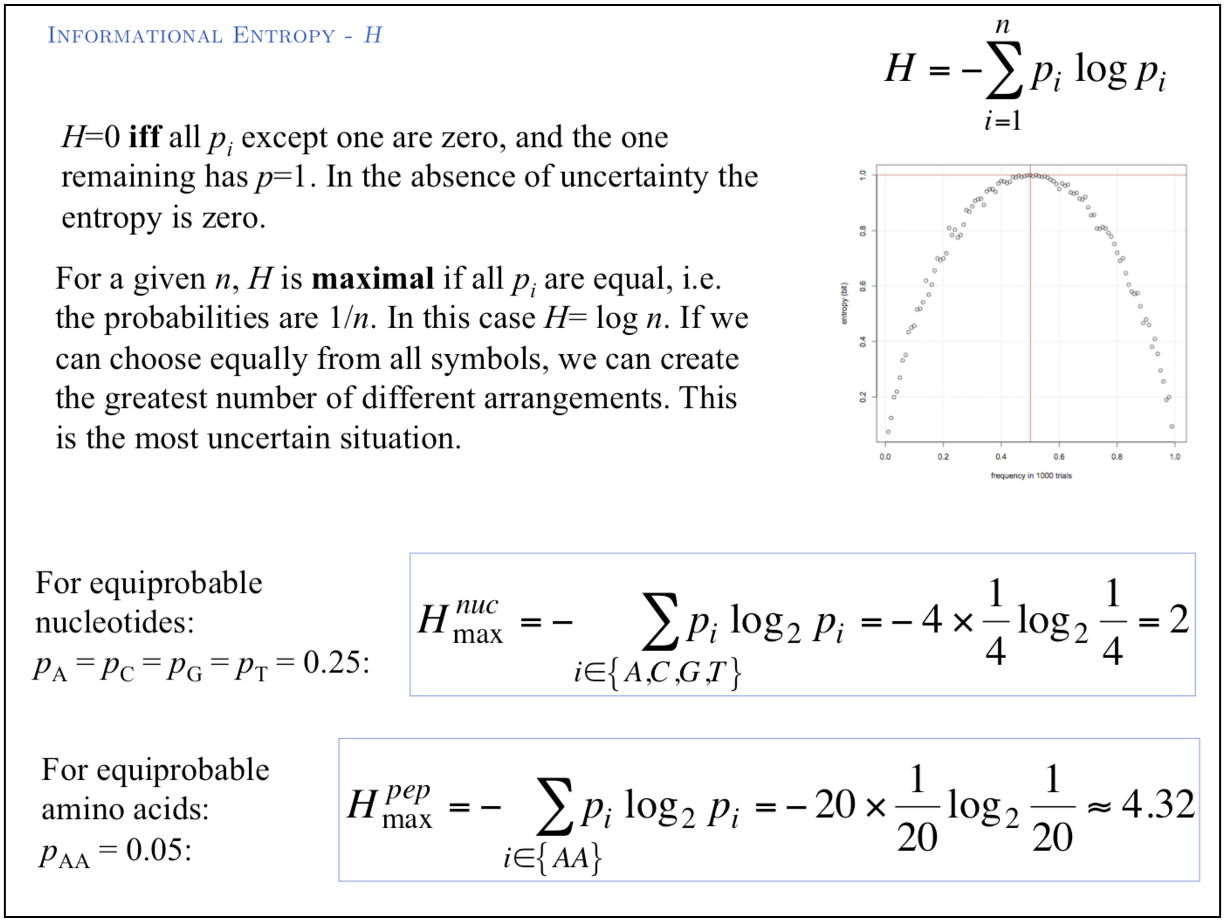
Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

Where K is a positive constant [... which] merely amounts to a choice of a unit of measure.

The form of H will be recognized as that of **entropy** as defined in certain formulations of statistical mechanics where p_i is the probability of a system being in cell i of its phase space.

It turns out the there is only one mathematical expression with this property. The quantity H is now known as (informational) entropy.



Informational Entropy - H

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

H=0 **iff** all p_i except one are zero, and the one remaining has p=1. In the absence of uncertainty the entropy is zero.

For a given n, H is **maximal** if all p_i are equal, i.e. the probabilities are 1/n. In this case H= log n. If we can choose equally from all symbols, we can create the greatest number of different arrangements. This is the most uncertain situation.

For equiprobable nucleotides:
$p_{\mathrm{A}} = p_{\mathrm{C}} = p_{\mathrm{G}} = p_{\mathrm{T}} = 0.25$:

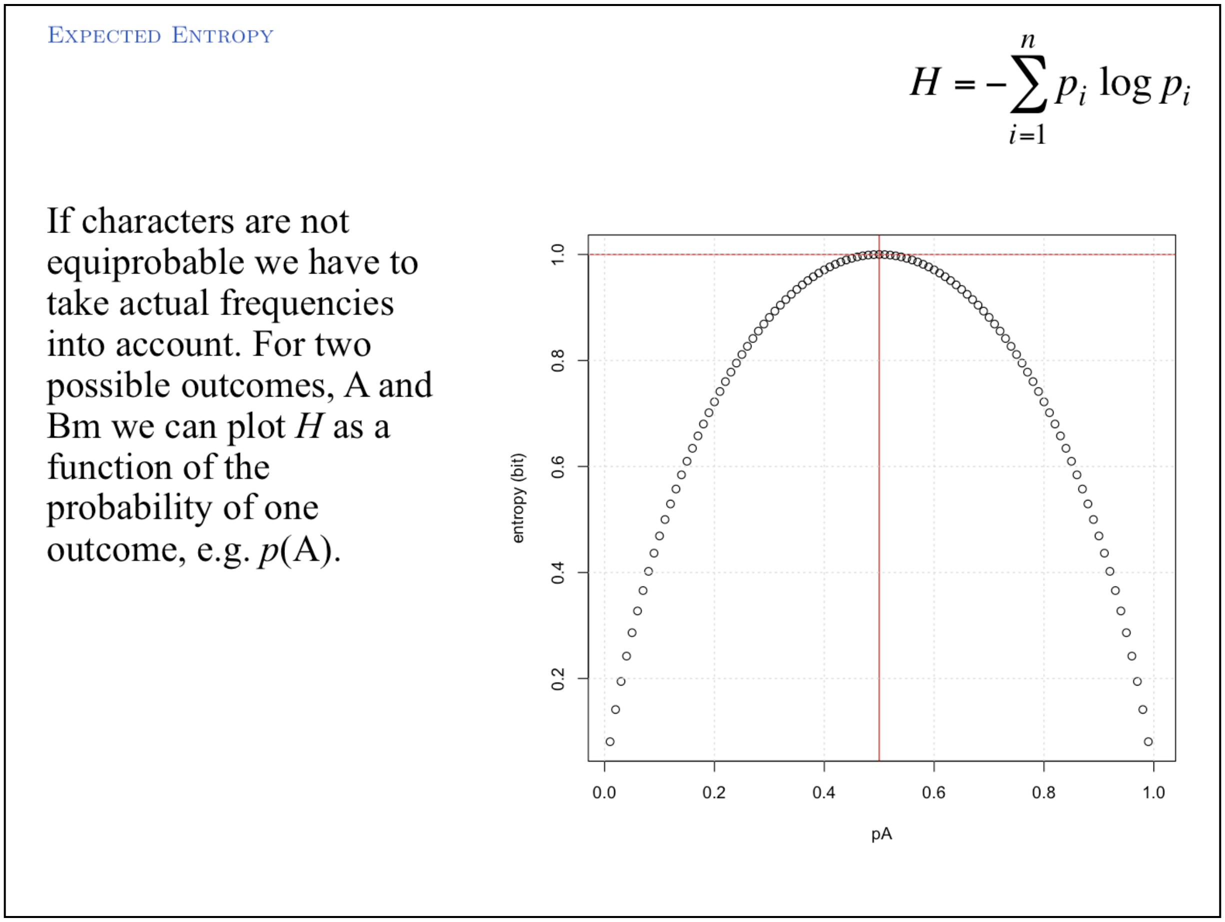
$$H_{\max}^{nuc} = -\sum_{i \in \{A,C,G,T\}} p_i \log_2 p_i = -4 \times \frac{1}{4} \log_2 \frac{1}{4} = 2$$

For equiprobable amino acids:
$p_{\mathrm{AA}} = 0.05$:

$$H_{\max}^{pep} = -\sum_{i \in \{AA\}} p_i \log_2 p_i = -20 \times \frac{1}{20} \log_2 \frac{1}{20} \approx 4.32$$

This is simple to compute.

If n states are equally probable, $H = \log\, n$.



Expected Entropy

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

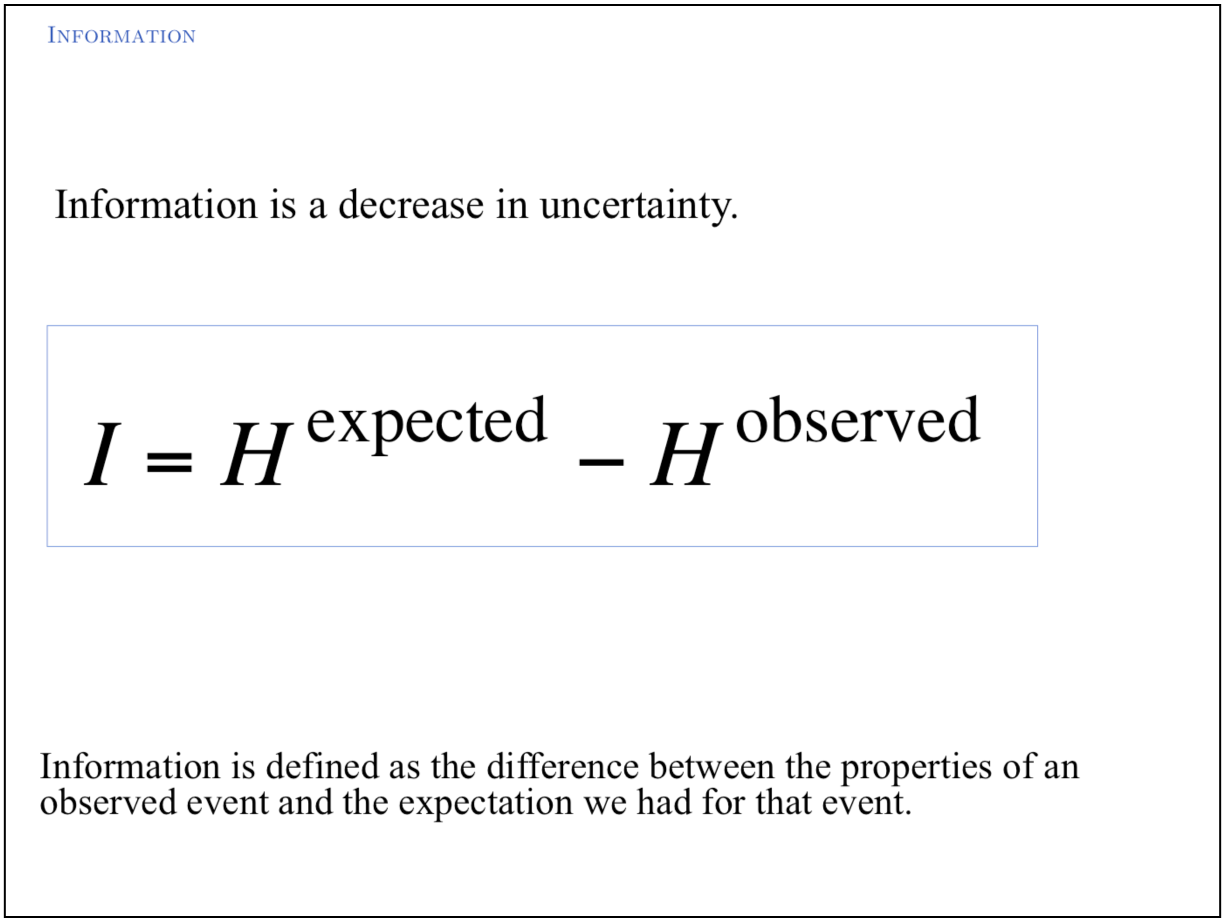
If characters are not equiprobable we have to take actual frequencies into account. For two possible outcomes, A and Bm we can plot *H* as a function of the probability of one outcome, e.g. *p*(A).

If we have two possible outcomes A and B, *H* depends on the probabilites of *A* and *B* (or *A* and 1-*A*, which is the same).

We can plot the entropy for pA = 0.1, pB= 0.9; pA=0.2, pB=0.8 ... etc.

Entropy is zero when either outcome has zero probability (0 log0 + 1 log1= 0).

Entropy is maximal when both outcomes are equiprobable i.e pA = pB = ½.



Information

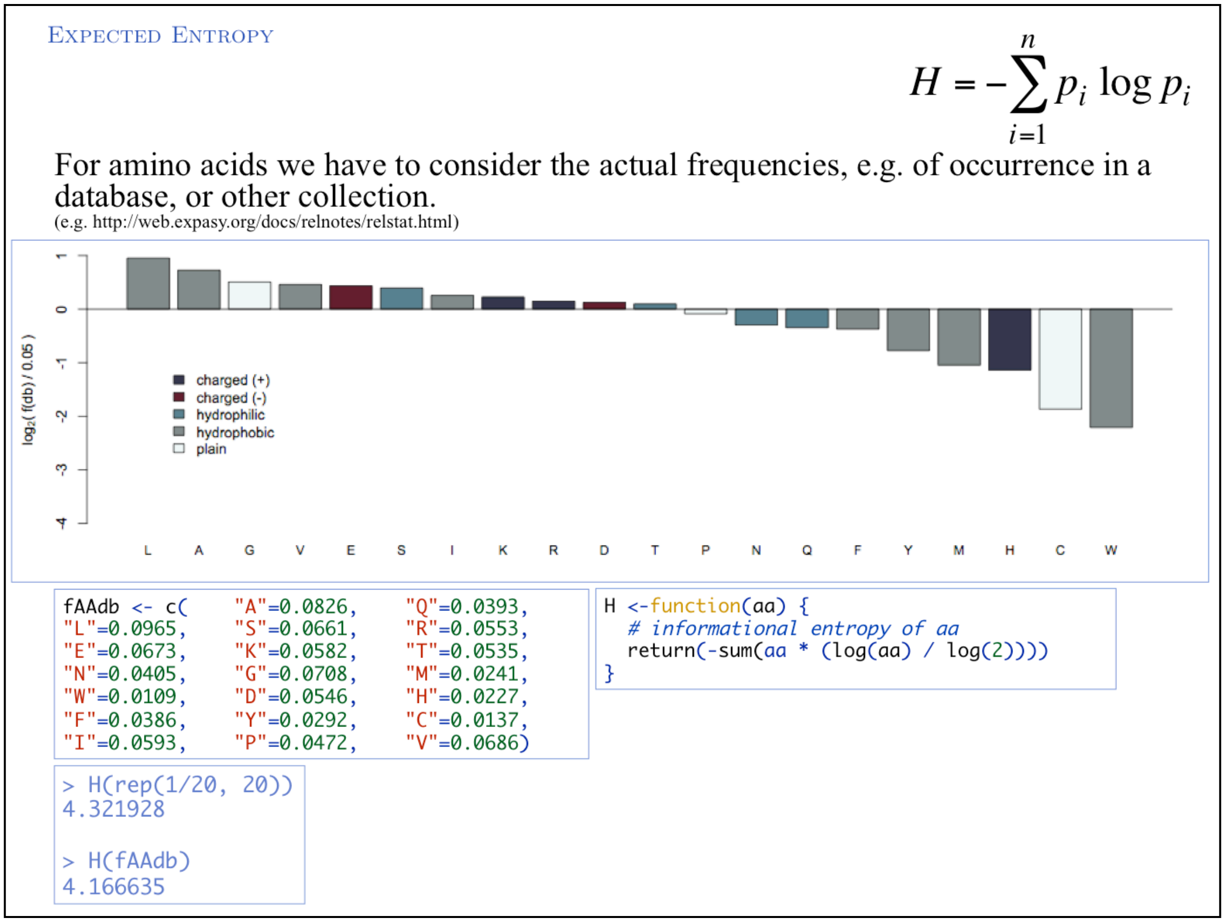
Information is a decrease in uncertainty.

$$I = H^{\text{expected}} - H^{\text{observed}}$$

Information is defined as the difference between the properties of an observed event and the expectation we had for that event.

Finally we get to define **information**:

Information is the difference between the entropy of a distribution that we expect, and the entropy that we actually observe.



Expected Entropy

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

For amino acids we have to consider the actual frequencies, e.g. of occurrence in a database, or other collection.

(e.g. http://web.expasy.org/docs/relnotes/relstat.html)

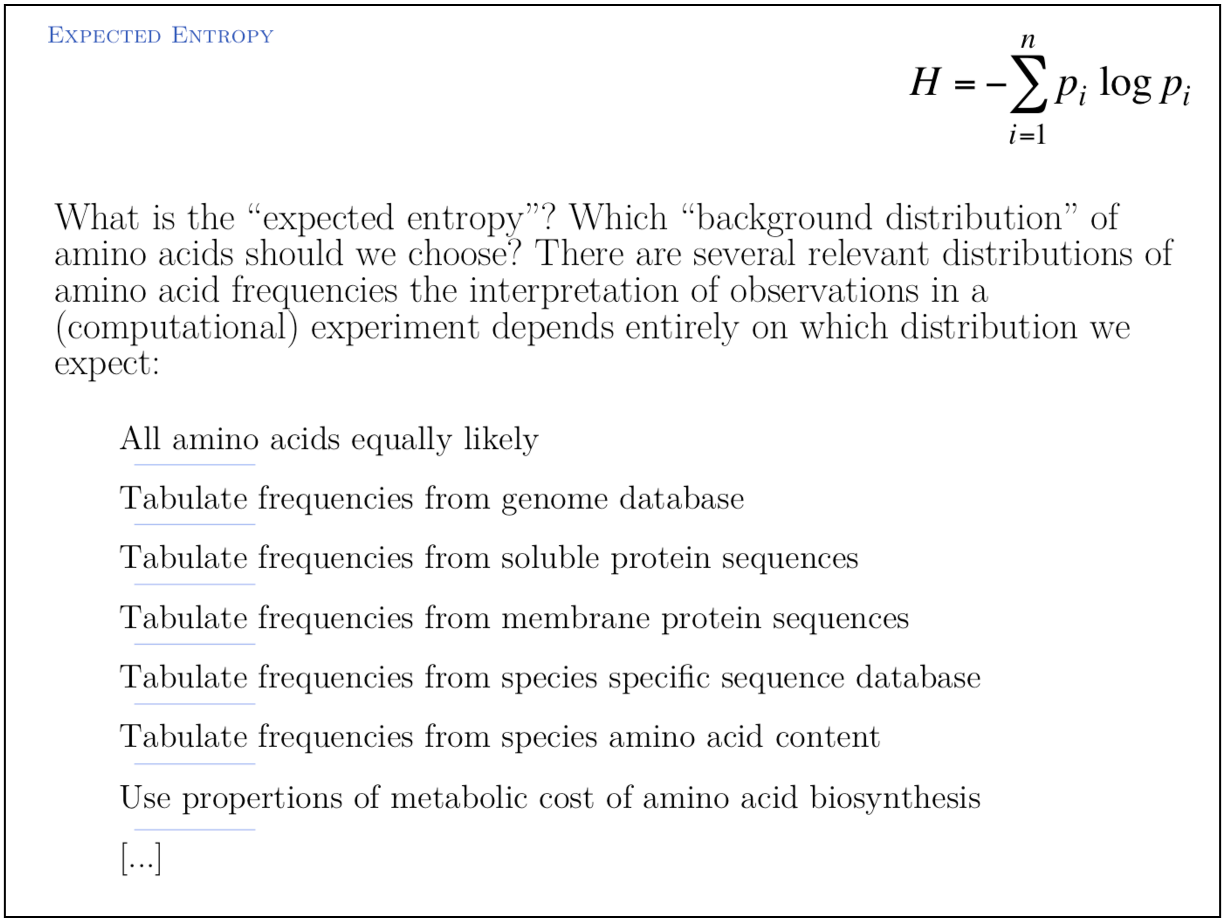
```
fAAdb <- c(     "A"=0.0826,     "Q"=0.0393,
"L"=0.0965,     "S"=0.0661,     "R"=0.0553,
"E"=0.0673,     "K"=0.0582,     "T"=0.0535,
"N"=0.0405,     "G"=0.0708,     "M"=0.0241,
"W"=0.0109,     "D"=0.0546,     "H"=0.0227,
"F"=0.0386,     "Y"=0.0292,     "C"=0.0137,
"I"=0.0593,     "P"=0.0472,     "V"=0.0686)
```

```
H <-function(aa) {
  # informational entropy of aa
  return(-sum(aa * (log(aa) / log(2))))
}
```

```
> H(rep(1/20, 20))
4.321928

> H(fAAdb)
4.166635
```

If we want to measure the information of amino acid distributions, we need to define the expected background distribution. Assuming all amino acids are equally likely is usually not a good assumption. Often we use the frequencies of amino acids that we observe in a sequence database instead.



Expected Entropy

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

What is the “expected entropy”? Which “background distribution” of amino acids should we choose? There are several relevant distributions of amino acid frequencies the interpretation of observations in a (computational) experiment depends entirely on which distribution we expect:

All amino acids equally likely

Tabulate frequencies from genome database

Tabulate frequencies from soluble protein sequences

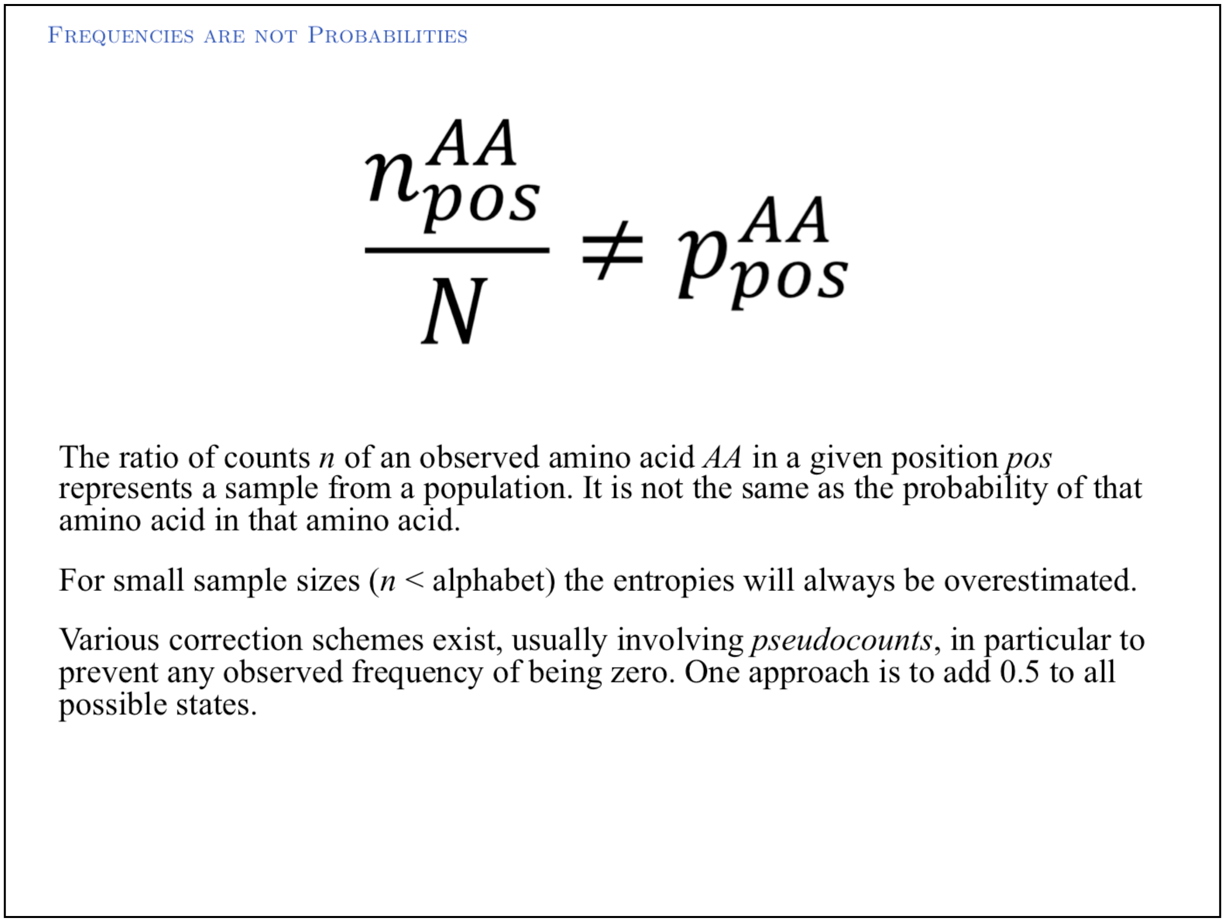
Tabulate frequencies from membrane protein sequences

Tabulate frequencies from species specific sequence database

Tabulate frequencies from species amino acid content

Use propertions of metabolic cost of amino acid biosynthesis

[...]



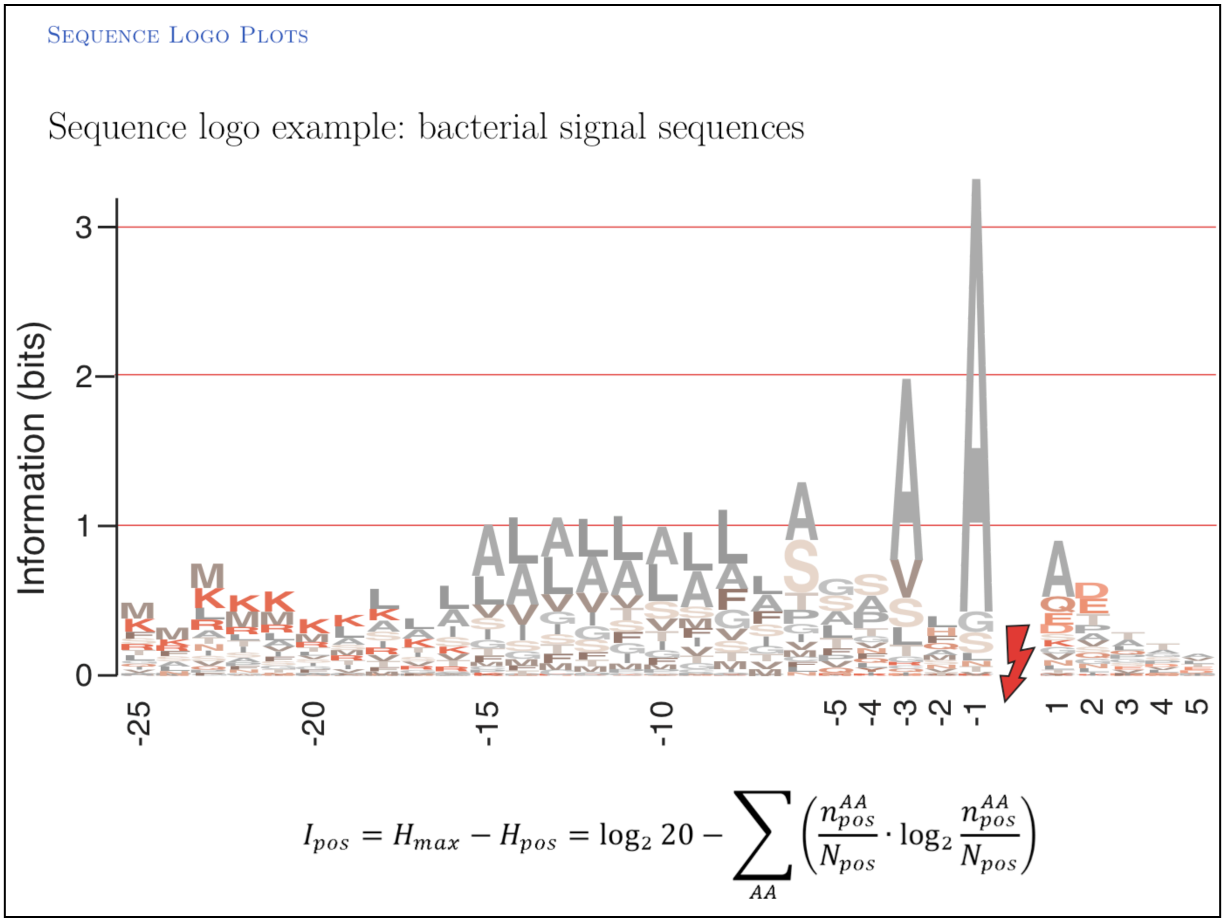
Frequencies are not Probabilities

$$\frac{n_{pos}^{AA}}{N} \neq p_{pos}^{AA}$$

The ratio of counts *n* of an observed amino acid *AA* in a given position *pos* represents a sample from a population. It is not the same as the probability of that amino acid in that amino acid.

For small sample sizes (*n* < alphabet) the entropies will always be overestimated.

Various correction schemes exist, usually involving *pseudocounts*, in particular to prevent any observed frequency of being zero. One approach is to add 0.5 to all possible states.



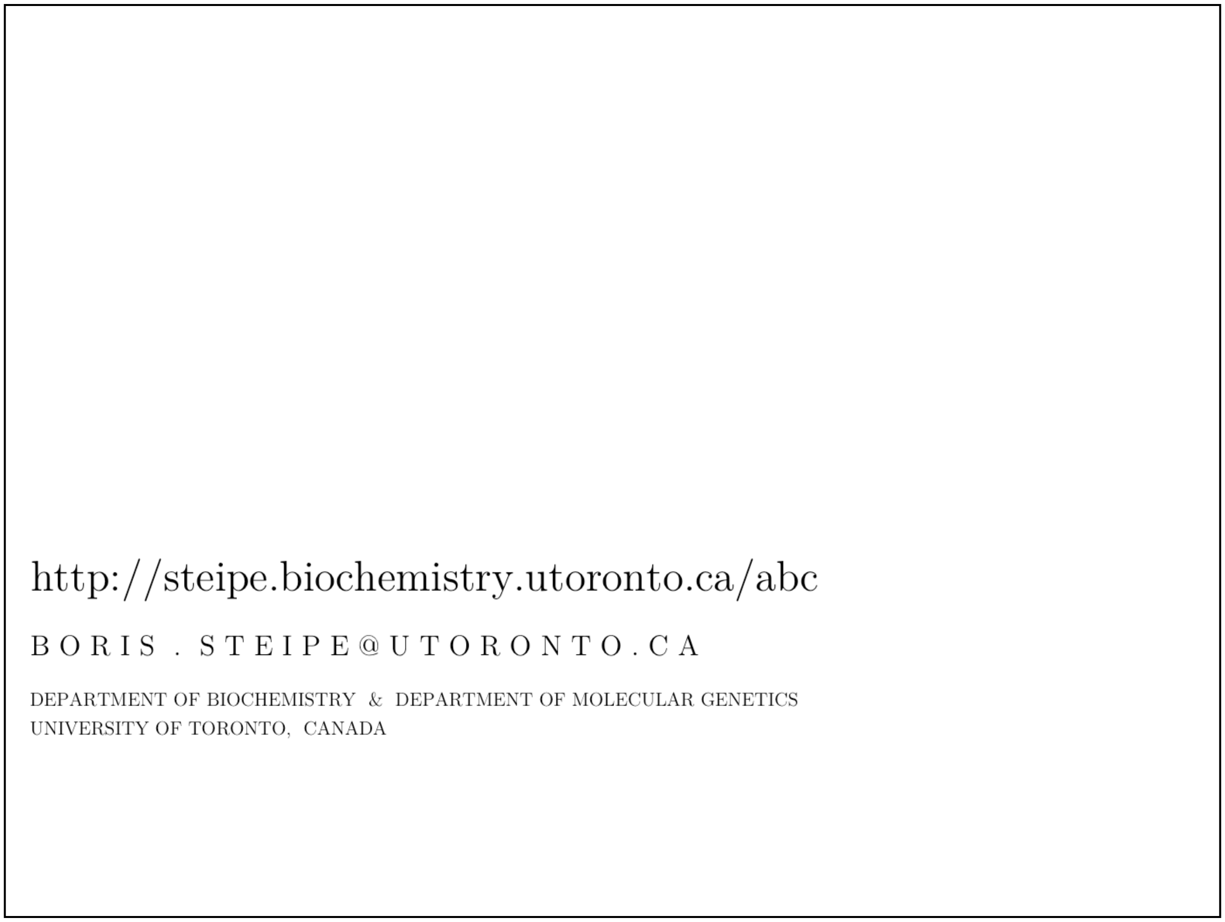
Sequence Logo Plots

Sequence logo example: bacterial signal sequences

$$I_{pos} = H_{max} - H_{pos} = \log_2 20 - \sum_{AA} \left(\frac{n_{pos}^{AA}}{N_{pos}} \cdot \log_2 \frac{n_{pos}^{AA}}{N_{pos}} \right)$$

Sequence logos plot features of aligned sets of sequences. Each column corresponds to a position in an alignment, the height of each stack corresponds to the **information** calculated for the residues that are observed in that position, and the height of each letter in a stack corresponds to its frequency in that position. This emphasizes the conserved positions, and displays **what** is conserved.

For this plot, bacterila signal sequences were aligned on the signal-peptidase cleavage site. Their common features include a positively charged N-terminus, a hydrophobic helical stretch and a small residue that precedes the actual cleavage site.



http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA