# Lecture 8: Information Theory

## I. Introduction

**Information** and **uncertainty** are terms that describe any process that selects one or more objects from a set of objects. Proteins do this all the time. A soluble enzyme must select the proper substrate from a soup of competing molecules; RNA polymerase must select a specific sequence of DNA to begin transcription, a transmembrane ion channel must select the correct ion to transport, etc. We can describe protein function in terms of the information transfer that occurs during binding of a ligand and the process of carrying out the function.

Information and uncertainty are related but they are opposite in a sense. For a protein to recognize a ligand it has to have multiple interactions with that ligand - perhaps (1) a hydrogen bond, (2) a favorable charge-charge interaction, and (3) some favorable VDW interactions due to complementary geometry. If one of these interactions is recognized (e.g. the hydrogen bond) the protein still has uncertainty as to whether or not the ligand is the correct one. Once another correct interaction occurs (e.g. the charge interaction) the uncertainty of the ligand identity is decreased and we can say that the protein received some information. In other words, the information received by the protein is a *decrease* in the uncertainty. This relationship between information and uncertainty is a subtle but important one; a subtraction in the *before* and *after* uncertainties is necessary to define the information transferred.

In the following we will designate uncertainty as $H$ and information as $I$. Thus, for the above case,

$$I = (H_{Hbond} - H_{Hbond,\,charge}) = -\Delta H$$

where $I$ is the information transferred, $H_{Hbond}$ is the uncertainty in ligand identity after the first contact, $H_{Hbond,charge}$ is the uncertainty left after two correct contacts are made, $-\Delta H$ is the decrease in uncertainty (the negative sign indicates that as uncertainty decreases, information is gained).

A good way to look at this kind of information is that the information $I$ inherent in any given arrangement of matter (or energy) is related to the number of possible *outcomes* when making choices about the identity of the matter. For a coin flip, there are two possible outcomes and we say that the coin provides one bit of information or that we need one bit of information to know the identity (heads or tails). The more possible arrangements (in the above example, the different specific contacting groups on the ligand) the more information needed to identify that particular ligand.

After each new atomic contact between an enzyme and ligand a choice is made by the protein: Is this the substrate, yes or no? In the example above there are three types of contacts, so before the initial binding the uncertainty about the identity of the ligand (as far as the protein is concerned) is related to eight possible outcomes as described next.

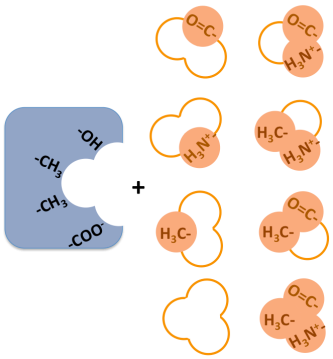
The figure at right illustrates the eight possible ligand outcomes (identities, orange). Uncertainty is usually quantified using the base 2 logarithm,

$$H = \log_2(8) = 3 \text{ bits}$$

After the protein recognizes the first contact (e.g., -OH---O=C-, Hbond) it still has uncertainty of four possible outcomes (identities). After the first contact is recognized,

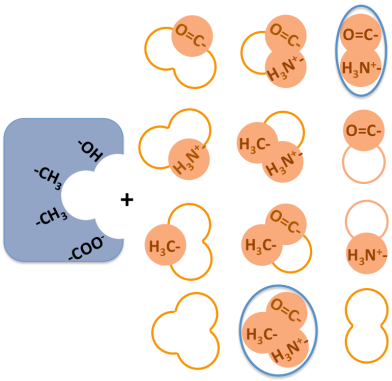
$$H = \log_2(4) = 2 \text{ bits}$$

and there are 2 bits of uncertainty left.




Now consider the case where the binding specificity of our enzyme is somewhat promiscuous - it recognizes two different substrates, one that has all three contacts (hydrogen bond, charge-charge, VDW complementarity, at bottom) and another one that does not have the VDW complementarity (upper right).

As illustrated in the figure at right, although there appear to be nine different outcomes (identities) some look equivalent to the protein. We need to extend our definition of uncertainty to take this equivalence into account, i.e. we need to consider the *probability* of any set of features required to identify the substrate.




Previously the uncertainty has been defined as $\log_2(N)$, where $N$ is the number of different outcomes. To relate this to the probability of *equivalent* outcomes we can consider the following rearrangements,

$$\log_2(N) = -\log_2(N^{-1})$$
$$\log_2(N) = -\log_2(1/N)$$
$$\log_2(N) = -\log_2(P)$$

where $P = 1/N$ is the probability that any equivalent outcome will be observed. The uncertainty the protein has about observing the $i^{th}$ equivalent outcome upon binding is defined by analogy with the above relationship ($H = \log_2(N)$) to be

$$H_i = -\log_2 P_i$$

If $P_i$ approaches 0, then the protein will have great uncertainty about observing the $i^{th}$ equivalent outcome and therefore $H_i$ approaches $\infty$. On the other hand, if $P_i = 1$, then the protein won't be uncertain at all about observing the $i^{th}$ outcome and $H_i = 0$.

The average uncertainty $H$ for all the possible equivalent outcomes available to the protein is the sum of the product of each equivalent outcome times its probability,

$$H = \sum P_i H_i$$

And by substituting for $H_i$ from above we obtain,

$$H = -\sum P_i \log_2 P_i$$

This is Shannon's famous general formula for uncertainty. You will frequently see this type of uncertainty defined as **Shannon entropy**. Other common definitions of Shannon entropy are:

- the measure of average uncertainty in a random variable.
- the average number of bits needed to describe a random variable.

We will continue to call it uncertainty to avoid confusion with Boltzmann entropy as described below. Also, you will see the same expression with opposite sign written as the definition of **Shannon information**,

$$I = \sum P_i \log_2 P_i$$

When proteins function, they receive information from the ligands and other proteins they interact with. Of course, proteins also contain a lot of information in their primary, secondary and tertiary structures. Many treatments of information theory in biology follow the information flow from DNA to protein structure, and we could have a whole course on this topic, but I want to concentrate on how proteins receive and use information from ligand binding in their function.

The units for uncertainty or information here are **bits**/contact because we used the log base 2 in the formulae. Beware that some treatments of information will use natural log (ln) and then the units are **nats** instead of bits. It doesn't matter as long as you are consistent within the same treatment.

A good general summary of the relationship between Shannon uncertainty ($H$) and information content of a discrete variable $X$ with possible values $\{x_i, ..., x_n\}$ and probabilities $p(x_i)$, $(i=1, ..., n)$ as expressed in various equation formats you will encounter is the following:

$$H(X) = \sum_{i=1}^{n} p(x_i) I(x_i) = \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

where $b$ is the base of the logarithm used and $I(x_i)$ denotes the information content of $x_i$. If $b = 2$, the units are in bits, if $b = e$, the units are in nats.

**II. Entropy and information**. If all this seems unfamiliar let's turn to Boltzmann entropy. You will remember the famous equation written on Boltzmann's tombstone,

$$S = k \ln W$$

Here $S$ is a measure of how much we can infer about the arrangement of a system on the basis of its distribution. The larger $S$, the less information a distribution conveys about the arrangement of the system. Notice how we just introduced the concept of information into Boltzmann theory.

One definition for $W$ that we encountered previously is the general expression for calculating the weight of a configuration, $N!/n_1!n_2!...n_x!$ where $N$ is the total number of systems, $n_i$ is the number of systems in a particular energy state, and $x$ is the total number of states. Substituting this definition in the Boltzmann equation above,

$$S = k \ln(N!/n_1!n_2!...n_x!)$$
$$S = k \ln N! - k \ln n_1!... - k \ln n_x!$$

and using Stirling's approximation,

$$S \approx (Nk \ln N - N) - (n_1 k \ln n_1 - n_1)... - (n_x k \ln n_x - n_x)$$
$$S = -k \sum n_i \ln n_i$$

let $P_i = n_i/N$, the probability that a system occurs in a particular state,

$$S = -Nk \sum P_i \ln P_i$$

where the units of entropy here are joules/$K$•microstate. This is a generalization of Boltzmann's formula for the case when all energy states are not equally probable. This looks very similar to Shannon's equations above except that Boltzmann used the natural log (ln). We can relate Shannon information to Boltzmann's entropy as follows. Remembering that $\log_2 X = \ln X/\ln 2$ then,

$$I = \sum P_i \log_2 P_i = \frac{\sum P_i \ln P_i}{\ln 2}$$

$$\ln 2\, I = \sum P_i \ln P_i$$

Substituting this definition of $I$ into Boltzmann's equation we obtain,

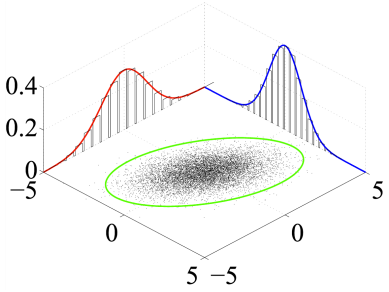
$$S = -Nk \ln 2\, I$$

Where $Nk$ can be considered just a scaling factor. In other words, for a molecular system (e.g. a protein) to make choices (gain information) its entropy must decrease. See **Appendix 1: Four Entropies** for a comparison of different definitions of entropies.

**III. Mutual information**. The usefulness of information theory to the typical questions we have in biophysics of macromolecular function comes into play when we consider the **mutual information** of components within our system. We especially want to understand how the dynamics of a protein are used in its function. For example, how is the binding of an allosteric effector on one side of a protein communicated to the active site or to residues important in quaternary structure changes? *We want to know the mutual information shared between these different sites on the protein*.

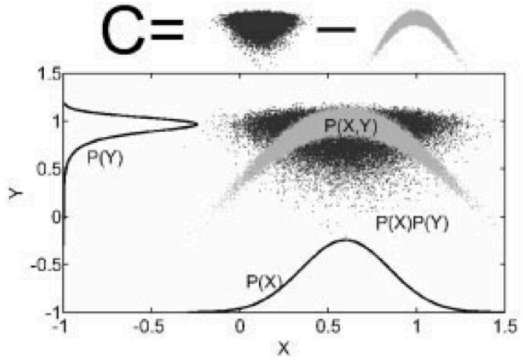
To understand mutual information, we start with the concept of a **joint probability distribution**. Wikipedia has a nice graphic to illustrate a joint probability distribution of two random variables ($X$, $Y$) with **marginal distributions** between -5 and +5 (see next figure). Black points in the green ellipse in the figure below represent samples from the joint probability distribution, $p(x,y)$, of these two variables; the red and blue curves represent the marginal distributions of values for each variable, $p(x)$ and $p(y)$.




The joint probability distribution for this case simply means the probability of any two values being found together in a combined dataset of both marginal distributions. If the two marginal distributions are independent and uncorrelated the joint probability distribution is equal to the product of the two marginal distributions,

$$p(x,y) = p(x)p(y).$$

We want to quantify the correlation (C) between the variables $X$, $Y$ as the *deviation* between both sides of the above equation as illustrated in the next figure.




In the figure at left the black dots represent the joint probability distribution of two *independent* distributions, $p(X)p(Y)$; the gray dots represent the joint probability distribution of two *correlated* distributions, $p(X,Y)$. The difference between these two distributions gives a measure of the correlation.

The mutual information of two variables such as those above is defined as,

$$I(X;Y) = \sum_{y} \sum_{x} p(x,y) \log_2 \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

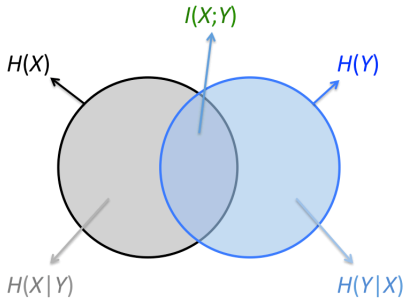
where $p(x,y)$ is a joint probability distribution and $p(x)$ and $p(y)$ are marginal distributions. In other words, the average information that $Y$ gives us about $X$ is the average *decrease* in uncertainty of $X$ by observing $Y$. If $p(x,y) = p(x)p(y)$ you can see that $I(X;Y) = 0$, i.e., there is no mutual information. *Only if the joint probability distribution deviates from what is expected from random association is there information gained from the joint probability distribution.*

The application of mutual information to quantify the relationships of protein conformational dynamics is considered an improvement on the use of straightforward correlation analysis. See **Appendix 2: Covariance and correlation** for a review of the definitions of covariance and correlation. Estimates of correlations from the Pearson correlation coefficient are only strictly valid if the atomic fluctuations are colinear, a limitation that is frequently not met with protein conformational dynamics. Also, the use of covariance implies a Gaussian approximation of the underlying atomic fluctuations which is not always the case.

We can also define mutual information in terms of uncertainty. If $H(X)$ is the uncertainty of a single random variable, and the **conditional uncertainty** $H(X|Y)$ is the uncertainty of one random variable conditional upon knowledge of another then the mutual information shared by X and Y is,

$$I(X;Y) = H(X) - H(X|Y)$$

This relationship can be illustrated with Venn diagrams such as shown below. When using Venn diagrams in this field we consider *subtraction* rather than the usual *addition*. For example, the following diagram indicates the mutual information common to two variables $X$, $Y$. Arrows stemming from the perimeter of a circle refer to the area inside the *whole* circle. Arrows stemming from the interior of a region refer to the area of only that *region* of the circle. Conditional relationships are represented by subtraction, e.g. the set representing $H(X|Y)$ results from subtracting the set representing $H(Y)$ from the set representing $H(X)$. The mutual information $I(X;Y)$ is defined by subtraction of $H(X|Y)$ from $H(X)$ as in the formula above.



You will also see various other treatments that describe **joint uncertainty**, $H(X,Y)$ which is the sum of the two Venn circles above. We will ignore this for now.

Finally, the above discussion is concerned with two one-dimensional sets of variables. In the case of three dimensional variables such as the atomic coordinates of a protein's atoms the correlations can be more complicated and the formulae are slightly different. We may also want to consider higher order correlations to include paths of information flow. But the concepts underlying these more complicated cases are still based on the fundamental concepts outlined above.

**Appendix 1: Four entropies.**

*Boltzmann entropy:* $S_B = -Nk\sum P_i \ln P_i$

where $N$ = number of systems, $P_i = n_i/N$ = probability that a system occurs in a particular microstate ($n_i$ = microstate of a system), $k$ = Boltzmann's constant. Boltzmann was interested in the microscopic state of a gaseous particle system and the probability of it being in a particular energy state.

*Shannon entropy:* $H = -\sum P_i \ln P_i$

where $P_i$ is probability that a specific item is part of a description (e.g. a molecular contact describing a substrate, or specific letter in a message). Shannon was interested in the uncertainty (entropy) in the content of a message.

*Gibbs entropy:* $S_G = -k\int_\Omega p(x,t)\ln(p(x,t))dx$

where $\Omega$ = phase space, $p(x,t)$ = the density function of systems in $\Omega$ at time $t$, $p(x,t)dx$ = the number of systems whose states lie in the region $(x, x+dx)$. Gibbs was interested in an ensemble of many systems evolving over time.

*Clausius entropy:* $\Delta S = S_{final} - S_{initial} = \int(\delta Q_R/T)$

where $Q_R$ is the heat of a system undergoing a reversible change at temperature $T$. Also called the **thermodynamic entropy**.

**Appendix 2: Covariance and correlation.**

*Variance* $$s_x^2 = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}$$

($n$-1 implies sample variance, not population variance.)

*Standard Deviation* $$s_x = \sqrt{s_x^2}$$

*Covariance* $$s_{xy} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1}$$

(covariance is concerned only with the sign of the relationship, positive or negative, not the strength)

*Pearson correlation coefficient* $$r = \frac{s_{xy}}{s_x s_y}$$

$r$ = covariance/product of standard deviations

(correlation is the normalized covariance or strength of the covariance)