An Introduction to Information Theory and Entropy[1]

1. Measuring complexity

Researchers in the field of complexity face a classic problem: How can we tell that the system we are looking at is actually a complex system? Should we even be studying such a system? Of course, in practice, we will study the systems that interest us, for whatever reasons, so the problem identified above tends not to be a real problem. On the other hand, having chosen a system to study, we might well ask: *How complex is this system?*

In this more general context, we probably want at least to be able to compare two systems, and be able to say that system ***A*** is more complex than system ***B***. Eventually, we probably would like to have some sort of numerical rating scale.

We can't expect to be able to come up with a single universal measure of complexity. The best we are likely to have is a measuring system useful by a particular observer, in a particular context, for a particular purpose.

Our focus here will be on measures related to how surprising or unexpected an observation, or event, is. This approach has been described as *information theory*.

2. Some probability background

There are two main notions of *probability* of an event happening. These are:

[1] Based on references at the end of the manuscript.

Ivo D. Dinov UCLA Statistics
http://www.stat.ucla.edu/~dinov Courses & Students

A ***frequentist*** version of probability: In this version, we assume we have a set of possible events, each of which we assume occurs some number of times. Thus, if there are N distinct possible events $(x_1, x_2, \ldots, x_N)$, no two of which can occur simultaneously, and the events occur with frequencies $(n_1, n_2, \ldots, n_N)$, we say that the probability of event x_i is given by $P(x_i) = \frac{n_i}{\sum_{j=1}^{N} n_j}$. This definition has the nice property that $\sum_{i=1}^{N} P(x_i) = 1$.

An *observer relative* (***Bayesian***) version of probability: In this version, we take a statement of *probability* to be an assertion about the belief that a specific observer has of the occurrence of a specific event. Note that in this version of *probability*, it is possible that two different observers may assign different probabilities to the same event. Furthermore, the *probability* of an event is likely to change as we learn more about the event, or the context of the event.

In some cases, we may be able to find a reasonable correspondence between these two views of probability. In particular, we may sometimes be able to understand the *observer relative* version of the probability of an event to be an approximation to the *frequentist* version, and to view new knowledge as providing us a better estimate of the relative frequencies.

Some probability basics, where A and B are events:

$$P(\sim A) = P(A^c) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

We will often denote $P(A \cap B)$ by $P(A, B)$. If $P(A, B) = 0$, we say A and B are mutually exclusive events.

Conditional probability: $P(A \mid B)$ is the probability of A, given that we know B occurred. The joint probability of both A and B is given by: $P(A, B) = P(A \mid B)\, P(B)$. Since $P(A, B) = P(B, A)$, we have Bayes' Theorem:

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A),$$
$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}.$$

If two events A and B are such that $P(A \mid B) = P(A)$, we say that the events A and B are ***independent***. From Bayes' Theorem, we will also have that $P(B \mid A) = P(B)$, and furthermore, $P(A, B) = P(A \mid B)\, P(B) = P(A)\, P(B)$. This last equation is often taken as the definition of ***independence***.

We have in essence begun here the development of a mathematical methodology for drawing inferences about the world from uncertain knowledge. We could say that our observation of the coin showing heads gives us *information* about the world. We now develop a formal mathematical definition of the *information* content of an event, which occurs with a certain probability.

3. Axiomatic Development of Information Theory

We now want to develop a usable measure of the ***information*** we get from observing the occurrence of an event having probability ***p***. Our first reduction is to ignore any particular features of the event, and only observe whether or not it happened. In essence this means that we can think of the event as observance of a symbol whose probability of occurring is ***p***.

We will thus be defining the ***information*** in terms of the probability ***p***.

The following represent a set of reasonable axioms for an ***information*** measure $I(p)$:

1. Information is a non-negative quantity: $I(p) \geq 0$.
2. If an event has probability 1, we get no information from the occurrence of the event: $I(1)=0$.
3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two informations: $I(p1*p2) = I(p1) + I(p2)$.
4. We will want our *information* measure to be a **continuous** (and, in fact, monotonic) function of the probability (slight changes in probability should result in slight changes in *information*).

Corollaries of these axioms include:

1. $I(p^2) = I(p*p) = I(p) + I(p) = 2*I(p)$.
2. Thus, $I(p^n) = n*I(p)$, by induction.
3. $I(p) = I((p^{1/m})^m) = m * I(p^{1/m})$, so $I(p^{1/m}) = (1/m)*I(p)$ and thus, in general, $I(p^{n/m})=(n/m)*I(p)$.
4. And thus, by continuity, we get, for $0 < p \leq 1$, and $0 < a \Rightarrow I(p^a) = a*I(p)$.
5. From these, we can derive the nice property of *information* measure:

   $$\boldsymbol{I(p) = -\log_b(p) = \log_b(1/p)}$$

   for some log-base b. The base b determines the units we are using. Of course, we can change the units by changing the base, using the formulas, for

   $b_1, b_2, x > 0$, $\log_{b1}(x) = \log_{b2}(x) / \log_{b1}(b_2)$.

Thus, using different bases for the logarithm results in *information* measures which are just constant multiples of each other, corresponding with measurements in different units:

1. $\log_2$ units are *bits* (from *binary*)
2. $\log_3$ units are *trits* (from *trinary*)
3. $\log_e$ units are *nats* (from *natural logarithm*) (We commonly use $\ln(x)=\log_e(x)$)
4. $\log_{10}$ units are *Hartleys*, after *R.V.L. Hartleys, 1942*.

Unless we want to emphasize the units, we need not bother to specify the base for the logarithm, and simply write $log(p)$. Typically, we think in terms of $log_2(p)$.

Example: Suppose we flip a *fair* coin once. The outcomes are events H and T each with probability ½, and thus a single flip of a coin gives us $-log_2(1/2) = 1$, bit of *information* (whether the outcome is a H or T).

Flipping a fair coin n times (or, equivalently, independently flipping n fair coins) gives us $-log_2((1/2)^n) = log_2(2^n) = n * log_2(2) = n$ bits of *information*. We could randomly generate (see http://socr.stat.ucla.edu/) a sequence of 25 flips as, for example:

{HTHHTTHTHHHTHTTTHTHHHTHTT}

or, using 1 for H and 0 for T, the 25 bits

{10110010111010000101110100}.

We thus get the nice fact that n flips of a fair coin gives us n bits of *information*, and takes n binary digits to specify. That these two quantities are the same reassures us that we have done a reasonable axiomatic definition of *information measure*.

4. Some Entropy Theory

Suppose now that we have n symbols $\{a_1, a_2, ..., a_n\}$, and some source is providing us with a stream of these symbols. Suppose further that the source emits the symbols with probabilities $\{p_1, p_2, ..., p_n\}$, respectively. For now, we also assume that the symbols are emitted independently (successive symbols do not depend in any way on past symbols).

What is the average amount of *information* we get from each symbol we see in the stream? What we really want here is a weighted average. If we observe the symbol a_i, we will be getting $log(1/p_i)$ *information* from that particular observation. In a long run of (say N) observations, we will see (approximately) $N * p_i$ occurrences

of the symbol a_i (in the frequentist sense, that's what it means to say that the probability of seeing a_i is p_i). Thus, in the N (independent) observations, we will get total information I of

$$Total\ Information\ (TI) = \sum_{i=1}^{n} N \times p_i \times \log\left(\frac{1}{p_i}\right)$$

And therefore, the **average information** we get per symbol observed will be

$$I = \frac{1}{N}\sum_{i=1}^{n} N \times p_i \times \log\left(\frac{1}{p_i}\right) = \sum_{i=1}^{n} p_i \times \log\left(\frac{1}{p_i}\right)$$

Note that $\lim_{x \to 0} x \log\left(1/x\right) = 0$, so we can, for our purposes, define $p_i * log(1/p_i)$ to be 0 when $p_i = 0$. This brings us to a fundamental definition. This definition is essentially due to Shannon in 1948, in the seminal papers in the field of information theory. As we have observed, we have defined *information* strictly in terms of the probabilities of events. Therefore, let us suppose that we have a set of probabilities (a probability distribution $P = \{p_1, p_2, ..., p_n\}$).

Definition: We define the (Shannon-Wiener) *entropy* of the distribution P by:

$$H(P) = \sum_{k=1}^{n} p_k \times log\left(\frac{1}{p_k}\right) = -\sum_{k=1}^{n} p_k \times log\left(p_k\right) \qquad (1)$$

There is an obvious generalization of the entropy for continuous, rather than discrete, probability distribution $P(x)$:

$$H(P) = -\int P(x) \times log\left(P(x)\right) dx \qquad (2)$$

Another way to think about this is in terms of *expected value*. Given a PDF/PMF $P(x)$ we can define the *expected value* of an associated function $F(x)$ by:

$$E(F(X)) = \int F(x) \times P(x) dx .$$

With this definition, we have that: $H(P) = E(I(p))$. In other words, the *entropy* of a probability distribution is just the *expected value* of the *information* measure of that distribution.

We'll discuss the following few **important points**:

1. What properties does the function $H(P)$ have? For example, does it have extrema, and if so where?
2. Is *entropy* a reasonable name for this? In particular, the name *entropy* is already in use in physics/thermodynamics. How are these uses of the term related to each other?
3. What can we do with this new tool?

5. The Gibbs inequality

First, note that the (natural-log) function **ln(x)** has derivative $1/x$. From this, we find that the tangent to $ln(x)$ at $x=1$ is the line $y = x - 1$. Further, since $ln(x)$ is concave down, we have, for $x>0$, that $ln(x) \leq x - 1$, with equality only when $x = 1$.

Now, given two probability distributions,

$$P = \{p_1, p_2, \ldots, p_n\} \text{ and } Q = \{q_1, q_2, \ldots, q_n\},$$

where $p_k, q_k \geq 0$ and $\sum_{k=1}^{n} p_k = \sum_{k=1}^{n} q_k = 1$, we have

$$\sum_{k=1}^{n} p_k \ln\left(\frac{q_k}{p_k}\right) \leq \sum_{k=1}^{n} p_k \left(\frac{q_k}{p_k} - 1\right) = \sum_{k=1}^{n} (q_k - p_k) = 1 - 1 = 0, \qquad (3)$$

with equality only when $p_k = q_k$ for all k. It is easy to see that the inequality actually holds for any log-base, not just base e.

We can now use the *Gibbs inequality* to find the probability distribution, which maximizes the *entropy function*. Suppose $P = \{p_1, p_2, \ldots, p_n\}$ is a probability distribution. We have

$$\begin{aligned} H(P) - \log(n) &= \sum_{k=1}^{n} \left[p_k \log\left(1/p_k\right) \right] - \log(n) \\ &= \sum_{k=1}^{n} \left[p_k \log\left(1/p_k\right) \right] - \log(n) \sum_{k=1}^{n} p_k \\ &= \sum_{k=1}^{n} \left[p_k \left(\log\left(1/p_k\right) - \log(n) \right) \right] \\ &= \sum_{k=1}^{n} \left[p_k \log\left(\frac{1/n}{p_k}\right) \right] \\ &= \sum_{k=1}^{n} \left[p_k \log\left(\frac{q_k}{p_k}\right) \right] \leq 0 \end{aligned} \qquad (4)$$

with equality only when $p_k = \frac{1}{n}$ for all k. The last step is the application of the Gibbs inequality (3). What this means is that:

$$0 \leq H(P) \leq \log(n) \qquad (5).$$

In particular, if for some k_o, $p_{Ko}=1$ and $p_k=0$ for all $k \neq k_o$, we have $H(P) = 0$. On the other end of the spectrum, the entropy $H(P) = \log(n)$ (maximum possible entropy) only when all of the events (outcomes) have the same probability, $p_k = \frac{1}{n}$. That is, the maximum of the entropy function is $log($of the number of possible events/outcomes), and occurs when all the events are equally likely. This illustrates the entropy as a measure of uncertainty ➔ ***high entropy means lots of uncertainty*** and ***low entropy yields high certainty*** about the outcome of the process/experiment.

Example: How much information is obtained by a single neuropsychiatric (NP) test? First, the maximum information occurs if all outcomes/scores/results of the NP test have equal probability to be observed (e.g., in an AD vs. NoAD test, on average half the subjects should end up having AD and the other have should not have dementia) if we want to **maximize the information given by the NP test**. Here are several common situations indicating the conditions for obtaining the maximum information from a single NP test result, we use equation (1):

Experiment/Process Type	Max Information [plug in equation (1) $p_k=1/n$]
Binary Test (AD vs. NoAD)	1 bit = $\log_2(2)$
Five-level test results: [Extreme(E), Severe(S), Moderate(M), MCI, Normal(N)]	2.3 bits = $\log_2(5)$
Twelve-level test results: E, E^-, S^+, S, S^-, …, MCI, N	3.6 bits = $\log_2(12)$

Thus, using +/–'s gives the patients/doctors about 1.3 more bits of information, per test-level, than without using +/–'s, and about 2.6 bits per grade more than binary (AD vs. NoAD) type test results. Which is naturally expected as we actually have more information available in addition to presence or absence of AD NP symptoms.

Example: The genetic code provides us with sequences constructed from 4 symbols (A, C, G, T). The maximum average information per symbol is **$\log_2(4)=2$ bits**. If the source provides codons (blocks of 3 of these symbols), then the maximum average information is 6 bits per block, as $I(p^k) = kxI(p)$, $p = ¼$. If we used different units, e.g., log_{10}, the max entropy will be *4.159 nats* per block.

Remarks:

1. First, these definitions of *information* and *entropy* may not match with some other uses of the terms. For example, if we know that a source will, with equal probability, transmit either the complete text of Hamlet or the complete text of Macbeth (and nothing else), then receiving the complete text of Hamlet provides us with precisely 1 bit of information. Suppose a book contains ASCII characters. If the book is to provide us with *information* at the maximum rate, then each ASCII character will occur with equal probability – it will be a random sequence of characters.

2. Second, it is important to recognize that our definitions of *information* and *entropy* depend only on the probability distribution. In general, it won't make sense for us to talk about the *information* or the *entropy* of a source without specifying its probability distribution.

3. Beyond that, it can certainly happen that two different observers of the same data stream have different models of the source, and thus associate different probability distributions to the source. **The two observers will then assign different values to the *information* and *entropy* associated with the source**. This observation accords with our intuition: two people listening to the same tune can get very different information from the music. For example, without appropriate music background, one person may get excited, another one may get bored, yet another one may fall asleep. The first listener who enjoys music may assign non equal probabilities to each sound/chord/epochs as he/she may anticipate much of where the tune goes. On the contrary, the musical composition may sound as (random) unsynchronized collection of chords (e.g., abstract jazz) and hence the amount of information comprehended by this listener will be significantly higher, as the probabilities that he/she assigns to each note/chord are roughly equal.

A physical Example (Gas Particles): Let us consider a simple model for an idealized gas. Suppose a cubical volume V contains gas made up of N point particles. Assume also that through some mechanism, we can determine the location of each particle sufficiently well as to be able to locate it within a box with sides 1/100 of the sides of the containing volume V. There are 10^6 of these small boxes within V.

A frequentist probability model for this system is obtained by measuring the number of particles in each of the 10^6 small boxes at one fixed time, and assigning a probability p_k of finding a gas particle in the small box by counting the number of particles n_k in the box, and dividing by N. That is, $p_k = n_k/N$. From this probability distribution model, we can calculate the entropy:

$$H(P) = \Sigma_{k=1}^{N} p_k \log\left(\frac{1}{p_k}\right)$$
$$= \Sigma_{k=1}^{10^6} \frac{n_k}{N} \log\left(\frac{N}{n_k}\right)$$

There are a couple of special cases to consider (representing the extrema of the values of the entropy).

1. If the particles are *evenly distributed* among the 10^6 boxes, then we will have that each $n_k=N/10^6$, and in this case the entropy will be:

$$H(P) = \Sigma_{k=1}^{10^6} \frac{1}{10^6} \log\left(10^6\right) = \log\left(10^6\right) = 6 \times \log(10)$$

   This case obviously presents a maximum entropy configuration.

2. At the opposite side of the spectrum we have all 10^6 particles sitting in exactly one small box, and the entropy of each of those configurations is:

$$H(P) = \Sigma_{k=1}^{10^6} p_k \log\left(\frac{1}{p_k}\right) = 0,$$
$$as\ p_{k_o} = 1\ and\ p_k = 0\ for\ k \neq k_o$$

   This case obviously presents a minimum entropy configuration.

Notice that these two calculated entropies of the system depend in a strong way on the relative scale of measurement. For example, if the particles are evenly distributed, and we increase our accuracy of measurement by a factor of *10* (i.e., if each small box is *1/1000* of the side of *V*), then the calculated maximum entropy, in the first case, will be $log(10^9)$ instead of $log(10^6)$.

In addition, for physical systems, we know that quantum limits (e.g., Heisenberg uncertainty relations) will give us a bound on the accuracy of our measurements, and thus a more or less natural scale for doing entropy calculations. On the other hand, for macroscopic systems, we are likely to find that we can only make relative rather than absolute entropy calculations.

Third, suppose we generalize our model slightly, and allow the particles to move about within *V*. A *configuration* of the system is then simply a list of 10^6 numbers $1 \le b_k \le N$ (i.e., a list of the numbers of particles in each of the small boxes). Suppose that the motions of the particles are such that for each particle, there is an equal probability that it will move into any given new small box during one (macroscopic) time step. How likely is it that at some later time we will find the system in a high entropy configuration? How likely is it that if we start the system in a low entropy configuration, it will stay in a low entropy configuration for an appreciable length of time? If the system is not currently in a maximum entropy configuration, how likely is it that the entropy will increase in succeeding time steps (rather than stay the same or decrease)?

Recall the binomial coefficients (number of arrangements of n objects taken k at a time, where the order does not matter, combinations) $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$. And the Stirling's approximation of $n!$: $n! = \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$, for large n.

There are 10^6 configurations with all the particles sitting in exactly one small box, and as we shown above, the entropy of each of those configurations is $H(P)=0$. These are obviously minimum entropy configurations.

If we now consider pairs of small boxes, the number of configurations with all the particles evenly distributed between two boxes is $\binom{10^6}{2} = 5\times 10^{11}$, which is large. The entropy of each of these configurations is

$$H(P \mid All-evenly-in-2-boxes) = \frac{1}{2}\log(2) + \frac{1}{2}\log(2) = \log(2).$$

The total number of system configurations, in terms of the number of particles within a small box, is at least $5*10^{11} + 10^6$. If we start the system in a configuration with entropy *0*, then the probability that at some later time it will be in a configuration with entropy $H(P) \geq \log(2)$ will be larger than $[5*10^{11}] / [5*10^{11} + 10^6] > 1-10^{-5}$, as $P(Event) = \frac{|\,Event\,|}{|\,S\,|}$ and $\frac{a+x}{a+b+x} > \frac{a}{a+b}, \forall a,b,x>0$. Here, *a=P(All-in-2-boxes)*, *b=P(all-in one-box)* and *x=P(all-in-more-than-2-boxes)*.

As an example at the other end, consider the number of configurations with the particles distributed almost equally, except that half the boxes are short by one particle, and the rest have one extra particle. The number of such configurations is:

$$\binom{10^6}{10^6/2} \approx 10^{3*10^5}$$

Each of these configurations has entropy essentially approximately equal to $log(10^6)$. From this, we can conclude that if we start the system in a configuration with entropy

of *0* (i.e., all particles in one box), the probability that later it will be in a higher entropy configuration will be larger than $1-10^{-3*10^5} \approx 1$.

Similar arguments (with similar results in terms of probabilities) can be made for starting in any configuration with entropy appreciably less than $log(10^6)$ (the maximum). In other words, it is overwhelmingly probable that as time passes, macroscopically, the system will increase in entropy until it reaches the maximum.

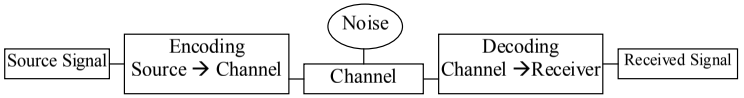
In many respects, these general arguments can be thought of as a *proof* (or at least an explanation) of a version of the **second law of thermodynamics**:

*Given any macroscopic system, which is free to change configurations, and given any configuration with entropy less than the maximum, there will be overwhelmingly many more accessible configurations with higher entropy than lower entropy, and thus, with probability indistinguishable from **1**, the system will (in macroscopic time steps) successively change to configurations with higher entropy until it reaches the maximum.*

7. Shannon's communication theory

In some classic 1948 papers, Claude Shannon laid the foundations for contemporary *information*, *coding*}, and *communication* theory.
He developed a general model for

communication systems, and a set of theoretical tools for analyzing such systems His basic model consists of three parts: a sender (or source), a channel, and a receiver (or



sink). In addition, the model also includes encoding and decoding elements, and noise within the communication channel.

In Shannon's discrete model, it is assumed that the source provides a stream of symbols selected from a finite alphabet $A = \{a_1, a_2, \ldots, a_n\}$, which are then encoded. The code is sent through the channel (and possibly corrupted by noise). At the other end of the channel, the receiver will decode, and derive information from the sequence of symbols.

Given a source of symbols and a channel with noise (in particular, a probability model for these elements), we can talk about the **capacity** of the channel. The general model Shannon worked with involved two sets of symbols, the *input symbols* and *the output symbols*. Let us say the two sets of symbols are $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$. Note that we do not necessarily assume the same number of symbols in the two sets. Given the noise in the channel, when symbol b_k comes out of the channel, we can not be sure which a_l, was put in. The channel is characterized by the set of probabilities
$\{P(a_l \mid b_k)\}_{l,k}$.

We can then consider various related information and entropy measures. First, we can consider the information we get from observing a symbol b_k. Given a probability model of the source, we have an *a priori* estimate $P(a_l)$ that symbol a_l will be sent next. Upon observing b_k, we can revise our estimate to $P(a_l \mid b_k)$. The change in our information (the *mutual information*) will be given by:

$$I(a_l\,; b_k) = \log\left(\frac{1}{P(a_l)}\right) - \log\left(\frac{1}{P(a_l \mid b_k)}\right) = \log\left(\frac{P(a_l \mid b_k)}{P(a_l)}\right) \qquad (6)$$

We have the properties of this functional:

1. $I(a_l; b_k) = I(b_k; a_l)$
2. $I(a_l; b_k) = \log(P(a_l \mid b_k)) + I(a_l)$
3. $I(a_l; b_k) \leq I(a_l)$

If a_l and b_k are independent (i.e., $P(a_l \mid b_k) = P(a_l)$), then $I(a_l; b_k) = 0$. What we often times want is to average the mutual information over all the symbols:

$$\begin{aligned} I(A; b_k) &= \sum_{l \in A} P(a_l \mid b_k) \times I(a_l\,; b_k) \\ &= \sum_{l \in A} P(a_l \mid b_k) \times \log\left(\frac{P(a_l \mid b_k)}{P(a_l)}\right) \\ I(a_l\,; B) &= \sum_{k \in B} P(b_k \mid a_l) \times I(b_k\,; a_l) \\ &= \sum_{k \in B} P(b_k \mid a_l) \times \log\left(\frac{P(b_k \mid a_l)}{P(b_k)}\right) \end{aligned} \qquad (7)$$

Therefore

$$\begin{aligned} I(A; B) &= \sum_{l \in A} P(a_l) \times I(a_l\,; B) \\ &= \sum_{l \in A} \sum_{k \in B} P(a_l) \times P(a_l \mid b_k) \times \log\left(\frac{P(a_l\,; b_k)}{P(a_l) \times P(b_k)}\right) \\ &= \sum_{l \in A} \sum_{k \in B} P(a_l\,; b_k) \times \log\left(\frac{P(a_l\,; b_k)}{P(a_l) \times P(b_k)}\right) \end{aligned} \qquad (8)$$

We have the properties: $I(A; B) \geq 0$ and $I(A; B) = 0$, if and only if A and B are independent.

Definition: Conditional Entropy is defined by

$$H(A \mid B) = \sum_{l \in A} \sum_{k \in B} P(a_l \mid b_k) \times \log\left(\frac{1}{P(a_l \mid b_k)}\right) \qquad (9)$$

Notice that:

$$H(A) = \sum_{l \in A} P(a_l) \times \log\left(\frac{1}{P(a_l)}\right)$$
$$H(B) = \sum_{k \in B} P(b_k) \times \log\left(\frac{1}{P(b_k)}\right) \qquad (10)$$
$$H(A,B) = \sum_{l \in A} \sum_{k \in B} P(a_l; b_k) \times \log\left(\frac{1}{P(a_l; b_k)}\right)$$

And $H(A, B) = H(A) + H(B|A) = H(B) + H(A|B)$;

$I(A; B) = H(A) + H(B) - H(A, B) = H(A) - H(A|B) = H(B) - H(B|A) \geq 0.$

If we are given a channel, we could ask what is the maximum possible information can be transmitted through the channel. We could also ask what mix of the symbols $\{a_l\}$ we should use to achieve the maximum bandwidth. In particular, using the definitions above, we can define the **Channel Capacity**, ***C***, to be:

$$C = \max_{P(a)} \left(I(A; B) \right)$$

We have the nice property that if we are using the channel at its capacity, then for each of the a_l, $I(a_l;B) = C$, and thus, we can maximize channel use by maximizing the use for each symbol independently.

Theorem: [Shannon] For any channel, there exist ways of encoding input symbols such that we can simultaneously utilize the channel as closely as we wish to the capacity, and at the same time have an error rate as close to zero as we wish.

This is actually quite a remarkable theorem. We might *naively* guess that in order to minimize the error rate, we would have to use more of the channel capacity for error

detection/correction, and less for actual transmission of information. Shannon showed that it is possible to keep error rates low and still use the channel for information transmission at (or near) its capacity.

Unfortunately, Shannon's proof has a couple of downsides. The **first** is that the proof is **non-constructive**. It doesn't tell us how to construct the coding system to optimize channel use, but only tells us that such a code exists. The **second** is that in order to use the capacity with a low error rate, we **may have to encode very large blocks of data**. This means that if we are attempting to use the channel in real-time, there may be time lags while we are filling buffers. There is thus still much work possible in the search for efficient coding schemes.

Among the things we can do is look at natural coding systems (such as, for example, the DNA coding system, or neural systems) and see how they use the capacity of their channel. It is not unreasonable to assume that evolution will have done a pretty good job of optimizing channel use.

8. Application of Entropy to modeling DNA sequences

Let us apply some of these ideas to the (general) problem of analyzing genomes. We can start with an example such as the comparatively small genome of *Escherichia coli*, strain K-12, substrain MG1655, version M52. This example has the convenient features:

1. It has been completely sequenced.
2. The sequence is available for downloading: http://www.genetics.wisc.edu/
3. Annotated versions are available for further work.
4. It is large enough to be interesting (somewhat over 4 mega-bases, or 4 million nucleotides), but not so huge as to be completely unwieldy.
5. This data file begins with:

>gb|U00096|U00096 Escherichia coli
K-12 MG1655 complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAG
AGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGG
TGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACC
AAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA

In this exploratory project, our goal will be to apply the *information* and *entropy* ideas outlined above to genome analysis. Our first step is to generate a **random genome** of comparable size to compare things with. We can use SOCR, Excel, SAS, R, C++, Java or other languages/programs to generate a file containing a random sequence of about 4 million letters *A, C, G, T*. In the actual genome, these letters stand for the nucleotides **adenine (A), cytosine (C), guanine (G), and thymine (T).**

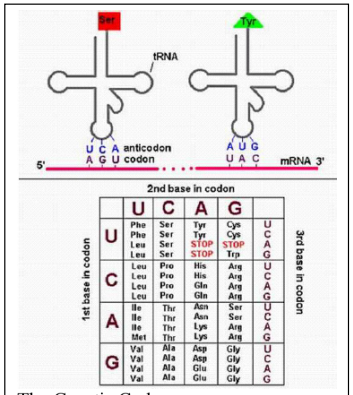
There are other approaches to this process, e.g., randomly shuffling an actual observed genome (thus maintaining the relative proportions of As, Cs, Gs, and Ts). Part of the justification for this methodology is that actual (identified) coding sections of DNA tend to have a ratio of $\frac{|C|+|G|}{|A|+|T|} \neq 1$.

One can hope that important stretches of DNA will have entropy different from other stretches. Of course, as noted above, the entropy measure depends in an essential

way on the probability model attributed to the source. We will want to try to build a model that catches important aspects of what we find interesting or significant.

We will want to use our knowledge of the systems in which DNA is embedded to guide the development of our models. On the other hand, we probably don't want to constrain the model too much. Remember that information and entropy are measures of unexpectedness. If we constrain our model too much, we won't leave any room for the unexpected!

We know, for example, that simple repetitions have low entropy. But if the code being used is redundant (sometimes called *degenerate*), with multiple encodings for the same symbol (as is the case for DNA codons), what looks to one observer to be a random stream may be recognized by another observer (who knows the code) to be a simple repetition.



The Genetic Code.
http://www.accessexcellence.org/

The coding sequences for peptides and proteins are encoded via codons, that is, by sequences of blocks of triples of nucleotides. Thus, for example, the codon ***AGC*** on mRNA (messenger RNA) codes for the amino acid ***serine*** (or, if we happen to be reading in the reverse direction, ***CGA***, it might code for ***alanine***). On DNA, *AGC* codes for *UCG* or *CGA* on the mRNA, and thus could code for *cysteine* or *arginine*.

Amino acids specified by each codon sequence on mRNA.

A = adenine	G = guanine	C = cytosine	T = thymine	U = uracil

Amino acid key for the above Figure:

Ala = Alanine	Arg=Arginine	Asn=Asparagine	Asp=Aspartic acid
Cys=Cysteine	Gln=Glutamine	Glu=Glutamic acid	Gly=Glycine
His=Histidine	Ile=Isoleucine	Leu=Leucine	Lys=Lysine
Met=Methionine	Phe=Phenylalanine	Pro=Proline	Ser=Serine
Thr=Threonine	Trp=Tryptophane	Tyr=Tyrosine	Val=Valine

As a first step consider each of the three-nucleotide codons as a distinct symbol. We can then **take a chunk of genome and estimate the probability of occurrence of each codon by simply counting and dividing by the length**. At this level, we are assuming we have no knowledge of where codons start, and so in this model, we assume that *readout* could begin at any nucleotide. We thus use each three adjacent nucleotides.

For example, given the DNA chunk (length=34):

AGCTTTTCATTCTGACTGCAACGGGCAATATGTC

Our codon count yields (in lexicographical order!):

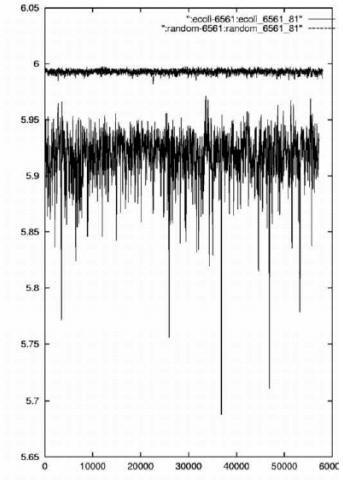
AAT 1 AAC 1 ACG 1 ACT 1 AGC 1
ATA 1 ATG 1 ATT 1 CAA 2 CAT 1
CGG 1 CTG 2 CTT 1 GAC 1 GCA 2
GCT 1 GGC 1 GGG 1 GTC 1 TAT 1
TCA 1 TCT 1 TGA 1 TGC 1 TGT 1
TTC 2 TTT 2

Ivo D. Dinov
http://www.stat.ucla.edu/~dinov
UCLA Statistics
Courses & Students

We can then estimate the entropy of this sequence by:

$$\sum_{l=1}^{27} p_l \times \log_2\left(\frac{1}{p_l}\right) = 4.7\ bits.$$

The maximum possible entropy for this chunk would be: $\log_2(32) = 5$ bits.



Comparison of the entropy for *E. coli* genome (spiky curve) and a random genome sequence.

Suppose we want to find interesting sections (features) in the genome. As a starting place, we can slide a window over the genome, and estimate the entropy within the window. **The plot below shows the entropy estimates for the *E. coli* genome, within a window of size 3^8=6561. The window is slid in steps of size 3^4=81**. This results in 57,194 snapshots, one for each placement of the window. For comparison, the values for a random genome are also shown.

At this level, we can make the simple observation that **the actual genome values are quite different from the comparative random string**. The values for *E. coli* range from about 5.8 to about 5.96, while the random values are clustered quite closely above 5.99 (the maximum possible is $\log_2(4^3)=\log_2(64) = 6$. Recall, ***I(p) = -log$_b$(p) = log$_b$(1/p)*** and ***p=1/64***).

9. Measures of Dimensionality and relation to Entropy

A useful generalization of entropy (as a measure of complexity) was developed by the Hungarian mathematician *A. Renyi*. The ***Renyi Entropy*** is defined as the moments of order q of a probability distribution $P=\{p_i\}$:

$$S_q = \frac{1}{q-1} \log \Sigma_i \, p_i^q \qquad (11)$$

Taking the limit as $q \to 1$, we get:

$$S_1 = \Sigma_i \, p_i \log\left(\frac{1}{p_i}\right)$$

The last expression is exactly the entropy we previously defined. So, S_q is a generalized entropy for any real number q. The limit of S_1, as $q \to 1$ is because:

$$\lim_{q \to 1} \frac{\log \Sigma_i \, p_i^q}{q-1} = \lim_{q \to 1} \frac{\frac{\partial}{\partial q}\left(\log \Sigma_i \, p_i^q\right)}{\frac{\partial}{\partial q}(q-1)}$$

$$\text{and } \frac{\partial}{\partial q}\left(p_i^q\right) = p_i^q \log(p_i) \xrightarrow{q \to 1} p_i^q \log(p_i)$$

Using the ***Renyi Entropy***, we can then define a generalized *dimension* associated with a data set. Suppose a data set is distributed among bins of **diameter *r***, we can let p_i be the probability that a data item falls in the i^{th} bin (estimated by counting the data elements in the bin, and dividing by the total number of items). We can then, define a **dimension** (for each q):

$$D_q = \lim_{r \to 0} \left(\frac{1}{q-1} \frac{\log \Sigma_i \, p_i^q}{\log(r)} \right) \qquad (12)$$

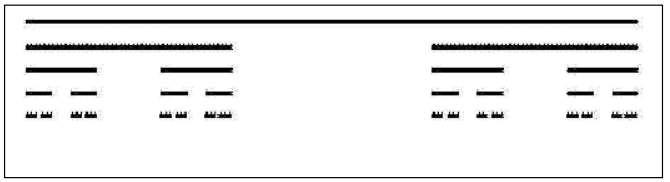
Why do we call this a generalized *dimension*? Consider first $D_{q=0}$. We define $p_i^0 = 0$, when $p_i=0$. Also, let N_r be the number of non-empty bins of diameter r it takes to cover the data set. Then we have:

$$D_0 = \lim_{r \to 0} \left(\frac{\log \Sigma_i \, p_i}{\log\left(\frac{1}{r}\right)} \right) = \lim_{r \to 0} \left(\frac{\log N_r}{\log\left(\frac{1}{r}\right)} \right)$$

Definition: D_0 is the ***Hausdorff dimension***, which is sometimes called the *fractal dimension* of the set.

Examples:

1. **1D:** Consider the unit interval $[0,1]$. Let $r_k = \frac{1}{2^k}$. Then $N_{rk} = 2^k$, and
$$D_0 = \lim_{r \to 0} \left(\frac{\log\left(2^k\right)}{\log\left(2^k\right)} \right) = 1$$
2. **2D:** Consider the unit square $[0,1]x[0,1]$. Again, let $r_k = \frac{1}{2^k}$. Then $N_{rk} = 2^{2k}$, and $D_0 = \lim_{r \to 0} \left(\frac{\log\left(2^{2k}\right)}{\log\left(2^k\right)} \right) = 2$.
3. **1D⟷2D?** Consider the **Cantor set**: The construction of the Cantor set is done by induction.



The Cantor set is what remains from the interval after we have removed middle thirds countably many times. It is an uncountable set, with measure (**length**) **0**.

For this set we will let $r_k = \frac{1}{3^k}$. Then $N_{rk} = 2^k$, and

$$D_O = \lim_{r \to 0} \left(\frac{\log\left(2^k\right)}{\log\left(3^k\right)} \right) = \frac{\log(2)}{\log(3)} = 0.631$$

The Cantor set is a traditional example of a **fractal**. It is self-similar, and has Hausdorff dimension of *0.631*, which is strictly greater than its (integer) *topological dimension 0*.

Some nonlinear dynamical systems have trajectories which are locally the product of a Cantor set with a manifold (i.e., *Poincare* sections are generalized Cantor sets).

Properties of D_q:

1. *Monotonicity*: If $q_1 \le q_2$, then $D_{q_1} \le D_{q_2}$.
2. *Fractal Calculations:* If the set is strictly self-similar with equal probabilities $p_i = 1/N$, then the calculations are trivial and we do not need to take the limit as $r \to 0$, since

$$D_q = \lim_{r \to 0} \left(\frac{1}{q-1} \frac{\log\left(N \times \left(1/N\right)^q \right)}{\log(r)} \right) = \frac{\log(N)}{\log\left(\frac{1}{r}\right)} = D_o$$

This is the case, for example, for the Cantor set.

3. *Information Dimension*:

$$D_1 = \lim_{r \to 0} \left(\frac{\Sigma_i \, p_i \, \log\left(1/p_i\right)}{\log(r)} \right) \qquad (13)$$

The numerator is just the entropy of the probability distribution.

4. *Correlation Dimension*: This dimension is related to the probability of finding two elements of the set within a distance r of each other.

$$D_2 = \lim_{r \to 0} \left(\frac{\log\left(\Sigma_i \, p_i^2\right)}{\log(r)} \right) \qquad (14)$$

10. Mutual Information

Consider a system with the input X and output Y (X, Y random variables). How can we measure the uncertainty about X after observing Y? Let's define **conditional entropy of X with given Y**:

$$H(X \mid Y) = H(X, Y) - H(Y).$$

$$H(X,Y) = -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \log(f_{X,Y}(x,y)) dx dy \qquad (14)$$

$$H(Y) = -\int_{-\infty}^{\infty} f_Y(y) \log(f_Y(y)) dy$$

The conditional entropy $H(X \mid Y)$ represents the amount of the uncertainty remaining about the system input X after the system output Y has been observed. The next claim

is intuitively clear: The difference $H(X)-H(X|Y)$ must represent the uncertainty about the system input that is resolved by observing the system output:

$$I(X,Y)=H(X)-H(X|Y) \quad (15)$$

Mutual information can qualitatively be thought of as a measure of how well one image explains the other, and is maximized at the optimal alignment. It can be expressed in the following form:

$$I(A,B)=H(A)+H(B)-H(A,B)=\sum_a\sum_b P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \quad (16)$$

The conditional probability *P(b|a)* is the probability that *B* will take the value *b* given that *A* has the value *a*. The **conditional entropy** is therefore the average of the entropy of *B* for each value of *A*, weighted according to the probability of getting that value of *A*:

$$H(B|A)=-\sum_{a,b}P(a,b)\log P(b|a)=H(A,B)-H(A) \quad (17)$$

Thus the equation for mutual information can be rewrite as:

$$I(A,B)=H(A)-H(B|A)=H(B)-H(A|B) \quad (18)$$

Registration by maximization of mutual information therefore involves finding the transformation that makes image *A* the best possible predictor for image *B* within the region of overlap.

The advantage of mutual information over joint entropy is that it includes the entropies of the separate images. Mutual information and joint entropy are computed for the overlapping parts of the images and the measures are therefore sensitive to the size and the contents of overlap. A problem that can occur when using joint entropy on its own is that low values (normally associated with a high degree of alignment) can be found for complete misregistrations. For example, when transforming one

image to such an extent that only an area of background overlaps for the two images, the joint histogram will be very sharp, there is only one peak from background.

Mutual information is better equipped to avoid such problems, because it includes the marginal entropies *H*(*A*) and *H*(*B*). These will have low values when the overlapping part of the images contains only background and high values when it contains anatomical structure. The marginal entropies will thus balance the measure somewhat by penalizing for transformations that decrease the amount of information in the separate images. Consequently, mutual information is less sensitive to overlap than joint entropy, although not completely immune.

11. Normalized Mutual Information

The size of the overlapping part of the images influences the mutual information measure in two ways. First of all, a decrease in overlap decreases the number of samples, which reduces the statistical power of the probability distribution estimation. Secondly, the mutual information measure may actually increase with increasing misregistration (which usually coincides with decreasing overlap). This can occur when the relative areas of object and background even out and the sum of the marginal entropies increases, faster than the joint entropy. Studholme *et al.* proposed a normalized measure of mutual information, which is less sensitive to changes in overlap:

$$NMI=\frac{H(A)+H(B)}{H(A,B)} \quad (19)$$

Maes et al. have suggested the use of the Entropy Correlation Coefficient (ECC) as another form of normalized mutual information. *NMI* and *ECC* are related in the following manner:

$$ECC=\frac{2I(A,B)}{H(A)+H(B)}=2-2/NMI$$

12. Kullback-Leibler divergence

It could be useful to define a **distance** between two vector distributions. If $f_{\mathbf{X}}(\mathbf{x})$ is a distribution of the vector $\mathbf{X}$, and $g_{\mathbf{X}}(\mathbf{x})$ is a different distribution of the $\mathbf{X}$, than the distance between these distributions can be written as

$$D_{f(x)\|g(x)} = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \log \frac{f_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})} d\mathbf{x} \qquad (20)$$

For a single image, the entropy is normally calculated from the image intensity histogram in which the probabilities $\{p_1, p_2, p_3, \ldots, p_n\}$ are the histogram entries. If all voxels in an image have the same intensity a, the histogram contains a single non-zero element with probability of 1, and the entropy of this image is 0.

If this uniform image were to include some noise, then the histogram will contain a cluster of non-zero entries around a peak at the average intensity value. So the addition of noise to the image tends to equalize the probabilities, which increases the entropy. One consequence is that interpolation of an image may smooth the image, which can reduce the noise, and consequently 'sharpen' the histogram. This sharpening of the histograms reduces entropy.

Application of entropy for intramodality image registration: The goal now is to calculate the **entropy of a difference image**. If two perfectly aligned identical images are subtracted the result is an entirely uniform image that has zero entropy. For **two images that differ by noise, the histogram will be blurred, giving higher entropy**, as is shown in the Figure below. Any misregistration, however, will lead to edge artifacts that further increase the entropy. **Very similar images can therefore be registered by iteratively minimizing the entropy of the difference image**.

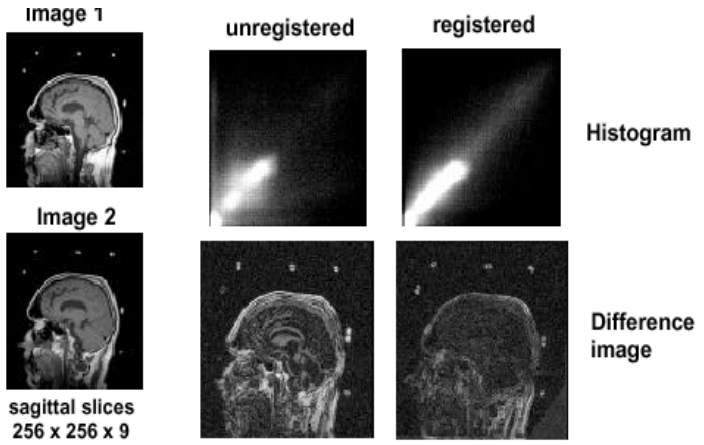


Figure 1 Histogram is blurred if the images are not aligned (Guido Gerig)

12. Joint entropy

Joint entropy measures the amount of information in the two images combined. If these two images are totally unrelated, then the joint entropy will be the sum of the entropies of the individual images. The more similar the images are, the lower the joint entropy compared to the sum of the individual entropies. The concept of joint entropy can be visualized using a joint histogram calculated from the images, as shown in Figure below. For all voxels in the overlapping regions of the images we plot the intensity of this voxel in image *A* against the intensity of the corresponding voxel in image *B*. The joint histogram can be normalized by dividing by the total number of voxels *N*, and regarded as a joint probability density function (PDF) *P(a; b)* of images *A* and *B*. The number of elements in the PDF can either be determined by the range of intensity values in the two images, or from a partitioning of the intensity space into "bins".

Definition: The **joint entropy *H(A,B)*** is therefore given by:

$$H(A,B) = -\sum_{a}\sum_{b} P(a,b)\log P(a,b) \qquad (21)$$

where *a, b* represent the original image intensities or the selected intensity bins. As can be seen from the Figure below, the joint histograms disperse or *blur* with increasing misregistration and thus increases the entropy.

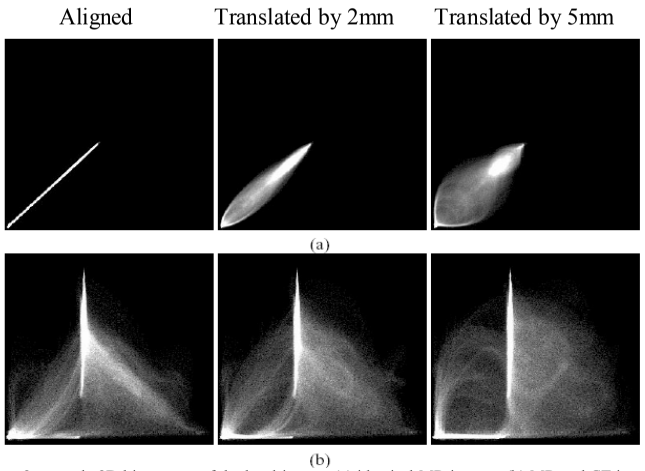


Figure 2 example 2D histograms of the head images (a) identical MR images, (b) MR and CT images (Hill et al., "Voxel similarity measures for automated image registration," *Visualization in Biomedical Computing 199*4, vol. Proc. SPIE 2359, pp. 205–216, 1994.)

Ivo D. Dinov UCLA Statistics
http://www.stat.ucla.edu/~dinov Courses & Students

13. References:

[1] Brillouin, L., Science and information theory Academic Press, New York, 1956.

[2] Brooks, Daniel R., and Wiley, E. O., Evolution as Entropy, Toward a Unified Theory of Biology, Second Edition, University of Chicago Press, Chicago, 1988.

[3] Campbell, Jeremy, Grammatical Man, Information, Entropy, Language, and Life, Simon and Schuster, New York, 1982.

[4] Cover, T. M., and Thomas J. A., Elements of Information Theory, John Wiley and Sons, New York, 1991.

[5] DeLillo, Don, White Noise, Viking/Penguin, New York, 1984.

[6] Feller, W., An Introduction to Probability Theory and Its Applications, Wiley, New York,1957.

[7] Feynman, Richard, Feynman lectures on computation, Addison-Wesley, Reading, 1996.

[8] Gatlin, L. L., Information Theory and the Living System, Columbia University Press, New York, 1972.

[9] Haken, Hermann, Information and Self-Organization, a Macroscopic Approach to Complex Systems, Springer-Verlag, Berlin/New York, 1988.

[10] Hamming, R. W., Error detecting and error correcting codes, Bell Syst. Tech. J. 29 147, 1950.

[11] Hamming, R. W., Coding and information theory, 2nd ed, Prentice-Hall, Englewood Cli_s, 1986.

[12] Hill, R., A first course in coding theory Clarendon Press, Oxford, 1986.

[13] Hodges, A., Alan Turing: the enigma Vintage, London, 1983.

[14] Hofstadter, Douglas R., Metamagical Themas: Questing for the Essence of Mind and Pattern, Basic Books, New York, 1985

[15] Jones, D. S., Elementary information theory Clarendon Press, Oxford, 1979.

[16] Knuth, Eldon L., Introduction to Statistical Thermodynamics, McGraw-Hill, New York, 1966.

[17] Landauer, R., Information is physical, Phys. Today, May 1991 23-29.

[18] Landauer, R., The physical nature of information, Phys. Lett. A, 217 188, 1996.

[19] van Lint, J. H., Coding Theory, Springer-Verlag, New York/Berlin, 1982.

[20] Lipton, R. J., Using DNA to solve NP-complete problems, Science, 268 542–545, Apr. 28, 1995.

[21] MacWilliams, F. J., and Sloane, N. J. A., The theory of error correcting codes, Elsevier Science, Amsterdam, 1977.

[22] Martin, N. F. G., and England, J. W., Mathematical Theory of Entropy, Addison-Wesley, Reading, 1981.

[23] Maxwell, J. C., Theory of heat Longmans, Green and Co, London, 1871.

[24] von Neumann, John, Probabilistic logic and the synthesis of reliable organisms from unreliable components, in automata studies(Shanon,McCarthy eds), 1956 .

[25] Papadimitriou, C. H., Computational Complexity, Addison-Wesley, Reading, 1994.

[26] Pierce, John R., An Introduction to Information Theory – Symbols, Signals and Noise, (second revised edition), Dover Publications, New York, 1980.

[27] Roman, Steven, Introduction to Coding and Information Theory, Springer-Verlag, Berlin/New York, 1997.

[28] Sampson, Je_rey R., Adaptive Information Processing, an Introductory Survey, Springer-Verlag, Berlin/New York, 1976.

[29] Schroeder, Manfred, Fractals, Chaos, Power Laws, Minutes from an Infinite Paradise, W. H. Freeman, New York, 1991.

[30] Shannon, C. E., A mathematical theory of communication Bell Syst. Tech. J. 27 379; also p. 623, 1948.

[31] Slepian, D., ed., Key papers in the development of information theory IEEE Press, New York, 1974.

[32] Turing, A. M., On computable numbers, with an application to the Entscheidungsproblem, Proc. Lond. Math. Soc. Ser. 2 42, 230 ; see also Proc. Lond. Math. Soc. Ser. 2 43, 544, 1936.

[33] Zurek, W. H., Thermodynamic cost of computation, algorithmic complexity and the information metric, Nature 341 119-124, 1989.

[34] Buzug T. M. and Weese J., “Image registration for DSA quality enhancement”, *Computerized Imaging Graphics* 22 103 1998.

[35] Tom Carter’s Notes: http://cogs.csustan.edu/~tom/