

# Computational Statistics with Application to Bioinformatics

Prof. William H. Press  
Spring Term, 2008  
The University of Texas at Austin

## Unit 21: Information Theory Characterization of Distributions

# Unit 21: Information Theory Characterization of Distributions (Summary)

- Entropy of a message
  - various definitions
    - log of the number of at-all-probable messages
    - length of maximally compressed message
    - log cutdown on hypothesis space, given a symbol
    - e-folding rate of capital increase in a fair game
  - di-, tri- (etc.) graph entropy bounds true entropy
    - interesting biological signals in di-peptide distribution in proteome
- Various entropies from a joint distribution are all related
  - for example,
    - entropy of the joint distribution (i.e., flattened)
    - conditional entropy
    - mutual entropy
  - positivity of mutual entropy
    - implies decreasing conditional entropy when you know the conditions
    - mutual entropy is the betting value of side-information
    - Kelly's formula for proportional betting
    - genomic CG richness is predictive of proteomic AA choice
- Kullback-Leibler distance
  - how different are two distributions
    - if you try to compress a message in one with a code from the other
    - or, if you are trying to distinguish them by log-odds on observed data
    - or, in Kelly's formula, if you are competing with the bookie on estimating the true odds

## Information Theory Characterization of Distributions

As functioning machines, proteins have a somewhat modular three-dimensional (tertiary) structure. But the [more-or-less] complete instructions for making a protein are a one-dimensional sequence of characters representing amino acids.



lactate dehydrogenase,  
showing alpha helices and beta  
sheets

For example:

261 characters, each in {A-Z} minus {BJOUXZ} (20 amino acids)

**MAAACRSVKGLVAVITGGASGLGLATAERLVGQGASAVLLDLPNSG  
GEAQAKKLGNNCVFAPADVTSEKDVQTALALAKGKFG RVDVAVNCA  
GIAVASKTYNLKKGQTHTLEDFQRVLDVNLMGTFNVIRLVAGEMGQN  
EPDQGGQRGVIINTASVAAFEGQVGQAAYSASKGGIVGMTLPIARDL  
APIGIRVMTIAPGLFGTPLLTSLEKVCNFLASQVPFPSRLGDPAEYAH  
LVQAIENPFLNGEVIRLDGAIRMQP\***

(I picked this randomly in the human genome. A sequence search shows it to be “hydroxysteroid (17-beta) dehydrogenase “.)

How many proteins of length 261 are there?  $20^{261}$  ? Yes, in a sense, but...

Shannon's key observation is that, if the characters in a message occur with unequal distribution  $p_i$ , then, for long messages, there is quite a sharp divide between rather probable messages and extremely improbable ones. Lets estimate the number of probable ones.

(The  $\log_2$  of this number is the information content of the message, in bits.)

We estimate as follows

$$2^B \approx \frac{M!}{\prod_i (M p_i)!}$$

← number of shuffled messages
← number of rearrangements of identical symbols i

$$\begin{aligned}
 B \ln 2 &\approx M \ln \left( \frac{M}{e} \right) - \sum_i (M p_i) \ln \left( \frac{M p_i}{e} \right) \\
 &= \cancel{M \ln \left( \frac{M}{e} \right)} - M \left( \cancel{\sum_i p_i} \right) \ln \left( \frac{M}{e} \right) - M \sum_i p_i \ln p_i \\
 &\equiv M H(\mathbf{p})
 \end{aligned}$$

← entropy in nats

If you take all logs base 2, you get entropy in bits.  
1 nat = 1.4427 bits.

## Interpretations of entropy:

1. It's the (binary) message length of the maximally compressed message.

Because, just send a binary serial number among all the probable messages. (And do something else for the improbable ones – which will never happen and negligibly affect the mean length!)

2. It's the expected log cut-down in the number of remaining hypotheses with a feature distributed as  $\mathbf{p}$ , if we do an experiment that measures  $i$

$$\langle \ln p_i \rangle = \sum_i p_i \ln p_i = -H(\mathbf{p})$$

This is a figure of merit for experiments if, by repeated experiments, we want to get the number of remaining hypotheses down to 1.

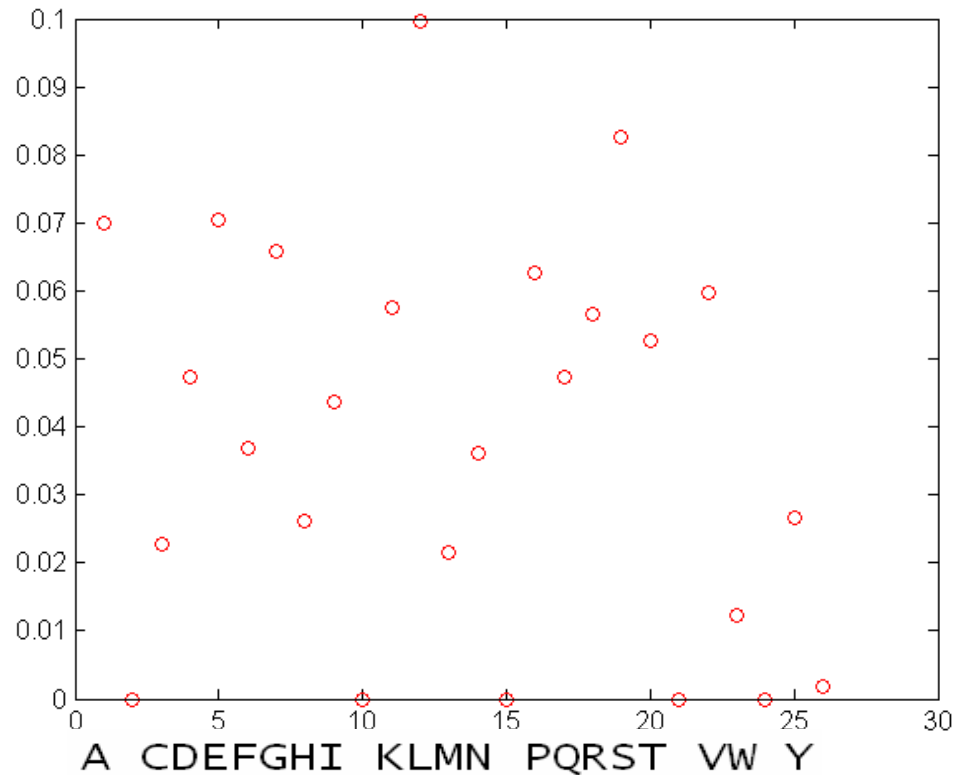
3. It's the e-folding (or doubling) rate of capital for a fair game about which you have perfect predictive information.

$$\text{payoff (odds)} \rightarrow \langle o_i \rangle = p_i o_i = 1$$

(This seems fanciful, but will make more sense when we discuss the case of partial predictive information.)

# What actually is the distribution of amino acids in human proteins?

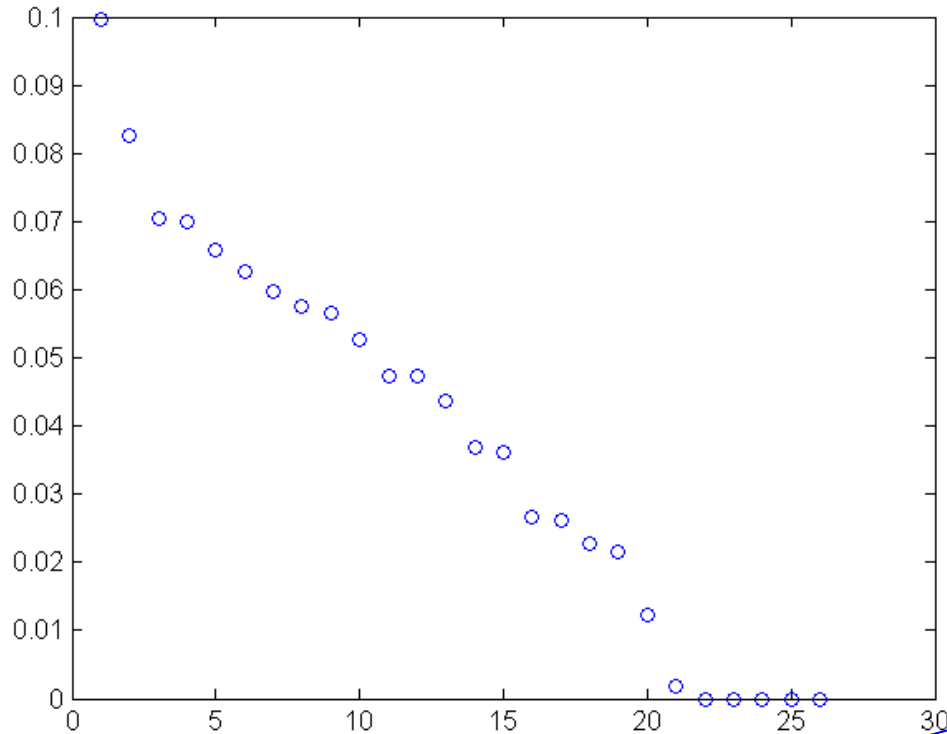
```
load 'aadi st_mono.txt'; ← (file on course web site)  
mono = aadi st_mono ./ sum(aadi st_mono(:));  
plot(mono(1:26), 'or')
```



A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid (Aspartate)
C	Cysteine
Q	Glutamine
E	Glutamic acid (Glutamate)
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
S	Serine
T	Threonine
W	Tryptophan
Y	Tyrosine
V	Valine

Plot distribution in descending order. Also calculate entropy:

```
plot(sort(mono(1:26), 'descend'), 'ob')
```



Notice that we flatten any structure in x when calculating the entropy.

```
entropy2 = @(x) sum(-x(:). *log(x(:)+1.e-99))/log(2);
```

```
h2bound = log(20)/log(2)
```

```
h2mono = entropy2(mono)
```

```
h2bound =  
4.3219
```

```
h2mono =  
4.1908
```

maximum entropy that 20 characters could have

actual (single peptide) entropy of the AA's

Actually, the single peptide (“monographic”) entropy is only a bound on the true entropy of proteins, because there can be (and is) multiple symbol nonrandomness.

Standard compression programs also bound the entropy, sometimes well, sometimes not:

Directory of D:\staticbio\prot\*

4/11/08	12: 18	9, 753, 363	___A_	proteomeHG17. txt
4/14/08	17: 45	5, 554, 389	___A_	proteomeHG17. zip
4/11/08	12: 18	5, 554, 186	___A_	proteomeHG17_1. txt. gz

$8 \times 5554186 / 9753363 = 4.556$  (yuck! not as good as our monographic bound of 4.191)

Let’s look at the dipeptide (digraph) and tripeptide (trigraph) distribution.

```
load 'aadi st_di . txt' ;
di = aadi st_di ./ sum(aadi st_di (:));
h2di = entropy2(di)
h2di =
    8.3542
```

$8.3542 / 2 = 4.177$

```
load 'aadi st_tri . txt' ;
tri = aadi st_tri ./ sum(aadi st_tri (:));
h2tri = entropy2(tri)
```

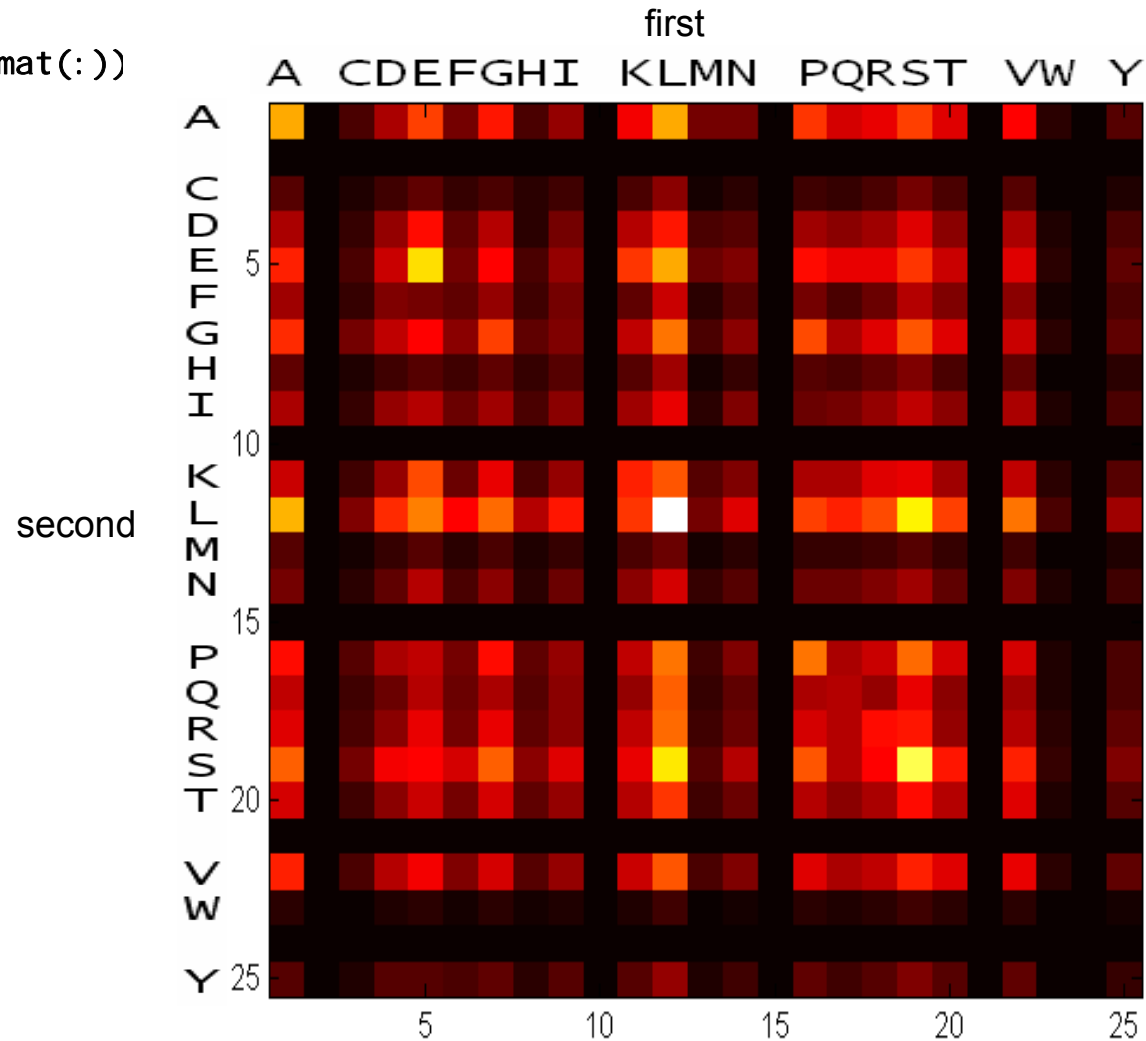
```
h2tri =
    12.5026
```

$12.5026 / 3 = 4.168$

(We’ll see in a minute that it’s a mathematical theorem that these have to decrease – but they don’t have to decrease much!)

Actually it's interesting to look at the dipeptide distribution

```
di mat = reshape(di , 32, 32);  
image_di = 64*di mat ./max(di mat(:))  
image(image_di (1: 25, 1: 25))  
colormap(' hot' )
```



But how much of this structure is due to the monographic distribution alone?

```

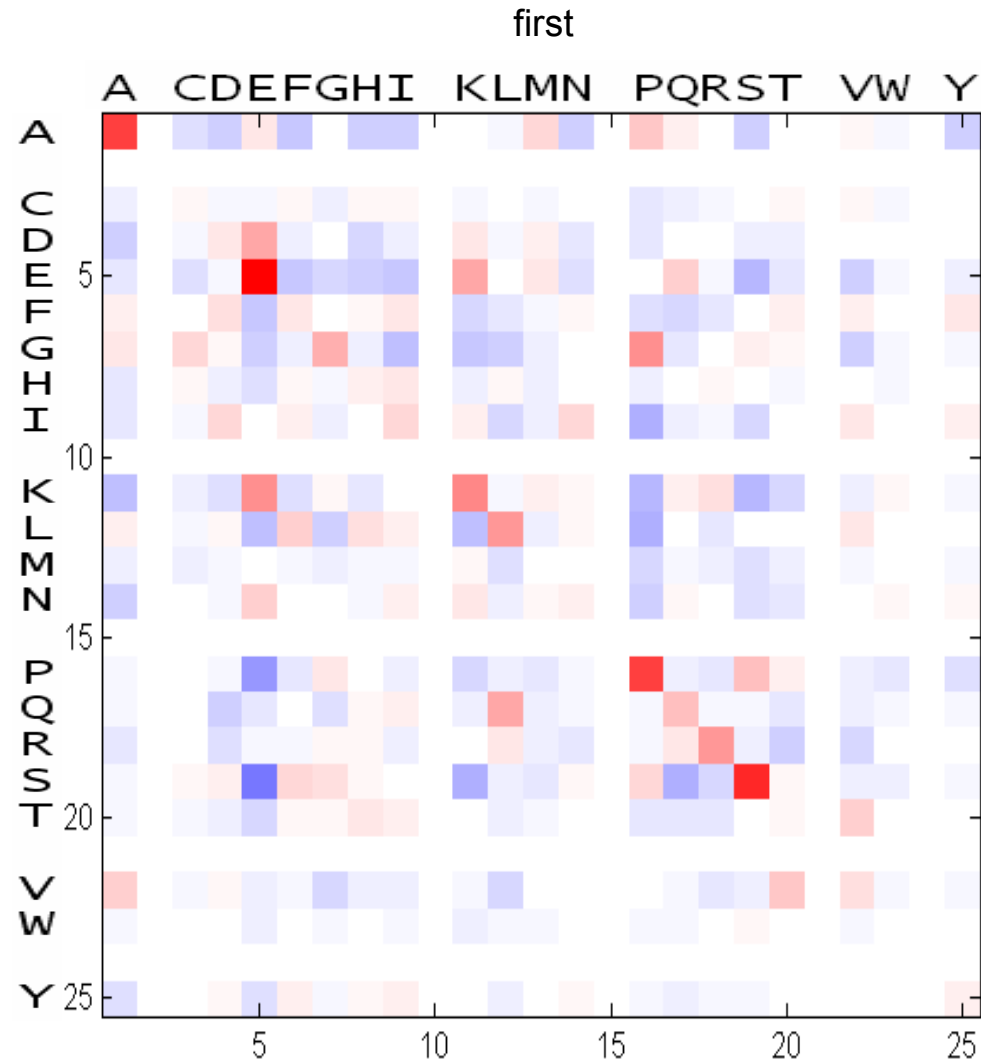
di screp = di mat - mono * mono';
i mage_di screp = (32/max(di screp(:))) * di screp + 32;
i mage(i mage_di screp(1: 25, 1: 25));
genecol ormap = [mi n(1, (1: 64)/32); 1-abs(1-(1: 64)/32); mi n(1, (64-(1: 64))/32)]';
col ormap(genecol ormap)

```

Interesting biology: AA's like to repeat. Is this AA chemistry or genomic stuttering? And what's going on among S, E, P, and K?

A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid (Aspartate)
C	Cysteine
Q	Glutamine
E	Glutamic acid (Glutamate)
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
S	Serine
T	Threonine
W	Tryptophan
Y	Tyrosine
V	Valine

second



Is there more we can say about this picture information theoretically?

So far, we have the monographic entropy ( $H = 4.1908$  bits) and the digraph entropy ( $H = 8.3542$  bits).

But the digraph entropy is flattened – doesn't know about rows and columns:

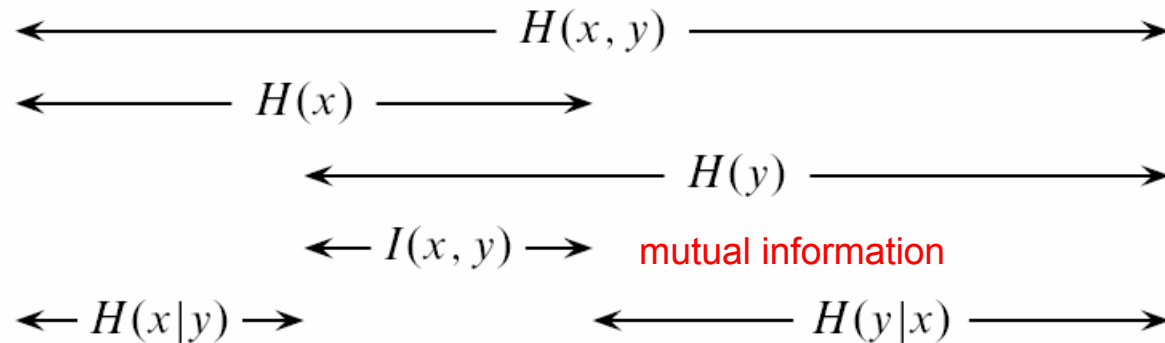
$$H(x, y) = - \sum_{i,j} p_{ij} \ln p_{ij}$$

Let's try to capture something with more structure. The conditional entropy is the expected (average) entropy of the second character, *given* the first:

$$\begin{aligned} H(y|x) &= - \underbrace{\sum_i p_i}_{\text{expectation}} \underbrace{\sum_j \frac{p_{ij}}{p_i} \ln \frac{p_{ij}}{p_i}}_{\text{entropy of one column}} = - \sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_i} \\ &= H(x, y) + \sum_i \left( \sum_j p_{ij} \right) \ln p_i. \\ &= H(x, y) - H(x) \qquad \qquad \qquad 4.1642 \text{ bits} \end{aligned}$$

So the conditional entropy, while (as we will see) useful, follows directly from the monographic and digraphic entropies.

In fact there are a bunch of relations, all easy to prove:



$$H(x) - H(x|y) = H(y) - H(y|x) \equiv I(x, y) \quad \text{0.0266 bits}$$

$$I(x, y) = \sum_{i,j} p_{ij} \ln \left( \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} \right)$$

Proof that mutual information always positive:

$$\begin{aligned} H(y|x) - H(y) &= - \sum_{i,j} p_{ij} \ln \frac{p_{ij} / p_{i\cdot}}{p_{\cdot j}} \\ &= \sum_{i,j} p_{ij} \ln \frac{p_{\cdot j} p_{i\cdot}}{p_{ij}} \\ &\leq \sum_{i,j} p_{ij} \left( \frac{p_{\cdot j} p_{i\cdot}}{p_{ij}} - 1 \right) \\ &= \sum_{i,j} p_{i\cdot} p_{\cdot j} - \sum_{i,j} p_{ij} \\ &= 1 - 1 = 0 \end{aligned}$$

You might wonder if a quantity as small as 2.7 centibits is ever important. The answer is yes: It is a signal that you could start to detect in  $1/0.027 \sim 40$  characters, and easily detect in  $\sim 100$ .

Mutual information has an interesting interpretation in game theory (or betting)

side information:

Outcome  $i$  with probability  $p_i$  is what you can bet on at odds  $1/p_i$

But you also know the value of another feature  $j$  that is partially informative

In other words, you know the matrix  $p_{ij}$

and it's neither diagonal (perfect prediction) nor rank-one (complete independence)

example:  $i$  is which horse to bet on,  $j$  is which jockey is riding

What is your best betting strategy?

$b_{ij}$  fraction of assets you bet on  $i$  when the side info is  $j$

$$\sum_i b_{ij} = 1, \quad 0 \leq j \leq J - 1$$

maximize the return on assets per play:

$$W = \left\langle \ln \frac{b_{ij}}{p_{i\cdot}} \right\rangle = \sum_{i,j} p_{ij} \ln \frac{b_{ij}}{p_{i\cdot}}$$

we can do this by Lagrange multipliers, maximizing the Lagrangian

$$\mathcal{L} = \sum_{i,j} p_{ij} \ln \frac{b_{ij}}{p_{i\cdot}} - \sum_j \lambda_j \left( \sum_i b_{ij} - 1 \right)$$

$$\mathcal{L} = \sum_{i,j} p_{ij} \ln \frac{b_{ij}}{p_{i\cdot}} - \sum_j \lambda_j \left( \sum_i b_{ij} - 1 \right)$$

$$0 = \frac{\partial \mathcal{L}}{\partial b_{ij}} = \frac{p_{ij}}{b_{ij}} - \lambda_j$$

$$b_{ij} = \frac{p_{ij}}{\lambda_j} = \frac{p_{ij}}{p_{\cdot j}}$$

This is the famous “proportional betting” formula or “Kelly’s formula”, first derived by Kelly, a colleague of Shannon, in 1956. You should bet in linear proportion to the probabilities conditioned on any side information.

$$W = \sum_{i,j} p_{ij} \ln \left( \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} \right) = I(x, y)$$

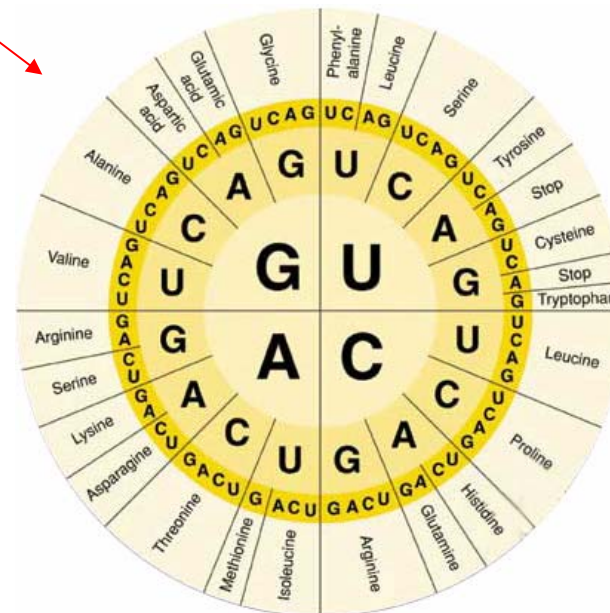
So your expected gain is the mutual information between the outcome and your side information!

Back in the proteome, let's do another example:

It's normally assumed that proteins are selected for maximum fitness, with the genome just recording that selection (and providing the mechanism for random mutations of the proteins).

If so, we should not expect a protein's choice of AAs to be affected by a purely genomic parameter like local CG (vs. AT) richness. The latter should affect only synonymous codon choice, not AA choice.

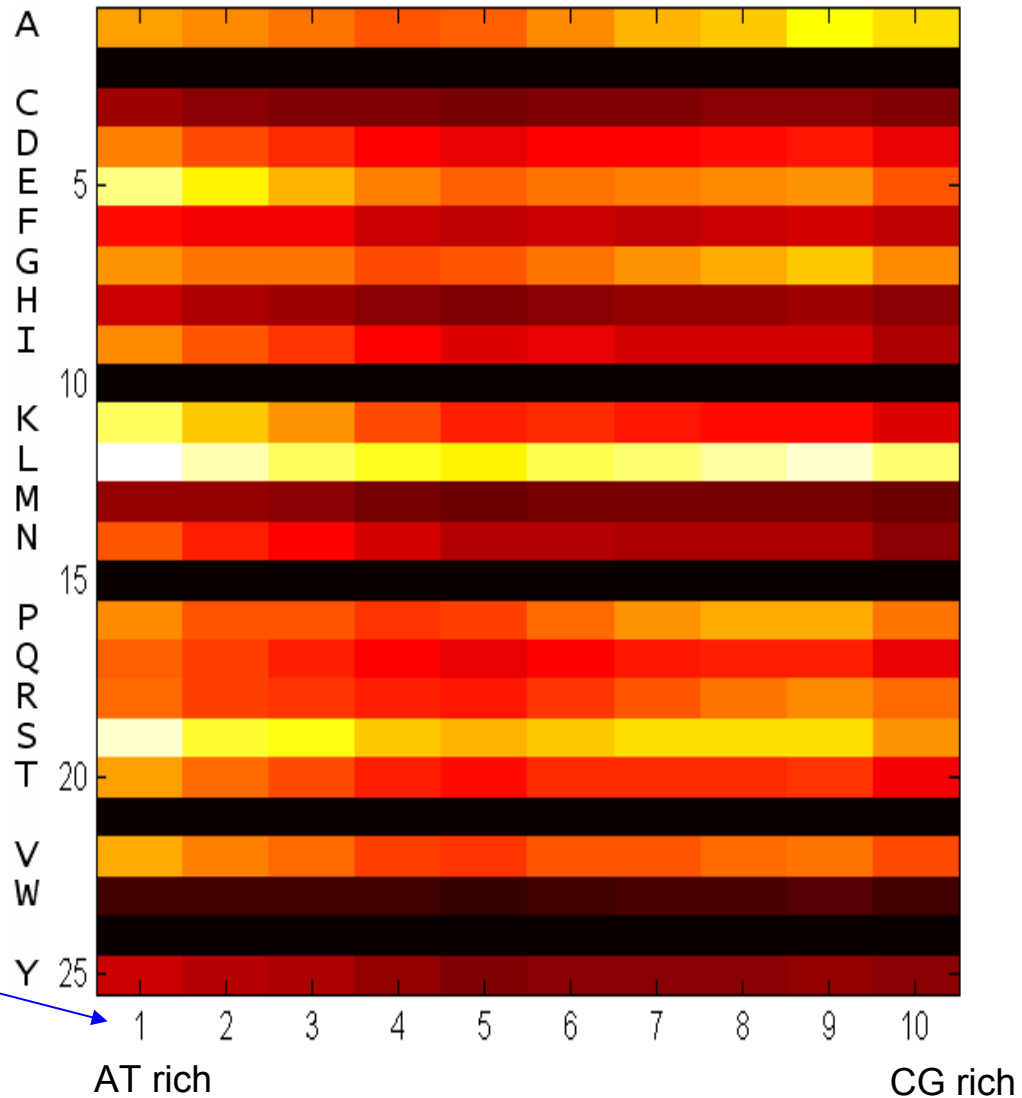
The null hypothesis is thus that CG richness should not be predictive of the AA distribution.



```
load 'aadi st_monocg.txt';
monocg = aadi st_monocg ./ sum(aadi st_monocg(:));
monocgmat = reshape(monocg, 32, 10);
image(64*monocgmat(1:25, :). /max(monocg(:)))
colormap('hot')
```

(As before, this isn't very informative, because it conflates the mono- and di-signals.)

these are 10 roughly equal "bins" of CG richness, as measured by the 3<sup>rd</sup> codon distribution separately for each protein



```

load 'aadi st_cg.txt';
cg = aadi st_cg ./ sum(aadi st_cg(:));
di screp = monocgmat - mono * cg';
image_di screp = (32/max(di screp(:))) * di screp + 32;
image(image_di screp(1: 25, :));
colormap(genecolormap)

```

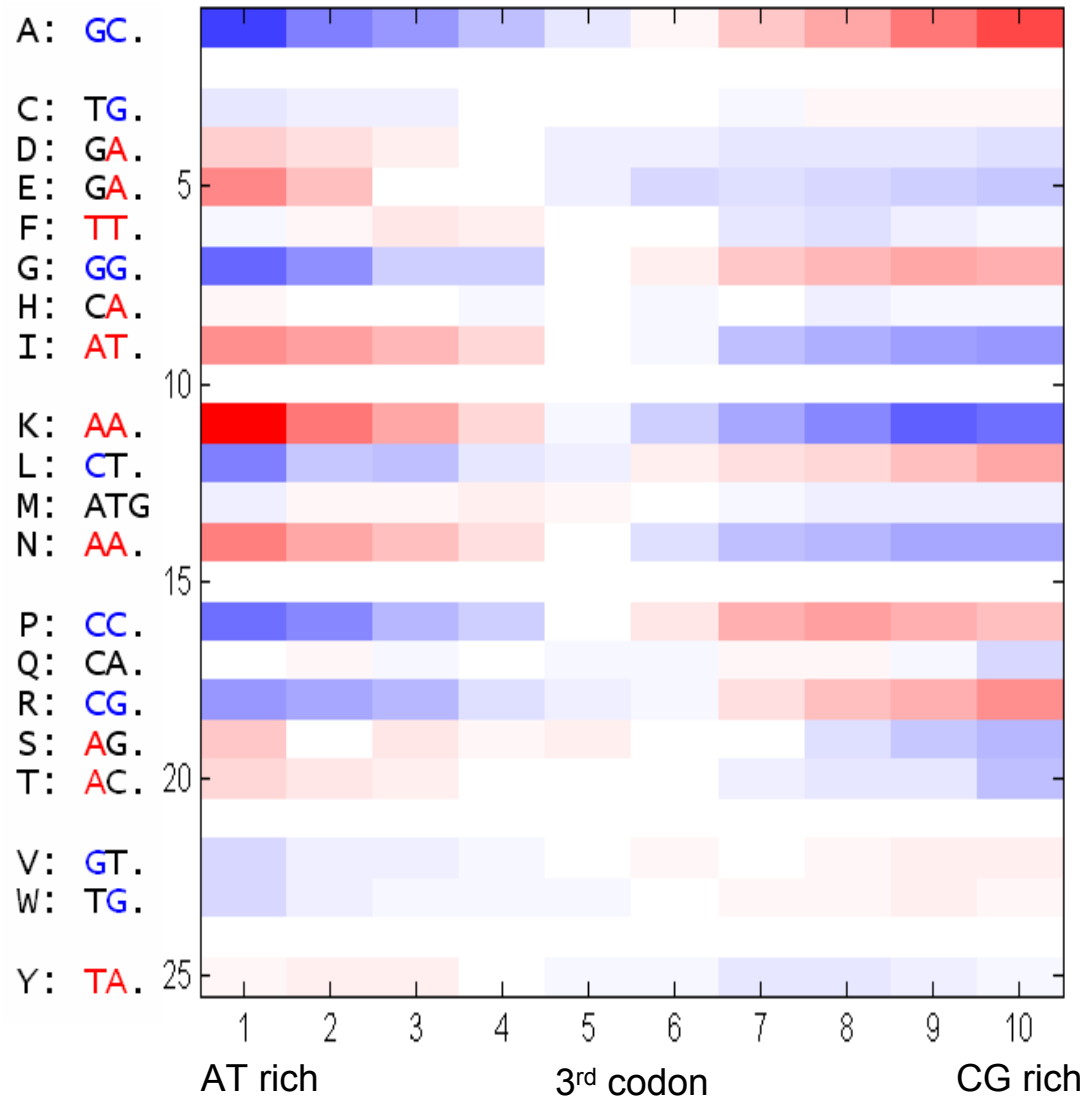
```

h2cg = entropy2(cg)
h2monocg = entropy2(monocg)
h2cg =
    3.3134
h2monocg =
    7.4969

```

$I(\text{"CG"}, \text{"AA"})$   
 $= 4.1908 + 3.3134 - 7.4969$   
 $= 0.0073$  bits

One could check (e.g., by resampling), that this is highly statistically significant. But that is also wonderfully clear from the systematics, which almost completely follow the genetic code if you assume a priority ordering  $A > G > C > T$



So codon preferences definitely do affect the choice of AA!

Finally, the Kullback-Leibler distance is an information theoretic measure of how different are two distributions (“distance” from one to the other).

A.k.a. “relative entropy”.

$$D(\mathbf{p} \parallel \mathbf{q}) \equiv \sum_i p_i \ln \frac{p_i}{q_i}$$

Notice that it's not symmetric. It also doesn't have a triangle inequality. So it's not a metric in the mathematical sense.

But at least it's always positive!

$$-D(\mathbf{p} \parallel \mathbf{q}) = \sum_i p_i \ln \left( \frac{q_i}{p_i} \right) \leq \sum_i p_i \left( \frac{q_i}{p_i} - 1 \right) = 1 - 1 = 0$$

Interpretations:

1. It's the extra length needed to compress  $\mathbf{p}$  with a code designed for  $\mathbf{q}$

$$-\sum_i p_i \ln q_i = H(\mathbf{p}) + \sum_i p_i \ln \frac{p_i}{q_i} \equiv H(\mathbf{p}) + D(\mathbf{p} \parallel \mathbf{q})$$

2. It's the average log odds (per character) of rejecting the (false) hypothesis that you are seeing  $\mathbf{q}$  when you are (actually) seeing  $\mathbf{p}$

$$\mathbb{E} \mathcal{L} = \frac{p(\text{Data} \mid \mathbf{p})}{p(\text{Data} \mid \mathbf{q})} = \prod_{\text{data}} \frac{p_i}{q_i}$$

3. It's your expected capital gain when you can estimate the odds of a fair game better than the person offering (fair) odds, and when you bet by Kelly's formula

$$W = \langle \ln(b_i o_i) \rangle = \sum_i p_i \ln(b_i o_i)$$

$$b_i = p_i$$

$$o_i = 1/r_i$$

calculation similar to the previous gives

$$W = \langle \ln(b_i o_i) \rangle = \sum_i p_i \ln \frac{q_i}{r_i} = D(\mathbf{p} \parallel \mathbf{r}) - D(\mathbf{p} \parallel \mathbf{q})$$

Turns out that if the house keeps a fraction  $(1 - f)$ , the requirement is

$$D(\mathbf{p} \parallel \mathbf{r}) - D(\mathbf{p} \parallel \mathbf{q}) > -\ln f$$

Betting is a competition between you and the bookie on who can more accurately estimate the true odds, as measured by Kullback-Leibler distance.