Opinionated Lessons in Statistics

by Bill Press

#37 A Few Bits of Information Theory

Information Theory quantifies the information in “messages”.
The messages can be natural, not just made by humans.

lactate dehydrogenase, showing alpha helices and beta sheets

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.

For example:

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

MAAACRSVKGLVAVITGGASGLGLATAERLVGQGASAVLLDLPNSG
GEAQAKKLGNNCVFAPADVTSEKDVQTALALAKGKFGRVDVAVNCA
GIAVASKTYNLKKGQTHTLEDFQRVLDVNLMGTFNVIRLVAGEMGQN
EPDQGGQRGVIINTASVAAFEGQVGQAAYSASKGGIVGMTLPIARDL
APIGIRVMTIAPGLFGTPLLTSLPEKVCNFLASQVPFPSRLGDPAEYAH
LVQAIIENPFLNGEVIRLDGAIRMQP

(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

number of shuffled messages

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

number of rearrangements of identical symbols i

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

entropy in nats

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

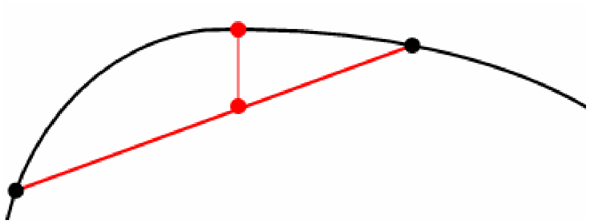
If you take all logs base 2, you get entropy in bits.
1 nat = 1.4427 bits.

$$H(\mathbf{p}) = -\sum_{i=1}^{N} p_i \ln p_i$$

Evidently positive for all **p**’s.

Minimum value zero when a single p_i=1.

Maximum when all the p_i’s are equal:



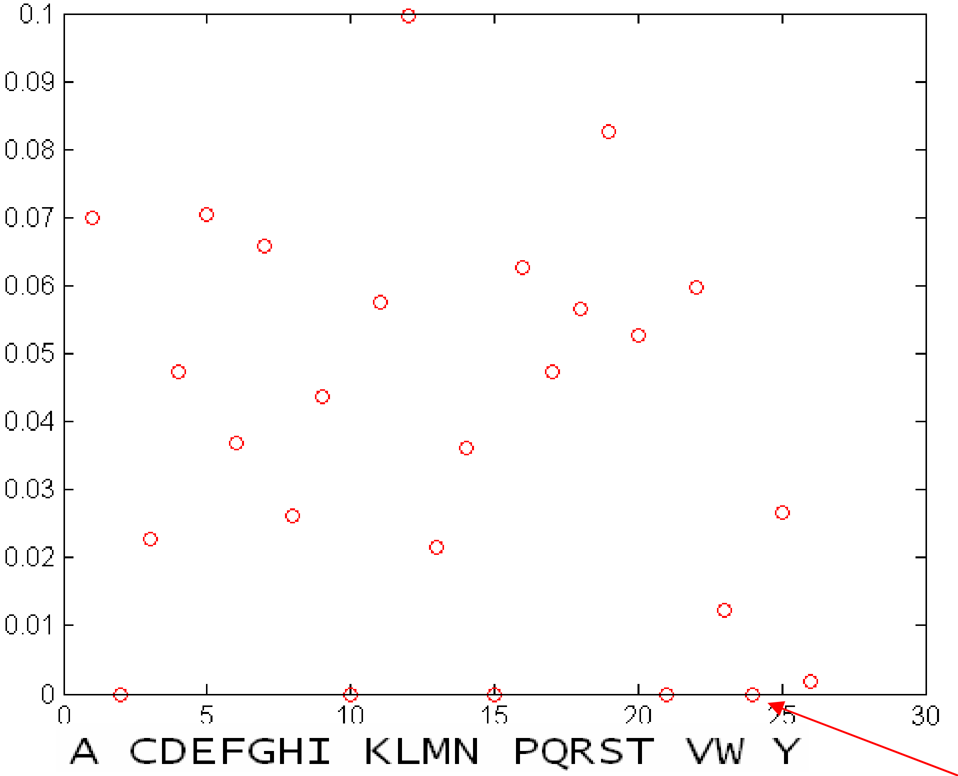
$$\mathcal{L} = -\sum_i p_i \ln p_i + \lambda \left(\sum_i p_i - 1 \right)$$

$$0 = \frac{\partial \mathcal{L}}{\partial p_j} = -\ln p_j - 1 + \lambda$$

$$\Rightarrow \ln p_j = \lambda - 1 = \text{constant}$$

$$\max(H) = \ln N$$

Example: what is the distribution of amino acids in human proteins?

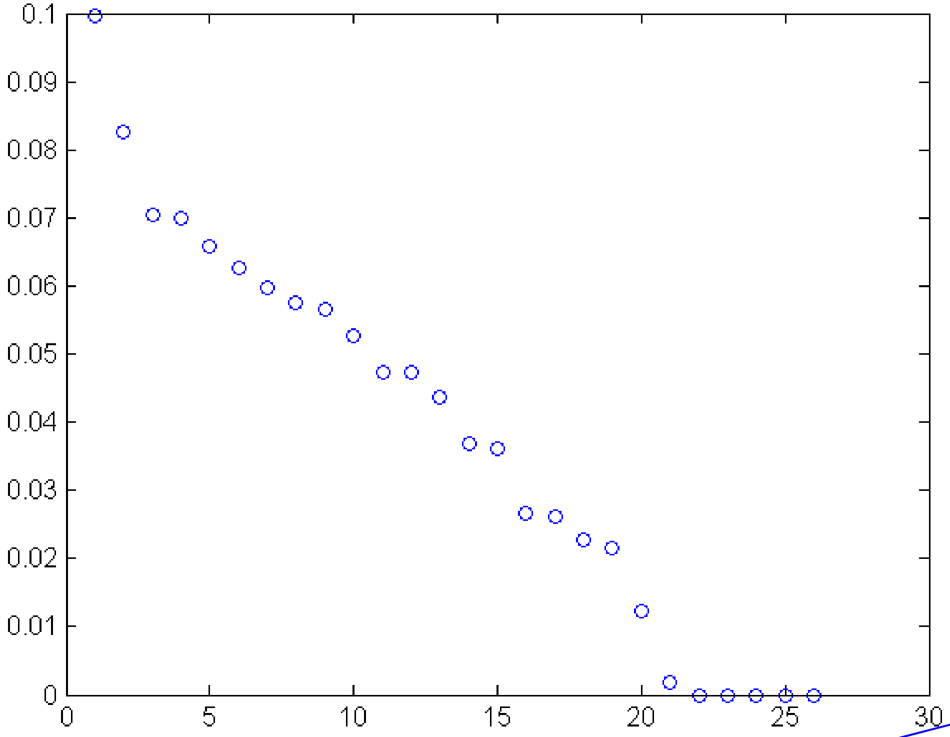


A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid (Aspartate)
C	Cysteine
Q	Glutamine
E	Glutamic acid (Glutamate)
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
S	Serine
T	Threonine
W	Tryptophan
Y	Tyrosine
V	Valine

Zeros just mean that there’s no AA with that letter abbreviation!

Plot distribution in descending order and calculate entropy:

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



So the answer to "how many likely proteins are there of length 261" (as a fraction of what is combinatorially possible):

$$\frac{\left(2^{4.19}\right)^{261}}{20^{261}} = 4.31 \times 10^{-11}$$

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 <u>bound</u> on the true entropy of proteins, because there can be (and is) multiple symbol nonrandomness.

Standard compression programs 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
```

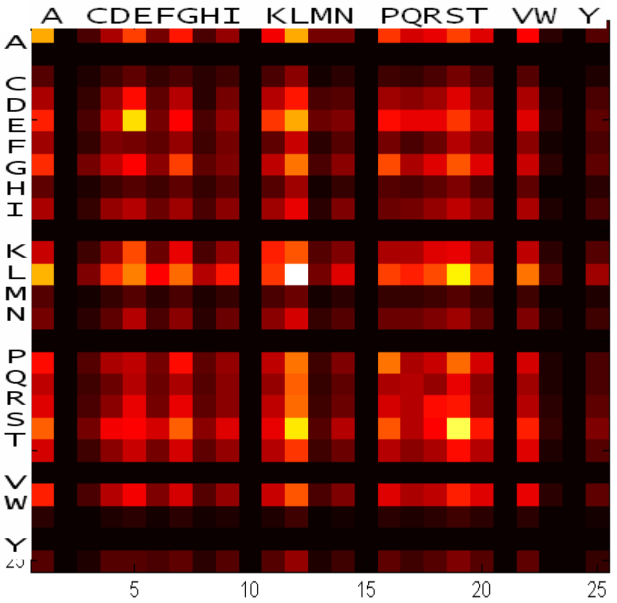
8 x 5554186 / 9753363 = 4.556 (yuck! not as good as our monographic bound of 4.191)

Let's look at the digraph entropy:

400 p_i's adding up to 1

```
h2di = entropy2(di)
h2di =
    8.3542
```

8.3542 / 2 = 4.177



And the trigraph entropy:

8000 p_i's adding up to 1

```
h2tri = entropy2(tri)
h2tri =
    12.5026
```

12.5026 / 3 = 4.168

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

"True" entropy would be the limit of the n-graph entropies.

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

Interpretations of the entropy of a distribution:

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.

payoff (odds)

$$\langle o_i \rangle = p_i o_i = 1$$

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