ABSTRACT

Individual protein binding sites on DNA can be measured in bits of information. This information is related to the free energy of binding by the second law of thermodynamics, but binding kinetics appear to be inaccessible from sequence information since the relative contributions of the on- and off-rates to the binding constant, and hence the free energy, are unknown. However, the on-rate could be independent of the sequence since a protein is likely to bind once it is near a site. To test this, we used surface plasmon resonance and electromobility shift assays to determine the kinetics for binding of the Fis protein to a range of naturally occurring binding sites. We observed that the logarithm of the off-rate is indeed proportional to the individual information of the binding sites, as predicted. However, the on-rate is also related to the information, but to a lesser degree. We suggest that the on-rate is mostly determined by DNA bending, which in turn is determined by the sequence information. Finally, we observed a break in the binding curve around zero bits of information. The break is expected from information theory because it represents the coding demarcation between specific and nonspecific binding.

R. K. Shultzaberger, L. R. Roberts, I. G. Lyakhov, I. A. Sidorov, A. G. Stephen, R. J. Fisher, and T. D. Schneider, Correlation between binding rate constants and individual information of *E. coli* Fis binding sites, *Nucleic Acids Res.* **35** 5275-5283, 2007.

For further information: http://tinyurl.com/9tbph

THEORY

How is the Information of a DNA Binding Site related to Binding Rates?

1. **Information is related to binding energy by the Second Law of Thermodynamics**

$$R_i \propto -\Delta G \tag{1}$$

2. **Binding energy is related to the binding constant:**

$$\Delta G \propto \log K_{\mathrm{D}} \tag{2}$$

3. **The binding constant is a function of the on and off rates:**

$$K_{\mathrm{D}} = k_{\mathrm{off}}/k_{\mathrm{on}} \tag{3}$$

4. Once a protein is at a binding site, it will frequently bind irrespective of how strong the binding is, so **the on-rate, k_{on} should be roughly constant**

5. **Combining the above**

$$R_i \propto -\log k_{\mathrm{off}} \tag{4}$$

The more information a binding site has, the larger the number of contacts it can make with the protein and correspondingly the more difficult it becomes for thermal noise to separate the two once they are bound together. The off rate is strongly dependent on the detailed binding contacts since all of these have to be broken to release the protein.

Sequence Logo of Fis DNA Binding Sites

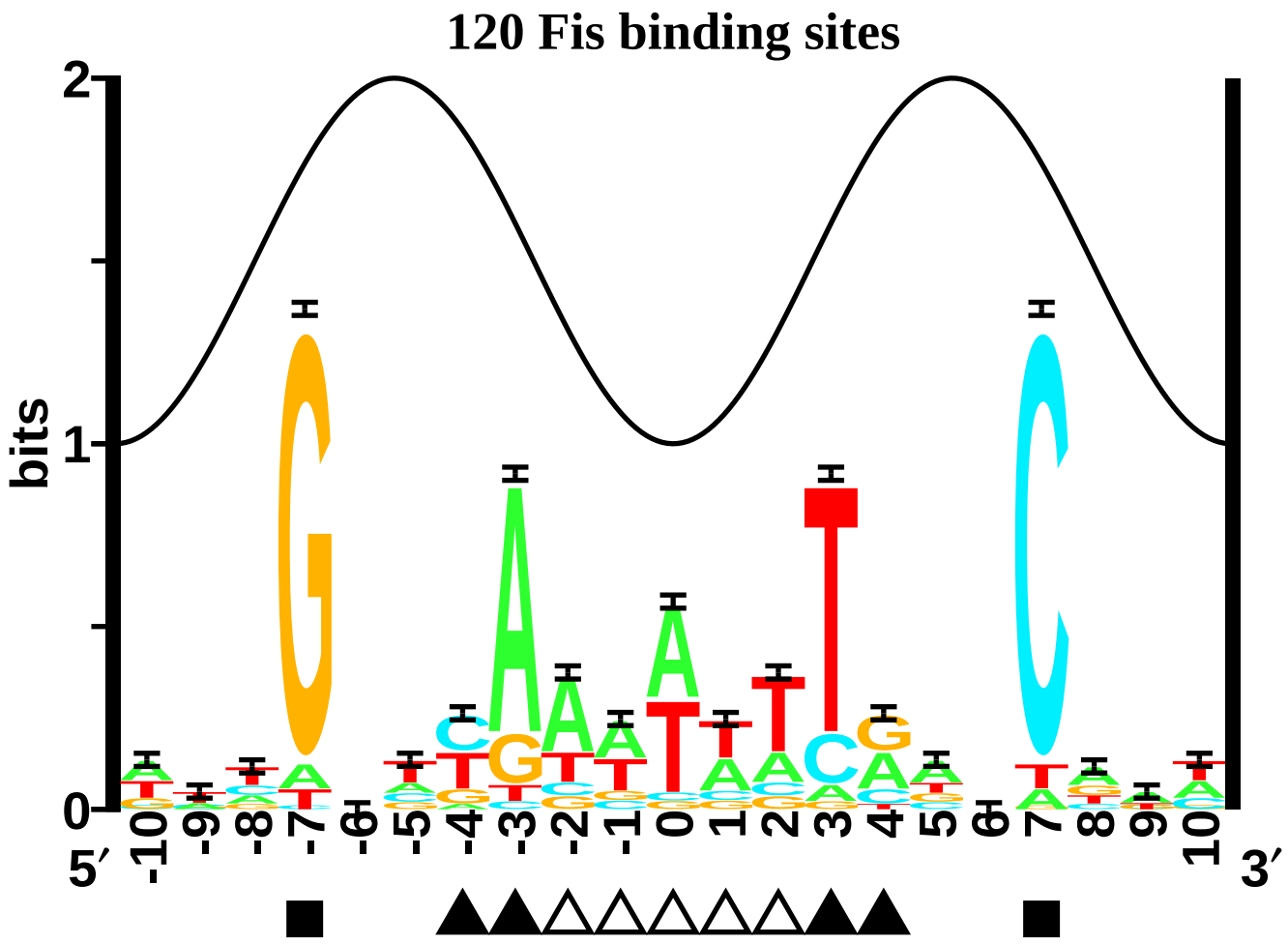

Figure 1: Sequence logo for the Fis protein.
The heights of letters in each stack are proportional to the frequency of each base at that position. The height of the stack is the information content for that position. The total conservation, summed for all positions in the range -7 to $+7$ is $R_{sequence} = 7.18 \pm 0.23$ bits per site, which is also the average of the individual information of all of the sites. The sine wave above the logo represents the 10.6 bp helical twist of B-form DNA. The positions presumably bound by the D-helices from the major groove at ± 7 are marked with squares, the pyrimidine/purine steps that kink the DNA at ± 4 and ± 3 are marked with filled triangles, and the A/T bases that allow bending into the minor groove are marked with open triangles.

Fis DNA Binding Sites of Varioius Strengths

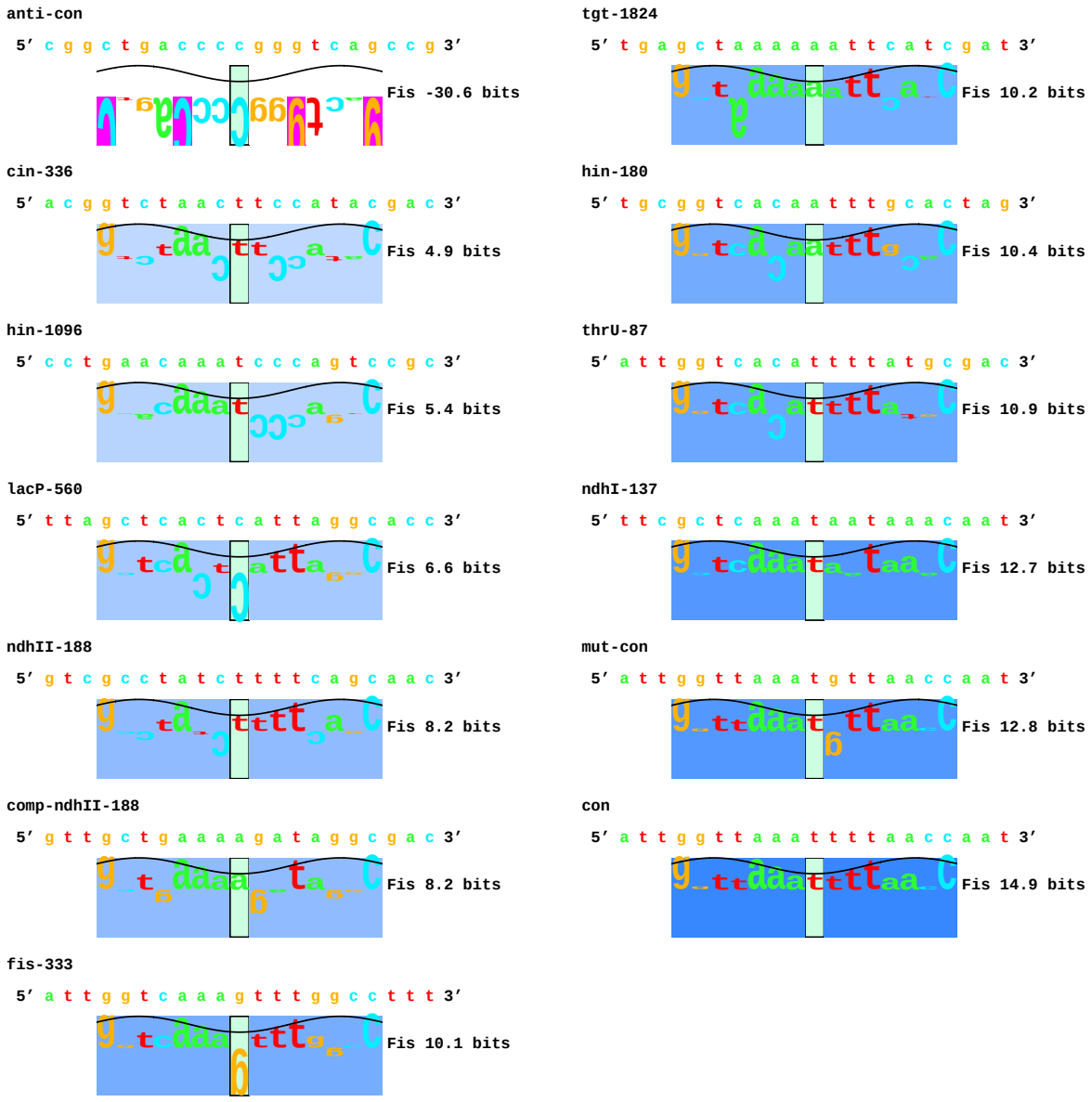

Figure 2: Sequences used for analysis.

Information analysis for individual sequences was computed and displayed using sequence walkers. Those positions that are favored according to the $R_{iw}(b,l)$ weight matrix (contribute positive information) are represented by bases above the x-axis, whereas those bases that are not favored (contribute negative information) are below the x-axis. The height of each base is its information contribution to the site strength. The sum of all base heights is R_i for the sequence, and this is given on the right of the sequence walker. The sequences are sorted by their strength in bits and the saturation of a colored rectangle behind each walker is proportional to that strength. As in Fig. 1, the sine wave above the walker represents the 10.6 bp helical twist of B-form DNA.

Surface Plasmon Resonance of Fis DNA Binding Off-Rates

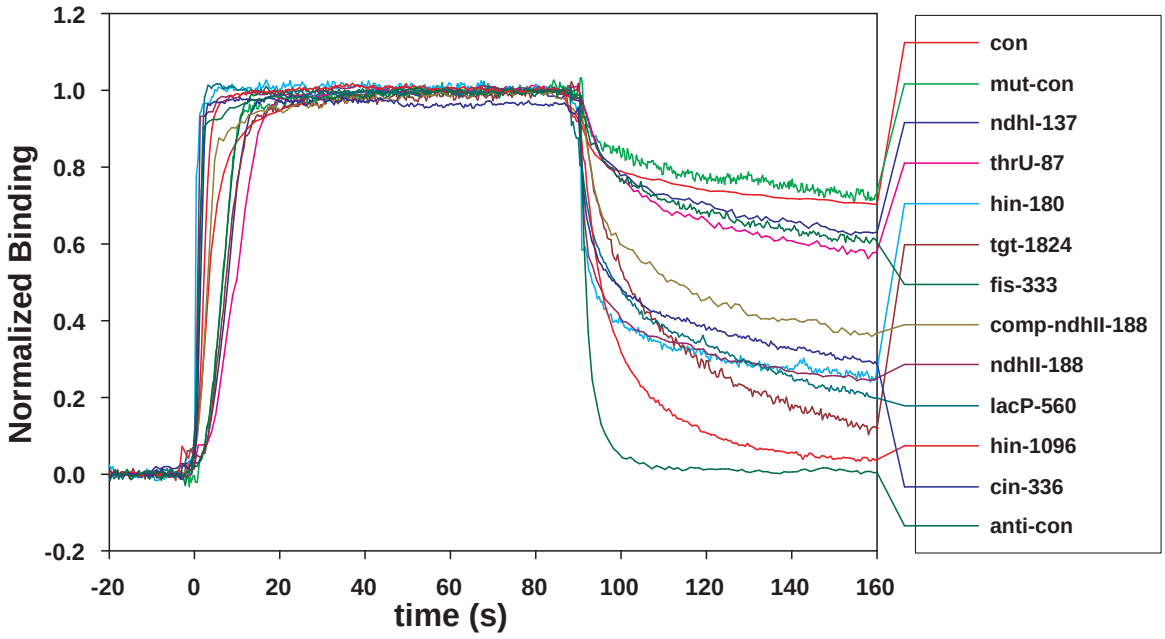

Figure 3: Sensogram of Fis bound to different DNA sequences.
All curves were normalized so that saturation of the chip is set to 1. At time zero, Fis was washed onto the SPR chip. At time 90 seconds, Fis was washed off the chip. The stability measurements were determined from the curve after 90 seconds.

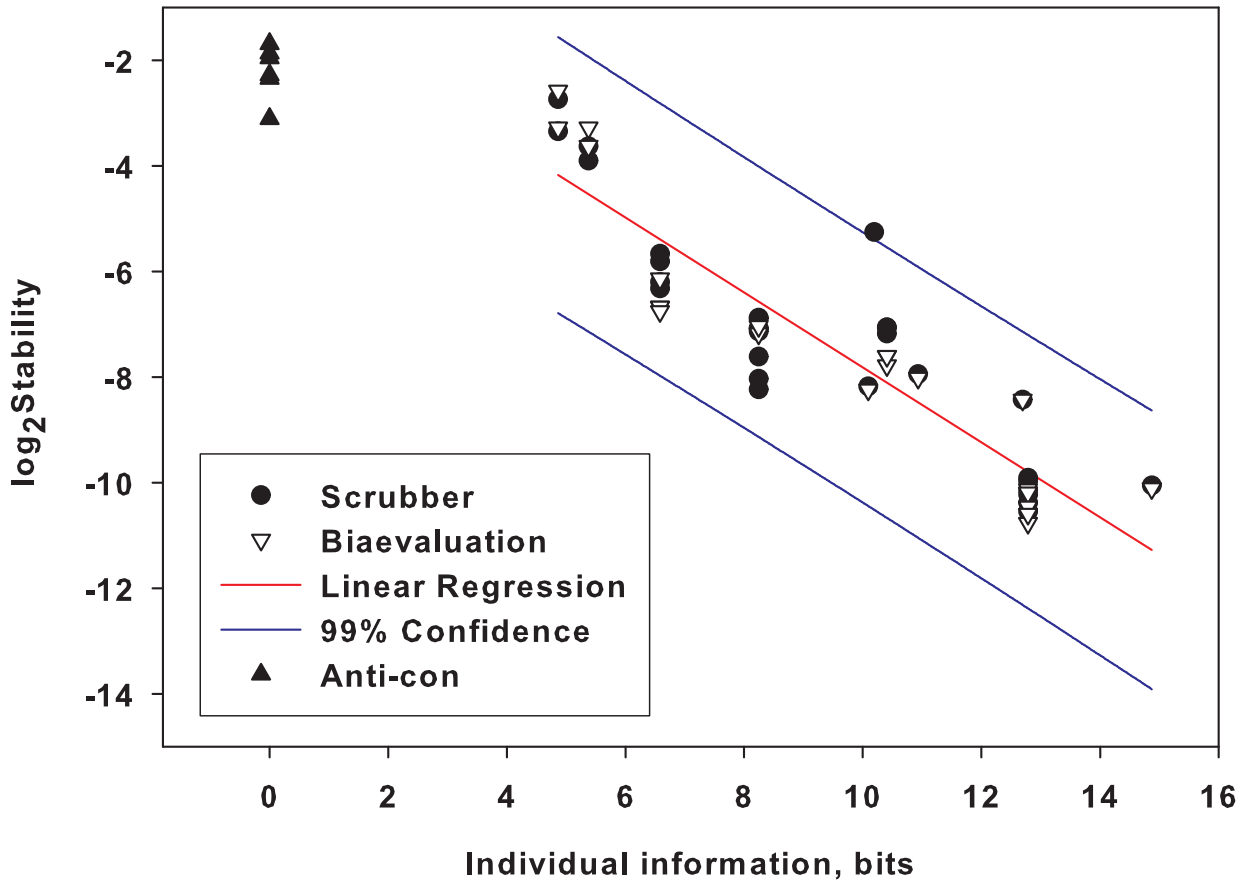

Figure 4: Binding site information is correlated to stability.
For each sequence described in Fig. 2 we plotted the stability *vs.* the information R_i. Scrubber and Biaevaluation are two implementations of curve fitting by a single exponential decay describing the dissociation. Both were used to evaluate all of the data and slight differences were observed from small deviations in the start and stop points chosen for analysis. We plot each measurement independently. Although the anti-con oligo is presumably non-specific at -30.6 bits, we plotted it as having 0 bits of information. All points at zero bits are for the anti-con oligo. The regression line (excluding the anti-con) is shown as a red line ($r^2 = -0.84$). 99% confidence limits for the regression are shown with blue lines. The equation for the regression line is $\log_2(\text{Stability}) = -0.70 \times \text{Individual information} - 0.84$.

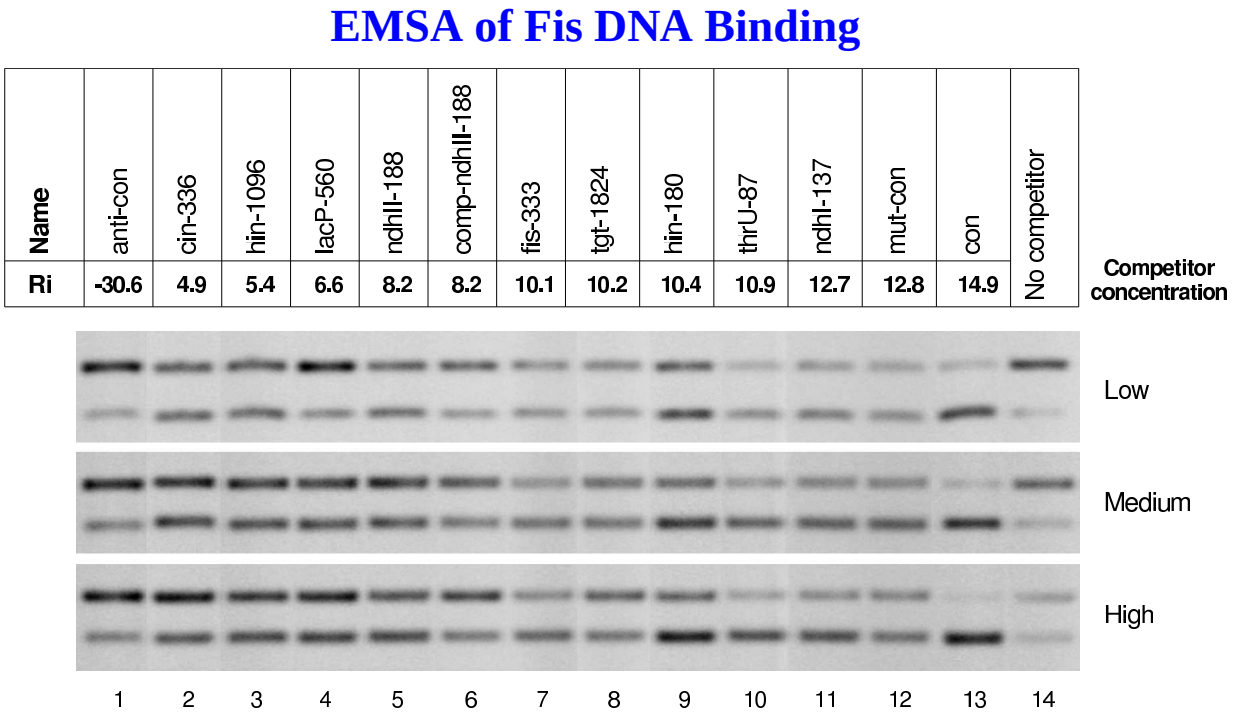

Figure 5: Competition Electrophoretic Mobility Shift Assay with three different concentrations of oligos containing different Fis binding sites. (See Supplementary Material Fig. 1 for the sequences.) For each concentration, the top band is Fis bound to the consensus 5′ 6-FAM labeled oligo and the bottom band is unbound labeled oligo (see Materials and Methods). The competitor concentrations shown are approximately: 1.0μ M low, 1.5μ M medium, 2.0μ M high; the exact values for each competitor are given in Supplementary Materials. Lanes 1 to 13: competitor oligos 1 to 13; Lane 14: no competitor.

A Breakpoint in the Fis DNA Binding Curve

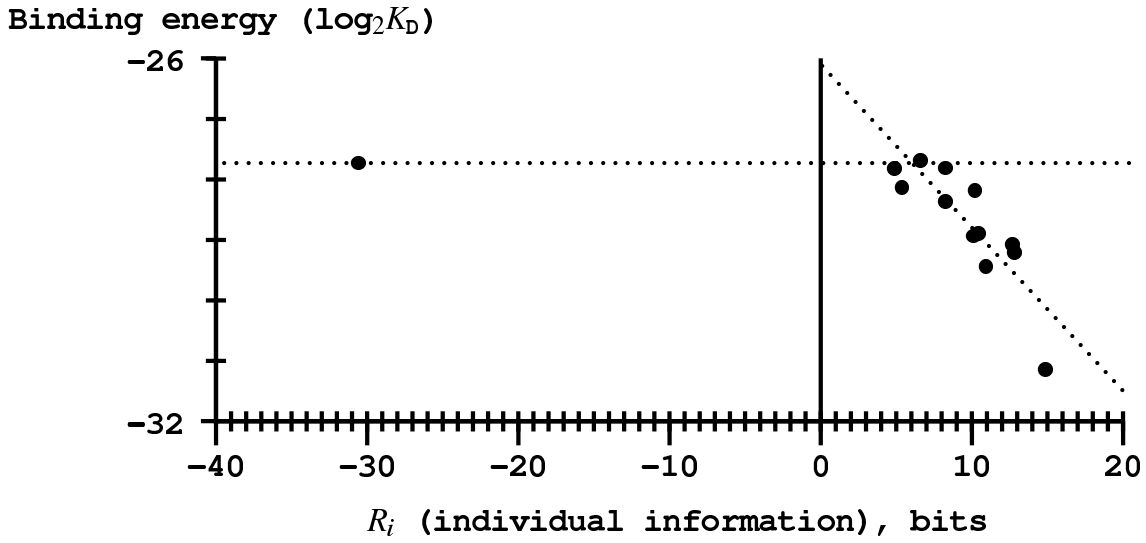

Figure 6: Binding energy is linearly related to binding site information for positive information binding sites but apparently flat for sites with negative information. The curve appears to break near zero bits. The average K_D values were normalized so that the Hin-180 sequence has the published value of Hin-D, 2×10^{-9} M. Excluding the anti-consensus at -30.6 bits, the regression line is has $r^2 = 0.73$.

Why Communications Have Breakpoints

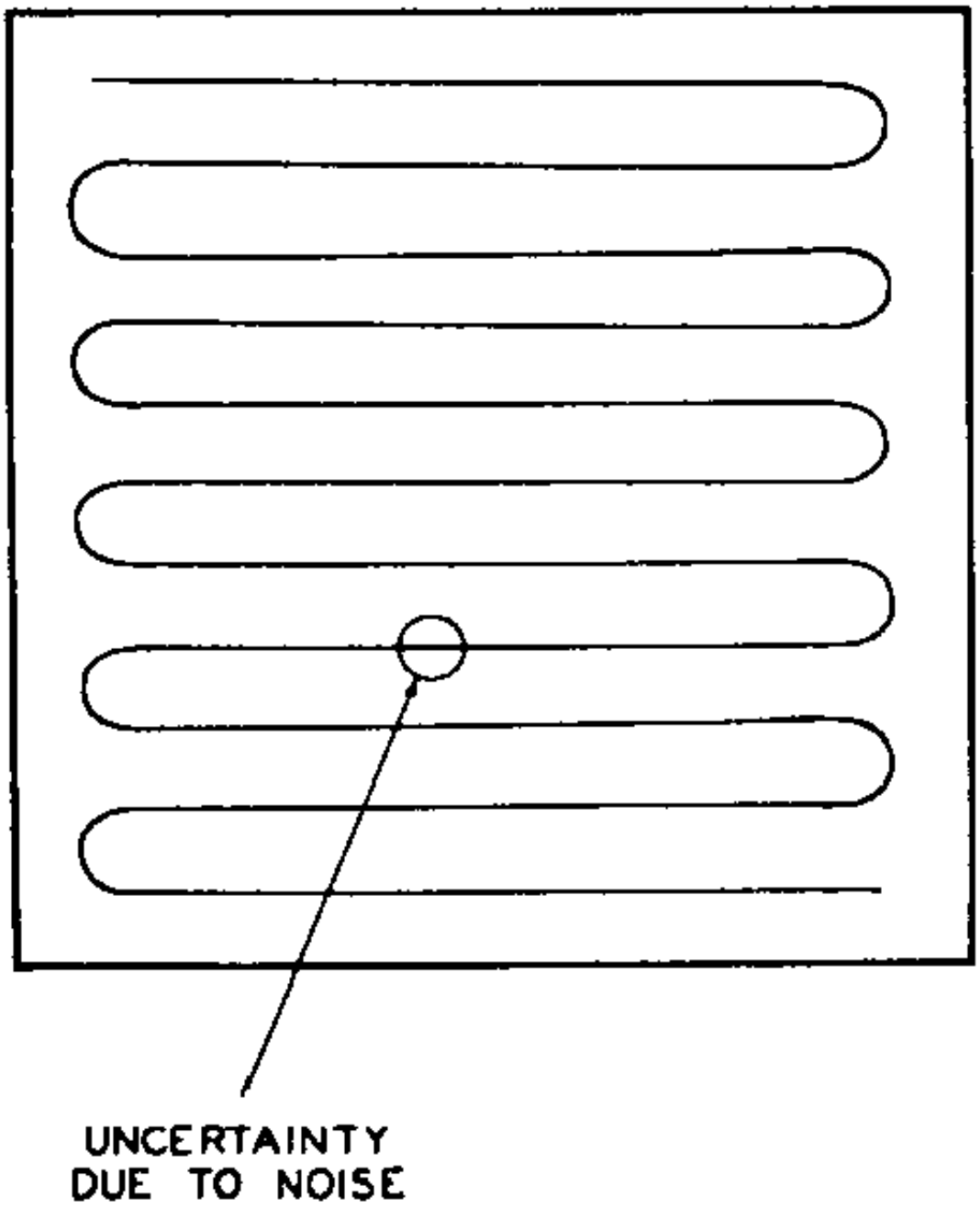

Figure 7: Mapping a line into a square. All communications maps signals into a higher dimensional space. Folding the signal into the space causes breaks between different messages. The same mathematics applies to binding sites, so the observed break is expected from information theory. The figure is from C. E. Shannon Communication in the Presence of Noise *Proc. IRE*, **37** 10-21, 1949.

CONCLUSIONS

Significance of the "Breakpoint Paper": **The binding rate constants and binding energy curves break at zero bits of information. This observation implies that the DNA binding protein interaction is a multi-dimensional coded system rather than a single dimensional chemical system as usually proposed in standard thermodynamics and biochemistry.**

Summary: **A breakpoint in the DNA binding energy/information curve shows that DNA binding is coded and not a simple binding reaction.**

Background: **Information theory is well suited to studying DNA binding by proteins, and this is generally recognized by the wide use of sequence logos, which were invented in this laboratory. The theory is about averages and not necessarily about the kinetics of binding. However, if the on-rate is essentially constant because the protein binds when it is close to its site, then the log of the off-rate should be related to the information in a binding site. In collaboration with Robert Fisher's laboratory, we tested this hypothesis on the Fis protein which we had characterized previously.**

Advance: **We discovered that the log of the off rate is indeed related to the information in a binding site, but so is the on-rate. We believe that the Fis protein requires DNA bending and that DNA bendability is related to the information in the binding site.**

We also observed that the linear relationship between the log of the off rate (or the binding energy) and the information 'breaks' at zero bits of information. This result is expected from a version of the Second Law of Thermodynamics which predicts that sequences that bind a protein have positive information while those that are not sites have information less than zero. The break in the curve at this point indicates that the theory is correct. Previous models of DNA binding do not predict the breakpoint.

Implications: **In 1949, Claude Shannon predicted break effects in communications systems. Observing the same kind of break in DNA binding strongly implies that we can understand DNA/protein interactions using the vast array of mathematical tools used to develop modern communications systems. That is, the result implies that a precise nanotechnology of DNA binding may be possible to engineer using knowledge we already have. The result is general and should apply to any specific interactions between molecules. This is a paradigm shift because the classical view is that molecular interactions are one dimensional. The new view is that the interactions are multidimensional and have evolved codes. If we can crack those codes we will have a general nanotechnology for molecular interactions.**