Genome 540 Discussion

Conor Camplisson

January 12th, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

Outline

- Related topics
  - Entropy
  - Information Theory

- Homework 2 overview

- Homework 1 & 2 questions

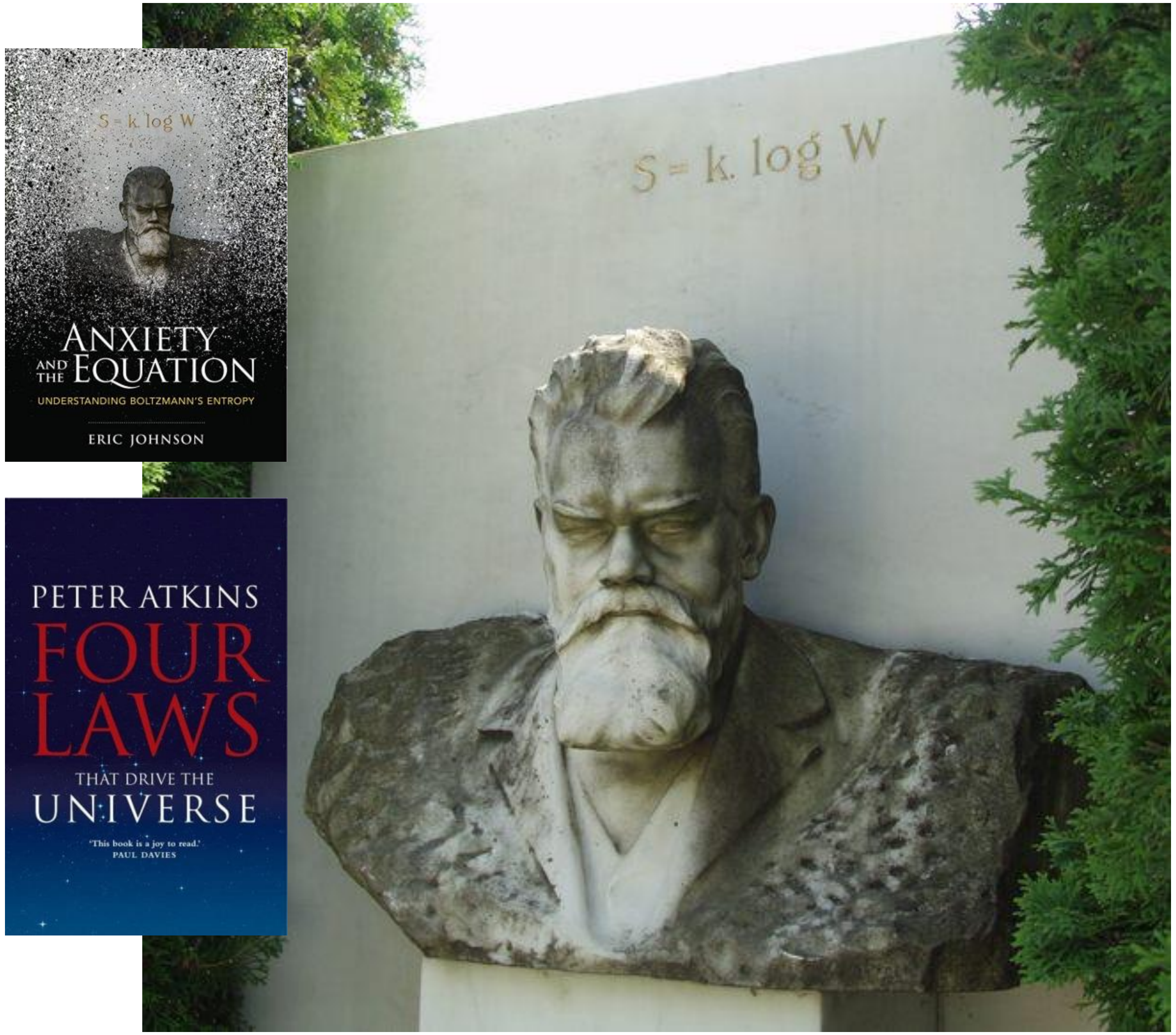
Outline

- Related topics
  - Entropy
  - Information Theory

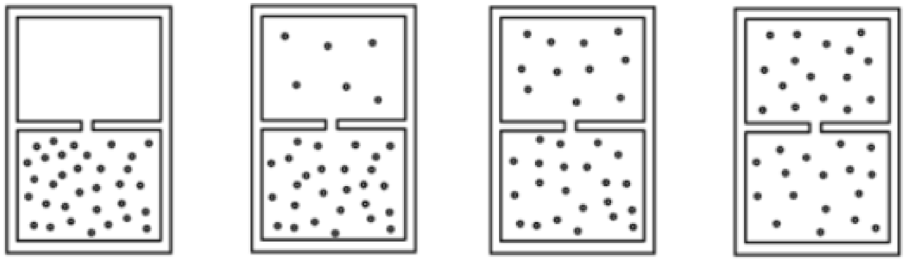
- Homework 2 overview

- Homework 1 & 2 questions

Entropy: microstates, macrostates



Why isn't all the air on one side of the room?



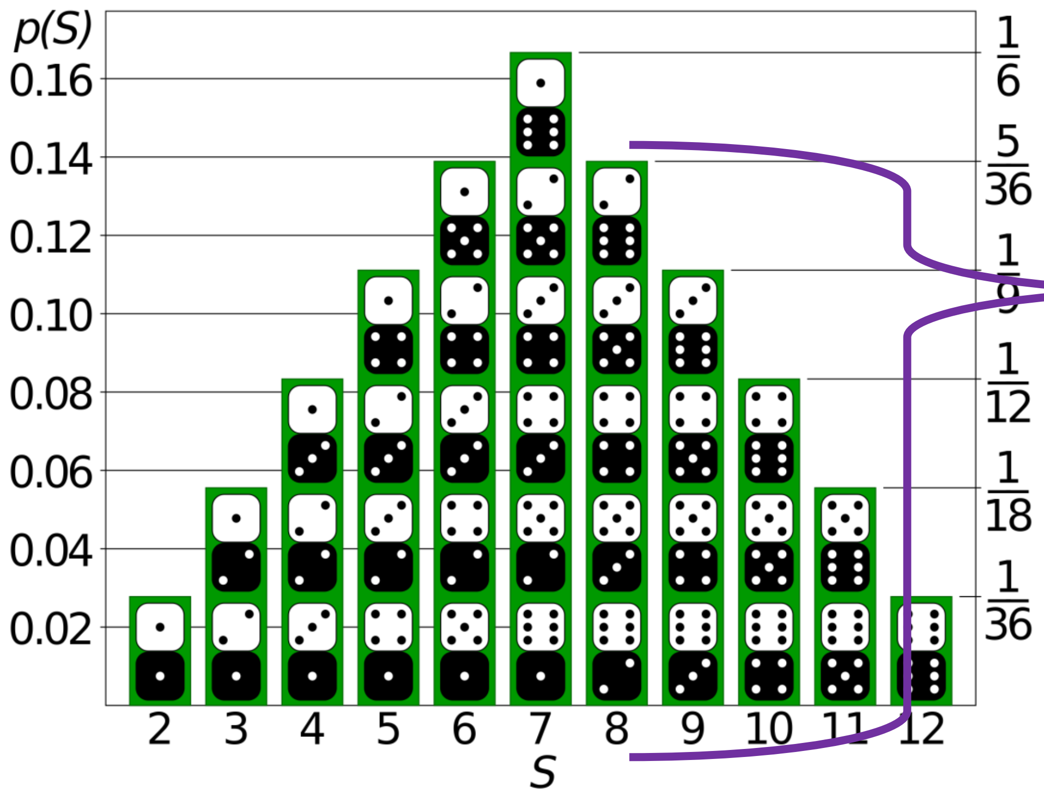
Macrostate:	0%	15%	30%	50%
Microstates:	1	few	many	overwhelming, only observed

52 cards → 52! = 1.55e66 orderings

Macrostate	# Microstates
In order	1
Out of order	~1.55e66

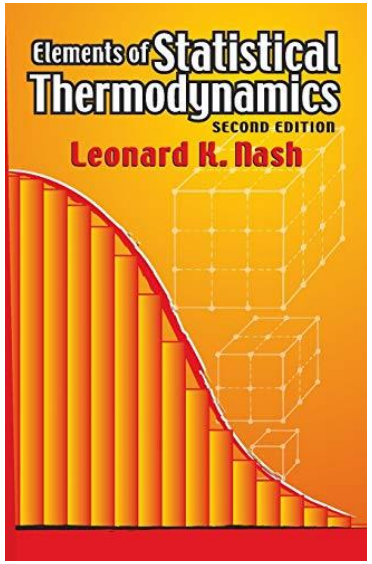
Entropy: microstates, macrostates

Rolling two dice



Microstates (configurations)

Macrostates (outcomes)



- Entropy is directly related to probability
- Microstate statistics explain macro phenomena
  - Quantum → thermodynamics, gases
  - Quantum → chemical equilibrium, kinetics

Information Theory

sending information over a noisy channel

Information Theory

sending information over a noisy channel

Harry Nyquist

Certain Topics in Telegraph Transmission Theory

H. NYQUIST, MEMBER, A. I. E. E.

Classic Paper

Bell Laboratories

Communication in the Presence of Noise*

CLAUDE E. SHANNON†, MEMBER, IRE

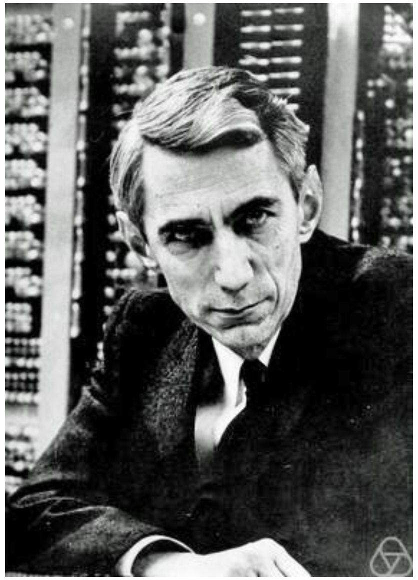
I. Introduction

ENERAL COMMUNICATIONS system is hown schematically in Fig. 1. It consists essen- ially of five elements.

information source. The source selects one mes- n a set of possible messages to be transmitted to ving terminal. The message may be of various r example, a sequence of letters or numbers, as aphy or teletype, or a continuous function of , as in radio or telephony.

e transmitter. This operates on the message in y and produces a signal suitable for transmis- he receiving point over the channel. In teleph-

- Radio astronomy
- Transistor
- LASER
- Photovoltaic cell
- Charge-coupled device (CCD)
- UNIX, C, C++, AWK, others
- 9 Nobel Prizes
- *Information Theory*



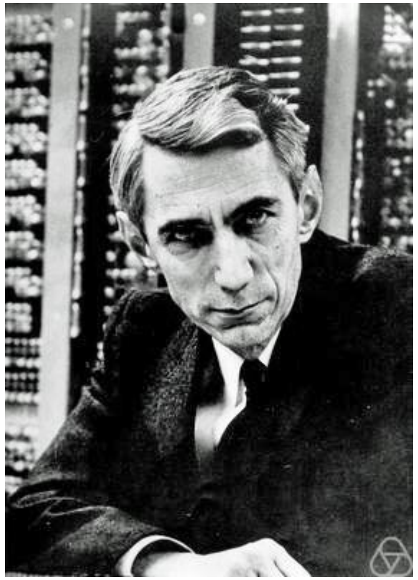
Claude Shannon

Information Theory

Quantifying information

Heads ✗

Tails ✓



Claude Shannon

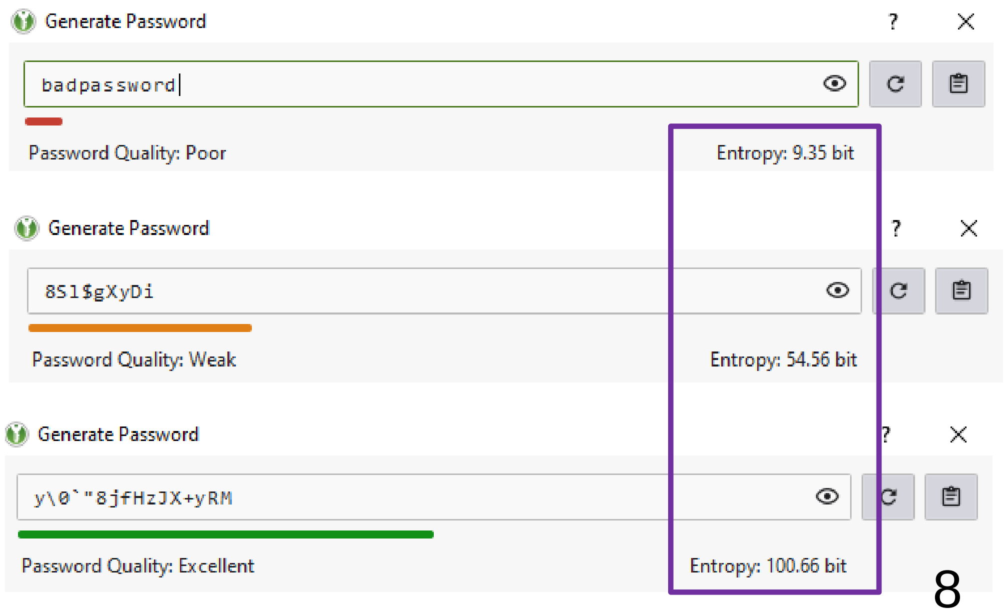
Information entropy

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

H(toss) = -(p(heads) * log2 (p(heads) + p(tails) * log2(p(tails))

H(toss) = -(0.5*log2(0.5) + 0.5*log2(0.5)) = **1.0 bit**

"binary digit" → "bit"



Information gain

What do you learn from a coin flip?

Heads ✗

Tails ☑

Information gain

$$I(x) = -\log_2(P(X=x))$$

Fair coin: $I(\textit{Tails}) = -\log_2(1/2) =$ **1 bit**

Less probable events are more informative!

prob_heads	info_gain
0.10	3.32
0.20	2.32
0.30	1.74
0.40	1.32
0.50	1.00
0.60	0.74
0.70	0.51
0.80	0.32
0.90	0.15
1.00	0.00

Information gain

I'm thinking of a card...

hint	num_cards	total_cards	prob	info_bits
red	26	52	0.50	1.00
not_face	40	52	0.77	0.38
heart	13	52	0.25	2.00
5	4	52	0.08	3.70
5_hearts	1	52	0.02	5.70

Information gain

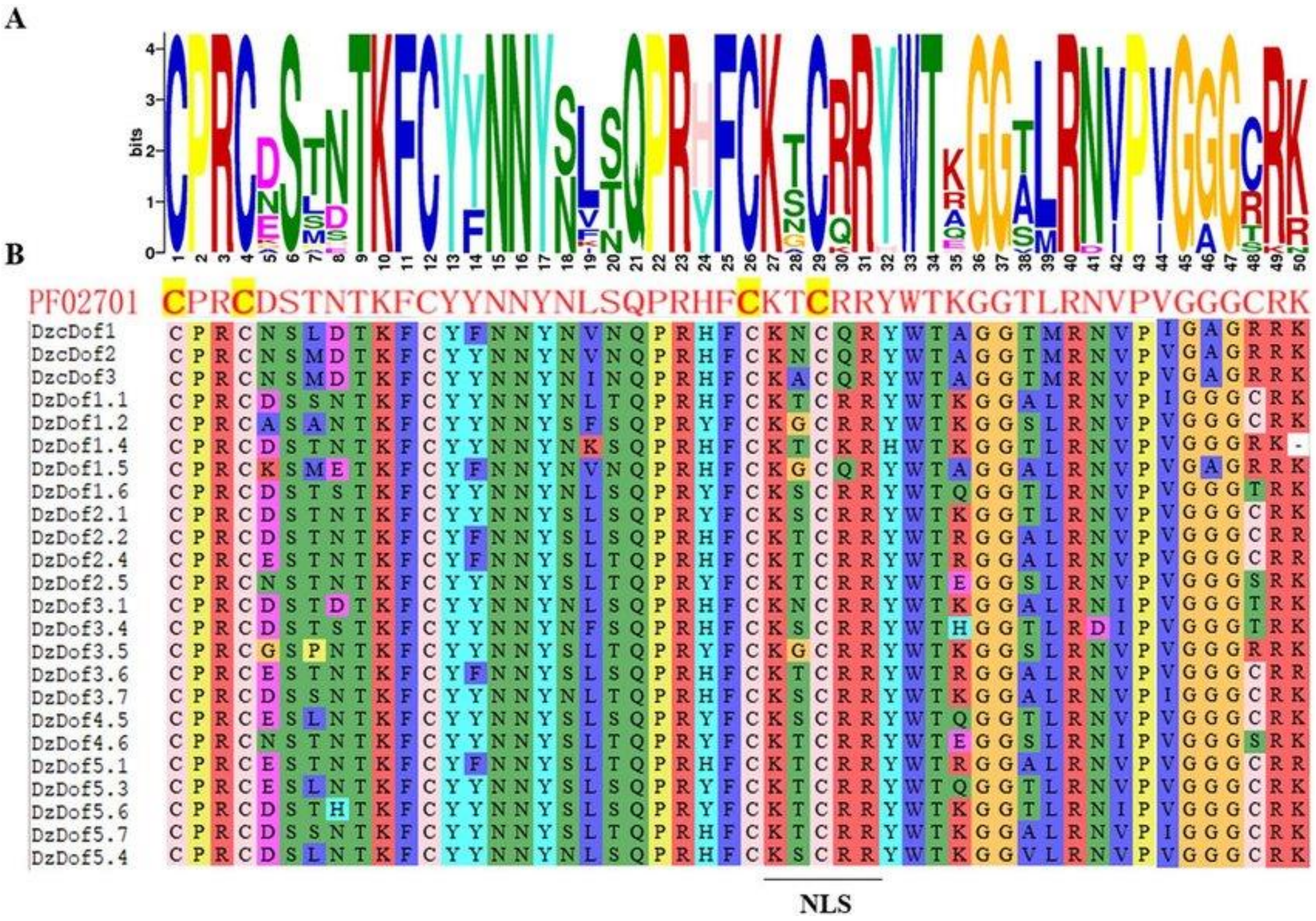
$$I(x) = -\log_2(P(X=x))$$

Entropy, information, and probabilities are linked

Less probable events are more informative!

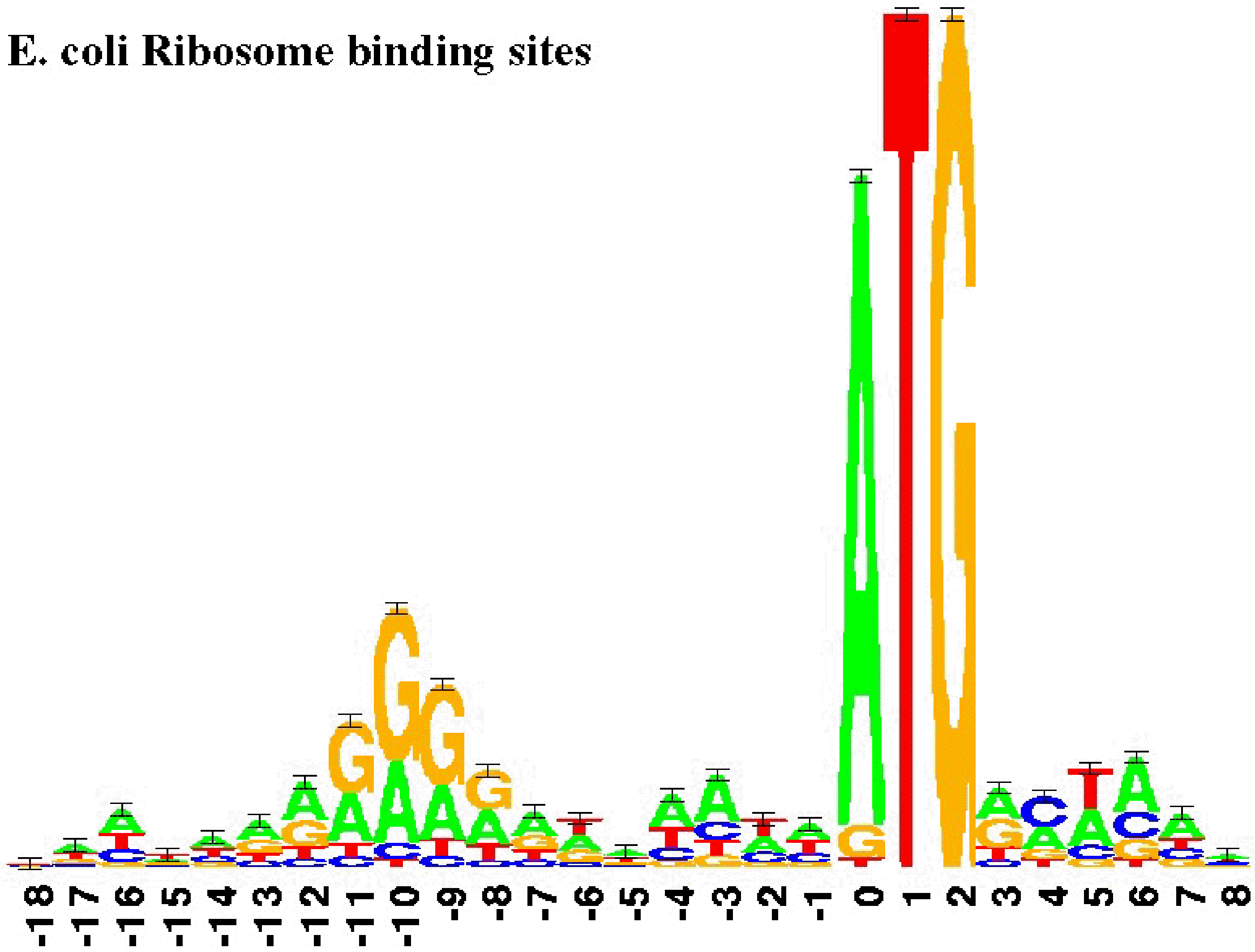
Information ~ a loss in entropy / a resolution of uncertainty

Sequence Logos



- Population of sequences
  - Nucleotide, amino acid sequences
- Information entropy at each site
  - Evolution *selects* a residue
  - Loss of entropy at that site
- Visualize both identity and importance

Sequence Logos

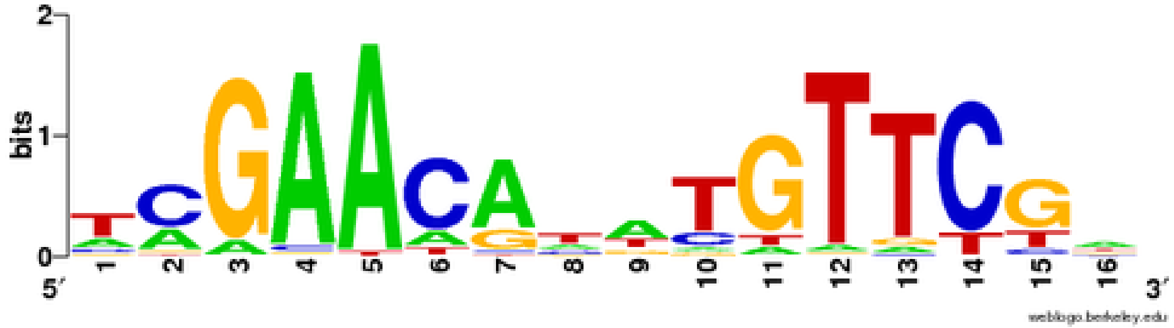


- Population of sequences
  - Nucleotide, amino acid sequences

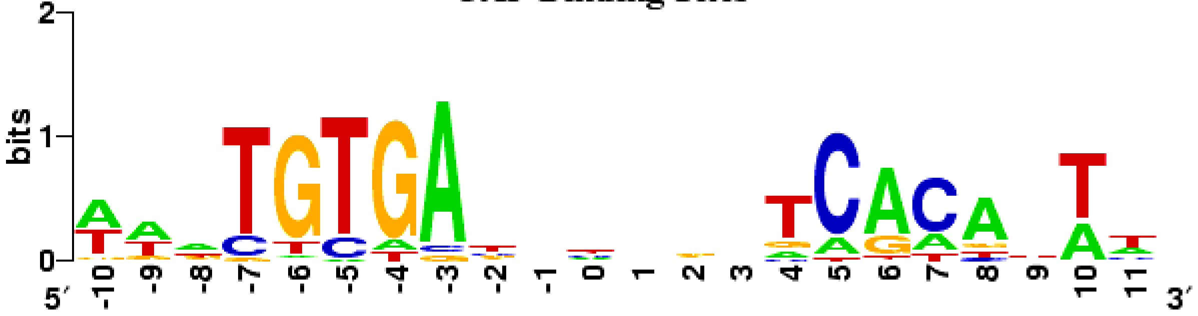
- Information entropy at each site
  - Evolution *selects* a residue
  - Loss of entropy at that site

- Visualize both identity and importance

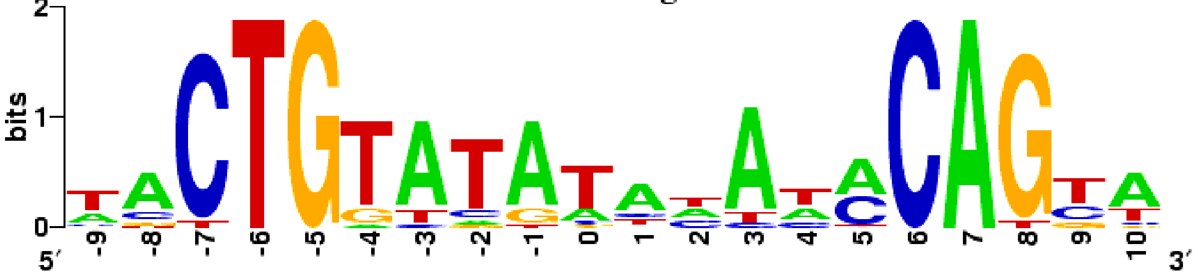
Sequence Logos



CAP Binding Sites



19 lexA Binding Sites



Y-axis = loss of entropy ~ information

Identity of a *random* residue out of {A, C, G, T} contains 2 bits of info:

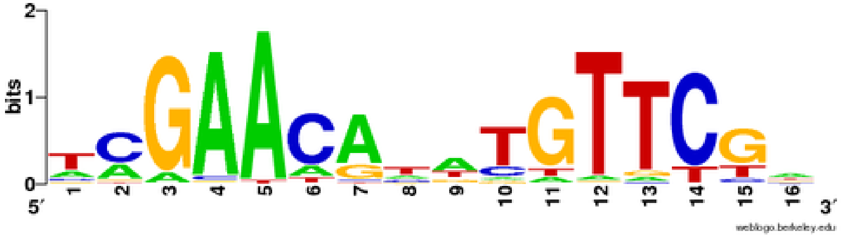
$$-\log2(1/4) = 2.0$$

A value of 0.0 means no entropy was lost, uniform probabilities {A, C, G, T}

A value of 2.0 means all entropy was lost, identity is p=1.00 for selected residue

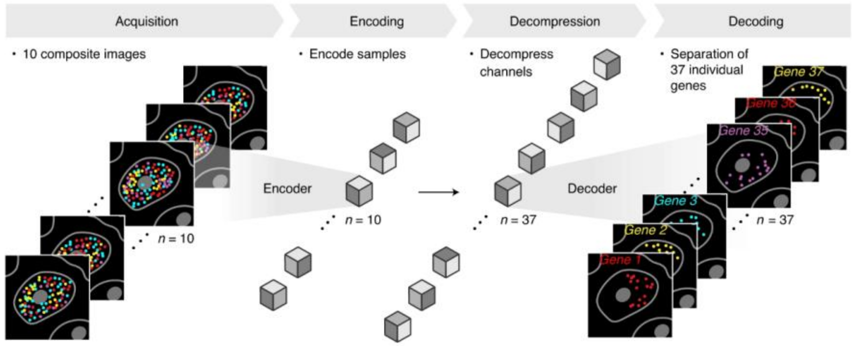
Examples of info theory's use in research

Sequence logos



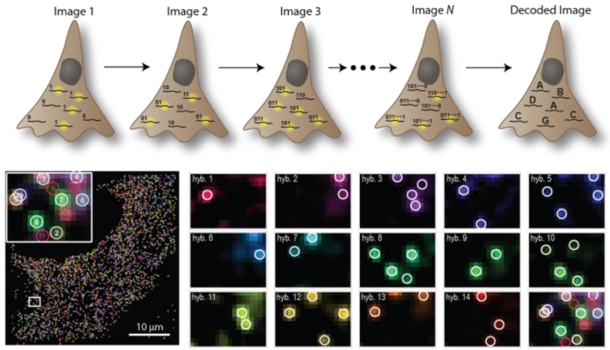
Visualize loss of sequence entropy at sites

Compressed sensing & FISH



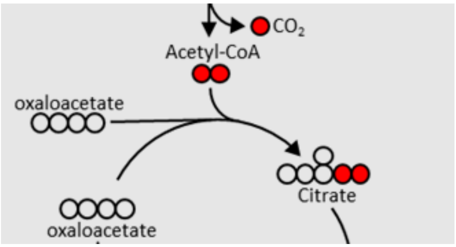
Gather a few microscope images, impute many distinct images

MERFISH



Linear Block Code theory (Hamming), error-correcting codes

Metabolic pathway analysis



Maximize information gain when choosing carbon atoms to trace

Outline

- Related topics
  - Entropy
  - Information Theory
- Homework 2 overview
- Homework 1 & 2 questions

Homework 2 Overview

Part one: write a new program

- read in a file in FASTA format
- determine the frequencies of the nucleotides and dinucleotides (based on the forward strand) and the length of the sequence
- produce three simulated sequences based on the length and nucleotide or dinucleotide frequency of the original sequence
  - ‘Equal frequency’ model
  - Order 0 Markov model
  - Order 1 Markov model
- output three files in FASTA format containing the simulated sequence

Homework 2 Overview

order-0 Markov

“Equal frequency” model

```
Nucleotide Frequencies:
A=0.2500
C=0.2500
G=0.2500
T=0.2500
```

```
Fasta 1: CP003913.fna
>gi|440453185|gb|CP003913.1| Mycoplasma pneumoniae M129-B7, complete genome
*=816373
A=249201
C=162924
G=163697
T=240551
N=0

Nucleotide Frequencies:
A=0.3053
C=0.1996
G=0.2005
T=0.2947
```

order-1 Markov

Dinucleotide Count Matrix:

	A	C	G	T
A	A=98512	50763	47914	52012
C	C=53047	36681	26746	46450
G	G=40870	37148	36764	48915
T	T=56772	38332	52273	93173

CGACTA

Dinucleotide Frequency Matrix:

A=0.1207	0.0622	0.0587	0.0637	
C=0.0650	0.0449	0.0328	0.0569	= 1
G=0.0501	0.0455	0.0450	0.0599	
T=0.0695	0.0470	0.0640	0.1141	

Conditional Frequency Matrix:

A=0.3953	0.2037	0.1923	0.2087	= 1
C=0.3256	0.2251	0.1642	0.2851	= 1
G=0.2497	0.2269	0.2246	0.2988	= 1
T=0.2360	0.1594	0.2173	0.3873	= 1

Homework 2 Overview

Part two: run your HW1 program on three simulated genomes

- Run your HW1 program three times, using as input:
  - Human 10-Mb segment + simulated ‘equal frequency’ genome
  - Human 10-Mb segment + simulated Mouse Markov-0
  - Human 10-Mb segment + simulated Mouse Markov-1
- Given observed matches between the Human and simulated genomes, what can you conclude about the statistical significance of matches between the orthologous mouse and human regions in homework 1?

Outline

- Related topics
  - Entropy
  - Information Theory

- Homework 2 overview

- Homework 1 & 2 questions

Homework 1&2 Questions ?

p_{10}	AAACCGTACACTGGGTTCAAGAGATTTCCC
p_{11}	AACCGTACACTGGGTTCAAGAGATTTCCC
p_{28}	AAGAGATTTCCC
p_{17}	ACACTGGGTTCAAGAGATTTCCC
p_{12}	ACCGTACACTGGGTTCAAGAGATTTCCC
p_{1}	ACCTGCACTAAACCGTACACTGGGTTCAAGAGATTTCCC
p_{7}	ACTAAACCGTACACTGGGTTCAAGAGATTTCCC
p_{19}	ACTGGGTTCAAGAGATTTCCC
p_{29}	AGAGATTTCCC
p_{31}	AGATTTCCC
p_{33}	ATTTCCC
p_{27}	CAAGAGATTTCCC
	⋮

Observed Dinuc Freqs

	A	C	G	T
A	0.099	0.051	0.069	0.078
C	0.073	0.052	0.011	0.069
G	0.059	0.043	0.052	0.050
T	0.066	0.059	0.072	0.098

Reminders

- Homework 1 due this Sunday Jan. 15, 11:59 pm
- Homework 2 will be posted tonight (Jan. 12)