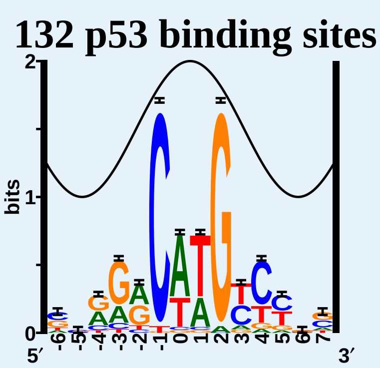
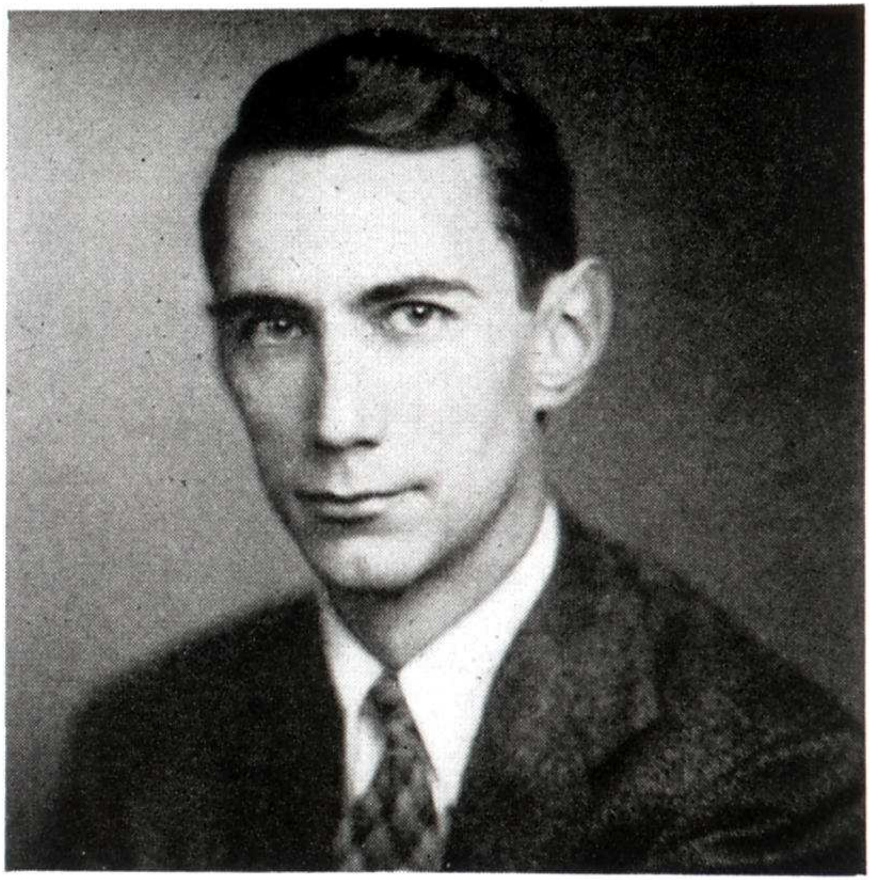
Why Do Restriction Enzymes Prefer 4 and 6 Base DNA Sequences?

Thomas D. Schneider, Ph.D.
Vishnu Jejjala, Ph.D.

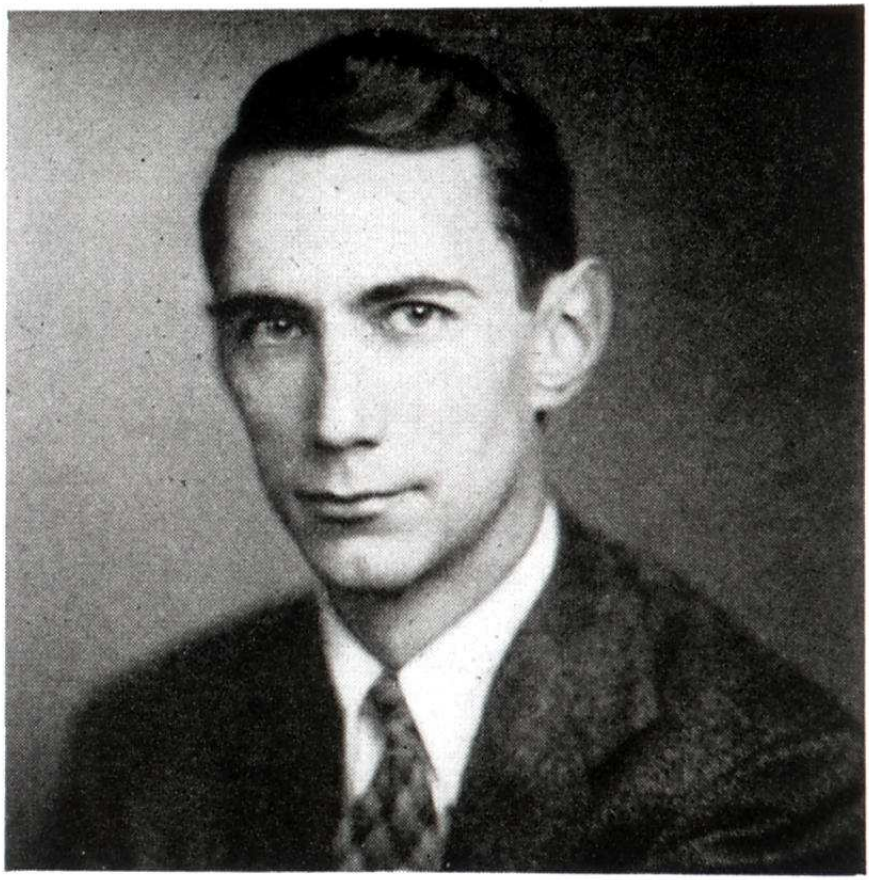
Molecular Information Theory Group
Center for Cancer Research
RNA Biology Laboratory
National Cancer Institute
Frederick, MD 21702-1201
and
University of the Witwatersrand
Johannesburg, South Africa





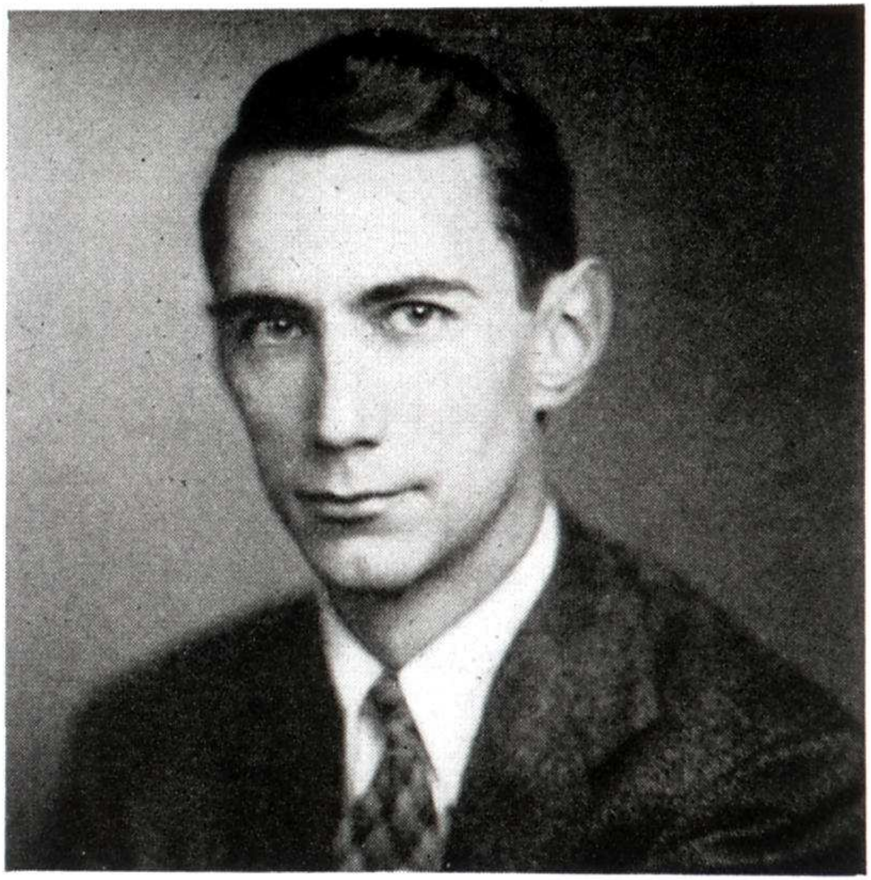
CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory



CLAUDE E. SHANNON

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.
2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measures entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.
3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

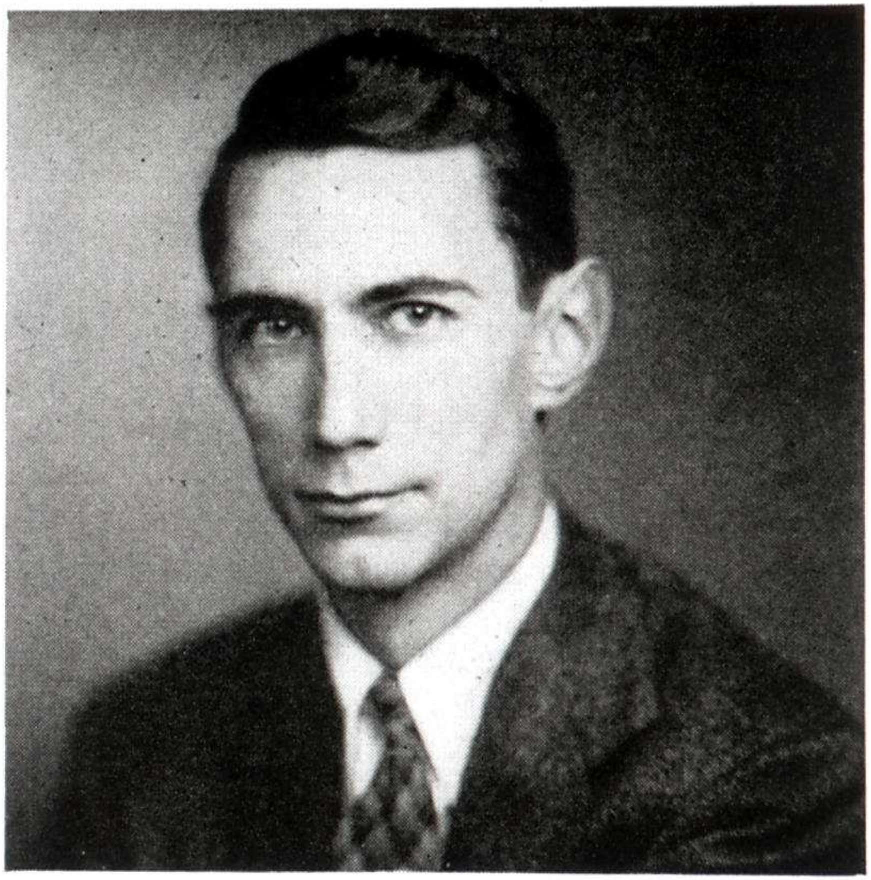
The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\begin{aligned}\log_2 M &= \log_{10} M / \log_{10} 2\\ &= 3.32 \log_{10} M,\end{aligned}$$

[1]Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Trans.*, v. 47, April 1928, p. 617.
[2]Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

1

- April 30, 1916 - February 24, 2001
- Founded Information Theory
- Important papers: 1948



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory
- Important papers: 1948 , 1949

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental p
proximately a message
to or are correlated acc
aspects of communicat
message is one *selecte*
possible selection, not j
If the number of m
can be regarded as a n
choices being equally
function. Although thi
statistics of the messa
essentially logarithmic
The logarithmic me

1. It is practically n
of relays, etc., te
adding one relay
logarithm of this
doubles the loga

2. It is nearer to ou
tuitively measure
two punched car
channels twice tl

3. It is mathematic
rithm but would

The choice of a log
base 2 is used the resu
J. W. Tukey. A device
information. *N* such de
If the base 10 is used tl

[1]Nyquist, H., "Certain Fa
Telegraph Transmission The
[2]Hartley, R. V. L., "Trans

Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Classic Paper

A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two "function spaces," and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concerning expansion and compression of bandwidth and the threshold effect. Formulas are found for the maximum rate of transmission of binary digits over a system when the signal is perturbed by various types of noise. Some of the properties of "ideal" systems which transmit at this maximum rate are discussed. The equivalent number of binary digits per second for certain information sources is calculated.

I. INTRODUCTION

A general communications system is shown schematically in Fig. 1. It consists of essentially of five elements.

1) An Information Source: The source selects one message from a set of possible messages to be transmitted to the receiving terminal. The message may be of various types; for example, a sequence of letters or numbers, as in telegraphy or teletype, or a continuous function of time $f(t)$, as in radio or telephony.

2) The Transmitter: This operates on the message in some way and produces a signal suitable for transmission to the receiving point over the channel. In telephony, this operation consists of merely changing sound pressure into a proportional electrical current. In telegraphy, we have a encoding operation which produces a sequence of dots, dashes, and spaces corresponding to the letters of the message. To take a more complex example, in the case of multiplex PCM telephony the different speech functions must be sampled, compressed, quantized and encoded, and finally interleaved properly to construct the signal.

3) The Channel: This is merely the medium used to transmit the signal from the transmitting to the receiving point. It may be a pair of wires, a coaxial cable, a band of radio frequencies, etc. During transmission, or at the receiving terminal, the signal may be perturbed by noise or distortion. Noise and distortion may be differentiated on the basis that distortion is a fixed operation applied to the signal, while noise involves statistical and unpredictable

This paper is reprinted from the PROCEEDINGS OF THE IRE, vol. 37, no. 1, pp. 10–21, Jan. 1949.
Publisher Item Identifier S 0018-9219(98)01299-7.

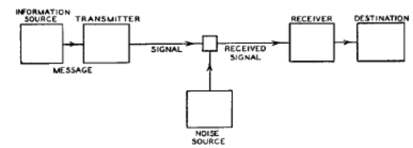


Fig. 1. General communications system.

perturbations. Distortion can, in principle, be corrected by applying the inverse operation, while a perturbation due to noise cannot always be removed, since the signal does not always undergo the same change during transmission.

4) The Receiver: This operates on the received signal and attempts to reproduce, from it, the original message. Ordinarily it will perform approximately the mathematical inverse of the operations of the transmitter, although they may differ somewhat with best design in order to combat noise.

5) The Destination: This is the person or thing for whom the message is intended.

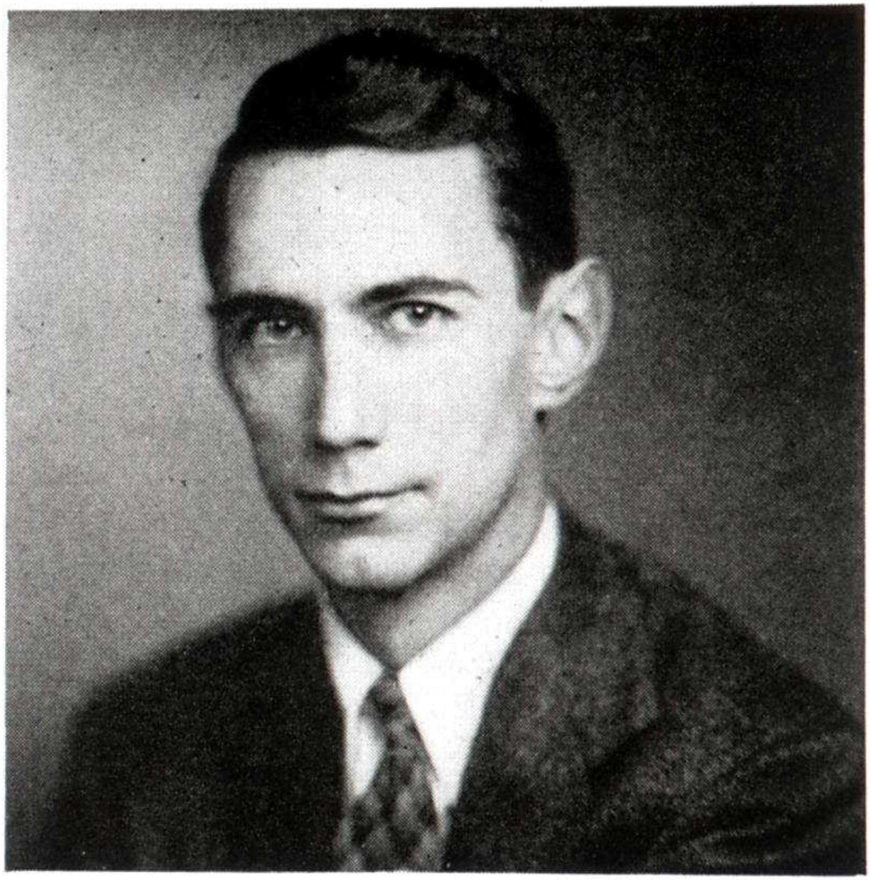
Following Nyquist[1] and Hartley,[2] it is convenient to use a logarithmic measure of information. If a device has n possible positions it can, by definition, store $\log_b n$ units of information. The choice of the base b amounts to a choice of unit, since $\log_b n = \log_b c \log_c n$. We will use the base 2 and call the resulting units binary digits or bits. A group of m relays or flip-flop circuits has 2^m possible sets of positions, and can therefore store $\log_2 2^m = m$ bits.

If it is possible to distinguish reliably M different signal functions of duration T on a channel, we can say that the channel can transmit $\log_2 M$ bits in time T. The *rate* of transmission is then $\log_2 M/T$. More precisely, the *channel capacity* may be defined as

$$C = \lim_{T\to\infty} \frac{\log_2 M}{T}. \quad (1)$$

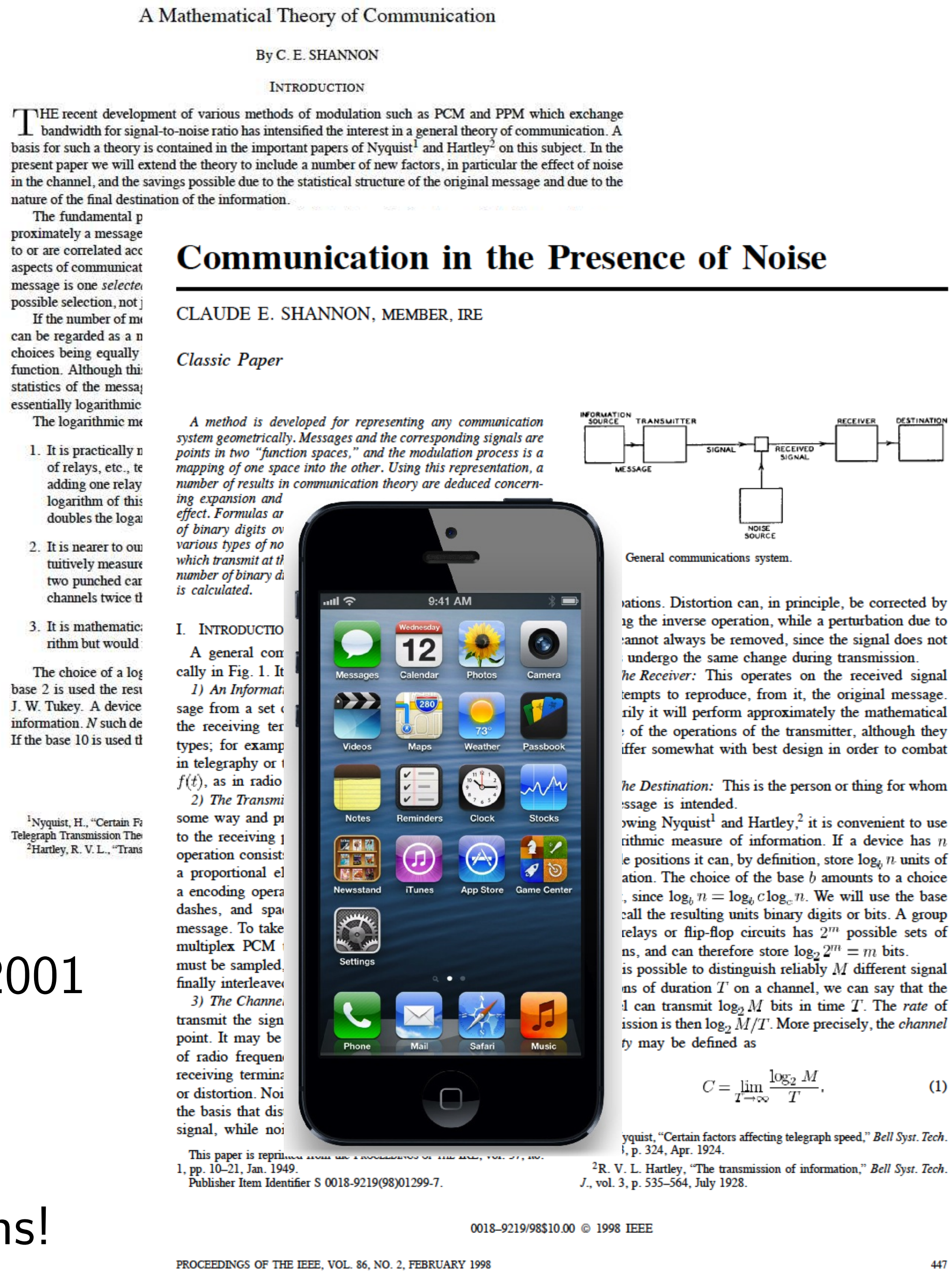
[1]H. Nyquist, "Certain factors affecting telegraph speed," *Bell Syst. Tech. J.*, vol. 3, p. 324, Apr. 1924.

[2]R. V. L. Hartley, "The transmission of information," *Bell Syst. Tech. J.*, vol. 3, p. 535–564, July 1928.



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory
- Important papers: 1948 , 1949
- Result: modern communications!



A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Classic Paper

A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two "function spaces," and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concern-

General communications system.

$$C = \lim_{T \to \infty} \frac{\log_2 M}{T}. \quad (1)$$

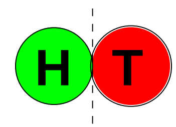
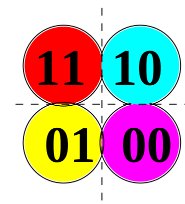
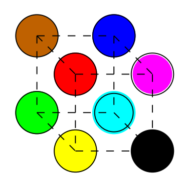
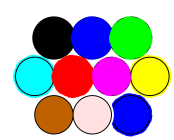
Publisher Item Identifier S 0018-9219(98)01299-7.

²R. V. L. Hartley, "The transmission of information," *Bell Syst. Tech. J.*, vol. 3, p. 535–564, July 1928.

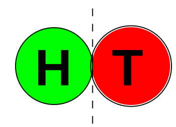
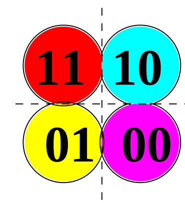
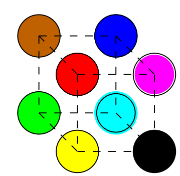
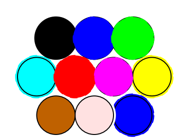
0018–9219/98$10.00 © 1998 IEEE

PROCEEDINGS OF THE IEEE, VOL. 86, NO. 2, FEBRUARY 1998 447

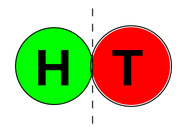
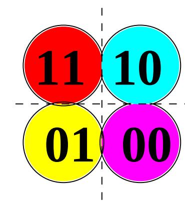
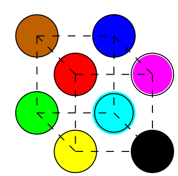
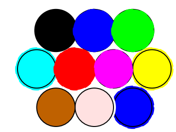
Information Theory: One-Minute Lesson

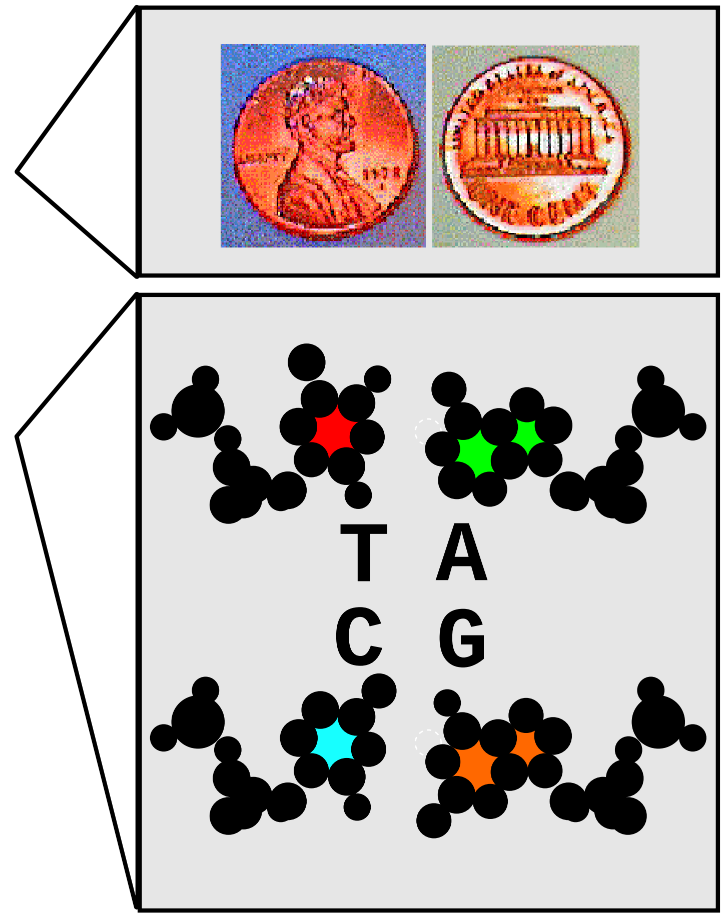
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

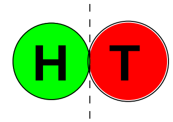
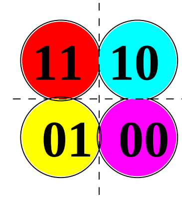
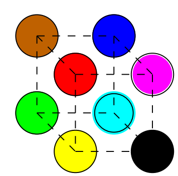
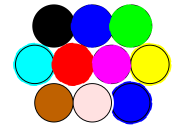
Information Theory: One-Minute Lesson

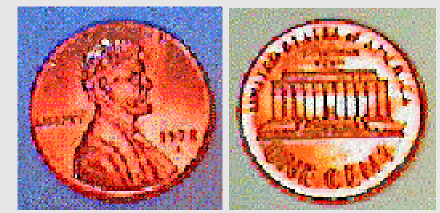
number of symbols	number of bits	example
M	B	
2	1	 
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

Information Theory: One-Minute Lesson

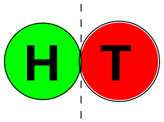
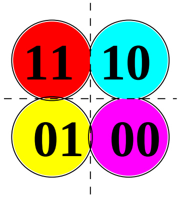
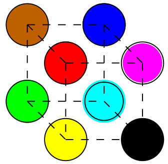
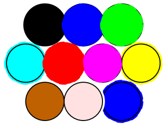
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

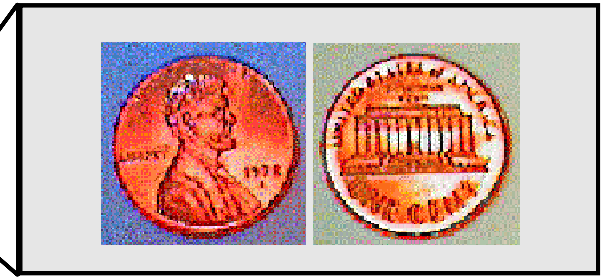


number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

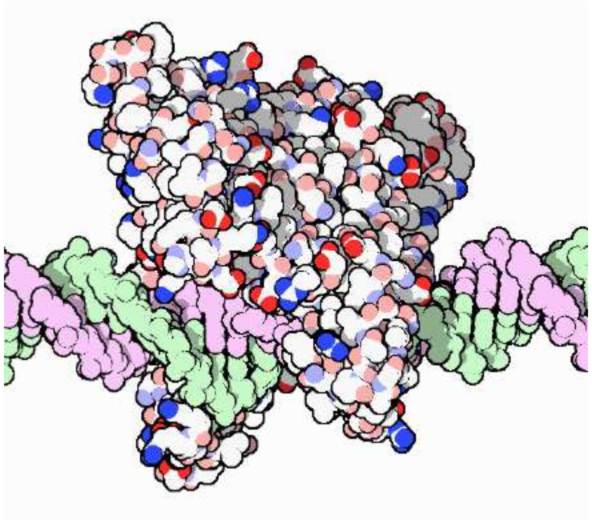


Information Theory: One-Minute Lesson

number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	



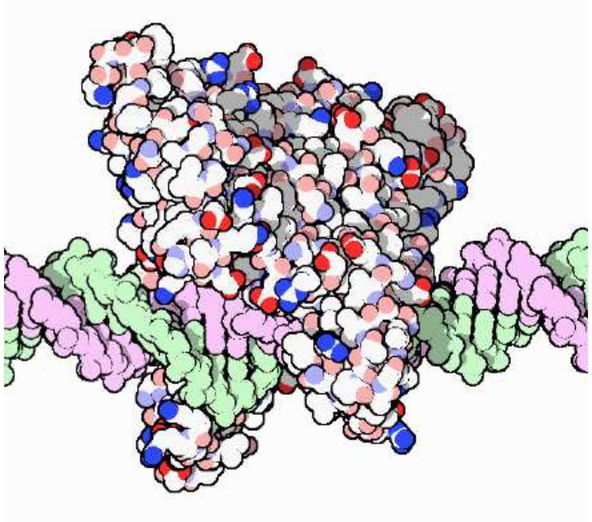
Information of EcoRI DNA Binding

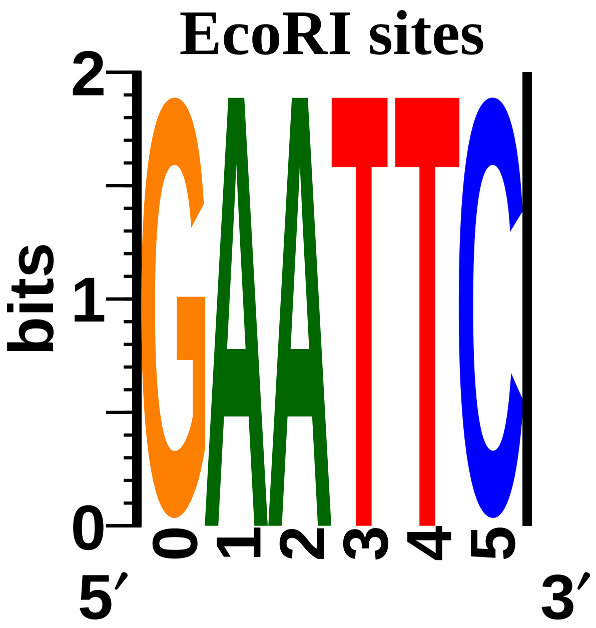


- EcoRI - restriction enzyme

Information of EcoRI DNA Binding

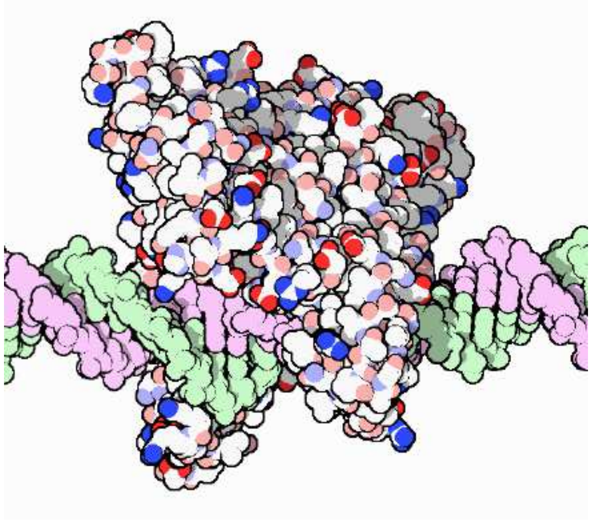
- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$

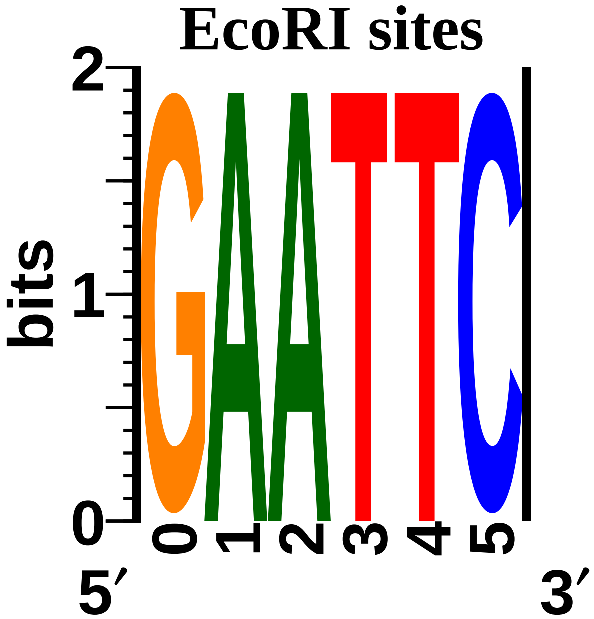




Information of EcoRI DNA Binding

- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$
- information required:
  6 bases $\times$ 2 bits per base $=$ $\boxed{\text{12 bits}}$

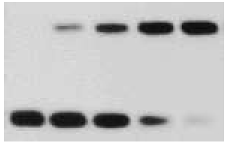




Energy Dissipation by EcoRI

- Measured specific binding constant:

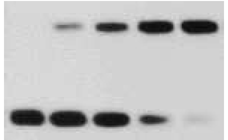
$$K_{spec} = 1.6 \times 10^5$$



Energy Dissipation by EcoRI

- Measured specific binding constant:

$$K_{spec} = 1.6 \times 10^5$$



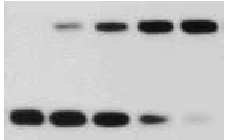
- Average energy dissipated by one molecule as it binds:

$$\Delta G^{\circ}_{spec} = -k_{\mathsf{B}} T \ln K_{spec} \qquad \text{(joules per binding)}$$

Energy Dissipation by EcoRI

- Measured specific binding constant:

$$K_{spec} = 1.6 \times 10^5$$



- Average energy dissipated by one molecule as it binds:

$$\Delta G^\circ_{spec} = -k_{\mathsf{B}} T \ln K_{spec} \qquad \text{(joules per binding)}$$

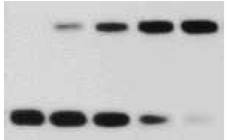
- The Second Law of Thermodynamics as a conversion factor:

$$\mathcal{E}_{min} = k_{\mathsf{B}} T \ln 2 \qquad \text{(joules per bit)}$$

Energy Dissipation by EcoRI

- Measured specific binding constant:

$$K_{spec} = 1.6 \times 10^5$$



- Average energy dissipated by one molecule as it binds:

$$\Delta G^\circ_{spec} = -k_\mathsf{B} T \ln K_{spec} \qquad \text{(joules per binding)}$$

- The Second Law of Thermodynamics as a conversion factor:

$$\mathcal{E}_{min} = k_\mathsf{B} T \ln 2 \qquad \text{(joules per bit)}$$

- Number of bits that could have been selected:

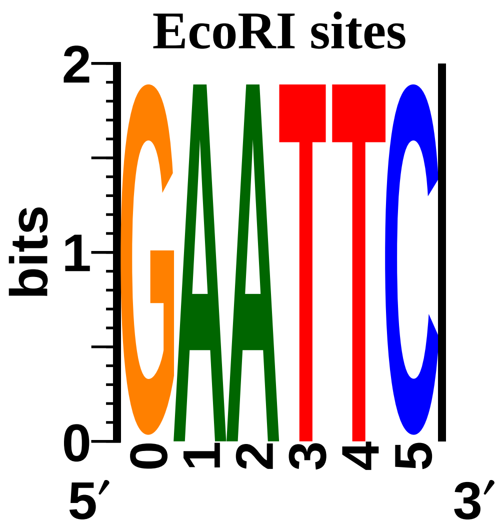
$$\begin{aligned} R_{energy} &= -\Delta G^\circ / \mathcal{E}_{min} \\ &= k_\mathsf{B} T \ln K_{spec} / k_\mathsf{B} T \ln 2 \\ &= \log_2 K_{spec} \qquad \Leftarrow \text{SO SIMPLE!} \\ &= \boxed{\text{17.3 bits per binding}} \end{aligned}$$

Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices

Information/Energy = Efficiency of EcoRI

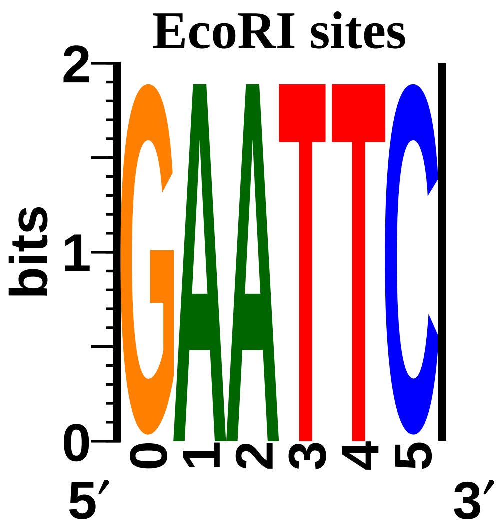
EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.



Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.

Efficiency is
'WORK' DONE / ENERGY DISSIPATED

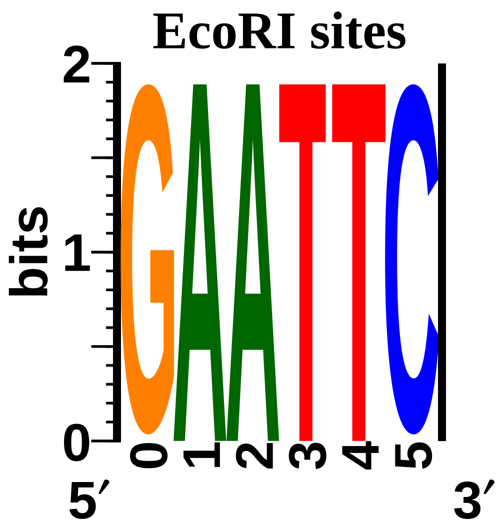


Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.

Efficiency is
'WORK' DONE / ENERGY DISSIPATED

$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$



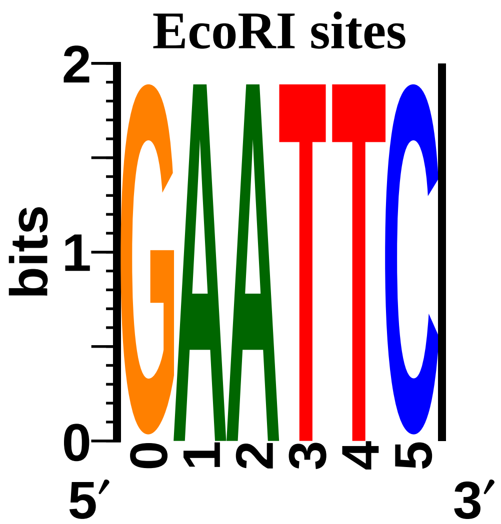
Information/Energy = Efficiency of EcoRI = 70%

EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.

Efficiency is
'WORK' DONE / ENERGY DISSIPATED

$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

The efficiency is 70%.



Information/Energy = Efficiency of EcoRI = 70%

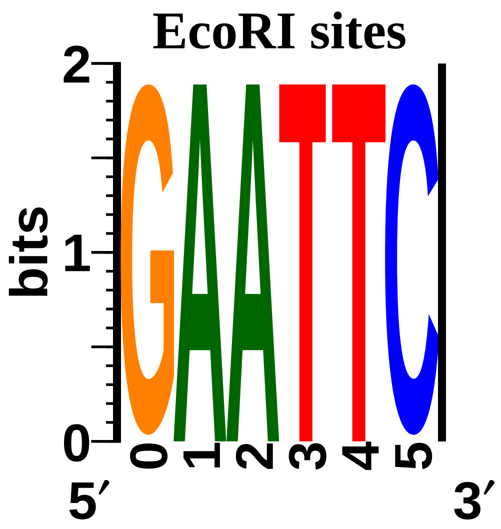
EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.

Efficiency is
'WORK' DONE / ENERGY DISSIPATED

$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

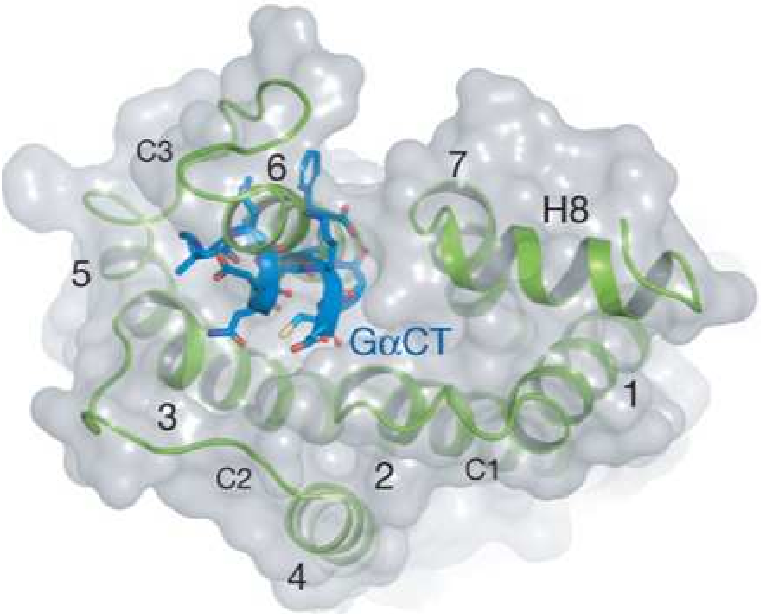
The efficiency is 70%.

18 out of 19 DNA binding proteins give ∼70% efficiency.

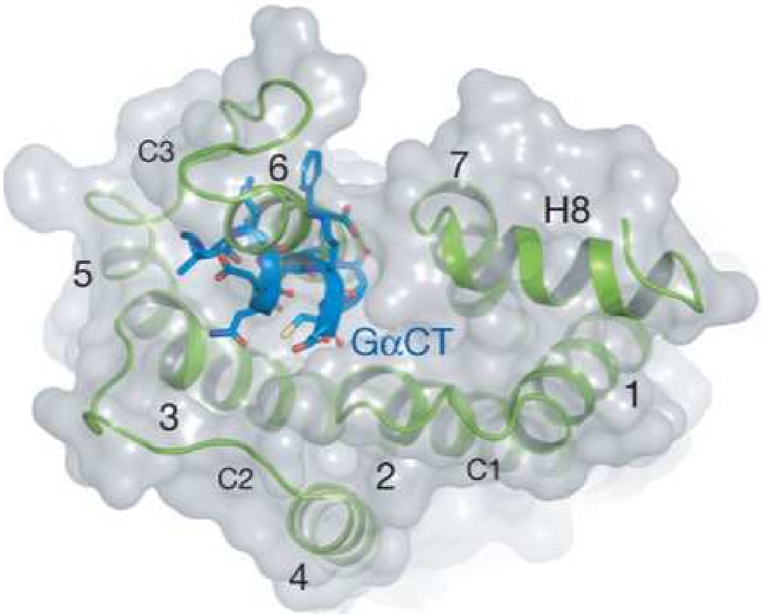


Rhodopsin Shape Change

Dark State



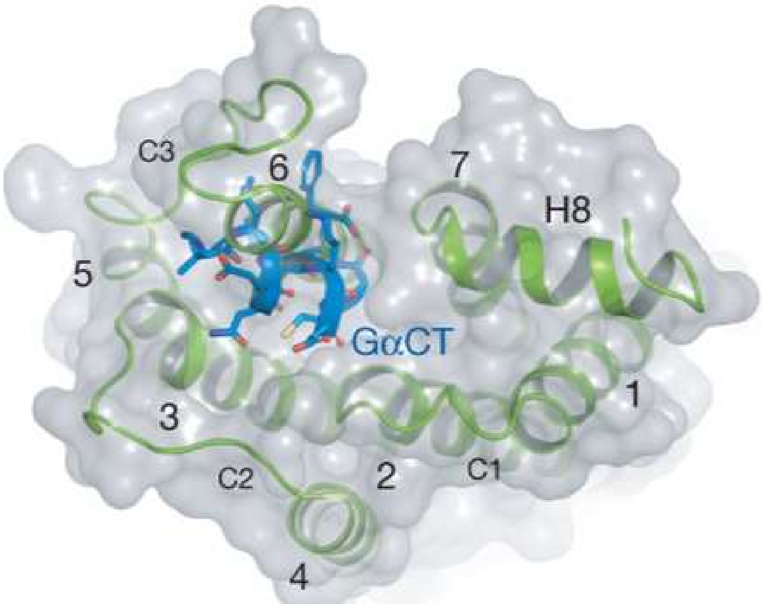
Dark State



$$\mathbf{h}\nu$$

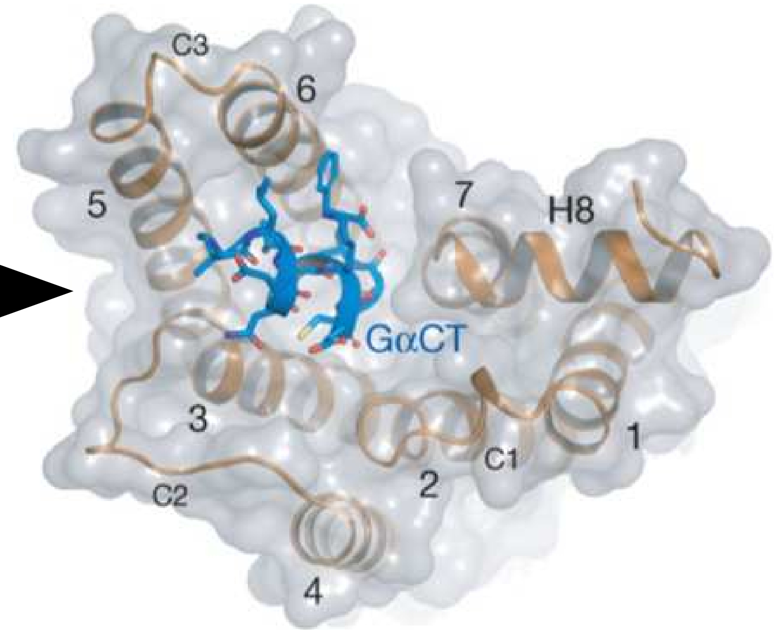
Rhodopsin Shape Change

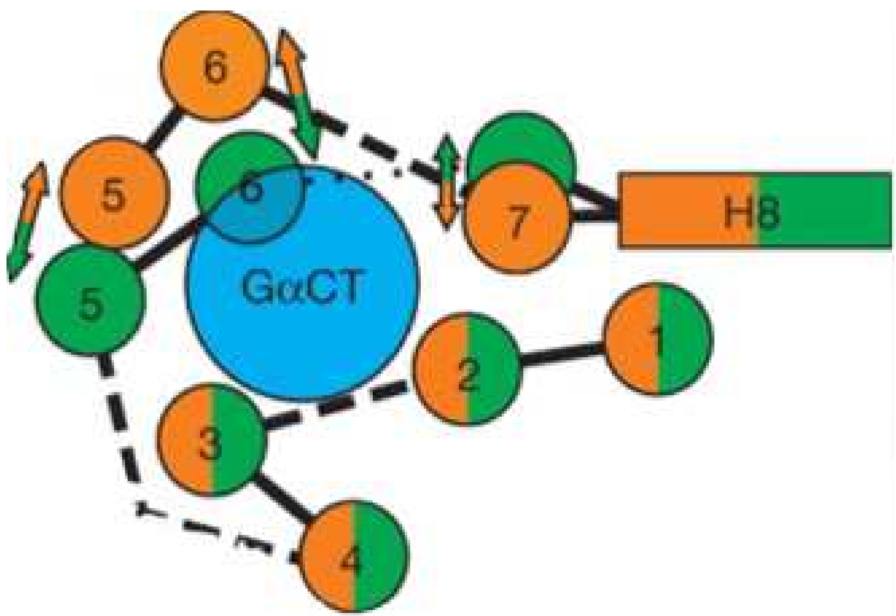
Dark State



$\mathbf{h}\nu$ →

After Photon - Light State



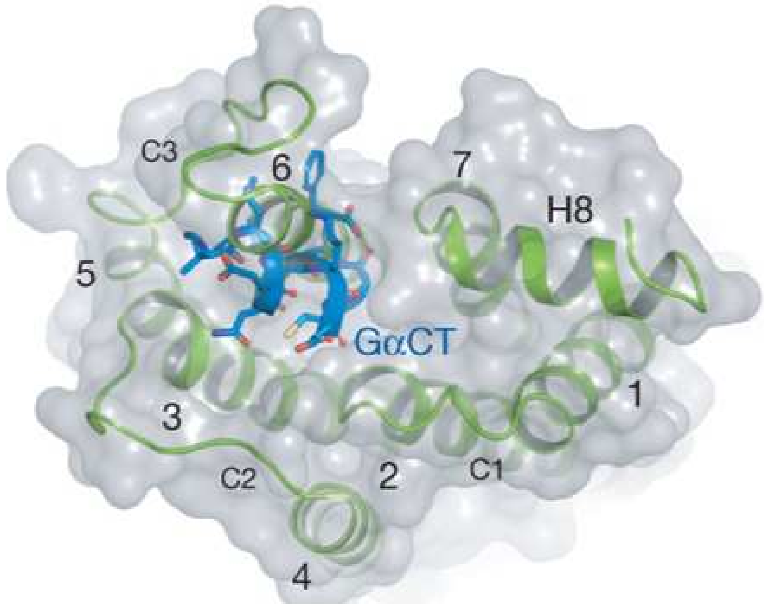


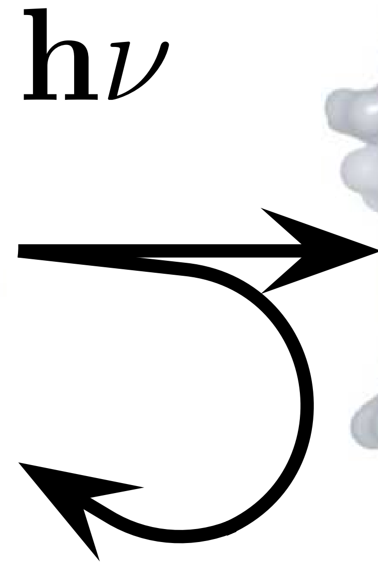
P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

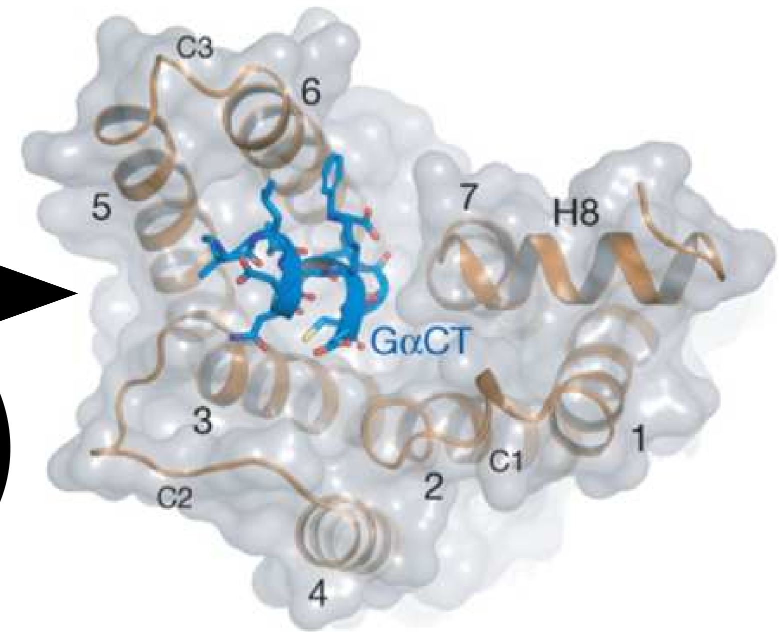
Rhodopsin Shape Change

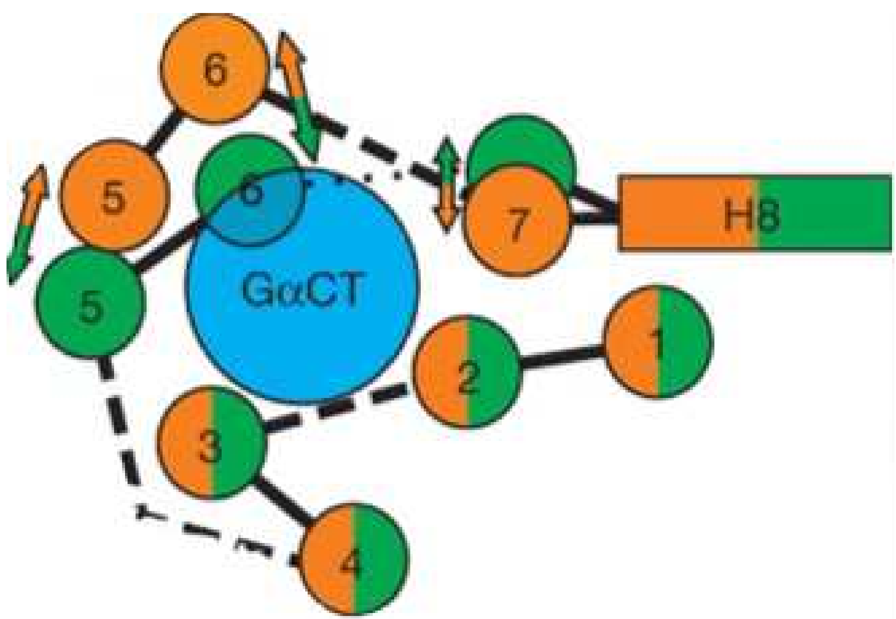
Dark State

After Photon - Light State







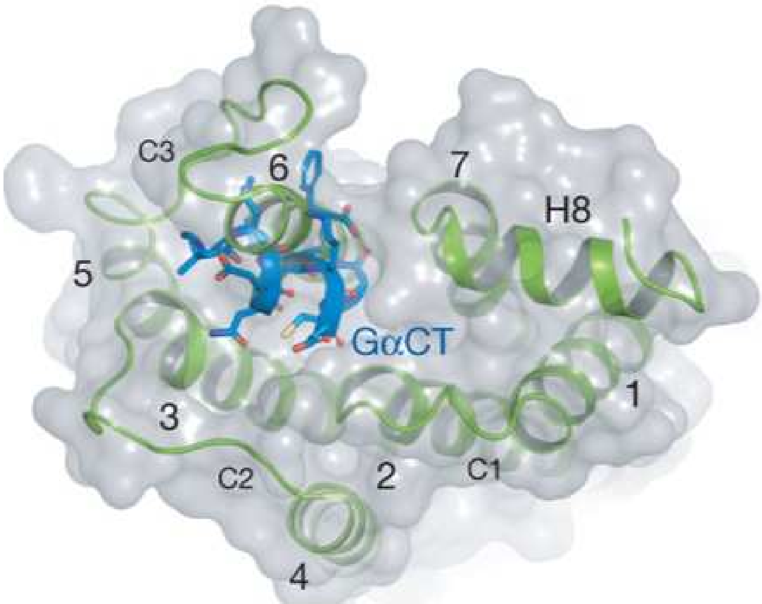


P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Rhodopsin Shape Change

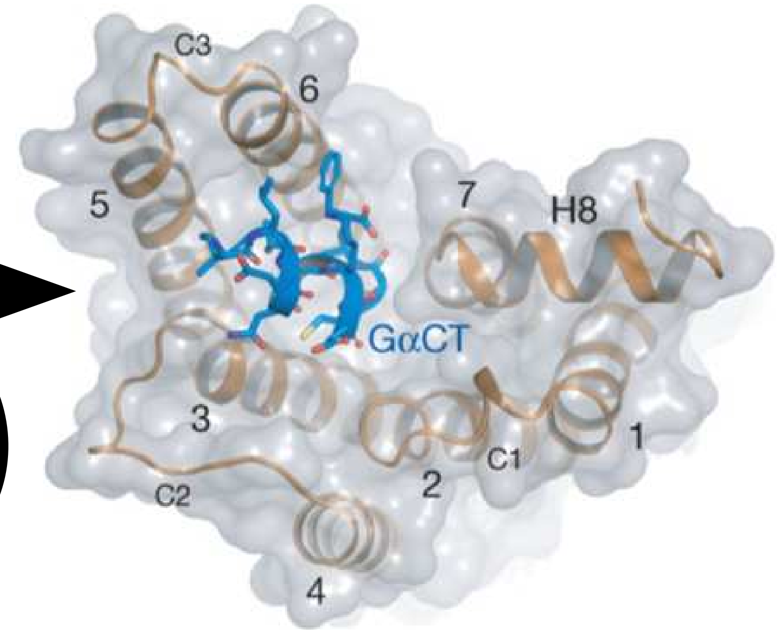
Dark State

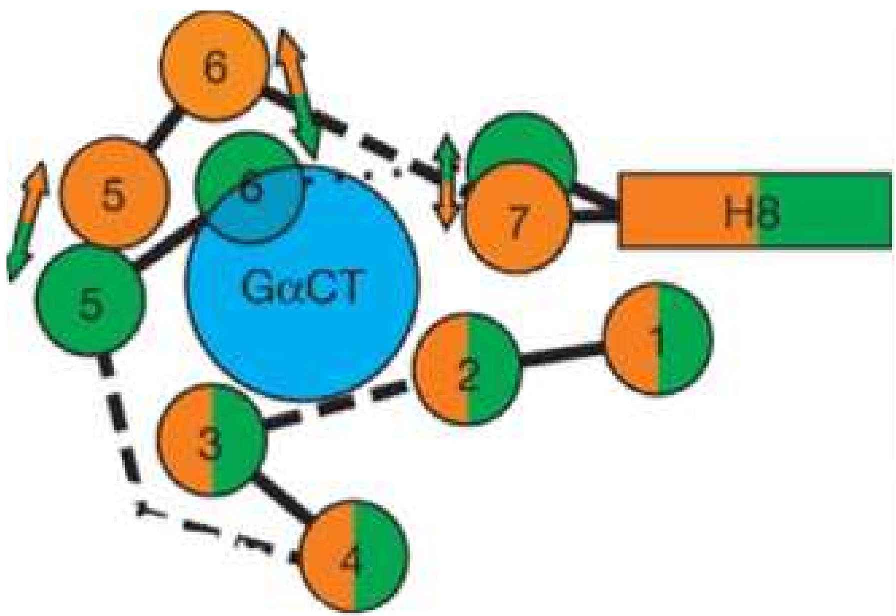
After Photon - Light State



hν

70%



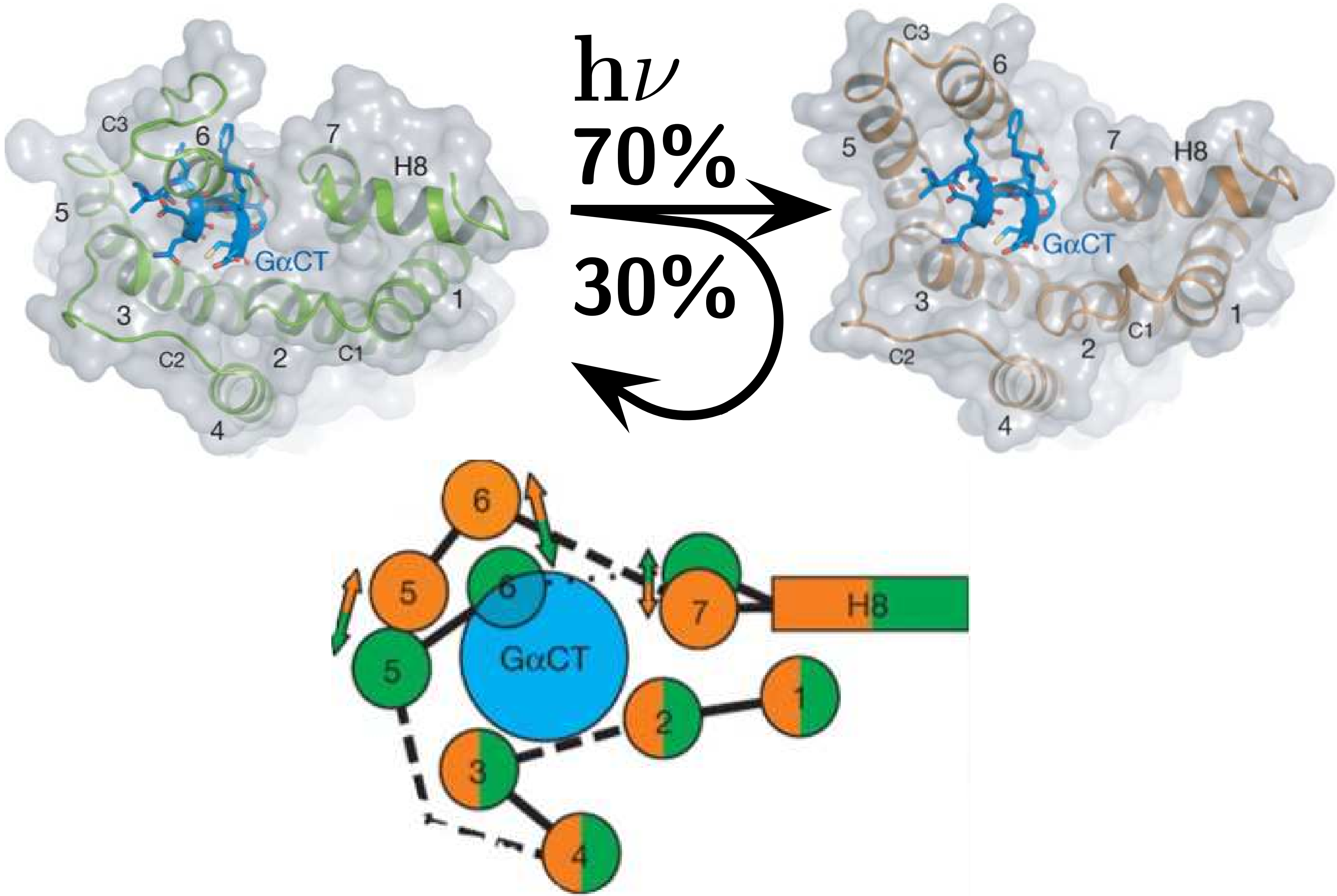


P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Rhodopsin Shape Change

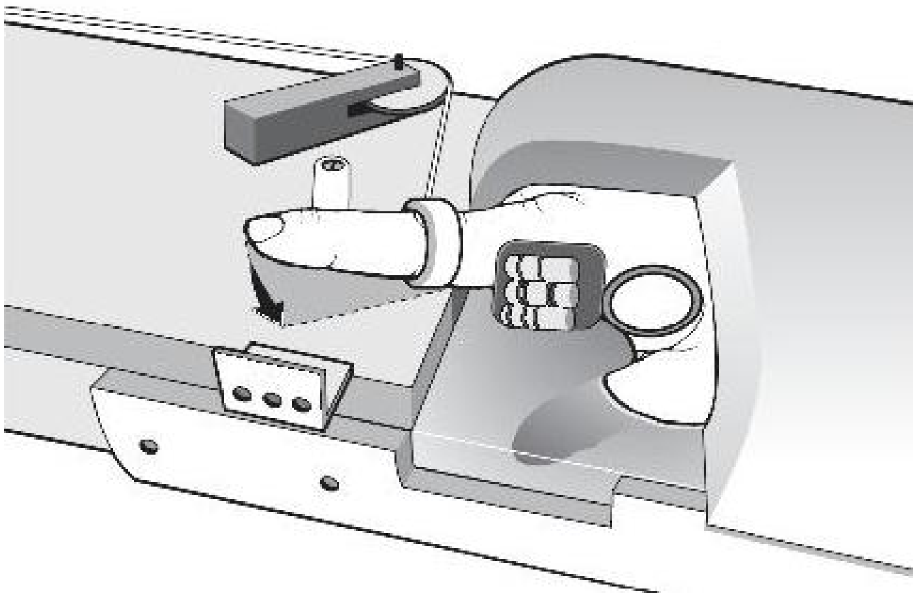
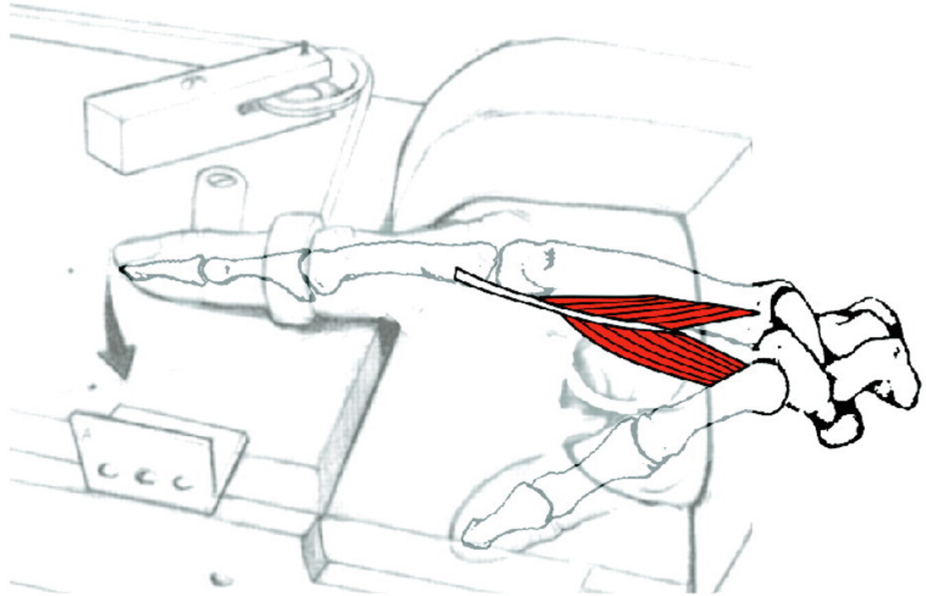
Dark State

After Photon - Light State



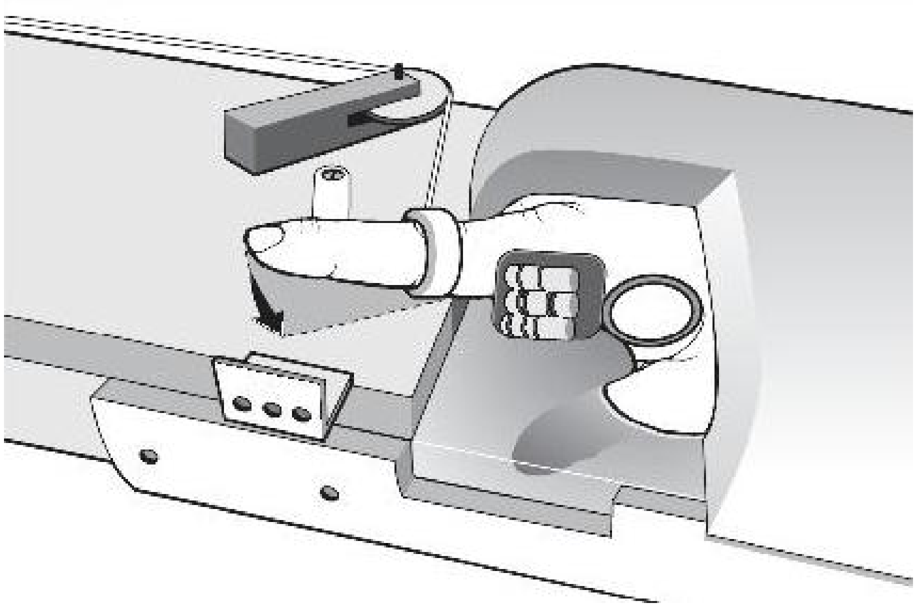
P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

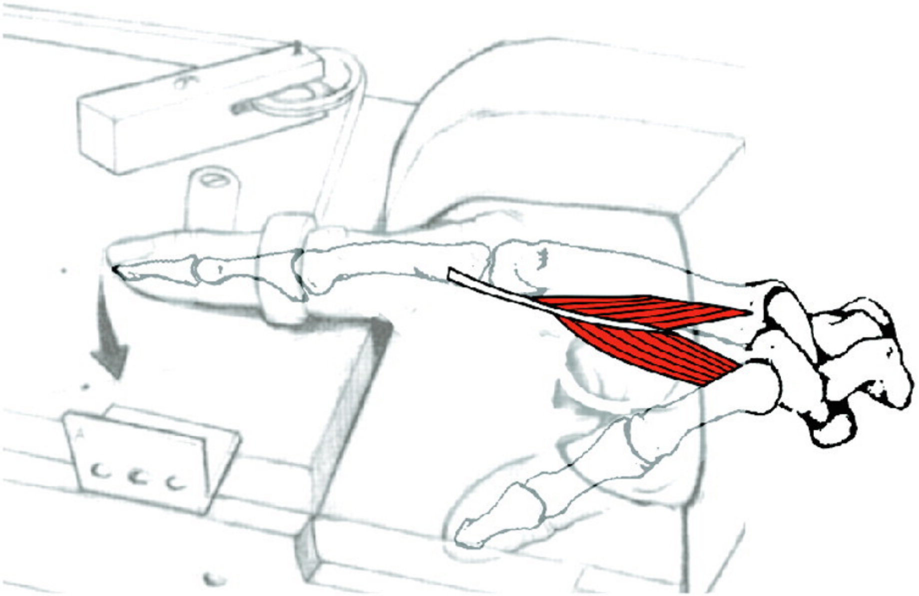
Efficiency of Muscle

- Experiments by Kushmerick's lab since (at least) 1969

Efficiency of Muscle

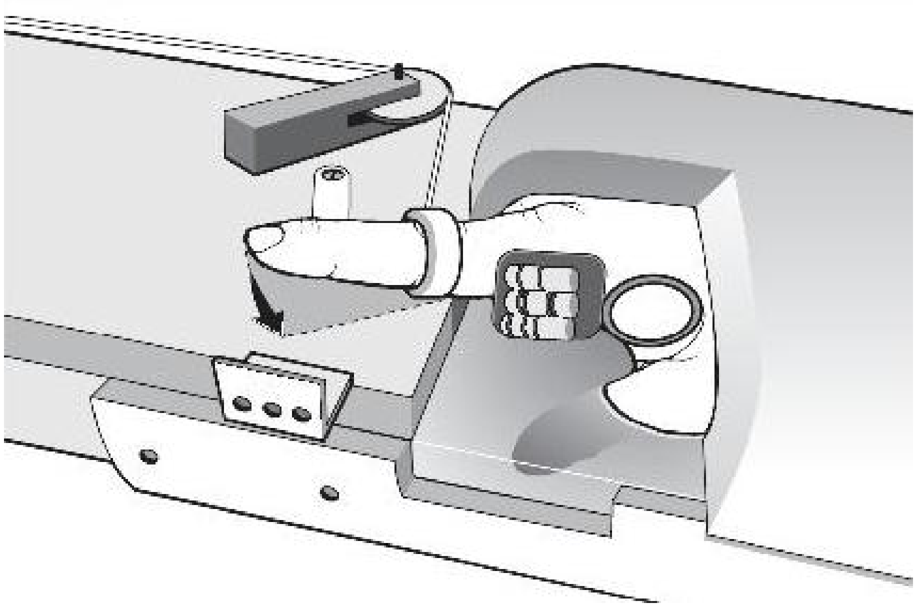


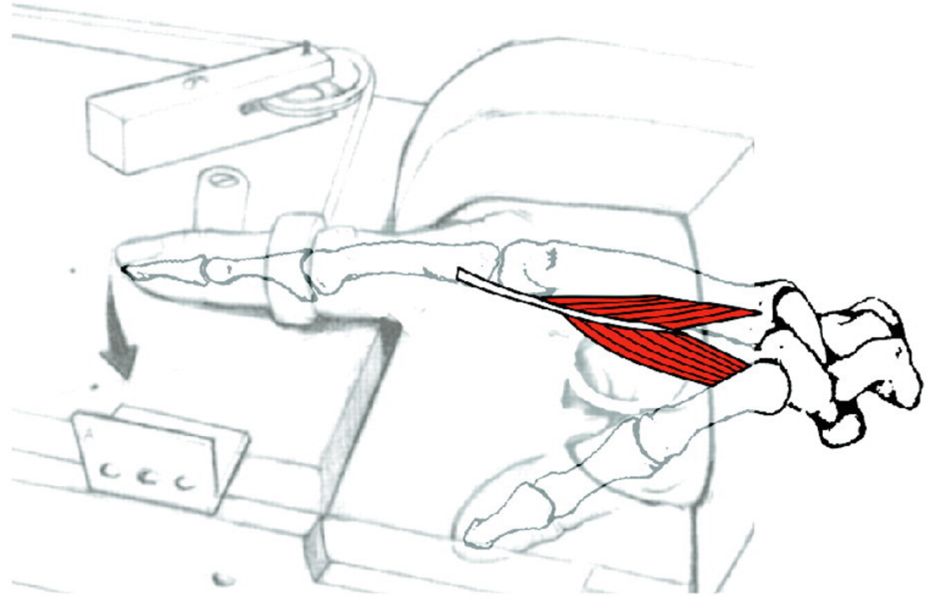


- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011

http://dx.doi.org/10.1113/jphysiol.2007.146829 http://dx.doi.org/10.1242/jeb.052985

Efficiency of Muscle

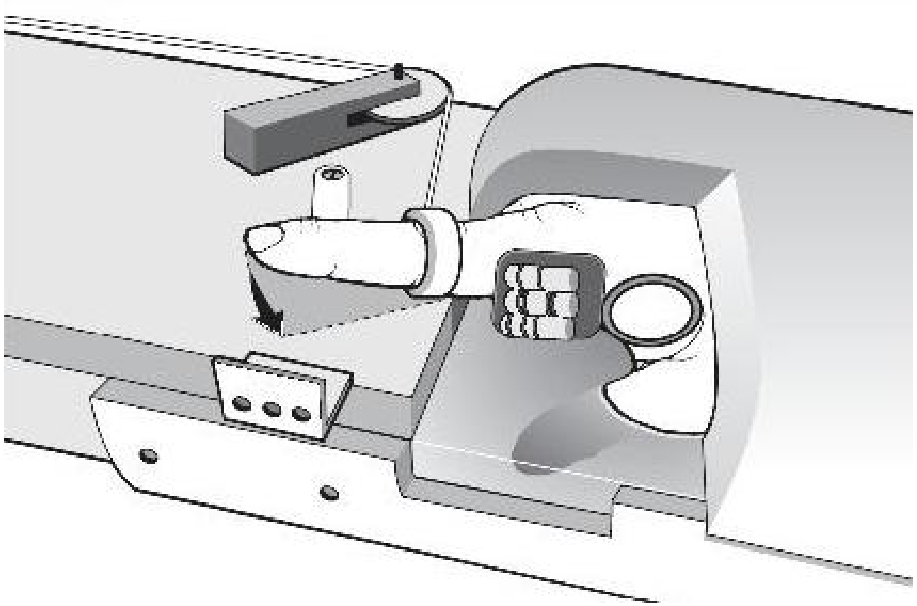


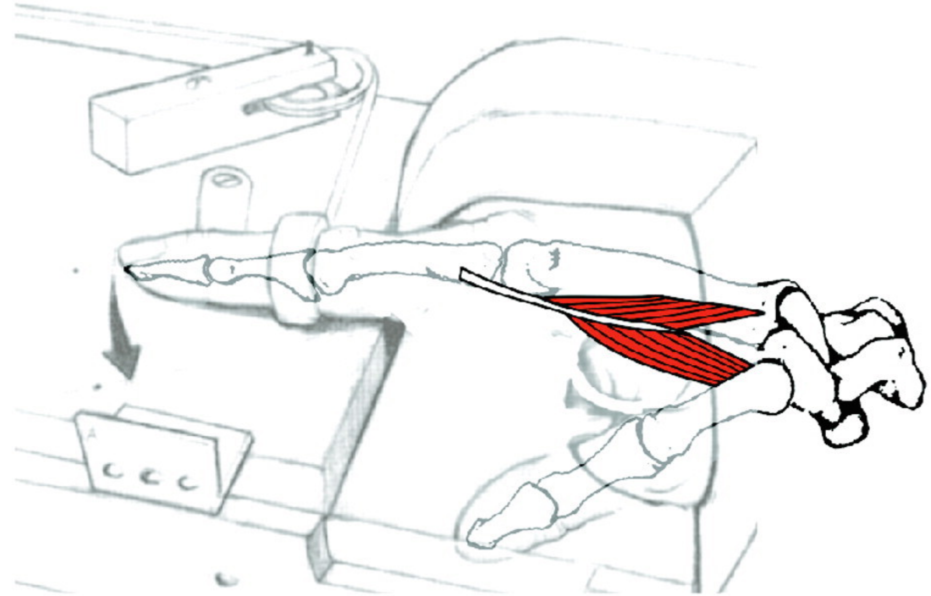


- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011
- Weight lifting gives work done

http://dx.doi.org/10.1113/jphysiol.2007.146829 http://dx.doi.org/10.1242/jeb.052985

Efficiency of Muscle

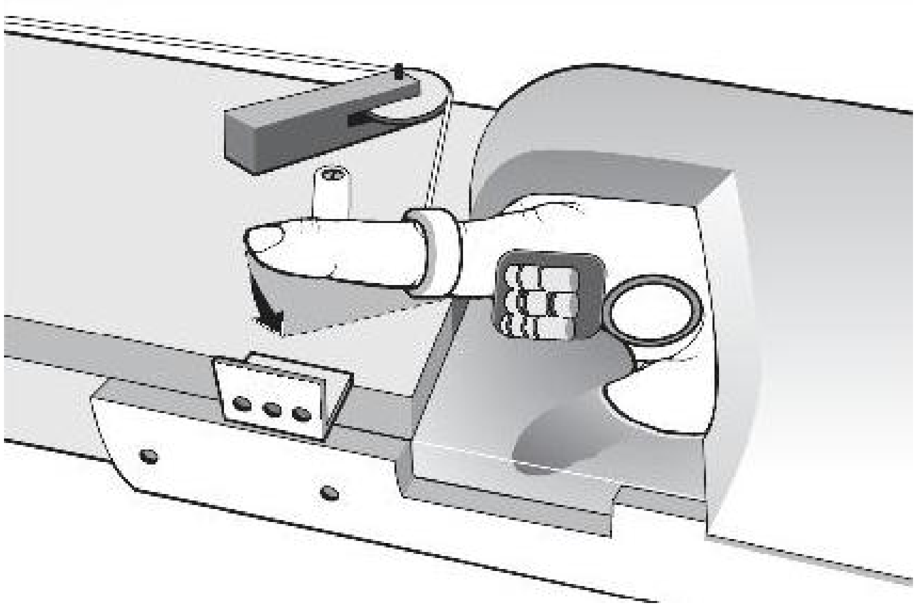


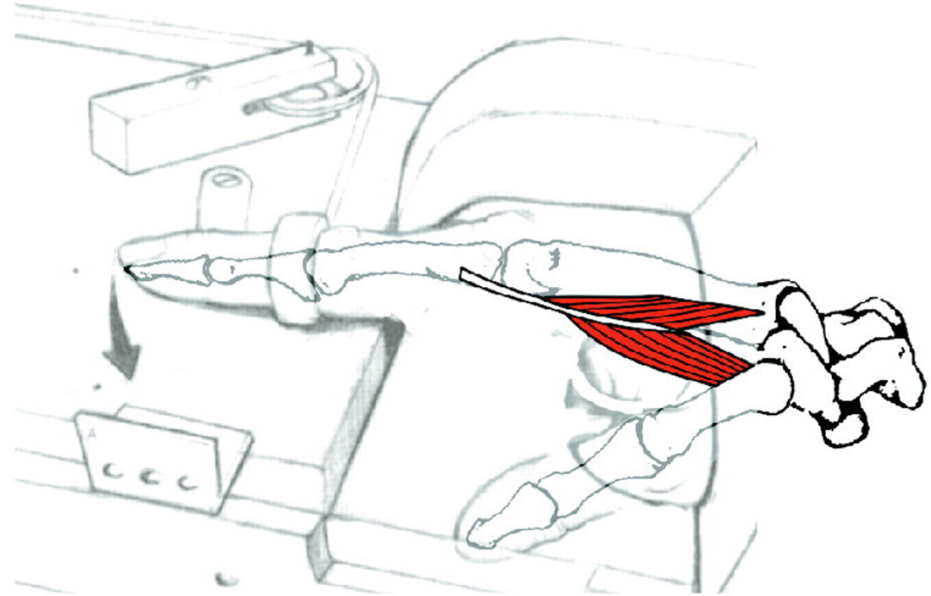


- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011
- Weight lifting gives work done
- NMR coil gives ATP = energy used

http://dx.doi.org/10.1113/jphysiol.2007.146829 http://dx.doi.org/10.1242/jeb.052985

Efficiency of Muscle

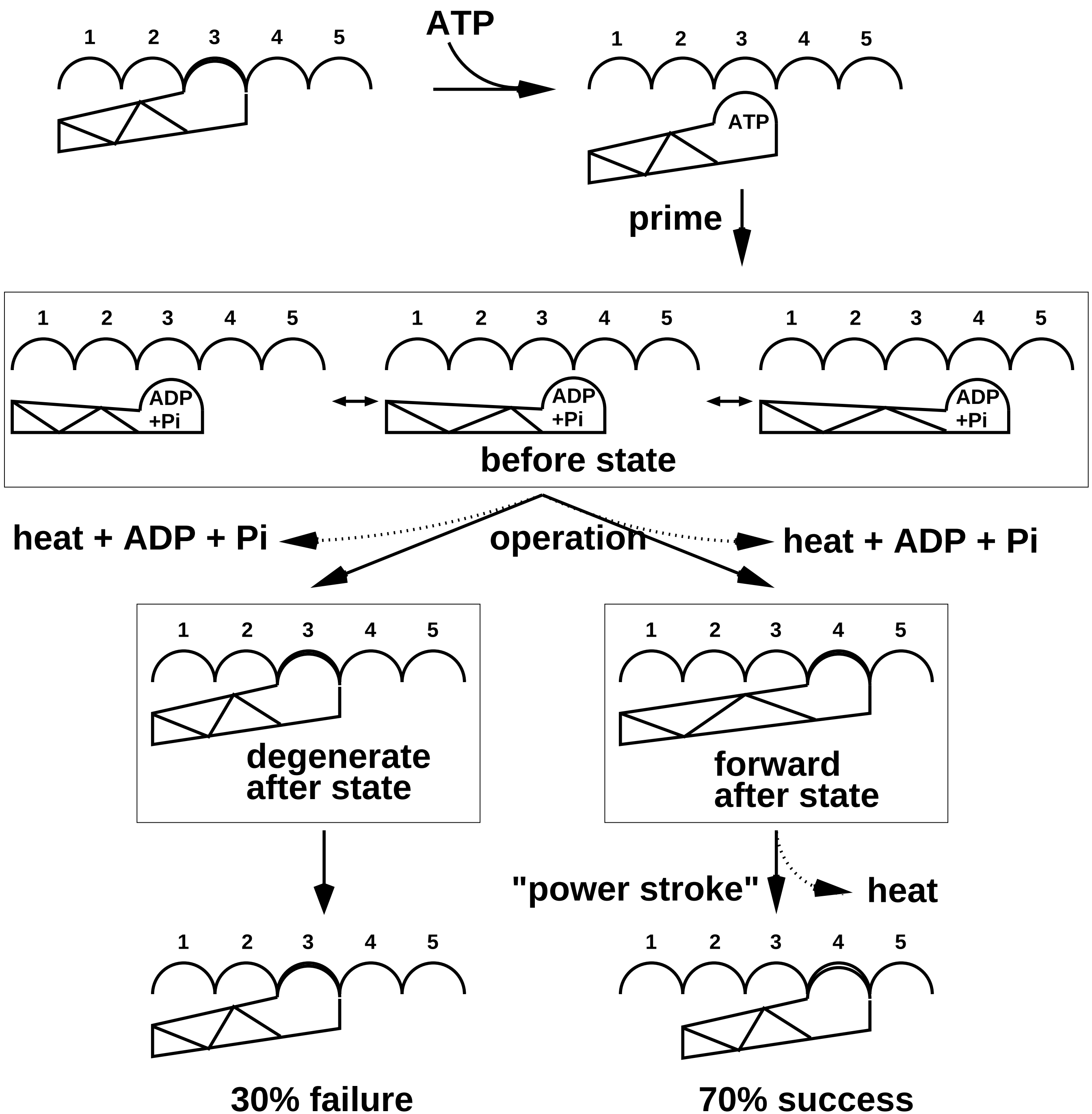




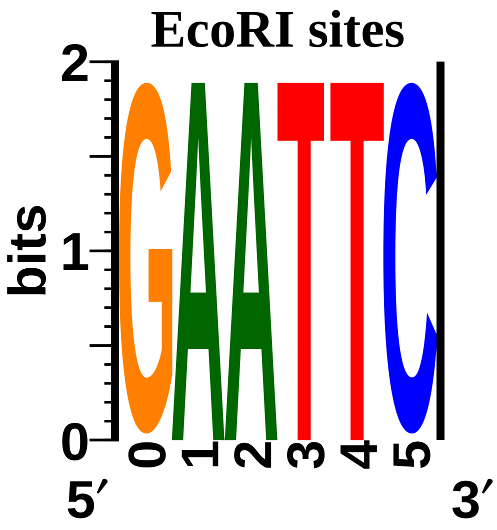
- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011
- Weight lifting gives work done
- NMR coil gives ATP = energy used
- Efficiency: 0.68 ± 0.09

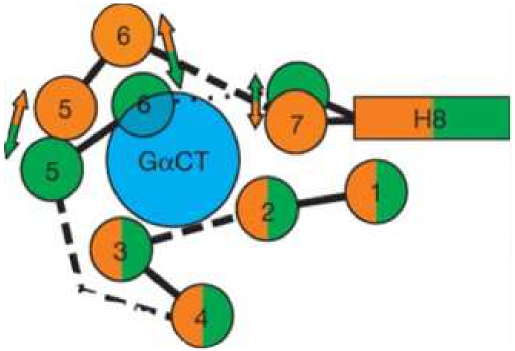
http://dx.doi.org/10.1113/jphysiol.2007.146829 http://dx.doi.org/10.1242/jeb.052985

Tom's Model of Muscle Mechanism



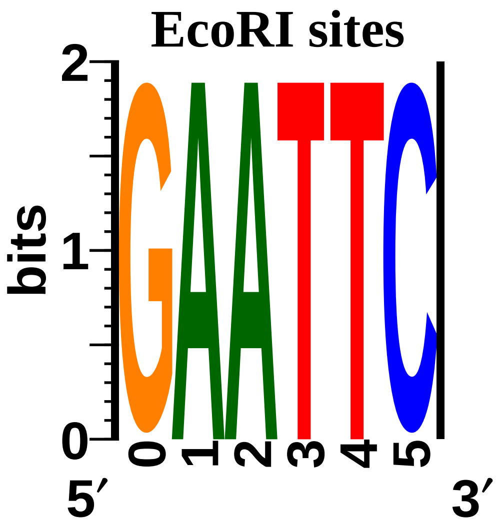
70% efficiency appears widely in biology:

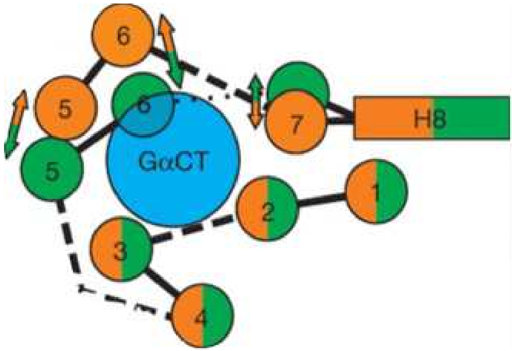




70% efficiency appears widely in biology:

- DNA - protein binding

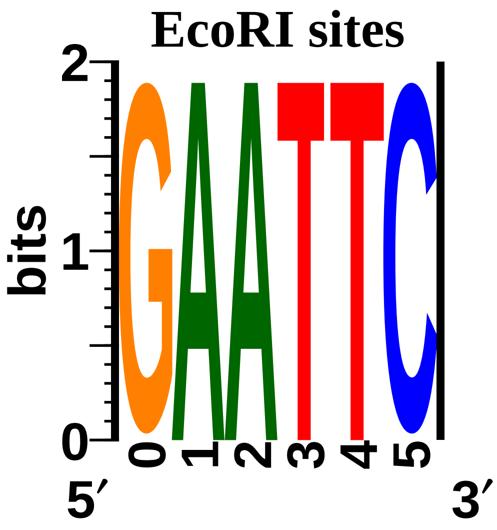


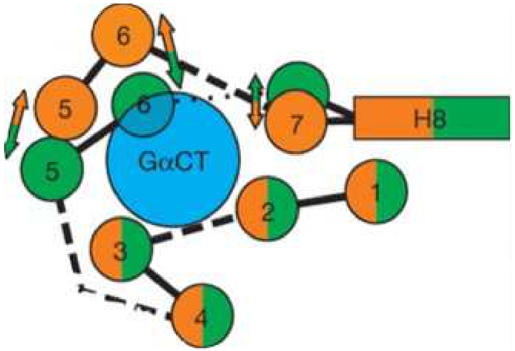


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

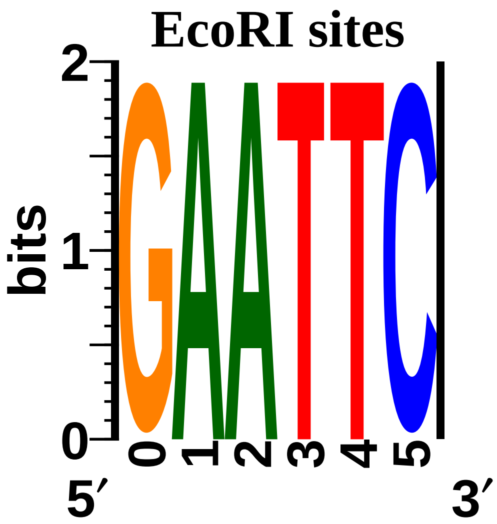
- DNA - protein binding
- rhodopsin

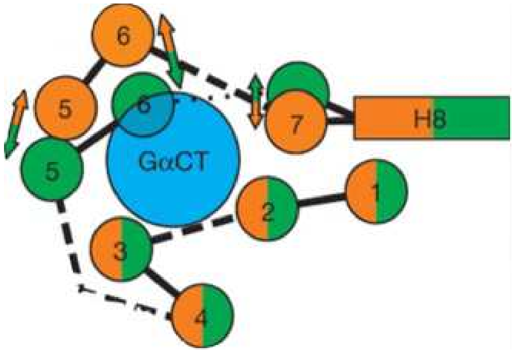




70% efficiency appears widely in biology:

- DNA - protein binding
- rhodopsin
- muscle

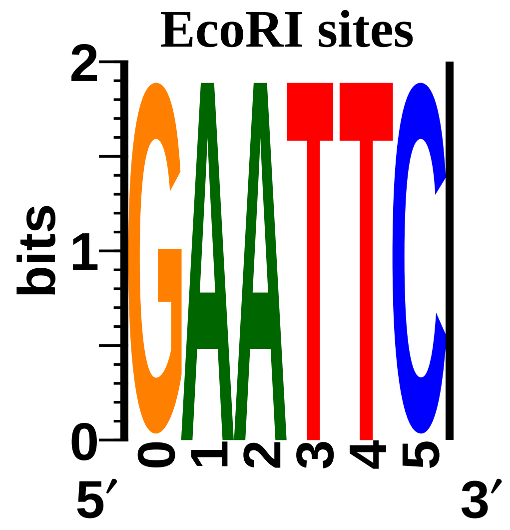


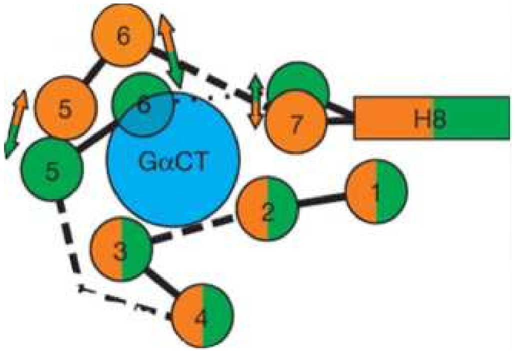


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

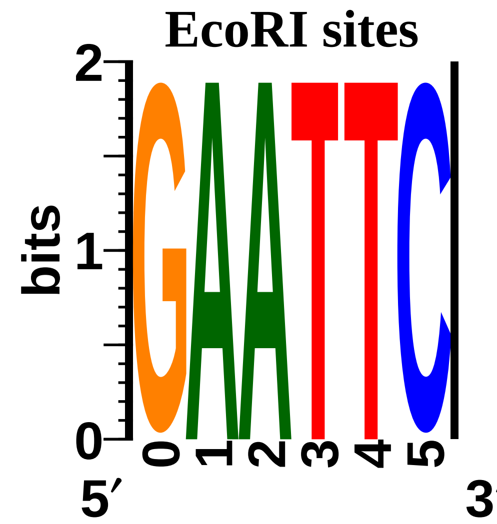
- DNA - protein binding
- rhodopsin
- muscle
- other systems



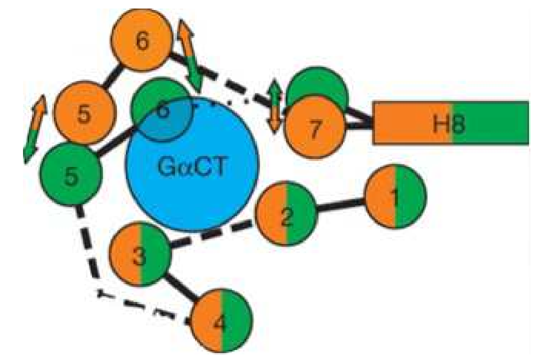


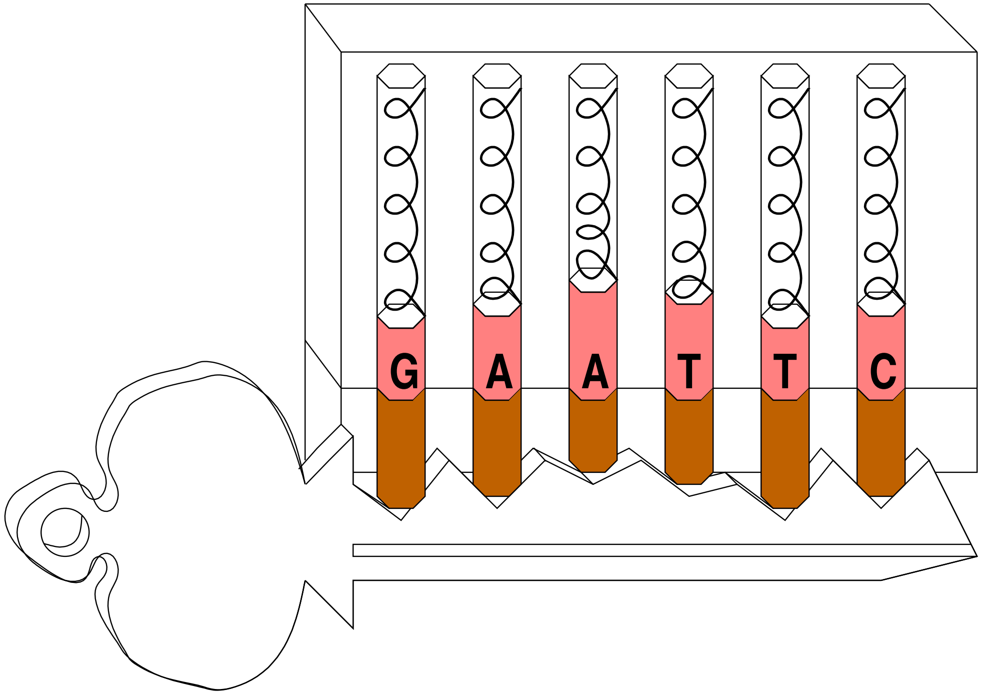
70% efficiency appears widely in biology:

- DNA - protein binding
- rhodopsin
- muscle
- other systems



Why 70% efficiency?

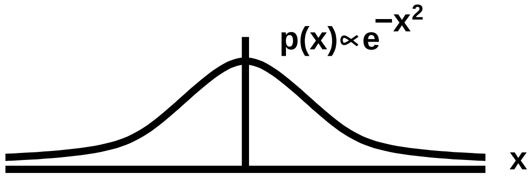




Like a key in a lock
which has many independent pins,
it takes many numbers
to describe the vibrational state
of a molecular machine

- Pin motion x has a Gaussian distribution:

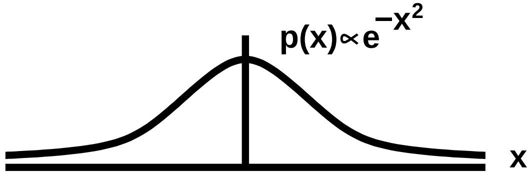
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ = mean, σ = standard deviation

- Pin motion x has a Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



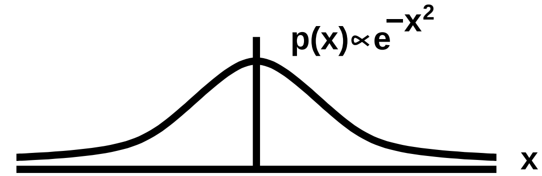
μ = mean, σ = standard deviation

- Gaussian distributions are generated by the sum of many small random variables

Gaussians

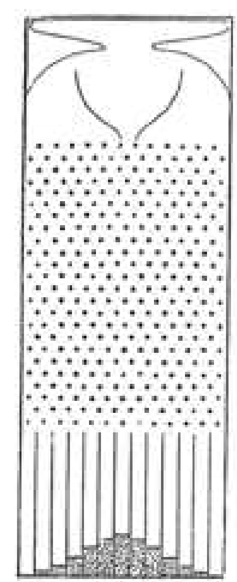
- Pin motion x has a Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



  μ = mean, σ = standard deviation

- Gaussian distributions are generated by the sum of many small random variables
- Drunkard's walk: Galton's quincunx device!



http://www.youtube.com/watch?v=xDIyAOBa_yU

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{2}$$

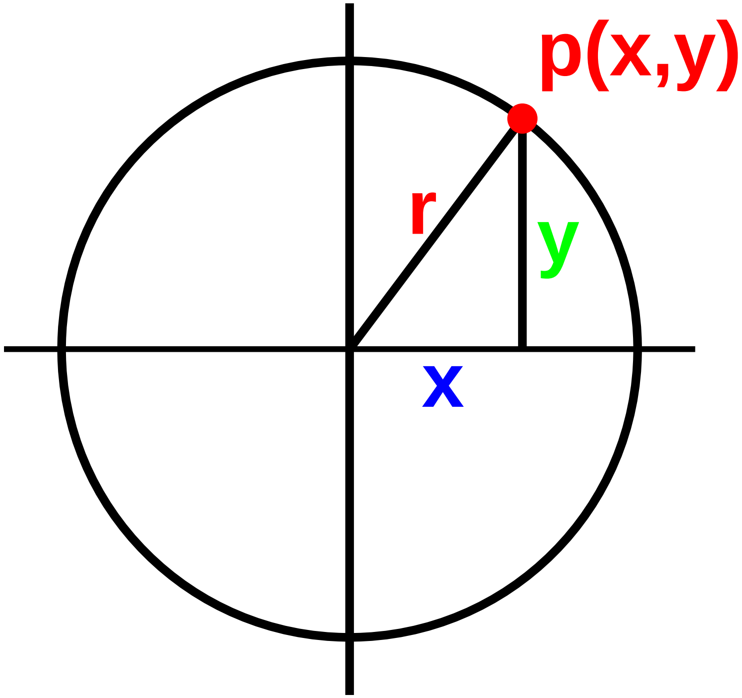
Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$

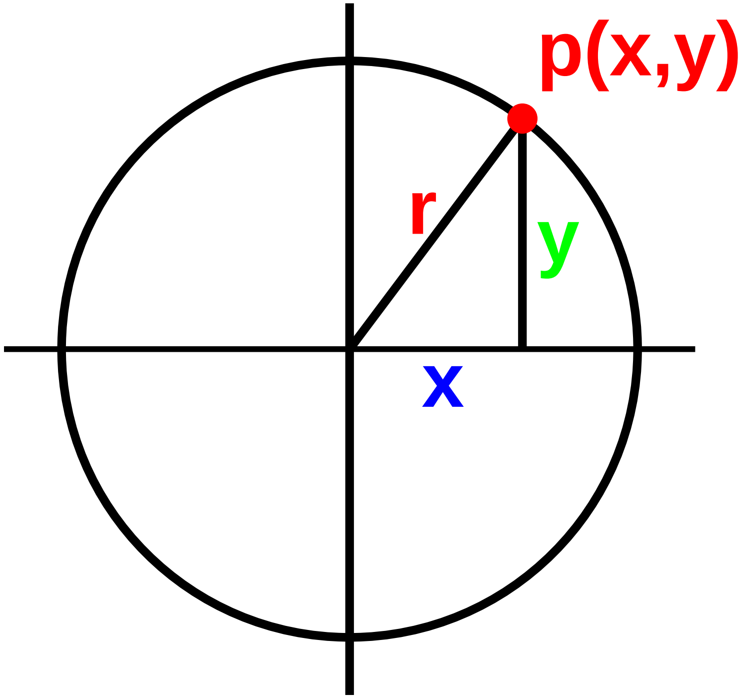
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

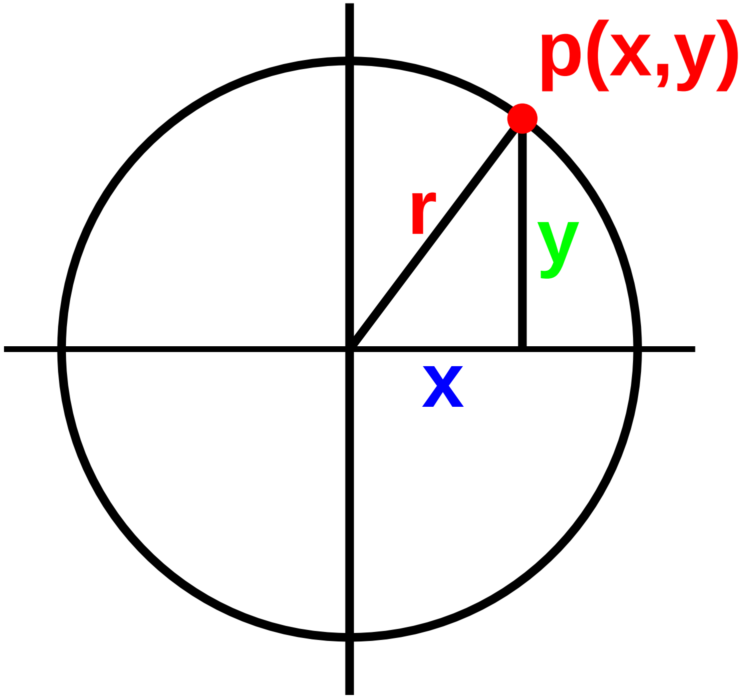
$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$



$$p(x,y) = p(x) \times p(y) \quad (3)$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$

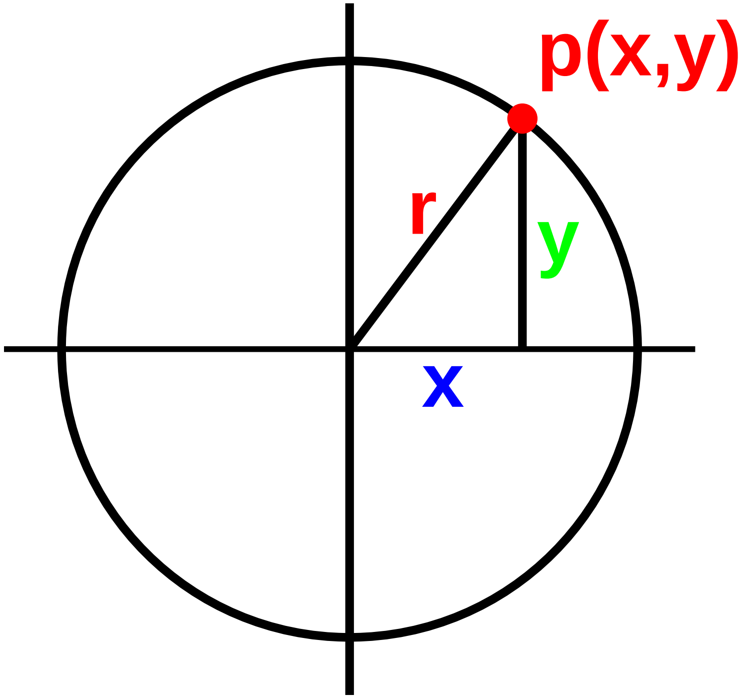


$$p(x,y) = p(x) \times p(y) \quad (3)$$

$$\propto e^{-x^2} \times e^{-y^2} \quad (4)$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$



$$p(x,y) = p(x) \times p(y) \quad (3)$$

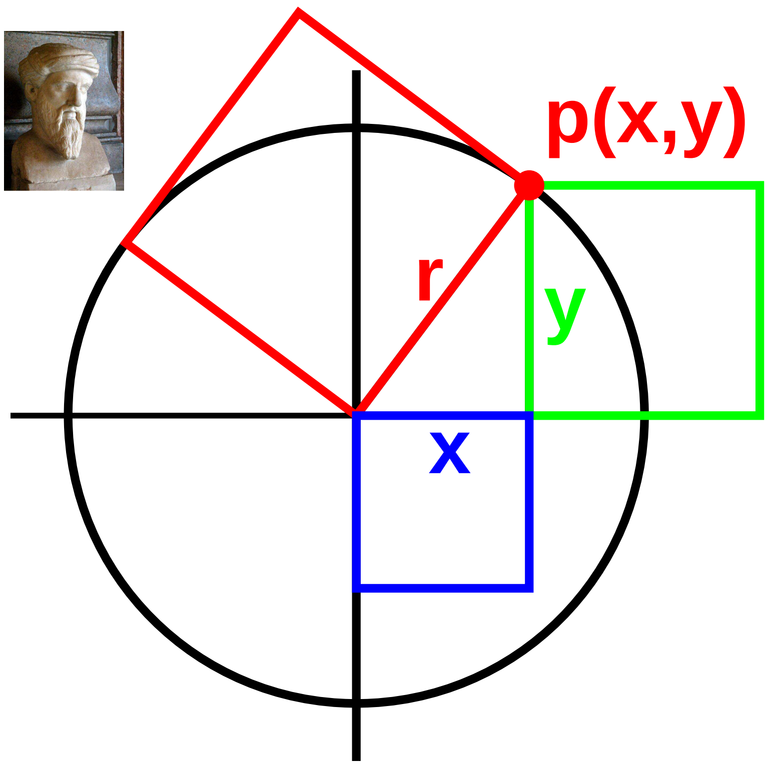
$$\propto e^{-x^2} \times e^{-y^2} \quad (4)$$

$$\propto e^{-(x^2+y^2)} \quad (5)$$

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$



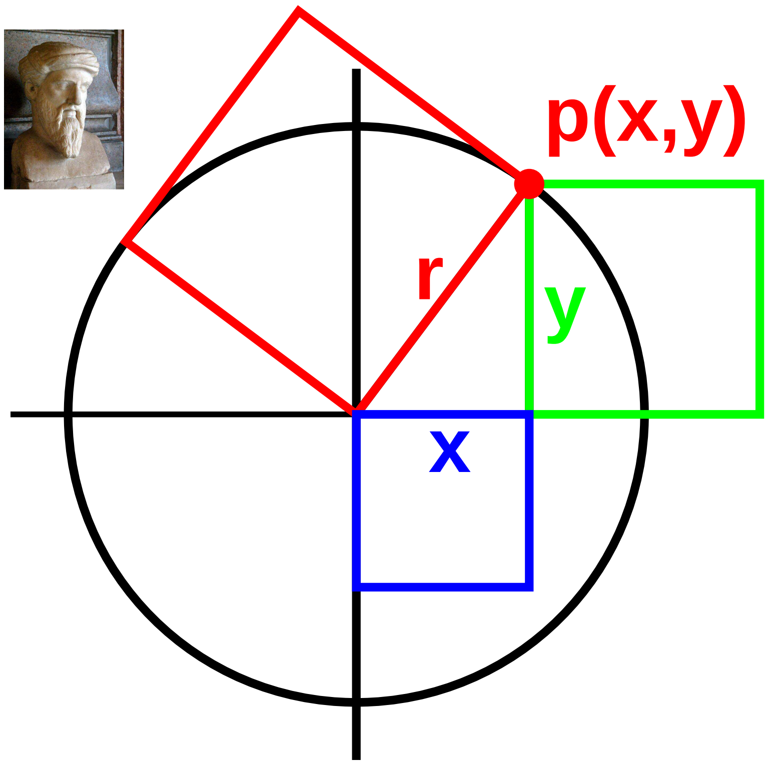
$$\begin{aligned} p(x,y) &= p(x) \times p(y) && (3) \\ &\propto e^{-x^2} \times e^{-y^2} && (4) \\ &\propto e^{-(x^2+y^2)} && (5) \\ &\propto e^{-r^2} && (6) \end{aligned}$$

credit: http://en.wikipedia.org/wiki/Pythagoras

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \qquad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \qquad (2)$$



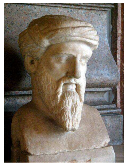
$$\begin{aligned} p(x,y) &= p(x) \times p(y) && (3) \\ &\propto e^{-x^2} \times e^{-y^2} && (4) \\ &\propto e^{-(x^2+y^2)} && (5) \\ &\propto e^{-r^2} && (6) \end{aligned}$$

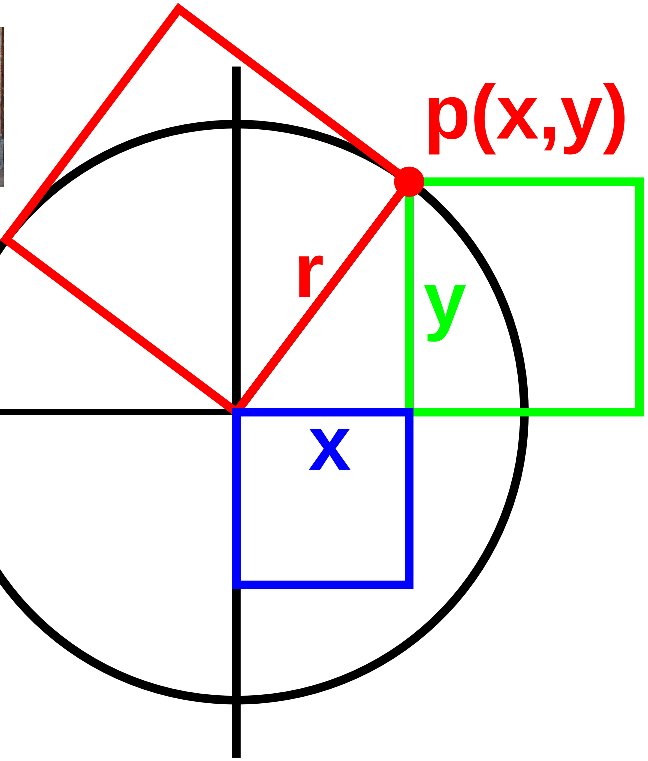
If $p(x,y)$ is a constant, then r is a constant.

credit: http://en.wikipedia.org/wiki/Pythagoras

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$

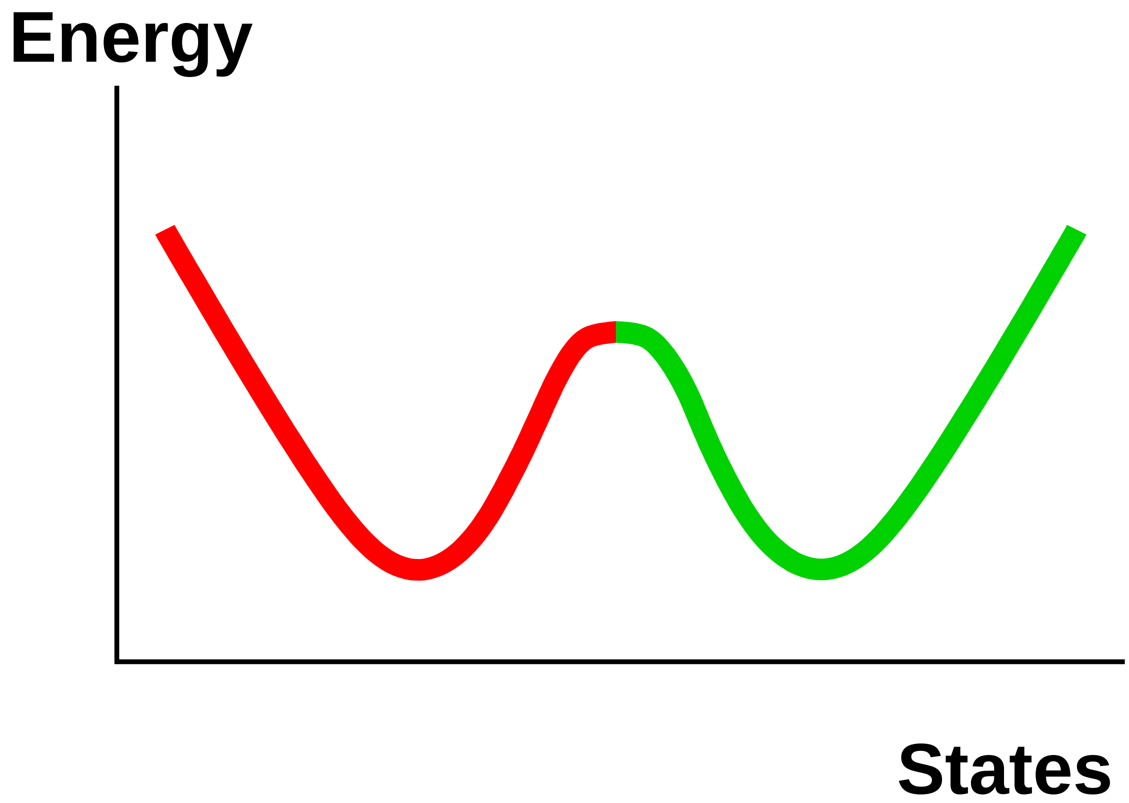




$$\begin{aligned} p(x,y) &= p(x) \times p(y) && (3) \\ &\propto e^{-x^2} \times e^{-y^2} && (4) \\ &\propto e^{-(x^2+y^2)} && (5) \\ &\propto e^{-r^2} && (6) \end{aligned}$$

If $p(x,y)$ is a constant, then r is a constant.

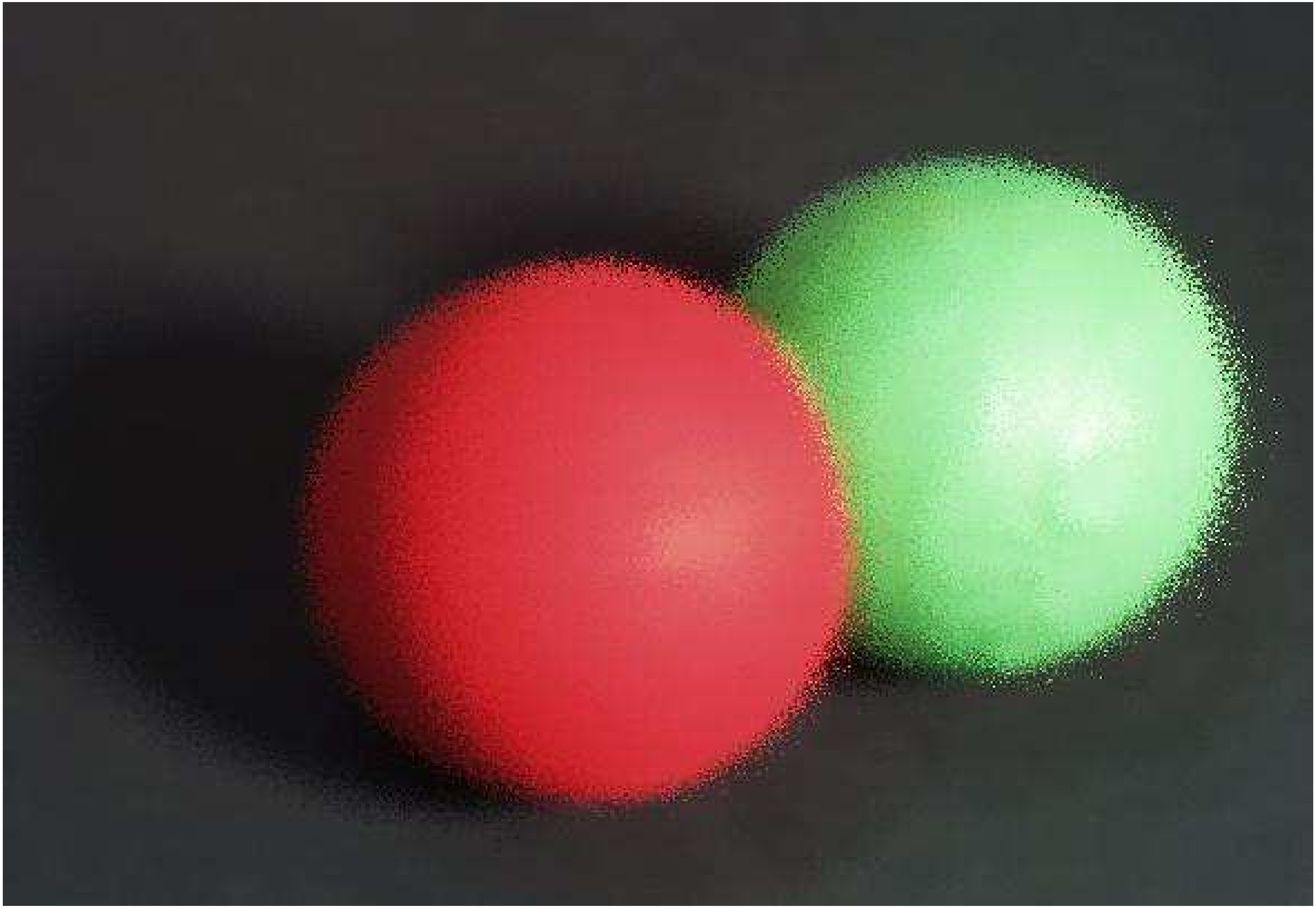
Circular distribution!



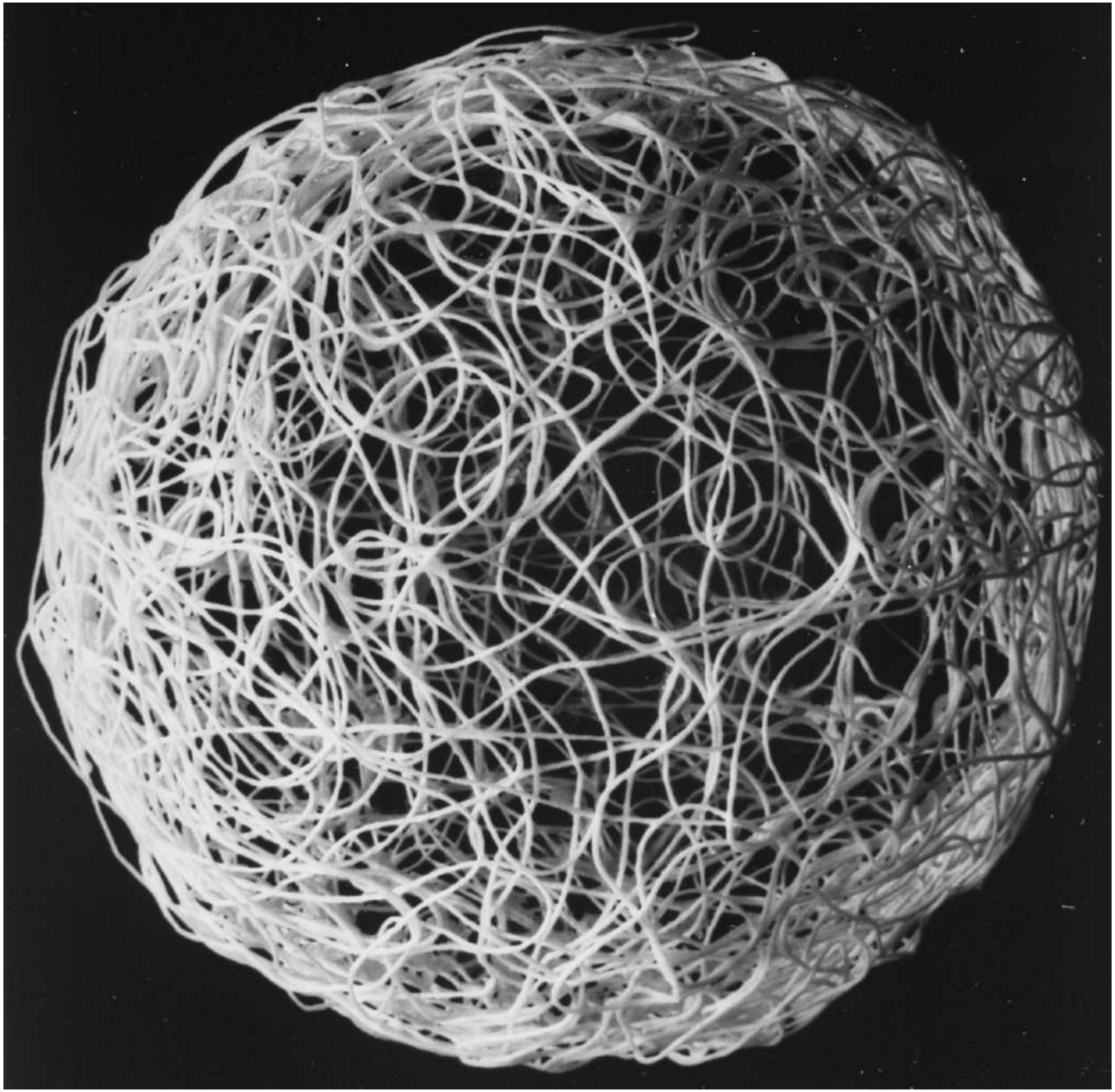
1 dimension is too simple!

Bowls in 2 Dimensions

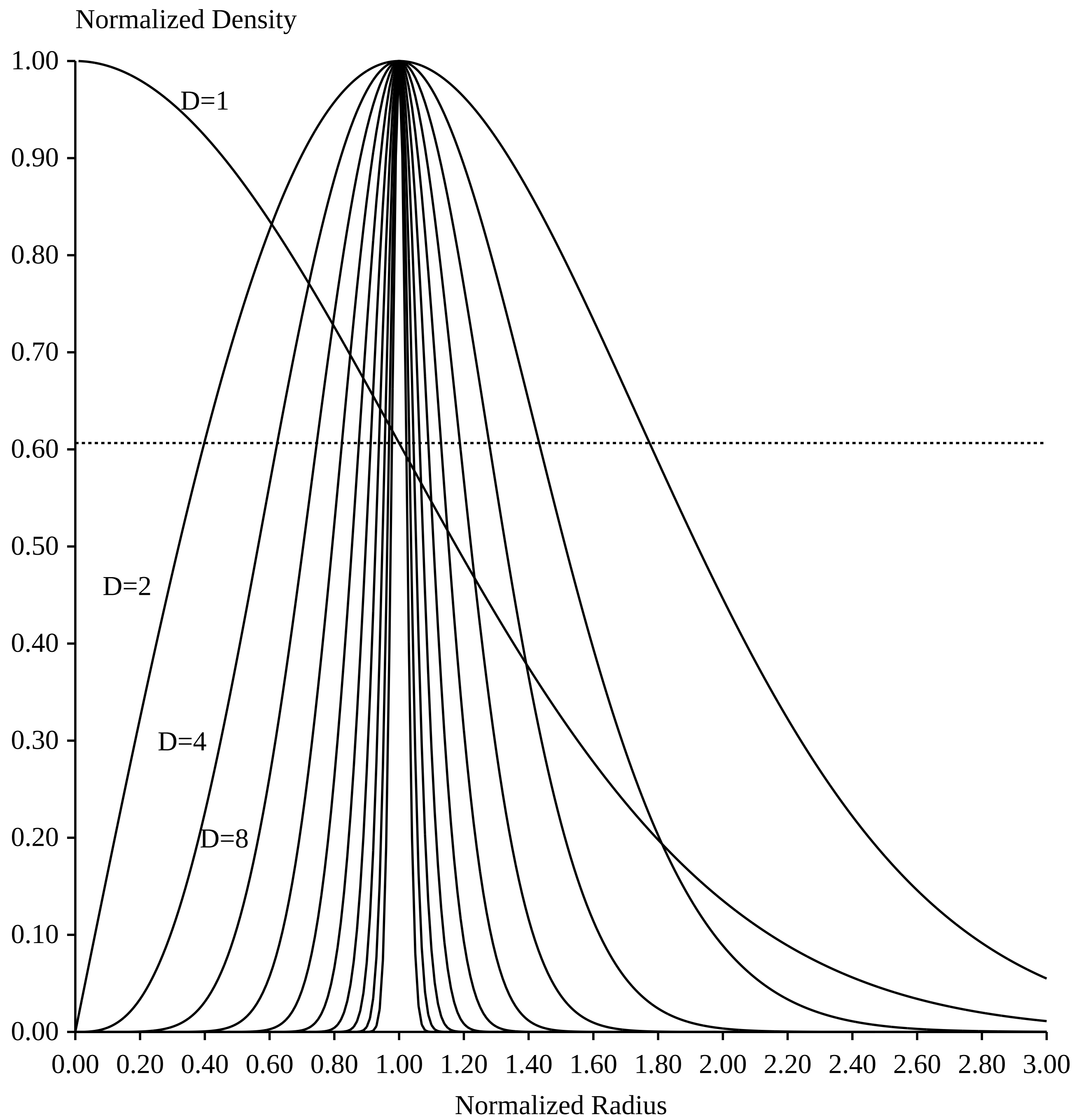
Spheres in 3 Dimensions



N Dimensional Sphere



Spheres tighten in high dimensions



Good Sphere Packing

- Good packing of spheres gives a molecule the capacity to make selections efficiently

Good Sphere Packing

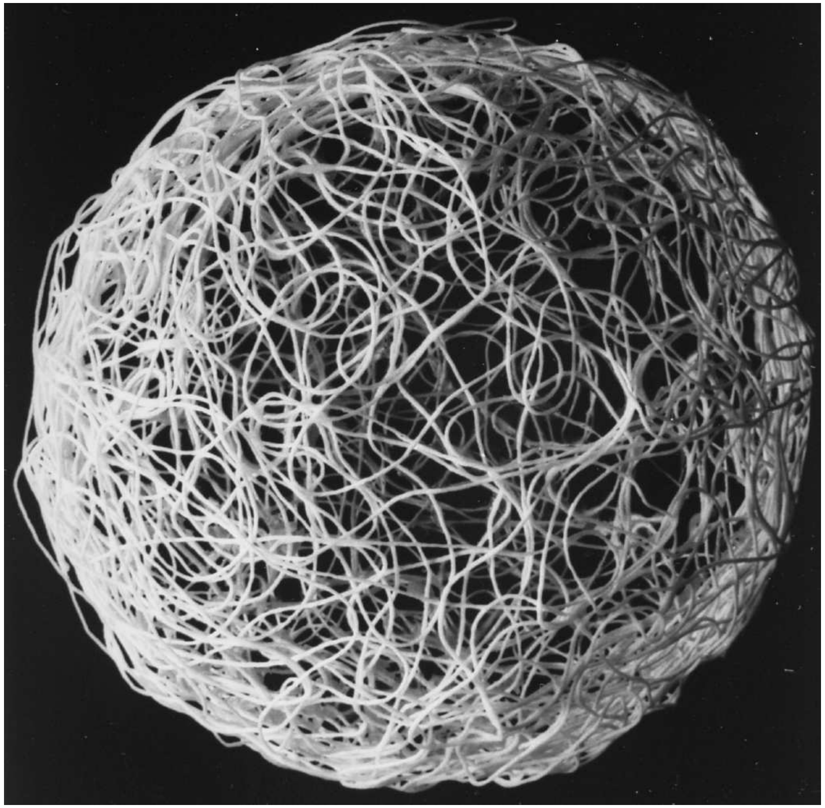
- Good packing of spheres gives a molecule the capacity to make selections efficiently
- Shannon's 1949 paper: each gumball is a message

Good Sphere Packing

- Good packing of spheres gives a molecule the capacity to make selections efficiently
- Shannon's 1949 paper: each gumball is a message
- For a molecule each gumball is a state

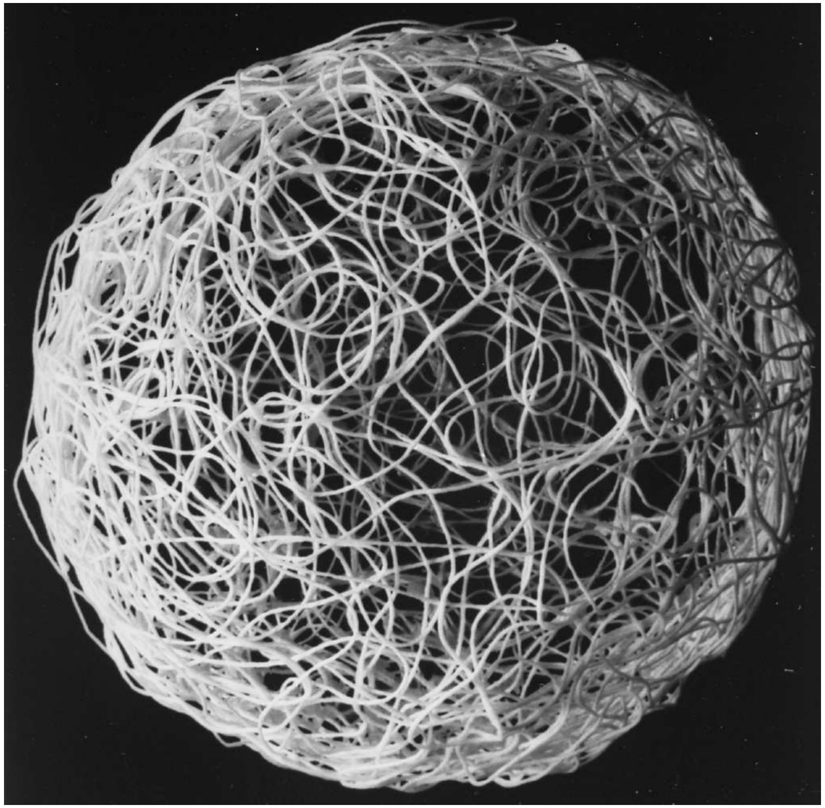
N Dimensional Sphere Separation

Degenerate Sphere

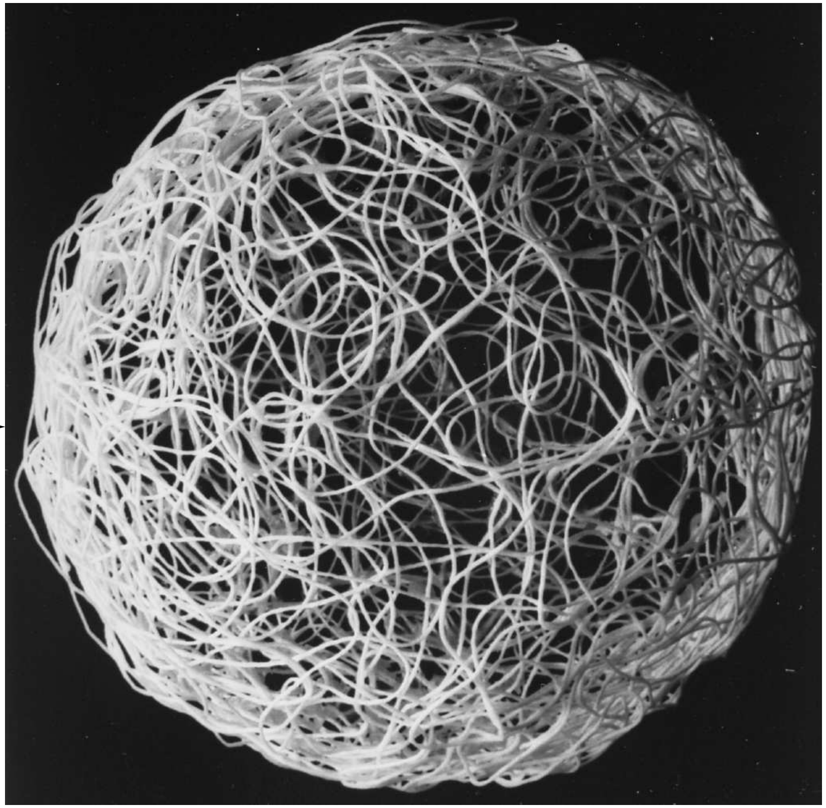


N Dimensional Sphere Separation

Degenerate Sphere



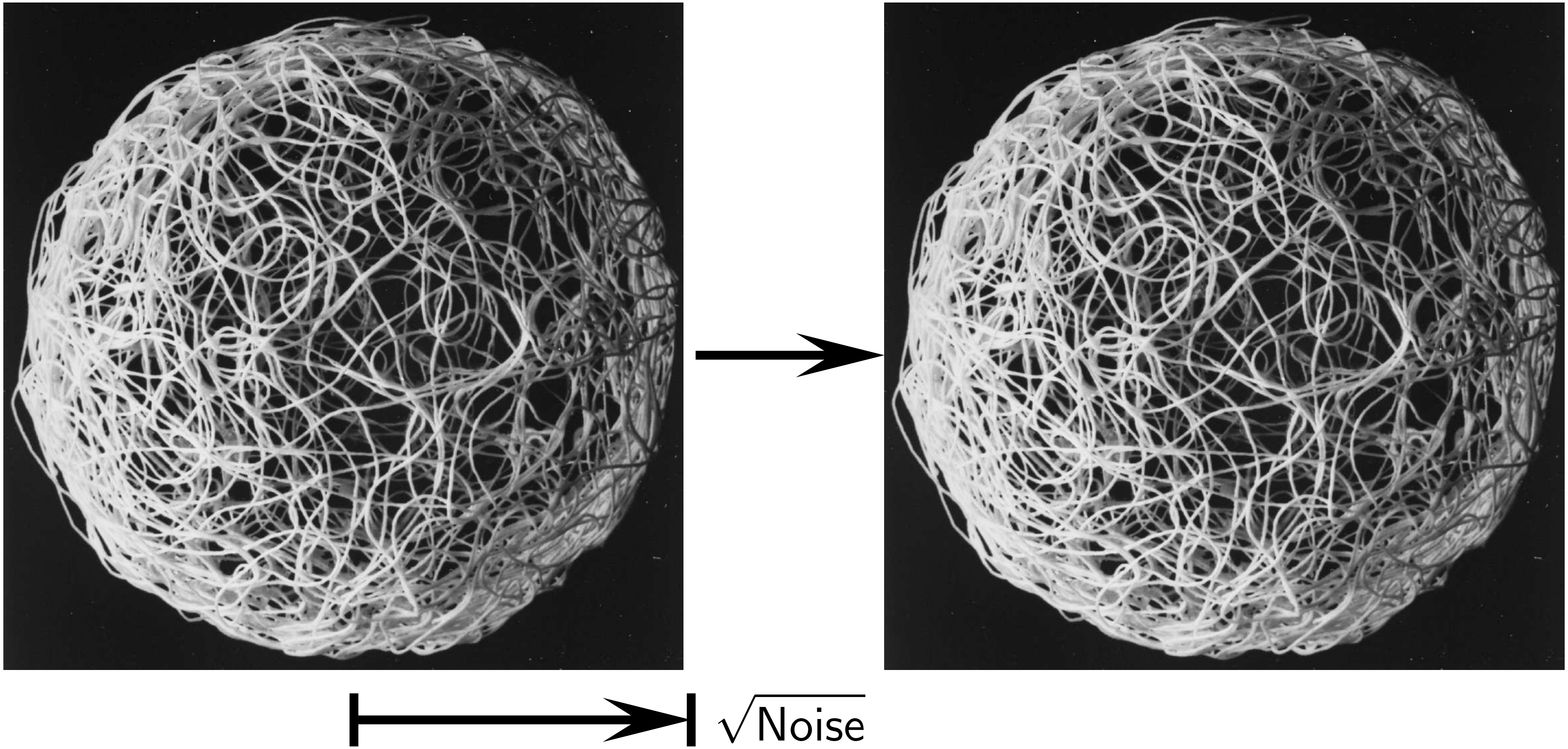
Forward Sphere



N Dimensional Sphere Separation

Degenerate Sphere

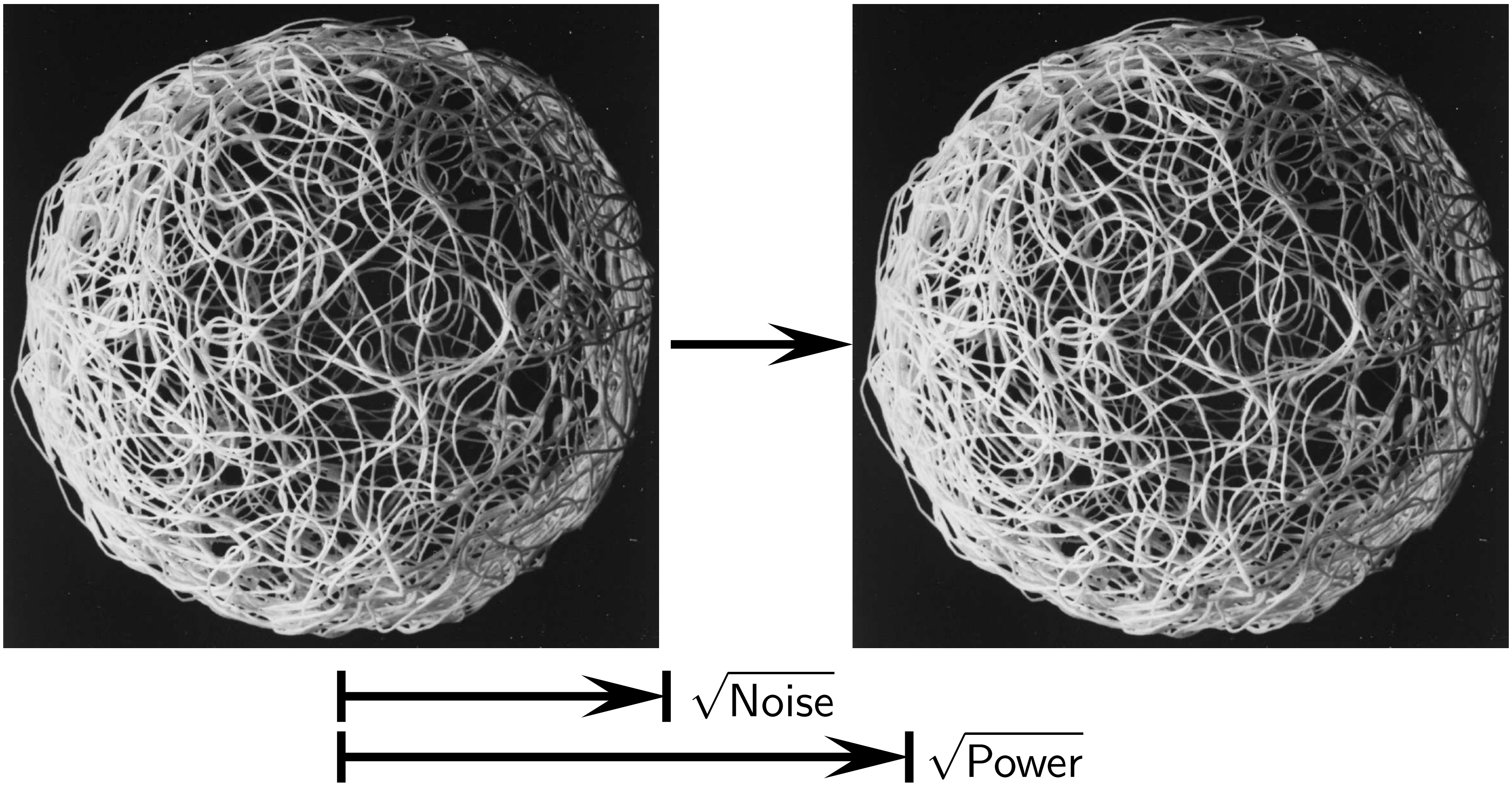
Forward Sphere



N Dimensional Sphere Separation

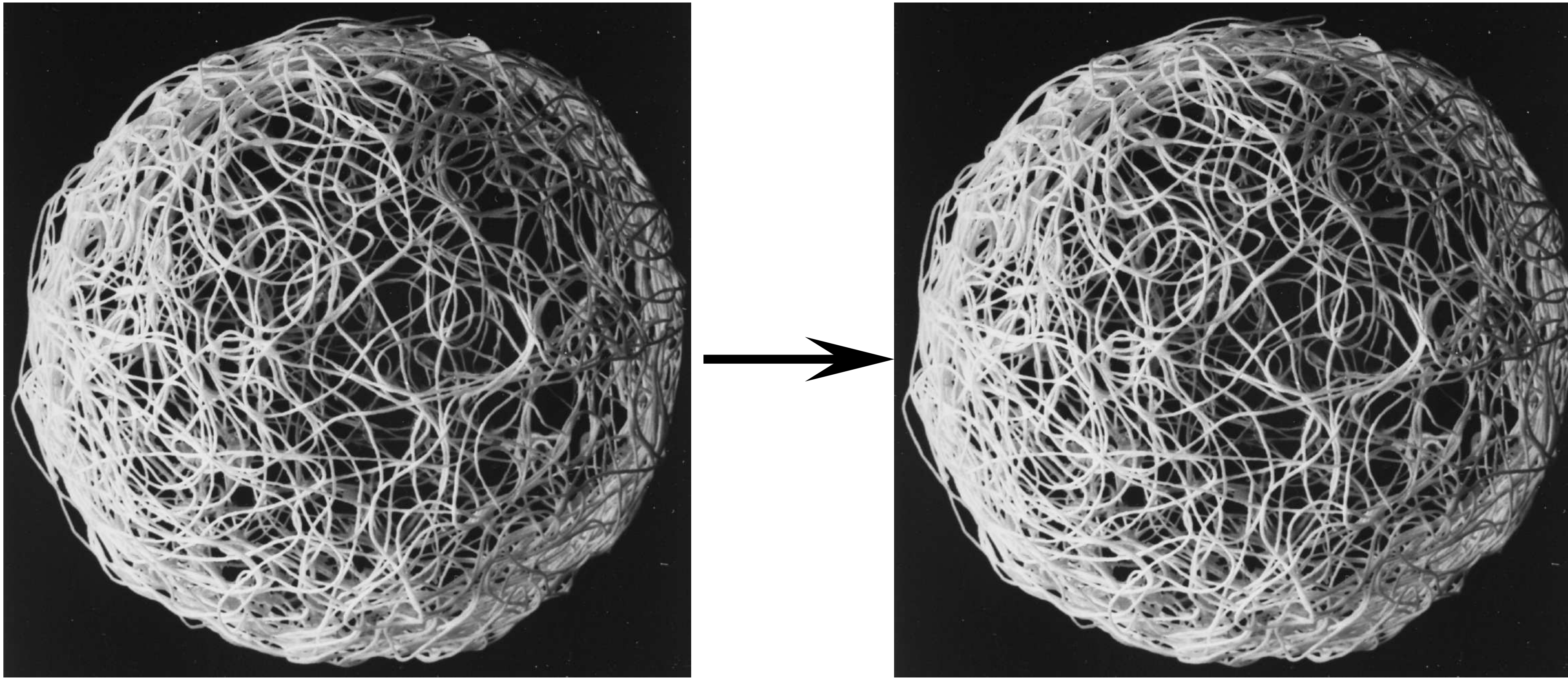
Degenerate Sphere

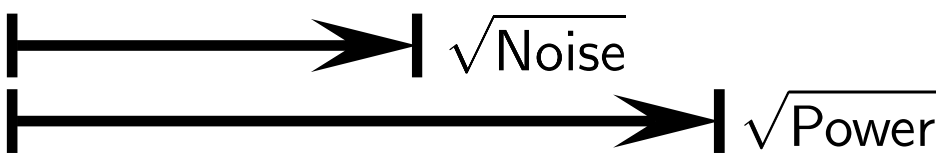
Forward Sphere



N Dimensional Sphere Separation

Degenerate Sphere   Forward Sphere

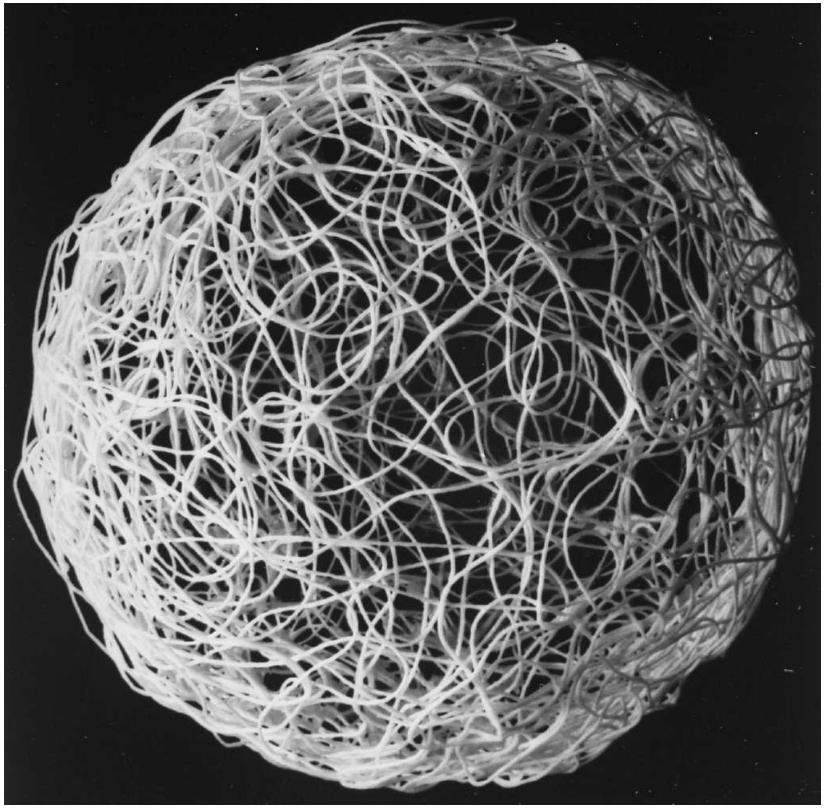




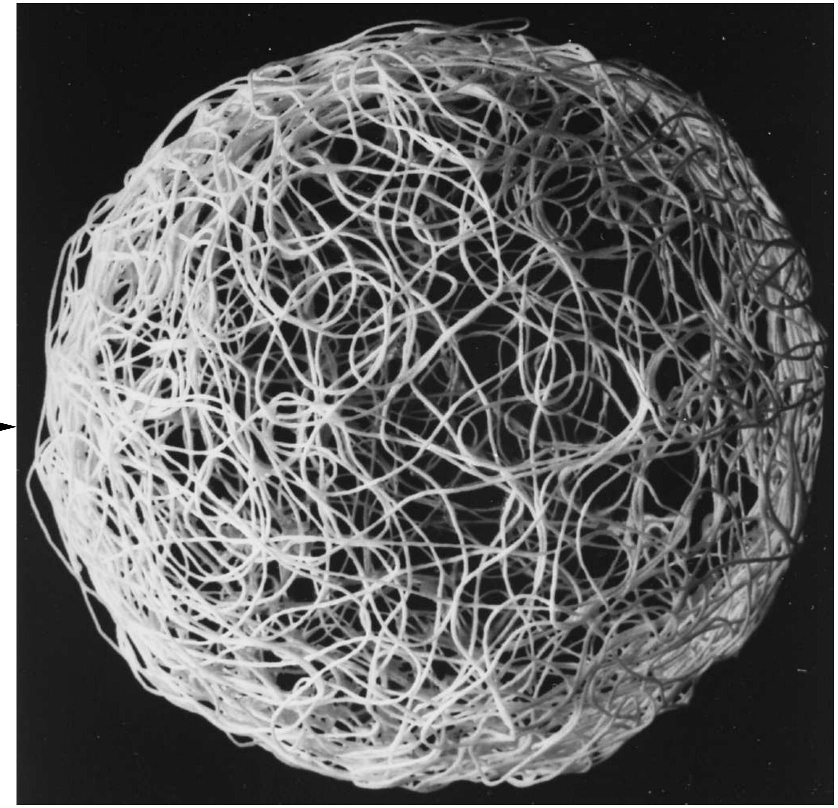
Energy dissipated to escape the Degenerate Sphere must exceed the Noise

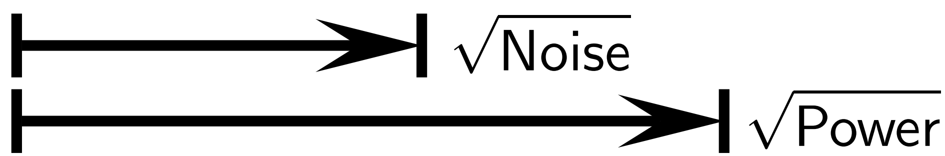
N Dimensional Sphere Separation

Degenerate Sphere



Forward Sphere



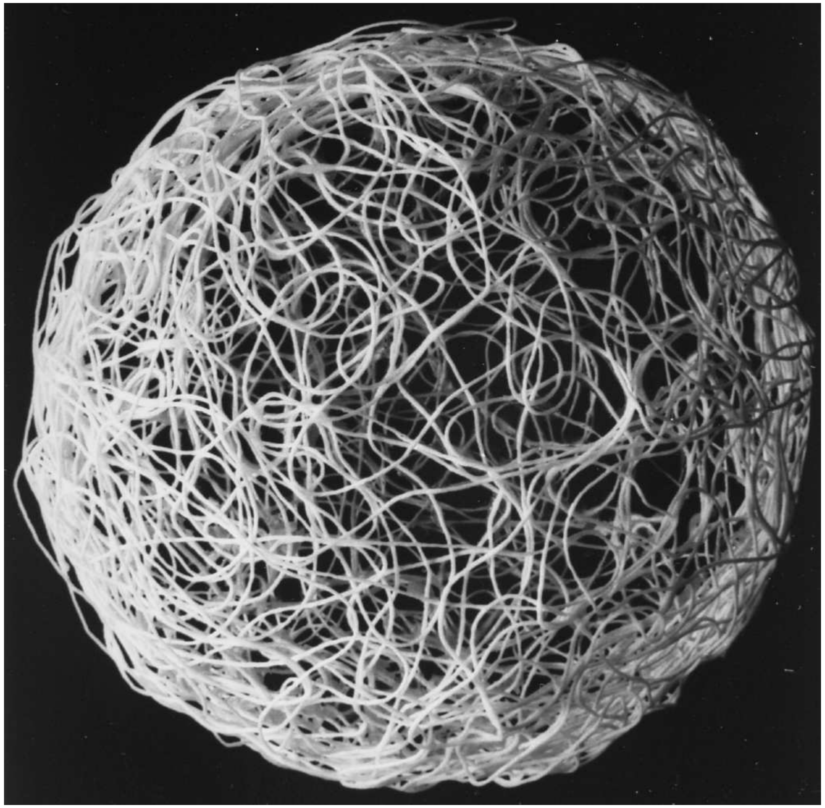


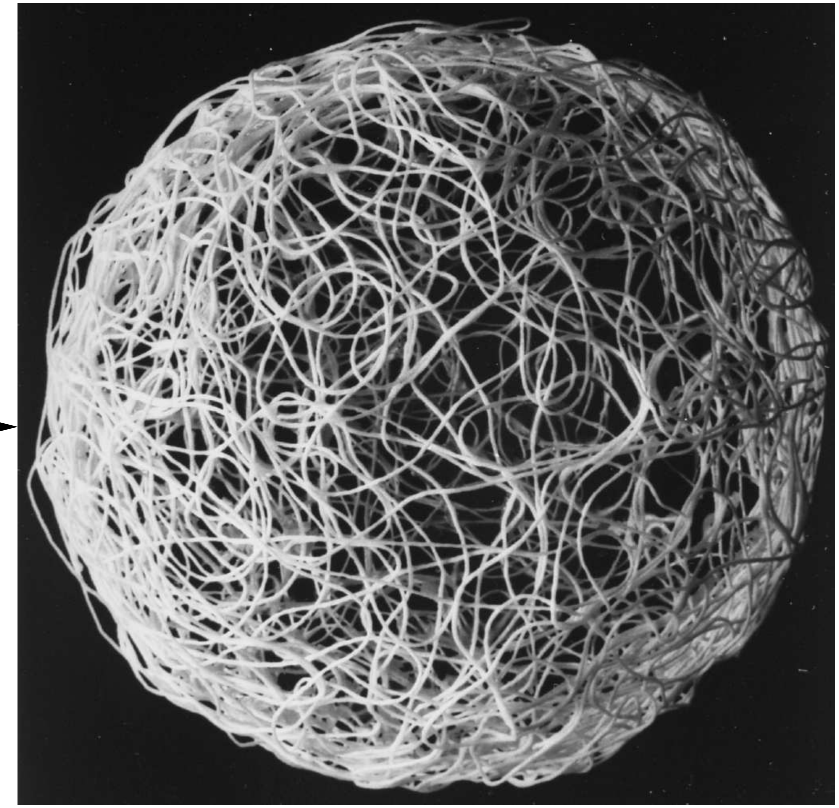
Energy dissipated to escape the Degenerate Sphere must exceed the Noise

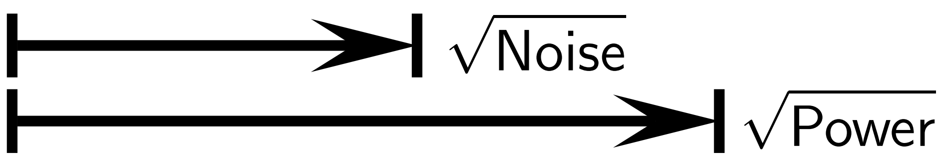
$$\sqrt{\text{Power}} > \sqrt{\text{Noise}}$$

N Dimensional Sphere Separation

Degenerate Sphere | Forward Sphere





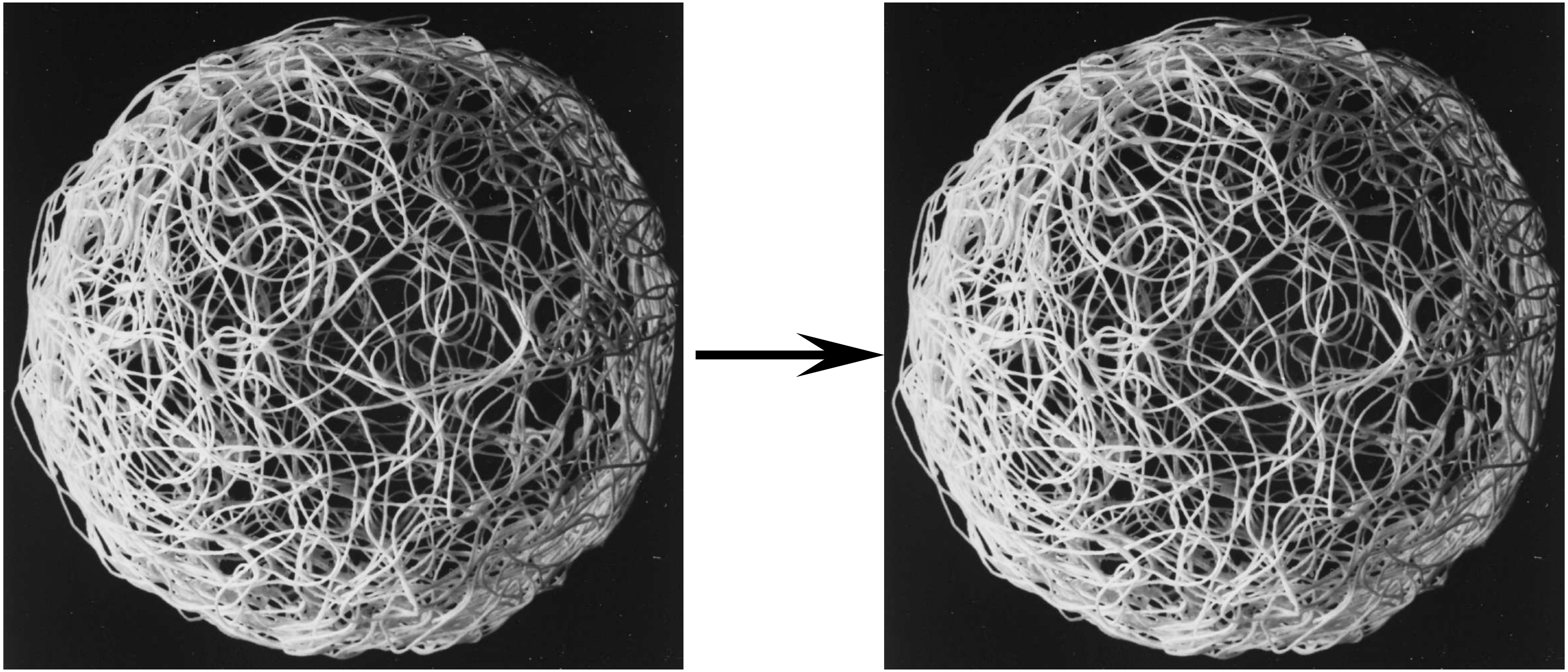


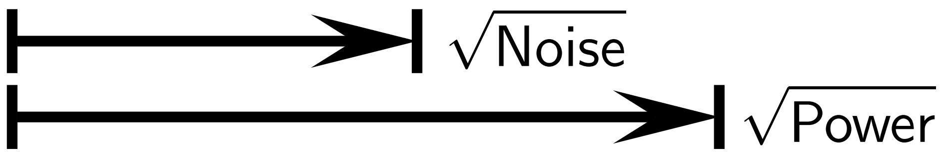
Energy dissipated to escape the Degenerate Sphere must exceed the Noise

$\sqrt{\text{Power}} > \sqrt{\text{Noise}}$ SO Power $>$ Noise

N Dimensional Sphere Separation

Degenerate Sphere　　　　Forward Sphere

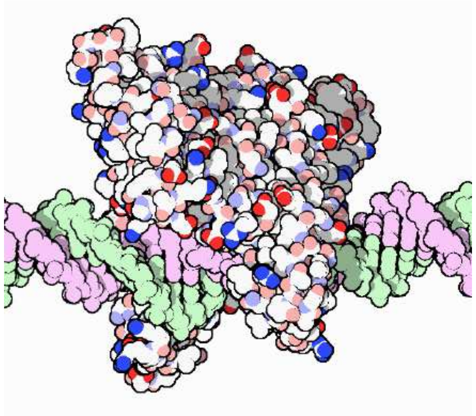




Energy dissipated to escape the Degenerate Sphere must exceed the Noise

$\sqrt{\text{Power}} > \sqrt{\text{Noise}}$ SO Power $>$ Noise SO Power/Noise > 1

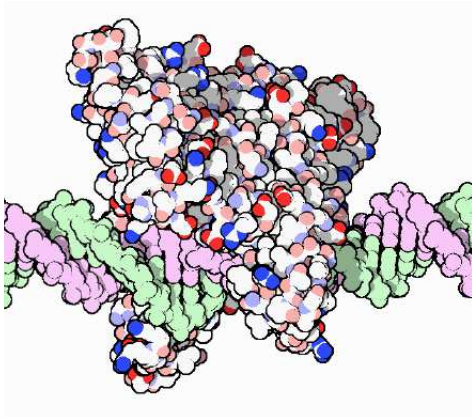
Theoretical Isothermal Efficiency



- For molecular states of molecules with d_{space} 'parts' P energy is dissipated for noise N and

  $C = d_{space} \log_2(P/N + 1) \leftarrow$ machine capacity

Theoretical Isothermal Efficiency



- For molecular states of molecules with d_{space} 'parts' P energy is dissipated for noise N and

  $C = d_{space} \log_2(P/N + 1) \leftarrow$ machine capacity

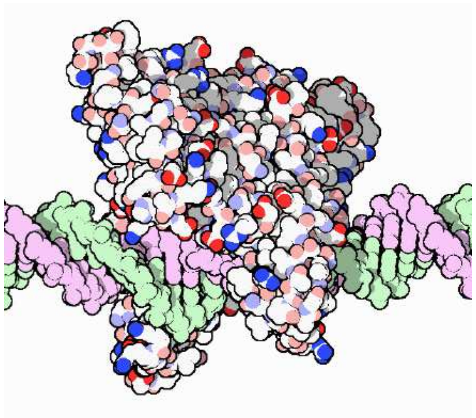
  $\epsilon_t = \frac{\ln\left(\frac{P}{N}+1\right)}{\frac{P}{N}} \leftarrow$ molecular efficiency

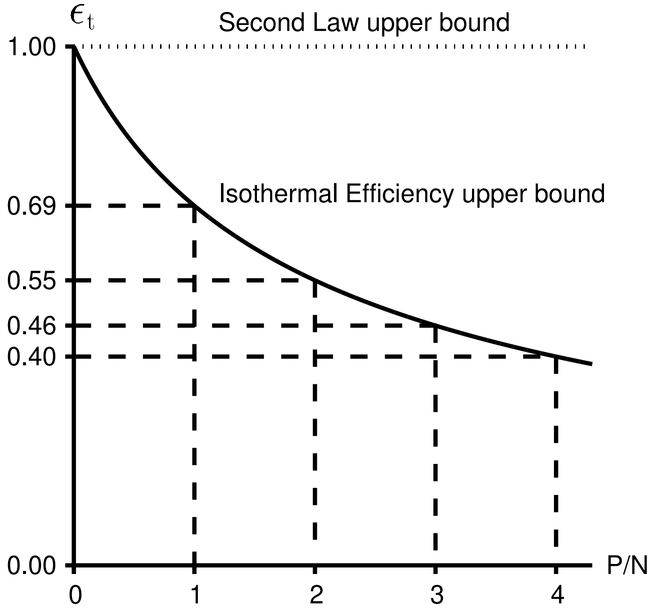
Theoretical Isothermal Efficiency

- For molecular states of molecules with d_{space} 'parts' P energy is dissipated for noise N and

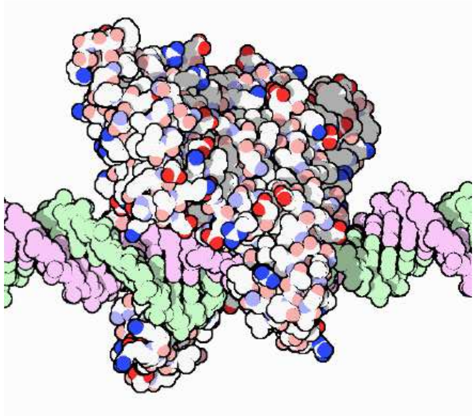
$$C = d_{space}\log_2(P/N+1) \leftarrow \text{machine capacity}$$

$$\epsilon_t = \frac{\ln\left(\frac{P}{N}+1\right)}{\frac{P}{N}} \leftarrow \text{molecular efficiency}$$





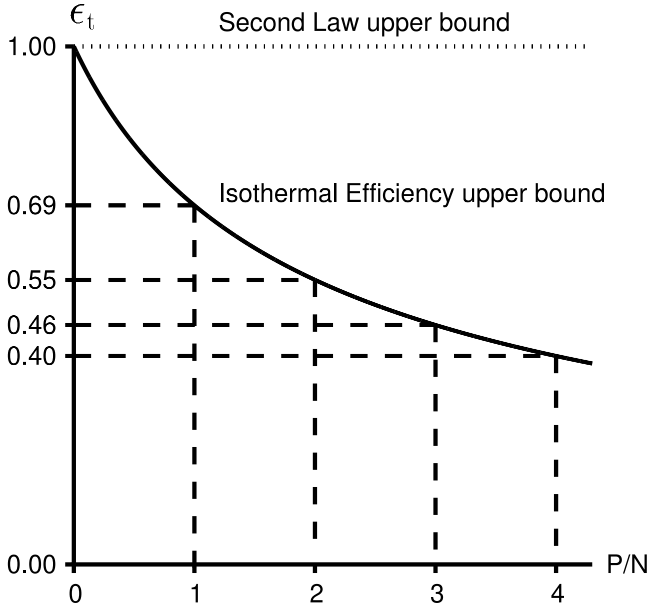
The curve is an upper bound



- For molecular states of molecules with d_{space} ‘parts’ P energy is dissipated for noise N and

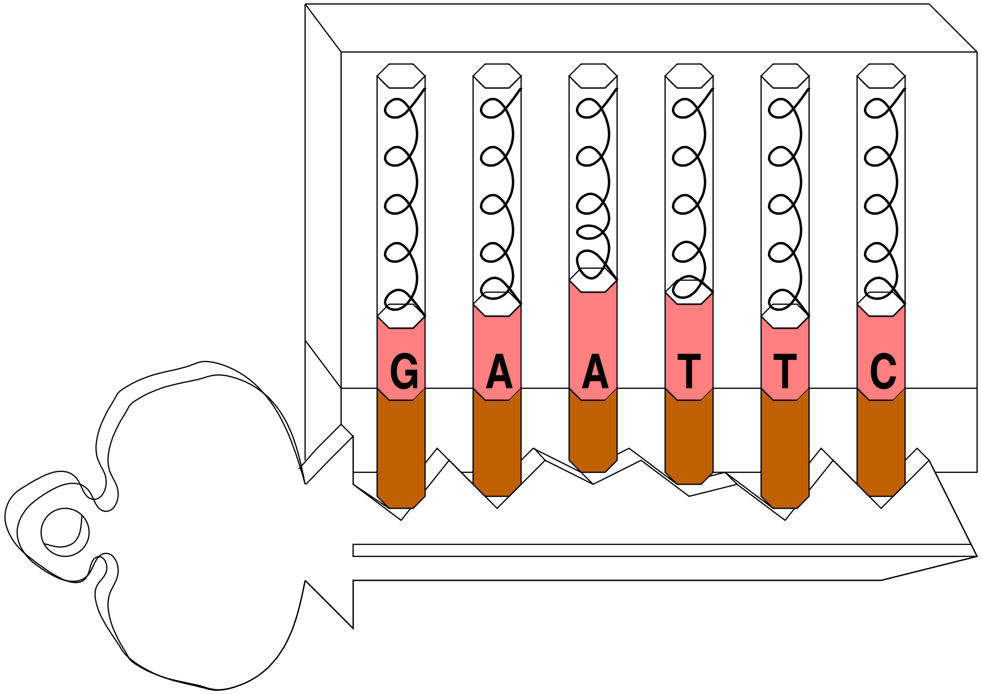
$$C = d_{space} \log_2(P/N + 1) \leftarrow \text{machine capacity}$$

$$\epsilon_t = \frac{\ln\left(\frac{P}{N}+1\right)}{\frac{P}{N}} \leftarrow \text{molecular efficiency}$$



The curve is an upper bound

- If $P/N = 1$ the efficiency is 70%!



Like a key in a lock
which has many independent pins,
it takes many numbers
to describe the vibrational state
of a molecular machine

A Dimensionality Equation

Channel capacity of molecular machine:

$$C = d_{space} \log_2 \left(\frac{P}{N} + 1 \right) \quad \text{(bits per operation)} \tag{7}$$

A Dimensionality Equation

Channel capacity of molecular machine:

$$C = d_{space} \log_2 \left(\frac{P}{N} + 1 \right) \quad \text{(bits per operation)} \tag{7}$$

Maximum of the information R is the channel capacity:

$$C \geq R. \tag{8}$$

A Dimensionality Equation

Channel capacity of molecular machine:

$$C = d_{space} \log_2 \left(\frac{P}{N} + 1 \right) \quad \text{(bits per operation)} \tag{7}$$

Maximum of the information R is the channel capacity:

$$C \geq R. \tag{8}$$

Dimensionality of the coding space:

$$D = 2d_{space} \tag{9}$$

since there are both a phase and an amplitude for each of the independent oscillator pins that describe the motions of a molecule at thermal equilibrium.

A Dimensionality Equation

Channel capacity of molecular machine:

$$C = d_{space} \log_2 \left(\frac{P}{N} + 1 \right) \quad \text{(bits per operation)} \tag{7}$$

Maximum of the information R is the channel capacity:

$$C \geq R. \tag{8}$$

Dimensionality of the coding space:

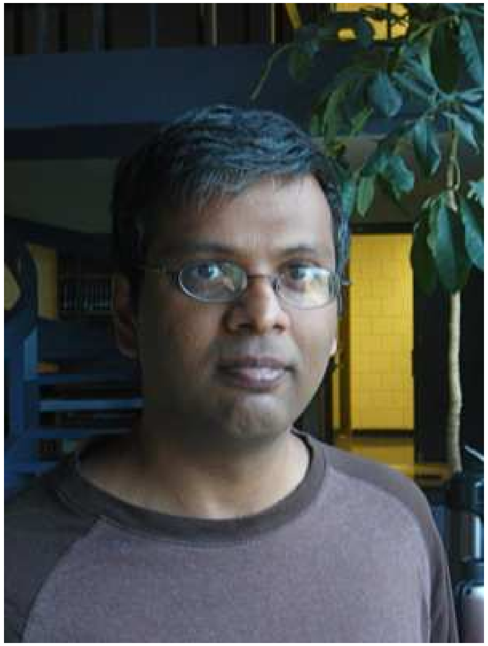
$$D = 2 d_{space} \tag{9}$$

since there are both a phase and an amplitude for each of the independent oscillator pins that describe the motions of a molecule at thermal equilibrium.

Combining equations (7), (8) and (9) gives a lower bound for the dimensionality:

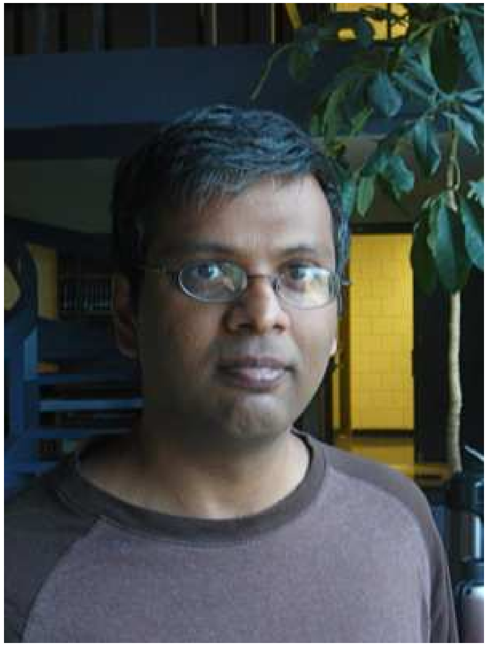
$$D \geq \frac{2R}{\log_2 \left(\frac{P}{N} + 1 \right)}. \tag{10}$$

Vishnu's Observation



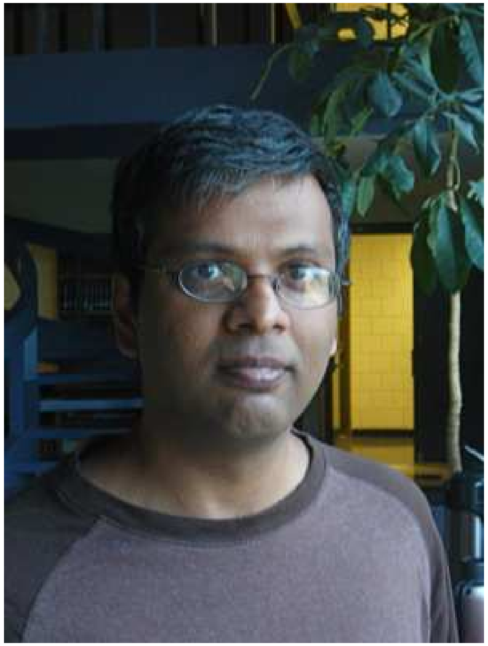
- In 1993 Vishnu Jejjala,
  THEN a graduated high school student
  from the NCI-Frederick Student Intern Program (SIP)

Vishnu's Observation



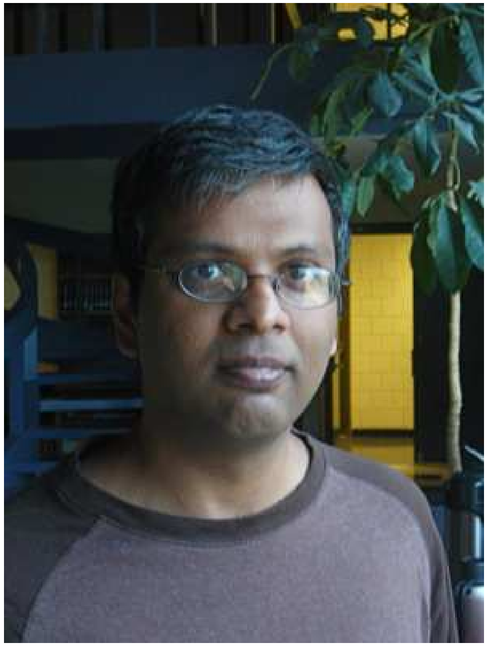
- In 1993 Vishnu Jejjala,
  THEN a graduated high school student
  from the NCI-Frederick Student Intern Program (SIP)
  NOW a string theory physicist
  pointed out that the equation gives a lower bound on the dimension.

Vishnu's Observation



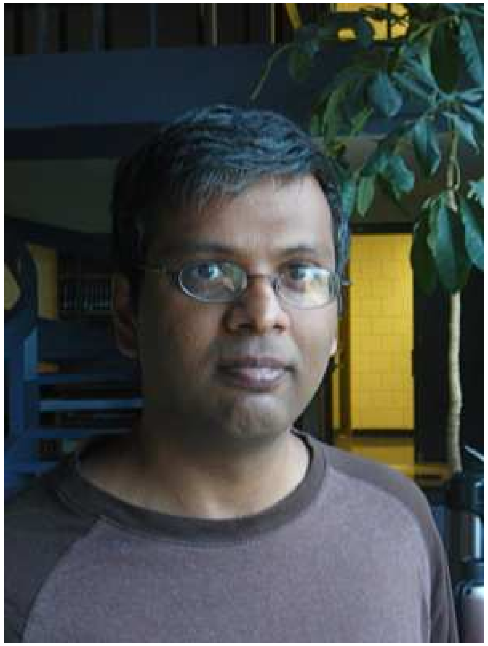
- In 1993 Vishnu Jejjala,
  THEN a graduated high school student
  from the NCI-Frederick Student Intern Program (SIP)
  NOW a string theory physicist
  pointed out that the equation gives a lower bound on the dimension.
- Vishnu suggested there could be **another equation for an upper bound**.

Vishnu's Observation



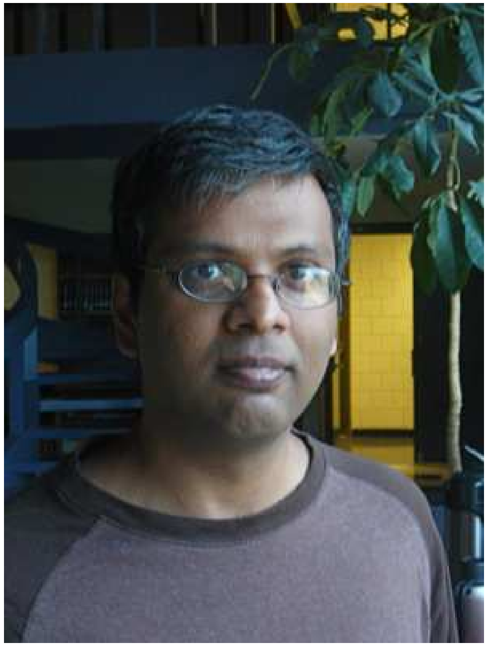
- In 1993 Vishnu Jejjala,
  THEN a graduated high school student
  from the NCI-Frederick Student Intern Program (SIP)
  NOW a string theory physicist
  pointed out that the equation gives a lower bound on the dimension.
- Vishnu suggested there could be **another equation for an upper bound**.
- He suggested that **the two bounds might converge** to give one number.

Vishnu's Observation



- In 1993 Vishnu Jejjala,
  THEN a graduated high school student
  from the NCI-Frederick Student Intern Program (SIP)
  NOW a string theory physicist
  pointed out that the equation gives a lower bound on the dimension.
- Vishnu suggested there could be **another equation for an upper bound**.
- He suggested that **the two bounds might converge** to give one number.
- He set out to find that equation.

Vishnu's Observation



- In 1993 Vishnu Jejjala,
  THEN a graduated high school student
  from the NCI-Frederick Student Intern Program (SIP)
  NOW a string theory physicist
  pointed out that the equation gives a lower bound on the dimension.
- Vishnu suggested there could be **another equation for an upper bound**.
- He suggested that **the two bounds might converge** to give one number.
- He set out to find that equation.
- He did not succeed.

18 years later ...

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.
- Each degree of freedom carries $\frac{1}{2}k_{\mathrm{B}}T$ energy.

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.
- Each degree of freedom carries $\frac{1}{2}k_{\mathrm{B}}T$ energy.
- But since they are independent, degrees of freedom are dimensions D!

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.
- Each degree of freedom carries $\frac{1}{2}k_{\mathrm{B}}T$ energy.
- But since they are independent, degrees of freedom are dimensions D!
- For n atoms there are $3n - 6$ degrees of freedom.

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.
- Each degree of freedom carries $\frac{1}{2}k_{\mathrm{B}}T$ energy.
- But since they are independent, degrees of freedom are dimensions D!
- For n atoms there are $3n - 6$ degrees of freedom.
- A molecule can use only some of these: $D \ll 3n - 6$.

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.
- Each degree of freedom carries $\frac{1}{2}k_{\mathrm{B}}T$ energy.
- But since they are independent, degrees of freedom are dimensions D!
- For n atoms there are $3n - 6$ degrees of freedom.
- A molecule can use only some of these: $D \ll 3n - 6$.
- So the relevant thermal noise energy flowing through a molecule is:

$$N = \tfrac{1}{2}k_{\mathrm{B}}TD \qquad \text{(joules per mmo)} \tag{11}$$

(mmo $=$ molecular machine operation)

Key Discovery: July 2011

- Tom was reading a 1983 paper by the famous information theorist Jaynes.
- Jaynes pointed out that
  the total energy in a molecule
  depends on the number of degrees of freedom.
- Each degree of freedom carries $\frac{1}{2}k_{\mathrm{B}}T$ energy.
- But since they are independent, degrees of freedom are dimensions D!
- For n atoms there are $3n - 6$ degrees of freedom.
- A molecule can use only some of these: $D \ll 3n - 6$.
- So the relevant thermal noise energy flowing through a molecule is:

$$N = \tfrac{1}{2}k_{\mathrm{B}}TD \qquad \text{(joules per mmo)} \tag{11}$$

  (mmo = molecular machine operation)
- Tom already had this equation in 1991!

Rearranging the Equation to get the Dimensionality

- Tom's 70% discovery implies that the energy a molecule dissipates to make selections must exceed this thermal noise:

$$P > N \tag{12}$$

Rearranging the Equation to get the Dimensionality

- Tom's 70% discovery implies that the energy a molecule dissipates to make selections must exceed this thermal noise:

$$P > N \tag{12}$$

So plugging in N:

$$P > N = \frac{1}{2} k_{\mathsf{B}} T D \tag{13}$$

Rearranging the Equation to get the Dimensionality

- Tom's 70% discovery implies that the energy a molecule dissipates to make selections must exceed this thermal noise:

$$P > N \tag{12}$$

So plugging in N:

$$P > N = \tfrac{1}{2} k_{\mathsf{B}} T D \tag{13}$$

Rearrange:

$$\frac{P}{\frac{1}{2} k_{\mathsf{B}} T} > D. \tag{14}$$

That's an upper bound on the dimensionality!

Rearranging the Equation to get the Dimensionality

- Tom's 70% discovery implies that the energy a molecule dissipates to make selections must exceed this thermal noise:

$$P > N \tag{12}$$

So plugging in N:

$$P > N = \tfrac{1}{2} k_{\mathrm{B}} T D \tag{13}$$

Rearrange:

$$\frac{P}{\frac{1}{2} k_{\mathrm{B}} T} > D. \tag{14}$$

That's an upper bound on the dimensionality!

Vishnu was right!
There is an equation
for the upper bound!

Convert to more useful form - Part 1 - Definitions

- The energy available in coding space for making selections is the free energy:

$$P = -\Delta G^{\circ} \quad \text{(joules per mmo)} \tag{15}$$

Convert to more useful form - Part 1 - Definitions

- The energy available in coding space for making selections is the free energy:

$$P = -\Delta G^{\circ} \quad \text{(joules per mmo)} \tag{15}$$

- The maximum bits that can be gained for that free energy is

$$R_{energy} \equiv -\Delta G^{\circ} / \mathcal{E}_{min} \quad \text{(bits per mmo)} \tag{16}$$

Convert to more useful form - Part 1 - Definitions

- The energy available in coding space for making selections is the free energy:

$$P = -\Delta G^{\circ} \quad \text{(joules per mmo)} \tag{15}$$

- The maximum bits that can be gained for that free energy is

$$R_{energy} \equiv -\Delta G^{\circ}/\mathcal{E}_{min} \quad \text{(bits per mmo)} \tag{16}$$

- Use the second law of thermodynamics as an ideal conversion factor between energy and bits:

$$\mathcal{E}_{min} = k_{\mathsf{B}} T \ln 2 \quad \text{(joules per bit)} \tag{17}$$

Convert to more useful form - Part 1 - Definitions

- The energy available in coding space for making selections is the free energy:

$$P \quad = \quad -\Delta G^{\circ} \qquad \text{(joules per mmo)} \tag{15}$$

- The maximum bits that can be gained for that free energy is

$$R_{energy} \equiv -\Delta G^{\circ}/\mathcal{E}_{min} \qquad \text{(bits per mmo)} \tag{16}$$

- Use the second law of thermodynamics as an ideal conversion factor between energy and bits:

$$\mathcal{E}_{min} \quad = \quad k_{\mathsf{B}} T \ln 2 \qquad \text{(joules per bit)} \tag{17}$$

- A measured isothermal efficiency, $\epsilon_r < \epsilon_t$, is defined by the information gained, R, *versus* the information that could be gained for the given energy dissipation, R_{energy}:

$$\epsilon_r \quad = \quad R/R_{energy} \tag{18}$$

Convert to more useful form - Part 2 - Substitutions

- combining equations (15) to (18) gives

$$
\begin{aligned}
P &= \mathcal{E}_{min} R_{energy} && (19)\\
&= k_{\mathsf{B}} T \ln 2 \, R_{energy} && (20)\\
&= k_{\mathsf{B}} T R \ln 2 / \epsilon_r . && (21)
\end{aligned}
$$

Convert to more useful form - Part 2 - Substitutions

- combining equations (15) to (18) gives

$$\begin{aligned} P &= \mathcal{E}_{min} R_{energy} && (19) \\ &= k_{\mathsf{B}} T \ln 2 \, R_{energy} && (20) \\ &= k_{\mathsf{B}} T R \ln 2 / \epsilon_r . && (21) \end{aligned}$$

- Inserting this result into equation (14) gives

$$\frac{2R \ln 2}{\epsilon_r} > D \quad (22)$$

Convert to more useful form - Part 2 - Substitutions

- combining equations (15) to (18) gives

$$
\begin{aligned}
P &= \mathcal{E}_{min} R_{energy} && (19)\\
&= k_{\mathsf{B}} T \ln 2\, R_{energy} && (20)\\
&= k_{\mathsf{B}} T R \ln 2 / \epsilon_r . && (21)
\end{aligned}
$$

- Inserting this result into equation (14) gives

$$\frac{2R\ln 2}{\epsilon_r} > D \qquad (22)$$

- Equation (22) is an

upper bound on the dimensionality
as a function of the information gain R
and the isothermal efficiency ϵ_r.

Bounds on the dimensionality of molecular machines

- Combining the lower bound (10) with the upper bound (22)

$$\frac{2R}{\log_2\left(\frac{P}{N}+1\right)} \leq D < \frac{2R\ln 2}{\epsilon_r} \qquad (23)$$

Bounds on the dimensionality of molecular machines

- Combining the lower bound (10) with the upper bound (22)

$$\frac{2R}{\log_2\left(\frac{P}{N}+1\right)} \leq D < \frac{2R\ln 2}{\epsilon_r} \tag{23}$$

- To simplify terminology, define $\rho = P/N$

Bounds on the dimensionality of molecular machines

- Combining the lower bound (10) with the upper bound (22)

$$\frac{2R}{\log_2\left(\frac{P}{N}+1\right)} \leq D < \frac{2R\ln 2}{\epsilon_r} \tag{23}$$

- To simplify terminology, define $\rho = P/N$
- Notice that $\log_2(\rho+1) = \frac{\ln(\rho+1)}{\ln 2}$

- Combining the lower bound (10) with the upper bound (22)

$$\frac{2R}{\log_2\left(\frac{P}{N}+1\right)} \leq D < \frac{2R\ln 2}{\epsilon_r} \tag{23}$$

- To simplify terminology, define $\rho = P/N$
- Notice that $\log_2(\rho+1) = \frac{\ln(\rho+1)}{\ln 2}$
- So

$$\frac{2R\ln 2}{\ln\left(\rho+1\right)} \leq D < \frac{2R\ln 2}{\epsilon_r} \tag{24}$$

A beautifully symmetrical equation!

$$\frac{2R\ln 2}{\ln(\rho+1)} \le D < \frac{2R\ln 2}{\epsilon_r}$$

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho+1)} \leq D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho + 1)} \leq D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

- R - the information gain in bits

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho+1)} \leq D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

- R - the information gain in bits
- ϵ_r - the real isothermal efficiency

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho+1)} \leq D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

- R - the information gain in bits
- ϵ_r - the real isothermal efficiency
- ρ - the normalized energy dissipation

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho+1)} \le D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

- R - the information gain in bits
- ϵ_r - the real isothermal efficiency
- ρ - the normalized energy dissipation

As a molecular machine evolves:

$\rho \to 1$ — The left hand side converges to $2R$.

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho+1)} \leq D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

- R - the information gain in bits
- ϵ_r - the real isothermal efficiency
- ρ - the normalized energy dissipation

As a molecular machine evolves:

$\rho \to 1$	The left hand side converges to $2R$.
$\epsilon_r \to \ln(2)$	The right hand side converges to $2R$.

Pincers on the dimensionality of molecular machines

$$\frac{2R\ln 2}{\ln(\rho+1)} \leq D < \frac{2R\ln 2}{\epsilon_r}$$

The dimensionality of a molecular machine is bounded on the two sides by three factors:

- R - the information gain in bits
- ϵ_r - the real isothermal efficiency
- ρ - the normalized energy dissipation

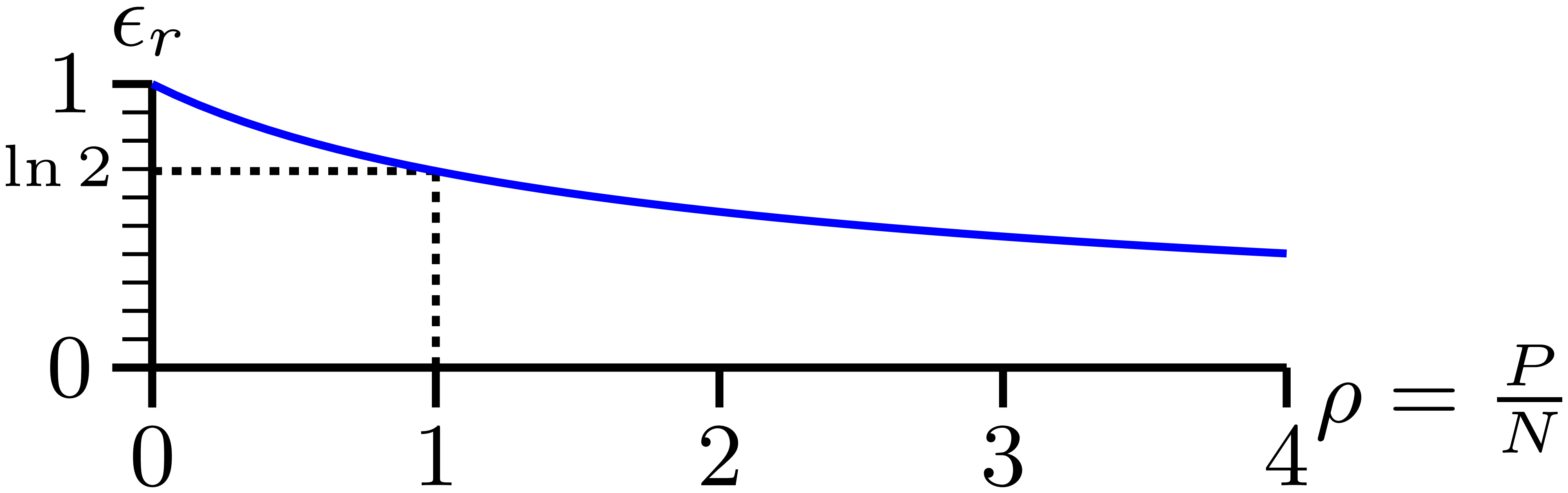
As a molecular machine evolves:

$\rho \to 1$	The left hand side converges to $2R$.
$\epsilon_r \to \ln(2)$	The right hand side converges to $2R$.

BOTH SIDES converge to $2R$!

Vishnu was right
about the convergence!

Efficiency curve bounds

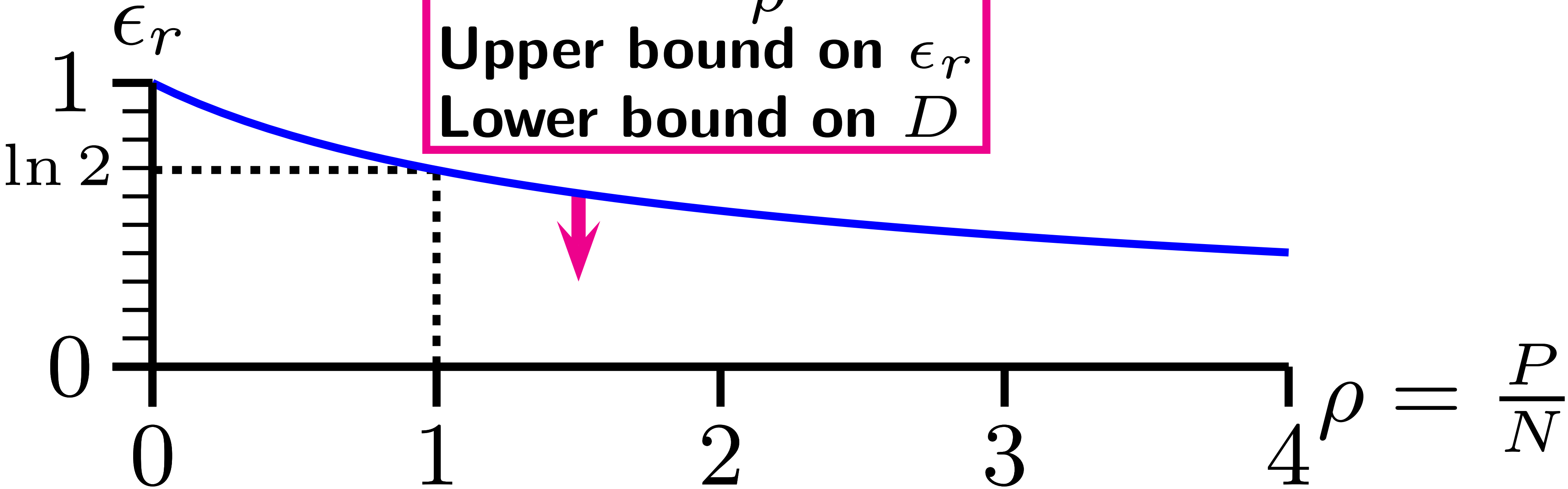


Efficiency curve bounds

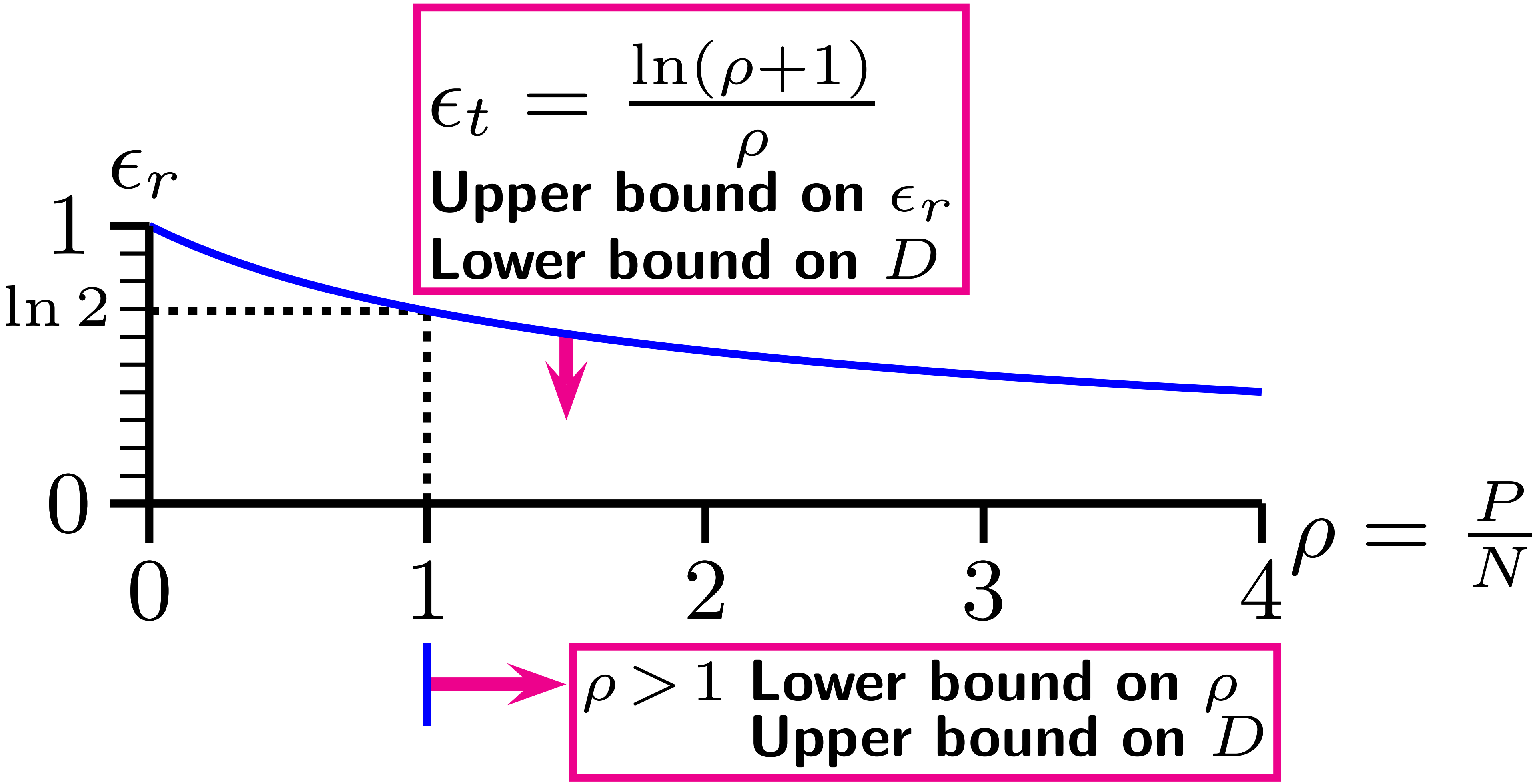
$$\epsilon_t = \frac{\ln(\rho+1)}{\rho}$$

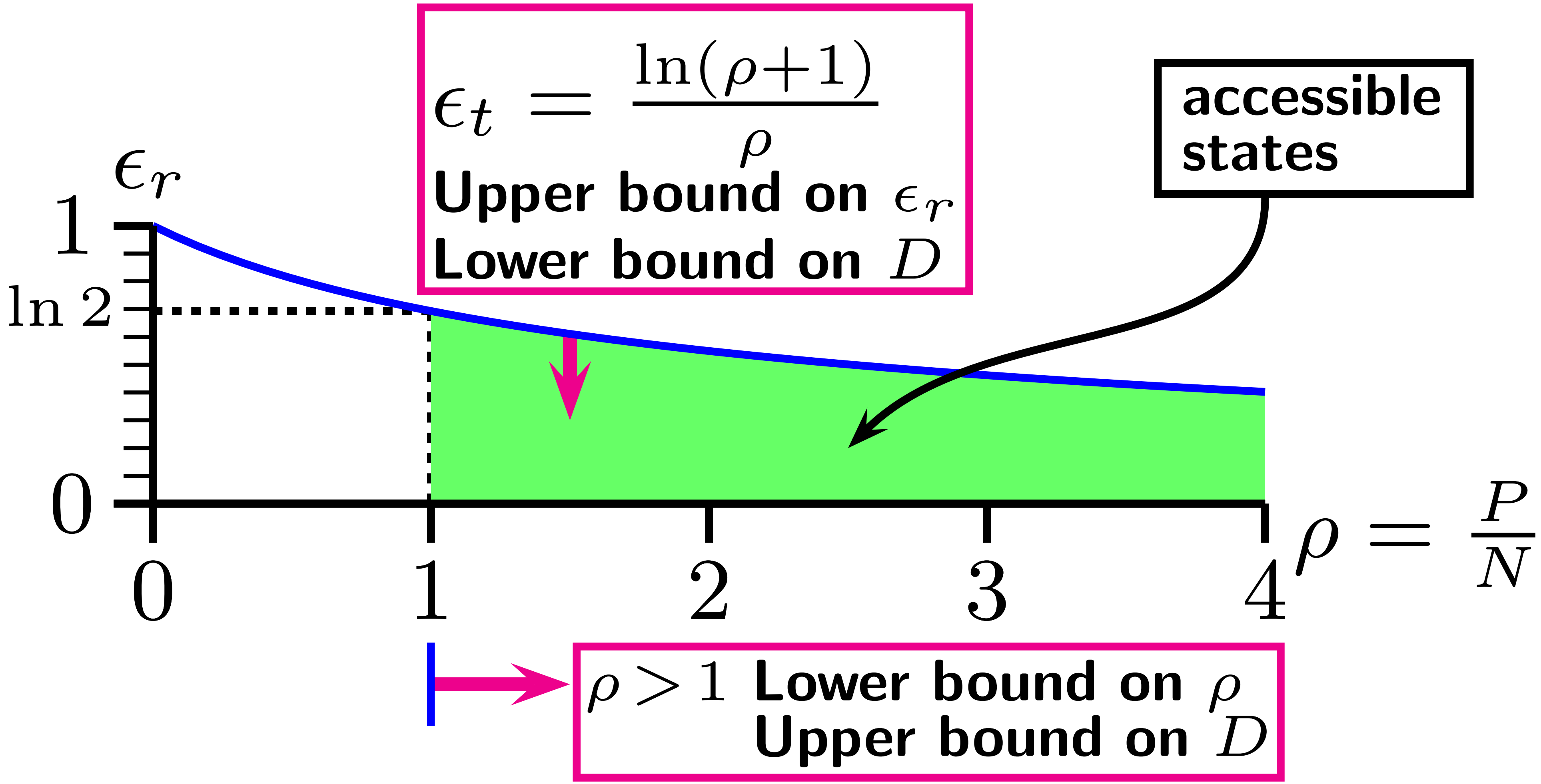
Upper bound on ϵ_r

Lower bound on D



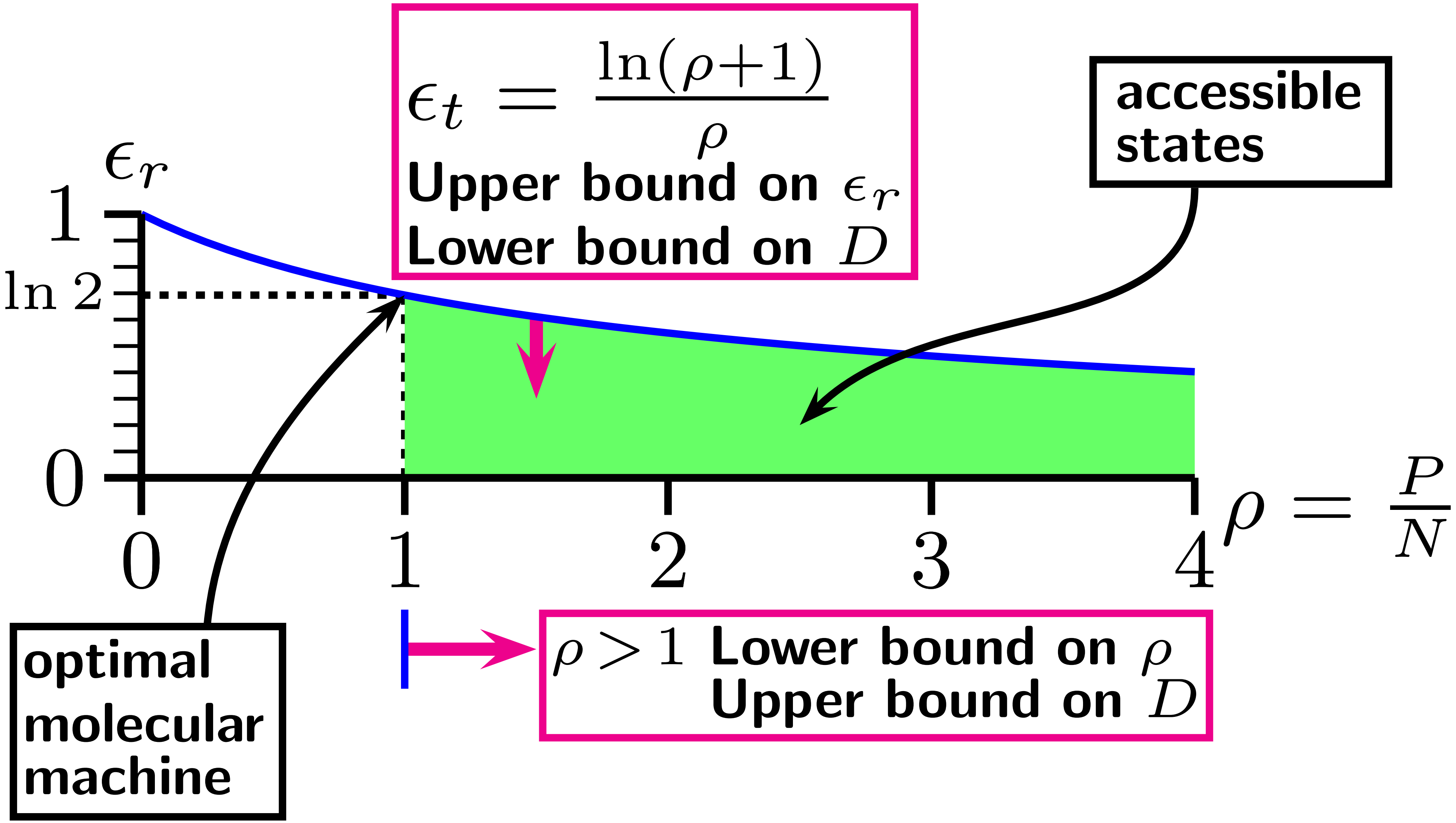
Efficiency curve bounds

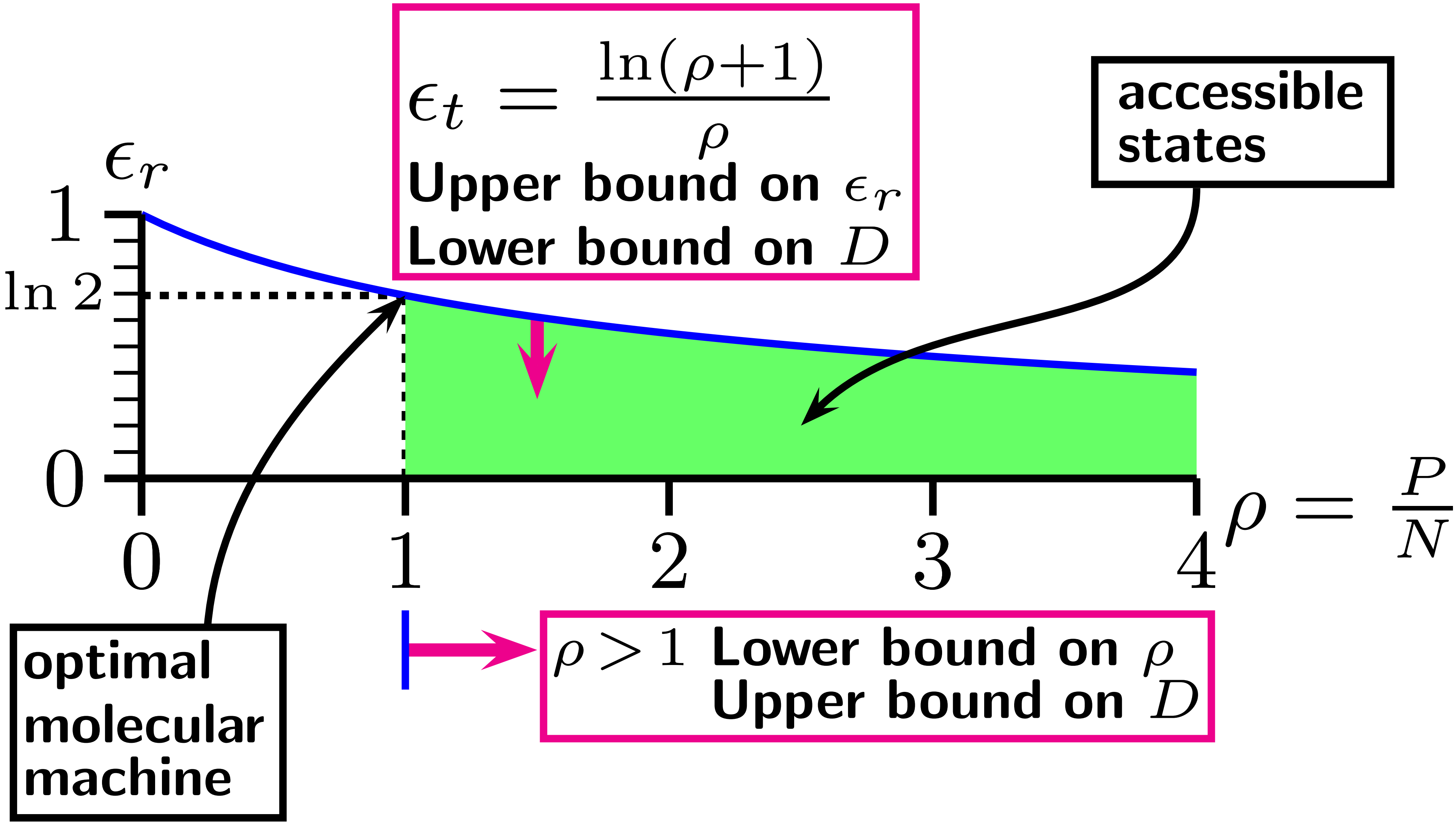




$\epsilon_t = \frac{\ln(\rho+1)}{\rho}$
Upper bound on ϵ_r
Lower bound on D
accessible states
ϵ_r
1
$\ln 2$
0
0
1
2
3
4
$\rho = \frac{P}{N}$
$\rho > 1$ Lower bound on ρ
Upper bound on D

Efficiency curve bounds





$D = 2R$ **when the molecular machine is optimal**

If a molecular machine has evolved to optimum, then the dimensionality is

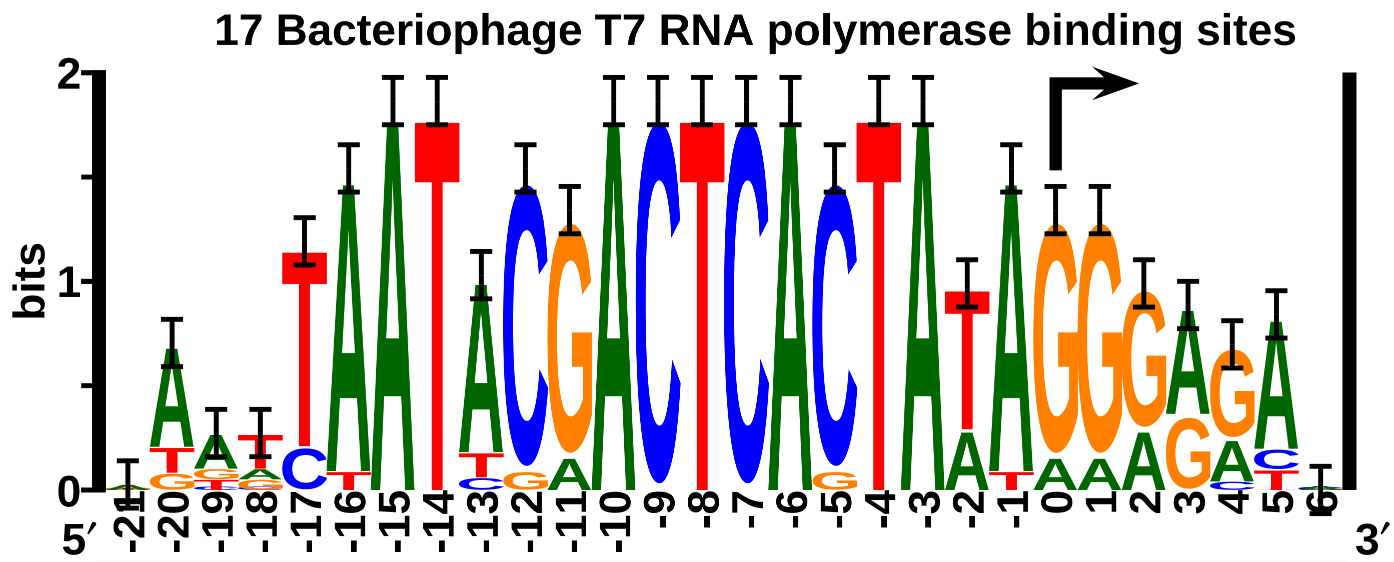
$$D = 2R$$

If a molecular machine has evolved to optimum, then the dimensionality is

$$D = 2R$$

Let's calculate D for restriction enzymes!

Sequence Logo

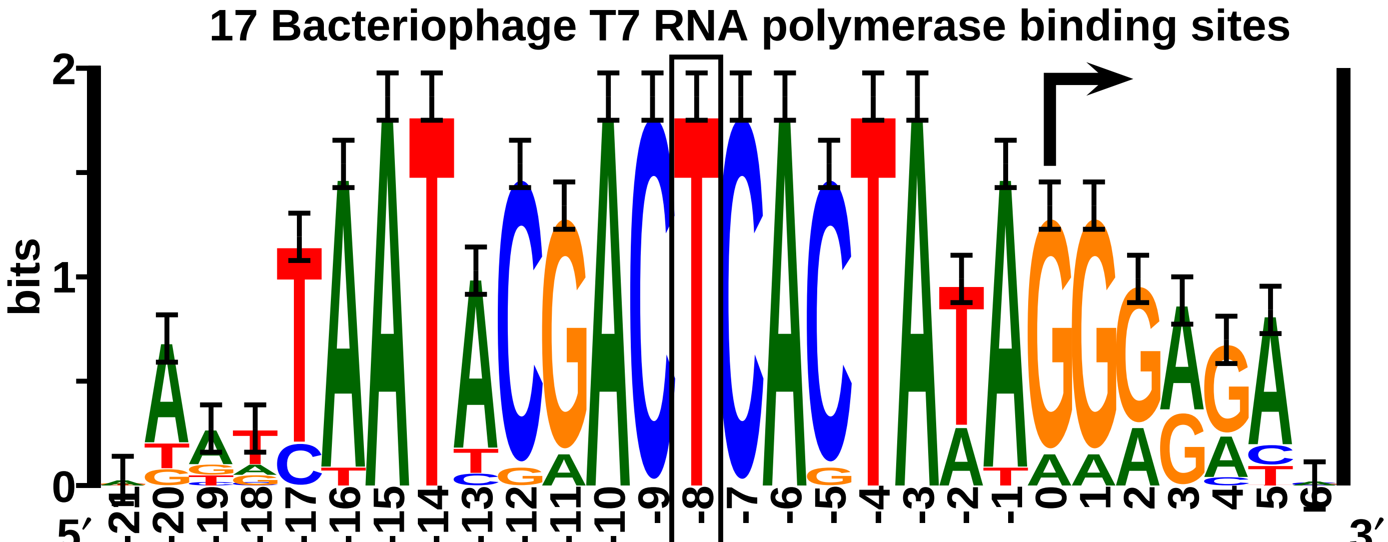


Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

```
1 ttattaatacaactcactataaggagag
2 aaatcaatacgactcactatagagggac
3 cggttaatacgactcactataggagaac
4 gaagtaatacgactcagtatagggacaa
5 taattaattgaactcactaaagggagac
6 cgcttaatacgactcactaaaggagaca
```

6 of 17 sites

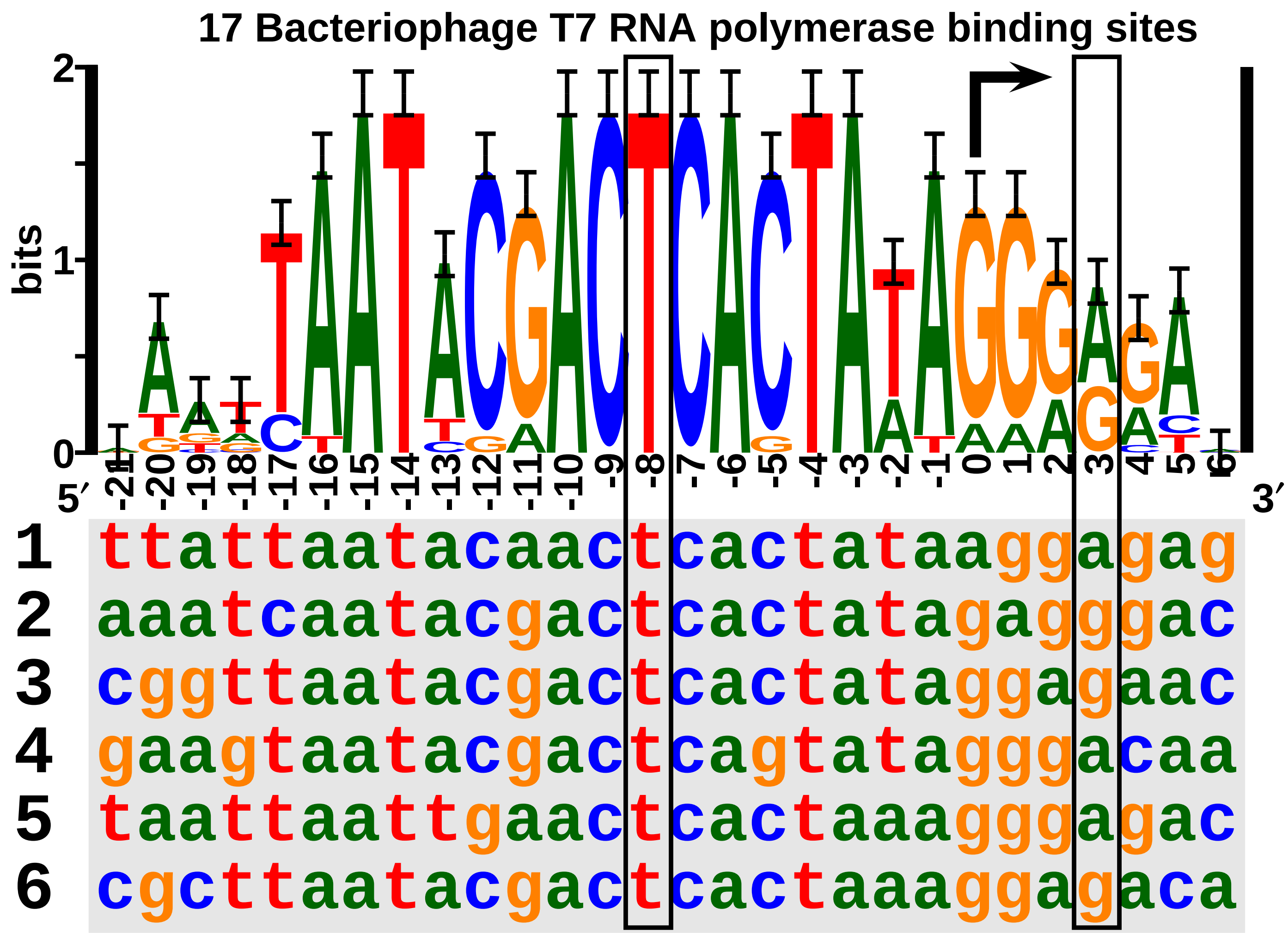
Sequence Logo



1 ttattaatacaactcactataaggagag
2 aaatcaatacgactcactatagagggac
3 cggttaatacgactcactataggagaac
4 gaagtaatacgactcagtatagggacaa
5 taattaattgaactcactaaagggagac
6 cgcttaatacgactcactaaaggagaca

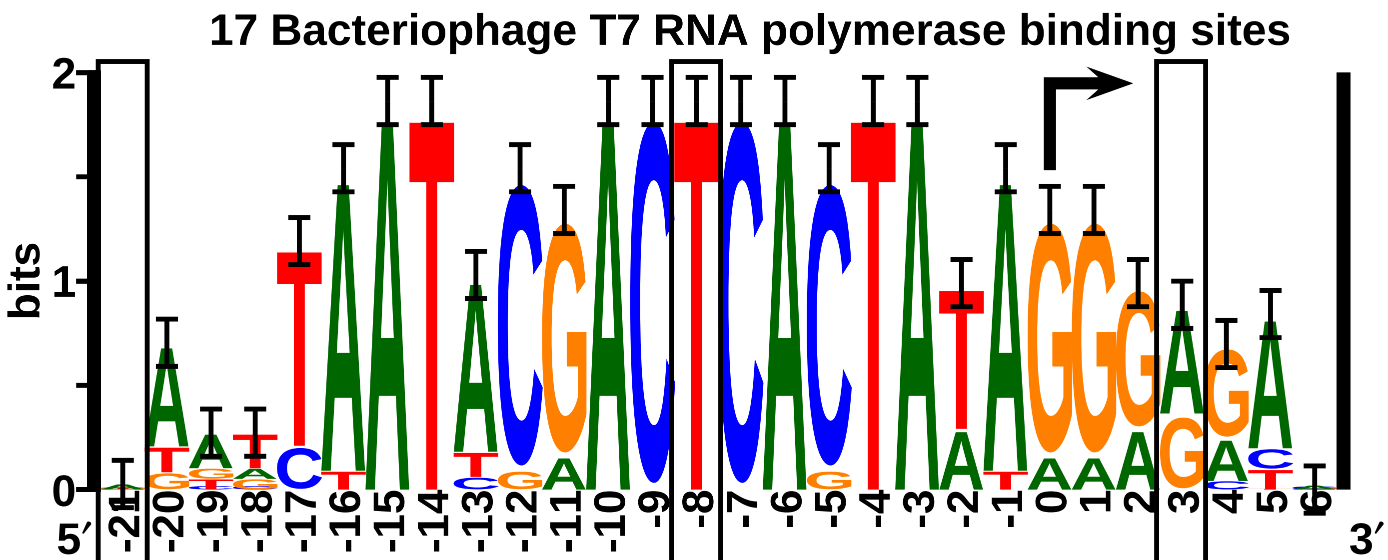
6 of 17 sites

Schneider & Stephens *Nucl. Acids Res.* **18**: 6097-6100 1990



6 of 17 sites

Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990



Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

1 ttattaatacaactcactataaggagag
2 aaatcaatacgactcactatagagggac
3 cggttaatacgactcactataggagaac
4 gaagtaatacgactcagtatagggacaa
5 taattaattgaactcactaaagggagac
6 cgcttaatacgactcactaaaggagaca

6 of 17 sites

Example:

Example:

- EcoRI

Example:

- EcoRI
- $5'$ G$\downarrow$AATTC $3'$

Example:

- EcoRI
- $5'$ G$\downarrow$AATTC $3'$
- 6 bases: selecting 1 in 4.

Example:

- EcoRI
- $5'$ G↓AATTC $3'$
- 6 bases: selecting 1 in 4.
- Uncertainty before binding: 2 bits
  Uncertainty after binding: 0 bits
  Decrease in uncertainty: 2 bits

Example:

- EcoRI
- $5'$ G↓AATTC $3'$
- 6 bases: selecting 1 in 4.
- Uncertainty before binding: 2 bits
  Uncertainty after binding: 0 bits
  Decrease in uncertainty: 2 bits
- $6 \times (2 - 0) = 12$ bits

Example:

- EcoRI
- $5'$ G$\downarrow$AATTC $3'$
- 6 bases: selecting 1 in 4.
- Uncertainty before binding: 2 bits
  Uncertainty after binding: 0 bits
  Decrease in uncertainty: 2 bits
- $6 \times (2 - 0) = 12$ bits
- $12 \times 2 = 24$ dimensions

Example:

Example:

- $5'$ GTY↓RAC $3'$ HincIII

Example:

- $5'$ GTY↓RAC $3'$ HincIII
- $5'$ GT AC $3'$
- GT AC :$(2-0) \times 4 = 8$ bits

total $= 8$

Example:

- $5'$ GTY↓RAC $3'$ HincIII
- $5'$ GTY↓RAC $3'$
- GT AC $:(2-0) \times 4 = 8$ bits

  Y↓R $:(2-1) \times 2 = 2$ bits

  total $= 8 + 2$

Example:

- $5'$ GTY↓RAC $3'$ HincIII
- $5'$ GTY↓RAC $3'$
- GT AC $:(2-0)\times 4 = 8$ bits

  Y↓R $:(2-1)\times 2 = 2$ bits

  total $= 8 + 2 = 10$ bits

Example:

- $5'$ GTY↓RAC $3'$ HincIII
- $5'$ GTY↓RAC $3'$
- GT AC $:(2-0)\times 4 = 8$ bits

  Y↓R $:(2-1)\times 2 = 2$ bits

  total $= 8 + 2 = 10$ bits
- 10 bits$/2 = 5$ compressed bases

Example:

- $5'$ GTY$\downarrow$RAC $3'$ HincIII
- $5'$ GTY$\downarrow$RAC $3'$
- GT AC $:(2-0)\times 4 = 8$ bits

  Y$\downarrow$R $:(2-1)\times 2 = 2$ bits

  total $= 8 + 2 = 10$ bits
- 10 bits$/2 = 5$ compressed bases
- 10 bits$\times 2 = 20$ dimensions

Example:

Example:

- $5'$ VCW $3'$ RlaI

Example:

- $5'$ VCW $3'$ RlaI
- $5'$ C $3'$
- C : $2 - 0 = 2$ bits

total $= 2$

Example:

- $5'$ VCW $3'$ RlaI
- $5'$ CW $3'$
- C : $2 - 0 = 2$ bits

  W = A/T: $2 - 1 = 1$ bit

total $= 2 + 1$

Example:

- $5'$ VCW $3'$ RlaI
- $5'$ VCW $3'$
- C : $2 - 0 = 2$ bits

  W $=$ A/T: $2 - 1 = 1$ bit

  V $=$ A,C,G: $2 - \log_2 3 \approx 0.42$ bits

  total $= 2 + 1 + 0.42$

Example:

- $5'$ VCW $3'$ RlaI
- $5'$ VCW $3'$
- C : $2 - 0 = 2$ bits

  W = A/T: $2 - 1 = 1$ bit

  V = A,C,G:

  $2 - \log_2 3 \approx 0.42$ bits

  total $= 2 + 1 + 0.42 = 3.42$ bits

Example:

- $5'$ VCW $3'$ RlaI
- $5'$ VCW $3'$
- C : $2 - 0 = 2$ bits

  W $=$ A/T: $2 - 1 = 1$ bit

  V $=$ A,C,G:

  $2 - \log_2 3 \approx 0.42$ bits

  total $= 2 + 1 + 0.42 = 3.42$ bits
- 3.42 bits $/2 = 1.71$ compressed bases

Example:

- $5'$ VCW $3'$ RlaI
- $5'$ VCW $3'$
- C : $2 - 0 = 2$ bits

  W $=$ A/T: $2 - 1 = 1$ bit

  V $=$ A,C,G:

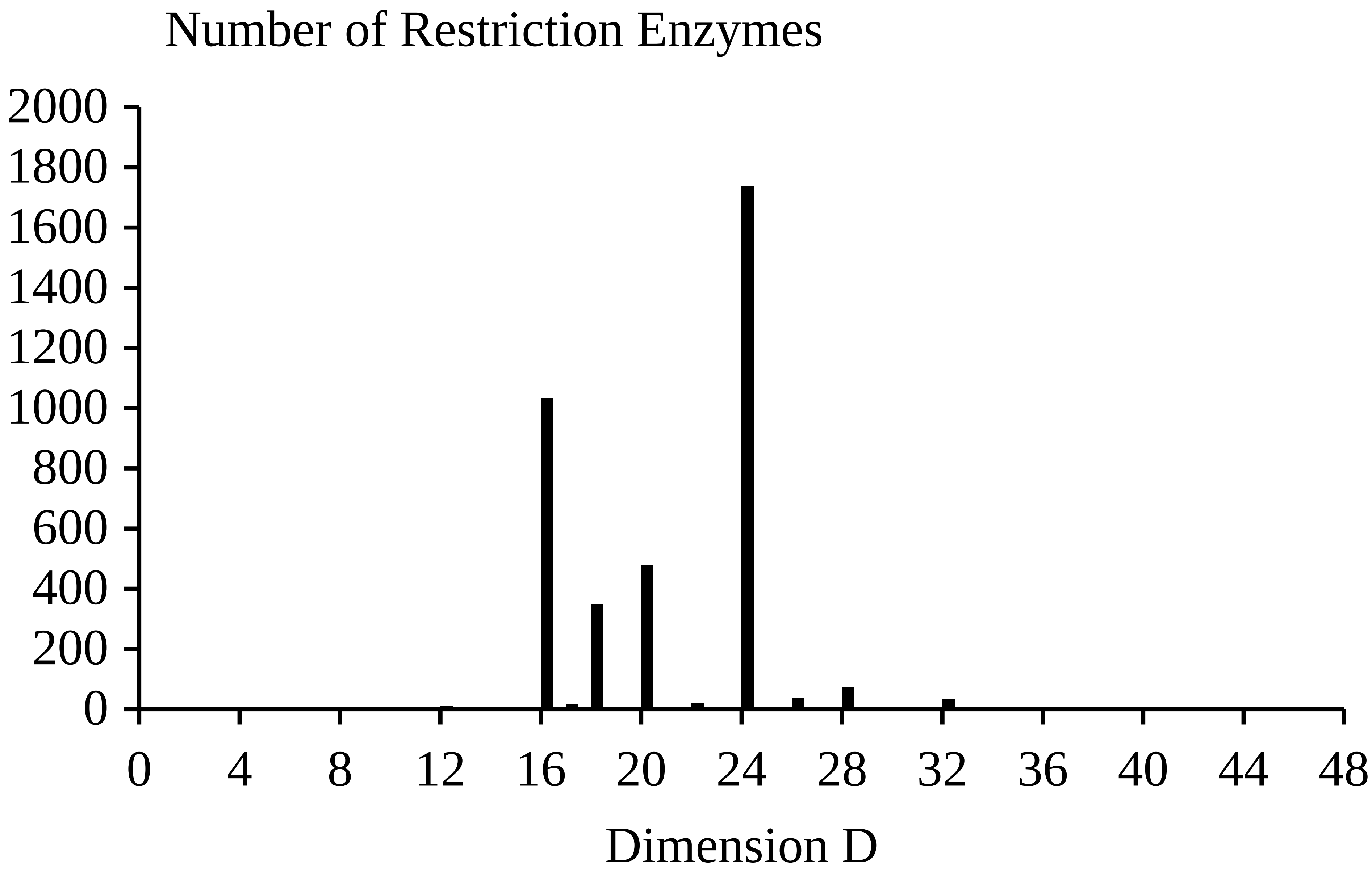
  $2 - \log_2 3 \approx 0.42$ bits

  total $= 2 + 1 + 0.42 = 3.42$ bits
- 3.42 bits $/2 = 1.71$ compressed bases
- 3.42 bits $\times 2 = 6.83$ dimensions

Restriction Enzyme Coding Space Dimensionality

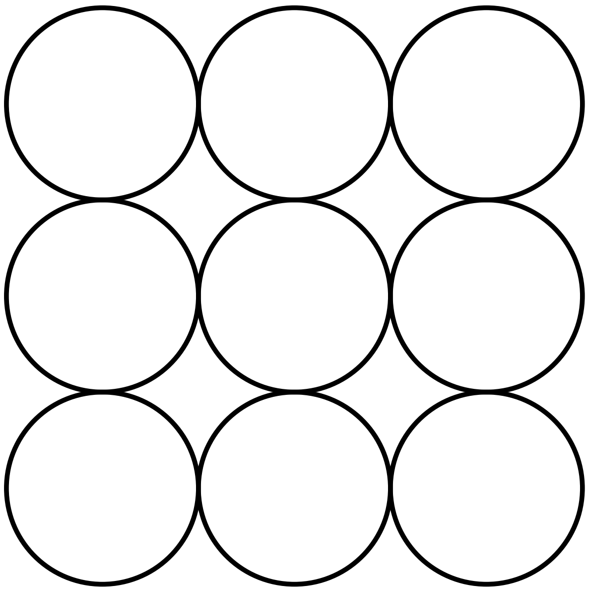
	Example Restriction Enzyme	Sequence	Compressed Bases, $\lambda = R/2$	Bits R (pins)	Dimension $D = 2R$	Number N
	MspJI	CNNR(9/13)	1.50	3.00	6.00	1
$\Rightarrow$	RlaI	VCW	1.71	3.42	6.83	1
	SgeI	CNNGNNNNNNNNN↓	2.00	4.00	8.00	5
	AspBHI	YSCNS(8/12)	2.50	5.00	10.00	1
	SgrTI	CCDS(10/14)	2.71	5.42	10.83	2
	CviJI	RG↓CY	3.00	6.00	12.00	9
	LpnPI	CCDG(10/14)	3.21	6.42	12.83	1
	M.NgoMXV	GCCHR	3.71	7.42	14.83	1
	TaqI	T↓CGA	4.00	8.00	16.00	1034
	Bsp1286I	GDGCH↓C	4.42	8.83	17.66	15
	AvaII	G↓GWCC	4.50	9.00	18.00	346
	Hin4I	(8/13)GAYNNNNNVTC(13/8)	4.71	9.42	18.83	1
$\Rightarrow$	HincII	GTY↓RAC	5.00	10.00	20.00	480
	PpuMI	RG↓GWCCY	5.50	11.00	22.00	20
$\Rightarrow$	EcoRI	G↓AATTC	6.00	12.00	24.00	1738
	PspXI	VC↓TCGAGB	6.42	12.83	25.66	1
	RsrII	CG↓GWCCG	6.50	13.00	26.00	37
	SgrAI	CR↓CCGGYG	7.00	14.00	28.00	73
	KpnBI	CAAANNNNNNRTCA	7.50	15.00	30.00	2
	SfiI	GGCCNNNN↓NGGCC	8.00	16.00	32.00	34

3802 restriction enzymes from Rich Roberts' Restriction Enzyme Database, REBASE


Number of Restriction Enzymes
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
4
8
12
16
20
24
28
32
36
40
44
48
Dimension D

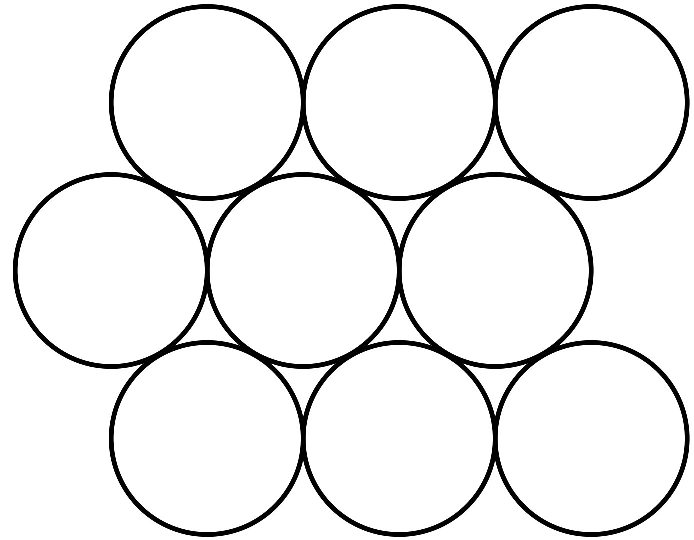
Packing in 2 Dimensions

Square Packing



$\pi r^2/(2 \times r)^2 = 79\%$
of the plane filled

Hexagonal Packing



$\pi/\sqrt{12} = 91\%$
of the plane filled

Packing in 3 Dimensions

hexagonal sphere packing

Packing in 3 Dimensions

hexagonal sphere packing

green balls reveal square packing

Packing in 3 Dimensions


hexagonal sphere packing


green balls reveal square packing


face centered cubic packing

Packing in 3 Dimensions

hexagonal sphere packing

green balls reveal square packing

face centered cubic packing

more spheres show square packing

What happens in higher dimensions?

Discovery of the Leech Lattice

Leech, John (1964), "Some sphere packings in higher space", Canadian Journal of Mathematics 16: 657-682

SOME SPHERE PACKINGS IN HIGHER SPACE

JOHN LEECH

Introduction. This paper is concerned with the packing of equal spheres in Euclidean spaces $[n]$ of $n > 8$ dimensions. To be precise, a *packing* is a distribution of spheres any two of which have at most a point of contact in common. If the centres of the spheres form a lattice, the packing is said to be a *lattice packing*. The densest lattice packings are known for spaces of up to eight dimensions **(1, 2)**, but not for any space of more than eight dimensions. Further, although non-lattice packings are known in [3] and [5] which have the same density as the densest lattice packings, none is known which has greater density than the densest lattice packings in any space of up to eight dimensions, neither, for any space of more than two dimensions, has it been shown that they do not exist.

In Part 1 the densest lattice packings in [4] and [8] are generalized to packings, not all lattice packings, in $[2^m]$, in which each sphere touches

$$(2+2)(2+2^2)(2+2^3)\ldots(2+2^m)$$

others. This gives packings in [16] in which each sphere touches 4320 others, which may be the densest in this space. For $m > 4$ the corresponding packings are unlikely to be the densest, though they seem to be the densest yet constructed.

In Part 2 some different analogies to the densest lattice packing in [8] are considered, which lead to new packings in [12] and [24]. In [12] this does not lead to any packing as dense as K_{12} **(5)**, though it leads to new co-ordinates for some known packings. In [24] a dense lattice packing is found in which each sphere touches 98256 others. Other packings in up to 23 dimensions are found as sections of this packing in [24].

In Part 3 the densities of these packings are compared with Rogers' upper bound **(10)**. This comparison is also made for the known densest lattice packings in up to eight dimensions for which it has not been made before. The numbers of spheres touched are compared with Coxeter's upper bound **(4)**. For the packings in $[2^m]$ the density and the number of spheres touched are of a much smaller order of magnitude than Rogers' and Coxeter's upper bounds as $m \to \infty$. The packings in up to 24 dimensions are closer to the upper bounds, though not so close as in from 3 to 8 dimensions, that in [8] being especially close.

Received June 26, 1963.

Leech, John (1967), "Notes on sphere packings", Canadian Journal of Mathematics 19: 251–267

NOTES ON SPHERE PACKINGS

JOHN LEECH

These notes are to supplement my paper **(4)**, and should be read in conjunction with it. Both are divided into three parts, and in these notes the section numbers have a further digit added; thus §1.41 here supplements §1.4 of **(4)**. References by section numbers are always to **(4)** or to the present notes, but references to other papers are numbered independently.

The principal results of these notes are the following. New sphere packings are given in $[2^m]$, $m \geqslant 6$, and in [24], which are twice as dense as those of §§1.6, 2.3. Others are given in $[2^m]$, $m \geqslant 5$, with the same density as those of §1.6, but in which each sphere touches fewer other spheres than in the earlier packings. Acknowledgment is made of lattice packings in $[2^m]$ given by Barnes and Wall **(2)** in anticipation of §1.6 (in which the packings were lattice packings only for $m \leqslant 6$). Denser packings than those of §2.4 are given for [11], as a section of K_{12}, and for [22] and [23], as sections of the new packing in [24]; that in [11] is due to Barnes **(1)**. A proof is given that the packing in [16] (§1.2) is a section of those in [24] (§2.3 or §2.31). Table I supersedes those of **(4)**, giving values of Rogers' bound for the density of packings, and Coxeter's bound for the number of spheres that may touch any one, for spaces of up to 24 dimensions, in comparison with the best figures achieved by known packings in these spaces.

1. Packings in $[2^m]$

1.31. This section and the next two are preliminary to the constructions of sphere packings in §§1.61–1.63 below. It was shown in §1.3 that any two rows of 2^m binary digits having k-parity differ in at least 2^k places. It follows that any row having exactly k-parity differs in at least 2^k places from any row having $(k+1)$-parity. We now investigate whether there exist rows differing in more than 2^k places from every row having $(k+1)$-parity.

For $k = 0$ there is clearly no such row, since every row either has 1-parity (simple even parity) or can be altered in any one place so as to give it 1-parity. There is also no such row for $k = 1$. If a row does not have 1-parity, then it can be altered in one place to give it 2-parity; we reverse the digit whose position, when expressed as an integer in the binary scale, has 1's in just those positions whose significance corresponds to those binary constituent rows (§1.3) of the given row which do not have 1-parity. If a row has 1-parity, we can alter it in two places so as to give it 2-parity; one may be chosen arbitrarily, and the other is then uniquely determined as above for a row not having 1-parity.

Received June 30, 1965. Revised version received August 5, 1966.

Discovery of the Leech Lattice

Leech, John (1964), "Some sphere packings in higher space", Canadian Journal of Mathematics 16: 657-682

SOME SPHERE PACKINGS IN HIGHER SPACE

JOHN LEECH

Introduction. This paper is concerned with the packing of equal spheres in Euclidean spaces $[n]$ of $n > 8$ dimensions. To be precise, a *packing* is a distribution of spheres any two of which have at most a point of contact in common. If the centres of the spheres form a lattice, the packing is said to be a *lattice packing*. The densest lattice packings are known for spaces of up to eight dimensions **(1, 2)**, but not for any space of more than eight dimensions. Further, although non-lattice packings are known in [3] and [5] which have the same density as the densest lattice packings, none is known which has greater density than the densest lattice packings in any space of up to eight dimensions, neither, for any space of more than two dimensions, has it been shown that they do not exist.

In Part 1 the densest lattice packings in [4] and [8] are generalized to packings, not all lattice packings, in $[2^m]$, in which each sphere touches

$$(2 + 2)(2 + 2^2)(2 + 2^3) \ldots$$

others. This gives packings in [16] in which ea… which may be the densest in this space. For m … are unlikely to be the densest, though they … constructed.

In Part 2 some different analogies to the densest lattice packing in [8] are considered, which lead to new packings in [12] and [24]. In [12] this does not lead to any packing as dense as K_{12} **(5)**, though it leads to new co-ordinates for some known packings. In [24] a dense lattice packing is found in which each sphere touches 98256 others. Other packings in up to 23 dimensions are found as sections of this packing in [24].

In Part 3 the densities of these packings are compared with Rogers' upper bound **(10)**. This comparison is also made for the known densest lattice packings in up to eight dimensions for which it has not been made before. The numbers of spheres touched are compared with Coxeter's upper bound **(4)**. For the packings in $[2^m]$ the density and the number of spheres touched are of a much smaller order of magnitude than Rogers' and Coxeter's upper bounds as $m \to \infty$. The packings in up to 24 dimensions are closer to the upper bounds, though not so close as in from 3 to 8 dimensions, that in [8] being especially close.

Received June 26, 1963.

The Best Sphere Packing is in 24 dimensions!

Leech, John (1967), "Notes on sphere packings", Canadian Journal of Mathematics 19: 251–267

NOTES ON SPHERE PACKINGS

JOHN LEECH

These notes are to supplement my paper **(4)**, and should be read in conjunction with it. Both are divided into three parts, and in these notes the section numbers have a further digit added; thus §1.41 here supplements §1.4 of **(4)**. References by section numbers are always to **(4)** or to the present notes, but references to other papers are numbered independently.

The principal results of these notes are the following. New sphere packings are given in $[2^m]$, $m \geqslant 6$, and in [24], which are twice as dense as those of §§1.6, 2.3. Others are given in $[2^m]$, $m \geqslant 5$, with the same density as those of §1.6, but in which each sphere touches fewer other spheres than in the earlier packings. Acknowledgment is made of lattice packings in $[2^m]$ given by Barnes and Wall **(2)** in anticipation of §1.6 (in which the packings were lattice packings only for $m \leqslant 6$). Denser packings than those of §2.4 are given for [11], as … and for [22] and [23], as sections of the new packing in [24]; …nes **(1)**. A proof is given that the packing in [16] … in [24] (§2.3 or §2.31). Table I supersedes those of …rs' bound for the density of packings, and Coxeter's … spheres that may touch any one, for spaces of up to 24 dimensions, in …ison with the best figures achieved by known packings in these spaces.

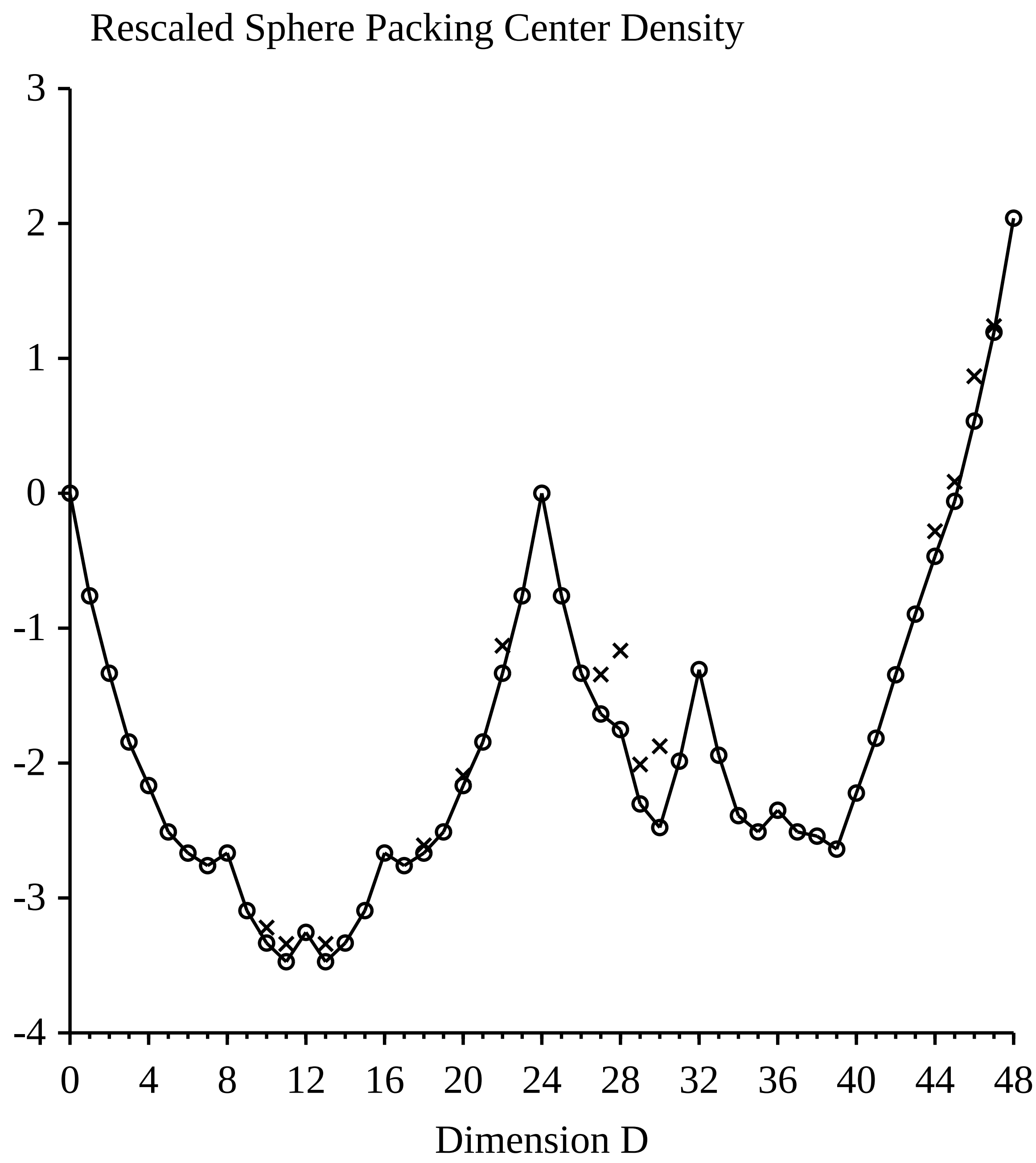
1. Packings in $[2^m]$

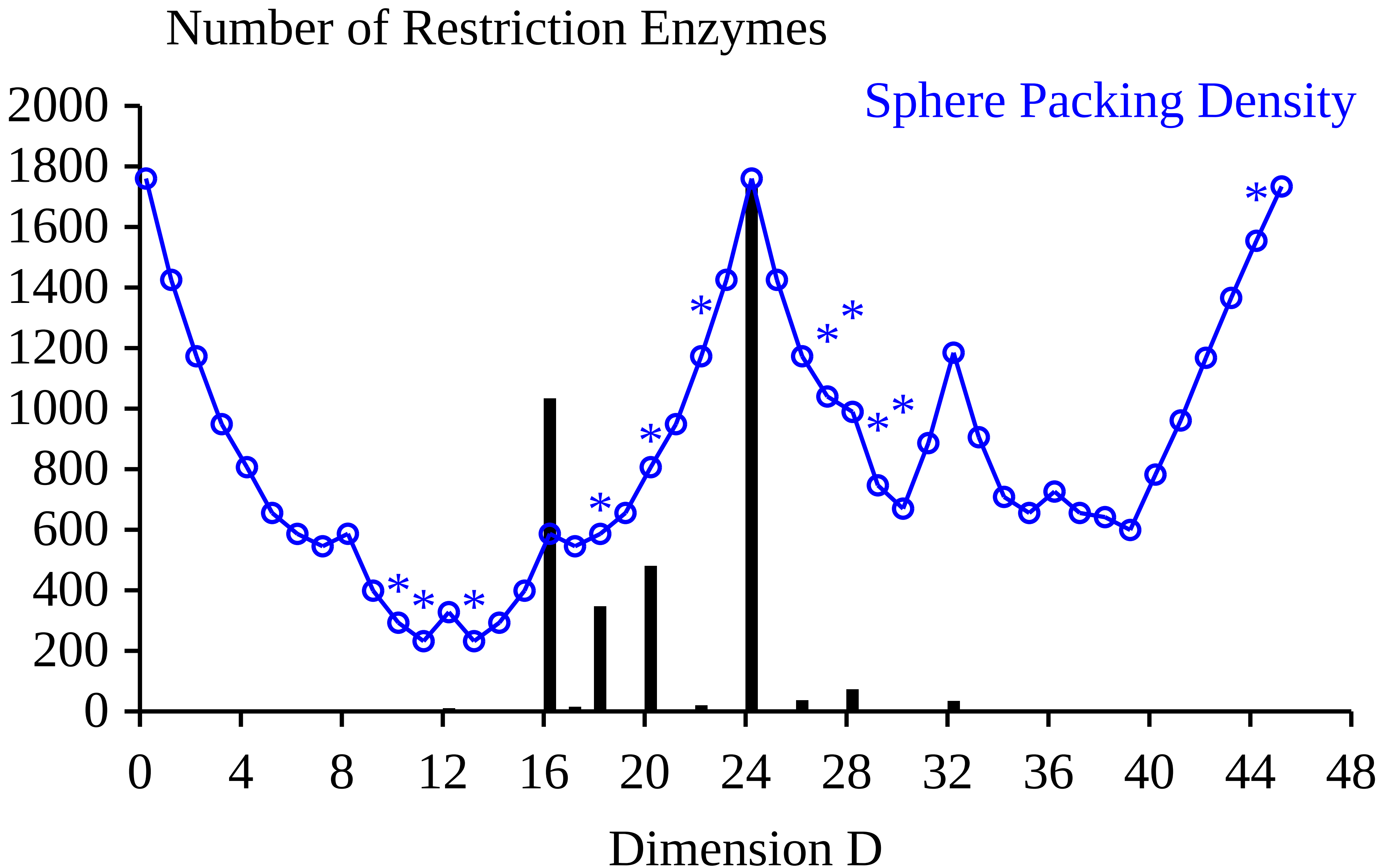
1.31. This section and the next two are preliminary to the constructions of sphere packings in §§1.61–1.63 below. It was shown in §1.3 that any two rows of 2^m binary digits having k-parity differ in at least 2^k places. It follows that any row having exactly k-parity differs in at least 2^k places from any row having $(k + 1)$-parity. We now investigate whether there exist rows differing in more than 2^k places from every row having $(k + 1)$-parity.

For $k = 0$ there is clearly no such row, since every row either has 1-parity (simple even parity) or can be altered in any one place so as to give it 1-parity. There is also no such row for $k = 1$. If a row does not have 1-parity, then it can be altered in one place to give it 2-parity; we reverse the digit whose position, when expressed as an integer in the binary scale, has 1's in just those positions whose significance corresponds to those binary constituent rows (§1.3) of the given row which do not have 1-parity. If a row has 1-parity, we can alter it in two places so as to give it 2-parity; one may be chosen arbitrarily, and the other is then uniquely determined as above for a row not having 1-parity.

Received June 30, 1965. Revised version received August 5, 1966.

Best Sphere Packing




Number of Restriction Enzymes
Sphere Packing Density
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
4
8
12
16
20
24
28
32
36
40
44
48
Dimension D

Leech Lattice Modem

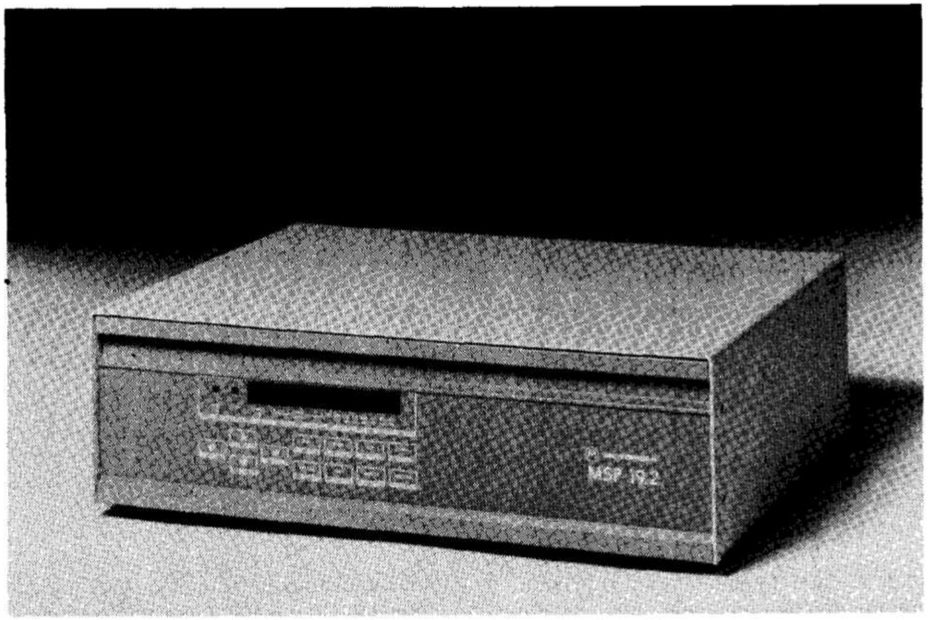


Fig. 1. A hardware prototype of the 19 200 bit/s Leech modem.

Leech Lattice Modem

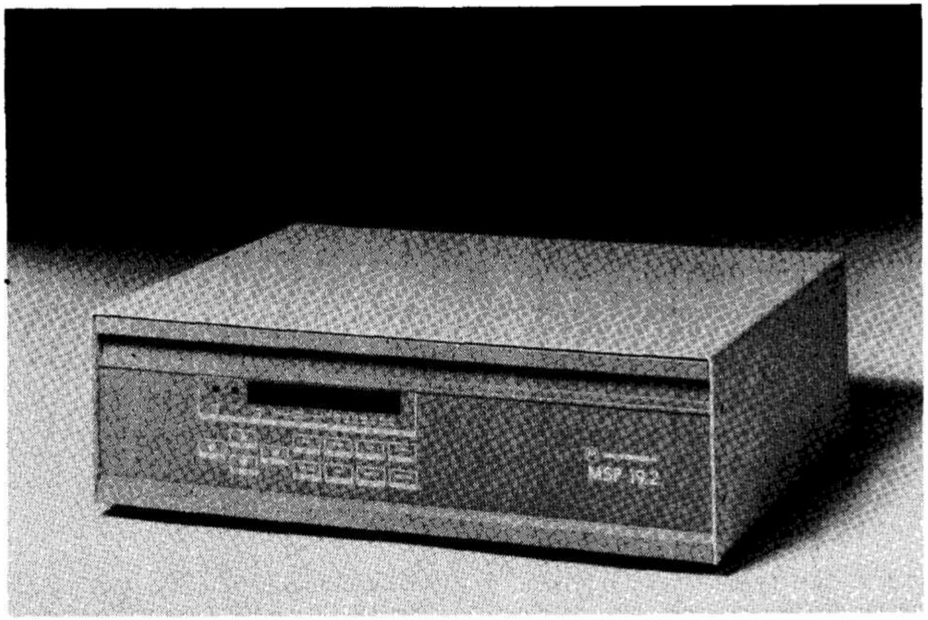


Fig. 1. A hardware prototype of the 19 200 bit/s Leech modem.

- Leech Lattice 19,200 bit/sec modem built by Motorola

Lang & Longstaff, IEEE J. on Selected Areas in Comm. 7:968–973, 1989

Leech Lattice Modem

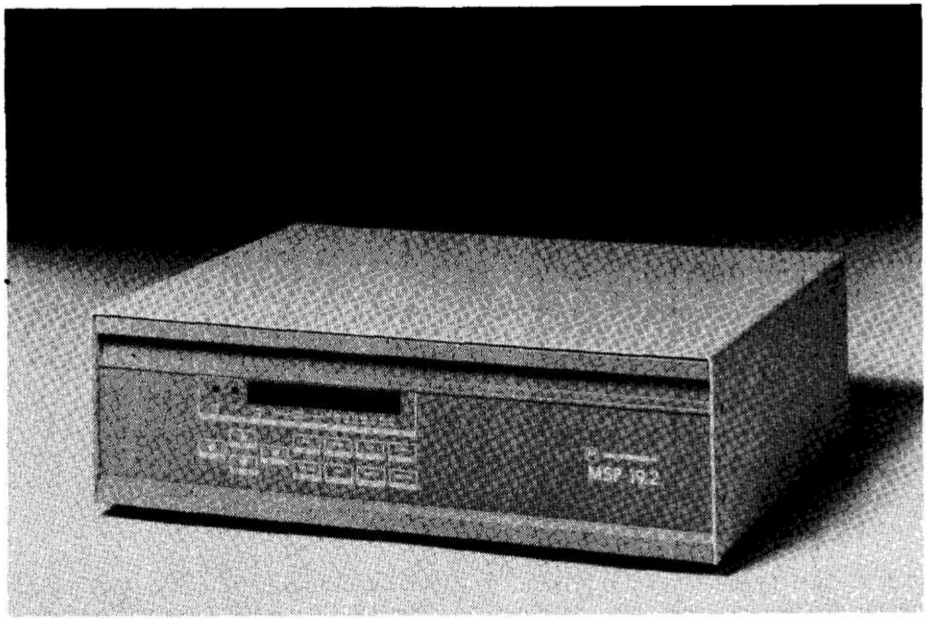
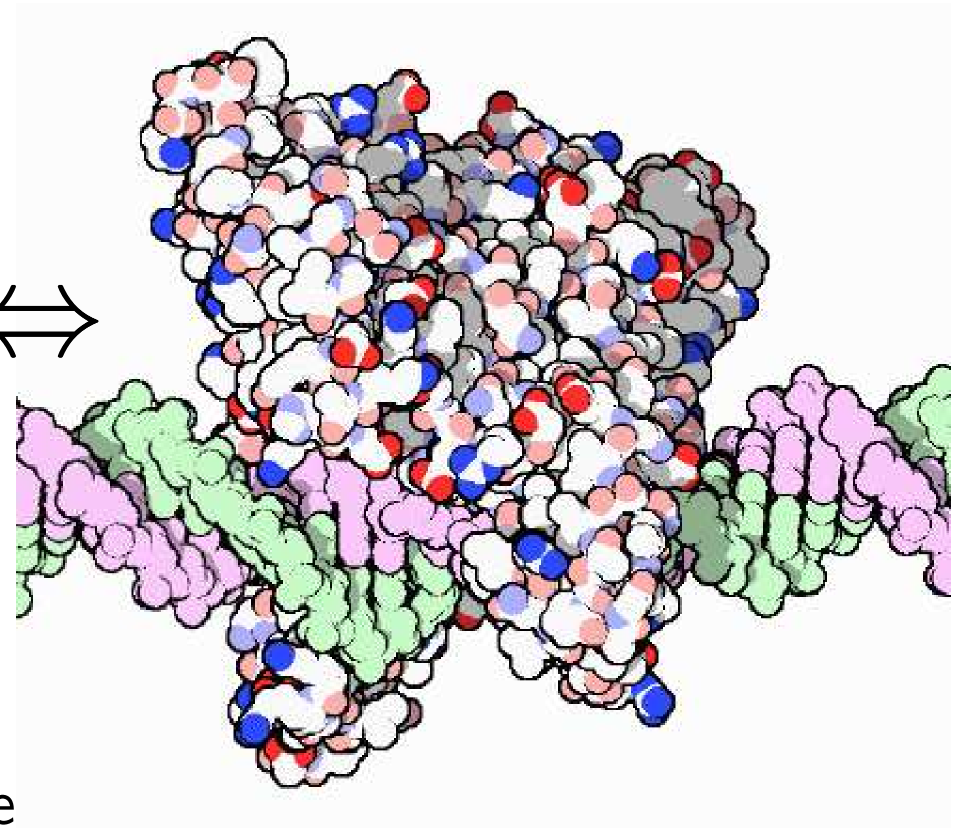


Fig. 1. A hardware prototype of the 19 200 bit/s Leech modem.

$\Leftrightarrow$



- Leech Lattice 19,200 bit/sec mode
- EcoRI decodes one of 4096 patterns probably using the Leech Lattice

Lang & Longstaff, IEEE J. on Selected Areas in Comm. 7:968–973, 1989

Leech Lattice Modem

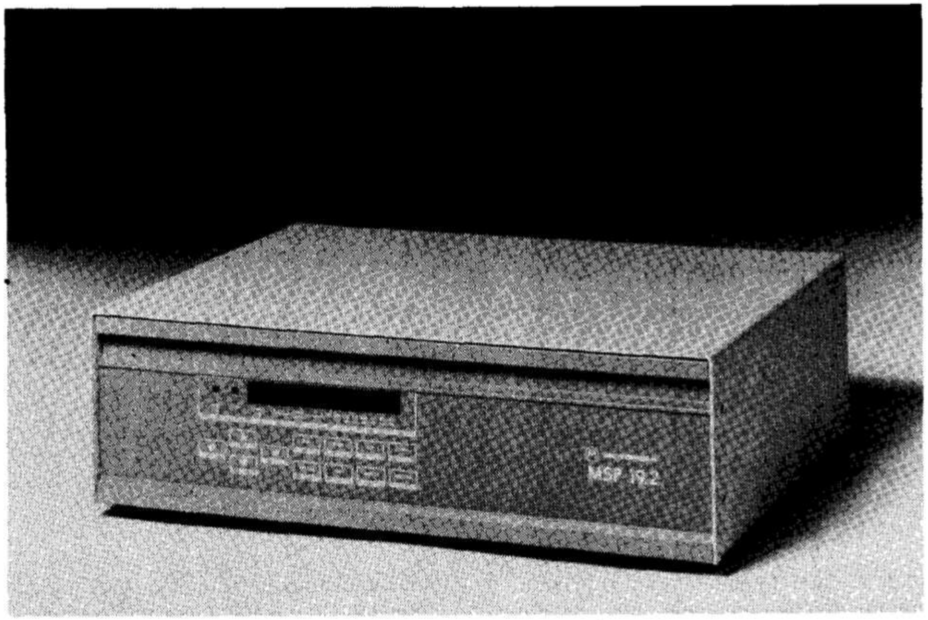
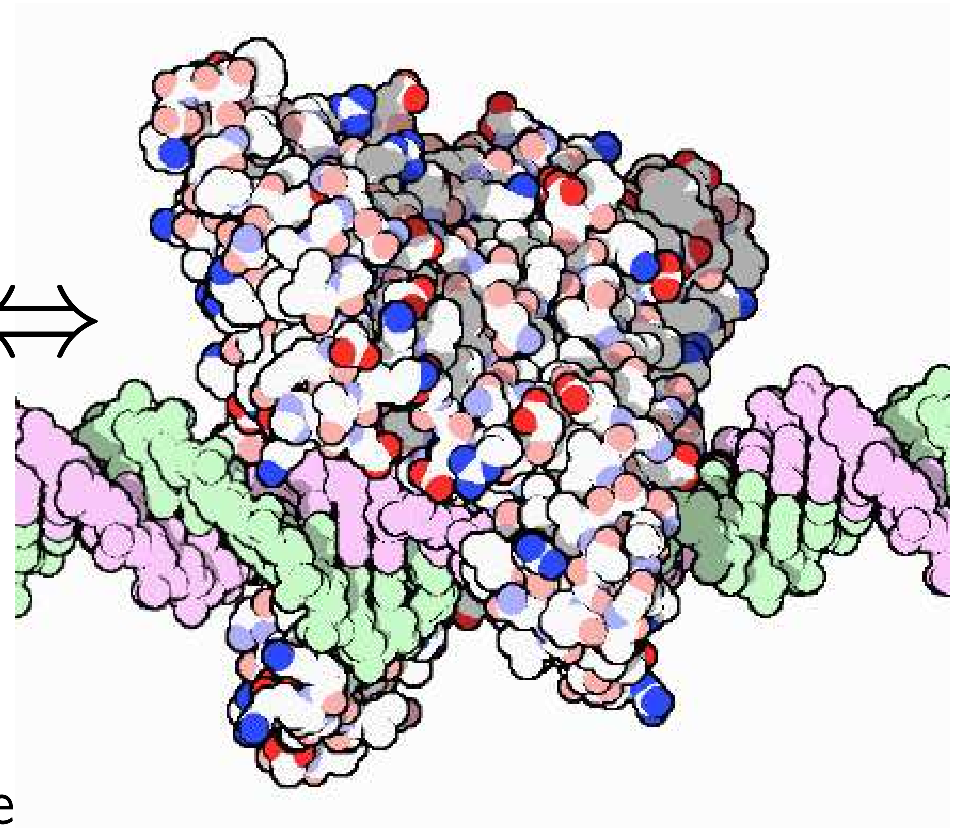


Fig. 1. A hardware prototype of the 19 200 bit/s Leech modem.

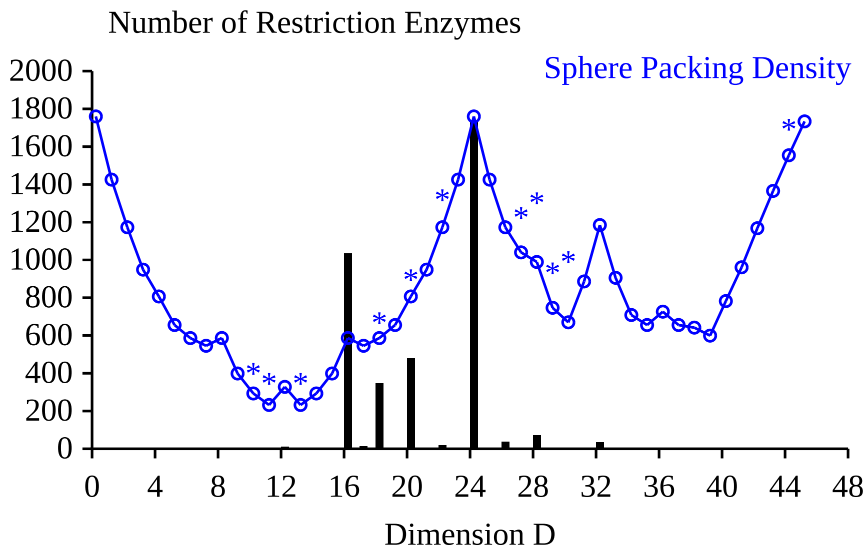
$\Leftrightarrow$

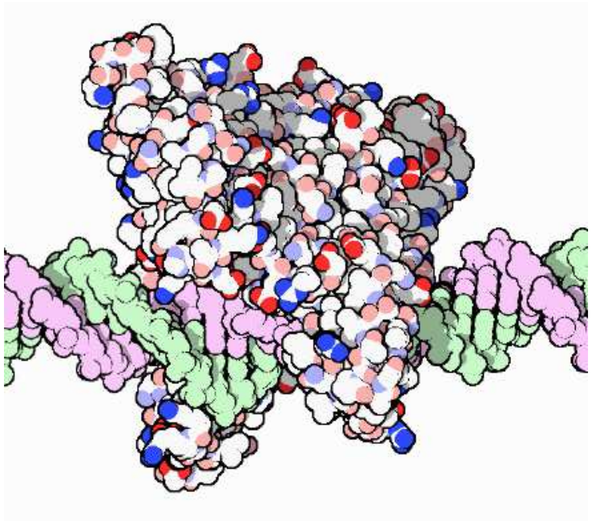


- Leech Lattice 19,200 bit/sec mode
- EcoRI decodes one of 4096 patterns probably using the Leech Lattice
- Single molecules could be used to build a modem, in theory.

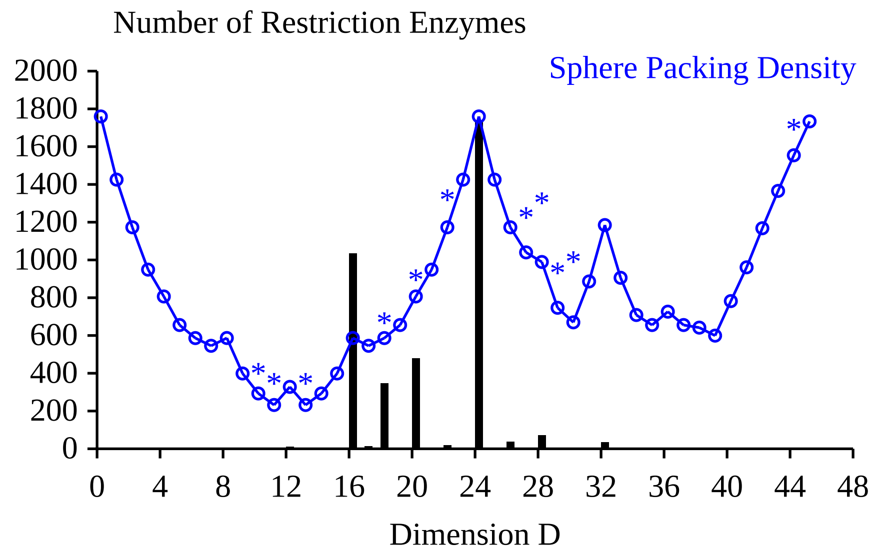
Lang & Longstaff, IEEE J. on Selected Areas in Comm. 7:968–973, 1989

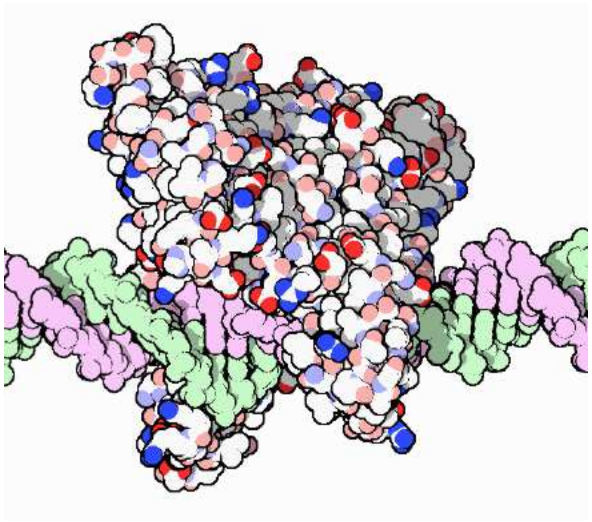
Conclusions





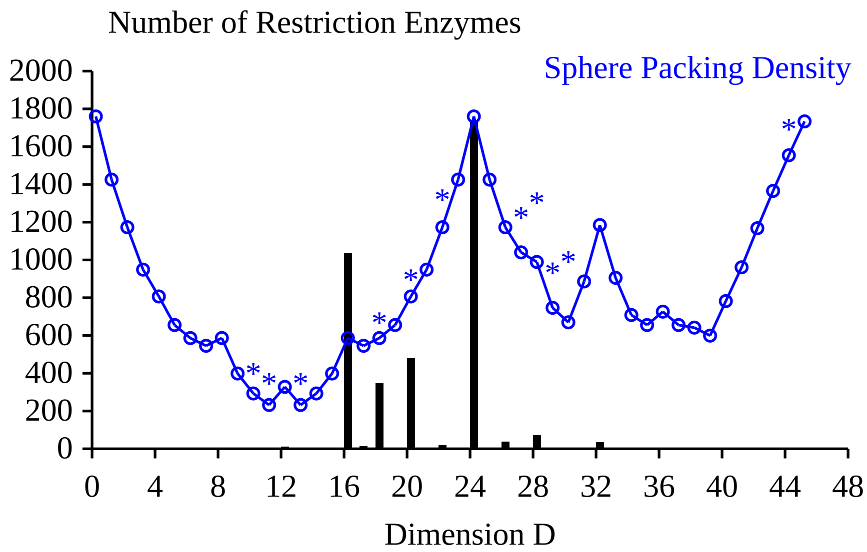
Conclusions

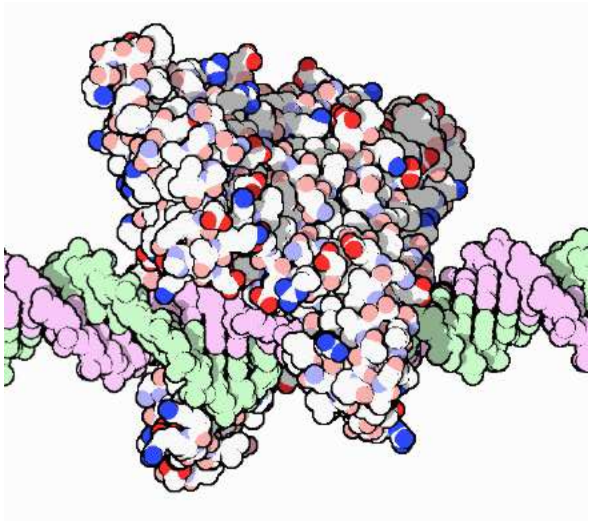




- 6 base cutting restriction enzymes work in 24 dimensions
  4 base cutting restriction enzymes work in 16 dimensions
  Both have good lattice packings.

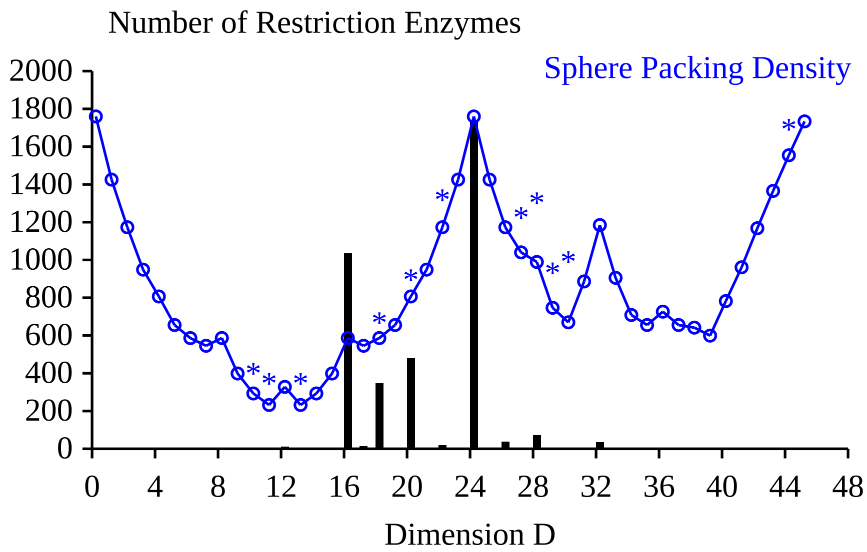
Conclusions

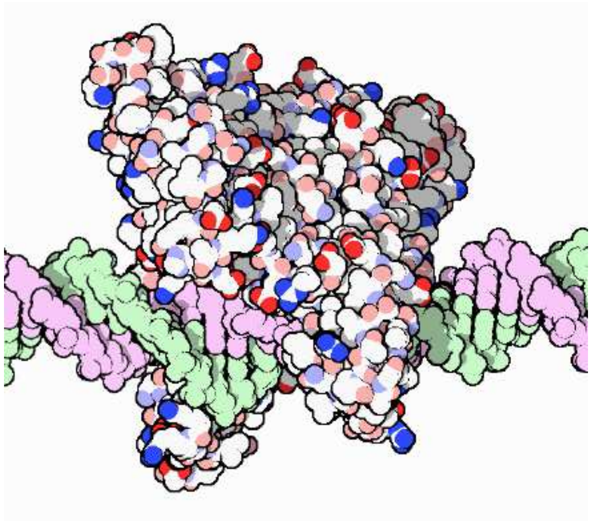




- 6 base cutting restriction enzymes work in 24 dimensions
  4 base cutting restriction enzymes work in 16 dimensions
  Both have good lattice packings.
- Restriction Enzymes probably use the 24 dimensional Leech Lattice

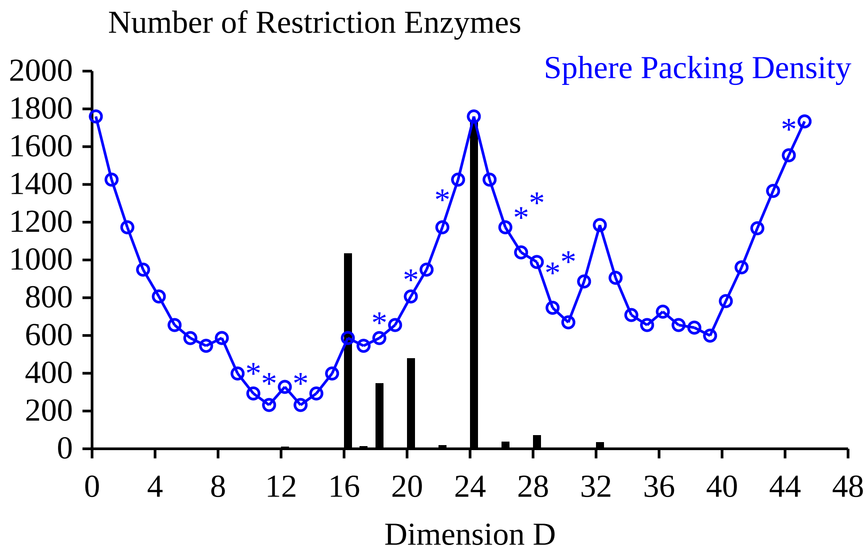
Conclusions

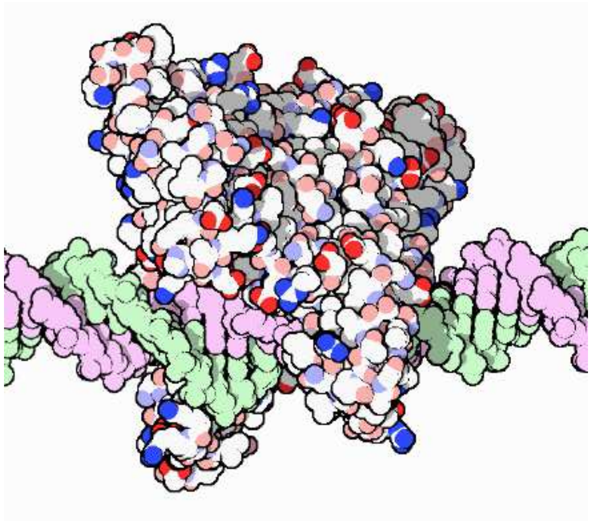




- 6 base cutting restriction enzymes work in 24 dimensions
  4 base cutting restriction enzymes work in 16 dimensions
  Both have good lattice packings.
- Restriction Enzymes probably use the 24 dimensional Leech Lattice
- Good packing can explain why restriction enzymes are so precise

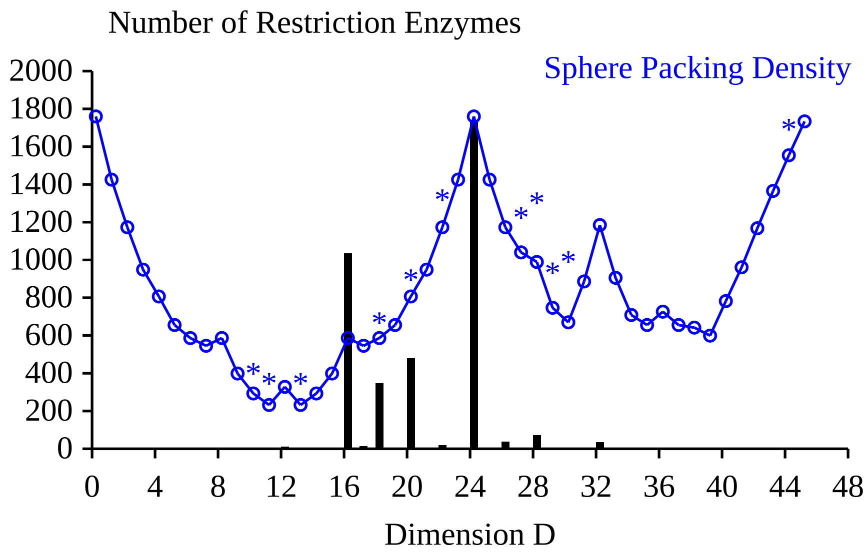
Conclusions

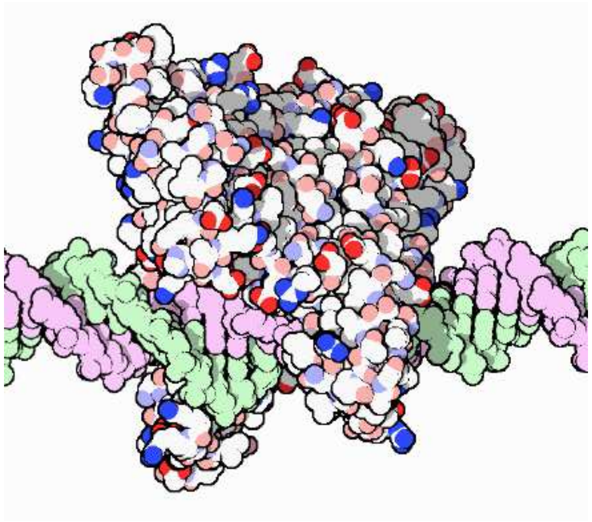




- 6 base cutting restriction enzymes work in 24 dimensions
  4 base cutting restriction enzymes work in 16 dimensions
  Both have good lattice packings.
- Restriction Enzymes probably use the 24 dimensional Leech Lattice
- Good packing can explain why restriction enzymes are so precise
- Dimensionality is a clue to mechanism

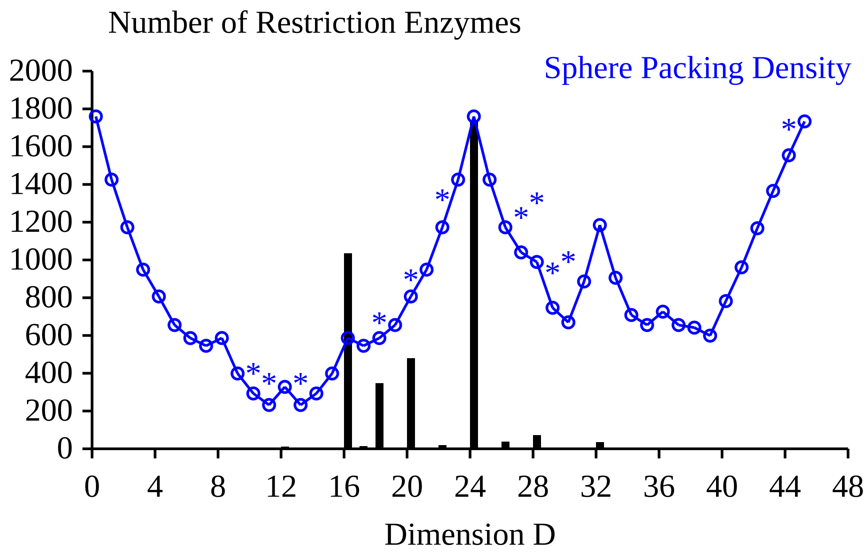
Conclusions

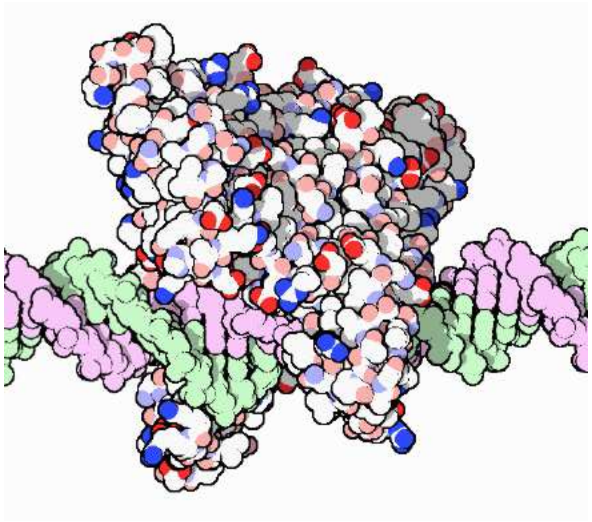




- 6 base cutting restriction enzymes work in 24 dimensions
  4 base cutting restriction enzymes work in 16 dimensions
  Both have good lattice packings.
- Restriction Enzymes probably use the 24 dimensional Leech Lattice
- Good packing can explain why restriction enzymes are so precise
- Dimensionality is a clue to mechanism
- Leech Lattice coding is used for modern communications

Conclusions





- 6 base cutting restriction enzymes work in 24 dimensions
  4 base cutting restriction enzymes work in 16 dimensions
  Both have good lattice packings.
- Restriction Enzymes probably use the 24 dimensional Leech Lattice
- Good packing can explain why restriction enzymes are so precise
- Dimensionality is a clue to mechanism
- Leech Lattice coding is used for modern communications
- Single molecules probably can be used to do communications!

Acknowledgements

- N. J. A. Sloane
- Rich Roberts
- Soren Brunak and Technical University of Denmark
- Michael Smith, NCI
- Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.
- At Queen Mary, University of London Vishnu Jejjala was supported by the UK Science and Technology Facilities Council under grant ST/G000565/1.
- Vishnu Jejjala, School of Physics, University of the Witwatersrand, Johannesburg, South Africa.
  Support: South African Research Chairs Initiative of Department of Science and Technology and National Research Foundation.

Version

version = 1.62 of latticetalk.tex 2020 Jan 21