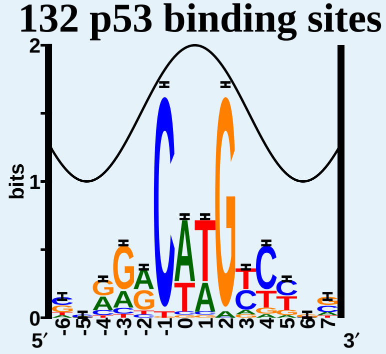
Why is the Genetic Code Degenerate?

Thomas D. Schneider, Ph.D.

Molecular Information Theory Group
Center for Cancer Research
Gene Regulation and Chromosome Biology Laboratory
National Cancer Institute
National Institutes of Health
Frederick, MD


132 p53 binding sites

The Genetic Code is Degenerate

First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

The Genetic Code is Degenerate

First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code
translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

The Genetic Code is Degenerate

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

The Genetic Code is Degenerate

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

The Genetic Code is Degenerate

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

The Genetic Code is Degenerate

Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code
translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

$4 \times 4 \times 4 = 4^3 = 64$ codons
BUT only 20 amino acids

The Genetic Code is Degenerate

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

$4 \times 4 \times 4 = 4^3 = 64$ codons
BUT only 20 amino acids

Where we are going:

> The Genetic Code is degenerate because it has distinct states.

The Genetic Code is Degenerate

Second base in codon

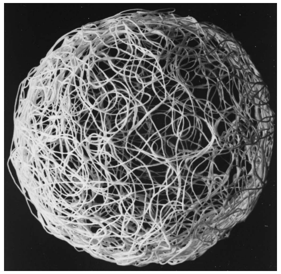
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

$4 \times 4 \times 4 = 4^3 = 64$ codons
BUT only 20 amino acids

Where we are going:

The Genetic Code
is degenerate
because it has
distinct states.



The Genetic Code is Degenerate

Second base in codon

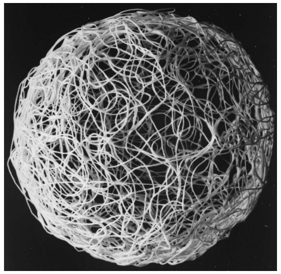
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code translates:
3 nucleotides in RNA
(a codon)
to one amino acid in a protein

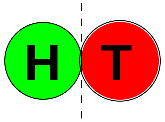
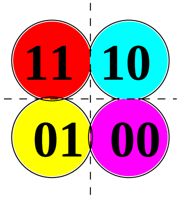
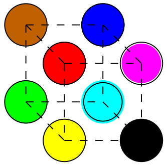
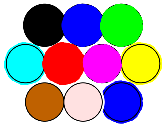
$4 \times 4 \times 4 = 4^3 = 64$ codons
BUT only 20 amino acids

Where we are going:

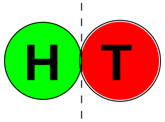
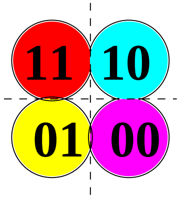
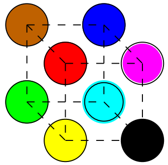
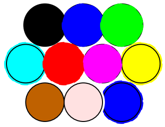
The Genetic Code
is degenerate
because it has
distinct states.



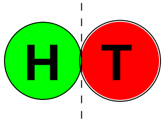
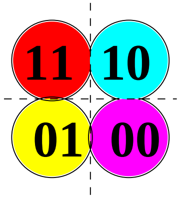
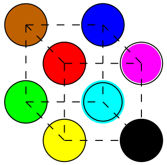
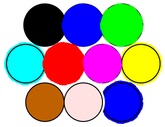
Information Theory: One-Minute Lesson

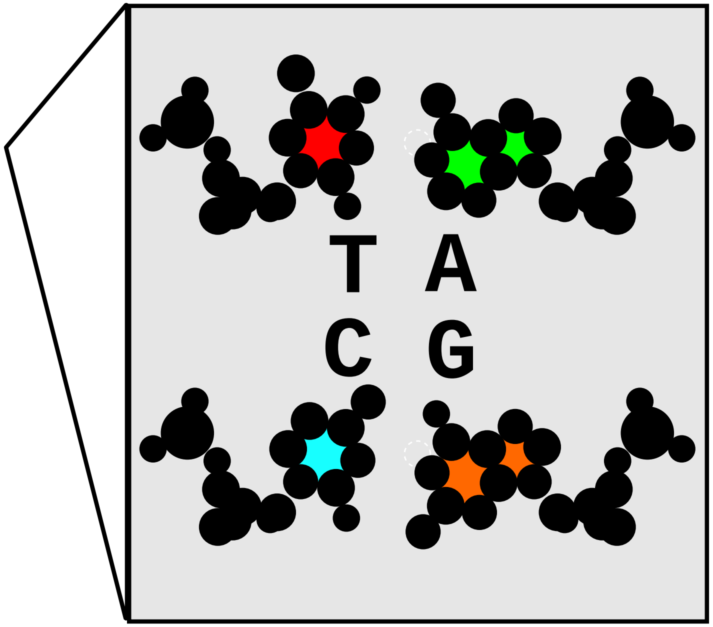
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

Information Theory: One-Minute Lesson

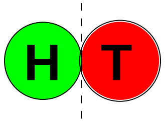
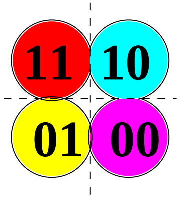
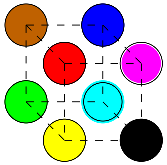
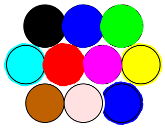
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

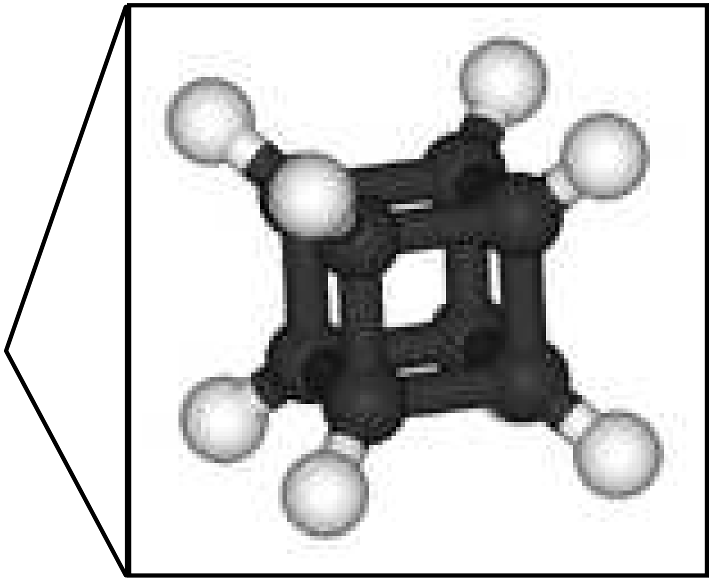
Information Theory: One-Minute Lesson

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

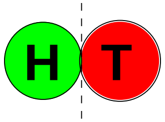
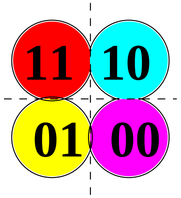
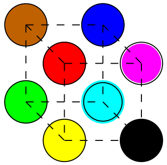
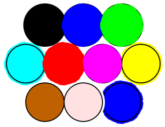


Information Theory: One-Minute Lesson

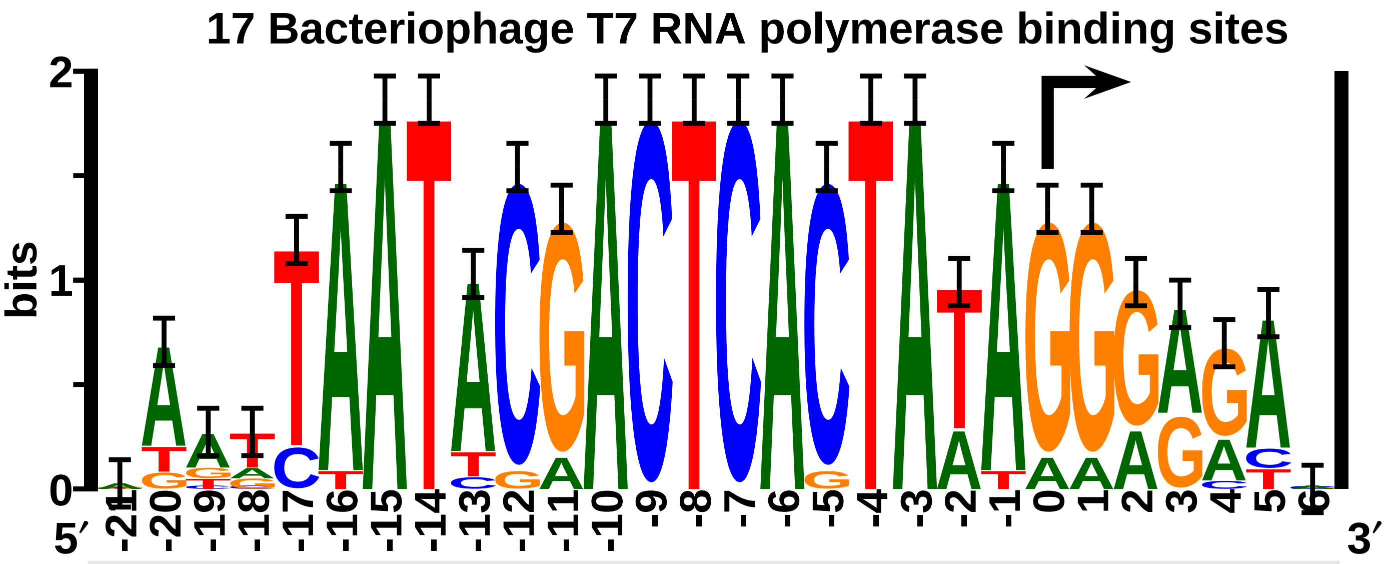
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	



Information Theory: One-Minute Lesson

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

Sequence Logo



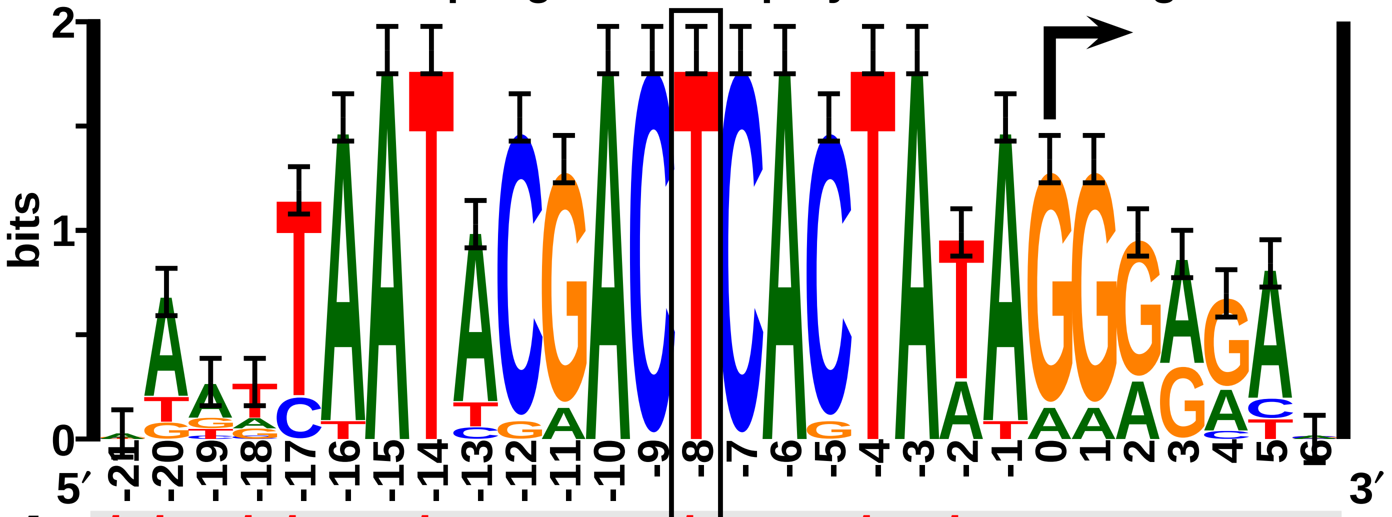
Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

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

6 of 17 sites

Sequence Logo

17 Bacteriophage T7 RNA polymerase binding sites



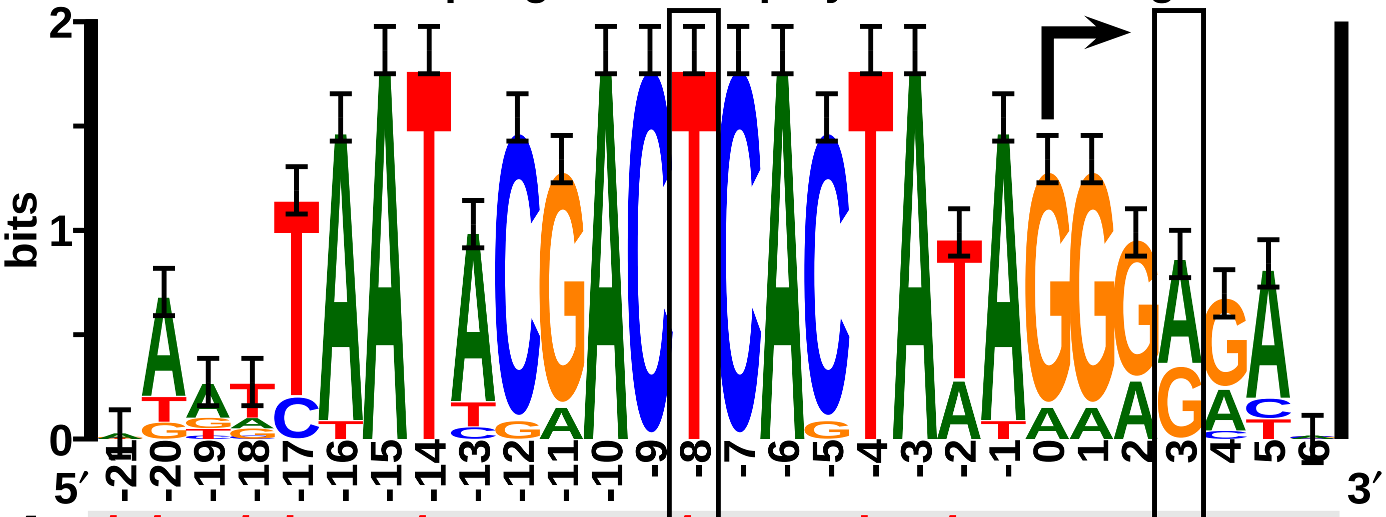
Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

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

6 of 17 sites

Sequence Logo

17 Bacteriophage T7 RNA polymerase binding sites

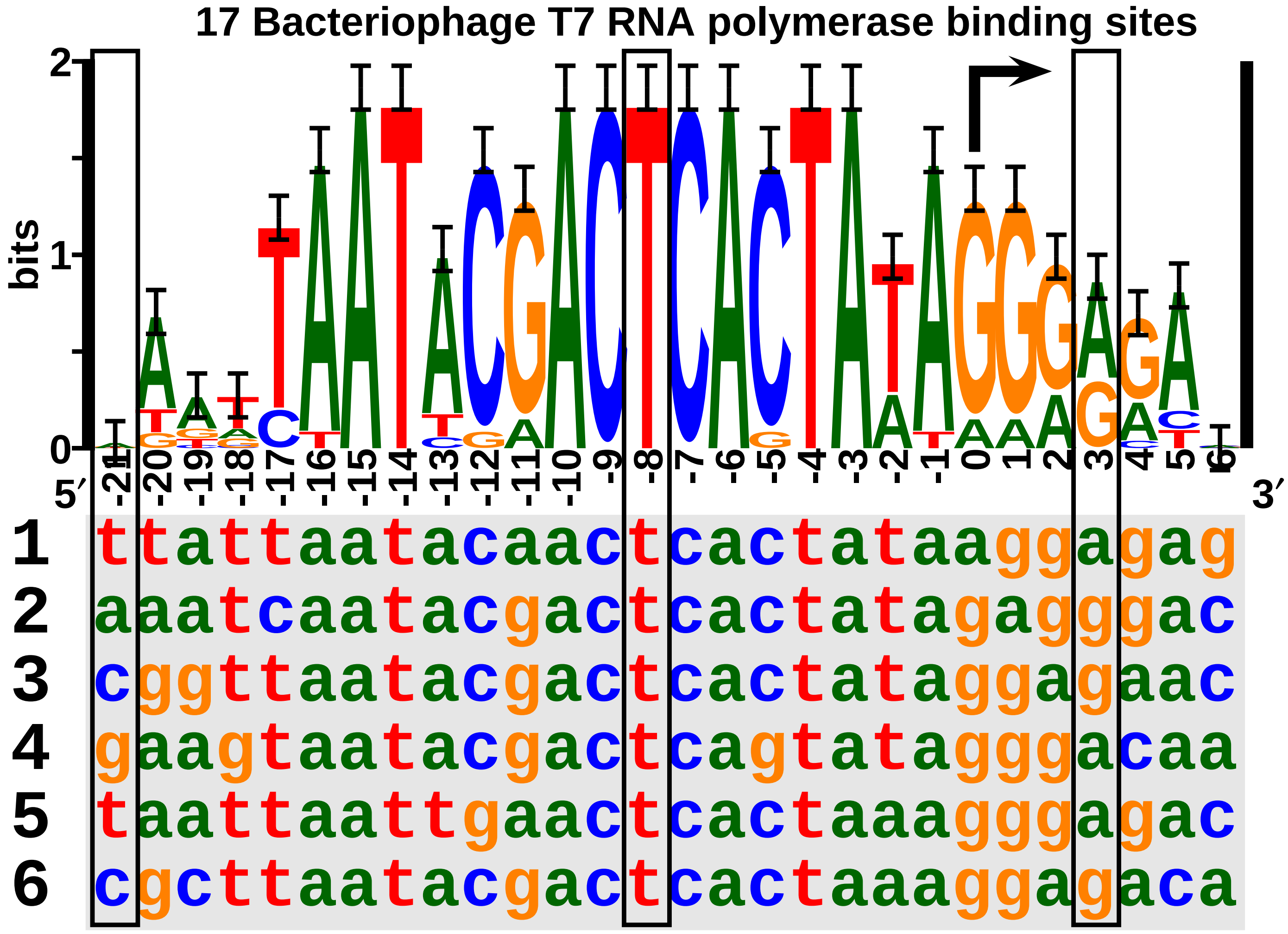


Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

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

6 of 17 sites

Sequence Logo



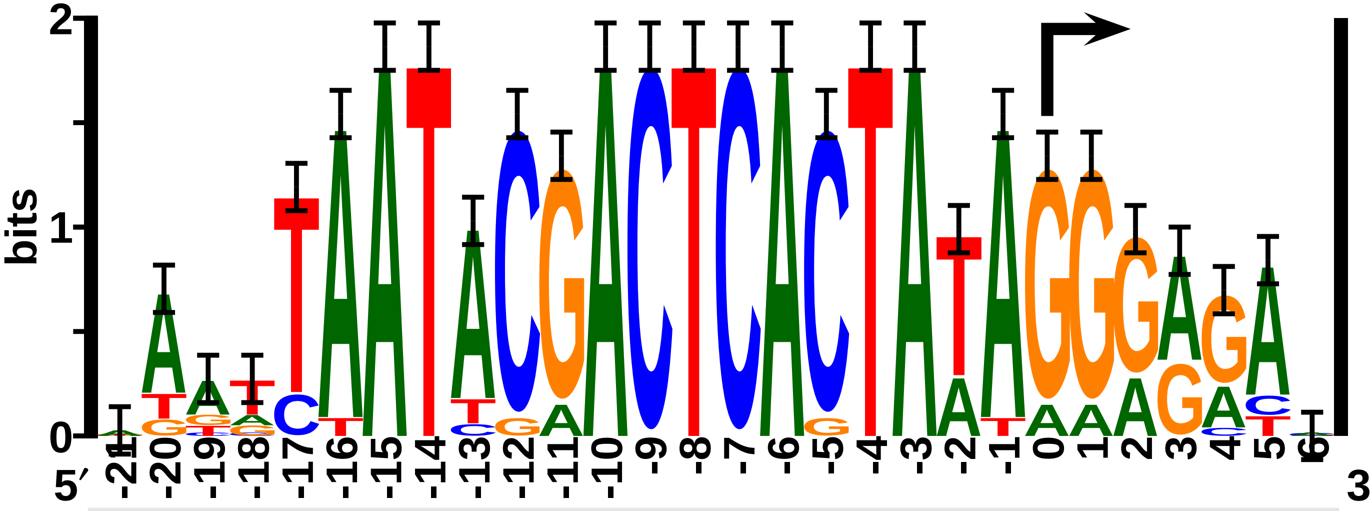
1 ttattaatacaactcactataaggagag
2 aaatcaatacgactcactatagagggac
3 cggttaatacgactcactataggagaac
4 gaagtaatacgactcagtatagggacaa
5 taattaattgaactcactaaagggagac
6 cgcttaatacgactcactaaaggagaca

6 of 17 sites

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

Sequence Logo and Sequence Walker

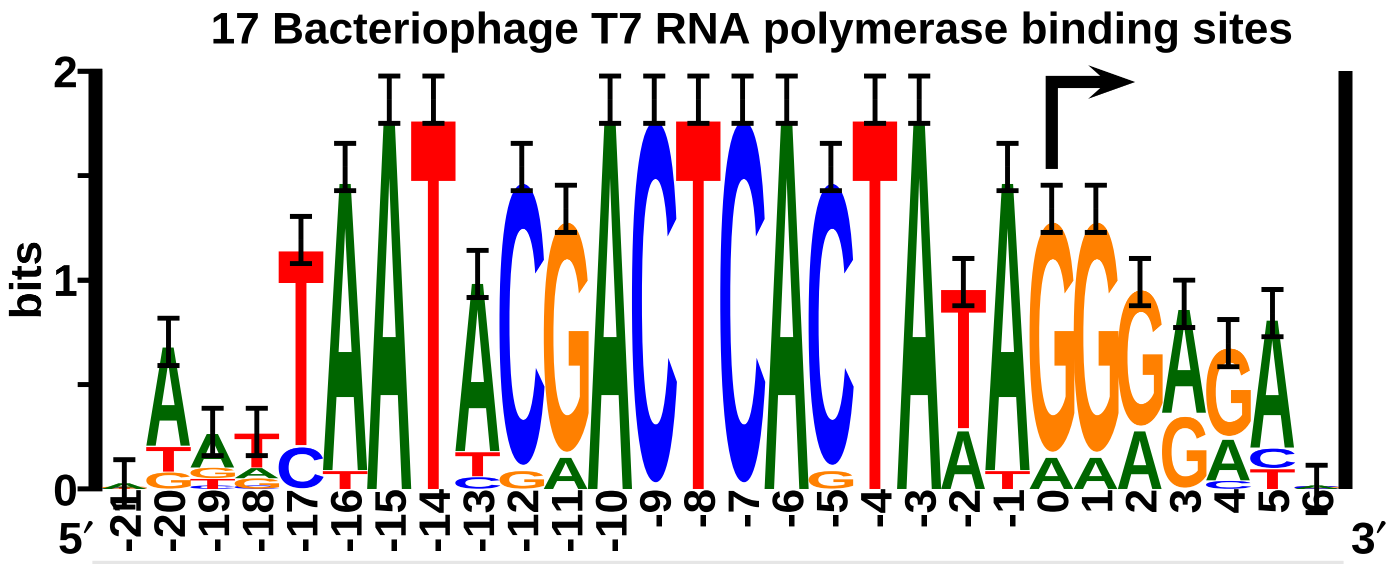
17 Bacteriophage T7 RNA polymerase binding sites



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

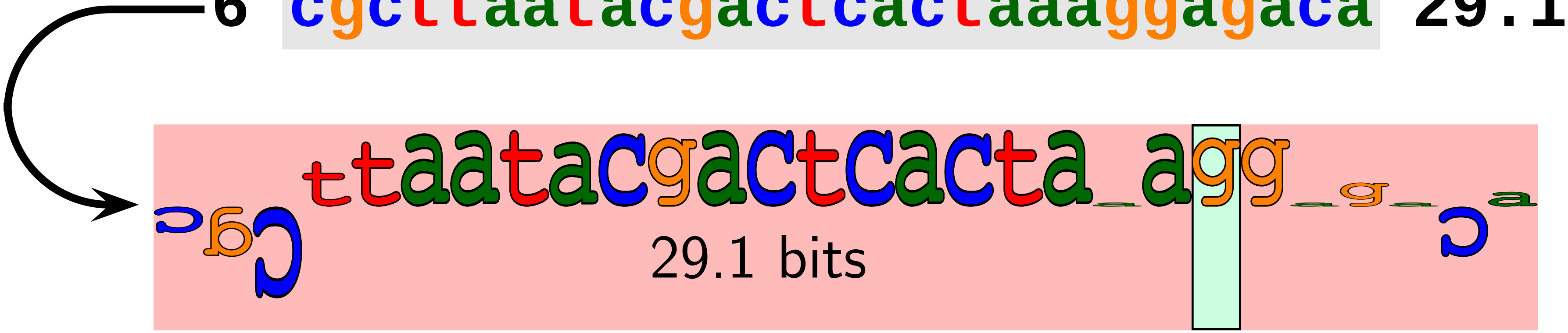
		Bits
1	ttattaatacaactcactataaggagag	33.3
2	aaatcaatacgactcactatagagggac	37.4
3	cggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaagggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1

Sequence Logo and Sequence Walker



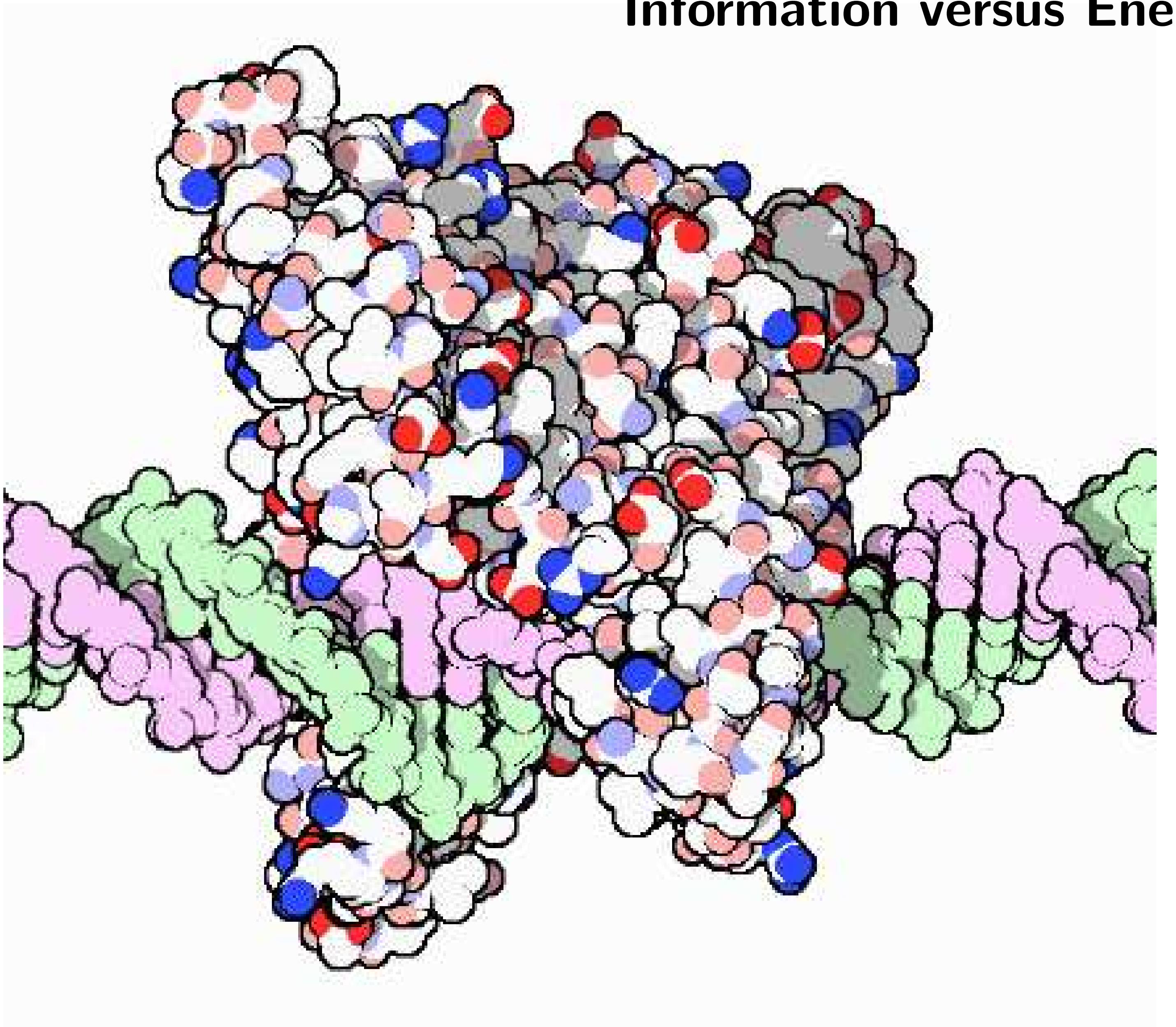
Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

		Bits
1	ttattaatacaactcactataaggagag	33.3
2	aaatcaatacgactcactatagagggac	37.4
3	cggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaagggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1



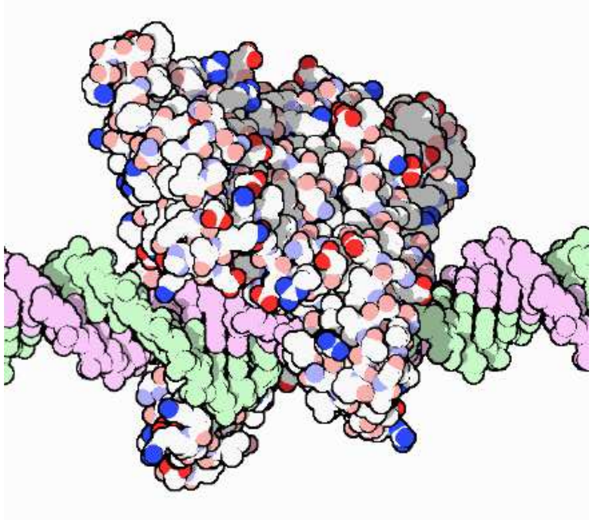
Sequence
Walker
Patent
5,867,402

Information versus Energy



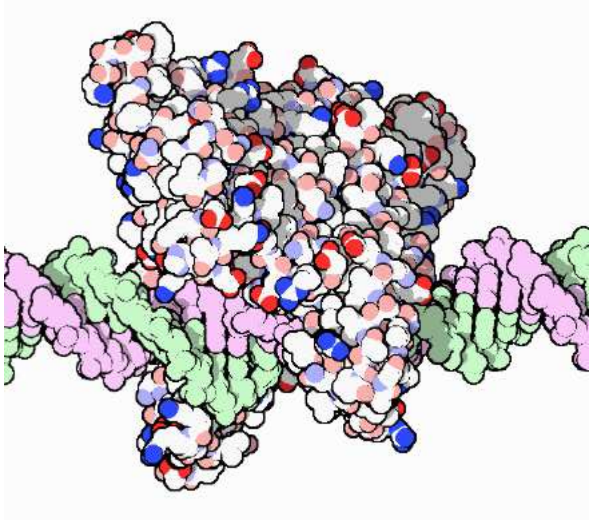
Information of EcoRI DNA Binding

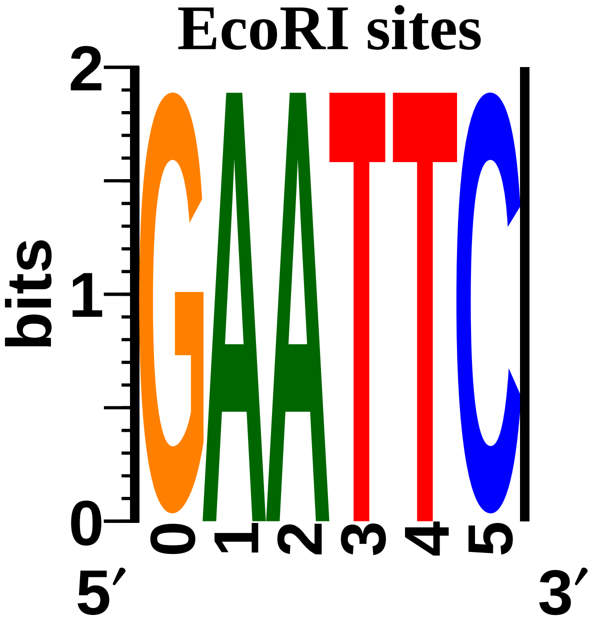
- EcoRI - restriction enzyme



Information of EcoRI DNA Binding

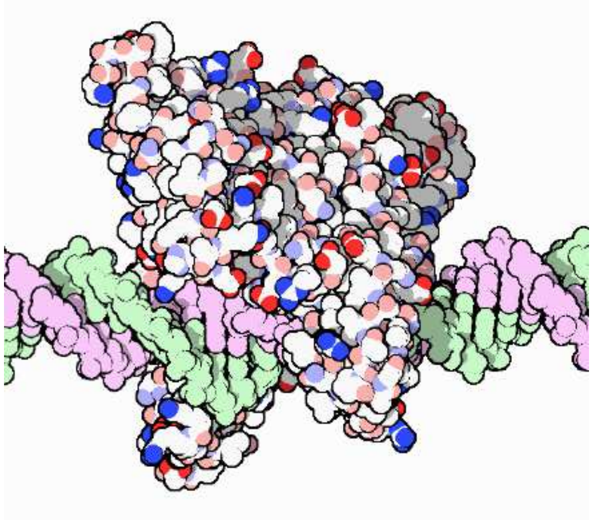
- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$

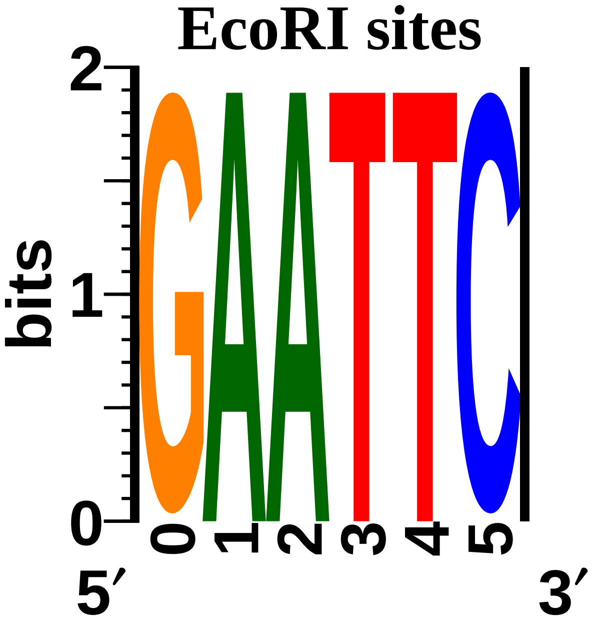




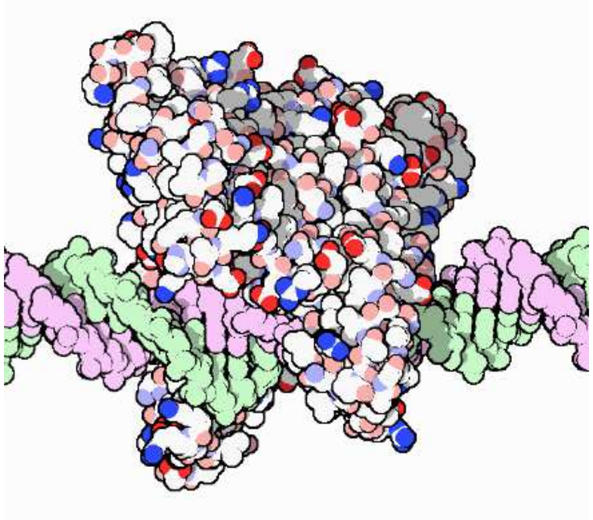
Information of EcoRI DNA Binding

- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$
- $4^6 = 4096$ possible DNA hexamers

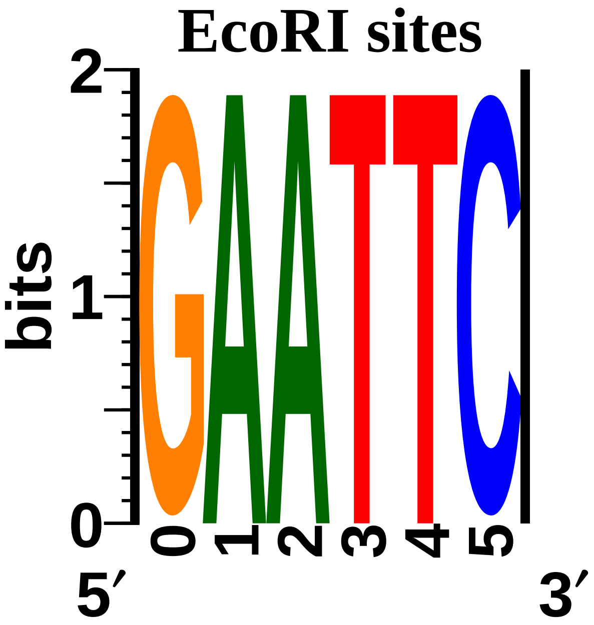




Information of EcoRI DNA Binding



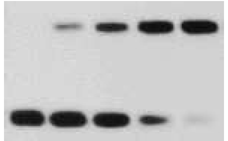
- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$
- $4^6 = 4096$ possible DNA hexamers
- information required:
  $\log_2 4096 = 12$ bits
  or
  6 bases $\times$ 2 bits per base $=$ $\boxed{\text{12 bits}}$



Energy Dissipation by EcoRI

- Measured specific binding constant:

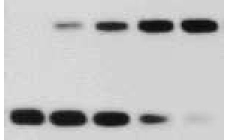
$$K_{spec} = 1.6 \times 10^5$$



Energy Dissipation by EcoRI

- Measured specific binding constant:

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



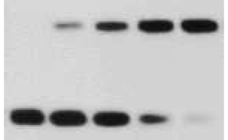
- Average energy dissipated by one molecule as it binds:

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

Energy Dissipation by EcoRI

- Measured specific binding constant:

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



- Average energy dissipated by one molecule as it binds:

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

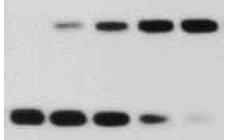
- The Second Law of Thermodynamics as a conversion factor:

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

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:

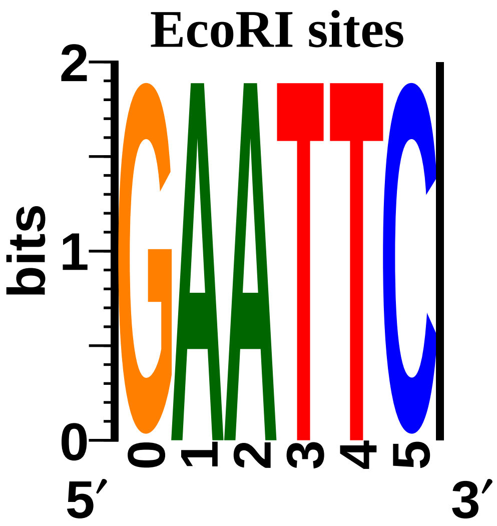
$$\begin{aligned} R_{energy} &= -\Delta G^\circ / \mathcal{E}_{min} \\ &= k_\mathsf{B} T \ln K_{spec} / k_\mathsf{B} T \ln 2 \\ &= \log_2 K_{spec} \qquad \Leftarrow \text{SO SIMPLE!} \\ &= \boxed{\text{17.3 bits per binding}} \end{aligned}$$

Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices

Information/Energy = Efficiency of EcoRI

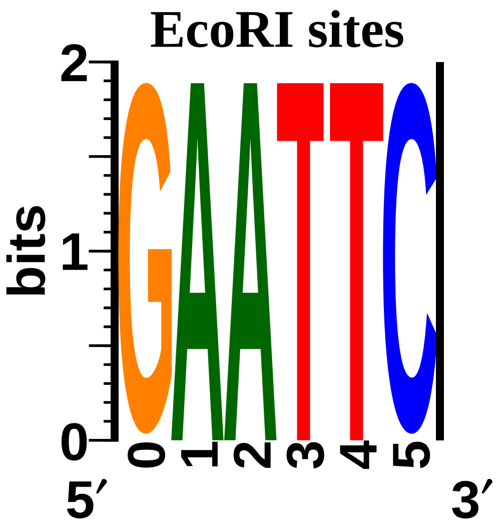
EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.



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

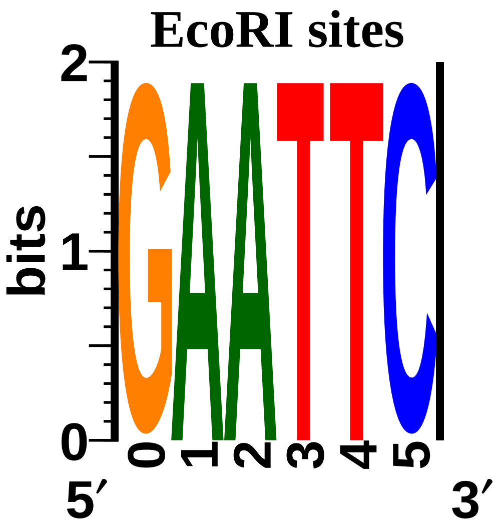


Information/Energy = Efficiency of EcoRI

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

Efficiency is
'WORK' DONE / ENERGY DISSIPATED

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



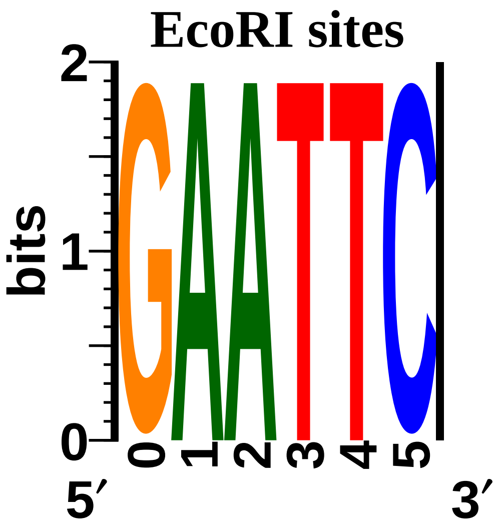
Information/Energy = Efficiency of EcoRI = 70%

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

Efficiency is
'WORK' DONE / ENERGY DISSIPATED

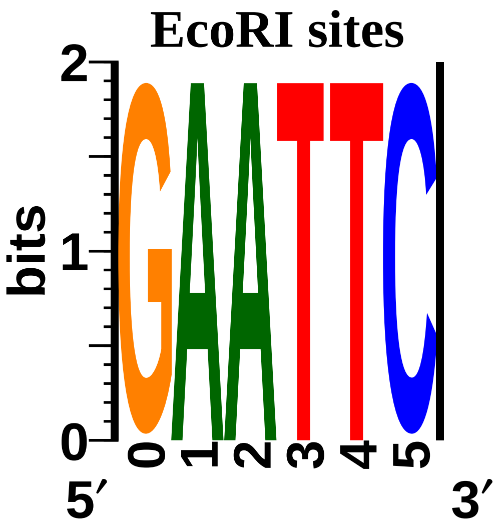
$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

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

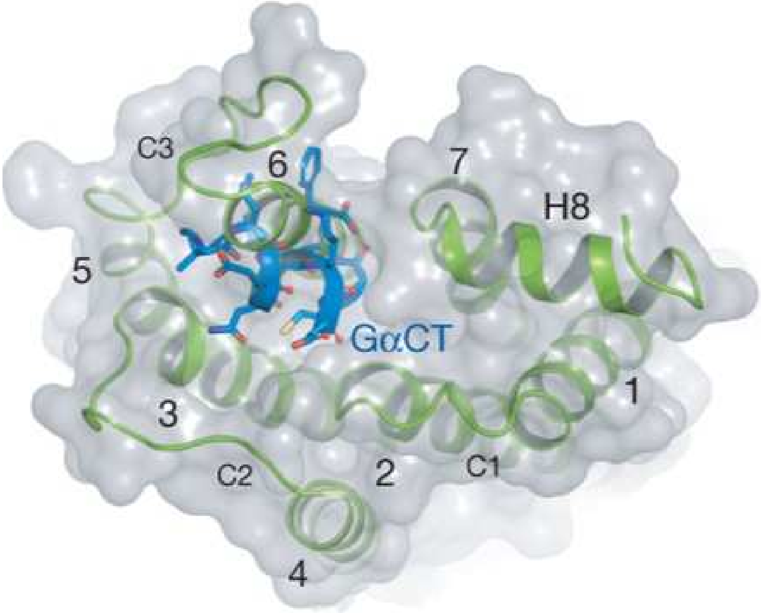
$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

The efficiency is 70%.

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

Rhodopsin Shape Change

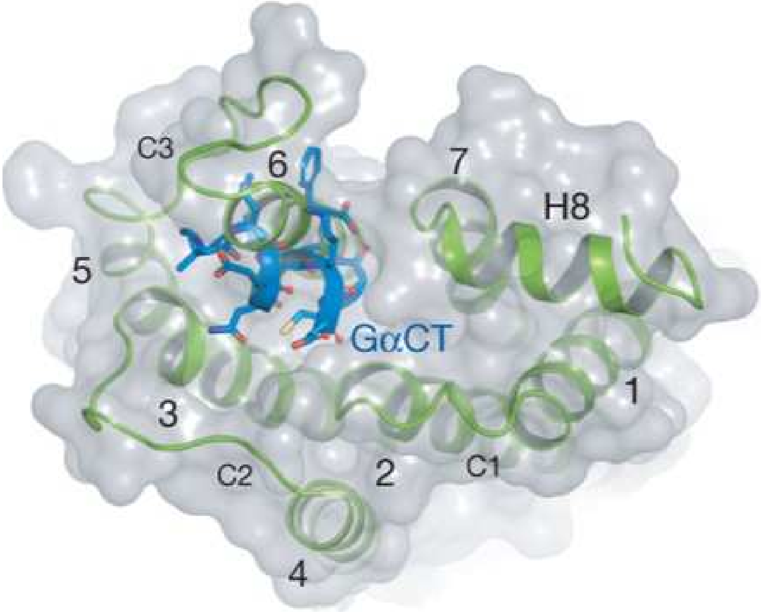
Dark State



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

Rhodopsin Shape Change

Dark State

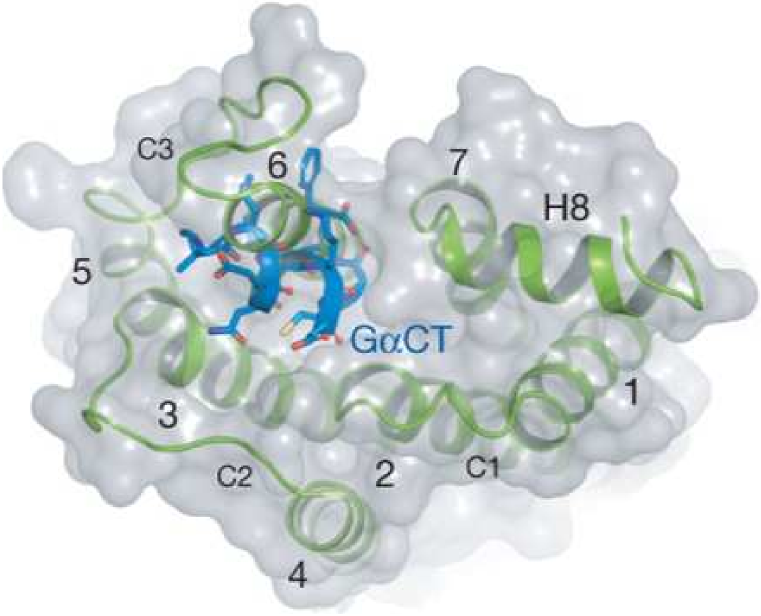


$$\mathbf{h}\nu$$

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

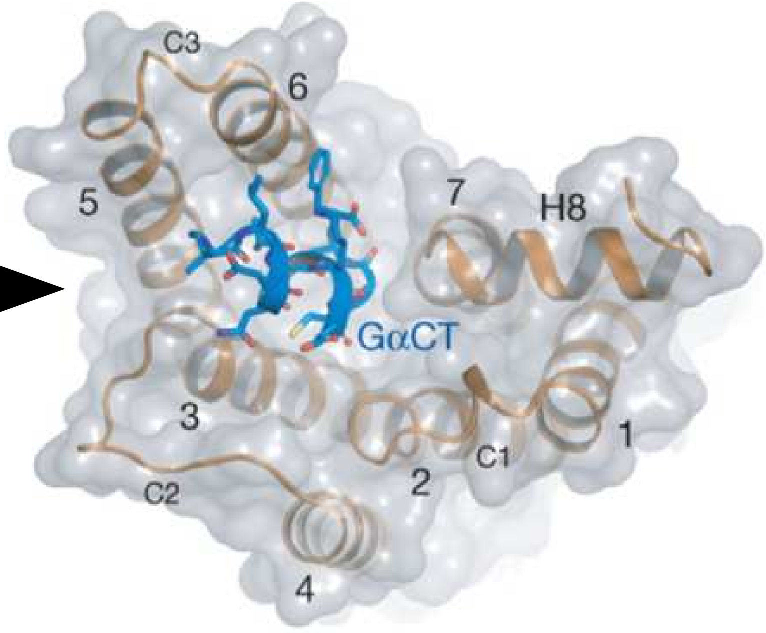
Rhodopsin Shape Change

Dark State



$$\xrightarrow{\mathbf{h}\nu \;\; \mathbf{70\%}}$$

After Photon - Light State

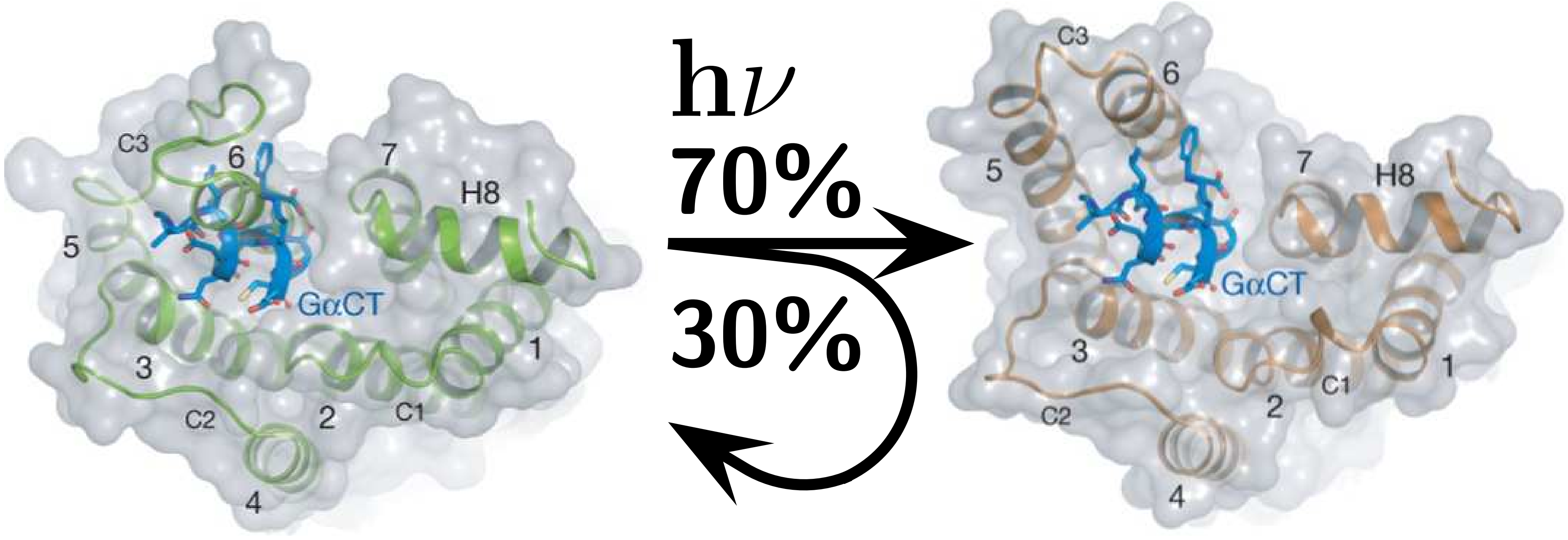


P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Rhodopsin Shape Change

Dark State

After Photon - Light State



$h\nu$

70%

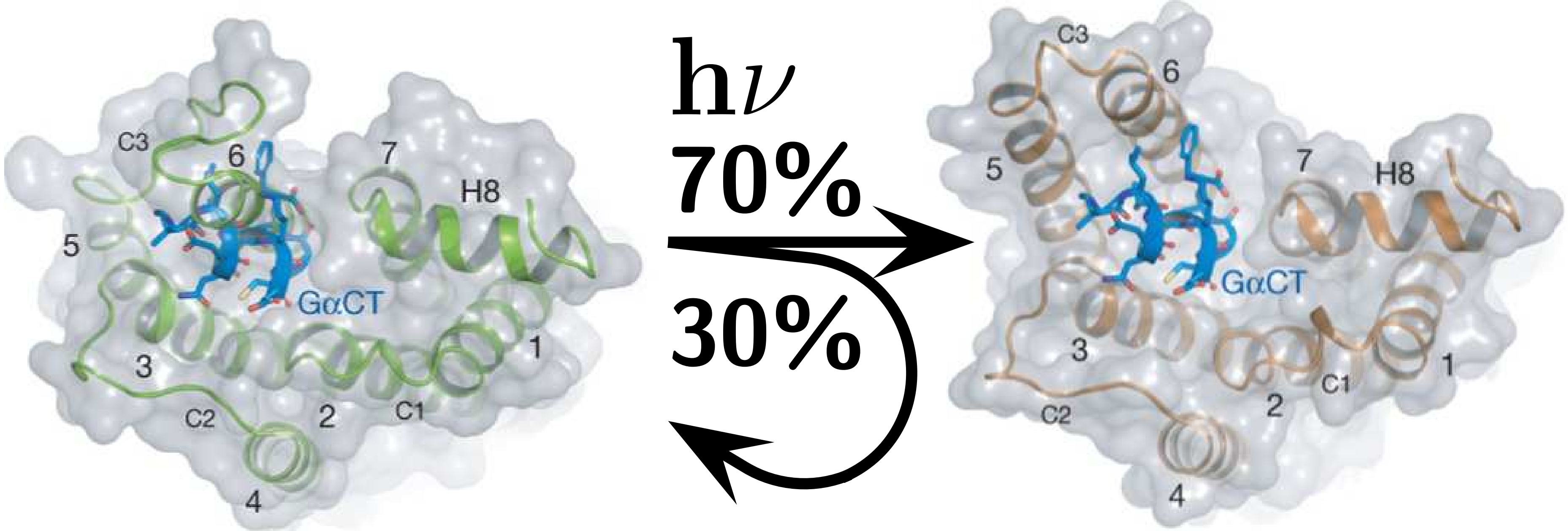
30%

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

Rhodopsin Shape Change

Dark State

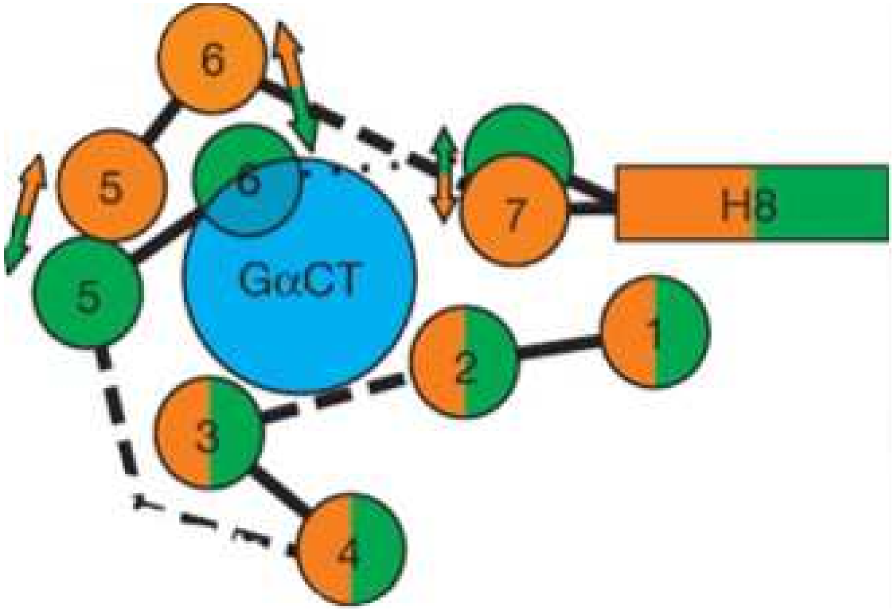
After Photon - Light State



$\mathrm{h}\nu$

70%

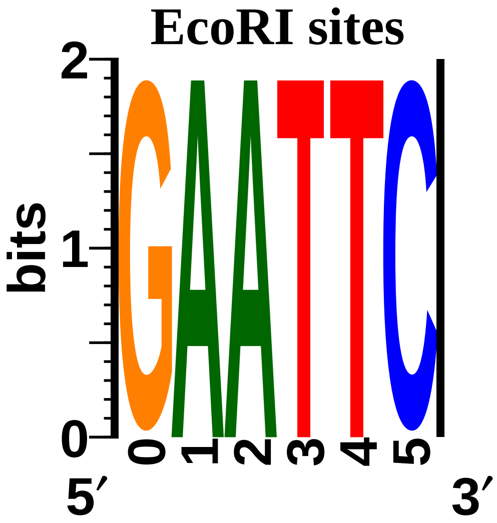
30%

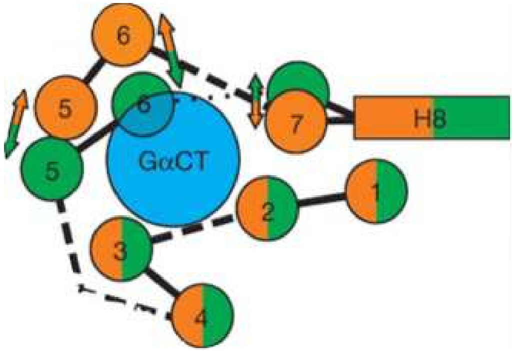


P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

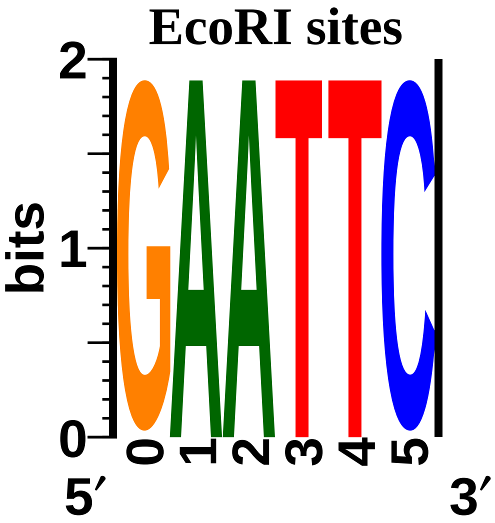


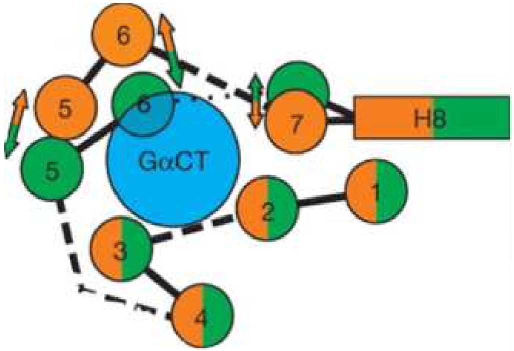


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

- DNA - protein binding

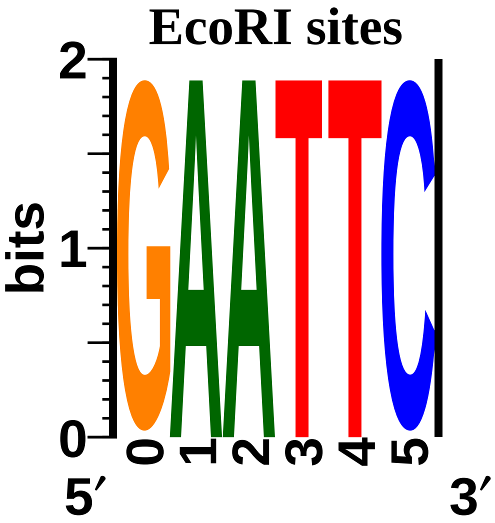


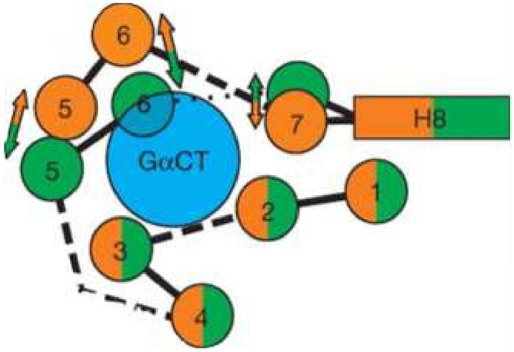


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

- DNA - protein binding
- rhodopsin

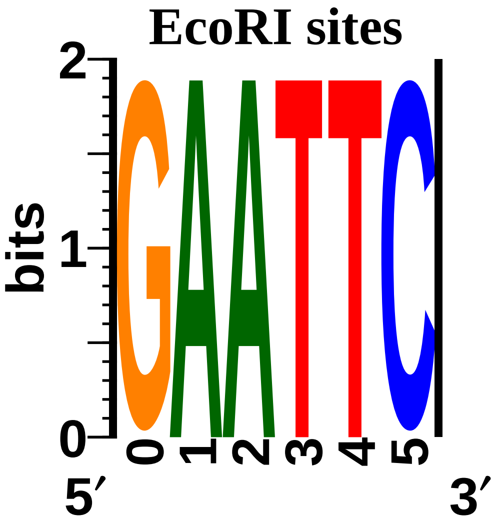


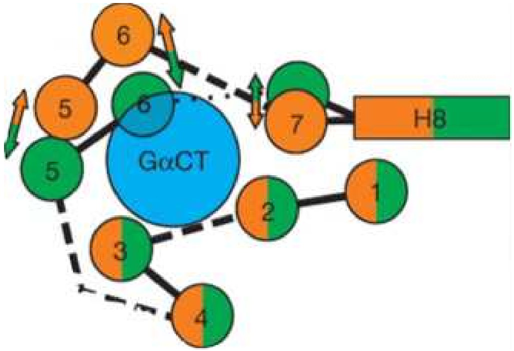


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

- DNA - protein binding
- rhodopsin
- muscle

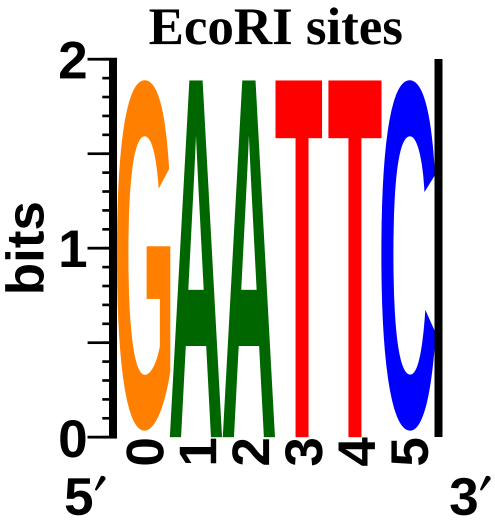


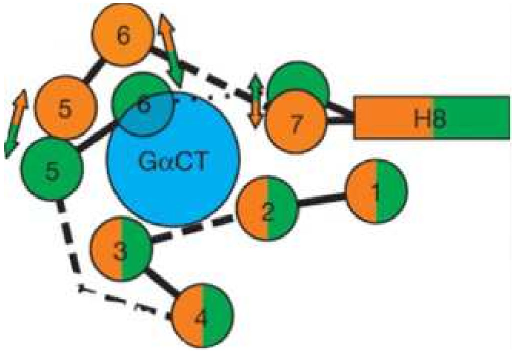


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

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

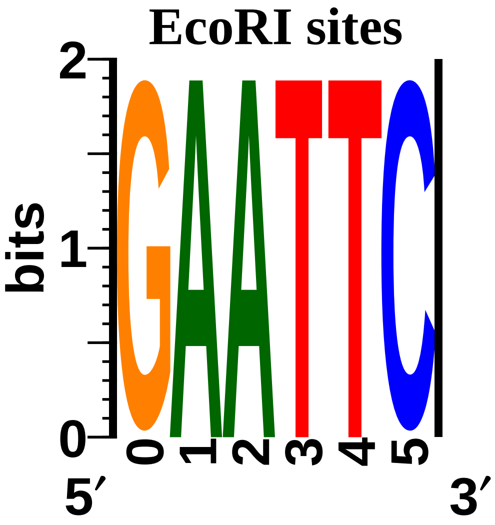




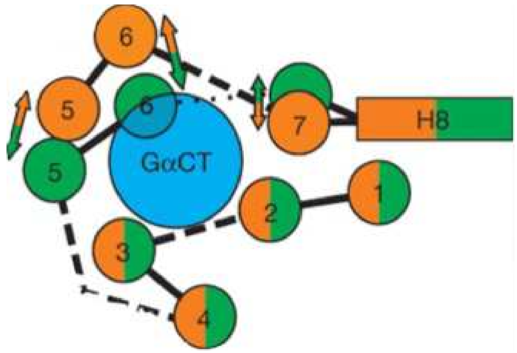
Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

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



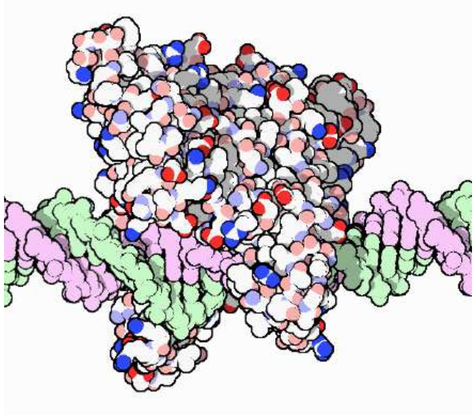
Why 70% efficiency?



Theoretical Isothermal Efficiency

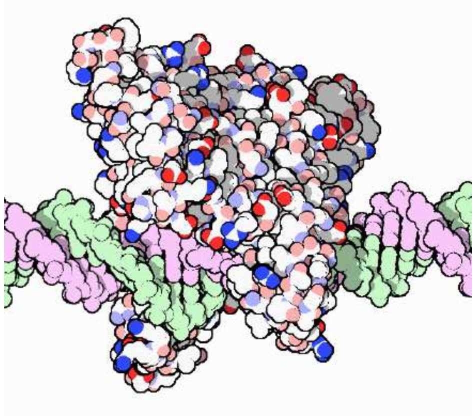
- For molecular states of molecules with d_{space} 'parts' P_y energy is dissipated for noise N_y and

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



T. D. Schneider Nucl. Acids Res. 38: 5995-6006, 2010

Theoretical Isothermal Efficiency



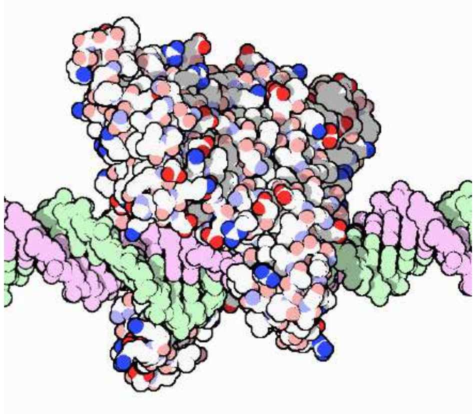
- For molecular states of molecules with d_{space} ‘parts’ P_y energy is dissipated for noise N_y and

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

  $$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow \text{molecular efficiency}$$

T. D. Schneider Nucl. Acids Res. 38: 5995-6006, 2010

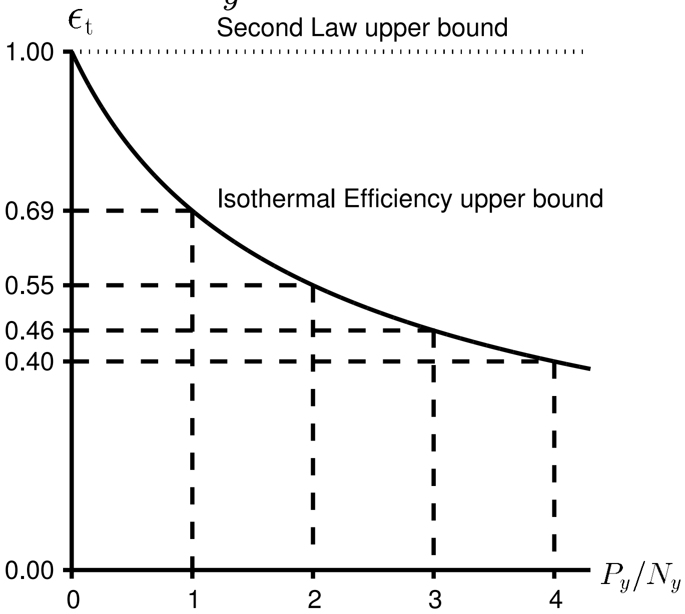
Theoretical Isothermal Efficiency



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

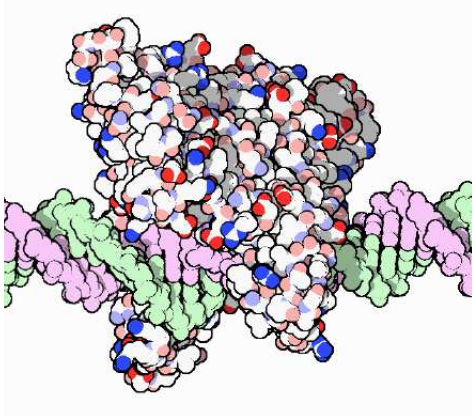
  $C = d_{space} \log_2(P_y/N_y + 1) \leftarrow$ machine capacity

  $\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow$ molecular efficiency



The curve is an upper bound

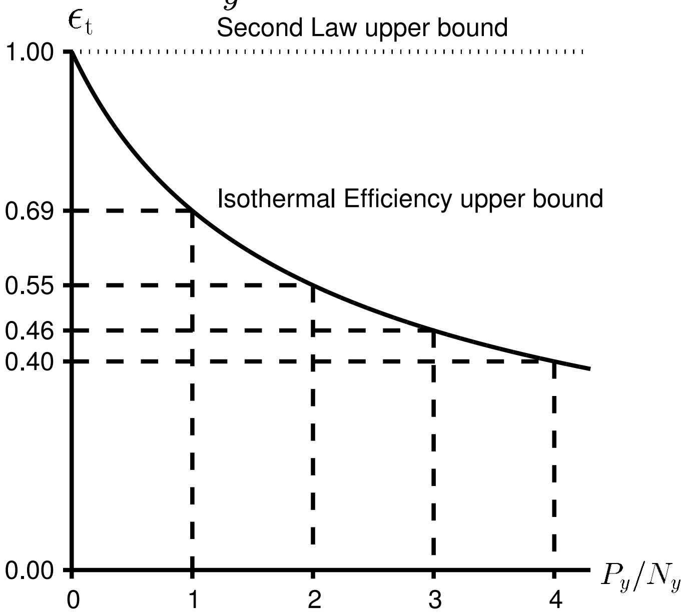
Theoretical Isothermal Efficiency



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

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

  $\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow$ molecular efficiency



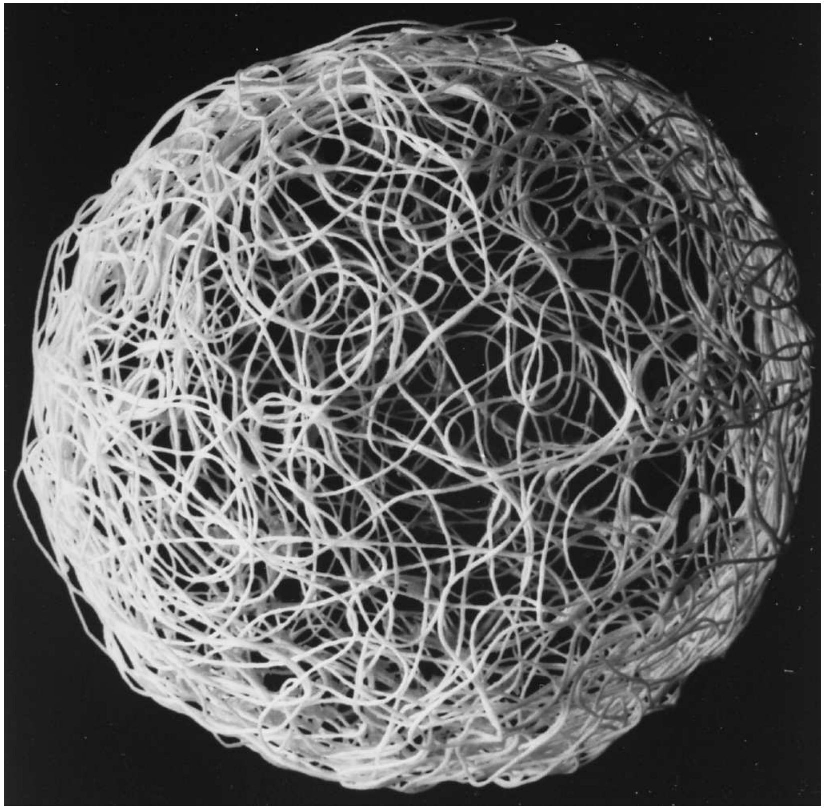
The curve is an upper bound

- $\boxed{\text{If } P_y/N_y = 1 \text{ the efficiency is 70\%!}}$

T. D. Schneider Nucl. Acids Res. 38: 5995-6006, 2010

N Dimensional Sphere Separation

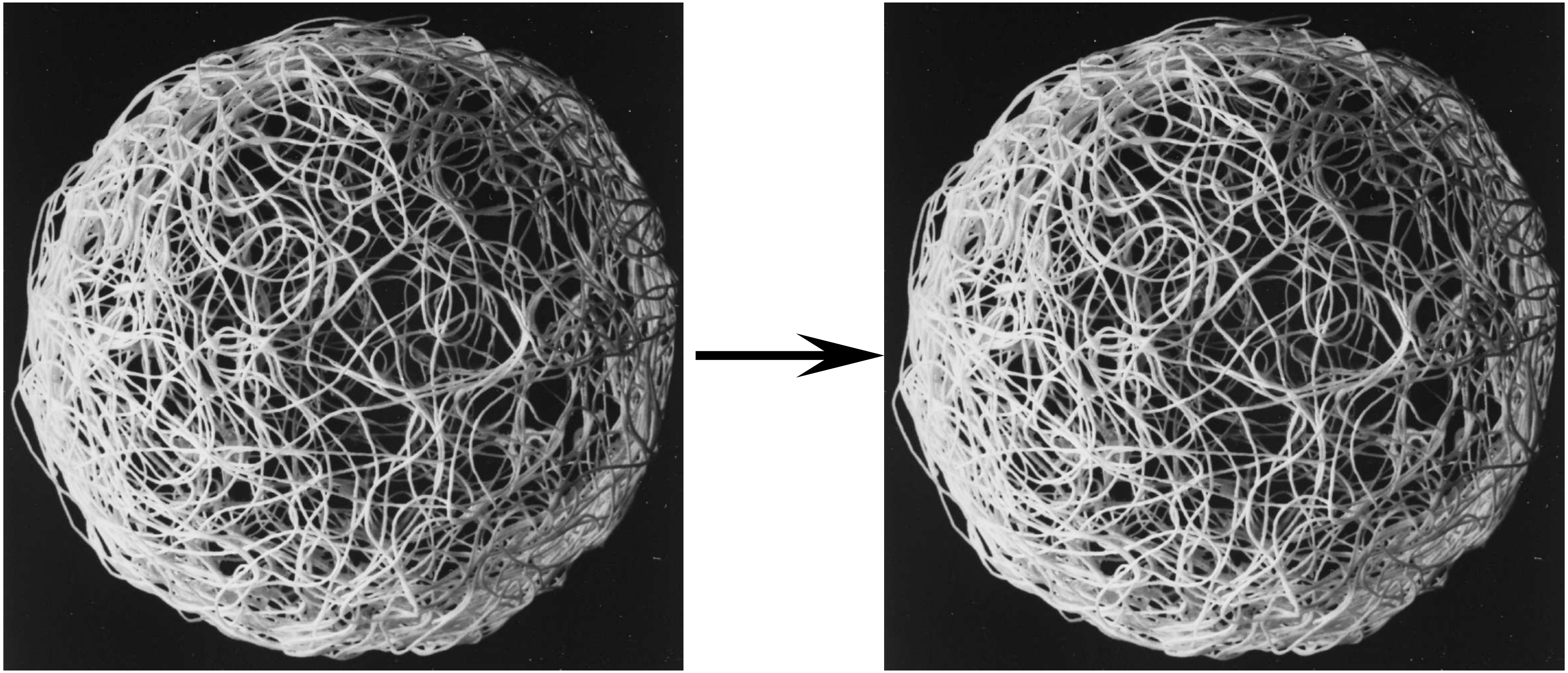
Degenerate Sphere



N Dimensional Sphere Separation

Degenerate Sphere

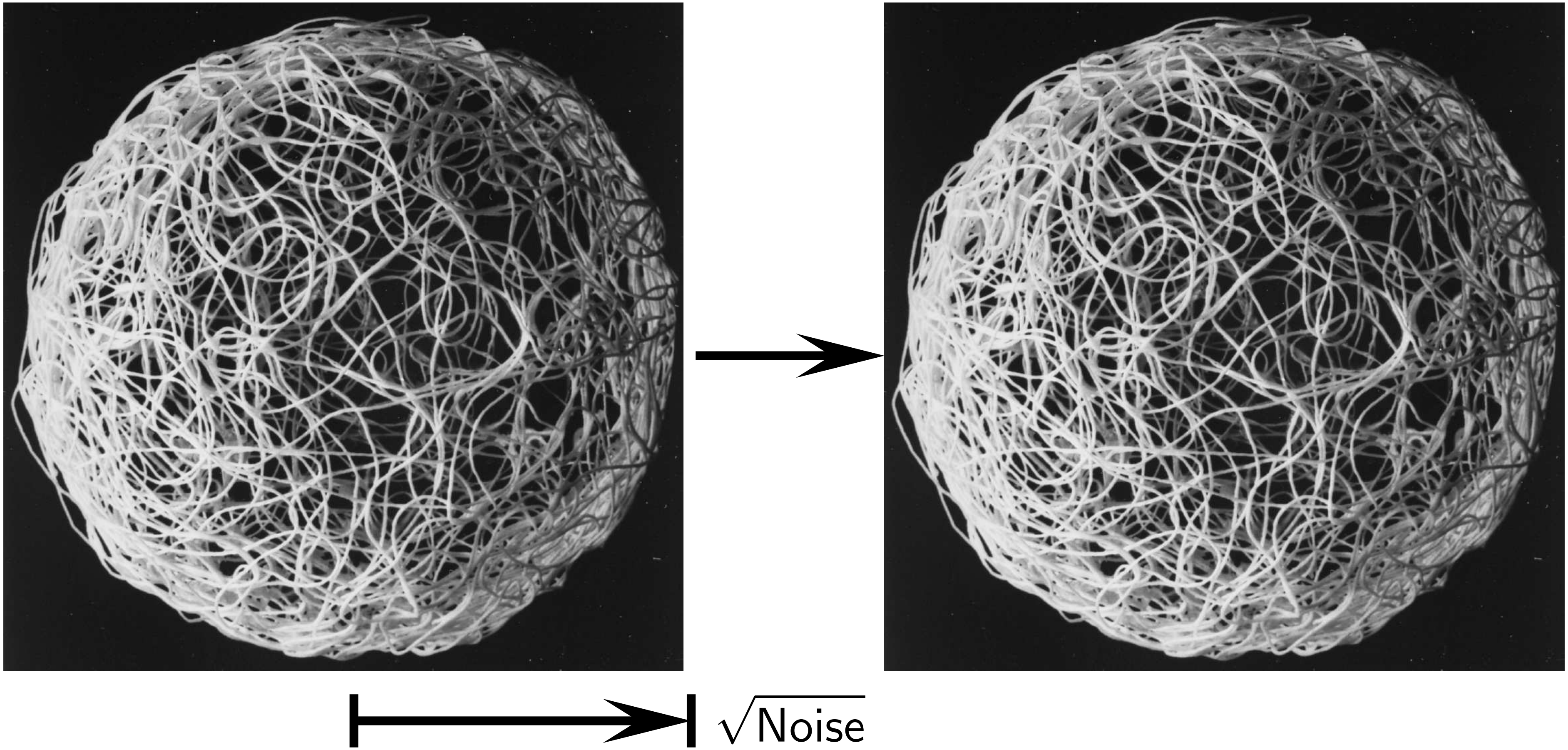
Forward Sphere



N Dimensional Sphere Separation

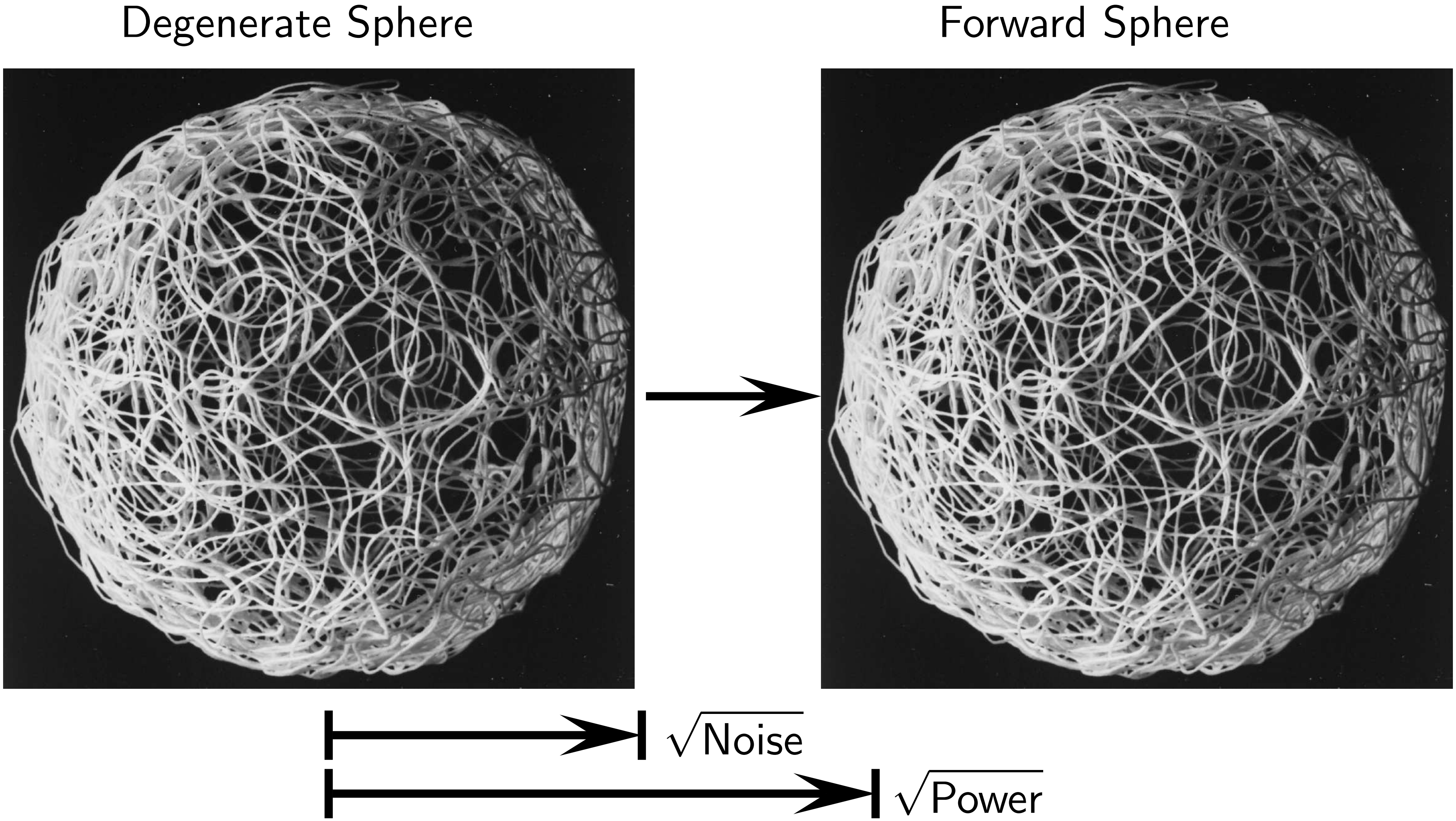
Degenerate Sphere

Forward Sphere



$\sqrt{\text{Noise}}$

N Dimensional Sphere Separation



Degenerate Sphere

Forward Sphere

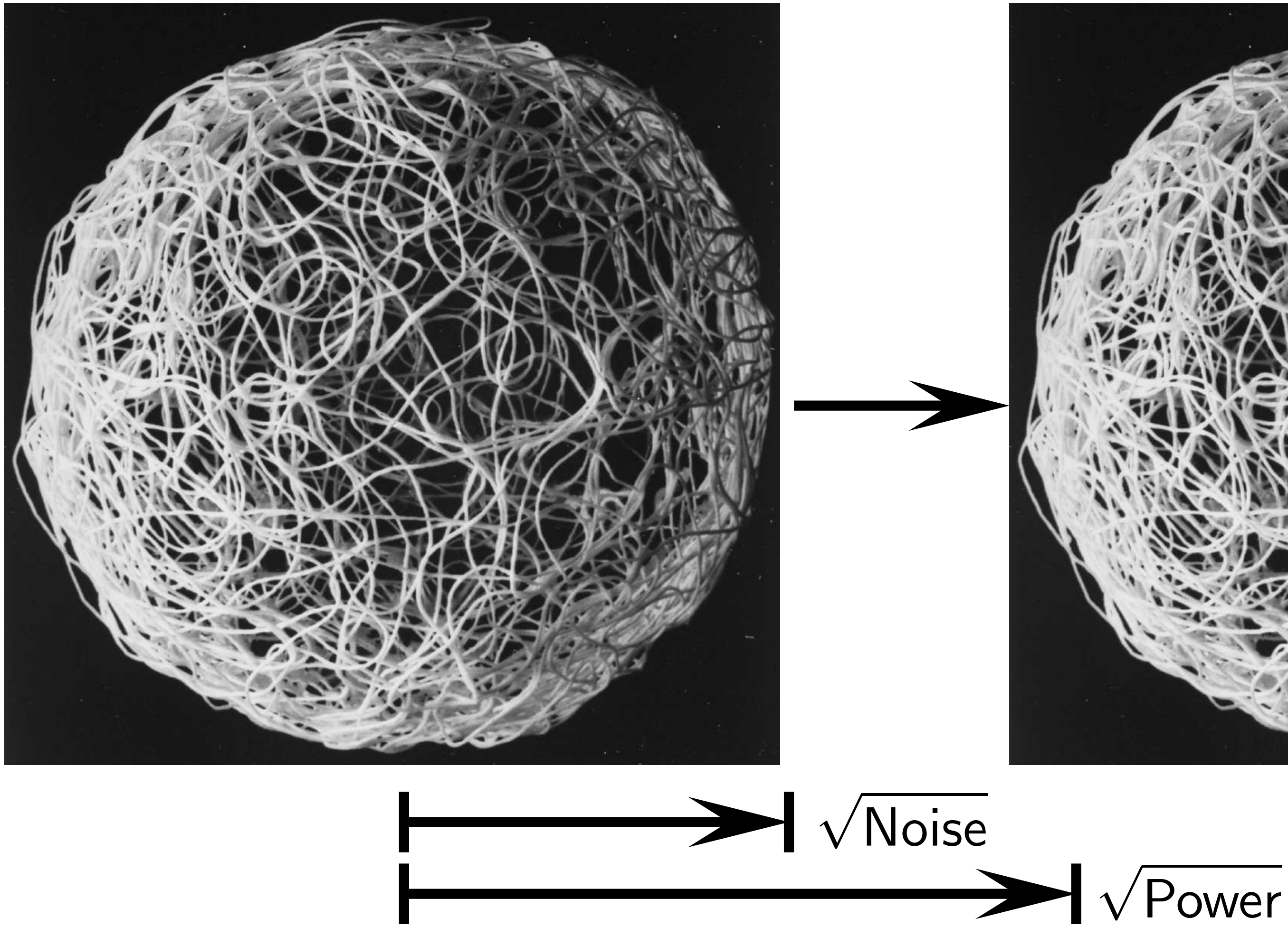
$\sqrt{\text{Noise}}$

$\sqrt{\text{Power}}$

N Dimensional Sphere Separation

Degenerate Sphere

Forward Sphere

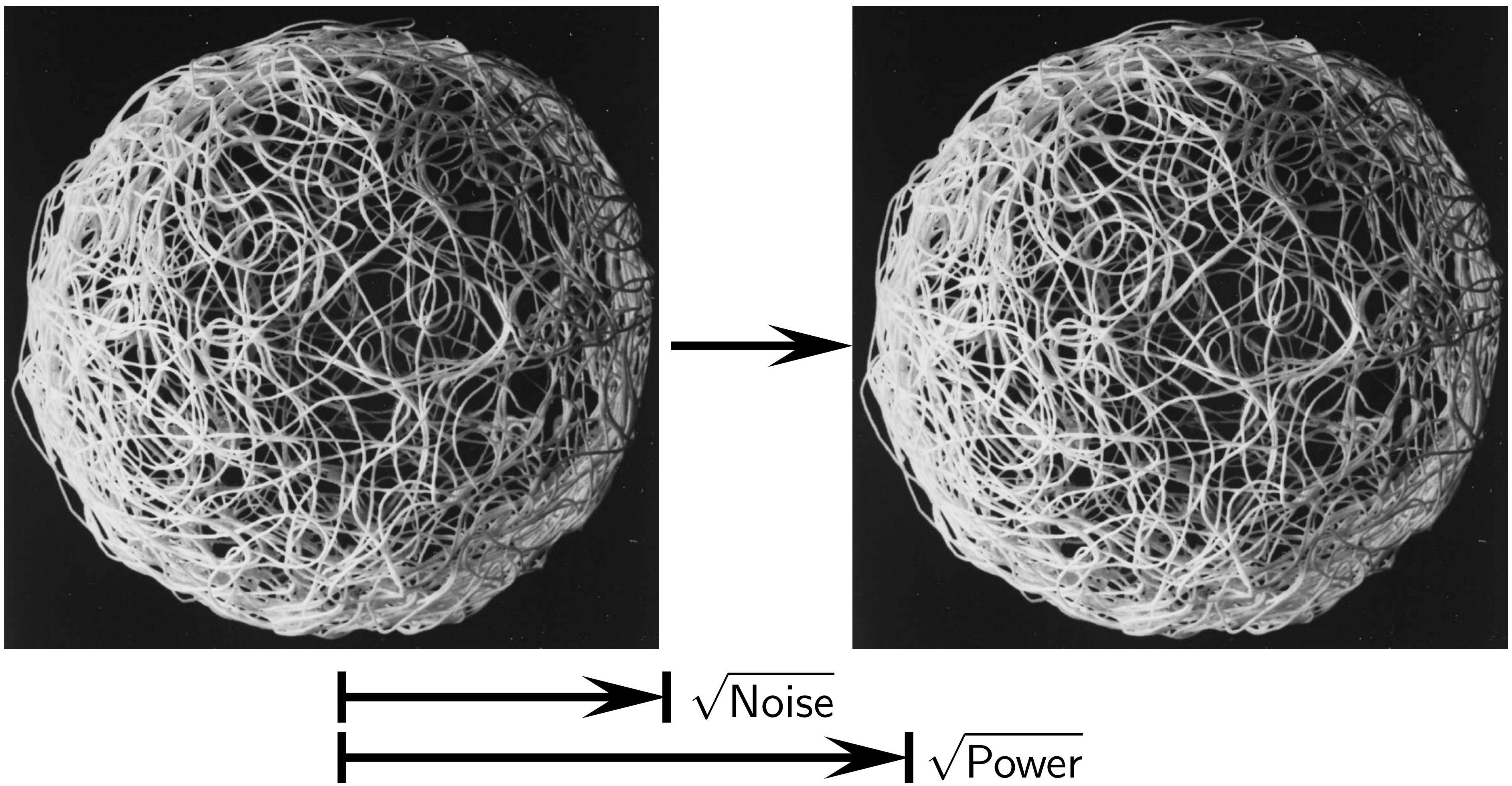


Energy dissipated to escape the Degenerate Sphere must exceed the Noise

N Dimensional Sphere Separation

Degenerate Sphere

Forward Sphere



Energy dissipated to escape the Degenerate Sphere must exceed the Noise

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

Why is the Genetic Code Degenerate?

The Genetic Code

First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

64 codons

$\log_2 64 = 6$ bits/amino acid

The Genetic Code

First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

64 codons

$\log_2 64 = 6$ bits/amino acid

20 amino acids

$\log_2 20 = 4.3$ bits/amino acid

Efficiency of The Genetic Code

Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

64 codons

$\log_2 64 = 6$ bits/amino acid

20 amino acids

$\log_2 20 = 4.3$ bits/amino acid

Compute Efficiency

$$\begin{aligned} \epsilon_r &= \frac{\log_2 \text{actual choices}}{\log_2 \text{maximum choices}} \\ &= \frac{4.3}{6} = 0.72 \end{aligned}$$

Efficiency of The Genetic Code

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

64 codons

$\log_2 64 = 6$ bits/amino acid

20 amino acids

$\log_2 20 = 4.3$ bits/amino acid

Compute Efficiency

$$\begin{aligned} \epsilon_r &= \frac{\log_2 \text{actual choices}}{\log_2 \text{maximum choices}} \\ &= \frac{4.3}{6} = 0.72 \end{aligned}$$

The Genetic Code fits the theory!

Amino Acid Frequencies

A	114882992
C	19056074
D	73332522
E	84344300
F	52828061
G	91113903
H	29753791
I	75133404
K	71121318
L	130161413
M	29818802
N	57427084
O	8
P	67078118
Q	53820991
R	78100977
S	100354324
T	75562140
U	477
V	87249674
W	16751452
Y	40544232

Refine the Calculation

Obtain actual amino acid frequencies from the 50% sequence identity non-redundant Protein Information Resource (PIR) UniRef50 database, January 2011.

$n = 1{,}240{,}702{,}008 = 1.2 \times 10^9$ amino acids

Amino Acid Frequencies

A	114882992
C	19056074
D	73332522
E	84344300
F	52828061
G	91113903
H	29753791
I	75133404
K	71121318
L	130161413
M	29818802
N	57427084
O	8
P	67078118
Q	53820991
R	78100977
S	100354324
T	75562140
U	477
V	87249674
W	16751452
Y	40544232

Refine the Calculation

Obtain actual amino acid frequencies from the 50% sequence identity non-redundant Protein Information Resource (PIR) UniRef50 database, January 2011.

$n = 1{,}240{,}702{,}008 = 1.2 \times 10^9$ amino acids

Compute the uncertainty:

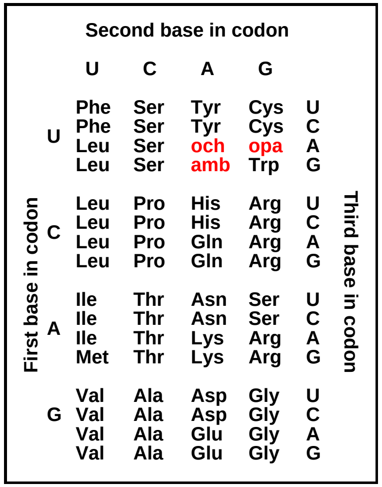
$$\begin{aligned} H_{\mathsf{aa}} &= -\sum_{\mathsf{aa}\,=\,\mathsf{A}}^{\mathsf{Y}} P_{\mathsf{aa}} \log_2 P_{\mathsf{aa}} \quad \text{bits per amino acid} \\ &= 4.170 \quad \text{bits per amino acid} \end{aligned}$$

That's what is actually accomplished by translation.

Translational Efficiency

Compute the efficiency:

$$\epsilon_r = \frac{4.170}{6}$$



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Translational Efficiency

Compute the efficiency:

$$\begin{aligned}\epsilon_r &= \frac{4.170}{6} \\ &= 0.6949 \text{ Measured efficiency}\end{aligned}$$

Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Translational Efficiency

Compute the efficiency:

$$\begin{aligned} \epsilon_r &= \frac{4.170}{6} \\ &= 0.6949 \text{ Measured efficiency} \\ \epsilon_t &= 0.6931 \text{ Theoretical maximum} = \ln(2) \\ & 0.0018 \text{ difference} \end{aligned}$$

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Since this comes from > 1 billion amino acids, 0.2% excess is significant!

Translational Efficiency

Compute the efficiency:

$$
\begin{aligned}
\epsilon_r &= \frac{4.170}{6} \\
&= 0.6949 \text{ Measured efficiency} \\
\epsilon_t &= 0.6931 \text{ Theoretical maximum} = \ln(2) \\
& \ 0.0018 \text{ difference}
\end{aligned}
$$

Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Since this comes from > 1 billion amino acids, 0.2% excess is significant!

Theory violation! ...What's Missing?

- Rare amino acids don't contribute much.

Translational Efficiency

Compute the efficiency:

$$\begin{aligned}
\epsilon_r &= \frac{4.170}{6} \\
&= 0.6949 \text{ Measured efficiency} \\
\epsilon_t &= 0.6931 \text{ Theoretical maximum} = \ln(2) \\
& \ 0.0018 \text{ difference}
\end{aligned}$$

		Second base in codon				
		U	C	A	G	Third base in codon
First base in codon	U	Phe	Ser	Tyr	Cys	U
		Phe	Ser	Tyr	Cys	C
		Leu	Ser	och	opa	A
		Leu	Ser	amb	Trp	G
	C	Leu	Pro	His	Arg	U
		Leu	Pro	His	Arg	C
		Leu	Pro	Gln	Arg	A
		Leu	Pro	Gln	Arg	G
	A	Ile	Thr	Asn	Ser	U
		Ile	Thr	Asn	Ser	C
		Ile	Thr	Lys	Arg	A
		Met	Thr	Lys	Arg	G
	G	Val	Ala	Asp	Gly	U
		Val	Ala	Asp	Gly	C
		Val	Ala	Glu	Gly	A
		Val	Ala	Glu	Gly	G

Since this comes from > 1 billion amino acids, 0.2% excess is significant!

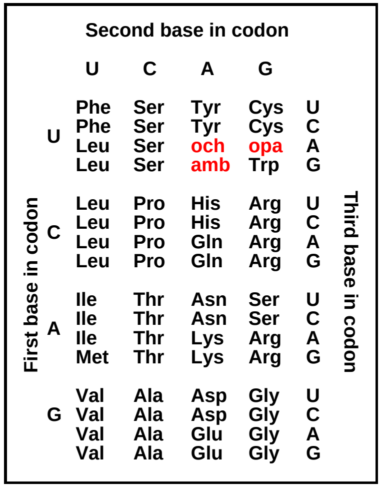
Theory violation! . . . What's Missing?

- Rare amino acids don't contribute much.
- Removing the stop codons reduces the maximum from 6 bits to $\log_2 61 = 5.931$ bits and the efficiency would be $4.170/5.931 = 0.7031$, so this makes the situation worse and does not explain the discrepancy.

Translational Efficiency

Compute the efficiency:

$$
\begin{aligned}
\epsilon_r &= \frac{4.170}{6} \\
&= 0.6949 \text{ Measured efficiency} \\
\epsilon_t &= 0.6931 \text{ Theoretical maximum} = \ln(2) \\
& 0.0018 \text{ difference}
\end{aligned}
$$



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

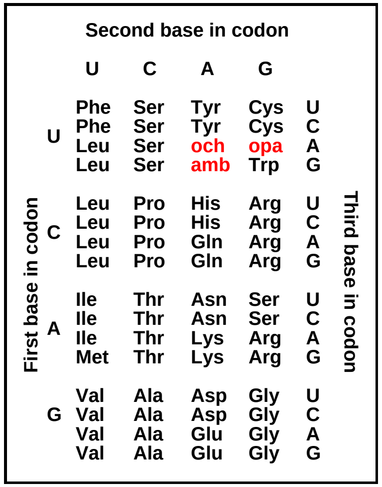
Since this comes from > 1 billion amino acids, 0.2% excess is significant!

Theory violation! . . . What's Missing?

- Rare amino acids don't contribute much.
- Removing the stop codons reduces the maximum from 6 bits to $\log_2 61 = 5.931$ bits and the efficiency would be $4.170/5.931 = 0.7031$, so this makes the situation worse and does not explain the discrepancy.
- Translational error rate was not accounted for?

Efficiency of the Genetic Code

Theory Violation! What's missing?
Error rate of transcription/translation was not accounted for.
See if we can compute it.



	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Efficiency of the Genetic Code

Theory Violation! What's missing?
Error rate of transcription/translation was not accounted for.
See if we can compute it.

Compute Error Rate

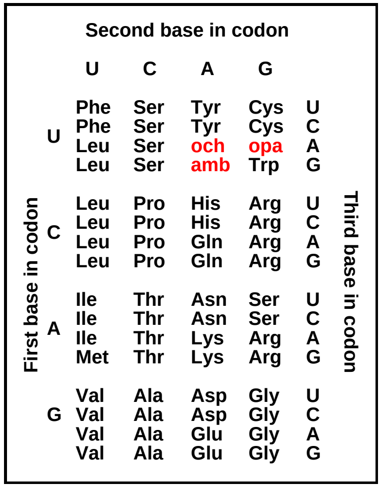
Proper Computation:

$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Efficiency of the Genetic Code

Theory Violation! What's missing?
Error rate of transcription/translation was not accounted for.
See if we can compute it.



First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Compute Error Rate

Proper Computation:

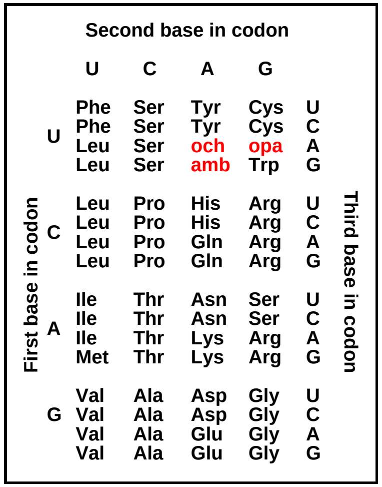
$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

Average probability of misincorporation, P_{error} determines the information lost:

$$H_{\text{error}} = [-P_{\text{error}} \log_2 P_{\text{error}}] + [-(1 - P_{\text{error}}) \log_2 (1 - P_{\text{error}})]$$

Efficiency of the Genetic Code

Theory Violation! What's missing?
Error rate of transcription/translation was not accounted for.
See if we can compute it.



	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Compute Error Rate

Proper Computation:

$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

Average probability of misincorporation, P_{error} determines the information lost:

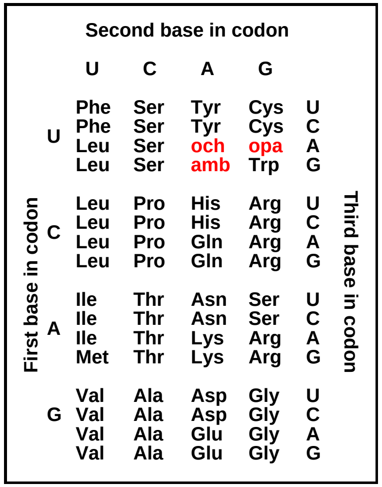
$$H_{\text{error}} = [-P_{\text{error}} \log_2 P_{\text{error}}] + [-(1 - P_{\text{error}}) \log_2 (1 - P_{\text{error}})]$$

Solving gives the **theoretically predicted error rate of translation**:

$$P_{\text{error}} = 0.94 \times 10^{-4} \approx 1 \times 10^{-3}$$

Efficiency of the Genetic Code

Theory Violation! What's missing?
Error rate of transcription/translation was not accounted for.
See if we can compute it.



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Compute Error Rate

Proper Computation:

$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

Average probability of misincorporation, P_{error} determines the information lost:

$$H_{\text{error}} = [-P_{\text{error}} \log_2 P_{\text{error}}] + [-(1 - P_{\text{error}}) \log_2 (1 - P_{\text{error}})]$$

Solving gives the **theoretically predicted error rate of translation**:

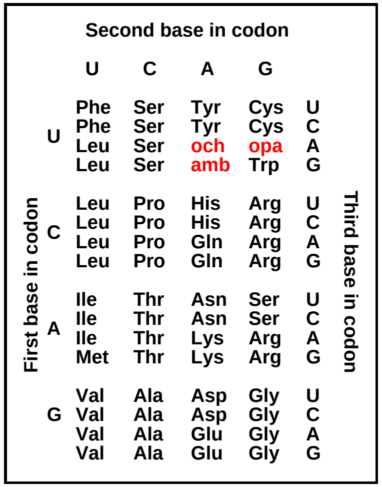
$$P_{\text{error}} = 0.94 \times 10^{-4} \approx 1 \times 10^{-3}$$

Experimental data from Parker (1989) gave:

$$5 \times 10^{-5} \text{ to } 3 \times 10^{-3},$$

$$\text{average } \approx (1 \pm 1) \times 10^{-3}$$

Efficiency of the Genetic Code



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Theory Violation! What's missing?
Error rate of transcription/translation was not accounted for.
See if we can compute it.

Compute Error Rate
Proper Computation:

$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

Average probability of misincorporation, P_{error} determines the information lost:

$$H_{\text{error}} = [-P_{\text{error}} \log_2 P_{\text{error}}] + [-(1 - P_{\text{error}}) \log_2 (1 - P_{\text{error}})]$$

Solving gives the **theoretically predicted error rate of translation**:

$$P_{\text{error}} = 0.94 \times 10^{-4} \approx 1 \times 10^{-3}$$

Experimental data from Parker (1989) gave:

$$5 \times 10^{-5} \text{ to } 3 \times 10^{-3},$$
$$\text{average } \approx (1 \pm 1) \times 10^{-3}$$

The theory correctly predicts the error rate of translation

Efficiency of the Genetic Code

Combine:

Frequencies of >1 billion amino acids

Second base in codon

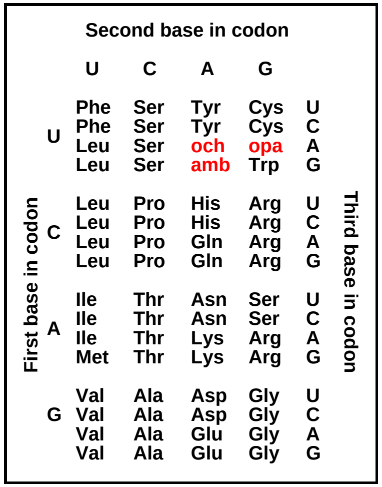
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Efficiency of the Genetic Code

Combine:
Frequencies of >1 billion amino acids
with
The known translational error rate, 1×10^{-3}

First base in codon	Second base in codon				Third base in codon
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Efficiency of the Genetic Code



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Combine:

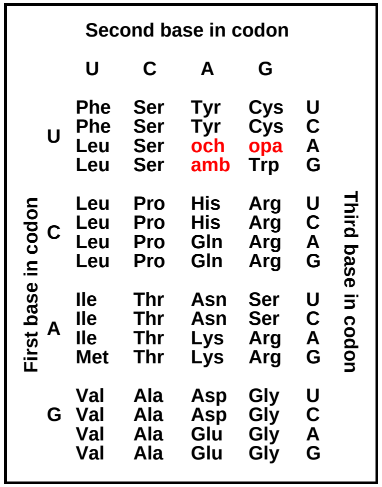
Frequencies of >1 billion amino acids

with

The known translational error rate, 1×10^{-3}

$$(H_{aa} - H(P_{\text{error}}))/6 \;=\; 0.69304765 = \text{ measured efficiency}$$

Efficiency of the Genetic Code



	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Combine:

Frequencies of >1 billion amino acids

with

The known translational error rate, 1×10^{-3}

$$(H_{aa} - H(P_{\text{error}}))/6 \;=\; 0.69304765 = \text{measured efficiency}$$

$$\ln(2) \;=\; 0.69314718 = \text{theoretical efficiency}$$

Efficiency of the Genetic Code

First base in codon	Second base in codon: U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Combine:

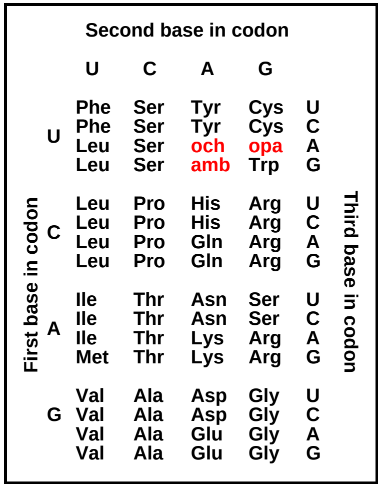
Frequencies of >1 billion amino acids

with

The known translational error rate, 1×10^{-3}

$$\begin{aligned}(H_{aa} - H(P_{\text{error}}))/6 &= 0.69304765 = \text{ measured efficiency}\\ \ln(2) &= 0.69314718 = \text{ theoretical efficiency}\\ \Delta &= 0.\underline{0000}9953 = \text{ difference}\end{aligned}$$

Efficiency of the Genetic Code



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Combine:
Frequencies of >1 billion amino acids
with
The known translational error rate, 1×10^{-3}

$$\begin{aligned}
(H_{aa} - H(P\text{error}))/6 &= 0.69304765 = \text{measured efficiency} \\
\ln(2) &= 0.69314718 = \text{theoretical efficiency} \\
\Delta &= 0.\underline{0000}9953 = \text{difference}
\end{aligned}$$

The theory matches the data to 4 decimal places!

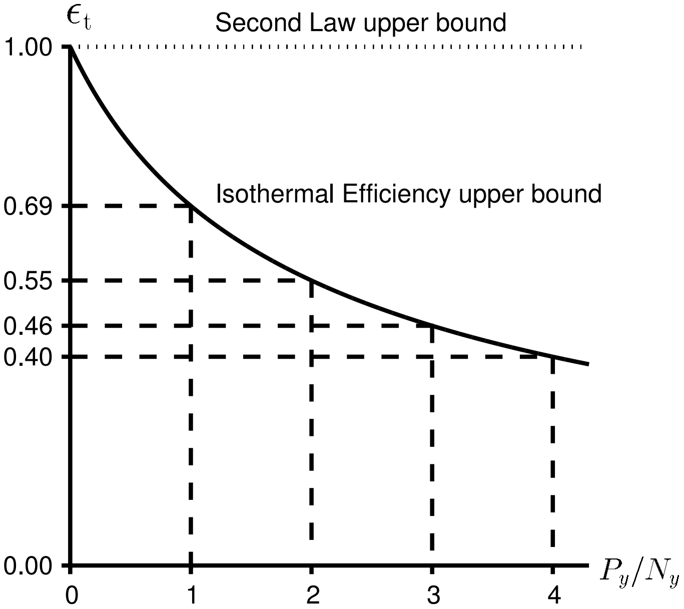
Why the Genetic Code is Degenerate

- **The genetic code has an isothermal efficiency at** $\ln 2 = 0.693$

Why the Genetic Code is Degenerate

- **The genetic code has an isothermal efficiency at** $\ln 2 = 0.693$
- **. . . so the genetic code is optimally efficient molecular machine**

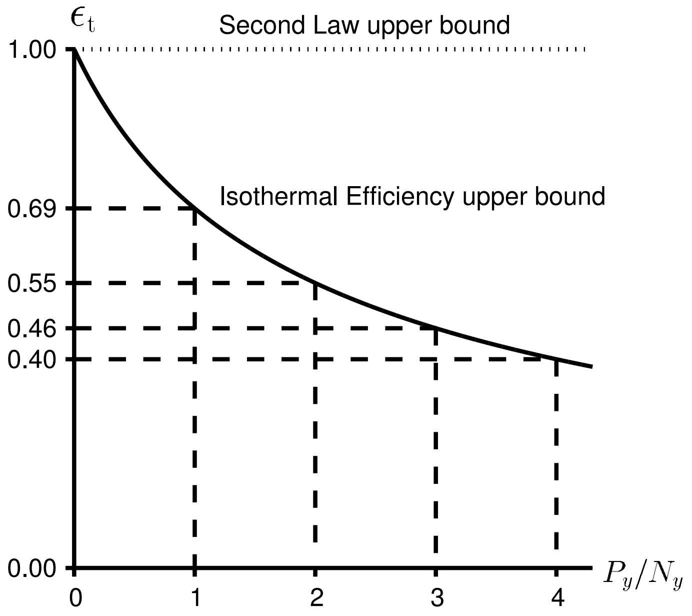
$$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}}$$



Why the Genetic Code is Degenerate

- **The genetic code has an isothermal efficiency at** $\ln 2 = 0.693$
- **. . . so the genetic code is optimally efficient molecular machine**

$$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}}$$

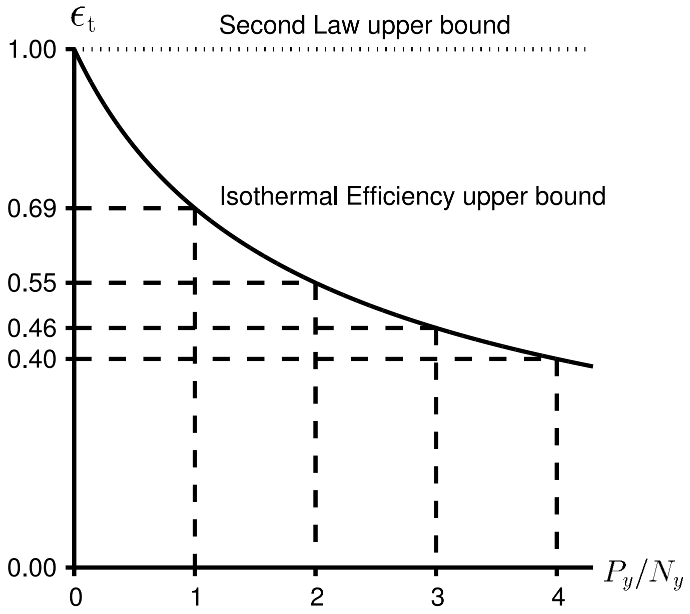


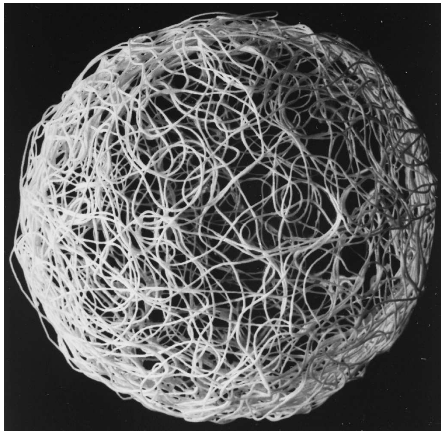
- **. . . so** $P_y/N_y \geq 1$

Why the Genetic Code is Degenerate

- **The genetic code has an isothermal efficiency at** $\ln 2 = 0.693$
- **. . . so the genetic code is optimally efficient molecular machine**

$$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}}$$



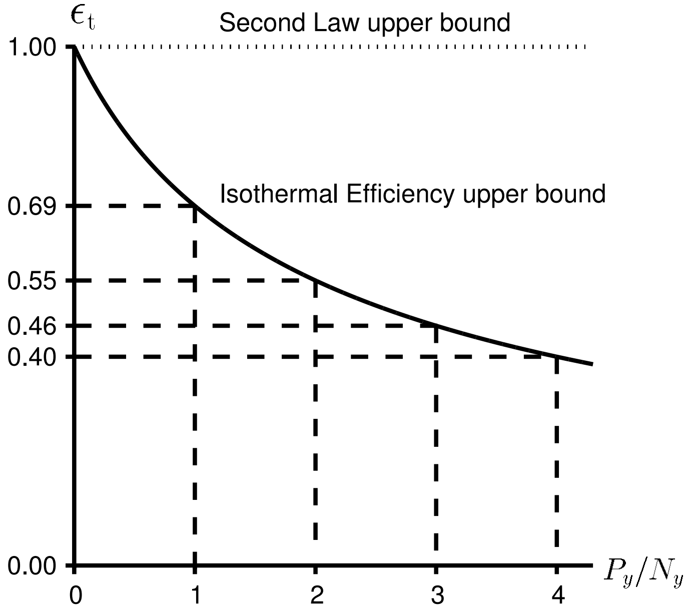


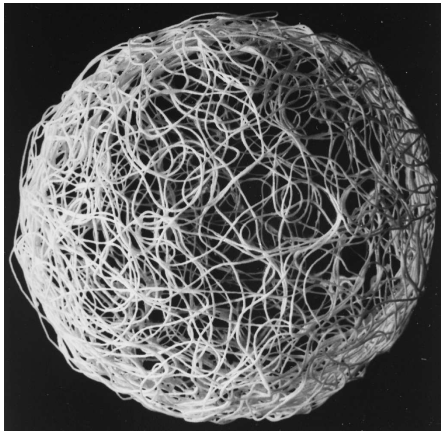
- **. . . so** $P_y/N_y \geq 1$
- **. . . so the amino acid states are distinct high dimensional spheres**

Why the Genetic Code is Degenerate

- The genetic code has an isothermal efficiency at $\ln 2 = 0.693$
- . . . so the genetic code is optimally efficient molecular machine

$$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}}$$



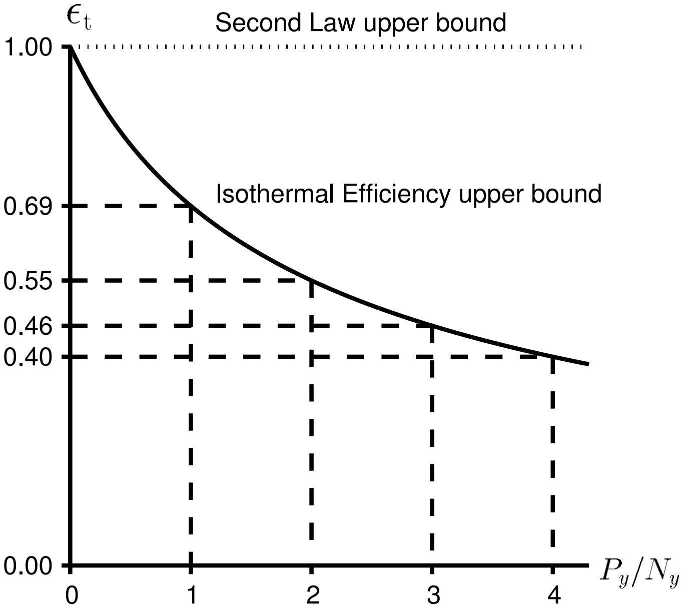


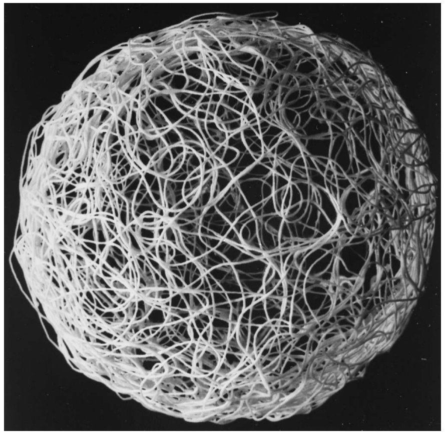
- . . . so $P_y/N_y \geq 1$
- . . . so the amino acid states are distinct high dimensional spheres
- . . . and there is good sphere packing: the spheres do not intersect.

Why the Genetic Code is Degenerate

- The genetic code has an isothermal efficiency at $\ln 2 = 0.693$
- . . . so the genetic code is optimally efficient molecular machine

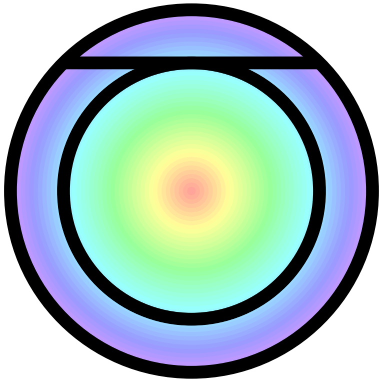
$$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}}$$





- . . . so $P_y/N_y \geq 1$
- . . . so the amino acid states are distinct high dimensional spheres
- . . . and there is good sphere packing: the spheres do not intersect.
- The price for having distinct states is ‘degeneracy’.

Acknowledgments



Herbert A. Schneider (1922-2009)

Andrzej Ehrenfeucht

John Spouge, **Peter Rogan**, **John Garavelli**

Andrzej Konopka, Martin Bier, Ilya Lyakhov, Danielle Needle, Peyman Khalichi, Carrie Paterson, Ryan Shultzaberger, Amar Klar, Peter Lemkin, Barry Zeeberg, Lynn Bayer, Zehua Chen, Blake Sweeney, Bert Gold, Sorina Eftim, Mikhail Kashlev, Alex Mitrophanov, Peter Thomas, Hong Qian, Baris E. Suzek, Jeff Strathern and Peter Wills

National Institutes of Health, National Cancer Institute

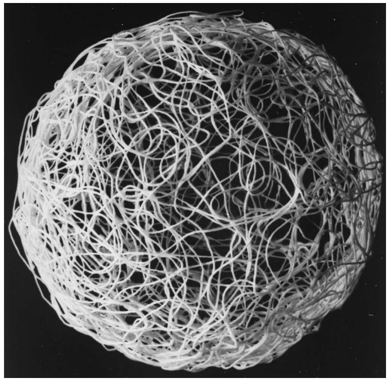
DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

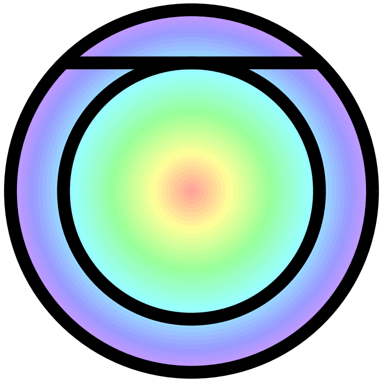
NATIONAL INSTITUTES OF HEALTH

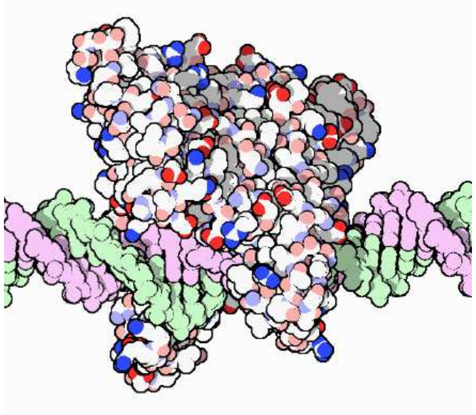
NATIONAL® CANCER INSTITUTE

Web site:

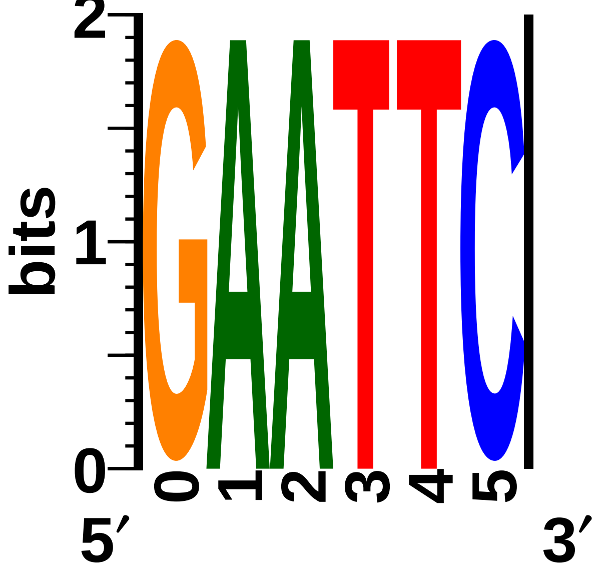
TinyURL.com/tomschneider





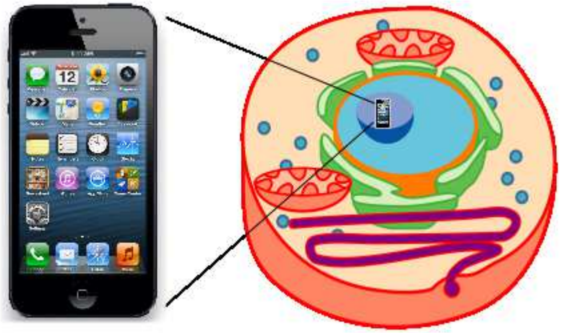


EcoRI sites



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G



Version

version = 1.58 of code15.tex 2014 Apr 29