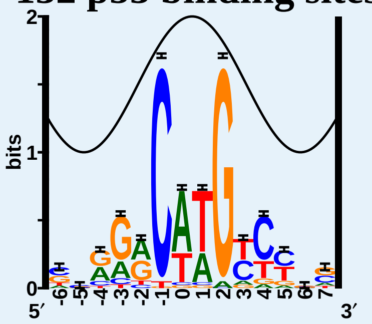
Information Theory and Molecular Biology

Thomas D. Schneider, Ph.D.

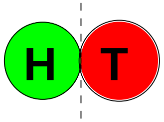
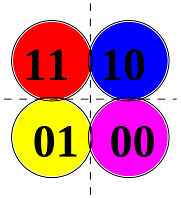
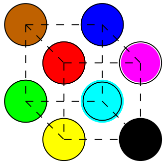
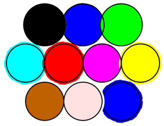
National Cancer Institute at Frederick
Gene Regulation and Chromosome Biology Laboratory
Molecular Information Theory Group



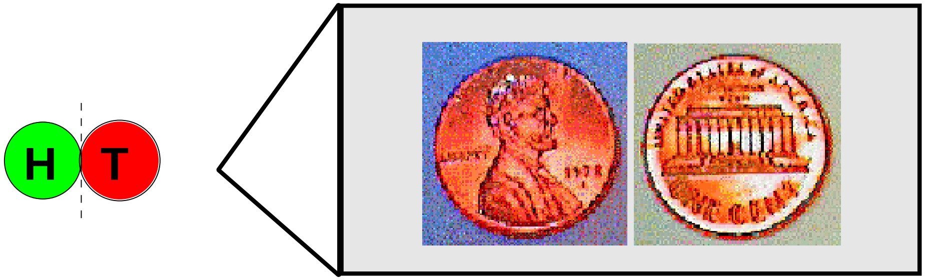
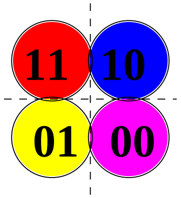
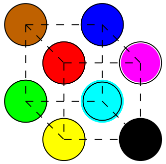
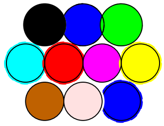
132 p53 binding sites

El Duomo, Florence, Italy

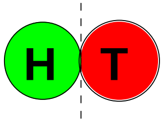
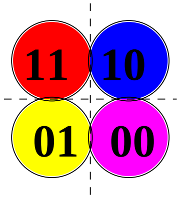
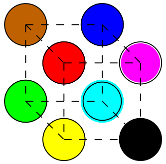
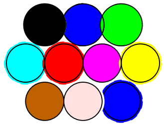
Information Theory: One-Minute Lesson

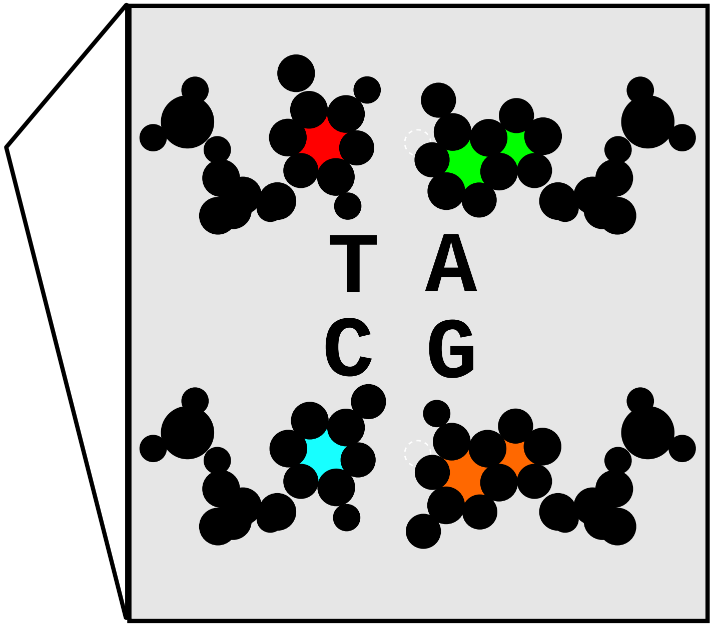
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

Information Theory: One-Minute Lesson

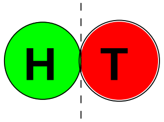
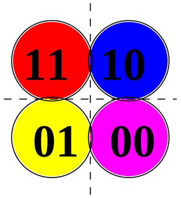
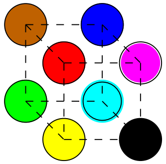
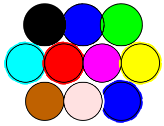
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

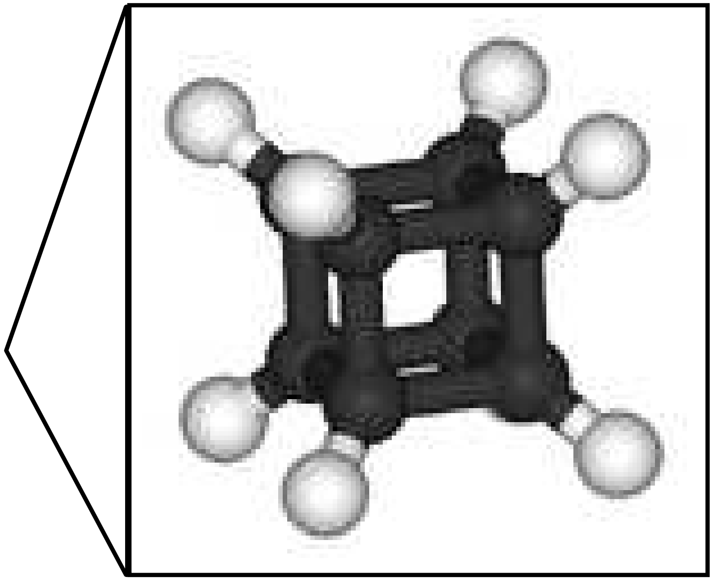
Information Theory: One-Minute Lesson

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

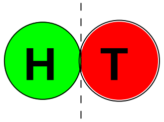
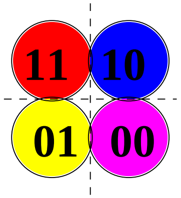
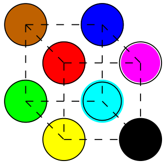
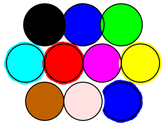


Information Theory: One-Minute Lesson

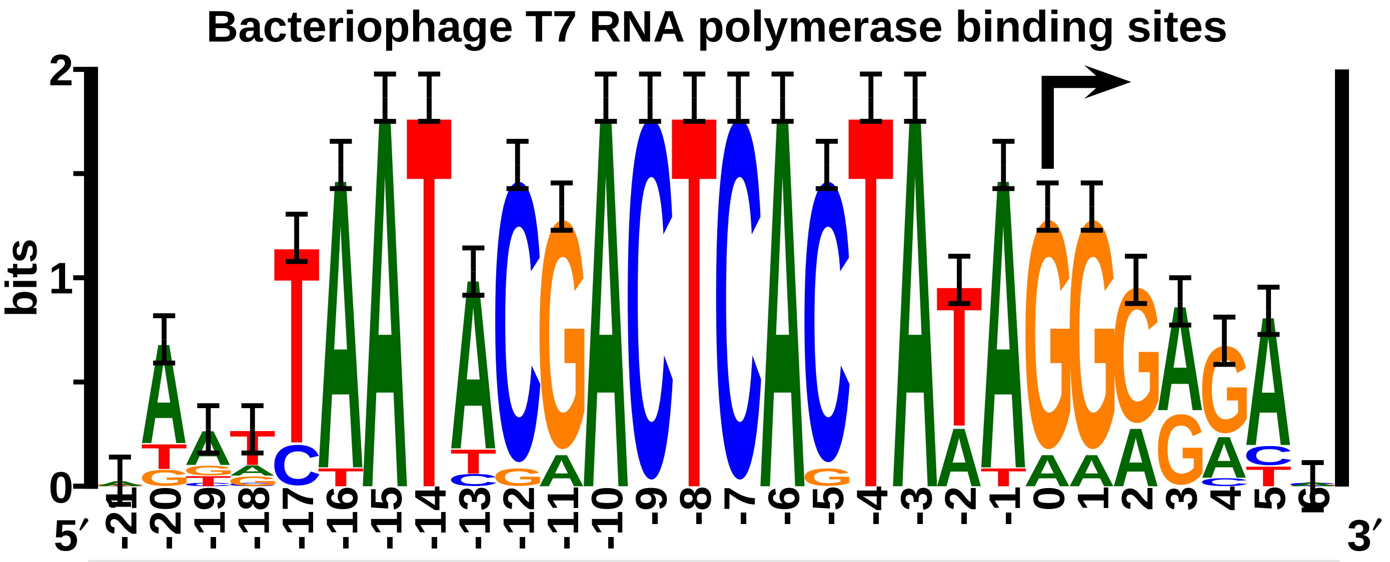
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	



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


Bacteriophage T7 RNA polymerase binding sites

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

1 ttattaatacaactcactataaggagag

2 aaatcaatacgactcactatagagggac

3 cggttaatacgactcactataggagaac

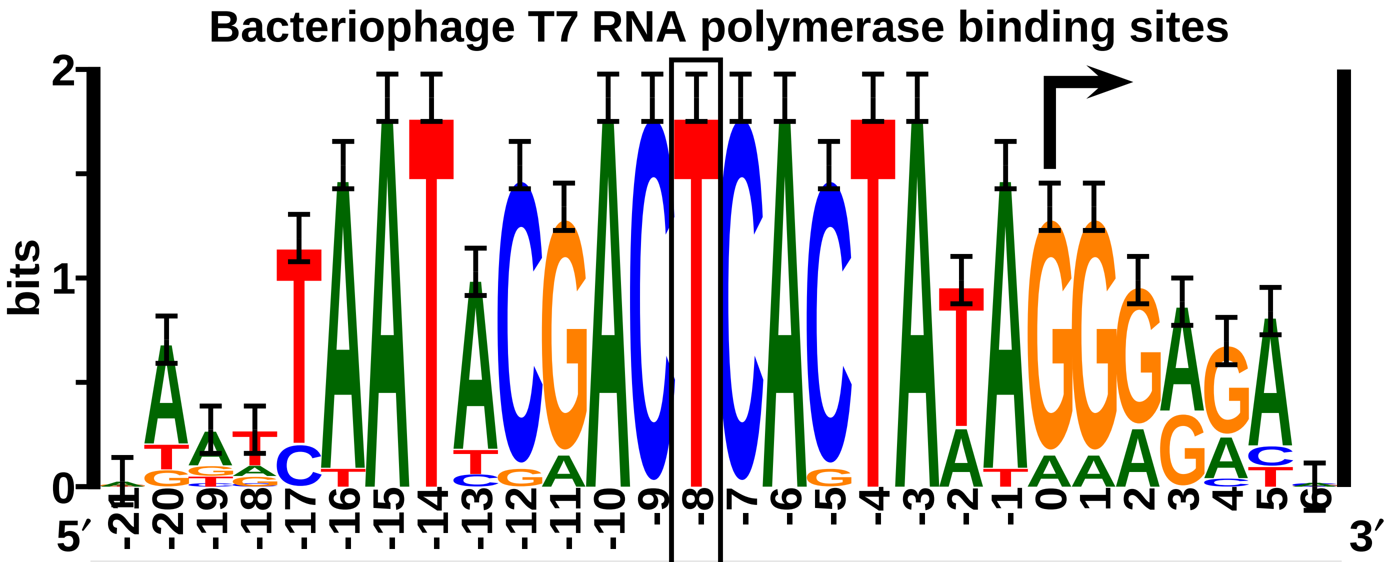
4 gaagtaatacgactcagtatagggacaa

5 taattaattgaactcactaaagggagac

6 cgcttaatacgactcactaaaggagaca

6 of 17 sites

Sequence Logo

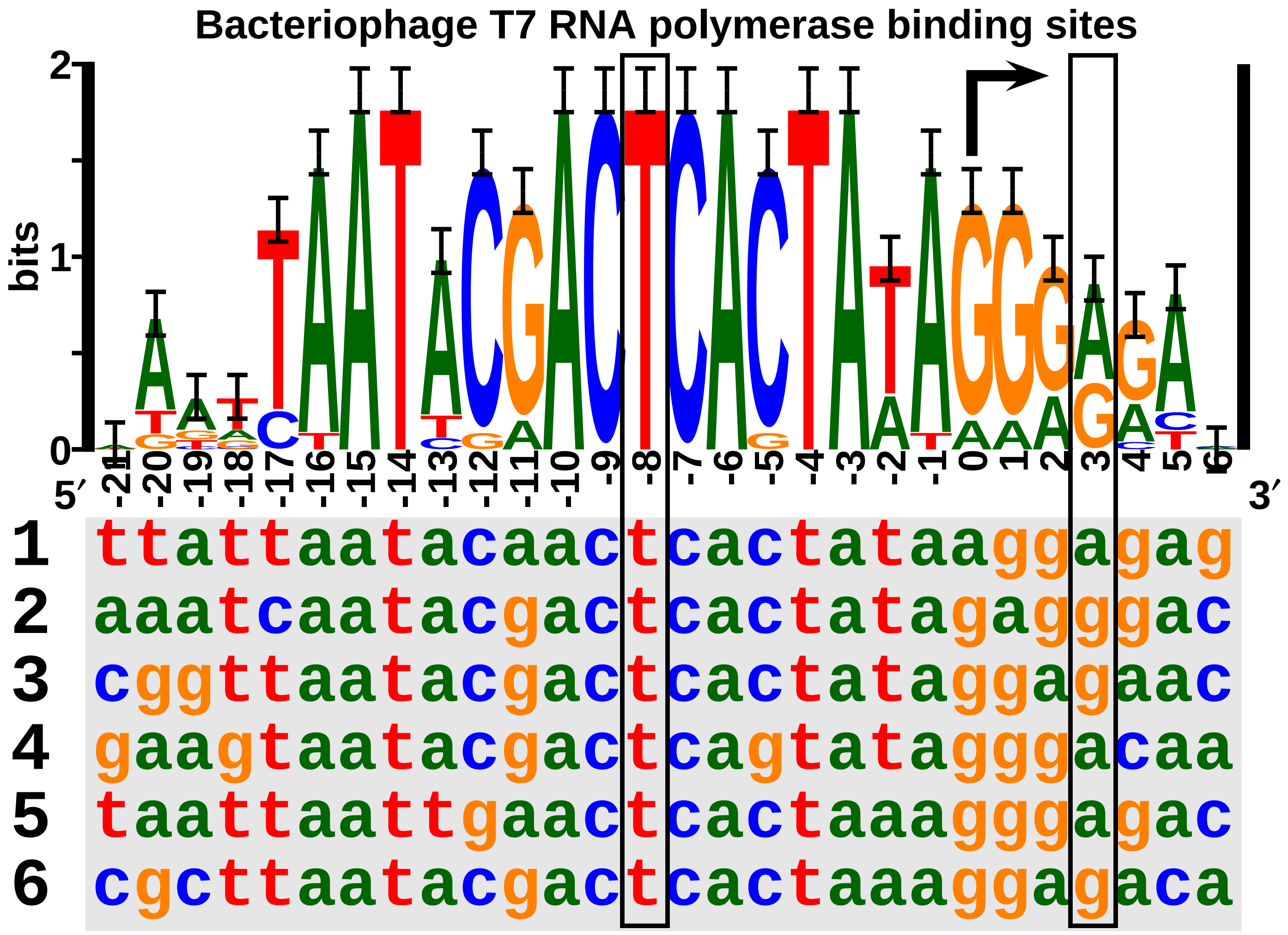


Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

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

6 of 17 sites

Sequence Logo


Bacteriophage T7 RNA polymerase binding sites

1 ttattaatacaactcactataaggagag

2 aaatcaatacgactcactatagagggac

3 cggttaatacgactcactataggagaac

4 gaagtaatacgactcagtatagggacaa

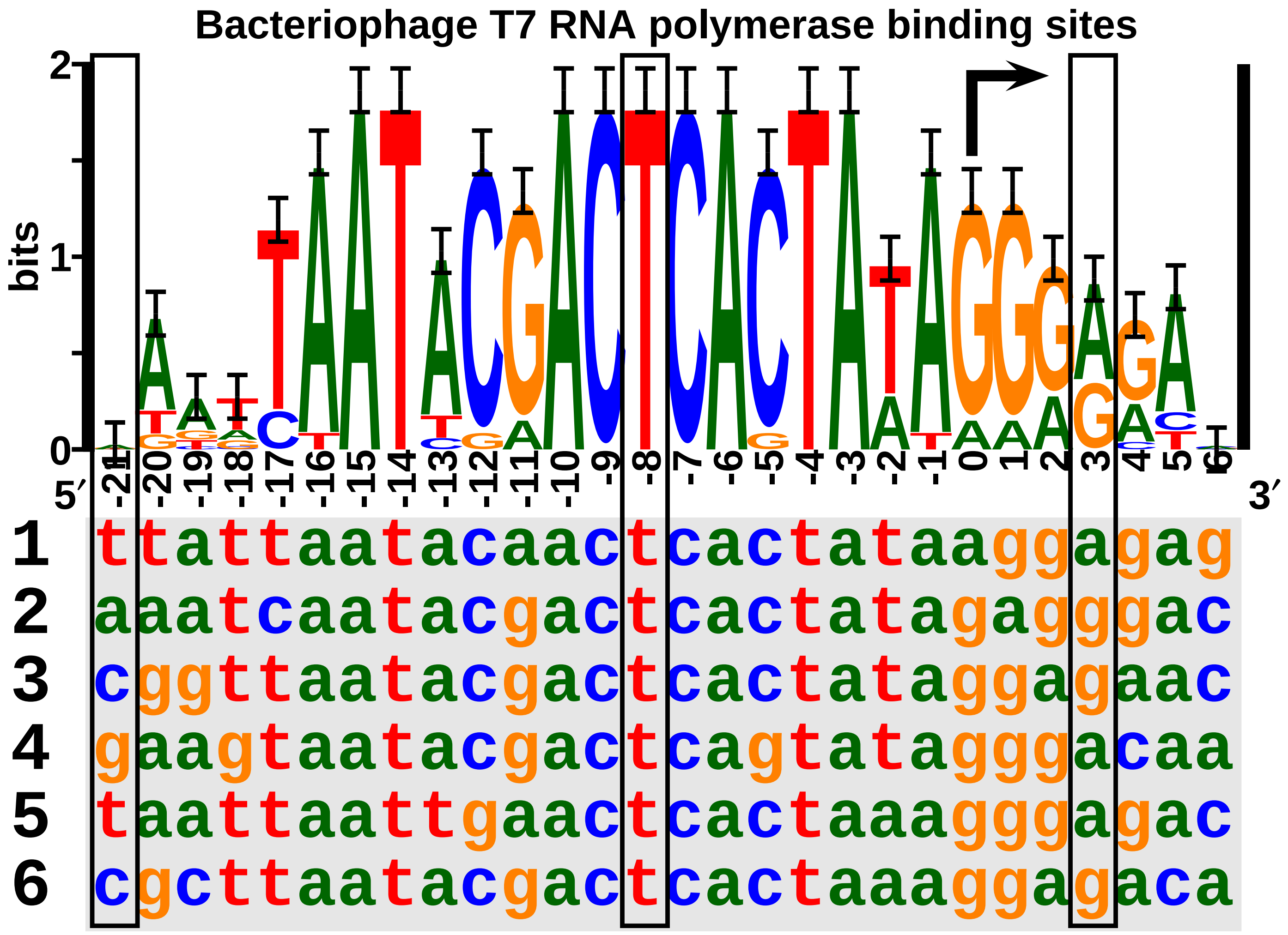
5 taattaattgaactcactaaagggagac

6 cgcttaatacgactcactaaaggagaca

6 of 17 sites

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

Sequence Logo



1 ttattaatacaactcactataaggagag

2 aaatcaatacgactcactatagagggac

3 cggttaatacgactcactataggagaac

4 gaagtaatacgactcagtatagggacaa

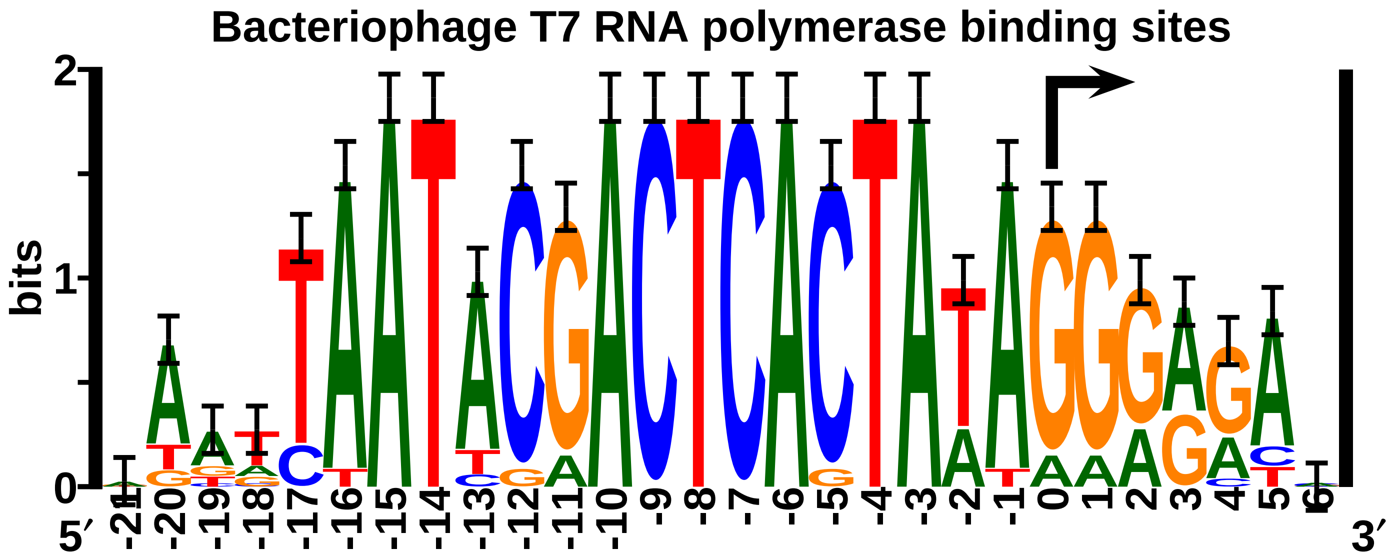
5 taattaattgaactcactaaagggagac

6 cgcttaatacgactcactaaaggagaca

6 of 17 sites

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

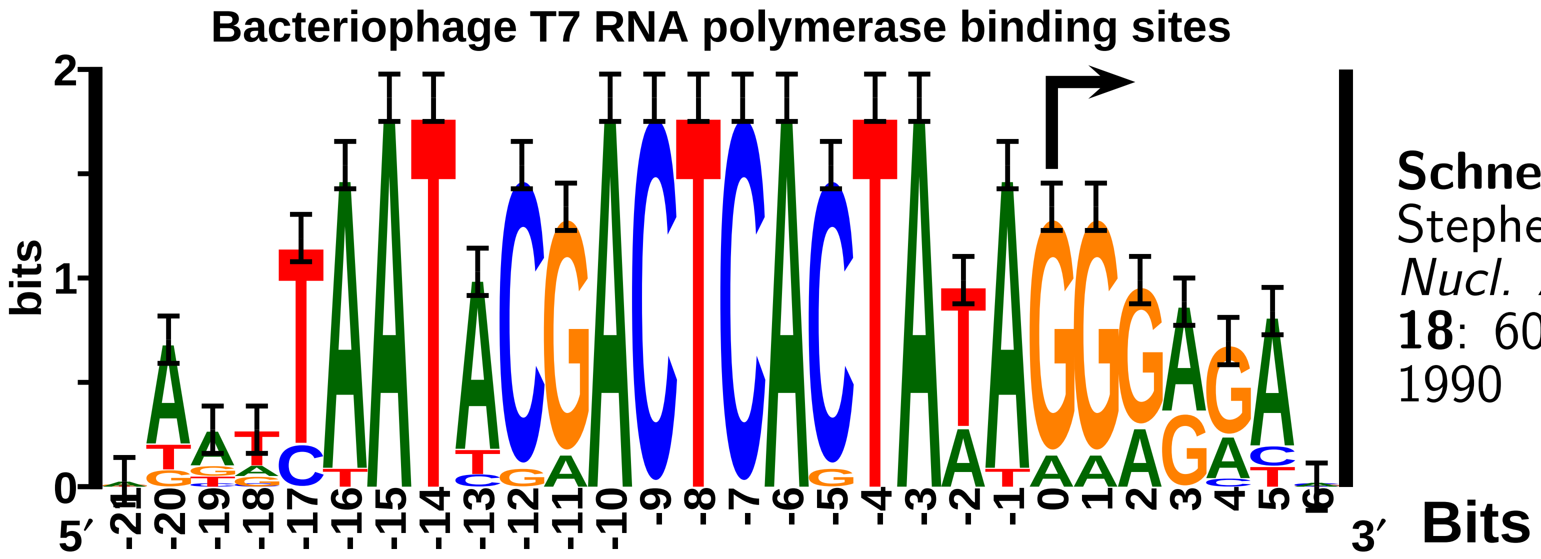
Sequence Logo and Sequence Walker



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



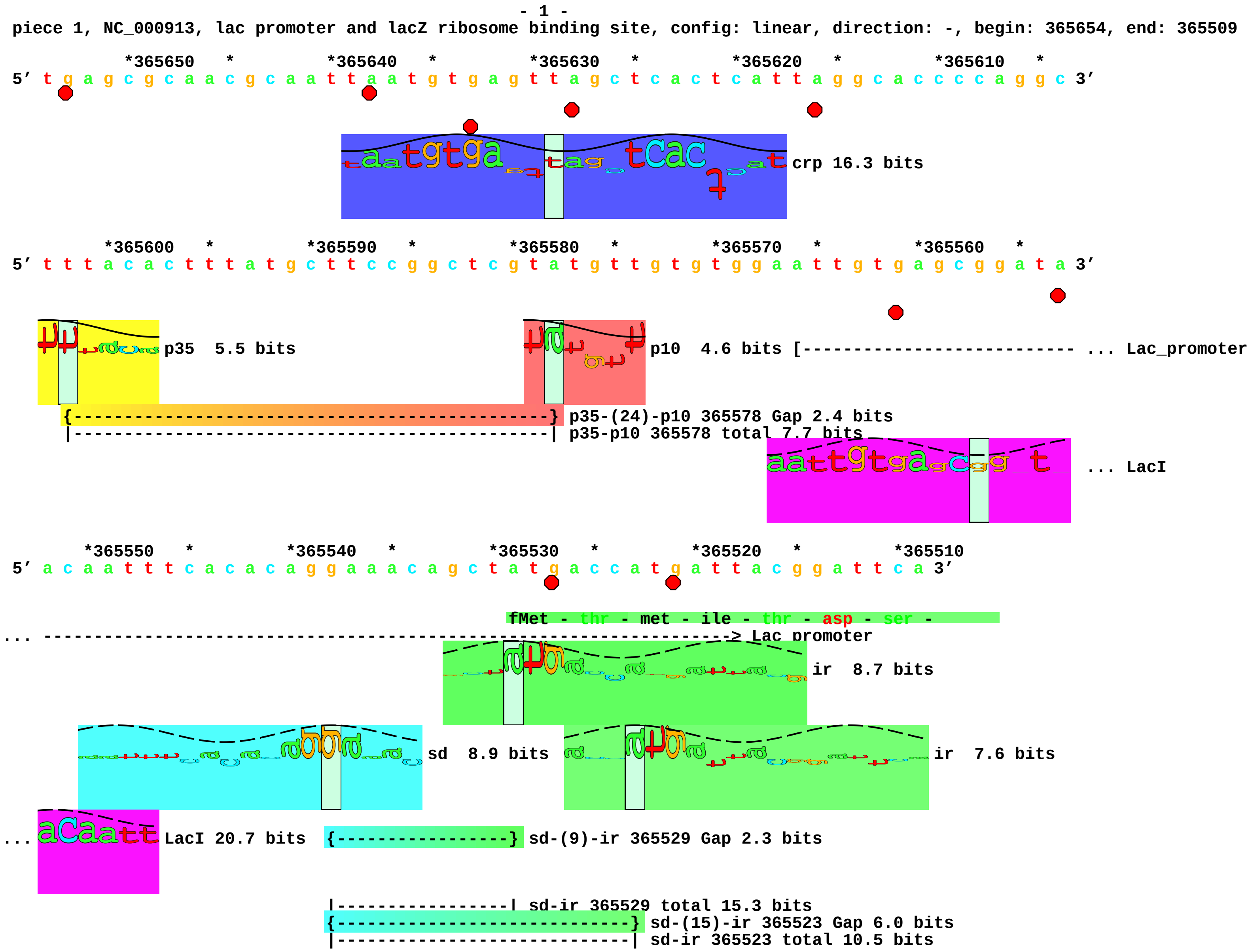
Bacteriophage T7 RNA polymerase binding sites

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 Walker
Patent
5,867,402

Sequence Walkers in the Lac Promoter



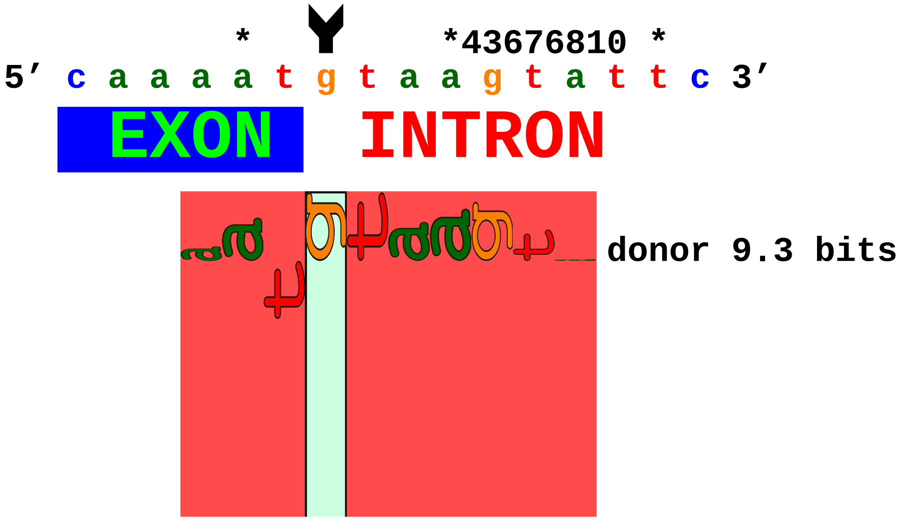
```
                                  - 1 -
piece 1, NC_000913, lac promoter and lacZ ribosome binding site, config: linear, direction: -, begin: 365654, end: 365509

          *365650    *         *365640    *         *365630    *         *365620    *         *365610    *
5' t g a g c g c a a c g c a a t t a a t g t g a g t t a g c t c a c t c a t t a g g c a c c c c a g g c 3'

                                                                                crp 16.3 bits

         *365600    *         *365590    *         *365580    *         *365570    *         *365560    *
5' t t t a c a c t t t a t g c t t c c g g c t c g t a t g t t g t g t g g a a t t g t g a g c g g a t a 3'

              p35   5.5 bits                      p10   4.6 bits [-------------------------- ... Lac_promoter
  {-----------------------------------------------} p35-(24)-p10 365578 Gap 2.4 bits
  |-----------------------------------------------| p35-p10 365578 total 7.7 bits
                                                                                   ... LacI

        *365550    *         *365540    *         *365530    *         *365520    *         *365510
5' a c a a t t t c a c a c a g g a a a c a g c t a t g a c c a t g a t t a c g g a t t c a 3'

                                            fMet - thr - met - ile - thr - asp - ser -
... ------------------------------------------------------------------> Lac promoter
                                                            ir   8.7 bits

                                sd   8.9 bits                                ir   7.6 bits

...         LacI 20.7 bits  {-----------------} sd-(9)-ir 365529 Gap 2.3 bits

                            |-----------------| sd-ir 365529 total 15.3 bits
                            {-----------------------------} sd-(15)-ir 365523 Gap 6.0 bits
                            |-----------------------------| sd-ir 365523 total 10.5 bits
```

Predicting splicing mutations using information theory

- Xeroderma Pigmentosum-Variant:
  defective postreplication repair
  predisposes to skin cancers
  on UV radiation

Predicting splicing mutations using information theory

- Xeroderma Pigmentosum-Variant: defective postreplication repair predisposes to skin cancers on UV radiation
- POLH exon 6 splice donor site



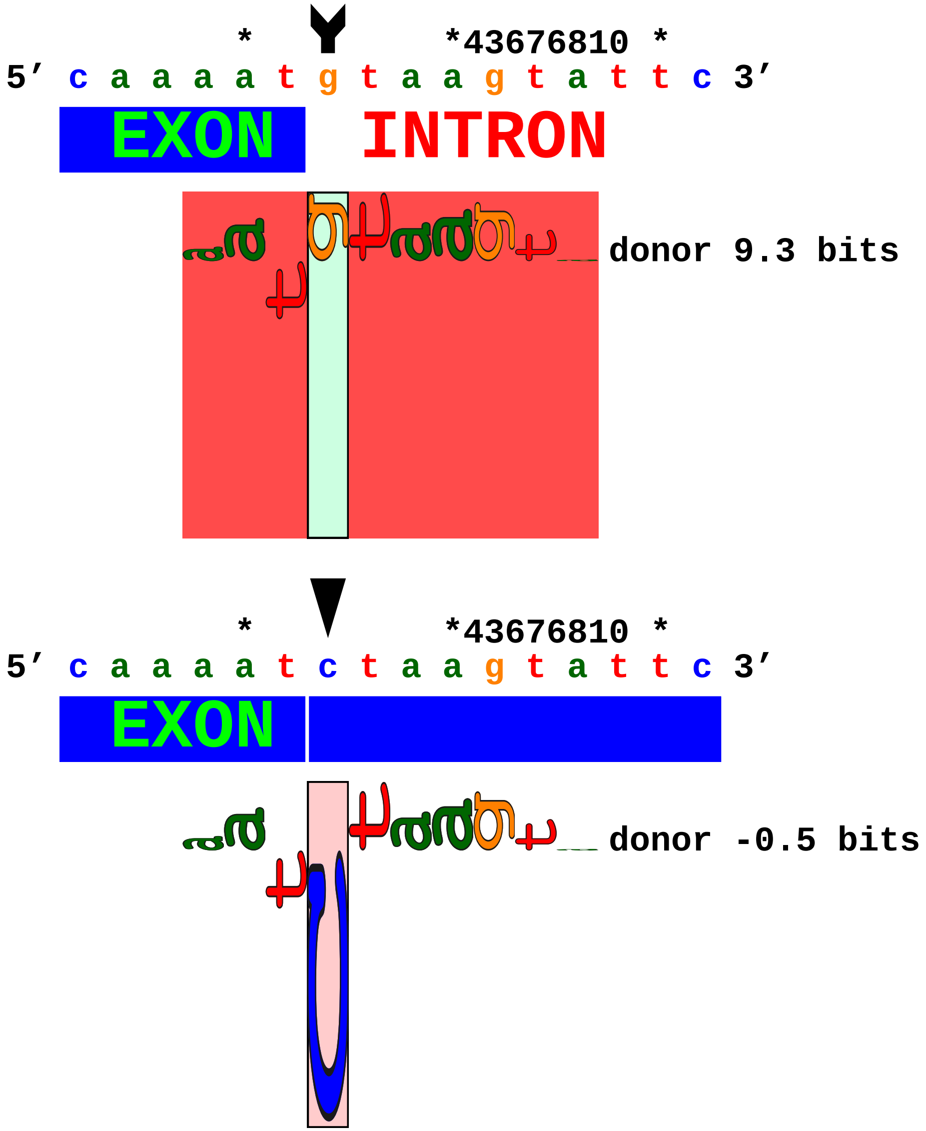
)

Predicting splicing mutations using information theory

- Xeroderma Pigmentosum-Variant: defective postreplication repair predisposes to skin cancers on UV radiation

- POLH exon 6 splice donor site

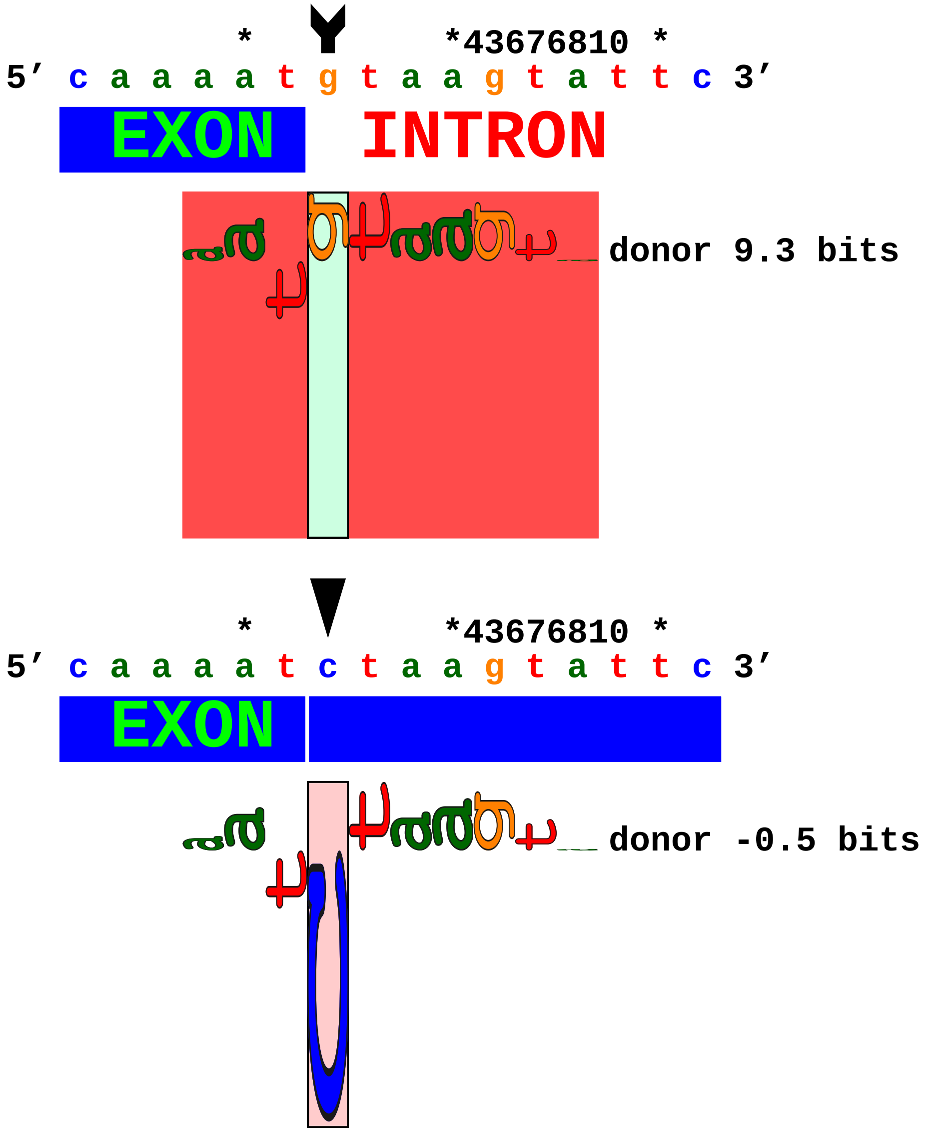
- G → C change observed in a patient. Can this explain the disease?



Inui, . . . , **Schneider** and Kraemer,
J Invest Dermatol. 128:2055-68 (2008)

Predicting splicing mutations using information theory

- Xeroderma Pigmentosum-Variant: defective postreplication repair predisposes to skin cancers on UV radiation
- POLH exon 6 splice donor site
- G → C change observed in a patient. Can this explain the disease?
- Second law of thermodynamics: < 0 bits is not a site



Inui, . . . , **Schneider** and Kraemer,
J Invest Dermatol. 128:2055-68 (2008)

Predicting splicing mutations using information theory

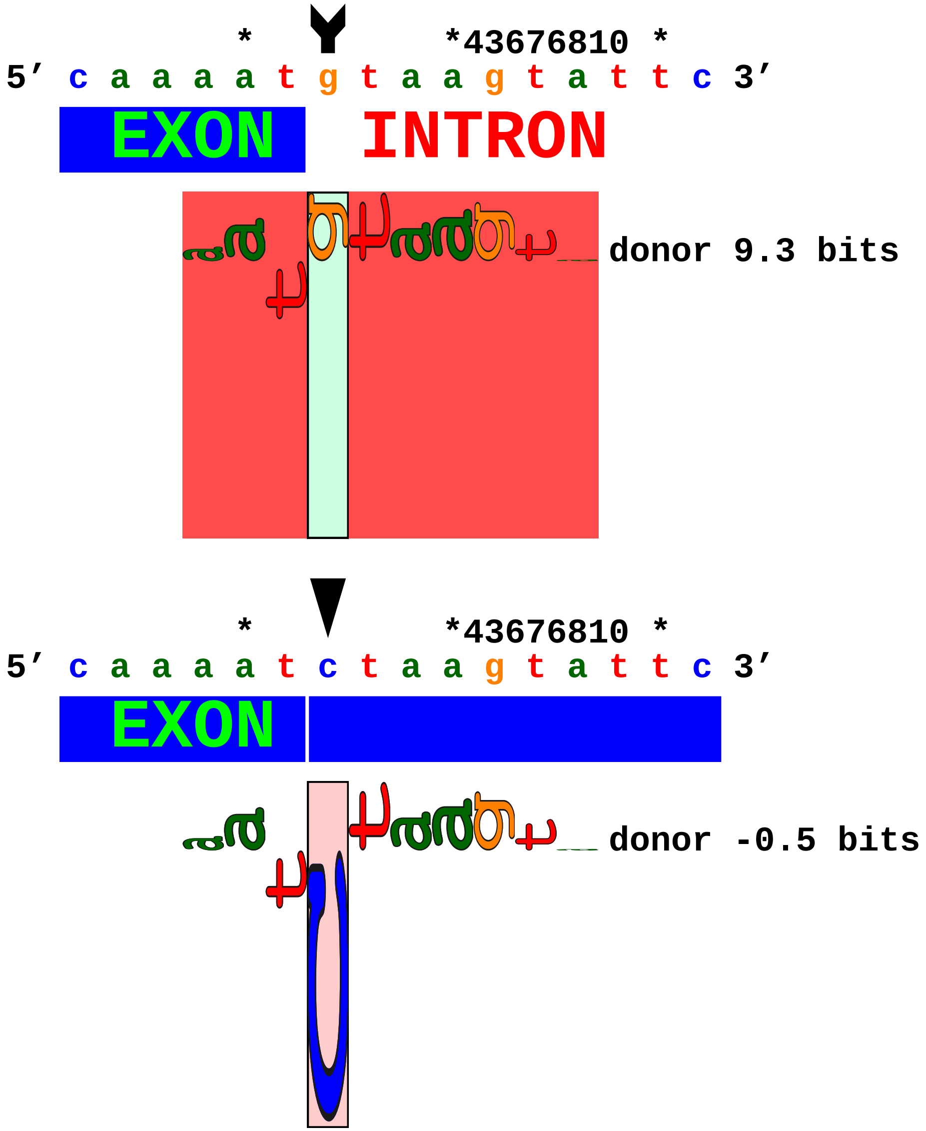
- Xeroderma Pigmentosum-Variant: defective postreplication repair predisposes to skin cancers on UV radiation

- POLH exon 6 splice donor site

- G $\rightarrow$ C change observed in a patient. Can this explain the disease?

- Second law of thermodynamics: < 0 bits is not a site

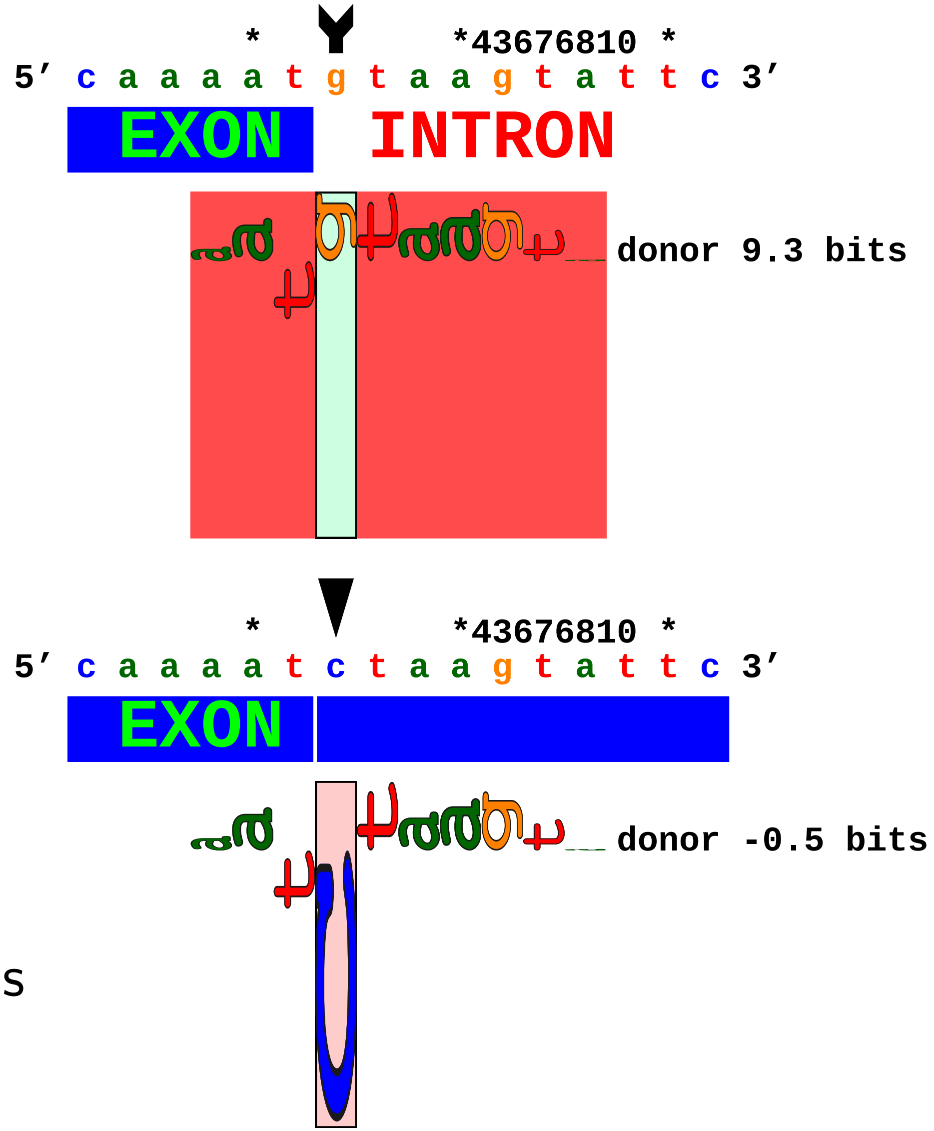
- Information theory explains the disease



Inui, . . . , **Schneider** and Kraemer,
J Invest Dermatol. 128:2055-68 (2008)

Predicting splicing mutations using information theory

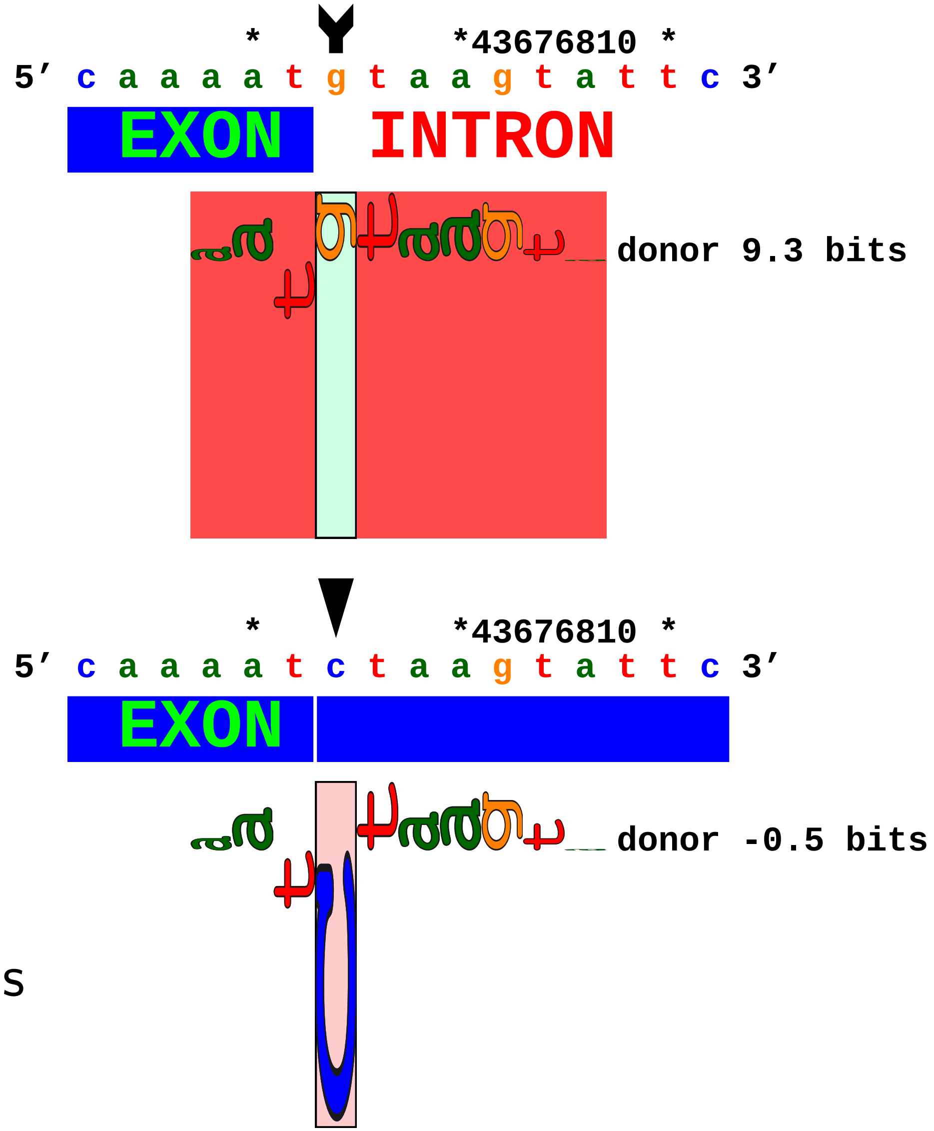
- Xeroderma Pigmentosum-Variant: defective postreplication repair predisposes to skin cancers on UV radiation
- POLH exon 6 splice donor site
- G → C change observed in a patient. Can this explain the disease?
- Second law of thermodynamics: < 0 bits is not a site
- Information theory explains the disease
- 175 papers published since 2004 using this technique



Inui, . . . , **Schneider** and Kraemer,
J Invest Dermatol. 128:2055-68 (2008)

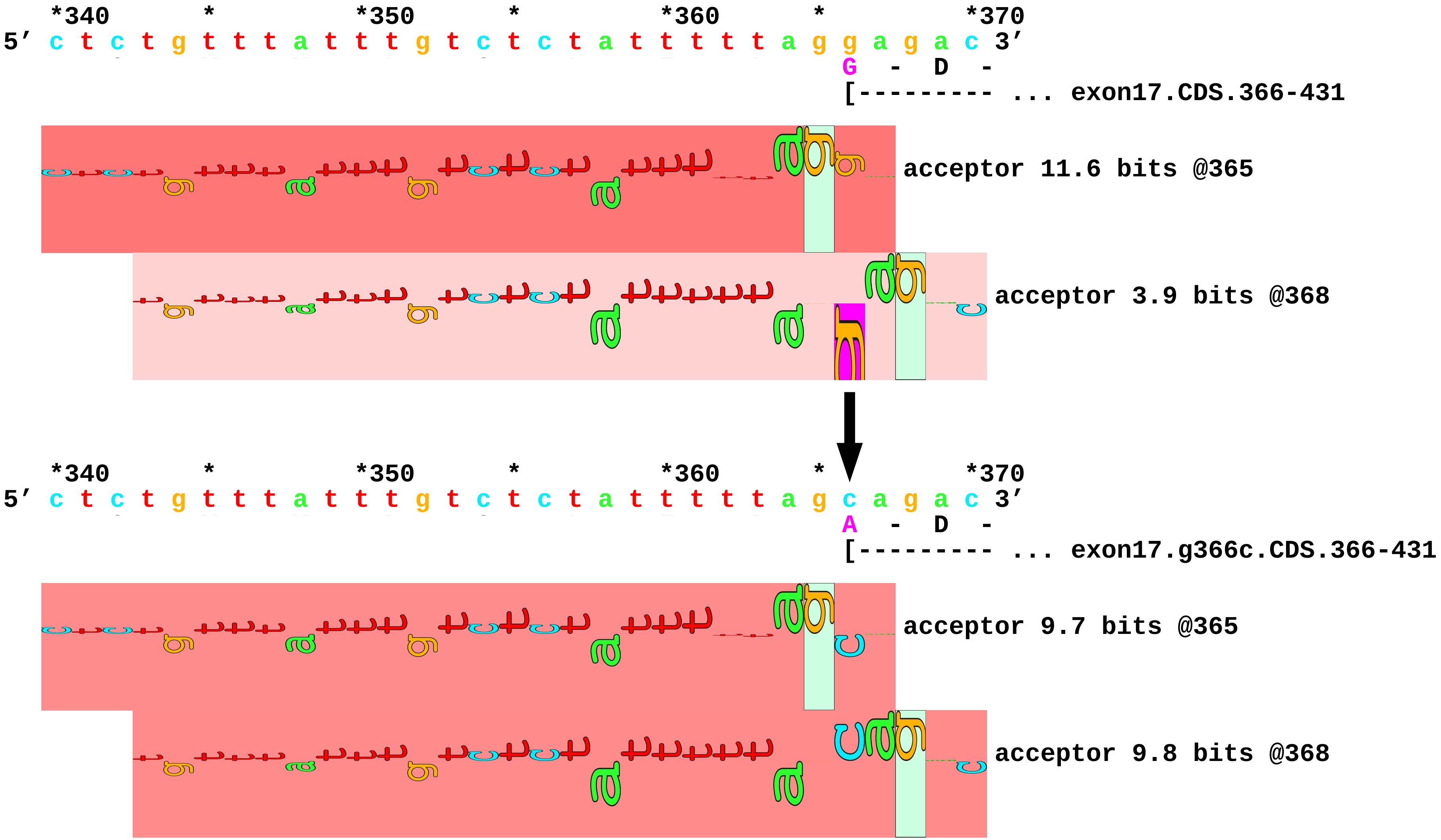
Predicting splicing mutations using information theory

- Xeroderma Pigmentosum-Variant: defective postreplication repair predisposes to skin cancers on UV radiation
- POLH exon 6 splice donor site
- G → C change observed in a patient. Can this explain the disease?
- Second law of thermodynamics: < 0 bits is not a site
- Information theory explains the disease
- 175 papers published since 2004 using this technique
- Collaborators:
  Dr. Kenneth Kraemer (NIH, NCI, CCR)
  Dr. Peter Rogan (Univ. Western Ontario)



Inui, . . . , **Schneider** and Kraemer, J Invest Dermatol. 128:2055-68 (2008)

Medical Applications of Sequence Walkers



Mutation G863A: Stargardt disease = age-related macular degeneration

http://alum.mit.edu/www/toms/g863a.html

Discovering p53/p63/p73 controlled genes

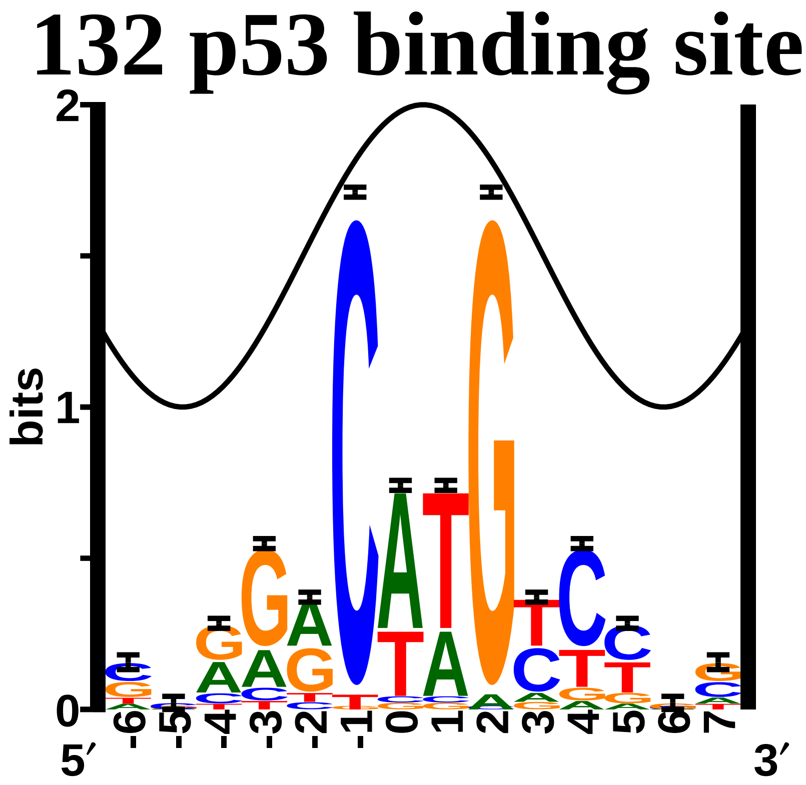
- p53 - transcriptional regulator controlling cell cycle

Discovering p53/p63/p73 controlled genes

- p53 - transcriptional regulator controlling cell cycle
- p53 signaling is inactivated in 50% of human cancers

Discovering p53/p63/p73 controlled genes

- p53 - transcriptional regulator controlling cell cycle
- p53 signaling is inactivated in 50% of human cancers
- Natural model built from proven sites



Lyakhov, Annangarachari and **Schneider**
Nucleic Acids Res. 36:3828-33 (2008)

Discovering p53/p63/p73 controlled genes

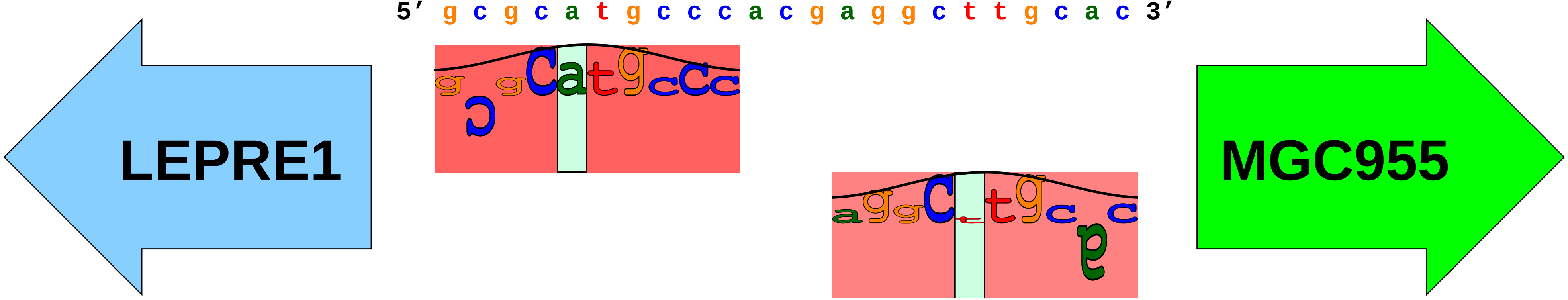
- Natural model was used to **predict 16 previously unidentified p53 controlled genes** on human chromosomes 1 and 2

Discovering p53/p63/p73 controlled genes

- Natural model was used to **predict 16 previously unidentified p53 controlled genes** on human chromosomes 1 and 2
- **15 novel genes confirmed** by EMSA, promoter assays, qPCR. Controlled by p53 or related family members p63 or p73.

Discovering p53/p63/p73 controlled genes

- Natural model was used to **predict 16 previously unidentified p53 controlled genes** on human chromosomes 1 and 2
- **15 novel genes confirmed** by EMSA, promoter assays, qPCR. Controlled by p53 or related family members p63 or p73.
- LEPRE1 and MGC955 dual promoter



Lyakhov, Annangarachari and **Schneider**
Nucleic Acids Res. 36:3828-33 (2008)

Discovery of a 7^{th} Bacteriophage λ Operator

- **Bacteriophage λ - a paradigm for gene control > 50 years**

Discovery of a 7^{th} Bacteriophage λ Operator

- **Bacteriophage λ - a paradigm for gene control > 50 years**

a good testing ground for theory

Discovery of a 7^{th} Bacteriophage λ Operator

- Bacteriophage λ - a paradigm for gene control > 50 years

a good testing ground for theory

- Only 6 known λ operators, bound by CI and Cro proteins

Discovery of a 7^{th} Bacteriophage λ Operator

- **Bacteriophage λ - a paradigm for gene control > 50 years**

a good testing ground for theory

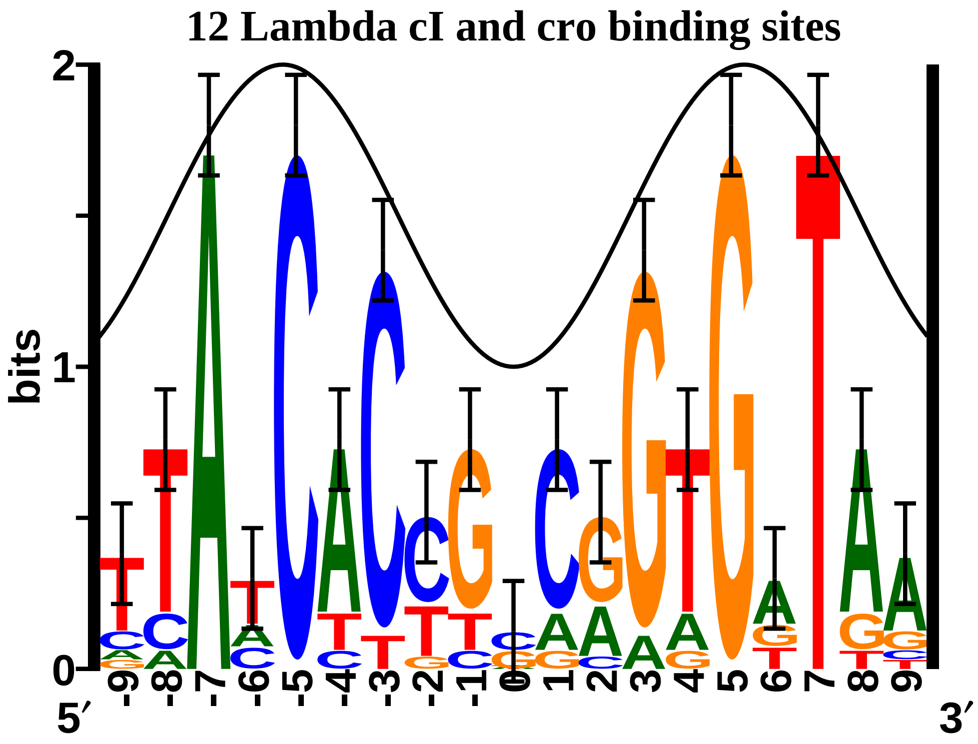
- **Only 6 known λ operators, bound by CI and Cro proteins**
- **Using the consensus and mismatch counting: cannot find more**

Discovery of a 7^{th} Bacteriophage λ Operator

- Bacteriophage λ - a paradigm for gene control > 50 years

a good testing ground for theory

- Only 6 known λ operators, bound by CI and Cro proteins
- Using the consensus and mismatch counting: cannot find more
- Make a sequence logo

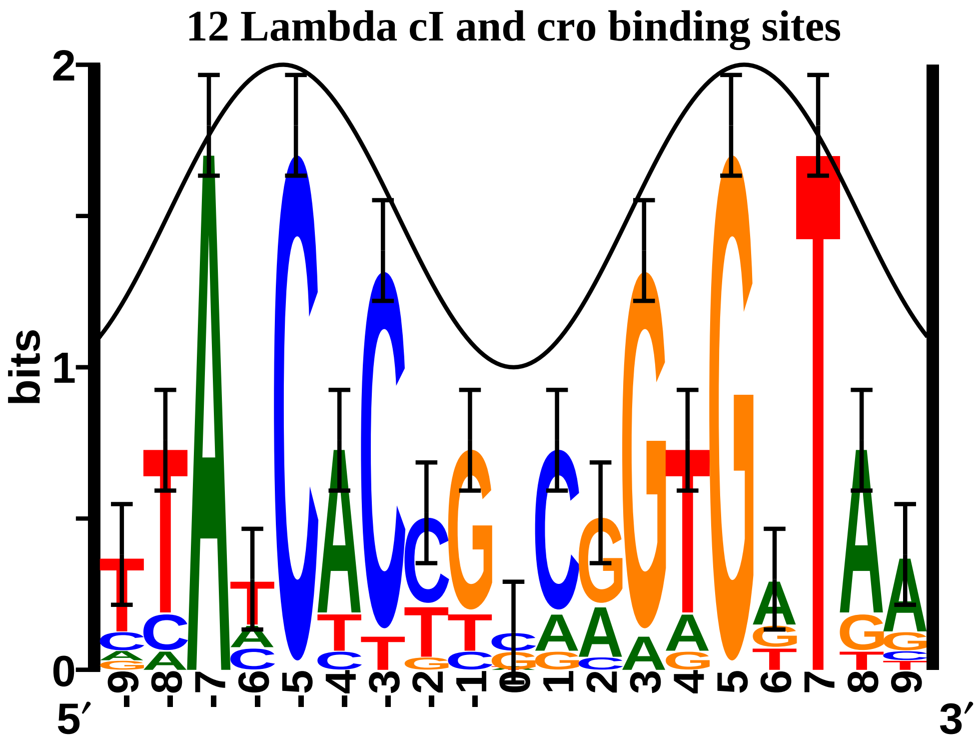


Discovery of a 7^{th} Bacteriophage λ Operator

- Bacteriophage λ - a paradigm for gene control > 50 years

a good testing ground for theory

- Only 6 known λ operators, bound by CI and Cro proteins
- Using the consensus and mismatch counting: cannot find more
- Make a sequence logo



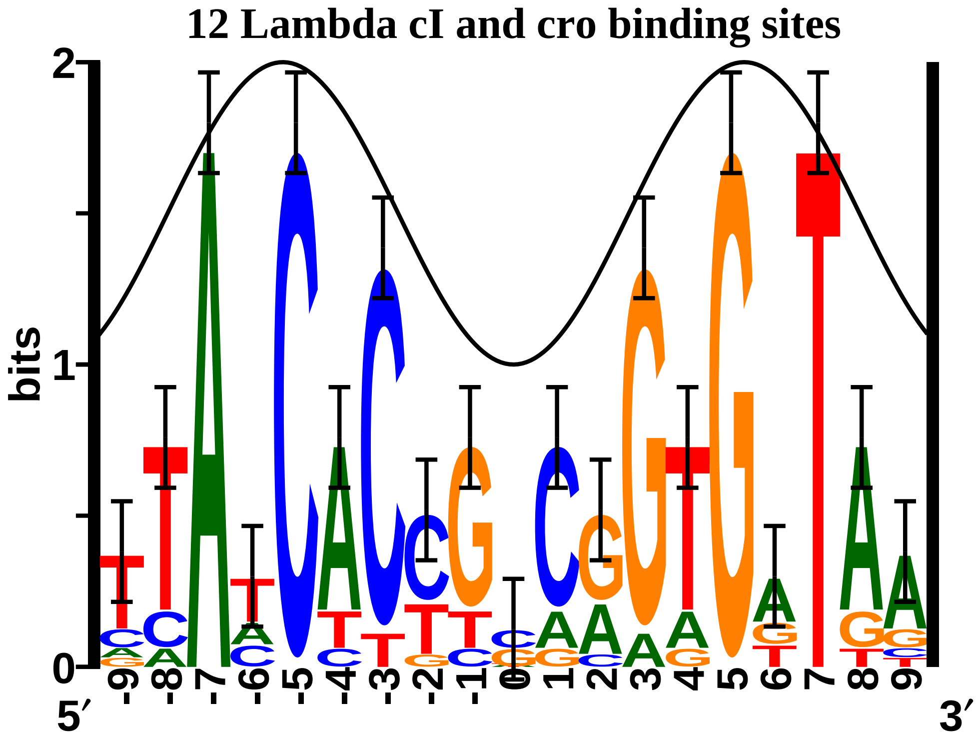
- Search λ using information theory model

Discovery of a 7^{th} Bacteriophage λ Operator

- Bacteriophage λ - a paradigm for gene control > 50 years

a good testing ground for theory

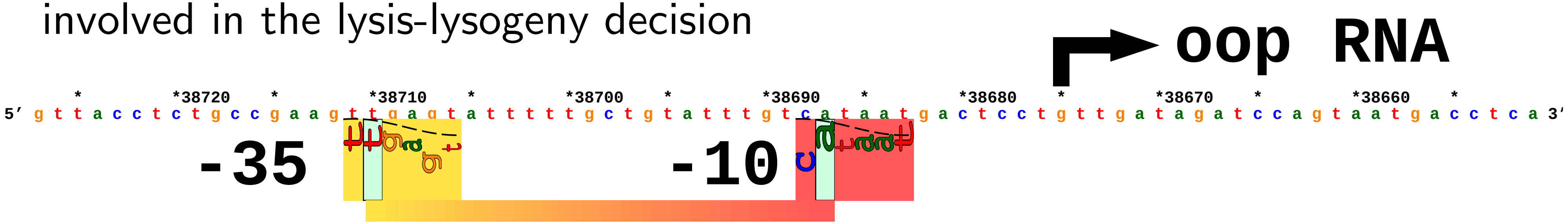
- Only 6 known λ operators, bound by CI and Cro proteins
- Using the consensus and mismatch counting: cannot find more
- Make a sequence logo

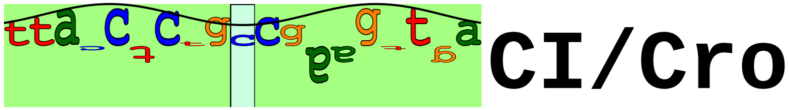


- Search λ using information theory model
- A 7^{th} Operator found!

Bacteriophage λ Oop promoter: controlled by CI/Cro?

oop RNA is antisense to the $3'$ end of *cII* mRNA
involved in the lysis-lysogeny decision

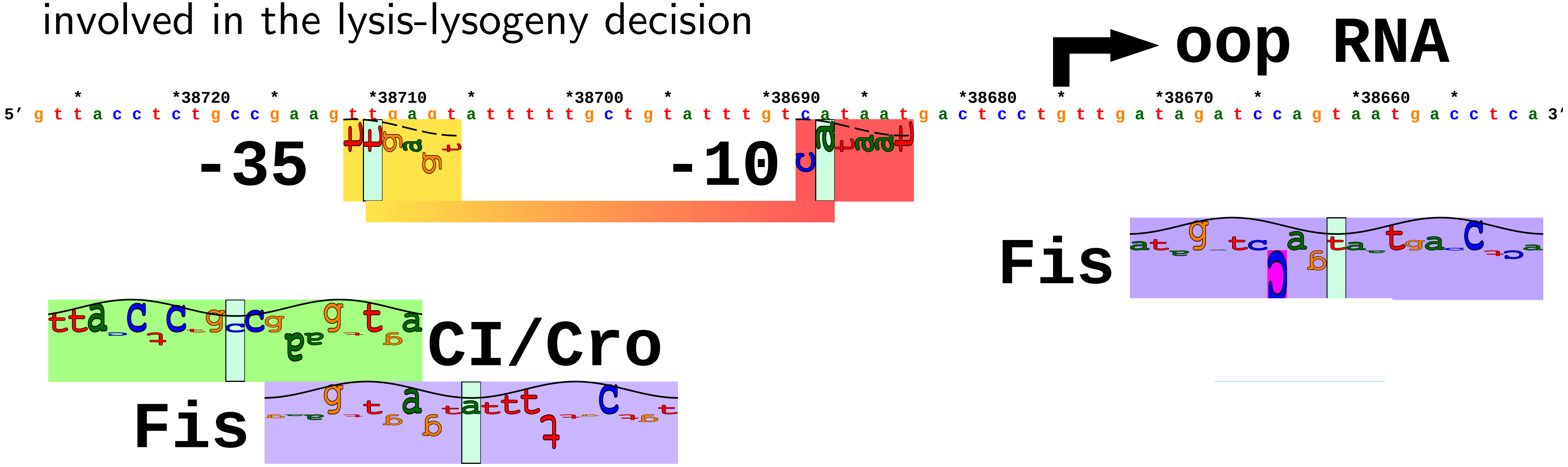




CI/Cro	λ switch to lytic growth	predicted

Bacteriophage λ Oop promoter: controlled by CI/Cro?

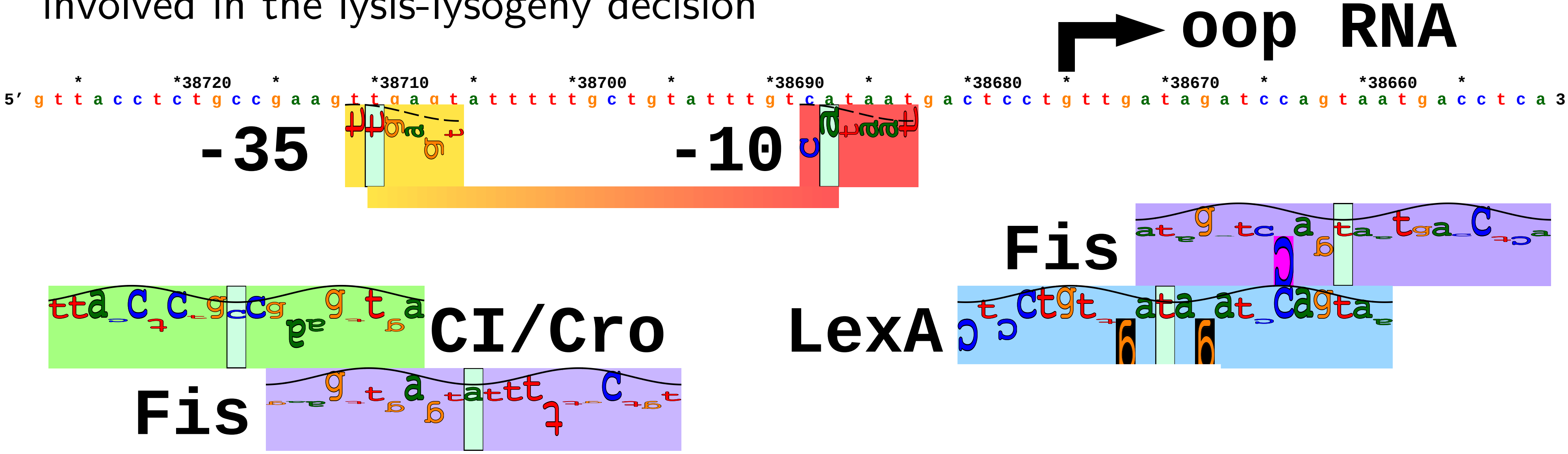
oop RNA is antisense to the $3'$ end of *cII* mRNA
involved in the lysis-lysogeny decision



CI/Cro	λ switch to lytic growth	predicted
Fis	Nutrients	predicted

Bacteriophage λ Oop promoter: controlled by CI/Cro?

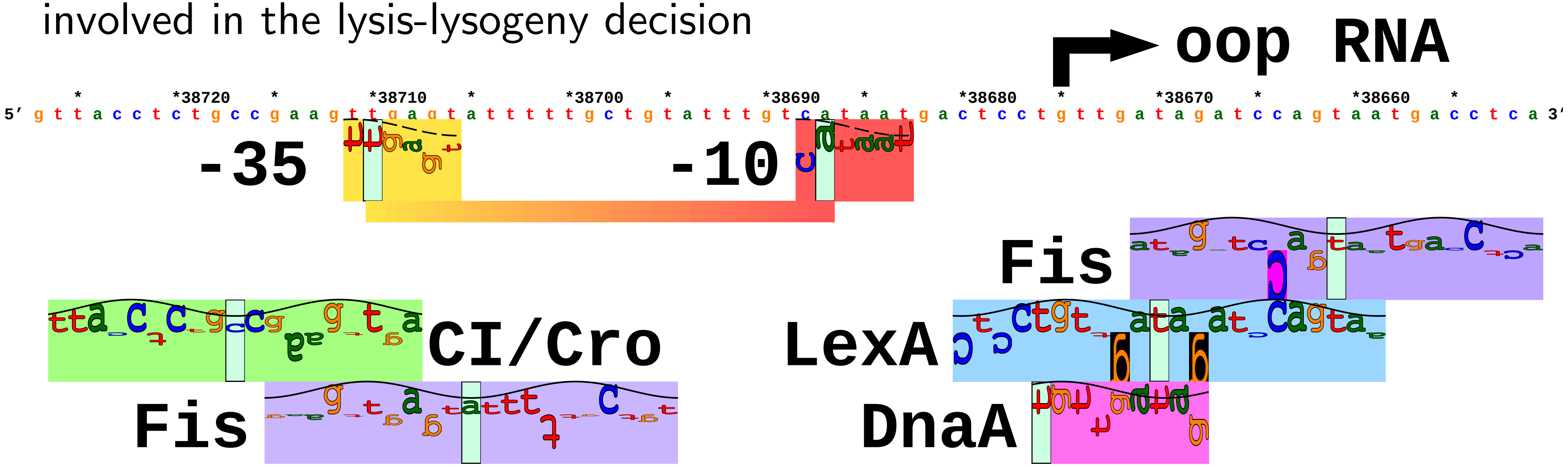
oop RNA is antisense to the $3'$ end of *cII* mRNA
involved in the lysis-lysogeny decision



CI/Cro	λ switch to lytic growth	predicted
Fis	Nutrients	predicted
LexA	DNA Damage	known

Bacteriophage λ Oop promoter: controlled by CI/Cro?

oop RNA is antisense to the $3'$ end of *cII* mRNA
involved in the lysis-lysogeny decision

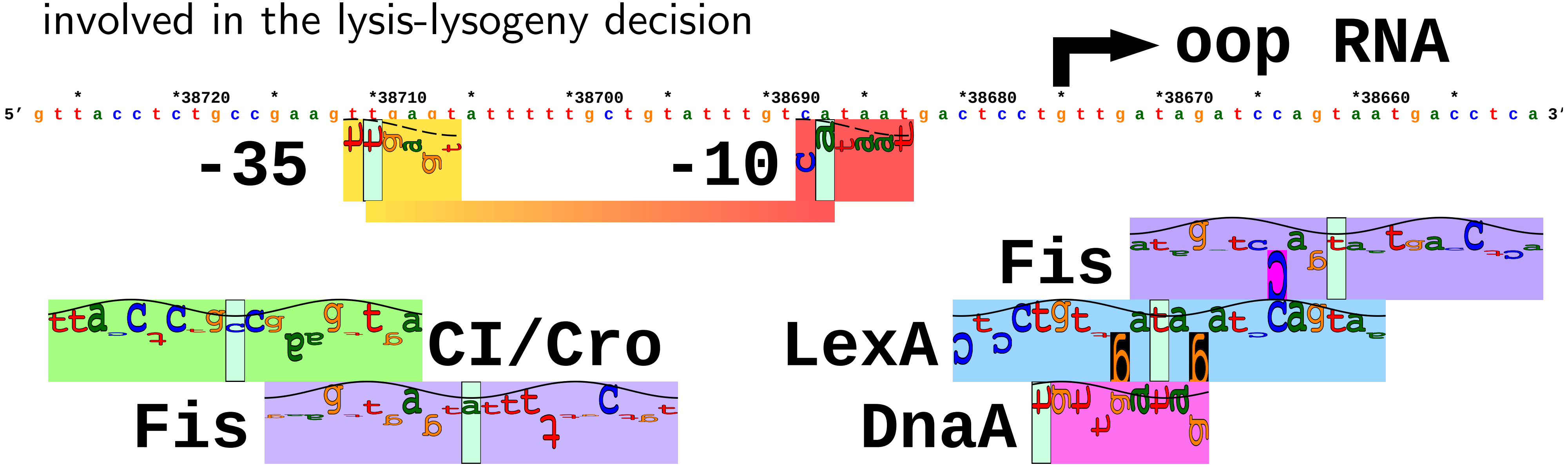


CI/Cro	λ switch to lytic growth	predicted
Fis	Nutrients	predicted
LexA	DNA Damage	known
DnaA	Cell replication	predicted

- **4 new sites predicted by information theory**

Bacteriophage λ Oop promoter: controlled by CI/Cro?

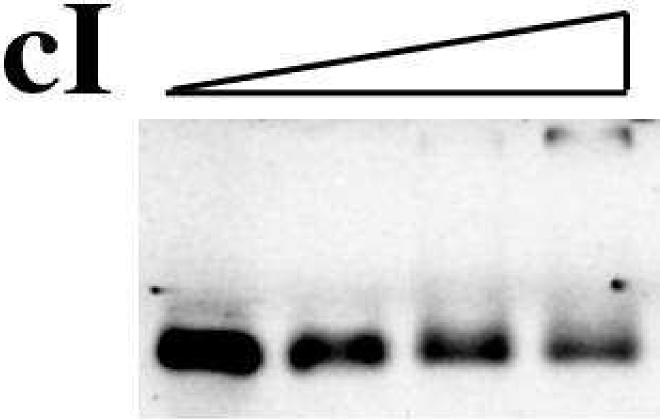
oop RNA is antisense to the $3'$ end of *cII* mRNA
involved in the lysis-lysogeny decision



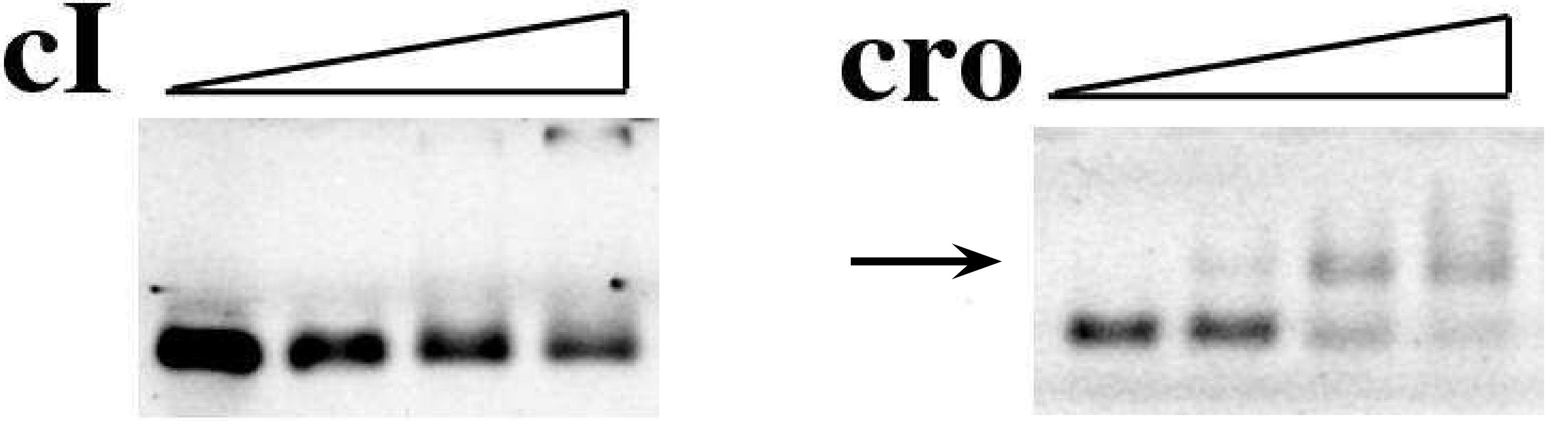
CI/Cro	λ switch to lytic growth	predicted
Fis	Nutrients	predicted
LexA	DNA Damage	known
DnaA	Cell replication	predicted

- **4 new sites predicted by information theory**
- **A cell-state detection/control center?**

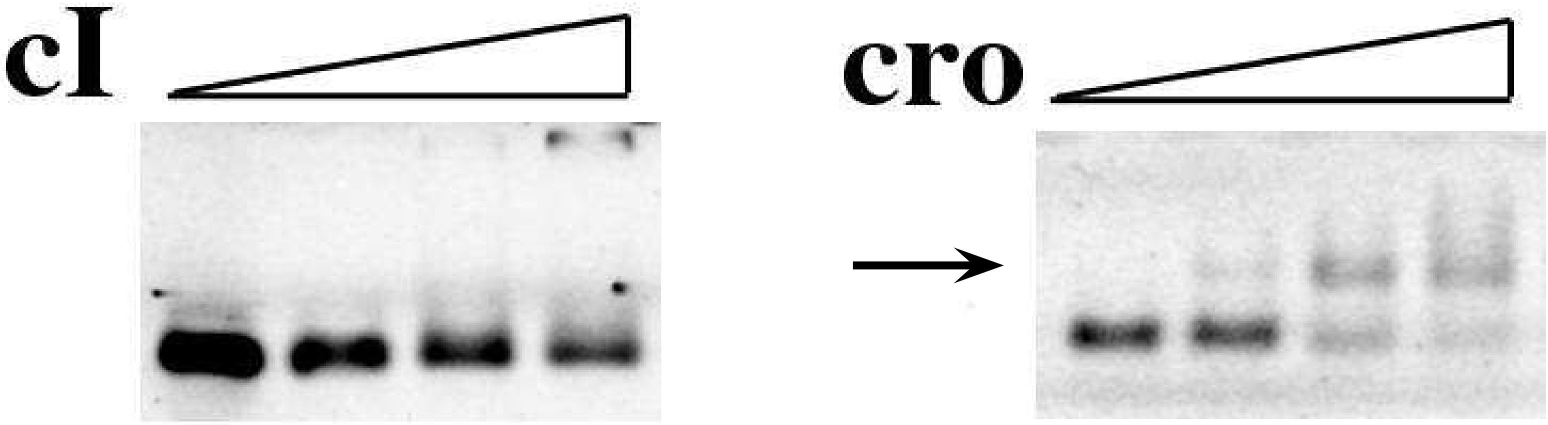
Bacteriophage λ Oop promoter Cro site



Bacteriophage λ Oop promoter Cro site

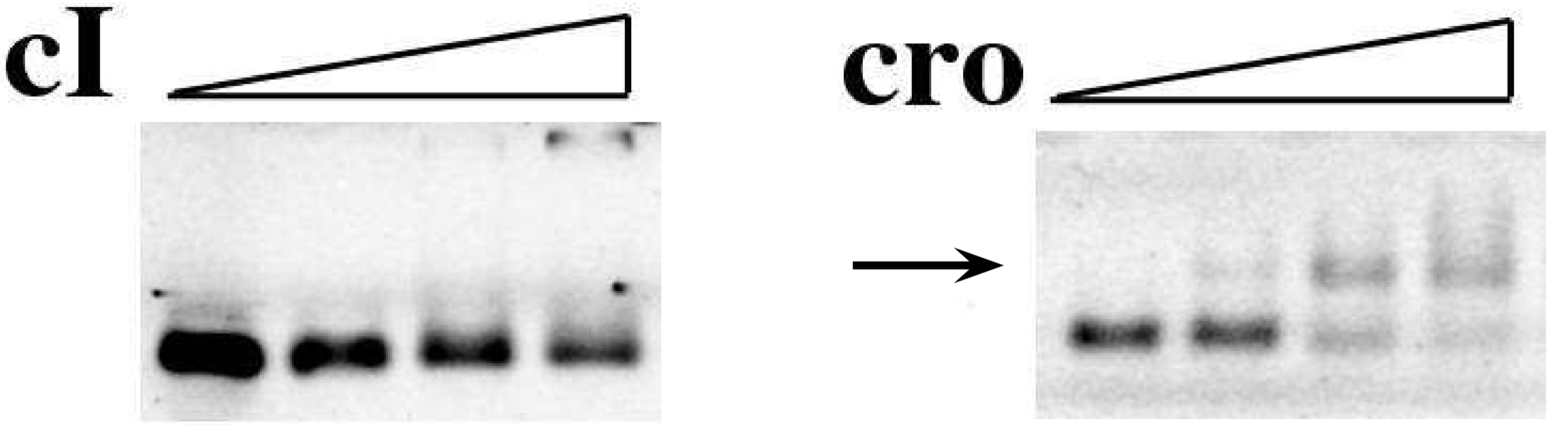


Bacteriophage λ Oop promoter Cro site



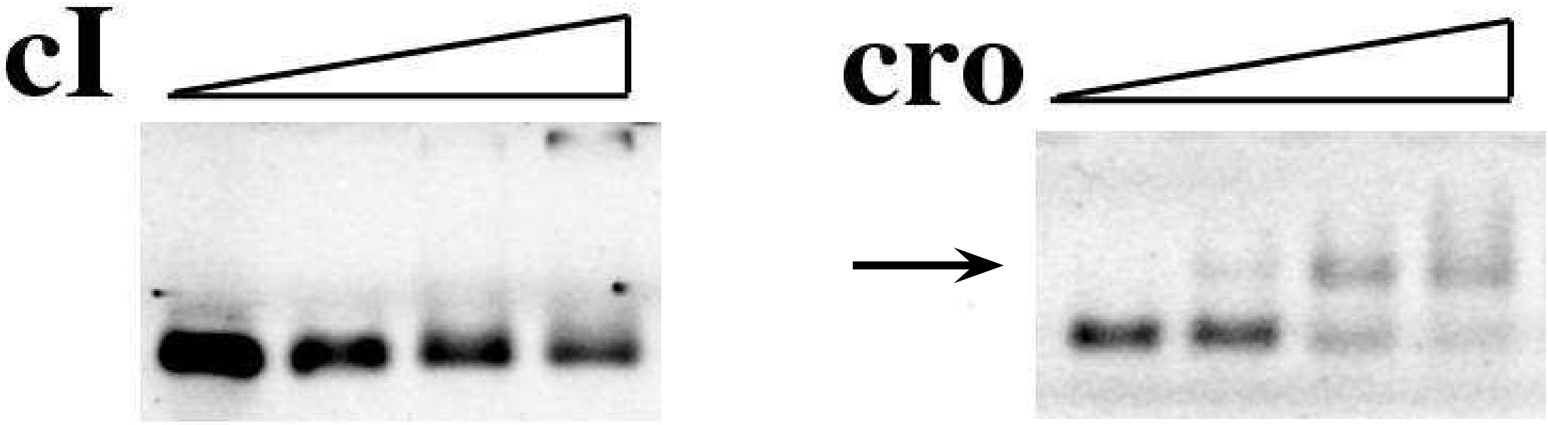
- **The 7th λ Operator is a Cro site, NOT a CI site.** $\boxed{\textbf{Prediction confirmed}}$

Bacteriophage λ Oop promoter Cro site



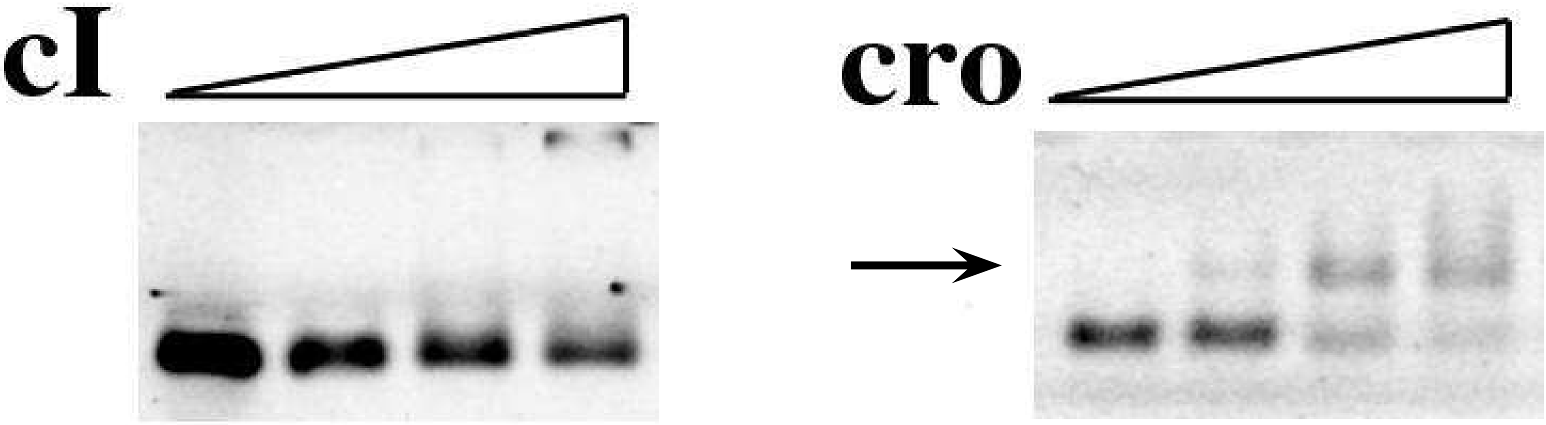
- The 7th λ Operator is a Cro site, NOT a CI site. Prediction confirmed
- Test information theory predictions to further confirm theory

Bacteriophage λ Oop promoter Cro site



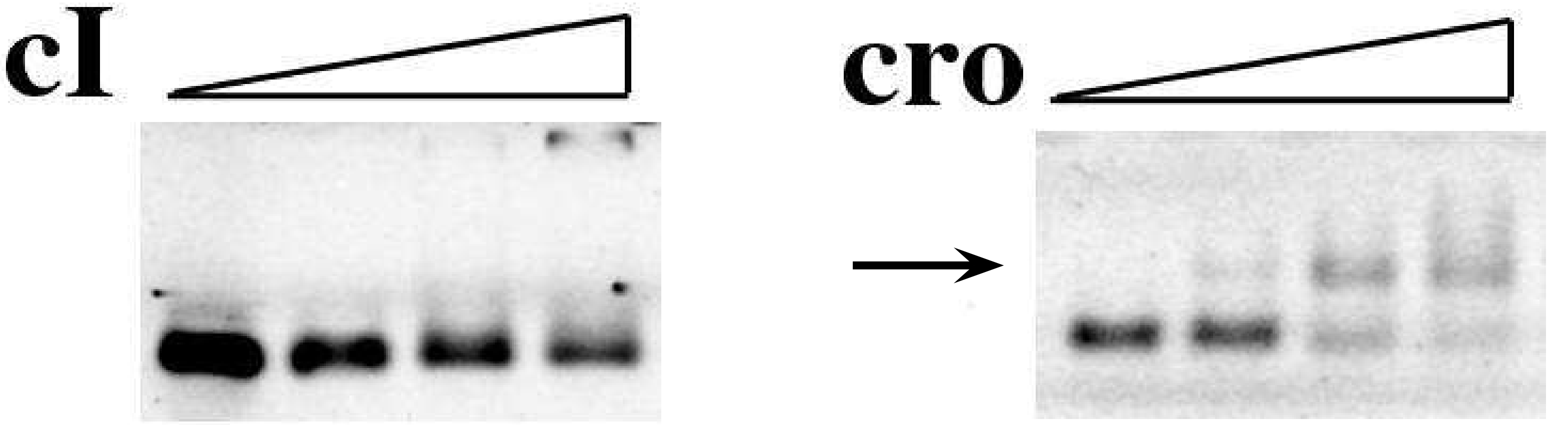
- The 7th λ Operator is a Cro site, NOT a CI site. Prediction confirmed
- Test information theory predictions to further confirm theory
- *in vivo* experiments in progress: knock out Cro and Fis sites

Bacteriophage λ Oop promoter Cro site



- The 7th λ Operator is a Cro site, NOT a CI site. $\boxed{\text{Prediction confirmed}}$
- Test information theory predictions to further confirm theory
- *in vivo* experiments in progress: knock out Cro and Fis sites
- live phage knockouts by recombineering planned: affects lysogeny?

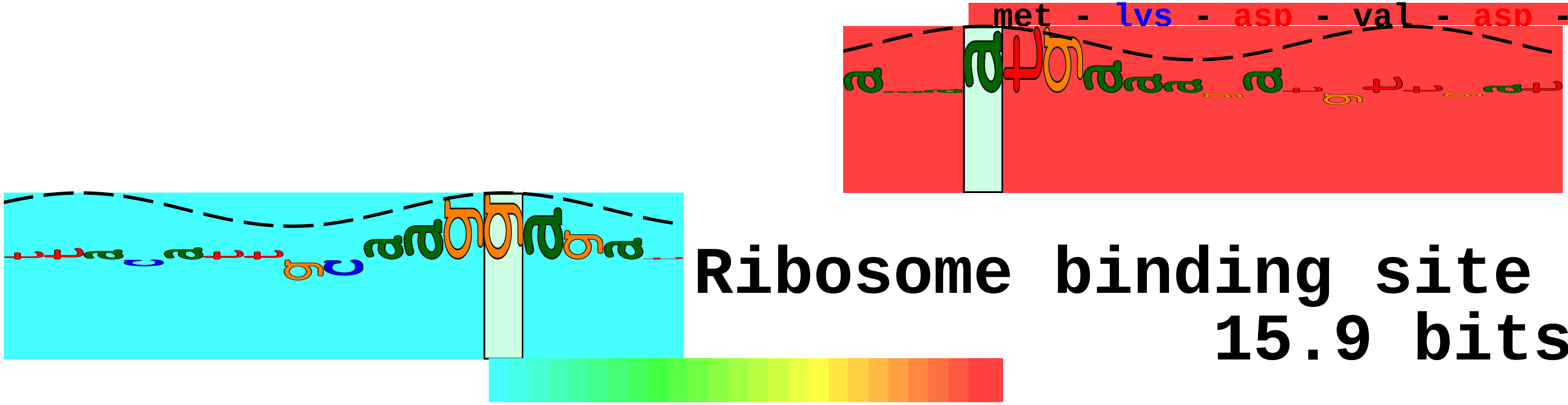
Bacteriophage λ Oop promoter Cro site



- The 7th λ Operator is a Cro site, NOT a CI site. Prediction confirmed
- Test information theory predictions to further confirm theory
- *in vivo* experiments in progress: knock out Cro and Fis sites
- live phage knockouts by recombineering planned: affects lysogeny?
- Collaborator: Dr. Don Court (NCI) - λ expert, invented recombineering

Predicting Small Open Reading Frames in *E. coli*

```
   *         *2987940   *         *2987930   *         *2987920   *         *2987910   *
5′ t t a c a t t g c a a g g a g a t g t c t a a a a t g a a a g a t g t t g a t c a a a t c 3′
                                                       met - lys - asp - val - asp - gln - ile -
```



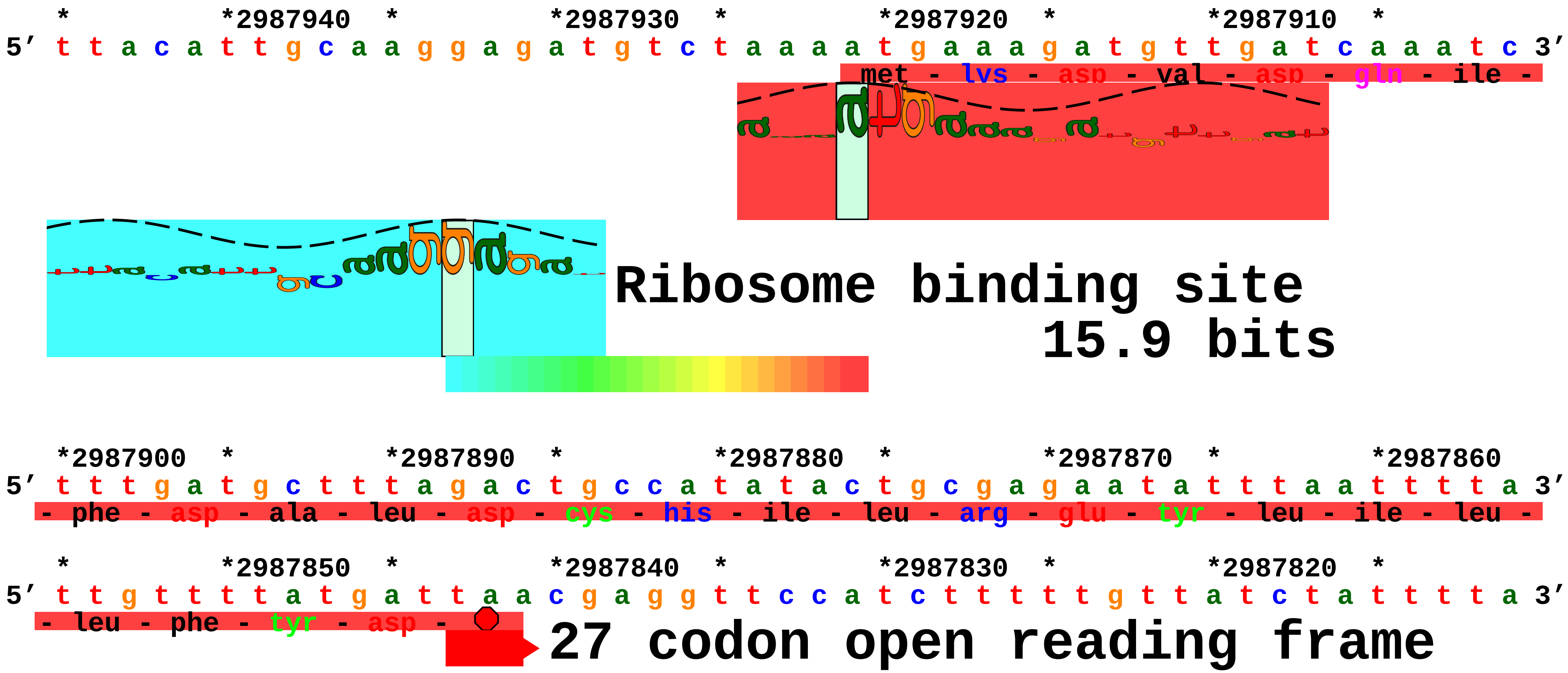
Ribosome binding site
15.9 bits

```
   *2987900   *         *2987890   *         *2987880   *         *2987870   *         *2987860
5′ t t t g a t g c t t t a g a c t g c c a t a t a c t g c g a g a a t a t t t a a t t t t a 3′
 - phe - asp - ala - leu - asp - cys - his - ile - leu - arg - glu - tyr - leu - ile - leu -

   *         *2987850   *         *2987840   *         *2987830   *         *2987820   *
5′ t t g t t t t a t g a t t a a c g a g g t t c c a t c t t t t t g t t a t c t a t t t t a 3′
 - leu - phe - tyr - asp -
```

27 codon open reading frame

Collaborators: Dr. Gisela Storz (NIH, NICHD, Bethesda, MD),
Dr. Kenneth Rudd (University of Miami School of Medicine, Miami, FL),
Dr. Matt Hemm (Towson University, MD)

Predicting Small Open Reading Frames in *E. coli*



* *2987940 * *2987930 * *2987920 * *2987910 *

5′ t t a c a t t g c a a g g a g a t g t c t a a a a t g a a a g a t g t t g a t c a a a t c 3′

met - lys - asp - val - asp - gln - ile -

Ribosome binding site
15.9 bits

*2987900 * *2987890 * *2987880 * *2987870 * *2987860

5′ t t t g a t g c t t t a g a c t g c c a t a t a c t g c g a g a a t a t t t a a t t t t a 3′

- phe - asp - ala - leu - asp - cys - his - ile - leu - arg - glu - tyr - leu - ile - leu -

* *2987850 * *2987840 * *2987830 * *2987820 *

5′ t t g t t t t a t g a t t a a c g a g g t t c c a t c t t t t t g t t a t c t a t t t t a 3′

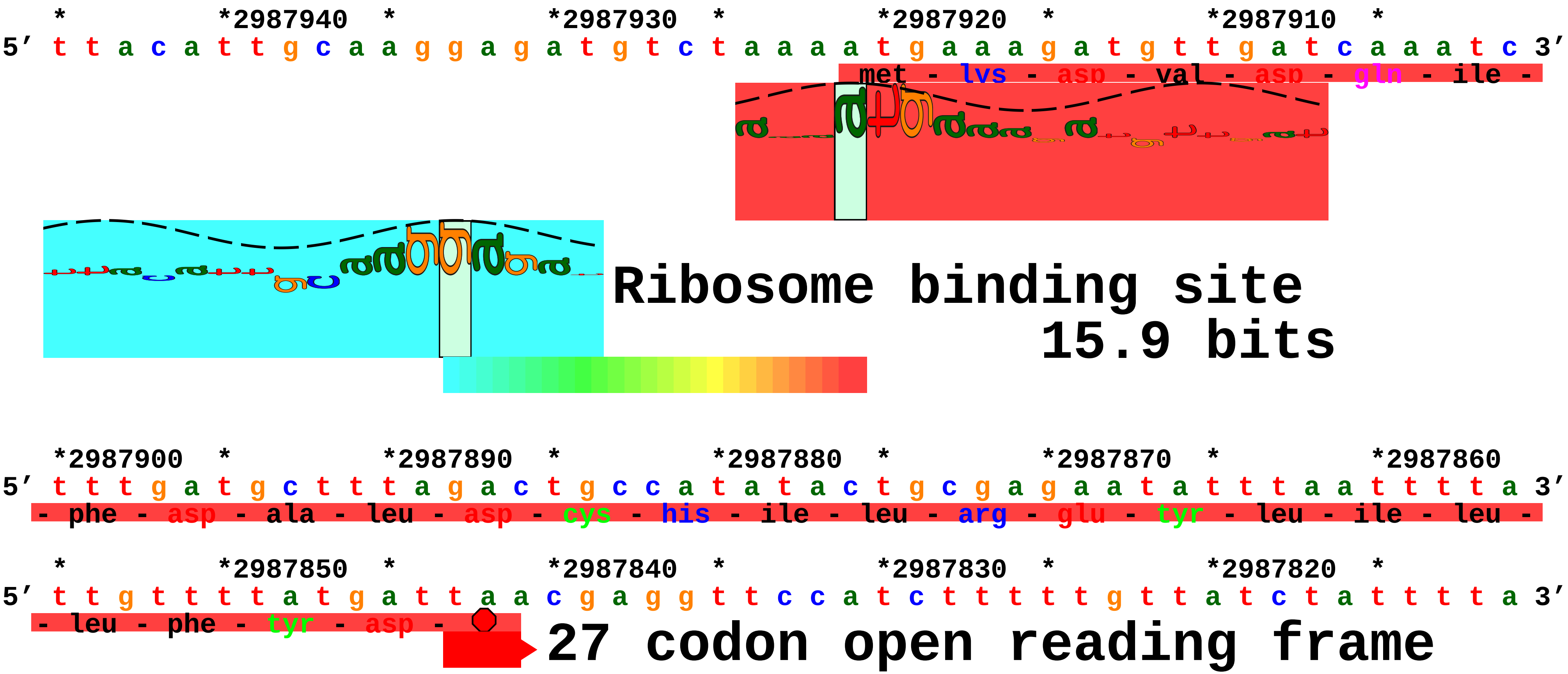
- leu - phe - tyr - asp -

27 codon open reading frame

- > 2000 **information theory predictions; test 24**

**Collaborators: Dr. Gisela Storz (NIH, NICHD, Bethesda, MD),
Dr. Kenneth Rudd (University of Miami School of Medicine, Miami, FL),
Dr. Matt Hemm (Towson University, MD)**

Predicting Small Open Reading Frames in *E. coli*



Ribosome binding site
15.9 bits

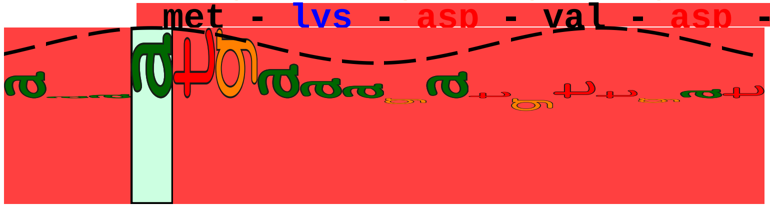
27 codon open reading frame

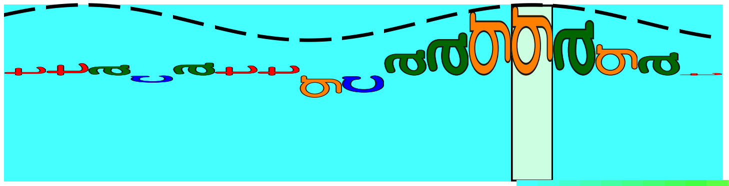
- > 2000 information theory predictions; test 24
- tested by sequential peptide affinity (SPA) tag

Collaborators: Dr. Gisela Storz (NIH, NICHD, Bethesda, MD),
Dr. Kenneth Rudd (University of Miami School of Medicine, Miami, FL),
Dr. Matt Hemm (Towson University, MD)

Predicting Small Open Reading Frames in *E. coli*

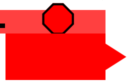
```
   *         *2987940  *         *2987930  *         *2987920  *         *2987910  *
5' t t a c a t t g c a a g g a g a t g t c t a a a a t g a a a g a t g t t g a t c a a a t c 3'
                                                       met - lys - asp - val - asp - gln - ile -
```





Ribosome binding site
15.9 bits

```
   *2987900  *         *2987890  *         *2987880  *         *2987870  *         *2987860
5' t t t g a t g c t t t a g a c t g c c a t a t a c t g c g a g a a t a t t t a a t t t t a 3'
 - phe - asp - ala - leu - asp - cys - his - ile - leu - arg - glu - tyr - leu - ile - leu -

   *         *2987850  *         *2987840  *         *2987830  *         *2987820  *
5' t t g t t t t a t g a t t a a c g a g g t t c c a t c t t t t t g t t a t c t a t t t t a 3'
 - leu - phe - tyr - asp -
```

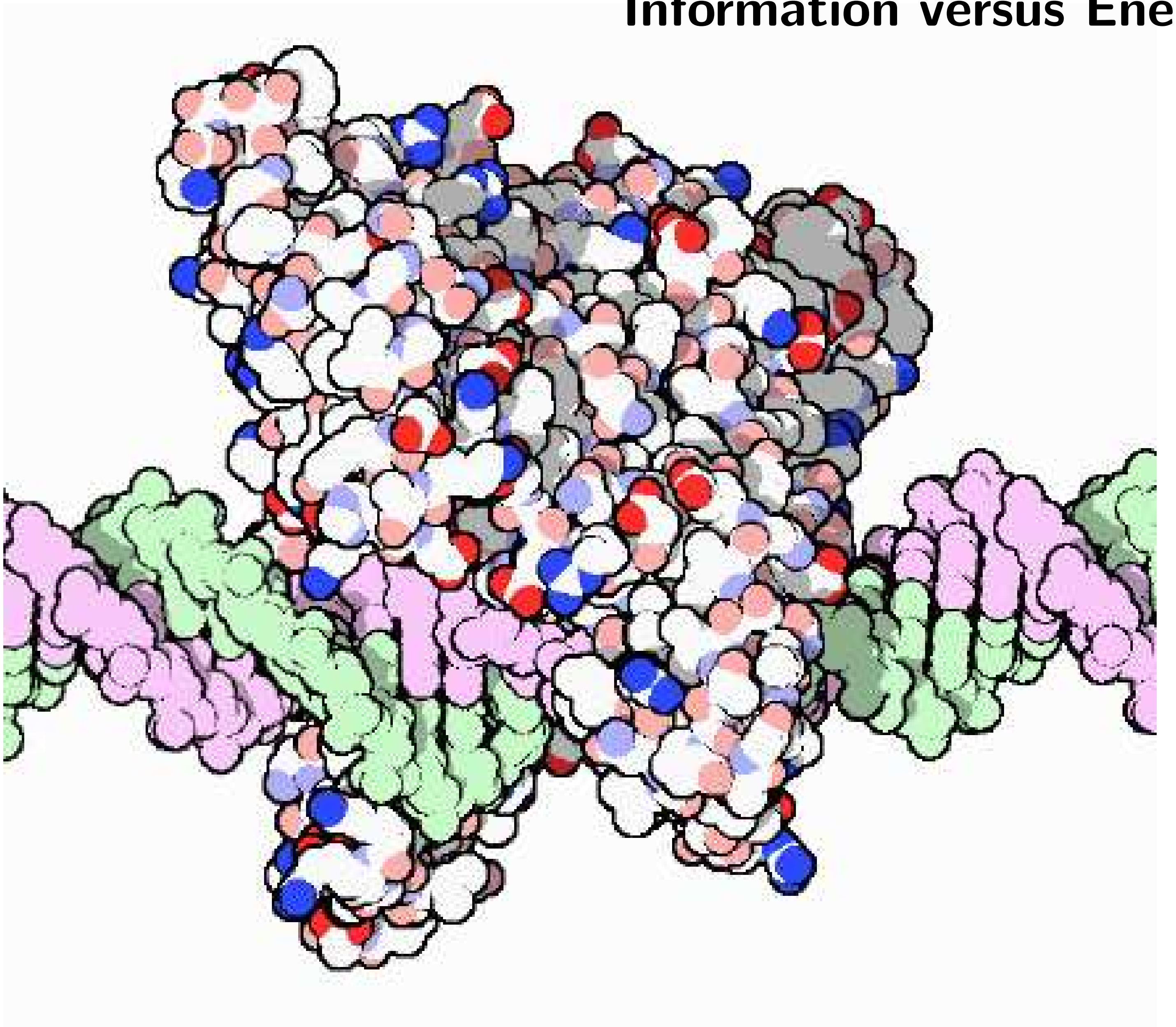


27 codon open reading frame

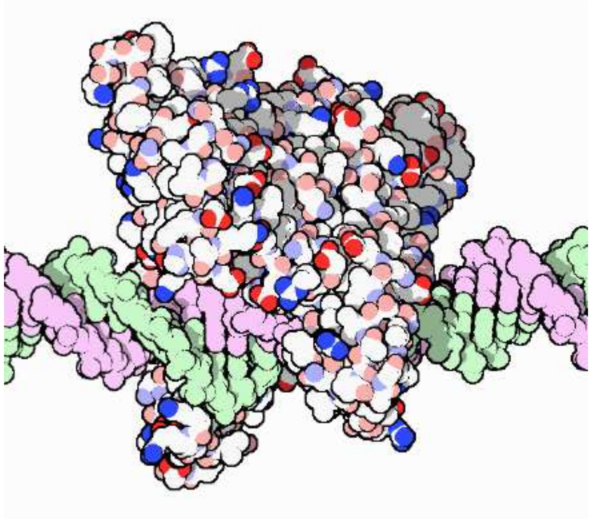
- > 2000 information theory predictions; test 24
- tested by sequential peptide affinity (SPA) tag
- 18 new genes < 50 aa long

Collaborators: Dr. Gisela Storz (NIH, NICHD, Bethesda, MD),
Dr. Kenneth Rudd (University of Miami School of Medicine, Miami, FL),
Dr. Matt Hemm (Towson University, MD)

Information versus Energy



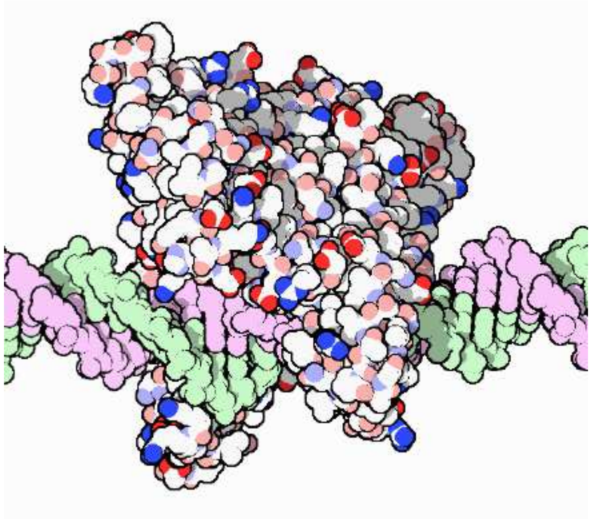
Information of EcoRI DNA Binding

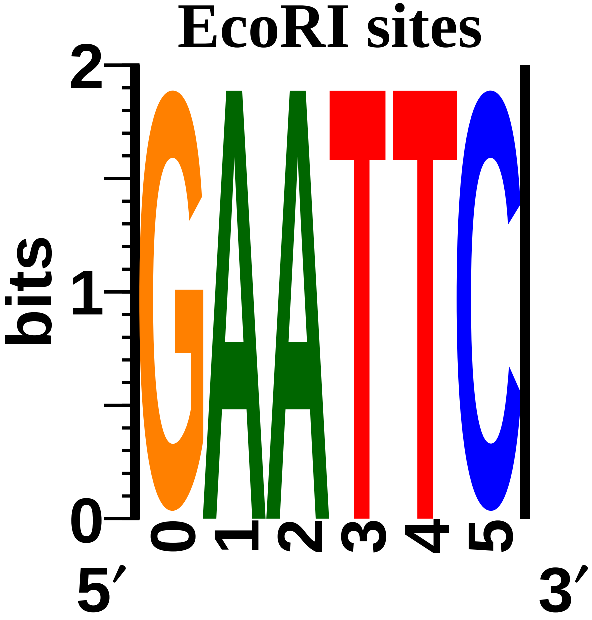


- EcoRI - restriction enzyme

Information of EcoRI DNA Binding

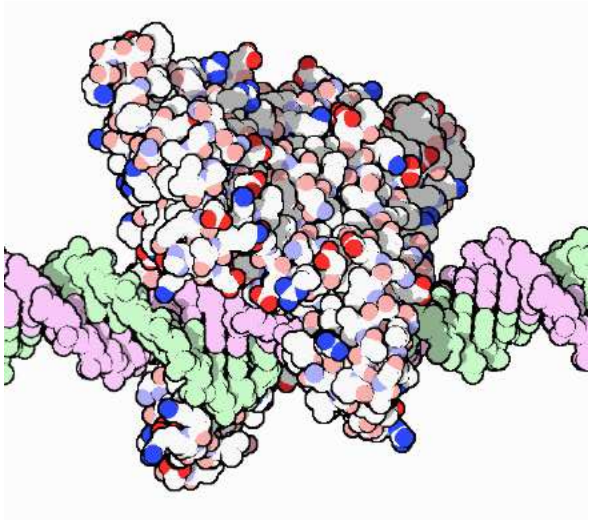
- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$

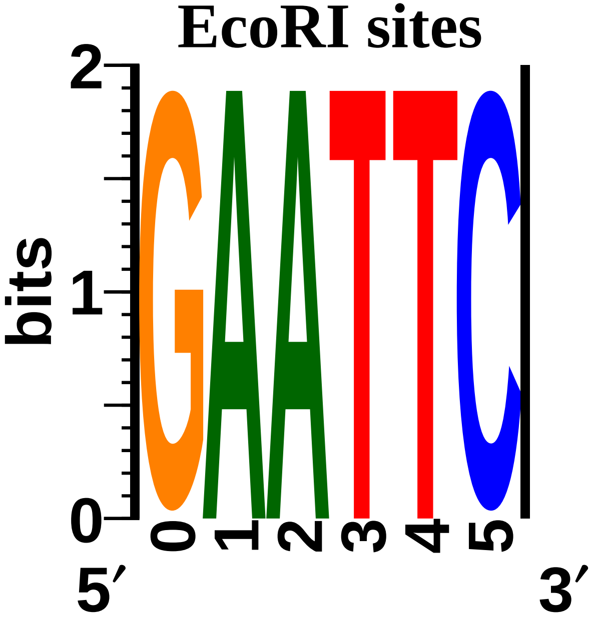




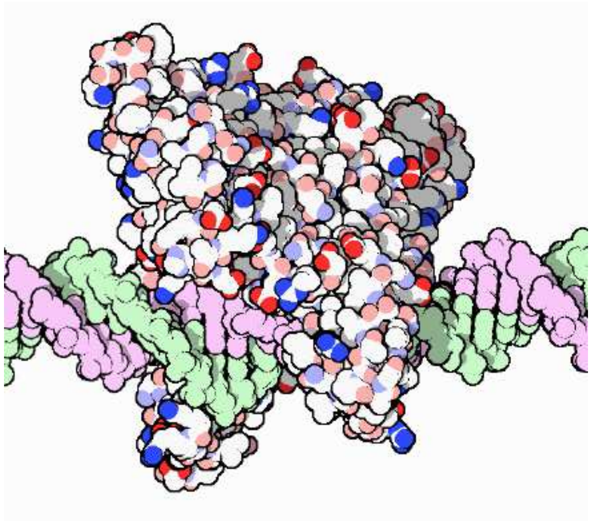
Information of EcoRI DNA Binding

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

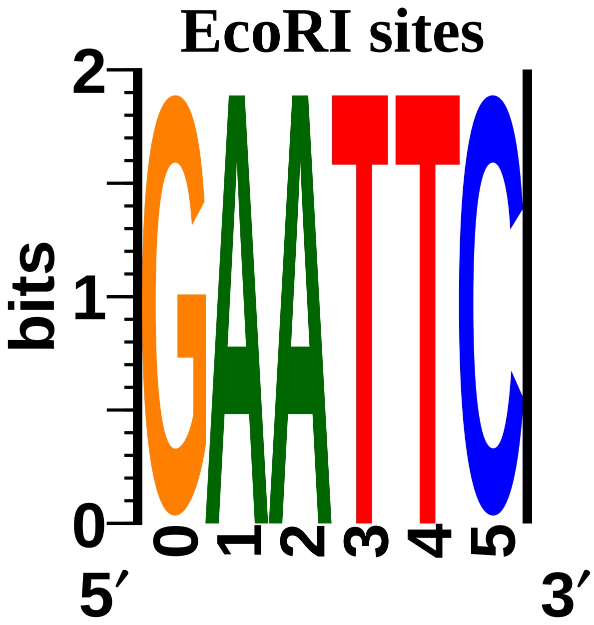




Information of EcoRI DNA Binding



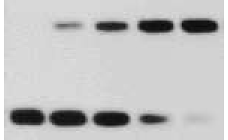
- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$
- $4^6 = 4096$ possible DNA hexamers
- information required:
  $\log_2 4096 = 12$ bits
  or
  6 bases $\times$ 2 bits per base $=$ $\boxed{\text{12 bits}}$



Energy Dissipation by EcoRI

- Measured specific binding constant:

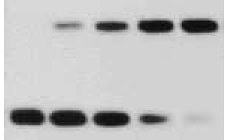
$$K_{spec} = 1.6 \times 10^5$$



Energy Dissipation by EcoRI

- Measured specific binding constant:

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



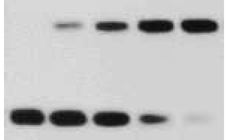
- Average energy dissipated by one molecule as it binds:

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

Energy Dissipation by EcoRI

- Measured specific binding constant:

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



- Average energy dissipated by one molecule as it binds:

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

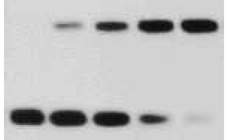
- The Second Law of Thermodynamics as a conversion factor:

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

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:

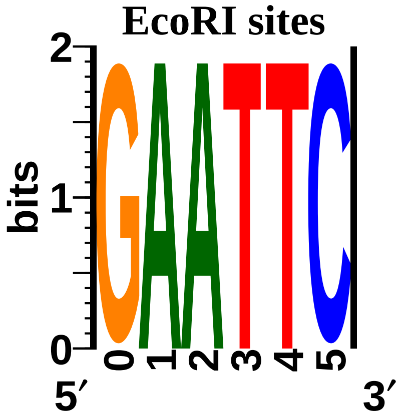
$$\begin{aligned} R_{energy} &= -\Delta G^\circ / \mathcal{E}_{min} \\ &= k_{\mathsf{B}} T \ln K_{spec} / k_{\mathsf{B}} T \ln 2 \\ &= \log_2 K_{spec} \qquad \Leftarrow \text{SO SIMPLE!} \\ &= \boxed{\text{17.3 bits per binding}} \end{aligned}$$

Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices

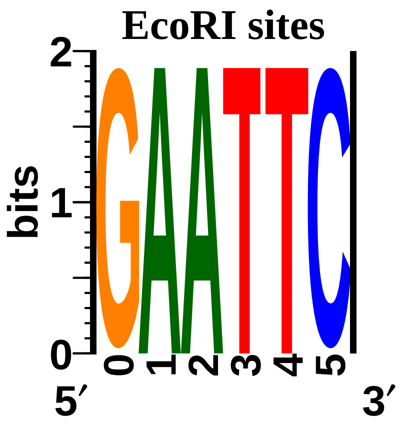
Information/Energy = Efficiency of EcoRI

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



Information/Energy = Efficiency of EcoRI

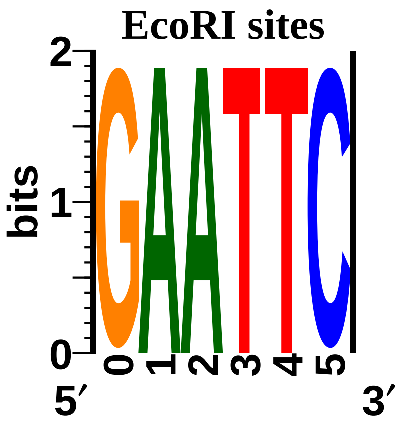
EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.



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.

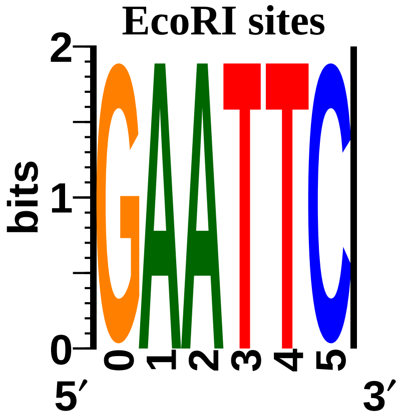


Efficiency is
‘WORK’ DONE / ENERGY DISSIPATED

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

Information/Energy = Efficiency of EcoRI = 70%

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

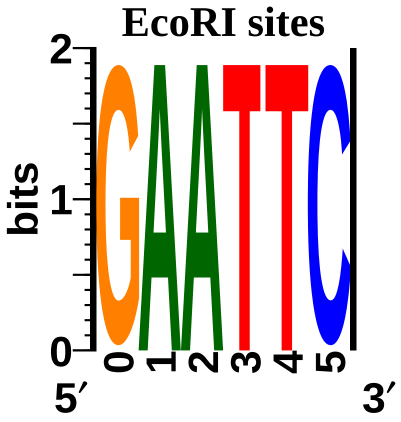


Efficiency is
'WORK' DONE / ENERGY DISSIPATED

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

The efficiency is 70%.

Information/Energy = Efficiency of EcoRI = 70%



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

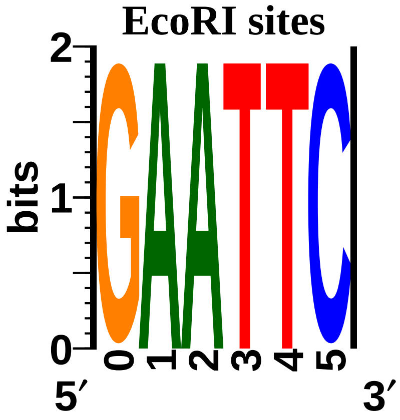
Efficiency is
'WORK' DONE / ENERGY DISSIPATED

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

The efficiency is 70%.

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

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

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

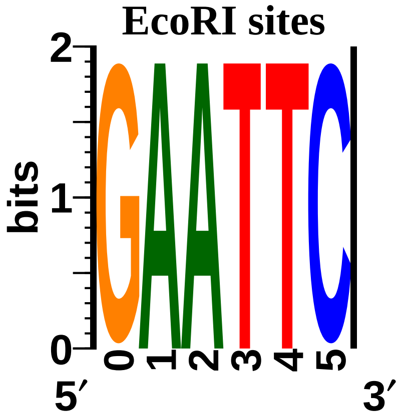
The efficiency is 70%.

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

**70% efficiency also appears widely in biology:
rhodopsin, muscle and other systems.**

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

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

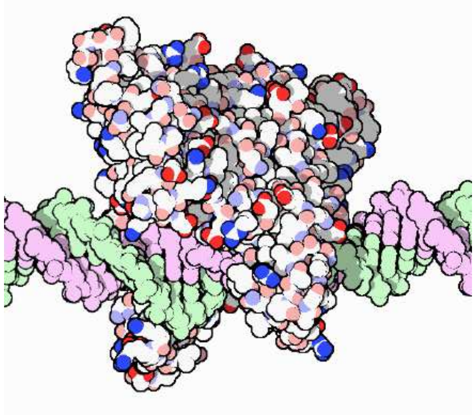
The efficiency is 70%.

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

**70% efficiency also appears widely in biology:
rhodopsin, muscle and other systems.**

Why 70% efficiency?

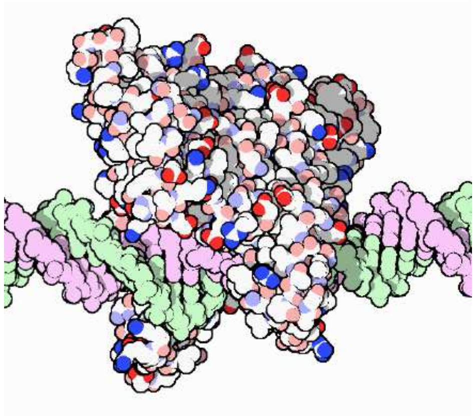
Theoretical Efficiency



- 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

Theoretical Efficiency

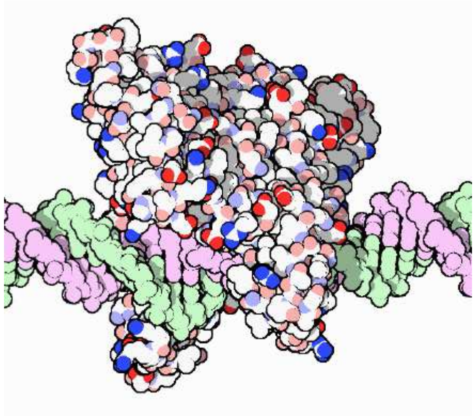


- For molecular states of molecules with d_{space} 'parts' P_y energy is dissipated for noise N_y and

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

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

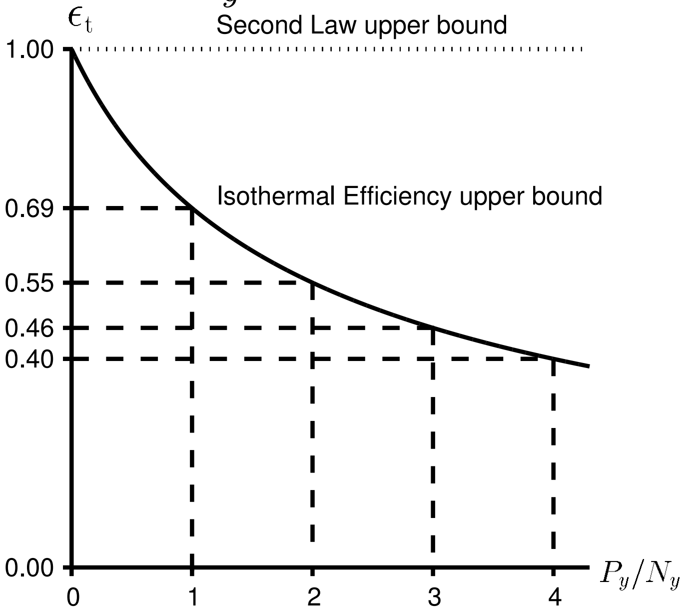
Theoretical Efficiency



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

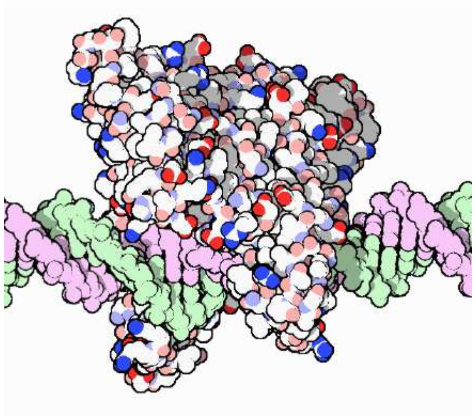
  $C = d_{space} \log_2(P_y/N_y + 1) \leftarrow$ machine capacity

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



The curve is an upper bound

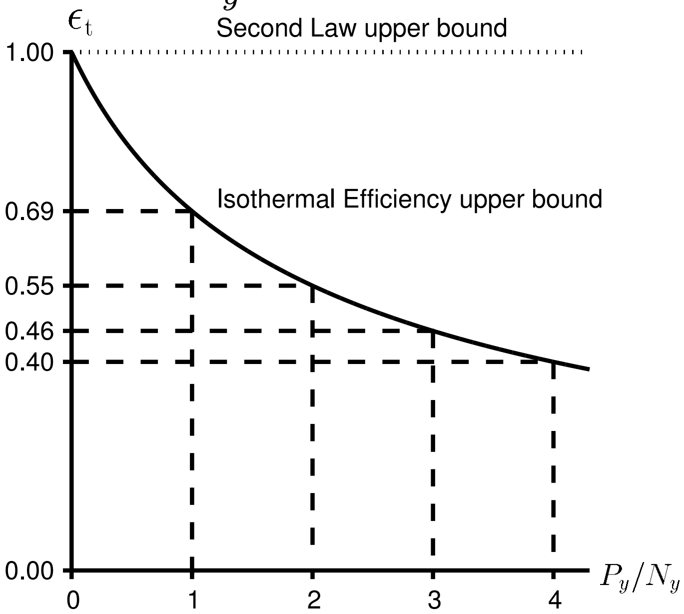
Theoretical Efficiency



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

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

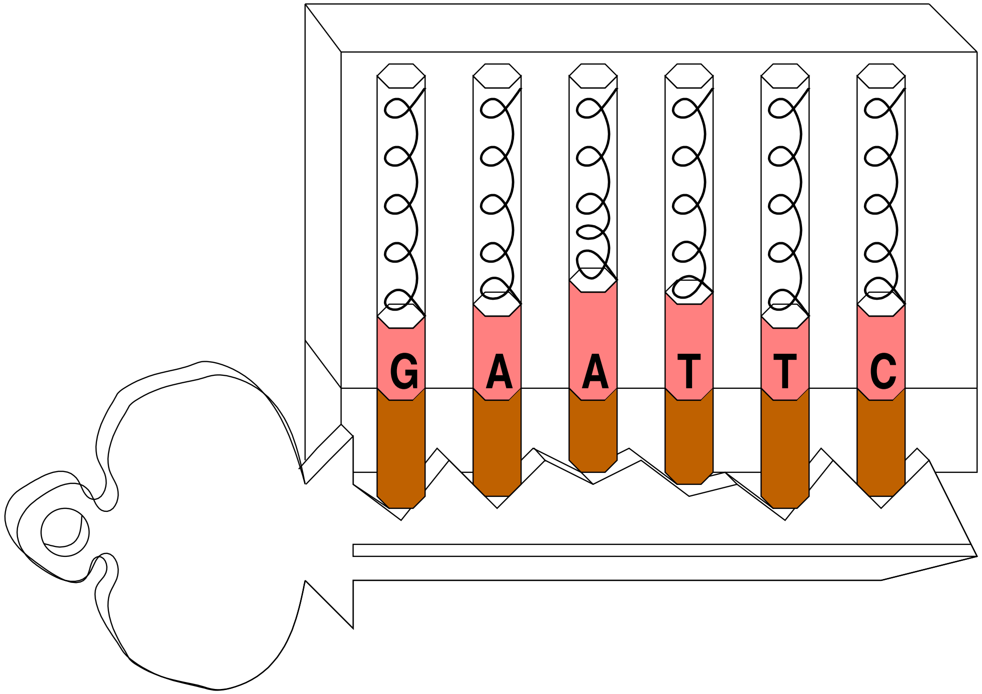
$$\epsilon_t \leq \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow \text{molecular efficiency}$$



The curve is an upper bound

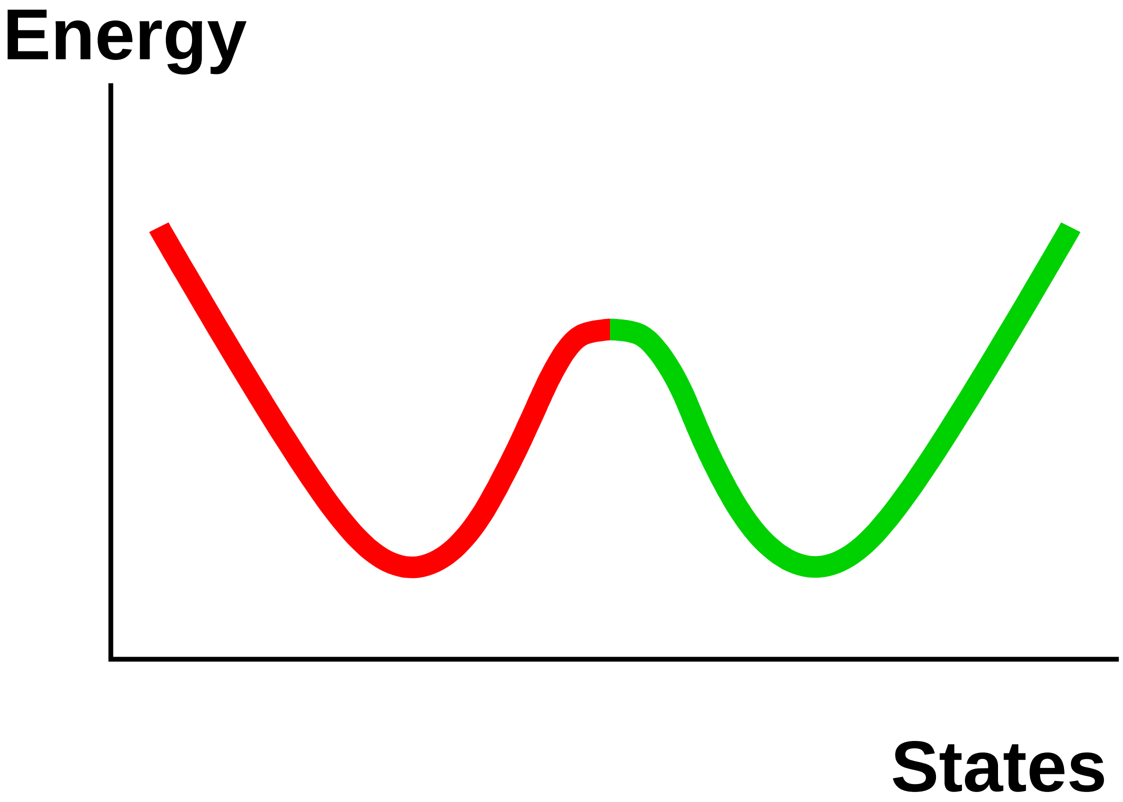
- $\boxed{\text{If } P_y/N_y = 1 \text{ the efficiency is 70\%!}}$

Lock and Key



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

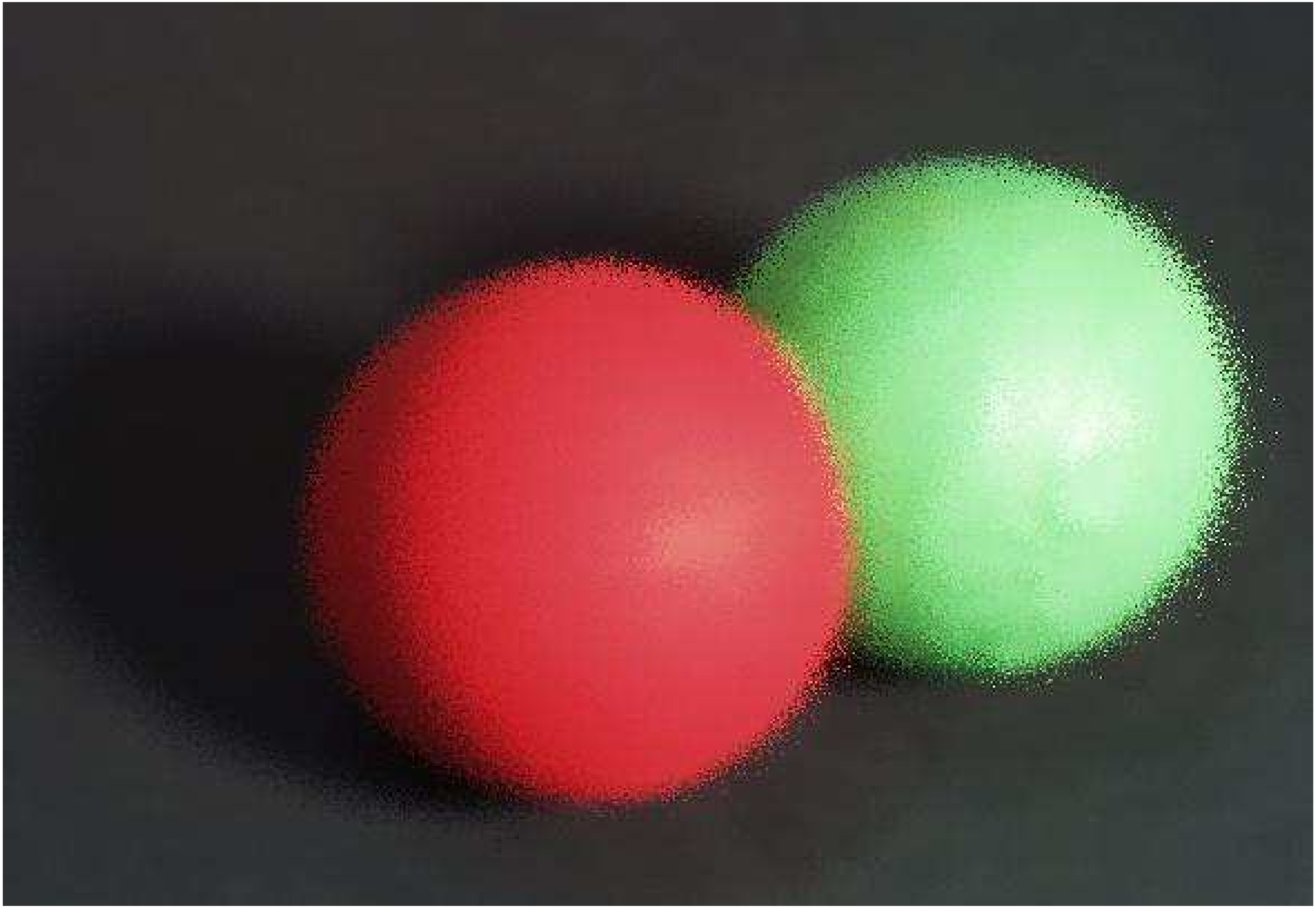
1 Dimension



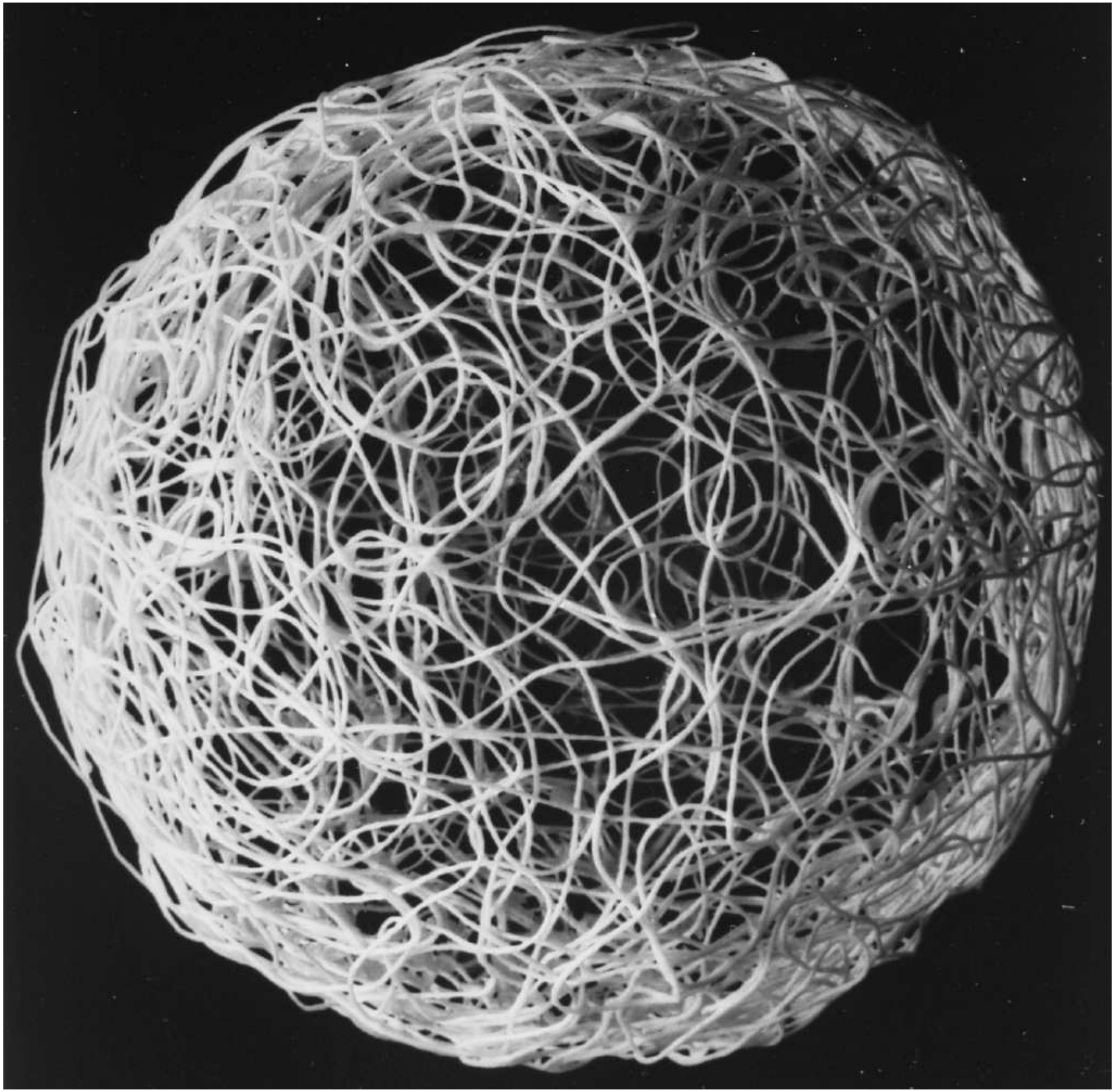
1 dimension is too simple!

Bowles in 2 Dimensions

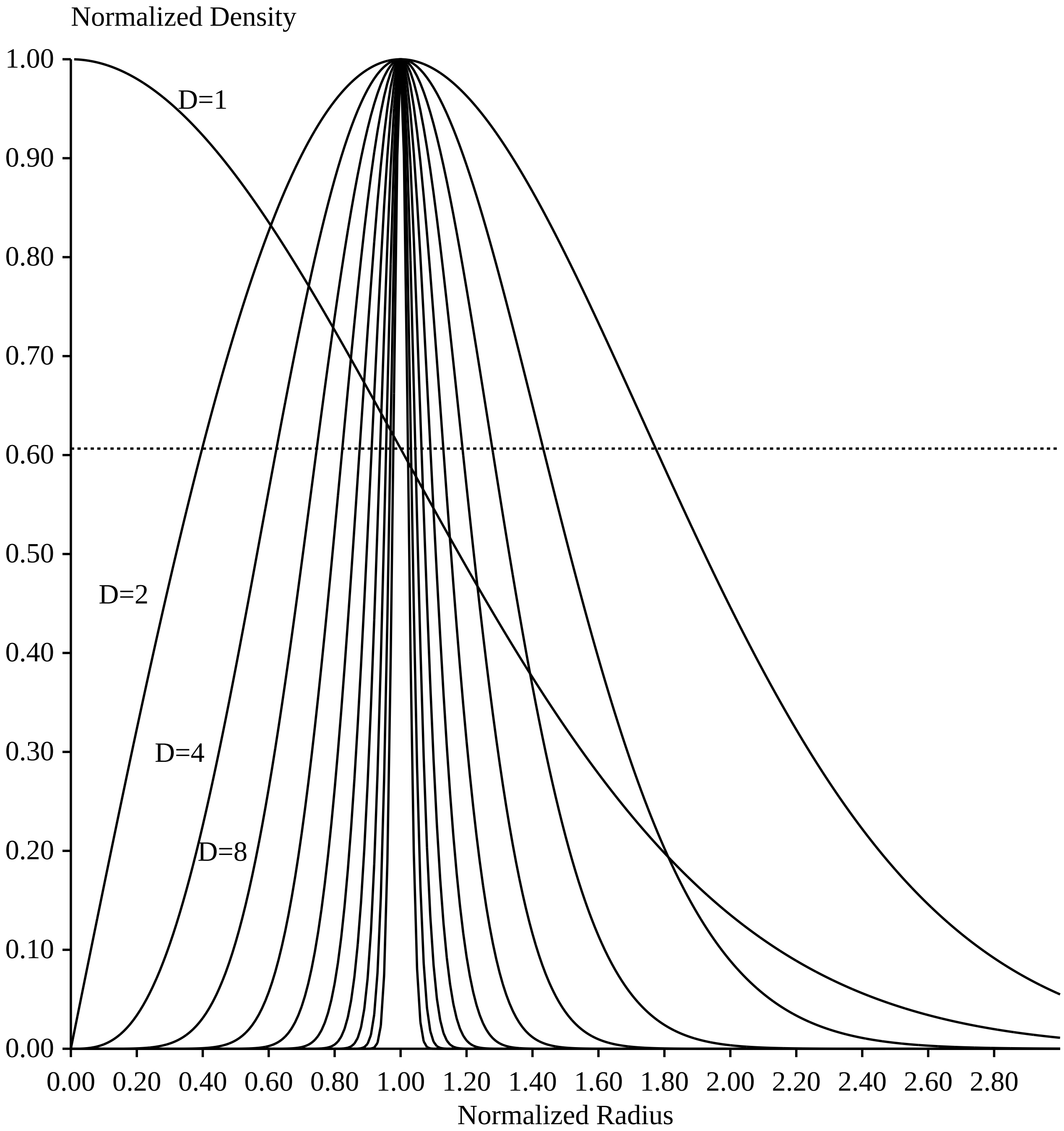
Spheres in 3 Dimensions



N Dimensional Sphere



Spheres tighten in high dimensions



Energy and velocity

$$\text{Energy} = \frac{1}{2}\text{Mass} \times \text{velocity}^2$$

Energy and velocity

$$\text{Energy} = \frac{1}{2}\text{Mass} \times \text{velocity}^2$$

$$\text{Energy in the molecule} = \text{Noise} = \text{N}$$

Energy and velocity

$$\text{Energy} = \frac{1}{2}\text{Mass} \times \text{velocity}^2$$

$$\text{Energy in the molecule} = \text{Noise} = \text{N}$$

$$\text{maximum velocity} \propto \sqrt{\text{N}}$$

Energy and velocity

$$\text{Energy} = \frac{1}{2}\text{Mass} \times \text{velocity}^2$$

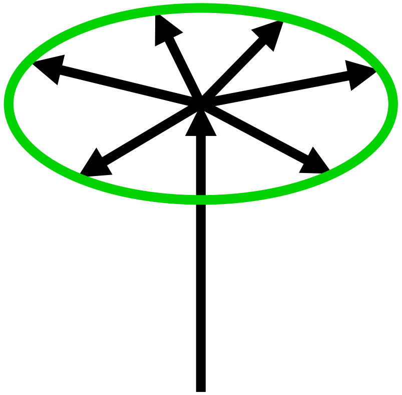
$$\text{Energy in the molecule} = \text{Noise} = \text{N}$$

$$\text{maximum velocity} \propto \sqrt{\text{N}}$$

$$\text{sphere radius} \propto \sqrt{\text{N}}$$

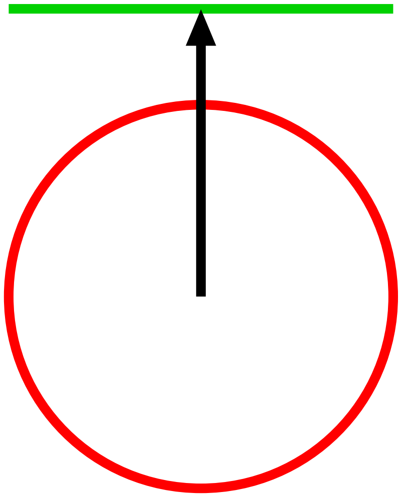
Sphere Packing

hyperdirection



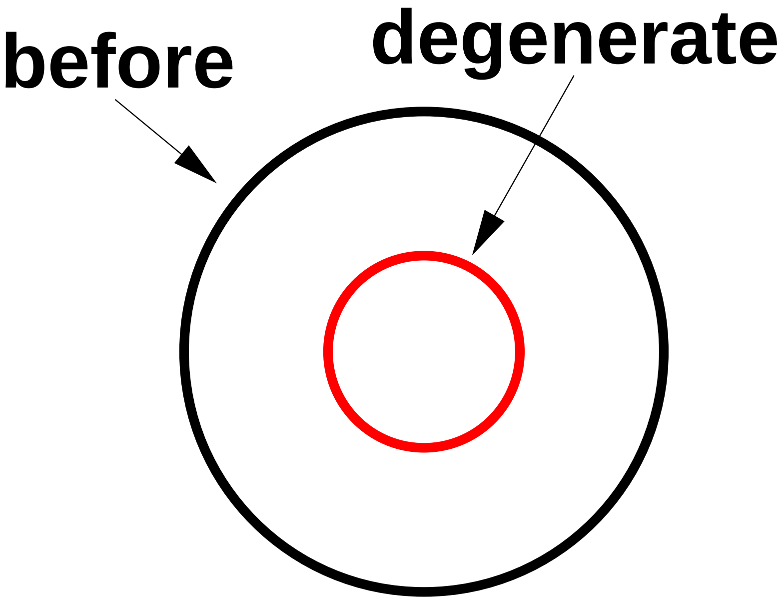
In 100 dimensions
99% of the thermal noise
is at right angles
to a given direction!

two



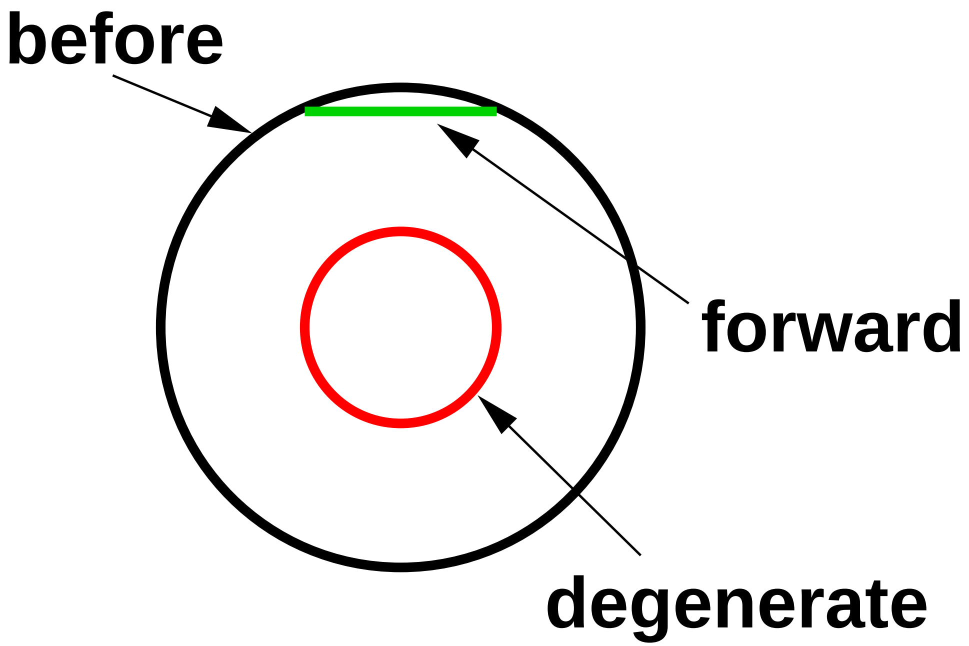
Two spheres in
high dimensional space

degenerate



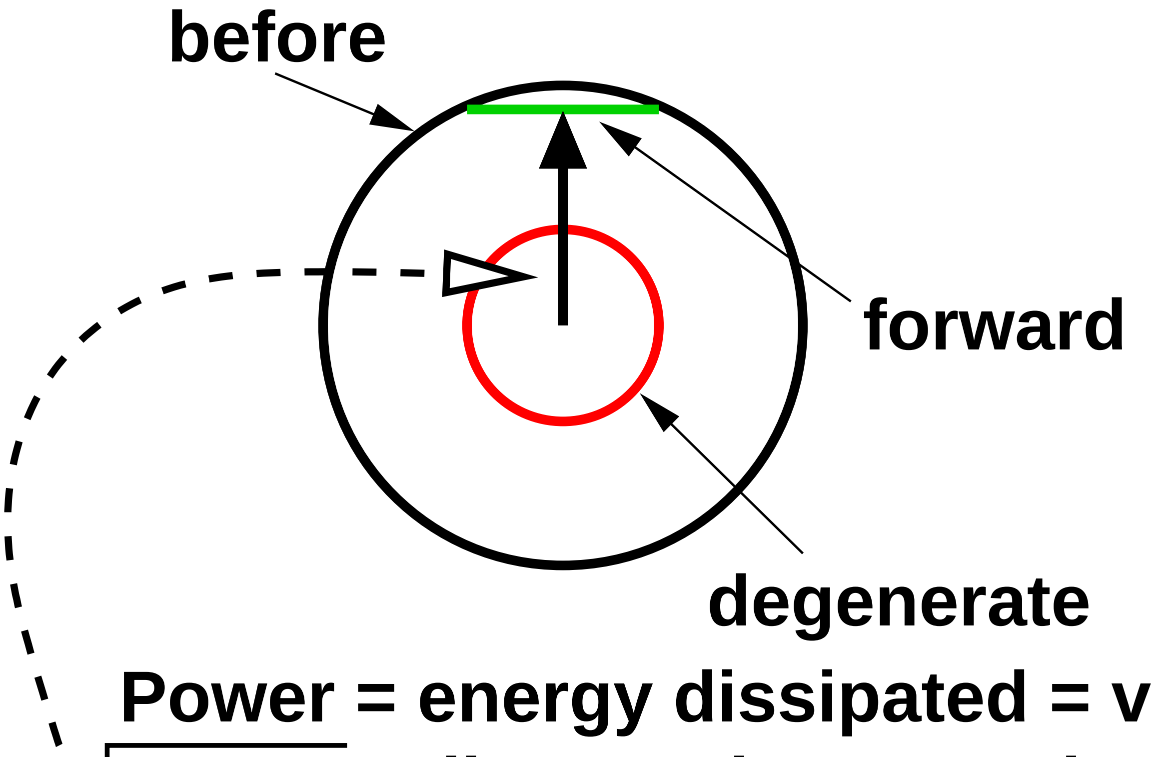
Hypothesis:
there is a sphere
in the middle
of the before sphere

forward



To do useful selections
the molecular machine
must avoid the degenerate sphere
It must choose the forward sphere

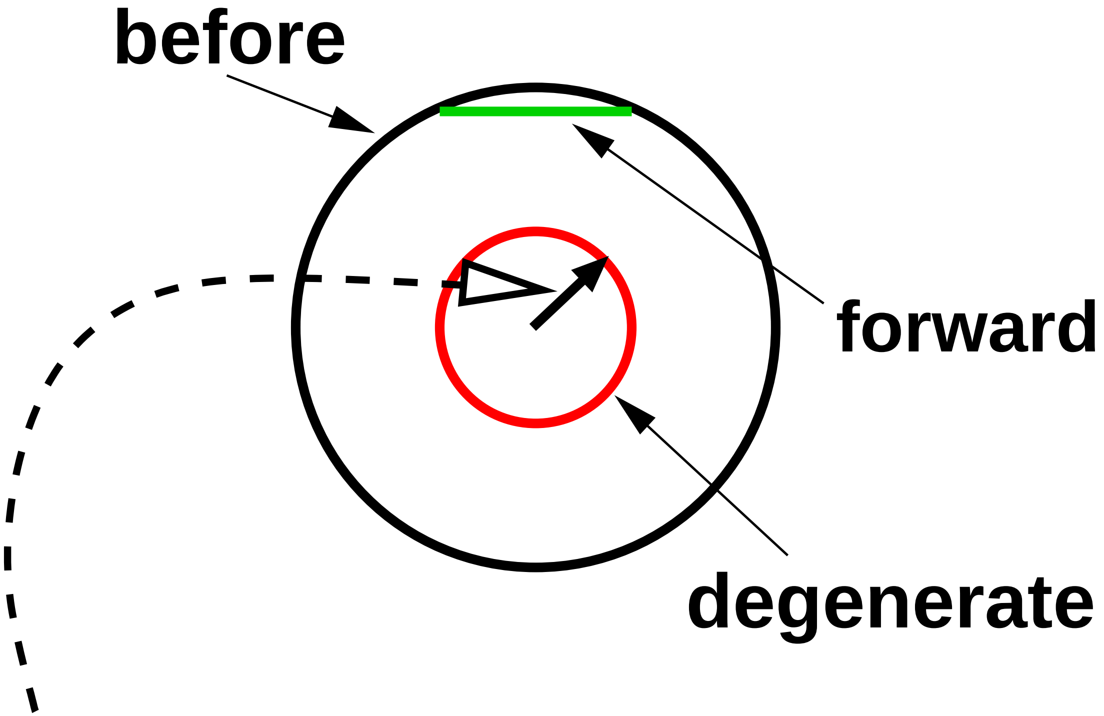
power



Power = energy dissipated = velocity^2

$\sqrt{\text{Power}}$ = distance between the forward and the degenerate sphere centers

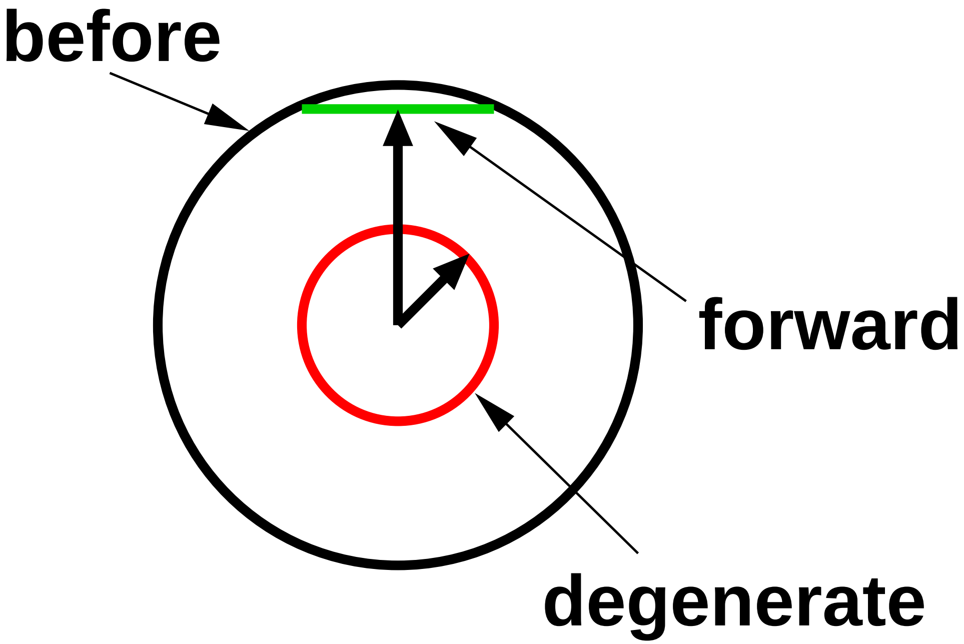
noise



$\sqrt{\text{Noise}}$ = degenerate sphere radius

**Thermal noise determines
the radius of the degenerate sphere**

criterion

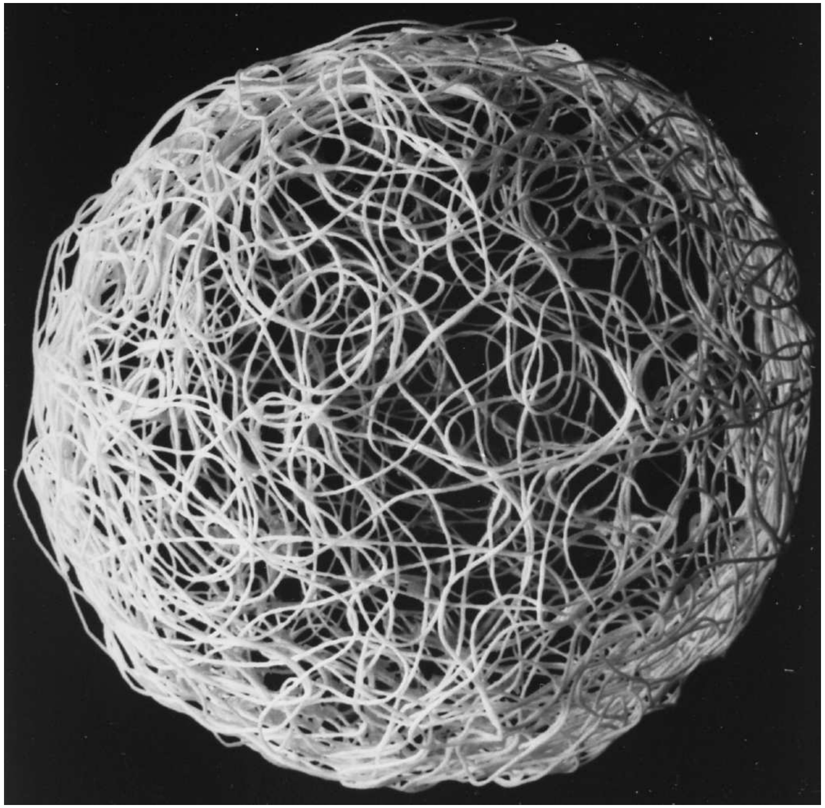


**Criterion for distinct states:
forward does not touch degenerate**

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

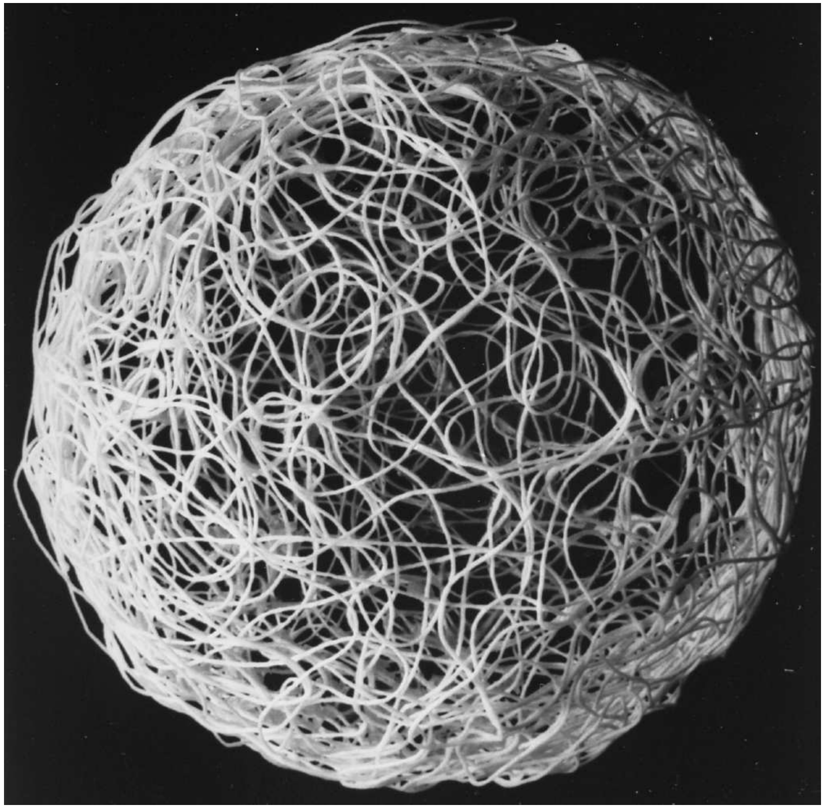
N Dimensional Sphere Separation

Degenerate Sphere

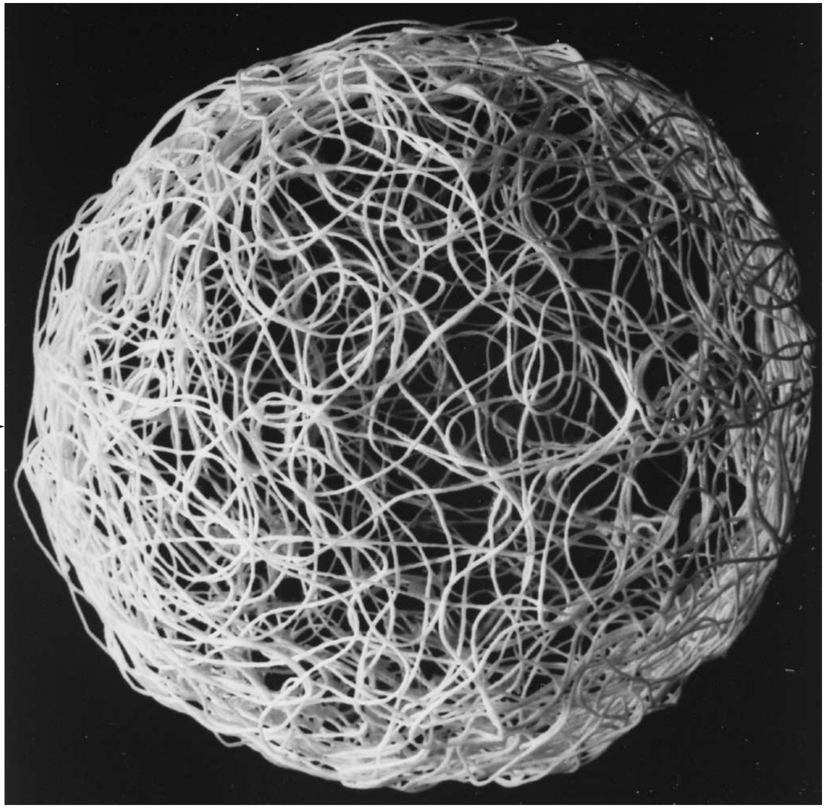


N Dimensional Sphere Separation

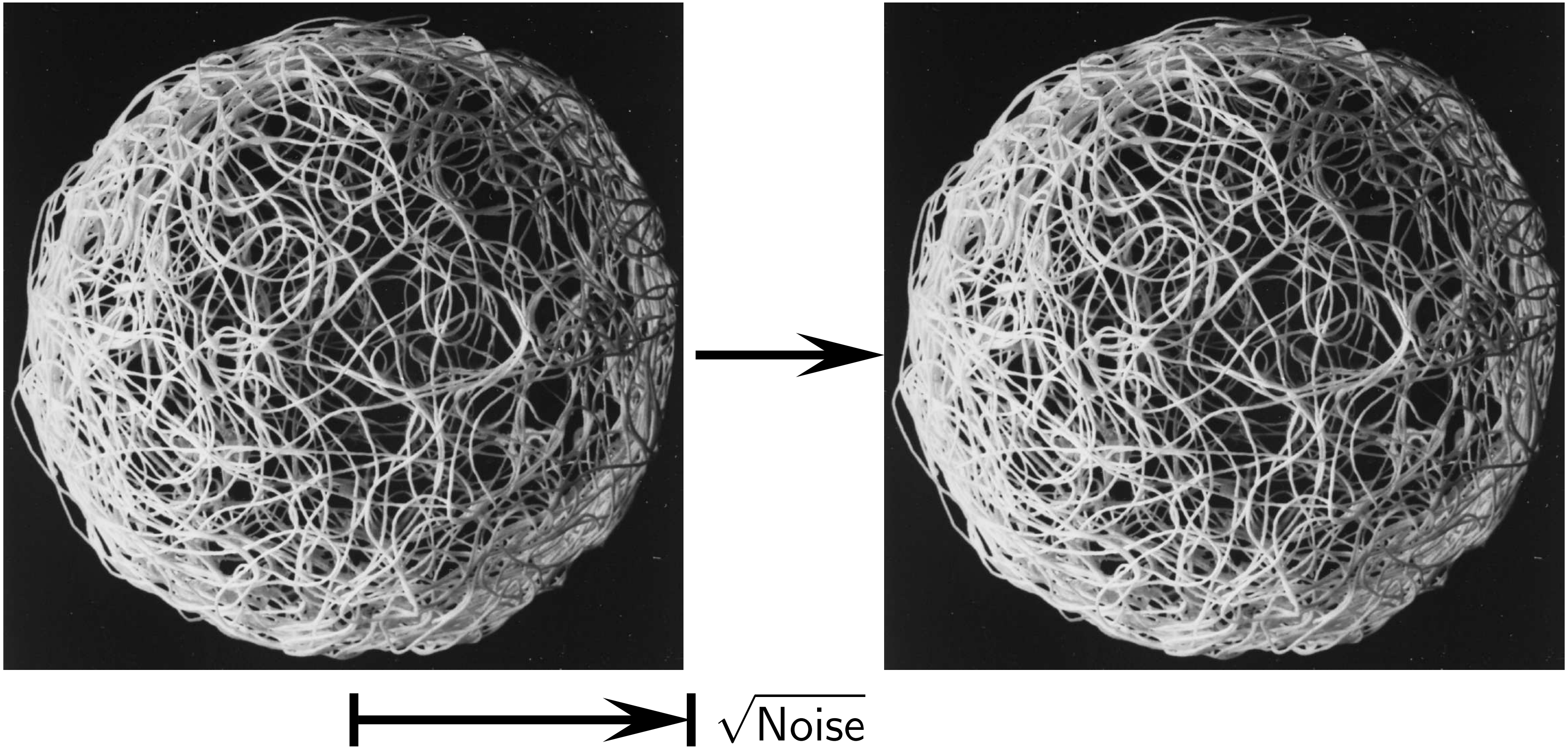
Degenerate Sphere



Forward Sphere



N Dimensional Sphere Separation

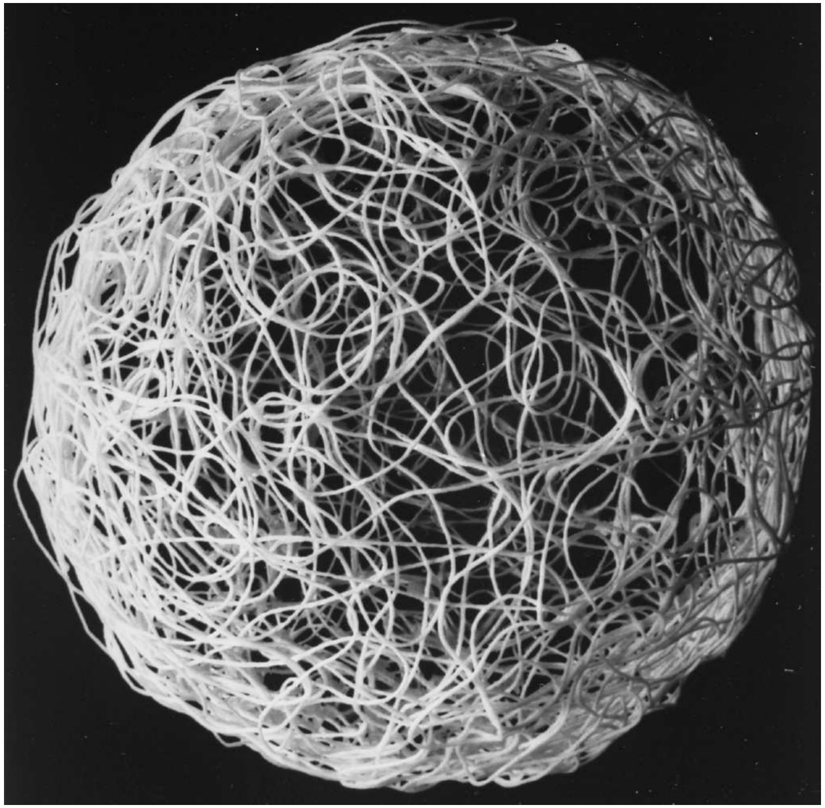


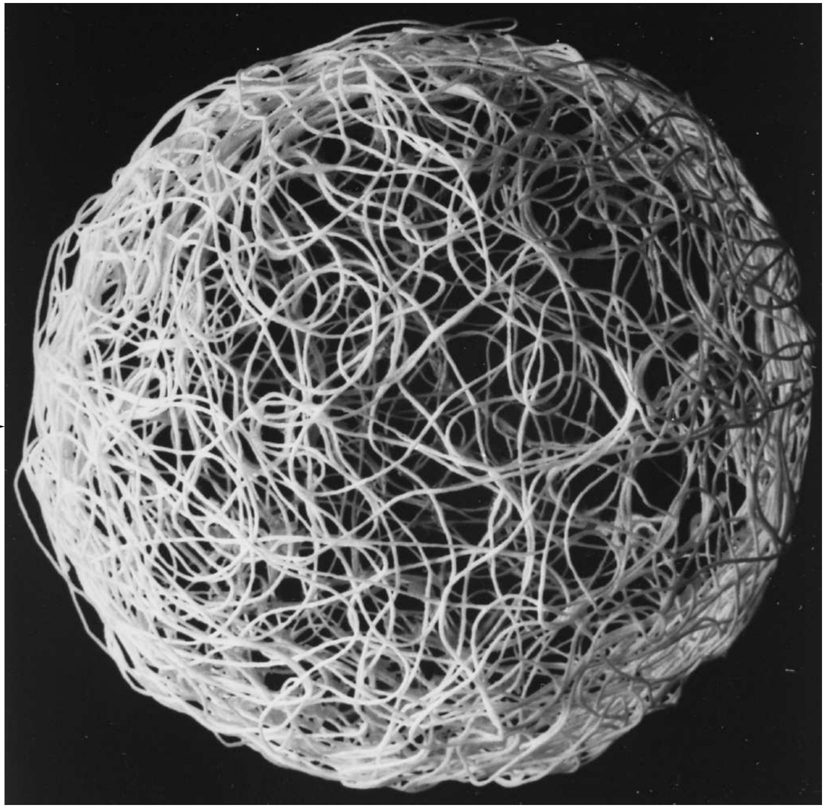
Degenerate Sphere

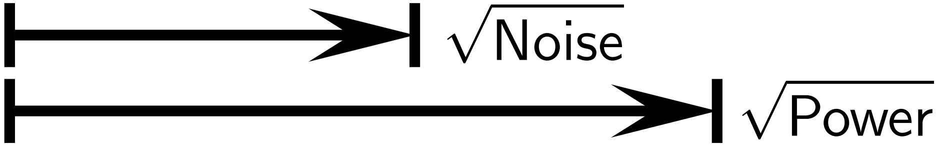
Forward Sphere

$\sqrt{\text{Noise}}$

N Dimensional Sphere Separation

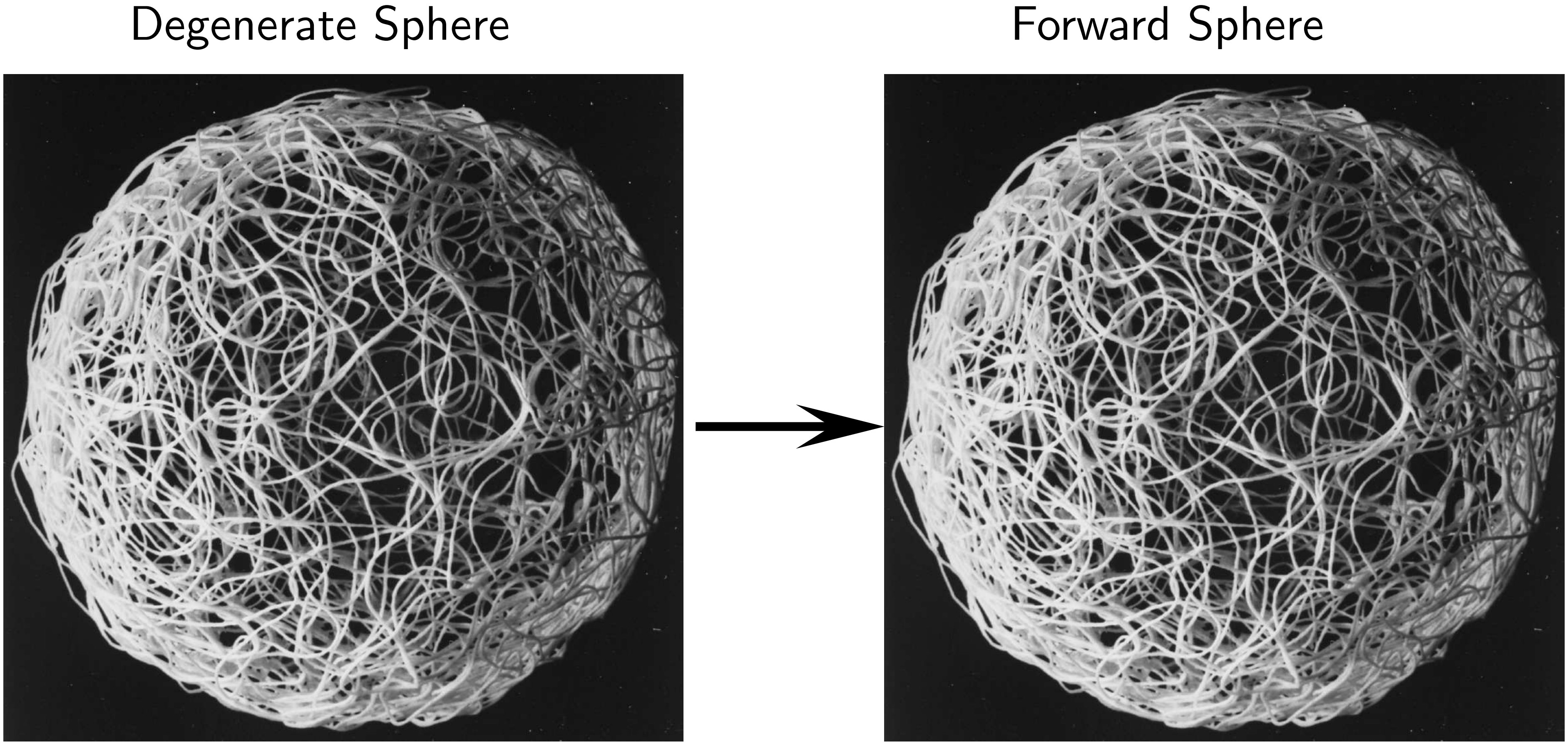

Degenerate Sphere


Forward Sphere

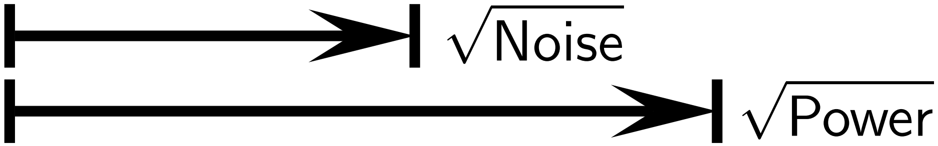

$\sqrt{\text{Noise}}$

$\sqrt{\text{Power}}$

N Dimensional Sphere Separation


Degenerate Sphere

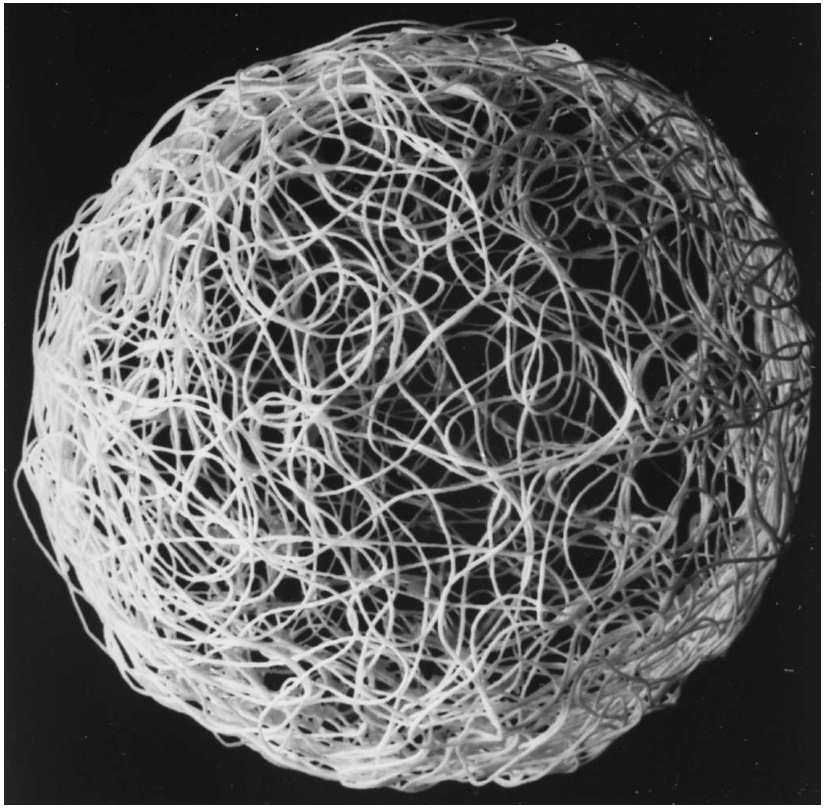
Forward Sphere



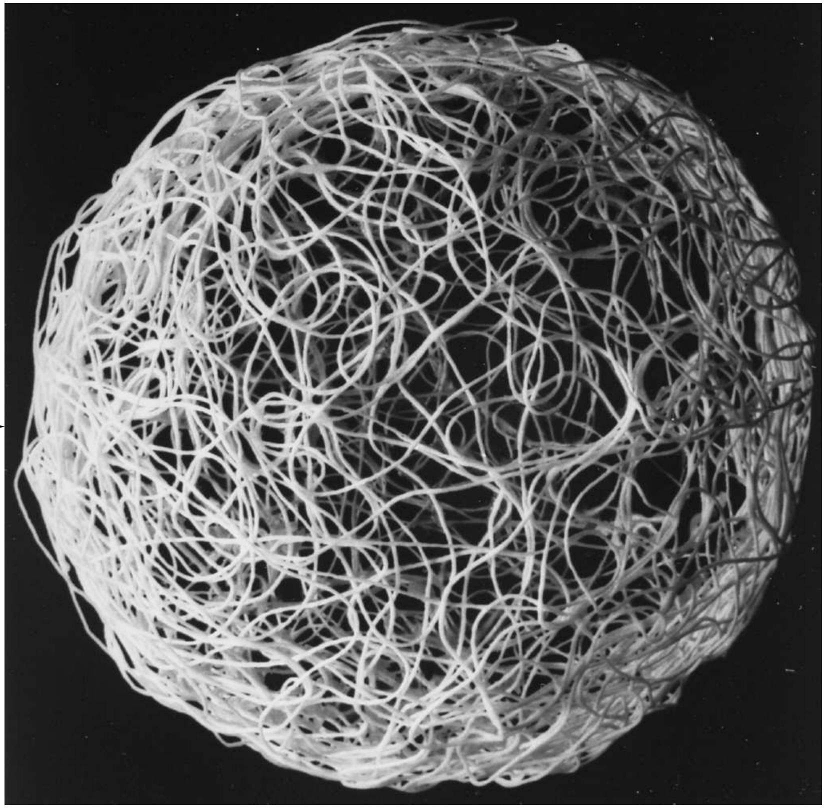
Energy dissipated to escape the Degenerate Sphere must exceed the Noise

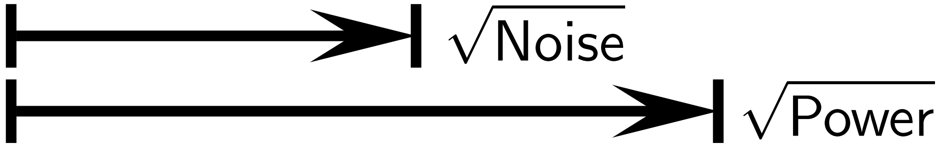
N Dimensional Sphere Separation

Degenerate Sphere



Forward Sphere





$\sqrt{\text{Noise}}$

$\sqrt{\text{Power}}$

Energy dissipated to escape the Degenerate Sphere must exceed the Noise

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

CONSEQUENCES OF THE DEGENERATE SPHERE HYPOTHESIS

The geometry gives:

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

CONSEQUENCES OF THE DEGENERATE SPHERE HYPOTHESIS

The geometry gives:

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

so

$$\frac{\text{Power}}{\text{Noise}} > 1$$

CONSEQUENCES OF THE DEGENERATE SPHERE HYPOTHESIS

The geometry gives:

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

so

$$\frac{\text{Power}}{\text{Noise}} > 1$$

which when plugged into the efficiency formula:

$$\epsilon_t \equiv \frac{\mathcal{E}_{min}}{\mathcal{E}} = \frac{\ln\left(\frac{\text{Power}}{\text{Noise}} + 1\right)}{\frac{\text{Power}}{\text{Noise}}} \qquad \frac{\text{(joules per bit)}}{\text{(joules per bit)}}$$

CONSEQUENCES OF THE DEGENERATE SPHERE HYPOTHESIS

The geometry gives:

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

so

$$\frac{\text{Power}}{\text{Noise}} > 1$$

which when plugged into the efficiency formula:

$$\epsilon_t \equiv \frac{\mathcal{E}_{min}}{\mathcal{E}} = \frac{\ln\left(\frac{\text{Power}}{\text{Noise}} + 1\right)}{\frac{\text{Power}}{\text{Noise}}} \qquad \frac{\text{(joules per bit)}}{\text{(joules per bit)}}$$

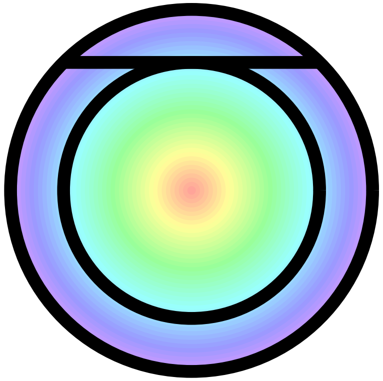
gives:

$$\epsilon_t = \ln 2 \approx 0.6931$$

CONSEQUENCES OF THE DEGENERATE SPHERE HYPOTHESIS

The geometry gives:

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



so

$$\frac{\text{Power}}{\text{Noise}} > 1$$

which when plugged into the efficiency formula:

$$\epsilon_t \equiv \frac{\mathcal{E}_{min}}{\mathcal{E}} = \frac{\ln\left(\frac{\text{Power}}{\text{Noise}} + 1\right)}{\frac{\text{Power}}{\text{Noise}}} \qquad \frac{\text{(joules per bit)}}{\text{(joules per bit)}}$$

gives:

$$\epsilon_t = \ln 2 \approx 0.6931$$

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

Why is the Genetic Code Degenerate?

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

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

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

$$\epsilon_r = \frac{\log_2 \text{actual choices}}{\log_2 \text{maximum choices}} = \frac{4.3}{6} = 0.72$$

The Genetic Code fits the theory!

Amino Acid Frequencies

A	105312381
C	17427433
D	67454442
E	77603281
F	48627269
G	83989735
H	27315242
I	69538797
K	65592680
L	119947552
M	27534150
N	53024966
O	10
P	61536653
Q	49569998
R	71591890
S	91898484
T	69490771
U	397
V	80381739
W	15430467
Y	37433671

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	105312381
C	17427433
D	67454442
E	77603281
F	48627269
G	83989735
H	27315242
I	69538797
K	65592680
L	119947552
M	27534150
N	53024966
O	10
P	61536653
Q	49569998
R	71591890
S	91898484
T	69490771
U	397
V	80381739
W	15430467
Y	37433671

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:

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

That's what is actually accomplished by translation.

Translational Efficiency

Compute the efficiency:

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

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

Translational Efficiency

Compute the efficiency:

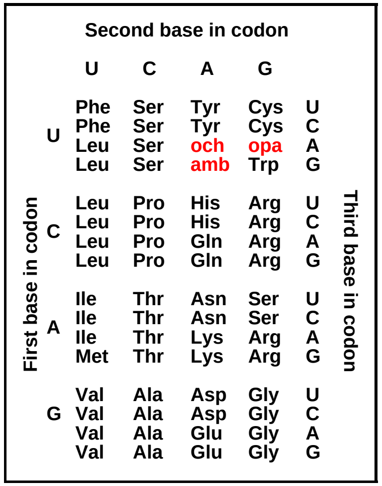
$$\begin{aligned}\epsilon_r &= \frac{4.170}{6} \\ &= 0.6949 \text{ Measured efficiency}\end{aligned}$$

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

Translational Efficiency

Compute the efficiency:

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



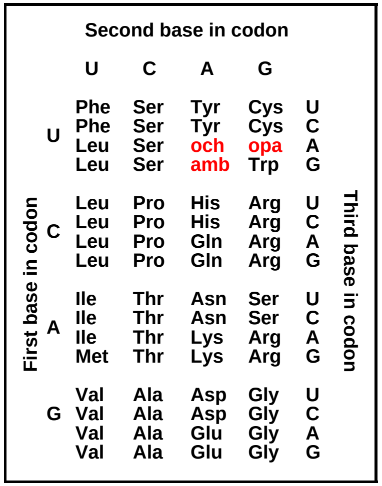
	Second base in codon				
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg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

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

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 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

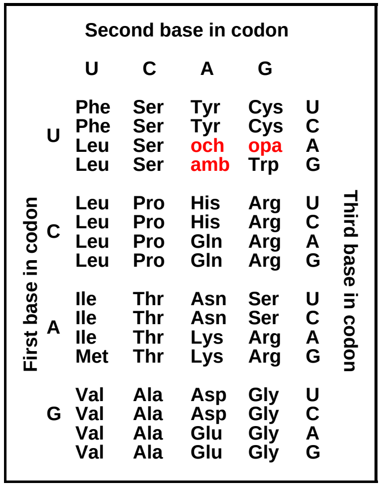
Since this comes from > 1 billion amino acids, 0.2% excess is significant!

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:

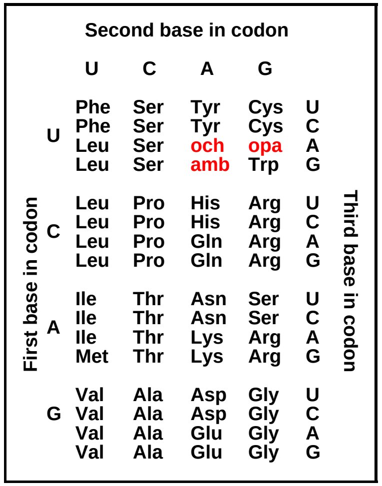
$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

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.



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:

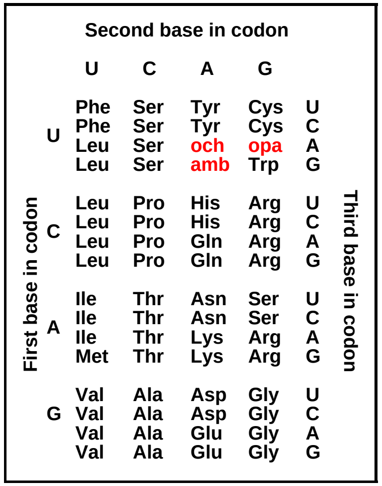
$$\epsilon_r = \frac{H_{\text{before}} - H_{\text{after}}}{6} = \frac{4.170 - H_{\text{error}}}{6} = \ln 2$$

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

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

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

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:

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

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

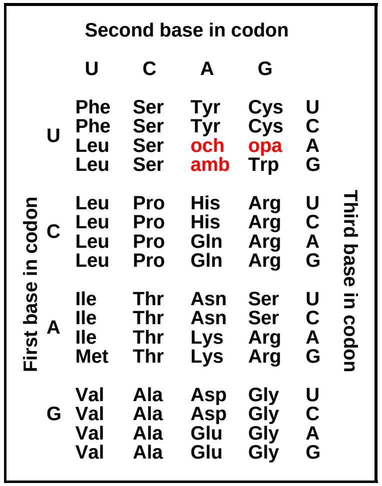
Efficiency of the Genetic Code

Combine:

Frequencies of 1 billion amino acids

with

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



	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

	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:

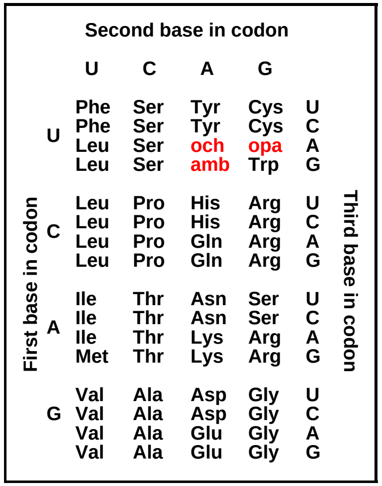
Frequencies of 1 billion amino acids

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}

$$\begin{aligned}(H_{aa} - H(P_{\text{error}}))/6 &= 0.69304765 = \text{measured efficiency}\\ \ln(2) &= 0.69314718 = \text{theoretical efficiency}\end{aligned}$$

Efficiency of the Genetic Code

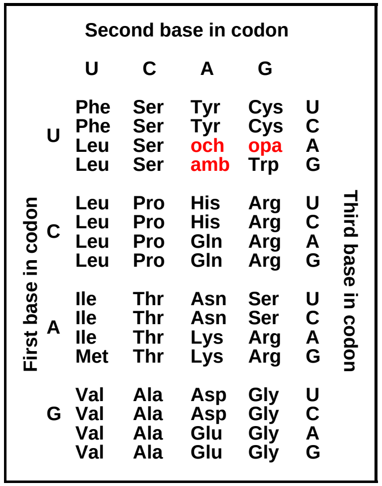
Second base in codon

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}

$$\begin{aligned} (H_{aa} - H(P_{\text{error}}))/6 &= 0.69304765 = \text{ measured efficiency} \\ \ln(2) &= 0.69314718 = \text{ theoretical efficiency} \\ \Delta &= 0.\underline{000}099530 = \text{ difference} \end{aligned}$$

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}

$$\begin{aligned}(H_{aa} - H(P_{\text{error}}))/6 &= 0.69304765 = \text{ measured efficiency}\\ \ln(2) &= 0.69314718 = \text{ theoretical efficiency}\\ \Delta &= 0.\underline{0000}99530 = \text{ difference}\end{aligned}$$

The theory matches the data to 4 decimal places!

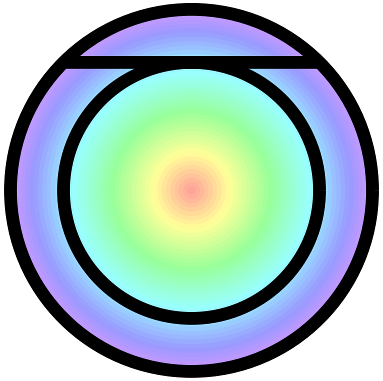
Significance of 70% efficiency

- **Establishes a novel mathematical field of biology**

Significance of 70% efficiency

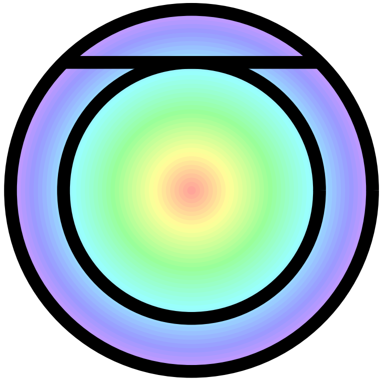
- **Establishes a novel mathematical field of biology**
- **70% efficiency implies:**

Significance of 70% efficiency



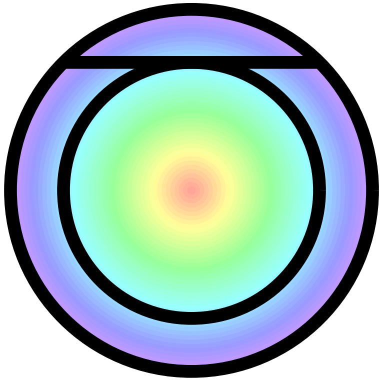
- **Establishes a novel mathematical field of biology**
- **70% efficiency implies:**
    - **Molecular machines function at channel capacity**

Significance of 70% efficiency



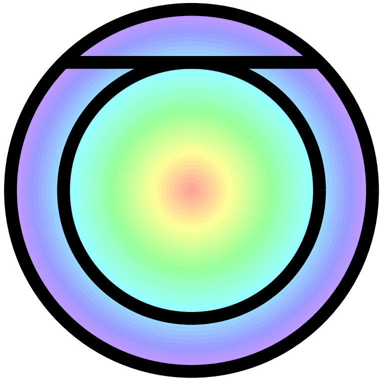
- **Establishes a novel mathematical field of biology**
- **70% efficiency implies:**
  - **Molecular machines function at channel capacity**
  - **Molecular machines are coded**

Significance of 70% efficiency



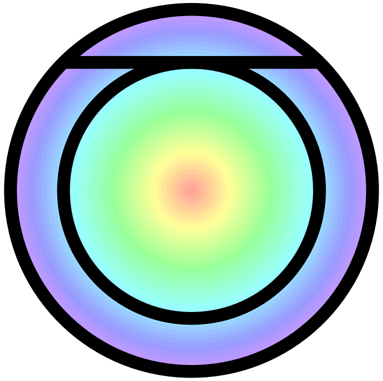
- Establishes a novel mathematical field of biology
- 70% efficiency implies:
  - Molecular machines function at channel capacity
  - Molecular machines are coded
  - Coding explains the low error rates in molecular biology

Significance of 70% efficiency



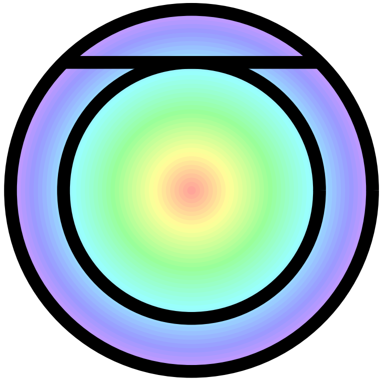
- Establishes a novel mathematical field of biology
- 70% efficiency implies:
  - Molecular machines function at channel capacity
  - Molecular machines are coded
  - Coding explains the low error rates in molecular biology
- Uses in research

Significance of 70% efficiency



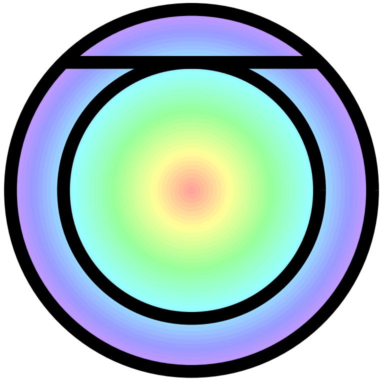
- Establishes a novel mathematical field of biology
- 70% efficiency implies:
  - Molecular machines function at channel capacity
  - Molecular machines are coded
  - Coding explains the low error rates in molecular biology
- Uses in research
  - Predict specific binding constants of proteins on DNA from sequences

Significance of 70% efficiency



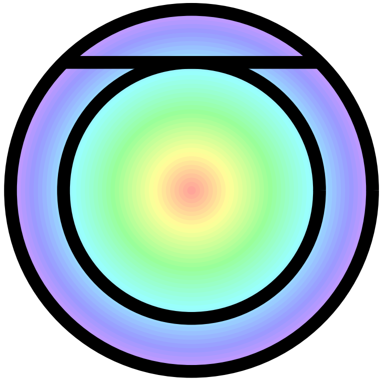
- Establishes a novel mathematical field of biology
- 70% efficiency implies:
    - Molecular machines function at channel capacity
    - Molecular machines are coded
    - Coding explains the low error rates in molecular biology
- Uses in research
    - Predict specific binding constants of proteins on DNA from sequences
    - Anomalies that do not match the theory unveil new phenomena

Significance of 70% efficiency



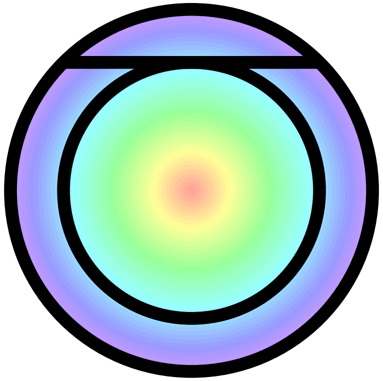
- Establishes a novel mathematical field of biology
- 70% efficiency implies:
    - Molecular machines function at channel capacity
    - Molecular machines are coded
    - Coding explains the low error rates in molecular biology
- Uses in research
    - Predict specific binding constants of proteins on DNA from sequences
    - Anomalies that do not match the theory unveil new phenomena
- Practical applications

Significance of 70% efficiency



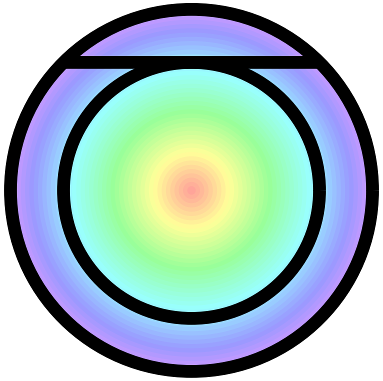
- Establishes a novel mathematical field of biology
- 70% efficiency implies:
  - Molecular machines function at channel capacity
  - Molecular machines are coded
  - Coding explains the low error rates in molecular biology
- Uses in research
  - Predict specific binding constants of proteins on DNA from sequences
  - Anomalies that do not match the theory unveil new phenomena
- Practical applications
  - Understanding how molecules use energy

Significance of 70% efficiency



- Establishes a novel mathematical field of biology
- 70% efficiency implies:
    - Molecular machines function at channel capacity
    - Molecular machines are coded
    - Coding explains the low error rates in molecular biology
- Uses in research
    - Predict specific binding constants of proteins on DNA from sequences
    - Anomalies that do not match the theory unveil new phenomena
- Practical applications
    - Understanding how molecules use energy
    - Designing robust molecular devices that function with few errors

Significance of 70% efficiency



- Establishes a novel mathematical field of biology
- 70% efficiency implies:
  - Molecular machines function at channel capacity
  - Molecular machines are coded
  - Coding explains the low error rates in molecular biology
- Uses in research
  - Predict specific binding constants of proteins on DNA from sequences
  - Anomalies that do not match the theory unveil new phenomena
- Practical applications
  - Understanding how molecules use energy
  - Designing robust molecular devices that function with few errors i.e. designing nanotechnologies at the engineering limit

Acknowledgments

Herbert A. Schneider (1922-2009)

Acknowledgments

Herbert A. Schneider (1922-2009)

John Spouge
Peter Rogan
John Garavelli

Acknowledgments

Herbert A. Schneider (1922-2009)

John Spouge
Peter Rogan
John Garavelli

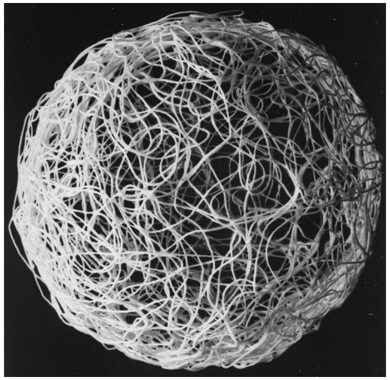
Martin Bier, Ilya Lyakhov, Danielle Needle, Peyman Khalichi, Carrie Paterson, Ryan Shultzaberger, Amar Klar, Peter Lemkin, Barry Zeeberg, Lynn Bayer, Zehua Chen, Blake Sweeney, Bert Gold, Sorina Eftim, Mikhail Kashlev, Alex Mitrophanov, Peter Thomas, and Hong Qian

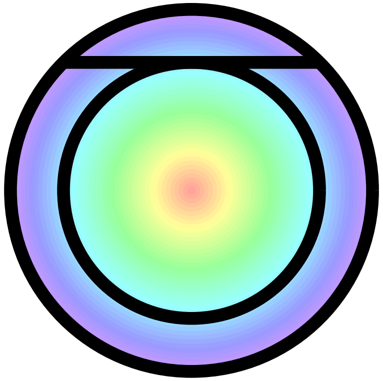
National Institutes of Health, National Cancer Institute

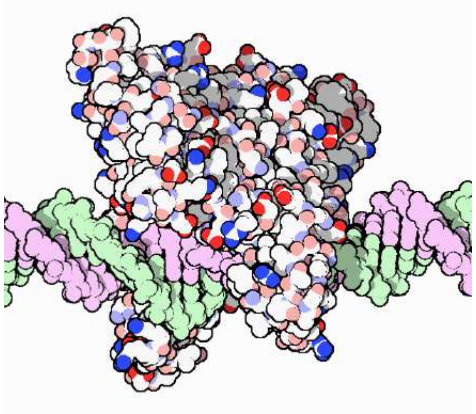
DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

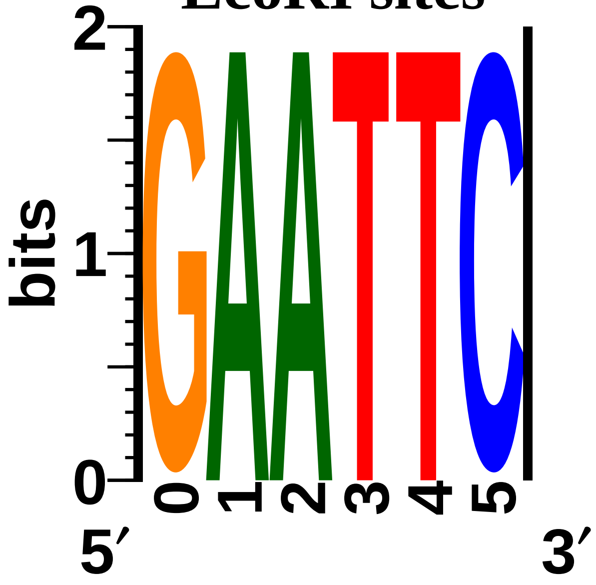
Web site:

TinyURL.com/tomschneider








EcoRI sites

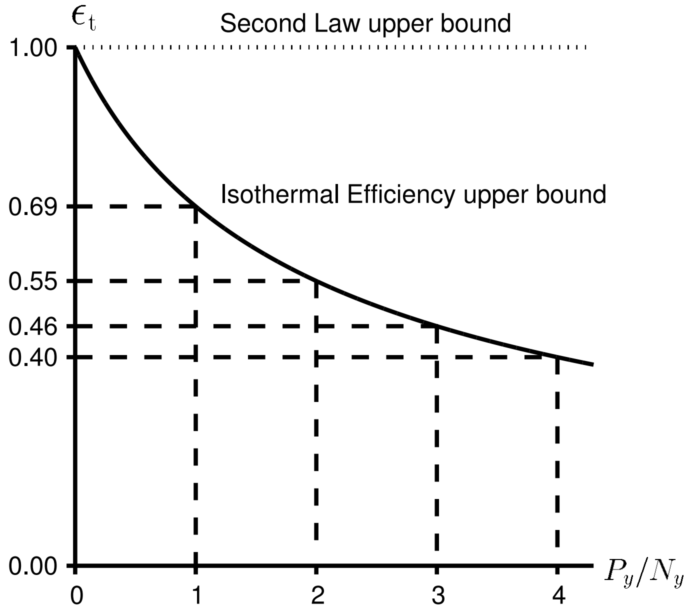
Second base in codon

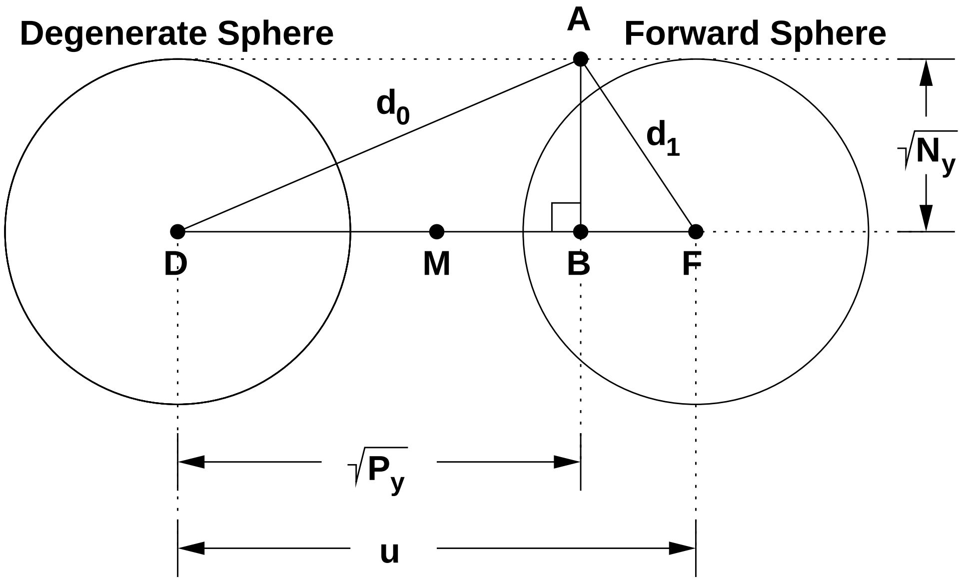
First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	och	opa	A
	Leu	Ser	amb	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val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

Version

version = 1.33 of codetalk.tex 2011 Feb 25

Proof that $P_y > N_y$, $\epsilon < \ln(2)$





buffer zone: $u > 2\sqrt{N_y}$ (0)

distance² from A to D: $d_0^2 = \sqrt{P_y}^2 + \sqrt{N_y}^2 = P_y + N_y$ (1)

distance² from A to F: $d_1^2 = (u - \sqrt{P_y})^2 + \sqrt{N_y}^2$ (2)

decoding to forward sphere: $d_1 < d_0$ (3)

(1) and (2) into square of (3): $\sqrt{P_y} > u/2$ (4)

from (0) and (4): $\sqrt{P_y} > \sqrt{N_y}$ **so** $P_y > N_y$ (5)

$$\epsilon = \frac{\ln\left(\frac{P_y}{N_y} + 1\right)}{\frac{P_y}{N_y}} \quad \textbf{so} \quad \epsilon < \ln(2) \approx 0.6931$$

More Information Theory - 1

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \tag{6}$$

More Information Theory - 1

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\begin{aligned} \log_2 M \\ &= -\log_2 1/M. \end{aligned} \tag{6}$$

$1/M$ is like the probability of a symbol.

More Information Theory - 1

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \tag{6}$$

$$= \quad -\log_2 1/M.$$

$1/M$ is like the probability of a symbol.

If the probabilities P_i of different symbols, i, are not equal, then the **surprisal** is:

$$u_i \equiv -\log_2 P_i. \tag{7}$$

how surprised one is to see a symbol

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \quad (8)$$

$$P_{\text{silent}} = 1023/1024 \quad (9)$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{8}$$

$$P_{\text{silent}} = 1023/1024 \tag{9}$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \ \text{bits} \tag{10}$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \ \text{bits} \tag{11}$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{8}$$

$$P_{\text{silent}} = 1023/1024 \tag{9}$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \ \text{bits} \tag{10}$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \ \text{bits} \tag{11}$$

The **average surprisal** is called the **uncertainty**, H:

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}}$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \quad (8)$$

$$P_{\text{silent}} = 1023/1024 \quad (9)$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \text{ bits} \quad (10)$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \text{ bits} \quad (11)$$

The **average surprisal** is called the **uncertainty**, H:

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}} + P_{\text{silent}} \times \text{surprisal}_{\text{silent}} \quad (12)$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{8}$$

$$P_{\text{silent}} = 1023/1024 \tag{9}$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \ \text{ bits} \tag{10}$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \ \text{ bits} \tag{11}$$

The **average surprisal** is called the **uncertainty**, H:

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}} + P_{\text{silent}} \times \text{surprisal}_{\text{silent}} \tag{12}$$

$$H = P_{\text{ring}} \times \Big(-\log_2(P_{\text{ring}})\Big) + P_{\text{silent}} \times \Big(-\log_2(P_{\text{silent}})\Big) \tag{13}$$

More Information Theory - 3

For M symbols use the sum ($\sum$) notation:

$$H \;=\; \sum_{i=1}^{M} P_i \times (\text{surprisal for} P_i) \tag{14}$$

More Information Theory - 3

For M symbols use the sum ($\sum$) notation:

$$H = \sum_{i=1}^{M} P_i \times (\text{surprisal for} P_i) \tag{14}$$

$$= \sum_{i=1}^{M} P_i \times (-\log_2 P_i) \tag{15}$$

More Information Theory - 3

For M symbols use the sum ($\sum$) notation:

$$
\begin{aligned}
H &= \sum_{i=1}^{M} P_i \times (\text{surprisal for} P_i) && (14) \\
&= \sum_{i=1}^{M} P_i \times (-\log_2 P_i) && (15) \\
&= -\sum_{i=1}^{M} P_i \log_2 P_i \quad \text{bits per symbol} && (16)
\end{aligned}
$$

More Information Theory - 4

$\boxed{\text{Information is a decrease in uncertainty}}$

$$R = H_{\text{before}} - H_{\text{after}} \tag{17}$$

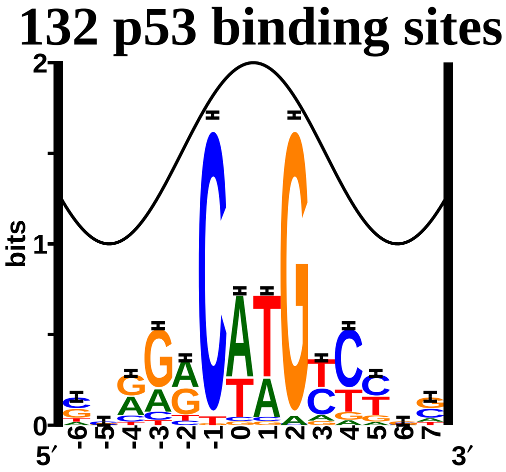
More Information Theory - 4

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \tag{17}$$

Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \tag{18}$$



More Information Theory - 4

Information is a decrease in uncertainty

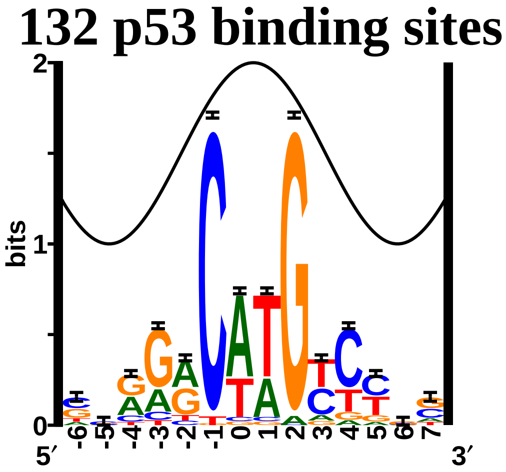
$$R = H_{\text{before}} - H_{\text{after}} \tag{17}$$

Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \tag{18}$$

and

$$\begin{aligned} H_{\text{after}} &= \text{uncertainty of bases} \\ &= -\sum_{base=A}^{T} P_{base} \log_2 P_{base} \end{aligned} \tag{19}$$



More Information Theory - 4

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \qquad (17)$$

Example a sequence logo is computed from equiprobable bases before:

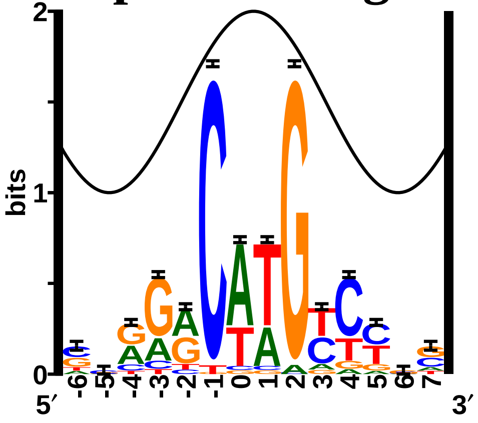
$$H_{\text{before}} = 2 \text{ bits/base} \qquad (18)$$

and

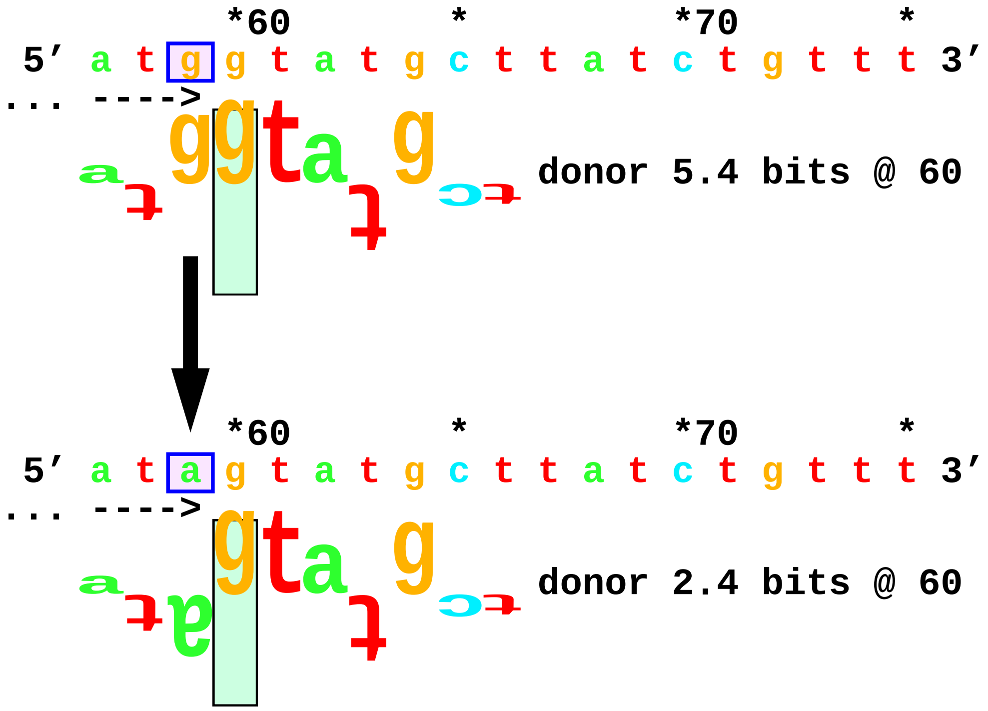
$$\begin{aligned} H_{\text{after}} &= \text{uncertainty of bases} \\ &= -\sum_{base=A}^{T} P_{base} \log_2 P_{base} \qquad (19) \end{aligned}$$

Note: with only one base, $H_{\text{after}} = 0$
so $R = 2$ bits/base.

132 p53 binding sites



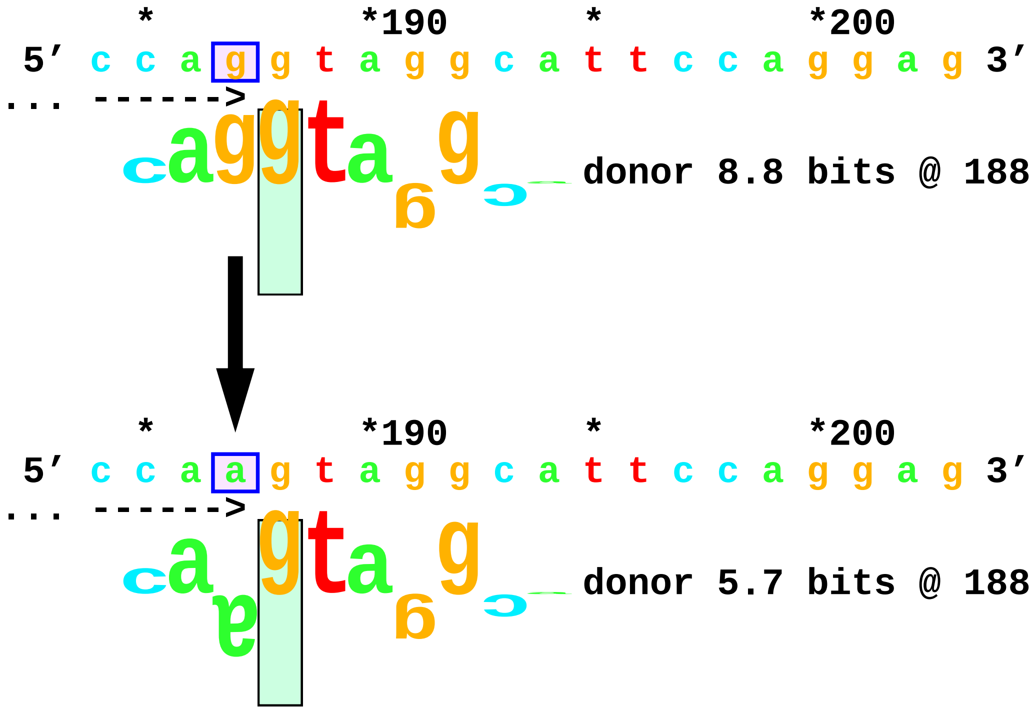
Splice Junction Mutation by Sequence Walkers



COL1A2 gene results in 50% exon skipping and Ehlers-Danlos syndrome

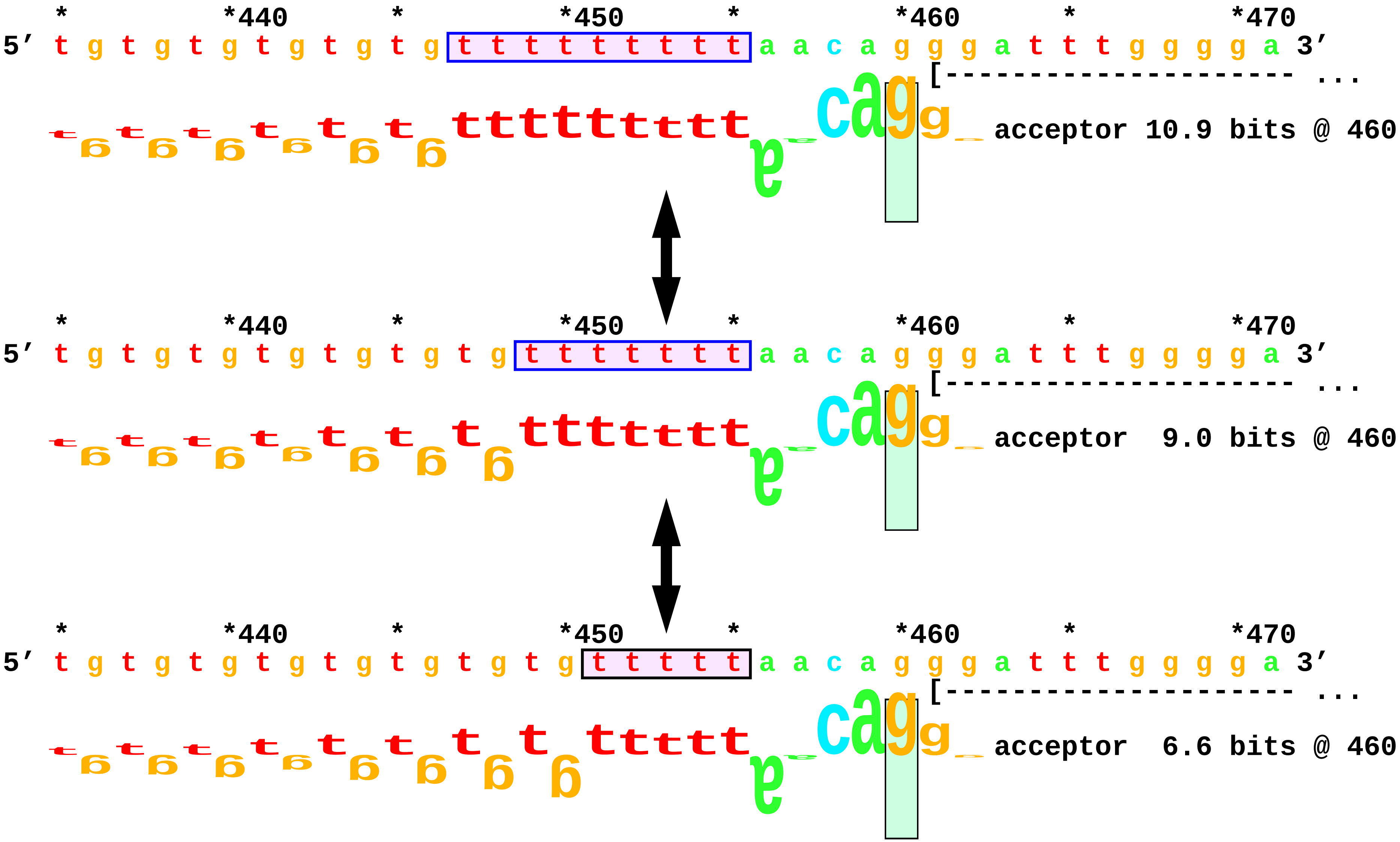
Rogan, Faux, Schneider, Human Mutation 12: 153-171 (1998)

Leaky Splice Junction Mutation



Lysosomal lipase gene [LIPA]
mild cholesterol ester storage disease with 4-9% enzymatic activity
Rogan, Faux, Schneider, Human Mutation 12: 153-171 (1998)

Polymorphic Variation Affects Splicing

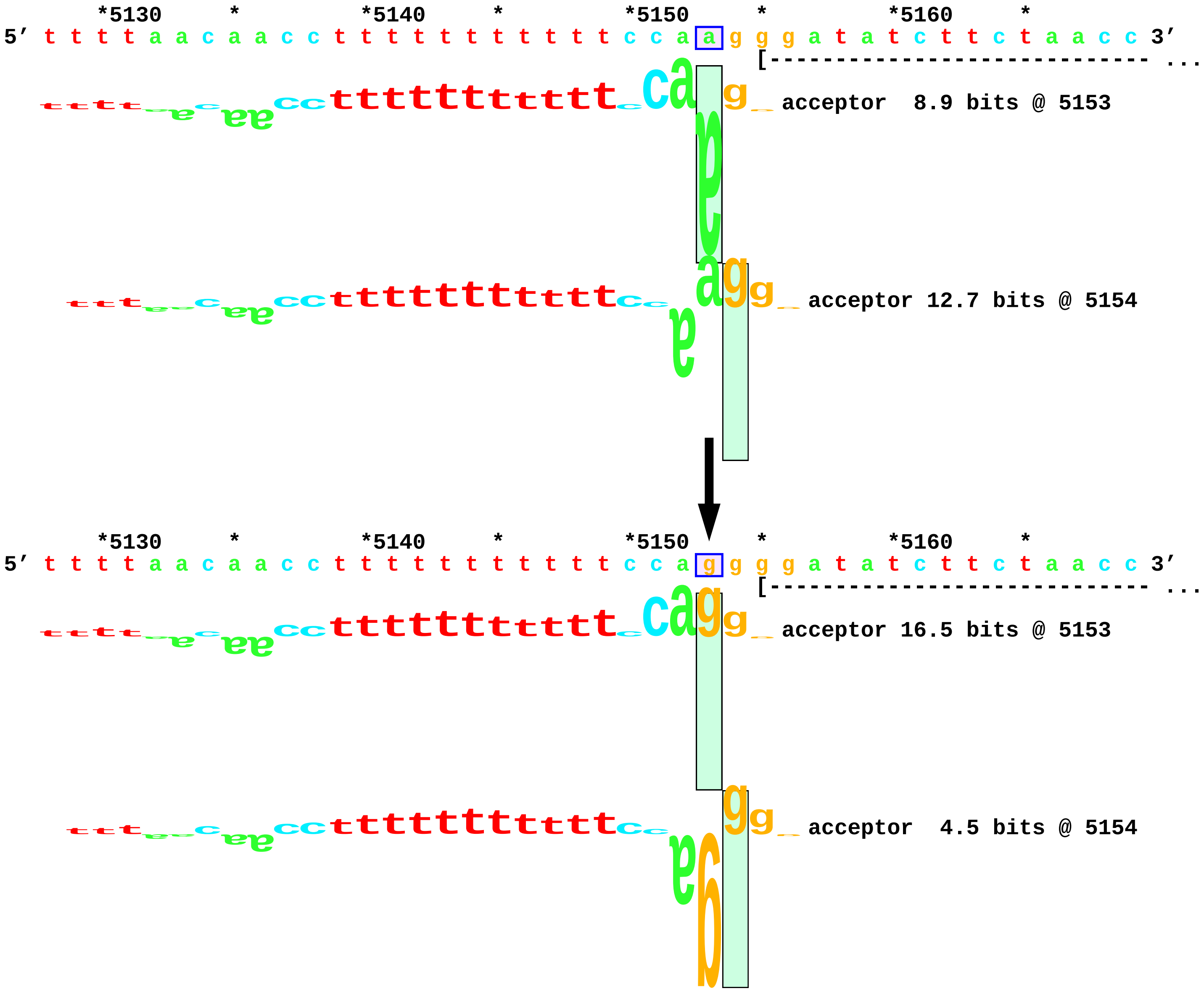


```
   *          *440        *          *450        *          *460        *          *470
5′ t g t g t g t g t g t g t t t t t t t t t a a c a g g g a t t t g g g g a 3′
                                                       [---------------------- ...
                                                         acceptor 10.9 bits @ 460

   *          *440        *          *450        *          *460        *          *470
5′ t g t g t g t g t g t g t g t t t t t t t a a c a g g g a t t t g g g g a 3′
                                                       [---------------------- ...
                                                         acceptor  9.0 bits @ 460

   *          *440        *          *450        *          *460        *          *470
5′ t g t g t g t g t g t g t g t g t t t t t a a c a g g g a t t t g g g g a 3′
                                                       [---------------------- ...
                                                         acceptor  6.6 bits @ 460
```

Cystic fibrosis transmembrane regulator [CFTR]

Rogan, Faux, Schneider, Human Mutation 12: 153-171 (1998)

Cryptic Site Generation and Mutation of Natural Site



Iduronidase synthetase gene [IDS]

Rogan, Faux, Schneider, Human Mutation 12: 153-171 (1998)