ICCT in biology at the molecular and cellular level – some steps in unveiling the protection and prioritization in the DNA

Dawit Nigatu and **Werner Henkel**
(collaboration with Patrick Sobetzko and **Georgi Muskhelishvili**)

Jan 21, 2020

Transmission Systems Group
Jacobs University Bremen
`http://trsys.faculty.jacobs-university.de/`

BioTICC NSF workshop, Alexandria, VA

JACOBS
UNIVERSITY

Outline

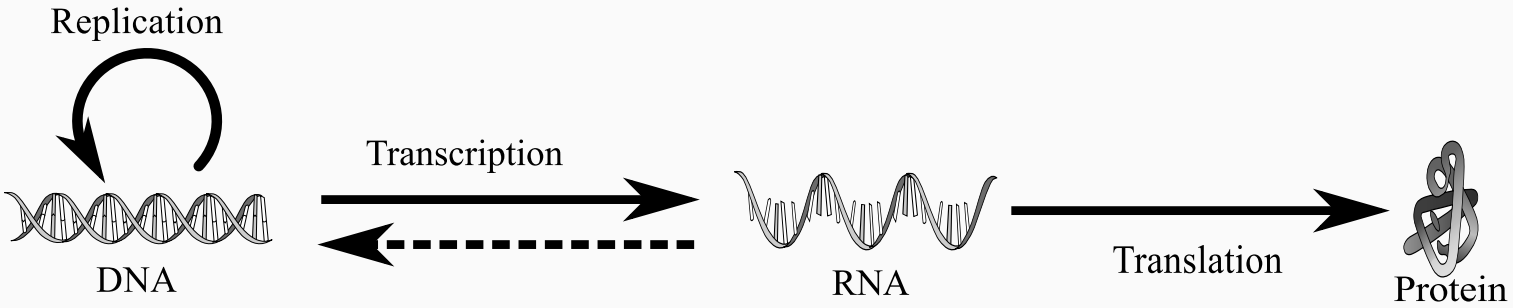
- Mapping Between Codons and Amino Acids
  - ✓ A Channel Model from a Mutation Matrix
  - ✓ Set Partitioning
- Digital Information and Thermodynamic Stability
- Essential and Non-essential Genes
- Conclusion and Discussing Open Questions

Mapping Between Codons and Amino Acids

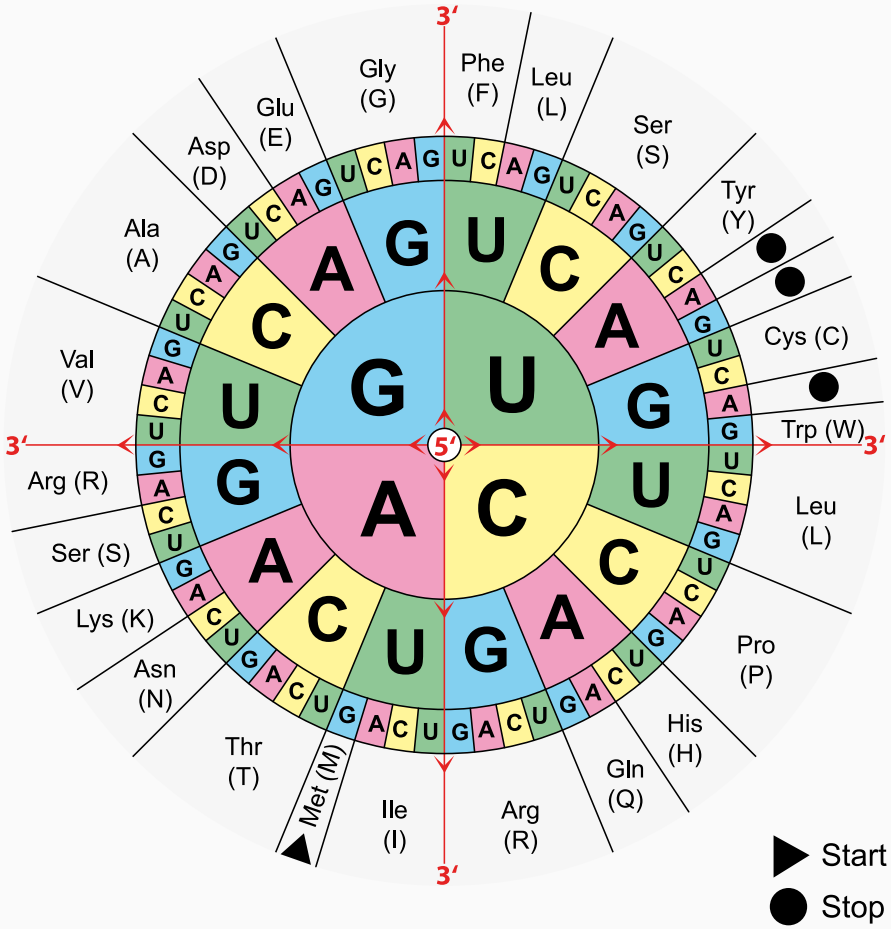
Background and Motivation

Flow of biological information

Modified from refs [1] & [2]



The Genetic Code



The genetic code chart [3]

[a]

The genetic code

- Degenerate: synonymous codons provide redundancy
- Optimal: minimizing substitution and frame-shift errors
- "One in a million": outperforms randomly generated codes

[a] **Freeland1998.**

Substitution Matrices

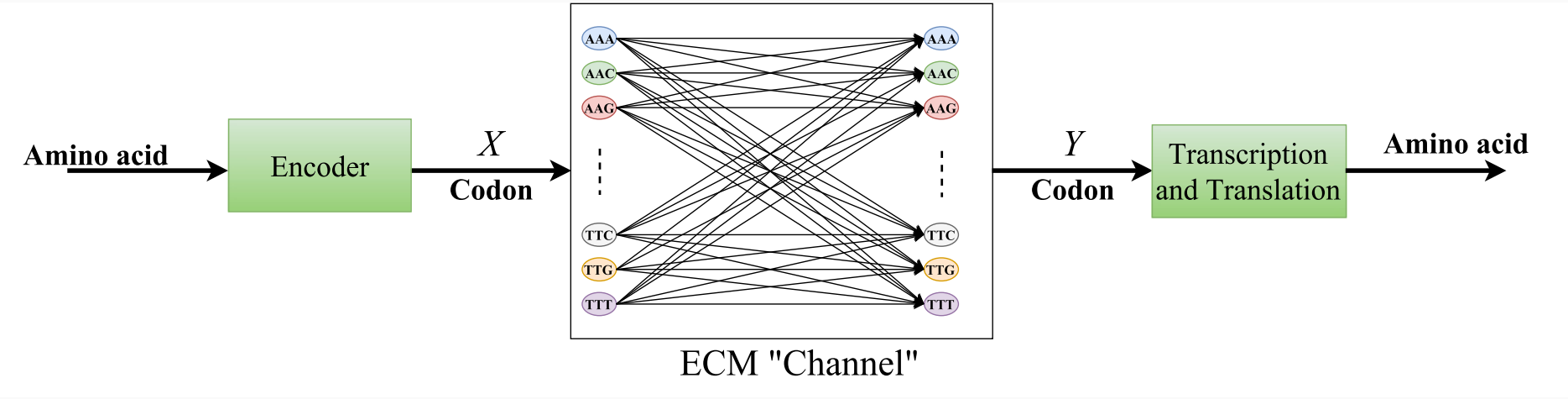
- Nucleotide-based models
  - Jukes and Cantor, Kimura ...
- Protein-based models
  - PAM, BLOSUM, WAG, ...
- Codon-based models
  - Empirical codon mutation (ECM), Goldman and Yang, ...

ECM matrix

- Proposed by Schneider et al.[1] in 2005
- 17,502 alignments from five vertebrates
- Estimated from 8.3 million aligned codons

[1] **schneider2005empirical**.

ECM Channel Model



ECM "Channel" [a]

- Mutation matrix describes a channel transition probability matrix $\mathbf{P}(y|x)$
- Using SVD for matrix exponentiation

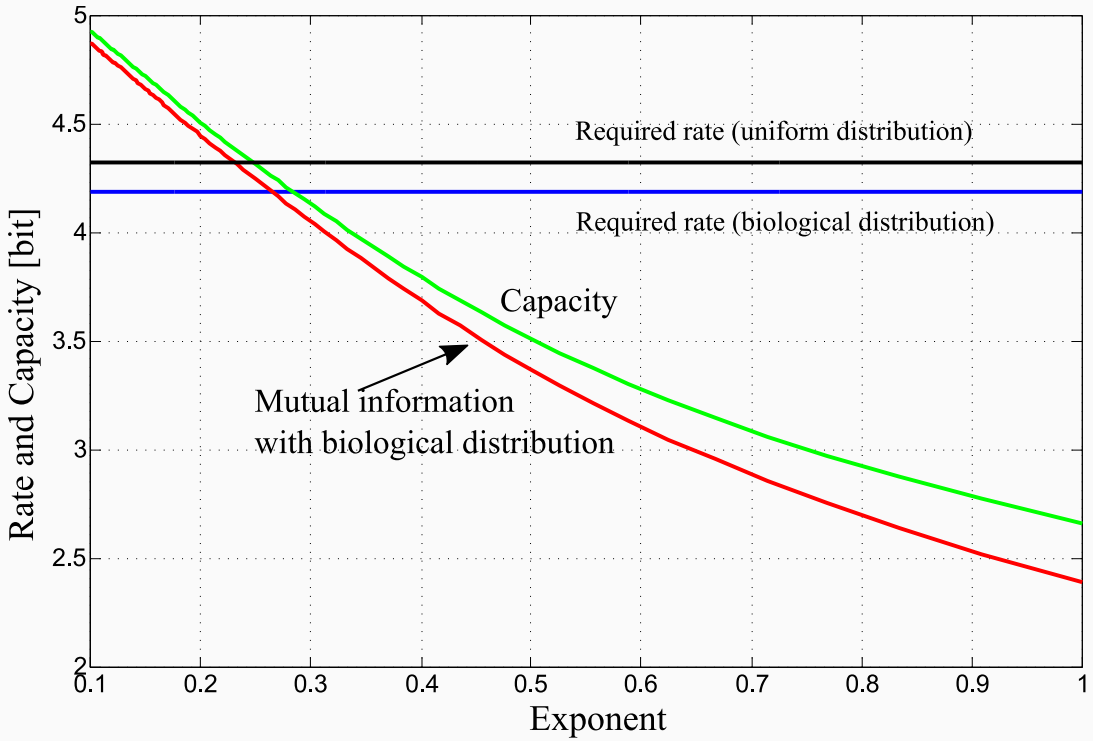
$$[\mathbf{P}(y|x)]^F = \mathbf{U}(\mathbf{\Sigma})^F \mathbf{V}^\mathsf{T},$$

where $\mathbf{U}, \mathbf{V}$ are unitary matrices and $\mathbf{\Sigma}$ is a diagonal matrix

Find the optimal exponent for error-free transmission

[a] nigatu2014empirical

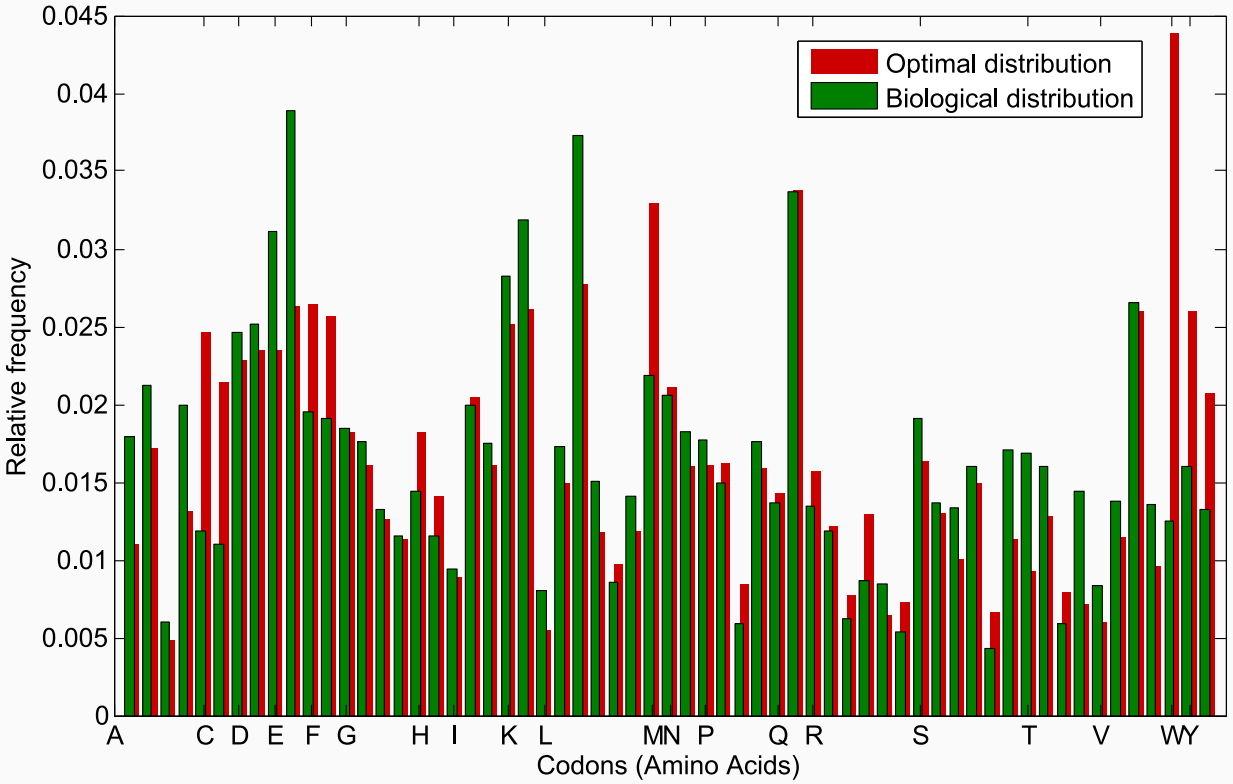
Capacity of the ECM Channel



- Optimal exponent = 0.26
  $\Rightarrow$ Mutation rate = 29%
- Capacity curve is very close to the mutual information curve
  $\Rightarrow$ **The biological distribution is optimally "chosen"**

Biological distribution $\approx$ Optimal distribution

Results



Observations

- $D_{KL}(\text{observed}||\text{optimal}) = \mathbf{0.0926}$ bit

  $\Rightarrow$ Comparable with $D_{KL}(N(\mu;\sigma)||N(\mu;2\sigma))$

- Distribution among synonymous codons is similar

Mutation vs. Chemical Distances

Grantham's[2] chemical distance matrix

- **Composition, polarity, and molecular volume**
- 20×20 distance matrix

Compare the mutation and chemical distance matrices

Classical multidimensional scaling (CMDS)

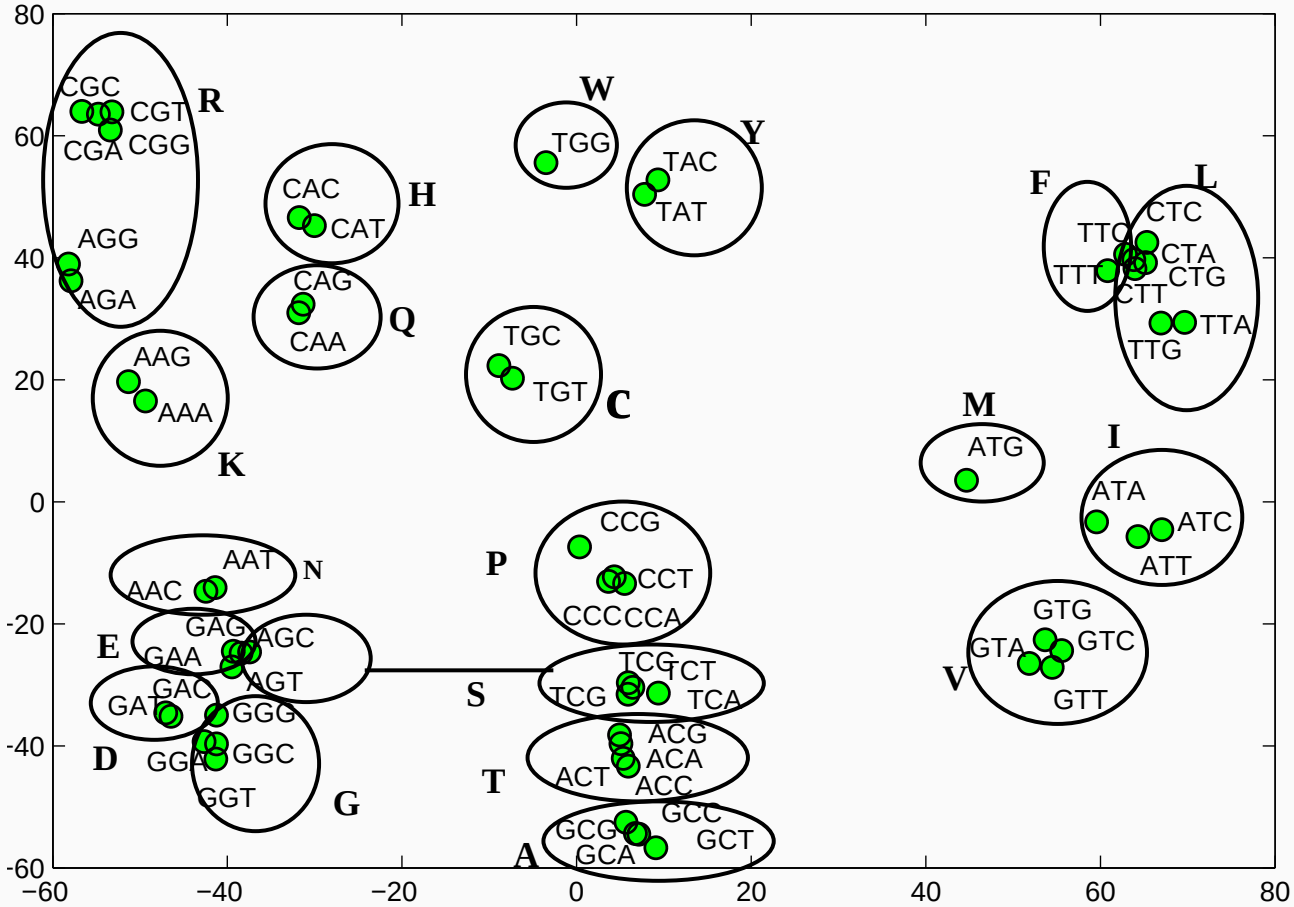
- **Given pairwise dissimilarities, reconstruct a map that preserves distances**
- ECM matrix: 61×61 probability matrix

  $\Rightarrow$ pairwise point distances are computed assuming a Gaussian i.i.d. "channel"

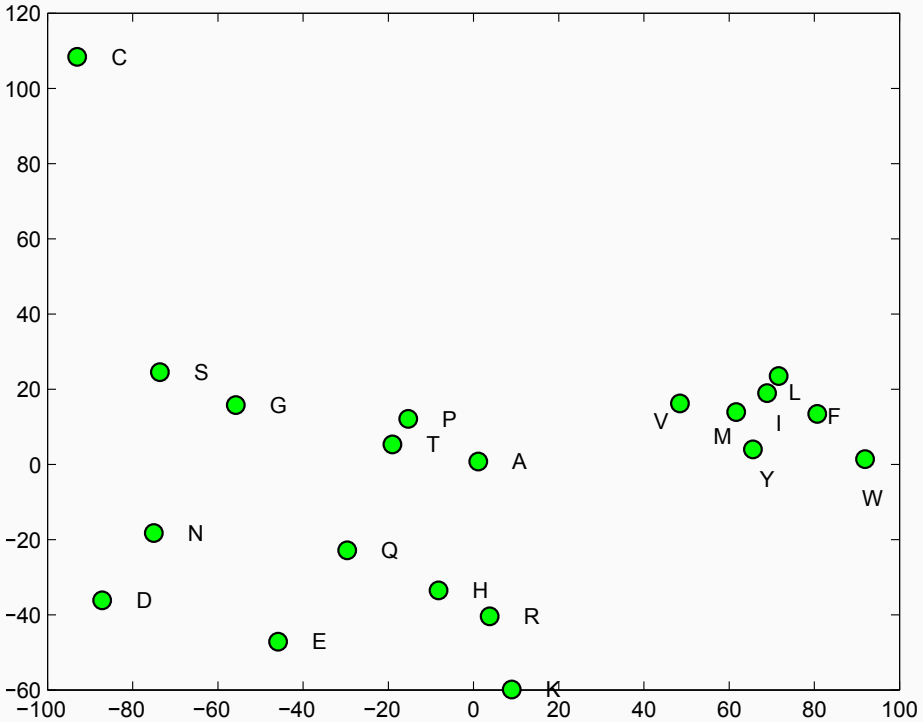
$$P_{ij} = \frac{1}{2}\operatorname{erfc}\left(\frac{D_{ij}}{\sqrt{2}\sigma}\right)$$

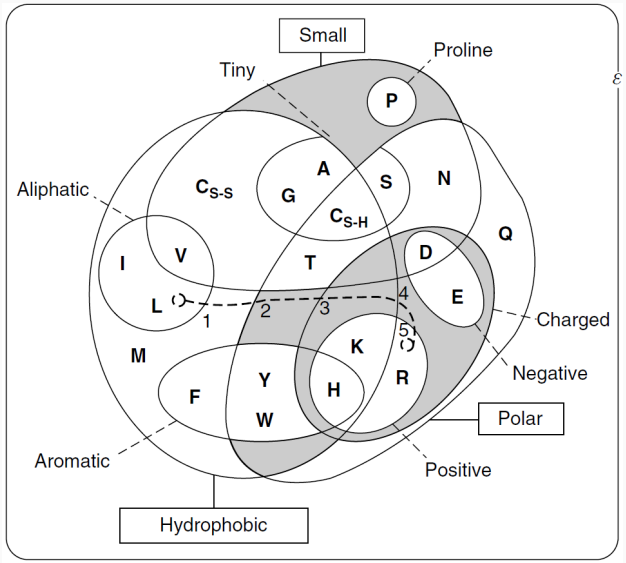
[2] **Grantham1974**.

2D-view of the codon distance matrix



2D-View of the Chemical Distance Matrix





Taylor classification of amino acids [4]

- Synonymous codons are clustered together
- Highly probable mutation are between chemically similar amino acids

Inconsistencies

Large chemical distance but small mutation distance:

- **C** with **“all others”**
- **G** with **E**
- **S** with **{P,T,A}**
- **{D,N}** with **E**
- **{D,N}** with **G**
- **{Q,H}** with **{W,Y}**
- **K** with **N**

Small chemical distance but large mutation distance:

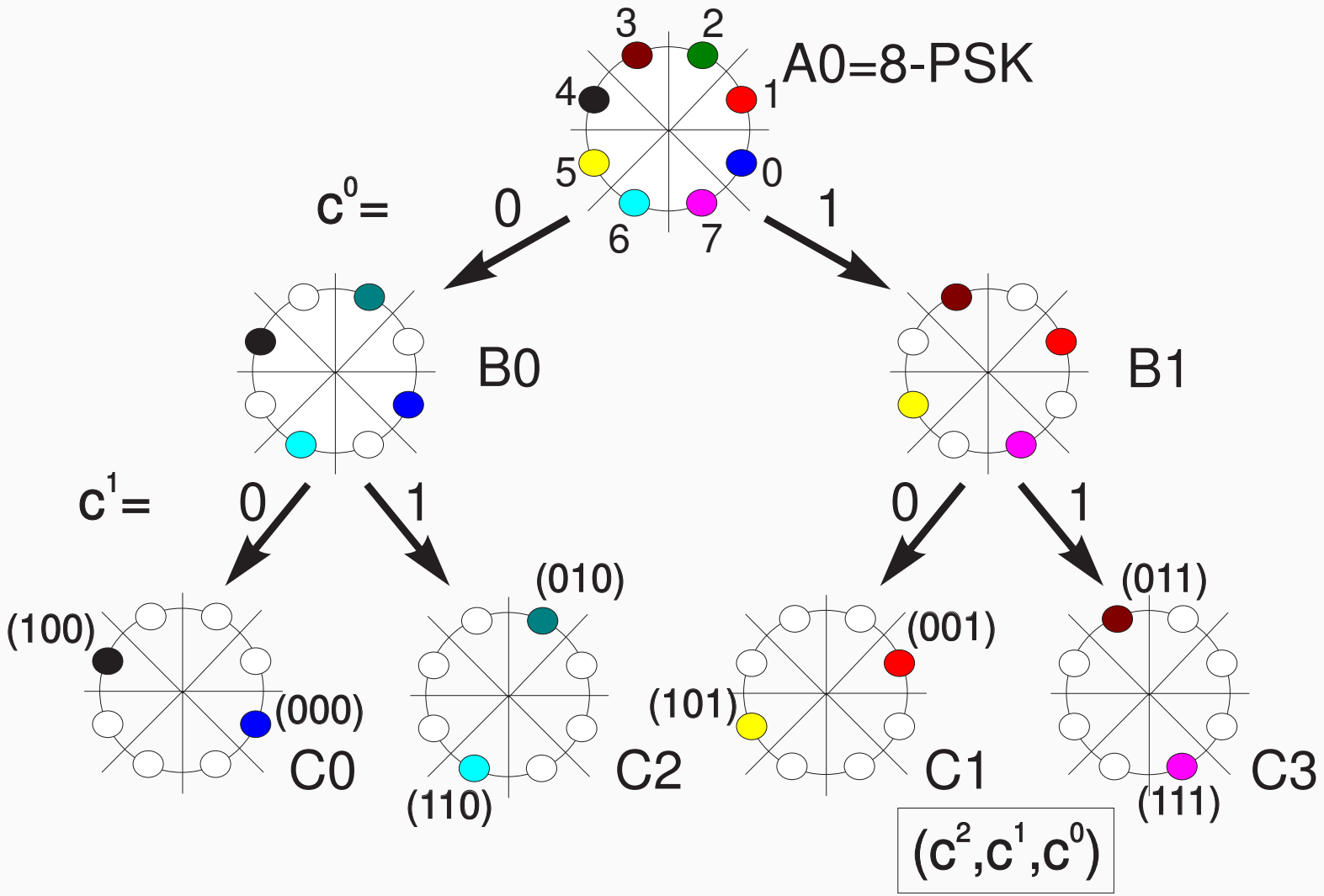
- **{W,Y}** with **{F,L,M,I,V}**
- **{P,T,A}** with **{Q,H,R}**

Explaining the inconsistencies?

- Another level of error protection (Coded Modulation, Multilevel Coding)

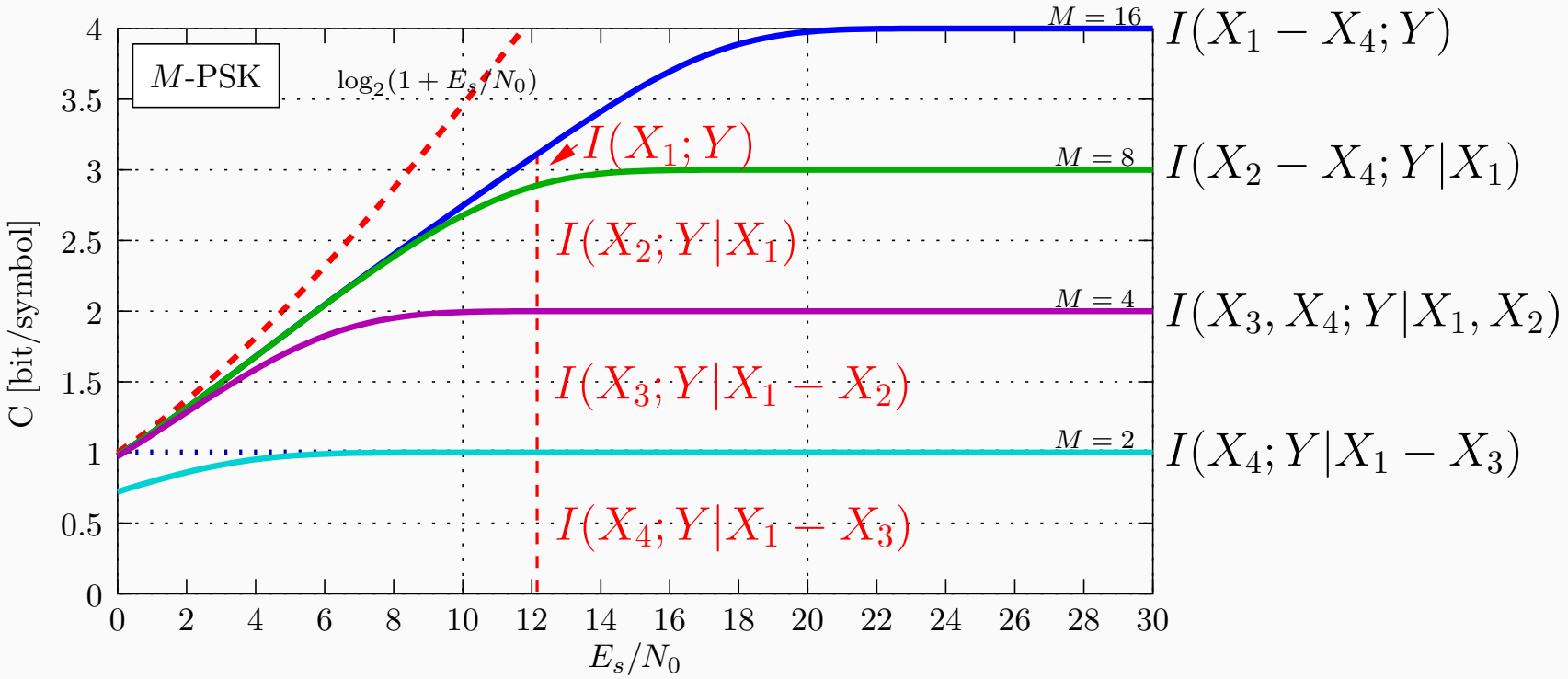
Set Partitioning

Ungerböck's mapping by set partitioning



Multilevel code rates

- Every level is protected with a separate code
- Following the *Chain Rule*, code rates are obtained as the differences between neighboring capacity curves

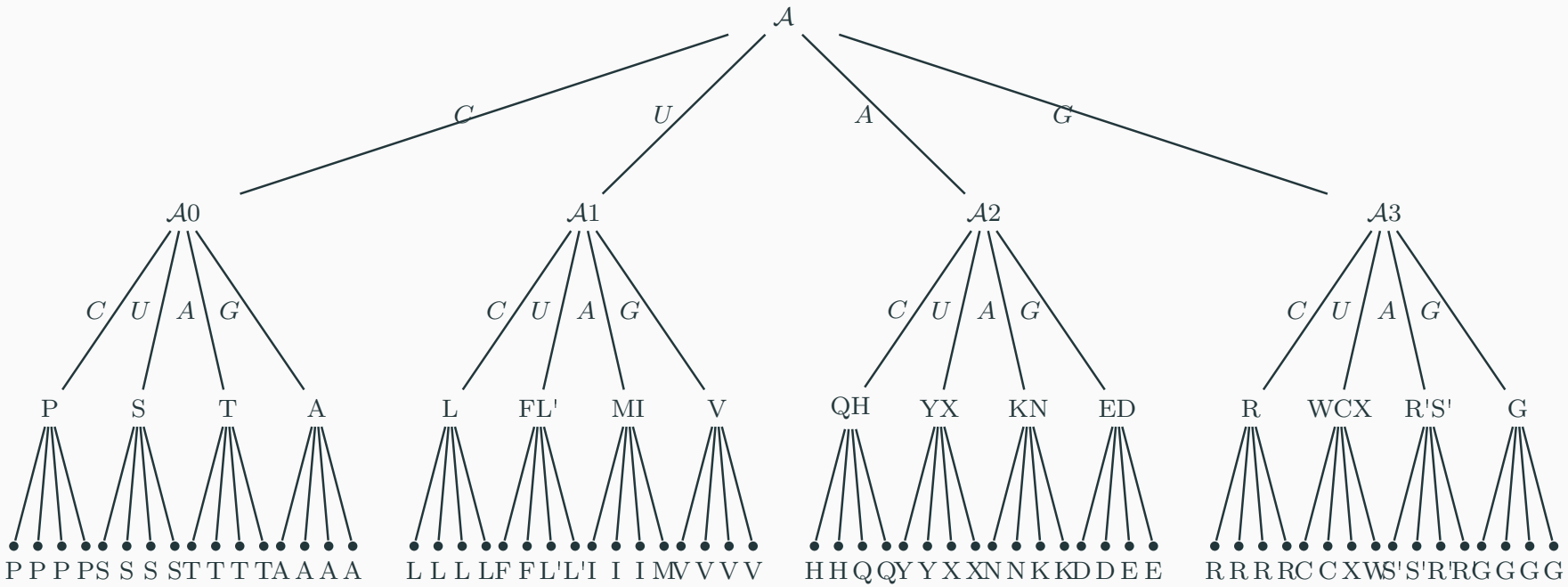


Set Partitioning of the Genetic Code

4-ary set partitioning

- **Block partitioning is preferred:** closest points are similar
- **Start with the second position:** it is the most informative

2nd → 1st → 3rd

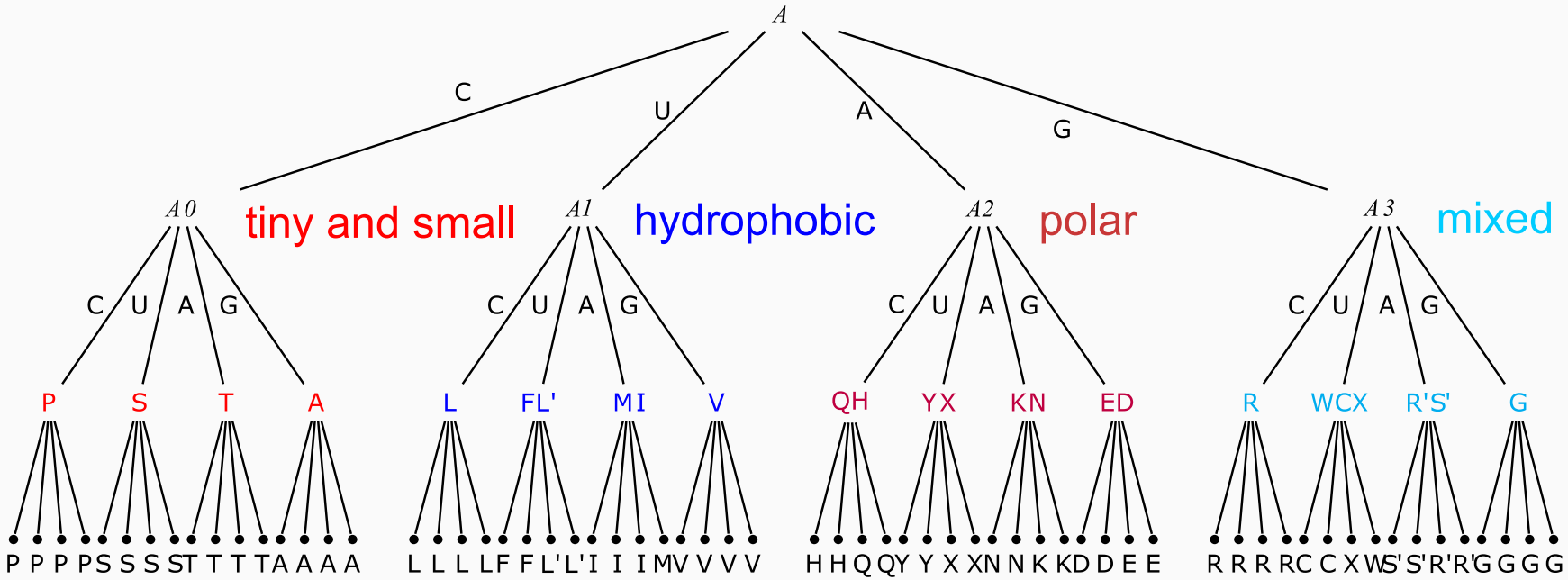


1st Partition Level

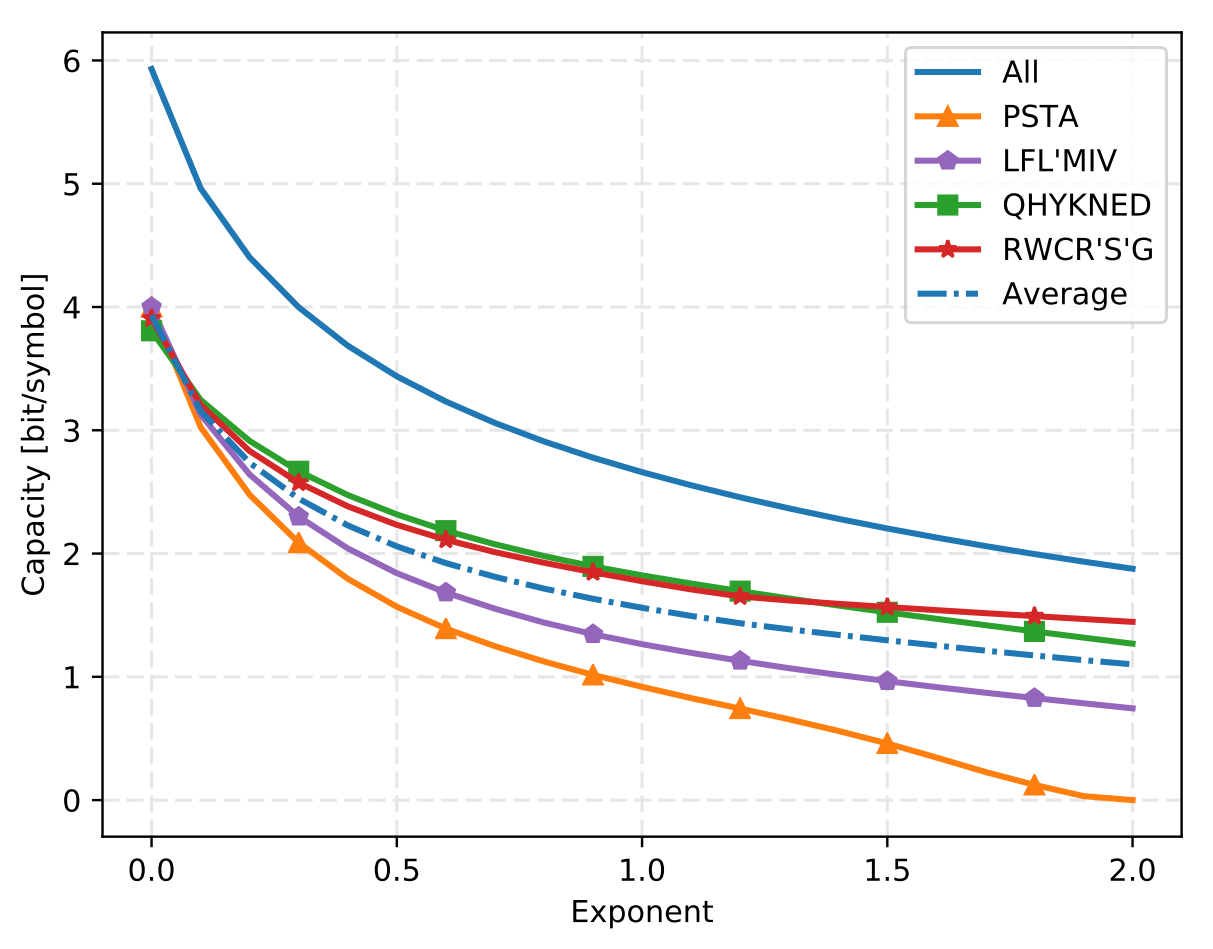
- $\mathcal{A}$ is the set of all codons
- X_1, X_2, and X_3 are the three codon positions

$$I(Y;X) = I(Y;X_1,X_2,X_3)$$
$$= \underbrace{I(Y;X_2)}_{\text{1st partition}} + \underbrace{I(Y;X_1|X_2)}_{\text{2nd partition}} + \underbrace{I(Y;X_3|X_1,X_2)}_{\text{3rd partition}}$$

$$\mathcal{A}0 = \mathcal{A}(x_2 = C)$$
$$\mathcal{A}1 = \mathcal{A}(x_2 = U)$$
$$\mathcal{A}2 = \mathcal{A}(x_2 = A)$$
$$\mathcal{A}3 = \mathcal{A}(x_2 = G)$$

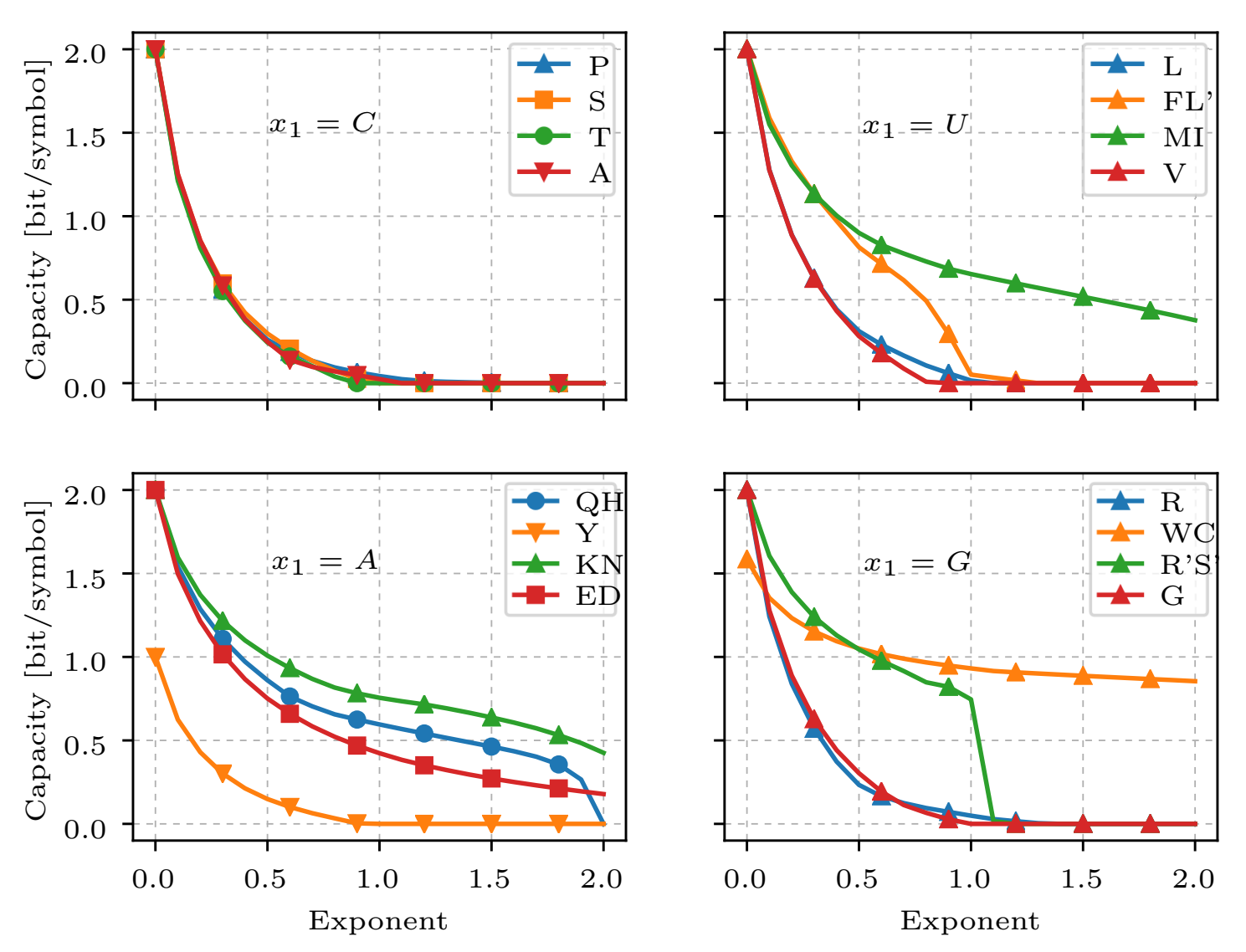


Capacities at the 1st Partition



- {P,S,T,A} sub group relatively smaller information

Capacities at the 2nd Partition



- Synonymous codons → small inter-distances → vanishing capacities
- {W, C} → high capacity even for large "SNR"

Level Capacities

The level capacity C^1 of the 1st partition level

$$C^1 = I(Y; X_2) = I(Y; X_1, X_2, X_3) - I(Y; X_1, X_3|X_2)$$

Similarly, the capacity of the 2nd partition level, C^2,

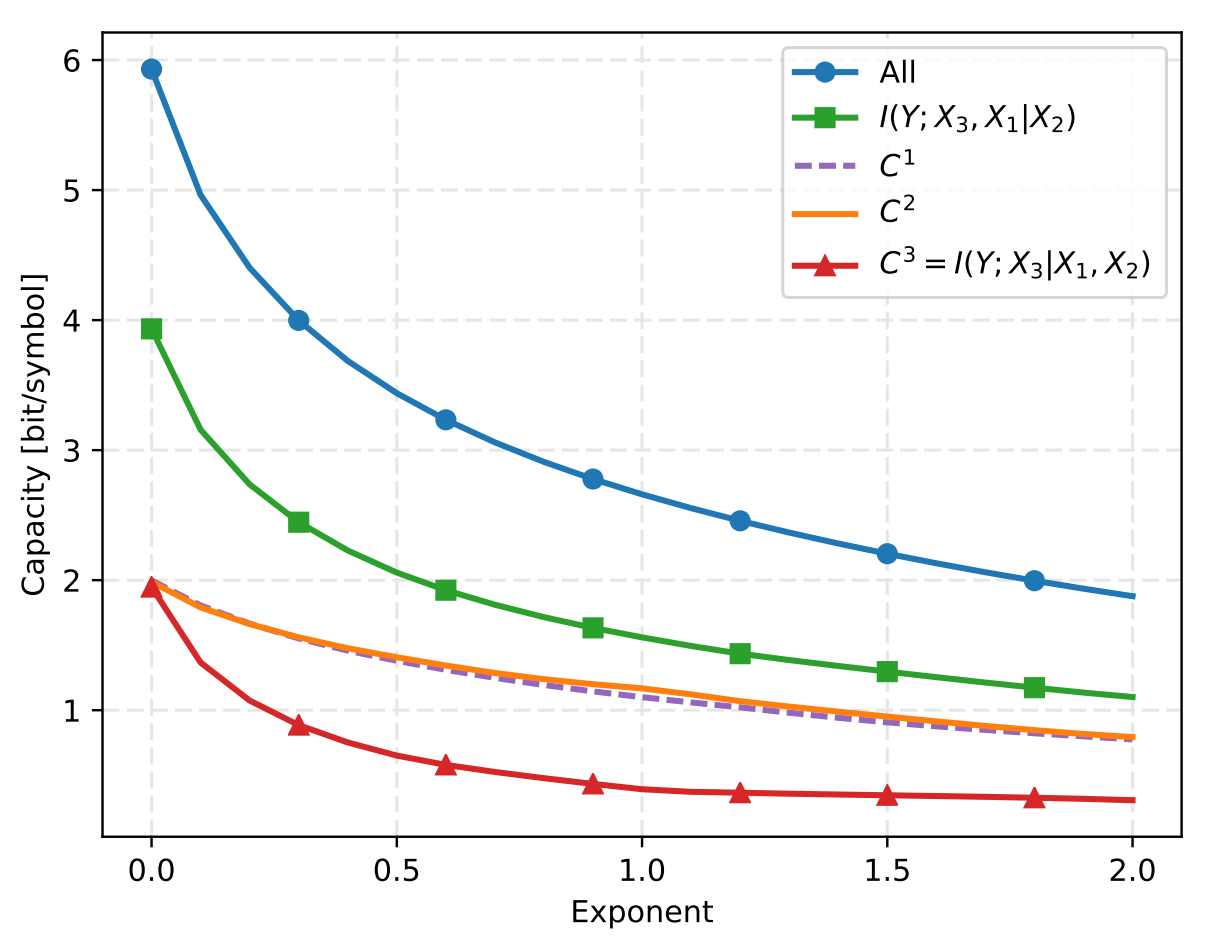
$$C^2 = I(Y; X_1|X_2) = I(Y; X_1, X_3|X_2) - I(Y; X_3|X_1, X_2)$$

where

$$I(Y; X_3|X_1, X_2) = \mathbb{E}_{x_1,x_2}\{I(Y; X_3|x_1, x_2)\}$$

C^1 and C^2 specify the required code rates
$C^3 = I(Y; X_3|X_1, X_2)$

Level Capacities



What does $C^1 = C^2$ mean?

How to transmit 4 symbols using a channel capacity of 1 bit?

Digital Information and Thermodynamic Stability in Bacteria

Digital information

- Information to encode proteins and RNA molecules
- Apparent from the quaternary alphabet

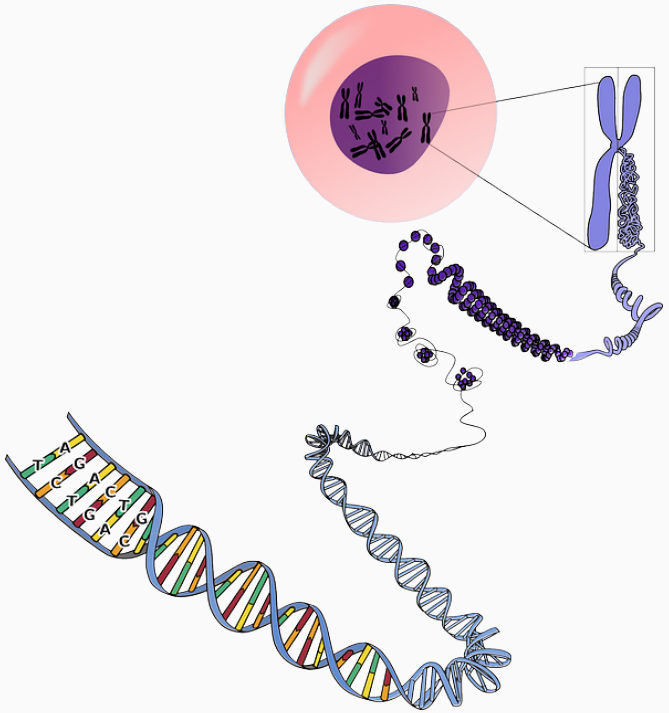
Digital and analog information

Digital information

- Information to encode proteins and RNA molecules
- Apparent from the quaternary alphabet

A co-existent “analog” information

- Defined by sequence-dependent physicochemical properties of the DNA polymer
- Dynamic structural and topological variations
- Facilitating and regulating the gene expression, chromosome compaction, and replication



Shannon and Gibbs entropies

DNA Sequence

ATCGGTAACCCGGTAGGTAACGGTATT.....

Shannon's block entropy for a block size of N symbols is

$$H_N = -\sum_i P_s^{(N)}(i) \log_2 P_s^{(N)}(i)$$

Shannon and Gibbs entropies

DNA Sequence

ATCGGTAACCCGGTAGGTAACGGTATT.....

Shannon's block entropy for a block size of N symbols is

$$H_N = -\sum_i P_s^{(N)}(i) \log_2 P_s^{(N)}(i)$$

The Gibbs entropy is given by

$$S_G = -k_B \sum_i P_G(i) \ln P_G(i)$$

k_B is the Boltzmann constant

Shannon entropy $\rightarrow$ digital information
Gibbs entropy $\rightarrow$ analog information $\rightarrow$ thermodynamic stability

Stability of DNA

- Stacking between adjacent bases
- Hydrogen bonding between complementary bases

AGTGGTAACCC
TCACCATTGGG

Stability quantified by energy values

- SantaLucia's unified free energy parameters for base pairs ($N = 2$)
- For $N > 2$, neighboring base steps are added

Stability of DNA

- Stacking between adjacent bases
- Hydrogen bonding between complementary bases

AGTGGTAACCC
TCACCATTGGG

Stability quantified by energy values

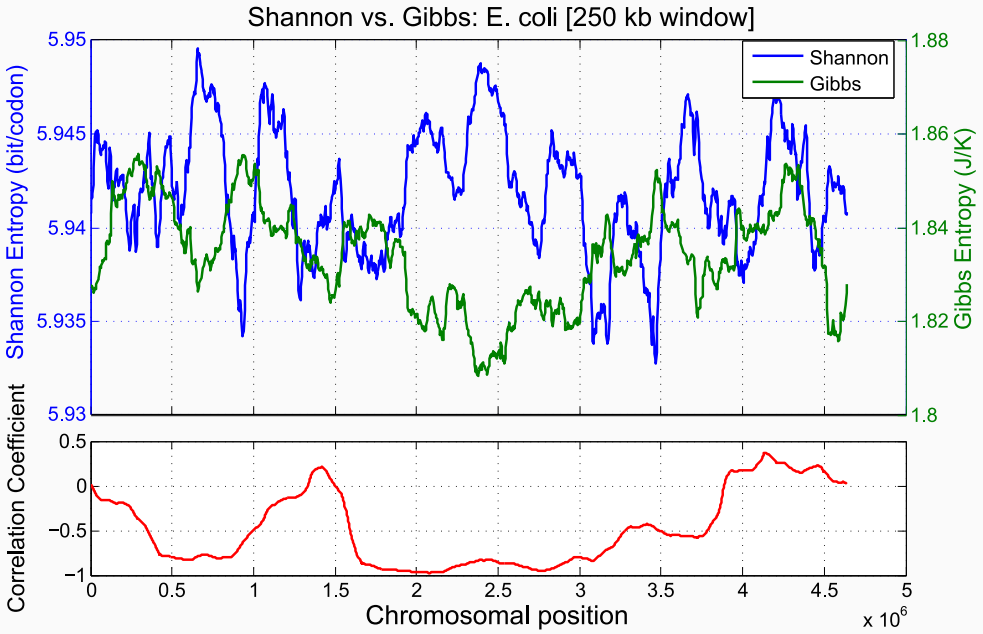
- SantaLucia's unified free energy parameters for base pairs ($N = 2$)
- For $N > 2$, neighboring base steps are added

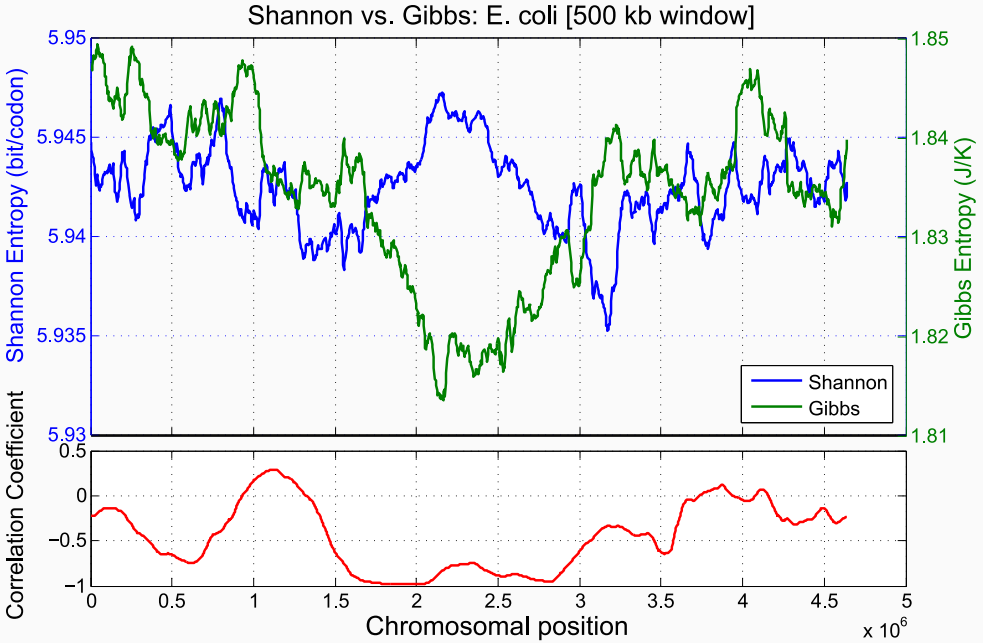
Gibbs entropy ⇒ measure of thermodynamic stability

Energies are assumed to be distributed according to the Boltzmann distribution

$$P_G(i) = \frac{n_i e^{-\frac{E(i)}{k_B T}}}{\sum_j n_j e^{-\frac{E(j)}{k_B T}}}$$

Shannon vs. Gibbs entropy in *E. coli*





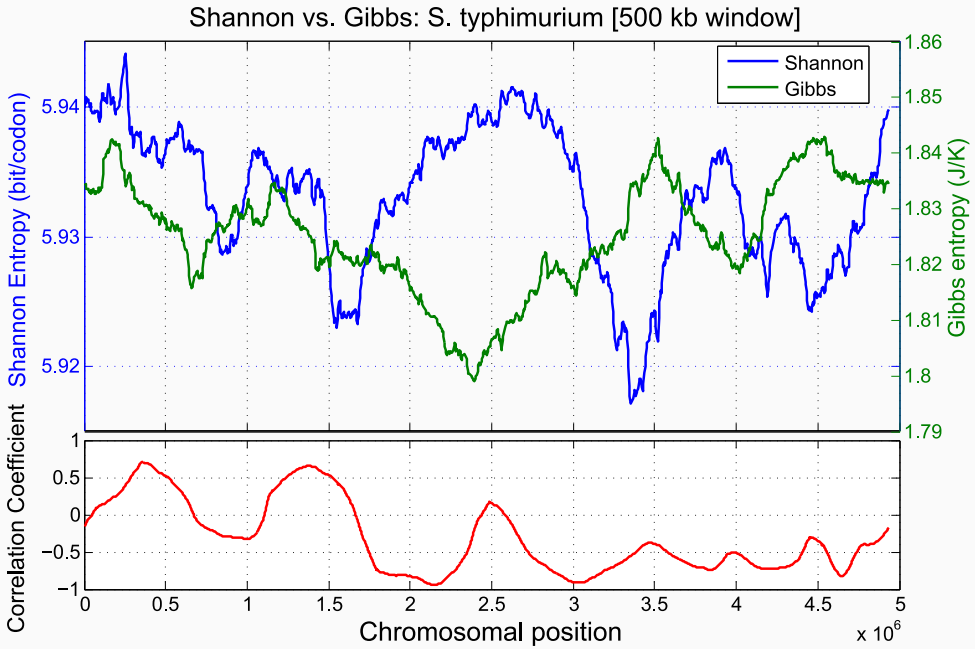
A sliding window approach

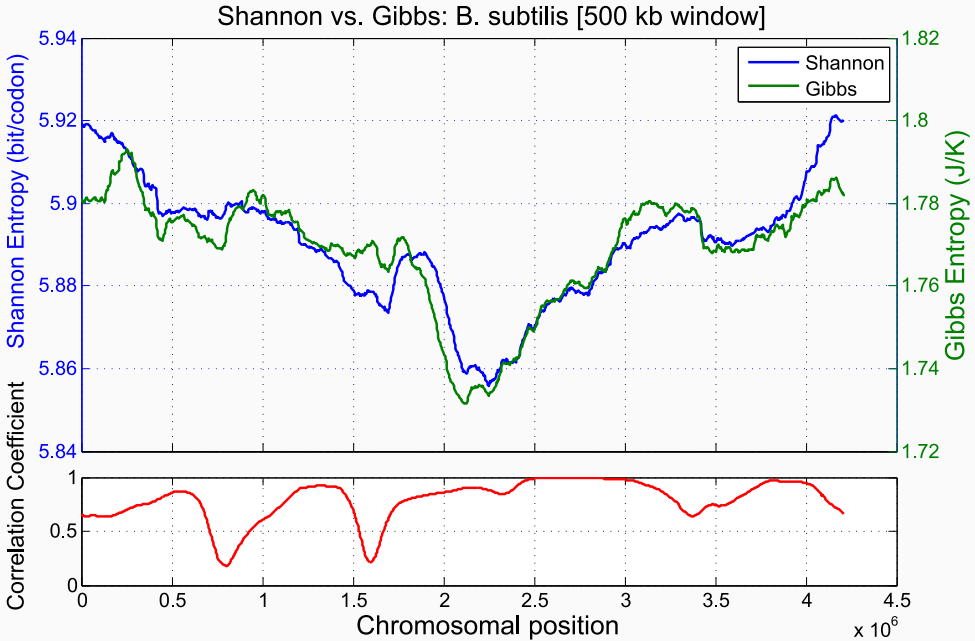
- 4 kb shifts
- block size $= 3$
- Oric $\rightarrow$ Ter $\rightarrow$ Oric

Observations

- Ter: less stable and more random
- Shannon and Gibbs entropies: mostly anti-correlated

Shannon vs. Gibbs entropy in *B. subtilis* and *S. typhimurium*





Parameters

- 4 kb shifts
- Window size = 500 kb
- Oric → Ter → Oric

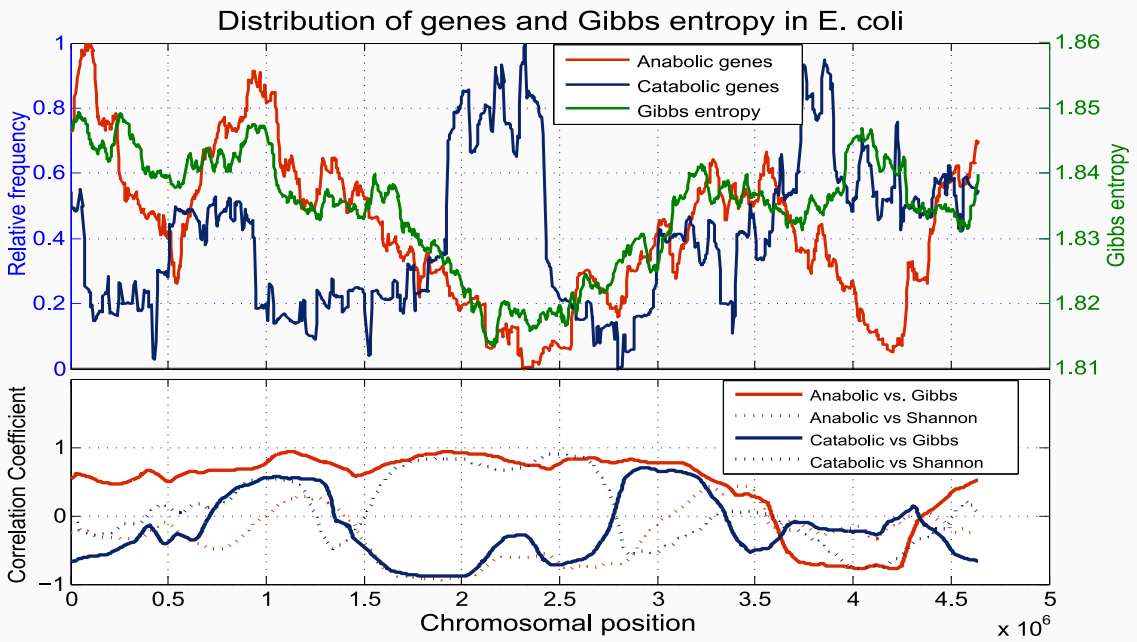
Phylogeny

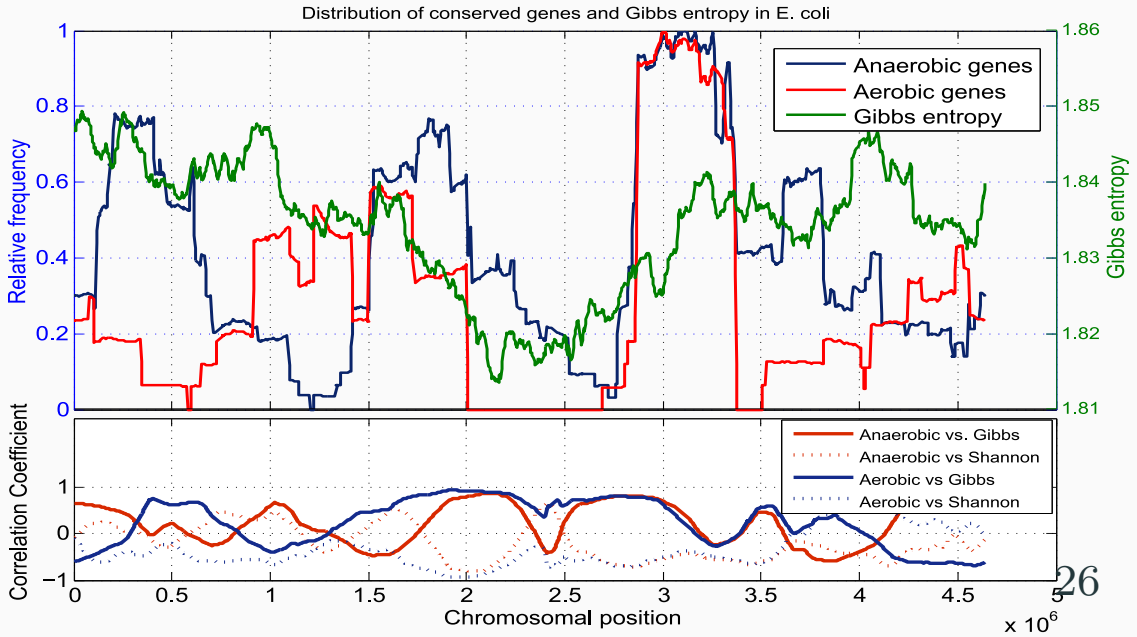
- *S. typhi* close to *E. coli*
- *B. subtilis* is more distant and gram-positive

Functional classes of genes

Functional classes of genes

- Anabolic genes: biosynthesis of macromolecules
- Catabolic genes: degradation of macromolecules
- Aerobic genes: aerobic respiration
- Anaerobic genes: anaerobic respiration





Essential and Non-essential Genes

Features

Information-Theoretic features

- Mutual Information (MI)
- Conditional Mutual Information (CMI)
- Entropy (E)
- Markov Model (M)

Non-IT features

- GC content, length, and GC3
- Close-to-stop
- Number and position of stop codons in the other ORFs

Mutual Information (MI)

- Widely used in computational biology and bioinformatics
  - Identification of coding and non-coding DNA (Grosse et al., 2000)
  - As a phylogenetic metric (**date2003discovery**)
  - Genomic signature (**bauer2008average**)
  - SNP identification (**hagenauer2004genomic**)

Mutual Information between X and Y

$$I(X,Y) = \sum_{x \in \Omega} \sum_{y \in \Omega} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

$\Omega = \{A, T, C, G\}$.

For a given gene:

- Mutual Information between consecutive bases
- Probabilities estimated from frequencies
- $P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$ as a feature

Conditional Mutual Information (CMI)

- CMI measures conditional dependency between two variables

Conditional Mutual Information is defined as

$$I(X;Y|Z) = \sum_{z\in\Sigma} P(z) \sum_{x\in\Omega}\sum_{y\in\Omega} P_k(x,y|z) \log_2 \frac{P(x,y|z)}{P(x|z)P(y|z)}$$

$$= \sum_{x\in\Omega}\sum_{y\in\Omega}\sum_{z\in\Sigma} P(x,z,y) \log_2 \frac{P(z)P(x,z,y)}{P(x,z)P(z,y)}$$

$\Omega = \{A, T, C, G\}$.

For a given gene:

- CMI between 1st (X) and 3rd (Y) positions conditioned on the 2nd (Z).
- Probabilities estimated from frequencies
- $P(x,z,y) \log_2 \frac{P(z)P(x,z,y)}{P(x,z)P(z,y)}$ as a feature

Markov Model (M)

Assumption: The gene sequence is generated by a Markov source

Order estimation → Construct the Markov chains → Score the genes

CMI based Markov order estimator (Papapetrou2013)

Markov chain of order L

$$P(x_n|x_{n-1},\dots,x_{n-L},x_{n-L-1},\dots) = P(x_n|x_{n-1},\dots x_{n-L})$$

Observation:
for any m,
If $m \leq L \rightarrow$ CMI > 0
If $m > L \rightarrow$ CMI$= 0$

Transition probabilities of the Markov chains

- Two Markov chains of order m_E and m_N
- Transition probabilities estimated

If the gene sequence is $b_1, b_2, b_3, ..., b_N$, the score is calculated as

$$Score = \sum_{i=1}^{N-\hat{L}} P(b_i b_{i+1} \dots b_{i+\hat{L}}) \log_2\left(\frac{P(b_{i+\hat{L}}|b_i b_{i+1} \dots b_{i+\hat{L}-1})}{P(b_{i+\hat{L}})}\right)$$

Classifier design and evaluation

Machine learning algorithms

- Support Vector Machine (SVM)
- Random Forest

Performance evaluation

- Area Under the ROC Curve (AUC)
- 15 bacteria, 1 archeaon, and 4 eukaryotes

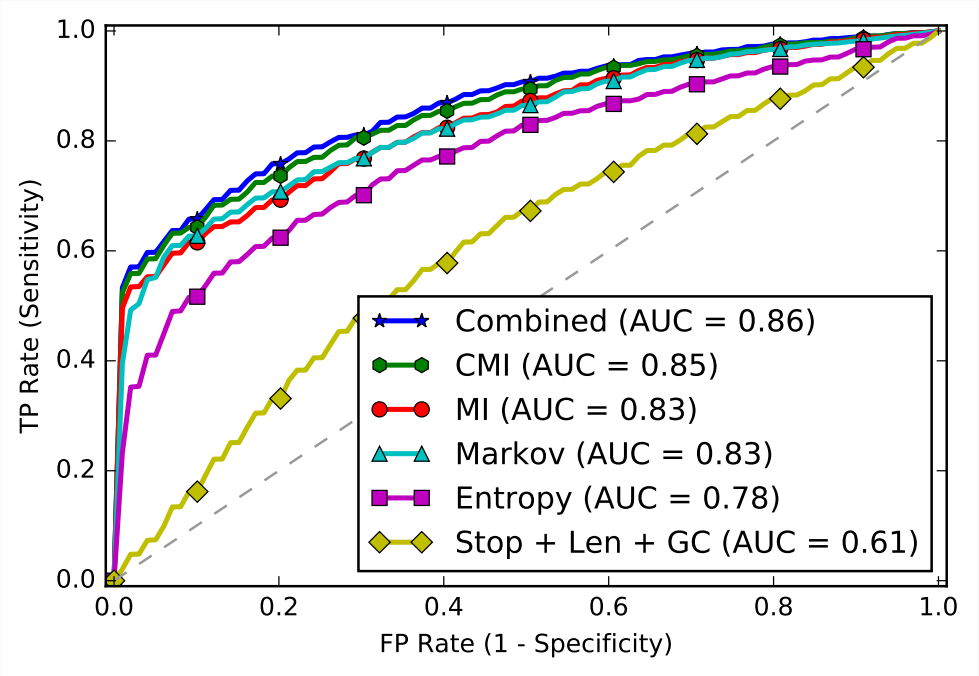
Unbalanced datasets

- $\#EGs \ll \#NEGs \rightarrow$ Undersampling

Prediction approaches

- Intra-organism prediction
  - 80 % training
  - 20 % testing
- Cross-organism prediction
  - pairwise
  - leave-one-species-out

Intra-organism predictions



E. coli

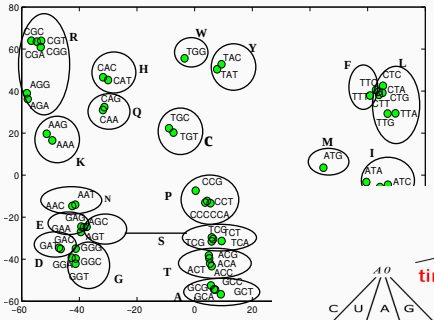
- **296 EGs** and **4077 NEGs**
- **Markov order**: 5

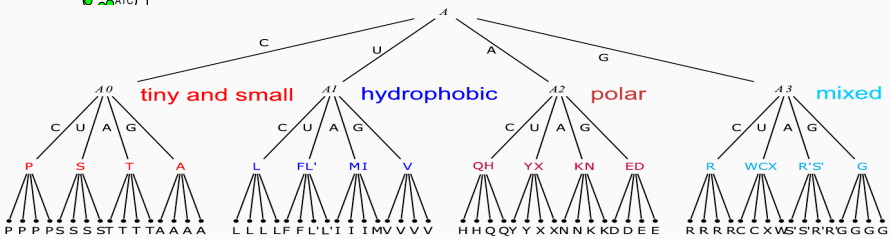
Comparisons

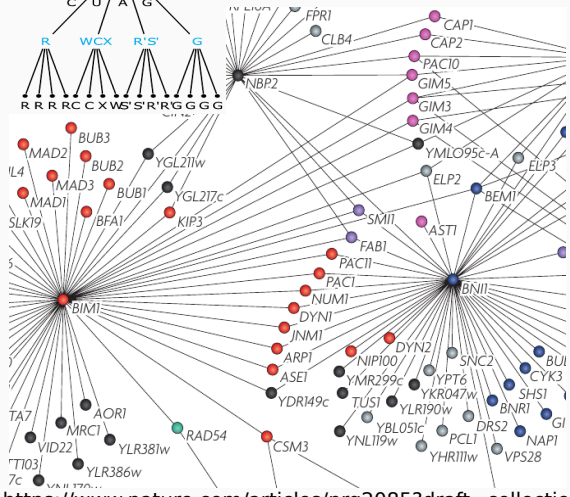
- Ning et al. (sequence composition)→ 0.82
- Li et al. (inter-nt distance)→ 0.80
- Yu et al. (fractals) → 0.75

Conclusions and Discussing Open Problems

Conclusions and open problems







https://www.nature.com/articles/nrg2085?draft=collection