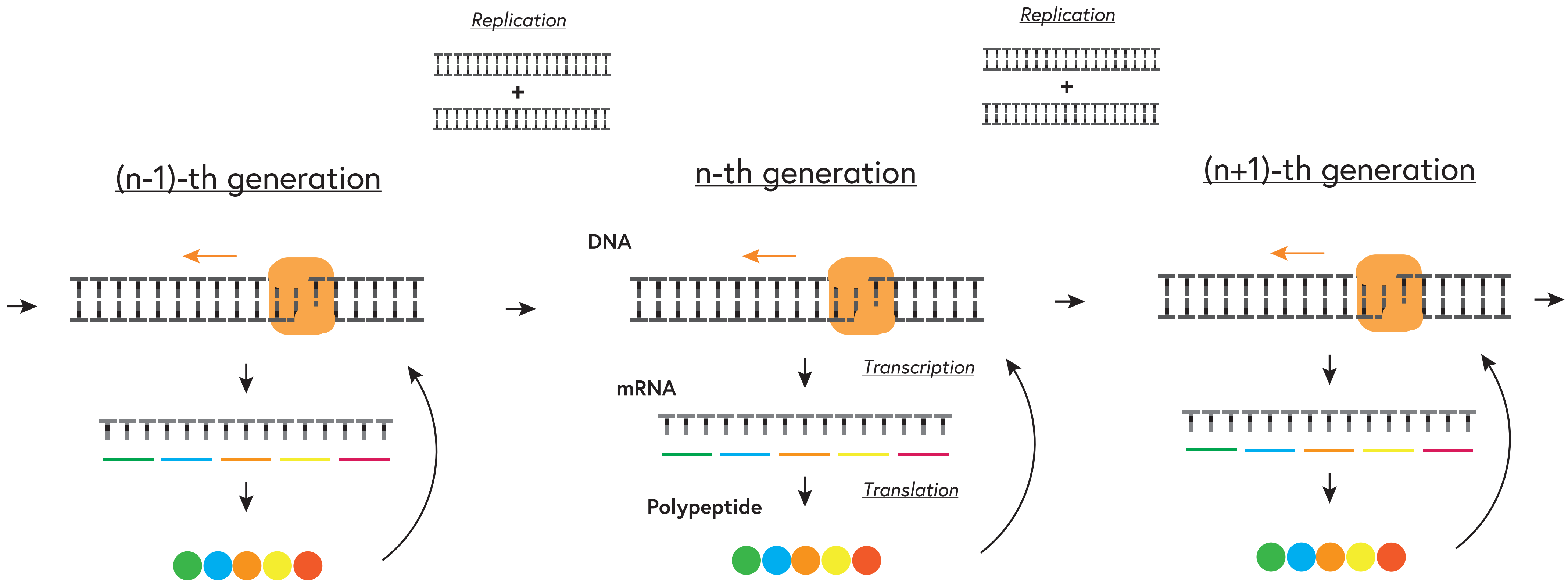
Towards An Extension of Artificial Communication System for Self-Organized System

HARUKA MIYAZAWA

Ph.D. Candidate, Osaka University
Academic Memebr, Athens Institution for Education & Research
Alumni, RIKEN Quantitative Biology Center

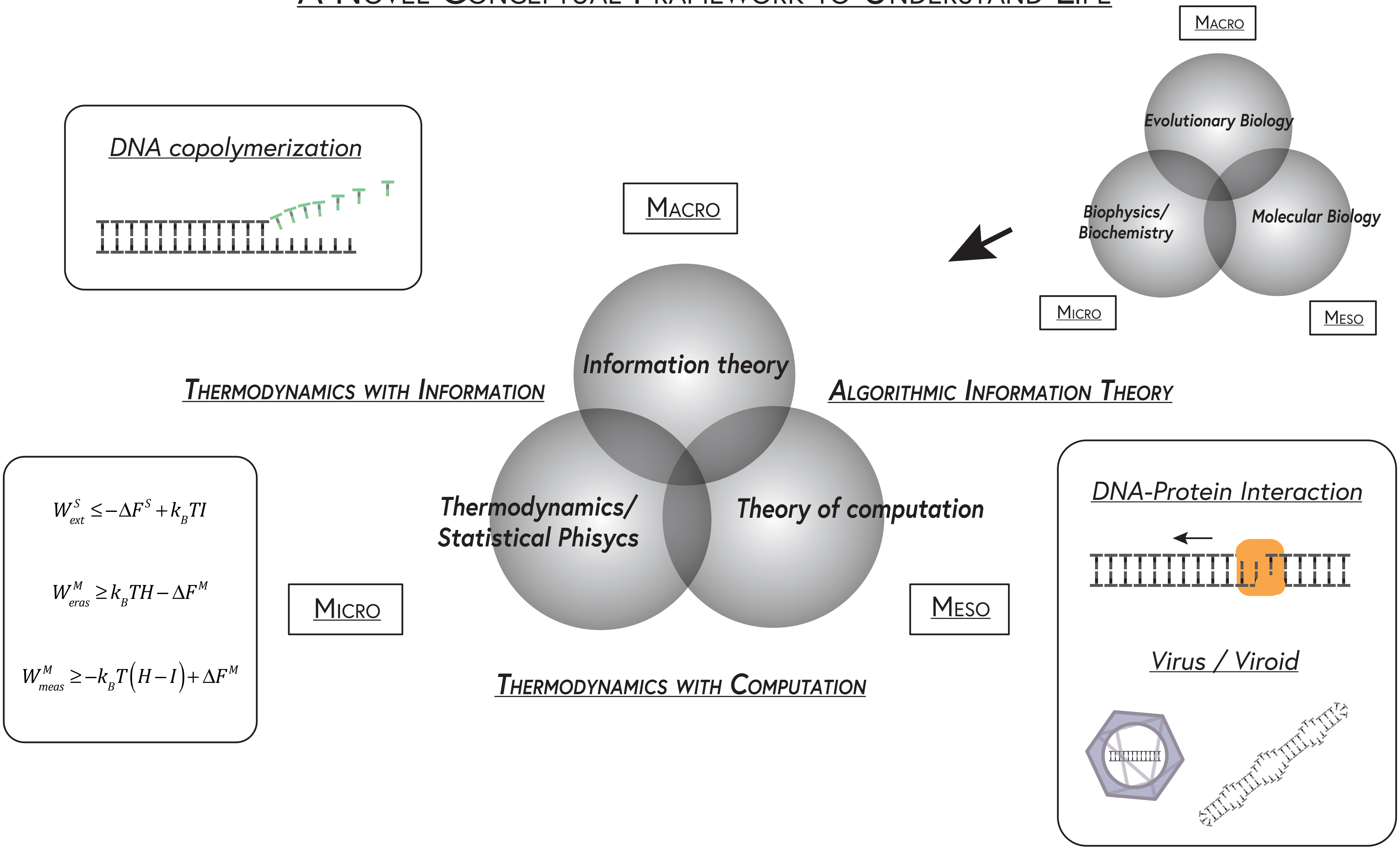
RIKEN

A Big Question to Be Solved in 21st Century



Q. How can we 'mathematically-well' extract the essence of biology beyond its physical features or constraints?

A Novel Conceptual Framework to Understand Life



Recent Applications of Information Theory to Life Science

- *"Nonequilibrium generation of information in copolymerization processes"*

  Andrieux, D. and Gaspard, P. Proc Natl Acad Sci U S A. 105: 9516 (2008).

  Introduces thermodynamical equations for copolymerization processes of nucleic acids.

- *"Robustness and Compensation of Information Transmission of Signaling Pathways"*

  Uda, S. et al. Science 341 (6145), 558-561. doi: 10.1126/science.1234511 (2013).

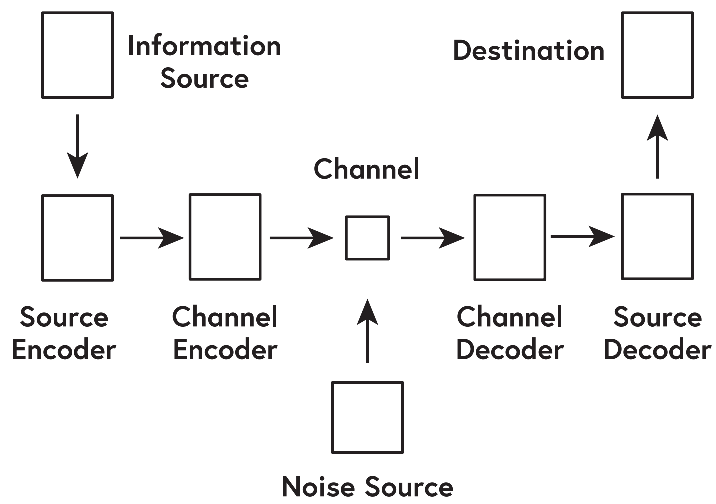
  Calculates the mutual information transmitted through signaling pathways.

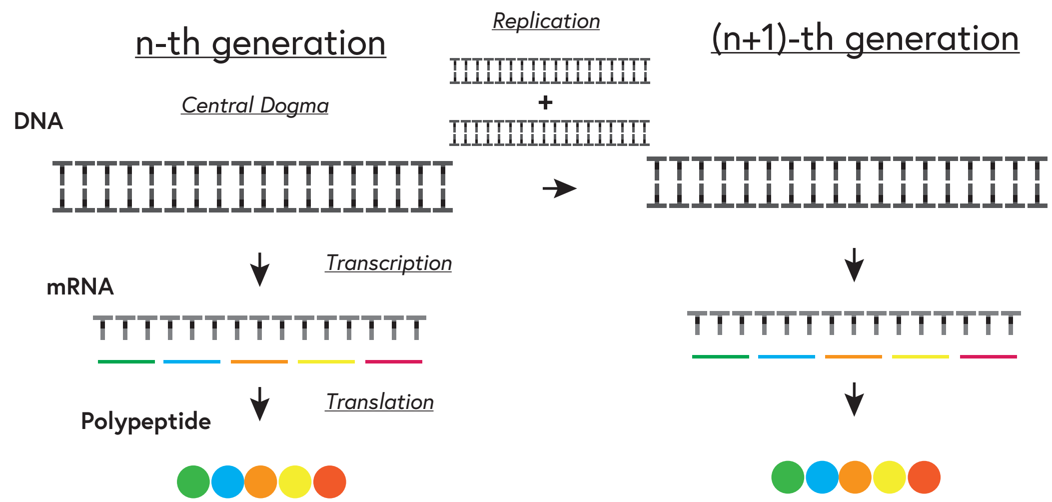
- *"Maxwell' s demon in biochemical signal transduction with feedback loop."*

  Ito, S. and Sagawa, T. Nat. Commun. 6:7498 doi: 10.1038/ncomms8498 (2015).

  Elucidates that transfer entropy gives the upper bound of robustness of signal transduction against environmental fluctuations based on the second law of thermodynamics with information.

A Comparison between Artificial Communication System and Biological System





Information Source	*DNA or RNA*	*Cellular State (Polypeptides) of n-th generation*
Encode/Encoder	*Replication/DNA polymerase*	?
Channel	*Intracellular Space*	?
Decode/Decoder	?	*Central Dogma/RNA polymerase, transfer RNA*
Destination	?	*Cellular State (Polypeptides) of (n+1)-th generation*

Artifical Communication System v.s. 'Self-Organized' Communication System

- *Artificial Communication System*

  A communication system, which can be artificially assigned with optimal coding system.

  Source alphabet and code alphabet should be predetermined.

  Shannon entropy gives a lower bound of average code length.

  ex.) Standard information theory, Multi-user information theory

- *'Self-Organized' Communication System*

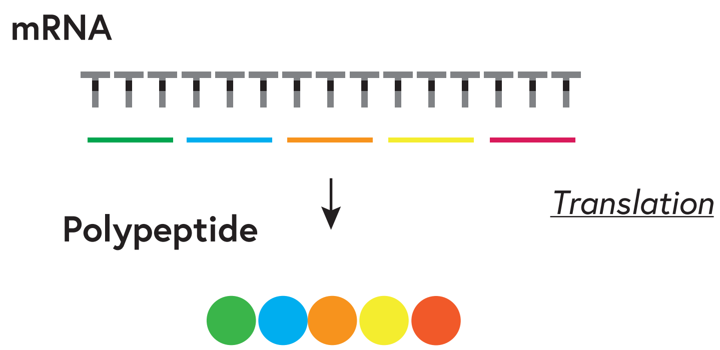
  A communication system, which seems to have spontaneously evolved its coding system.

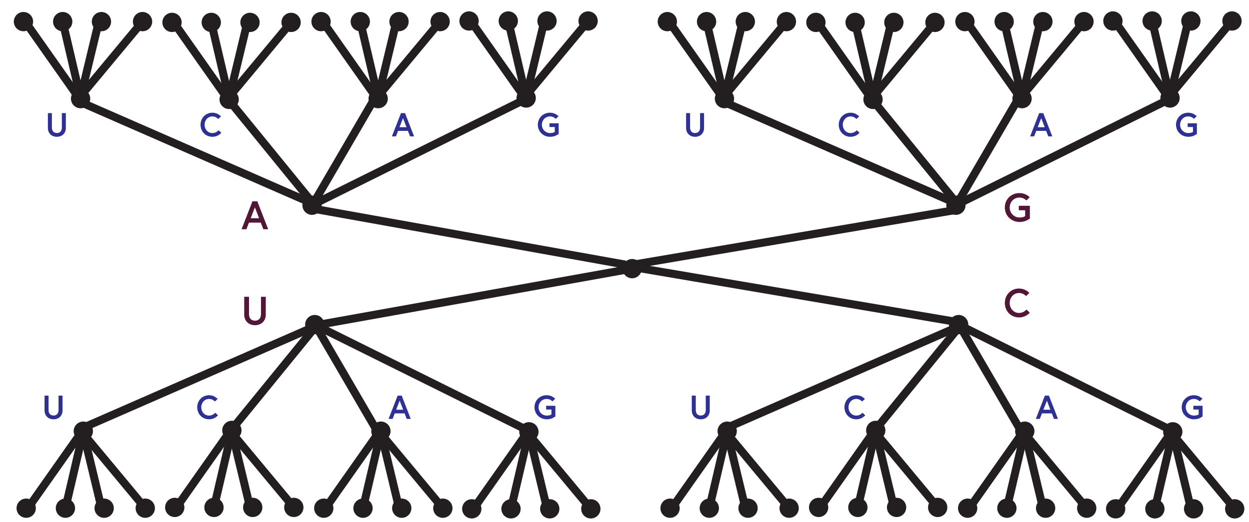
  Source alphabet / Code alphabet cannot be predetermined.

  What kind of entropy gives a lower bound of average code length?

  ex.) Biological systems (central dogma), Natural language

Genetic Code (Codon Table) as a K-ary Code Tree





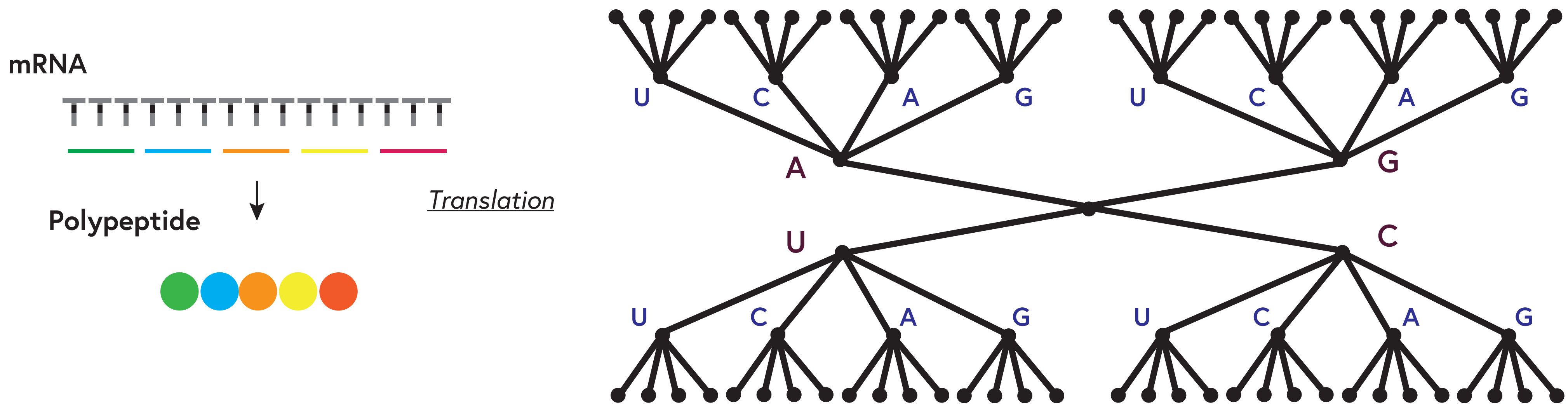
1st	2nd: U	2nd: C	2nd: A	2nd: G
U	F: UUU, UUC L: UUA, UUG	S: UCU, UCC, UCA, UCG	Y: UAU, UAC UAA, UAG	C: UGU, UGC UGA W: UGG
C	L: CUU, CUC, CUA, CUG	P: CCU, CCC, CCA, CCG	H: CAU, CAC Q: CAA, CAG	R: CGU, CGC, CGA, CGG
A	I: AUU, AUC, AUA M: AUG	T: ACU, ACC, ACA, ACG	N: AAU, AAC K: AAA, AAG	S: AGU, AGC R: AGA, AGG
G	V: GUU, GUC, GUA, GUG	A: GCU, GCC, GCA, GCG	D: GAU, GAC E: GAA, GAG	G: GGU, GGC, GGA, GGG

- *Information source: Amino Acids, Code alphabet: a,t(u),g,c*
- *Uniquely decodable code*

  All codes have same description length (Block code).
- *Instantaneously decodable code*

  Any codes are not prefix part of other codes (Prefix code).

Genetic Code (Codon Table) as a K-ary Code Tree



Information Source	*DNA or RNA*	*Cellular State (Polypeptides) of n-th generation*
Encode/Encoder	*Replication/DNA polymerase*	?
Channel	*Intracellular Space*	?
Decode/Decoder	?	*Central Dogma/RNA polymerase, transfer RNA*
Destination	?	*Cellular State (Polypeptides) of (n+1)-th generation*

One Parameter Extension of Shannon Entropy

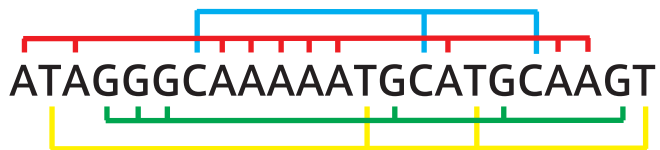
	Boltzmann-Gibbs-Shannon	*q-generalization (Tsallis)*
Statistical Independence	*Independent*	*Non-Independent*
Axioms	*Shannon Khinchin axioms* • *Shannon additivity*	*(q-)generalized Shannon Khinchin axioms* • *(q-)generalized Shannon additivity*
Entropy	$S_1(p_1,\cdots,p_n)=-k\sum_{i=1}^{n}p_i\ln p_i$	$S_q(p_1,\cdots,p_n)=\frac{1-\sum_{i=1}^{n}(p_i)^q}{\phi(q)}$
Additivity	$S_1(X,Y)=S_1(X)+S_1(Y)$	$S_q(X,Y)=S_q(X)+S_q(Y)-\phi(q)S_q(X)S_q(Y)$
exponential	$\exp(x)$	$\exp_q x:=\left[1+(1-q)x\right]^{\frac{1}{1-q}}$
logarithm	$\ln x$	$\ln_q x:=\frac{x^{1-q}-1}{1-q}$

Application of Tsallis Statistics to DNA Sequences

- *"Long-range correlations in nucleotide sequences"*

  Peng, C.-K. et al. Nature 356 (168). doi: 10.1038 (1992)

  'DNA walk' extracts the power-law features of intron-containing genes and nontranscribed regulatory DNA sequences within genome.



- *"Superstatistical model of bacterial DNA architecture."*

  Bogachev, M. I. et al. Sci. Rep. 7, 43034; doi: 10.1038/srep43034 (2017)

  Internucleotide interval distribution within genome from Archea to Homo sapience follows q-exponential distribution.

- *"Complex multifractal nature in Mycobacterium tuberculosis genome."*

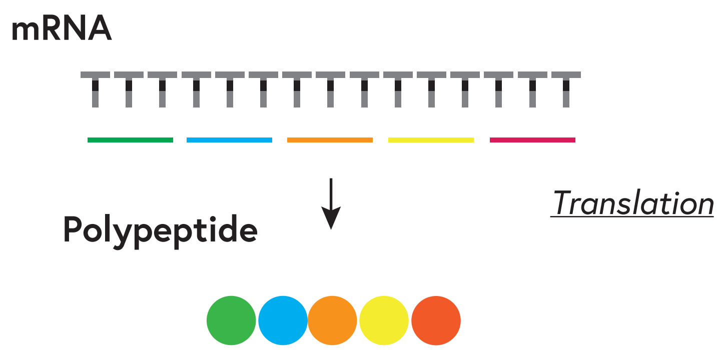
  Mandal, S. et al. Sci. Rep. 7, 46395; doi: 10.1038/srep46395 (2017).

  Multifractal analyses conducted on the highly polymorphic region of each gene in Mycobacterum tuberculosis genome.

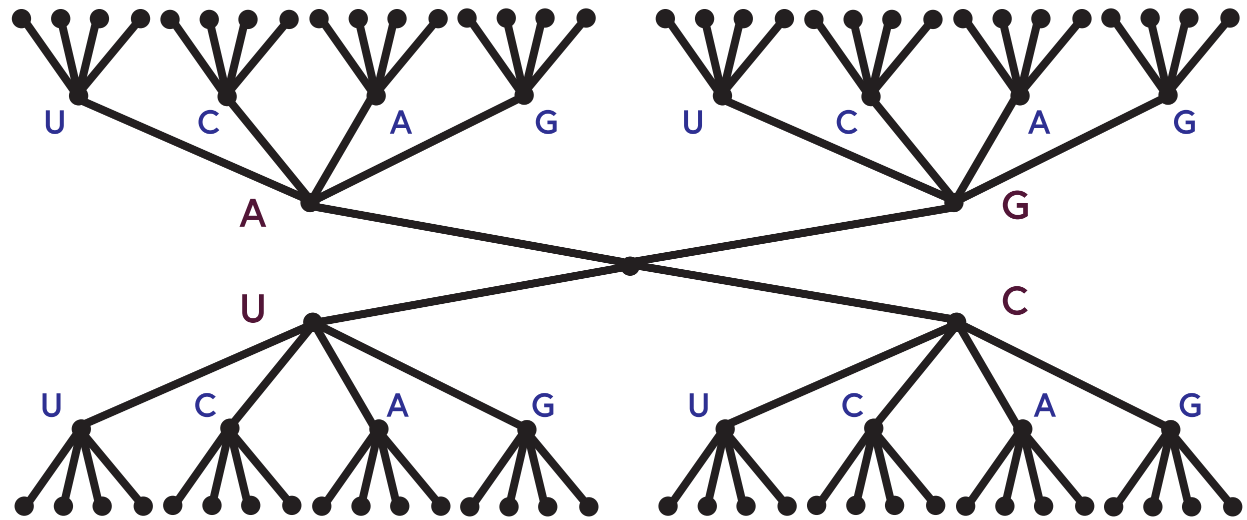
- *"Dynamic Organization of Chromatin Domains Revealed by Super-Resolution Live-Cell Imaging."*

  Nozaki, T. et al. Molecular Cell., ; doi: 10.1016/j.molcel2017.06.018 (2017)

Genetic Code (Codon Table) as a K-ary Code Tree



1st \ 2nd	U	C	A	G
U	F: UUU, UUC L: UUA, UUG	S: UCU, UCC, UCA, UCG	Y: UAU, UAC UAA, UAG	C: UGU, UGC UGA W: UGG
C	L: CUU, CUC, CUA, CUG	P: CCU, CCC, CCA, CCG	H: CAU, CAC Q: CAA, CAG	R: CGU, CGC, CGA, CGG
A	I: AUU, AUC, AUA M: AUG	T: ACU, ACC, ACA, ACG	N: AAU, AAC K: AAA, AAG	S: AGU, AGC R: AGA, AGG
G	V: GUU, GUC, GUA, GUG	A: GCU, GCC, GCA, GCG	D: GAU, GAC E: GAA, GAG	G: GGU, GGC, GGA, GGG

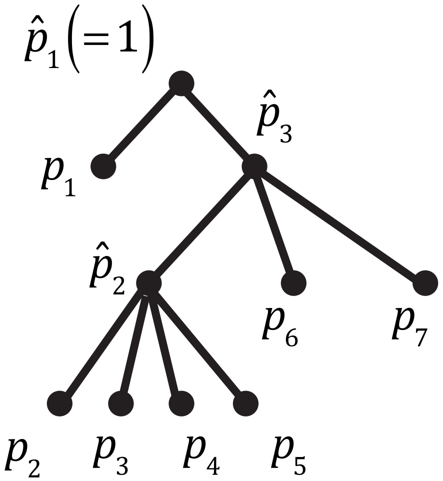


- *Information source: Amino Acids, Code alphabet: a,t(u),g,c*
- *Uniquely decodable code*

  All codes have same description length (Block code).
- *Instantaneously decodable code*

  Any codes are not prefix part of other codes (Prefix code).

Tsallis Entropy for Q-Generalized Code Tree



K-ary code tree	*q-generalized K-ary code tree*
$\ln K_j \le \ln K$	$\left(\hat{p}_j\right)^{q-1} \ln_q K_j \le \ln_q K$
$\frac{S_1(p_1,\cdots,p_N)}{\ln K} \le \sum_{i=1}^{N} p_i l_i$	$\frac{S_q(p_1,\cdots,p_N)}{\ln_q K} \le \sum_{i=1}^{N} p_i l_i$

$$N=7, D=2, q=1.5 \qquad \hat{p}_1 = 1, \hat{p}_2 = \frac{5}{16}, \hat{p}_3 = \frac{15}{32}$$

$$\left(\hat{p}_j\right)^{q-1} \ln_q K_j \le \ln_q K = \ln_q 2$$

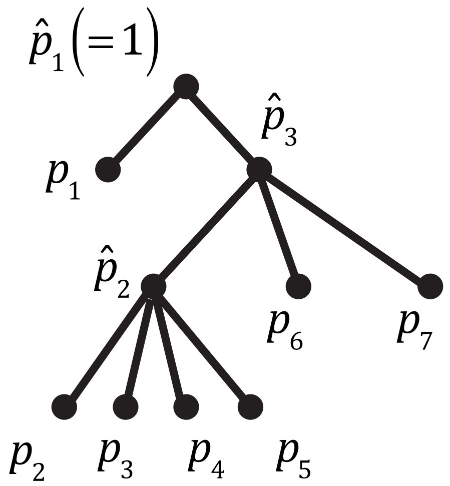
$$(p_1,p_2,p_3,p_4,p_5,p_6,p_7) = \left(\frac{17}{32},\frac{1}{32},\frac{2}{32},\frac{3}{32},\frac{4}{32},\frac{2}{32},\frac{3}{32}\right)$$

$$\frac{S_1(p_1,\cdots,p_7)}{\ln 2} = 2.16 \qquad \frac{S_q(p_1,\cdots,p_7)}{\ln_q 2} = 1.62$$

- *"Tsallis entropy as a lower bound of average description length for the q-generalized code tree."*

Suyari, H. Proceedings of 2007 IEEE International Symposium on Information Theory, pp.901-905, 2007

Tsallis Entropy for Q-Generalized Code Tree

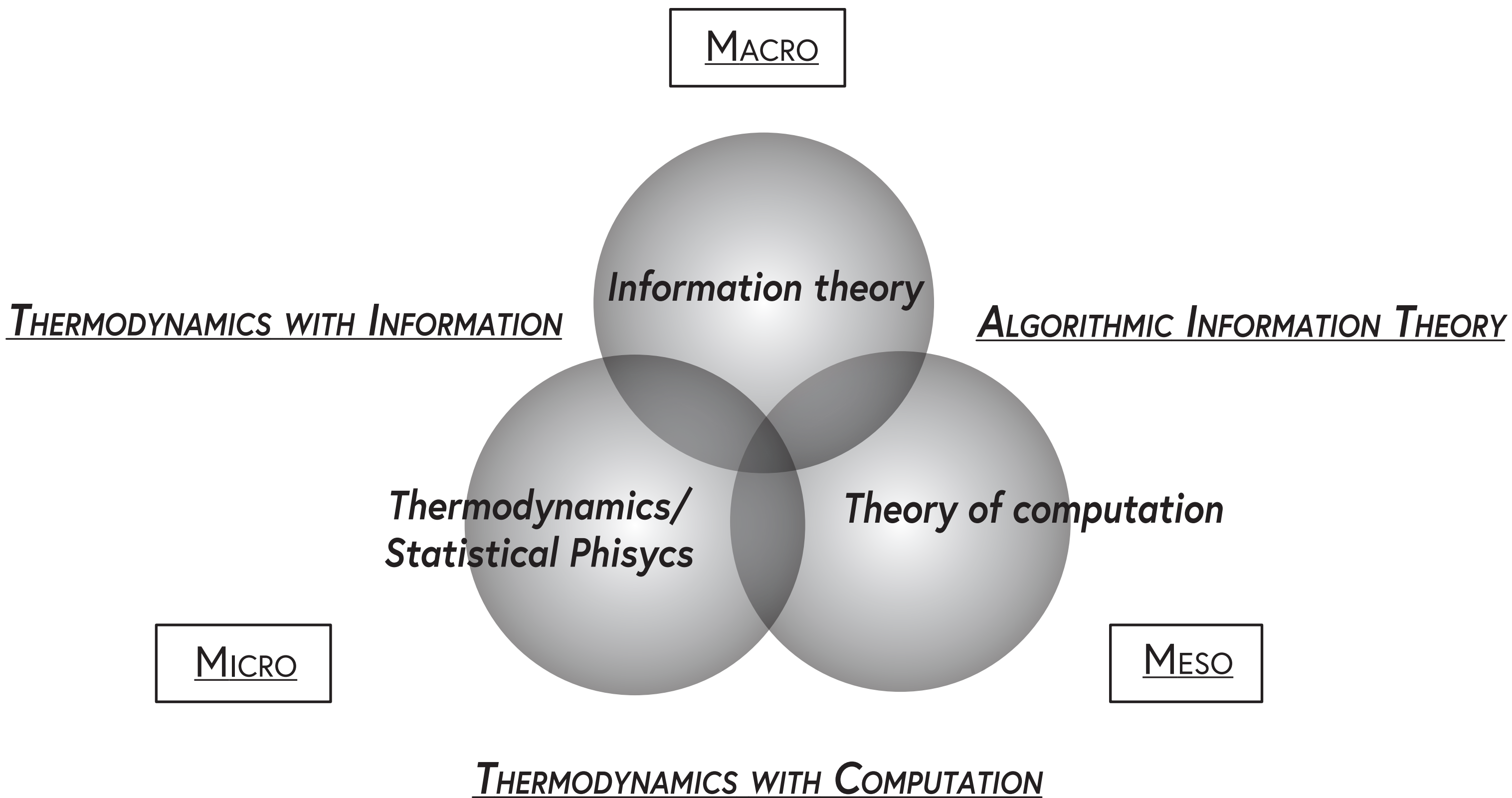


K-ary code tree	*q-generalized K-ary code tree*
$\ln K_j \leq \ln K$	$\left(\hat{p}_j\right)^{q-1} \ln_q K_j \leq \ln_q K$
$\dfrac{S_1\left(p_1,\cdots,p_N\right)}{\ln K} \leq \sum_{i=1}^{N} p_i l_i$	$\dfrac{S_q\left(p_1,\cdots,p_N\right)}{\ln_q K} \leq \sum_{i=1}^{N} p_i l_i$

- *For* $0 < q \leq 1$, $K_j \leq K$ *holds. Shannon entropy gives a lower bound of average code length.*
- *For* $q > 1$, *there is a case of* $K_j > K$, *in which Tsallis entropy gives a lower bound.*

- *The diversity of code alphabet increases.*
- *When the nuber of source alphabet is fixed, it is possible to make its average code length shorter.*
- *The feature that the nuber of code alphabet cannot be predetermined can be preferable for the case when the number of source alphabet cannot be predermined (Central Dogma, Natural Language).*

A Novel Conceptual Framework to Understand Life




September 8th, 2017
Thank you very much for your patience!
TOKYO SKY 'TREE'