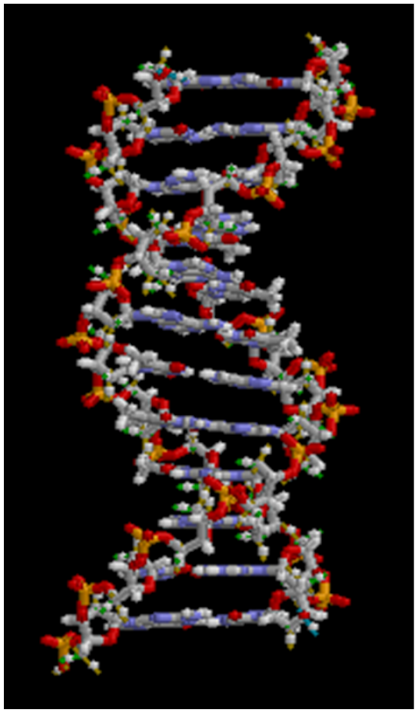
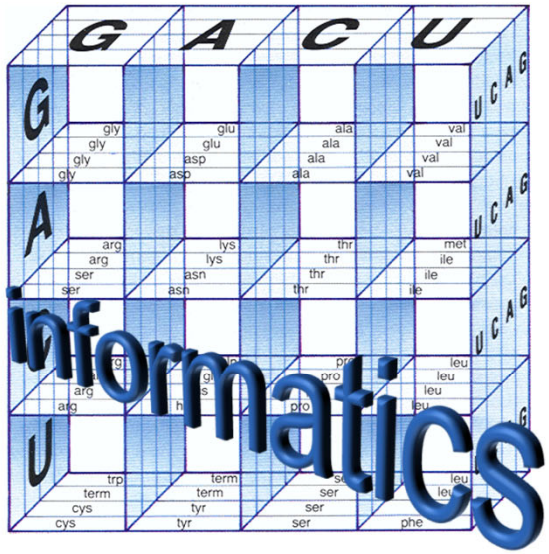
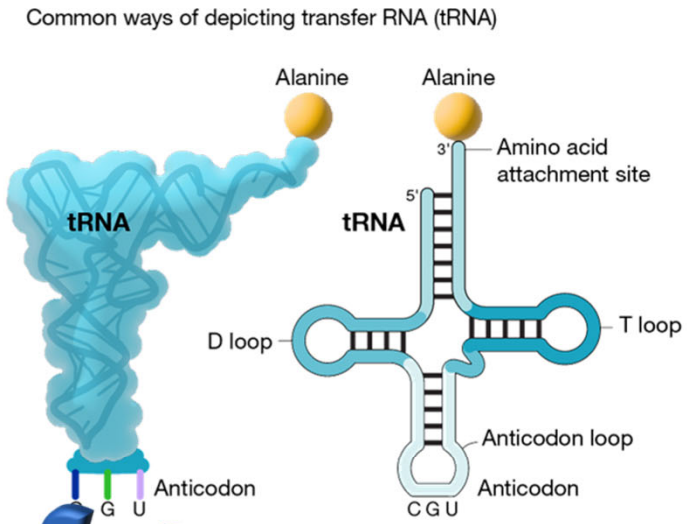
lecture 7: the informatics of DNA





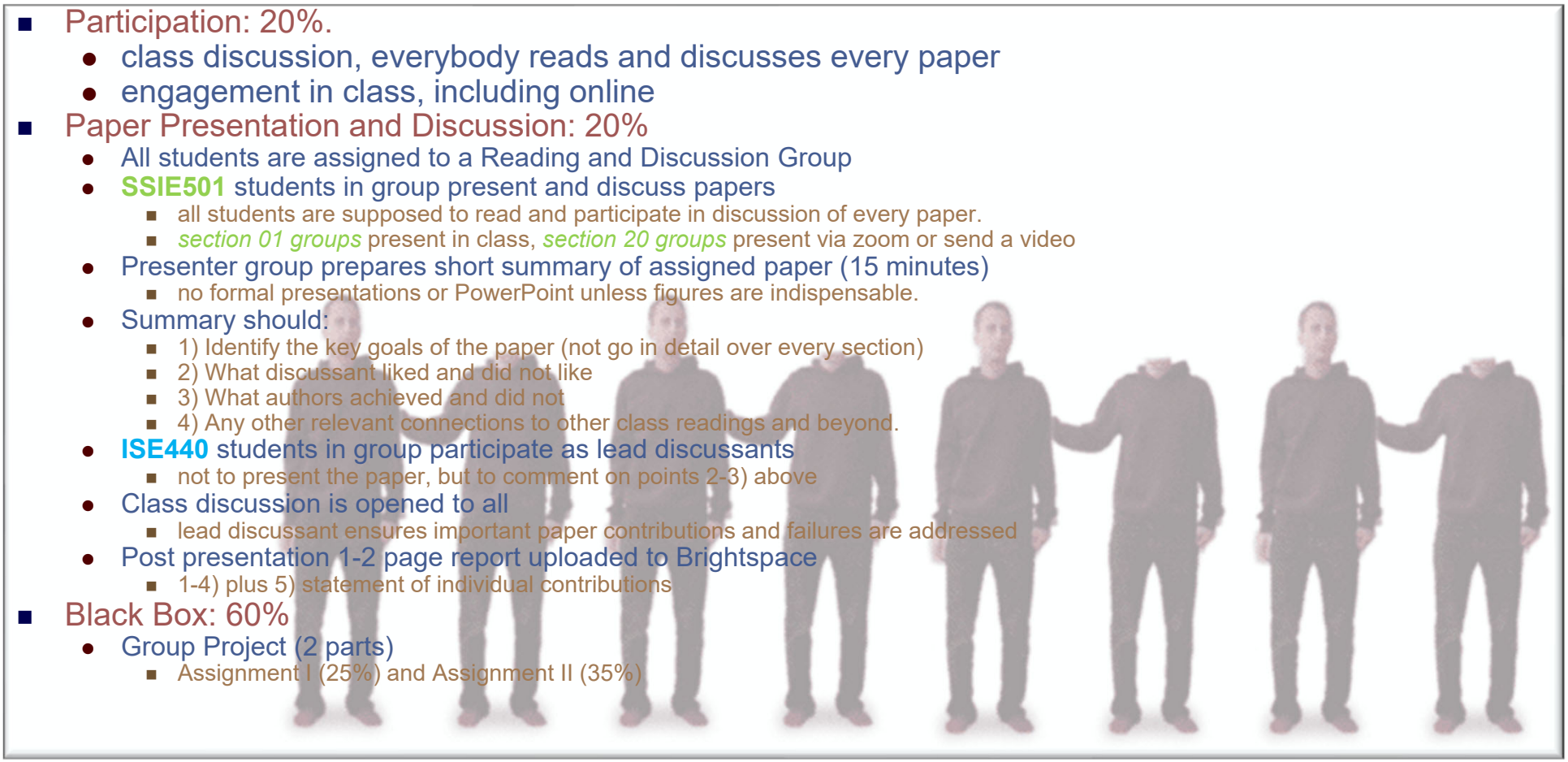


The informatics of DNA

rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

evaluation

- Participation: 20%.
  - class discussion, everybody reads and discusses every paper
  - engagement in class, including online
- Paper Presentation and Discussion: 20%
  - All students are assigned to a Reading and Discussion Group
  - **SSIE501** students in group present and discuss papers
    - all students are supposed to read and participate in discussion of every paper.
    - *section 01 groups* present in class, *section 20 groups* present via zoom or send a video
  - Presenter group prepares short summary of assigned paper (15 minutes)
    - no formal presentations or PowerPoint unless figures are indispensable.
  - Summary should:
    - 1) Identify the key goals of the paper (not go in detail over every section)
    - 2) What discussant liked and did not like
    - 3) What authors achieved and did not
    - 4) Any other relevant connections to other class readings and beyond.
  - **ISE440** students in group participate as lead discussants
    - not to present the paper, but to comment on points 2-3) above
  - Class discussion is opened to all
    - lead discussant ensures important paper contributions and failures are addressed
  - Post presentation 1-2 page report uploaded to Brightspace
    - 1-4) plus 5) statement of individual contributions
- Black Box: 60%
  - Group Project (2 parts)
    - Assignment I (25%) and Assignment II (35%)



BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

The Black Box: Due: October 6th, 2023

Herbert Simon: Law discovery means only finding **pattern** in the data; whether the pattern will continue to hold for new data that are observed subsequently will be decided in the course of **testing the law**, not discovering it. The **discovery process** runs from particular facts to general laws that are somehow induced from them; the **process of testing** discoveries runs from the laws to predictions of particular facts from them [...] To explain why the patterns we extract from observations frequently lead to correct predictions (when they do) requires us to face again the problem of **induction**, and perhaps to make some hypothesis about the uniformity of nature. But that hypothesis is neither required for, nor relevant to, the theory of discovery processes. […] By separating the question of pattern detection from the question of prediction, we can construct a **true normative theory of discovery**-a logic of discovery.

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

The Black Box: Due: October 6th, 2023

Herbert Simon: Law discovery means only finding **pattern** in the data; whether the pattern will continue to hold for new data that are observed subsequently will be decided in the course of **testing the law**, not discovering it. The **discovery process** runs from particular facts to general laws that are somehow induced from them; the **process of testing** discoveries runs from the laws to predictions of particular facts from them [...] To explain why the patterns we extract from observations frequently lead to correct predictions (when they do) requires us to face again the problem of **induction**, and perhaps to make some hypothesis about the uniformity of nature. But that hypothesis is neither required for, nor relevant to, the theory of discovery processes. […] By separating the question of pattern detection from the question of prediction, we can construct a **true normative theory of discovery**-a logic of discovery.

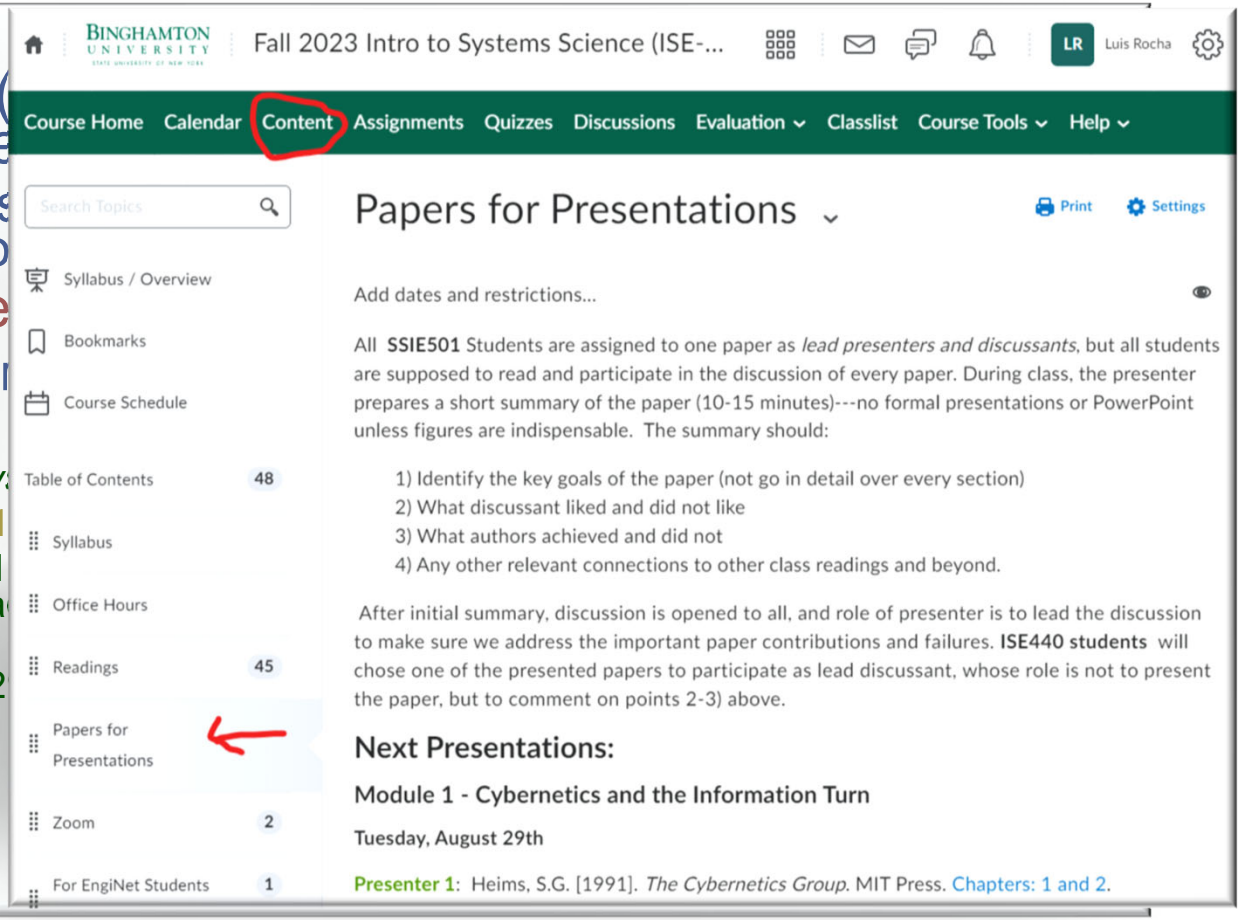
BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

next readings (check brightspace)

- Paper Presentation: 20%
  - Present (501) and lead (501&440) the discussion of an article related to the class materials
  - *section 01* presents in class, *section 20* (Enginet) posts videos on Brightspace (exceptions possible)
- Tuesday, September 19th or Thursday September 21st
  - Module 2: Systems Science
    - Reading and Discussion Group 3 (Enginet)
      - Sarah Donovan, Nicole Dates, et al:
        - Klir, G.J. [2001]. *Facets of systems Science*. Springer. Chapters 1 and 2.
          - Optional:
          - Rosen, R. [1986]. "Some comments on systems and system theory". *Int. J. of General Systems*, **13**: 1-3. Available in: Klir, G.J. [2001]. *Facets of systems Science*. Springer. pp: 241-243.
          - Wigner, E.P. [1960], "The unreasonable effectiveness of mathematics in the natural sciences". Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Comm. Pure Appl. Math*, **13**: 1-14.
        - Klir, G.J. [2001]. *Facets of systems Science*. Springer. Chapter 3.
- Future Modules
  - See brightspace

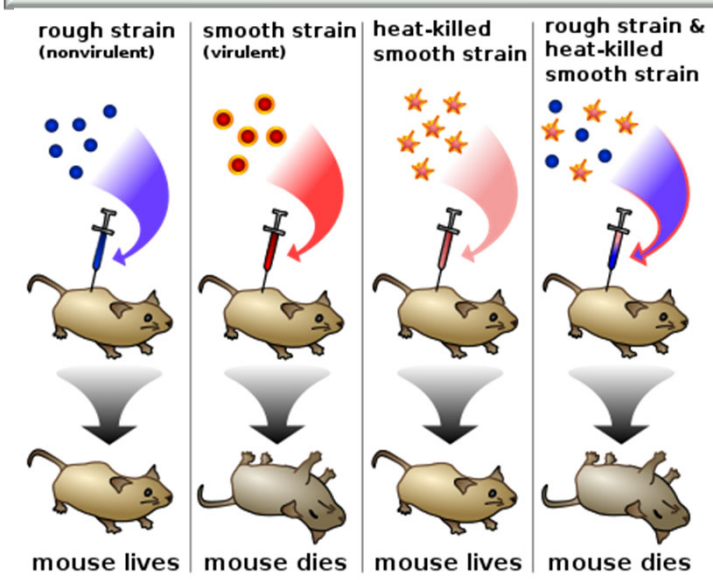
more upcoming readings (check brightspace)

- Paper Presentation: 20%
  - Present (501) and lead (related to the class mate
  - *section 01* presents in class Brightspace (exceptions po
- Module 2: Systems Science
  - Reading and Discussion Gr
    - Emma Bachyrycz, et al:
      - Klir, G.J. [2001]. *Facets of sy*
        - Optional: Klir, G.J. [2001
      - Schuster, P. (2016). The end efficiency of computational fa
      - Von Foerster, H., P. M. Mora 2026." Science **132**(3436):12
- Future Modules
  - See brightspace



identifying the loci of genetic information

- Frederick Griffith's experiment
  - In 1928: Identified a "transforming principle"
- Avery's experiment
  - Oswald Avery, Colin MacLeod, and Maclyn McCarty
  - 1944: DNA as the loci of "transformation"
    - Chemically knocking off various cellular constituents until trying DNA
    - Considerable resistance in the community accepting this result until the early 1950's (Schrodinger, Delbruck, phage group)



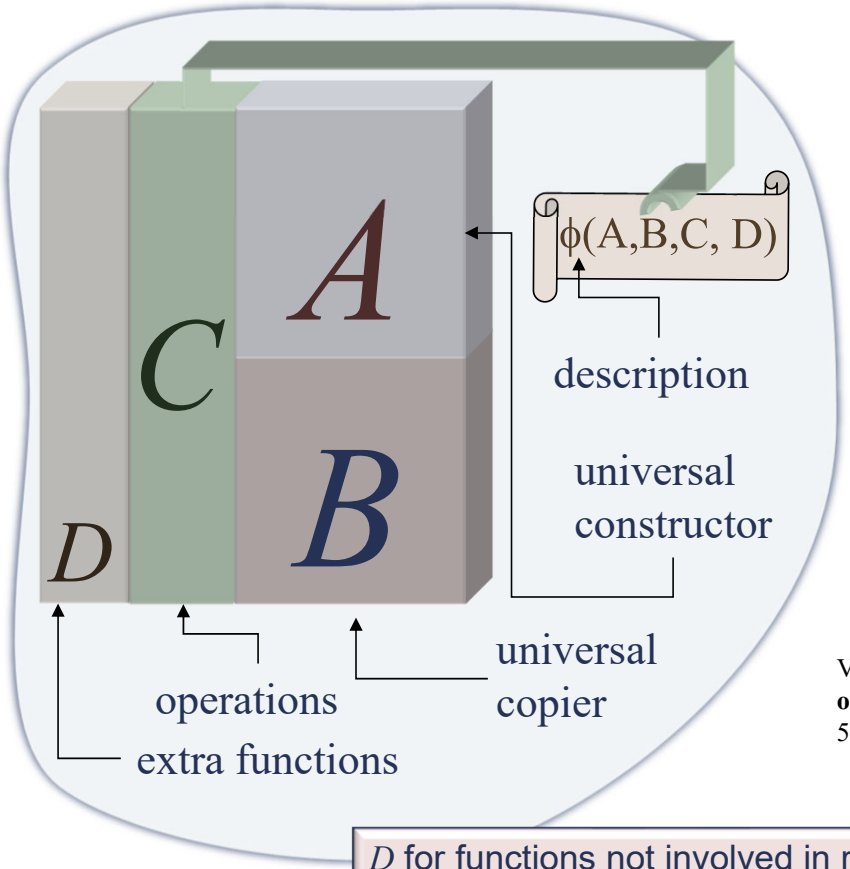
2 different strains of pneumococcus bacteria

initially not well accepted (No auto-catalysis with DNA)

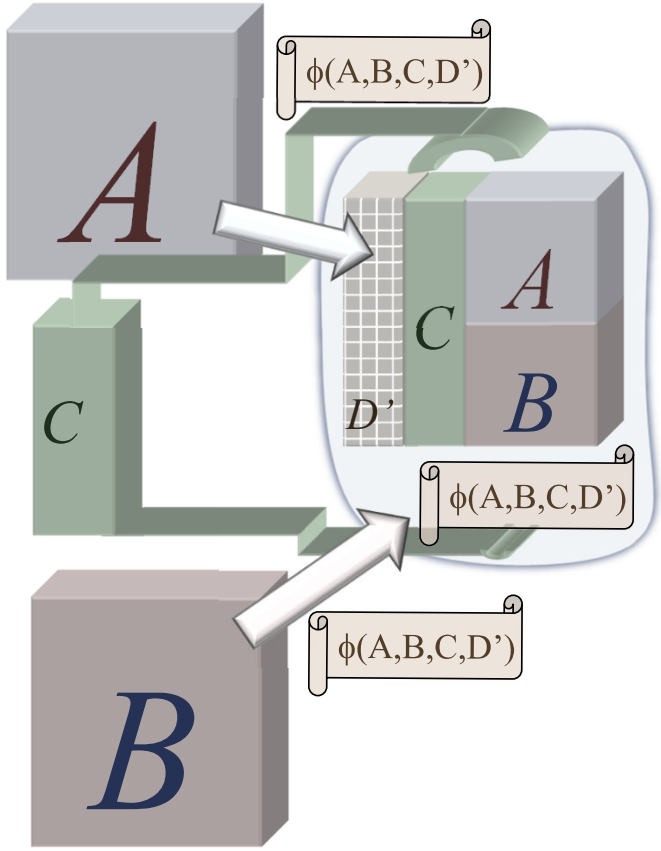
BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

Von Neumann's generalization of Turing's tape

as a general principle (system) of evolution or **open-ended complexity**



Von Neumann, J. [1949]. "**Theory and organization of complicated automata**." 5 lectures at University of Illinois



D for functions not involved in reproduction
Mutations in *D* can be propagated *vertically*
Leads to **open-ended evolution**

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

Schrodinger vs. Von Neumann

self-replication vs. decoupled, encoded information

Von Neumann, J. [1949]. "**Theory and organization of complicated automata**." 5 lectures at University of Illinois

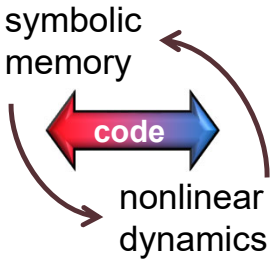
Brenner, Sydney. [2012]. "Life's code script." *Nature* **482** (7386): 461-461.

"Turing invented the stored-program computer, and von Neumann showed that the description is separate from the universal constructor. This is not trivial. Physicist Erwin Schrödinger confused the program and the constructor in his 1944 book *What is Life?,* in which he saw chromosomes as "*architect's plan and builder's craft in one*". This is wrong. The code script contains only a description of the executive function, not the function itself." (Sydney Brenner)

two roles of information
data/program (Turing)
passive/active (Von Neumann)
description/construction-function (Pattee)
genotype/phenotype (Biology)

semiotic closure (semiotic coupling)

fundamental principle of *organized complexity*
Leads to **open-ended evolution**
General principle that includes *Natural Selection*
Von Neumann described this scheme **before** structure of DNA molecule was identified in 1953 by Watson & Crick



Howard Pattee

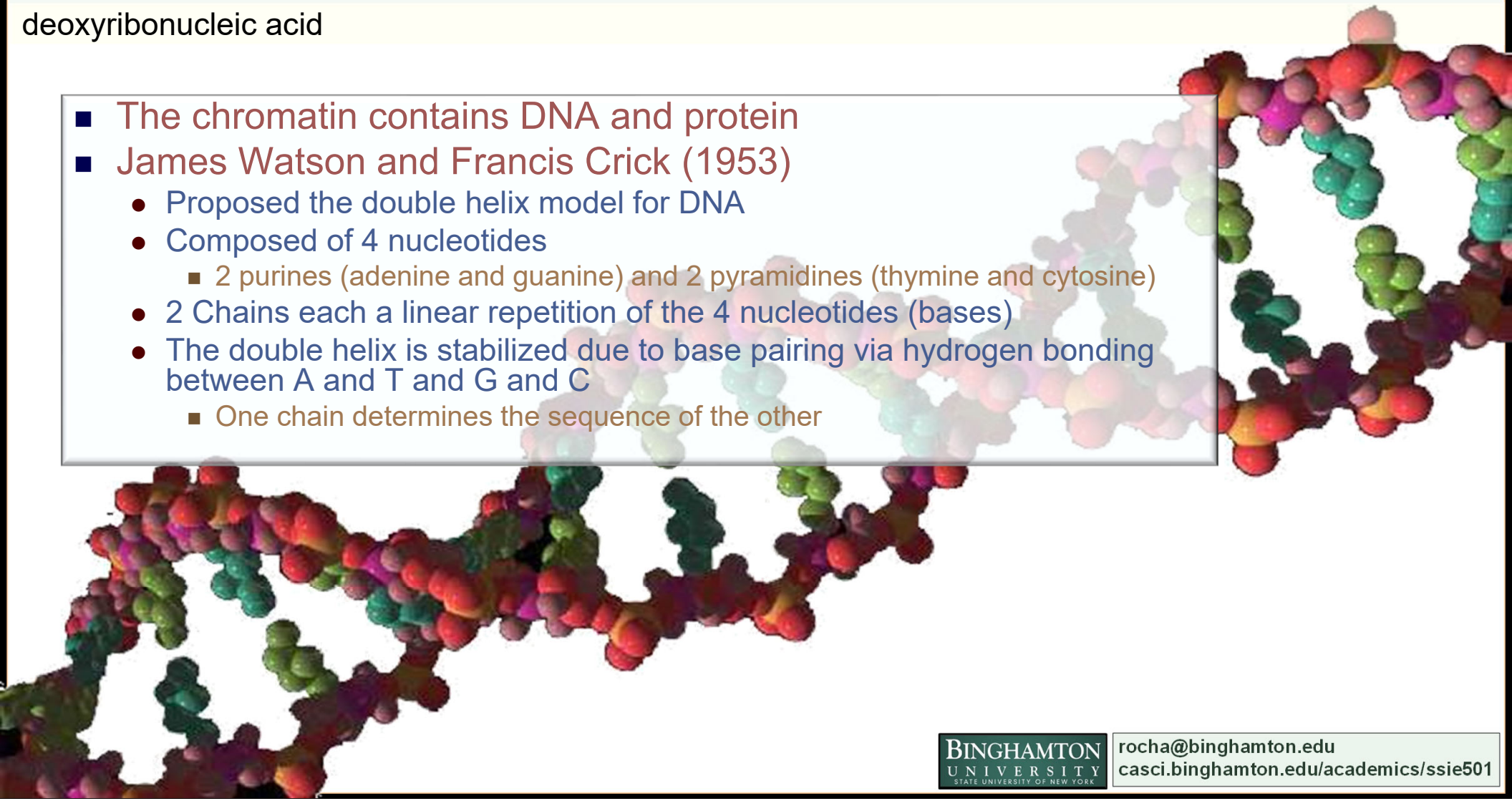
Pattee, HH [2001] *Biosystems* **60** (1):5-21

Rocha, L.M. & W. Hordijk [2005] *Artificial Life* **11**:189 - 214.
Rocha, L.M. [2001] *Biosystems* **60**: 95-121.
Rocha, L.M. [1996] *Systems Research* **13**: 371-384.

BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

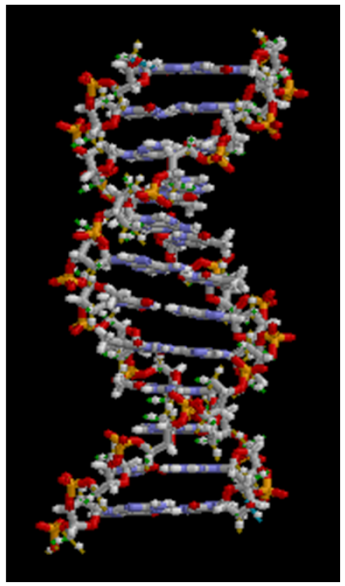
deoxyribonucleic acid

- The chromatin contains DNA and protein
- James Watson and Francis Crick (1953)
  - Proposed the double helix model for DNA
  - Composed of 4 nucleotides
    - 2 purines (adenine and guanine) and 2 pyramidines (thymine and cytosine)
  - 2 Chains each a linear repetition of the 4 nucleotides (bases)
  - The double helix is stabilized due to base pairing via hydrogen bonding between A and T and G and C
    - One chain determines the sequence of the other



BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

a molecular language system: nucleotide “bases” (the genotype “tape”)



Purine (R) → Adenine (A), Guanine (G)

Nucleotides

Pyrimidine (Y) → Cytosine (C), Thymine (T), Uracil (U)

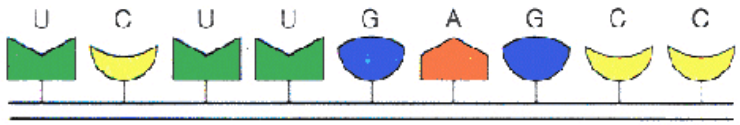
4 Letter Alphabet
DNA: A, G, C, T
RNA: A, G, C, U

Form sequences that can store information

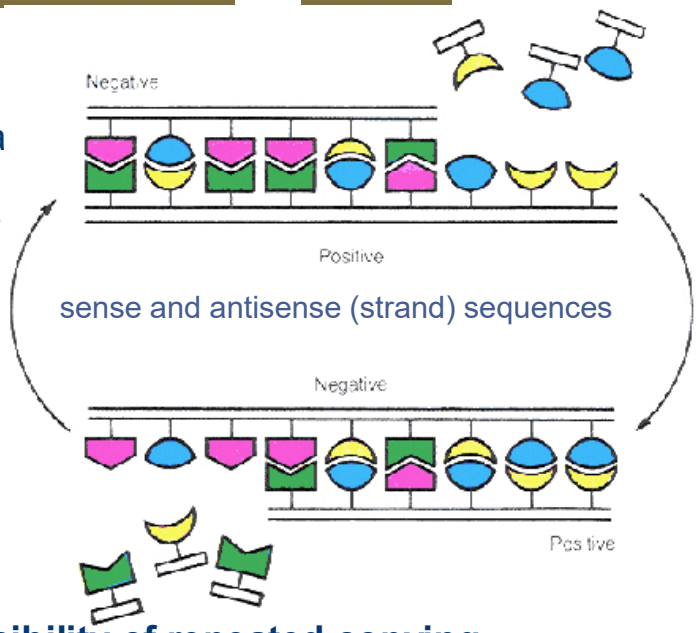
Linear molecules with a phosphate-sugar backbone (deoxyribose and ribose)

Complementary base pairing
(Hydrogen-bonding between purines and pyrimidines)

A-T (U) G-C



Figures from Eigen [1992] . *Steps Towards Life*.

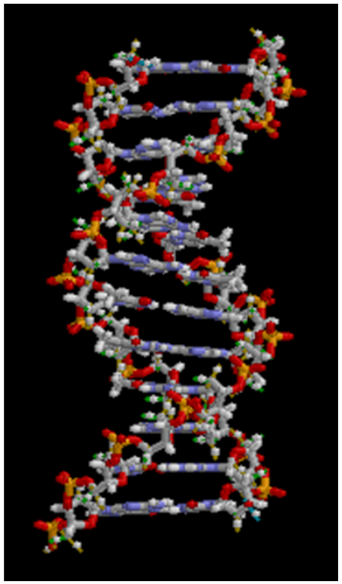


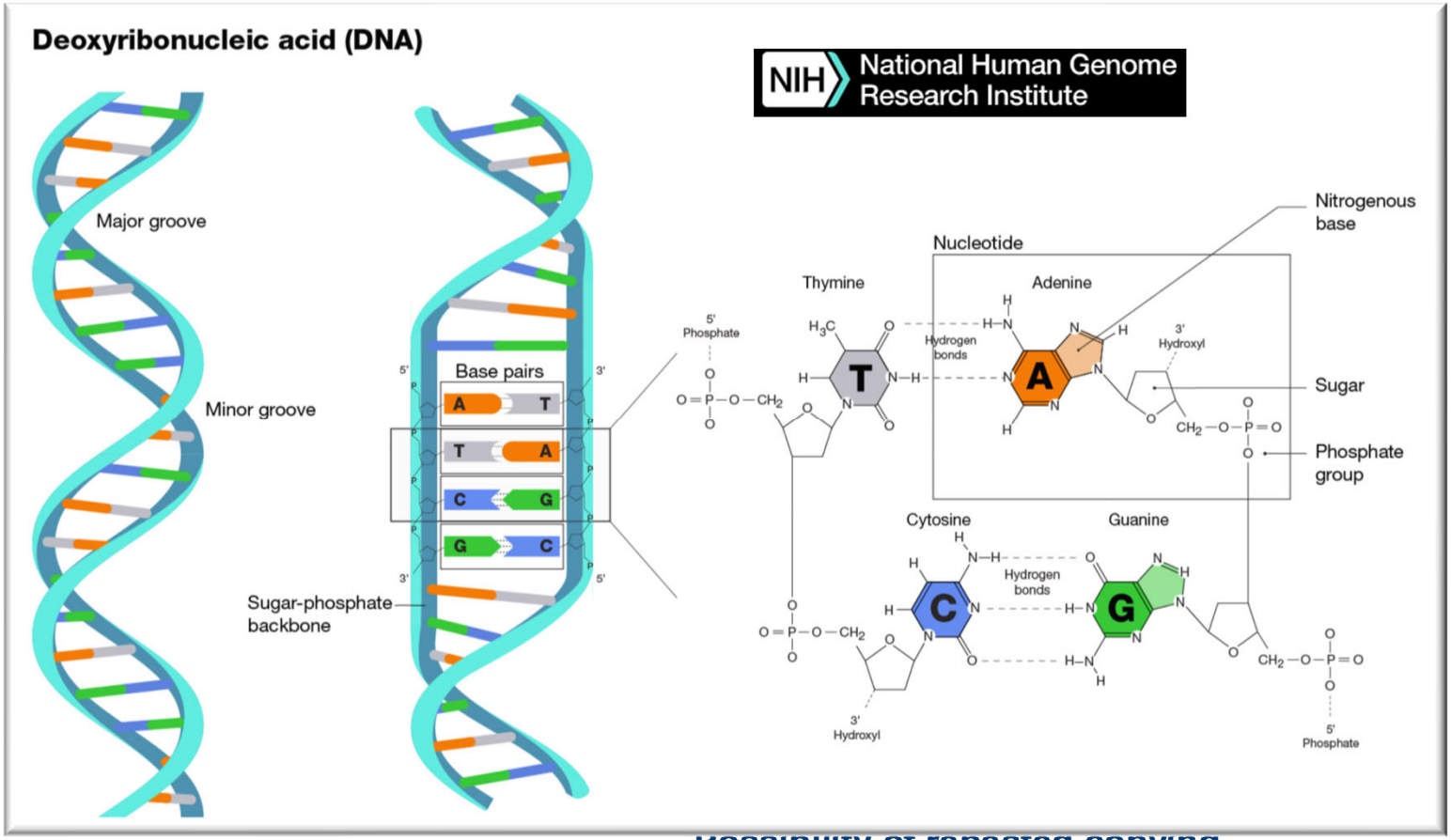
Requirements for structural information

Possibility of repeated copying

BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

a molecular language system: nucleotide “bases” (the genotype “tape”)

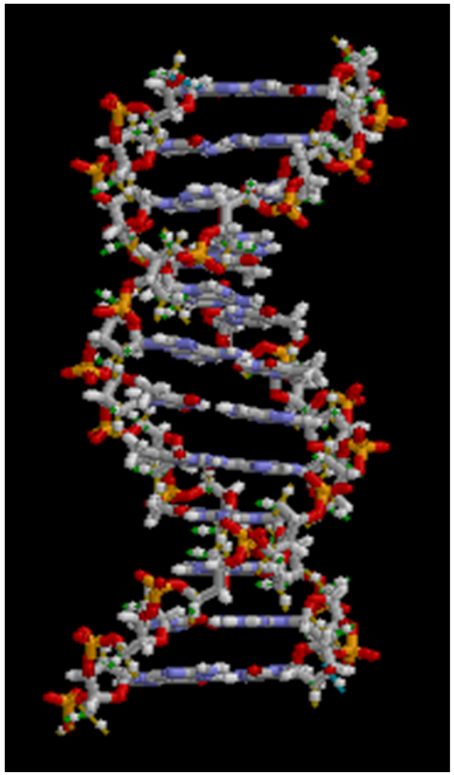


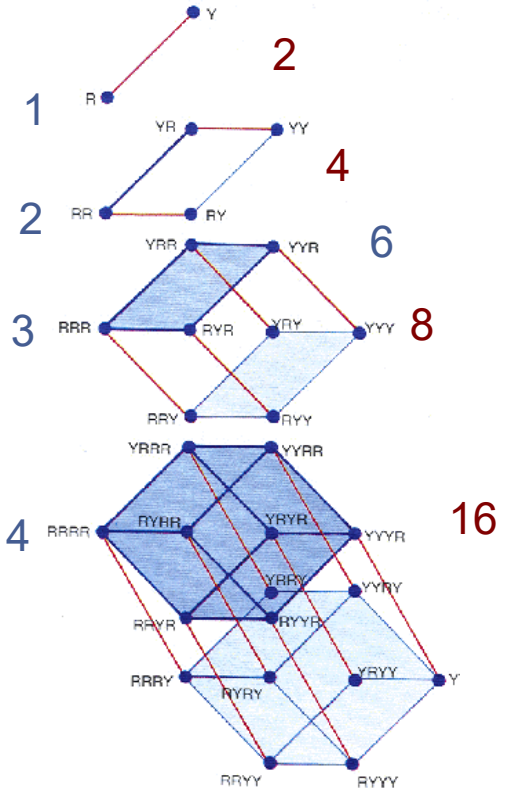


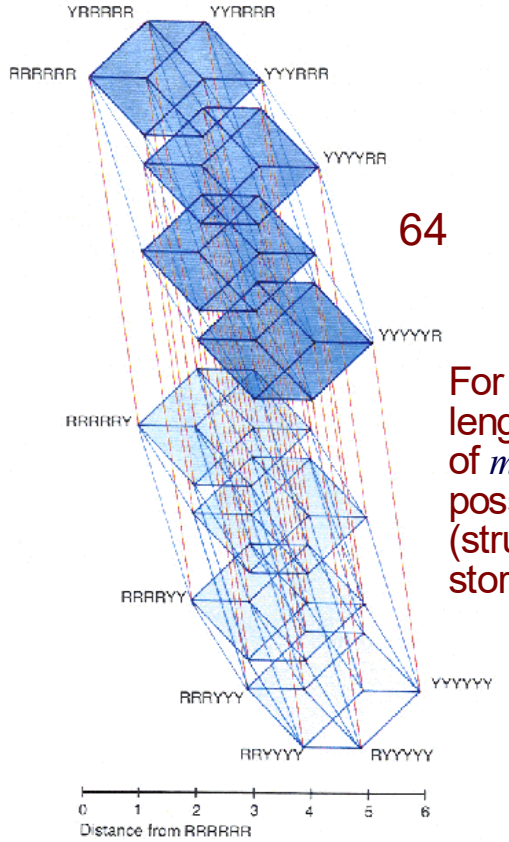
Possibility of repeated copying

BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

the genotype “tape” encodes an enormous amount of information







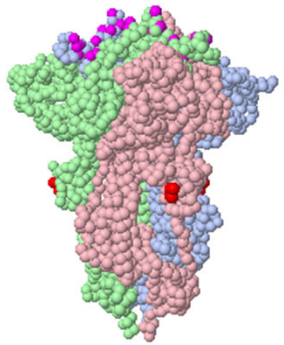
64

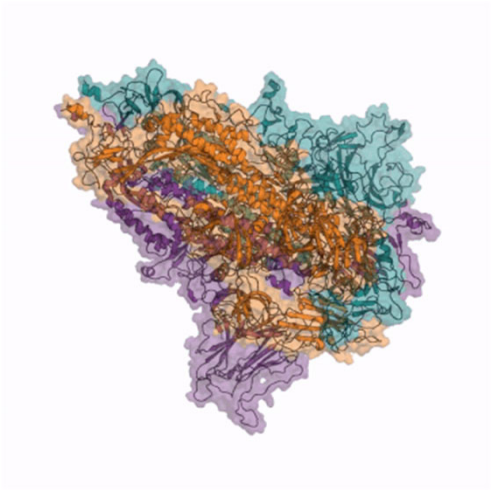
For a sequence of length n, composed of m-ary symbols, m^n possible values (structures) can be stored

Figures from Eigen [1992] . *Steps Towards Life*.

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

functional products that build up (self-organize) the phenotype





Polypeptide chains of aminoacids

Primary Structure

Folding

3-dimensional structure

Secondary and tertiary bonds

- In proteins, it is the 3-dimensional structure that dictates function
  - The specificity of enzymes to recognize and react on substrates
- The functioning of the cell is mostly performed by proteins
  - Though there are also ribozymes

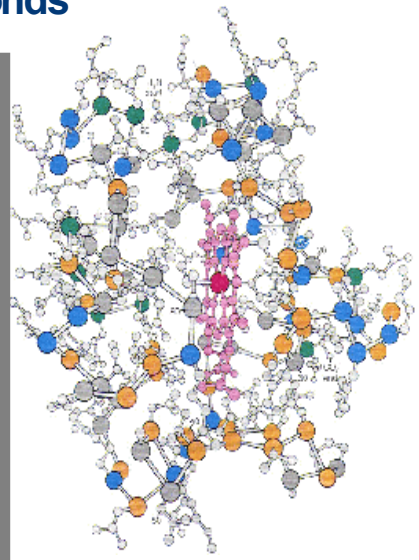


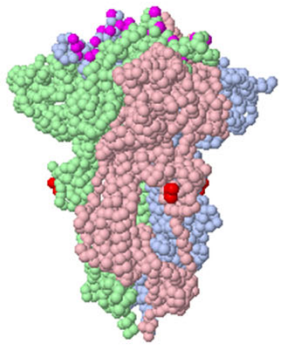
Table 1.4. Amino acid codes

Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Asn or Asp
Glx	Z	Gln or Glu
Sec	U	Selenocysteine
Unk	X	Unknown

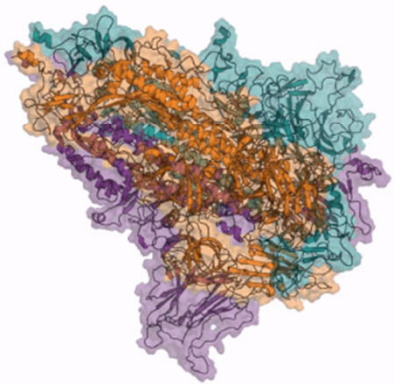
Figures from Eigen [1992] . *Steps Towards Life*.

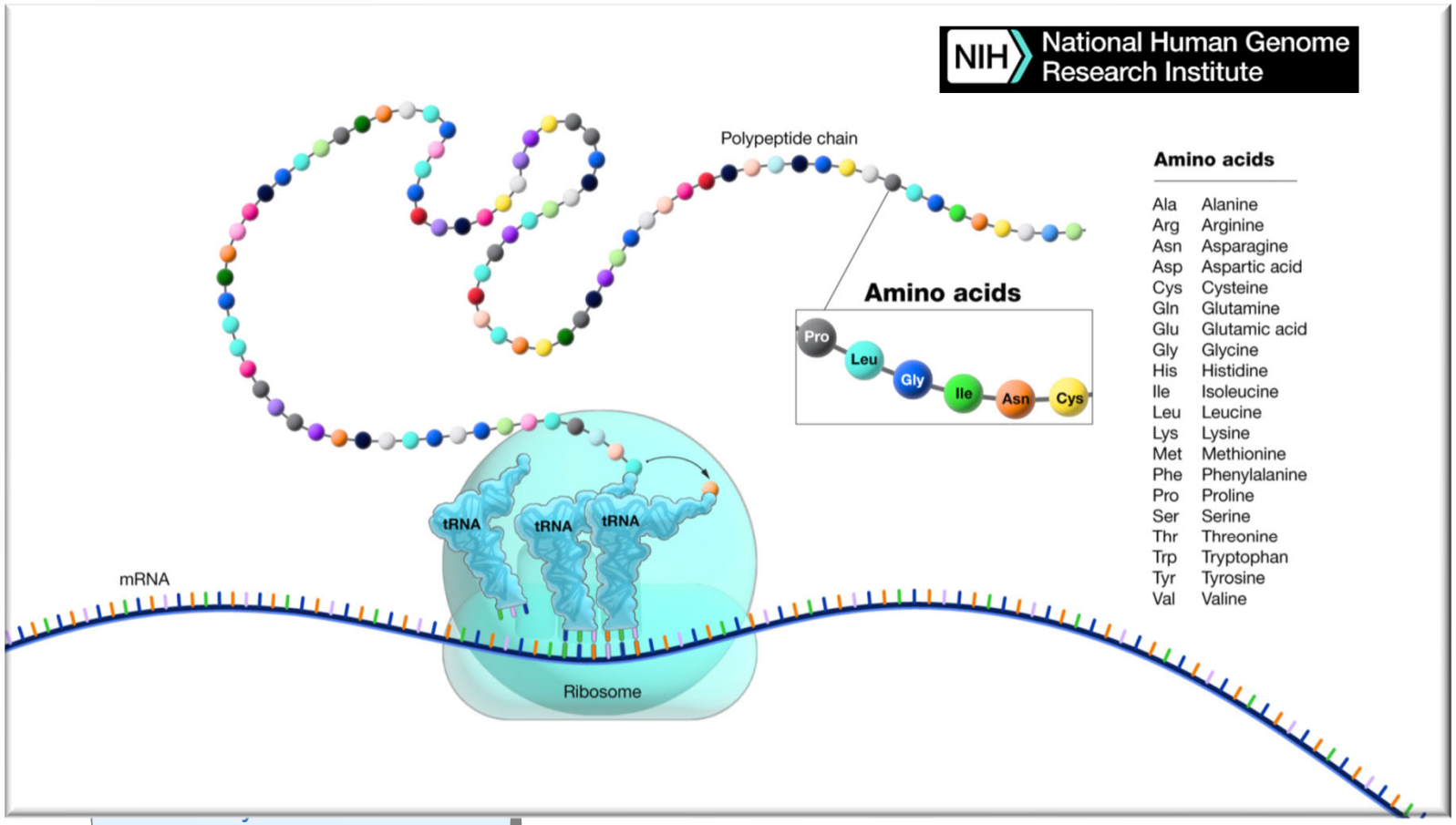
BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

functional products that build up (self-organize) the phenotype



Jmol

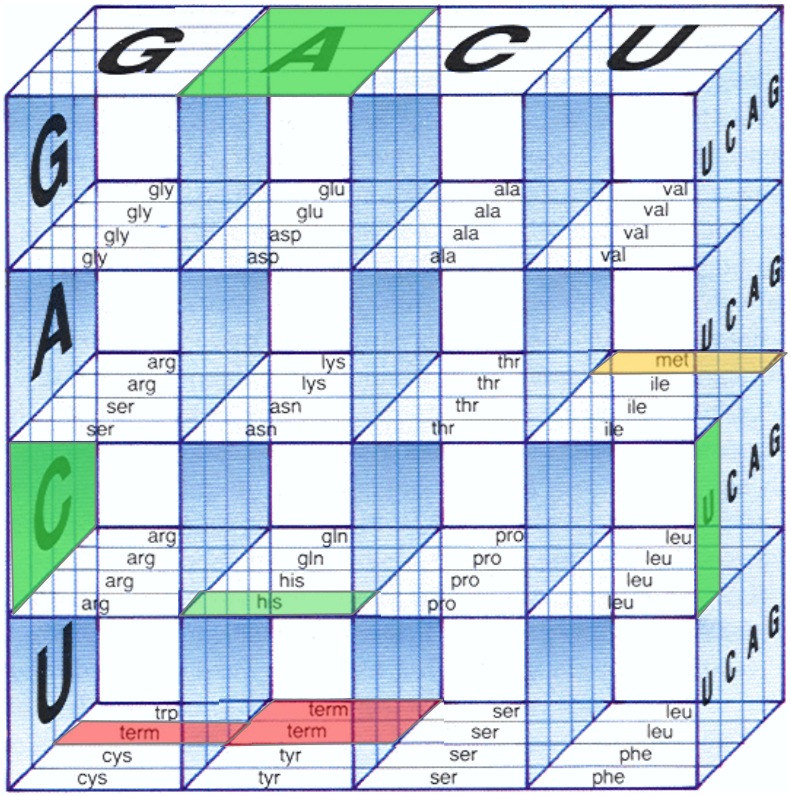




BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

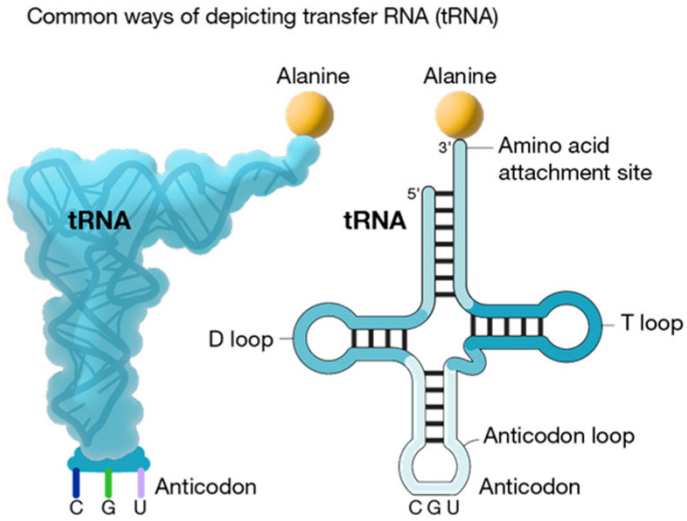
between genotype and phenotype

Triplets of 3 Nucleotides can define 64 possible codons, but only 20 amino acids are used (redundancy)



Figures from Eigen [1992] . *Steps Towards Life*.

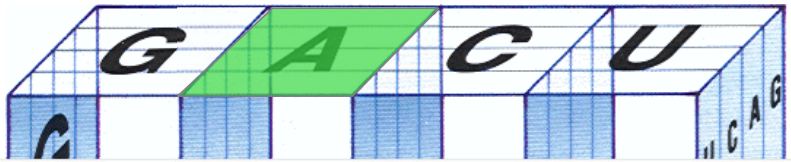
- The genetic code maps information stored in the genome into functional proteins
  - Triplet combinations of nucleotides into amino acids



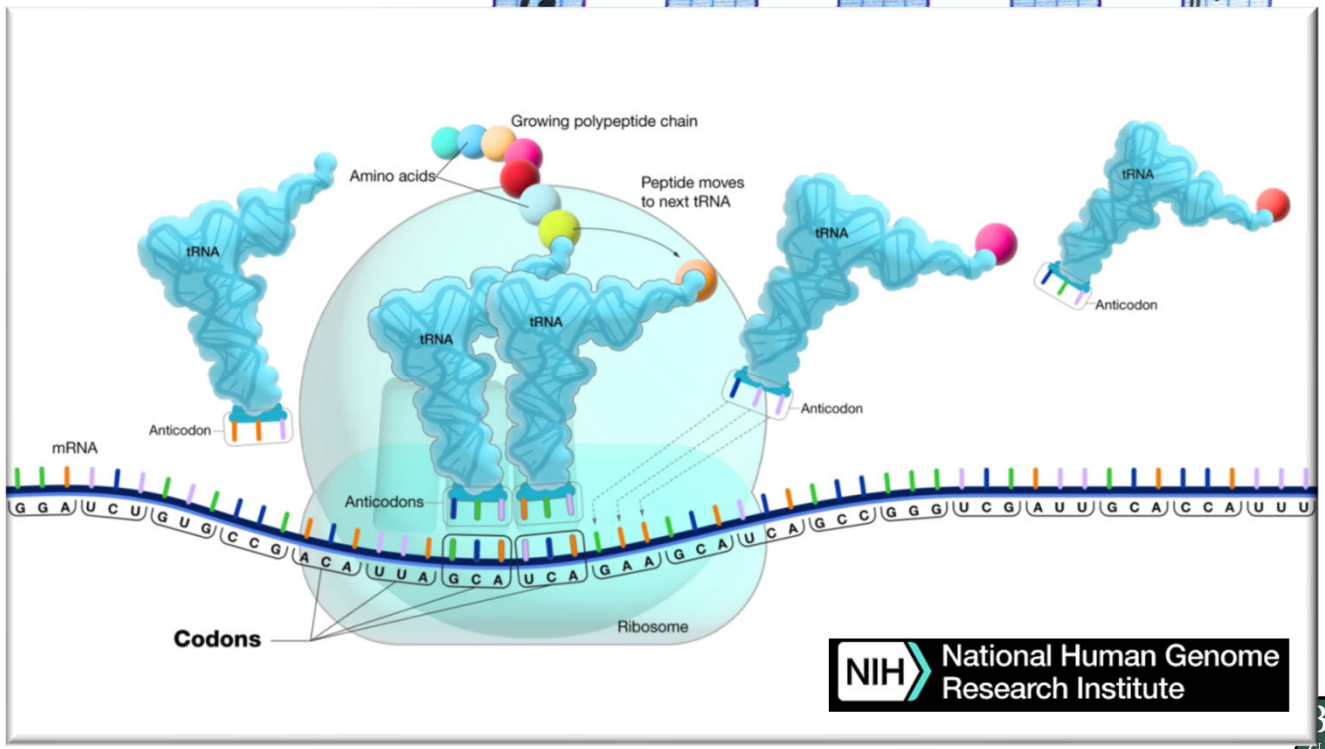
BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

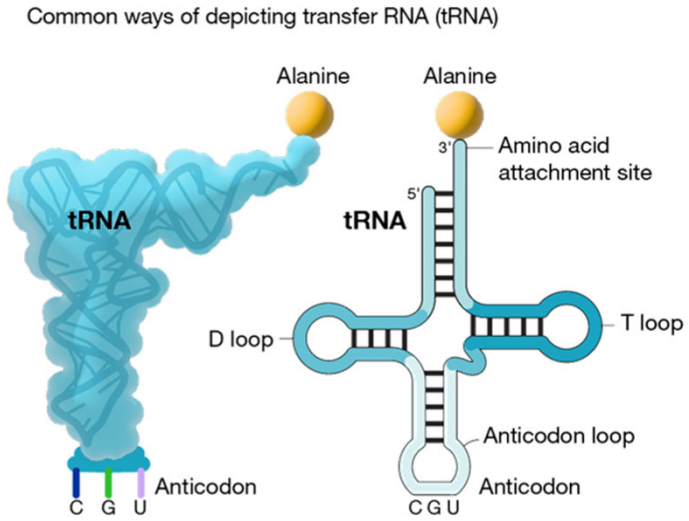
between genotype and phenotype

Triplets of 3 Nucleotides can define 64 possible codons, but only 20 amino acids are used (redundancy)



- The genetic code maps information stored in the genome into functional proteins
  - Triplet combinations of nucleotides into amino acids

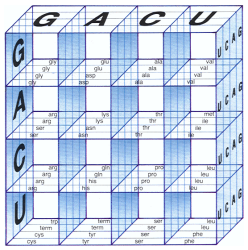




BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

semiotic closure: genetic information at work

genotype/phenotype



Genotype

DNA

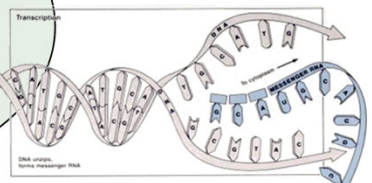
Memory

C

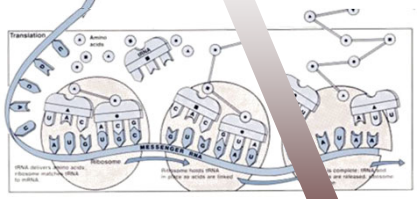
syntax

transcription

RNA

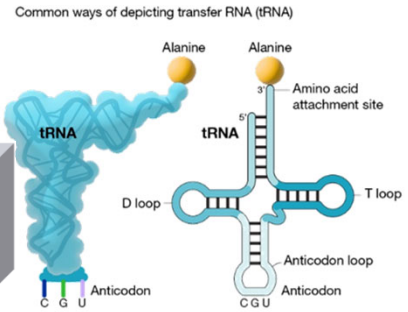


translation



code

A



semantics

C

replication

B

Inherited variation

selection

amino acid chains

Development, regulation

pragmatics

Germ cell line

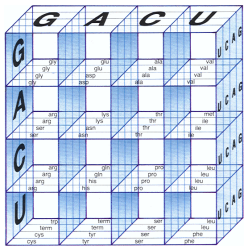
D

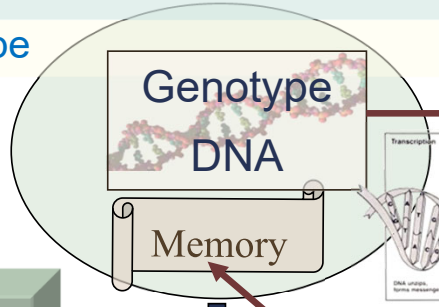
environmental ramifications

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

semiotic closure: genetic information at work

genotype/phenotype

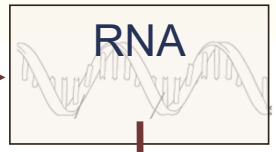


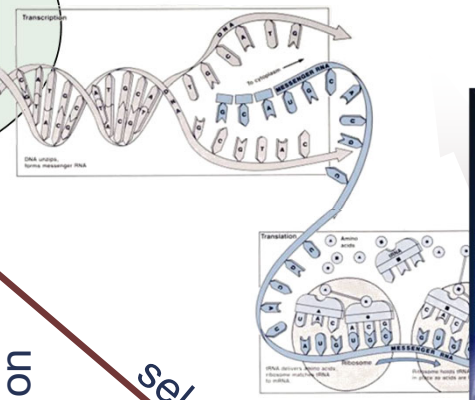


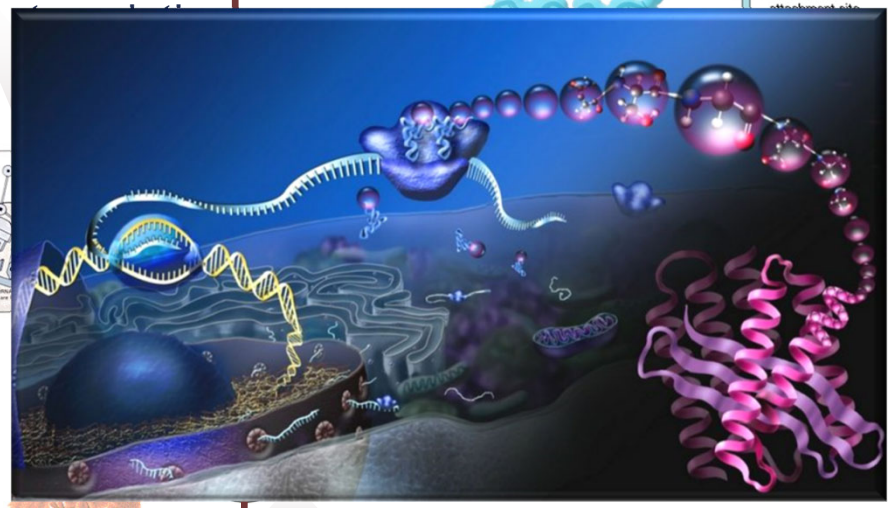
C

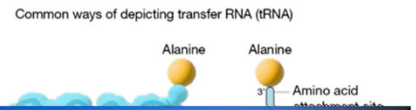
syntax

transcription







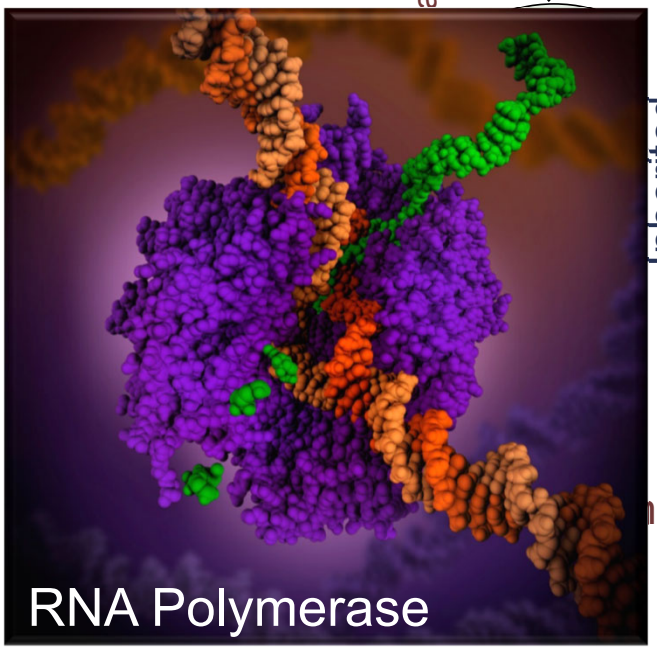


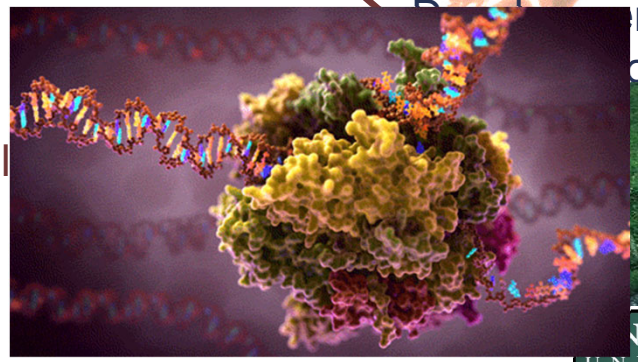
C

ation

Inherited variation

selection

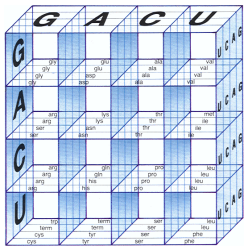




BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

semiotic closure: genetic information at work

genotype/phenotype



Genotype

DNA

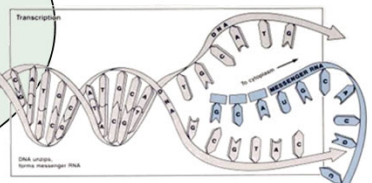
Memory

C

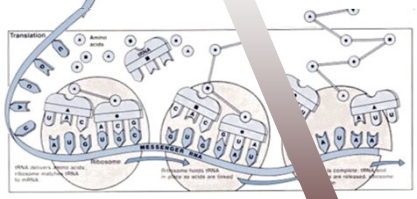
syntax

transcription

RNA

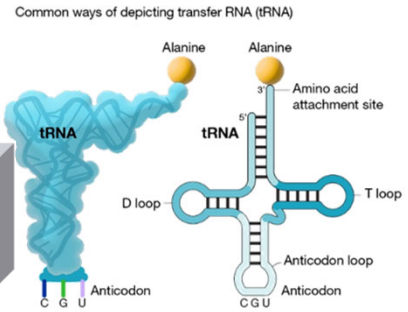


translation



code

A



semantics

C

replication

B

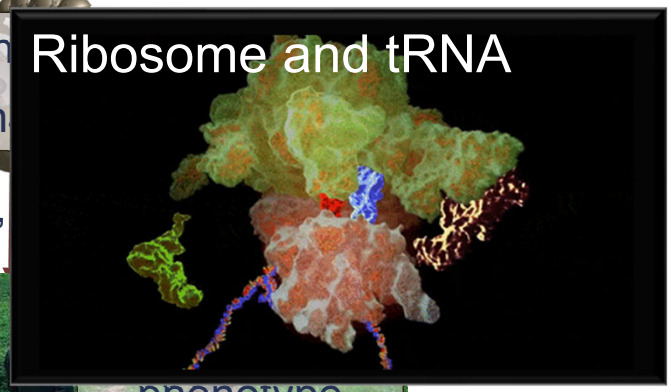
Inherited variation

selection

amin

ch

Ribosome and tRNA



Development, regulation

pragmatics

Germ cell line

D

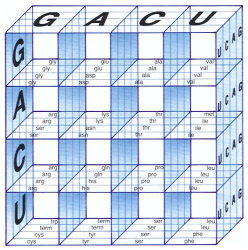
environmental ramifications

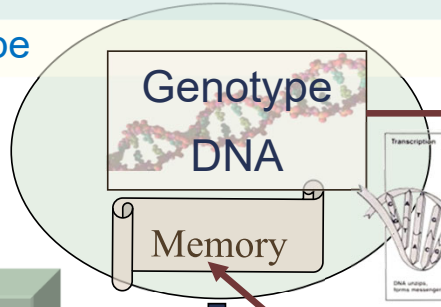
BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

semiotic closure: genetic information at work

genotype/phenotype





C

syntax

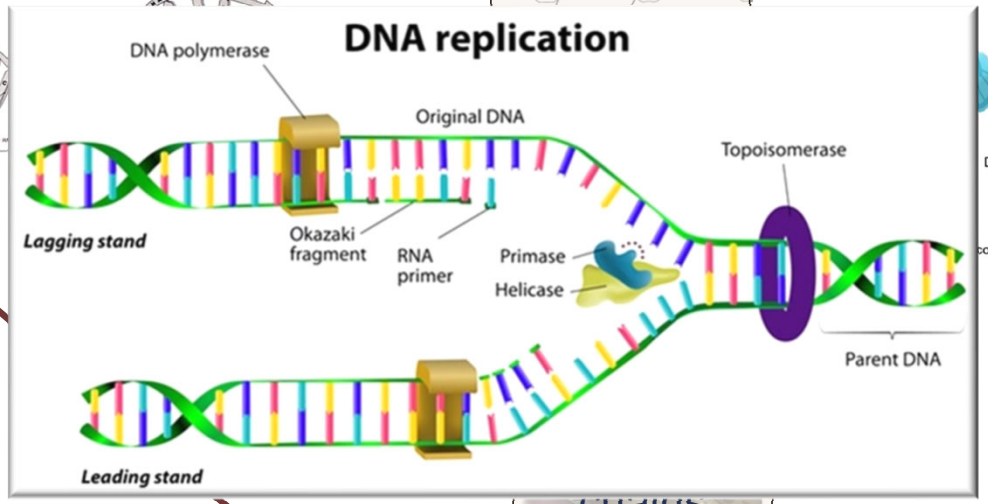
transcription

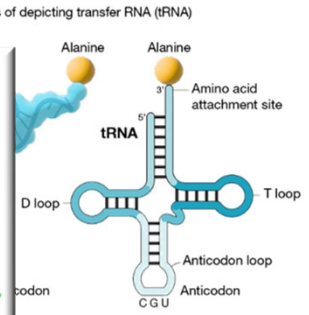
C

replication

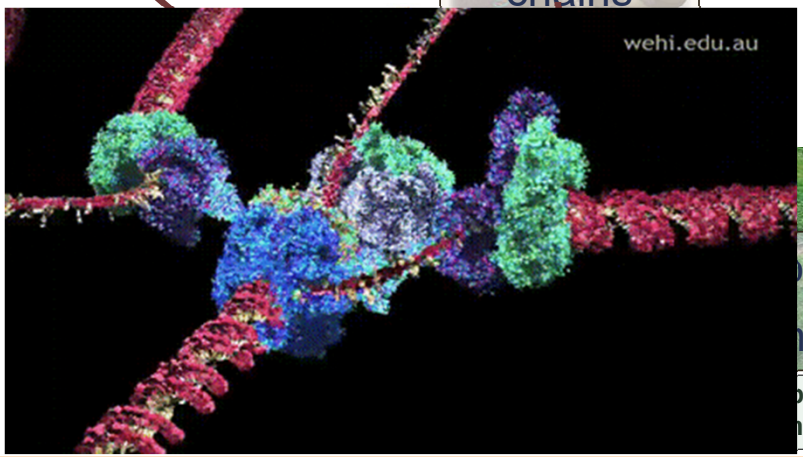
B

Inherited variation

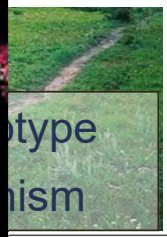




chains



Germ cell line



binghamton.edu
nghamton.edu/academics/ssie501

importance of the “external tape”

in biology

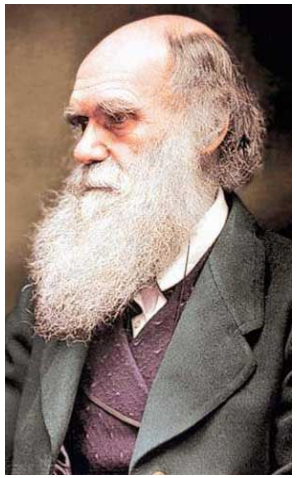
- The “information turn”
  - Unlike Schrödinger, Turing and Von Neumann had no direct effect on molecular biology
  - But the “external tape” separated from the constructor (semiotic closure) has become an unavoidable **principle of organization of biocomplexity**
  - A new synthesis?
    - In 1971 Brenner: “in the next twenty-five years we are going to have to teach biologists another language still, […] where a science like physics works in terms of laws, or a science like molecular biology, to now, is stated in terms of mechanisms, maybe now what one has to begin to think of is algorithms. Recipes. Procedures.”

“The concept of the gene as a symbolic representation of the organism — a ***code script*** — is a fundamental feature of the living world and must form the kernel of biological theory. […] at the core of everything are the tapes containing the descriptions to build these special Turing machines.” (Sydney Brenner)

Brenner, Sydney. [2012]. “Life’s code script.” *Nature* **482** (7386): 461-461.

BINGHAMTON UNIVERSITY STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

fundamental principle of organisms as *cybernetic mechanisms*



BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
rocha@binghamton.edu
casci.binghamton.edu/academics/ssie501

readings

- Class Book
  - Klir, G.J. [2001]. *Facets of systems science*. Springer.
- Papers and other materials
  - Reading and Discussion Group 3 (Enginet)
    - Klir, G.J. [2001]. *Facets of systems Science*. Springer. Chapters 1 and 2.
      - Optional:
      - Rosen, R. [1986]. "Some comments on systems and system theory". *Int. J. of General Systems*, **13**: 1-3. Available in: Klir, G.J. [2001]. *Facets of systems Science*. Springer. pp: 241-243.
      - Wigner, E.P. [1960], "The unreasonable effectiveness of mathematics in the natural sciences". Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Comm. Pure Appl. Math*, **13**: 1-14.
    - Klir, G.J. [2001]. *Facets of systems Science*. Springer. Chapter 3.