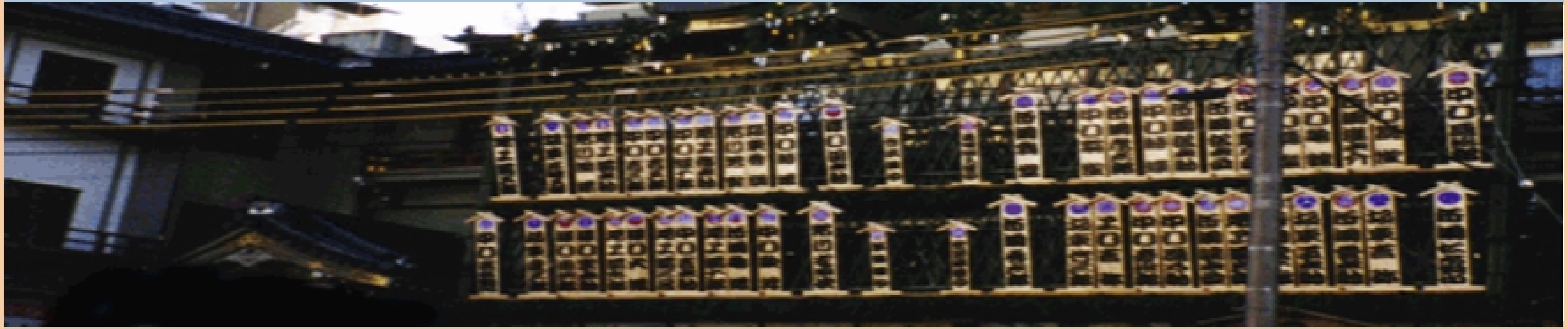
brief introduction to bioinformatics and computational biology



from systems science to systems biology

luis m. rocha
Indiana university
school of informatics and cognitive science program
901 East Tenth Street, Bloomington IN 47408
and
Instituto Gulbenkian de Ciencia
Computational and Mathematical Biology
Oeiras, Portugal

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Cognitive Science Program
the biocomplexity institute
INDIANA UNIVERSITY

summary

- Information Processes in Biology
- Systems Biology, Computational Biology, Bioinformatics
- Synthetic, Multi- Disciplinary Approach to Biology
- Grand Challenges of Systems Biology
- Components of Bioinformatics & Computational Biology
- Some traditional components of Bioinformatics
- Literature Discussion and Useful Resources

rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2006

from systems science to post-genome informatics

> The word "system" is almos never used by itself; it is generally accompanied by an adjective or other modifier: physical system; biological system; social system [...] The adjective describes what is specific and particular; i.e., it refers to the specific "thinghood" of the system; the "system" describes those properties which are independent of this specific "thinghood." [Rosen, 1986]

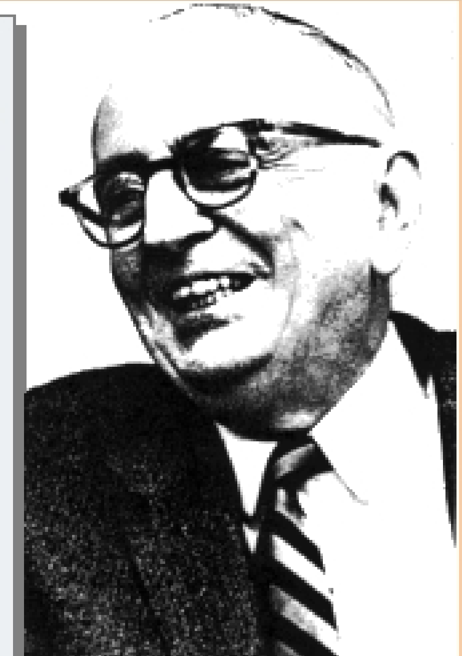
- Systems Science is the methodology used to study *systemhood* not *thinghood* properties in Nature.
  - General Principles of Life (and other systems)
  - Modeling and Simulation of systems measured from and validated in real things.
  - It accumulates knowledge via Mathematical and Computational analysis of classes of systems, models, and problems.
    - Dynamical Systems, Automata Theory, Pattern Recognition, etc.
- Interdisciplinary Meta-Methodology
  - Comparative, Integrative, Non-reductionist
- Historically Related to Cybernetics
  - Complex Systems, Artificial Life

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

dealing with general principles of complex systems

- Weaver [1948] identified 3 types of problems in Science
  - Organized Simplicity: systems with small number of components
    - Classical mathematical tools: calculus and differential equations
  - Disorganized Complexity: systems with large number of erratic components
    - Stochastic, Statistical Methods
  - Organized Complexity: systems with a fair number of components with some functional identity
    - When the behavior of components depends on the organization and function of the whole
    - Techniques depend on Computer Science and Informatics. Require massive combinatorial searches, simulations, and knowledge integration.
    - The realm of Systems Science
  - Complex Systems are systems of many components which cannot be completely understood by the behavior of their components.
    - Complementary models, Hierarchical Organization, Functional decomposition [See Klir, 1991]



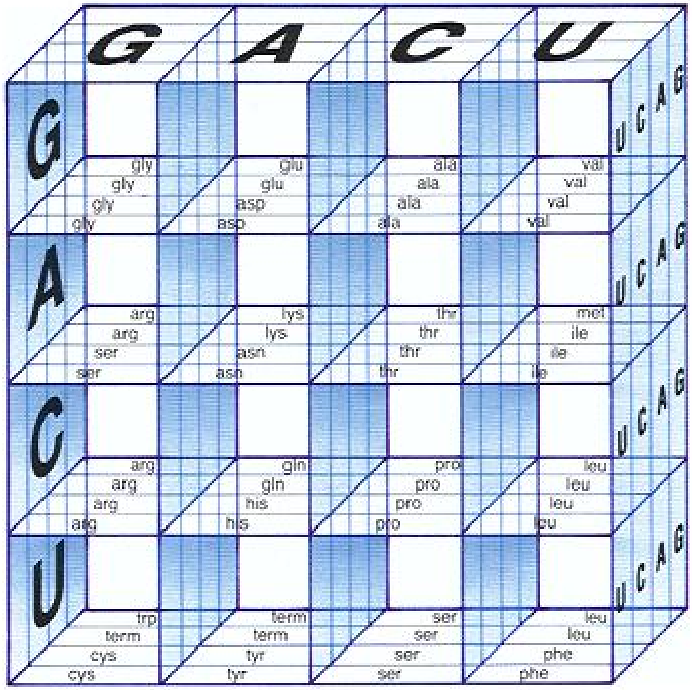
rocha@indiana.edu
http://informatics.indiana.edu/rocha

Systems Biology

And its Involvement with Systems Science

- People
  - Von Bertalanffy [1952, 1968], Mesarovic [1968], Rosen [1972, 1978, 1979, 1991], Pattee [1962, 1979, 1982, 1991, 2001], Maturana and Varela [1980], Kauffman [1991], Conrad [1983], Matsuno [1981], Cariani [1987].
  - Leading Journal: *Biosystems*
- Biology is the most Fundamental Inspiration for Systems Science
  - Cybernetics and Control Theory derive Feedback Control from the physiological concept of Homeostasis
  - Automata Theory, Artificial Intelligence, Artificial Life derived from attempts (by Turing, McCulloch and Pitts) to study the behavior of the Brain and Evolution (Von Neumann)
  - Self-Organizing, Autopoiesis, Complex Adaptive Systems, Artificial Life, Embodied Cognition from developmental and evolutionary biology.
- But Systems Science has had a Small impact in the practice of Biology
  - Due to a large gap between theoretical and experimental biologists.
    - Systems-based theoretical Biology versus a reductionist view
    - Theoretical biology has had more impact on other areas (AI, Alife, Complexity, Systems Science) than Biology itself.

general principles and metaphors from life



Life and Information

biologically
Inspired
computing

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

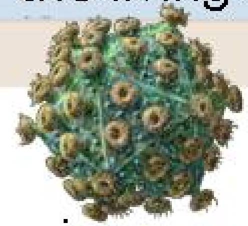
INDIANA
UNIVERSITY

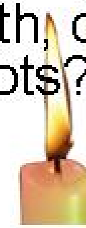
the living organization?

how to identify it?

- **List of properties**
  - Growth
  - Metabolism
  - Reproduction
  - Adaptibility
  - Self-maintenance (autonomy)
  - Self-repair
  - Reaction
  - Evolution
  - Choice
- **Threshold of complexity**
  - Categorization and Control
  - Function (self-reference)
  - Open-ended evolution
  - Information

Is life Fuzzy?





viruses, candle flames, the Earth, certain robots?

Is there a synthetic criteria? How general can it be?

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

the roles of information

in the living organization

- organisms act according to information they perceive in an environment
- organisms reproduce and develop from genetic information
  - genetic information is ***transmitted*** "vertically" (inherited) in phylogeny and cell reproduction, and ***expressed*** "horizontally" within a cell in ontogeny and plain functioning
- Self-reference
  - Information relevant to organism: ***function***
  - Only in ***reference*** to an organism does a piece of DNA ***function*** as a gene
  - Biology is contextual, physics is universal

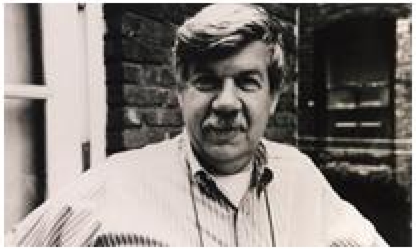
"Life is a dynamic state of matter organized by information". Manfred Eigen [1992]

"Biology and physics have nothing to do with each other because biological evolution is essentially historical, and physical laws must be independent of history". Ernst Mayer

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

emergence and explanation

- impossibility of epistemological reduction of the properties of a system to its components
  - Wave-particle duality
  - Information and function are contextual and historical
    - “Clockness”: many possible material implementations
    - Several biological designs for similar function (e.g. flying)
    - The function of DNA does not lie in its dynamic (bio-chemical) characteristics



“First, nothing in biology contradicts the laws of physics and chemistry; any adequate biology must be consonant with the ‘basic’ sciences. Second, the principles of physics and chemistry are not sufficient to explain complex biological objects because new properties emerge as a result of organization and interaction. These properties can only be understood by the direct study of the whole, living systems in their normal state. Third, the insufficiency of physics and chemistry to encompass life records no mystical addition, no contradiction to the basic sciences, but only reflects the hierarchy of natural objects and the principle of emergent properties at higher levels of organization”. Stephen Jay Gould

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

How much is specific bio-chemistry?

- Can there be several implementations of life?
  - To study life do we need to find and synthesize the necessary threshold of complexity?
    - Hard and wet Artificial Life
  - Or is it enough to simulate the behavior of life?
    - Soft Artificial Life
- Important to study the living organization
  - What can be abstracted and implemented in a different medium?
  - Understanding organization and design principles
    - Scientific advancement of the essential principles of life
      - Systems Biology, Artificial Life
    - Solving engineering and design problems
      - Bio-inspired computing

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

Information Processes in Biology

- **Genetic System**
  - Construction (expression, development, and maintenance) of cells ontogenetically: horizontal transmission
  - Heredity (reproduction) of cells and phenotypes: vertical transmission
- **Immune System**
  - Internal response based on accumulated experience (information)
- **Nervous and Neurological system**
  - Response to external cues based on memory
- **Language, Social, Ecological, Eco-social, etc.**

"Life is a complex system for information storage and processing". Minoru Kanehisa [2000]

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

information units in biology

memory, structural and functional

- Mendelian Gene
  - Hereditary unit responsible for a particular characteristic or trait
- Molecular Biology Gene
  - Unit of information expression via ***transcription*** and ***translation***
    - Horizontal information expression (semantics, active)
- Genome
  - Unit of information transmission via ***DNA replication***
    - Vertical information transmission (syntactic, passive)
  - Set of genes in the chromosome of a species
- Genotype
  - Instance of the genome for an individual
- Phenotype
  - Expressed and developed genotype
    - Genes have different alleles
- Transcriptome
  - Set of expressed genes (mRNA transcripts) in a given context
- Proteome
  - Set of proteins that are encoded and expressed by a genome

"Biology is the science of life that aims at understanding both functional and structural aspects of living organisms". Minoru Kanehisa [2000]

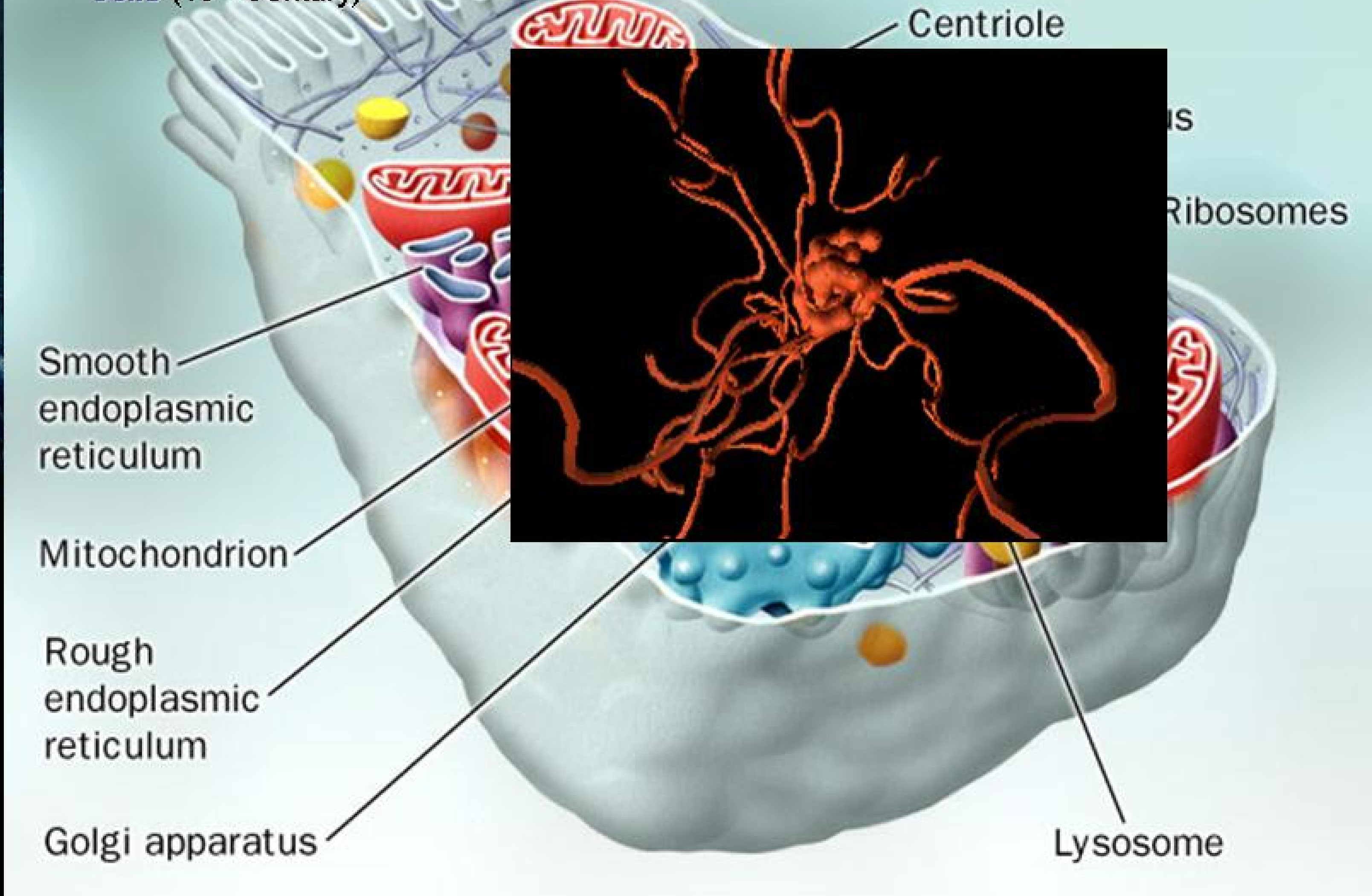
What about evolution and development?

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

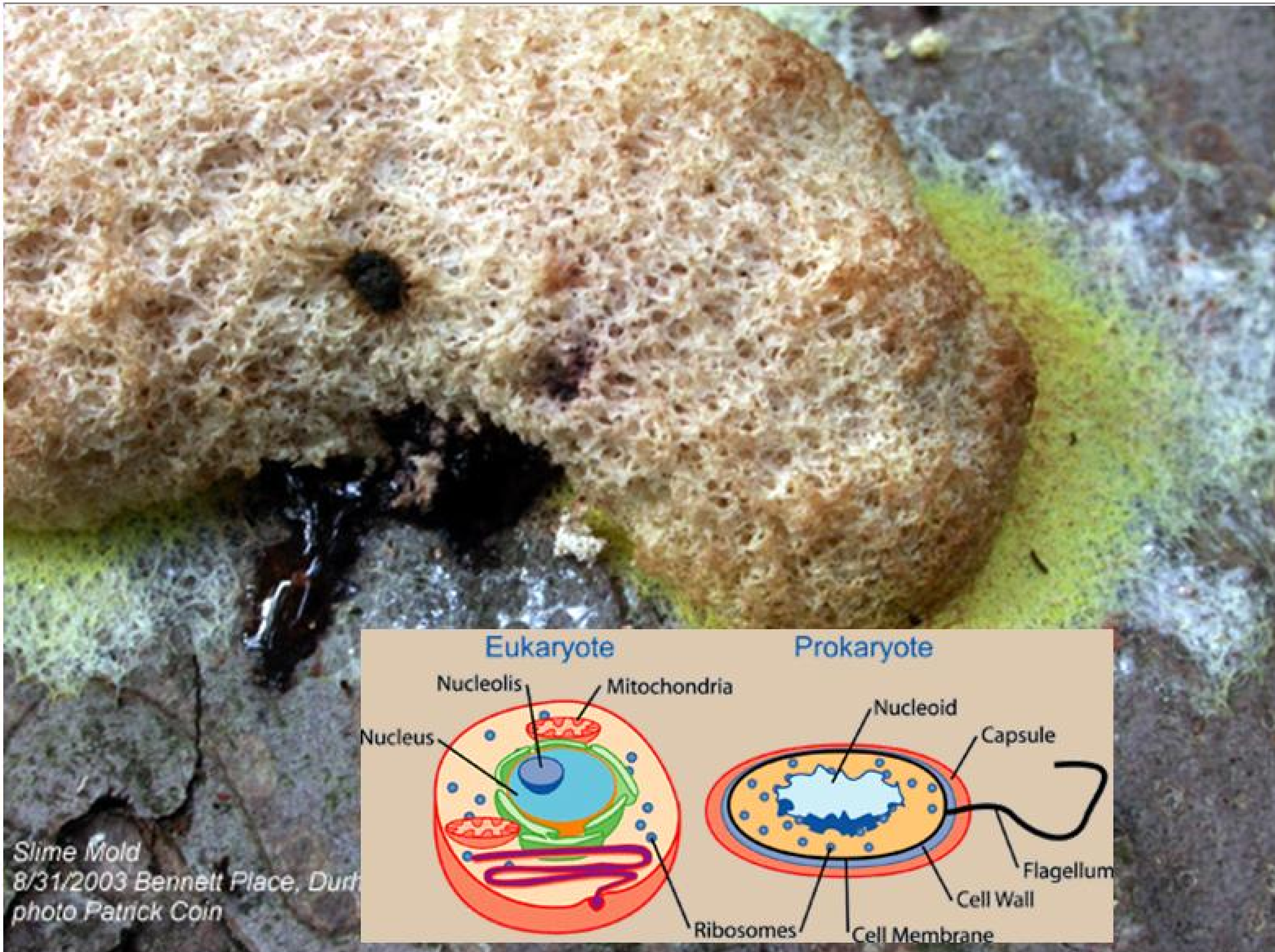
the eukaryotic cell

Structure and organelles

Life is made up of structural and functional units called *cells* (19th Century)



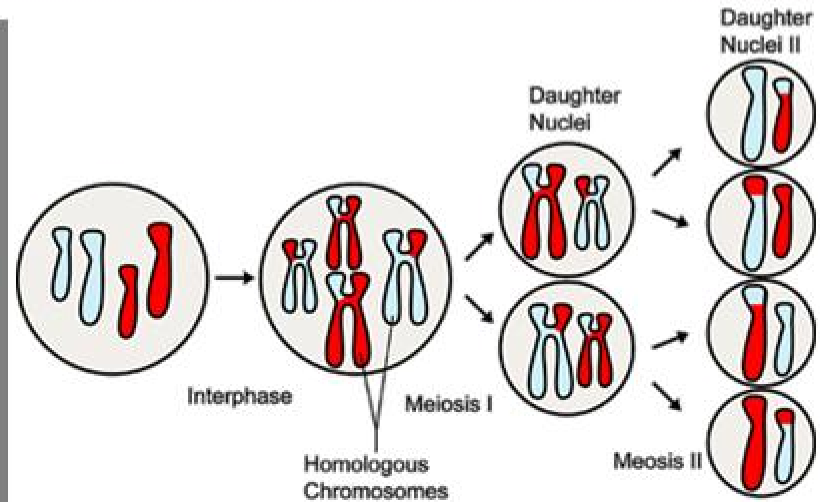
more about cells



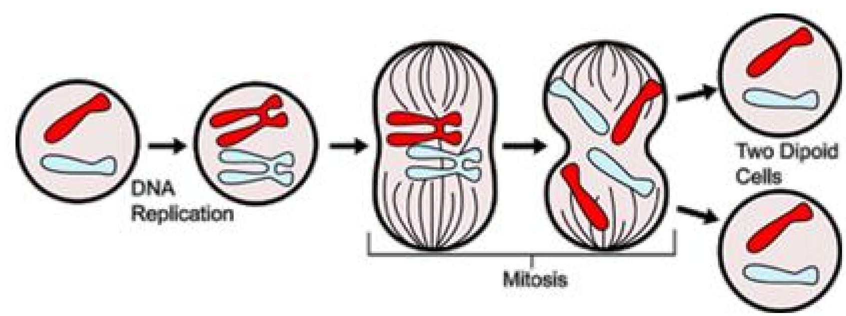
biologically
inspired
computing

Cell division

- Meiosis
  - Diploid cell's genome is replicated once and split twice
  - produces four haploid (***germ***) cells each with half the chromosomes
  - Sexual reproduction combines germ cells from two individuals to produce diploid (***zygocyte***) cells
  - Vertical genetic information transmission
    - Offspring with a new genotype
- Mitosis
  - Eukaryotic cell separates its duplicated genotype into two identical halves
    - somatic cells in multicellular organisms
  - Horizontal genetic information expression
    - development



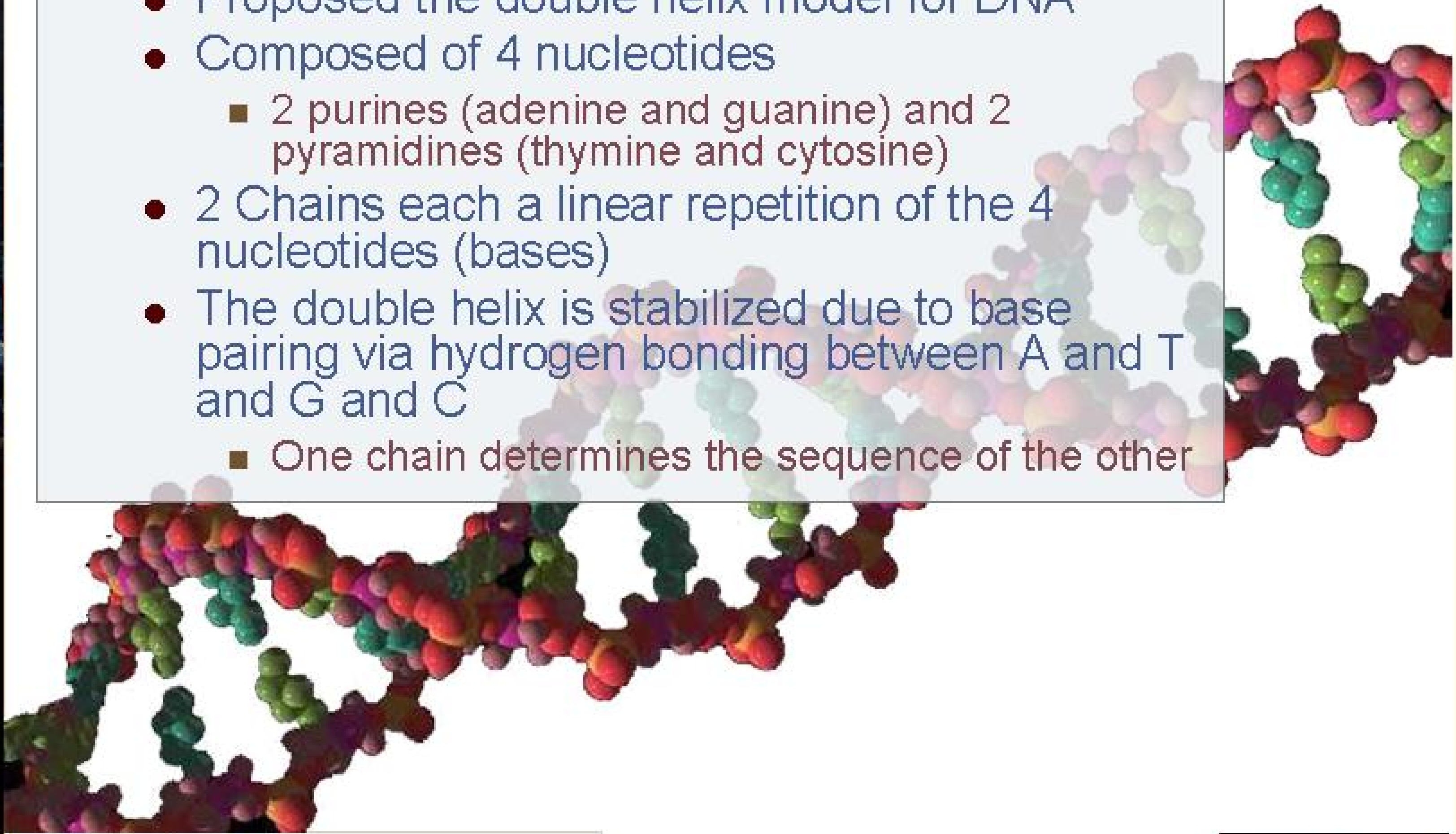
Crossovers may occur in meiosis



DNA

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



rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

a molecular language system

Purine (R) → Adenine (A), Guanine (G)

Nucleotides

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

Uracil (U)

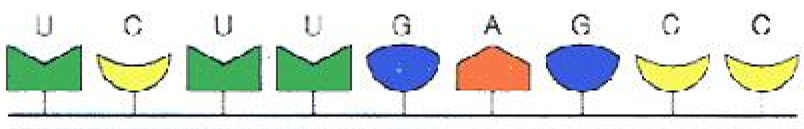
4 Letter Alphabet
DNA: A, G, C, T
RNA: A, G, C, U

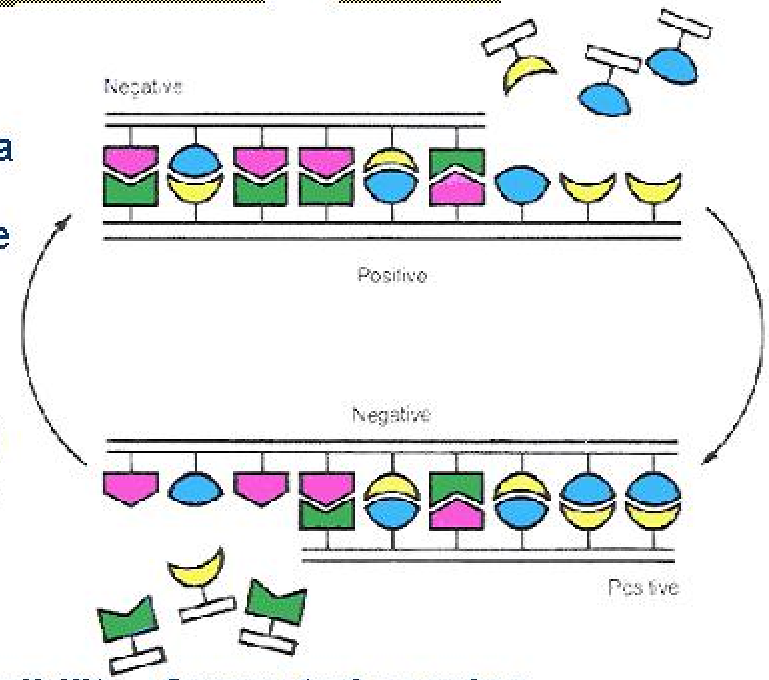
Form sequences that can store information

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

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

A-T (U)   G-C





Requirements for structural information

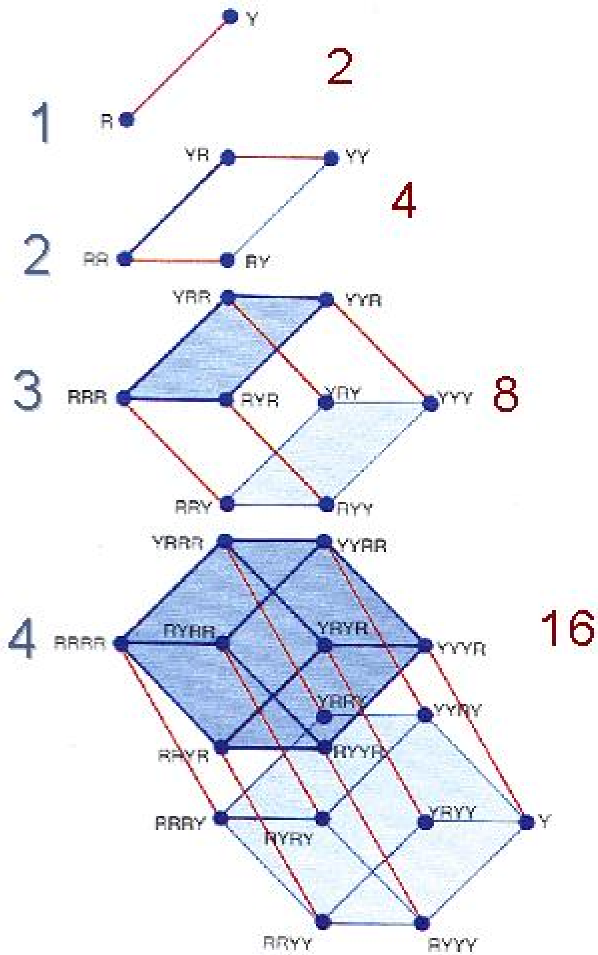
Possibility of repeated copying

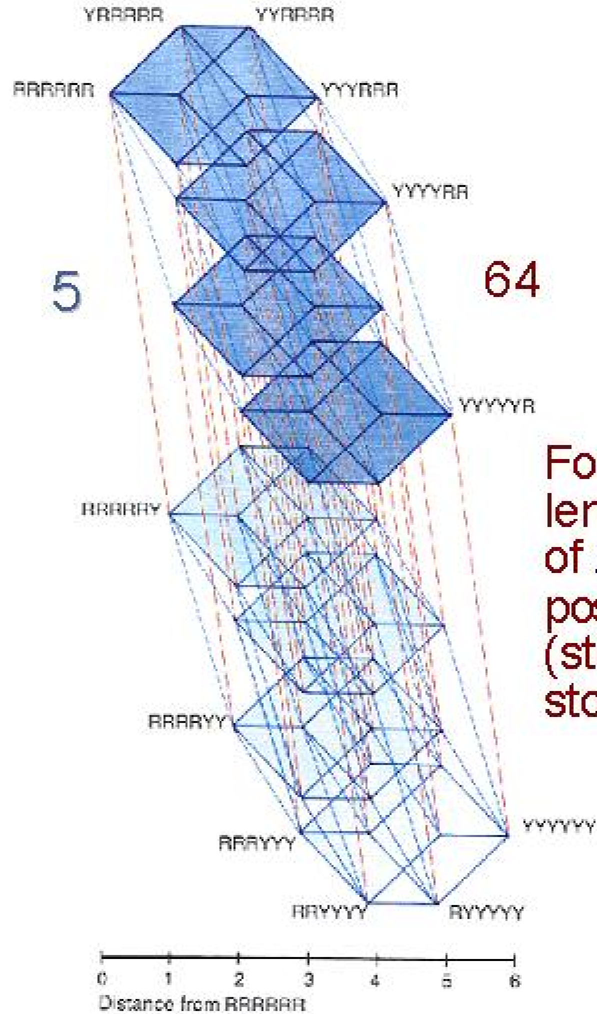
rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

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

INDIANA UNIVERSITY

Information and Sequence Space





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.*

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

Proteins

functional products

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

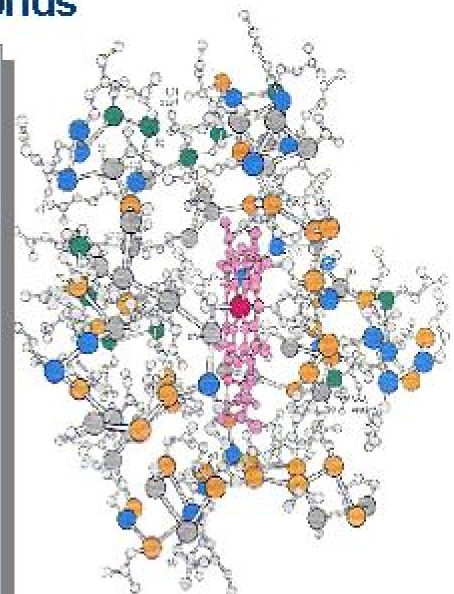


Table 1.4. Amino acid codes

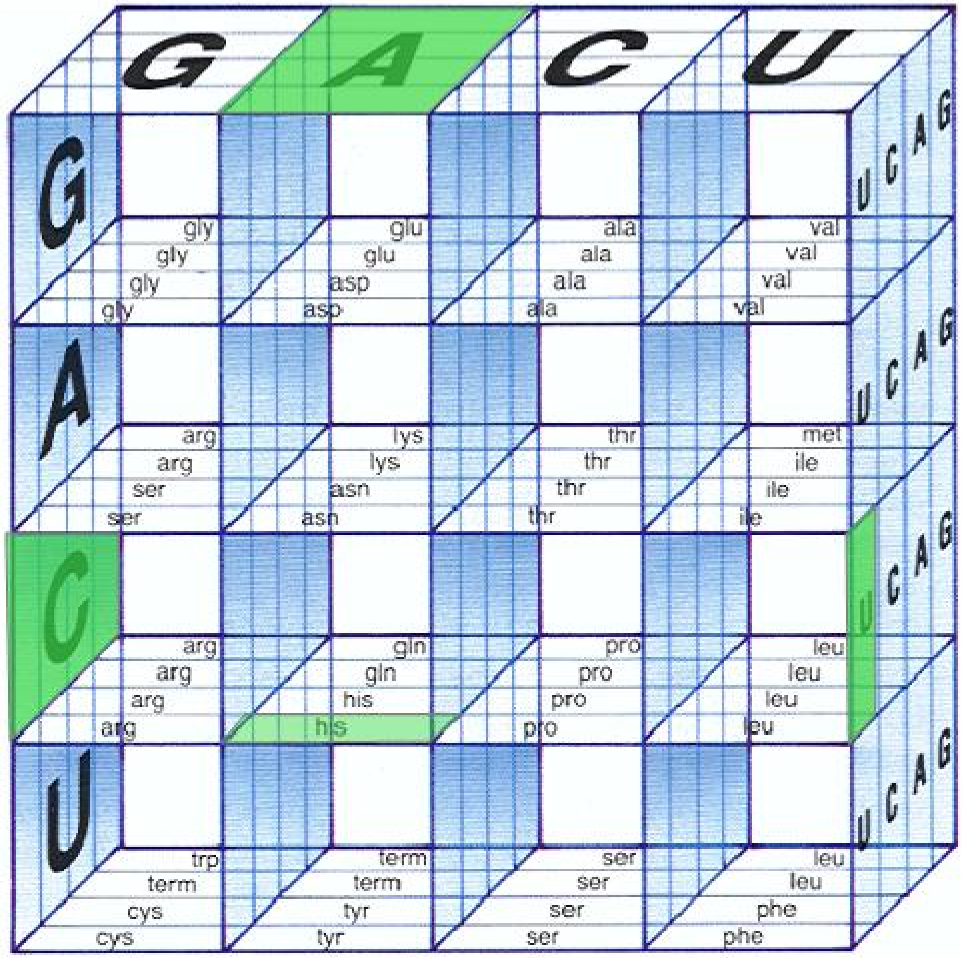
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
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.*

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

The Genetic Code

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



Triplets of 4 Nucleotides can define 64 possible codons, but only 20 amino acids are used (redundancy)

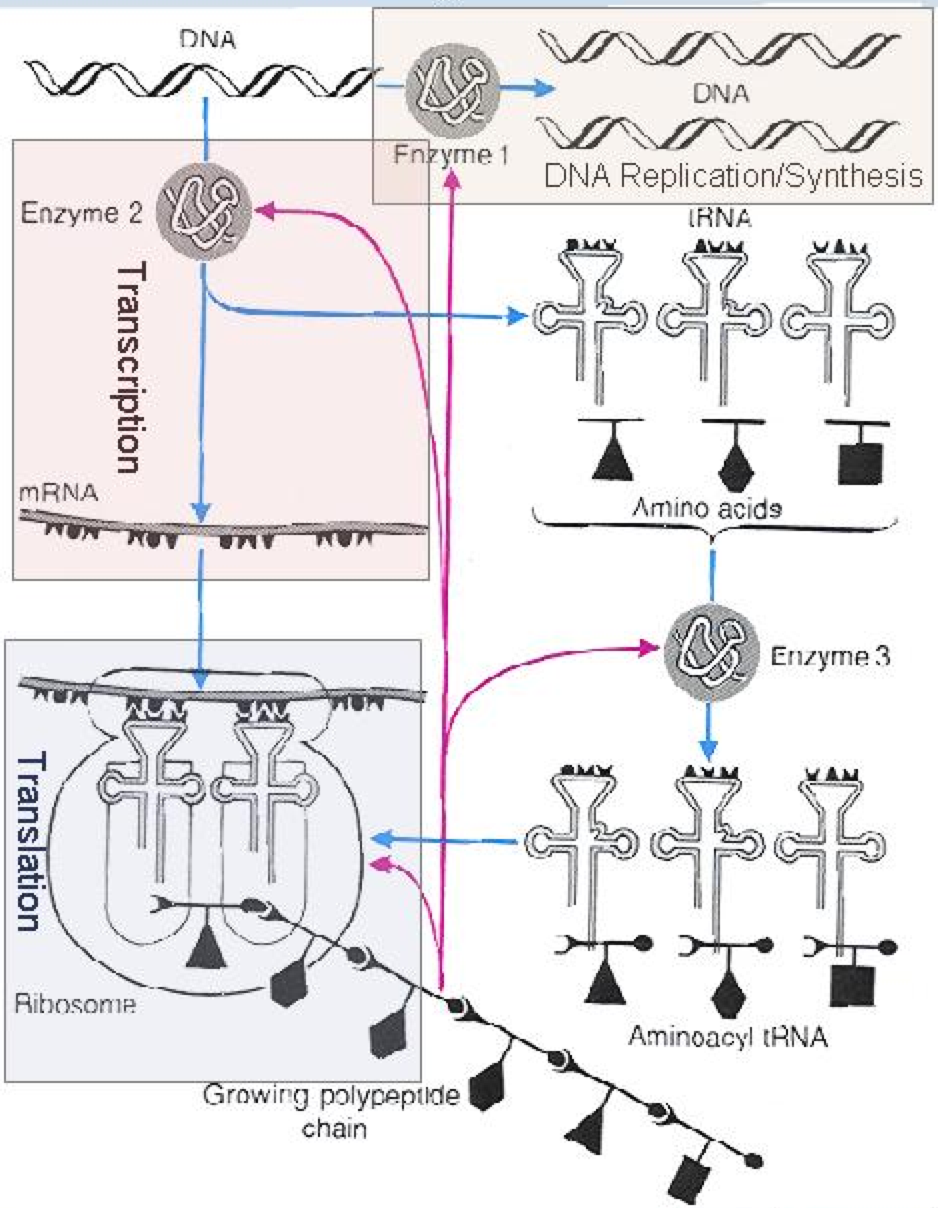
biologically
Inspired
computing

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

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

the genetic code at work

- Reproduction
  - DNA Polymerase
- Transcription
  - RNA Polymerase
- Translation
  - Ribosome
- Coupling of AA's to adaptors
  - Aminoacyl Synthetase

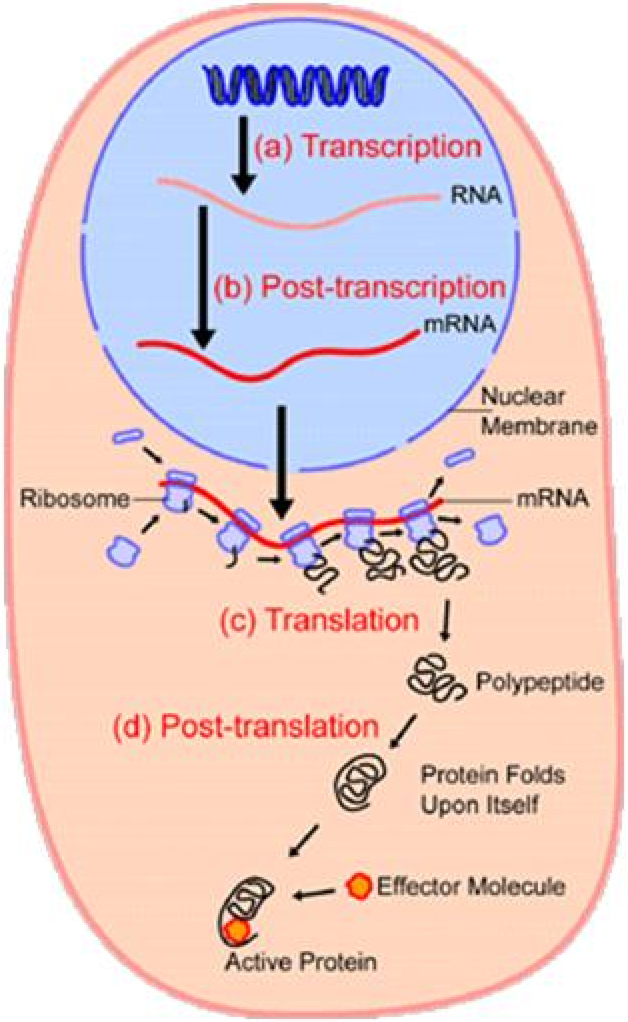


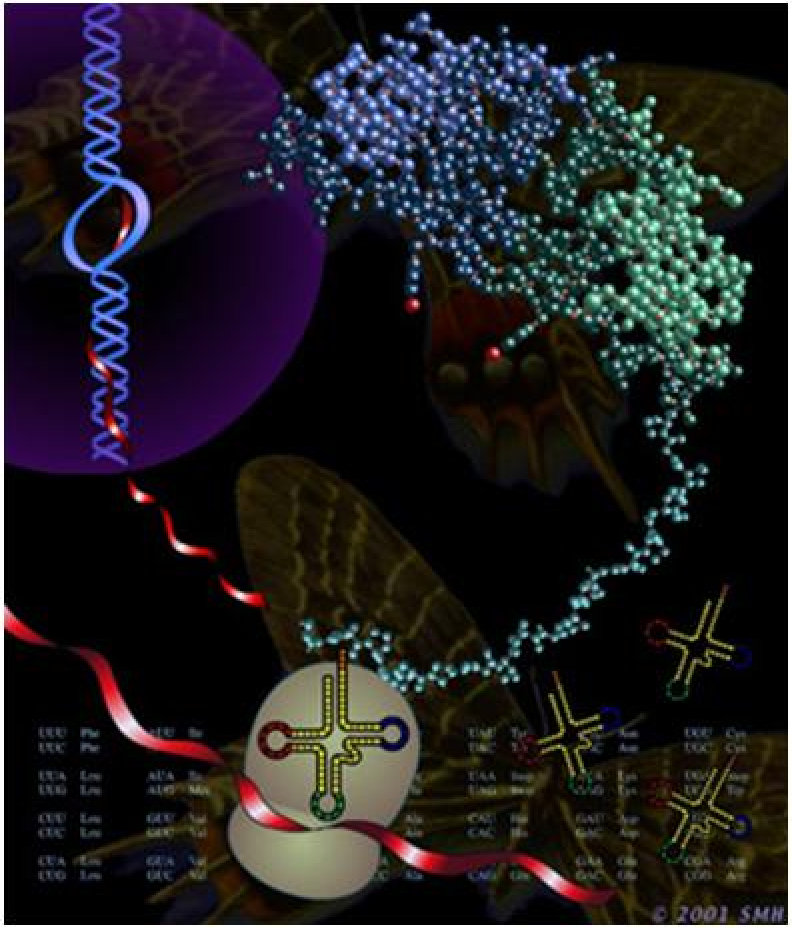
Figures from Eigen [1992]. *Steps Towards Life.*

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

information expression in the genetic system

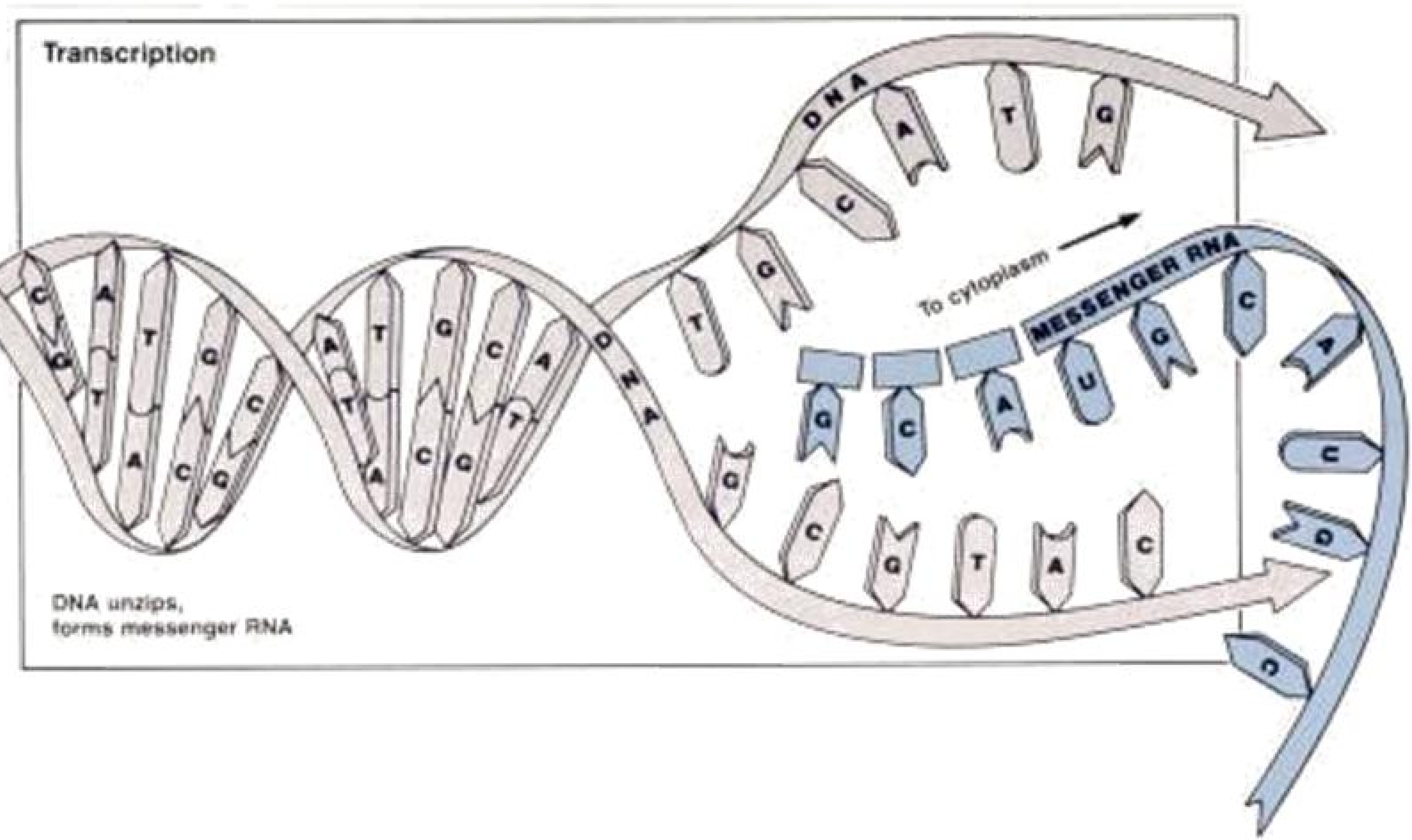
transcription and translation



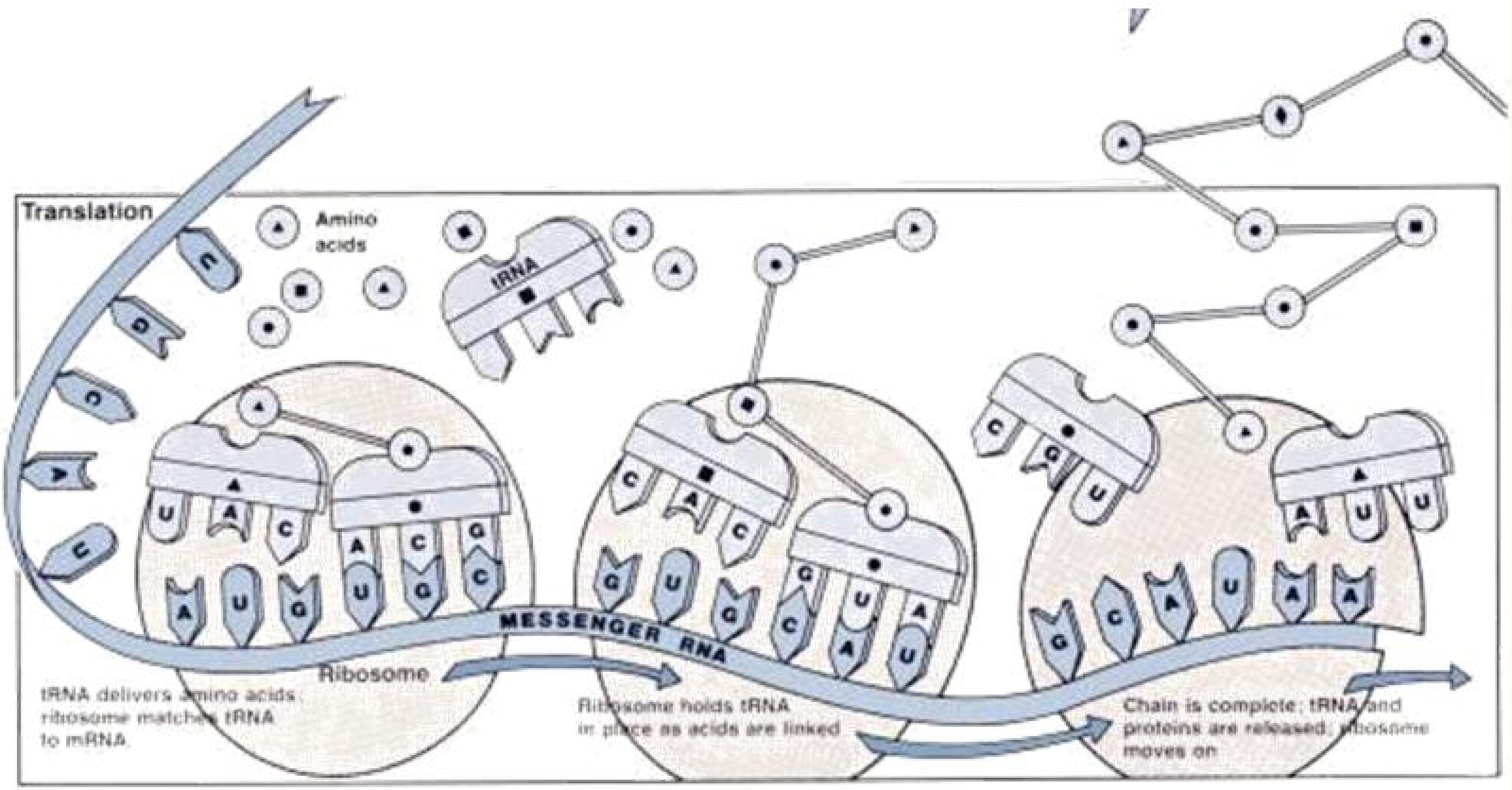


rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

generating a message to be expressed



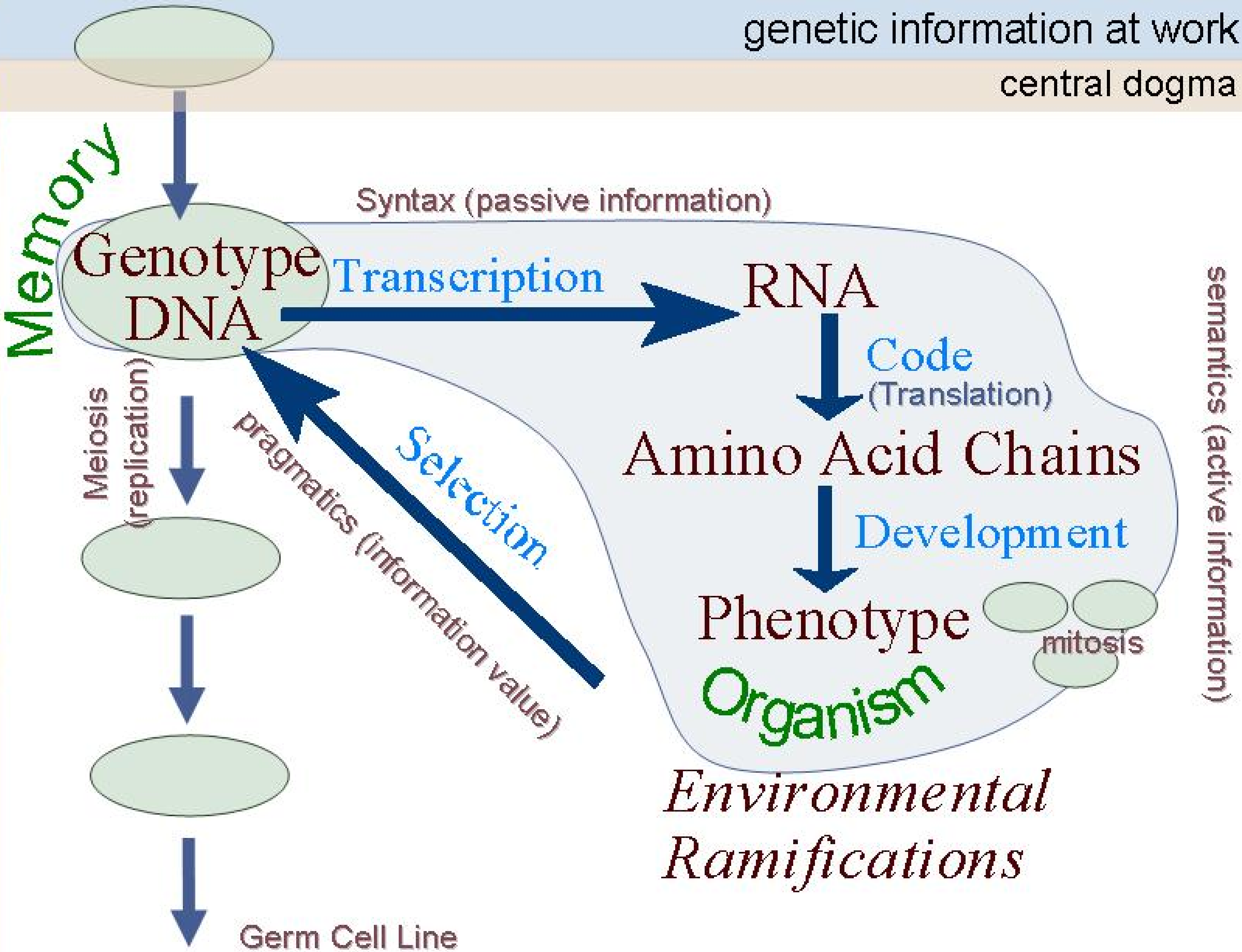
constructing (decoding) the message



rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

genetic information at work

central dogma



rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

life-as-it-could-be

- Chris Langton
  - Artificial Life can contribute to theoretical biology by locating *life-as-we-know-it* within the larger picture of *life-as-it-could-be*
  - life as a property of the *organization* of matter, rather than a property of the matter which is so organized
    - The way information is processed
  - Whereas biology has largely concerned itself with the material basis of life, Artificial Life is concerned with the formal basis of life.
    - views an organism as a large population of *simple* machines
    - *Synthetic approach* or *emergent behavior*

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

scientific approaches of life

- Analytical
  - Reduction to (non-living) components
    - Reductionism
  - Life is complicated chemistry
  - Tied to specific materiality
  - Does not allow emergence
    - Function, control, measurement, categorization, information are unnecessary "illusions"

- Synthetic
  - Construction from components
    - Holist
  - Life is Organization
    - Networks of components
  - Universal or implementation independent
  - Emergence
    - "bottom-up" approach

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

goals of alife

- Hard Alife
  - Logical mechanisms of life
  - Discover and synthesize the design principles of life
    - Threshold of complexity
    - Lists of characteristics
- Soft or weak Alife
  - To simulate life
  - Compare design principles of life with simulations
  - Extract design principles to solve problems
    - Bio-inspired computing
- Bottom-up methodology

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic

Systemhood

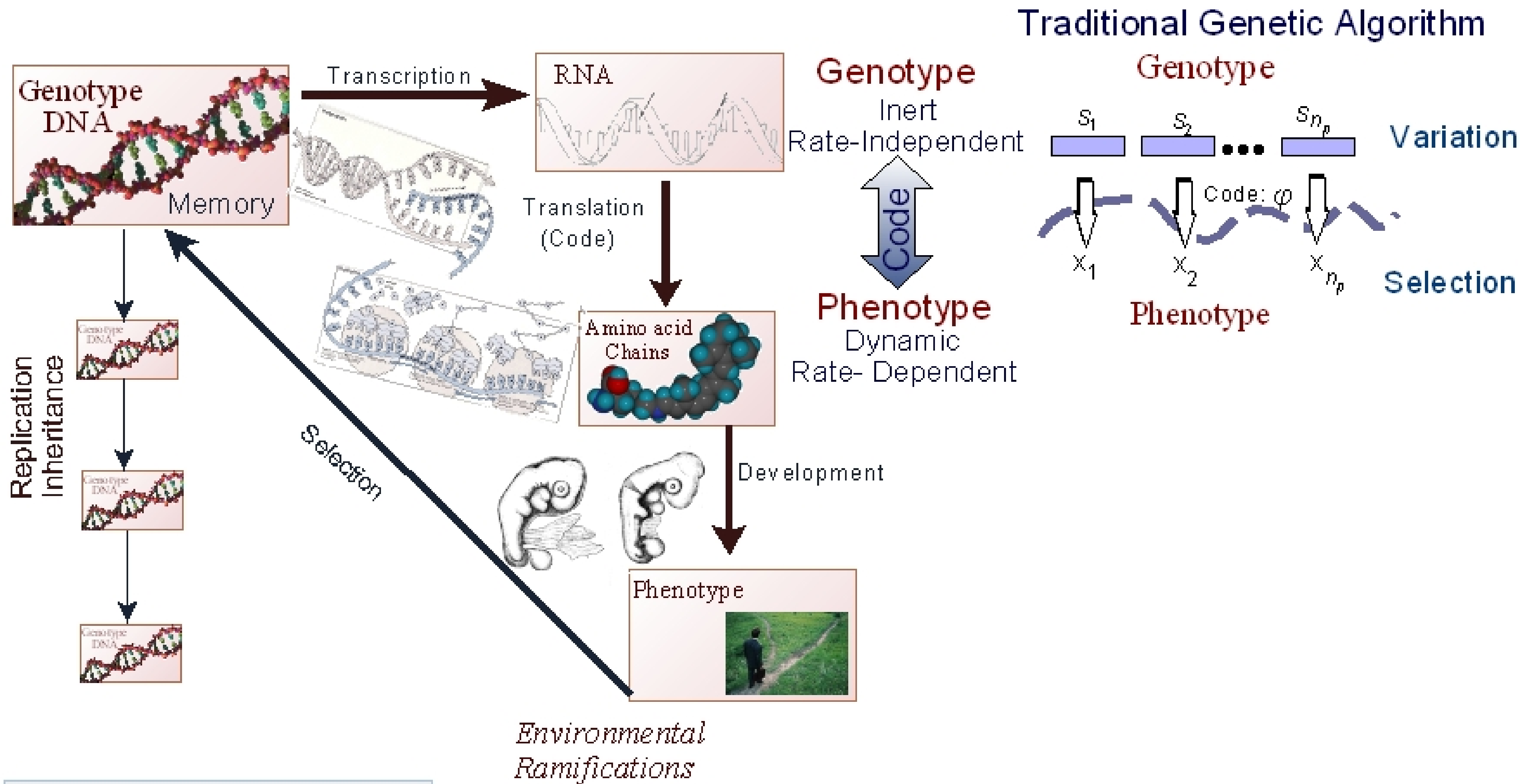
- A system possesses ***systemhood*** and ***thinghood*** properties
  - Thinghood refers to the specific material that makes up the system
  - Systemhood are the abstracted properties
    - E.g. a clock can be made of different things, but there are implementation-independent properties of “clockness”
  - Systems science deals with the implementation-independent aspects of systems
    - Robert Rosen, George Klir…

rocha@indiana.edu
http://informatics.indiana.edu/rocha/i-bic
INDIANA UNIVERSITY

informatics
luis rocha 2006

traditional view of gene function

genotype/phenotype mapping in artificial life



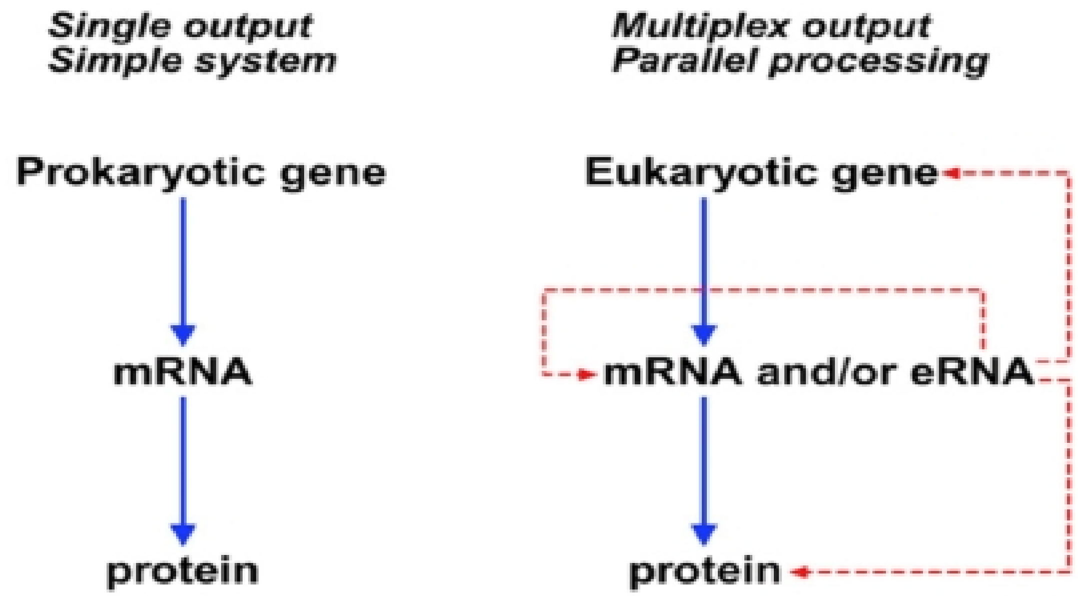
rocha@indiana.edu
http://informatics.indiana.edu/rocha

non-coding RNA

ncRNA: a regulatory hidden layer in Eukaryots

- Evidence for non-protein coding RNA (ncRNA) in complex organisms (higher eukaryotes)
  - "ncRNA dominates the genomic output of the higher organisms and has been shown to control chromosome architecture, mRNA turnover and the developmental timing of protein expression, and may also regulate transcription and alternative splicing."
    – Mattick, J. S. (2003). *BioEssays*. **25**: 930-939
  - A ***Hidden Layer*** of Non-protein-coding RNAs in Complex Organisms.

- Two types of genetic information
  - ***mRNA*** for proteins
  - ***ncRNA*** for RNA products
- Three types of genes in eukaryotes
  - Encoding only proteins
  - Encoding only ncRNA
  - Encoding both
- Many types of ncRNA
  - tRNA, rRNA, SnoRNA, miRNA, siRNA, eRNA, etc.



Mattick, J. S. [2001]. *EMBO Reports* 2, 11, 986–991
Mattick, J. S. And V. Makunin [2005]. *Human Molecular Genetics* 14, 11, R121-R132

RNA editing

u-insertion (via gRNA)

Ser Gly Glu Lys

AuGuuuCGuuGuAGAuuuuuAuuAuuuuuuuuAuuA

Mer Phe Arg Cys Arg Phe Leu Leu Phe Phe Leu Leu

Gln Glu Gly Arg Gly Lys •••

CAGGAGGGCCGUGGAuAAG

Gln Glu Gly Arg Gly STOP

Bass, B.L. (Ed.) (2001). *RNA Editing.*

Benne, R. (Ed.) (1993). *RNA Editing: The Alteration of Protein Coding Sequences of RNA.*

- RNA Editing: post-transcriptional alteration of genetic information
  - can be performed by ncRNA structures (and proteins).
- U-Insertion/deletion RNA Editing (mitochondria of kinetoplastid protozoa)
  - involve small *guide RNAs* (gRNA) complementary to the target mRNA
- Adenosine (A) to inosine (I) Substitition (higher Eukaryotes)
  - Inosine (I) is read as guanosine (G) in translation
  - Involve enzymes: *adenosine deaminases acting on RNA* (ADAR)/ *RNA Editase*
    - Implicated in epilepsy, Parkinson's Disease, depression, etc.

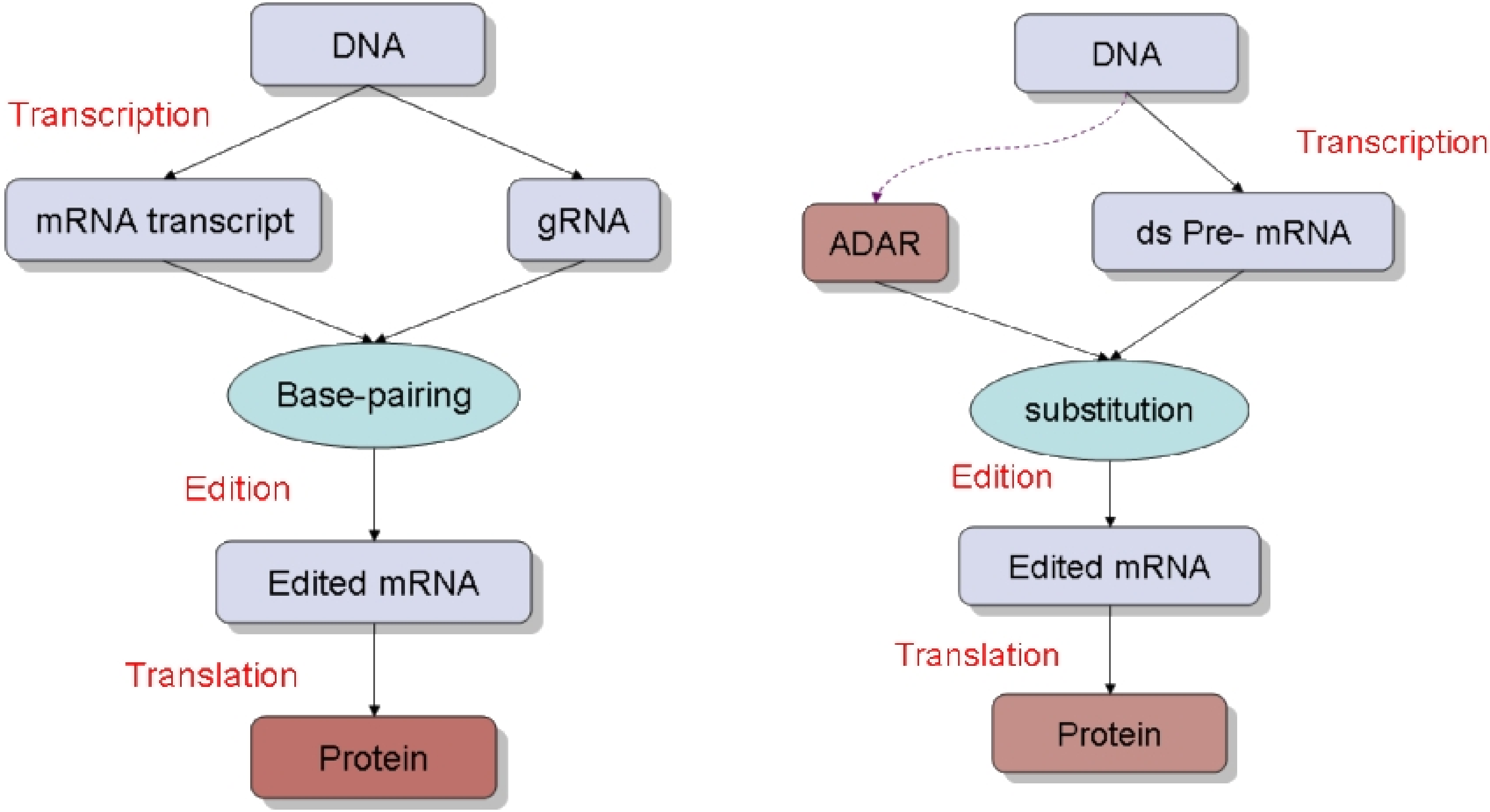
informatics
luis rocha 2006

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

RNA editing

U-insertion and A-to-I subtitution

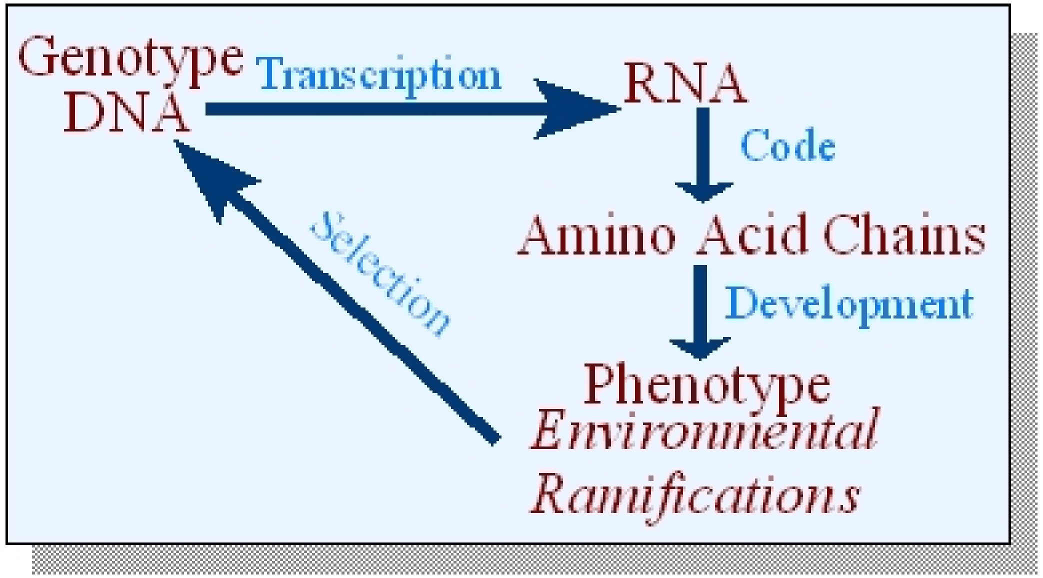


informatics
luis rocha 2006
rocha@indiana.edu
http://informatics.indiana.edu/rocha
INDIANA UNIVERSITY

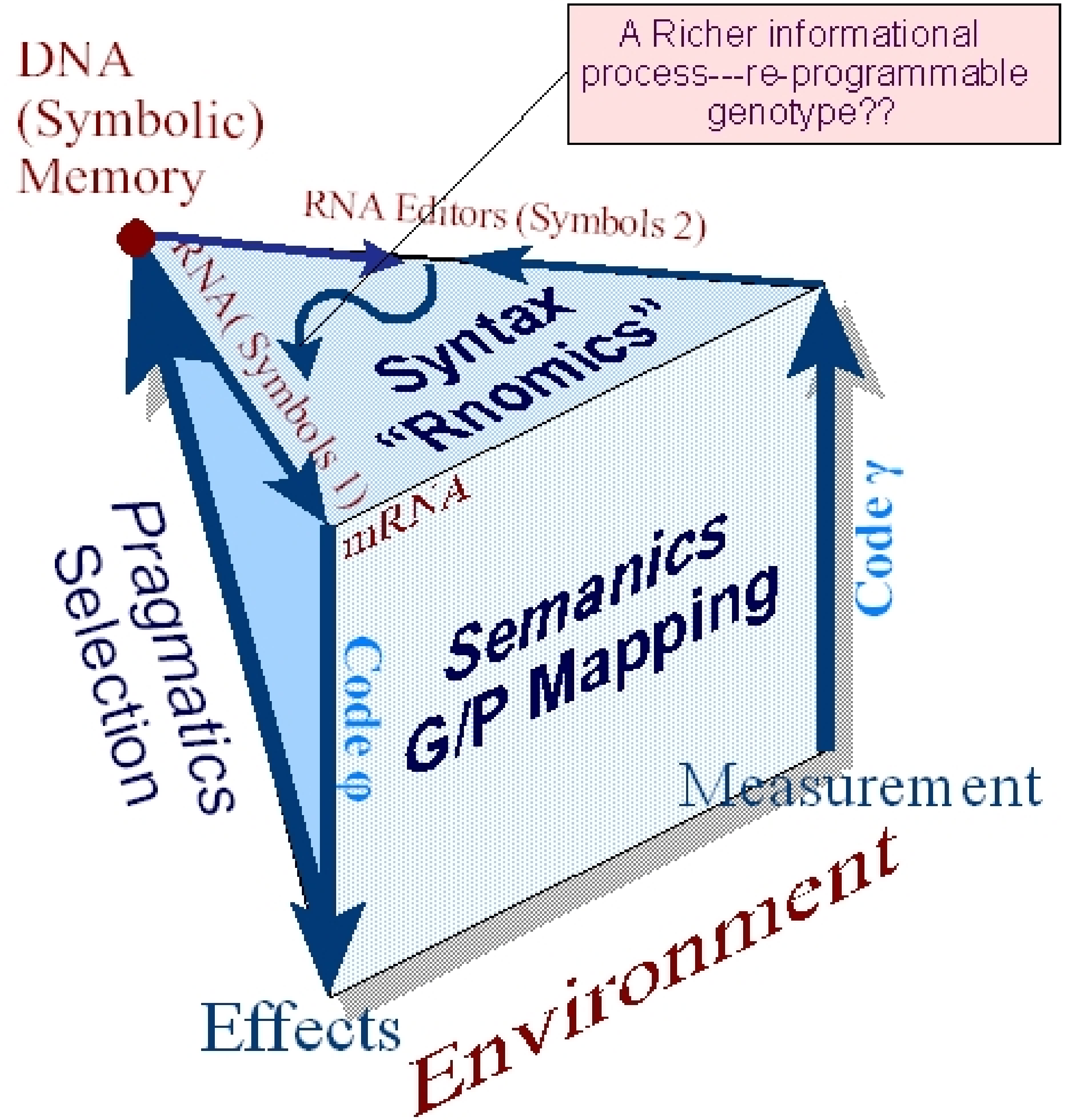
informatics
luis rocha 2006

RNA editing modulates gene expression

genes may encode different proteins depending on environment



- Only mutations that occur during DNA replication can become permanent and heritable
- RNA Editing may produce different mRNA's (and thus proteins), but editions are not inherited.
  - What is inheritable, and subjected to variation, is the genetic material (both coding and non-coding) which is ultimately selected and transmitted to the offspring of the organism



rocha@indiana.edu
http://informatics.indiana.edu/rocha

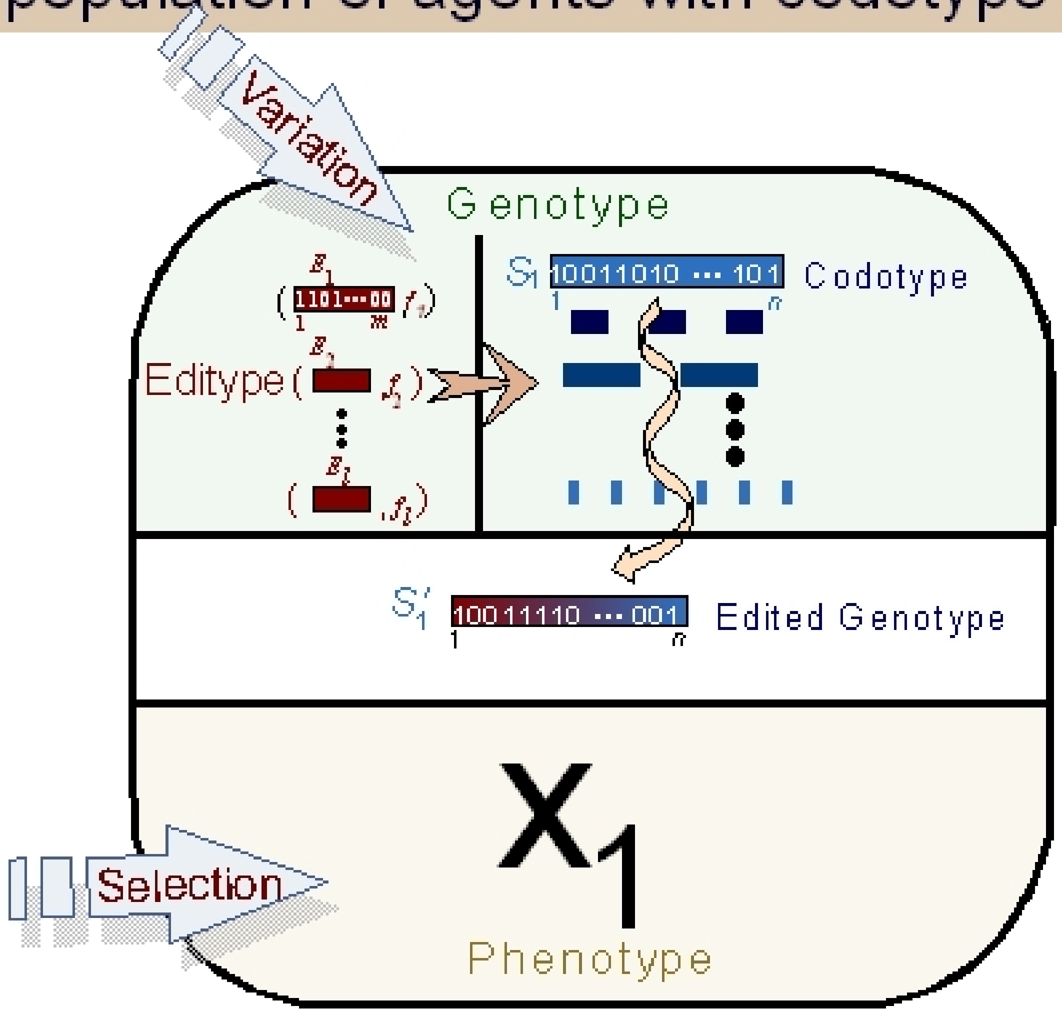
Rocha, Luis M. [1995]. ***ECAL95*** *LNCS/* **929** . pp. 368-382.

INDIANA UNIVERSITY

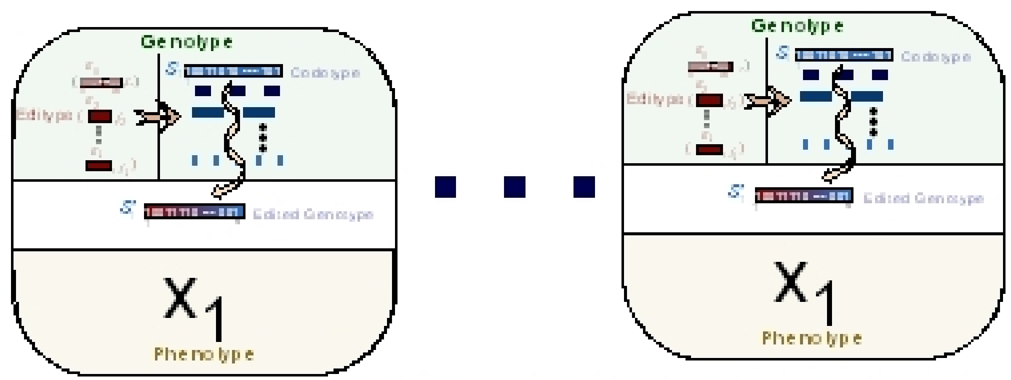
informatics
luis rocha 2006

agent-based model of genotype editing

population of agents with codotype and editype



- Genome contains both coding and non-coding portions
  - Codome and Editome (Editosome)
- For each agent
  - Codotype edited by editype before "translation"
- Modeling pre-translation information (syntactic) processes
  - no RNA/DNA distinction
  - a process of ***non-inheritable alteration of genotypes via edition***, not any specific type of RNA Editing.
  - Not mutation
- co-evolution of editype and codotype
  - Not in the EC sense of independenp populations
  - Independent variation



rocha@indiana.edu
http://informatics.indiana.edu/rocha

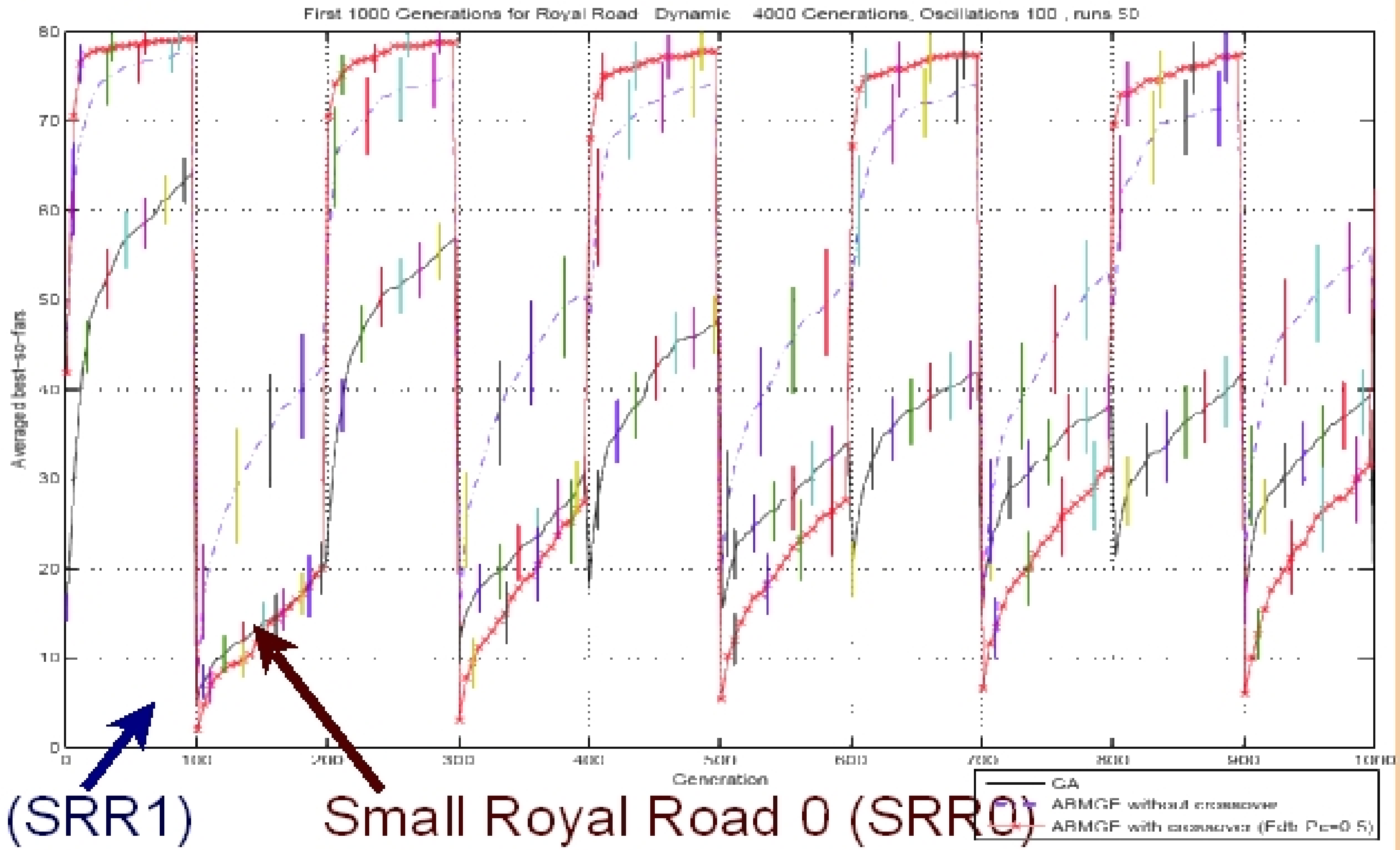
Rocha, L.M. [2006]. *Alife X*. MIT Press. pp. 105-111

INDIANA UNIVERSITY

oscillatory royal road function

dramatic environmental changes

- Oscillation period
  - **100** (50, 200) generations
- First 1000 generations
  - Same parameters as in static case



Small Royal Road 1 (SRR1)

s_1 = 11111***********************************; $c_1 = 10$
s_2 = *****11111******************************; $c_2 = 10$
s_3 = **********11111*************************; $c_3 = 10$
s_4 = ***************11111********************; $c_4 = 10$
s_5 = ********************11111***************; $c_5 = 10$
s_6 = *************************11111**********; $c_6 = 10$
s_7 = ******************************11111*****; $c_7 = 10$
s_8 = ***********************************11111; $c_8 = 10$

Small Royal Road 0 (SRR0)

s_1 = 00000***********************************; $c_1 = 10$
s_2 = *****00000******************************; $c_2 = 10$
s_3 = **********00000*************************; $c_3 = 10$
s_4 = ***************00000********************; $c_4 = 10$
s_5 = ********************00000***************; $c_5 = 10$
s_6 = *************************00000**********; $c_6 = 10$
s_7 = ******************************00000*****; $c_7 = 10$
s_8 = ***********************************00000 $c_8 = 10$

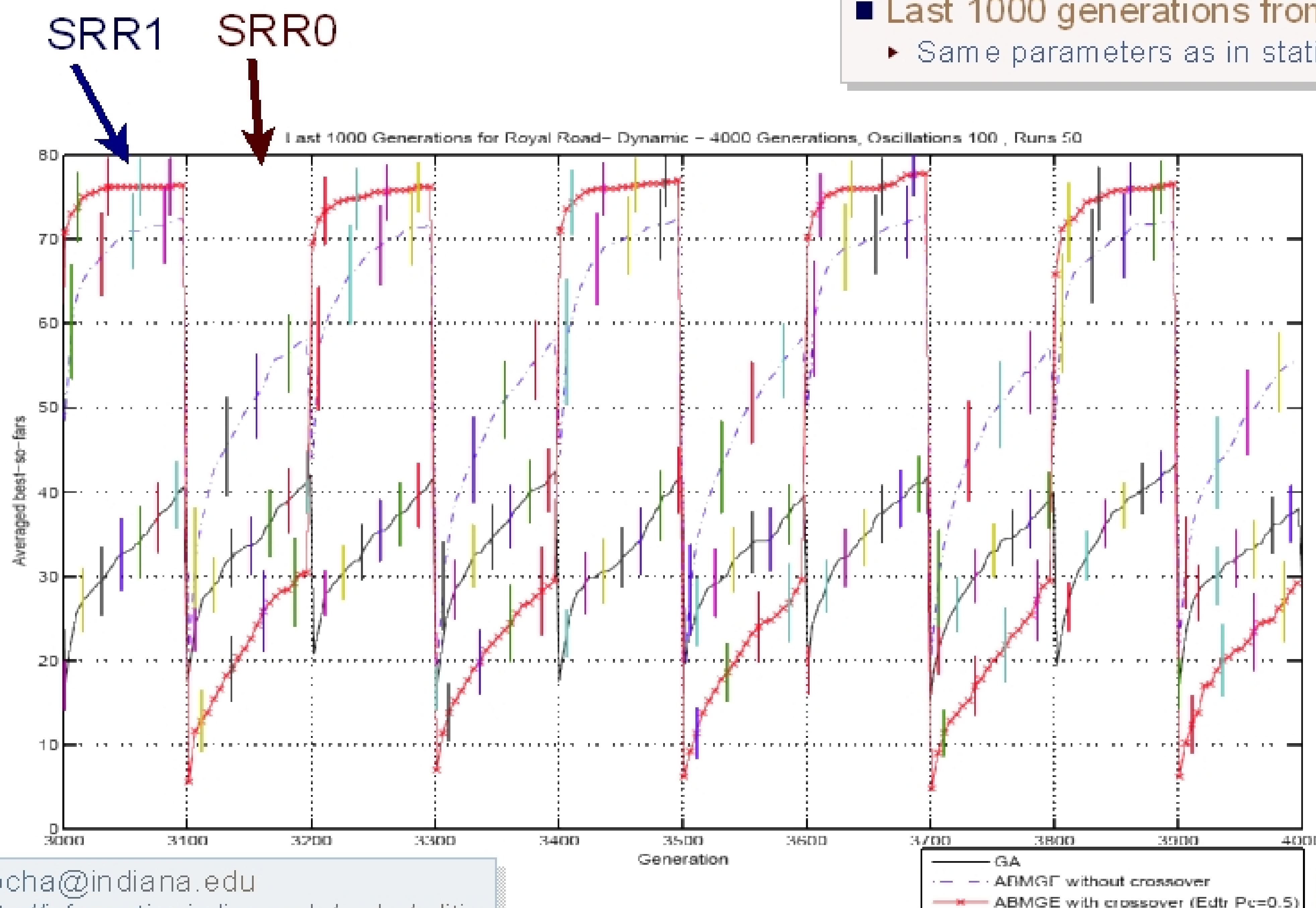
Rocha, L.M. [2006]. *Alife X*. MIT Press. pp. 105-111

informatics
luis rocha 2006
rocha@indiana.edu
http://informatics.indiana.edu/rocha
INDIANA UNIVERSITY

oscillatory royal road function

dramatic environmental changes

- Oscillation period
  - 100 generations
- Last 1000 generations fromm 4000
  - Same parameters as in static case



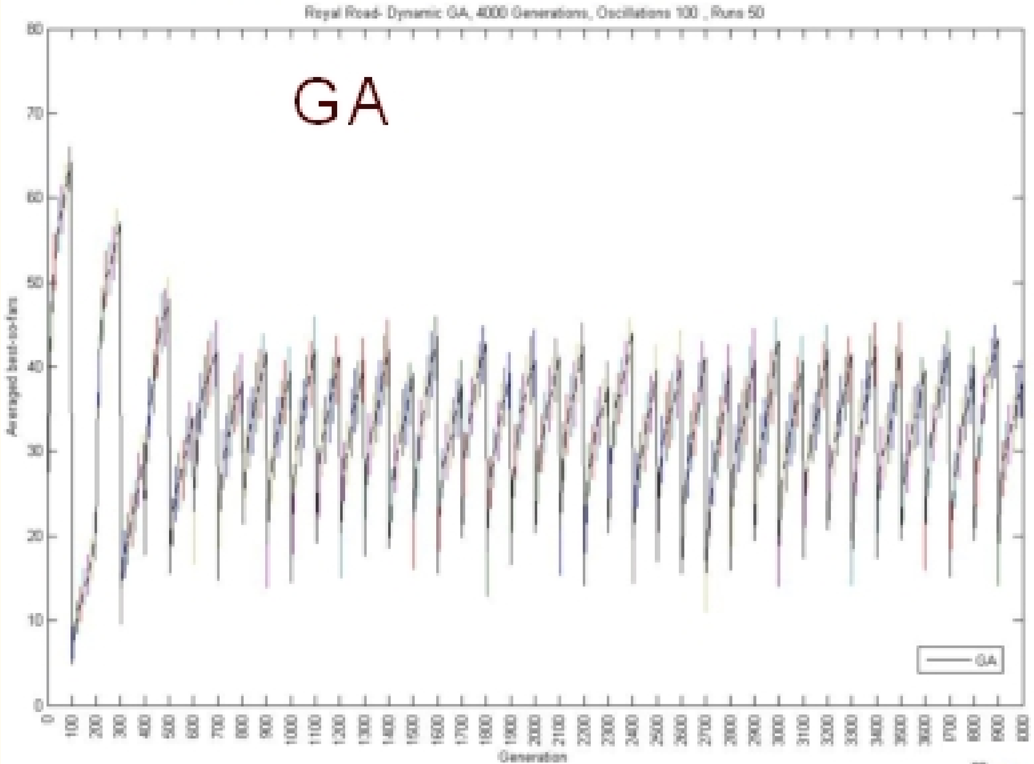
informatics
luis rocha 2006

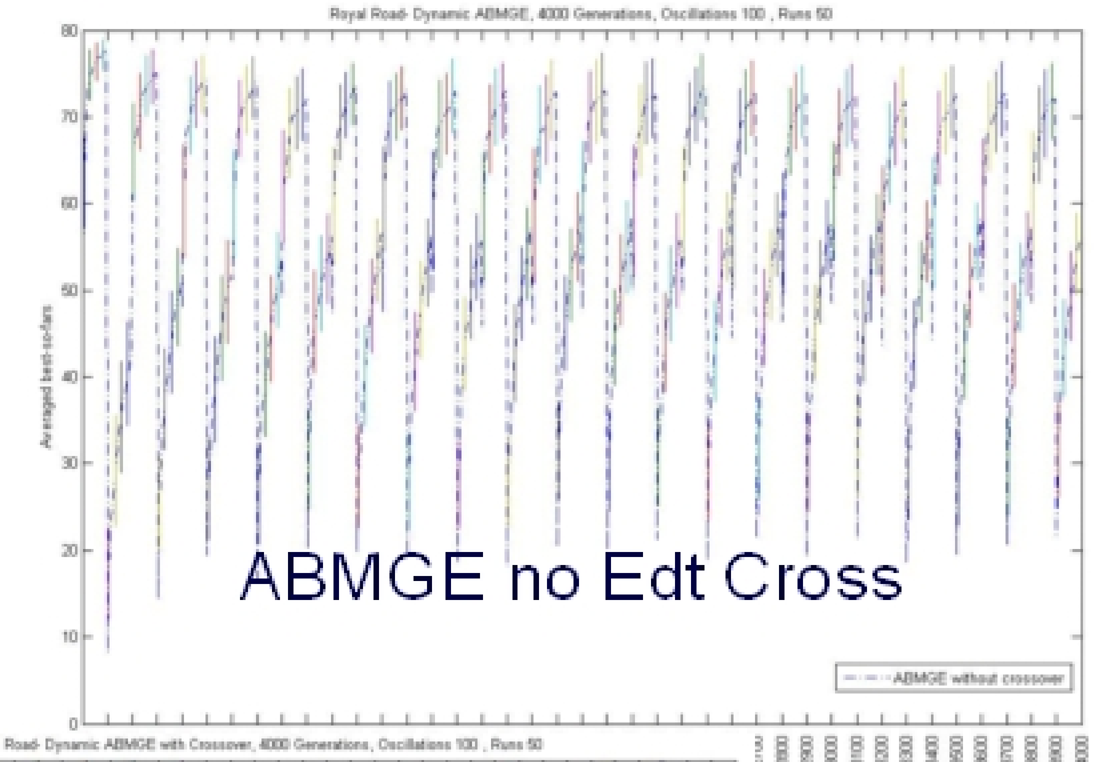
rocha@indiana.edu
http://informatics.indiana.edu/rocha/editing

INDIANA UNIVERSITY

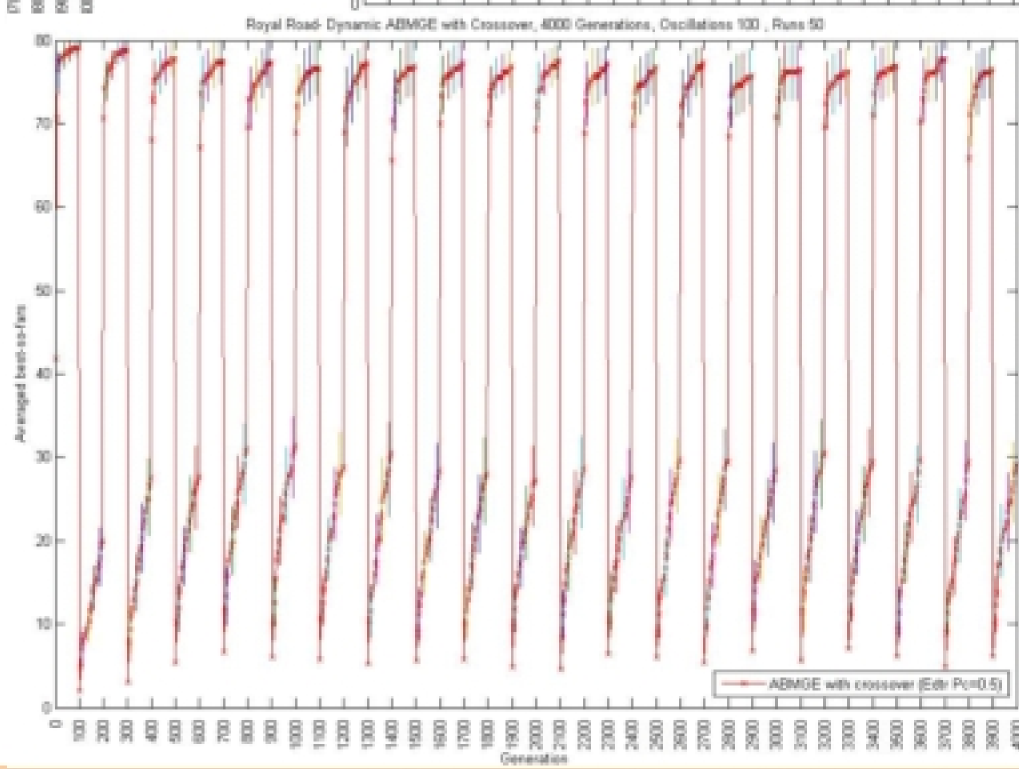
Oscillatory royal road

behavior of 3 algorithms





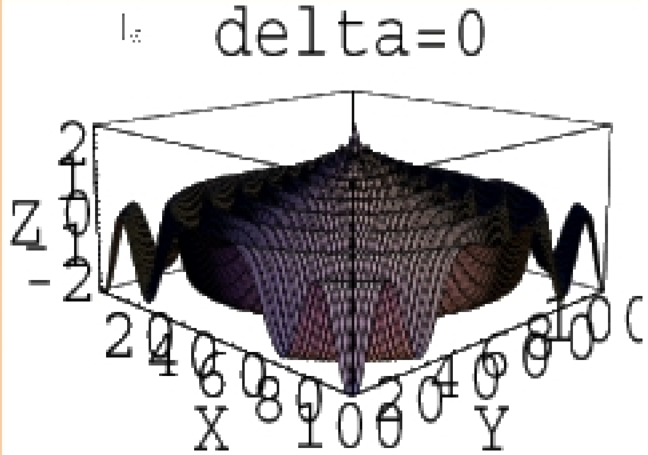
ABMGE

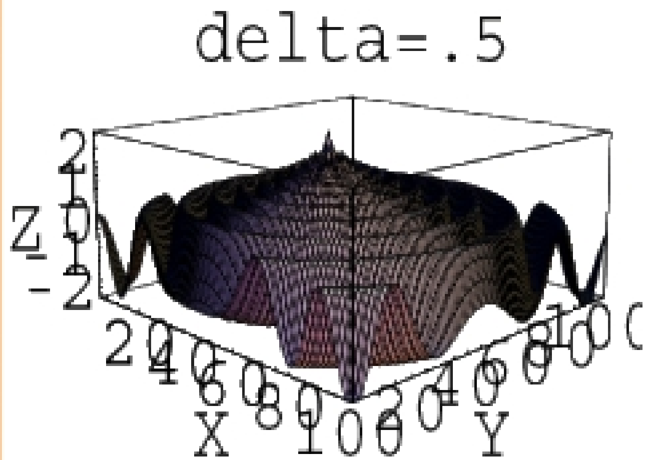


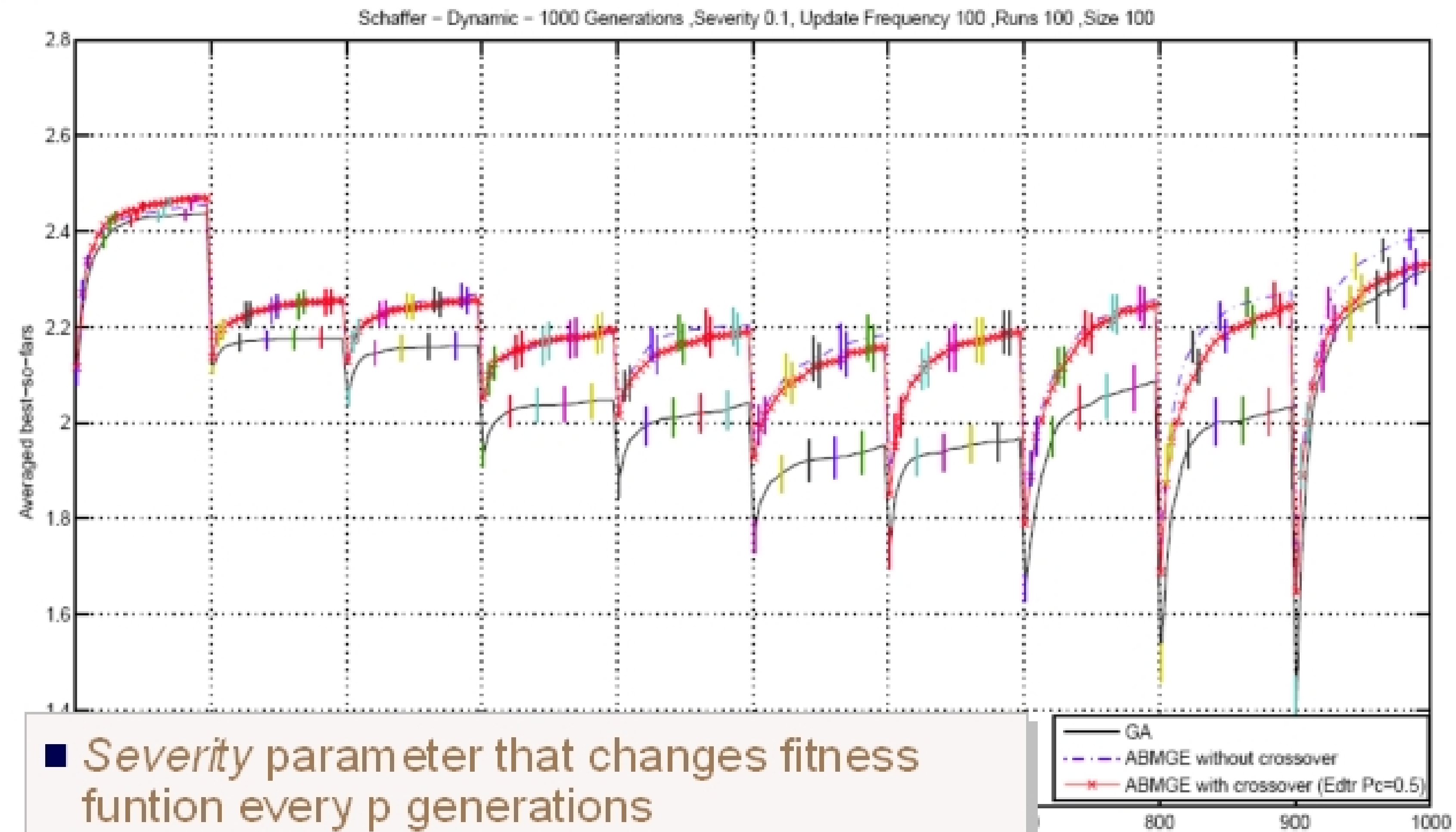
rocha@indiana.edu
http://informatics.indiana.edu/rocha/editing

informatics
luis rocha 2006

dynamic Shaffer function







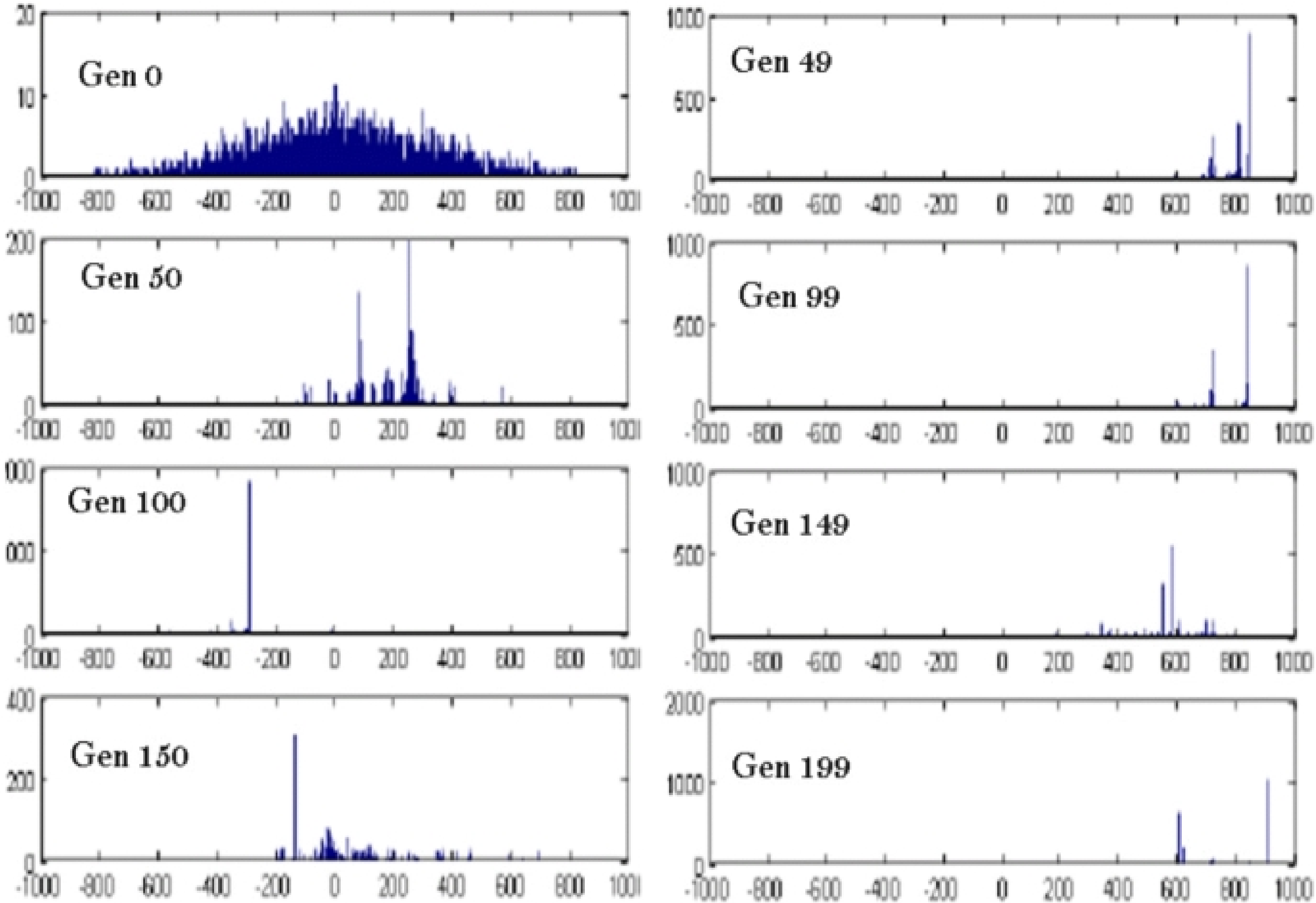
- *Severity* parameter that changes fitness funtion every p generations
  - **100** (50, 200)
  - Linear and **jumping** dynamics
- 50 agents
  - 1000 generations
  - 100 runs

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Rocha, L.M. [2006]. ***Alife X***. MIT Press. pp. 105-111

100 run fitness distribution

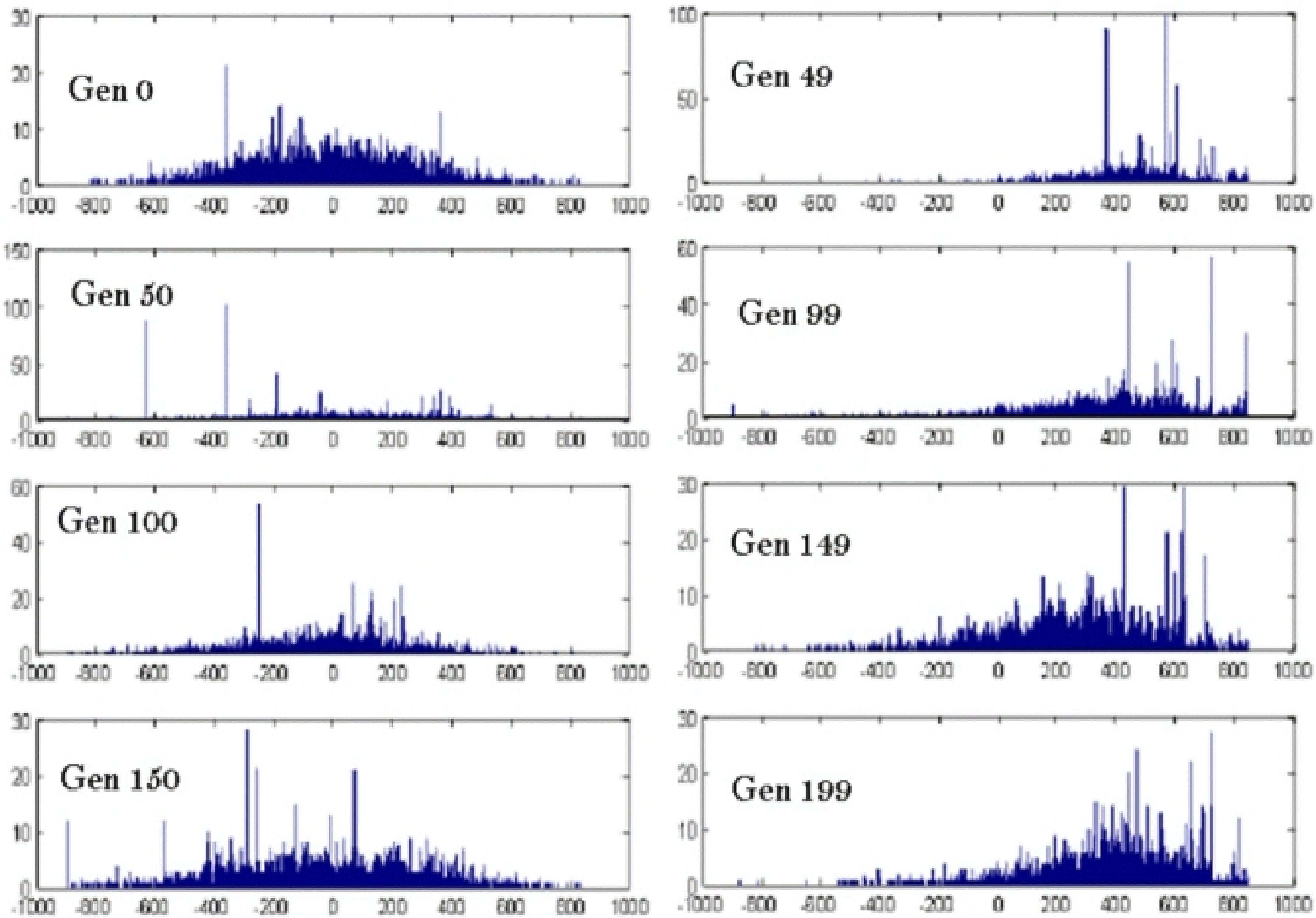
dynamic Schwafel function example (GA)



rocha@indiana.edu
http://informatics.indiana.edu/rocha

100 run fitness distribution

dynamic Schwafel function example (ABMGE)

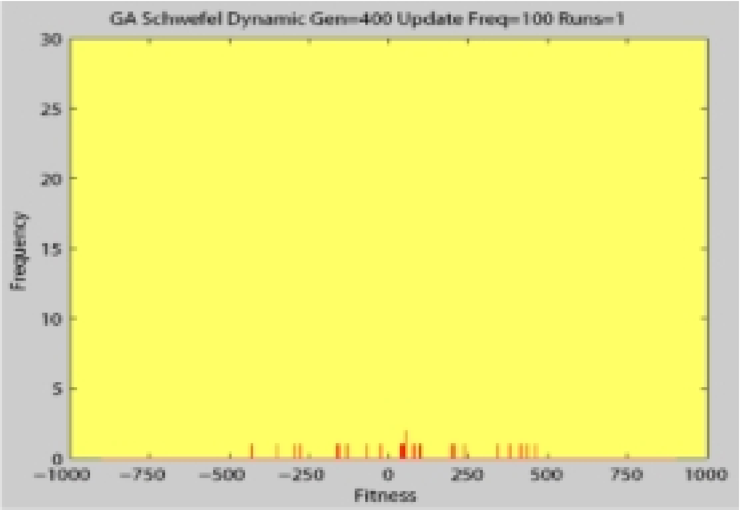


rocha@indiana.edu
http://informatics.indiana.edu/rocha

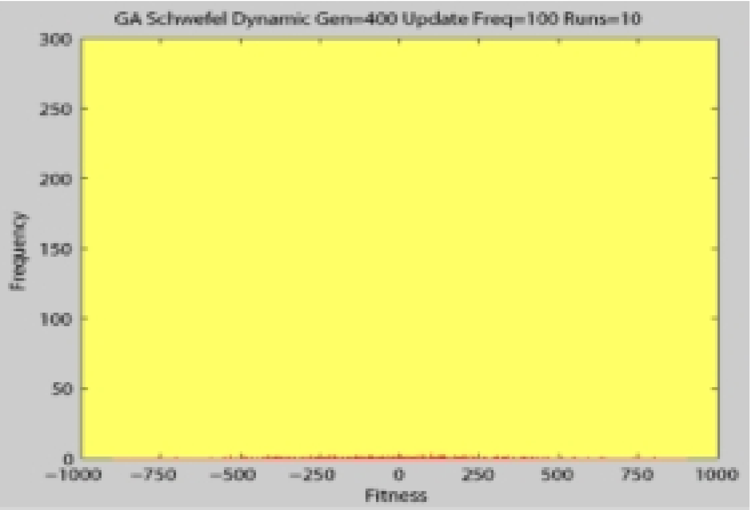
exploration and exploitation with genotype editing

dynamic Schwafel function fitness distribution videos

informatics
luis rocha 2006

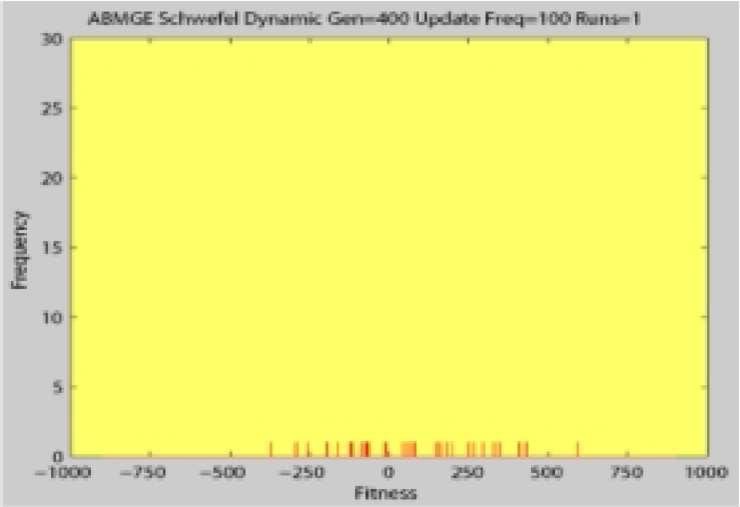


GA

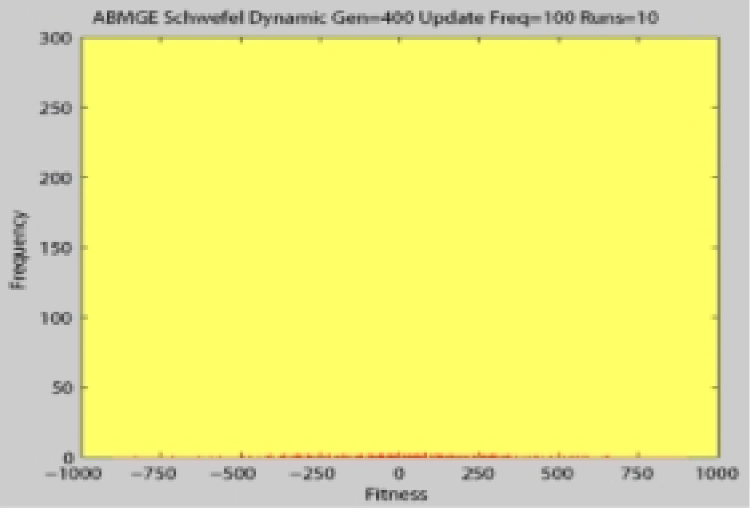


1 run

10 run



ABMGA



rocha@indiana.edu
http://informatics.indiana.edu/rocha/editing

INDIANA UNIVERSITY

the gap between experimental reductionism vs. Systems view

The only consensus found among biologists about their subject is that biological systems are complicated, by any criterion of complexity that one may care to specify. [Rosen, 1972]

- Biology must simplify organisms to study them – some type of abstraction or modeling is needed.
  - External (Functional) description (favored by Systems Thinking)
    - *Blackbox*, input-output behavior of observables
    - Tells us what the system does
    - Function depends on repercussions in an environment
  - Internal (structural) description (favored by Experimentalists)
    - State description, trajectory behavior
    - Tells us how the system does what it does
    - Structural information can be measured for any component
  - Ideally, we would like to move between the two descriptions
    - But in Biology, the structural states we can measure, are not obviously related to the observed functional activities (and vice versa).
    - Thus, *Systems Biology* has mostly been relegated to deal with evolutionary problems, and Experimental Biology to increase our knowledge of the molecular components of organisms

informatics
luis rocha 2006
rocha@indiana.edu
http://informatics.indiana.edu/rocha

Why Structural Reductionism is Not Sufficient

Destruction of Dynamical Properties

- **Naive Structural Decomposition**
  - Breaks an organism into simpler components, gathers information about those, and attempts to assemble information about the organism from the components
  - But some properties of the original system cannot be reconstructed from components
    - E.g. the crucial stability properties of 3-body system cannot be reconstructed from knowledge of 2-body or 1-body constituents – the dynamics is destroyed.
    - Think what this means for the methodologies of molecular biology!

http://faculty.ifmo.ru/butikov/Projects/Collection2.html
http://www.freewebz.com/vitaliy/triApplet/triGrav.html
http://www.dynamical-systems.org/threebody

rocha@indiana.edu
http://informatics.indiana.edu/rocha

How To Close the Gap

Coupling Structural Data with Functional Decomposition

- Biological Systems require “function-preserving” and “dynamics-preserving” Decompositions
  - In biology, the same physical structure typically is simultaneously involved in several functional activities
    - E.g. unlike airplanes, birds use the same structure (wing) as both propeller and airfoil
  - We must allow the simplifying decompositions to be dictated by system dynamics
    - Iterative Design of Experiments from Knowledge of Dynamics
    - Data accumulated from experiments based on naive structural decompositions are simply the first iteration!
  - Search for Global Patterns and Juxtaposed Functional Modes
    - E.g. studying global patterns of antigens rather than specific molecular interactions [Coutinho et al]
    - Spectral, PCA-like, Fourrier Analysis approaches
  - Build IntegrativeTechnology to Disseminate and Utilize Structural Data – for a diverse group of scientists

Integrative Link for bridging Experimental and Systems Biology

- Genome Informatics initially as enabling technology for the genome projects
  - Support for experimental projects
  - Genome projects as the ultimate reductionism: search and characterization of the function of information building blocks (genes)
- Post-genome informatics [Kanehisa 2000] aims at the synthesis of biological knowledge from genomic information
  - Towards an understanding of basic principles of life (while developing biomedical applications) via the search and characterization of *networks* of building blocks (genes and molecules)
    - The genome contains information about building blocks but, given the knowledge of Systems Biology, it is naive to assume that it also contains the information on how the building blocks relate, develop, and evolve.
  - Interdisciplinary: biology, computer science, mathematics, and physics

"Life is a complex system for information storage and processing". Minoru Kanehisa [2000]

Enabling a Systems Approach to Biology

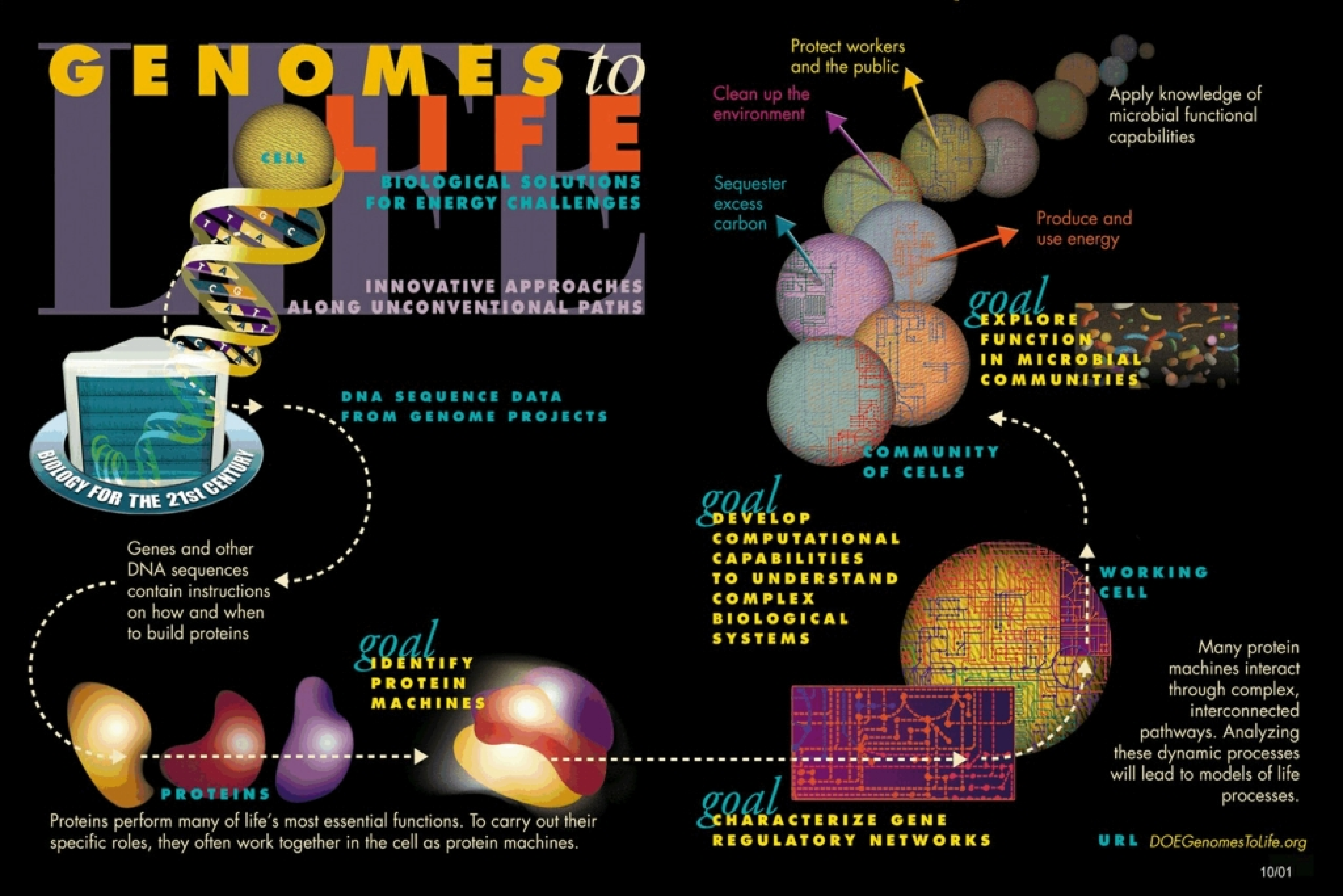
- Not just support technology but involvement in the systematic design and analysis of experiments
  - *Functional genomics*
  - Where, when, how, and why of gene expression
  - *Post-genome informatics* aims to understand biology at the molecular network level using all sources of data: sequence, expression, diversity, etc.
  - Cybernetics, Systems Theory, Artificial Life, Complex Systems approach to Theoretical Biology
  - *Synthetic Biology*: to engineer novel life forms and bio-technology
- Grand Challenge of Computational Biology
  - Given a complete genome sequence, reconstruct in a computer the functioning of a biological organism
    - Regards Genome more as set of initial conditions for a dynamic system, not as complete blueprint (Pattee, Rosen, Atlan). The genome can be contextual and dynamically accessed and even modified by the complete network of reactions in the cell (e.g. editing).
    - Uses additional knowledge for integration comparative analysis: Comparative Biology
- Grand Challenge of Sythetic Biology
  - If we understand it, we can build it!
    - intentional design of real biological systems
    - reversal of aging and innovative medical treatments such as beneficial bacterial infections programmed to augment immunity

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Post-Genome Informatics or the "New" Systems Biology

- *Systems biology* is a unique approach to the study of genes and proteins which has only recently been made possible by rapid advances in computer technology. Unlike traditional science which examines single genes or proteins, systems biology studies the complex interaction of all levels of biological information: genomic DNA, mRNA, proteins, functional proteins, informational pathways and informational networks to understand how they work together. Systems biology embraces the view that most interesting human organism traits such as immunity, development and even diseases such as cancer arise from the operation of complex biological systems or networks.
  - Institute for Systems Biology: http://www.systemsbiology.org
  - Kitano Symbiotic Systems Project: http://www.symbio.jst.go.jp/
- The "New" Systems Biology is not novel per se, it is rather a result of new enabling technology for doing "Old" Systems Biology
  - But it is finally allowing experimentalists to work with theorists.

rocha@indiana.edu
http://informatics.indiana.edu/rocha


GENOMES to LIFE
CELL
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS
BIOLOGY FOR THE 21st CENTURY
DNA SEQUENCE DATA FROM GENOME PROJECTS
Genes and other DNA sequences contain instructions on how and when to build proteins
goal IDENTIFY PROTEIN MACHINES
PROTEINS
Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.
goal CHARACTERIZE GENE REGULATORY NETWORKS
goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS
WORKING CELL
Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.
COMMUNITY OF CELLS
goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES
Sequester excess carbon
Clean up the environment
Protect workers and the public
Produce and use energy
Apply knowledge of microbial functional capabilities
URL DOEGenomesToLife.org
10/01

informatics
luis rocha 2006

Needs of Systems Biology

- **Experimental Side**
  - Improving cellular measurement methods
    - High-throughput identification of the components of protein complexes; Parallel, comparative, high-throughput identificationof DNA fragments among microbial communities and for community characterization; Whole-cell imaging including in vivo measurements; Better Separtion techniques.
  - Measurements Based on Functional Decompositions
    - Functional assays? Flexible, fast, novel experimental design based on informatics results.
- **Computational Side**
  - Integrative Technology
    - Standardized formats, databases, and visualization methods
    - Automated collection, integration and analysis of biological data
    - Algorithms for genome assembly and annotation and measurement of protein expression and interactions;
  - Simulation Technology
    - Improved methods for distributed simulation, analysis, and visualization of complex biological pathways;
    - Prediction of emergent functional capabilities of microbial communities

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

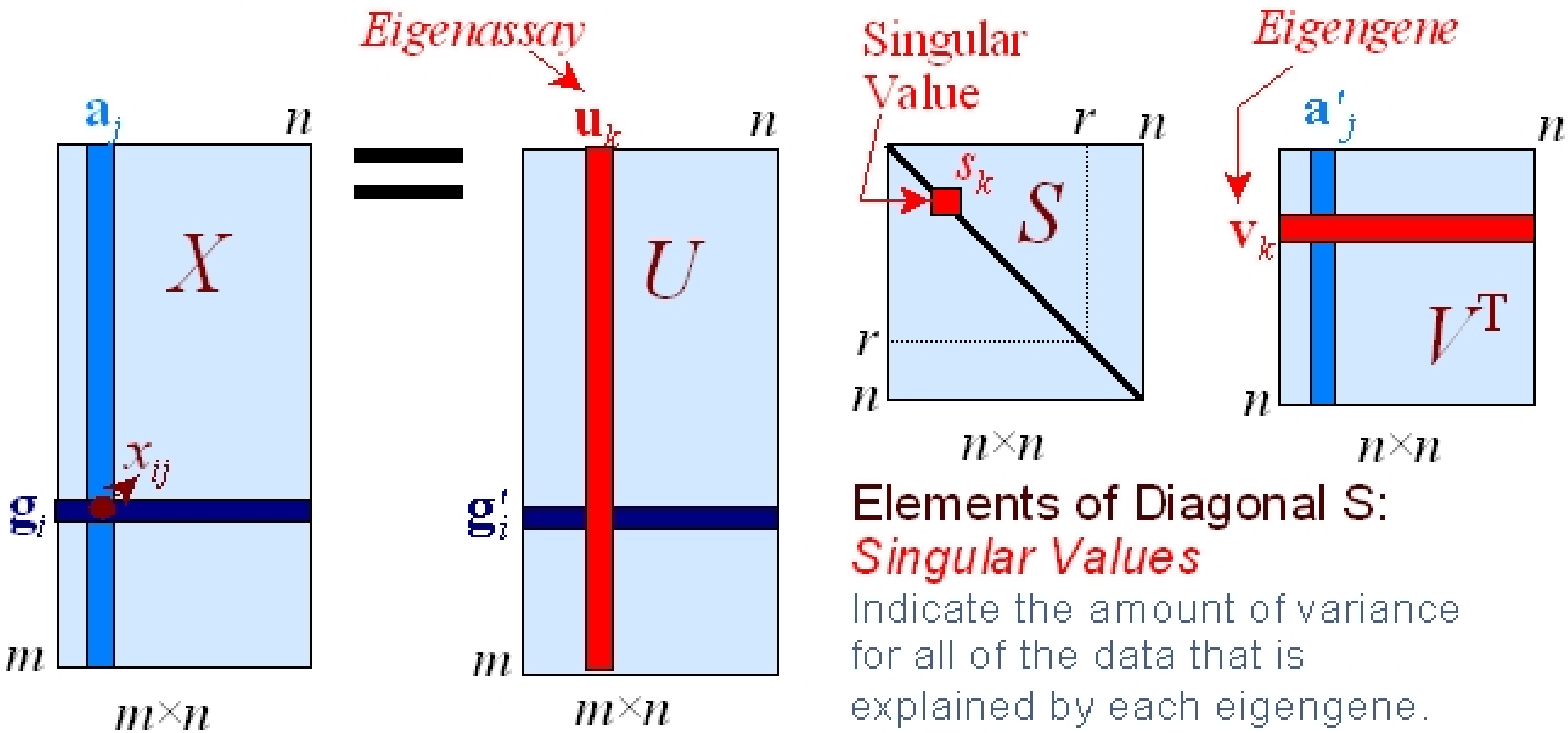
Needs of Systems Biology

Continuation

- Modeling Side
  - Algorithms for Discovery of Global Patterns and Juxtaposed Functional Modes
    - Pattern Recognition, data-mining, "Spectral" methods.
  - Network Models and Analysis
    - Predictive Models based on biochemical pathways of observed networks
    - Simplification Strategies for Network Modeling
    - Reduction of possible cell-behaviors from steady-state models of metabolic network models
    - High-Perforemance Algorithms to allow whole-system Kinetic models

singular value decomposition

for microarray analysis



Rows of V^T: *eigengenes* (colums are time steps)

Each gene's expression pattern is a linear combination of the eigengene patterns.

$$\mathbf{g}_i = \sum_{k=1}^{r} u_{ik} s_k \mathbf{v}_k, \quad i{:}1,\ldots,m$$

Elements of Diagonal S: *Singular Values*

Indicate the amount of variance for all of the data that is explained by each eigengene.

$$X = USV^T$$

Gene Expression Matrix: Columns are assays (time steps) and rows are genes

Columns of U: *eigenassays* (rows are genes) **describe how each component contributes to a single gene's expresssion pattern**

$$\mathbf{a}_j = \sum_{k=1}^{r} v_{jk} s_k \mathbf{u}_k, \quad j{:}1,\ldots,n$$

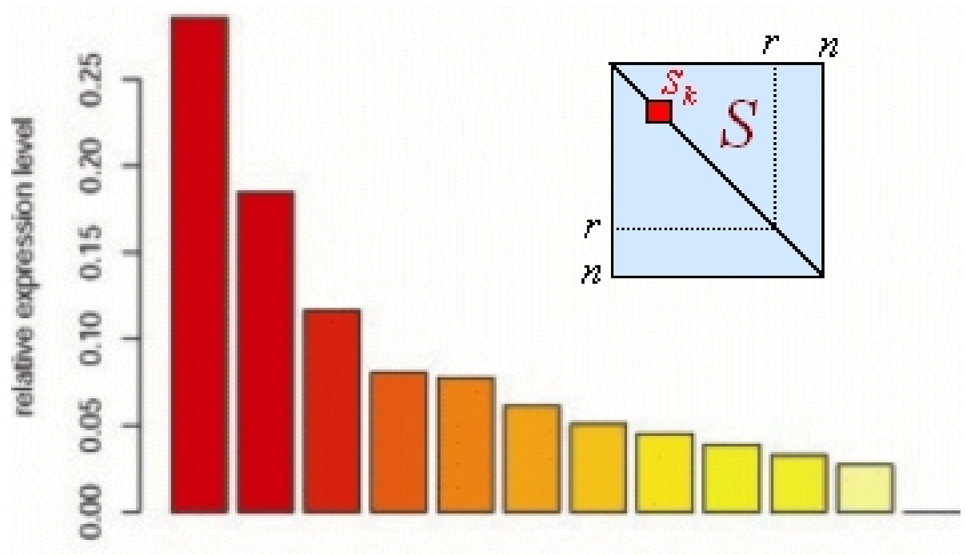
Wall, Rechtsteiner and Rocha [2002]. "Singular value decomposition and principal component analysis". In *Understanding and Using Microarray Analysis Techniques: A Practical Guide*. D.P. Berrar, W. Dubitzky, M. Granzow, eds.

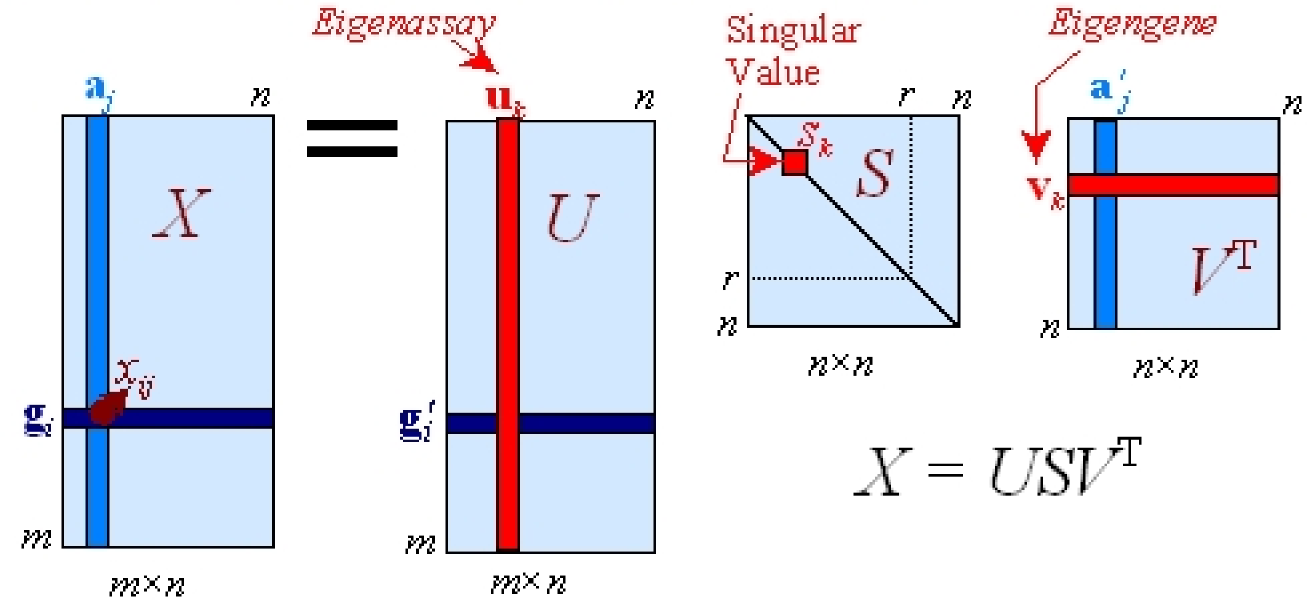
rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2006

singular value decomposition

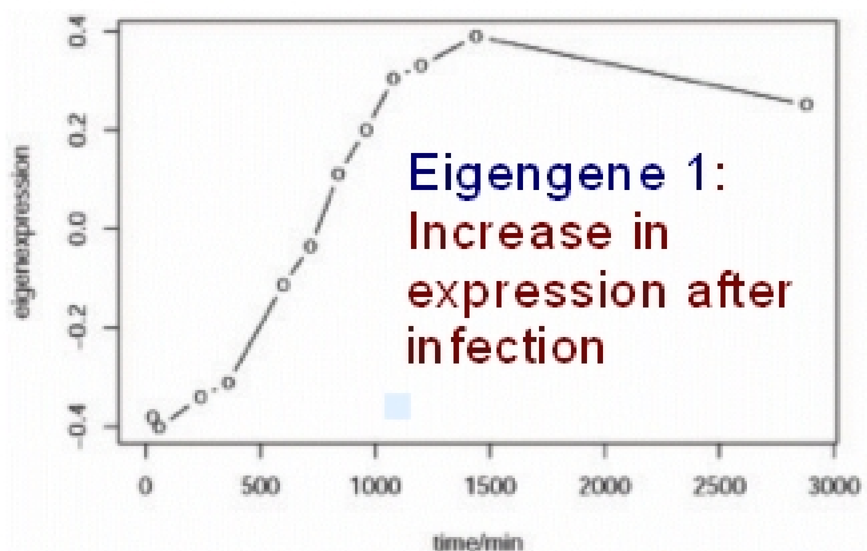
gene expression (13000 genes) after infection with herpes virus

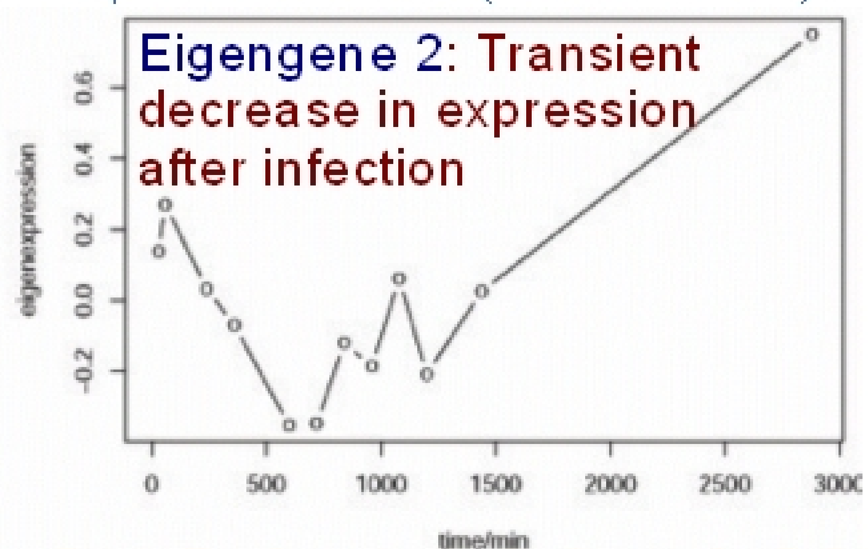


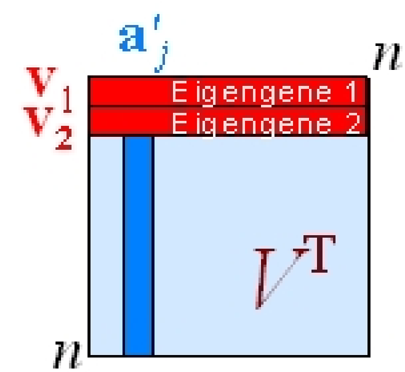


$$X = USV^T$$

12 point time series (30min - 48hrs)







Challacombe, J., A. Rechtsteiner, G. Gottardo, L.M. Rocha, E.P. Brown, T. Shenk, M. Altherr, T. Brettin [2004]. "Evaluation of the host transcriptional response to human cytomegalovirus infection". *Physiol. Genomics*. **10**.1152

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

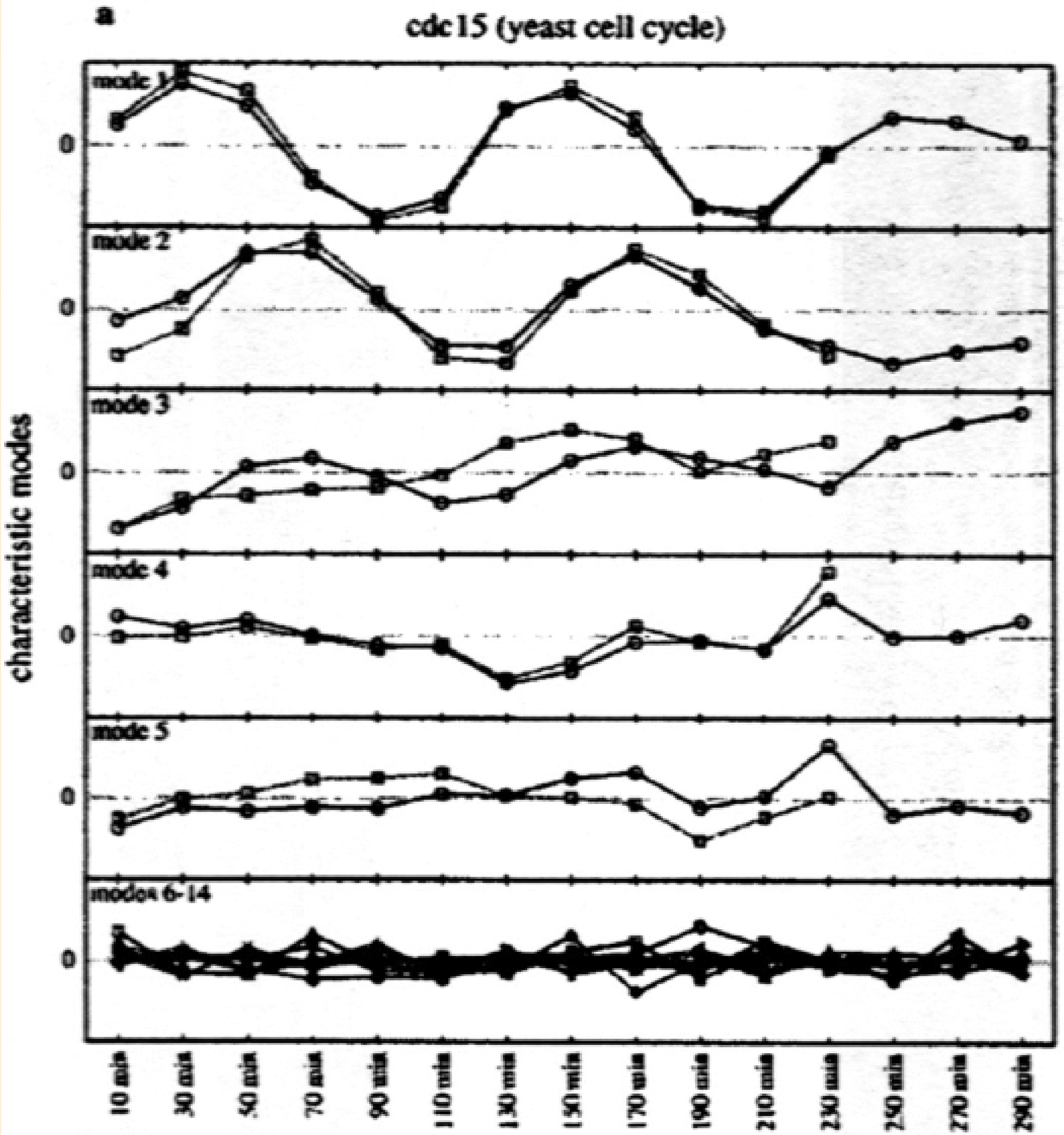
Data-Mining of Global Patterns

Discovery of Juxtaposed Functional Modes

- Gene Expression Modes
  - Cluster analysis provides little insight into inter-relationships among groups of co-regulated genes. Tends to demand separated grupings.
  - Component (“spectral”) analysis yields a description of superposed behavior of gene expression networks, rather than a partition.
    - PCA, SVD, etc.
    - Holter et al [2000] compares the superposed components to the characteristic vibration modes of a violin string which entirely specify the tone produced
  - Holter et al [2000] compared SVD analysis of yeast cdc15 cell-cycle [Spellman et al 1998] and sporulation [Chu et al, 1998] data sets, as well as the data set from serum-treated human fibroblasts [Iyer et al, 1999].
    - Essential temporal behavior is captured by first 2 modes (sine and cosine)
    - Large group of genes with same sinosoidal period but dephased

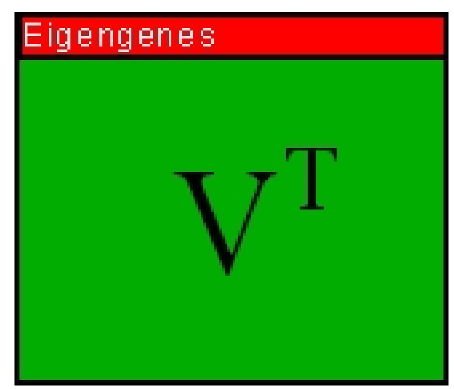
rocha@indiana.edu
http://informatics.indiana.edu/rocha

Holter et al SVD Analysys

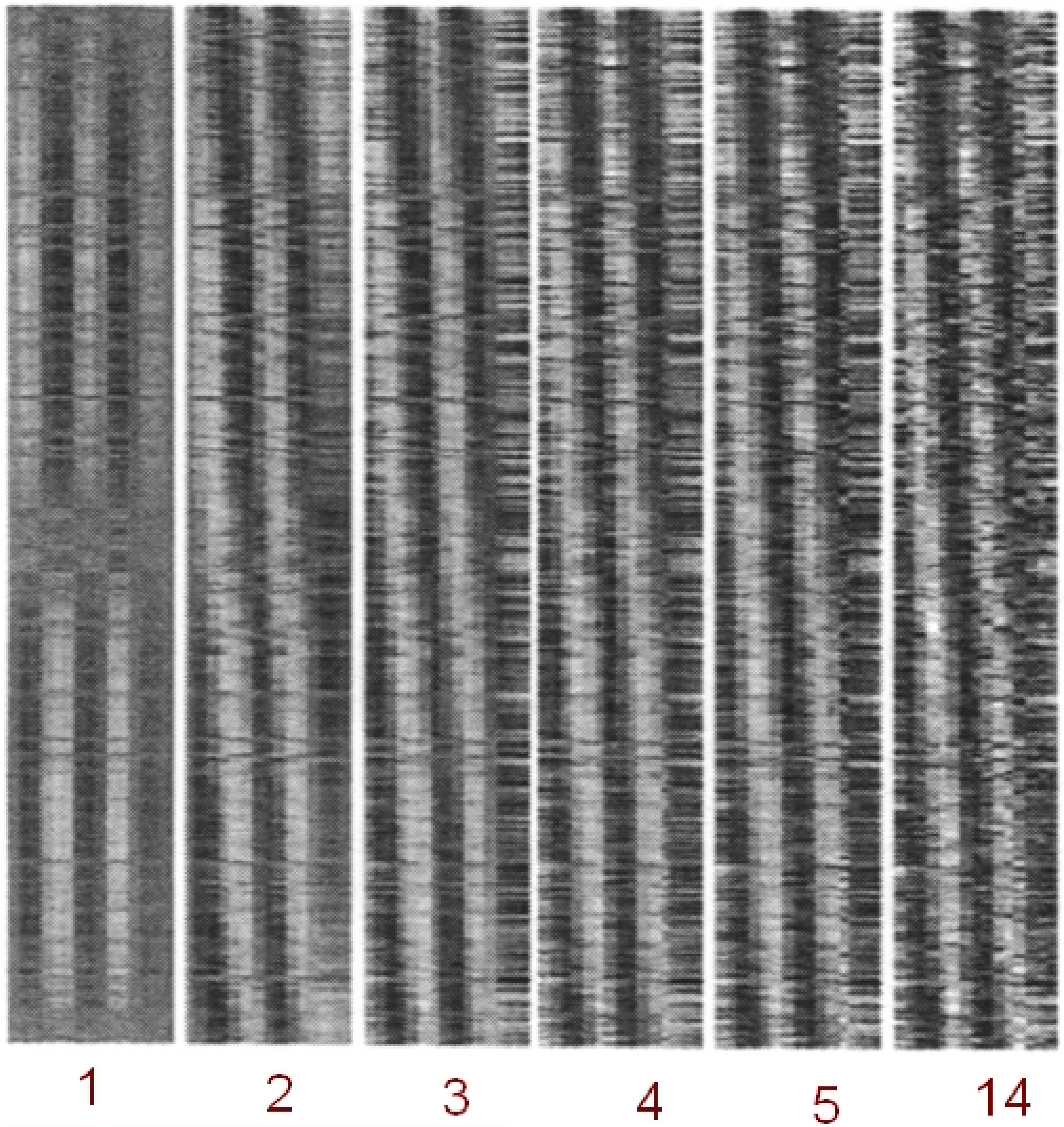


- 800 genes by 15 (12) time measurements
- 2 dominant modes
  - Approximately sinusoidal and out of phase
  - Less synchronized as cell enters 3rd cycle
  - If only 12 points are used, third SV loses relevance, but 2 first components remain largely unchanged

Eigengene: rows of V^T (each column is a time instance)



cdc15 Reconstruction with k-highest modes



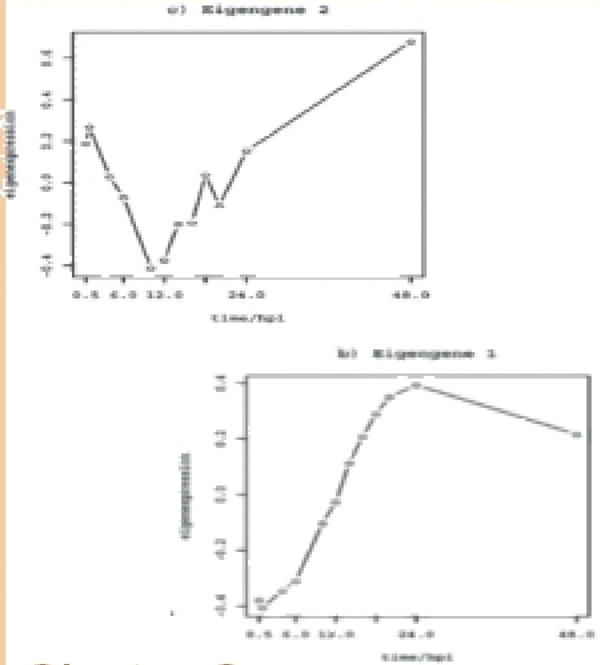
Rows are genes
Columns are time points

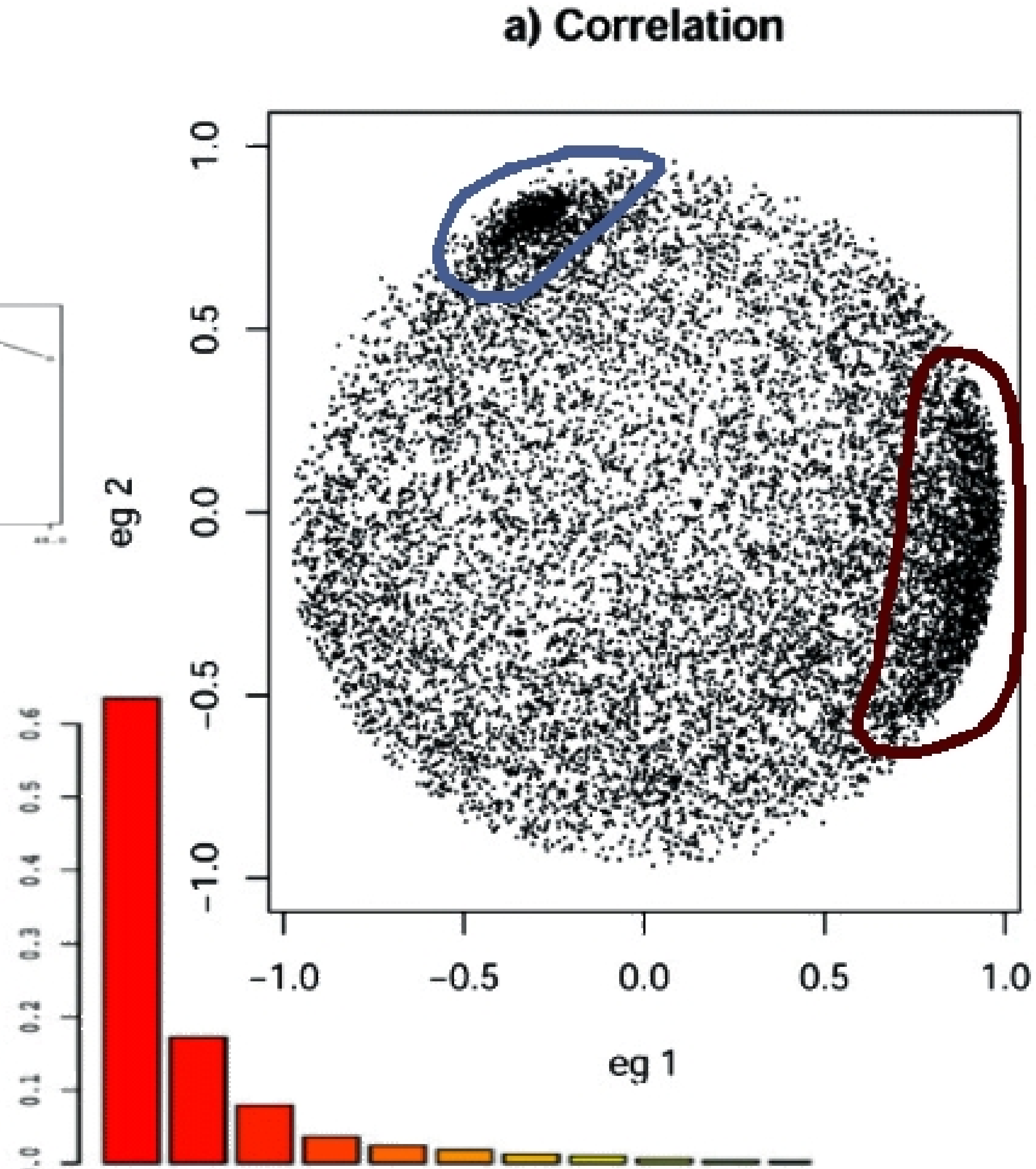
It implies an undelying simplicity in genetic response

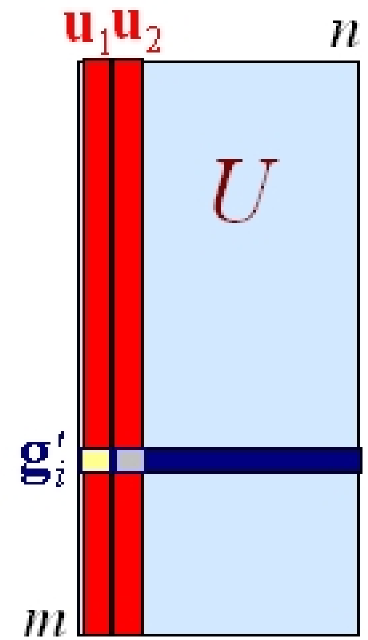
informatics
luis rocha 2006

rocha@indiana.edu
http://informatics.indiana.edu/rocha

eigenassay coefficient plot: human cytomegalovirus infection







Cluster 2:
Genes involved in immune system regulation, signal transduction and cell adhesion. Also mainly in cluster 2, genes targeted by HCMV's immune evasion strategies.

Cluster 1:
Genes involved in ***transcriptional regulation***, oncogenesis and cell cycle regulation. Also mainly in cluster 1, genes involved in the host response to HCMV infection.

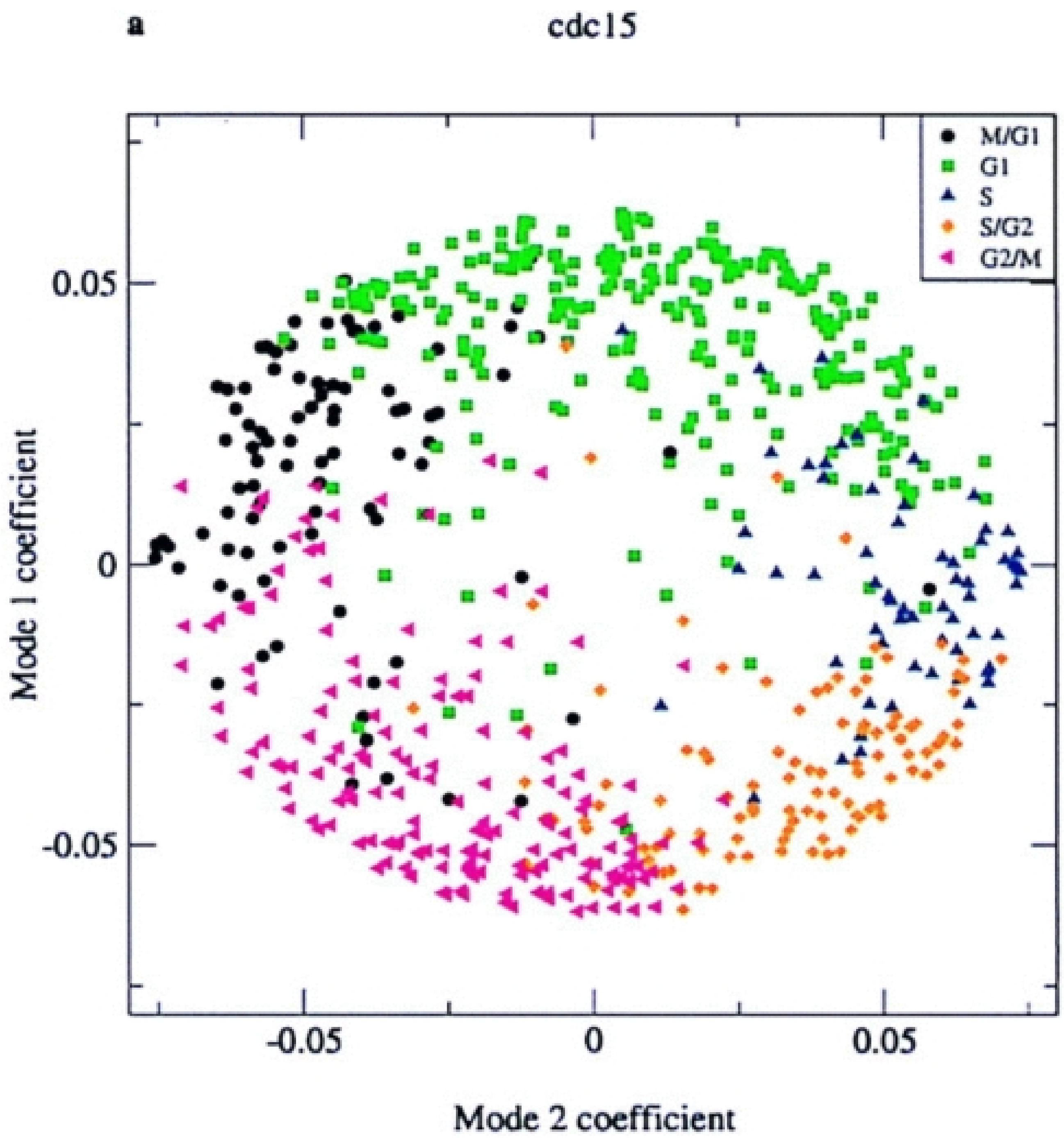
rocha@indiana.edu
http://informatics.indiana.edu/rocha

Challacombe et al. *Physiol. Genomics*. **10**.1152. 2004.

INDIANA UNIVERSITY

Eigenarray Coefficient Plot

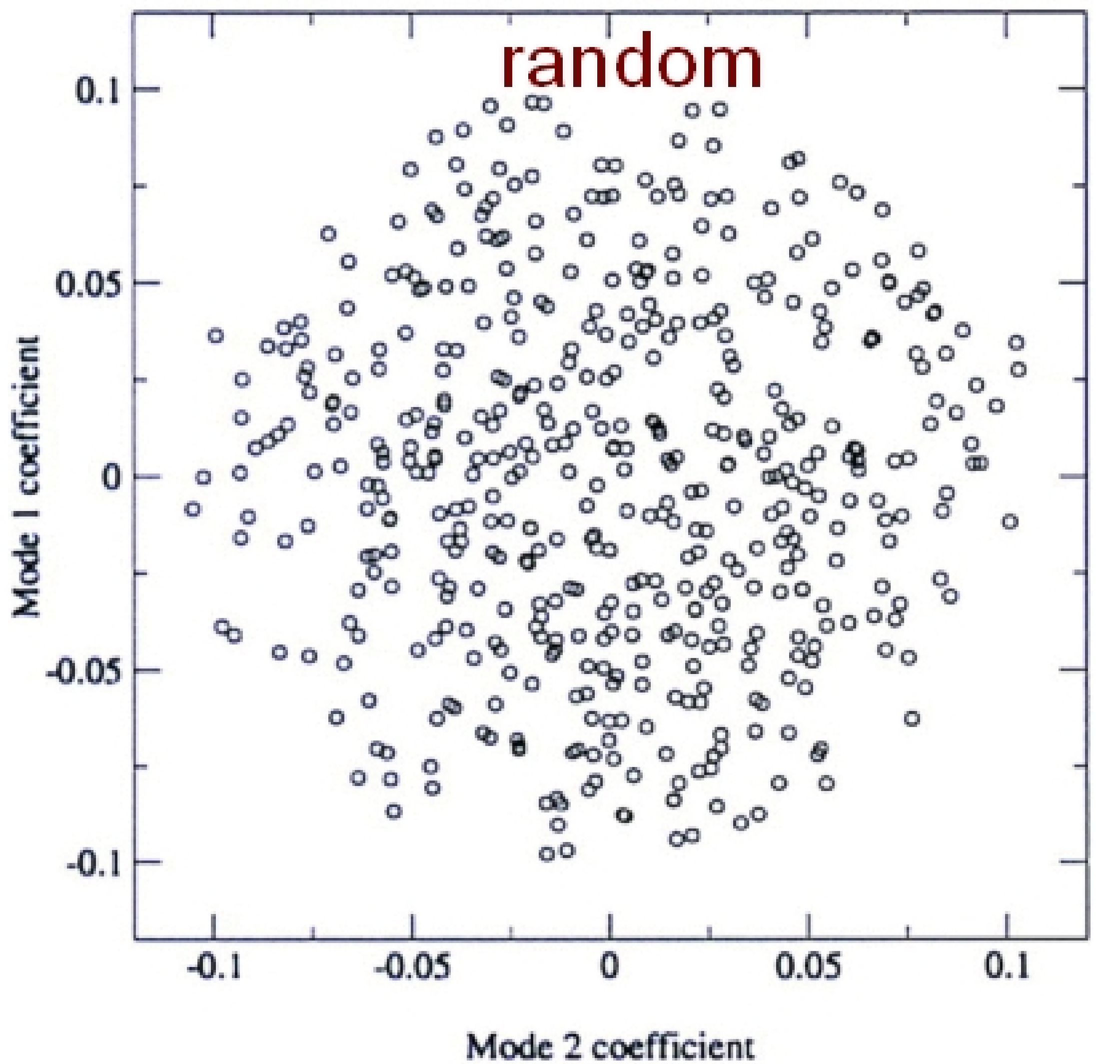
Plot of the coefficients of the first 2 modes for all genes

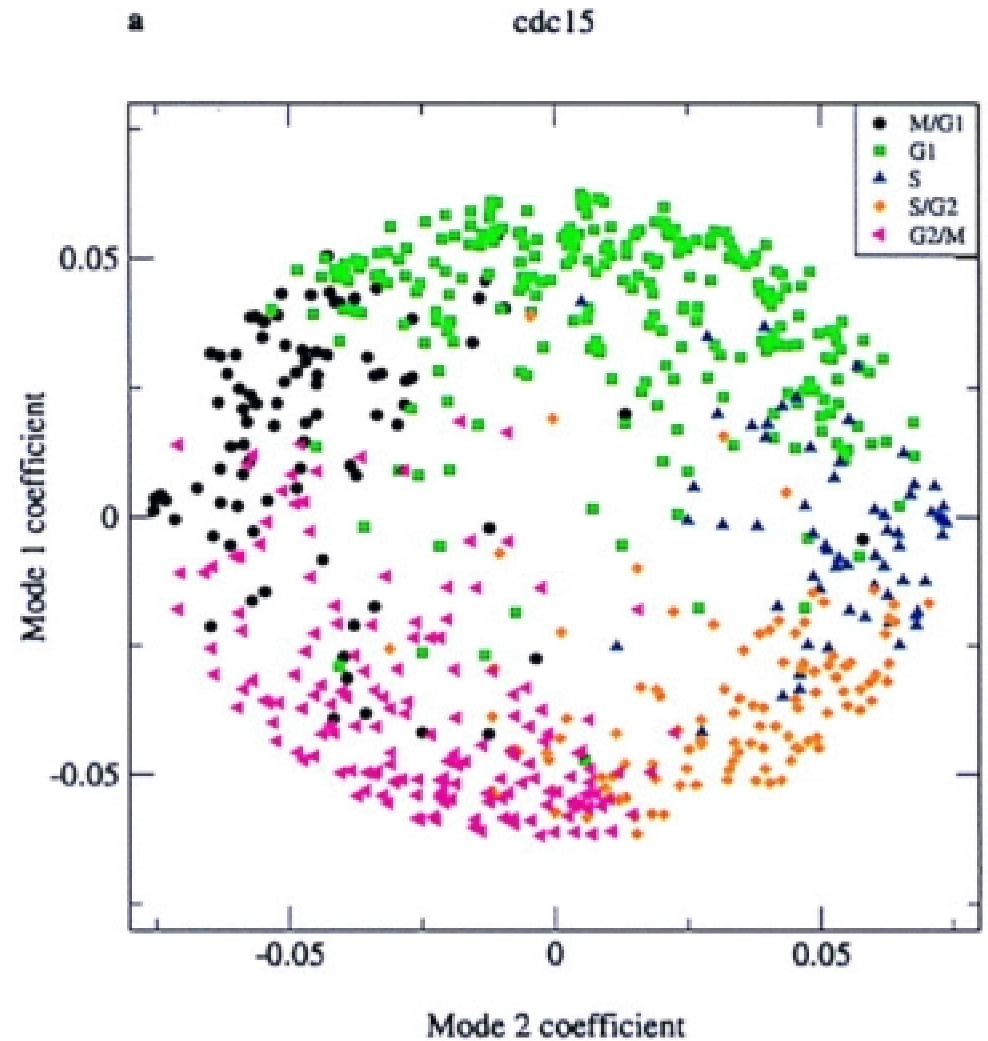


- Clusters of genes by other methods cluster in these plots, but the temporal progression in the cell cycle and in the course of sporulation is more evident in the SVD analysis
- Holter et al conclude that genes are not activated in discrete groups or blocks, as historically implied by the division of the cell cycle into phases or the sporulation response into tempotal groups.There is a continuity in expression change

Eigenarray Coefficient Plot

Random data





Fill most of the plot because genes are not very correlated with components. A circle implies equal contribution from each component (rather than an elipse)

rocha@indiana.edu
http://informatics.indiana.edu/rocha

leading publication and conferences sources

Most common areas

- Journals
  - Bioinformatics
  - BMC Bioinformatics
  - Journal of Theoretical Biology
  - PNAS
  - Biosystems
  - Genome Research
  - IEEE Transactions on Computational Biology and Bioinformatics
- Conferences
  - Intelligent Systems for Molecular Biology (ISMB)
  - Research in Computational and Molecular Biology (RECOMB)
  - Pacific Symposium on Biocomputing (PSB)
- Areas
  - Genome Analysis
  - Sequence Analysis
  - Systems Biology
  - Data and Text Mining
  - Structural Bioinformatics
  - Gene Expression
  - Genetics and Population Analysis

Curriculum For Computational Biology

Graduate Study in Computational Biology: The Gulbenkian CB PhD program

- **Direction**
  - Marie-France Sagot (Program Director)
  - Jorge Carneiro (Program Deputy-Director)
  - Luis Rocha (Collaboratorium Director)
- **Background**
  - Knowledge of empirical sciences (Physics, Chemistry, Biology) and quantitative technical disciplines (programming, appplied mathematics, statistics)
    - Introduction Module to catch up on biology, modeling, and CS
- **Regular Syllabus**
  - Training in Biology
    - Molecular evolution and sequences
      - Theory of evolution and population genetics, sequence alignment, from pairwise to multiple, from genes to genomes, molecular phylogeny
    - Structures (DNA, RNA and proteins)
      - Introduction to biomolecular structures, determination and visualisation, biomolecular structure mechanics, dynamics, prediction and design

Curriculum For Computational Biology

Graduate Study in Computational Biology: The Gulbenkian CB PhD program

- Regular Syllabus (Cont)
  - Training in Biology
    - Genome structure
    - Genome evolution and genome dynamics
    - Function classification
    - Transcriptomics and proteomics
    - Networks
      - Generic aspects, Protein interaction networks, Metabolic networks, Genetic networks
    - Systems Biology
      - Population biology, epidemiology and immune system, Cytoskeleton and cell morphogenesis, motion and chemotaxis modelling, Development and whole organism modelling, Evolutionary development, Computational Neurobiology
  - Training in Computer Science
    - Algorithms in computational biology
    - Statistical data mining and machine learning
    - Database management systems, knowledge systems and integration
    - Introduction to dynamical systems

http://bc.igc.gulbenkian.pt/pdbc/syllabus.htm

informatics
luis rocha 2006

What's to come

- Monday, June 19, 2006
  - 10:00 – 13:00 **Introduction:** ***From Bioinformatics to Systems Biology***
    - by Luis M. Rocha, Indiana University and Instituto Gulbenkian de Ciencia
  - 14:30 – 15:30 ***Microarray Data Analysis with Data Mining and Machine Learning Methods***
    - by Miguel Rocha and Isabel Rocha, Universidade do Minho
  - 15:30 – 16:30 ***Modeling and Optimization of Metabolic and Regulatory Networks in Systems Biology*** by Miguel Rocha and Isabel Rocha, Universidade do Minho
- Tuesday, June 20, 2006
  - 10:00 – 11:00 ***GENE-CBR: a Case-Based Reasoning Tool for Cancer Diagnosis using Microarray Datasets***
    - by Florentino Fernández Riverola, Universidad de Vigo.
  - 11:00 – 12:00 ***Bibliome Informatics***
    - by Luis M. Rocha, Indiana University and Instituto Gulbenkian de Ciencia
  - 12:00 – 13:00 ***Machine Learning Methods for Computational Proteomics and Beyond***
    - by Pierre Baldi, University of California, Irvine

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY