

complex systems modeling



from computational biology to artificial life

luis m. rocha

CCS3 - modeling, algorithms, and informatics
los alamos national laboratory, MS B256
and
instituto gulbenkian de ciencia, Oeiras

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



Luis Rocha
2004



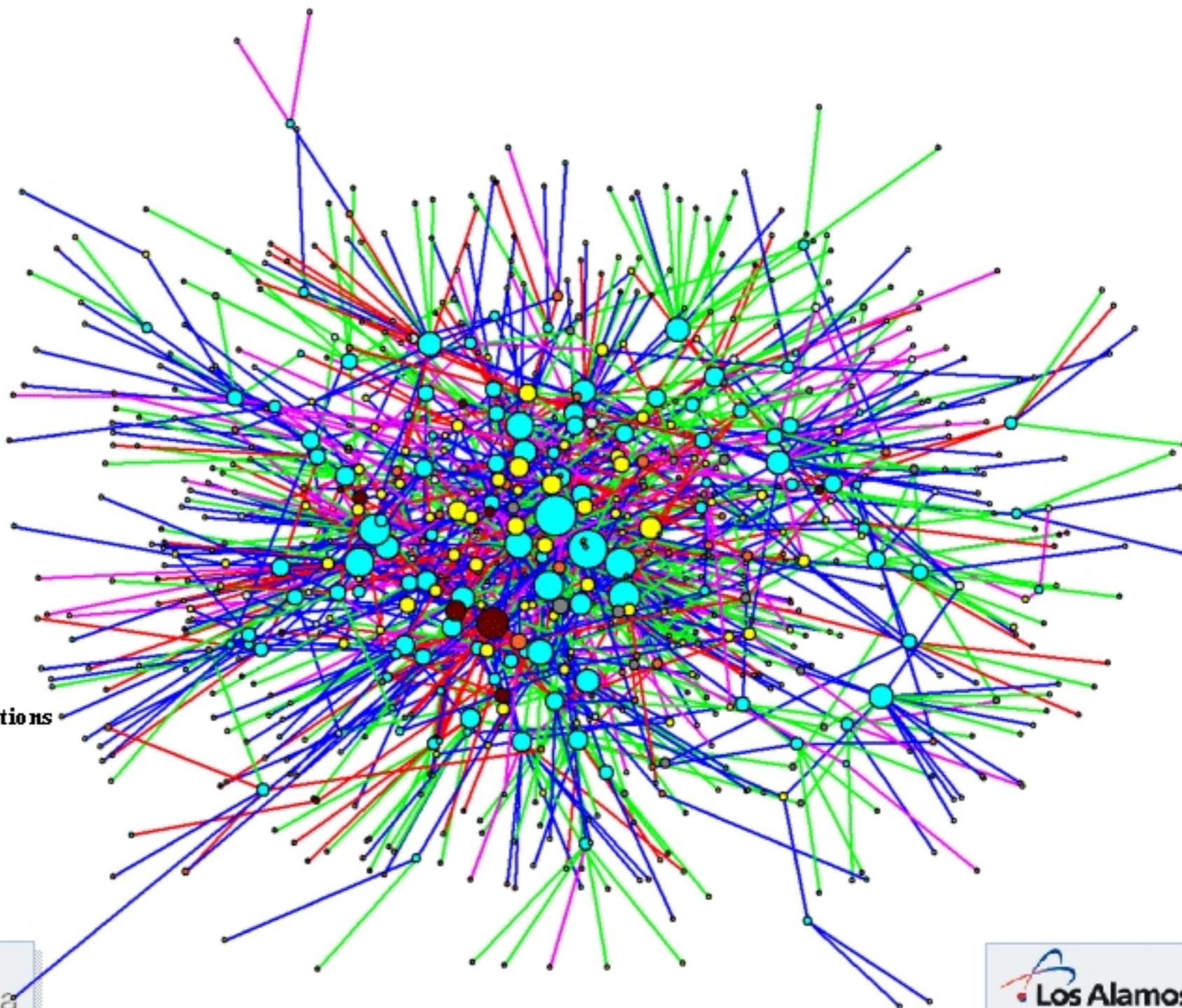
collaboration in the life sciences

- R&D in Biotechnology requires collaborations across multiple types of organizations
 - ▶ Walter Powell, Jason Owen-Smith, Douglas White, Kenneth Kopout
 - ▶ Work described in the following papers:
 - “Interorganizational collaboration and the locus of innovation: networks of learning in Biotechnology”. *Administrative Science Quarterly* 41(1):116-45.
 - “Practicing polygamy with good taste: the evolution of interorganizational collaboration in the Life sciences”.
 - “A comparison of U.S. and European University-Industry relations in the Life Sciences”
- Study of network dynamics to understand the evolution of the field

Luis Rocha
2004

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

1998 all ties



Tie Key

- RED**= R&D
- GREEN**= FINANCE
- BLUE**= COMMERCIALIZATION
- MAGENTA**= LICENSING

Node Key

- CYAN**= DBF
- ORANGE**= Public Research Organizations
- BROWN**= Gov't
- YELLOW**= Pharma
- GRAY**= VC
- WHITE** = Other

Node size = standardized network degree

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



Luis Rocha
2004



1998 new ties

Luis Rocha
2004

Tie Key

RED = R&D

GREEN = FINANCE

BLUE = COMMERCIALIZATION

MAGENTA = LICENSING

Node Key

CYAN = DBF

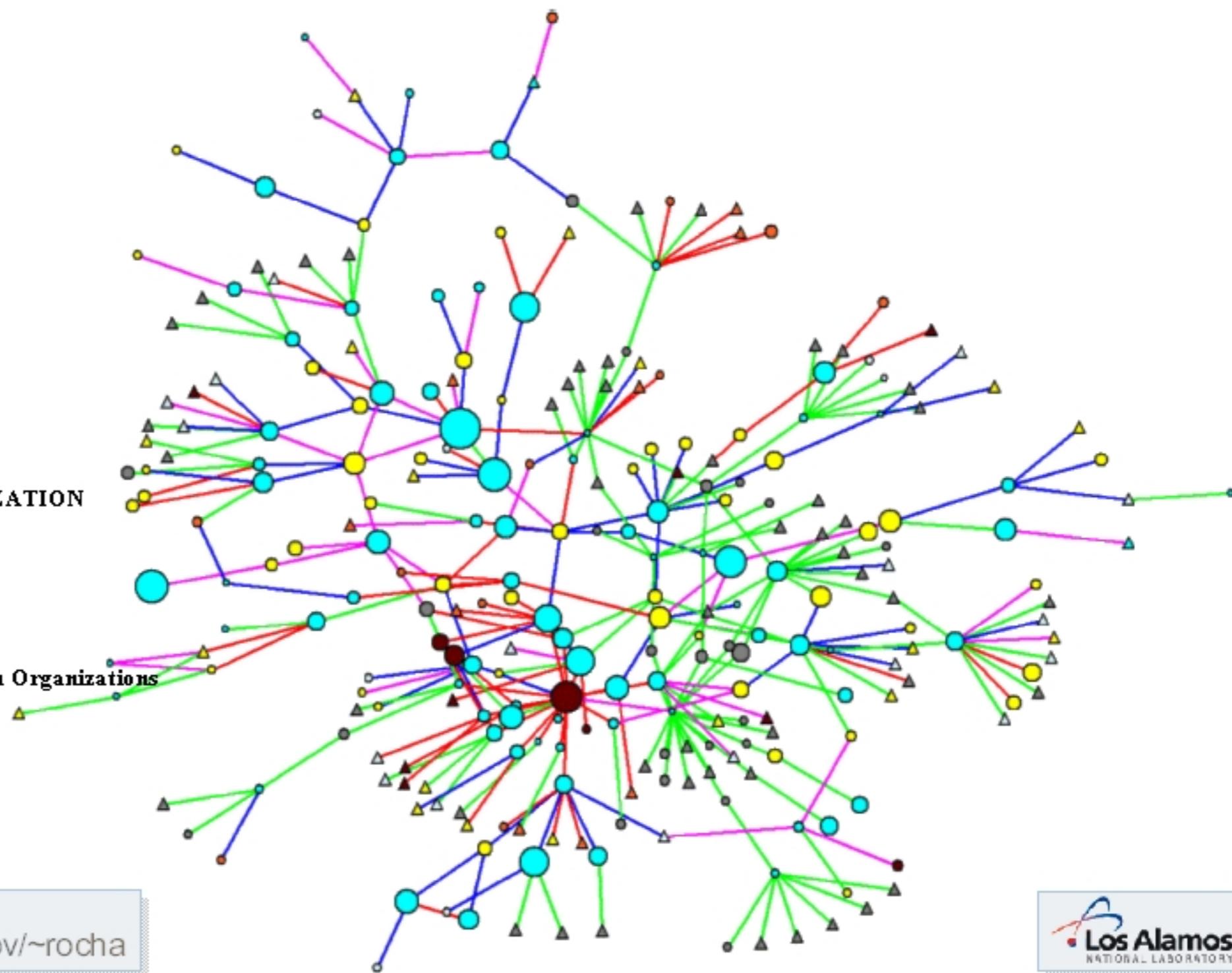
ORANGE = Public Research Organizations

BROWN = Gov't

YELLOW = Pharma

GRAY = VC

WHITE = Other

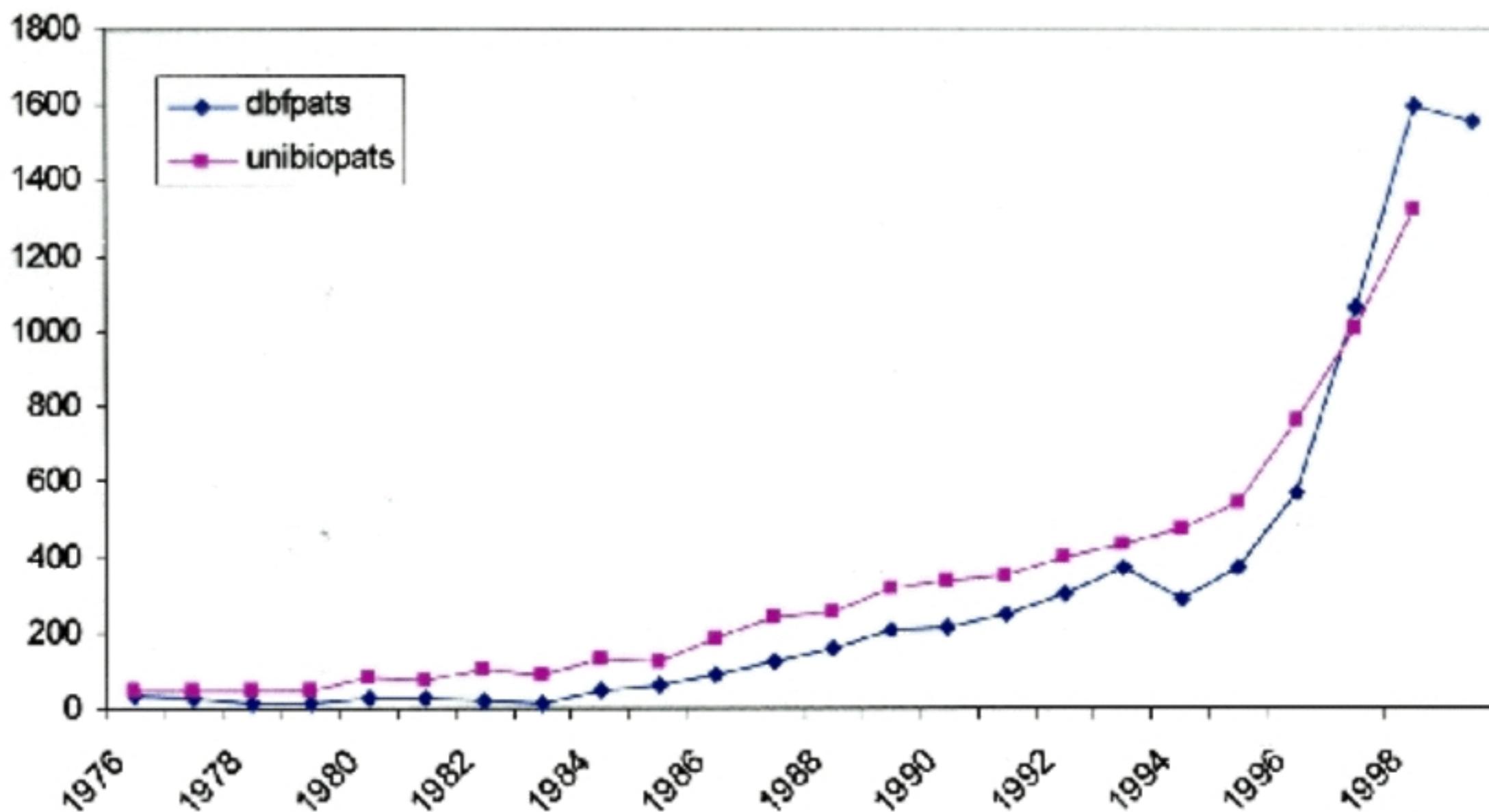
Node size = standardized
network degreerocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

of dedicated biotech firm (DBF) relationships

- **Early years: Lattice-relationship with universities and large corporations**
 - ▶ Dedicated Biotech Firms (DBFs) very small to bring medications to the market
 - ▶ Large corporations lacked closeness to universities
- **Later years: Collaboration remained**
 - ▶ Firms became stronger to produce medications
 - ▶ Corporations created in-house molecular biology research
 - ▶ Goal of utilizing innovation networks
- **General Observations**
 - ▶ Liability of disconnectedness
 - Less-linked organizations most likely to fail
 - ▶ DBFs had to both “make news and be in on the news”
 - Generate novel contributions and have capability to evaluate what other organizations were doing
 - ▶ Pathway to centrality on network through R&D collaborations
 - Preference for diversity
 - ▶ Dynamic environment
 - DBFs not renewing or expanding collaborations lost centrality
 - ▶ Venture Capital somewhat counter cyclical
 - VC adverse to risk of rejection of drug candidates
 - More enthusiasm for IT
 - Government funds more R&D and VC specializes in finance

universities and firms constitute the same community

DBF and Uni Biopats only 1976-1999



Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



■ U.S.

- ▶ Public Research Organizations, DBFs, Universities conduct R&D across multiple therapeutic areas and stages
- ▶ Robust and diverse national network of collaborations
- ▶ Academic mobility and incentives for entrepreneurship
- ▶ Generalist regional clusters
 - Attracts quality researchers
- ▶ Greater interdisciplinary outlook
 - From the blurring of boundaries between basic and applied research and competition for funding requiring interdisciplinary teams
 - E.g. Bioinformatics and Computational Biology

■ E.U.

- ▶ Regional specialization and Narrow competencies
- ▶ Less diverse group of public institutes working in a smaller number of therapeutic areas
- ▶ Institutes develop local ties to DBFs working on similar problems
- ▶ Cross-national linkages depend large Pharma Corps
- ▶ More difficult industry-university relations
 - Legal restrictions and cultural predispositions
- ▶ Need to integrate basic research and clinical development; foster collaborations among Universities, DBFs, public institutes and large Pharma



Luis Rocha
2004



The Mathematical and Computational Biology Collaboratorium

to foster international and inter-institutional collaboration and re-integration

- Open organization, designed to enable intense cooperation with national and international institutions
 - ▶ An enabling hub in a collaborative network of research, educational, and industrial institutions
 - ▶ Synergy between MCBC and associated PhD Program in CB
 - ▶ Goals:
 - Provide facilities for visiting scientists to collaborate.
 - Host informatics technology for off-site collaboration and research in Mathematical and Computational Biology.
 - Provide bioinformatics services for both academic and industrial partners
 - Attract quality students, for PhD program on computational biology (PCB), and facilitate the re-integration of these students in the Portuguese research community.
 - ▶ Hosted by Instituto Gulbenkian de Ciencia

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004

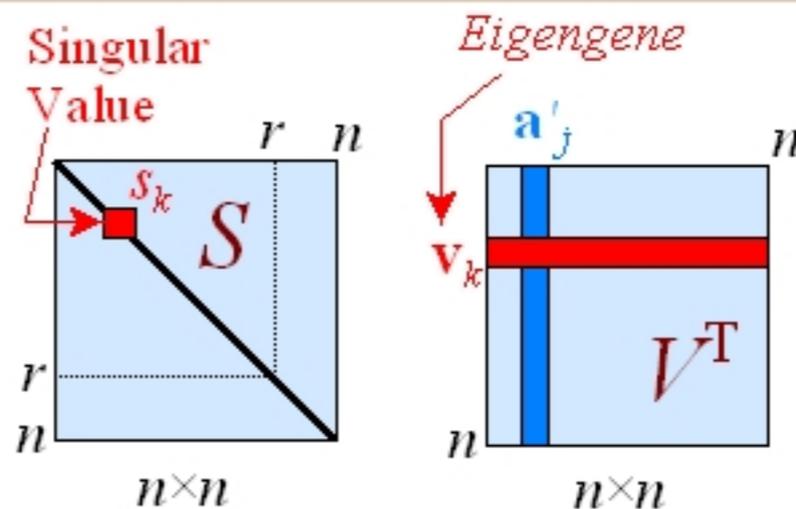
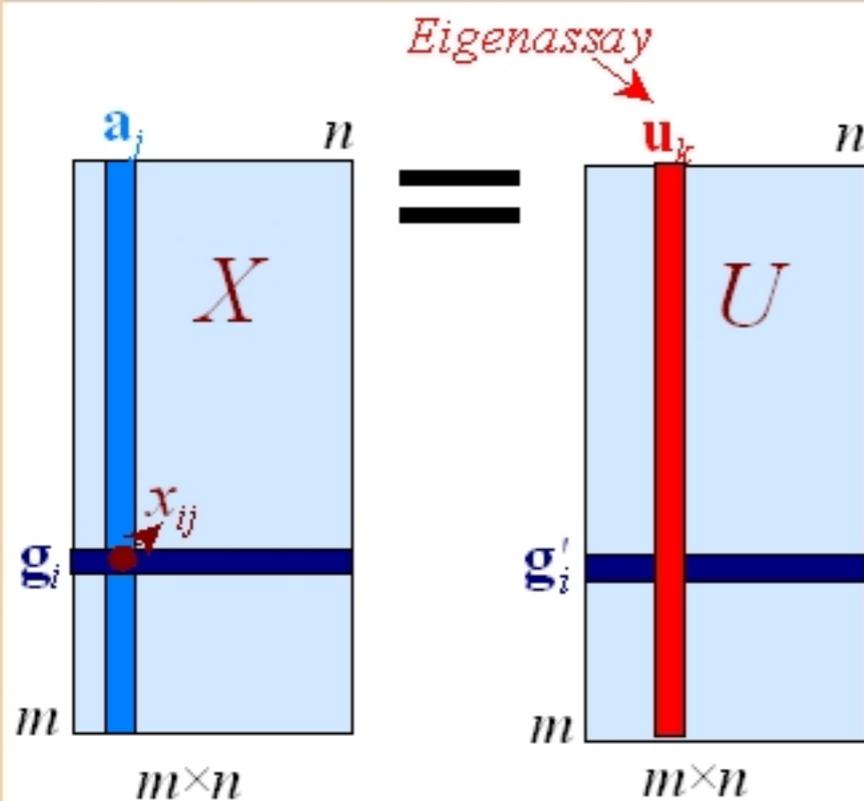


network identity and bottom-up methodology

- A **complex system** is any system featuring a large number of interacting components (agents, processes, etc.) whose aggregate activity is nonlinear: not derivable from the summations of the activity of individual components
- **Network identity**: Components form aggregate structures or functions that requires more explanatory devices than those used to explain the components
 - ▶ Genetic networks, Immune networks, Neural networks, Social insect colonies, Social networks, Distributed Knowledge Systems, Ecological networks
- **Bottom-up Methodology**
 - ▶ In **Modeling and Simulation**: Study of Simple Rules that Produce Complex Behavior
 - Agent-based Simulation, Evolutionary Computation, Swarm Modeling, Artificial Immune Systems
 - ▶ In **Analysis**: Discovery of Global Patterns of behavior
 - Dynamics-preserving decompositions
 - Spectral methods (SVD, Fourier Analysis)
 - Study Graphs built from simple associations amongst components
 - E.g. co-occurrence and Aggregate behavior, Power laws, small world, semi-metric behavior
 - Integration of knowledge from several sources
 - Use of electronic knowledge databases to discover biological function
 - ▶ For **Applications**: Construction of emergent global behavior
 - Building blocks based on lower-level, simple statistical relations
 - Artificial systems endowed with adaptability

singular value decomposition

microarray analysis



Rows of V^T : **eigengenes** (columns are time steps)
 Each gene's expression pattern is a linear combination of the eigengene patterns.

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

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 expression pattern

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

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.

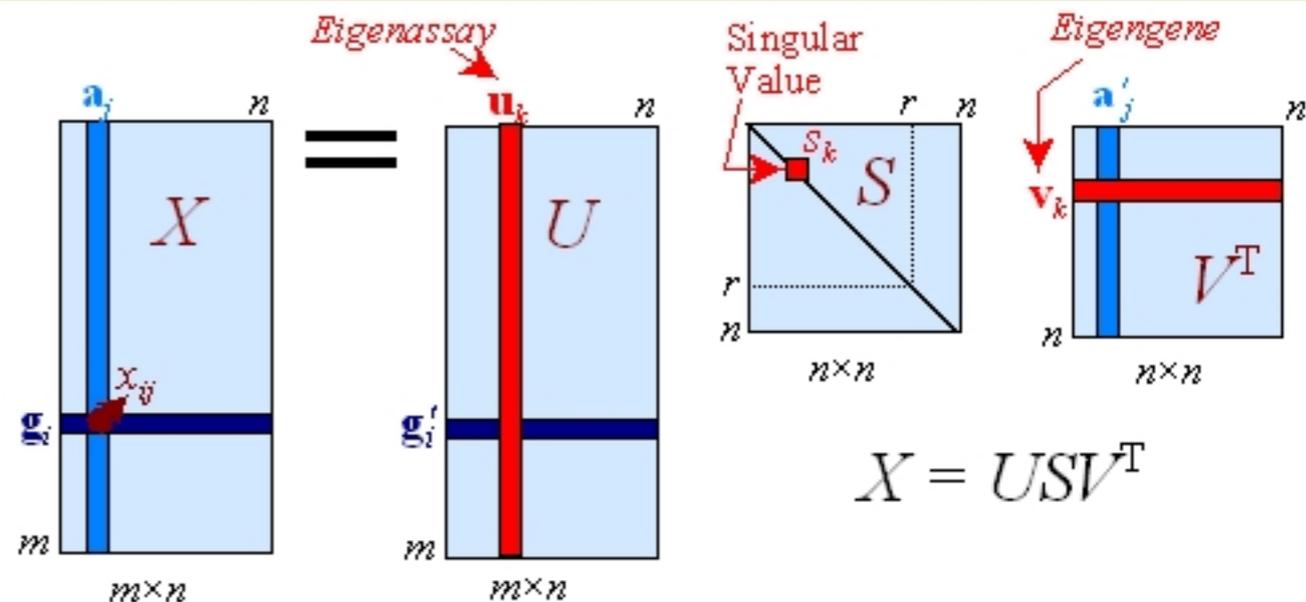
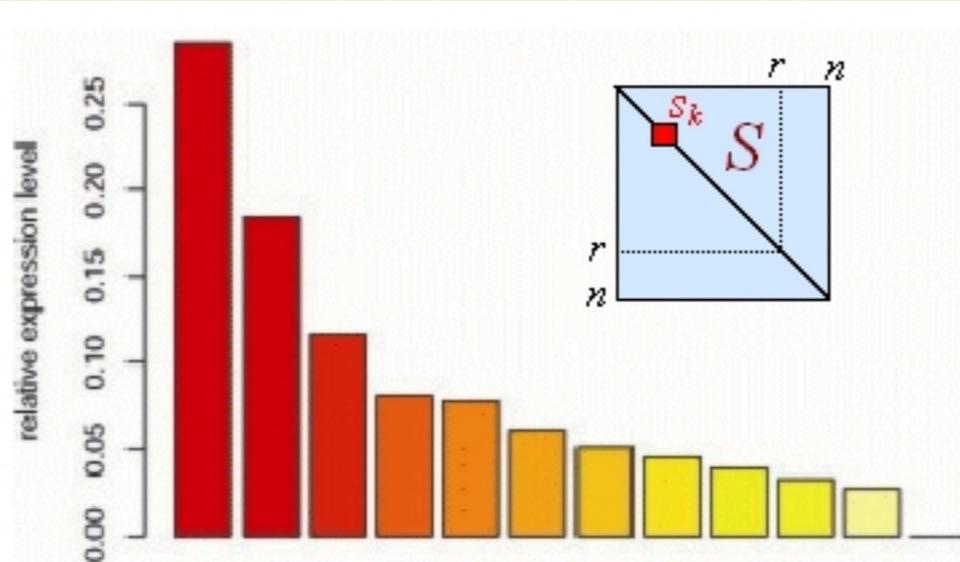
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
 2004

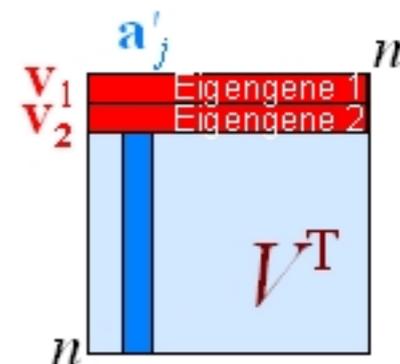
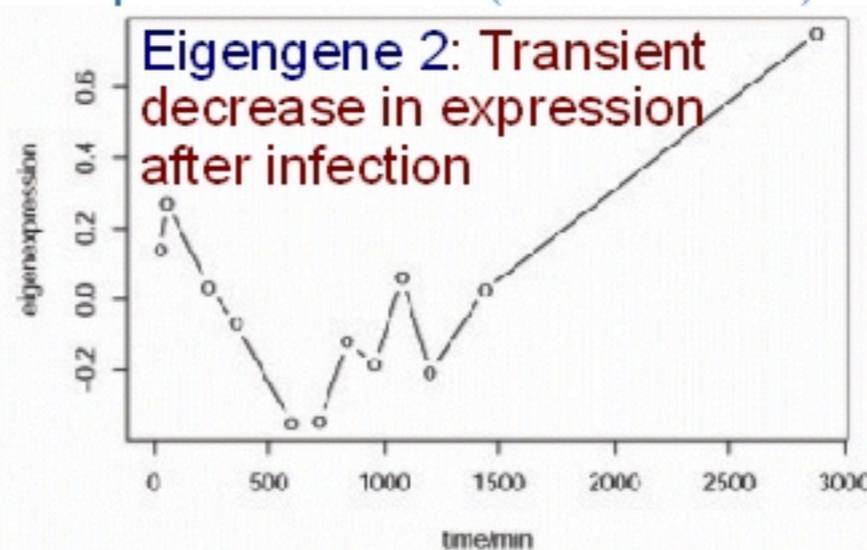
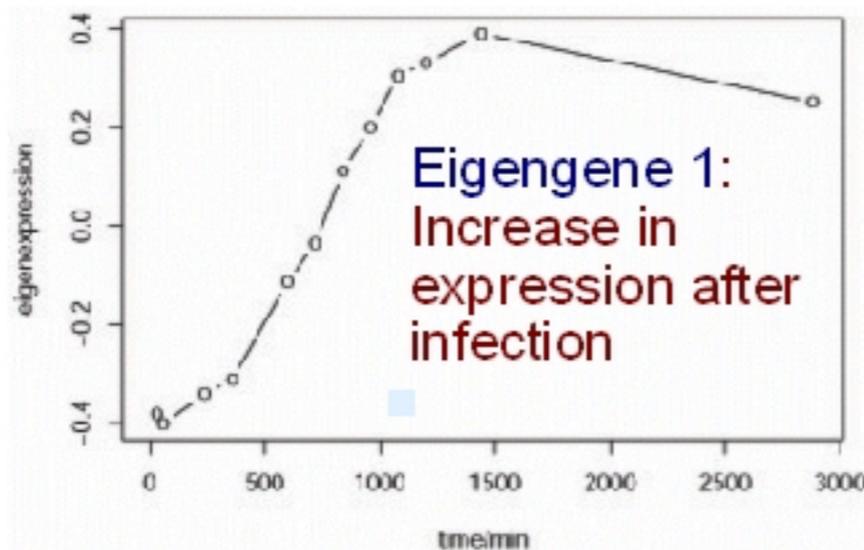


singular value decomposition

gene expression (13000 genes) after infection with herpes virus



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

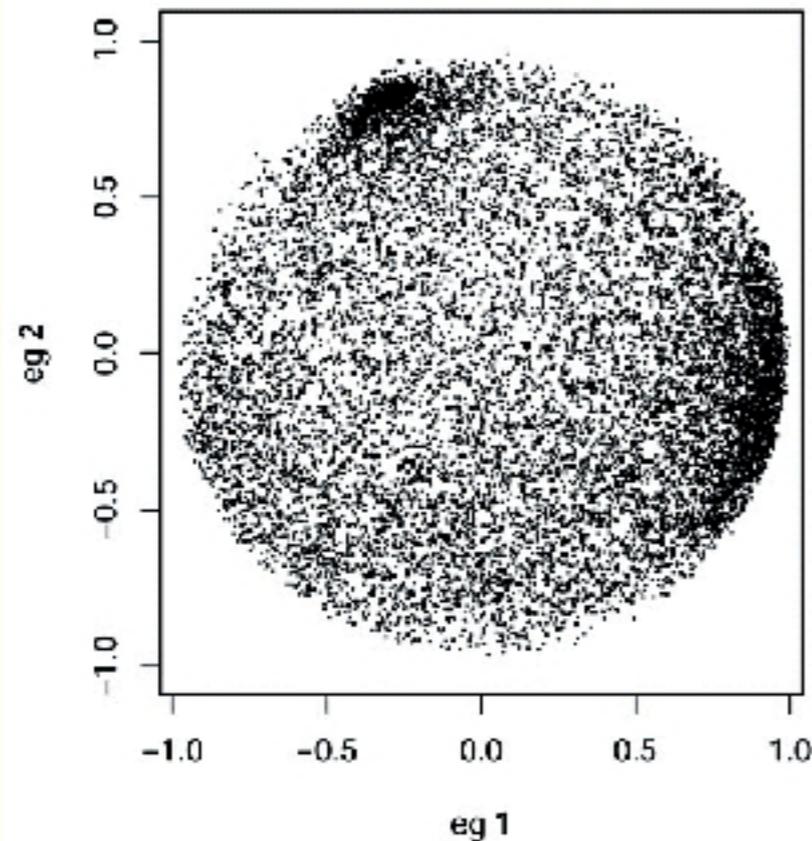
rocha@lanl.gov
http://www.c3.lanl.gov/~rocha

Luis Rocha
2004

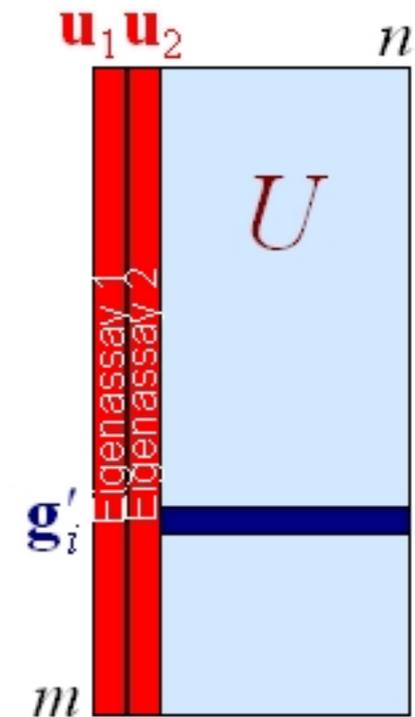
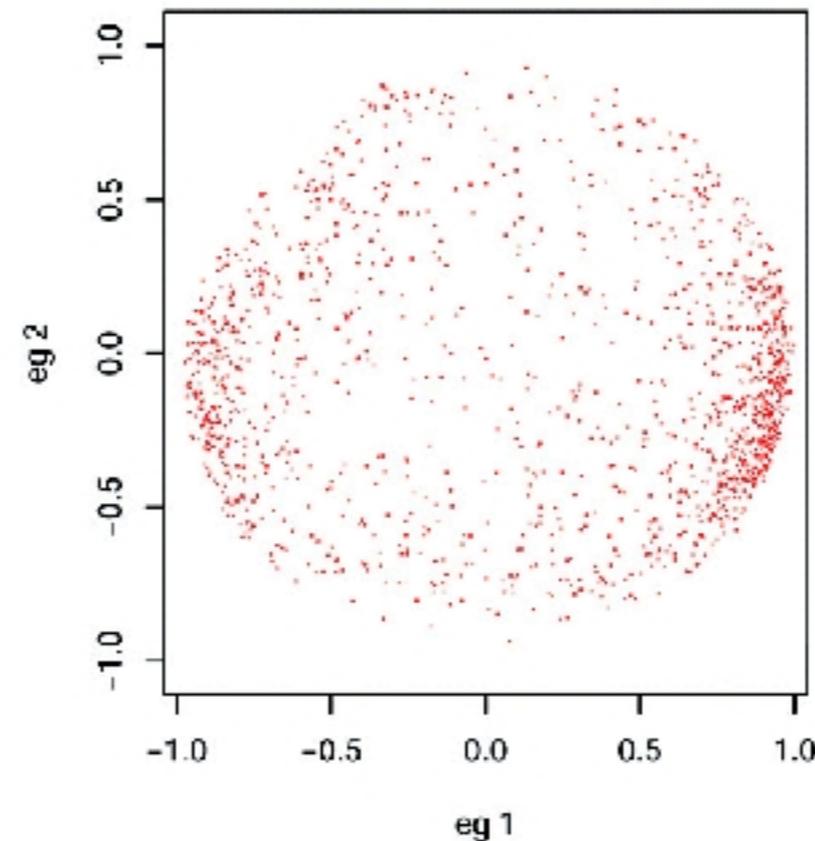
biological discovery via SVD

eigenassay coefficient plot: human cytomegalovirus infection

a) Correlation



b) Selected probe sets



We identified several functional categories important to host cell responses to HCMV infection.
dChip plus SVD

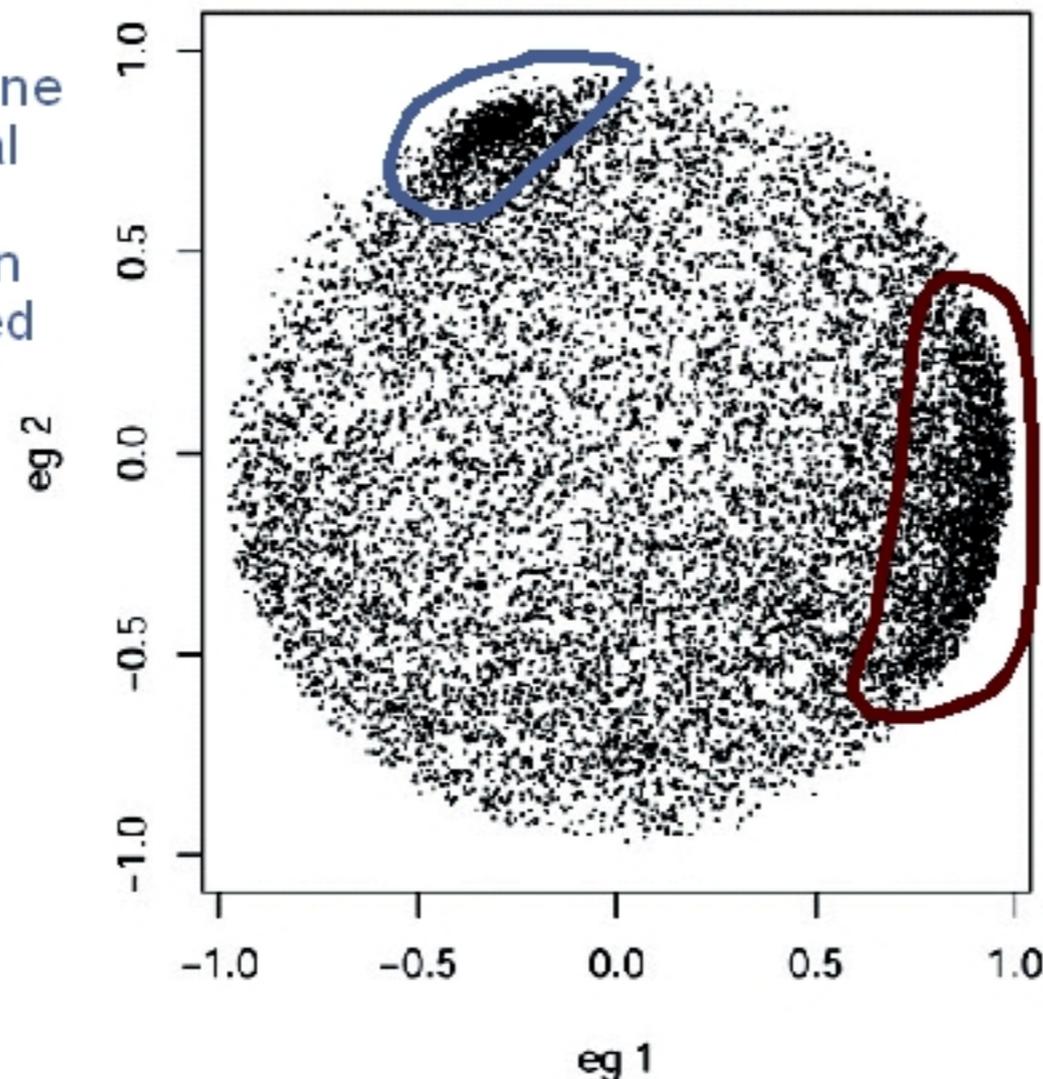
Princeton group (Brown et al (2001) *J Virol*, 75, 12319-12330.) found ~1200 genes that showed significant changes in expression

GeneChip plus (3) fold change in expression (at at least 2 consecutive time points)

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

eigenassay coefficient plot: human cytomegalovirus infection

a) Correlation



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.



Luis Rocha
2004



density estimation for filtering and analysis of time-series

of gene expression profiles on 2 orthonormal Directions

■ Data Filtering

- ▶ Serial Correlation (auto-correlation) Test
 - Removes gene expression profiles which are serially random

■ Choose a 2-D correlation subspace

- ▶ Singular value decomposition on remaining gene profiles
 - Choose subspace as a pair of interesting singular vectors
- ▶ We could choose other directions
 - Specific assays, abstract assays, etc.

■ Detect filtering boundary in space

- ▶ Genes at the center (periphery) of space are least correlated with directions of interest and tend to be uniformly distributed (clustered)
- ▶ Find boundary that detects a change from uniform to clustered behavior
 - Track the distribution of polar angles, and its difference from uniform (g)
 - choose the radius where we find the largest rate of change of g .

■ Choose regions of higher density

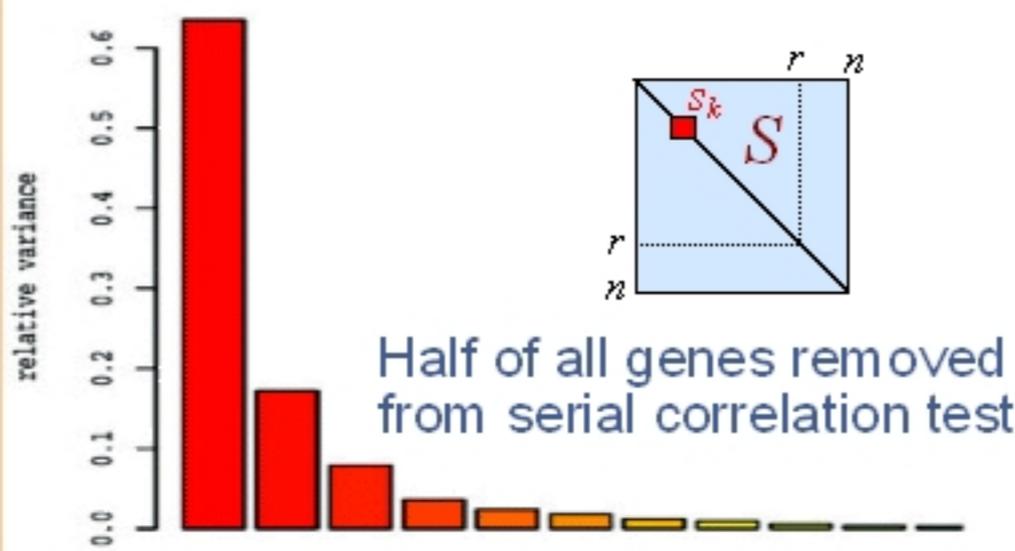
- ▶ Track density of polar angles, choose clusters

Rechtsteiner, A., R. Gottardo, L.M. Rocha, and M.E. Wall [2003]. "Singular Value Decomposition for Analysis of Gene Expression". *Currents in Computational Molecular Biology*. (RECOMB 2003), pp. 275-276.

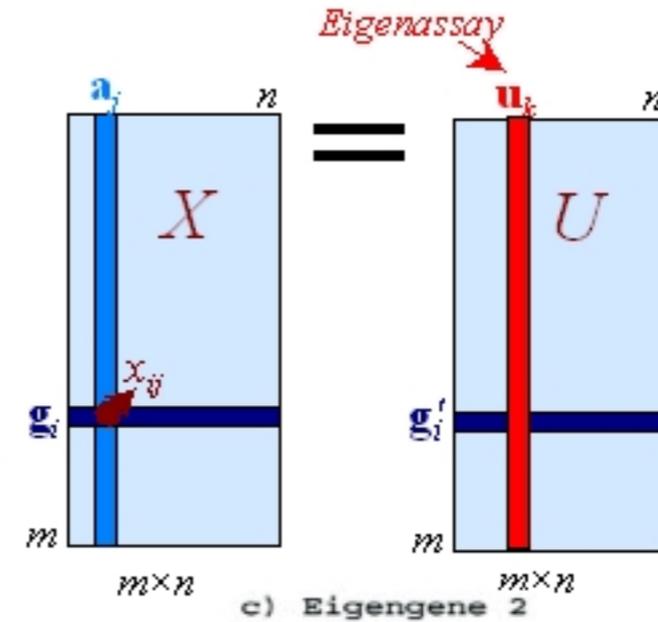
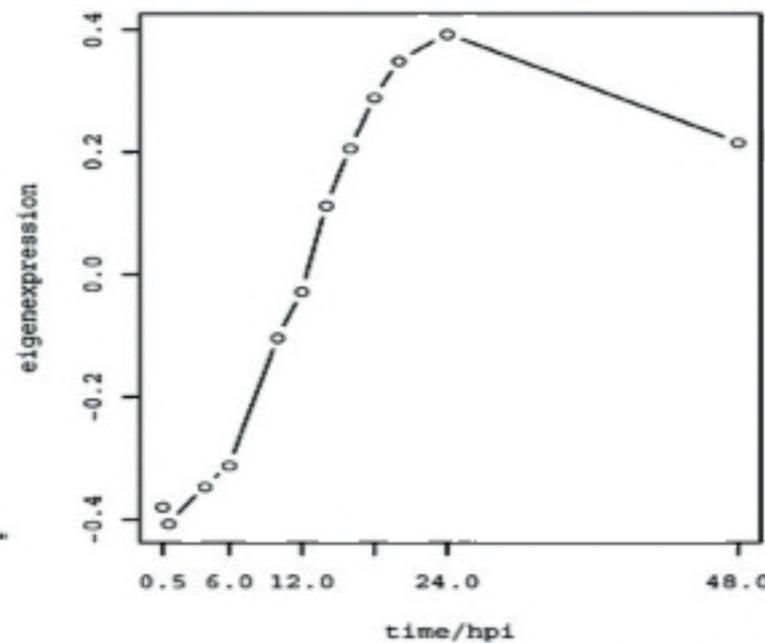
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

SVD + auto-correlation filtering

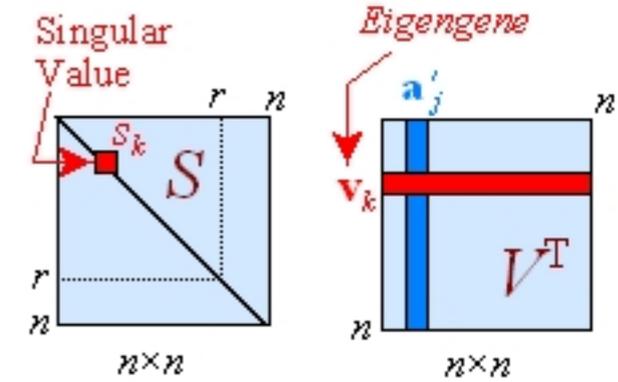
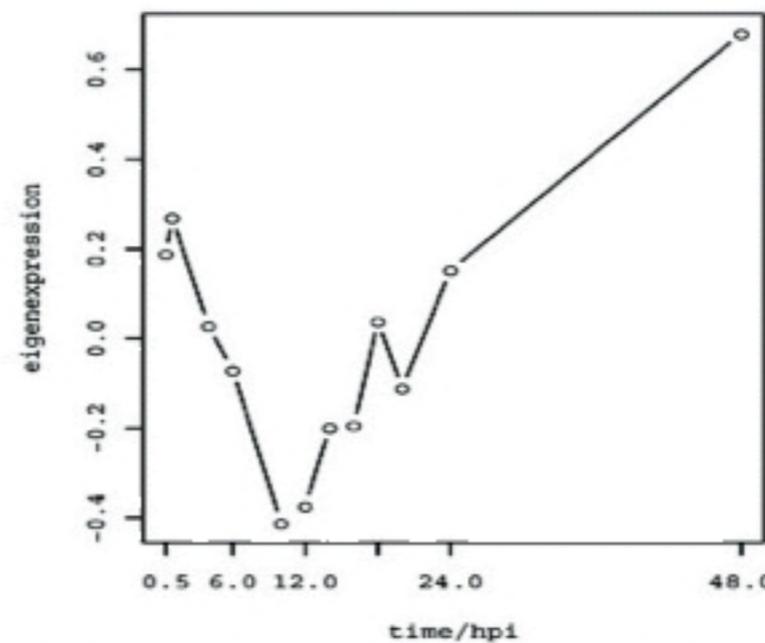
13000 human genes after infection with herpes virus



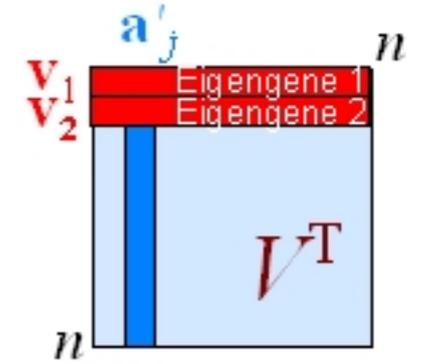
b) Eigengene 1



c) Eigengene 2



$$X = USV^T$$



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Rechtsteiner, A. and L.M. Rocha [2004]. "Validation and annotation of co-expressed clusters with LSA in MeSH Space". *RECOMB 2004*. In Press.

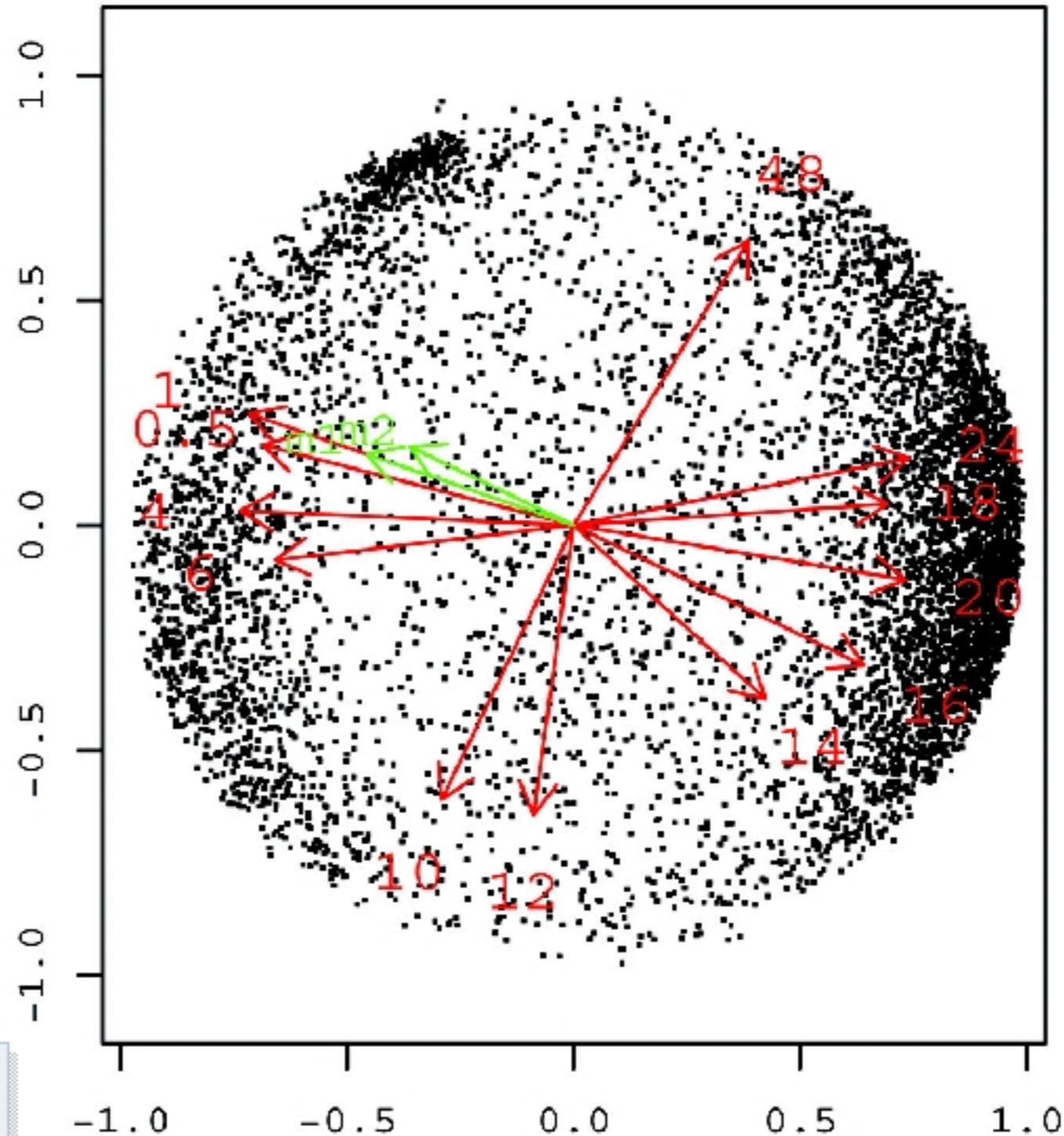


Luis Rocha
2004



time-series in SVD space

assays are organized almost perfectly sequentially



Rechtsteiner, A. and L.M. Rocha [2004]. *RECOMB 2004*. In Press.

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004



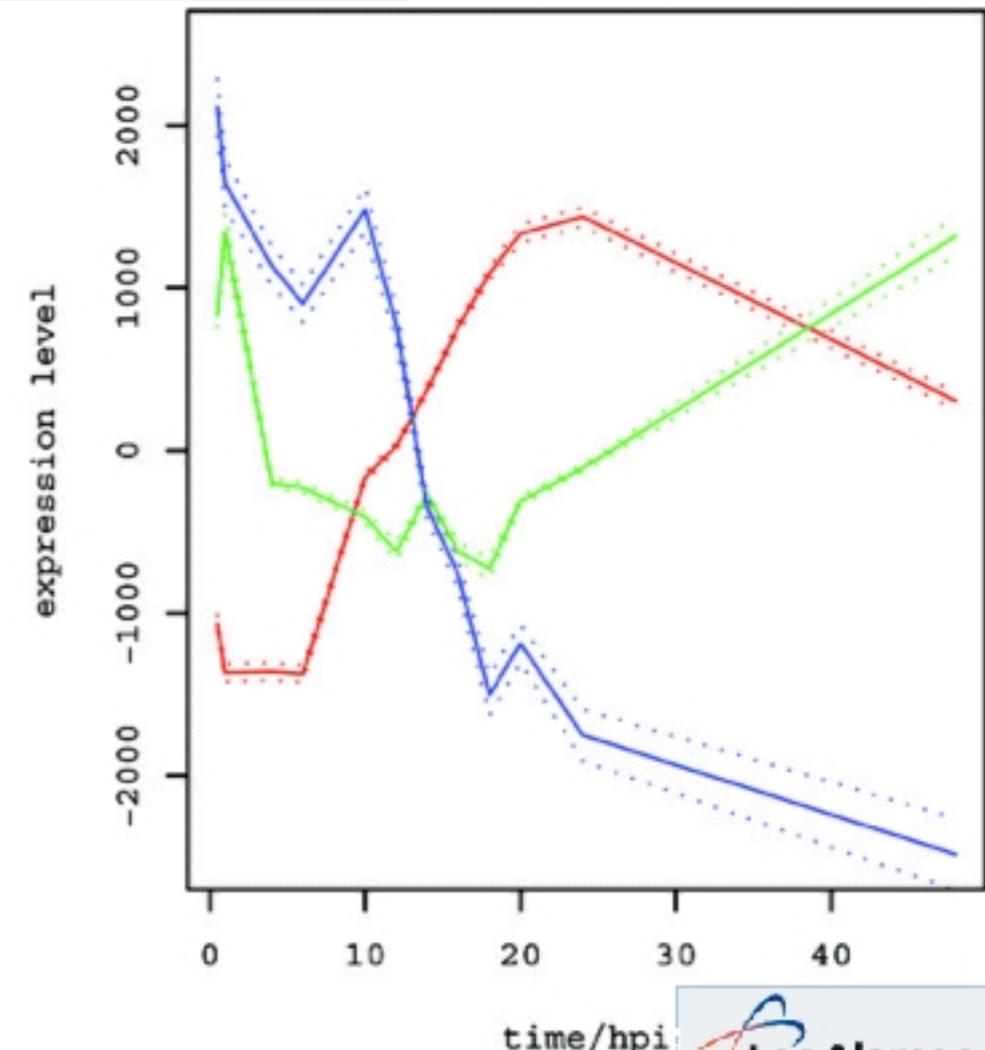
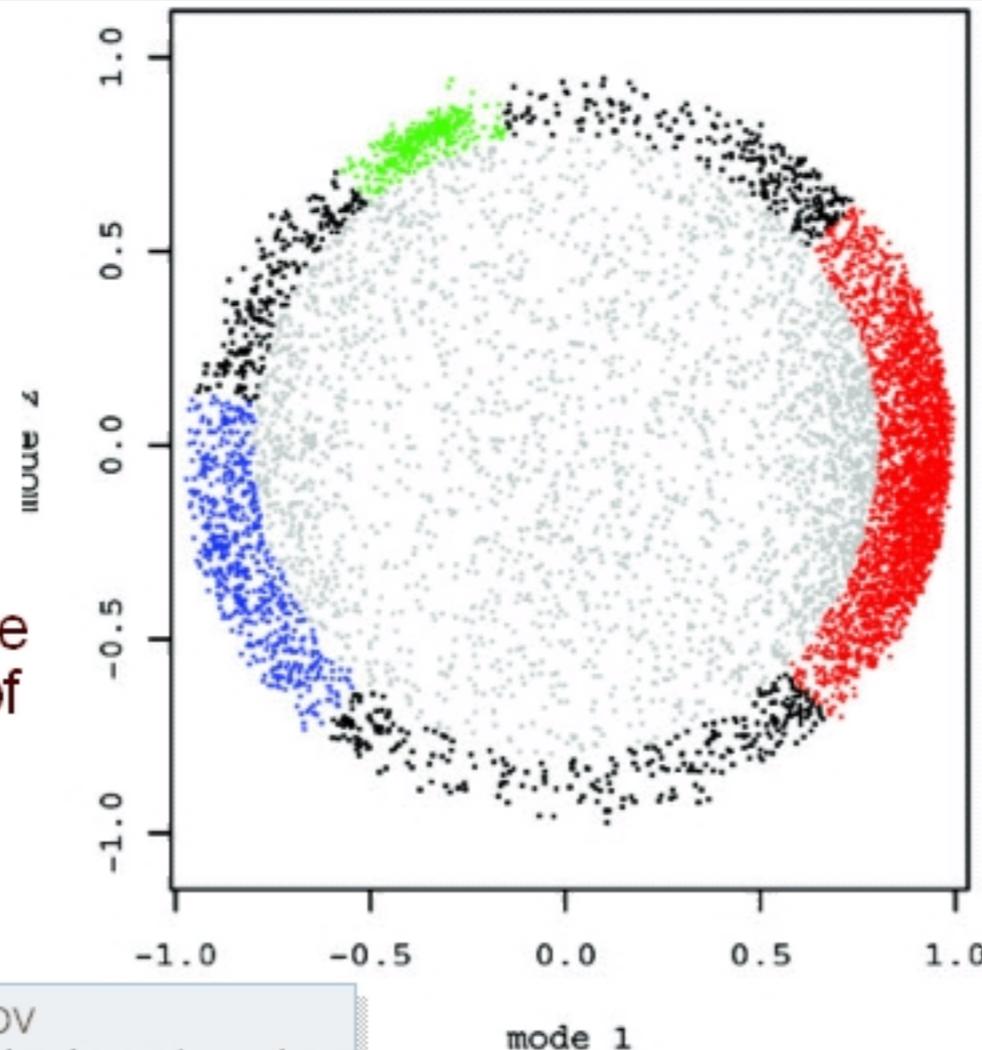
density estimation of polar angles

in SVD subspace (after serial correlation filtering)

- **Boundary in space**
 - ▶ largest rate of change of polar angle density from uniform
- **Choose regions of higher density**
 - ▶ By density of polar angles

Rechtsteiner, A. and L.M. Rocha [2004]. *RECOMB 2004*. In Press.

Luis Rocha
2004



What is the
Function of
genes in
clusters?

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



Summary

■ A Hierarchical Thesaurus (Ontology?)

- ▶ “National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity”
 - <http://www.nlm.nih.gov/mesh/>
 - Most general levels contain very broad headings such as "Anatomy" or "Mental Disorders." More narrow levels contain headings such as "Ankle" and "Conduct Disorder."
- ▶ Numbers
 - 21,973 descriptor headings.
 - Plus 132,123 Supplementary Concept Records within a separate chemical thesaurus.
 - Plus thousands of cross-references.
- ▶ Applications
 - Used by NLM for indexing articles in MEDLINE.
 - average of 10 headings per paper.
 - Source of the descriptors used in NLM's *Index Medicus*
- ▶ Updated continuously by its staff of 10



Luis Rocha
2004



Browse from Tree Top

- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
- Psychiatry and Psychology [F]
- Biological Sciences [G]
- Physical Sciences [H]
- Anthropology, Education, Sociology and Social Phenomena [I]
- Technology and Food and Beverages [J]
- Humanities [K]
- Information Science [L]
- Persons [M]
- Health Care [N]
- Geographic Locations [Z]

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



Luis Rocha
2004





Luis Rocha
2004



■ Chemicals and Drugs [D]

- ▶ Inorganic Chemicals [D01] +
- ▶ Organic Chemicals [D02] +
- ▶ Heterocyclic Compounds [D03] +
- ▶ Polycyclic Hydrocarbons [D04] +
- ▶ Environmental Pollutants, Noxae, and Pesticides [D05] +
- ▶ Hormones, Hormone Substitutes, and Hormone Antagonists [D06] +
- ▶ Reproductive Control Agents [D07] +
- ▶ **Enzymes, Coenzymes, and Enzyme Inhibitors [D08] +**
- ▶ Carbohydrates and Hypoglycemic Agents [D09] +
- ▶ Lipids and Antilipemic Agents [D10] +
- ▶ Growth Substances, Pigments, and Vitamins [D11] +
- ▶ **Amino Acids, Peptides, and Proteins [D12] +**
- ▶ Nucleic Acids, Nucleotides, and Nucleosides [D13] +
- ▶ Neurotransmitters and Neurotransmitter Agents [D14] +
- ▶ Central Nervous System Agents [D15] +
- ▶ Peripheral Nervous System Agents [D16] +
- ▶ Anti-Inflammatory Agents, Antirheumatic Agents, and Inflammation Mediators [D17] +
- ▶ Cardiovascular Agents [D18] +
- ▶ Hematologic, Gastrointestinal, and Renal Agents [D19] +
- ▶ Anti-Infective Agents [D20] +
- ▶ Anti-Allergic and Respiratory System Agents [D21] +
- ▶ Antineoplastic and Immunosuppressive Agents [D22] +
- ▶ Dermatologic Agents [D23] +
- ▶ Immunologic and Biological Factors [D24] +
- ▶ Biomedical and Dental Materials [D25] +
- ▶ Specialty Chemicals and Products [D26] +
- ▶ Chemical Actions and Uses [D27] +

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Lada Adamic and Eytan Adair (HP Labs)

Demo for retrieving genes relevant to MeSH terms

<http://www-idl.hpl.hp.com/cgi-bin/genelit>

Demo for Biomedical Abbreviation Dictionary

<http://www-idl.hpl.hp.com/cgi-bin/abbrevs/search.cgi>

Demo for finding Biomedical Abbreviations

<http://www-idl.hpl.hp.com/cgi-bin/abbrevs/search2.cgi>

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

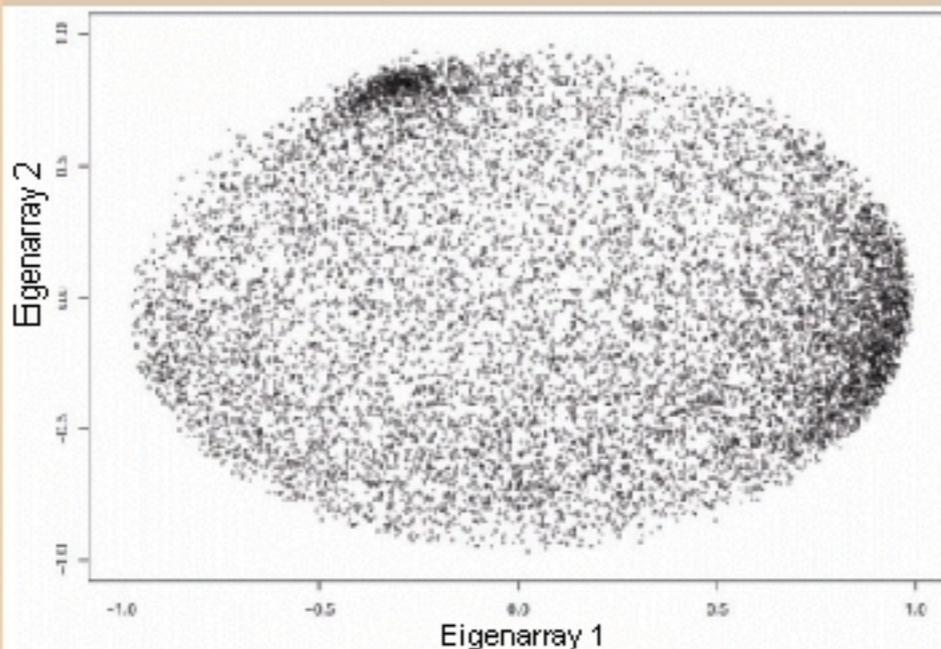


Luis Rocha
2004



MeshGene: Automatic Functional Annotation

Of sets of genes



Extract

Sets of genes most correlated and anti-correlated with each eigenarray (or cluster)

Checked likelihood that they are gene symbols (whether or not they co-occur with words like DNA, gene, etc.)

Genbank IDs
Gene symbols

Extract

Medline Docs

MeSH Headings

Identify MeSH terms which occur in at least 10% of documents mentioning some gene
Statistical measure significance of each individual gene for each MeSH term.

No significance (score=0) Mesh term co-occurs in 10% of documents for a given target gene

Significance (score>0) Mesh term co-occurs in more than 10% of documents for a given target gene. Score is number of standard deviations from the mean, assuming normal distribution

Collaboration with **Lada Adamic**, **Eytan Adair**, and **Bernardo Huberman** at *HP Labs*, Palo Alto.

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004

Interface and Results (2nd Mode in SVD of Herpes Data)

E05 Investigative Techniques: Immune Adherence Reaction	HLAA(7,3.30), GLI3(5,4.62), WISP3(5,5.26), CR1(13,19.88),
H01 Natural Sciences: Membrane Potentials	SLC6A12(2,2.04), RFC4(86,2.22), CASP9(1,2.45), KCNN4(33,26.30),
D02 Organic Chemicals: Taurine	RPE(1,2.38), SLC6A12(5,25.17), MUC5B(4,2.49), GYPE(1,2.68),
A15 Hemic and Immune Systems: Antigen-Presenting Cells	FCGR1A(5,2.78), HLAA(83,7.34), GLI3(53,10.20), WISP3(53,11.96),
D08 Enzymes, Coenzymes, and Enzyme Inhibitors: DNA-Directed RNA Polymerase	RFC4(72,7.17), MRPL23(2,4.03), POLR2J(4,8.21), POLRMT(1,12.18),
A11 Cells: Centromere	CAMK2A(1,21.04), POLR2J(1,3.46), CDC14A(1,2.61), LMX1B(1,4.27),
D24 Immunologic and Biological Factors: Antigens, CD4	CD48(53,18.70), TTTY1(1,6.21), IL2RB(2,2.01), TNFRSF7(7,4.22),
B04 Viruses: Herpesvirus 4, Human	HLAA(76,5.11), SPN(18,8.42), IL2RB(3,4.22), RFXANK(23,12.93),
D24 Immunologic and Biological Factors: Receptors, Interleukin-2	CD48(23,5.27), WISP3(39,3.12), IL2RB(15,17.38), TNFRSF7(9,4.80),
G05 Genetics: Genome, Human	PSMB7(1,12.70), GYPE(2,2.96), TNP1(1,4.45), HTR4(1,10.34),
A11 Cells: Sarcolemma	CBARA1(1,8.12), STX8(1,5.38), MUC5B(16,13.82), KCNN4(1,2.84),
C17 Skin and Connective Tissue Diseases: Panniculitis	SPN(1,6.28), GLI3(1,2.48), WISP3(1,2.79), IL2RB(1,18.28),
D12 Amino Acids, Peptides, and Proteins: Myosins	CBARA1(4,17.71), TPM2(1,5.61), SPN(4,2.64), TCN1(4,3.76),
D01 Inorganic Chemicals: Sodium	SLC6A12(6,7.94), MUC5B(67,13.15), KCNN4(8,5.77), SERPINB3(11,2.53),
D02 Organic Chemicals: Sulfhydryl Compounds	BNC(1,5.38), MUC5B(8,2.43), SKIL(5,13.53), LPO(17,8.00),
D12 Amino Acids, Peptides, and Proteins: Ankyrins	OSBP(1,8.07), CDKN2D(3,14.48), RPE(1,2.96), NCOR2(1,3.62),

Co-occurrence matrix
MeSH terms are scaled
according to their Inverse
Frequency

	M (MeSH Terms)
G (Genes)	$M \times G$

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

<http://www-idl.hpl.hp.com/meshgene>



Luis Rocha
2004



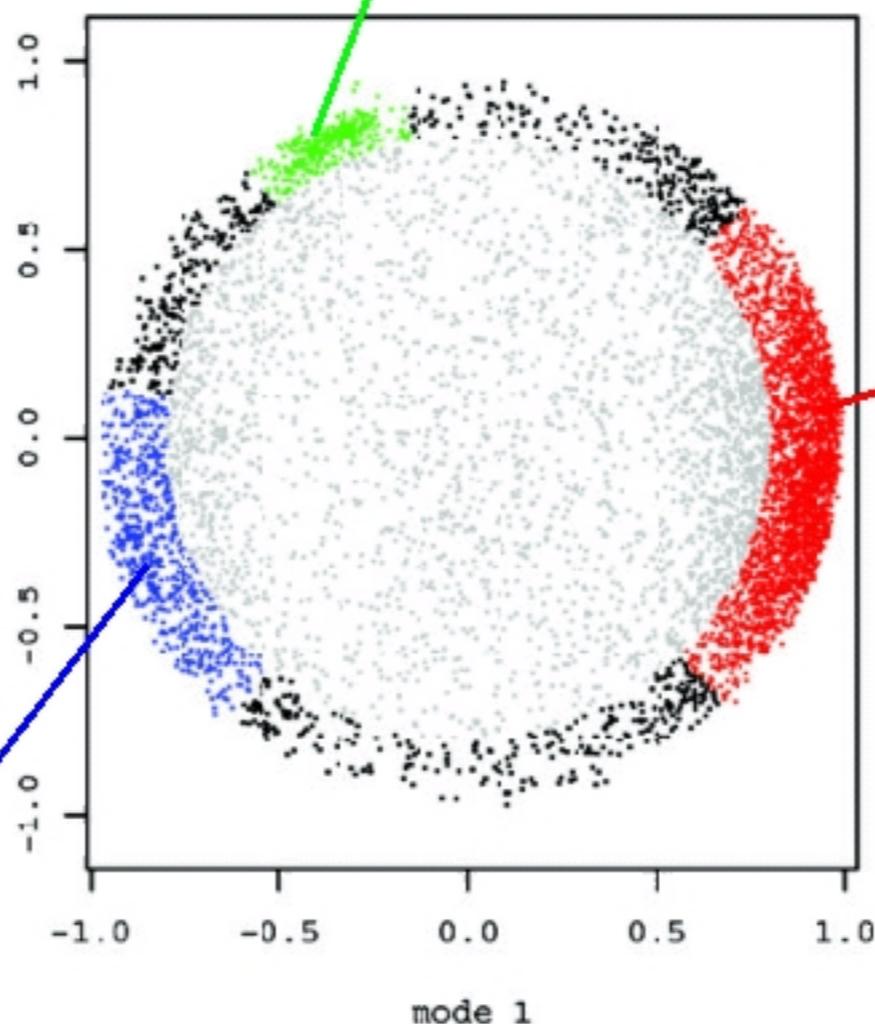
automatic functional gene annotation

using the biomedical literature

What is the
Function of genes
in clusters?

Rechtsteiner, A. and L.M.
Rocha [2004]. *RECOMB*
2004. In Press.

	$M(\text{MeSH Terms})$
$G(\text{Genes})$	$M \times G$



	$M(\text{MeSH Terms})$
$G(\text{Genes})$	$M \times G$

	$M(\text{MeSH Terms})$
$G(\text{Genes})$	$M \times G$

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004

SVD of Gene/MeSH co-occurrence

for each gene co-expression cluster: uncovering "functional themes"

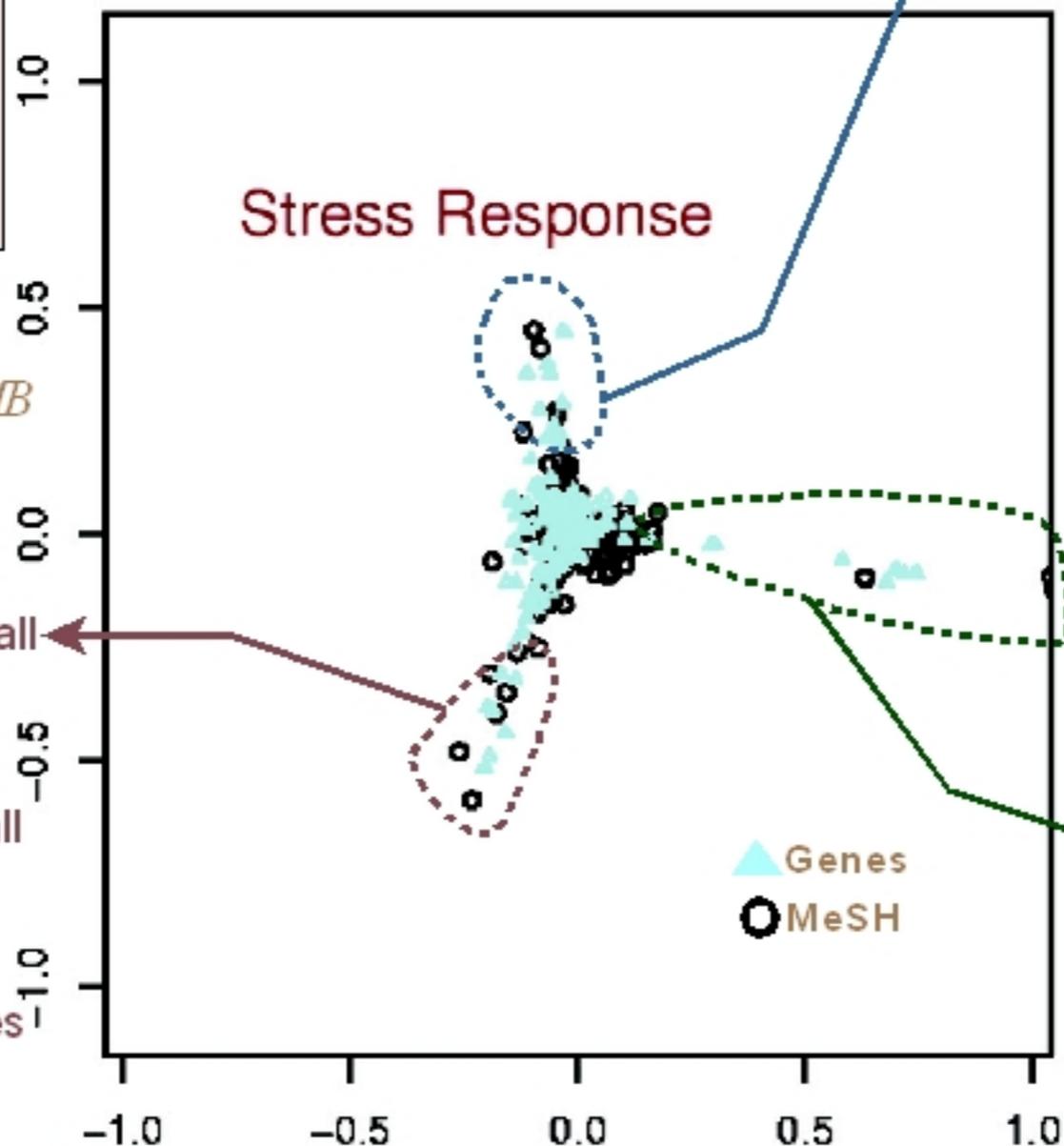
	$M(\text{MeSH Terms})$
$G(\text{Genes})$	$M \times G$

Rechtsteiner, A. and L.M. Rocha [2004]. *RECOMB 2004*. In Press.

Spliceosomes
 RNA- Binding Proteins
 Ribonucleoprotein, U1 -Small Nuclear
 Ribonucleoproteins
 RNA Helicases
 Ribonucleoprotein, U2 Small Nuclear
 Ribonucleoproteins, Small Nuclear
 RNA Nucleotidyltransferases
 Autoantigens

mode 3

cluster 1



Heat-Shock Proteins
 Heat-Shock Proteins 70
 Chaperonin 60
 Molecular Chaperones
 GroEL Protein
 Heat-Shock Proteins 90
 Chaperonins
 Protein-Serine-Threonine Kinases
 Stress
 Chaperonin 10

Ligases
 Ubiquitins
 SUMO-1 Protein
 Fungal Proteins
 Cell Cycle Proteins
 Cysteine Endopeptidases
 Polyubiquitin
 Peptide Synthases
 Bloom Syndrome
 Werner Syndrome
 Multienzyme Complexes

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
 2004

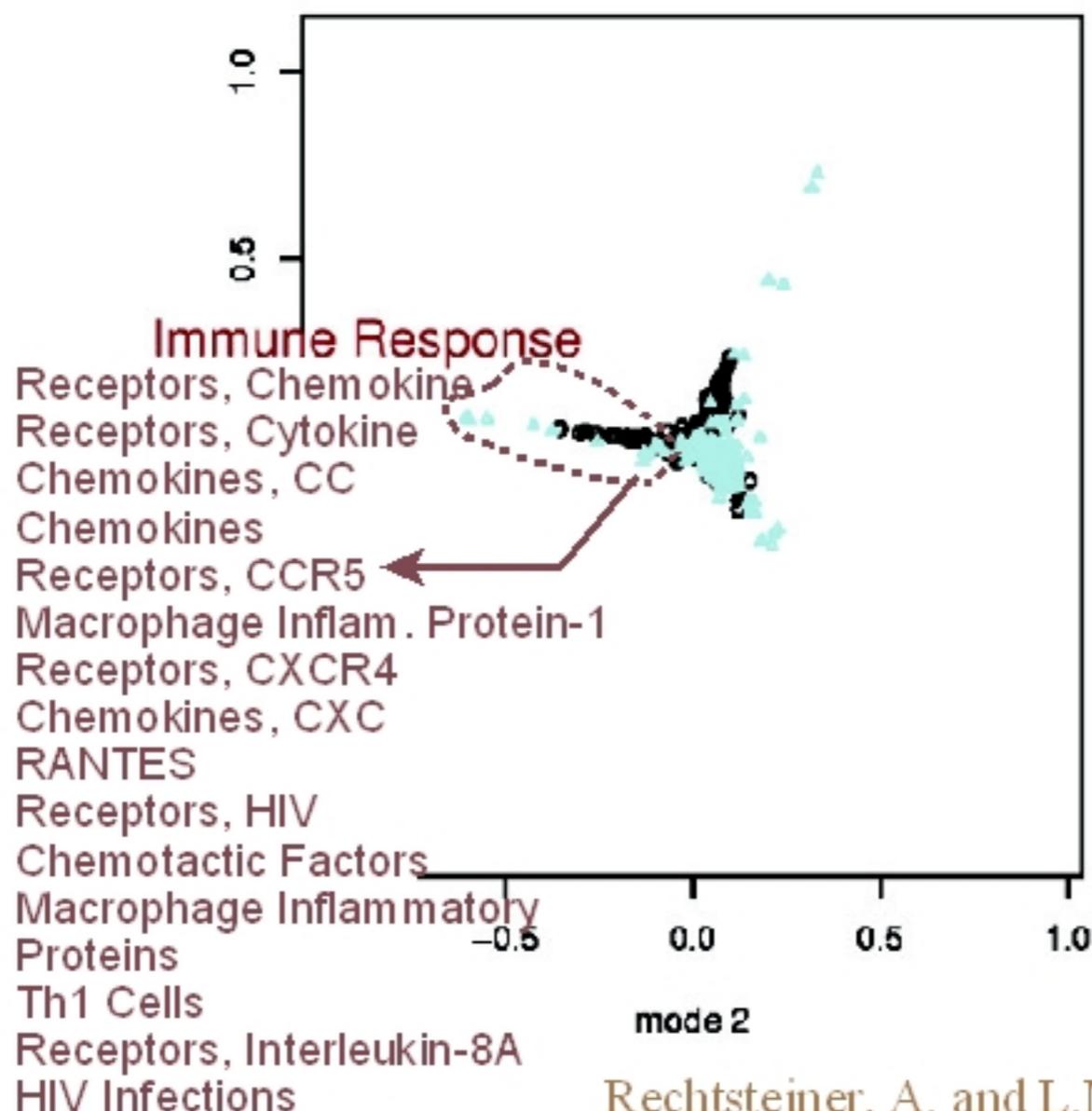


SVD of Gene/MeSH co-occurrence

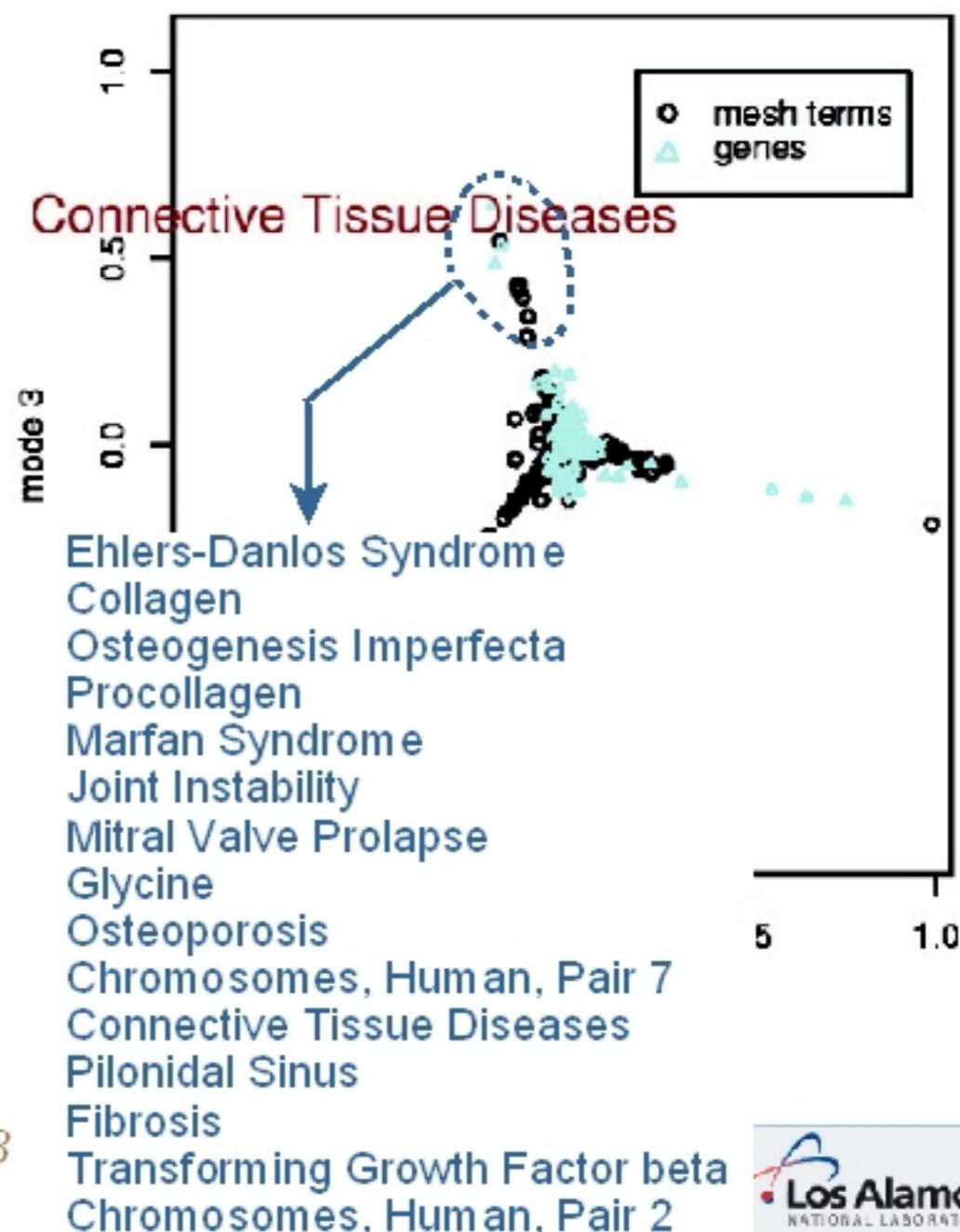
for each gene co-expression cluster: uncovering "functional themes"

Luis Rocha
2004

cluster 2



cluster 3



Rechtsteiner, A. and L.M.
Rocha [2004]. *RECOMB*
2004. In Press.

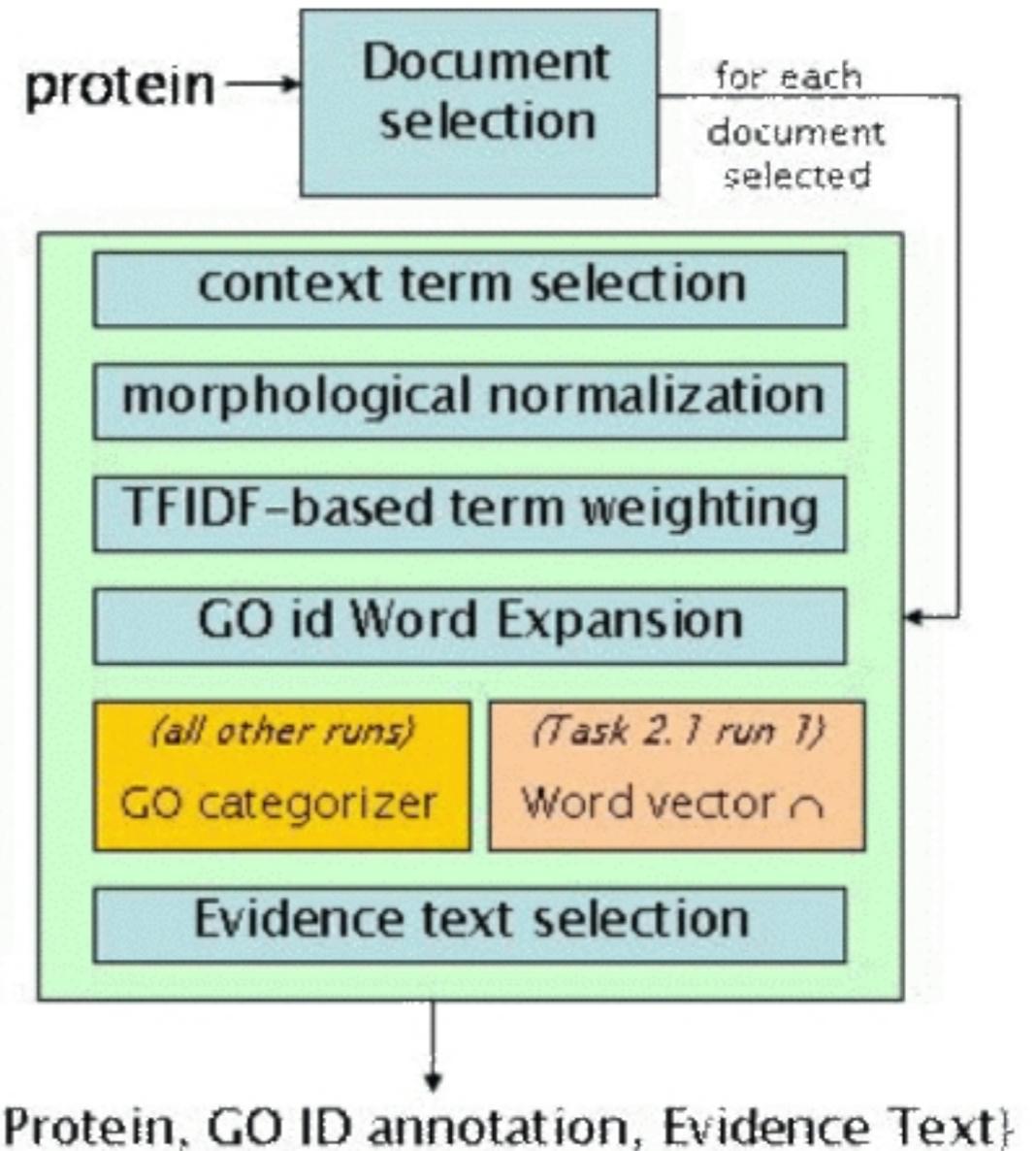
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



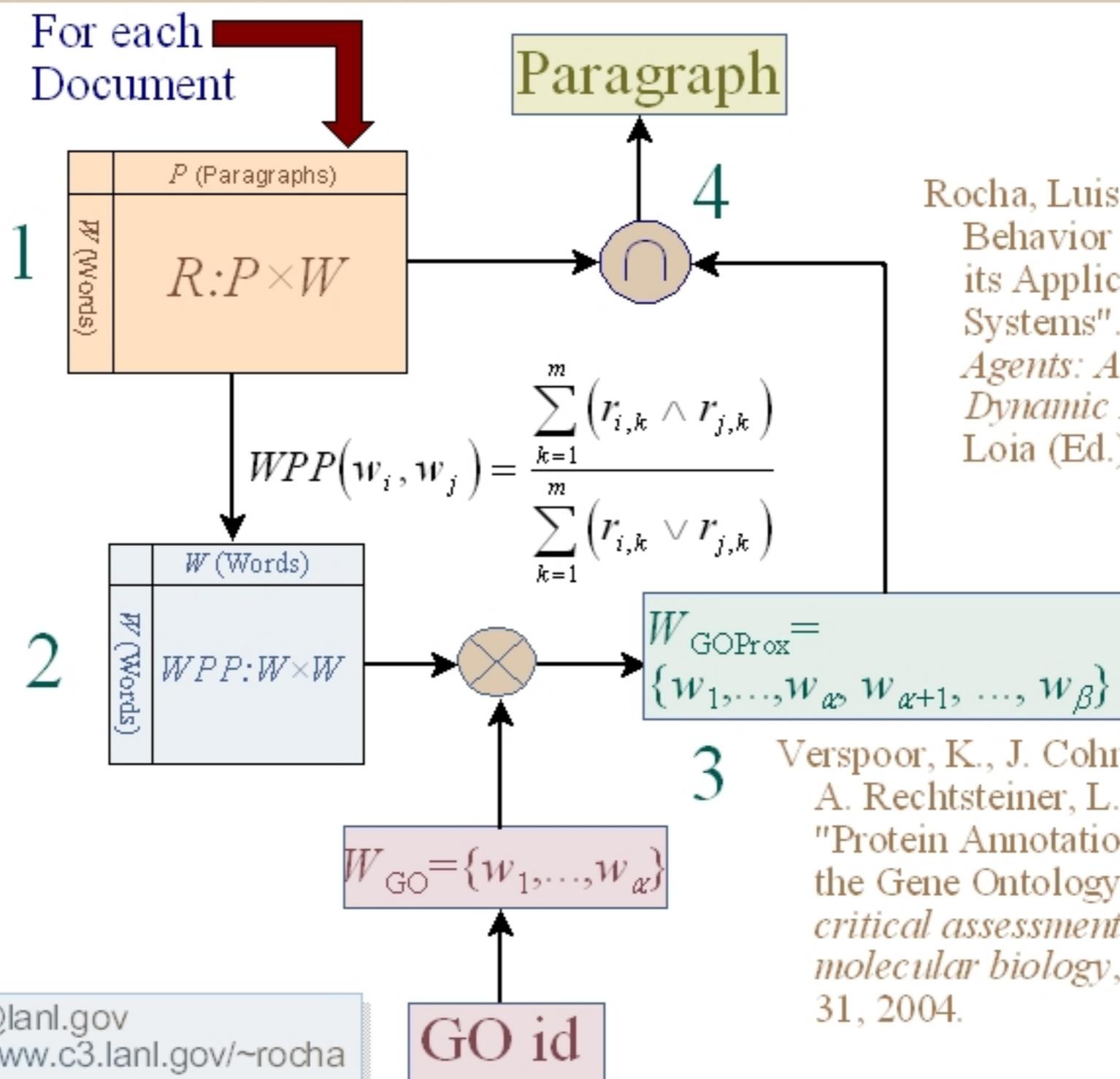
Task 2

- Given a document, discover the portion of text most appropriate to annotate the protein's function, and produce appropriate Gene Ontology node for annotation
 - ▶ Learning set: triplets (protein, document, GO id)
 - ▶ Test set: documents

Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2004]. "Protein Annotation as Term Categorization in the Gene Ontology". *EMBO Workshop: A critical assessment of text mining methods in molecular biology*, Granada, Spain, March 28-31, 2004.



based on proximity measure



Rocha, Luis M. [2002]. "Semi-metric Behavior in Document Networks and its Application to Recommendation Systems". In: *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. V. Loia (Ed.) IOS Press, pp. 137-163.

3 Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2004]. "Protein Annotation as Term Categorization in the Gene Ontology". *EMBO Workshop: A critical assessment of text mining methods in molecular biology*, Granada, Spain, March 28-31, 2004.

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

GO id

Luis Rocha
 2004

Task 2.1 Results

Proximity-based run



User, Run	# results	"perfect"	"generally"
4, 1	1048	268 (25.57%)	74 (7.06%)
5, 1	1053	166 (15.76%)	77 (7.31%)
5, 2	1050	166 (15.81%)	90 (8.57%)
5, 3	1050	154 (14.67%)	86 (8.19%)
7, 1	1057	272 (25.73%)	154 (14.57%)
7, 2	1864	43 (2.31%)	40 (2.15%)
7, 3	1703	66 (3.88%)	40 (2.35%)
9, 1	251	125 (49.80%)	13 (5.18%)
9, 2	70	33 (47.14%)	5 (7.14%)
9, 3	89	41 (46.07%)	7 (7.87%)
10, 1	45	36 (80.00%)	3 (6.67%)
10, 2	59	45 (76.27%)	2 (3.39%)
10, 3	64	50 (78.12%)	4 (6.25%)
14, 1	1050	303 (28.86%)	69 (6.57%)
15, 1	524	59 (11.26%)	28 (5.34%)
15, 2	998	125 (12.53%)	69 (6.91%)
17, 1	413	83 (20.10%)	19 (4.60%)
17, 2	458	7 (1.53%)	0 (0.00%)
20, 1	1048	301 (28.72%)	57 (5.44%)
20, 2	1050	280 (26.72%)	60 (5.73%)
20, 3	1050	239 (22.76%)	59 (5.62%)

Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2004]. "Protein Annotation as Term Categorization in the Gene Ontology". *EMBO Workshop: A critical assessment of text mining methods in molecular biology*, Granada, Spain, March 28-31, 2004.

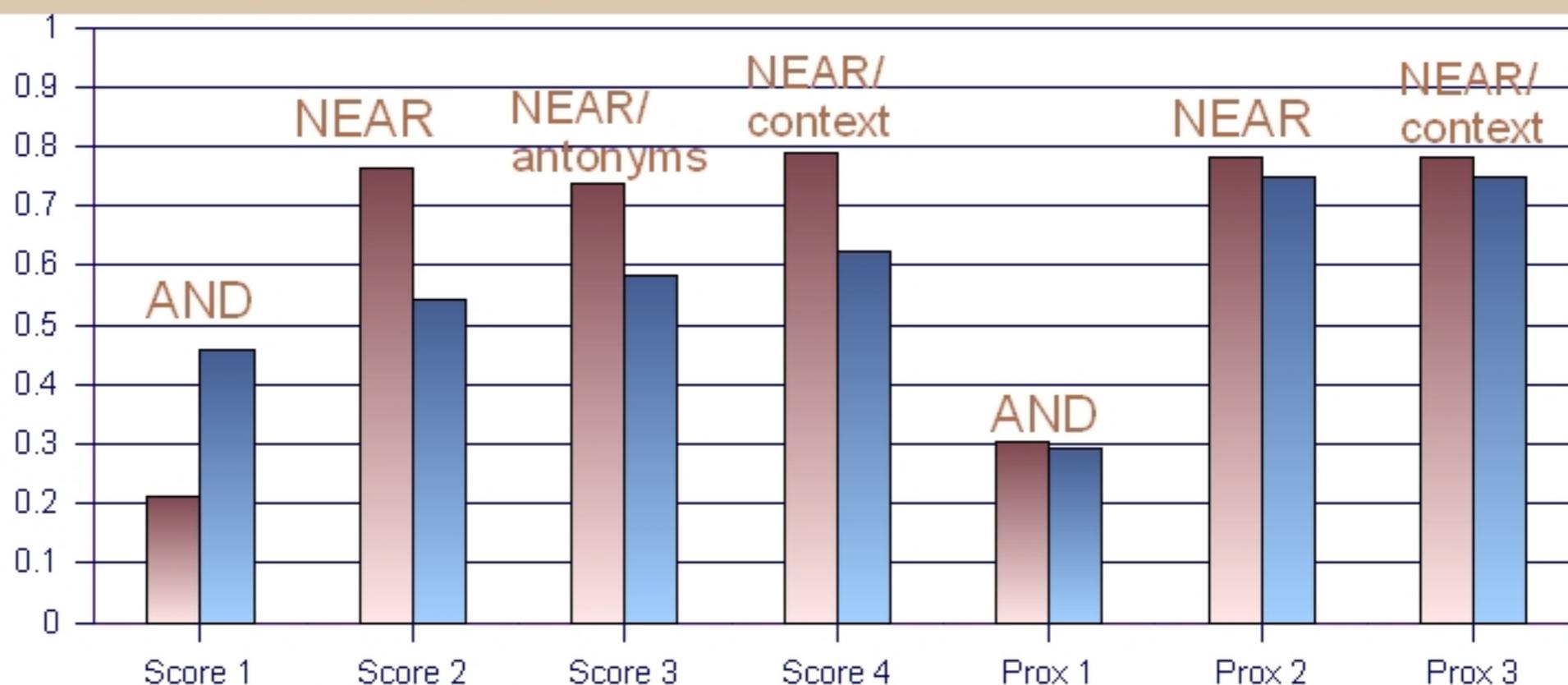
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



Luis Rocha
2004



values obtained with FaCSO



Precision: probability that an identified association is relevant

Recall: probability that an association has been identified given that it is relevant

$$precision = \frac{||\{relevant\} \cap \{retrieved\}||}{||\{retrieved\}||}$$

$$recall = \frac{||\{retrieved\} \cap \{relevant\}||}{||\{relevant\}||}$$

Rocha, Luis M. and Andreas Rechtsteiner [2003]. "Fast Cheap and Synthetic Oracle (FACSO): Proximity Measures to capture Expert Knowledge in the Bibliome". *Pacific Symposium on Biocomputing 2003*.

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

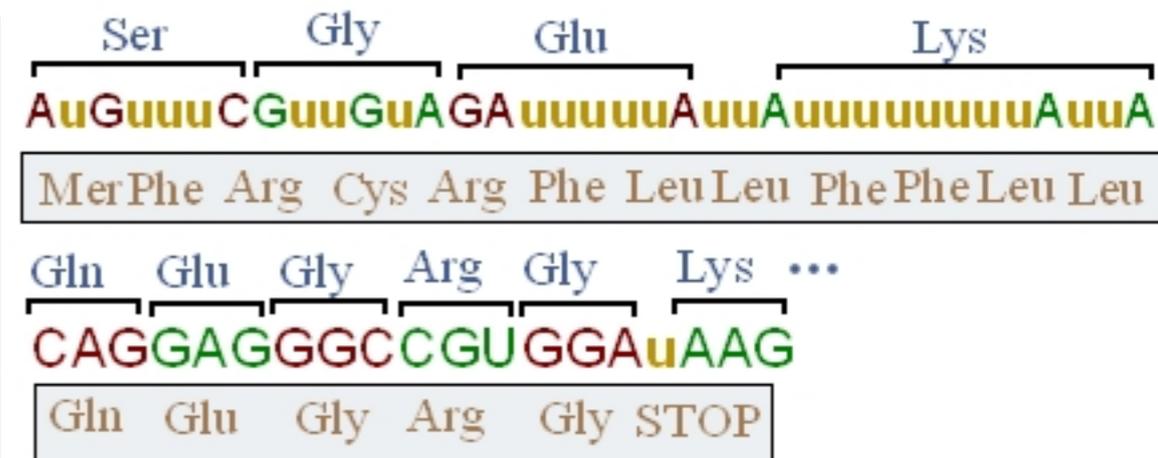
Luis Rocha
2004



artificial evolutionary models with genotype edition

■ RNA Editing

- ▶ Posttranscriptional alteration of mRNA
 - Usually performed by non-protein coding RNA
- ▶ U-Insertion/deletion, C-insertion, C to U substitution, etc.
- ▶ **Control of Developmental Processes** from environmental cues
 - In Trypanosomes, Evolution of parasites, Neural receptor channels in rats.
 - Rats with ADAR1 knocked out, do not develop past first embryonic stages
- ▶ Developing evolutionary models to investigate the evolutionary advantages of genotype editing and to produce more robust GA



Rocha, Luis M. [1995]. "Contextual Genetic Algorithms: Evolving Developmental Rules." *Lecture Notes in Artificial Intelligence*. 929, pp. 368-382.

Rocha, Luis M. [1998]. "Selected Self-Organization and the Semiotics of Evolutionary Systems." In: *Evolutionary Systems: Biological and Epistemological Perspectives on Selection and Self-Organization*. S. Salthe, G. Van de Vijver, and M. Delpo (eds.). Kluwer Academic Publishers, pp. 341-358.

Huang, Chien-Feng and Luis M. Rocha [2003]. "Exploration of RNA Editing and Design of Robust Genetic Algorithms". *2003 Congress on Evolutionary Computation (CEC)*, Canberra, Australia, December 2003. R.Sarker et al (Eds). IEEE Press, pp. 2799-2806.

Huang, Chien-Feng and Luis M. Rocha [2004]. "A Systematic Study of Genetic Algorithms with Genotype Editing". *GECCO 2004*. In Press.

Rocha, Luis M. and Chien-Feng Huang [2004]. "The Role of RNA Editing in Dynamic Environments." *Artificial Life 9*. In Press.

Rocha, Luis M. and Chien-feng Huang [2004]. "An agent-based model of genotype editing". *8th International Conference on Parallel Problem Solving from Nature (PPSN VIII)*. Submitted.

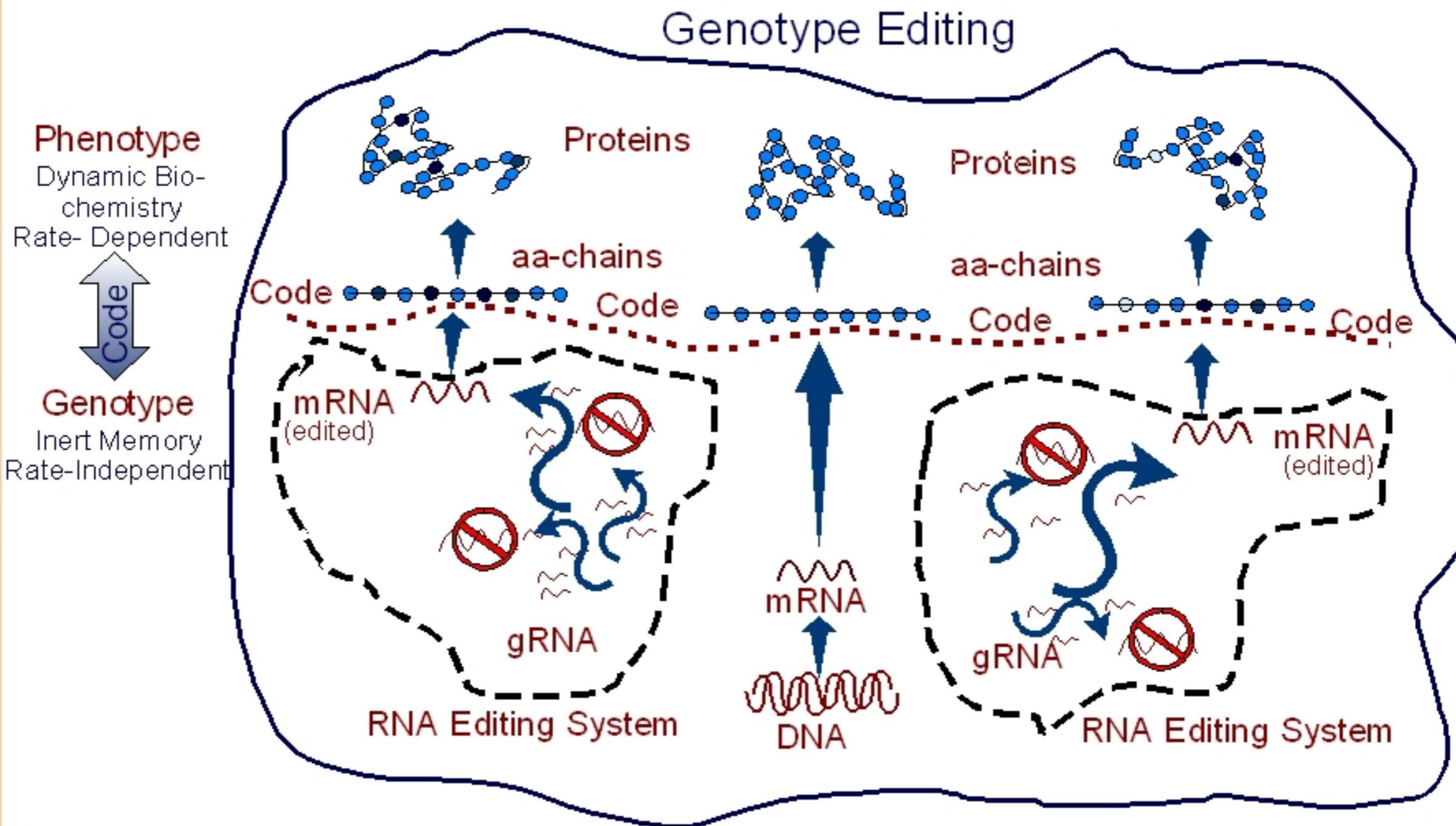
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004



RNA editing acts on genotype

based on guide-RNAs



Luis Rocha
2004

Phenotype
Dynamic Bio-chemistry
Rate-Dependent



Genotype
Inert Memory
Rate-Independent



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



genetic algorithms with edition

editing operators

Luis Rocha
2004

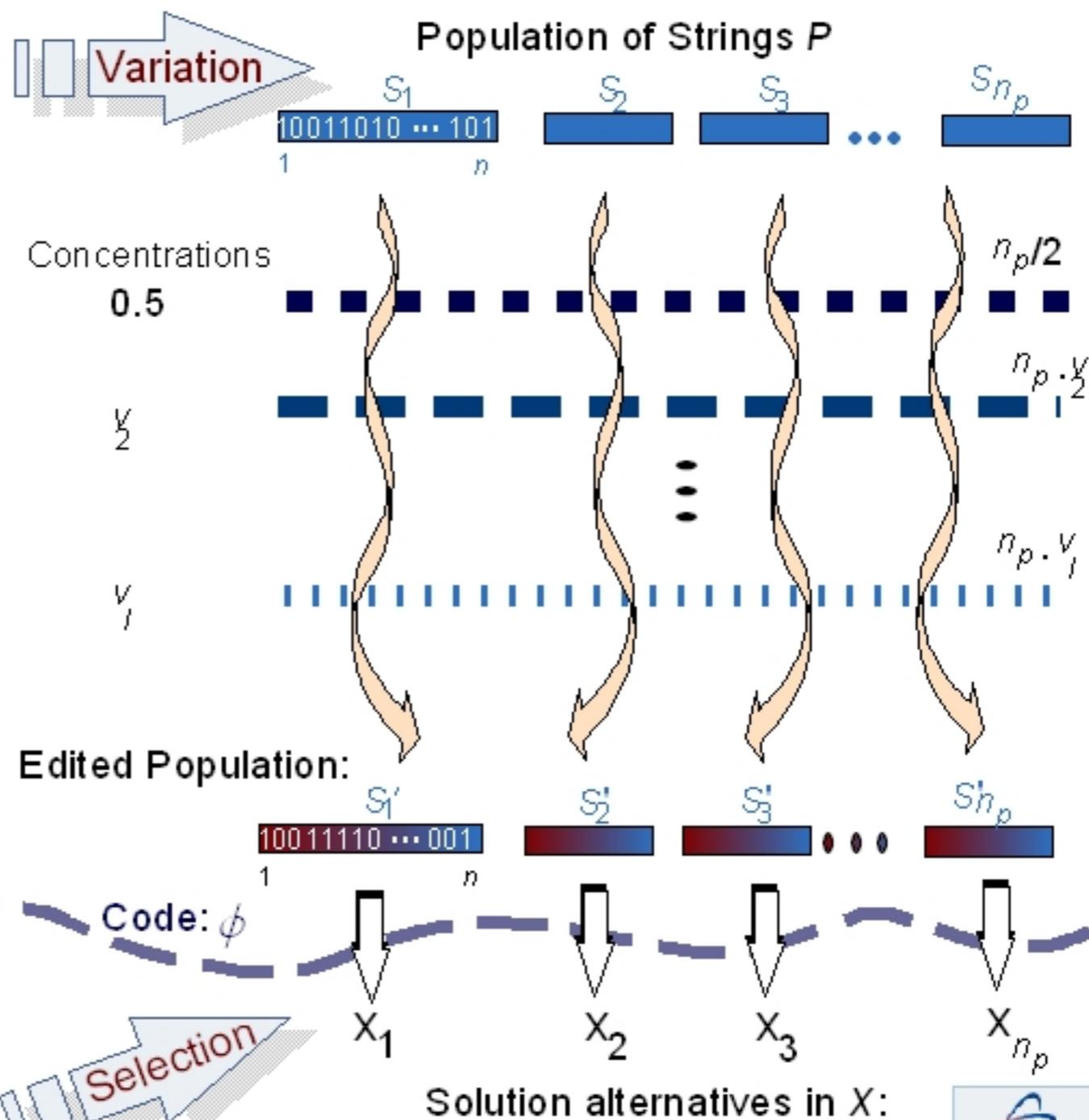
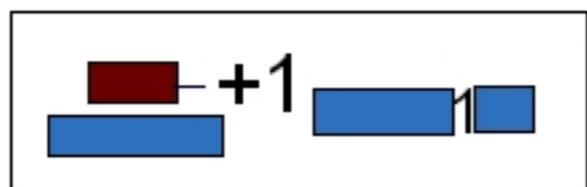
Family of l editors $(\mathcal{E}, \mathcal{F})$

$(\underbrace{1^*01\dots*0}_m, \text{Add}_1)$

$(\underbrace{}_{E_2}, f_2)$

$(\underbrace{}_{E_l}, f_l)$

Ministrings functions



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

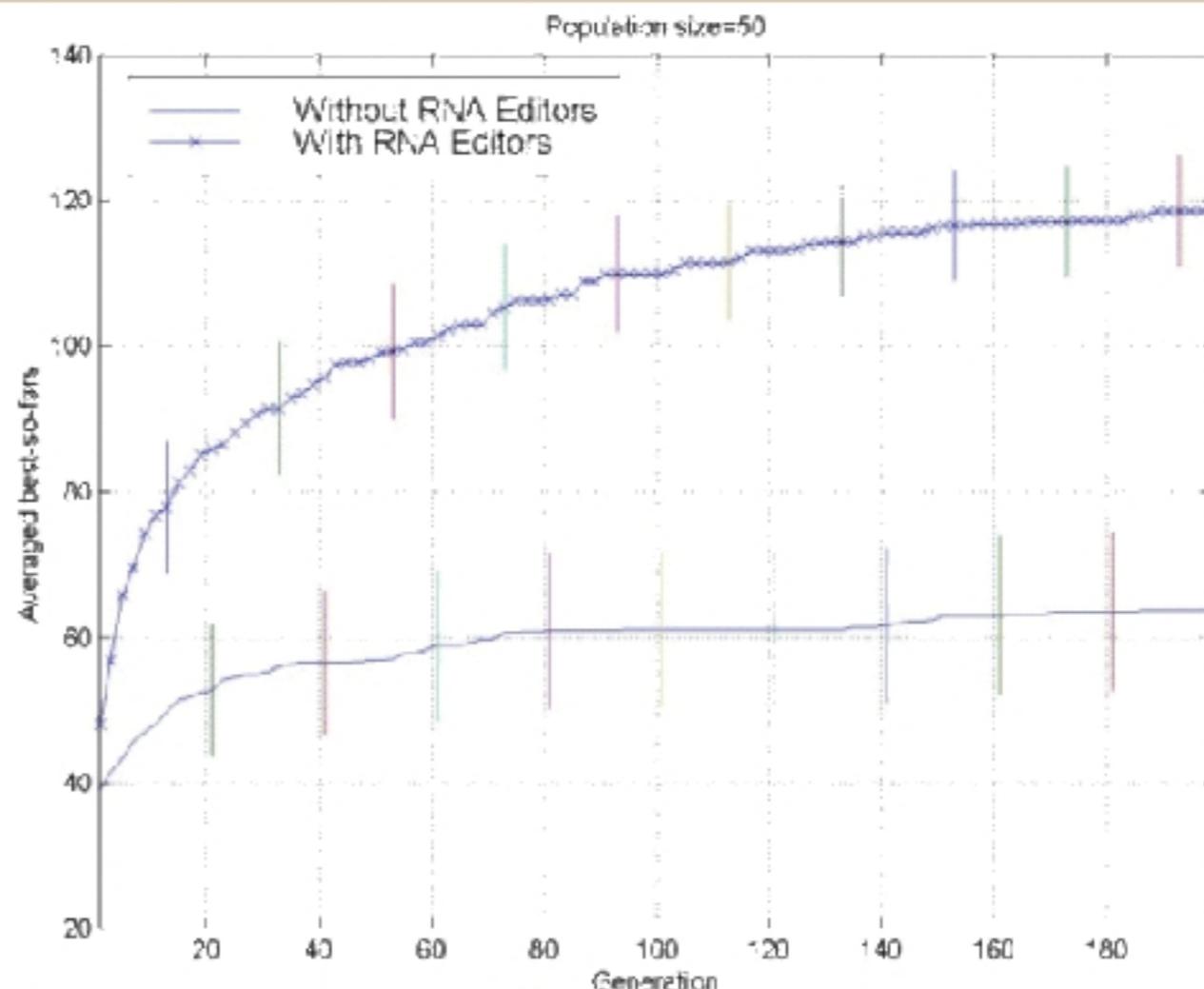
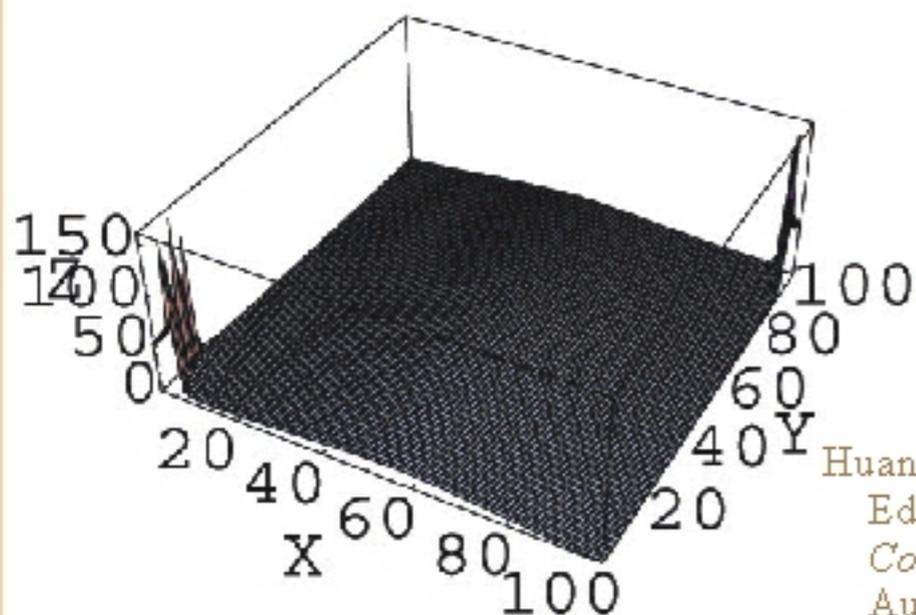


optimal control test problem

effects of genotype edition (static environment)

■ Optimal control problems

- ▶ Existence of multiple local optima in the area of interest.
- ▶ Premature convergence.
 - Height of the hill is much lower than the spikes, but it occupies most of the search space
 - Most of the population individuals attracted to the hilltop



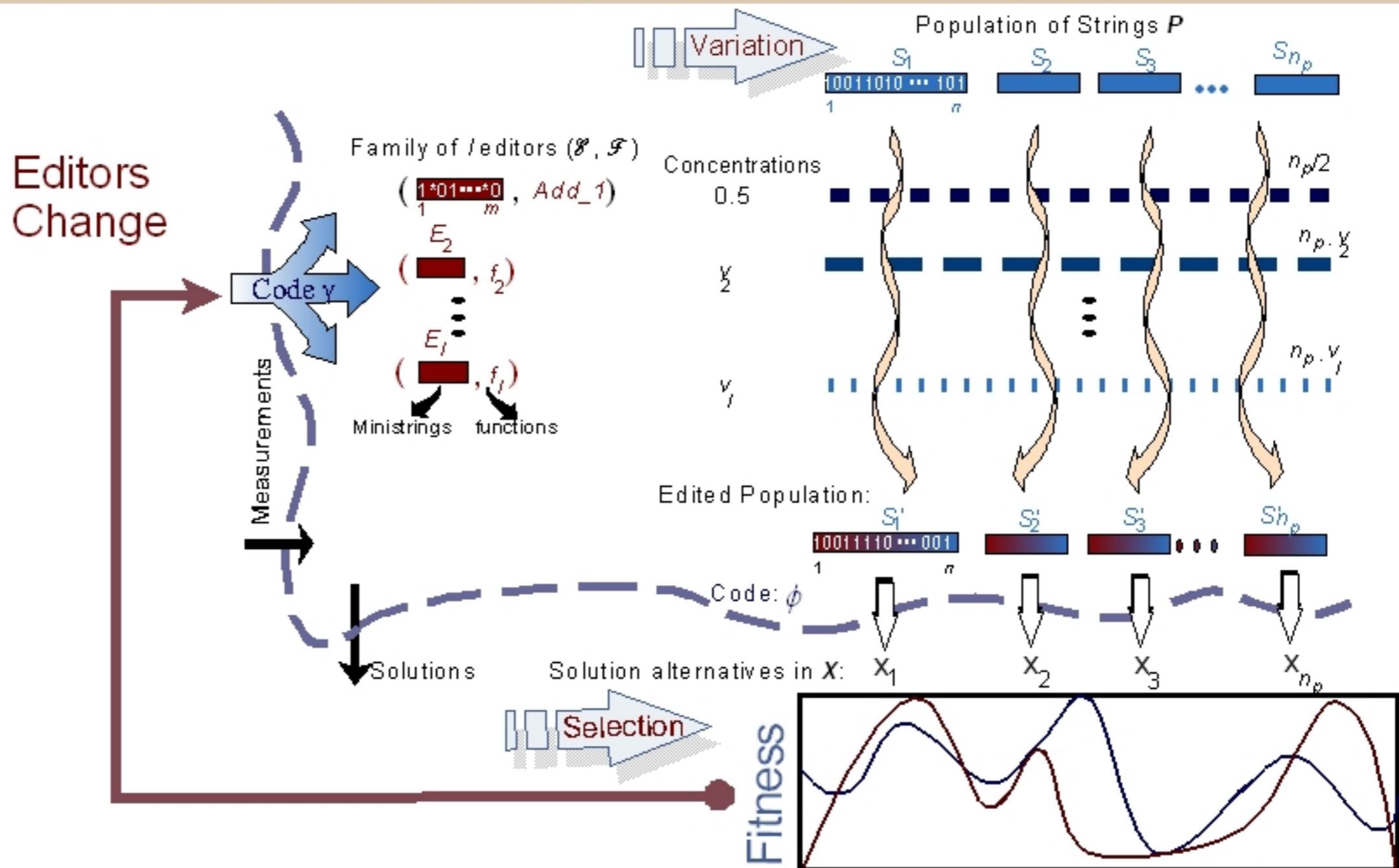
Huang, Chien-Feng and Luis M. Rocha [2003]. "Exploration of RNA Editing and Design of Robust Genetic Algorithms". *2003 Congress on Evolutionary Computation (CEC)*, Canberra, Australia, December 2003. R.Sarker et al (Eds). IEEE Press, pp. 2799-2806.

Huang, Chien-Feng and Luis M. Rocha [2004]. "A Systematic Study of Genetic Algorithms with Genotype Editing". *GECCO 2004*. In Press.

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

contextual genetic algorithms

linking editor concentrations to environment: phenotypic plasticity



Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

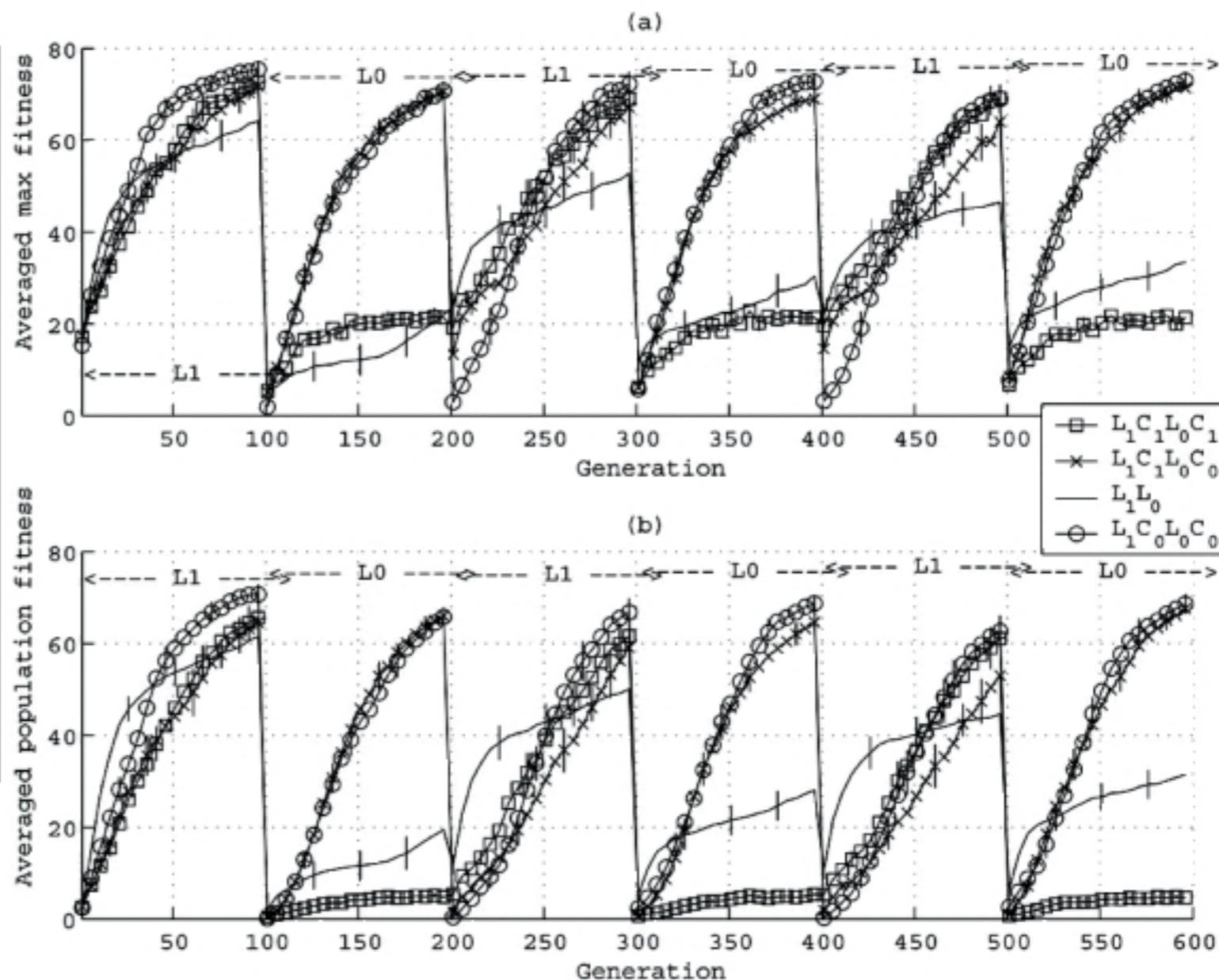
Environmental demands change



oscillating fitness functions

period: 50 generations (mean fitness)

- **Fitness oscillates**
 - ▶ Every 100 generations
 - ▶ Two royal roads
- **Genotype edition can help in two ways:**
 - ▶ Discovering editors which facilitate adaptation in both environments
 - ▶ Alternating editor concentrations for greater phenotypic plasticity



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Rocha, Luis M. and Chien-Feng Huang [2004]. "The Role of RNA Editing in Dynamic Environments." *Artificial Life 9*. In Press.

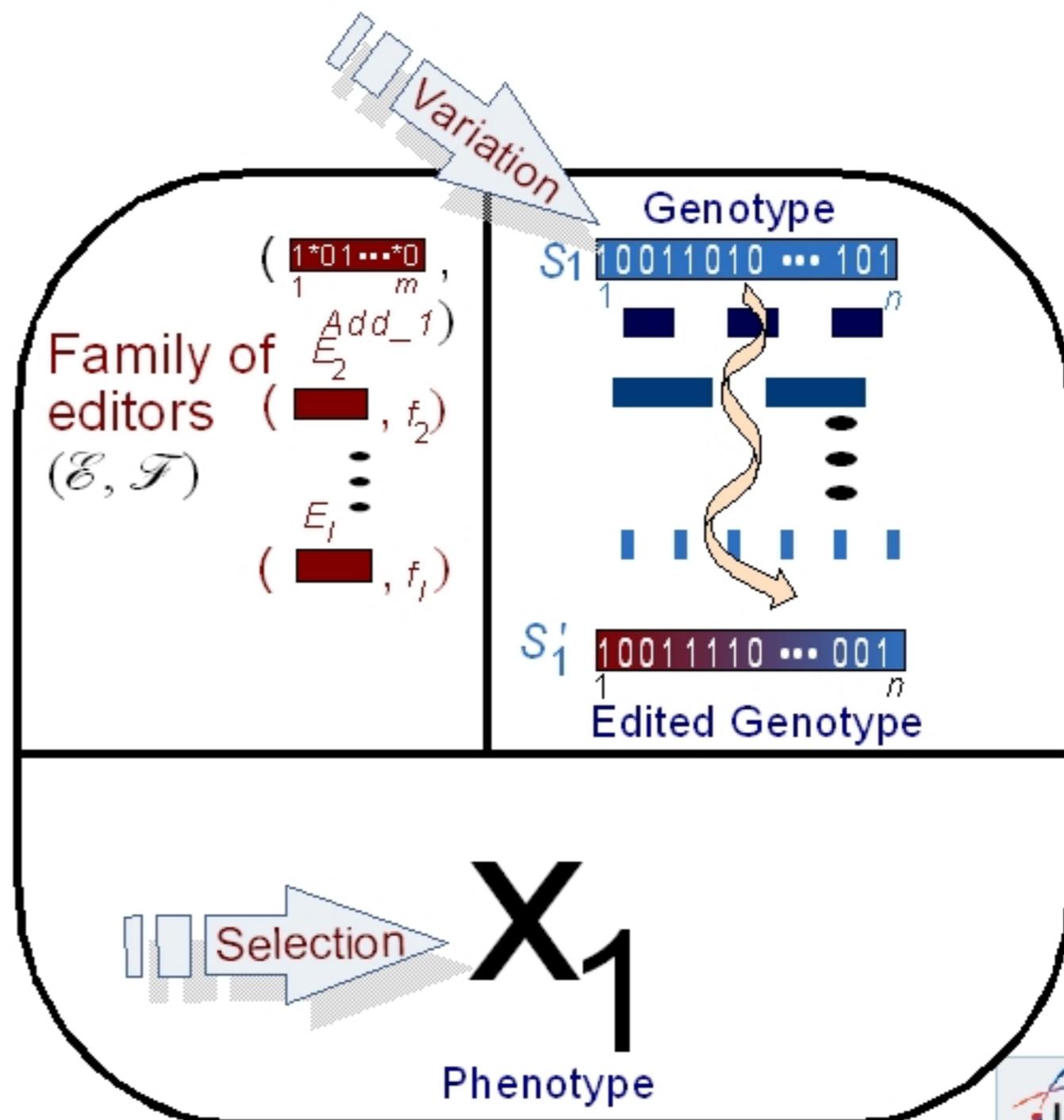
Luis Rocha
2004

agent-based model of genotype editing

heterogeneous edition

- Tested on Royal Road testbed

Rocha, Luis M. and Chien-feng Huang [2004]. "An agent-based model of genotype editing". *8th International Conference on Parallel Problem Solving from Nature (PPSN VIII)*. Submitted.

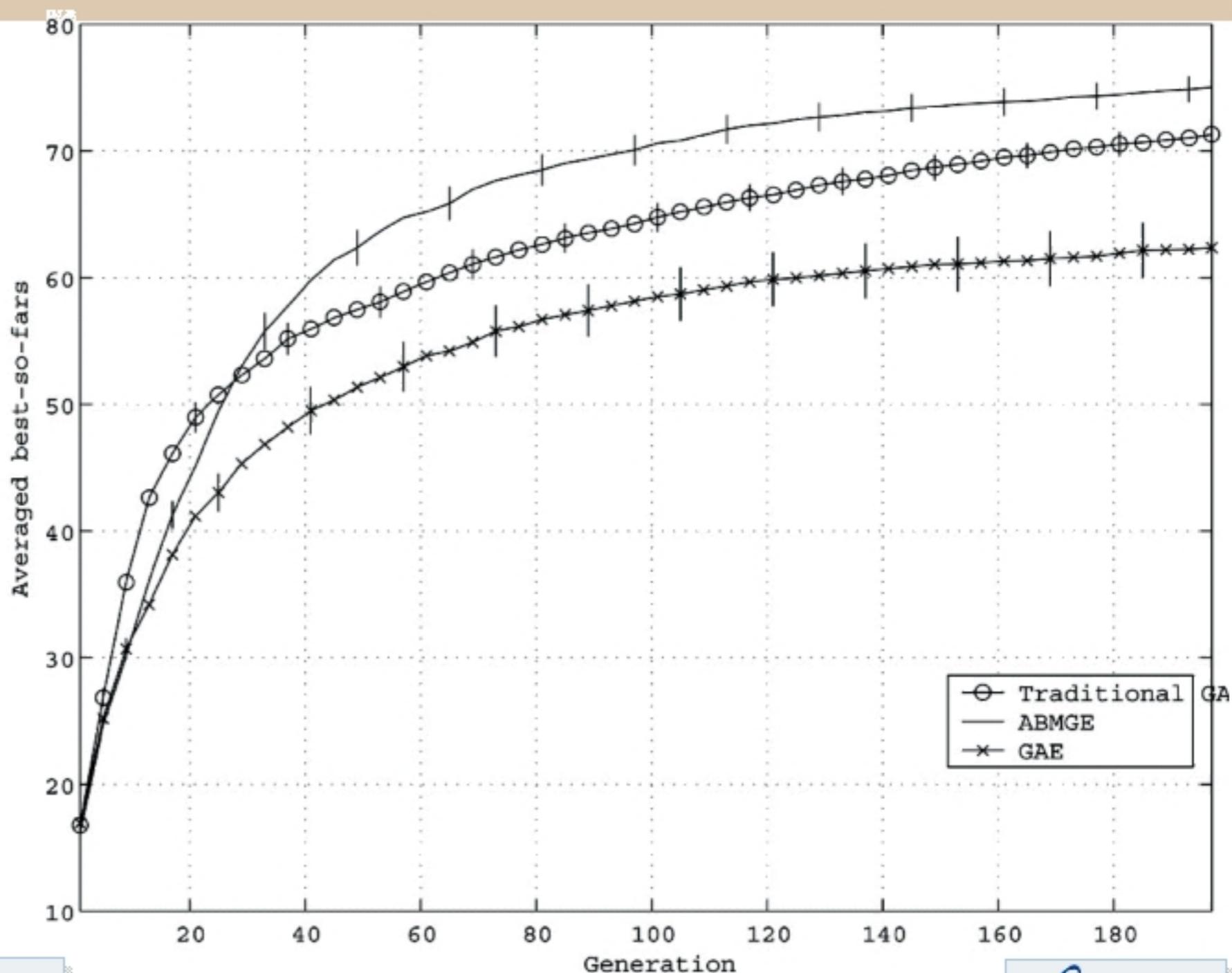


rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004

royal road results

Rocha, Luis M. and Chien-feng Huang [2004]. "An agent-based model of genotype editing". *8th International Conference on Parallel Problem Solving from Nature (PPSN VIII)*. Submitted.



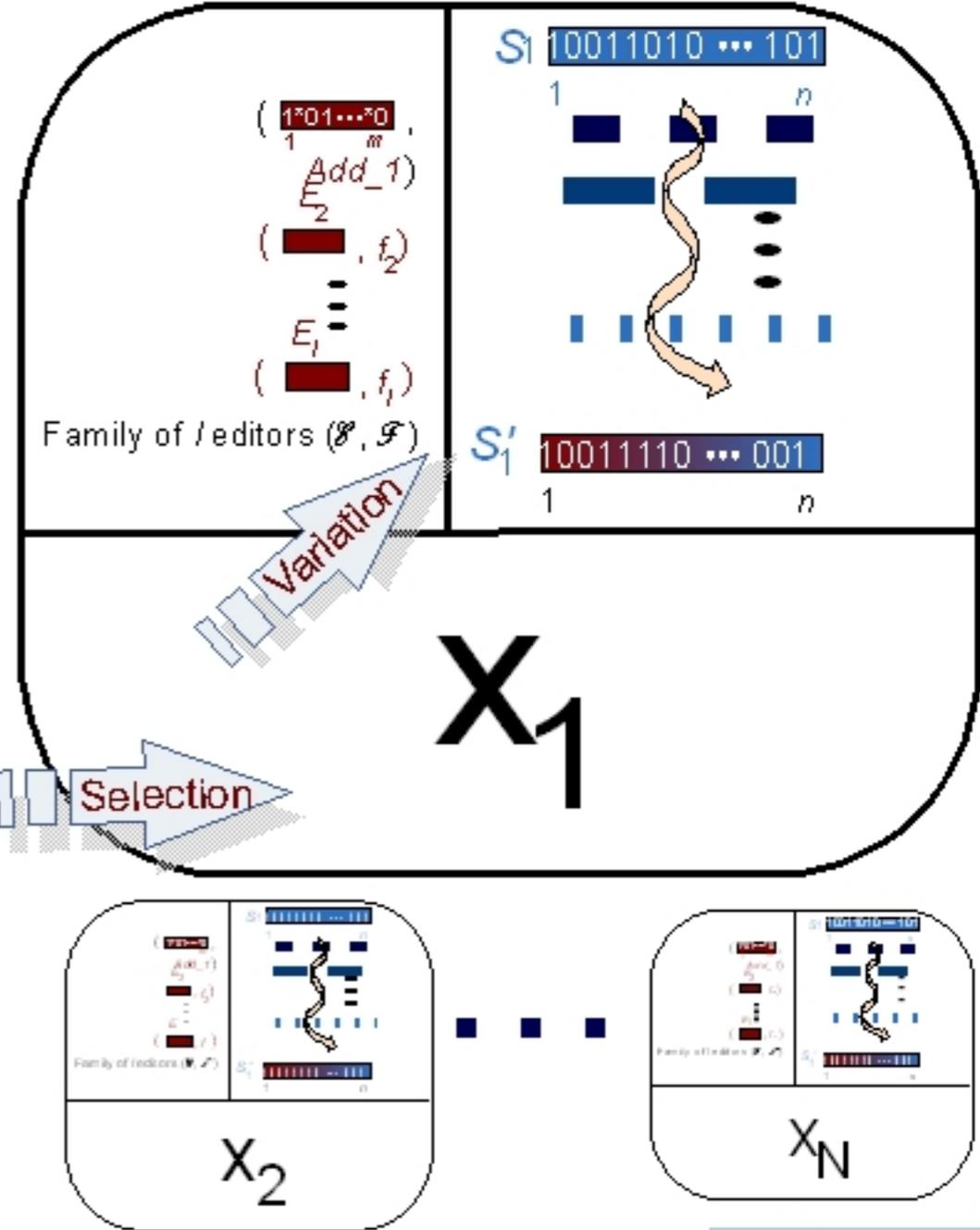
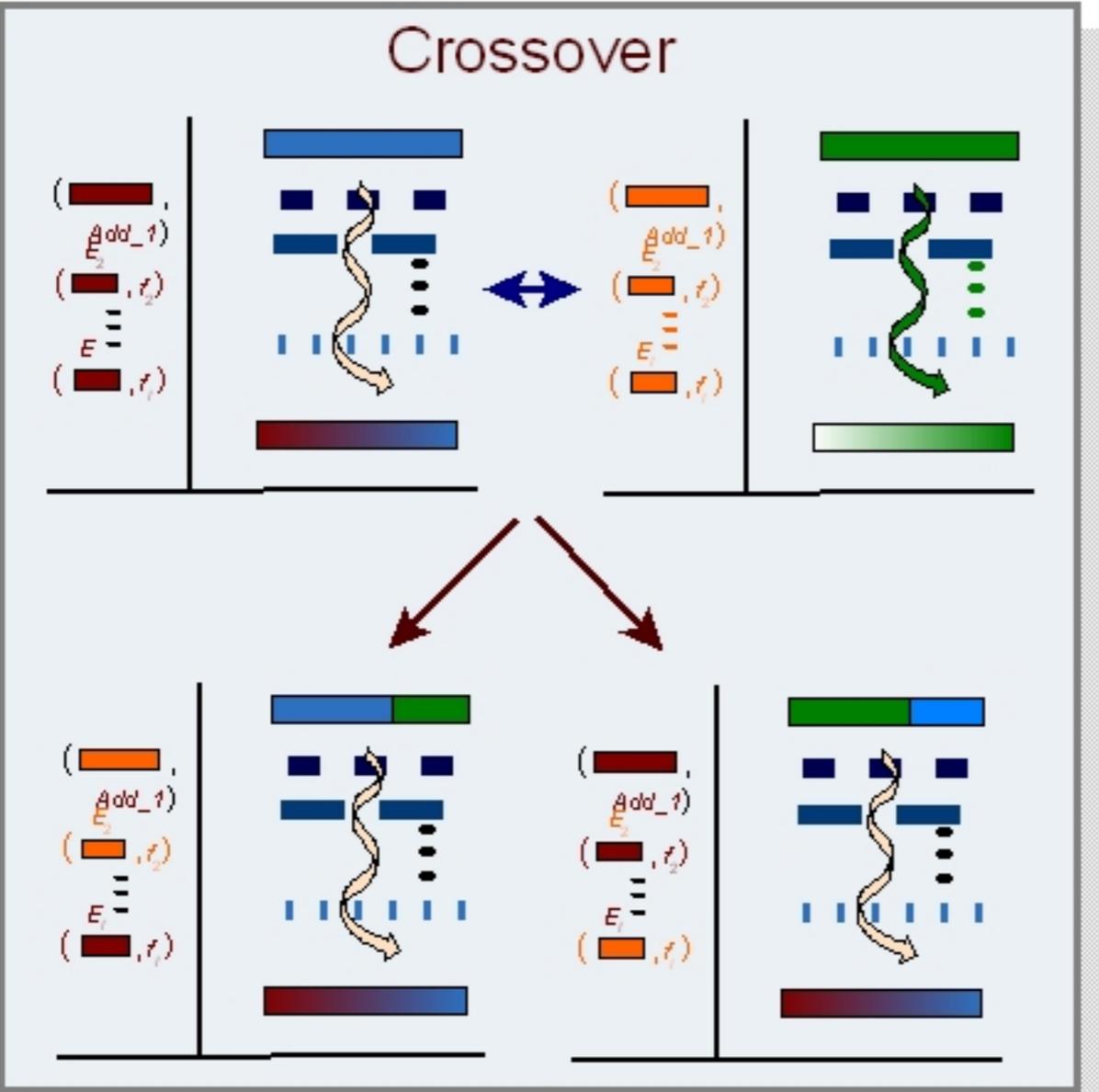
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Luis Rocha
2004



agent-based model of editing

co-evolving editors



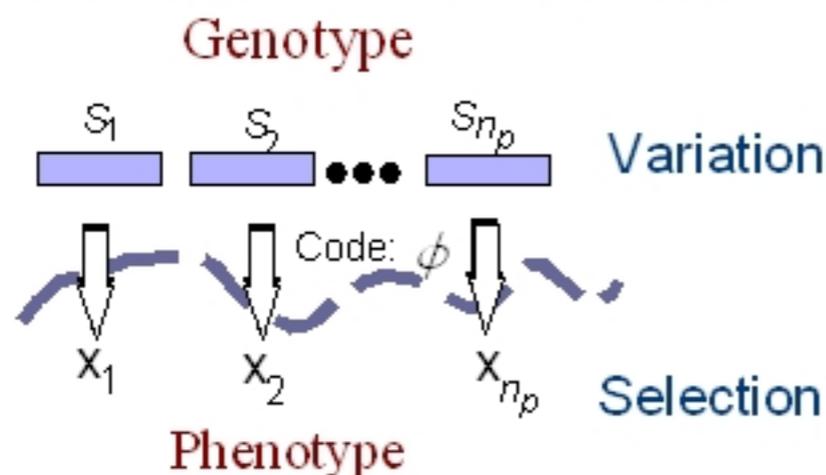
Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

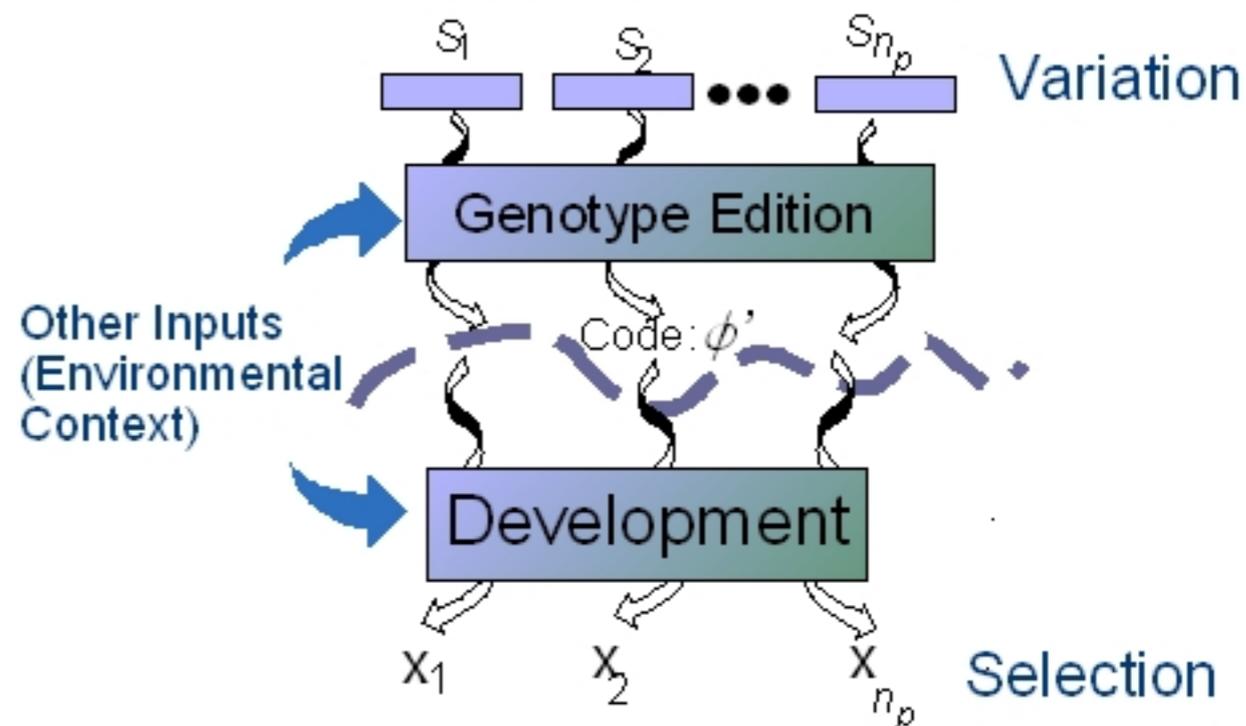


Traditional Genetic Algorithm



Used for *optimization* of solutions for different problems. Uses the syntactic operators of *crossover* and *mutation* for variation of encoded solutions, while selecting best solutions from generation to generation. Holland, 1975; Goldberg, 1989; Mitchell, 1995.

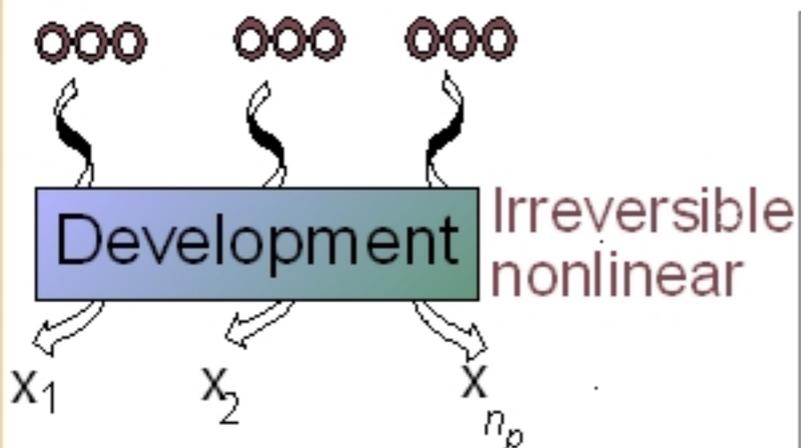
Contextual Genetic Algorithm



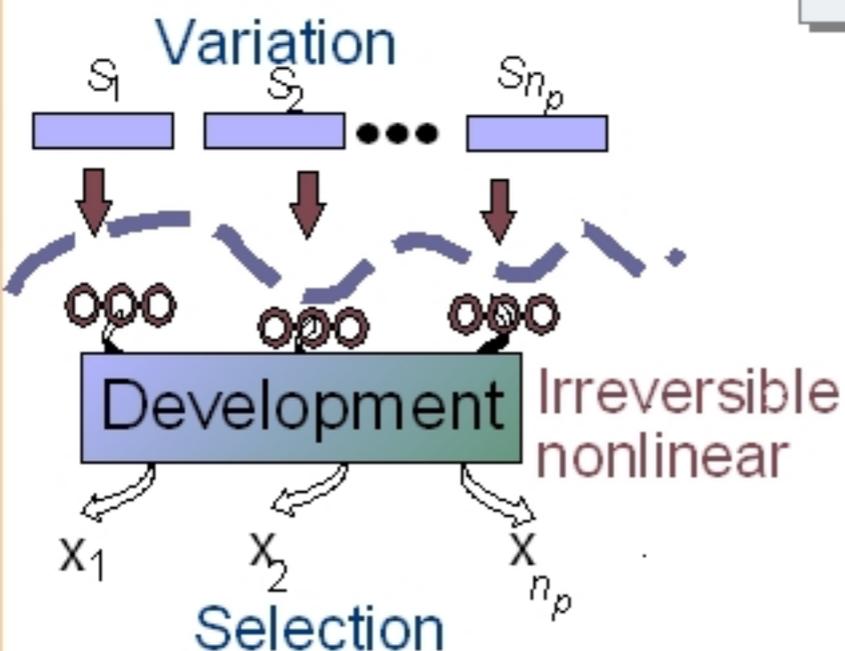
- **Indirect Encoding**
 - ▶ Irreversible
- **Contextual Dependencies**
- **Genotype Edition**
 - ▶ Extra Variation, One-to-many encoding, Co-evolution of contextual gene regulation
- **Development**
 - ▶ Smaller Genotypes, Self-Organization

simulations of evolving eagents

contrasting coded and uncoded reproduction



Variation and Selection



■ Reproduction via self-inspection (includes template reproduction)

- ▶ Hypothetical reproduction of agents based solely on amino acid chains is possible
 - Instead of a ribosome another set of organic machinery would copy amino acid chains
- ▶ Copy components to assemble new agent
 - Same set of components may produce different agents
 - Lamarckian Reproduction

■ Reproduction via a Genotype/Phenotype code

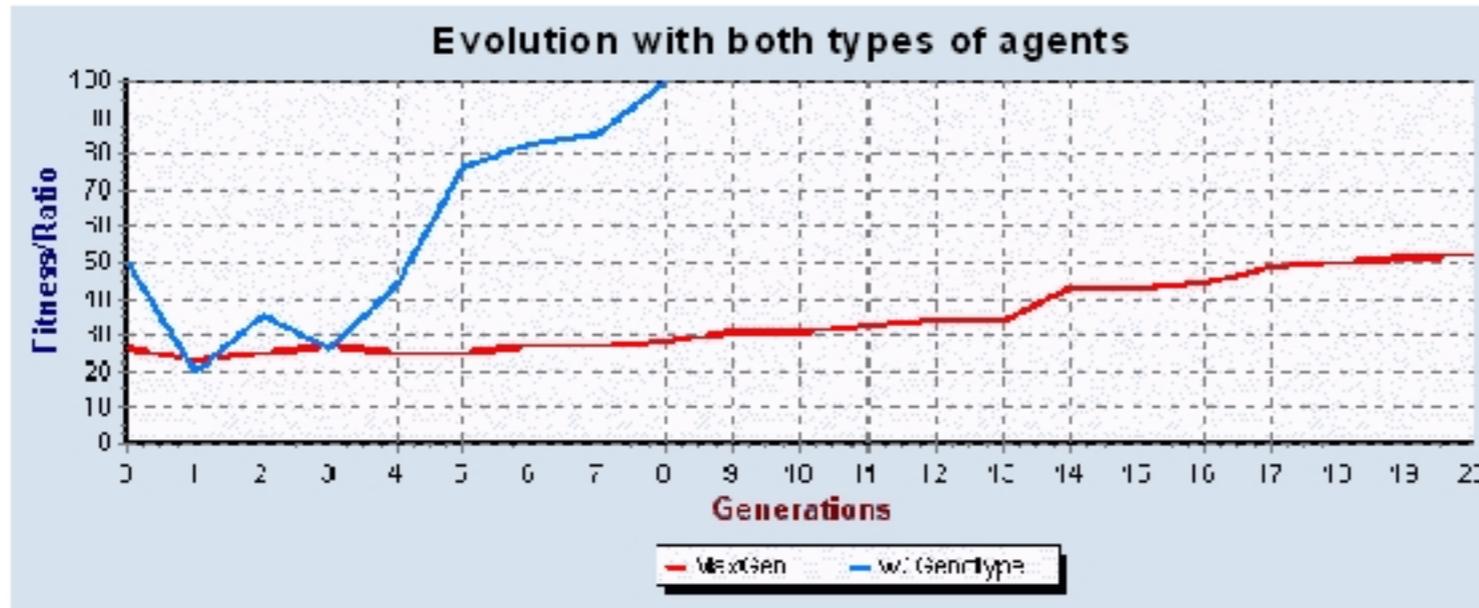
- ▶ Can consistently produce any configuration from a stable, inheritable description
- ▶ Variation on Genotype not on phenotypes
- ▶ Genotype as a stable repository of initial conditions to reliably reproduce evolving agents

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

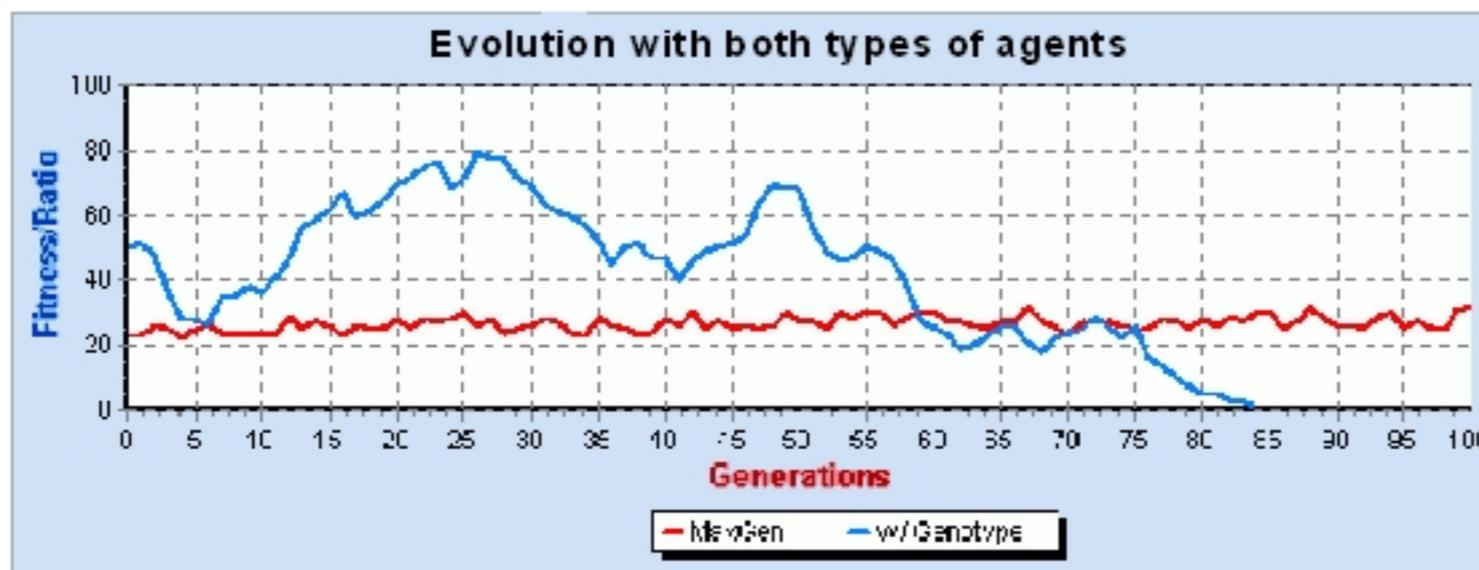
Rocha, L.M [2001]. *Biosystems*.
60, 95-121

coded vs. noncoded agents

simulations of evolutionary potential



Under most conditions and types of evolutionary algorithms, coded agents overtake the population in a small number of generations.



With too much genetic variation, the stability of genotypes is lost, resulting in occasional taking over of the population by noncoded agents.



Luis Rocha
2004

