rocha@lanl.gov

Integrative Technology for a Systems Biology

At the Los Alamos National Laboratory

Luis M. Rocha

CCS3 - Modeling, Algorithms, and Informatics

Los Alamos National Laboratory, MS B256

Los Alamos, NM 87501

e-mail: rocha@lanl.gov or rocha@santafe.edu

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

rocha@lanl.gov

Los Alamos National Laboratory

From the Manhattan to the "Genomes to Life" Project

- U.S. Department of Energy (DOE) Laboratory managed by the University of California.
  - Annual Budget ≈ US$1.200.000.000
  - One of the largest multidisciplinary research institutions in the World.
    - ≈ 6.800 U.C employees plus 2.800 contracted personnel.
    - 1/3 of researchers are Physicists, 1/4 Engineers, 1/6 Chemists and Material Scientistcs. The remainder (1/4), works in Mathematics, Computer and Computational Science, Biology, Geoscience and other disciplines.
    - External scientists (from Academia and Industry), as well as students, come to Los Alamos to work in research projects (basic and applied) developing technology for future applications.

Luis Rocha
2001

rocha@lanl.gov

Systems Biology at LANL

Genomes To Life Program: DOEGenomesToLife.org

- DOE 10 year program on Systems Biology
  - the next step of the Genome Project
  - From whole-genome sequences, build a systemic understanding of complex living systems
  - Systems approach to Computational Biology
  - DOE Mission: produce energy, sequester excess atmospheric carbon that contributes to global warming, clean up environments contaminated from weapons production, protect people from energy byproducts (e.g. radiation) and from the threat of bioterrorism.
  - Interdisciplinary: Biology, Mathematics, Computer and Computational Science, Engineering, Physics, etc.
- 4 Goals:
  - Identify and characterize molecular machines of life
  - Characterize gene regulatory networks
  - Characterize the functional repertoire of complex microbial communities
  - Develop computational methods and capabilities to advance understanding and predict behavior of complex biological systems

Luis Rocha
2001

GENOMES to LIFE

ACCELERATING BIOLOGICAL DISCOVERY

A NEW PROGRAM PROPOSED BY THE U.S. DEPARTMENT OF ENERGY

DNA SEQUENCE DATA FROM GENOME PROJECTS

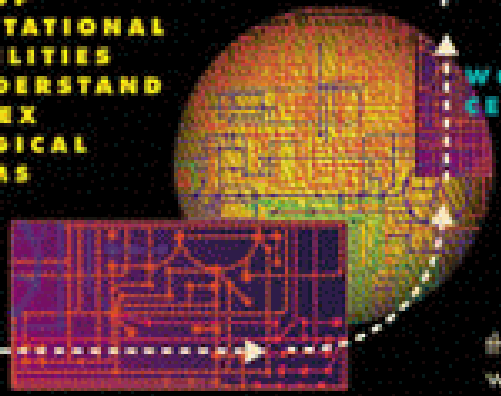
Genes and other DNA sequences contain instructions on how and when to build proteins

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
IDENTIFY PROTEIN MACHINES

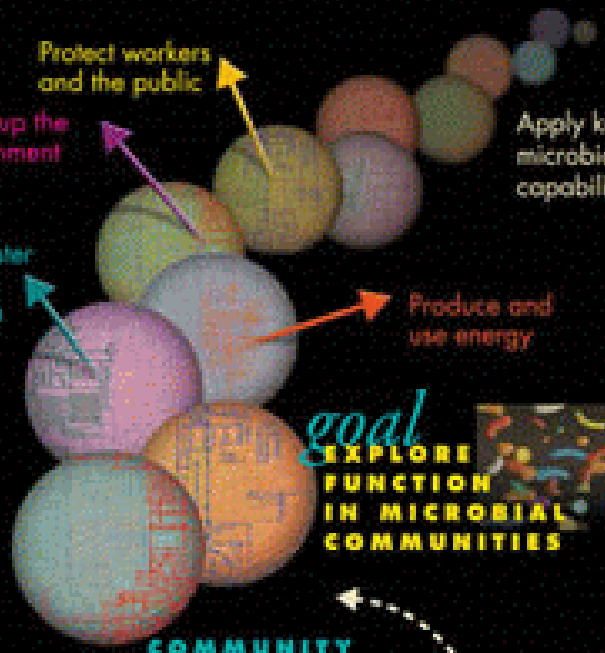
goal
DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS



WORKING CELL

Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to a model of a living cell.

goal
CHARACTERIZE GENE REGULATORY NETWORKS



goal
EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

COMMUNITY OF CELLS

URL DOEGenomesToLife.org

4/01

rocha@lanl.gov

Systems Biology

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 is the methodology used to study *systemhood* not *thinghood* properties in Nature.
  - 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

Luis Rocha
2001

rocha@lanl.gov

Systems Science

Dealing with 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 os 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@lanl.gov

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].
- 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 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.

Luis Rocha
2001

rocha@lanl.gov

Modeling Biological Systems

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

Luis Rocha
2001

rocha@lanl.gov

Why Structural Reductionism is Not Sufficient

Destruction of Dynamical Properties

- Naive Structural Decomposition
  - Breaks an organism into simpler components, gathers information about those, and atempts 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!

rocha@lanl.gov

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]
    - PCA-like, Fourrier Analysis approaches
  - Build IntegrativeTechnology to Disseminate and Utilize Structural Data – for a diverse group of scientists

rocha@lanl.gov

BioInformatics and Computational Biology

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

Luis Rocha
2001

rocha@lanl.gov

Post-Genome informatics

Enabling a Systems Approach to Biology

- Not just support technology but involvement in the systematic, iterative design and analysis of experiments
  - *Functional genomics*: analysis of gene expression patterns at the mRNA and protein levels, as well as analysis of polymorphism, mutation patterns and evolutionary considerations.
  - Where, when, how, and why of gene expression
  - *A*ims to understand biology at the molecular network level using all sources of data: sequence, expression, diversity, etc.
- Grand Challenge: Given a complete genome sequence, reconstruct in a computer the functioning of a biological organism

Luis Rocha
2001

rocha@lanl.gov

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.

Luis Rocha
2001

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

rocha@lanl.gov

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

Luis Rocha
2001

rocha@lanl.gov

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-Performeance Algorithms to allow whole-system Kinetic models

rocha@lanl.gov

Systems Biology

On-going work at LANL (Complex Systems Modeling)

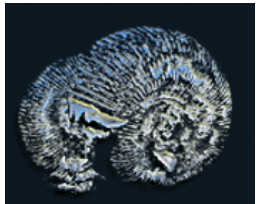
- Data-mining of Functional Global Patterns
  - Discovery of Juxtaposed temporal patterns in GE data (cell-cycle)
    - Comparison between clustering, SVD (PCA), and Gene Shaving. Mapped weaknesses of gene shaving with artificial and real data. Testing better methods for characterization of temporal processes such as Fourier analysis. (Michael Wall, Andreas Rechtsteiner, Deborah Rocha)
    - Association Rules for GE data: Generalized AR into an exhaustive search of itemsets, and inclusion of uncertainty. (Deborah Rocha)
    - Prediction of temporal processes using Klir's Mask Analysis (Cliff, Joslyn, Andreas Rechtsteiner, Deborah Rocha)
- BioKnowledge Systems
  - Representations of Biological Data
  - Latent Databases
  - Collaborative and Recommendation Systems
  - Automated Analysis of Whole Databases of Publications and data-sets

rocha@lanl.gov

Data-Mining of Global Patterns

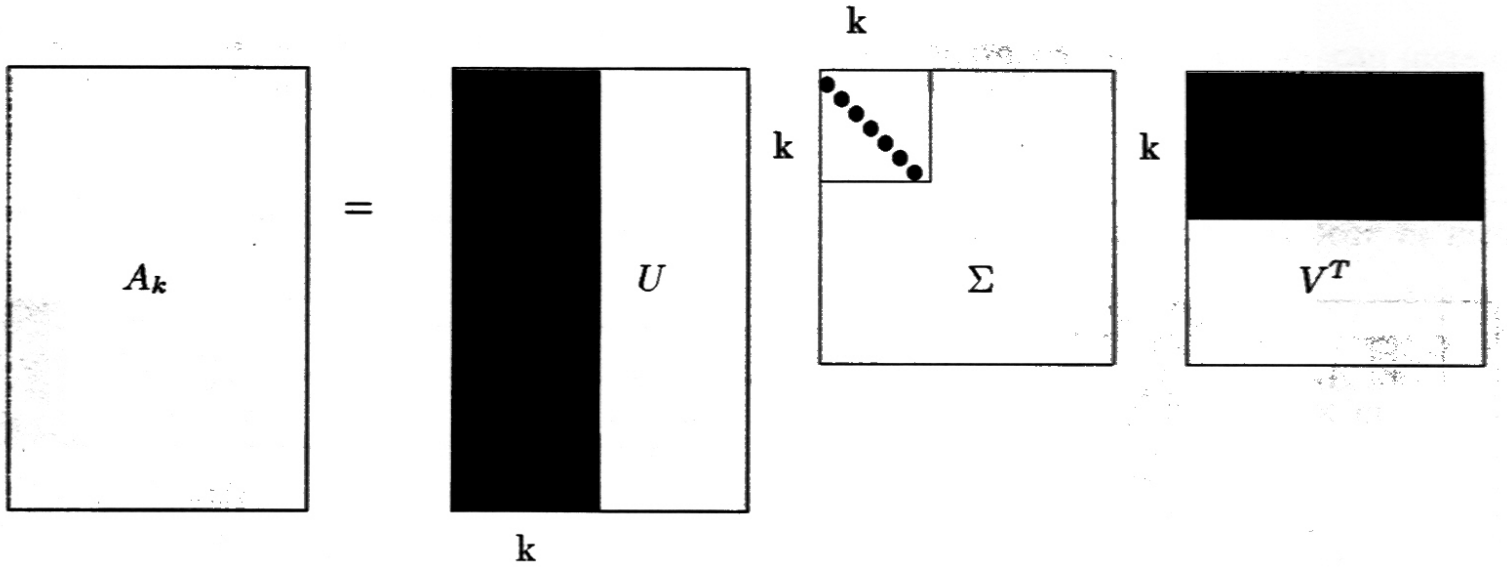
Discovery of Juxtaposed Functional Modes

- 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@lanl.gov

SVD for Gene Expression



$m \times n$ $\quad$ $m \times r$ $\quad$ $r \times r$ $\quad$ $r \times n$

Columns are time steps and rows are genes

Columns of U are eigenarrays (rows are genes)

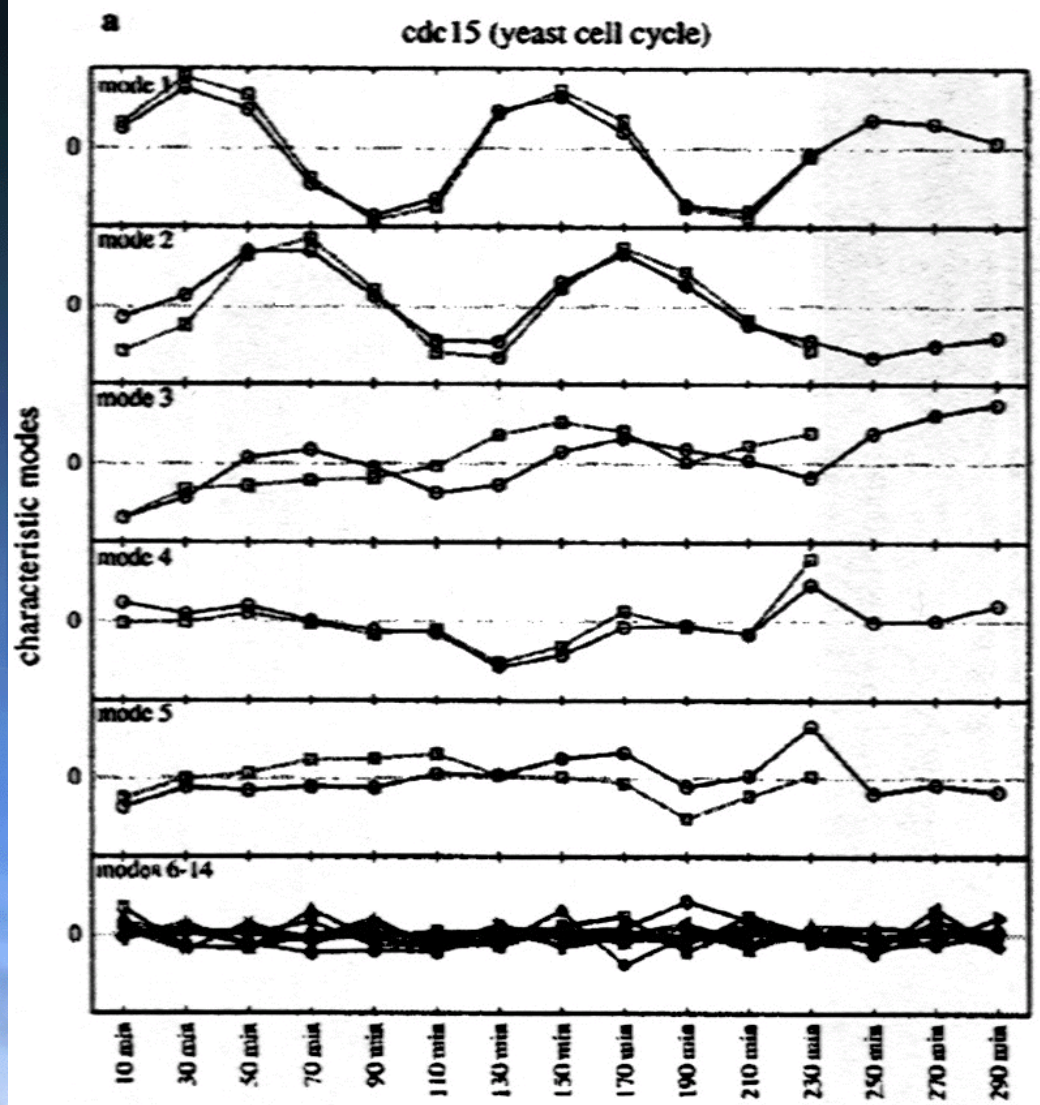
Rows of V^T are eigengenes (colums are time steps)

Luis Rocha
2001

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

rocha@lanl.gov

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)

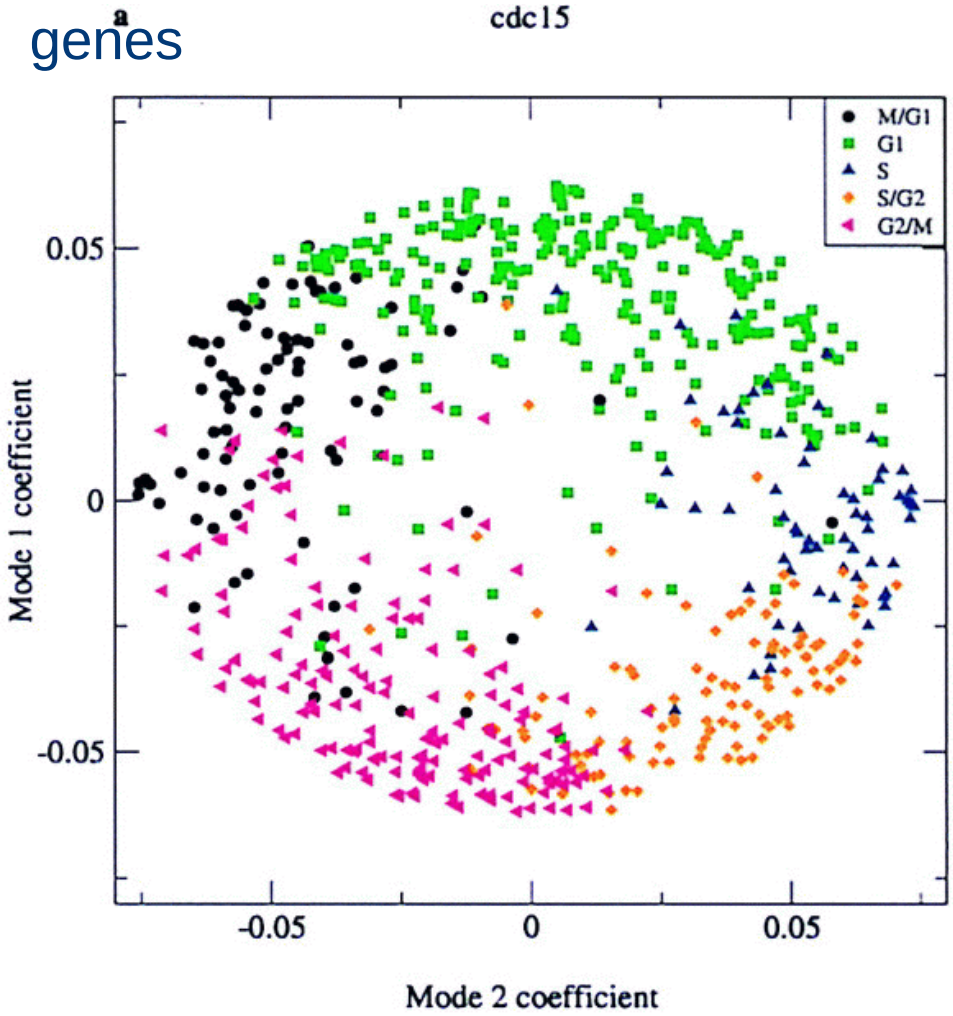
Luis Rocha
2001

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

rocha@lanl.gov

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

Luis Rocha
2001

rocha@lanl.gov

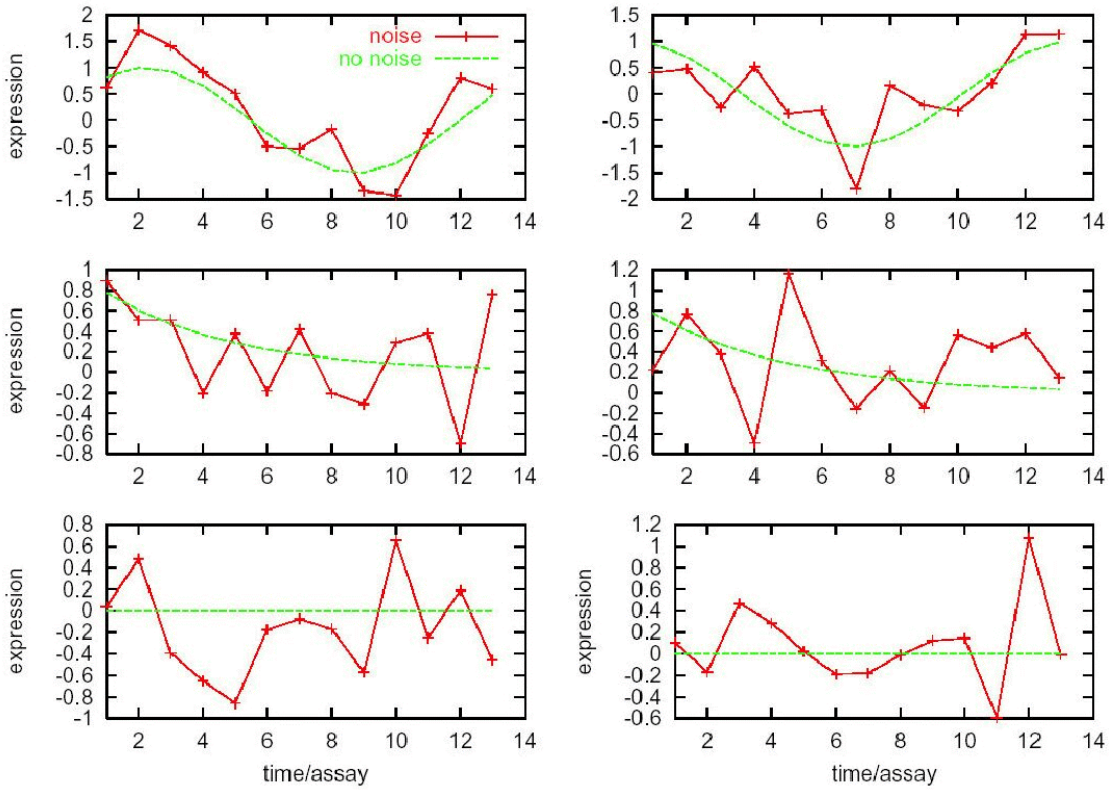
SVD and Functional Decomposition

- Sorting GE data according to the coefficients of genes and arrays in eigengenes and eigenarrays gives a global picture of expression dynamics
  - Genes and arrays are classified into groups of similar regulation and function or similar cellular state and biological phenotype respectively
  - Wall et al [2001], clusters eigenarray coefficients. Better than traditional clustering since genes affected by the same regulator are clustered together irrespective of up or down regulation
- Spectral approaches allow us to filter out the effects of particular eigengenes/eigenarrays
  - Selective discovery of functional patterns
- Aid to the functional simplification necessary for a Systems Biology

rocha@lanl.gov

Discovering Hidden Functional Expression Modes

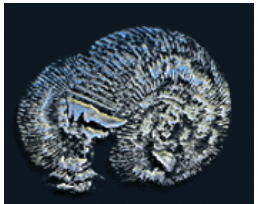
Comparison of SVD Methods with Artificial and Real Data



- Andreas Rechtsteiner
- Artificial data based on yeast cell cycle data.
  - 700 genes with sine wave expression profile
    - Unit amplitude random phase
  - 50 genes exponential decay and 50 genes exponential growth
  - 5200 random genes

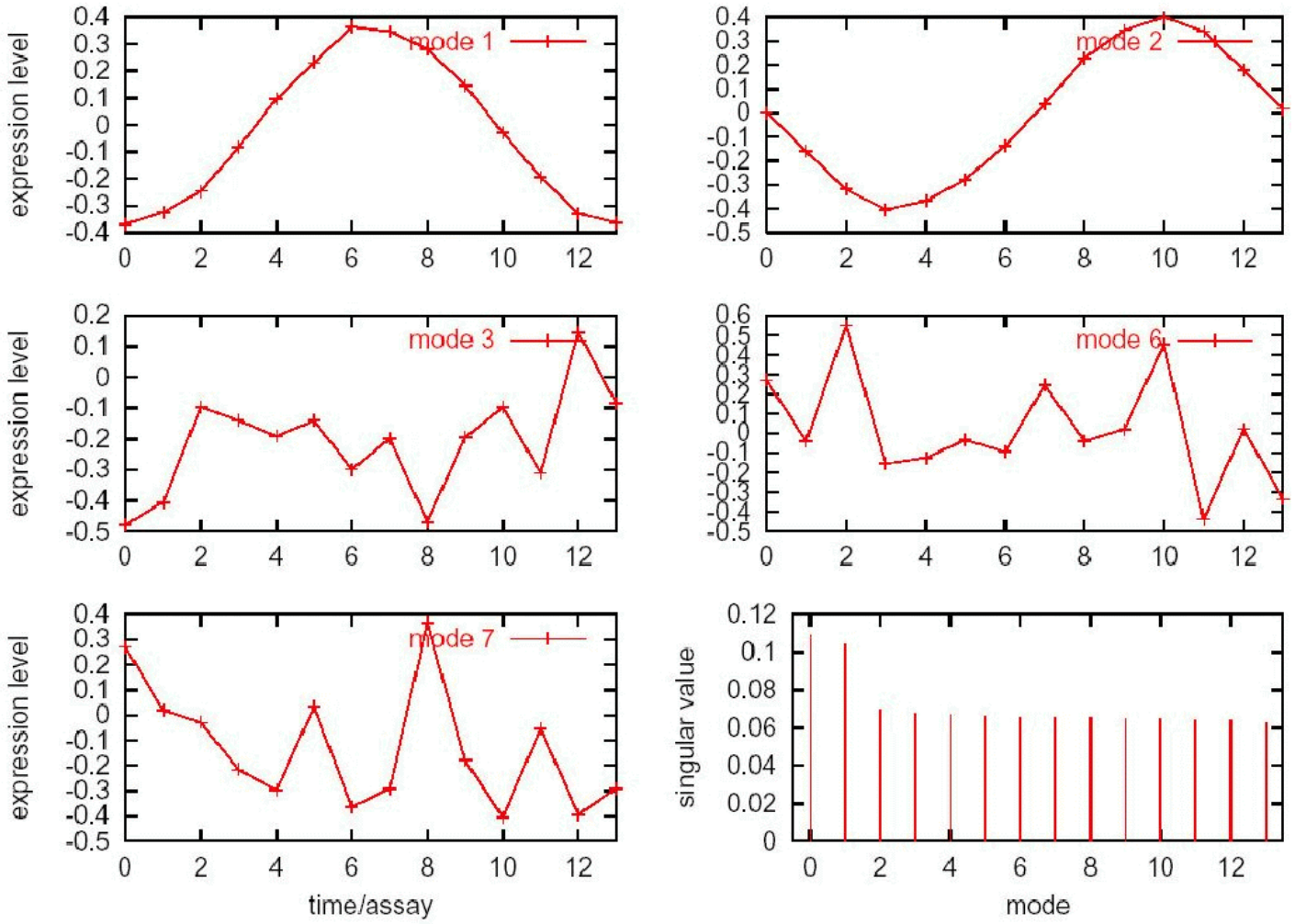
Luis Rocha
2001

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



rocha@lanl.gov

SVD of Artificial Data Set



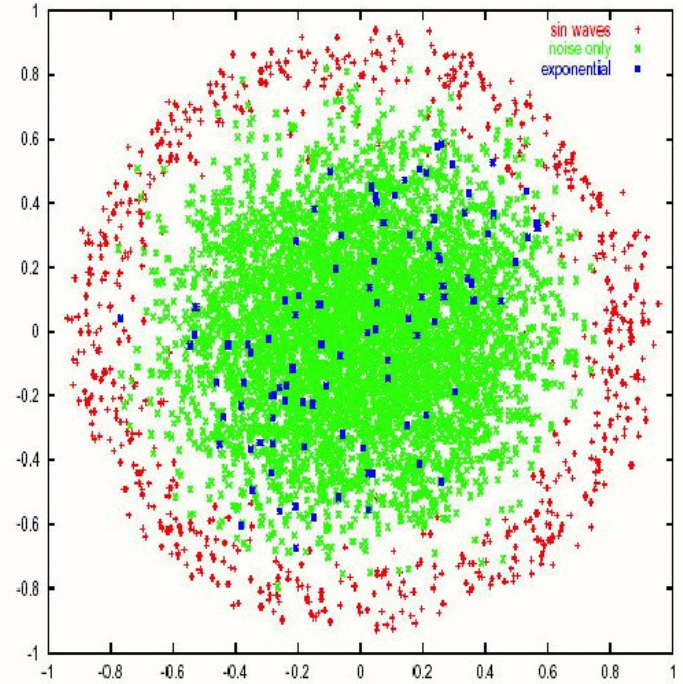
Luis Rocha
2001

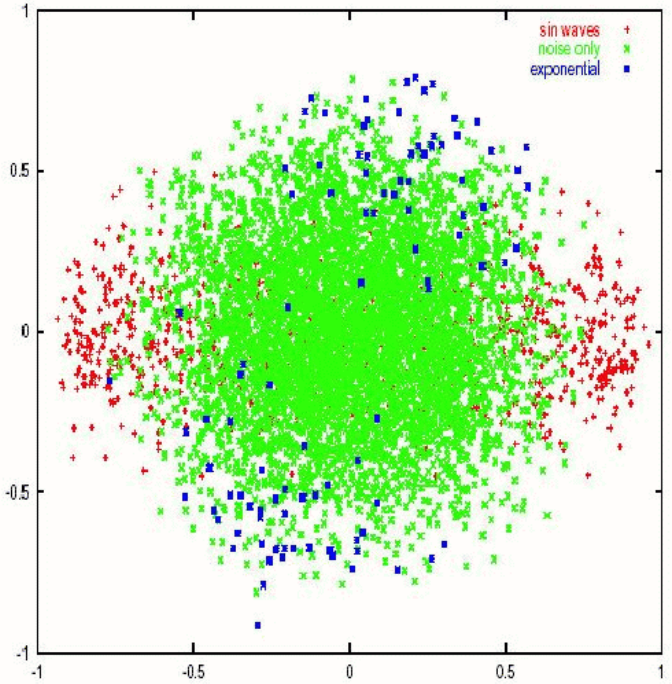
http://www.c3.lanl.gov/~rocha

rocha@lanl.gov

SVD Mode Plot

Need for More Iterative Spectral Methods





- Gene Shaving and Clustering do not even find the full sinusoisal component
- Exploring Iterative Variations to Extract Weaker Signals

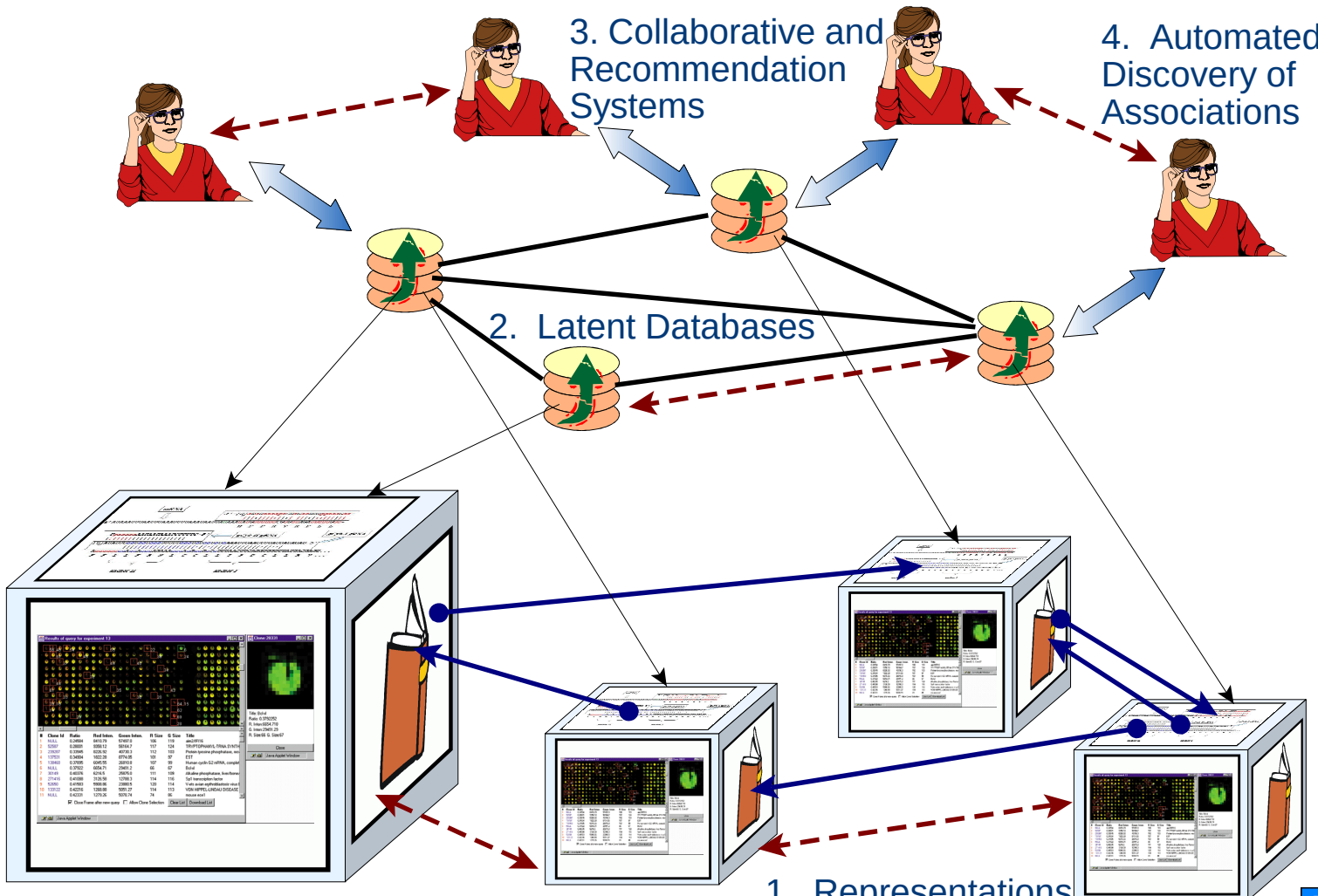
Luis Rocha
2001

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

rocha@lanl.gov

BioKnowledge Systems

Collaborative Scientific Environments



Luis Rocha
2001

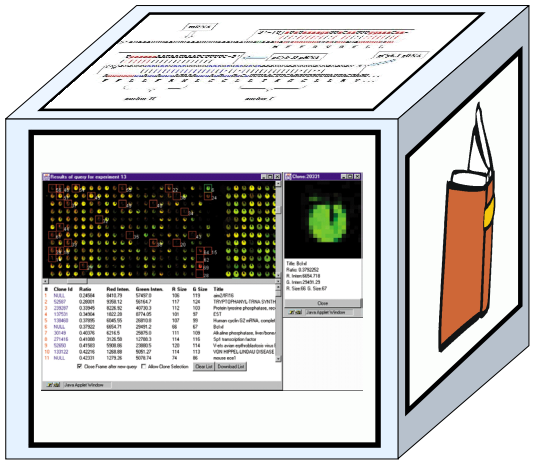
http://www.c3.lanl.gov/~rocha

Los Alamos National Laboratory

rocha@lanl.gov

1. Representations of Biological Data

Data Objects and Object Architectures



- Objects capable of grouping data sets, reports, code, etc.
  - Networked, proactive containers
  - Nelson's Buckets: Intelligent Data Agents
  - Object Management Group
- Specify specific needs of biological data
  - Genomic, Immunological, Epidemiological, etc.

Luis Rocha
2001

rocha@lanl.gov

Representations Of Biological Data

Semantic Markup and Exchange Protocols

- To facilitate retrieval, linking, and intelligent behavior of data objects there is a need to characterize data according to the needs of users.
  - Standards based on XML and UML
    - GEML (Gene Expression Markup Language)
    - GeneX (NCGR)
    - SBML (Systems Biology Markup Language)
  - Domains can be conceptualized as ontologies
    - Bio-ontologies Consortium
    - BioPathways Consortium
  - Exchange protocols
    - Based on RDF(S) (Resource Description Format Schema)
    - Ontology Interchange Layer (OIL)
    - For biological data: EcoCyc and TAMBIS.
- Aim is to select and develop appropriate representations for biological data data.

rocha@lanl.gov

GEML

Example of a Pattern File

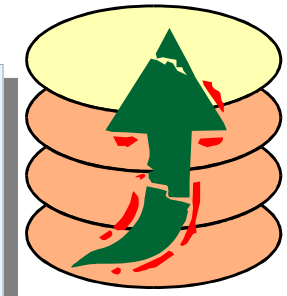
```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE project SYSTEM "GEMLPattern.dtd">
<project name="Hsapiens-421205160837" date="07-12-1999 12:43:48" by="jzsmith" company="JZSmith Technologies" >
    <pattern name="Hsapiens-421205160837">
        <reporter  name="XV186450"  systematic_name="XV186450"
                   active_sequence="TCTCACTGGTCAGGGGTCTTCTCCC"  start_coord="159">
            <feature number="6878">
                <position x="0.3333" y="0.508374" units="inches" />
            </feature>
            <gene primary_name="XV186450" systematic_name="XV186450" >
                <accession database="n/a" id="XV186520" />
            </gene>
        </reporter>
        <reporter name="T89593" systematic_name="T89593"
                   active_sequence="TACAGTGTCAGAATTAACTGTAGTC"  start_coord="201"  >
            <feature number="6879">
                <position x="0.340707" y="0.508374" units="inches" />
            </feature>
            <gene primary_name="T89593" systematic_name="T89593" >
                <accession database="n/a" id="T89593" />
            </gene>
        </reporter>
        <!-- Total Number of Reporters: 2 -->
    </pattern>
    <printing date="07-12-1999 12:43:48" printer="IJS 3" type="INKJET"
              pattern_name="Hsapiens- 421205160837" >
        <chip barcode="JZS123456781" />
        <chip barcode="JZS123456782" />
        <chip barcode="JZS123456783" />
        <chip barcode="JZS123456784" />
    </printing
</project>
```

Luis Rocha
2001

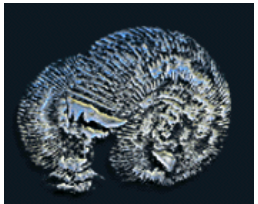
rocha@lanl.gov

2. Latent Databases of Biological Data

- Latent databases discover implicit, higher-order associations among stored objects
  - Latent Semantic Analysis
  - Analysis of Graph Structure
    - Links, Distance Functions and Metrics
  - Clustering
  - Works at several levels
    - Within objects, groups of objects, and the entire corpus
- In Information Retrieval latent associations are extracted from the relation between documents and keyterms



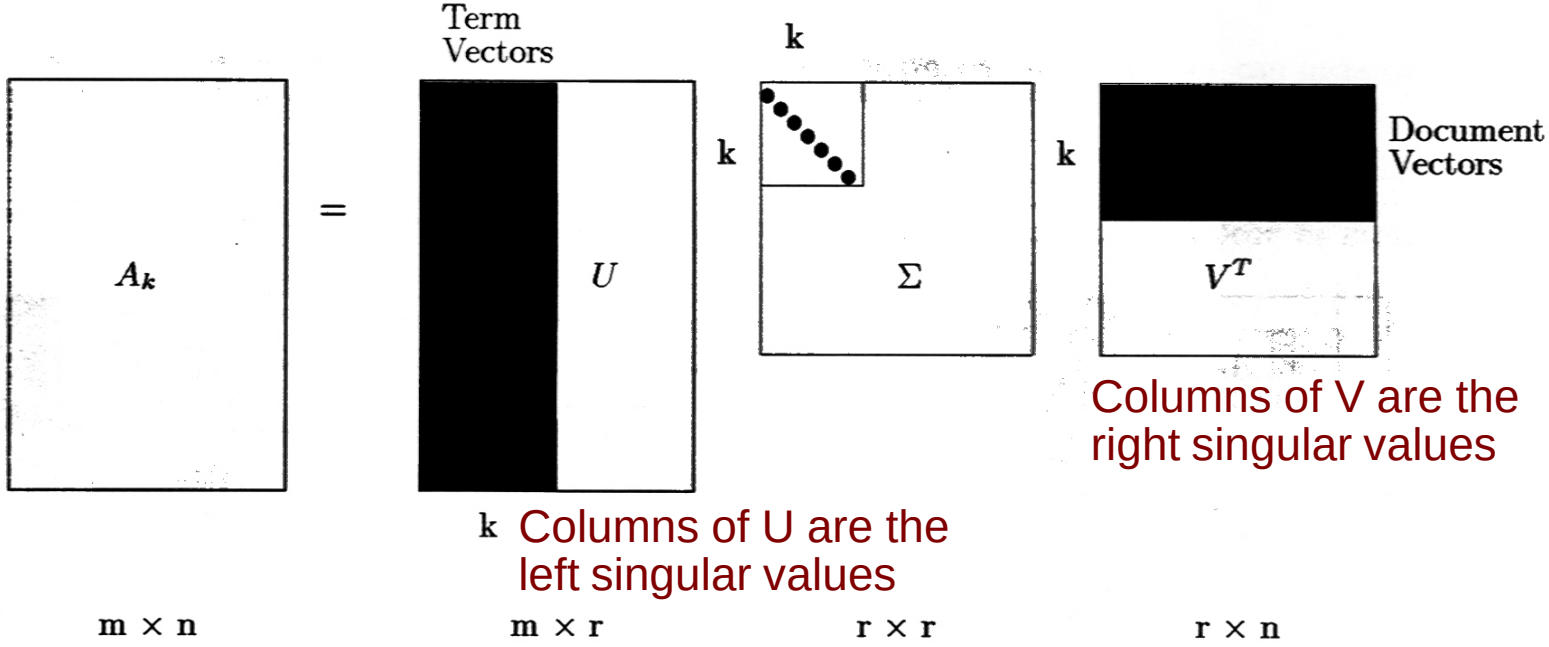
Luis Rocha
2001



rocha@lanl.gov

Latent Databases

Using Singular Value Decomposition



SVD allows us to obtain the lower rank approximations that best approximate the original matrix. What is lost by losing weaker singular values, is unnecessary noise. The underlying, essential structure of associations between keyterms and records is kept

Luis Rocha
2001

rocha@lanl.gov

Latent Databases

Keyword × Documents Relation is stored as a lower k SVD representation

Example: Small database from 17 books reviewed by SIAM Review

Underlied words are keyterms

Label	Titles
B1	A Course on Integral Equations
B2	Attractors for Semigroups and Evolution Equations
B3	Automatic Differentiation of Algorithms: Theory, Implementation, and Application
B4	Geometrical Aspects of Partial Differential Equations
B5	Ideals, Varieties, and Algorithms – An Introduction to Computational Algebraic Geometry and Commutative Algebra
B6	Introduction to Hamiltonian Dynamical Systems and the N-Body Problem
B7	Knapsack Problems: Algorithms and Computer Implementations
B8	Methods of Solving Singular Systems of Ordinary Differential Equations
B9	Nonlinear Systems
B10	Ordinary Differential Equations
B11	Oscillation Theory for Neutral Differential Equations with Delay
B12	Oscillation Theory of Delay Differential Equations
B13	Pseudodifferential Operators and Nonlinear Partial Differential Equations
B14	Sinc Methods for Quadrature and Differential Equations
B15	Stability of Stochastic Differential Equations with Respect to Semi-Martingales
B16	The Boundary Integral Approach to Static and Dynamic Contact Problems
B17	The Double Mellin-Barnes Type Integrals and Their Applications to Convolution Theory

Luis Rocha
2001

rocha@lanl.gov

16×17 Keyterm × Document Matrix

Terms	Documents																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
nonlinear	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ordinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

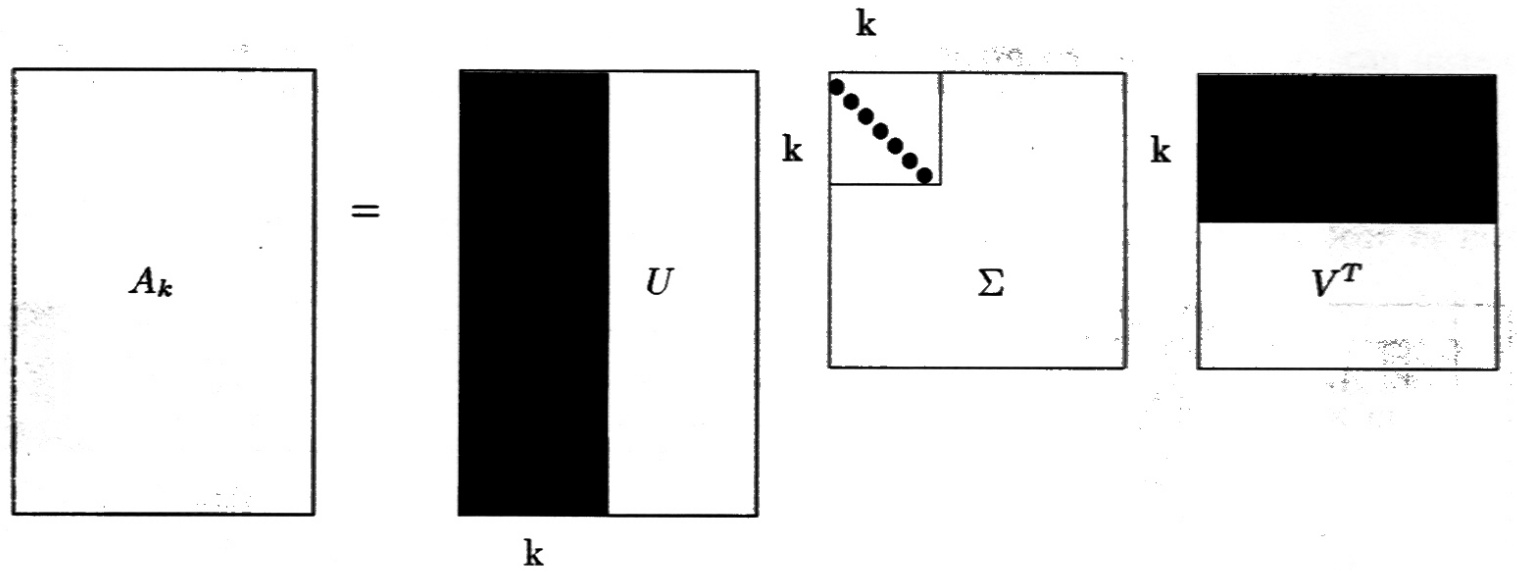
Think of a database of GEA data sets (instead of or in addition to documents), indexed by relevant genes (instead or in addition to keyterms)

Luis Rocha
2001

rocha@lanl.gov

Term × Document SVD

m=17, n=16



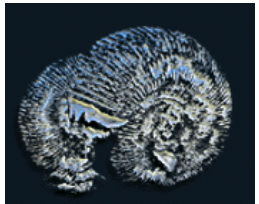
Columns are terms and rows are documents

Columns of U are eigenterms (rows are documents)

Rows of V^T are eigendocuments (columns are terms)

Luis Rocha
2001

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



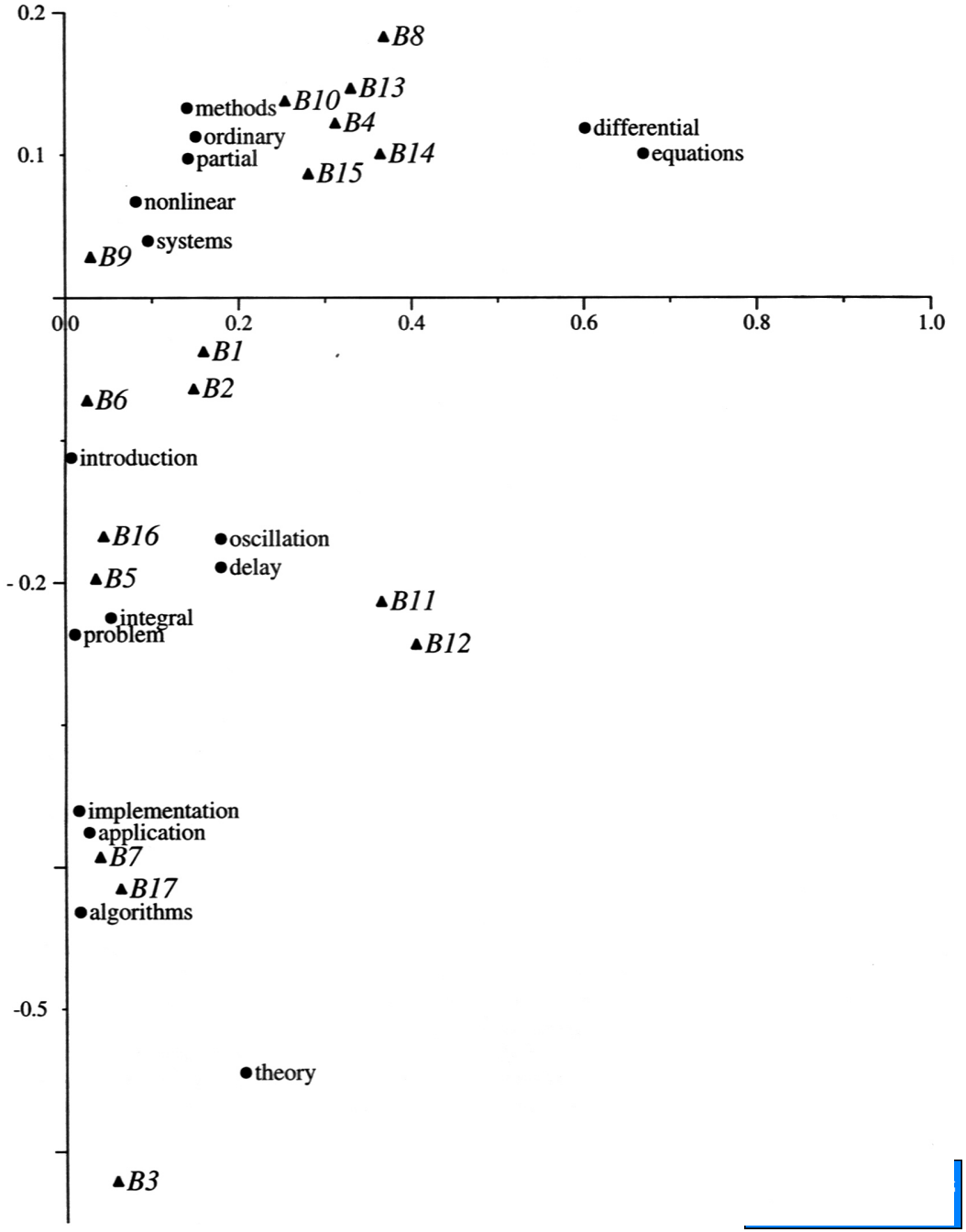
rocha@lanl.gov

SVD Aprox for k=2

Document and terms are plotted according to coefficients in the derived 2 eigenterms and eigendocuments

X seems to be about "differential equations" while y about more general algorithms and applications

Again think of the gene/dataset analogy



Luis Rocha
2001

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

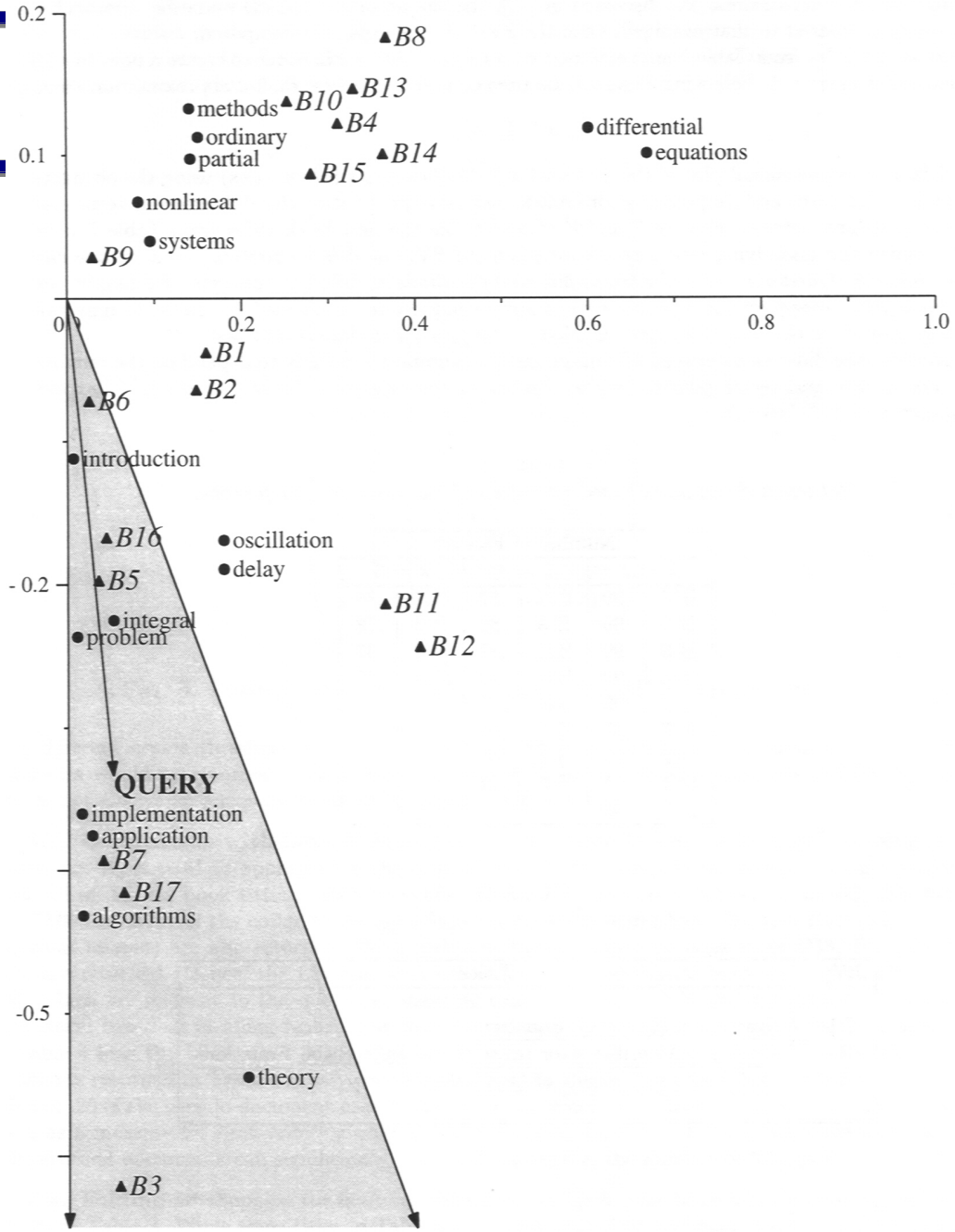
rocha@lanl.gov

Keyterm Query

Retrieval

Can also retrieve documents close to a set of other documents

For a database of datasets, this would mean that we would retrieve those data sets most relevant to study the genes in a query



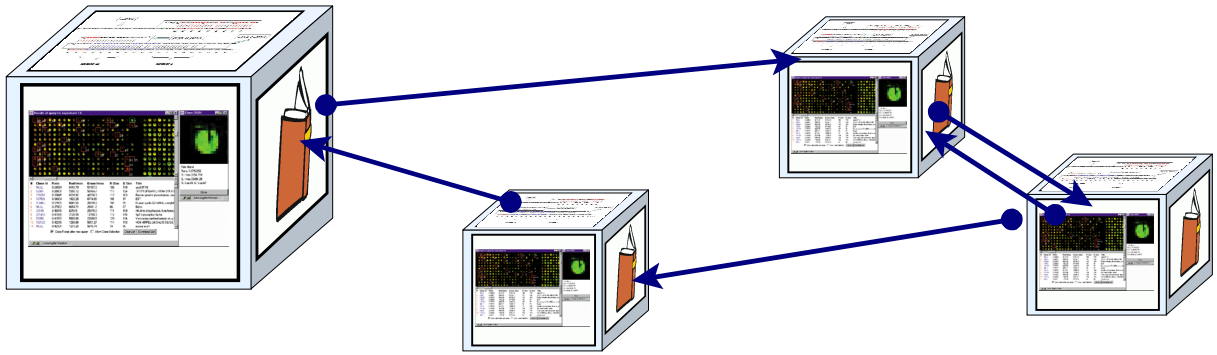
Luis Rocha
2001

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

rocha@lanl.gov

Publication Databases

Are Networks of Documents, datasets, software, etc.



- Associative Networks can be constructed for Publications, People, Keywords:
  - e.g. gene networks
- Associative weights (proximity) between chosen items are derived from
  - Co-occurrence, Co-citation, and other relations between documents containing chosen items

Luis Rocha
2001

Los Alamos
National Laboratory

rocha@lanl.gov

Distance Functions

From Selected Relations in Document Databases

- Document × Document
  - Document Distance according to Co-Citation or Hyperlink
- Document × Keyterms
  - Keyterm Distance
- Document/Dataset × Gene Expression
  - Gene Distance
- Document × Author
  - Author Distance (Collaboration Network)

$$ksp(k_i,k_j)=\frac{\sum_{k=1}^{m}(a_{i,k}\wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k}\vee a_{j,k})}=\frac{N_{\cap}(k_i,k_j)}{N_{\cup}(k_i,k_j)}$$

(Keyword Semantic Proximity)

$$d_{ksp}(k_i,k_j)=\frac{1}{ksp(k_i,k_j)}-1$$

(Keyword Semantic Distance)

Luis Rocha
2001

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

Los Alamos
National Laboratory

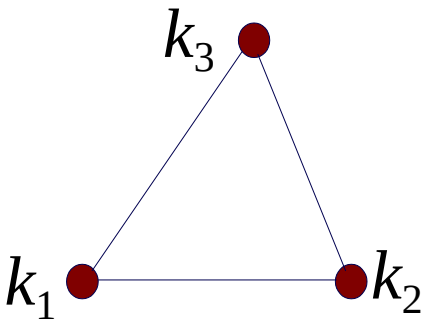
rocha@lanl.gov

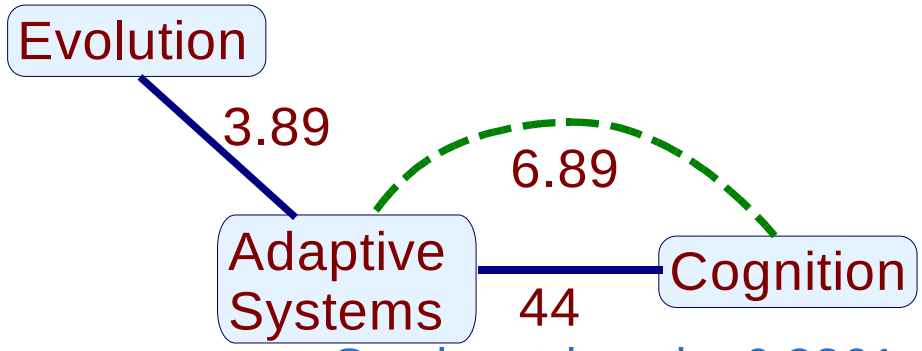
Measuring Semi-Metric Behavior

Semi-metric ratios

$$d_{ksp}(k_i, k_j) = \frac{1}{ksp(k_i, k_j)} - 1$$

(Keyword Semantic Distance)





Semi-metric ratio: 6.3861

$$d(k_1, k_2) \leq d(k_1, k_3) + d(k_3, k_2)$$

Metric

$$d(k_1, k_2) > d(k_1, k_3) + d(k_3, k_2)$$

Semi-metric

Luis Rocha
2001

rocha@lanl.gov

Measuring Semi-Metric Behavior

Semi-metric Measures

- Semi-metric ratio
  - Absolute measure of indirect distance reduction

$$s\left(k_i, k_j\right) = \frac{d_{direct}\left(k_i, k_j\right)}{d_{indirect}\left(k_i, k_j\right)}$$

- Relative Semi-metric ratio
  - Distance reduction against maximum contraction

$$rs\left(k_i, k_j\right) = \frac{d_{direct}\left(k_i, k_j\right) - d_{indirect}\left(k_i, k_j\right)}{d_{max} - d_{min}} = \frac{d_{direct}\left(k_i, k_j\right) - d_{indirect}\left(k_i, k_j\right)}{d_{max}}$$

- Below Average Ratio
  - Captures semi-metric distance reductions which contract to below the average distance for a given node. Captures some of the cases of initial ∞ distance

$$b\left(k_i, k_j\right) = \frac{\overline{d_{k_i}}}{d_{indirect}\left(k_i, k_j\right)}$$

Luis Rocha
2001

rocha@lanl.gov

Trends in Collections

ARP Database

500 Keywords

leukemia	myocardi	Semi-metric Ratio	272.1996	Relative Semi-metric	0.4981	1
hormon	thin	Semi-metric Ratio	214.0797	Relative Semi-metric	0.9953	2
care	excit	Semi-metric Ratio	213.5900	Relative Semi-metric	0.9953	3
gene	equat	Semi-metric Ratio	205.7649	Relative Semi-metric	0.9951	4
film	transcript	Semi-metric Ratio	204.5103	Relative Semi-metric	0.9951	5
spectroscopi	care	Semi-metric Ratio	194.3478	Relative Semi-metric	0.9949	6
transcript	thin	Semi-metric Ratio	193.0644	Relative Semi-metric	0.9948	7
pressur	t-cell	Semi-metric Ratio	190.8173	Relative Semi-metric	0.9948	8
film	mutat	Semi-metric Ratio	186.8350	Relative Semi-metric	0.9946	9
vascular	catalyst	Semi-metric Ratio	185.3671	Relative Semi-metric	0.9946	10
film	endoth	Semi-metric Ratio	183.0219	Relative Semi-metric	0.9945	11
film	macrophag	Semi-metric Ratio	180.4128	Relative Semi-metric	0.9945	12
nonlinear	nerv	Semi-metric Ratio	177.6419	Relative Semi-metric	0.9944	13
film	clone	Semi-metric Ratio	175.6775	Relative Semi-metric	0.9943	14
mutat	equat	Semi-metric Ratio	175.1138	Relative Semi-metric	0.9943	15
film	secretion	Semi-metric Ratio	174.9438	Relative Semi-metric	0.9943	16
thin	endoth	Semi-metric Ratio	173.8007	Relative Semi-metric	0.9942	17
pressur	leukemia	Semi-metric Ratio	172.2365	Relative Semi-metric	0.9942	18
thin	macrophag	Semi-metric Ratio	171.4462	Relative Semi-metric	0.9942	19
film	mortal	Semi-metric Ratio	169.9975	Relative Semi-metric	0.9941	20
clone	thin	Semi-metric Ratio	167.1643	Relative Semi-metric	0.9940	21

95% Semi-metric
35% Below Average

leukemia	myocardi	Under Average	77.7665
thin	nitric	Under Average	50.5213
equat	messenger-rna	Under Average	42.3600
chemotherapi	myocardi	Under Average	41.6956
nonlinear	nerv	Under Average	40.1634
film	risk	Under Average	40.0989
equat	transcript	Under Average	39.9494
equat	clone	Under Average	39.9156
film	hormon	Under Average	39.6987
equat	gene-express	Under Average	37.0435

Luis Rocha
2001

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

rocha@lanl.gov

Extracting Networks from Publications

Literature

- Masys, D.R. et al [2001]."Use of keywords hierarchies to interpret gene expression patterns." *Bioinformatics*. Vol. 17, no. 4, pp. 319-326
  - Derived a measure of conceptual similarity between genes based on co-occuring keywords (in the Medical Subject Headings Index) associated to documents detailing gene expression patterns in MEDLINE.
- Jenssen, T.K. et al [2001]. “A literature network of human genes for high-throughput analysis of gene expression”. *Nature Genetics*, Vol. 28, pp. 21-28.
  - Similar to above, but produced a web tool to navigate to conceptual gene space: http://www.PubGene.org.
- Stapley, B.J. and G. Benoit [2000]. “Biobibliometrics: Information retrieval and visualization from co-occurrence of gene names in Medline abstracts”. *Pac. Symp. Biocomput*. Vol. 5, pp. 529-540.
  - Defined measures of similarity used by work above.
- Andrade, M.A. and P. Bork. “Automated extraction of information in molecular biology”. *FEBS Letters*, Vol. 476, pp. 12-17.
  - Review of data mining, Information Retrieval, and Text mining techniques for molecular biology databases.

rocha@lanl.gov

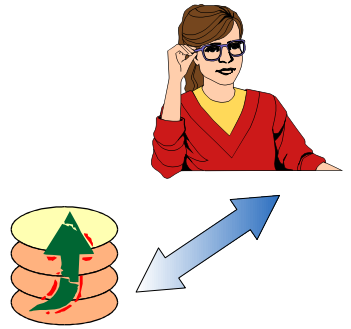
Gene Networks From Publications

- Strength of these techniques
  - Since much of the publication records discuss clinical data, the derived gene networks offer a global picture of gene associations in an integrated clinical observation space. This can supplement gene networks derived from biochemical observations.
  - *At Worst*: powerful tool for biologists to navigate new literature knowledge about the subsets of genes they are interested in.
  - *At Best*: such conceptual gene networks identify actual empirical associations
- Current literature uses the simplest similarity analysis of derived networks
  - We are utilizing more sophisticated Information Retrieval and Recommendation technology as well as graph-theoretical analysis of associative networks extracted from publication databases.
    - E.g. Latent database architectures and metric analysis of derived distance functions

rocha@lanl.gov

3. Collaborative and Recommendation Systems

Adaptive and Collective Behavior



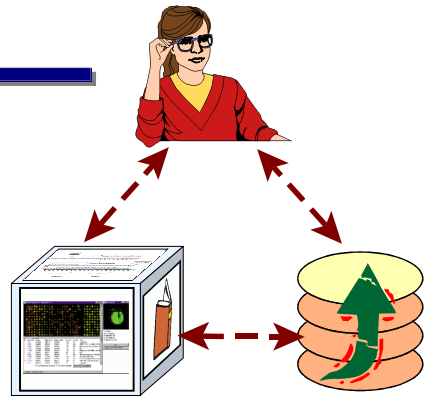
- Tasks of complex biological problems are tackled by large teams and communities.
  - The behavior of these communities can itself be harvested to discover associations between data-sets, hypothesis, etc.
- Recommendation systems use the collective behavior of users (plus latent relations) to discover, categorize, and recommend resources and fellow researchers.

Luis Rocha
2001

rocha@lanl.gov

4. Automated Discovery of Associations

Analysis of New Associations



- All 3 previous subcomponents aim at building the cabability of automatic generation of associations:
  - 1. Produces intelligent data containers that keep both author-supplied and automatic associations
  - 2. Produces databases that discover latent associations, distance functions, and reduce dimensionality
  - 3. From collective behavior, associations are produced at all levels.
- However, this automatic generation of associations should be itself harvested
  - For generating hypothesis about biological processes to help design new experiments (new decompositions)
  - Discovery of communities of interest
  - DKS as research tools in addition to information retrieval and recommendation systems.