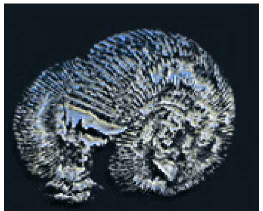


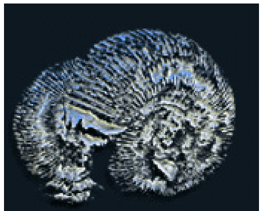
rocha@lanl.gov

Introduction to Bioinformatics

A Systems Biology Approach

Luis M. Rocha
Complex Systems Modeling
CCS3 - Modeling, Algorithms, and Informatics
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545

COMPUTER & COMPUTATIONAL SCIENCES



Introduction to Bioinformatics

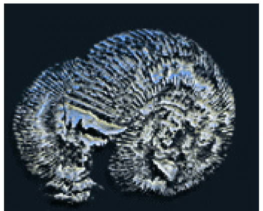
Course Layout: March 11-15, 2002

- Monday: *Bioinformatics Practice*
  - Pedro Fernandes
- Tuesday: *From Bioinformatics to Systems Biology*
  - Luis Rocha
- Wednesday: *DNA Chip Technology*
  - Michael Wall
- Thursday: *Network Inference*
  - Patrik D'haeseleer
- Friday: *Integrative Technology for Computational Biology*
  - Luis Rocha

From Bioinformatics to Systems Biology

Layout

- Systems Biology
- Synthetic, Multi- Disciplinary Approach to Biology
- Grand Challenges of Systems Biology
- Full Curriculum for Bioinformatics
- Some traditional components of Bioinformatics:
  - Sequence Analysis, Similarity Search, Motif Search, Data-driven vs. Knowledge-based Functional Interpretation, Sequence Alignment, Dynamic Programming for Sequence Alignment Optimization, Similarity Database Search, basics of FASTA Method, Simulated Annealing and Genetic Algorithms for Multiple Sequence Alignment, etc
- Literature Discussion and Useful Resources

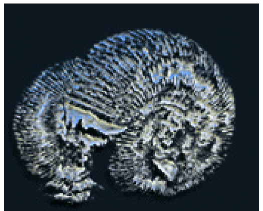


Systems Biology

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]

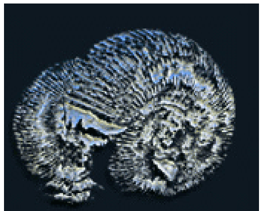
- Systems Science is the methodology used to study *systemhood* not *thinghood* properties in Nature.
  - Modeling and Simulation of systems measured from and validated in real things.
  - It accumulates knowledge via Mathematical and Computational analysis of classes of systems, models, and problems.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



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 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]



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.

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]

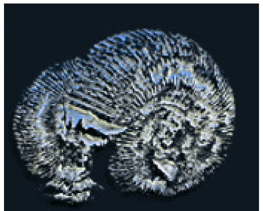
- Biology must simplify organisms to study them – some type of abstraction or modeling is needed.
  - External (Functional) description (favored by Systems Thinking)
    - *Blackbox*, input-output behavior of observables
    - Tells us what the system does
    - Function depends on repercussions in an 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

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://www.dynamical-systems.org/threebody/



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

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

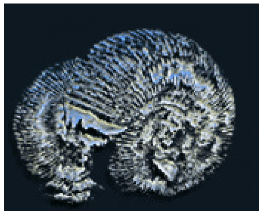
Post-Genome informatics

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

Post-Genome Informatics or the "New" Systems Biology

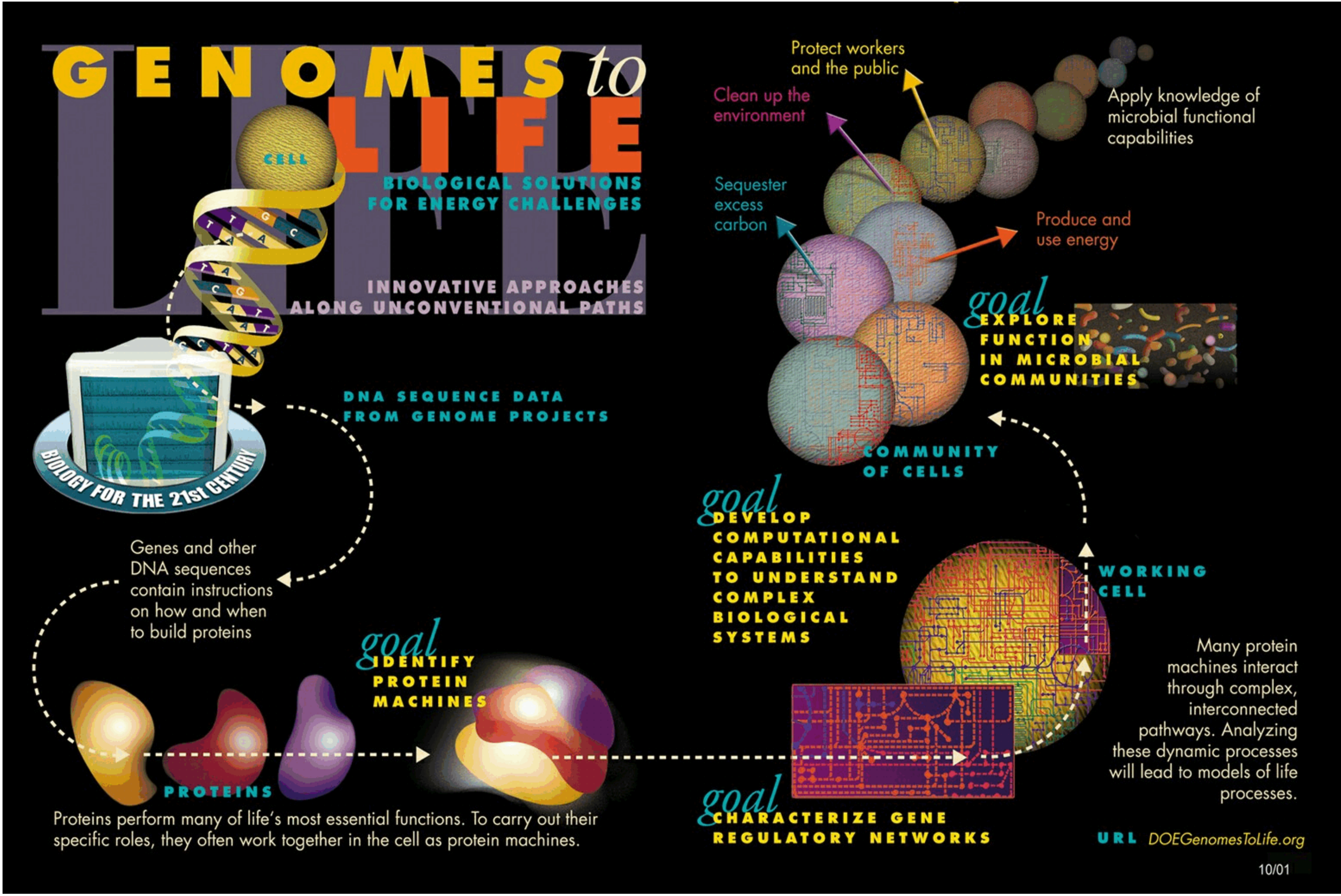
- *Systems biology* is a unique approach to the study of genes and proteins which has only recently been made possible by rapid advances in computer technology. Unlike traditional science which examines single genes or proteins, systems biology studies the complex interaction of all levels of biological 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.



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



GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

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 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

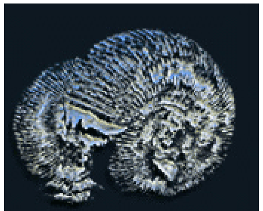
Protect workers and the public

Produce and use energy

Apply knowledge of microbial functional capabilities

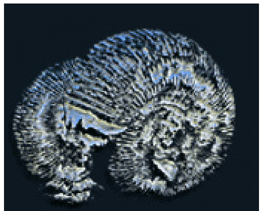
URL *DOEGenomesToLife.org*

10/01



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@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-Perforemance Algorithms to allow whole-system Kinetic models

Luis Rocha
2002

rocha@lanl.gov

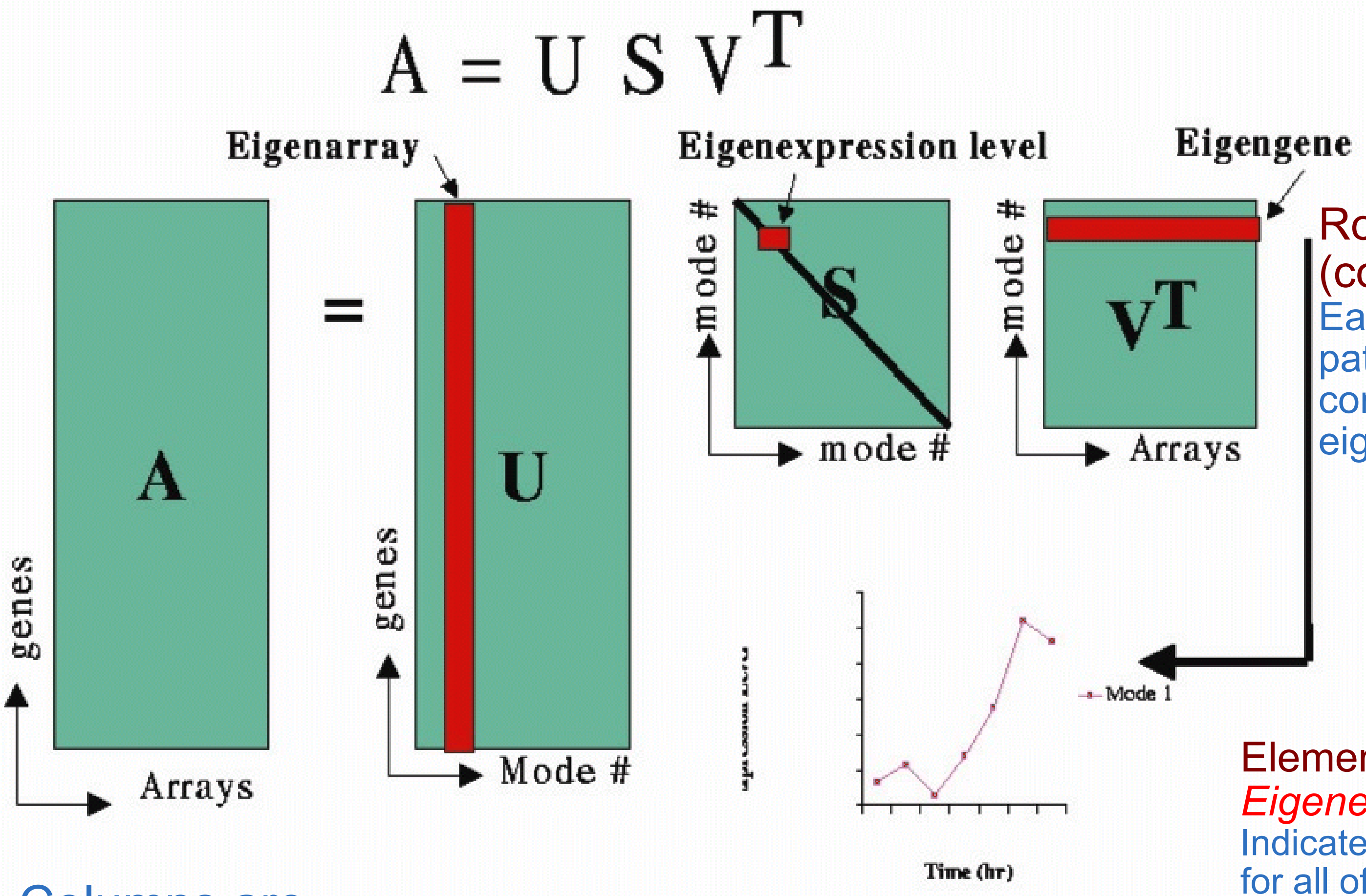
Systems Biology

On-going work at LANL (CCS)

- 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)
    - Network Inference (John Ambrosiano, Michael Wall)
    - 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)
- Integrative Technology
  - Representations of Biological Data
  - Latent Databases
  - Collaborative and Recommendation Systems
  - Automated Analysis of Whole Databases of Publications and data-sets

rocha@lanl.gov

SVD for Gene Expression



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

Elements of Diagonal S: *Eigenexpression level* Indicate the amount of variance for all of the data that is explained by each eigengene.

Columns are arrays (time steps) and rows are genes

Columns of U: *eigenarrays* (rows are genes) describe how each eigengene contributes to a single gene's expresssion pattern (coefficients in a linear expansion).

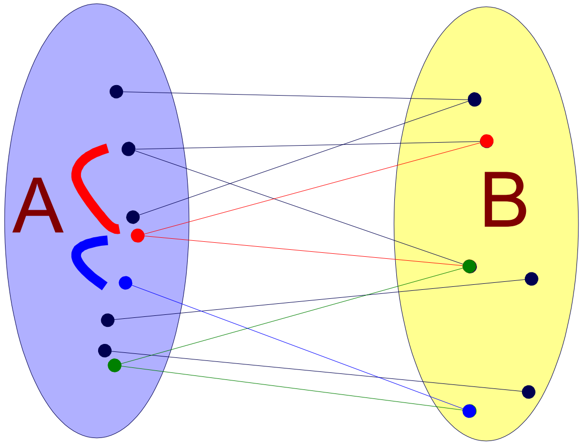
Luis Rocha
2002

rocha@lanl.gov

Singular Value Decomposition

What does it do? Higher-Order “Clustering”

Also Known as: Principal Components Analysis.



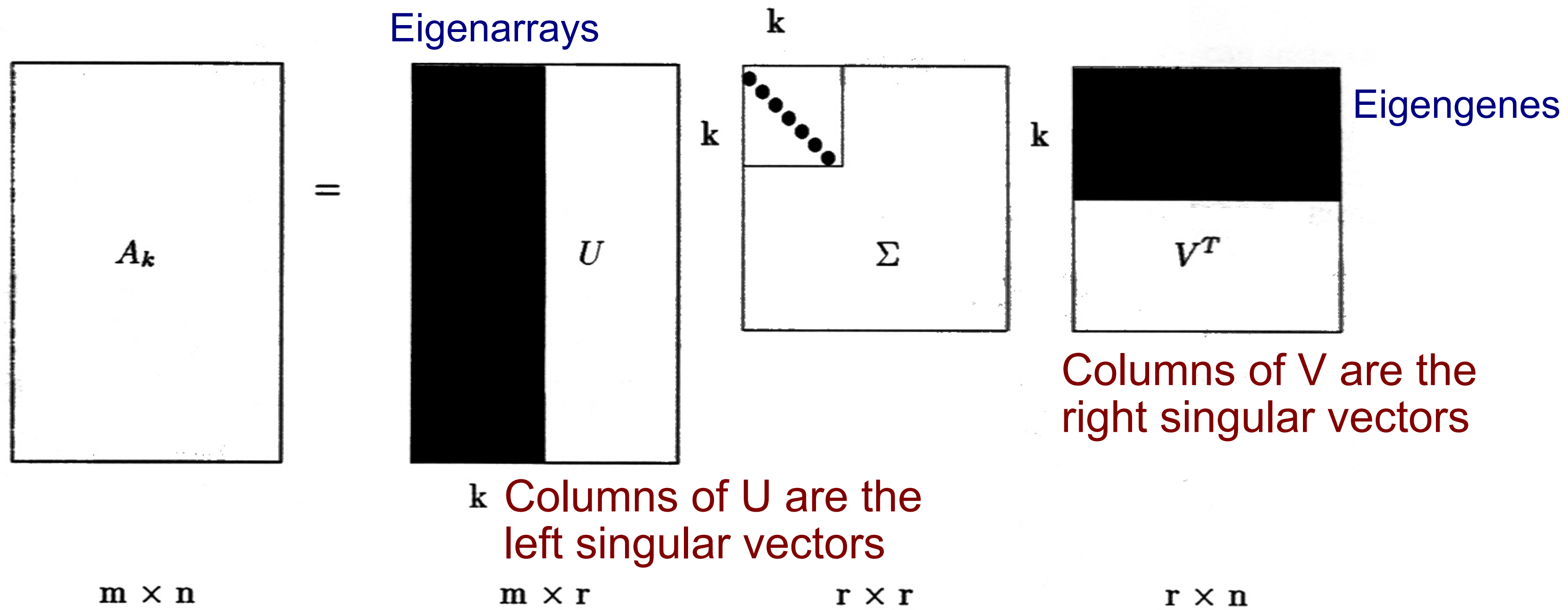
- Given a relation (a matrix) between 2 sets of distinct objects. SVD is used to discover the implicit higher-order structure in the relation
  - Keyterms by Documents, Genes by Arrays
  - Higher-order means indirect relationships: Those associations between the two types of objects which are not evident by individual associations.

- In Language and IR most words have many meanings (polysemy) and there are several possible words to express the same concept (synonymy)
  - SVD is used to identify the several meanings of words and “cluster” the words that express the same concept.
- For gene expression data, we expect to find genes which participate in several networks (gene functional polysemy) and different genes to participate in the same networks (gene functional synonymy)
  - Clustering usually demands strict inclusion (except for Fuzzy)

Luis Rocha
2002

rocha@lanl.gov

SVD for Lower Rank Approximations



Columns of V are the right singular vectors

Columns of U are the left singular vectors

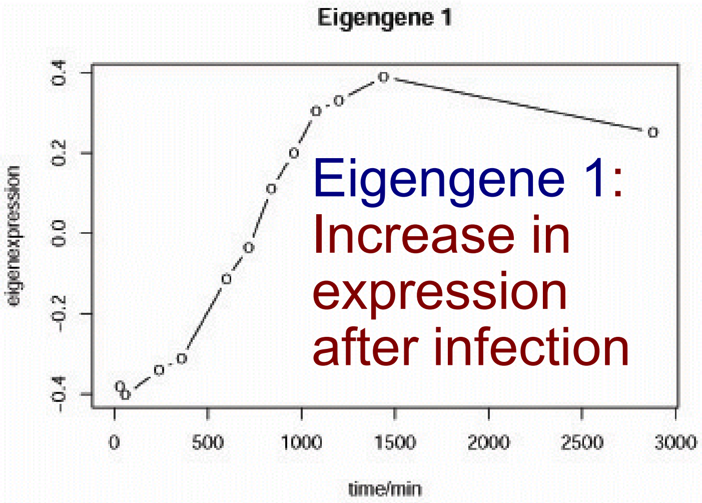
SVD allows us to obtain the lower rank approximations that best approximate the original matrix. What is lost by losing weaker singular values, is believed to be unnecessary noise. The underlying, essential structure of associations between genes and arrays is preserved. Neural Networks and other classifiers perform better on the decomposed, lower dimensionality data (yeung, 2001???)

http://linneus20.ethz.ch:8080/2_2_1.html

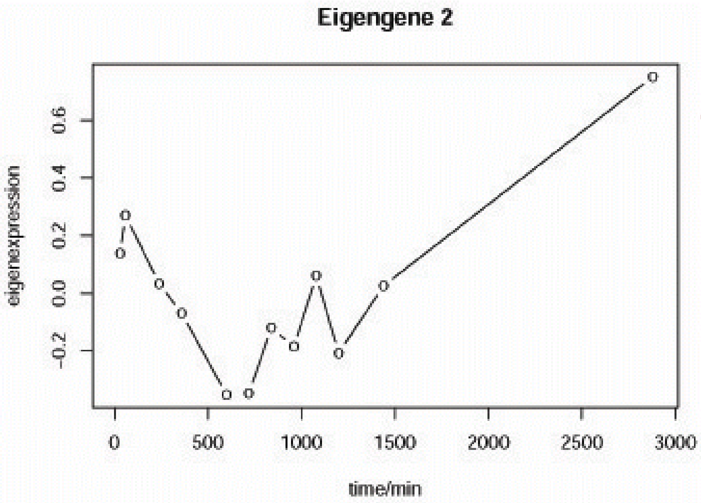
http://fonsg3.let.uva.nl/praat/manual/Principal_component_analysis.html

SVD of Time-Dependent Expression Data

Gene expression (13000 genes) after infection with herpes virus

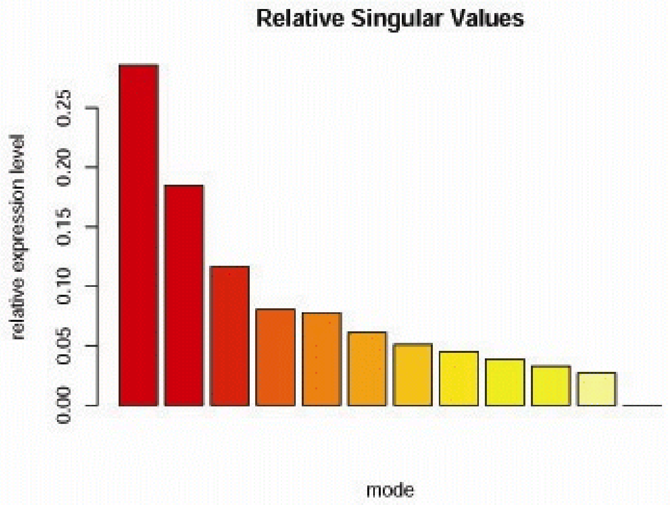


Eigengene 1: Increase in expression after infection



Eigengene 2: Transient decrease in expression after infection

12 point time series (30min - 48hrs)



- Genes whose expression is positively (negatively) correlated with Eigengene 1 are genes whose expression is increased (decreased) after infection with Herpes virus
- Genes whose expression is positively (negatively) correlated with Eigengene 2 are genes whose expression is transiently decreased (increased) after infection with Herpes virus.
- The singular value spectrum shows that the signal cannot be explained by just the first few modes

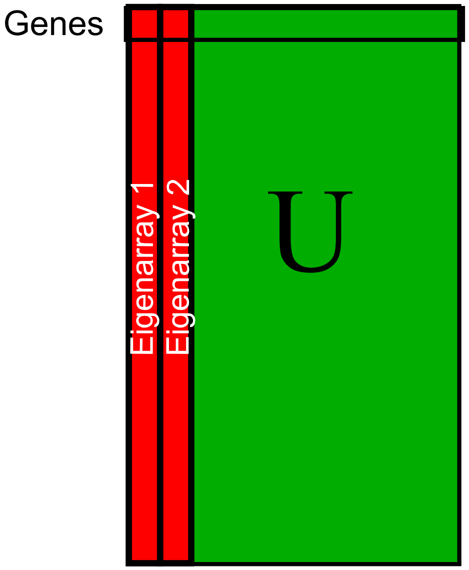
http://www.c3.lanl.gov/~rocha/bioinformatics

rocha@lanl.gov

COMPUTER & COMPUTATIONAL SCIENCES

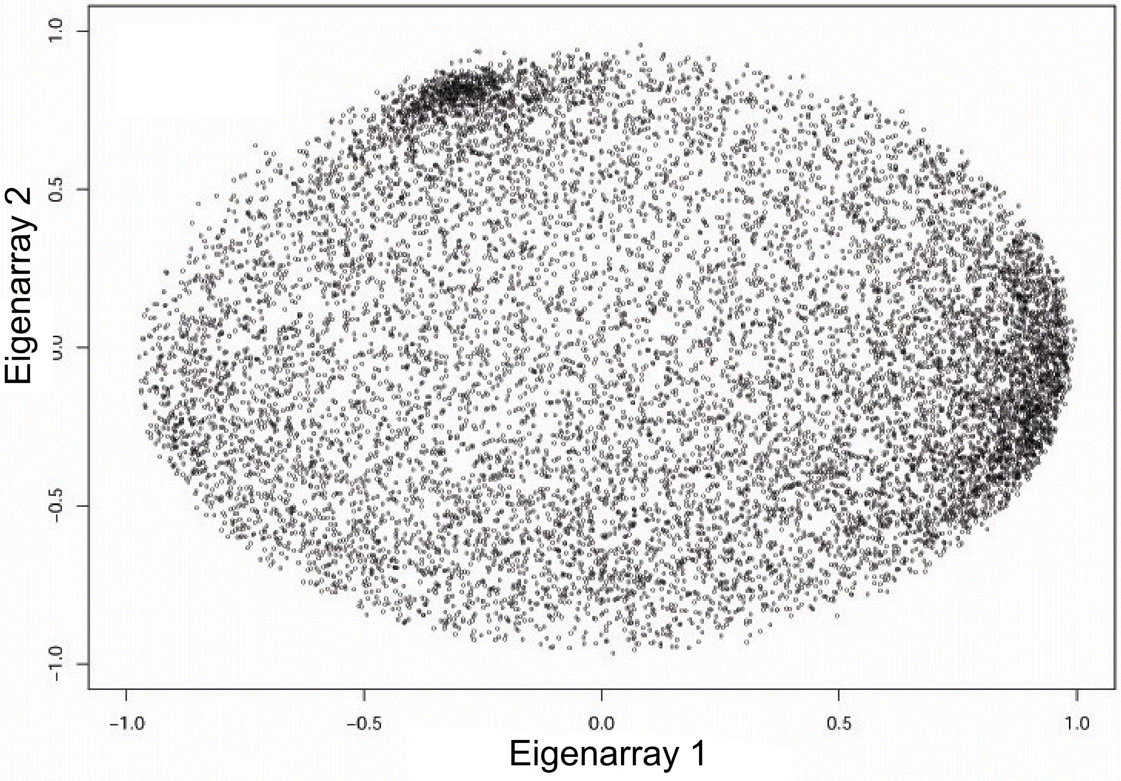
Biological Discovery via SVD

Eigenarray Coefficient Plot



LANL group found a second feature with interesting biological associations

genes involved in transcription regulation, immune response, oncogenesis as well as growth factors/cytokines and their receptors



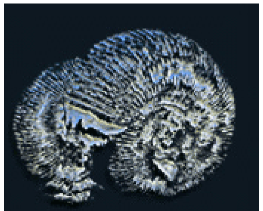
Princeton group (Shenk's lab) found ~1200 genes that showed significant changes in expression

at least 3 fold change in expression at at least 2 consecutive time points

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

Luis Rocha
2002

Los Alamos
NATIONAL LABORATORY

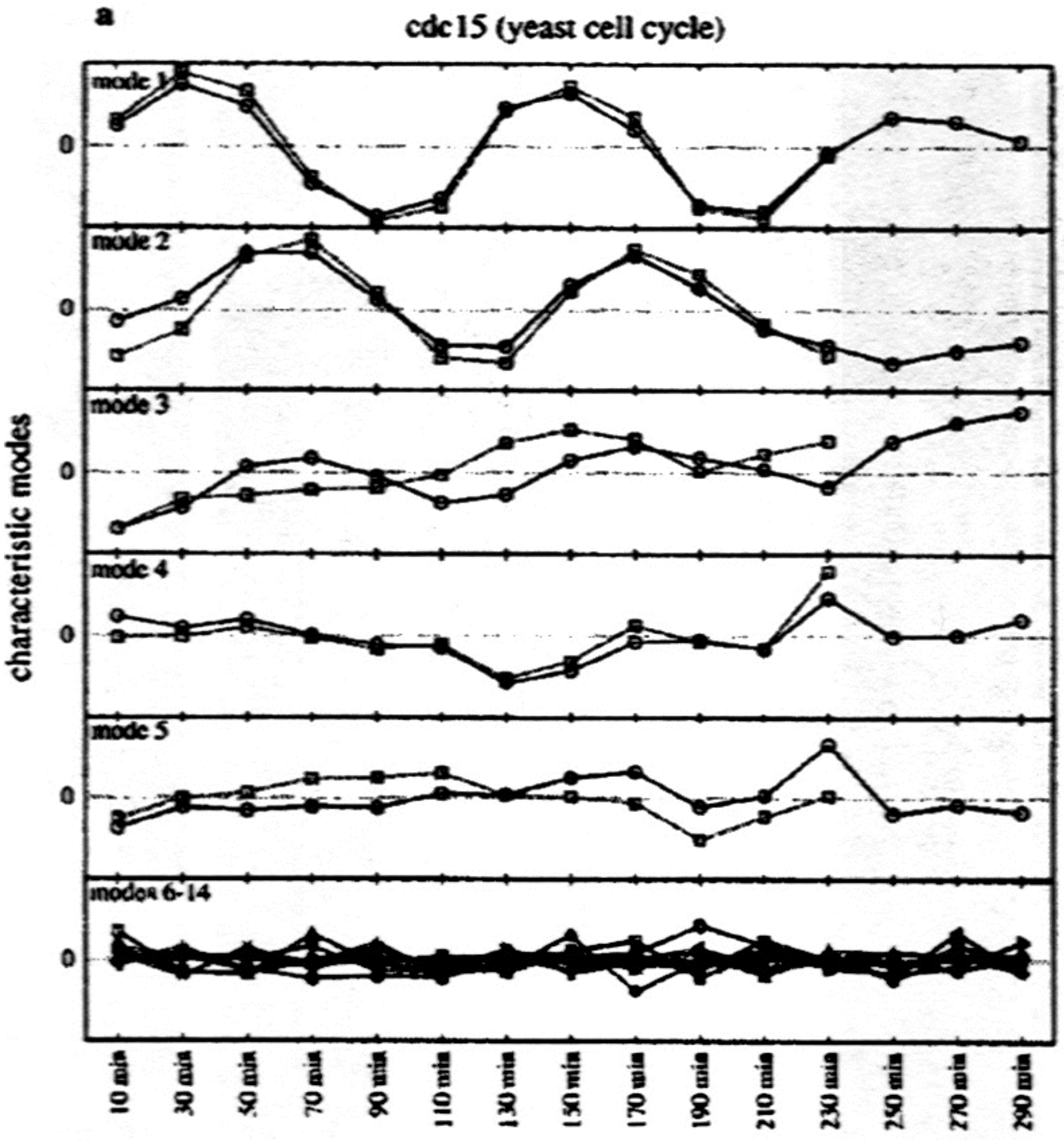


Data-Mining of Global Patterns

Discovery of Juxtaposed Functional Modes

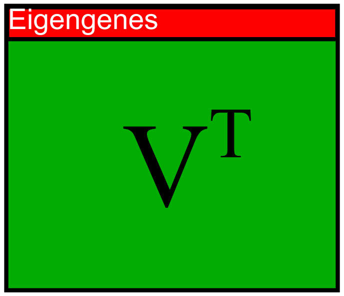
- Gene Expression Modes
  - Cluster analysis provides little insight into inter-relationships among groups of co-regulated genes. Tends to demand separated grupings.
  - Component (“spectral”) analysis yields a description of superposed behavior of gene expression networks, rather than a partition.
    - 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

Holter et al SVD Analysys



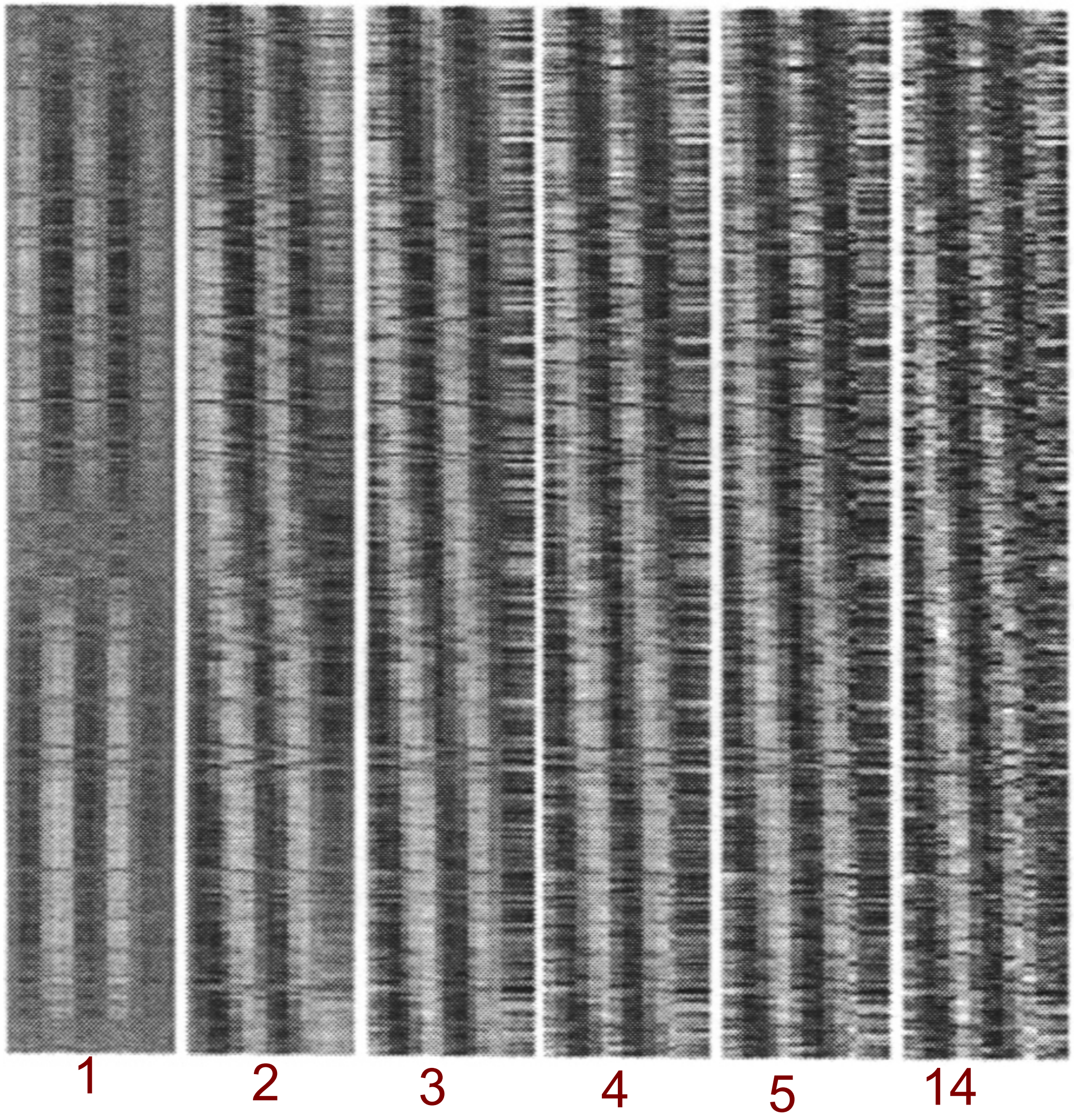
- 800 genes by 15 (12) time measurements
- 2 dominant modes
  - Approximately sinusoidal and out of phase
  - Less synchronized as cell enters 3rd cycle
  - If only 12 points are used, third SV loses relevance, but 2 first components remain largely unchanged

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


Eigengenes

V^T

cdc15 Reconstruction with k-highest modes



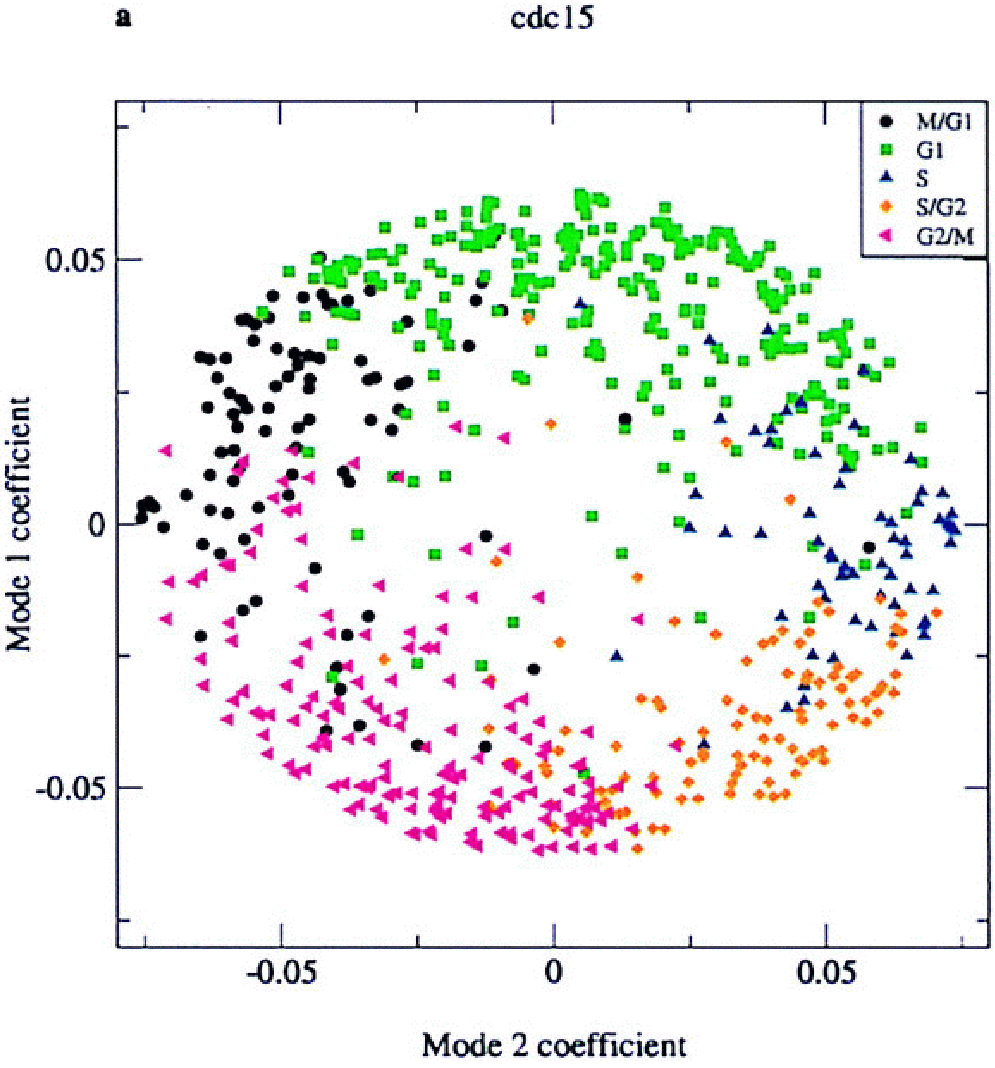
Rows are genes
Columns are time points

It implies an undelying simplicity in genetic response

rocha@lanl.gov

Eigenarray Coefficient Plot

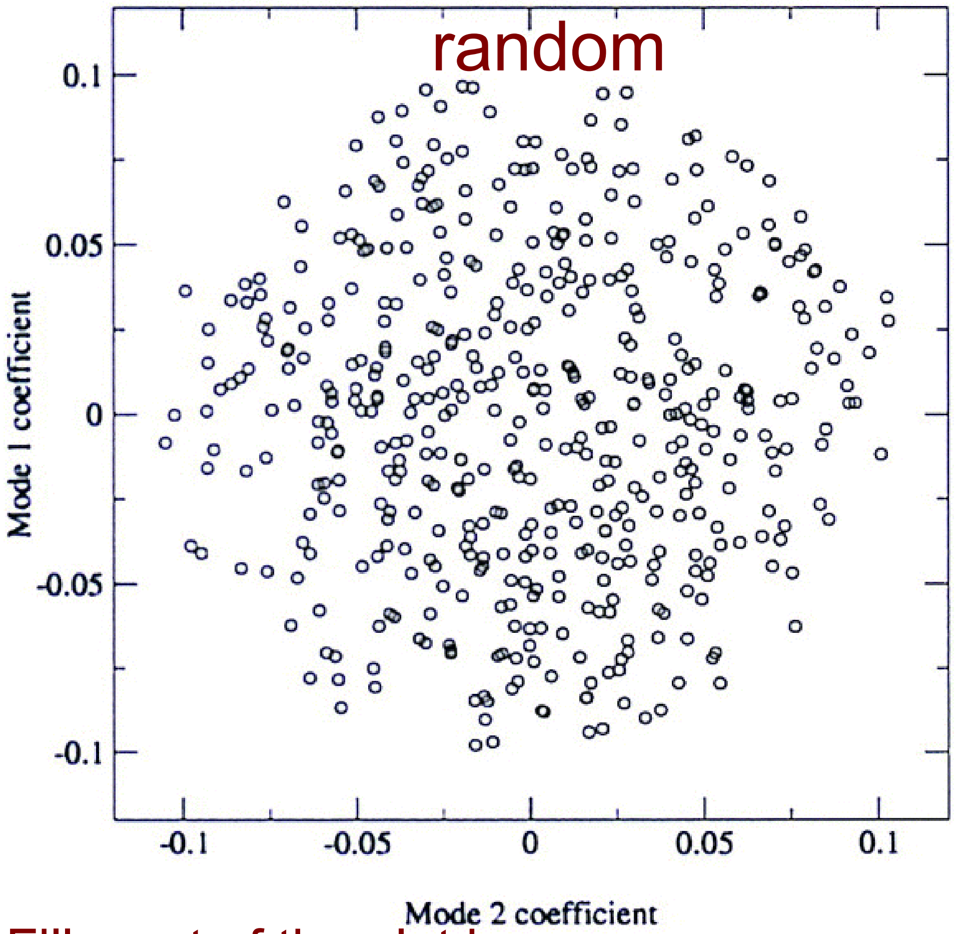
Plot of the coefficients of the first 2 modes for all genes

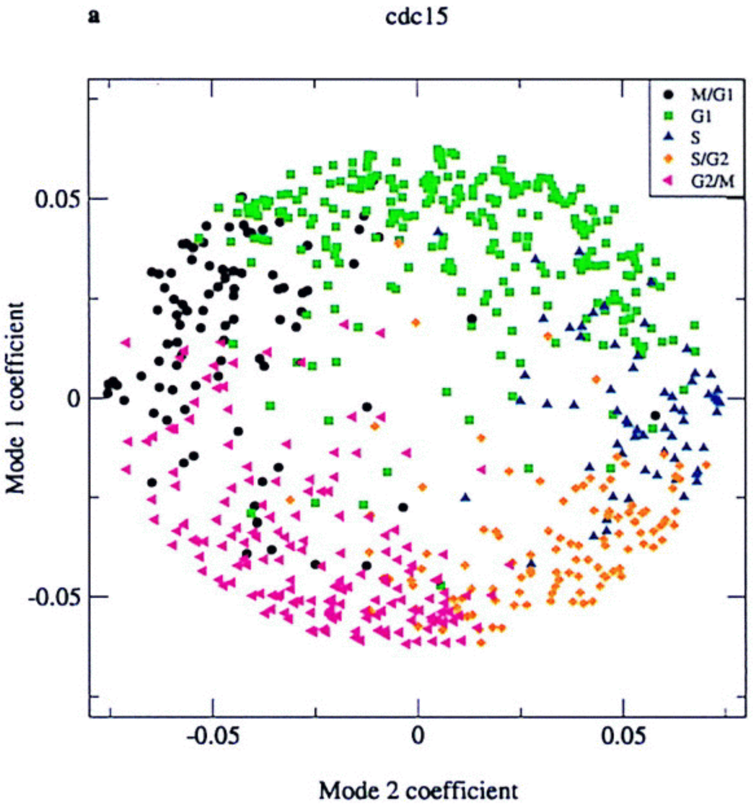


- Clusters of genes by other methods cluster in these plots, but the temporal progression in the cell cycle and in the course of sporulation is more evident in the SVD analysis
- Holter et al conclude that genes are not activated in discrete groups or blocks, as historically implied by the division of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

Eigenarray Coefficient Plot

Random data





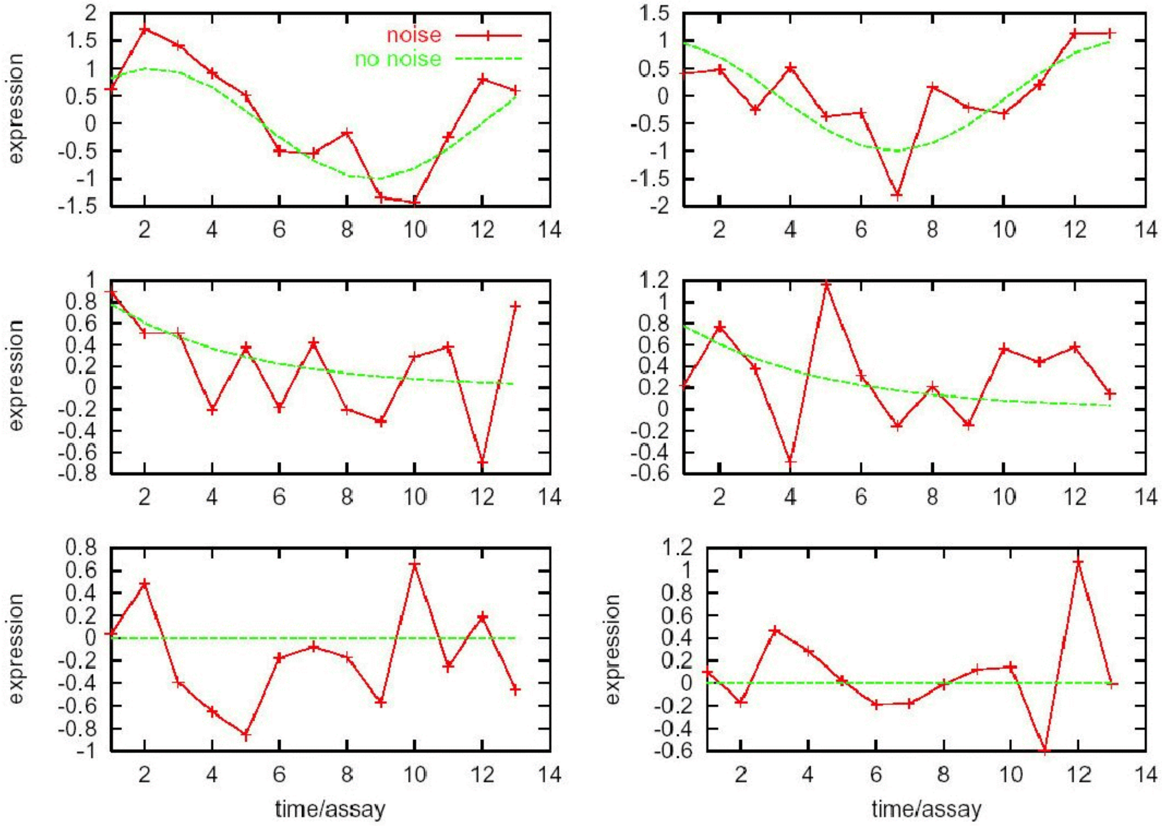
Fill most of the plot because genes are not very correlated with components. A circle implies equal contribution from each component (rather than an elipse)

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
  - Discovers "superposed" gene expression behavior. The overall behavior identified by eigengenes does not describe a particular gene or the average of a cluster, but rather a separable component of the integrated behavior of the colection. The same gene can be correlated with several eigengenes.

Discovering Hidden Functional Expression Modes

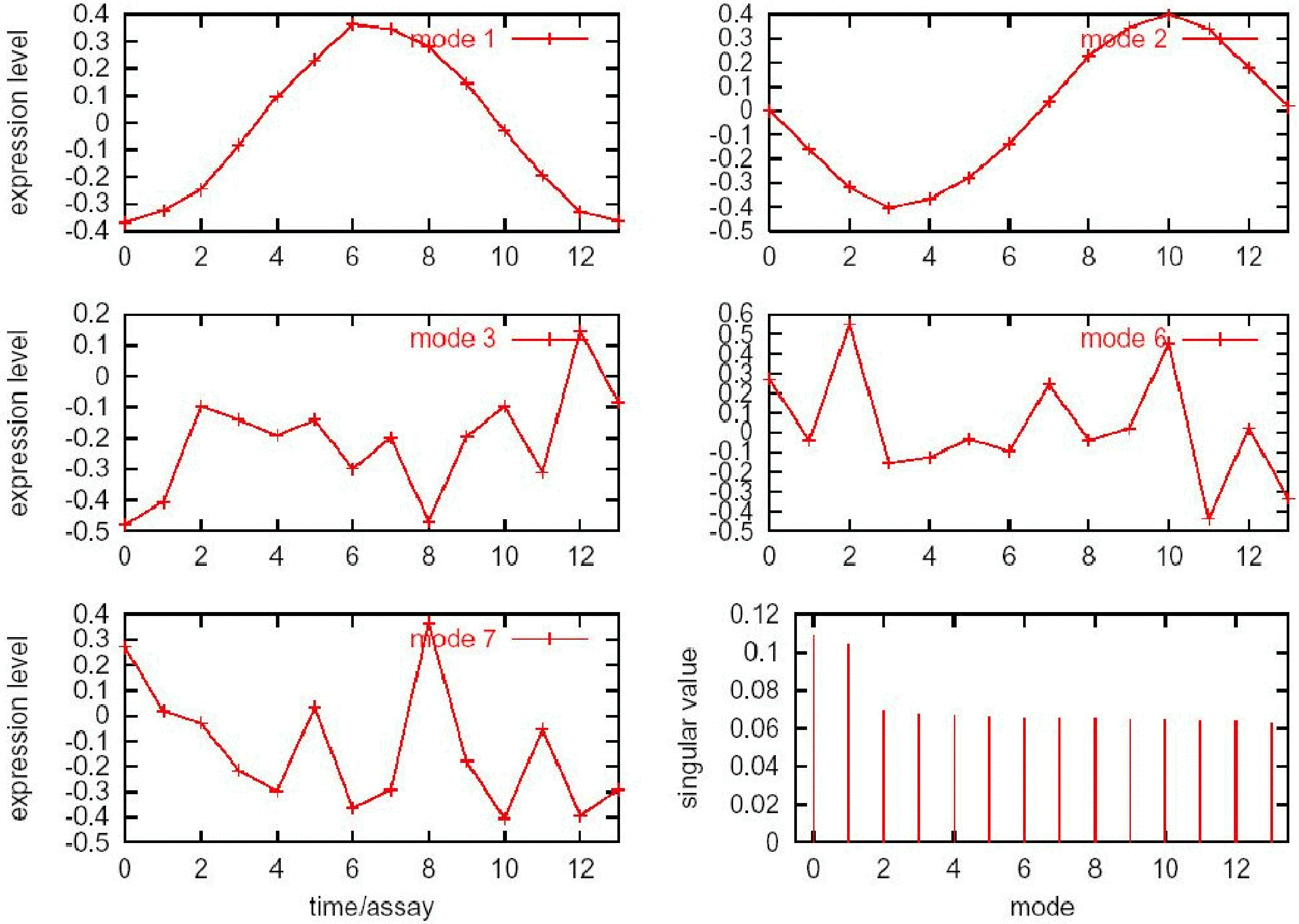
Comparison of SVD Methods with Artificial and Real Data



- 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

COMPUTER & COMPUTATIONAL SCIENCES

SVD of Artificial Data Set

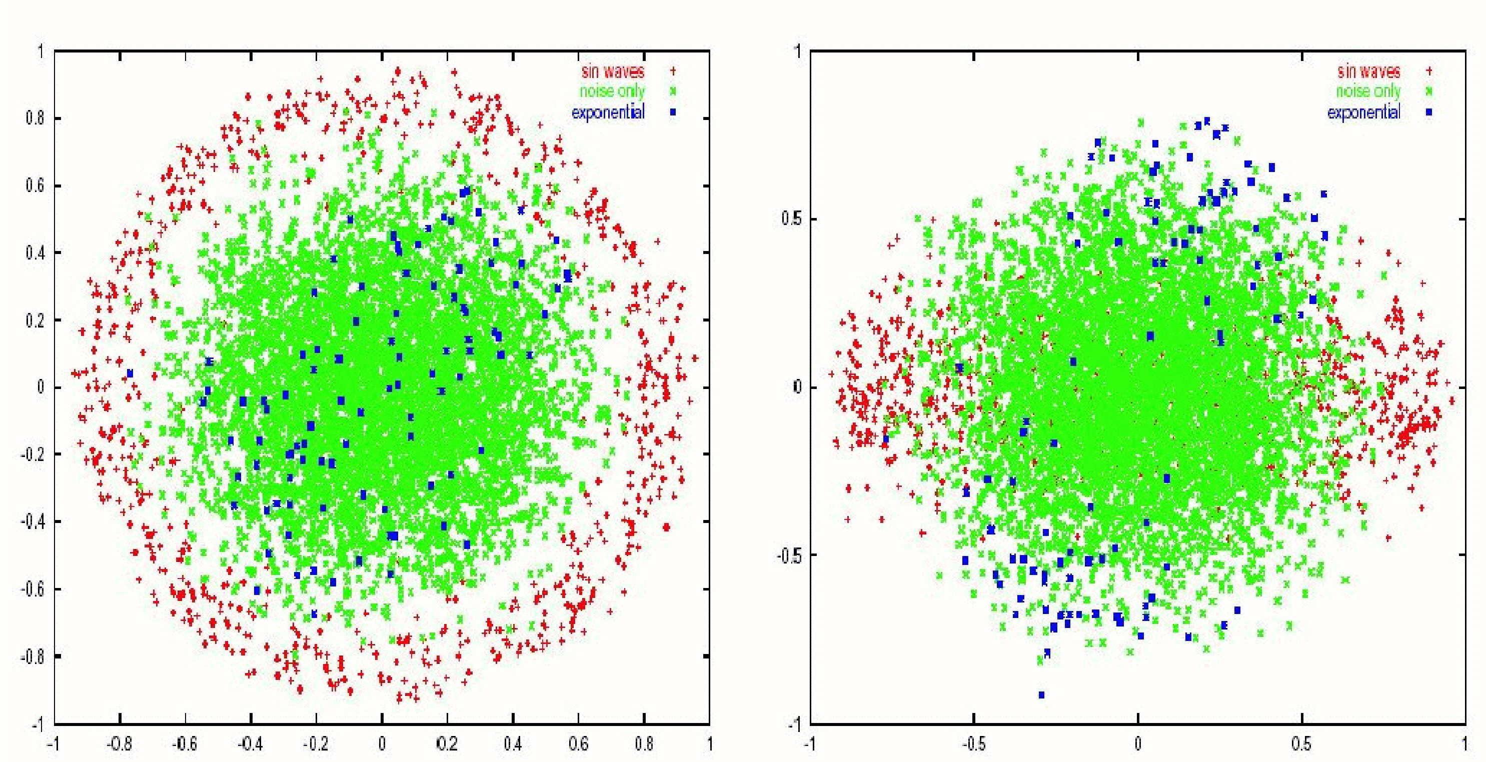


rocha@lanl.gov

COMPUTER & COMPUTATIONAL SCIENCES

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
2002

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

Los Alamos NATIONAL LABORATORY

Bioinformatics as Systems Biology

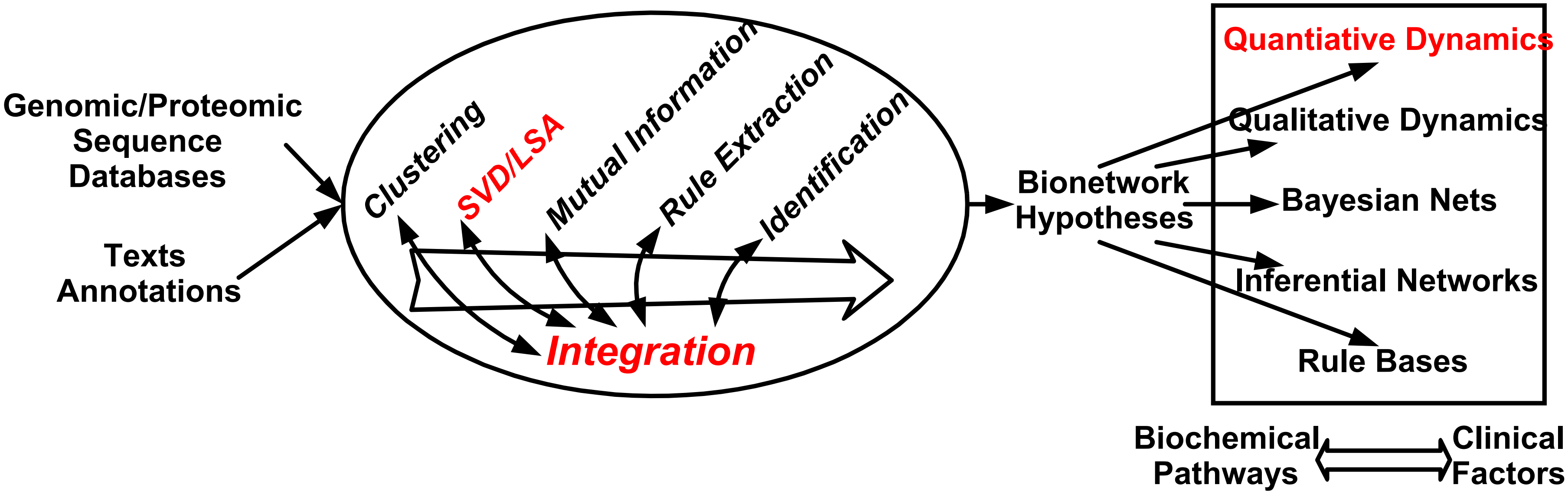
A Synthetic Multi-Disciplinary Approach to Biology

- 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, Complex Systems approach to Theoretical Biology
- Grand Challenge: 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

rocha@lanl.gov

Systems Biology

CCS Stance: Integration and Bionetwork Hypothesis



Gene Expression Analysis discovers patterns of expression behavior in groups of genes:

- numerical expression values without functional or semantic characterization

The biological reasons of gene groupings must be ascertained by biologists

- Need to be able to integrate knowledge about a large number of possible underlying biological mechanisms for a large number of genes in microarrays

Integration of available sources of functional knowledge

- databases with biomedical publications and data

COMPUTER & COMPUTATIONAL SCIENCES

Luis Rocha
2002

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

Los Alamos
NATIONAL LABORATORY

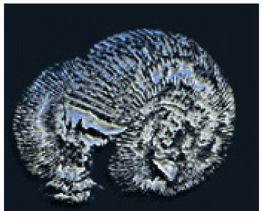
Curriculum For Bioinformatics

Graduate Study in Computational Biology

- Background
  - Knowledge of empirical sciences (Physics, Chemistry, Biology) and quantitative technical disciplines (programming, appplied mathematics, statistics)
- Graduate Program (adaptive):
  - Training in Biology
    - Basic theoretical concepts and experimental method
    - Courses: Molecular Biology, Genetics, Cell Biology, Immunology, Epidemiology, Neurology, etc...
  - Training in Computer Science
    - Programming,data structures, databases, web technology, robotics and automation, optimization, Artificial Intelligence and Life, Simulation, Autonomous Systems
  - Mathematics
    - Statistics, probability, stochastics processes, dynamical systems, measures of complexity and uncertainty, graph theory
  - Ethics
    - Privacy, Security, Technology and Social Issues, bioterrorism

http://www.smi.stanford.edu/projects/helix/bmi214/

Altman, R.B. (1998). Bioinformatics. 14, pp. 549-550



Computational Biology

Fundamental Concepts

- Pairwise Sequence Alignment and Multiple Sequence Alignment
  - Dynamic Programming, Simulated Annealing, Similarirty Matrices
- Hidden Markov Models
  - Alignment, Prediction
- Phylogenetic Trees
- Combinatorics
  - Sequencing
- RNA World
  - Structure Prediction
- Sequence feature extraction and annotation
- Proteomics
  - Homology Modeling, molecular dynamics, structure prediction
- Database integration and Design
- Optimization
  - Expectation Maximization, Monte Carlo Methods, Simulated Annealing, Gradient-based methods
- Dynamic programming, Bounded Search Algorithms, Cluster Analysis, Machine Learning, Bayesian Inference, Support Vector Machines, etc. etc.

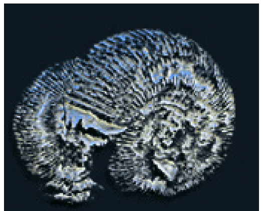
http://www.bioinf.man.ac.uk/ember/documentation.html

rocha@lanl.gov

Bioinformatics and Biomedicine

- Bioinformatics efforts that appear to be wholly geared towards basic science are likely to become relevant to clinical informatics in the coming decade. For example, DNA sequence information and sequence annotations will appear in the medical chart with increasing frequency. The algorithms developed for research in bioinformatics will soon become part of clinical information systems.
  - Linking of biomedical data for “clinical genomics”
  - Altman [1998]. Bioinformatics in Support of Molecular Medicine.

Luis Rocha
2002



Traditional Components of Bioinformatics

- Sequence Analysis
- Similarity Search and Motif Search
- Data-driven vs. Knowledge-based Functional Interpretation
- Sequence Alignment
- Dynamic Programming for Sequence Alignment Optimization
- Basics of FASTA Method
- Simulated Annealing and Genetic Algorithms for Multiple Sequence Alignment
- Basics of BLAST
- Hidden Markov Models
- Suffix Trees for Sequence Alignment
- Evolutionary Trees.

Sequence Analysis

Uncovering higher structural and functional characteristics from nucleotide and amino acid sequences

Data-Driven approach rather than first-principles equations.
Assumption:when 2 molecules share similar sequences, they are likely to share similar 3D structures and biological functions because of evolutionary relationships and/or physico-chemical constraints.

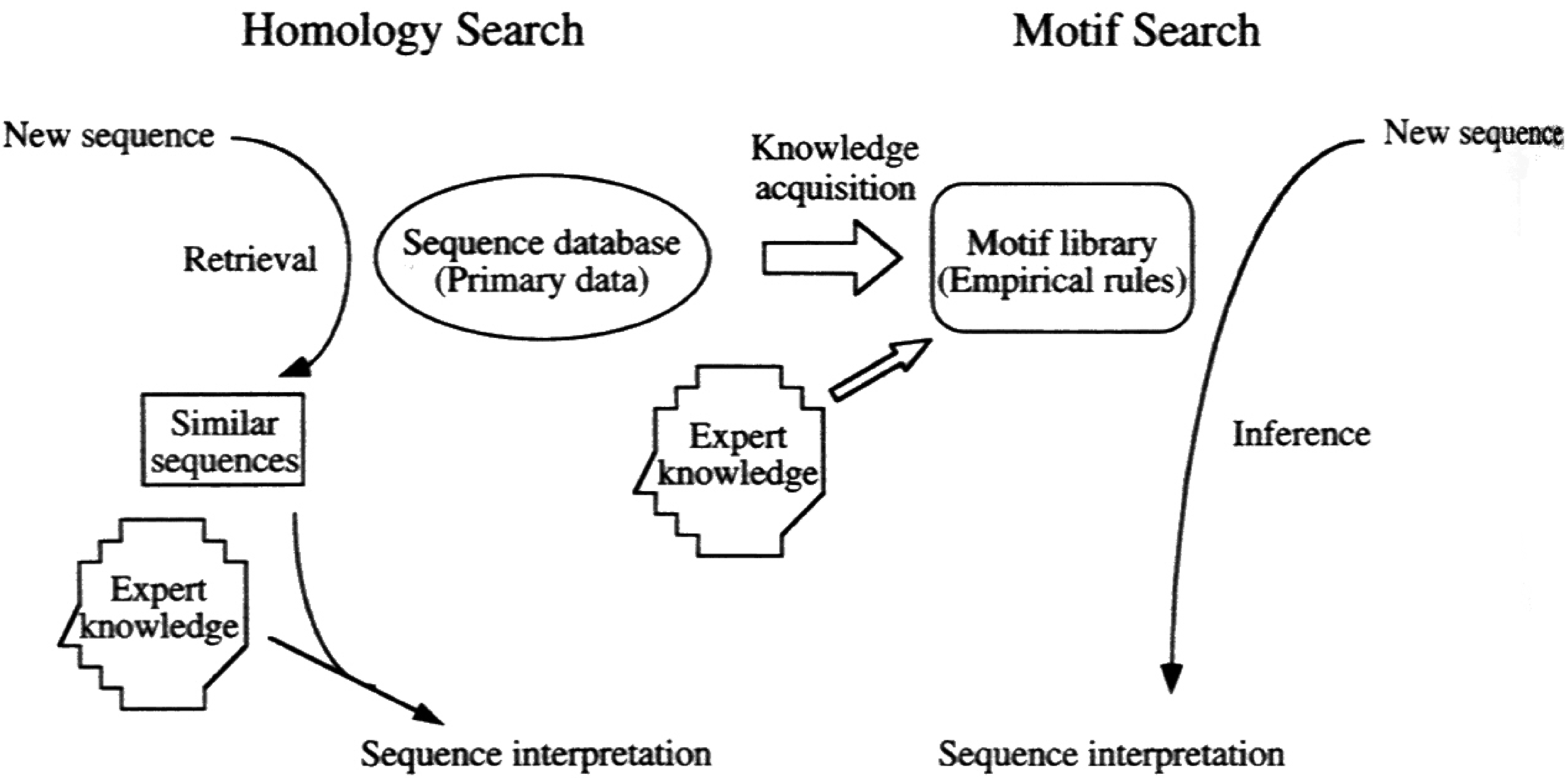
- Similarity (Homology) Search
  - Pairwise and multiple sequence alignment, database search, phylogenetic tree reconstruction, Protein 3D structure alignment
    - Dynamic programming, Simulated annealing, Genetic Algorithms, Neural Networks
- Structure/function prediction
  - Ab initio: RNA secondary and 3D structure prediction, Protein 3D structure prediction
  - Knowledge-based: Motif extraction, functional site prediction, cellular localization prediction, coding region prediction, protein secondary and 3D structure prediction
    - Discriminant analysis, Neural Networks, Hidden Markov Model, Formal Grammars

Similarity Search vs. Motif Search

Data-driven vs. Knowledge-based Functional Interpretation

- Similarity (Homology) Search
  - A query sequence is compared with others in a database. If a similar sequence is found, and if it is responsible for a specific function, then the query sequence can potentially have a similar function.
    - Like assuming that similar phrases in a language mean the same thing.
- Motif Search (Knowledge-based)
  - A query sequence is compared to a motif library, if a motif is present, it is an indication of a functional site.
    - A Motif is a subsequence known to be responsible for a particular function (often interaction sites with other molecules)
    - A Motif library is like a dictionary of sequence-function relationships: PROSITE (http://www.expasy.ch/sprot/prosite.html)
    - Unfortunately there are no comprehensive motif libarries for all types of functional properties

Similarity Search vs. Motif Search



Homology Search

Motif Search

New sequence

Retrieval

Sequence database (Primary data)

Similar sequences

Expert knowledge

Sequence interpretation

Knowledge acquisition

Motif library (Empirical rules)

Expert knowledge

New sequence

Inference

Sequence interpretation

Sequence Similarity Search

Sequence Alignment

- Produce the optimal (global or local) alignment of symbols that best reveals the similarity between 2 sequences (strings).
    - Minimizing gaps, insertions, and deletions while maximizing matches between elements using a scoring scheme

ALIGNMENT OF 2 STRINGS:

```
POST GENOME INFORMATICS IS THE FUTURE
GENOME HAS A FUTURE

POST GENOME INFORMATICS IS THE FUTURE
#####GENOME ##HAS A########### FUTURE
#####GENOME #####HAS###### A## FUTURE
```

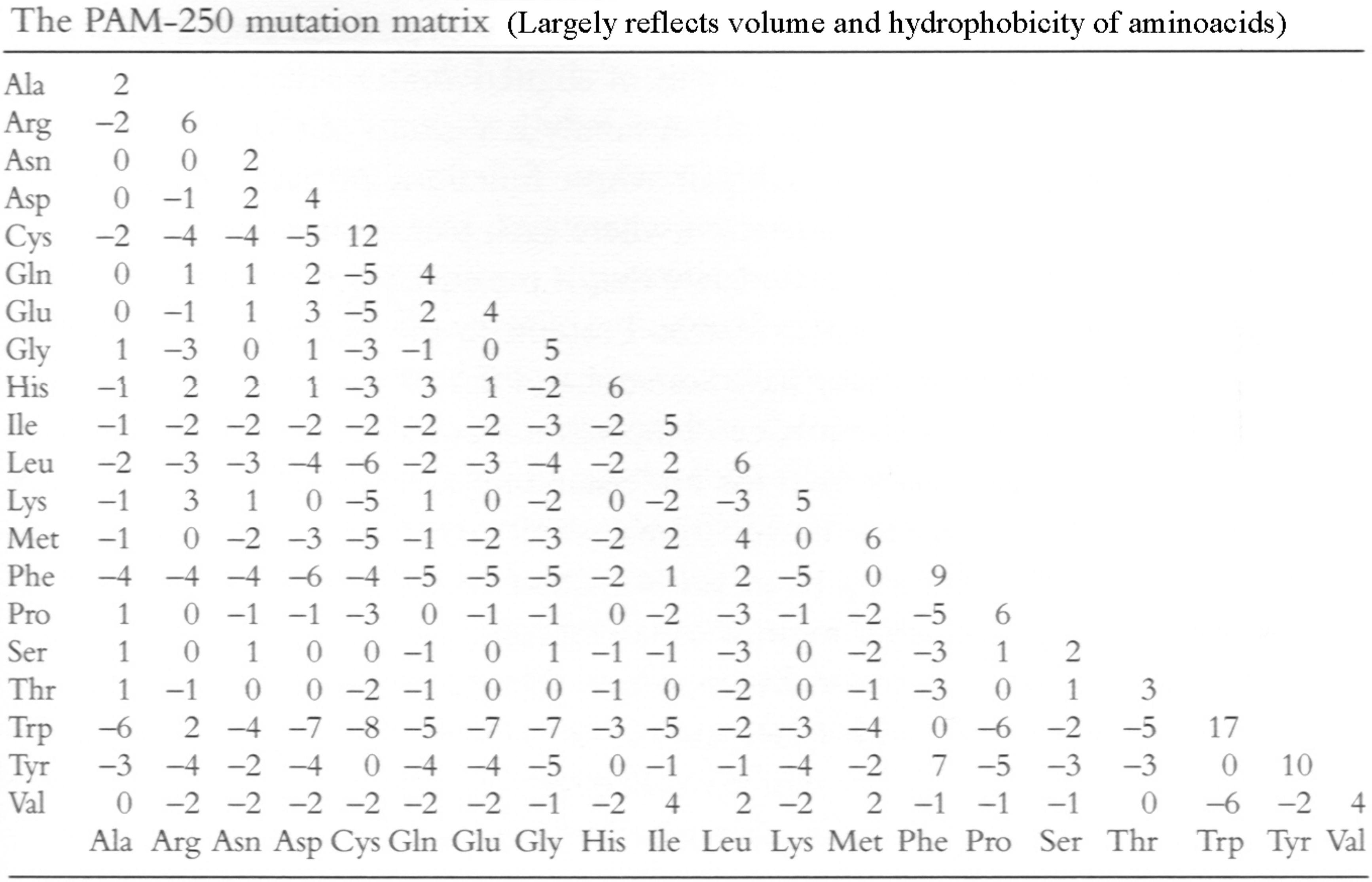
Sequence Similarity Search

Sequence Alignment in Biology

- Produce the optimal (global or local) alignment that best reveals the similarity between 2 sequences.
  - Minimizing gaps, insertions, and deletions while maximizing matches between elements.
  - DNA (RNA)
    - 4 (nucleoptide) symbol alphabet + gap
    - **TTGACAC**
    - **TTTACAC**
  - Proteins
    - 20 (aminoacid) symbol alphabet + gap
    - **RKVA--GMAKPNM**
    - **RKIAVAAASKPAV**
  - An emprirical measure of similarity between pairs of elements is needed (substitution scoring scheme)
    - Such as the amino acid mutation matrix

Dayhoff et al [1978] collected data for accepted point mutations (frequency of mutation) (PAMs) from groups of closely related proteins. Different matrices reflect different properties of amino acids (e.g. volume and hydrophobicity)
AAIndex: www.genome.ad.jp/dbget/aaindex.html

Mutation Matrix as Substitution Table



The PAM-250 mutation matrix (Largely reflects volume and hydrophobicity of aminoacids)

Ala	2																			
Arg	–2	6																		
Asn	0	0	2																	
Asp	0	–1	2	4																
Cys	–2	–4	–4	–5	12															
Gln	0	1	1	2	–5	4														
Glu	0	–1	1	3	–5	2	4													
Gly	1	–3	0	1	–3	–1	0	5												
His	–1	2	2	1	–3	3	1	–2	6											
Ile	–1	–2	–2	–2	–2	–2	–2	–3	–2	5										
Leu	–2	–3	–3	–4	–6	–2	–3	–4	–2	2	6									
Lys	–1	3	1	0	–5	1	0	–2	0	–2	–3	5								
Met	–1	0	–2	–3	–5	–1	–2	–3	–2	2	4	0	6							
Phe	–4	–4	–4	–6	–4	–5	–5	–5	–2	1	2	–5	0	9						
Pro	1	0	–1	–1	–3	0	–1	–1	0	–2	–3	–1	–2	–5	6					
Ser	1	0	1	0	0	–1	0	1	–1	–1	–3	0	–2	–3	1	2				
Thr	1	–1	0	0	–2	–1	0	0	–1	0	–2	0	–1	–3	0	1	3			
Trp	–6	2	–4	–7	–8	–5	–7	–7	–3	–5	–2	–3	–4	0	–6	–2	–5	17		
Tyr	–3	–4	–2	–4	0	–4	–4	–5	0	–1	–1	–4	–2	7	–5	–3	–3	0	10	
Val	0	–2	–2	–2	–2	–2	–2	–1	–2	4	2	–2	2	–1	–1	–1	0	–6	–2	4
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

rocha@lanl.gov

Dynamic Programming

For Sequence Alignment Optimization

Optimal alignment maximizing the number of matched letters

Score function: 1 for match, 0 for mismatch, 0 for insertion/deletion

3 matches, 2 mismatches, 2 gap insertions = 3

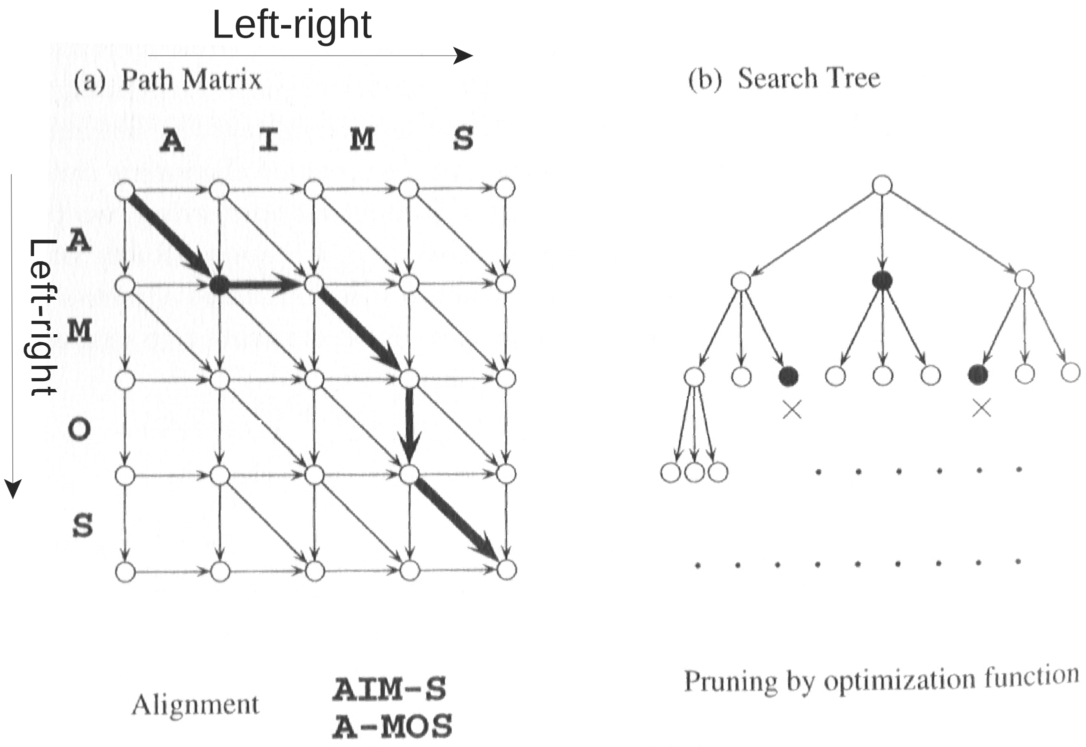
```
AIMS   →   AIM-S
AMOS       A-MOS
```

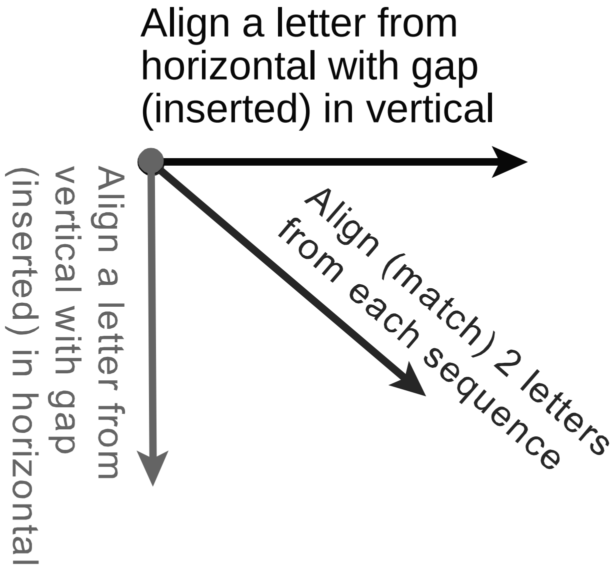
Dynamic programming is a very general optimization technique for problems that can recursively be divided into two similar problems of smaller size, such that the solution to the larger problem can be obtained by piecing together the solutions to the two subproblems. Example: shortest path between 2 nodes in a graph.

The first mathematical treatment is due to Richard Bellman (1957)

Dynamic Programming

Path Matrix

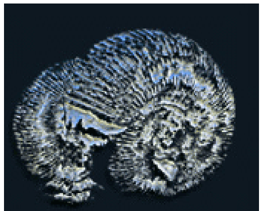



Align a letter from horizontal with gap (inserted) in vertical

Align (match) 2 letters from each sequence

Align a letter from vertical with gap (inserted) in horizontal

A path starting at the upper-left corner and ending at the lower-right corner of the path matrix is a global alignment of the two sequences. The optimal alignment is the optimal path in the matrix according to the score function for each of the 3 path alternatives at each node. Most path branches are pruned out locally according to the score function.



Global Sequence Alignment

With Dynamic Programming

- Score Function *D* (to optimize) sum of weights at each alignment position from a substitution matrix *W*
  - Nucleotide sequences
    - Arbitrary weights: a fixed value for a match or mismatch irrespective of the types of base pairs
  - Amino acid sequences
    - Needs to reveal the subtle sequence similarity. Substitution matrix constructed from the amino acid mutation frequency adjusted for different degrees of evolutionary divergence (since the table is built for closely related sequences)

$W_{s(i),t(j)}$ Weigth for aligning (Substituting) element i from sequence s with element j of sequence t

d Weigth for a single element gap

$$D_{i,j} = \max(D_{i\text{-}1,j\text{-}1} + W_{s(i),t(j)}, D_{i\text{-}1,j} + \mathrm{d}, D_{i,j\text{-}1} + \mathrm{d})$$

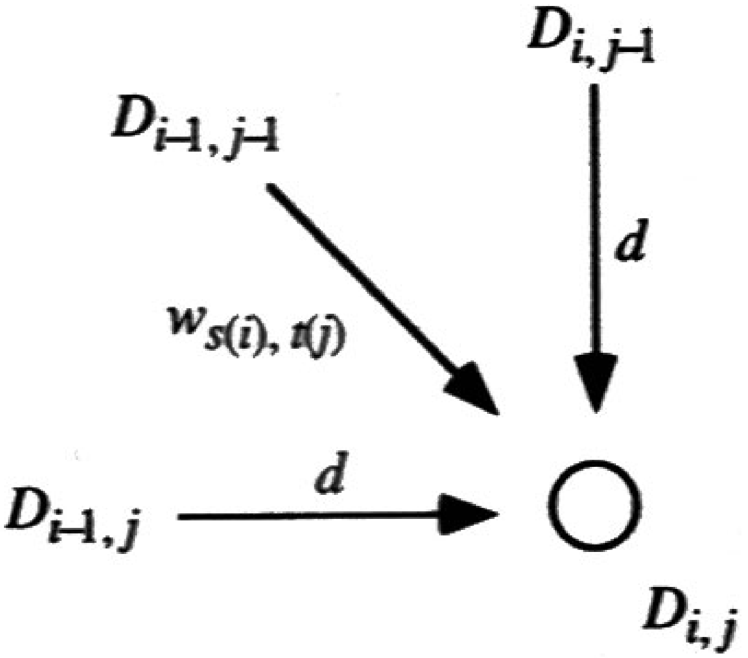
$$D_{0,0} = 0,\ D_{i,0}= id\ (i=1\ldots n),\ \ D_{0,j}= jd\ (j=1\ldots m)$$

rocha@lanl.gov

Global Alignment

$$D_{i,j} = \max(D_{i\text{-}1,j\text{-}1} + W_{s(i),t(j)},\ D_{i\text{-}1,j} + \mathrm{d},\ D_{i,j\text{-}1} + \mathrm{d})$$

$$D_{0,0} = 0,\ D_{i,0} = id\ (i=1...n),\ \ D_{0,j} = jd\ (j=1...m)$$



Starting at $D_{1,1}$, repeatedly applying the formula, thefinal $D_{n,m}$ is the optimal value of the score function for the alignment. The optimal path is reconstructed from the stored values of matrix D by tracing back the highest local values

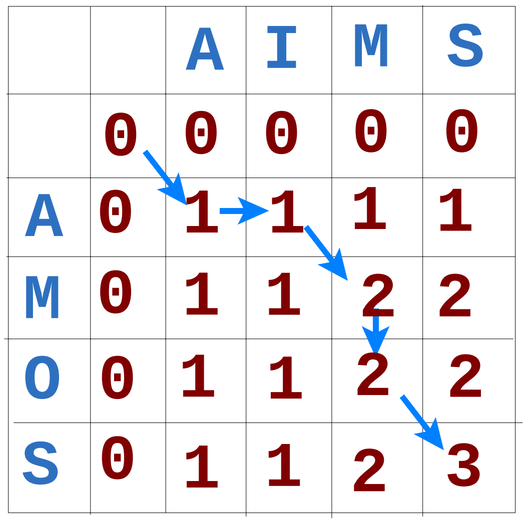
Number of operations proportional to the size of the matrix $n \times m$: $O(n^2)$

Needleman and Wunsch algorithm introduces a gap length dependence with a gap opening and elongation penalty.

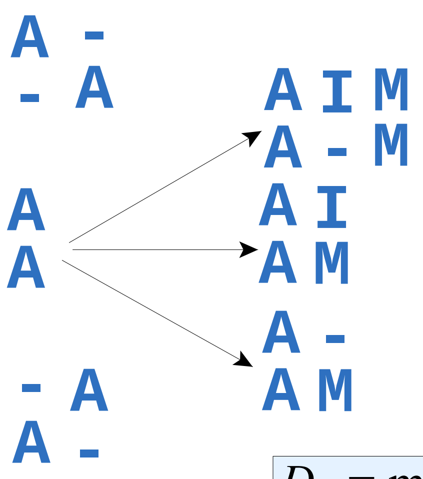
rocha@lanl.gov

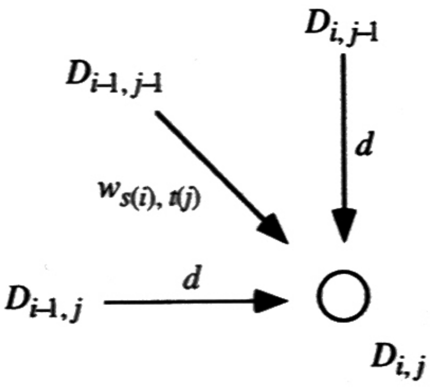
Global Alignment

Toy Example: Maximization



		A	I	M	S
	0	0	0	0	0
A	0	1	1	1	1
M	0	1	1	2	2
O	0	1	1	2	2
S	0	1	1	2	3

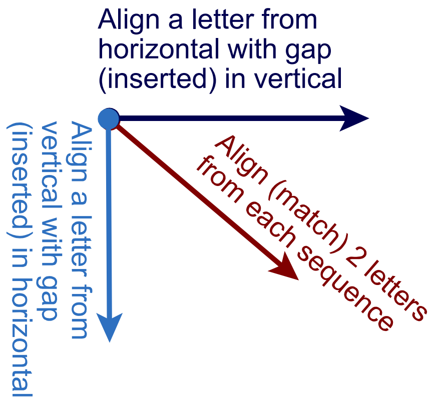




Score function: 1 for match, 0 for mismatch, 0 for gap

$$D_{i,j} = \max(D_{i-1,j-1} + W_{s(i),t(j)},\ D_{i-1,j} + \mathrm{d},\ D_{i,j-1} + \mathrm{d})$$

$$D_{0,0} = 0,\ D_{i,0} = id\ (i=1\ldots n),\ D_{0,j} = jd\ (j=1\ldots m)$$


Align a letter from horizontal with gap (inserted) in vertical
Align a letter from vertical with gap (inserted) in horizontal
Align (match) 2 letters from each sequence

```
AIM-S
A-MOS
```

3 matches, 2 mismatches, 2 gap insertions = 3

Backtrack: Maintains Pointer of previous max path

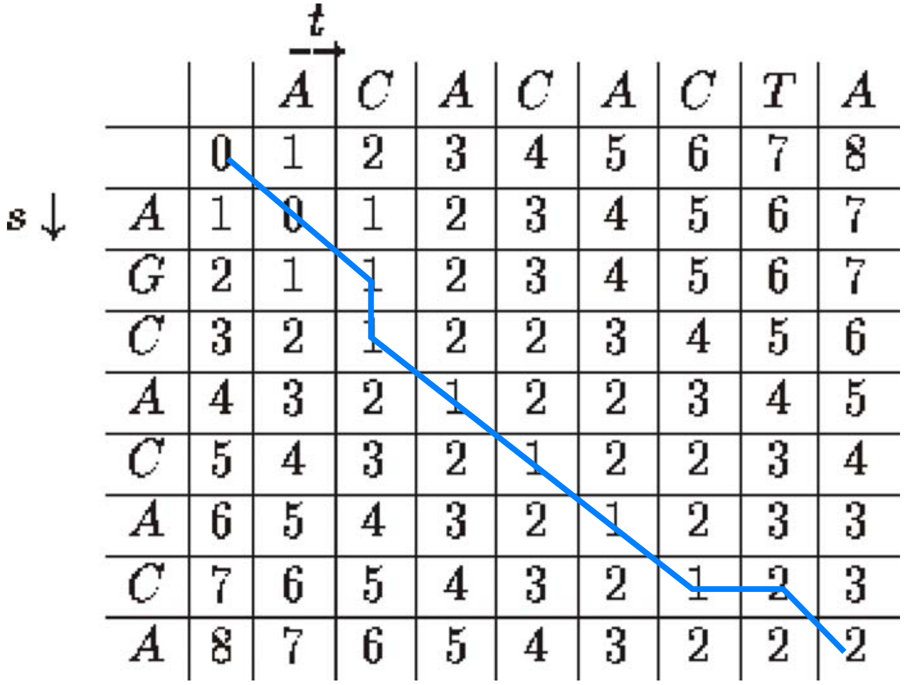
Several Optimal Alignments are possible. Backtracking can be computationally expensive if all branches are pursued. Making arbitrary decisions on what pointers to follow, then the computation complexity is O(N). For DP is O(N^2)

Luis Rocha
2002

Sequence Alignment

Nucleotide Sequence: Minimization

$$s = AGCACACA,\ t = ACACACTA$$



		A	C	A	C	A	C	T	A
	0	1	2	3	4	5	6	7	8
A	1	0	1	2	3	4	5	6	7
G	2	1	1	2	3	4	5	6	7
C	3	2	1	2	2	3	4	5	6
A	4	3	2	1	2	2	3	4	5
C	5	4	3	2	1	2	2	3	4
A	6	5	4	3	2	1	2	3	3
C	7	6	5	4	3	2	1	2	3
A	8	7	6	5	4	3	2	2	2

Score function: 0 for match, 1 for mismatch, 1 for gap

```
A-CACACTA
AGCACAC-A
```

Aminoacid sequence alignment has much more complicated substitution scores

http://merlin.mbcr.bcm.tmc.edu:8001/bcd/Curric/PrwAli/node3.html

Local Alignment

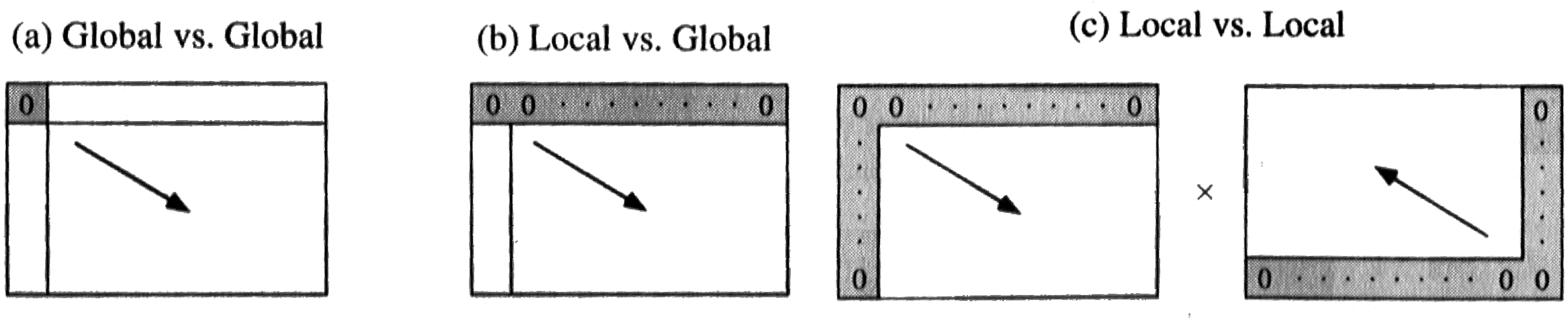
Goal :Alignment of subsequences

$$D_{i,j} = \max(D_{i-1,j-1} + W_{s(i),t(j)}, D_{i-1,j} + \mathrm{d}, D_{i,j-1} + \mathrm{d})$$

$$D_{0,0} = 0, D_{i,0}= id\ (i=1...n),\ D_{0,j}= jd\ (j=1...m)$$

$$\boxed{D_{0,j}= 0\ (j=1...m)}$$

Any letter in the horizontal sequence can be a starting point without any penalty: detects multiple matches within the horizontal sequence containing multiple subsequences similar to the vertical sequence


(a) Global vs. Global

(b) Local vs. Global

(c) Local vs. Local

Local Alignment

Smith-Waterman Local Optimality Algorithm

$$D_{i,j} = \max(D_{i-1,j-1} + W_{s(i),t(j)},\ D_{i-1,j} + \mathrm{d},\ D_{i,j-1} + \mathrm{d})$$

$$D_{0,0} = 0,\ D_{i,0} = id\ (i=1\ldots n),\ \ D_{0,j} = jd\ (j=1\ldots m)$$

$$D_{i,j} = \max(D_{i-1,j-1} + W_{s(i),t(j)},\ D_{i-1,j} + \mathrm{d},\ D_{i,j-1} + \mathrm{d},\ 0)$$

$W_{s(i),t(j)} > 0$ match $\qquad W_{s(i),t(j)} < 0$ mismatch $\qquad d < 0$

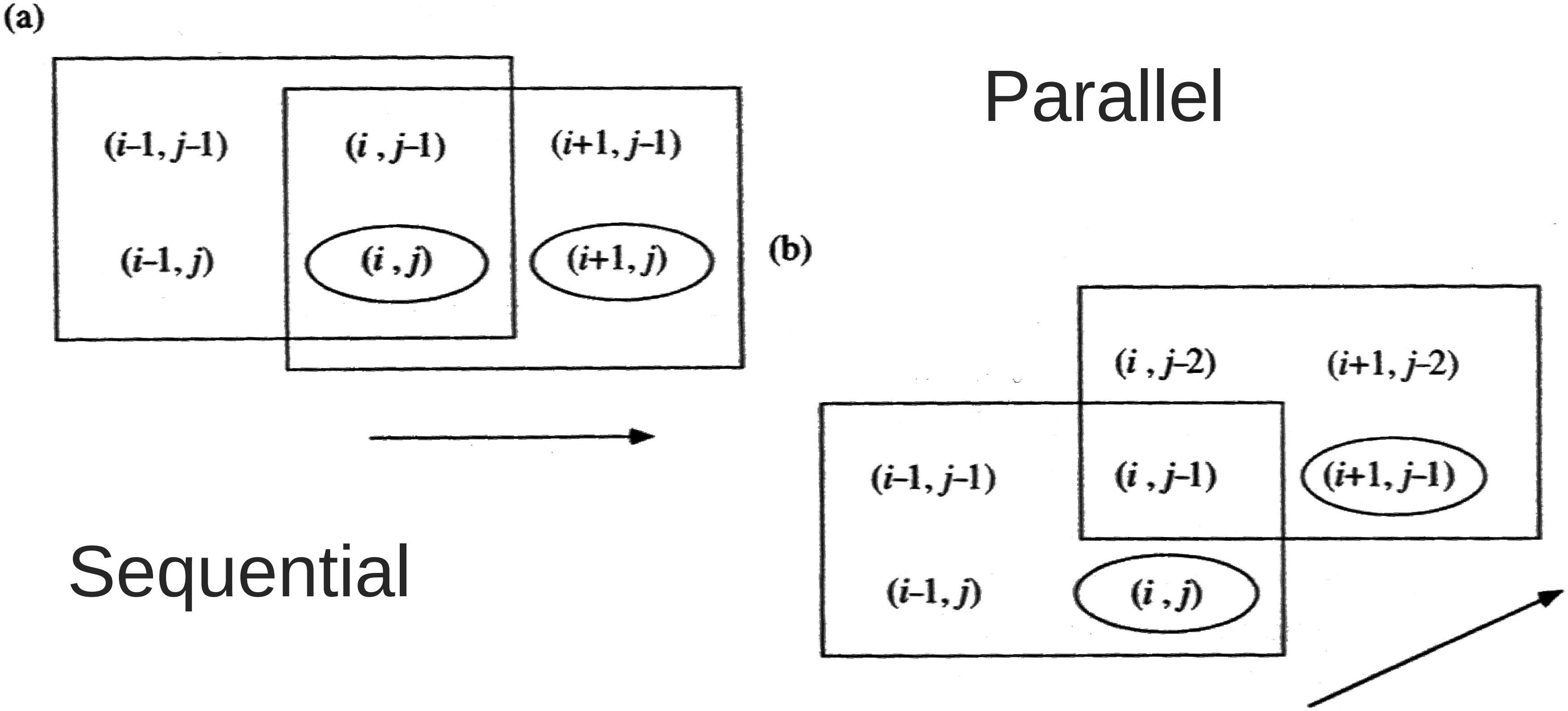
Forces local score for match to be non-negative and for mismatch to be negative. Optimal path is not entered, but clusters of favourable local alignment regions. Trace back starts at the matrix element with maximum score.

http://www.cse.ucsc.edu/research/kestrel/runkestrel.html

Similarity Database Search

Parallelized Dynamic Programming

Number of operations in DP is proportional to the size of the matrix *n*×*m*: $O(n^2)$ – a lot for a large database of sequences!



Parallel

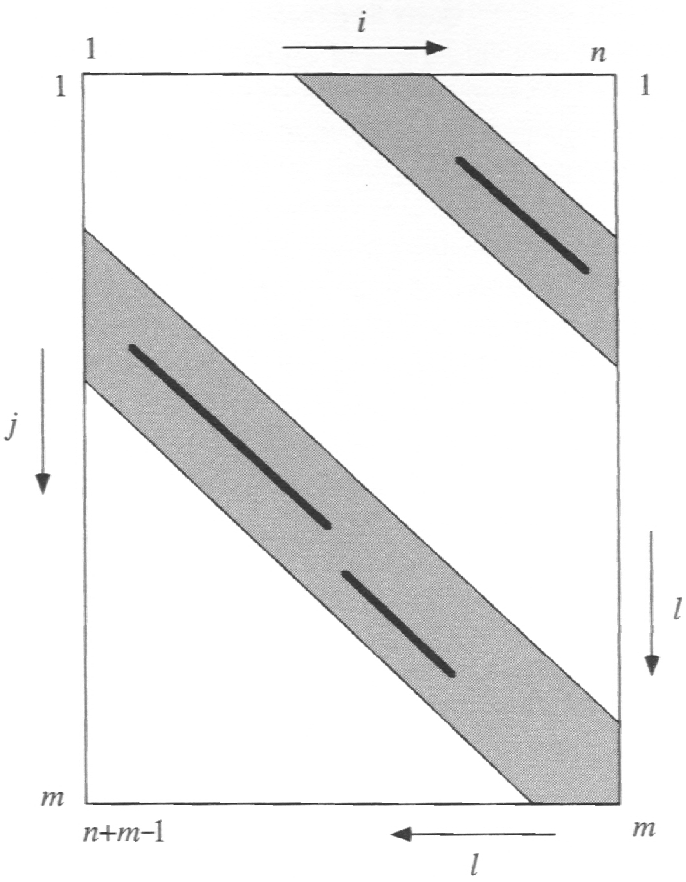
Sequential

FASTA Method

Dot Matrix Reduces DP Search Area

	A	I	M	S
A	*			
M			*	
O				
S				*

Dot Matrix



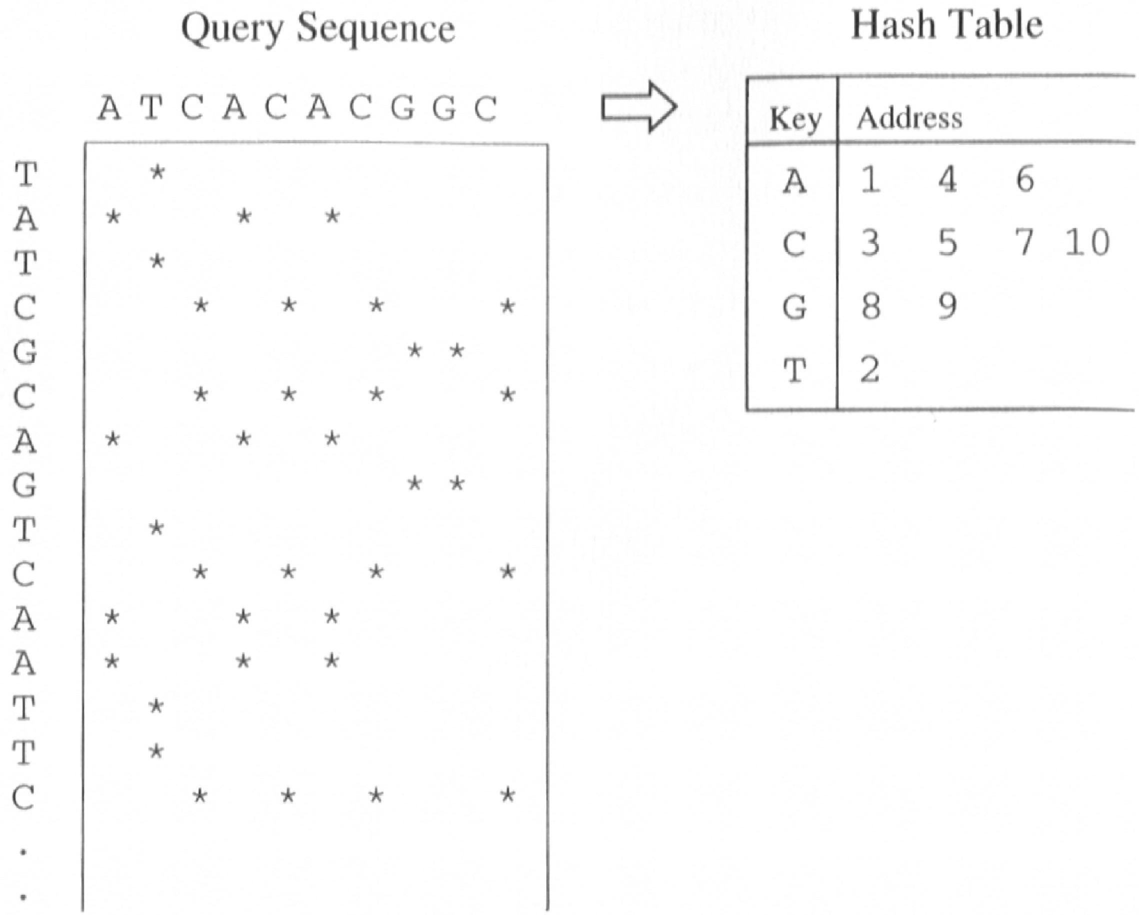
The dot matrix can be used to recognize local alignments which show as diagonal stretches or clusters of diagonal strectches. DP can be used only for the portions of the matrix around these clusters – a limited search area.

rocha@lanl.gov

FASTA

Hashing the Dot Matrix

Better than BLAST
for DNA Sequences



Query Sequence

	A	T	C	A	C	A	C	G	G	C
T		*								
A	*			*		*				
T		*								
C			*		*		*			*
G								*	*	
C			*		*		*			*
A	*			*		*				
G								*	*	
T		*								
C			*		*		*			*
A	*			*		*				
A	*			*		*				
T		*								
T		*								
C			*		*		*			*
.										
.										

Hash Table

Key	Address			
A	1	4	6	
C	3	5	7	10
G	8	9		
T	2			

Rapid access to stored data items by hashing. Sequences are stored as hash (look-up) tables. This facilitates the sequence comparison to produce a dot matrix. 4 times faster for nucleotide sequences: the number of operations is proportional to the mean row size of the hash table (times dots entered), which is on average 1/4 of the sequence.

http://www.ebi.ac.uk/fasta33/

rocha@lanl.gov

FASTA

More Details

```
position   1 2 3 4 5 6 7 8 9 10 11
protein 1  n c s p t a . . . . .
protein 2  . . . . . a c s p r k
```

amino acid	position in		offset
	protein 1	protein 1	pos 1 - pos2
a	6	6	0
c	2	7	-5
k	-	11	
n	1	-	
p	4	9	-5
r	-	10	
s	3	8	-5
t	5	-	

Note the common offset for the 3 amino acids c,s and p
A possible alignment is thus quickly found -

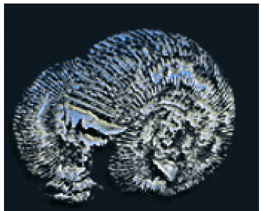
```
protein 1 n c s p t a
            | | |
protein 2 a c s p r k
```

Number of comparisons: O(n)
in DP it is O(n^2)

Usually with words (k-tuples)
length is typically 1 or 2 for protein sequences and 5-20 (6) for nucleotide sequences

The larger the k-tuple chosen, the more rapid but less thorough, a database search is.
AC ≠ AG are mismatch, not partial match

Words that have the same offset position reveal a region of alignment between the two sequences.



Statistical Significance

Is the similarity found biologically significant?

Because good alignments can occur by chance alone, the statistics of alignment scores help assess the significance. We know that the average alignment score for a query sequence with fixed length *n* increases with the logarithm of length *m* of a database sequence. Thus, the distribution of sequence lengths in the database can be used to estimate empirically the value of the expected frequency of observing an alignment with high score.

Another idea is to use the Z-test:

$$Z = \frac{S - \mu}{\sigma}$$

S is the optimal alignment score between 2 sequences

Each sequence is randomized k times (preserving the composition) and new optimal alignment is computed: s1, s2, ..., sk with mean μ and standard deviation σ. If the score distribution is normal, Z values of 4 and 5 correspond to threshold probabilities of 3×10^{-5} and 3×10^{-6}. However, the distribution typically decays exponentially in S rather than S^2 (as in the normal distribution). Thus, a higher Z value should be taken as a threshold for significant similarity.

Multiple Alignment

Simultaneous Comparison of a Group of Sequences

- Reasons for Multiple Alignment
  - Summarize classes of related proteins (motifs)
  - Assess conservation over several proteins
  - Establish Evolutionary Relationships
    - History of proteins in evolution
  - Help model 3D strucures
    - What other aminoacids are possible?

- DP can be expanded to a n-dimensional search space.
  - Exhaustive search is manageable for 3, and for a limited portion of the space for up to 7 or 8 sequences.
- Heuristics and approximate algorithms
  - Compute score for sequences A-C, from A-B, and B-C
    - which is in general different from the optimal A-C.
  - Hierarchical Clustering of a set of sequences, from a distance matrix computed from pairwise sequence alignment

Hierarchical Clustering

Given a set of N items and an NxN distance matrix:

1. Assign each item to its own cluster, producing N clusters, each containing just one item.
2. Find the closest pair of clusters and merge them into a single cluster.
3. Compute distances between the new cluster and each of the old clusters.
4 . Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Distances between clusters:

Single-link clustering: shortest distance from any member of one cluster to any member of the other cluster.

Complete-link clustering: farthest distance from any member of one cluster to any member of the other cluster.

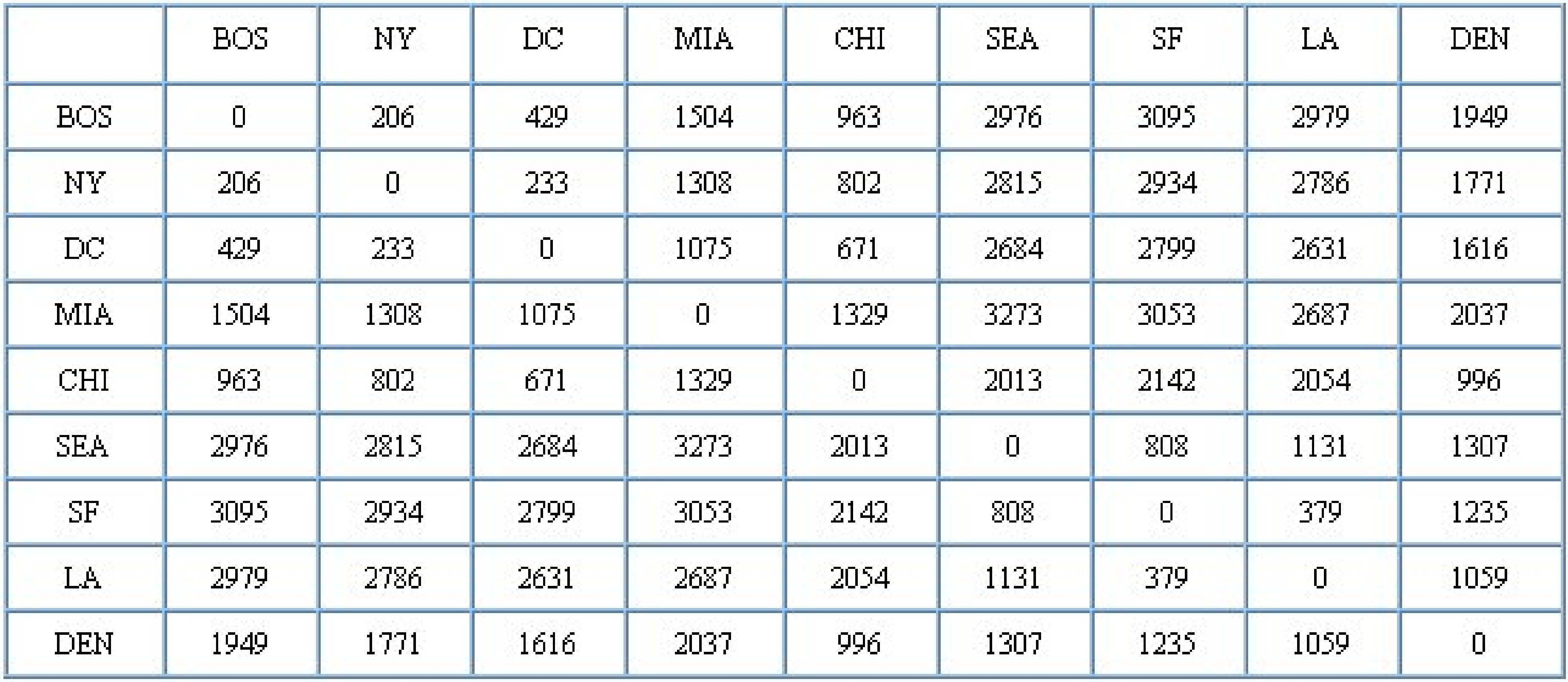
Average-link clustering: average distance from one cluster to the other cluster.

rocha@lanl.gov

Hierarchical Clustering

Single-link clustering: shortest distance from any member of one cluster to any member of the other cluster.

City Example



	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Given a set of N items and an NxN distance matrix:

1. Assign each item to its own cluster, producing N clusters
2. Find the closest pair of clusters and merge them into a single cluster.
3. Compute distances between the new cluster and each of the old clusters.
4 . Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

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

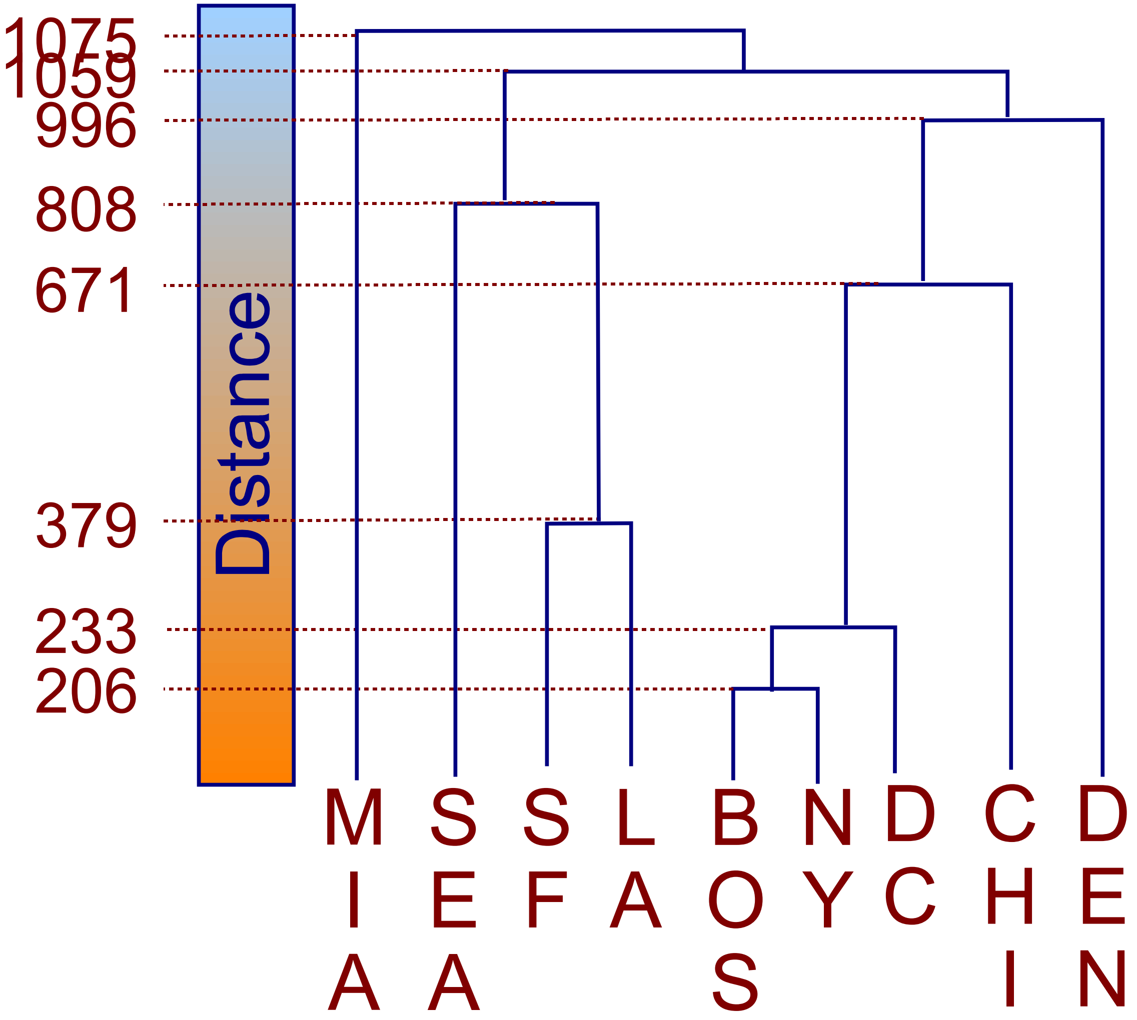
Luis Rocha
2002

rocha@lanl.gov

Hierarchical Clustering

City Example: Dendogram

http://www.analytictech.com/networks/hiclus.htm

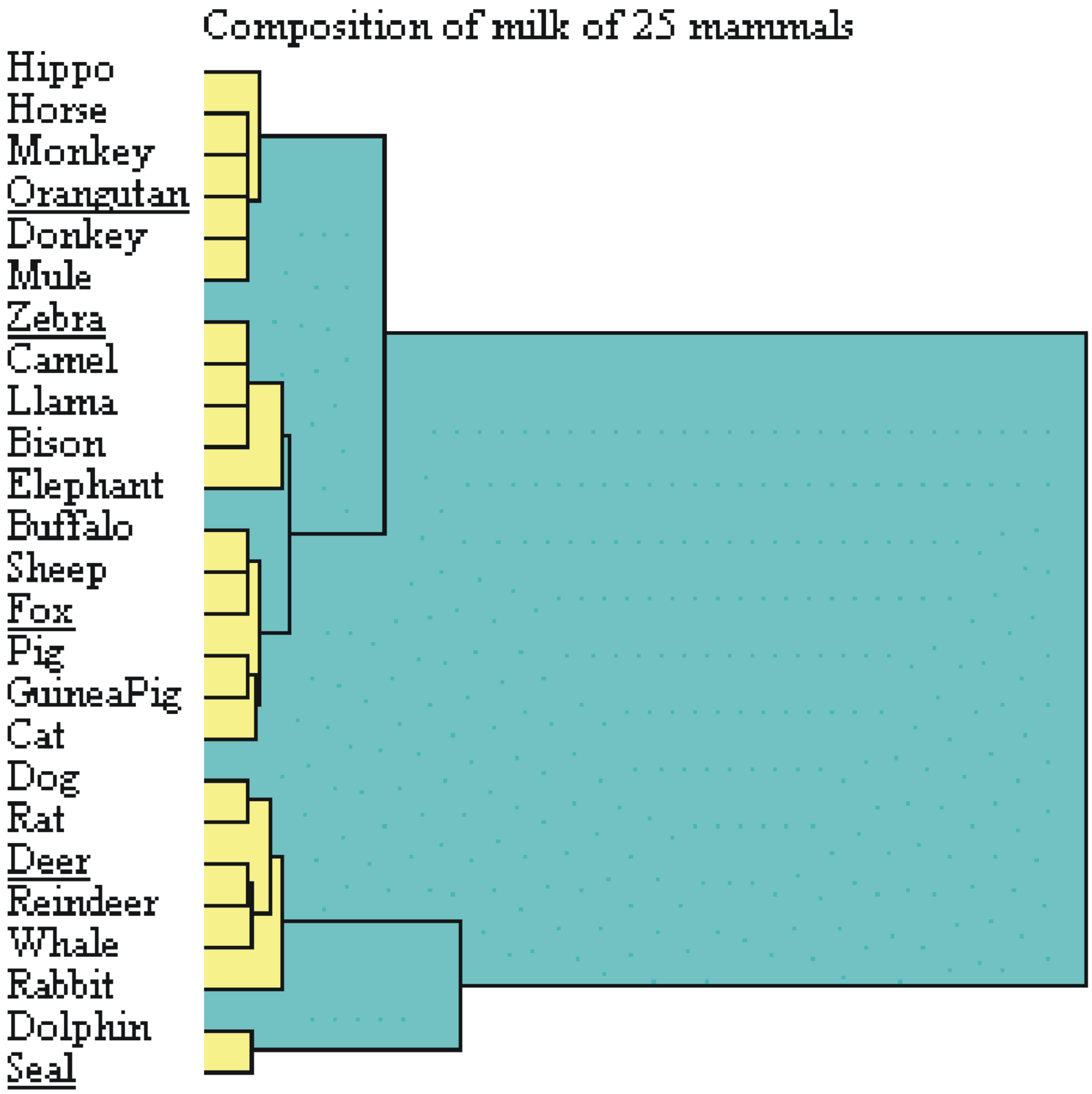


Single-link clustering: shortest distance from any member of one cluster to any member of the other cluster.

Luis Rocha
2002

Hierarchical Clustering

Mammal Milk Example

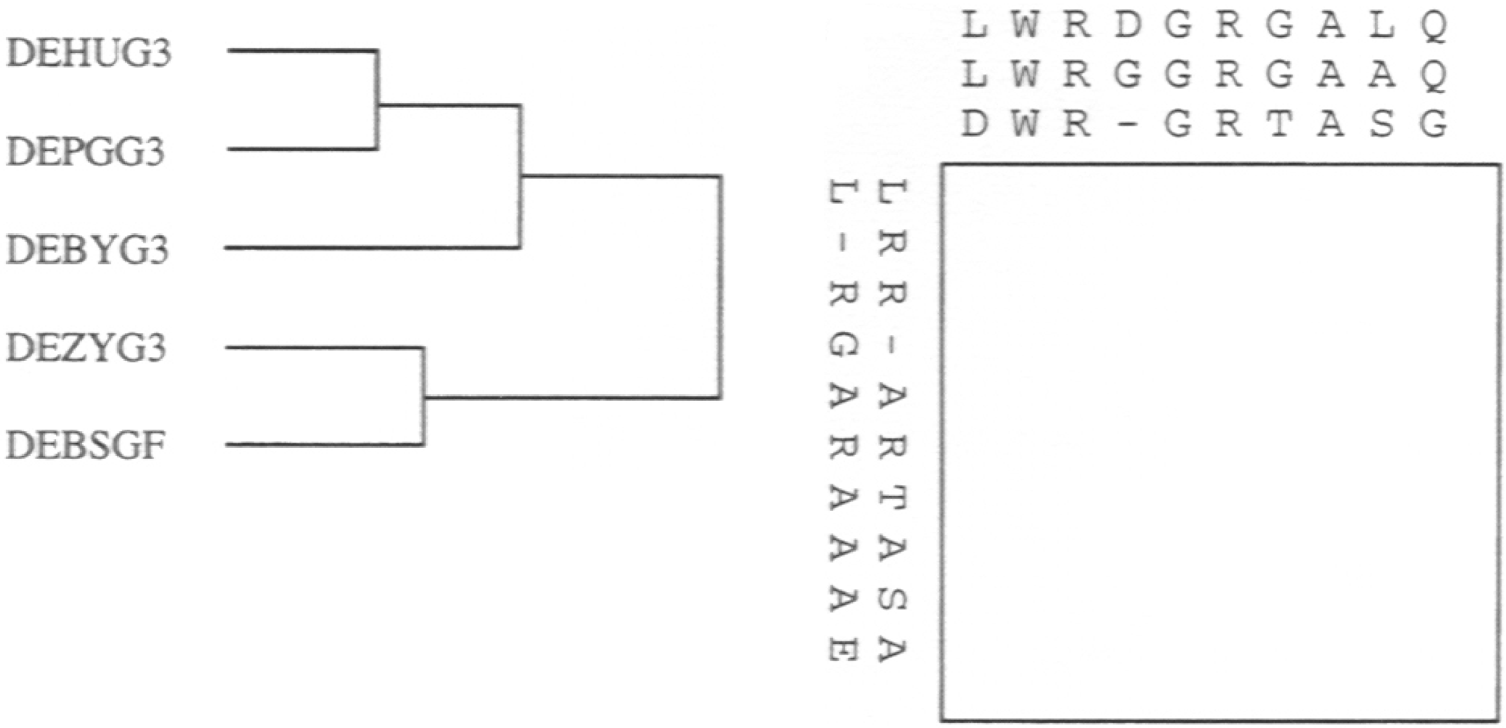


http://www.clustan.com/hierarchical_cluster_analysis.html

Multiple Sequence Alignment

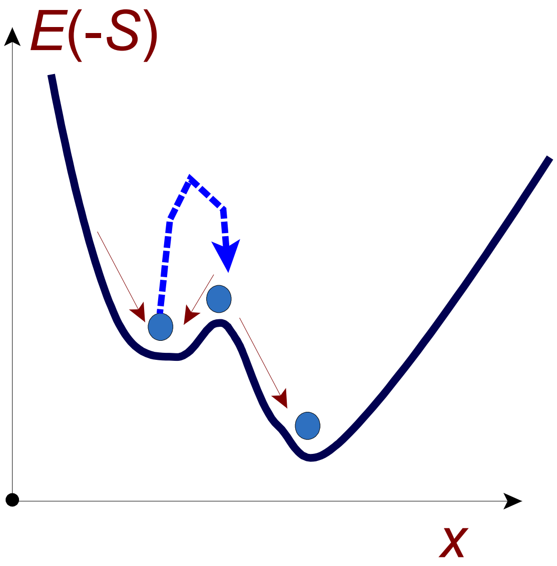
With Hierarchical Clustering

- Distance matrix computed from optimal pairwise sequence alignment
- Followed by computation of the alignment between groups of sequences without changing the predetermined alignment within each group.
  - Or using iterative procedure



Simulated Annealing

For Multiple Alignment



- SA is a stochastic method to search for global minimum in the optimization of functions to be minimized.
  - Starting with a given alignment for a set of sequences, a small random modification is repeatedly introduced and a new score is calculated. When the score is better (negative energy function), it is accepted.
  - Would Not escape local minima
- A stochastic unfavourable modification is accepted with (Metropolis Monte Carlo) probability:

$$p = e^{(-\Delta E/T)}$$

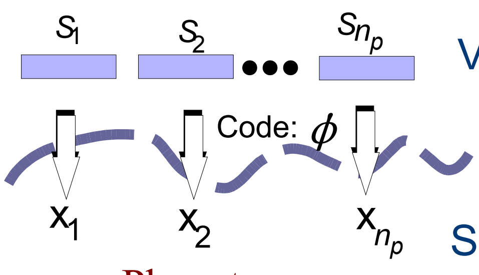
  - ΔE is the increment of the energy function from the modification. T is a simulated temperature parameter. The probability is calculated until equilibrium is reached. Then the temperature is lowered, and so on.
- Global miniumum is guaranteed for infinite MMC steps and infinitesimal ΔT.
  - Success depends onTi, Tf, ΔT, and # of MMC steps

Genetic Algorithms

For Multiple Sequence Alignment

Traditional Genetic Algorithm

Genotype



Variation

Selection

Phenotype

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.

- GAs are another stochastic method used for optimization.
  - Solutions to a problem are encoded in bit strings.
  - The best decoded solutions are selected for the next population (e.g. by roulette wheel or Elite)
  - Variation is applied to selected new population (crossover and mutation).

Other Bioinformatics Technology

Major Components not Fully Discussed

- BLAST
  - Heuristic algorithm for sequence alignment that incorporates good guesses based on the knowledge of how random sequences are related.
- Prediction of structures and functions
  - Neural Networks and Hidden Markov Models

Literature

Bioinformatics Overviews

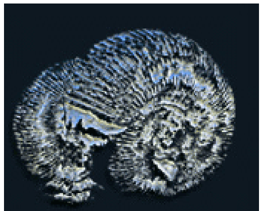
- Kanehisa, M. [2000]. *Post-Genome Informatics*. Oxford University Press.
- Waterman, M.S. [1995] *Introduction to Computational Biology*. Chapman and Hall.
- Baldi. P. and S. Brunak [1998]. *Bioinformatics: The Machine Learning Approach*. MIT Press.
- Wada, A. [2000]. “Bioinformatics – the necessity of the quest for ‘first principles’ in life”. *Bioinformatics*. V. 16, pp. 663-664. (http://bioinformatics.oupjournals.org/content/vol16/issue8)
- Altman, R.B. [1998]. A Curriculum for Bioinformatics: The Time is Ripe. Bioinformatics 14(7):549-550,
- Altman, R.B. [1998]. Bioinformatics in Support of Molecular Medicine. In C.G. Chute, Ed., 1998 AMIA Annual Symposium, Orlando, FL, 53-61. 1998.
- Altman’s Biomedical Informatics course: http://www.smi.stanford.edu/projects/helix/bmi214/
- EMBER Bioinformatics Resources: http://www.bioinf.man.ac.uk/ember/documentation.html

Systems Science and Complex Systems

- von Bertallanfy [1968] General System Theory. Foundations, Development, Applications, New York 1968
- Cariani, Peter [1989]. On The Design of Devices With Emergent Semantic Functions. PhD Dissertation. State University of New York at Binghamton.
- Conrad, Michael [1983]. Adaptability. Plenum Press.
- Kauffman, S. [1993]. The Origins of Order: Self-Organization and Selection in Evolution. Oxford university Press.
- Klir, George, J. [1991]. Facets of Systems Science. Plenum Press.
- Mesarovic, MD: (1968) "Auxiliary Functions and Constructive Specification of Gen. Sys.", /Mathematical Systems Theory, v. 2:3
- Pattee, Howard H. [1982]."Cell psychology: an evolutionary approach to the symbol-matter problem." Cognition and Brain Theory. Vol. 5, no. 4, pp. 191-200.
- Rosen, Robert [1991]. Life Itself. Columbia University Press.

New Systems Biology

- Institute for Systems Biology: http://www.systemsbiology.org
- Kitano Symbiotic Systems Project: http://www.symbio.jst.go.jp/



Literature

Dynamic Programming and Sequence Alignment

- Bellman, R.E. [1957] *Dynamic Programming*. Princeton University Press, Princeton,
- Bertsekas, D. [1995]. *Dynamic Programming and Optimal Control*. Athena Scientific.
- Needleman, S. B. and Wunsch, C. D. [1970]. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”.*J. Mol. Biol*., 48,443-53.
- Giegerich, R. [2000]. “A systematic approach to dynamic programming in bioinformatics”. *Bioinformatics*. V. 16, pp. 665-677.
- Sankoff, D. [1972]. Matching sequences under deletion/insertion constraints. *Proc. Natl. Acad. Sci*. USA, 69,4-6.
- Sellers, P. H [1974]. “On the theory and computation of evolutionary distances”. *SIAM J. Appl. Mat .,* 26,787-793.
- Sellers, P. H. [1980]. The theory and computation of evolutionary distances: pattern recognition. *Algorithms*, 1,359-73.
- Smith, T. F. and Waterman, M. S. [1981] . “Identification of common molecular subsequences”. *J.Mol. Biol*., 147,195--7.
- Goad, W. B. and Kanehisa, M. I. [1982]. “Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and Symmetries”. *Nucleic Acids Res*., 10, 247-63.
- Scientific Computation (Gaston Gonnet) http://linneus20.ethz.ch:8080/, section on DP: http://linneus20.ethz.ch:8080/4_6.html
- Probabilistic Dynamic Programming and Multiple Alignments (gaston Gonnet): http://www.inf.ethz.ch/personal/gonnet/papers/ProbAncSeq/node13.html
- Pairwise Sequence Alignment: http://merlin.mbcr.bcm.tmc.edu:8001/bcd/Curric/PrwAli/prwali.html
  - Pairwise Alignment via Dynamic Programming http://merlin.mbcr.bcm.tmc.edu:8001/bcd/Curric/PrwAli/node3.html
- Hardware Protein Database Search using Local Alignment (Smith-Waterman algorithm): http://www.cse.ucsc.edu/research/kestrel/runkestrel.html

Literature

- Similarity Matrices
  - Dayhoff, M. 0., Schwartz, R. M. and Orcutt, B.C. [1978] “A model of evolutionary change in proteins”. In *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3 (ed. M. 0. Dayhoff), pp. 345--52. National Biomedical Research Foundation, Washington, DC.
  - Henikoff, S. and Henikoff, J. G. [1992]. Amino acid substitution matrices from protein blocks. *Proc. Natl.Acad. Sci*. USA,89, 10915--19.
- FASTA algorithm and BLAST algorithm
  - Wilbur, WJ. and Lipman, D.J. [1983]. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl.Acad. sci*. USA, 80,726-30.
  - Lipman, D.J. and Pearson, W R. [1985]. Rapid and sensitive protein similarity searches. *Science*, 227,1435-41.
  - Altschul, S. F., Gish, W, Miller, W, Myers, E. W, and Lipman, D.J. [1990]. Basic local alignment search tool. *J. Mol. Biol*., 215,403-10.
  - Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W, and Liprnan, D.J. [1997]. Gapped BLAST and PSI-BLAST:a new generacion of protein database search programs. *Nucleic Acids Res*., 25, 3389--402.
  - FASTA: http://www.ebi.ac.uk/fasta33/ , http://vega.igh.cnrs.fr/bin/fasta-guess.cgi
- Statistical Significance
  - Karlin, S. and Altschul, S. F. [1990]. Methods for assessing the statiscical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. sci.* USA, 87 . 2264-8.
  - Pearson, W R. [1995]. Comparison of methods for searching protein sequece databases. *Protein sci*.,4, 1145--60.

Literature

- Simulated Annealing
  - Ishikawa, M. et al [1993]. “Multiple sequence alignment by parallel simulated annealing. Compt. *Appl. Biosci*. 9, 267-73.
  - Bertsimas, D. and J. Tsitsiklis [1993]. Simulated Annealing. *Statis. Sci*. 8, 10-15.
  - Kirkpatrick, S. C.D. Gelatt, and M.O. Vecchi [1983]. Optimization by simulated annealing. *Science*. 220, 671-680.
- Genetic Algorithms
  - Goldberg, D.E. [1989]. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.
  - Holland, J.H. [1975]. *Adaptation in Natural and Artificial Systems*. University of Michigan Press.
  - Holland, J.H. [1995]. *Hidden Order: How Adaptation Builds Complexity*. Addison-Wesley.
  - Mitchell, Melanie [1996]. *An Introduction to Genetic Algorithms*. MIT Press.