rocha@lanl.gov

Introduction to Bioinformatics

A Complex Systems Approach

Luis M. Rocha
Complex Systems Modeling
CCS3 - Modeling, Algorithms, and Informatics
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545
rocha@lanl.gov or rocha@santafe.edu

rocha@lanl.gov

Bioinformatics: A Complex Systems Approach

Course Layout

- Monday: *Overview and Background*
  - Luis Rocha
- Tuesday: *Gene Expression Arrays* – Biology and Databases
  - Tom Brettin
- Wednesday: *Data Mining and Machine Learning*
  - Luis Rocha and Deborah Stungis Rocha
- Thursday: *Gene Network Inference*
  - Patrik D'haeseleer
- Friday: *Database Technology, Information Retrieval and Distributed Knowledge Systems*
  - Luis Rocha

rocha@lanl.gov

Bioinformatics: A Complex Systems Approach

Overview and Background

- A Synthetic Approach to Biology
  - Information Processes in Biology
    - Biosemiotics
  - Genome, DNA, RNA, Protein, and Proteome
    - Information and Semiotics of the Genetic System
  - Complexity of Real Information Proceses
    - RNA Editing and Post-Transcription changes
  - Reductionism, Synthesis and Grand Challenges
  - Technology of Post-genome informatics
    - Sequence Analysis: dynamic programming, simulated anealing, genetic algorithms
  - Artificial Life

Luis Rocha
2001

rocha@lanl.gov

Information Processes in Biology

Distinguishes Life from Non-Life

Different Information Processing Systems (memory)

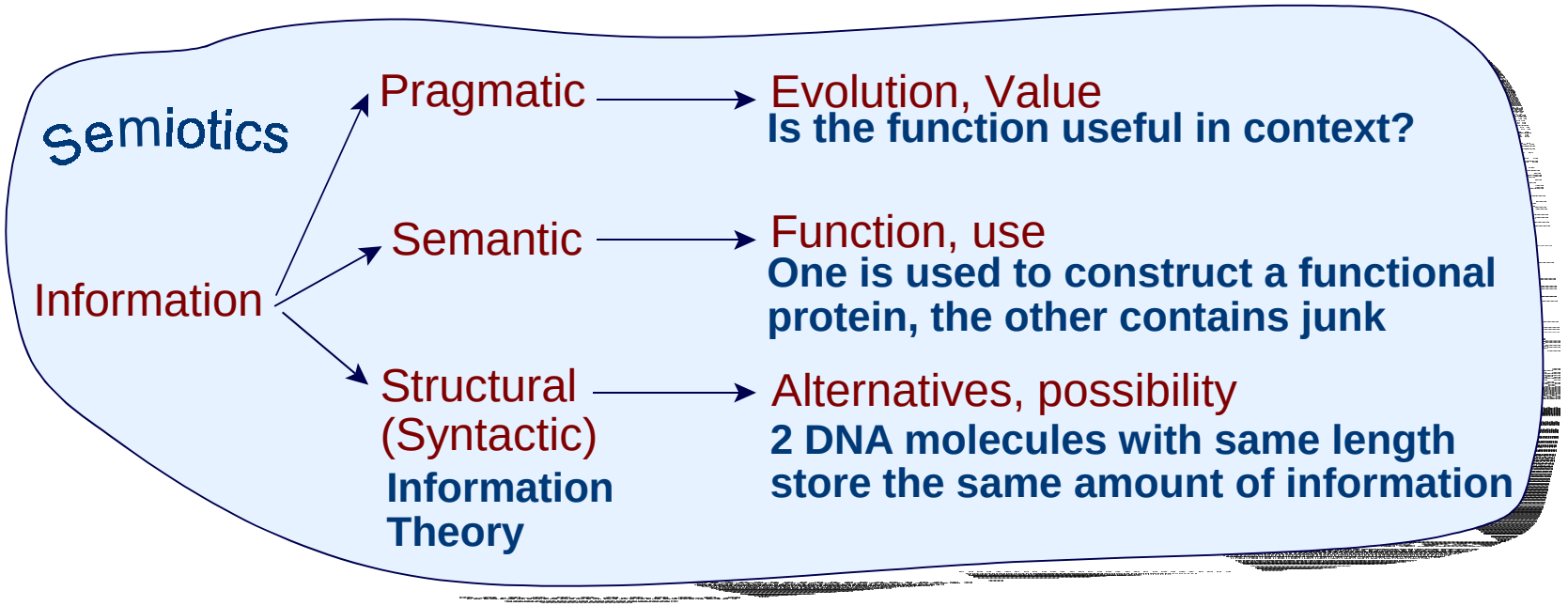
- Genetic System
  - Construction (expression, development, and maintenance) of cells ontogenetically: horizontal transmission
  - Heredity (reproduction) of cells and phenotypes: vertical transmission
- Immune System
  - Internal response based on accumulated experience (information)
- Nervous and Neurological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

Luis Rocha
2001

rocha@lanl.gov

What is Information?

Choice, alternative, memory, semiosis....



For Discrete Memory Structures !!

What does information mean in continuous domains?

Luis Rocha
2001

rocha@lanl.gov

Biology and Biosemiotics

The Study of the Semiosis of Life

Biology is the science of life that aims at understanding the *structural*, *functional*, and *evolutionary* aspects of living organisms

Biosemiotics is the study of informational aspects of biology in their *syntactic*, *semantic*, and *pragmatic* dimensions.

Genomics research has focused mostly on the syntactic (structural) dimension. Bioinformatics is an important tool for a more complete Biosemiotics

Luis Rocha
2001

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

Los Alamos National Laboratory

rocha@lanl.gov

Genomics and Proteomics

Information and Expression Units

- Mendelian Gene
  - Hereditary unit responsible for a particular characteristic or trait
- Molecular Biology Gene
  - Unit of (structural and functional) *information expression* (via Transcription and Translation)
- Genome
  - Set of genes in the chromosome of a species
  - Unit of (structural) *information transmission* (via DNA replication)
- Genotype
  - Instance of the genome for an individual
- Phenotype
  - Expressed and developed genotype
- Proteome
  - (Dynamic) Set of proteins that are encoded and expressed by a genome

rocha@lanl.gov

Nucleic Acids as Information Stores

Nucleotides (bases) as linguistic symbols

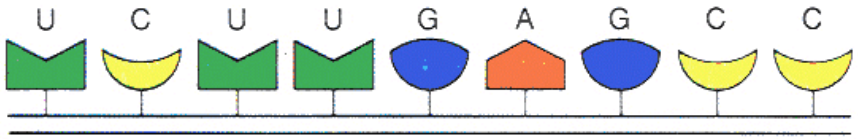
Purine (R) → Adenine (A), Guanine (G)

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

Nucleotides

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

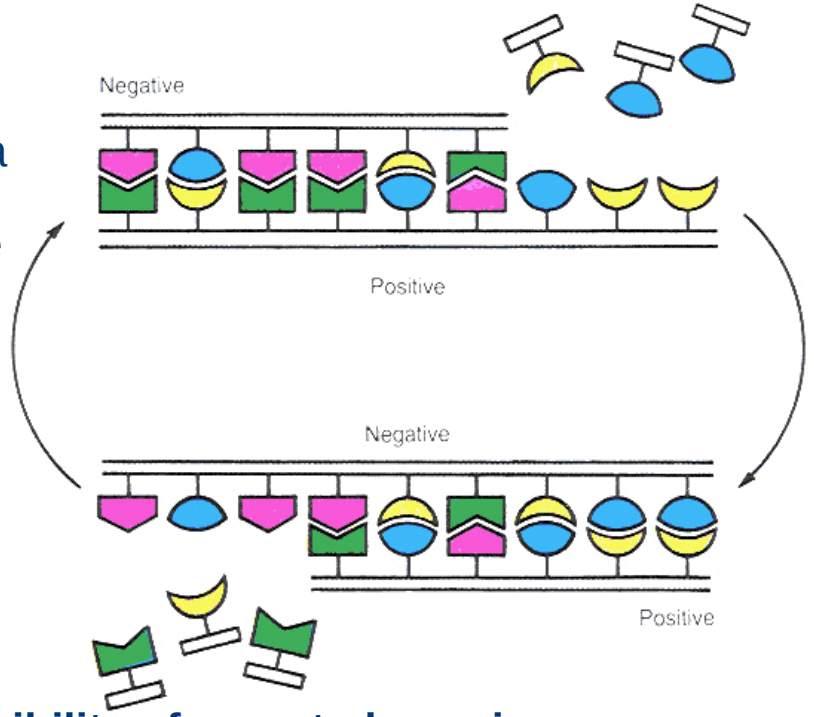
Form sequences that can store information



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

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

A-T (U) G-C



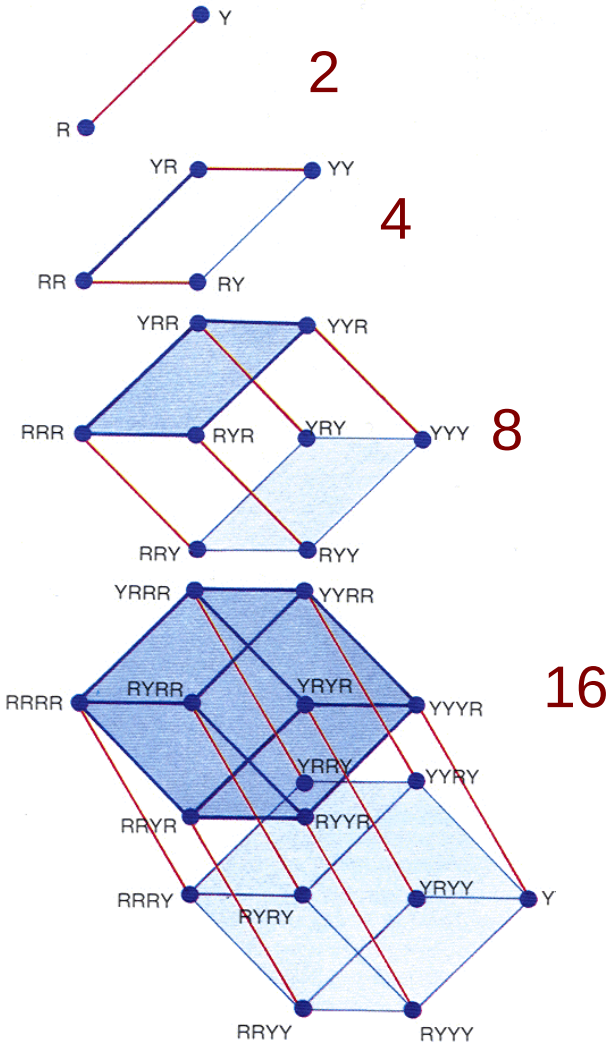
Requirements for structural information

Possibility of repeated copying

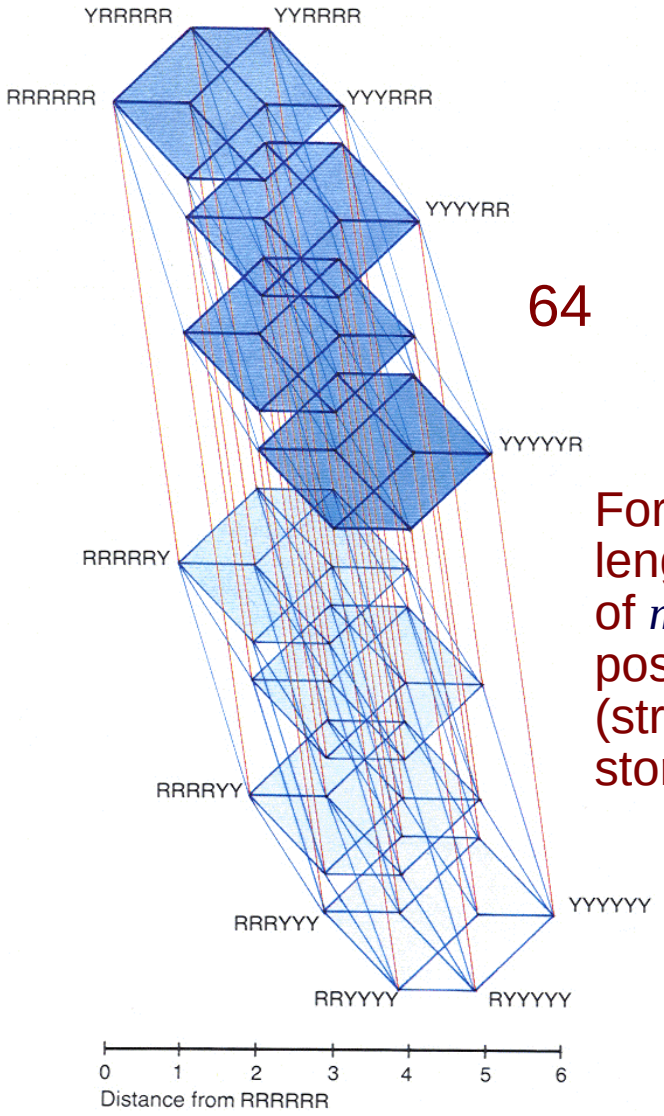
Luis Rocha
2001

rocha@lanl.gov

Information and Sequence Space



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



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

Luis Rocha
2001

Los Alamos National Laboratory

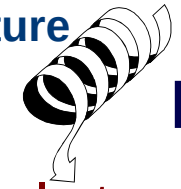
rocha@lanl.gov

Proteins: Functional Products

Sequences of Amino acids via peptide bonds

Polypeptide chains of aminoacids

Primary Structure



Folding

3-dimensional structure

Secondary and tertiary bonds

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

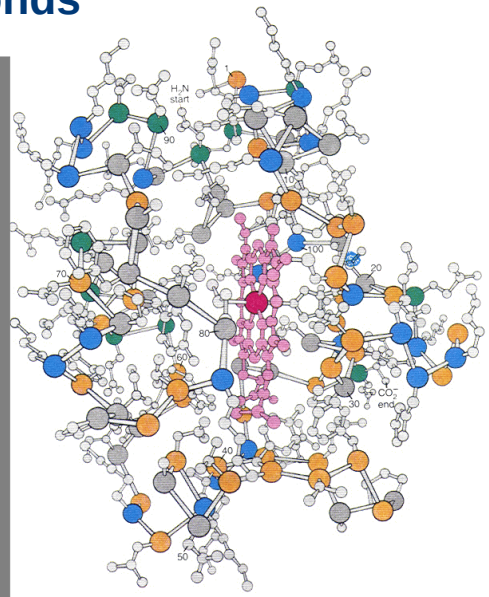


Table 1.4. Amino acid codes

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

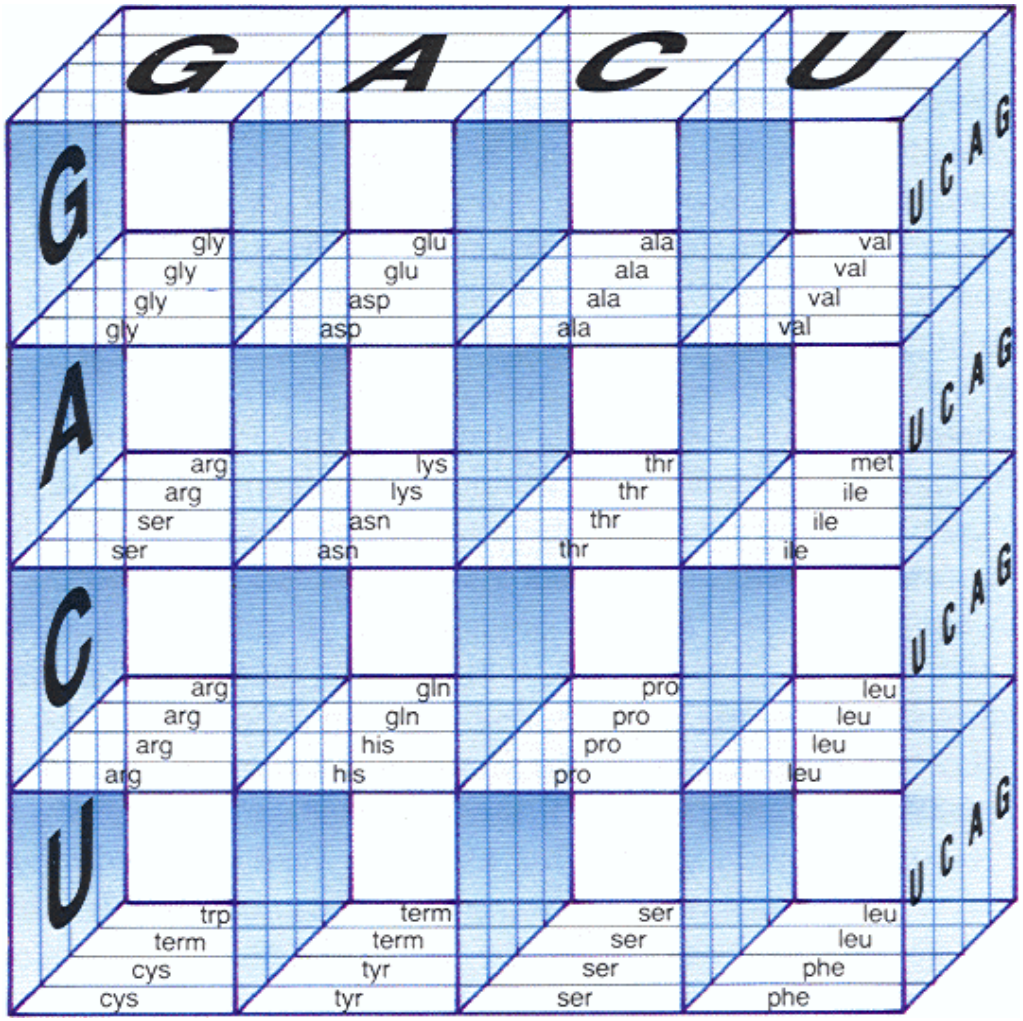
Luis Rocha
2001

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

Los Alamos National Laboratory

rocha@lanl.gov

The Genetic Code



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

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

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

Luis Rocha
2001

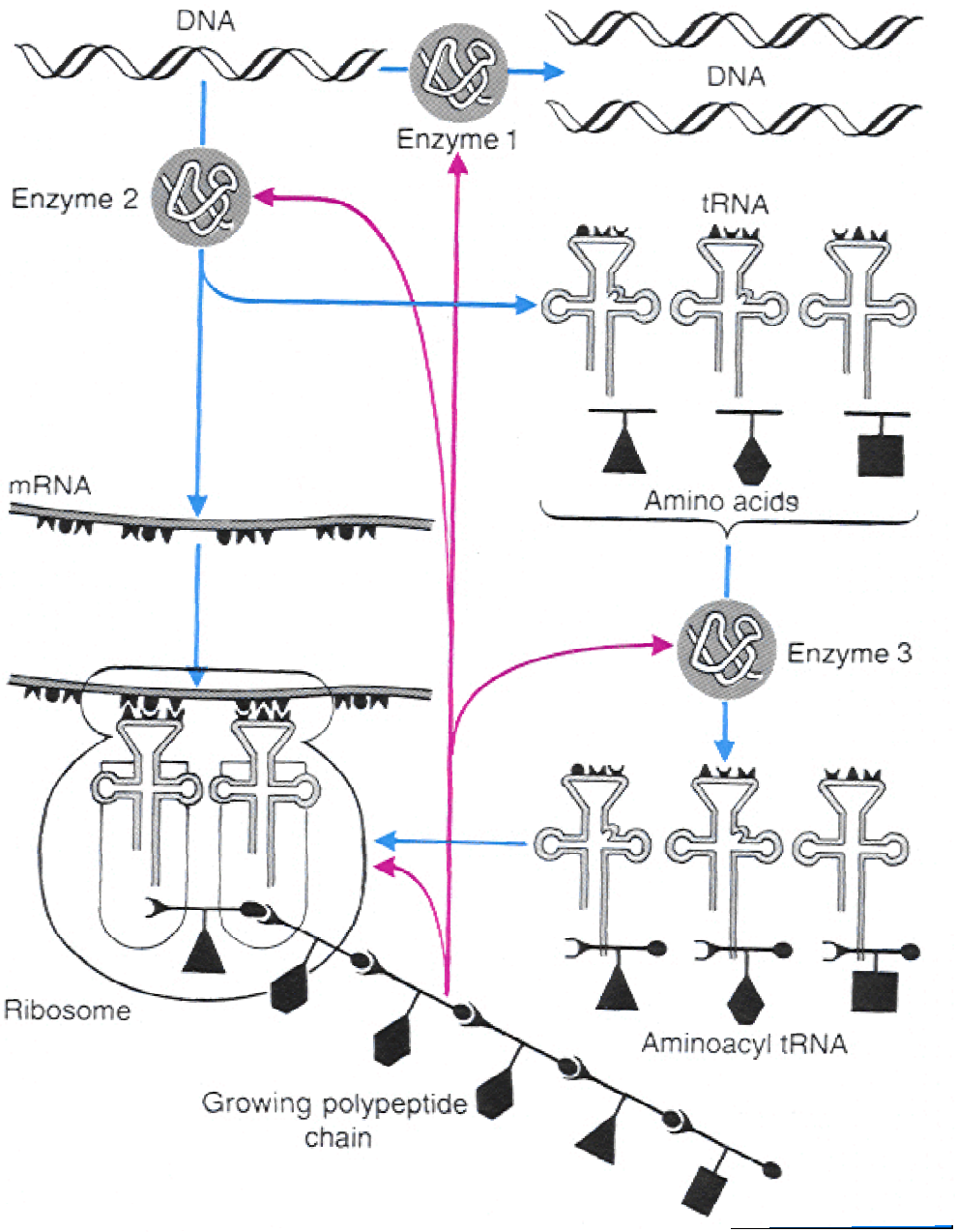
Los Alamos National Laboratory

rocha@lanl.gov

The genetic code at work

Structural and Functional Information

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



Luis Rocha
2001

rocha@lanl.gov

Variations of Genetic Codes

Table 1.6. Variation of genetic codes

	T1	T2	T3	T4	T5	T6	T9	T10	T12	T13	T14	T1
CUU	Leu	–	Thr	–	–	–	–	–	–	–	–	–
CUC	Leu	–	Thr	–	–	–	–	–	–	–	–	–
CUA	Leu	–	Thr	–	–	–	–	–	–	–	–	–
CUG	Leu	–	Thr	–	–	–	–	–	Ser	–	–	–
AUU	Ile	–	–	–	–	–	–	–	–	–	–	–
AUC	Ile	–	–	–	–	–	–	–	–	–	–	–
AUA	Ile	Met	Met	–	Met	–	–	–	–	Met	–	–
AUG	Met	–	–	–	–	–	–	–	–	–	–	–
UAU	Tyr	–	–	–	–	–	–	–	–	–	–	–
UAC	Tyr	–	–	–	–	–	–	–	–	–	–	–
UAA	*Stop*	–	–	–	–	Gln	–	–	–	–	Tyr	–
UAG	*Stop*	–	–	–	–	Gln	–	–	–	–	–	Gl
AAU	Asn	–	–	–	–	–	–	–	–	–	–	–
AAC	Asn	–	–	–	–	–	–	–	–	–	–	–
AAA	Lys	–	–	–	–	–	Asn	–	–	–	Asn	–
AAG	Lys	–	–	–	–	–	–	–	–	–	–	–
UGU	Cys	–	–	–	–	–	–	–	–	–	–	–
UGC	Cys	–	–	–	–	–	–	–	–	–	–	–
UGA	*Stop*	Trp	Trp	Trp	Trp	–	Trp	Cys	–	Trp	Trp	–
UGG	Trp	–	–	–	–	–	–	–	–	–	–	–
AGU	Ser	–	–	–	–	–	–	–	–	–	–	–
AGC	Ser	–	–	–	–	–	–	–	–	–	–	–
AGA	Arg	*Stop*	–	–	Ser	–	Ser	–	–	Gly	Ser	–
AGG	Arg	*Stop*	–	–	Ser	–	Ser	–	–	Gly	Ser	–

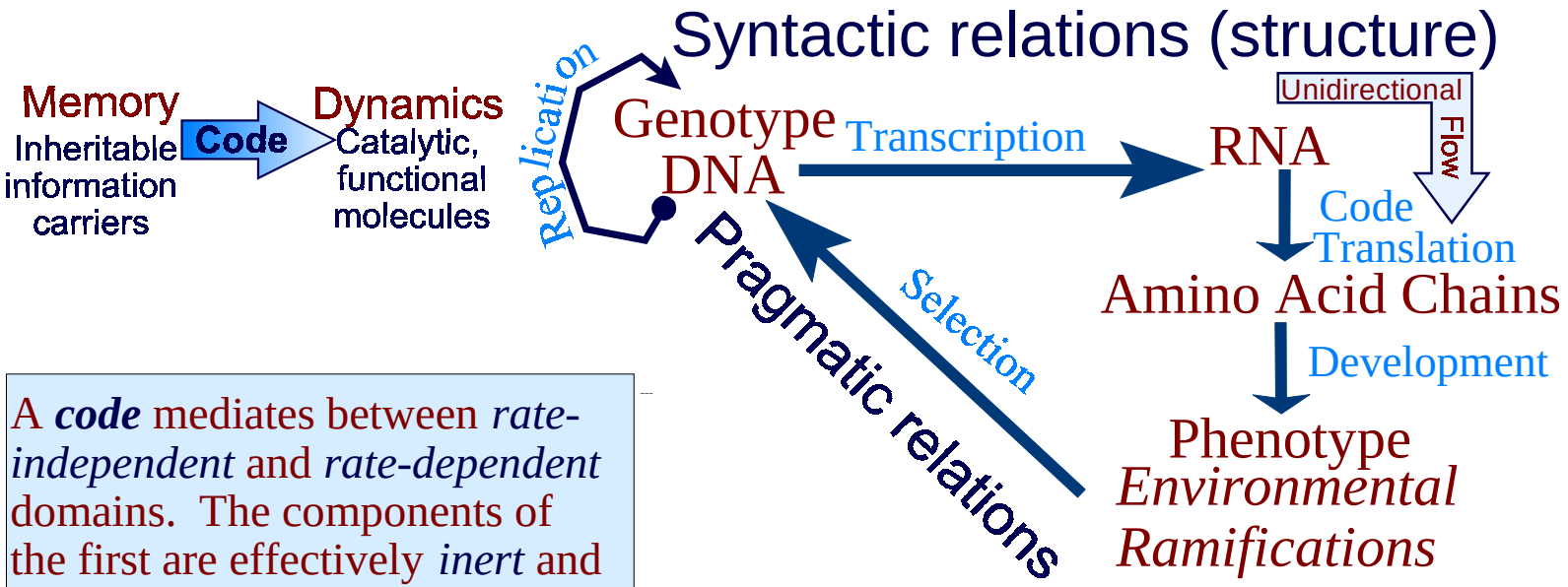
T1, Standard code; T2, vertebrate mitochondrial code; T3, yeast mitochondrial code; T4, mould, proto zoan, and coelenterate mitochondrial code and mycoplasma and spiroplasma code; T5, invertebrate m tochondrial code; T6, ciliate, dasycladacean and hexamita nuclear code; T9, echinoderm mitochondri code; T10, euplotid nuclear code; T12, alternative yeast nuclear code; T13, ascidian mitochondrial cod T14, flatworm mitochondrial code; T15, blepharisma nuclear code.

Luis Rocha
2001

rocha@lanl.gov

The Semiotics of the Genetic System

The Central Dogma of Information Transmission



A ***code*** mediates between *rate-independent* and *rate-dependent* domains. The components of the first are effectively *inert* and used as *memory stores* (structural information, descriptions, etc.) While the components of the second are *dynamic* (functional) players used to directly *act* in the world (e.g. enzymes). sa2.html.

"Genetic information is not expressed by the dynamics of nucleotide sequences (RNA or DNA molecules), but is instead mediated through an arbitrary coding relation that translates nucleotide sequences into amino-acid sequences whose dynamic characteristics ultimately express genetic information in an environment." sa2.html.

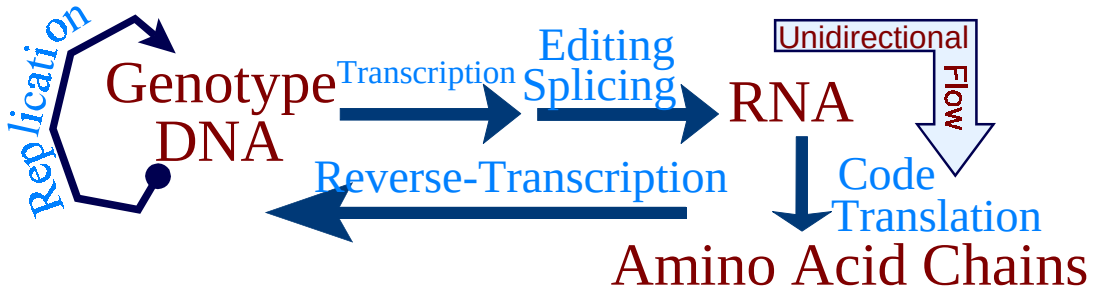
Luis Rocha
2001

rocha@lanl.gov

Real Information Processes in The Genetic System

A More Complex Picture of Syntactic Operations

- Reverse-Transcription
  - Retroviruses store genetic information in genomic RNA rather than DNA, so to reproduce they require reverse transcription into DNA before replication
- Complex Transcription of DNA to RNA before translation
  - Intron Removal and Exon Splicing (deletion operation)
  - RNA Editing (insertion and replacement operation)
- Do not challenge the Central Dogma but increase the complexity of information processing

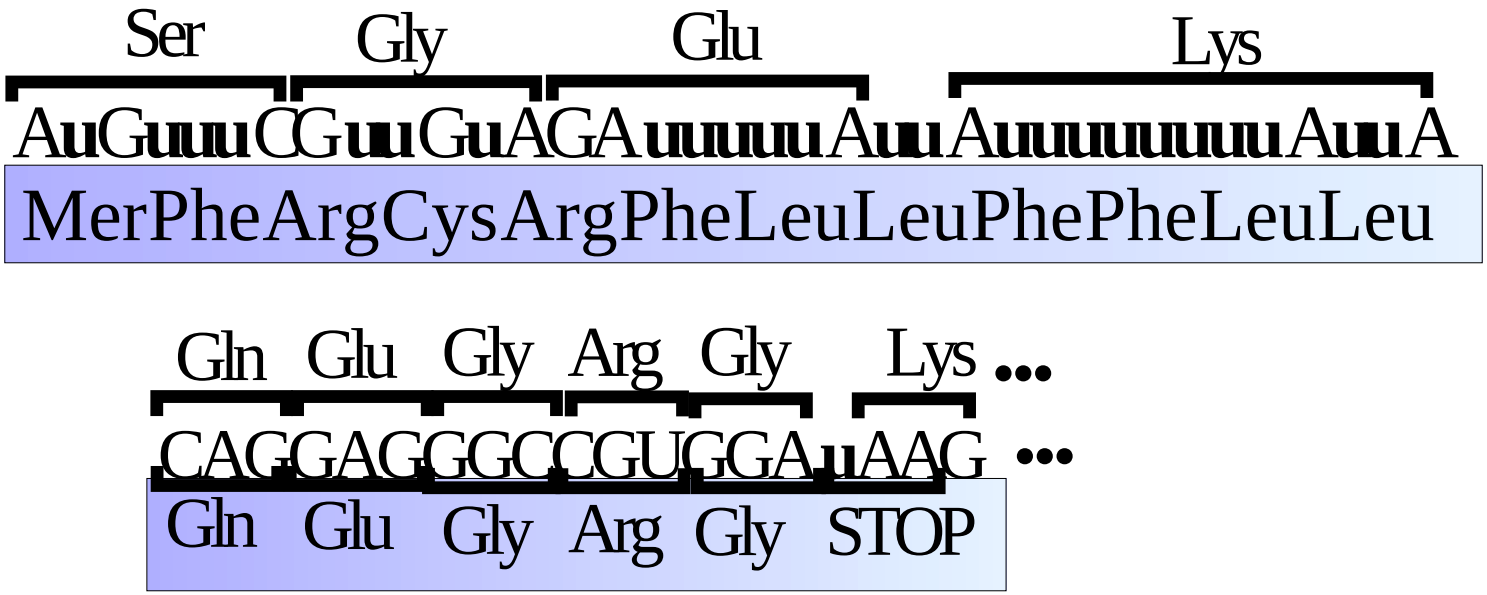


Luis Rocha
2001

rocha@lanl.gov

RNA Editing

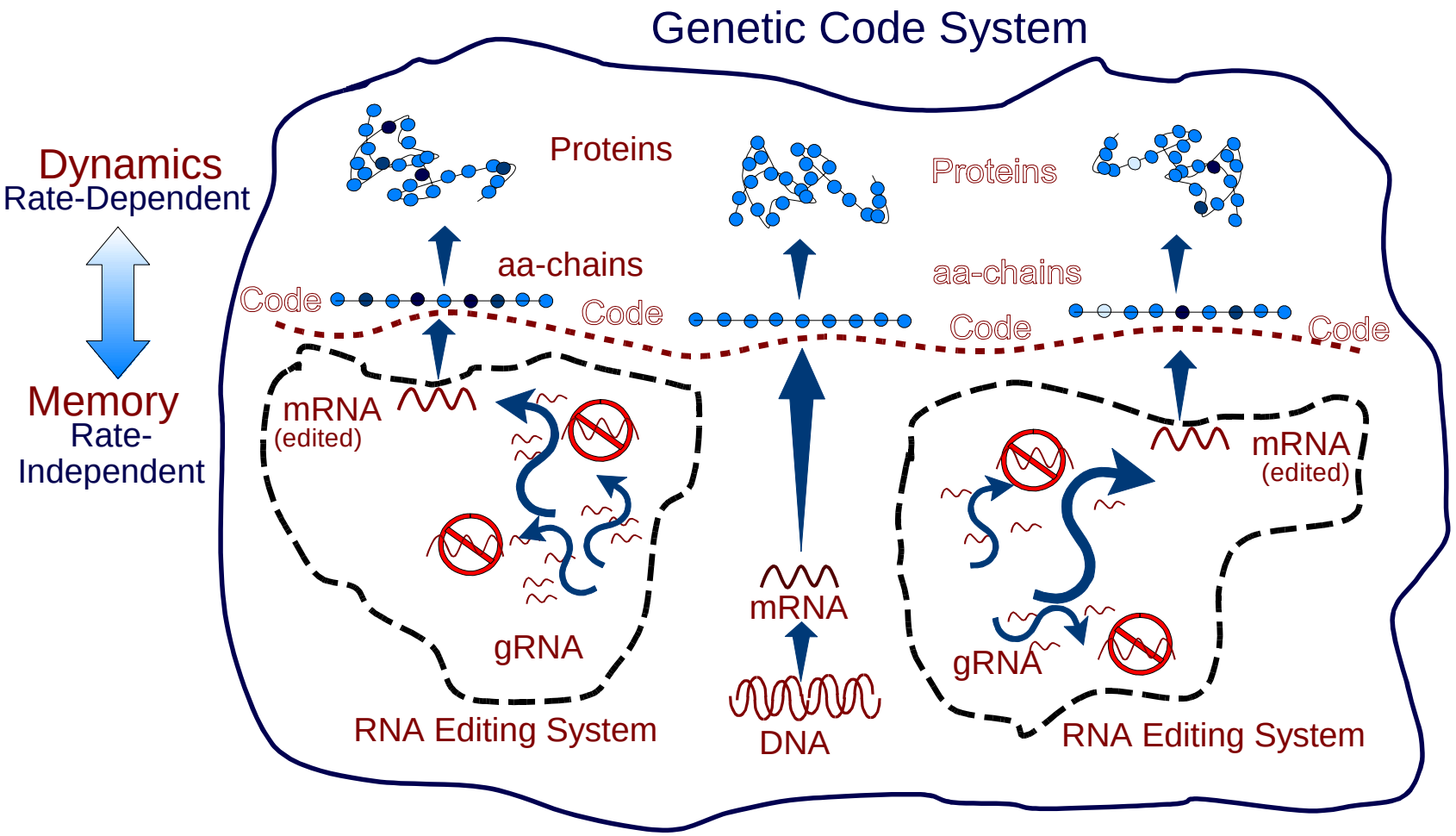
Example



Luis Rocha
2001

rocha@lanl.gov

RNA Editing acts on Memory (syntax)



Luis Rocha
2001

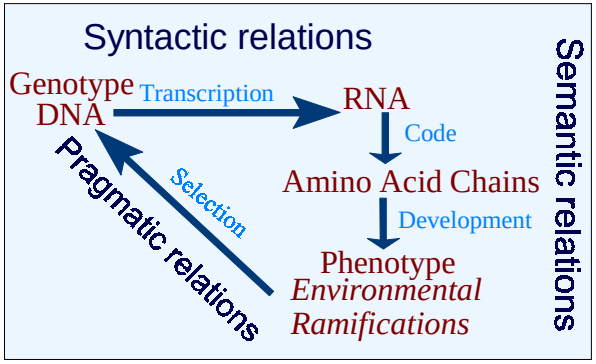
http://www.c3.lanl.gov/~rocha/bioinformatics

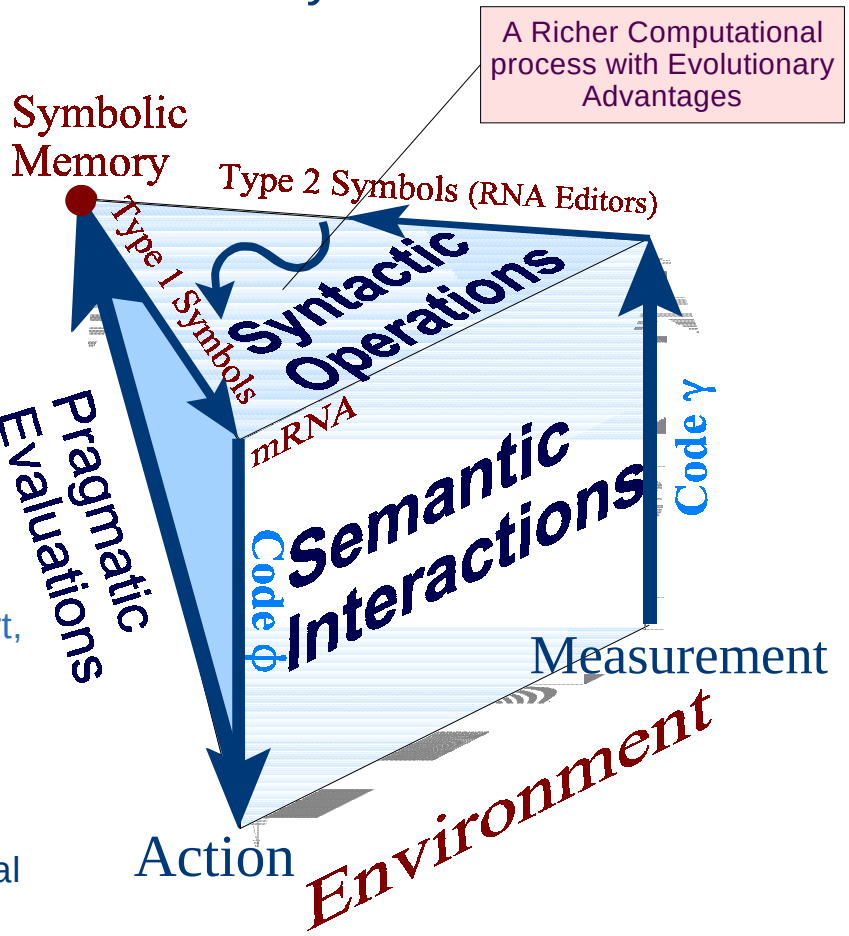
Los Alamos
National Laboratory

rocha@lanl.gov

RNA Editing as a Measurement Code

Expanding the Semiotics of the Genetic System





- Suggested Process of Control of Development Processes from environmental cues
  - In Trypanosomes: Benne, 1993; Stuart, 1993. Evolution of parasites: Simpson and Maslov, 1994. Neural receptor channels in rats: Lomeli et al, 1994
  - Metal ion switch (with ligase and cleavage activities) in a single RNA molecule used to modulate biochemical activity from environmental cues. Landweber and Pokrovskaya, 1999

ises.html and e95_abs.html

Luis Rocha
2001

Los Alamos National Laboratory

rocha@lanl.gov

Post-Translation

Complex Dynamic Interactions

- Rate-dependent expression products: non-linear, environmentally dependent, development
  - Catalysis, metabolism, cell regulation
- Protein folding though thermodynamically reversible in-vitro, is expected to depend on complex cellular processes
  - E.g. chaperone molecules
- Prediction of protein folded structure and function from sequence is hard
- Biological function is not known for roughly half of the genes in every genome that has been sequenced
  - Lack of technology
  - The genome itself does not contain all information about expression and development (Contextual Information Processing)

rocha@lanl.gov

Bioinformatics

A Synthetic Multi-Disciplinary Approach to 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)
  - Deals with syntactic information alone
- Post-genome informatics aims at the synthesis of biological knowledge (full semiosis) 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 (syntactic) information about building blocks but it is premature to assume that it also contains the information on how the building blocks relate, develop, and evolve (semantic and pragmatic information)
  - Interdisciplinary: biology, computer science, mathematics, and physics

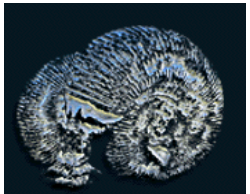
Luis Rocha
2001

rocha@lanl.gov

Bioinformatics as Biosemiotics

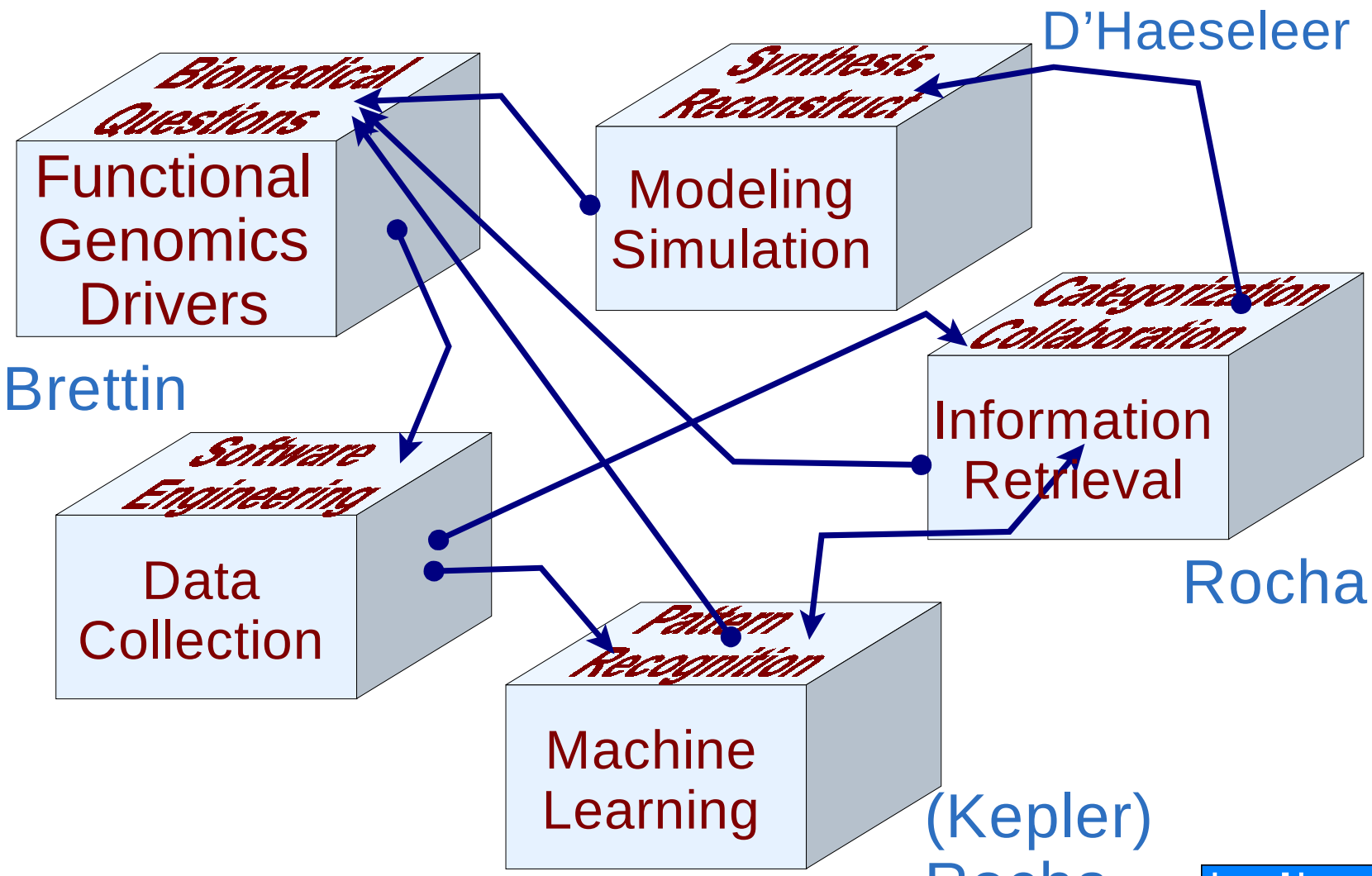
A Synthetic Multi-Disciplinary Approach to Biology

- Not just support technology but involvement in the systematic design and analysis of experiments
  - *Functional genomics*: analysis of gene expression patterns at the mRNA (syntactic information) and protein (semantic information) levels, as well as analysis of polymorphism, mutation patterns and evolutionary considerations (pragmatic information).
    - Using and developing computer science and mathematics
  - 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 contextuall and dynamically accessed and even modified by the complete network of reactions in the cell (e.g. editing).
  - Uses additional knowledge for comparative analysis: Comparative Biology
    - e.g. reference to known 3D structures for protein folding prediction, or reference databases across species



rocha@lanl.gov

Components of Bioinformatics



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

Luis Rocha
2001

Los Alamos National Laboratory

rocha@lanl.gov

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.

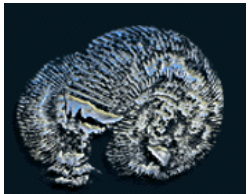
- Similarity (Homology) Search
  - Pairwise and multiple sequence alignment, database search, phylogenetic tree reconstruction, Protein 3D structure alignment
    - Dynamic programming, Simulated annealing, Genetic Algorithms, Neural Networks
- Structure/function prediction
  - Ab initio: RNA secondary 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

rocha@lanl.gov

Similarity Search vs. Motif Search

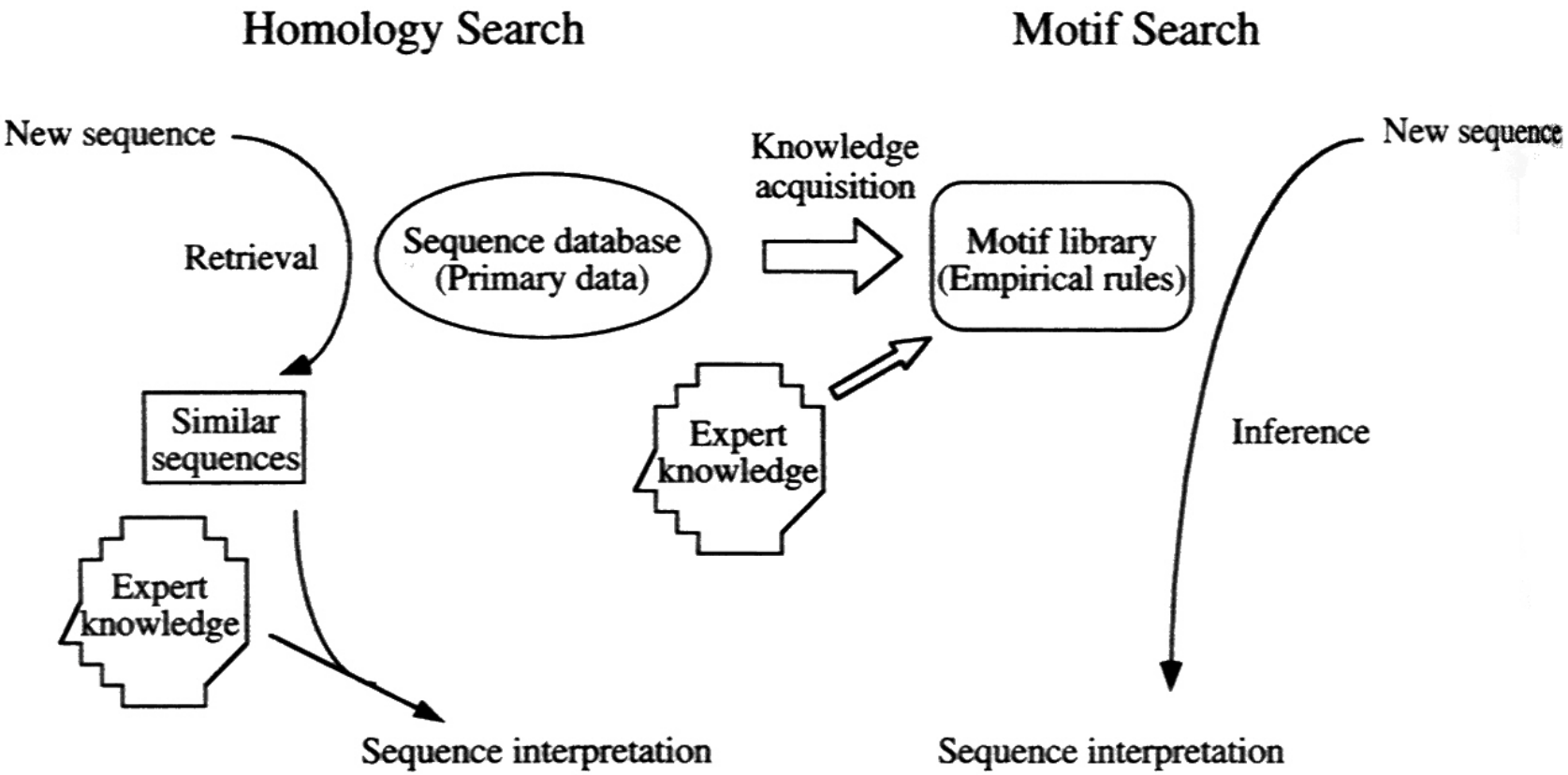
Data-driven vs. Knowledge-based Functional Interpretation

- Similarity (Homology) Search
  - A query sequence is compared with others in 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 (interaction sites with other molecules)
    - A Motif library is like a dictionary
    - Unfortunately there are no comprehensive motif libarries for all types of functional properties



rocha@lanl.gov

Similarity Search vs. Motif Search



Luis Rocha
2001

rocha@lanl.gov

Sequence Similarity Search

Sequence Alignment

- 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.
  - 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

rocha@lanl.gov

Mutation Matrix as Substitution Table

Table 2.6. The PAM-250 mutation matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
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

Luis Rocha
2001

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**

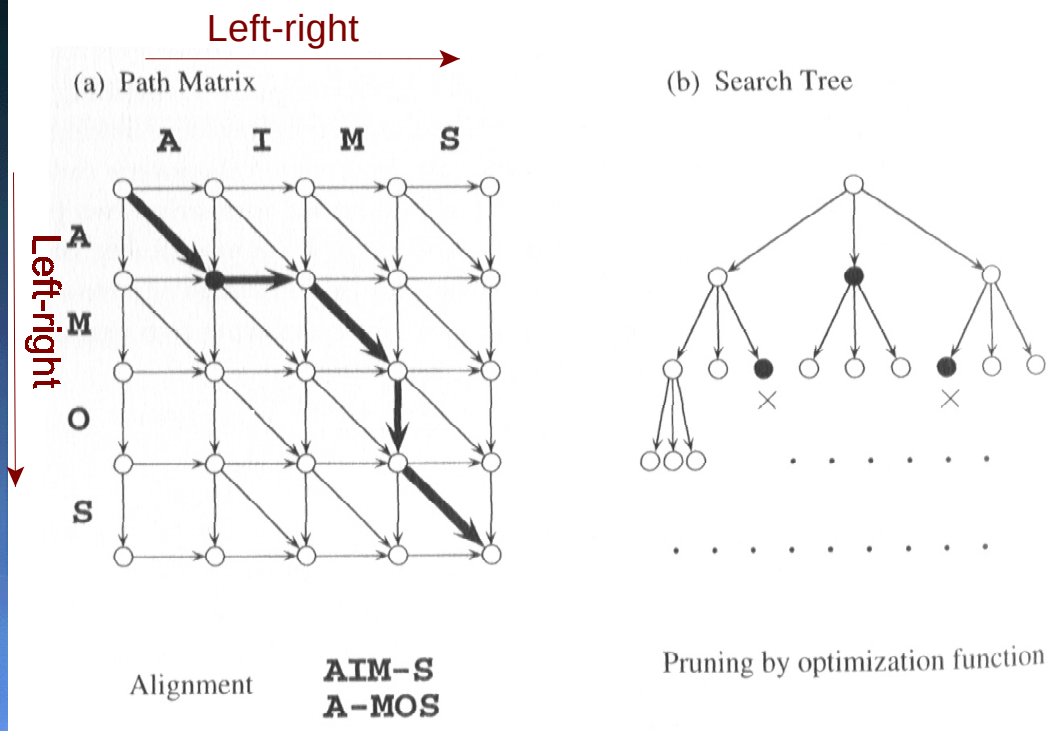
AMOS → **A-MOS**

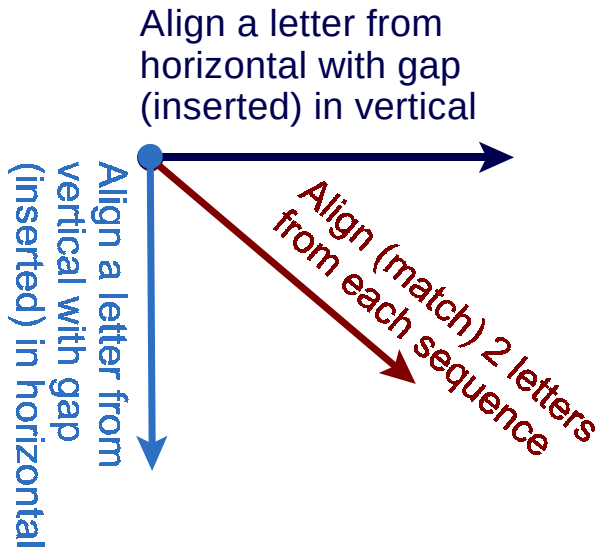
Dynamic programming is a very general optimization technique for problems that can recursively be divided into two similar problems of smaller size, such that the solution to the larger problem can be obtained by piecing together the solutions to the two subproblems. Example: shortest path between 2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

rocha@lanl.gov

Dynamic Programming

Path Matrix





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.

Luis Rocha
2001

rocha@lanl.gov

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)

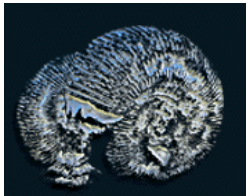
$W_{s(i),t(j)}$ Weigth for aligning (Substituting) element i from sequence s with element j of sequence t

d Weigth for a single element gap

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

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

Luis Rocha
2001

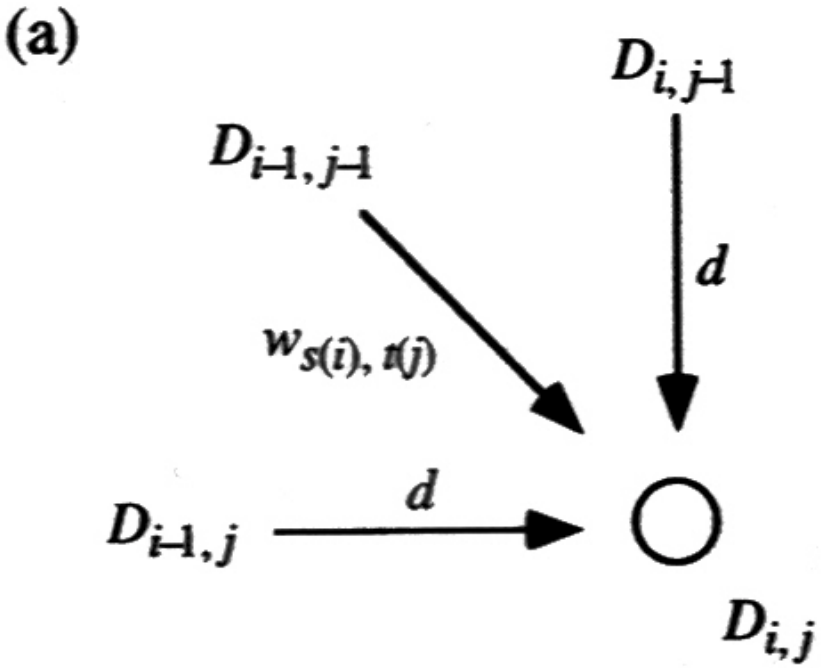


rocha@lanl.gov

Global Alignment

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

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



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.

Luis Rocha
2001

rocha@lanl.gov

Local Alignment

Alignment of subsequences

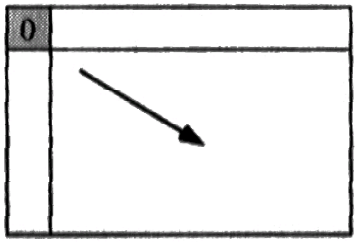
$$D_{i,j} = \max(D_{i-1,j-1} + W_{s(i),t(j)}, D_{i-1,j} + \mathrm{d}, D_{i,j-1} + \mathrm{d})$$

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

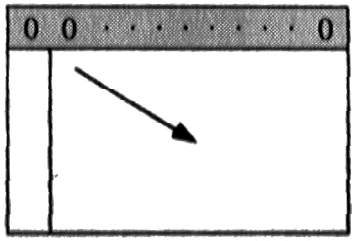
$D_{0,j} = 0\ (j=1...m)$

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

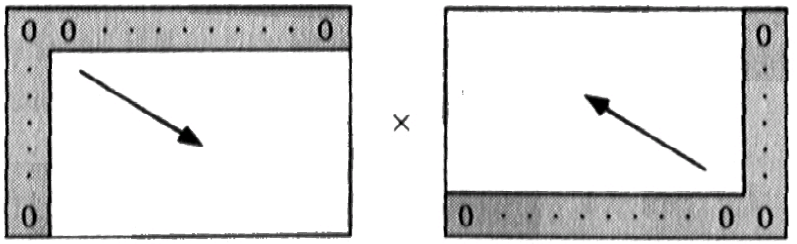
(a) Global vs. Global



(b) Local vs. Global



(c) Local vs. Local



Luis Rocha
2001

rocha@lanl.gov

Local Alignment

Smith-Waterman Local Optimality Algorithm

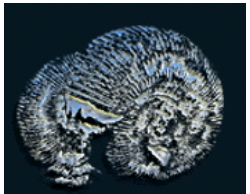
$$D_{i,j} = \max(D_{i-1,j-1} + W_{s(i),t(j)}, D_{i-1,j} + \mathrm{d}, D_{i,j-1} + \mathrm{d})$$

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

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

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

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

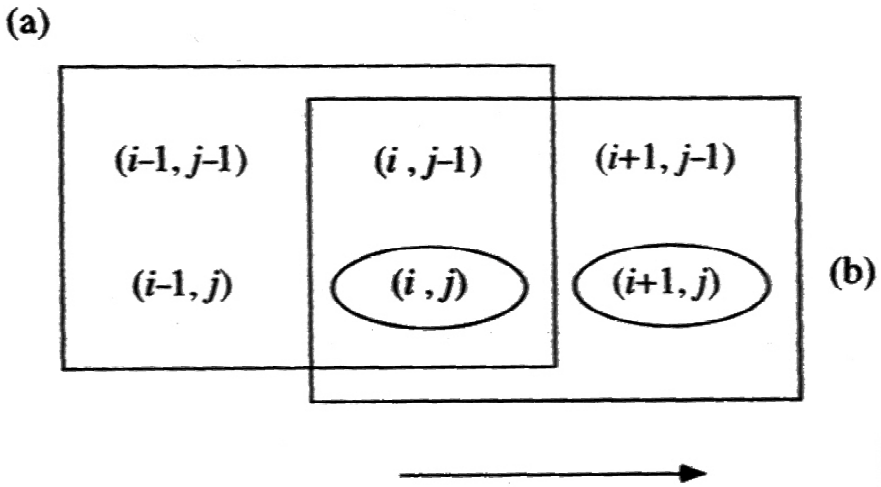


rocha@lanl.gov

Similarity Database Search

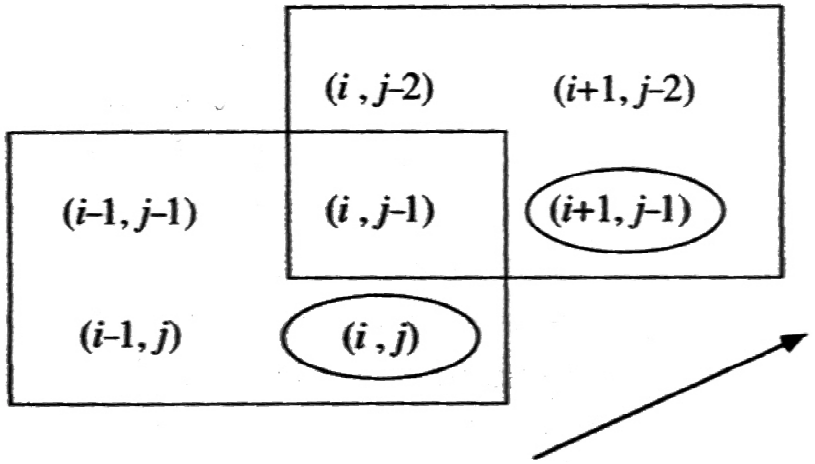
Parallelized Dynamic Programming

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



Sequential

Parallel



Luis Rocha
2001

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

Los Alamos National Laboratory

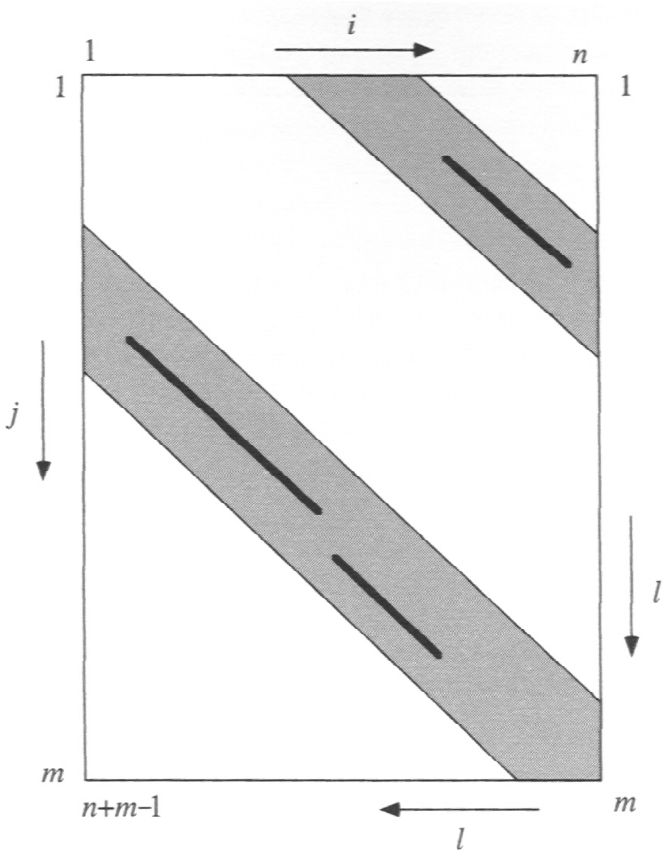
rocha@lanl.gov

FASTA Method

Dot Matrix Reduces DP Search Area

```
  AIMS
A *
M   *
O
S     *
```

Dot Matrix



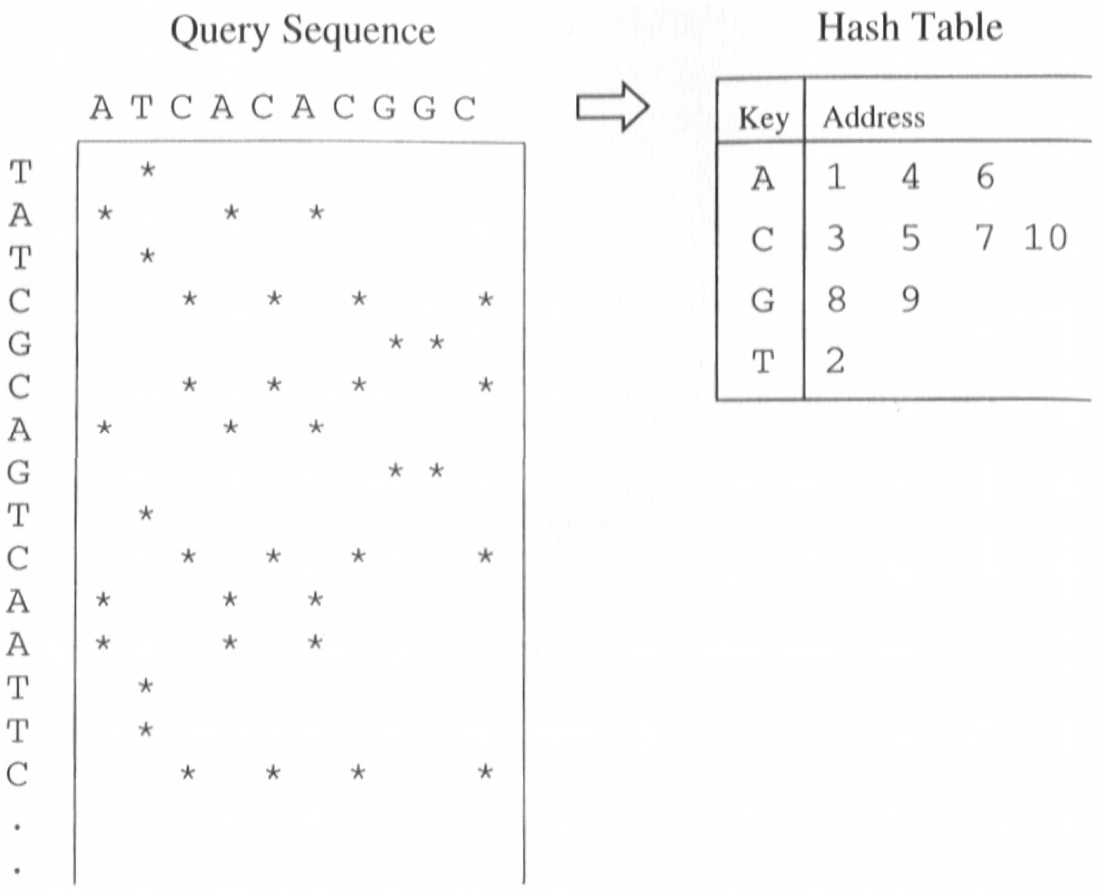
The dot matrix can be used to recognize local alignments which show as diagonal stretches or clusters of diagonal strectches. DP can be used only for the portions of the matrix around these clusters – a limited search area.

Luis Rocha
2001

rocha@lanl.gov

FASTA

Hashing the Dot Matrix



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 to the mean row size of the hash table (times dots are entered), which is on averahe 1/4 of the sequence.

Luis Rocha
2001

rocha@lanl.gov

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 between 2 sequences

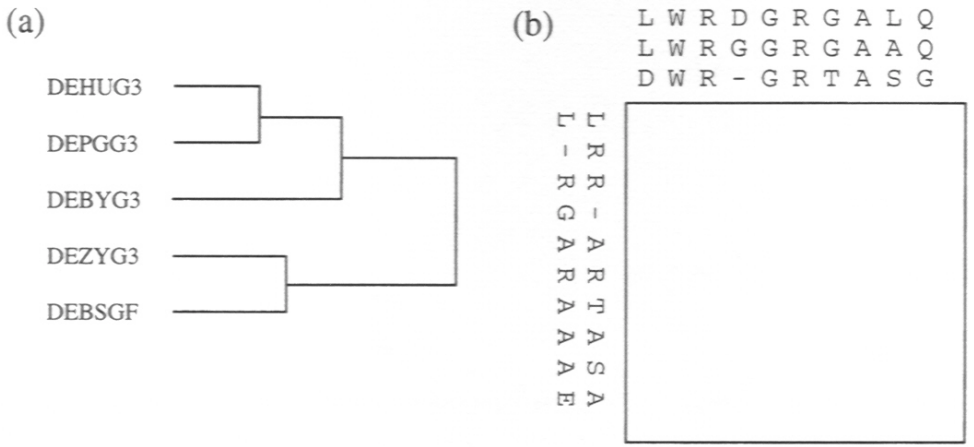
Each sequence is randomized k times (preserving the composition) and new optimal alignment is computed: s1, s2, ..., sk with mean μ and standard deviation σ. If the score distribution is normal, Z values of 4 and 5 correspond to threshold probabilities of 3×10^{-5} and 3×10^{-6}. However, the distribution 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.

rocha@lanl.gov

Multiple Alignment

Simultaneous Comparison of a Group of Sequences

- 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, followed by computation of the alignment between groups of sequences without changing the predetermined alignment within each group.

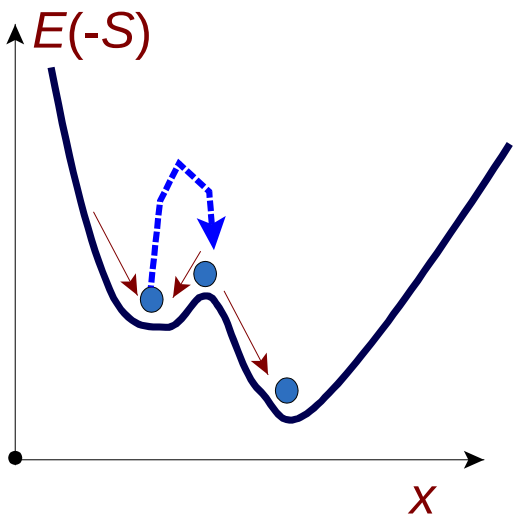


Luis Rocha
2001

rocha@lanl.gov

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

Luis Rocha
2001

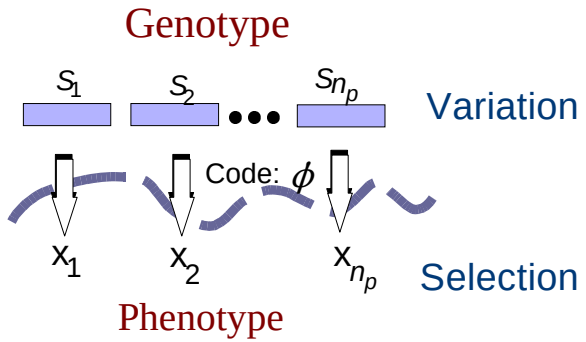
http://www.c3.lanl.gov/~rocha/bioinformatics

rocha@lanl.gov

Genetic Algorithms

For Multiple Sequence Alignment

Traditional Genetic Algorithm



- 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).

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.

rocha@lanl.gov

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

rocha@lanl.gov

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)
- Dynamic Programming and Sequence Alignment
  - 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.

Luis Rocha
2001

rocha@lanl.gov

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.
- 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.

rocha@lanl.gov

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.

rocha@lanl.gov

Literature

- Biosemiotics
  - Emmeche, Claus [1994]. *The Garden in the Machine: The Emerging Science of Artificial Life*. Princeton University Press.
  - Hoffmeyer, Jesper [2000]."Life and reference." *Biosystems*. In Press.
  - 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.
  - Rocha, Luis M. [1996]."Eigenbehavior and symbols." *Systems Research*. Vol. 13, No. 3, pp. 371-384.
  - Rocha, Luis M. [2000]."Syntactic Autonomy: or why there is no autonomy without symbols and how self-organizing systems might evolve them." In: *Closure: Emergent Organizations and Their Dynamics*.. J.L.R. Chandler and G. Van de Vijver (Eds.). *Annals of the New York Academy of Sciences*. Vol. 901, pp.207-223.
  - http://www.c3.lanl.gov/~rocha/pattee
  - Rocha, Luis M. [2001]. “Evolution with material symbol systems”. Biosystems.

rocha@lanl.gov

Bioinformatics Technology

Gene Expression Focus

- Biology Driver
- Gene Expression Databases
- Statistical and Machine Learning Analysis
- Network Analysis and Modeling
- Database Technology, Information Retrieval, and Recommendation

Luis Rocha
2001