

10.555

***Bioinformatics: Principles,
Methods and Applications***

MIT, Spring term, 2004

Lecture 1:

Introduction to the course.

Definitions, perspectives, general overview.

Rudiments of dynamic programming

Course goals-requirements

★ Goals

- Not code writing
- Not algorithm optimization
- Provide appreciation-familiarity-expertise with available computational tools of sequence analysis, pattern discovery, analysis of cell physiology
- Define important problems in bioinformatics-systems biology
- Suggest process for problem solving by illustrative case studies involving important biological-physiological problems

★ Requirements

- Advanced course
- None formally
- Applied probability- statistics most important background

More Information at:

★ MIT Class:

<http://web.mit.edu/cheme/gnswebpage/index.shtml>

Look for course 10.555

We have entered a period of rapid change

... For the times, they are a changin'

Bob Dylan

***How is phenotype controlled by the genes?
Nobody knows, least of all the machines.
Medicine will thrive if we can discover the means***

***To merge our knowledge and information
And find genes' intent and control by environment***

For the times, they are a changin'.

Drivers of change (1)

- **Genomics**: Sequencing and study of genomes
 - ✱ More than 100 (150) genome species sequenced: *Comparative* studies
 - ✱ All industrially important organisms will be sequenced within the next 1-2 years
 - ✱ Impact of Genomics:
 - ✱ ***Generation of data about the cellular phenotype: Link with genotype (expression)***
 - ✱ ***Need of integration in a meaningful framework***
 - ✱ ***Use of these data in discovery***

Drivers of change (2)

- **Heavy investment in the life sciences**
 - ✱ Excess of \$200B in the USA
 - ✱ Mostly health-oriented, however, equally applicable to other fields
 - ✱ **Applied molecular biology**
 - ✱ Construct optimal genetic background
 - ✱ Controls at the genetic level
 - ✱ Optimal cell-bioreactor combinations

Drivers of change (3)

- **Enormous opportunities**
 - ✱ Medicine
 - ✱ Chemicals (*cell factories*)
 - ✱ Materials (special properties, PHA's, PHB)
 - ✱ Pharmaceuticals (chirality)
 - ✱ Fuels (bio-diesel), ethanol, CO₂, methane-methanol
 - ✱ Environmental
 - ✱ Plants (resistance, oils, enrichment, polymers)

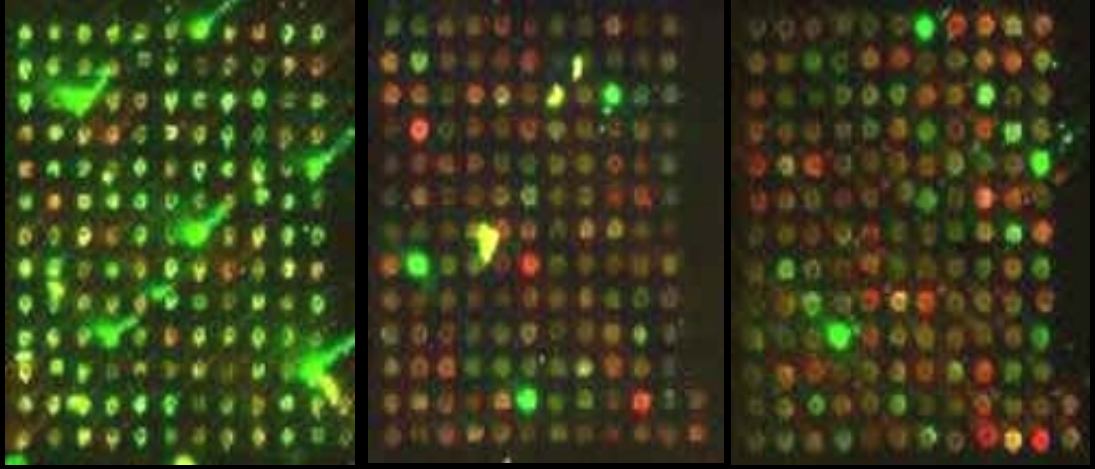
Drivers of Bioinformatics (1)

- **Genomics**

Drivers of Bioinformatics (2)

- Technologies probing the expression, proteomic and metabolic phenotype generating large volumes data
 - ✱ DNA microarrays (*gene chips*)
 - ✱ 2-D gel electrophoresis
 - ✱ Protein chips or HPLC methods combined with maldi/seldi tof spectrometry for protein characterization
 - ✱ Isotope label distributions

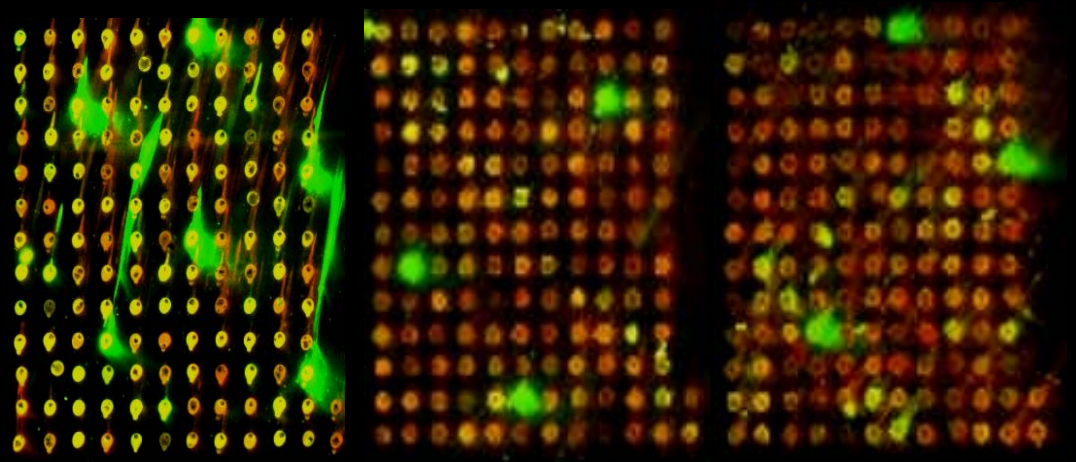
0.4% Pine-Sol
mid-exponential phase



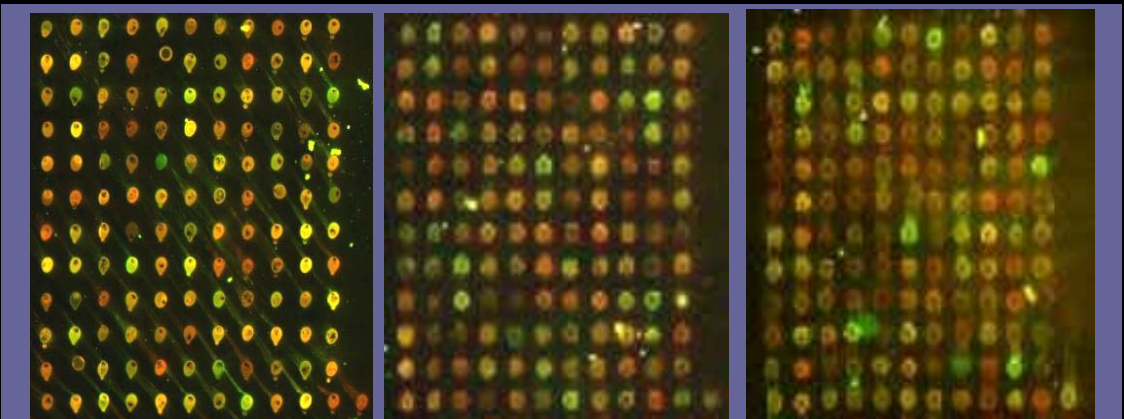
Same culture
+3 hours



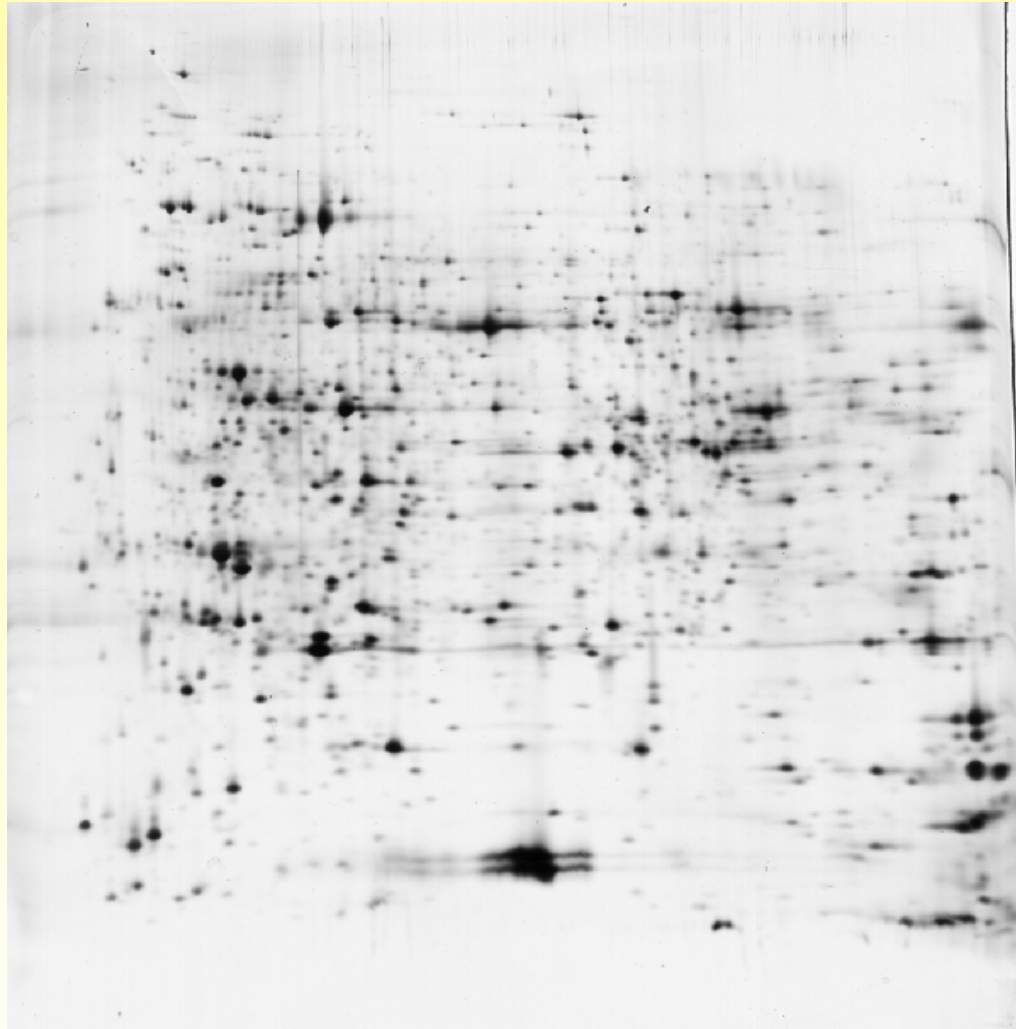
0.4% Pine-Sol
early stationary phase



Control
early stationary phase



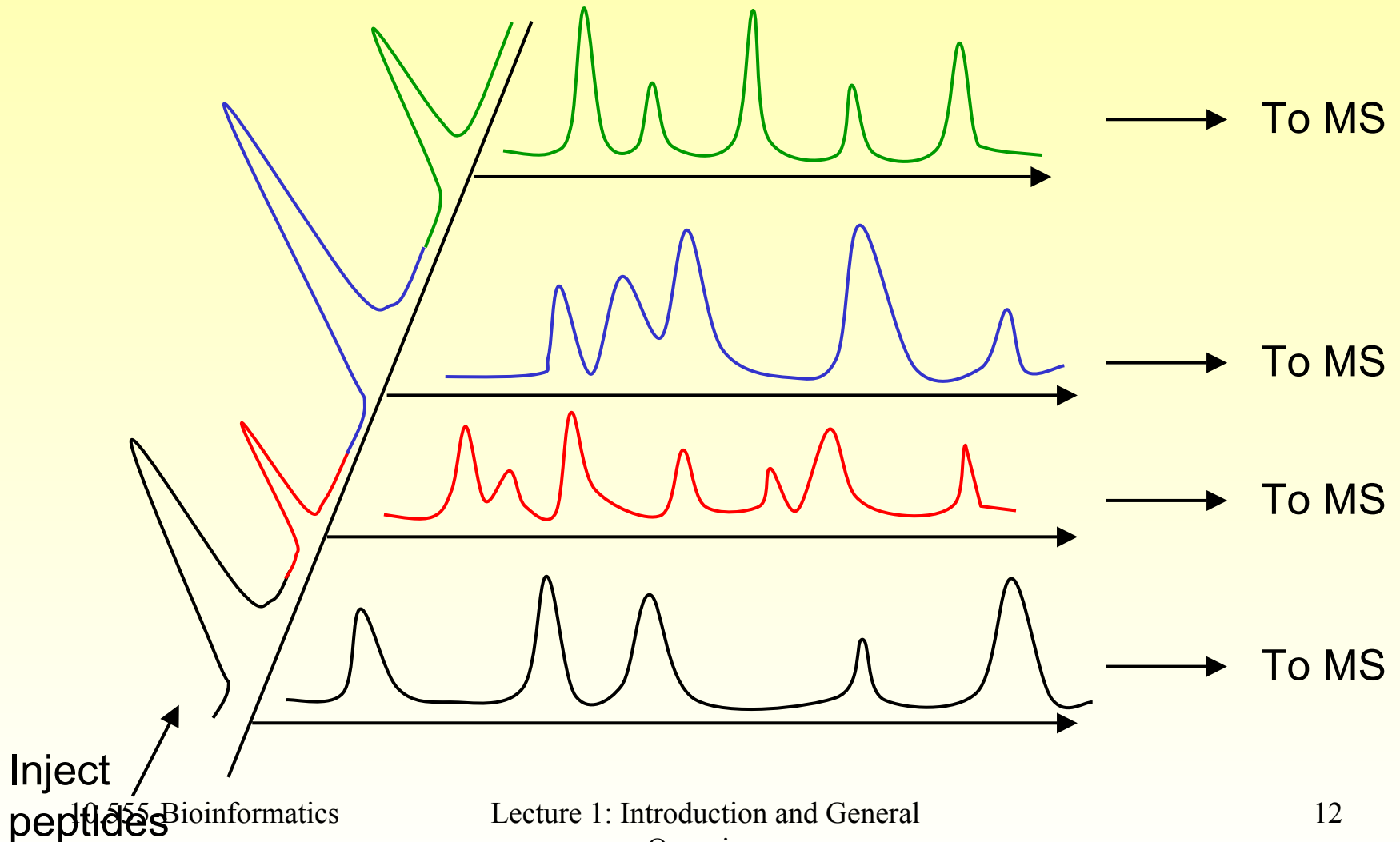
Example of 2-D gels



2-Dimensional Liquid Chromatography

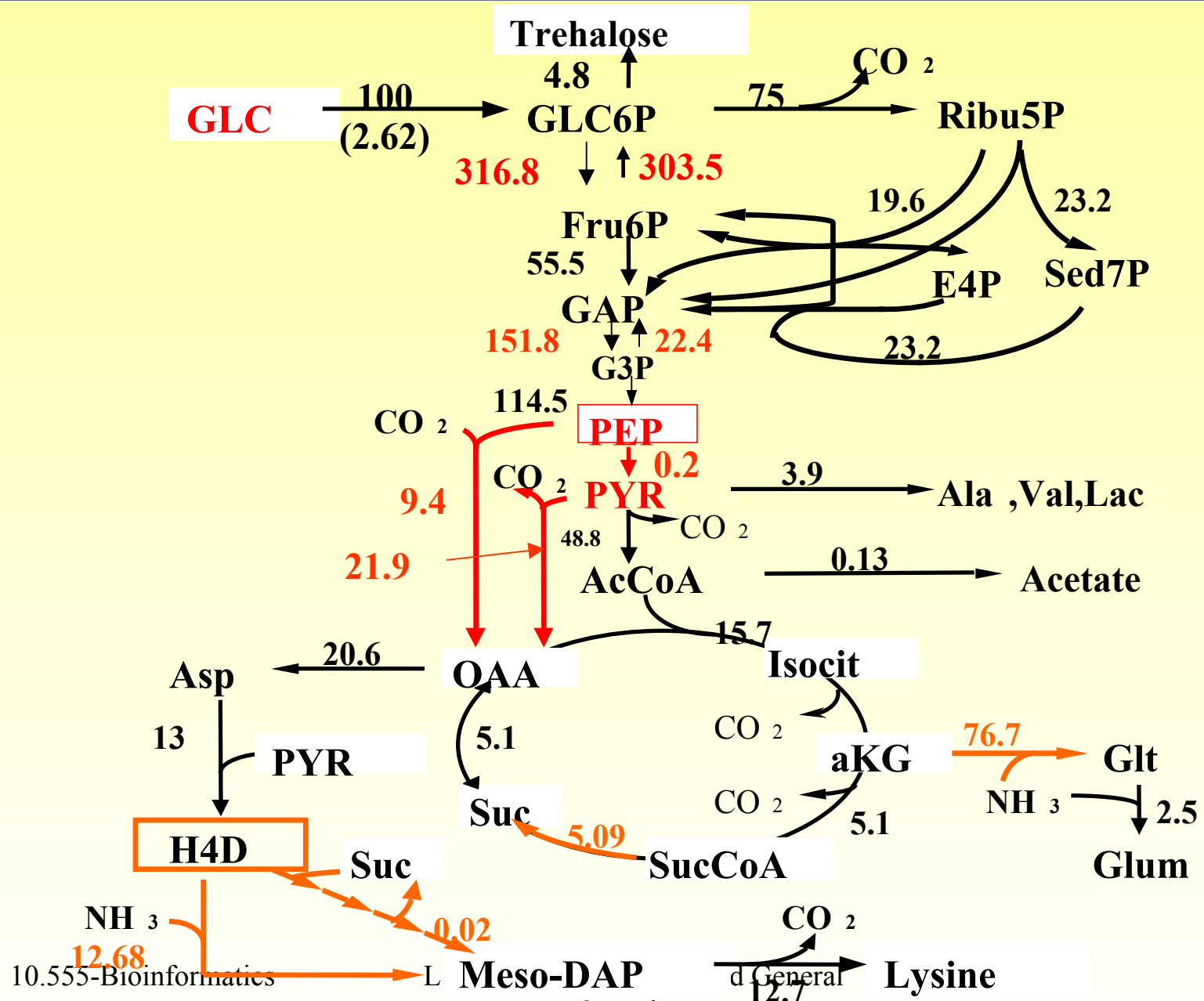
1. Ion Exchange

2. Size exclusion



Fluxes estimated from isotopomer balancing

(locally optimal solution - value of objective: 0.021)



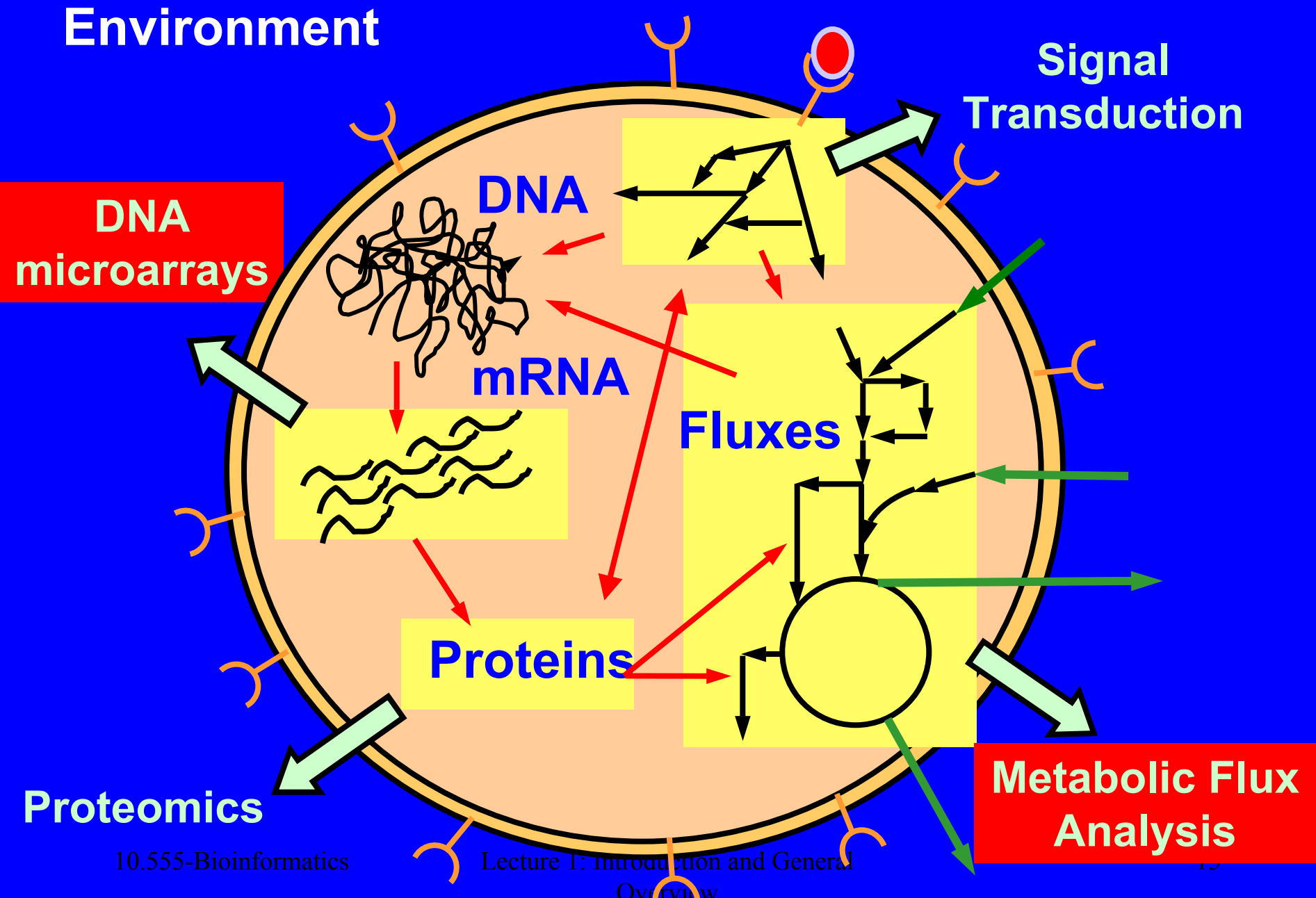
Bioinformatics

The process and methods applied to the **upgrade** of the information content of biological measurements

More specifically:

The utilization of sequence, expression, proteomic and physiological data to identify characteristic patterns of disease and elucidate mechanisms of gene regulation, signal transduction, flux control and overall cell physiology

Probing cellular function



Central Goal

- **Genome sequences and technologies generating large volumes of data**

- ✱ DNA microarrays (*gene chips*)
- ✱ Proteomics
- ✱ Small metabolite measurements
- ✱ Isotope labeling

Challenge:

Utilize genomic information and above data to elucidate gene regulation, metabolic flux control, and overall cell physiology

Importance:

✱ Elucidate genetic regulation and flux controls in their entirety instead of one at a time

Bioinformatics course

Structured to address these driving forces.

It comprises three basic units:

I. Genomics: Sequence-driven problems

II. Data-driven problems

**III. Functional Genomics and Issues of
Systems Biology**

I. Sequence-driven problems

★ Central theme: *Sequence comparisons*

- Metrics of comparison
- Scoring matrices
- Score evaluation
- Comparison by *alignment* vs. comparison by *entries of bio-dictionaries (Teiresias)*

★ Emphasis

- Computer Science
- Applied probability
- Statistics
- Comparative genomics

I. Sequence-driven problems

● **GENOMICS**

- **Fragment reassembly for chromosome sequence**
- **Identify *Open Reading Frames***
- **Identify gene splicing sites (introns)**
- **Gene annotation (*inter-genomic comparisons*)**
- **Determine sequence patterns of *regulatory sites*.
Important in understanding gene regulation and expression characteristics of tissues, disease, phase of development, etc.**
- **Gene regulation (using expression and other data)**

I. Sequence-driven problems (cont'd)

● PROTEOMICS

- Identification of functional domains in protein sequences
- Single, multiple protein alignment (homology)
- Sequence-structure, sequence-function relationships (structural bioinformatics)
- Pattern discovery (phylogeny, remote associations)
- Framework for the analysis of signaling networks
 - **Pathway interactions**
 - **Effect of pathway convergence-divergence on signaling**
 - **Controls of signaling (kinetic, regulatory)**
 - **Integration**

II. Data-driven problems

- ★ **Central theme: *Information upgrade***
- ★ **Transcriptional data (microarrays). Use to:**
 - **Identify *discriminatory genes***
 - **Define clusters of co-expressed genes**
 - **Correlations of gene expression**
 - **Sample classification**
 - ❖ **Disease diagnosis**
 - ❖ **Therapy prognosis**
 - ❖ **Evaluation of side effects**

Data-driven problems (cont'd)

- ★ **Metabolic rate-isotopic tracer data. Use to:**
 - Determine fluxes
 - Study flux control (metabolic engineering)
- ★ **Mass Spectrometry data of proteins. Use to:**
 - Identify proteins and protein fragments
 - Protein quantification

III. Problems of Systems Biology- Physiology- Functional Genomics

Central theme:

***Interactions of cellular parts
Networks
Associations of data***

Methods:

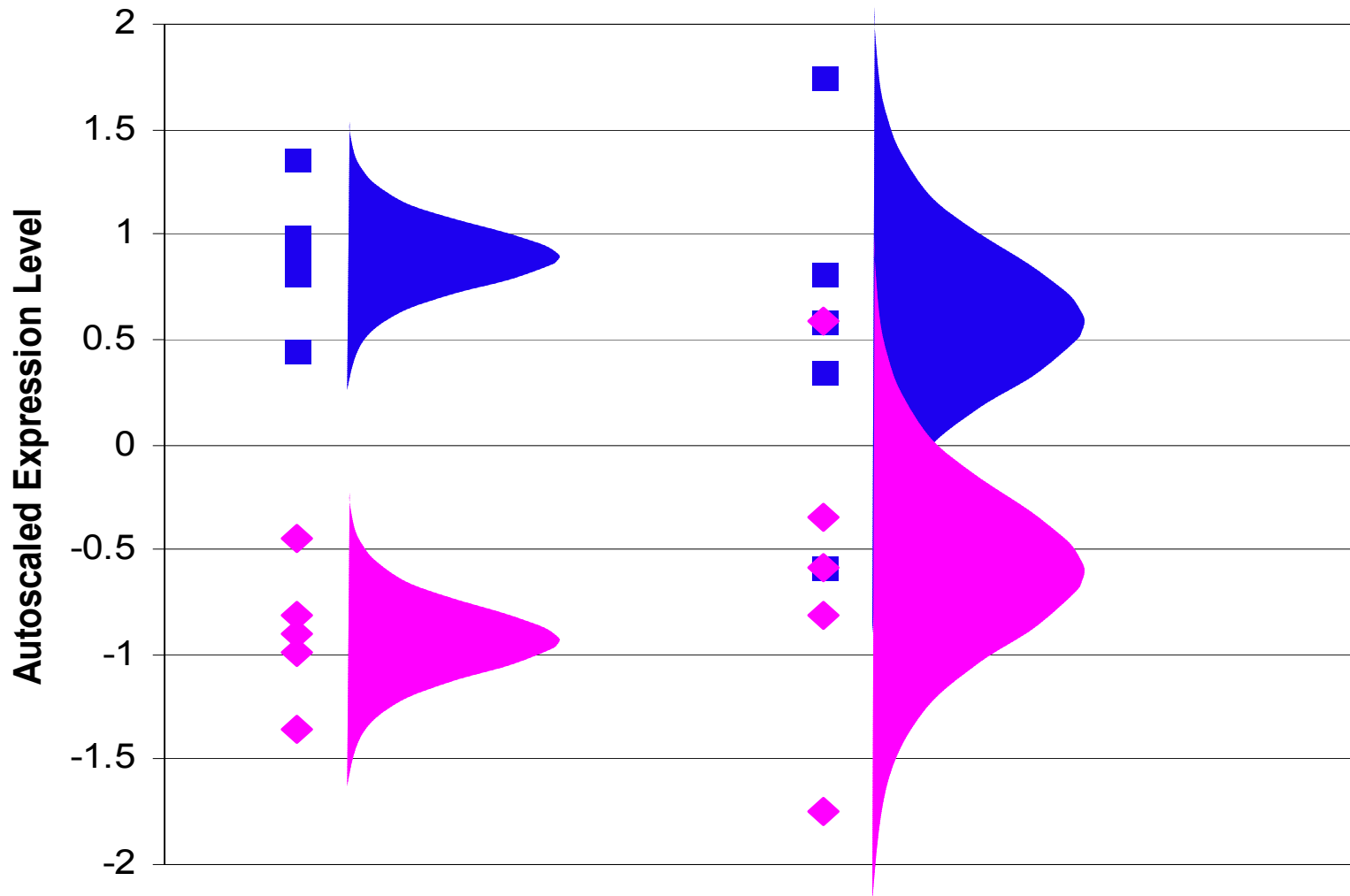
***Network analysis (balances)
Projections (PCA, FDA, etc.)
Correlations (PLS, PCR, etc.)
Pattern Discovery in data***

Interesting problems

Microarray data analysis

- Identify discriminatory genes
- Decipher gene regulation
- Elucidate gene regulatory networks
- Use transcriptional data to discover characteristic patterns of gene expression for different tissue types, disease, etc.

Example: Use microarray expression data to discover discriminatory genes

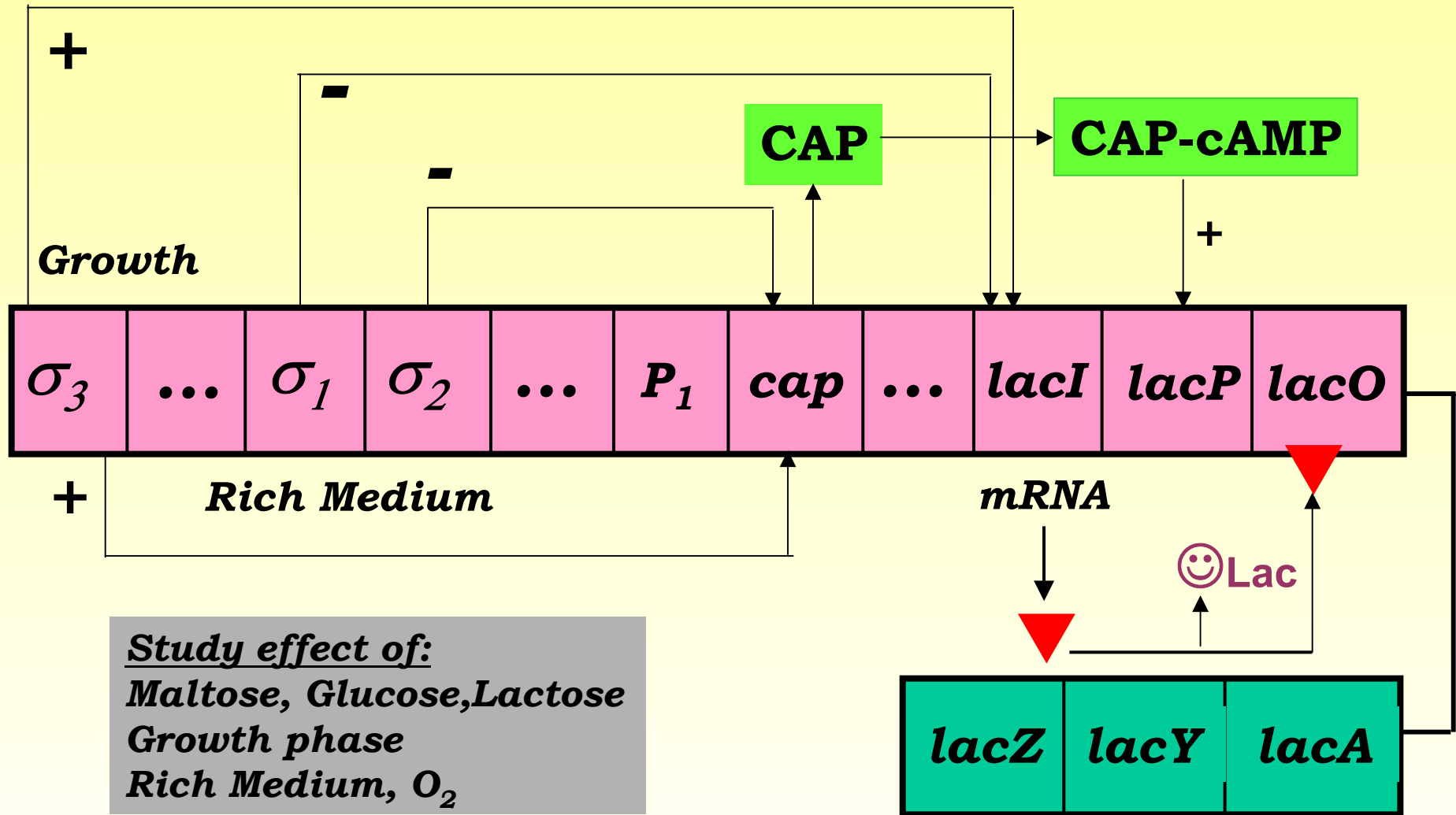


Interesting problems

Microarray data analysis

- Identify discriminatory genes
- Decipher gene regulation
- Elucidate gene regulatory networks
- Use transcriptional data to discover characteristic patterns of gene expression for different tissue types, disease, etc.

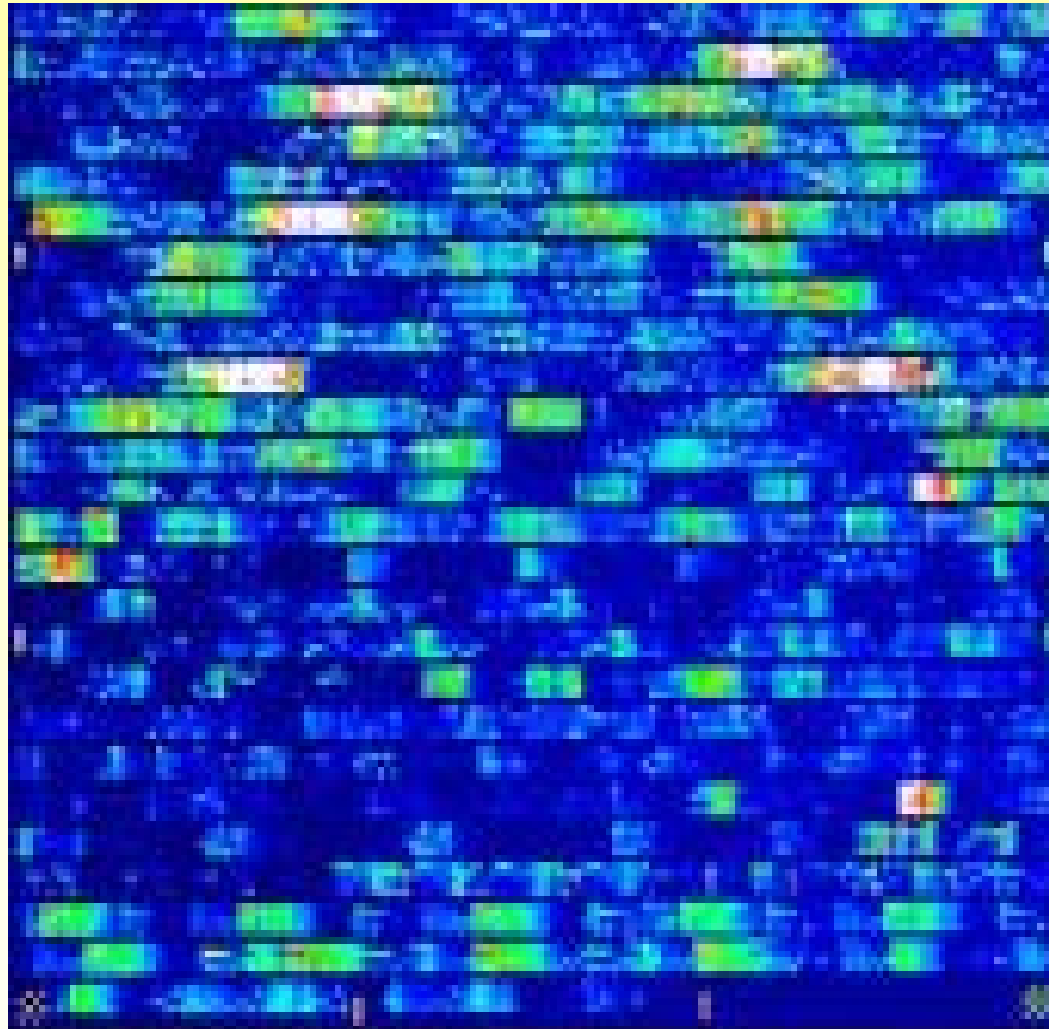
Interesting problems: Decipher structure of gene regulation: Modified lac operon



Expected profiles in differential gene expression experiments

	Glc, Grth	Glc,Lac, No grth	Glc, Rich Medium	Mal,Lac Growth	Glc, Stat.	Grth. Rich Md.
σ_3	●	●		●	●	●
σ_1	●	●		●	●	●
σ_2	●	●		●	●	●
cap_1	●	●		●	●	●
$lacI$	●	●		●	●	●
$lacZ$	●	●		●	●	●
$lacY$	●	●		●	●	●
$lacA$	●	●		●	●	●
cap_2	●	●		●	●	●
$malI$	●	●		●	●	●
$malZ$	●	●		●	●	●
$malY$	●	●		●	●	●
$malA$	●	●		●	●	●

Example of gene chip data



Deciphering microarray data (cont'd)

Step 2: Determine structure of interactions

	P_1	P_2	P_3	P_4	GC_1	GC_2	GC_3	GC_4	GC_5	...
C_1	H	L	L	L	I	R	0	I	I	...
C_2	H	H	H	L	I	R	R	I	0	...
C_3	H	H	L	H	R	R	0	I	R	...

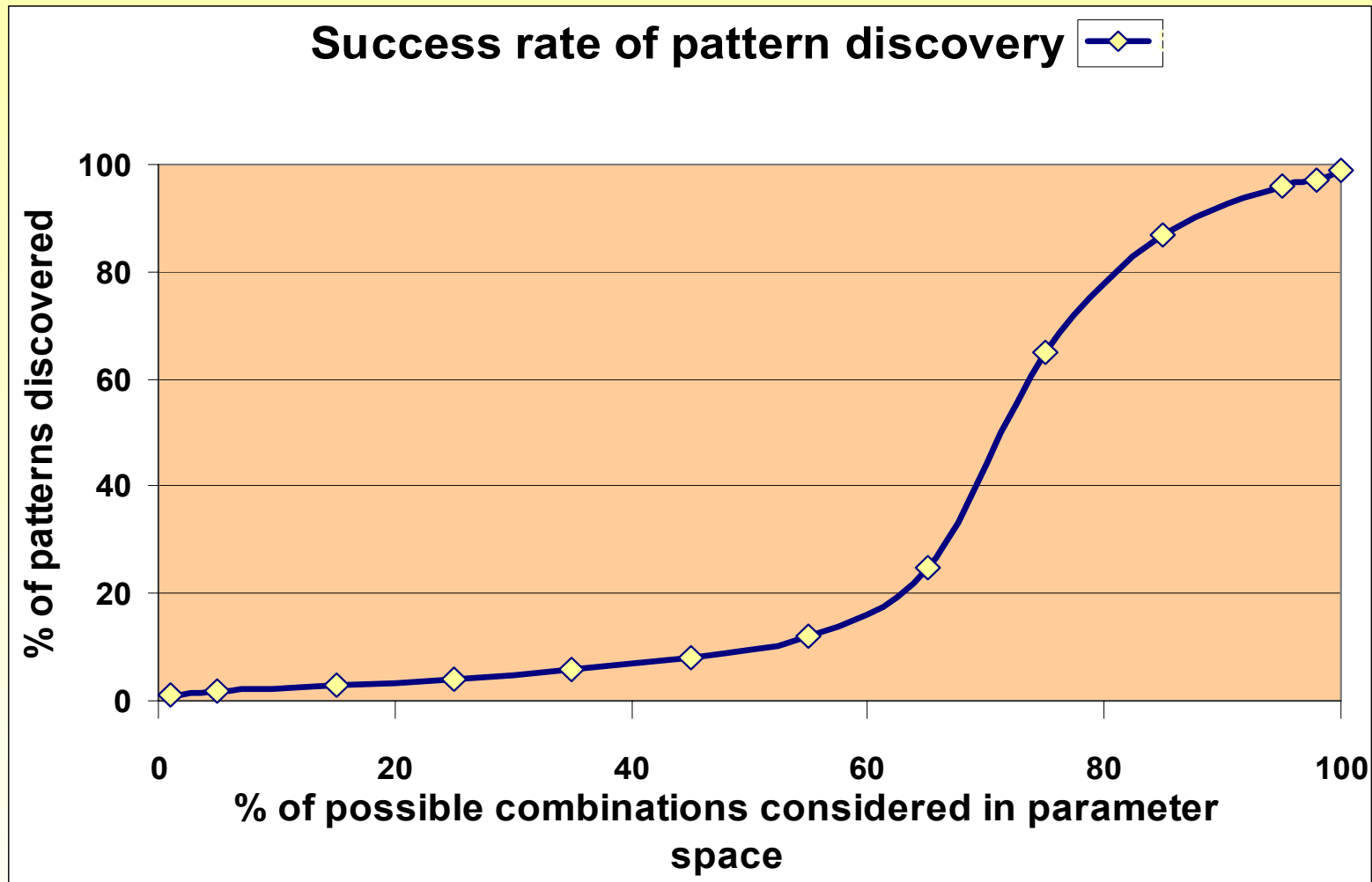
Look for most generic *patterns*:

W W H W	I R I W W I R
L W H W	W I I R W W R
W W W H	W W I R R W I

Apply: Decision trees, Teiresias, pattern discovery

Deciphering microarray data (cont'd)

Simulation results with *lac* operon



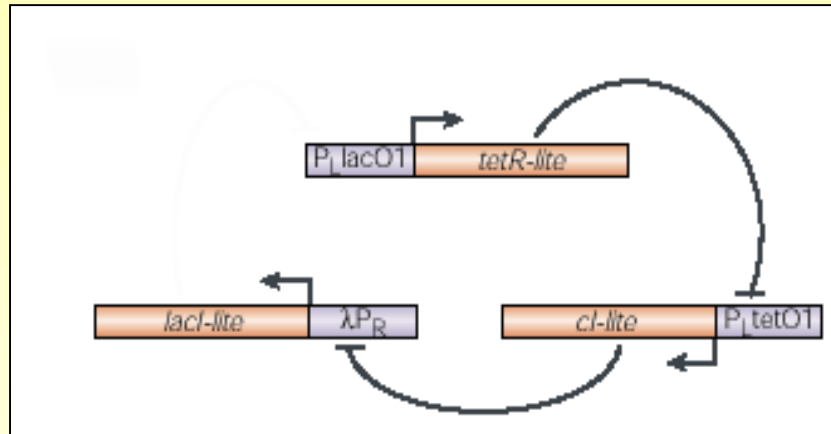
Interesting problems

Microarray data analysis

- Identify discriminatory genes
- Decipher gene regulation
- Elucidate gene regulatory networks
- Use transcriptional data to discover characteristic patterns of gene expression for different tissue types, disease, etc.

Interesting problems of Systems Biology- Physiology-Functional Genomics

★ Genetic Regulatory Networks



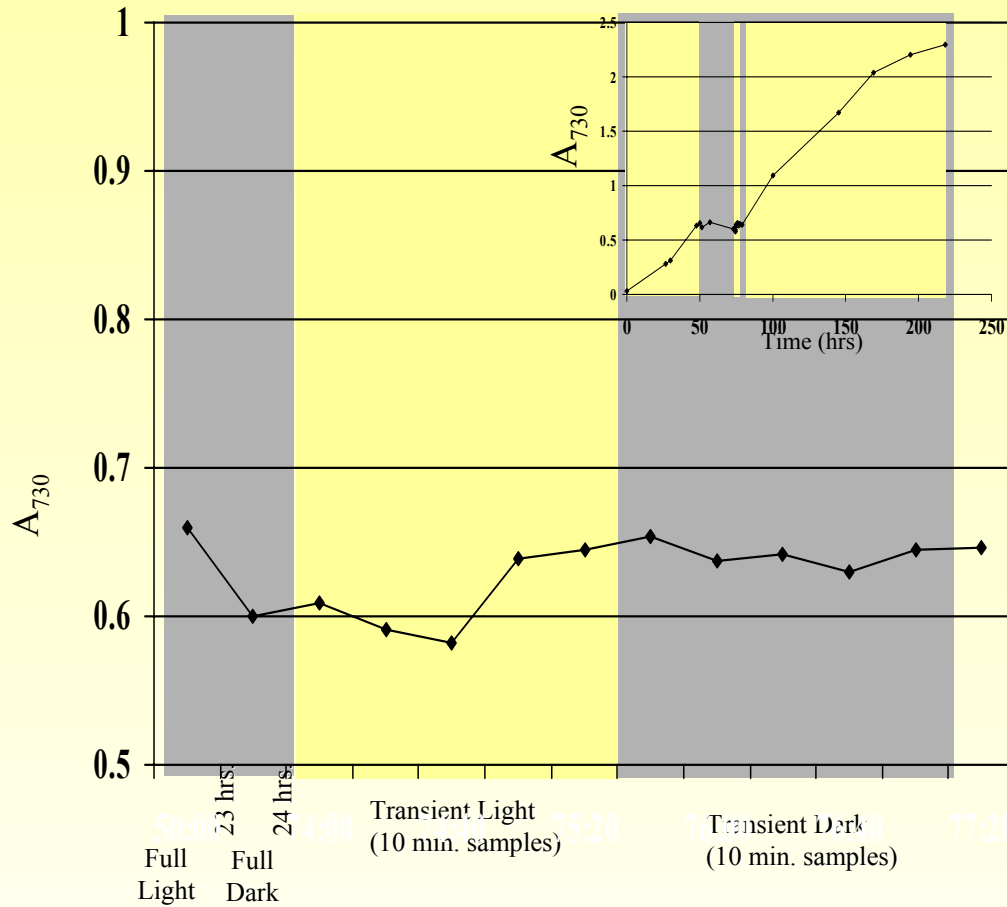
- Apply time-lagged *correlations* in transcriptional data
- Apply *pattern discovery* in upstream sequence regions of co-expressed genes

Interesting problems

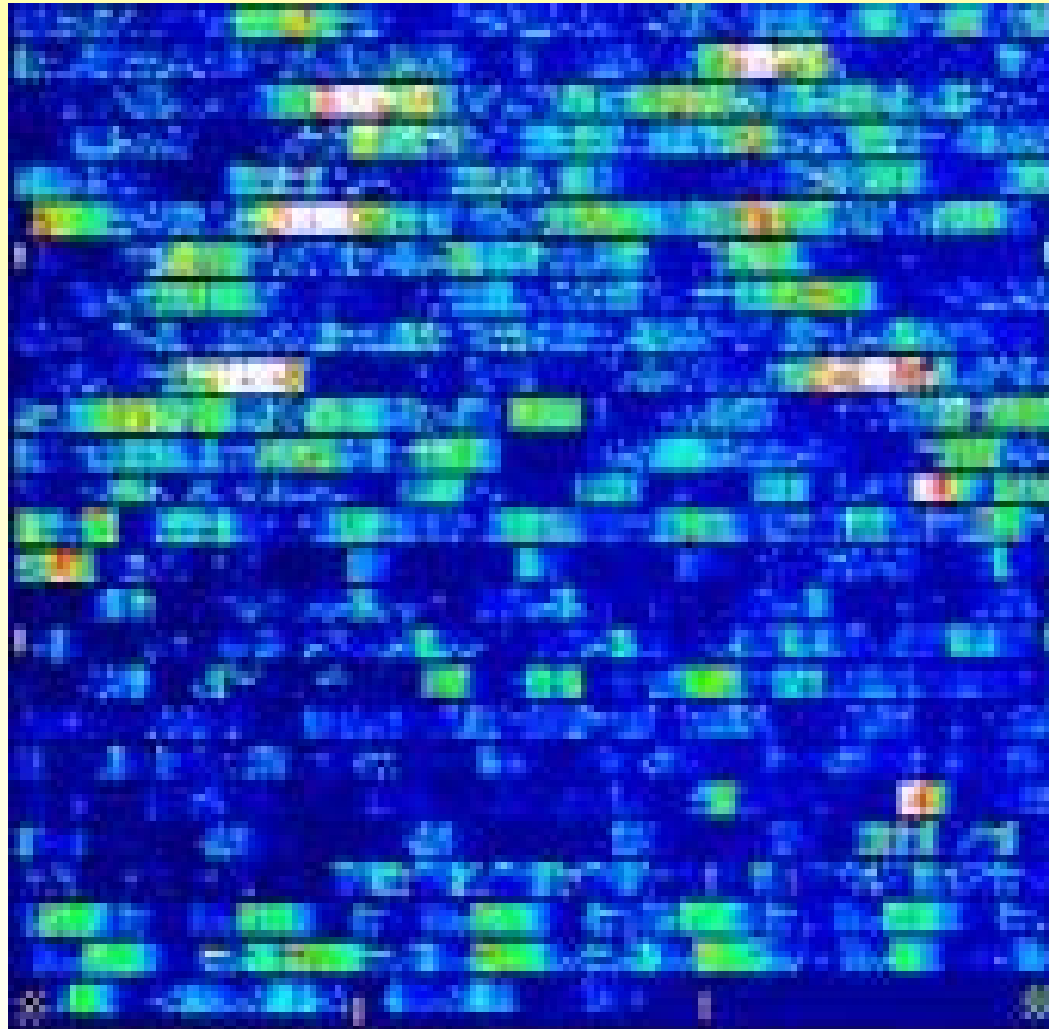
Microarray data analysis

- Identify discriminatory genes
- Decipher gene regulation
- Elucidate gene regulatory networks
- Use transcriptional data to discover characteristic patterns of gene expression for different tissue types, disease, etc.

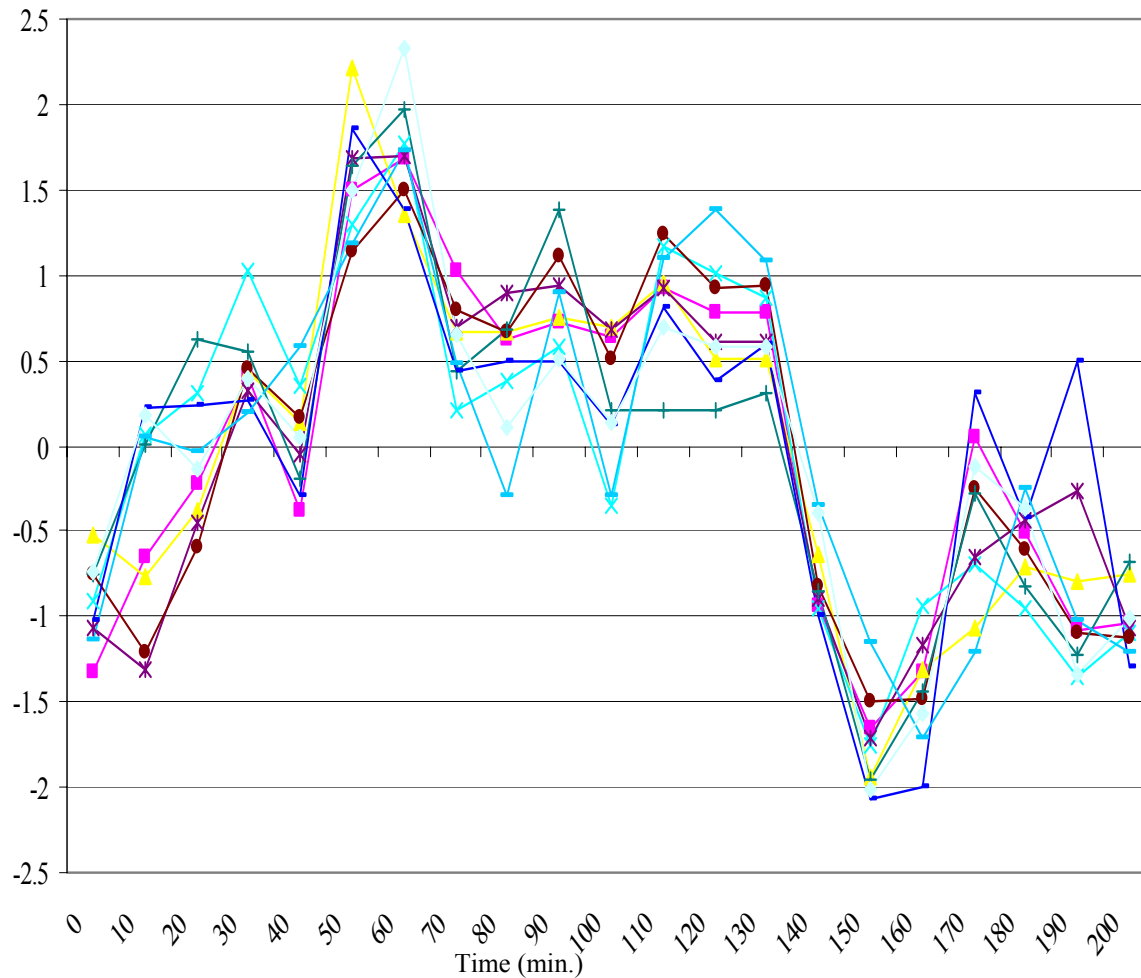
Example: Defining the physiological state from the expression phenotype



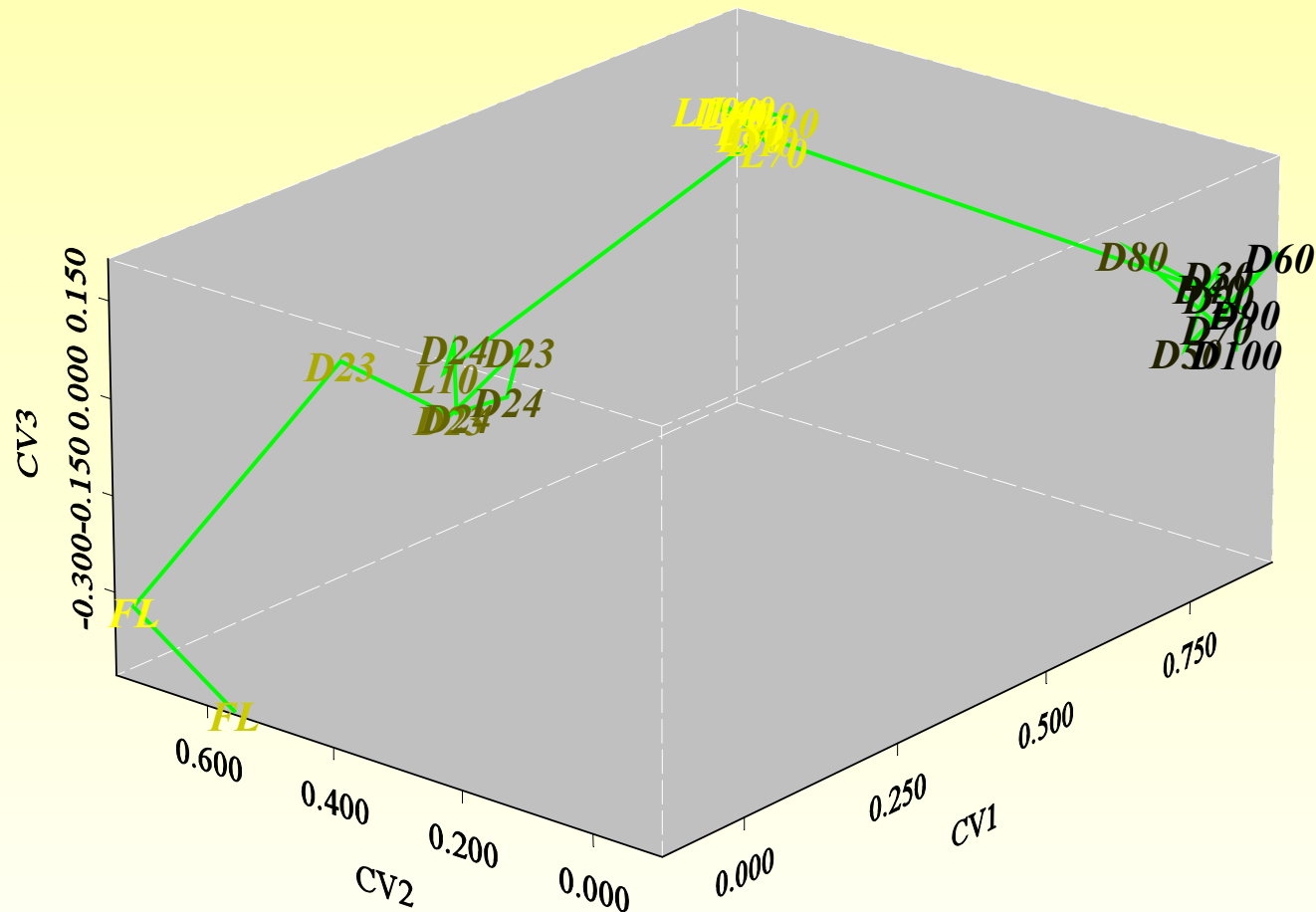
Example of gene chip data

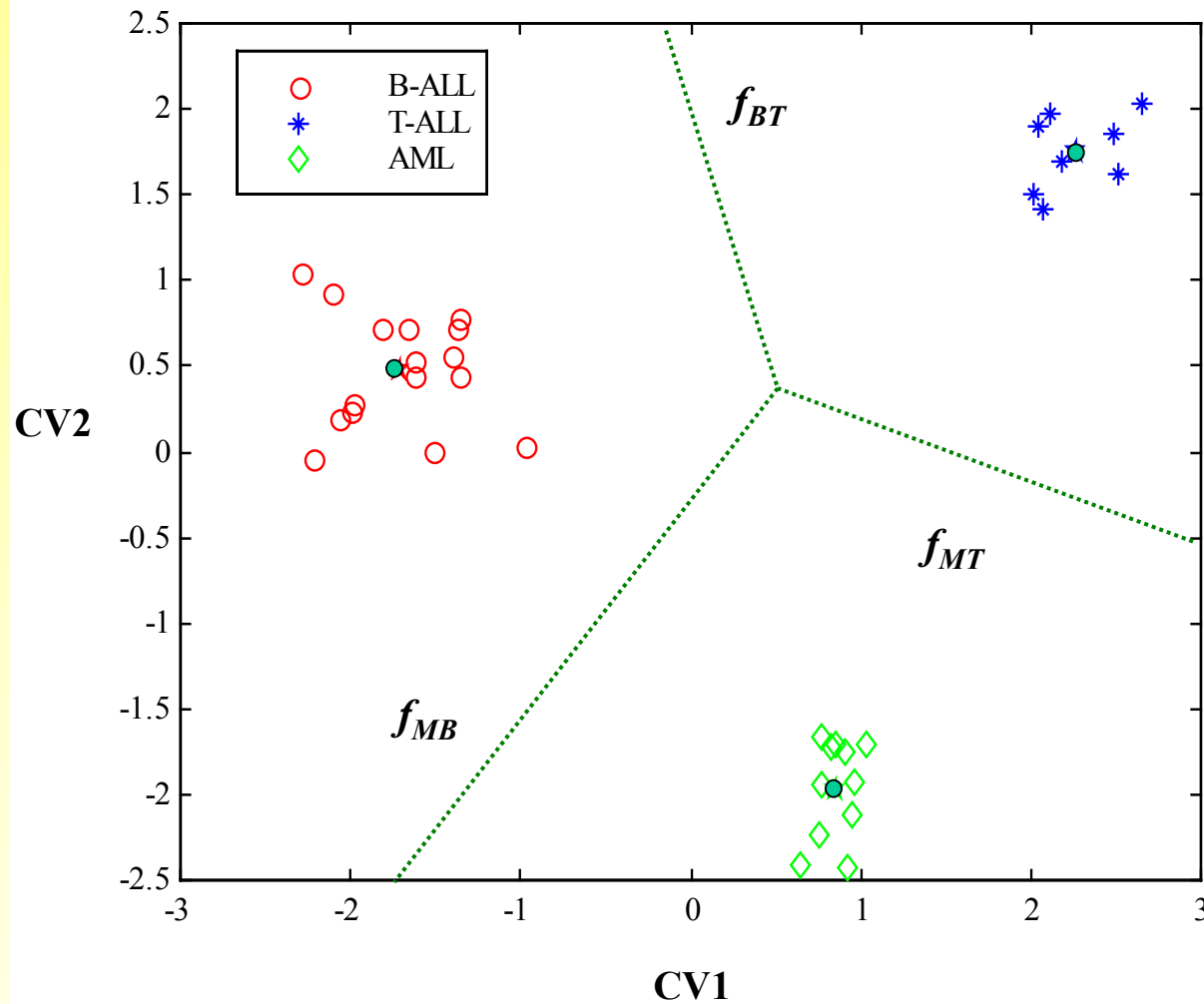


Example (cont'd)



Dimensionality reduction of the expression phenotype space by CDA (10 minute time delay)





Decision boundaries

f_{BT} : between B-ALL and T-ALL

f_{MT} : between T-ALL and AML

f_{MB} : between AML and B-ALL

$$f_{mk} : \mathbf{y} = \mathbf{V}^T \mathbf{x}$$

$$(\bar{\mathbf{y}}_m - \bar{\mathbf{y}}_k)^T \mathbf{y} - \frac{1}{2}(\bar{\mathbf{y}}_m - \bar{\mathbf{y}}_k)^T (\bar{\mathbf{y}}_m + \bar{\mathbf{y}}_k) \geq 0$$

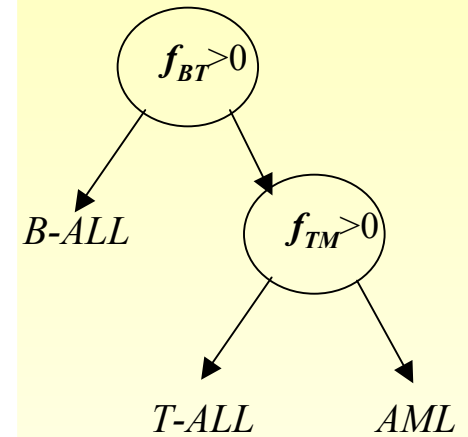


Figure 5: Physiological state domains are defined by a series of decision boundaries for the projected values and group means in the reduced CV plane.

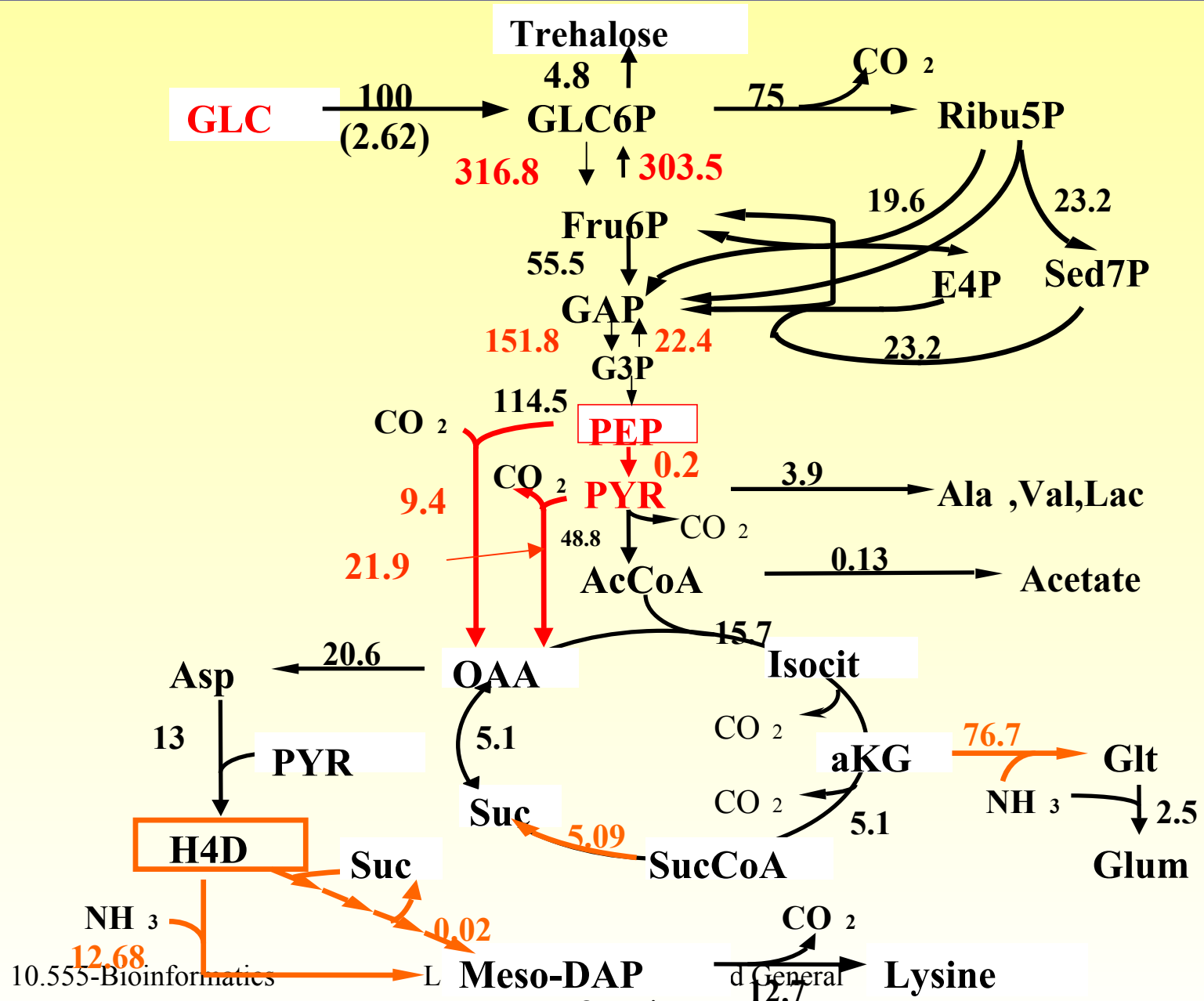
Interesting problems (cont'd)

- **METABOLISM (Metabolic Engineering)**

- ✱ **Determination of *in vivo* metabolic fluxes from metabolite and isotopomer balances**
- ✱ **Determine rate controlling steps in metabolic networks to achieve directed flux amplification**
- ✱ **Establish link between the expression phenotype and the metabolic phenotype as determined by the fluxes**
- ✱ **Discover patterns in fluxes**
- ✱ **Framework for the analysis of metabolic networks**
 - **Pathway interactions**
 - **Controls of flux (kinetic, regulatory)**
 - **Integration**

Fluxes estimated from isotopomer balancing

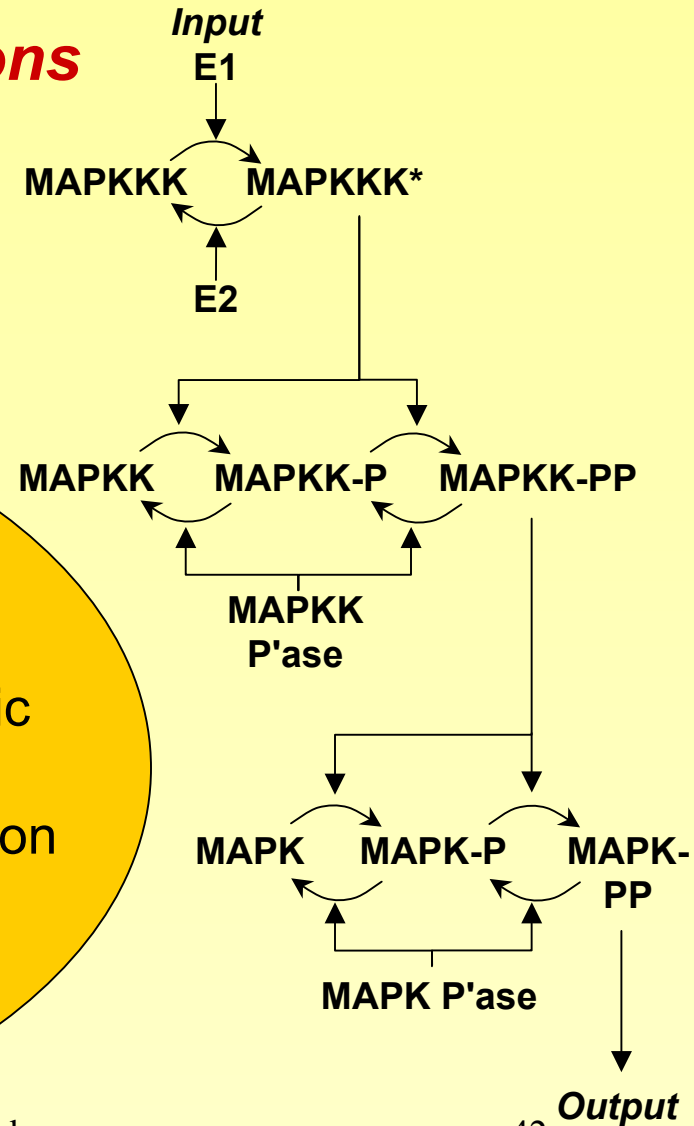
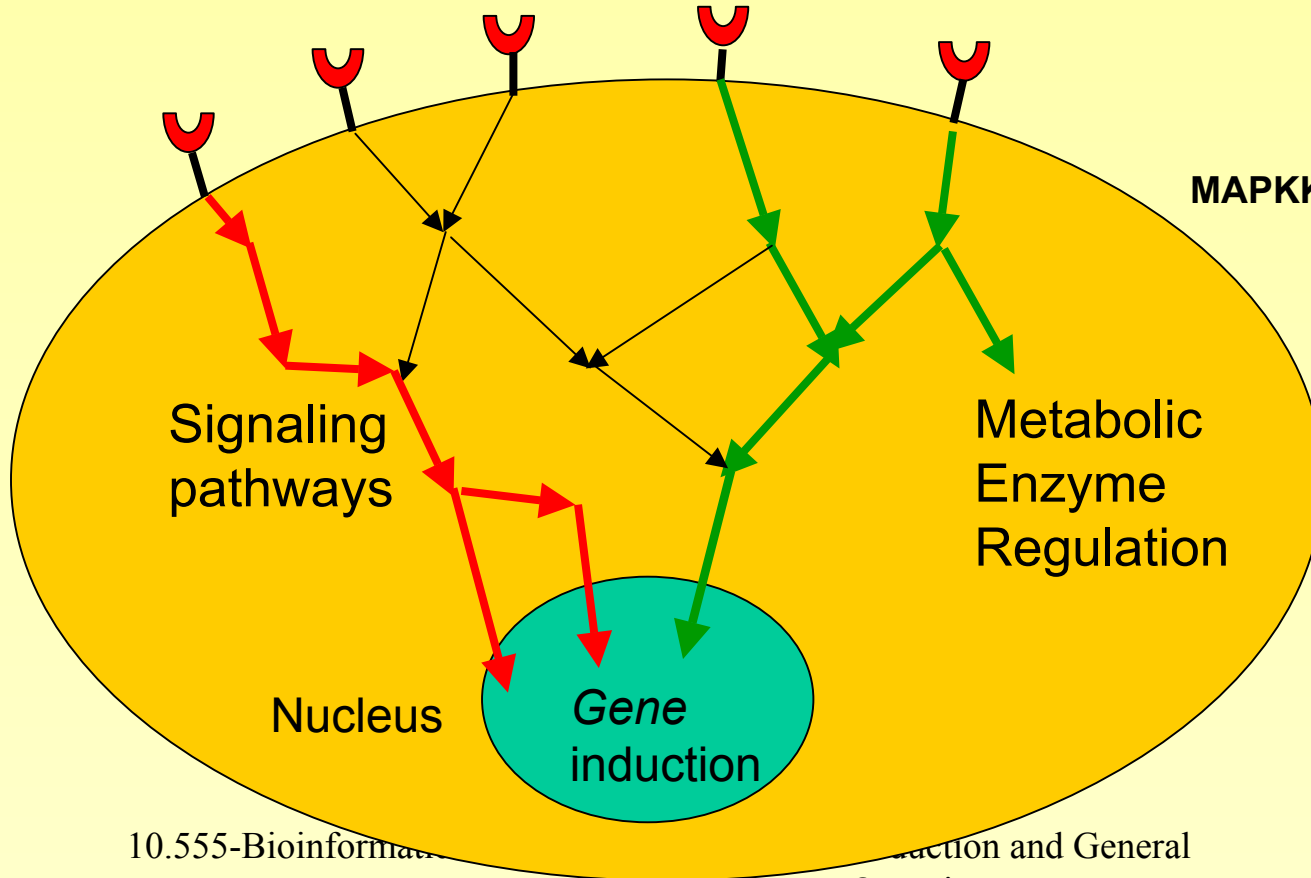
(locally optimal solution - value of objective: 0.021)



Interesting problems: Signal transduction pathways

★ Analysis of structure and interactions of signal transduction pathways

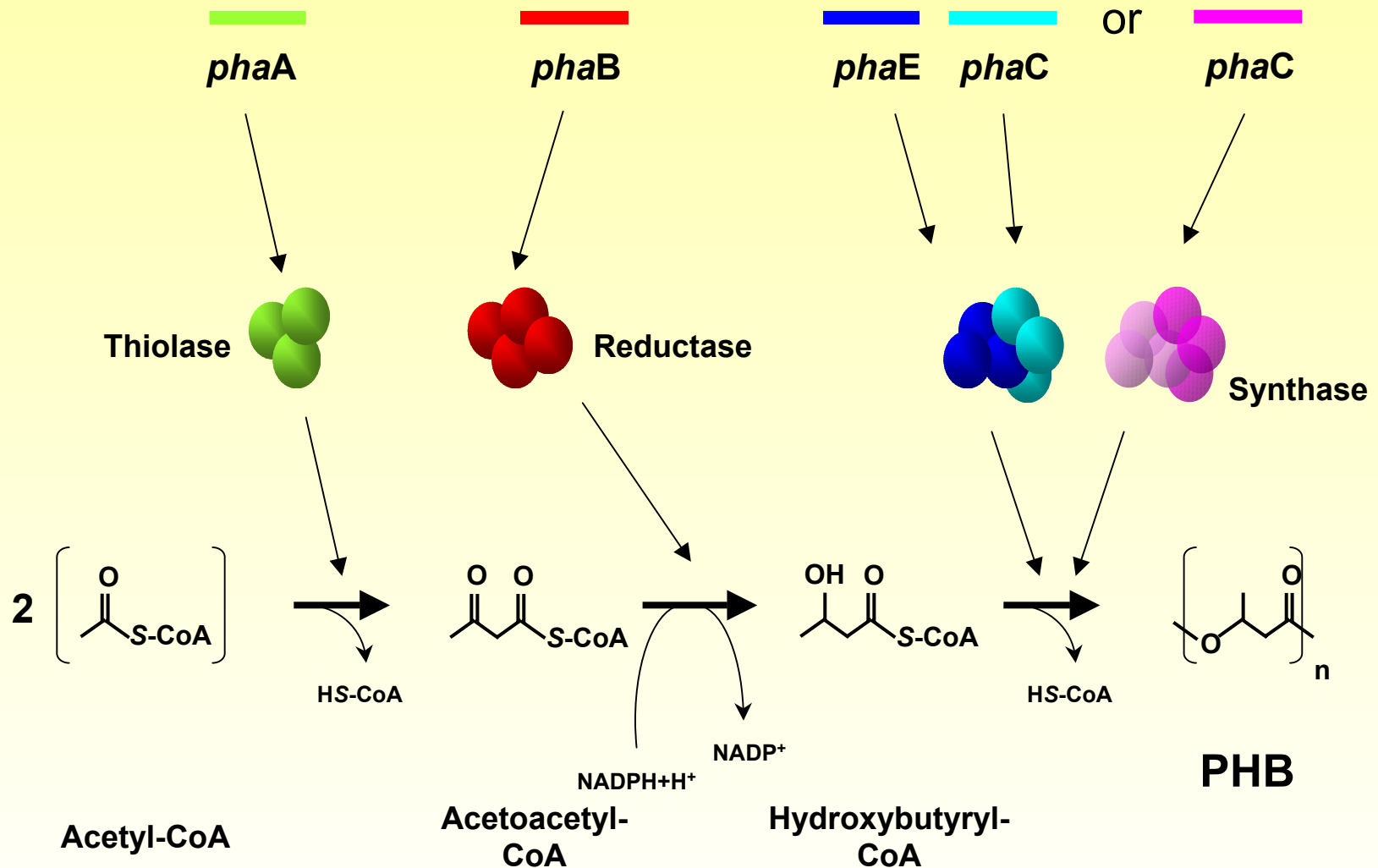
Receptor-Ligand binding



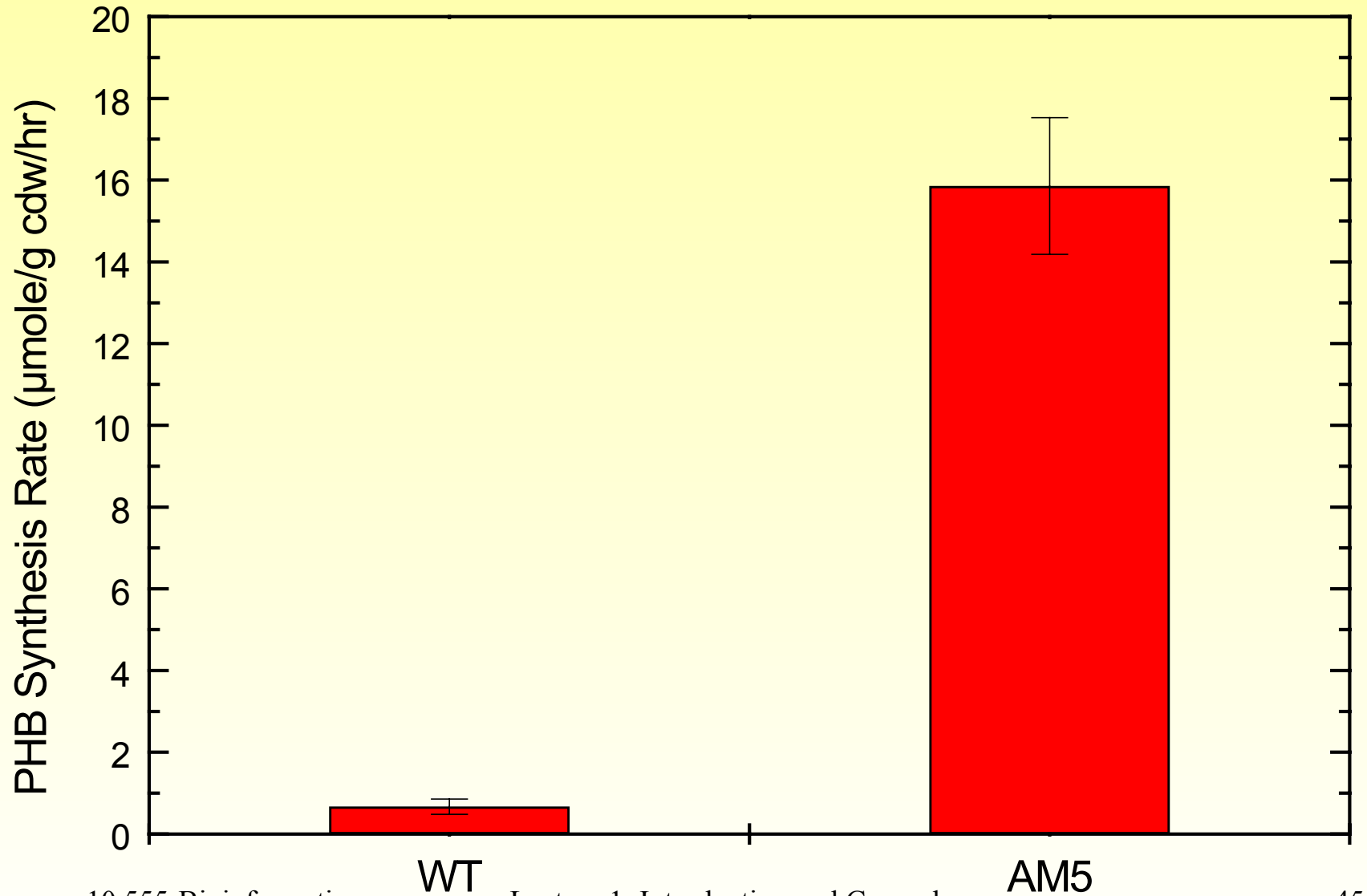
Problems of Systems Biology-Physiology- Functional Genomics

- * Identification of genes that *collectively***
 - **Correlate with a phenotype**
 - **Are critical for important physiological properties**

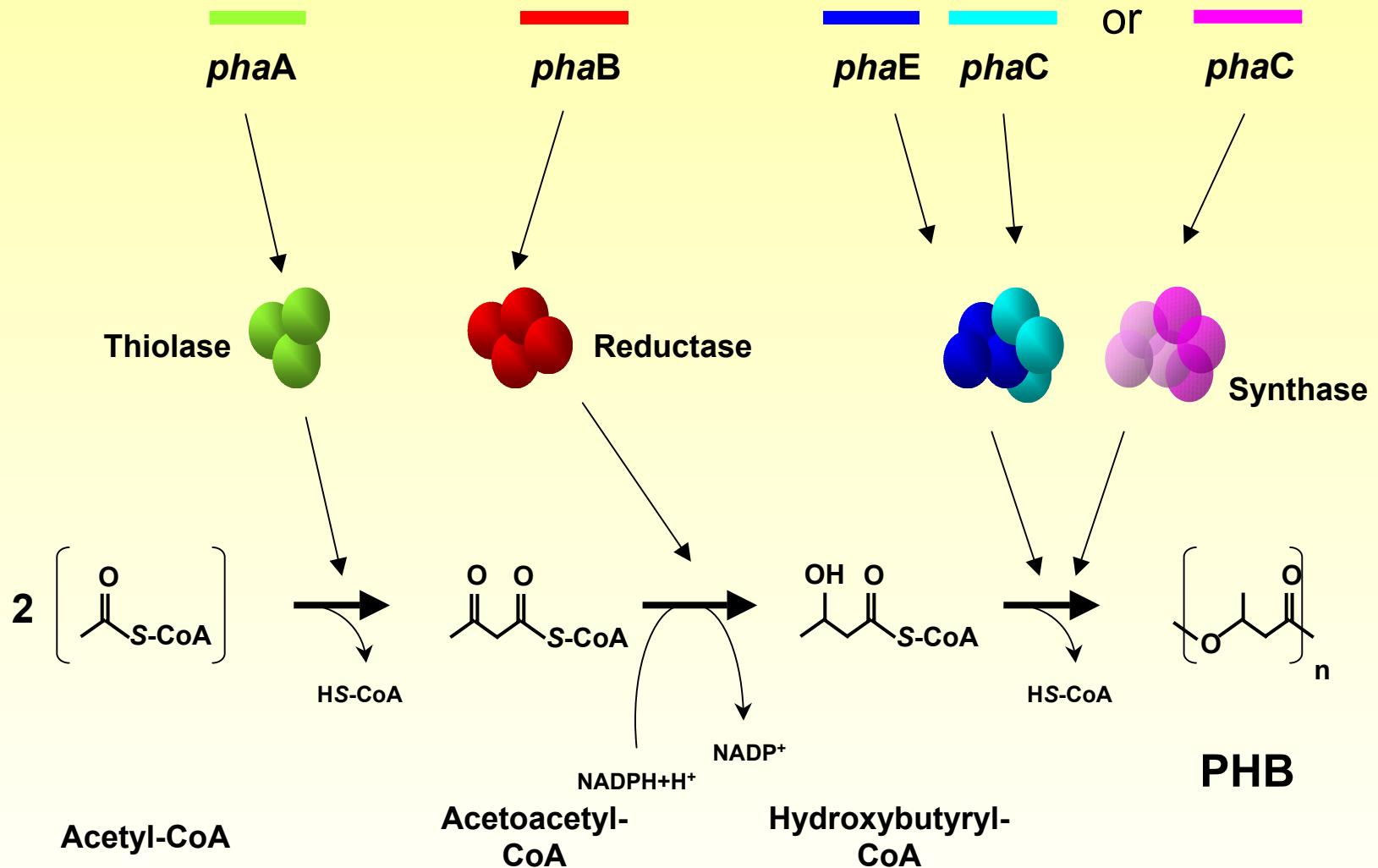
PHA/PHB biosynthesis genes



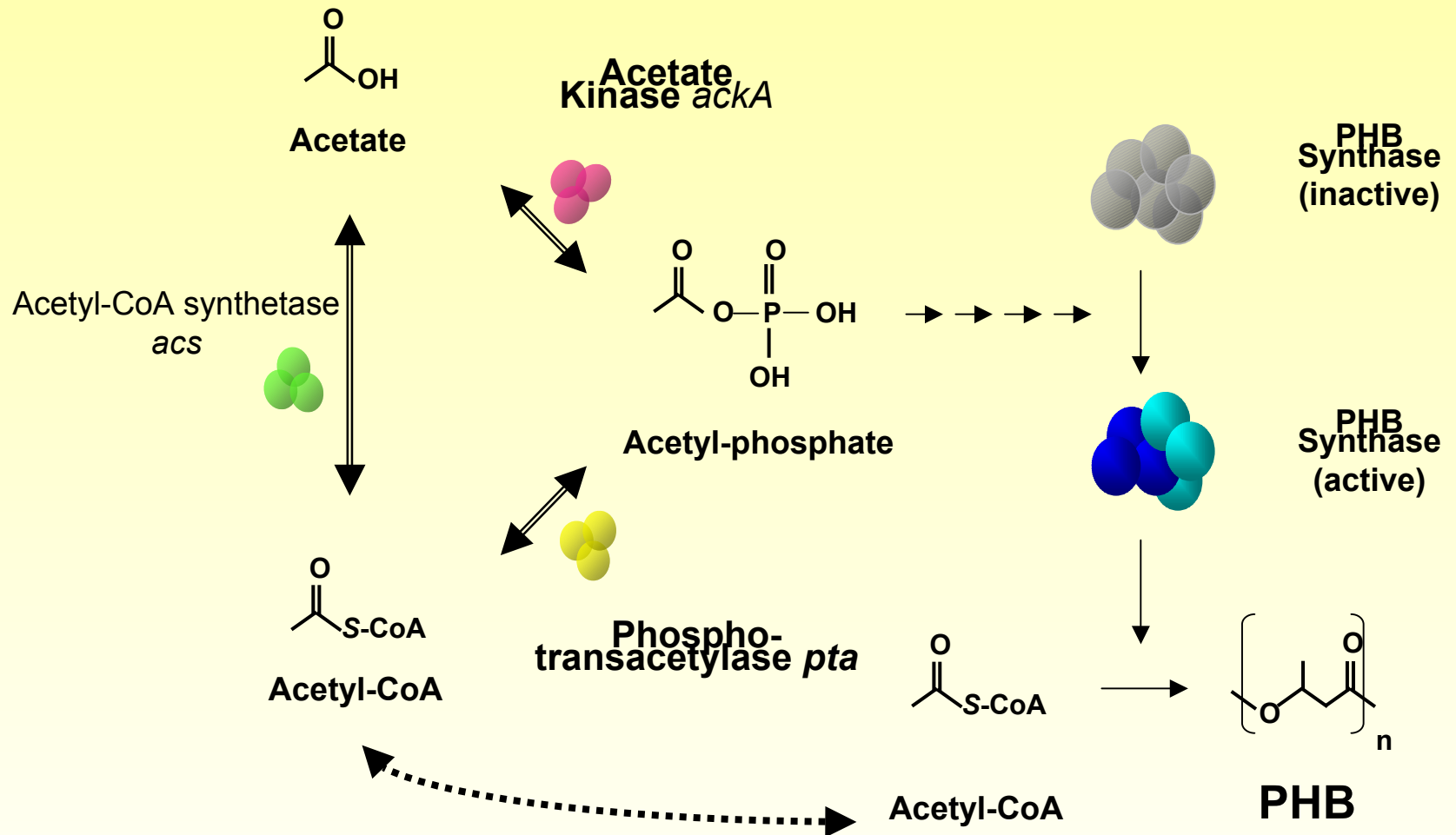
Drug Discovery (cont'd)



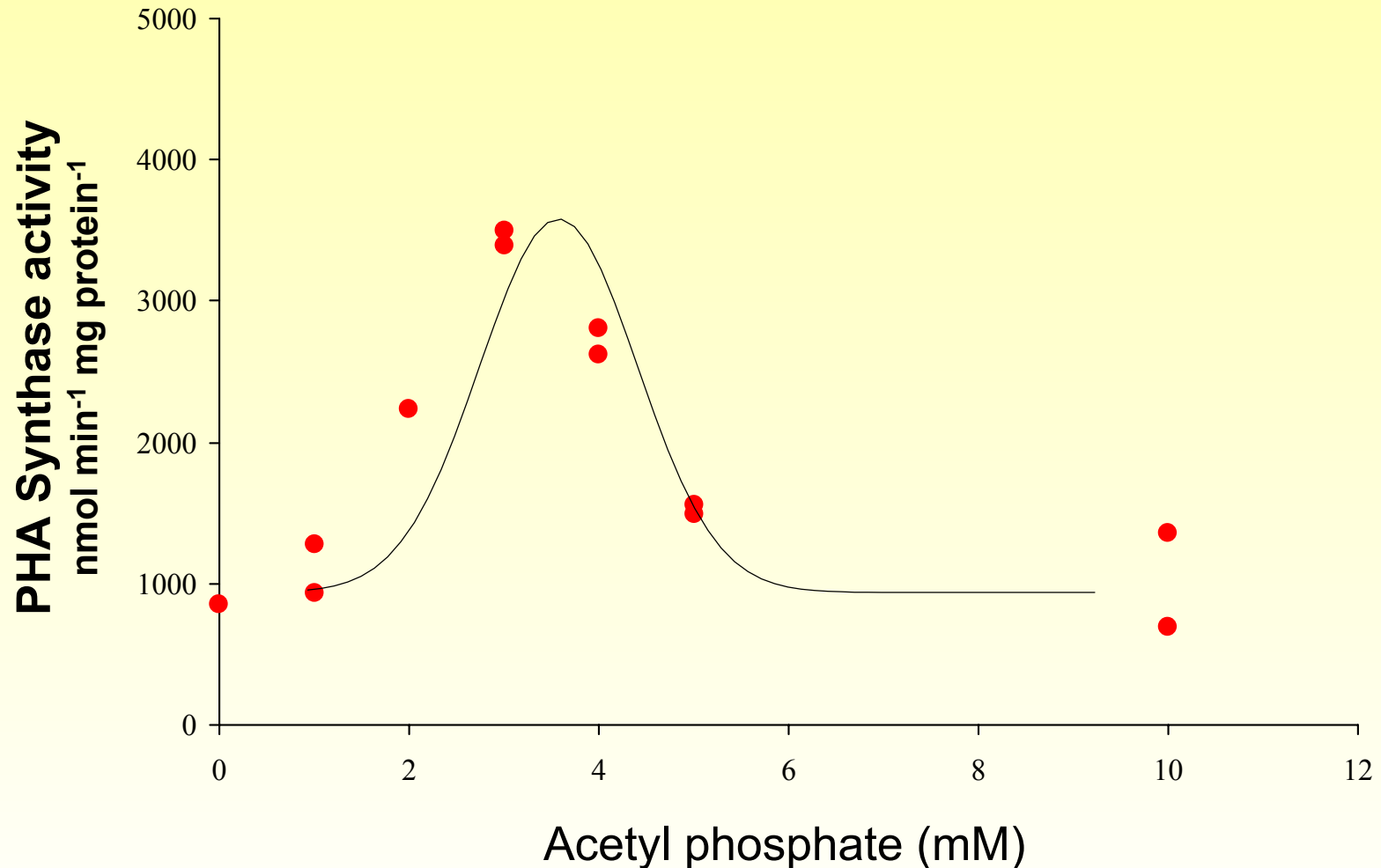
PHA/PHB biosynthesis genes



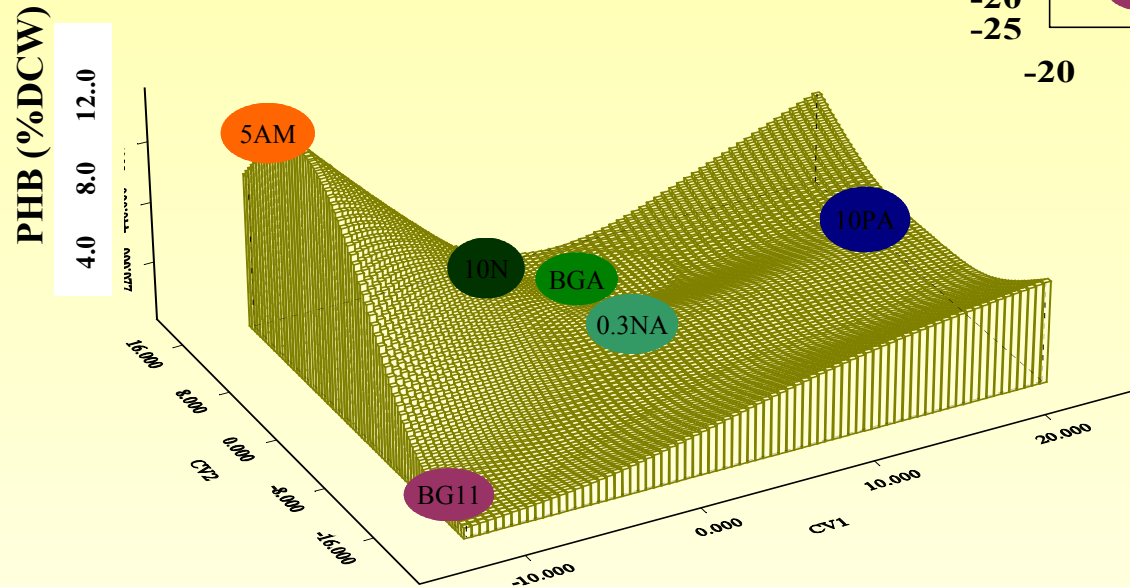
Regulation of PHA biosynthesis in *Synechocystis* sp.



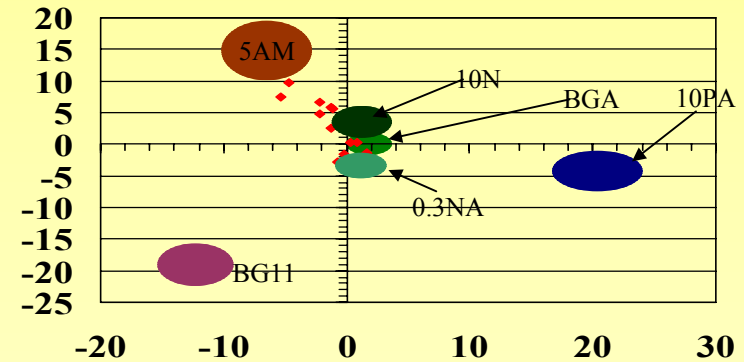
Acetyl phosphate dependent activation of the *Synechocystis* sp. PHA synthase



(A)



(B)



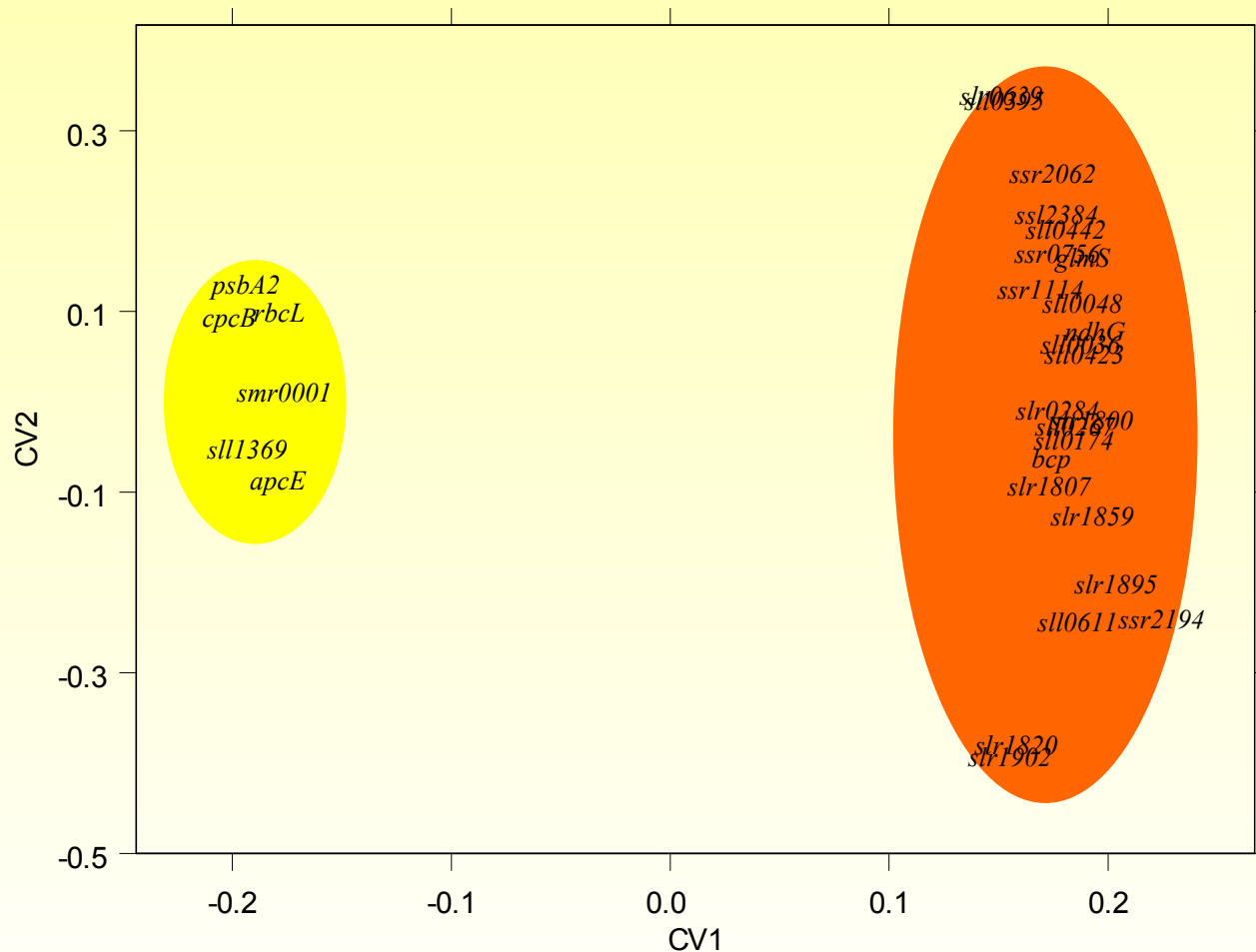
(C)

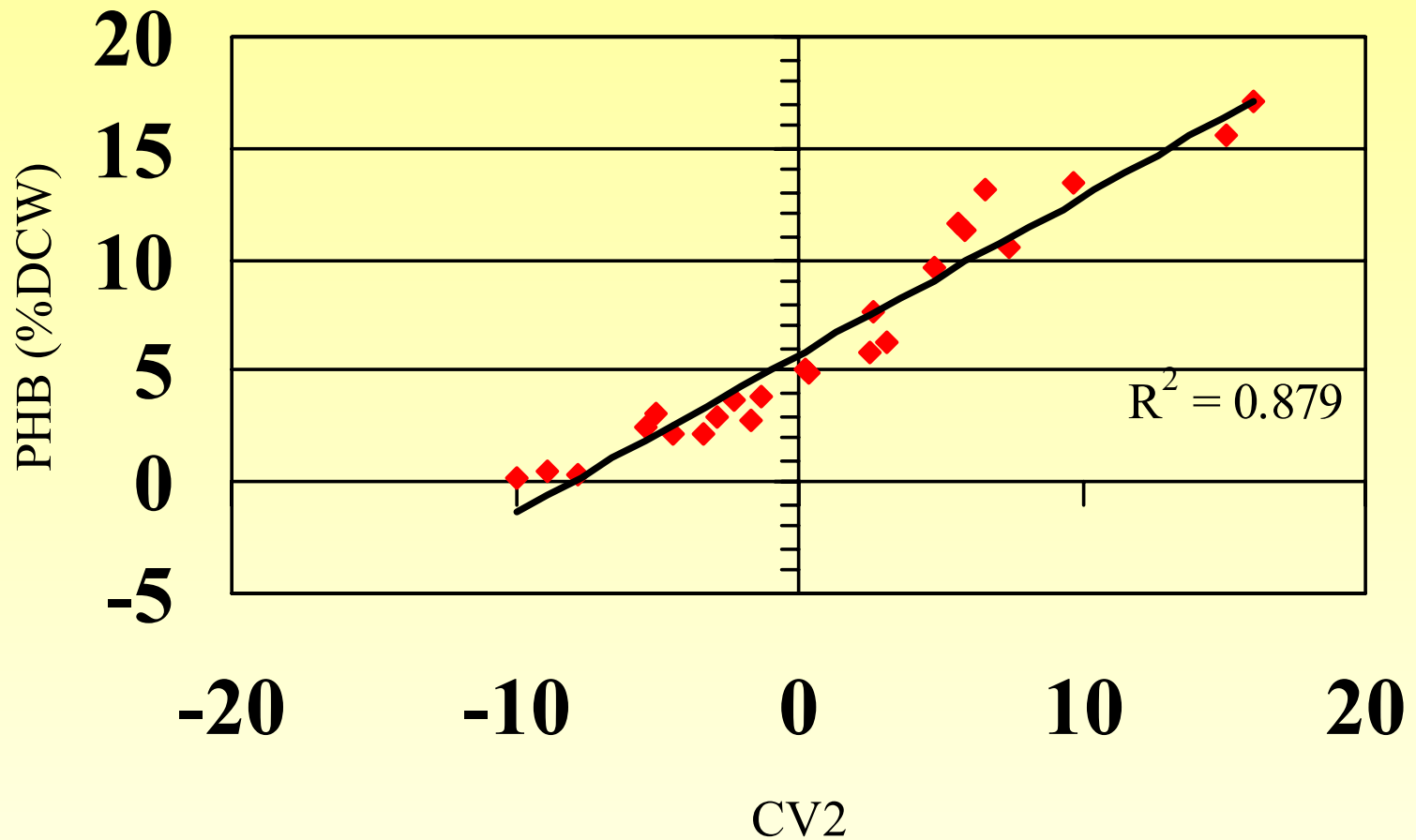
Dimensionality reduction by Fisher Discriminant Analysis

$$\text{Projections: } CV_1 = a_1g_1 + a_2g_2 + a_3g_3 + \dots + a_ng_n$$

$$CV_2 = b_1g_1 + b_2g_2 + b_3g_3 + \dots + b_ng_n$$

Plotting the loadings *(i.e., the coefficients of linear projection)*

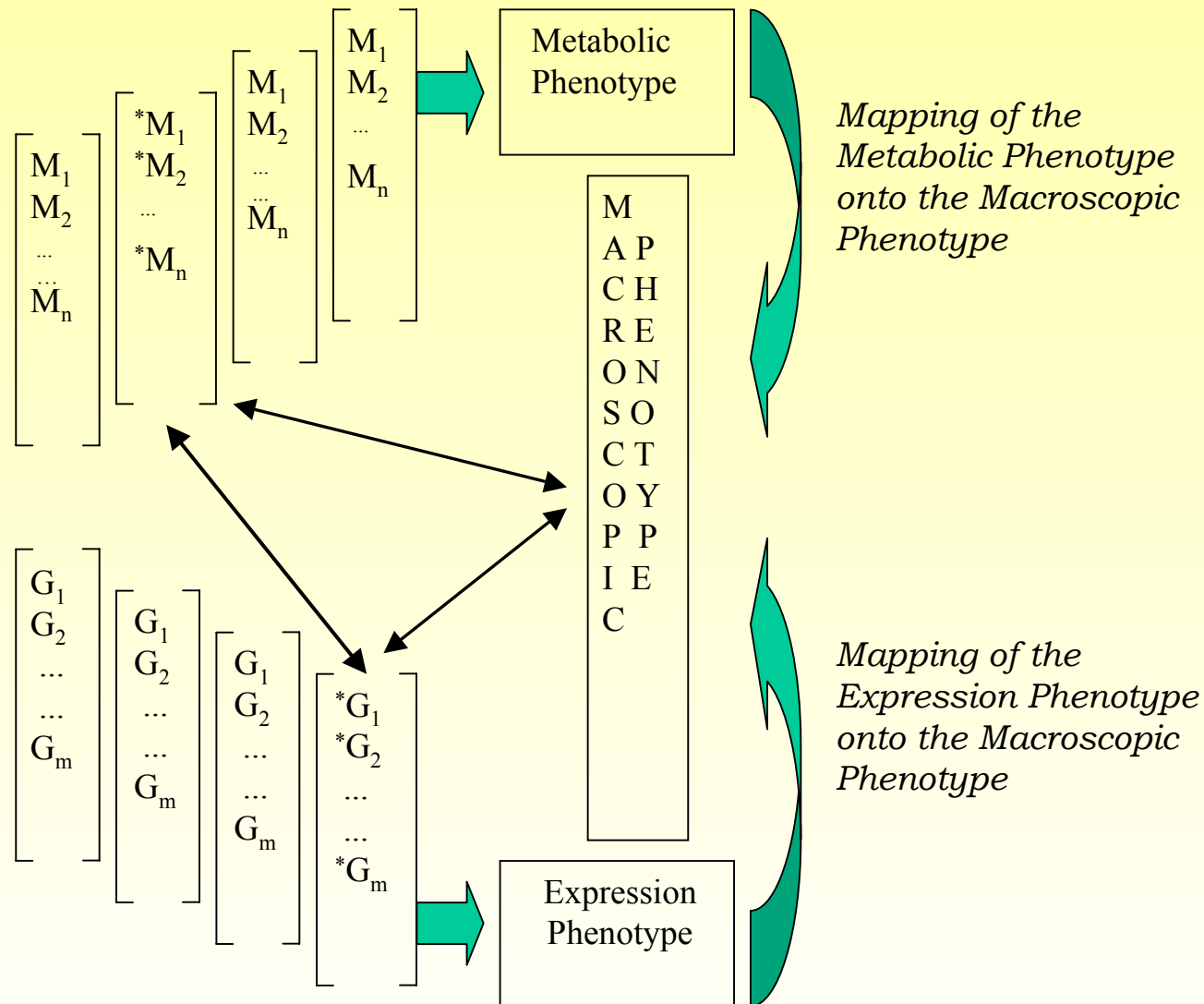




Correlational structure of PHB accumulation

$$\text{PHB} = c_1g_1 + c_2g_2 + c_3g_3 + \dots + c_ng_n$$

Linking the Metabolic and Expression Phenotypes



Bioinformatics

The process and methods applied to the **upgrade** of the information content of biological measurements

More specifically:

The utilization of sequence, expression, proteomic and physiological data to identify characteristic patterns of disease and elucidate mechanisms of gene regulation, signal transduction, flux control and overall cell physiology

In summary

- ★ **Bioinformatics is driven by genomics and data**
- ★ **These driving forces define three essential units (or types of problems) of instruction:**
 - **Sequence-driven problems**
 - **Data-driven problems**
 - **Problems of physiology or systems biology**
- ★ **While initially focused on sequence analysis, the future belong to successful linkage of genomics and *physiology***
- ★ **This is a young field very much in flux with many problems in search of good solutions**

10.555 Bioinformatics: Principles, Methods and Applications

Instructors

Gregory Stephanopoulos and Isidore Rigoutsos

Special lectures by Dr. Joanne Kelleher

Assistant-Grading: Adrian Fay, afi@mit.edu

9 units (H), Class meets Tuesday 2-5pm, Room 66-154

COURSE SYLLABUS

Part I: INTRODUCTION

Lecture 1: February 3

- Historical perspectives, definitions
- Impact of genomics on problems in molecular and cellular biology; need for integration and quantification, contributions of engineering
- Overview of problems. Sequence driven and data driven problems
- Rudiments of dynamic programming with applications to sequence analysis
- Overview of course methods
- Integrating cell-wide data, broader issues of physiology

Lecture 2: February 10 (Assignment 1, due February 24)

- Primer on probabilities, inference, estimation, Bayes theorem
- Dynamic programming, application to sequence alignment, Markov chains, HMM

No class on February 17 (Monday schedule-President's day)

Lecture 3: February 24 (Assignment 2, due on March 2)

- Primer on Biology (the units, the code, the process, transcription, translation, central dogma, genes, gene expression and control, replication, recombination and repair)
- Data generation and storage
- Schemes for gene finding in prokaryotes/eukaryotes
- Primer on databases on the web
- Primer on web engines

Lecture 4: March 2 (Assignment 3, due March 9)

- Some useful computer science (notation, recursion, essential algorithms on sets, trees and graphs, computational complexity)
- Physical mapping algorithms
- Fragment assembly algorithms
- Comparison of two sequences
- Dynamic programming revisited
- Popular algorithms: Smith-Waterman, Blast, Psi-blast, Fasta

Lecture 5: March 9 (Assignment 4, due on March 16)

- Building and using scoring matrices
- Multiple sequence alignment
- Functional annotation of sequences
- Micro RNA, Antimicrobial peptides

Lecture 6: March 16 (Assignment 5, due March 30)

- Pattern discovery in biological sequences
- Protein motifs, profiles, family representations, tandem repeats, multiple sequence alignment and sequence comparison through pattern discovery- Functional annotation-Promoter site recognition

No class on March 23: Spring Break

Lecture 7: March 30 (Assignment 6, due on April 13): Physiology

- Primer on cell physiology. Definition at the macroscopic level
- Molecular cell physiology. Interactions of pathways, cells, organs. Measurements and their integration
- Distribution of kinetic control. Rudiments of metabolic control analysis (MCA)

PART III: UPGRADING EXPRESSION AND METABOLIC DATA

Lecture 8: April 6: Fluxes

- MCA continued. Analysis of metabolic pathways, fluxes, the *metabolic phenotype*
- Methods for metabolic flux determination
- Methods for underdetermined systems

Lecture 9: April 13 (Assignment 7, due on April 27): Microarrays

- Monitoring gene expression levels. DNA microarrays
- Data collection, error analysis, normalization, filtering
- Novel applications of DNA microarrays
- Analysis of gene expression data
- Clustering methods: Coordinated gene expression
- Identification of discriminatory genes
- Determination of discriminatory gene expression patterns. Use in diagnosis
- Data visualization. Reconstruction of genetic regulatory networks

No class on April 20: Patriots Day

Lecture 10: April 27: Linkage

- Linking the metabolic and expression phenotypes
- Quantitative predictive models of cell physiology from gene expression
- Linkage by partial least squares
- Systems biology

Lecture 11: May 4

- Signaling and signal transduction pathways
- Measurements in signaling networks
- Integrated analysis of signal transduction networks

Lecture 12: May 18

- Putting it all together
 - Project presentations
- Lecture 1: Introduction and General Overview

HOMEWORKS

There will be five Problem Sets on the methodologies and computational algorithms covered in the course, as follows:

Problem Set – 1: Material of Lectures 1, 2

Problem Set – 2: Material of Lecture 3

Problem Set – 3: Material of Lecture 4

Problem Set – 4: Material of Lecture 5

Problem Set – 5: Material of Lecture 6

Problem Set – 6: Material of Lectures 7,8

Problem Set – 7: Material of Lectures 9,10

PROJECTS

The students, in groups of 2-3, will carry out a project on a course-related subject of their own choosing, or from a list of suggested topics. The groups must be formed topics selected by April 6, 2004. An oral presentation of the project by the group members will take place on May 18, 2004, at which time the final report on the project will be also due.

GRADE

There will be no mid-term or final exams. The grade in the course will be based on the homeworks, the group project, and the oral presentation, with the following weights:

Homeworks (40 %); Written project report (35 %); Oral presentation (25 %)

CLASS NOTES and REFERENCES

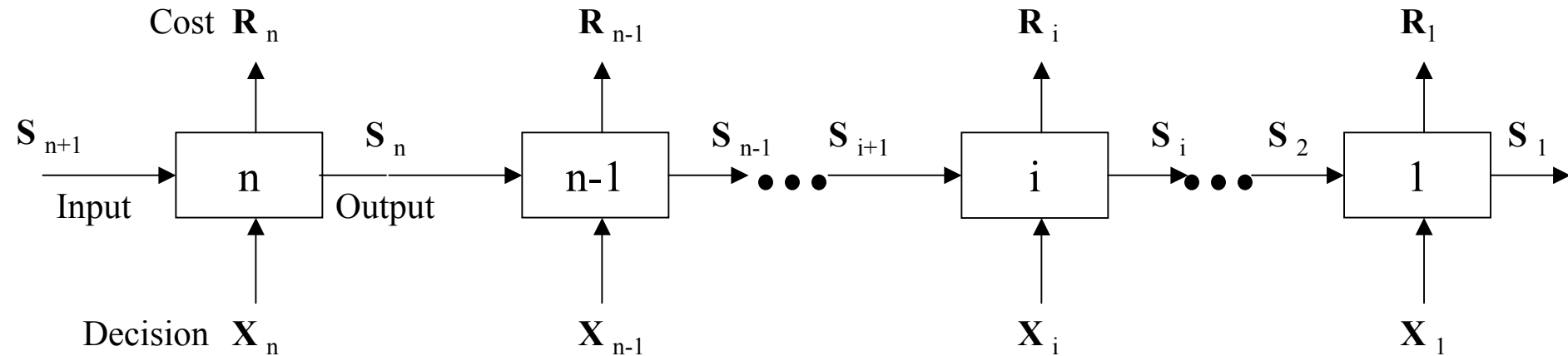
Lecture notes will be posted on the web. Additionally, the course will draw from a series of published papers and the following list of books, which are recommended as references:

References

- *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, D. Gusfield, Cambridge University Press, ISBN: 0521585198
- *Fundamental concepts of bioinformatics*, D.E. Krane and M. L. Raymer, ISBN: 0-8053-4633-3
- *Introduction to probability*, D.P. Bertsekas and J.N. Tsitsiklis, ISBN:1-886529-40-X
- *Genetics, a Molecular Approach*, T.A.Brown, Chapman & Hall, ISBN: 0412447304.
- *Introduction to Computational Molecular Biology*, J.Setubal and J.Meidanis, PWS Publishing Company, ISBN: 0534952623.
- *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, A.D.Baxevanis and B.F.F.Ouellette, Wiley-Interscience, ISBN: 0471191965.
- *Bioinformatics: The Machine Learning Approach*, P. Baldi and S. Brunal, MIT Press, ISBN: 0-262-02442-X
- *Introduction to Computational Biology: Maps, Sequences, Genomes*, M.S.Waterman, Chapman & Hall, ISBN: 0412993910.
- *Biological Sequence Analysis: Probabilistic Models of proteins and Nucleic Acids*, R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Cambridge University Press, ISBN: 0-521-62041
- *Bioinformatics: Methods and Protocols*, S. Misener and S.A. Krawetz (editors), Humana Press, ISBN: 0-89603-732-0
- *Bioinformatics Basics: Applications in Biological Science and Medicine*, H.H. Rashidi and L.K. Buehler, CRC Press, ISBN: 0-8493-2375-4
- *Introduction to Protein Structure?*, C.Branden and J.Tooze, Garland Publishing Inc., ISBN: 0815302703.
- *Molecular Biotechnology: Principles and Applications of Recombinant DNA*, B.R.Glick and J.JPasternak, ASM Press, ISBN: 1555811361.
- *Introduction to Proteins and Protein Engineering*, B.Robson and J.Garnier, Elsevier Science Publishers, ISBN: 0444810471.
- *Computational Molecular Biology: An algorithmic approach*, Pavel Pevzner, MIT Press, ISBN: 0262161974
- *Metabolic Engineering: Principles and Methodologies*, G. Stephanopoulos, A. Aristidou and J. Nielsen, Academic Press, ISBN: 0-12-666260-6

Additional references of web-based material will be distributed to the students during the course.

DYNAMIC PROGRAMMING



$$S_i = t_i(S_{i+1}, X_i)$$

$$R_i = r_i(S_{i+1}, X_i)$$

Overall cost function: $f(R_1, R_2, \dots, R_n)$

- *Separability*: $f = \sum R_i$
- *Monotonicity*

Task: Optimize f

Overall Optimization solved by

- breaking down the problem to the sum of its parts using the recurrence relation:

$$f_i^*(S_{i+1}) = \underset{X_i}{\text{opt}} [R_i(S_{i+1}, X_i) + f_{i-1}^*(S_i)] \quad \text{where } * \text{ denotes optimum value}$$

- keeping track of the optimum decision at each stage, X_i^*

Solve a problem by using already computed solutions for smaller, but similar problems

SEQUENCE DRIVEN PROBLEMS

- DNA Sequencing
 - Identification of similar prefixes and suffixes for given sequences
- Comparing two sequences: Homology searches in databases
 - Identification of similar substrings
- Multiple alignment of sequences: Homologies in related genes/proteins
 - Identification of short substrings/motifs
- Determination of introns and exons
- Construction of Phylogenetic trees
- Structure prediction from sequences
 - RNA secondary structure prediction
 - Protein folding

Application of DP to sequence problems

COMPARISON OF SEQUENCES

Input, S_{i+1}

- Alignment of preceding prefix

Output, S_i

- Alignment of preceding prefix and current element

Decision, X_i

- Alignment with an element in the sequence
- Alignment with a gap

Cost Function, R_i

- reward for exact match
- penalty for mismatch
- penalty for matching with a gap or a break in the sequence

COMPARISON OF SEQUENCES

$s = \text{AAAC}$ $t = \text{AGC}$

DP definition of the problem: Recursive

For the alignment of the $s[1:i]$ and $t[1:j]$, we have:

Decision, X

- Align $s[1:i]$ with $t[1:j-1]$, and match a space with $t[j]$
- Align $s[1:i-1]$ with $t[1:j-1]$, and match $s[i]$ with $t[j]$
- Align $s[1:i-1]$ with $t[1:j]$, and match $s[i]$ with a space

Cost, R

- reward for exact match, pm
- penalty for mismatch, mm
- penalty for matching with a gap or a break in the sequence, sp

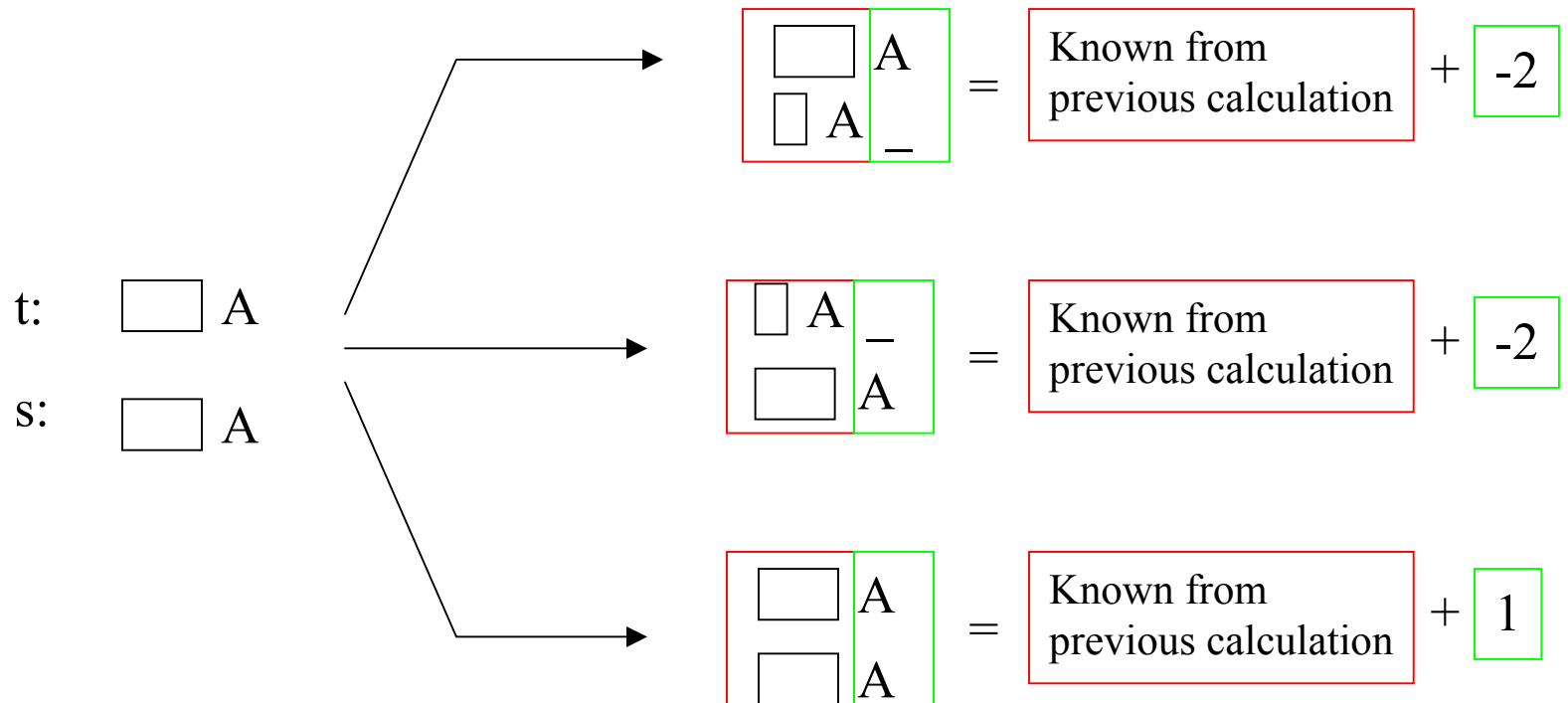
$$f^*(s[1:i], t[1:j]) = \max \begin{cases} f^*(s[1:i], t[1:j-1]) - sp \\ f^*(s[1:i-1], t[1:j-1]) + pm, \text{ if } s[i] = t[j] \\ f^*(s[1:i-1], t[1:j-1]) - mm, \text{ if } s[i] \neq t[j] \\ f^*(s[1:i-1], t[1:j]) - sp \end{cases}$$

ELUCIDATION OF ALGORITHM

s = _AAAC t = _AGC

Consider 2nd element of s and t, and the decisions possible:

Basic calculation approach



sp = -2, mm = -1, pm = 1

Sequences can be expanded at will by adding more spaces

ELUCIDATION OF ALGORITHM

Matrix based calculation

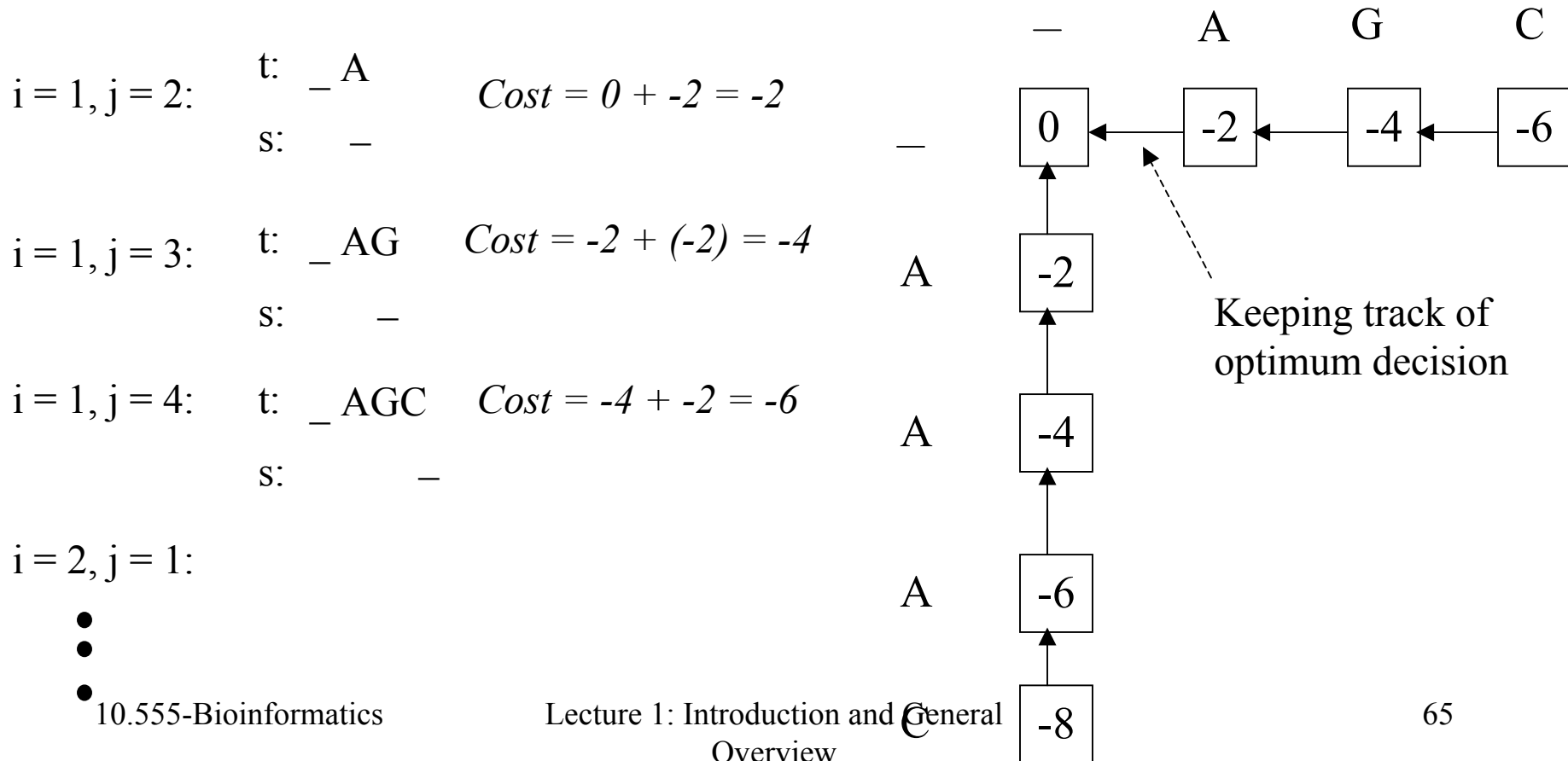
s = _AAAC

t = _AGC

Step 1: Initialization of s[1] and t[1]

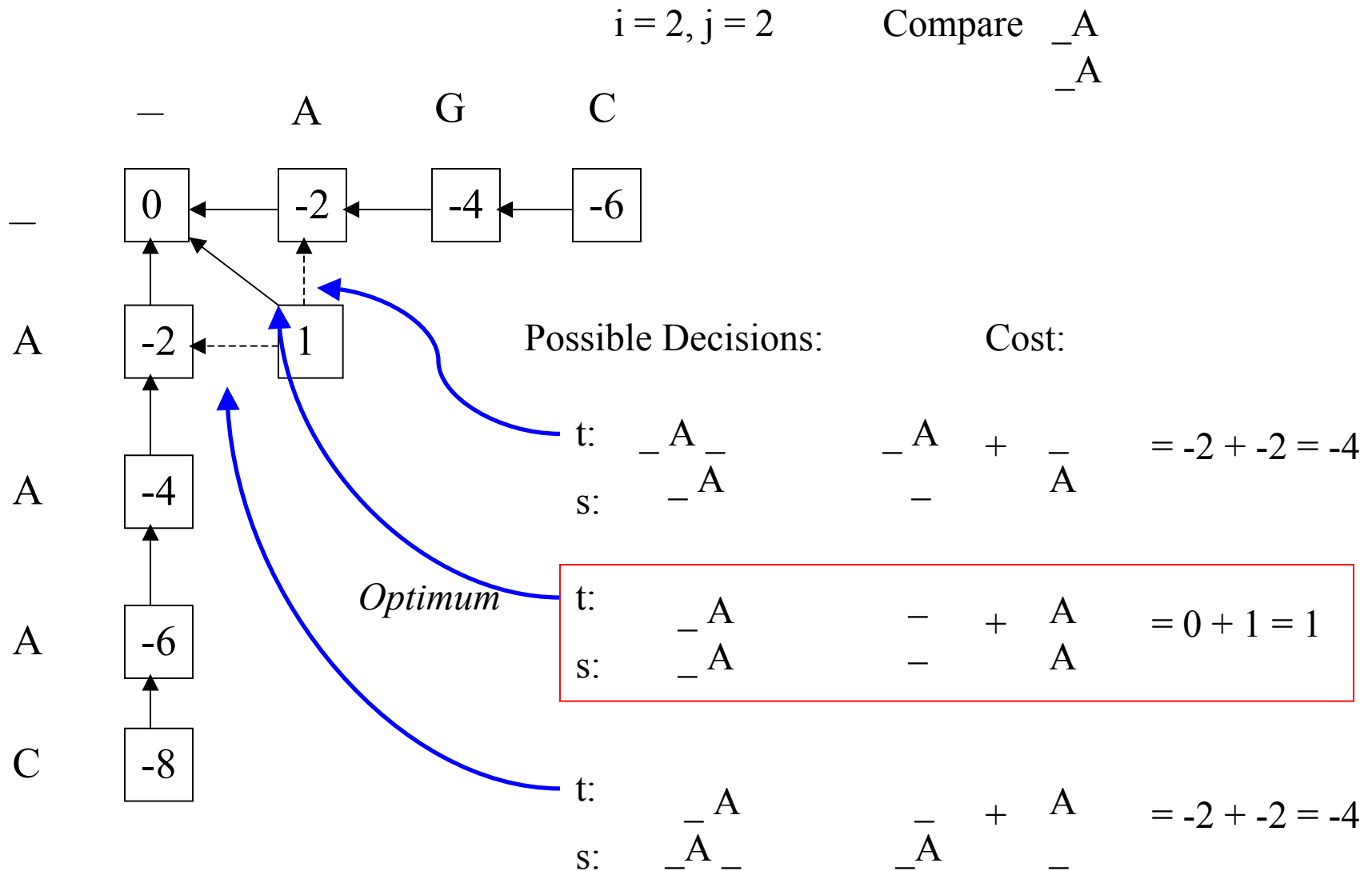
Cost of alignment of _ and _ = 0

Step 2: Initialization of first row and column



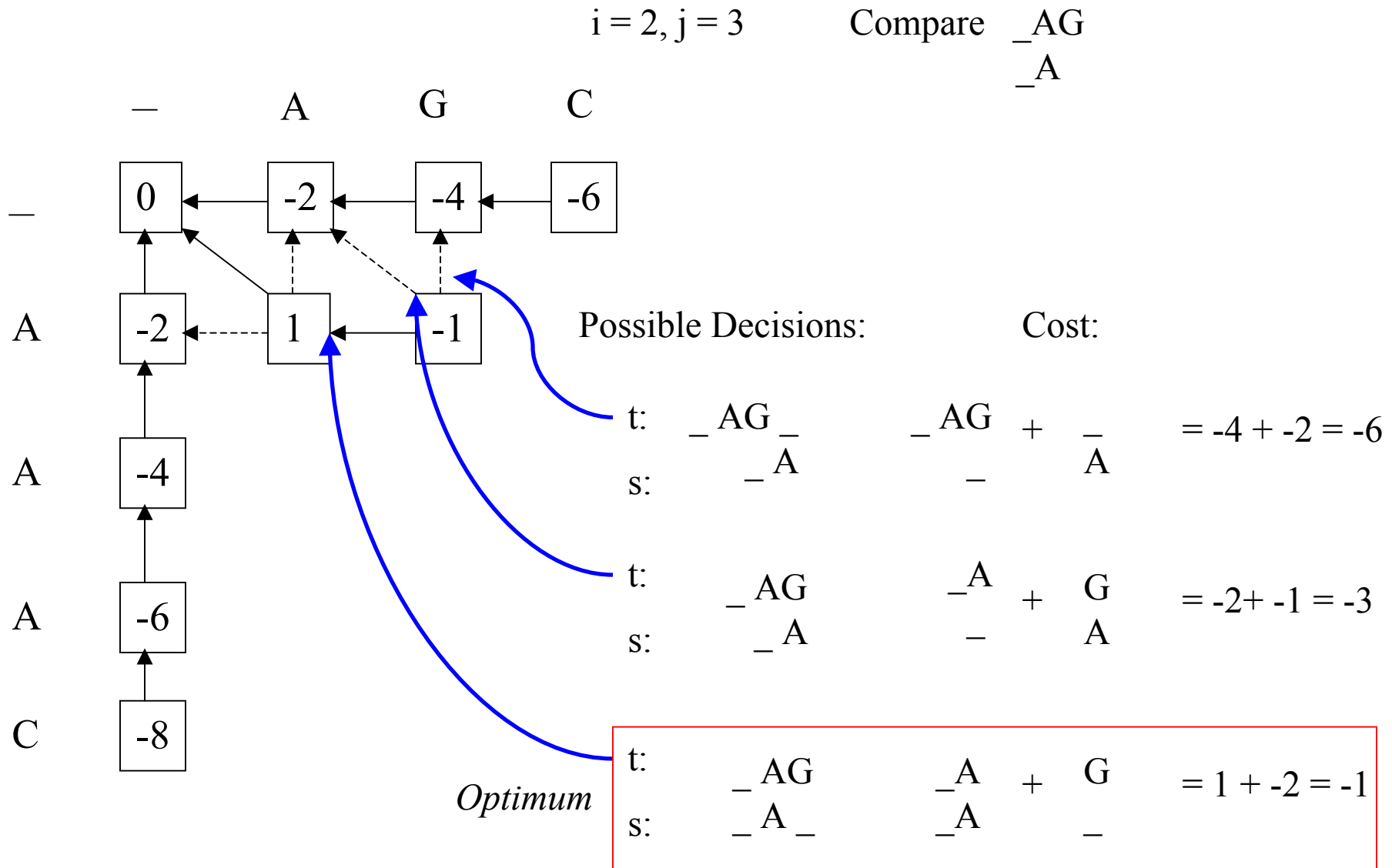
ELUCIDATION OF ALGORITHM

Step 3(0): Apply recursive optimization



ELUCIDATION OF ALGORITHM

Step 3(1): Apply recursive optimization

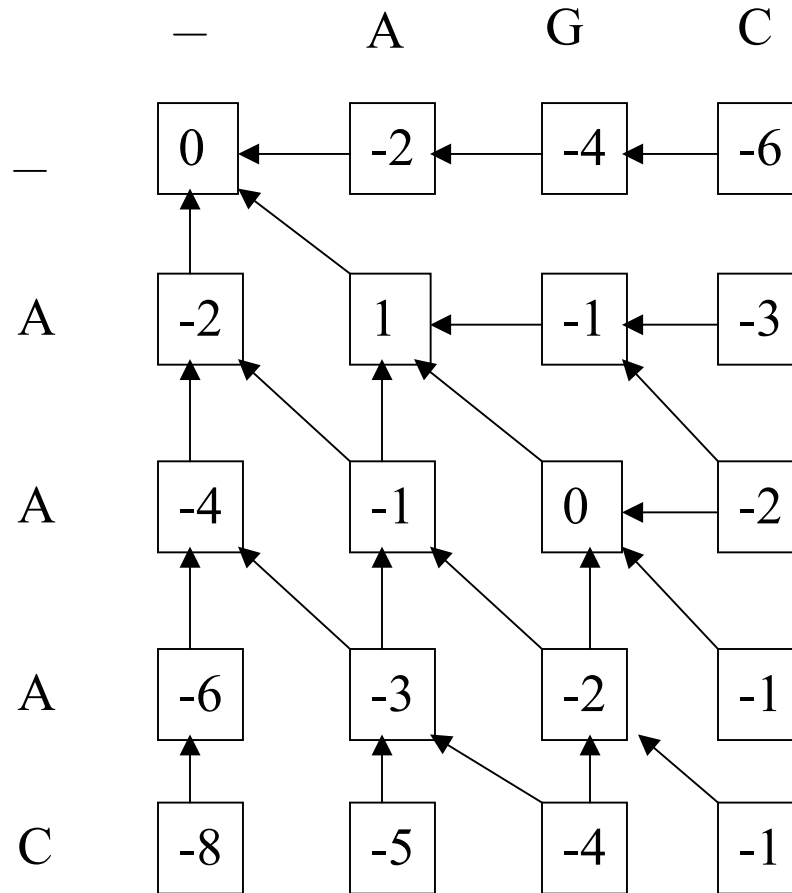


ALGORITHM: SUMMARY

s = AAAC

t = AGC

Bidimensional array for computing optimal alignments

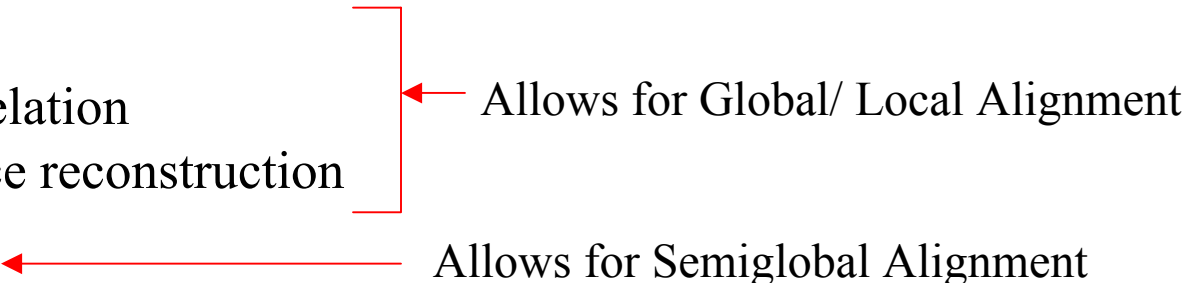


Arrows keep track of optimal decision

Reconstruction of optimal alignment done by retracing the arrows from the last corner

MODIFICATIONS TO THE BASIC ALGORITHM

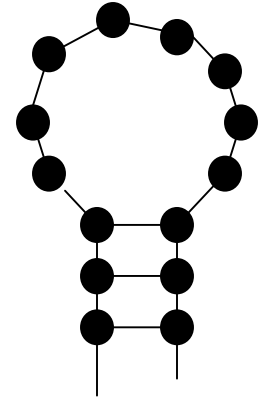
Parameters (allow for modification of the basic algorithm):

- Initialization
 - Recurrence scoring relation
 - Start point in sequence reconstruction
 - Scoring of spaces
 - Multiple sequence alignment
- 
- Diagram illustrating parameter groupings:
- A red bracket groups the first three items: Initialization, Recurrence scoring relation, and Start point in sequence reconstruction. A red arrow points from this bracket to the text "Allows for Global/ Local Alignment".
 - A red arrow points from the item "Scoring of spaces" to the text "Allows for Semiglobal Alignment".

Other modifications:

- Cost function R can be a matrix, with mm and pm being vectors
- Gap penalty can be a function instead of a fixed value

RNA SECONDARY STRUCTURE PREDICTION



Cost function: Free Energy of structure

Dynamic programming applied under certain **assumptions**:

- Free energy of structure is the sum of its parts

Implies

- independent base pairs, or
 - consider adjacent base pairs, but ignore free energies of bases not part of any loop
-
- Ignore the prediction of knots in the structure
 - Only certain types of loops allowed

Under these assumptions, and with an appropriate cost function, a DP based algorithm can be used to compute the optimal structure