10.555
Bioinformatics: Principles, Methods and Applications

MIT, Spring term, 2003

Lecture 1:

Introduction to the course.
Definitions, perspectives, general overview.
Rudiments of dynamic programming

Course goals-requirements

- **Goals**
  - Not code writing
  - Not algorithm optimization
  - Appreciation-familiarity-expertise with available computational tools
  - Define important problems in bioinformatics-systems biology
  - Creativity in solving important biological-physiological problems
- **Requirements**
  - Advanced course
  - None formally
  - Applied probability- statistics most important background

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'.

Jay Bailey

Drivers of change (1)

- **Genomics: Full genome sequencing**
  - More than 100 species completed
  - *All* industrially important organisms will be sequenced within the next 1-2 years
  - Impact of Genomics to (Bio)ChE:
    - ***Generation of data about the cellular phenotype***
    - ***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, CO_2, methane-methanol
  - Environmental
  - Plants (resistance, oils, enrichment, polymers)

Drivers of Bioinformatics (4)

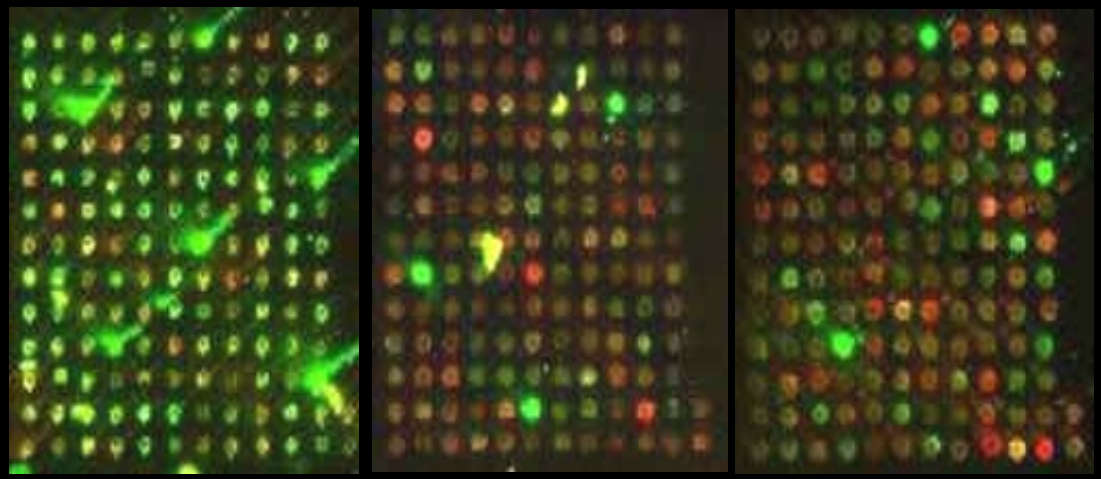
- Technologies probing the expression, proteomic and metabolic phenotype generating large volumes data
  - DNA microarrays *(gene chips)*
  - 2-D gel electrophoresis
  - Protein chips or HPLC methods combined with maldi/seldi tof spectrometry for protein characterization
  - Isotope label distributions

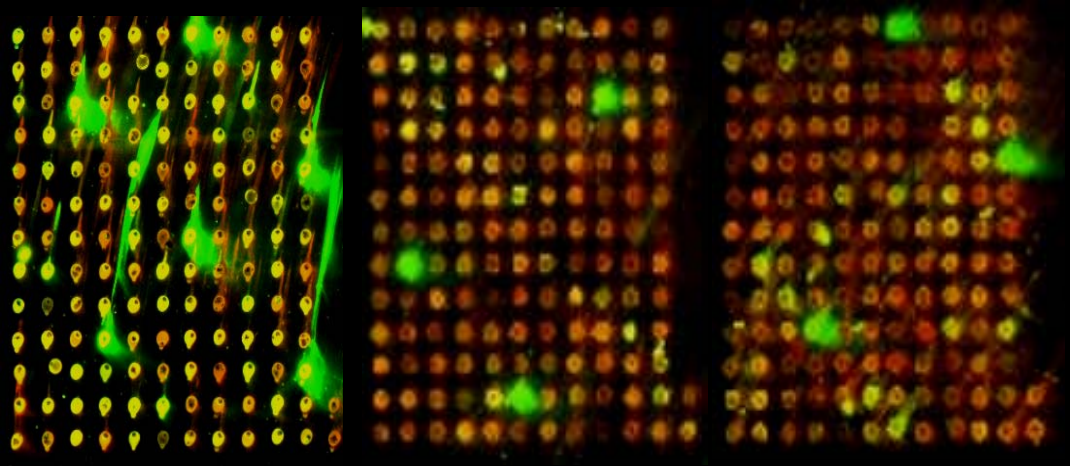
0.4% Pine-Sol
mid-exponential phase

Same culture
+3 hours

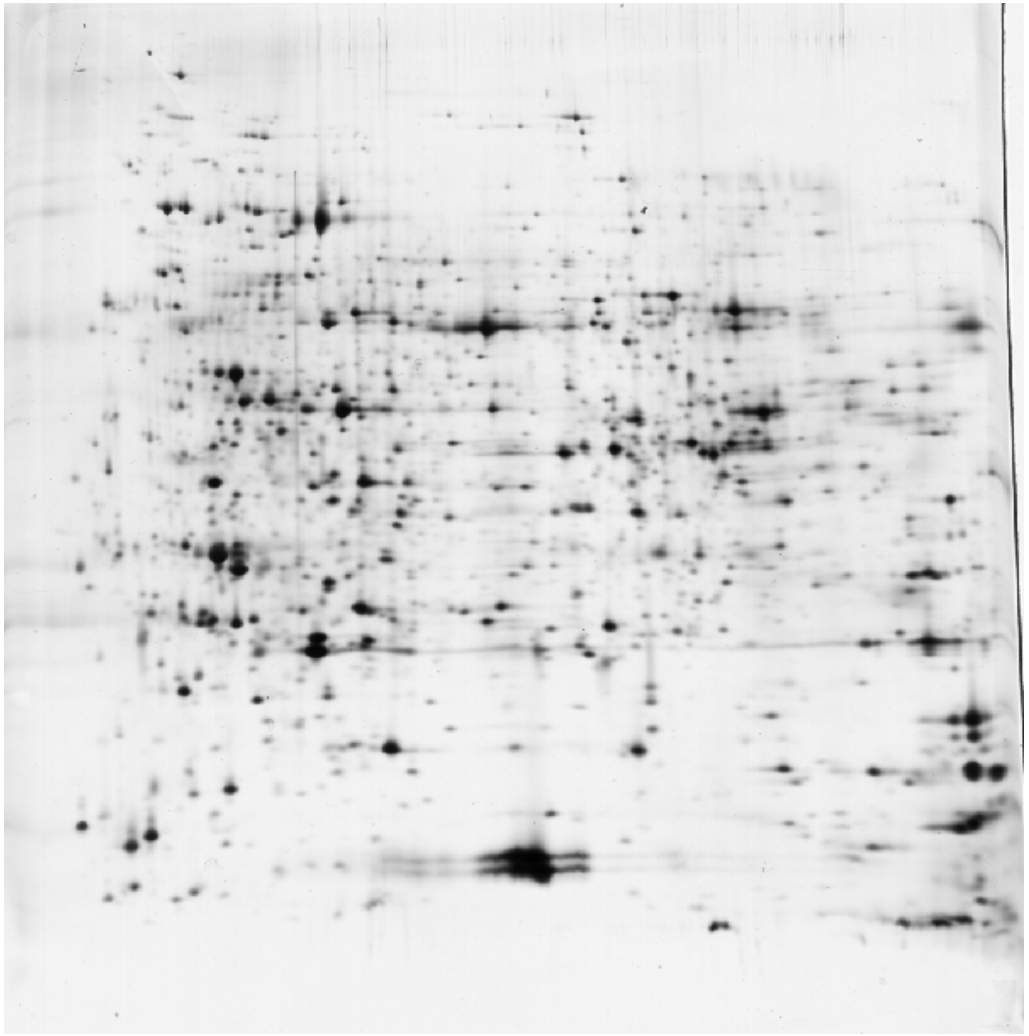
0.4% Pine-Sol
early stationary phase

Control
early stationary phase

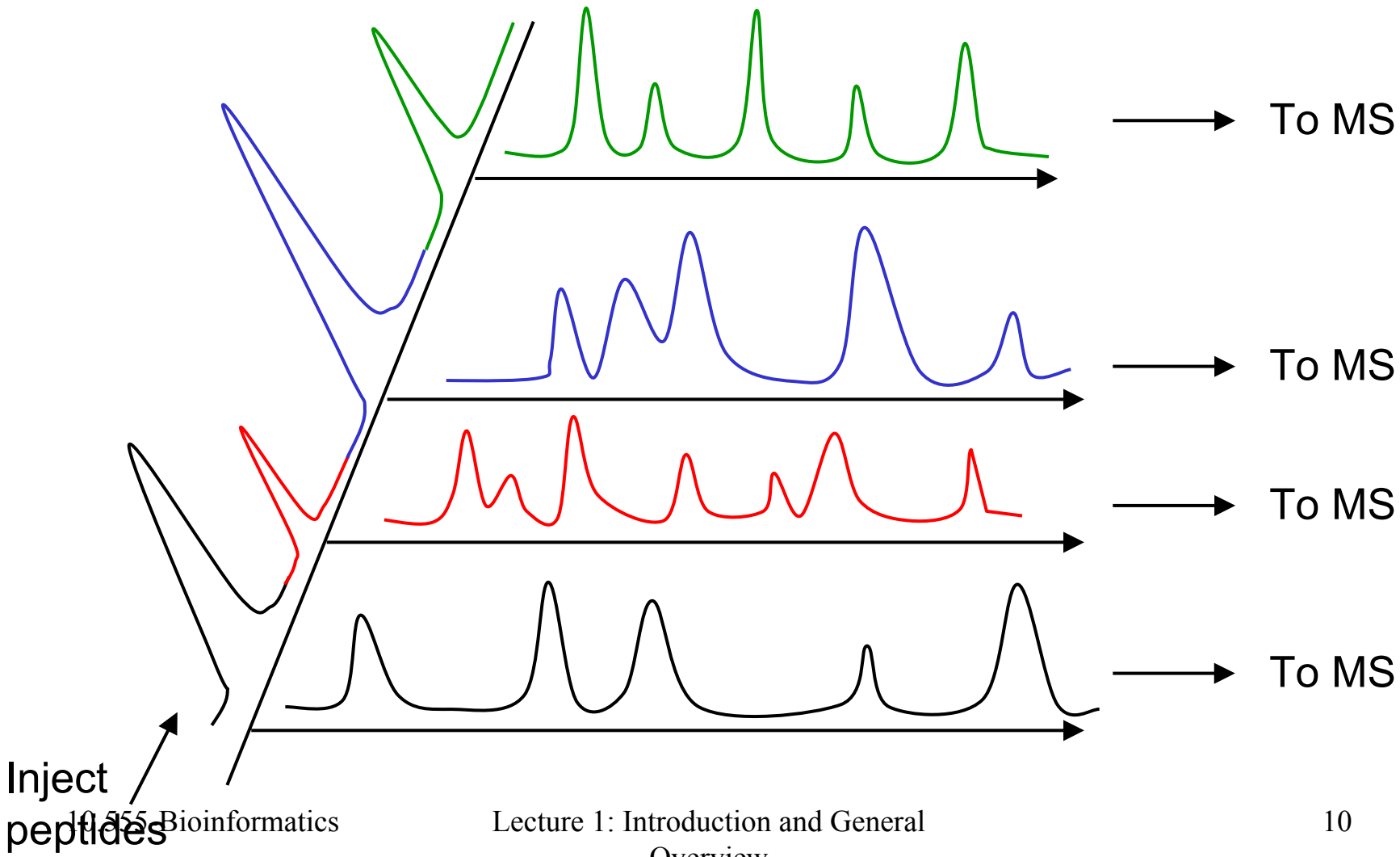




Example of 2-D gels

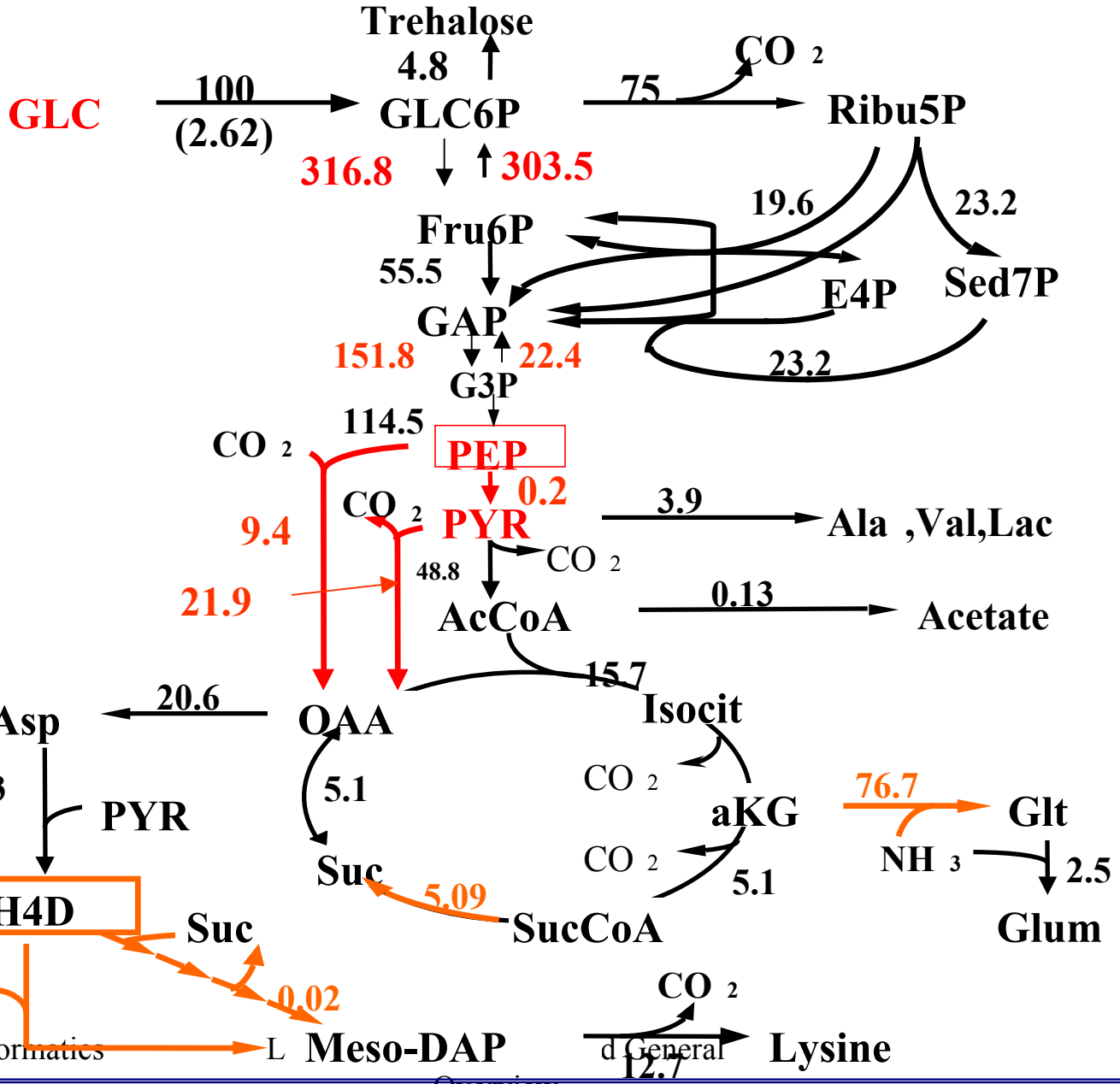


2-Dimensional Liquid Chromatography

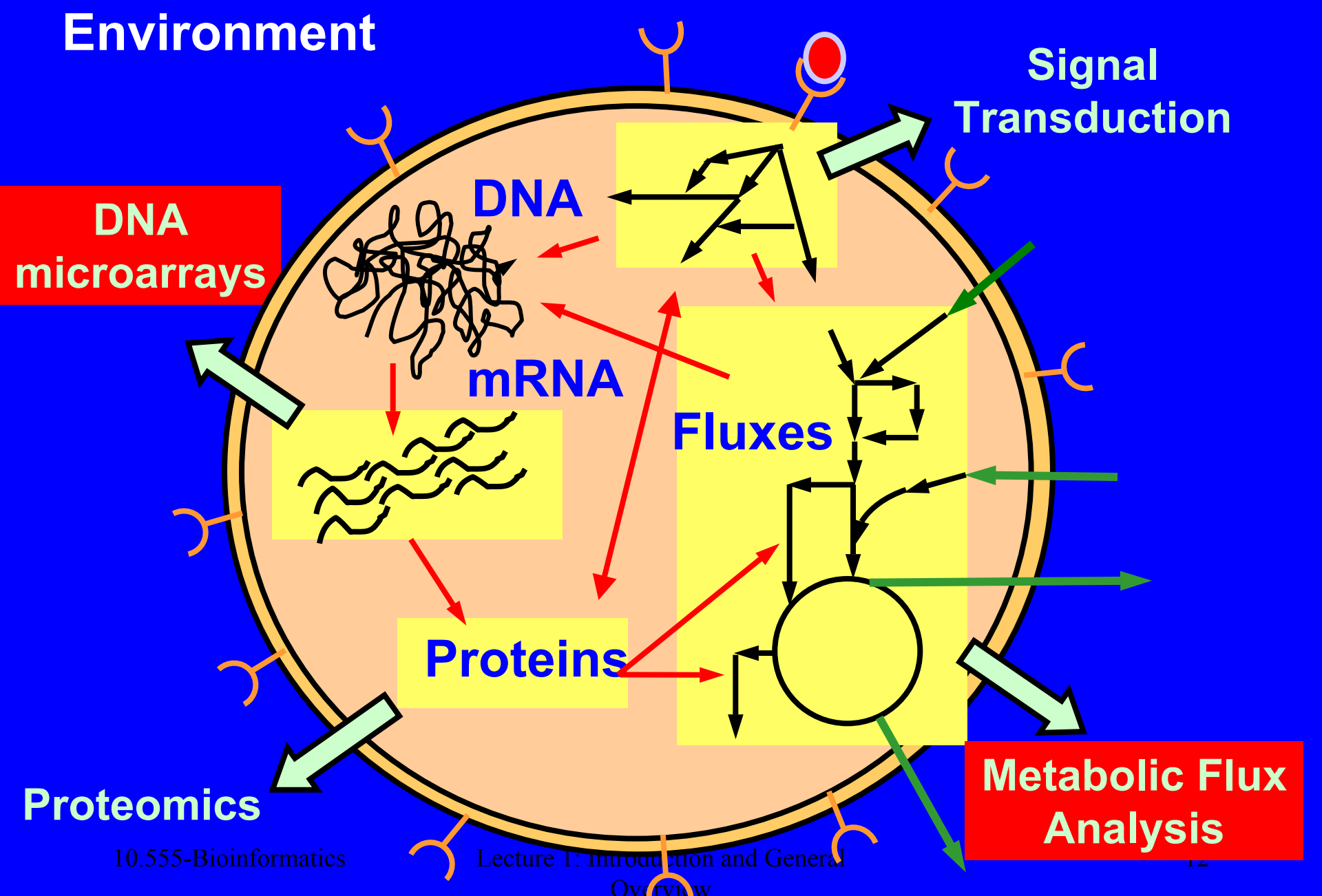


Fluxes estimated from isotopomer balancing

(locally optimal solution - value of objective: 0.021)



Probing cellular function



Central Goal

- **Technologies generating large volumes of data**
  - DNA microarrays *(gene chips)*
  - 2-gel electrophoresis
  - Protein chips combined with maldi/seldi tof spectrometry for protein characterization
  - 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***

Structured to address these driving forces.

It comprises three basic units:

- **Genomics: Sequence-driven problems**
- **Data-driven problems**
- **Functional Genomics and Issues of Systems Biology**

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)**

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

Sequence-driven problems

- **Central theme: *Sequence comparison***
  - **Metrics of comparison**
  - **Scoring matrices**
  - **Score evaluation**
  - **Comparison by *alignment vs.* comparison by *entries of bio-dictionaries (Teiresias)***
- **Emphasis**
  - **Computer Science**
  - **Applied probability**
  - **Statistics**

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

Problems of Systems Biology- Physiology- Functional Genomics

Central theme:

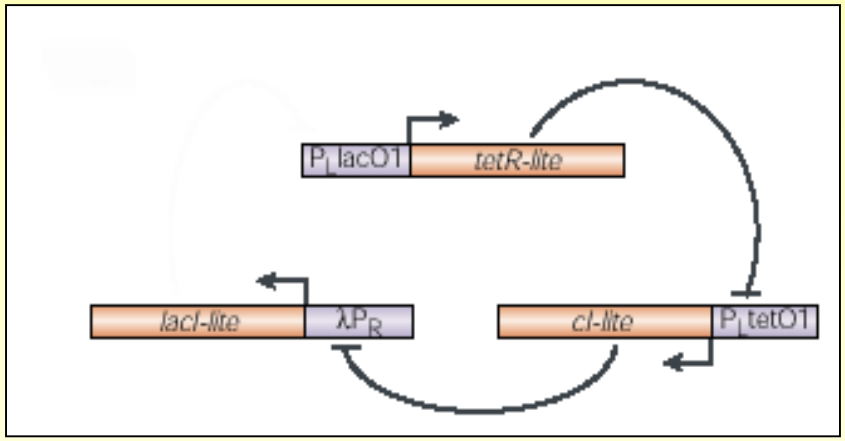
Interactions of cellular parts
Networks
Associations of data

Methods:

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

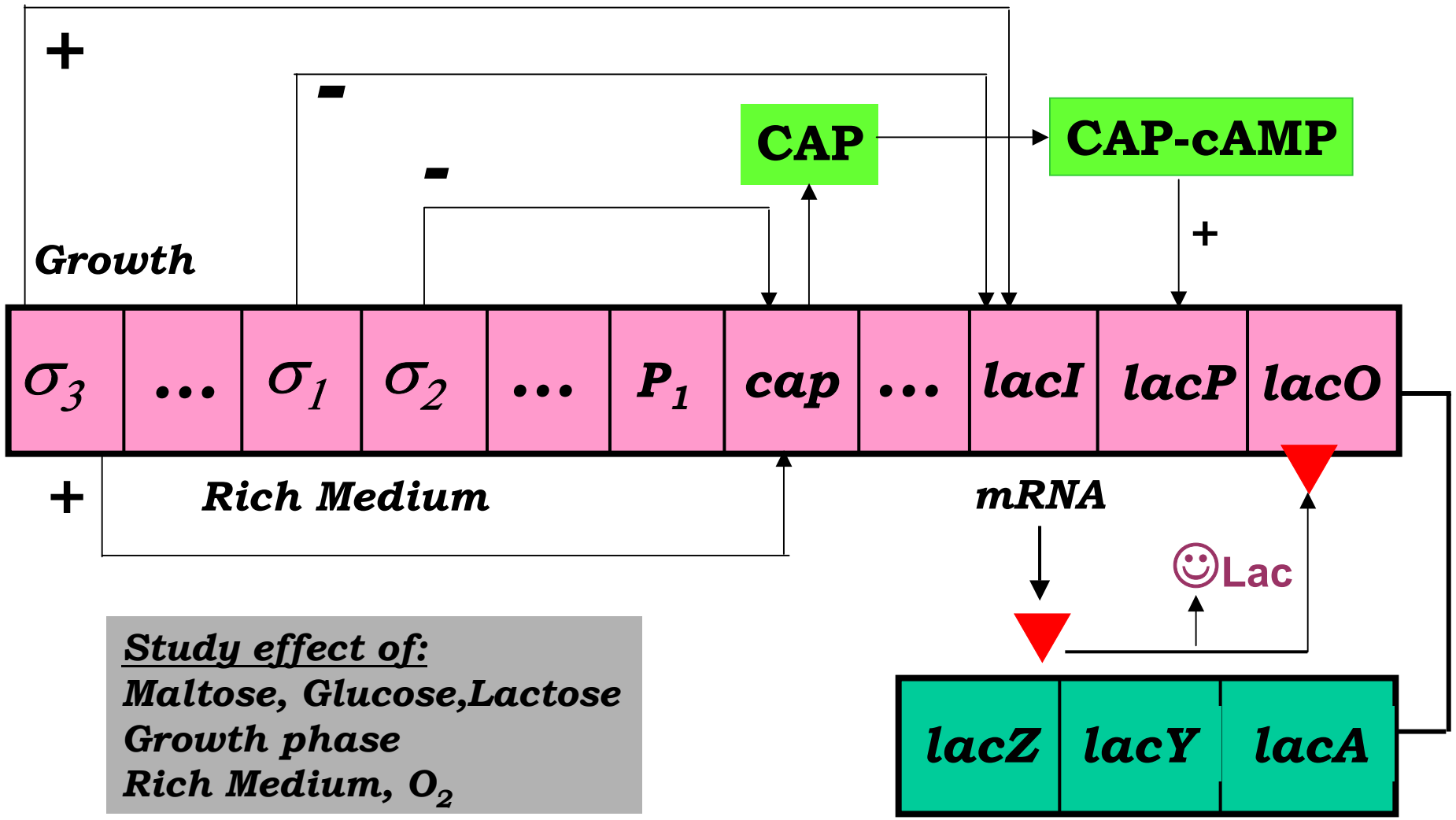
Problems of Systems Biology-Physiology-Functional Genomics

✸ Genetic Regulatory Networks



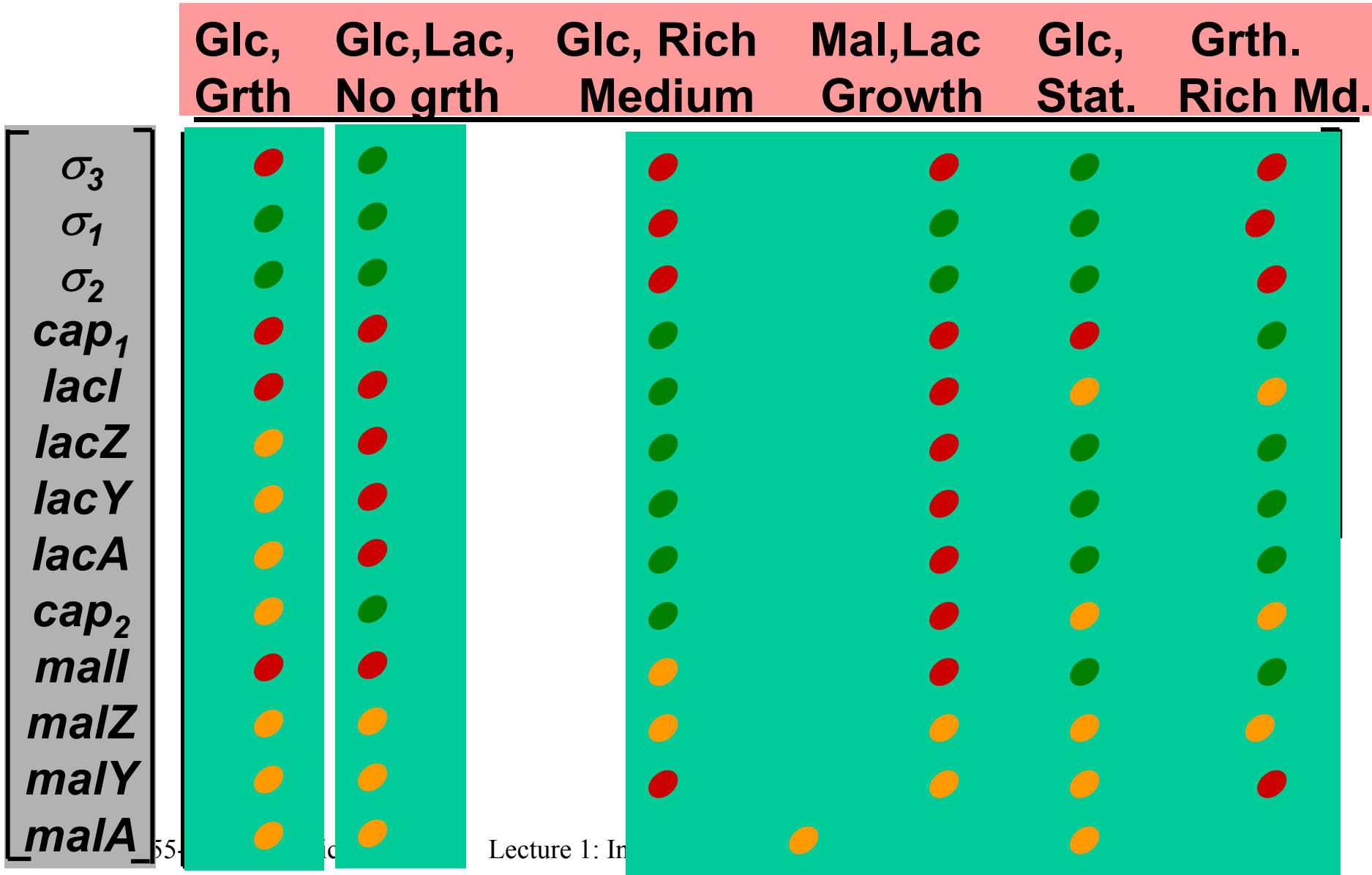
- Apply time-lagged *correlations* in transcriptional data
- Apply *pattern discovery* in upstream sequence regions of co-expressed genes

Illustrative example: Modified lac operon



Study effect of:
Maltose, Glucose,Lactose
Growth phase
Rich Medium, O_2

Expected profiles in differential gene expression experiments



Questions

1. Identify genes that exhibit significant variation in expression
   - Use *Mean Hypothesis Testing* (MHT)

2. *Cluster* genes in groups with similar expression characteristics
   - Apply *Clustering methods*

3. Extract structural information on genetic controls
   - Combinatorial pattern discovery and *ad hoc* methods

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	...

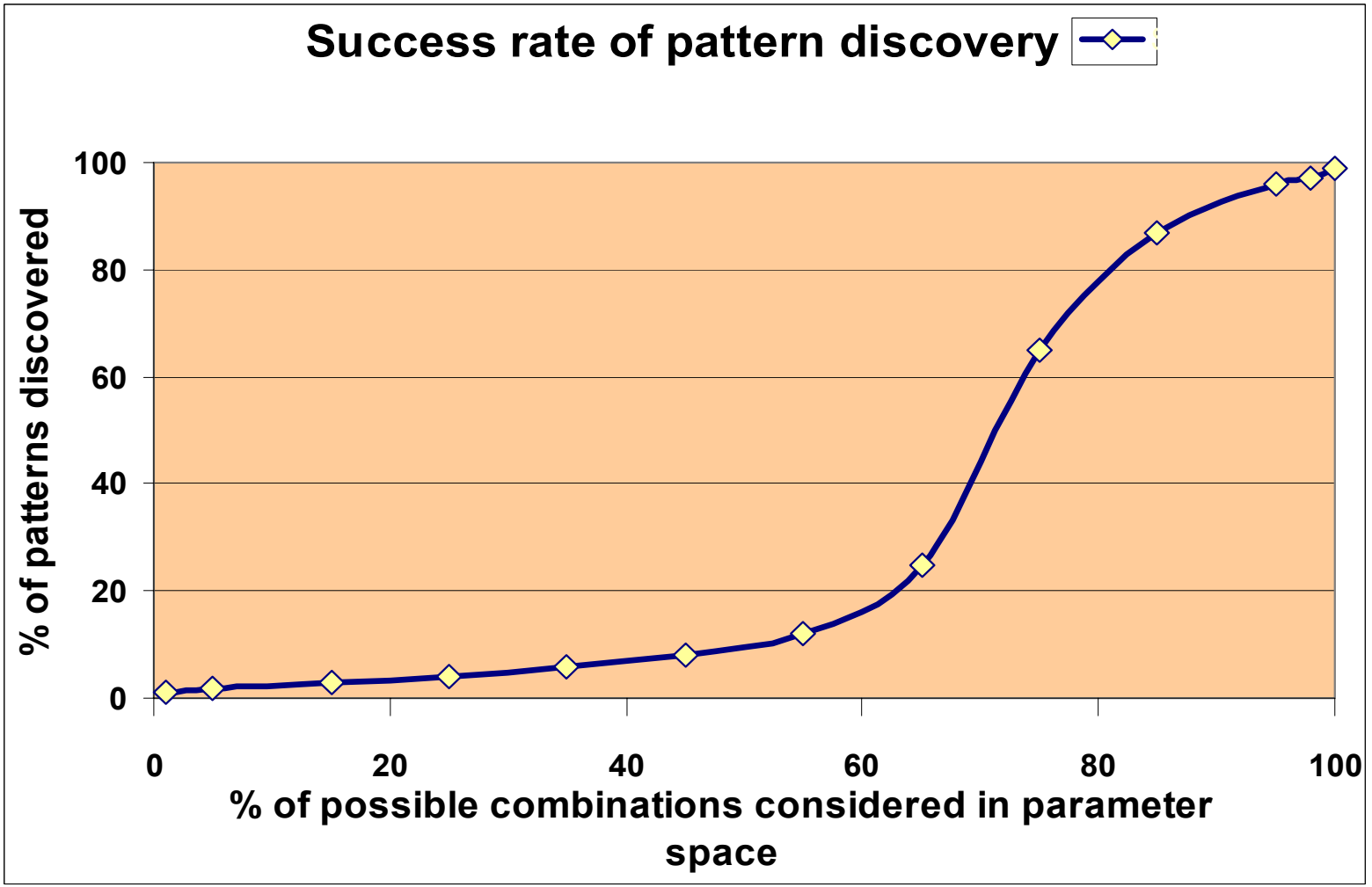
Look for most generic *patterns*:

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



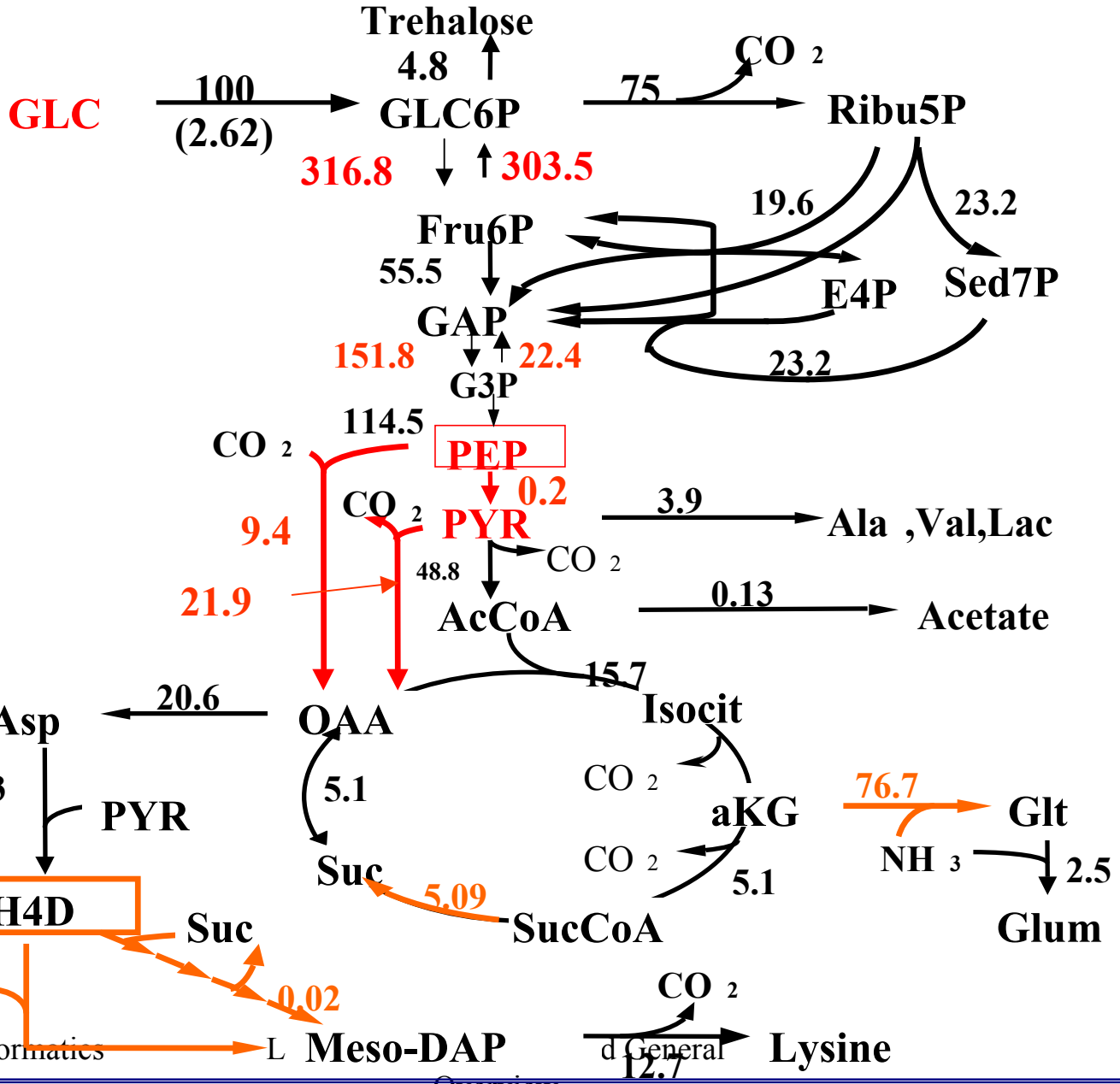
Interesting *problems (cont'd)*

- **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**

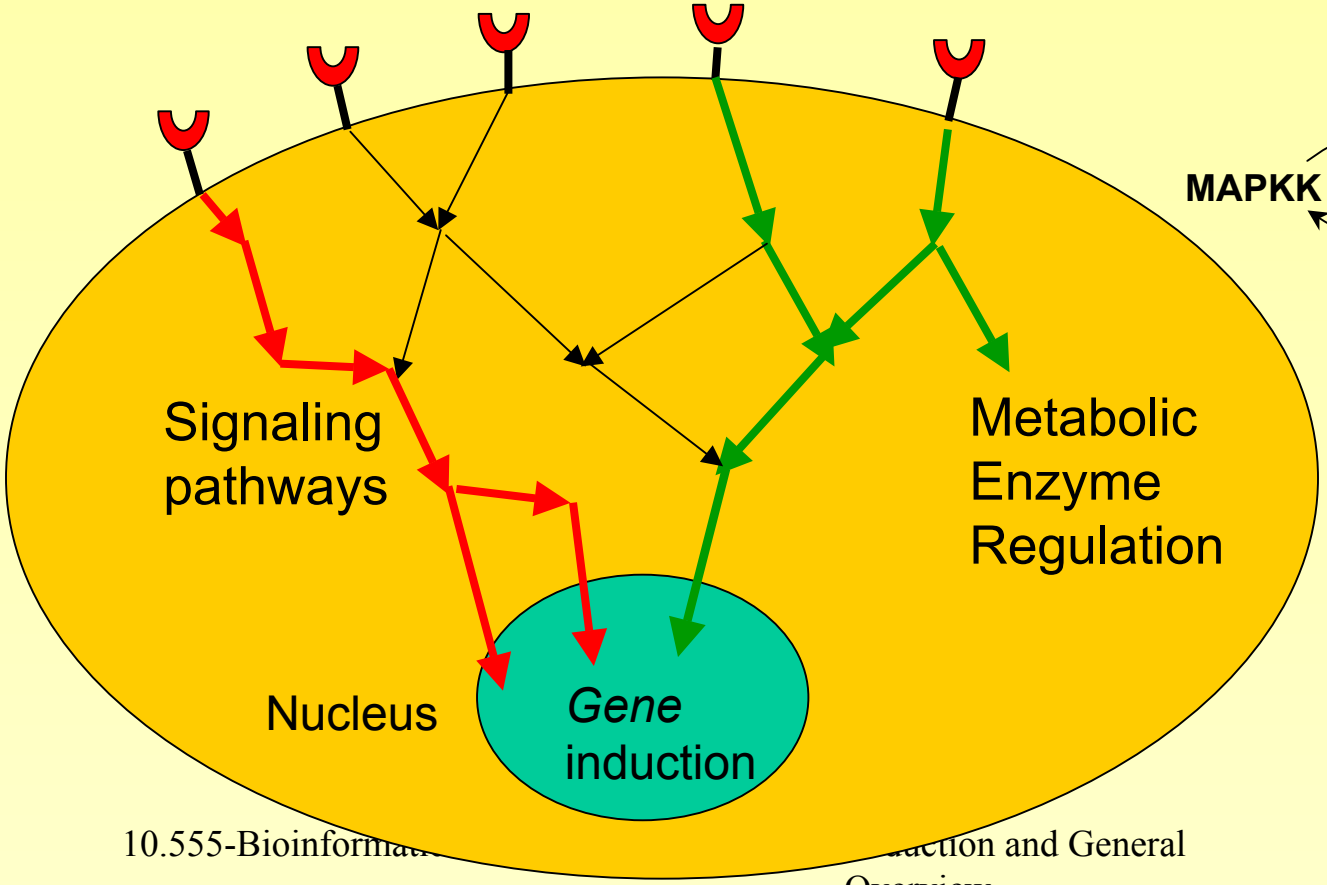
Fluxes estimated from isotopomer balancing

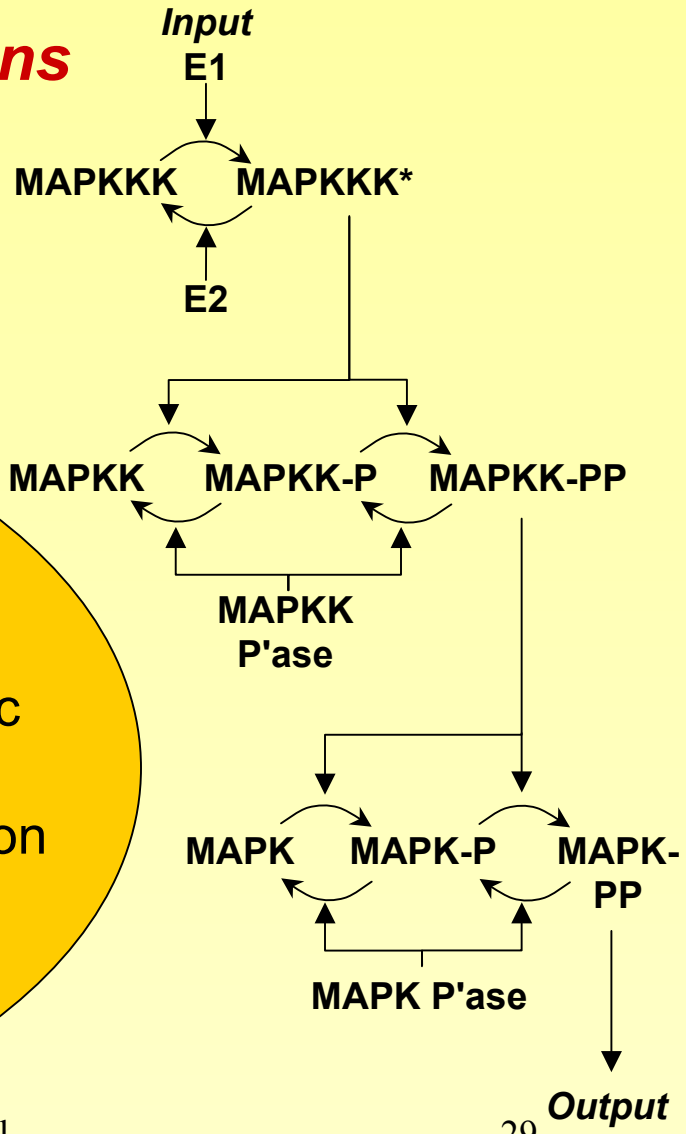
(locally optimal solution - value of objective: 0.021)



Problems of Systems Biology-Physiology-Functional Genomics

- **Analysis of structure and interactions of signal transduction pathways**



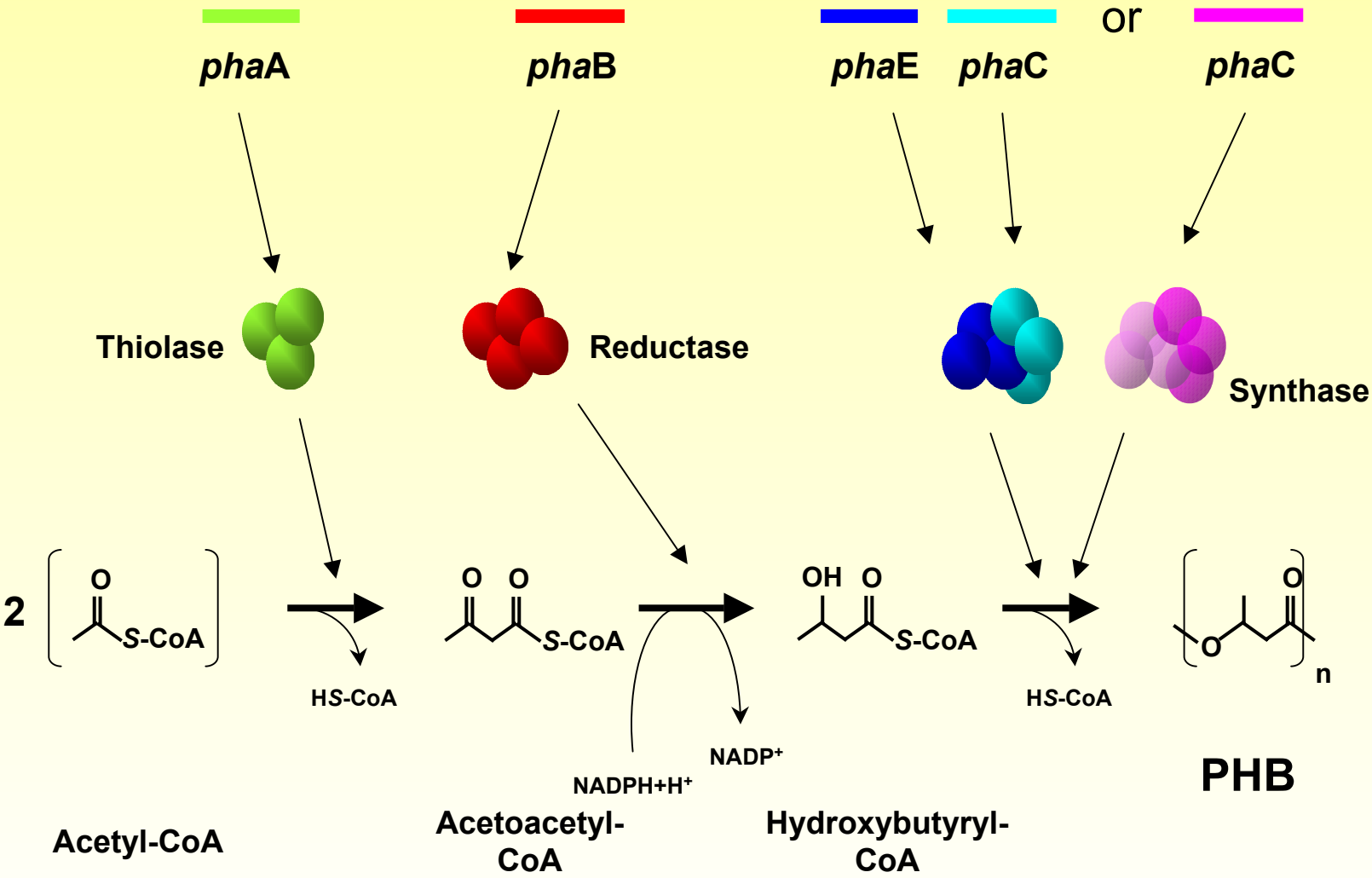


Problems of Systems Biology-Physiology-Functional Genomics

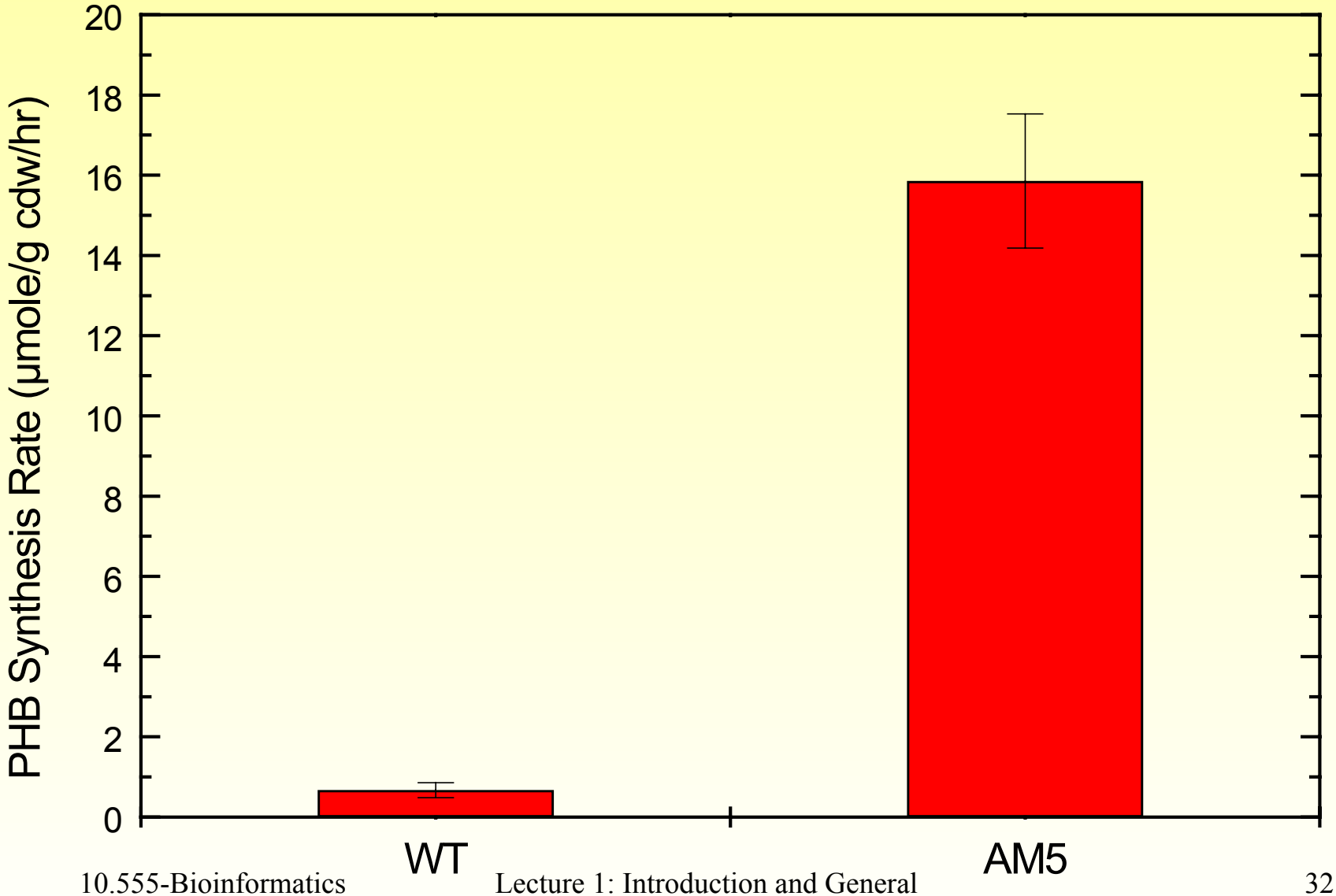
- **Analysis of Metabolic Networks**
  - Determine fluxes
  - Study flux control (metabolic engineering)
  - Distribution of kinetic control in more than one reaction steps (MCA)

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

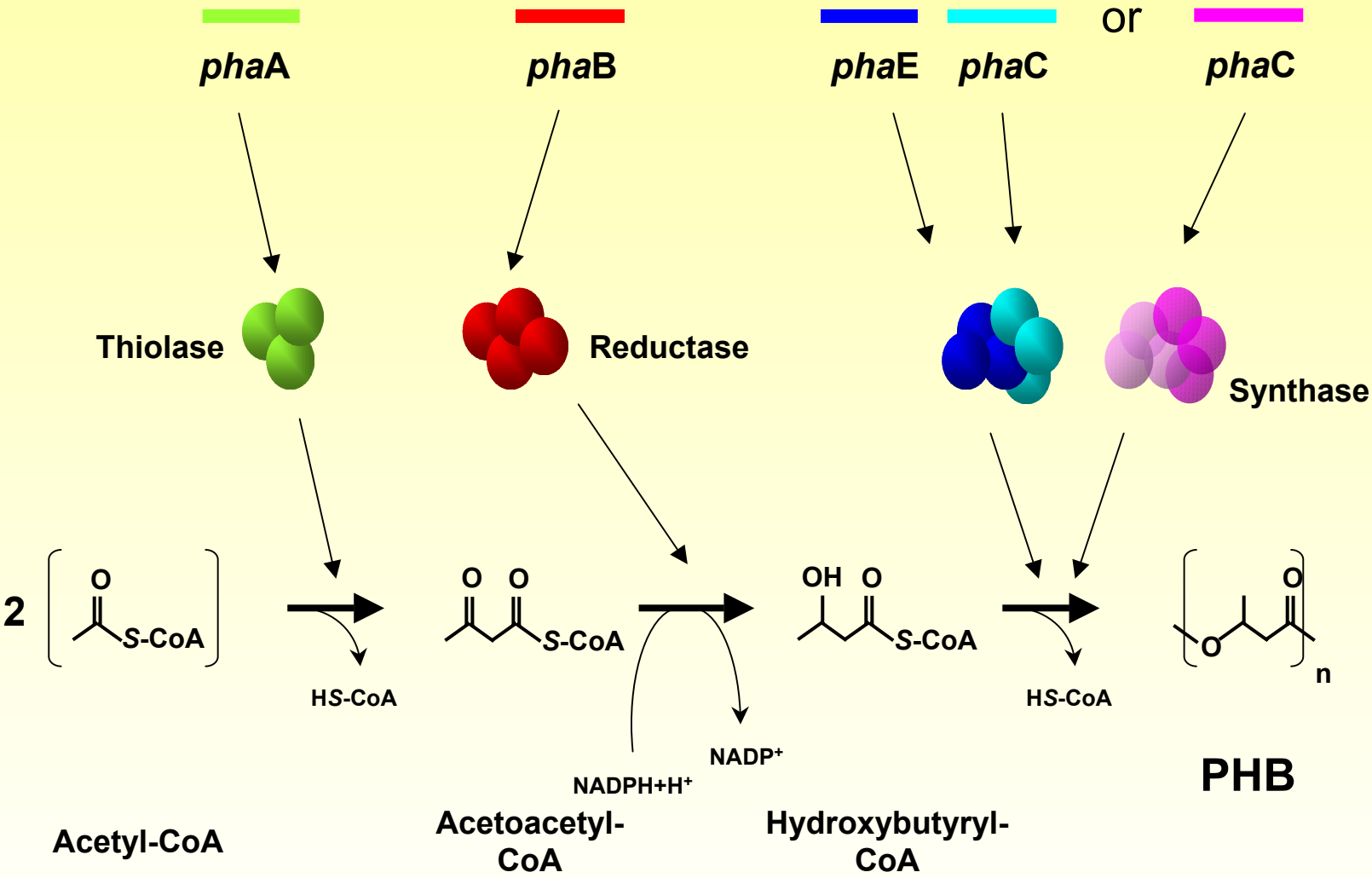
PHA/PHB biosynthesis genes



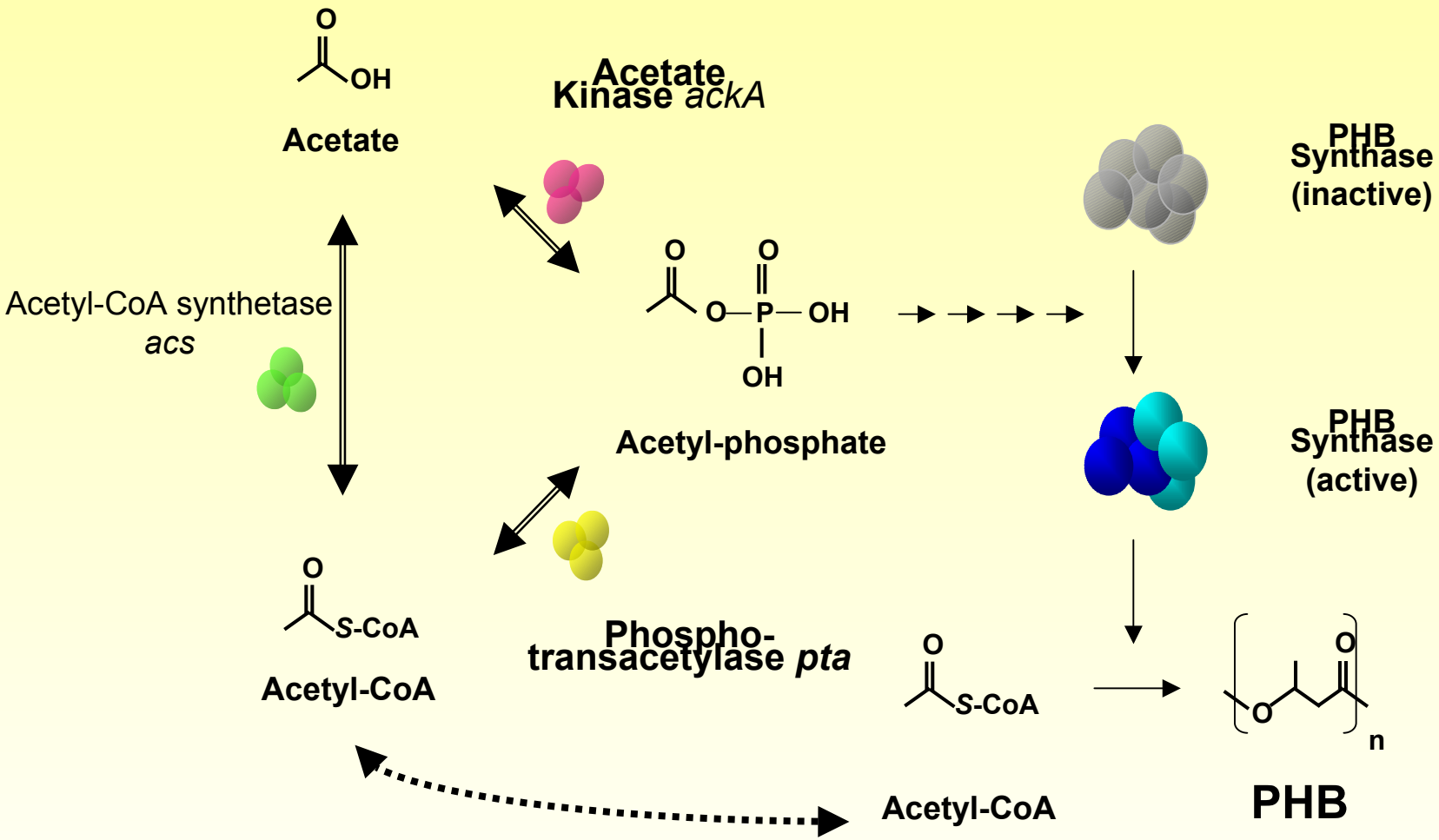
Drug Discovery (cont'd)



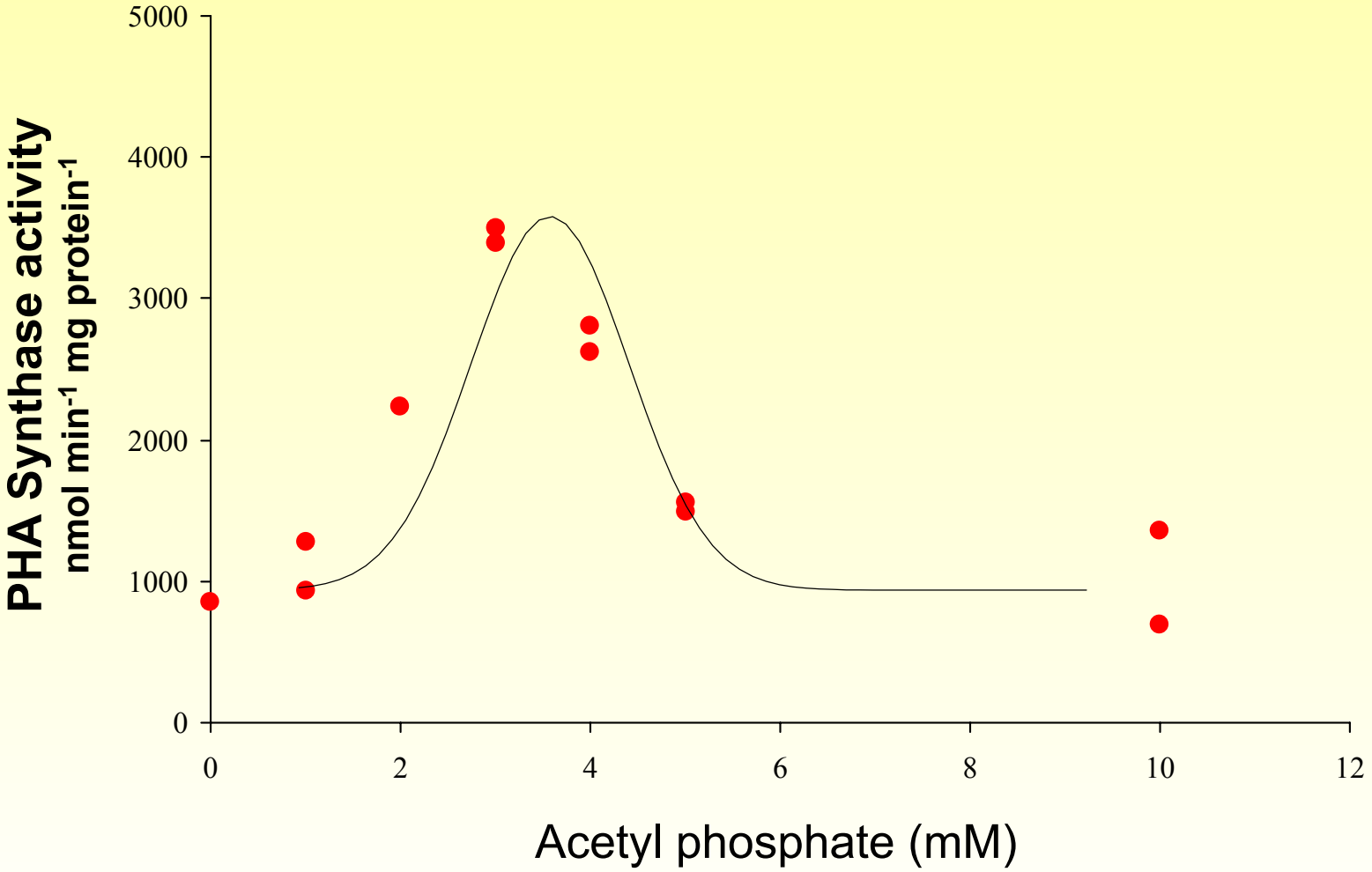
PHA/PHB biosynthesis genes



Regulation of PHA biosynthesis in *Synechocystis* sp.

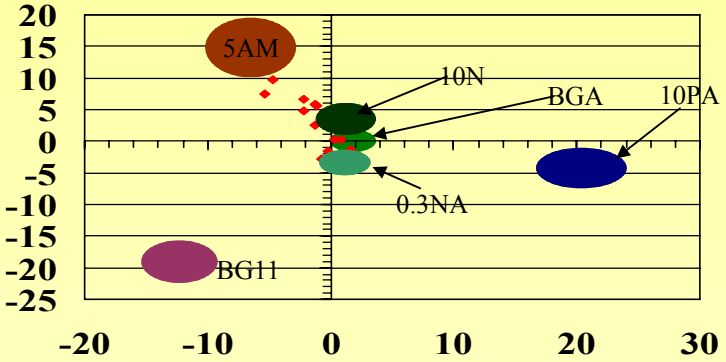


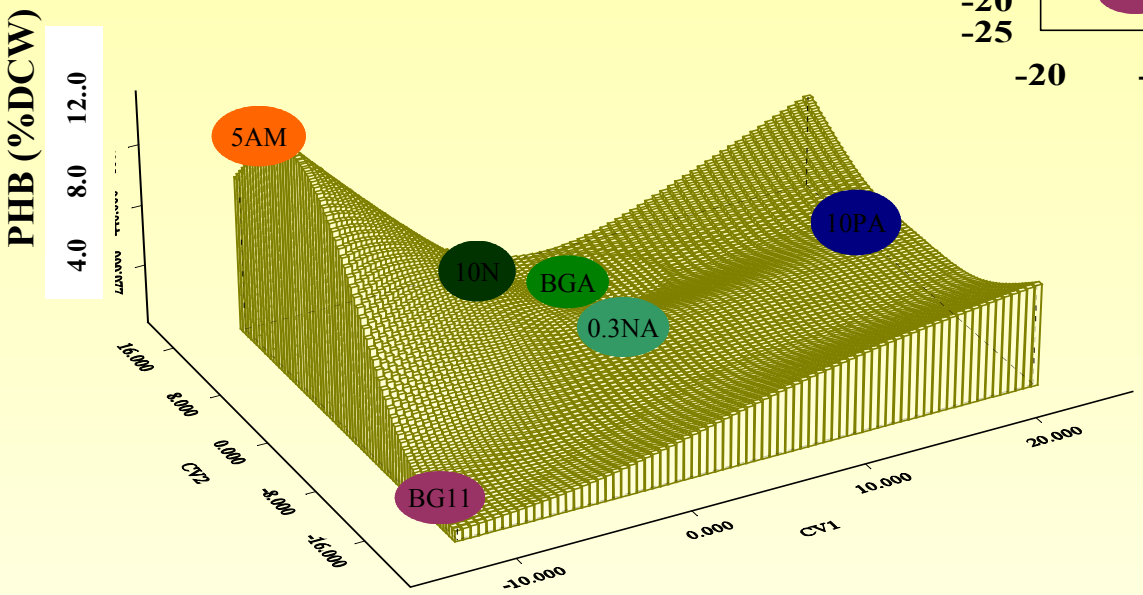
Acetyl phosphate dependent activation of the *Synechocystis* sp. PHA synthase



(A)

(B)





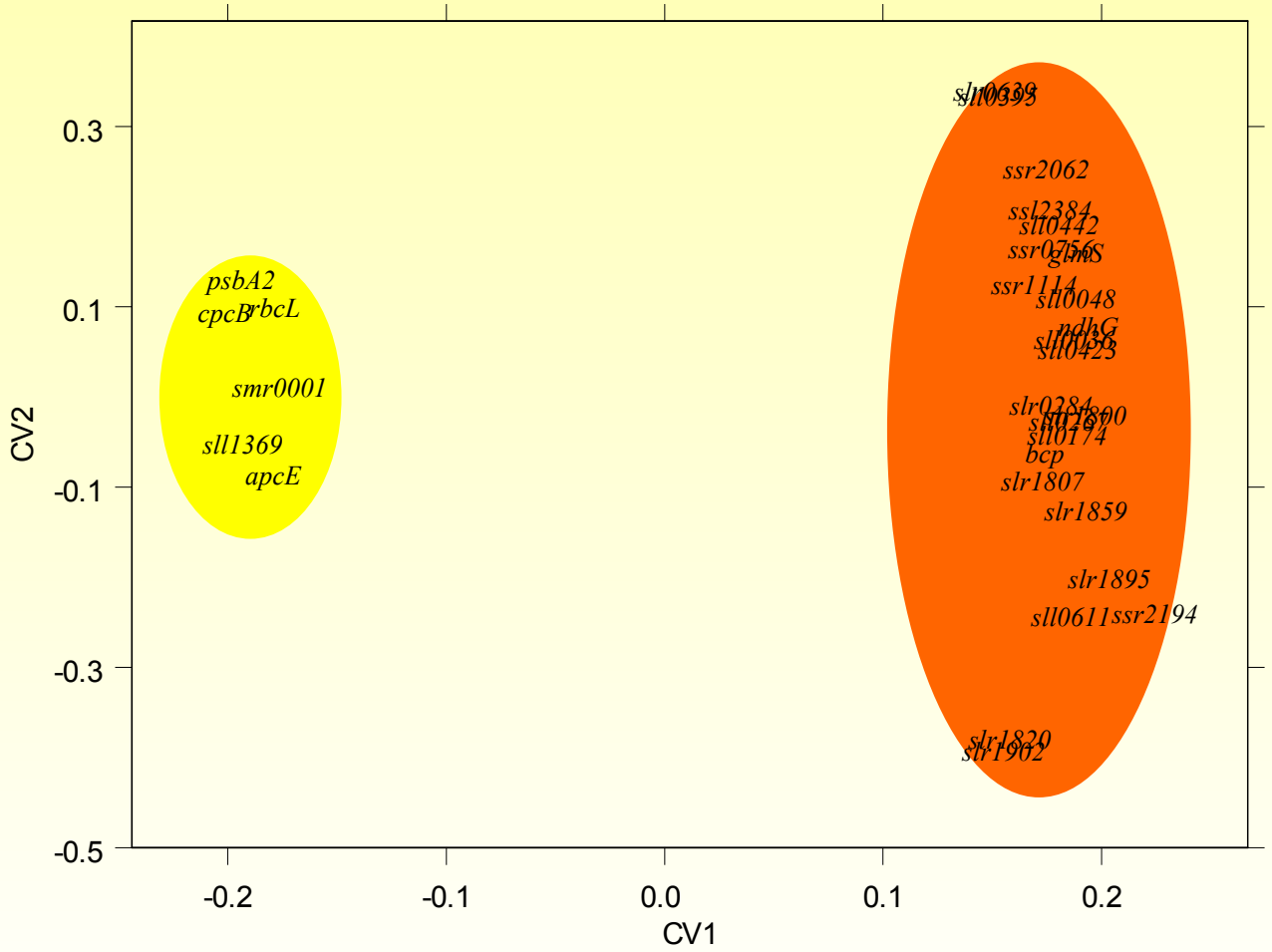
(C)

Dimensionality reduction by Fisher Discriminant Analysis

Projections: $CV_1 = a_1g_1 + a_2g_2 + a_3g_3 + \ldots + a_ng_n$

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

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



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**
- **Of the various types of problems those of physiological nature are most suitable for ChE**
- **This is a young field very much in flux with many problems in search of good solutions**

Spring Semester 2003

10.555 Bioinformatics: Principles, Methods and Applications

Instructors

Gregory Stephanopoulos and Isidore Rigoutsos
Special lectures by Dr. Joanne Kelleher

9 units (H), Class meets Tuesday 2-5pm, Room 56-154

COURSE SYLLABUS

Part I: INTRODUCTION

Lecture 1: February 4 (Assignment 1, due February 25)

- 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 11

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

Part II: SEQUENCE DRIVEN PROBLEMS

No class on February 19 (Monday schedule-President's day)

Lecture 3: February 25 (Assignment 2, due on March 4)

- 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 4 (Assignment 3, due March 11)

- Some useful computer science (notation, recursion, essential algorithms on sets, trees and graphs, computational complexity)
- Physical mapping algorithms
- Fragment assembly algorithms

Lecture 5: March 11 (Assignment 4, due on March 18)

- Comparison of two sequences
- Dynamic programming revisited
- Building and using scoring matrices
- Popular algorithms: Smith-Waterman, Blast, Psi-blast, Fasta
- Multiple sequence alignment
- Functional annotation of sequences

Lecture 6: March 18 (Assignment 5, due April 1)

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

No class on March 25: Spring Break

Lecture 7: April 1 (Assignment 6, due on April 22)

- 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 8

- MCA continued. Analysis of metabolic pathways, fluxes, the *metabolic phenotype*
- Methods for metabolic flux determination
- Linking the metabolic and expression phenotypes

No class on April 15: Patriots Day

Lecture 9: April 22 (Assignment 7, due on May 6)

- Importance of metabolic fluxes in deciphering metabolic controls.
- Linking the metabolic and expression phenotypes
- Use of isotopic tracers for flux determination
- Isotopic Spectral Analysis (ISA)

Lecture 10: April 29

- 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

Lecture 11: May 6

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

Lecture 12: May 13

- Putting it all together
- Project presentations

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 8, 2003. An oral presentation of the project by the group members will take place on May 13, 2003, 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

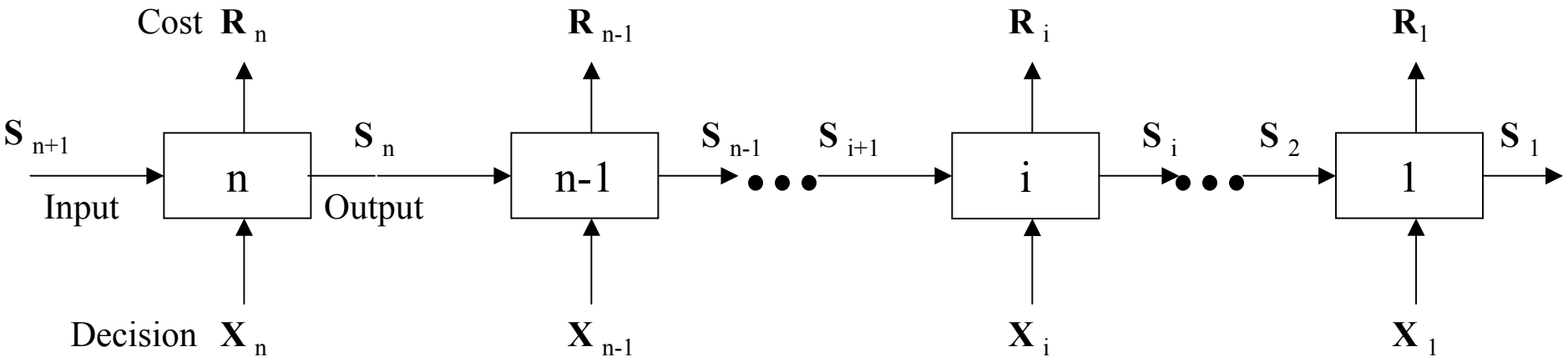
Lecture notes will be posted on the web. Additionally, the course will draw from a series of published papers and the following list of books, which are recommended as references:

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



$\mathbf{S}_i = \mathbf{t}_i(\mathbf{S}_{i+1}, \mathbf{X}_i)$

$\mathbf{R}_i = \mathbf{r}_i(\mathbf{S}_{i+1}, \mathbf{X}_i)$

Overall cost function: $f(R_1, R_2, .. R_n)$
- *Separability*: $f = \Sigma R_i$
- *Monotonicity*

Task: Optimize f

Overall Optimization solved by
- breaking down the problem to the sum of its parts using the recurrence relation:

$$f_i^*(\mathbf{S}_{i+1}) = \underset{\mathbf{X}_i}{\text{opt}}[\, \mathbf{R}_i(\mathbf{S}_{i+1}, \mathbf{X}_i) + f^*_{i-1}(\mathbf{S}_i) \,] \qquad \textit{where * denotes optimum value}$$

- 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, $\mathbf{S_{i+1}}$
- Alignment of preceding prefix

Output, $\mathbf{S_i}$
- Alignment of preceding prefix and current element

Decision, $\mathbf{X_i}$
- Alignment with an element in the sequence
- Alignment with a gap

Cost Function, $\mathbf{R_i}$
- reward for exact match
- penalty for mismatch
- penalty for matching with a gap or a break in the sequence

COMPARISON OF SEQUENCES

s = AAAC t = 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*

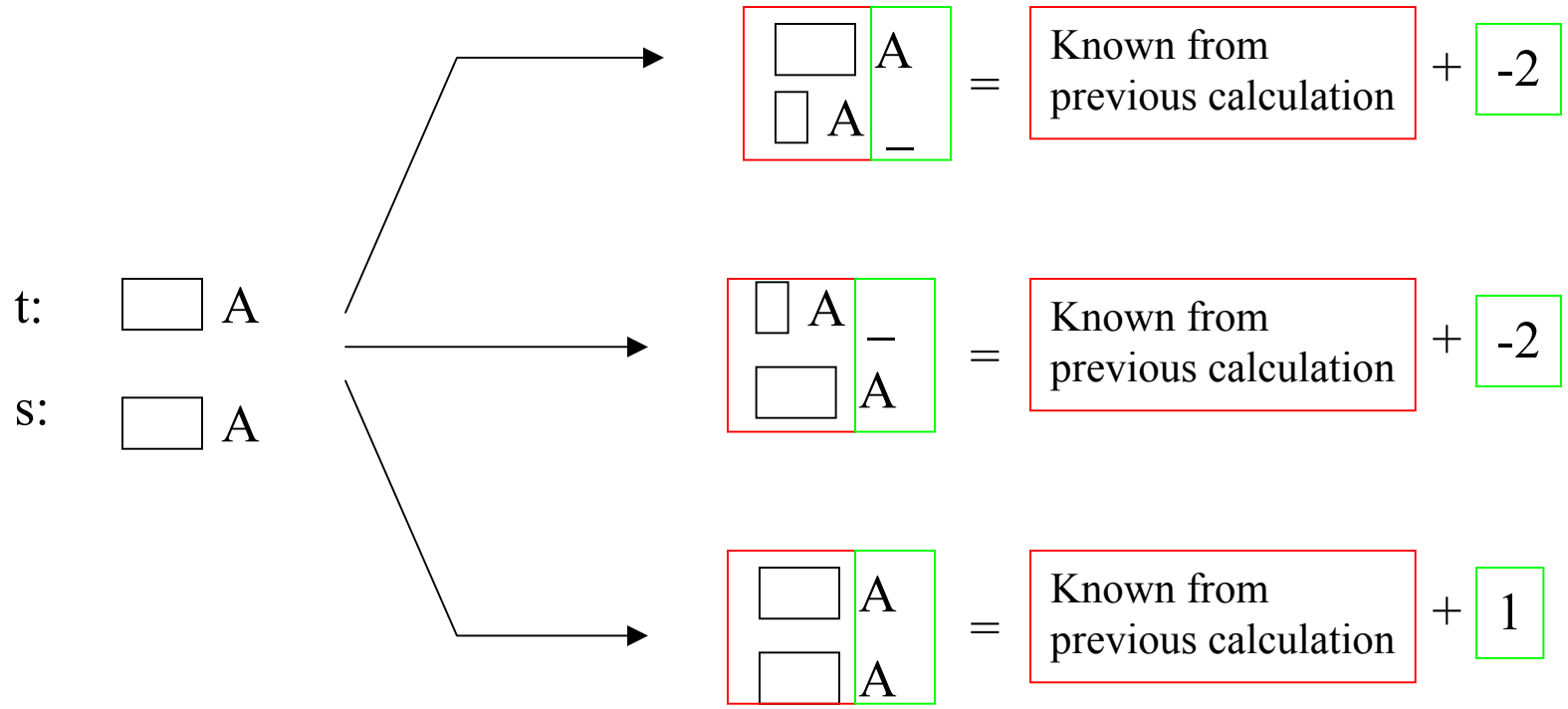
$$f^*(s[1:i], t[1:j]) = \max \begin{cases} f^*(s[1:i], t[1:j-1]) - sp \\ f^*(s[1:i-1], t[1:j-1]) + pm, \text{ if } s[i] = t[j] \\ f^*(s[1:i-1], t[1:j-1]) - mm, \text{ if } s[i] \neq t[j] \\ f^*(s[1:i-1], t[1:j]) - sp \end{cases}$$

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

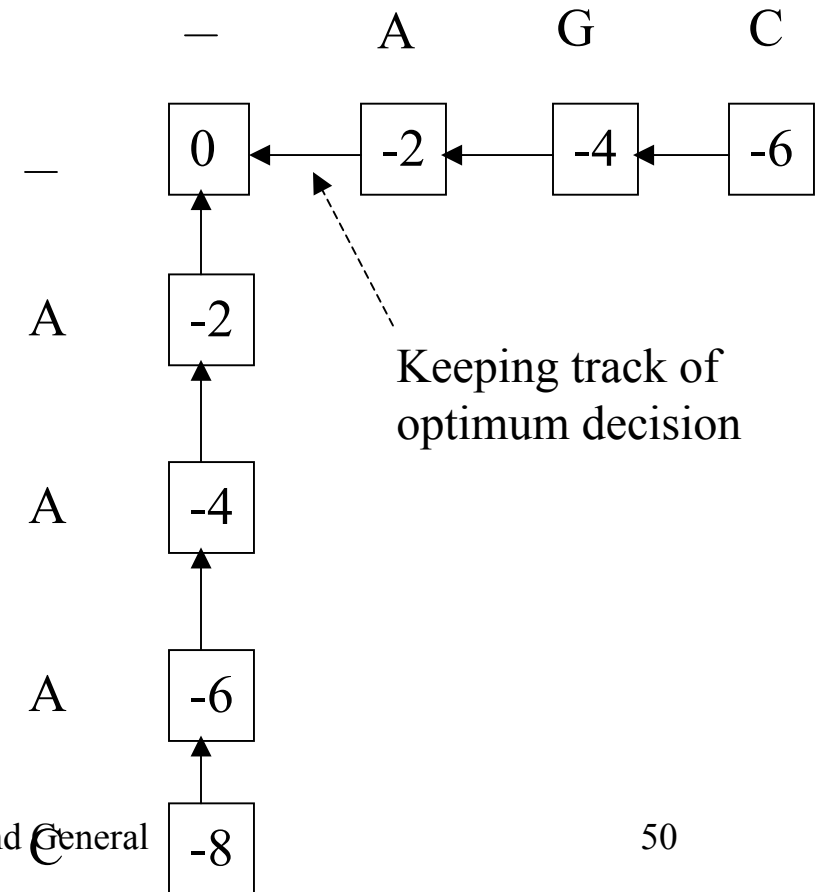
i = 1, j = 2: t: _ A *Cost = 0 + -2 = -2*
s: –

i = 1, j = 3: t: _ AG *Cost = -2 + (-2) = -4*
s: –

i = 1, j = 4: t: _ AGC *Cost = -4 + -2 = -6*
s: –

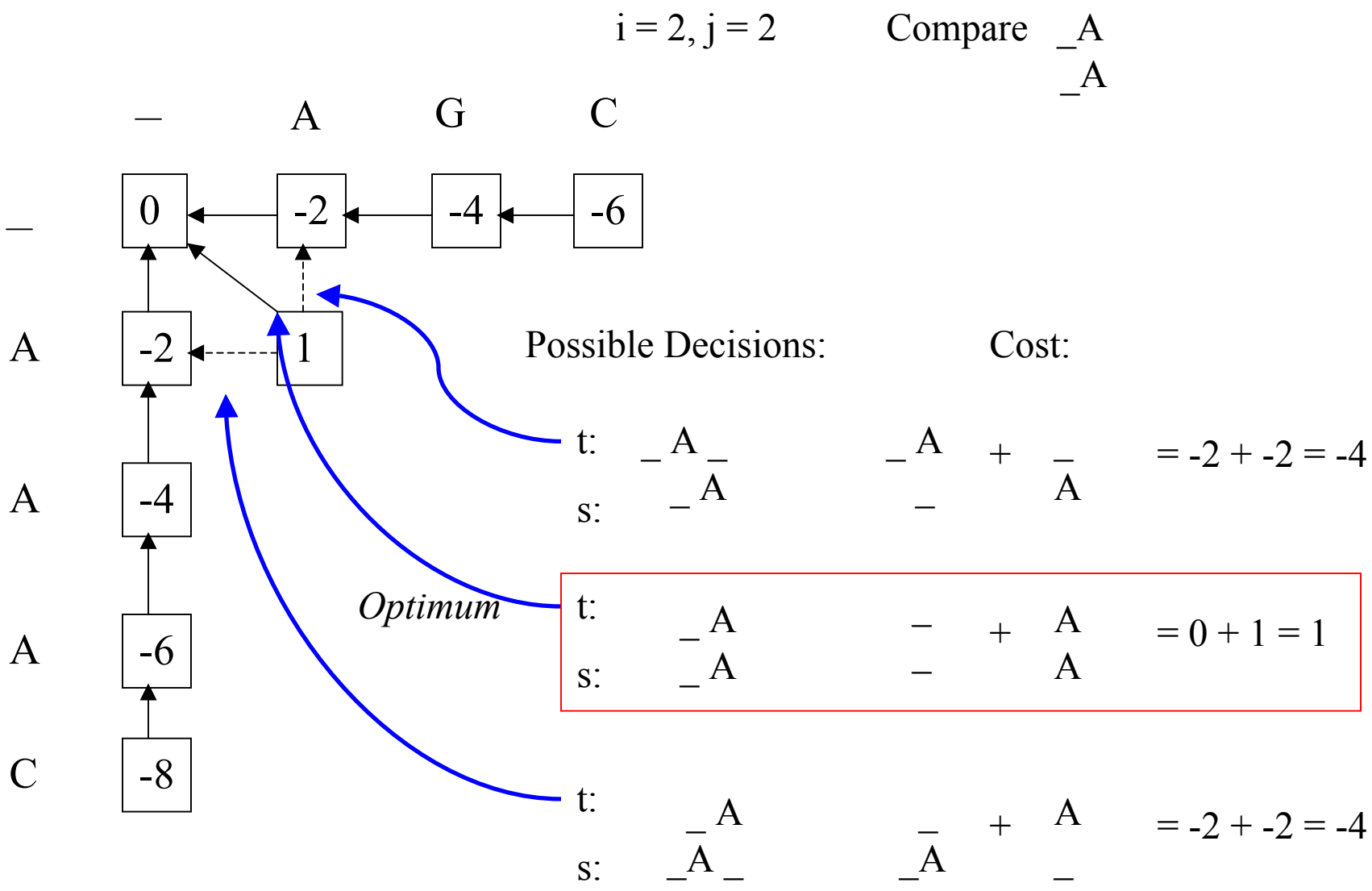
i = 2, j = 1:

•
•
•



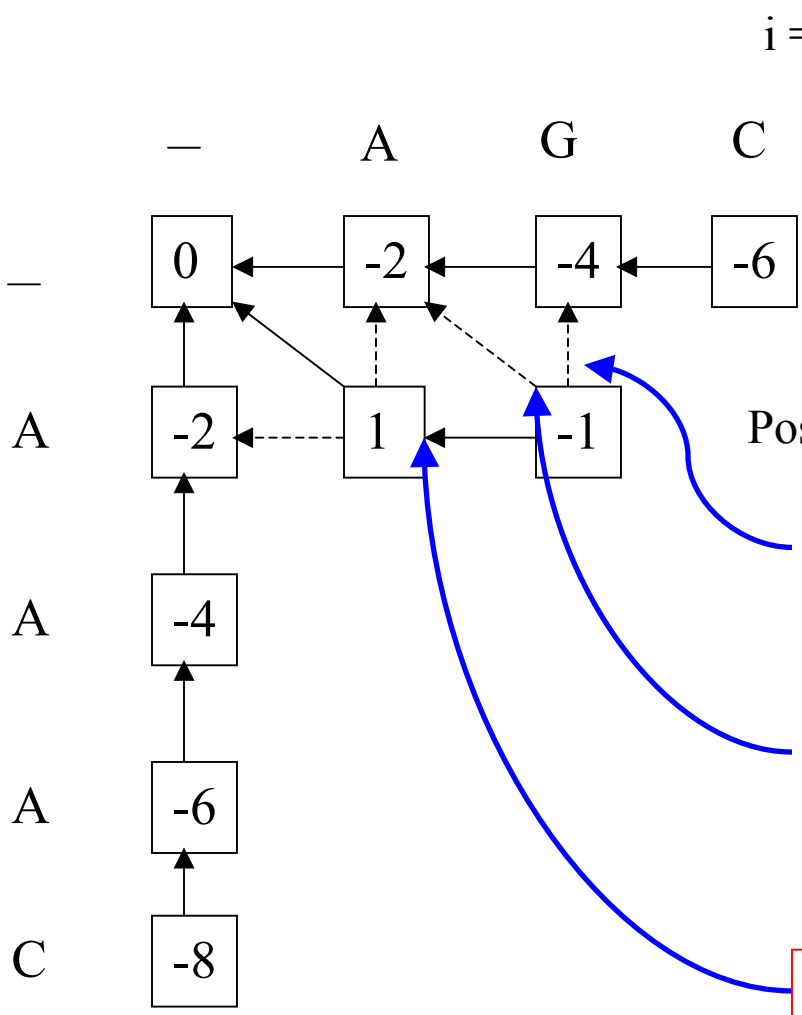
ELUCIDATION OF ALGORITHM

Step 3(0): Apply recursive optimization



ELUCIDATION OF ALGORITHM

Step 3(1): Apply recursive optimization

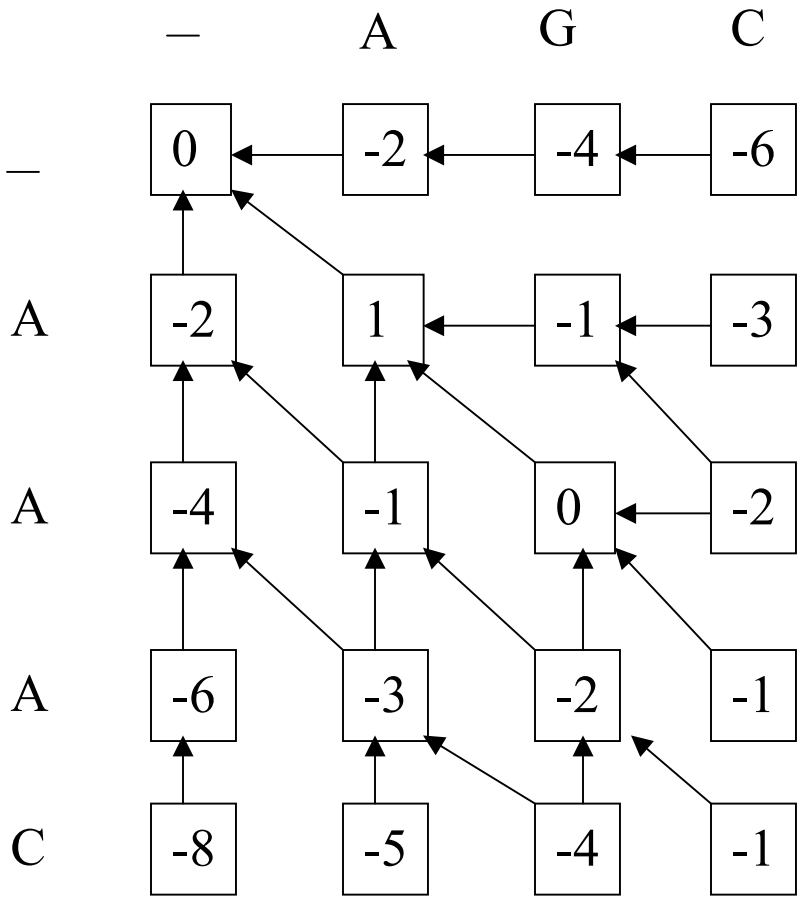


And so on, and so forth ...

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

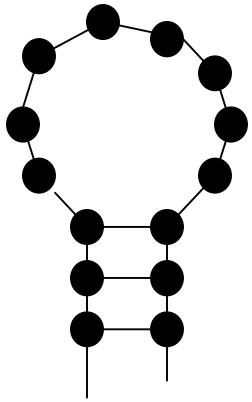
Allows for Global/ Local Alignment

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