10.555
Bioinformatics: Principles, Methods and Applications

MIT, Spring term, 2003

Lecture 7:

- Using sequence analysis tools to solve problems

- Physiology: Definitions and measurements at the cellular, molecular and organismal levels

10.555
Solving sequence problems

Problem: Discover primary sequence features that are critical for a particular gene function

- Promoter binding sites
- Enhancers
- Transcription factors
- Directing proteins to specific pathways (secretion)
- Endowing proteins with a particular property
- Unknown genes

How do we characterize the sequences we seek?

These sequences,

- May be over-represented
- Are very different and so they can be distinguished from the background noise
- Look at databases (Transfac, regulons, etc.)
- Do smart experiments to screen some of these sequences (transcriptional studies, other)

Be careful with "obvious" things

For example, in eukaryotes:

- Regulatory sequences may be located quite far from the corresponding coding region, upstream or downstream
- Need not be in the same orientation as the coding sequence
- There can be great variability in the binding sites of a single factor (not well understood)

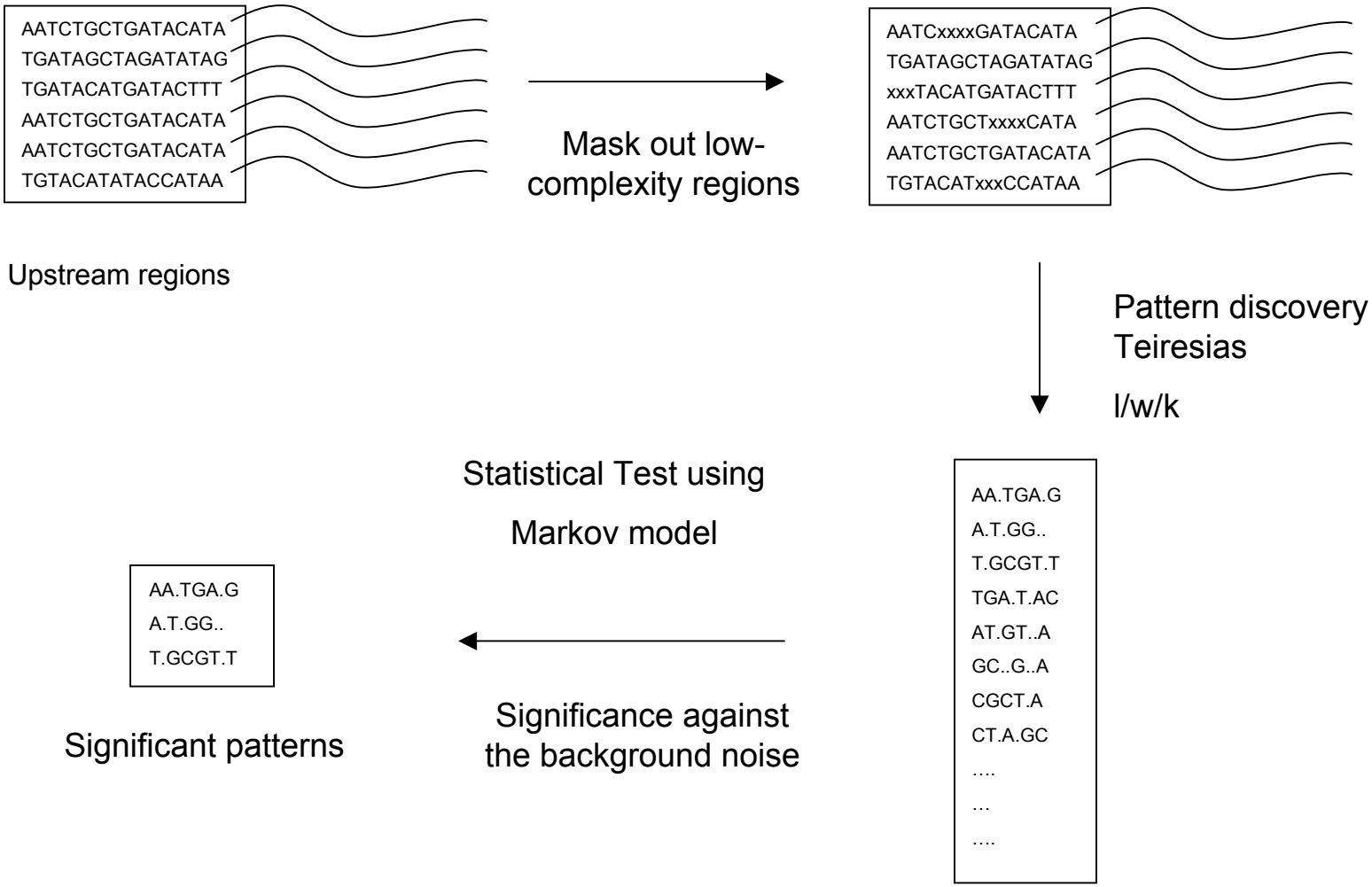
After *searching* literature and *working with databases* you may find

- Most known sites lie within 800bp upstream of structural genes
- Number of well conserved bases in the sites of a single t-factor is typically 6-8
- There are 0-11 wild-cards in the middle

e.g. $AGGN^{0-11}CGC$

- From a database: the above description matches 70% of their consensus motifs
- Size of known sites: 8 – 50 (median 17)
- Known sites are randomly located in upstream regions
- Poly A's and poly T's are over-represented (A,T:30%, G,C: 20%)

The Overall Scheme



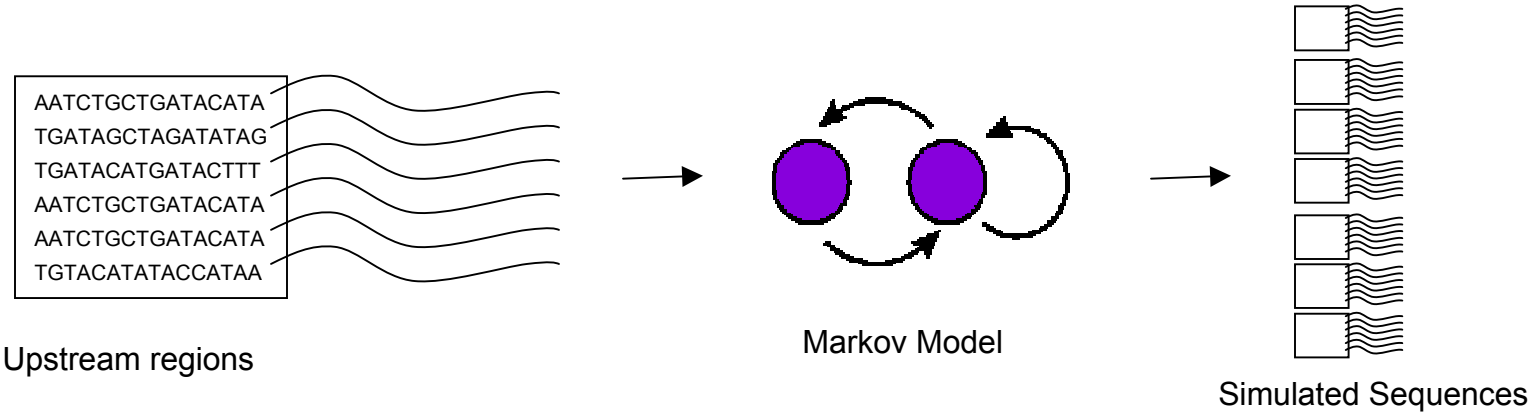
Select sequences to do pattern discovery

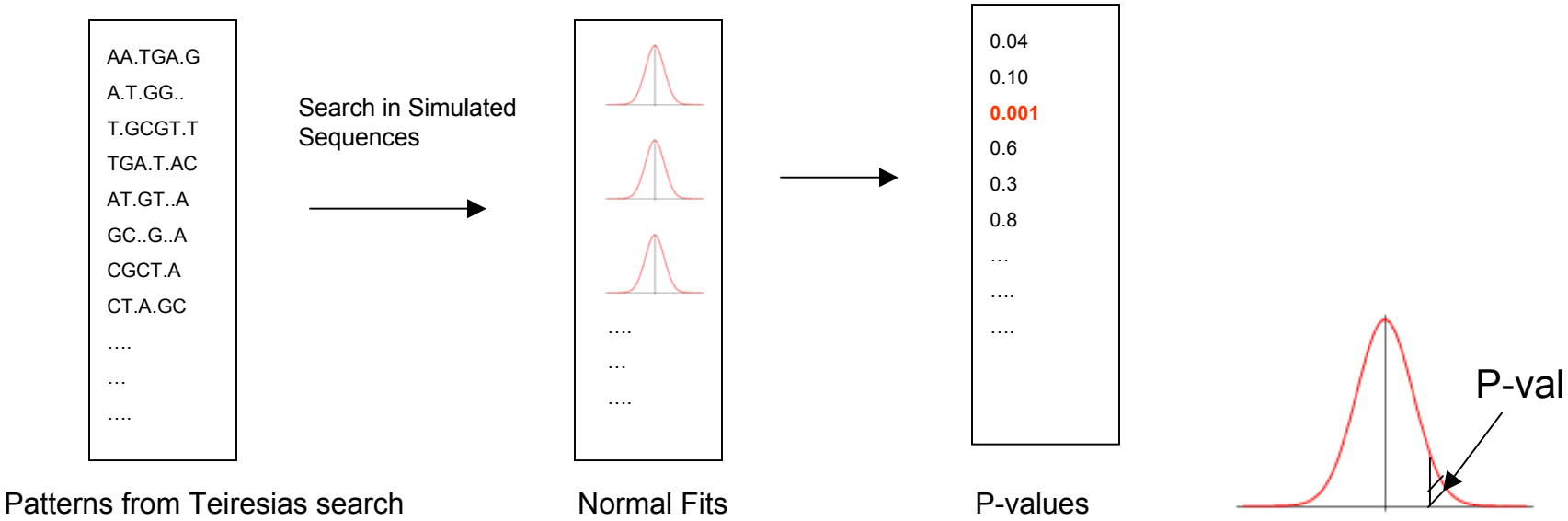
- All genes in genome
  - Computationally intensive
  - Get a lot of junk
- Clusters of genes sharing property related to the property investigated

Finding Patterns - Teiresias

- l / w / k
  - l -> 6 to 8 CTT….A.TG
  - w -> 17 to 19
  - k -> ?
- Heuristic approach (lot of overlaps)
  - Specify k, find all patterns >=k
- Top down approach (avoids overlaps, but may also lose some patterns)
  - Find all patterns with maximal support
  - Collect them, mask them
  - Drop support and repeat

Statistical Significance Test





First Question: ***Does a Short DNA Stretch Come from a CpG Island?***

Table of Transition Probabilities for CpG Islands

Model +	A	C	G	T
A	.180	.274	.426	.120 = 1
C	.171	.368	.274	.188 = 1
G	.161	.339	.375	.125 = 1
T	.079	.355	.384	.182 = 1

Table of Transition Probabilities for Regions with no CpG Islands

Model -	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Calculate the Log-Odds ratio for a chain *x:*

$$S(x)=\log_2 \{[P(x/model+]/[P(x/model-]\} = \Sigma_i \log_2 \{a^+_{x(i-1)x(i)} / a^-_{x(i-1)x(i)}\} = \\ = \Sigma_i \log_2 \beta_{x(i-1)x(i)}$$

Scores S(x) allow discrimination of a model (+) against another (-)

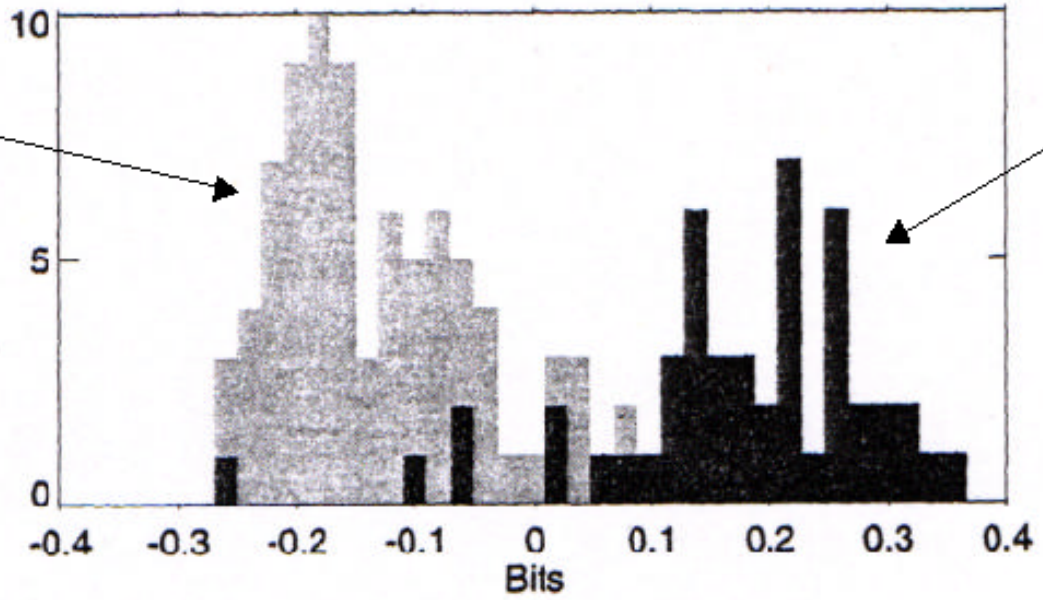
First Question: *Does a Short DNA Stretch Come from a CpG Island?*

Likelihood Ratios

β	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

Test a Given Stretch of DNA

Regions free of CpG Islands



CpG Islands

Parameter estimation, model selection

Problem: Two models, M_1 and M_2 can be compared by comparing their probabilities $P(M_1/D)$ and $P(M_2/D)$. The *best* model *in its class* is found by determining the set of parameters w maximizing the posterior probability p(M/D), or

$$\mathbf{Min(-\log P(M/D) = -\log P(D/M) - \log P(M) + \log P(D)}$$

This is called ***MAP estimation*** **(Maximum *a posteriori*)**

P(D) is a normalizing constant independent of optimization. If the prior P(M) is uniform over all models then the above problem is reduced to the following *Maximum Likelihood (ML)* maximization *(ML estimation):*

$$\mathbf{Min\ (-\log P(D/M)}$$

Case study: Find transcription factor

- **It regulates 9 genes**
- **Database lists binding sites for each gene**

SCPD database - GCN4

Get regulated genes | Get sites | Get consensus | Get matrix | Get affinity data

Get genomewise distribution | Sort by copy No | Sort by function category

>YBR248C	AAGAGTCAG
>YBR248C	TTGAGTCAT
>YCL030C	TGACTA
>YCL030C	TGACTC
>YCL030C	ACAGTGACTCACGTTT
>YCL030C	TGACTC
>YCL030C	CAGTCA
>YDR354W	ATGATTCAT
>YDR354W	TTGACTCTC
>YDR354W	ATGACTAAT
>YER086W	TGAGTG
>YER086W	AAGTCA
>YER086W	GAGTCA
>YMR108W	TGATTC
>YMR300C	TTGACTCTT
>YMR300C	ATGACTGCT
>YMR300C	ATGAATAAT
>YOL058W	TGACTCA
>YOL058W	GAGTCAT
>YOL140W	TGACTCA
>YOR202W	TGACTC
>YOR202W	ATGACTCTT
>YOR202W	TTACTC
>YOR202W	TGACGA
>YOR202W	AAGTCA
>YOR202W	TAGTCA
>YOR202W	GAGTCA

Reported Consensus –TGA.T.

Case study: Find transcription factor

- **It regulates 9 genes**
- **Database lists binding sites for each gene**
- **Teiresias: 5 /16 / 7 parameters**
- **Found ~23000 patterns**
- **Developed statistical model: 3rd order Markov for randomly distributed sequences**
- **Found ~2,000 motives with $p<0.01$**
- **No poly A's or poly T's**
- **Pattern closest to the consensus (TGA.T.) had p value of 0.003**

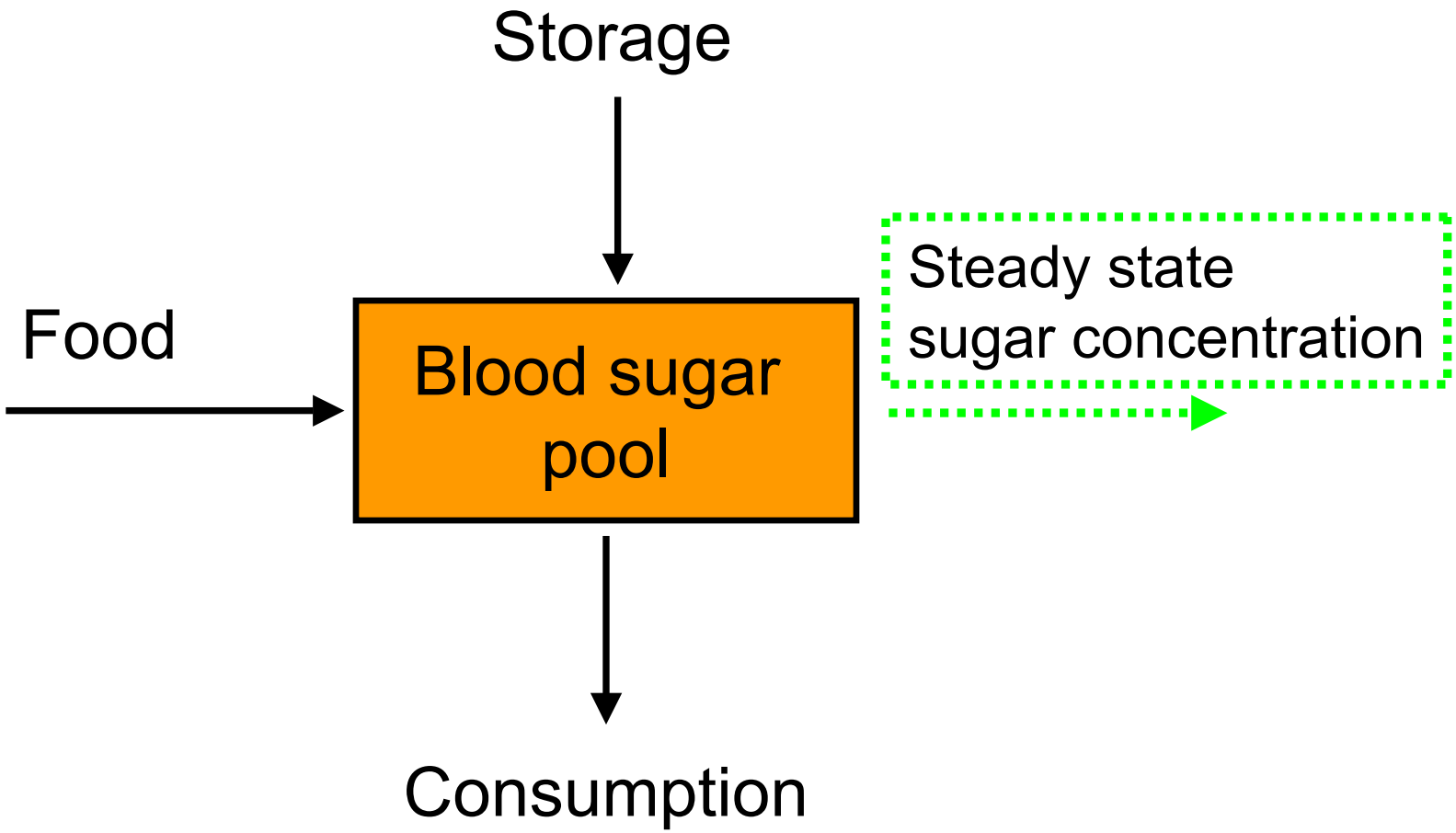
Defining and understanding Physiology

PHYSIO - LOGY

Physical state **Logos = Subject of inquiry/expertise**

- **Describe the state of living cells and organisms**
- **Understand the mechanisms by which homeostasis is achieved in organisms**

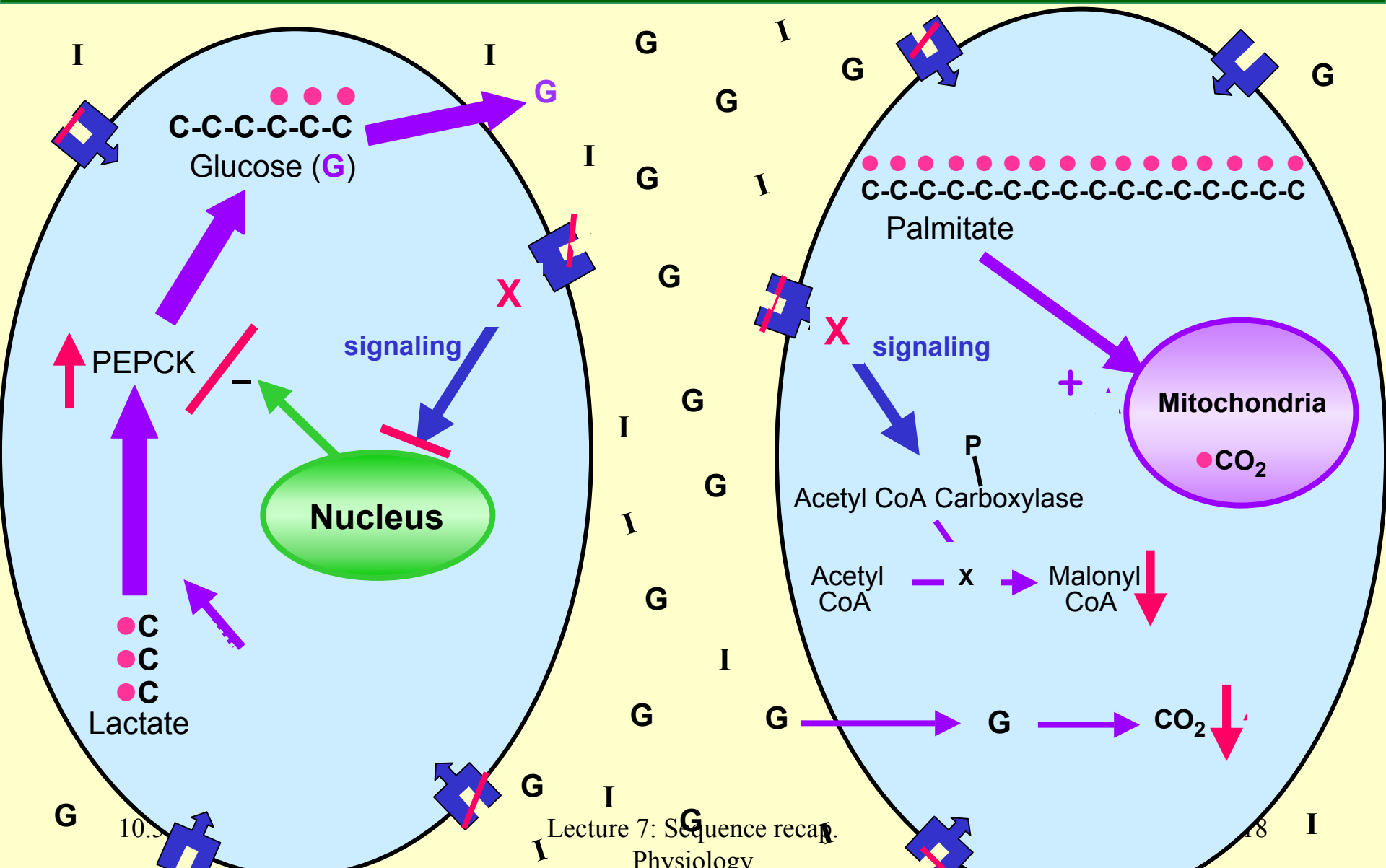
Controlling blood sugar level



HEPATIC INSULIN RESISTANCE, POOR RESPONSE TO INCREASED PLASMA GLUCOSE

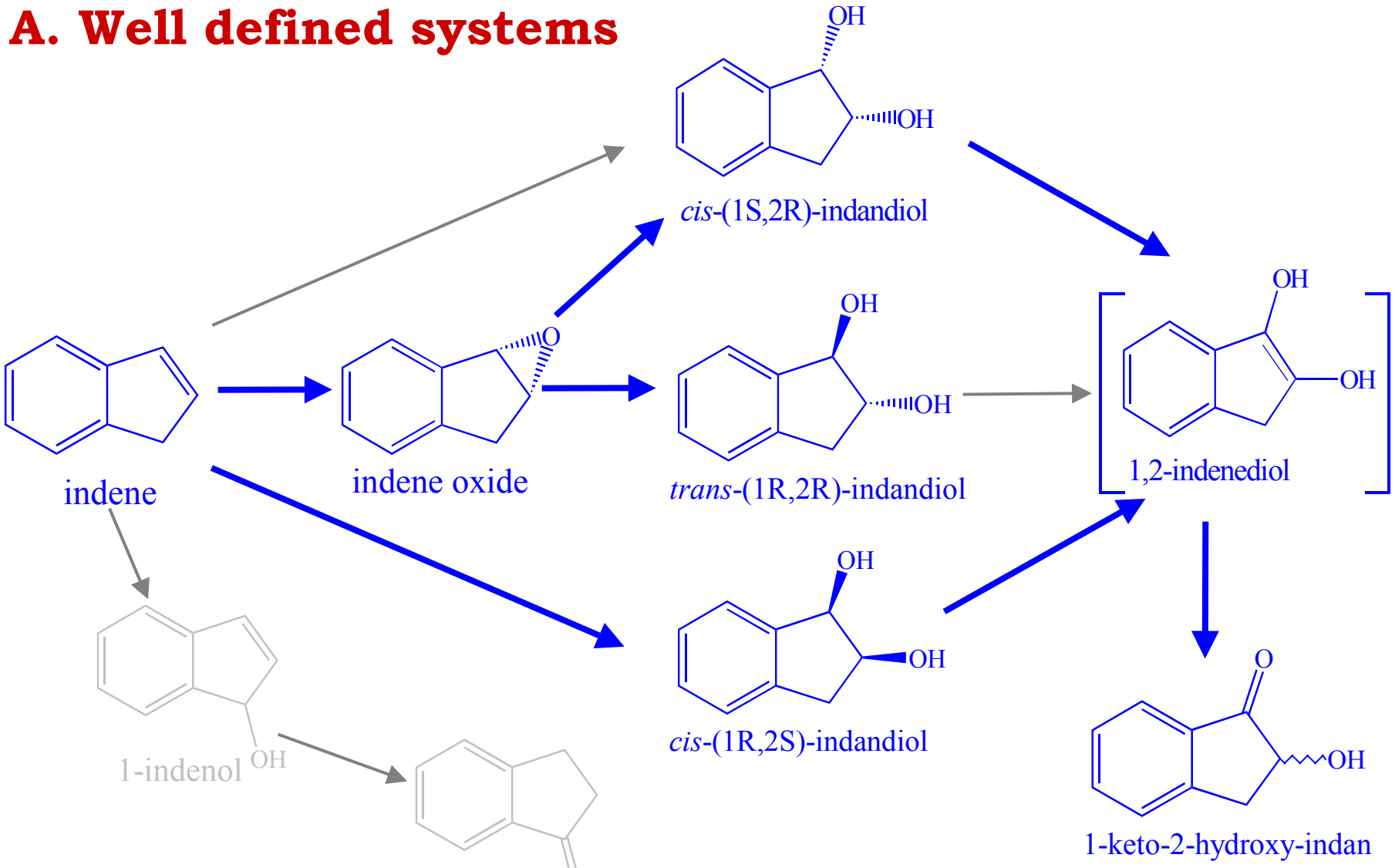
HPERGLYCEMIA VIA INCREASED PEPCK GENE EXPRESSION

HPERGLYCEMIA VIA DECREASED ACETYL COA CARBOXYLASE & MALONYL COA



Methods depend on available measurements

A. Well defined systems



Methods depend on measurements available

B. Systems that are not so well understood

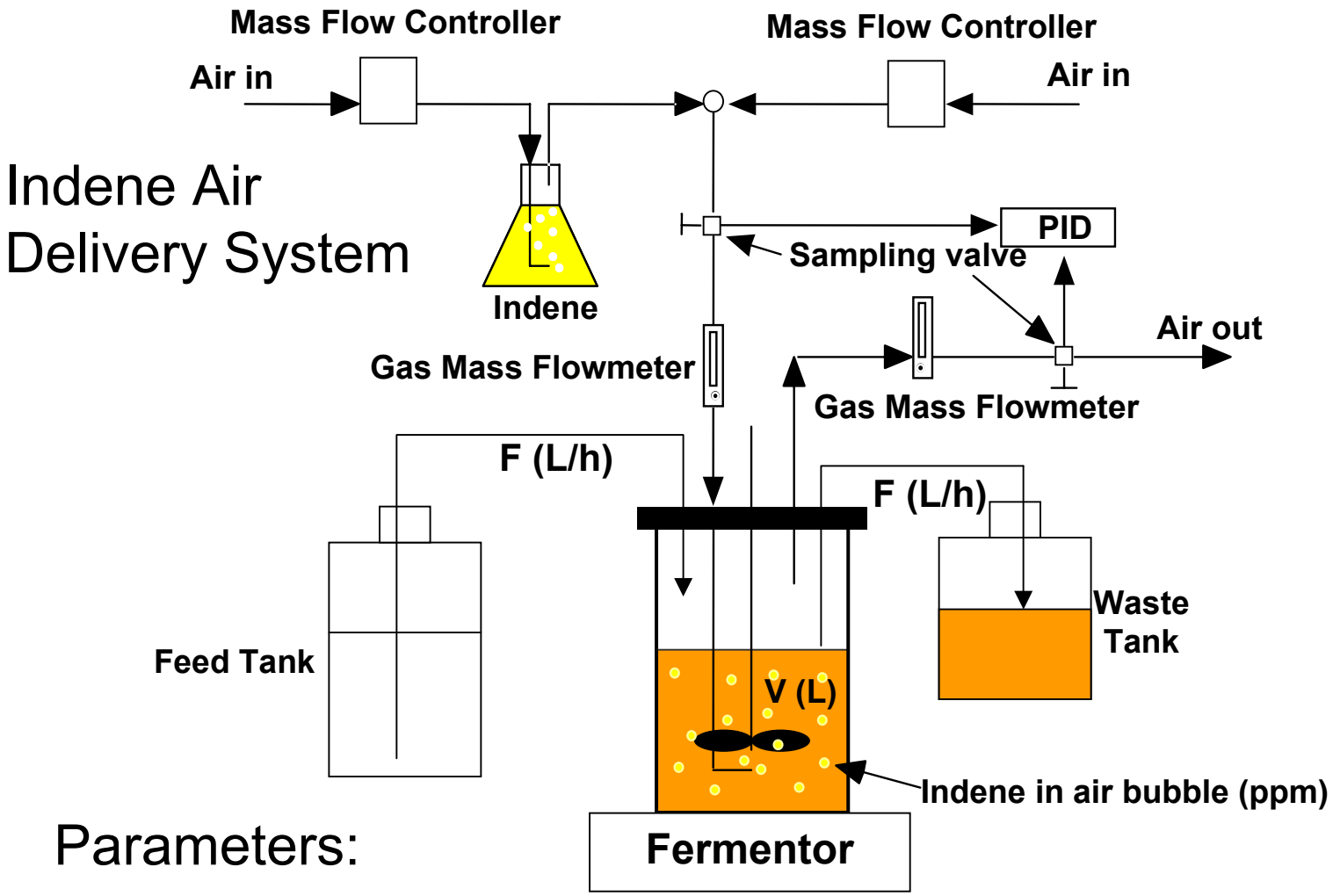
- **Different approaches are required**

- **Makes no sense to pursue system description at ultimate level of detail**

Two broad categories of systems:

- **Cells**
- **Tissues and whole organisms**

Where do we study cell physiology? Bioreactors: Continuous, batch and fed-batch



Parameters:

- Dilution Rate, D=F/V
- Indene Air Feed Concentration

Bioreactor balances

See notes

Bioreactor balances: Summary

Remember goal: **Study physiology of cells**

This means develop a system and protocol to allow the measurement of important physiological variables:

Specific growth rate: μ (h^{-1})
Specific rate of substrate uptake: q_s (g glc/g cells * h)
Specific secretion rates: q_p (g of P/g cells * h)

Similarly for other measurable extracellular metabolites

These variables provide little intracellular insight!

Intracellular measurements

- **Metabolite measurements**
  - **Concentrations**
  - **Isotopic tracer distributions (using labeled substrates)**
    - **^{13}C enrichment of specific metabolite carbons (NMR)**
    - **Mass isotopomers (GC-MS)**
    - **Radioisotopes**
- **Proteins**
  - **Specific proteins**
  - **Protein profiles (proteomics)**
  - **Modified protein fractions (phosphorylated, glycosylated, etc.)**
- **mRNA transcripts (DNA microarrays)**

Use of measurements

- Reporters of intracellular state
- In conjunction with models
  - Predictive
  - Descriptive
    - Calculation of informative parameters

This use requires additional knowledge, usually of mechanistic nature. Such knowledge is more readily available in cellular systems

- Use as profiles characteristic of physiological states (molecular physiology)