Quantitative Molecular Biology

PHYS 176/276
Instructor: Terry Hwa
Winter 2021

What is quantitative biology?

→ quantitative biology ≠ biology + numbers/equations
  ≠ application of quant tools to bio

→ use numbers as clutches to gain predictive understanding

Why quantitative biology?

-- because biology is quantitative
-- needed to formulate and test falsifiable predictions
-- demanded by synthetic biology

Role of theory

• formulate expectation and predictions (via quantitative model)
• guide the design of new experiments and technology
• power: the generality of (falsifiable) ideas, not necessarily math
  [e.g., Copernicus, Darwin, Einstein]
• “cost”: the simplifying assumptions, not necessarily forced by math,
  but required in order to reveal principles

→ This course: quantitative molecular biology of bacteria
Life of a bacterium:

molecular composition: \( \text{CH}_{1.5} \text{O}_{0.35} \text{N}_{0.24} \) (+S, P, Mg, Fe, …)

algal photosynthesis:
\[
\text{CO}_2 + \text{H}_2\text{O} + \text{N}_2 + \text{photons} \rightarrow \text{biomass} + \text{O}_2
\]

\( E. \) coli (minimal medium):
\[
\text{glucose} + \text{NH}_3 \rightarrow \text{biomass} + \text{CO}_2
\]

Learning from the growth curve

- \( \text{OD}_{600} = \) biomass content
  
  \[1 \text{ OD}\cdot\text{ml} = 0.5\text{mg CDW}\]

- saturation OD \( \rightarrow \) yield

- (lag: transition from pre-shift phase)
growth of E. coli

Can we predict GR & yield?

environmental factors: nutrient types & conc, temperature, pH, osmolarity, drugs, ...
genetic factors: enzymes & regulation

What does it take to replicate a cell?

protein = defined sequence of 20 amino acids

protein synthesis: ribosomes

What does it take to replicate a cell?

condition-dependent

water

biomass

DNA

other

RNA

protein

condition-dependent

biomass

DNA

other

RNA

protein

water

biomass

DNA

other

RNA

protein

condition-dependent

water

biomass

DNA

other

RNA

protein

condition-dependent

water

biomass

DNA

other

RNA

protein

condition-dependent
protein = defined sequence of 20 amino acids

protein synthesis: ribosomes

>85% of all RNA up to 1/3 of all proteins
protein = defined sequence of 20 amino acids

protein synthesis: ribosomes

amino acids & ATP from metabolic reactions

metabolism
• sequester & breakdown nutrients
  - derive energy
  - generate carbon precursors
  - sequester N, S, P, metals
  
  glycolysis
  glucose (6C)
  2x pyruvate (3C)

  respiration
  32x ATP

  fermentation
  8x ATP

  2x acetate + 2x CO₂

  but many organisms use fermentation even with oxygen (Crabtree effect); why?

biosynthesis
  ("precursors" to "building blocks")
  - amino acid
  - nucleic acid
  - lipids
  - co-enzyme (or 'co-factor')

  IlvBN, IlvGM, IlvIH

  IlvC

  IlvD

  IlvE

  NH₃

  H₂O

  H⁺, NADPH

  valine

  2x e⁻ C-H bond to e⁻
metabolism
• sequester & breakdown nutrients
  – derive energy
  – generate carbon precursors
  – sequester N, S, P, metals
• biosynthesis of building blocks
  – amino acid
  – nucleic acid
  – lipids
  – co-enzymes
• degradation/recycling (e.g., mRNA)
• typical biochemical reaction:
  \[ S + C \cdot b \rightleftharpoons S \cdot b + C \]
  
  \[ S: \text{substrate} \]
  \[ b: \text{component (e.g., CH}_3\text{, NH}_2, \text{e') } \]
  \[ C: \text{co-enzyme} \]
  
  (needed for difficult reactions)

  ➔ most reactions catalyzed by enzymes (proteins)
  ➔ flux of the products and “by-products” need to be balanced

metabolic control via coordinated regulation of enzyme abundance/activity

feedback inhibition
• 1st reaction of pathway often inhibited by product
• same enzymes used for synthesis of valine and isoleucine
• must have enzymes responding differently to different products (isozymes)
• in E. coli K-12, ilvG is defective
  ➔ valine sensitivity in minimal media
  ➔ \( \alpha \)-ketobutyrate toxicity (repressed by isoleucine)
protein = defined sequence of 20 amino acids

protein synthesis: ribosomes

\[ J_i = k_i [RS_i] \frac{[a_i]}{K_{a,i} + [a_i]} \cdot \frac{[tR_i]}{K_{tR,i} + [tR_i]} \]

~100 variables
~500 parameters

free AA & uncharged tRNAs

protein = defined sequence of 20 amino acids

protein synthesis: ribosomes

free AA & uncharged tRNAs
protein = defined sequence of 20 amino acids

protein synthesis: ribosomes

~100 variables
~500 parameters

~500 enzymes
~500 metabolites
~5000 parameters

dependence on temp, pH, osmolarity, ...

regulation: when and how much proteins to make

~10000 parameters

parameter explosion

How to deal with exploding no. of parameters?

Newtonian Mechanics

\[
\frac{d^2 \vec{r}_i}{dt^2} = \vec{f}_{ij}(\vec{r}_i - \vec{r}_j)
\]

Need moles of parameters:

\[
\vec{r}_i(t = 0) = \ldots
\]

\[
\vec{v}_i(t = 0) = \ldots
\]

Thermodynamics

\[PV = nRT\]
How to deal with exploding no. of parameters?

Newtonian Mechanics \[\rightarrow\] Thermodynamics

Statistical Mechanics

\[\rho(q, p; t) \rightarrow \rho_{eq}(\mathcal{H}(q, p))\]

Molecular/Cell Biology \[\rightarrow\] Physiology

Quantitative Systems Biology
This course: explore cellular strategies of dimension reduction using bacterial mechanisms

Molecular aspect: Central Dogma + regulation

- tsx initiation control by transcription factors (TF)
- tsl initiation control by sRNA and RNA-binding proteins
- tsx termination control by sRNA and anti-terminators
- control of mRNA and protein degradation
- control of enzyme activity by metabolites

how does it all work together? what can be predictively manipulated? what parameters are needed?

sugar, NH₃, O₂
molecular biology of transcription (RNA synthesis)

transcriptional initiation

- elongation and termination

transcriptional initiation control

- modulation of RNAp-promoter affinity via activators and repressors

→ net result: rate of transcriptional initiation dependent on cellular conc of activators and repressors controlled by metabolites and signaling molecules
Quantitative aspect

Dimension

- DNA: 2 nm x 2 nm x 3.4 nm/turn
- small proteins: (few nm)\(^3\) or ~10nt
- protein complexes, (10-20 nm)\(^3\) or 30 ~ 60nt
- cell size: 1 \(\mu\)m\(^2\) x 3 \(\mu\)m (condition dependent)
- concentration: 1 molecule/cell ~ 1nM (=0.6/ \(\mu\)m\(^3\))
- intracellular diffusitivity of protein: ~10 \(\mu\)m\(^2\)/sec
(can usually be regarded as ‘well mixed’)

abundance

- ribosomes: ~ 20,000 (52 proteins each)
- RNAp ~ 1,000 (a portion available)
- proteins: 2x10\(^6\) (TF: 10 ~ 1,000 / type)
- mRNA: 3% of RNA
  - 0.1 ~ 100 copies/cell;
  - peaked at 2 ~ 3 copies / cell

### 25% of bacterial dry mass is concerned with gene expression

<table>
<thead>
<tr>
<th>Component</th>
<th>Dry Cell Mass (%)</th>
<th>Molecules /cell</th>
<th>Different types</th>
<th>Copies of each type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Membrane</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DNA</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>mRNA</td>
<td>1</td>
<td>1,500</td>
<td>600</td>
<td>2-3</td>
</tr>
<tr>
<td>rRNA</td>
<td>3</td>
<td>200,000</td>
<td>60</td>
<td>&gt;3,000</td>
</tr>
<tr>
<td>tRNA</td>
<td>16</td>
<td>38,000</td>
<td>2</td>
<td>19,000</td>
</tr>
<tr>
<td>Ribosomal proteins</td>
<td>9</td>
<td>10(^6)</td>
<td>52</td>
<td>19,000</td>
</tr>
<tr>
<td>Soluble proteins</td>
<td>46</td>
<td>2.0 x 10(^6)</td>
<td>1,850</td>
<td>&gt;1,000</td>
</tr>
<tr>
<td>Small molecules</td>
<td>3</td>
<td>7.5 x 10(^6)</td>
<td>800</td>
<td></td>
</tr>
</tbody>
</table>

Warning: many of these numbers are dependent on growth condition!
rates

- transcription: elongation ~50 nt/s
- translation: ~ 16 aa/s
- transcription-translation coupling: infrequently translated mRNA cleaved
- mRNA half-life: < 5 min

- protein half-life: from cell-doubling time (passive decay) down to a few min (active proteolysis)

RNA synthesis:

- transcriptional initiation

- elongation and termination

heavily transcribed genes coding ribosomal RNA
3. **Transcription and translation rates:** In parts (a)-(f), deduce the transcription and translation rates for typical (non-ribosomal) genes in the exponential growth phase. Take the doubling time to be 45 min, and use the average copy number of a gene to be 2. Please report your numerical answers as well as the mathematical expressions.

<table>
<thead>
<tr>
<th>Table 1: Pertinent values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{\text{max}}$</td>
</tr>
<tr>
<td>$\tau_{\text{act}}$</td>
</tr>
<tr>
<td>$L_{\text{RNA}}$</td>
</tr>
<tr>
<td>$R_{\text{RNA}}$</td>
</tr>
<tr>
<td>$t_{\text{mRNA}}$</td>
</tr>
<tr>
<td>$t_{\text{mRNA}}$</td>
</tr>
<tr>
<td>$T$</td>
</tr>
<tr>
<td>$\mu$</td>
</tr>
<tr>
<td>$g$</td>
</tr>
<tr>
<td>$G_{\text{E. coli}}$</td>
</tr>
<tr>
<td>$l_{\text{gen}}$</td>
</tr>
</tbody>
</table>

(a) Given the maximal speed of transcriptional elongation by RNAp ($\sim 48$ nt/sec) and the physical size of RNAp (covers $\sim 55$ nt), find the maximal rate at which full-length mRNA transcripts can be synthesized.

(b) Given the half-life of 2 min for a typical transcript, what is the maximal copy number for each type of such transcripts in the steady state (of balanced exponential growth)?

**“Systems Biology”**

**Scope and focus:**

- Biological systems whose **functions** are derived from the interaction of **many sub-components**
- Ex: from macromolecular assemblies to ecological communities
- Central focus: subcellular and cellular processes, e.g., genetic circuits, protein interaction networks

- Long-term goals:
  - Mapping out the complete wiring diagram of the cell
  - Quantitative, predictive computational model of the cell
Circuit diagram as system-level descriptor?

- Circuit diagram supplemented by component parameters provides a concise quantitative description of the system.
- Circuit topology not necessarily predictive of system function; need to know the properties of the nodes; parameter explosion.

<table>
<thead>
<tr>
<th></th>
<th>electronic circuits</th>
<th>genetic circuits</th>
</tr>
</thead>
<tbody>
<tr>
<td>components</td>
<td>simple &amp; well-characterized; many (~10⁹); fast (10⁻⁹ sec)</td>
<td>heterogeneous, most rates unknown; few (~1000); slow (&gt;10 min)</td>
</tr>
<tr>
<td>connectivity</td>
<td>physical interconnect between well-insulated components (1~2 inputs per node)</td>
<td>multiply-connected (1~10 inputs per node); regulation at all stages</td>
</tr>
<tr>
<td>network complexity</td>
<td>iterated cascades from complex network wiring</td>
<td>combinatorial signal integration from complex molecular control</td>
</tr>
</tbody>
</table>

Experimental & Computational Approaches

- **Traditional mol bio:** one gene, one process (e.g., A activates B)
  - high throughput methods
  - bioinformatic analysis
  - one/few genes, multiple processes (e.g., txn init/term, post-tsl modification, degradation, small genetic circuits)
  - quantitative expts and modeling
  - design and synthesis of biomolecular systems

- **Systems biology:** many components, many processes (e.g., predictive modeling of the cell)
  - qualitative circuit diagram linking many nodes
  - quantitative model of individual nodes
  - physiology
  - many genes, one/few process(es) (e.g., transcriptomics, metabolomics, …)
scope of this course

- focus on simple systems (bacterial gene regulation)
- role of theory, modeling, and computation
- coarse-grained description at multiple scales (telescoped description)
- quantitative connections between molecular mechanisms
  and physiological (functional or behavioral) characteristics
- power of functional and mechanistic constraints

❖ course content
  - molecular interactions: ligand-protein, protein-DNA, and protein-protein
  - transcriptional initiation control: activation, repression, and combinatorial
  - post-transcriptional control: attenuation, termination, degradation
  - modeling genetic circuits: bistability and oscillation
  - stochastic gene expression and phenotype
  - growth physiology and metabolic control