Advanced Algorithms and Models for Computational Biology

Introduction to cell biology, genomics, development, and probability

Eric Xing

Lecture 2, January 23, 2006
Reading: Chap. 1, DTM book

Introduction to cell biology, functional genomics, development, etc.
Model Organisms

Bacterial Phage: T4
Bacteria: E. Coli

The Budding Yeast: *Saccharomyces cerevisiae*
The Fission Yeast:  
*Schizosaccharomyces pombe*

- SMALL: ~ 250 µm
- TRANSPARENT
- 959 CELLS
- 300 NEURONS
- SHORT GENERATION TIME
- SIMPLE GROWTH MEDIUM
- SELF- FERTILIZING HERMAPHRODITE
- RAPID ISOLATION AND CLONING OF MULTIPLE TYPES OF MUTANT ORGANISMS

The Nematode:  
*Caenorhabditis elegans*

- SMALL: ~ 250 µm
- TRANSPARENT
- 959 CELLS
- 300 NEURONS
- SHORT GENERATION TIME
- SIMPLE GROWTH MEDIUM
- SELF- FERTILIZING HERMAPHRODITE
- RAPID ISOLATION AND CLONING OF MULTIPLE TYPES OF MUTANT ORGANISMS
The Fruit Fly: Drosophila Melanogaster

Normal

Ubx mutant

The Mouse

transgenic for human growth hormone
Prokaryotic and Eukaryotic Cells

A Close Look of a Eukaryotic Cell

The structure:

The information flow:
A variety of plasma membrane receptor proteins bind extracellular signaling molecules and transmit signals across the membrane to the cell interior.

(c) Tyrosine kinase-linked receptors (erythropoietin, interferons)
Signal Transduction Pathway

Functional Genomics and X-omics
A Multi-resolution View of the Chromosome

DNA Content of Representative Types of Cells

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of base pairs (millions)</th>
<th>Number of encoded proteins</th>
<th>Number of chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PROKARYOTIC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycoplasma genitalum (Bacterium)</td>
<td>0.58</td>
<td>470</td>
<td>1</td>
</tr>
<tr>
<td>Helicobacter pylori (Bacterium)</td>
<td>1.67</td>
<td>1590</td>
<td>1</td>
</tr>
<tr>
<td>Haemophilus influenza (Bacterium)</td>
<td>1.83</td>
<td>1743</td>
<td>1</td>
</tr>
<tr>
<td><strong>EUKARYOTIC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (yeast)</td>
<td>12</td>
<td>5885</td>
<td>17</td>
</tr>
<tr>
<td>Drosophila melanogaster (insect)</td>
<td>165</td>
<td>13,601</td>
<td>4</td>
</tr>
<tr>
<td>Caenorhabditis elegans (worm)</td>
<td>97</td>
<td>19,999</td>
<td>6</td>
</tr>
<tr>
<td>Homo sapiens (human)</td>
<td>2900</td>
<td>30,000 TO 40,000</td>
<td>23</td>
</tr>
<tr>
<td>Arabidopsis thaliana (plant)</td>
<td>125</td>
<td>25,498</td>
<td>10</td>
</tr>
</tbody>
</table>
Functional Genomics

- The various genome projects have yielded the complete DNA sequences of many organisms.
  - E.g. human, mouse, yeast, fruitfly, etc.
  - Human: 3 billion base-pairs, 30-40 thousand genes.
- Challenge: go from sequence to function,
  - i.e., define the role of each gene and understand how the genome functions as a whole.

Regulatory Machinery of Gene Expression

- Diagram of gene regulatory sequences, gene regulatory proteins, TATA box, promoter, start of transcription, general transcription factors, RNA polymerase.
Classical Analysis of Transcription Regulation Interactions

“Gel shift”: electrophoretic mobility shift assay ("EMSA") for DNA-binding proteins

* Protein-DNA complex

* Free DNA probe

Advantage: sensitive
Disadvantage: requires stable complex; little "structural" information about which protein is binding

Modern Analysis of Transcription Regulation Interactions

- Genome-wide Location Analysis (ChIP-chip)

(b)

Advantage: High throughput
Disadvantage: Inaccurate
Gene Regulatory Network

Biological Networks and Systems Biology

Systems Biology:
understanding cellular event under a system-level context
Genome + proteome + lipome + ...

Protein-protein Interaction networks

Metabolic networks
Gene Regulatory Functions in Development

Temporal-spatial Gene Regulation and Regulatory Artifacts

A normal fly

Hopeful monster?
Gene Regulation and Carcinogenesis

The Pathogenesis of Cancer

Normal  BCH  DYS

CIS  SCC
Genetic Engineering: Manipulating the Genome

- Restriction Enzymes, naturally occurring in bacteria, that cut DNA at very specific places.

Recombinant DNA
Transformation

- **Recombinant plasmid**
- **E. coli chromosome**
- **Transformed cell survives**
- **Mix E. coli with plasmids in presence of CaCl$_2$; heat pulse**
- **Culture on nutrient agar plates containing ampicillin**
- **Cells that do not take up plasmid die on ampicillin plates**

Formation of Cell Colony

- **Cell multiplication**
- **Colony of cells, each containing copies of the same recombinant plasmid**
How was Dolly cloned?

- Dolly is an exact genetic replica of another sheep.

2. The making of Dolly

Definitions

- **Recombinant DNA**: Two or more segments of DNA that have been combined by humans into a sequence that does not exist in nature.
- **Cloning**: Making an exact genetic copy. A clone is one of the exact genetic copies.
- **Cloning vector**: Self-replicating agents that serve as vehicles to transfer and replicate genetic material.
Software and Databases

- NCBI/NLM Databases Genbank, PubMed, PDB
  - DNA
  - Protein
  - Protein 3D
  - Literature

Entrez

Introduction to Probability
Basic Probability Theory Concepts

- A **sample space** $\mathcal{S}$ is the set of all possible outcomes of a conceptual or physical, repeatable experiment. ($\mathcal{S}$ can be finite or infinite.)
  - E.g., $\mathcal{S}$ may be the set of all possible nucleotides of a DNA site: $\mathcal{S} = \{A, T, C, G\}$

- A **random variable** is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)
  - E.g., seeing an "A" at a site $\implies X=1$, o/w $X=0$.
  - This describes the true or false outcome a random event.
  - Can we describe richer outcomes in the same way? (i.e., $X=1, 2, 3, 4$, for being A, C, G, T) --- think about what would happen if we take expectation of $X$.

- Unit-Base Random vector
  - $X_\mathcal{S} = [X_A, X_T, X_G, X_C]^T$, $X_\mathcal{S} = [0,0,1,0]^T$ $\implies$ seeing a "G" at site $i$

Basic Prob. Theory Concepts, ctd

- **(In the discrete case)**, a probability distribution $P$ on $\mathcal{S}$ (and hence on the domain of $X$) is an assignment of a non-negative real number $P(s)$ to each $s \in \mathcal{S}$ (or each valid value of $x$) such that $\sum_{s \in \mathcal{S}} P(s) = 1$. ($0 \leq P(s) \leq 1$)
  - intuitively, $P(s)$ corresponds to the frequency (or the likelihood) of getting $s$ in the experiments, if repeated many times
  - call $\theta_j = P(s)$ the parameters in a discrete probability distribution

- A probability distribution on a sample space is sometimes called a **probability model**, in particular if several different distributions are under consideration
  - write models as $\mathcal{M}_j$, $\mathcal{M}_s$, probabilities as $P(X|\mathcal{M}_j)$, $P(X|\mathcal{M}_s)$
  - e.g., $\mathcal{M}_j$ may be the appropriate prob. dist. if $X$ is from "splice site", $\mathcal{M}_s$ is for the "background".
  - $\mathcal{M}$ is usually a two-tuple of (dist. family, dist. parameters)
Discrete Distributions

- Bernoulli distribution: Ber(\(p\))
  \[ P(X) = \begin{cases} 1 - p & \text{for } X = 0 \\ p & \text{for } X = 1 \end{cases} \implies P(X) = p^X (1 - p)^{1-X} \]

- Multinomial distribution: Mult(1, \(\theta\))
  - Multinomial (indicator) variable: \(X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}\), where \(X_j = 1\) w.p. \(\theta_j\), \(\sum_{j=1}^k \theta_j = 1\).
  \[ P(X) = \prod_{j=1}^k \theta_j^{X_j} = \theta^X \]

- Multinomial distribution: Mult(\(n\), \(\theta\))
  - Count variable: \(X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}\), where \(\sum_{j=1}^k X_j = n\)
  \[ P(X) = \frac{n!}{X_1!X_2!\cdots X_k!} \theta_1^{X_1} \theta_2^{X_2} \cdots \theta_k^{X_k} \]

Basic Prob. Theory Concepts, ctd

- A continuous random variable \(X\) can assume any value in an interval on the real line or in a region in a high dimensional space
  - \(X\) usually corresponds to a real-valued measurements of some property, e.g., length, position, …
  - It is not possible to talk about the probability of the random variable assuming a particular value \(\implies P(X) = 0\)
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval
    \[ P(X \in [x_1, x_2]) \]
  \[ P(X < x) = P(X \in (-\infty, x]) \]

- The probability of the random variable assuming a value within some given interval from \(x_1\) to \(x_2\) is defined to be the area under the graph of the probability density function between \(x_1\) and \(x_2\).
  - Probability mass: \(P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x) dx\), note that \(\int_{-\infty}^{\infty} p(x) dx = 1\).
  - Cumulative distribution function (CDF): \(P(X) = P(X < x) = \int_{-\infty}^{x} p(x) dx\)
  - Probability density function (PDF): \(p(x) = \frac{d}{dx} P(x)\)
Continuous Distributions

- Uniform Probability Density Function
  \[ p(x) = \frac{1}{b-a} \quad \text{for} \quad a \leq x \leq b \]
  \[ = 0 \quad \text{elsewhere} \]

- Normal Probability Density Function
  \[ p(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
  - The distribution is symmetric, and is often illustrated as a bell-shaped curve.
  - Two parameters, \( \mu \) (mean) and \( \sigma \) (standard deviation), determine the location and shape of the distribution.
  - The highest point on the normal curve is at the mean, which is also the median and mode.
  - The mean can be any numerical value: negative, zero, or positive.

- Exponential Probability Distribution
  - density: \( p(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \),
  - CDF: \( P(X \leq x_0) = 1 - e^{-\frac{x_0}{\mu}} \)

Statistical Characterizations

- **Expectation**: the center of mass, mean value, first moment:
  \[ E(X) = \begin{cases} 
  \sum_{i=1}^{n} x_i p(x_i) & \text{discrete} \\
  \int x p(x) dx & \text{continuous} 
  \end{cases} \]
  - Sample mean:
    \[ \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \]

- **Variance**: the spreadness, second moment:
  \[ Var(X) = \begin{cases} 
  \sum_{i=5}^{n} (x_i - E(X))^2 p(x_i) & \text{discrete} \\
  \int (x - E(X))^2 p(x) dx & \text{continuous} 
  \end{cases} \]
  - Sample variance
    \[ \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \]
Basic Prob. Theory Concepts, ctd

- **Joint probability:**
  - For events $E$ (i.e. $X=x$) and $H$ (say, $Y=y$), the probability of both events are true:
    \[ P(E \text{ and } H) := P(x,y) \]

- **Conditional probability**
  - The probability of $E$ is true given outcome of $H$
    \[ P(E \text{ and } H) := P(x|y) \]

- **Marginal probability**
  - The probability of $E$ is true regardless of the outcome of $H$
    \[ P(E) := P(x) = \sum_y P(x,y) \]

- Putting everything together:
  \[ P(x|y) = \frac{P(x,y)}{P(y)} \]

---

Independence and Conditional Independence

- Recall that for events $E$ (i.e. $X=x$) and $H$ (say, $Y=y$), the conditional probability of $E$ given $H$, written as $P(E|H)$, is
  \[ P(E \text{ and } H)/P(H) \]
  (= the probability of both $E$ and $H$ are true, given $H$ is true)

- $E$ and $H$ are (statistically) independent if
  \[ P(E) = P(E|H) \]
  (i.e., prob. $E$ is true doesn't depend on whether $H$ is true); or equivalently
  \[ P(E \text{ and } H) = P(E)P(H). \]

- $E$ and $F$ are conditionally independent given $H$ if
  \[ P(E|H,F) = P(E|H) \]
  or equivalently
  \[ P(E,F|H) = P(E|H)P(F|H) \]
Representing multivariate dist.

- Joint probability dist. on multiple variables:
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1)P(X_2 | X_1)P(X_3 | X_2)P(X_4 | X_1, X_2)P(X_5 | X_1, X_2, X_3)P(X_6 | X_1, X_2, X_3, X_4) \]

- If \( X_i \)'s are independent: \( P(X_i | \cdot) = P(X_i) \)
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)P(X_6) = \prod_i P(X_i) \]

- If \( X_i \)'s are conditionally independent, the joint can be factored to simpler products, e.g.,
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1)P(X_2 | X_1)P(X_3 | X_2)P(X_4 | X_1, X_2)P(X_5 | X_1, X_2, X_3)P(X_6 | X_1, X_2, X_3, X_4) \]

- The **Graphical Model** representation

---

The Bayesian Theory

- The Bayesian Theory: (e.g., for date \( D \) and model \( M \))

  \[ P(M|D) = \frac{P(D|M)P(M)}{P(D)} \]

  - the posterior equals to the likelihood times the prior, up to a constant.

  - This allows us to capture uncertainty about the model in a principled way