

# Gene Expression and Their Computational Analysis

## An Overview of Gene Expression

## Gene Expression

- A biological cell's function is thought of as determined by the proteins that are "expressed" in it
  - Encoding (DNA) is the same for every cell
    - Four different molecules: A, T, C, G
    - A is complementary to T
    - C is complementary to G
    - DNA molecules are polarized and thus have directions; the 5'-end and the 3'-end
  - Watson-Crick Complementarity
    - DNA sequence and its reverse complement are paired

## DNA

- A DNA molecule is a phosphate attached on a sugar backbone
- The numbers assigned to the carbon molecules of the backbone indicate direction; they are charged molecules: 5'-end and 3'-end.
- There are four molecules, A, C, G, and T.

## Complementarity of DNA

- A is complementary to T; C is to G
- 5'-AAAACCCCGTGTGT-3' is complementary anti-parallel to 5'-ACACACGGGGTTTT-3' and they hybridize with each other to form:



## RNA-DNA

- There are four molecules: A, C, G, and U.
- DNA 5'-AAAACCCCGTGTGT-3' is complementary anti-parallel to RNA 5'-ACACACGGGGUUUU-3' and they hybridize with each other to form:



## Gene Expression

- The protein synthesis process consists of:
  - DNA -> RNA (transcription)
    - Four different molecules: A, U, C, G
  - RNA -> Protein (translation)
    - Consecutive triples (codon) of RNA encode one amino acid molecule:
- Genes contain parts that are not used in encoding of proteins, which are removed during transcription to RNA

## Gene Expression

- *Ideally*, if you can identify all the proteins that are “expressed” in a cell and count them, you get a pretty good idea of what’s going on in the cell
  - Factors exist that control expression after transcription, e.g., microRNA, so now it is known that the quantity of RNA may not directly correlate with the quantity of protein

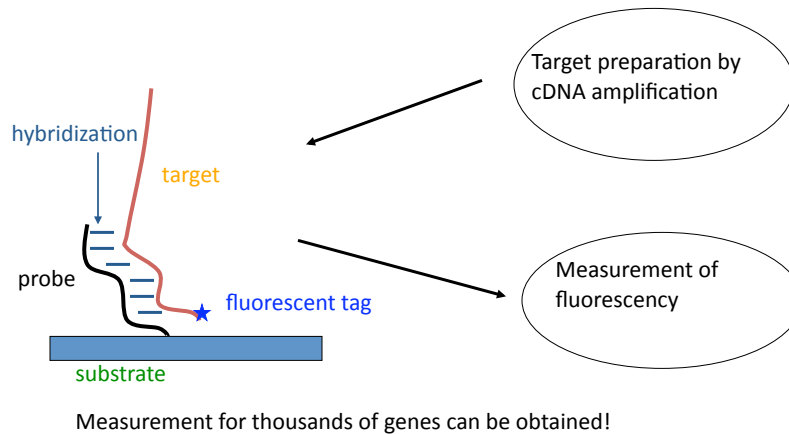
## Capturing RNA

- Proteins are hard to capture and there aren't many molecules
- RNA molecules can be transcribed backed to DNA; DNA can be amplified; DNA can be dyed and captured

## Array Technologies

- Separation of RNA from cells
- Creation of DNA whose transcription matches the RNA ... called cDNA (copy DNA)
  - RNA reverse transcriptase
- Creation of many copies of the DNA (PCR ... Polymerase Chain Reaction)
  - There are enzymes that are able to assemble reverse complement of DNA sequences
- Dye the copy DNA sequences
- Trap the sequences on probes tethered on a glass surface
- Measure the reflection of light at certain frequencies

## Microarray Technology... mRNA Quantification



3/24/09

University of Miami

11

## Various Gene Expression Quantification Methods

- Whole sequence methods (cDNA arrays; two-color systems)
  - Use the entire cDNA sequence as the probe
    - Must know the exact sequence
    - Lower density of probes on surface so as to avoid probe-probe interactions
- Sequence fragment methods (Affymetrics arrays; Agilent arrays)
  - Design a collection of probes
    - A probe might be shared among genes
    - The set as a whole is unique to its corresponding gene

## Important Points about Gene Expression Data

- The values are not absolute; but are relative
  - Within a single array data, the values can be compared against each other
  - For multiple array data, the values can be compared assuming that the total amount is equal
  - Can assume that a certain group of genes has consistent expression (“housekeeping genes”)
- Much room for errors
  - The speed of PCR is dependent on the base length of DNA
  - Replication error
  - Human error
  - Expression is probabilistic

## Important Points (cont'd)

- Gene expression of one sample is a vector  $A = (a_1, \dots, a_n)$  where the indices 1 through n correspond to the genes of interest
- Gene expression of multiple samples is a collection of vectors  $A_i = (a_{i1}, \dots, a_{in})$ ,  $1 \leq i \leq m$ , where m is the number of samples

## Exploration of Gene Expression Data

- Fundamental data layout ... principal component analysis
- Identifying genes that exhibit similar expression patterns ... clustering
- Correlating gene expression with a certain cell type/condition ... classification

## Principal Component Analysis

## Principal Component Analysis

- Understanding underlying structure of the data
- Covariance matrix  $C$  of data  $A_i = (a_{i1}, \dots, a_{in})$  where  $1 \leq i \leq m$ :
  - The  $[pq]$  entry is the average of  $(a_{ip} - x_p)(a_{iq} - x_q)$  of all  $i$ ; where  $x_p$  is the average of  $a_{kp}$ ,  $1 \leq k \leq m$
  - An eigenvector is a vector  $V = (v_1, \dots, v_n)$  such that  $CV = \lambda v$  for some constant  $\lambda > 0$ . Since  $C$  is symmetric (the  $[pq]$  entry is identical to the  $[qp]$  entry), such vectors do exist.
  - Normalize the vector so that its norm (the square root of the sum of the squares of the entries) is 1.

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{1n} \\ c_{21} & c_{22} & c_{2n} \\ c_{n1} & c_{n2} & c_{nn} \end{pmatrix}$$

$$c_{pq} = c_{qp} = \frac{1}{n} \sum_{i=1}^n \left( a_{pi} - \frac{1}{n} \sum_{k=1}^n a_{pk} \right) \left( a_{qi} - \frac{1}{n} \sum_{k=1}^n a_{qk} \right)$$

$$(v_1 \quad \dots \quad v_n)C = \left( \sum_{k=1}^n c_{k1}v_1 \quad \dots \quad \sum_{k=1}^n c_{kn}v_n \right) = (\lambda v_1 \quad \dots \quad \lambda v_n)$$

## PCA (cont'd)

- The importance of an eigenvector is measured by the value  $\lambda$ .
- Each input data is decomposed into a weighted sum of the eigenvectors.
- Given a data  $A = (a_1, \dots, a_n)$  and an eigenvector  $V = (v_1, \dots, v_n)$ ,  $V$ 's weight in  $A$  is defined to be the sum of  $v_i a_i$  for all values of  $i$ . This is called the projection of  $A$  onto  $V$ .

## PCA (cont'd)

- Use the first two or three components
- Project the data on each component and use the projection to plot the data
- May delineate underlying data structure