

# Summer Institute for Training in Biostatistics - 2005

## Lecture II: Statistics and Genomics and Introduction to Categorical Data Analysis

Sündüz Keleş

Department of Statistics and of Biostatistics & Medical Informatics

## Statistics and Genomics

- DNA microarray experiments. Relate biological and clinical outcomes to genome-wide measures of transcript (mRNA) levels, DNA copy number.
  - Predict survival from transcript levels or DNA copy number for thousands of genes.
  - Relate sequence motifs in regulatory control regions to transcript levels.
- Genetic networks. Infer gene-gene relationships (e.g., based on microarray expression measures, co-citation in literature), protein-protein relationships, gene-protein relationships.
- Genetic mapping using single nucleotide polymorphisms (SNP). Relate biological and clinical outcomes (phenotypes) to genotypes at thousands of genetic markers  $\implies$  Which genes affect phenotypes of interest?

## Statistics and Genomics

**DNA.** Sequencing, assembly, alignment, annotation, gene finding, **TF binding site detection**, Phylogeny, Polymorphism.

**RNA.** Structure prediction, splice site detection, **transcript levels**.

**Protein.** Alignment, phylogeny, structure prediction, functional pathways.

Additionally, biological/clinical covariates and outcomes **biological meta data**.

## Statistics and Genomics

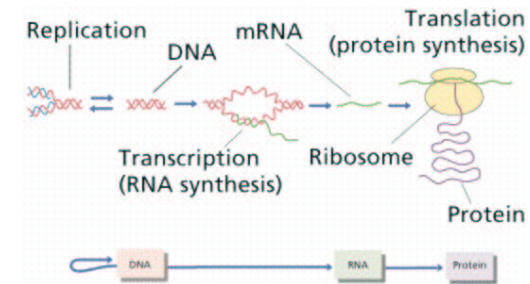
- Identification of regulatory motifs in DNA sequences. From unaligned DNA sequence data, identify start and base composition of transcription factor binding sites.
- Protein structure prediction. Predict 3-D structure of proteins from 1-D amino acid sequence.
- Drug discovery. Screening chemical compounds based on thousands of variables corresponding to chemical properties of compounds (e.g., molecular weight, number of filled orbitals). Which compounds bind to a particular target?

## Statistics and Genomics

Current problems in genomics involve analysis and comprehension of large multivariate data sets. Statistical and computational methods are essential for the reliable and efficient analysis of these large multivariate datasets.

C. Tilstone (2003). DNA microarrays: Vital statistics. *Nature*, 424: 610-612.

## Gene expression: the flow of genetic information from DNA via RNA to Protein



↔ Behind the scenes: **GENE REGULATION!**

## Regulatory motif finding

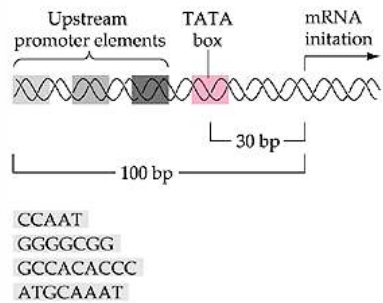
- Regulation of gene expression eukaryotes: DNA binding proteins (transcription factors) and their binding sites.
- Several data and approaches are available for this purpose now: based on (1) **single species sequence data**; (2) **gene expression and single species sequence data**; (3) multiple-species sequence data and gene expression from a single species; (4) genome tiling arrays (ChIP-chip data).

## Eukaryotic gene regulation

### Regulatory components:

- Specific DNA sequences in the vicinity of the gene to be regulated (*cis-acting elements*, e.g. promoters, enhancers, more generally regulatory motifs.),
- DNA binding proteins (transcription factors) encoded by other genes (*trans-acting elements*).

Transcription factor binding sites (**regulatory motifs**) are short sequences (5-25 base pairs) from the alphabet  $\{A, C, G, T\}$ .



## How to represent binding sites?

- **Consensus sequences**  
e.g. **CGGNNNNNNNNNNNCGG**,  $N \in \{A, C, G, T\}$  for GAL4.
- **Position weight matrices (PWMs or PSSMs (Position specific scoring matrices))**: A 4 by *width of the motif* matrix where position  $(j, w)$  represents the frequency of observing nucleotide  $j$  at position  $w$  of the DNA motif (Stormo, 1982).  
Corresponds to fitting an independent multinomial model at each position of the binding site.

## Challenges of regulatory motif identification:

- Variable location of the sites: upstream, downstream, within introns.
- Degeneracy of the sites.

### GAL4 binding sites for different yeast genes (from SCPD):

>YBR019C **CGGCGATACCTTCACCG**

>YBR020W **CGGGCGACGATTACCG**

>YLR081W **CGGAGCGTAGGCGGCCG**

## Multinomial distribution

The multinomial distribution is an extension of the binomial distribution involving joint probabilities (a generalization from 2 outcomes to  $k$  outcomes).

Consider a statistical experiment, e.g., throwing a dice  $n$  times, with  $k$  possible outcomes  $E_1 \cdots E_k$ , with respective probabilities  $p_1, p_2, \dots, p_k$ . The multinomial distribution is the joint probability distribution of the set of random variables  $X_1, \dots, X_k$ , where  $X_i$  is the number of occurrences of  $E_i$ . It has a probability mass function of the form

$$Pr(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

where  $\sum_{i=1}^k x_i = n$ ,  $\sum_{i=1}^k p_i = 1$ .

## Multinomial distribution

The first term represents the number of ways to distribute  $x_1$  outcomes of  $E_1$ ,  $x_2$  outcomes of  $E_2$ ,  $\dots$ ,  $x_k$  outcomes of  $E_k$  among  $n$  trials.

The term  $p_1^{x_1} \dots p_k^{x_k}$  is the probability that there are  $x_1$  outcomes of  $E_1$ ,  $x_2$  outcomes of  $E_2$ ,  $\dots$ ,  $x_k$  outcomes of  $E_k$ .

The product of these two terms is the probability that in  $n$  trials, there are  $x_1$  outcomes of  $E_1$ ,  $x_2$  outcomes of  $E_2$ ,  $\dots$ ,  $x_k$  outcomes of  $E_k$ .

## Fitting a multinomial distribution to one position of the binding site

Say we have  $n$  examples of a length  $w = 7$  binding site. We imagine that each position of the binding site is generated according to the following experiment:

For position  $i$ , we have a die with 4 sides representing  $A$ ,  $C$ ,  $G$ , and  $T$  with probabilities  $p_i(A), p_i(C), p_i(G), p_i(T)$ . We roll the corresponding die  $n = 7$  times and record the outcomes. Say, we observe  $T$  6 times and  $G$  1 time. Based on these data, we *estimate* the base probabilities for this position to be  $\hat{p}_i(A) = \hat{p}_i(C) = 0$ ,  $\hat{p}_i(G) = 1/7$ , and  $\hat{p}_i(T) = 6/7$ .

## Multinomial distribution assumptions

- Individual trials are independent.
- Outcomes are mutually exclusive and all inclusive.

## Position weight matrices (PWMs)

**E.g.** Position weight matrix of GAL4:

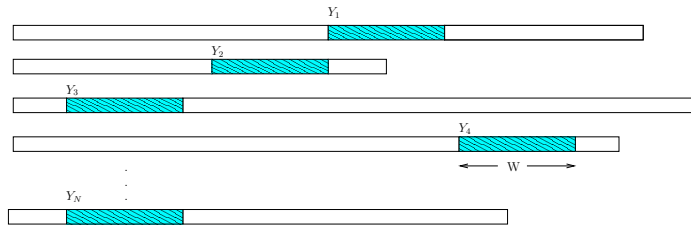
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0	0	0	1/4	.	.	.	.	.	.	.	.	.	1/4	0	0	0
C	1	0	0	1/4	.	.	.	.	.	.	.	.	.	1/4	1	0	0
G	0	1	1	1/4	.	.	.	.	.	.	.	.	.	1/4	0	1	1
T	0	0	0	1/4	.	.	.	.	.	.	.	.	.	1/4	0	0	0

⇒ An independent multinomial distribution for each position.

⇒ Forms the basis of commonly used motif finding methods based on multinomial mixture models (Lawrence & Reilly (1990), MEME of Bailey & Elkan (1994)).

## Sequence data

$N$  *unaligned* sequences  $\vec{X}_i = (X_{i,1}, \dots, X_{i,L_i})$ ,  $i = 1, \dots, N$ , where  $L_i$  is the length of the  $i$ th sequence.



⇒ The start positions  $Y_i$  are hidden!

## Data and Multinomial mixture models

$N$  *unaligned* sequences  $\vec{X}_i = (X_{i,1}, \dots, X_{i,L_i})$ ,  $i = 1, \dots, N$ , where  $L_i$  is the length of the  $i$ th sequence and  $X_{i,l} \in \{A, C, G, T\}$ .

**Full data.**  $\{(\vec{X}_1, \vec{Y}_1), (\vec{X}_2, \vec{Y}_2), \dots, (\vec{X}_N, \vec{Y}_N)\}$ , where  $\vec{Y}_i = (Y_{i1}, \dots, Y_{i,L_i-W+1})$  and  $Y_{i,l} = I(\text{motif starts at position } l \text{ in sequence } i)$ .

**Observed data.**  $\{\vec{X}_1, \vec{X}_2, \dots, \vec{X}_N\}$ .

Common approaches are based on two component **multinomial mixture models** (Lawrence & Reilly (1990), MEME of Bailey & Elkan (1994)).

## Statistical question

Given a set of genes that are potentially co-regulated, identify shared regulatory motif(s) from their unaligned upstream control regions, i.e., in regions roughly 600-1,000 base pairs from the gene start site in lower eukaryotes such as yeast.

- **Start site** of the motif in each sequence.
- Base composition of the motif, i.e., **nucleotide frequencies** at each position of the motif.
- **Width** of the motif.
- Incorporation of the available **structural information**, i.e., supervising the search. [Advance]

## Multinomial mixture model for sequence data

**One motif per sequence model:** Lawrence and Reilly (1990); *OOPS* model of Bailey and Elkan's MEME (1994).

- Sites are distributed independently with

$$P_0 = (p_{01}, \dots, p_{04}) \quad \text{for background sites,}$$

$$P_w = (p_{w1}, \dots, p_{w4}) \quad \text{for position } w \text{ in the motif, } w \in \{1, \dots, W\}$$

- Only one motif per sequence, i.e.,  $\sum_l Y_{il} = 1$ .
- Uniform start site distribution.  $P(Y_{il} = 1) = 1/(L_i - W + 1)$ .

## Multinomial mixture model for sequence data

Extension: Zero or one motif per sequence model: ZOOPS model of Bailey and Elkan's MEME (1994). Introduces another hidden random variable:

$$Z_i = \begin{cases} 1 & \text{if sequence } i \text{ has one copy of the motif } \sum_l Y_{il} = 1 \\ 0 & \text{if sequence } i \text{ has no copy of the motif } \sum_l Y_{il} = 0 \end{cases}$$

Extension: TCM model of MEME: Allows any number of occurrences per sequence.

⇒ Model parameters,  $\theta = (P_0, P_1, \dots, P_W)$  in *OOPS* and *ZOOPS* are estimated with the **maximum likelihood** method. Various implementations use the **EM algorithm** or **Gibbs sampling**.

⇒ *TCM* model is a **heuristic**, applies *ad hoc* modifications to the M-step of the EM algorithm.

## Extracting upstream regulatory sequences

- *Saccharomyces* Genome Database:  
<http://www.yeastgenome.org/>.
- Regulatory Sequence Analysis Tools (RSAT):  
<http://rsat.ulb.ac.be/rsat/>.

## An example from baker's yeast

Our biologist collaborator gives us the following gene IDs: YCL027W, YDR461W, YFL026W, YNL145W; and tells us that these genes are expressed under very similar conditions and are possibly regulated by the transcription factor STE12. The scientific question of interest whether we can identify a common binding site in upstream control regions of these genes.

**Tasks:** (1) Extract regulatory control regions of these genes; (2) Identify common sequence elements in these regions.

## Find regulatory motifs using the multinomial mixture model

- A popular implementation is known as MEME at  
<http://meme.nbcr.net/meme/website/intro.html>.

## Sequence logos

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position.

In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

$$I_j = \log_2 4 - \sum_{i \in \{A,C,G,T\}} p_j(i) \log_2 p_j(i)$$

is the information at  $j$ -th position.

## Introduction to categorical data analysis

Chi-square goodness-of-fit test and Chi-square test of independence for contingency tables [On the board].

## Sequence logos

A web based application for generating sequence logos is available at <http://weblogo.berkeley.edu/logo.cgi>

Generating the sequence logo of the first candidate motif:

```
> MFA2
ATGAAAC
> STE2
ATGAAAC
> STE2
ATGAAAC
>STE2
ATGAAAC
>MFA1
ATGAAAC
> FUS1
ATGAAAC
> FUS1
ATGAAAC
> MFA2
AGGAAAC
```

## Example: Chi-square goodness-of-fit test

```
> chisq.test(c(6, 5, 46, 43), p = c(1/16, 1/16, 7/16, 7/16))
```

Chi-squared test for given probabilities

```
data: c(6, 5, 46, 43)
```

```
X-squared = 0.3886, df = 3, p-value = 0.9426
```

Cannot reject the null hypothesis that the base composition at the 6th position is  $p(A) = p(C) = 1/16$ ,  $p(T) = p(G) = 7/16$ .

### Example: Chi-square test of homogeneity

```
> chisq.test(matrix(c(71, 34, 45, 42), 2, 2), correct = F)
```

Pearson's Chi-squared test

```
data: matrix(c(71, 34, 45, 42), 2, 2)
```

```
X-squared = 5.0264, df = 1, p-value = 0.02496
```

Reject the null hypothesis that there is no association between the motif occurrence and differential expression. We can further recommend the biologist to experimentally verify the association.

## References

### Microarrays

- Schulze A & Downward J. (2001) *Navigating gene expression using microarrays a technology review*. Nature Cell Biology. 3:E190-E195. [http://www.nature.com/ncb/journal/v3/n8/full/ncb0801\\_e190.html](http://www.nature.com/ncb/journal/v3/n8/full/ncb0801_e190.html)
- [Next lecture] Golub et al. (1999) *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science. 1999 Oct 15;286(5439):531-7. <http://www.sciencemag.org/cgi/content/full/286/5439/531>

### Statistics

- Nolan & Speed (2000). *Mathematical Statistics through Applications*. [Chapter 4, Appendix B]

### GO Annotations

- [Next lecture] Biconductor Tutorial, Part II, MGED6 (<http://www.bioconductor.org/workshops/MGED6/MGED6IIx6.pdf>).