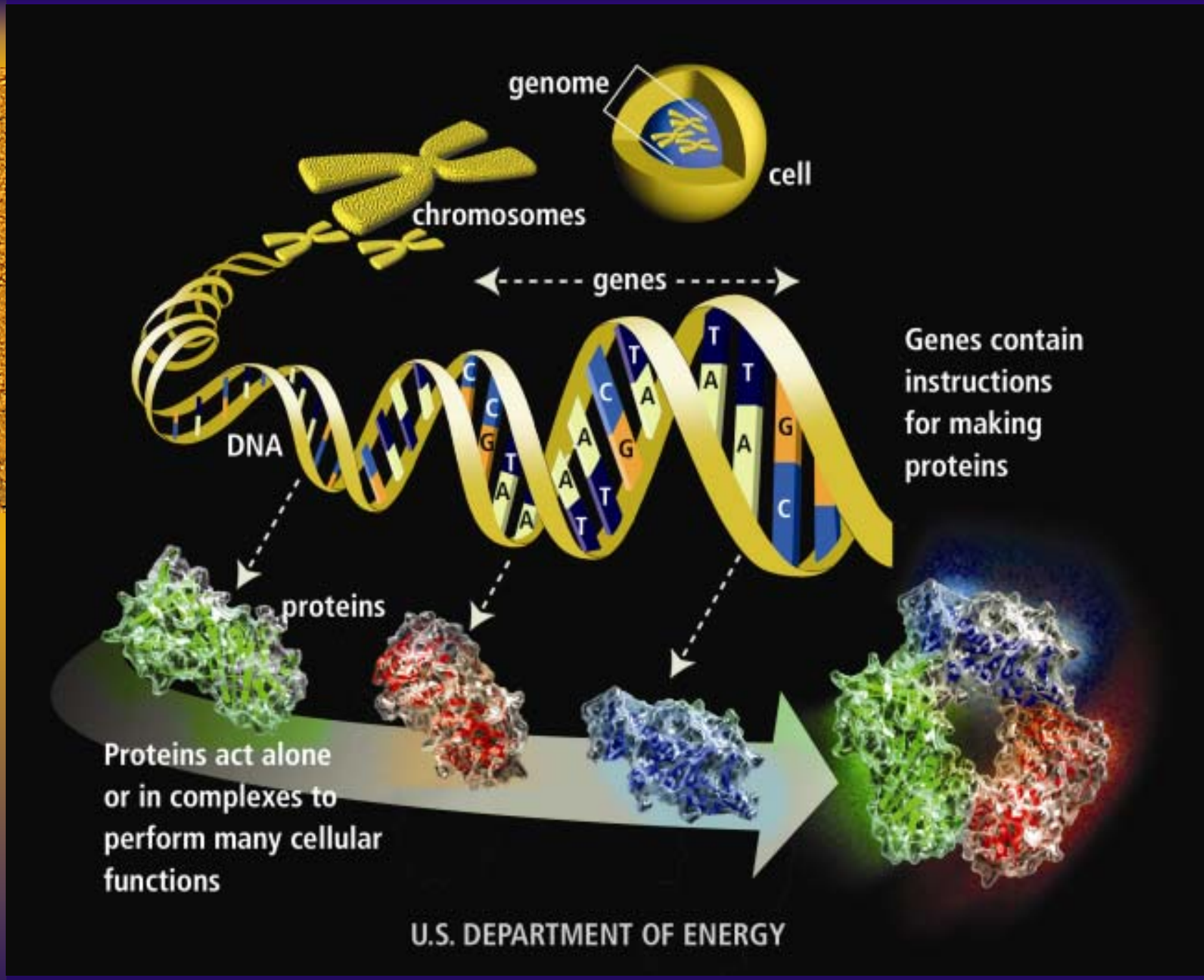
Experimental Design for Gene Expression Microarray

Jing Yi

18 Nov, 2002


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

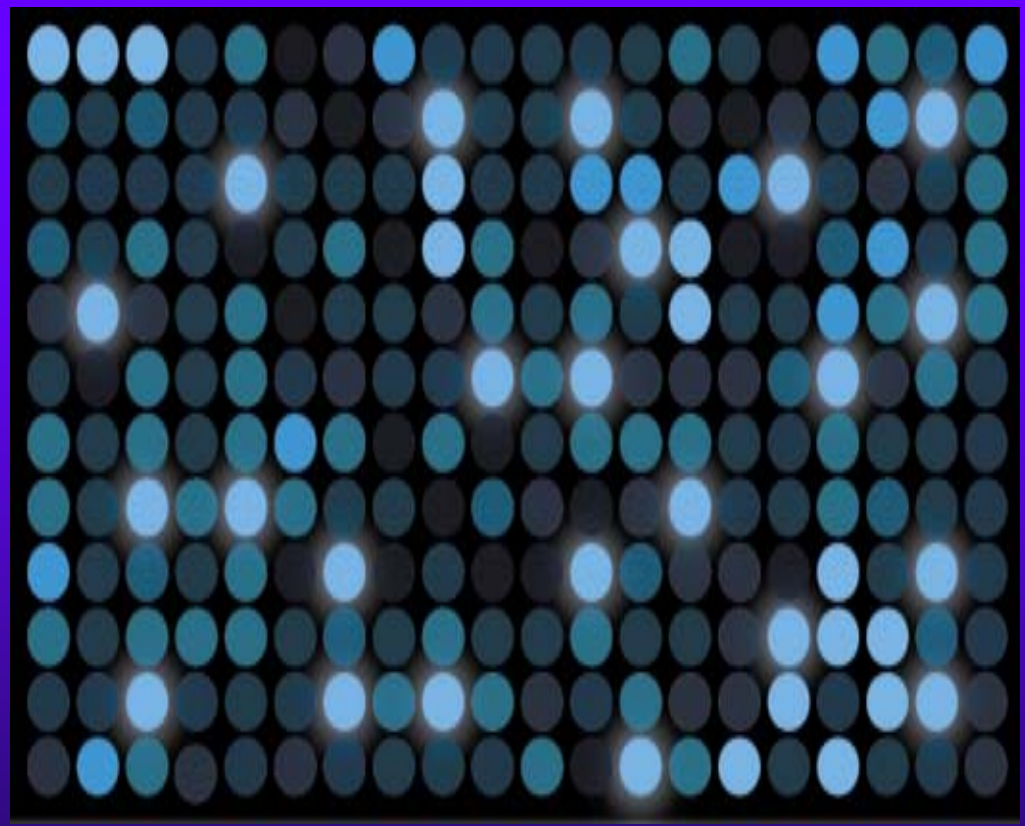
Human Genome Project

- The HGP continued emphasis is on obtaining by 2003 a complete and highly accurate reference sequence(1 error in 10,000 bases).
- The total number of the genes is estimated at 30,000 to 40,000, much lower than previous estimated 80,000 to 140,000.
- The human genome contains 3164.7 million chemical nucleotide bases.(A,T,C,G).

- The order of almost all(99.9%)nucleotide bases is exactly the same in all people.
- The functions are unknown for more than 50% of the discovered genes.
- The world is really mysterious , tantalizing and unexplored.

Exploring the New World of the Genome with DNA Microarrays

- The genome project has revitalized exploration in the biological research.
- DNA microarrays provide a simple and natural vehicle for exploring the genome in a way both systematic and comprehensive.
- Exploration of the genome using the DNA microarray should narrow the gap in our knowledge of gene function and molecular biology between the currently favored model organisms and other species.



Spotted cDNA Microarray

- Samples of DNA clones with known sequence content are spotted and immobilized onto a glass slide or other substrate.
- Pools of purified mRNA from cell populations are reverse-transcribed into cDNA and labeled with one of the two fluorescent dyes: “Red” and “ Green”.
- They are then mixed and hybridized with the arrayed DNA spots.

Why DNA Microarray?

- It is (relatively) cheap. The marginal cost per copy of the yeast genome microarray is about $20.
- It is flexible and universal.
- It is fast. The total time required to print 150 copies of an array of 12,000 genes is about a day.
- It is user-friendly: convenient, solid format of microarray slide and non-radioactive,non-toxic hybridization solution.

Beginning Projected Use of Microarray Technology

- The first experiment with microarrays were time-series studies:

  General idea: when a gene of unknown function ends up in a cluster of genes with known function,one has a valuable clue as to the function of the unknown gene.

- Clustering ideas similarly have been used to classify tissue samples according to their global patterns of gene expression.

  Perou et al.(1999)uses gene expression to classify human breast cancer……

Two Main Factors in Microarray Experiment Designs

We will only discuss two-color cDNA microarray experiment here.

- The design of array itself: Decide which DNA probes are to be printed on the solid substrate.
- The allocation of mRNA samples to the microarrays.

The First Aspect

The choice of which DNA probes to print onto the solid substrate usually is determined:

- By the cDNA libraries(i.e the collections of cDNA clones)available to them.
- By the genes whose expression levels the biologist wants to measure.

Use of Controls

Advice is sometimes sought from statistics on the use of controls:

- Negative controls: blank spots;spots with cDNA from very different species,….
- Positive controls: “housekeeping” gene is ubiquitously expressed at more or less constant level.
- Hybridization : success or failure?
- Normalization: microarray sample pool; ScoreCard system.

Sources of Variation in Microarray Experiments

- The simplest microarray experiment look for changes in gene expression across a single factor of interest.
- Four basic experiment factors:

Varieties(V) – the categories of a factor of interest

Genes(G) – spotted sequences

Dyes(D)

Arrays(A)

Sources of Variation in Microarray Experiments - II

- Array-gene interaction(AG): “spot”effects.
- Dye-gene effects(DG): arise if there are differences in the dyes that are gene-specific.
- Variety-gene interaction(VG):reflect difference in expression for particular variety and gene combination.

  ---Identifying genes whose expression changes in different varieties means identifying non-zero differences in VG effects.

Graphical Representation

Multi-digraph: a directed graph with multiple edges as shown below:

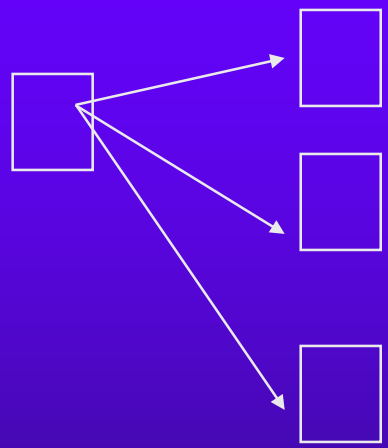
- Vertices or nodes: target mRNA samples
- Edges or arrows: hybridizations between two mRNA samples.
- Green-labeled sample at the tail; Red-labeled at the head of the arrow.

Two Advantage of Graphical Design Illustration

- They quickly and clearly communicate the setup of a design.
- They allow one to easily evaluate certain design properties.

Design comparisons

The "Reference"design: studying v varieties of a factor of interest.



Use one dye to label the reference variety and the other dye to label the varieties of interest.

Drawback of the Reference Design

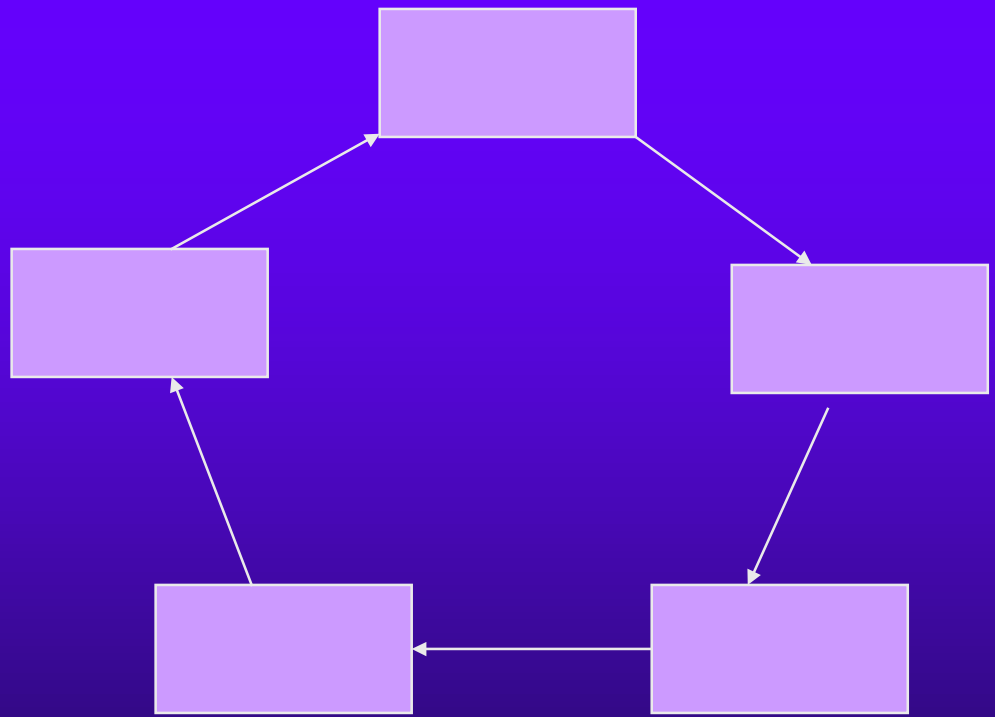
- Variety effects are completely confounded with dye effects.
  - ---The effect of interest VG is completely confounded with DG effects.
- We need to assume there are no gene-specific dye effects.
- No degrees of freedom remain to estimate error.

Degree of Freedom

V varieties and n genes produces 2vn observations. The mean and the array, variety and gene main effects account for 2v+(n-1) degrees of freedom. VG accounts for v(n-1) degrees of freedom. If AG effects must be accounted for, they comprise the final (v-1)(n-1) degrees of freedom. No degree of freedom remain to estimate the error.

Loop Design

V=5 varieties



About the Loop Design

- Using the same number of arrays as the reference design, loop collects twice the data on the varieties of interest.
- Each variety is labeled once with the red and green dyes,so VG effects are not confounded with DG effects.
- n-1 degree of freedom provides a basis for estimating error variation.

Drawback of the Loop Design

- Each sample must be labeled with both the red and green dyes, effectively doubling the number of labeling reactions.
- Because microarray technology is new, now the opinion is this extra effort is worthwhile.

Model Assumptions

- There exists a transformation of microarray data on which the effects are additive.(e.g. log scale)

- The same set of genes is spotted on each array in an experiment:

  ---gene effects are orthogonal to all effects of these factors.

  ---two group of effects:

  "Global": involve A,D and V.

  "Gene-specific": involve G.

ANOVA Model

Each gene is spotted only once per array in the three models

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + \varepsilon_{ijkg}$$

Including the array-gene effects:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + \varepsilon_{ijkg}$$

Including the dye-gene effects:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg}$$

Contrast of Interest

- Microarrays are useful for studying the relative expression of genes across samples.
- The effects of interest are VG interaction. Specially the contrasts of interest:

$$(VG)_{k_1 g} - (VG)_{k_2 g}$$

for fixed gene g and pairs of varieties k1~=k2.

- The least-squares estimate(in first model) is:

$$\frac{n-1}{n}\left(\frac{1}{r_{k_1}} + \frac{1}{r_{k_2}}\right)\sigma^2$$

Which model is better?

- The first model is inadequate because of spot-spot variation on arrays. The other two account for this with AG effects.
- "Even" design: the degree of every node is even in the graphical representation.

  From Euler's theorem, every even graph has a circuit that traverses every edge exactly once.
- If VG and DG are orthogonal, the problem of choosing a good design considering the third model reduces to the problem considering the second.

Recommendations

- Choose an even design so that varieties can be balanced w.r.t dyes.
- Among even designs, look for a design that is efficient for comparing gene expression across varieties while accounting for spot-spot variation.
- Balance and replication.Keep in mind!!!

Fundamental Principles of Good Design

- **Balance** ensures the effects of interest are not confounded with other sources of variation.
- **Replication** improves the precison of estimates and provides degree of freedom for error estimation.

(Fisher, 1951)

General A-Optimality

When suppose comparisons between all pairs of varieties are of equal interest, a reasonable criterion for evaluating design is A-Optimality:

$$\frac{1}{\binom{v}{2}} \sum_{k_1 \neq k_2} \text{var}\ ((V\hat{G})_{k_1 g} - (V\hat{G})_{k_2 g})$$

This criterion favors design minimize the average variance of a contrast of interest.

- Forming the information matrix and get its eigenvalues μ , the A-optimality criterion becomes:

$$\frac{n-1}{n}\,\frac{2\sigma^2}{v-1}\sum_{i=2}^{v}\frac{1}{\mu_i}$$

- Result: Under A-Optimality, loop designs are more efficient than common reference designs.

Summary

- We introduced the linear models as a starting point for studying microarray experimental design. More general assumption of gene-dependent var…..
- We treat all effects are fixed.
- We want to connect microarray experimental design with classical results.
- We have used A-Optimality criterion to evaluate designs.

Reference

- M.Kathleen Kerr, Gary A. Churchill.Experimental design for Gene Microarrays.
- Patrick O. Brown, David Botstein. Exploring the new world of the genome with DNA microarrays.
- http://www.ornl.gov/hgmis
- Yee Hwa Yang, Terence P. Speed. Design and Analysis of Comparative Microarray Experiments