

Statistical Methods in Bioinformatics

CS 594/680

Arnold M. Saxton

Department of Animal Science

UT Institute of Agriculture

Bioinformatics:

Interaction of

Biology/Genetics/Evolution/Genomics

Computer Science/Algorithms/Database

Statistics/Probability

Purpose: address complex biological questions involving large amounts/types of information.

Statistics:

Theory and methods for making scientific decisions from noisy data.

Scientific Method:

Create hypothesis

Ho: $t_1 = t_2$

Design experiment

RBD split-plot

Collect data on
population

Interpret

statistical analysis

Probability:

Closely related to statistics, but different.

Consider

Bayesian networks

Blast sequence analysis

Human genome

Replicated observations are missing.

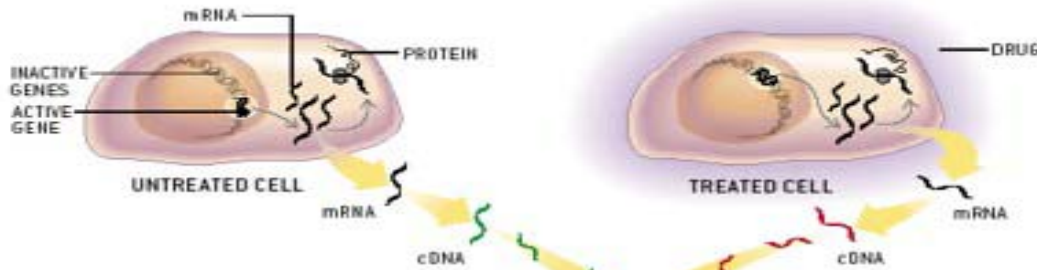
No inferences to a biological population.

[Pietro Lio 2003 Bioessays 25:266-273.]

"Most of the data currently available in the biological literature is qualitative. The need for statistics will grow with the availability of quantitative data,

We will then be able to apply the tools of statistical modeling and computational biology to explain how transferred genes and specific mutations serve to reprogram the “integrated circuit of the bacterial cell” so as to manifest pathogens and mutants."

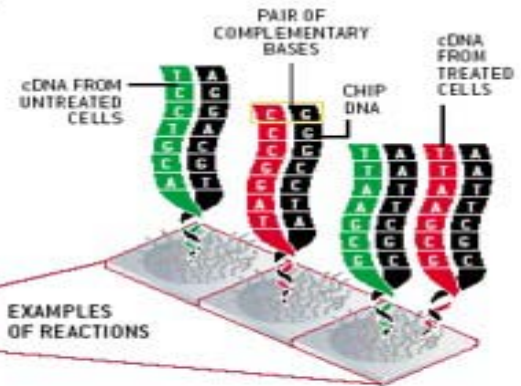
Microarray Example



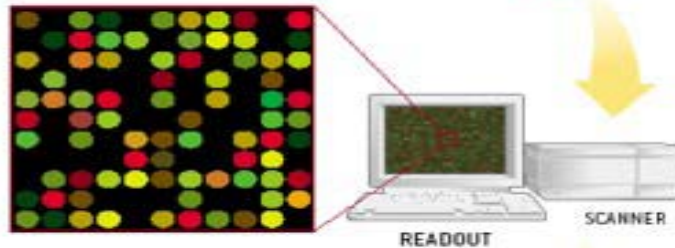
3 Transcribe the mRNA into more stable complementary DNA [cDNA] and add fluorescent labels—green to cDNAs derived from untreated cells, red to those from treated cells.

4 Apply the labeled cDNAs to the chip. Binding occurs when cDNA from a sample finds its complementary sequence of bases on the chip [detail at right]. Such binding means that the gene represented by the chip DNA was active, or expressed, in the sample.

2 Obtain two samples of liver cells; apply the drug to one sample. Then, from each sample, collect molecules of messenger RNA [mRNA]—the mobile copies of genes and the templates for protein synthesis in cells.

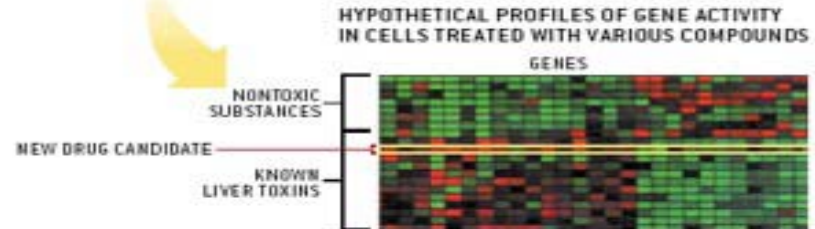


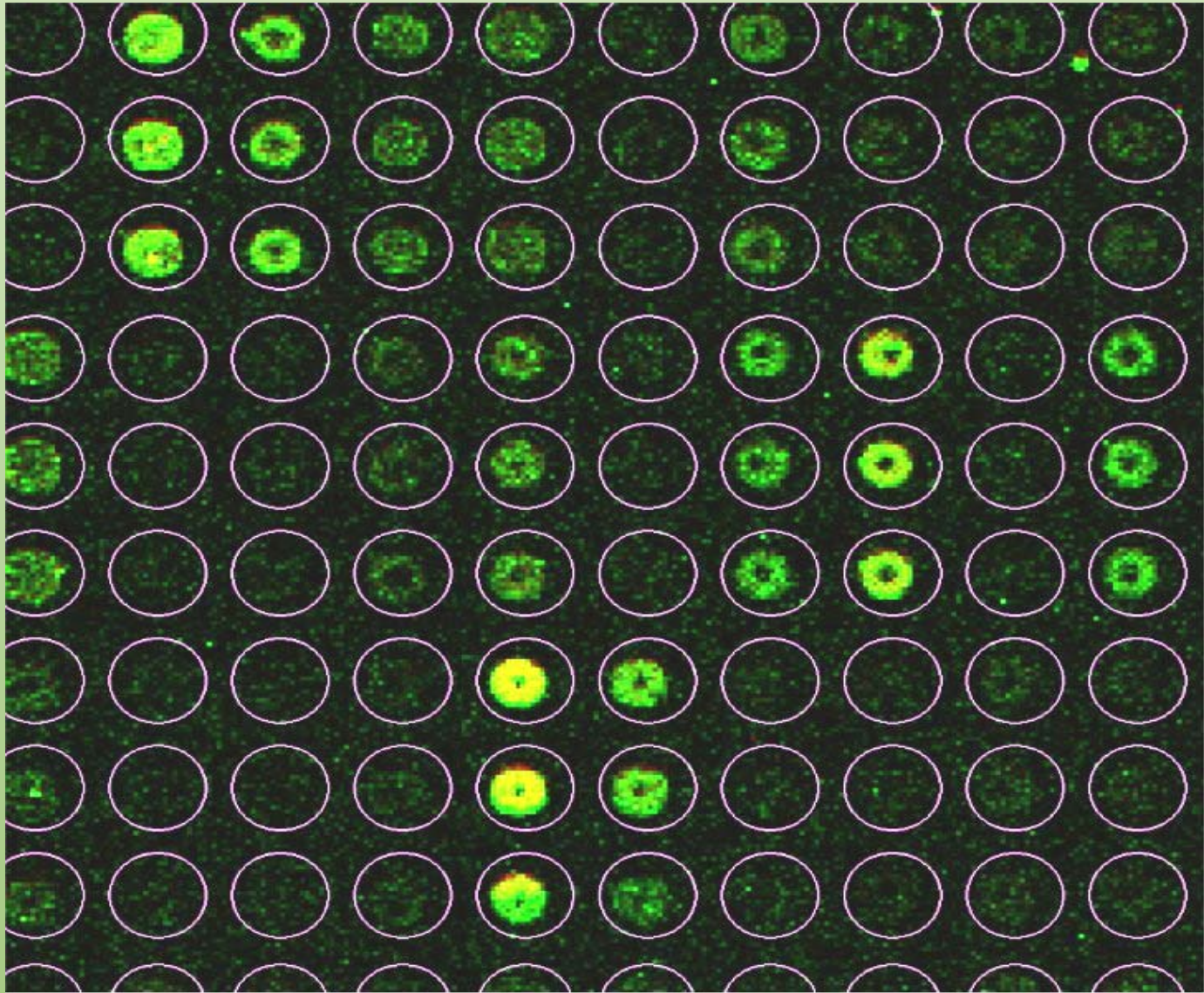
- GENE THAT STRONGLY INCREASED ACTIVITY IN TREATED CELLS
- GENE THAT STRONGLY DECREASED ACTIVITY IN TREATED CELLS
- GENE THAT WAS EQUALLY ACTIVE IN TREATED AND UNTREATED CELLS
- GENE THAT WAS INACTIVE IN BOTH GROUPS



5 Put the chip in a scanner. Have a computer calculate the ratio of red to green at each spot (to quantify any changes in gene activity induced by the drug) and generate a color-coded readout.

6 Determine whether any genes responded strongly to the drug in ways known to promote or reflect liver damage. Or compare the overall expression pattern produced by strong responders with the patterns produced when those genes react to known liver toxins (right). Close similarity would indicate that the new candidate was probably toxic as well. In the diagram, each box represents a single gene's response to a compound.





Array	Signal	Background	Dye
40	277	257	Green
40	248	235	Green
43	192	189	Green
43	198	168	Green
55	2292	4010	Green
55	2130	4268	Green
69	1007	350	Green
69	978	404	Green
70	663	585	Green
70	792	536	Green
72	1161	820	Green
72	1287	816	Green
40	1090	984	Red
40	690	732	Red
43	113	60	Red
43	101	59	Red
55	467	297	Red
55	478	639	Red
69	7901	240	Red
69	7220	330	Red
70	632	169	Red
70	735	143	Red
72	3615	933	Red
72	2972	890	Red

Statistical Experimental Design

Replication – need to observe multiple occurrences of treatments so mean can be accurately estimated.

Variation must be controlled, by blocking or covariates.

Microarrays differ, biological samples differ, etc, so all these sources of variation should be replicated or controlled.

In this experiment:

Each array has both dyes, so we can block on array.

Each array represents one person's gene expression, so multiple arrays allows multiple biological units – replication.

Statistical Models

$$y_{ijk} = \mu + a_j + d_i + a * d_{ij} + e_{ijk}$$

Compare to any scientific model, like $E=mc^2$

Statistical models must address sources of variation, and must test effects of interest.

They contain parameters, and unexplained error.

Statistical Hypothesis Test

Ho: red mean=green mean

	Ho is true in population	Ho is false in population
Observed data say accept Ho	Correct	Type II error (<20%)
Observed data say reject Ho	Type I error (usually set to 5%)	Power (want this to be > 80%)

For programmers...a SAS program for

$$y_{ijk} = \mu + a_j + d_i + a * d_{ij} + e_{ijk}$$

```
proc mixed data=one;  
  class dye array;  
  model signal = dye /outp=rrr;  
  random array array*dye;  
  lsmeans dye/pdiff;  
run;
```

If the model above is used to statistically analyze the example data, we get

Covariance Parameter	Estimates
Cov Parm	Estimate
array	232902
dye*array	4150322
Residual	46190

Type 3 Tests of Fixed Effects

Effect	DF	Num DF	Den DF	F Value	Pr > F
dye	1	5		1.09	0.3439

Least Squares Means

Effect	dye	Estimate	Error Standard	DF	t Value	Pr > t
dye	G	935.42	856.96	5	1.09	0.3248
dye	R	2167.83	856.96	5	2.53	0.0526

Interpretation

The P-value of .3439 is the probability of observing G-R differences of (935-2137) or larger, GIVEN that H_0 is true.

So if green and red are truly equal, we have a 34% chance of seeing a difference at least as large as 935-2137.

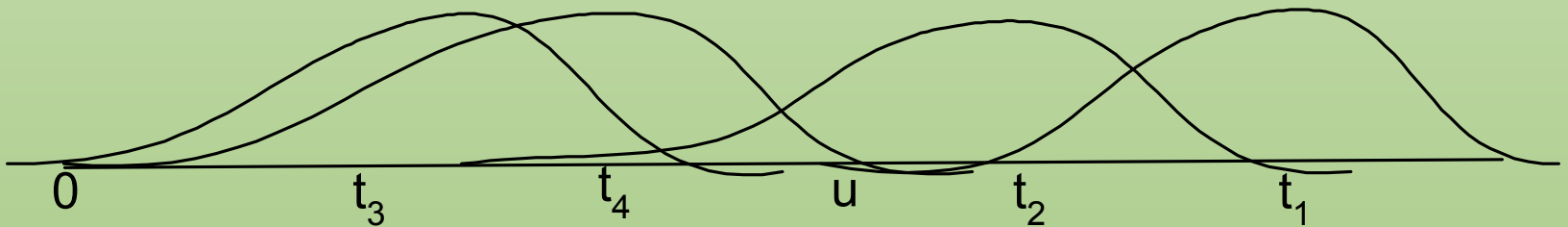
We conclude this chance is fairly likely, so green does equal red – there is no differential gene expression.

In fact only if the P-value gets below .05, with a 5% Type I error rate, do we conclude there is a difference.

Statistical distributions

All of the above statements are based on a probability distribution.

In this example, a normal distribution is used.



More testing concepts

This was just one gene. Remember we are doing this statistical process on 15,000+ genes. With each test, there is a 5% chance of a Type I error – false positive.

After making 100 independent tests, the chance of at least one false positive is .994. Methods such as "false discovery rate" have been proposed to avoid this.

What about the opposite, a false negative, or too little statistical power? Maybe there really is a red-green difference, and we failed to detect it. To make sure power is high, increase sample size and reduce or control variability.

Statistical Concept Summary

Replication

Blocking

Type I error

Power

These all work together to produce
"mostly" correct decisions in the
presence of experimental variation.

Ag Applications Advertisement

Major uses for genomics

Evolution

Human medicine

Agriculture - World population is likely to double in 50 years. Need to double food production from essentially the same land, water and nutrient resources.

Genomic data is in early phases.

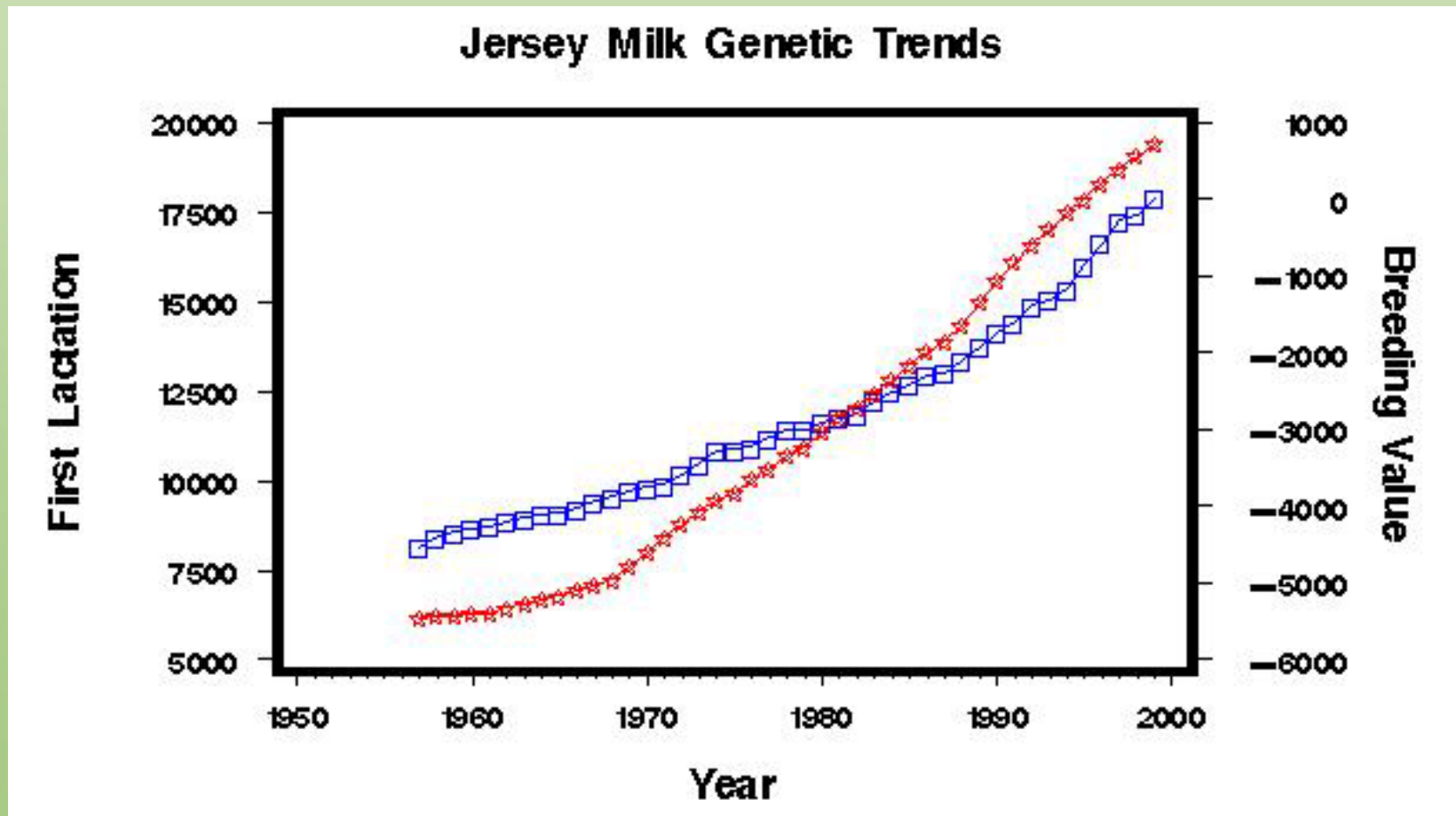
Sequences for rice, poplar trees are available.

Poultry and cattle will be done in a few years.

Lots of opportunities!



Quantitative genetics traditionally has used pedigree and phenotype to predict genetic value.



Genetic evaluation is a statistical process that controls known environment differences (herd, year) and estimates animal differences while accounting for pedigree relationships.

$$y_{ij} = \mu + env_i + animal_j + e_{ij}$$

Implementation is done nationally for cattle, with up to 21 million observations.

This requires linear algebraic solution of sparse matrix equations, taking up to a week of computer time.

Imagine how complex the models will be as genotype information is collected, and ultimately what if we had sequence data on every individual?

$$y_{ijklmnopqrstuv} = \mu + env_i + SNP1_j + SNP2_k + \dots + SNP3000000_j + e_{ij}$$

Lots of opportunities for good science to be done!

Some Web Resources to Explore Statistics for Bioinformatics

Courses

<http://stat-www.berkeley.edu/users/juliab/141C/index.s03.html>

<http://www.biostat.utsa.edu/~kgarnett/bioinformatics/teaching.html>

<http://biomaps.rutgers.edu/course515.html>

Presentations

www.bioss.sari.ac.uk/~jim/bbsrc/bbsrc.ppt

Texts

Ben Hui Liu 2004 Statistical Genomics and Bioinformatics

Ewens and Grant 2001 Statistical Methods in Bioinformatics