Experimental Design For Microarray Experiments

Robert Gentleman, Denise Scholtens
Arden Miller, Sandrine Dudoit

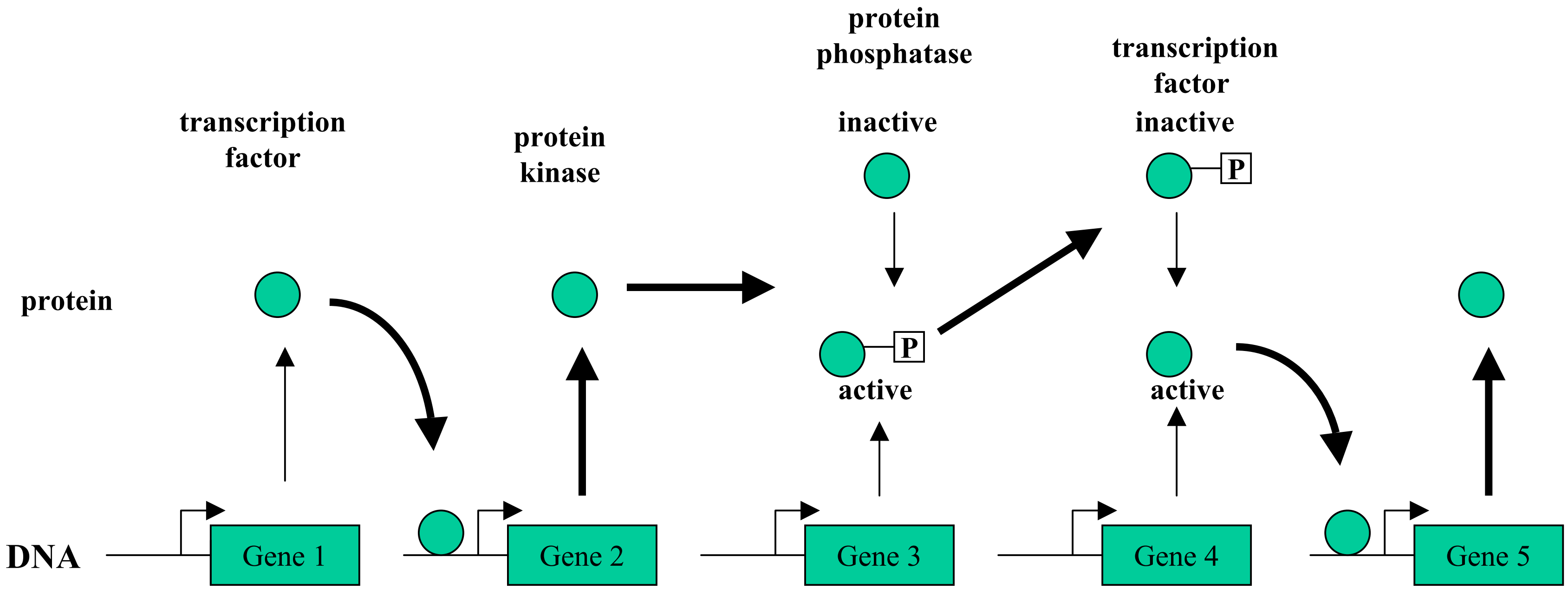
Complexity of Genomic data

- the functioning of cells is a complex and highly structured process
- tools are being developed that allow us to explore this functioning in a multitude of different ways
- these include expression of RNA, expression of proteins and many other processes

Complexity of Genomic Data

- in the next slide we show a stylized biochemical pathway (adapted from Wagner, 2001)
- there are transcription factors, protein kinase and protein phosphatase reactions

An example of the interactions between some genes (adapted from Wagner 2001)



Overview

- Wagner (2001) suggests that the holy grail of functional genomics is the reconstruction of genetic networks
- in this tutorial we examine some methods for doing this in factorial genome wide RNA expression experiments
- such experiments are easy to carry out and are becoming widespread, tools for analyzing them are badly needed

Overview

- while much of the early microarray data have been observational there are many different experiments that can be carried out
- we consider some simple factorial experiments and their analysis
- we assume that there are two factors of interest, F_1 and F_2

Factorial Experiments

- we can obtain expression data on the balanced application of the factors, under the four conditions
  - nothing
  - F_1 alone
  - F_2 alone
  - F_1 and F_2

Factorial Experiments

- if more factors are of interest then fractionally replicated factorial designs should be considered
- biological replication, while not essential is certainly helpful

Factorial Experiments

- the observed data consist of measured levels of mRNA at each of these conditions on patients or model organisms such as cell lines, yeast or mice
- the questions of interest are typically which genes are directly affected by the two factors F_1 and F_2

Factorial Experiments

- we do not just observe changes in the genes that have been directly affected by the factors (primary targets)
- we also observe changes in any other genes whose expression levels are affected by changes in the primary targets (secondary targets)

Gene Effects

- a factor can either inhibit or enhance the production of mRNA for any gene
- the inhibition or enhancement of mRNA production for any given gene can affect mRNA production for other genes either through inhibition or enhancement

Targets

- we define a *target* of a factor to be a gene whose expression of mRNA is altered by the presence of the factor
- a *primary* target is a target that is directly affected by the factor
- a *secondary* target is a target whose expression is altered only via the effects of some other gene (can be traced back to one or more primary targets)

Factorial Experiments

- these experiments can be contrasted with those proposed by Wagner (2001)
- he proposed perturbing each gene in the genome of interest and observing the gene specific effects
- in our experiments we observe genome wide changes and hence less specific information

Factorial Experiments

- here we consider carrying out very few experiments
- the two methods can be complimentary since the results of the genome wide study could be used to design several single gene experiments

Some Examples of Experiments

- methylation: inhibits transcription of specific genes
- if a factor that demethylates the genome were available then one could, in principle determine which genes were methylated (or affected by mythylated genes)
- however we could not determine which genes were primary and which were secondary targets

Some Examples

- many cellular reactions are carried out using energy that is provided by the ADP-ATP phosphorylation mechanism
- if a simple mechanism was available for halting this mechanism then that could be used as a factor in these experiments and information on genes whose transcription is affected by phosphorylation could be identified

Some Examples

- the addition of a second factor (say one such as cyclohexamide, CX, that inhibits translation) will often allow us to isolate the primary factors from the secondary factors
- a simple (but not quite accurate) way to think of the data is as follows
  - N – F_1 (N forms a baseline for just F_1)
  - F_2 – F_1+F_2 (F_2 forms a baseline for F_1+F_2)

Inference

- if the effect of F_1 is the same in the presence and absence of F_2 then it is possibly a primary candidate
- this is especially true in the case of CX (since it has halted most translation)
- we can similarly find potential primary targets of F_2 by reversing the argument

Limitations

- while we may identify genes that are potentially primary targets and those that are potentially secondary targets we cannot identify specific gene gene interactions
- the experiments proposed by Wagner could do that
- the use of relevant meta-data, biological and publication, seems pertinent and could help resolve some of the interactions

Limitations

- a direct corollary of the preceding limitations is the fact that we cannot identify feed back loops
- that is genes, or sets of genes whose regulation is self—controlled
- we can observe the effects but not attribute them

Complications

- complications include the fact that the both F_1 and F_2 will have effects on the cells and their functioning other than those we are interested in
- we could see effects due to either of them because of chemical interactions etc.
- for simplicity we will assume this does not happen

An Experiment

- we now consider a two factor experiment involving CX in detail

The Experiment

- there are two factors, E is known affect transcription of various genes (some known, some unknown)
- CX is known to stop all translation (with very few exceptions)
- the design is a classical factorial design with two factors and we are interested in the main effects and interactions

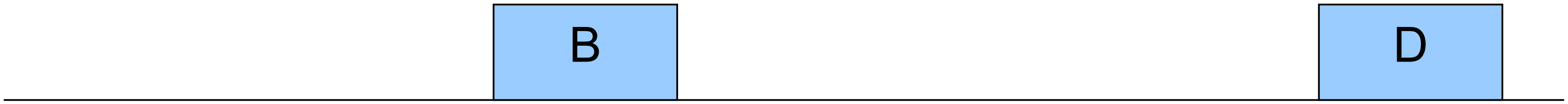
The Experiment

- we identify as targets all genes whose expression of mRNA is affected by the application of E
- a target can be either primary or secondary
  - primary if E directly affects expression of mRNA
  - secondary if mRNA production is affected by some other gene (can be traced back to a primary target)

Scenario 1

- assume that there are two related genes, B and D
- neither is expressed initially, but E causes B to be expressed and this in turn causes D to be expressed
- the addition of CX by itself may not affect expression of either B or D
- conditions with CX and E present will have elevated levels of mRNA_B and low levels of mRNA_D

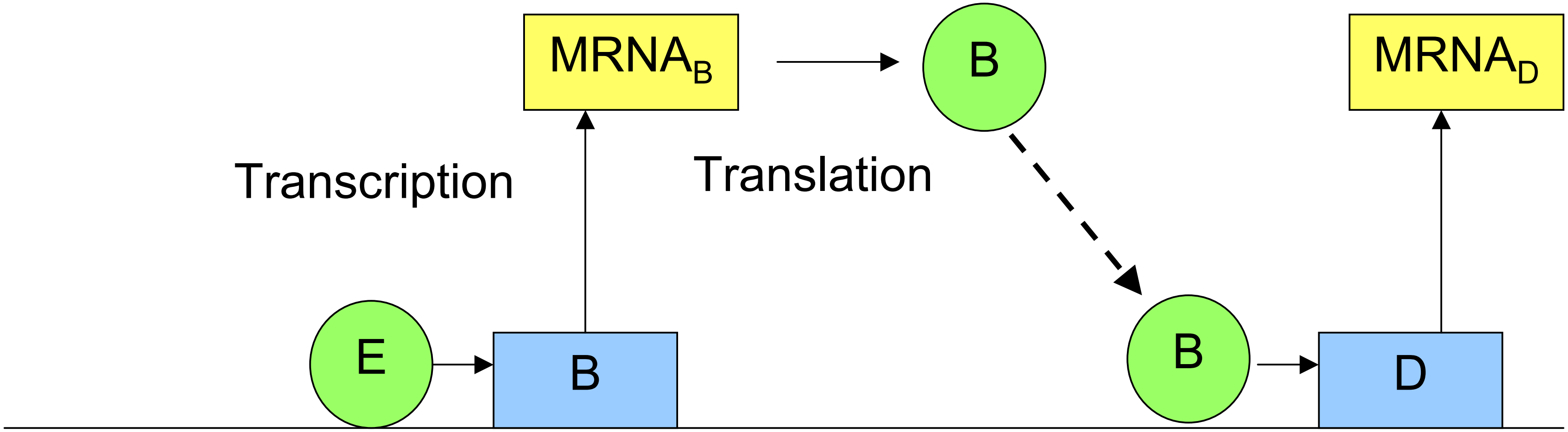
No Factors applied



GeneB is not active

Gene D is not active

E only



MRNA$_B$

B

MRNA$_D$

Transcription

Translation

E

B

B

D

B is a Primary
Target of E

D is a Secondary
Target of E

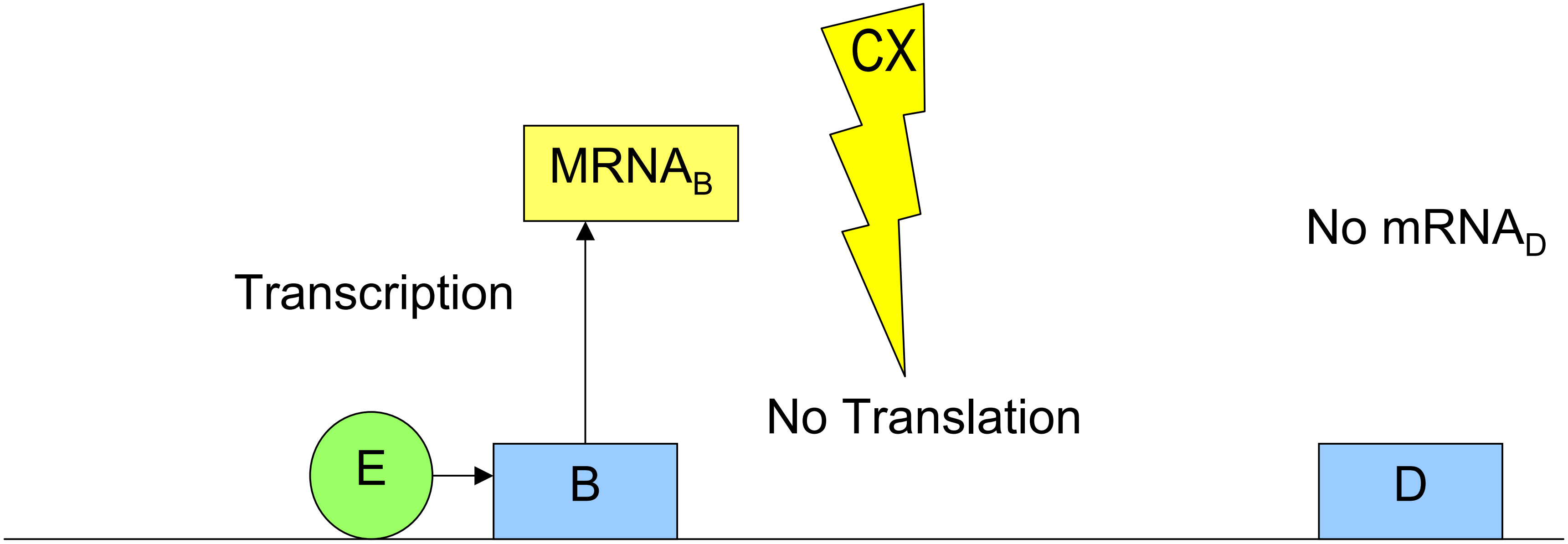
Production of mRNA$_B$
is enhanced by E

Production of mRNA$_D$
is enhanced by B

Interpretation: Scenario 1

- for both genes B and D we expect to see significant regression coefficients for the presence of E
- note that while we show a direct relationship between the expression of B and of D we cannot detect such a relationship from these data (its purpose is purely pedagogical)

E and CX both present



MRNA$_B$

CX

Transcription

No mRNA$_D$

No Translation

E

B

D

B is a Primary Target

Production of mRNA$_B$ is enhanced by E

Production of mRNA$_D$ is decreased (prevented)

Interpretation: Scenario 1

- in the presence of both CX and E we see increased expression of $mRNA_B$ but not of $mRNA_D$
- this will be one of the principles we can use to differentiate between primary targets of E (such as B) and secondary targets of E (such as D)

Interpretation: Scenario 1

	$mRNA_B$	$mRNA_D$
Nothing	Low	Low
E	High	High
CX	Low(?)	Low (?)
E and CX	High	Low

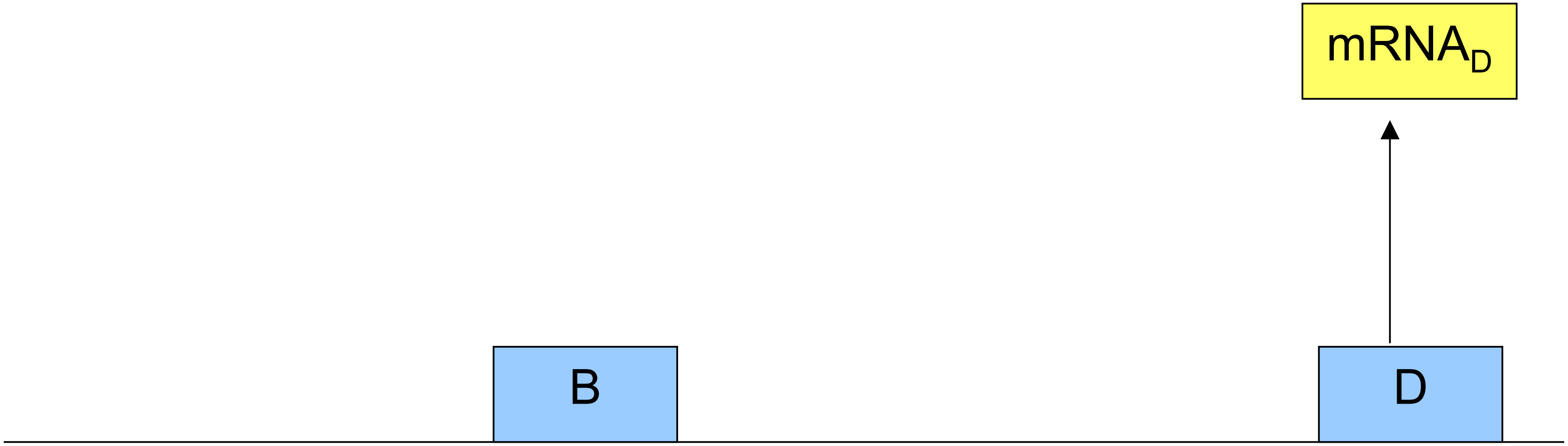
Suppressors: Scenario 2

- we now consider a similar setting where the effect of the gene being enhanced by E is to suppress the other gene D
- initially $mRNA_D$ is being produced and $mRNA_B$ is not
- the addition of E causes the production of $mRNA_B$ and hence the inhibition of $mRNA_D$

Suppressors: Scenario 2

- CX by itself may reduce production of $mRNA_D$
- CX and E together will yield levels of $mRNA_B$ that are high, and levels of $mRNA_D$ that are the same as those observed with CX alone

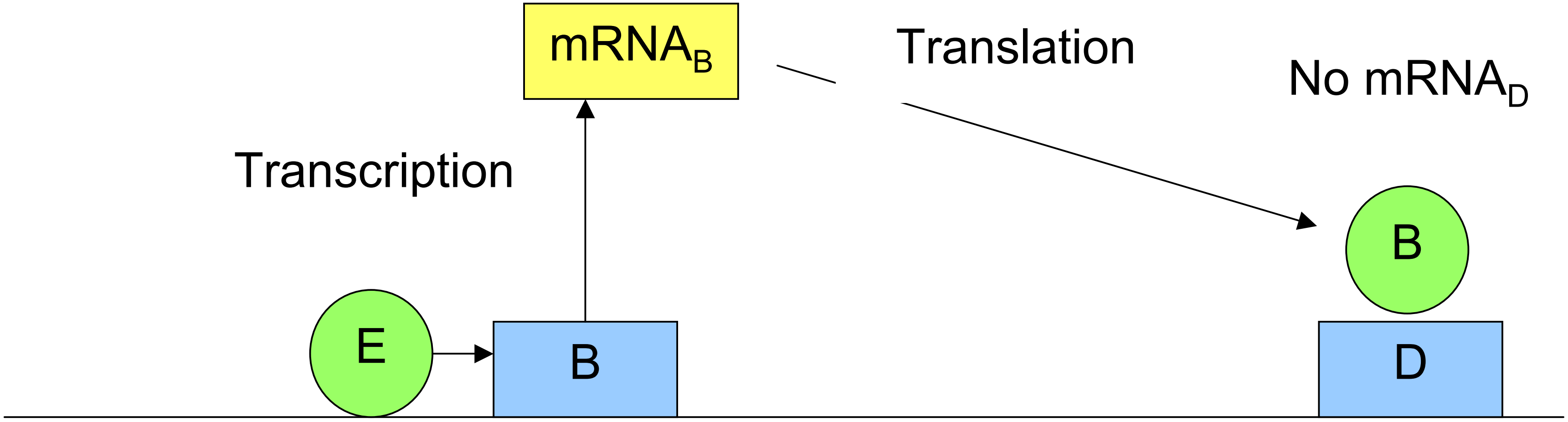
Normal Conditions



B is not active

Production of $mRNA_D$

Introduction of E



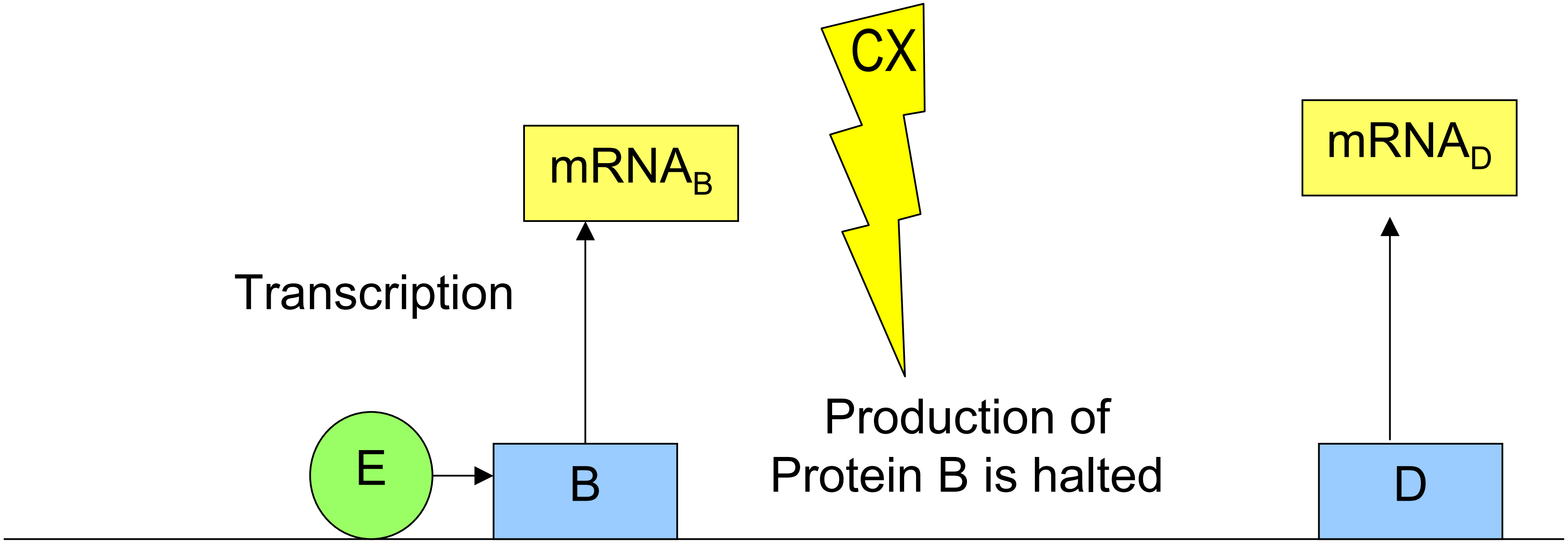
B is a Primary
Target of E

D is a Secondary
Target of E

Production of mRNA$_B$
is enhanced by E

Production of mRNA$_D$
is suppressed by B

Both E and CX present



CX

mRNA$_B$

Transcription

E

B

Production of
Protein B is halted

mRNA$_D$

D

B is a Primary
Target of E

D is a Secondary
Target of E

Production of mRNA$_B$
is enhanced by E

Production of mRNA$_D$
is restored

Interpretation:Scenario 2

	mRNA_B	mRNA_D
Nothing	Low	High
E	High	Low
CX	Low (a)	High(b)
E and CX	High	High(b)

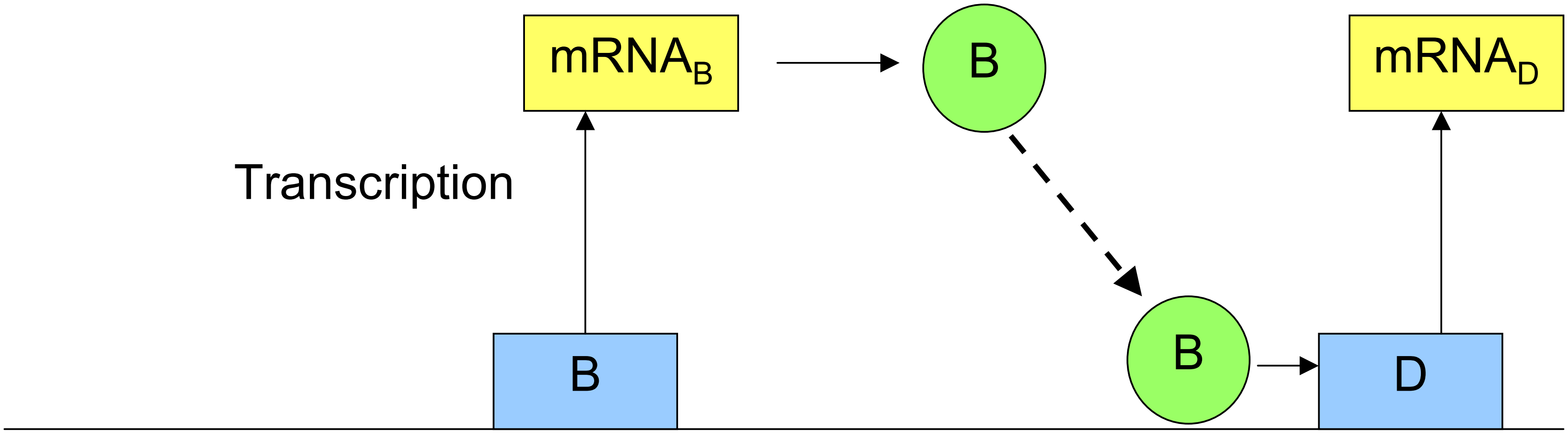
Interpretation: Scenario 2

- the level of mRNA$_D$ when both CX and E are present should be the same as the amount that is present when CX alone is present
- this could be different than the amount when both factors are absent
- mRNA$_D$ could be translationally controlled and so it will be affected by CX

One more example: Scenario 3

- here genes B and D are active, protein B is enhancing production of D
- E inhibits production of mRNA_B, which in turn affects production of D
- CX alone decreases production of mRNA_D, B may be unchanged
- CX and E together will result in decreases in the levels of both mRNA_B and mRNA_D

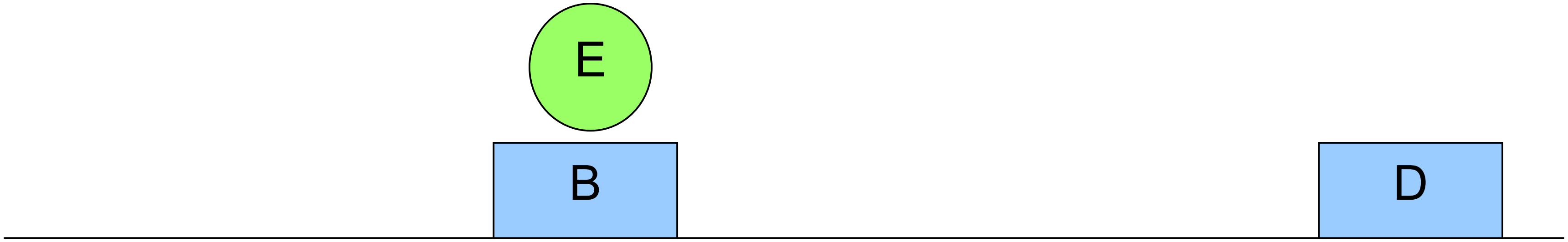
Normal State



B is active

Production of mRNA$_D$ is enhanced by B

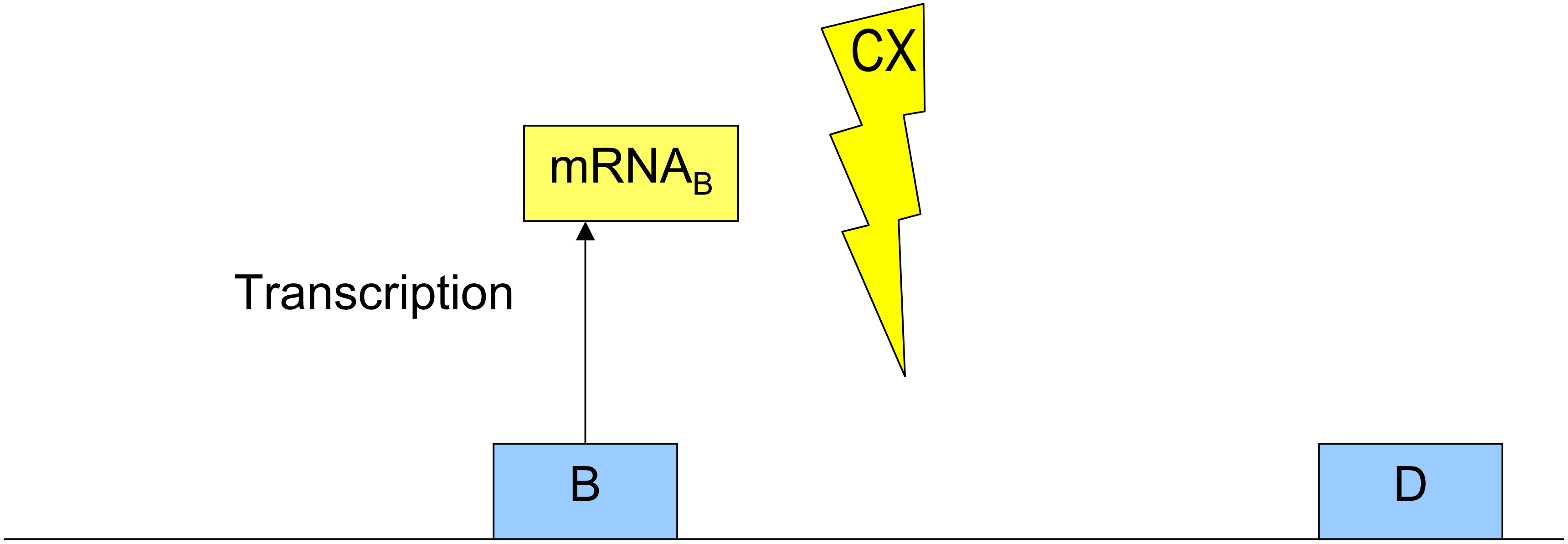
Addition of E



B is suppressed
by E

Production of mRNA$_D$
is also suppressed

Addition of CX



CX

$mRNA_B$

Transcription

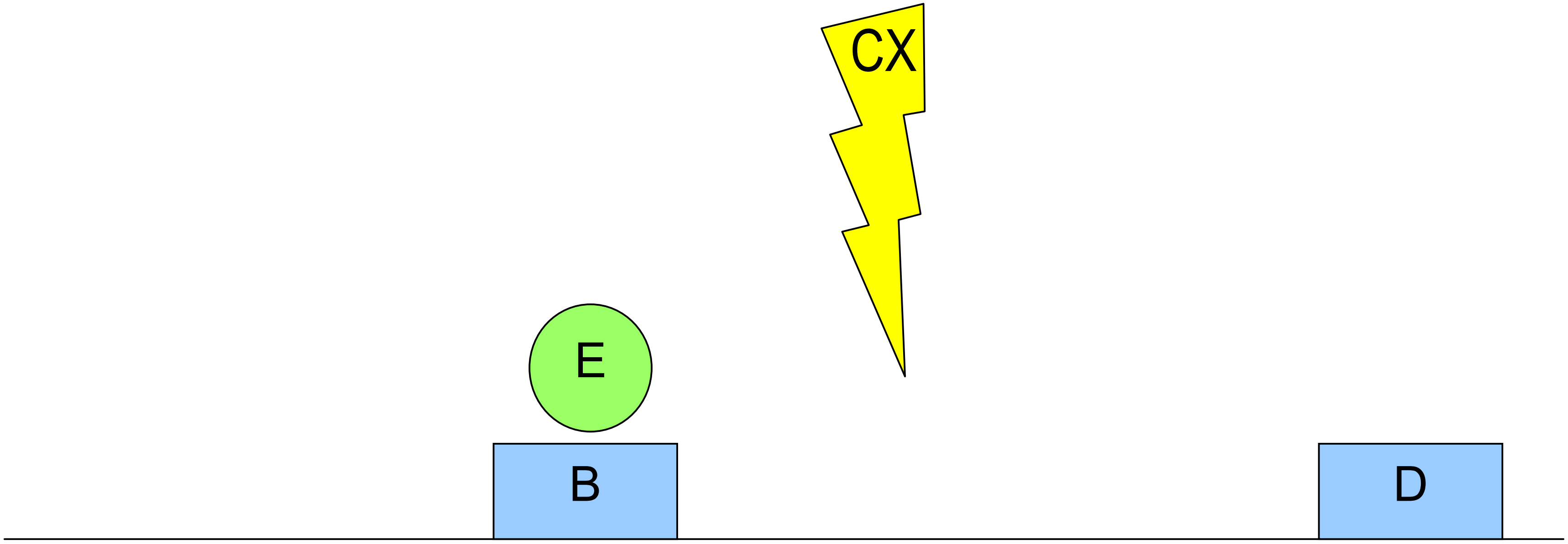
B

D

Production of $mRNA_B$

Production of $mRNA_D$
is halted

Addition of E and CX



Production of mRNA$_B$
is halted

Production of mRNA$_D$
is halted

Interpretation: Scenario 3

	mRNA$_B$	mRNA$_D$
Nothing	High	High
E	Low	Low
CX	High	Low
E and CX	Low	Low

The Experiment

- a microarray experiment can detect changes in the level of mRNA and for both $mRNA_B$ and $mRNA_D$
- but there is a difference, B is a primary target of E, while D is a secondary target of E

Inference

- we are experimenting with a closed, functioning system
- there is no true baseline
- these two facts complicate the analysis and inference in many ways

Inference

- if gene X is any target for E the level of $mRNA_X$ might not change when E is added
- $mRNA_X$ might already be being made as fast as possible, so addition of E has no effect (if we had a true baseline we could eliminate this)
- production of $mRNA_X$ might already be suppressed by some other compound

Inference

- the introduction of CX provides a form of baseline
- since (among other things) CX halts translation we should be able to use the presence or absence of CX to find out about primary versus secondary targets

The Experiment

- if we assume that there is a linear model for the observed expression value (possibly on transformed data) it is:

$$y_{ig} = \mu_g + \beta_{Eg} x_{1i} + \beta_{CXg} x_{2i} + \beta_{E:CX,g} x_{1i} x_{2i} + \varepsilon_{ig}$$

- where *i* indexes chips and *g* indexes genes
- x_1 indicates the presence of E and x_2 indicates the presence of CX

The Experiment

- for any gene we can interpret the coefficients in the linear model as follows
- the parameter β_E can be interpreted as the effect of E
- genes for which β_E is different from zero are potential *targets*
- as noted previously not all targets will have β_E different from zero

The Experiment

- the parameter β_{CX} can be interpreted as the effect due to CX
- if β_{CX} is different from zero indicates that production of mRNA is translationally regulated
- the interpretation of $\beta_{E:CX}$ is more difficult

The Experiment

- we now refer back to the preceding scenarios to determine sets of conditions that will allow us to identify both primary and secondary targets

Scenario 1

	Primary	Secondary
β_1	> 0	> 0
β_2	$= 0$	$= 0$
β_3	$= 0$	$- \beta_1$

Scenario 2

	Primary	Secondary
β_1	> 0	< 0
β_2	$= 0$	$= 0$
β_3	$= 0$	$- \beta_1$

Scenario 3

	Primary	Secondary
β_1	< 0	< 0
β_2	$= 0$	< 0 ($\approx \beta_1$)
β_3	$= 0$	$-\beta_1$

Primary Targets

- consider the case where we have only CX and CX+E
- since CX halts all translation then any differences between the condition where CX alone is present and CX+E is present should indicate primary targets of E

Primary Targets

- this is equivalent to testing the hypothesis:

  H_0: $\mu+\beta_E+\beta_{CX}+\beta_{E:CX} = \mu+\beta_{CX}$

- another equivalent hypothesis is

  H_0: $\mu+\beta_E+\beta_{CX}+\beta_{E:CX} = \mu+\beta_{CX}$

Primary Targets

- genes for which the hypothesis

  H_0: $\mu+\beta_E+\beta_{CX}+\beta_{E:CX} = \mu+\beta_{CX}$

  is rejected are candidates for primary targets
- those with β_E different from zero but for which we do not reject H_0 are secondary targets
- this holds for all Scenarios discussed above

Primary Targets

- we can identify primary targets in at least two different ways
  - fold change, look at ratio of the means of the CX arrays with the CX+E arrays
  - use a linear model and estimate the contrasts (possibly then estimate the ratio)

Secondary Targets

- a secondary target should have the property that β_1 is not zero
- this means that E had some observed effect on expression of the gene
- and that we did not determine that it was a primary target

Other information

- what other information is available from the experiment?
- it seems likely that some inference may be drawn from the relationship between β_E and $\beta_{E:CX}$, their signs and their significance levels

Limitations

- while we can identify primary and secondary targets there is no way to determine the relationship between any two genes
- a corollary of this is that it is not possible to identify feedback loops using these data

Gene Filtering

- some reduction by filtering out genes that are not expressed or that are not affected by the factors will help reduce the computation
- this is problematic since we have only 2 observations at each level of the factors
- our approach was to compute an average for each data pair

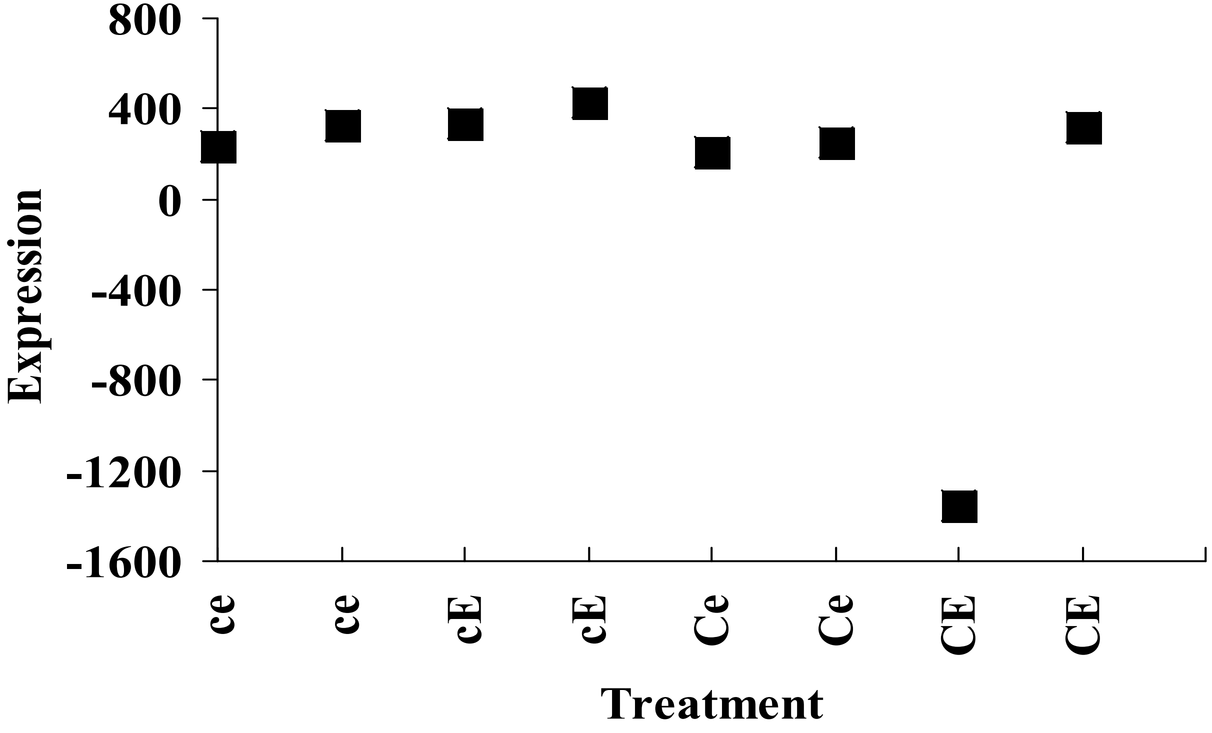
Gene Filtering

- thus for any gene we have four averages (a$_i$)
- if the maximum of the four averages for a given gene was less than 100 the gene was filtered and not analysed further

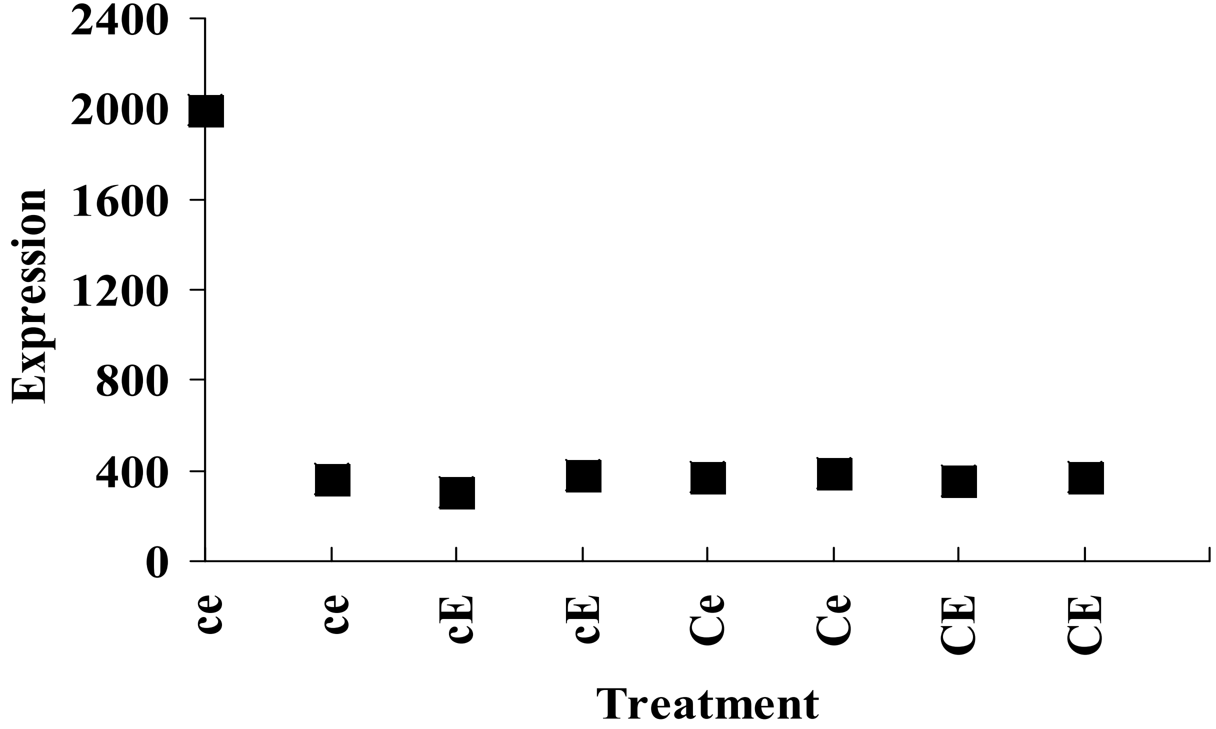
Outlier Detection

- the detection of outliers in factorial experiments is difficult
- the residuals from the fit of the linear model must satisfy a number of constraints and hence are not suitable for outlier detection
- however, outlier detection is important since the presence of outliers will inflate the estimated variance and hence decrease our ability to detect significant effects

Examples of single outliers



Examples of single outliers



Outlier Detection

- when there are replicate chips a simple but effective procedure can be employed
- Miller and Scholtens (xxxx) propose using the following process
- put in some pictures of the outliers/etc
- tables indicating the preliminary results

Outlier Detection

- we presume that the expression at some set of experimental conditions is Normally distributed with mean μ and variance σ^2
- so that the difference $d_\mathrm{i} = x_{\mathrm{i}1} - x_{\mathrm{i}2}$, is

  N(μ, $2\sigma^2$)
- then the ratio, $\dfrac{d_i^2}{\sum_{j \neq i} d_j^2 / 3}$, is $F_{1,3}$ and we

Outlier Detection

- we use a p-value of $4*P(F_{1,3}>f)$ to adjust for the fact that we have used the maximum of the d_i's in our calculation

Relevance

- in most cases there is some literature on genes that are likely to be affected by the different factors
- it is prudent to obtain this information and examine its consistency with the experimental data

Relevance

- there is a great deal of metadata available
- this includes references in published literature
- relationships through protein—protein interactions
- known promoter inhibitor relationships
- these data can all be used to further explore and understand the experimental data

References

- *How to reconstruct a large genetic network in fewer than n^2 easy steps*, Wagner, A., Bioinformatics, 2001, 1183—1197.