Limitations of massively parallel technologies

Zoltan Szallasi, MD

Children’s Hospital Informatics Program
www.chip.org
New technology

All problems will be solved within a couple of years

Realistic Expectations (limitations)
Limitations: (you want to make predictions)

Accuracy – noise

Sensitivity - completeness

Inherent limitations –
(think about unpredictability > chaos)
NOISE:
- what is noise? (and what is signal?)
- noise as an inherent feature of complex systems
- noise in continuous and discrete measurements
- noise as the limitation of the technology
- what can be done about noise?
  Statistics
  Normalization as a way to deal with systematic errors
c: an unwanted signal or a disturbance (as static or a variation of voltage) in an electronic device or instrument (as radio or television); 
broadly: a disturbance interfering with the operation of a usually mechanical device or system

d: electromagnetic radiation (as light or radio waves) that is composed of several frequencies and that involves random changes in frequency or amplitude

e: irrelevant or meaningless data or output occurring along with desired information
Noise may turn out to be an important signal !!!!

- Penzias and Wilson >>> cosmic background radiation

- discovery of the chemotherapeutic agent cis-platinum
What we perceive as noise/error might be a key component of biological processes:

1) Mutations in evolution
2) “Junk” DNA
3) Asymmetric cell division may contribute to differentiation
4) Stochastic fluctuations may be important for the stability of complex physicochemical systems
Genetic networks are stochastic systems:
1) A couple of hundred copies of a given transcription factor/nucleus
2) Intracellular environment is the not a free solution
3) Reaction kinetics is often slow etc.

Comment in:

Stochastic gene expression in a single cell.
Elowitz MB, Levine AJ, Siggia ED, Swain PS.
-measuring population averaged data.

That is true even if single cells are quantified due to stochasticity > two cells can get from a given state to another one via different paths
Noise in measurements

There is no measurement without noise - (it is the accuracy/sensitivity of your measurement that is low)

For continuous variables it is expected to obtain data with a certain “spread”
Consequently: Statistics was invented

- 0.5, -0.3, 0.2, 1.4, -1.5…..etc what is the true value of the observed variable ?
- Did the variable change due to a given treatment? Etc.

Lots of measurements and/or fairly good idea about the nature of the noise (e.g. normal distribution)
Statistical analysis in biology:

1) What is the true value of a given parameter?

2) the most common analysis – Bayesian

3) You don’t believe the measurements >> normalization

4) There are too many numbers >> permutation etc.
Biological measurements are often expensive !!!!!!!!!

A large number of papers relating to cancer were published in Nature/Science ….. based on single microarray measurements

STATISTICS

Reliable numbers cannot be produced without replicates
The central problem:

In massively parallel biological measurements quantitative or qualitative calls are supposed to be made on a large number of heterogeneous variables using only a few replicates.
Noise of continuous variables, e.g. microarray measurements

Tissue or Tissue under influence

---

RNA

......

cDNA/cRNA
Tagged with fluorescent dye

---

Microarray of genes aka gene chips

---

Ideally: 1 copy of a given RNA will produce 1 unit of a specific signal  !!!!!!!!!!!!!!
1) cDNA produced from RNA (initiation of RT step, RT might drop off etc.)
2) cRNA produced in the presence of fluorescent dyes (cRNA production is not linear, Dye incorporation)
3) Breaking down cRNA into small pieces
4) hybridization/cross hybridization

final signal = $\sum$ (all of the above)
The situation is further complicated by other experimental issues >>> two-color cDNA microarray
Ratio is influenced on background calculations

equal amounts of labelled cDNA samples

There is no truly blank spot !!!! Background
**Perfect match probe cells**

**Mismatch probe cells**

**Fluorescence Intensity Image**

Reference sequence Spaced DNA probe pairs

.. TGTGATGGTGGAATGGTGTCAGAAGGACTCCTATGTGGGTGACGGAGGCC ...

**mRNA reference sequence**

```
A A T G G G T C A G A A G G A C T C C T A T G T G G G T G
A A T G G G T C A G A A C G A C T C C T A T G T G G G T G
```

**Perfect match Oligo**

**Mismatch Oligo**
Data representation

If we express gene expression measurements as “per unit RNA” then decrease in the level of a given message unavoidably leads to a relative increase in the level of other messages.
Distribution of probe intensities of several Affymetrix data sets belonging to the same set of experiment.

Systematic error

\[ \text{Density (x = x[1], from = 4, to = 16)} \]

\[ \text{N = 131822 \quad Bandwidth = 0.1128} \]

Normalization
Normalization – You don’t believe the numbers

1) “most or certain things do not change”

2) Error model
Shifting the means or medians and adjusting the distributions by Cubic spline fit/ Lowess etc. (Overfitting !!!)
cDNA microarray: the R/G ratios are intensity dependent

Values should scatter about zero.

Courtesy of Natalie Thorne. Used with permission.
Overview of normalization:
- to correct for systematic errors

1) Choose a set of elements that will be used
   - housekeeping genes
   - special control genes etc.

2) Determine the normalization function
   - global mean/median normalization
   - intensity dependent normalization
Microarray Gene Expression Data Society
www.mged.org
Intensity dependent normalization by error models

Error model: Rocke, Vingron

Low concentrations → $x = \mu + \varepsilon$

High concentrations → $x = \mu e^{\eta}$

$x = \mu e^{\eta} + \varepsilon$

$\eta \sim N(0, \sigma_{\eta}^2)$

$\varepsilon \sim N(0, \sigma_{\varepsilon}^2)$
Noise will limit the useful information content of measurements:

A reliable detection of 2-fold differences seems to be the practical limit of massively parallel quantitation.

(estimate: optimistic and not cross-platform)
A rational experiment will sample gene-expression according to a time-series in which each consecutive time point is expected to produce at least as large expression level difference as the error of measurement: approximately 5 min intervals in yeast, 15-30 min intervals in mammalian cells.
Limitations: (you want to make predictions)

Accuracy – noise

Sensitivity - completeness

Inherent limitations –
   (think about unpredictability > chaos)
Sensitivity – completeness

How many parameters are we measuring?

How many parameters should we measure?
How many bionodes?

Cautious estimate: on the order of $1-2 \times 10^5$

10,000-20,000 active genes per cell

< 3 posttranslational modifications/protein in yeast

3-6 (?) posttranslational modifications/protein in humans

The number of bionodes is probably less than 10 times the number of genes

Splice variants $<$ modules
The coverage of microarray chips and proteomics keeps increasing >>>> complete genome
Holland MJ. Transcript abundance in yeast varies over six orders of magnitude.
J Biol Chem. 2002

Sensitivity: 2 copies/cell
MOST transcripts are not seen by microarray


Transcript abundance in yeast varies over six orders of magnitude.
Holland MJ.
The utmost goal of technology:

Single copy/ single cell

BUT even if you measure everything accurately there might be problems with predictions
Even a relatively simple set of ODEs can produce a rather strange behavior.

Edward Lorenz – 3 linked ODEs produced a behavior very sensitive to the initial conditions. (Chaos theory, Bifurcations etc.)

Small changes in the initial conditions can cause huge changes at later time points
The problem of way too many correlated numbers:

Can this be
due to chance?
- Analytical solution

- Computational solution:
  Permute and look for similar patterns
In some cases analytical solution may exist
Six breast cancer cell lines yielded 13 consistently mis-regulated
genes (H-cadherin, S1002A, keratin 5 etc.)

Can this be due to chance?

“E” different cell lines
“N”-gene microarray
\( M_i \) genes mis-regulated in the “\( i \)”-th cell line,
K consistently mis-regulated across all E cell lines.

What is the probability that the K genes were mis-regulated
by chance?
This translates into a simple combinatorics problem

**BUT !!!** - what if more genes are involved
Distribution of pair-wise correlation coefficients in cancer associated gene expression data
The problem of way too many correlated numbers is a particularly nasty one.

Significance can be off by orders of magnitude when comparing completely random permutations with “structural permutations”
Noise in discrete measurements: DNA sequences

Measurement error: Sequencing errors (0.1%-1%)

Solution: sequence a lot
AAATAA ACTCGGTGACCAAAAAAGAGTGATGAGGATAGATGTCAGAATGGTTGCTAAGGCACCTATTATTAGGTCGCTTTATTAGTTTTT CATGCCGTACATTGCACCTGGCAGACCTTGGCCTTATTTTCTCTGTACATTTTTATTTTTCCCGCTGTGCTGCGCGGTTAGTTACACTGCGTTG TGTATTTCGCTGTGCACCGGAGGTCTGCCTGTTAAGCGATGTGGTAGG GCCACGTGTTTGCTTCTAGAGTGGCCTCTCGCTCTTTATTACCTC GCGGTTTTGCTTCTAGAGTGCCTCTCGGCTCTTTTTTATTACCTCG CGCCTTGTCAAATAGGTTTTTACCTCGCAAGGGATATAAGAAAC GTTCGCCGCGCGCTTCCGCCTGACAATAAAAATCTTGATGGGACAGCG GTTATACCAGG

3 billion
-Find genes, introns, exons, transcription factor binding sites etc.
Help can be found --- cDNA libraries etc.

**BUT**


~1600 ACTUALLY transcribed antisense transcriptional Units


As much as one order of magnitude more of the genomic sequence is transcribed than accounted for by the predicted and characterized exons.
TF binding site: TGGGACT

It can also be: TGCACCT

TGG/CACT

TCG/CNCT

Try to add constraints –
1) Within –500 bp from the ATG
2) Tends to cluster in the same region
Even if you do all this you will find that many “obviously” TF binding site-looking sequences do not function as such. (due to higher level DNA organization etc.)

AND

You often do not know what sequence to start with.
1. Statistical overrepresentation

2. Cross-species conservation

3. Using artificial intelligence/Machine learning
   Hidden Markov models for exon/intron/gene identification (GENIE)

You define the rules
Sequencing and comparison of yeast species to identify genes and regulatory elements.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES.

*S. cerevisiae*  *S. bayanus*  *S. mikatae*  *S. paradoxus*

**Number of genes ~ 5,500**
Figure 1  Aligned ORFs across four species. A 50-kb segment of *S. cerevisiae* chromosome VI aligned with orthologous contigs from each of the other three species. Predicted ORFs are shown as arrows pointing in the direction of transcription. Orthologous ORFs are connected by dotted lines and are coloured by the type of correspondence: red for 1-to-1 matches, blue for 1-to-2 matches and white for unmatched ORFs. Sequence gaps are indicated by vertical lines at the ends of contigs, with the estimated size of each gap shown by the length of the hook. See Supplementary Information for 250 such figures tiling the complete *S. cerevisiae* genome.
Slow and rapid evolution:

YBR184W – 32% nucleotide and 13% aa identity

MATa2 - 100 % nucleotide and 100 % aa identity  !!!!!!!!
Figure 8: Conservation in the GAL1-GAL10 intergenic region. Multiple alignment of the four species shows a strong overlap between functional elements and stretches of conservation. Arrows denote conserved and divergent positions in the multiple alignment. Blue arrows indicate the start and transcriptional orientation of the mating ORAs. Conserved DNase I footprints are boxed and labelled according to the bound factor.

Courtesy of Eric Lander. Used with permission.
$XYZn_{(0-21)}ABC$

Intergenic conservation
Intergenic vs. genic conservation
Upstream vs. downstream conservation

A given motif is also enriched in front of genes with similar function
<table>
<thead>
<tr>
<th>Discovered motif</th>
<th>Location</th>
<th>MCS</th>
<th>Best category</th>
<th>CCS</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>yCGTnnnmRYGAY</td>
<td>5'</td>
<td>36.2</td>
<td>ChiP: Abf1</td>
<td>90</td>
<td>Known: Abf1</td>
</tr>
<tr>
<td>RTTACCGRRM</td>
<td>5'</td>
<td>34.3</td>
<td>ChiP: Reb1</td>
<td>38</td>
<td>Known: Reb1</td>
</tr>
<tr>
<td>gcGATGAGmgtcaw</td>
<td>5'</td>
<td>24.7</td>
<td>Exp. cluster 74</td>
<td>62</td>
<td>Known: Esr1 GATGAG</td>
</tr>
<tr>
<td>TSGGCGGCTAVW</td>
<td>5'</td>
<td>23.4</td>
<td>GO meiosis</td>
<td>10</td>
<td>Known: Lme6/Not180</td>
</tr>
<tr>
<td>RTCAGGTGV</td>
<td>5'</td>
<td>17.6</td>
<td>ChiP: Cbf1</td>
<td>27</td>
<td>Known: Cbf1/Phe1</td>
</tr>
<tr>
<td>WTATYTYAACADG</td>
<td>3'</td>
<td>17.4</td>
<td>Exp. cluster 16, downstream</td>
<td>25</td>
<td>New mitochondrial down</td>
</tr>
<tr>
<td>GRRAAAAATTITTCACT</td>
<td>5'</td>
<td>15.6</td>
<td>Exp. cluster 74</td>
<td>37</td>
<td>Known: Esr2</td>
</tr>
<tr>
<td>TCGCAAnAttGGAAGA</td>
<td>5'</td>
<td>13.8</td>
<td>ChiP: Mcm1</td>
<td>29</td>
<td>Known: Mcm1</td>
</tr>
<tr>
<td>CTTCTCCTTTTCCY</td>
<td>5'</td>
<td>13.5</td>
<td>GO filamentation</td>
<td>7</td>
<td>New filamentation</td>
</tr>
<tr>
<td>TYYTCCAGA</td>
<td>5'</td>
<td>12.5</td>
<td>Exp. cluster 86</td>
<td>5</td>
<td>Known: Xbp1 (Hsf1-co-oc)</td>
</tr>
<tr>
<td>TTTTCGGCG</td>
<td>5'</td>
<td>12.0</td>
<td>ChiP: Sw4</td>
<td>21</td>
<td>Known: Sw4 fixed gap</td>
</tr>
<tr>
<td>TTTTGCAG</td>
<td>5'</td>
<td>12.0</td>
<td>ChiP: Mbp1</td>
<td>18</td>
<td>New Sw4 variable gap</td>
</tr>
<tr>
<td>TTTTTCGGT</td>
<td>5'</td>
<td>12.0</td>
<td>ChiP: Sw4</td>
<td>11</td>
<td>Known: Sw4 variable gap</td>
</tr>
<tr>
<td>GGGATGCCG</td>
<td>5'</td>
<td>11.8</td>
<td>GO filamentation</td>
<td>11</td>
<td>New Mbp1/Sw6</td>
</tr>
<tr>
<td>GGGATGCCG</td>
<td>5'</td>
<td>11.8</td>
<td>ChiP: Fkh2</td>
<td>26</td>
<td>New filamentation</td>
</tr>
<tr>
<td>TTTTTGCACC</td>
<td>5'</td>
<td>11.0</td>
<td>GO proteolysis</td>
<td>11</td>
<td>New filamentation</td>
</tr>
<tr>
<td>TTGTTCATnTTT</td>
<td>5'</td>
<td>10.8</td>
<td>ChiP: Fkh2</td>
<td>28</td>
<td>Known: Fkh1/2</td>
</tr>
<tr>
<td>RvAACCCTD</td>
<td>5'</td>
<td>10.3</td>
<td>-</td>
<td>-</td>
<td>Known: AFl</td>
</tr>
<tr>
<td>WCCGGCTGGCGT</td>
<td>5'</td>
<td>10.2</td>
<td>ChiP: Mbp1</td>
<td>17</td>
<td>New, double Mbp1</td>
</tr>
<tr>
<td>GGGGTnACCC</td>
<td>5'</td>
<td>10.0</td>
<td>ChiP: Reb1</td>
<td>9</td>
<td>New Reb1 palindrome</td>
</tr>
<tr>
<td>GnnATGTTGQGCTGT</td>
<td>5'</td>
<td>9.9</td>
<td>ChiP: Fkh1</td>
<td>6</td>
<td>Known: Rap1</td>
</tr>
<tr>
<td>TTTTGGTGCA</td>
<td>5'</td>
<td>9.9</td>
<td>ChiP: Sum1</td>
<td>14</td>
<td>Known: Maxe</td>
</tr>
<tr>
<td>TTCAnCAGGG</td>
<td>5'</td>
<td>9.8</td>
<td>-</td>
<td>-</td>
<td>New no category</td>
</tr>
<tr>
<td>ATTAATATATTATATATnatta</td>
<td>3'</td>
<td>9.5</td>
<td>GO filamentation</td>
<td>6</td>
<td>New no category</td>
</tr>
<tr>
<td>SCGnHAGS</td>
<td>5'</td>
<td>8.8</td>
<td>Exp. cluster 37</td>
<td>6</td>
<td>New no category</td>
</tr>
<tr>
<td>ACAGCCGCGG</td>
<td>5'</td>
<td>8.6</td>
<td>Exp. cluster 46</td>
<td>6</td>
<td>New expression cluster 3</td>
</tr>
<tr>
<td>DGCGCAGG</td>
<td>5'</td>
<td>8.1</td>
<td>ChiP: Met31</td>
<td>8</td>
<td>Known: Mig1b</td>
</tr>
<tr>
<td>SKGTTGAGc</td>
<td>5'</td>
<td>8.1</td>
<td>ChiP: Met31</td>
<td>5</td>
<td>Known: Met31</td>
</tr>
<tr>
<td>TTTTn19GCKCG</td>
<td>5'</td>
<td>8.1</td>
<td>ChiP: Met31</td>
<td>5</td>
<td>Known: Met31</td>
</tr>
<tr>
<td>HRCCCTYWnD</td>
<td>5'</td>
<td>8.1</td>
<td>ChiP: Met31</td>
<td>5</td>
<td>Known: Met31</td>
</tr>
<tr>
<td>TCKCtCnnnGGING</td>
<td>5'</td>
<td>7.8</td>
<td>Exp. cluster 8</td>
<td>22</td>
<td>Known: Msm2/4</td>
</tr>
<tr>
<td>GTGTCAAGAA</td>
<td>5'</td>
<td>7.3</td>
<td>ChiP: Msm1</td>
<td>15</td>
<td>Known: Msm1 (htar tRNA)</td>
</tr>
<tr>
<td>RGTGTTTTCGG</td>
<td>5'</td>
<td>7.1</td>
<td>ChiP: Msm1</td>
<td>15</td>
<td>Known: Msm1</td>
</tr>
<tr>
<td>TTC1MGGAGA</td>
<td>5'</td>
<td>7.1</td>
<td>ChiP: Msm1</td>
<td>7</td>
<td>New Msm1</td>
</tr>
<tr>
<td>YCCGSgG3S</td>
<td>5'</td>
<td>7.0</td>
<td>ChiP: Msm1</td>
<td>10</td>
<td>Known: Hsf1</td>
</tr>
<tr>
<td>CnCCCTTTTAC</td>
<td>5'</td>
<td>6.7</td>
<td>GO filamentation</td>
<td>9</td>
<td>New filamentation</td>
</tr>
<tr>
<td>OCSCGCAAGG</td>
<td>5'</td>
<td>6.6</td>
<td>-</td>
<td>-</td>
<td>New no category</td>
</tr>
<tr>
<td>SKTTCCTT</td>
<td>5'</td>
<td>6.4</td>
<td>ChiP: Leu3</td>
<td>8</td>
<td>Known: Leu3</td>
</tr>
<tr>
<td>CTCGCCCTAT</td>
<td>5'</td>
<td>6.4</td>
<td>GO filamentation</td>
<td>7</td>
<td>New filamentation</td>
</tr>
<tr>
<td>GCCCGG</td>
<td>5'</td>
<td>6.3</td>
<td>Exp. cluster 8</td>
<td>11</td>
<td>New filamentation</td>
</tr>
<tr>
<td>SGGGGCGFB</td>
<td>5'</td>
<td>6.3</td>
<td>GO filamentation</td>
<td>10</td>
<td>New filamentation</td>
</tr>
</tbody>
</table>

Courtesy of Eric Lander. Used with permission.