Statistical and mathematical challenges of the emerging technologies

Terry Speed
Department of Statistics, UC Berkeley
Division of Genetics & Bioinformatics, WEHI

Mathematical Challenges in Systems Biology
University of Warwick, October 28, 2004
What systems biology means to me?

My principal interest in systems biology: *cis*-regulatory transcription networks.

What I see..

*L Hood / ISB*. Pathways and networks play a key role. Specific biological systems (e.g. galactose utilization in yeast) are systematically perturbed. Many responses are monitored, e.g. gene expression microarray and proteomic data for 1000s genes/proteins. These are typically integrated with stored data, e.g. protein-protein interaction or pathway data. New hypotheses are then formulated, new experiments conceived, and so on.
What systems biology means to me, cont.

**EH Davidson/H Bolouri.** The *cis*-regulatory system of *endo16* in the sea urchin. A computational model representing the “genetically mandated logic functions that the system dictates, …how the time-varying kinetic inputs are processed ....into kinetic outputs”. Not a d.e. model, nor a Boolean model.

**H McAdam/L Shapiro/A Arkin.** Modelling the *E. coli trp* and *lac* operon, the phage *T7* life cycle, the lytic/lysogenic pathways in phage λ. Markov process (s.p.d.e.) models.

**JJ Tyson and colleagues.** Modelling the yeast cell cycle: ~30 o.d.e’s, ~30 other equations, several rules.
What systems biology means to me, fin.

H Kitano: integrated process of computational modelling, systems analysis, technology development for experiments, and qualitative experiments (cf L Hood above)

S Miyano: system component identification, prediction, control, design. Many examples in the bioinformatics literature.

A mix of users and non-users of high-throughput technologies. From now on h- tt’s only!
Some half-baked philosophy

Much of our break-out discussion last night centered around the interplay of data and modelling. There is a third player: methods. Some comments.

Many people like to study models…which is fine
Others like to study math/stat/comp methods…also fine

Most of the scientists I interact with have questions and data they believe relevant to the questions. So my world is:

Questions, data and statistics (to help answer the question using the data, and assess the uncertainty in the answer).

To me, at this point in my life, models and methods are means to this end, not such interesting ends in themselves.
Two quotes from the other Doyle

“I have no data. It is a capital mistake to theorise before one has data.” (Scandal in Bohemia)

"Data! data! data!" he cried impatiently. "I can't make bricks without clay.” (The Adventure of The Copper Beeches)
What are the emerging technologies?

**Hybridization-based assays:** microarrays/chips, (Affymetrix, Agilent, Amersham, Appliedbiosystems), bead arrays (Illumina), a thousand sources of home-made arrays

**Protein assays:** many variants, see a few slides on (Clontech)

**PCR based assays:** qrt-pcr (AB)

**DNA tag-based:** Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS)

**Composites:** ChIP-chip
What are the emerging technologies, cont?

Mass spectrometry: for detecting presence (MS, MS/MS) or relative concentration of proteins (ICAT); for measuring relative mRNA abundance (MassARRAY QGE).

FACS: Fluorescence-activated cell sorting

Plus many many things I’ve just read about: assays to detect protein phosphorylation state, to identify protein interactions (Y2H), Imaging of living cells: confocal, FRET,..(Cellomics KineticScan); to measure metabolite concentrations, ...
Millions of copies of a specific oligonucleotide probe synthesised in situ. 

≈500,000 different complementary probes
Protein analysis on a proteomic scale
Phizicky et al, Nature Insight Review 2003

Source of the next 3 figures.
A glimpse of data to come.
(a) Serum probes
- Cell lysates
- Living cells

(b) Protein probes
- Nucleic acid probes
- Drug probes
- Enzymes

Protein expression level
Protein profiling
Diagnostics

Protein binding properties
Pathway building
Drug discovery
Post-translational modification
High-throughput yeast 2-hybrids
Detecting phosphorylated tyrosines using GFP and FRET
Reverse-phase protein lysate arrays (Nishizuka et al, PNAS 2003: 100 (24) 14229)

Staining with SYPRO Ruby for total protein

Each row consists of 2 10-fold dilutions of an NCI-60 cell Line or the control pool.

Representative candidate Abs prescreened for specificity by Western blots

Negative control
To monitor and analyze multiple, individual cells with subcellular resolution. Data reveals heterogeneity of kinetic responses and enables researchers to make better decisions by observing correlations that are not revealed in well-average responses.
Some characteristics of the emerging technologies

• Many measurements per sample (100s-10s of 1,000s)

• Highly sensitive, and so systematic effects due to time, place, reagents, personnel, etc can be visible

• Reproducibility can easily be compromised

• Amenable to careful experimental design

• Strong trend towards collecting more and more data: microfabrication, parallelism (1…96…1,000…50,000)
Some characteristics, cont.

• Sophisticated *instrumentation*, which *requires* understanding by the statistician/mathematician

• **Systematic** features of the data are at least as important as the random ones

• Often more like *industrial* process than single investigator lab research; usually not in control

• Integration of many data types…..*databases*
Affymetrix High Throughput Assay
From now on, S.C. = Statistical Challenge
S.C. 1: Low-level analyses

• Preprocessing, including removal of artifacts & normalization

• Summarizing signals (typically fluorescent or chemiluminescent) to give measurements

• Assessing quality of measurements

• Assessing consistency and strengths/weaknesses of different measurements of the same quantities (e.g. microarrays, qrt-pcr)

There are many challenges here, continuing ones.
A SELDI mass spectrum of human serum
Image analysis and probe quantification for GeneChip data

A very quick look
Affymetrix GeneChip® Probe Sets

Probes are 25-mers selected from target sequence.

5K-40K target fragments are interrogated by probe sets of 11-20 sequences called Perfect Match probes.

There are also corresponding MisMatch probes.
The rawest of raw data
500,000 probes x ~50 pixels each @ 16 bits/pixel
~ 50Mb tiff
Reduction to the .cel file: remove boundary pixels and use 75th quantile of remainder
Main image analysis issue: checking grid alignment
The raw data for our purposes (11 Mb ASCII)
A glance at some “raw” GeneChip data
20 probe spike-in set across 10 arrays

PM intensity

Concentration = array
16 probe spike-in set: 6 non-responding
16 probe non-spiked set constant across 59 arrays
16 such sets (out of 20,000)

Normalizing and summarizing this chip data is challenging.
Why (how) we normalize

Density of PM probe intensities for Spike-In chips

- After Quantile Normalization
Average and SD of 5 SN19 Samples stratified by average expression
HG_U133 Chips

Distribution of Average Expression by Stratum

Distribution of SD of Expression by Stratum
Brief comments on QA/QC for Affy chip data
Comparison of protein and mRNA measurements

Affymetrix chips cf Clontech Ab arrays (374 targets)
Comparison of male and female brain tissue (from DLPFC).

We compare log ratios of the genes with log ratios of the proteins, by looking at the difference of log ratios vs their means, and vs mean intensities, i.e. instead of plotting log fold change $M$ of one platform against that of the other, we plot the differences against their average, and again vs mean overall intensity $A$. We hope for tight scatter about 0.
Pre-processed protein M-values vs Affy gene M-values, colored by protein A-values (orange: A>11, red: A>13)

Nothing much here!
Differences of logratios vs mean intensities

Or here!
Possible reasons for no relation:

There simply is no such relationship between proteins and genes - not even one meeting the standards for quick & dirty approximations...

There is a relationship yet to be revealed, but only for a subset of proteins and genes which are obscured by others, or which are not contained in our sets.

The form of the relationship strongly depends on the actual gene.

Too few replicates (2).
Possible reasons, cont.

Protein data too noisy (antibody chips are still referred to as semi-quantitative)

Not enough differential expression

A problem with one or more brain samples

Errors in the matching between proteins and Affy IDs

Multiple matches obscuring existing relations.
S.C. 2: Intermediate inferences

- Discovering genes/proteins that are regulated, that interact, ...

- Discovering co-regulated genes/proteins

- Identifying pathways, processes that are perturbed, regions that interact, ...

- Composite forms of the foregoing

- Design questions: which experiments should be done?

- Much more of the same .. **continuing challenges**
In a microarray simulation study to examine the ability of statistical methods to identify perturbed genes, 50% false negative rates are typical. False positive rates vary widely depending on cut-offs and replication, but can be very high.

Y2H false positive rates may be about 50%.

Some of you may need to recalibrate your notion of measurement. These technologies are frequently more like screens in molecular biology than measurement in physical and engineering science and industry. However, they are improving, and .....data are data.
S.C. 3: Inferring models

Defining the models is the first challenge

Many targets: for biochemical reactions, gene transcription kinetics, cellular physiology, metabolic control...

Many types: Graphs, Networks (Boolean + extensions, Bayesian, Dynamic Bayesian), ODEs/PDEs, SDEs

The major problem: connecting the model (class) with the data in a reasoned (preferably statistical) manner, to estimate parameters (and associated uncertainty).

Very very little of this so far: a great (and never ending) challenge
Combining sequence, expression, and interaction data is beginning to occur, e.g. in discovering *cis*-regulatory transcription modules, assigning proteins to classes.

Some formal mathematical work already exists, e.g. using kernels and semi-definite programming, but I have yet to see it in the context of systems biology (that doesn’t mean it isn’t out there).

Another great and continuing challenge.
Case study 1: Yeast galactose utilization
Ideker et al., Science 292, 2001

a) Existing knowledge about the pathway organized.
b) Wildtype and 9 yeast deletion mutants were grown in the presence and absence of galactose, and 20 expression microarray experiments conducted. Abundance data collected (ICAT, MS/MS) on proteins from wt+gal and wt-gal conditions. Perturbed genes identified and clustered.
c) The existing pathway was enlarged to include data from protein-protein and protein-DNA databases, and displayed as a physical interaction network. Results from b) added.
d) Results of predictions from the existing model noted, and novel hypotheses (e.g. on double mutants) formulated and tested. Existing model enlarged.
Fig 4 from Ideker et al: yellow=DNA-protein, blue=protein-protein
Comments on this case study

• There was no mathematical model or mathematical analysis in this paper.

• Statistics entered directly in determining which genes and subnetworks were perturbed, and indirectly in interpreting the MS/MS data and in seeking binding sites as part of identifying putative interactions.

• Computation was restricted to graphical layout and network display, not inference.
Case study 2: Plant response to a pathogen
Glazebrook *et al* Plant J 34 ‘03, Wildermuth *et al* ‘04

a) Existing knowledge about the signal transduction network controlling defence (several interacting pathways, including SA, ET, JA) summarized.

b) Expression profiling of wt and 14 mutants at one or more times post-infection carried out, and perturbed genes identified and clustered along with mutants.

c) Some predictions were made tested on data from further assays (qrt-pcr, metabolites).

d) Conclusions summarized in a network. Model?
Case study 3: gene regulatory for development
Davidson et al Science 295 2002

I see their modelling, but their data and (deep) understanding of the system does not seem to have made use of “emerging technologies”
A systems approach proposed to predicting outcomes of therapies that target up- and downstream elements of the extended Epidermal Growth Factor Receptor (EGFR) signaling pathway. A very small part of this is on the next slide. The system to be used will be ~60 well-studied cell lines derived from normal and malignant breast tissues. A huge amount of data will be collected on these cell lines. Drugs will be added to them and their gene expression profiles will be studied over time using:

- The new Affymetrix High throughput Assay
- KineticScan (Cellomics)
- Reverse-phase protein microarrays: spot protein lysates e.g. 5,000 at a time and add antibodies
- Panels of siRNAs to knock-out genes
- ~200 genes will be sequenced in each cell line,
- and more.
Some of the EGFR pathways
Closing comment: forward, reverse, or..

Much of the systems biology research I have seen exemplifies general approaches in canonical examples, rather than asking what a systems approach can tell us about a specific system of interest to a random biologist, e.g. plant response to pathogen.

Ways of attacking a given problem from a system viewpoint need to be developed, not just illustrations of systems ideas in well chosen contexts. Perhaps the greatest challenge of all: developing tools and approaches to systems biology for general use.
Acknowledgements

WEHI
Tony Papenfuss
Phil Hodgkin

UC Berkeley
Mary Wildermuth
Julia Brettschneider
Adam Arkin
Paul Spellman (Joe Gray)

Both places: my many low level colleagues
Regulation of gene expression

1. Transcriptional control
2. RNA processing control
3. Transport control
4. mRNA stability control
5. Translation control
6. Post-translational control

DNA → primary RNA transcript → mRNA

NUCLEUS → CYTOSOL

mRNA → protein

Active/inactive protein

Nuclear envelope
Transcriptional Regulation in Eukaryotic Cells

RNA polymerase II complex

CTD

Snf2
Sin4
Med6
Srb4
Srb6
Srb10
TFIIE
Kin28
RNAP II
TFIIF
TFIIH
TFIID

Transcription factors

SAGA
Hybridization in short oligo arrays

Target

Probes on probe cell
A Mass Spectrometer

“An analytical device that **determines the molecular weight of chemical compounds** by separating molecular ions according to their mass-to-charge ratio (m/z)”

- **Ionisation**
  - molecular weight = 600
  - abundance = 50%
- **Separation**
  - molecular weight = 400
  - abundance = 20%
- **Detection**
  - molecular weight = 300
  - abundance = 30%
A pair of duplicate SELDI spectra

Spectra from blood serum, after baseline subtraction
MS/MS

• To gain more information about the detected masses:
  – one of the detected products is selected;
  – we induce a collision with a gas to break it;
  – we do a second MS to identify the pieces.

  – different molecules of the same substance can split in different ways.
  – in each molecule, only the pieces that retain one of the charge will be observed and present in the spectrum; the others are discarded.
Use of MS for protein identification (1)

- Easiest way: find the molecular weight of a protein.
- Allows us to rule out many of the putative proteins.
- Not enough discrimination: many different proteins have the same weight because of the number of way the amino acids can be combined.

Ex: masses of proteins with only 5 amino acids can range between 285 and 930, but there are $20^5 = 3.2$ M possible proteins of length 5, that is about 5000 different proteins for each mass. If the length increases, the situation gets even worse.
To get more specific information, use Peptide Mass Fingerprinting (PMF):
- select one protein (e.g., do a 2D-GEL and excise one spot);
- digest it with a protease (usually trypsin);
- do a MS on the result; the resulting list of masses is the protein’s fingerprint;
- search for the fingerprint in a database.

Example: the fingerprint of MALKCGIRGGSRPFLRATSKASRSDD is (333, 336, 406, 448, 462, 889)

Just as with a real fingerprint, you cannot identify the owner just using the fingerprint; you need it to be in the database. If this is not the case, PMF does not help you.
Use of MS for protein identification (3)

- Digest the protein, then use MS/MS
- In the collision process, the peptides will usually break at the peptidic bond, which is the weakest bond.
- Example: GPFNA, mass = 57+97+147+114+71=486 Da
Analysis of MS/MS spectra

- If the spectrum has a good quality, you can interpret it directly, without a database («de novo sequencing»).
- Otherwise, you can do a database search
  - Most well known search algorithms: Mascot and Sequest
  - Search in a peptide sequences database the peptides that are the more likely to produce the observed spectrum.
  - Compute a score for every entry in the database and return the top scoring peptides.

SEQUEST v.27 (rev. 11), (c) 1993-2000
(M+H)+ mass = 1580.5800 ~ 3.0000 (+2), fragment tol = 0.5, MONO/MONO
# amino acids = 228010264, # proteins = 715871 , # matched peptides = 205749
(M* +16.000) (C# +105.000) Enzyme:Trypsin (2)

<table>
<thead>
<tr>
<th>#</th>
<th>(M+H)+</th>
<th>XCorr</th>
<th>Ions</th>
<th>Reference</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1580.8324</td>
<td>4.7460</td>
<td>22/26</td>
<td>sp</td>
<td>Q9WUB3</td>
</tr>
<tr>
<td>2</td>
<td>1581.7746</td>
<td>2.7348</td>
<td>15/24</td>
<td>tn</td>
<td>CAC46405</td>
</tr>
<tr>
<td>3</td>
<td>1581.8497</td>
<td>2.5998</td>
<td>14/24</td>
<td>tr</td>
<td>Q89202</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

71
Crosslink protein to DNA \textit{in vivo} with formaldehyde

Break open cells and shear DNA

Immunoprecipitate

Reverse-crosslink
Genome wide location analysis (II)

Genomic DNA  IP-enriched DNA

Target Amplification
Ligation-mediated PCR

Yeast cDNA microarrays
6,361 intergenic DNA sequences

Data analysis
Hosts and Pathogens Compete for Resources

- Carbon, Nitrogen, Iron
- Specialized Products

Small Molecules

Signal Transduction & Regulatory Pathways

Altered Metabolism & Physiology
Plants and Microbes employ Common Small Molecules in their Interactions

Plant

ROS  NO

Et  SA  JA

Pathogen

↓

Signal Transduction & Regulatory Pathways

↓

Altered Metabolism & Physiology
# Modeling Plant Defense Against Pathogens

- Biosynthetic pathways of small molecules that regulate host defense pathways
  - overlay pathogen and host annotated genomes
- Architecture of Host Regulatory Pathways
  - overlay response to different pathogens
- Global changes to metabolism and physiology
  - couple with physiological responses
- Mathematical Formulation
  - metabolic control analysis, phenotype phase plane analysis
Arabidopsis thaliana as a model plant

- Susceptible to full range of pathogens--bacterial, fungal, viral, insects
- Exhibits non-host resistance, gene-for-gene resistance, and susceptible responses to pathogens
- Mounts major known defenses--e.g. synthesis of phytoalexins, PR genes, etc.
- Extensive genomic and genetic resources
  - Small genome (130 MB), sequenced!
  - Mutants, including knockouts (SALK, AFGC)
  - Tools--DNA chips, Transformation protocols, etc.
Model Pathogen: *Erysiphe orontii*

- Fungal biotroph *Erysiphe orontii*
  - causes powdery mildew (visible symptoms, scoring system)
  - infection dose can be controlled and limited to single spore
  - infection limited to epidermal cells
  - virulent on *Arabidopsis* ecotype Columbia
  - collection of *Arabidopsis* mutants with altered susceptibility to *E. orontii*
  - common agronomic and ornamental pathogen
Modeling *Arabidopsis* Defense Against *Erysiphe*

- **Disease Score** = \( f(\text{RC}, \text{AD}) \)
  - \( \text{AD} = \) Active Defenses
  - \( \text{AD} = f(\text{JA}, \text{Et}, \text{SA}) \)

- Active Defenses
  - ROS, NO
  - JA, Et, SA

- Signal Transduction & Regulatory Pathways
- Altered Metabolism & Physiology
- Disease Outcome

- Resistance <-> Susceptibility

Define critical small molecules
Define biosynthetic pathways
Overlay metabolic pathways from annotated genomes
Search for all reactions/enzymes that use a given compound