UNIVERSITY OF OXFORD

Systems Biology Doctoral Training Centre

Oxford Physics

Machine Learning for Systems Biology

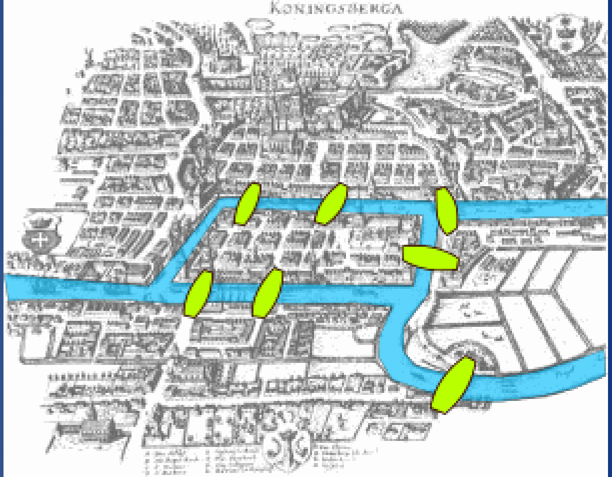
Sumeet Agarwal

(joint work with Gabriel Villar and Nick Jones)

A general perspective on network science

- The study of graphs and networks goes back at least to Euler. People from a wide range of disciplines have contributed: Mathematicians, Computer Scientists, Electrical Engineers, Sociologists, Physicists, Statisticians...

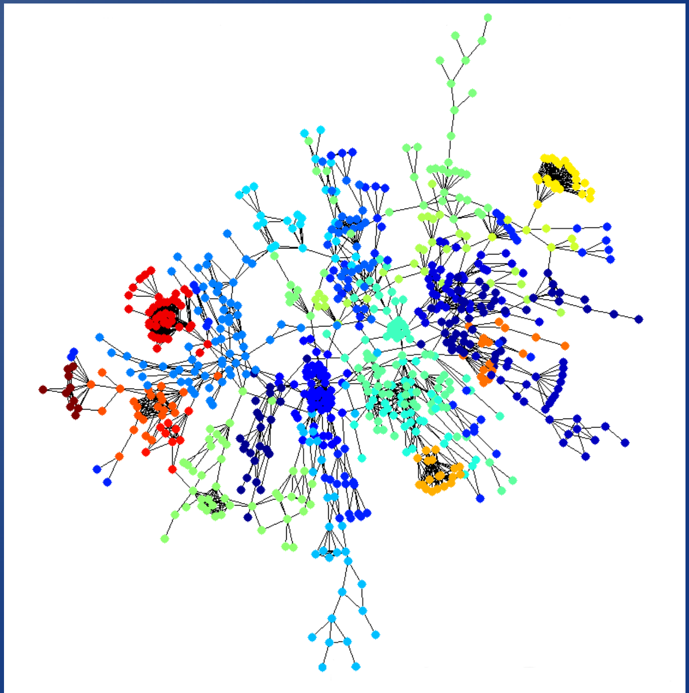
- This has led to a fragmented literature, with inconsistent terminology and frequent reinvention of concepts and methodologies



Courtesy: Wikipedia

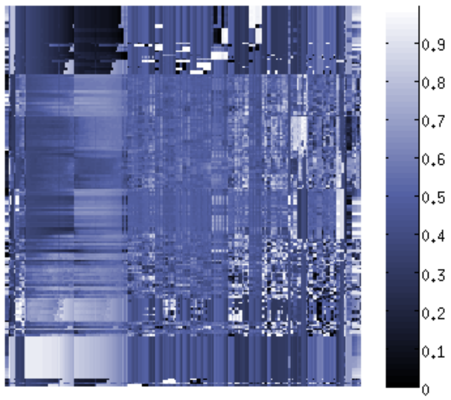
- Our aim is to utilise the power of computing and machine learning techniques to construct a comprehensive database of networks and network algorithms, and use this to systematically investigate patterns of relationships between different kinds of networks and metrics/features

- This kind of data-driven approach may allow us to choose the most relevant features for a given task, motivate appropriate network models, and in general answer the question: What are the best ways of thinking about networks?

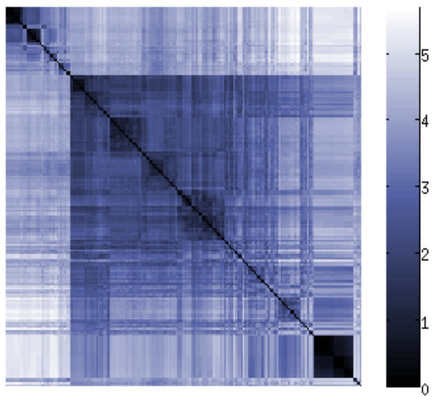


Comparative network analysis

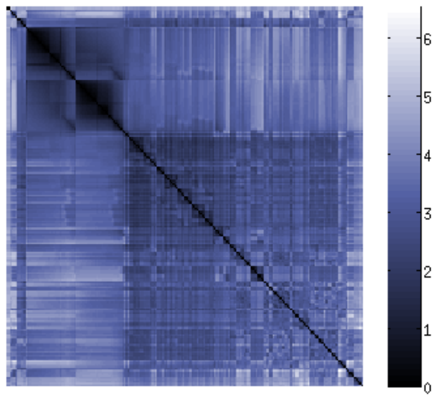
- An attempt to study network properties at a rather abstract level, using computing power to automate many different analytic procedures across many different networks
- This gives us a matrix of networks versus metrics/features, which can be mined to identify features and networks of interest, cluster them into ‘families’, learn predictive models for system phenotype etc.
- It is a way of organising and systematising the diverse range of network analysis techniques to give us a better sense of the current state of the field



Data matrix:
networks vs. metrics



Correlation matrix:
networks vs. networks



Correlation matrix:
metrics vs. metrics

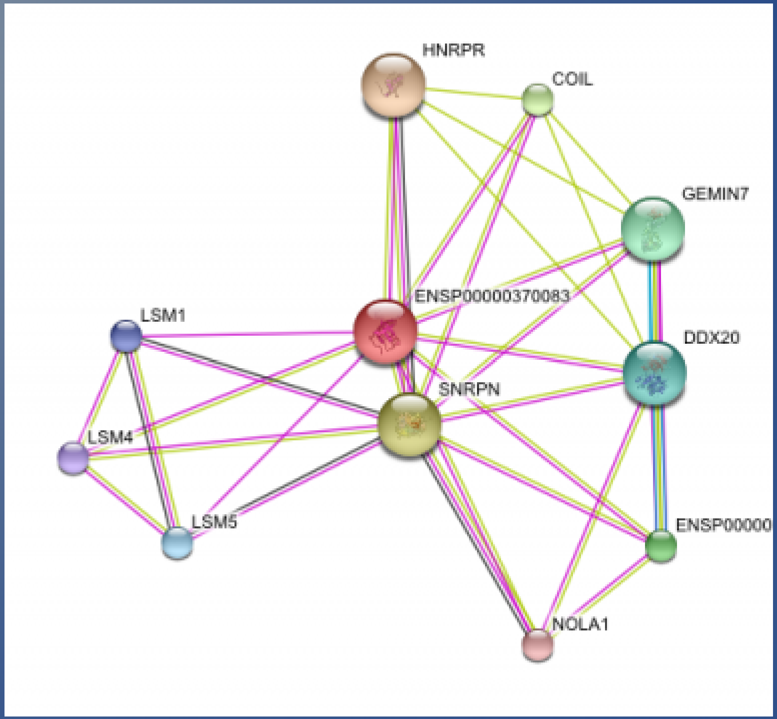
What kinds of networks do we study?

- Network representations have been used to study a wide variety of data:
  - Technological networks (railways, telephone lines, internet)
  - Information networks (WWW, cell phones, e-mail)
  - Social networks (friendship/kinship, Facebook, Twitter)
  - Biological networks:
    - Ecological
    - Neural
    - Subcellular (metabolic, protein-protein, gene regulation)
- We attempt to gather as many data sets as we can from different sources, and also construct synthetic data sets for comparative purposes

What kinds of metrics do we study?

Simple numeric features: size, assortativity (degree correlations), mean path length

Summaries of feature distributions over nodes/links: degree, centrality measures, clustering coefficient

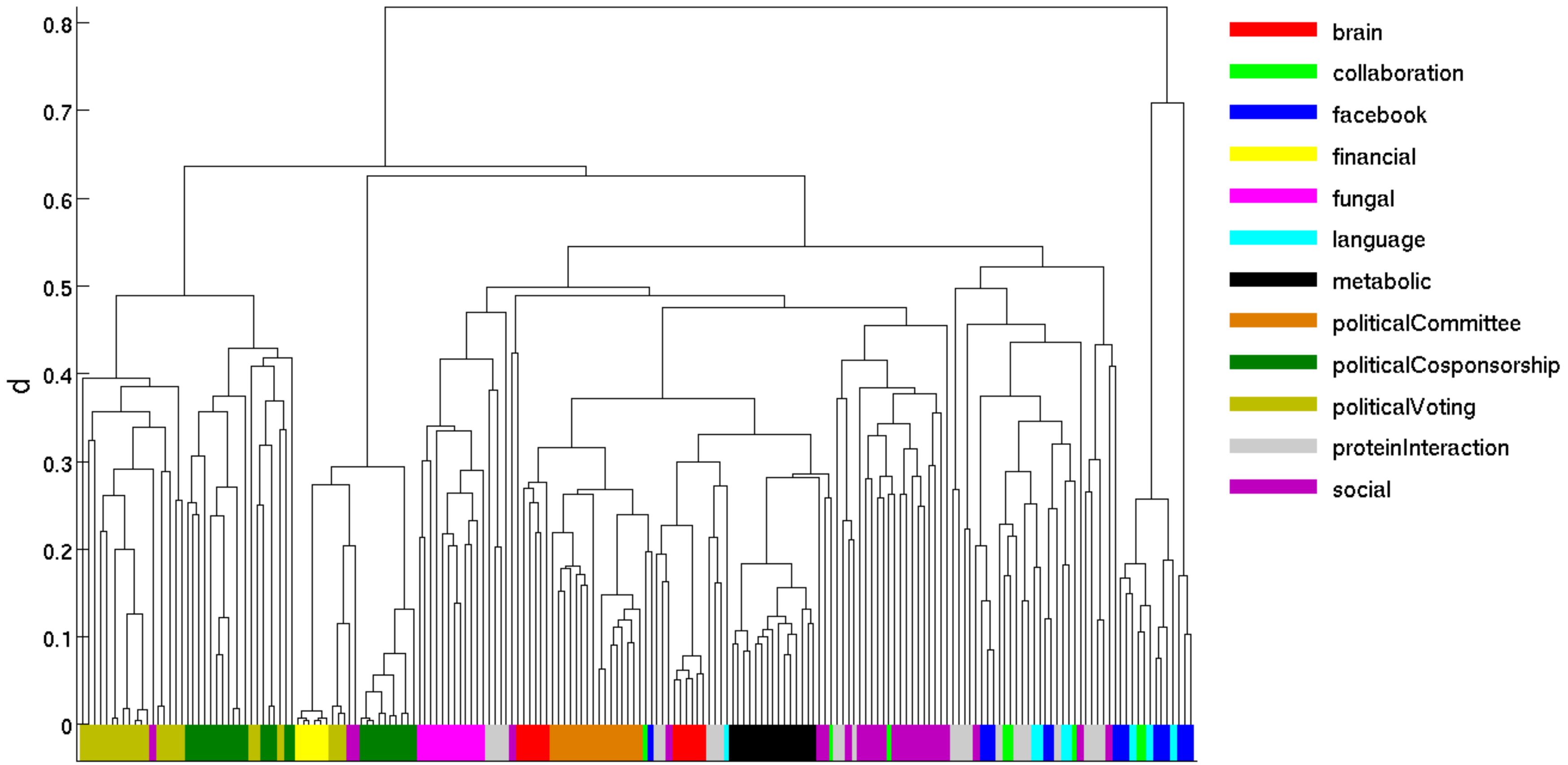


Community structure: partition entropy, modularity, coarse-grained networks

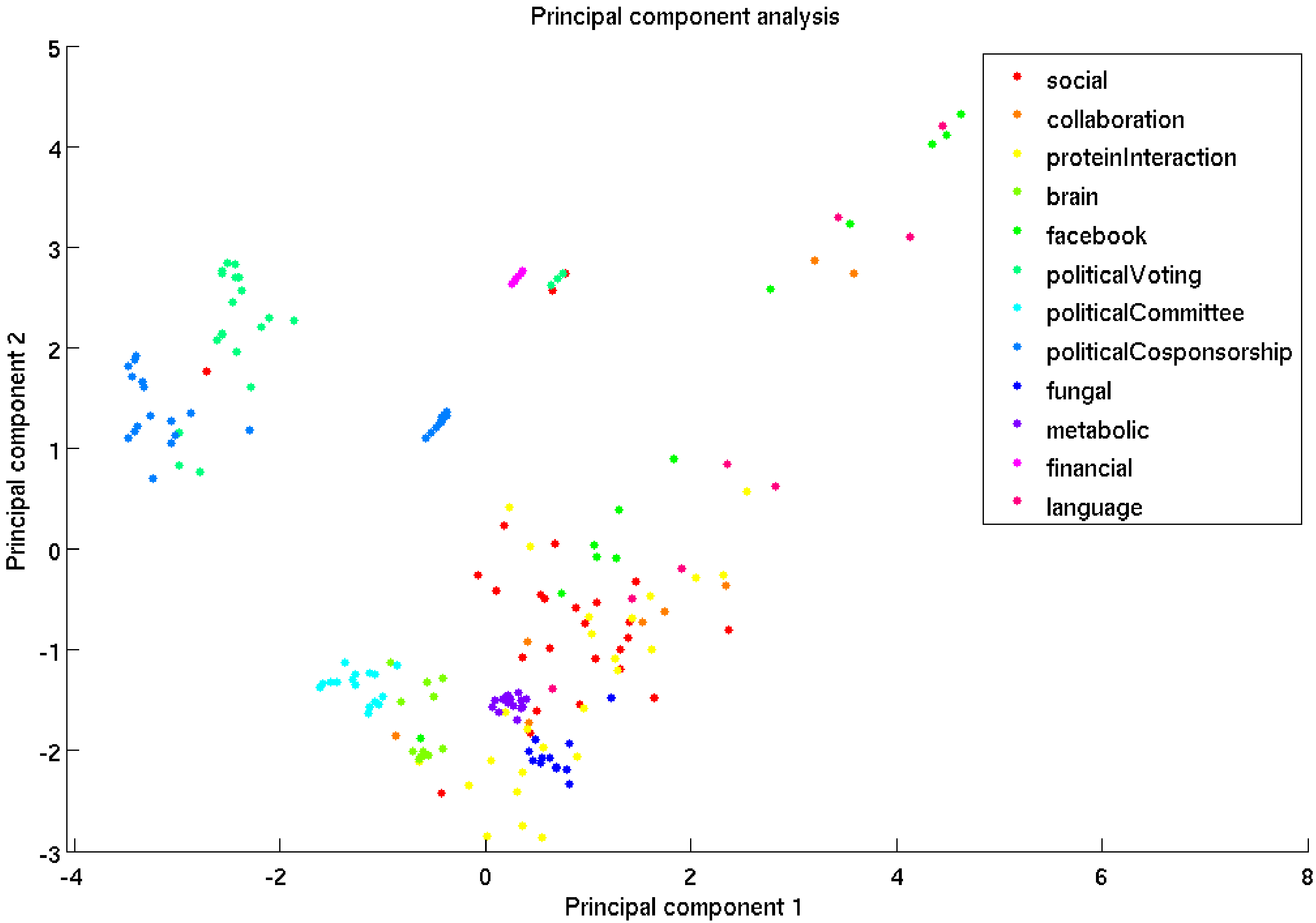
Model fits: how well the network is explained by a certain generative model (preferential attachment, duplication and divergence)

Other quantities such as motif counts, linear algebra operations (eigenvectors, Laplacian) on adjacency matrix

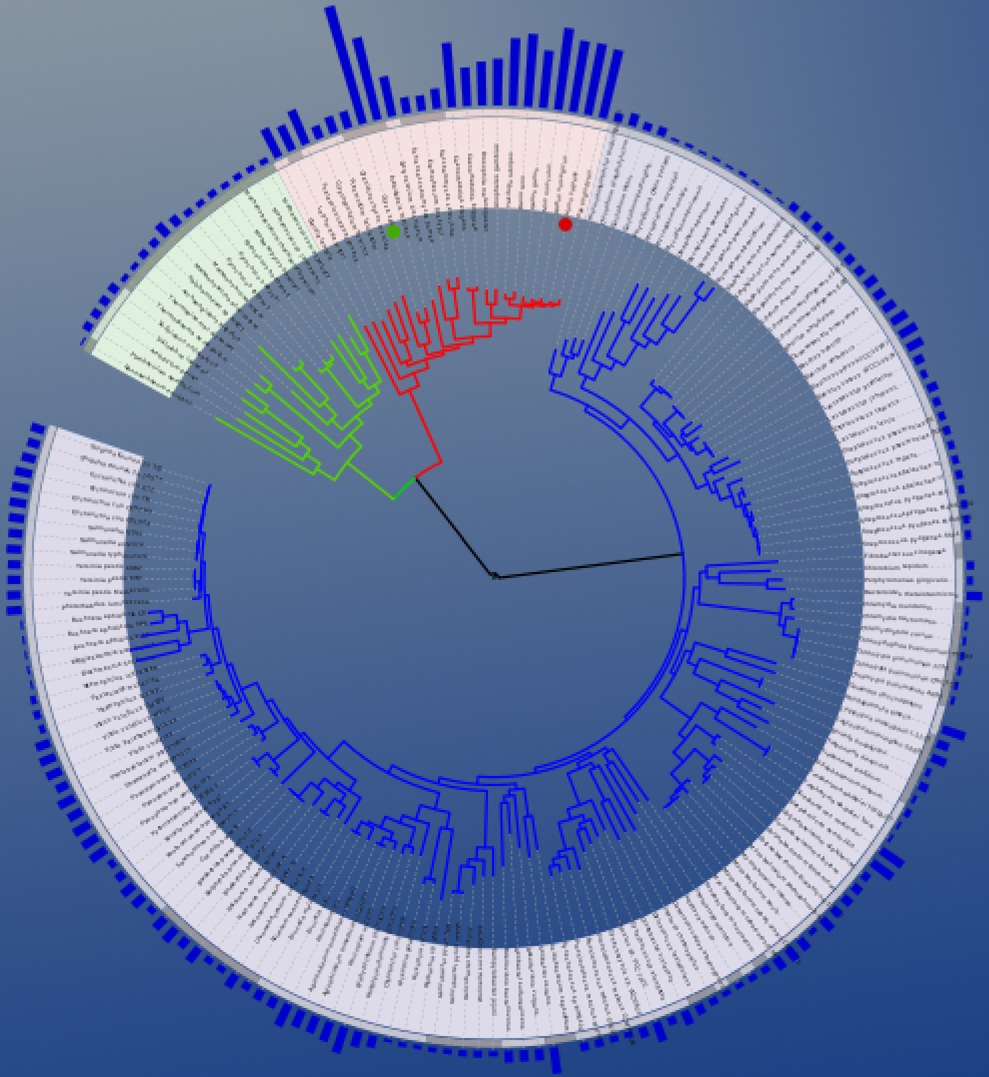
Network Families: Single linkage clustering



Network Families: Principal Component Analysis



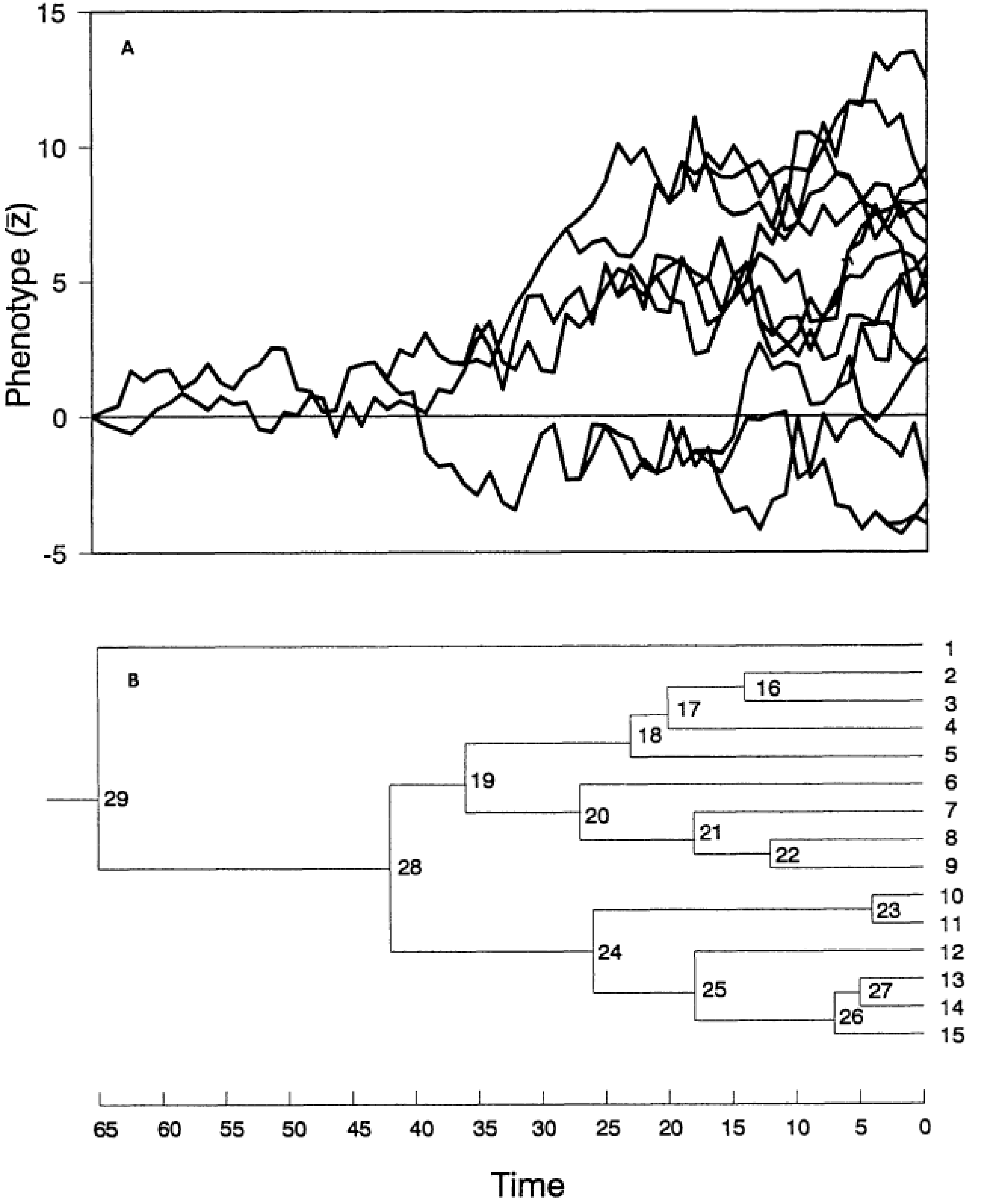
Example: Phylogenetic Comparative Methods



Courtesy: Wikipedia

- We can use features of biological networks in conjunction with independent evolutionary phylogenies to search for 'phylogenetic signals', i.e., properties that are most conserved in closely related species
- The idea is to assume a statistical process governing the evolution of any given trait (e.g., Brownian motion), and compute the likelihood of seeing the observed distribution of trait values at the leaves of the tree

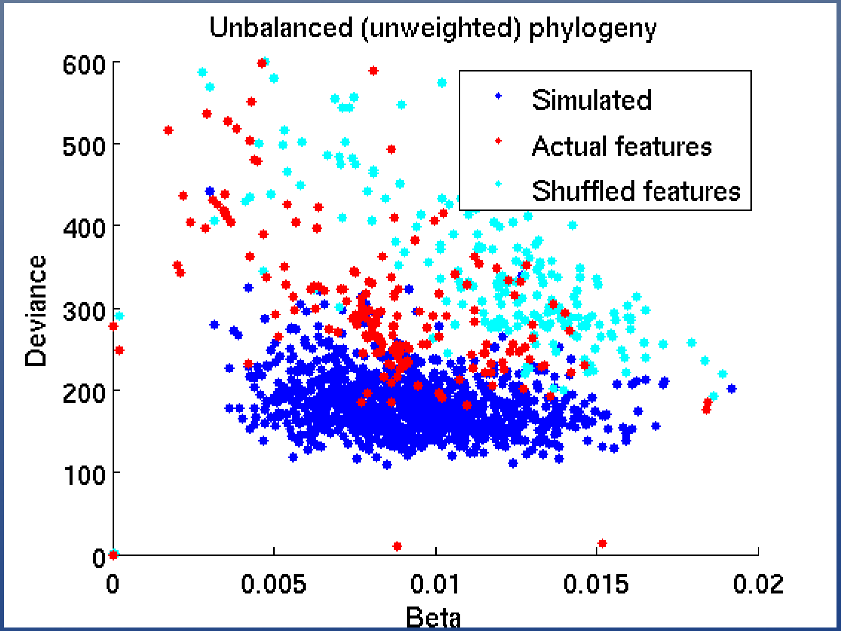
- We attempted to fit a Brownian motion model of evolution (V = βt + ε) to 272 real-valued network metrics computed on 450 metabolic networks from 158 different genuses, using a phylogeny taken from the Tree of Life

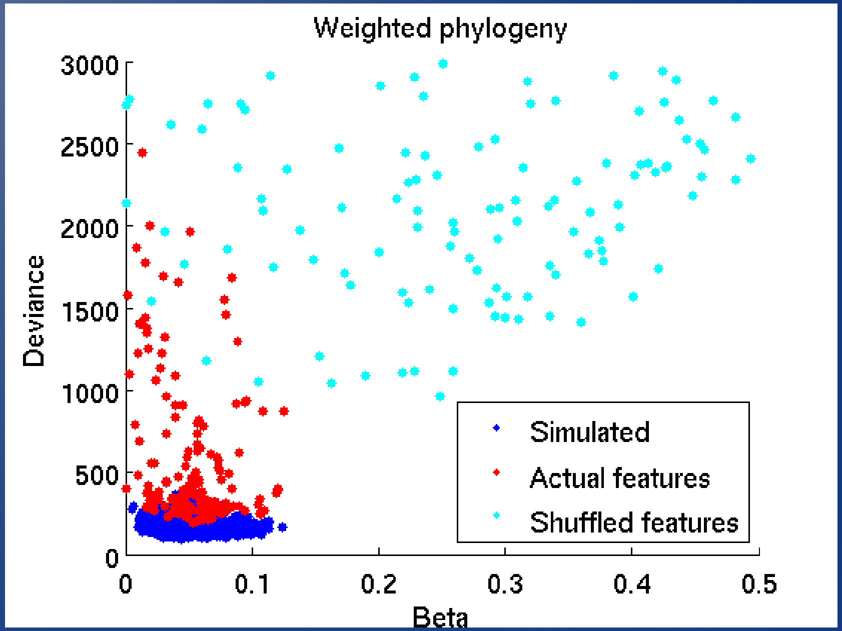


(Emilia P. Martins, *Am. Nat.* 1994)

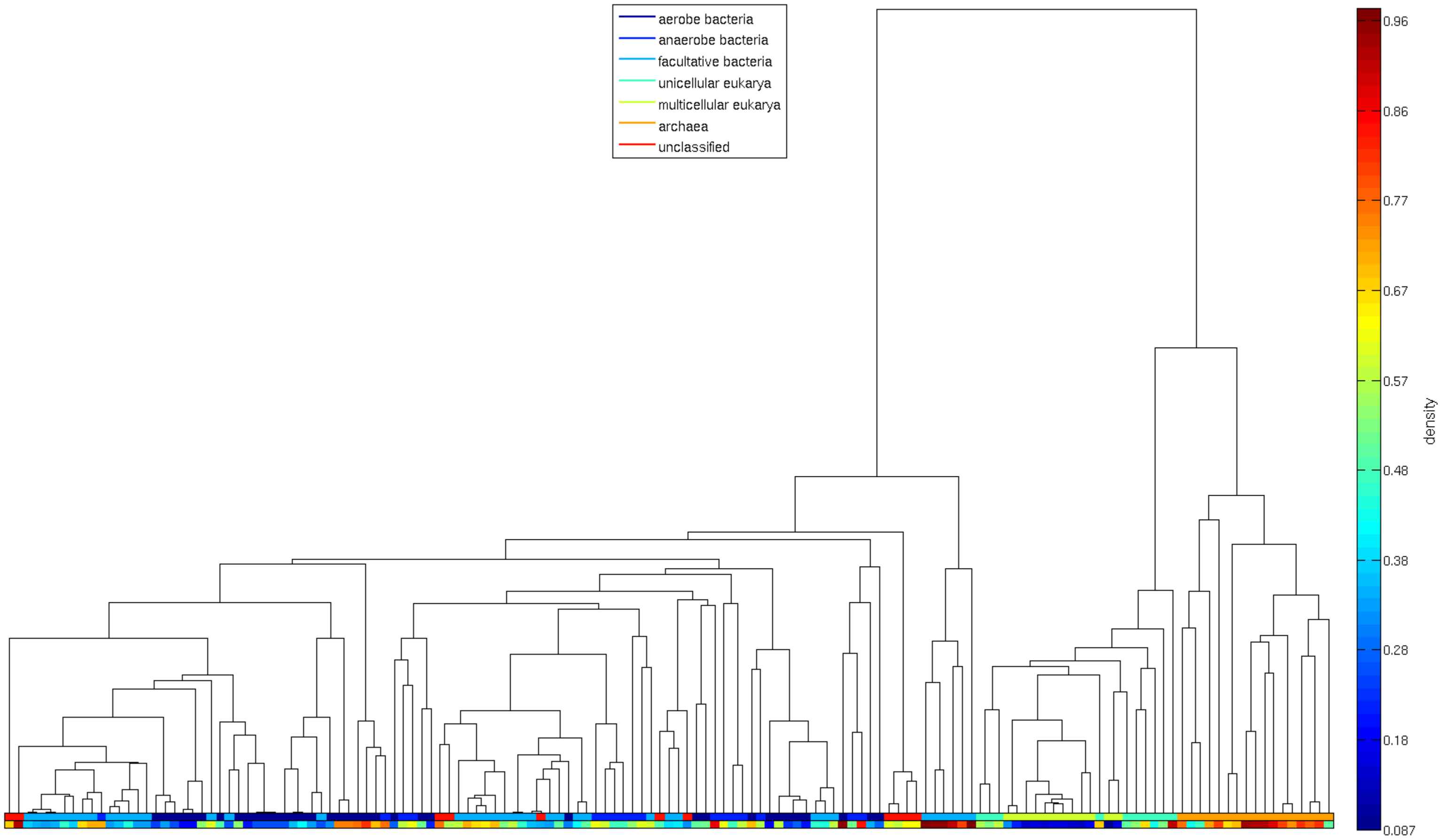
A realistic phylogeny gives significant feature correlations

- An unbalanced version of the tree (with no branch weights) was compared with a weighted version (based on actual estimates of evolution times)
- We used deviance (sum of sqaures of the residuals, ε) as a measure of the goodness-of-fit of the model for each metric/feature



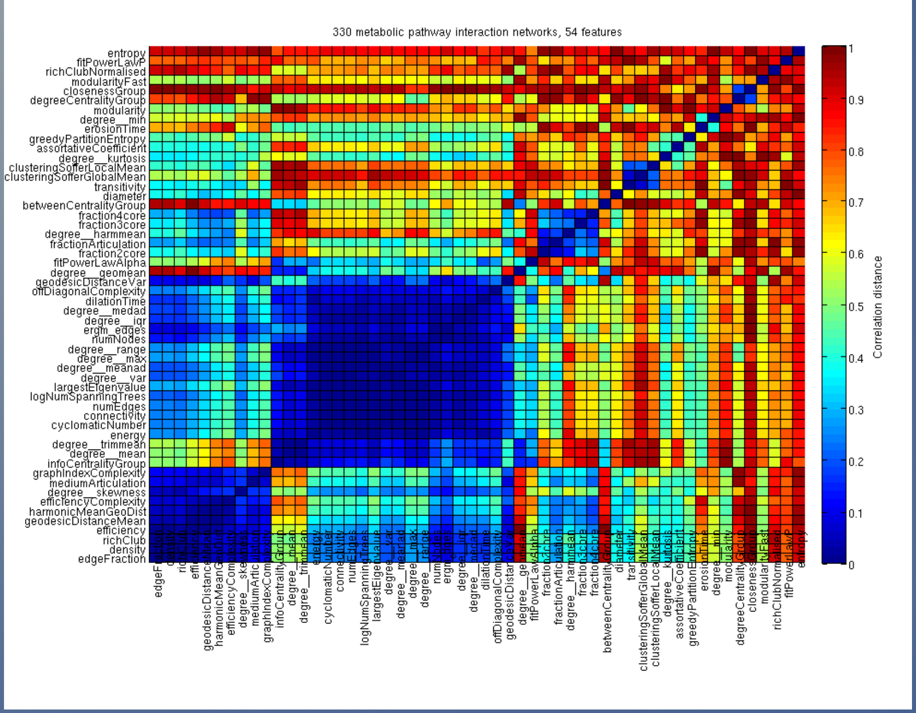


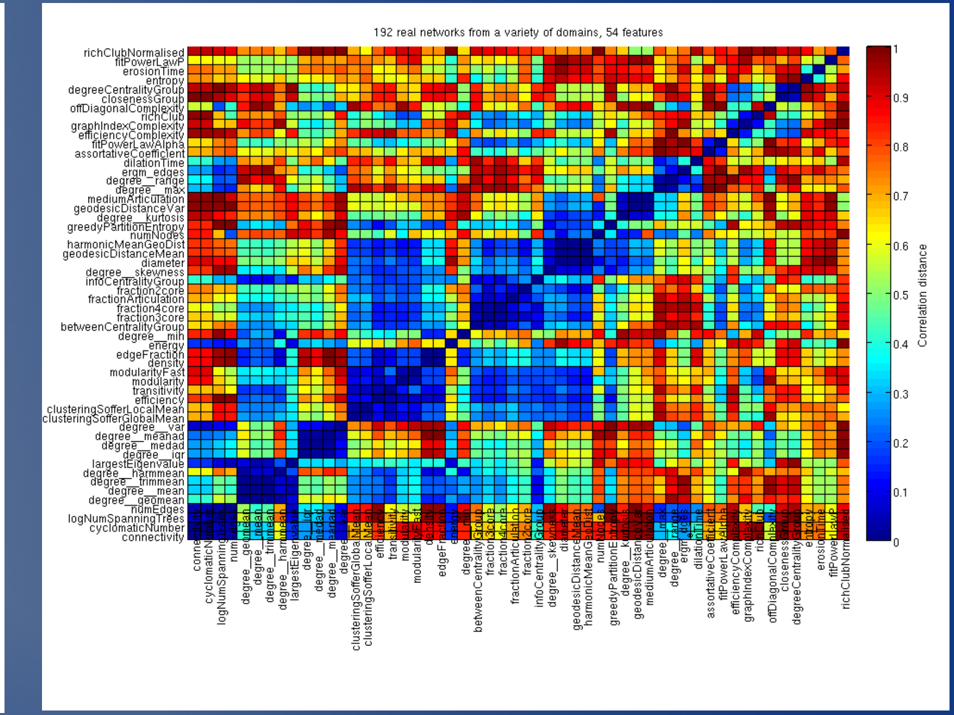
How do networks features vary across the phylogeny?



- Such approaches can be thought of as one way of resolving a debate over the nature of biological taxonomy: pheneticism (Linnæus) vs. cladism (Darwin)

Feature correlations: pointers to 'simplicity' in nature?


330 metabolic pathway interaction networks, 54 features


192 real networks from a variety of domains, 54 features

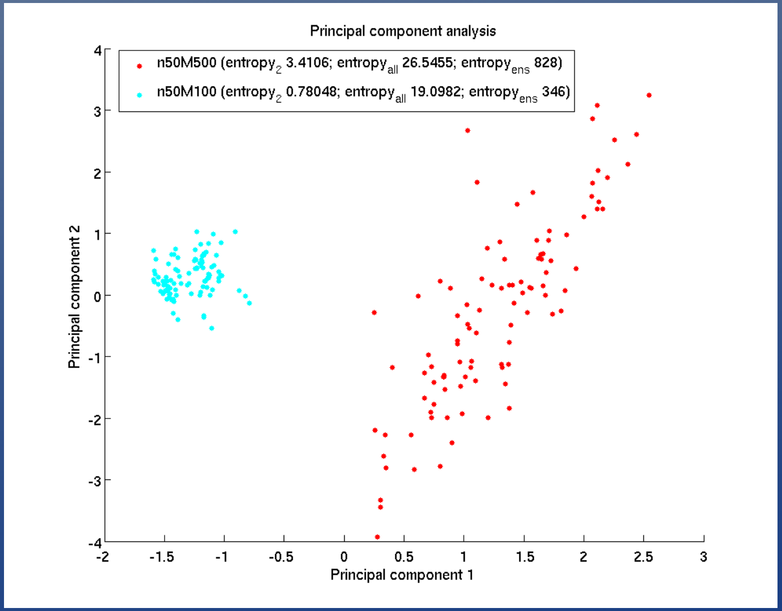
- For restricted classes of networks, many generically different ways of thinking about or characterising networks appear to become degenerate
- Perhaps functional network classes sit on low-dimensional manifolds in the high-dimensional structure space
- One way to think of this is that real-world network categories have relatively low entropy, because they have evolved under entropy-lowering constraints. Can we use such observations to actually recover the underlying generative constraints or mechanisms?

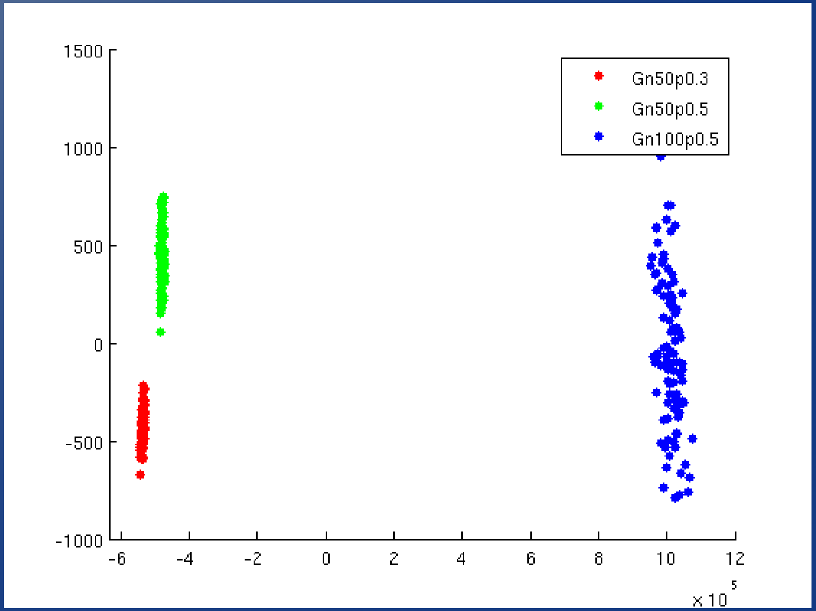
An 'empirical' measure for network entropy?

- We can think of a model or ensemble of networks as specifying a probability distribution over all possible networks; and thus we can define the entropy of this distribution in the standard way. For simple models this can be computed analytically. E.g., for the ensemble G(N,L) (networks with N nodes and L links), the entropy is given by

$$H = -\Sigma\, p_i \log p_i = \log {}^{N(N-1)/2}C_L$$

- Using our method we can also generate a sample from a given ensemble, embed it in a feature space and compute its empirical entropy that way. How do these two measures of entropy match up?



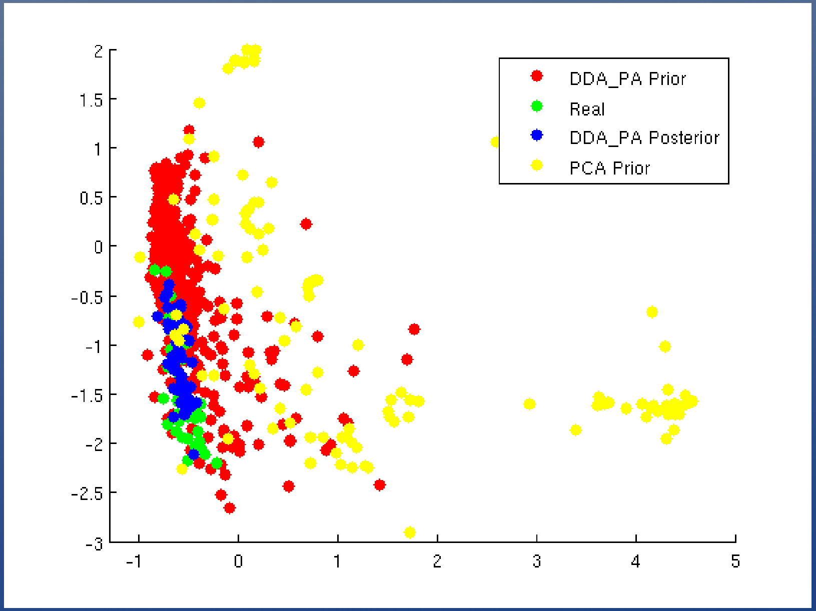


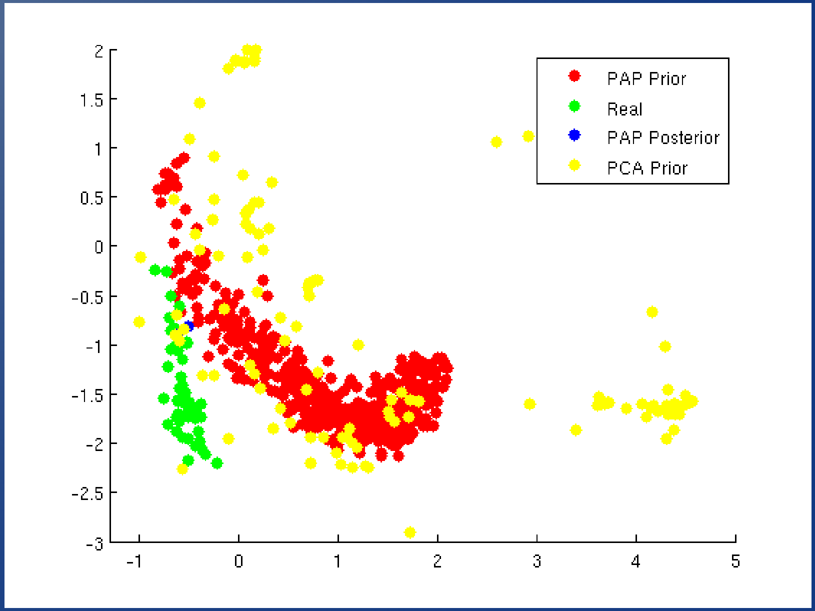
Recovering network models

- The fact that our low-dimensional network embedding allows us to estimate entropy suggests that we could use this for fitting appropriate models to real networks, using the related approach of Approximate Bayesian Computation:

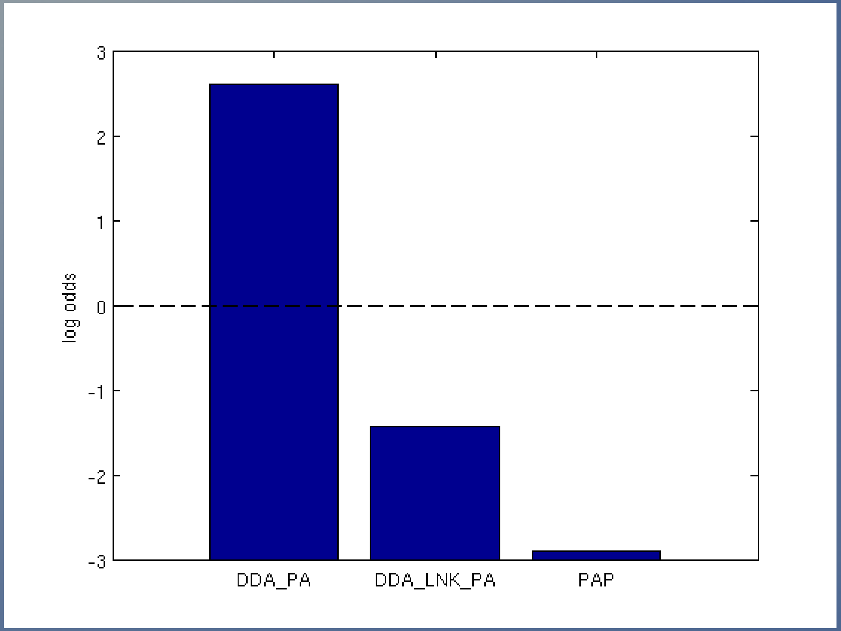
$$P(M|D) \sim P(D|M).P(M)$$

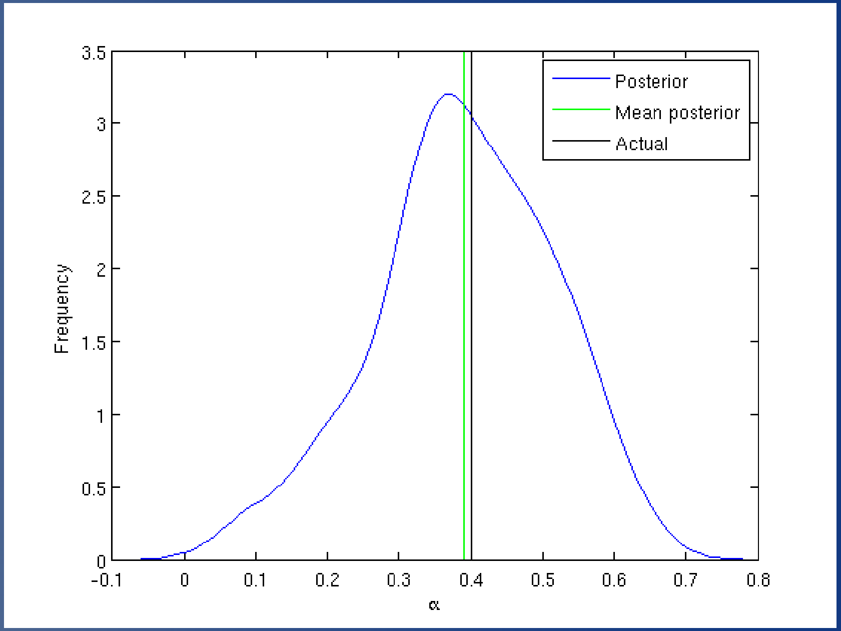
- We have tried generating synthetic networks using a model proposed for the evolution of protein-protein interaction networks, to see how well we can recover the model and its parameters

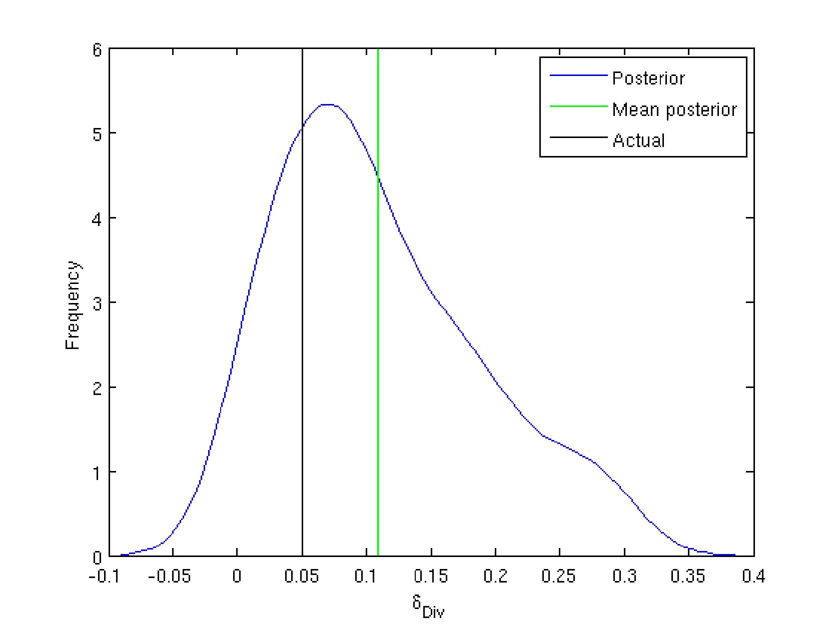


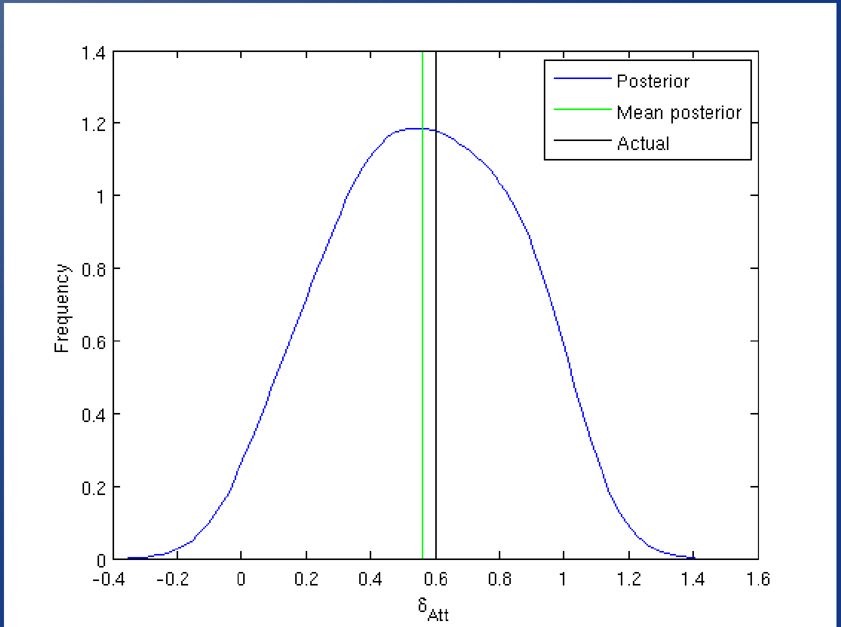


Recovering models with parameters





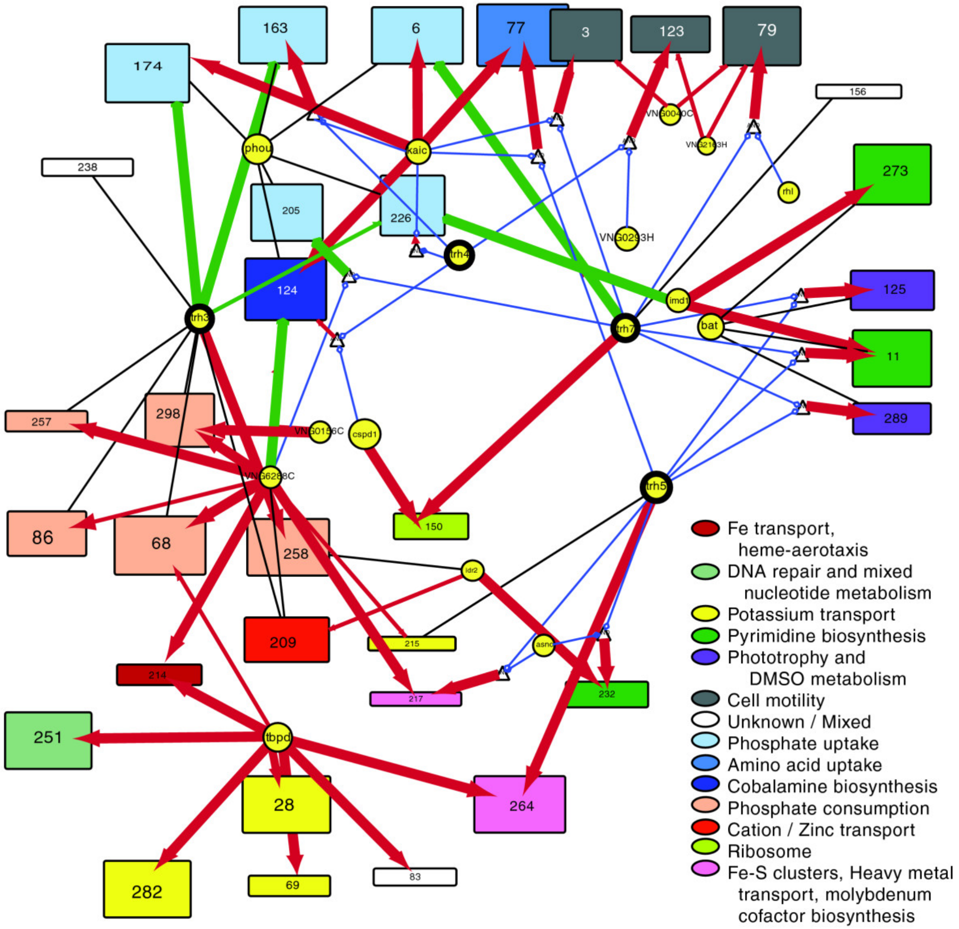




Conclusions

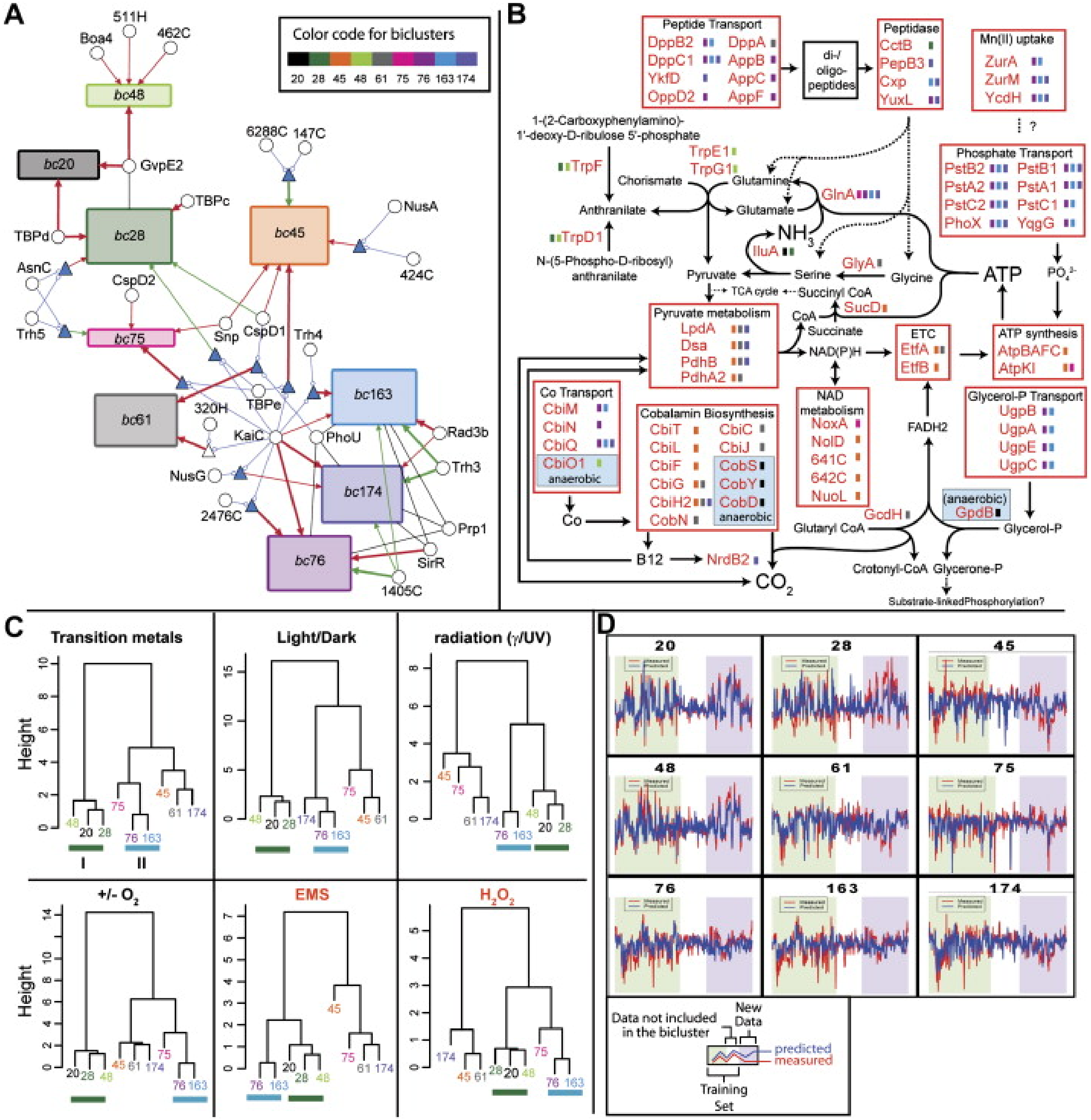
- Our approach is an attempt at systematically comparing and categorising a variety ways of measuring network structure and properties, and also looking at robustness and scaling properties of different metrics
- A data-driven approach to examining large numbers of networks and metrics is useful for feature selection in classification tasks, identifying redundant metrics and matching real-world networks to appropriate generative models
- Quantifying the significance of biological network features in the context of evolutionary phylogenies provides one approach towards the problem of establishing relationships between network structure and function
- We have demonstrated several different applications of the framework, corresponding to different ways of relating network structure to behaviour/complexity; ultimately it provides a tool which can give meaningful results only in the context of an appropriately framed scientific question

Dynamics and Inference on Biological Networks



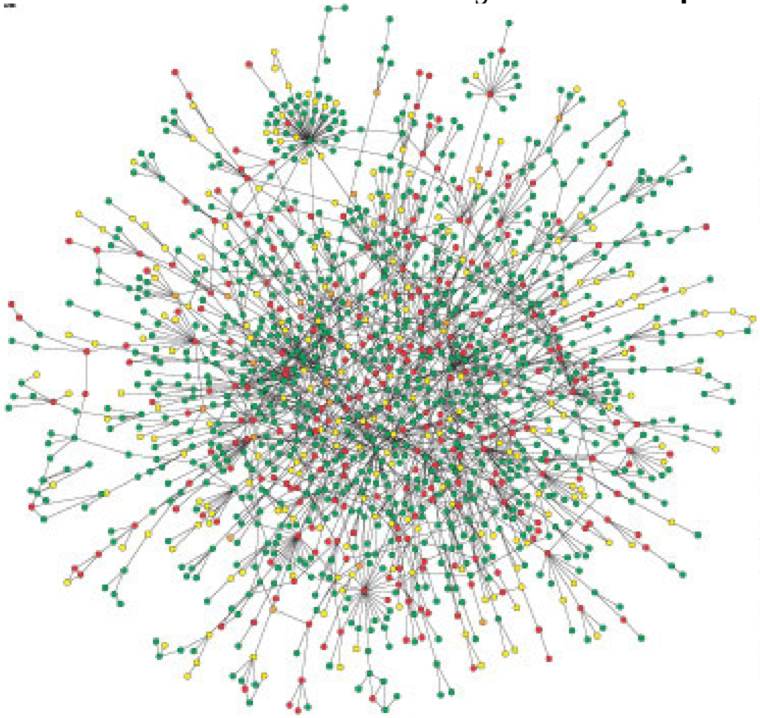
What do we mean by ‘dynamics’?

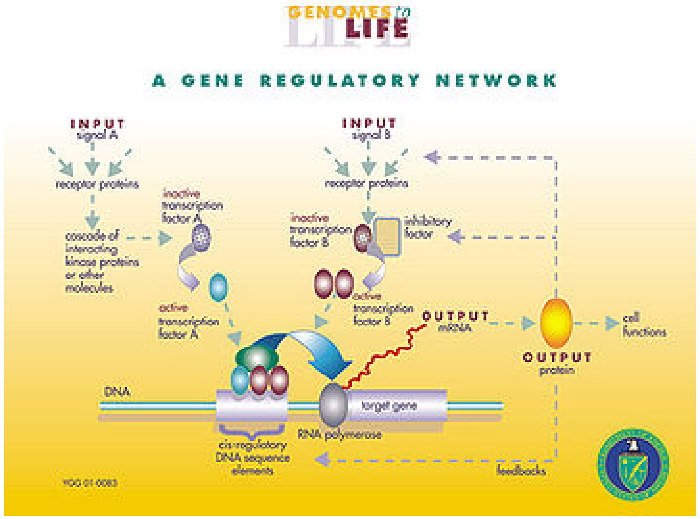
- Basically: things are changing with time
- For networks, each node may represent a time-varying quantity, such as gene expression levels
- Links may also change with time, or have a weight/strength dependent on the endpoints



Why study dynamics?

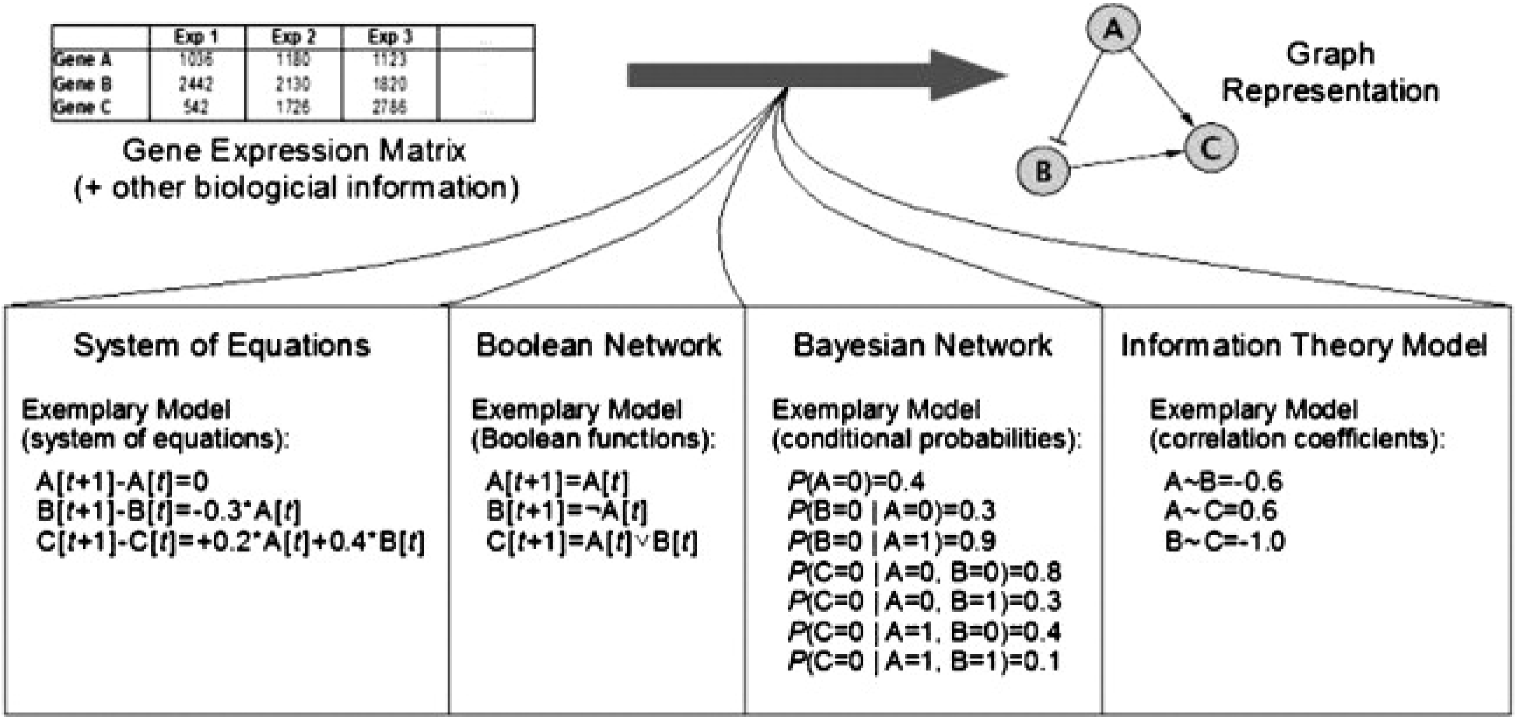
- Most networks/systems in the real world do change over time
- Studying dynamics tells us about certain properties of the system: both global (steady states, attractors) and local (causality)
- A model of dynamics can be used to make predictions about the future, and also about how the system responds to perturbation





Mathematical Representation

- The most preferred method, if feasible, is ordinary differential equations (ODEs), but for most large-scale systems we have to use appropriate simplifications for tractability



	Exp 1	Exp 2	Exp 3	...
Gene A	1036	1180	1123	...
Gene B	2442	2130	1820	...
Gene C	542	1726	2786	...

Gene Expression Matrix
(+ other biologicial information)

System of Equations

Exemplary Model
(system of equations):

$A[t+1]-A[t]=0$
$B[t+1]-B[t]=-0.3^{*}A[t]$
$C[t+1]-C[t]=+0.2^{*}A[t]+0.4^{*}B[t]$

Boolean Network

Exemplary Model
(Boolean functions):

$A[t+1]=A[t]$
$B[t+1]=\neg A[t]$
$C[t+1]=A[t]\vee B[t]$

Bayesian Network

Exemplary Model
(conditional probabilities):

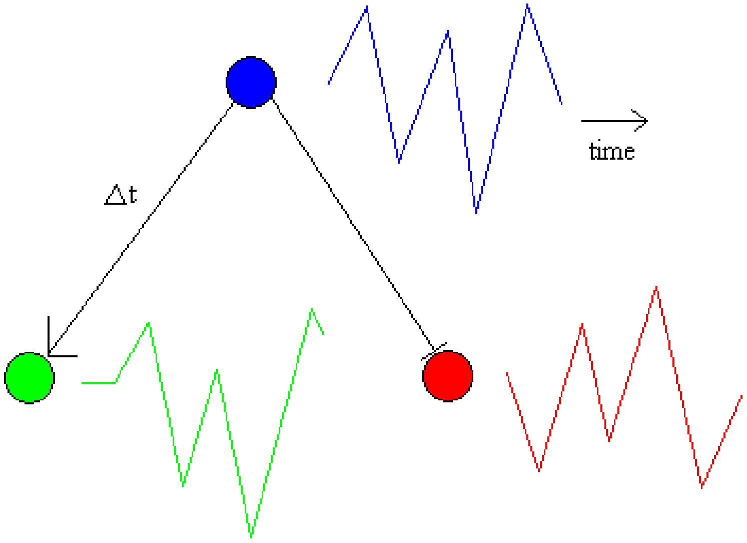
$P(A=0)=0.4$
$P(B=0 \mid A=0)=0.3$
$P(B=0 \mid A=1)=0.9$
$P(C=0 \mid A=0, B=0)=0.8$
$P(C=0 \mid A=0, B=1)=0.3$
$P(C=0 \mid A=1, B=0)=0.4$
$P(C=0 \mid A=1, B=1)=0.1$

Information Theory Model

Exemplary Model
(correlation coefficients):

A~B=-0.6
A~C=0.6
B~C=-1.0

Link with time series

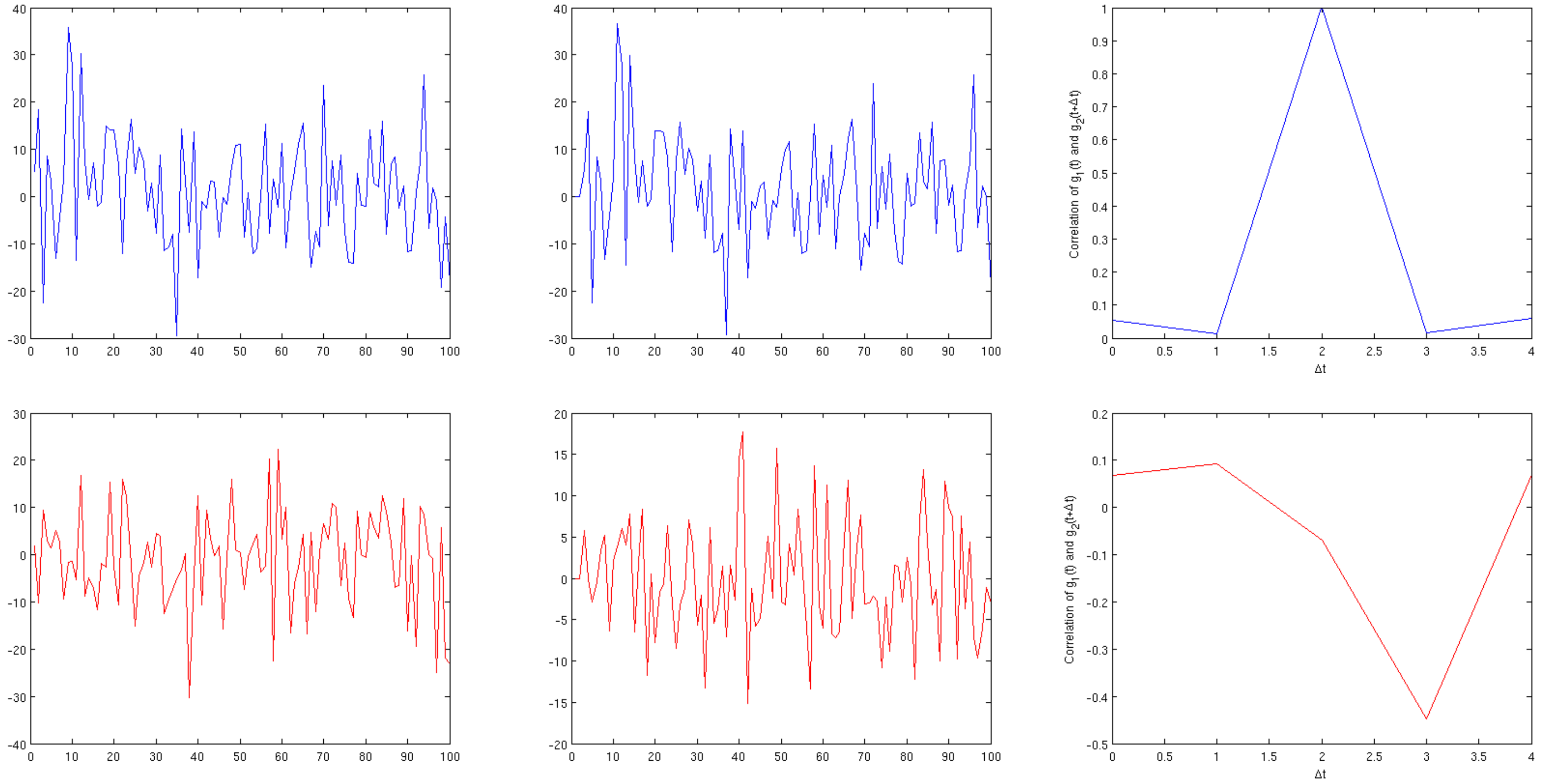


- The behaviour at each node can be described by a time series tracking an appropriate quantity (body weight, gene expression,...)
- In practice, raw observations are often in the form of time series data, which can be used (possibly combined with other information) to construct a plausible network
- Statistical techniques for time series are widely used for this purpose

Intermission

How to do Inference in Two Easy Steps

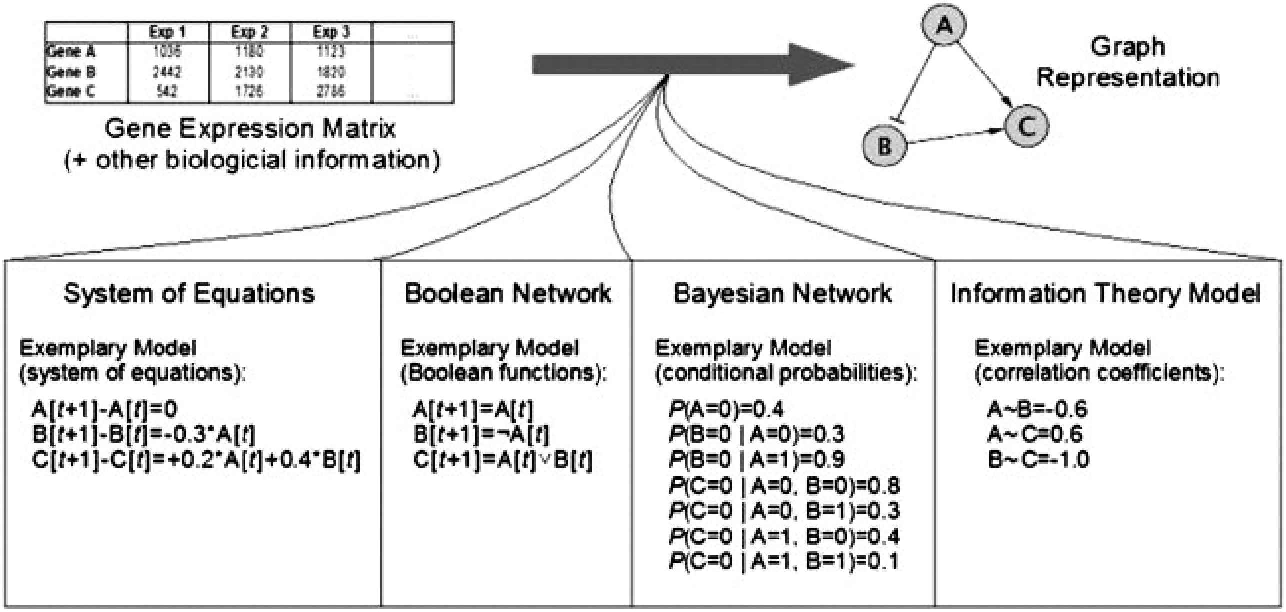
Gene 1 Gene 2 Cross-Correlograms



What can you infer about the relation between the two genes in each case?

Causality in networks

- Is a somewhat vexed philosophical notion; the classic 'correlation vs. causation' conundrum
- For our purposes, it just means we're trying to understand, at some given level (say proteins in a cell) whether changes in one entity lead to changes in another
- We can represent causal relationships in a network, and model them more precisely mathematically



	Exp 1	Exp 2	Exp 3	...
Gene A	1036	1180	1123	...
Gene B	2442	2130	1820	...
Gene C	542	1726	2786	...

Gene Expression Matrix
(+ other biologicial information)

System of Equations

Exemplary Model
(system of equations):

$A[t+1]-A[t]=0$
$B[t+1]-B[t]=-0.3*A[t]$
$C[t+1]-C[t]=+0.2*A[t]+0.4*B[t]$

Boolean Network

Exemplary Model
(Boolean functions):

$A[t+1]=A[t]$
$B[t+1]=\neg A[t]$
$C[t+1]=A[t]\vee B[t]$

Bayesian Network

Exemplary Model
(conditional probabilities):

$P(A=0)=0.4$
$P(B=0 \mid A=0)=0.3$
$P(B=0 \mid A=1)=0.9$
$P(C=0 \mid A=0, B=0)=0.8$
$P(C=0 \mid A=0, B=1)=0.3$
$P(C=0 \mid A=1, B=0)=0.4$
$P(C=0 \mid A=1, B=1)=0.1$

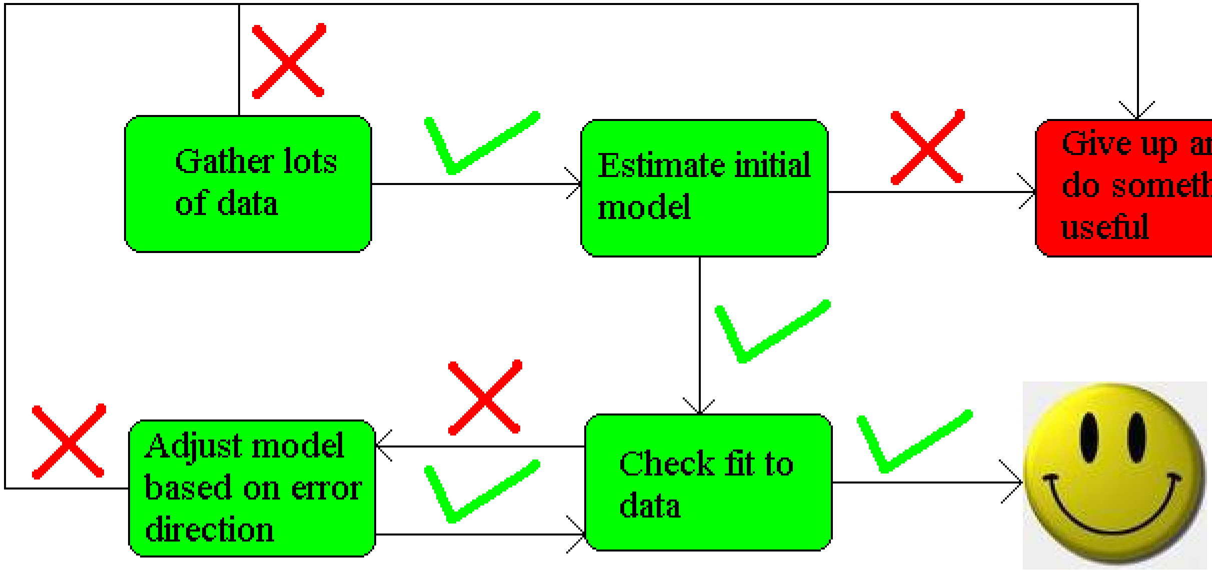
Information Theory Model

Exemplary Model
(correlation coefficients):

$A\sim B=-0.6$
$A\sim C=0.6$
$B\sim C=-1.0$

Inference techniques

- We've already seen a simple example of how one might do this
- Two parts: inferring the network structure, and inferring the parameters or weights (which depend on the type of model used)
- Large number of techniques, but essential idea of all of them is to do statistical analysis of large quantities of experimental data
- Often involve iteration: infer a particular network, see how well it can reproduce your data, then attempt to adjust structure/parameters to improve this; stir and repeat until desired consistency is reached



Gather lots of data

Estimate initial model

Give up and do something useful

Check fit to data

Adjust model based on error direction

Qualitative Modelling

- We can define qualitative relations between variables and attempt to learn a model based on these
- The variables themselves become qualitative: e.g., instead of tracking the actual expression level of a gene, we just represent it by a certain number of discrete states, say ON and OFF
- Relational learning techniques like Inductive Logic Programming can be used to infer such models from data

Example

- Suppose we have a gene whose expression level Y is regulated by an activator (level X1) and a repressor (level X2)
- Then a qualitative model for it might be:

```
DERIV(Y, DY)            // DY is the derivative
MPLUS(X1, ProdY)        // Production is incr. fn.
MMINUS(X2, ProdY)       // Decreasing fn. of X2
MPLUS(Y, DegY)          // Degradation rate
ADD(DY, DegY, IncrY)       // Net change is sum
```

Advantages and Challenges

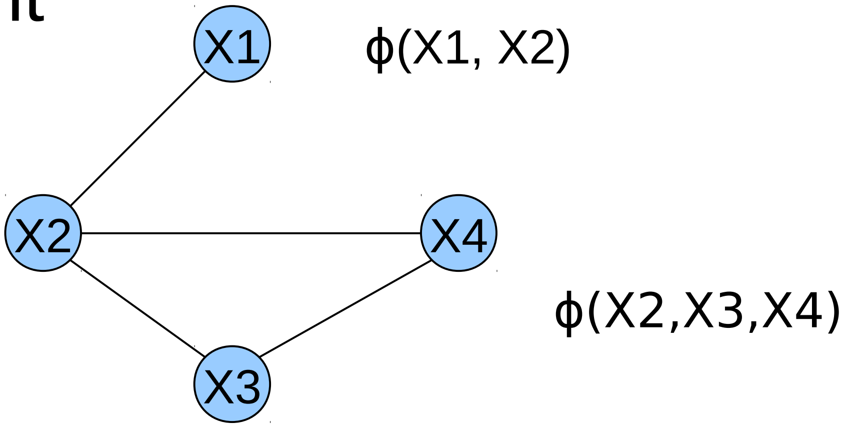
- Qualitative models are one way of dealing with highly noisy expression data sets, by abstracting away the precise measurements
- Have to come up with an appropriate discretisation of variables
- This approach has worked well for small-scale models, but will it scale to thousands of genes? Do we have enough data?

Probabilistic Models

- Another approach is to attempt to model joint probability distributions over gene expression levels
- Since the full joint distribution over thousands of genes will be not be learnable from realistically-sized datasets, we need to partition it in some way
- One way of doing this is to use a Markov Random Field (MRF) model

MRFs

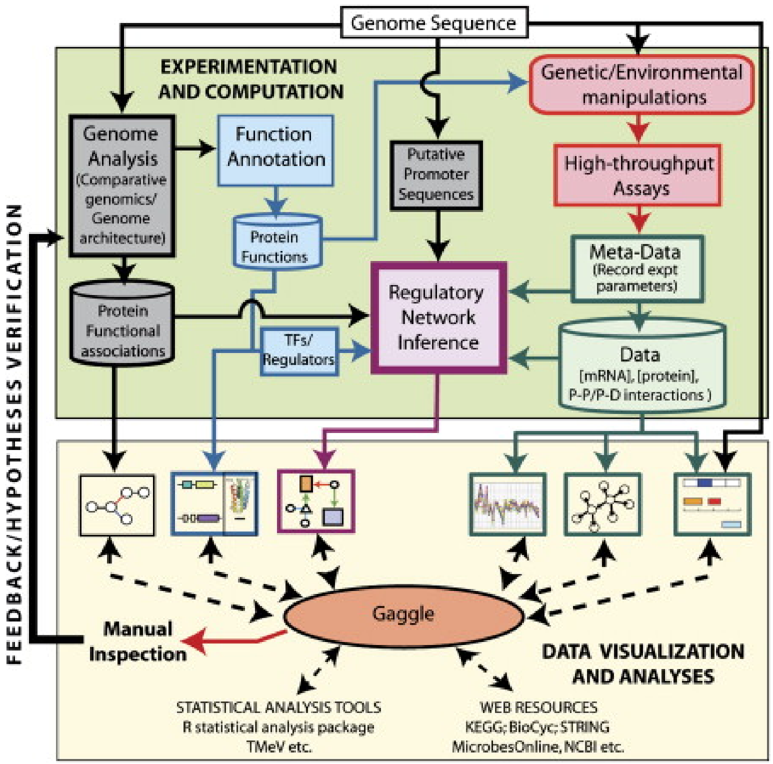
- We define a graph of linkages/correlations between different genes, based on domain knowledge
- The graph is partitioned into “components”, and a distribution function is learnt independently for each component

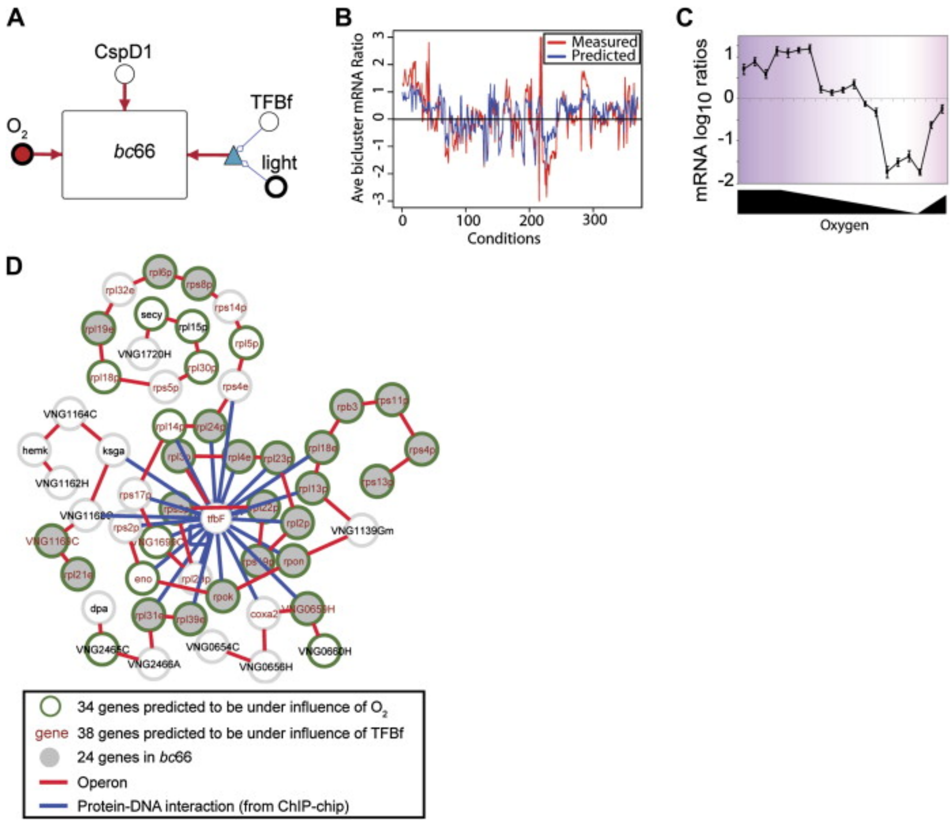


Example: gene regulatory network for *H.salinarum*

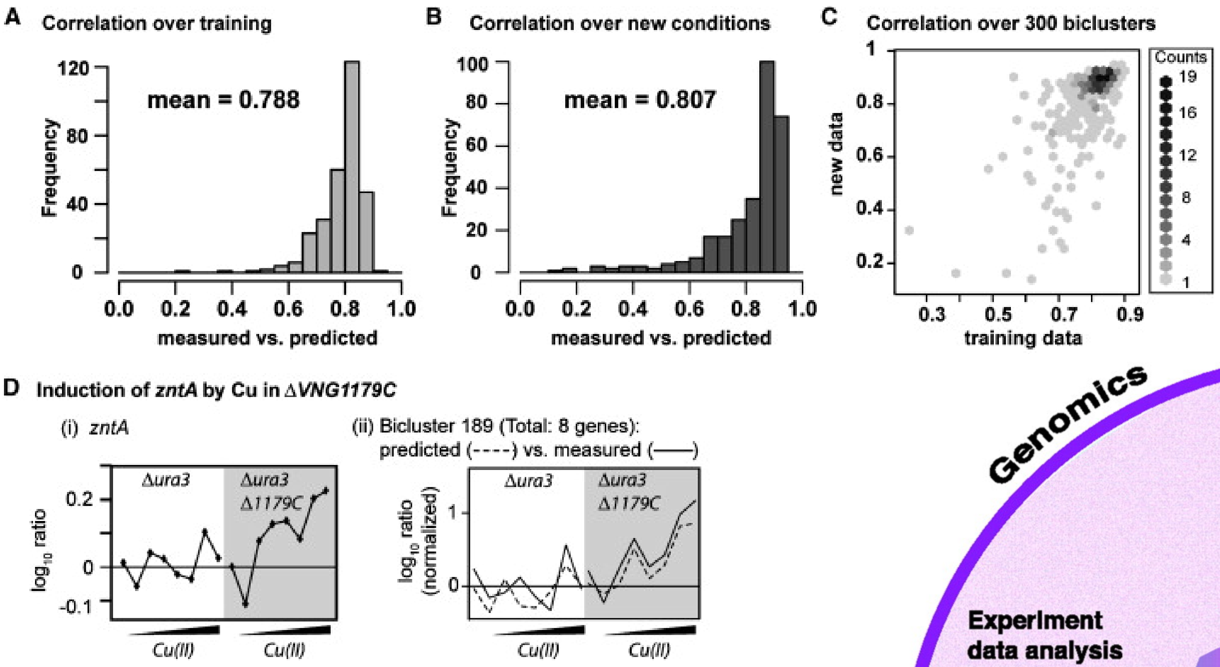
(Bonneau *et al*. 2007)

For a novel, largely uncharacterised organism, using a large number of microarray experiments combined with homology and other information, a remarkably successful attempt at creating a network that explains the data


Genome Sequence
EXPERIMENTATION AND COMPUTATION
Genome Analysis (Comparative genomics/ Genome architecture)
Function Annotation
Protein Functions
Putative Promoter Sequences
Genetic/Environmental manipulations
High-throughput Assays
Meta-Data (Record expt parameters)
Protein Functional associations
TFs/ Regulators
Regulatory Network Inference
Data [mRNA], [protein], P-P/P-D interactions
FEEDBACK/HYPOTHESES VERIFICATION
Manual Inspection
Gaggle
DATA VISUALIZATION AND ANALYSES
STATISTICAL ANALYSIS TOOLS
R statistical analysis package
TMeV etc.
WEB RESOURCES
KEGG; BioCyc; STRING
MicrobesOnline, NCBI etc.

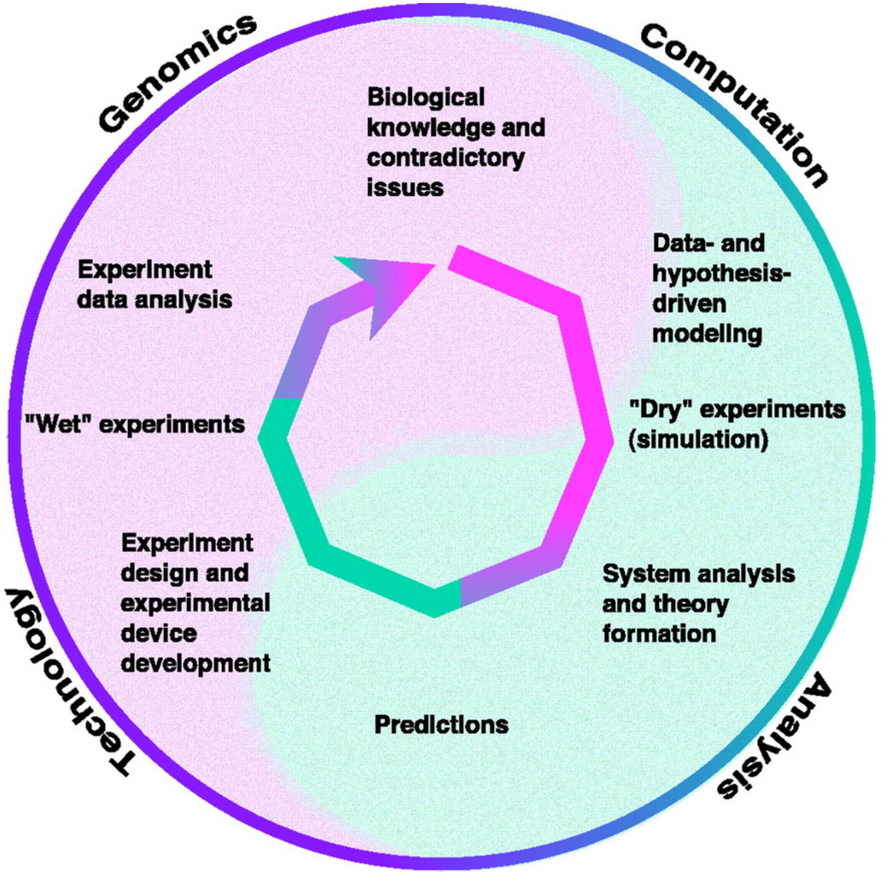


Predictive Systems Biology



(Bonneau *et al*. 2007)

An important reality check for any model is: can it make accurate predictions of how the system will behave in novel circumstances? This can be regarded as a major goal of “systems biology”, and network-based models have a key role



Genomics
Computation
Analysis
Technology
Biological knowledge and contradictory issues
Data- and hypothesis-driven modeling
"Dry" experiments (simulation)
System analysis and theory formation
Predictions
Experiment design and experimental device development
"Wet" experiments
Experiment data analysis

References

- Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biology* 2006, **7:**R36.
- Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang DE, Diruggiero J, Johnson CH, Hood L, Baliga NS. A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. *Cell* 2007, **131:**1354-1365.
- H. Jeong, S. P. Mason, A.-L. Barabási & Z. N. Oltvai. Lethality and centrality in protein networks. *Nature* 2001, **411**:41-42.
- Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren and Reinhard Guthke. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* 2009, **96:**86-103.
- Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth & Marc Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004, **430:**88-93.
- Hiroaki Kitano. Systems Biology: A Brief Overview. *Science* 2002, **295:**1662-1664.
- Sydney Brenner. Sequences and consequences. *Phil. Trans. R. Soc. B* 2010, **365:**207-212.
- Denis Noble. Biophysics and systems biology. *Phil. Trans. R. Soc. A* 2010, **368:**1125-1139.

Acknowledgements

- Dan Fenn
- Ben Fulcher
- Anna Lewis
- Max Little
- Aziz Mithani

International Postdoc Fellowships for Outstanding Researchers

India - Imperial College Biomathematics Bridge

Physics, CS, EE and Mathematics

Please google:

Biomathematics Bridge

To apply: tinyurl.com/k58enec

Excellence in Research

Indian Institutes | Imperial College