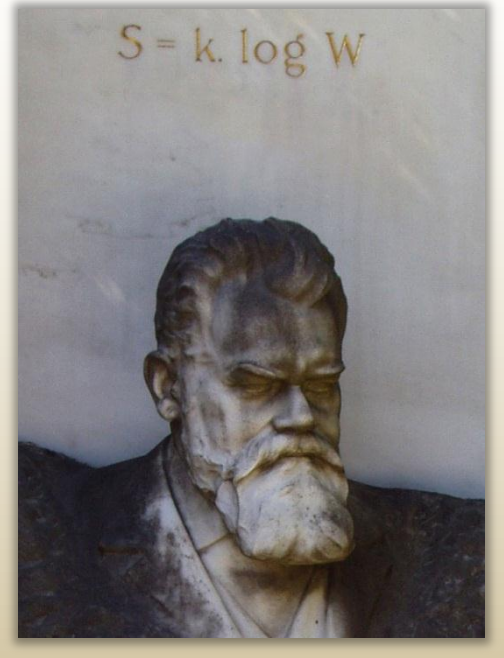
# Network Entropy and the Cancer Cell

Local and global cancer hallmarks
and applications

# Contents

1. Cancer research in the 21st century
2. Entropy in gene expression
3. Entropy in copy number
4. Conclusions



## Three reasons why we haven't cured cancer

1. Cancer consists of an unknown number of sub-types whose origins and conditions are not well known: which genes and cellular functions are important?

Daily Mail

**Landmark British study that could revolutionise breast cancer treatment: It turns out it's actually TEN different diseases**

By FIONA MACRAE

**PUBLISHED:** 18:00, 18 April 2012 | **UPDATED:** 17:48, 19 April 2012

Comments (79) | Share | +1 3 | Tweet 0 | Like 160

Breast cancer is effectively ten different diseases, according to breakthrough research that could revolutionise treatment.

The biggest study of its kind in the world has classified the country's most common cancer into ten separate types.

The finding brings doctors closer to the holy grail of tailoring treatments to individual women. The rewriting of the rule book on breast cancer could also lead to new drugs and better diagnostic tests.

BBC

**Breast cancer rules rewritten in 'landmark' study**

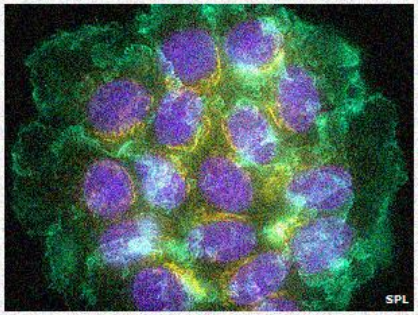
By James Gallagher
Health and science reporter, BBC News

**What we currently call breast cancer should be thought of as 10 completely separate diseases, according to an international study which has been described as a "landmark".**

The categories could improve treatment by tailoring drugs for a patient's exact type of breast cancer and help predict survival more accurately.

**The study in Nature** analysed breast cancers from 2,000 women.

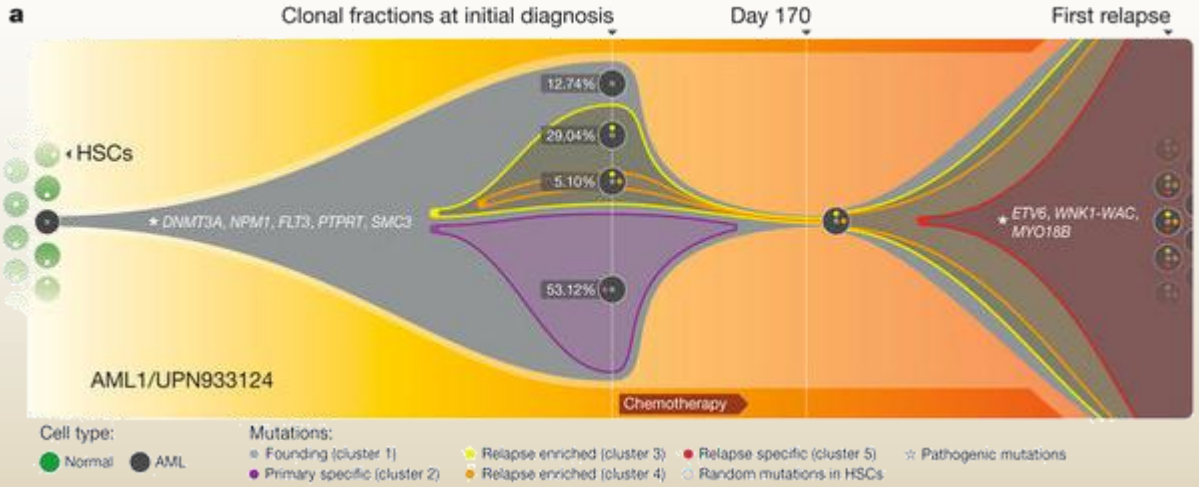
It will take at least three years for the findings to be used in hospitals.




Breast cancer cells should be classified into one of 10 different diseases, say researchers.

# Three reasons why we haven't cured cancer

2. A given tumor tends to consist of many different sub-populations. Thus any treatment tends to be like cutting off some of the heads of a hydra.

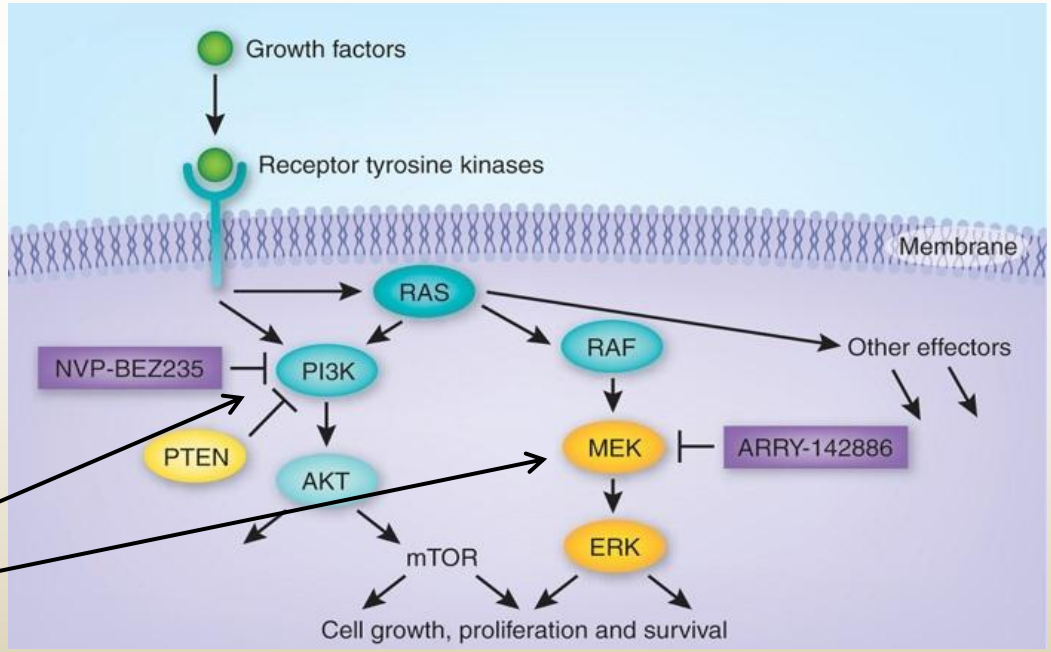



A relatively minor subpopulation of tumor cells survives chemotherapy and arises to become the dominant sub-type at relapse

*Ding et al, Nature 2012*

# Three reasons why we haven't cured cancer

3. Most often cancer is driven by alterations in a complex gene network. The biological picture of the network state is still largely incomplete.




In ordinary tumors with RAF activated you only need to inhibit MEK to put the cancer into remission.

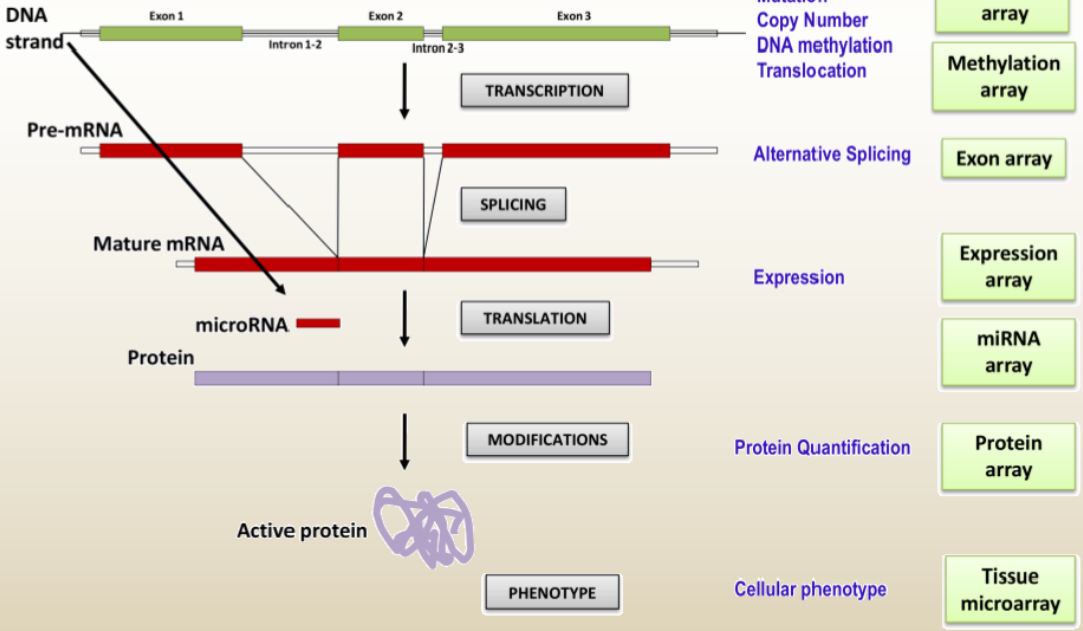
But when RAS is mutant, you have to inhibit PI3K and MEK!

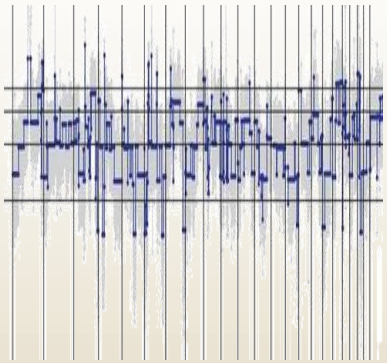
In this case knowing the transduction pathway involved gives biological understanding to why treatments work. Unfortunately, these are not known for most pathways!

*Engelmann et al, Nature Med. 2008*

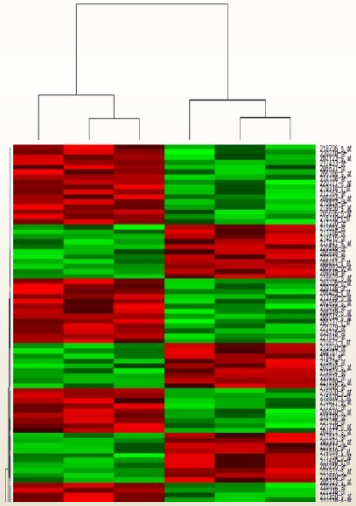
# Modern tools in cancer genomics

- We now have the technology to study cancer at the level of entire genomes:



Copy number

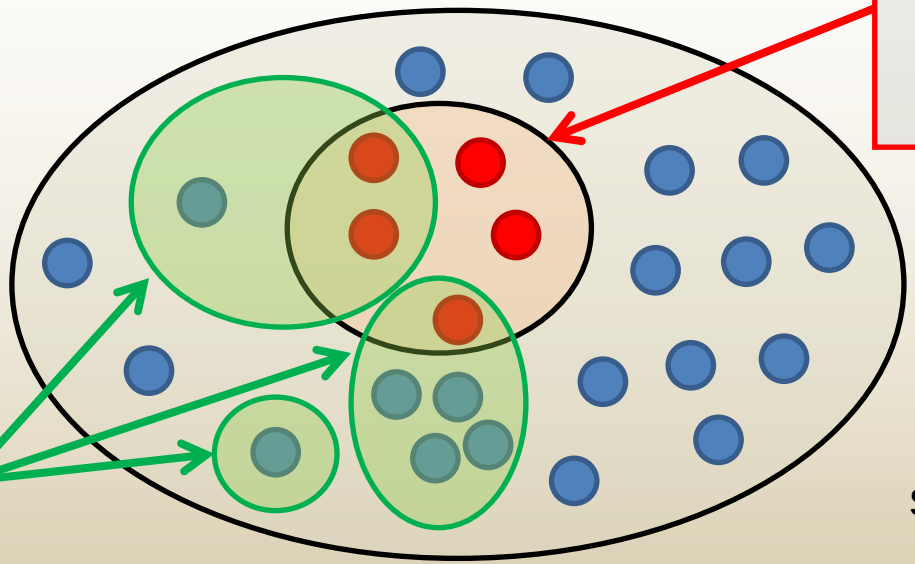


Gene expression

# Finding cancer genes: the traditional approach

- Given genomic disease data it is natural to ask which parts of the genome are different in the disease and how that has affected the biological function of the system.

Check to see which known biological functions of the cell contain significantly many of these genes. Control for false discovery!



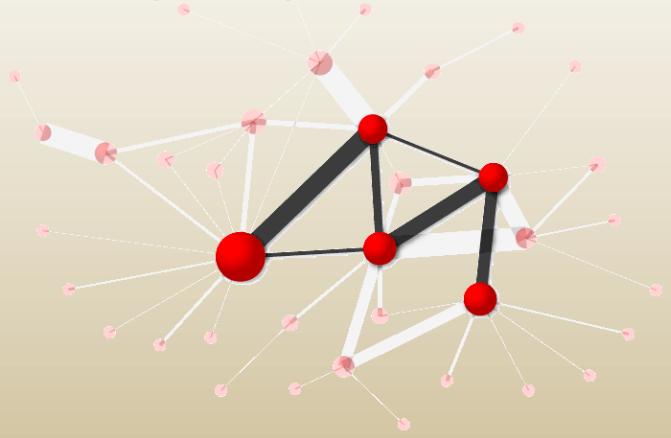
Rank genes according to size of change and select those such that only 5% are expected to be false positives.

The network structure is ignored in this approach

GSEA
Gene Set Enrichment Analysis

# Network approaches in cancer

- It is now known that most cancers are not caused by malfunctions in a single protein.
- In fact most tumors are characterized by hundreds of alterations (copy number, DNA methylation, …), often of different unknown levels of importance (drivers vs passengers).
- To address this challenge, biological data are often combined with interaction network models, giving rise to so-called integrated approaches.



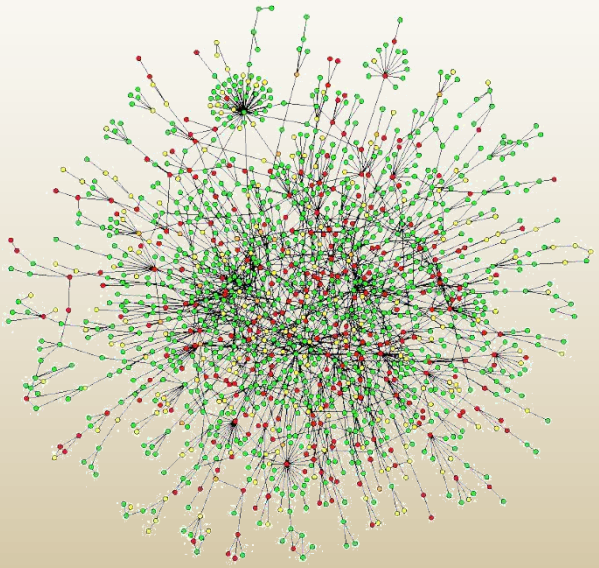
# The protein interaction network (*PIN*)

- The approximately 20,000 genes in the human genome are synthesized into proteins. These proteins interact, although we believe ourselves only to know a small proportion of all interactions; perhaps only 10%.
- The current interaction network models are obtained from agglomerating results of vast amounts of experiments.
    - *http://www.pathwaycommons.org/*

Human Protein Reference Database

Reactome

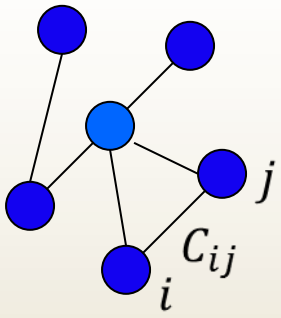
BioGRID



It is believed that around 50% of the interactions in the yeast protein interaction network are known.

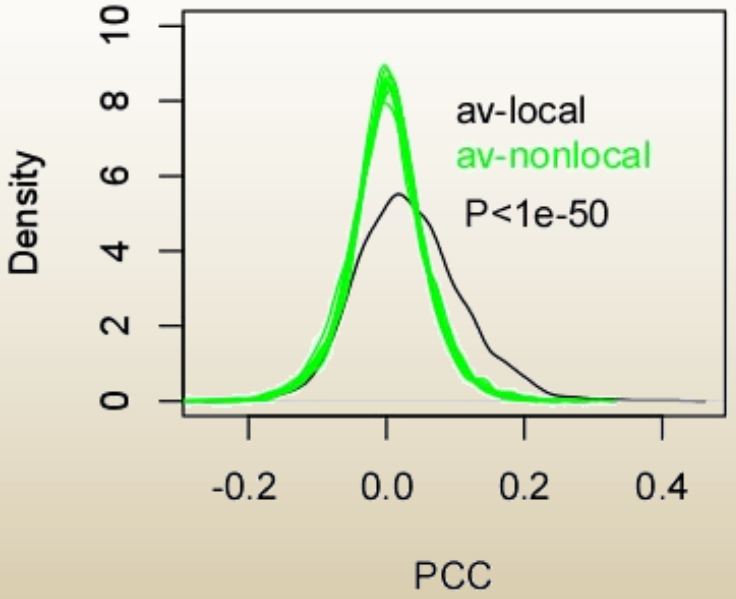
# Integrated expression-PIN networks

- Proteins that interact are in general more correlated at the expression (mRNA) level:



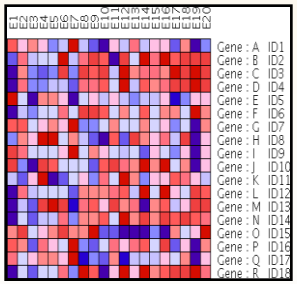

$$PCC_i = \frac{\sum_{j \in N(i)} C_{ij}}{k_i}$$

Example: normal bladder tissue (around 40 samples)



# Integrated expression-PIN networks in cancer

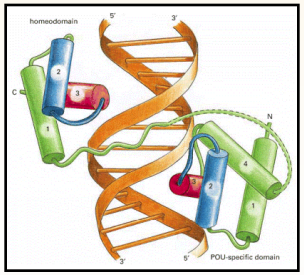
*Normal gene expression*



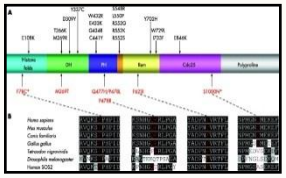
PINs (knowledge & predicted)

*Experiment*

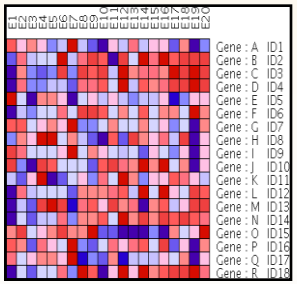

*Domain Analysis*

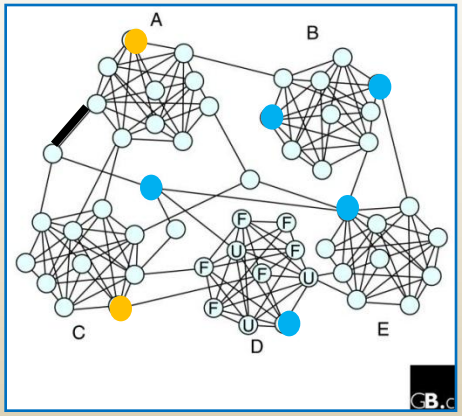



*Homology (common ancestry)*



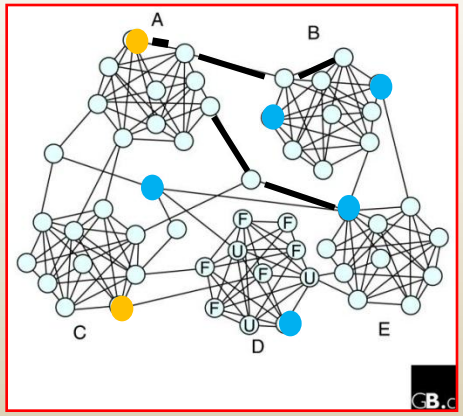
*Cancer gene expression*






***Normal network***

Which local and global network properties best define the cancer phenotype?




***Cancer network***

# Advantages of network approaches

- **Improved interpretation:** networks provide a biological context to interpret data like gene expression and copy number. For example, the network approach has uncovered subnetworks whose individual properties help to make predictions about cancer.

*Taylor IW Nat Biotech 2009*

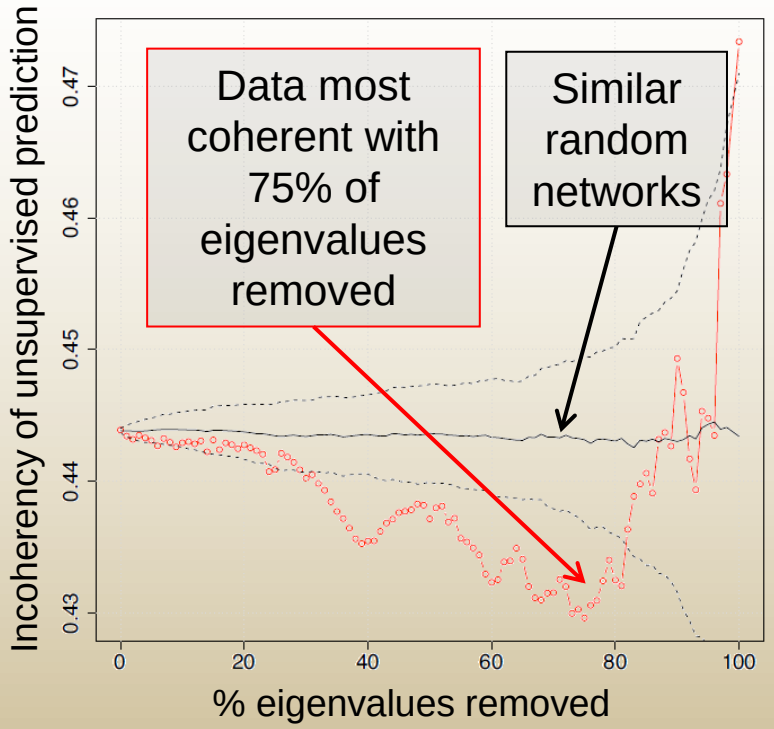
*Teschendorff BMC Syst Bio 2010*

- **Noise reduction**: using the eigendecomposition of the graph Laplacian may help remove noise in the data.

*Rapaport BMC Bioinformatics 2007*

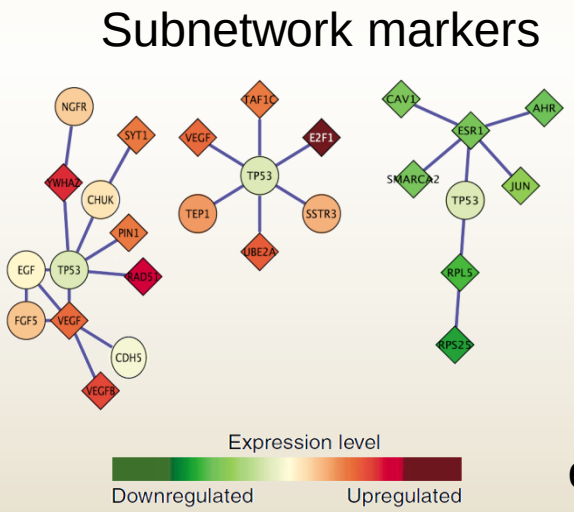
- **Improved prognostic models**: network allows more robust identification of relevant prognostic markers.

*Chuang Mol Syst Bio 2007*



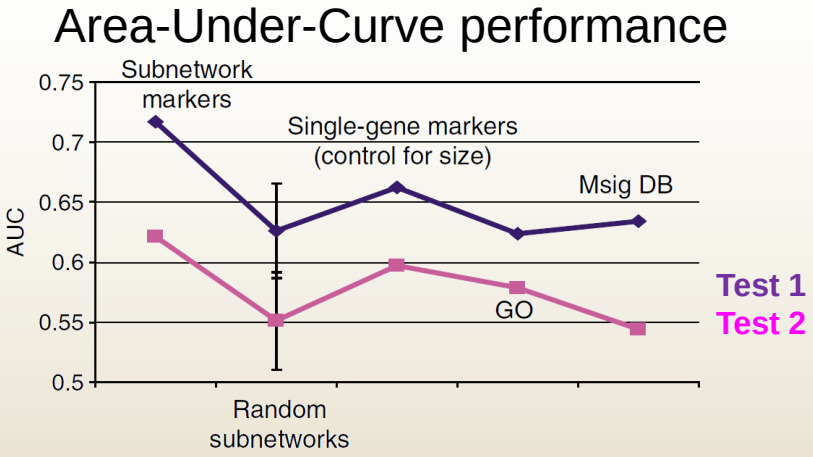
# A network approach to improving classification

- Chuang et al used the mean expression levels of subnetworks of the PIN as features in a classifier to predict whether breast cancers would spread (metastasize).




Classifier

Build classifier (mean expression levels of subnetwork markers)

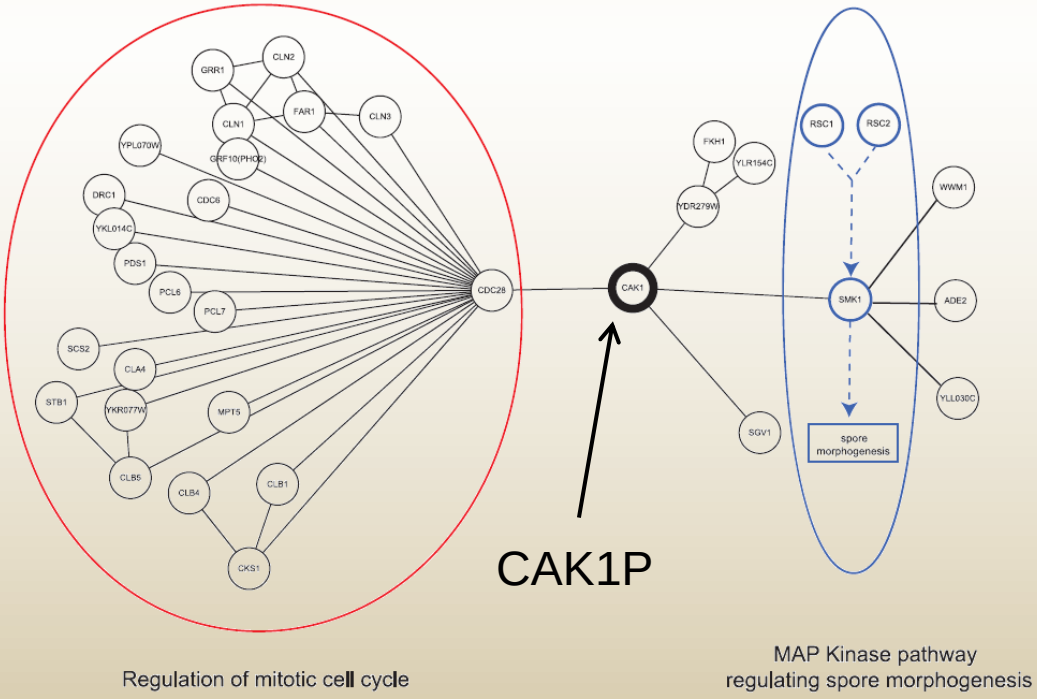



*Chuang et al, Mol Syst Bio, 2007*

# Properties of individual genes

- Unsurprisingly, simple topological properties of proteins in the PIN have been associated with function.
- In yeast, it was found that betweenness is a more effective indicator of “essentiality” (i.e. the organism either dies or ceases growth without it) than degree.

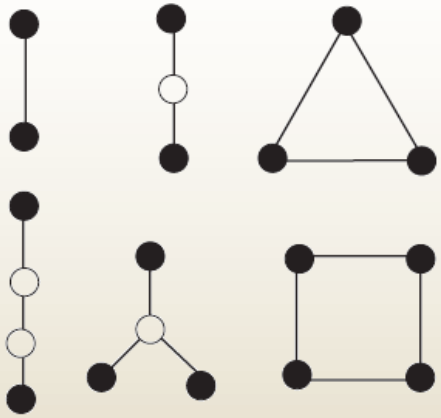
The protein CAK1P in yeast, an essential, but low-degree protein of high betweenness.



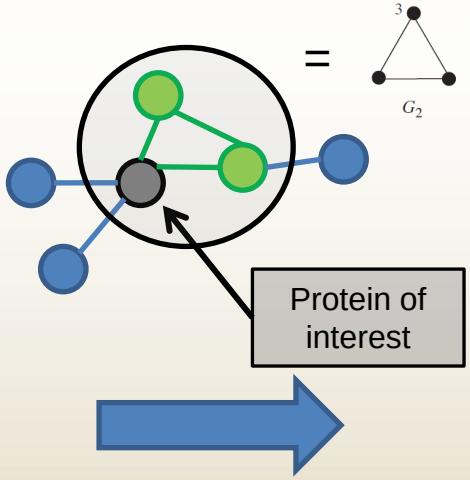

*Gerstein et al, PLoS Comp Bio 2007*

# Properties of individual genes

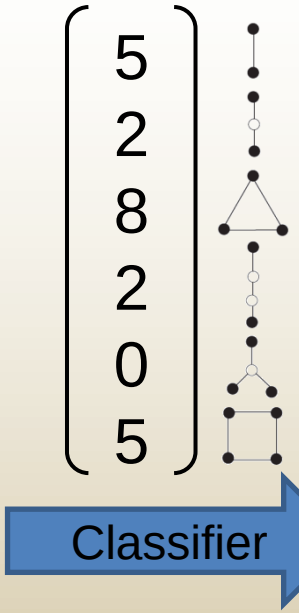
- By representing the structure of the network near a protein in terms of some of the different possible sub-networks containing it (viewed as sub-networks of the PIN), it is possible to predict the protein's function.



Pick a subset of the set of networks (graphs)




Compute the number of copies of each of these networks that include your protein of interest




Classifier

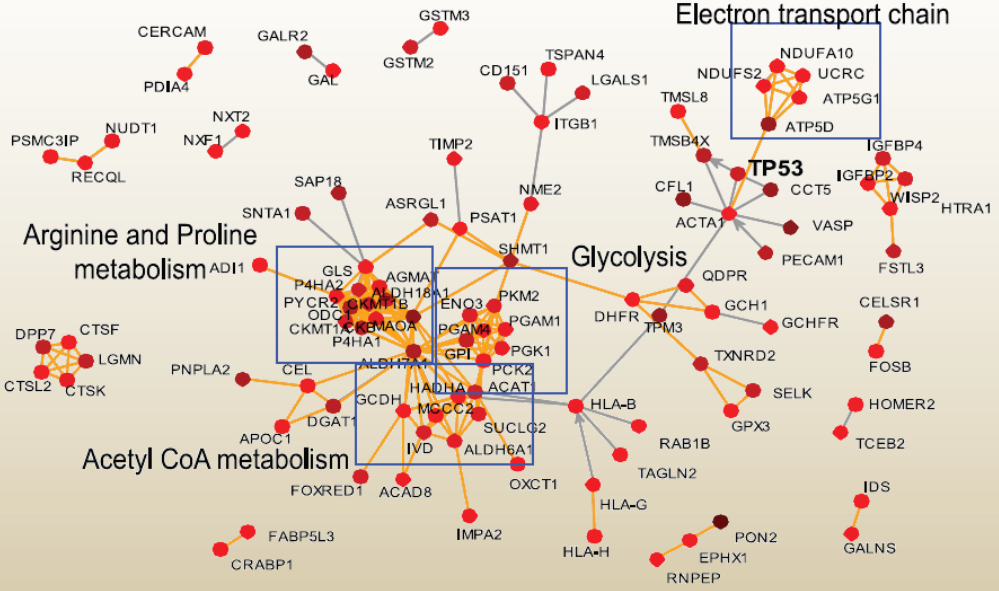

Predict what sort of thing the protein does: e.g. aid in killing the cell, fix the DNA, communicate between cells, etc.

*Kuchalev et al, J R Soc Interface 2010*

# Dynamics on the PIN

- Given weights on either the vertices (genes) or edges (interactions) in the PIN, there are several natural constructions giving rise to random walks.
- Komurov used the invariant measures (after adjustment) of walks coming directly from the expression-PIN to understand response to DNA damage.

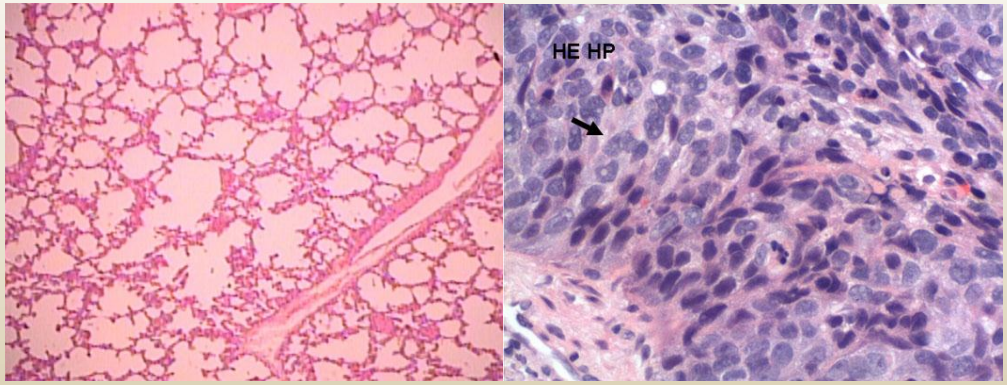
Invariant measure re-scaled to avoid great bias towards hubs




*Komurov et al, PLoS Comp Bio 2010*

# Motivation

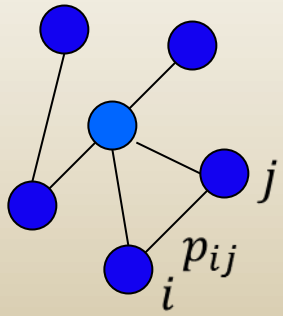
- Want to find a new framework to find genes responsible for “driving” cancers.
- At present rather little is known that distinguishes generic normal tissue cells and cancerous cells at the systems level, in particular of protein-interactions and expression (mRNA) levels.
- We investigate this by looking to see if changes in the information content across various notions of what might be called “molecular entropy” to distinguish normal and cancer tissues.




Normal lung tissue  Cancerous lung tissue

# A random walk on the PIN

- Given interacting proteins $i$ and $j$, their correlation $C_{ij}$ across samples of a given phenotype gives a weight on the edge between them.
- Scaling this to be non-negative, as $w_{ij} = (1+ C_{ij})/2$ (or perhaps $|C_{ij}|$), we obtain a random walk on the PIN by normalizing to obtain a probability $p_{ij}$ of walking from $i$ to $j$



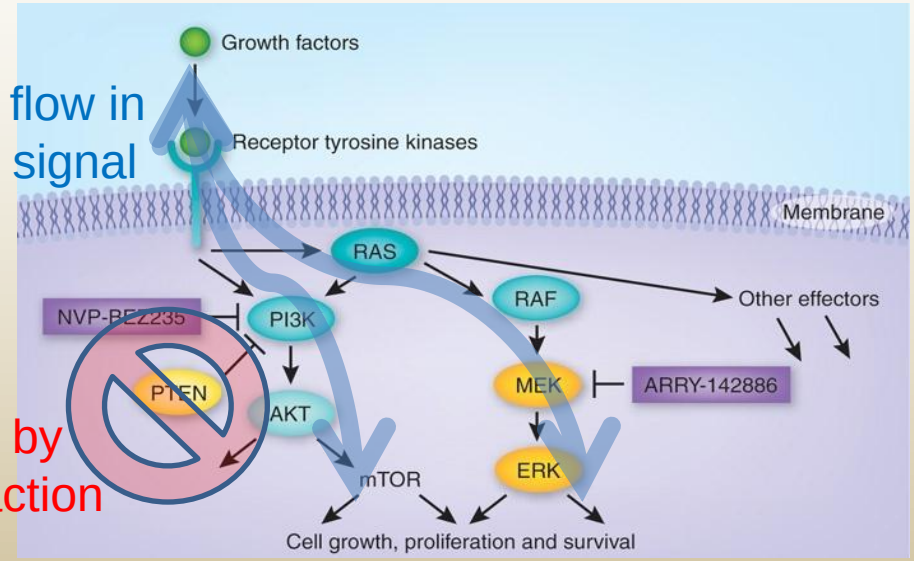

$$p_{ij} = \begin{cases} \dfrac{w_{ij}}{\sum_{j\sim i} w_{ij}} & \deg i \geq 1 \\ 0 & \deg i = 0 \end{cases}$$

# The choice of walk

- With the choice of edge weight $w_{ij} = (1+ C_{ij})/2$ we force the random walk to flow like information would flow through the system; in the direction of signal transduction paths and away from inactive or possibly inhibitory interactions.
- The choice of edge weight $w_{ij} = |C_{ij}|$ is also valid, but treats inhibition and activation equally.

Information flow in direction of signal

Flow blocked by inhibitory interaction



# Molecular entropy

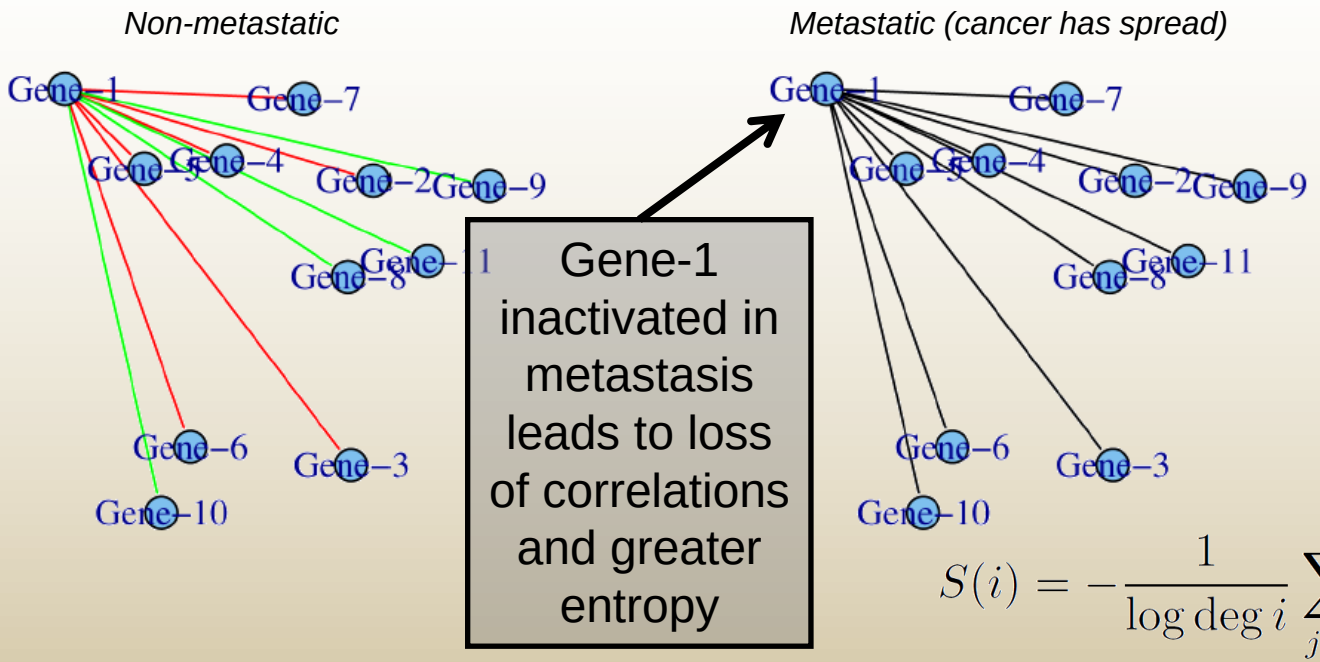
- Observe that the walk probabilities emanating from a vertex (protein) gives a probability distribution corresponding to that gene.
- We refer to the Shannon entropy of this, i.e.

$$S(i) = -\frac{1}{\log \deg i} \sum_{j \sim i} p_{ij} \log p_{ij}$$

  as the **molecular flux entropy** of the protein *i.*
- The Shannon entropy in a sense measures the **predictability** of the walk: on a finite set of values, a uniform distribution is least predictable, and a distribution always taking the same value the most.

# Information loss in the cancer cell

- Significant increases were observed in the transitions from normal to cancerous to metastatic cancers (those which have spread). This was observed to be the case across several cancer types.




$$S(i) = -\frac{1}{\log \deg i} \sum_{j \sim i} p_{ij} \log p_{ij}$$

*Teschendorff & Severini BMC Systems Biology 2010*

# Information loss in the cancer cell

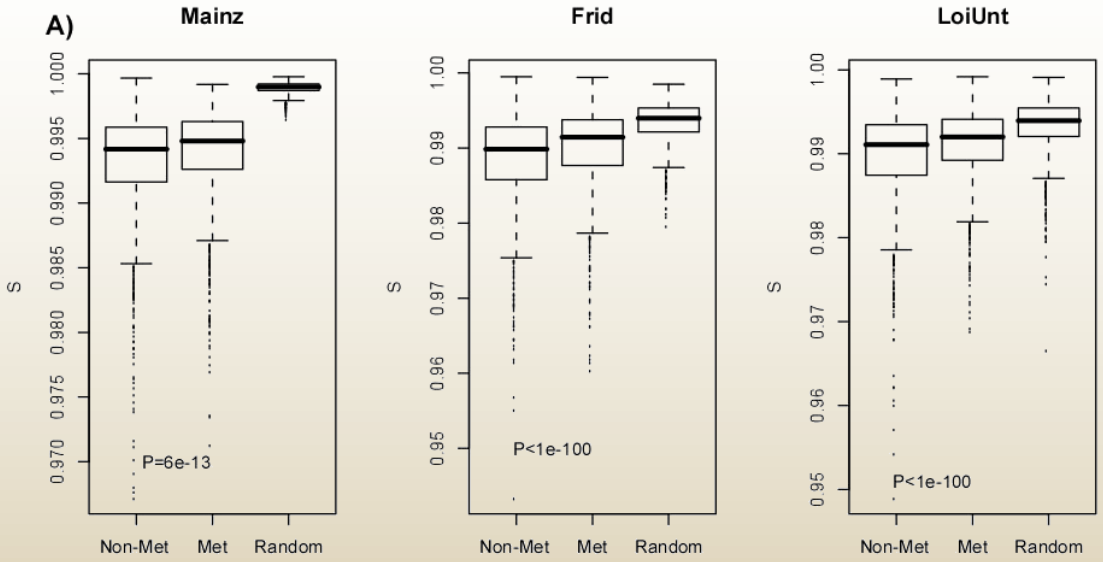
- Compared to the *average local correlation*, (i.e. the mean correlation of a node with its nearest neighbors) and the *mean local absolute correlation*, the molecular entropy was a better distinguisher of normal, cancer and metastatic cancers.

| Wilcox Test | Bladder (n=1952) | Pancreatic (n=1758) | Prostate (n=1758) |
|---|---|---|---|
| $dC$ | 0.96 | 0.39 | 0.59 |
| $d\|\overline{C}\|$ | 3e-10 | <1e-50 | <1e-50 |
| $dS$ | 2e-13 | <1e-50 | <1e-50 |

TABLE 1. One tailed paired Wilcoxon and Binomial test $p$-values comparing the statistical distribution of local measures of disruption in information flow in cancer versus normal PIN-mRNA networks across three different cancer normal tissue cohorts. $dC$ denotes the difference in mean local correlation, $d|C|$ denotes the difference in mean local absolute correlation and $dS$ denotes the difference of local Shannon entropy. The number of pairs (nodes) in the tests, $n$, corresponding to the number of nodes in the network with degree $\geq 10$ are given.

# Information loss in the cancer cell

- As a sanity check: the Shannon entropy was also higher in randomised null networks in which the expression values were randomly permuted among nodes in the PIN.




Entropy is higher in primary breast cancer tumors that have spread (metastasized)

*Teschendorff & Severini BMC Systems Biology 2010*

# Dynamical entropy over longer distances

- The previous notion of molecular entropy obtained weights from the correlations in two different phenotypes and computed the disorder of these (after slight adjustment) looking only at immediate neighbors.
- Observe that the $k^{\text{th}}$ power of the stochastic matrix $p = (p_{ij})$ gives the information flow over distances of length $k$.
- The total information flow of various distances between two genes can be obtained by taking a choice of linear combination of powers of $p$.

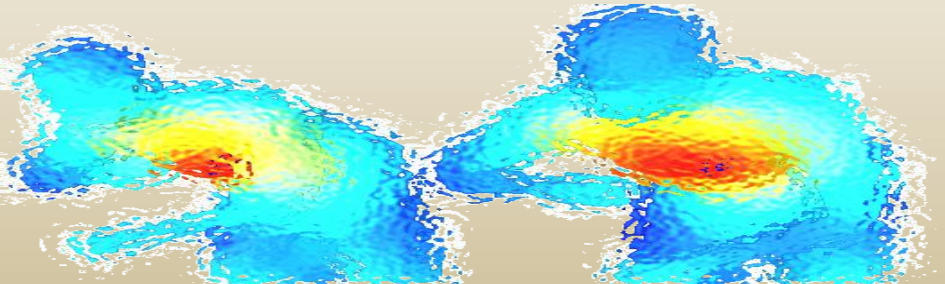
# A natural combination of powers

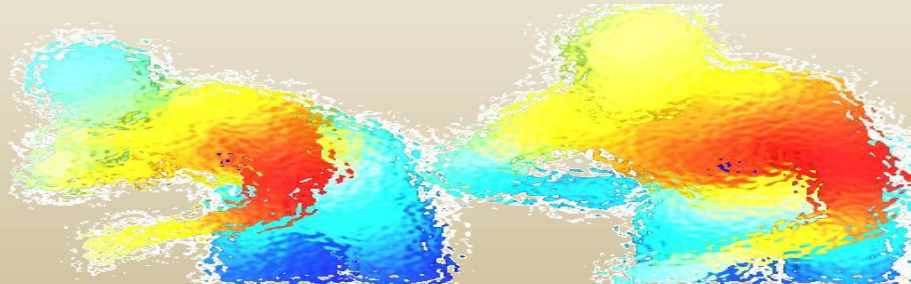
- Motivated by statistical physics, we introduce a temperature parameter $t$ and consider the family of matrices $K(t)$ obtained as

$$K_{ij}(t) = \frac{1}{e^t - 1} \sum_{\ell=1}^{\infty} \frac{t^\ell}{\ell!} \left(p^\ell\right)_{ij}$$

- This series converges and satisfies a modified heat diffusion equation

$$\partial_t K_{ij}(t) = -K_{ij}(t)\left(\delta_{ij} - p_{ij}\right) + \frac{1}{e^t - 1}\left(p_{ij} - K_{ij}(t)\right)$$





# Heating it up

- In the “hot” temperature limit, this indeed approximates a solution of the heat diffusion equation

$$\partial_t K_{ij}(t) = -K_{ij}(t)\left(\delta_{ij} - p_{ij}\right)$$

- For each $t$ we may write down a “global” entropy $S(t)$ of the information flux as

$$S(t) = -\frac{1}{\log Q}\sum_{ij} K_{ij}(t)\log K_{ij}(t)$$

where $Q$ is the number of non-zero entries of $K$.

- We refer to this as the global **flux entropy**.

*West et al Submitted 2012*

# Covariance entropy

- Before studying the flux entropy further, we introduce another notion recently studied.
- The **covariance entropy** quantifies the degree of similarity between samples as determined by their Pearson correlations, first considered in the context of gene expression by van Wieringen and van der Vaart.
- They argue that with the accumulation of copy number aberrations in cancer, there should be an increase in genomic entropy. For this to be cancerous, this should be reflected in expression.

*van Wieringen et al Bioinforrmatics 2011*

# When does entropy in copy number lead to entropy in expression?

**DNA copy number** $X$ — ***S(N)*** **<** ***S(C)***

**mRNA expression** $Y = g(X)$ — ***S(N)*** **<** ***S(C)***

***S(N)*** Entropy in normal phenotype

***S(C)*** Entropy in cancer phenotype

*Proposition:*

Let $X_1$ and $X_2$ be random variables with symmetric and zero-centred densities $f_1$ and $f_2$, respectively, and $g(.)$ a strict monotone function. Then

$$S(X_1) \leq S(X_2) \Longrightarrow S\big(g(X_1)\big) \leq S(g(X_2))$$

*Van Wieringen et al Bioinformatics 2010*

# Computing covariance entropy

- Modelling gene expression profiles across $n$ samples of $g$ genes as a multivariate normal $Y \sim N(\mu,\Sigma)$, the entropy $H(Y)$ is given by

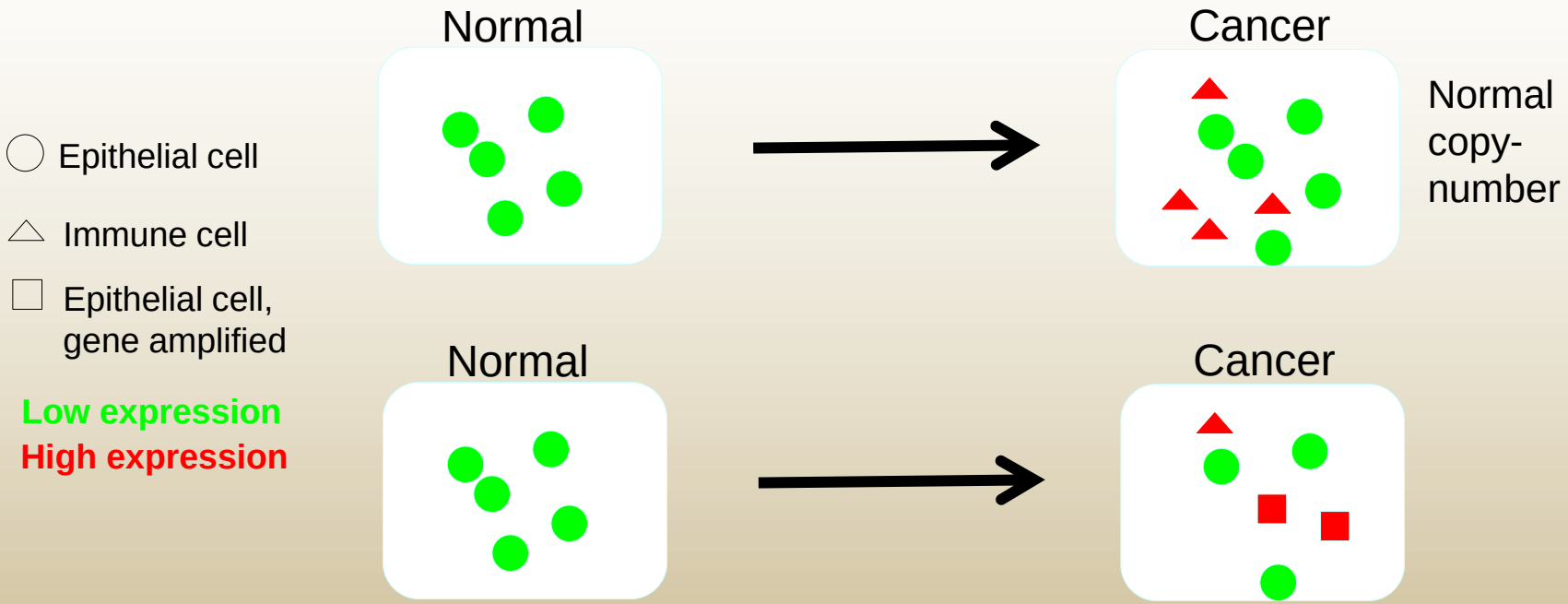
$$H(Y) = \frac{1}{2}\log(\det\Sigma) + \frac{g}{2}(1+\log 2\pi)$$

- Letting $\Sigma_i$ be the covariance matrix restricted to gene $i$ leads to the **local covariance entropy**:

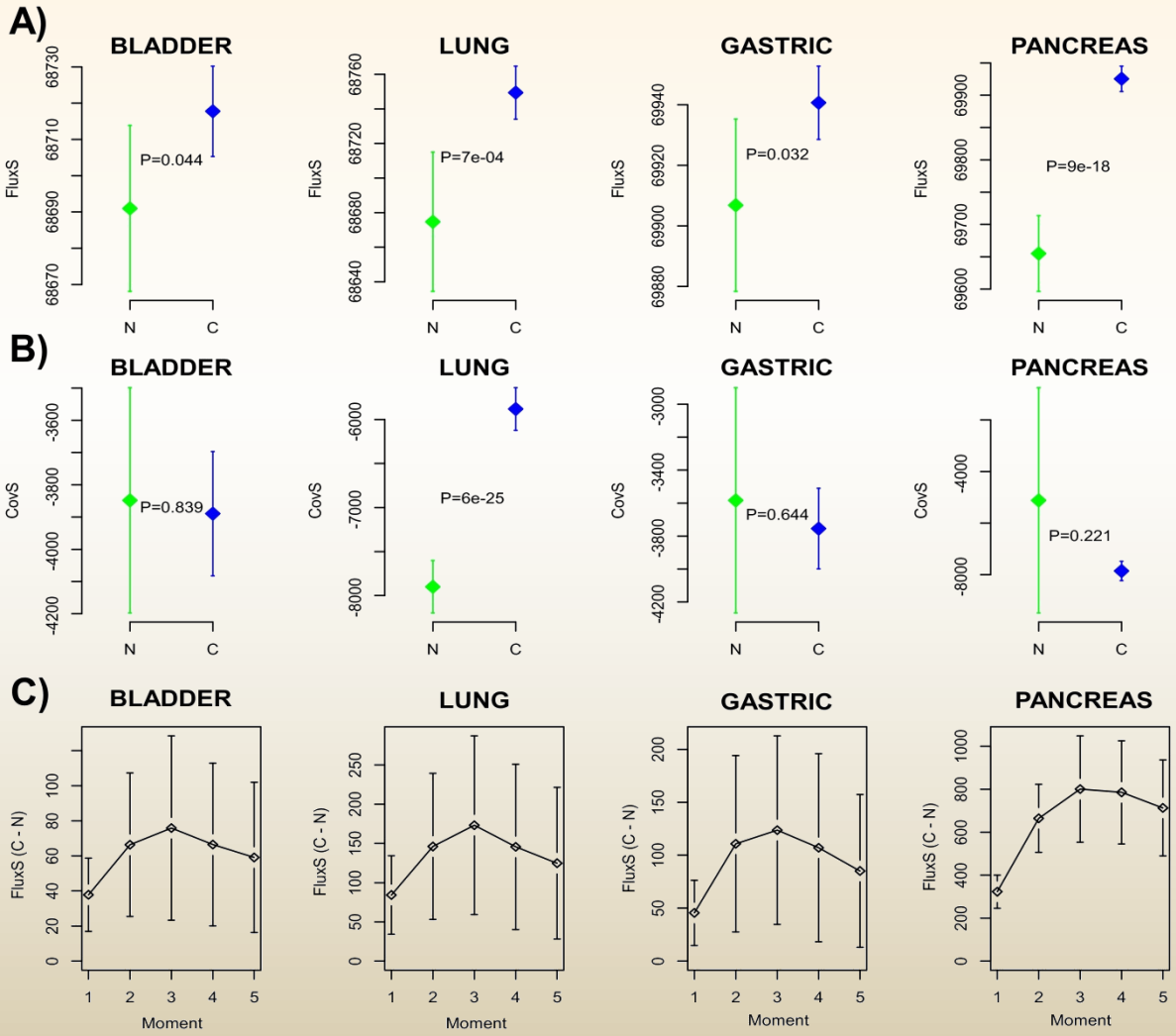
$$H_i = \frac{1}{2}\log\det\Sigma_i$$

# A caveat

- The monotone relationship between copy number $X$ and gene expression $Y = g(X)$ may be severely compromised if the measured expression levels are over a mixture of tumour cells and stromal (non-tumor cells).



# Global information entropy increases



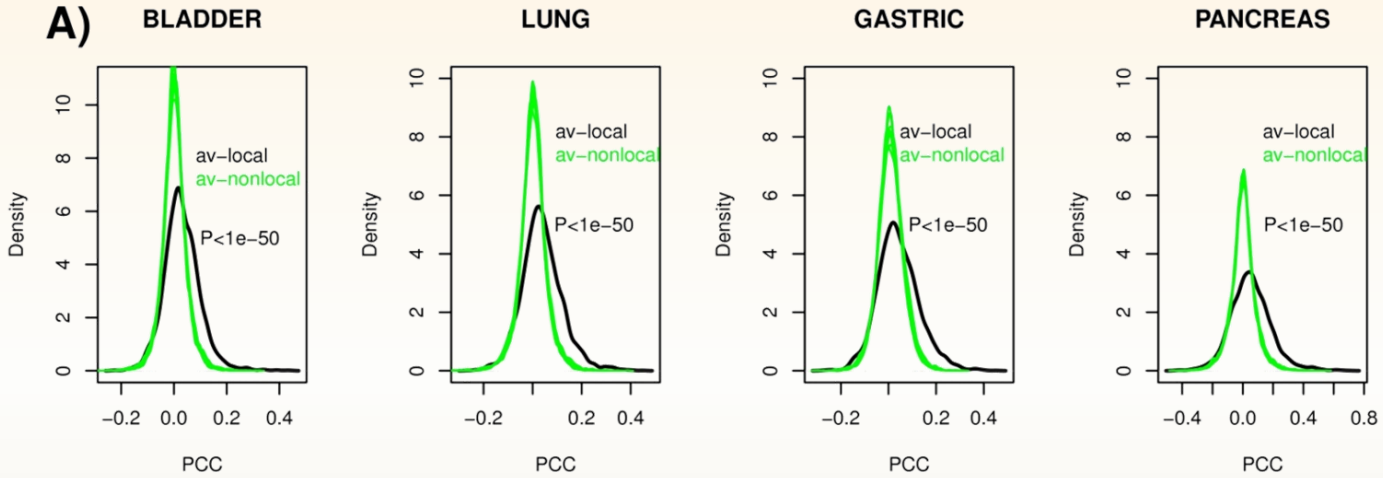

Flux entropy is increased in cancer relative to normal tissue.

Covariance entropy changes inconsistent.

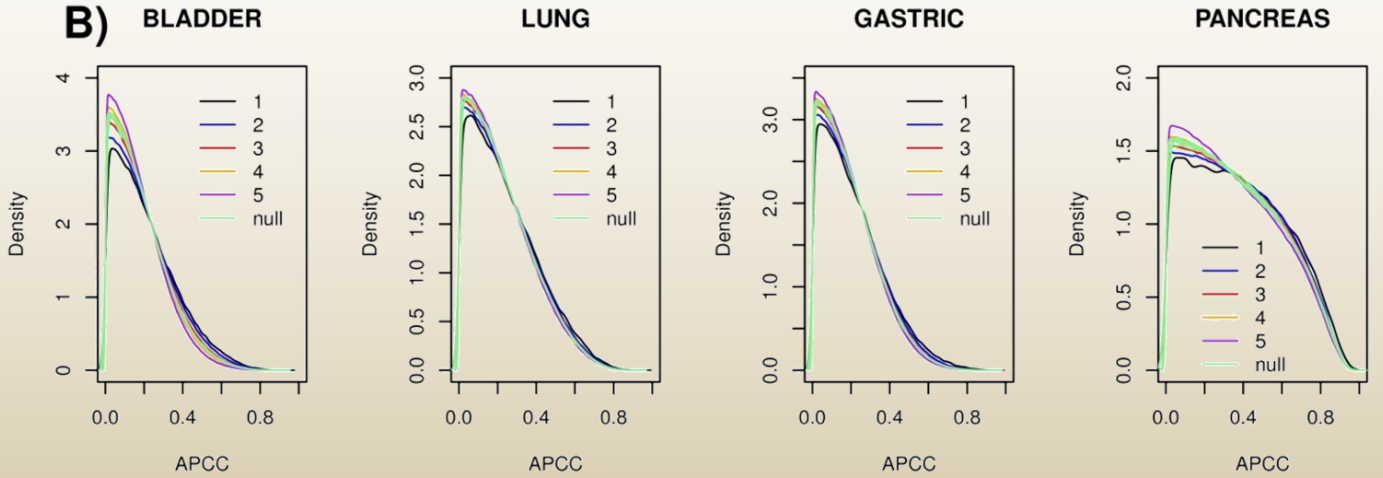
Maximum difference attained for paths in the PIN up to length 3.

*West et al Submitted 2012*

# Correlation path length scales



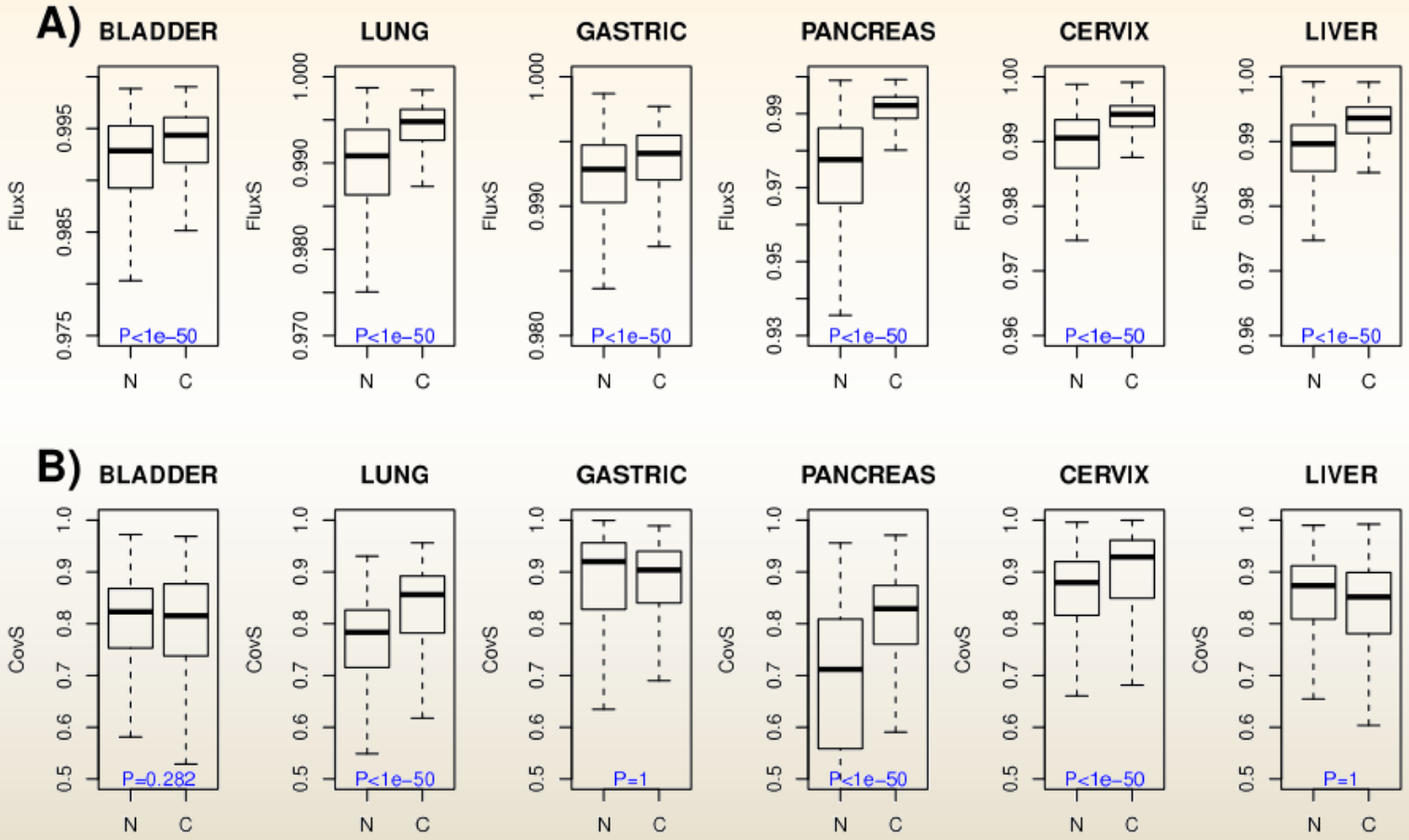

Local correlations stronger in every case.




Natural correlation length scale on the mRNA-PIN is 3.

*West et al Submitted 2012*

# Local flux and covariance entropies




Local flux entropies increased in cancer.

Local covariance entropies not always increased.

Consistent and significant increases in local flux entropy found from normal to cancer tissue.

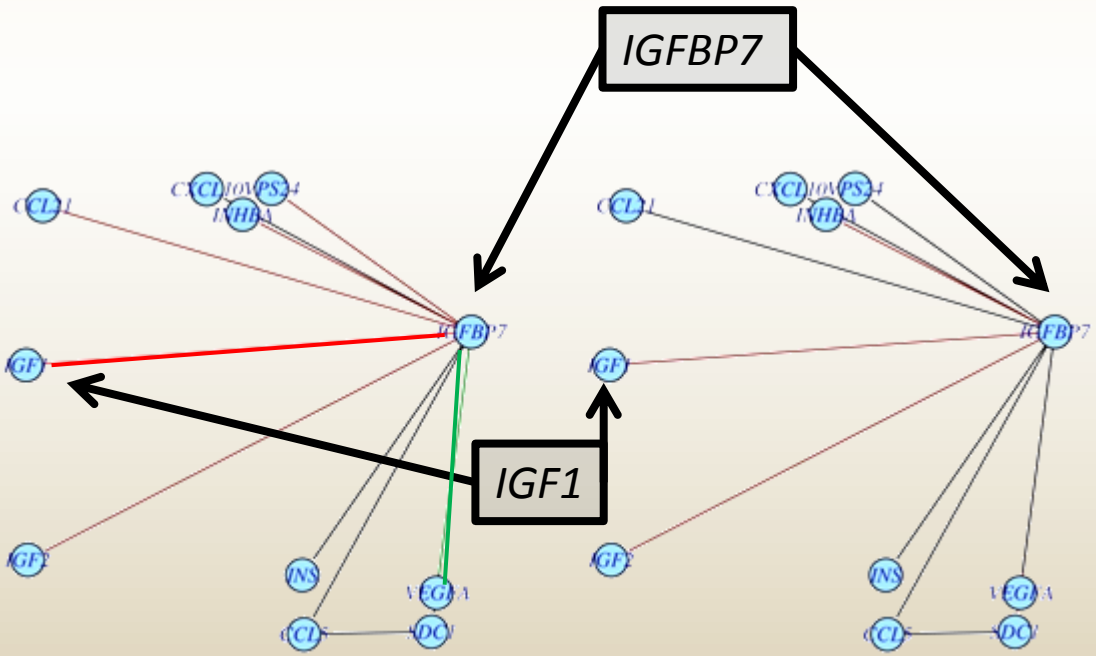
*West et al Submitted 2012*

# Differential expression and entropy

- **Hypothesis**: alterations in genes “driving” the cancer should lead to disruptions in local and global gene expression patterns causing changes in entropy.
- **Consequences**:
  - Genes whose inactivation confers selective advantage (tumor suppressors) should tend to show increases in local flux entropy.
  - Genes that become activated in cancer (oncogenes) can be expected to show reductions.

# Verifying changes in expression

- Genes significantly over-expressed in cancer ($P < 0.05$) show reductions in flux entropy when compared to genes which were under-expressed.




- Binding of *IGFBP7* to *IGF1* is reduced and there is reduced suppression of VEGFA.
- This leads to increased angiogenesis (growth of new blood vessels; a hallmark of cancer) in the metastatic phenotype.
- *IGFBP7* is putative tumor suppressor; lower expression in cancer. *Wajapeyee et al Cell 2008, Oh Y et al J Biol Chem 1996*

*Teschendorff & Severini BMC Systems Biology 2010*

# Interpreting flux entropy

- The increased flux entropy in cancer may endow cancer cells with the flexibility to adapt to the strong selective pressures of the tumor microenvironment.
- The **fluctuation theorem** of Demetrius et al asserts

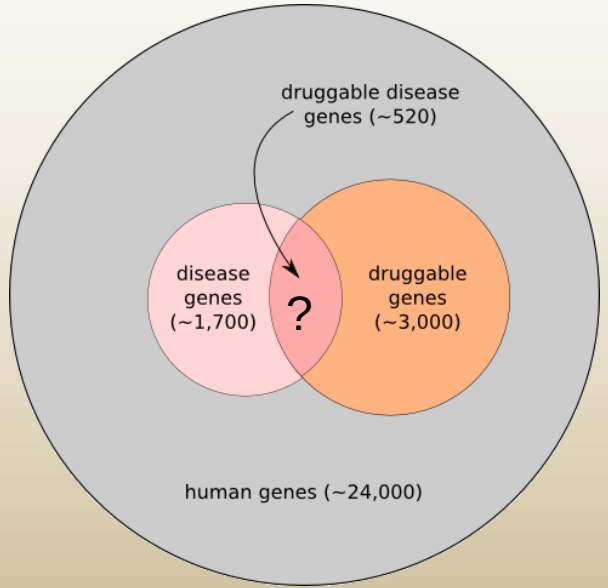
$$\Delta R \Delta S > 0$$

i.e. there is a correlation between changes in network entropy $S$ and robustness $R$. As such it is possible that cancer alterations leading to significant increases in flux entropy may contribute to the dynamical robustness of such cancer cells.

*Demetrius et al 2004 Physics A 346*

# Therapeutic applications

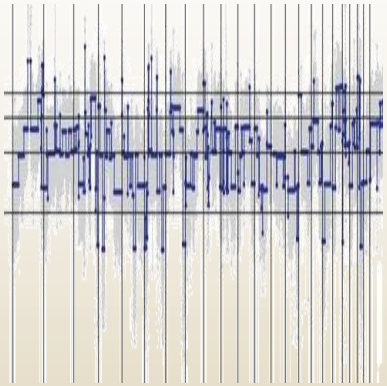
- A great problem is that important cancer-related genes are not directly druggable.
- In these cases it may be possible to use differential flux entropy to identify neighboring viable drug targets that also exhibit significant reductions in flux entropy.
- This computational strategy could therefore guide certain therapeutic strategies that aim to select drug targets within the same oncogenic pathway.




Often genes are not druggable because there are too many similar genes and drugs tend to affect all of them.

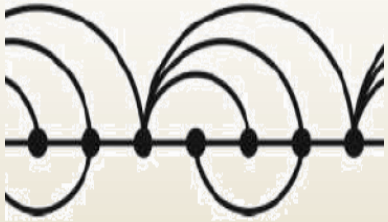
## Motivation

- In this section we use a “combinatorial smoothing” obtained by associating a network known as a visibility graph to copy number data viewed as a time series



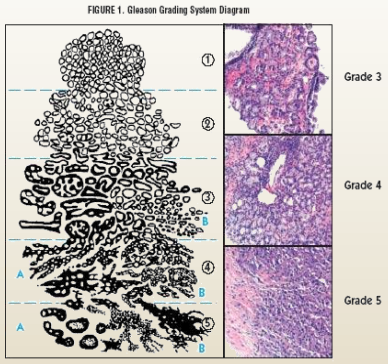
Copy number

Visibility algorithm



Visibility graph

Approx. entropy




Information about phenotype

# Approximate entropy of networks

- Much practical work has considered an entropic origin for various properties of complex networks, including biodiversity in ecological networks, the emergence of degree-degree correlations and communities in social and biological networks.
- These approaches focus mainly on the ubiquitous Shannon entropy, and to help complete the picture we introduce a notion based around the approximate entropy.

# Approximate entropy

- The approximate entropy considered by Pincus exists as a finite-sized statistic of the Eckmann-Ruelle entropy proposed to measure the complexity of a system with time evolution.
- The approximate entropy $\mathrm{ApEn}(m,r,N)$ of a time series $u$ of length $N$ is constructed from the image $X_m$ of the Takens map $x(t) = u((t),\ldots,u(t+m-1))$.

With $$\Phi_m(r) = -\frac{1}{|X_m|}\sum_{x\in X_m}\log\left(\frac{|\{y\in X_m : \|x-y\|_\infty \le r\}|}{|X_m|}\right)$$

define $$\mathrm{ApEn}\,(m,r,N) = \Phi_{m+1}(r) - \Phi_m(r)$$

# The Rukhin estimate

- Intuition is gained due to the combinatorial interpretation of approximate entropy, due to Rukhin: given a sequence $u$ of length $N$ on $S$ symbols $\{0, 1, \ldots, S-1\}$, let $\nu(I)$ be the frequency with which block $I$ occurs. Denoting

$$\widetilde{\Phi}_m = -\sum_{I \in \{0,1,\ldots,S-1\}^m} \nu(I) \log \nu(I)$$

  the estimate is $\widetilde{\mathrm{ApEn}}(m) := \widetilde{\Phi}_{m+1} - \widetilde{\Phi}_m$

- Fact: this is a.s. a good approximation of the ordinary approximate entropy of $u$.

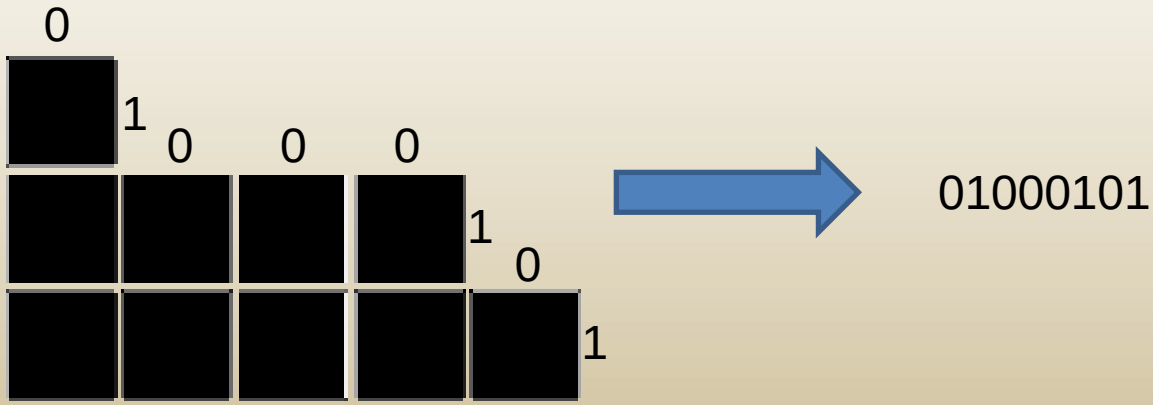
*Rukhin, J Appl Probab 2000*

# Upshot of the Rukhin estimate

- As such, on a finite set of symbols, the approximate entropy adds something new, measuring something distinct from the ordinary statistical moments (mean, variance, … ) and the Shannon entropy of the sequence.
- It can be (almost surely!) thought of measuring how much data a choice of universal data compressor would use to store the object.

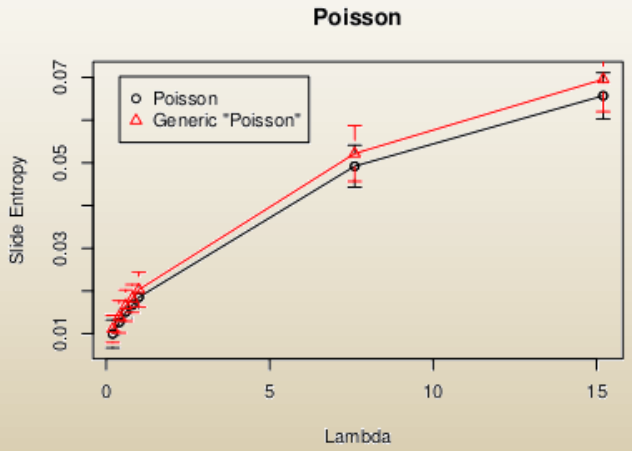
# Approximate entropy of a degree sequence

- As such for a network, we introduce the **slide entropy** which is the approximate entropy of a binary sequence encoding something like an “infinitesimal” disorder of the finite degree sequence.
- The slide construction is simple: draw the degree distribution as a partition diagram and trace along it from left to right. When you go horizontally one unit write down “0” and vertically one unit, write down “1”.



# Analytics for the slide entropy

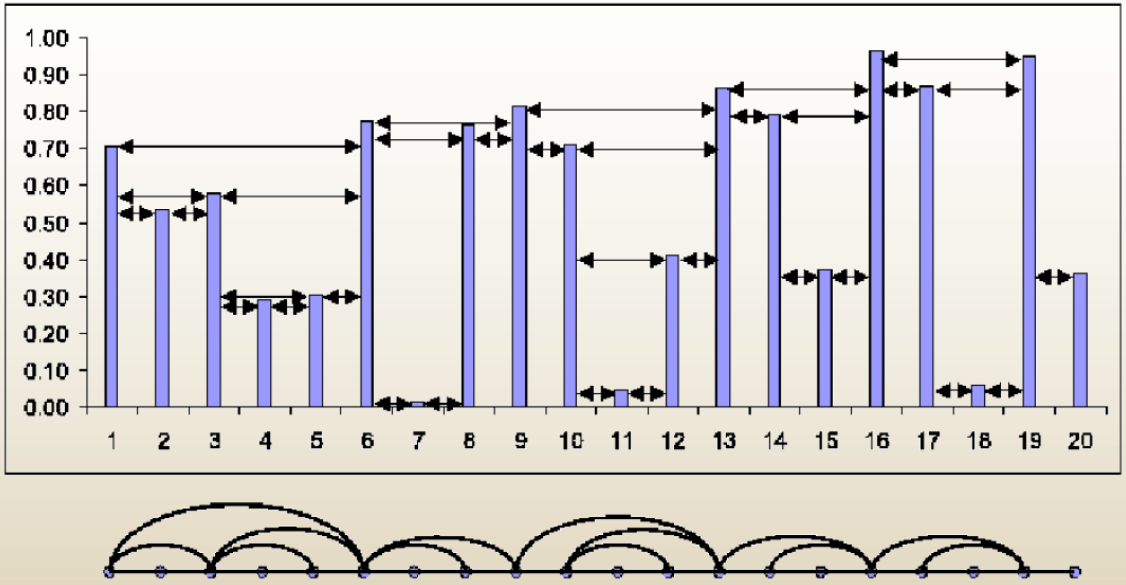
- In nice circumstances (almost surely!) approximate entropy of this slide sequence is recovered asymptotically by a Shannon entropy relating the distribution of 0s and 1s in the associated sequence, yielding an analytic formula of empirically measured reasonable accuracy, say for Poisson networks.
- Almost surely isn't always…




*West et al, Phys Rev E 2012*

# Visibility graphs

- Visibility graphs are networks associated to time series that capture features of the time series in their topology.




Constructing a horizontal visibility graph

Time series

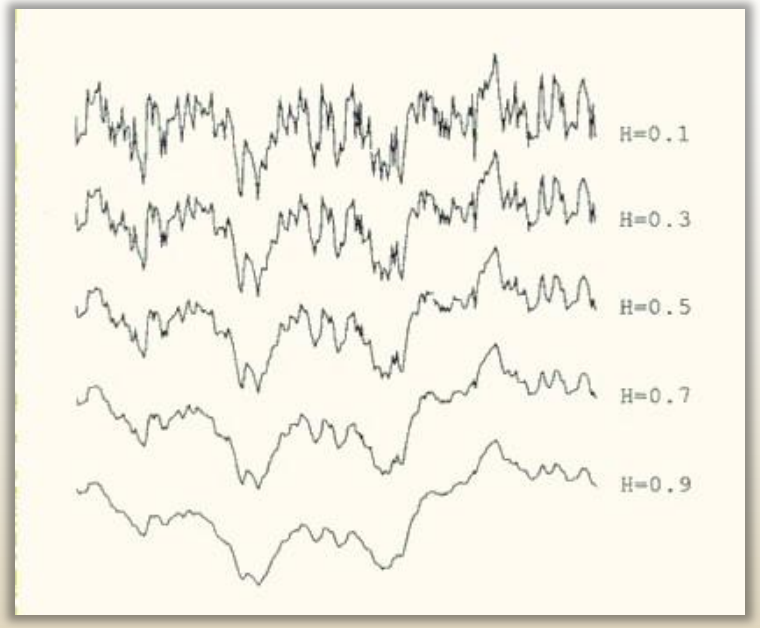
Visibility algorithm

Visibility graph

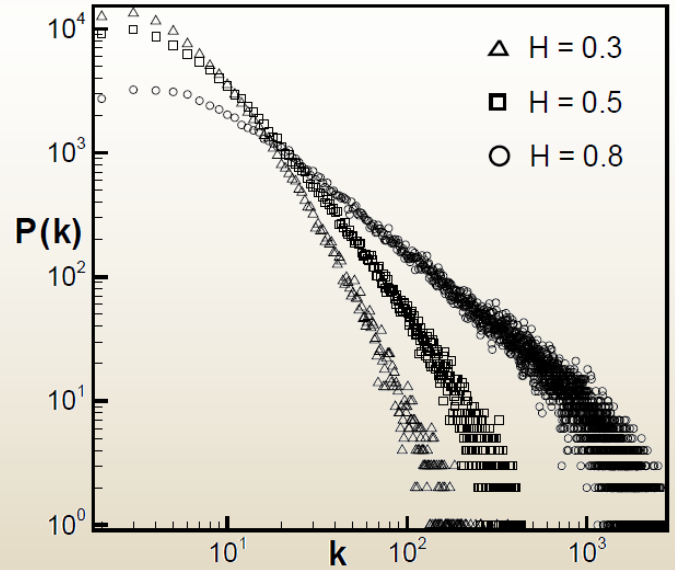
*Lacasa & Toral, Phys Rev E 2010*

# An example application of visibility graphs

- A fractional Brownian motion process of Hurst exponent $H$ gives rise to a scale free visibility graph of scale free parameter $3 - 2H$.



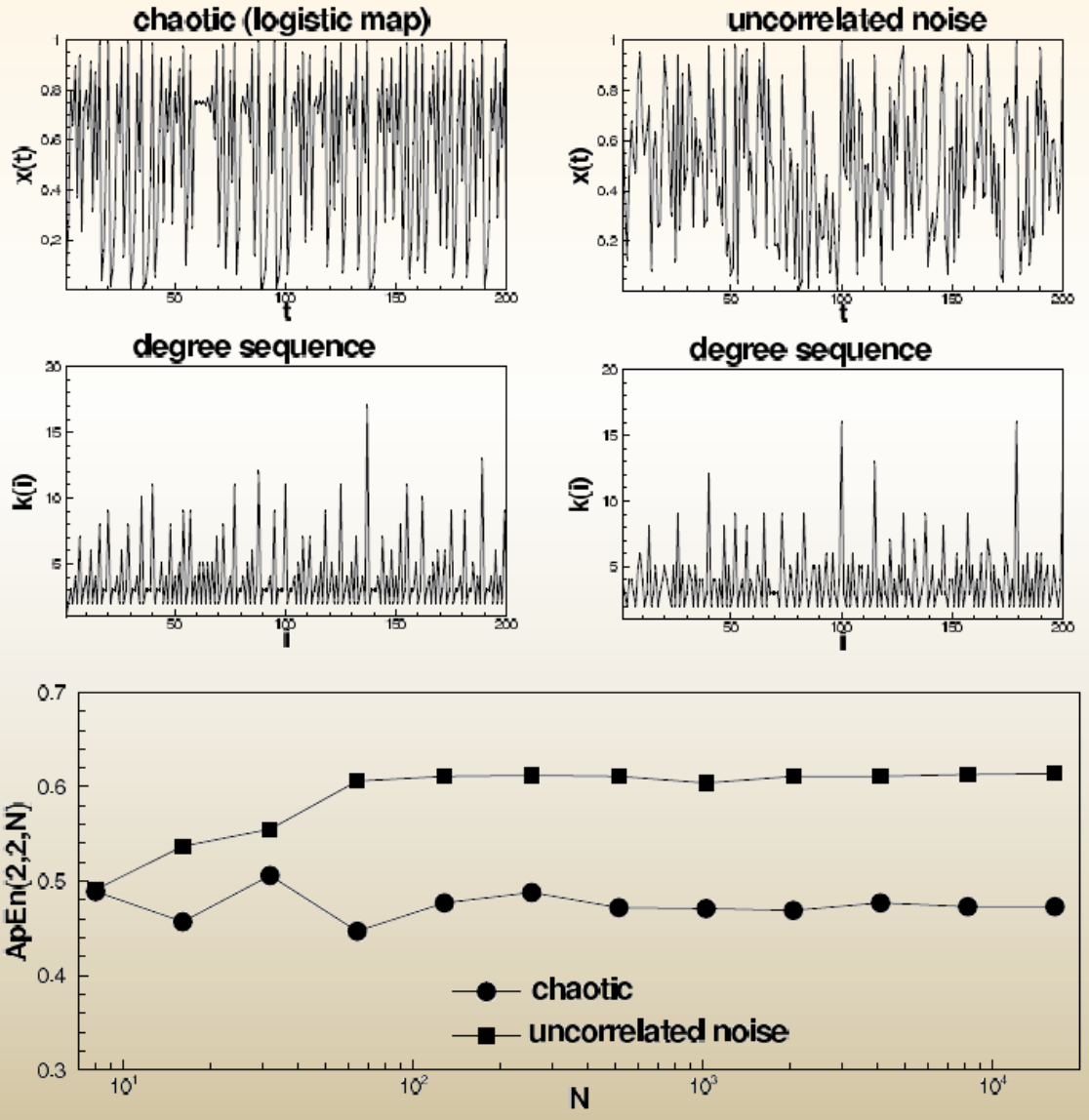

Fractional Brownian motion ($H$).




Degree distribution of associated visibility graph

*Lacasa et al, arXiv:0901.0888v1 2009*

# Approximate entropy of visibility graphs



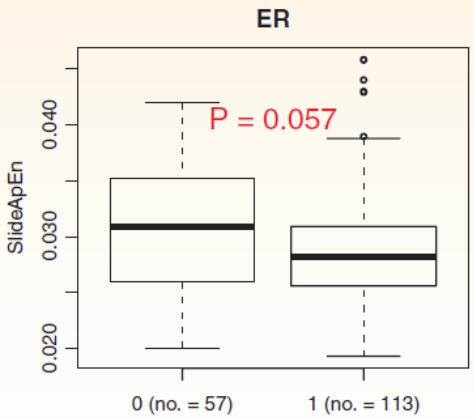

Uncorrelated white noise gives rise to visibility graphs of maximal ApEn (it was known previously also to maximize the Shannon entropy of the degree distribution).

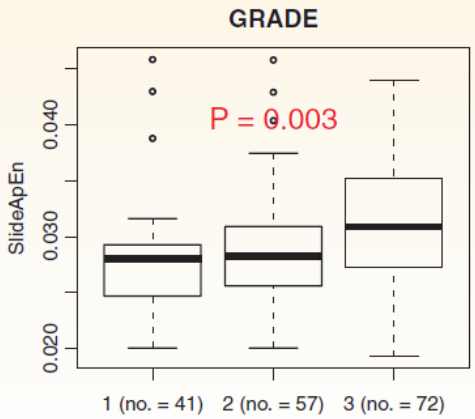
The ApEn of visibility graphs associated to chaotic maps reaches a non-zero value, reminiscent of the underlying attractor of the dynamics.

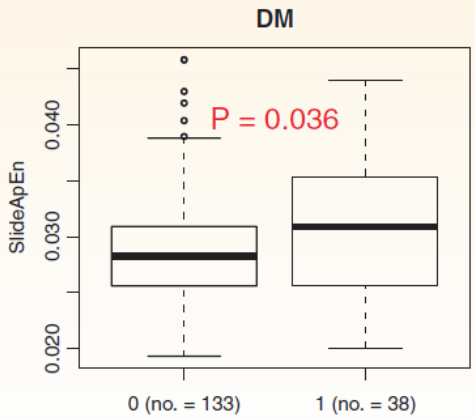
As such, the network structure also inherits the complexity of the time series.

*West et al, Phys Rev E 2012*

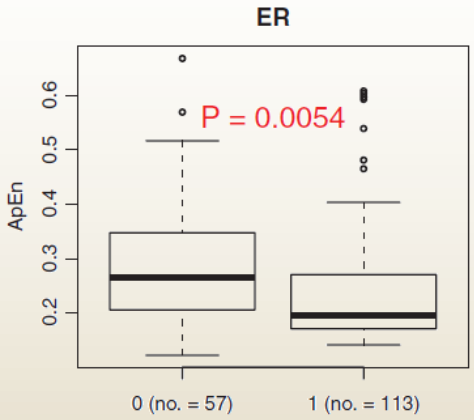
# Approximate entropy of copy number data



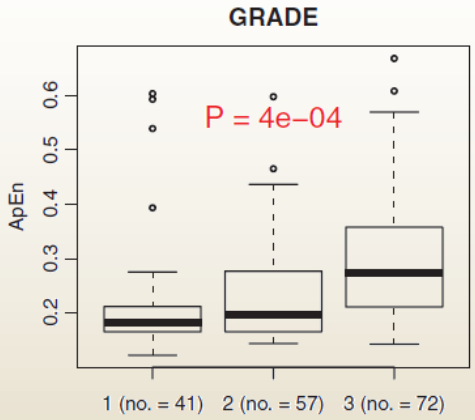



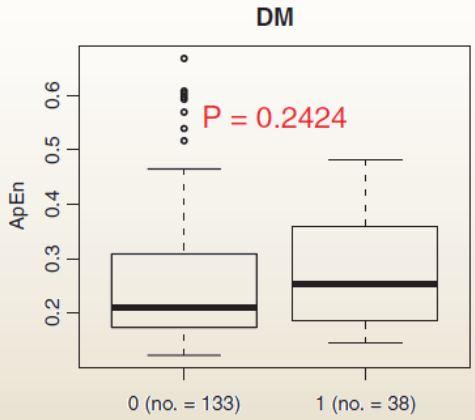

Significant increases in approximate entropy of slide sequence .




Is the cancer over-expressing estrogen receptors?




How fast growing is the cancer?




Has the cancer spread?

N.B. ER-status, Cancer GRADE and Distal Metastasis all correlate in cancers, but remain distinct phenotypic properties.

*West et al, Phys Rev E 2012*

## Summary

1. The **flux entropy** combines network information with gene expression to provide a hallmark of cancer.
2. It is a more consistent indicator than more basic considerations of the correlations, and similarity of samples even when restricted to the network.
3. It may be useful in guiding therapeutic target selection and helps to indicate genes driving cancers.
4. The **approximate entropy** of the slide sequence associated to copy number data might help to distinguish tumor grade and metastasis.

# References


1. "***On dynamical network entropy in cancer***". James West, Ginestra Bianconi, Simone Severini and Andrew E. Teschendorff. arXiv:1202.3015v1 [q-bio.MN].
2. "***Approximate entropy of network parameters***", Physical Review E 2012. James West, Lucas Lacasa, Simone Severini and Andrew E. Teschendorff. arXiv:1201.0045v1 [cond-mat.dis-nn].

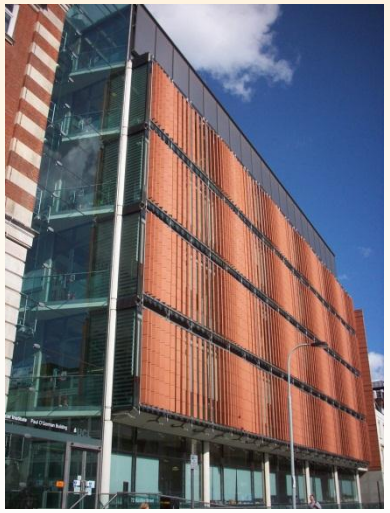
# Acknowledgements

Andrew Teschendorff

Simone Severini

Lucas Lacasa

Ginestra Bianconi



EPSRC

CANCER RESEARCH UK

BBSRC
bioscience for the future

MRC Medical Research Council