Theoretical Physics Methods for Computational Biology.

Third lecture

M. Caselle
Dip di Fisica Teorica, Univ. di Torino

Berlin, 06/04/2006

Third lecture: Theoretical questions in genome biology

- DNA correlators.
- Graph theory for protein interactions networks and regulatory networks.

Analysis of DNA correlations.

Can we identify, using entropy methods and correlation functions, the different types of DNA (coding, non-coding, repeats...) in a typical genome?

Goals:

- Study of global, large scale properties of DNA sequences (HMM are better for short scale properties)
- distinguish coding from non-coding
- find out hidden structures and regularities among different species

Index

- Long range behaviour (100-100000 bp): domain structure
- Short range behaviour:(1-15 bp) binding sites and period three of the coding regions.

Tools:

- Correlators
- Entropy and related quantities

Correlators

The basic measure of correlation among bases in a DNA sequence are the 16 correlation functions between all 16 possible base-pairs:

$$\{\Gamma_{\alpha\beta}(d)\} \equiv \{\Gamma_{AA}(d), \Gamma_{AC}(d), \ldots, \Gamma_{GT}(d), \Gamma_{TT}(d)\}$$

each defined as the correlation between nucleotide α and another nucleotide β separated by a distance d:

$$\Gamma_{\alpha\beta}(d) \equiv P_{\alpha\beta}(d) - P_{\alpha.}P_{.\beta} \qquad \alpha, \beta = \{A, C, G, T\},$$

where $P_{\alpha\beta}(d)$ is the joint probability of observing α and β separated by a distance d, $P_{\alpha.} \equiv \sum_{\beta'} P_{\alpha\beta'}(d)$ and $P_{.\beta} \equiv \sum_{\alpha'} P_{\alpha'\beta}(d)$ are the density for nucleotide α and β, respectively.

Because of the relations:

$$\sum_{\alpha} \Gamma_{\alpha\beta}(d) = \sum_{\beta} \Gamma_{\alpha\beta}(d) = 0$$

the number of independent correlation functions (not yet considering any other symmetries) is actually 9.

Estimation of correlations from a sequence with finite length

Since the correlation at longer distances is typically small, it is important to use the best possible estimator to measure the correlation. Otherwise, the error due to a finite sample size can be as large as the correlation value itself. The simplest and most popular choice is the **Frequency-count estimator**:

The probability of an event a (P_a) is estimated as

$$\widehat{(P_a)}_{freq} = \frac{N_a}{N}.$$

where N_a= the number of counts for a and N= total number of counts.

The frequency-count estimator for the correlation function $\Gamma_{\alpha\beta}(d)$ is:

$$\widehat{\Gamma_{\alpha\beta}(d)}_{freq} = \frac{N_{\alpha\beta}(d)}{N} - \frac{N_{\alpha\cdot}}{N} \cdot \frac{N_{\cdot\beta}}{N}$$

Symmetries

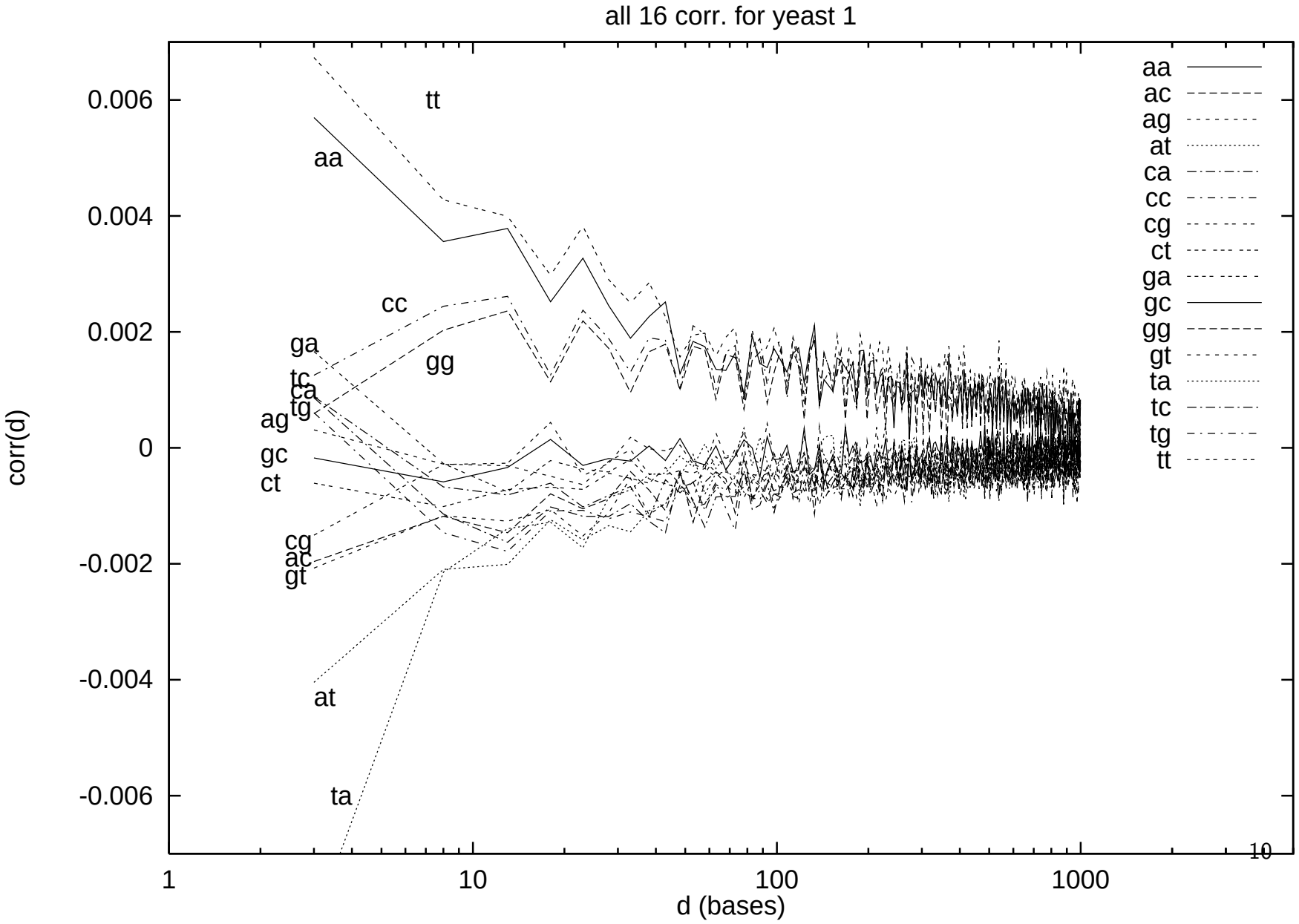
1 Strand complementarity:

DNA sequences are double-stranded with nucleotides on one strand complementary to those on the other. As a result, $\Gamma_{\alpha\beta}(d)$ on one strand (in 5' $\rightarrow$ 3' direction) is exactly the same with the $\Gamma_{\bar{\beta}\bar{\alpha}}(d)^{opposite}$ on the opposite strand in the opposite direction (but also in 5' $\rightarrow$ 3' direction viewed from that strand). For example, $\Gamma_{CT} = \Gamma_{AG}^{opposite}$

2 Strand symmetry:

It was observed that $\Gamma_{\alpha\beta}(d)$ on one strand is *approximately* equal to $\Gamma_{\alpha\beta}(d)^{opposite}$ on the opposite strand in the opposite direction. This suggests the idea of a "strand symmetry". Combining strand symmetry with strand complementarity, we have $\Gamma_{\alpha\beta}(d) \approx \Gamma_{\bar{\beta}\bar{\alpha}}(d)$ on one strand. For example, $\Gamma_{CT} \approx \Gamma_{AG}$. This approximate symmetry reduces the number of independent correlation functions from 9 to 6.

Let us see an example: in the following figure the 16 correlation functions $\Gamma_{\alpha\beta}(d)$ for d from 1 to 1000 are shown for the budding yeast chromosome 1. Correlations at 5 neigbouring distances (e.g. d=1,2,3,4,5) are averaged to smooth the curve. From [W. Li 1997].


all 16 corr. for yeast 1
corr(d)
d (bases)
0.006
0.004
0.002
0
-0.002
-0.004
-0.006
1
10
100
1000
10
tt
aa
cc
ga
gg
tc
ca
tg
ag
gc
ct
cg
ac
gt
at
ta
aa
ac
ag
at
ca
cc
cg
ct
ga
gc
gg
gt
ta
tc
tg
tt

Considerations

1] **Larger symmetries**:

It is evident from the figure that $\Gamma_{AA}(d) \approx \Gamma_{TT}(d)$ and $\Gamma_{CC}(d) \approx \Gamma_{GG}(d)$, with all other cross-correlations roughly similar to each other. This is indeed a general phenomenon. It suggests that more approximate symmetries are present in the data. In particular the simplest way to understand the data is to assume that correlations are approximately the same under simultaneous $A \to T$ and $T \to A$ transformation (e.g. $\Gamma_{AG} \approx \Gamma_{TG}$) and separately an analogous $C \to G$, $G \to C$ transformation. ("binary" approximation)

2] **Anticorrelation**

it is also evident that while correlations between the same nucleotide are always positive, those between different nucleotides are usually negative ("anticorrelation") or compatible with zero. Models have been proposed to understand this behvaiour. The simplest one is the so called "domains structure" model [Li et al. 1994]

3] **Power law behaviour**

The most interesting and surprising feature of these correlations is that they show a power law behaviour. Which is the origin of this "critical" (from the statistical mechanics point of view) behaviour? Also in this case many models have been proposed. The simplest option is to assume a fractal structure of the domains

There are claims that this behaviour is different between coding and non-coding regions. Attempt have been done to use it to distinguish between the two.

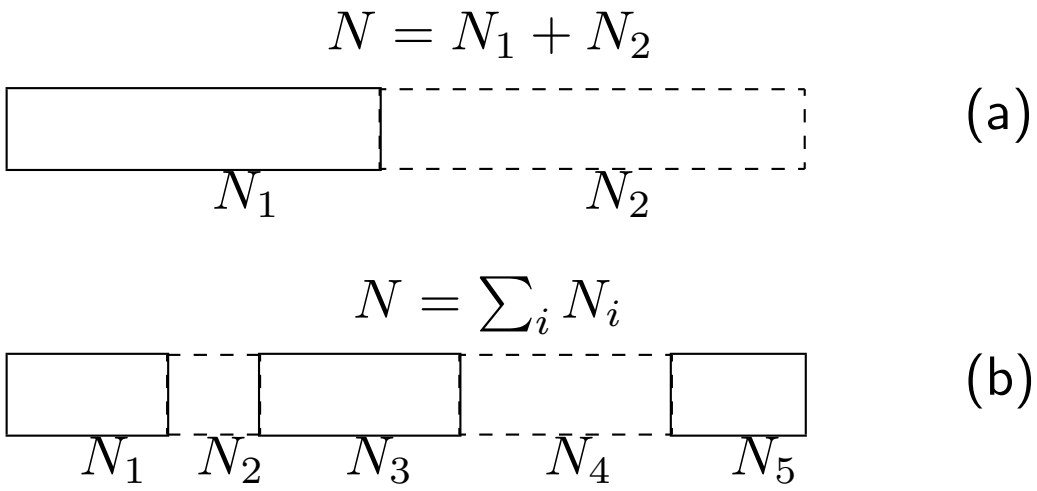


Figure 1: Domain struture model: Illustration of the situation when the sequence can be decomposed into (a) two sub-regions, or (b) many sub-regions, with each sub-region being a white noise, but different base compositions.

Two Sub-region Case: Assume the sequence is divided into two regions of same size $N/2$ and the extreme situations in which the two domains are made only of nucleotides of type α (first domain) and of type β (second domain). The mean values will be:

$$\frac{N_\alpha}{N} = \frac{N_\beta}{N} = \frac{1}{2}$$

If the distance d is $d << N$ we have with a good approximation that also

$$\frac{N_{\alpha,\alpha}}{N} = \frac{N_{\beta,\beta}}{N} = \frac{1}{2} - \frac{d}{N} \sim \frac{1}{2}$$

while

$$\frac{N_{\alpha,\beta}}{N} = \frac{d}{N} \sim 0$$

Thus

$$\Gamma_{\alpha,\alpha} = \Gamma_{\beta,\beta} \sim \frac{1}{2} - \frac{1}{4} = \frac{1}{4} > 0$$

while

$$\Gamma_{\alpha,\beta} = \frac{d}{N} - \frac{1}{4} \sim -\frac{1}{4} < 0$$

A power low d-dependence could appear if we are in presence of a hierarchy of domains (see fig (b)) which have all the possible sizes.

Let us see few other examples in the case of Prokaryotes (taken from [M. de Sousa Vieira 1999] web link: http://xxx.sissa.it/abs/cond-mat/?9905074)

Fig. 3(a)

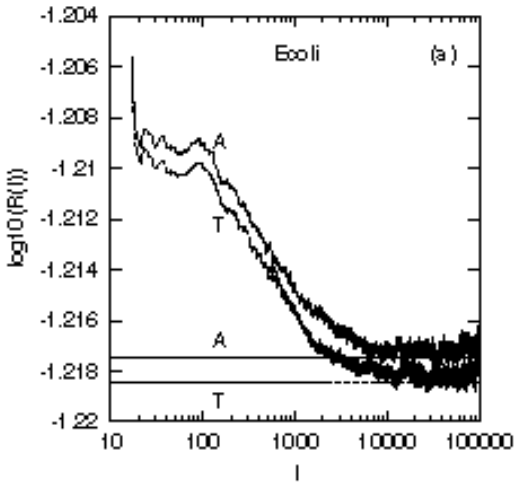


Fig. 3(b)

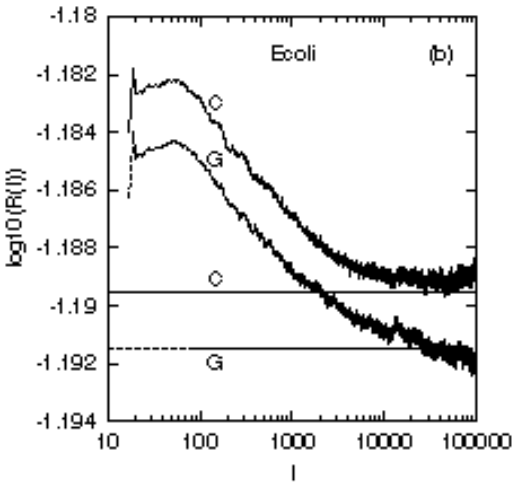


Fig. 3(c)

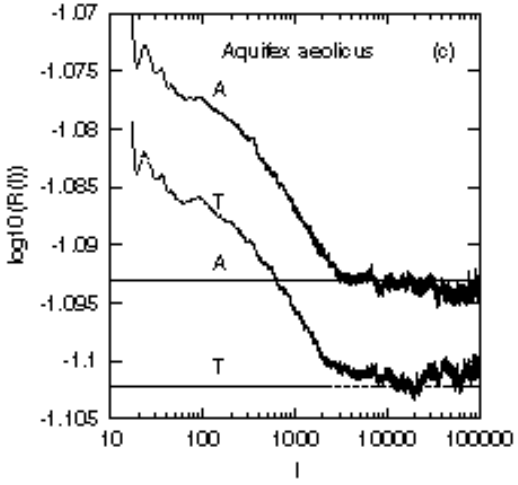


Fig. 3(d)

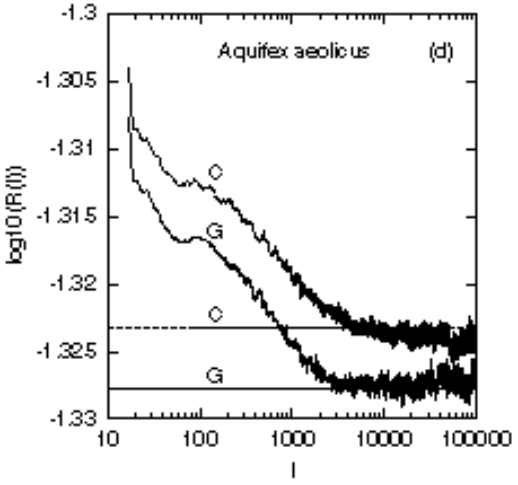


Fig. 3(e)

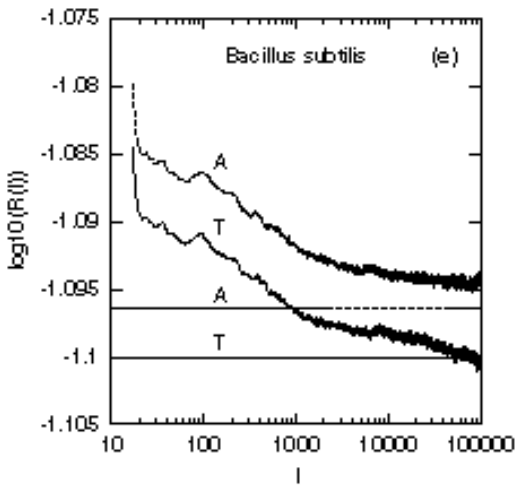


Fig. 3(f)

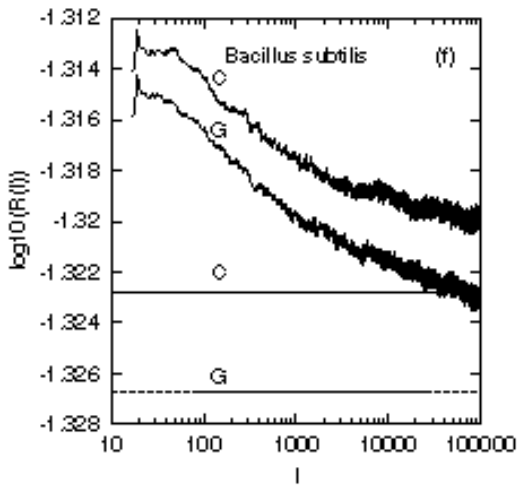


Fig. 3(g)

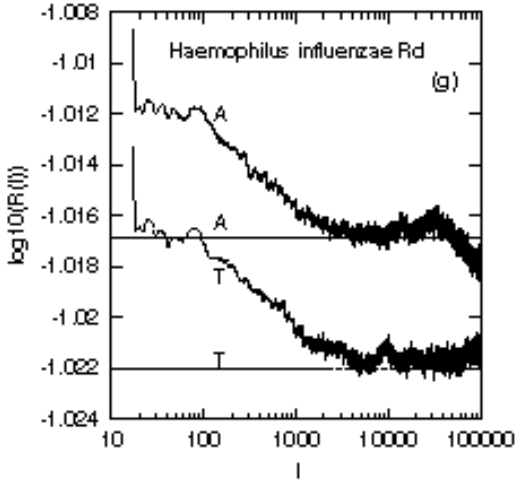
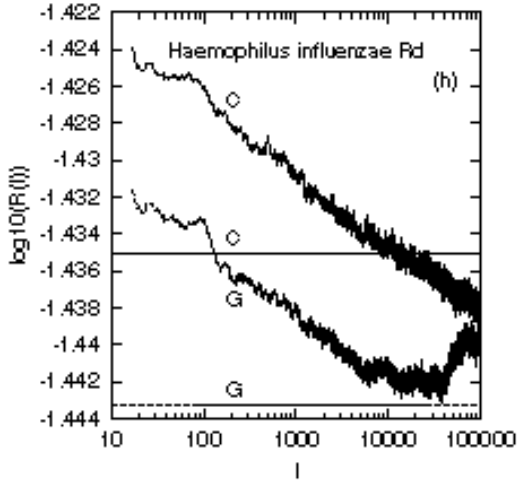


Fig. 3(h)



Generalized Entropies

Goal: use Entropy–like quantities as tools to identify sequence structures.

Definitions: In order to describe the structure of a given string of length L using an alphabet of λ letters $\{A_1 A_2 \dots A_\lambda\}$ we introduce the following notations:

Let $A_1 A_2 \dots A_n$ be the letters of a given substring of length $n \leq L$. We define the probability to find in a string a block of length n (subword of length n) with the letters $A_1 \dots A_n$ as

$$p^{(n)}(A_1 \dots A_n).$$

Entropies: We shall use the following quantities:

1. Shannon Entropy:

$$H_1 = -\sum_{i=1}^{\lambda} p^{(1)}(A_i) \log p^{(1)}(A_i) \ ,$$

a few results: for white noise $H_1 = log(\lambda)$, which is a maximum. If all the symbols are the same, $H_1 = 0$.

2. Shannon word entropy

the entropy per word of length n is given by

$$H_n = -\sum p^{(n)}(A_1 \dots A_n) \log p^{(n)}(A_1 \dots A_n) \ ,$$

The sum is over λ^n entries. The limit

$$H_{met} = \lim_{n \to \infty} \frac{H_n}{n}$$

is usually called metric Entropy. In case of white noise it should be $\frac{H_n}{n} = log(\lambda)$ See figures below

3. The differential Entropy

   This quantity essentially describes the uncertainty of the letter following a block of length n

$$h_n = H_{n+1} - H_n \ ,$$

4. The entropy of the source (related to the Kolmogorov–Sinai entropy)

$$h = \lim_{n \to \infty} h_n \ .$$

5. The mutual information, also called transinformation Let us define the probability of having a pair with $(n-2)$ arbitrary letters in between as

$$p^{(n)}(A_1, A_n) = p^{(n)}(A_1 \ ? \ ? \ ? \ ? \ A_n) \ .$$

Then we may define the mutual information as:

$$I(n) = \sum_{A_i A_j} p^{(n)}(A_i, A_j) \log \left(\frac{p^{(n)}(A_i, A_j)}{p^{(1)}(A_i) \cdot p^{(1)}(A_j)} \right) \ ,$$

This quantity can be understood as follows: Consider a compound system (X, Y) consisting of two susystems X and Y. Let p_i denote the probability of finding system X in state i and q_j that of finding system Y in state j . Let p_{ij} denote the probability of finding the compound system in state (i,j). Then the entropies of X, Y and (X, Y) are:

$$H_1(X) = -\sum_{i=1}^{\lambda} p_i \log p_i \ ,$$

$$H_1(Y) = -\sum_{j=1}^{\lambda} q_j \log q_j \ ,$$

$$H_1(X, Y) = -\sum_{i,j=1}^{\lambda} p_{ij} \log p_{ij} \ ,$$

Then:

- if the two systems are scorrelated $H(X) + H(Y) = H(X, Y)$
- we can apply the formalism to a sequence assuming that X and Y are two bases located at a distance, say, n and using the probability defined above.
- then it is easy to see that

$$I(n) = H(X) + H(Y) - H(X, Y)$$

Let us see two applications.

First a systematic study in Eukaryotes and Prokaryotes of the behaviour of H_n/n versus n In particular in the following three figures one can see Fig. 1 : I, IV, VI and XV yeast chromosomes.

Fig. 2 $rpxx$, $ecoli$, $tpal$, $mtub$, $mpneu$, $mgen$ bacteria genomes.

Fig. 3 comparison of human 22 chromosome, genomes of bacteria: $synecho$, $aful$, $hinf$, $hpyl$ and XV, XIII, VI, IX yeast chromosomes.

One can see an impressive decrease of the entropy as larger words are taken into account.

Second: Use of the mutual information function to identify coding from non coding regions. With this indicator the period 3 in the coding region becomes evident (taken from [Grosse et al., 2000]

fig. 1

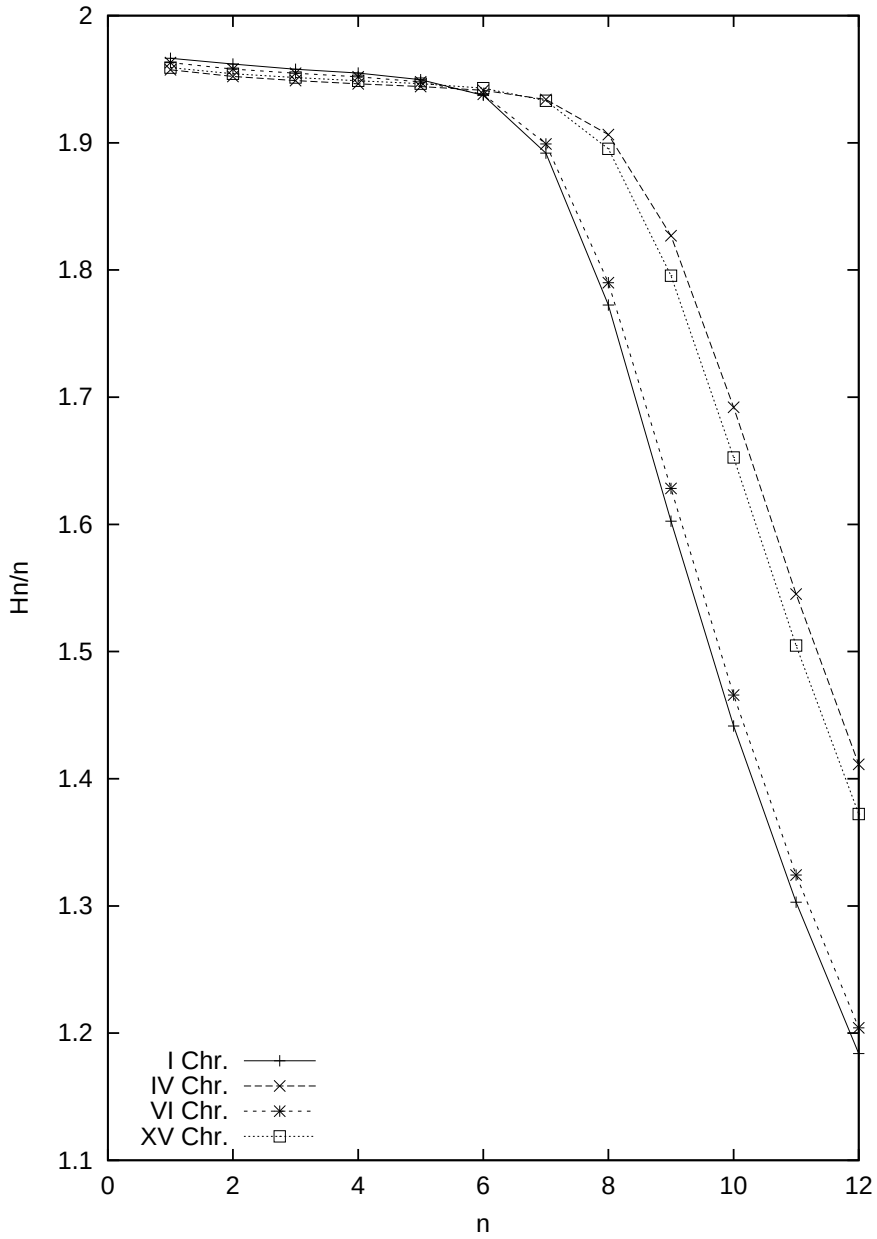

2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
Hn/n
0
2
4
6
8
10
12
n
I Chr.
IV Chr.
VI Chr.
XV Chr.

fig. 2

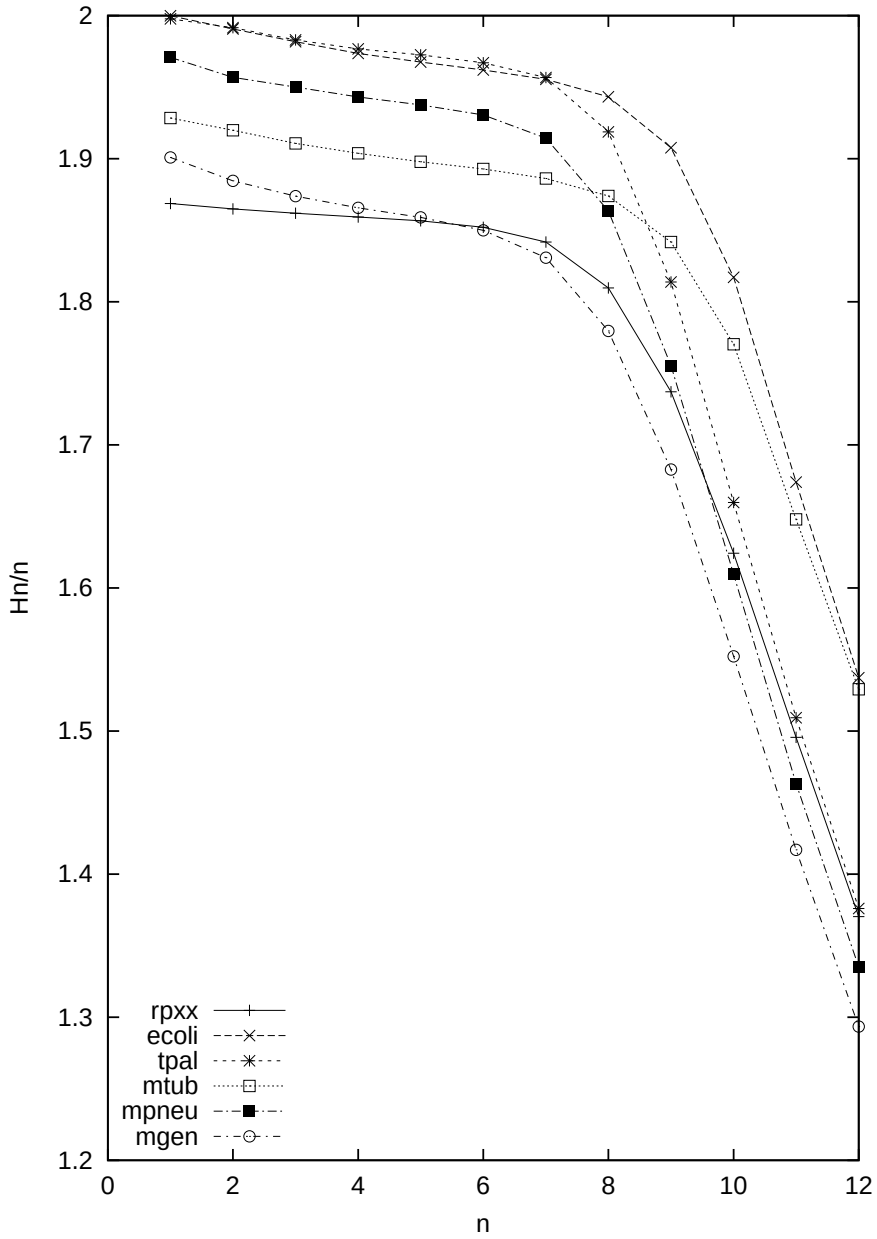
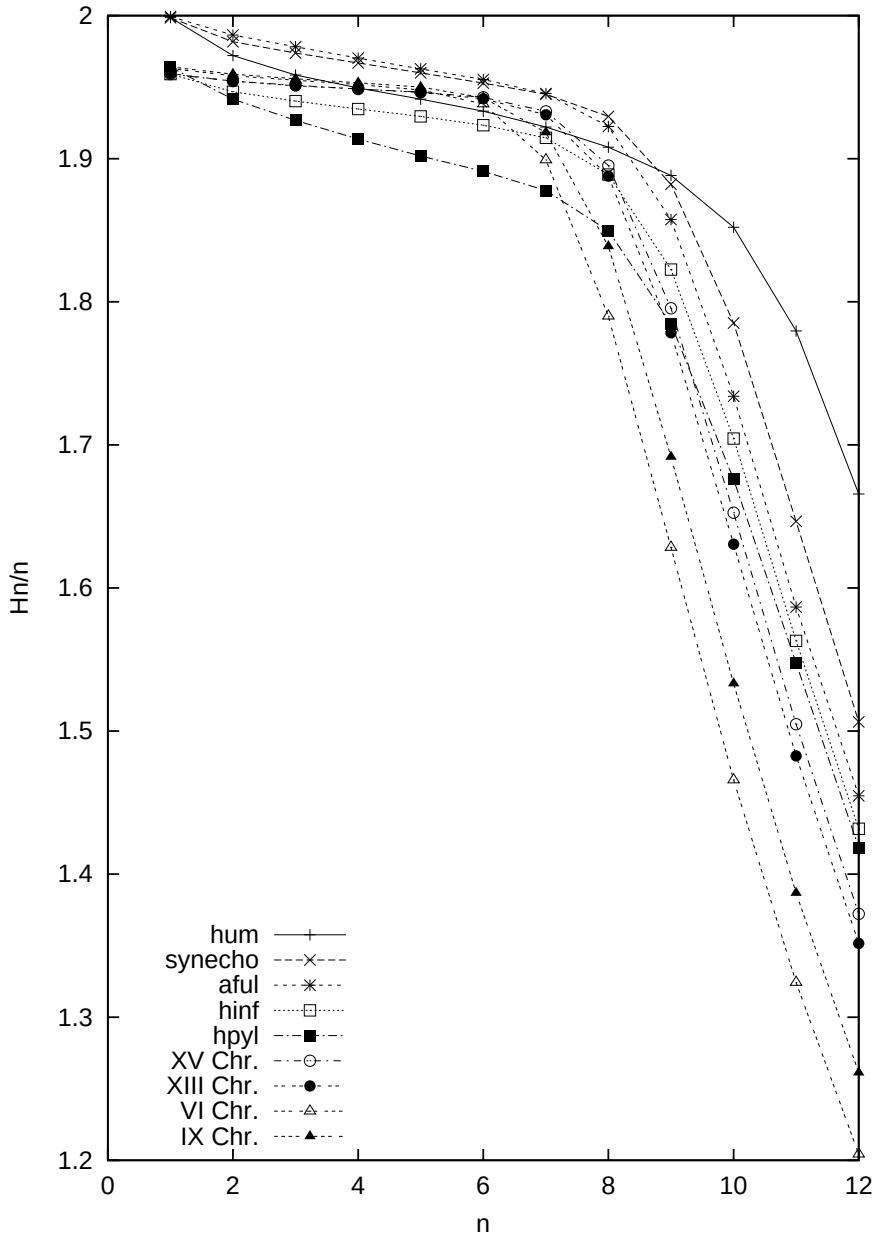

2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
Hn/n
0
2
4
6
8
10
12
n
rpxx
ecoli
tpal
mtub
mpneu
mgen

fig. 3


Hn/n
n
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
0
2
4
6
8
10
12
hum
synecho
aful
hinf
hpyl
XV Chr.
XIII Chr.
VI Chr.
IX Chr.

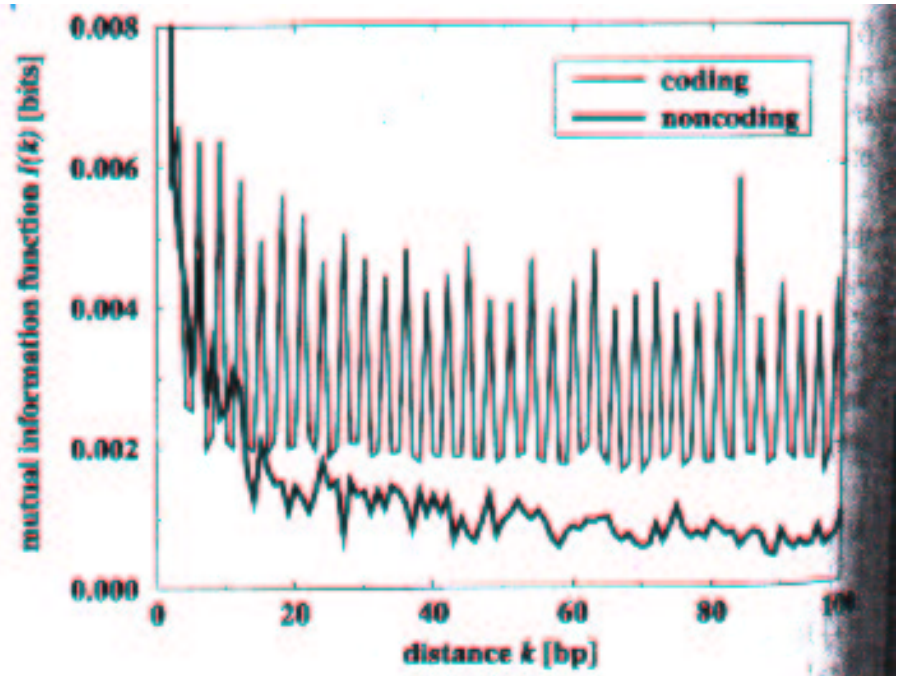
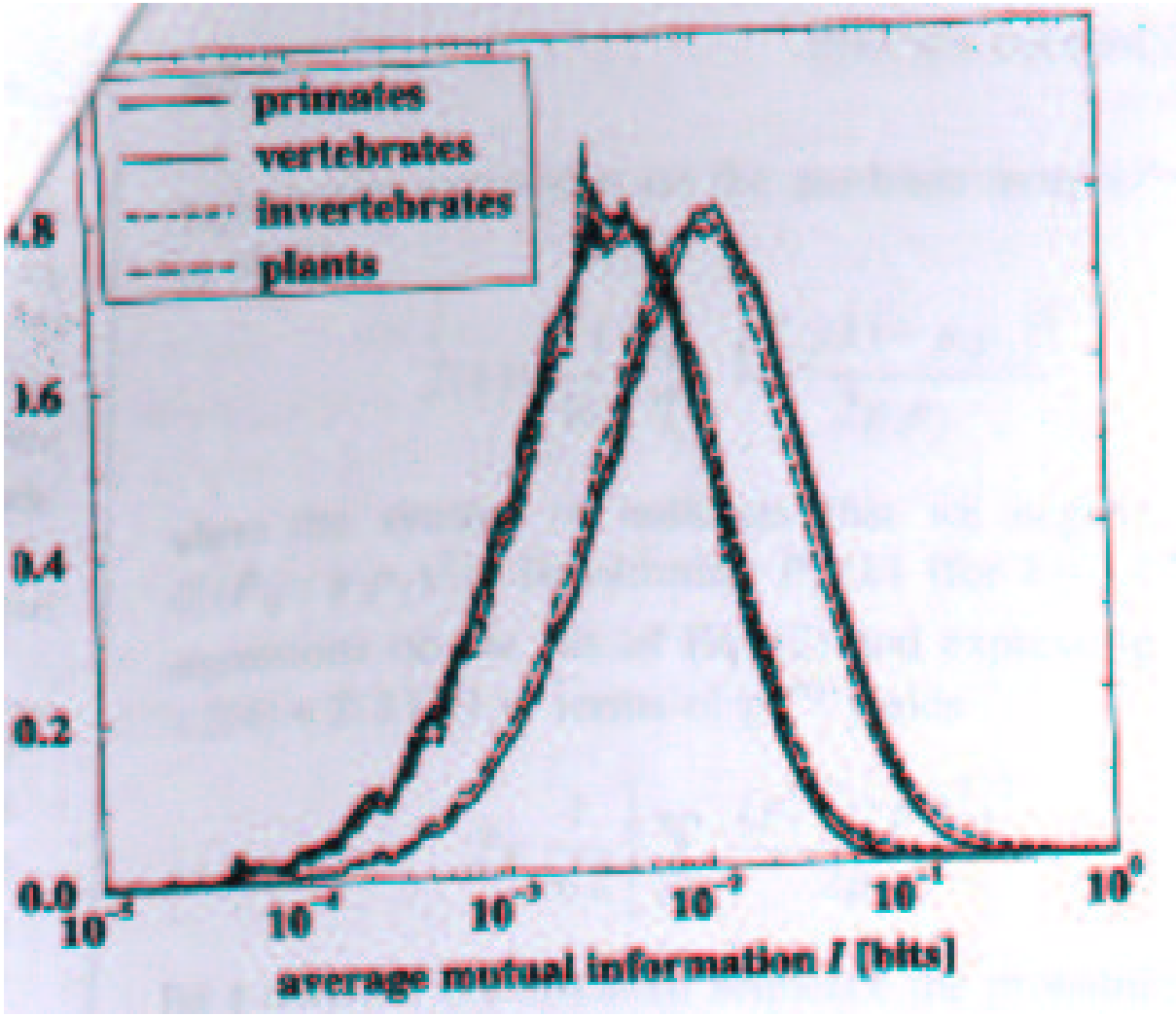


FIG. 1. Mutual information function, $\mathcal{I}(k)$, of human coding (thin line) and noncoding (thick line) DNA, from GenBank release 111 (Ref. [10]). We cut all human, non-mitochondrial DNA sequences into non-overlapping fragments of length 500 bp, starting at the 5'-end. We compute the mutual information function of each fragment, correct for the finite length effect (Ref. [13]), and display the average over all mutual information functions (of coding and noncoding DNA separately). We find that for noncoding DNA $\mathcal{I}(k)$ decays to zero as k increases, while for coding DNA $\mathcal{I}(k)$ shows persistent period-3 oscillations.

Defining the average mutual information as follows:

$$\hat{I} = \frac{1}{3}[I(3) + I(4) + I(5)]$$

one finds, sampling the DNA chain into nonoverlapping fragments of 54 bp [Grosse et al. 2000]


primates
vertebrates
invertebrates
plants
.8
1.6
0.4
0.2
0.0
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
average mutual information *I* [bits]

Means(variance) of $log_{10}(\hat{I})$ for coding and noncoding DNA

	Noncoding	Coding
Primates	−2.52 (0.31)	−2.04 (0.30)
Nonprimate vertebrates	−2.54 (0.39)	−2.06 (0.30)
Vertebrates	−2.53 (0.34)	−2.05 (0.30)
Invertebrates	−2.50 (0.33)	−2.04 (0.32)
Animals	−2.52 (0.34)	−2.05 (0.31)
Plants	−2.48 (0.31)	−2.09 (0.31)

References

[1] S. Zoubak, O. Clay, G. Bernardi (1996), "The gene distribution of the human genome", Gene, 174, 95-102.

[2] B.T.M. Korber, R.M. Farber, D.H. Wolpert, A.S. Lapedes (1993), "Covariation of mutation in the V3 loop of HIV-1: an information theoretic analysis", Proceedings of National Academy of Sciences, 90, 7176-7180.

[3] H. Herzel, I. Grosse (1995), "Measuring correlations in symbol sequences", Physica A, 216, 518-542.

[4] H. Herzel, I. Grosse (1997), "Correlations in DNA sequences - the role of protein coding segments", Physical Review E, 55, 800-810.

[5] I. Grosse et al (2000), "Species indipendence of mutual information in coding and noncoding DNA", Physical Review E, 61, 5624.

[6] W. Li, K. Kaneko (1992b), "DNA correlations" (scientific correspondence), Nature, 360, 635-636.

[7] W. Li, "The study of correlation structures of DNA sequences: a critical review." http://babbage.sissa.it/abs/adap-org/9704003

[8] W. Li, T. Marr, K. Kaneko (1994), "Understanding long-range correlations in DNA sequences", Physica D, 75, 392-416 [erratum, 82, 217 (1995)].

[9] G. Bernardi (1989), "The isochore organization of the human genome", Annual Review of Genetics, 23, 637-661.

Graphs and Biology

Graphs turn out to be a clever way to organize data in Biology. They are in particular efficient to disentangle hidden collective behaviours and above all to detect universal behaviours. The typical questions which one can pose to a graph description of a biological problem are:

- To which class of networks the graph in which we are interested belongs?
- Which is the degree of universality of the observed pattern?
- Which are the biological implications of the relevant graph observables? i.e. what can we learn from the graph?

Main observables

From the adjacency matrix one can easily obtain some general observables, which, even if very simple, contain a lot of important information on the structure and properties of the graph:

- Clusters: disjoint connected components of the graph.
- Percolating cluster: giant connected cluster (of size comparable with the number of vertices)
- Number of clusters: the function $N_c(c)$ which counts the number of clusters of size c in the graph

- Degree of a vertex (also called connectivity): number of links of a given vertex. Notation: $z(i)$. The mean connectivity is denoted by z.

- Connectivity distribution: The function $N_k(k)$ which counts the number of vertices with degree k

Non trivial observables

There are also other quantities which are slightly more difficult to evaluate and also describe important (and less obvious) properties of the graph.

- Clustering coefficient C

  Mean probability that two nearest neighbours of a given vertex are also nearest neighbours among them. For a random graph this quantity can be evaluated exactly: $C = z/N$ and is very small. In most of real networks it is a much larger number.

- Diameter

  Mean distance (evaluated over all the possible pairs of vertices). Recall that the distance between two vertices is defined as the minimum number of links needed to join the two vertices. It is rather large in regular graphs. In most of real networks (and in random graphs) it is much smaller.

- Assortativity The assortativity A of a network is the linear correlation coefficient between the connectivities of the nodes on the two ends of a link.
  - $A > 0$: (assortative network) nodes are preferentially linked to nodes of similar connectivity
  - $A < 0$: (dissortative network): high connectivity nodes are preferentially linked to low-connectivity ones

Roughly speaking A measures the probability that two highly connected vertices are connected between them.

A dissortative graph is composed by isolated hubs. Removal of a high-connectivity node strongly disrupts the topology of the network $\longrightarrow$ highly connected nodes are "essential".

An assortative graph is composed by a core of highly connected hubs. Removal of a hub does not affect the topology of the network $\longrightarrow$ high degree of "redundancy".

Universality classes

There are several classes of graphs. In particular we concentrate here on three families:

- random graph
- "exponential" graph
- "power-like" graph

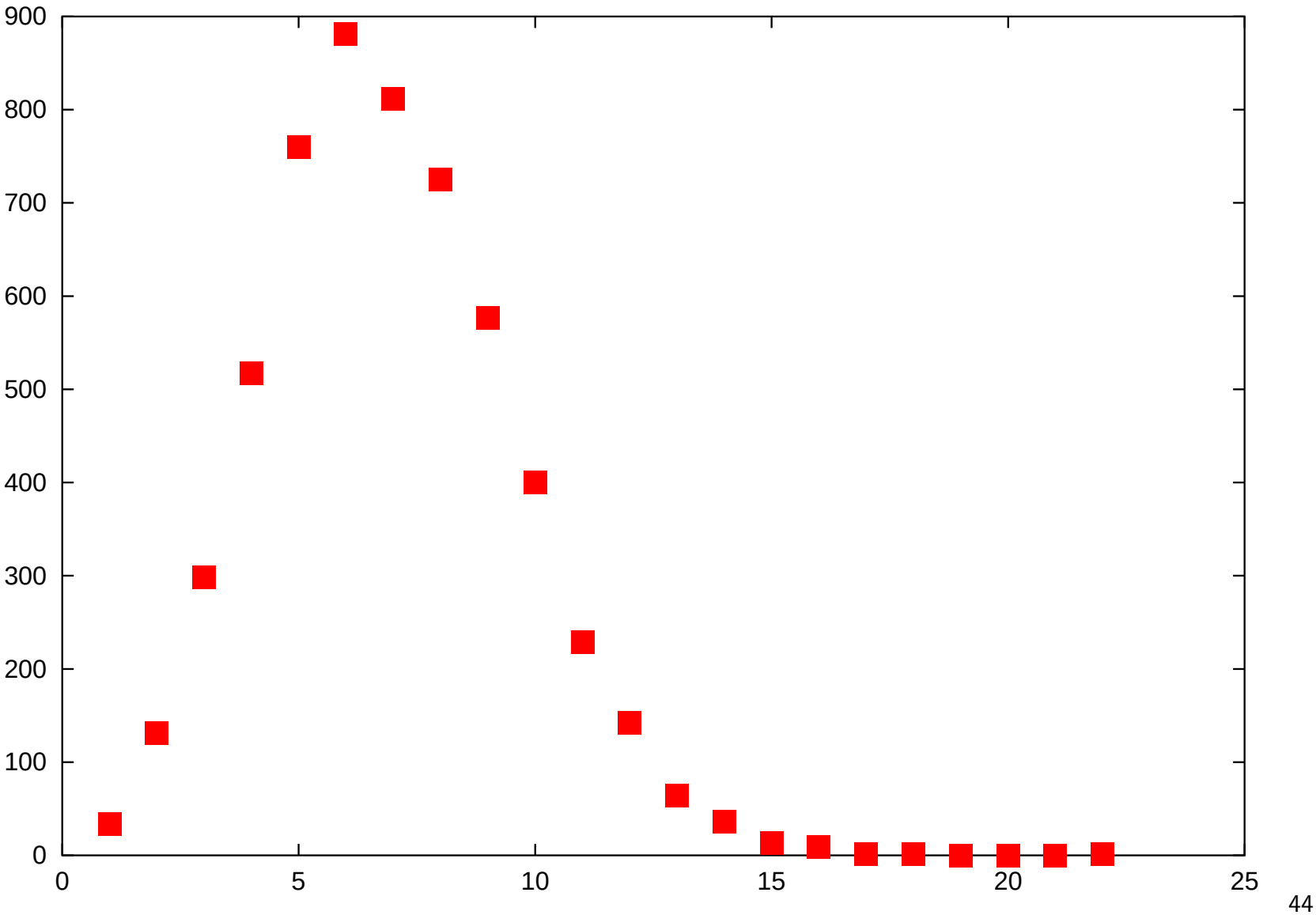
They should be considered as "reference graphs" against which we compare the real graph in which we are interested.

Random graph

The random graph is obtained by randomly switching on links between the vertices, with a probability p. By increasing p one smoothly moves from the empty graph made of N disjoint vertices: $(p = 0, \quad , N_l = 0, \quad N_c = N)$ to the complete graph: $(p = 1, \quad N_l = N(N-1)/2, \quad N_c = 1)$.

The number of links for a given value of p is $N_l = pN(N-1)/2$ hence $z = p(N-1)$

The connectivity distribution is (in the large N limit) a Poisson distribution


900
800
700
600
500
400
300
200
100
0
0
5
10
15
20
25

$$N_k = \binom{N}{k} p^k (1-p)^{N-k} \sim \frac{z^k e^{-z}}{k!}$$

The distribution has a peak for $k = z$ i.e. most of the vertices has degree z and only an exponentially small number of vertices has a very large or very small degrees.

Remarkable feature of the random graph: there is a percolation threshold above which a percolating cluster appears.

Exponential graphs.

The connectivity distribution is in this case:

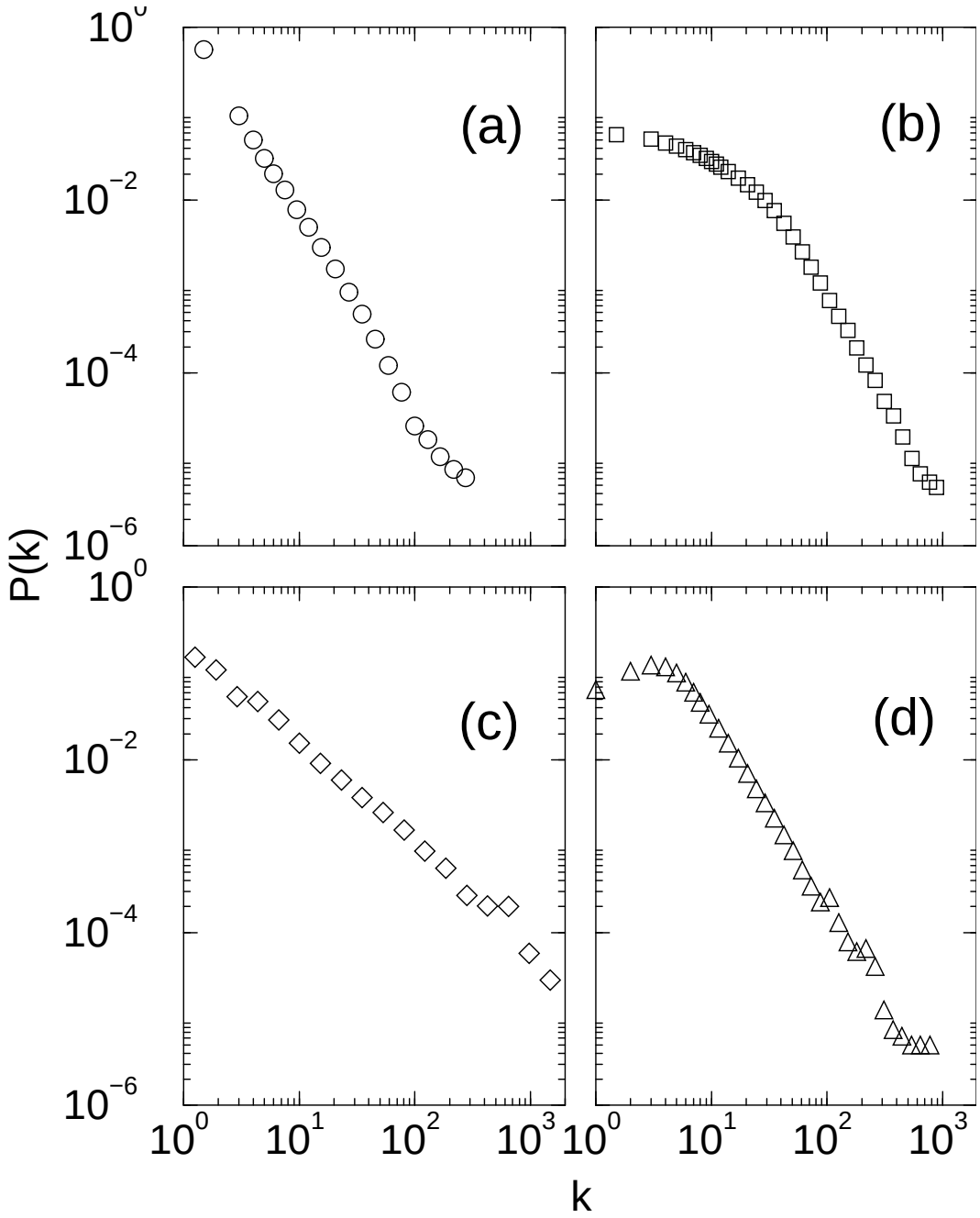
$$N_k = Ce^{-k/\xi}$$

As for the random graph large (with respect to ξ) degrees are always suppressed. However small degrees in this case are always enhanced.

However many real world networks are not of random or exponential type. The easiest way to see this is to look at the connectivity distribution which turns out to be power like.

See the following examples taken from Albert and Barabasi, Rev. Mod. Phys. 74, 47 (2002).

- (a) Internet at the router level
- (b) Movie actor collaborations network
- (c) High energy physics coauthorship
- (d) Neuroscience coauthorship.


(a)
(b)
(c)
(d)
P(k)
k
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{0}
10^{1}
10^{2}
10^{3}

Power-like graphs.

The connectivity distribution is:

$$N_k = Ck^{-\tau}e^{-k/\xi} \qquad (k > 0)$$

with C a suitable normalization constant.

If ξ is large enough there is a window in which the connectivity distribution follows a power like decay with exponent $-\tau$. The exponential cutoff at large values of k is a consequence of the finite number of vertices of the graph. The typical feature of this class is that it contains a small number of highly connected vertices (hubs) and a large number of peripherical vertices with small degree.

The scale-free model
(Barabasi and Albert, Science 286:589 1999)

A model (almost) as simple as the Erdos-Renyi random graph, giving rise to a connectivity distribution decaying as a power law. The ingredients are

- Growth: new nodes are added to the network one at a time and linked to existing nodes

- Preferential attachment: new nodes are attached preferentially to existing nodes of high connectivity

$$P(k) \propto k^{-\gamma} \qquad \gamma = 3$$

Protein Networks

Let us see as an example the network of protein-protein interaction in S. Cerevisiae.

Ref: A.-L. Barabasi et al. Nature 411 (2001) 41

In this case the vertices of the graph are proteins and the links the (experimentally validated) protein-protein interactions. The network has 1870 vertices and 2240 links. From the adjacenty matrix it is easy to reconstruct the cluster decomposition. It turns out that the graph is composed by a giant cluster of 1458 proteins (78%); 4 isolated clusters with 7 proteins and 168 clusters with a number of component less or equal to 6.

The connectivity distribution is power like and follows the law:

$$N_k = a k^{-\tau} e^{-k/\xi}$$

with $\xi \sim 20$ e $\tau \sim 2.4$.

The same pattern is found if one studies:

- a different set of interaction in yeast

  Ref: Schiwokswki et al. Nature Biotechnol. 18 (2000) 1257

  made of 1825 proteins and 2238 links

- the protein interaction network in H. Pylori

  Ref: Rain et al. Nature 409 (2001) 211

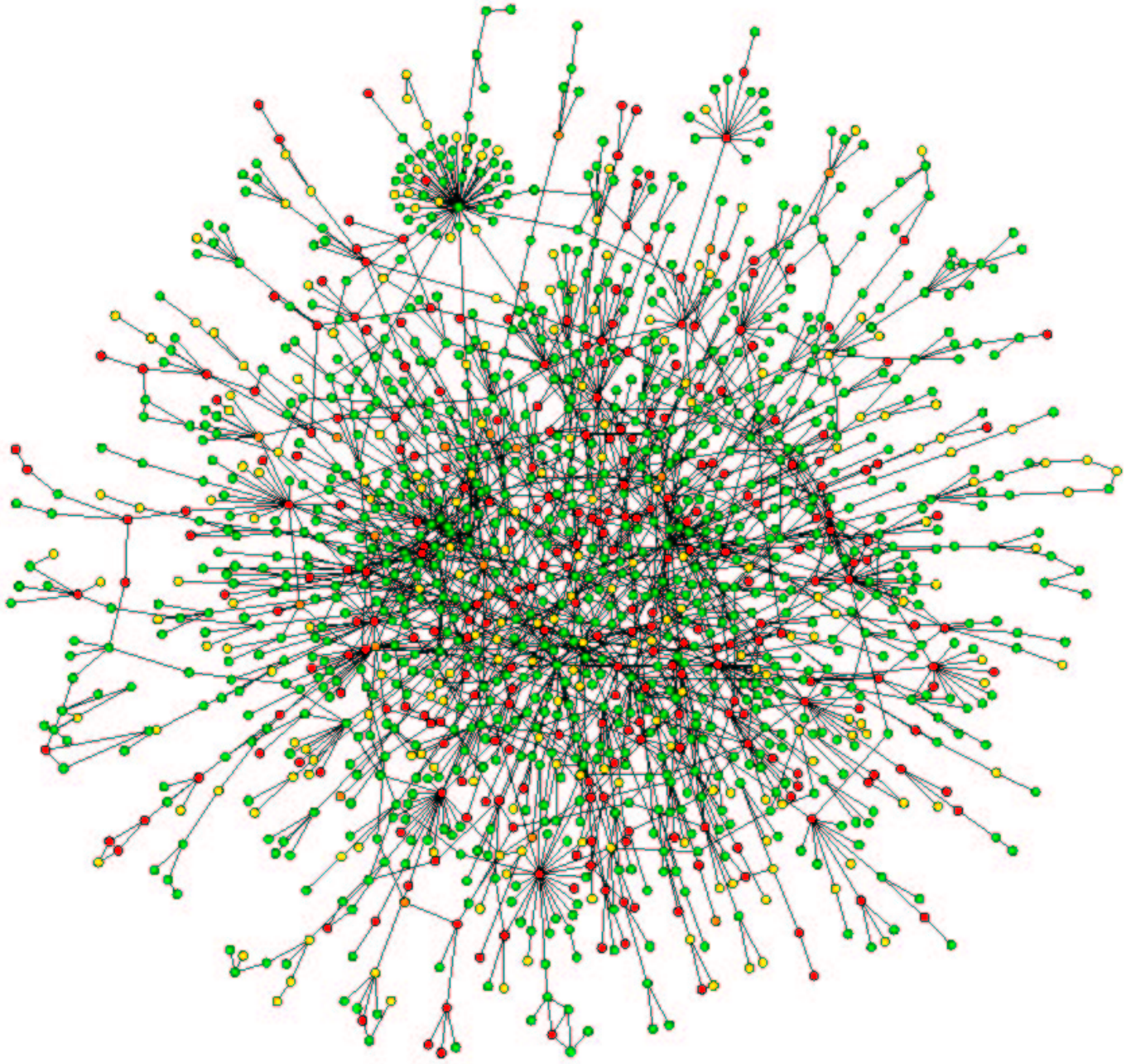
  made of 724 proteins and 1407 links

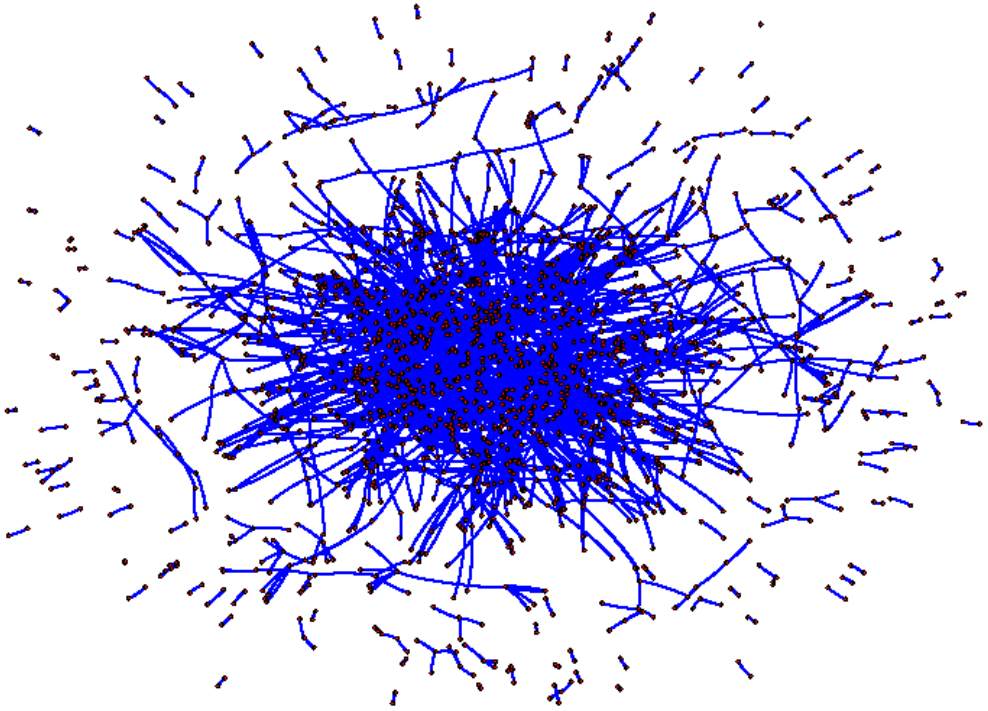
- The network of interaction between nuclear proteins only in yeast (see file 0205380.pdf).

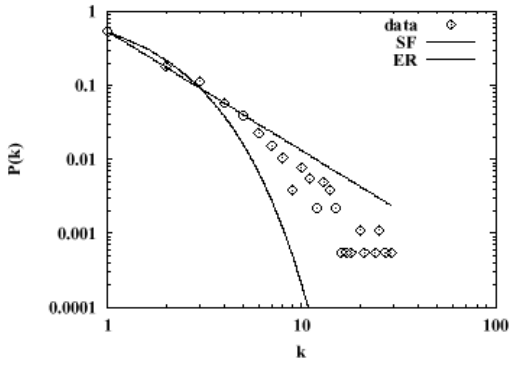
  Ref: Maslov and Sneppen, cond-mat/0205380

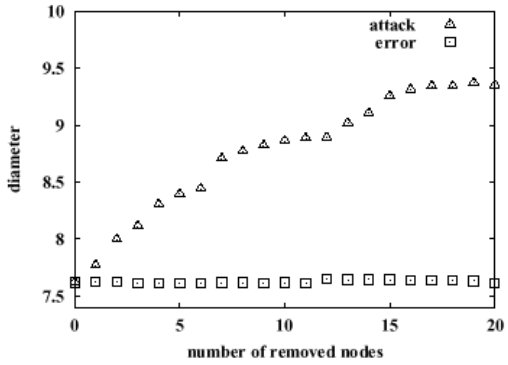
  in this case we have 329 proteins and 318 links

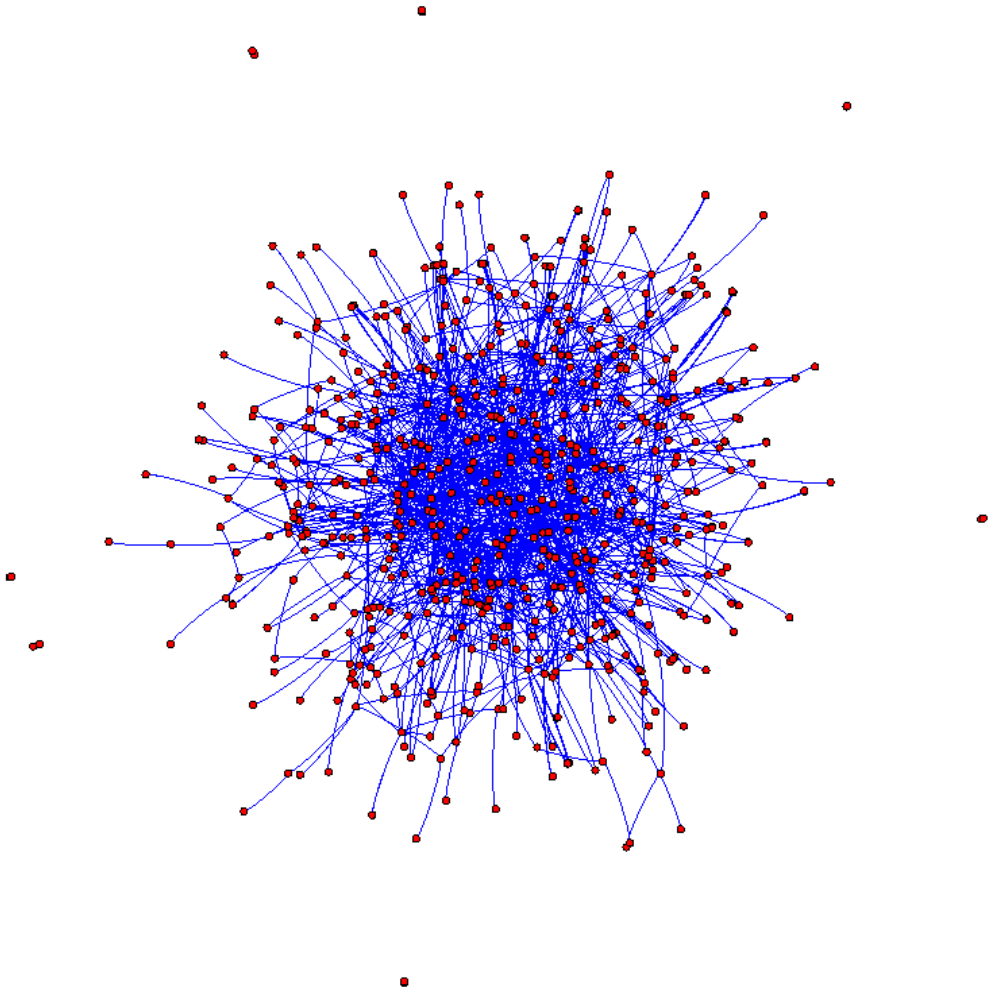
Universality seems to be fulfilled

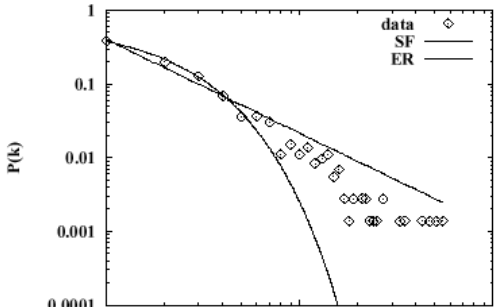


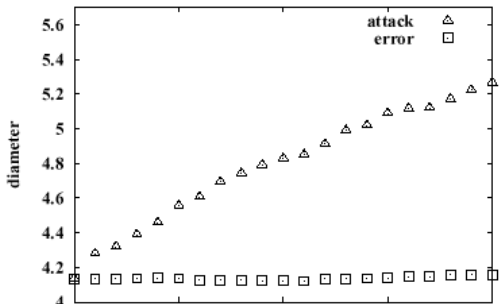



data
SF
ER
P(k)
k
1
0.1
0.01
0.001
0.0001
1
10
100


attack
error
diameter
number of removed nodes
10
9.5
9
8.5
8
7.5
0
5
10
15
20




data
SF
ER
P(k)
1
0.1
0.01
0.001
0.0001


attack
error
diameter
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4

From this analysis we extract three suggestions which could be of biological relevance and can be tested directly on the graph:

1] Proteins with larger connectivity are the **lethal** ones, i.e. are those without which the cell cannot survive. To test this hypothesis one must study the correlation between connectivity and lethality.

This hypothesis is indeed confirmed by the data:

93% of proteins has a connectivity less or equal 5, but only 21% of them is lethal. Similarly only 0.6% has connectivity greater than 15 but 62% of them turns out to be lethal. This means that highly connected proteins have a much larger probability to be lethal.

2] Highly connected proteins are indispensable to keep the network topology. If we selectively eliminate these proteins the diameter of the

network increases dramatically while with a random elimination process almost nothing happens.

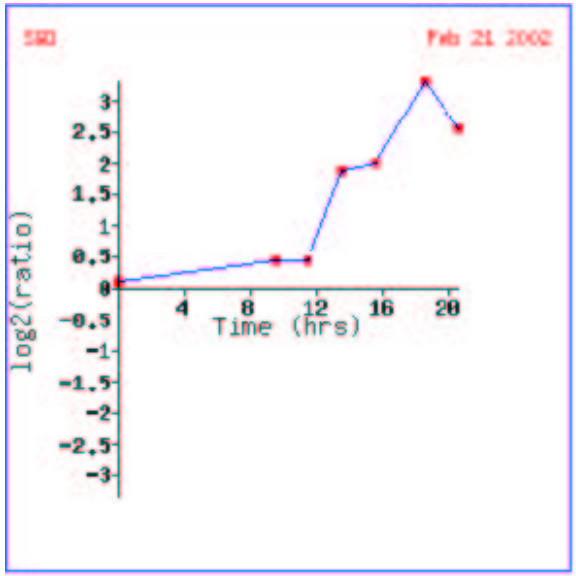
3] The number of links between highly connected vertices is smaller than expected on a random basis.

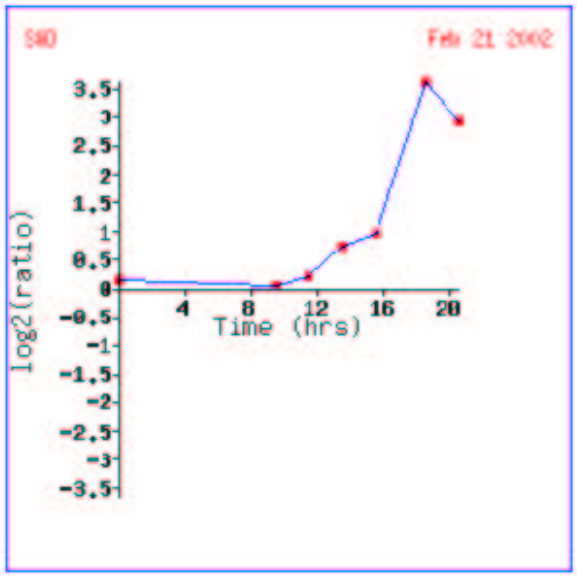
This means that the protein network is a "dissortative graph". This effect decreases the likelihood of cross talk between different functional modules of the cell, and increases the overall robustness of the network by localizing effects of deleterious perturbations.

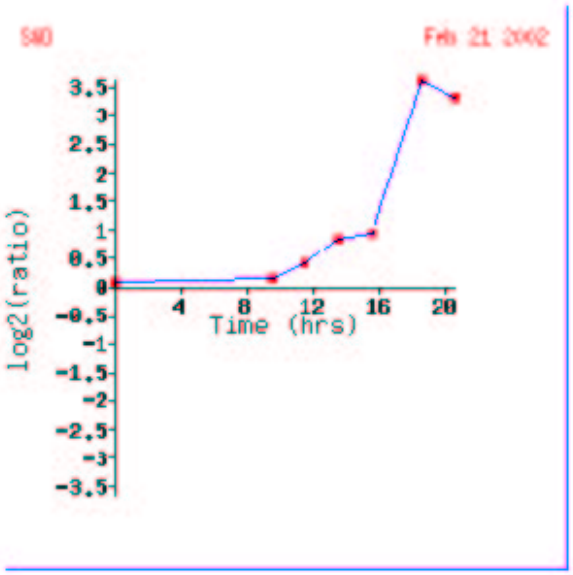
Coregulation networks

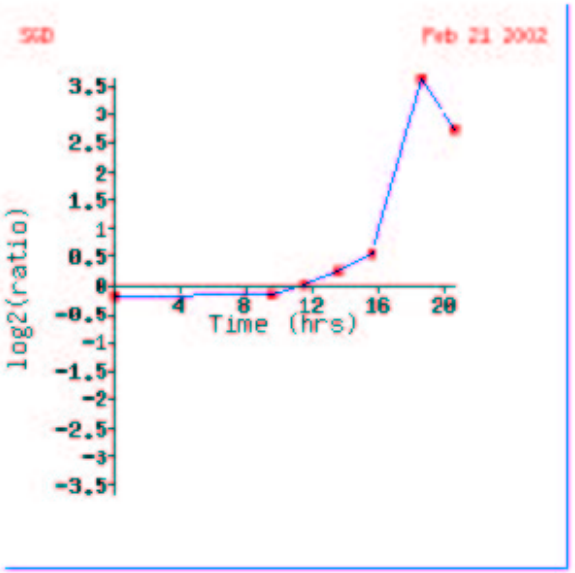
Construct a network whose nodes are the genes, and a link is formed between two genes whenever they are coregulated (e.g. the correlation coefficient between their expression profiles is higher than a certain cutoff).

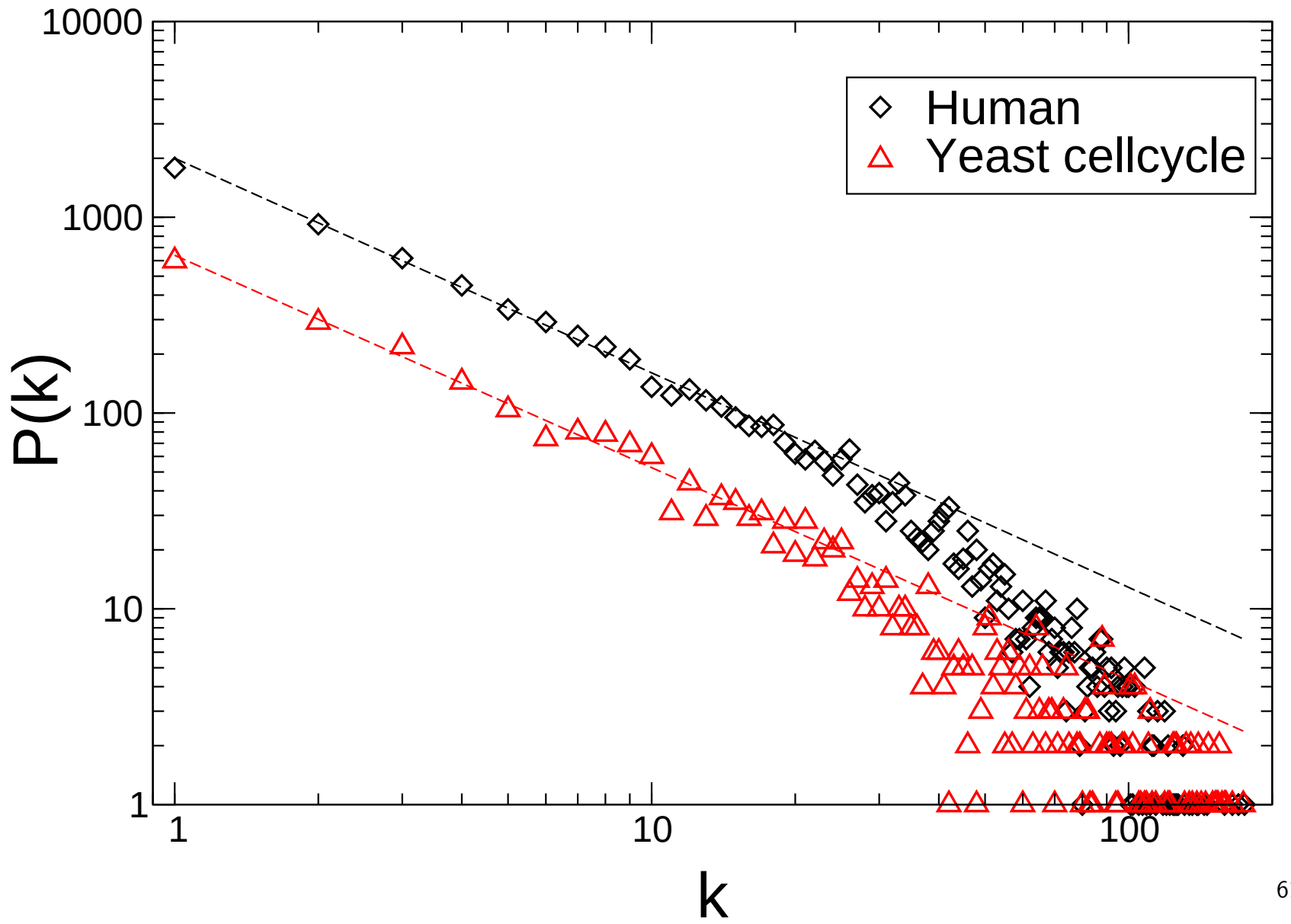
Coregulation: two genes are coregulated if their expression profiles are similar (as measured e.g. by linear correlation coefficient).


SBD
Feb 21 2002
log2(ratio)
Time (hrs)


SBD
Feb 21 2002
log2(ratio)
Time (hrs)


SBD
Feb 21 2002
log2(ratio)
Time (hrs)


SBD
Feb 21 2002
log2(ratio)
Time (hrs)


10000
1000
100
10
1
P(k)
1
10
100
k
Human
Yeast cellcycle

Also in this network a strong correlation between centrality and lethality can be found.
Example: of all genes in the network, $\sim 18\%$ are essential. Of the 138 genes with 60 or more neighbors, 84 ($\sim 61\%$) are essential.

However a remarkable difference with respect of the protein network is that the coregulation network is strongly *assortative*: high connectivity nodes are connected to each other, so that eliminating a central node does not disrupt the topology of the network.

The hubs form a highly connected core of "central genes"

Central genes

An explanation can be found by biological analysis of the central genes in the coregulation networks:

- Being central is a property of a set of genes, largely independent of the experimental conditions.
- Central genes are involved in basic cellular functions (e.g. ribosome biogenesis and protein synthesis).
- Central genes are evolutionarily conserved: most central genes are common to yeast and human

The analysis of the coregulation network allowed us to identify a tightly coregulated core of essential and evolutionarily conserved genes, simply by looking at the connectivity distribution.