Graph theory analysis of genomics problems: community analysis of fragile sites correlations and of pseudogenes alignements

Angela Re[a], Ivan Molineris[a], Michele Caselle[a,b]

[a] *Department of Theoretical Physics of the University of Torino and I.N.F.N., Via Pietro Giuria 1, I-10125 Torino, Italy*

[b] *Interdepartmental Centre on Complex Systems in Biology and Molecular Medicine, Via Accademia Albertina 13, I-10125 Torino, Italy*

e–mail: angelare@to.infn.it, molineri@to.infn.it, caselle@to.infn.it

Abstract

Graph theory offers the ideal framework to model biological systemic properties. Recently these methods were succesfully applied in proteomics and in the study of metabolic networks. In this paper we want to show that these same tools are equally powerful also to address genomic problems, like alignement networks or the networks obtained by looking at suitable correlators of chromosomic features. We shall in particular address two examples. In the first example we shall study human common fragile sites (CFS), a class of "hypersensitive" segments of DNA. The interest to CFS is motivated by their largely debated role in cancerogenesis. In order to functionally characterize them we developed a novel genome-wide approach based on graph theory and Gene Ontology vocabulary. We obtain a few non trivial results fitting with largely accepted knowledges and a more recently advanced proposal about the role of CFS in tumor cell biology. The second application is a preliminary work on a potential new type of transcriptional regulatory mechanism. It involves pseudogenes which are non-functional copies of genes. This mechanism should imply similarity between the upstream sequences of genes and pseudogenes. We constructed the upstream similarity network in the budding yeast S.Cerevisiae. Network properties suggest that pseudogenes-mediated regulation could be a common feature in eukariotic organisms.

Keyword: fragile site, pseudogene, functional annotation, gene regulation, graph theory

1 Introduction

In the past few years we have seen an increasing interest for the so called "systemic" approach to biological problems. At the basis of the systemic approach is the idea that it is only by looking at the network of interactions of a living system as a whole that one may hope to understand the functional role of its various components.

One of the main mathematical tools to perform this type of analysis is graph theory, and indeed we saw in these last few years an impressive progress also in this direction, with a lot of new results in graph theory and, as a consequence, in our theoretical understanding of complex networks.

In this contribution we want to discuss two distinct examples of applications of graph theory to complex biological problems which are exactly along this line. Our networks refer to two very different biological problems in two different organisms but both can be modelled in terms of complex networks based on similarity measures. In both cases a careful theoretical analysis (and in particular the identification of the underlying community structure of the network) allows to obtain a few non trivial result and to guess which are the biological mechanisms which shape the networks in which we are interested.

The first application focuses on human common fragile sites. They are "hyper-sensitive" segments of DNA, they are said to be "expressed" when they appear as gaps or breaks on chromosomes. Despite long efforts, the understanding of the mechanisms of their instability and their functional characterization are still largely incomplete [1, 2]. Here we ask if the "similarity" (as defined below) observed among fragile sites patterns of expression implies functional interactions among the genes that are contained in fragile sites. We find that such genes tend to be specialized in function and we speculate that their co-regulation could contribute to the correlated expression patterns of fragile sites.

The second application studies a potential new class of regulatory mechanisms at the level of the transcription process in the budding yeast S. Cerevisiae. According to this hypothesis, pseudogenes would act as regulators of their corresponding coding mRNAs. A few experimental evidences of such a mechanism do exist [3, 4]. Here we carry on a large-scale sequence analysis to quantify the statistical significance of suitable features that should underlie the action of this mechanism. Positive results of our study provide actual support for a new model of transcriptional fine tuning guided by psudogenes. We suggest that it could explain observed pseudogenes' deviations from the neutral evolution model.

This paper is organized as follows. After a short introduction on graph theory (section 2) we shall discuss the application of these methods to the study of common fragile sites (section 3) and of a possible regulatory role of pseudogenes (section 4).

2 Graph theoretical background

The aim of this section is to give a short account on a few simple tools which turns out to be of great importance in the analysis of biological problems. It is important to stress that we shall discuss only a very small portion of the impressive amount of results which have been obtained in this sector in these last few years. For a more complete and detailed account of these results and for updated reviews on graph theory we refer the reader to [5,6].

We shall discuss here two main classes of observables: those related to the properties of vertices and those related to the community analysis of the graph.

2.1 Vertex properties

All along the paper we shall use the well know Erdos-Renyi random graph model as "reference model", i.e. as the "null hypothesis" with which we shall compare our findings. The idea underlying our whole analysis is that departure from the predictions of the Erdos-Renyi random graph model should indicate a potential biological relevance of the observable under study. For this reason we shall close this section with a brief summary of known properties of random graphs.

2.1.1 Degree

If one is interested to discuss the properties of the vertices of a graph the first observable one must address is the degree of a vertex which is the number of links connected to such vertex in a network. The degree distribution of a graph is a powerful tool to organize graphs into families with different properties: (power-like graphs versus exponential graphs). We shall denote in the following as z_i the degree of the vertex i and as z the mean degree. As we shall see below the probability of finding a vertex of degree k in an Erdos-Renyi random graph is given by a Poisson distribution.

2.1.2 Betweenness

A more sophisticated indicator of the properties of a vertex is betweenness. Betweenness is a measure of the extent to which a node lies on the paths between others. Following the standard definition, we define the betweenness of a node i as the fraction of shortest paths between pairs of nodes in the network that pass through i. This quantity is interesting also because it allows to estimate the so called "centrality" of the vertex. Vertices with high centrality are expected to play a more important role with respect to the remaining vertices in the life of the network.

2.1.3 Clustering coefficient

The property of clustering (which is also sometimes called *network transitivity*) is one of the most powerful tools to identify non random features in biological networks. It can be measured using the clustering coefficient C. It is essentially the mean probability that two vertices that are network neighbours of the same other vertex are also neighbours. In an Erdos-Renyi random graph C can be easily evaluated (for more details see next section). High values of the ratio between the clustering coefficient that we find and the Erdos-Renyi one would mean strong tendency of vertex to cluster among them.

2.2 Comparison with the random graph hypothesis

The Erdos-Renyi random graph is the simplest possible model for a network. It depends on two only parameters: the number of vertices n and the probability p of connecting two vertices with an edge. Actually this model describes not a single graph but an ensemble (in the sense of statistical mechanics) of graphs in which a graph with exactly n vertices and m edges appears with probability $p^m(1-p)^{M-m}$ where $M = \frac{n(n-1)}{2}$ is the number of pairs of vertices of the graph (and hence the maximum possible number of edges). The most important feature of the model is the presence at a particular value of p of a phase transition called percolation transition in which suddenly a giant connected component appears in the graph. This transition occurs exactly at $z = 1$ (where z is the mean degree of the graph and is given by $z = p(n-1)$. The appearance of a giant connected component at z far below the percolation threshold is a highly non trivial result.

Another important feature of random graphs is that, due to their simplicity, is rather easy to evaluate a number of important graph theoretical quantities. In our analysis we use the aforementioned probability of a vertex having a degree k, $p_k = \binom{n}{k} p^k (1-p)^{n-k} \cong \frac{z^k e^{-z}}{k!}$ and the mean clustering coefficient which (for an undirected graph) is defined as $\langle C \rangle = \frac{\sum_{i=1}^{n} C_i}{n}$ where $C_i = \frac{2|\{e_{jk}\}|}{K_i(K_i-1)}$ where e_{ij} denotes an edge between vertices v_k and v_j which are among the nearest neighbours of the vertex v_i (degree K_i)

2.3 Community structure analysis

2.3.1 Connected components

The very first step of any graph theoretical analysis of a network is the reconstruction of its connected components. We extracted such connected components by using the standard Hoshen-Kopelman algorithm [7]. However it is by know well understood that inside a large enough connected component of a graph there may be a highly non trivial organization in so called "communities". Roughly speaking a community is a subgraph of the network with a large number of interconnections among its vertices and a rather small amount of links joining it with the remaining part of the graph.

2.3.2 The Newman Algorithm

To reconstruct the community structures of the networks that we shall study we applied the agglomerative hierarchical clustering algorithm proposed by Newman [8]. The starting step of the algorithm is the extreme structure in which each vertex is isolated. Then the algorithms proceeds by joining communities together in pairs if as a result of this fusion there is an increase in the modularity coefficient Q (see next section for the exact definition). The best partition of the network in communities corresponds to the maximal value of Q

2.3.3 Validation of the community structure

A powerful tool to test if a particular partition in communities is meaningful or not is the so called "modularity coefficient" $Q = \sum_i (e_{ij} - a_i^2)$ where e_ij is the fraction of edges in the network that connect vertices of the community i with those of the community j and $a_i = \sum_j e_{ij}$. Roughly speaking Q measures the fraction of edges which lie within the community minus the expected value for the same quantity in a random graph, thus for a random graph $Q = 0$ while larger values of Q indicate a significant departure from a random distribution of the edges. In practice already values of $Q \geq 3$ indicate a well defined community structure in the network.

3 Common fragile sites in a systemic perspective

Common fragile sites (CFS) are peculiar regions of DNA showing a high rate of breakage and/or recombination events. Such events imply both intracellular DNA exchange and external DNA viral integration. CFS are said to be "expressed" when they show one of the above mentioned events. These regions are termed "common" since they exist in almost all the individuals, hence they do not denote by themselves a pathological status of the cell. They have been studied mainly in human and mouse [9], but are expected to exist in all higher eukariotes. There are evidences that these CFS are conserved by evolution (at least as far as human-mouse comparison is concerned) and thus it is likely that they have some important functional role which however has yet to be understood.

Recently a lot of interest has been attracted by these CFS in view of a possible non trivial relationship between their expression and tumour development [10]. The main open issue is if CFS have a positive or negative role in tumour development: one would like to understand if tumour benefits from fragile site instability or if instead fragile sites act as "sensors" to elicit, by altered expression of their genes, cellular response against hazards at preliminary stages.

To address the intriguing issue a deeper understanding of the cellular function of CFSs is needed. Motivated by a few recent discoveries [11] about the correlation between two frequently expressed fragile sites, we decided to extend such a correlation analysis to a genome wide scale. To understand the relevant patterns of correlations on such a large scale a graph theoretical analysis of the network of correlated CFS turned out to be mandatory. We performed our analysis in three steps:

- we constructed the network of co-expressed CFS,
- we isolated the relevant communities inside this network,
- we looked for possible functional correlations among the genes insides the communities using Gene Ontology [12].

Let us discuss in more detail these three points:

3.1 The network of co-expressed CFS

For each pair of fragile sites we studied the linear correlation coefficient of their expression patterns and selected only those pairs with a correlation higher than a given threshold. We set three thresholds; they correspond to

α	Q
0.1%	0.573
1%	0,461
5%	0.359

Table 1: Modularity coefficient Q at the significance level for fragile site correlation set to $\alpha = 0.1\%$, $\alpha = 1\%$ and $\alpha = 5\%$.

correlators which respectively have a (Bonferroni corrected) probability of 0.1%, 1% and 5% to appear by chance.

The data which we used for our analysis are the expression patterns of 137 fragile sites on a sample of 60 subjects reported in [11]. Raw data and experimental procedures to gather them are described in detail in [13] to which we refer the interested reader.

Co-expression data were represented as a network where nodes stand for fragile sites and links between couples of nodes are added if such fragile sites exhibit a significant correlation coefficient. Networks at different thresholds are reported in Fig 1.

3.2 Community analysis

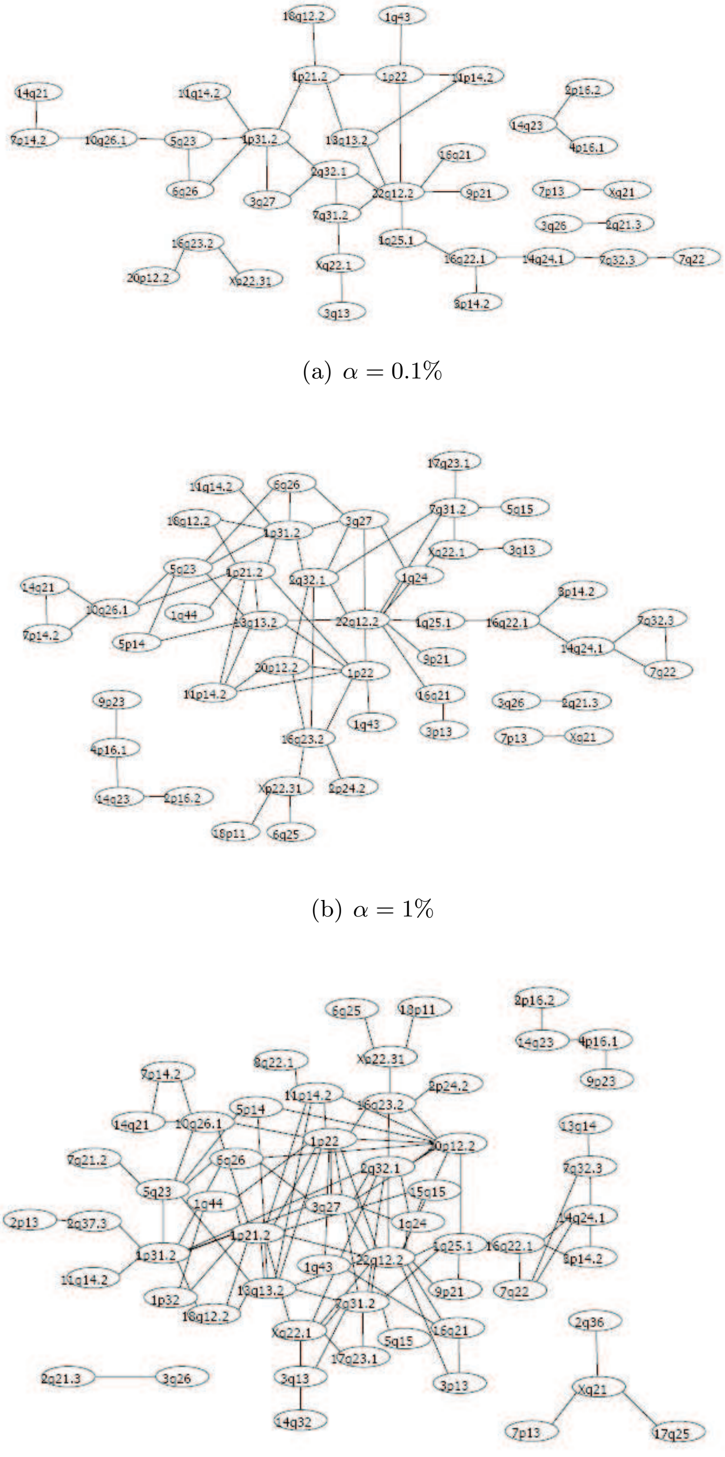
We then measured the three vertex observables discussed in 2: degree, betweenness and clustering coefficient. The most remarkable result of this analysis was that in all our graphs (i.e. at all the thresholds) the values for the clustering coefficient values were much higher (about 30 times) than the Erdos-Renyi ones.

This result prompted us to analyse the community organization of the giant connected component in all three networks. High Q values quantify the tendency of the three networks to be divided into two communities. Q values are listed in Tab 1.

These findings strongly suggest that the co-expression networks should hint to some kind of functional interactions among the genes located at correlated fragile sites.

3.3 Functional analysis using Gene Ontology

Functional analysis was performed using the Gene Ontology database. Gene Ontology (GO) [12] provides a dynamic and controlled annotation framework for describing gene products. GO (http://www.geneontology.org/,



(a) $\alpha = 0.1\%$

(b) $\alpha = 1\%$

(c) $\alpha = 1\%$

Figure 1: Visualization of the network based on correlated expression patterns for fragile sites.

version 3.1191) includes three extensive subontologies describing molecular function (the biochemical activity of a gene product), biological process (the biological goal a gene product contributes to) and cellular component (the cellular place where the biological activity of a gene product is exerted). Individual terms are organized as a directed acyclic graph, in which the terms form the nodes in the ontology and the arcs the relationships. Descendent terms are related to their parent terms by "is-a" relationships or "part-of" relationships. In contrast to simpler hierarchical structures, one node in a directed acyclic graph may have multiple parents. This allows for a more flexible and detailed description of biological functions.

We used GO to give a functional meaning to our communities. More precisely we collected the sets of genes mapped to fragile sites belonging to the connected components and their communities (at all thresholds) and looked for categories of biological process and molecular function defined in GO which were significantly enriched in these sets. We performed an exact Fisher's test to check whether the term appeared in the set significantly more often than expected by chance. The full list of genes associated to the few reliable GO terms at the highest threshold is provided in Tab 1.

3.4 Results

The most comprehensive GO function including 34 genes located at 10 out of 27 connected fragile sites turned out to be "cytokine activity". Cytokines act as mediators of innate and adaptive immune responses by controlling cell growth and division. As a result of our analysis we suggest that correlated expression at fragile sites may derive from a co-regulated expression of their genes. The alterations constantly observed at or near these genes would be produced by cellular processes connected with their co-regulation [14, 15]. In this respect it is interesting to notice that immune gene expression has been recently shown to be epigenetically regulated [16].

A second interesting result is that a surprising high proportion of genes at correlated fragile sites are implicated in cancer. According to a challenging viewpoint, fragile site expression may protect against cancer at early stages [17–19]. Genomic integrity would be ensured by the aberrations occurring at fragile genes that would act as sensors to elicit cell-cycle arrest or death. We believe that fragile sites are not located by chance within or near our highlighted genes, but take part with these genes to the mechanism that regulate the cellular response to DNA damage [20]. This proposal was based on some known genes located in proximity of highly expressed fragile sites such as STS at FRAXB and Wwox at FRA16D.Remarkably enough we found that these genes were connected together in one of our communities thus further supporting the idea of a common interaction among them.

4 Upstream similarity network in yeast

Pseudogenes are defined as DNA sequences of former functional genes made nonfunctional by severe mutations. Operationally, pseudogenes are usually identified by their disrupted open reading frames (ORFs), which are homologous to functional genes. Since some pseudogenes exhibit features suggesting a non-neutral molecular evolution, it is plausible that, at least some of the pseudogene, have some still unknown functional role [21]; moreover a specific molecular function for a pseudogene has been found in some cases. S.A. Korneev et al. have shown that neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from a NOS pseudogene in Lymnaea stagnalis [3]. Hirotsune et al. [4] have found that the expression of the *Makorin1* gene in Mus Musculus is controlled by one of its pseudogene copies, *Makorin1-p1.* Even though it is not completely clear how this regulative interaction is exploited, the authors experimentally demonstrated that in this process the first 700 base pairs of mRNA of the pseudogene, which are very similar to those of the gene, play a fundamental role. Fig 2 shows two ways in which this might happen.

The upstream similarity network could be a powerful tool to perform a genome wide analysis of regulative interactions like that described for Makorin1 and Makorin1-p1. There are two possible reasons for this:

- the sequences that we study are upstream of the translation starting site, thus they include the 5'UTR region of the gene and if the mechanisms discussed in Fig 2 are correct they require a high degree of similarity between the 5'UTR regions in competition;

- if a gene and a pseudogene share some kind of regulatory interaction they should also be themselves coregulated, in order to be simultaneously expressed. Thus it is likely that they share a common regulatory pattern in their promoter regions (which is also included in the upstream sequences which we select).

We chose the well known and relatively simple genome of Saccharomyces Cerevisiae for our preliminary work. The subject of our study is a network whose edges consist of all the pairs of genes which present an upstream similarity above a given significance cut-off. Each entry annotated in the Saccha-

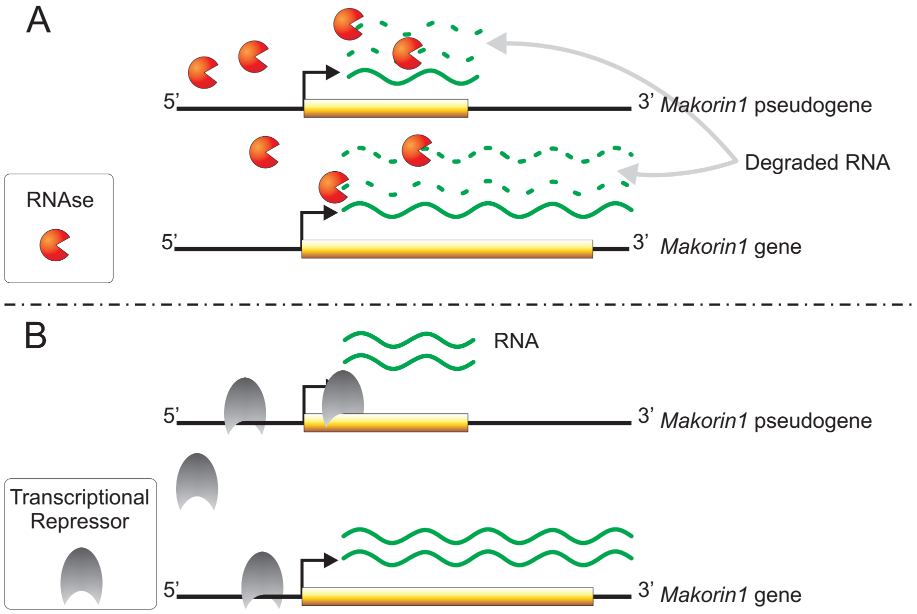


Figure 2: Plausible mechanism of gene-pseudogene interaction. **A**) A RNA-mediated mechanism: here, messenger RNA copies of the pseudogene and gene compete for a destabilizing protein that binds a crucial 700-nucleotide region near the beginning of the mRNAs. This destabilizing protein might be an RNA-digesting enzyme (RNAse). **B**) A DNA-mediated mechanism: here, regulatory elements of the pseudogene and gene, located in the same region as above, compete for transcriptional repressors.

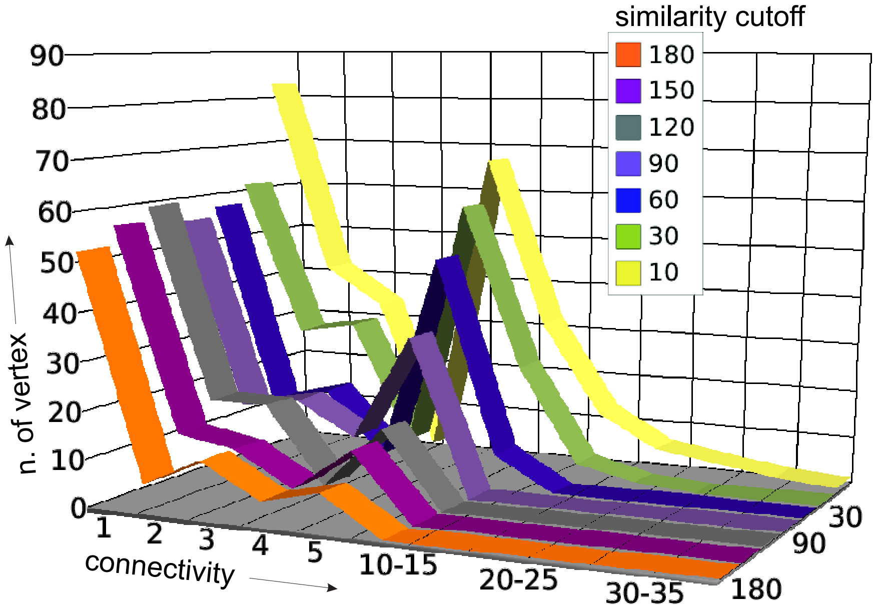


Figure 3: Connectivity distribution at different similarity cutoff w_c.

romyces Genome Database (SGD) [22] as open reading frame or pseudogene is as vertex of the network. For all of the 6612 vertices we selected 500 bases upstream of translation starting site (we call this sequence "upstream"). In case of superpositions with other ORF's, we cut the sequence so that only the non-coding nucleotides between the two ORF were included. We aligned every pair of sequences obtained in this way using NCBI-BLAST [23] and defined the similarity between two sequences as the opposite of the base 10 logarithm of the e-value supplied by the program for their best local alignment. For partially overlapping upstreams we included the alignment in the graph only if the overlapping portion did not contribute significantly to the alignment score.

We put an edge e_w between each couple of vertex (v_1, v_2); the weight w of this edge is the upstream similarity of the two vertex v_1 and v_2. Therefore we consider a set of unweighted graphs, each characterized by a given cutoff w_c, in which each edge e_w survives only if its weight w is greater of w_c.

As one can easily expect the number of vertices n is a decreasing function of w_c. For instance we have: $n = 501$ for $w_c = 5$, $n = 287$ for $w_c = 10$ and $n = 133$ for $w_c = 90$; the medium connectivity z varies in the same way from $z = 12,8$ for $w_c = 5$ to $z = 5.7$ for $w_c = 10$ and $z = 2.8$ for $w_c = 90$.

comunity label	community size	go term	P-value
ERR	3	phosphopyruvate hydrase activity	6.26E-04
COS	8	storage vacuole litic vacuole	1.07E-02 1.07E-02
ASP	4	asparagine activity cellular response to nitrogen starvation	9.29E-07 9.29E-07
THI	3	thiamin biosyntesis	1.88E-07

Table 2: Overrepresented GO terms in graph comunity.

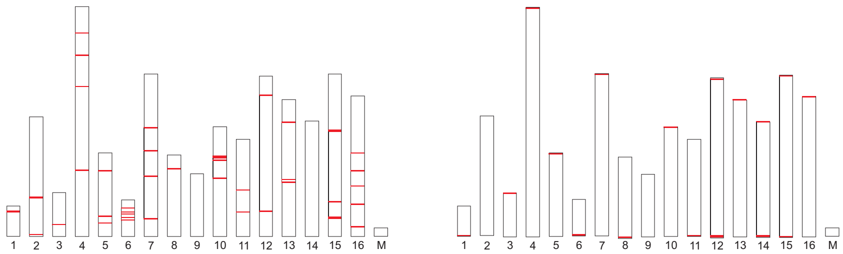
As shown in Fig 3, the connectivity distribution does not correspond to that predicted by standard random graph theory or by scale-free models [24]. This is mainly due to the presence of peaks in the distribution with high connectivity given by the presence of subgroups of highly interconnected vertices. Because of the small size of the network, further considerations about this fact cannot be made, but we hope to obtain statistical evidences for this observation by analysing the same type of upstream similarity network for the mammalians genomes.

The graph with cutoff $w_c = 10$ presents 54 connected components, one of these is very populated (77 vertices) and 15 have size bigger than 4; the giant component is made by 3 groups of vertices which may be immediately identified as distinct "communities" since they present a large number of inside connections while are weakly connected with the remaining vertices [8]. These communities are splitted in distinct connected components in the graph with cutoff $w_c = 20$, in this case however their size is smaller.

Analyzing the Gene Ontology annotations of genes belonging to the same community or component we observed a significant enrichment of similar functional annotations (some example in Table 2). This is not strange since most of the communities of the graph roughly coincide with known families of genes.

The genes related to the 30% of the communities with size bigger than 4 are placed at the extremity of the chromosomes (some example in Fig 4) suggesting that genes with highly similar upstreams can be produced by events of duplication in telomerics zones.

The set of vertices of the network with cutoff 10 includes 8% of consid-



(a) Dispersed community (b) Subtelomeric community

Figure 4: Chromosome localization of ORF (red hyphen) with highly similar upstream sequences.

ered sequences, still in the same set appears the 40% of the pseudogenes (6 pseudogenes annotated as so in SGD out of 14 present in the set of the upstream sequences at the beginning). In the same way the set of selected genes is enriched of dubious ORF's and spurious sequences (that have codon compositions not characteristic of genuine genes and did not yield detectable protein products [25]); some of these genes could indeed be yet unrecognized expressed pseudogenes.

As a negative test we also constructed, following the same procedure outlined above, the similarity graph of the coding portions of genes. In this second case we found a definitely smaller fraction of pseudogenes and spurious or dubious ORF's in the graph. This fact could indicate the presence of an evolutive pressure which favours the similarity between the upstream sequence of the pseudogene and that of its relative gene. This signature is compatible with the regulative mechanism of the Makorin1 and Makorin1-p1 pair discussed above and could suggest a wider presence of this type of regulation even in organisms as simple as yeast. We are presently extending this analysis to other eukaryotes and in particular to vertebrates in order to give a more reliable statistical basis to the above observation.

Acknowledgement:

The authors thank Ferdinando Di Cunto, Paolo Provero, Isabella Sbrana and Davide Corà for useful discussions. This work is partially supported by the Fund for Investiments of Basic Research (FIRB) from the Italian Ministry of the University and Scientific Research, No. RBNE03B8KK-006.

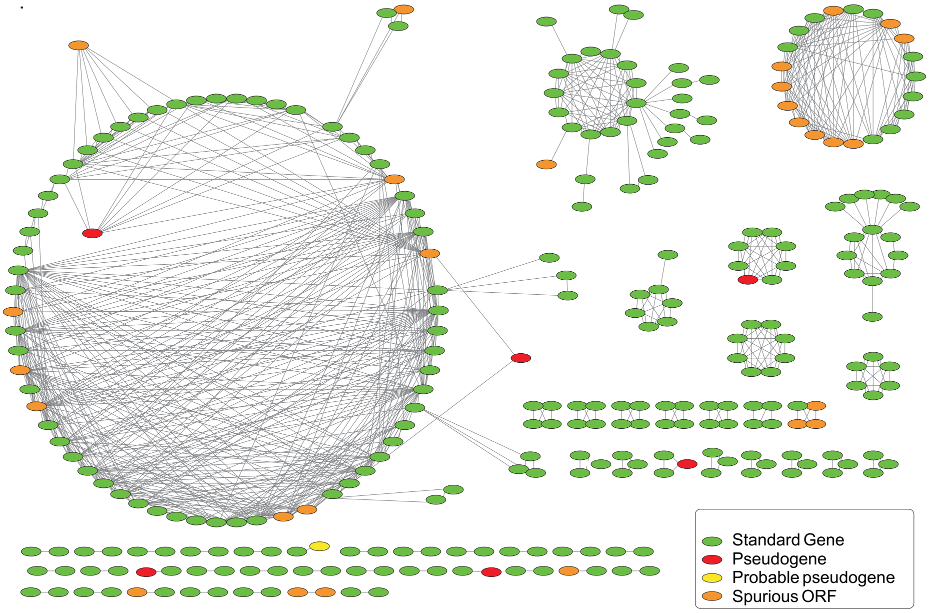


Figure 5: Visual representation of the upstream similarity network

References

[1] T. Glover, Common fragile sites., Cancer Lett. 232 (2006) 4–12.

[2] S. Corbin, M. Neilly, R. Espinosa, E. Davis, T. Mckeithan, M. L. Beau, Identification of unstable sequences within the common fragile site at 3p14.2: implications for the mechanism of deletions within fragile histidine triad gene/common fragile site at 3p14.2 in tumors., Cancer Res. 62 (2002) 3477–8.

[3] S. Korneev, J. Park, M. O'Shea, Neuronal expression of neutral nitric oxide synthase (nnos) protein in suppressed by an antisenserna transcribed from an nos pseudogene., J. Neurosci. 19 (1999) 7711–7720.

[4] S. Hirotsune, N. Yoshida, A. Chen, L. Garrett, F. Sugiyama, S. Takahashi, K. ichi Yagami, A. Wynshaw-Boris, A. Yoshiki, An expressed pseudogene regulates the messenger-rna stability of its homologous coding gene, Nature 423 (2003) 91–96.

[5] M. Girvan, M. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (2002) 7821–7826.

[6] O. Mason, M. Verwoerd, Graph theory and networks in biology, arXiv:q-bio.MN 0604006.

[7] J. Hoshen, R. Kopelman, Percolation and cluster distribution. cluster multiple labeling technique and critical concentration algorithm, Phys. Rev. B (1976) 3438–3445.

[8] M. Newman, Fast algorithm for detecting community structure in networks, Physical Review E (69).

[9] A. Matsuyama, T. Shiraishi, F. Trapasso, T. Kuroki, H. Alder, M. Mori, K. Huebner, C. Croce, Fragile site orthologs fhit/fra3b and fhit/fra14a2: evolutionarily conserved but highly recombinogenic, Proc Natl Acad Sci USA 100 (2003) 14988–93.

[10] D. Liopoulos, G. Guler, S. Han, T. Druck, M. Ottey, K. McCorkell, K. Huebner, Roles of fhit and wwox fragile genes in cancer, Cancer Lett. 232 (2006) 27–36.

[11] I. Sbrana, F. Veroni, M. Nieri, A. Puliti, R. Barale, Chromosomal fragile sites fra3b and fra16d show correlated expression and association with failure of apoptosis in lymphocites from patients with thyroid cancer, Genes Chromosomes Cancer 45 (2006) 429–36.

[12] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, R. Apweiler, The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology, Nucl. Acids Res. 32 (2004) 262–266.

[13] A. Re, D. Corà, A. Puliti, M. Caselle, I. Sbrana, Correlated fragile site expression allows the identification of candidate fragile genes involved in immunity and associated with carcinogenesis, BMC Bioinformatics.

[14] T. Ito, R. Amakawa, M. Inaba, T. Hori, M. Ota, K. Nakamura, M. Takebayashi, M. Miyaji, T. Yoshimura, K. Inaba, S. Fukuara, Plasmacytoid dendritic cells regulate the cell responses through ox40 ligand and type 1 infs, J Immunol. 172 (2004) 4253–9.

[15] G. Loots, R. Locksley, C. Blankespoor, Z. Wang, W. Miller, E. Rubin, K. Frazer, Identification of a coordinate regulator of interleukins 4, 13 and 5 by cross-species sequence comparisons, Science 288 (2000) 136–40.

[16] S. Reiner, Epigenetic control in the immune response, Hum Mol Genet. Spec 1 (2005) 41–6.

[17] L. O'Keefe, R. Richards, Common chromosomal fragile sites and cancer: focus on fra16d, Cancer Lett. 232 (2006) 37–47.

[18] J. Bartkova, Z. Horejsi, K. Koed, A. Kramer, F. Tort, K. Zieger, P. Guldberg, M. Sehested, J. M. Nesland, C. Lukas, T. Orntoft, J.Lukas, J. Bartek, Dna damage response as a candidate anti-cancer barrier in early human tumorigenesis, Nature 434 (2005) 864–70.

[19] P. Hoglund, Dna damage and tumor surveillance: one trigger for two pathways., Sci STKE 317.

[20] S. Gasser, S. Orsulic, E. Brown, D. Raulet, The dna damage pathway regulates innate immune system ligands of the nkg2d receptor, Nature 436 (2005) 1186–90.

[21] O. Podlaha, J. Zhang, Nonneutral evolution of the transcribed pseudogene makorin1-p1 in mice, Mol Biol Evol 21 (2004) 2202–2209.

[22] J. Cherry, C. Adler, C. Ball, S. Chervitz, S. Dwight, E. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, D. Botstein, SGD: Saccharomyces Genome Database, Nucl. Acids Res. 26 (1998) 73–79.

[23] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucl. Acids Res. 25 (1997) 3389–3402.

[24] R. Albert, Scale-free networks in cell biology, J Cell Sci 118 (2005) 4947–4957.

[25] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, J. S. Weissman, Global analysis of protein expression in yeast, Nature 425 (2003) 737–741.