IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10366-y

Colloquia: CSFI 2008

Network approaches to Genome-Wide Association studies

D. Remondini[1](*), E. Verondini[1], F. Lescai[3], I. Zironi[1]
and G. Castellani[2]

[1] *Dipartimento di Fisica, Università di Bologna and INFN, Sezione di Bologna - Italy*
[2] *DIMORFIPA, Università di Bologna and INFN, Sezione di Bologna - Bologna, Italy*
[3] *Dipartimento di Patologia Sperimentale, Università di Bologna - Bologna, Italy*

(ricevuto il 15 Maggio 2009; pubblicato online il 3 Agosto 2009)

Summary. — In the framework of large-scale genotypic studies (describing the distribution of allele frequencies inside human genome) we characterize the Linkage Disequilibrium (LD) matrix as a network of relationships between alleles. We propose a suitable matrix discretization threshold, after a characterization of the distribution of noisy values inside LD matrix. We compare the main network parameters of a real LD matrix with two null models (Erdos-Renyi random network and a rewiring of the original network), in order to highlight the peculiar features of the LD network. We conclude stating the need of adequate computing tools for handling the high-dimensional data coming from Genome-Wide genotyping datasets.

PACS 87.18.Vf – Systems biology.
PACS 87.18.Wd – Genomics.
PACS 87.19.xk – Genetic diseases.

1. – Description

In the framework of genotypic studies, a mainstream topic is the quantitative characterization of the relative allele frequencies inside genome. These studies have been applied to characterize the evolution of human populations [1, 2], by looking at how patterns of allele frequencies are distributed geographically. More recently, these studies have been extended to the characterization of whole human genomes over large amounts of samples, characterizing specific populations (*e.g.*, Africans or Caucasians, see `www.hapmap.org`) or groups with specific stratification (*e.g.*, people sharing the same pathology [3]). These data are allowing the study of the so-called *complex trait diseases*, in which the pathology (and its degree of severity) is a phenotypic trait due to a combination of several genetic factors (possibly hidden inside single-nucleotide mutations) that are not harmful

(*) E-mail: `daniel.remondini@unibo.it`

if found singularly. The challenge is open, but a great effort must be undertaken, both from an experimental point of view, and also for the mathematical and informatics tools necessary for the storage and treatment of the huge amounts of high-dimensional data obtained from such experiments.

The topic of Genome-Wide Association studies is referred to the characterization of specific allelic profiles that can be associated to a particular pathology, in comparison with a *background* allelic frequency obtained from healthy samples, considering a sampling over the whole chromosomic set. Such studies produce data with a dimensionality in the order of $4 \cdot 10^5$ (corresponding to about 20 single-nucleotide samplings per gene) or larger. These studies concentrate on Single-Nucleotide Polymorphisms (SNPs), that may lead to changes in the aminoacidic sequence in the proteins encoded in genes, or to changes in binding affinity for transcription factors in the non-coding regions surrounding genes, thus altering genes functionality to a different extent.

A typical measurement of deviations of SNP frequency from random recombination is Linkage Disequilibrium (D'):

$$D'_{AB} = \frac{P_A B - P_A \cdot P_B}{D_{\max}}\,, \tag{1}$$

in which A and B are two (of the possible four) alleles at two different chromosomal locations. It is a measurement of co-occurence of SNPs in couples, in which $D_{\max}$ is a normalization constant that sets the maximum/minimum LD values to $+1/-1$. The sign is usually not important, the information is encoded in significant deviations from zero. In general, for statistical purposes, only high values of LD are considered ($|\mathrm{LD}| \simeq 1$), and low values ($|\mathrm{LD}| < 0.25$) are considered as *noise*, leaving a large *gray zone* unexplored.

We aim at characterizing LD matrices with a network-based approach, applying a suitable discretization that removes the noisy (unreliable) values. Considering a low-dimensional sample dataset (number of nodes $N \simeq 1000$ SNPs), the most common network parameters are considered, and the relative distributions are compared with two null models, a random network (in the Erdos-Renyi sense) with the same average connectivity as the original matrix, and a matrix obtained by original link reshuffling, that instead preserves single-node connectivity degree [4].

2. – Analysis

As a first step, we estimate the region of noisy values (to be removed) by looking at the probability density of the LD coefficients. We consider the range of values ≤ 0.3, in which the distribution is clearly Gaussian (figure not shown), for estimating the noise parameters (essentially σ). Thus, we obtain a (symmetric unweighted) adjacency matrix A by setting to zero the LD matrix values below $3\sigma = 0.385$, and setting to 1 the remaining. The main network parameters (connectivity degree K, betweenness centrality BC, clustering coefficient C, average nearest-neighbour connectivity K_{NN}) for each node are calculated, and the relative probability density functions are compared with two null models: a random network R with the same average connectivity of A, and a more structured matrix, the *rewired network* W, obtained by randomly reshuffling the links in A, thus preserving single-node connectivity degree.

Concerning network parameters, average clustering results much higher in A and W ($C_A = 0.415$ and $C_W = 0.4$, respectively) as compared to the completely randomized counterpart ($C_R = 0.1563$, as expected from random graph theory [5]: $C_{ER} = \langle K \rangle / N =$

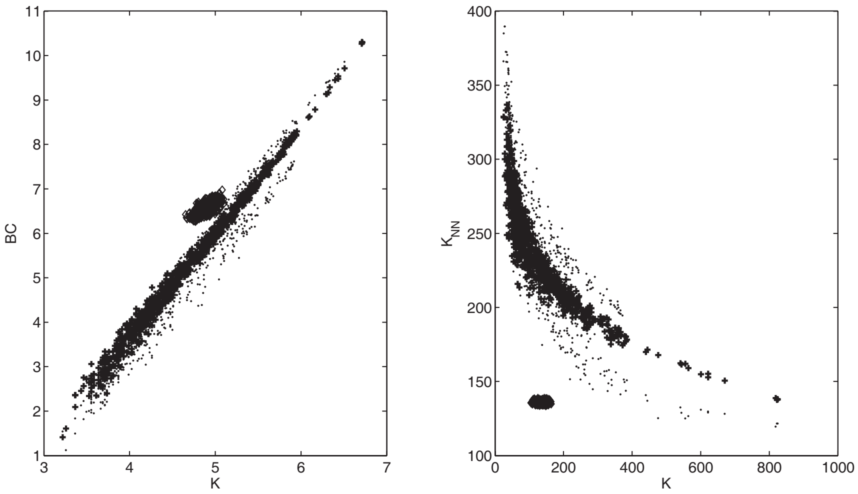

Fig. 1. – Plot of BC *vs.* K (left panel) and K_{NN} *vs.* K (right panel). Full circles: A; crosses: W. The "small clouds" in both plots (with diamonds as markers) represent the R node parameters.

0.1562 in our case), reflecting a higher level of interconnectivity between nodes. Also the BC and K_{NN} values (plotted in fig. 1 *vs.* K values) are more widely distributed as compared to R network values: the real network is more heterogeneous, reflecting a higher level of *hierarchy* among nodes.

The LD network peculiar structure appears more markedly in the plot of the joint distribution of K_{NN} and K (fig. 1, right). In a purely random network, nearest-neighbour connectivity is independent from node connectivity [5]; for the LD matrix the situation is that of a *dissassortative* network, meaning that few highly connected nodes (*hubs*) are more preferentially connected with less connected nodes. Disassortativity is conserved also after rewiring, and this can be explained by a probabilistic argument: since the less connected nodes are much more than the hubs, and such distribution is preserved by the rewiring algorithm used, the same situation remains the most probable to occur. Some differences can be observed nonetheless: as compared to the rewired network, there is a "branch" of nodes (in the plot) with more connected nearest neighbours and another with less connected ones. This can be due to a *stratification* of connectivity as a function of node distance, reflecting the metric structure inside the genome: close nodes in the network are referred to close SNPs in the chromosomes, thus it is more likely that they are in strong LD as compared to more distant SNPs. As can be seen in fig. 2 such structure is present in A, but is completely destroyed by rewiring in W, becoming indistinguishable from the random matrix R.

3. – Discussion: scaling up to GWA studies

In this paper we have shown the starting point for the characterization of LD matrices as networks, with the limitation of considering a small subset of elements (about a factor of 100 smaller) as compared to a whole genome study. Some of the most common network features have been studied, that allow to easily recover the main structural properties

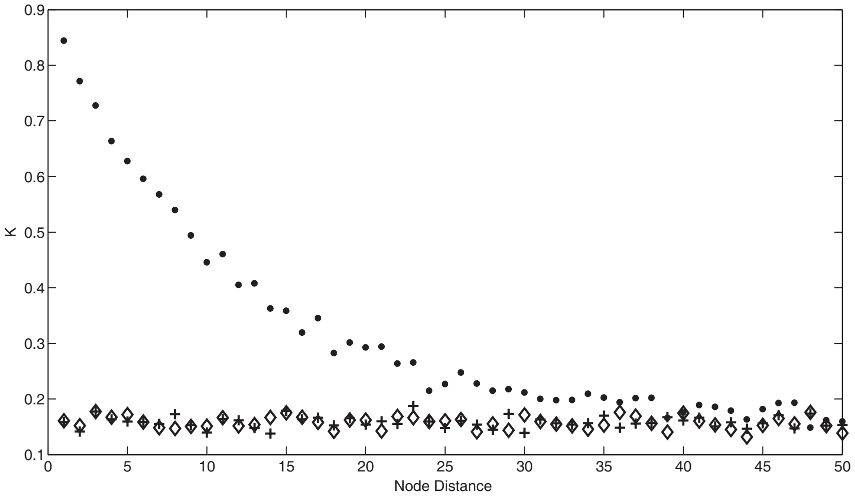

Fig. 2. – Plot of average connectivity as a function of node distance, for the first 50 neighbours. Full circles: A; crosses: W; diamonds: R.

of LD matrices, such as high heterogeneity reflecting node hierarchy, and the presence of a *metrics* due to spatial relationships of SNPs inside genes and chromosomes. The computational effort required for dealing with whole genome datasets is relevant, since most of the centrality measures require algorithms that scale with the number of nodes as N^3. Moreover, more complex algorithms that search for community structures, *i.e.* network modules of strictly connected nodes, explode exponentially with the number of nodes. Thus new strategies (or possibly new algorithms) are required, and high-speed computing over parallel architectures may represent a necessity for such a task.

* * *

DR was supported by "Progetto Strategico d'Ateneo 2006"; DR and GC by INFN SHEILA National Grant.

REFERENCES

[1] Cavalli-Sforza L. L., Barrai I. and EDWARDS A. W., *Cold Spring Harb Symp. Quant. Biol.*, **29** (1964) 9.

[2] Cavalli-Sforza L. L., *Sci. Am.*, **231** (1974) 80.

[3] Tian C. *et al.*, *PLoS ONE*, **3** (2008) e3862.

[4] Maslov S. and Sneppen K., *Science*, **296** (2002) 910.

[5] Barabasi L. and Alberts R., *Rev. Mod. Phys.*, **74** (2002) 47.