HAL open science

Google matrix analysis of bi-functional SIGNOR network of protein-protein interactions

Klaus M. Frahm, Dima Shepelyansky

To cite this version:

Klaus M. Frahm, Dima Shepelyansky. Google matrix analysis of bi-functional SIGNOR network of protein-protein interactions. Physica A: Statistical Mechanics and its Applications, 2020, 559, pp.125019. 10.1016/j.physa.2020.125019 . hal-02297365

HAL Id: hal-02297365

https://hal.science/hal-02297365v1

Submitted on 22 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0378437120305318
Manuscript_f53a491c93c8581ec5485e6b99df991e

Google matrix analysis of bi-functional SIGNOR network of protein-protein interactions

Klaus M. Frahm[a], Dima L. Shepelyansky[a]

[a] *Laboratoire de Physique Théorique, IRSAMC, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France*

Abstract

Directed protein networks with only a few thousand of nodes are rather complex and do not allow to extract easily the effective influence of one protein to another taking into account all indirect pathways via the global network. Furthermore, the different types of activation and inhibition actions between proteins provide a considerable challenge in the frame work of network analysis. At the same time these protein interactions are of crucial importance and at the heart of cellular functioning. We develop the Google matrix analysis of the protein-protein network from the open public database SIGNOR. The developed approach takes into account the bi-functional activation or inhibition nature of interactions between each pair of proteins describing it in the frame work of Ising-spin matrix transitions. We also apply a recently developed linear response theory for the Google matrix which highlights a pathway of proteins whose PageRank probabilities are most sensitive with respect to two proteins selected for the analysis. This group of proteins is analyzed by the reduced Google matrix algorithm which allows to determine the effective interactions between them due to direct and indirect pathways in the global network. We show that the dominating activation or inhibition function of each protein can be characterized by its magnetization. The results of this Google matrix analysis are presented for three examples of selected pairs of proteins. The developed methods work rapidly and efficiently even for networks with several million of nodes and can be applied to various biological networks.

KEYWORDS: PageRank, protein-protein interactions, directed networks, Ising spin

1. Introduction

Protein-protein interactions (PPI) are at the heart of information processing and signaling in cellular functions. It is natural to present and analyze these PPI by presenting them as a directed network of actions between proteins (or network nodes). The simplest case of action is activation or inhibition so that such networks can be considered as bi-functional. The development of related academic databases of PPS networks with an open public access is a challenging task with various groups working in this direction (see e.g. [1], [2], [3], [4], [5]). A typical example is the SIGNOR directed network of PPI links for about 4000 proteins of mammals and 12000 bi-functional directed links as reported by [2].

On the scale of the past twenty years, modern society has created a variety of complex communication and social networks including the World Wide Web (WWW), Facebook, Twitter, Wikipedia. The size of these networks varies from a several millions for Wikipedia to billions and more for Facebook and WWW. The description of generic features of these complex networks can be found e.g. in [6].

An important tool for the analysis of directed networks is the construction of the Google matrix of Markov transitions and related PageRank algorithm invented by Brin and Page in 1998 for ranking of all WWW sites (see [7], [8]). This approach has been at the foundations of the Google search engine used world wide. A variety of applications of Google matrix analysis to various directed networks is described by [9].

Here we apply recently developed extensions of Google matrix analysis, which include the REduced GOogle MAtriX (REGOMAX) algorithm [10] and the LInear Response algorithm for GOogle MAtriX (LIRGOMAX) [11], to the SIGNOR PPI network. The efficiency of these algorithms has been demonstrated for Wikipedia networks of politicians [10] and world uni-

URL: `http://www.quantware.ups-tlse.fr/dima` (Dima L. Shepelyansky)

© 2020 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

versities [11], [12] and multi product world trade of UN COMTRADE database [13]. Thus it is rather natural to apply these algorithms to PPI networks which have a typical size being significantly smaller than Wikipedia and WWW.

From a physical view-point the LIRGOMAX approach corresponds to a small probability pumping at a certain network node (or group of nodes) and absorbing probability at another specific node (or group of nodes). This algorithm allows first to determine the most sensitive group of nodes involved in this pumping-absorption process tracing a pathway connecting two selected proteins. In a second stage one can then apply the REGOMAX algorithm and obtain an effective reduced Google matrix, and in particular effective interactions, for the found subset of most sensitive nodes. These interactions are due to either direct or indirect pathways in the global huge network in which is embedded the selected relatively small subset of nodes.

The REGOMAX and LIRGOMAX algorithms originate from the scattering theory of nuclear and mesoscopic physics, field of quantum chaos and linear response theory of electron transport [10], [11].

We point out that the analysis of the SIGNOR PPI network already found biological applications reported by [14], [15], [16], [17]. The detailed review of various applications of the PPI signaling networks is given by [18]. However, the Google matrix analysis has not been used in these studies.

The challenging feature of PPI networks is the bi-functionality of directed links which produce activation or inhibition actions. While in our previous analysis of SIGNOR network by [19] this feature was ignored, here we apply the Ising-PageRank approach developed in [20] for opinion formation modeling. In this Ising-type approach the number of nodes in the PPI network is doubled, with a (+) or (−) attribute for each protein, and the links between doubled nodes are described by 2×2 matrices corresponding to activation or inhibition actions.

In this work we apply the LIRGOMAX and REGOMAX algorithm to the bi-functional PPI network of SIGNOR. We show that this approach allows to determine the effective sensitivity with direct and indirect interactions between a selected pair of proteins. As particular examples we will choose three protein pairs implicating the *Epidermal growth factor receptor (EGFR)* which is considered to play an important role in the context of lung cancer (see e.g. [21], [22]).

The interest to apply computer science methods, such as the PageRank algorithm, to PPI networks is growing (see e.g. the recent review [23]) and we hope that the Google matrix algorithms described in this work will attract the interest of biologists working with PPI networks. In addition to the methods described in [23] these algorithms allow to take into account the bi-functional nature of PPI network links and focus the investigation on a specific group of proteins taking into account all their direct and indirect interactions via the global network.

The paper is constructed as follows: in Section 2 we describe the construction of Google matrix from links between proteins and related LIRGOMAX and REGOMAX algorithms, in Section 3 we characterize data sets and the Ising-PPI-network for bi-functional interactions between proteins, results are presented in Section 4 and the conclusion is given in Section 5. Appendix provides additional matrix data and executable code for the described algorithms for the SIGNOR Ising-PPI-network.

The Google matrix data and executive code of described algorithms are available at http://www.quantware.ups-tlse.fr/QWLIB/google4signornet/ .

2. Methods of Google matrix analysis

2.1. Google matrix construction

The Google matrix G of N nodes (proteins or proteins with (+)/(−) attribute) is constructed from the adjacency matrix A_{ij} with element 1 if node j points to node i and zero otherwise. The matrix G has the standard form $G_{ij} = \alpha S_{ij} + (1 - \alpha)/N$ (see [7], [8], [9]), where S is the matrix of Markov transitions with elements $S_{ij} = A_{ij}/k_{out}(j)$ and $k_{out}(j) = \sum_{i=1}^{N} A_{ij} \neq 0$ being the out-degree of node j (number of outgoing links); $S_{ij} = 1/N$ if j has no outgoing links (dangling node). The parameter $0 < \alpha < 1$ is known as the damping factor with the usual value $\alpha = 0.85$ [8] which we use here. For the range $0.5 \leq \alpha \leq 0.95$ the results are not sensitive to α [8], [9]. A useful view on this G matrix is given by the concept of a random surfer, moving with probability α from one node to another via one of the available directed links or with a jump probability $(1 - \alpha)$ to any node.

The right PageRank eigenvector of G is the solution of the equation $GP = \lambda P$ for the leading unit eigenvalue $\lambda = 1$ [8]. The PageRank $P(j)$ values represent positive probabilities to find a random surfer on a node j ($\sum_j P(j) = 1$). All nodes can be ordered by decreasing probability P numbered by PageRank index $K = 1, 2, ...N$ with a maximal probability at $K = 1$ and minimal at $K = N$. The numerical computation of $P(j)$ is done efficiently with the PageRank iteration algorithm

described by [8]. The idea of this algorithm is simply to start with some initial, sum normalized, vector $P^{(0)}$ of positive entries, e.g. being $1/N$ for simplicity, and then to iterate $P^{(n+1)} = G\,P^{(n)}$ which typically converges after $n = 150 - 200$ iterations (for $\alpha = 0.85$).

It is also useful to consider the original network with inverted direction of links. After inversion the Google matrix G^* is constructed via the same procedure with $G^* P^* = P^*$. The matrix G^* has its own PageRank vector $P^*(j)$ called CheiRank [24], [9]. Its values give probabilities to find a random surfer of a given node and they can be again ordered in a decreasing order with CheiRank index K^* with highest P^* at $K^* = 1$ and smallest at $K^* = N$. On average, the high values of P (P^*) correspond to nodes with many ingoing (outgoing) links [8], [9].

2.2. Reduced Google matrix (REGOMAX) algorithm

The REGOMAX algorithm is described in detail by [10, 12, 19]. It allows to compute efficiently a "reduced Google matrix" $G_{\rm R}$ of size $N_r \times N_r$ that captures the full contributions of direct and indirect pathways appearing in the full Google matrix G between N_r nodes of interest selected from a huge global network with $N \gg N_r$ nodes. For these N_r nodes their PageRank probabilities are the same as for the global network with N nodes, up to a constant multiplicative factor taking into account that the sum of PageRank probabilities over N_r nodes is unity. The computation of $G_{\rm R}$ determines a decomposition of $G_{\rm R}$ into matrix components that clearly distinguish direct from indirect interactions: $G_{\rm R} = G_{\rm rr} + G_{\rm pr} + G_{\rm qr}$ [10]. Here $G_{\rm rr}$ is given by the direct links between the selected N_r nodes in the global G matrix with N nodes. We note that $G_{\rm pr}$ is rather close to the matrix in which each column is approximately proportional to the PageRank vector P_r, satisfying the condition that the PageRank probabilities of $G_{\rm R}$ are the same as for G (up to a constant multiplier due to normalization). Hence, in contrast to $G_{\rm qr}$, $G_{\rm pr}$ doesn't give much new information about direct and indirect links between selected nodes.

The most interesting role is played by $G_{\rm qr}$, which takes into account all indirect links between selected nodes happening due to multiple pathways via the global network of nodes N (see [10]). The matrix $G_{\rm qr} = G_{\rm qrd} + G_{\rm qrnd}$ has diagonal ($G_{\rm qrd}$) and non-diagonal ($G_{\rm qrnd}$) parts with $G_{\rm qrnd}$ describing indirect interactions between selected nodes. The exact formulas for all three components of $G_{\rm R}$ are given in [10]. It is also useful to compute the weights $W_{\rm R}$, $W_{\rm pr}$, $W_{\rm rr}$, $W_{\rm qr}$ of $G_{\rm R}$ and its 3 matrix components $G_{\rm pr}$, $G_{\rm rr}$, $G_{\rm qr}$ given by the sum of all its elements divided by the matrix size N_r. Due to the column sum normalization of $G_{\rm R}$ we obviously have $W_{\rm R} = W_{\rm rr} + W_{\rm pr} + W_{\rm qr} = 1$.

We note that the matrix elements of $G_{\rm qr}$ may have negative values (only the full reduced matrix $G_{\rm R}$ should have positive elements; $G_{\rm rr}$ also has only positive matrix elements) but these negative values are found to be small for the Ising-PPI-networks and do not play a significant role. A similar situation for Wikipedia networks is discussed by [10], [11].

2.3. LIRGOMAX algorithm

The detained description of the LIRGOMAX algorithm is given by [11]. It performs an infinitely weak ε-probability injection (pumping) at one node (a protein or a protein with (+)/(−) attribute) and absorption at another node of interest. This process is described by the modified PageRank iteration $P^{(n+1)} = G\,F(\varepsilon, P^{(n)})$ where the vector valued function $F(\varepsilon, P)$ has the components $P(i) + \varepsilon$ for i being the index of the injection/pumping node, $P(j) - \varepsilon$ for j being the index of the absorption node and simply $P(k)$ for all other nodes k. In this way the vector $F(\varepsilon, P)$ is also sum normalized if P is sum normalized and obviously $F(0, P) = P$ is the identity operation. In [11] a more general version of $F(\varepsilon, P)$ was considered with potentially different prefactors for the ε contributions, injection/absorption at possibly more than two nodes and an additional renormalization factor to restore the sum normalization (which is automatic in the simple version). However, for the applications in this work the above given simple version of $F(\varepsilon, P)$ is sufficient.

In principle one can solve iteratively the above modified PageRank iteration formula which converges at the same rate as the usual PageRank iteration algorithm and provides a modified ε-depending PageRank $P(\varepsilon)$. Then one can compute the linear response vector $P_1 = dP(\varepsilon)/d\varepsilon|_{\varepsilon=0} = \lim_{\varepsilon\to 0}[P(\varepsilon) - P_0]/\varepsilon$ where P_0 is the PageRank obtained for $\varepsilon = 0$. However the naive direct evaluation of this limit is numerically not stable in the limit $\varepsilon \to 0$. Fortunately as shown by [11] it is possible to compute P_1 directly in an accurate and efficient way by solving the inhomogeneous PageRank equation

$$P_1 = G\,P_1 + V_0 \quad , \quad V_0 = G\,W_0 \qquad (1)$$

where the vector W_0 has only two non-zero components for the two particular injection or absorption nodes $W_0(i) = 1$ or $W_0(j) = -1$ respectively. Therefore a more explicit expression for the vector V_0 appearing in (1) is $V_0(k) = G_{ki} - G_{kj}$ (for all nodes k). We mention that the three vectors P_1, V_0 and W_0 are orthogonal to the vector $E^T = (1, \ldots, 1)$ composed of unit entries, i.e.

$\sum_k P_1(k) = \sum_k W_0(k) = \sum_k V_0(k) = 0$. Furthermore, all of these vectors, especially P_1 have *real* positive or negative entries (note that in general eigenvectors of a non-symmetric real matrix may be complex).

A formal solution of the inhomogeneous PageRank equation is: $P_1 = \sum_{n=0}^{\infty} G^n V_0 = (\mathbf{1} - G)^{-1} V_0$ which is well defined since V_0, when expanded in the basis of (generalized) eigenvectors of G, does NOT have a contribution of P_0 (the only eigenvector of G with eigenvalue 1) such that the singularity of the matrix inverse does not constitute a problem. Of course numerically, we compute P_1 in a different way, as described by [11] one can iterate the equation $P_1^{(n+1)} = G P_1^{(n)} + V_0$ with $P_1^{(0)} = 0$ which converges with the same rate as the usual PageRank iteration.

We note that a propagator somewhat similar to the above expression $P_1 = (\mathbf{1} - G)^{-1} V_0$, namely $\tilde{P} = (\mathbf{1} - \gamma G)^{-1} V_{init}$, has been used in [25] as the ImpactRank of specific nodes related to an initial probability localized on a certain initial node described by the initial vector V_{init}. However, this ImpactRank used $\gamma < 1$ so that there was no singularity in denominator and also $\tilde{P}$ represented a certain stationary probability distribution while P_1 represents a deviation from the stationary distribution of PageRank probability P. In fact the propagator discussed in [23] corresponds to the ImpactRank case [25] with $\gamma < 1$ thus being qualitatively different from the LIRGOMAX propagator considered here.

In a similar as the PageRank P_0 is characterized by the index K we introduce the index K_L by ordering $|P_1|$ such that $K_L = 1$ corresponds to the node with largest value of $|P_1|$ and $K_L = N$ to the node with smallest value of $|P_1|$. Once P_1 is computed for the pair of chosen injection/absorption nodes we determine the 20 top nodes with strongest negative values of P_1 and further 20 top nodes with strongest positive values of P_1 which constitute a subset of 40 nodes which are the most significant nodes participating in the pathway between the pumping node i and absorbing node j. We also require that these two particular nodes i and j belong to this subset. If this is not automatically the case we replace the node at total position 20 (position 20 for strongest negative values of P_1) with the absorption node j and/or the node at total position 40 (position 20 for strongest positive values of P_1) with the injection node i. This situation happens once for the absorption node of the third example below which has a very low ranking position $K_L \approx 2000$ for $|P_1|$.

In general from a physical/biological point of view we indeed expect that the two particular injection/absorption nodes belong automatically to the selected subset of most sensitive nodes. However, there is no simple or general mathematical argument for this.

Using this subset of top nodes in the K_L ranking we then apply the REGOMAX algorithm to compute the reduced Google matrix and its components and in particular we determine the effective direct and indirect interactions of this reduced network. The advantage of the application of LIRGOMAX at the initial stage is that it provides an automatic and more rigorous procedure to determine an interesting subset of protein nodes related to the pumping between nodes i and j instead of using an arbitrary heuristic choice for such a subset.

3. Data sets and Ising-PPI-network construction

We use the open public SIGNOR PPI network [2] (April 2019 release for human, mouse and rat). This network contains $N = 4341$ nodes (proteins) and $N_\ell = 12547$ directed hyperlinks between nodes. Each protein (node) is described by their name and identifier.

A new interesting feature of this PPI directed network is that its hyperlinks have activation and inhibition actions. For some links the functionality is unclear and then they are considered to be neutral. This feature rises an interesting mathematical challenge for the Google matrix description of such bi-functional networks. To meet this challenge we use the Ising-PageRank approach developed by [20] for a model of opinion formation on social networks. In this approach each node is doubled getting two components marked by (+) and (−). The activation links point to the (+) components and inhibition links point to the (−) components. Such transitions between doubled nodes are described by 2×2 block matrices σ_+ (σ_-) matrices with entries 1 (0) in the first row and 0 (1) in the second row as for Ising spin-1/2 (see details described in Appendix). A neutral transition is described by 2×2 matrix σ_0 with all elements being $1/2$. Thus for this Ising-network (doubled-size network) we have doubled number of node $N = 8682$ and the total number of hyperlinks being $N_\ell = 27266$; among them there are $N_{act} = 14944$ activation links, $N_{inh} = 7978$ inhibition links and $N_{neut} = 4344$ neutral links ($N_\ell = N_{act} + N_{inh} + N_{neut}$). From this weighted Ising-PPI-network with $N_\ell = 27266$ nodes we construct the Google matrix following the standard rules described by [8], [9] and also given above.

Below we apply the Google matrix analysis taking into account the bi-functionality PPI and illustrate the efficiency of the LIRGOMAX and REGOMAX algorithms for the SIGNOR Ising-PPI-network.

The details of Ising-PPI-network construction, its main statistical properties and an executable code for

the described algorithms are provided in Appendix and in [26]. Below we discuss the results obtained with the LIRGOMAX and REGOMAX algorithms for three examples of specific pathways between two specific proteins.

4. Results

Here we present results obtained with LIRGOMAX and REGOMAX algorithms for pathways between several pairs of selected proteins.

4.1. Case of pathway EGFR - JAK2 proteins

As a first example we choose the node *EGFR P00533 (+)* for injection (pumping) and *JAK2 O60674 (-)* for absorption. It is known that mutations affecting the protein EGFR expression or activity could result in lung cancer (see e.g. [21]; [22]). This protein interacts with the protein JAK2 whose mutations have been implicated in various types of cancer. We argue that the injection (pumping) at *EGFR P00533 (+)* and absorption at *JAK2 O60674 (-)* should involve certain variations of the PageRank probability, represented by P_1, showing interactions between various proteins actively participating in the pathway from *EGFR P00533 (+)* to *JAK2 O60674 (-)*. The pumping process can be viewed as a result of disease development and absorption as a certain mutation of this disease into another one.

The global PageRank indices of these two nodes are $K = 90$ (PageRank probability $P(90) = 0.0009633$) for *EGFR P00533 (+)* and $K = 470$ (PageRank probability $P(470) = 0.0003444$) for *JAK2 O60674 (-)*. As described above in the LIRGOMAX computations we choose the vector in V_0 which appears in the inhomogeneous PageRank equation (1) as $V_0 = G\,W_0$ with $W_0(K = 90) = +1$, $W_0(K = 470) = -1$ and $W_0(K) = 0$ for all other values of the Kindex K. We remind that both W_0 and V_0 are orthogonal to the left leading eigenvector $E^T = (1, \ldots, 1)$ of G according to the general description of the LIRGOMAX algorithm given above and in [11].

For comparison we let us note that the top 4 PageRank nodes are $K = 1$ ($P(1) = 0.003041$) for *CASP3 P42574 (+)*, $K = 2$ ($P(2) = 0.002821$) for *NOTCH1 P46531 (+)*, $K = 3$ ($P(3) = 0.002433$) for *PIK3CD O00329 (-)*, $K = 4$ ($P(4) = 0.002413$) for *CTNNB1 P35222 (-)* (other values/data are available at [26]).

Similar to the two Wikipedia examples analyzed by [11] the LIRGOMAX algorithm selects the proteins mostly affected by injection/absorption process with 20 most positive (EGFR block) and 20 most negative

Table 1: Top 20 nodes of strongest negative values of P_1 (index number $i = 1, \ldots, 20$) and top 20 nodes of strongest positive values of P_1 (index number $i = 21, \ldots, 40$) with P_1 being created as the linear response of the PageRank of the Ising-PPI-network with injection (or pumping) at *EGFR P00533 (+)* and absorption at *JAK2 O60674 (-)*; K_L is the ranking index obtained by ordering $|P_1|$ and K is the usual PageRank index obtained by ordering the PageRank P_0.

i	K_L	K	Node name
1	1	30	JAK2 O60674 (+)
2	4	470	JAK2 O60674 (-)
3	5	554	IFNGR2/INFGR1 SIGNOR-C142 (+)
4	6	354	ARHGEF1 Q92888 (+)
5	7	631	APOA1 P02647 (+)
6	8	956	CSF2RA/CSF2RB SIGNOR-C212 (+)
7	9	57	STAT1 P42224 (+)
8	10	204	MAP3K5 Q99683 (-)
9	12	1008	STAT4 Q14765 (+)
10	13	825	CCR2 P41597 (+)
11	14	2377	PRMT5 O14744 (-)
12	15	2378	STAM Q92783 (+)
13	16	1482	EPOR P19235 (+)
14	17	1117	CSF2RA P15509 (+)
15	18	959	ITGAL P20701 (+)
16	19	1968	CTLA4 P16410 (-)
17	20	2058	STAP2 Q9UGK3 (+)
18	21	2024	ITGB2 P05107 (+)
19	22	532	EZH2 Q15910 (-)
20	23	1196	GTF2I P78347 (+)
21	2	29	GRB2 P62993 (+)
22	3	172	FES P07332 (+)
23	11	90	EGFR P00533 (+)
24	27	3	PIK3CD O00329 (-)
25	30	126	CBL P22681 (+)
26	31	136	EGFR P00533 (-)
27	32	648	EZR P15311 (+)
28	36	38	PTK2 Q05397 (+)
29	37	456	GAB1 Q13480 (+)
30	39	424	BCR P11274 (-)
31	42	124	PIK3R1 P27986 (+)
32	43	26	PLCG1 P19174 (+)
33	44	58	SHC1 P29353 (+)
34	45	88	ESR1 P03372 (+)
35	46	2398	VAV2 P52735 (+)
36	47	746	SHC3 Q92529 (+)
37	48	291	ERBB2 P04626 (+)
38	49	888	ERBB3 P21860 (+)
39	51	1109	NCK1 P16333 (+)
40	52	1531	CRK P46108 (-)

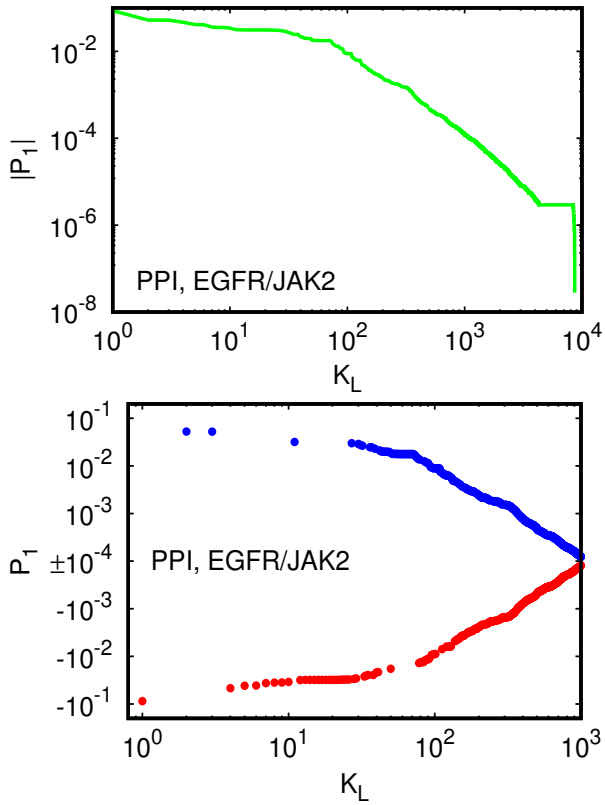

Figure 1: Linear response vector P_1 of PageRank for the Ising-PPI-network with injection (or pumping) at *EGFR P00533 (+)* and absorption at *JAK2 O60674 (-)*. Here K_L is the ranking index obtained by ordering $|P_1|$ from maximal value at $K_L = 1$ down to minimal value. Top panel shows $|P_1|$ versus K_L in a double logarithmic representation for all N nodes. Bottom panel shows a zoom of P_1 versus K_L for $K_L \leq 10^3$ in a double logarithmic representation with sign; blue data points correspond to $P_1 > 0$ and red data points to $P_1 < 0$.

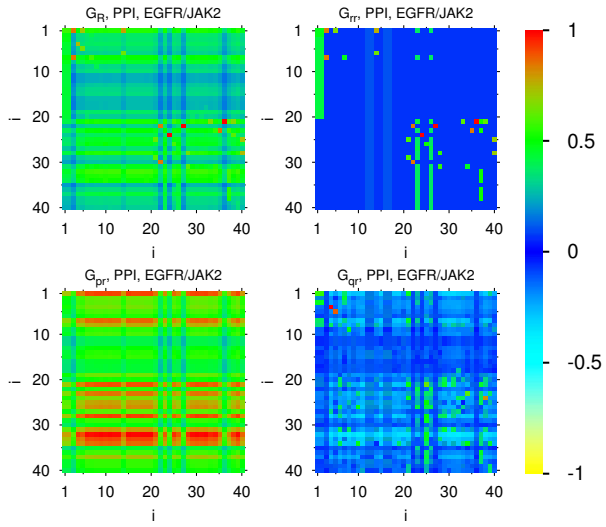

Figure 2: Reduced Google matrix components G_R, G_{pr}, G_{rr} and G_{qr} for Ising-PPI-network and the subgroup of nodes given in Table 1 corresponding to injection at *EGFR P00533 (+)* and absorption at *JAK2 O60674 (-)* (see text for explanations). The axis labels correspond to the index i used in Table 1. The relative weights of these components are $W_{pr} = 0.761$, $W_{rr} = 0.220$, and $W_{qr} = 0.019$. The values of the color bar correspond to $\mathrm{sgn}(g)(|g|/\max|g|)^{1/4}$ where g is the shown matrix element value. The exponent 1/4 amplifies small values of g for a better visibility.

(JAK2 block) values of P_1 shown in Table 1. Here the pumped protein *EGFR P00533 (+)* is on the third position in its block of positive P_1 values ($i = 23$) and with $K_L = 11$ (where K_L is the ranking index obtained by ordering the components of $|P_1|$) while the protein with absorption *JAK2 O60674 (-)* has the second position in its block of negative P_1 values ($i = 2$) with $K_L = 4$. Thus these two nodes are not at the first positions in their respective blocks but still they are placed at very high positions.

The dependence of $|P_1|$ of the index K_L is shown in the top panel of Figure 1. The decay of $|P_1|$ is relatively slow for $K_L \leq 40$ followed by a more rapid drop for $K_L > 40$. The bottom panel shows the dependence of positive (blue) and negative (red) values of P_1 on K_L. We note that the top absolute values $|P_1|$ for blue and red components have comparable values being of the order of $|P_1| \sim 0.1$ for approximately $K_L \leq 40$. However, in this range the number of positive (blue) values of P_1 is significantly smaller compared to the number of negative (red) values of P_1. This point can also be seen from the column of K_L values in Table 1. Another feature visible from Table 1 is that the number of proteins with negative component (−) is significantly smaller than those with a positive component (+) (5 for $1 \leq i \leq 20$ and 4 for $21 \leq i \leq 40$. We return to the properties of positive and negative components a bit later.

After the selection of most significant 40 nodes of the pathway between the two injection/absorption proteins (see Table 1) we apply the REGOMAX algorithm which determines all matrix elements of Markov transitions between these 40 nodes including all direct and indirect pathways via the large global Ising-PPI-networks network with 8682 nodes.

The reduced Google matrix G_R and its three components G_{pr}, G_{rr}, G_{qr} are shown in Figure 2 for proteins of Table 1 ($1 \leq i \leq 40$). The weight of the component G_{pr} is $W_{pr} = 0.761$ being not so far from unity but this value is below $W_{pr} \approx 0.95$ appearing usually in Wikipedia networks [10]; [11]. We attribute this to a significantly smaller number of links per node being $\ell = N_\ell/N \approx 3.1$ for the Ising-PPI-network while for the English Wikipedia network of 2017 we have $\ell \approx 22.5$ [11]. Indeed, the weight $W_{rr} = 0.220$ of direct transi-

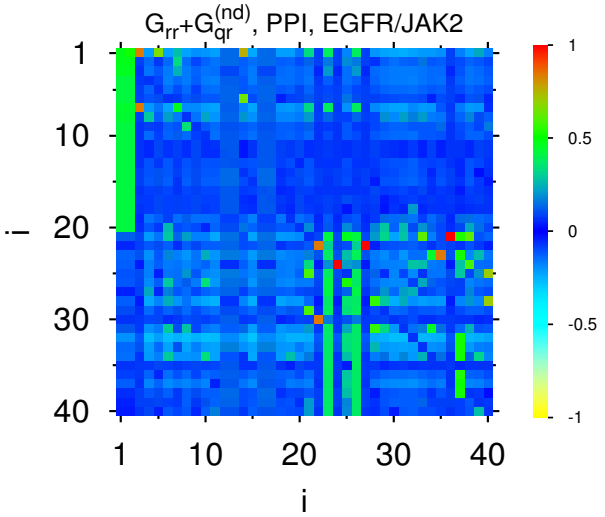

Figure 3: Same as in Fig. 2 but for the matrix $G_{\rm rr} + G_{\rm qrnd}$, where $G_{\rm qrnd}$ is obtained from $G_{\rm qr}$ by putting its diagonal elements at zero; the weight of these two components is $W_{rr+qrnd} = 0.227$.

tions of $G_{\rm rr}$ is significantly larger than the corresponding values for the Wikipedia case with $W_{\rm rr} \approx 0.04$. However, the weights $W_{\rm qr} = 0.019$ are comparable for both reduced networks.

The matrix structure of direct transitions $G_{\rm rr}$ has a clear two block structure with dominant transitions inside each block associated to EGFR and JAK2 with only 4 significant matrix elements from the EGFR to the JAK2 block. These matrix elements correspond to links from EGFR (±) to JAK2 (+) and STAT1(+) and have the same value $g \approx 0.0167$ while all other matrix elements (of this EGFR to JAK2 block) are very small with the value $g \approx 1.73\times10^{-5}$ corresponding to the minimal value $(1-\alpha)/N$ in G related to the damping factor $\alpha = 0.85$.

The matrix $G_{\rm pr}$ (which is exactly of rank 1) has a very simple structure with all columns being (approximately) proportional to the (local) PageRank of $G_{\rm R}$ (which is itself proportional to the global PageRank projected onto the subset of 40 nodes) and one clearly sees that the strong horizontal red lines correspond to index positions i of Table 1 where the corresponding index K is quite low below $\sim$ 100 corresponding to a relatively high PageRank position. The full reduced matrix $G_{\rm R}$ is numerically dominated by $G_{\rm pr}$ (but less clearly as for typical Wikipedia cases) and has at first sight a similar structure as $G_{\rm pr}$ but with somewhat smaller values. However, some of the strongest direct links (from $G_{\rm rr}$) are also visible. Similarly to the Wikipedia network of politicians as discussed in [10] both matrix components $G_{\rm R}$ and $G_{\rm pr}$ are not very usefully to identify the indirect links.

The indirect links are visible in the matrix $G_{\rm qr}$. As explained and shown mathematically in [10] they correspond to pathways where a given node i_1 of the small subset points to a certain node outside the subset (in the big surrounding PPI network) which itself points eventually to another node outside the subset and comes after a finite number of iterations finally back to a different node i_2 inside the subset. This provides an indirect link from i_1 to i_2 and the weight or strength of this indirect link is characterized by the value of the matrix element $(G_{\rm qr})_{i_2,i_1}$. According to Figure 2 there are now also significant interactions between the two blocks of EGFR and JAK2 for the matrix $G_{\rm qr}$, sometimes with negative values (note that the matrix elements of $G_{\rm qr}$ may be negative). Figure 3 shows the the sum of the two components $G_{\rm rr} + G_{\rm qrnd}$ ($G_{\rm qrnd}$ corresponds to $G_{\rm qr}$ without its diagonal elements) which confirms this observation. Actually, we consider that the elements of $G_{\rm rr} + G_{\rm qrnd}$ describe best the combined direct and indirect links for the given subset.

Due to the contribution of indirect transitions there are additional transitions between these two blocks where the four strongest additional elements of $G_{\rm qr}$ have values $g = 0.0106$ (*GRB2 P62993 (+)* to *JAK2 O60674 (+)*); $g = 0.0099$ (*GRB2 P62993 (+)* to *STAT1 P42224 (+)*); $g = 0.0059$ (*GAB1 Q13480 (+)* to *GTF2I P78347 (+)*); $g = 0.0039$ (*PIK3R1 P27986 (+)* to *GTF2I P78347 (+)*). There are also 11 additional transitions with $g > 0.1$. Thus even if the weight of $G_{\rm qr}$ is not high it provides important new indirect interactions between proteins from the EGFR block to the JAK2 block.

The situation is even more striking when we consider the transitions from the JAK2 block to the EGFR block. There are no direct links between these blocks in this direction from the global network but due to the construction of the Google matrix described above there are still numerically very small values g for the matrix elements of $G_{\rm rr}$ due to dangling nodes (nodes with no outgoing links) with $g = 1/N \approx 1.15 \times 10^{-4}$ (in certain columns) or due to the damping factor term $(1-\alpha)/N \approx 1.73 \times 10^{-5}$ (for the other columns). On the other side concerning the indirect links described by $G_{\rm qr}$ we find rather significant transitions from the JAK2 block to the EGFR block with the four largest values: $g \approx 0.0122$ (*CCR2 P41597 (+)* to *EGFR P00533 (-)* and to *ESR1 P03372 (+)*); $g \approx 0.006$ (*CSF2RA/CSF2RB SIGNOR-C212 (+)* to *ESR1 P03372 (+)* and to *PIK3R1 P27986 (+)*). There are also 9 additional transitions with $g > 0.001$. Complete data files for the matrix elements of matrix components (for all examples) are

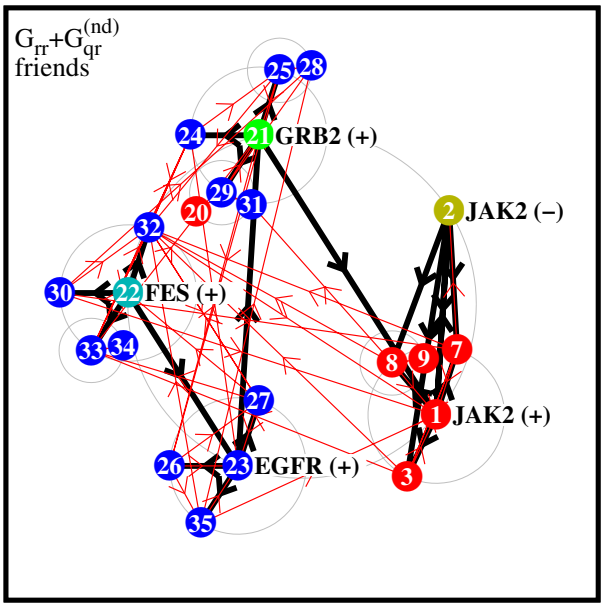

Figure 4: Network of friends for the subgroup of nodes given in Table 1 corresponding to injection at *EGFR P00533 (+)* and absorption at *JAK2 O60674 (-)* constructed from the matrix $G_{\rm rr} + G_{\rm qrnd}$ using 4 top (friends) links per column (see text for explanations).

available at [26].

It is convenient to present the interactions between proteins, generated by the matrix elements of the sum of two components $G_{\rm rr} + G_{\rm qrnd}$ from Figure 3, in the form of a network shown in Figure 4. To construct the network of effective friends, we select first five initial nodes which are placed on a (large) circle: the two nodes with injection and absorption (*EGFR (+)* (injection node, blue) and *JAK2 (-)* (absorption node, olive) and three other nodes with a rather top position in the K_L ranking: *JAK2 (+)* (related to *JAK2 (-)* with $K_L = 1$, $i = 1$, red), *GRB2 (+)* (with $K_L = 2$, $i = 21$, green) and *FES (+)* (with $K_L = 3$, $i = 22$, cyan). For each of these five initial nodes we determine four friends by the criterion of largest matrix elements (in modulus) in the same column, i.e. corresponding to the four strongest links from the initial node to the potential friends. The friend nodes found in this way are added to the network and drawn on circles of medium size around their initial node (if they do not already belong to the initial set of 5 top nodes). The links from the initial nodes to their friends are drawn as thick black arrows. For each of the newly added nodes (level 1 friends) we continue to determine the four strongest friends (level 2 friends) which are drawn on small circles and added to the network (if there are not already present from a previous level). The corresponding links from level 1 friends to level 2 friends are drawn as thin red arrows.

Each node is marked by the index i from the first column of Table 1. The colors of the nodes are essentially red for nodes with strong negative values of P_1 (corresponding to the index $i = 1, \ldots, 20$) and blue for nodes with strong positive values of P_1 (for $i = 21, \ldots, 40$). Only for three of the initial nodes we choose different colors which are olive for *JAK (-)*, green for *GRB2 (+)* and cyan for *FES (+)*. This procedure generates the directed friendship network shown in Figure 4.

The obtained network of Figure 4 has a rather clear separation between the two blocks related to EGFR (mainly blue nodes) and JAK2 (mainly red nodes). There is only one link of first level (black arrow) from the EGFR block (GRB2 (+)) to the JAK2 block (JAK2 (+)), Of course, there are other strong direct transitions from the EGFR block to the JAK2 block as described above, but these links are weaker than the 4 closest friends and therefore they do not appear in the network structure of Figure 4. However, we see that there are many links between the two blocks on the secondary level of red arrows.

The block of JAK2 (red nodes) is rather compact with only 6 nodes (one red node at $i = 20$ is more linked to the EGFR block). In contrast the EGFR block contains 15 (blue) nodes showing that this group of proteins is characterized by broader and more extensive interconnections. We think that such a network presentation provides a useful qualitative image of the effective interactions between the two groups of proteins.

Network figures, for this example and the other two examples discussed below, constructed in the same way using the other matrix components $G_{\rm R}$, $G_{\rm rr}$ or $G_{\rm qr}$ (instead of $G_{\rm rr} + G_{\rm qrnd}$) or using strongest matrix elements in rows (instead of columns) to determine follower networks are available at [26].

4.2. *Magnetization of proteins of EGFR - JAK2 pathway*

In the Ising-PPI-network each protein is described by two components which can be considered as spin up or down state. The PageRank probability of a protein is given by the sum of probabilities of its two components with $P(j) = P_+(j) + P_-(j)$. It can be shown that due to the structure of the matrix transitions given by the matrices $\sigma_+, \sigma_-, \sigma_0$ the sum of probabilities $P(j)$ for a given protein j is the same as for the directed PPI network without doubling (see Appendix). Thus the activation or inhibition links in the Ising-PPI-network of doubled size only redistribute PageRank probability for a given protein between up and down components. The physical meaning of these up and down

component probabilities P_+ and P_- is qualitatively related to the fact that on average the PageRank probability P of a node is proportional to the number of ingoing links. Thus P_+ is proportional to the number of ingoing activation links and P_- is proportional to the number of ingoing inhibition links. Thus we can characterize each node by its normalized magnetization $M(j) = (P_+(j) - P_-(j))/(P_+(j) + P_-(j))$. By definition $-1 \leq M(j) \leq 1$. Big positive values of M mean that this protein has mainly ingoing activation links while big negative values mean that this protein has mainly inhibition ingoing links. In principle, we can also study the magnetization of CheiRank probability of proteins given by $M^*(j) = (P^*_+(j) - P^*_-(j))/(P^*_+(j) + P^*_-(j))$ but we keep this for further investigations. We note that $M(j)$ and $M^*(j)$ represent the normalized values which are independent of the total probability $P(j), P^*(j)$. Thus the magnetization of nodes of the reduced Google matrix remains the same as in the global network.

We take all different 38 proteins present in Table 1 and consider their magnetization (this number is smaller than 40 since for few proteins both (+) or (−) components are present in this Table). All these 38 proteins are listed in Table 2 with their local PageRank and CheiRank indices K and K^*. The distribution of these 38 proteins on the PageRank-CheiRank plane is shown in Figure 5 and the colors of the square boxes presents the values of $M(j)$ (see caption of Figure 5). The three proteins with the strongest positive magnetizations are *PLCG1 P19174* ($M = 0.8959$), *GRB2 P62993* ($M = 0.8899$), *FES P07332* ($M = 0.8719$) and with the strongest negative values are *BCR P11274* ($M = -0.7799$), *PIK3CD O00329* ($M = -0.7328$), *PRMT5 O14744* ($M = -0.3527$). In total there are only 5 proteins of Table 2 with negative magnetization values. We attribute this to the fact that the number of inhibition links is smaller than the number of activation ones. We think that the magnetization of proteins can provide new interesting information about the functionality of proteins.

4.3. Examples of other protein pathways

We also consider two other proteins pairs for injection (pumping) and absorption which we analyzed in the same way. Again we compute the vector $V_0 = G\,W_0$ where W_0 has only two non-zero components being 1 at the pumping node and -1 at the absorption node, we solve the inhomogeneous PageRank equation (1) to obtain the linear response vector P_1 from which we determine a set of 40 nodes composed with 20 strongest negative and 20 strongest positive values. In order to ensure that the two initial injection and absorption nodes also

Table 2: Group of 38 nodes of the single protein network obtained from the group of Table 1 by removing the (+) and (−) attributes. K (K^*) represent the local rank indices obtained from the PageRank (CheiRank) ordering of the single protein network. The index i is the same as in Table 1 where the two values $i = 2$ and $i = 26$ do not appear here since they correspond to the two nodes where both components (+) and (−) are present in Table 1.

K	K^*	i	Node name
1	34	24	PIK3CD O00329
2	3	1	JAK2 O60674
3	1	23	EGFR P00533
4	2	32	PLCG1 P19174
5	10	21	GRB2 P62993
6	11	28	PTK2 Q05397
7	8	34	ESR1 P03372
8	5	7	STAT1 P42224
9	7	33	SHC1 P29353
10	13	8	MAP3K5 Q99683
11	4	25	CBL P22681
12	25	31	PIK3R1 P27986
13	6	37	ERBB2 P04626
14	21	22	FES P07332
15	12	5	APOA1 P02647
16	9	19	EZH2 Q15910
17	27	4	ARHGEF1 Q92888
18	30	29	GAB1 Q13480
19	28	3	IFNGR2/INFGR1 SIGNOR-C142
20	24	30	BCR P11274
21	22	40	CRK P46108
22	26	27	EZR P15311
23	15	6	CSF2RA/CSF2RB SIGNOR-C212
24	19	38	ERBB3 P21860
25	33	36	SHC3 Q92529
26	18	10	CCR2 P41597
27	14	39	NCK1 P16333
28	31	15	ITGAL P20701
29	17	9	STAT4 Q14765
30	23	14	CSF2RA P15509
31	16	20	GTF2I P78347
32	37	16	CTLA4 P16410
33	36	13	EPOR P19235
34	20	18	ITGB2 P05107
35	38	17	STAP2 Q9UGK3
36	32	11	PRMT5 O14744
37	35	12	STAM Q92783
38	29	35	VAV2 P52735

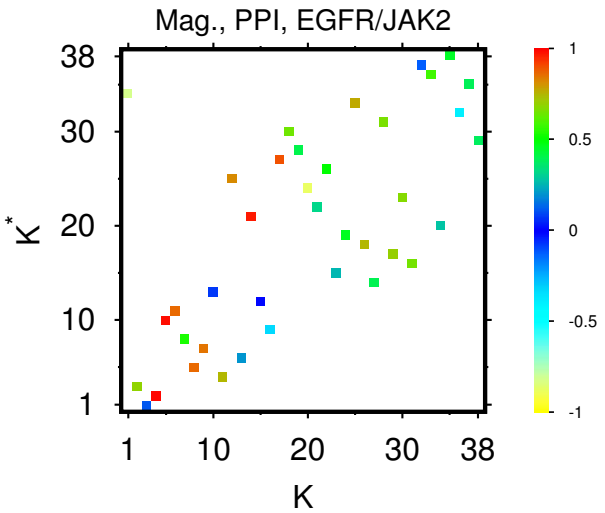

Figure 5: PageRank "magnetization" $M(j) = (P_+(j) - P_-(j))/(P_+(j) + P_-(j))$ of proteins of Table 2 shown on the PageRank-CheiRank plane (K, K^*) of local indices; here j represents a protein node in the initial single protein network and $P_\pm(j)$ are the PageRank components of the Ising-PPI-network (see text). The values of the color bar correspond to $M/\max|M|$ with $\max|M| = 0.896$ being the maximal value of $|M(j)|$ for the shown group of proteins.

belong to this subset we eventually replace the node at position 20 for strongest positive and/or negative values with the injection and/or absorption node respectively. Here we only show and discuss the list of the obtained subsets and the effective network schemes corresponding to Table 1 and Figure 4 for these two examples while Tables and Figures analogous to Table 2, Figures 1, 2, 3, 5 are given in Appendix.

First we discuss the case of injection at *MAP2K1 Q02750 (+)* and absorption at *EGFR P00533 (-)*. The protein MAP2K1 is a member of the dual-specificity protein kinase family that acts an integration point for multiple biochemical signals. There is no direct link between MAP2K1 and EGFR. The global PageRank indices of these two nodes are $K = 84$ (PageRank probability $P(84) = 0.0009794$) for *MAP2K1 Q02750 (+)* and $K = 136$ (PageRank probability $P(136) = 0.0007817$) for *EGFR P00533 (-)*. The subset of most sensitive proteins obtained from the LIRGOMAX algorithm for this protein pair is given in Table 3. These proteins are different from those of Table 1. We note now that the injection and absorption proteins have lower positions in the rank indices K_L and i of Table 3. We attribute this somehow unexpected result of the P_1 ranking to rather nontrivial vortex flows on the Ising-PPI-network.

The friendship network for this case is shown in Fig-

Table 3: Same as in Table 1 but for injection (pumping) at *MAP2K1 Q02750 (+)* and absorption at *EGFR P00533 (-)*.

i	K_L	K	Node name
1	15	172	FES P07332 (+)
2	16	29	GRB2 P62993 (+)
3	17	90	EGFR P00533 (+)
4	18	3	PIK3CD O00329 (-)
5	19	126	CBL P22681 (+)
6	20	648	EZR P15311 (+)
7	21	136	EGFR P00533 (-)
8	22	38	PTK2 Q05397 (+)
9	23	30	JAK2 O60674 (+)
10	24	57	STAT1 P42224 (+)
11	26	456	GAB1 Q13480 (+)
12	27	424	BCR P11274 (-)
13	29	9	PI3K SIGNOR-C156 (+)
14	32	26	PLCG1 P19174 (+)
15	33	746	SHC3 Q92529 (+)
16	34	2398	VAV2 P52735 (+)
17	35	291	ERBB2 P04626 (+)
18	36	15	STAT3 P40763 (+)
19	37	888	ERBB3 P21860 (+)
20	38	40	JAK1 P23458 (+)
21	1	125	CEBPA P49715 (+)
22	2	144	MAPK14 Q16539 (-)
23	3	54	GSK3B P49841 (+)
24	4	543	TAL1 P17542 (-)
25	5	74	CASP9 P55211 (-)
26	6	16	PPARG P37231 (-)
27	7	1491	ARRB2 P32121 (+)
28	8	156	MAPK3 P27361 (+)
29	9	84	MAP2K1 Q02750 (+)
30	10	246	MAPK1 P28482 (+)
31	11	80	IRS1 P35568 (-)
32	12	1	CASP3 P42574 (+)
33	13	523	KIF3A Q9Y496 (+)
34	14	7	ERK1/2 SIGNOR-PF1 (+)
35	25	528	ERG P11308 (+)
36	28	106	MEF2C Q06413 (+)
37	30	826	ANGPT2 O15123 (+)
38	31	290	TEK Q02763 (+)
39	78	181	CPT1B Q92523 (+)
40	86	20	JUN P05412 (+)

ure 6 (the construction method is the same as Figure 4). The 5 proteins of the initial large circle are EGFR (-) (olive), FES (+) (red), MAP2K1 (+) (cyan), MAPK14 (-) (green), CEBRPA (+) (blue). In this network we find a number of strong indirect links from the block of *MAP2K1 Q02750 (+)* (blue nodes) to *EGFR P00533 (-)* (red nodes) for which there is no direct link (e.g. from $i = 21$ to $i = 14$ proteins of Table 3). In the opposite direction from red to blue nodes there are only two strong direct matrix elements of G_{rr} being from *PI3K SIGNOR-C156 (+)* $i = 13$ to *IRS1 P35568 (-)* $i = 21$ with $g = 0.08501$ and from *STAT3 P40763 (+)* $i = 18$ to *CASP3 P42574 (+)* $i = 32$ with $g = 0.03543$ with all other elements being below 1.8×10^{-5}. However, in this direction there are 9 new indirect links with elements $g > 0.01$ and 20 with $g > 0.005$. This results in a rather dense network with many links shown in Figure 6. From the network structure we see that the proteins $i = 25, 40$ of the blue block are more closely related with proteins of the red block and inversely the proteins $i = 10, 18, 20$ of the red block are more closely related with proteins of the blue block.

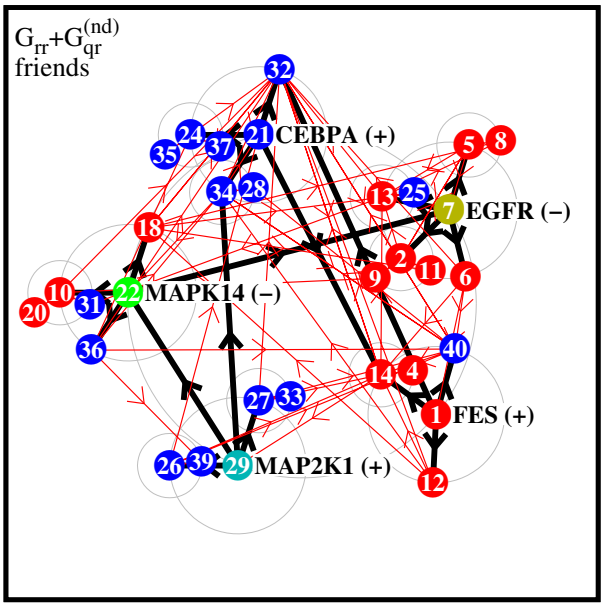

Figure 6: Same as Figure 4 but for the pathway of Table 3.

As a further example we also briefly discuss the pathway generated by injection at *EGFR P00533 (+)* and absorption at *PIK3CA P42336 (-)*. These two proteins are conventional bio markers of lung cancer (see e.g. [22]). The global PageRank indices of these two nodes are $K = 90$ (PageRank probability $P(90) = 0.0009633$) for *EGFR P00533 (+)* and $K = 1604$ (PageRank probability $P(1604) = 0.0001366$) for *PIK3CA P42336 (-)*.

The most sensitive proteins obtained by the LIRGOMAX algorithm, are shown in Table 4. However, now the absorption node *PIK3CA P42336 (-)* has a very low value (in modulus) of P_1 ($P_1 = -4.59 \times 10^{-5}$, $K_L = 2806$) and does initially not belong to the group of nodes with 20 top strongest negative values. Therefore we replace the node *AKT3 Q9Y243 (+)* ($K_L = 70$) which was initially selected for $i = 20$ by the absorption node *PIK3CA P42336 (-)*. Furthermore, also its (+) component *PIK3CA P42336 (+)* ($P_1 = -0.004546$ and $K_L = 138$) does not appear in Table 4 showing that the influence of *EGFR P00533 (+)* on the protein *PIK3CA P42336* is rather low.

The friendship network structure of shown in Figure 7 shows a clear separation between the two blocks of positive (blue) and negative (red) P_1 values. However, some proteins of one block happen to be closer to proteins of the other block (e.g. proteins $i = 10, 14$ from the red block are closer to the blue block and blue block protein $i = 29$ is closer to the proteins of the red block). We also note that concerning the links from the blue to the red block there are 9 significant direct transitions (matrix elements of G_{rr} larger than 0.01) and 35 significant indirect *and* direct transitions (matrix elements of $G_{rr} + G_{qrnd}$ larger than 0.01). For the opposite direction of transitions from the red to the blue block the increase is less significant but still there are new transitions due to indirect pathways (2 significant transitions for G_{rr} and 3 for $G_{rr} + G_{qr}$). The significance of indirect transitions is also well visible in the friendship network of Figure 7 with many red arrows between the two blocks.

The same results for the original list, where the node *AKT3 Q9Y243 (+)* at position $i = 20$ has not been replaced by *PIK3CA P42336 (-)*, are available at [26].

5. Discussion

In this work we describe the properties of Google matrix analysis of the bi-functional SIGNOR PPI network from [2]. The main elements of this approach are: the activation and inhibition actions of proteins on each-other are described by Ising spin matrix transitions between the protein components in the doubled size Ising-PPI-network; the recently developed LIRGOMAX [11] algorithm determines the most sensitive proteins on the pathway between two selected proteins with probability injection (pumping) at one protein and absorption at another protein; the set of most sensitive proteins are analyzed by the REGOMAX algorithm which treats efficiently all direct and indirect interactions in this subset taking into account all their effective interactions

Table 4: Same as in Table 1 but for injection (pumping) at *EGFR P00533 (+)* and absorption at *PIK3CA P42336 (-)*.

i	K_L	K	Node name
1	1	203	BTK Q06187 (+)
2	2	19	AKT SIGNOR-PF24 (+)
3	3	14	AKT1 P31749 (+)
4	4	100	AKT2 P31751 (+)
5	5	63	MTOR P42345 (+)
6	6	24	PtsIns(3,4,5)P3 CID:24755492 (+)
7	7	80	IRS1 P35568 (-)
8	8	23	RAC1 P63000 (+)
9	10	330	PI3K SIGNOR-C156 (-)
10	11	9	PI3K SIGNOR-C156 (+)
11	38	1014	TEC P42680 (+)
12	39	970	BMX P51813 (+)
13	62	1587	ITK Q08881 (+)
14	63	154	PIK3CB P42338 (+)
15	65	1672	DAPP1 Q9UN19 (+)
16	66	1076	PLCG2 P16885 (+)
17	67	56	mTORC1 SIGNOR-C3 (+)
18	68	1196	GTF2I P78347 (+)
19	69	75	BAD Q92934 (-)
20	2806	1604	PIK3CA P42336 (-)
21	9	172	FES P07332 (+)
22	12	29	GRB2 P62993 (+)
23	13	90	EGFR P00533 (+)
24	14	3	PIK3CD O00329 (-)
25	15	136	EGFR P00533 (-)
26	16	126	CBL P22681 (+)
27	17	30	JAK2 O60674 (+)
28	18	57	STAT1 P42224 (+)
29	19	648	EZR P15311 (+)
30	20	456	GAB1 Q13480 (+)
31	21	424	BCR P11274 (-)
32	22	58	SHC1 P29353 (+)
33	23	746	SHC3 Q92529 (+)
34	24	2398	VAV2 P52735 (+)
35	25	40	JAK1 P23458 (+)
36	26	291	ERBB2 P04626 (+)
37	27	888	ERBB3 P21860 (+)
38	28	7	ERK1/2 SIGNOR-PF1 (+)
39	29	1028	JAK1/STAT1/STAT3 SIGNOR-C120 (+)
40	30	303	STAT1/STAT3 SIGNOR-C118 (+)

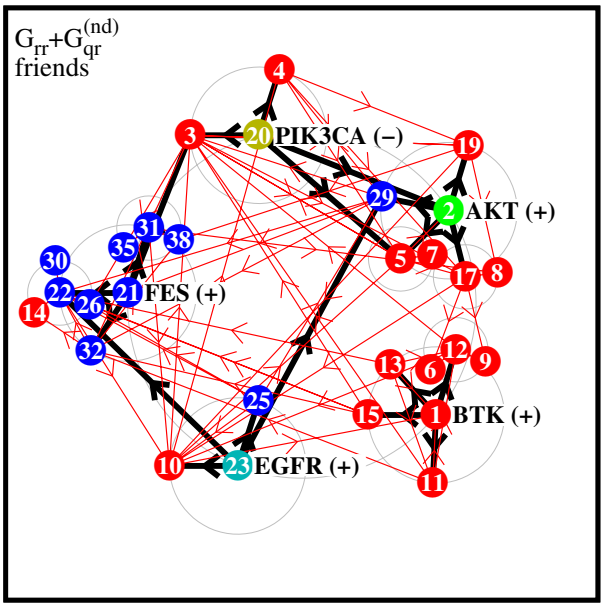

Figure 7: Same as Figure 4 but for the pathway of Table 4.

through the global PPI network. We illustrated the efficiency of this approach on several examples of two selected proteins. The obtained results show the efficiency of the LIRGOMAX and REGOMAX algorithms. We also show that the bi-functionality of protein-protein interactions leads to a certain effective magnetization of proteins which characterizes their dominant action on the global PPI network.

The executive codes and reduce Google matrix data are open and publicly available at [26] and interested researchers can easily study any example of a pathway between any pair of proteins from the SIGNOR network.

The described LIRGOMAX and REGOMAX algorithms can be applied also to other type of biological networks (e.g. metabolic networks discussed by [27]).

We mention that the described Google matrix algorithms have been tested for networks with 5 million nodes and thus they can operate efficiently on other PPI networks of significantly larger size (e.g. MetaCore network [28] which has several tens of thousands of nodes and about 2 million links). Thus we expect that the Google matrix approach, or in short Googlomics, will find broad applications for the analysis of protein-protein interactions.

6. Acknowledgments

This work was supported in part by the Programme Investissements d'Avenir ANR-11-IDEX-0002-02,

reference ANR-10-LABX-0037-NEXT (project THETRACOM). This work was granted access to the HPC resources of CALMIP (Toulouse) under the allocation 2019-P0110.

APPENDIX

A.1. Statistical properties of the SIGNOR protein-protein interactions network PPI

Using the Signor database a network of $N = 4341$ proteins with $N_\ell = 12547$ interactions was created. In a first version, called the "single protein network", the links do not contain the information if the interaction corresponds to activation, inhibition or is neutral/unknown. As usual we first construct an adjacency matrix with entries $A_{ij} = 1$ if there is a link from node $j \to i$ and $A_{ij} = 0$ if there is no such link. However, in certain rare cases there are multiple types of links between two proteins (e.g. activation and inhibition) in which case we choose A_{ij} being a multiplicity factor of 2 or 3 (instead of the usual entry 1). Once the adjacency matrix is fixed the Google matrix of this (single) protein network is constructed in the usual way: column sum normalization, taking into account the effect of dangling nodes (nodes with no outgoing link) by replacing each zero column by a uniform column with entries $1/N$ and with the application of the standard damping factor $\alpha = 0.85$.

In Fig. A.1 we show the PageRank P (CheiRank P^*) for this single network versus the corresponding rank index K (K^*) showing a typical decay (roughly) comparable to a power law $P \sim 1/K^\beta$ ($P^* \sim 1/(K^*)^\beta$) with $\beta \approx 0.7$ (0.8) for $K \geq 100$ ($K^* \geq 10$). Fig. A.2 shows the density of nodes in the PageRank-CheiRank plane (K, K^*) and the positions of the subgroup of nodes corresponding to Table 2 for this network.

To take into account the information about the nature of the links we use the approach of the Ising-PageRank to construct a larger network where each node is doubled with two labels (+) and (−). To construct the doubled "Ising" network of proteins each unit entry of the initial adjacency matrix is replaced by 2×2 matrices which are:

$$\sigma_+ = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad , \quad \sigma_- = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \quad , \quad \sigma_0 = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{A.1}$$

where σ_+ applies to "activation", σ_- to "inhibition" and σ_0 to "neutral" or "unknown". For the rare cases with multiple types of links between two proteins we use the sum of the corresponding σ matrices which increases the weight of the adjacency matrix elements. After this the corresponding Google matrix is constructed in the usual way. The doubled Ising protein network corresponds to $N_I = 8682$ nodes and $N_{I,\ell} = 27266$ links (according to the non-zero entries of the used σ matrices).

Now the PageRank vector (of this doubles Ising network) has components $P_+(j)$ and $P_-(j)$. Due to the particular structure of the σ matrices (A.1) one can show analytically the exact identity $P(j) = P_+(j) + P_-(j)$ where $P(j)$ is the PageRank of the initial single protein network. For this we have to replace in Eq. (4) of [20] the value n_i by n_{ij} with $n_{ij} = 1, 0, 1/2$ for the matrix σ_+, σ_- or σ_0 respectively. The additional dependence of n_{ij} on j takes into account that the choice of the σ matrix may be different for each link (and is not identical inside each row as it was the case for the model used in [20]. Then the analytical argument of this work also applies in exactly the same way to the case of the doubled Ising protein network. We have also numerically verified that the identity $P(j) = P_+(j) + P_-(j)$ holds up to numerical precision ($\sim 10^{-13}$).

As in [20] we introduce the PageRank "magnetization" by:

$$M(j) = \frac{P_+(j) - P_-(j)}{P_+(j) + P_-(j)} \tag{A.2}$$

for a node j. The dependence of $M(j)$ on nodes is shown in Fig. A.3 for the whole network and in Fig. 5 for the subgroup of nodes of Table 2.

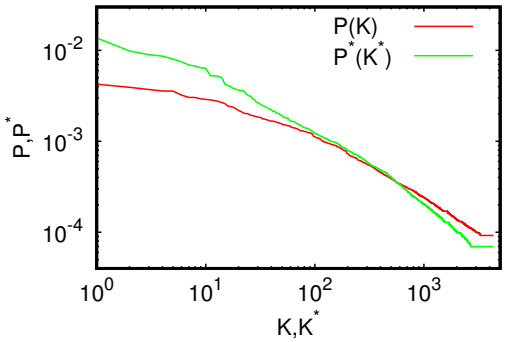

Figure A.1: PageRank $P(K)$ and CheiRank $P^*(K^*)$ for the single protein network.

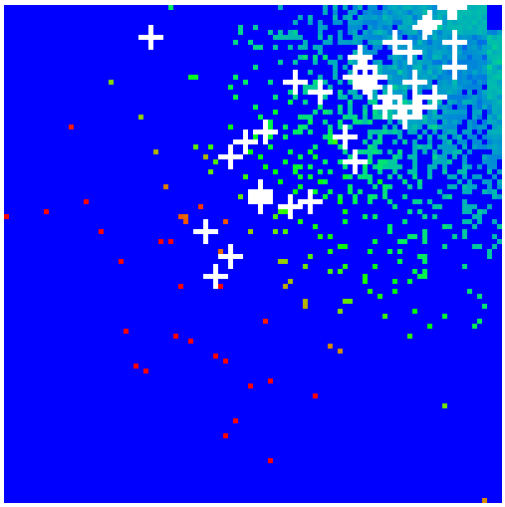

Figure A.2: Density of nodes $W(K, K^*)$ of the single protein network on PageRank-CheiRank plane (K, K^*) averaged over 100×100 logarithmically equidistant grids for $0 \leq \ln K, \ln K^* \leq \ln N$, the density is averaged over all nodes inside each cell of the grid, the normalization condition is $\sum_{K,K^*} W(K, K^*) = 1$. The color bar of Fig. 2 applies (for positive values) and its values correspond to $(W/\max W)^{1/4}$. In order to increase the visibility large density values have been reduced to (saturated at) 1/16 of the actual maximum density. The x-axis corresponds to $\ln K$ and the y-axis to $\ln K^*$. The white crosses show the positions of the 38 nodes of Table 2 and in Fig. 5.

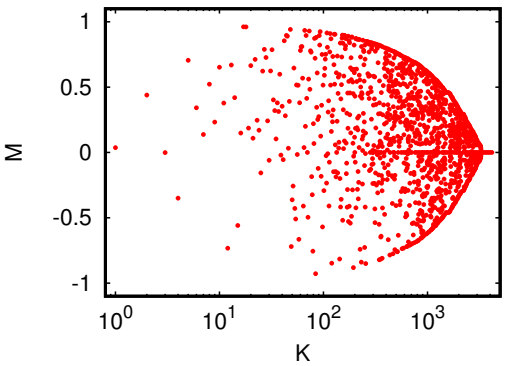

Figure A.3: PageRank "magnetization" $M(j) = (P_+(j) - P_-(j))/(P_+(j) + P_-(j))$ in the Ising-PPI-network; here j is the node index and $K(j)$ is the PageRank index of the initial SIGNOR network (without node doubling).

A.2. Pathway from *MAP2K1 Q02750 (+)* to *EGFR P00533 (-)*

Here we present additional figures and table for this pathway discussed in subsection 4.3. Table A.1 gives the proteins (extracted from Table 3) for which the magnetization M is presented in Figure A.7.

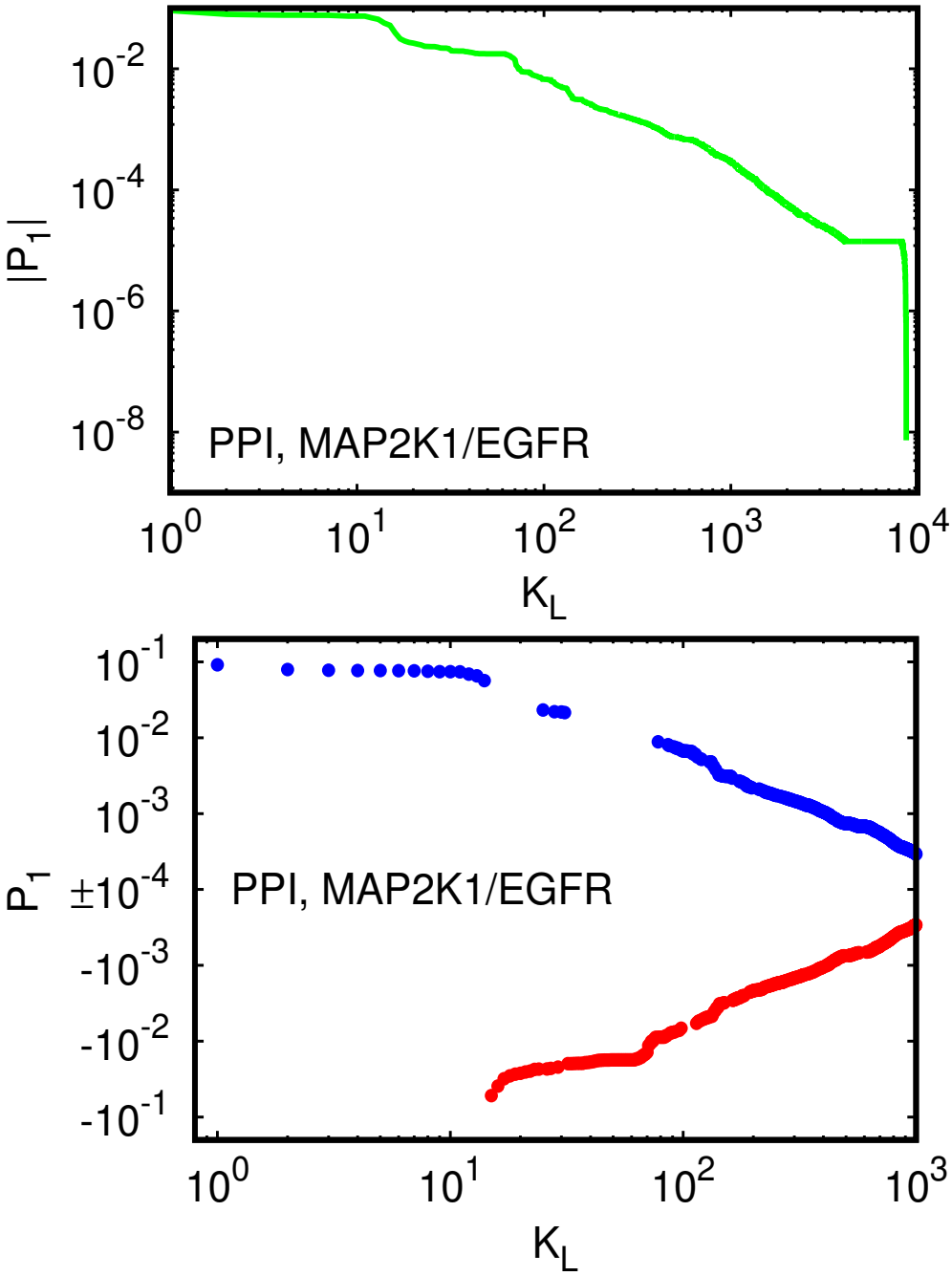

Figure A.4: Same as in Fig. 1 but for the pathway from *MAP2K1 Q02750 (+)* to *EGFR P00533 (-)*.

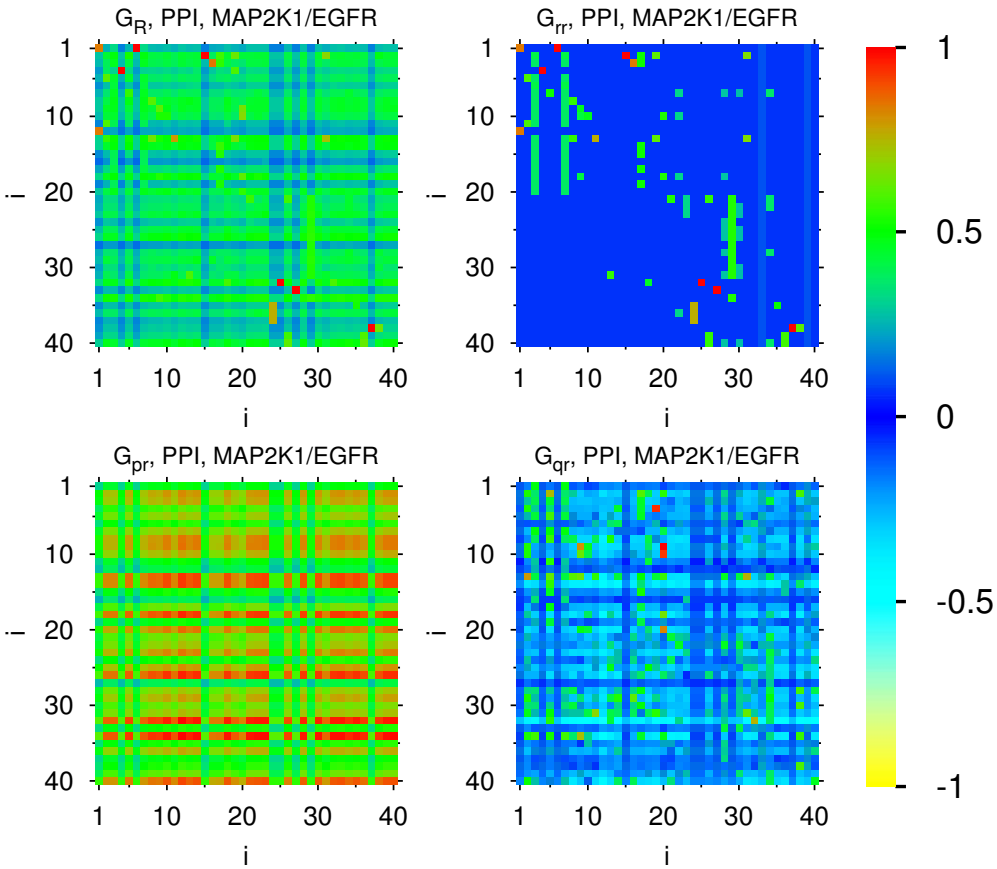

Figure A.5: Same as in Fig. 2 but for the pathway from *MAP2K1 Q02750 (+)* to *EGFR P00533 (-)*.

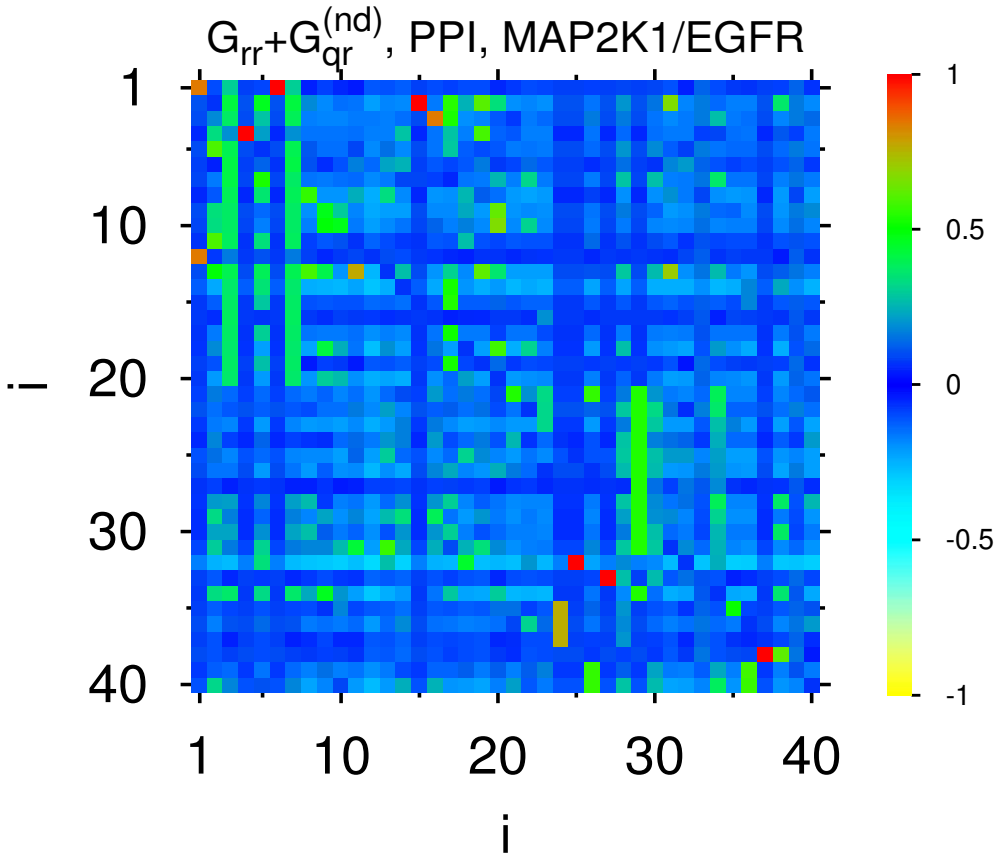

Figure A.6: Same as in Fig. 3 but for the pathway from *MAP2K1 Q02750 (+)* to *EGFR P00533 (-)*.

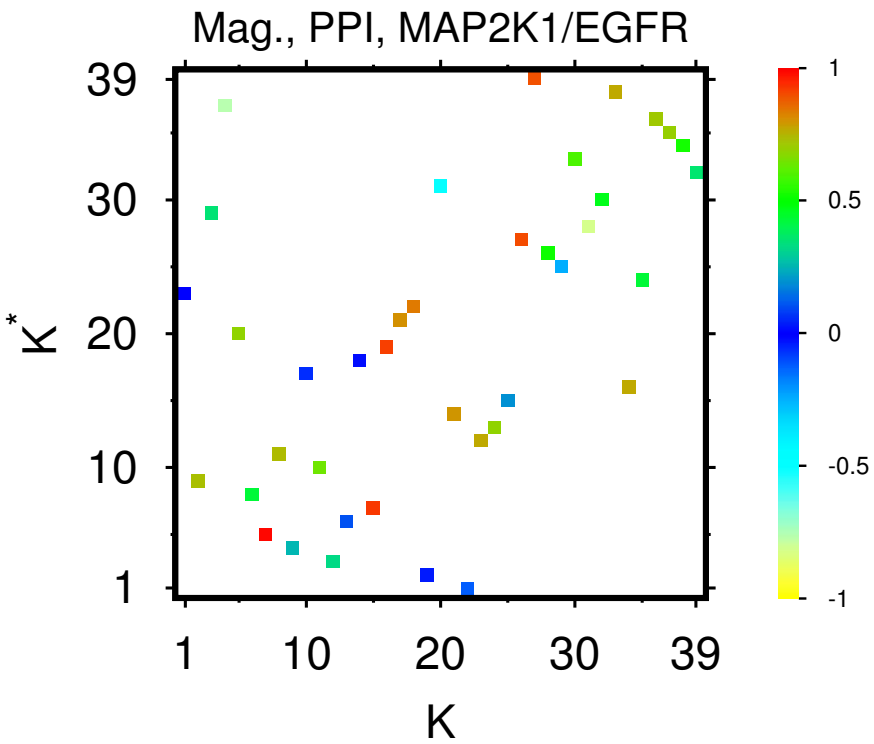

Figure A.7: Same as in Fig. 5 but for the pathway from *MAP2K1 Q02750 (+)* to *EGFR P00533 (-)* with proteins from Table A.1; the maximal magnetization used in the color bar normalization is M_{max} = 0.961

Table A.1: Same as in Table 2 but for injection (pumping) at *MAP2K1 Q02750 (+)* and absorption at *EGFR P00533 (-)*. The index i is the same as in Table 3 where two values do not appear here since they correspond to the two nodes where both components (+) and (−) are present in Table 3.

K	K^*	i	Node name
1	23	26	PPARG P37231
2	9	32	CASP3 P42574
3	29	25	CASP9 P55211
4	37	4	PIK3CD O00329
5	20	13	PI3K SIGNOR-C156
6	8	18	STAT3 P40763
7	5	34	ERK1/2 SIGNOR-PF1
8	11	40	JUN P05412
9	4	22	MAPK14 Q16539
10	17	36	MEF2C Q06413
11	10	9	JAK2 O60674
12	3	23	GSK3B P49841
13	6	3	EGFR P00533
14	18	21	CEBPA P49715
15	7	14	PLCG1 P19174
16	19	2	GRB2 P62993
17	21	8	PTK2 Q05397
18	22	20	JAK1 P23458
19	2	28	MAPK3 P27361
20	31	31	IRS1 P35568
21	14	10	STAT1 P42224
22	1	30	MAPK1 P28482
23	12	29	MAP2K1 Q02750
24	13	5	CBL P22681
25	15	17	ERBB2 P04626
26	27	1	FES P07332
27	39	39	CPT1B Q92523
28	26	38	TEK Q02763
29	25	24	TAL1 P17542
30	33	11	GAB1 Q13480
31	28	12	BCR P11274
32	30	6	EZR P15311
33	38	33	KIF3A Q9Y496
34	16	35	ERG P11308
35	24	19	ERBB3 P21860
36	36	15	SHC3 Q92529
37	35	37	ANGPT2 O15123
38	34	27	ARRB2 P32121
39	32	16	VAV2 P52735

A.3. Pathway from *EGFR P00533 (+)* to *PIK3CA P42336 (-)*

Here we present additional figures and table for this pathway discussed in subsection 4.3. Table A.2 gives the proteins (extracted from Table 4) for which the magnetization M is presented in Figure A.11.

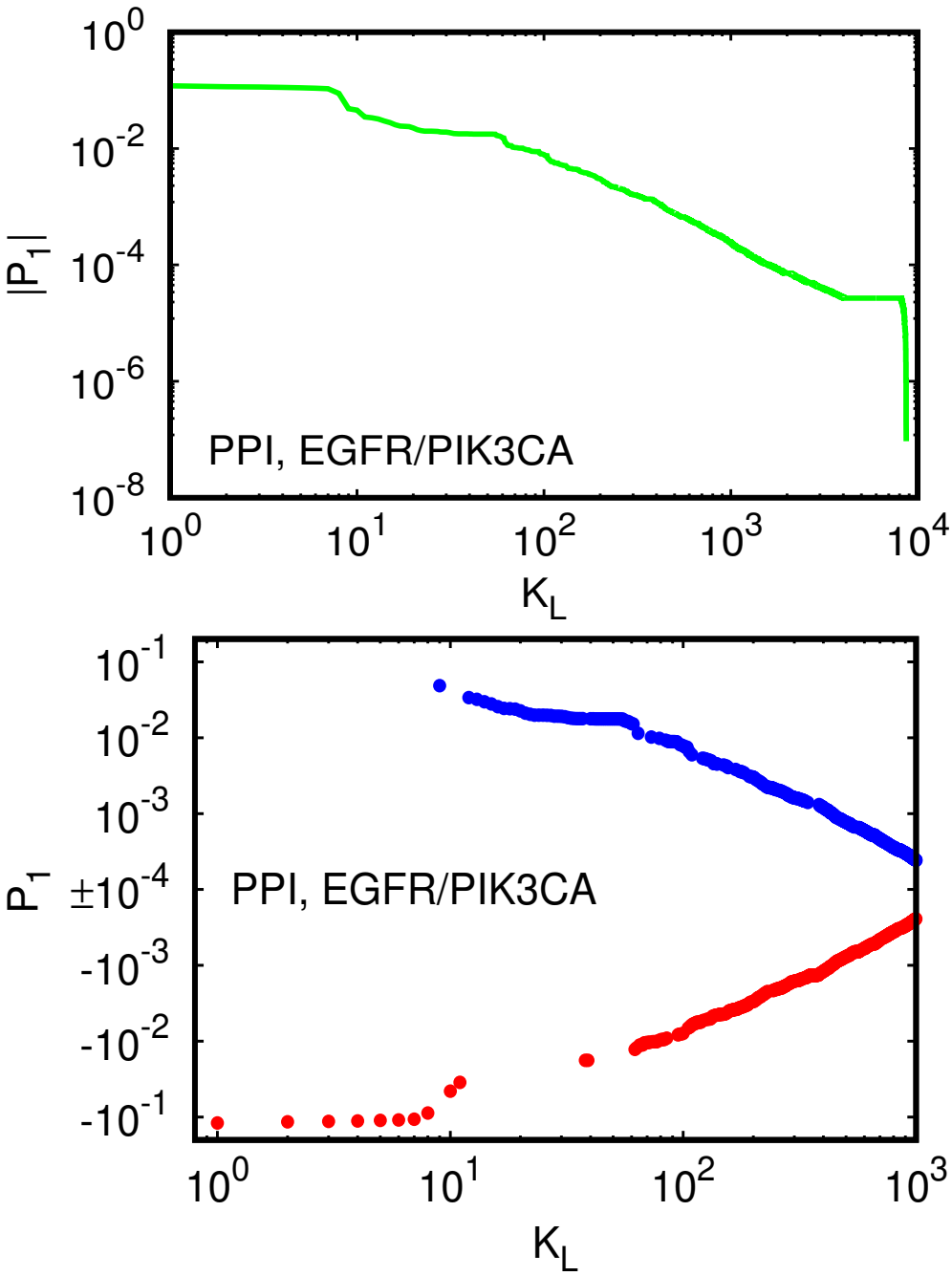

Figure A.8: Same as in Fig. 1 but for the pathway from *EGFR P00533 (+)* to *PIK3CA P42336 (-)*.

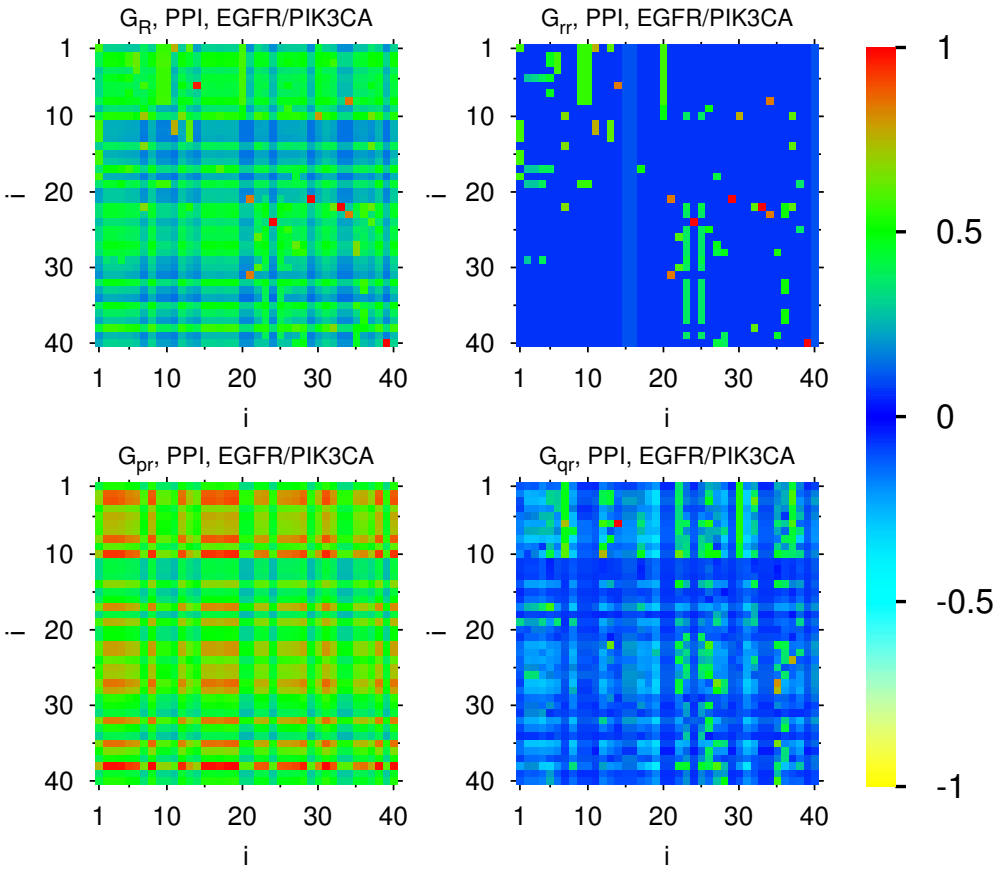

Figure A.9: Same as in Fig. 2 but for the pathway from *EGFR P00533 (+)* to *PIK3CA P42336 (-)*.

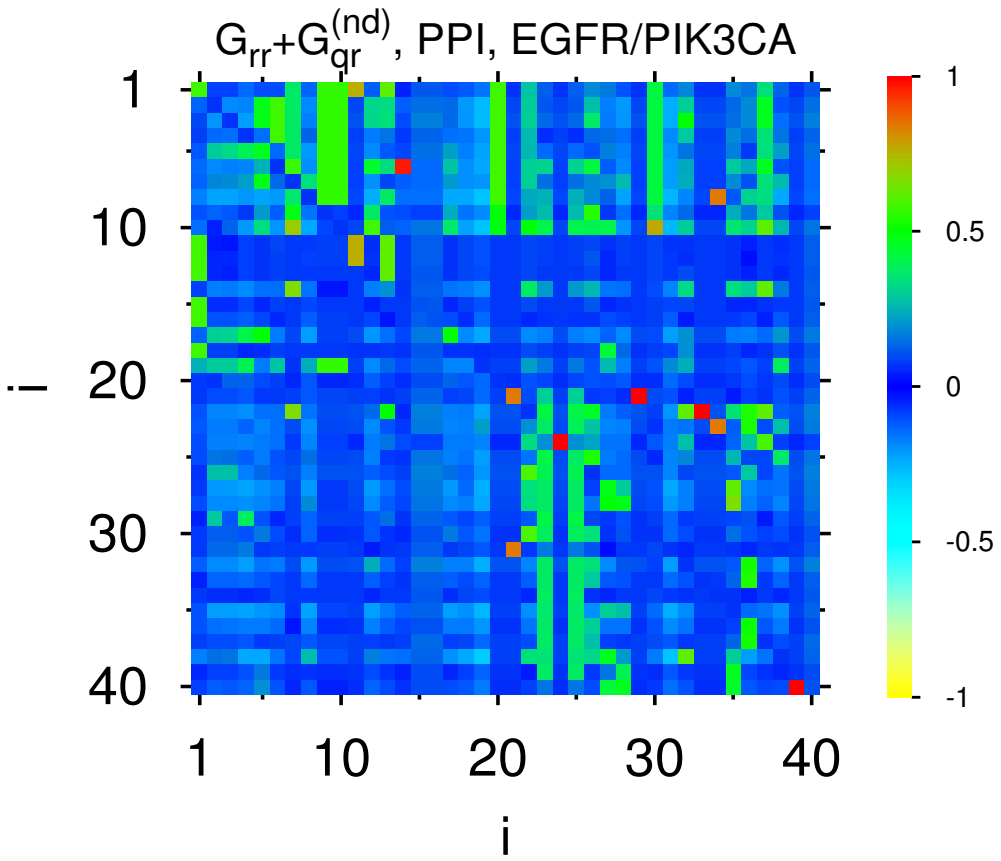

Figure A.10: Same as in Fig. 3 but for the pathway from *EGFR P00533 (+)* to *PIK3CA P42336 (-)*.

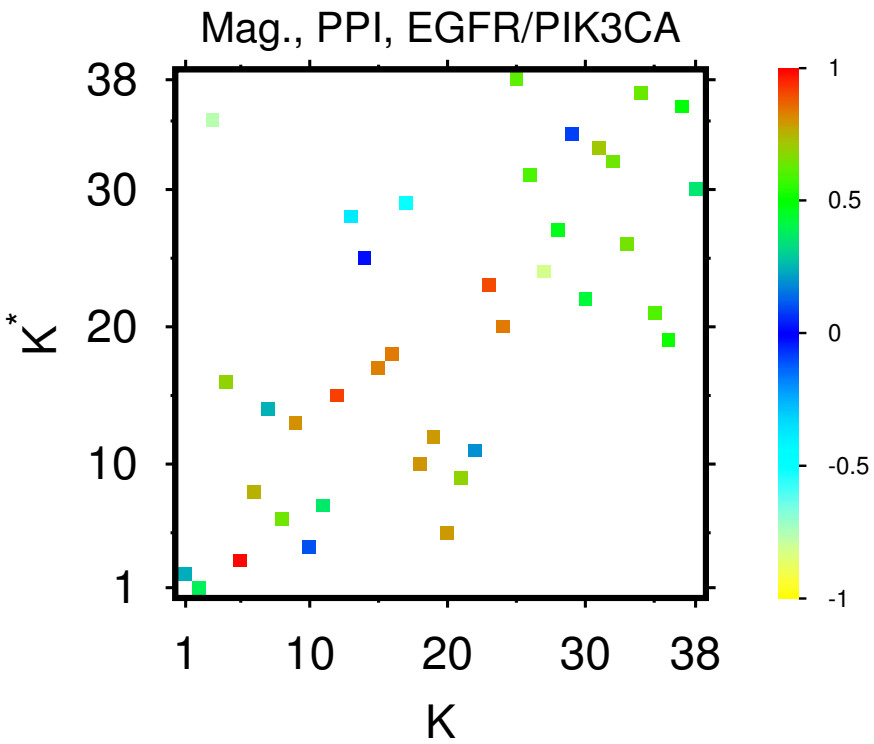

Figure A.11: Same as in Fig. 5 but for the pathway from *EGFR P00533 (+)* to *PIK3CA P42336 (-)* with proteins from Table A.2; the maximal magnetization used in the color bar normalization is $M_{max} = 0.961$

Table A.2: Same as in Table 2 but for injection (pumping) at *EGFR P00533 (+)* and absorption at *PIK3CA P42336 (-)*. The index i is the same as in Table 4 where two values do not appear here since they correspond to the two nodes where both components (+) and (−) are present in Table 4.

K	K^*	i	Node name
1	2	2	AKT SIGNOR-PF24
2	1	3	AKT1 P31749
3	35	24	PIK3CD O00329
4	16	9	PI3K SIGNOR-C156
5	3	38	ERK1/2 SIGNOR-PF1
6	8	6	PtsIns(3,4,5)P3 CID:24755492
7	14	17	mTORC1 SIGNOR-C3
8	6	27	JAK2 O60674
9	13	8	RAC1 P63000
10	4	23	EGFR P00533
11	7	5	MTOR P42345
12	15	22	GRB2 P62993
13	28	19	BAD Q92934
14	25	14	PIK3CB P42338
15	17	20	PIK3CA P42336
16	18	35	JAK1 P23458
17	29	7	IRS1 P35568
18	10	28	STAT1 P42224
19	12	32	SHC1 P29353
20	5	4	AKT2 P31751
21	9	26	CBL P22681
22	11	36	ERBB2 P04626
23	23	21	FES P07332
24	20	1	BTK Q06187
25	38	40	STAT1/STAT3 SIGNOR-C118
26	31	30	GAB1 Q13480
27	24	31	BCR P11274
28	27	29	EZR P15311
29	34	39	JAK1/STAT1/STAT3 SIGNOR-C120
30	22	37	ERBB3 P21860
31	33	33	SHC3 Q92529
32	32	12	BMX P51813
33	26	11	TEC P42680
34	37	16	PLCG2 P16885
35	21	18	GTF2I P78347
36	19	13	ITK Q08881
37	36	15	DAPP1 Q9UN19
38	30	34	VAV2 P52735

References

[1] A. Liberzon *et al.*, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27** (2011) 1739-1740.

[2] L. Perfetto *et al*. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* **44** Database issue (2016) D548-D554.

[3] M. Kanehisa *et al.*, KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45** Database issue (2017) D353-D361.

[4] A. Fabregat *et al*, The Reactome pathway knowledgebase. *Nucleic Acids Res.* **46** Database issue (2018) D649-D655.

[5] D.N. Splender *et al.*, The Reactome pathway knowledgebase. *Nucleic Acids Res.* **46** Database issue (2018) D661-D667.

[6] S. Dorogovtsev, *Lectures on complex networks.* (Oxford University Press, Oxford) 2010.

[7] S. Brin and L.Page, The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* **30** (1998) 107-117.

[8] A.M. Langville and C.D.Meyer, *Google's PageRank and beyond: the science of search engine rankings*, (Princeton University Press, Princeton) 2006.

[9] L. Ermann *et al*., Google matrix analysis of directed networks. *Rev. Mod. Phys.* **87** (2015) 1261-1310.

[10] K.M. Frahm *et al*., Wikipedia mining of hidden links between political leaders. *Eur. Phys. J. B* **89** (2016) 269.

[11] K.M. Frahm,K.M. and D.L. Shepelyansky, Linear response theory for Google matrix. *arXiv:1908.08924 [cs.SI]* (2019).

[12] C. Coquide *et al*. World influence and interactions of universities from Wikipedia networks. *Eur. Phys. J. B* **92** (2019) 3.

[13] C. Coquide *et al*., Influence of petroleum and gas trade on EU economies from the reduced Google matrix analysis of UN COMTRADE data. *Eur. Phys. J. B* **92** (2019) 171.

[14] F. Sacco *et al*., Deep proteomics of breast cancer cells reveals that metformin rewires signaling networks away from a pro-growth state. *Cell Systems* **2** (2016) 159-171.

[15] X.K. Lun *et al*., Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry. *Nature Biothech.* **35** (2017) 164-172.

[16] K. Kanhaiya *et al*., Controlling directed protein interaction networks in cancer. *Sci. Reports* **7** (2017) 10327.

[17] C. Dimitrakopoulos *et al*., Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34(14)** (2018) 2441-2448.

[18] B.M. Invergo and P. Beltrao, Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data. *Essays Biochemistry* **62** (2018) 525-534.

[19] J. Lages *et al*., Inferring hidden causal relations between pathway members using reduced Google matrix of directed biological networks. *PLoS ONE* **13(1)** (2018) e0190812.

[20] K.M. Frahm and D.L. Shepelyansky, Ising-PageRank model of opinion formation on social networks. *Physica A* **526** (2019) 121069.

[21] G. Bethune *et al*, Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J Thorac Dis.* **2** (2010) 48.

[22] T.N. Zamay *et al*., Current and prospective protein biomarkers of lung cancer. *Cancers MDPI* **9** (2017) 155.

[23] L. Cowen *et al*., Network propagation: a universal amplifier of genetic associations. *Nature Rev. Genetics* **18** (2017) 551-562.

[24] A.D. Chepelianskii, Towards physical laws for software architecture *arXiv:1003.5455 [cs.SE]* (2010).

[25] K.M. Frahm *et al*., Google matrix of the citation network of Physical Review. *Phys. Rev. E* **89** (2014) 052814.

[26] K.M. Frahm and D.L. Shepelyansky, Google matrix analysis for SIGNOR network. Available at *http://www.quantware.ups-tlse.fr/QWLIB/google4signornet/*. Accessed August (2019).

[27] C. Frainay *et al*., MetaboRank: network-based recommendation system to interpret and enrich metabolomics results. *Bioinformatics* **35(2)** (2019) 274-283.

[28] MetaCore (Clarivate Analytics). Available at *https://portal.genego.com/*. Accessed August (2019).