Collective intelligence defines biological functions in Wikipedia as communities in the hidden protein connection network

Andrei Zinovyev[1*+], Urszula Czerwinska[1], Laura Cantini[1,2], Emmanuel Barillot[1], Klaus M. Frahm[3], Dima L. Shepelyansky[3+],

1 Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, F-75005, Paris, France
2 Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, Paris Sciences et Lettres Research University, 75005 Paris, France
3 Laboratoire de Physique Théorique, IRSAMC, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

* corresponding author: andrei.zinovyev@curie.fr
+ co-senior authors

Abstract

English Wikipedia, containing more than five millions articles, has approximately eleven thousands web pages devoted to proteins or genes most of which were generated by the Gene Wiki project. These pages contain information about interactions between proteins and their functional relationships. At the same time, they are interconnected with other Wikipedia pages describing biological functions, diseases, drugs and other topics curated by independent, not coordinated collective efforts. Therefore, Wikipedia contains a directed network of protein functional relations or physical interactions embedded into the global network of the encyclopedia terms, which defines hidden (indirect) functional proximity between proteins. We applied the recently developed reduced Google Matrix (REGOMAX) algorithm in order to extract the network of hidden functional connections between proteins in Wikipedia. In this network we discovered tight communities which reflect areas of interest in molecular biology or medicine. Moreover, by comparing two snapshots of Wikipedia graph (from years 2013 and 2017), we studied the evolution of the network of direct and hidden protein connections. We concluded that the hidden connections are more dynamic compared to the direct ones and that the size of the hidden interaction communities grows with time. We recapitulate the results of Wikipedia protein community analysis and annotation in the form of an interactive online map, which can serve as a portal to the Gene Wiki project.

Introduction

Wikipedia is a unique knowledge resource containing a collection of approximately 5.5 millions articles in its English version, connected with each other by approximately 122 millions links (data from year 2017). Studying the large graph of Wikipedia hyperlinks with a focus on a particular subset of pages can provide interesting insights about certain topics. Thus, for example, Wikipedia networks were explored to establish the top historical figures of human history over 15 centuries [11], the geopolitical relations between countries [9], the leading world universities [7], world influence of infectious and cancer diseases [27, 28]. Hierarchical structure of Wikipedia was revealed through application of network community detection algorithms [20]. The variety of applications of Wikipedia in academic research was reviewed in [24, 26].

Wikipedia is a resource curated by a decentralized community effort (collective intelligence), which also includes semi-automated page generation from other structured resources. With time, automatically generated

pages are modified by the community and hyperlinked with the rest of the encyclopedia and external Internet. In this way, potentially any structured resource can be imported into Wikipedia, profit from continuous collective annotation by the Wikipedia editors and eventually become tightly embedded into the global Wikipedia knowledge graph.

Such an effort was made in the past for representing human genes in Wikipedia, the Gene Wiki project, `https://en.wikipedia.org/wiki/Gene_Wiki`. By 2008, a massive import of approximately 8000 gene-specific pages from Entrez Gene database was made, which boosted the community-based annotation of genes [17]. The initial set of protein page "stubs" have been complemented by adding the knowledge about protein-protein interactions, represented by hyperlinks between pages. Thus, in 2009, 3389 protein pages were connected by the 12628 most confident interactions from BioGrid database [16]. In 2011, it was estimated that 10369 protein pages in Wikipedia were annotated by 37578 PubMed citations with about 200 new citations added each month [15]. In 2009, Wikipedia protein pages of Wikipedia have been edited with a rate of approximately 1000 non-bot edits/month [16]. In 2016, the Gene Wiki project was complemented by the mechanism of Wikidata for better structuring the infoboxes of protein and gene pages [4]. As of today, Wikipedia pages related to proteins have become tightly integrated with the pages of common interest, describing diseases, drugs, biological functions, general culture phenomena. For example, the "BRCA1" page in Wikipedia is linked by such pages as "Oncogene", "Mastectomy", "Joe DiMaggio", "Breast cancer", "Carleton College", "DNA repair" (selected from the top 20 at `https://en.wikipedia.org/w/index.php?title=Special:WhatLinksHere/BRCA1`).

In this article, we were interested in studying how the knowledge about interactions between proteins is represented in Wikipedia. We also focused on how this knowledge is interconnected with the rest of the encyclopedia, serving a constantly updated corpus of annotation texts. Since the major bulk of direct protein-protein interactions has been automatically imported from existing databases of molecular interactions, the principal interest is in studying the topology of hidden, indirect connections between proteins through the rest of the Wikipedia graph. In order to study this topology, we used the recently developed methodology of reduced Google Matrix, which was already applied before for inferring hidden causal relations in a subnetwork of interacting proteins, embedded into a global network of protein-protein regulations [19].

The performed network analysis used the PageRank algorithm, which is at the foundation of the Google search engine [3, 21], and other properties of the Google matrix employed for analysis of various types of directed networks [12]. The recent approach of reduced Google matrix (REGOMAX) [13, 14] allows establishing indirect interactions between the selected nodes of interest taking into account all hidden pathways between these nodes via the remaining part of global network with a huge number of nodes. This REGOMAX algorithm originates from the scattering theory of nuclear and mesoscopic physics and field of quantum chaos.

Using the REGOMAX formalism, we characterized the topology of hidden connections between proteins in Wikipedia and identified the features distinguishing this topology from the network of direct connections. Following the recipe of the well-known proverb *Tell me who your friends are and I will tell you who you are* we characterize the function of specific hidden protein communities using their *friendship network* of links.

Results

WikiPedia networks of direct and hidden connections between proteins

We first listed all English Wikipedia page titles containing a description of a protein or gene and having a link with at least one other Wikipedia page. This resulted in 4899 protein page titles from the global Wikipedia network of $N = 5416537$ titles with $N_l = 122232932$ hyperlinks (for 2017 version and $N = 4212493$, $N_l = 101611732$ for 2013 version [8]). Using the global Wikipedia graph, we extracted the subnetwork of oriented direct hyperlinks between protein pages, which we will call in further the "network of direct links" between proteins. Then, we applied the reduced Google Matrix algorithm in order to quantify hidden links

between all pairs of proteins (see Methods). As a result, for each oriented protein pair, a weight was assigned representing the strength of the hidden connection through the rest of the Wikipedia network. We filtered out the hidden connections having small weights, by setting a threshold such that the resulting network would have the largest connected component (LCC) with the same average connectivity (number of nodes divided by the number of edges) as the LCC of the network of direct links. We will refer to the remaining links as "strong hidden connections". The resulting number of edges is indicated in Figure 1,B. Overall, strong hidden links tend to connect less proteins than the direct ones. Most of both direct and hidden links form one largest connected component (LCC) comprising more than 96-97% of total number of links. The number of direct and hidden links grew from 2013 to 2017 (in 4 years) of 14% and 32%, respectively, showing that the strong hidden connections form LCC increasing in size with time.

A noticeable structural difference between direct and hidden network concerns the number of their bidirectional links (when two protein pages point to each other reciprocally): 35% in the direct network and only 10% in the hidden one, see Figure 1. This might reflect the way the information about physical interactions between proteins was populated in Wikipedia, where the large part of interactions were considered non-oriented, so if protein A has protein B in the list of protein with which it interacts, then B should have A in its corresponding list.

Hidden protein links are explained by the existence of connected Wikipedia page sequences (paths) of hyperlinks through the rest of the Wikipedia network that connect two protein pages. We were interested in quantifying how many Wikipedia pages separated two protein pages associated via a strong hidden link. We found out that for the absolute majority of strong hidden links the shortest path length was equal to 2, compared to 3 or 4 for a randomly chosen protein pair, or to 5 and more for a randomly picked pair of Wikipedia pages (10000 page pairs have been sampled in order to estimate the distribution), see Figure 1,C. The shortest path length itself, however, is not a good measure of hidden link between two nodes of the graph, since it does not reflect the global topology of the Wikipedia network, e.g., the total number of shortest paths, connecting two nodes. By contrast, the weight of the link, estimated through the application of the reduced Google matrix, reflects the global topology of the rest of network and the probability to arrive from one protein page to an other one via a random walk through the rest of the graph.

Comparing the networks of direct and hidden protein links with existing pathway databases

We checked how many links extracted from Wikipedia matches known regulations or physical interactions between proteins, described in existing pathway databases. With this aim, we compared the reconstructed protein connection networks with two protein networks, resulted from the systematic collection of the protein interactions. One such database, SIGNOR, is characterized by a relatively small size and contains a set of highly confident interactions [25]. Another database, Pathway Commons [5], is an assembly of protein-protein interactions and regulations from multiple databases and computational predictions (including BioGrid [30]). Therefore, it is larger in size, but it potentially contains many spurious interactions, observed only in a certain context or predicted by computational biology methods. In order to compare interactions between networks, each Wikipedia protein page title was matched to a standard HUGO gene symbol.

The overlap between the links extracted from Wikipedia and the pathway databases was not very strong but highly significant. For SIGNOR, we found 1714 proteins in common with Wikipedia. These proteins were connected in SIGNOR by 4026 interactions from which we found 861 direct and 170 hidden links matched in Wikipedia. For Pathway Commons, we found 4768 proteins in common with Wikipedia protein page list. These proteins were connected in Pathway Commons by 269665 interactions from which we found 8212 direct and 2563 hidden links matched in Wikipedia. Taking as a null hypothesis that any two proteins from the Wikipedia network could be connected, this gives statistical significance for a Fisher's exact test with p-values of the order of 10^{-300}, in all comparisons. This overlap is, however, not completely surprising, given that a number of direct interactions between proteins were automatically imported from pathway databases. At the

same time, we found many direct and hidden connections between proteins from Wikipedia not found in existing databases, which reflects relative independence and non-redundancy of two sources.

We verified subsequently if the connectivity distribution for the nodes matched between Wikipedia network and pathway databases is similar. The comparison showed a significant correlation between the matched node connectivities (see Figure 1,D, which was much higher for the network of direct links (Pearson R=0.4 and Pearson R=0.58 for SIGNOR and Pathway Commons correspondingly) compared to the network of hidden links (Pearson R=0.18 and Pearson R=0.19 correspondingy). This correlation was determined, to a large extent, by the existence of common hubs in two types of networks. For example, BRCA1 was the top connected protein in the network of direct links form Wikipedia version 2017, and it ranks 35 in the top connected proteins in the network of Pathway Commons.

We further checked which interaction types are more present in those links which were matched between Wikipedia protein network and a pathway database. In order to do this, for each interaction type t, we first computed the fraction of matched interactions $f_t = I_t^W / I_t^{PD}$, where I_t^{PD} is the total number of links in a pathway database PD of type t and I_t^W is the number of these interactions matched in Wikipedia protein network. When computing I_t^{PD}, we limit the network only to those proteins common between Wikipedia and a pathway database. In order to compare direct and hidden connections, we used the relative fraction value $f_t^{rel} = f_t / \sum_t f_t$, which is shown in Figure 1, E. From this comparison it emerges that some interaction types have higher chance to be found in the Wikipedia network (e.g., "down-" or "up-regulates activity" interaction type in SIGNOR). We also detected a difference between direct and hidden interactions with respect to which interaction type they match more frequently. For example, for the interaction type "catalysis-precedes" of Pathway Commons there is almost three-fold increase in the relative frequency of match with hidden interactions, while for the "interacts-with" type the relative match frequency is much higher for direct interactions. Also, it seems that the hidden interactions between proteins in the Wikipedia network reveal more frequently co-participation of proteins in a complex, compared to direct interactions.

Community structure of the Wikipedia network of hidden connections between proteins

Simple visual inspection of the 2D force-directed layouts of networks of direct and hidden connections shows existence of relatively small scale compact communities in the network of hidden connections (Figure 2,A). We compared the two networks, using three network topology measures, namely connectivity distribution, average clustering coefficient distribution and average neighbourhood connectivity distribution (Figure 2,B-D). For all three measures, the networks of direct and hidden interactions resulted to be similar for the nodes with large (more than 20) number of neighbours. At the same time, nodes having smaller number of neighbours (less than 20), are characterized by larger local connectivity in the case of the network of hidden protein connections. This is particularly pronounced for the average clustering coefficient which equals 0.19 and 0.35 respectively for the direct and hidden connection networks, for the nodes having 10 neighbours (Figure 2,C). This analysis allowed us to conclude that the hidden protein connection network is characterized by the presence of communities, with a characteristic size of 10-20 protein pages. We thus hypothesized that these communities could be matched to the biological functions implicitly defined in Wikipedia through the community-based effort.

Annotated hidden protein connection community map

Following our conclusion about the presence of small communities in the hidden network, we clustered the network of hidden protein connections using Markov Cluster Algorithm [10]. 274 and 289 communities were identified in the network of hidden protein connections computed from the English Wikipedia graph from 2013 and 2017 correspondingly (only those communities having size at least 4 pages were kept in this phase). The maximum community size was 148 and 187 correspondingly for 2013 and 2017 Wikipedia versions. Despite this size of the largest community, overall the others resulted to be smaller with average community size 8.6

and 9.7 in 2013 and 2017 correspondingly.

Using HUGO symbols matched to the Wikipedia protein page titles, we performed function enrichment analysis for all communities using ToppGene [6]. The results of this analysis are available online at [1]. We found that most of the communities had clear enrichment in one of the biological functions or in a biological pathway. Thus, the geometric mean q-value of the most significant enrichment in a Gene Ontology-related term was 10^{-19} (for community sizes in at least 10 proteins), and in a Pathway term it was 10^{-16}. The exceptionally large community with 187 nodes (2017 version) had enrichment in Gene Ontology terms "cytokine activity" (q-value=10^{-30}), "leukocyte proliferation" (q-value=10^{-55}), "adaptive immune response" (q-value=10^{-47}), pathway terms "Cytokine-cytokine receptor interaction" (q-value=10^{-50}), "Hematopoietic cell lineage" (q-value=10^{-39}) and other multiple immune system-related terms. It was also strikingly enriched in the MSigDB HALLMARK [22] gene sets: e.g., "Genes up-regulated during transplant rejection" (q-value=10^{-52}), "Genes defining inflammatory response "(q-value=10^{-28}).

Alternatively to the use of enrichment analysis, the biological function defined by a community could be identified by looking at the direct connections through neigbouring Wikipedia pages. For each link inside the community we extracted titles from the global Wikipedia graph along the shortest oriented paths of length 2 connecting the connected pair of proteins. This defined an augmented community network, with Wikipedia pages corresponding not only to protein pages but also to the Wikipedia titles through which the shortest paths had gone through. An example of such an augmented network is shown in Figure 2,E. We ranked the set of nodes by their local connectivity in the augmented network, and used the most connected page title for labeling the community. For example, the augmented network shown in Figure 2,E was labeled in this way as "Coagulation". Also, among the most strongly locally connected nodes there were such titles as "Haemophilia", "Warfarin", "Heparin", "Liver" and others. The three most connected proteins in the augmented network shown in Figure 2,E were "Protein C", "Thrombin", "Factor X".

Following this strategy, we annotated each community by the Wikipedia title, having the largest local connectivity in the augmented network. In some cases, we manually changed this title, selecting among the 10 most connected titles, which would have a better match for the enriched biological function. Afterwards, we counted the number of hidden links between the nodes in each community from the initial network of hidden protein connections. In this way we constructed an abstracted graph of communities and oriented links between them, which we visualized in Cytoscape [29], using force-directed layout, Figure 3. This map shows the repertoire of biological functions described in English Wikipedia by groups of pages forming relatively compact subnetworks in the global graph of hyperlinks. As one can see, the central place in this map is taken by "Immune system", "Apoptosis", "Cell cycle", "Insulin/Glycolysis", "Mitogen-activated protein kinase kinase", "Cell migration" and other communities which correspond to well studied biological functions. There exist relatively large protein page communities collecting proteins characterized by a presence of a particular domain such as "CARD domain", "RING finger domain", "SH2/SH3 domain", "C2 domain". Interestingly, the map is characterized by a meaningfull hierarchy of functions. For example, 4 communities annotated by the names of the major DNA repair pathways ("Non-homologous end joining", "Nucleotide excision repair", "Base excision repair", "Fanconi anemia") point to the large community annotated as "DNA repair".

We provide the hidden protein connection community map in interactive form, using NaviCell Google Maps-based platform for annotated network visualization [2,18]. The online map of hidden protein interactions in Wikipedia is available from `http://navicell.curie.fr/pages/maps_wikipediacommunity.html`. The map can be queried for a protein name or a part of the Wikipedia page title. All community node annotations are hyperlinked to the corresponding Wikipedia pages. Therefore, the interactive map serves as a convenient portal to the set of Wikipedia pages related to proteins and associated pages. Moreover, the map can be used for molecular data visualization, using NaviCell data analysis toolbox and binding to major programming languages (R, Python, Java) [2]

Evolution of WikiPedia protein network between 2013 and 2017

We compared the changes in the direct and hidden protein connections, between two versions of English Wikipedia (2013 vs 2017). We found that 96% of direct connections did not change in four years, while only 71% hidden connections remained unchanged in the same period of time (Figure 4),A. This indicates that the Wikipedia network of protein connections evolves more slowly through the curation of pages devoted to proteins compared to more dynamic modifications of the information in the associated pages from the network neighbourhood (for example, pages describing molecular mechanisms of diseases or pages devoted to the systematic description of protein families).

From the reduced Google matrix analysis we know the relative PageRanks of proteins which were not exactly the same between two Wikipedia versions, despite good overall correlation (Figure 4,B). Thus, we found that a significant number of proteins strongly improved their PageRanks in 2017 (Supplementary File 1). For example, MGMT gene changed its PageRank from 1856 to 174 (more than ten-fold) and FANCB gene changed its PageRank from 3240 to 351 (almost ten-fold). Such drastic changes in PageRanks might indicate recent interest in studying these genes which led to intense curation of the associated pages.

We verified if these proteins were enriched in a particular biological function. To answer this question, 181 proteins whose PageRank improved more than two-fold were tested for enrichment in reference gene sets using ToppGene. In the top of the list of the enriched categories we found such Gene Ontologies as "fibroblast growth factor receptor binding" ($p=10^{-6}$), "damaged DNA binding" ($p=10^{-5}$), "response to radiation" ($p=10^{-10}$), "aging" ($p=10^{-9}$), "DNA repair complex" ($p=10^{-7}$), "transcription factor complex" ($p=10^{-6}$). Among MSigDB signatures, the top enriched was "Genes involved in DNA repair, compiled manually by the authors" (19 genes, $p=10^{-10}$). Overall, it shows significant recent editing efforts in the part of Wikipedia related to DNA repair, which led to higher hidden connectivity between pages in this area. At the same time, 46 genes loosing their PageRank position more than 2-fold did not show any strong enrichment in Gene Ontologies or other reference gene sets (e.g., none has passed the corrected p-value threshold 0.001).

We matched the communities obtained in 2013 and 2017 versions of Wikipedia by computing the Jaccard index for the overlap between the set of the genes composing them. We defined a match, if the Jaccard index was reciprocally maximal between two community sets aka it is done for defining orthologous genes in evolutionary bioinformatics [31]. Overall, 189 communities could be matched in this way with a minimum threshold for the intersection in 3 proteins. We observed a consistent increase between the matched community sizes between 2013 and 2017 versions, starting from the community size in 10 protein pages, Figure 4,D.

The abstracted map of hidden protein connection communities shows emergence of some communities in the 2017 version of English Wikipedia which can not be matched in 2013 version. Examples are "RING finger domain", "SWI/SNF", "ATPase", "Bcl-2 family", "Integrin", "Fanconi anemia" communities (Figure 3). Hypothetically, this also indicates an active curation efforts happening between 2013 and 2017 in the Wikipedia pages related to these functions or the pages directly connected to them.

Materials and Methods

Direct network of protein connections

Global network of links between English Wikipedia pages was extracted using in-house web crawler, for 2013 and 2017 years [Dima, please give details]. Protein and gene pages have been identified by querying for the presence of "Infobox protein" and "Infobox gene" Wikipedia templates in the page text. Those pages not having any outgoing or incoming links have been filtered out.

Google matrix construction

The Google matrix G of a directed network with N nodes (titles) and N_l hyperlinks is constructed from the adjacency matrix A_{ij} with elements 1 if node (title) j points to title (node) i and zero otherwise. The matrix elements have the standard form $G_{ij} = \alpha S_{ij} + (1-\alpha)/N$ [3, 12, 21] 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 the damping factor. We use the standard value $\alpha = 0.85$ [21] noting that for the range $0.5 \leq \alpha \leq 0.95$ the results are not sensitive to α [12, 21]. For a random surfer, moving from one title to another, the probability to jump to any title is $(1-\alpha)$.

The right PageRank eigenvector of G is the solution of the equation $GP = \lambda P$ for the unit eigenvalue $\lambda = 1$. The PageRank $P(j)$ values give positive probabilities to find a random surfer on a node j ($\sum_j P(j) = 1$). We order all nodes 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 algorithm described in [3, 21].

Reduced Google matrix algorithm

The REGOMAX algorithm is described in detail in [13, 14]. It allows to compute efficiently a "reduced Google matrix" of size $N_r \times N_r$ that captures the full transitions of direct and indirect pathways happening in the full Google matrix between N_r nodes of interest.
For the selected N_r nodes their PageRank probabilities remain 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 $G_{\rm R}$ provides a decomposition of $G_{\rm R}$ into matrix components that clearly distinguish direct from indirect interactions: $G_{\rm R} = G_{rr} + G_{\rm pr} + G_{\rm qr}$ [13]. Here G_{rr} is given by the direct links between selected N_r nodes in the global G matrix with N nodes. In fact, $G_{\rm pr}$ is rather close to the matrix in which each column is given by the PageRank vector P_r, ensuring that PageRank probabilities of $G_{\rm R}$ are the same as for G (up to a constant multiplier). Hence $G_{\rm pr}$ does not provide much information about direct and indirect links between selected nodes. The most nontrivial and interesting role is played by $G_{\rm qr}$, which takes into account all indirect links between selected nodes appearing due to multiple pathways via the global network nodes N. The exact formulas for all three components of $G_{\rm R}$ are given in [13, 14].

The efficiency of the REGOMAX approach has been demonstrated for various Wikipedia networks [7, 9, 13, 27, 28], protein networks from SIGNOR database [19], and the multiproduct world trade network from UN COMTRADE database [7].

All matrix data and PageRank vectors for the reduced Google matrix of $N_r = 4899$ proteins are available at [1] for Wikipedia versions of 2013 and 2017, together with the global Wikipedian networks.

Network of hidden protein connections

The network of hidden protein connections is obtained from the component $G_{\rm qr}$ of the reduced Google matrix $G_{\rm R}$ by keeping only matrix elements being larger than a certain critical cutoff value. This value if determined from the condition of having the same connectivity value in the Largest Connected Components of both networks.

Defining hidden communities by clustering

For the networks of hidden protein connections we applied Markov Clustering Algorithm (MCL) implemented in ClusterMaker plugin for Cytoscape [23] with default parameters (granularity=2.0, edge weight cutoff=1.0, number of iterations=16, maximum residual value=0.001).

Functional enrichment analysis

The functional enrichment analysis was performed using ToppGene [6] and recapitulating the results in the form of an interactive web-page, available at [1]. In the automatically produced summary of the enrichment results for each hidden protein community, one of the reference set per category is displayed but only if it overlapped with the query set in at least $k = 5$ genes and only if the corrected for multiple testing q-value did not exceed $s = 10^{-8}$.

Discussion

We studied the network of protein-protein interactions embedded into the graph of hyperlinks between Wikipedia pages. We focused on comparing direct hyperlinks between protein pages (most of which were automatically imported from existing molecular interaction databases) and hidden links through the rest of the Wikipedia graph. The hidden links were identified by using the reduced Google Matrix approach.

The most striking conclusion from this analysis is the existence of pronounced small-scale (10-20 proteins on average) clusters (communities) in the network of hidden protein connections. The absolute majority of these clusters have rather clear biological meaning which was quantified by the functional enrichment analysis. This is in contrast with the previous conclusions about the power law-like distribution of connectivity of protein pages in the global Wikipedia network. Existence of such clusters (communities) points out to emergence, due to collective intelligence of Wikipedia editors, of relatively well defined groups of proteins sharing the common biological function (such as cell cycle), structural feature (such as SH2/SH3 domain) or other common topics. These clusters are generally not present and can not be deduced from the network of direct interactions.

Interestingly, one can easily deduce the biological function of the community by looking at the titles of the pages most tightly connected in the the augmented network of pages linking the community proteins. Using this labeling, we created an abstracted interactive online map of connections between the protein communities, which can serve a portal to the Gene Wiki Wikipedia project.

We characterized the evolution of the network of hidden protein connections and its community structure between two snapshots of Wikipedia in 2013 and 2017 years. We showed that the nature of hidden protein connections is much more dynamic compared to the direct links. A clear trend has been noticed on the faster relative increase of the number of hidden connections such that they combine more proteins in one largest connected component. Interestingly, we show that there are more proteins that drastically (by few folds) improved their PageRanks in 2017 compared to those who drastically lost their PageRanks inside the global Wikipedia network. We found that the Wikipedia topics being improved in connectivity were related to DNA repair and damage. Most of the hidden connection network communities between 2013 and 2017 can be matched in terms of maximally reciprocal Jaccard index quantifying their intersection. We show that the matched communities have larger size on average in 2017 compared to the 2013 network.

Altogether, these observations indicate increasing integration of the Gene Wiki project into the global Wikipedia context, a trend which will certainly persist in the future. It remains an interesting question what can be a practical use of of the protein function definition derived from the Wikipedia structure. Another interesting question is how to use the insights obtained from analysing the topology of hidden protein connections, in order to guide further evolution of the Gene Wiki project. For example, it would be interesting to identify missing biological functions or topics which do not yet form tight clusters in the Wikipedia network.

Acknowledgements

This research is supported in part by the MASTODONS-2016/2017 CNRS project APLIGOOGLE (see http://www.quantware.ups-tlse.fr/APLIGOOGLE/) and in part (for KMF and DLS) by the Pro-

gramme Investissements d'Avenir ANR-11-IDEX-0002-02, reference ANR-10-LABX-0037-NEXT (project THETRACOM).

Supplementary Files

The Supplementary files are available from [1].
Supplementary File 1 - Table containing computed PageRanks of Wikipedia protein pages within the reduced network, definitions of hidden protein connection communities.
Supplementary File 2 - Cytoscape sessions containing networks of direct and hidden connections between proteins, in 2013 and 2017.

References

1. Wikiprotein networks web-page, http://www.quantware.ups-tlse.fr/QWLIB/wikiproteinnets/ (2019).
2. E. Bonnet, E. Viara, I. Kuperstein, L. Calzone, D. Cohen, E. Barillot, and A. Zinovyev. NaviCell Web Service for network-based data visualization. *Nucleic Acids Research*, 43(W1), 2015.
3. S. Brin, L. Page, S. Brin, and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
4. S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. Putman, J. Leong, C. Naik, P. Pavlidis, L. Schriml, B. M. Good, and A. I. Su. Wikidata as a semantic framework for the Gene Wiki initiative. *Database*, 2016:baw015, 2016.
5. E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(Database issue):D685–90, 2011.
6. J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(Web Server):W305–W311, 2009.
7. C. Coquidé, J. Lages, and D. L. Shepelyansky. World influence and interactions of universities from Wikipedia networks. *The European Physical Journal B*, 92(1):3, 2019.
8. D.L.Shepelynasky. Wikipedia networks: quantware articles and data sets, http://www.quantware.ups-tlse.fr/QWLIB/wikinets/ (2017).
9. S. El Zant, K. Jaffrès-Runser, and D. L. Shepelyansky. Capturing the influence of geopolitical ties from Wikipedia with reduced Google matrix. *PLoS ONE*, 13(8):e0201397, 2018.
10. A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–84, 2002.
11. Y.-H. Eom, P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky. Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. *PLoS ONE*, 10(3):e0114825, 2015.
12. L. Ermann, K. M. Frahm, and D. L. Shepelyansky. Google matrix analysis of directed networks. *Reviews of Modern Physics*, 87(4):1261–1310, 2015.
13. K. M. Frahm, K. Jaffrès-Runser, and D. L. Shepelyansky. Wikipedia mining of hidden links between political leaders. *The European Physical Journal B*, 89(12):269, 2016.
14. K. M. Frahm and D. L. Shepelyansky. Reduced google matrix. arXiv:1602.02394 [physics.soc], 2016.

15. B. M. Good, E. L. Clarke, L. de Alfaro, and A. I. Su. The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Research*, 40(D1):D1255–D1261, 2012.

16. J. W. Huss, P. Lindenbaum, M. Martone, D. Roberts, A. Pizarro, F. Valafar, J. B. Hogenesch, and A. I. Su. The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Research*, 38(suppl_1):D633–D639, 2010.

17. J. W. Huss, C. Orozco, J. Goodale, C. Wu, S. Batalov, T. J. Vickers, F. Valafar, A. I. Su, and A. I. Su. A gene wiki for community annotation of gene function. *PLoS biology*, 6(7):e175, 2008.

18. I. Kuperstein, D. Cohen, S. Pook, E. Viara, L. Calzone, E. Barillot, and A. Zinovyev. NaviCell: A web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Systems Biology*, 7, 2013.

19. J. Lages, D. Shepelyansky, and A. Zinovyev. Inferring hidden causal relations between pathway members using reduced Google matrix of directed biological networks. *PLoS ONE*, 13(1), 2018.

20. A. Lancichinetti, M. I. Sirer, J. X. Wang, D. Acuna, K. Körding, and L. A. N. Amaral. High-Reproducibility and High-Accuracy Method for Automated Topic Classification. *Physical Review X*, 5(1):011007, 2015.

21. A. N. Langville and C. D. C. D. Meyer. *Google's PageRank and beyond : the science of search engine rankings*. Princeton University Press, 2012.

22. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. Mesirov, and P. Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 2015.

23. J. H. Morris, L. Apeltsin, A. M. Newman, J. Baumbach, T. Wittkop, G. Su, G. D. Bader, and T. E. Ferrin. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12(1):436, 2011.

24. F. Å. Nielsen. Wikipedia Research and Tools: Review and Comments. *SSRN Electronic Journal*, 2012.

25. L. Perfetto, L. Briganti, A. Calderone, A. Cerquone Perpetuini, M. Iannuccelli, F. Langone, L. Licata, M. Marinkovic, A. Mattioni, T. Pavlidou, D. Peluso, L. L. Petrilli, S. Pirrò, D. Posca, E. Santonico, A. Silvestri, F. Spada, L. Castagnoli, and G. Cesareni. SIGNOR: a database of causal relationships between biological entities. *Nucleic acids research*, 44(D1):D548–54, 2016.

26. J. M. Reagle. *Good faith collaboration : the culture of Wikipedia*. MIT Press, 2010.

27. G. Rollin, J. Lages, and D. Shepelyansky. Wikipedia network analysis of cancer interactions and world influence. *bioRxiv*, 527879, 2019. https://www.biorxiv.org/content/early/2019/01/23/527879.

28. G. Rollin, J. Lages, and D. L. Shepelyansky. World Influence of Infectious Diseases From Wikipedia Network Analysis. *IEEE Access*, 7:26073–26087, 2019.

29. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003.

30. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535–9, 2006.

31. R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science (New York, N.Y.)*, 278(5338):631–7, 1997.

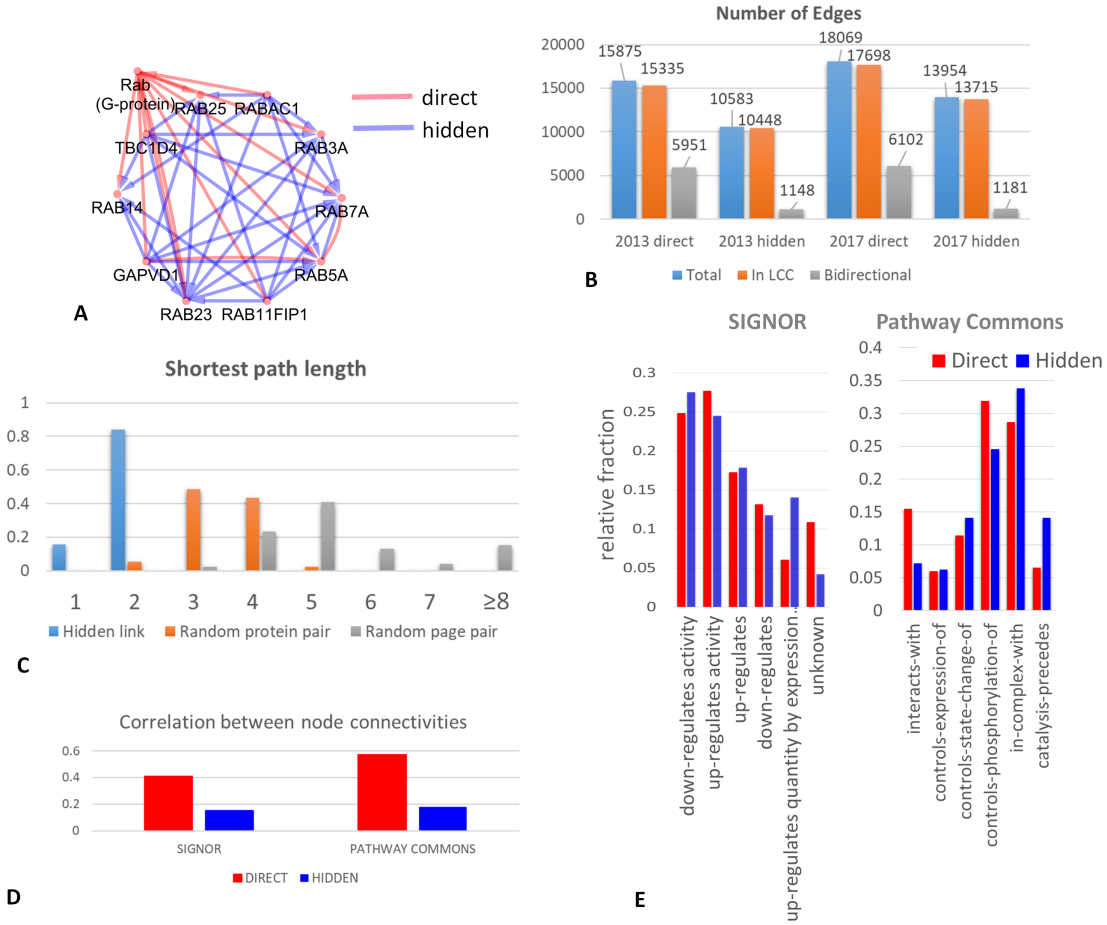

Fig 1. Characterizing the networks of direct and hidden connections between proteins in the global Wikipedia network. A) Example of a distinction between direct and hidden interactions. The page "Rab (G-protein)" links to a set of pages for the proteins from the family. Hidden interactions (in blue) connects the whole family into an almost complete clique graph, making them a tight community. B) Number of edges in the networks of direct and hidden protein connections. Number of bidirectional links is separately show. C) Quantifying the shortest path length in the global Wikipedia network between proteins connected by a hidden link, random protein pair and random Wikipedia page pair. D) Correlation between node connectivities in the networks of direct (or hidden) protein connections extracted from Wikipedia and two pathway databases (SIGNOR and Pathway Commons). E) Relative fraction of link types found in the networks extracted from Wikipedia (direct and hidden) and in two pathway databases (SIGNOR and Pathway Commons).

.

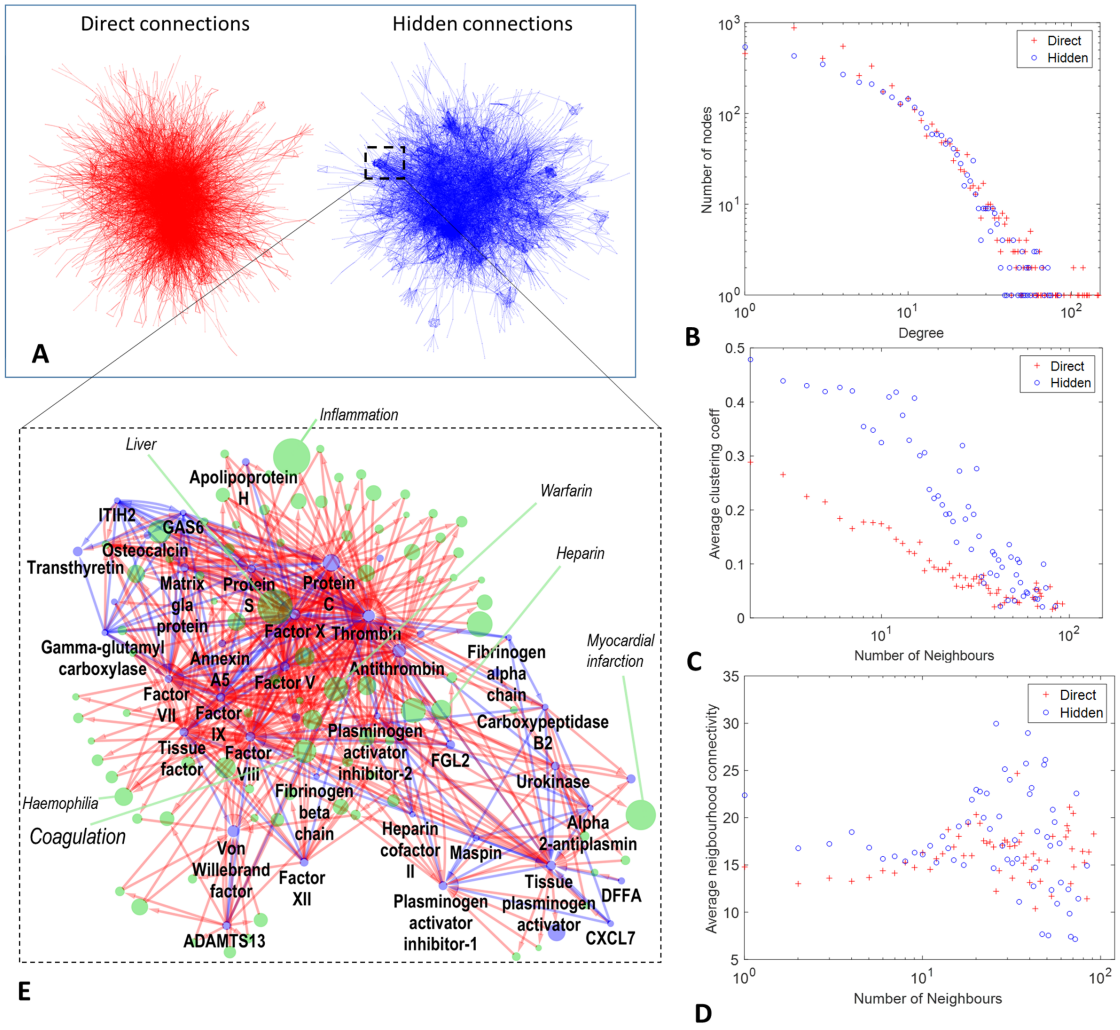

Fig 2. Network of hidden interactions is characterized by relatively well defined communities as compared to the network of direct interactions. A) Force-directed layouts of the networks of direct and hidden connections. B-D) Comparison of two networks in terms of connectivity, average clustering coefficient, average neighborhood connectivity distribution. E) One of the communities in the network of hidden interactions is shown together with direct links to the Wikipedia pages connecting the protein pages.

.

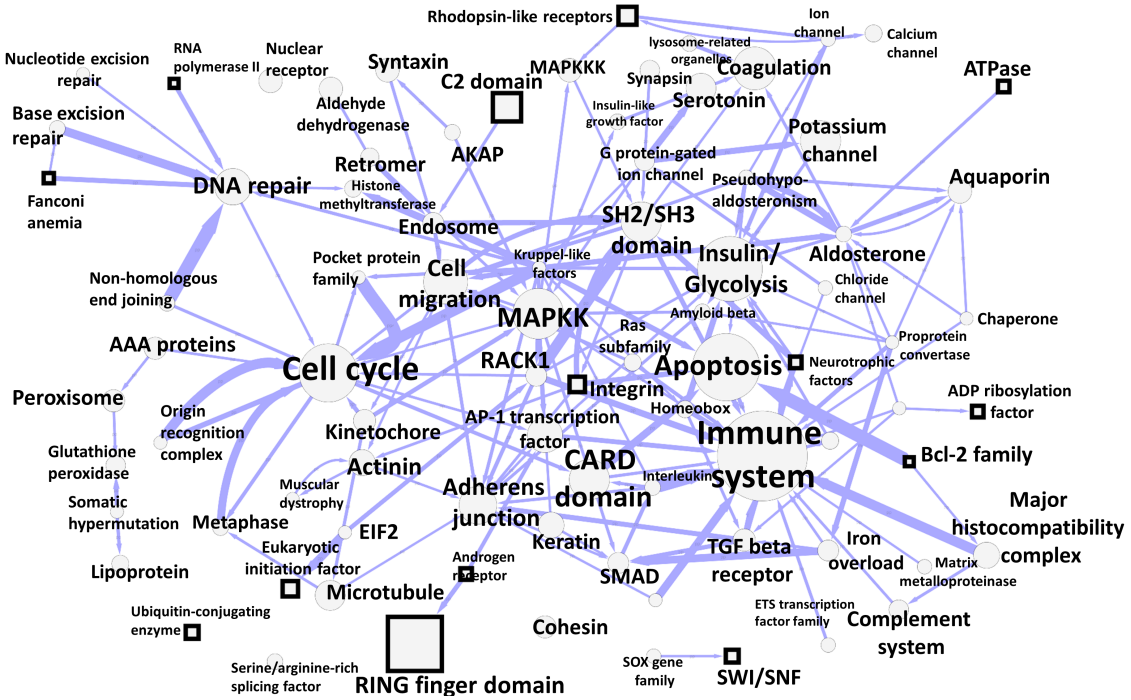

Fig 3. Map of connections between communities of hidden protein connections through the global Wikipedia network. The size of the nodes reflect the number of proteins in the community. Square nodes signify those communities identified in the 2017 Wikipedia network which do not have a match in the 2013 Wikipedia network. The width of the line reflects the number of oriented links between the communities. The node is labeled by the title of the Wikipedia page associated with the community by the connector links, having the strongest internal connectivity. The interactive version of this map is available from `http://navicell.curie.fr/pages/maps_wikipediacommunity.html`.

.

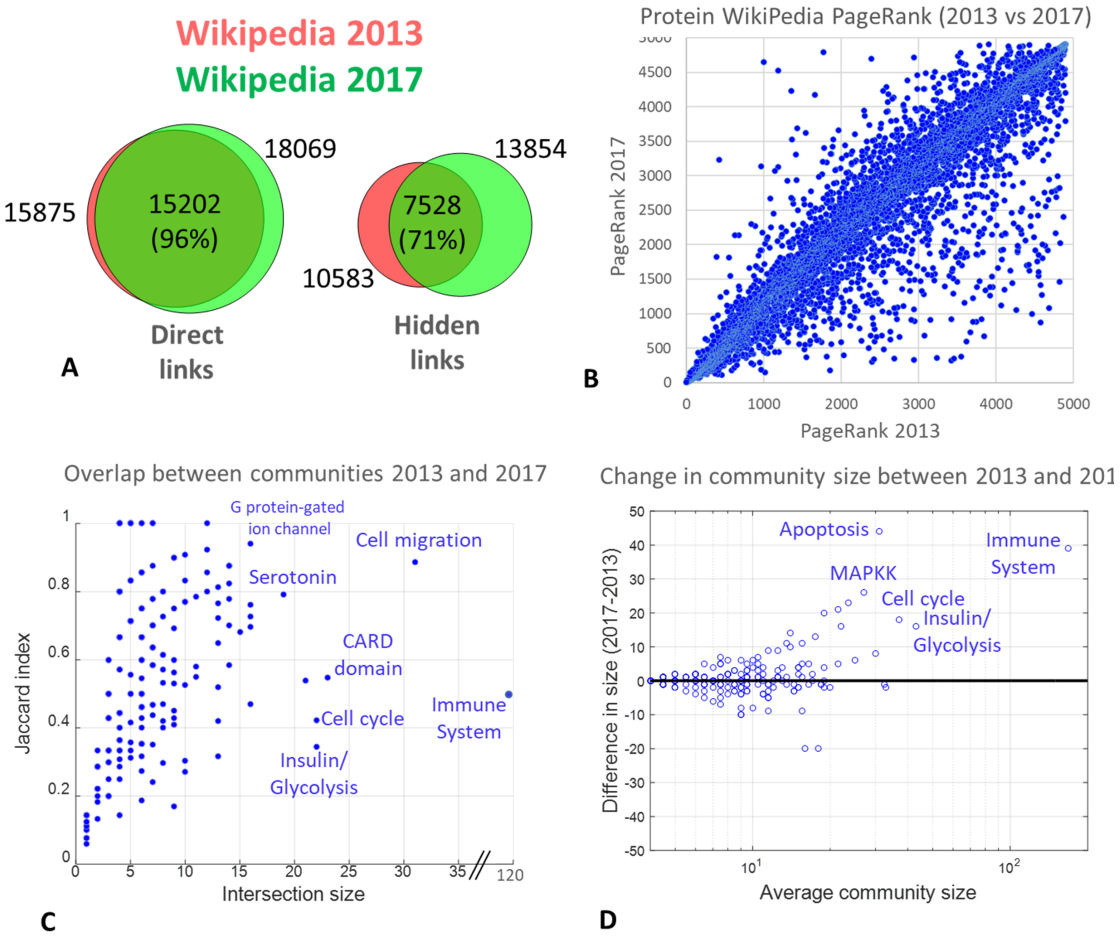

Fig 4. Evolution of the networks of protein connections within the global Wikipedia network between 2013 and 2017. A) Overlap between links for the network of direct and hidden protein connections in two versions of Wikipedia. B) Changes in the PageRanks in the reduced Google matrix for protein pages, compared between 2013 and 2017. C) Overlap between matched communities of hidden protein connections extracted from two versions of Wikipedia. D) Change in matched hidden interaction community size between two versions of Wikipedia.

.