Integrative Biology

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c4ib00125g

Network wiring of pleiotropic kinases yields insight into protective role of diabetes on aneurysm

Anida Sarajlić,[a] Vladimir Gligorijević,[a] Djordje Radak[b] and Nataša Pržulj*[a]

Received 1st June 2014,
Accepted 28th July 2014

DOI: 10.1039/c4ib00125g

www.rsc.org/ibiology

Recent studies suggest a protective role of diabetes in the development of aneurysm, but the biological mechanisms behind this are still unknown. This type of association is not present in the case of diabetes and atherosclerosis despite similar risk factors for aneurysm and atherosclerosis. We postulate the existence of genes that disrupt the pathways needed for the onset of aneurysm in the presence of diabetes. Motivated by the significance of genetic interactions in understanding disease–disease associations, we tackle this problem by integrating protein–protein interaction and genetic interaction data, *i.e.*, we examine the biological pathways related to the three diseases that contain genes involved in the following genetic interactions: one gene in a genetic interaction is part of a diabetes pathway, the other gene is part of an aneurysm, or an atherosclerosis pathway. We create a protein–protein interaction sub-network that contains disease pathways described above. We then use a "brokerage" measure – a topological measure that identifies proteins in this sub-network whose removal severely affects the interconnectedness of their neighbourhood, enabling such proteins to disrupt the pathway they are in. We identify a set of proteins with high brokerage values and find this set to be enriched in biological functions, including cell-matrix adhesion, which facilitates mechanisms that have already been suggested as possible causes of diabetes–aneurysm association. We further narrow down our set to 16 proteins that are involved in an aneurysm or an atherosclerosis pathway and are encoded by genes participating in genetic interactions with a gene in a diabetes pathway. This set is enriched in kinases and phosphorylation processes, with two pleiotropic kinases that are involved in both aneurysm and atherosclerosis pathways. Kinases can turn on or off proteins, explaining how functional changes of such proteins could result in the disruption of pathways. So if in an aneurysm-related pathway a gene is turned off, the onset of the disease could be prevented. However, mutations of pleiotropic genes could have effects only on one of the traits, which explains why pleiotropic kinases that are involved in both aneurysm and atherosclerosis pathways could disrupt aneurysm pathways explaining the reduced risk of aneurysm in diabetes patients, but not affect the atherosclerosis pathways.

Insight, innovation, integration

The main contribution of this paper is identification of 16 genes that uncover biological mechanisms behind the relationship between aneurysm, atherosclerosis and diabetes. We address this problem using a computational approach and topology analysis of molecular interaction networks. We use high-throughput molecular network data to create a disease sub-network (consisting of pathways related to the three diseases) by integrating protein–protein interaction and genetic interaction data. Then, we use computational methods to analyse the topology of the disease sub-network, resulting in finding genes responsible for the relationship between the three diseases. In particular, we use the Simmelian brokerage measure to identify genes with such local topology that can explain how these genes can disrupt the disease pathways.

1 Introduction

Abdominal aortic aneurysm (AAA) is a permanent dilatation of the abdominal aorta and a leading cause of death amongst the population of older men.[1] Several studies suggest the protective role of diabetes in the development of aneurysm.[2,3] De Rango *et al.* showed that progression of small AAA is 60% lower in patients who suffer from diabetes.[3] Prakash *et al.* also confirmed that diabetes is associated with a decreased rate of hospitalization due to thoracic aortic aneurysms (TAAD).[2] This seems paradoxical, as diabetes is known to predispose cardiovascular diseases: peripheral, coronary, and cerebrovascular diseases.[3,4]

[a] *Department of Computing, Imperial College London, 180 Queen's Gate, SW7 2AZ London, UK. E-mail: natasha@imperial.ac.uk; Tel: +44-(0)207-594-1516*
[b] *Institute for Cardiovascular Diseases Dedinje, Belgrade, Serbia*

Also, vascular diseases are the principal cause of death and disability in people with diabetes and a common macro-vascular manifestation for this is atherosclerosis.[5] Note that atherosclerosis shares similar risk factors with aneurysm, such as male gender, increasing age, hyperlipidemia, and hypertension, and as such was considered as an underlying pathogenesis in AAA.[1,6] However, a decreased prevalence of AAA in patients with diabetes may suggest that atherosclerosis is an associated feature and not a cause of aneurysm.[1] Hence, we explore possible mechanisms behind the protective role of diabetes in the development of aneurysm and why there is no similar diabetes–atherosclerosis association, as published work in this area is still inconclusive.[3] Therefore, to tackle this problem we use high-throughput molecular network data.

A number of large-scale biological data sets exist as a result of recent advances in high-throughput techniques. Large-scale molecular data include information on interactions among biological macromolecules and metabolites, such as protein–protein interactions (PPI), genetic interactions, enzyme–substrate relationships and pathway maps. Network representations of such interaction data enable graph theoretic approaches to be applied to help identify topological properties which are different from that expected at random, revealing the connection between a specific topological characteristic and a related biological function or phenotype, such as disease. PPI networks, where nodes correspond to proteins and edges are placed between two proteins if they physically interact, are networks with commonly explored topology. It was shown that proteins that are closer in the PPI network are more likely to perform the same function,[7] which was later used for inferring functions of unannotated proteins: the direct neighbourhoods of proteins,[8] n-neighbourhoods of proteins,[9] and shared neighbours of proteins[10] were examined looking for the most common functions among annotated direct neighbours. The local topology around a protein in a PPI network was summarized into a topological ''signature'' of a protein – *graphlet degree vector* (GDV),[11] and the similarity of these protein ''signatures,'' or GDV similarity, is a good indicator of proteins belonging to the same protein complex, performing the same biological function, that are coexpressed, that are involved in the same diseases, and that are part of the same sub-cellular components.[11] This topological measurement of similarity was used for predicting new melanogenesis related genes that were phenotypically validated.[12] There is a growing trend in studying the topology of molecular networks that yield insights into human disease. For example, networks in cardiovascular disease have recently been used as a platform for better understanding of the complexity behind the disease.[13–15]

We use both the human PPI network and the genetic interaction network to find an explanation for the protective role of diabetes in aneurysm and why a similar relationship is not present in the case of diabetes and atherosclerosis. We hypothesize that a functional change of a protein on a pathway that is important for aneurysm could disrupt this pathway preventing the onset of the disease. We suspect that a mutation of a gene in a pathway involved in diabetes is related to a functional change of a protein in an aneurysm-related pathway, explaining the protective role of diabetes in the development of aneurysm. This is why we integrate PPI data with information from the human genetic interaction network. In a genetic interaction network nodes correspond to genes in the network and edges represent functional associations between them: an interaction between two genes occurs when the result of simultaneous mutations in the genes is not just a combination of phenotypes of single mutations.[16] It has been shown that genetic interactions are critical for understanding disease evolution[17] and a key to capturing disease–disease associations from molecular interaction data.[18] Although by definition a genetic interaction between two genes does not indicate a direct interaction, it can indicate how strongly the function of one gene depends on the presence of the other, *i.e.*, it can indicate how much is the phenotype of one mutation modified by the presence of the second mutation.[19] Even the order in which mutations occur is in some cases likely defined by the genetic interactions.[17] One such example in cancer progression is when P53 dysfunction usually precedes BRCA loss of function generating synthetic viability.[17] In the case of genetic interaction between genes whose protein products directly interact, a mutation in one protein that affects a physical interaction can be compensated by a mutation of its interacting partner, for example, proteins S12 and L19 in *Salmonella typhimurium*.[20]

The methodology of our study is presented in Fig. 1.

We first identify pathways that play a role in the formation of the three diseases, as described in Section 4.1. Then, we use information from the human genetic interaction network to single out pathways that contain genes (henceforth, we use terms protein and gene interchangeably), which take part in genetic interactions such that one interacting gene is part of a diabetes-related pathway while the other is part of an aneurysm- or an atherosclerosis-related pathway. We use selected pathways to create a disease-related sub-network of the human PPI network, as described in Section 4.2.

In search of genes whose change in functionality could disrupt a pathway, we rely on the network topology and look for genes in this disease PPI sub-network with a local topology that could explain a gene's high ''destructiveness'' for the related pathway. The Simmelian brokerage measure[21] captures the cohesion of a neighbourhood of a node and measures the importance of the node for maintaining interconnectedness of its neighbourhood. Using this measure, as described in the Methods, we identify a set of ''broker'' genes and find this set to be statistically significantly enriched in biological functions that facilitate mechanisms that have already been suggested as possible causes of diabetes–aneurysm association. We narrow down this set to 16 genes that are on aneurysm- or atherosclerosis-related pathways and participate in genetic interactions with genes from diabetes-related pathways. We find this set to be enriched in kinases and in the biological function of phosphorylation. This confirms our hypothesis that identified proteins could disrupt the pathways, in particular, kinases can switch on and off proteins on an aneurysm-related pathway, which can lead to prevention of aneurysm formation. Importantly, two kinases from the set that are on both aneurysm- and

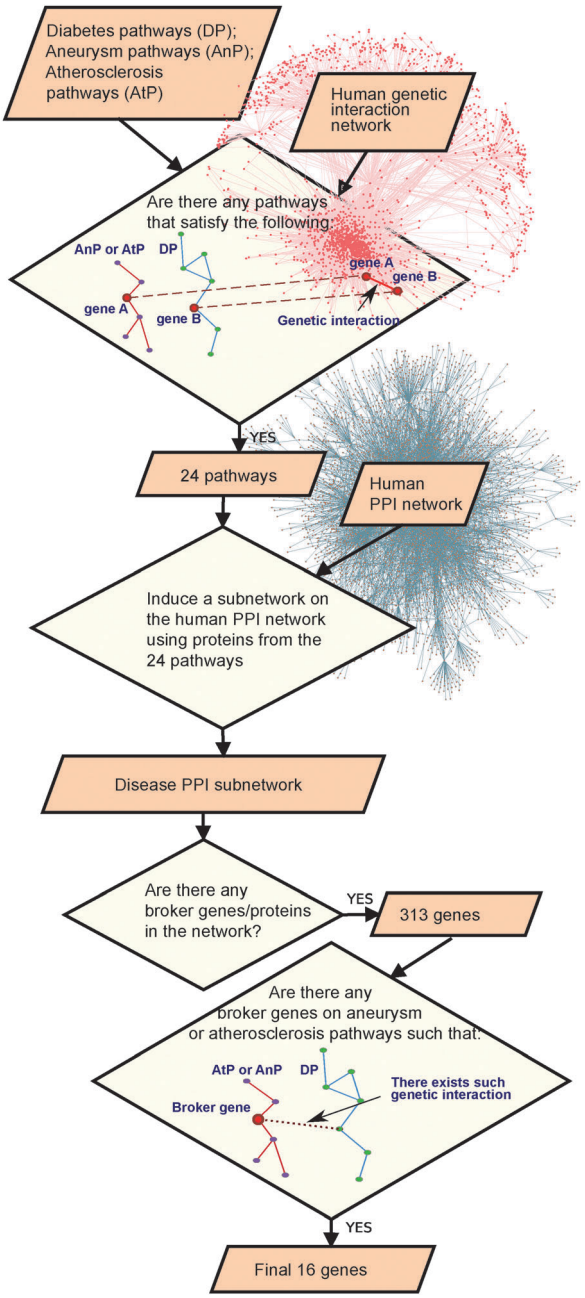

Fig. 1 Work-flow of the study.

atherosclerosis-related pathways are pleiotropic, explaining why a mutation of such genes could disrupt an aneurysm-related pathway but not affect the atherosclerosis-related pathway.

2 Results and discussion

In addition to known diabetes pathways recovered from KEGG, we identify pathways significant for each of the two diseases, aneurysm and atherosclerosis, by examining which KEGG pathways are statistically significantly enriched in genes related to the diseases. Note that the same pathways can be involved in several diseases. For example, the cytokine–cytokine receptor interaction pathway hsa04060 is enriched both in aneurysm and atherosclerosis genes (see Table 1). This is not specific to diseases that we study here as it is well known that some pathways are involved in many diseases, *e.g.* the MAPK signaling pathway has been involved in many human diseases including Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis and various types of cancer.[22] Out of the pathways related to the three diseases, we examine the 24 pathways that contain at least one gene involved in genetic interactions with one interacting gene on a diabetes pathway and the other on an aneurysm or atherosclerosis pathway. The 24 pathways, together with their KEGG IDs, are listed in Table 1.

Using these pathways we create the disease PPI sub-network as described in the section Disease PPI sub-network. In search of genes that can compromise pathways, we rely on the disease PPI sub-network topology to find *broker genes* that can disrupt their neighbourhood's interconnectedness. We describe such a gene property using the Simmelian brokerage measure[21] (detailed in Section 4.3). To our knowledge, this measure is the only topological measure that quantifies the importance of a node for maintaining the connectivity between its neighbouring nodes. High brokerage of a node implies high topological importance for the connectivity between nodes in its neighbourhood. In other words, if the functionality of a protein that has a high brokerage score would be altered, this would influence the interconnectedness of the protein's neighbourhood, which in our disease PPI sub-network is a part of the pathway in which this protein plays a role.

We find brokers in the disease PPI sub-network by identifying statistically significant brokerages of nodes in the disease PPI sub-network, as described in Section 4.4. Bins with statistically significant p-values (<0.01) are presented in Fig. 2.

We suspect that identified broker genes, due to their importance for the interconnectedness of their neighbourhoods in the disease PPI sub-network, can lead to disabling signal transduction, or completion of chain reactions in the pathways.

In the disease PPI sub-network we find 313 proteins with statistically significant brokerage. Using the DAVID[23,24] database we examine their functional enrichment and find this set to be enriched in a number of GO biological processes including phosphorylation (p-value = 5.3×10^{-31}), as well as vascular development (p-value = 5.1×10^{-10}) and regulation of cell-matrix adhesion (p-value = 3.9×10^{-10}). Cell-matrix adhesion, *i.e.*, binding of a cell to the extracellular matrix (ECM), plays an important role in regulation of many processes, such as cell adhesion, tissue homeostasis, and wound healing.[25] Matrix metalloproteinases (MMPs), proteolytic enzymes, exhibit increased activity in the human aneurysmal tissue.[1] MMP-2, which is among the 313 genes, takes part in the breakdown of the matrix proteins, including elastin, and therefore influences degradation of the vessel wall in aneurysm. However, in diabetes, there is a reduced degradation of the matrix that results in an increased matrix volume.[26] Concentrations of MMP-2 and MMP-9 are reduced in coronary arteries of diabetic patients and it has been postulated that the reduction of MMPs activity can slow down the matrix

Table 1 The 24 pathways containing genes that participate in specific genetic interactions

Pathway name	KEGG ID	Disease
Colorectal cancer	hsa05210	An
MAPK signaling pathway	hsa04010	An
Viral myocarditis	hsa05416	An
Type I diabetes mellitus	hsa04940	D, At
Pathways in cancer	hsa05200	An, At
Vascular smooth muscle contraction	hsa04270	An
Type II diabetes mellitus	hsa04930	D
Maturity onset diabetes of the young	hsa04950	D
Cytokine–cytokine receptor interaction	hsa04060	An, At
Dilated cardiomyopathy	hsa05414	At
Graft-*versus*-host disease	hsa05332	At
Systemic lupus erythematosus	hsa05322	At
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	hsa05412	At
Focal adhesion	hsa04510	At
Jak-STAT signaling pathway	hsa04630	At
Asthma	hsa05310	At
Hypertrophic cardiomyopathy (HCM)	hsa05410	At
Hematopoietic cell lineage	hsa04640	At
Toll-like receptor signaling pathway	hsa04620	At
PPAR signaling pathway	hsa03320	At
NOD-like receptor signaling pathway	hsa04621	At
Prion diseases	hsa05020	At
Allograft rejection	hsa05330	At
Chemokine signaling pathway	hsa04062	At

The first column: the 24 pathways that contain genes that are part of genetic interactions with one gene in a diabetes pathway and the other in an aneurysm or an atherosclerosis pathway. The second column: KEGG ID of the pathway. The third column: disease to which the pathway is related to (An denotes aneurysm, At denotes atherosclerosis, D denotes diabetes).

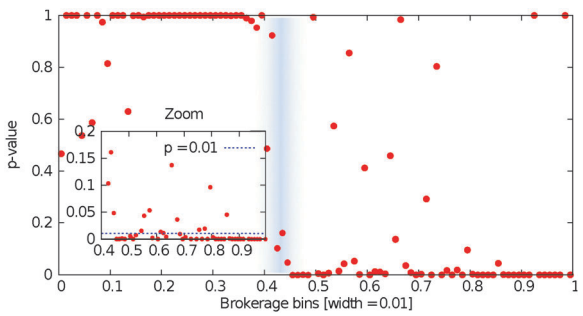

Fig. 2 Statistically significant brokerage values. *X*-axis: brokerage values in bins of 0.01. *Y*-axis: *p*-value that corresponds to the probability of obtaining the same or higher numbers of proteins (as counted in the disease PPI sub-network) in the bin by chance. Inset in the bottom left: red dots under the blue dotted line correspond to the statistically significant bins (p-values < 0.01). The shaded blue line highlights the natural barrier reflecting the difference between the statistical significance of low brokerage values and statistical significance of high brokerage values.

loss, which is necessary for the pathogenesis of aneurysm.[1] This validates that the presented methodology identifies genes enriched in biological processes that have already been proposed as causes of diabetes-aneurysm association.

Out of the 313 genes we identify 16 genes that, in addition to taking part in aneurysm or atherosclerosis pathways, also take part in genetic interactions with genes from diabetes pathways. We postulate that among the 313 broker genes in the disease PPI sub-network, these 16 genes are the most likely to be responsible for the observed relationships between the diseases. Namely, as discussed in the Introduction, genetic interactions can point to pairs of genes such that mutation of one interacting partner can be indicative of a functional change of the other interacting partner. In that sense, if there is a genetic interaction between a gene on a diabetes pathway and a broker gene on an aneurysm or an atherosclerosis pathway, then a mutation of a gene involved in a diabetes pathway can be related to a functional change of a broker gene on aneurysm or an atherosclerosis pathway. The 16 genes, their brokerage values, and KEGG IDs of the related pathways are presented in Table 2.

Recall that the number of pathways related to atherosclerosis is much higher than the number of pathways related to aneurysm (as listed in Tables 3 and 4). This is a consequence of a higher number of genes that are known to be related to atherosclerosis in comparison to the number of genes that are known to be related to aneurysm, as detailed in Section 4.1.2. Therefore, the ratio of the number of identified broker genes on aneurysm pathways and the number of identified broker genes on atherosclerosis pathways might be influenced by this difference in the size of available input data for the two diseases. With this in mind, note that the 16 genes that we further analyse are accurately identified. With additional data available in the future, possibly including biologically validated networks of pathways responsible for the two diseases, our methodology would be useful for identifying additional broker genes.

Using the DAVID database, we check functional enrichment of the 16 genes from Table 2. There are 8 kinases among the 16 genes: PIK3CG, MAP2K4, CDK2, GSK3A, RPS6KA5, BRAF, MAPK7, and MAP2K7. We use the hyper-geometric cumulative distribution to calculate the *p*-value that corresponds to the probability of finding 8 or more kinases among the 16 genes purely by chance. Since there are 151 kinases among 958 genes in the disease PPI sub-network, 8 out of 16 genes being kinases

Table 2 The 16 broker genes participating in specific genetic interactions

Gene name	Brok.	Degree	Pathways (KEGG ID)
MAPK7	1.0	7	hsa04010 (AN)
PPP3CA	1.0	4	hsa04010 (AN)
RPS6KA5	0.83	6	hsa04010 (AN)
MAPK8IP2	0.58	13	hsa04010 (AN)
GSK3A	0.83	4	hsa04062 (AT)
HSPA5	0.7	5	hsa05020 (AT)
PIK3CG	0.95	7	hsa05200 (AN,AT), hsa05210 (AN), hsa04630 (AT), hsa04062 (AT), hsa04620 (AT), hsa04510 (AT)
RAC1	0.84	29	hsa05200 (AN,AT), hsa04010 (AN), hsa05416 (AN), hsa05210 (AN), hsa04510 (AT), hsa04620 (AT), hsa04062 (AT)
CDK2	0.60	36	hsa05200 (AN,AT)
ACTG1	0.58	4	hsa05416 (AN), hsa04510 (AT), hsa05410 (AT), hsa05412 (AT), hsa05414 (AT)
HDAC1	0.48	49	hsa05200 (AN,AT)
CCND1	0.48	16	hsa05200 (AN,AT), hsa05416 (AN), hsa05210 (AN), hsa04630 (AT), hsa04510 (AT)
MAP2K7	0.48	19	hsa04010 (AN), hsa04620 (AT)
MAP2K4	0.46	22	hsa04010 (AN), hsa04620 (AT)
BRAF	0.46	16	hsa04270 (AN), hsa05200 (AN,AT), hsa04010 (AN), hsa05210 (AN), hsa04062 (AT), hsa04510 (AT)
CREBBP	0.46	49	hsa05200 (AN,AT), hsa04630 (AT)

The first column: the 16 genes that have statistically significant brokerage, that are on aneurysm or atherosclerosis pathways and that participate in genetic interactions such that one gene in the interaction is part of a diabetes pathway, while the other is part of an aneurysm or an atherosclerosis pathway. The second column: brokerage of the corresponding gene. The third column: the degree of the corresponding gene in the disease PPI subnetwork. The fourth column: KEGG IDs of pathways in which the gene takes part. We additionally denote pathways with: (AN) for aneurysm-related pathway, and (AT) for atherosclerosis-related pathway.

is statistically significant, p-value = 0.0013. To make sure that finding kinases is not just the consequence of a possibly high number of kinases among the 313 broker genes, we also calculate the statistical significance of finding 8 or more kinases among the 16 genes when taking only 313 broker genes as the background set. There are 66 kinases among the 313 genes, so finding 8 or more kinases among the 16 genes is again statistically significant, p-value = 0.008. Out of the 16 genes, 9 are involved in phosphorylation: PIK3CG, BRAF, MAP2K4, CDK2, RPS6KA5, CCND1, GSK3A, MAPK7, MAP2K7 (p-value = 1×10^{-4}). Clearly, all of the above listed 8 kinases are among them, as kinases can be turned on or off by phosphorylation (adding phosphate groups). Phosphorylation usually results in a functional change of the target protein, cellular location, or association with other proteins. That can lead to rewiring of pathways that these kinases participate in, which in the case of an aneurysm pathway could disrupt the onset of aneurysm.

The question remains why broker genes from our set that are kinases on an atherosclerosis pathway would not disrupt the onset of atherosclerosis. To answer this we check if any of the 16 genes have pleiotropic traits. Pleiotropy occurs when a gene influences multiple traits, for example, because the gene encodes a protein that is used for two or more functions, or has different functions in different tissues.[27] We find that PIK3CG phosphorylates phosphatidylinositol 4,5-bisphosphate to generate PIP3, which plays a pleiotropic role in regulating membrane signaling.† Pleiotropic activities of GSK3 have made it a therapeutical target for treatment of various human diseases, including type 2 diabetes.[28] It is also known that mutations that result from the pleiotropic effects of BRAF can lead to different transcriptional changes.[29] Also, MAPK7 has pleiotropic functions.[30] A mutation in a pleiotropic gene can have an effect on just one of its traits, or on all of them.[27] Two of these genes, BRAF and PIK3CG, are present both in aneurysm and atherosclerosis pathways (see Table 2), and since they genetically interact with diabetes-related genes this may explain why a mutation on such genes would influence the development of aneurysm and not atherosclerosis in diabetic patients.

The identified set of the 16 genes should further be explored in the search for exact mechanisms behind the protective role of diabetes in the development of aneurysm. The most likely candidate genes are MAPK7 and PPP3CA, as their brokerage values equal 1 (see Table 2), suggesting the high destructive potential on the pathways that they take part in. In fact, brokerage value 1 means that inactivity of MAPK7 or PPP3CA would completely destroy connectivity in their neighbourhoods. Note that MAPK7 and PPP3CA are involved in the MAPK signaling pathway, which is related to aneurysm, therefore their functional change can disable the signaling process that plays a role in formation of this disease. Although both genes have been already linked to aneurysm,[31,32] we here uncover that they may also play an important role in the diabetes–aneurysm relationship.

3 Conclusions

We address an important issue of why patients with diabetes do not develop aneurysm, but do develop atherosclerosis when the two diseases have similar risk factors. We integrate PPI and genetic interaction data.

† http://www.phosphosite.org/proteinAction.do?id=3655&showAllSites=true

Aiming to identify a set of genes responsible for the protective role of diabetes in the development of aneurysm, we focus on topological properties of genes in the PPI sub-network of pathways of the three diseases, identified by integrating information from genetic interactions. We apply the topological measure of Simmelian brokerage to find genes that have high potential for disrupting their neighbourhoods' connectivities, meaning that functional changes on such genes would result in disabling the pathways that they are part of. To the best of our knowledge, this topological measure has not previously been used for exploring disease relationships, or biological processes related to pathway functioning. Using this approach, we identify a set of 313 genes enriched in GO biological processes that facilitate mechanisms behind these particular disease relationships. Since genetic interactions involve pairs of genes such that a mutation on one gene is related to a functional change of the other,[19] out of the 313 genes we identify 16 genes on aneurysm and atherosclerosis pathways that take part in genetic interactions with genes from diabetes pathways. We suggest these genes hold the answer for the relationships between the three diseases and we encourage further research in this direction. We are encouraged by finding out that 8 out of the identified 16 genes are kinases (a statistically significant enrichment) that may act as switches in the related pathways. Also, two of the kinases that are both on aneurysm and atherosclerosis pathways are pleiotropic, explaining why these genes could disable onset, formation and progression of aneurysm, but enable atherosclerosis.

4 Materials and methods

4.1 Datasets

4.1.1 Biological networks. We obtain the human PPI network from BioGRID,[33] release 3.2.106, September 2013. We analyze the largest connected component of the network. To reduce noise we remove ubiquitin as the most connected protein in the network, since proteins with a large number of non-specific interaction partners might seriously bias the network topology leading to biased results. The resulting PPI network has 13 410 proteins (nodes) and 116 552 interactions (edges).

We downloaded the human genetic interaction (GI) network from BioGRID in September 2013 (release 3.2.106). The network contains 986 genes and 1295 genetic interactions. To increase coverage we also constructed a *predicted* human GI network using new GI data on direct positive and negative genetic interactions in *S. cerevisiae* from Boone Lab,‡ which they gave to us in September 2013.[34]. The yeast GI network contains 4365 genes and 266 750 interactions. Then, we use information on homologous genes between *H. sapiens* and *S. cereviseiae* from Homologene database,§ version *build67*, downloaded in January 2014. There are 1568 human genes that are yeast homologs. We create a *predicted* human GI network as follows: for each genetic interaction between yeast genes, we create a genetic interaction between their corresponding human homologs. This network of predicted human genetic interactions contains 1088 genes and 34 160 genetic interactions between them. We merge the human GI network from BioGRID with the predicted human GI network, resulting in the final network of human genetic interactions containing 1983 genes and 35 454 interactions. In this manuscript we refer to this network as the human genetic interaction (GI) network.

4.1.2 Disease genes. We obtain a list of genes involved in aneurysm using several sources to increase coverage: KEGG DISEASE database,[35] OMIM database[36] and Disease Ontology Lite.¶ We find in total 53 genes related to aneurysm, out of which 37 are present in the human PPI network.

We find genes involved in atherosclerosis in the OMIM database and Disease Ontology (DO) Lite. We find in total 205 atherosclerosis related genes, out of which 184 are present in the human PPI network.

We obtain genes involved in diabetes from KEGG DISEASE database, OMIM database and Disease Ontology Lite. To increase coverage, we also include genes from the following pathways in the KEGG PATHWAY database: type I diabetes mellitus, type II diabetes mellitus, and Maturity onset diabetes of the young. We find in total 503 diabetes genes, out of which 423 are present in the human PPI network. All data on disease genes are downloaded in November 2013.

4.1.3 Pathways. We downloaded all pathways relevant for diabetes mellitus from KEGG PATHWAY database in November 2013: type I diabetes mellitus (hsa04940), type II diabetes mellitus (hsa04930), and maturity onset diabetes of the young (hsa04950). These pathways have 47, 48, and 25 genes in the human PPI network, respectively. The KEGG Pathway database does not list a set of pathways directly related to aneurysm, so we identify pathways that may play a role in formation of this disease by checking the enrichment of all available KEGG pathways in genes known to be involved in this disease. Among all 282 pathways from KEGG, we find 8 pathways that are statistically significantly enriched in aneurysm genes (p-value threshold of 0.05). The obtained pathways and their KEGG IDs are listed in Table 3.

Henceforth, we refer to these pathways as ''aneurysm pathways''. Similarly, we identify 23 ''atherosclerosis pathways,'' listed in Table 4.

4.2 Disease PPI sub-network

We postulate that a mutation of a gene on a diabetes pathway is related to a functional change of a protein on an aneurysm pathway, such that it would disable the aneurysm pathway from causing the disease. A question remains why diabetes does not have a similar effect on atherosclerosis. As discussed in the Introduction, genetic interactions can point us to gene pairs such that a gene mutation on one gene can be indicative of a change in another gene's function. Hence, we identify pairs of genes involved in genetic interactions such that at least one

‡ http://www.utoronto.ca/boonelab/
§ http://www.ncbi.nlm.nih.gov/homologene
¶ http://django.nubic.northwestern.edu/fundo

Table 3 Pathways related to aneurysm

Pathway	KEGG ID	*p*-value
Pathways in cancer	hsa05200	2.1×10^{-3}
Cytokine–cytokine receptor interaction	hsa04060	4.5×10^{-3}
Vascular smooth muscle contraction	hsa04270	1.2×10^{-2}
Intestinal immune network for IgA production	hsa04672	1.9×10^{-2}
MAPK signaling pathway	hsa04010	2.6×10^{-2}
Viral myocarditis	hsa05416	3.7×10^{-2}
ECM–receptor interaction	hsa04512	5.0×10^{-2}
Colorectal cancer	hsa05210	5.0×10^{-2}

The first column: pathways that are statistically significantly enriched in genes related to aneurysm. The second column: KEGG ID of the pathway. The third column: *p*-value of statistical significance of the enrichment.

gene is from a diabetes pathway while the other is from an atherosclerosis or an aneurysm pathway. We find 31 genes that take part in such genetic interactions. We find that 24 pathways involved in one or more of the 3 diseases contain these 31 genes. We create a sub-network of the human PPI network on all proteins from these 24 pathways: the set of edges in the sub-network consists of all the edges in the PPI network that connect the proteins from the 24 pathways. Such a sub-network that contains all edges from the original big network is called an *induced* sub-network and we say that we *induce* a sub-network when we create a sub-network in this way. This sub-network contains 958 proteins and 3370 interactions. Henceforth, we refer to it as the ''disease PPI sub-network.''

4.3 Brokerage measure

Simmelian brokerage[21] is a measure that describes the significance of a node for the interconnectedness of its local neighbourhood in the network. For a node i, it is calculated as follows: $B_i = k_i - (k_i - 1)E_i$, where k_i is the degree of node i, and E_i is the ''local efficiency'' of node i in the network, calculated as:

$$E_i = \frac{1}{k_i(k_i-1)} \sum_{l \in N_i} \sum_{m \in N_i, m \neq l} \frac{1}{d_{lm}}, \quad (1)$$

where N_i denotes the neighbourhood of node i (the sub-network induced on the first neighbours of node i), and d_{lm} denotes the distance between nodes l and m. The local efficiency is normalised to $0 \leq E_i \leq 1$, so that the ''local brokerage'' of a node, B_i, takes values: $1 \leq B_i \leq k_i$. By definition, brokerage values for the nodes with degree 1 are equal to zero.

To be able to compare proteins based on their brokerage in the disease PPI sub-network, we normalize the described brokerage measure by scaling to the range [0,1], as follows: $B_{i,n} = \frac{B_i - 1}{k_i - 1}$, where $B_{i,n}$ is the normalized brokerage of node i.

Note that a high node degree does not implicate high brokerage (see first two rows of Table 2).

4.4 Finding statistically significant brokerage values

We calculate the brokerage values for all nodes of degree higher than 2 in the disease PPI sub-network. We assign nodes into bins in increments of 0.01 of brokerage values. We only take into account genes with degree higher than 2 for the following reasons. We are not interested in nodes with degree 1, as such local topology cannot confirm or refuse our hypothesis (we are looking for nodes whose removal will affect the interconnectedness of its first neighbours, and node with degree one has just one first neighbour). Also, there are 100 genes in the disease PPI sub-network with degree 2 whose neighbours are not directly connected. This means that their normalized

Table 4 Pathways related to atherosclerosis

Pathway	KEGG ID	*p*-value
Cytokine–cytokine receptor interaction	hsa04060	5.9×10^{-10}
Type I diabetes mellitus	hsa04940	5.9×10^{-7}
Toll-like receptor signaling pathway	hsa04620	9.2×10^{-7}
Hematopoietic cell lineage	hsa04640	4.4×10^{-5}
Allograft rejection	hsa05330	2.2×10^{-4}
Complement and coagulation cascades	hsa04610	2.7×10^{-4}
Graft-*versus*-host disease	hsa05332	3.4×10^{-4}
NOD-like receptor signaling pathway	hsa04621	7.7×10^{-4}
ECM–receptor interaction	hsa04512	1.0×10^{-3}
Focal adhesion	hsa04510	3.9×10^{-3}
Hypertrophic cardiomyopathy (HCM)	hsa05410	4.8×10^{-3}
Chemokine signaling pathway	hsa04062	6.3×10^{-3}
Intestinal immune network for IgA production	hsa04672	6.9×10^{-3}
PPAR signaling pathway	hsa03320	6.9×10^{-3}
Dilated cardiomyopathy	hsa05414	7.4×10^{-3}
Prion diseases	hsa05020	1.0×10^{-2}
Systemic lupus erythematosus	hsa05322	1.1×10^{-2}
Pathways in cancer	hsa05200	3.1×10^{-2}
Asthma	hsa05310	3.4×10^{-2}
Autoimmune thyroid disease	hsa05320	3.7×10^{-2}
Jak-STAT signaling pathway	hsa04630	3.8×10^{-2}
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	hsa05412	3.9×10^{-2}
Cell adhesion molecules (CAMs)	hsa04514	4.5×10^{-2}

The first column: pathways that are statistically significantly enriched in genes related to atherosclerosis. The second column: KEGG ID of the pathway. The third column: *p*-value of statistical significance of the enrichment.

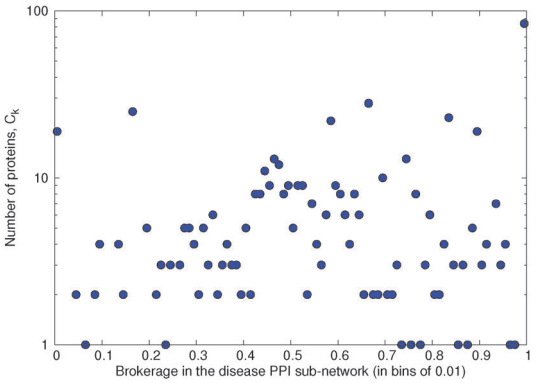

Fig. 3 Distribution of brokerage values in the disease PPI sub-network. *X*-axis: brokerage values in bins of 0.01. *Y*-axis: the number of proteins in the disease PPI sub-network that have a given brokerage.

brokerage equals 1. The number of such proteins is higher than the number of the remaining proteins in the disease PPI sub-network whose local wiring is non-trivial and yields brokerage scores of 1, so inclusion of degree 2 nodes would introduce noise to our analysis. The brokerage distribution is shown in Fig. 3.

In the remainder of this section, we explain how we model the disease PPI sub-network and identify statistically significant brokerage values.

4.4.1 Modeling the disease PPI sub-network. We generate 60 random networks with the same number of nodes and edges as in the disease PPI sub-network for each of the six commonly used random network models (totaling 60 × 6 = 360 random networks): Erdos–Renyi random graphs (ER),[37] Erdos–Renyi random graphs with the same degree distribution as the data (ER-DD),[38] Geometric Random Graphs (GEO),[39] Geometric Random Graphs with Gene Duplications and Mutations (GEO-GD),[40] Scale free Barabasi–Albert type networks (SF–BA),[41] and stickiness index based networks (STICKY).[42]

To find the best fitting network model, we compare the disease PPI sub-network with these random networks using the *graphlet degree distribution agreement* (*GDDA*) measure:[43] GDDA measures how similar the networks are in terms of distributions of small induced subgraphs – *graphlets*.[44] The arithmetic average of scaled and normalized distributions of all 73 graphlets results in a GDDA value in the range [0,1]. We use GDDA since it is a very sensitive measure for comparing the network structure.[43,45] The average GDDA values obtained for the GEO-GD, GEO, STICKY, SF-BA, ER-DD and ER network models are 0.85, 0.839, 0.825, 0.777, 0.755 and 0.673, with standard deviations of 0.01, 0.007, 0.007, 0.005, 0.006 and 0.008, respectively. Hence, GEO-GD and GEO models both provide a good fit to the disease PPI sub-network based on the best average GDDA value. Hence, we choose the GEO-GD random network model for modeling the disease PPI sub-network.

4.4.2 Statistically significant brokerage values. We find statistically significant brokerage values by using GEO-GD as a well-fitting network model to the disease PPI sub-network. We generate 1000 GEO-GD networks with the same number of nodes and edges as the disease PPI sub-network and calculate their brokerage distributions, again including only nodes with degree higher than 2. For each bin k and the corresponding node count, C_k, in the distribution shown in Fig. 3 for the disease PPI sub-network, we calculate the *p*-value that corresponds to the probability of obtaining C_k or more nodes in this bin by chance. We do this by comparing C_k for the disease PPI sub-network with the corresponding node counts in the 1000 GEO-GD networks. We identify the statistically significant brokerage bins by using the threshold of 0.01 (*p*-value). We further examine the proteins with the brokerage scores in the statistically significant bins.

Note that when performing this statistical analysis, we have used different bin sizes. Comparing the results, the bin size of 0.01 resulted in the most natural barrier between statistical significance of low brokerage values and high brokerage values (see Fig. 2). This bin size also resulted in the smallest number of bins whose statistical significance strongly deviates from the statistical significance of their neighbouring bins (scattered dots in Fig. 2). Therefore we report the results obtained using the bin size of 0.01.

Acknowledgements

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, the Serbian Ministry of Education and Science Project III44006, and ARRS project J1-5454.

References

1 S. Shantikumar, R. Ajjan, K. Porter and D. Scott, *Eur. J. Vasc. Endovasc.*, 2010, **39**, 200–207.
2 S. K. Prakash, C. Pedroza, Y. A. Khalil and D. M. Milewicz, *J. Am. Heart Assoc.*, 2012, **1**, jah3-e000323, DOI: 10.1161/JAHA.111.000323.
3 P. D. Rango, P. Cao, E. Cieri, G. Parlani, M. Lenti, G. Simonte and F. Verzini, *J. Vasc. Surg.*, 2012, **56**, 1555–1563.
4 S. R. Preis, S.-J. Hwang, S. Coady, M. J. Pencina, R. B. D'Agostino, P. J. Savage, D. Levy and C. S. Fox, *Circulation*, 2009, **119**, 1728–1735.
5 M. Creager, T. Lüscher, F. Cosentino and J. Beckman, *Circulation*, 2003, **108**, 1527–1532.
6 M. Patel, D. Hardman, C. Fisher and M. Appleberg, *J. Am. Coll. Surg.*, 1995, **181**, 371–382.
7 R. Sharan, I. Ulitsky and R. Shamir, *Mol. Syst. Biol.*, 2007, **3**, 88, DOI: 10.1038/msb4100129.
8 B. Schwikowski and P. Uetz, *Nat. Biotechnol.*, 2000, **18**, 1257–1261.
9 H. N. Chua, W.-K. Sung and L. Wong, *Bioinformatics*, 2006, **22**, 1623–1630.
10 M. P. Samanta and S. Liang, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 12579–12583.
11 T. Milenković and N. Pržulj, *Cancer Inf.*, 2008, **4**, 257–273.

12 T. Milenković, V. Memišević, A. Ganesan and N. Pržulj, *J. R. Soc., Interface*, 2010, **7**, 423–437.
13 A. Lusis and J. Weiss, *Circulation*, 2010, 157–170.
14 W. R. MacLellan, Y. Wang and A. J. Lusis, *Nat. Rev. Cardiol.*, 2012, 172–184.
15 A. Sarajlić and N. Pržulj, *BioMed Res. Int.*, 2014, **2014**, 527029, DOI: 10.1155/2014/527029.
16 R. Mani, R. P. S. Onge, J. L. Hartman, G. Giaever and F. P. Roth, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 3461–3466.
17 A. Ashworth, C. J. Lord and J. S. Reis-Filho, *Cell*, 2011, **145**, 30–38.
18 M. Žitnik, V. Janjić, C. Larminie, B. Zupan and N. Pržulj, *Sci. Rep.*, 2013, **3**, 3202.
19 M. C. Bassik, M. Kampmann, R. J. J. Lebbink, S. Wang, M. Y. Hein, I. Poser, J. Weibezahn, M. A. Horlbeck, S. Chen, M. Mann, A. A. Hyman, E. M. Leproust, M. T. McManus and J. S. Weissman, *Cell*, 2013, **152**, 909–922.
20 B. Lehner, *Trends Genet.*, 2011, **27**, 323–331.
21 V. Latora, V. Nicosia, P. Panzarasa, *CoRR*, 2012, abs/1211.0719.
22 E. K. Kim and E.-J. Choi, *Biochim. Biophys. Acta, Mol. Basis Dis.*, 2010, **1802**, 396–405.
23 D. W. Huang, B. T. Sherman and R. A. Lempicki, *Nat. Protoc.*, 2008, **4**, 44–57.
24 D. W. Huang, B. T. Sherman and R. A. Lempicki, *Nucleic Acids Res.*, 2009, **37**, 1–13.
25 A. Berrier and K. Yamada, *J. Cell. Physiol.*, 2007, **213**, 565–573.
26 P. Norman, T. Davis, M. Le and J. Golledge, *Connect. Tissue Res.*, 2007, **48**, 125–131.
27 F. W. Stearns, *Genetics*, 2010, **186**, 767–773.
28 J. Van Wauwe and B. Haefner, *Drug News Perspect.*, 2003, **16**, 557–565.
29 S. Pavey, P. Johansson, L. Packer, J. Taylor, M. Stark, P. M. Pollock, G. J. Walker, G. M. Boyle, U. Harper, S.-J. Cozzi, K. Hansen, L. Yudt, C. Schmidt, P. Hersey, K. A. O. Ellem, M. G. E. O Rourke, P. G. Parsons, P. Meltzer, M. Ringner and N. K. Hayward, *Oncogene*, 2004, **23**, 4060–4067.
30 D. Bond and E. Foley, *PLoS Pathog.*, 2009, **5**.
31 Y. Wang, C. Barbacioru, D. Shiffman, S. Balasubramanian, O. Iakoubova, M. Tranquilli, G. Albornoz, J. Blake, N. Mehmet, D. Ngadimo, K. Poulter, F. Chan, R. Samaha and J. A. Elefteriades, *PLoS One*, 2007, **2**, e1050, DOI: 10.1371/journal.pone.0001050.
32 B. Bakir-Gungor and O. U. Sezerman, *PLoS One*, 2013, **8**, e57022.
33 A. Chatr-aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. G. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. M. Rust, M. S. Livstone, R. Oughtred, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2013, **41**, 816–823.
34 We thank Charlie Boone for giving us his unpublished complete set of genetic interactions in baker's yeast.
35 M. Kanehisa, S. Goto, Y. Sato, M. Furumichi and M. Tanabe, *Nucleic Acids Res.*, 2012, **40**, D109–D114.
36 A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic Acids Res.*, 2005, **33**, D514–D517.
37 P. Erdös and A. Rényi, *Publ. Math., Debrecen*, 1959, **6**, 290.
38 M. Newman, *Networks: an introduction*, Oxford University Press, Inc., 2010.
39 M. Penrose, *Random Geometric Graphs (Oxford Studies in Probability)*, Oxford University Press, USA, 2003.
40 N. Przulj, O. Kuchaiev, A. Stevanovic and W. Hayes, *Pacific Symposium on Biocomputing*, 2010, pp. 178–189.
41 A.-L. Barabási and R. Albert, *Science*, 1999, **286**, 509–512.
42 N. Przulj and D. J. Higham, *J. R. Soc., Interface*, 2006, **3**, 711–716.
43 N. Pržulj, *Bioinformatics*, 2007, **23**, e177–e183.
44 N. Pržulj, D. G. Corneil and I. Jurisica, *Bioinformatics*, 2004, **20**, 3508–3515.
45 W. Hayes, K. Sun and N. Przulj, *Bioinformatics*, 2013, **29**, 483–491.