Applied Network Science

RESEARCH

Open Access

Geometric characterisation of disease modules

Franziska Härtner[1], Miguel A. Andrade-Navarro[2] and Gregorio Alanis-Lobato[2*]

*Correspondence:
galanisl@uni-mainz.de
[2]Faculty of Biology, Johannes Gutenberg Universität, Institute of Molecular Biology, Ackermannweg 4, 55128, Mainz, Germany
Full list of author information is available at the end of the article

Abstract
There is an increasing accumulation of evidence supporting the existence of a hyperbolic geometry underlying the network representation of complex systems. In particular, it has been shown that the latent geometry of the human protein network (hPIN) captures biologically relevant information, leading to a meaningful visual representation of protein-protein interactions and translating challenging systems biology problems into measuring distances between proteins. Moreover, proteins can efficiently communicate with each other, without global knowledge of the hPIN structure, via a greedy routing (GR) process in which hyperbolic distances guide biological signals from source to target proteins.

It is thanks to this effective information routing throughout the hPIN that the cell operates, communicates with other cells and reacts to environmental changes. As a result, the malfunction of one or a few members of this intricate system can disturb its dynamics and derive in disease phenotypes. In fact, it is known that the proteins associated with a single disease agglomerate non-randomly in the same region of the hPIN, forming one or several connected components known as the disease module (DM). Here, we present a geometric characterisation of DMs. First, we found that DM positions on the two-dimensional hyperbolic plane reflect their fragmentation and functional heterogeneity, rendering an informative picture of the cellular processes that the disease is affecting. Second, we used a distance-based dissimilarity measure to cluster DMs with shared clinical features. Finally, we took advantage of the GR strategy to study how defective proteins affect the transduction of signals throughout the hPIN.

Keywords: Protein interactions, Hyperbolic geometry, Disease modules, Greedy routing, Systems biology, Network medicine

Introduction

Regardless of whether they represent the Internet or associations between proteins, people or airports; complex networks share many topological features (Albert and Barabási 2002), which suggests that similar rules govern their formation. Various models, aimed at mimicking the evolution and growth of these networks, assume the existence of a geometry underlying their structure and shaping their topology (Aste et al. 2005; Aste et al. 2012; Barthélemy 2011; Dall and Christensen 2002; Ferretti and Cortelezzi 2011; Krioukov et al. 2010; Papadopoulos et al. 2012). Recent work has shown that if such a geometry is hyperbolic, the emergence of the observed architecture of complex networks can be naturally explained by a distance-dependent connection probability between nodes in this metric space (Bianconi and Rahmede 2017; Ferretti et al. 2014; Krioukov et al. 2010; Papadopoulos et al. 2012; Wu et al. 2015).

In the native representation of complex networks in the two-dimensional hyperbolic plane $\mathbb{H}^2$, the N network nodes are enclosed inside a circle of radius $R \sim \ln N$, each one lying at polar coordinates (r_i, θ_i) (Krioukov et al. 2010). These positions must ensure that connected nodes are close to each other and disconnected nodes are far apart. According to the definition of hyperbolic distance $d_{\mathbb{H}^2}(i,j) \approx r_i + r_j + 2\ln(\theta_{ij}/2)$ (Alanis-Lobato and Andrade-Navarro 2016; Krioukov et al. 2010; Papadopoulos et al. 2012), high-degree nodes are close to the centre of $\mathbb{H}^2$ because they need to be nearby many other nodes.

The embedding of real networks to hyperbolic space has shed light on their function and community organisation (Alanis-Lobato et al. 2018; Allard and Serrano 2018; Boguñá et al. 2010; García-Pérez et al. 2016; Serrano et al. 2012). For example, information routing overheads could be alleviated if the latent geometry of the Internet is used to guide packets between computers (Boguñá et al. 2010). Also, the geometry of the *E. coli* and human metabolic networks has put forward a new view of the definition of and interdependence between biochemical pathways (Serrano et al. 2012).

Of special interest for the present work is the analysis of the latent geometry of the human protein interaction network (hPIN). Alanis-Lobato and colleagues found that the hyperbolic map of the hPIN constitutes a meaningful and useful two-dimensional depiction of proteins and their interactions (Alanis-Lobato et al. 2018). The inferred radial coordinates of proteins hint at their evolutionary origin, whereas angular sectors group proteins with related biological functions and cellular localisations. In addition, hyperbolic distances can be used as likelihood scores for the prediction of biologically plausible protein-protein interactions (PPIs). Finally, Alanis-Lobato et al. showed that proteins can efficiently communicate with each other, without knowledge of the whole hPIN structure, by means of a greedy routing process in which hyperbolic distances guide biological signals from membrane receptors to transcription factors in the nucleus (Alanis-Lobato et al. 2018).

It is thanks to the effective transduction of signals throughout the hPIN that the cell operates, communicates with other cells and reacts to environmental stresses (Vinayagam et al. 2011). Therefore, defective or dysregulated proteins can disrupt PPIs, clog important signalling pathways and cause disease phenotypes (Taylor and Wrana 2012). In fact, it has been reported that the proteins associated with a single disease agglomerate non-randomly in the same region of the hPIN, forming one or several connected components known as the disease module (DM) (Agrawal et al. 2018; Menche et al. 2015). Consequently, disease-related proteins are more likely to have PPIs with each other than with random proteins. This particular connectivity pattern has been exploited to prioritise other proteins that may be related to a disease of interest (Cowen et al. 2017; Ghiassian et al. 2015; Köhler et al. 2008; Lage et al. 2007; Wu et al. 2008).

The above prompted us to analyse disease-associated proteins from a geometric perspective and to investigate how the latent geometry of the hPIN can reflect and expand our current knowledge about the organisation of DMs. We also took advantage of the greedy routing protocol to study the impact of disease proteins on the function of the hPIN.

Results

Topology and geometry of DMs

After the construction of a high-quality hPIN, its embedding to $\mathbb{H}^2$ (see Methods and Additional file 1: S1 and S2) and the evaluation of the embedding (see Methods, Additional file 2: Figures S1 and S2), we proceeded to analyse the topological and geometrical properties of DMs formed by the products of genes associated with 157 different diseases (see Methods and Additional file 1: S3).

In agreement with previous studies (Agrawal et al. 2018; Menche et al. 2015), we found that proteins associated with a single disease (see Fig. 1a) form several connected components in the hPIN, instead of a unique connected entity (see Fig. 1b). Nevertheless, the fraction of proteins contained in the largest of such components (LCC, see Fig. 1a) is significantly greater than expected by chance (see Methods and Fig. 1c). Besides, the average

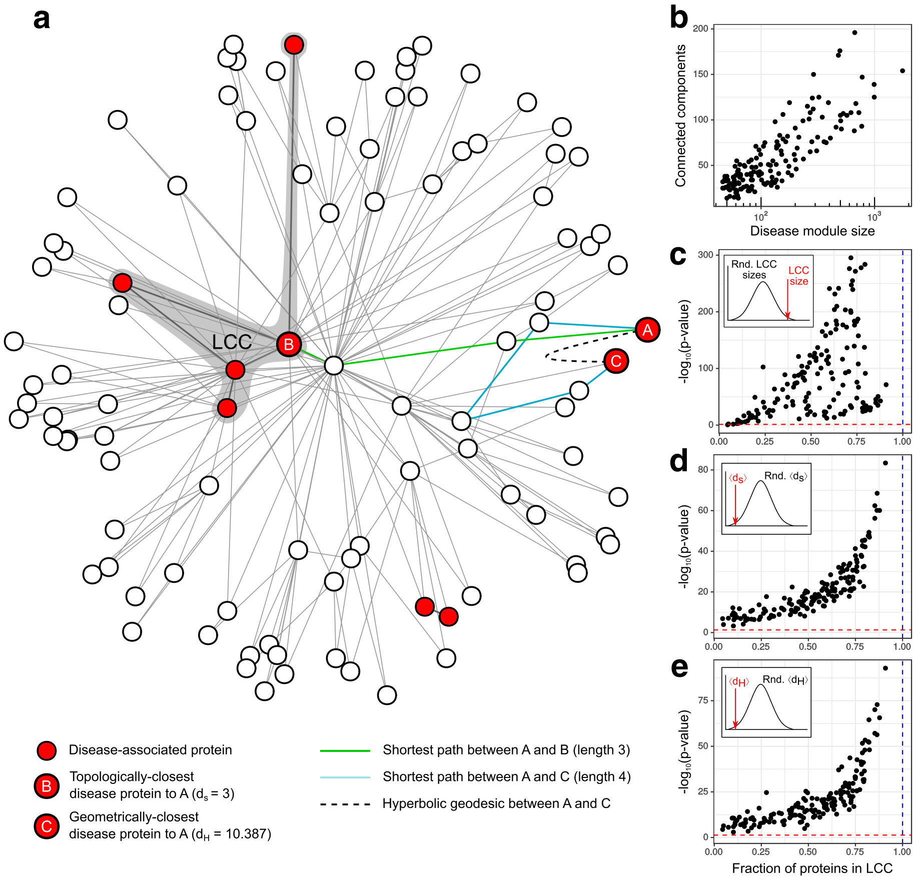

Fig. 1 Topology and geometry of disease modules. **a** A toy protein network embedded in $\mathbb{H}^2$ and the gene products associated with a disease highlighted in red. The resulting DM is fragmented in 4 components and the largest one (LCC) has size 5. The topologically-closest protein to A is B ($d_s = 3$) and its geometrically-closest protein is C ($d_H = 10.387$), although it is 4 hops away from A. **b** The 157 studied DMs are split into several connected components in the hPIN. **c** The LCC size of each DM was compared with a random distribution of LCC sizes via a z-test. **d** The average of the shortest paths from each disease protein to its topologically-closest other DM member ($\langle d_s \rangle$) was compared with random expectation via a z-test. **e** The average of the hyperbolic distances from each disease protein to its geometrically-closest other DM member ($\langle d_H \rangle$) was compared with random expectation via a z-test. The red lines correspond to the significance level $\alpha = 0.05$ and the blue ones to the situation in which all proteins associated with a disease form a single connected component

of the shortest paths from each disease protein to its closest other DM member ($\langle d_s \rangle$, see Fig. 1a) is smaller than expected by chance (see Methods and Fig. 1d), which indicates that, although DMs are fragmented, their components are made up of topologically nearby proteins (Menche et al. 2015).

To investigate how the above DM connectivity patterns are reflected in $\mathbb{H}^2$, we computed, for each of the n disease proteins in a DM, the hyperbolic distance to the closest other protein in the same module (d_H, see Fig. 1a). The n resulting distances were averaged ($\langle d_H \rangle$) and compared with a distribution of average hyperbolic distances resulting from randomly sampling as many proteins from the hPIN as there are in each considered DM (see Methods). Figure 1e shows that, in all 157 cases, $\langle d_H \rangle$ is significantly shorter than expected by chance. This result points out that in $\mathbb{H}^2$ DMs are fragmented in components formed by at least two hyperbolically nearby proteins. While this appears to be a predictable result, it is important to note that the DM member that is topologically-closest to a protein in the same module can be different from the geometrically-closest one (see Fig. 1a).

DMs are formed by functionally distinct submodules

The observed level of DM fragmentation has been attributed to the incompleteness of the hPIN and the limited knowledge of disease genes (Menche et al. 2015). While these two factors may certainly contribute to DM splitting, disease-related genes belong to broad functional categories (Jimenez-Sanchez et al. 2001) and locate to distant cellular compartments, which makes their PPI unlikely (Thul et al. 2017). As such, DMs are most probably formed by functionally distinct submodules that, together, contribute to the disease phenotype.

In $\mathbb{H}^2$, neighbouring proteins play roles in very similar biological processes and lie in the same organelles (see Fig. 2a, b and Methods). Furthermore, Alanis-Lobato et al. reported that the angular component of hyperbolic space captures the functional and spatial organisation of the cell (Alanis-Lobato et al. 2018) (see Additional file 2: Figure S2b). Therefore, we partitioned the angular dimension into several sectors, according to the over-represented protein class within them (see Methods), and studied whether the distribution of inferred angular coordinates of disease proteins does hint at their functional modularity.

Figure 2c and Additional file 1: S3 promptly highlight the protein class heterogeneity of DMs, together with interesting correlations between protein function and features of the disease. For example, almost all diseases exhibit a strong receptor (Rec) component. Since most drug-targets are membrane Recs, this could be the result of study biases against membrane proteins (Brito and Andrews 2011). However, this also points at the crucial role of signalling in cell operation (Vinayagam et al. 2011): celiac disease and rheumatoid arthritis, the DMs with the highest peaks at the Rec-enriched sector, are both characterised by altered inflammatory signalling pathways (Coenen et al. 2009). On the other hand, a high peak in the sector enriched for RNA-binding proteins (RBPs) may be indicative of problems in post-transcriptional gene regulation (Lukong et al. 2008). In fact, myelodysplastic syndromes have been linked with the malfunction of different components of the spliceosome and frequently develop into myeloid leukaemia, the DM with the highest RBP-related peak (Turner and Monzón-Casanova 2017).

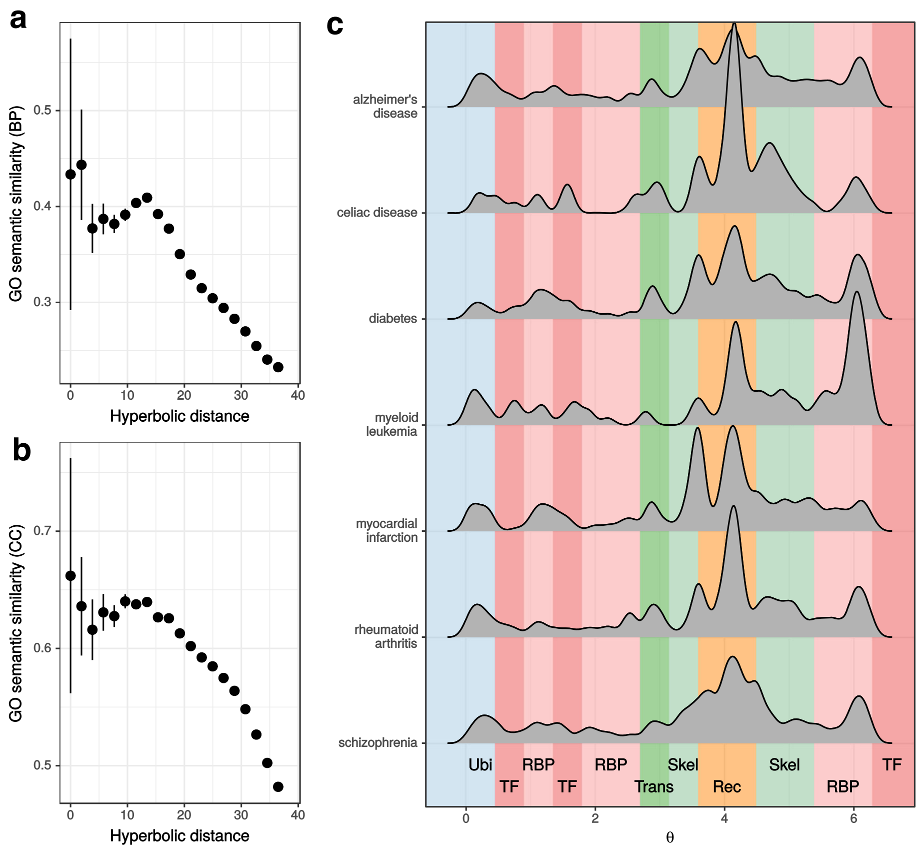

Fig. 2 Disease modules split into functionally distinct submodules. **a**. Gene Ontology (GO) semantic similarity of Biological Process (BP) terms that annotate proteins pairs spaced not further than the indicated hyperbolic distance. **b**. Same as **a** but for Cellular Component (CC) terms. **c** Angular distribution of disease proteins associated with 7 illnesses from our gene-disease association dataset. The coloured backgrounds indicate that one of 6 considered protein classes is over-represented in that angular range (Ubi: proteins involved in ubiquitination/proteolysis, TF: transcription factors, RBP: RNA-binding proteins, Trans: transporters, Skel: constituents of the cytoskeleton, Rec: receptors)

Distance-based clustering of DMs

A pathophysiology-based classification system of human disorders is pivotal in diagnosis and research (Baird 2013). The International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) is an effort of the World Health Organisation to categorise morbid entities according to established criteria (WHO 2016). Despite the fact that, in practice, the ICD has become the standard classification, it has to be periodically revised (WHO 2016). Revisions remain a challenging task: categories must have a strong scientific basis, many diseases seem to be a complicated mix of disorders and many lack archetypal biomarkers, which means that their diagnosis still depends on subjective criteria (Baird 2013). Could DM features aid in this task?

Menche et al. introduced a measure of topological overlap between DMs to identify their shared clinical characteristics (Menche et al. 2015). This measure of separation compares the average shortest path between each DM member and the nearest protein within the module, $\langle d_s(A,A)\rangle$ and $\langle d_s(B,B)\rangle$, to the average shortest path between each DM member and the nearest protein in the other module, $\langle d_s(A,B)\rangle$: $s_s(A,B) = \langle d_s(A,B)\rangle - (\langle d_s(A,A)\rangle - \langle d_s(B,B)\rangle)/2$. We translated this definition to the geometric position of DMs in $\mathbb{H}^2$ by considering the average hyperbolic distance between each DM member and the closest protein in the same, $\langle d_H(A,A)\rangle$, or another module, $\langle d_H(A,B)\rangle$ (see Methods).

We used the distance-based separation between DMs, $s_H(A,B)$, to construct a matrix of pairwise module separations. Then, we resorted to non-centred Minimum Curvilinear Embedding (ncMCE) for the non-linear dimensionality reduction of our 157 DMs through this matrix (Cannistraci et al. 2013) (see Methods). Figure 3 shows that the use of $s_H(A,B)$ leads to the distinction of ICD-10 categories that group pathobiologically and clinically similar illnesses. This is in spite of the observed fragmentation and functional heterogeneity of DMs (see Figs. 1 and 2). We quantified the quality of the resulting clustering with the concordance score (C-score, see Methods) and found that it is close to what $s_s(A,B)$ achieves and better than using just intersections between the members of DMs (see Fig. 3, Additional file 2: Figure S4 and the Methods). Moreover, using the $s_H(A,B)$ matrix as the basis for a hierarchical clustering of DMs recapitulates the disease types exposed by ncMCE (see Additional file 2: Figure S5 and the Methods).

Figure 3 also highlights some DMs whose ICD-10 classification does not correspond to their ncMCE cluster. Yet, we found reasonable explanations for all these DMs. For example, non-alcoholic fatty liver disease and steatohepatitis (a more extreme variety of the former) are both strongly associated with insulin resistance and metabolic syndromes

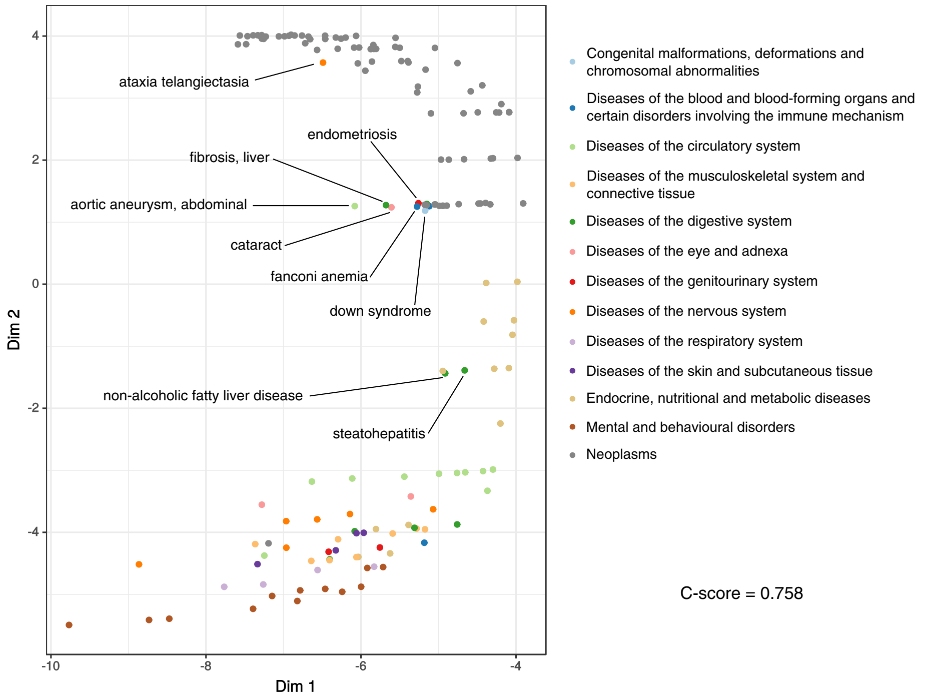

Fig. 3 Distance-based clustering of disease modules. ncMCE, a nonlinear dimensionality reduction algorithm, was applied to the matrix of pairwise hyperbolic distance-based DM separations. The resulting two-dimensional projection of the data separates DMs according to the disease type reported in the ICD-10. Some interesting DMs that were not correctly clustered are highlighted. The concordance score (C-score) over Dim 2 for this projection is 0.758. The points corresponding to *endometriosis* and *down syndrome* were slightly moved to make them visible

(Tolman and Dalpiaz 2007), which explains why they are clustered with endocrine and metabolic disorders (ICD-10 classifies them as diseases of the digestive system).

The other cases involve conditions clustered with cancers but categorised differently by the ICD-10. Ataxia telangiectasia (AT) is considered a neurodegenerative disease but it is linked to a high risk for cancer (Gilmore 2014). Mutations in the gene coding for the kinase ATM were identified as the main cause of AT (Gilmore 2014). ATM is mainly involved in the recruitment of DNA repair complexes in response to single or double strand breaks, hence the relationship between AT and neoplasms (Cremona and Behrens 2013).

Endometriosis is a disease characterised by the presence of endometrial-like tissue outside the uterine cavity, causing a chronic inflammatory response (Heidemann et al. 2013). Women with atypical endometriosis have a significantly increased risk of certain forms of ovarian cancer. Likewise, women with ovarian cancer are more likely to have the disease (Vercellini et al. 2013). In addition, there is also increasing awareness that women with endometriosis have a higher risk for developing breast cancer (Roy et al. 2015; Pontikaki et al. 2015).

Liver fibrosis is the result of constant inflammation of the liver, leading to excessive accumulation of extracellular matrix proteins, especially collagen. Severe liver fibrosis can lead to hepatocellular carcinoma (Uehara et al. 2013; Sakurai and Kudo 2013).

Down syndrome (DS) is known to have several associated haemopoietic conditions, leukaemia amongst them. Children with DS have a 10-20 times higher risk of having leukaemia than non-DS children (Bruwier and Chantrain 2011; Xavier et al. 2009).

Fanconi anaemia (FA) is a rare bone marrow failure disease leading to impaired DNA repair response. This disease also leads to haematologic changes. So far, mutations in 19 DNA damage response genes are known to either cause FA or increase the risk to have it (Bogliolo and Surrallés 2015). Depending on the affected genes, FA increases the risks for developing leukaemia, solid tumours, breast and ovarian cancer (Alter 2014; Bogliolo and Surrallés 2015).

Finally, we did not find conclusive associations between abdominal aortic aneurysm and cataracts with neoplasms. However, smoking is the strongest risk factor of the former (Kent 2014), which could explain its concomitance with lung cancer (Blochle et al. 2008). For the latter, the association between early-onset cataracts and insufficient anti-oxidative activity inspired a study in which early-onset cataract and healthy patients were followed for several years to estimate the incidences of cancer. The result was a two-fold higher cancer risk for the early-onset cataract cohort (Chiang et al. 2014).

Impact of disease proteins on cellular function

Signal transduction is the process of translating external signals at the cell membrane to specific responses within the cell. In this manner, cells react to environmental stresses by switching on certain genes and turning off others (Vinayagam et al. 2011). Interestingly, membrane receptors are only aware of their direct interaction partners, not the entire hPIN structure. Nonetheless, they manage to relay an external signal to the appropriate neighbour, so that it reaches transcription factors (TFs) in the nucleus that up-regulate target genes (Vinayagam et al. 2011).

The efficiency of signal transduction throughout the hPIN can be studied by means of its latent geometry and a navigation strategy known as greedy routing (GR) (Kleinberg

2000). In GR, inferred node coordinates are used as addresses that guide a signal from a source to a target node. The source checks amongst its direct neighbours for the hyperbolically closest to the target and sends the signal there, the recipient node does the same and the process is repeated until the signal reaches the target. If the signal is sent to a previously visited node (i.e. it falls into a loop), the routing is considered unsuccessful (Boguñá et al. 2009; Ortiz et al. 2017; Papadopoulos et al. 2010). GR efficiency (i.e. network navigability) is quantified as the fraction of successful routing events from a sufficiently large number of randomly chosen source-target pairs (Krioukov et al. 2010).

We measured the reference efficiency of the hPIN by considering 500 source-target pairs and repeating the measurement 100 times. Figure 4a shows that the median GR efficiency of the hPIN is 0.8 and the median hop stretch (length of the greedy path between two nodes divided by length of their shortest path) is close to 1. This figure also highlights the biological importance of signal transduction: GR efficiency

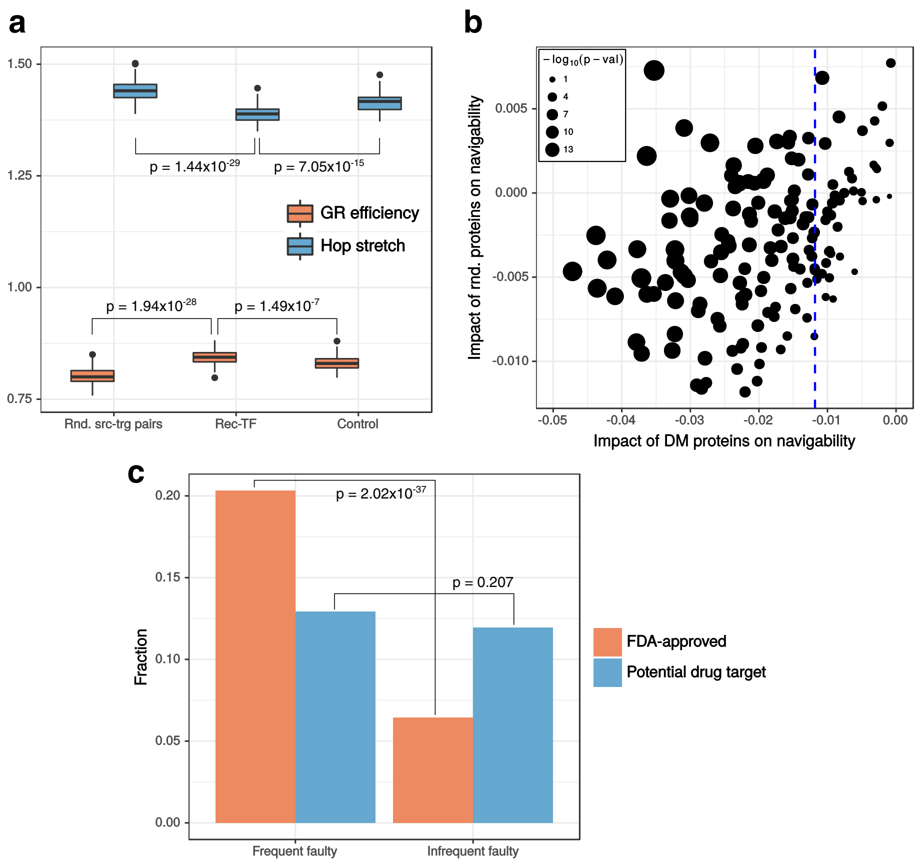

Fig. 4 Disease proteins affect signal transduction. **a** Reference greedy routing (GR) efficiency and hop stretch of the hPIN (Rnd. src-trg pairs) compared to the efficiency (hop stretch) achieved if signals are routed between receptors (Rec) and transcription factors (TF) or between proteins that are neither Recs nor TFs, but that have degrees similar to their counterparts (Control). **b** Impact on GR efficiency (navigability) dealt by faulty proteins sampled from the set of proteins associated with a DM or from a pool of proteins with similar degrees. The blue dashed line indicates the value of the highest impact (i.e. the most negative) from the latter case. **c** Frequently used faulty proteins are more likely to be FDA-approved drug targets or potential drug targets than infrequent faulty proteins. *P*-values are the result of a Fisher's test comparing the proportion of FDA-approved or potential drug targets in frequent vs infrequent faulty proteins

(hop stretch) for Rec-TF pairs is significantly bigger (smaller) than the reference and what is achieved by considering proteins that are neither Recs nor TFs, but have degrees similar to their counterparts (see Additional file 2: Figure S6 and the Methods).

The communication channels between Recs and TFs are so important that almost all known diseases are linked to their dysfunction (Thul et al. 2017). We studied the impact of disease proteins on GR efficiency through the introduction of faulty proteins to the GR process (Alanis-Lobato et al. 2018). Faulty proteins drop any signals they receive, making routing unsuccessful, and model the effects caused by mutations or insufficient protein levels. For each DM, we measured the average of 50 GR efficiencies resulting from routing signals between 500 source-target pairs. Based on a study estimating 3-24 homozygous disease-causing mutations per individual (Xue et al. 2012), we introduced 20 faulty proteins to each one of the 50 experiments. Faulty proteins were sampled from the members of a DM or chosen at random from a pool of proteins with similar degrees. In each case, we quantified the impact on navigability as the difference between the resulting average efficiency and the reference efficiency of the hPIN (0.82). The more negative the difference, the higher the impact (see Methods).

Figure 4b shows that disease proteins deal a higher impact on network navigability than faulty proteins sampled randomly from the hPIN (see Methods). Besides, for 89.17% of the DMs, this impact is significantly greater than expected (see Methods). Interestingly, the functional enrichment analysis (see Methods) of proteins associated with these DMs (mostly neoplasms and diseases of the nervous system) revealed that they are mainly receptors whose failure affects gene transcription and apoptosis (see Additional file 2: Figure S7a). By contrast, proteins associated with the remaining 10.83% DMs (mostly diseases of the connective tissue and metabolic disorders) are receptors whose failure affects inflammatory responses and cell homeostasis (see Additional file 2: Figure S7b). These results further support our assumption that GR simulates signal transduction events from Recs to TFs.

Given the high percentage of DMs whose members exert a negative impact on GR, recurring gene products in the lists of faulty proteins from the above experiments could hint at the biological role played by proteins responsible for such an impact. We focused on faulty proteins reappearing more than the upper quartile of the frequency distribution (frequent faulty) and found that they are mostly the products of genes with enzymatic and receptor activity (see Additional file 2: Figure S8a), whereas infrequent faulty proteins are involved in more heterogeneous functions (see Additional file 2: Figure S8b). Since current FDA-approved drugs are mainly directed to enzymes and receptors (Brito and Andrews 2011; Thul et al. 2017), we reasoned that this methodology could be used to identify potential drug targets. Indeed, a significant fraction of frequent faulty proteins are reported FDA-approved drug targets (see Fig. 4c, Methods and Additional file 1: S4). Also, although not significant, the fraction of frequent faulty proteins that are considered potential drug targets by the Human Protein Atlas (Thul et al. 2017) is higher than the fraction of infrequent faulty proteins with the same property (see Fig. 4c and Additional file 1: S4).

Potential drug targets are proteins whose structure, biochemical aspects and associated pathways make them candidate druggable proteins (Thul et al. 2017). Intriguingly, some of the most frequent faulty proteins that we identified are already being investigated for their therapeutic potential, which endorses the value of the proposed GR-based approach for drug target prioritisation (see Additional file 1: S4). For example, Fienberg and colleagues are trying to selectively inhibit the domains of the ACE protein to treat fibrosis and hypertension (Fienberg et al. 2018); Ding et al. are developing dual inhibitors of EGFR and PI3Kα as an approach against tumours (Ding et al. 2018); similarly, early-phase trials suggest that therapies targeting the PI3K/AKT/mTOR pathway can be used in patients with advanced cancers (Janku et al. 2013). Finally, it has been shown that directly targeting STAT3 with piperlongumine has positive and potent effects against breast cancer (Bharadwaj et al. 2014).

Conclusions

The representation of the human protein interaction network in the two-dimensional hyperbolic plane has been shown to be both meaningful and useful: inferred node coordinates convey information about protein evolution and function, whereas hyperbolic distances can be used to identify potential protein interactions and simulate signalling events (Alanis-Lobato et al. 2018).

In this paper, we report yet another scenario in which the latent geometry of the hPIN proves useful, namely the network-based analysis of disease-associated proteins. First, we found that the geometric position of disease modules reflects their fragmentation and functional heterogeneity, underscoring the complexity of human disorders. Visualisation of the angular distribution of DM members, together with the protein classes over-represented in different sectors of $\mathbb{H}^2$, renders an informative picture of the cellular processes that the disease is affecting. Second, we used a hyperbolic distance-based dissimilarity measure to cluster DMs. The resulting clusters are in good agreement with the standard ICD-10 and bring out unexpected but reasonable relationships between disorders. Finally, we studied how defective proteins affect the efficient routing of signals throughout the hPIN. Interestingly, proteins that were frequently considered faulty in our experiments represent known or potential drug targets.

While our geometric characterisation of DMs was carried out on a proteome-scale high-quality network, human PPI maps are still incomplete (Luck et al. 2017). In consequence, the projection to $\mathbb{H}^2$ can change if more interactions or proteins are considered. To test the robustness of our findings to network topology changes, we repeated our experiments on an independent PPI dataset and observed the same trends (see Additional file 2: Figures S11–S14, Additional file 1: S5-S6 and the Methods). The consistency between the two independent analyses reasserts the validity of our results.

Improvements in the sensitivity and scalability of PPI detection methods will ultimately lead to a more complete picture of the network of protein interactions that take place in the human cell (Luck et al. 2017). This reference PPI map will allow for a more accurate depiction of the network in hyperbolic space, which in turn will represent a more powerful tool for the analysis of protein and cellular functions. Based on the present study, we anticipate that the integration of the hPIN geometry with gene-disease association data will play a key role in advancing our understanding of human disease and defining the druggable proteome.

Methods

Gene-disease associations

We obtained gene-disease associations from DisGeNET v5.0 (Piñero et al. 2017), filtered out gene-disease pairs supported by less than 3 publications and diseases with less than 50 associated genes. To avoid redundancies, we merged diseases with very similar lists of associated genes. For this, we constructed a Jaccard similarity matrix between diseases to cluster them hierarchically. Upon inspection of the resulting dendrogram, we merged diseases that ended up in the same cluster after cutting the tree at height 6. If X and Y represent the sets of genes associated with two different diseases, their Jaccard similarity is obtained with $J(X, Y) = |X \cap Y|/|X \cup Y|$ (Jaccard 1912).

The described filtering and merging process resulted in 157 diseases. We assigned them to their corresponding disease type using Revision 10 of the ICD (WHO 2016).

Construction of the hPIN

We built the hPIN with high-quality interactions from the Human Integrated Protein-Protein Interaction rEference (HIPPIE) v2.0 (Alanis-Lobato et al. 2017). Each PPI in HIPPIE has an assigned confidence score. Based on reference estimates of the hPIN size (Venkatesan et al. 2009), we considered the 155,000 interactions with the highest scores. After discarding self-interactions, merging redundant PPIs by considering their maximum score and extracting the network's LCC, we finally arrived to an hPIN formed by 150,212 interactions between 14,788 proteins. Protein names were translated to their corresponding Entrez ID and gene symbol with the MyGene.info web service (Xin et al. 2016).

In addition, we verified the validity of our results in an independent network dataset: the LCC formed by PPIs from the HINT database (19/02/18 snapshot of binary interactions) (Das and Yu 2012). The HINT network is comprised of 47,181 interactions between 10,370 proteins.

Identification of proteins classes and drug targets

We integrated information from several resources to identify proteins with TF, receptor, transporter or RNA-binding activity; as well as constituents of the cytoskeleton and proteins involved in ubiquitination/proteolysis. Our pool of TFs comes from the Animal Transcription Factor Database 2.0 (Zhang et al. 2015), the census of human TFs (Vaquerizas et al. 2009) and the Human Protein Atlas (Uhlen et al. 2015). From the latter, we also collected constituents of the cytoskeleton, proteolysis- and cancer-related proteins, receptors, transporters, RBPs, FDA-approved and potential drug targets. Additional receptors and transporters were taken from the Guide to Pharmacology (Southan et al. 2015). We also took into account RBPs from the RBP census (Gerstberger et al. 2014).

Mapping the hPIN to hyperbolic space

We mapped the hPIN to the two-dimensional hyperbolic plane using LaBNE+HM (Alanis-Lobato et al. 2016b), a method that combines manifold learning (Alanis-Lobato et al. 2016a) and maximum likelihood estimation (Papadopoulos et al. 2015) for fast yet accurate embeddings. We used the method implemented in R's package NetHypGeom (https://github.com/galanisl/NetHypGeom) with parameters $\gamma = 2.644$, $T = 0.827$ and

$w = 2\pi$. We mapped the HINT network to $\mathbb{H}^2$ using the same method with parameters $\gamma = 2.115$, $T = 0.882$ and $w = 2\pi$.

Evaluation of the hyperbolic map

In order to ensure that the inferred geometry of the hPIN and HINT was compatible with their network topology, we considered four network-structure-related criteria: (i) We divided the range of hyperbolic distances between proteins into 20 bins and, for each one, computed the fraction of protein pairs that are connected in the network (i.e. the connection probability for a distance window). We checked if these empirical connection probabilities agree with those predicted by the Popularity-Similarity model (PSM). The PSM is a model of network formation in which nodes connect with each other if their hyperbolic distance is small enough (Papadopoulos et al. 2012). (ii) We checked if expected node degrees $\langle k_i \rangle = \sum_{j \neq i} p_{ij}$ were similar to actual node degrees. $p_{ij} = 1/\left[1 + e^{(x_{ij} - R)/2T}\right]$ is the probability that node i forms a link with node j and depends on the hyperbolic distance x_{ij} described in the Introduction. R is the radius of the hyperbolic disc containing the network and T is the network temperature. (iii) We checked, if the clustering coefficient of the protein networks was similar to the average clustering of 10 artificial networks generated with the PSM, using the same topological properties of the hPIN and HINT. (iv) Finally, we checked whether greedy routing success rates and hop stretches were similar to those achieved in artificial networks generated with the PSM, using the same topological properties of the hPIN and HINT.

In addition, we validated the biological interpretation that Alanis-Lobato and colleagues reported for inferred protein coordinates (Alanis-Lobato et al. 2018). For this, we assigned all network proteins to six different age groups using FastaHerder2 (Mier and Andrade-Navarro 2016) and looked at the distribution of inferred radial coordinates for each one. Also, we studied how proteins from different classes (i.e. TFs, Recs, RBPs, etc.) agglomerate in the angular dimension of $\mathbb{H}^2$.

Statistical tests for DM topology and geometry

Using a z-test and a significance level $\alpha = 0.05$, we compared the size of the LCC that each disease module forms in the hPIN with a distribution of 1,000 random LCC sizes formed by as many proteins as there are in each module, but sampled uniformly at random from the set of all network nodes (see inset in Fig. 1c).

For each of the n disease proteins in a DM, we determined the shortest path and hyperbolic distance to the closest other protein in the same module (d_s and d_H, respectively). The n resulting distances were averaged ($\langle d_s \rangle$ and $\langle d_H \rangle$) and compared with a distribution of 1000 average d_s and d_H resulting from randomly sampling as many proteins from the hPIN as there are in each considered DM (see insets in Fig. 1d, e). A z-test and a significance level $\alpha = 0.05$ were employed in both cases.

GO semantic similarities

We divided the range of hyperbolic distances between proteins into 20 bins and, for each one, computed the average Gene Ontology (GO) semantic similarity of protein pairs within the distance window. GO semantic similarities are a valuable means to quantify the level of similitude between the GO annotations associated with two genes. We used the R package GOSemSim (Yu et al. 2010) to calculate Wang similarities from the

Biological Process and Cellular Compartment aspects of GO. We decided to use Wang's index because it was formulated specifically for the GO (Wang et al. 2007).

Protein classes in the angular dimension

We split the range of inferred protein angular coordinates into 15 bins and, for each one, we carried out six Fisher's tests to determine the most over-represented protein class in the bin from the six considered classes (see above). If two or more adjacent bins were enriched for the same protein class, the whole range covered by the bins was regarded as representative for that class. We chose 15 bins as it was the minimum number necessary for all the 6 protein classes to be represented.

Disease module separation and clustering

The separation between two disease modules A and B in hyperbolic space is inspired by the shortest path-based measure proposed by Menche and colleagues (Menche et al. 2015). The separation is defined as follows:

$$s_H(A,B) = \langle d_H(A,B)\rangle - \frac{\langle d_H(A,A)\rangle + \langle d_H(B,B)\rangle}{2}$$

The more negative $s_H(A,B)$, the greater the overlap between modules A and B. We compared this distance-based measure with the original shortest path-based one $s_s(A,B)$ and with the Jaccard distance between DMs. The latter compares the size of the intersection between the members of two modules to the size of their union: $J(A,B) = 1 - |A \cap B|/|A \cup B|$ (Jaccard 1912).

The resulting pairwise separations and Jaccard distances between DMs were used for their unsupervised classification via ncMCE and hierarchical clustering with Ward's linkage (Ward Jr 1963). We used ncMCE, a non-linear dimensionality reduction algorithm, because it is parameter-free and excels at emphasising small differences between samples. Its two dimensional projections cluster similar observations along the y-axis and outline their diversity on x (Alanis-Lobato et al. 2015; Cannistraci et al. 2013). To measure the ability of ncMCE to separate DMs into reference categories (ICD-10) along the y-axis, we employed the C-score. The C-score ranges from 0 to 1, where 0 corresponds to a completely random clustering and 1 to a perfect ordering of samples along a single dimension, i.e. disease types appear one after the other with no DMs belonging to one, mixed with the other (Alanis-Lobato et al. 2015; Zagar et al. 2011).

Navigability impact of faulty proteins

GR of signals involved 100 experiments with 500 sources and 500 targets each. Sources and targets were selected at random from the hPIN or from a pool of Recs or TFs. We used pools with the same amount of Recs and TFs (500 randomly-selected) to make sure that the observed effects were not due to different abundances of these protein types in the hPIN. In addition, we formed pools of non-Recs and non-TFs with degrees similar to the ones exhibited by the actual members of each class (Control experiments). The resulting Rec-TF efficiencies and hop stretches were compared to reference and Control via Mann-Whitney U tests.

For each DM, we calculated the impact on navigability as the average of the 50 efficiencies resulting from routing signals between 500 source-target pairs (with 20 faulty proteins sampled from the set of DM members) minus the reference average efficiency of the hPIN (0.82). The same was done for faulty proteins sampled from a pool of proteins with degrees similar to the DM members'. To assess if the difference between both cases was significant, we computed the 50 impacts separately, compared them with a Mann-Whitney U test and considered a significance level $\alpha = 0.05$.

GO and REACTOME enrichment analyses were carried out with R's package FunEnrich (https://github.com/galanisl/FunEnrich). *P*-values were corrected with Benjamini-Hochberg's method.

To compare the proportion of frequent faulty proteins that are FDA-approved or potential drug targets with the proportions in the infrequent faulty protein set, we used a Fisher's test and considered a significance level $\alpha = 0.05$.

Hardware used for experiments

We executed the experiments presented in this paper on a Lenovo ThinkPad 64-bit with 7.7 GB of RAM and an Intel Core i7-4600U CPU @ 2.10 GHz × 4, running Ubuntu 16.04 LTS. The only exceptions were the greedy routing experiments, which we executed on nodes with 30 GB of RAM, within the Mogon computer cluster at the Johannes Gutenberg Universität in Mainz.

Additional files

Additional file 1: Supplementary files. (ZIP 3154 kb)

Additional file 2: Supplementary information. (PDF 5687 kb)

Acknowledgements
The authors would like to thank the Zentrum für Datenverarbeitung at the Johannes Gutenberg Universität in Mainz, which made memory-intensive experiments possible in the Mogon computer cluster.

Availability of data and materials
The network data used in this paper, the inferred hyperbolic coordinates of its proteins, their class memberships and the gene-disease associations are provided as Additional file 1.

Authors' contributions
FH and GAL designed, implemented and carried out the experiments. FH and GAL wrote the manuscript, incorporating comments, contributions and corrections from MAAN. GAL and MAAN supervised the research. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Faculty for Physics, Mathematics and Computer Science, Johannes Gutenberg Universität, Institute of Computer Science, Staudingerweg 7, 55128, Mainz, Germany. [2]Faculty of Biology, Johannes Gutenberg Universität, Institute of Molecular Biology, Ackermannweg 4, 55128, Mainz, Germany.

Received: 30 March 2018 Accepted: 28 May 2018
Published online: 18 June 2018

References

Agrawal M, Zitnik M, Leskovec J (2018) Large-scale analysis of disease pathways in the human interactome. In: Pacific Symposium on Biocomputing, vol 23. World Scientific Publishing Company, Singapore. pp 111–122

Alanis-Lobato G, Andrade-Navarro MA (2016) Distance distribution between complex network nodes in hyperbolic space. Compl Syst 25(3):223–236

Alanis-Lobato G, Cannistraci CV, Eriksson A, Manica A, Ravasi T (2015) Highlighting nonlinear patterns in population genetics datasets. Sci Rep 5:8140. https://doi.org/10.1038/srep08140

Alanis-Lobato G, Mier P, Andrade-Navarro MA (2016a) Efficient embedding of complex networks to hyperbolic space via their Laplacian. Sci Rep 6:30,108. https://doi.org/10.1038/srep30108

Alanis-Lobato, G, Mier P, Andrade-Navarro MA (2016b) Manifold learning and maximum likelihood estimation for hyperbolic network embedding. Appl Netw Sci 1(1):10. https://doi.org/10.1007/s41109-016-0013-0

Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH (2017) HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res 45(D1):D408–D414. https://doi.org/10.1093/nar/gkw985

Alanis-Lobato G, Mier P, Andrade-Navarro MA (2018) The latent geometry of the human protein interaction network. Bioinformatics bty206. https://doi.org/10.1093/bioinformatics/bty206

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97

Allard A, Serrano MA (2018) Navigable maps of structural brain networks across species. ArXiv e-prints 1801.06079. https://arxiv.org/abs/1801.06079

Alter BP (2014) Fanconi anemia and the development of leukemia. Best Pract Res Clin Haematol 27(3-4):214–221. https://doi.org/10.1016/j.beha.2014.10.002

Aste T, Di Matteo T, Hyde S (2005) Complex networks on hyperbolic surfaces. Physica A 346(1-2):20–26

Aste T, Gramatica R, Di Matteo T (2012) Exploring complex networks via topological embedding on surfaces. Phys Rev E 86(3):036,109. https://doi.org/10.1103/PhysRevE.86.036109

Baird G (2013) Classification of diseases and the neurodevelopmental disorders: the challenge for dsm-5 and icd-11. Dev Med Child Neurol 55(3):200–201. https://doi.org/10.1111/dmcn.12087

Barthélemy M (2011) Spatial networks. Phys Rep 499(1-3):1–101. https://doi.org/10.1016/j.physrep.2010.11.002

Bharadwaj U, Eckols TK, Kolosov M, Kasembeli MM, Adam A, Torres D, Zhang X, Dobrolecki LE, Wei W, Lewis MT, Dave B, Chang JC, Landis MD, Creighton CJ, Mancini MA, Tweardy DJ (2014) Drug-repositioning screening identified piperlongumine as a direct stat3 inhibitor with potent activity against breast cancer. Oncogene 34:1341. https://doi.org/10.1038/onc.2014.72

Bianconi G, Rahmede C (2017) Emergent hyperbolic network geometry. Sci Rep 7:41,974. https://doi.org/10.1038/srep41974

Blochle R, Lall P, Cherr GS, Harris LM, Dryjski ML, Hsu HK, Dosluoglu HH (2008) Abdominal aortic aneurysms. Am J Surg 196(5):697–702. https://doi.org/10.1016/j.amjsurg.2008.07.011

Bogliolo M, Surrallés J (2015) Fanconi anemia: a model disease for studies on human genetics and advanced therapeutics. Curr Opin Genet Dev 33:32–40. https://doi.org/10.1016/j.gde.2015.07.002

Boguñá M, Krioukov D, Claffy KC (2009) Navigability of complex networks. Nat Phys 5(1):74–80. https://doi.org/10.1038/nphys1130

Boguñá M, Papadopoulos F, Krioukov D (2010) Sustaining the Internet with hyperbolic mapping. Nat Commun 1(62). https://doi.org/10.1038/ncomms1063

Brito GC, Andrews DW (2011) Removing bias against membrane proteins in interaction networks. BMC Syst Biol 5(1):169. https://doi.org/10.1186/1752-0509-5-169

Bruwier A, Chantrain CF (2011) Hematological disorders and leukemia in children with down syndrome. Eur J Pediatr 171(9):1301–1307. https://doi.org/10.1007/s00431-011-1624-1

Cannistraci CV, Alanis-Lobato G, Ravasi T (2013) Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. Bioinformatics 29(13):i199–i209. https://doi.org/10.1093/bioinformatics/btt208

Chiang CC, Lin CL, Peng CL, Sung FC, Tsai YY (2014) Increased risk of cancer in patients with early-onset cataracts: A nationwide population-based study. Cancer Sci 105(4):431–436. https://doi.org/10.1111/cas.12360

Coenen MJH, Trynka G, Heskamp S, Franke B, van Diemen CC, Smolonska J, van Leeuwen M, Brouwer E, Boezen MH, Postma DS, Platteel M, Zanen P, Lammers JWWJ, Groen HJM, Mali WPTM, Mulder CJ, Tack GJ, Verbeek WHM, Wolters VM, zHouwen RHJ, Mearin ML, van Heel DA, Radstake TRDJ, van Riel PLCM, Wijmenga C, Barrera P, Zhernakova A (2009) Common and different genetic background for rheumatoid arthritis and coeliac disease. Hum Mol Genet 18(21):4195–4203. https://doi.org/10.1093/hmg/ddp365

Cowen L, Ideker T, Raphael BJ, Sharan R (2017) Network propagation: a universal amplifier of genetic associations. Nat Rev Genet 18(9):551–562. https://doi.org/10.1038/nrg.2017.38

Cremona CA, Behrens A (2013) ATM signalling and cancer. Oncogene 33(26):3351–3360. https://doi.org/10.1038/onc.2013.275

Dall J, Christensen M (2002) Random geometric graphs. Phys Rev E 66(1):016,121. https://doi.org/10.1103/PhysRevE.66.016121

Das J, Yu H (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 6(1):92

Ding HW, Deng CL, Li DD, Liu DD, Chai SM, Wang W, Zhang Y, Chen K, Li X, Wang J, Song SJ, Song HR (2018) Design, synthesis and biological evaluation of novel 4-aminoquinazolines as dual target inhibitors of egfr-pi3kα. Eur J Med Chem 146:460–470. https://doi.org/10.1016/j.ejmech.2018.01.081

Ferretti L, Cortelezzi M (2011) Preferential attachment in growing spatial networks. Phys Rev E 84(1):016,103. https://doi.org/10.1103/PhysRevE.84.016103

Ferretti L, Cortelezzi M, Mamino M (2014) Duality between preferential attachment and static networks on hyperbolic spaces. Europhys Lett 105(3):38,001. https://doi.org/10.1209/0295-5075/105/38001

Fienberg S, Cozier GE, Acharya KR, Chibale K, Sturrock ED (2018) The design and development of a potent and selective novel diprolyl derivative that binds to the n-domain of angiotensin-i converting enzyme. J Med Chem 61(1):344–359

García-Pérez G, Boguñá M, Allard A, Serrano MA (2016) The hidden hyperbolic geometry of international trade: World Trade Atlas 1870–2013. Sci Rep 6:33,441. https://doi.org/10.1038/srep33441

Gerstberger S, Hafner M, Tuschl T (2014) A census of human RNA-binding proteins. Nat Rev Genet 15(12):829–845. https://doi.org/10.1038/nrg3813

Ghiassian SD, Menche J, Barabási AL (2015) A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol 11(4):e1004,120. https://doi.org/10.1371/journal.pcbi.1004120

Gilmore EC (2014) DNA repair abnormalities leading to ataxia: shared neurological phenotypes and risk factors. Neurogenetics 15(4):217–228. https://doi.org/10.1007/s10048-014-0415-z

Heidemann LN, Hartwell D, Heidemann CH, Jochumsen KM (2013) The relation between endometriosis and ovarian cancer - a review. Acta Obstet Gynecol Scand 93(1):20–31. https://doi.org/10.1111/aogs.12255

Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytol 11(2):37–50

Janku F, Hong DS, Fu S, Piha-Paul SA, Naing A, Falchook GS, Tsimberidou AM, Stepanek SL Vanda MMoulder, Lee JJ, Luthra R, Zinner RG, Broaddus RR, Wheler JJ, Kurzrock R (2013) ssessing pik3ca and pten in early-phase trials with pi3k/akt/mtor inhibitors. Cell Rep 6(2):377–387. https://doi.org/10.1016/j.celrep.2013.12.035

Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. Nature 409:853–855

Kent KC (2014) Abdominal aortic aneurysms. N Engl J Med 371(22):2101–2108. https://doi.org/10.1056/NEJMcp1401430

Kleinberg JM (2000) Navigation in a small world. Nature 406(6798):845–845

Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4):949–958. https://doi.org/10.1016/j.ajhg.2008.02.013

Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguñá M (2010) Hyperbolic geometry of complex networks. Phys Rev E 82(3):036,106. https://doi.org/10.1103/PhysRevE.82.036106

Lage K, zKarlberg EO, Størling ZM, Ólason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25(3):309–316. https://doi.org/10.1038/nbt1295

Luck K, Sheynkman GM, Zhang I, Vidal M (2017) Proteome-Scale Human Interactomics. Trends Biochem Sci 42(5):342–354. https://doi.org/10.1016/j.tibs.2017.02.006

Lukong KE, Chang Kw, Khandjian EW, Richard S (2008) RNA-binding proteins in human genetic disease. Trends Genet 24(8):416–425. https://doi.org/10.1016/j.tig.2008.05.004

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL (2015) Uncovering disease-disease relationships through the incomplete interactome. Science 347(6224):1257,601. https://doi.org/10.1126/science.1257601

Mier P, Andrade-Navarro MA (2016) FastaHerder2: four ways to research protein function and evolution with clustering and clustered databases. J Comput Biol 23:270–278. https://doi.org/10.1089/cmb.2015.0191

Ortiz E, Starnini M, Serrano MA (2017) Navigability of temporal networks in hyperbolic space. Sci Rep 7:15,054. https://doi.org/10.1038/s41598-017-15041-0

Papadopoulos F, Krioukov D, Boguñá M, Vahdat A (2010) Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In: INFOCOM, 2010 Proceedings IEEE. pp 1–9. https://doi.org/10.1109/INFCOM.2010.5462131

Papadopoulos F, Kitsak M, Serrano MA, Boguñá M, Krioukov D (2012) Popularity versus similarity in growing networks. Nature 489(7417):537–540. https://doi.org/10.1038/nature11459

Papadopoulos F, Aldecoa R, Krioukov D (2015) Network geometry inference using common neighbors. Phys Rev E 92(2):022,807. https://doi.org/10.1103/PhysRevE.92.022807

Piñero J, Bravo A, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 45(D1):D833–D839. https://doi.org/10.1093/nar/gkw943

Pontikaki A, Sifakis S, Spandidos DA (2015) Endometriosis and breast cancer: A survey of the epidemiological studies. Oncol Lett 11(1):23–30. https://doi.org/10.3892/ol.2015.3895

Roy D, Morgan M, Yoo C, Deoraj A, Roy S, Yadav V, Garoub M, Assaggaf H, Doke M (2015) Integrated bioinformatics, environmental epidemiologic and genomic approaches to identify environmental and molecular links between endometriosis and breast cancer. Int J Mol Sci 16(10):25,285–25,322. https://doi.org/10.3390/ijms161025285

Sakurai T, Kudo M (2013) Molecular link between liver fibrosis and hepatocellular carcinoma. Liver Cancer 2(3-4):365–366. https://doi.org/10.1159/000343851

Serrano MA, Boguñá M, Sagués F (2012) Uncovering the hidden geometry behind metabolic networks. Mol BioSyst 8(3):843–850. https://doi.org/10.1039/c2mb05306c

Southan C, Sharman JL, Benson HE, Faccenda E, Pawson AJ, Alexander S, Buneman OP, Davenport AP, McGrath JC, Peters JA, Spedding M, Catterall WA, Fabbro D, Davies JA, NC-IUPHAR (2015) The IUPHAR/BPS guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res 44(D1):D1054–D1058. https://doi.org/10.1093/nar/gkv1037

Taylor IW, Wrana JL (2012) Protein interaction networks in medicine and disease. Proteomics 12(10):1706–1716. https://doi.org/10.1002/pmic.201100594

Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, Bäckström A, Danielsson F, Fagerberg L, Fall J, Gatto L, Gnann C, Hober S, Hjelmare M, Johansson F, Lee S, Lindskog C, Mulder J, Mulvey CM, Nilsson P, Oksvold P, Rockberg J, Schutten R, Schwenk JM, Sivertsson A, Sjöstedt E, Skogs M, Stadler C, Sullivan DP, Tegel H, Winsnes C, Zhang C, Zwahlen M, Mardinoglu A, Pontén F, von Feilitzen K, Lilley KS, Uhlén M, Lundberg E (2017) A subcellular map of the human proteome. Science 356(6340):eaal3321. https://doi.org/10.1126/science.aal3321

Tolman KG, Dalpiaz AS (2007) Treatment of non-alcoholic fatty liver disease. Ther Clin Risk Manag 3(6):1153–1163

Turner M, Monzón-Casanova E (2017) Rna-binding proteins mind the gaps. Nat Immunol 18:146–148. https://doi.org/10.1038/ni.3662

Uehara T, Ainslie GR, Kutanzi K, Pogribny IP, Muskhelishvili L, Izawa T, Yamate J, Kosyk O, Shymonyak S, Bradford BU, Boorman GA, Bataller R, Rusyn I (2013) Molecular mechanisms of fibrosis-associated promotion of liver carcinogenesis. Toxicol Sci 132(1):53–63. https://doi.org/10.1093/toxsci/kfs342

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CAK, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F (2015) Tissue-based map of the human proteome. Science 347(6220):1260,419. https://doi.org/10.1126/science.1260419

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10(4):252–263. https://doi.org/10.1038/nrg2538

Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási AL, Vidal M (2009) An empirical framework for binary interactome mapping. Nat Methods 6(1):83–90. https://doi.org/10.1038/nmeth.1280

Vercellini P, Viganó P, Somigliana E, Fedele L (2013) Endometriosis: pathogenesis and treatment. Nat Rev Endocrinol 10(5):261–275. https://doi.org/10.1038/nrendo.2013.255

Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE (2011) A directed protein interaction network for investigating intracellular signal transduction. Sci Signal 4(189):rs8. https://doi.org/10.1126/scisignal.2001699

Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23(10):1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Ward Jr JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845

WHO (2016) International Statistical Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2016/en. Accessed 5 Feb 2018

Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4:189. https://doi.org/10.1038/msb.2008.27

Wu Z, Menichetti G, Rahmede C, Bianconi G (2015) Emergent complex network geometry. Sci Rep 5:10,073. https://doi.org/10.1038/srep10073

Xavier AC, Ge Y, Taub JW (2009) Down syndrome and malignancies: A unique clinical relationship. J Mol Diagn 11(5):371–380. https://doi.org/10.2353/jmoldx.2009.080132

Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, Stupp GS, Putman TE, Ainscough BJ, Griffith OL, Torkamani A, Whetzel PL, Mungall CJ, Mooney SD, Su AI, Wu C (2016) High-performance web services for querying gene and variant annotation. Genome Biol 17(1). https://doi.org/10.1186/s13059-016-0953-9

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91(6):1022–1032. https://doi.org/10.1016/j.ajhg.2012.10.015

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26(7):976–978. https://doi.org/10.1093/bioinformatics/btq064

Zagar L, Mulas F, Garagna S, Zuccotti M, Bellazzi R, Zupan B (2011) Stage prediction of embryonic stem cell differentiation from genome-wide expression data. Bioinformatics:2546–2553. https://doi.org/10.1093/bioinformatics/btr422

Zhang HM, Liu T, Liu CJ, Song S, Zhang X, Liu W, Jia H, Xue Y, Guo AY (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res 43(D1):D76–D81. https://doi.org/10.1093/nar/gku887