OPEN ACCESS Freely available online

PLOS ONE

# Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action

**Ruggero Gramatica[1,2], T. Di Matteo[1], Stefano Giorgetti[2], Massimo Barbiani[2], Dorian Bevec[2], Tomaso Aste[3]***

**1** Department of Mathematics, King's College London, London, United Kingdom, **2** Therametrics AG, Stans, Switzerland, **3** Department of Computer Science, University College London, London, United Kingdom

## Abstract

We introduce a methodology to efficiently exploit natural-language expressed biomedical knowledge for repurposing existing drugs towards diseases for which they were not initially intended. Leveraging on developments in Computational Linguistics and Graph Theory, a methodology is defined to build a graph representation of knowledge, which is automatically analysed to discover hidden relations between any drug and any disease: these relations are specific paths among the biomedical entities of the graph, representing possible Modes of Action for any given pharmacological compound. We propose a measure for the likeliness of these paths based on a stochastic process on the graph. This measure depends on the abundance of indirect paths between a peptide and a disease, rather than solely on the strength of the shortest path connecting them. We provide real-world examples, showing how the method successfully retrieves known pathophysiological Mode of Action and finds new ones by meaningfully selecting and aggregating contributions from known bio-molecular interactions. Applications of this methodology are presented, and prove the efficacy of the method for selecting drugs as treatment options for rare diseases.

**Citation:** Gramatica R, Di Matteo T, Giorgetti S, Barbiani M, Bevec D, et al. (2014) Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action. PLoS ONE 9(1): e84912. doi:10.1371/journal.pone.0084912

**Editor:** Renaud Lambiotte, University of Namur, Belgium

**Received** June 20, 2013; **Accepted** November 28, 2013; **Published** January 9, 2014



**Funding:** The authors have no support or funding to report.

**Competing Interests:** The authors declare competing financial interests: RG, DB, SG and MB are employed by Therametrics (formerly Mondobiotech AG) and detain stock options of the Company. The presented methodology is part of the research tools currently employed and licensed to Therametrics AG. The methodology described in this paper, concerning the use of the semantic approach to investigate public knowhow and building a knowledge network, alongside the exploitation of certain graph theory instruments to discover emergent patterns, has been filed by RG at the European International Patent Office (application PCT/EP2013/050056). The financial competing interests do not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: t.aste@ucl.ac.uk

## Introduction

In pharmaceutical research the subject of *drug repurposing* is rapidly raising significant interest. Repurposing means redirection of clinically advanced or marketed products into certain diseases rather than in the initially intended indications. A significant advantage of repurposing drugs is their demonstrated clinical pharmacological efficacy and safety profile. Repurposing is especially interesting in the area of life-threatening Rare or Orphan diseases with high unmet medical need. The hypothesis for drug repurposing is based on the drugs' side effects profiles, indicating interaction with more than one cellular target. These pathway interactions open up the opportunity to exploit existing medicines towards other diseases.

Extensive data sets describing drug effects have been published globally, resulting in a huge amount of information publically available in large on-line collections of bio-medical publications such as PubMed (http://www.ncbi.nlm.nih.gov/pubmed/).

This is an opportunity for literature-based scientific discovery; see [1–15,54], [2] and [3]. However, important pieces of information regarding chemical substances, biological processes and pathway interactions are scattered between publications from different communities of scientists, who are not always mutually aware of their findings. In order to generate a working hypothesis from such a body of literature, a researcher would need to read thoroughly all the relevant publications and to pick among them the relevant items of information. Search engines help scientists in this endeavour, but are unable to semantically aggregate information from different sources, leaving all the initiative to researchers; complex relation-focused and graph-like representations (*ontologies*) have been extensively produced and used to fill the gap, since their introduction for the Semantic Web; see [16] and [17]. Yet ontologies need to be man-made and they are difficult to integrate each other and to maintain; see [18].

Here we propose an approach to literature-based research ultimately based on the *distributional hypothesis of linguistic theory* (see [19] and [20]) - whose analysis relates the statistical properties of words association to the intrinsic meaning of a concept - and *network theory* (see [21,22,54]) - a collection of versatile mathematical tools for representing interrelated concepts and analyse their connections structure.

Main aim of this work is to provide a methodology for creating network knowledge representations, capturing the essential entities occurring in a variety of publications and connecting them into a graph whenever they co-occur in a given sentence. The knowledge graph thus created can then be analysed in order to identify and

rank statistically relevant *indirect connections* among prospect medicines and diseases. We show that with a suitable set of concepts, specifically compiled in a dictionary, the linked biochemical entities in the network can be connected along paths that mimic a chain of reasoning and lead to prospect inferences about the mechanism of action of a chemical substance in the pathophysiology of a disease.

In this paper we introduce a method to rank the relevance of the inferences, introducing a measure based on a stochastic process (random walk) defined on the graph: this measure takes into account all paths connecting two concepts and uses the abundance and redundancy of these paths, together with their weights, as a measure of the strength of the overall relation between the concepts.

The paper is organized as follows: we first discuss the construction of the knowledge graph; we then introduce the tools to analyse this graph and extract possible mechanisms of action; relevant emerging indirect links between peptides and sarcoidosis are then discussed in the 'Results' section. A clinical reader might skip the methodological part and go directly to the 'Results' section.

## Methods

### Construction of the knowledge graph

In the field of linguistics it is commonly accepted that the meaning of a word must be inferred by examining its occurrences over large *corpora* of text. Adopting this perspective (see [23–25]), one can say that the meaning of a word ultimately depends on the words it mostly goes along with: this is the basis of the so-called "Distributional Hypothesis" introduced by Firth in 1957. The general idea shows that there is a correlation between distributional similarity and meaning similarity, which allows exploiting the former in order to derive the latter. This hypothesis suggests the assumption that concepts occurring in the same unit of text are in some way semantically related. Let us note that co-occurrence is nowadays a common method to find a relationship between biomedical concepts; co-occurrence methods are commonly used to discover new and hidden relations, following the seminal work of Swanson (see [15], [25] and the more detailed works [26–30]). Some authors (e.g. see [31–36]) use networks to map specific biomedical entities such as protein-protein interactions, gene regulatory events and links between proteins and phosphorylation or genes interactions. Our aim is to build and use a co-occurrence network of biomedical concepts to produce inferences that are new hypotheses for drug repurposing. The key idea exploited in this paper is that hopping through this knowledge network and drawing a path between any two non-adjacent concepts can be interpreted as suggesting a possible "sentence" that has never actually been uttered but that can implicitly carry a new and correct idea.

Let us here start by describing in details how our knowledge network of peptides, related biological processes and rare diseases is built (see Fig. 1). We have extracted three million PubMed paper abstracts – out of a total of more than 20 million – using keyword searches on a list of 1606 concepts, comprising 127 peptides, 300 rare diseases and 1179 other biological entities such as chemical compounds, proteins, receptors, enzymes, hormones and physiological entities (e.g. cells, organs, tissues, pathways, processes). Every abstract has then been broken down into its constituent sentences. The full list of concepts and the full list of paper IDs are provided as File S1 and File S2. Entity recognition has been carried out on every sentence, following a dictionary-based approach (see [12,25]). Specifically, we built a dictionary, enriching each item of our biomedical item list with a set of acronyms, synonyms and other identifying phrases gathered from MeSH (Medical Subject Headings), Orpha.net and the "cope with cytokines" web site. Of course different concepts may share some of the identifying expressions. This is the *polysemy* problem, i.e. the capacity for a word or a phrase to have multiple meanings that leads to the necessity of disambiguation (see [25], [37]). This is a very complex problem in general and, to tackle with disambiguation, we employed a version of the Lesk algorithm (see [38]). Whenever disambiguation fails, we have chosen to keep both the possible concepts: this option reduces precision but maximizes recall – i.e. the quantity of relevant concepts that are retrieved. Many other errors in the detection of co-occurrences arise beyond the ones due to failed disambiguation: a sentence boundary may be misplaced, one of the occurrences may be a false positive or the occurrences may be just part of a list (and therefore not semantically related). It is expected though that as more and more papers are analysed the meaningful co-occurrences will outgrow the spurious ones: in fact "real" co-occurrences are repeated consistently as more and more literature is considered, while spurious ones become statistically insignificant because the same concept is linked randomly to a great number of other concepts. In a figurative manner we may think of a "noise" in the co-occurrence detection that becomes negligible as a large number of papers are considered. We thus obtain a co-occurrence network (Fig. 2a) where the biomedical concepts of our dictionary are the nodes and the co-occurrence frequency is the weight of the edges. The resulting network is sparse with a small number of links (158,428) compared to the complete graph (12.7%) but, nonetheless, only 30 concepts are not connected to the giant component of the network, thus comprising 1576 nodes (98.13% of the total). The diameter (i.e. the maximum distance between any two nodes) of this network is 4 with an average path length (i.e. average distance between any two nodes; see [9,10]) of 1.95. It is observed that the graph contains *hubs* interpreted as physiological processes typical of diseases (e.g. inflammation, proliferation, necrosis), immune system-related items (e.g. white blood cells, cytokines) and the major organs – especially the ones dealing with chemical elaboration of drugs (e.g. kidney, liver). A number of direct connections of "*peptides – diseases*" are present, such as ANGIOTENSIN – SARCOIDOSIS or ANGIOTENSIN – DIABETIC NEPHROPATHY (see Fig. 2b). The relations between those peptides and diseases are already known as we expected on the ground that they appear together in a predicate. Indeed, Angiotensin is known to worsen Sarcoidosis symptoms, while it is of aid in diabetic nephropathy. These features are interpreted as a positive feedback on the meaningfulness of the knowledge graph.

### Analysis of the knowledge graph

Once the Knowledge Graph is built, we are in the position to analyse it in order to highlight new scientifically analysable relations between a peptide and a rare disease. We search for indirect relations in the network (Fig. 2a) and therefore for a *path* (see [21,22]) between a peptide and rare disease (Fig. 2b). Since all nodes in the network are connected, these paths always exist: the challenge is to rank them (in order to find the most significant ones) and to explore and choose those paths that suggest understandable and yet non-trivial inferences.

Shorter paths must be considered more relevant, as more steps introduce new levels of indirection and magnify the effects of randomness and noise. Yet the paths cannot be too short, because they must be "verbose" enough to suggest a rationale to indicate the biological *Mechanisms of Action* (MoA), i.e. a specific biochemical interaction through which a drug substance produces its

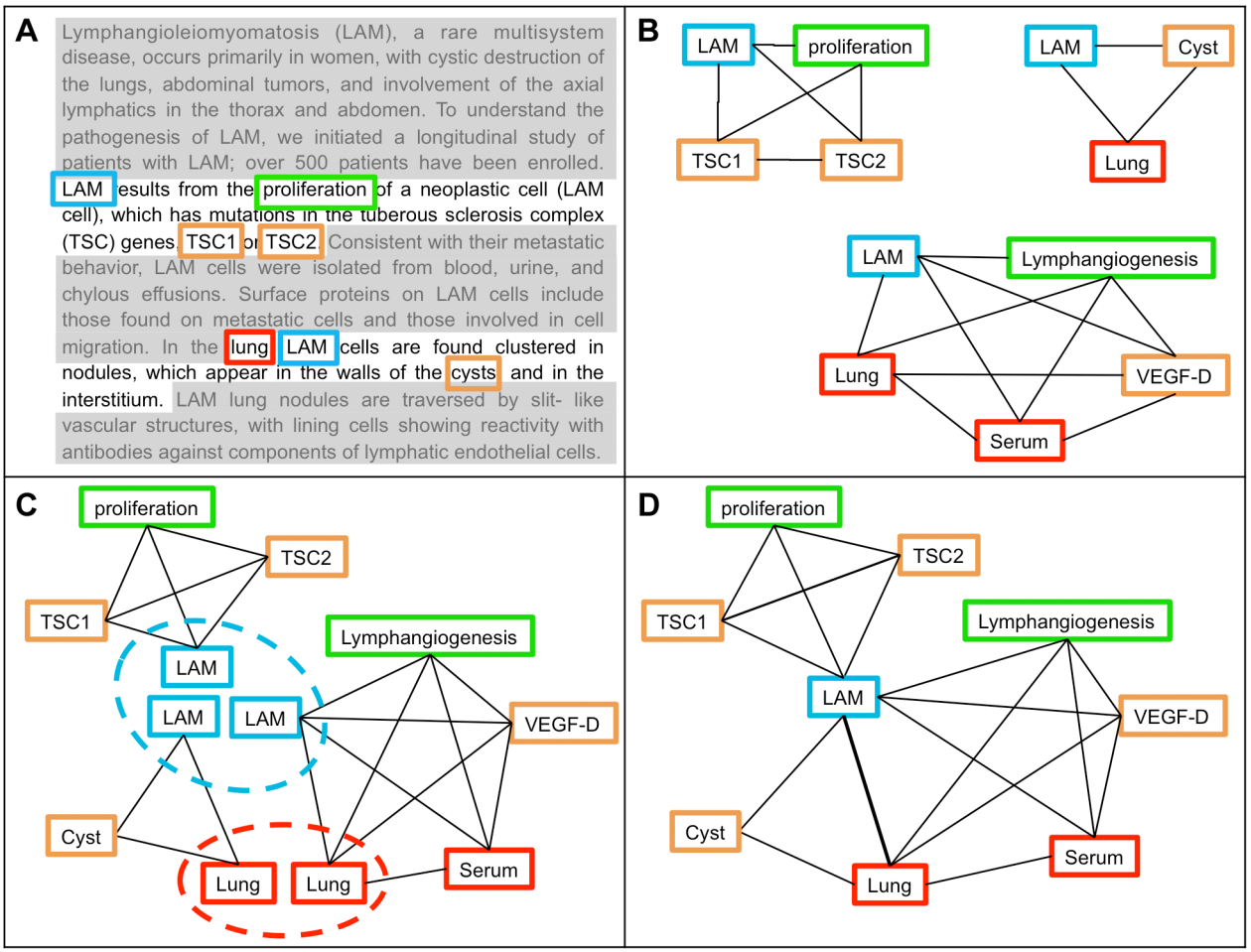


**Figure 1. Conceptual outline of the knowledge graph building process.** (A) Every document is split into its constituent sentences and each of them is scanned to identify expressions registered on the dictionary. In the figure, two sentences are highlighted and the matching expressions are enclosed in coloured boxes. Every one of these expressions is associated to a concept in the dictionary. (B) The concepts co-occurring in a sentence are connected pairwise. A sentence is therefore abstracted as a complete graph where the occurring concepts are the nodes and a single co-occurrence is a link. The weight of a link is increased if more instances of the same co-occurrence are present. (C) The sentence graphs are then merged in such a way that each node (concept) appears only once in the graph. In the figure it is evident that the «LAM» node (abbreviation for Lymphangioleiomyomatosis – a rare disease) appears in every graph and the «Lung» node in two of them. (D) The result of the merging is a new graph – which is no more complete – where the weight of the link is associated to the frequency of the same co-occurrence.
doi:10.1371/journal.pone.0084912.g001

pharmacological effect amongst molecular targets like cell receptors, proteins or enzymes; in other words the MoA explains why and how a drug substance works. Specifically, when dealing with peptides, the MoA, that we aim to replicate, is the one where a peptide binds to its specific receptors, thus activating or modulating a physiological process involved in the disease. To achieve such characteristics, we consider specific interactions (links) among nodes, filtering out unwanted information. For instance, a peptide may be connected to any node but since we look for mechanisms of actions, only links in the form of *peptide–cell receptor* are allowed and therefore considered in the graph. Similarly, a receptor can be either involved in a pathway or influence directly a biological process, thus only links in the form *cell receptor–process* or *cell receptor–protein* are allowed. To this purpose, every item in our dictionary is assigned to one of the following categories: AMINOACID, BACTERIA, CELL, DISEASE, DRUG, ENZYME, GENE, HORMONE, NUCLEOTIDE, ORGAN, OTHER, PATHWAY, PEPTIDE, PROCESS, PROTEIN, RECEPTOR, VITAMIN.

From the mathematical perspective, co-occurrences define the coefficients of the *similarity matrix* $A$ representing the weighted graph. Through a suitable normalization of $A$ we are able to find a probabilistic interpretation for the link weights. Specifically, posing

$$w_{ij} = \frac{a_{ij}}{\sum_k a_{ik}}$$

where $a_{ij}$ is the $i,j$ element of the similarity matrix, which is zero if the vertices $i,j$ are not directly connected and equal to the edge weight otherwise (see also [54]). Therefore we interpret the components $w_{ij}$ as the conditional probability $p(j|i)$ of finding concept $j$ in a co-occurrence containing concept $i$. Since the coefficients of the matrix W are in the range (0,1], we can also introduce a *dissimilarity measure*

$$d(i,j) = -\log(W)$$

which is correctly defined in the range $d(i,j) \in [0,\infty)$ and,

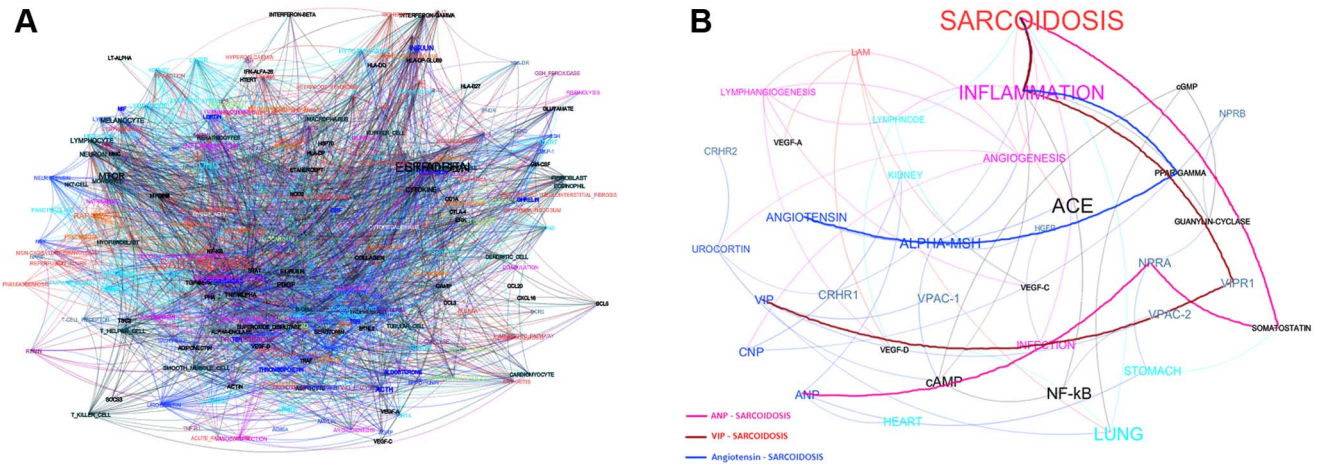


**Figure 2. Paths identification and selection.** (A) This figure shows a version of the graph – simplified for illustration purposes – built focusing onto 300 concepts and with 200,000 documents. (B) This figure shows three automatically retrieved and meaningful paths, identifying three – out of five – prospect candidate peptides for sarcoidosis. The paths are depicted in a further simplified version of the graph obtained from the first one by filtering out nodes not relevant to the paths.
doi:10.1371/journal.pone.0084912.g002

oppositely to weights returns larger values for smaller similarities. This distance representation allows immediate application of the available algorithms for computing *shortest paths* (see [22]). With this definition, the shortest path between two any given nodes $i$ and $j$ represents the most probable path (and therefore in our interpretation, the most probable MoA) connecting them. In fact, the shortest path is the one $\pi_{ij}=iv_1\ldots v_kj$ that minimizes the total distance

$$D(\pi_{ij})=d(i,v_1)+d(v_1,v_2)+\ldots+d(v_k,j)$$

therefore, we have

$$D(\pi_{ij})=-\log w_{iv_1}-\ldots-\log w_{v_kj}=-\log\Big(\prod_{r,s\in\pi_{ij}} w_{rs}\Big) \quad (1)$$

Since $w_{ij}$ are conditional probabilities, the above equation (1) is a product of conditional probabilities (a Markov chain). Therefore the conditional probability associated to the shortest path is maximized.

## Ranking of paths using random walk

We have seen that shortest paths maximize the probability of a single MoA, but strong indirect connections between a given peptide and a given disease may arise also from a set of paths which are smaller in weight but that contribute in larger numbers. We have therefore devised a different ranking algorithm for a peptide-disease correlation that considers all paths connecting the two concepts and uses the abundance and redundancy of these paths, together with their weights, as a measure of the strength of the overall relation between the concepts.

This can be achieved by measuring the average number of time-steps required to go from one vertex to the other in the network, assuming that a walker is moving at random and that at each discrete time-step it jumps from a vertex to one of its neighbours with a probability which depends on the number of available links and to their weights. This *random walker* produces a distance that is a function of both the length and the abundance of paths [55,56,57].

Intuitively, imagine two nodes connected by one short (one step) path and many longer ones. A random walker trying the route many times will tread the longer paths more often therefore perceiving a "long" distance. Instead, if the end points are connected with a lot of medium-sized paths, the walker will tread those most of the times and thus perceiving a distance shorter than the previous one. A common-world example for conveying this idea: imagine a drunkard trying to go home. He is likely to make many mistakes at the crossroads effectively selecting the next lane at random. He is more likely to get home sooner if many roads converge to his destination rather than if only a short one goes there and the others lead astray.

From the computational perspective, the random walk distances can be computed by pure algebraic means. The computation is carried out defining a vector, where each component is the likelihood that at a given time a random walker is on a given node. The step-by-step evolution of this vector is a representation of the shifting distribution of these walkers in the nodes in their random wandering.

The probability to walk from vertex $i$ to vertex $j$ is defined in the random walk theory by the *transfer matrix* P, computed from the similarity matrix $A$ with the formula:

$$p_{ij}=\frac{a_{ij}}{\sum_k a_{ik}}=w_{ij}\ i,j=1,\ldots,N$$

which is exactly the matrix we have previously denoted W. It has been shown (see [39]) that the random walk distances of two nodes $i$ and $j$ are given by:

$$d(i,j)=\sum_{k=1}^{N}\left[\frac{1}{I-B(j)}\right]_{ik}$$

where $I$ is the identity matrix and B is a square matrix identical to P having posed

$$B(j)_{ij}=0\ \forall i$$

.The *random walk distance* built this way is non-symmetric, but for

our purposes we symmetrise it by taking the average of the two directions.

$$d_s(i,j) = \frac{1}{2}[d(i,j) + d(j,i)]$$

This distance defines an implicit ranking measure for each couple of distinct nodes and therefore between any peptide-disease couple.

Such a measure can be interpreted as the probability of finding that path, and thus the MoA, within the document base.

## Results

In this paper we show examples of rationales produced by our methodology with regard to a) the granulomatous disease *Sarcoidosis* and its pulmonary pathology, and b) Imatinib, a targeted-therapy agent against cancer cells, well known for its apoptosis action.

Sarcoidosis is a disease in which abnormal collections of chronic inflammatory cells form as nodules (granulomas) in multiple organs. Sarcoidosis is present at various level of severity in all-ethnic and racial groups and is mainly caused by environmental agents in people with higher genetic sensitivity. The disease is a chronic inflammatory disease that primarily affects the lungs but can affect almost all organs of the body. Sarcoidosis is a complex disease displaying incorrect functionalities within immune cells, cytokines, and antigenic reactions; see [40]. Fig. 3 shows a subgraph of the knowledge network comprising the concepts related with Sarcoidosis.

We were interested in using peptides to treat Sarcoidosis. Therefore a number of rationales have been obtained from a pool of peptides against sarcoid pathologies, and the most relevant findings are listed below, ranked according to the random walk distance:

1. **VIP** – VIPR1 – INFLAMMATION – **SARCOIDOSIS**
2. **α-MSH** – HGFR – INFECTION – **SARCOIDOSIS**
3. **CNP** – NPRB – GUANYLIN_CYCLASE – INFLAMMATION – **SARCOIDOSIS**

### The Match VIP – SARCOIDOSIS

Vasoactive Intestinal Peptide - VIP (also known as Aviptadil), is an endogenous human peptide. It is predominantly localized in the lungs where it binds specific receptors (VPAC-1, VPAC-2), which transform the signal into an increased production of intracellular cyclic adenosine monophosphate (cyclic AMP or cAMP), as well as into the inhibition of translocation of NF-κB from cytoplasm into the nucleus. This process regulates the production of various cytokines responsible for the inflammatory reaction, such as TNF-α. Hence, VIP is responsible for preventing or attenuating a wide variety of exaggerated pro-inflammatory activities; see [41].

The path in Fig. 4 shows that VIP is affecting the inflammation processes related to Sarcoidosis.

The scientific evidence clearly suggests VIP as a potential treatment option for Sarcoidosis: the system has been able to retrieve the main receptor of VIP and its relevance in the inflammation process.

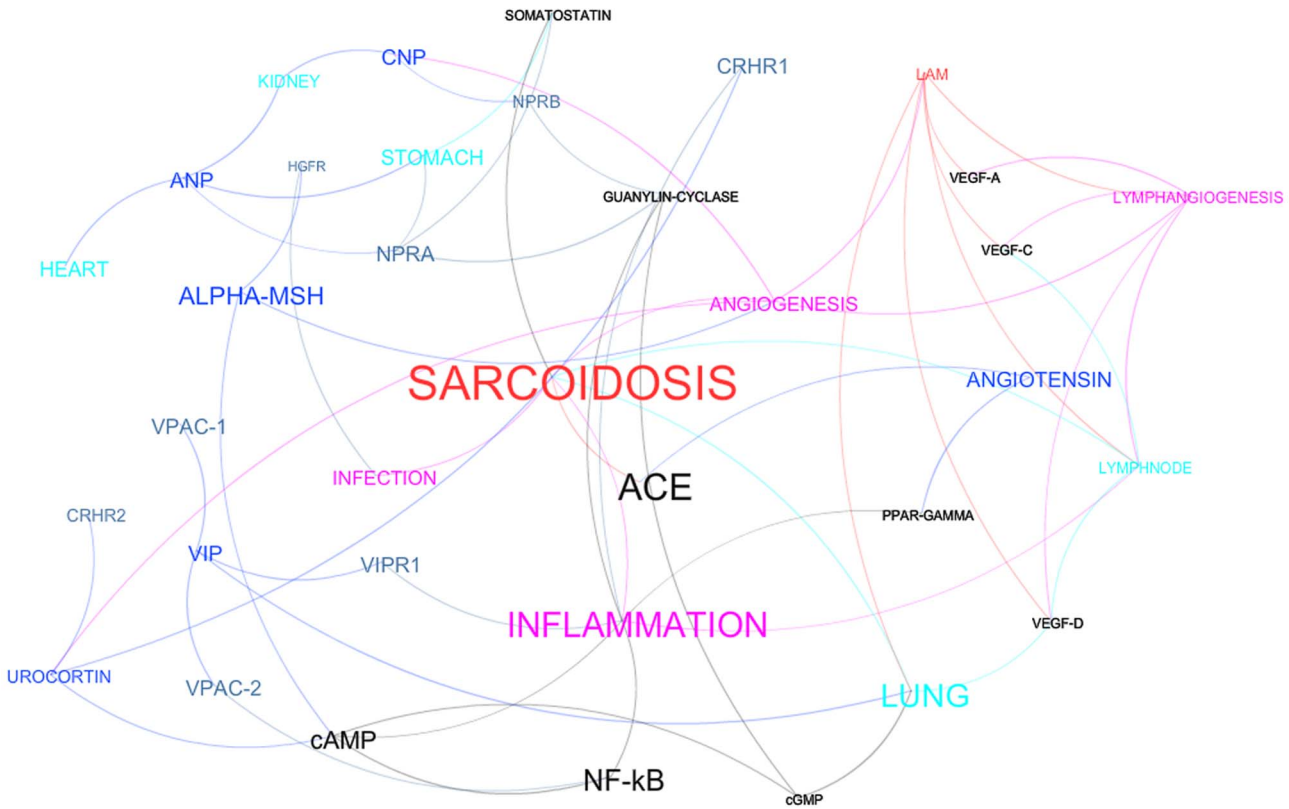


**Figure 3. The Sarcoidosis knowledge network.** A portion of the knowledge network showing the neighbourhood of Sarcoidosis. The figure is intended as a bird-eye view of the entities the system detected as related with Sarcoidosis
doi:10.1371/journal.pone.0084912.g003

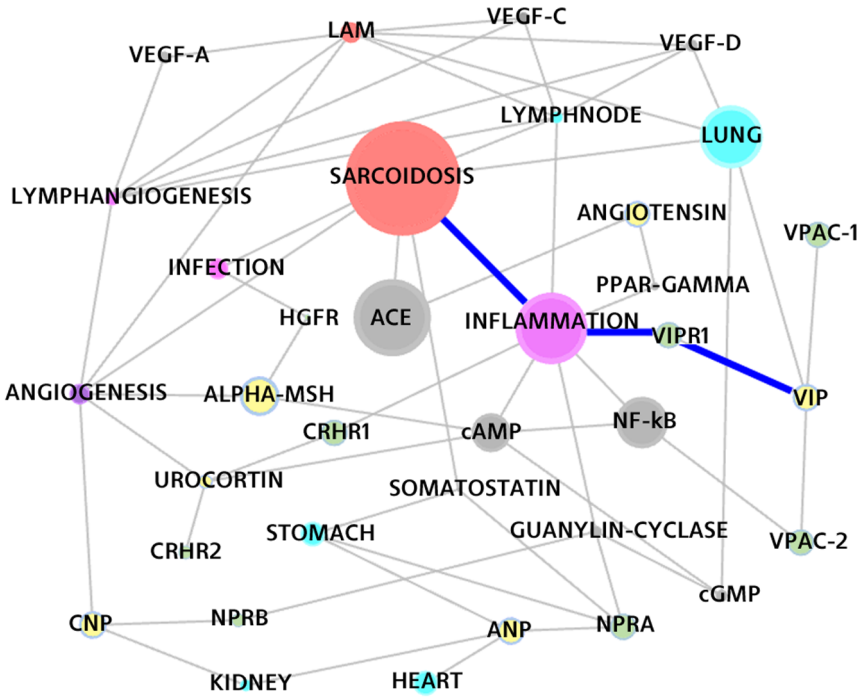


**Figure 4. The VIP – SARCOIDOSIS path and other closely related concepts.**
doi:10.1371/journal.pone.0084912.g004

## The Match ALPHA-MSH – SARCOIDOSIS

α-Melanocyte Stimulating Hormone (α-MSH) is an endogenous peptide originally described for stimulating melanogenesis, mainly for the pigmentation of the skin. Later it gained roles in feeding behaviour, sexual activity, immune responses, inflammation and fibrosis. Upon binding to its specific cell surface receptors it increases production of cAMP in the target cells and triggers four signalling pathways leading to the disruption of the transcription of several pro-inflammatory mediators genes; see [41].

In addition, α-MSH also regulates the MET proto-oncogene expression in both melanoma cells and in normal human melanocytes. The MET proto-oncogene encodes for the Hepatocyte Growth Factor Receptor (HGFR) that is involved in melanocyte growth and melanoma development; see [42].

There is evidence of interrelation between Epstein-Barr Virus (EBV) infection and MET proto-oncogene expression, and at date several infection agents have been suggested to have an implication as cause of Sarcoidosis.

A role for a transmissible agent is also suggested by the finding of granulomatous inflammation in patients without Sarcoidosis who received heart transplantation from donors who had Sarcoidosis; see [43] and [44].

The system sees both these processes (as apparent from Fig. 5), assigning a better ranking to the second one. α-MSH is another candidate for the treatment of the sarcoid-pathology due to this double action.

## The Match CNP – SARCOIDOSIS

CNP (C-type Natriuretic Peptide) is a human peptide, which elicits a number of vascular, renal, and endocrine activities, regulating blood pressure and extracellular fluid volume. When CNP binds to its receptor, NPRB, on the cell surface it activates a cell signalling through a Guanyl cyclase that increases intracellular cGMP level activating specific pathways ultimately modifying cellular functions. cGMP is known for its potent vasodilatory action in pulmonary vessels. Depending on the tissues involved, however, some of its effects are directly opposite to those of cAMP, which is a potent inhibitor of proinflammatory tumor necrosis factor (TNF-α) synthesis; see [45].

The inference subtended by the path in Fig. 6 is sound and correctly traces a biological process. Yet CNP is not considered a treatment option for Sarcoidosis because of its potential negative side effects profile due to its systemic vasodilatory characteristics.

## The Match Imatinib – Creutzfeldt-Jakob disease

Imatinib (commercialized under the name GLEEVEC) is a rationally designed pyridylpyrimidine derivative, and a highly potent and selective competitive tyrosine kinase inhibitor, especially effective in the inhibition of kinases c-Abl (Abelson proto-oncogene), c-kit, and PDGF-R (platelet-derived growth factor receptor); see [46] and [47]. These kinases are enzymes involved in cellular signal transduction processes, whose dysregulation may lead to malfunctioning of cells and disease processes, as exemplified in a variety of hyperproliferative disorders and cancers. Imatinib has been regulatory approved for chronic myelogenous leukemia (CML), gastrointestinal stromal tumors (GISTs), aggressive systemic mastocytosis (ASM), hypereosinophilic syndrome (HES), chronic eosinophilic leukemia (CEL), dermatofibrosarcoma protuberans, and Acute Lymphoblastic Leukemia (ALL).

Exploiting our methodology we looked for rationales for the redirection of Imatinib; on the basis of the results of the stochastic measure, the system indicates the neurodegenerative transmissible spongiform encephalopathies – exemplified by the Creutzfeldt-Jakob disease (CJD) – as promising targets for this drug. Transmissible spongiform encephalopathies are caused by the aberrant metabolism of the prion protein (PrP). Prions are seemingly infectious agents without a nucleic acid genome. Prion diseases belong to the group of neurodegenerative diseases

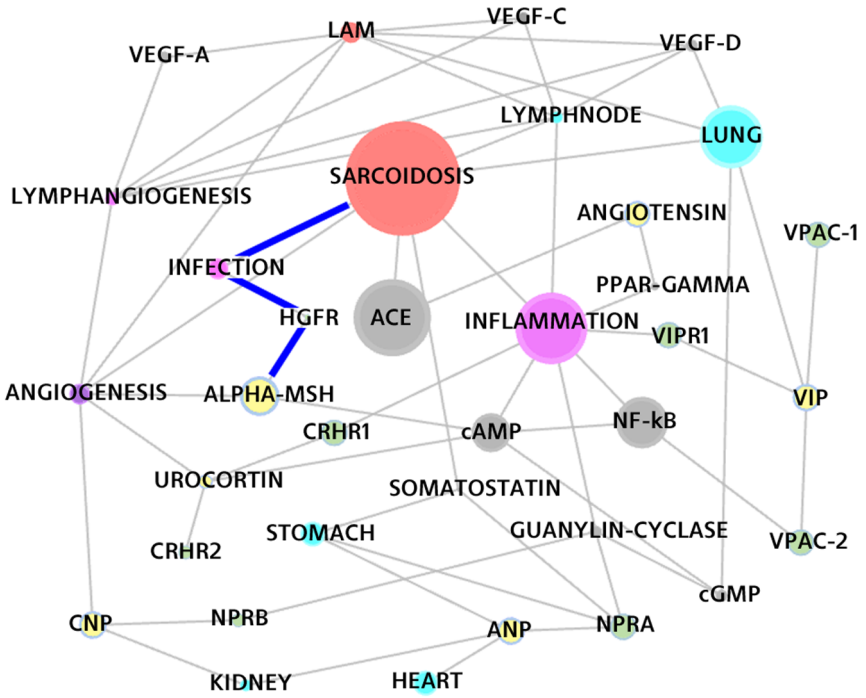


**Figure 5. The α-MSH – SARCOIDOSIS path and other closely related concepts.**
doi:10.1371/journal.pone.0084912.g005

acquired by exogenous infection and have a long incubation period followed by a clinical course of progressive dementia, myoclonal ataxia, delirious psychomotor excitement, and neuronal death; see [48].

Moreover, the system selects the path (see Fig. 7) that indicates the kinase c-Abl effect on cell-apoptosis as key MoA for redirecting Imatinib towards CJD. In fact, the c-Abl tyrosine kinase is found to be over-activated in neurodegenerative diseases like Alzheimer's disease and Parkinson's diseases, and overexpression of active c-Abl in adult mouse neurons results in neurodegeneration and neuroinflammation; see [49]. There is clear experimental evidence that activation of c-Abl leads to neuronal cell death and neuronal

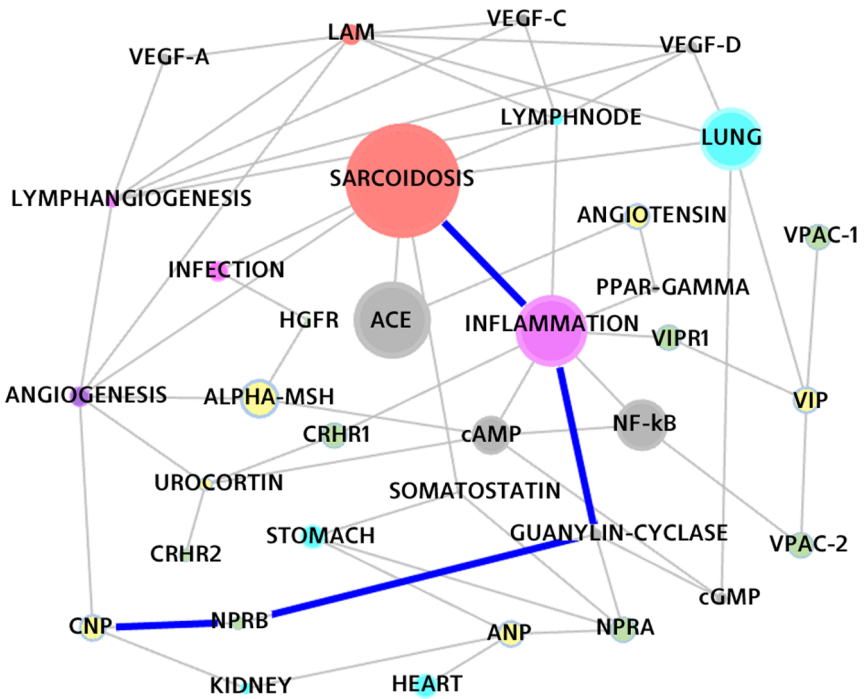


**Figure 6. The CNP – SARCOIDOSIS path and other closely related concepts.**
doi:10.1371/journal.pone.0084912.g006

apoptosis in experimental Creutzfeldt-Jakob disease; see [50]. Imatinib has been shown to prevent c-Abl kinase induced apoptosis in animal models of neurodegeneration; see [51]. Finally, Imatinib was shown to clear prion-infected cells in a time and dose-dependent manner from misfolded infectious protein without influencing the normal biological features of the healthy PrP, and Imatinib activated the lysosomal degradation of pre-existing misfolded PrP; see [52]. This provides a sound rationale for the proposed redirection.

The system indicated also Imatinib as a treatment option for pulmonary arterial hypertension (PAH), via its potent inhibitory effect on the PDGF Receptor (PDGF-R). For the indication PAH, the drug is however not approved.

## Conclusions

A double-layer methodology is presented, consisting of semantic analysis – leveraging on developments of Computational Linguistics – and graph analysis – exploiting Graph Theory and Stochastic Process Theory tools. This methodology has allowed the screening of more than 3 million abstracts from PubMed-published biomedical papers and the detection of relevant concepts identified by dictionary-defined expressions; concepts have been mapped as nodes of a graph, whose links are defined by co-occurrence of concepts across roughly 30 million of sentences. Specifically, the pathophysiological connections between peptides and diseases have been detected in order to provide inferences for biomedical rationales for drug repurposing.

The proposed methodology provides an effective instrument to detect different MoAs of peptides and drugs; though it may not capture the full-detail of the MoAs, it succeeds in making them recognizable by a short chain of biomedical entities. Moreover, the graph representations of biomedical knowledge seen above produces a sound and meaningful representation of the many interrelated concepts of the biomedical discipline; such methodology successfully allows both the validation of existing rationales and the discovery of new ones, a feat usually left to serendipity and intuition. We have translated the scientific rationales in relevant clinical trial settings into new potential treatment options for the affected patients in Sarcoidosis.

Our methodology confirmed the result of an open clinical phase II study, where we treated 20 patients with histologically proven Sarcoidosis and active disease with nebulized VIP for 4 weeks. This study is the first to show that VIP has clear, positive, immune-regulatory effects in sarcoid patients without any obvious side effects and without systemic immuno-suppression. VIP should therefore be developed as an attractive therapeutic option for patients with pulmonary Sarcoidosis; see [53]. We have initiated a clinical ex-vivo trial to prove α-MSH in a sarcoid pathology. Preliminary data clearly suggest a beneficial outcome of the experimentation (unpublished data), clearly suggesting α-MSH as another potential treatment option for this pathology.

Moreover, the case for Imatinib as a treatment option for the Creutzfeldt-Jakob disease shows how the system is able to produce a sound scientific rationale also for non-peptide drugs and with a mechanism of action quite different from the others, thus proving a much wider applicability.

Results are more noteworthy if the relative slimness of the dictionary is taken into account. Better representations are to be expected defining more detailed and more comprehensive dictionaries. Furthermore, Graph Theory tools provide quite an interesting arsenal of instruments to analyse a complex network of nodes (biological and medical concepts) and highlight hidden inferences across biochemical compounds, clinical data and medical concepts.

As it is apparent from this presentation the specific field of application enters the methodology in the broad selection of the document base and in the definition of the dictionary: the inner mechanism of knowledge representation and analysis is quite independent of it.

We would like to stress that here we have provided only very general characterization of the knowledge network and focused onto very well consolidated tools of analysis. But the field of complex networks is currently under massive development,

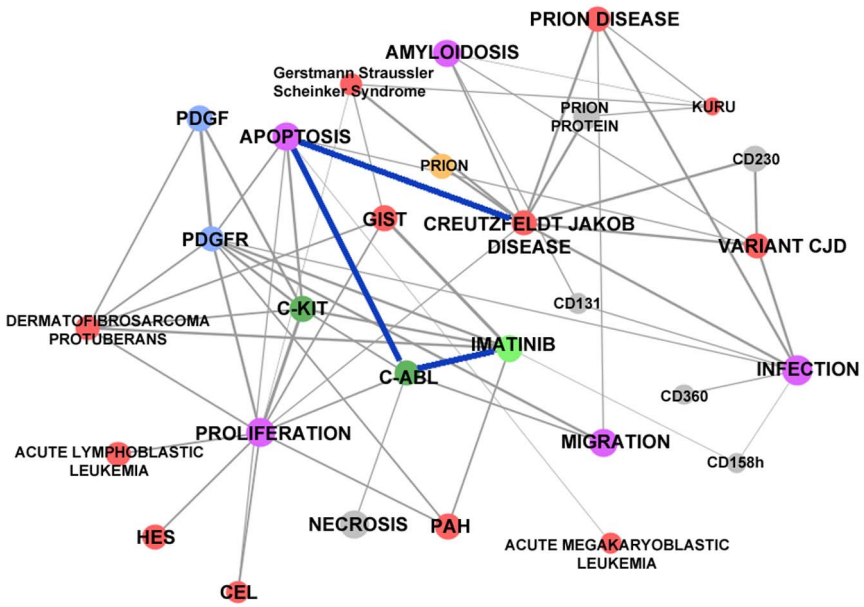


**Figure 7. Imatinib (GLEEVEC) – Creutzfeldt-Jakob Disease path and other closely related concepts.**
doi:10.1371/journal.pone.0084912.g007

providing ever more subtle indicators of graph features and related techniques of analysis. We therefore think that our reliance on the combination of the knowledge network inference with our random walk rankings poses this method in the best position to exploit this development and may well prove to make it mainstream in the field of text mining.

This methodology can be applied to other fields: for sure it can be extended over broader biomedical research, transcending peptides to study other chemical compounds and also focusing on diseases other than rare. We think it can be applied to any field of research – even outside natural science – provided that a suitable amount of literature is available and that the main issue be the association of a great number of particular facts and observation that do not yet fit into an already understood and comprehensive scheme.

## Supporting Information

**File S1 List of all concepts.**
(TXT)

**File S2 List of all paper IDs.**
(DATA)

## Acknowledgments

Tiziana Di Matteo and Ruggero Gramatica gratefully acknowledge support of this work by COST TD1210 project.

## Author Contributions

Conceived and designed the experiments: TA RG TDM SG MB DB. Performed the experiments: TA RG TDM SG MB DB. Analyzed the data: TA RG TDM SG MB DB. Contributed reagents/materials/analysis tools: TA RG TDM SG MB DB. Wrote the paper: TA RG TDM SG MB DB.

## References


1. Swanson D (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30: 7–18.
2. Swanson DR (1993) Intervening in the life cycles of scientific knowledge. Libr Trends 41(4): 606–631.
3. Swanson DR, Smalheiser NR (1997) An interactive system for finding complementary literatures. Artif Intell 91: 183–203.
4. Srinivasan P, Libbus B (2004) Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics 20 (suppl 1): i290–i296.
5. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74(2): 289–298.
6. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB (2007) Frontiers of biomedical text mining: current progress. Brief Bioinform 8(5): 358–375.
7. Rzhetsky A, Seringhaus M, Gerstein M (2008) Seeking a new biology through text mining. Cell 134(1): 9–13.
8. Erhardt RA, Schneider R, Blaschke C (2006) Status of text-mining techniques applied to biomedical text. Drug Discov Today 11(7): 315–325.
9. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A (2011) Literature mining, ontologies and information visualization for drug repurposing. Brief Bioinform 12(4): 357–368.
10. Wild DJ, Ding Y, Sheth AP, Harland L, Gifford EM, et al. (2012) Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research. Drug Discov Today 17(9): 469–474.
11. Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7(2): 119–129.
12. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R (2012) Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet 13(12):829–839.
13. Lesk AM (2005) Introduction to Bioinformatics. 3rd Ed. Oxford: Oxford University Press.
14. Cohen KB, Hunter L (2008). Getting started in text mining. PLOS Comput Biol 4(1): e20.
15. Džeroski S, Langley P, Todorovski L (2007) Computational discovery of scientific knowledge. Berlin: Springer Berlin Heidelberg.
16. Gruber T (1995) Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Int J Hum-Comput St 43(5–6): 907–928.
17. Maedche A, Staab S (2001) Ontology learning for the Semantic Web. IEEE Intell Syst 16(2): 72–79.
18. Klein M (2001) Combining and relating ontologies: an analysis of problems and solutions. Available: http://ceur-ws.org/Vol-47/ONTOL2-Proceedings.pdf Accessed 4 June 2012.
19. De Saussure F (2011) Course in General Linguistics. New York City: Columbia University Press.
20. Firth JR (1957) Papers in Linguistics 1934–1951. Oxford: Oxford University Press.
21. Caldarelli G (2007) Scale-Free Networks: Complex Webs in Nature and Technology. Oxford: Oxford University Press.
22. Newman M, Barabasi AL, Watts DJ (2011) The structure and dynamics of networks. Princeton: Princeton University Press.
23. Manning C, Schütze H (1999) Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press.
24. Sahlgren M (2008) The Distributional Hypothesis. Riv Linguist 20 (1):33–53.
25. Ananiadou S, McNaught J (2006). Text mining for biology and biomedicine. London: Artech House.
26. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform, 74(2): 289–298.
27. Yetisgen-Yildiz M, Pratt W (2006) Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 39(6): 600–611.
28. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, et al. (2010) Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLOS Comput Biol 6(9): e1000943.
29. Loging W, Rodriguez-Esteban R, Hill J, Freeman T, Miglietta J (2012) Cheminformatic/bioinformatic analysis of large corporate databases: Application to drug repurposing. Drug Discov Today Ther Strateg 8(3): 109–116.
30. Lekka E, Deftereos SN, Persidis A, Persidis A, Andronis C (2012) Literature analysis for systematic drug repurposing: a case study from Biovista. Drug Discov Today Ther Strateg 8(3): 103–108.
31. Blaschke C, Andrade MA, Ouzounis CA, Valencia A (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. Proc Int Conf Intell Syst Mol Biol 7: 60–67.
32. Hunter L, Lu Z, Firby J, Baumgartner WA, Johnson HL, et al. (2008) OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. BMC bioinformatics 9(1): 78.
33. Oda K, Kim JD, Ohta T, Okanohara D, Matsuzaki T, et al. (2008) New challenges for text mining: mapping between text and manually curated pathways. BMC bioinformatics (suppl 3): S5.
34. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K (2005) Beyond the clause: extraction of phosphorylation information from medline abstracts. Bioinformatics (suppl 1): i319–i327.
35. Yuan X, Hu ZZ, Wu HT, Torii M, Narayanaswamy M, et al. (2006) An online literature mining tool for protein phosphorylation. Bioinformatics 22(13): 1668–1669.
36. Saric J, Jensen LJ, Rojas I (2005) Large-scale extraction of gene regulation for model organisms in an ontological context. In Silico Biol 5(1): 21–32.
37. Jurafsky D, Martin JH (2002) Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Upper Saddle River: Pearson Prentice Hall.
38. Navigli R (2009) Word sense disambiguation: A survey. ACM Comput Surv 41(2): 10.
39. Zhou H (2003) Network Landscape from a Brownian particle's perspective. Phys. Rev. E 67:041908.
40. Müller-Quernheim J (1998) Sarcoidosis: immunopathogenetic concepts and their clinical application. Eur Respir J 12(3): 716–738.
41. Gonzalez-Rey E, Chorny A, Delgado M (2007) Regulation of immune tolerance by anti-inflammatory neuropeptides. Nature Rev Immunol 7:52–63.
42. Beuret L, Flori E, Denoyelle C, Bille K, Busca R, et al. (2007) Up-regulation of MET Expression by alpha-Melanocyte-stimulating Hormone and MITF Allows Hepatocyte Growth Factor to Protect Melanocytes and Melanoma Cells from Apoptosis. J Biol Chem 282(19): 14140–1414
43. Luo B, Wang Y, Wang XF, Gao Y, Huang BH, et al. (2006) Correlation of Epstein-Barr virus and its encoded proteins with Helicobacter pylori and expression of c-met and c-myc in gastric carcinoma. World J of Gastroenterol 12(12): 1842–1848.
44. Eishi Y, Suga M, Ishige I, Kobayashi D, Yamada T, et al. (2002) Quantitative Analysis of Mycobacterial and Propionibacterial DNA in Lymph Nodes of Japanese and European Patients with Sarcoidosis. J Clin Microbiol 40 (1): 198–204.
45. Nakayama T (2005) The genetic contribution of the natriuretic peptide system to cardiovascular diseases. Endocr J 52 (1): 11–21.
46. Buchdunger E, Zimmermann J, Mett H, Meyer T, Müller M, et al. (1996) Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. Cancer Res 56:100–104.
47. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, et al. (2000) Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase. Science 289: 1938–1942.

48. Marandi Y, Farahi N, Sadeghi A, Sadeghi-Hashjin G (2012) Prion diseases – current theories and potential therapies: a brief review. Folia Neuropathol 50(1): 46–49.
49. Schlatterer SD, Acker CM, Davies P (2011) c-Abl in Neurodegenerative Disease. J Mol Neurosci 45(3): 445–452.
50. Jesionek-Kupnicka D, Kordek R, Buczyñski J, Liberski PP (2001) Apoptosis in relation to neuronal loss in experimental Creutzfeldt-Jakob disease in mice. Acta Neurobiol Exp 61: 13–19.
51. Cancino GI, Toledo EM, Leal NR, Hernandez DE, Yévenes LF, et al. (2008) STI571 prevents apoptosis, tau phosphorylation and behavioural impairments induced by Alzheimer's β-amyloid deposits. Brain 131:2425–2442.
52. Ertmer A, Gilch S, Yun SW, Flechsig E, Klebl B, et al. (2004) The Tyrosine Kinase Inhibitor STI571 Induces Cellular Clearance of PrPSc in Prion-infected Cells. J Biol Chem 279(40):41918–41927.
53. Prasse A, Zissel G, Lützen N, Schupp J, Schmiedlin R, et al. (2010) Inhaled Vasoactive Intestinal Peptide Exerts Immunoregulatory Effects in Sarcoidosis. Am J Respir Critic Care Med 182: 540–548.
54. Jin W, Srihari RK, Wu X (2007) Mining concept associations for knowledge discovery through concept chain queries. In: Zhou ZH, Li H, Yang Q editors. Advances in Knowledge Discovery and Data Mining. Berlin: Springer Berlin Heidelberg. pp. 555–562.
55. Li Y, Jagdish CP (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics 26(9): 1219–1224.
56. Chen X, Ming-Xi L, Gui-Ying Y (2012) Drug-target interaction prediction by random walk on the heterogeneous network. Mol BioSyst 8(7): 1970–1978.
57. Goñi J, Avena-Koenigsberger A, Velez de Mendizabal N, van den Heuvel MP, Betzel RF, et al. (2013) Exploring the Morphospace of Communication Efficiency in Complex Networks. PLoS One 8(3):e58070.