Dana Yeo
Biomedin 231 Final Paper
Submission Date: 3/10/2014

# Drug Repurposing

**Motivation:**

The biopharmaceutical industry currently faces a major problem - production has not kept pace with the huge increase in R&D spending. Despite numerous new technologies which promised to reduce R&D costs (such as combinatorial chemistry, genomics, and structure-based drug design), the productivity problem has persisted and has forced pharma companies to become more creative. One such strategy that has been adopted has been trying to find new uses for, and better versions of, pre-existing drugs. This strategy is often referred to as drug repurposing, repositioning, or reprofiling.

Historically, attempts to minimize the amount of R&D which goes into a drug have been associated with increased risk. Drug repositioning, however, offers the unique possibility of avoiding many of the pitfalls associated with shrinking the R&D timeline. Development risk is significantly reduced in drug repurposing since repositioning candidates have usually already been through multiple stages of clinical development and have well-defined safety and kinetic profiles. Additionally, many hurdles such as chemical optimization, toxicology, formulation, etc have already been done and can easily be bypassed. These factors allow repurposed drugs to skip past the first several years of development substantially increasing its risk-versus-reward trade-off.

Having said this, drug repositioning is not without it's challenges. Oftentimes, drug repositioning required novel designs for clinical trials and the benefits of having well-defined safety and pharmacokinetic data of drug candidates are undermined by the lack of understanding of the underlying mechanism of action. Furthermore, the original toxicology or kinetic data may no longer meet modern regulatory standards. Finally, the process of beating out the competition in the realm of drug repurposing can be quite a bit more complex than is the case with *de novo* discovery. There are two cases to be considered when discussing the competitive aspect of repositioning - firstly, the case where composition-of-matter (COM) IP is held by another group and secondly, the case where the compound is off-patent and is now considered generic. In the first case, an agreement must be made with the party which holds the COM IP while in the latter case, the repositioner can just use the patent to provide substantial barriers to entry it the drug has never been marketed.

Given the disenchantment of many venture capitalists in regard to investing in *de novo* drug development, drug repurposing has grown increasingly more attractive to venture capitalists for firms of all sizes ranging from fresh startup to publicly-traded pharmaceutical powerhouses. Additionally, the party pushing drug repositioning is shifting from pharmaceutical companies (which have not traditionally organized which are conductive to repositioning) to biotechnology companies which possess the combination of incentives and

institutional flexibility [1].

**Emergence of Network-Based Inferences (NBI):**

The first workflow for selecting some combination of drugs to repurpose was presented in 2010. Protein interaction information was extracted using natural language processing (NLP) and imported into a relational database as a network of molecular interactions with each interaction having an annotation corresponding to a set of references. Such a model made it possible to use graph theory in the computational analysis of repurposing.

This initial workflow contained two parts. The first part was based on a thorough analysis of the literature. To find drugs which may be active against a particular disease pathway, the consensus disease pathway was manually constructed using information from reviews and collections of canonical pathways. A consensus disease pathway was then blasted against a database (ChemEffect) to search for chemicals which may inhibit multiple proteins along said pathway. The second component of the approach involves analysis of gene expression data and the use of a unique Sub-Network Enrichment Analysis (SNEA) found in Pathway Studio. Preliminary findings suggest SNEA can suggest novel therapeutics which are not described anywhere in the literature. These new threads of information can be used to further enrich the consensus disease pathway [2].

After the success of the initial protein repurposing strategy outlined above, effort was made to establish protein-protein networks on a larger scale. In 2010, an open-source protein network databases named PROMISCUOUS was published on the web. Initially, PROMISCUOUS contained greater than 25,000 drugs (including retired and experimental drugs) compiled from public resources via manual curation and data mining. At its core, PROMISCUOUS was the first exhaustive network-focused resource of protein-protein and protein-drug interactions enriched with side-effects and structural information which aimed to provide a comprehensive data set upon which basic graph theory methods could be applied. The network contained in PROMISCUOUS could be access one of two ways: firstly, by searching a specific drug or alternatively, by searching by a specific metabolic or signalling pathway. Drugs, targets, and side-effects in PROMISCUOUS are represented as nodes in a network with edges representing types of relationships between them [3]. Network accession of this type makes intuitive sense. To repurpose a drug, one might begin with either a specific drug in mind (due to availability, etc) or a specific disease-state in mind. In either case, the representation of the protein network described by PROMISCUOUS is not only highly descriptive but sufficient to begin making inferences and build hypotheses.

At this point resources such as PROMISCUOUS were broadly-based and while, descriptive, failed to articulate which facet of the network had the most predicting power. For example, is structural similarity more

powerful in selecting a viable drug repurposing candidate or network similarity? To try and tackle this problem, algorithms based specifically on different drug/target characteristics were developed and tested. For example, Keiser *et al* predicted new targets for known drugs using chemical two-dimensional structural similarity while receptor-based methods like reversing docking have also been applied. Unfortunately, some of these methods (in particular, receptor-based strategies) are not usable unless the both the drug and target's 3-dimensional structure are known.

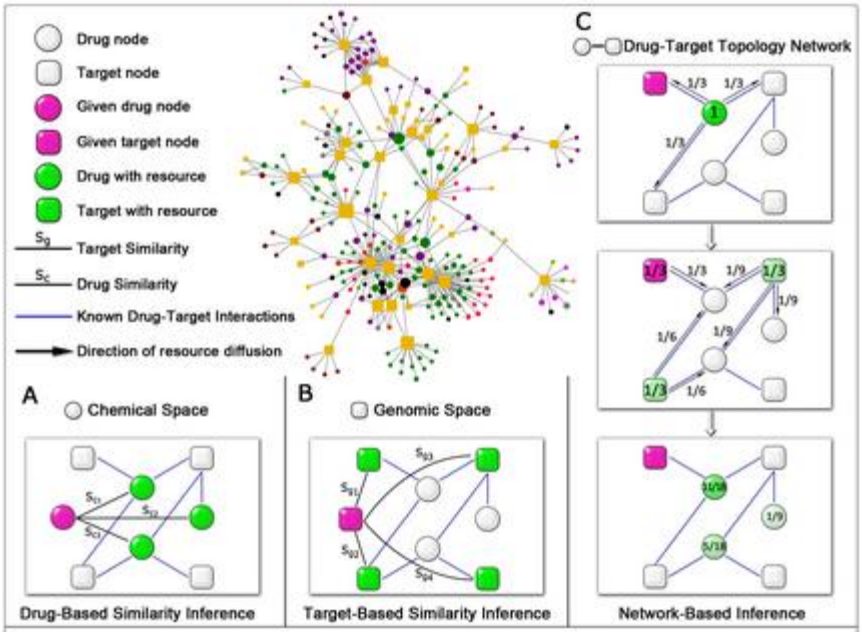


**Figure 1:** Graph schematics of DBSI, TBSI, and a drug-target topology Network

In an effort to glean which trait of a network is most descriptive a comparative study was done to discern the accuracy of different repurposing strategies. The three that were compared were drug-based similarity inference (DBSI) via drug-drug two dimensional structural similarity, target-based similarity inference (TBSI) via target-target genomic sequence similarity, and network-based inference (NBI) (**Figure 1**). In the study, NBI used only known drug-target bipartite network topology similarity to predict unknown drug-target inference (**Figure 2**). The process is reminiscent of mass diffusion in physics across the drug-target network.

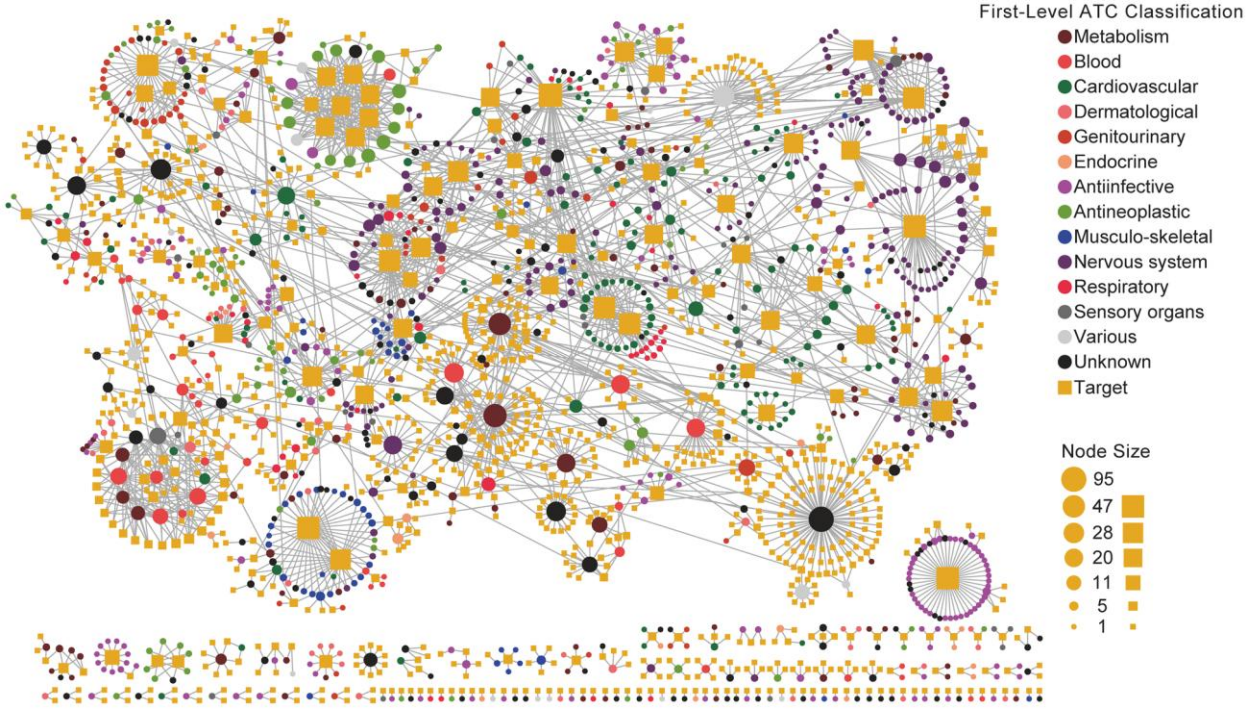


**Figure 2:** Bipartite drug-target Network

Using 4 well-defined data sets and running 30 simulations using 10-fold cross-validation, it was observed that all the methods had high true positive rates (TPR) and low false positive rates (FPR). Having said that NBI gave the best TPR value at any FPR value suggesting that of the three methods, NBI has the highest predictability. Notably, all methods outperformed other methods reported elsewhere. The essential difference between the three methods is the definition of similarity. In DBSI, similarity is defined by the chemical 2D structure of respective drugs whereas in considering TBSI, similarity is defined by genomic sequence similarity of the target. Finally in NBI, similarity is defined by drug-target network topology. Interestingly, NBI is the only method which fails to consider the structural similarity of the target or the drug and is, in some ways, the simplest method, using only drug-target interactions topology to infer new potential DTIs. In many ways NBI inherents the best of traditional drug discovery methods and chemical biology methods. It can prioritize drugs for a given target or prioritize candidate targets for a given drug based on recommendations from the user. Through matrix transposition, it is also possible to prioritize new potential targets for a given drug demonstrating it's promise as a powerful tool in drug repositioning. It is, however, important to note a major weakness in the NBI method. Since NBI only used drug-target interaction information, the method cannot predict targets for a new drug without known target information in the training set [4].

**The Cost-Benefit of Network Analysis:**

It is worthwhile to consider why networks are considered at all in regards to drug repositioning. Networks provide a description and a strong compromise between extreme reductionism and the paradigm of "knowledge is everything." In order for networks to be a suitable model in drug repositioning, we must consider the following factors: the definition of network nodes, edge and edge weights, the quality of data, and data refinement based on genetic variability, aging, and environmental factors.

Networks are often considered mathematical constructions or representations. Take, for example, graphs. When considering networks in drug-design is is important to be cautious of oversimplifying the model. Nodes in disease networks are not just points, but macromolecules containing within themselves sub-networks. To encapsulate the complexity of the sub-network structure of each point it is often useful to consider edge direction, sign (activation or inhibition), conditionality, and the number of dynamically changing quantities in each point [5].

**Foray into Tri-Partite Graphs and Network Analysis Algorithms:**

In the study previously listed, it was found that an NBI bipartite drug-target network outperformed NBSI and TBSI in statistical tests. Yet, as aforementioned the NBI network has many limitations. Firstly the NBI network from before was a bipartite graph. While there were a number of methods and measures which had been proposed as to how to find hidden connections within a bipartite network, they all suffer from the same problem - these algorithms are only applicable to monopartite networks with a single node type. A bipartite (or multipartite) network needed to be projected into a monopartite space. Unfortunately, network projection often results in loss of information, especially for low-degree nodes. In a paper published in 2012, Lee *et al* derived a new prediction method called the Shared Neighborhood Scoring Algorithm (SNS) to calculate the probability of a link existence between two nodes of interest by evaluating the connections of their neighbors.

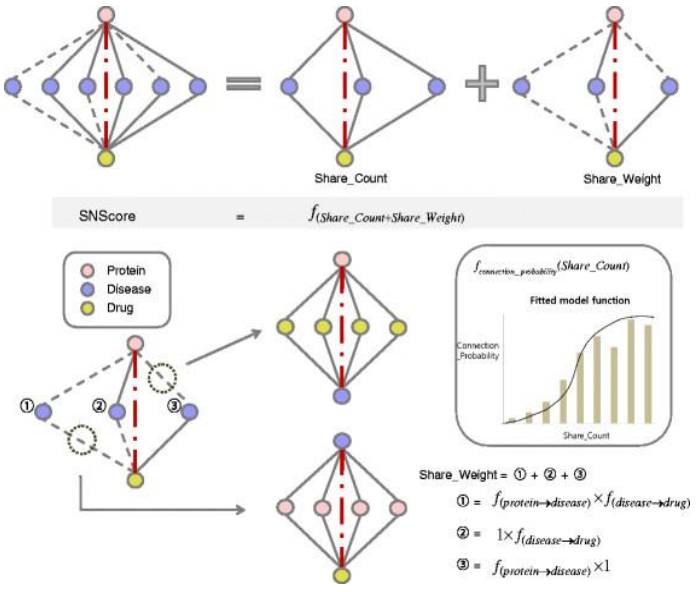


**Figure 3:** Schematic representation of SNS reduction for network analysis

To do this, Lee *et al* first established a database called "PharmaDB" (akin to a more well-defined PROMISCUOUS) which is a tripartite

pharmacological network database consisting of three types of nodes (up from a bipartite graph): human diseases, FDA-approved drugs and druggable chemicals, and proteins. The new type of node introduced in PharmaDB is the disease node. The proteins in the network contain therapeutic targets, disease-associated proteins, and drug-metabolizing proteins.

The basis of the SNS algorithm is as follows: the probability of connection between two nodes is directly proportional to the number of shared nodes between them,. Via this principle, the connection probability of two indirectly linked nodes can be computed, called the Shared Neighborhood Score. This score can bring attention to missing linkages and thus have significant repercussions in drug repositioning and repurposing. This is essentially applying the transitivity rule in graph theory to discover missing knowledge.

To calculate probabilistic connectedness via the SNS algorithm, one must consider both the "Shared Nodes Count" and "Shared Nodes Weight" of all the nodes between the two primary nodes of interest called A and B (Figure 3). "Shared Nodes Count" defines the number of in-between nodes connecting A and B and is directly correlated to the probability of connection between them. "Shared Nodes Weight" is the weight between 2 nodes as defined by a score of 1 for any pair of directly connect nodes, an indirect score for any pair of distantly connected nodes, or a virtual score for two completely disconnected nodes. The connection probability for A and B is the fraction of directly connected pairs among the total number of pairs for a given "Shared Nodes Count". The Share Neighborhood Score (SNS) is then the sum of the "Shared Nodes Count" and "Shared Nodes Weight", the product of each weight of links bridging A and B. For each relational dimension of the graph (i.e. drug-protein interactions, protein-disease interactions , and drug-disease interactions), the SNS can take a specific set of values which has been normalized based on the connecting probability function of the score neighborhood distribution [6].

**Applications of Sequencing and Genomics:**

Given the pervasiveness of genomics and sequencing, it comes as no surprise that harnessing the information reservoir of DNA would play a large role in redefining the networks and algorithms used in predicting new drug targets. While the assemblage of NBI strategies have allowed for the statistical prioritization of new drug targets with retrospective validation, all techniques derived from a network-based approach require prior knowledge of the disease, drugs, phenotypes, and regulation of pathways of interest. One approach which has taken advantage of modern sequencing has been the generation of genome-wide mRNA signatures on the theoretical basis that drug-cell signatures that anti-correlate with disease signatures can have therapeutic value. MRNA expression can be seen as lower-level function and by associating sets of mRNA data with disease, it is possible to derive novel therapeutic indications for approved drugs entirely from *in silico* studies and incorporate genome-wide complexity in the process. By modeling, in some way, the full complexity of the

system, this technique fundamentally differs from the combinatorial approach of reducing drug repositioning to a list of prioritized molecular mechanisms and, perhaps more importantly, does not try to understand nor require the understanding of the underlying molecular pathways. On the basis of mRNA sampling, gene expression classifiers have emerged to categorize patient samples and are trained to distinguish between phenotypic states and serve as archetypal emergent properties. Similar to mRNA microanalysis is the use of gene set enrichment analyses (GSEAS), which provide even more unbiased genome-wide analysis of gene expression. These two approaches suggest that intermediate genomic state have emergent properties unseen at higher and lower levels of biological complexity and have not been fully exploited. Undoubtedly, these metrics have much potential in drug repurposing [7].

To demonstrate the effectiveness of gene expression data in drug repurposing the landmark papers by Butte *et al* employed significance analysis of microarrays to obtain a signature of genes that were either significantly up or down regulated for a set of 100 diseases. After comparing every disease signature against every drug expression signature, Butte *et al* were able to create a drug-gene expression profile where each drug-disease pair was assigned a score based on profile similarity. A negative score implied that the drug had the potential for a therapeutic effect on the disease [8].

After applying their methodology, Butte *et al* found that topiramate, a drug used to treat epilepsy, showed a stronger score for treating Crohn's disease than the current medication. Topiramate was also expected to be one of the strongest predictors of inflammatory bowel disease (IBS). Rats with IBS treated with topiramate displayed less diarrhea, lower levels of inflammation, reduced incidences of ulceration, and hardier colon mucosal layers. The group also found a number of additional potential therapeutics for treating lung adenocarcinoma. Due to it's availability and strong side-effect profile, the authors tested cimetidine in lung adenocarcinoma cells. Cells treated with cimetidine showed reduced growth comparable to the leading chemotherapeutic. In essence, the use of microarray gene analysis by the others show that analysis of public gene expression databases can markedly contribute towards the effort of drug repurposing [8].

**Techniques using Machine Learning:**

Using gene expression data or drug-to-disease inference networks, while powerful, have significant limitations. Gene expression data can be extremely noisy (given that the publicly-available expression data often comes from patients being treated with multiple drugs simultaneously) and the variability in the complexity and sparsity of data avaliable for diseases necessary to build drug-to-disease inference networks impede the potential of these approaches. A viable alternative mRNA microarray analysis and inference

networks is the application of machine learning algorithms on data which completely lacks information on disease states and, instead, focuses exclusively on drug characteristics. One such machine-learning classification algorithm published in the journal of *Chemoinformatics* in 2013 works on an integration of many layers of information including similarities in chemical structures (via computing the distance between corresponding binary fingerprints), molecular targets (based on known common targets and their distances across the global human protein network), and post-treatment gene expression signatures. After consolidating and reducing these metrics into a single layer, Napolitano *et al* used these drug characteristics to train a multi-class Support Vector Machine (SVM) classifier. Mismatches between known and predicted drug classifications are intentionally interpreted as potential alternative therapeutic indications. The classifier reached extremely high accuracy levels (>78%) and produced reliable hints for drug repurposing [9].

It is important, however, to note that a classifier can only obtain knowledge over known samples and cannot predict new ones. Secondly, the backwards use of the classifier results where correct classifications are used to evaluate the quality of misclassifications indicates a new application of classifiers for the future [9].

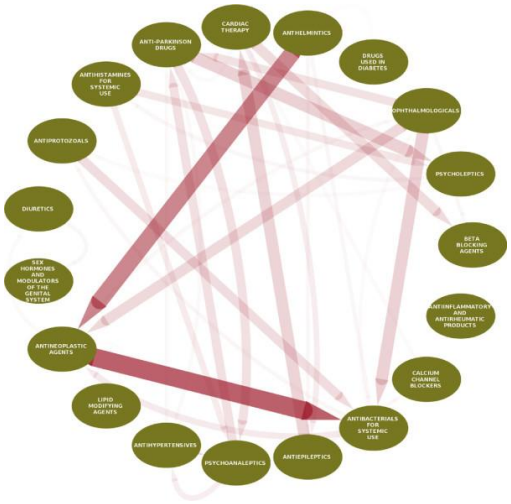


**Figure 4;** Graph representation of the repositioning of many drug groups with other groups. Arrow direction references the repositioning of one class to another. Thickness and color intensity reflect a score-weighted sum of reclassification events.

Another similar machine learning approach was published in 2013 which used a weighted k-nearest neighbor algorithm and a large margin method (to weigh and consolidate information from heterogeneous sources) to make predictions. The algorithms, named SLAMS, is rooted on the concept that similar drugs treat similar diseases. Given two drugs x and y, if x and y are found to be similar and drug x is used to treat disease z, then it follows that drug y is a drug repositioning candidate for z. As in the previous machine learning algorithm, the characteristics of the disease are not considered. However, unlike the previous method, SLAMS is extensive and can easily expand to n dimensions [10].

In this particular study SLAMs used metrics similar to the ones used by Napolitano *et al* - a drug pair's chemical structure (via a binary fingerprint which indicates the presence/absence of a structural motif), the target protein sets' sequence similarity based their Smith-Waterman sequence alignment score, and the pairwise side-effect similarity score using simple counting. Ultimately, the optimized model used k = 20 for

10-fold cross-validation. The results of pitting SLAMS against algorithms which rely heavily on averages or logical regression, were heavily in favor of SLAMS. SLAMS predictions overlapped with drug-disease associations currently being tested in clinical trials.

**Figure 5:** Schematic representation of extensibility of SLAMS

The primary selling point of SLAMS, again, is it's extensible nature. The algorithm allows for the easy integration of additional information sources and, in addition, ranks the value of each source based on their contribution to the prediction thus paving the way for even more reliable drug repositioning [10].

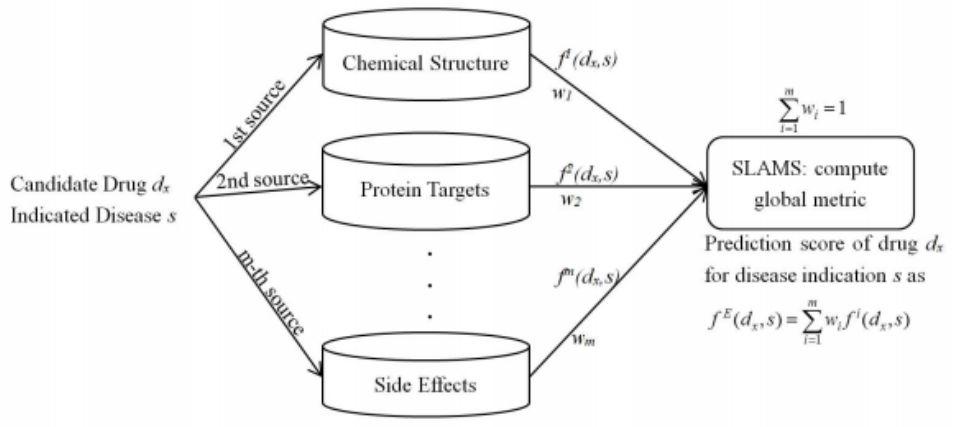


**Fig. 1.** Illustration of SLAMS Framework

**Future direction:**

Given the growing number of strategies and successes in drug repurposing (consider Pfizer's Viagra), a number of companies including Ore Pharmaceuticals, Biovista, SOM Biotech, Numedicus, Melior Discovery are all tackling drug repositioning on a systemic basis. Beyond the for-profit sphere, groups like *Cures within Reach* strive to repurpose generic drugs.

While it is difficult to say which drug repurposing method will ultimately prove to be the holy grail of repositioning, it is easy to appreciate the growing pervasiveness and importance of drug repurposing in the cost-benefit of analysis of pharmaceutical development and within the context of an exponentially growing biological knowledge-base.

**Bibliography:**

[1] http://www.nature.com/nrd/journal/v3/n8/full/nrd1468.html

[2] http://www.worldscientific.com/doi/abs/10.1142/S0219720010004732

[3] http://nar.oxfordjournals.org/content/39/suppl_1/D1060.short

[4] http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002503#pcbi-1002503-g001

[5] http://www.sciencedirect.com/science/article/pii/S0163725813000284

[6] http://www.biomedcentral.com/1752-0509/6/80/

[7] http://stm.sciencemag.org/content/3/96/96ps35.short

[8] http://www.nature.com/nrd/journal/v10/n10/full/nrd3565.html

[9] http://link.springer.com/article/10.1186%2F1758-2946-5-30

[10] http://astro.temple.edu/~tua87106/ecml13.pdf