Automated Hypothesis Generation Based on Mining Scientific Literature

Scott Spangler[*,1], Angela D. Wilkins[*,3], Benjamin J. Bachman[3], Meena Nagarajan[1], Tajhal Dayaram[3], Peter Haas[1], Sam Regenbogen[3], Curtis R. Pickering[2], Austin Comer[2], Jeffrey N. Myers[2], Ioana Stanoi[1], Linda Kato[1], Ana Lelescu[1], Jacques J. Labrie[1], Neha Parikh[3], Andreas Martin Lisewski[3], Lawrence Donehower[3], Ying Chen[1], Olivier Lichtarge[3]

[1]IBM Research
San Jose, California

[2]The University of Texas MD
Anderson Cancer Center
Houston, Texas

[3]Baylor College of Medicine
Houston, Texas

ABSTRACT

Keeping up with the ever-expanding flow of data and publications is untenable and poses a fundamental bottleneck to scientific progress. Current search technologies typically find many relevant documents, but they do not extract and organize the information content of these documents or suggest new scientific hypotheses based on this organized content. We present an initial case study on KnIT, a prototype system that mines the information contained in the scientific literature, represents it explicitly in a queriable network, and then further reasons upon these data to generate novel and experimentally testable hypotheses. KnIT combines entity detection with neighbor-text feature analysis and with graph-based diffusion of information to identify potential new properties of entities that are strongly implied by existing relationships. We discuss a successful application of our approach that mines the published literature to identify new protein kinases that phosphorylate the protein tumor suppressor p53. Retrospective analysis demonstrates the accuracy of this approach and ongoing laboratory experiments suggest that kinases identified by our system may indeed phosphorylate p53. These results establish proof of principle for automated hypothesis generation and discovery based on text mining of the scientific literature.

*These authors contributed equally

KDD '14, August 24 - 27 2014, New York, NY, USA
Copyright 2014 ACM 978-1-4503-2956-9/14/08…$15.00.
http://dx.doi.org/10.1145/2623330.2623667

Categories and Subject Descriptors

I.2.6 [**Learning**]: Concept Learning and Knowledge Acquisition

General Terms

Algorithms, Experimentation.

Keywords

Text Mining; Scientific Discovery; Hypothesis Generation.

1. INTRODUCTION

The pace of scientific publications is growing at an exponential rate [18] with over 50 million papers published so far [16], and over a million additional articles published annually [4]. This means on average a new article being published every 30 seconds. Within specific fields there may be tens of thousands of papers published every year—far more than any individual scientist can keep up with. In biomedical research, papers on specialized topics often run in the tens of thousands and topic areas contain orders of magnitude more. For example, over 70,000 papers have been published on a single protein, the tumor suppressor p53 [13]. Proteins are the fundamental machinery of the cell; understanding them is critical to advances in biology and medicine, and yet no scientist can possibly assimilate, recall and accurately process all of the known facts and relationships that could be relevant to discovering unknown protein functions, identifying relationships between proteins, or elucidating the role a particular protein may play in disease. Even recognizing new questions that should be asked can be a challenge. Instead, only a sliver of the relevant knowledge guides hypotheses: an approach that is deeply wasteful. This fundamental bottleneck is pervasive in biology and representative of every area of human endeavor in which there is a mushrooming mismatch between raw information and our analytic abilities.

Our goal is to accelerate scientific progress by combining mining, visualization, and analytics, with the hope to integrate all available content, identify the facts that are relevant to a given query, and from these facts suggest hypotheses that are new, interesting, testable and likely to be true.

Baylor College of Medicine and IBM Research have engaged in a long-term partnership to define scientific goals, build the necessary infrastructure, software, and algorithms and test the strengths and limitations of our discovery capabilities through direct experiments. We see this as a multi-year effort, but in the process of designing our approach and applying some very basic algorithms to the data, we made some early and surprising discoveries: even relatively simple approaches, when applied across enough data and with the benefit of expert domain knowledge, may quickly lead to significant discoveries in a complex domain. This paper describes our initial approach for this Knowledge Integration Toolkit (**KnIT**) for discovery and the early findings it has already led to so far.

KnIT embodies a three-phase process of Exploration, Interpretation, and Analysis. The Exploration phase surveys the relevant unstructured information, designs text queries, and extracts relevant documents for entities of interest. The entities of interest in this paper are a particular set of human proteins called kinases, detailed in Section 3. In general, each entity is modeled as a point in an abstract feature space, where the features of an entity correspond to its aggregate "text signature" in a corpus.

Next, the Interpretation phase builds a connected graph that represents the similarity relationship among entities. This helps domain experts visualize hidden connections between entities based on most of the known features and properties discussed in the literature. Coloring overlaid on top of this graph may reveal sub-graphs that correspond to clusters, typically with respect to some important property of interest. Critically, a sub-graph may contain a few "deviant" entities not known to possess the property of interest; because these entities are surrounded by others with a specific property, it is plausible to hypothesize that the deviant entities share it as well.

Finally, the Analysis phase globally diffuses annotation information among entities to rank order the best entity candidates for further experimentation of novel annotation predictions. In doing this we may compare the candidates derived analytically and rank-order the best candidates for further experimentation. The graph from the Interpretation phase can be used as a sanity check to ensure that the results correspond to common sense. Ultimately the domain expert can choose to verify only those annotation candidates that are the most analytically probable, experimentally testable, and of direct interest to the problem at hand, such as those suggesting a novel component of a disease's mechanism or a potential drug target.

Some past approaches have inferred formulas from experimental data [17]. Others have deduced direct connections based on a set of indirect associations that are obtained from highly structured, manual annotations of a corpus, such as MeSH annotations of MEDLINE data [30], hypothesis generation from unstructured text has been a hit-or-miss manual process [31] that is heavily dependent upon serendipity. Our approach leverages mining techniques for unstructured text to automatically discover hidden similarities between entities based on a corpus of scientific articles. The hope is that this approach will be robust and scalable even as entities and their multi-dimensional features create complex network relationships far beyond what human scientists can reason over, generating hypotheses that would otherwise elude domain experts.

In the rest of the paper we describe our proof of concept case study: a particular application of this methodology to a protein of singular importance across biology, the tumor suppressor p53. After discussing its role in cancer research, we describe our representation of the p53 literature and related proteins of interest. Next, we explain our knowledge visualization strategy for these proteins using a similarity graph, followed by a presentation of the analytical reasoning approach based on information diffusion. We then show in a series of retrospective validation studies that we are indeed able to create meaningful and accurate predictions. Finally, in a bona fide example of discovery, we present describe laboratory experiments that confirm several new p53 interactions predicted by KnIT, providing a proof of concept that will help direct future biological research.

2. IMPACT ON SOCIETY

In the spirit of this year's KDD emphasis on Social Good, we briefly place our work in a broader societal context. Humanity is facing a fundamental information bottleneck that overwhelms cognitive capabilities inherited from 160 million years of mammalian evolution. For most of this time, brains only had to cope with the perils posed by the physical world and by social competition. But as human language emerged, and with it the ability to pass complex abstract information across generations, a new evolutionary demand took form that was different from past constraints: now the recording, teaching and learning of facts and knowledge yielded advantages. Even so, as long as education remained oral, the body of knowledge was bounded by and matched to the cognitive abilities of each generation. Once written language emerged, however, a mere 4000 years ago, this constraint broke. Now facts and knowledge could accumulate *ex vivo* regardless of whether we could absorb them intellectually: by antiquity the library of Alexandria may have held as many as 100,000 scrolls, and by 2010 Google estimated that 130 million books were in existence. With the automated large-scale production of industrial, social, and scientific data upon us, the exponential divergence between information collections and human understanding is now as brutally vexing to all as it doubtless already was to the Alexandrian scholars.

Of course, humans overcome biological shortcomings by inventing tools, and already, many such computational aids exist. However, as empowering as they are compared to just a decade ago, their first-generation limitations are also inescapable. How often do we examine the fourth Google page, or even the third one for that matter? How deep into the literature do we read when a Medline search returns 200 papers, from the last 18 months? How much network traffic must we track to detect a lethal cancer cell, nascent electrical grid instability, or a terror target amidst anarchist chatter?

In this study, we tackle these questions. Starting from an immense corpus of knowledge, some in text and some in Big Data, we aim to extract relevant facts, represent them, and reason over them in order to generate new hypotheses that we can then test experimentally to arrive at a validated discovery. The tools and the computational framework that we develop for this purpose are entirely general, but in order to demonstrate that it may lead to immediate practical discoveries, we are focusing here on biology.

3. THE PROBLEM OF P53 KINASES

All human cells throughout an individual's body contain roughly the same genome, that is, the DNA molecules which represent the blueprints of biology inherited from one's parents. These blueprints contain the information necessary to create tens of thousands of different proteins, which are the molecular machines that are fundamental to all of cellular biology, performing a wide range actions such as metabolizing nutrients, allowing a cell to respond to its environment, and even controlling the quantities, or "expression levels", of other proteins. In this paper we investigate a particularly important protein, p53, which is often referred to as "the guardian of the genome" and is implicated in many biological processes and diseases including cancer [7; 9; 19; 22; 24; 25; 29]. As an individual grows and ages, cells must repeatedly make copies of their own genomes, which eventually results in degradation of the information contained therein. When enough errors accumulate, it is possible for a cell to enter a broken, cancerous state in which it grows continuously, damaging nearby tissue and causing harm to the organism. The p53 protein is a major player in the cell's natural defense against entering such a state: p53 responds to the detection of genomic problems by increasing the expression of hundreds of other proteins to try to fix the errors, or, if that isn't possible, it can even cause a cell to destroy itself, saving the neighboring cells and the life of the individual. One way that p53 is able to react to such problems is due to signals from a set of proteins that chemically modify p53 in response to different conditions. Each p53 modification, of which there are over 50, acts as an on/off switch, causing p53 to have one response or another [12]. The most common type of modification among all proteins, including p53 in particular, is phosphorylation, in which a phosphate molecule (PO_4^{3-}) is bonded to a specific atom in a protein molecule. The class of proteins that carries out the addition of phosphate molecules are known as kinases, which are increasingly the target of promising cancer treatments for use in these signaling mechanisms [27]. Drugs can affect the behavior of specific kinases, which can produce specific reactions in the proteins they phosphorylate, with the goal being to activate the cell's innate cancer-fighting abilities. Knowing which proteins are kinases is a well-solved problem [21]; however, knowing which proteins are modified by each kinase, and therefore which kinases would make good drug targets, is a difficult and unsolved problem. There are over 500 known human kinases and tens of thousands of possible proteins they can target. Biochemical experiments require months to establish a single novel kinase-protein relationship, and then years to fully elucidate the relationship's biological impact. Only 33 of the 500+ kinases are currently known to modify p53 [8; 12; 15; 23], but it is likely that there are many such relationships that remain unknown. In this paper, we asked as a proof of principle whether KnIT can discover novel p53 kinases.

4. REPRESENTING KINASES

To approach this problem, we first note that there are over 240,000 papers that mention one or more of 500+ known human kinases in their Medline abstract. An avid reader capable of absorbing 10 papers per day would need 70 years to go through this relevant literature—a completely unrealistic feat. Instead, however, we mine text so as to create a model for each kinase that represents all the terms present in the abstracts of the papers that specifically mention that kinase. In aggregate, and ignoring issues of errors and uncertainty, the words in these abstracts are assumed to be a useful signature of kinase features, such as details about biological process, molecular function, cellular component and specific interactions.

KnIT collects and labels the abstracts to be mined using queries against a text index of all Medline abstracts. There is one OR query for each kinase that includes the kinases canonical name along with its synonyms taken from [6; 11; 32]. KnIT submits the queries and downloads all abstracts that match each kinase up to query size. (A few kinases have well over 10,000 abstracts, which is far more than is needed to develop an accurate model.) In order to explore KnIT's predictive properties, we used several different Medline searches. In our initial exploration, we searched all kinases but removed abstracts that make any reference either to p53 or to a second kinase. We thus excluded data that would trivialize the predictions. This left us with 259 kinases in all. Of these, 23 were known to be p53 kinases.

Next we create a numeric representation that encapsulates all we know about each kinase relative to every other kinase. To facilitate this process we represent the documents in a vector space model. That is, each document is a vector of weighted frequencies of its features (words and phrases) [26]. We emphasize words with high frequency in a document, and normalize each document vector to have unit Euclidean norm.

The words and phrases that make up the document feature space are determined by counting the number of documents in which each word appears and identifying the words with the highest counts. A standard "stop word" list is used to eliminate words such as "and", "but", and "the". The top N words are retained in the first pass, where the value of N may vary depending on the length of the documents, the number of documents and the number of categories to be created. In our experiments we found that N=20000 is sufficient for the categories and documents used in this domain. After selecting the words in the first pass, we make a second pass to count the frequency of the phrases that occur using these words. A phrase is considered to be a sequence of two words occurring in order without intervening non-stop words. We again prune to keep only the N most frequent words and phrases. This becomes the feature space. A third pass through the data indexes the documents by their feature occurrences. We experimented with various methods of weighting term occurrences in this matrix and eventually determined that a Term Frequency–Inverse Document Frequency weighting (TF-IDF) [26] yielded the best overall prediction accuracy.

4.1 Kinase Space Visualization through Relative Neighborhood Graphs

Once we have a feature space we create a representation of each kinase by averaging the feature vectors of all documents that contain the kinase. This is the kinase centroid. Next we calculate a distance matrix that measures the distance between each kinase and every other kinase in the space. Such matrices are fine for computers to read and calculate properties over, but notoriously difficult for a domain expert to interpret in order to get a sense of the data's underlying validity and meaning. Interpretability is important, because an expert must be confident and insightful when proposing new hypotheses. Thus some way must be found to convert the numbers into a meaningful picture of kinase-kinase relationships.

A network graph is one approach that is often tried [10] but this requires determining when two nodes should be considered connected, when in fact all nodes are connected at some level of similarity threshold. We could pick an arbitrary similarity cutoff and draw lines whenever similarity was greater than the specified limit, but that assigns more meaning to the absolute value of the distance metric employed here than is strictly warranted. Relative distance is the more important concept to convey here.

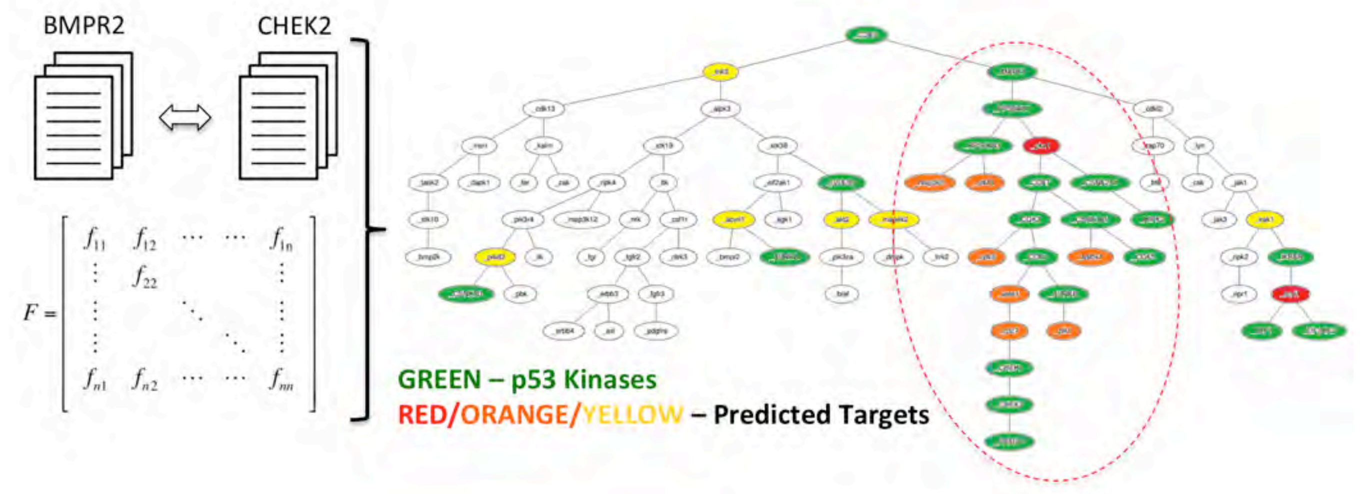
BMPR2
CHEK2
$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & \cdots & f_{1n} \\ \vdots & f_{22} & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & \cdots & f_{nn} \end{bmatrix}$
GREEN – p53 Kinases
RED/ORANGE/YELLOW – Predicted Targets

Figure 1 Kinases are clustered based on their literature distance. The clustered p53 kinases (green) suggest new kinases that may also phosphorylate p53.

To design the graph used in KnIT, we switch the goal around. What does a maximally communicative graph look like? First of all it should be minimally connected, in other words containing one less arc than the number of nodes. Secondly it should be a tree because trees are easy to navigate and communicate information based on distance from the root, which is often helpful. Third, the tree should spread connections out fairly evenly among the nodes to avoid extreme situations where one node is connected to all the others, a very uninteresting graph. This leads us to the conclusion that a binary tree (or at least low n-ary) would be highly advantageous if it can be drawn so as to accurately represent the distance matrix. To create a binary tree we must choose a meaningful root node. Does any particular entity stand out for this honor? In fact, there is one property unique and important in the text vector space—namely typicality. There is one entity whose vector is closest to the average of all the vectors. This will be the root. Now as we move down the tree we will naturally go towards less typical (more unusual nodes). This turns out to be a very intuitive concept to grasp, visually.

Algorithm 1 is used to create an n-ary similarity tree from the set of entities, where each entity is represented as a feature vector. The root of the tree is the "most typical" entity, and typicality decreases with increasing distance from the root, so that the leaves of the tree are the "least typical", i.e., unique outliers. The algorithm first computes the "most typical" feature vector as the average of the entities, and then calls the closestToFV function to select the entity closest to this typical feature vector as the root of the similarity tree. The algorithm also initializes *candidates* as a singleton set comprising the root node; here *candidates* are the set of nodes currently in the tree at which to potentially attach new child nodes. The algorithm next uses the closestPair function to find the pair (e,c) such that e belongs to *entities*, c belongs to *candidates*, and distance(e,c) is minimized over all such pairs. (Thus e is the closest entity to the current set of *candidates* and c is the candidate closest to e.) Then e is added to the tree as a child of c. Moreover, e is removed from *entities*, the set of entities that have not yet been added to the tree, and added to *candidates*. If adding e as a child to c increases the number of c's children to the limit n, then c is removed from *candidates* to ensure that its n-ary property will not be violated in the future. The algorithm continues to add elements of entities to the tree in a similar manner, until there are no more entities to add. The functions closestToFV and closestPair use Euclidean distance and break ties randomly.

In Figure 1, we show an example of this kinase network diagram with n=2. Green nodes are p53 kinases, red/orange/yellow nodes are hypothesized new p53 kinases based on their similarity to known p53 kinases. What is remarkable about this visual representation is that the green "kinase" nodes tend to be clumped together, even though the algorithm knows nothing about p53 kinases. This tends to lend credence to the supposition that those nodes in the midst of the green clumps are also likely to be p53 kinases.

Algorithm 1 *Create an n-ary similarity tree from a set of entities*

```
Input: entities, n
Output: n-ary similarity tree
mostTypicalFV = average(entities)
root = closestTo FV(entities, mostTypical FV)
entities.remove(root)
candidates = {root}
while not entities.isEmpty()
    (e, c) = closestPair(entities, candidates)
    c.addChild(e)
    if c.numChildren() == n then
        candidates.remove(c)
    end if
    candidates.add(e)
    entities.remove(e)
end while
Return: root
```

5. SELECTING CANDIDATE P53 KINASES

The visualization tool provides a set of kinases that may phosphorylate p53. However, some sort of principled ranking

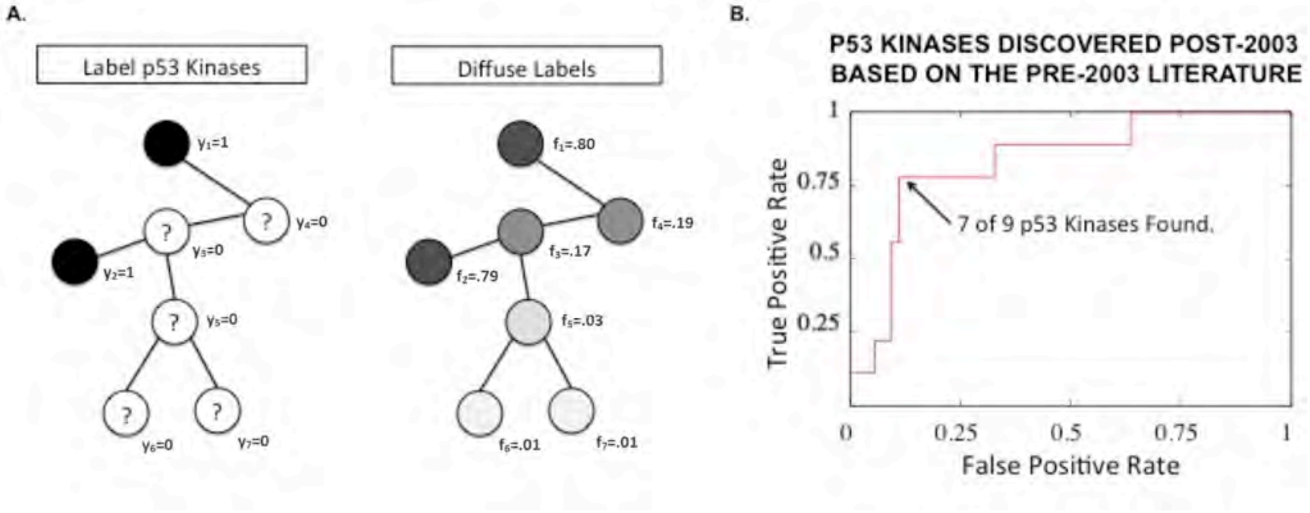

Figure 2 A) Example of graph diffusion on a network. Given a set of edges and labeled nodes, information is diffused to find additional candidate nodes for the annotation in question. B) Retrospective validation of literature vector models with graph diffusion to predict p53 kinases. Using only papers from before 2003, predictions were made for 64 other kinases, 9 of which are now known to be true positives but were not known in 2003.

scheme is needed in order to prioritize the kinases for further experimentation. To provide such a scheme, our initial prototype uses graph diffusion [34]. Graph diffusion is a semi-supervised learning approach for classification based on labeled and unlabeled data. It takes known information (initial labels) and then constrains the new labels to be smooth in respect to a defined structure (e.g. a network). In our case, we know which kinases phosphorylate p53 (initial labels); we would like to know which other proteins phosphorylate p53 (final labels). The distance matrix based on the literature gives us the structure of our kinase network. The initial labels are extracted from current knowledge found in review articles [8; 12; 15; 23].

5.1 Graph Construction and Diffusion

Graph diffusion propagates information among network nodes following the edges between them (Figure 2A). Here, human kinases are the nodes and the distance matrix of literature similarity between each kinase provides the edges. To formulate the kinase network, we defined edges between each kinase and the top ten most closely related kinases. This cutoff was determined empirically by cross-validation performance.

We can represent our knowledge of protein function as y, a fixed binary vector of labels with y_i representing whether protein i phosphorylates p53. We seek to identify a new set of continuous labels, f (*i.e.* how likely a kinase is to phosphorylate p53) by diffusing the known information in a network. We can solve for f by minimizing the sum of the loss and smoothing functions [3]:

$$(f-y)^T(f-y)+\mu f^T Lf$$

The first term, the loss function, represents the difference between initial y and final labels f. During diffusion, this function regulates and prevents the loss of the initial labels. The second term, the smoothing function, represents the smoothness of the new labels f in the context of the Laplacian matrix L. The Laplacian matrix [5] is the matrix representation of the kinase network and defined $L = D - A$. The adjacency matrix, denoted A, specifies if kinase i is connected to kinase j where $A(i,j)=1$ if the entities are connected and $A(i,j)=0$ otherwise. The degree matrix, D, is a diagonal matrix given by

$$D_{ii} = \sum_j A(i,j)$$

The diffusion coefficient μ balances the loss of the initial labels against the smoothness. The previous equation has a closed form solution [3]:

$$f = (I+\mu L)^{-1} y$$

where I is the Identity matrix. We set the diffusion coefficient μ to the inverse of the Laplacian's norm

$$\mu = \frac{1}{\max_i(\sum_{k=1}^{n} |L_{ik}|)}$$

This value insures that the Hessian is positive definite and the above function is convex [20]. In order to identify new kinases, we then look at the new labels f where the labels with the largest increase will be our targets.

5.2 Computational Validation of Performance

Leave-one-out cross-validation eliminates an observation from the original data set, and then tests whether predictions based on the remaining information can recover the observation that was removed. Each one of the known p53 kinases was relabeled as an unknown and the performance of the prediction method was evaluated by whether it could recover the information. Due to the limited availability of negative information in biology, those with no known p53 activity were treated as negatives for the purposes of computational validation.

We often used Receiver Operating Characteristic (ROC) curves to measure performance. In more detail, the diffusion scores in *f*, together with a threshold *t*, can be used to classify each protein as a “positive” (i.e., it phosphorylates p53) or a “negative” (no phosphorylation). Protein *i* is predicted to be positive if $f_i > t$ and a negative otherwise. Recall that the True Positive Rate (TPR) is the fraction of the True Positives found over the total possible Positives. Similarly, the False Positive Rate (FPR) is a fraction of the False Positives found over the total possible Negatives. Thus in our setting the ROC is the cumulative plot constructed by calculating TPR and FPR at all possible binary classification thresholds *t*. The area under this curve (AUC) summarizes prediction performance: an AUC of 1.0 represents a perfect predictor while a random prediction would lead a diagonal line and an AUC of 0.5.

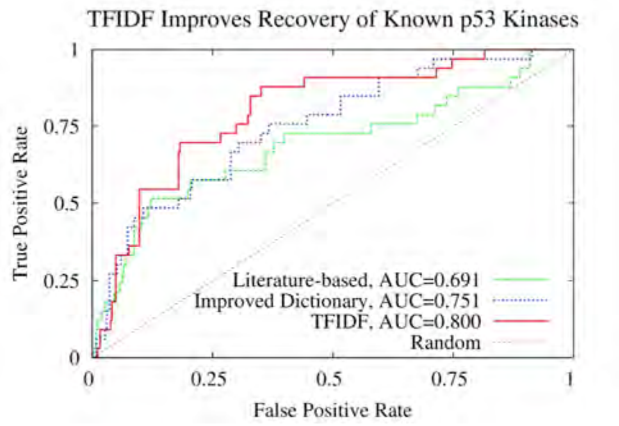

Figure 3 Graph diffusion was performed with leave-one-out-cross validation to assign scores to 33 kinases known to target p53 as well as 368 kinases not known to target p53. Of the 509 kinases included in analysis, we only included the 401 kinases reaching a minimum document count threshold. (Black) represents the average performance of a random algorithm with no predictive value. (Green) is the original algorithm’s performance. (Blue) shows the improvement due to dictionary synonym refinement. (Red) is with the inclusion of the TFIDF weighting scheme.

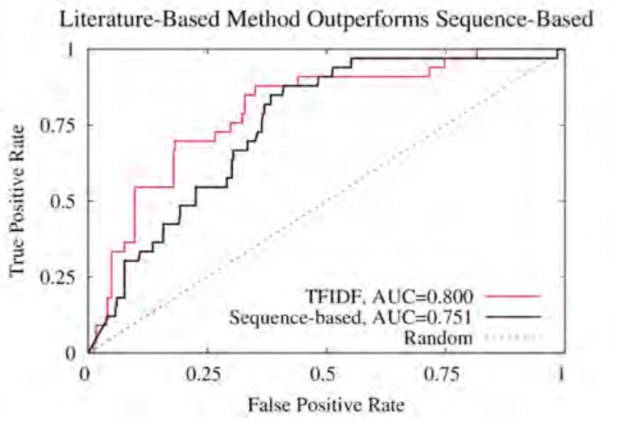

Figure 4 Comparison to sequence-based method. (Red) is predictive power of the literature vector model in a leave-one-out cross-validation analysis of known p53 kinases. (Black) is the performance of a method that uses biological sequence information to do the same task. (Black-dotted) represents the performance of a random model.

6. RESULTS

In order to test the algorithm, we first applied it in a retrospective analysis to show whether recent annotations of new p53 kinases occurring *after* a certain date could be predicted from a model that only took into account papers written *before* that date, at a time when these discoveries of p53 kinases were still unknown. Next we asked whether some variations in the algorithm could improve p53 kinase prediction as we compared its performance to the common approach used most typically to identify functionally similar proteins in biology. Finally, we expanded our analysis to a larger set of proteins to test scalability. The positive results in these retrospective controls led to a prospective study of *bona fide* predictions, discussed in section 6.

6.1 Retrospective study

If KnIT is effective, it should predict from papers published prior to a given date events that were only discovered after that date. To test this hypothesis, we mined the literature up to 2003, when only half of the 33 currently known p53 kinases had been discovered. Because we filtered out confusing abstracts that mentioned multiple different kinases and p53, the kinase search space became small: only 74 kinases. But among these 74, ten were known to phosphorylate p53 in 2003, nine were found at a later date, and the remaining 55 are for simplicity assumed to not phosphorylate p53. A kinase distance matrix and a literature vector model was developed for these 74 kinases, the ten p53 kinases that were known prior to 2003 were labeled as such, and these labels were propagated to the other 64 kinases by global graph diffusion from which we could now rank the 64 kinases by the likelihood they targeted p53. Strikingly, an ROC curve shows that seven of the nine true positives are readily predicted with this algorithm (Precision= 0.54 at Recall=0.77, with an AUC of 0.840); see Figure 2B. This time-stamped study shows that back in 2003, we could have automatically predicted many of the p53 kinases that were discovered in the subsequent decade by combining text mining with feature analysis and graph-based diffusion as KnIT does. This result is remarkable considering that for simplicity we only used a limited subset of the least ambiguous abstracts, which restricted us to studying only 74 rather than about 500 kinases.

6.2 Further Analyses of Algorithm

In order to test KnIT on a larger scale, we next sought to include more abstracts, regardless of their date or mention of multiple kinases and p53 information. This time to mitigate the most obvious problems related to incomplete or noisy information, we only filtered out those kinases having fewer than 20 abstracts. This allowed us to model 401 kinases, rather than the previous 74 kinases. Instead of a time-stamped retrospective analysis, we performed leave-one-out cross-validation of all 33 currently known p53 kinases. Specifically, a literature vector model and then a kinase distance matrix were built and we then performed 33 separate experiments. In each, one of the 33 p53 kinases was relabeled “unknown” rather than “p53 kinase,” then the remaining 32 p53 kinases were diffused globally to yield a ranking for the “unknown” p53 kinase. We then diffused the label information of all 33 kinases to obtain scores for all non-p53 kinase (401-33=368). Once each kinase has a label score, we can then calculate the ROC curve treating the known p53 kinases as positives. The area under ROC curve was shown to be significant at 0.691 (Figure 3, green curve).

We first constructed a list of gene synonyms extracted form from NCBI[2; 33], UniProt KnowledgeBase[32], and Hugo Gene Nomenclature Committee [11]. We noted that some kinases have ambiguous synonyms that are likely to return more incorrect abstracts than correct ones when querying Medline:

BTK has "AT" as a synonym

TYRO3 has "TIF" as a synonym

ARUKA has "AURA" as a synonym

RIPK1 has "RIP" as a synonym (confused with "repeat induced point mutation")

ITK has "EMT" as a synonym (confused with "epithelial-mesenchymal transition")

MOK has "RAGE" as a synonym

BMP2K has "BIKE" as a synonym

KIT has a problematic main gene name

FYN has “SYN” as a synonym (common acronym for synthetic)

To improve on results, we removed these vague terms. A second issue we considered was that many of the kinases shared synonyms names (i.e. PAK1 is a synonym for PKN1, but PAK1 is a distinctly different kinase). We removed redundant gene names within the protein kinase, unless the gene name was the Hugo Approved Gene Name (As is the case of PAK1). Removing these ambiguous synonyms yields a more accurate representation of each kinase’s content. After removing problematic synonyms, we resubmitted our queries and performed the analysis again with the new set of abstracts for each kinase. This improved kinase dictionary now increased the AU(ROC) from 0.691 to 0.751 (Figure 3, blue curve).

Next, we experimented with different weighting schemes for term frequencies when creating the distance matrix. We determined that a TFIDF [26] weighting, which gives a stronger weight to shared words that occur less, dramatically improves the AUC=0.800 (Figure 3, red curve). In contrast, increasing the max query size from 1600 to 3200 abstracts had virtually no impact on performance (data not shown). We also compared KnIT to a common approach for predicting biological function: protein sequence alignment. Every protein exists as a sequence of amino acid molecules, which can be represented by a series of letters using a 20-character alphabet, one letter for each type of building block in the sequence. It is well known that similar sequence often, though not always, suggests similar function of two proteins. The amino acid sequence of each kinase was aligned with the set of positives, and those with low e-value as reported by BLAST [1] were predicted to be more likely to target p53. (Roughly speaking, the BLAST e-value indicates the probability that sequences align purely by chance.) A comparison of the leave-one-out performance of these methods is shown in Figure 4. These leave-one-out control studies show over a nearly fully representative set of kinases that that KnIT is superior to a sequence-based approach with respect recovering knowledge of a true p53 kinase when this knowledge is erased. Although leave-one-out studies should not be over-interpreted, this positive result is consistent with the possibility that KnIT is predictive.

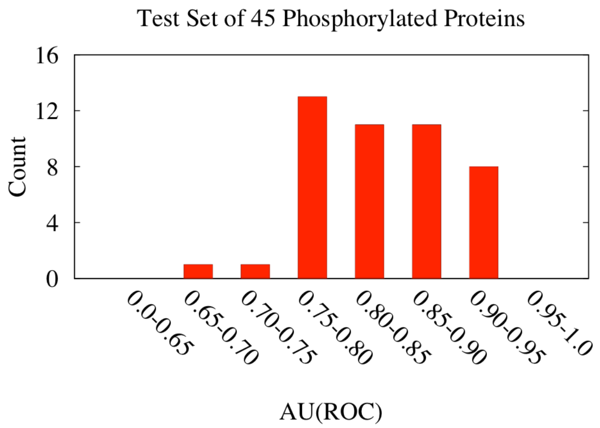

Figure 5 Summary of 45 different AU(ROC) performance measures for predictions on each of 45 different kinase targets, each with a unique set of labels for the search space of 500+ known kinases.

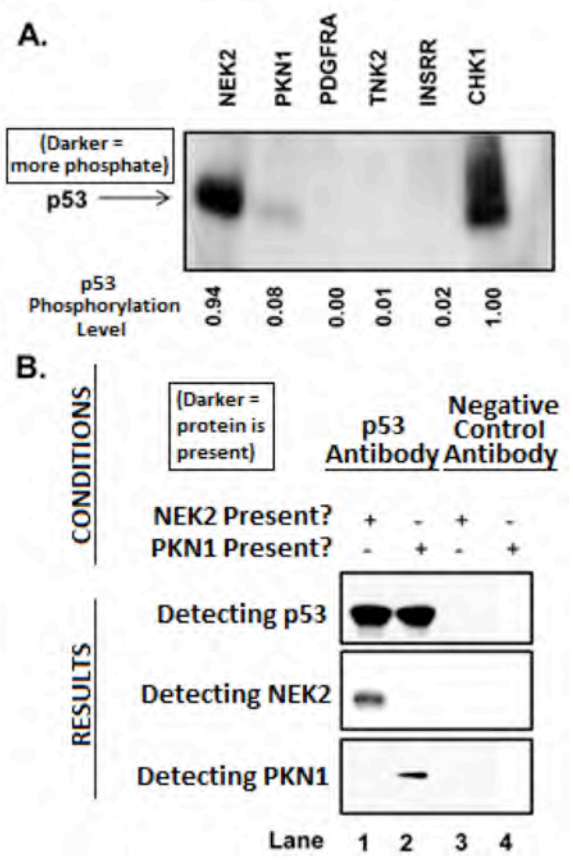

Figure 6 Experimental validation of candidate p53 kinases as *bona fide* p53 kinases. (A) *In vitro* kinase assay demonstrates phosphorylation of p53 by top ranked candidate kinases PKN1 and NEK2. Relative levels of p53 phosphorylation are indicated for each kinase normalized to positive control CHK1. Though the signal is weak for PNK1, subsequent experiments lend further support. (B) PKN1 and NEK2 shown to interact with p53 *in vivo.* A p53 antibody isolates p53 and any proteins bound to it. Antibodies detect the presence of candidate kinases in this isolate.

6.3 Large-scale Study

As a further test of scalability we next asked if KnIT could also predict kinase activity for targeted proteins other than p53. In previous experiment we exploited review articles from p53 experts. In order to extend to the larger scale analysis, we exploited the kinase-to-target knowledge base built for human kinases by PhosphoSitePlus [14]. These data are manually curated for the modifying protein, the type of modification, and the subject of the modification. So far, the curators of PhosphoSitePlus have reliably identified over 14,000 relationships between kinases and target proteins by reading the available literature. Among these, we narrowed this study to 45 unique human proteins, each modified by at least eight unique kinases belonging to our kinase network. It is likely that some sets of kinases may be easier to predict than others, for example if some have simpler biological mechanisms, therefore rather than normalizing and plotting all predictions onto a single ROC curve, each protein was evaluated independently, and the AUC was calculated for each one separately. Figure 5 shows a histogram summarizing the performance of these 45 ROC curves. We found that in all cases that the predictions were significant (AU(ROC)>0.65) and the average AUC was 0.835. These data show that KnIT does not uniquely apply to p53, but that it can extend to make predictions in many other human proteins.

6.4 Experimental Validation

Retrospective analyses are suggestive but never proof of discovery. For the latter it is critical to predict an observation that has never been made and then assess its truth in the laboratory. The algorithms described above were used to rank 252 kinases (those with at least 20 publications) by likelihood of being p53 kinases. As expected, most kinases known to phosphorylate p53 were near the top of the rankings list. Five kinases on the list not known to phosphorylate p53 were then tested by biochemical and molecular biology experiments for their ability to interact with and phosphorylate p53. Two kinases (PKN1 and NEK2) were chosen from near the top of the list and three kinases were chosen from the bottom half of the list (TNK2, INSRR, and PDGFRA). Two validation assays are shown here: an *in vitro* kinase assay and a cellular co-immunoprecipitation (IP) assay.

In Figure 6A, the *in vitro* kinase assay, p53 is combined with a kinase and a radioactive source of phosphate, gamma-^{32}P-ATP. Then a technique known as electrophoresis is used to separate the components of the mixture by weight to different positions in a gel. Because the weight of p53 is known (53 kilodaltons), we can check for radioactivity of anything that weights exactly that amount. If the predicted relationship is real, the kinase will add the radioactive phosphate to p53, and electrophoresis will move that radioactivity to a specific position in the gel, which is detectable by standard instruments. Note that in Figure 6A top ranked kinase candidates NEK2 and PKN1 exhibit a labeled p53 band, as does a known p53 kinase CHK1 [28] (used here as a positive control). The PKN1 band is faint relative to the others, but in subsequent experiments, the interaction was found to be robust. In contrast, low ranked p53 kinase candidates PDGFRA, TNK2, and INSRR exhibit no 53 kilodalton band, indicating they are unlikely to be p53 kinases.

Figure 6B shows a different approach for validating the predictions. The goal of this assay is to show that there is a physical protein-protein interaction between p53 and the predicted kinases. This is considered a strong indication that the kinase is likely to target p53. Human cells containing p53 and a candidate kinase are generated and analyzed. Proteins from the cell are isolated and a p53-specific antibody is then added; an antibody is a substance that will bind to and isolate a specific protein, in this case p53, along with any protein that is bound to it, in this case - if the prediction is correct - the kinase being tested. This isolate is then separated by size, and an additional antibody is used to test for the presence of each candidate kinase. In Figure 6B, each column, or "lane", represents a different set of experimental conditions. Lane 1 shows that p53 was indeed bound to NEK2, with lanes 2 and 3 as controls that show both NEK2 and the p53 antibody must be present to achieve this result. Lane 2 shows that p53 was bound to PKN1.

These two sets of experiments argue strongly that computationally predicted top p53 kinase candidates PKN1 and NEK1 are indeed true p53 kinases.

7. CONCLUSION AND FUTURE WORK

This early study tackles a basic problem that is challenging progress in every field of human intellectual activity: we have become much better at generation of information than at its integrative analysis. This leads to deep inefficiencies in translating research into progress for humanity. No scientist can keep up with the unrelenting flow of new studies and results, even within specialized fields. While intuition and selective reading in a highly narrowed field of work are essential and can certainly lead to breakthroughs, they are also likely to lead most scientists at one time or another towards deeply unproductive hypotheses that might have benefitted from more comprehensive insights into the data that was available but which we could not find an opportunity to learn about. Specialization also inherently limits the opportunities to find common grounds at the interface between fields, even though these interfaces are often are among the deepest areas of scientific synthesis.

Baylor College of Medicine and IBM Research have joined forces for the purpose of combining talents and technologies in many diverse fields to accelerate scientific discoveries that lead to improved patient outcomes. This research represents the first stage in this collaborative effort and as such it proves the principle that mining past literature is a viable strategy for predicting previously unknown biological events. We have shown that p53 kinases predicted with our text mining methods are supported by laboratory findings. In the future, it should be possible to make many other kinds of predictions on a much larger scale as our infrastructure and capabilities ramp up.

In the future our team will focus on a wider area of proteins and functions, building up comprehensive networks of interactions and predicting where new connections ought to exist based on everything else that is known. We believe this will ultimately accelerate the pace of cancer discoveries by an order of magnitude and allow scientists to come to a much more complete understanding of the mechanisms behind this disease. We also feel that the general approach of mining literature to identify hidden relationships between entities is not confined to cancer or even to biology, but is a general tool that can be applied in almost any science. The potential for dramatic acceleration of discovery in all sciences holds out the possibility of tremendous benefits for human health and for societal progress in the coming years. Given the enormous challenges facing science today on a global scale, with ever more complex systems of entities and networks of relationships, the acceleration of discovery is not only desirable, but also indispensable for human flourishing.

8. ACKNOWLEDGMENTS

This work was supported by the Robert and Janice McNair Foundation, DARPA (N66001-14-1-4027), National Science Foundation (NSF CCF-0905536, NSF DBI-1062455), National Institutes of Health (NIH-GM079656), and was supported in part by the IBM Accelerated Discovery Lab.

9. REFERENCES

[1] ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., and LIPMAN, D.J., 1990. Basic local alignment search tool. *J Mol Biol 215*, 3 (Oct 5), 403-410. DOI= http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

[2] ASHBURNER, M., BALL, C.A., BLAKE, J.A., BOTSTEIN, D., BUTLER, H., CHERRY, J.M., DAVIS, A.P., DOLINSKI, K., DWIGHT, S.S., EPPIG, J.T., HARRIS, M.A., HILL, D.P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J.C., RICHARDSON, J.E., RINGWALD, M., RUBIN, G.M., and SHERLOCK, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet 25*, 1 (May), 25-29. DOI= http://dx.doi.org/10.1038/75556.

[3] BELKIN, M., MATVEEVA, I., and NIYOGI, P., 2004. Regularization and Semi-supervised Learning on Large Graphs. In *Learning Theory*, J. SHAWE-TAYLOR and Y. SINGER Eds. Springer Berlin Heidelberg, 624-638. DOI= http://dx.doi.org/10.1007/978-3-540-27819-1_43.

[4] BJÖRK, B.-C., ROOSR, A., and LAURI, M., Global annual volume of peer reviewed scholarly articles and the share available via different open access options. In *Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing*, Toronto, Canada.

[5] CHUNG, F.R.K., 1997. Spectral Graph Theory American Mathematical Society.

[6] COORDINATORS, N.R., 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res 42*, 1 (Jan 1), D7-D17. DOI= http://dx.doi.org/10.1093/nar/gkt1146.

[7] DA COSTA, C.A., SUNYACH, C., GIAIME, E., WEST, A., CORTI, O., BRICE, A., SAFE, S., ABOU-SLEIMAN, P.M., WOOD, N.W., TAKAHASHI, H., GOLDBERG, M.S., SHEN, J., and CHECLER, F., 2009. Transcriptional repression of p53 by parkin and impairment by mutations associated with autosomal recessive juvenile Parkinson's disease. *Nat Cell Biol 11*, 11 (Nov), 1370-1375. DOI= http://dx.doi.org/10.1038/ncb1981.

[8] DAI, C. and GU, W., 2010. p53 post-translational modification: deregulated in tumorigenesis. *Trends Mol Med 16*, 11 (Nov), 528-536. DOI= http://dx.doi.org/10.1016/j.molmed.2010.09.002.

[9] DERDAK, Z., VILLEGAS, K.A., HARB, R., WU, A.M., SOUSA, A., and WANDS, J.R., 2013. Inhibition of p53 attenuates steatosis and liver injury in a mouse model of non-alcoholic fatty liver disease. *J Hepatol 58*, 4 (Apr), 785-791. DOI= http://dx.doi.org/10.1016/j.jhep.2012.11.042.

[10] GOH, K.I., CUSICK, M.E., VALLE, D., CHILDS, B., VIDAL, M., and BARABASI, A.L., 2007. The human disease network. *Proc Natl Acad Sci U S A 104*, 21 (May 22), 8685-8690. DOI= http://dx.doi.org/10.1073/pnas.0701361104.

[11] GRAY, K.A., DAUGHERTY, L.C., GORDON, S.M., SEAL, R.L., WRIGHT, M.W., and BRUFORD, E.A., 2013. Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res 41*, Database issue (Jan), D545-552. DOI= http://dx.doi.org/10.1093/nar/gks1066.

[12] GU, B. and ZHU, W.G., 2012. Surf the post-translational modification network of p53 regulation. *Int J Biol Sci 8*, 5, 672-684. DOI= http://dx.doi.org/10.7150/ijbs.4283.

[13] HAGER, K.M. and GU, W., 2014. Understanding the non-canonical pathways involved in p53-mediated tumor suppression. *Carcinogenesis*(Feb 3). DOI= http://dx.doi.org/10.1093/carcin/bgt487.

[14] HORNBECK, P.V., KORNHAUSER, J.M., TKACHEV, S., ZHANG, B., SKRZYPEK, E., MURRAY, B., LATHAM, V., and SULLIVAN, M., 2012. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res 40*, Database issue (Jan), D261-270. DOI= http://dx.doi.org/10.1093/nar/gkr1122.

[15] JENKINS, L.M., DURELL, S.R., MAZUR, S.J., and APPELLA, E., 2012. p53 N-terminal phosphorylation: a defining layer of complex regulation. *Carcinogenesis 33*, 8 (Aug), 1441-1449. DOI= http://dx.doi.org/10.1093/carcin/bgs145.

[16] JINHA, A.E., 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing 23*, 3 (//), 258-263. DOI= http://dx.doi.org/10.1087/20100308.

[17] LANGLEY, P., BRADSHAW, G., and SIMON, H., 1983. Rediscovering Chemistry with the Bacon System. In *Machine Learning*, R. MICHALSKI, J. CARBONELL and T. MITCHELL Eds. Springer Berlin Heidelberg, 307-329. DOI= http://dx.doi.org/10.1007/978-3-662-12405-5_10.

[18] LARSEN, P.O. and VON INS, M., 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics 84*, 3 (Sep), 575-603. DOI= http://dx.doi.org/10.1007/s11192-010-0202-z.

[19] LI, M., HE, Y., DUBOIS, W., WU, X., SHI, J., and HUANG, J., 2012. Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. *Mol Cell 46*, 1 (Apr 13), 30-42. DOI= http://dx.doi.org/10.1016/j.molcel.2012.01.020.

[20] LISEWSKI, A.M. and LICHTARGE, O., 2010. Untangling complex networks: risk minimization in financial markets through accessible spin glass ground states. *Physica A 389*, 16 (Aug 15), 3250-3253. DOI= http://dx.doi.org/10.1016/j.physa.2010.04.005.

[21] MANNING, G., WHYTE, D.B., MARTINEZ, R., HUNTER, T., and SUDARSANAM, S., 2002. The protein kinase complement of the human genome. *Science 298*, 5600 (Dec 6), 1912-1934. DOI= http://dx.doi.org/10.1126/science.1075762.

[22] MAY, P. and MAY, E., 1999. Twenty years of p53 research: structural and functional aspects of the p53 protein. *Oncogene 18*, 53 (Dec 13), 7621-7636. DOI= http://dx.doi.org/10.1038/sj.onc.1203285.

[23] MEEK, D.W. and ANDERSON, C.W., 2009. Posttranslational modification of p53: cooperative integrators of function. *Cold Spring Harb Perspect Biol 1*, 6 (Dec), a000950. DOI= http://dx.doi.org/10.1101/cshperspect.a000950.

[24] MULLER, P.A. and VOUSDEN, K.H., 2013. p53 mutations in cancer. *Nat Cell Biol 15*, 1 (Jan), 2-8. DOI= http://dx.doi.org/10.1038/ncb2641.

[25] NATHANSON, J.W., YADRON, N.E., FARNAN, J., KINNEAR, S., HART, J., and RUBIN, D.T., 2008. p53 mutations are associated with dysplasia and progression of dysplasia in patients with Crohn's disease. *Dig Dis Sci 53*, 2 (Feb), 474-480. DOI= http://dx.doi.org/10.1007/s10620-007-9886-1.

[26] SALTON, G. and MCGILL, M.J., 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

[27] SHAWVER, L.K., SLAMON, D., and ULLRICH, A., 2002. Smart drugs: tyrosine kinase inhibitors in cancer therapy. *Cancer Cell 1*, 2 (Mar), 117-123.

[28] SHIEH, S.Y., AHN, J., TAMAI, K., TAYA, Y., and PRIVES, C., 2000. The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites. *Genes Dev 14*, 3 (Feb 1), 289-300.

[29] SIGANAKI, M., KOUTSOPOULOS, A.V., NEOFYTOU, E., VLACHAKI, E., PSARROU, M., SOULITZIS, N., PENTILAS, N., SCHIZA, S., SIAFAKAS, N.M., and TZORTZAKI, E.G., 2010. Deregulation of apoptosis mediators' p53 and bcl2 in lung tissue of COPD patients. *Respir Res 11*, 46. DOI= http://dx.doi.org/10.1186/1465-9921-11-46.

[30] SRINIVASAN, P., 2004. Text mining: generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol. 55*, 5, 396-413. DOI= http://dx.doi.org/10.1002/asi.10389.

[31] SWANSON, D.R., 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med 30*, 1 (Autumn), 7-18.

[32] UNIPROT, C., 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res 41*, Database issue (Jan), D43-47. DOI= http://dx.doi.org/10.1093/nar/gks1068.

[33] WHEELER, D.L., CHURCH, D.M., FEDERHEN, S., LASH, A.E., MADDEN, T.L., PONTIUS, J.U., SCHULER, G.D., SCHRIML, L.M., SEQUEIRA, E., TATUSOVA, T.A., and WAGNER, L., 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res 31*, 1 (Jan 1), 28-33.

[34] ZHOU, D., BOUSQUET, O., WESTON, J., and SCHOLKOPF, B., 2004. Learning with local and global consistency. In *Adnvaces in Neural Information Processing Systems (NIPS) 16* MIT, 321-328.