Bioinformatics and Biosystems 2006, Vol. 1, No. 1, pp. 51-60

# Visualization for Digesting a High Volume of the Biomedical Literature

**Changsu Lee[1], Jinah Park[2], Jong C. Park[3]**

**[1,3]KAIST, 373-1 Guseong-dong, Yuseong-gu, Daejeon, South KOREA**

**[2]ICU (Information and Communications University), 119 Munjiro, Yuseong-gu, Daejeon, South KOREA**

**[1]burning@nlp.kaist.ac.kr, [2]jinah@icu.ac.kr, [3]park@cs.kaist.ac.kr**

## Abstract

The paradigm in biology is currently changing from that of conducting hypothesis-driven individual experiments to that of utilizing the results of a massive data analysis with appropriate computational tools. We present LayMap, an implemented visualization system that helps the user to deal with a high volume of the biomedical literature such as MEDLINE, through the layered maps that are constructed on the results of an information extraction system. LayMap also utilizes filtering and granularity for an enhanced view of the results. Since a biomedical information extraction system gives rise to a focused and effective way of slicing up the data space, the combined use of LayMap with such an information extraction system can help the user to navigate the data space in a speedy and guided manner. As a case study, we have applied the system to datasets of journal abstracts on 'MAPK pathway' and 'bufalin' from MEDLINE. With the proposed visualization, we have successfully rediscovered pathway maps of a reasonable quality for ERK, p38 and JNK. Furthermore, with respect to *bufalin*, we were able to identify the potentially interesting relation between the Chinese medicine *Chan su* and apoptosis with a high level of detail.

**Keywords**: visualization, information extraction, biomedical literature

## Introduction

The amount of information available in biomedical domain is not only already enormous, but is also growing extremely fast. At the same time, much attention is currently being drawn to a new paradigm that utilizes various information technologies that specialize in dealing with such a high volume of information in an efficient manner. In particular, the paradigm in biology is shifting from that of conducting hypothesis-driven individual experiments to that of utilizing the results of massive data analysis with appropriate computational tools (Fields 2001).

It is well known that the vast majority of the data in biomedical domain is in the form of unstructured and usually English texts. And related resources such as MEDLINE are receiving much attention recently as an important target for biomedical information. Since information from such resources is not readily digestible for individual researchers due to its high volume, we can expect that information extraction (IE) systems, which help extract meaningful facts from such resources, are likewise crucially important, as witnessed by a number of related studies (Park 2001, Friedman 2001, Yandell 2002, Hirschman 2002).

Although IE systems drastically reduce the search space for interesting pieces of undiscovered knowledge by producing facts of specified patterns from unstructured resources, they do not work as a *direct* means for discovering such pieces of knowledge. This is due to the fact that such pieces of knowledge are often not retrievable, but inferable, from relevant facts of a first level, which is the major concern for researchers in the field of knowledge discovery.

We believe that a visualization system plays an important role in complementing the functions of an IE system by guiding the user in his/her inference process over the extracted facts. Visualization systems enhance the level of interaction with the user, in particular by stimulating the high pattern recognizing abilities of the user, making it possible to understand data of a high complexity and to lead to consequent knowledge discovery (Hinneburg 1999, Ankerst 2000, Shneiderman

Corresponding Author: Jinah Park (Email: park@cs.kaist.ac.kr)
This work was supported by MOST/KOSEF through AITrc and Grants for Interdisciplinary Research (R01-2005-000-10824-0).

2002). They have been proposed for a wide range of fields in biology for the analysis of complex data, such as sequence analysis (Mayor 2000), protein-protein interaction analysis (Lee 2002, 2004) and metabolic pathway analysis (Goesmann 2002).

In this paper, we propose LayMap, an implemented visualization system that helps the user to digest such a high volume of the biomedical literature in MEDLINE. LayMap utilizes, in addition to an information extraction system, filtering, granularity and layered visualization as a novel technique. Each layer represents a different level of detail, making it possible to digest a high volume of the literature in an efficient way. Since biomedical information extraction systems give rise to a more focused and effective way of slicing up the data space, the combined use of LayMap with such an information extraction system can help the user to navigate the data space in a speedy and guided manner.

As a case study, we have applied the system to datasets of 2977 journal abstracts from MEDLINE: 2914 abstracts on 'MAPK pathway' and 'human', and 63 abstracts on 'bufalin' and 'human'. We used 2 layers, one of which shows a low level of detail for 2914 abstracts, whereas the other shows a high level of detail for 63 abstracts. Among others, we have successfully rediscovered pathway maps of a reasonable quality for ERK and p38 mainly based on the visualization, and the pathway map of a comparable quality for JNK with some domain-specific inference. With respect to *bufalin*, we were able to identify the potentially significant, but a not yet well-known relation between the Chinese medicine *Chan su* and apoptosis with a high level of detail.

The rest of this paper is organized as follows. Section 2 reviews related work in biomedical visualization. Section 3 introduces our technique of visualizing the user's exploration over the MEDLINE literature and of integrating the visualization system and an information extraction system. Section 4 presents a case study of knowledge discovery through visualization and analysis of mitogen-activated protein kinase (MAPK) pathways by the proposed system. In Section 5, we discuss the discovered pieces of knowledge with a domain expert. The last section shows a summary of this paper and discusses possible directions for future research.

## Related Work

Visualization systems are in use in a biomedical domain for a number of purposes. We discuss here only those that are applied to molecular interactions. Koike (2000) introduced a graphic editor for the analysis of complex signal transduction pathways. They have implemented CUtenet, a Java application for the representation, visualization and analysis of signal transduction pathways of eukaryote that have both a biochemical level and a logical level. However, since CUtenet is a graphic editor, it provides those functions of *managing* the data, rather than *digesting* the data. Thus it is designed to operate on a more-or-less curated data set, unlike our system that can also operate on a coarse data set.

Mrowka (2001) presents a visualization system for protein-protein interactions, by taking into account the functional grouping and the neighboring distance of proteins for the visualization. However, the system only takes a precompiled data set for visualization, unlike ours that allows the users to seek out for interesting data subsets.

Some visualization systems are also utilized in conjunction with information extraction systems. Feldman *et al.* (2002) claim that the user can easily navigate a large collection of biomedical documents by combining information extraction and visualization. They also claim that text mining needs pre-processing, and that information extraction is suitable for such pre-processing. However, the proposal does not allow for any cycle between information extraction and visualization, unlike ours that emphasizes an active cycle between the two to deal with a high volume of literature.

Friedman *et al.* (2001) present GENIES, a natural language processing module in GeneWays, which extracts molecular pathway information from literature and manages the resulting body of knowledge. However, their visualization system does not show the consideration for scalability. In particular, it does not provide a way of dealing with a high volume of data that are produced by an information extraction system for an even higher volume of the literature.

Wong (2001) describes PIES, a system for the extraction, modification, and management of protein-protein interaction pathways. He argues that since the results of IE are not perfect, it is important that we tune them with visualization. As such, the visualization system in PIES supports the modification for each node. However, he does not propose a method to deal with a high volume of data either.

## Methods

Section 3.1 describes the overall organization of the proposed system and explains how to combine information extraction and visualization. In particular, it discusses several issues related to the visualization of the results of information

extraction. Section 3.2 explains the filtering and granularity techniques that are utilized by LayMap, in order to visualize the results that are extracted from a large volume of the literature. In particular, we have made it possible to represent a different level of detail by allowing the user to set the level of filtering and granularity. Section 3.3 presents a layered visualization technique as a way of representing different levels of detail at the same time. The technique utilizes the notion of layers to deal with different levels of detail, where with controlled transparency, the user can focus on the layers one by one, thus effectively visualizing a large volume of the literature.

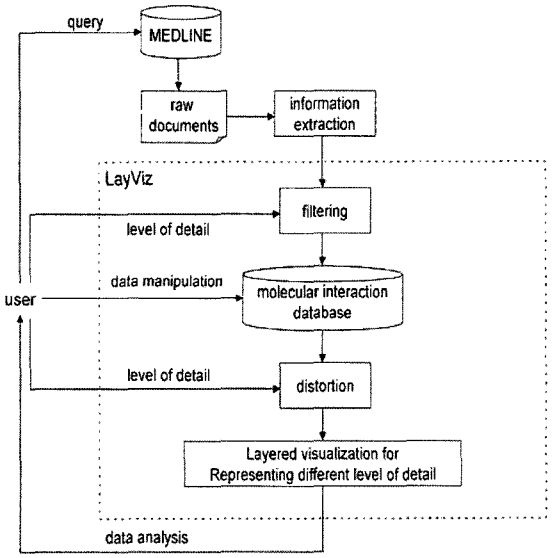



**Figure 1.** System Outline

### Visualizing Molecular Interactions from IE

The target data space is MEDLINE in our study. The user may first use PubMed queries to collect relevant documents from MEDLINE. The more specific a query term is, the more probable it is to retrieve the document set that meets the users' intention. Nevertheless, it is often quite difficult to know in advance the query terms that are specific enough for the purpose at hand, since it is usually the case that the user does not have sufficient knowledge as to what to look for. In this case, the user may instead make up successively more specific, and thus more effective, queries by utilizing the earlier search results that are collected by less specific queries.

The binary relations, such as *inhibit, activate*, or *bind*, between molecules are produced as the output of an IE system, which are then provided as input to the visualization system. Among them, some are visualized and some filtered out. The user may make up more specific queries or the queries that contain the name of the molecule on the user's interest, then repeat the processes described above. Figure 1 describes these processes pictorially.

We place the results of information extraction on a 2-D space. We assume that the results are represented as binary relations between terms, i.e., English words or phrases. Extracted terms are represented as nodes, and the relations as edges. Relations are color-coded as shown in Table 1. All the relations except for *bind, interact, associate*, and a few other relations, grouped as *others*, are directed and represented by a cone.

**Table 1.** Color assignment by relation type

| relation | color |
|---|---|
| bind, interact, associate | purple |
| activate, simulate, induce, accelerate | red |
| inhibit, block, prevent | green |
| phosphorylate | bluish green |
| component, part-of | yellow |
| others | blue |

We used the public domain program GraphViz by AT&T for the placement of objects (Gasner 1993), and the layout algorithm *neato*. As shown in Figure 1, the results of information extraction are first filtered and then converted into an input file for GraphViz. The *neato* algorithm in GraphViz generates a text file containing the $x$ and $y$ coordinates of each node. This text file is then provided as an input to LayMap. Figure 2 shows an example layout by *neato*, where the extracted results are shown for documents collected from MEDLINE with the keywords 'MAPK' and 'human'.

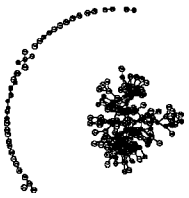


**Figure 2.** A layout with GraphViz using neato

Figure 2 shows that a large number of connected components that are rather small are on the verge. These are due to the artifact that is introduced when the IE system processes the source natural language texts without resolving the naming problem. To see the nature of this problem, consider the example sentences below. (The relation to the right of the arrow

corresponds to the output of the IE system on the input sentence on the left.)

- HGF-induced survival correlates with Akt activity and is inhibited by the specific PI3-kinase inhibitor LY294002 ... (PMID: 11134526) → **inhibit** (LY294002, the specific PI3-kinase)
- We also found that the colony formation of CEF is strongly induced by a constitutively active PI3K mutant, and that a PI3K inhibitor, LY294002 ... (PMID: 10852971) → **inhibit** (LY294002, a PI3K)

We can see that LY294002 is a shared argument by both of the results, and that both of the results are about the same inhibit relation on *LY294002.* In addition, a morphological analysis of *a PI3K* and *the specific PI3-kinase* suggests that both expressions are in fact about the same objects even though they are different in form. We will discuss why visualization sheds light to problems of this kind in Section 4.2.

**Filtering and Granularity Techniques**

LayMap provides a filtering mechanism that blocks out certain pieces of information from IE, for instance those arguments that do not appear more than a given number times, say 3, in the whole dataset. This mechanism is utilized for the following reason. When there is a huge collection of documents, the amount of information that the IE system produces may also become quite large. This means that the number of entities the visualization system must handle can become quite large as well. While those molecules that do not show up as frequently as the others have a varied significance[1)], the visualization system may have a lesser burden to display so many entities by filtering them out, with understood risks in doing so.

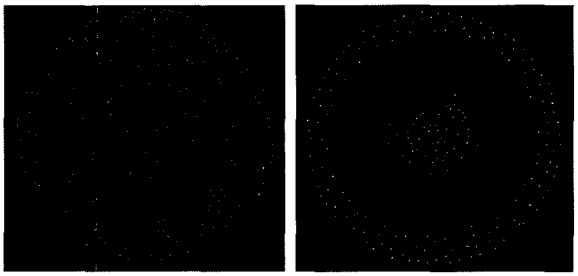
The visualization system with a lighter burden allows the user to operate it a lot easier, and consequently makes it possible for the user to discover much more information. In general, the more common the keywords are, the larger the resulting collection of documents will be. When a novice user wishes to use a visualization system for the exploration over MEDLINE, it is quite likely that he/she starts with a common keyword combination. Thus the filtering mechanism is helpful to novice users, as well as experts, for effective exploration.

We represent the frequency of particular relations in the literature by the thickness of the corresponding edges. This way, we may gain crucial insights into much studied molecules. We may also emphasize with thickness such pieces of information that the user must pay special attention to. By appropriately distorting the frequency information, we have allowed the user to set a primary focus on a particular subset of the large volume of data.

A comparison between the visualization system that utilizes both filtering and granularity and one that does not is depicted in Figure 3. It visualizes the extracted results from the documents collected from MEDLINE with the keyword combination '*Ras* AND *Raf* AND *GTP*'. Figure 3-(2) visualizes subsets of data, where both arguments appear in the documents more than twice. A large interconnected network, located in the bottom-right corner of (1), is at the center of (2). We can see that many small connected components do not show up in (2).

We can see in Figure 3-(2) that the networks with higher interconnectivity have thicker edges. Krauthammer *et al.* (Krauthammer 2002) claim that important molecules have a high interconnectivity. This supports our prediction that more frequently appearing molecules in the documents are more important.

The user can specify the minimum number of times that a particular node must appear during filtering. He/she can likewise specify the thickness on a certain frequency during granularity. Through this process, the user can choose the level of detail for effective visualization. For instance, we can set the level of filtering high when there is a huge amount of the literature and we need to hide a lot of the detail, say the minimum frequency of 10. We can opt out the filtering altogether, when the size of the available literature is manageable, or when we desire a closer look at the available literature. As for granularity, we can also set the level of thickness sensitive when we want to have a detailed control over the change of frequency, and less sensitive when we do not.



(1) (2)

**Figure 3.** A comparison between the visualization system that utilizes both filtering and granularity and one that does not

1) For instance, they may be due to the aforementioned naming problem. In other cases, they may play an important role as crucial pieces of evidence for certain knowledge discovery.

**Visualizing Different Levels of Detail**

LayMap utilizes the filtering and granularity techniques (discussed in Section 3.2) to help in setting the level of detail for visualization depending on the user's preference. However, in order to map out the large amount of the literature effectively, we may need to cut down on the details for some of the data, and zoom in on for some other. For instance, when we visualize a large volume of the literature on cancer, we can set to visualize the low level of detail until we want more details on *CIN 3*. We can then set a high level of detail for *CIN 3*, effectively digesting a large volume of the literature. We propose layered visualization for different levels of detail of this kind. In this subsection, we explain the core algorithm for maintaining layered visualization with an example below.

1. Let a set of nodes {p1, p2, p3, p4, p5, p6, p7, p8, p9} constructed from the initial query (p1) be A. Suppose that the user wanted a low level of detail for the bird's eye view of the results. Suppose that the set of nodes after strong filtering is A' = {p1, p2, p3, p4, p5}.
2. If the user comes to have an interest in p2 and p3 after examining the analysis result of the visualization, he/she may make a more refined query, such as (p2, p3). Let a set of nodes {p2, p3, q2, q3} constructed from the second query be B. Suppose that we do not apply filtering so that we use a high level of detail for visualization. Let B' be the set of nodes.
3. We use the neato algorithm again for the union of A' and B'. We move only a set of entities that are contained in the set B' - A' = {q2, q3} onto a different layer. When we move entities, we change only their values in the y coordinate, not in the x and z coordinates (cf. Figure 4). In Figure 4, entities that are contained only in A' are colored in grey.
4. Repeat steps 1 - 3, displaying the results at each separate layer along the +y direction, until the user is satisfied with the chosen level of details.

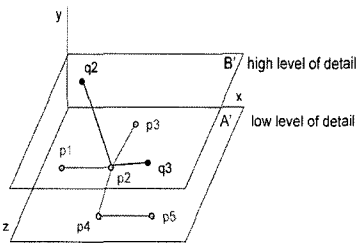



**Figure 4.** Visualization for Representing Different Levels of Detail

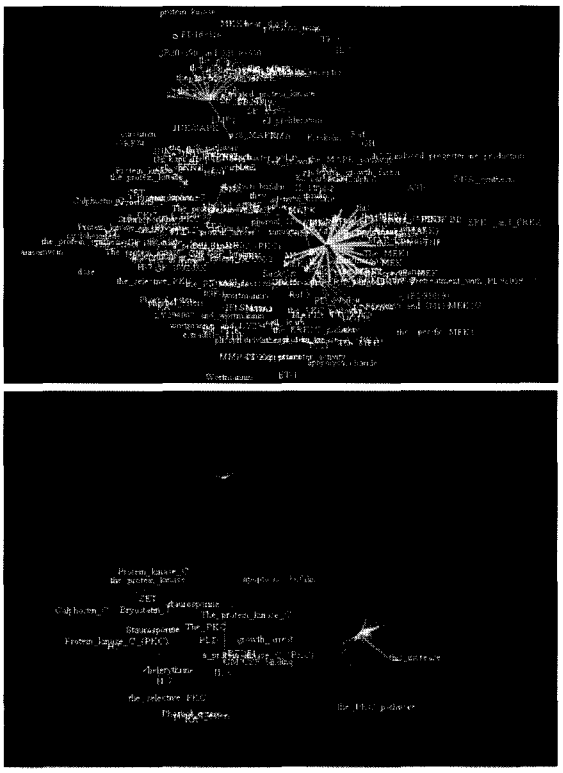


**Figure 5.** Comparison: Using Transparency Technique (converted to images in reverse video to enhance readability)

The user can repeat the process above to make more than 2 layers, and set each layer with a different level of detail. Each node is, however, visualized only once in the entire layers, and when the node is shared by more than one layer, it shows up only at the lowest layer. In order to avoid confusion when the user examines the entire layers at the same time, we may need to place the multi-layer nodes in only one of the layers. However, it is more general to start with a low level of detail and to proceed to a higher level of detail, than the other way around, when we visualize a large volume of the literature. It is thus quite helpful to place the multi-layer nodes in the lowest layer, so that the number of edge crossings can be reduced accordingly.

We can thus set out to focus on each level of detail one by one, as well as on different levels of detail at the same time. If the user wants to concentrate on a set of data in either A' or B', a different set of data in another layer may confuse the user. In order to avoid this problem, we make all the layers transparent except those that the user has an interest in. In addition, since the user can choose any layer that is in their interest, the selective concentration on the whole data

space becomes possible. When we represent a layer opaque, the related nodes in other layers are also represented opaque. For instance, when we represent the layer B' opaque in Figure 4, the related nodes are p2, q2, and q3. Figure 5 depicts the comparison between the visualization technique with the transparency module and that without. We can see that the use of transparency is an effective measure for helping the user to concentrate on a subset of the whole data. Figure 5 shows a visualization example in two layers, where the lower one shows a low level of detail, and the upper one a high level of detail. The figure on the top in Figure 5 represents none of the layers opaque, whereas the figure at the bottom shows the lower layer transparent and the upper one opaque.

A case study on mitogen-activated protein kinase pathways will be shown in the next section, attesting to the thesis that the proposed visualization indeed helps the user to explore MEDLINE.

## Application Examples

We conducted a case study for the visualization of molecular interaction information on MAPK. For the implementation of a prototype system, we used Visual C++ and OpenInventor. The system provides interactive visualization for the users to move, zoom in and out, and rotate objects.

### Settings

We utilize BioIE (Kim and Park 2004) for the information extraction of molecular interaction information from MEDLINE. The sample input and output of BioIE are shown in Figure 6.

An MAPK pathway is an important signal transduction pathway for the regulation of cell responses. It is composed of three modules: MKKK (MAPK Kinase Kinases), MKK (MAPK Kinases) and MAPK. A variety of extracellular signals such as growth factors, stresses, and differentiation factors activate MKKK, MKK, and MAPK, in that sequence. MAPKs regulate growth, survival and apoptosis through various proteins including *Chop* and *c-Jun* (Garrington 1999, Widmann 1999).

In this section, we show how visualization may help to rediscover MAPK-related established knowledge and indicate the possibility of MAPK-related novel knowledge. For this purpose, we will show the process of rediscovering some pieces of the MAPK-related knowledge, which include the following information:

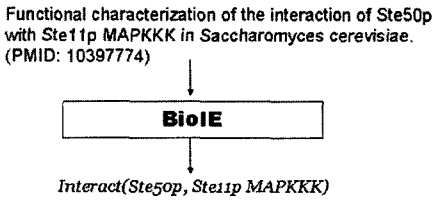



**Figure 6.** Sample input and output information of BioIE

- That mammalian MAPK pathways are of four types: ERK, JNK, p38, and MKK5/ERK, and
- that ERK, JNK, and p38 regulate cell responses as follows: ERK regulates growth and differentiation; JNK regulates growth, differentiation, survival, and apoptosis; and p38 regulates cytokine production, and apoptosis.

### Keyword: MAPK pathway AND human

We have assumed that the user does not have any prior knowledge of MAPK pathways, and used the keyword combination 'MAPK pathway' and 'human' for MEDLINE to collect 2914 journal articles. From these journal abstracts, BioIE extracted 9651 pieces of molecular interaction information. We have used the keyword *human* as the terms on MAPK are dependent upon the species it applies. We used a low level of detail for the bird's eye view of the data. For this purpose, we set the filtering high, and the granularity low. We also set the minimum number of occurrences to 3 for filtering, and raised the thickness by one per 10th frequency increase for granularity. The total number of visualized relations is 486. Figure 7 shows a snapshot of the visualization system. It shows two characteristic sub-networks, whose centers have noticeably many connections with thick edges, as identified on the left and on the right. Figure 8 (a) and (b) show close-up views of the two sub-networks.

### ERK pathway:

Figure 8(a) is a closed-up view of Figure 7 (1). Figure 8 (a) shows PD98059 at the center of the network, with green edges spreading out from it. Notice that the green edges denote the inhibition relation. The network shows that there are a number of nodes with various labels, such as 'the ERK pathway', 'the ERK', 'ERK1/2', 'the ERK1/2 pathway', 'MAPK', 'MAP kinase', 'the MAPK kinase (MEK)', 'the MAPK kinase 1', 'MEK-1', and 'proliferation', that are all "inhibited" by PD98059. A short digress is in order. If we consider the fact that the source English sentences for such relations are full of virtually synonymous expressions, this phenomenon at hand would be of particular concern, since it may present an im-

portant clue as to the real nature of the target relations. In other words, when inspected closely, the English sentences under consideration may indicate that those extracted relations between the pairs X and Y, X' and Y, X" and Y, and so on, are in fact the same relations between [X] and Y, where [X] is a generic name for all X, X', X" and so on. This is quite likely, due to the naming problem.

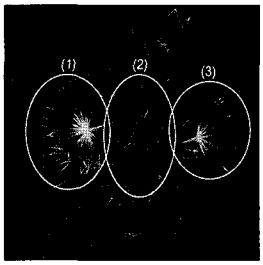
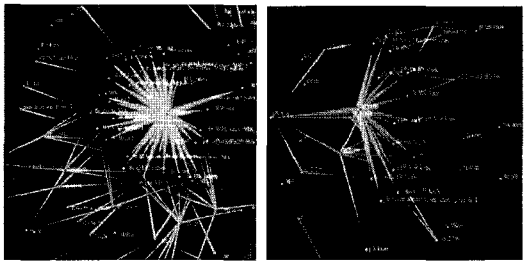



**Figure 7.** Visualization: keywords of 'MAPK pathway AND human'



(a) (b)

**Figure 8.** Visualization of (a) part number 1 and (b) 3

Coming back to the multitude of nodes with different labels, all of which are related to a single node, i.e., PD98059, the naming problem gives us a strong indication that ERK may in fact be synonymous to MAPK. We know that this is an established fact, but the users may not know this for sure at this point. In addition, with the prior knowledge of the syntactic constructions of English expressions such as 'the MAPK kinase (MEK)', we know that MEK is synonymous to MKK, or the MAPK kinase. We are now ready to put together the jigsaw puzzle pieces. For instance, we may expect that the MAPK kinase (MEK), the MAPK kinase 1, and MEK-1 are all morphological variations of MEK-1, and that MEK-1 is synonymous to MKK. We may also expect that PD98059 is an ERK1/2 inhibitor, as well as an MEK-1 inhibitor. Finally, we predict that ERK1/2 controls proliferation, since PD98059 inhibits proliferation.

When we examine the nodes that represent ERK more closely, we can classify those that inhibit the ERK pathway (green edges) and those that activate it (red edges). Among those that activate the ERK pathway are: PDGF-BB, TNF, ET-1, GHRH, PDT, NS-398, 5-HT, and RSV, and that inhibit the ERK pathway are: Akt and U-0126. We see that since growth factors activate the ERK pathway, the ERK pathway comes to control growth. In addition, we see that it is Raf-1 that phosphorylates and activates MEK-1. We may thus conclude that it is af-1 that corresponds to MKKK, since MEK-1 corresponds to KK in the ERK pathway.

**Figure 9**: Visualization of part number 2

**JNK pathway:**

Figure 9 is a close-up view of Figure 7 (2). Figure 9 shows three major sub-networks, with centers JNK, apoptosis, and MAPK. Since apoptosis and MAPK are general terms, they do not constitute the pieces of knowledge on MAPK that we need to pay attention to. Therefore, it is safe to say that we can just look into the sub-network with JNK at its center.

The JNK-centered sub-network shows that the molecules that activate JNK include: TPA, TNF-alpha (Tumor necrosis factor-alpha), ASK1, IGF-1, and anisomycin. We (the users) do not know yet which functions as MKK and, likewise, which functions as MKKK. But we (the users) know that, since growth factors activate the JNK pathway, the JNK pathway controls growth. And we see that quercetin inhibits JNK, and that JNK phosphorylates c-Jun. The JNK-centered sub-network, in comparison to the ERK-centered sub-network discussed earlier and the p38-centered sub-network to be discussed, is rather small. This indicates that JNK-related studies are not conducted as actively as p38- and ERK1/2-related studies.

We can see that the 'bufalin' at the center of Figure 9 activates MAPK and that it has a relation with apoptosis. This observation allows us to predict that bufalin activates an MAPK that controls apoptosis. This prediction will be borne out by an analysis of the visualized results, whose inputs are extracted by BioIE from the documents collected with the keyword combination 'bufalin' and 'human'. Further details will be shown in Section 4.3. We were able to figure out the fact that apoptosis is an important aspect of life through the size

of the network with apoptosis at the center in Figure 9, but we were not able to figure out the relation between MAPK and apoptosis yet. We have therefore chosen to examine in closer detail those nodes that are related to both MAPK and apoptosis, e.g., EGF, bufalin, and TNF-alpha.

**p38 pathway:**

Figure 8 (b) is a closed-up view of Figure 7 (3). Figure 8 (b) shows that SB203580 is at the center of the sub-network. All the edges in this sub-network are green, and those connected to SB203580 are 'p38 kinase', 'p38', 'the specific p38', 'the p38 MAP kinase', 'p38(MAPK)', and 'the p38', mostly morphological variations of p38 MAPK.

Among those that inhibit p38 MAPK are PD169316 and SB202190, in addition to SB203580. On the other hand, the sub-network shows that activators of p38 MAPK include MKK6, palytoxin, TGF-beta, CO, IL-1, IL-2, and arsenite. And, by an analysis of the labels, we see that MKK6 is the MKK in the p38 pathway. Furthermore, the fact that IL-1 (interleukin-1) and IL-2 (interleukin-2) activate the p38 pathway suggests that the p38 pathway controls cytokine production.

**Keyword: bufalin AND human**

We collected 63 MEDLINE abstracts with the keyword combination 'bufalin' and 'human'. BioIE extracted 108 relations from them. We have set the data in a high level of detail to identify the MAPK through which bufalin induces apoptosis. We have thus set the filtering very low and the granularity very high (i.e., sensitive). The minimum number of occurrences for filtering is set to 1 (i.e., no filtering is applied), and the level of thickness is set to one step up for each 3rd increase in the frequency for granularity.

Figure 10 shows the existing results with a low level of detail and the extracted results with the keyword combination 'bufalin' and 'human' with a high level of detail. We make the lower layer transparent in order to help the user concentrate on the new pieces of knowledge. Figure 11 shows a close-up view of the results from the second query, 'bufalin AND human'.

We can see that bufalin activates JNK and that it induces apoptosis through the nodes of *bufalin* and *JNK* and *the induction of apoptosis*. We can also see that bufalin activates AP-1, that it has a relation with inflammatory cytokine, and that it inhibits testosterone production. The nodes of bufalin and *Chan su,* located at the bottom of Figure 11, indicate that bufalin is a component of the Chinese medicine *Chan su.* Through the results of the analysis of Figure 11, we suspect that *Chan su,* which includes bufalin as an active component, may activate JNK and thus may induce apoptosis. This suspected relation between *Chan su* and apoptosis gives biologists a direction for future experiments. This also gives rise to a good cooperation model for accelerated biological knowledge discovery, in the sense that it provides a combined view of a guided plan for experiments through the proposed visualization and the experiment skill by biologists.

**Figure 10.** Visualization: keyword combination 'bufalin AND human'

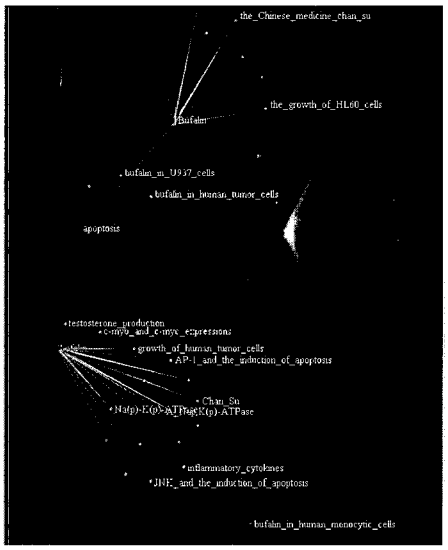



**Figure 11.** Close-up View

We then looked into the question of whether or not the suspicion on the relation between *Chan su* and apoptosis is factually grounded. We collected the documents from MEDLINE with the keyword combination '*Chan su* AND *apoptosis*'. We have discovered 6 abstracts. However, all of these abstracts mentioned that bufalin, an active component of *Chan su*, induces apoptosis, but none of them mentioned a direct relationship between *Chan su* and apoptosis. This shows

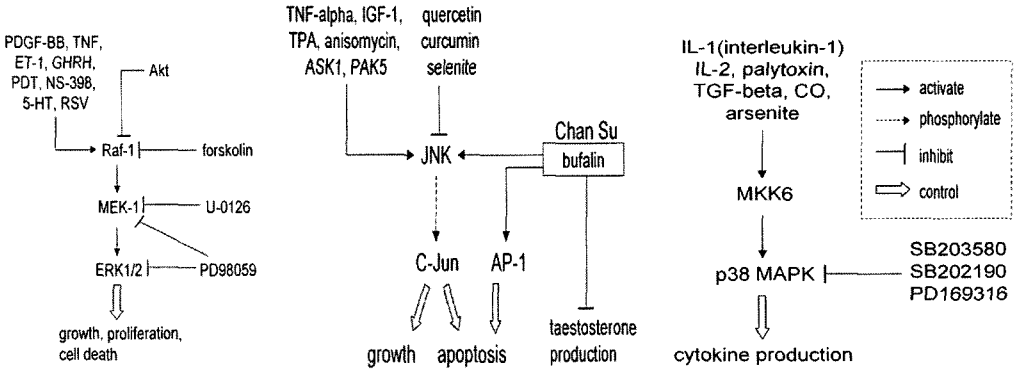



**Figure 12.** ERK, JNK and p38 pathways

that the proposed visualization helps the user to find a clue for biomedical knowledge discovery by effectively visualizing the user's exploration process in the massive sea of data space, or MEDLINE.

Figure 12 summarizes what we have found with respect to the ERK pathway, the JNK pathway, and the p38 pathway, respectively. We need to pay attention to JNK pathway, which contains a suspected relation between the Chinese medicine *Chan su* and apoptosis.

## Discussion

We have shown that by utilizing visual analyses, we can effectively gain pieces of knowledge on MAPK pathways from a high volume of the biomedical literature in MEDLINE. The (re-)discovered pieces of knowledge are consulted with a domain expert in *kinases*, who confirmed that they represent fairly common pieces of knowledge for domain experts in kinases.

Among others, we have seen that PD98059 and SB203580 are the most frequently discussed inhibitors of ERK1/2 and p38. By using the granularity technique, or varying the thickness of the edges, according to their frequency, we were able to identify ERK1/2 and p38 with MAPK through these inhibitors.

We have rediscovered only "common" pieces of knowledge from the perspective of domain experts in kinases from 2914 journal abstracts from the user's initial keyword, 'MAPK' and 'human'. This is due to the fact that our intention has been different from the intention of a domain expert, in the sense that we have chosen to use a low level of detail only for the extracted results on 'MAPK' and 'human' for the purpose of mapping out the entire network. A domain expert may want to set a high level of detail with a more refined keyword combination for the desired pieces of knowledge.

However, the domain expert shows a strong interest on the suspected relation between the Chinese medicine *Chan su* and apoptosis. We were able to discover this relation as we have utilized different levels of detail in an effective manner. In particular, a low level of detail was utilized to identify *bufalin*, which is related to both MAPK and apoptosis, and a high level of detail was utilized to identify the suspected relation between Chan su and apoptosis on the extracted relations from the MEDLINE literature with the keyword 'bufalin'.

## Conclusion

In this paper, we have proposed LayMap, an implemented visualization system that helps the user to digest a high volume of the biomedical literature in MEDLINE. LayMap has utilized not only an information extraction system, but also filtering, granularity and a novel visualization technique that involves layering. Each layer represents a different level of detail so that a high volume of the literature is effectively digested. In particular, we have shown that the implemented system enables the users to rediscover the pieces of existing knowledge and to discover the potentially novel knowledge about molecular interaction information on MAPK. The discovered pieces of knowledge on MAPK can hardly be obtained without a proper visualization technique for a high volume of literature, such as filtering, granularity and visualizing different levels of detail.

The future work includes applications of the proposed visualization technique and the corresponding implementation to cancer-related diseases that are currently at the center of attention in biology. It also includes the usability study for domain

experts.

## References


[1] Ankerst M, Ester M, Kriegel HP (2000) Towards an effective cooperation of the user and the computer for classification. *Knowledge Discovery and Data Mining,* pages 199-188.

[2] Feldman R, Regev Y, Finkelstein-Landau M, Hurvitz E, Kogan B (2002) Mining biomedical literature using information extraction. *Current Drug Discovery,* pages 19-23.

[3] Fields S (2001) The interplay of biology and technology. *PNAS*, 98(18):10051-10054.

[4] Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A (2001) Genies: a natural language processing system for the extraction of molecular pathways from journal articles. In *Proc. of ISMB*, 17:S74-S82.

[5] Gansner ER, Koutsofios E, North SC, Vo KP (1993) A technique for drawing directed graphs. *Software Engineering*, 19(3):214-230.

[6] Garrington TP, Johnson GL (1999) Organization and regulation of mitogen-activated protein kinase signaling pathways. *Current Option in Cell Biology*, 11:211-218.

[7] Goesmann A, Haubrok M, Meyer F, Kalinowski J, Giegerich R (2002) Pathfinder: Reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, 18(1):124-129.

[8] Hinneburg A, Keim DA, Wawryniuk M (1999) HD-eye: Visual mining of high-dimensional data. *IEEE Computer Graphics and Applications*, 19:22-31.

[9] Hirschman L, Park JC, Tsujii J, Wong L, Wu CH (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553-1561.

[10] Kim J and Park JC (2004) BioIE: Retargettable information extraction and ontological annotation of biological pathways from literature, *Journal of Bioinformatics and Computational Biology,* 2(3):551-568.

[11] Koike T and Rzhetksy A (2000) A graphic editor for analyzing signal-transduction pathways. *GENE*, 259:235-244.

[12] Krauthammer M, Kra P, Iossifov I, Gomez SM, Hripcsak G, Hatzivassiloglou V, Friedman C, Rzhetsky A (2002) Of truth and pathways: chasing bits of information through myriads of articles. In *Proc. of ISMB*, 18:S249-S257.

[13] Lee C, Park J, Park JC (2002) BiopathwayBuilder: Nested 3D visualization system for complex molecular interactions. *Genome Informatics*, 13:447-448.

[14] Lee C, Park J, Park JC (2005) A graphic tool for curating molecular interaction networks from the literature, *Computers in Biology and Medicine*, 35:555-564.

[15] Mayor C, Brundo M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) Vista: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics,* 16(11):1046-1047.

[16] Mrowka R (2001) A java applet for visualizing protein-protein interaction. *Bioinformatics*, 17(7):669-670.

[17] Park JC, Kim HS, Kim JJ (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Proc. of PSB*, 6:396-407.

[18] Shneiderman B (2002) Inventing discovery tools: Combining information visualization with data mining. *Information Visualization*, 1:5-12.

[19] Widmann C, Gibson S, Jarpe MB, Johnson GL (1999) Mitogen-activated protein kinase: Conservation of a three-kinase module from yeast to human. *Physiological Reviews*, 79(1):143-180.

[20] Wong L (2001) PIES, a protein interaction extraction system. In *Proc. of PSB*, 6:520-531.

[21] Yandell MD, Majoros WH (2002) Genetics and natural language processing. *Nature Reviews Genetics,* 3(8):601-610.