Boyack, Kevin W., Mane, Ketan and Börner, Katy. (2004). Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. IV2004 Conference, London, UK, pp. 965-971.

Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research

Kevin W. Boyack[†], Ketan Mane[‡], Katy Börner[‡]
[†] VisWave LLC, Albuquerque, NM 87122
[‡] School of Library and Information Science, Indiana University, Bloomington, IN 47405
{boyack@viswave.com, kmane@indiana.edu, katy@indiana.edu}

Abstract

What is the structure of the research reported on melanoma? How has it evolved over the last 40 years? Which parts of this research field are correlated with the study of genes and proteins? Are there sudden increases in the number of occurrences of certain gene or protein names, reflecting a surge of interest? How are genes, protein and papers interconnected via co-occurrence patterns?

This paper aims to provide answers to these questions by analyzing a data set consisting of papers from Medline, genes from the Entrez Gene database, and proteins from UniProt. Word burst detection and co-occurrence analyses were both performed. The spatial layout algorithm VxOrd was applied to create the very first map that shows papers, genes, and proteins and their co-occurrence relationships. The results were validated by five domain experts leading to a number of interesting facts pertaining to structure and dynamics of the melanoma research field.

1. Introduction

Given the explosive growth of biomedical databases, a large and diverse number of approaches have been suggested to automatically extract information and knowledge out of data. There are information retrieval tools, text mining tools, clustering tools, categorization tools, and text summarization tools. In addition, a number of information extraction, semantic annotation, and knowledge discovery tools have been developed. A recent review of those tools can be found in [1]. Many of the proposed approaches and tools draw on research done in not only statistics and linguistics, but also bibliometrics and social network analysis. Information visualization techniques [2] are frequently applied to manage the complexity of data, information and knowledge, and to communicate results to diverse stakeholders.

Rather than discovering information and knowledge from data, the work presented here aims to give researchers a more global view of the structure and dynamics of a research domain. The resulting 'birds eye picture' view is intended to show opportunities for collaboration and to minimize unfruitful duplication of research despite the increasing specialization of science.

In this paper we will present approaches that can help scientists to answer questions such as: What is the structure of the research reported on a particular field? How has it evolved over the course of its history? Which parts in this research field study what biological entities (e.g., gene and proteins)? Are there sudden increases in the number of occurrences of certain biological entities reflecting a surge of interest? How are biological entities and papers reporting our knowledge on them interconnected?

We demonstrate our approaches on a data set containing 53,804 papers, 299 genes and 367 proteins related to research on melanoma. Kleinberg's burst detection algorithm [3] and advanced knowledge domain visualization techniques [4] are applied to characterize the structure and dynamics of this research field over the last 40 years.

2. Process and Dataset Characterization

The process of generating a map that shows the association linkages between papers, genes, and proteins in a common context is identical to the process used in many other cases for literature alone [4], and is as follows:

1. Collection of appropriate data records, in this case papers, genes, and proteins related to melanoma,
2. Calculation of pairwise similarities between records,
3. Ordination, or layout of the records, based on calculated similarities,
4. Visualization and exploration of the data, enabling characterization and analysis of the data.

All four steps are explained in detail subsequently.

2.1. Data collection

Three types of data related to melanoma were retrieved for this study: papers, genes, and proteins. First,

the published literature, a total of 54,016 records over the period from 1960 until the date of the query, Feb. 11, 2004, related to melanoma was collected from Medline[1] using a general search on the single term 'melanoma.' Of these, a few records were later excluded from the data set due to improper formatting or being incomplete records. 53,804 papers were retained for analysis. We feel confident that the Medline data adequately represents published knowledge on melanoma given the breadth of the original query.

Second, genes for 'melanoma' were obtained through query at the Entrez Gene database[2], which returned a list of 304 genes. From the list of genes, we obtained 299 unique gene names, e.g., CMM. These gene names, along with their gene aliases (e.g. LOC385488), were used as the query list to obtain gene-paper association data. We looked for all instances of the gene names and aliases in the titles, abstracts, MeSH (Medical Subject Headings) terms, and substance lists from the 53,804 Medline paper records. A total of 107 of the 299 genes were found in the Medline data. Thus, 192 of the genes retrieved from the Entrez Gene database had no mention in the set of Medline records. The distribution of gene name occurrences by Medline field is shown in Table 1. Abstracts were by far the richest source of matches for the gene names. Titles were next richest source of matches, although there were more unique genes in the substance list than in the titles. In fact, if the matching query had been restricted to just abstracts and substances, no genes would have been missed. The output of this process was a list of the unique gene-paper association pairs.

Table 1. Gene-paper relational data.

Source	# Genes found	# Occurrences	# Genes unique to source
Titles	66	704	0
Abstracts	97	2374	25
MeSH terms	4	154	0
Substances	40	578	9

Third, the Universal Protein Resource UniProt[3], a central repository for protein information, was used to obtain information on the proteins associated with melanoma. This database combines the Swiss-Prot (115,000 entries), TrEMBL (700,000 entries), and PIR databases (283,000 entries), providing access to a comprehensive list of proteins.

Query for the term 'melanoma' resulted in 566 hits. The unique protein names were identified from the list, resulting in 367 proteins. In order to get the best possible matches, we modified the proteins names to the form they would be referred to in the Medline fields. Two levels of modification were done to the protein names before matching:

- Specific match – for example, '60S ribosomal protein L23a' was shortened to 'L23a'
- Broad match – in order to get a partial match between, for example, the proteins 'Melanocortin 1 receptor variant C315R' and 'Melanocortin 1 receptor variant F45L', each name was modified to 'Melanocortin 1' to get a match based on the functionality of the protein.

Case specificity was not maintained in this study for either genes or proteins; thus "Bcl2" and "BCL2" were considered to be the same gene. As with the genes, we looked for all instances of the proteins in the titles, abstracts, MeSH terms, and substance lists from the 53,804 Medline paper records. A total of 121 of the 367 proteins were found in the Medline data. The distribution of gene name occurrences by field is shown in Table 2. For protein-paper association pairs, although titles, MeSH terms, and substances add a significant number of mentions, they only add 4 unique proteins to the set found in the abstracts. Yet, the new association pairs are helpful in that they expand the number of papers connected to the proteins, and thus enrich the data set.

Table 2. Protein-paper relational data.

Source	# Proteins found	# Occurrences	# Proteins unique to source
Titles	92	2648	3
Abstracts	116	7722	22
MeSH terms	22	2988	0
Substances	52	2268	1

2.2. Calculation of similarities

Similarities between the various records – papers, genes, and proteins – were calculated in three parts. First, given that the papers dominate the map, and thus form the backbone for the entire data set, paper-paper similarities were calculated. This was done using a cosine similarity based on co-occurrence of MeSH terms as

$$SIM_{p1,p2} = \frac{M_{p1,p2}}{\sqrt{M_{p1} M_{p2}}}$$

where M_{p1} is the number of MeSH terms for paper *p1* and $M_{p1,p2}$ is the number of co-occurring MeSH terms for papers *p1* and *p2*. Of the 32,319 unique MeSH terms occurring 2 or more times, 36 common, non-specific terms were removed prior to calculating the similarity values (see Table 3). Many of the removed terms are Medline "check tags." In the future we will consider

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Search&DB=gene
[3] http://www.pir/uniprot.org/index.shtml

Table 3. MeSH terms removed from similarity calculation.

Rank	Term	# occur		Rank	Term	# occur
1	Human	44161		20	Aged, 80 And Over	2791
2	Female	21073		21	Time Factors	2491
3	Male	19598		23	Child	2444
4	Support, Non-U.S. Gov't	15420		26	Follow-Up Studies	2018
5	Middle Aged	14466		27	Molecular Sequence Data	1878
6	Animals	13760		28	Cells, Cultured	1771
7	Adult	13185		31	Retrospective Studies	1632
8	Aged	11455		32	Immunohistochemistry	1597
9	Mice	9858		35	Mice, Nude	1372
10	Support, U.S. Gov't, P.H.S.	8127		37	Amino Acid Sequence	1261
11	Tumor Cells, Cultured	5986		38	Survival Rate	1260
12	English Abstract	4673		39	Child, Preschool	1247
13	Comparative Study	4531		40	Base Sequence	1219
14	Adolescent	3707		41	Support, U.S. Gov't, Non-P.H.S.	1191
15	Prognosis	3617		45	Mice, Inbred Balb C	1098
16	Cell Line	3387		46	Rats	1092
18	Diagnosis, Differential	3050		50	Treatment Outcome	1015
19	Mice, Inbred C57Bl	2846		61	Infant	854

removing all check tags. Also, our experience with many data types and many data sets indicates that use of the full similarity matrix is not necessary. Rather, use of the top few similarities per record is sufficient to characterize the map. Thus, in this case, after calculation of the full paper-paper similarity matrix, only the top 15 similarities per paper were used.

Gene-gene, protein-protein, and gene-protein similarities were calculated from the lists of gene/protein-paper pairs as

$$SIM_{g1,g2} = \frac{P_{g1,g2}}{\sqrt{P_{g1} P_{g2}}}$$

where P_{g1} is the number of papers referring to gene/protein $g1$ and $P_{g1,g2}$ is the number of papers referring to both $g1$ and $g2$. Similarities between gene/proteins and papers were calculated as

$$SIM_{p1,g1} = \frac{1}{\sqrt{P_{g1} G_{p1}}}$$

where G_{p1} is the number of genes/proteins referred to in paper $p1$. The value of "one" in the numerator for this similarity is appropriate in a co-occurrence sense given that each gene/protein – paper combination occurs only once. In addition, for genes and proteins mentioned in many papers, we did not want them to unduly influence the positions of the papers, but rather to be placed within the literature map in appropriate positions.

The three sets of similarities were combined in one file (simple concatenation since there were no duplicate node pairs between files) and used in a single layout calculation. The resulting map is discussed in section 3.2.

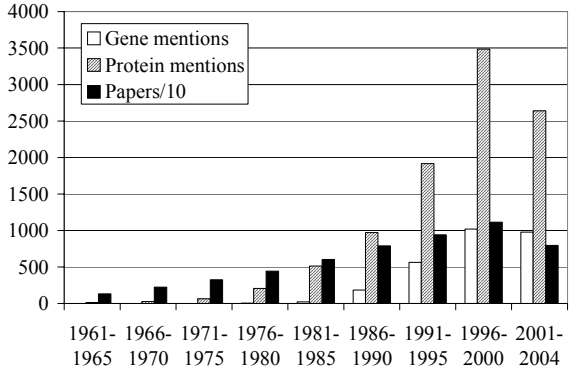

Figure 1. Numbers of melanoma-related papers, gene-mentions, and protein-mentions by time period.

3. Data Analysis and Visualization

3.1. Temporal and burst analysis

The overall history of the magnitude of melanoma research is shown in Figure 1. Here the numbers of papers, and numbers of mentions of genes and proteins are given for different time periods. While growth in overall melanoma research, as indicated by the number of papers, has been increasing at a relatively steady rate over the 40-year time period, the growth in gene and protein mentions has been dramatic. Protein work started before the gene work, and has been more prominent than the

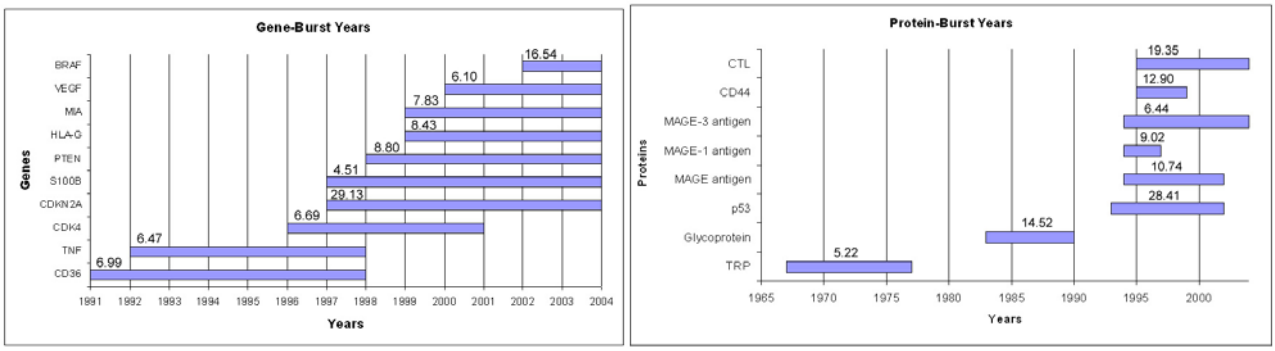

Figure 2. Gene and protein names and burst time intervals.

gene work. But the protein work no longer seems to be growing as fast as the gene-related work.

In addition to the simple temporal analysis, Kleinberg's burst detection algorithm [3] was applied to identify sudden interests in research on certain genes or proteins. The algorithm analyzes streams of time-sorted records (here publications) to find features that have high intensity over finite/limited durations of time periods. Rather than using raw frequencies of the occurrences of words, the algorithm employs a probabilistic automaton whose states correspond to the frequencies of individual words. State transitions correspond to points in time around which the frequency of the word changes significantly.

As a result, the algorithm generates a ranked list of the most significant word bursts in the stream, together with the intervals of time in which they occurred. This can serve as a means of identifying topics or concepts that rose to prominence over the course of the stream, were discussed actively for a period of time, and then faded away.

For the analysis, the complete set of 54,016 papers was used and the burst analysis was applied over titles and MeSH terms. Altogether 6,041 bursty words were identified. From these words we selected those that matched gene or protein names from the Entrez and Uniport databases. A total of 10 genes and eight proteins bursted. The time intervals in which they bursted are depicted in Figure 2.

The gene burst diagram reveals two categories of genes: melanoma-specific genes and proteins, and genes or proteins that were explored as a possible treatment for melanoma owing to their success in treating some other cancer, as follows:

Specific genes	MIA, S100B, CDKN2A, CDK4, VEGF, BRAF
Non-specific genes	HLA-G, PTEN, TNF, CD36
Specific proteins	TRF
Non-specific proteins	CD44, P53

3.2. Paper-gene-protein map

Using the similarity file described in section 2.2, a map of papers, genes, and proteins was generated. Layout of the data was done using VxOrd [5], a proprietary algorithm [6] that uses force-directed placement, a density field for repulsion, boundary jumping, and edge cutting. Our experience with many studies, both published [7-10] and unpublished, is that VxOrd preserves both global and local structure for large graphs (>10,000 nodes) from many different data types. The resulting layout is shown in Figure 3.

To the best of our knowledge, this is the first map which combines the three different element types, papers, genes, and proteins, and that can show five different types of networks: paper-gene, paper-protein, gene-gene, protein-protein, and gene-protein.

Figure 3 shows the main research areas covered by melanoma research over the last 40 years. Gray dots represent publications. Genes and proteins are given in blue and red respectively. Labeling was done by hand after exploration of the map using VxInsight [8]. The different research areas can be grouped into two main categories: 1) applied medical sciences (left side) and 2) basic molecular sciences (right side). Applied medical science work occurs at the organism level and rarely involves the study of molecular entities. In the basic molecular science studies, more research is carried out for genes and proteins. Interestingly, papers in the applied science portions of the map are less numerous than their molecular science counterparts.

Given the three different node types (papers, genes, and proteins) and their diverse associations, a number of association networks can be mapped within a common context. Figure 4 shows the gene-paper and gene-gene networks. Links in the left image indicate that a gene (white dot) was mentioned in a paper (grey dot). Links between two genes in the right image indicate that they have been co-mentioned in the same (one or more) paper and may possibly provide information on gene interactions related to melanoma. The gene 'CMM' in the

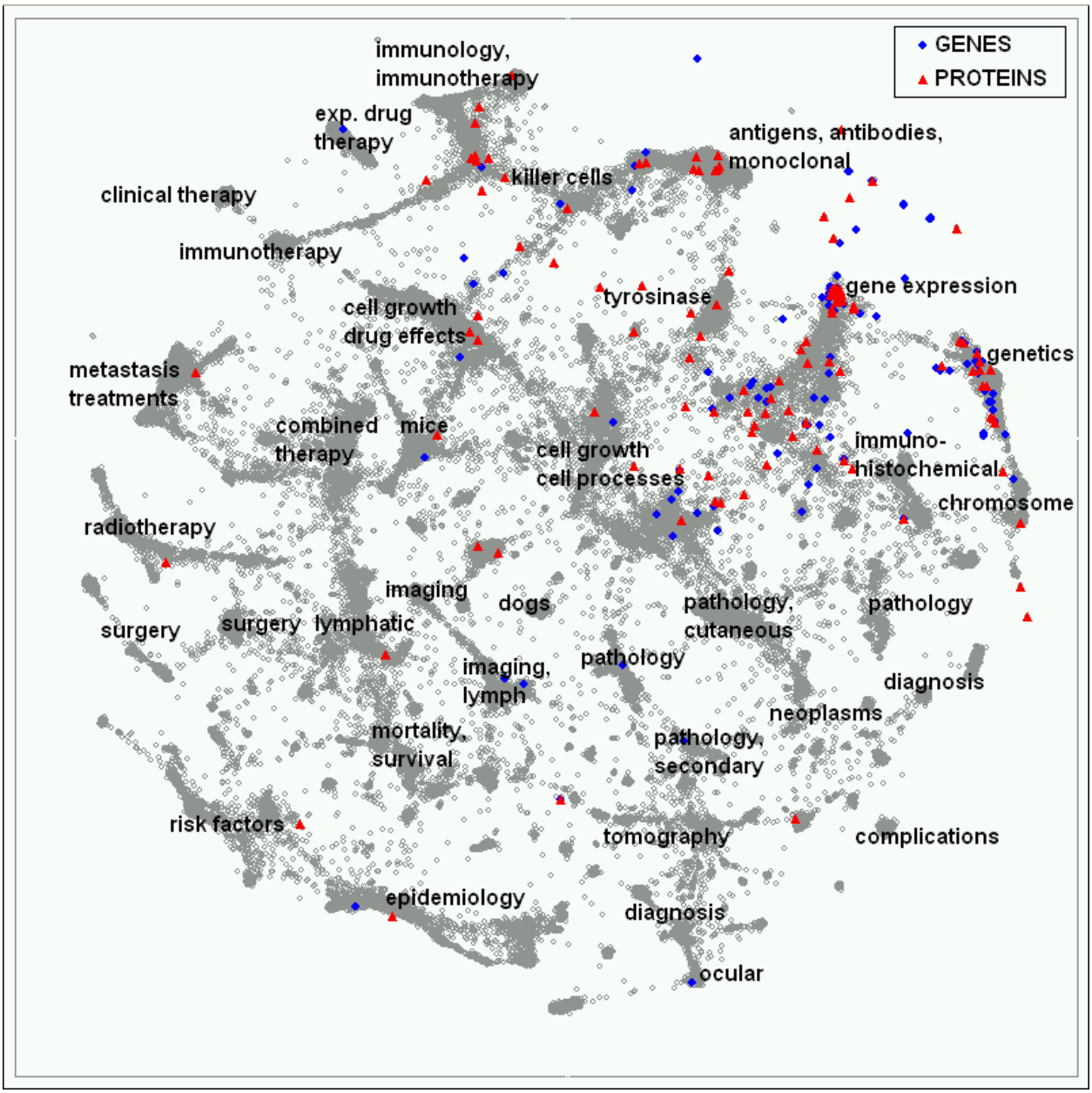

Figure 3. Melanoma paper-gene-protein map.

cancer/incidence region is one of the most often mentioned genes. In Figure 4 (left) its links to papers in the network are shown as red arrows.

The gene-gene network in Figure 4 (right) is obviously much smaller than the gene-paper network, and is focused in the region previously identified with molecular sciences. Paper-protein, protein-protein, or gene-protein networks could be viewed and explored in a similar fashion.

4. Validation by Experts

In order to validate how well the results and visualizations match human judgment we consulted five biologists. The informal validation sessions lasted about 30-40 minutes. Experts were shown four sets of information on a computer screen: 1) Figure 2, 2) Figure 3, 3) four decade-long time slices from the paper-gene-protein map, 4) the five network association maps (including the two in Figure 4). In general, the experts

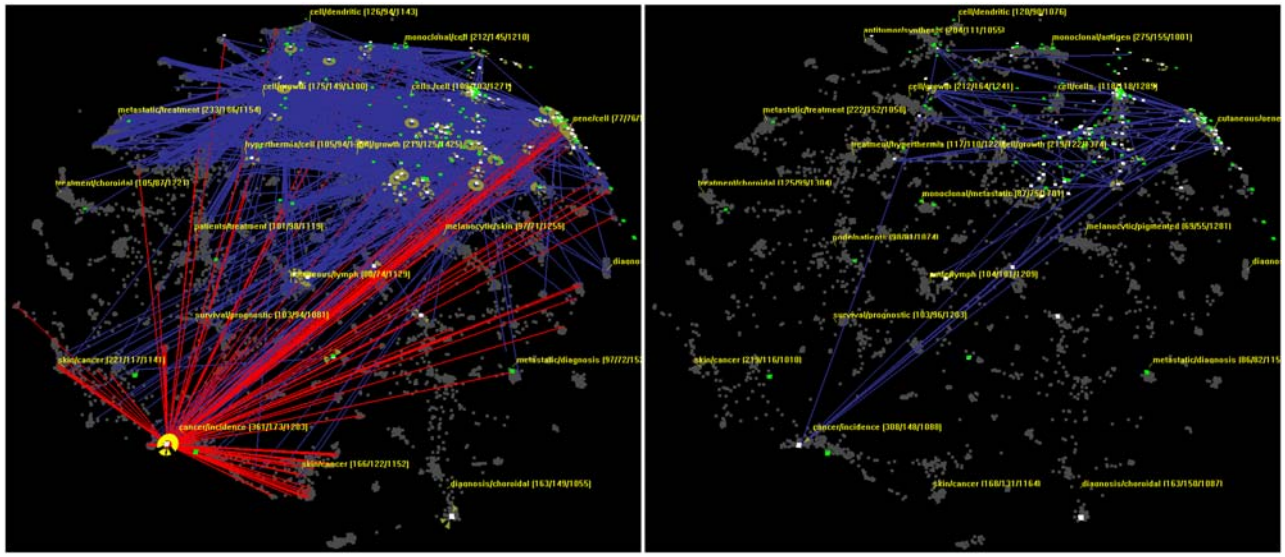

Figure 4. Gene-paper (left) and gene-gene (right) networks overlaid on the melanoma paper-gene-protein map.

remarked that this is a very good way to look at the domain.

4.1. Burst Analysis

The experts were shown Figure 2 and told that the 40-year data set was analyzed for sudden increases in the usage of gene and protein names. The experts classified the genes and proteins into the specific and general genes shown in section 3.1. The success rate of a particular gene for curing a disease leads researchers to experiment with other pre-known genes. The experts felt that the set of genes that bursted appeared to be more related to melanoma than the protein set.

4.2. Paper-Gene-Protein Map

Each expert was shown Figure 3, along with an explanation of the data and process used to generate the map. They were told that the papers close to a certain area label are related to this type of research and that genes and proteins were placed close to the papers in which they are mentioned. Subsequently, they were asked to interpret this map.

Three experts segregated the data into two regions – applied science vs. basic science. One of the experts mentioned that the majority of research has been done in basic science and that there needs to be a shift towards applied science where we see only scattered research efforts today.

The fifth expert classified the map into two categories – organism level vs. molecular level. By organism level this expert referred to treatments within the applied science domain. By molecular level the expert meant the basic science level dealing with genes and proteins.

4.3. Time Series Analysis

In order to understand the structure and dynamics of melanoma research and its evolution over the last 40 years, a series of time slices from the paper-gene-protein map was shown to the experts. The time slices were visualized in VxInsight and captured for use by the experts, and covered the decades 1964-1973, 1974-1983, 1984-1993, and 1994-2003.[4] The experts were asked to compare the time-series plots while thinking aloud. The subsequent description of the evolution of melanoma research was compiled from the explanations of these maps by these five experts.

For the first time slice, 1964-1973, all experts except one identified Chemotherapy as an emerging area for the possible treatment of cancer. One expert pointed out the dominance of diagnostic and immunity based approaches.

In the following decade, 1974-1983, chemotherapy gained immense popularity and seemed a most viable treatment for cancer. In addition, cell research was conducted in parallel to understand the cause of the disease. A shift towards molecular technology led to the development of methods for tagging cancerous cells using antigens. This in turn increased the number of 'monoclonal studies' – one of the major research areas – as an alternative to chemotherapy. In contrast to chemotherapy, monoclonal treatments remove only cancerous cells, but retain healthy cells.

[4] Figures for the time slices were too large for the paper, but are available upon request from the authors.

The subsequent decade, 1984-1993, shows a high percentage of research on 'metastasis' behavior of cancer. While former studies focused on cancer that is affecting a particular region of the body, many studies in the time period also aimed to find out how to stop cancer that is spreading throughout the body. As lymph nodes are indicators for early cancer detection during the metastasis phase of the cancer, the number of research papers in this area increased as well. Due to the human genome project initiative, chromosomal research became more widespread.

In the most recent decade, 1994-2003, interest in mapping the human genome and the availability of sequence data boosted gene-expression and mutation studies. Research on lymph nodes increased further. In sum, the time series nicely shows the different evolutionary stages of melanoma research.

4.4. Association Data

Experts were told what associations among papers, genes and proteins are shown in the five maps. The map shown in Figure 4 (left) was used to explain the connections of one gene, here the CMM gene, to papers.

Experts found the high number of connections to the 'CMM' gene surprising and mentioned that this gene might be mentioned in many papers that address demographic issues which is not the case for most of the other genes.

They noticed the high density of the protein network and explained it with the fact that researchers had a head start in proteins studies as compared to gene studies. Given that proteins are the functional units of cells that are primarily responsible for cell interactions they are more attractive for study.

5. Discussion and Future Directions

This paper presented a very first attempt to map a 'network ecology', namely, the interrelations among papers, genes and proteins in order to answer the questions stated in the introduction.

We believe that a global view of how many different research results are interconnected, what areas are currently unknown to mankind, and a means to quickly filter out relevant material are essential to biology today.

Systems such as Arrowsmith that supports the discovery of links between two literatures within Medline [11], or literature based methods for identifying gene-disease relations [12] are first steps in the right direction. However, in order to gain a big picture view more than one or two entities (genes, proteins, diseases, papers) need to be correlated and managed.

In future work we plan to further evaluate the accuracy of gene and protein placement, investigate changes in the similarity measure that will increase gene and protein placement accuracy, and collaborate with melanoma specialists to look more closely at the inferred gene-gene, protein-protein, and gene-protein networks with regards to their explanatory and predictive power.

In addition, we plan to map major institutions in geographic space based on zip code information. This will help identify the (changing) research focus of institutions and the importance of geographical space for collaboration and information diffusion.

Acknowledgements

We would like to thank Kranthi Varala, Stuart Young, Anne Prieto, Richard Repasky and Susanne Ragg for their expert input during the evaluation of the data mining and visualization results. This work is supported by a National Science Foundation CAREER Grant under IIS-0238261 and NSF grant DUE-0333623.

References

[1] R. Mack and M. Hehenberger, "Text-based knowledge discovery: Search and mining of life-sciences documents," *Drug Discovery Today*, vol. 7, 2002, pp. S89-S98.

[2] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA: Morgan Kaufmann, 1999.

[3] J. M. Kleinberg, "Bursty and heirarchical structure in streams," presented at 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2002, pp. 91-101.

[4] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," in *Annual Review of Information Science & Technology*, vol. 37, B. Cronin, Ed. Medford, NJ: Information Today, Inc./ASIST, 2003, pp. 179-255.

[5] G. S. Davidson, B. N. Wylie, and K. W. Boyack, "Cluster stability and the use of noise in interpretation of clustering," presented at 7th IEEE Symposium on Information Visualization, San Diego, CA, 2001, pp. 23-30.

[6] B. N. Wylie, "Method using a density field for locating related items for data mining." United States Patent 6,424,965: Sandia Corporation, 2002.

[7] K. W. Boyack, B. N. Wylie, G. S. Davidson, and D. K. Johnson, "Analysis of patent databases using VxInsight.," presented at New Paradigms in Information Visualization and Manipulation '00, McLean, VA, 2000

[8] K. W. Boyack, B. N. Wylie, and G. S. Davidson, "Domain visualization using VxInsight for science and technology management.," *Journal of the American Society for Information Science and Technology*, vol. 53, 2002, pp. 764-774.

[9] S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie, and G. S. Davidson, "A gene expression map for Caenorhabditis elegans," *Science*, vol. 293, 2001, pp. 2087-2092.

[10] M. Werner-Washburne, B. Wylie, K. Boyack, E. Fuge, J. Galbraith, J. Weber, and G. Davidson, "Comparative analysis of multiple genome-scale data sets," *Genome Research*, vol. 12, 2002, pp. 1564-1573.

[11] D. R. Swanson, N. R. Smalheiser, and A. Bookstein, "Information discovery from complementary literatures: Categorizing viruses as potential weapons," *Journal of the American Society for Information Science and Technology*, vol. 52, 2001, pp. 797-812.
[12] L. A. Adamic, D. Wilkinson, B. A. Huberman, and E. Adar, "A literature based method for identifying gene-disease connections," presented at IEEE Computer Society Bioinformatics Conference, 2002, pp. 109-117.