A Literature Based Method for Identifying Gene-Disease Connections

Lada A. Adamic, Dennis Wilkinson, Bernardo A. Huberman, Eytan Adar
HP Laboratories, Palo Alto, CA
{ladamic, dennisw, huberman, eytan}@hpl.hp.com

Abstract

We present a statistical method that can swiftly identify, from the literature, sets of genes known to be associated with given diseases. It offers a comprehensive way to treat alias symbols, a statistical method for computing the relevance of the gene to the query, and a novel way to disambiguate gene symbols from other abbreviations. The method is illustrated by finding genes related to breast cancer.

1. Introduction

The number of medical articles accessible electronically is growing at an unprecedented rate, a fact that makes it impossible for individuals to keep up with the pace of available information. At the same time, it presents a unique opportunity to gather information automatically, summarizing vast quantities of data in a statistical fashion and allowing for the discovery of new and relevant connections among pieces of dispersed information. Such information can then be used to test new hypotheses or aid a human in editing summaries of available results [1].

In this paper we present an identification and summarization method that relies on statistical techniques to extract gene sets relevant to a query such as a particular disease. It offers a comprehensive way to treat alias symbols, a statistical method for computing the relevance of a gene or a group of genes to the query, and a novel way to disambiguate gene symbols from other abbreviations.

Natural language processing (NLP) methods have previously been used to automatically extract gene names and gene and protein interactions from text [2-12]. For example, rule-based systems use part-of speech information and keywords to discover and tag names. These techniques can be computationally intensive and unsuitable to apply to large datasets such as the complete set of Medline abstracts. Moreover, they frequently seek to find relationships between a small, restricted set of genes and/or are applied to a few, select articles. Our work uses statistical methods to find the relevant set among all known human genes and identify the correlations between them quickly from the entire Medline database. This set of genes could in turn be used by a NLP algorithm or one using templates [13] to identify the nature of the interactions.

Methods for extracting keywords related to diseases [14] have a tendency to sometimes return overly general terms, and do not focus on genes in particular. The use of the "term frequency, inverse document frequency" (TFIDF) metric [13,14] requires a further thresholding step, since the method can rank very infrequent terms quite highly. In contrast, our statistical method captures only highly relevant genes, ones that only occur frequently in a specific context.

Previous work on extraction of gene mentions from Medline [15] has displayed high error rates due to abbreviations or words being mistaken for gene symbols. Instead, by focusing on statistically relevant symbols and further disambiguating their meanings, we are able to significantly reduce the error rate while also identifying sets of genes relevant to a particular condition.

In what follows we present the statistical basis of the method and illustrate its applicability to the Medline database. We then discuss a powerful disambiguation technique that allows for the identification of genes associated with given diseases. We present results that exhibit the power and simplicity of the method. In addition we compare our method with existing data and outline further extensions.

2. A Statistical Basis for Gene Relevance to Diseases

Since the information needed to explain the relevance of a gene to a disease is often present in the articles of the biological literature, detailed syntactic analysis of every article could in principle yield all, or almost all, of the information needed to explain the connection between a set of genes and diseases. However, such analysis is not only computationally prohibitive but also error prone when using natural language techniques.

Our method, on the other hand, relies on the statistical analysis of gene-disease occurrences in the biomedical literature, rather than exhaustive analyses of given articles.

Consider a set *S* of *N* articles from the biomedical literature and assume that a subset *s* mentions a particular disease, i.e. leukemia. A particular gene, *A* which has no correlation with leukemia, should occur in the same proportion in *s* as it does in all of *S*. Therefore, the probability that with no correlations one would find a co-occurrence of the gene *A* and the word leukemia would be simply *s/S*. It follows from probability theory that the probability of *n* co-occurrences can be simply computed from the Binomial distribution, as well as the expected number of occurrences and the associated variance.

A number of co-occurrences much greater than that computed from the binomial distribution would certainly indicate a strong correlation, which is what one seeks. Moreover, this metric can be used to roughly quantify the strength of the correlations of different genes based on known, published results. A gene whose relationship to a disease has only recently been documented would hopefully show as relevant to that disease, but perhaps not score as highly as a gene whose relationship is established and well documented.

We extended this statistical method to determine the relevance of a pair of genes to a disease. A pair that occurs much more frequently in the context of a disease is likely to be relevant to the disease. One can further examine to what extent the two genes are complementary. That is, when one gene is mentioned in an article related to a disease, another gene is likely to be mentioned as well, indicating that a strong link exists between the two genes. Of course one is not limited to studying only single genes or gene pairs. If a sufficient number of articles mention a larger number of genes, the same method can be applied to any size set of genes.

3. The Method

The outline of the method is shown in Figure 1. As shown, we first gathered gene symbols, both official and alias (see below), listed by 3 online gene databases: HUGO (Human Genome Organization) [16], OMIM (Online Mendelian Inheritance in Man) [17], and LocusLink (an online database of gene loci) [18]. We obtained titles and abstracts published between 1960 and 2001 from over 11 million Medline records.

We performed an automated search of the abstract and title of the Medline records to produce a "PMID/gene list" which for each document, identified by a unique PMID (PubMed Identifier) number, lists the different gene symbols that occurred in the abstract or title. In the current system we did not search for the full name of each gene, only its symbol, and we did not count how many times a particular symbol occurred in each article, just whether it occurred.

Additionally, we recorded whether each article's abstract or title contained a word or words pertaining to a particular disease or gene expression pathway. For example, if we were focusing on leukemia, we searched for the words "leukemia" or "leukaemia".

3. 1 Alias Symbols

Because the identification and categorization of human genes on a large scale is at an early stage, the nomenclature system is still haphazard. In some cases, a particular symbol is used in different contexts to denote 10 or more distinct genes; conversely, a particular gene may be represented by many symbols. Indeed, as of April 2002, HUGO alone lists some 13,800 redundant gene symbols.

Obviously, this unfortunate situation makes it more difficult to identify individual gene relevance to given diseases and to detect important gene pairs. To mitigate the confusion, we made use of the "official" symbol for each gene, as designated by HUGO, OMIM or Locuslink. These three databases have chosen one symbol per gene to be official, while the other symbols used to refer to the

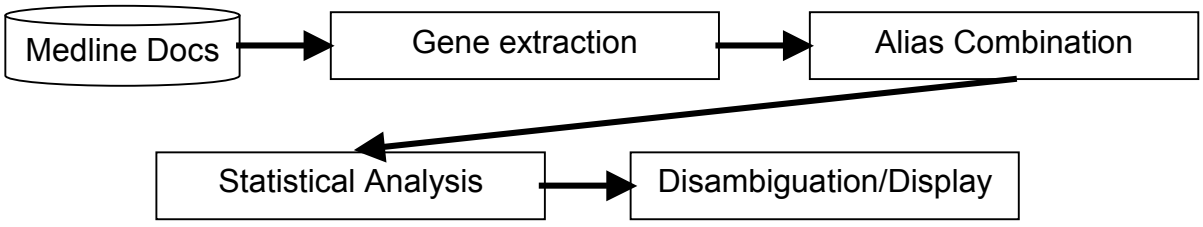

Figure 1. A general flow chart of the method. Medline documents are analyzed for gene symbols, the aliases are combined, and statistical relevance is determined given different disease contexts. Finally, a disambiguation and display step visualizes the results.

same gene are known as "aliases". We attempted, where possible, to replace any mention of an alias symbol with the official symbol, using the following scheme.

First, we counted all occurrences of gene names (official symbols and aliases) within the entire article set and within a focus subset (e.g. those article which mention leukemia). For each alias occurrence, we added to the count of both the alias and the official gene or genes it represented. For example, if the symbol OS, an alias for MID1, occurred in 49 articles, while MID1 occurred in 3, we would have a count of 52 for MID1. We kept track of the fact that 49 of the counts for MID1 originated with OS to be able to relate back to the articles and to modify the document gene lists as described below. Because OS frequently stands for "overall survival", it is important to keep track of its contribution to MID1's counts, as MID1 could otherwise erroneously be related to a disease.

We then modified our PMID/gene lists for the entire set and the focus subset to account for alias symbols. For each alias symbol, there were four possibilities:

1. The alias symbol represented only one official symbol, and the official symbol appeared independently (that is, its count was greater than its alias' count). For this case, we replaced all mentions of the alias in question in our PMID/ gene lists with the official symbol.
2. The alias symbol represented more than one official symbol, but only one of these official symbols occurred independently within the set or subset. Here we replaced the alias symbols with the official symbols which had counts.

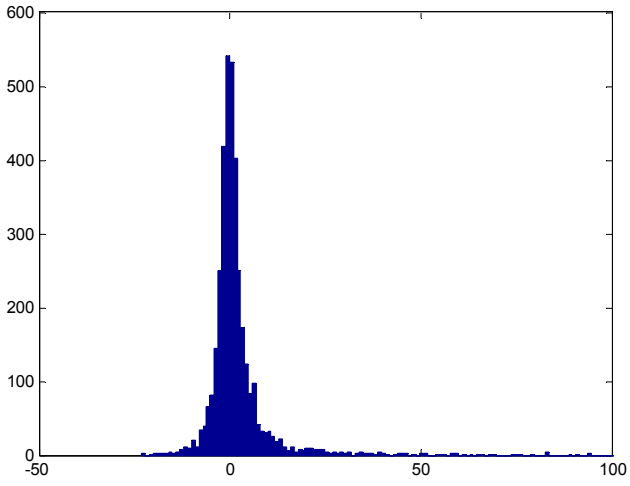

Figure 2. Distribution of correlation strengths between leukemia and various genes mentioned with leukemia in articles.

3. The alias symbol represented one or more official symbols, but none of these official symbol ever occurred independently within the set or subset. Clearly in this case the official symbol was not yet widely accepted by researchers, so it was more reasonable to refer only to the alias symbol.
4. The alias symbol represented more than one official symbol, and at least two of these had independent occurrences within the subset. In this case we could not decide, without syntactic analysis of the abstract or title text, which official symbol was meant by the paper authors, and so we kept the alias symbol.

Because we count only one occurrence of a gene per abstract, an alias and an official symbol occurring in the same abstract contribute together only a single count. In all cases, we kept the information about where the counts originally came from and indicated this in our results. For example, let's say our results implicate an obscure official symbol, which almost always appeared as the well-used alias symbol, in some disease. The original counts would show the user that 95% of the time that the gene was mentioned in connection with the disease, it was mentioned as the alias and not as the newer official symbol, hopefully mitigating possible confusion.

3.2 Measuring the Relevance of Genes to Diseases

Using the counts obtained after adding contributions from alias symbols, we compared the frequency of occurrence of a gene name in the set of all Medline articles (S_0) to the frequency with which the gene occurred in the focus subset (S_1) of articles which mentioned the disease. As stated earlier, this provides a measure of the relevance of the gene to the disease.

Focusing on "acute myeloid leukemia" (AML), consider for example the gene RUNX1, which our measure shows to be most tied to AML. The official HUGO symbol RUNX1 stands for "runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene). Aliases for the gene RUNX1 include AML1, CBFA2, AMLCR1, and PEBP2AB.

Putting for the moment aside that the name of the symbol RUNX1 itself reveals its strong connection to AML, we can establish the connection in a purely statistical fashion. The symbol RUNX1 occurs in 480 of the 20,909 articles mentioning AML and containing a gene symbol, and 590 times in the 3 million articles containing gene symbols. Next we compute how unlikely it would be to see the

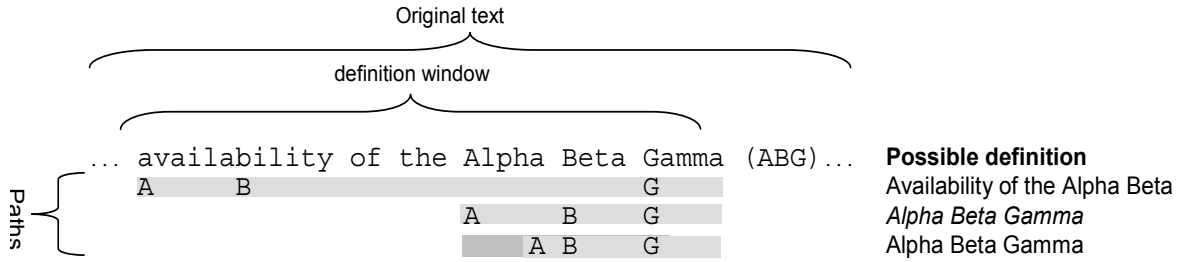

Figure 3: To extract the "definition" for an acronym or abbreviation in the original text, a window is constructed in which the definition may be found. The characters of the acronym, ABG, are aligned to the text. Three of these paths are illustrated in the figure. Each path represents a different definition, which is scored independently. The most likely candidate, *Alpha Beta Gamma*, is retained as the definition.

number of gene mentions in S_L, given how frequently the gene is mentioned overall.

The expected number of mentions of RUNX1 (assuming it is uncorrelated to leukemia) in S_L is given by the Binomial distribution $E[n_{RUNX1}] = N_L * p_{RUNX1}$. p_{RUNX1} = 0.0003 is the fraction of documents in all of Medline mentioning AML1, and N_L is the number of documents in S_L. The standard deviation is given by $\sigma(n_{RUNX1}) = \sqrt{N_L * (1 - p_{RUNX1}) * p_{RUNX1}}$. We measure the strength of the relationship (c_{RUNX1}) between RUNX1 and leukemia by measuring how much the observed number of RUNX1 documents deviates from the expected number had the draw been random.

$$c_{RUNX1} = \frac{n_{RUNX1} - E[n_{RUNX1}]}{\sigma(n_{RUNX1})} \qquad (1)$$

We find that c_{RUNX1} = 195.68, a very high value. We have used the normal approximation to the binomial distribution, valid in the case of large N. Using the normal distribution we can also find that the probability that 480 or more RUNX1 documents are found among a random draw of 20,909 documents is less than 10^{-16}. Our finding is consistent with the fact that AML1 is one of the most common targets of chromosomal translocations implicated in acute myeloid leukemia [19].

Most genes, however, show little or negative correlation with leukemia as demonstrated in Figure 2, which shows the values of c_{AML} for all genes which occur in S_L. The figure lacks those genes which occur in the database, but do not occur in S_L at all. They would populate the negative correlation side of the figure.

3.3 Relevance of Gene Pairs

We can further explore the relevance of gene pairs as opposed to individual genes. If a gene pair occurs more frequently in S_L than in the entire document collection, then the pair is considered relevant to S_L. For example, we find that the CFBF-MYH11 pair occurs 30 times with AML, and 37 times overall, giving the pair a relevance score of 48.83 to AML.

One can also compute to what extent the two genes are complementary, that is if they predominantly act together with regard to the disease. We compare the number of times each gene occurs in S_L separately to the number of occurrences together.

Let p_A (p_B) be the fraction of documents with gene A(B) in S_L. Then if A and B are uncorrelated, the probability of finding them together is $p_{AB}=p_A*p_B$. From here on, we proceed just as we did for the link of a single gene to leukemia. Consider again the genes CBFB and MYH11, which have an unusually high complementarity. CBFB occurs 40 times in S_L and MYH11 occurs 78 times, yet a full 30 of those occurrences are joint. The probability of this occurring is very small, and we obtain a complementarity score of 75.63 given by

$$c_{AB} = \frac{n_{AB} - E[n_{AB}]}{\sigma(n_{AB})} = \frac{n_{AB} - N_L * p_{AB}}{\sqrt{N_L * (1 - p_{AB}) * p_{AB}}} =$$

75.63

Searching though the literature we find why CBFB and MYH11 are complementary to such an extent: "In human acute myeloid leukemia samples with chromosome 16 inversion, a fusion gene *CBFB-MYH11* is created and expressed. This novel gene includes most of the CBFB

disease	c_G	$n_{DISEASE}$	n_{ALL}
colon cancer	33.30	83	1039
ACCEPT from definitions	30.1% had definitions., 96% matched	**match**: , deleted in colon cancer :15 (0.51), deleted in colorectal cancer :4 (0.78), deleted in colon carcinoma :2 (0.77), deleted colon cancer :1 (0.34), deleted colorectal carcinoma :1 (0.88), deletion :1 (0.24) **no match**: , dextran coated charcoal :1 (0.13)	
breast cancer	47.90	179	1039
REJECT from definitions	29.6% had definitions, 11% matched	**match**: , deleted in colon cancer :4 (0.51), deleted in colorectal cancer :2 (0.78) **no match**: , dextran coated charcoal :32 (0.13), dextran coated charcoal method :7 (0.12), dextran coated charcoal assay :2 (0.12), dextran coated charcoal technique :2 (0.11), dextrancoated charcoal :1 (0.13), dextrose coated charcoal :1 (0.09), dextran coated charcoal assays :1 (0.12), conventional radiochemical :1 (0.00)	

Table 1. Evaluation of the symbol DCC as a possible reference to the "deleted in colon cancer" gene for two diseases: breast and colon cancer. The number of occurrences and the matching score (0 to 1, low to high) is given after each extracted definition of the symbol.

gene, a hematopoietic transcription factor, and the last half of MYH11"[1].

3.4 Disambiguating Potential Gene Symbols

Once we identified symbols with high relevance to a disease, we still need to address the problem of polysemy, where gene symbols may be identical to abbreviations for other common terms. In fact, the most challenging aspect of extracting gene names from the text is determining which symbols actually correspond to genes. For example, *Emergency Room* and *Estrogen Receptor* are both represented by the acronym *ER* (an alias for the gene ESR1). Thus, to ensure that the statistical analysis is not weakened by non-gene symbol instances, it is necessary to take further steps to clean the data. We tackled this problem from two directions: using cues from the text that identify a particular gene, and when these cues were absent, calculating an overall likelihood that the symbol represents a gene.

Authors frequently offer cues about the meaning of a symbol when they list a definition followed by the symbol itself in parenthesis. We utilize these definitions to eliminate abbreviations that are possible gene aliases, but that are rarely used to refer to a gene.

To do so, we first determine different possibilities for acronym expansions. Once this information is in hand, the second step establishes which expansion relates to the gene. Below, we briefly describe the mechanism behind our approach. The full details of this process are beyond the scope of this paper, but are fully described in [20].

Several researchers [21-23] have attempted to tackle the problem of abbreviation and acronym definition from text. Acromed [24], the most successful of these methods, yields results with 98% precision but requires complex natural language processing. Our approach, which is much simpler to implement, generates comparable results (95%+ accuracy). The algorithm which is illustrated in Figure 3, constructs possible alignments (which we call *paths*) that potentially define an abbreviation within the text. The test for paths is done in a *definition window*, which contains several words preceding the first occurrence of the abbreviation. In the figure, we show three potential paths (although there are others in this example) that may define the abbreviation. Each path is scored by various rules in order to find the most likely path. One rule, for example, gives a higher score to paths in which the abbreviation letters occur at the start of words. (In Figure 3, the second path is then scored higher that the third.) Although simple and few, the rules generate the correct answer repeatedly.

Once we have a definition, it is possible to compare it to the known definition for the gene. Unfortunately, while we would like every correct definition for a gene such as

[1] http://www.umassmed.edu/pgfe/faculty/castilla.cfm

PR to be *progesterone receptor*, other definitions we have found include: *progesterone receptors* (plural), *progesterone* (no receptor), or *progestron receptor* (a spelling variant). To address this problem we apply a tri-gram based comparison to determine relationships between the "true" definition and others. In this approach, each definition is broken up into three letter chunks called tri-grams. Progesterone receptor, for example, is composed of the tri-grams *pro, rog, oge, ges*, etc. The similarity between the true definition and proposed definition is:

$$similarity = \frac{|| A \bigcap B ||+1}{\sqrt{|| A || +1} * \sqrt{|| B || +1}}$$

Where the numerator is the number of intersecting tri-grams between the true definition, *A*, and the proposed definition, *B*. The denominator is a normalization factor based on the number of tri-grams in both definitions. The resulting similarity value is then compared to a threshold (.2 in our case) to decide if the two definitions are sufficiently related.

Finally, in order to decide whether or not to include a symbol in the final statistics, we determine if there is enough evidence (enough definitions), and further compare the good (matching) to bad (non-matching) definitions. If this ratio is high enough, the gene symbol is accepted; otherwise the algorithm determines that the symbol is rarely used to represent a gene, and it is rejected or down-weighted in the overall statistics.

When an insufficient number of definitions is present, we calculate the likelihood that a symbol represents a gene by comparing the number of article titles and abstracts containing the symbol as well as words such as *gene, DNA, inhibit*, and *express*, to the total number of articles in which the symbol occurs. The higher this ratio, r_G, the greater the likelihood that any given instance of the symbol is a gene reference. While r_G is often an adequate indicator, using definitions when available yields much higher quality results relevant to the context. For example, the ratio r_G for the symbol DCC is only 0.46. This information alone does not allow us to judge with certainty whether the symbol DCC refers to a gene 'deleted in colon cancer' in any given article

Table 1 shows how definitions can be used to disambiguate the symbol DCC in two contexts, one of breast cancer and the other of colon cancer. Although the symbol occurs twice as often in documents dealing with breast cancer, our algorithm allows us to recognize that DCC in the context of colon cancer stands for the "deleted in colon cancer" gene, but most often stands for "dextran coated charcoal" in the breast cancer context. Dextran coated charcoal assay is the preferred method used to quantify the presence of estrogen and progesterone receptors in breast cancer tissue. This makes the symbol DCC highly relevant to breast cancer, but the gene DCC itself relates to breast cancer to a lesser extent. By analyzing the definitions accompanying the symbol, we were able to give opposite classifications (*accept* for the colon cancer context, and a *reject* for breast cancer) for DCC in two different contexts.

4. Results

Next we present results using the methods described in the above section, as applied to breast cancer. To summarize, we extract gene symbols from the literature, folding aliases into official symbols, and computing the relevance of each official symbol to breast cancer. We then eliminate non-gene symbols using contextual clues, such as whether the symbol has an overall likelihood to be representing a gene or whether its accompanying definitions match the official or alias gene names.

In order to evaluate our algorithm, we compared the selected genes by our algorithm to a human edited breast cancer gene database[2]. Of the 58 entries in the human edited database, 46 had a significant score ($s_G > 2$) which identified them as relevant to breast cancer. 3 genes, CTSD (cathepsin D), PLG (plasminogen) and COL18A1 (endostatin) were almost always mentioned in Medline text by their full names as opposed to symbols and hence were not selected by our algorithm. Future versions of the algorithm could of course include gene names as well as symbols, but this would require an additional disambiguation method. The remaining 9 genes had low or negative scores because there were too few articles supporting a connection, or because the symbol was obscured by a common acronym. Two of the gene symbols had fewer than 5 articles mentioning them in connection with breast cancer, with no further articles published in the past 10 years. Those connections might be quite weak.

On the other hand, our scoring method was able to produce a much more extensive list of genes connected to breast cancer. Table 2 shows the ten genes most relevant to breast cancer in order of relevance given by the function given in Equation 1[3]. We note that TFF1, CEA, MUC1 are scored as highly relevant, and play roles in diagnosis and treatment of breast cancers. A potential user

[2] http://tyrosine.biomedcomp.com

[3] a more complete list can be found at http://www.hpl.hp.com/shl/papers/genelit/index.html

of such information would be a human wanting to edit a list of known genes connected to breast cancer.

At the same time, we were able to eliminate those acronyms which while highly relevant to breast cancer, do not represent genes. Examples include FAC and CAF (5-fluorouracil, Adriamycin, cyclophosphamide chemotherapy), SLN (sentinel lymph node), OS (overall survival), ILC (invasive lobular carcinoma), TNM (tumor node metastasis). A list is shown in Table 3. In addition to filtering out non-gene symbols, we can also use the disambiguation techniques to resolve a symbol that is an alias for multiple genes. For example, using the definitions 'estrogen receptor' we were able to resolve ER to ESR1 (estrogen receptor 1) as opposed to EREG (epiregulin).

The quality of the output of the algorithm is heavily dependent on the completeness of the input. For example, the ki-67 antigen is associated with the gene MKI67. But having no listing for ki-67 from the HUGO, OMIM or LocusLink databases, we interpret the symbol KI as an alias for PSME3. Similarly, PSCP is an alias of BRCA1. Even though our database contained the alias, there was no definition present. When PSCP was mentioned, the definition "papillary serous carcinoma of the peritoneum" was tested and did not match the name of the official symbol BRCA1. In the case of PSCP, the error had no effect because the single occurrence of PSCP did not offset the strong statistical significance of the 1,300 BRCA1 mentions. These two sources of error will play a smaller role once more protein symbols and alias names are incorporated into databases. Another source of error is mistaking a cell line symbol such as MX1 for a gene symbol. Additional analysis of text and classification using mesh headings may eliminate false positives from cell lines.

official symbol and score	aliases	c_G	official or alias name	BC gene database
BRCA1:	PSCP	276.54	breast cancer 1, early onset	**x**
ESR1	ER	253.36	estrogen receptor	**x**
BRCA2		213.86	breast cancer 2, early onset	**x**
ERBB2	HER2, NEU, NGL	205.18	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2,neuro/glioblastoma derived oncogene homolog (avian)	**x**
PGR	PR	148.69	progesterone receptor	**x**
TFF1	PS2, PNR2, BCEI	106.91	trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)	
TP53	P53, TRP53	96.90	tumor protein p53 (Li-Fraumeni syndrome)	**x**
EGFR	ERBB, S7	88.29	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	**x**
CEACAM5	CEA	67.32	carcinoembryonic antigen-related cell adhesion molecule 5	
MUC1	PUM EMA PEM	58.90	mucin 1, transmembrane	

Table 2. Genes with the highest relevance score c_G to breast cancer. All aliases occurring more than once are listed, as well as the relevance score, the gene name, and whether the gene was listed in the human edited breast cancer gene database.

official symbol and score	symbol in text	official or alias name	most common meaning in text
FANCC, c =56.98:	FAC(152)	Fanconi anemia, complementation group C	fluorouracil doxorubicin and cyclophosphamide
SCYA27, c =56.89	ILC (88)	il11ra-locus chemokine	(infiltrating, invasive) lobular carcinoma
CMD1A, c =56.80	IDC (154)	cardiomyopathy, dilated 1A (autosomal dominant)	(infiltrating, invasive) ductal (carcinoma, cancer),
DCC, c =55.00	DCC (202)	deleted in colorectal carcinoma	dextran coated charcoal
PCAF, c = 53.34	CAF (161)	p300/CBP-associated factor	cyclophosphamide (adriamycin, doxorubicin) and 5 fluorouracil
ODZ1, c =52.56	TNM (349)	odz, odd Oz/ten-m homolog 1(Drosophila)	tumor node metastasis:7
MID1,, c =51.87,	OS (300)	midline 1 (Opitz/BBB syndrome)	overall survival
SLN, c = 48.71	SLN (132)	sarcolipin	sentinel lymph node
PRKWNK1 c=40.03	RFS (153)	protein kinase, lysine deficient 1	(relapse, recurrence) free survival
BCS1L, c = 36.28,	BCS (91)	BCS1-like (yeast)	breast conserv(ing,ation) surgery, breast cancer survivors

Table 3. Acronyms highly relevant to 'breast cancer' which were automatically filtered as not representing gene symbols. The score c_G, along with the number of times the symbol occurs in the breast cancer set (N_{BC}) and overall (N_{ALL}) are listed. The second column lists the most commonly occurring official or alias symbol. The shading of the second column corresponds to r_G, the overall likelihood that the symbol represents a gene. The last two columns list the gene name and the definitions found in text for comparison.

6. Conclusions and Future Work

The explosive growth in the number of medical publications and databases available electronically calls for novel ways of accessing, summarizing, and extracting knowledge from the data. In this spirit, we have developed an algorithm to automatically extract genes from text such as Medline titles and abstracts and determine their relevance to a particular topic. We presented a way to count alias symbols and to determine whether the frequency with which a particular gene is mentioned in a given context is statistically significant. We also presented a novel method for disambiguating gene symbols from acronyms. In our sample study of breast cancer genes, we found that the algorithm identified most breast cancer genes from a human edited database, as well as identifying many additional genes that have been tied to breast cancer in the literature. The algorithm was also able to discard acronyms which were relevant to breast cancer but did not represent gene symbols. Future work will further improve upon this method by incorporating full gene names and MESH headings. It would optionally use weights to favor more recent documents. This modification would emphasize recently discovered gene-disease connections as well as well-established ones that are being restated in the current literature.

For a demonstration of various aspects of our system, including the gene pair algorithm, see http://www.hpl.hp.com/shl/papers/genelit/index.html.

7. References

[1] M. V. Blasoklonny and A. B. Pardee, "Unearthing the gems," *Nature*, 416, 373.(2002)

[2] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing

system for the extraction of molecular pathways from journal articles," *Bioinformatics,* 2001 Jun;17 Suppl 1:S74-82.
[3] M. Andrade and P. Bork, "Automated extraction of information for molecular biology," *FEBS Lett.*, 476, 12 (2000).
[4] S.K. Ng and M. Wong, "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts," *Genome Informatics*, 1999 Dec 14-15;10:104-112.
[5] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using BLAST for identifying gene and protein names in journal articles," *Gene*. 2000 Dec 23;259(1-2):245-52.
[6] C. Blaschke, M. Andrade, C. Ouzounis and A. Valencia, "Automated extraction of biological information from scientific text: protein-protein interactions," *Intelligent Systems for Molecular Biology*, Heidelberg, 60 (1999).
[7] J. Thomas, D. Milward, C. Ouzounis, S. Pulman and M. Carroll, "Automated extraction of protein interactions from scientific abstracts," *Pac. Symp. Biocomput.*, 538 (2000).
[8] B. Stapley and G. Benoit, "Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts,". *Pac. Symp. Biocomput.*, 526 (2000).
[9] K. Humphreys, G. Demetriou and R. Gaizauskas. "Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures,". *Pac. Symp. Biocomput.*, 505 (2000).
[10] D. Proux, F. Rechenmann, L. Julliard, V. Pillet, and B. Jacq, "Detecting gene symbols and names in biological texts : a first step toward pertinent information extraction," *9th Workshop on Genome Informatics*, 72-80 (1998).
[11] T. C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter, "EDGAR: extraction of drugs, genes and relations from the biomedical literature," *Pac. Symp. Biocomput.*, 217 (2000).
[12] J. Pustejovsky, J. Castaño, J. Zhang, B. Cochran, and M. Kotecki, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," *Pac. Symp. Biocomput.*, 2002.
[13] M. Stephens, M. Palakal, S. Mukhopadhyay, and R. Raje. "Detecting Gene Relations from MEDLINE Abstracts,". *Pac. Symp. Biocomput.*, 2001.
[14] M. Andrade and A. Valencia, "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families,". *Bioinformatics* 14(7): 600-607, 1998.
[15] T.-K. Jenssen, A. Laegreid, J. Komorowski and E. Hovig. "A literature network of human genes for high-throughput analysis of gene expression," *Nat Genet*. 28(1): 21-28, 2001.
[16] H.M Wain,, M. Lush, F. Ducluzeau, and S. Povey, "Genew: The Human Nomenclature Database*," Nucleic Acids Research* 2002, Jan 1;30(1):169-71.
[17] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000.
[18] K.D. Pruitt, and D.R. Maglott, "RefSeq and LocusLink: NCBI gene-centered resources," *Nucleic Acid Research* 2001, Jan 1;29(1):137-40.
[19] J.L. Huret. AML1 (acute myeloid leukemia 1). Atlas Genet Cytogenet Oncol Haematol. December 1997 .URL : http://www.infobiogen.fr/services/chromcancer/Genes/AML1.html
[20] E. Adar, "Simple Abbreviation Definition, Clustering and Disambiguation," unpublished manuscript
[21] L. S. Larkey, et. al., "Acrophile: an automated acronym extractor and server," *Proceedings of the Fifth ACM Conference on Digital Libraries*, June 2 – 7, 2000, San Antonio, TX.
[22] J. Pustejovsky, et. al., "Extraction and Disambiguation of Acronym-Meaning Pairs in Medline," *10th World Congress on Health and Medical Informatics*, September 2-5, 2001, London.
[24] S. Yeates, et. al. "Using Compression to Identify Acronyms in Text," *Data Compression Conference 2000*, March 28 – 30, 2000, Snowbird, Utah.