Published in final edited form as:
Comput Biol Chem. 2006 December ; 30(6): 425–433.

Link Test—A Statistical Method for Finding Prostate Cancer Biomarkers

Xutao Deng[1], **Huimin Geng**[2], **Dhundy R. Bastola**[3], and **Hesham H. Ali**[1]

1 *College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA {xdeng, hali}@mail.unomaha.edu*

2 *Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE 68198, USA {huimingeng, dbastola}@unmc.edu*

3 *Department of Pediatrics, University of Nebraska Medical Center, Omaha, NE 68198, USA {huimingeng, dbastola}@unmc.edu*

Abstract

We present a new method, *link-test*, to select prostate cancer biomarkers from SELDI mass spectrometry and microarray data sets. Biomarkers selected by link-test are supported by data sets from both mRNA and protein levels, and therefore results in improved robustness. Link-test determines the level of significance of the association between a microarray marker and a specific mass spectrum marker by constructing background mass spectra distributions estimated by all human protein sequences in the SWISS-PROT database. The data set consist of both microarray and mass spectrometry data from prostate cancer patients and healthy controls. A list of statistically justified prostate cancer biomarkers is reported by link-test. Cross-validation results show high prediction accuracy using the identified biomarker panel. We also employ text mining approach with OMIM database to validate the cancer biomarkers. The study with link-test represents one of the first cross-platform studies of cancer biomarkers.

Keywords

Microarray; Mass spectrometry; Biomarker; Prostate Cancer; Text Mining

1. INTRODUCTION

Biomarkers usually refer to specific genes and their products which are biochemical features or facets that can be used to measure the progress of disease or the effects of treatment. Finding accurate biomarkers is a key to early diagnosis and successful treatment of many otherwise incurable diseases. A handful of established biomarkers such as prostate specific antigen (PSA) for prostate cancer and cancer antigen-125 (CA-125) for ovarian cancer are routinely used for disease monitoring. However, the relatively low specificity of those biomarkers makes them unsuitable for population cancer screening (Diamandis, 2004).

Microarray and mass spectrometry technologies have emerged to bring hopes for discovering biomarkers and building diagnosis models. Microarray and mass spectrometry technologies

Correspondence: Xutao Deng, xdeng@mail.unomaha.edu, Phone Number: 1-402-554-6004, Fax Number: 1-402-554-3284

have been commonly used in studying genome and proteome activities, respectively, and they serve as a pair of complementary tools. The large-scale, widely accessible nature made them extremely appealing for biomarker finding. For example, numerous studies have been performed using microarrays (Liu et al., 2005;Golub et al., 1999;Statnikov et al., 2005;Singh et al., 2002) or mass spectrometry (Lilien et al., 2003;Petricoin et al., 2002[a]2002,;Wagner et al., 2004;Liu and Li, 2005). These studies have reported more than 90% positive predictive value (PPV) when using mass spectrometry biomarkers as diagnosis indicators, and about 80% PPV when using microarrays. These exciting results show performance superior to current clinical biomarkers such as PSA for prostate cancer diagnosis. Although the biotechnology behind one is fundamentally different from that of the other, the strategies for biomarker finding and predictive model building using mass spectrometry and microarray are similar. They can be considered as a three-step data mining procedure:

1. Data generation and preprocessing: both healthy and ill patients' data are collected; the data are usually preprocessed by normalization, outlier detection, baseline correction (in mass spectrometry), etc.
2. Computational biomarker extraction: standard tools, such as analysis of variance (ANOVA), t-test, principal component analysis (PCA), and genetic algorithm (GA) can be used to select a small set of features; the features are genes or their protein products in microarrays, and are mass spectrum peaks in mass spectrometry.
3. Classification model building: standard classification tools such as support vector machine (SVM), decision trees (DT), k-nearest neighbors (kNN), etc., are routinely used to build predictive models based on selected biomarkers.

Mass spectrometry is considerably faster, cheaper, and more accessible than microarrays. (See introduction in Liebler, 2001;Siuzdak, 2003). As a result, it has received more attention lately, especially for clinical applications. From the data mining point of view, the feature-selection step (the second step), can be regarded as a preprocessing step for the classification step (the third step). As long as the classification accuracy achieved a high level, the biomarkers themselves are no longer important for practice. However, unlike microarray biomarkers, the mass spectrometry biomarkers are described only by their mass-to-charge ratio (*m/z*) values without further identification and annotation.

Our focus in this study is on the biomarker extraction step. The goal of biomarker extraction is to focus only on a small panel of important genes/proteins of a huge set of genes/proteins, or mass spectra from a highly mixed sample. This step is very important not only because it is the basis for building an effective predictive model but also because finding biomarkers could significantly enhance our understanding of the mechanism and treatment of diseases. However, there are technology limitations or computational artifacts in this, which have been extensively discussed in Diamandis (2004),Conrads et al. (2003), and Sorace et al. (2003). For example, several studies showed inconsistent set of biomarkers extracted for prostate cancer (Diamandis 2004;Sorace et al, 2003;Baggerly, 2004). Lacking confirmation of disease-specific biomarkers posed a huge problem in the clinical application of both mass spectrometry and microarray data (Lyons-Weiler, 2005;Pepe et al., 2001).

To get consistent and more reliable biomarkers, we want to cross-link microarray and mass spectrometry data, instead of using only one of them as in the previous studies. Our basic idea is to associate microarray and mass spectrometry biomarkers to cross-validate their existence by the evidence from each. We first extract biomarkers independently from each data set using existing feature-selection methods. This step results in two lists of biomarkers. One, for microarrays, is a list of genes or its protein products. The other, for mass spectrometry, contains mass spectrum peaks with associated *m/z* values. We have to differentiate the biomarkers in microarrays and mass spectrometry. In microarrays, a biomarker is a specific gene which may

have already been sequenced and annotated. In mass spectrometry, however, a biomarker is a mass spectrum peak which is usually not corresponding to an entire intact protein; rather, it is a set of possible peptides that happen to be of the same mass or within a region. Then we can use the gene/protein biomarker list to query against the mass spectrometry biomarker list (or vice versa) to construct the relationships between the two. Next, a fundamental question is, what is the level of significance of these associations between microarray markers (protein sequences) and mass spectrometry markers (mass peaks). In this paper, we develop a statistical test procedure to provide an answer. For convenience, we call this test *link-test*.

The paper is organized as follows. In the next section, we illustrate the overall design for biomarkers extraction. Each data preprocessing component is briefly described. Section 3, the core of the paper, is devoted to the link-test. In this section, we formalize the problem and provide an analytic solution for the link-test. We also show the results by using the link-test to associate microarray and mass spectrometry biomarkers of prostate cancer data. Text mining validation for the selected biomarkers are provided in section 4. In section 5, we conclude the paper with the findings and a brief discussion.

2. OVERALL STUDY DESIGN

The overall design of this study is illustrated in Figure 1. Microarray and mass spectrometry data are first processed independently, and candidate biomarkers are extracted for each type of data. Then both microarray markers and mass spectrum markers are associated by link-test. The goal for this step is to confirm the biomarkers from each source. Finally the confirmed biomarkers are used in building classifiers to predict new samples with observed mass spectra and microarray profiles.

2.1 Data Description

The detailed description of microarray data can be found in Singh et al. (2002), and the data set can be downloaded at http://www-genome.wi.mit.edu/MPR/prostate. This set of data contains high-quality gene-expression profiles obtained from 52 prostate tumor samples and 50 prostate non-tumor samples. It was collected using oligonucleotide microarrays containing probes for 12 600 human genes and Expressed Sequence Tags (ESTs). It is important to note that the mass spectrometry samples are from serum while the microarray samples are from prostate tissues.

The mass spectrometry data description can be found in Petricoin et al. (2002[b]), and the data set can be downloaded at http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. The data set contains 69 cancer samples (26 samples with PSA level 4–10 ng/mL, and 46 samples with PSA level greater than 10 ng/mL), and 63 normal samples with no evidence of cancer (PSA level less than 1 ng/mL). This set of data was collected using an H4 protein chip and a Ciphergen PBS1 SELDI-TOF (Surface-Enhanced Laser Desorption and Ionization Time of Flight) mass spectrometer. The spectra were exported with the baseline subtracted. The range of *m/z* values is from 0 to 20 000.

2.2 Mass Spectrum Peak Detection

The raw mass spectra for each sample are composed of 15 154 (x, y) pairs. x axis records m/z values with corresponding intensity on y axis. Therefore, we have 15 154 features for only 132 samples. Obviously the number of features is too large to build a reliable diagnosis model. Peak detection is the first step in reducing the number of features. Peaks are basically the features with local maximum intensities. Current peak detection jobs are usually done by the software bundled with a spectrometer so that the algorithms are hidden from users. The algorithm we use is a very simple one: we register all the m/z values with local maximum

intensity which exceeds user specified thresholds. We use both absolute threshold (intensity from baseline) and relative threshold (intensity from the left and right hill feet of the peak). Both thresholds are empirically set by a human annotator at 0.3.

2.3 Mass Spectra Peak Alignment

We applied the time-warping algorithm (Wang and Isenhour, 1987) to aligning the peaks extracted from each sample. The time-warping algorithm employs dynamic programming and is very similar to the global sequence alignment algorithm (Needleman and Wunsch, 1970). After peak detection and alignment, the mass spectra still contain 6 467 features (aligned peaks) with *m*/*z* value above 1 000. The number of features is further reduced to 5 709 by requiring that a peak must be observed in at least two samples to avoid noise peaks. Note that other spectra alignment algorithms are also good candidate for this task (Yu et al., 2006;Wong et al., 2005).

2.4 Biomarker (Gene, Mass Spectrum Peak) Extraction

In this step, we extract informative (differently expressed) biomarkers from the preprocessed data. The methods for extracting mass markers and gene markers are essentially the same. We use the t-statistic with permutation test (Golub et al., 1999). For each candidate gene or mass spectrum, we compute the t-statistic using the two group labels. Then we randomly permute the labels 10 000 times to see whether the t-statistic is significantly correlated with class labels. The level of significance α for an individual test is set at 0.0005. Note that the multiple statistical tests could result in many false biomarkers by chance. To overcome this problem, we use Bonferroni correction to adjust the significance level. This step yields 1 398 significant mass peaks (908 overexpressed and 490 underexpressed in cancer samples) and 436 genes (261 overexpressed and 175 underexpressed in cancer samples) as pre-biomarkers. Among the 436 genes, we identified 240 that have complete sequence information in NCBI Entrez Database (http://www.ncbi.nlm.nih.gov/Database/). The 240 genes and their description can be accessed at http://bioinformatics.ist.unomaha.edu/xdeng/cbacsuppl.txt.

3. Link-Test

Link-test sets out to detect those peptides that are NOT showing random behaviour as our potential biomarkers. These non-random mass spectrum peaks are more likely to be originated from those genes that are detected using microarrays, and therefore, link our observations between the two kinds of biomarkers. In this step, we construct the association between the microarray markers and mass spectrum markers. Ideally, if the mass markers are from whole intact proteins, we can simply compare the *m/z* values with the molecular weights of microarray markers derived from their sequences, and the match (also called hit, link) between the two should be able to confirm the existence of each. Unfortunately, this is not likely to be the case and studies showed that serum proteins are primarily composed of small protein segments (Adkins et al., 2002) The *m/z* values of all 151 peaks are less than 10 000 Da since the low-energy mass spectrometry data mainly consist of singly-charged ions. However, of all human protein sequences from the SWISS-PROT database (http://www.ebi.ac.uk/swissprot/,Bairoch et al., 2005), only 2.78% of proteins (348 out of 12 484) have a mass less than 10 000 Da. This suggests that most of the mass markers are fragmented peptides instead of intact proteins. In order to construct the associations between a query microarray marker and mass markers, we consider the match between all possible fragments (peptides) of a given protein and all possible mass peaks. However, we found that this resulted in many hits, especially when the mass matching tolerance is large, e.g. ±1 Da. Now the question is how to determine which hits between microarray markers and mass markers are statistically significant, and which are not.

3.1 Null Hypothesis

For a given microarray marker *P*, by computing all of its possible peptides (subsequences) and using them to query the mass markers, we observed that certain peptides match to a certain mass marker *m*. We could construct the following test:

Null Hypothesis H_0—The match between a protein *P* and a peak *m* is purely random. In other words, the chance of *P* matching to *m* is equal to the chances that other proteins match to *m*.

Alternative Hypothesis H_1—The match between a protein *P* and a peak *m* is NOT random. In other words, the chance of *P* matching to *m* is NOT equal to the chances that other proteins match to *m*.

If we find a microarray marker *P* or its derived peptides have a molecular weight equal to mass marker *m*, the link-test is to determine whether this match is likely due to chance (H_0) or significantly unlikely due to chance (H_1). This link-test is weaker than testing whether the peak *m* is from protein *P* or not. Nevertheless, the link-test is mathematically manageable while the latter test can be justified only by experimental study.

The first step towards the test is to estimate the parameters $\theta(m)$, the probability of a mass biomarker with mass *m* generated by a random peptide under null hypothesis H_0. To properly scale $\theta(m)$, we also require that the given peptide could generate a peak with mass *m*. See below for explanation.

We use all human protein sequences from SWISS-PROT to estimate this parameter. Assume we have *R* protein sequences in the database. The length of each sequence is denoted as n_1, $n_2,\ldots, n_R$. For a peak with mass *m*, the length of peptides that could generate this peak falls between L_1 and L_2, which are defined as:

$$L_1 = \left\lceil \frac{m}{186.07932} \right\rceil, \; L_2 = \left\lfloor \frac{m}{57.02147} \right\rfloor \tag{1}$$

where the two constants are the monoisotopic masses for the largest (Tryptophan) and smallest (Glycine) amino acid residue, respectively. The total number of peptides that could generate peak *m* in the database should be

$$N(m) = \sum_{l=L_1}^{L_2} \sum_{i=1}^{R} (n_i - l + 1). \tag{2}$$

Among $N(m)$ peptides, the number of peptides that have exact mass *m*, denoted by $E(m)$, can be computed from all protein sequences in SWISS-PROT. Using the maximum likelihood principle, we have the estimator for $\theta(m)$ as in Eq. (3)

$$\theta(m) = \frac{E(m)}{N(m)}. \tag{3}$$

In fact, each *m* is associated with an accuracy threshold δ (a small interval such as 1Da).

The parameter θ can also be generalized to deal with a mass interval $m_\delta = [m-\delta, m+\delta]$,

$$\theta(m_\delta) = \frac{E(m_\delta)}{N(m_\delta)} \tag{4}$$

where $N(m_\delta)$ and $E(m_\delta)$ are the total number of peptides and the exact number of peptides which may produce the mass in the interval $[m-\delta, m+\delta]$, respectively.

Figure 2 shows the frequency $E(m_\delta)$ and $-\ln \tilde{\theta}(m_\delta)$ estimated from SWISS-PROT. For visual reasons, we show only a segment of molecular weight (represented on x axis) for each graph. We can see that not all mass markers have an equal chance to be hit under the purely random hypothesis H_0. In other words, the association between a peptide and a peak m can be differentiated. This observation is the foundation of our link-test. Figure 2a shows the distribution of $E(m_\delta)$ when $\delta = 0.01$. We observed interesting periodic patterns. The same pattern is reflected in Figure 2b and 2c, which are the distributions of $-\ln \tilde{\theta}(m_\delta)$ with $\delta = 0.01$ and $\delta = 1$ respectively. Comparing Figure 2b and 2c, we can see that the values of $\tilde{\theta}(m_\delta)$ are greatly impacted by the setting of δ. It is understandable that when δ increases, $-\ln\tilde{\theta}(m_\delta)$ decreases considerably, since $E(m_\delta)$ increases greatly as δ increases. There are main theme trend line and periodical pattern showed in Figure 3c. In the main theme, $-\ln\tilde{\theta}(m_\delta)$ increases with the weight m, which suggests that the larger the m value of a mass marker, the lower the chance to be hit. Also notice in Figure 2c that the amplitude of the periodic wave decreases as m increases, which suggests that the larger the m values of mass makers, the more difficult they are to be discriminated from each other.

3.2 P-values

Having the values for parameter θ, we can build a test for the original null hypothesis H_0. Recall $\theta(m_\delta)$ is the conditional probability of any random peptide that happens to have a mass at the interval $[m-\delta, m+\delta]$. This is equivalent to viewing $\theta(m_\delta)$ as the probability of success for a Bernoulli trial in testing whether a peptide could happen to be within the mass interval.

Given a microarray marker P with the length n_P, the total number of its possible peptides that could generate a mass m_δ is

$$N_P(m_\delta) = \sum_{l=L_1}^{L_2} (n_P - l + 1) \tag{5}$$

where L_1 and L_2 are calculated as in Eq. (1). We can see that the link-test for a pair of biomarkers (P, m_δ) can be viewed as a binomial test with the probability of success $\theta(m_\delta)$ and the number of trials $N_P(m_\delta)$. The test statistic for the pair (P, m_δ) is to test the probability that the protein P finds a match at peak $m\delta$, and the P-value for this test can be expressed as:

$$\begin{aligned} \text{P-value} &= p(P \text{ links to } m_\delta \text{ under null hypothesis}) \\ &= p(\text{Protein } P \text{ with length } n_p \text{ produce at least one peptide that is within } m_\delta) \\ &= 1 - \binom{N_P(m_\delta)}{0} \theta(m_\delta)^0 (1 - \theta(m_\delta))^{N_P(m_\delta)} \\ &= 1 - (1 - \theta(m_\delta))^{N_P(m_\delta)} \end{aligned} \tag{6}$$

The P-value increases when $\theta(m_\delta)$ increases and $N_P(m_\delta)$ increases. Intuitively, if the P-value is extremely small, we could say that the observed link between the two biomarkers is unlikely to be random so that there may exist certain relationships between the two biomarkers in the link. The distribution of P-values of the statistic test that protein P links to mass interval m_δ is graphed in Figure 3. In Figure 3a, the P-values are plotted as a function of the input microarray markers and mass markers (δ =0.01 Da). Certain regions on the curve are more significant than others. In order to see the pattern clearly, we fix one variable and look at the P-values' changing with the other variable. In Figure 3b, when the protein length is fixed at 2 500 residues, the distribution of P-value shows a dwindling wave as the mass increases. In Figure 3c, the P-value

decreases when protein length increases. We also plotted the curve for δ =1 Da (the plot is not shown here); basically, every point on the curve is near zero. To determine the level of significance α of a link, care must be taken to adjust for the effect of multiple tests. Suppose we have a total number of K mass markers extracted from mass spectrometry data. For a given protein, there will be K possible link-tests between the protein and mass markers. Again, Bonferroni correction can be used here; $\alpha_{individual} = 1 - (1 - \alpha_{overall})^{1/K}$, where $\alpha_{overall}$ is the user-specified overall significance level.

In conclusion, the procedure of link-test is summarized as follows:

Input: a list of microarray markers and a list of mass markers

For each microarray marker P:

1. Generate all possible peptides and compare those peptides with all mass markers within the tolerance level δ.
2. For each matched peak m_δ, refer to Figure 3 to get the P-value p(P links to m_δ).
3. If the P-value is less than significance level $\alpha_{overall}$, output P and m_δ.

Output: a list of biomarker pairs (microarray markers, mass markers) which have been confirmed by link-test.

3.3 Biomarker Results

We use C++ to implement all the components in Figure 1. From the protein and mass pre-biomarkers identified by preprocessing and feature selection steps, link-test identified 18 pairs (13 unique microarray markers matched with 16 unique mass) of biomarkers between proteins and masses, when $-\ln\alpha_{individual} = 5$. This result is illustrated in Table 1. From the 13 proteins, five are ribosomal proteins. CDH12 are calcium-dependent cell-cell adhesion molecules that may be involved in the metastasis and invasion of cancer. KLK2 is a close family member of PSA (also named KLK3) which is a well established prostate cancer indicator.

To test the classification accuracy, we use SVM^{lignt} (Joachims, 1999) with a linear kernel as the classifier. We applied the 16 unique mass markers to train SVM with 5-fold cross-validation on prostate cancer samples. The classification accuracy we obtained is 85.3 ± 1.9, which is comparable to the original report (Petricoin et al., 2002[b]). We need to point out that, a wrapper (based on searching) is used in the work of Petricoin et al. (2002[b]), and the strategy is to search for biomarkers that maximize certain classifiers. However, our approach, instead of maximizing classifiers, select biomarkers from multiple data sources (microarray and mass spectrometry), and therefore is classifier-independent and less likely to be cryptic.

4. Validation with Text Mining of OMIM

Our objective is to identify prostate-cancer-related genes from OMIM (Online Mendelian Inheritance in Man, 2000) records and use them as evidence to confirm previously identified prostate-cancer-related genes in Table 1. *Named Entity Tagging* (NET) is a popular text mining approach which searches through all OMIM records for the terms related to prostate cancer and returns those records containing the terms as candidate prostate-cancer-related genes (de Bruin and Martin, 2002). The NET approach depends on the established human annotations of OMIM and therefore limits its use in finding potential links between phenotypes and genes.

To find potential links between genes and phenotypes, we start with the NET approach to obtain a list of gene records containing the search terms as the *seed records*, and then we construct a *record graph*, which is a graph theory representation of all OMIM records. Each OMIM record

is modeled as a graph node. For any pair of records A and B, whenever one mentions the other, there is an undirected, unweighted edge between them. Otherwise, there is no edge between them. Having the graph constructed, then we use the seed records to search the record graph and find the *minimum distance* (or minimum number of edges) of each record node, from the seed record nodes. The minimum distances are interpreted as the *degree* of association between each OMIM record and the phenotype of prostate cancer. The strategy is flow-charted in Figure 4.

We use an existing program CGMIM (Bajdik et al., 2005) to perform NET on OMIM. A list of synonyms for prostate cancer types are displayed below:

prostate| prostatic|	cancer, carcinoma, leukaemia, leukemia, lymphoma, malignancy, melanoma, myeloma, neoplasm, tumor, tumour

Of all of 17 251 OMIM records, 167 records are identified by CGMIM as seed records. By constructing the record graph, we get 15 056 nodes (records) in a major subgraph and 2 195 nodes disconnected from the major subgraph. The degrees of all OMIM records show a typical exponential distribution with average degree of 6.98 (Figure 5).

The 167 seed records are denoted as a set S with cardinality represented as $|S|$. The set of all records in OMIM is denoted as R. We apply Dijkstra's algorithm (Dijkstra, 1959;Cormen et al., 2001) to search the shortest distance from every seed record $s \in S$ to every record node t $\in R$ in the record graph. The shortest distance from s to t is denoted as $D(s, t)$. Then we calculate the minimum distance $Min\_D(t)$ for each record node t, from all seed records $s \in S$, as follows

$$Min\_D(t) = \min_{s \in S} (D(s, t)), \quad t \in R \tag{7}$$

$Min\_D(.)$ is a natural metric to describe the association of specific genes to prostate cancer with greater distance values impling lower association. By definition, the $Min\_D(.)$ values of the seed records equal 0.

The distribution of $Min\_D(.)$ is shown in Figures 6 and the $Min\_D(.)$ values for the 13 candidate biomarkers are shown in Table 1. Seven of the genes either lack OMIM entry or are not in the major subgraph. Among the remaining six genes in the major subgraph, KLK3 is a seed record with $Min\_D(.)$ equal 0; SLC39A6 has $Min\_D(.)$ value of 1; and the other four have $Min\_D(.)$ values of 2. According to Figure 6, Binomial test $Bin(6, 0.37)$ shows that finding the six biomarkers with $Min\_D(.)$ less than or equal to 2 is marginally significant (P-value = 0.062). From Figure 6, we observe that the minimum distance $Min\_D(.)$ of all records is approximately normally distributed with mean 2.29 and standard deviation 0.84. The sample average $Min\_D(.)$ for the six candidate biomarkers is 1.50. Two-tailed Z-test shows that 1.50 is significantly less than the population mean 2.29 (P-value = 0.021). Both statistical tests suggest that the biomarkers extracted using the link-test method are supported by OMIM text mining.

5. Conclusions

We develop a new method for extracting biomarkers from combined microarray and mass spectrometry data sets. The core of this study is the development of a statistical test procedure for detecting the level of significance between a specific microarray marker detected by microarray and a specific mass peak presented in the mass spectrum of a mixture of serum protein fragments. Our method builds relationships between the biomarkers at both transcriptomic and proteomic levels which help cross-validating the biomarkers. The identified biomarker panel performs well in terms of prediction accuracy and it is also supported by text

mining results. This study is among the first attempts for cross-platform cancer biomarker analysis.

Mass spectrometry intensities are not a reliable measurement of protein concentration, so the models for extracting biomarkers from mass spectrometry data sets are not fully quantitative. This may partially explain the inconsistent biomarkers found in the literature (Diamandis, 2004;Sorace and Zhan, 2003;Baggerly *et al*., 2004). A better way to find consistent and reliable biomarkers, instead of using mass spectrometry technology alone, is to use microarrays (more quantitative) to find gene (protein) biomarkers first, and then use them to pull out the confirmed mass markers.

The choice of peak threshold is a common problem in many mass spectrometry-based analysis. An overly large threshold may cause too many signal peaks be lost. On the other hand, too many noise peaks will be included when the threshold is set too low. Therefore, an appropriate peak threshold may impact the sensitivity and specificity of our method. This threshold is set by an experienced mass spectrometry annotator at this stage of research. Other parameter choices in the pre-processing step have to be carefully examined to ensure that reasonable biomarkers could be identified in the link test.

Besides the cleavages of proteins in serum, the mature expressed proteins undergo many post-translational modifications. These post-translational modifications could impact the links between the mass peaks and the genes. Our method could be enhanced by incorporating post-translational modification information from the SWISS-PROT database. We are also interested in expanding link- test to the problem of peptide mass fingerprinting in which multiple peptides are matched to multiple mass peaks.

Acknowledgements

This work was supported by the NIH grant number P20 RR16469 from the IMBRE program of the National Center for Research Resources.

References

Adkins JN, Varnum SM, Auberry KJ, Moore RJ, Angell NH, Smith RD, Springer DL, Pounds JG. Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. Mol Cell Proteomics 2002;12:947–55. [PubMed: 12543931]

Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI mass spectrometry patterns in serum: comparing proteomic data sets from different experiments. Bioinformatics 2004;20(5):777–785. [PubMed: 14751995]

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:D154–D159. [PubMed: 15608167]

Bajdik CD, Kuo B, Rusaw S, Jones S, Brooks-Wilson A. CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC Bioinformatics 2005;6(1):78. [PubMed: 15796777]

Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. Introduction to Algorithms. 2. MIT Press and McGraw-Hill; 2001.

Conrads TP, Zhou M, Petricoin EF 3rd, Liotta L, Veenstra TD. Cancer diagnosis using proteomic patterns. Expert Rev Mol Diagn 2003;3(4):411–420. [PubMed: 12877381]

de Bruin B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Medical Informatics 2002;67:7–18.

Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. Mol Cell Proteomics 2004;3.4:367–378. [PubMed: 14990683]

Dijkstra EW. A Note On Two Problems In Connexion With Graphs. Numerische Mathematik 1959;1:S 269–271.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–537. [PubMed: 10521349]

Joachims, T. Making large-Scale SVM Learning Practical. In: Schölkopf, B.; Burges, C.; Smola, A., editors. Advances in Kernel Methods - Support Vector Learning. MIT-Press; 1999.

Liebler, DC. Introduction to Proteomics: Tools for the New Biology. Humana Press; New Jersey: 2001.

Lilien RH, Farid H, Donald BR. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. J of Comp Bio 2003;10(6):925–946.

Liu J, Li M. Finding cancer biomarkers from mass spectrometry data by decision lists. J Comp Bio 2005;12(7):971–979.

Liu JJ, Cutler G, Li W, Pan Z, Peng S, Hoey T, Chen L, Ling XB. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 2005;21(11):2691–2697. [PubMed: 15814557]

Lyons-Weiler J. Standards of Excellence and Open Questions in Cancer Biomarker Research: An Informatics Perspective. Cancer Informatics 2005;1(1):1–7.

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48(3):443–453. [PubMed: 5420325]

Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). 2000. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93(14):1054–61. [PubMed: 11459866]

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002a;359(9306):572–577. [PubMed: 11867112]

Petricoin EF, Ornstein DK, Paweletz CP, et al. Serum proteomic patterns for the detection of prostate cancer. J Natl Cancer Inst 2002b;94(20):1576–1578. [PubMed: 12381711]

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203–209. [PubMed: 12086878]

Siuzdak, G. The Expanding Role of Mass Spectrometry in Biotechnology. MCC Press; San Diego: 2003.

Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4:24. [PubMed: 12795817]

Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21(5):631–643. [PubMed: 15374862]

Wagner M, Naik DN, Pothen A, Kasukurti S, Devineni RR, Adam BL, Semmes OJ, Wright GL. Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics 2004;5:26. [PubMed: 15113409]

Wang CP, Isenhour TL. Time-warping algorithm applied to chromatographic peak matching gas chromatography/Fourier transform infrared/mass spectrometry. Anal Chem 1987;59:649–654.

Wong JW, Cagney G, Cartwright HM. SpecAlign--processing and alignment of mass spectra datasets. Bioinformatics 2005;21(9):2088–90. [PubMed: 15691857]

Yu W, Wu B, Lin N, Stone K, Williams K, Zhao H. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. Comput Bio Chem 2006;30(1):27–38. [PubMed: 16298163]

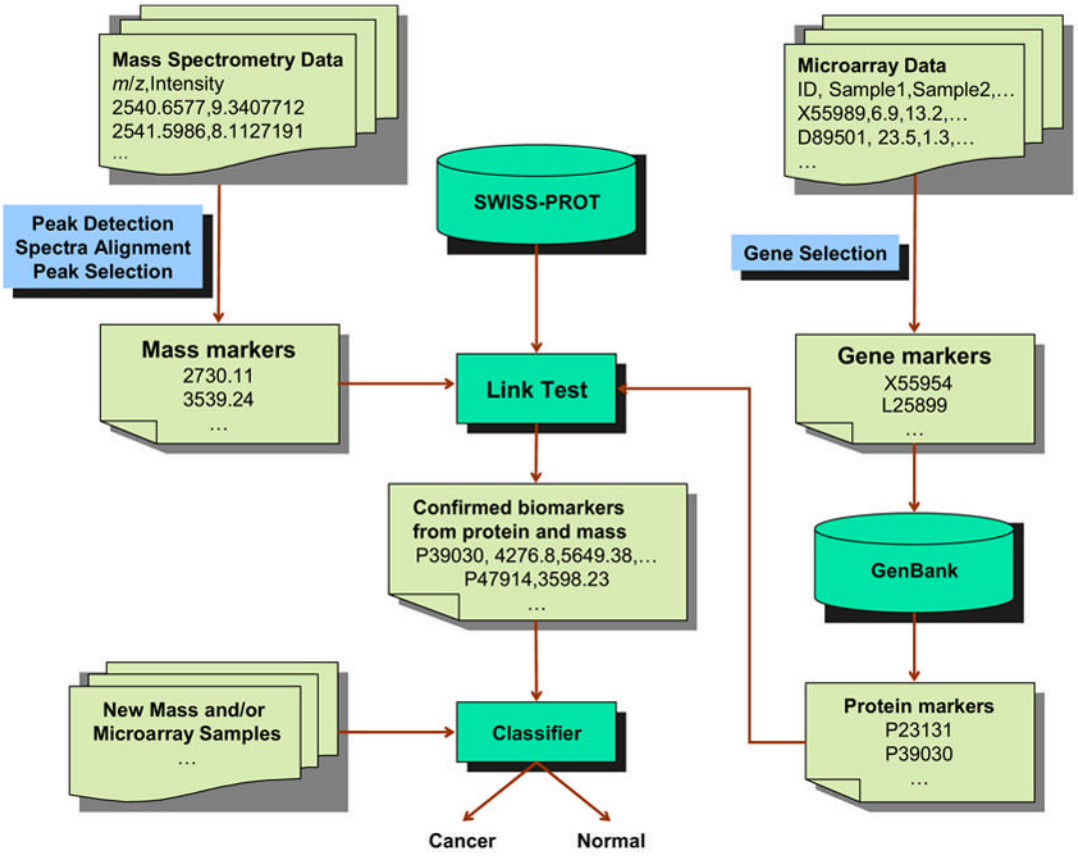

Figure 1.

Flowchart of biomarkers extraction and their application in disease prognosis. Microarray and mass spectrometry data are first pre-processed independently, and differentially expressed candidate biomarkers are extracted for each type of data. Then link tests were applied to the microarray markers and mass spectrum markers to identify significant biomarkers for building a classifier. Unknown samples can then be classified using the trained classifier.

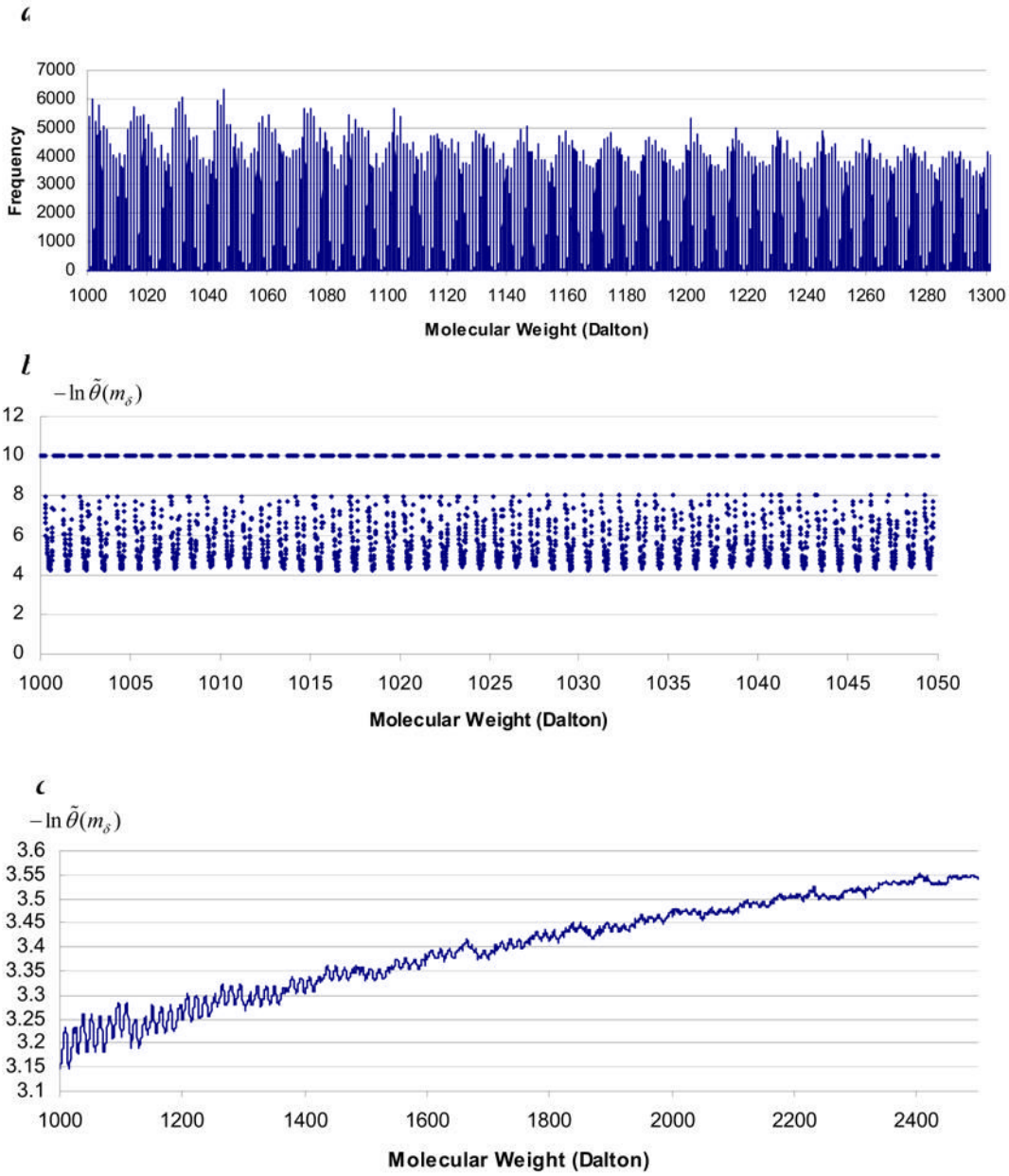

Figure 2.

The distribution of $\theta(m_\delta)$ and frequency $E(m_\delta)$. ***a***. A segment of $E(m_\delta)$ distribution shows periodic behavior ($\delta = 0.01$). ***b***. Periodic distribution of $\theta(m_\delta)$ ($\delta = 0.01$), where the value 10 on the *y* axis denotes infinity, $+\infty$. ***c***. Periodic distribution of $\theta(m_\delta)$ ($\delta = 1$), where the trend line of $-\ln\theta$ (m_δ) increases as the molecular weight *m* increases.

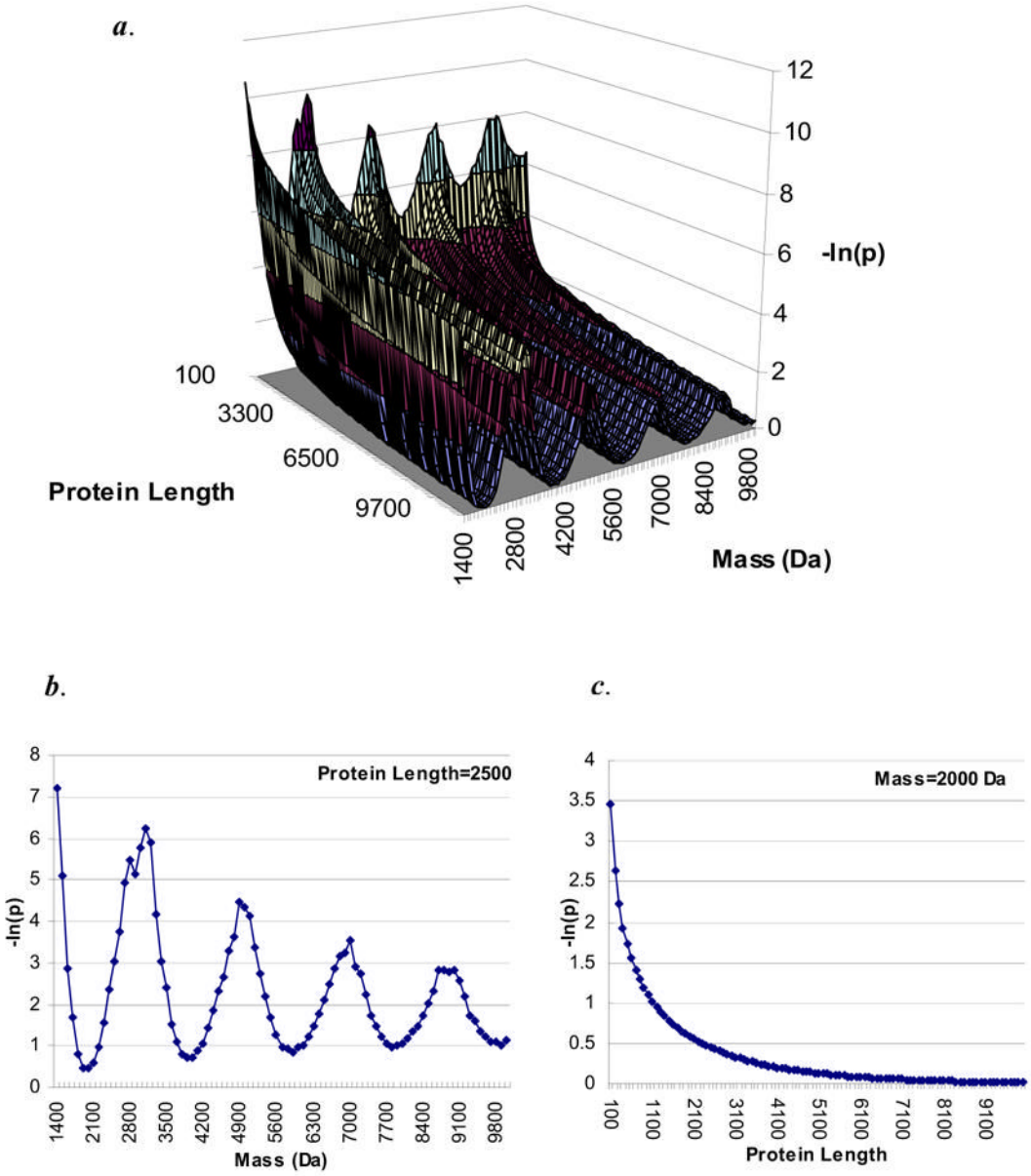

Figure 3.

The distribution of P-value, $\delta = 0.01$. ***a***. P-value depends on the length of protein and mass marker. ***b***. The distribution of P-value when the protein length is fixed at 2 500 residues. ***c***. P-value decreases with protein length with fixed mass (mass = 2 000 Da).

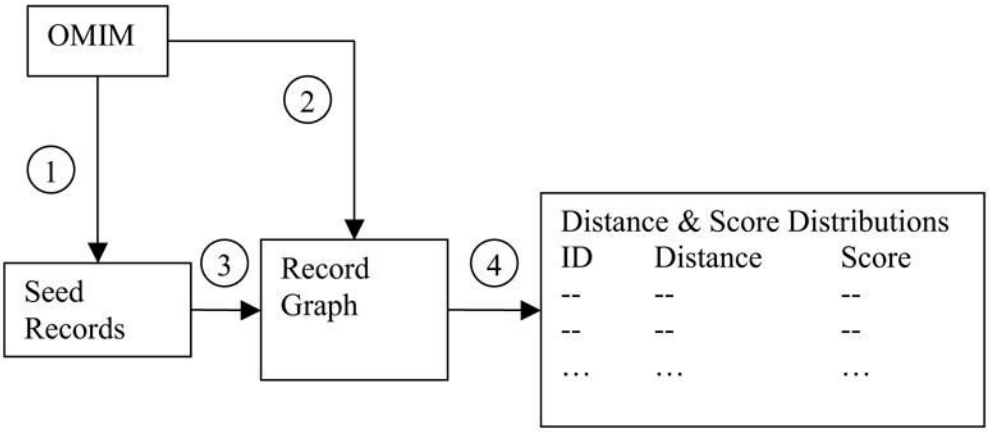

Figure 4.

The flow chart of text mining OMIM records for finding prostate cancer genes. 1. Search seed records using NET by identifying the keywords; 2. Construct record graph from the OMIM data base; 3. Query record graph using seed record with the Dijkstra's algorithm; 4. Generate distributions of the minimum distances and Bayesian scores for all records in OMIM.

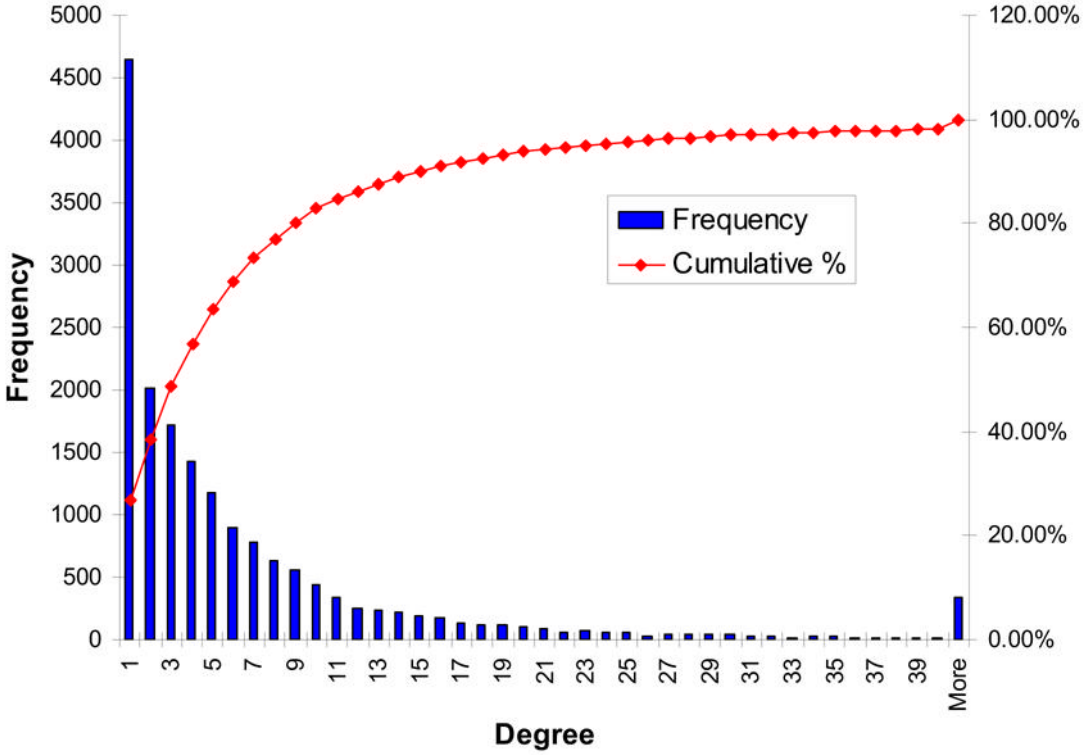

Figure 5.

The distribution of nodes' degrees of OMIM record graph. The histogram represents the distribution of nodes' degrees of the record graph. The curve represents the cumulative distribution (%) of the nodes' degrees.

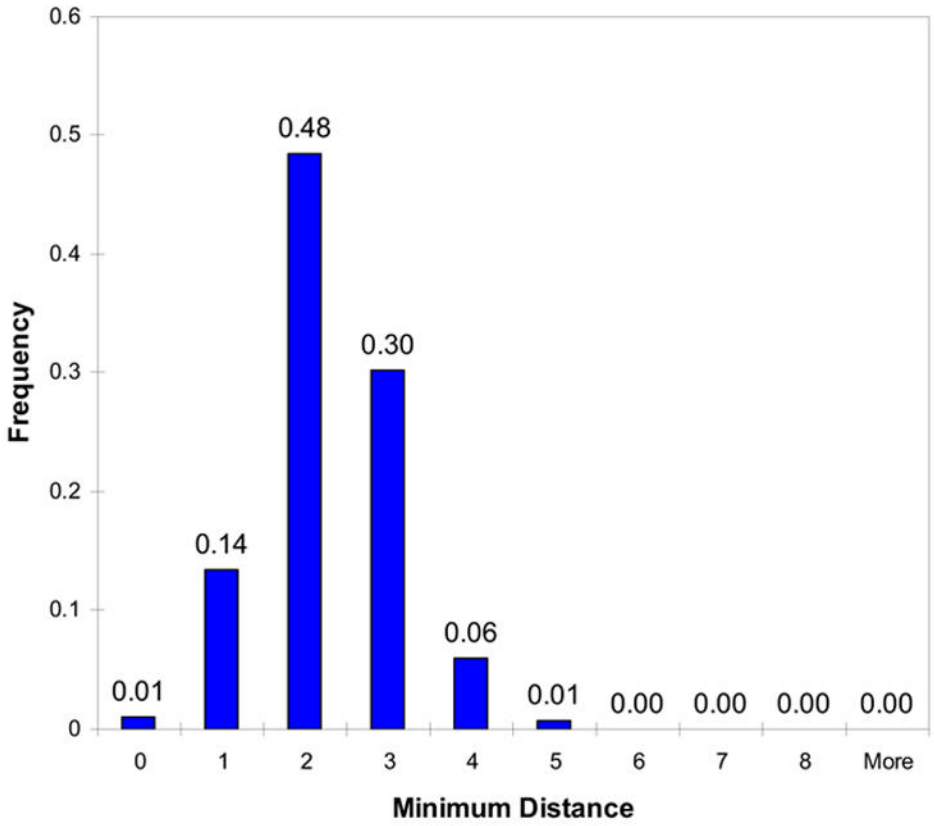

Figure 6.

Distribution of the Minimum Distance of OMIM records of the major subgraph. The minimum distances from seed records to all the records in the major subgraph were calculated using Dijkstra's algorithm.

Table 1

Significant biomarkers found by link-test †.

Microarray marker	Length	Mass Marker	-ln*p*	OMIM ID	OMIM Distance
RPL14: ribosomal protein L14	228	6901.92 4238.47	6.03243 5.8982	N/A	N/A
RPL12: 60S ribosomal protein L12	168	3459.02*	5.9712	180475	N/A
RPL4 : ribosomal protein L4	434	4216.62	6.90635	180479	N/A
TMED3: transmembrane emp24 domain containing 3	110	4246.98 4240.90	5.05342 5.7951	N/A	N/A
RPS4X: 40S ribosomal protein S4, Y isoform 1	133	1914.14 4209.35 3360.93	5.08932 5.23715 5.43597	312760	2
KLK2: Kallikrein 2 precursor (Tissue kallikrein 2)	132	1809.10	5.60852	147960	0
RPL35: 60S ribosomal protein L35	63	6908.12 3362.01	8.93158 8.21264	N/A	N/A
Tspan-1: Tetraspanin-1	122	1850.17*	7.04008	N/A	N/A
NDUFV2: NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa	126	2004.19	5.21845	600532	2
PTPLA: protein tyrosine phosphatase-like, member a	145	3459.02*	6.16651	N/A	N/A
CDH12: cadherin 12, type 2 preproprotein	582	1850.17*	5.33553	600562	2
STK39: STE20/SPS1-related proline-alanine rich protein kinase (Ste-20 related kinase)	275	1971.73	5.96336	607648	2
SLC39A6: solute carrier family 39 (zinc transporter), member 6	1044	2048.72	5.32459	608731	1

† $\alpha_{overall}$=0.1, $\alpha_{individual}$=6.74e-3, the average number of mass markers $K \approx 15$.

* also found in another microarray marker