INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

EFFICIENT AND EFFECTIVE APPROACHES FOR GENE EXPRESSION AND METHYLATION DATA

S. Nithya[1], Dr. S. Uma[2]

[1]*Student, Department of computer science and engineering, Hindusthan institute of technology, Coimbatore*
[2]*Head of the department, Department of information technology, Hindusthan institute of technology, Coimbatore*

Abstract: - Microarray is a powerful technology which could concurrently discover the stages of thousands of transcripts such as genes among diverse experimental constraints or tissue samples. Association rule mining algorithm generally used for discovering the relationship among high dimensional gene dataset. In the existing scenario, ranking of association rule mining algorithms are utilized for taking the decision from the gene dataset. The new technique introduced in this scenario named as Rank-Based Weighted Association Rule Mining (RANWAR) to rank the rules using two novel rule-interestingness measures. They are rank-based weighted condensed support and weighted condensed confidence measures to overcome this issue. Such kinds of measures are usually based on the rank of genes. Using the rank, we assign weight to each item generates much less number of frequent item sets than the state-of-the-art association rule mining algorithms. However this scenario has issue with utility of the information loss and hence efficiency of the scenario gets reduced significantly. In the proposed system, we use temporal mining algorithm to discover the most frequent temporal itemsets. To increase the mining utility performance we propose Utility Sequence Pattern Algorithm (USPA). It is used to reduce the information loss by extracting the significant features from the specified temporal dataset. Micro Genetic Algorithm (MGA) is focused on the sub feature selection by mining the more relevant features. The experimental result concludes that the proposed system is more efficient utility sequential pattern algorithm. Thus proposed system is better than the existing system in terms of fast computation, reduction in information loss, and memory efficiency rather the existing system.

Introduction

Data mining is mainly used for the extraction of knowledge from huge volume of data ware house. It is an analytic procedure developed to discover the huge volume of data in search of reliable model and systematic relationships among variables. Then it can be vision as a result of ordinary development of information in improvement of functionalities such as data selection, database formation, data management and data investigation. It is the procedure where intelligent techniques in order to mine information patterns from databases, data repositories and other information storages.

The data mining concept is categorized into two classes such as descriptive and predictive. Descriptive mining tasks describe the common properties of the information in the database. Predictive extraction tasks execute inference on the present data to construct predictions. Data mining methods are classified to the types of databases extracted, the kinds of knowledge mined, the techniques used, or the applications adapted. This question language is developed to support AdHoc as well as communicative data. The functionalities are class descriptions, correlations, forecast, and cluster and outlier analysis. Frequent mining patterns are used to detect the patterns which are frequently occurred in the database and correlations of data are also discovered.

Gordon et.al [1] suggested the linear models and empirical bayes techniques for assessing differential expression in microarray experiments. This paper handled the issue of recognizing differentially expressed genes in designed microarray. The main purpose of this scenario is to improve the hierarchical model into practical method for common microarray experiments along with arbitrary numbers of samples. However this has issue with huge variations in the final prediction results.

Saurav Malik et.al [2] proposed integrated statistical and rule mining methods in gene expression dataset analysis. This scenario is utilized statistical test and association rule mining approaches to predict the various gene or sample classes for the given gene expression data. In this work, the author suggested a new rule based classification approach to generate the various rules based on the ranking concept. But still this research has issue with high dimensional dataset in few cases.

M.K.Ghose et.al [3] discussed about association rule mining in genomics. Association rules are applied extensively in the marketing industrial analysis and medical data analysis. The association rule is defined to extract the most frequent itemsets or patterns from the given database. It is pair of disjoint itemsets and the rules are used to identify the two disjoint itemsets. However it takes long time for training process in case of large dataset.

P.K. Vaishali et.al [4] presented application of data mining and soft computing in bioinformatics. The soft computing methods are such as neural network algorithms, optimization algorithm like genetic approach and fuzzy logic. This research scenario is focused on the problems associated to data mining methods in bioinformatics. The above mentioned methods are used to detect the hidden patterns from the large dataset. However it has issue with lack in ontology concepts and hence the results are not superior.

Feng Tao et.al [5] suggested weighted association rule mining by using weighted significance and support framework. This research is handled the problems of detecting significant binary relationships in transaction datasets along with weighted setting. The main objective is to guide the extraction focus to those significant associations linking items along with significant weights. A novel approach is named as weighted association rule mining which is improved depends on the model. This algorithm is more efficient and scalable but however it is expensive scenario.

Ke Sun et.al [6] presented mining weighted association rules without preassigned weights. Association rule mining technique is focused to discover the huge transaction database which is used to reveal the implicit relations between data attributes. This research is introduced w-support and it is not required the preassigned weights. By using the link based models it takes the quality of transactions. However it has issue with accuracy of the results.

Paolo Palmerini et.al [7] discussed enhancing the Apriori algorithm for frequent set counting. In this research, Apriori class of data mining approach is introduced to solve the frequent set counting issue. The main objective of this research is achieving the optimization of reduction in time complexity of Apriori approach and pruned the unnecessary itemsets more effectively. However it has issue with high dimensional datasets.

Asha Thomas et.al [8] discussed expression profiling of cervical cancers in various stages to discover gene signatures during progression of the disease. Cervical carcinoma is most frequent cancer affecting women worldwide reporting for more than half a million per year. This scenario is used gene expression profiling to analyze variations in cervical cancer. It is capable to discover potential biomarkers but still it has issue with expression of the gene in cervical cancer has not been reported.

2. Materials and Methods

Preprocessing

In this module, the preprocessing step is performed to improve the classification results more effectively. For the specified dataset, the instances and tuples are assigned with weights values. The main aim of the preprocessing is to examine the missing values, replicate handling, flat pattern filtering and pattern standardization. It is also focused on the removal of noise rates for the given dataset and thus the size of the dataset is reduced significantly.

ARM and naïve bayes method

In this process, the rules are generated which is focused on the gene selection among the number of genes. It generates all combination of gene data that have support count value above certain threshold named as minsupport. Then generate those combinations large itemsets and all other combinations of which are not meeting the threshold small itemsets. The approach makes multiple passes over the database. It determines and selects the most frequent gene points from the dataset. But it does not suitable for larger dataset hence it reduced the accuracy of results significantly. The bayes method is used to select the most similar gene data and it is used to reduce the frequent items compare the previous approach. But this approach is also failed to reduce most informative genes among several gene dataset.

Ranking procedure

In this module, there are two novel rule measures such as weighted condensed support (wcs) and weighted condensed confidence (wcc) for the given dataset. The assigned weights w and w_i is attached forever gene g_i. The pair of (g_i and w_i) instances is described as a weighted gene and the weight of gene g_i in the kth sample is identified by w_{ki}. If the gene g_i provides in the k-th transaction (s_k), then value of w_{ki} is the weight of the gene g_i otherwise, value of becomes zero. In other words,

$$w_{ki} = \begin{cases} w_{i,} & if \text{ gi } \in s_k \\ 0, & otherwise \end{cases} \qquad (1)$$

For given dataset, differential expression values of genes are computed on the statement of independency of genes. If all the specified genes are dependent then particular possibilities of the genes are invalid. This process is mainly focused to discover the itemset transaction weight based on the independent of gene dataset. The weight value of gene is computed by the ranked value of gene.

$$W_k(Z)=\prod_{i=1}^{Q} \forall g_i \in Z, Q = |Z| \ \ wk_i \qquad (2)$$

Where$W_k(Z)$ denotes itemset-transaction weight of itemset for k-th transaction, wk_i refers to the weight of gene gi for the kth transaction.

Assigning weight value for each gene

In the given gene expression dataset the weight values are allocated to every gene. Initially filtering methods are used to remove the low variance gene features. The weights of genes have been computed in order to weights of any two repeated ranked genes. The weight value ranges are among 0 to 1. If the number of genes are represented by n then it can identified as $w_{i,}1 \leq i < n$ which is evaluated through a rank function.

$$w_i = \frac{1}{n} * (n - (r_i - 1)) \qquad (3)$$

Where r_i is ranked gene, and n is number of genes

Data discretization

In this process, r represents genes and c represents samples in the given microarray dataset. For discretization idea the utilized standard k-means clustering algorithm is used. Initially, we select former cluster center uniformly at random from all the data points (X). Then, for each data-point, this approach is used to calculate the distance among the data-point and neighbor center which is already selected. This process is to determine the possibility function depends on the following formula:

$$D(y')^2/(\sum_{y\in X} D(y)^2 \qquad (4)$$

Repeat the procedure for number of cluster centers. By this way, the preliminary centers are selected which is utilized for the standard k-means clustering. This data discretization is denoting the treated and normal genes respectively.

Recognition of Frequent Item set and Rule Mining

In this process, the association rule mining concept is focused to execute the most frequent item sets depends on the top k ranked gene sequences. Initially this method is estimated the wcs of 1-itemsets then denote the frequent singleton gene dataset. In same way, this method is computing the 2-itemsets and 3-itemsets then recognize the frequent 2 itemsets as well as frequent 3 itemsets respectively. The specified rules are mined from these itemsets generation.

```
Procedure RANWAR
Normalize the data matrix D using zero mean normalization
Estimate the gene rank for the given dataset
Assign the weights to all genes based on the rank
Transpose the normalized data matrix
Choose the initial seed values by using k means clustering'
Discretize the transposed matrix applying standard k-means clustering sample-wise.
Apply post discretization technique
Initialize k=1
Determine frequent 1-itemset
Repeat the procedure
K=k+1
Generate the candidate itemsets
For each candidate itemsets, c∈ $CI_k$
Compute the ecs value
If wcs >=min_wsupp then
$FI_k$ ← $[FI_k; c]$
Generate the rule from frequent itemset
Find the wcc
For each rule do
If wcc(r)>= min_wconf then
Save the results
Rulesupp←wcs(r) and Ruleconf ←wcc(r)
End if
End for
End if
End for
Until(FIk=null)
End procedure
```

Utility Sequential Pattern Algorithm

This algorithm is mainly used to focus on the information loss by increasing the utility itemset in maximum.

```
Process the incremental database and if p is a leaf node then return
Examine the projected database S(v(t)) once to:
put I-Concatenation items into ilist, or
put S-Concatenation items into slist and Calculate the utility value Uyz of the item iyz and add this utility value to the sequence utility suy of the sequence seqy .
Eliminate unpromising items in ilist and slist
for each item i in ilist do
(t' , v(t ')) ← I-Concatenate(p,i)
if umax(t' ) ξ then
Output t'
USpan(t' , v(t' ))
```

```
11. for each item i in slist do
12. (t’ , v(t ‘)) ← S-Concatenate(p,i)
13. if umax(t ) ξ then
14. output t’
15. USpan(t’ , v(t’ ))
16. Find all high utility sequential patterns with i as their prefix item by finding HUS(i, sd(i), r) procedure
17. Return
```

Result and discussion

In this section, the performance metrics are compared by using existing and proposed techniques. The performance metrics are such as accuracy, precision, recall and f-measure. The existing ARM, Naïve bayes and RANWAR and proposed USPA algorithm is used to classify the multiple measurement of specified dataset. However the existing method has shown the lower performance in the gene classification results. The proposed USPA method has shown the higher performance in the gene classification results. The experimental result concludes that the proposed system is better than the existing system. An experimental result shows that the proposed method achieves superior performance in terms of precision, recall, f-measure and accuracy metrics.

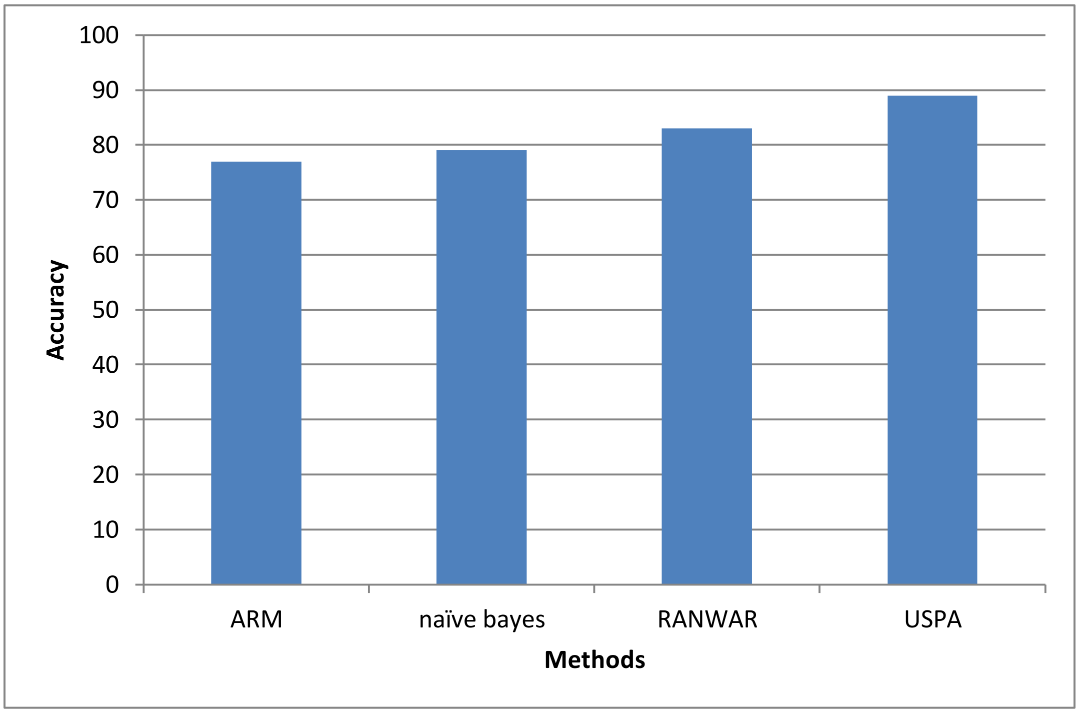
3.1. Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative}+\text{False positive}+ \text{False negative}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.



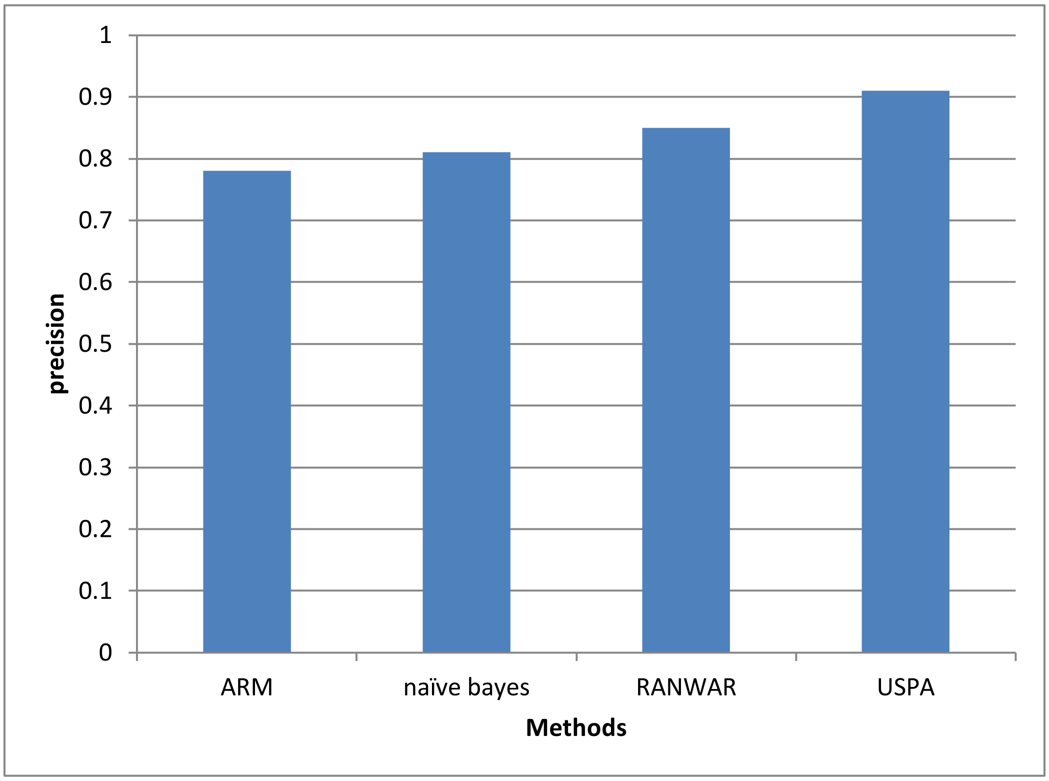
From the above graph we can observe that the comparison of existing and proposed system in terms of accuracy metric. In x axis we plot the methods and in y axis we plot the accuracy values. In existing scenario, the accuracy values are lower by using ARM, NB and RANWAR algorithms. The accuracy value of existing scenario is 77, 79 and 83 % using ARM, NB and RANWAR methods respectively. In proposed system, the accuracy value is higher by using the USPA algorithm. The accuracy value of proposed scenario is 89%. From the result, we conclude that proposed system is superior in performance.

Precision

The precision is calculated as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive}+\text{False positive}}$$

Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant. In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).



From the above graph we can observe that the comparison of existing and proposed system in terms of precision metric. In x axis we plot the methods and in y axis we plot the precision values. In existing scenario, the precision values are lower by using ARM, NB and RANWAR algorithms. The precision value of existing scenario is 0.78, 0.81 and 0.85 using ARM, NB and RANWAR methods respectively. In proposed system, the precision value is higher by using the USPA algorithm. The precision value of proposed scenario is 0.91. From the result, we conclude that proposed system is superior in performance.

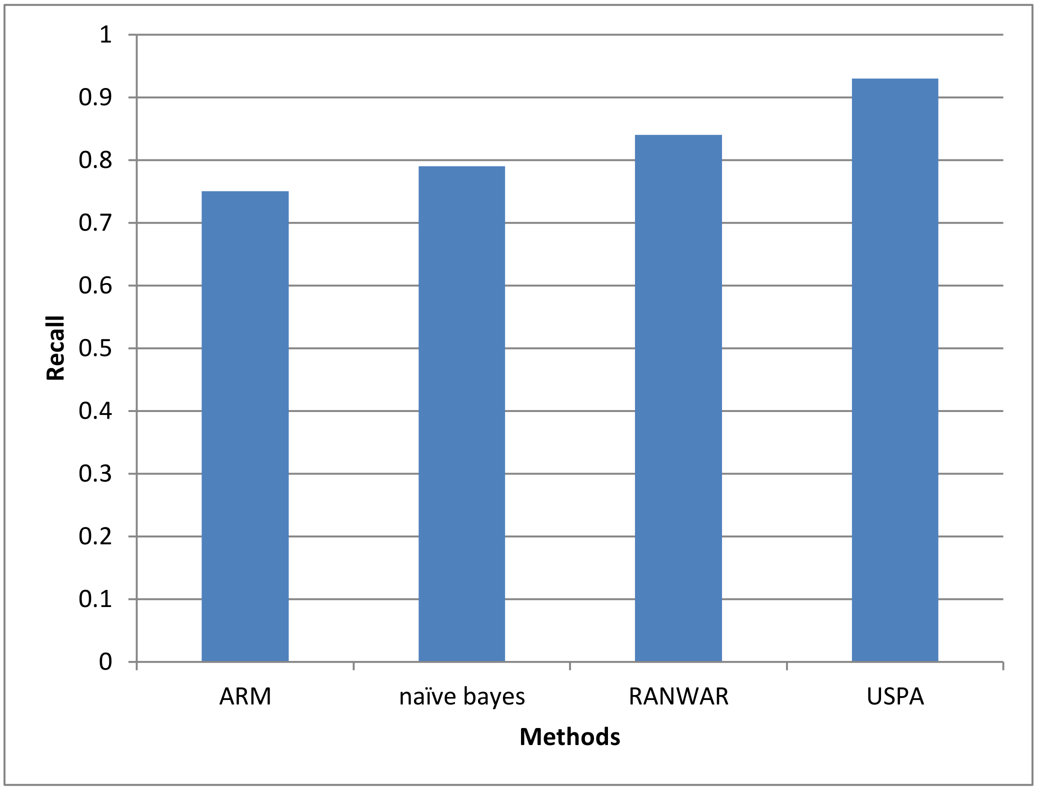
Recall

The calculation of the recall value is done as follows:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive}+\text{False negative}}$$

The comparison graph is depicted as follows:

Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).



From the above graph we can observe that the comparison of existing and proposed system in terms of recall metric. In x axis we plot the methods and in y axis we plot the recall values. In existing scenario, the recall values are lower by using ARM, NB and RANWAR algorithms. The recall value of existing scenario is 0.75, 0.79 and 0.84 using ARM, NB and RANWAR methods respectively. In proposed system, the recall value is higher by using the USPA algorithm. The recall value of proposed scenario is 0.93. From the result, we conclude that proposed system is superior in performance.

Conclusion

In this section, we conclude that our proposed system yields higher performance in terms of better identification gene results. In the existing scenario, the ARM and naïve bayes method introduced which focused on the gene selection based on the frequent count and probability conditions respectively. Then a weighted rule-mining algorithm has been developed using the measures especially for medical gene expression dataset. The rank-based weighted condensed support and weighted condensed confidence measures are used to select the most repeated gene sets. Such kinds of measures are usually based on the rank of genes. By using the rank, we allocate weight to each item generates much less number of frequent item sets than the state-of-the-art association rule mining algorithms. To avoid the information loss, in the proposed system UPSA is sued which is increasing the system accuracy greater. Micro genetic algorithm is used for efficient feature selection which is used to improve higher accuracy. The proposed algorithm provides higher precision, recall and accuracy values for gene classification results. From the experimental result, the method proved that the proposed RANWAR is better than existing ARM, naïve bayes and RANWAR methods.

REFERENCES

1. G. Smyth, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Stat. Appl. Genet. Mol. Biol.*, vol. 3, no. 1, p. 3, 2004.
2. S. Mallik *et al*., "Integrated analysis of gene expression and genomewide DNA methylation for tumor prediction: An association rule miningbased approach," in *Proc. 2013 IEEE Symp. Comput. Intell. Bioinformat. Comput. Biol. (CIBCB)*, Singapore, pp. 120–127.
3. M. Anandhavalli, M. K. Ghose, and K. Gauthaman, "Association Rule Mining in Genomics," *Int. J. Comput. Theory Eng.*, vol. 2, no. 2, pp. 1793–8201, 2010.
4. Vaishali, P. K., and A. Vinayababu. "Application of Data mining and Soft Computing in Bioinformatics." *International Journal of Engineering Research and Applications (IJERA), ISSN*: 2248-9622.
5. Tao, Feng, Fionn Murtagh, and Mohsen Farid. "Weighted association rule mining using weighted support and significance framework." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003.
6. Sun, Ke, and Fengshan Bai. "Mining weighted association rules without preassigned weights." *Knowledge and Data Engineering, IEEE Transactions on* 20.4 (2008): 489-495.
7. Perego, Raffaele, Salvatore Orlando, and P. Palmerini. "Enhancing the apriori algorithm for frequent set counting." *Data Warehousing and Knowledge Discovery*. Springer Berlin Heidelberg, 2001. 71-82.
8. Premalatha, S., and C. Usha Nandhini. "Efficiently Generating The Rank Based Weighted Association Rule Mining Using Apriori Algorithm In High Biological Database." (2015).
9. Thomas, Asha, et al. "Expression profiling of cervical cancers in Indian women at different stages to identify gene signatures during progression of the disease." *Cancer medicine* 2.6 (2013): 836-848.