USE OF RELEVANCE NETWORKS IN KNOWLEDGE DISCOVERY FROM MEDICAL DATABASES

[1]Senthil Kumar Ramadoss, [2]Puviarasan N

[1]Research Scholar, [2]Professor & Head
Department of Computer and Information Science,
Annamalai University, TN, India

Abstract: While i*ncreasing amounts of data exist in medical databases and multiple variables are measured for each case in a data set, there exists an underlying relationship between all pairs of variables, some highly correlated and some not. This report describes a technique that creates networks of related variables, or* relevance *networks,* by dropping links with either too weak correlation or too few data points to defend the relationship. The paper describes how applying this methodology to the domain of laboratory results allows the generation of meaningful relations between types of laboratory tests. These relations could be used as the basis of further exploratory research.*

Index Terms – **Database Knowledge Discovery, Medical Database, Relevance Networks, Data Mining, knowledge Discovery Models, Strategic knowledge Discovery.**

I. INTRODUCTION & BACKGROUND

In an unsupervised manner, extract meaningful relationships between variables in large medical databases in order to generate hypotheses to be studied with targeted research. With increasing data in medical databases, medical data mining is growing in popularity. Some of these analyses use techniques from the machine learning literature, including inducing propositional rules from databases using rough sets [1], using these rules in an expert system [2], using Bayes models to find similar cases [3], using a finite-mixture-augmented naïve Bayes model to classify cases [4], and constructing decision trees and neural networks to classify cases. These applications have focused on supervised techniques in very specific domains, where cases are labeled and classifications are induced or trained. Unsupervised techniques applied to medicine include AutoClass, a system using mixture models to determine optimal classes. [5]

Bayesian networks [6] have traditionally been used to model conditional probabilities between variables in the medical domain. However, there are three reasons why using Bayesian networks is difficult for unsupervised learning. First, computing the structure of a Bayesian network without any prior assignment is computationally intractable. Second, updating and maintaining the conditional probabilities in networks with cycles is difficult, and networks with cycles should not necessarily be excluded.

Third, computing and updating conditional probability is hard when continuous variables are used instead of discrete values (for instance, 100, 210, or 345 mg/dl versus "high" or "normal"). In addition to these three problems, information on the specific relatedness between variables may be lost when these variables are forced into discrete values.

Our purpose was to exploit existing electronic databases for unsupervised medical knowledge discovery without a prior model or information. Observations collected within labs, physical examination, history, and gene expressions can be expressed as continuous variables describing human physiology at a point in time. These variables may be related to each others in several ways: (1) directly through physiology, such as serum concentration of bicarbonate and alveolar partial pressure of carbon dioxide; (2) related through mathematical formulae, such as absolute neutrophil count and percentage of neutrophils; (3) related indirectly through hidden variables, such as how thyrotropin releasing hormone controls thyroxine level through thyroid stimulating hormone; and (4) related through synonymy, such as somatomedin C and insulin-like growth factor-1 both referring to the same molecule. Although the relations discovered in a medical database may not be of high quality compared to a prospective study, or even a comprehensive retrospective chart review, we felt that the hypotheses generated can be used to fuel further clinical investigations. Our goal was to identify candidate models and systems of these putative relationships for further exploration. Specifically, we wanted to ascertain the relationships between laboratory tests, to see if an unsupervised technique could discover the physiologic, mathematical, and other classes of relationships between types of tests.

II. METHODS

Construction of Table of Simultaneous Laboratory Measurements

We first created a list of all patients registered at Children's Hospital between November 1998 and February 1999. We then created a list of all clinical laboratory tests performed on these patients along with results and date and time of specimen collection. Laboratory tests that did not result in a numeric value at least three or more times were excluded.

Each laboratory measurement was considered in the context of an interval of time for which the result was considered to be valid. These intervals were calculated without *a priori* knowledge. For each *type* of laboratory test, a pass was made through the list of results and the minimum amount of time between any two sequential results for a patient was calculated. This minimum time was considered the *valid interval* for that lab test. For example, if a shortest interval for sequential hematocrit measurements on any single patient was when it was measured on patient X at 8 AM and 10 AM on the same day, then the valid interval for all hematocrit results for all patients was set to be 2 hours. We acknowledge that these valid intervals could have been determined more accurately and appropriately with an expert-constructed knowledge base of half-lives and other physiologic parameters; however, such a technique does not realistically scale well when hundreds of laboratory tests are to be analyzed.

For each patient, the list of laboratory results was cross-tabulated into a table, so that each type of laboratory test was placed in a separate column. Laboratory results with overlapping valid intervals of time were placed in the same rows. Thus, laboratory tests that were not performed simultaneously, but whose valid intervals overlapped, could be compared against each other in the same row. As shown in figure 1, this allowed for tests with longer valid intervals to be compared to tests with shorter valid intervals. We felt this reduced the bias of similarly-sampled routine tests being exclusively compared against each other. The resulting cross-tabulated tables were sparse arrays, in that not every laboratory test was present in every intersecting valid interval. Unknown values were treated differently than zero values.
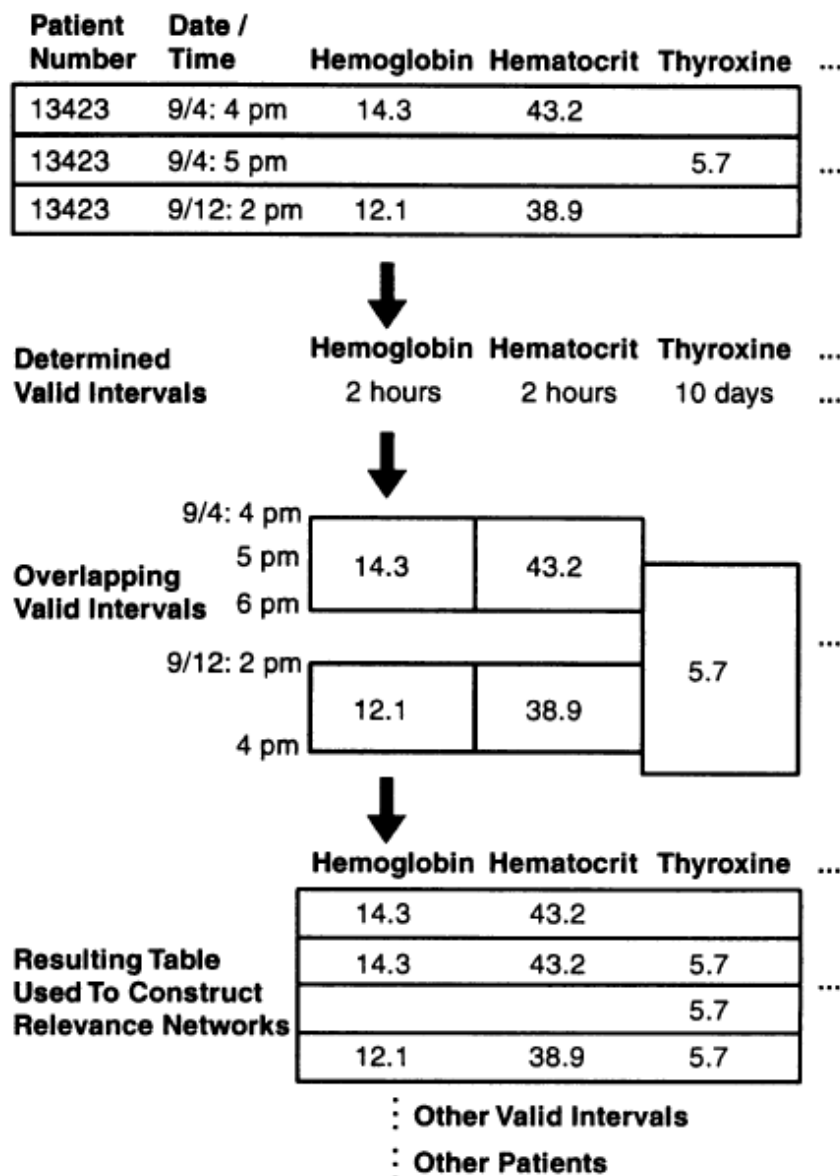

Figure 1: Methodology for creating table of overlapping laboratory results

Creation of Relevance Networks

Each column in the table was treated as a separate variable. We first performed an analysis between each pair of variables. If there were three or more unique rows where each of the pair of variables was present, we attempted to fit a linear model between them. For each pair wise comparison, we stored the correlation coefficient, r, measuring the quality of fit of the linear model, and the number of intersecting intervals, n. The result of this was that almost every variable was connected to every other variable by a linear model of varying quality, by which we mean the n and r2 of each correlation varied widely.

To split this nearly completely-connected network, a threshold n and r2 were chosen. Those links representing linear models with r2 under the threshold or models constructed with fewer points of data than the threshold n, were dropped. This led to the breakup of the completely connected network into smaller islands, where connections were stronger than the chosen thresholds. These islands, or relevance networks, were then displayed in a graphical manner, and the relationship of the variables in each relevance network was analyzed manually.

Graphically, each relevance network is represented separately. Each node represents a variable. The thickness of the link between variables is drawn proportionally to the r2 of the link. The number of surrounding lines around the link (or the "jiggle" of the link) is proportional to the n of the link. A legend summarizing these graphical features is shown in figure 2.

The number of data points used to create the model, or n, increases in the table from left to right, and is represented by the number of surrounding lines, or the "jiggle" of the link. The r^2 of the model increases in the table from top to bottom, and is represented by the thickness of the line. Not shown, 0 placed next to a line indicates a linear model with negative slope (negative r).
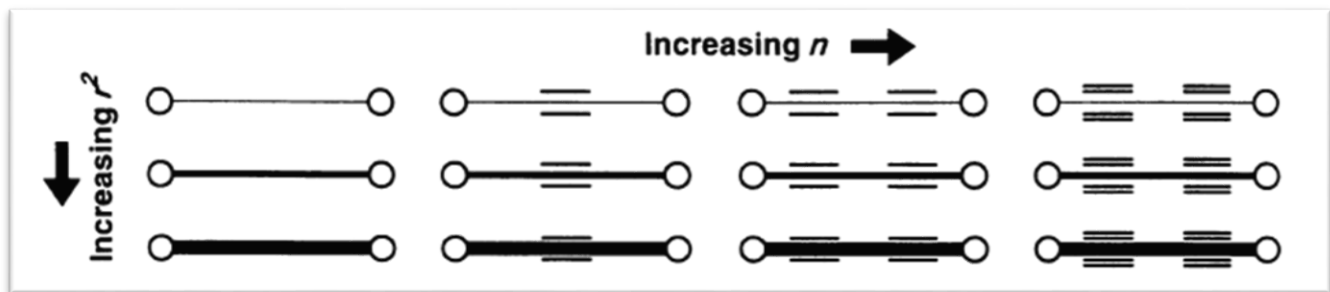
Figure 2: Graphical representation of models between two variables

III. RESULTS

A total of 5,158 patients were found to be registered during the 4 month period. These patients had 798 different types of laboratory tests performed that resulted in numeric values, making a total of 410,514 distinct results. When laboratory tests performed fewer than 3 times were excluded, 642 types of tests remained. After the cross-tabulation, the resulting array measured 642 (types of tests) by 28,566 (intersecting valid intervals). This computation was performed by a program written in Java querying an Oracle 8 database, and running on a 266 megahertz Pentium II processor. The process took under 15 minutes of wall-clock time.

Creation of the pairwise r and n arrays was performed by a C language program on a Sun Ultra HPC 5000 server running Solaris, taking 30 minutes of wall-clock time. Each resulting array was square, measuring 642 columns on each side. The creation of the relevance networks was performed by a program written in Matlab, running in under 10 seconds on a 266 megahertz Pentium II processor. Constructing the graphical representation of these relevance networks took under 15 seconds. Setting the threshold r2 at 0.6 and n at 50 resulted in the relevance networks seen in figure 3.

The overall model detected both positive and negative correlations. The resulting relevance networks shown here are not surprising; we show them because it is reassuring that what is being found is consistent with basic human physiology. However, as the r2 and n are lowered to less conservative values, other relevance networks appear that are less obvious, and remain for further clinical investigation. Each link in each network represents a putative connection belonging to taxonomy of five types of relationships: identity or synonymy, mathematical, physiologic, pathologic, and causal. An example of the first is seen in network 16, where serum fibrinogen level is linked to estimated fibrinogen level. These two variables represent an identical concept and distribution. Other links of this type are in network 4, where serum chloride is related to whole blood chloride and in network 10 where two separate labels exist for arterial oxygen saturation measurements. Network 6 demonstrates the second type of link: mathematical. Prothrombin time is linked to the international normalized ratio, a mathematical relation that normalizes the prothrombin time using laboratory controls.

In network 9, hematocrit is calculated based on red blood cell count. Many physiologic links were found in the data set. In network 8, increasing serum blood sugar level is positively correlated with increasing serum insulin level. In the sweat test for cystic fibrosis, sodium and chloride are transported together and this is shown in network 18. In network 15, the amount of hemoglobin in red cells is correctly shown to be related to the size of the red cells. Network 19 shows the inverse relationship between the partial pressure of carbon dioxide and pH.A pathologic type of link is shown in network 7.

Erythrocyte sedimentation rate is a nonspecific indicator of inflammation, and its link to alpha-1 antitrypsin, an acute phase protein properly models this relationship in inflammatory conditions. Dehydration from diarrhea is a common condition in our patient $n = 5$ population, and hypertonic hypernatremia can be demonstrated by an increase in both sodium and osmolarity, shown in network 1. Network 5 shows aspartate aminotransferase level related to lactic acid dehydrogenase, both of which are elevated in hepatic disease. The laboratory tests include only a few variables relevant to human biology; therefore, few of the links are directly causal. One instance where the link is direct is when increasing blood sugar causes the beta-cells in the pancreas to release more insulin. This is properly demonstrated in network 8. The effect of the threshold r^2 and n on the number of $= 30$ nets created is shown in figure 4.

As the threshold r^2 is lowered, initially more small relevance nets are created. As the process continues, these scattered nets are merged finally into one large network. Thus, the threshold r^2 can be thought of as the fit, or *accuracy,* of the entire model. Lowering the threshold n has a different effect: more links are created having fewer data points to back up each model. Not shown in figure 4 is the $n = 5000$ size of the networks and how they vary with the threshold r^2 and n; with lower n, larger networks are formed at higher r^2. Statistically, the confidence interval for 7 is inversely proportional to n: as more points are included, the confidence interval for r^2 is narrower. [7] In this way, the threshold n can be thought of as the *consistency,* or acceptable degree of error in each model between variables and the overall model. By setting the two parameters of threshold r^2 and $n,$ one can select a model with a particular accuracy and consistency for the degree of belief, as diagrammed in figure **5.**
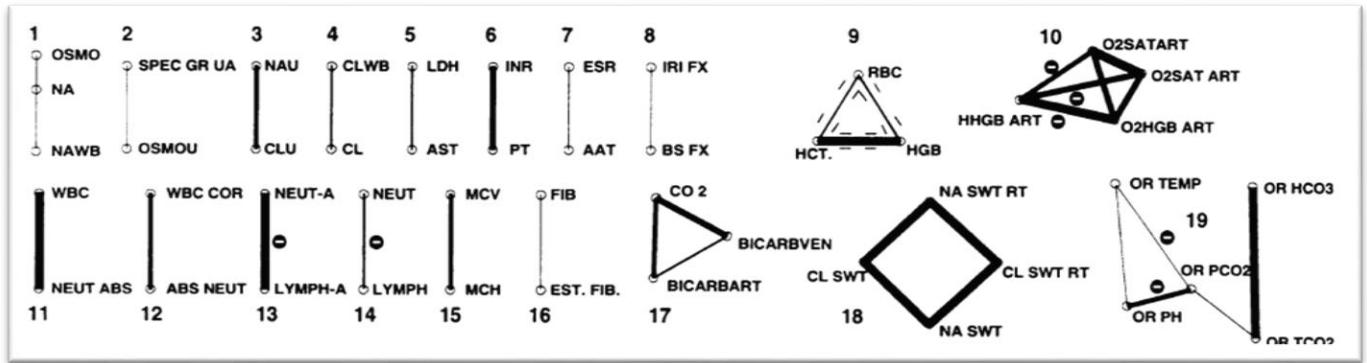
Figure 3: Relevance networks formed with threshold r^2 at 0.6 and n at 50

(Abbreviations: 1. OSMO: serum osmolarity, NA: serum sodium, NAWB: whole blood sodium; 2. SPEC GR UA: urine specific gravity, OSMOU: urine osmolarity; 3. NAU: urine sodium, CLU: urine chloride; 4. CLWB: whole blood chloride, CL: serum chloride; 5. LDH: lactic acid dehydrogenase, AST: aspartate aminotransferase; 6. INR: international normalized ratio, PT: prothrombin time; 7. ESR: erythrocyte sedimentation rate, AAT: alpha-1 antitrypsin; 8. IRI FX: insulin level, BS FX: blood sugar; 9. RBC: red blood cell count, HGB: hemoglobin, HCT: hematocrit; 10. O2SATART, 02SAT ART: arterial oxygen saturation, O2HGB ART: arterial oxyhemoglobin, HHGB ART: arterial deoxyhemoglobin; 11. WBC: white blood cell count, NEUT ABS: absolute neutrophil count in automated differential; 12. WBC COR: corrected white blood cell count, ABS NEUT: absolute neutrophil count in manual differential; 13. NEUT-A: percentage of neutrophils in automated differential, LYMPH-A: percentage of lymphocytes; 14. NEUT: percentage of neutrophils in manual differential, LYMPH: percentage of lymphocytes; 15. MCV: mean red cell corpuscular volume, MCH: mean corpuscular hemoglobin; 16. FIB: fibrinogen, EST FIB: estimated fibrinogen; 17. CO 2: serum bicarbonate, BICARBVEN: venous bicarbonate measured as blood gas, BICARBART: arterial bicarbonate; 18. NA SWT: sweat test left arm sodium, CL SWT: chloride, NA SWT RT: sweat test right arm sodium, CL SWT RT: chloride; 19. OR TEMP: operating room patient temperature, OR PH: blood gas pH, OR PCO2: blood gas partial pressure of carbon dioxide, OR TCO2: blood gas measured whole blood bicarbonate, OR HCO3: serum bicarbonate.)
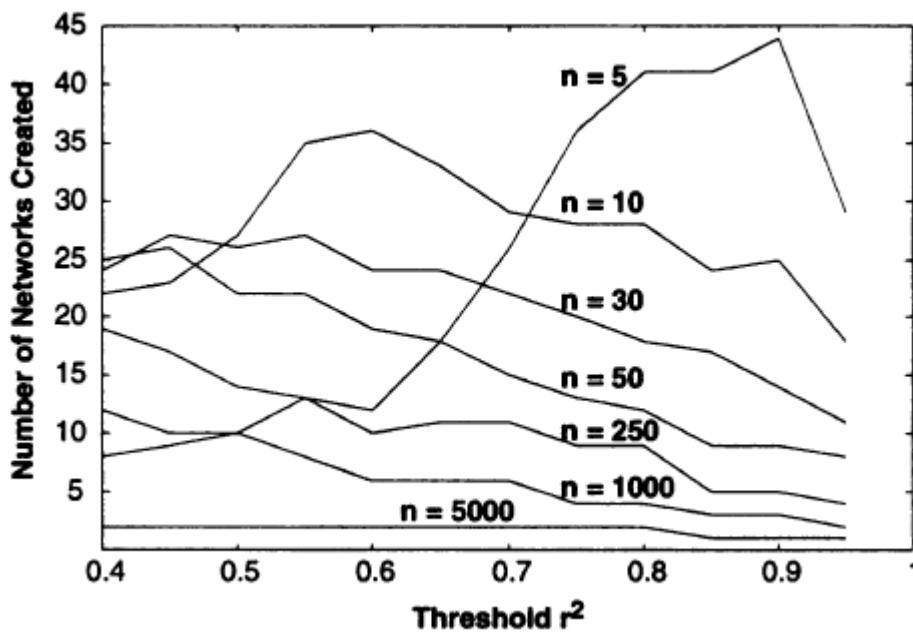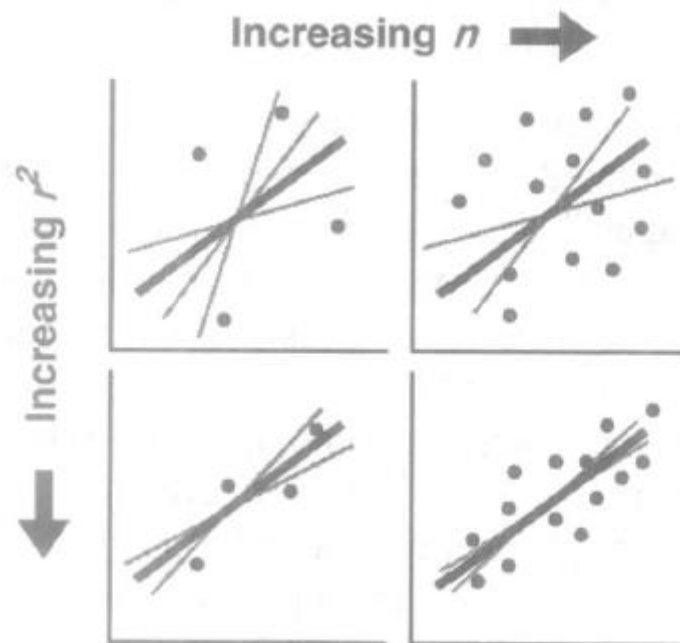

Figure 4: Changing the threshold r^2 and n affects the number of relevance networks created

Figure 5: Increasing r^2 and n

It may be noted from Figure 5 that an increasing $\textbf{n}$ narrows the confidence interval of r^2, improving the consistency of the linear models and an increasing r^2 means an improvement in the accuracy of fit of the linear models.

IV. DISCUSSION

By applying the technique of creating islands of variables with high cross-correlation coefficients, or relevance networks, to the laboratory results cross-tabulation, we were successfully able to generate valid physiologic, pathophysiologic and mathematical, and synonymic relationships hypotheses. Changing the threshold r^2 and n allows specificity in the accuracy and consistency of the generation of a range of networks with varying degrees of belief.

Creating relevance networks on a larger set of laboratory results would produce links backed by more data points; this could be used to automate linking of laboratory results classified under two labels (for instance, merging results from separate institutions). There are a few limitations with this technique. First, the relevance networks are necessarily undirected. Each link represents a hypothesis that a linear relationship exists between two variables; more complex casual mechanisms cannot be ascertained without *a priori* knowledge.

Related to this, each node or variable is not analyzed free from confounders. Confounders may exist in four ways: (1) categorical variables that may or may not be present in the model, but have an effect on intravariable relationships, such as the specific medical identifier number or gender; (2) continuous variables that are in the model, but act as discrete variables, as exemplified by the fact that a higher serum hCG does not make one "more pregnant"; (3) continuous variables that are hidden from the model, such as the myriad of socioeconomic variables and other physiologic parameters that confound use of initial blood glucose to determine length of stay for an admission for new onset diabetes; and (4) variables that directly or indirectly influenced the selection bias, such as the date used to select these laboratory results.

One may view the purpose of this methodology as to find as many such confounders as possible that may exist in the data set. A domain expert or additional data are still needed to ascertain the importance of each link and how direct or indirect each link is.

V. CONCLUSION

We envision at least four areas for expanding the use and development of relevance nets. First, although this technique works well on sparse matrices, it performs better on complete matrices with less missing data, such as those from RNA expression scanning arrays [8] or in the stock market domain. Second, other types of models besides linear may be used between variables. Third, picking a few highly connected variables from relevance networks for use as training features in classification engines might allow better classification with fewer features. Finally, rows in the data set that violate the model between two variables may be interesting to study as pathologic exceptions.

REFERENCES

[1] Ohrn A, Ohno-Machado L, Rowland T. Building manageable rough set classifiers. Proc Amia Symp 1998:543-7.
[2] Tsumoto S. Automated knowledge acquisition from clinical databases based on rough sets and attribute-oriented generalization. Proc Amia Symp 1998:548-52.
[3] Cooper GF, Buchanan BG, Kayaalp M, Saul M, Vries JK. Using computer modeling to help identify patient subgroups in clinical data repositories. Proc Amia Symp 1998:180-4.
[4] Monti S, Cooper GF. The impact of modeling the dependencies among patient findings on classification accuracy and calibration. Proc Amia Symp 1998:592-6.
[5] Cheeseman P, Stutz J. Bayesian Classification (AutoClass): Theory and Results. In: Fayyad UM, Piatetsky-Shapiro G, Symth P, Uthurusamy R, editors. Advances in Knowledge Discovery and Data Mining. Cambridge, Massachusetts: The MIT Press; 1996. p. 153-180.

[6] Heckerman D. Bayesian Networks for Knowledge Discovery. In: Fayyad UM, PiatetskyShapiro G, Symth P, Uthurusamy R, editors. Advances in Knowledge Discovery and Data Mining. Cambridge, Massachusetts: The MIT Press; 1996. p. 273-305.

[7] Kleinbaum DG, Kupper LL, Muller ICE. The Correlation Coefficient and Straight-Line Regression Analysis. In: Applied Regression Analysis and Other Multivariable Methods. Belmont, California: Duxbury Press; 1988. p. 90-91.

[8] Ramsay G. DNA chips: state-of-the art. Nat Biotechnol 1998;16(1):40-4.