International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 32343-32354
© Research India Publications
http://www.ripublication.com

Regulation Analysis of Time Series Gene Expression Data in PCOS Using Fuzzy Clustering Algorithm

[1]Murshidha.S.G.Taj and [2]Thangaraj.P

[1]Department of IT, Angel College of Engineering and Technology,
Angel Nagar, Dharapuram Main Road,
P.K.Palayam, Tirupur - 641665, Tamilnadu, India
+91-9842211430. murshidha.rizwan@gmail.com.
[2]Department of CSE, Bannari Amman Institute of Technology,
Alathukombai – Post
Sathyamangalam - 638 401Erode District, Tamil Nadu, India
.+91-9842720572. ctptr@yahoo.co.in.

Abstract

Microarray technology is widely used in monitoring thousands of gene expression levels. Time-series microarray data are gene expression values generated from microarray experiments within certain time intervals.Perfect analysis procedure is essential to extract useful information from the huge amount of gene expression data. In this work the internal connection between time points and the preservation of time locality in time course gene expression data is considered by combining Fuzzy C means (FCM) algorithm with centralization technique to overcome the limitation of existing methods.This proposed method deterministically finds quality clusters acclaiming probable study of gene function for the regulation process based on the experimental conditions.Poly Cystic Ovarian Syndrome (PCOS) data set is used to exemplify the main concepts conversed in the study and the results were compared with Fuzzy C Means algorithm.

Key words: Microarray technology, clustering algorithm, regulation, time series gene expression, fuzzy, data mining, PCOS

I. INTRODUCTION

Vast technological development in current scenario makes microarray techniques generate more and more biological data. This leads to extract useful information from the available junk data. Many data mining techniques have been proposed for varioustypes of data. Fuzzy c-means (FCM) is a clustering method which allows one piece of data to belong to twoor more clusters.In medical diagnostic systems, fuzzy c-means algorithm gives the better results than hard k means algorithm [1].Fuzzy clustering methods can be helpful for the medical experts inproblem-solving.Clustering of time series gene expression data is an enormous field with multiple numbers of clustering techniques along with different types of data. The frequent data sets used is yeast data set. This paperdeals with PCOS data set.

PCOS is a disease found common in women. This cause problems in menstrual cycle and make it difficult to get pregnant. It changes the outlook of the affected person. If it isn't treated, over time it can lead to serious health problems, such as diabetes and heart disease.Symptoms tend to be mild at first. The most common symptoms are Acne, weight gain and trouble losing weight, unwanted hair growth on the face, irregular periods, thinning hair on the scalp, some have no periods while others have excess bleeding, fertility problems, depression and the last stage is trouble in getting pregnant.There is no cure for PCOS. Hence, it needs to be managed to prevent further problems. There are many medications to control the symptoms of PCOS. Doctors most commonly prescribe the birth control pill for this purpose. Birth control pills regulate menstruation, reduce androgen levels, and help to clear acne. Other drugs can help with cosmetic problems. There also are drugs available to control blood pressure and cholesterol. Progestin and insulin-sensitizing medications can be taken to induce a menstrual period and restore normal cycles. Eating a balanced diet low in carbohydrates and maintaining a healthy weight can help lessen the symptoms of PCOS. Regular exercise helps weight loss and also aids the body in reducing blood glucose levels and using insulin more efficiently.

Although it is not recommended as the first course of treatment, surgery called ovarian drilling is available to treat PCOS. This involves laparoscopy, which is done under general anesthesia on an outpatient basis. A very small incision is made above or below the navel, and a small instrument that acts like a telescope is inserted into the abdomen. During laparoscopy, the doctor then can make punctures in the ovary with a small needle carrying an electric current to destroy a small portion of the ovary. The success rate is less than 50% and there is a risk of developing adhesions or scar tissue on the ovary [2] [3] [4].

1. *FSH (Follicle Stimulating hormone)*

To mature the follicles

2. *LH (Luteinizing hormone)*

Causes the ovulation

3. *Estrogen*

This controls the LH surge by a feedback regulation mechanism. It also prepares the uterus by thickening the lining.

4. Progesterone

A negative feedback with LH and FSH, prepares the uterus for implantation, and maintains uterine lining and drop in progesterone along with estrogen causes menstruation.

There are multiple parameters analyzed for finding negative or positive result of PCOS in a patient. Since many parameters are available there are chances of getting positive result but the result may be the other case. This limitation is over come in the proposed method. In 90% of the time series clustering algorithms, data preprocessing technique is neglected. The remaining 10% have utilized the preprocessing technique which is meant only for the noisy data resulting in the loss of some interesting patterns. In most of the cases internal connection between time points and the preservation of time locality in time course gene expression data is not considered. This leads to the formation of bad clusters nothing but unwanted cluster formation resulting in the unnecessary time consumption. With the proposed method the above mentioned issues are addressed by giving an optimal solution. Further this paper is divided into 5 sections. Definitions, related works, methods and materials, proposed method, evaluation, conclusion and future work respectively in section II, III, IV, V, VI, VII and VIII respectively.

II. DEFINITIONS

A.Time series gene expression data

This is obtained with multiple samples gathered from regular time intervals.

B. Co expressed genes

These are the one that shares similar expression patterns discovered by cluster analysis.

C. Co regulated genes

These are the one that are regulated by at least one common known transcription factor.

D. Gene Regulation

Gene regulation is a substrate for evolutionary change. Since control of the timing, location and amount of gene expression can have an intense effect on the functions of the gene in a cell or in a multi cellular organism. It is the process of turning genes on and off. During early development, cells begin to take on specific functions. Gene regulation ensures that the appropriate genes are expressed at the proper times. Itcan also help an organism respond to its environmentand is accomplished by a variety of mechanisms including chemically modifying genes and using regulatory proteins to turn gene on or off.

E. Types of Regulation

There are three types

(1) up-regulation,

(2) Down regulation and
(3) Neither up or down regulated.

F. Up-regulation

There is an increase in the number of receptors on the surface of target cells, making the cells more sensitive to a hormone or another agent.

G. Down-regulation

There is a decrease in the number of receptors on the surface of target cells, making the cells less sensitive to a hormone or another agent. Some receptors can be rapidly down regulated.

III. RELATED WORKS

The basic FCM is an iterative procedure method where cluster centers are updated. The time taken to partition the cluster is reduced with this fast FCM [5]. Many data preprocessing techniques are available in data mining. Normalization is one used for the transformation of data. Another new method called centralization is introduced [6]. With this two conditions are fulfilled and also the regulation of gene expression is well behaved. There are different patterns of co regulatedgenes available. Co regulation graph [7] is generated based on the binned matrix. Based on the aerobic conditions genes are positively and negatively clustered. In order to justify that fuzzy c means algorithm is best for very large data sets, single pass fuzzy c means, online fuzzy c means, weighted fuzzy c means, bit reduced fuzzy c means, kernel fuzzy c means are discussed [8]. Accuracy and feasibility is shown to be primary whereas efficiency is considered secondary. Fuzzy clustering is of two types hard and soft. High dimensional data can be clustered with the help of soft fuzzy clustering. For the energy design function of local minima a new convergence technique is introduced [9]. A concept from diffusion graph isutilized by comparing with genetic data sets and with the data set containing taxonomic measurements [10]. Also, many new interesting patterns are generated using yeast data sets by applying the combination of data preprocessing with supervised and unsupervised learning methods [11]. Various clustering algorithms such as hierarchical, k-means and self-organizing maps are discussed and they are compared with each other for the performance measures [12]. Power spectral analysis is combined with the traditional algorithms.Clustering can be done in two approaches. One is top-down and the other one is bottom-up approach. Pre pruning and tree based clustering method is used for finding maximal subspace of co regulation [13]. All the time series clustering and regulation is done based on the experimental conditions [14]. This method analyzes the similarities of gene expression patterns for different conditions. Some genes may be interrelated with the other biologically. Algorithms are developed to retrieve it from the network of co expressed genes [15]. Regulators play a vital role in gene expression. This has to be identified for proper utilization. Meaningful biological conditions can be drawn from time series gene expression data which depends strongly on regulation [16] instead of expression values. Many redundant patterns are observed due tonoise as it isinherent to microarray data [17].Hence similarity measures between profiles is essential. Although many similarity measures are available Pearson correlation coefficient [18] is found common in all the papers measuring similarity. The values are between -1 and +1 and If 1, 0 and -1, then Positive, no and negative correlation respectively. The ultimate goal of analyzing gene expression data is to identify relationship between co regulated gene pairs [19].

IV. METHODS AND MATERIALS

Binning is a preprocessing technique which has been applied in few papers giving a partial solution. Technique focuses on the noise and removes it. This is done by analyzing the more general effects of genes. Here gene expression data is binned in two levels (a) highly co expressed (up-regulated), and (b) inhibited (down-regulated) based on some precise static threshold. In some cases there may be a third level neither up or down regulated is not considered at all. This paper works on the mentioned limitation.

Time Series Gene Expression Data

The time series gene expression data consists of a matrix containing intensity data for a group of genes for certain time points. Let X_{ij} be the gene expression level representing the i^{th} gene at time point t_j, for i =1, …, p, and j =1, …, n, where p is the number of genes and n is number of time points. The overall structure of the timeseries data is shown in figure 1. Depending on the research problem under study, the entire data or a subset of the above matrix may be selected for the data analysis process.

Pre Processing the data

Though many algorithms have been developed for preprocessing the data, the internal connection between time points and the preservation of time locality in time course gene expression data is not considered to certain extent. PCOS data sets purely rely on the internal connection and time locality. In order to obtain this, centralization a preprocessing technique [6] is combined with the fuzzy c means algorithm to produce three different clusters. Out of the three clusters one will contain gene which is to be up regulated, the other with gene to be down regulated and the last one with normal genes falling into the category neither to be up nor to be down regulated.

Clustering gene expression data

First step in gene expression analysis is clustering. All types of gene clustering are purely based on the suitable similarity measures. This is the fundamental process of data mining. The best proximity measure is Euclidean distance for dense and continuous data. The value is between -1 and +1 in correlation coefficient which works better in linear but not suitable for curve linear[20].

Time-Series data point structure

Time points

$$X_{p*n} = \begin{array}{ccccc} T_1 & T_2 & T_3 & \dots & T_n \\ \hline X_{11} & X_{12} & X_{13} & \dots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2n} \\ & \dots & \dots & & \\ & \dots & \dots & & \\ X_{p1} & X_{p2} & X_{p3} & \dots & X_{pn} \end{array}$$

Fig. 1. Initial structure of the gene expression time series data

Fuzzy c means algorithm

Fuzzy c-means (FCM) is a method developed by Dunn in 1973 and improved by Bezdek in 1981. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. Each iteration results in the updating of membership and cluster centers.It is based on minimization of objective function (1).

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}^{m}\left\|X_i - V_j\right\|^2, 1 \le m < \infty \qquad (1)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j in (2)

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}, \mathrm{C_j} = \frac{\sum_{i=1}^{N}\mu_{ij}^{m}.x_i}{\sum_{i=1}^{N}\mu_{ij}^{m}} \qquad (2)$$

This iteration will stop when

$$max_{ij}\left\{\left|\mu_{ij}^{(k+1)} - \mu_{ij}^{(k)}\right|\right\} < \varepsilon \qquad (3)$$

In Eq.3 ε is a termination criterion between 0 and 1, whereas k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m.

The algorithm is composed of the following steps:

Step 1: Initialization

$$U = [u_{ij}] \text{ matrix}, U^{(0)}$$

Step 2: Calculation of centre vectors

$$C^{(k)} = [c_j] \text{ with } U^{(k)}$$

at k-step with the following formula

$$\mathrm{C_j} = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} . X_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}$$

Step 3: Updation of $U^{(k)}$, $U^{(k+1)}$ is done as shown below

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Step 3: If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then STOP; otherwise return to step 2.

V. PROPOSED METHOD

Combination of Centralization and Fuzzy c means algorithm

In the proposed approach, a new clustering algorithm is introduced which combines FCM algorithm with centralization technique [6]. This generates clusters based on the internal connection between the time points preserving the time locality in time course gene expression data. Let k_{ti}, $k_{tj} \epsilon$ S be two samples, with unknown constants of proportionality c_{ti},c_{tj}. Let $m_{g,k}^{*}$ denote the measured value, i.e. a number proportional to the signal intensity measured at the spot for g on the array for sample k and $b_{g,k}$ is the background noise. Let $G_{ti,tj}$ be the set of genes that are considered to be expressed and reliably measured in both the samples. The other genes are excluded since ratios of values that are dominated by background noise are incorrectly biased towards one. In order to estimate $q^{*}_{ti,tj}$,set of quotients (4) is used.

$$Q_{ti,tj} := \left\{ q_g | q_g := \frac{m_{g,k_{ti}}^{*} - b_{g,k_{ti}}}{m_{g,k_{tj}}^{*} - b_{g,k_{tj}}} \text{ }, g \ \varepsilon \ G_{ti,tj} \right\} \qquad (4)$$

The following are the advantages of using FCM along with centralization preprocessing technique

(1) A single gene can be assigned to more than one cluster. This helps for the co regulation process of gene expression data.

(2) PCOS data sets generate three types of clusters out of which one has to be down regulated, second one has to be up regulated and the last cluster describes the gene is not affected of PCOS as it can neither up nor down regulated.

Before the proposed method combination of centralization and FCM can be applied, a similarity matrix or distance matrix $f_{similar}$ is computed based on correlation coefficient. Then combination of centralization and FCM algorithm can cluster the genes according to the similarity matrix $f_{similar}$.

The combination of centralization and FCM algorithm is composed of the following steps

```
Algorithm:  Combination of centralization and FCM
Input   :  Time series gene expression matrix
Output :  Clusters to be up regulated, down regulated and normal (neither up nor down regulated).
Step 1: Compute the similarity or distance matrix $f_{similar}$
C=1;
For each gene $g_{ti}$
C++;
TI$^{th}$ gene is placed in cluster c;
For each gene tj<>ti
Compute the similarity ti$^{th}$ gene with tj$^{th}$ gene $f_{similar}$ (ti,tj) using correlation coefficient metric
If $f_{similar}$ (ti,tj)>threshold $\infty$ place tj in cluster c
End;
End;
Step 2: For each gene $g_{ti}$ estimate $q^*_{ti,tj}$ with the set of quotient

$Q_{ti,tj} := \left\{ q_g \middle| q_g := \frac{m^*_{g,k_i} - b_{g,k_i}}{m^*_{g,k_j} - b_{g,k_j}}, g \varepsilon G_{ti,tj} \right\}$

Step 3: Calculate the median $Q_{ti,tj}$ for the c clusters.
Step 4: Assign each data D to the down regulated or up regulated by finding the difference in its distance from the cluster centroid pairs $a_i$ and $a_j$:

$[\ d\ (D-a_{ti}) - d\ (D-a_{tj})\ ]$

Step 5: If the distance is less than some threshold $\infty$, $X_D$ is in down regulation and if it is greater than some threshold $\infty$, $X_D$ is in up regulation else it is in normal.
Step 6: Compute new median for each cluster C and iterate until there are no tasks.
```

The algorithm generates the clusters based on the internal connection between time points. It also preserves the time locality in the time course gene expression data. Using the correlation coefficient as the metric, the algorithm finds the possible number of clusters and the distance matrix based on the similarity between genes. Genes that are more similar are put in the same cluster. Each object is assigned either to be up regulated, down regulated or normal cluster based on the threshold value. The median of each cluster then taken as the centroid (pair) of that cluster.

VI. EVALUATION

Performance evaluation is done for the proposed approach using PCOS data sets. The proposed algorithm is compared with the FCM. The table 1 provides the complete details about the cluster structure, clustering patterns for FCM & CCFCM. The results show that the quality of clusters and computational time is better in the proposed CCFCM algorithm.

TABLE I EXPERIMENTAL REULTS FOR PCOS DATA SETS

Algorithm	**No. of Clusters**	**Computational Time**	
		O (n log n)	**O(n^2)**
FCM	2	0.7	0.4
Proposed approach CCFCM (combination of centralization technique with FCM)	3	0.9	0.6

In fig.2 only two clusters have been obtained. If the vale is -1 then the genes in that cluster has to be up regulated. In fig.3 three clusters have been obtained. The clusters lying in the 0 value need to be neither up nor need to be down regulated.

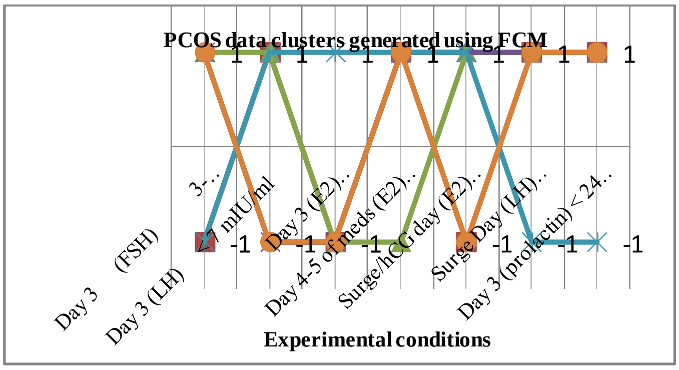

Fig. 2.Clustering using FCM

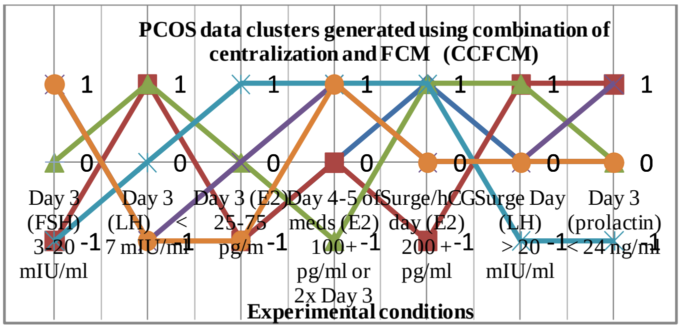

Fig. 3.Clustering using centralization and FCM (CCFCM)

VII. CONCLUSION

There are numerous algorithms for time series gene expression data have been proposed by taking various factors in to consideration. Every factor varies based on the biological conditions and the data sets. The proposed work deterministically finds quality clusters based on the experimental conditions. The main aspect of this algorithm is the consideration of internal connection between time points by preserving the time locality of time course gene expression data. To evaluate the performance of this proposed algorithm PCOS data sets have been used. The efficiency of the algorithm is proved by comparing the CCFCM with FCM algorithm.

VIII. FUTURE WORK

The efficiency of the proposed method can be enhanced further by implementing rough set theory and bi clustering algorithms. Enhancement in the efficiency of the proposed method can be tested by comparing the implemented algorithms. Also, the combinations of computational intelligence approaches can be implemented and tested.

REFERENCES

[1] SongulAlbayrak and Faith Amasyali, “Fuzzy C means clustering on medical diagonastic systems”, International XII.Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN 2003.

[2] http://www.pcosjournal.com/hormone-levels-ranges-and-what-they-mean

[3] http://www.womenshealth.gov/publications/our-publications/fact-sheet/polycystic-ovary-syndrome.html

[4] http://www.diabetes.org/living-with-diabetes/treatment-and-care/women/polycystic-ovarian-syndrome.html

[5] Ming-Chuan Hung and Don-Lin Yang, “An Efficient Fuzzy C-Means Clustering Algorithm”, IEEE, pg. no. 225 to 232, 2001.

[6] Alexander Zien, Thomas Aigner, Ralf Zimmer and Thomas Lengauer, “Centralization: a new method for the normalization of gene expression data”, Oxford University Press, Vol. 17 Suppl. 1 2001, Pages S323-S331.

[7] Kerstin Koch, Stefan Sch¨onauer, Ivy Jansen, Jan van den Bussche, Tomasz Burzykowski, “Finding Clusters of Positive and Negative Coregulated Genes in Gene Expression Data”, IEEE, pg. no. 93-99, 2007.

[8] Timothy C. Havens, James C. Bezdek, ChristopherLeckie, Lawrence O. Hall, and MarimuthuPalaniswami, “Fuzzy c-Means Algorithms for Very Large Data” IEEE Transactions On Fuzzy Systems, vol. 20, no. 6, December 2012.

[9] TurgayCelik and HweeKuan Lee, “Comments on “A Robust Fuzzy Local Information C-Means Clustering Algorithm”, IEEE transactions on image processing, vol. 22, no. 3, March 2013.

[10] OrnellaCominetti, AnastasiosMatzavinos, SandhyaSamarasinghe, Don Kulasiri, Sijia Liu, Philip K. Maini and RadekErban, “DifFUZZY: A fuzzy clustering algorithm for complex data sets”, International Journal of Computational Intelligence in Bioinformatics and Systems Biology 1(4) pp. 402-417, 2010.

[11] Famili, F., Liu, Z., Ouyang, J., Walker, P.R., Smith, B., O'Connor, M., Lenferink, A, “A Novel Data Mining Technique for Gene Identification in Time-Series Gene Expression Data”, published at the 16th European Conference on Artificial Intelligence (ECAI 2004), August 22-27, 2004. Valencia, Spain. NRC 47142.

[12] Wentao Zhao, ErchinSerpedin and Edward R. Dougherty, “Spectral Preprocessing for Clustering Time-Series Gene Expressions”, Hindawi Publishing Corporation, EURASIP Journal on Bioinformatics and Systems Biology, Volume 2009, Article ID 713248, 10 pages, doi:10.1155/2009/713248.

[13] Yuhai Zhao, Jeffrey Xu Yu, Guoren Wang, Lei Chen, Bin Wang, and Ge Yu, “Maximal Subspace Coregulated Gene Clustering”, IEEE transactions on knowledge and data engineering, vol. 20, no. 1, pg. no. 83-98, JANUARY 2008.

[14] Krzysztof Polanski, Johanna Rhodes,Claire Hill, Peijun Zhang, Dafyd J. Jenkins, Steven J. Kiddle, Aleksey Jironkin, Jim Beynon, Vicky Buchanan-Wollaston, SaschaOtt and Katherine J. Denby, “Wigwams: identifying gene modules co-regulated across multiple biological conditions”, BIOINFORMATICS, 2014, pages 1–9.

[15] Swarup Roy1*, Dhruba K Bhattacharyya2, Jugal K Kalita3, “Reconstruction of gene co-expression network from microarray data using local expressionpatterns”, Roy et al. BMC Bioinformatics 2014, 15(Suppl 7):S10.

[16] Martina Bremer1 andR. W. Doerge2, “The KM-Algorithm Identifies Regulated Genes in Time Series Expression Data”, Hindawi Publishing Corporation, Advances in Bioinformatics, Volume 2009, Article ID 284251, 10 pages.

[17] Attila Gyenesei, Ulrich Wagner, Simon Barkow-Oesterreicher, EtzardStolte and Ralph Schlapbach, “Mining co-regulated gene profiles for the detection of

functional associations in gene expression data", BIOINFORMATICS, Vol. 23 no. 15 2007, pages 1927–1935.

[18] Anindya Bhattacharya and Rajat K. De, "Bi-correlation clustering algorithm for determining a set of co-regulated genes" BIOINFORMATICS, Vol. 25 no. 21 2009, pages 2795–2801.

[19] Ya Zhang1, Hongyuan Zha2, James Z. Wang3, Chao-Hsien Chu4, "Gene Co-regulation vs. Co-expression", The Pennsylvania State University, University Park, PA 16802.

[20] http://inside.mines.edu/~ckarlsso/mining_portfolio/similarity.html.