Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 95 (2016) 40842-40844

Elixir ISSN: 2229-712X

Correlation Based Grouping and Ranking of Genes in Bipolar Disorder

K.Siva Sakthi[1], Dr.V.Jayaraj[2] and Dr.L.Nagarajan[3]

[1]Assistant Professor, Department of Computer Science, Adaikala Matha College, Vallam.

[2]Associate Professor, School of Computer Science and Engineering, Bharathidasan University.

[3]Professor and Head, Department of Computer Science, Adaikala Matha College, Vallam.

ARTICLE INFO

Article history:

Received: 5 May 2016;

Received in revised form:

2 June 2016;

Accepted: 7 June 2016;

Keywords

Bipolardisorder,

Genetics,

Mania,

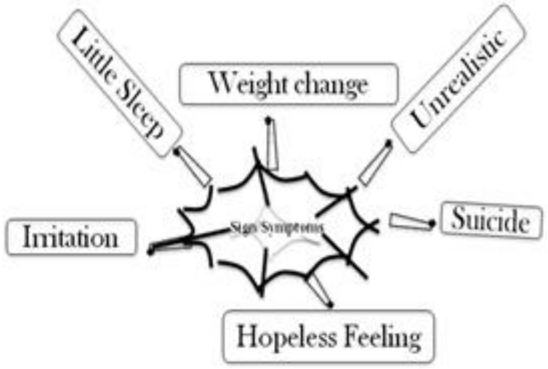
Orbitofrontal cortex.

ABSTRACT

Bipolar Disorder is a psychiatric disorder in which the core feature is pathological disturbance in mood rearranging from extreme mania to severe depression. Research analysis shows there is no specific genes are analyze which is directly associated with Bipolar Disorder. Diverse data mining technique is used to find out genes responsible for BD. It is found that genes present in orbit frontal cortex are differentially expressed in disease and control subjects. Gene Expression analyses of thousands of genes are studied from the orbit frontal cortex region of brain. Microarray technology is used to study approximately 22,283 mRNA transcripts from Orbit frontal cortex in brain region. Statistical t-tests are applied and 1,172 genes have obtained as significant genes. Correlation based feature selection is used to reduce the large dataset into small dataset.

Introduction

Pathophysiology of Bipolar Disorder also called manic-depression is still unknown however it is found that Bipolar Disorder has strong genetics component. Bipolar Disorder Patients characterized by two types of episodes of mania and Depression. Frequency of these two types of episode may vary from one day to one month. The patients has tendency to suicide in extreme depression whereas they feel fresh even after two hour sleep.



t-Test:

The t-test is commonly used to determine whether the mean of a population differs from known population.

There are three types of t-test

- One-sample t-test
- Paired sample t-test
- Independent sample t-test

One-sample t-Test:

One-sample t-test is used to compare mean value of a sample with known population mean. The one-sample t-test is used only for tests of the sample mean. Thus, our hypothesis tests whether the average of our sample (M) suggests that our students come from a population with a population.

In one sample t-test, we know the population mean. We draw a random sample from the population and then compare the sample mean with the population mean and make a statistical decision as to whether or not the sample mean is different from the population mean. We can use this analysis, for example, when we take a sample from the city and we know the mean of the country (population mean). If we want to know whether the city mean differs from the country mean, we will use the one sample *t*-test.

Steps

1.Set up the hypothesis:

A. Null hypothesis: assumes that there are no significance differences between the population mean and the sample mean.

B. Alternative hypothesis: assumes that there is a significant difference between the population mean and the sample mean.

C. Calculate the standard deviation for the sample by using this formula:

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Where,

S = Standard deviation

$\overline{X}$ = Sample mean

n = number of observations in sample

2. Calculate the value of the one sample t-test, by using this formula:

$$t = \frac{\overline{X} - \mu}{S}\sqrt{n}$$

Where,

t = one sample t-test value

μ = population mean

3. Calculate the degree of freedom by using this formula

Tele:
E-mail address: shaku.ngrkl@gmail.com

V=n-1

Where,

V= degree of freedom

4. Hypothesis testing: In hypothesis testing, statistical decisions are made to decide whether or not the population mean and the sample mean are different. In hypothesis testing, we will compare the calculated value with the table value. If the calculated value is greater than the table value, then we will reject the null hypothesis, and accept the alternative hypothesis.

Paired sample t-Test

Paired sample t-test is used to compare two means that are repeated measure of the same subject. Paired sample t-test is a statistical technique that is used to compare two population means in the case of two samples that are correlated. Paired sample t-test is used in before-after studies, or when the samples the matched pairs (or) when it is a case control study.

Steps

1. Set up hypothesis: We set up two hypotheses. The first is the null hypothesis, which assumes that the mean of two paired samples are equal. The second hypothesis will be an alternative hypothesis, which assumes that the means of two paired samples are not equal

2. Select the level of significance**:** After making the hypothesis, we choose the level of significance. In most of the cases, significance level is 5%, (in medicine, the significance level is set at 1%).

Following formula:

$$t = \frac{\bar{d}}{\sqrt{s^2/n}}$$

Where d bar is the mean difference between two samples, s² is the sample variance, n is the sample size and t is a paired sample t-test with n-1 degrees of freedom. An alternate formula for paired sample t-test is:

$$t = \frac{\Sigma d}{\sqrt{\dfrac{n\left(\Sigma d^2\right)-\left(\Sigma d\right)^2}{n-1}}}$$

3. Testing of hypothesis or decision making: After calculating the parameter, we will compare the calculated value with the table value. If the calculated value is greater than the table value, then we will reject the null hypothesis for the paired sample t-test. If the calculated value is less than the table value, then we will accept the null hypothesis and say that there is no significant mean difference between the two paired samples.

Independent sample t-test:

Independent sample t-test is used to compare the two mean from independent group. The independent samples t-test compares the means of two independent groups in order to determine whether there is statistical evidence that associated population means are significantly different. The independent samples t-test is a parametric test.

The unpaired t method tests the null hypothesis that the population means related to two independent, random samples from an approximately normal distribution are equal.

Assuming equal variances, the test statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{j=1}^{n_1}(x_j-\bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_i-\bar{x}_2)^2}{n_1+n_2-2}$$

- Where x bar 1 and x bar 2 are the sample means, s² is the pooled sample variance; n_1 and n_2 are the sample sizes

Assuming unequal variances, the test statistic is calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}$$

$$df = \frac{\left[\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right]^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1-1}+\dfrac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

$$s_1^2 = \frac{\sum_{j=1}^{n_1}(x_j-\bar{x}_1)^2}{n_1-1}$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2}(x_j-\bar{x}_2)^2}{n_2-1}$$

- Where x bar 1 and x bar 2 are the sample means, s² is the sample variance; n_1 and n_2 are the sample sizes.

P-Value

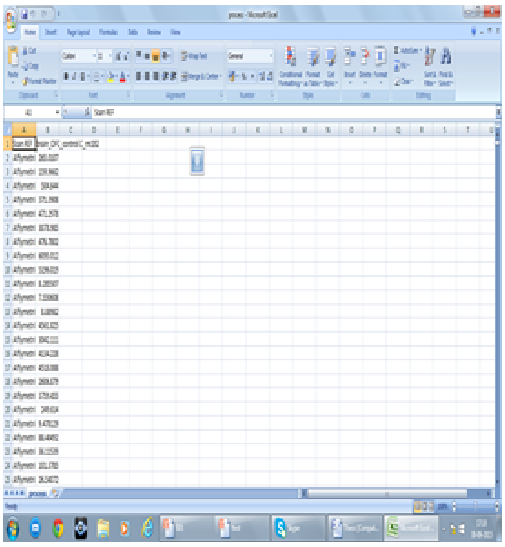
The P value is the level of marginal significance with in a statistical hypothesis test representing the probability of the occurrence of a given event.

The **P** value is used as an alternative to rejection points to provide the smallest level of significance at which null hypothesis would be rejected.

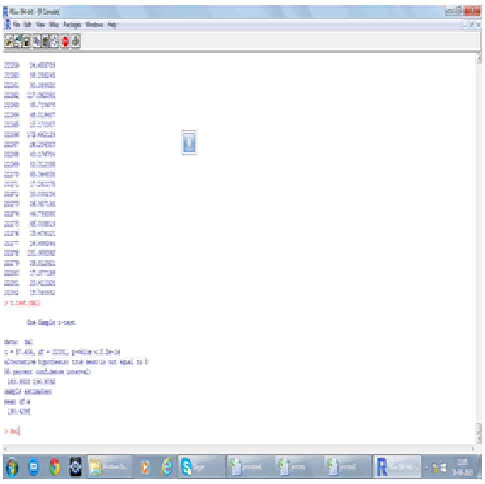
The *p*-value is a number between 0 and 1 and interpreted in the following way:

- A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- *p*-values very close to the cut-off (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

Sample Dataset



t-Test Result



```
> t.test([illegible])

        One Sample t-test

data:  [illegible]
t = 37.406, df = 22282, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 [illegible]
sample estimates:
mean of x
 [illegible]
```

It gives output for those genes whose p values are less than 0.05. That particular genes are consider as significant genes.

Conclusion

Microarray technology is used to study approximately 22,283 mRNA transcripts from Orbitofrontal cortex in brain region. Statistical t-tests are applied and 1,172 genes have obtained as significant genes. Correlation based feature selection is used to reduce the large dataset into small dataset. My feature work is to classify the number of normal, disease gene set and find out the **most** significant genes which plays a major role in causing Bipolar disorder and also find out the genes which are co-occurring with significant genes which causing disease.

Reference

- ➢Muhleisen, T. W. et al. Genome-wide association study reveals two new risk loci for bipolar disorder. Nature Communications, doi: 10.1038/ncomms4339, 12 March 2014.
- ➢Xu, W. et al. Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. BMC Medical Genetics, doi: 10.1186/1471-2350-15-2,4 January 2014.
- ➢Durairaj M. et al. "Data Mining Applications in Healthcare Sector: A Study", International Journal of Scientific and Technology Research, Vol. 2, ISSN 2277-8616.2013.
- ➢Hussan D. "Data Mining based Prediction of Medical data using K-means algorithm", Basrah Journal of Science(A), Vol. 30(1), 46-56. 2013.
- ➢Jain N. et al. "Data Mining Techniques: A Survey Paper", International Journal of Research in Engineering and Technology, Volume: 02, Issue: 11, eISSN: 2319-1163 | pISSN: 2321-7308. 2013.
- ➢Kharya S. "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", International Journal of Computer Science, Engineering and InformationTechnology (IJCSEIT), Vol. 2, No. 2. 2012.
- ➢Mittal P. et al. "Study and Analysis of Predictive Data Mining Approaches for Clinical Dataset", International Journal of Computer Applications, Volume 63, No. 3. 2013.
- ➢Naib M. et al. "Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining", International Journal of Computer Applications (0975 – 8887), Volume 96, No. 8. 2014.
- ➢M .Anandhavalli, M.K.Ghose, K.Gauthaman, Association Rule Mining in Genomics *International Journal of Computer Theory and Engineering 2010*.
- ➢Howard J. Edenberg, Daniel L. Koller, Xiaoling Xuei, Genome-Wide Association Study of Alcohol Dependence Implicates a Region on Chromosome 11 *ISBRA 2010*.
- ➢Jiawei Han and Micheline Kamber, Data Mining Concepts and Technique *Morgan Kufmann Publisher 2010*.
- ➢DovStekel,Microarray Bioinformatics, *Cambridge University Press 2008*.
- ➢GoodwinKaj.,KR manic. *DepressiveIllness*. New York, NY: Oxford University Press. 1990.