World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques): 53-59, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.29.dmsct.10

# Visualization of Chemical Space Using Principal Component Analysis

*B. Firdaus Begam and J. Satheesh Kumar*

Department of Computer Applications, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India

**Abstract:** Principal component analysis is one of the most widely used multivariate methods to visualize chemical space in new dimension by the chemist for analysing data. In Multivariate data analysis, the relationship between two variables with more number of characteristics can be considered. PCA provides a compact view of variation in chemical data matrix which helps in creating better Quantitative Structure Activity Relationship (QSAR) model. It highlights the dominating pattern in the matrix through principal component and graphical representation. This paper focuses on mathematical aspects of principal components and role of PCA on Maybridge dataset to identify dominating hidden patterns of drug likeness based on Lipinski RO5.

**Key words:** Principal Component Analysis · Load p lot · Score plot · Biplot

## INTRODUCTION

Principal Component Analysis (PCA) is a multivariate statistical approach to analyze data in lower dimensional space. PCA used vector space transformation technique to view datasets from higher-dimensional space to lower dimensional space. PCA is an application of linear algebra [2] which was first coined by Karl Pearson during 1901 [1]. PCA has been rediscovered in many diverse scientific fields by Fischer and MAcKenzie [2], Wolf [3] and Hoteling [4]. In 1960, PCA has been taken by Malinowski for chemical applications and later by many chemists [5]. PCA is one of the multivariate methods and it is a member of multidimensional factorial methods [6].

According to Jolliffe, "The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated and which are ordered so that the first few retain most of the variation present in all of the original variables" [7]. Goals of Principal component analysis are simplification, prediction, redundancy removal, feature extraction, un-mixing, data compression and other related areas. PCA gives a simplified view of larger datasets, by building a new model of selected objects and variables with minimal loss of information. It can also be used for prediction model as PCA acts as exploratory tool for data analysis by calculating the variance among the variables which are uncorrelated [8].

Chemical space (drug space) is defined as number of descriptors calculated for each molecule and stored in multidimensional space. Visualizing chemical space through lower dimensional space based on principal components [9]. Analysing the space to identify hidden or dominating patterns of drug-likeness molecules are done effectively by applying PCA method.

**Importance of PCA:** Data analysis through bivariate analysis provides the correlation among two variables/descriptors $X_i$ and $Y_i$. The pair wise correlations between descriptors are represented by Pearson correlation coefficient (r) lies between -1 and +1 as in eq (1). The degree of dependency or redundancy is analysed through the range. Correlation coefficient value +1 represents that the variables are positively correlated and correlation coefficient -1 represents that the variables are negatively correlated and zero coefficient represents that they are not correlated. The correlations among the variables are represented through scatter plot [10].

**Corresponding Author:** B. Firdaus Begam, Department of Computer Applications,
School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India.

Bivariate analysis is feasible for smaller datasets, but for larger datasets it is not suggested. To overcome the drawback of bivariate analysis on large datasets, multivariate analysis is used [8]. PCA is one of the multivariate methods and it is a member of multidimensional factorial methods [6].

$$\gamma = \frac{\sum_{k=1}^{n} [(x_i - (\bar{x}))(y_i - (\bar{y}))]}{\sqrt{\sum_{k=1}^{n} [(x_i - (\bar{x}))^2 (y_i - (\bar{y}))^2]}} \quad (1)$$

where, $\bar{x}$ and $\bar{y}$ represents the mean vector of $x$ and $y$ variables / descriptors.

Organization of the paper is follows, section 2 discuss about the principal component analysis and interpretation of principal components, section 3 describes materials and methods used for analysing drug-likeness dataset by decomposing principal components. Principal component analysis through load, score and biplot are discussed in results and discussion section 4. Section 5 concludes the paper.

**Principal Component Analysis:** The data matrix, $X_{m \times n}$ transformed as a product of *Tand P'* by applying PCA where *m* represents number of objects / molecules and *n* represents number of variables / descriptors. The data element $X_{ij}$ represents the data in the matrix *X* at the $i^{th}$ row and $j^{th}$ column. The output, *T and P'* are scalar vector of row and columns which gives various patterns hidden in the datasets [11, 13]. Dominating object patterns are interpreted by score plot over row values of *T* and dominating variables patterns are interperted by load plot over column values of *P'* as shown in figure 1 and both the vectors are orthogonal. PCA in matrix form is the least square model which is represented as given below equation (2).

$$X_{m \times n} = 1.\bar{X} + T_{m \times q}.P'n \times q + E \quad (2)$$

where, $P_i' P_j = 0 and\ T_i' T_j = 0\ for\ i \neq j.E$ represents the residue or noise in the data. $\bar{X}$ is mean vector.

Based on least square model, scores *T* can be viewed as the linear combination of the data with coefficients *P'*. Loadings *P'* can also be viewed as the linear combinations of data with coefficient *T*. PCA method is referred as called bilinear model (BLM). PCA is used to deal with set of descriptors based on their variances and covariance for internal analysis of data [12]. PCA analysis can be interpreted by identifying principal

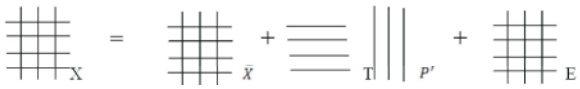


Fig. 1: Decomposition of PCA principal

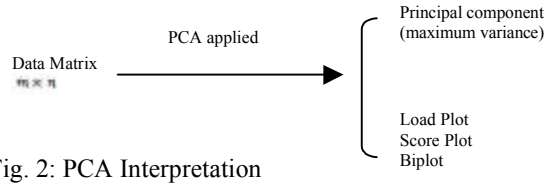


Fig. 2: PCA Interpretation

component and through graphical visualization as shown in figure 2. Opera *et al*. (2000) coined new name for navigating chemical space through PCA called Chemography [13].

Application of PCA can be extended for Cheminformatics for better prediction and analysis [13]. This research study focuses on role of PCA for efficient visualization of dominant patterns in chemical space [14].

## MATERIALS AND METHODS

**Data Sets:** Maybridge Drug likeness dataset contains 14,400 molecules based on highly popular method "rule of five" (RO5) proposed by Lipinski et al [16-17]. Rule of five is based on descriptors, molecular weight (MW), calculated octanol - water partition coefficient also called lipophilicity (ClogP - Calculated logP), hydrogen donors (HBD) and hydrogen acceptors (HBA) which has been widely applied to distinguish or predict drug-like and non-drug like chemical agent. Along with these descriptors rotatable bonds, flexibility of the molecule and polar surface area (PSA) descriptors are also considered. Visualizing the chemical space based on certain descriptors which are correlated in lower dimensional space is called Principal compound analysis [18-19].

The distributions of data based on each descriptor are represented in figure 3. As per Lipinski rule, molecular weight should be less than 500 Daltons (figure 3a), Calculated logP - ClogP descriptor not more than 5 number (figure 3b), hydrogen bond donor not more than 5 (figure 3c) and hydrogen bond acceptor not more than 10 bonds (figure 3d). Drug likeness of Maybridge dataset is analysed through PCA approach by implementing in MATLAB environment.

**Data Pre-Processing:** Pre-processing of the dataset is necessary to normalize or standardize the data for further processing. First, lengths of dataset vector are checked

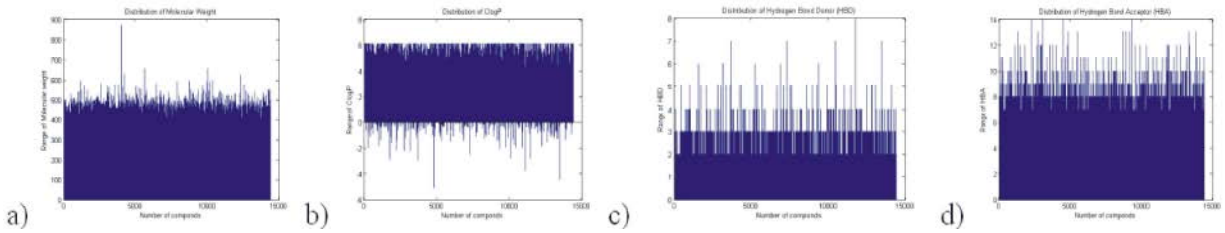


Fig. 3: Distribution of compounds based on a) Molecular Weight b) ClogP c) Hydrogen Bond Donor d) Hydrogen Bond Acceptor.

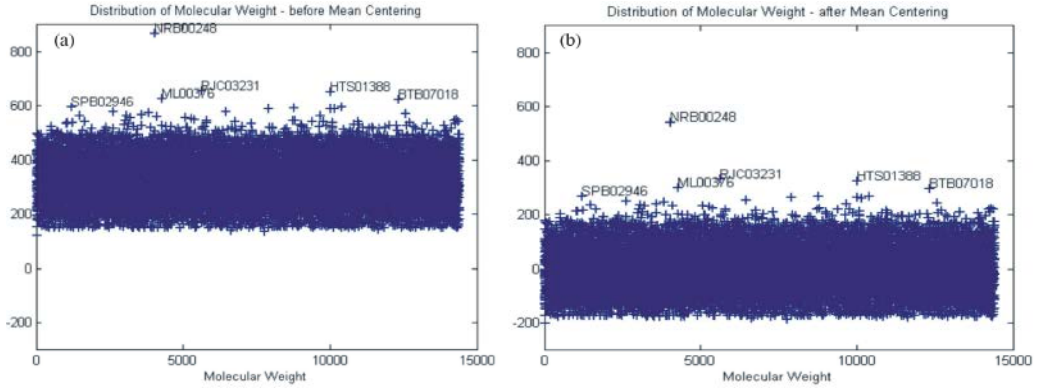


Fig. 4: a) Distribution of Molecular weight descriptor before and b) after mean centring. (Variables/Maybridge codes are marked in MATLAB environment).

whether they are of equal lengths or any data is missed. This is called holes in the data. It can be filled in two ways, one way is to guess the value based on knowledge from objects and variables exist in data set as it does not affect much for PC analysis. Another way is to analyse the data set with holes as PC algorithm can handle missing values [6].

Mean-centring scaling method is applied over dataset to standardize the data. Mean - centring approach is to standardize the deviation in the variables (columns) as it is collected under various conditions [6, 10]. The mean is to be calculated by sum of each variable and divide by number of columns of the variable (eq 3). The mean vector is subtracted from each variable (columns) which translates dimension of the dataset (eq 4) to be centred to zero (refer figure 4a and 4b).

$$X_i' = \sum_{i=1}^{n} \frac{x_i}{n} \tag{3}$$

$$X_{MCV} = \sum_{i=1}^{n} X_i - X_i' \tag{4}$$

where $X_i'$ represents mean of each column, $X_i$ represents $i^{th}$ element of data matrix $X$ and $n$ denotes number of elements in data matrix and MCV represents Mean Centred Vector.

**Generate Covariance Matrix:** The covariance matrix and correlation matrix is calculated to scale the matrix so that data with high variance are shrinked with low variance and vice versa. This approach has translated the data matrix or dataset with normalized variance (eq 5). The size of covariance matrix is $n \times n$. The covariance matrix © is symmetric as, *cov(x,y)=cov(y,x)* [11]. The correlation of un-standardized matrix is identical with covariance value of each pair of standardized variables [6].

$$C = \frac{1}{n}(X.X^T) \tag{5}$$

where $X^T$ represents transpose of data matrix $X$.

**Principal Component:** The next task of PCA is to identify new set of data model with orthogonal axes. This is done by calculating Eigen values and Eigen vectors (eq 6) from the symmetric covariance matrix which forms multidimensional rotation [15]. Eigen vectors are coefficients of principal components. The diagonal values of Eigen vector have been compared.

If some Eigen values are very small (zero) corresponding Eigen values and vectors are discarded which reduce the dimensionality on new basis [24]. This is done only when there is cost connected to the variable. For analysis all the values are considered. Eigen vectors ranked and based on which Eigen values are sorted. For $n \times n$ covariance matrix, there exist $n$ real Eigen values. The reason for ranking is that, largest Eigen values and vectors retains the important or useful information of dataset and remaining ones comprises noises in data matrix.

$$[PC \quad EV] = Eig(C) \tag{6}$$

where PC represents principal component (Eigen Value $eig_i$) and EV represents Eigen vector of covariance matrix C.

Eigen values are also called characteristics roots of dataset. This principal component visualizes the variance of each variable in the dataset called feature vector (eq 7). Eigen vectors provide the weights to compute the uncorrelated principal components. Component vector is constructed by taking on Eigen values. Eigen vectors are perpendicular so that data matrix can be interpreted in new perpendicular axis instead on X and Y axis as shown in Load plot.

$$Principal\ component = (eig_1, eig_2, ....eig_n) \tag{7}$$

**Features of Principal Component:** Principal components are calculated based on total variance of the dataset. First principal component represents the maximum total variances of variables. The second principal component is the next maximal total variances of variables compared to first principal component [16, 25].

$$factor\ score = X.\ PC \tag{8}$$

In principal component analysis each component is calculated based on maximum amount of variance which is not accounted by previous component. PC calculated for 7 descriptors in Maybridge dataset, variance and cumulative variance among PC are show in figure 5 and 6. The analysis results in degree of correlation of components are visualized, which are totally uncorrelated with each other. The data matrix is projected into new dimension by calculating factor scores as in equation (8) and geometrical interpreted through score plot as discussed in section 4.

| | EValue | %Variance | CVar |
|---|---|---|---|
| PC1 | 6707.21486 | 90.70742 | 90.70742 |
| PC2 | 680.71338 | 9.20587 | 99.91329 |
| PC3 | 4.11660 | 0.05567 | 99.96896 |
| PC4 | 0.93564 | 0.01265 | 99.98161 |
| PC5 | 0.76312 | 0.01032 | 99.99193 |
| PC6 | 0.32200 | 0.00435 | 99.99629 |
| PC7 | 0.27451 | 0.00371 | 100.00000 |

Fig. 5: Eigenvalues and variance of the principal component axes

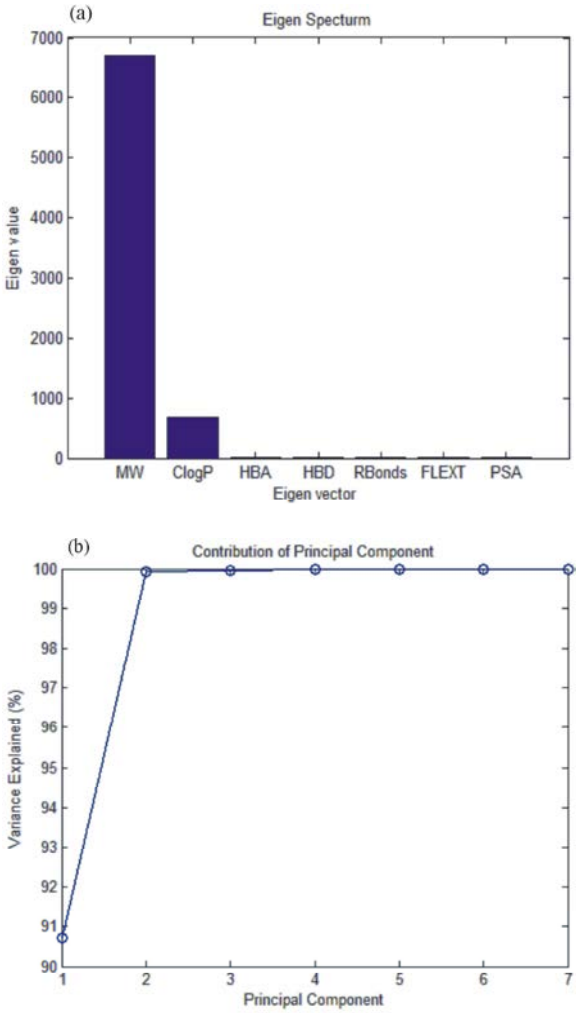


Fig. 6: a) Eigen Specturm b) Distribution of Variance of each PC

## RESULTS AND DISCUSSIONS

Principal components are very useful in reducing dimensionality and visualize various patterns of data hidden in datasets. PC provides valuable information of the dataset in new orthogonal axis to represent the data.

PCA projects the data along the direction of maximum dispersion of dataset. The direction of variance is determined by calculating Eigen vectors and Eigen values of covariance matrix. The geometrical representation of principal components shows the degree of variance through correlation [10] by Load, Score and Biplot.

Load plot depicts the interrelation among the variables / descriptors based on principal components. Score plot is used to analyse the similar patterns exists between the observation based on principal component [19-20]. Biplot visualize inter-relationships (variance and correlation) between the observations and variables are visualized through biplot in two-dimension space.

Load Plot: The load plot shows how much each variable contributes to each PC. It is used to identify which variable cause to be outliner and which variables are responsible for [17] classification. The x-axis in load plot denotes the coefficient for all variables making up PC1 and y-axis denotes the coefficient for all variables making up PC2.

The load plot shows the larger variance in molecular weight compared to other descriptors. Higher prediction in principal component (PC1) primarily depends on size parameters [18]. Geometrical representation of dependency of descriptors over PC is interpreted through load plot as shown in figure 7 based on PC1 and PC2. It shows that molecular weight is located on first PC. ClogP also contributes to right side of first PC. Other variables like, Hydrogen bond donor, Hydrogen bond acceptor, Rotatable bonds, flexibility and polar surface area are contributed to first PC but also to second PC as they are located near to the zero axis.

**Score Plot:** The principal component score plot is based on Eigen vectors derived for each observation. The mean of PC score is zero as it is linear combination of mean-centred data variables [17]. PC scores are used to detect multivariate outliners or to check multi-colliniarity in data variables. The principal coordinates are scores of data matrix $X$ in principal component space [8, 18]. Figure 8 represents score plot of first and second principal coordinates from principal score matrix with total variance of 99.9%. It highlights which molecule depends on PC1 and PC2. The molecules which depend on both PC1 and PC2 can be visualized through score plot. Score plot based on first few principal components will depict the most dominating patterns in dataset.

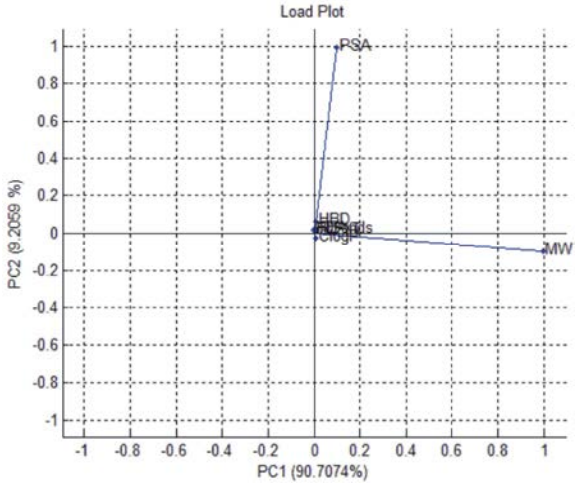


Fig. 7: Load plot of Principal component correlation with 90.70+9.20=99.90% of variance

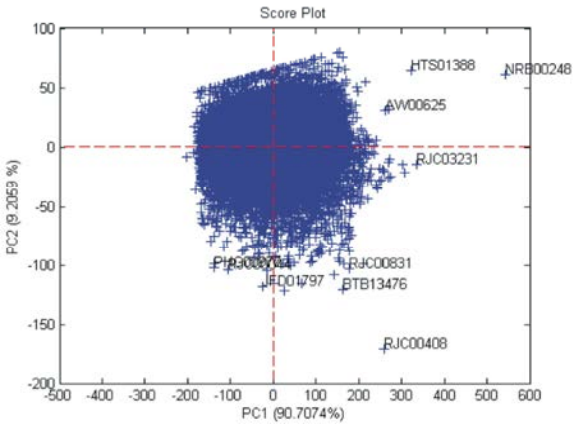


Fig. 8: Geometrical representation of score matrix based on PC1 and PC2 with 99.9% of variance

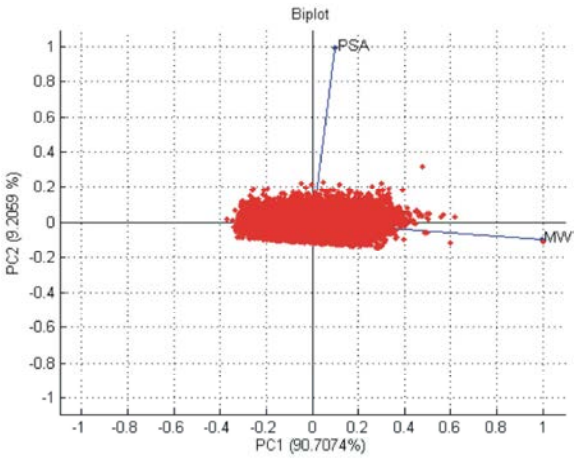


Fig. 9: Two dimensional Biplot of Maybridge drug likeness dataset based on PC1 and PC2.

**Biplot:** Biplots are the multivariate analogue of scatter plots which was defined by Gabriel during 1971 [21]. The biplot is a graphical representation which allows rapid visualization of the structure of the matrix [22]. The relationship between score and loading associated with any two principal components can be interpreted through biplot. In other words, biplot is superimposing score and load factors to interpret complete equivalence between variables and observations [23]. In data which have similar properties are grouped to the zero axes which indicate drug-likeness in dataset as shown in figure 9. The horizontal axis represents PC1 and vertical axis represents PC2. Drug molecules projected based on PC1 and PC2 are represented as red balls (rows). Blue lines (columns) show projection of descriptors based on PCs. The distance and cosine between red balls and blue line represents correlation among the two vectors (drug molecule and descriptors).

Load, Score and Biplot represents variance of principal components which highlights dominant or hidden patterns in dataset. The visualization of chemical space through geometrical representation results in better understanding.

## CONCLUSION

Principal component acts as a visualization tool for quantitative and qualitative analysis of huge compound libraries in a reduced dimensional space. This approach has been applied to compare or classify drug agents, natural products and combinatorial libraries. Principal component with higher variation and correlation helps in identifying dominant patterns in the dataset. This research work explains detailed study on PCA and application of PCA on chemical space. This paper also discusses efficient visualization of molecular descriptor with respect to principal components. Geometrical representations through Load, Score and Biplot have projected for better understanding of dominating patterns in chemical space. Also Load, score and biplot highlights diverse samples in lower dimension.

## REFERENCE


1. Pearson, K., 1901. On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 6(2): 559-572.
2. Fisher, R. and W. MacKenzie, 1923. Studies in crop variation. II. The manurial response of different potato varieties, Journal of Agricultural Science, 13: 311-320.
3. Wold, H., 1966. Nonlinear estimation by iterative least squares procedures, in F. David (Editor), Research Papers in Statistics, Wiley, New York, pp: 411-444.
4. Hotelling, H., 1933. Analysis of a complex of statistical variables into principal moments, Journal of Education Psychology, 24: 417-441.
5. Malinowski, F. and D. Howery, 1980. Factor Analysis in Chemistry, Wiley, New York.
6. Christophe, B.Y., Cordella, PCA: The Basic Building Block of Chemometrics, http://dx.doi.org/ 10.5772/51429
7. Jolliffe, I.T., 2002. Principal Component Anslysis, Second edition, Springer Series in Statistics, ISBN- 0-387-95442.
8. Svante Wold, Kim Esbensen and Paul Geladi, 1987. Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems, 2: 37-52.
9. Lakshmi B. Akella and David DeCaprio, 2010. Cheminformatics approaches to analyse diversity in compound screening libraries, current opinion in Chemical Biology, 4: 325-330.
10. Andrew R. Leach and Valerie J. Gillet, 2009. An Introduction to Cheminformatics, Springer.
11. Jürgen Bajorath, Chemoinformatics Concepts, Methods and Tools for Drug Discovery, Methods in Molecular Biology, pp: 275.
12. Josefin Rosen anders Lovgren, Thierry Kogej Sorel Muresan, Johan Gottfries and Anders Backlund, 2008. ChemGPS-NPWeb: chemical space navigation online, J Comput Aided Mol Des, Springer Science+Business Media B.V. DOI 10.1007/s10822-008-9255-y.
13. Stephen J. Haggarty, Paul A. Clemons, Jason C. Wong and Stuart L. Schreiber, 2004. Mapping Chemical Space Using Molecular Descriptors and Chemical Genetics: Deacetylase Inhibitors, Combinatorial Chemistry & High Throughput Screening, 7: 669-676.
14. Jean-Louis Reymond, Ruud van Deursen, Lorenz C. Blum and Lars Ruddigkeit, 2010. Chemical space as a source for new drugs, The Royal society of chemistry, Med. Chem. Commun., 1: 30-38.

15. Firdaus B. Begam and J. Satheesh Kumar, 2012. A Study on Cheminformatics and its Applications on Modern Drug Discovery, Procedia Engineering 38: 1264-1275.
16. Patrick W. Walters and Mark A. Murcko, 2002. Prediction of 'drug-likeness', Advanced Drug Delivery Reviews, 54: 255-271.
17. Tudor I. Oprea and Johan Gottfries, 2001. Chemography: The Art of Navigating in Chemical Space, J. Comb. Chem. 3: 157-166.
18. Christopher A. Lipinski, 2004. Lead- and drug-like compounds: the rule-of-five revolution, Drug Discovery Today: Technologies| Lead Profiling, (1) 4.
19. Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy and Paul J. Feeney, 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Advanced Drug Delivery Reviews, 46: 3-26.
20. Herve Abdi and Lynne J. Williams, 2010. Principal component analysis, John Wiley & Sons, Inc, 2.
21. Gabriel, K.R., 1971. The biplot graphic display of matrices with application to pricipal component analysis. Biometrika, 58; 453.
22. Gower, J.C. and Hand, D.J., 1996. Biplots. Monographs on Statistics and Applied Probability 54. Chapman & Hall, London.
23. John C. Gower, 2003. Unified Biplot Geometry, Developments in Applied Statistics, Metodološki zvezki, 19, Ljubljana: FDV.
24. Hossein Dehghan, Hamid Hassanpour and Ali A. Pouyan, 2012. ROI Analysis Using Harvard-Oxford Atlas in Alzheimer'sDisease Diagnosis Based on PCA, Iranica Joirnal of Energy and Environment, 3(3): 255-258.
25. Olewe, O.M., A.C. Odebode, O.J. Olawuyi and A.O. Akanmu, 2013. Correlation, Principal Component Analysis and Tolerance of Maize Genotypes to Drought and Diseases in Relation to Growth Traits, American-Eurasian J. Agric. & Environ. Sci., 13911): 1554-1561.