

A Regression Model using K-Means Algorithm to Screen 32 Compound Dataset as COX-2 Inhibitors

Chunduru Madhava Rao
Research Scholar
Rayalaseema University
Kurnool
Andhra Pradesh
India

Yesu Babu Adimulam, PhD
Professor and Head
Department of CSE
Sir C R Reddy College of Engg
Eluru, Andhra Pradesh
India

ABSTRACT

COX-2 provided a new class of anti inflammatory, analgesic and antipyretic drugs with significantly reduced side effects. It has been reported that inhibiting COX-2 could also be an important strategy for preventing or treating a number of cancers. A report with modified k-means clustering algorithm to cluster groups of compounds obtained from regression analysis along with few compounds which were non-tested against COX-2 and screened them using regression model. The regression model due to its high predictive ability can be utilized as an alternative aid to the costly and time consuming experiments for recognizing and determining compounds with high COX-2 binding affinity. Hence, a group of new derivatives from literature are subjected to screening utilizing the produced model. A set of 32 compounds with pyrazole ring as main nucleus was selected from a published review paper. In this work, a modification of k-means algorithm that efficiently searches data to cluster points by computing sum of squares within each cluster which makes the program to select the most promising subset of classes for clustering. From a set of 32 compounds, only the top 5 compounds are combined with 58 molecule data set to perform cluster analysis. From the analysis it is evidenced that k-means clustering algorithm is able to group data objects of all molecules based on the 3 centroids provided and all top 5 compounds appear to be centred on one spade whereas Celecoxib appeared in another cluster.

Keywords

COX-2, Cluster analysis, k-means, phylogenetic tree

1. INTRODUCTION

Cyclooxygenase (COX) is a rate-limiting enzyme that converts arachidonic acid to prostaglandins (PGs) and exists in two isoforms, COX-1 and COX-2 [1]. Cyclooxygenase 2 (COX-2) is induced at inflammation sites and produces pro-inflammatory prostaglandins and participates in inflammation-mediated cytotoxicity. Conventional non steroidal anti-inflammatory drugs (NSAIDs) nonspecifically inhibit cyclooxygenase-1 (COX-1), an enzyme critical to normal platelet function, and COX-2 which mediates inflammatory response mechanisms. Celecoxib, an anti-arthritic agent that inhibits COX-2 but spares COX-1 at therapeutic doses, is expected to have minimal effects on platelet function [2]. COX-2 is required for both the constitutive and mitogen-induced PGE₂ synthesis. Moreover, over-expression and persistent expression of COX-2 may be influenced by breast tumor hormone status and seem to be a feature of the aggressive, metastatic phenotype [3]. Classic NSAIDs preferentially inhibit COX-1 in vitro, but assessing

their gastrointestinal (GI) safety profile is critical. New compounds with a high selectivity for COX-2, especially those that are non-acidic, may be better tolerated in the GI tract [4].

This led to the hypothesis that inhibition of COX-1 by NSAIDs resulted in side effects such as ulcers and renal failure, whereas the anti inflammatory properties result from the inhibition of the inducible COX-2 [5]. Selective inhibition of COX-2 provided a new class of anti inflammatory, analgesic, and antipyretic drugs with significantly reduced side effects. It has been reported that inhibiting COX-2 could also be an important strategy for preventing or treating a number of cancers [6] and could be used to delay or slow the clinical expression of Alzheimer's disease [7].

The discovery of two lead compounds arising from distinct chemical classes, NS-398 [8] and DuP-697 [9], has led to two general classes of selective COX-2 inhibitors, the diarylheterocycles and the methanesulfonanilides. Later, the structures of some nonselective NSAIDs such as indomethacin, zomepirac, aspirin, and flurbiprofen have been successfully converted into selective COX-2 inhibitors [10]. Further, several compounds belonging to structurally different families such as depending on the central cyclic ring such as pyrroles [11], imidazoles [12], cyclopentenones [13], pyrazole [14], spiroheptene, spiroheptadiene [15], isoxazole [16], and thiophene have been reported as selective COX-2 inhibitors. In this paper, we report a modified k-means clustering algorithm to cluster groups of compounds obtained from regression analysis along with few compounds which were non-tested against COX-2 and screened them using regression model.

2. MATERIALS AND METHODS

2.1 Data set

A set of 58 compounds with dihydropyrazole sulphonamides and triaryl pyrazoline derivatives were considered [17] [18] [19] [20] from our earlier work. The inhibitory activities of these derivatives reported in terms of IC₅₀ in micromolar were transformed into their corresponding logarithmic values in order to overcome overlapping data.

2.2 Regression Model

The best regression model obtained from multiple linear regression analysis on a set of 58 COX-2 inhibitors resulted in few influential parameters towards COX-2 inhibition. The model described nearly 12 parameters being important to consider on a set of compounds for possible evaluation as COX-2 inhibitors. The equation was given below.

$$\begin{aligned} \text{Log}(1/\text{IC}_{50}) = & +0.062886 \times \text{DIPOLE} \\ & +0.115464 \times \text{LIPOLE} \\ & +0.104411 \times \text{MR} \\ & -0.956308 \times \text{KAPPA2} \\ & +0.662139 \times \text{F} \\ & -0.039715 \times \text{CL} \\ & -0.012485 \times \text{MW} \\ & +0.892437 \times \text{HBA} \\ & +1.415335 \times \text{HBD} \\ & -0.097028 \times \text{LOGP} \\ & -0.137041 \times \text{RB} \\ & +2.025566 \times \text{BALABAN} \\ & -7.17311 \end{aligned}$$

$r = 0.884$; $r^2 = 0.782$; $\text{adj } r^2 = 0.724$; $F = 13.47$;
 $p\text{-value} = 0.85$; $f\text{-value} = 0.511$; $\text{Chi}^2 = 4.774$; $n=58$

2.3 Screening non-tested compounds

From the regression model, it was observed that designing or screening compound libraries for new compounds or analogs with possible H-bond acceptor and donor substitutions, and increase in number of Fluorine atoms and decrease in molecular weight, logp and rotatable bonds on the molecule with concomitant increase in dipole, lipole and molar refractivity would enhance inhibitory activity against COX-2.

The proposed regression model due to its high predictive ability (unpublished results) can be utilized as an alternative aid to the costly and time consuming experiments for recognizing and determining compounds with high COX-2 binding affinity. The method can also be used to screen similar repertoire of compounds reported in various literature sources or chemical compound libraries to identify new potentially active compounds.

Hence, a group of new derivatives from literature are subjected to screening utilizing the produced model. We, therefore, searched for compounds with either pyrazoles, pyrazolines or other pyrazole derivatives in Archives of organic chemistry journal [www.arkat-usa.org]. Therefore, a set of 32 compounds with pyrazole ring as main nucleus was selected from a published review paper [21]. The structures are devoid of biological activity against COX-2, however, these pyrazole compounds are known to inhibit certain disease causing protein targets, for example, FactorXa, IXa, platelet aggregation inhibition, antidepressant, voltage-gated sodium channel blockers etc. All compounds screened for this study are provided with compound numbers as given in reference article (Table 1).

Table 1: Compounds along with predicted activity data after applying regression model on this 32 compound dataset.

Molecule No.	Predicted Activity (\log_1/IC_{50})	Calculated $\text{IC}_{50}(\mu\text{M})$
28.mol	-0.6172	4.1420
35.mol	-1.1829	15.2383
37.mol	-2.0143	103.3463
38.mol	0.5860	0.2594
42.mol	-2.2016	159.0788
52.mol	-2.2864	193.3873
57.mol	-1.0800	12.0230
58.mol	0.7703	0.1697
62.mol	0.1038	0.7874
64.mol	-2.3705	234.7178
65.mol	-2.1758	149.9144
66.mol	-1.5239	33.4115
67.mol	0.9942	0.1013
68.mol	1.9191	0.0120
69.mol	2.1406	0.0072
70.mol	0.6351	0.2317
72.mol	1.8572	0.0139
73a.mol	1.1774	0.0665
73b.mol	2.8803	0.0013
73c.mol	1.8022	0.0158
74a.mol	1.2406	0.0575
74b.mol	-0.1164	1.3074
74c.mol	1.9006	0.0126
75.mol	-0.5985	3.9675
76a.mol	-0.3190	2.0843
76b.mol	-1.7174	52.1630
77.mol	1.4140	0.0385
78.mol	-1.5899	38.8932
79.mol	-2.0934	123.9857
80.mol	-3.6364	4329.3848
81.mol	-2.8340	682.3371
90.mol	-0.8823	7.6259
Celecoxib	0.2472	0.5660

2.4 Clustering analysis

A modified k-means clustering analysis was implemented in the study. k-means clustering is an iterative clustering procedure works as a greedy algorithm for partitioning the n samples into k clusters and predefines the number of clusters. The algorithm begins by defining centroids, which are points in the dataset that eventually appear at the center of each cluster. A very common task in data analysis is that of grouping a set of objects into subsets such that all elements within a group are more similar among them than they are to the others. Therefore, the primary objective of this study is to cluster COX-2 inhibitors using modified k-means algorithm. Hence, in this paper, we present a modification of k-means algorithm that efficiently searches data to cluster points by computing sum of squares within each cluster which makes the program to select the most promising subset of classes for clustering. From a set of 32 compounds, only the top 5 compounds are combined with 58 molecule data set to perform cluster analysis.

3. RESULTS AND DISCUSSION

The k-means clustering algorithm is popular because it can be applied to relatively large sets of data. The user specifies the number of clusters to be found. The algorithm then separates the data into spherical clusters by finding a set of cluster centers, assigning each observation to a cluster, determining new cluster centers, and repeating this process.

A modified k-means python program was written and implemented to evaluate the clusters and assessing the probability of designed compounds to appear in a particular cluster thereby confirming the similarity of such compounds to appear in a cluster of approved drugs.

The data file is formatted as a comma delimited file .csv and the file was called from python. Further clustering code was implemented to perform k-means clusters on the dataset by calculating number of clusters for given centroids. The number of centroids can be changed depending on the data being clustered. In this case, 3 centroids are considered. Now, the k-means algorithm was implemented by finding distances between objects taken for the study. Once distances between objects are calculated, the sum of squares for the two objects from the centroid is calculated. Following this, means are calculated for all objects which are nearer to each other making it to study all possible objects and the List of such objects will be created by the program to iterate the process. Centroids are assigned and the distance is calculated from each centroid to the objects and clusters are identified. To find out better clusters, a minimum distance is calculated from each object to the nearest assigned centroid. Once each object is assigned to its nearest centroid, then the distances are recalculated by adjusting the centroids in such a way that the objects should have the shortest possible distance with the centroid. Likewise, a cluster index is created and appended to the objects. Finally, the k in k-means is calculated by randomly selecting k initial centroids.

The output of the algorithm is given below.
3 clusters

Cluster-1

```
[('5_A1.mol', 1.0), ('5_A3.mol', 2.09691), ('5_A4.mol', 1.82391), ('5_A5.mol', 1.22915), ('5_B5.mol', 0.67778), ('68compd.mol', 1.919127124), ('72compd.mol', 1.857245118), ('73acompd.mol', 1.177372651), ('73bcompd.mol', 2.880278247), ('73ccompd.mol', 1.802223722)],
```

Cluster-2

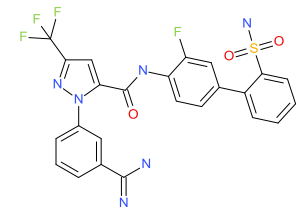
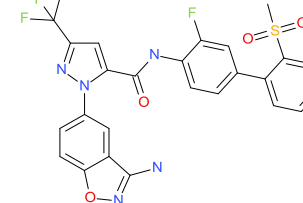
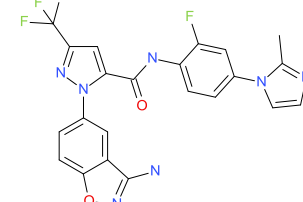
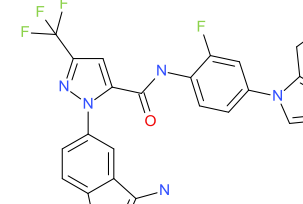
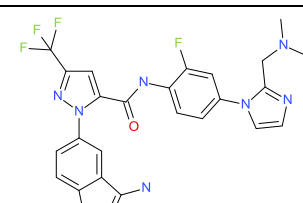
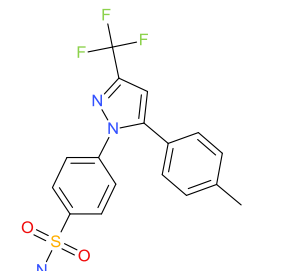
```
[('3_36.mol', -1.07518), ('3_38.mol', -1.18526), ('3_42.mol', -1.42062), ('3_45.mol', -1.50866), ('5_C1.mol', -1.82692), ('5_C2.mol', -2.23955), ('5_D2.mol', -1.97946)],
```

Cluster-3

```
[('1_8b.mol', -0.45332), ('1_8c.mol', -0.59988), ('1_8d.mol', -0.67302), ('1_8e.mol', -0.33646), ('1_8f.mol', -0.57287), ('1_8g.mol', -0.77379), ('1_8h.mol', -0.06446), ('1_8i.mol', -0.20952), ('1_8j.mol', -0.53275), ('1_8k.mol', -0.42651), ('1_8l.mol', -0.24551), ('1_8m.mol', -0.32428), ('1_8o.mol', -0.66932), ('1_8p.mol', -0.24304), ('2_10a.mol', -0.61595), ('2_10b.mol', -0.78604), ('2_10c.mol', -0.56229), ('2_10d.mol', -0.78604), ('2_10e.mol', -0.58995), ('2_10f.mol', -0.29667), ('2_10g.mol', -0.63649), ('2_10h.mol', -0.18184), ('2_10i.mol', -0.4216), ('2_10j.mol', 0.25181), ('2_10k.mol', 0.01323), ('2_10l.mol', -0.29447), ('2_10m.mol', -0.80209), ('3_30.mol', -0.92117), ('3_31.mol', -0.51587), ('3_32.mol', -0.39445), ('3_33.mol', -0.33445), ('3_34.mol', -0.41996), ('3_35.mol', -0.04922), ('3_37.mol', -0.68395), ('3_39.mol', -0.69197), ('3_41.mol', -0.54158), ('3_47.mol', 0.33724), ('3_48.mol', 0.48149), ('3_51.mol', -0.35793), ('3_52.mol', -0.91593), ('3_53.mol', -0.10037), ('3_54.mol', -0.83759), ('3_55.mol', -0.43775), ('5_C3.mol', -0.29447), ('5_C4.mol', -0.75205), ('5_D3.mol', -0.12057), ('Celecoxib', 0.24720)]
```

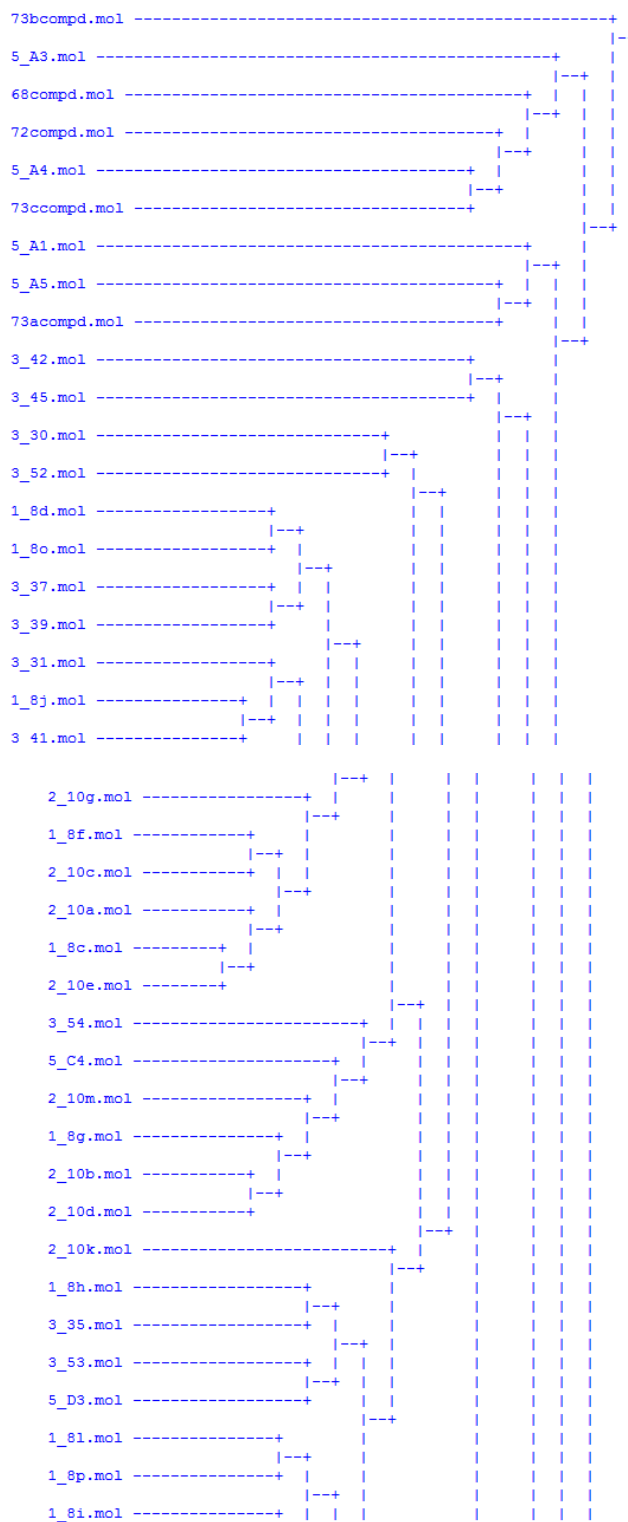
From the output, it is observed that the top 5 molecules viz., 68, 72, 73a, 73b and 73c appeared in Cluster-1, where the most active and potent compounds such as A1, A3, A4, A5 and B5 appeared. Interestingly, the approved drug Celecoxib appeared in cluster-3. From Table-2, it is evidenced that top 5 compounds surpassed the calculated IC50 value of drug Celecoxib (IC50: 0.566 μ M).

Table 2: Predicted activity and calculated IC₅₀ values of top 5 compounds from 32 novel, non-tested compound dataset from literature.

Compound No.	2D structure	MolWt	Predicted Activity (log 1/IC ₅₀)	Calculated IC ₅₀
68mol		546.54	1.919	0.012
72mol		559.53	1.857	0.013
73a		485.44	1.177	0.066
73b		514.49	1.159	0.0013
73cmol		528.52	1.802	0.016
Celecoxib		381.4	0.2472	0.565

Therefore, from the analysis it is evidenced that k-means clustering algorithm is able to group data objects of all molecules based on the 3 centroids provided. It is known that the clustering results are influenced by two factors namely, the choice of distance measure and the clustering algorithm [22]. The choice of algorithm depends on the data available and the purpose of clustering. Some algorithms result in clusters of similar size while others make clusters of

dissimilar size; some algorithms generate spherical clusters while other algorithms form elongated clusters, and few clustering algorithms are sensitive to outliers and so on. Therefore, a phylogenetic tree was processed on the dataset and presented here (Figure 1). It is observed that top 5 compounds appeared on one spade as evidenced by the k-means cluster segmentation.



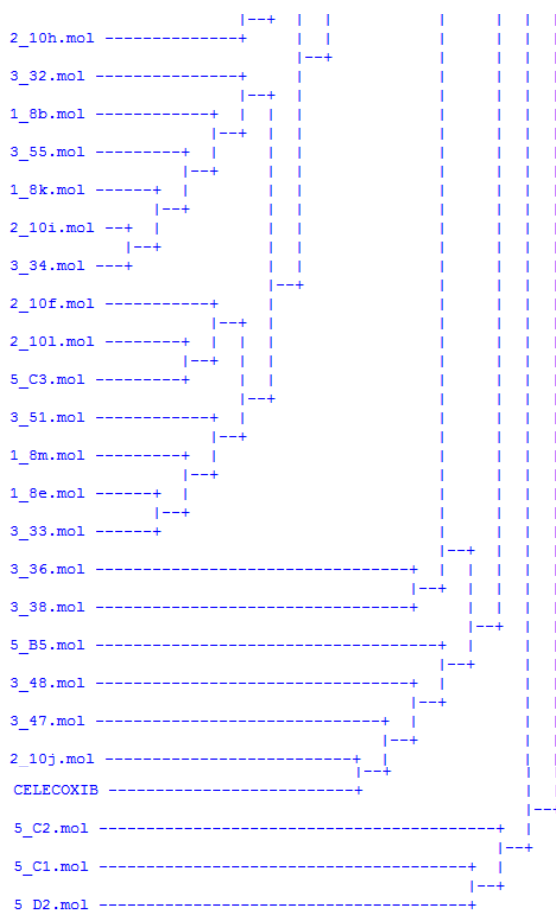


Figure 1: Phylogenetic tree representation of 58 molecule dataset along with top 5 compounds and approved drug Celecoxib

4. CONCLUSION

A modified k-means clustering algorithm was presented here to cluster groups of compounds obtained from regression analysis in combination with another set of compounds screened from literature to assess the possibility of compounds appearing in specific cluster. A 58 molecule dataset from our previous regression analysis was considered for cluster analysis which is known to possess COX-2 inhibition. The regression model obtained on this dataset was applied on a new set of 32 compounds which were not tested as COX-2 inhibitors in order to predict biological activities. A known approved drug Celecoxib was found to be less active than top 5 novel compounds screened from literature. Finally, a phylogenetic tree was presented to know the branches on which these novel compounds appear and hence can be used to further extend the study towards identifying more statistical parameters that would substantiate the identification of entirely novel, potent COX-2 inhibitors.

5. REFERENCES

- [1] Jouzeau JY, Terlain B, Abid A, Nedelec E, Netter P. Cyclo-oxygenase isoenzymes. How recent findings affect thinking about nonsteroidal anti-inflammatory drugs. *Drugs*. 1997;53(4):563-82
- [2] K. T. Moreland, J. D. Procknow, R. S. Sprague, J. L. Iverson, A. J. Lonigro and A. H. Stephenson. 2007. Cyclooxygenase (COX)-1 and COX-2 Participate in 5, 6-Epoxyeicosatrienoic Acid-Induced Contraction of Rabbit Intralobar Pulmonary Arteries *J. Pharmacol. Exp. Ther.*, 321(2): 446 – 454.
- [3] S. S. Reuben and E. F. Ekman. 2005. The Effect of Cyclooxygenase-2 Inhibition on Analgesia and Spinal Fusion *J. Bone Joint Surg. Am.*; 87(3): 536 – 542.
- [4] H Sano, T Hla, J A Maier, L J Crofford, J P Case, T Maciag, and R L Wilder. 1992. In vivo cyclooxygenase expression in synovial tissues of patients with rheumatoid arthritis and osteoarthritis and rats with adjuvant and streptococcal cell wall arthritis. *J Clin Invest.*; 89(1): 97–108
- [5] Reitz, D. B. 1995. Isakson, P. C. Cyclooxygenase-2 Inhibitors. *Curr. Pharm. Des*, 1, 211-220.
- [6] Subbaramaiah, K.; Zakim, D.; Weksler, B. B.; Dannenberg A. J. 1997. Inhibition of Cyclooxygenase: A Novel Approach to Cancer Prevention. *Proc. Soc. Exp. Biol. Med*, 216, 201-210.
- [7] Pasinetti, G. M. Cyclooxygenase and Inflammation in Alzheimer’s Disease. 1998. *Experimental Approaches and Clinical Intervention. J. Neurosci. Res*, 54, 1-6
- [8] Futaki, N.; Yoshikawa, K.; Hamasaka, Y.; Arai, I.; Higuchi, S.; Iizuka, H.; Otomo, S. NS398. 1993. a Novel Nonsteroidal Anti inflammatory Drug with Potent Analgesic and Antipyretic Effects, which Causes Minimal Stomach Lesions. *Gen. Pharmacol*, 24, 105-110.
- [9] Gans, K.; Galbraith, W.; Roman, R.; Haber, S.; Kerr, J.; Schmidt, W.; Smith, C.; Hewes, W.; Ackerman, N. 1990. Antiinflammatory and Safety Profile of DuP697, a Novel

- Orally Prostaglandin Synthesis Inhibitor. *J. Pharm. Exp. Ther.* 254, 180-187.
- [10] Kalgutkar, A. S.; Marnett, A. B.; Crews, B. C.; Rimmel, R. P.; Marnett, L. J. Ester and Amide Derivatives of the Nonsteroidal Anti inflammatory Drug, Indomethacin, as Selective Cyclooxygenase-2 Inhibitors. *J. Med. Chem.* 2000, 43, 2860-2870.
- [11] Khanna, I. K.; Weier, R. M.; Yu, Y.; Collins, P. W.; Miyashiro, J. M.; Koboldt, C. M.; Veenhuizen, A. W.; Currie, J. L.; Seibert, K.; Isakson, P. C. 1,2-Diarylpyrroles as Potent and Selective Inhibitors of Cyclooxygenase-2. *J. Med. Chem.* 1997, 40, 1619-1633.
- [12] Khanna, I. K.; Weier, R. M.; Yu, Y.; Xu, X. D.; Koszyk, F. J.; Collins, P. W.; Koboldt, C. M.; Veenhuizen, A. W.; Perkins, W.E.; Casler, J. J.; Masferrer, J. L.; Zhang, Y. Y.; Gregory, S. A.; Seibert, K.; Isakson, P. C. 1,2-Diarylimidazoles as Potent, Cyclooxygenase-2 Selective, and Orally Active Antiinflammatory Agents. *J. Med. Chem.* 1997, 40, 1634-1647.
- [13] Reitz, D. B.; Li, J. J.; Norton, M. B.; Reinhart, E. J.; Collins, J.T.; Anderson, G. D.; Gregory, S. A.; Koboldt, C. M.; Perkins, W.E.; Seibert, K.; Isakson, P. C. Selective Cyclooxygenase Inhibitors: Novel 1,2-Diarylcyclopentenes are Potent and Orally Active COX-2 Inhibitors. *J. Med. Chem.* 1994, 38, 3878-3881.
- [14] Enning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; et al. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1Hpyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* 1997, 40, 1347-1365.
- [15] Huang, H. C.; Li, J. J.; Garland, D. J.; Chamberlain, T. S.; Reinhart, E. J.; Manning, R. E.; Seibert, K.; Koboldt, C. M.; et al. Diarylspiro[2.4] heptenes as Orally Active, Highly Selective Cyclooxygenase-2 inhibitors: Synthesis and Structure-Activity Relationships. *J. Med. Chem.* 1996, 39, 253-266.
- [16] Talley J. J.; Brown, D. L.; Carter, J. S.; Graneto, M. J.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Rogers, R. S.; Shaffer, A. F.; Zhang, Y. Y.; Zweifel, B. S.; Seibert, K. 4-[5-Methyl-3- phenylisoxazol-4-yl]-benzenesulfonamide, Valdecoxib: a Potent and Selective Inhibitor of COX-2. *J. Med. Chem.* 2000, 43, 775-777.
- [17] Zhong Chen et al. Design, synthesis, biological evaluation and molecular modeling of dihydropyrazolesulfonamide derivatives as potential COX-1/COX-2 inhibitors. *Bioorganic & Medicinal Chemistry Letters* 25 (2015) 1947–1951.
- [18] Abdellatif, K.R.A. et al. Synthesis, cyclooxygenase inhibition, anti-inflammatory evaluation and ulcerogenic liability of novel triarylpyrazoline derivatives as selective COX-2 inhibitors, *Bioorganic & Medicinal Chemistry Letters* 2015; 25(24):5787-91.
- [19] Abdellatif, K.R.A. et al. Synthesis, cyclooxygenase inhibition, anti-inflammatory evaluation and ulcerogenic liability of new 1,3,5-triarylpyrazoline and 1,5-diarylpyrazole derivatives as selective COX-2 inhibitors. *Bioorganic & Medicinal Chemistry Letters* 2016; 26(2):406-12.
- [20] Yu, M., Yang, H., Wu, K., Ji, Y., Ju, L. and Lu, X., 2014. Novel pyrazoline derivatives as bi-inhibitor of COX-2 and B-Raf in treating cervical carcinoma. *Bioorganic & medicinal chemistry*, 22(15), pp.4109-4118.
- [21] Ruth Pérez-Fernández, Pilar Goya, and José Elguero. A review of recent progress (2002-2012) on the biological activities of pyrazoles. *ARKIVOC* 2014 (ii) 233-293.
- [22] Pablo A Jaskowiak et al. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics* 2014, 15(Suppl 2):S2 doi:10.1186/1471-2105-15-S2-S2.