PHYSICOCHEMICAL PROPERTIES OF MEDICINAL COMPOUNDS IN THE WORLD HEALTH ORGANISATION'S ANATOMICAL THERAPEUTIC CHEMICAL CLASSIFICATION

Francesca Pace, Frederick Lia and Claude Farrugia

Department of Chemistry, University of Malta, Msida MSD 2080, Malta

Introduction

The WHO's Anatomical Therapeutic Chemical (ATC) classification system is the most widely used drug classification system. Other classifications exist, based on different properties of the drug molecules, and with each system having its own advantages and limitations. The objective of this study was to propose a classification of the drugs in the ATC classification system, based on their physicochemical properties.

Methodology

A total of 2530 medicinal compounds was identified and six physicochemical parameters - molecular density, total surface area (TSA), polar surface area (PSA), Log P, parachor and molecular weight (MW) - generated using *ACD/ChemSketch*, *Calculator Plugins* and *chemicalize.org*. The data was statistically evaluated with *JMP software* using Multivariate Platform and Principal Component Analysis. K-Means Clustering was performed to propose a new PC classification system based on the physicochemical properties of the compounds. The PC and ATC classifications were compared using Supervised Linear – Canonical Discriminant Analysis, Fit Y by X and Artificial Neural Network Analysis.

Results and Discussion

Multivariate Platform

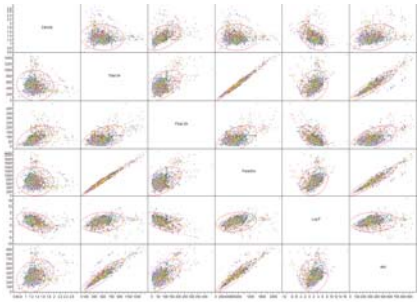


Figure 1: A scatterplot matrix explaining visually the correlation between the physicochemical properties of the medicinal substances. PSA has a strong positive correlation with both TSA and molecular density, while Log P has a strong inverse correlation with both PSA and molecular density.

Principal Component Analysis

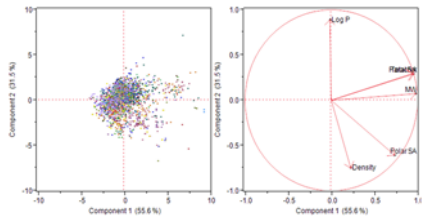
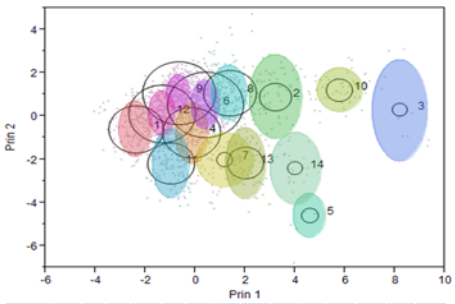


Figure 2: Score Plot and Factor Plot for all of the medicinal compounds classified according to their physicochemical properties. The first two principal components accounted for 87% of the total variance. The first component was characterised by major positive levels of MW, parachor, PSA and TSA, while the second component was characterised by major positive levels of Log P and major negative levels of molecular density and PSA.

K-Means Clustering



Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Count	248	85	16	272	23	469	22	227	440	54	181	365	106	18

Figure 3: K-Means Clustering for all the medicinal compounds based on their physicochemical properties, and the number of compounds in each cluster. The size of the circles is proportional to the count inside the cluster, and the shaded area is the 90% density contour around the cluster centre.

Classification Groups	Density	TSA	PSA	Parachor	Log P	MW
1	1.214	219.2	47.9	316.0	0.367	146.0
2	1.309	873.2	133.1	1241.4	3.598	589.3
3	1.267	1399.5	266.2	2019.3	1.552	922.9
4	1.454	434.5	94.9	640.4	0.960	330.3
5	1.677	832.0	341.0	1171.1	-8.154	616.3
6	1.273	555.2	71.9	805.1	2.858	381.2
7	2.461	396.2	79.0	698.7	3.501	586.9
8	1.238	683.8	80.7	977.0	3.884	454.0
9	1.138	471.9	34.4	677.8	3.620	298.7
10	1.295	1178.1	190.5	1637.9	3.529	777.4
11	1.755	294.6	104.6	424.3	-0.491	249.5
12	1.227	349.5	53.4	508.9	2.219	237.4
13	1.628	608.7	187.9	889.2	-1.979	473.5
14	2.018	733.1	214.0	1105.4	-0.387	722.8

Table 1: The means of the new PC classification groups, scaled from red at the smallest value, to yellow at the 50th percentile to green at the maximum value, showing the unique combination of physicochemical properties for each cluster of medicinal compounds.

Discriminant Analysis

Classification Type and Canonical Variable	Eigenvalue	Cumulative Percent of Variance (%)	Canonical Correlation
PC – Canonical 1	22.681	84.933	0.979
PC – Canonical 2	2.440	94.072	0.842
ATC – Canonical 1	0.313	52.632	0.488
ATC - Canonical 2	0.174	81.886	0.385

Table 2: Canonical Details for the PC and ATC classification expressed in terms of the first two canonical variables. The physicochemical parameters of the medicinal compounds are strongly correlated with the groups formed in the PC classification, but not with the groups in the ATC classification. The PC classification accounts for a greater percentage of the variance than the ATC classification.

Neural Network Analysis

Validation Method	Training Generalised R^2		Validation Generalised R^2		Percentage Misclassification	
	PC	ATC	PC	ATC	PC	ATC
Holdback	0.9997	0.4805	0.9987	0.3854	0.6%	69.9%
K-fold	0.9988	0.4457	0.9988	0.5172	1.8%	69.9%
Excluded Rows	0.9998	0.3337	0.9995	0.1753	0.2%	73.3%

Table 3: Training and validation of the Artificial Neural Network Analysis for the PC and ATC classifications. The PC classification exhibited a high degree of fit, with low misclassification rates, whereas the ATC classification fit the model poorly, with a low degree of fit and high misclassification rates.

Fit Y by X

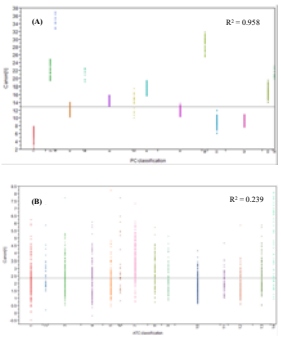


Figure 4: A One-way biplot illustrating the variation of the first canonical score with the groups found in the (A) PC and (B) ATC classification. The data in each group of the PC classification are closely packed implying similar physicochemical properties, and the R^2 value is close to 1, implying a good fit for the model. This is not the case for the ATC classification.

Conclusion

The three statistical analysis concluded that the ATC classification did not fit within the model analysed in this study. Therefore, the two classifications were statistically different from each other. The new PC classification system may be useful as a tool in assessing lead compounds; molecules falling within a physicochemical cluster of known drugs in the PC classification would probably present a lower risk of failure in the process of drug development.

References

- ACD/ChemSketch (Freeware) 2012, version 14.01, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2013
- Chemicalize.org by ChemAxon (http://www.chemicalize.org/)
- Calculator Plugins, Marvin 15.2.2.0, 2015, ChemAxon (http://chemaxon.com)
- JMP® 10.0.0, 2012, SAS Institute Inc., Cary, NC, 1989-2007