CHEMOMETRICS AND MULTIBLOCK METHODS FOR QUANTITATIVE STRUCTURE-ACTIVITY STUDIES OF ARTEMISININ ANALOGUES AND POLYCHLORINATED DIPHENYLETHERS

ROSMAHAIDA JAMALUDIN

UNIVERSITI TEKNOLOGI MALAYSIA

CHEMOMETRICS AND MULTIBLOCK METHODS FOR QUANTITATIVE STRUCTURE-ACTIVITY STUDIES OF ARTEMISININ ANALOGUES AND POLYCHLORINATED DIPHENYLETHERS

ROSMAHAIDA JAMALUDIN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Chemistry)

Faculty of Science
Universiti Teknologi Malaysia

SEPTEMBER 2015

I declare that this thesis entitled "*Chemometrics And Multiblock Methods For Quantitative Structure-Activity Studies of Artemisinin Analogues and Polychlorinated Diphenylethers*" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : Rosmahaida Jamaludin

Date : 29 September 2015

To Allah (SWT) and my beloved family

ACKNOWLEDGEMENT

The very first words that came through my mind was Syukur Alhamdulillah. It is of great honor for me to take this opportunity to thank many people whose contributions, helps and encouragement are so valuable throughout the journey.

First and foremost, I wish to extend my greatest appreciation to my supervisor, Professor Dr Mohamed Noor Hasan and my co-supervisor Dr Mohd Zuli Jaafar who have been instrumental in providing me beneficial guidance, valuable advices, kind supervision, patience and confidence on me throughout the course of the study. Sincere gratitude also goes to Dr Barry Lavine from Oklahoma State University, USA and Dr Neni Frimayanti from University Malaya for their professional assistance as well as interest towards the research. All the knowledge transfer and kindness will always be cherished in my heart.

Special thanks are due to all my colleagues, Alvin, Zalikha, Fatin and Bishir for their kind co-operation, undivided assistance as well as willingness to share knowledge and experiences. Wishing you good luck in your future.

Last but not least, I wish to forward my deepest gratitude to my husband, Hairil Anuar, my late father, Jamaludin Jaafar, my mother, Rosnah Juleh and my children, Amirul Hafidz, Akmal Darwisy, Alia Nurmaisarah, Adam Nurluqman and Ayman Nurqayyum for their understanding and strong encouragement. I love all of you. May Allah bless us.

ABSTRACT

Three major aspects of chemometrics have been investigated in this study namely Quantitative Structure-Activity Relationship (QSAR) and database mining, classification and multiblock methods. In the first analysis, 197 artemisinin compounds were divided into training set and test set together with structural descriptors generated by DRAGON 6.0 software had been used to develop three QSAR models. Statistics of the models were (r^2/ r_{test}^2) 0.790/0.853 for Forward Stepwise-Multiple Linear Regression (MLR), 0.807/0.789 for Genetic Algorithm (GA)-MLR and 0.795/0.811 for GA-Partial Least Square (PLS). The rigorously validated QSAR models were then applied to mine a chemical database which resulted in four potential new anti-malarial agents. The same artemisinin data set was then classified into active and less active compounds to develop reliable predictive classification models and to investigate the consequences of using various data splitting and data pre-processing methods on classification. Principal Component Analysis (PCA) and boundary plot had been utilized to visualize the four classifiers namely Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Linear Vector Quantization (LVQ) and Quadratic Discriminant Analysis (QDA). Kennard-Stone data splitting and standardization had produced better results in terms of percent correctly classified (% CC) compared to Duplex data-splitting and mean-centering. Moreover, LDA was found to be superior as compared to the other three classifiers with lower risk of over-fitting. Lastly, multiblock analysis methods such as Multiblock PLS and Consensus PCA have been implemented on polychlorinated diphenyl ethers (PCDEs) data set together with their respective descriptors blocked into three groups labelled as X_{1D}, X_{2D}, X_{3D} and a property block, Y which consists of log P_L (Pa, 25°C), log K_{OW} (25°C) and log S_{WL} (mol/L, 25°C). Their performance were then compared to single block methods that is PLS and PCA. The PLS models of each descriptor block with respect to each property were statistically best-fitted and well predicted with r_{train}^2 values greater than 0.96 while the r_{test}^2 values range from 0.86 to 0.98. It is interesting to note that the combination of the three descriptor blocks into a single block to produce Multiblock PLS super-scores (MBSS) model which was superior than Multiblock PLS block-scores (MBBS) yielded slightly better r_{train}^2 value and significantly better prediction with higher r_{test}^2 as compared to PLS model of individual descriptor block. In addition, three measures of block similarity such as Mantel Test, R_v coefficient and Procrustes analysis were used to investigate similarity and correlation between the blocks along with Monte Carlo simulations to determine their significance. Based on the similarity index between two blocks, X_{1D} descriptors resembled Y block better while X_{2D} was more correlated to X_{1D} block. In short, the chemometric methods had been applied successfully on both data sets using various descriptors generated by DRAGON software and yielded promising results beneficial not only in chemometrics area but also in drug design.

ABSTRAK

Tiga aspek utama bidang kimometrik telah disiasat dalam kajian ini iaitu kaedah Hubungan Kuantitatif Struktur-Aktiviti (QSAR) dan pangkalan data, klasifikasi dan multiblok. Dalam analisis yang pertama, 197 sebatian artemisinin telah dibahagikan kepada set latihan dan set ujian beserta deskriptor struktur yang dijanakan oleh perisian DRAGON 6.0 telah diguna untuk menghasilkan tiga model QSAR. Statistik model ialah (r^2/ r_{test}^2) 0.790/0.853 bagi kaedah Langkah Maju-Regresi Linear Berganda (MLR), 0.807/0.789 bagi Algoritma Genetik (GA)-MLR dan 0.795/0.811 bagi GA-Regresi Linear Separa (PLS). Model QSAR yang sah digunakan untuk mencari dalam pangkalan data kimia lalu menghasilkan empat bahan kimia baharu yang berpotensi sebagai agen anti malaria. Set data artemisinin yang sama kemudian dikelaskan kepada aktif dan kurang aktif untuk membina model klasifikasi, di samping menyiasat kesan penggunaan pelbagai teknik pemisahan dan pra-prosesan data terhadap klasifikasi. Analisis Komponen Prinsipal (PCA) dan plot sempadan telah digunakan untuk menggambarkan empat jenis model klasifikasi iaitu Mesin Vektor Sokongan (SVM), Analisis Pembezalayan Linear (LDA), Pengkuantuman Vektor Linear (LVQ) dan Analisis Pembezalayan Kuadratik (QDA). Kaedah Kennard-Stone dan pra-prosesan piawai telah menghasilkan keputusan yang lebih baik dari segi peratus pengkelasan yang betul (% CC) berbanding Duplex dan pra-prosesan purata-tengah. Di samping itu, LDA didapati lebih baik dengan risiko suaian lampau yang lebih rendah. Akhir sekali, analisis multiblok seperti Multiblok PLS dan konsensus PCA telah dijalankan ke atas set data poliklorin difenil eter (PCDEs) beserta dengan tiga kumpulan blok deskriptor masing-masing iaitu X_{1D}, X_{2D}, X_{3D} dan blok sifat, Y yang terdiri daripada log P_L (*Pa*, 25°C), log K_{OW} (25°C) and log S_{WL} (*mol/L*, 25°C). Prestasi kaedah ini seterusnya dibandingkan dengan kaedah blok tunggal iaitu PLS dan PCA. Model PLS setiap blok deskriptor terhadap setiap sifat secara statistiknya best-fitted dan ramalan baik dengan nilai r_{train}^2 lebih besar daripada 0.96 manakala nilai r_{test}^2 adalah dalam julat 0.86 hingga 0.98. Sesuatu yang menarik untuk diperhatikan bahawa gabungan tiga blok deskriptor ke dalam blok tunggal menghasilkan model Multiblok PLS Super-Skor (MBSS) yang lebih baik daripada Multiblok PLS Blok-Skor (MBBS) menghasilkan nilai r_{train}^2 dan r_{test}^2 yang lebih tinggi berbanding model PLS blok deskriptor individu. Sebagai tambahan, tiga pengukuran keserupaan blok seperti ujian *Mantel*, pekali R_v dan analisis *Procrustes* telah digunakan untuk menyiasat keserupaan dan korelasi antara blok diikuti simulasi *Monte Carlo* untuk menentukan kepentingannya. Berdasarkan indeks keserupaan antara dua blok, deskriptor X_{1D} lebih menyerupai blok Y manakala deskriptor X_{2D} mempunyai korelasi lebih kepada blok X_{1D}. Ringkasnya, kaedah kimometrik telah berjaya digunakan ke atas kedua-dua set data menggunakan pelbagai deskriptor yang dijanakan oleh perisian DRAGON dan menghasilkan keputusan bermanfaat bukan sahaja dalam bidang kimometrik tetapi juga bidang rekabentuk ubatan.