ชื่อเรื่องวิทยานิพนธ์	การวิเคราะห์โครงสร้างยา โดยวิธีการถดถอยกำลังสองน้อยที่สุดบางส่วน
ผู้เขียน	นางสาวปราณี คำแก้ว
ปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
คณะกรรมการที่ปรึกษาวิทยานิพนธ์	อ.ดร.ภัทรินี ไตรสถิตย์ ประธานกรรมการ ผศ.ดร.จีรยุทธ ไชยจารุวณิช กรรมการ ผศ.ดร.ปิยรัตน์ นิมมานพิภักดิ์ กรรมการ

บทคัดย่อ

งานวิจัยครั้งนี้ เป็นการวิเคราะห์โครงสร้างยา โดยวิธีการถดถอยกำลังสองน้อยที่สุดบางส่วน (PLS1 และ PLS2) เพื่อหาความสัมพันธ์ระหว่างคุณสมบัติต่างๆ ของโครงสร้างโมเลกุลของสารกับการออกฤทธิ์ทางยา โดยทำการศึกษาจากกรณีตัวอย่างของโรคภูมิคุ้มกันบกพร่อง หรือโรคเอดส์ จากฐานข้อมูล 3 ชุด ได้แก่ 1) โครงสร้างจากฐานข้อมูลธนาคารข้อมูลโปรตีน ซึ่งค่าการออกฤทธิ์ทางยาที่ใช้คือ ค่าคงที่ของสารยับยั้ง (pK_i) 2) โครงสร้างจากฐานข้อมูลสถาบันมะเร็งแห่งชาติ ซึ่งค่าการออกฤทธิ์ทางยาที่ใช้คือ ความเข้มข้นร้อยละ 50 ที่มีผลกระทบเทียบกับกลุ่มควบคุม (pEC_{50}) โดยใช้ฐานข้อมูลทั้งสองนี้เป็นชุดข้อมูลสร้างตัวแบบ (Training set) และ 3) โครงสร้างสารออกฤทธิ์ของสมุนไพรไทยจากฐานข้อมูลทางเคมี โดยใช้เป็นชุดข้อมูลทดสอบ (Test set) เสมือนเป็นข้อมูลที่ไม่เคยศึกษามาก่อน และต้องการพยากรณ์ค่าการออกฤทธิ์ทางยา

ผลการศึกษาพบว่า เมื่อทำการวิเคราะห์โดยวิธี PLS1 พบว่า 1) จากฐานข้อมูลธนาคารข้อมูลโปรตีน ตัวแบบที่ได้จากกลุ่มข้อมูลให้ค่าสัมประสิทธิ์การตัดสินใจพหุคูณที่มีการปรับค่า (R_a^2) มีค่าน้อย เช่น กลุ่มที่มีจำนวนตัวอย่าง 38 ตัวอย่าง แบ่งเป็น Training set 25 ตัวอย่าง และ Test set 13 ตัวอย่าง โดยที่มีตัวแปรอิสระ 192 ตัวแปร ตัวแปรตามคือ pK_i ได้ค่า R_a^2 เท่ากับ 0.2461 2) จากฐานข้อมูลสถาบันมะเร็งแห่งชาติ มีข้อมูลเพียงบางกลุ่มที่ตัวแบบที่สร้างขึ้น มีค่า R_a^2 มาก เช่น กลุ่มที่มีจำนวนตัวอย่าง 16 ตัวอย่าง แบ่งเป็นชุดข้อมูล Training set 11 ตัวอย่าง และ Test set 5 ตัวอย่าง โดยที่มีตัวแปรอิสระ 131 ตัวแปร ตัวแปรตามคือ pEC_{50} ได้ค่า R_a^2 เท่ากับ 0.9653 และ 3) จากชุด

ข้อมูลโครงสร้างสารออกฤทธิ์ของสมุนไพรไทยจากฐานข้อมูลทางเคมี เมื่อทำการพยากรณ์หาค่า pK_i พบว่าสารประกอบบางตัวมีค่าใกล้เคียงกับค่า pK_i ในทำนองเดียวกัน เมื่อทำการพยากรณ์หาค่า pEC_{50} พบว่าสารประกอบบางตัวมีค่าใกล้เคียงกับค่า pEC_{50}

เมื่อทำการวิเคราะห์โดยวิธี PLS2 พบว่า 1) จากฐานข้อมูลธนาคารข้อมูลโปรตีน ตัวแบบที่ได้จากกลุ่มข้อมูลให้ค่า R_a^2 มีค่าน้อย เช่น มีจำนวนตัวอย่าง 22 ตัวอย่าง แบ่งเป็น Training set 15 ตัวอย่าง และ Test set 7 ตัวอย่าง โดยที่มีตัวแปรอิสระ 191 ตัวแปร ตัวแปรตามคือ pK_i และ logP ได้ค่า R_a^2 เท่ากับ 0.1085 และ 0.6304 ตามลำดับ 2) จากฐานข้อมูลสถาบันมะเร็งแห่งชาติ มีข้อมูลเพียงบางกลุ่มที่ตัวแบบที่สร้างขึ้นให้ค่า R_a^2 มีค่าปานกลาง เช่น มีจำนวนตัวอย่าง 12 ตัวอย่าง แบ่งเป็นชุดข้อมูล Training set 8 ตัวอย่าง และ Test set 4 ตัวอย่าง โดยที่มีตัวแปรอิสระ 131 ตัวแปร ตัวแปรตามคือ pEC_{50} และ logP มีค่า R_a^2 เท่ากับ 0.4360 และ 0.6781 ตามลำดับ

จากผลการวิเคราะห์ทั้งสองวิธีข้างต้น (PLS1 และ PLS2) เมื่อตัวแบบที่สร้างขึ้นมีค่า R_a^2 มาก จะพบว่าสอดคล้องกับการที่พบว่าโครงสร้างเคมีมีลักษณะที่คล้ายคลึงกันมากเช่นกัน กล่าวได้ว่าตัวแบบจากกลุ่มข้อมูลชุดดังกล่าวนั้นสามารถนำไปเป็นข้อมูลที่ใช้ในการอ้างอิงต่อไปได้ ในทำนองเดียวกัน ถ้าตัวแบบที่ได้จากกลุ่มข้อมูลมีค่า R_a^2 น้อย ก็พบว่าสอดคล้องกับการพบโครงสร้างเคมีภายในกลุ่มที่มีลักษณะที่คล้ายคลึงกันน้อยเช่นกัน ถึงแม้ว่าการศึกษาครั้งนี้มีข้อจำกัดหลายประการ อย่างไรก็ตามพบว่า PLS1 และ PLS2 เป็นเครื่องมือที่เป็นประโยชน์ในการพยากรณ์ค่าการออกฤทธิ์ทางยา และสามารถนำไปประยุกต์ในการพยากรณ์คุณสมบัติทางเคมีของโครงสร้างที่คล้ายคลึงกัน

Thesis Title Analysis of Drug Structure by Partial Least Squares Regression Methods

Author Miss Pranee Kamkaew

Degree Master of Science (Applied Statistics)

Thesis Advisory Committee
Dr.Patrinee Traisathit — Chairperson
Asst.Prof.Dr.Jeerayut Chaijaruwanich — Member
Asst.Prof.Dr.Piyarat Nimmanpipug — Member

ABSTRACT

This research project is an analysis of drug structure by Partial Least Squares Regression Methods (PLS1 and PLS2). The objective is to find correlation between each structural descriptor and the activity of the selective HIV-1 inhibitors based on data from the following 3 sources.

1). Protein Data Bank (PDB) where inhibition constant (pK_i) was used as the activity.

2). National Cancer Institute (NCI) where the effect concentration 50% (pEC_{50}) was used as the activity to construct a training set.

3). Chemical Database where structures of effective substance contained in Thai herbs were used as test set under assumptions that they have never been previously studied and the activity has to be predicted.

Results from PLS1 analysis are as followed:

1). Model from The Protein Data Bank's database gave a low adjusted coefficient of multiple determinations (R_a^2). For example, a group containing 38 samples (25 samples for training set and 13 samples for test set) which has 192 independent variables and dependent variables was pK_i gave a $R_a^2 = 0.2461$.

2). Only a few groups of the model from The National Cancer Institute's database have a high R_a^2 such as a group that contained 16 samples (11 samples for training set and 5 samples for test set) which has 131 independent variables and dependent variables was pEC_{50} gave a R_a^2 = 0.9653

3). Prediction of pK_i value of Thai herbs' effective substance structure yielded number closed to that of pK_i on some component. The same thing happened when we tried to predict pEC_{50}.

PLS2 analysis yielded the following results:

1). Model from The Protein Data Bank's database gave a low R_a^2 such as 0.1085 and 0.6304 from a 22-sample-group (15 samples for training set and 7 samples for test set) which consists of 191 independent variables and pK_i and logP as dependent variables, respectively.

2). The National Cancer Institute's database showed only a few groups in which model gave a moderate R_a^2. For example, a 12-sample-group (8 training set samples and 4 test set samples) with 131 independent variables and pEC_{50} and logP as dependent variables gave R_a^2 values of 0.4360 and 0.6781, respectively.

Based on the results from both analytical techniques (PLS1 and PLS2), we can conclude that the chemical structures are very similar when model has a high R_a^2 value. Conversely, model from a low R_a^2 data set has less similarity. The model from this set of data can be used as a reference in the future. Despite of many limitations involved in this study, the PLS1 and PLS2 are very useful in the prediction of activity and can also be adapted to be used in the prediction of Chemical attribute in a similar structure.