A Comparison of Plant and Human Metabolites Using QSAR and Artificial Neural Networks

Michael Fox

Supervisors: Dr Irilenia Nobeli and Dr Adrian Shepherd

MRes Bioinformatics with Systems Biology

School of Crystallography
Birkbeck College
University of London
Malet Street
London WC1E 7HX

Abstract

Three groups of metabolites have been assembled. They comprise human metabolites taken from both the Human Metabolome Database (HMDB) and the KEGG database, and plant metabolites taken from the model plant organism *Arabidopsis thaliana*. A number of computational techniques borrowed from the field of chemoinformatics were used to identify the commonalities and differences between the plant and human metabolites. Most surprisingly, the study has revealed a large discrepancy between the human metabolites compiled by the HMDB and the human metabolites catalogued in KEGG. A subsequent QSAR analysis found that binary artificial neural networks trained to differentiate between plant and human metabolites performed more successfully on the data set consisting of Arabidopsis plant metabolites and KEGG human metabolites. Two different sets of molecular descriptors were used as the parameters for the artificial neural networks. Test results showed that the set of 32 2D 'VSA' descriptors as described in a paper by Labute (2000) out-performed the combined set of 26 low-correlated traditional 2D and 3D 'inductive' descriptors. The best performing neural networks were able to classify plant and human metabolites with a reasonable 85% accuracy and a Matthews coefficient of 0.7. Contrary to two studies by Karakoc involving human metabolites (2006 and 2007), the human metabolites used in this study were not shown to occupy an independent cluster in the QSAR descriptor space compared with the other classes of metabolites. Instead a 3D graph, which plotted the three principal components of the KEGG human and Arabidopsis data set, revealed a dense cluster of human and plant metabolites as well as a number of interesting outliers.

Contents

Acknowledgements

I would like to thank my supervisors Dr Irilenia Nobeli and Dr Adrian Shepherd for their help and support during the course of my project.

1 Introduction

The metabolism can be defined as the sum total of all the biochemical processes that take place within living cells. These processes interact to form molecular pathways. The small molecules that exist along these pathways are known as metabolites. Just as 'genome' refers to the complete set of genes in an organism, the word 'metabolome' can refer to the full complement of all small molecule metabolites found in a specific cell, organ or organism (Nobeli and Thornton, 2006). There are two distinct categories of approach towards the study of the metabolome as a whole. These approaches differ in terms of the number of compounds they analyse, their sensitivity, and the level of structural information they obtain.

The most common approach is known as *metabolite profiling* and seeks to analyse small numbers of known metabolites which exist along single pathways or within specific compound classes (for example, lipid, amino acid or sterol) and is concerned with the separation and identification of individual metabolites. The alternative to this is the *fingerprint* approach. First described by Oliver *et al*. (1998), it involves measuring the cellular concentrations of many metabolites at once to produce a 'metabolic snapshot'. Such snapshots can then be used to diagnose diseases (Brindle *et al*., 2002) or to understand the effects of environmental or genetic changes on the

organism. Both approaches most commonly use the techniques of mass spectrometry and nuclear magnetic resonance in their analysis.

The study of individual metabolites comes with its own set of problems and differs from the challenges present in the study of proteins. This is due to the fundamental differences which exist between these two types of organic compound. Firstly, whereas proteins can be linked to the genotype, there is no direct evolutionary relationship among metabolites. Secondly, while the primary sequence of a protein is restricted to an alphabet of 20 amino acids, a metabolite's constituents are only limited to the number of chemical scaffolds that are available to the organism, i.e. those scaffolds that can be constructed from the finite number of reactions discovered by evolution.

For bacteria and other simple organisms, the metabolome is considered to be much smaller than the genome or proteome. The metabolome of yeast, for example, is currently believed to involve some 1,200 enzymatic reactions and 650 metabolites. Similar numbers also exist for bacteria such as *Escherichia coli* and *Streptomyces coelicolor* (Kell, 2006). In plant organisms, the opposite is true. Plants produce an estimated 10^5 distinct small molecules (Fiehn, 2001).

Despite a large number of commercially available metabolomics databases, the examples in the public domain are currently not as comprehensive or indeed as interoperable as their equivalents for the genome and transcriptome. Mendes identifies several types of useful metabolomics databases (Mendes, 2002). Databases originally built for a different purpose but which catalogue the metabolites in a variety of

species already exist such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000) and PubChem (Wheeler *et al*., 2006). More recently, Chemical Entities of Biological Interest (ChEBI) is a development at the European Bioinformatics Institute which is building an ontological classification system specifically for small chemical compounds (Brooksbank *et al*., 2005). Another promising effort is the freely available Human Metabolome Database (HMDB) which stores small molecule metabolites found in the human body (Wishart *et al*., 2007).

Established chemoinformatics techniques originally developed for use on large scale chemical databases and for virtual screening libraries have found further application in numerous metabolite and other small-molecule studies. In particular, searching for structural and functional similarities amongst groups of metabolites is a data mining problem which is covered extensively in chemoinformatics research.

The classification of metabolites in this way has potentially many useful applications. One such use concerns the correlation between the similarity amongst metabolites and the protein domains they are known to bind to within metabolic pathways. Identifying such correlations can help to determine the function of proteins with known structure. It can also assist in designing ligands for known drug targets. A study by Nobeli *et al*. (2003) performed a cluster analysis of the set of 745 metabolites that were known to comprise the E. coli metabolome. These metabolites were then compared with a data set of ligands that were obtained from the crystal structures of proteins. The comparison showed that the set of ligands and the set of metabolites displayed remarkably similar physicochemical properties.

In a number of studies, artificial neural networks (ANNs) have been successfully used to generate non-linear QSAR models that can distinguish between different chemical compound groups. A paper by Cherkasov (2005a) reports how ANNs were trained to identify several categories of chemicals from a molecular data set comprised of 525 antimicrobial compounds, 959 general drugs and 1,202 drug-like chemicals. As well as demonstrating the high prediction accuracy of ANNs the study also showed that only a limited number of molecular descriptors were required to capture the structural features of the studied chemicals.

Another potential area of research concerns the similarity amongst metabolites generated along particular metabolic pathways. Hattori *et al*. (2003) performed a series of cluster analysis classifications of 9,383 compounds taken from the compound section of the KEGG biological database. The 2D structures of the chemical compounds were treated as labelled graphs with the atoms as vertices and the covalent bonds as edges. The data set contained 977 drug-related compounds, 2,649 secondary plant metabolites and 5,757 other metabolites. It was shown that a strong correlation existed between the structural similarity and the pathway connectivity of metabolites, with similar compounds linked together in localised regions of the metabolic pathways.

To affirm ANN modelling as a powerful tool for the classification of metabolites and other small molecules, a study published in 2006 by Karakoc and his team compared several data mining techniques with ANNs for their ability to accurately classify five different types of chemical compounds. The data set of small-molecules included 520 antimicrobial compounds, 959 general drugs, 1,202 drug-like chemicals, 562 bacterial

metabolites, and 1,104 human metabolites. ANNs were shown to provide the best performance results when compared with Linear Discriminant Analysis, Multiple Linear Regression and *k*-Nearest Neighbour approaches.

In a more recent study published in 2007, Karakoc *et al*. developed a number of binary classifiers based on the *k*-nearest neighbour algorithm. These classifiers were trained to distinguish between metabolic substances isolated from human, bacterial, plant and fungal cells. The set of plant metabolites consisted of 2,351 natural compounds characterised from plant isolates and taken from the commercially available AnalytiCon-Discovery Company database.

In both Karakoc studies, the results showed that only a limited overlap existed in the descriptor space between the group of human metabolites and the other groups of molecules. Crucially however, neither paper is clear on how many overlapping molecules were removed from the data set prior to the QSAR analysis. In particular, a considerable overlap would be expected between the metabolomes of plant, humans and bacteria. As much of the metabolism in animals, plants and microbes is focused on making amino acids, carbohydrates and lipids, a striking feature of metabolism is the similarity of the basic metabolic pathways in vastly different species.

Despite the cursory use of plant metabolites in the investigation by Karakoc *et al*. (2007) and Hattori *et al*. (2003), so far there have not been any significant studies which have sought to perform a comprehensive classification analysis of the biochemical compounds in plants. However, much has been made of the potentially useful information that might result. In a recent paper by Last *et al*. (2007), plants are

praised for their "traditional and crucial" roles in food, fibre and pharmaceuticals. Cataloguing all the metabolites that are synthesized by plants is cited as a way of providing a better understanding of the plant metabolome.

The initial aim of this study has been to conduct an analysis of metabolites taken from two organisms: Human and the model plant organism *Arabidopsis thaliana*. Arabidopsis is a small flowering plant, closely related to many crop plants such as cauliflower, cabbage and broccoli. It was the first higher plant to have its genome fully sequenced and is a member of the "Security Council of Model Genetic Organisms" (Garcia-Hernandez *et al*., 2002). Arabidopsis was chosen because of the wealth of compound and pathway information available on it through the Arabidopsis Information Resource (TAIR). Two sets of human metabolites have been compiled from two different databases, the HMDB and KEGG. Two sets of human metabolites were used because of the large discrepancy that was found to exist between the two databases.

Initially, artificial neural networks have been used to construct binary QSAR models which can distinguish between the two types of metabolites. ANNs have been used because of their high performance in similar investigations (Karakoc *et al*., 2006; Cherkasov, 2005a; Cherkasov and Jankovic, 2004). Furthermore, the method of principal component analysis has been used to project the data set of compounds into a 3D Euclidean space. Finally, the common molecular scaffolds and substituents which comprise the groups have been identified using structural similarity searches based on graph theory principles. In this investigation, particular attention has been paid to the extent of the overlap between the two groups of metabolites.

The following four chapters of this introduction delve into the scientific background surrounding the study. *Molecular Descriptors* introduces some of the methods for characterising the properties of molecules in order to produce effective QSAR models. *Artificial Neural Networks* focuses on the challenges present in the design and training of neural networks for the purpose of pattern recognition. *Representing 2D chemical structures* explains the different ways molecular structures can be encoded. Finally, *Similarity Searching* explains how mathematical research into graph theory can be employed to measure the structural similarities amongst a data set of small molecules.

1.1 Molecular Descriptors

1.1.1 Types of Descriptor

Molecular descriptors can be used to characterise the specific properties of the molecules within a data set. Such descriptors vary in the complexity of the information they encode and the time that is required to calculate them. A simple descriptor might be a count of such basic molecular features as hydrogen bond donors, hydrogen bond acceptors, ring systems, rotatable bonds and molecular weight. However, these basic descriptors are unlikely to offer much discriminating power and so are often used in conjunction with other more complex descriptors such as *hydrophobicity* and *molar refractivity*.

Hydrophobicity is an important property in determining the activity and transport of drugs and affects how tightly a molecule binds to a protein and its ability to pass

through the cell membrane. It is usually modelled using the logarithm of the partition coefficient between octanol and water, denoted as *logP*. Determining *logP* experimentally can be difficult, particularly for zwitterionic and very lipophilic or polar compounds (Leach and Gillet, 2005); hence there is much interest in the development of ways to estimate hydrophobicity values. Both hydrophobicity and molar refractivity require experimental data in order to be calculated precisely.

Topological indices such as the *Wiener index* are a category of descriptor based purely on algorithmic constructs (Lowell *et al*., 2001). Molecular structures are characterised according to their size, degree of branching and overall shape. Other descriptors such as *electrotopological state indices* encode electronic and topological characteristics for each atom individually rather than for whole molecules. These values are then combined into a whole-molecule descriptor by calculating the mean-square value over all atoms.

1.1.2 Data Verification

Once the descriptors have been calculated for a set of molecules, it is crucial to pre-process the data using several verification techniques before modelling can take place. One straightforward verification procedure which should be performed on every descriptor is to examine the spread of values for its data set. It is often important that the values of the descriptor follow a *normal* distribution. The coefficient of variation, which is equal to the standard deviation divided by the mean, can be used to assess the spread of the descriptor—the larger the coefficient of variation the better the spread of values. A manipulation of the data in some way may also be necessary prior to any

analysis. It is quite possible that the descriptors will have substantially different numerical ranges, in which case it is important that they are scaled appropriately. Otherwise, a descriptor with a large range of values is likely to overwhelm a descriptor with a small range, and this can bias the results.

The most important of all verification tasks is the elimination of possible cross-correlation between descriptors. This ensures each descriptor is independent from the others. As the number of molecular descriptors available to use increases, so too does the likelihood of finding chance relationships in the data (Labute, 2000). Several methods exist for identifying correlation amongst a set of descriptors. A *pairwise correlation matrix* is a standard method to use when considering a large number of descriptors. It works by quantifying the degree of correlation between all pairs of descriptors. The correlation coefficient, *R*, between a pair of descriptors is calculated using the following equation:

$$R = \frac{\sum_{k=1}^{N}\left[\left(x_{i,k} - mean(x_i)\right)\left(x_{j,k} - mean(x_j)\right)\right]}{\sqrt{\sum_{k=1}^{N}\left(x_{i,k} - mean(x_i)\right)^2 \sum_{k=1}^{N}\left(x_{i,k} - mean(x_i)\right)^2}}$$

Here, *k* is the number of values contained in each descriptor. Each entry *(i,j)* in the correlation matrix is the correlation coefficient between the descriptors x_i and x_j. The correlation coefficient values range from -1.0 to +1.0, with +1.0 indicating a perfect positive correlation and -1.0 indicating a perfect negative correlation. The preferred value of *R* is zero, which indicates that there is no relationship between the variables. In the study by Karakoc *et al*. (2006) which involved the use of molecular descriptors,

all sets that cross-correlated with a value of *R* greater than 0.9 were removed prior to neural network training. However, the paper does not make it clear whether descriptors that cross-correlated with a value of *R* less than -0.9 were also removed from the final descriptor set.

Principal Component Analysis (PCA) is a method for deriving a smaller set of new variables from the original set whilst preserving as much of the relevant information as possible (Gasteiger and Engel, 2003). In a multidimensional data set, each of the principal components is a linear combination of the original descriptors:

$$PC_1 = c_{1,1}x_1 + c_{1,2}x_2 + ... + c_{1,p}x_p$$

$$PC_2 = c_{2,1}x_1 + c_{2,2}x_2 + ... + c_{2,p}x_p$$

$$PC_i = c_{i,1}x_1 + c_{i,2}x_2 + ... + c_{i,p}x_p$$

Here PC_i is the ith principal component, c_{ij} is the coefficient of the descriptor x_j and p is the number of descriptors. The first principal component represents as much of the total variation of all variables within the data as possible, the second principal component accounts for as much of the variance in the data that is not already explained by the first principal component, and so on up to PC_i. Thus, the principal components are constructed in order of declining importance.

Principal components are calculated from a *variance-covariance matrix*. If **A** is a matrix comprised of *n* rows (one for each of the molecules in the data set) and *p* columns (one for each of the descriptors), the variance-covariance matrix is therefore

A multiplied by its transpose $\mathbf{A}^{\mathbf{T}}$ to make a nxn matrix. The eigenvectors of this new matrix represent the coefficients of the principal components. The first principal component corresponds to the largest eigenvalue, and so on. The eigenvalues indicate the proportion of the variance that is explained by each of the principal components. A rule of thumb is to only use the principal components that have eigenvalues greater than one (Leach and Gillet, 2005).

1.1.3 QSAR

It was Hammett who first correlated the electronic properties of organic acids and bases with their equilibrium constants and reactivity when he observed that adding substituents to the aromatic ring of benzoic acid had an orderly and quantitative effect on the dissociation constant (Leach and Gillet, 2005).

$$\log \frac{k}{k_0} = \sigma\rho$$

$$\log \frac{K}{K_0} = \sigma\rho$$

The Hammett equations shown above describe a linear free-energy relationship relating the reaction rate (k) or equilibrium constant (K) for many reactions involving benzoic acid derivatives with *meta*- and *para*-substituents to each other with just two parameters: a substituent constant (σ) and a reaction constant (ρ). The substituent constant is determined by the nature of the substituent and whether it is meta or para to the carboxylic acid ester group on the aromatic ring.

Hammett's success in treating the electronic effect of substituents on the rates and equilibria of organic reactions led Taft to apply the same principles to steric, inductive and resonance effects. The Hammett relation, as a rule, did not hold for the series of aliphatic compounds, so Taft incorporated steric substituent effects into Hammett's formula to produce the following equation:

$$\log \frac{K}{K_0} = \rho \sum_i \sigma^* + \delta \sum_i E_S$$

Here σ^* is a substituent constant depending only on its inductive (polar) influence, and E_S is the substituent constant reflecting its steric effect (Cherkasov, 2005b). Taft's substituent constants are the basis for the 'inductive' descriptors used later in this study. There have been many other extensions to Hammett's original formula. One key development was Swain and Lupton's suggestion that σ values could be written as a weighted linear combination of two components: a field component (F) and a resonance component (R) (Swain and Lupton, 1968). A review of the use of Hammett's equation which contains the more commonly accepted substituent constants is given in the paper "A Survey of Hammett Substituent Constants and Resonance and Field Parameters" (Hansch *et al.,* 1991).

A QSAR (Quantitative structure-activity relationship) attempts to find consistent relationships between the variations in the values of molecular descriptors and the biological activity for a series of compounds. These relationships can then be used to evaluate new chemical entities. QSARs based on Hammett's formula utilize the electronic properties of a chemical structure to produce molecular descriptors. When

Hammett-type relationships were applied to biological systems it was found that it was necessary to use additional molecular descriptors.

Hansch was the first to use QSARs to explain the biological activity of series of structural related molecules using descriptors related to electronic characteristics and to hydrophobicity (Hansch and Fujita, 1964). He proposed that biological activity could be related to the molecular structure via equations of the form:

$$\log(1/C) = k_1 \log P + k_2\sigma + k_3$$

In this equation, C is the molar concentration of compound that is required to produce a standard response in a given time and σ is the appropriate Hammett substitution parameter. Alternative approaches to QSARs based on Hansch's pioneering work generally take the form of a linear regression equation in the form:

$$y = Const + C_1P_1 + C_2P_2 + C_3P_3 + \ldots + C_iP_i$$

Here, the dependent variable y is the property being modelled (such as biological activity) and the independent variables $P_{1\text{-}i}$ are the molecular descriptors (such as *log P* or molar refractivity). The coefficients $C_{1\text{-}i}$ and the constant *Const* are calculated using multiple linear regression techniques. The aim of a multiple linear regression is to minimise the sum differences between the values predicted by the equation and the actual observations. The quality of the solution can be assessed using the squared correlation coefficient (R^2), which gives a value between zero and one. Suppose $\hat{y_i}$ are the values predicted by the linear regression equation by inserting the relevant

independent variables and y_i are the experimental observations that correspond to those variables (e.g., the biological activity of the molecule). The following can then be calculated:

Total Sum of Squares, TSS = $\sum_{i=1}^{N}(y_i - mean(y))^2$

Explained Sum of Squares, ESS = $\sum_{i=1}^{N}(\hat{y_i} - mean(y))^2$

Residual Sum of Squares, RSS = $\sum_{i=1}^{N}(y_i - \hat{y_i})^2$

Thus,

TSS=ESS+RSS

R^2 is given by the following relationships:

$$R^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS}$$

An R^2 of zero corresponds to a poor model where none of the variation in the observations is explained by the variation in the independent variables. An R^2 of one corresponds to a perfect model where all the variation is explained. The general goal of any QSAR experiment is to derive the most economical model with the best performance (Leach and Gillet, 2005).

1.1.4 Methods for Generating QSAR Models

There are many different statistical and computational methods for generating QSAR models. These include genetic algorithms, artificial neural networks, *k*-nearest neighbour, linear discriminative analysis and multiple linear regression. In this study binary QSAR models have been developed through the use of Artificial Neural Networks.

1.2 Artificial Neural Networks

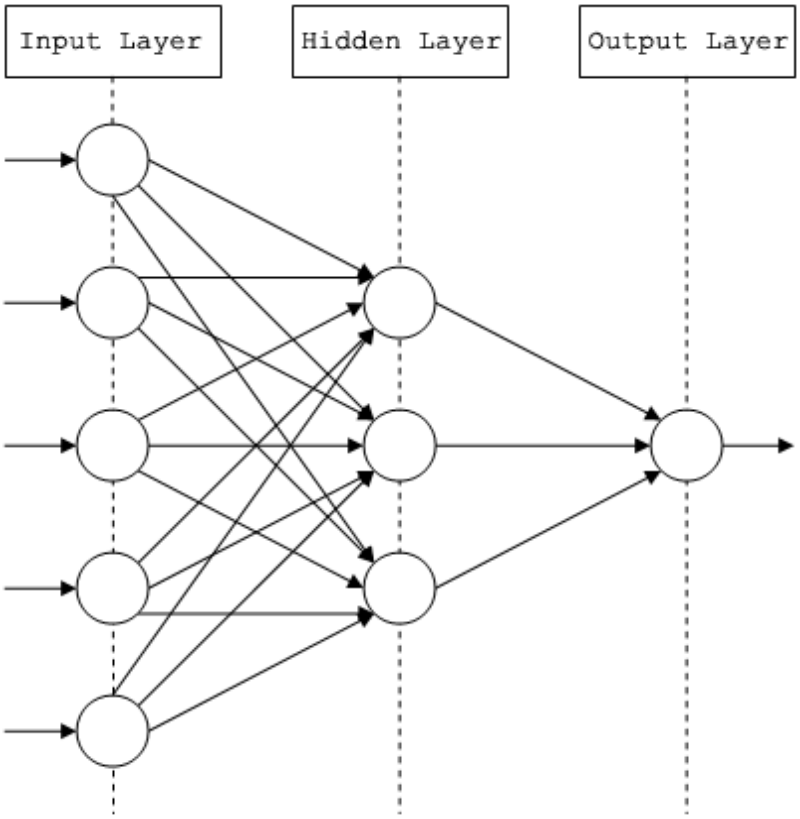
1.2.1 Multi-layer Perceptrons

Hansch and Fujita (1964) demonstrated the use of regression analysis in the design of QSAR models. However, such models assume a linear relationship between biological activity and the variation in the molecular descriptor values. This assumption can lead to a model with limited accuracy. Artificial Neural Networks (ANNs) have the advantage of being able to model non-linearity as well as being able to cope with a large number of potential descriptors.

ANNs are used extensively in many different fields and for many different purposes, ranging from financial forecasting to medical diagnosis. An early example of the use of ANNs in biology was a study published in 1987 by Qian and Sejnowski which used feed-forward neural networks to predict the secondary structure of globular proteins. Feed-forward neural networks consist of layers of artificial neurons with

connected channels running between all pairs of neurons in adjacent layers. These types of network are known as Multi-layer Perceptrons (MLPs) and are represented in Figure 1.1.

Figure 1.1. Schematic diagram of a fully-connected, three layer, feed-forward, multi-layer perceptron.

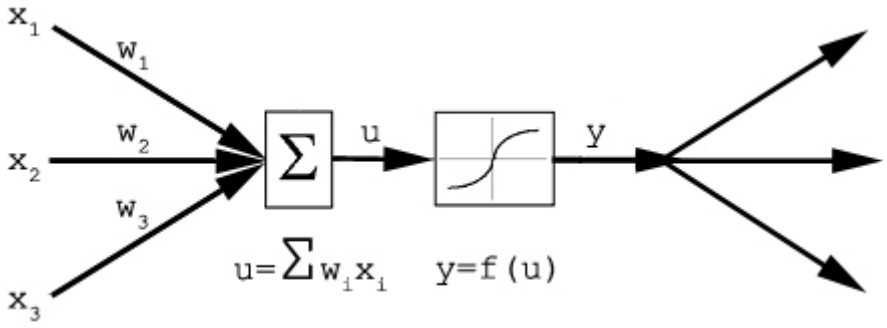


Each artificial neuron in the network (see Figure 1.2) is composed of four parts:

1. *Some **input weights** ($w_1, w_2, \ldots, w_i$), which weight the incoming signals ($x_1, x_2, \ldots, x_i$).*
2. *A **summation unit**, which collects the incoming signals ($u = \sum w_i x_i$).*

3. *A **non-linear function** of some kind, which processes the incoming signals, known as the activation function (y = f(u)).*
4. *An **emit unit**, to transmit the new signal (y).*

Figure 1.2. The configuration of an artificial neuron.



MLPs work by processing information in the following manner: At the input layer, each neuron accepts an input signal from a specific channel and delivers that signal to the other channels according to the delivery strength, which is quantified by the weights (there is no information processing for an input neuron). In the hidden layer, each neuron accepts the signals from the input neurons and likewise delivers the signal to the other channels, quantified by the weights. In the output layer, each neuron accepts the signals from the hidden neurons, processes them, and then makes a decision based on the magnitude of the processed signals (Bishop, 1995).

The presence of a hidden layer of nodes is a key feature, allowing the MLP to compute complex non-linear problems. There are many theoretic studies on the subject of MLP model structure, and several rules of thumb prevail (Bishop, 1995). These rules are often used in conjunction with trial-and-error testing in order to determine the most effective model structure. By running a variety of models, each

with a different number of hidden neurons, the model with the lowest error and the simplest structure can be found.

1.2.2 Training & Validation

Each neuron in the MLP uses a continuous activation function known as a *sigmoid function* to compute the signals it receives from other nodes. Because this activation function is differentiable, many powerful algorithms exist which can be used to train the network. The traditional method used for training is known as the *back-propagation algorithm*. The back-propagation algorithm is a supervised learning method. This means that the desired output must be known for each input pattern in the training set.

To initialise training, the weights between nodes are set to random values. Training the MLP is then a matter of optimising the weight using the back-propagation algorithm. First the MLP receives a set of inputs which are multiplied by each neuron's weights. The products are then summed for each neuron and the activation function is applied. The signals from the input neurons are then propagated forwards from the input layer to the output layer. At the next stage, the errors occurred at the output layer (the difference between the desired output and the actual output) are propagated backwards from the output layer to the input layer. The weights are then adjusted so that the next time the network sees the same set of inputs it will come closer to the desired output. This process is iterated for every set of input patterns in the training set. A particular drawback of the back-propagation algorithm is its slow

speed—the *scaled conjugate gradient* introduced by Møller in 1993 is one of a number of faster alternatives.

Once the neural network has been trained it can be used to predict the output values for a new set of input patterns. The ability of a trained neural network to work well on this novel data depends on its ability to generalise. There are two important issues affecting this ability; *under-fitting*, which means that the model is too simple, and *over-fitting*, which means that the model is too complex. To evaluate a model's ability to generalise well, its performance should be tested with an independent *validation* data set. By minimising the error on this validation set, over-fitting can be prevented. To prevent over-fitting again, this time to the validation set, the performance of the network and its reliability can be evaluated with a third independent *test* data set.

In practice, the availability of data with known output can be severely limited, so partitioning the data set may not always be possible for model comparison purposes. In such cases the procedure of *cross-validation* can be adopted. Here, the training set is divided at random into n distinct segments. The network is then trained using n-1 of the segments and its performance is evaluated using the remaining segment as a validation set. This process is repeated for each of the n possible choices for the segment which is omitted from the training process, and the test errors are averaged over all n results. This procedure allows the network to be trained using a high proportion of the available data, while also making use of all data points in evaluating the cross-validation error. The disadvantage of this approach is that it requires the training process to be repeated n times, which often requires large amounts of processing time (Bishop, 1995).

The results of a neural network are typically summarised in a table known as a *confusion matrix* or by drawing the *receiver operating characteristic* (ROC) curve for the trained network. ROC curves are a plot of the true positive fraction (the number of correctly classified positive inputs divided by the total number of positive inputs) against the false positive fraction (the number of misclassified negative inputs divided by the total number of negative inputs) for all possible threshold values (i.e. the cut-off point for positive/negative predictions). ROC curves are a good visual measure of system robustness with a large area under the ROC curve indicating good model performance.

1.2.3 Examples

In the study by Karakoc *et al*. (2006) five neural networks were used, one for each of the five groups of small-molecules under study. The neural networks had 41 input nodes each, which corresponded to the 41 QSAR descriptors pre-calculated for each molecule in the data set. The values for each descriptor were "normalised" within the range of 0 and 1. The design of each neural network consisted of a single layer of 10 hidden nodes and a single output node. The group of molecules that a neural network was being trained to recognise were assigned a target output of 1.0. Molecules from the remaining four groups were assigned an output of 0.0.

The initial weights of the neural networks were randomly assigned in a range between -1 and 1. The molecular data set was randomly separated into a non-overlapping training and testing set of 70% and 30% respectively. In total an average of 20

independent training and testing runs were conducted. A rigorous type of cross-validation known as a *leave-one-out* analysis was conducted to validate the developed neural network models. This involved testing each model a total of 4,346 times, once for every molecule in the molecular data set. The trained neural networks produced the following results:

Category	**Accuracy in training**	**Accuracy in testing**
'Antimicrobial Compounds vs. Others'	97%	97%
'Bacterial Metabolites vs. Others'	96%	95%
'General Drugs vs. Others'	88%	89%
'Drug-like Chemicals vs. Others'	93%	94%
'Human Metabolites vs. Others'	100%	100%

It is worthwhile noting that the *k*-nearest neighbour approach, which was also used to develop QSAR models for the same data in the study, showed a similar accuracy to the ANNs. A further investigation into the number of false-positive predictions produced by the models showed a remarkable similarity between the general drugs and the drug-like chemicals, as well as up to 56% cross-recognition between bacterial metabolites and antimicrobial compounds.

Neural networks were used by Cherkasov (2005a) to distinguish between antimicrobial compounds, conventional drugs and drug-like substances. In line with the study by Karakoc *et al*., a separate neural network was created for each molecular group. In addition, a single neural network to simultaneously recognise all three groups was also developed. This particular neural network consisted of a 30-10-2

configuration of nodes. However, when compared with the other neural networks, this model showed less sensitivity to the data.

1.3 Representing 2D Chemical Structures

Two-dimensional chemical structure diagrams are used extensively by chemists to represent chemical compounds. These diagrams lend themselves naturally to graph representation whereby vertices in the graph refer to the atoms in the chemical structure and edges refer to the bonds between the atoms. However, in order to search through and compare the molecular graphs of a large number of compounds using computational techniques, the graphs must first be represented in a machine readable format.

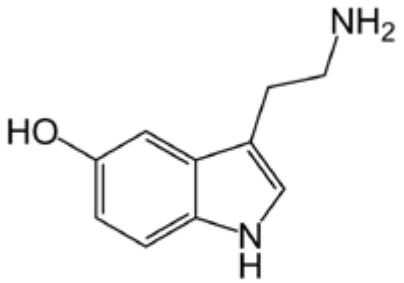
1.3.1 Connection Tables

A common method for reading a molecular graph to and from a computer is to use a *connection table*. Connection tables have been the predominant form of chemical structure representation in computer systems since the early 1980s (Gasteiger and Engel, 2003). A simple connection table consists of a list of the atomic numbers of the atoms in the molecules and a list of the bonds, specified as pairs of bonded atoms. Hydrogen atoms are not always explicitly included in the connection table so as to save on storage space. A connection table can be extended by adding other lists, such as lists of the free electrons and/or the charges on the atoms of the molecule.

1.3.2 The SMILES format

An alternative to using a connection table to represent a molecular graph is to use a *linear notation*. Linear notations represent the structure of chemical compounds as a linear sequence of symbols which indicate the sections of the chemical structure and the way they are connected together. The Simplified Molecular Input Line Entry Specification (SMILES) notation was created by David Weininger in 1986 for the purpose of chemical data processing (Weininger, 1988). The SMILES language has since found widespread distribution as a universal chemical nomenclature for the representation and exchange of chemical structure information. In SMILES, atoms are represented by their atomic symbol, which occurs in upper case for aliphatic atoms and lower case for aromatic atoms. Neighbouring atoms appear next to each along the sequence, with hydrogen atoms omitted. Double bonds are written using the symbol '=' and triple bonds using the symbol '#'. Rings are dealt with by breaking one of the bonds in the ring and appending an integer to the two atoms of the broken bond. The presence of a branch point is indicated using a left-hand bracket followed eventually by a right-hand bracket to indicate that all the atoms in the branch have been visited (see Figure 1.3 for an example).

Figure 1.3. The chemical diagram for the metabolite Serotonin and its corresponding SMILES string.



NCCC1=CNC2=C1C=C(O)C=C2

There are many ways to construct the SMILES strings for a given molecule depending on the atom which is used as a starting point and the path which is followed through the molecule. This means that there can be a large number of valid generic SMILES strings representing the same structure in a database. A *canonical* SMILES string consists of a unique ordering of the atoms for a given molecular structure. A widely used method for determining such a canonical order of the atoms is the *Morgan algorithm* (Morgan, 1965), which relies on the iterative calculation of "connectivity values" to enable differentiation between the atoms. Initially, each atom is assigned a connectivity value equal to the number of atoms connected to it. In subsequent iterations a new connectivity value is calculated as the sum of the neighbouring atoms' own connectivity values. This procedure is continued until the number of different connectivity values reaches a maximum. The atom with the highest connectivity value is then chosen as the first atom in the SMILES string, with its neighbours listed in the descending order of their own connectivity values.

The term *isomeric SMILES* collectively refers to SMILES strings written using rules that can specify isotopism, double bond configuration, and chirality. A molecule is said to be chiral if it contains an asymmetric center, known as a *chiral atom* or *chiral center*, and can thus occur in two non-superimposable mirror-image forms. The

simplest and most common kind of chirality is tetrahedral; four neighbouring atoms are arranged evenly about the central chiral atom. If all four neighbours are different from each other in any way, mirror images of the structure will not be identical. The two mirror images are known as *enantiomers* and are the only two forms that a tetrahedral centre can have. If two or more of the four neighbours are identical to each other, the central atom will not be chiral as the mirror images of the molecule can then be superimposed onto each other.

In the SMILES format, tetrahedral centres may be indicated by a simplified chiral specification using the symbol '@' or '@@' written as an atomic property following the atomic symbol of the chiral atom. There are many kinds of chirality other than tetrahedral which can also be represented using the SMILES format.

1.3.3 Binary Representation

Representing a molecule's chemical information in a binary form allows for rapid comparisons to be made between molecules in large databases. Such binary or *bitstring* representations of molecules consist of a sequence of '0's and '1's which can vary widely in size, from a few tens or hundreds of bits to several thousand bits.

1.3.4 Structural Keys

In the *structural key* method of binary representation, a single bit or multiple bits in the bitstring relate to a molecular substructure that exists in a pre-compiled *fragment dictionary* which the molecule is first checked against. If the substructure is present in

the molecule then the relevant bit is set to ‘1’. A ‘0’ in the bit string means that the substructure it relates to isn’t present in the molecule. Generating a structural key is time-consuming due to the number of substructure searches which need to be performed. For every molecule in the database, a substructure search for each pattern in the fragment dictionary must be made in order to create a string for that molecule. Examples of patterns contained in the fragment dictionary may include:

- *The presence or absence of an element.*
- *Unusual or important electronic configurations, e.g. triple-bonded nitrogen.*
- *Rings and ring systems.*
- *Common functional groups, e.g. hydrocarbons.*

1.3.5 Hashed Fingerprints

An alternative to the structural keys representation is to use *hashed fingerprints*. Like structural keys, hashed fingerprints consist of bitstrings, but unlike structural keys, a hashed fingerprint does not require a pre-defined fragment dictionary. The fingerprint is produced by generating all possible linear paths of connected atoms through the molecule that contain between 1 and a defined number of atoms (typically seven). For example, the molecule **OC=CN** would generate the following patterns:

0-bond paths:	**C**	**O**	**N**
1-bond paths:	**OC**	**C=C**	**CN**
2-bond paths:	**OC=C**	**C=CN**	
3-bond paths:	**OC=CN**		

Every path (or pattern) through the molecule, up to the path length limit, is generated. Because there is no pre-defined set of patterns, and due to the huge number of potential patterns, it is not possible to assign a particular bit in the bitstring to each pattern as is the case with structural keys. Instead, each pattern serves as the input to a second program which uses a hashing procedure to set a small number of bits (typically four or five) to '1' in the bitstring. Hashed fingerprints are particularly associated with *Daylight Chemical Information Systems, Inc* (see Further Information). Typically, organic molecules are set between 50 and 400 bits out of a total bit string length of 1,024. Like structural keys, simple Boolean operations on fingerprints can be used to compare the fingerprints of two separate molecules, which allows for extremely fast substructure and similarity searching.

1.4 Searching 2D Chemical Structures

In a database of chemical structures, simple tasks such as searching for a particular molecule to see if it is present can be achieved by directly comparing the hashed fingerprints of the molecules. Substructure and similarity searches are more complicated types of searching and require the use of sophisticated algorithms.

1.4.1 Substructure Searching

A common search requirement is the identification of molecules in the database which contain a specific substructure such as a particular functional group or sequence of atoms. When thinking in terms of a graph representation of molecules, substructure searching is equivalent to determining whether one graph is entirely contained with

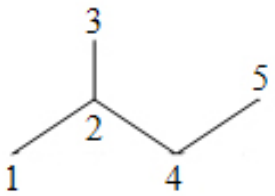
another—a problem known as *subgraph isomorphism*. A "brute-force" approach to subgraph isomorphism involves testing every possible way to map the nodes of the substructure graph onto those of the database molecule, and then testing whether the corresponding bonds also match (Leach and Gillet, 2005). However, this approach is *NP-complete* (Non-deterministic Polynomial time complete), meaning that the amount of time which is needed to compute the solution varies exponentially with the size of the problem (for subgraph isomorphism this corresponds to the number of nodes in the graph). Fortunately, there are more efficient methods for substructure searching available.

The *back-tracking mechanism* initially reported by Ray and Kirsch in 1957 is one of the earliest algorithms developed for the problem of subgraph isomorphism. The algorithm first attempts to map a node from the substructure graph to a matching node in the molecule graph based on atom types. It then attempts to map a neighbouring node of the substructure node to a neighbouring node of the molecule graph. This process is continued until all nodes have been successfully matched or until a match fails, at which point the algorithm back-tracks to the last successful match and attempts an alternative mapping. This procedure is known as depth-first searching.

The *Ullmann algorithm* is a more efficient back-tracking algorithm which works by using *adjacency matrices* to represent the graphs of the query substructure and the molecules in the database (Ullmann, 1976). The adjacency matrix of a molecular graph is a matrix with rows and columns corresponding to the atoms in the graph with a '1' or '0' in position (i, j) of the matrix depending on whether i and j are adjacent within the graph or not (Leach and Gillet, 2005). Therefore, for a simple chemical

structure without rings, the corresponding adjacency matrix will have ‘0’s along its diagonal (see Figure 1.4).

Figure 1.4. The adjacency matrix for a simple 5-atom “molecule”.



	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	1
6	0	0	0	0	1	0

The Ullmann algorithm works by first creating a new “matching” matrix with the number of columns equal to the number of atoms in the database molecule currently being searched and the number of rows equal to the number of atoms in the substructure being searched for. The elements in the matching matrix take the value ‘1’ if there is a match between the corresponding pair of atoms and ‘0’ otherwise. If each row in the matching matrix contains no more than one element equal to ‘1’ and each column contains no more than one element equal to ‘1’ then a unique mapping has been found.

The Ullman algorithm improves its efficiency by incorporating a *relaxation* step into its depth-first search. This consists of a highly discriminating “pruning” technique which takes into account the neighbouring atoms when considering possible matches

and prevents prolonged searches that cannot lead to a better solution than the currently existing one.

1.4.2 Similarity Searching

Substructure searching requires sufficient knowledge to be able to construct a meaningful substructure to search for. However, it may not always be known which the most relevant structures to search for are. Also it may be desirable to rank molecules in terms of the degree to which they compare to the substructure rather than simply whether they contain it or not. In such cases similarity searching may be a more suitable method. The rationale for similarity and substructure searching lies in the similar property principle which states that structurally similar molecules tend to have similar properties.

Many of the types of molecular descriptors described earlier, such as structural keys, atom-pairs and topological indices can be used as a means for computing the similarity between molecules. The most commonly used method for similarity searching is based on the hashed fingerprints form of 2D structure representation previously described. The similarity between two molecules represented by binary fingerprints can be quantified using the *Tanimoto coefficient*. The Tanimoto coefficient works by giving a measurement of the number of fragments in common between the two molecules. The similarity between the molecules A and B based on their hashed fingerprints is given by:

$$S_{AB} = \frac{c}{a + b - c}$$

Where there are *a* bits set to '1' in molecule *A*, *b* bits sets to '1' in molecule *B*, and *c* '1' bits common to both *A* and *B*. The value of the Tanimoto coefficient ranges from 0 to 1, where a value of 1 indicates that the molecules have identical fingerprint representations, and a value of 0 indicates that there is no similarity. One reason for the popularity of the Tanimoto coefficient is that it is not bias towards larger molecules, which tend to have more bits set to '1' in their bitstring compared to smaller molecules. This is because the equation for the coefficient has the affect of normalising the sizes of the molecule being compared (Leach and Gillet, 2005).

The challenges relating to searching within a database of molecular structures are dealt with in the area of mathematics known as *graph theory*. Research into applying established graph theory approaches to 2D chemical structures was started in Sheffield in 1979 when the graph representation of chemical structures meant that *isomorphism* algorithms could be used for searching (Mitchell *et al*., 1990).

In graph theory, a *subgraph* is defined as any subset of the nodes and edges in a graph (for instance, the molecular graph for benzene is a subgraph of the graph for aspirin). Isomorphism algorithms seek to compare the structural relationships which exist between pairs of graphs by identifying the various types of subgraph. Previously referred to *subgraph isomorphism* algorithms search for the presence of a particular subgraph within a graph and are therefore suitable for substructure searching. For similarity searching, *maximal common subgraph isomorphism* algorithms locate the largest section two structures have in common (Mitchell *et al*., 1990).

The maximum common subgraph (MCS) defines the largest set of atoms and bonds that two structures have in common. The number of atoms and bonds in the MCS can be used to compute a Tanimoto-like coefficient which quantifies the degree of similarity between the two compounds (Leach and Gillet, 2005). A number of both exact and approximate methods have been devised for identifying the MCS.

The first application of the MCS to database searching was described by Hagadone in 1992, where a two-stage approach was used. First a substructure search for pre-defined molecular fragments was used to provide an upper-bound to the size of the MCS. The database was then sorted according to the upper-bounds, thus restricting the more costly MCS calculation to those structures above a given threshold. More recently, a new algorithm called RASCAL has been described by Raymond and Willett (2002) which is able to perform tens of thousands of comparisons a minute. In this case, speed-up is achieved through the use of chemically relevant heuristics, a fast implementation of the clique detection process that underlies MCS detection, and the use of a very efficient screening process which prevents many of the more costly MCS comparisons from being performed.

Even for small molecules like metabolites, the application of graph-matching algorithms can be time consuming. In Nobeli *et al*. (2003) searching for the MCS between each molecule in the data set of 745 E. coli metabolites meant a total of 277,885 individual comparisons had to be made. A modified version of the Ullman algorithm was used to compute the similarity value for each pair. This algorithm was run on a dedicated 1.7 GHz processor with 1GB of memory. However, because of the

combinatorial nature of the problem only 71% of all the calculations could be completed in time.

2 Materials and Methods

2.1 Software

The Java code written specifically for this study consists of several packages which incorporate the Chemistry Development Kit (CDK) and the open source artificial neural network software that was used to create the QSAR models. The CDK is an open-source Java class library distributed under the GNU Lesser General Public License, and consists of data structures and algorithms intended as an aid in structural chemoinformatics and bioinformatics research. Initially, Java code was written for the purpose of extracting the collection of individual metabolites from the three metabolite databases. To keep track of the molecules and to be able to link them to their original data sources, all metabolites were given unique identifiers (in the case of HMDB and KEGG human metabolites, the original database compound IDs were used). In this project, the tools available in the CDK have been used whenever possible as part of a personal commitment to open-source software. The neural network software used to train the ANNs was based on a free open-source Java package, the details of which can be found in the References section. The complete set of Java code written for use in this study can be found on the supporting CD.

The same Scientific Vector Language (SVL) scripts that were used by Karakoc *et al*. (2006) were used in this study to calculate 'inductive' molecular descriptors and to

perform the fragmentation analysis. SVL is a simple scripting and application programming language with similarities to the statistical programming language R. Specifically, the scripts “inductive_descriptors.svl” and “fragmentation.svl” were used. These scripts are available for licensed users through the SVL exchange community (see References). The SVL scripts were run using Chemical Computing Group's Molecular Operating Environment (MOE), an interactive computing and molecular modelling tool that can be controlled through a graphical user and command line interface. A useful feature of the MOE is its ability to automatically construct 2D representations of molecules based on their SMILES notation.

2.2 Data Sets

To begin the study, three groups of metabolites were taken from three different online databases—the details of which are given below. Empty data fields from the source databases meant that SMILES information was not present for some metabolites. In these cases, as SMILES was the only means for determining the structure of the molecules, any metabolites without SMILES strings had to be ignored.

2.2.1 The Human Metabolome Database

The initial data set of human metabolites was extracted from the HMDB. The HMDB is a freely available electronic database containing information about small molecule metabolites found in the human body and is developed and maintained by the Wishart group at the University of Alberta. The database accumulates the information obtained by the Human Metabolome Project, which was launched in 2004 as part of

an effort to identify and quantify all detectable metabolites which exist in the human body. The criteria for inclusion in the HMDB are as follows.

- *The compound must weigh less than 1,500 Daltons (Da).*
- *The compound should be found at concentrations greater than 1 μM (either in normal or in diseased conditions) in one or more biofluids/tissues.*
- *The compound should be of biological origin.*

In the HMDB, some exceptions to these rules are made. These include biomedically important metabolites of low abundance (such as hormones, disease-associated metabolites, essential nutrients and signalling molecules), certain very common drugs (E.g. nicotine), and some ubiquitous food additives (E.g. vitamins) (Wishart *et al*., 2007).

Specific information on the compounds catalogued in the HMDB was taken from the HMDB's *MetaboCards* data set, which contains detailed information on 2,754 chemical structures. These are divided into the following taxonomy families:

- *Mammalian Metabolites*
- *Microbial Metabolites*
- *Drug Metabolites*
- *Cosmetic/Drug Additives*
- *Lipids*
- *Plant Metabolites*
- *Food Additives*

- *Drugs*
- *Synthetic/Industrial Chemicals*

The presence of 50 plant metabolites in the HMDB was surprising, but a closer inspection of the metabolites revealed that they were mostly flavanoids—a class of secondary plant metabolite—and that they originate from fruits and vegetables found in a typical human diet (For example, one such plant metabolite, *Naringin*, can be found in grapefruit.) Only the mammalian metabolites were extracted from the data set, and a total of 2,262 metabolites was obtained.

A single entry in the HMDB contains an average of 90 separate data fields with half of the information devoted to chemical or physicochemical data and the other half devoted to biological or biomedical data. The MetaboCards record two SMILES strings for each molecule; a SMILES string corresponding to an *isomeric* (chiral) metabolite structure and a SMILES string corresponding to a *canonical* (non-chiral) metabolite structure. The SMILES strings corresponding to the isomeric metabolite structure were used in preference to the canonical representation except when the isomeric representation was not present. This was done to retain chemical information concerning the isotopism, double bond configuration, and chirality of the molecules.

2.2.2 The Arabidopsis Information Resource

Plant metabolite information was taken from version 4.0 of the AraCyc pathway database. The pathway database is maintained by The Arabidopsis Information Resource (TAIR) which acts as a comprehensive and centralised resource for the plant

Arabidopsis thaliana (Garcia-Hernandez, 2002). The version of the database used in this study contains 285 metabolic pathways with a total of 1,879 unique genes assigned to the pathways, thus reflecting the current knowledge of the metabolism of Arabidopsis.

Name and structure information on plant metabolites was extracted from the compounds data file which contains all the compounds found in AraCyc pathways. 1,006 plant metabolites remained after duplicate entries were removed based on literal matches between SMILES strings.

2.2.3 KEGG

A second set of 881 human metabolites was taken from pathway information available in the KEGG database. Initiated by the Japanese human genome programme in 1995, KEGG integrates both biochemical and genetic information. It contains a catalogue of all the chemical compounds found within the cell, as well as reaction information and pathway maps describing the potential networks of metabolic activity. KEGG records many metabolic networks, mostly representing intermediary metabolism; a core portion of the metabolic network that is shared and conserved in many different organisms (Hattori *et al*., 2003). KEGG can also generate an idealised metabolic pathway for a particular organism by matching the proteins of that organism to enzymes within the reference pathways. The KEGG data set used was originally assembled for use in a study by Macchiarulo and Nobeli (2007).

2.3 Optimizing Molecular Structures

The MOE was used to create 3D conformations of the molecules based on 2D structure representations constructed from their original SMILES string. This was necessary because the 'inductive' descriptors as described in chapter 2.4 rely on such 3D representation. The molecule objects were first "washed" and then energy minimized using a series of algorithms. These were applied in the following order:

1) To put the molecules in a reasonably solvated state, the *Wash Molecule* application was used. During washing, hydrogen atoms and lone pairs are removed from all molecules. Counter ions, oxygen ions, simple acids and bases, common solvents, and isolated fragments of common salts are also removed. The bond lengths of the structures are scaled to approach equilibrium values and the atom ionization for all structures is set to formal charge. Then for each structure, acids are de-protonated and bases are protonated. Explicit hydrogen atoms are then added.
2) Atomic partial charges were computed using the MMFF94 forcefield (Halgren, 1996). The parameters in the MMFF family of forcefields are derived from computational data and have been known to perform well for a wide range of organic chemistry calculations (Karakoc *et al*., 2006; Labute, 2000).
3) Energy minimization of the structures was carried out in order to find the best nearby conformation. The MMFF94 forcefield was used to energy minimize the structures to an RMS gradient less than 0.1. Using gradient optimization, atoms are moved so as to reduce the net forces on them. The existing chirality of the molecules is preserved during this process.

2.4 Calculating Molecular Descriptors

The optimized molecule objects were used to calculate 184 2D molecular descriptors and 50 'inductive' descriptors using the MOE. In the MOE, 2D molecular descriptors are defined to be numerical properties that can be calculated from the connection table representation of a molecule. This includes elements, formal charges and bonds, but not atomic coordinates. The 2D descriptors were therefore not dependent on the conformation of the molecules.

The 'inductive' descriptors were calculated using a custom SVL script. These types of descriptor have been used successfully in previous artificial neural network-based QSAR studies to model the antimicrobial activity of organic compounds (Cherkasov and Jankovic, 2005b) and cationic peptides (Cherkasov and Jankovic, 2005b; Cherkasov and Jankovic, 2004), as well as to construct predictive binary models to recognise the substances involved in bacterial and human metabolism (Cherkasov and Jankovic, 2005a; Karakoc *et al*., 2006). A full definition of the descriptors as well as a review of other molecular modelling investigations which have been successfully performed with the same QSAR parameters can be found in the paper "'Inductive' descriptors: 10 successful years in QSAR" (Cherkasov and Jankovic, 2005b).

In summary, the common characteristic of the 'inductive' descriptors is that they are all related to atomic electronegativity (χ), covalent radii (R), and intramolecular distances (r) and can be derived from the formulas for steric Rs and 'inductive' σ^* parameters used to calculate Taft's substituent constants (see chapter 1.1.3) as well as

'inductive' electronegativity χ, 'inductive' partial charge ΔN, and 'inductive' analogues of chemical hardness η and softness s.

It was not possible to calculate 'inductive' descriptors for all molecules as the SVL script threw unknown exceptions for certain structures. This was thought not to have affected the accuracy of the descriptors for the molecules which could be calculated but nonetheless meant a further reduction in the size of the data set. The final number of molecular structures with all descriptors calculated was broken down as follows; 1,994 HMDB human metabolites, 843 KEGG human metabolites and 865 AraCyc plant metabolites.

2.5 Measuring Molecular Similarity

The next stage of the study involved identifying the overlap between the three molecule groups. The CDK facilitates a number of standard techniques for assessing the similarity between molecules based on their SMILES strings. For the remainder of the study, the original SMILES strings of the metabolites were replaced with those generated by the MOE from the optimized molecular conformations.

To perform the similarity test it was necessary to convert the SMILES representation of the metabolites into molecule objects that could be handled in the CDK. In the CDK, molecules are represented using a standard valence model which is capable of understanding the normal valences of organic compounds, and can fill in unspecified hydrogens as well as detect aromatic and anti-aromatic ring systems by counting the bonding electrons in a molecule. The CDK represents molecules as graphs consisting

of atom nodes and bond edges. Each atom object in the CDK has several properties which include its atomic number, atomic weight, charge, and the number of attached hydrogens. If the atom is a chiral center—such as a carbon, phosphorous or sulfur atom—it can also have chiral specifications. Bond objects between the atoms have the property of being single, double, triple or aromatic.

Once the SMILES strings had been converted into molecule objects in the CDK, it was possible to use the CDK's *Fingerprinter* class to generate hashed fingerprints for each of the 3,702 molecular structures. The Fingerprinter class works by generating all sequences of atoms up to 7 atoms long which are present in the molecule object. This is done using a depth-first search. The class then generates a hash code for each sequence using the built-in hash function of the Java programming language. This results in a binary vector where each bit indicates the presence or absence of a particular substructural fragment within the molecule.

Due to the hashing function, it is not 100 % guaranteed that different molecules will map to different fingerprints, but because the number of molecular patterns contained in the fingerprint is exhaustive, the chances are favourably high. In this study, the initial size of each fingerprint was set to 1,088, which was considered large enough to accurately represent each of the metabolite structures in the data set.

The generation of the fingerprints meant that the similarity between any two molecules in the data set could be measured using the Tanimoto coefficient. The Tanimoto coefficient was chosen because it is the most widely used similarity coefficient for binary fingerprints.

To remove any redundancy within each of the three groups; KEGG, HMDB and AraCyc, each molecule was compared to every other molecule in its own group. If two molecules shared a Tanimoto coefficient of 1.0, one of the two molecules was removed, leaving one remaining 'unique' molecule. This resulted in a data set of 1,491 HMDB human metabolites, 744 KEGG human metabolites, and 758 AraCyc plant metabolites. The three groups of metabolites were then separated into the following two data sets.

1) KEGG human metabolites vs. AraCyc plant metabolites.
2) HMDB human metabolites vs. AraCyc plant metabolites.

The overlap between the human and Arabidopsis metabolites was removed from both data sets prior to neural networking training. As the neural network's job is to separate the two molecule types within a multidimensional descriptor space, assigning the same molecule to both groups greatly reduces the performance of the classifier. It is much more interesting experimentally to remove the overlapping metabolites and analyse them separately. An analysis of the overlap between the three groups is provided in the results section.

2.6 Artificial Neural Networks

Unlike linear QSAR modelling ANN-based experiments are not always easy to interpret in conventional terms. Therefore to visually represent to some degree the separation between the human and plant metabolites within the molecular descriptor

space, a principal component analysis was performed on each of the data sets and the top three principal components were plotted using the MOE's 3D Plot application.

2.6.1 Choosing Molecular Descriptors

Before neural network training could take place, a strict elimination of the correlation between the molecular descriptors was performed. This was necessary because one or more highly correlated descriptors used in combination for training might result in an over-representation of the information they characterise and could adversely affect the performance of the neural network models. Values for all descriptors were first scaled within the range 0 – 1. This was done using the following equation for each value *d* in each set of descriptor values:

$$d^{new} = \frac{d - d^{\min}}{d^{\max} - d^{\min}}$$

Here d^{max} is the maximum value in the descriptor set and d^{min} is the minimum value. Based on these scaled values, a pairwise correlation matrix was calculated to quantify the degree of correlation, *R*, between every pair of descriptors. In order to choose which descriptors to remove from the set of descriptors, an algorithm was written that could systematically remove descriptors from the correlation matrix which cross-correlated with an *R* value that fell outside a set lower and upper bound. Five sets of descriptors were defined by gradually decreasing the range of acceptable values for *R*:

Descriptor Set 1: All descriptors with an *R* value greater than 0.9 and less than -0.9 were removed. A total of 100 descriptors remained.

Descriptor Set 2: All descriptors with an *R* value greater than 0.8 and less than -0.8 removed. A total of 68 descriptors remained.

Descriptor Set 3: All descriptors with an *R* value greater than 0.7 and less than -0.7 removed. A total of 51 descriptors remained.

Descriptor Set 4: All descriptors with an *R* value greater than 0.6 and less than -0.6 removed. Total of 30 descriptors remained.

Descriptor Set 5: All descriptors with an *R* value greater than 0.5 and less than -0.5 removed. A total of 25 descriptors remained.

A sixth descriptor set was compiled based purely from descriptors defined in the paper 'A Widely Applicable Set of Descriptors' (Labute, 2000). In this paper, three sets of "VSA" descriptors are defined and are all based on individual atomic contributions to a small number of molecular properties:

SlogP—VSA$_k$. 10 descriptors intended to capture hydrophobic and hydrophilic effects.

SMR—VSA$_k$. 8 descriptors intended to capture polarizability.

PEOE—VSA$_k$. 14 descriptors intended to capture direct electrostatic interactions.

Each of the three sets is related to the original QSAR descriptors used by Hansch *et al*. (1964) and Leo *et al*. (1969). Labute's paper demonstrates through a series of experiments that the new descriptors work well in QSAR models developed for sets of small organic molecules. The 32 VSA descriptors are also shown to be weakly correlated. This makes them ideal for use in neural network training. The decision was made to train some neural networks using the set of VSA descriptors and others using a similar number of non-VSA descriptors. The non-VSA descriptors were picked

from the weakly correlated descriptors in set 3. A visual inspection of the 51 descriptors in set 3 was performed by plotting the descriptor values in turn. Those descriptors whose graphs best resembled that of a normal distribution were retained. For Boolean descriptors such as the 2D descriptor 'reactive' (which indicates the presence or absence of reactive groups in the molecule), only those with a good distribution of true and false values were used. Any VSA descriptors in set 3 were ignored. Table 2.1 lists the final set of 16 2D descriptors and 10 'inductive' descriptors which comprised the non-VSA descriptor set.

Table 2.1. The set of non-VSA descriptors used in the study. The definitions of the 'inductive' descriptors are based on those given in the paper "'Inductive' descriptors: 10 successful years in QSAR" (Cherkasov and Jankovic, 2005b).

#	Descriptor	Type	Definition
1	logP(o/w)	2D	Log of the octanol/water partition coefficient (including implicit hydrogen atoms).
2	b_ar	2D	Number of aromatic bonds.
3	b_double	2D	Number of double bonds. (Aromatic bonds are not considered to be double bonds.)
4	b_rotR	2D	Fraction of rotatable bonds: number of rotatable bonds divided by the number of bonds between heavy atoms.
5	chiral_u	2D	The number of unconstrained chiral centers.
6	density	2D	Molecular mass density: molecular weight divided by *van der Waals* volume.
7	petitjeanSC	2D	Petitjean graph shape coefficient as defined in Petitjean (1992).
8	RPC+	2D	Relative positive partial charge: the largest positive q_i divided by the sum of the positive q_i (where q_i denotes the partial charge of atom i.)
9	RPC-	2D	Relative negative partial charge: the smallest negative q_i divided by the sum of the negative q_i (where q_i denotes the partial charge of atom i.)
10	vsa_base	2D	Approximation to the sum of *van der Waals* surface areas of basic atoms.
11	vsa_don	2D	Approximation to the sum of *van der Waals* surface areas of pure hydrogen bond donors (not counting basic atoms and atoms that are both hydrogen bond donors and acceptors such as -OH).
12	b_triple	2D	Number of triple bonds (aromatic bonds are not considered to be triple bonds.)
13	reactive	2D	Indicator of the presence of reactive groups. A non-zero value indicates that the molecule contains a reactive group.
14	a_nCl	2D	Number of chlorine atoms.
15	a_nI	2D	Number of iodine atoms.

16	a_nS	2D	Number of sulfur atoms.
17	Smallest_Neg_Hardness	'inductive'	Smallest atomic hardness among values for negatively charged atoms.
18	Smallest_Pos_Hardness	'inductive'	Smallest atomic hardness among values for positively charged atoms.
19	Smallest_Rs_i_mol	'inductive'	Smallest value of atomic steric influence in a molecule.
20	Smallest_Rs_mol_i	'inductive'	Steric influence on the most positively charged atom in a molecule.
21	Softness_of_Most_Neg	'inductive'	Atomic softness of an atom with the most negative charge.
22	Softness_of_Most_Pos	'inductive'	Atomic softness of an atom with the most positive charge.
23	Total_Charge_Formal	'inductive'	Sum of charges on all atoms of a molecule (formal charge of a molecule).
24	Total_Pos_Softness	'inductive'	Sum of softnesses of atoms with positive partial charges.
25	Total_Sigma_mol_i	'inductive'	Sum of inductive parameters for all atoms with a molecule.
26	Average_EO_Neg	'inductive'	Arithmetic mean of electronegativities of atoms with negative partial charge.

2.6.2 Training

The neural network software used in this study consisted of a standard feed forward network system which could be trained using the Scaled Conjugate Gradient (SCG) algorithm rather than the more standard back-propagation (the SCG algorithm requires orders of magnitude fewer iterations for a given network compared to back-propagation). The ANN configuration used consisted of one input node for each of the descriptors in the descriptor set, a single hidden layer, and a single output node. The two data sets of metabolite patterns were each randomly separated into non-overlapping training and testing sets of 70% and 30% respectively, with roughly the same proportion of positive and negative patterns in each set. Separation of the input patterns into training and testing groups was done to avoid over-fitting to the neural networks. The activity values of the metabolite patterns were set to 1.0 for plant metabolites, which represented a positive training example, and 0.0 for human metabolites, which represented a negative training example.

To determine the best performing ANN model structure, a series of neural networks were trained with a range of different numbers of hidden neurons in the hidden layer. The model structure tests showed that models with 9 hidden neurons produced the best results and hence all further ANNs were allocated 9 neurons in the hidden layer. A total of 20 independent training and testing runs were conducted for each ANN and the resulting training and testing statistics have been reported for the averaged network outputs. Due to the limited amount of metabolite data available for training the neural networks, the procedure of cross-validation was adopted in order to improve the performance of the models. More details on network training, validation and performance are given in the results section.

2.7 Fragmentation Analysis

To complement the ANN results, a fragmentation analysis was carried out so that the commonalities and differences between the three groups of metabolites could be understood in conventional terms of chemical structure. A custom SVL script was run in the MOE which calculated the most common structural scaffolds and substituents for each of the two data sets. The rules used to fragment the molecular structures are described in the paper "Drug-like index: a new approach to measure drug-like compounds and their diversity" (Xu and Stevenson, 2000).

In the paper the following seven rules are used to derive scaffolds and substituents:

1. A cap (substituent, or side chain) is an acyclic substructure with one attachment connecting to other structure building blocks.
2. A linker is an acyclic substructure with more than one attachment connecting to other structure building blocks.
3. A core is a cyclic substructure without linker or cap.
4. A drug structure consists of a scaffold with or without substituents (or caps).
5. A scaffold has at least one core. If it has more than one core, linkers should connect the cores.
6. A non-ring bond containing at least one ring-atom is called as “joint bond”. By breaking joint bond(s), a structure will fall into a set of building blocks, i.e. cores, linkers, and caps.
7. An unsaturated bond is treated as a pseudo-ring (a ring with only two vertexes) if it is attached to a cyclic system.

3 Results

3.1 Overlap between groups

The diagram in Figure 3.1 shows the extent of the overlap between the three groups of metabolites based on their 2D similarity, which was measured using the Tanimoto coefficient applied to hashed fingerprint representations of the molecules (two metabolites are said to overlap if they have a Tanimoto coefficient of 1.0.) What is most surprising about the results is the relatively small overlap of 500 molecules that exists between the two human metabolite groups. Since they are taken from the same organism, a much larger overlap was expected. There are many reasons that could

explain this lack of similarity, and so a more detailed analysis of the overlap was undertaken.

Many of the metabolites stored in the HMDB are taken originally from the KEGG database. The HMDB makes it easy to identify which metabolites come from KEGG by keeping a record of their KEGG compound IDs. For the 1,491 'unique' HMDB metabolites that were used in this study, a search showed that 993 had copies in the KEGG database. Of these, 382 corresponded to metabolites in the 'unique' set of 744 KEGG metabolites. A further searched showed that a total of 379 metabolites out of the 500 metabolite overlap between KEGG and HMDB consisted of molecules that were connected by accession numbers in both databases. Therefore, although the fingerprint method for measuring similarity may not be 100% accurate, there is still a considerable difference between the human metabolites taken from KEGG and those taken from the HMDB.

The diagram in Figure 3.1 also shows an area of mutual overlap between all three metabolite groups. It was found that the metabolites that existed within this overlap belonged to pathways common in both plants and humans which were synonymous with the basic metabolism. These pathways included those responsible for the synthesis of essential nucleotides purine and pyrimidine, as well as the degradation of amino acids such as lysine, isoleucine and tryptophan.

By contrast, it was found that metabolites in the non-overlapping set of Arabidopsis metabolites belonged to pathways specific to plants, such as those responsible for the synthesis of secondary plant metabolites like the flavanoids kaempferol and quercetin.

Plant-specific primary metabolites, such as brassin and gibberellin growth hormones and metabolites vital for photosynthesis such as carotenoids (a type of pigment in plants which absorb blue light), were also found.

The 248 human metabolites common to both the HMDB and KEGG but separate from the set of plant metabolites were shown to belong to such pathways as the pathway involved in the formation of C21-steroid hormones—steroid compounds containing 21 carbons which function as hormones in humans. Other metabolites in this set were involved in the pathways for androgen and estrogen metabolism as well as the production of amino acids such as tryptophan and tyrosine. The ten most common pathways for the various sections of the diagram are given in Table 3.1.

In the KEGG database, the metabolic pathways for tyrosine and tryptophan are referred to as "super-pathways" because each consists of a combination of interacting individual pathways. These pathways include many reactions found in humans and other mammals, such as the production of catecholamines like dopamine from tyrosine and thyroid hormones like thyroxine, or the production of serotononin/melatonin from tryptophan. This accounts for the high abundance of metabolites present in these pathways.

3.2 ANN Performance

The performance of the ANN-trained QSAR models was initially assessed based on the true/false positive and true/false negative predictions produced during the training and testing phases. For each of the four models, the average number of true/false

positives and true/false negatives was taken over 20 independent training and testing runs. The outcome is illustrated in Table 3.2. Specificity, sensitivity, accuracy, and positive and negative predictive values (PPV and NPV) have been calculated for each model result. The Matthews Correlation Coefficient (MCC) has also been calculated.

The MCC is generally regarded as a well balanced measure of the quality of a classification system which can be used even if the numbers of positive and negative patterns in the data set differ by a large margin (Matthews, 1975). It is given by the following equation:

$$MCC = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}}$$

In this equation T_P is the number of true positives, T_N the number of true negatives, F_P the number of false positives and F_N the number of false negatives. The MCC returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 the worst possible prediction.

Because of the limited number of training patterns in the data, a 10-fold cross-validation was carried out to try and improve ANN performance. Utilising this technique allowed for 100% rather than 70% of the data to be used for training. The results of the cross-validation are also shown in Table 3.2. As would be expected, these results show an overall improvement in the performance of all four models.

The results of the 10-fold cross-validation also serve to highlight the difference in performance between the ANNs trained with the 32 VSA descriptors compared with those trained using the set of 2D/'inductive' descriptors. It can be seen that the use of VSA descriptors greatly improves performance for both the HMDB vs. AraCyc and KEGG vs. AraCyc models.

The most striking outcome of the results is that ANN models trained on the KEGG vs. AraCyc data set out-performed those trained on the HMDB vs. AraCyc data set even though the number of training examples available in the data set was significantly smaller. The specificity of the HMDB vs. AraCyc-trained models was consistently high for both descriptor types, but sensitivity remained low. This is perhaps due to the much larger number of human metabolites/negative training patterns compared to plant metabolite/positive training patterns. The ANN model trained using the KEGG vs. AraCyc data set using VSA descriptors provides the best performance out of all the models. The 10-fold cross-validation result for this model produces an even percentage specificity, accuracy, PPV and NPV, as well as a good Matthews coefficient of 0.7.

Once neural network training had been completed, a principal component analysis was carried out on the VSA descriptor values for the KEGG vs. AraCyc dataset. The top three principal components were then plotted on a three-dimensional axis with points colour coded according to whether they were plant or human. This 3D principal component graph can be viewed in Figure 3.3(a). The graph shows no obvious plane of separation between the two types. This is in stark contrast to previous studies which have shown that human metabolites occupy a distinctive cluster in the 3D PCA space

relative to metabolites from other organisms (Karakoc *et al*., 2006 and 2007). There are, however, a number of significant plant outliers which can be seen more clearly in Figure 3.3(b). The majority of these outliers consist of long chain lipids that correspond to a pathway which is highly conserved among eukaryotes known as *dolichyl-diphosphooligosaccharide* or *mannosyl-chito-dolichol* biosynthesis. This pathway is of particular interest in humans because defects in the lycosyltransferases involved lead to congenital disorders of glycosylation (Aebi and Hennet, 2001). The 2D chemical structure representation of one of the outliers is given in Figure 3.4.

To illustrate the non-random nature of the developed QSAR ANN models, neural networks were trained using the two data sets when the corresponding Boolean activity values have been assigned to the set of plant and human metabolites in a random manner (but maintaining the same number of positive and negative examples). Results for the neural networks trained on this 'noise' data were averaged over 20 independent runs. The results are plotted in the ROC graph displayed in Figure 3.2. Also plotted are the ROC curves for the two best-performing ANN models for non-randomized data which were trained using VSA descriptors and 10-fold cross-validation. As the ROC graph illustrates, the ANNs trained on randomized data produce no meaningful results, while the ANNs trained on the real data produce good ROC curves. The relative performance of the two 'real' QSAR models is also demonstrated, with the model trained using the KEGG/AraCyc data set out-performing the model trained on the HMDB/AraCyc data set.

3.3 Fragmentation Analysis

The custom MOE script 'Fragmentation.svl' was used to separate the metabolite structures into their constituent "building blocks" according to the drug-like index approach (Xu and Stevenson, 2000). The five most abundant scaffolds for the two data sets; AraCyc vs. KEGG and AraCyc vs. HMDB, are shown in Tables 3.3 and 3.4 respectively (remember that the overlap between human and plant has been removed based on hashed fingerprints comparisons, so the group of plant metabolites in each data set is different). The overlap between the human and plant metabolites was also fragmented and the results for the top five elements included in the relevant table. The 'R*' annotations on the fragments denote the points where the scaffolds attach to other building blocks in the molecule.

It can be seen from the results in Table 3.3 that the five most common scaffolds are the same in both AraCyc and KEGG metabolites but with subtle differences in the order of the scaffolds and their percentage presence. The most dominant scaffolds are five and six-member sugar fragments. Also present amongst the five most common are 1,2,4-substituted aromatic rings, hydroxyl substructures and 1,6-disubstituted purines. The situation is similar in Table 3.4 where AraCyc and HMDB metabolites share the same top five scaffolds. Again there is a high abundance of five and six-member sugar fragments. Unlike in the previous data set, purine is not one of the five most common scaffolds for these molecules, but instead appears slightly further down the list. Para-substituted aromatic rings take the place of purine in the five most common scaffolds for AraCyc and HMDB. For both data sets, the overlapping metabolites contain very similar scaffolds to the ones found in the separate human and plant groups.

The most common substituents for the data sets are featured in Tables 3.5 and 3.6. It is the case with each data set that hydroxyl and methyl are the dominate functional groups for both the human and the plant metabolites, as well as being similarly abundant in the overlap between the two. The presence of carboxyl and phosphate groups can be found slightly further down the list. Not contained within the five most common substituents but represented in the top ten were phenyl groups.

The average number of scaffolds and substituents per chemical structure has been determined for the various groups by dividing the sum frequency of either scaffold or substituent fragments by the number of molecules in the group being measured. The values show that plants metabolites have a slightly higher average number of scaffolds than humans but that most of the metabolites in both groups tend to contain a single scaffold. The average number of substituents is also slightly higher for plants at around 3 or 4 distinct functional groups per metabolite.

The fragment analysis shows that human and plant metabolites are composed of the same basic building blocks, with the only difference being minor variations in the overall presence of the various scaffolds and substituents which comprise them. These results can be related to the large overlap which was observed between plant and human metabolites using the hashed fingerprint technique as well as the not inconsiderable number of metabolites which were incorrectly classified by the trained neural networks.

3.4 Tables and Figures

Figure 3.1. Venn diagram showing the overlap between the three groups of metabolites; 1,491 HMDB human metabolites, 744 KEGG human metabolites, and 758 AraCyc plant metabolites.

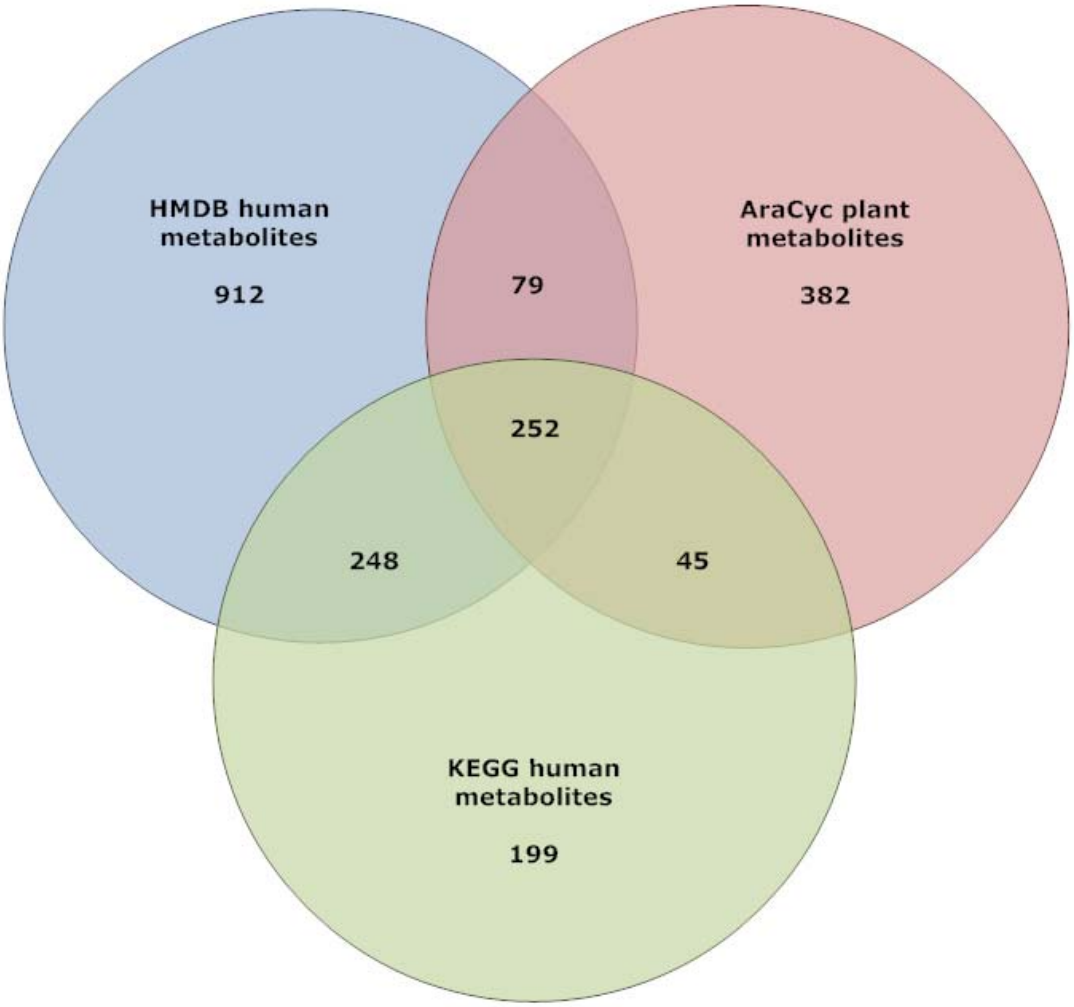


Figure 3.2. ROC parameters computed for the ANN models: KEGG human metabolites vs. AraCyc plant metabolites and HMDB human metabolites vs. AraCyc plant metabolites (using VSA descriptors and 10-fold cross-validation for both). Also shown are the ROC results for when activity values for both data sets are randomized.

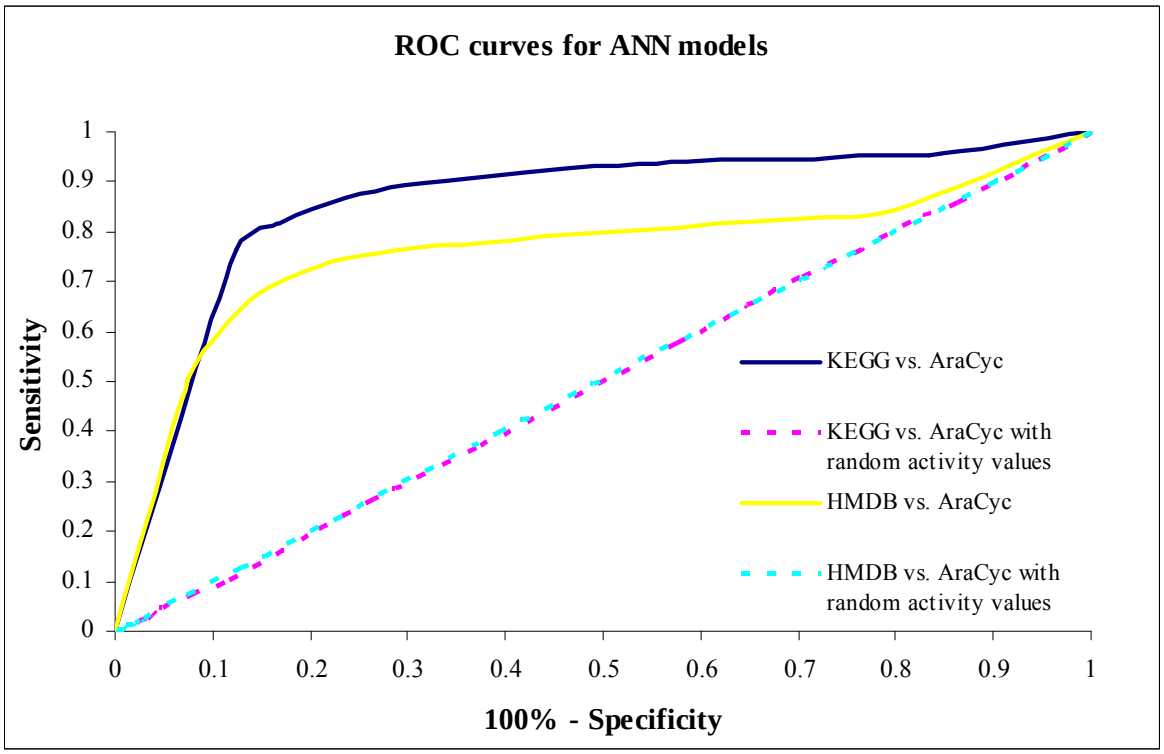
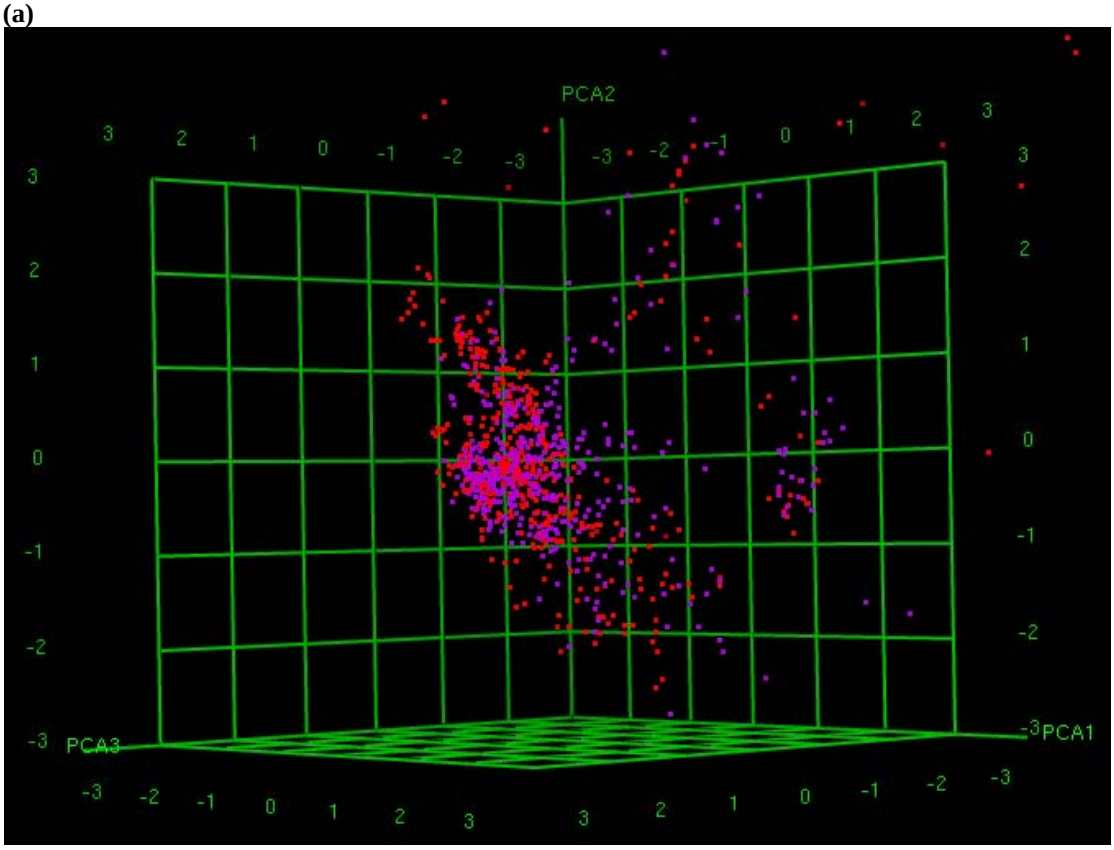


Figure 3.3. A plot of the 908 metabolites which formed the KEGG vs. AraCyc data set. The metabolites are plotted according to their values for the three principal components derived from 32 VSA descriptors. The red points correspond to Arabidopsis plant metabolites and the purple points correspond to KEGG human metabolites. Two images of the same graph are shown from different angles. Image **(b)** highlights the presence of a number of significant plant outliers.

(a)



(b)

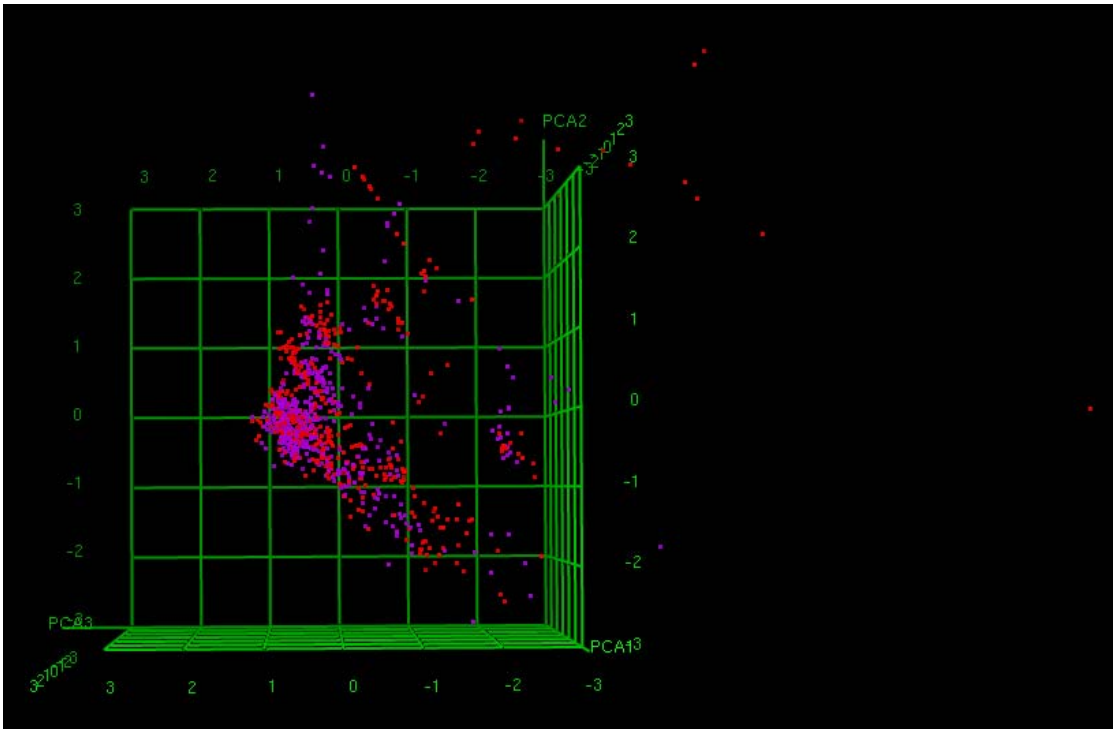


Figure 3.4. *(Mannosyl)7-(N-acetylglucosaminyl)2-diphosphodolichol* - an outlying Arabidopsis metabolite from the principal component graph in Figure 3.3. The metabolite belongs to the pathway responsible for the biosynthesis of *dolichyl-diphosphooligosaccharide* used to synthesize asparagine-linked glycans in eukaryotes.

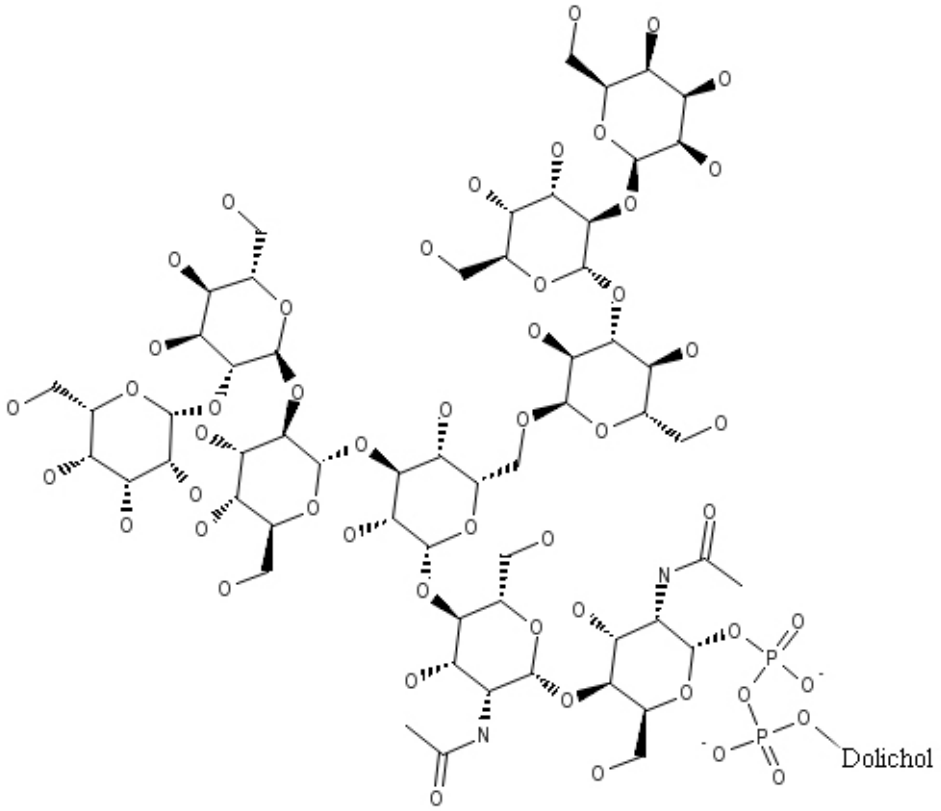


Table 3.1. A table displaying the ten most common pathways represented by the metabolites in the various overlapping/non-overlapping sections taken from the Venn diagram in Figure 3.1. The number in brackets at the end of each pathway name corresponds to the number of metabolites belonging to the pathway.

248 metabolite overlap between KEGG and HMDB not overlapping with AraCyc.	**252 common metabolite overlap between all three groups.**	**382 non-overlapping metabolites in AraCyc.**	**199 non-overlapping metabolites in KEGG.**
Tyrosine metabolism (27)	*Pyrimidine metabolism (KEGG) (33)*	*Gibberellin inactivation (13)*	*Tryptophan metabolism (18)*
C21-Steroid hormone metabolism (20)	*Purine metabolism (KEGG) (32)*	*Brassinosteroid biosynthesis I (11)*	*Purine metabolism (16)*
Androgen and estrogen metabolism (20)	*Glycine, serine and threonine metabolism (KEGG) (18)*	*Side chain elongation cycle aliphatic glucosinolates (Arabidopsis) (10)*	*Tyrosine metabolism (11)*
Neuroactive ligand-receptor interaction (19)	*Tryptophan metabolism (KEGG) (18)*	*Cytokinins 7-N-glucoside biosynthesis (10)*	*Folate biosynthesis (10)*
Arachidonic acid	*Glutamate*	*Glucosinolate*	*Androgen and estrogen*

metabolism (18)	*metabolism (KEGG) (18)*	*biosynthesis from homomethionine (9)*	*metabolism (9)*
Tryptophan metabolism (14)	*De novo biosynthesis of purine nucleotides (AraCyc) (11)*	*Phenylpropanoid biosynthesis (9)*	*Glycerophospholipid metabolism (9)*
Glycine, serine and threonine metabolism (11)	*(Deoxy)ribose phosphate degradation (AraCyc) (10)*	*Carotenoid biosynthesis (9)*	*C21-Steroid hormone metabolism (9)*
Glycerophospholipid metabolism (11)	*De novo biosynthesis of pyrimidine deoxyribonucleotides (AraCyc) (9)*	*13-LOX and 13-HPL pathway (8)*	*Fatty acid metabolism (8)*
Histidine metabolism (9)	*Tryptophan degradation (AraCyc) (8)*	*Monoterpene biosynthesis (7)*	*Bile acid biosynthesis (8)*
Arginine and proline metabolism (9)	*Lysine degradation II (AraCyc) (7)*	*Histidine biosynthesis (7)*	*Butanoate metabolism (6)*

Table 3.2. Statistics for the trained neural networks.

descriptor set	validation	true posit	true negat	false posit	false negat	spec	sens	accur	PPV	NPV	Matthews
					KEGG human metabolites vs. AraCyc plant metabolites						
Non-VSA	Training 70%	297	339	43	100	88.7%	74.8%	81.6%	87.3%	77.2%	0.64
	Testing 30%	118	136	28	53	83.2%	69.0%	76.0%	81.1%	72.0%	0.53
	10-fold	366	378	70	95	84.4%	79.4%	81.8%	83.9%	79.9%	0.64
VSA	Training 70%	344	334	48	53	87.5%	86.6%	87.0%	87.8%	86.3%	0.74
	Testing 30%	129	123	41	42	74.9%	75.3%	75.1%	75.8%	74.4%	0.50
	10-fold	392	381	66	69	85.2%	85.0%	85.1%	85.6%	84.7%	0.70
					HMDB human metabolites vs. AraCyc plant metabolites						
Non-VSA	Training 70%	211	1104	60	162	94.8%	56.6%	85.6%	77.8%	87.2%	0.58
	Testing 30%	79	459	40	82	92.1%	48.9%	81.5%	66.6%	84.8%	0.46
	10-fold	252	1070	90	175	92.2%	59.0%	83.3%	73.7%	85.9%	0.55
VSA	Training 70%	228	1114	50	145	95.7%	61.2%	87.3%	82.0%	88.5%	0.63
	Testing 30%	84	460	39	78	92.3%	51.9%	82.4%	68.4%	85.6%	0.49
	10-fold	280	1078	82	147	92.9%	65.6%	85.6%	77.3%	88.0%	0.62

Table 3.3. Most common distinct scaffolds for the data set of KEGG human metabolites vs. AraCyc plant metabolites.

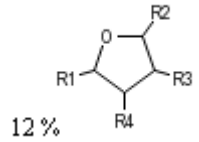
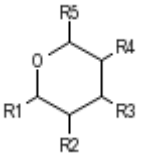
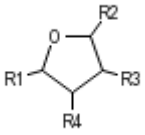
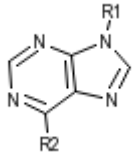
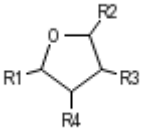
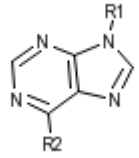
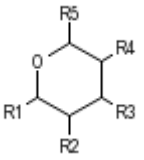
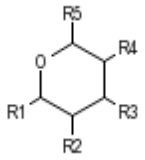
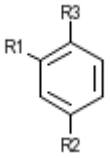
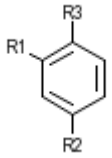
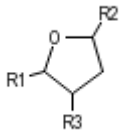
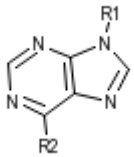
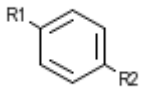
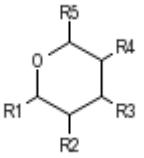
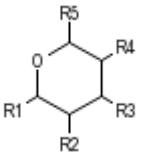
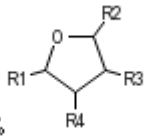
	KEGG	AraCyc	Overlap between KEGG and AraCyc
Average Number of Molecular Features	0.99	1.30	0.94
1	 12 %	 19%	 27%
2	 7%	 8%	 12%
3	 7%	 7%	 8%
4	 6%	 5%	 7%
5	 3%	 5%	 4%

Table 3.4. Most common distinct scaffolds for the data set of HMDB human metabolites vs. AraCyc plant metabolites.

	HMDB	AraCyc	Overlap between HMDB and AraCyc
Average Number of Molecular Features	1.07	1.41	0.84
1	 13%	 18%	 25%

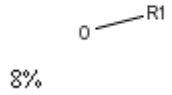
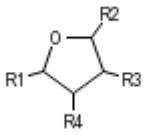
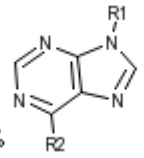
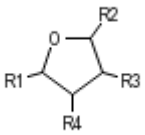
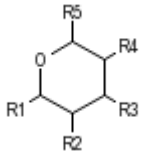
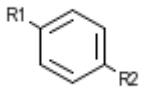
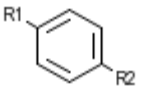
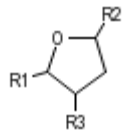
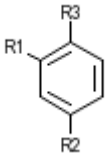
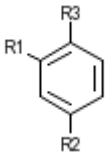
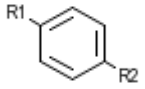
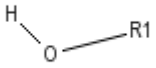
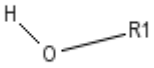
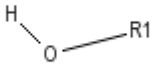
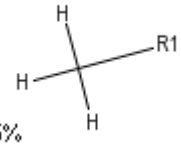
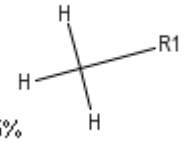
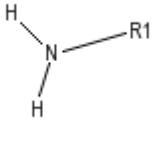
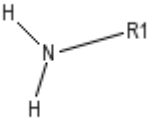
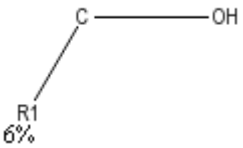
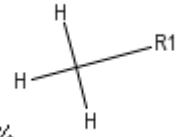
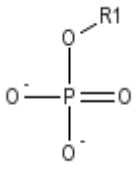
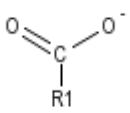
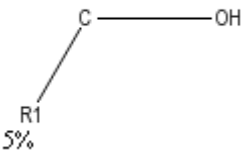
2	 8%	 8%	 14%
3	 7%	 7%	 8%
4	 6%	 4%	 6%
5	 5%	 4%	 4%

Table 3.5. Most common distinct substituents for the data set of KEGG human metabolites vs. AraCyc plant metabolites.

	KEGG	AraCyc	Overlap between KEGG and AraCyc
Average Number of Molecular Features	2.56	3.41	2.35
1	 35%	 39%	 38%
2	 16%	 16%	 10%
3	 6%	 6%	 10%
4	 4%	 6%	 5%

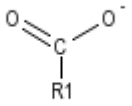
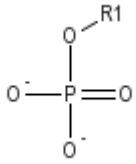
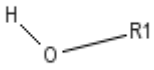
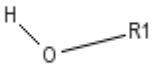
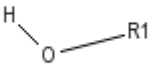
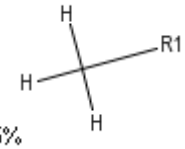
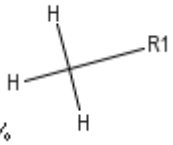
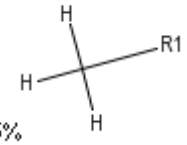
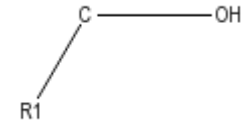
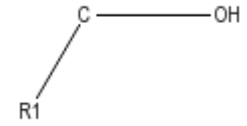
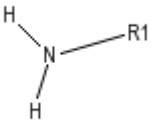
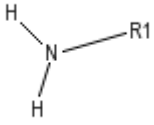
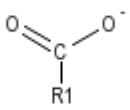
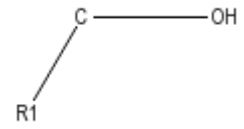
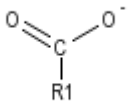
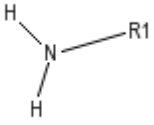
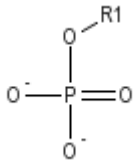
5	 3%	 3%	 5%

Table 3.6. Most common distinct substituents for the data set of HMDB human metabolites vs. AraCyc plant metabolites.

	HMDB	AraCyc	Overlap between HMDB and AraCyc
Average Number of Molecular Features	2.76	3.84	2.31
1	 33%	 40%	 36%
2	 16%	 14%	 15%
3	 5%	 7%	 7%
4	 4%	 6%	 5%
5	 3%	 3%	 4%

4 Discussion

This study has identified a discrepancy between the human metabolites compiled by the HMDB and those that exist in the KEGG compound database which partly serves to demonstrate the need for reliable data in molecular classification studies. This is an inherent problem with regards to the metabolome, as often what constitutes the metabolome depends on how a metabolite is defined. Wishart *et al*. (2007) estimate that if every small molecule in the human body were to be included in the human metabolome (be it food additive, plant extract, drug, drug derivative, toxin, cleaning agent or environmental contaminant)—from any source at any concentration level—the number of compounds might exceed 100,000. The criteria for a metabolite as defined by the Wishart group at the HMDB are given in chapter 2.2.1. Presumably, KEGG human metabolites picked for inclusion in the HMDB satisfied these entry requirements and others did not, which suggests a different set of rules are used to determine which metabolites are included in KEGG. Wittig and De Beuckelaer acknowledge the existence of similar inconsistencies between different metabolic pathway databases in their review paper (Wittig and De Beuckelaer, 2001).

Another point to consider is the unseasoned nature of metabolomics research compared to its 'omics' counterparts working with the genome, transcriptome and proteome. For example, 1,006 Arabidopsis metabolites were compiled for use in this study, which is far fewer than the 10^5 distinct small molecules that are estimated to occur in a single plant organism (Fiehn, 2001). The version of the AraCyc pathway database used only contained 285 metabolic pathways, many of which related to the basic metabolism, as evident from the comparison results in chapter 3.1. In fact it is only in recent years that plant-specific pathways such as carotenoid, brassinosteroid,

and giberellin biosyntheses have been added to the database (Garcia-Hernandez, 2002). A difficulty with analysing plant systems is that the secondary metabolism is heavily split among several cellular compartments and a combination of techniques are required to fully annotate the pathways (Fiehn, 2001). This and other complexities mean the full suite of metabolites synthesized by Arabidopsis is a long way from being discovered.

The small data sets used in this study meant that the effectiveness of the QSAR neural network models was in some sense restricted. Although performance was optimized using the 10-fold cross validation technique, a larger data set of training patterns would have been more desirable. Interestingly, although the HMDB vs. AraCyc data set contained more metabolite patterns, it didn't perform as well as KEGG vs. AraCyc (see Table 3.2). This is perhaps due to the larger proportion of human metabolites to plant metabolites biasing the results. Bearing this in mind, the results do not necessarily mean that the physicochemical properties of the KEGG metabolites differ more greatly from those of the Arabidopsis metabolites than the HMDB metabolites do. The neural network results also demonstrate that Labute's 32 VSA descriptors (Labute, 2000) work particularly well as metabolite classifiers and could be put to good use in similar future studies.

It is possible that the use of a Support Vector Machines (SVM) as an alternative form of supervised learning may have improved the results of the QSAR analysis. SVMs learn by creating a *hypersurface* within the space of all molecules in the training data. The hypersurface attempts to split the positive output molecules from the negative ones whilst also trying to keep a large margin between itself and the nearest molecules

(Bruce *et al*., 2007). SVMs differ from neural networks in that they are very efficient even with a large number of molecular descriptors.

The drawback of SVMs, neural networks and other supervised learning techniques is that they are primarily predictive tools that exist as black boxes, making it hard to determine the reason behind their behaviour. In QSAR modelling, this makes it difficult to ascertain which molecular properties are important. However, the neural network models trained in this study could have possible future application in predicting the otherwise unknown identity of small molecules. Alternatively they could be used to test new potential drug structures by predicting any interference with the normal metabolic pathways of humans.

Contrary to the two studies by Karakoc (Karakoc *et al*., 2006 and 2007), the human metabolites in this study were shown not to occupy an independent cluster in the QSAR descriptor space relative to the other classes of small molecules. The bacterial, fungal and plant metabolites used by Karakoc in his more recent paper displayed mutual overlap which was not apparent with the human metabolites. However, in my study a clear overlap between the metabolites of plants and humans has been observed. This draws into question the reliability of the methods and data sets used by all three studies and invites further investigation in order to shed some light on the contradictory nature of the results.

By contrast, the results of the fragmentation analysis in my study agree with Karakoc *et al*. 2006, where human metabolites are also shown to be composed largely of five and six-member sugar fragments, 1,6-disubstituted purines and exhibit a high

abundance of hydroxyl, methyl and carboxyl functional groups. This is perhaps not altogether surprising given the natural origin of the compounds, but nonetheless agrees with the words of Last *et al*. (2007) and others. However, contrary to the paper by Nobeli *et al*. (2003), where phosphate was found in over one-third of all known E. coli metabolites, the phosphate fragment in my fragmentation analysis exhibited a relatively low abundance and only appeared in an average of 5% of all plants and human metabolites.

This study also serves to highlight the difficulties inherent in comparing small molecules. In this study Tanimoto scores based on 2D hashed fingerprint descriptions of the molecules were used to identify the overlap between the three groups of metabolites. While this method is fast and easy to implement, it can lead to inaccuracies and will often see two molecules as identical even when they are not. This would have led to metabolites being removed from the data set unnecessarily which would in turn have had a detrimental effect on the performance of the neural networks, since there would be fewer metabolite patterns available for training.

One avenue for further work is to treat the molecules as graphs and apply graph-matching algorithms. The graph-matching approach does not suffer from the same inaccuracy as the hashed fingerprint approach. However, the complexity of the algorithms involved often means that it is not always possible to obtain an answer in a reasonable amount of time—even for small molecules like metabolites (as was the case in Nobeli *et al*., 2003). Another drawback is that the maximum common subgraph between two molecules is not necessarily meaningful in biological terms. However, it can still be used to compute a Tanimoto-like coefficient to measure

similarity. As graph-matching tasks are currently not dealt with by the CDK, any graph similarity comparisons would have to be performed with specially designed software. Detailed instructions for implementing a generic graph-matching algorithm suitable for comparing 2D chemical structures are given in Xu (1996).

To summarise, the results of this study have been consistent with existing knowledge about the basic metabolism of organisms, and have shown that Arabidopsis thaliana and human tend to share many of the molecules central to metabolism. This is reflected in the large overlap revealed by the 2D fingerprint comparisons and the tight cluster of metabolites in the QSAR descriptor space. However, a brief look at the metabolic pathways of the two organisms hints at significant differences in the way metabolites are linked together to form pathways, which could form the basis for further work. It is worth considering that many of the secondary Arabidopsis metabolites may not yet be available in the Arabidopsis Information Resource and so the data set used in this study does not yet accurately represent the majority of the Arabidopsis metabolome. This highlights a need for more publicly available compound information and centralised resources. Despite these data limitations, classification studies of the metabolome are an interesting complement to those which focus on the genome and proteome. Future investigations could benefit by integrating the knowledge obtained in studies such as the one I have conducted here with existing protein classification research.

5 References

Aebi, M. and Hennet, T. (2001). Congenital disorders of glycosylation: genetic model systems lead the way. Trends Cell Biol. 11(3), 136–141.

Bishop, C. M. (1996). Neural Networks for Pattern Recognition. Oxford University Press.

Brindle, J. T., *et al*. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. Nat. Med. 8, 1439–1444.

Brooksbank, C., Cameron, G., Thornton, J. (2005). The European Bioinformatics Institute's data resources: towards systems biology. Nucleic Acids Res. 33, D46–D53.

Bruce, C. L., Melville, J. L., Pickett, S. D., Hirst, J. D. (2007). Contemporary QSAR classifiers compared. J. Chem. Inf. Model. 47, 219-227.

Carhart, R. E., Smith, D. H., Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. J. Chem. Inf. Comput. Sci. 25, 64-73.

Cherkasov, A and Jankovic, B. (2004). Application of 'inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides. Molecules. 9, 1034-1052.

Cherkasov, A. (2005a). Can 'bacterial-metabolite-likeness' model improve odds of 'in silico' antibiotic discovery? J. Chem. Inf. Model. 46, 1214-1222.

Cherkasov, A. (2005b). 'Inductive' descriptors: 10 successful years in QSAR. Curr. Comput.-Aided Drug Des. 1, 21-42.

Clark, A. M., Labute, P., Santavy, M. (2005). 2D structure depiction. J. Chem. Inf. Model. 46, 1107-1123.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp. Funct. Genom. 2, 155–168.

Garcia-Hernandez, M., *et al*. (2002). TAIR: a resource for integrated Arabidopsis data. Funct. Integr. Genomics. 2, 239-253

Gasteiger, J. (Editor) and Engel, T. (Editor). (2003). Chemoinformatics: A Textbook. Wiley-VCH.

Hagadone, T. R. (1992). Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. J. Chem. Inf. Comput. Sci. 32, 515-521.

Halgren, T. A. (1996). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J. Comput. Chem. 17, 490-519.

Hansch, C. and Fujita, T. (1964). ρ-σ-π analysis – a method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. 86, 1616-1626.

Hansch, C., Leo, A., Taft, W. (1991). A survey of hammett substituent constants and resonance and field parameters. Chem. Rev. 91, 165-195.

Hattori, M., Okuno, Y., Goto, S., Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc. 125, 11853-11865.

Kanehisa, M., Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Karakoc, E., Sahinalp, SC., Cherkasov, A. (2006). Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. J. Chem. Inf. Model. 46, 2167-2182.

Karakoc, E., Sahinalp S. C., Cherkasov, A. (2007). Comparative QSAR analysis of bacterial-, fungal-, plant- and human metabolites. Pac Symp Biocomput. 12, 133-144.

Kell, DB. (2006). Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discovery Today. Vol 11, Numbers 23/24.

Labute, P. (2000). A widely applicable set of descriptors. J. Mol. Graphics Mod. 18, 464-477.

Last, R. L., Jones, D., Shachar-Hill, Y. (2007). Towards the plant metabolome and beyond. Nat. Rev. Mol. Cell Biol. 8(2), 167-74.

Leach, A. R. and Gillet, V. J. (2005). An Introduction to Chemoinformatics. Kluwer Academic Publishers.

Leo, A., Hansch, C., Church, C. (1969). Comparison of parameters currently used in the study of structure-activity relationships. J. Med. Chem. 12, 766-771.

Lipinski, CA., Lombardo, F., Dominy, BW., Feeney, PJ. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. Advan. Drug Deliv. Rev. 23, 3-25.

Lowell, H., Hall, L. H., Kier, L. B. (2001). Issues in representation of molecular structure: the development of molecular connectivity. J. Mol. Graphics Mod. 20, 4-18.

Macchiarulo, A. and Nobeli, I. (2007). Work in preparation.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. *405*, 442-451.

Mendes, P. (2002). Emerging bioinformatics for the metabolome. Brief Bioinform. 3, 134-145.

Mitchell, E. M., Artymiuk, P. J., Rice, D. W., Willett, P. (1990). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J. Mol. Biol. 212, 151-166.

Møller, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks. 6 (4), 525-533.

Monev, V. (2005). Introduction to similarity searching in chemistry. MATCH Commun. Math. Comput. Chem. 51, 7-38.

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. J. Chem. Doc. 5, 107-113.

Nobeli, I., Ponstingl, H., Krissinel, EB., Thornton, JM. (2003). A structure-based anatomy of the E. coli metabolome. J. Mol. Biol. 334, 697-719.

Nobeli, I. and Thornton, J. M. (2006). A bioinformatician's view of the metabolome. Bioessays. 28, 534-545.

Oliver, S. G., Winson, M. K., Kell, D. B., Baganz, F. (1998). Systematic functional analysis of the yeast genome. Trends. Biotechnol. 16, 373-378.

Qian, N. and Sejnowski, J. S. (1987). Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865-884.

Rang, H. P., Dale M. M., Ritter J. M. (1999). Pharmacology: Fourth Edition. Churchill Livingstone.

Ray, L. C. and Kirsch, R. A. (1957). Finding chemical records by digital computers. Science. 126, 814-819.

Raymond, J.W. and Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. J. Comput.-Aided Mol. Des. 16, 521-533.

Swain, C. G. and Lupton, E. C. (1968). Field and resonance components of susbtituent effects. J. Am. Chem. Soc. 90, 4328-4337.

Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. J. ACM. 23, 31-42.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31-36.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., *et al*. (2006). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 34, D173–D180.

Wiener, H. (1947). Structural determination of paraffin boiling point. J. Am. Chem. Soc. 69, 17-20.

Wishart, D.S., *et al*. (2007). HMDB: the Human Metabolome Database. Nucleic Acids Res. 35, D521-6.

Wittig, U. and De Beuckelaer, A. (2001). Analysis and comparison of metabolic pathway databases. Brief Bioinform. 2, 126-142.

Xu, J. and Stevenson, J. (2000). Drug-like index: a new approach to measure drug-like compounds and their diversity. J. Chem. Inf. Comput. Sci. 40, 1177-1187.

Xu, J. (1996). GMA: A generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. J. Chem. Inf. Comput. Sci. 36, 25-34.

Further Information

The AraCyc database:
http://www.arabidopsis.org/biocyc/index.jsp
The KEGG database:
http://www.genome.ad.jp/kegg/
Merck Index Database:
http://www.asu.edu/lib/resources/db/merck.htm
Chemical Computing Group, Inc:
http://www.chemcomp.com
SVL exchange community:
http://svl.chemcomp.com/index.php
DrugBank:
http://redpoll.pharmacy.ualberta.ca/drugbank/
ChEBI:
http://www.ebi.ac.uk/chebi/
Human Metabolome Database:
http://www.hmdb.ca/
Analyticon-Discovery Company:
http://www.ac-discovery.com/
Daylight Chemical Information Systems, Inc. Theory Manual:
http://www.daylight.com/dayhtml/doc/theory/index.html
Artificial Neural Network software in Java:
http://homepage.mac.com/jhuwaldt/java/Packages/NeuralNets/NeuralNets.html