A Support Vector Machine Based Model for Predicting Heparin-Binding Proteins Using XB Patterns as Features

Joseph Sirrianni, Zhichun Xiao, Wing Ning Li
Computer Science and Computer Engineering, University of Arkansas, Fayetteville, Arkansas, U.S.A

Abstract - Heparin is a highly sulphated and negatively charged polysaccharides belonging to the glycosaminoglycans(GAGs) family, used in medical treatments as an anticoagulant. Although many heparin-binding proteins have been identified, there are still many proteins needing to be classified as heparin-binding or not. Many studies have been aimed at prediction of heparin binding patterns in the primary structure of proteins, however, still no model has emerged which reasonably predicts proteins as heparin-binding or not. The main objective of this study is to predict heparin-binding proteins from their amino acid sequence information. A supervised learning algorithm based on support vector machine (SVM) is applied to two data sets; one training set used to create the model and one testing set used to validate and test accuracy of the model. For the testing set, the model achieves 75.36% accuracy in predicting heparin-binding proteins. The current model uses 66 XB patterns as features.

Keywords: Heparin, Heparin-binding proteins, Heparin-binding Motifs, Support vector machine, Prediction model.

2 Background

2.1 Biological background

Proteins are comprised of amino acids. These amino acids are strung together in an amino acid chain, which makes up the primary structure of a protein. The primary structure contains a lot of relevant data pertaining to the features and characteristics of its protein. An amino acid chain is a specific consecutive sequence of amino acids found in a protein. There are only 20 natural distinct amino acids found in proteins. For example, the amino acid sequence Alanine-Cysteine-Alanine-Glycine would correspond to the following string ‘ACAG’, where Alanine maps to the letter ‘A’, Cysteine maps to the letter ‘C’, and Glycine maps to the letter ‘G’.

Available structural information on heparin-binding proteins (HBPs) reveals that heparin binds to a binding pocket consisting of positively charged amino acids (lysine/arginine/histidine) [6]. An XB pattern is a string of X and B, where B stands for the 3 basic amino acids (lysine/arginine/histidine) and X stands for the remaining 17 of the 20 natural amino acids. Based on information given by the structures of fibroblast growth factors (FGFs) (proteins that interact with heparin during their cell signaling process), it is known that the selective distribution of the basic amino acids is important for heparin affinity and interaction. Based on the 3-dimensional structures of other heparin-binding proteins, consensus or signature heparin-binding sequences (strings) have been found to occur in these proteins that are thought to be required for their interaction with heparin. The two main pattern strings are XBBXBX and XBBBXXBX [7].

There is a strong interest in finding new heparin-binding proteins and peptide sequences in order to develop new treatment methods for many diseases. To support the above task, we have applied the Support Vector Machine (SVM) [3-5] approach, a supervised machine learning method, to build a prediction model. The model takes in the primary structure (amino acid sequence information) of a protein and determines if the protein is heparin-binding or not.
2.2 Computer science background

In the SVM based Supervised Machine Learning, samples are considered as points in a higher dimensional space. In this case, the samples will be individual proteins. Each point has a label that indicates to which of the two groups it belongs (in this case heparin-binding protein group and non-heparin-binding protein group). Further, a set of training or learning samples (usually half of the samples belong to one group and another half belong to the other group) are fed into the SVM learning algorithm. The algorithm attempts to build a hyper plane that separates the learning samples into the two groups. Samples belonging to the same group can be expected to reside on one side of the hyper plane and consequently the hyper plane becomes the classifier or the prediction model. It should be noted that any hyper plane partitions a higher dimensional space into two parts, on either sides of the hyper plane. To predict a new sample, the sample is transformed into a point in the higher dimensional space. Then depending on which side of the hyper plane the point ends up on, a prediction is made.

In a higher dimensional space, each dimension is a feature of the underlining sample (in this case proteins are samples). In building a SVM based prediction model for heparin-binding proteins, we need to decide what features of proteins to use. In this study, we have decided to use occurrences of different XB patterns as features, since, as stated above, the XB pattern occurrences in heparin-binding proteins seem to suggest that their presence is important to the potential to bind.

3 Contribution

Using XB patterns as features, a SVM based prediction model for heparin-binding proteins is proposed and developed. The prediction is based on sequence information or protein primary structure. The models achieve reasonable prediction results and support the research effort of finding new heparin-binding proteins and peptide sequences in order to develop new treatment methods for many diseases. As of now, we are unaware of any other heparin-binding predictive models existing currently.

4 Approach

Figure 1. The process of creating the predictive model from the initial training proteins.

The total process of creating the model is displayed in figure 1. The initial amino acid chain in FASTA format is converted into feature values. Then the new feature values are scaled and used to create the predictive SVM model.

4.1 Data sample collection

The protein’s primary structure information used for training and testing were extracted from the UniProt database [13]. One hundred and seventy four heparin binding proteins were gathered from the database of polyanion-binding proteins (DB-PABP) [16], and another one hundred and four non-binding proteins were gathered from the UniProt database. Tables 1 and 2 list examples of heparin binding and non-heparin binding proteins used in the testing and training. A complete list of the proteins used can be found in the appendix.

<table>
<thead>
<tr>
<th>Sequence Identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>731717</td>
<td>Protein NSG1</td>
</tr>
<tr>
<td>307129</td>
<td>hepatic lipase precursor</td>
</tr>
<tr>
<td>33959</td>
<td>[Homo sapiens]</td>
</tr>
<tr>
<td>1585498</td>
<td>diamine oxidase</td>
</tr>
<tr>
<td>340146</td>
<td>urokinase</td>
</tr>
</tbody>
</table>

Table 2. Examples of non-heparin binding proteins used in testing and training

<table>
<thead>
<tr>
<th>Sequence Identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>O35400</td>
<td>alcohol sulfotransferase</td>
</tr>
<tr>
<td>Q15125</td>
<td>cholesterol DELTA-isomerase</td>
</tr>
<tr>
<td>Q08462</td>
<td>adenylate cyclase</td>
</tr>
<tr>
<td>P08842</td>
<td>steryl-sulfatase</td>
</tr>
</tbody>
</table>
The training set for the models consisted of 140 total proteins, which is about half of the total protein sample size. The remaining proteins were part of the testing set, which was used to determine accuracy. Of the 140 training proteins, 70 were known to be heparin binding and the other 70 were known to be not heparin binding. The training set was evenly distributed between binding and non-binding proteins in order to ensure the models would be balanced in their construction.

4.2 Feature selection

A total of 66 features were used to map each protein to a point in a 66 dimensional space. The feature values are calculated based on the number of occurrences of each feature’s corresponding XB pattern in the protein’s primary amino acid sequence. Therefore, our model only uses a protein’s primary sequence data to make the heparin binding prediction.

Each protein has a primary structure, which is its sequence of amino acids. Each amino acid has its own generalized pH level. Heparin is one of the mostly negatively charged glycosaminoglycans, so it’s hypothesized that a larger volume of basic amino acids are necessary in a protein for it to bind. A protein’s sequence can be reduced to the two symbols B and X, where B represents the 3 basic amino acids (arginine, histidine, and lysine) and X represents the 17 non-basic amino acids [2, 9, 10, 11]. Recent studies have shown that heparin-binding proteins contain common XB pattern strings [8]. Thus, the XB patterns present in a protein should be relevant to its potential to bind. The XB patterns selected for the model are patterns that have been identified as being present in other heparin binding proteins.

When deriving the XB features of a protein, it first has its primary structure converted from a string of amino acid identifiers to a string of just Xs and Bs. Then, each of the 66 XB pattern occurrences is individually summed and set to the corresponding feature number (Feature 1 is the first XB pattern, Feature 2 is the second, and so on). A list of the first 7 features and their corresponding XB patterns is listed in Table 3. A full list of the 66 XB patterns can be requested from the authors.

<table>
<thead>
<tr>
<th>Feature Number</th>
<th>XB pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>XBXXXBXX</td>
</tr>
<tr>
<td>2</td>
<td>XBXBXXBX</td>
</tr>
<tr>
<td>3</td>
<td>XBBXBBX</td>
</tr>
<tr>
<td>4</td>
<td>XBBBXXBX</td>
</tr>
<tr>
<td>5</td>
<td>BBXXBBX</td>
</tr>
</tbody>
</table>

The software used for feature derivation consisted of python scripts. The python scripts transformed the protein data from FASTA format into the desired feature values.

4.3 Scaling the features

Every feature is represented by a number. In this case, the XB features count occurrences of XB patterns, so their feature value must be a positive integer. However, some XB patterns are much smaller than others and occur more often; this can result in some features values being much greater than others and overshadowing them in importance.

To handle the various possible ranges of the feature values, all of the features were scaled to fall within a range of [-1, +1]. This scaling helps the support vector machine balance out the importance of each feature, to avoid features with large values dominating ones with smaller values [3].

4.4 The support vector machine

In addition to the selected features, several aspects of the support vector machine affect the accuracy of the model. These aspects include the kernel function type, the function parameter, γ, and the soft margin parameter, C.

The kernel function translates the training data into higher dimensions so that the training points can become linearly separable. After testing several different kernel functions, the kernel function chosen for the SVM was the Gaussian radial basis function. This kernel takes in a specified parameter γ. The Gaussian function is as such:

\[ K(x_i, x_j) = \exp\left(-\gamma \| x_i - x_j \|^2 \right), \quad \gamma > 0. \tag{1} \]

The soft margin parameter, C, (also called the penalty parameter) determines how much variability in the computed hyperplane should be allowed. A larger C value will result in a tighter bound, while a smaller C value would result in a more relaxed bound [4].

In order to help determine optimal γ and C values, the grid tool from LIBSVM, a support vector machine library, was used [15]. The grid tool performs programmatic cross-validation checks with various combinations of γ and C values to determine which yields the best accuracy. These cross-validation checks are performed by separating the training data into several subsets and then running a test in which half of the subsets are considered the training set and the other half is considered the testing set. The cross validation model is created using the training subsets and tested against the testing
subsets in order to achieve accuracy [12]. The values of $(\gamma, C)$ which yielded the highest accuracy were used in the models.

### 4.5 Model validation

In order to validate the model a testing protein data set was identified consisting of the remaining 138 proteins not included in the training set. Of these, 105 of the proteins were known heparin binding proteins, while the other 34 were known heparin non-binding proteins.

A different set of LIBSVM functions was used to validate the model. The validation was performed by first transforming the testing protein’s sequence data into the features used by the model, scaling them, and then determining on which side of the dividing hyperplane each protein is located. The proteins were then assigned a label, either +1 or -1, to signify which set the protein was predicted to fall into, where the +1 set is the set of heparin binding proteins and the -1 set is the set of heparin non-binding proteins. Each newly predicted label is then compared to the protein’s known label. If the labels match, then the prediction is considered successful. If the labels do not match, the prediction is considered unsuccessful.

### 5 Results

For this research a model using 66 XB pattern features was created. The results of its accuracy are listed in Table 4 below.

<table>
<thead>
<tr>
<th>66 XB Pattern Features Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$ value</td>
<td>8</td>
</tr>
<tr>
<td>$C$ value</td>
<td>0.03125</td>
</tr>
<tr>
<td>Training Set Accuracy</td>
<td>99.28% (139/140)</td>
</tr>
<tr>
<td>Testing Set Accuracy</td>
<td>75.36% (104/138)</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>87.41% (243/278)</td>
</tr>
</tbody>
</table>

The model tended to do well (close to 100% accuracy) predicting the Training set. The Testing set was predicted with a high accuracy of 75.36%. The combined accuracy (combined totals of testing and training sets) across the entire dataset was 87.41%.

At this time we are unaware of any other heparin-binding prediction models, and thus have no precedent to compare these results to.

### 6 Conclusions and future work

A heparin-binding prediction model is proposed. It is a Support Vector Machine based model using protein primary structure as input. The model considers XB pattern frequency features at the present time. A preliminary prototype system is developed that allows a user to provide an amino acid sequence, pick a model, and then the system returns the prediction result to the user.

The models and the software described in this paper have demonstrated significant potential in advancing research efforts of identifying new heparin-binding proteins via the investigation and study of protein and peptide sequences as a means of developing new, more effective treatment methods for many diseases. By using XB patterns as features, this investigation provides additional insight into the role that XB patterns or motifs play in proteins that bind to heparin.

To improve the prediction accuracy, we will investigate the selection of other XB patterns and features beyond XB patterns such as chemical and physical properties of amino acids. Since protein interaction and protein binding take place in 3-dimensional space, including secondary and tertiary structural information would likely improve the accuracy of the prediction model and will be studied.

Additionally, other types of machine learning techniques may yield better results to this type of problem. However, given the limited dataset available for binding and non-binding proteins, finding other suitable techniques may be difficult.

### 7 Acknowledgements

This research was supported by an Honors College International Research Grant.

We would like to thank Dr. Kumar and Dr. Srinivas Jayanthi of the Chemistry and Biochemistry department of the University of Arkansas for providing the protein data and guidance with the biological aspect of this research.

### 8 References


