5-2016

A Support Vector Machine Base Model for Predicting Heparin-Binding Proteins Using Biological Metrics and XB Patterns as Features

Joseph W. Sirrianni
University of Arkansas, Fayetteville

A Support Vector Machine Base Model for Predicting Heparin-Binding Proteins Using Biological Metrics and XB Patterns as Features

Joseph W. Sirrianni

Bachelor of Science in Computer Science

University of Arkansas, Fayetteville

May 2016

INTRODUCTION

Heparin is a member of the glycosaminoglycan (GAG) family. Glycosaminoglycans are linear polysaccharides that participate in many biological processes by interacting with a wide range of proteins [1]. Heparin is a highly sulfated glycosaminoglycan found in most organisms and is widely used as an anticoagulant to treat thrombosis, thrombophlebitis, and embolism [2].

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.

Background

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. From a computer science standpoint, a protein's primary structure is often represented in the format of a string of capital letters, where each letter maps to one of the 20 natural amino acids. 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]. These two patterns and another 17 patterns are further investigated in their occurrences in heparin-binding proteins [8], as are other commonly occurring patterns.

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 belongs 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 different combinations of XB patterns and other biological measurements 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. Additionally, other biological indicators exist that could prove useful in the prediction. For example, heparin is very negatively charged; therefore it is reasonable to believe the proteins with a more positive charge may tend to bind more easily with Heparin. Thus, information regarding the charges of the protein are also taken as possible features, as well as other measurable information such as the amino acid chain's length, hydrophobicity index, et cetera.

Contribution

Using XB patterns and biological measurements 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.

Approach

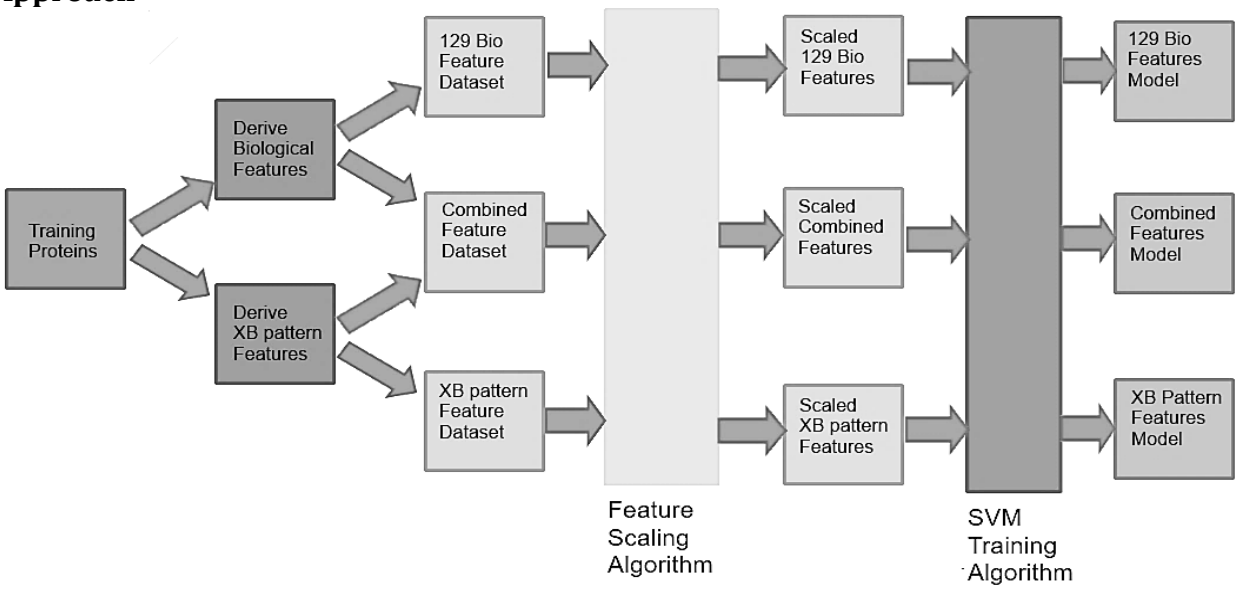

Figure 1 The process of creating the three predictive models from the initial training proteins.

Data sample collection

The proteins 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), 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 1 Examples of heparin binding proteins used in testing and training

Sequence Identifier	Description
731717	Protein NSG1
307129	hepatic lipase precursor
33959	interleukin 8 [Homo sapiens]
1585498	diamine oxidase
340146	urokinase

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

Sequence Identifier	Description
O35400	alcohol sulfotransferase
Q15125	cholestenol DELTA-isomerase
Q08462	adenylate cyclase
P08842	steryl-sulfatase
P25697	phosphoribulokinase

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.

Feature selection

For this research, two sets of features were used to create three predictive models. One model corresponds to the first feature set, another corresponds to the second feature set, and last model uses both feature sets combined. The two features sets were broken down by approach; one feature set used commonly occurring XB patterns and the other used likely biological metrics.

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

XB pattern feature set

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 [9, 10, 2, 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 and 19 XB patterns used can be found in the appendix.

Table 3 Example set of Features 1 – 7 and their XB patterns

Feature Number	XB pattern
1	XBXXBXBX
2	XBXBXXBX
3	XBBXBBX
4	XBBBXXBX
5	BBXXBBXX
6	XBBXXBBX
7	BXBXBBXB

Biological Metrics Feature set

The second feature set contained 129 features derived from the primary structures. These features were measured using the Python library Biopython[14]. Table 4 gives descriptions of the kinds of measurement used in the first 9 features. The remaining 120 features are 6 metrics applied for each of the 20 amino acids types present in the protein. A full list of the 129 features can be found in the appendix and their code used can be requested from the authors.

Table 4 Example descriptions for the first 9 biological features

Feature #	Desciption:
Feature #1	Charge of protein
Feature #2	Length of Protein Sequence
Feature #3	Molecular Weight
Feature #4	Aromaticity
Feature #5	Instability Index
Feature #6	Isoeletric Point
Feature #7	Helix Fraction
Feature #8	Turn Fraction
Feature #9	Sheet Fraction

Scaling the Features:

Every feature is represented by a number. For example, the XB features count occurrences of XB patterns, so their feature value must be a positive integer. However, the biological features may have vastly different values for their features. For example, the molecular weight feature will likely be a very large four or five digit number, while the Instability index feature will have a value less than 1. The large differences in the possible values can lead to higher value features overshadowing the smaller valued ones.

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].

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:

$$K(x_i, x_j) = \exp\left(-\gamma * \left|x_i - x_j\right|^2\right), \quad \gamma > 0.$$

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 (γ, C) which yielded the highest accuracy were used in the models.

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.

RESULTS

For this research 3 different models were created: one using the 129 biological metrics features, one using the 66 XB pattern features, and another using a combination of the 129 biological features and 19 of the most commonly occurring XB patterns in heparin-binding proteins. All of the models were trained using the same training set and all were tested against the same testing set. The results of their accuracies are listed in Table 5 below.

Table 5 Model Accuracies and parameter values.

	129 Bio Features Model	66 XB Pattern Features Model	Combined 129 Bio and 19 XB Feature model
γ value	1	8	2
C value	0.886123	0.03125	0.03125
Training Set Accuracy	100% (140/140)	99.28% (139/140)	100% (140/140)
Testing Set Accuracy	74.64% (103/138)	75.36% (104/138)	76.81% (106/138)
Combined Accuracy	87.41% (243/278)	87.41% (243/278)	88.49% (246/278)

All the models tended to do well (close to 100% accuracy) predicting the Training set. The Testing set was predicted with the highest accuracy in the combined model with an

accuracy of 76.81%, however the other two models were very close to it in accuracy as well. The combined accuracy across the entire dataset was around 88% across all the models.

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

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 models consider XB pattern frequency, biological indicators, or both as 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 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.

ACKNOWLEDGEMENTS

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

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

I would like to thank Dr. Zhichun Xiao for this preliminary testing results and scripts.

I would like to thank Dr. Wingning Li for being my mentor throughout this research. His guidance was invaluable.

REFERENCES

[1] Casu, B., Guerrini, M., and Torri. G. (2004) Structural and conformational aspects of the anticoagulant and anti-thrombotic activity of heparin and dermatan sulfate. Curr Pharm Des 10: 939-949.

[2] Gandhi N. S. and Mancera R. L. (2008) The Structure of Glycosaminoglycan and their interaction with Proteins, Chem Biol Drug Des 2008, 72:455-482

[3] C.-C. Chang and C.-J. Lin. (2011) LIBSVM A library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27.

[4] Shigeo Abe, Support Vector Machines for Pattern Classification, Second Edition, Springer, 2010

[5] C. Cortes and V. Vapnik. (1995) Support-vector network, Machine Learning, September 1995, Volume 20, Issue 3, pp273-297

[6]Arunkumar, A. I., Srisailam, S., Kumar, T. K., Kathir, K. M., Chi, Y. H., Wang, H. M., et al. (2002) Structure and stability of an acidic fibroblast growth factor from Notophthalmus viridescens. J Biol Chem, 277(48), 46424-46432

[7]Capila, I. and Linhardt, R. J. (2002) Angew. Chem. Int. Ed. Engl. 2002, 41, 391-412.

[8] Dempewolf, C., Morris J., Chopra M., Jayanthi S., Kumar T., and Li W. (2013) Identification of Consensus Glycosaminoglycan Binding Strings in Proteins Proc. International Conference on Information Science and Applications, 310-314.

[9] Arunkumar, A. I., Kumar, T. K., Kathir, K. M., Srisailam, S., Wang, H. M., Leena, P. S., et al. (2002) Oligomerization of acidic fibroblast growth factor is not a prerequisite for its cell proliferation activity. Protein Sci, 11(5), 1050-1061.

[10] Bae, J., Desai, U. R., Pervin, A., Caldwell, E. E., Weiler, J. M., and Linhardt, R. J. (1994) Interaction of heparin with synthetic antithrombin III peptide analogues. Biochem J, 301 (Pt 1), 121-129.

[11] Adjit Varki, Richard D Cummings, Jeffrey D Esko, Hudson H Freeze, Pamela Stanley, Carolyn R Bertozzi, Gerald W Hart, and Marilynn E Etzler, Essentials of Glycobiology, 2nd Edition, Cold Spring Harbor, 2009

[12] Hsu, C., Chang, C., and Lin, C. "A Practical Guide to Support Vector Classification" Internet: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, April 15th, 2010 [December 7, 2015].

[13] http://www.uniprot.org

[14] http://biopython.org/

[15] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Appendix

List of proteins used in this paper:

Training Set Proteins:

Heparin-binding Proteins:

>148747171 | adrenergic receptor, alpha 1a [Rattus norvegicus].
>730640 | 40S ribosomal protein S19.
>9082303 | amphoterin [Rattus norvegicus].
>2506872 | Fibronectin precursor (FN) (Cold-insoluble globulin) (CIG).
>37174 | thyroglobulin [Homo sapiens].
>6707677 | Probable eukaryotic translation initiation factor 3 30 kDa subunit (eIF-3 30 kDa) (eIF3j).
>56961643 | lipase, hepatic [Rattus norvegicus].
>112996 | Protein ABC1, mitochondrial precursor.
>220434 | heparin binding protein 44 [Mus musculus].
>74719685 | HSPC314.
>127115 | Midkine precursor (Retinoic acid-induced heparin-binding protein) (RI-HB).
>132876 | 60S ribosomal protein L30.
>2507249 | Cationic trypsin precursor (Beta-trypsin) [Contains: Alpha-trypsin chain 1; Alpha-trypsin chain 2].
>549657 | Uncharacterized TLC domain-containing protein YKL008C.
>758073 | complement factor H [Homo sapiens].
>549686 | Manganese resistance protein MNR2.
>139472860 | KCP [Human herpesvirus 8 type P].
>547759 | Tyrosine-protein kinase BTK (Bruton tyrosine kinase) (Agammaglobulinaemia tyrosine kinase) (ATK) (B cell progenitor kinase) (BPK).
>34420140 | probasin [Rattus norvegicus].
>5852072 | L-selectin [Homo sapiens].
>52782792 | Probable ATP-dependent RNA helicase DDX47 (DEAD box protein 47).
>4506629 | ribosomal protein L29 [Homo sapiens].
>74752707 | CDNA: FLJ21054 fis, clone CAS00538 (Syncoilin, intermediate filament 1).
>1730882 | Probable transcriptional regulatory protein YPL133C.
>22653668 | Bromodomain adjacent to zinc finger domain protein 2B (hWALp4).
>184820 | insulin-like growth factor binding protein 5 [Homo sapiens].
>586485 | Uncharacterized acyltransferase YBR042C.
>74762678 | Tigger transposable element-derived protein 1.
>549725 | NADH-cytochrome b5 reductase precursor (p34/p32) [Contains: NADH-cytochrome b5 reductase p34 form; NADH-cytochrome b5 reductase p32 form].
>189578 | plasminogen activator inhibitor 1.
>50401123 | Probable ATP-dependent helicase LGP2 (Protein D11Lgp2 homolog).
>401413 | von Willebrand factor precursor (vWF) [Contains: von Willebrand antigen 2 (von Willebrand antigen II)].
>204572 | heparin-binding growth associated molecule.
>465514 | Putative deoxyribonuclease YBL055C.
>554477 | neural cell adhesion molecule.
>731597 | OTU domain-containing protein 2.
>89276751 | alpha 1 type V collagen preproprotein [Homo sapiens].
>27923749 | Centaurin-alpha 2.
>4033508 | Annexin A5 (Annexin-5) (Annexin V) (Lipocortin V) (Endonexin II) (Calphobindin I) (CBP-I) (Placental anticoagulant protein I) (PAP-I) (PP4) (Thromboplastin inhibitor) (Vascular anticoagulant-alpha) (VAC-alpha) (Anchorin CII).
>52000648 | Ubiquitin D (Ubiquitin-like protein FAT10) (Diubiquitin).
>48146225 | LPL [Homo sapiens].

>1169344 | Dimethyladenosine transferase (S-adenosylmethionine-6-N&rsquo, N&rsquo -adenosyl(rRNA) dimethyltransferase) (18S rRNA dimethylase).
>34420913 | P-selectin [Homo sapiens].
>38045915 | pregnancy-associated plasma protein A preproprotein [Homo sapiens].
>42558987 | Liprin-beta-2 (Protein tyrosine phosphatase receptor type f polypeptide-interacting protein-binding protein 2) (PTPRF-interacting protein-binding protein 2).
>62988324 | tenascin N [Homo sapiens].
>399356 | tRNA pseudouridine synthase 3 (tRNA-uridine isomerase 3) (tRNA pseudouridylate synthase 3) (Depressed growth-rate protein DEG1).
>1708405 | Isocitrate dehydrogenase [NADP] (Oxalosuccinate decarboxylase) (IDH) (NADP(+)-specific ICDH) (IDP).
>731842 | Uncharacterized protein YIL091C.
>52783333 | Probable ribosome biogenesis protein RLP24 (Ribosomal protein L24-like).
>223002 | fibrin beta.
>731778 | Peroxiredoxin DOT5 (Thioredoxin reductase) (Nuclear thiol peroxidase) (nTPx) (Disrupter of telomere silencing protein 5).
>129724 | Platelet-derived growth factor B chain precursor (PDGF B-chain) (Platelet-derived growth factor beta polypeptide) (PDGF-2) (c-sis) (Becaplermin).
>6226778 | Bone morphogenetic protein receptor type IB precursor (CDw293 antigen).
>143811458 | Sentrin-specific protease 2 (Sentrin/SUMO-specific protease SENP2) (SMT3-specific isopeptidase 2) (Smt3ip2) (Axam2).
>4699844 | Chain A, Crystal Structure Of Heparin And Integrin Binding Segment Of Human Fibronectin.
>122742 | Heparin-binding growth factor 2 precursor (HBGF-2) (Basic fibroblast growth factor) (BFGF) (Prostatropin).
>38154680 | lactoferrin [Homo sapiens].
>74734322 | BM022.
>20149543 | placental growth factor, vascular endothelial growth factor-related protein [Homo sapiens].
>49066024 | Uncharacterized protein YCR016W.
>117168301 | laminin, alpha 1 precursor [Mus musculus].
>6175077 | Small inducible cytokine A5 precursor (CCL5) (T-cell-specific RANTES protein) (SIS-delta) (T cell-specific protein P228) (TCP228) [Contains: RANTES(3-68); RANTES(4-68)].
>12644361 | Seminal plasma protein BSP-30 kDa precursor (BSP-30K).
>183951 | heparin binding protein.
>731827 | Mitochondrial acidic protein MAM33, mitochondrial precursor.
>13549118 | platelet factor 4 [Homo sapiens].
>3915598 | DNA-3-methyladenine glycosylase (3-methyladenine DNA glycosidase) (ADPG) (3-alkyladenine DNA glycosylase) (N-methylpurine-DNA glycosylase).
>47523194 | sperm associated AWN protein [Sus scrofa].
>133796 | 40S ribosomal protein S15 (RIG protein).

Non-heparin binding proteins:

>Phosphatidylinositol 4-kinase alpha (P42356)
>Cullin-2, homo sapiens
>P79896|S-(hydroxymethyl)glutathione dehydrogenase|EC 1.1.1.284|Sparus aurata|Swiss-Prot
>Ras-related protein Rab-22A, Homo sapiens
>P98005|cytochrome-c oxidase|EC 1.9.3.1|Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)|Swiss-Prot
>Q92T88|methylenetetrahydrofolate dehydrogenase (NADP+)|EC 1.5.1.5|Rhizobium meliloti (strain 1021)|Swiss-Prot
>hexokinase
>C9M0C5|purine nucleosidase|EC 3.2.2.1|Lactobacillus helveticus DSM 20075|TrEMBL
>Protein FADD, Homo sapiens
>P31414|inorganic diphosphatase|EC 3.6.1.1|Arabidopsis thaliana|Swiss-Prot
>Nuclear factor NF-kappa-B p100 subunit, Homo sapiens

>Transforming growth factor beta-1, Homo sapiens
>neuron-specific vesicular protein calcyon, Homo sapiens
>Succinate dehydrogenase [ubiquinone] flavoprotein subunit (P31040)
>A9WLH9|Glutarate-CoA ligase|EC 6.2.1.6|Renibacterium salmoninarum (strain ATCC 33209 / DSM 20767 / JCM 11484 / NBRC 15589 / NCIMB 2235)|TrEMBL
>Inhibin beta A chain, Homo sapiens
>O35400|alcohol sulfotransferase|EC 2.8.2.2|Mus musculus|Swiss-Prot
>Catenin alpha-2, Homo sapiens
>Phosducin, Bos taurus
>Kinesin family member 5, Dictyostelium discoideum
>Q15125|cholestenol DELTA-isomerase|EC 5.3.3.5|Homo sapiens|Swiss-Prot
>Rap guanine nucleotide exchange factor 2, Homo sapiens
>Beta-arrestin-1, Homo sapiens
>Maleate isomerase (O24766)
>Q08462|adenylate cyclase|EC 4.6.1.1|Homo sapiens|Swiss-Prot
>Paxillin, Homo sapiens
>A7GPY3|kynureninase|EC 3.7.1.3|Bacillus cereus subsp. cytotoxis (strain NVH 391-98)|Swiss-Prot
>P28332|alcohol dehydrogenase|EC 1.1.1.1|Homo sapiens|Swiss-Prot
>BH3-interacting domain death agonist, Homo sapiens
> Insulin receptor substrate 1, Homo sapiens
>Acetoacetyl-CoA reductase/transferase (P07097)
>P17516|3alpha-hydroxysteroid dehydrogenase (B-specific)|EC 1.1.1.50|Homo sapiens|Swiss-Prot
>Protein TOB2, Homo sapiens
>P47985|ubiquinol-cytochrome-c reductase|EC 1.10.2.2|Homo sapiens|Swiss-Prot
>P22570|ferredoxin-NADP+ reductase|EC 1.18.1.2|Homo sapiens|Swiss-Prot
>Follistatin, Homo sapiens
>Cyclic AMP-responsive element-binding protein 3, Homo sapiens
>Phosphatidylinositol 4-kinase beta (Q9UBF8)
>Tumor necrosis factor, Homo sapiens
>P08842|steryl-sulfatase|EC 3.1.6.2|Homo sapiens|Swiss-Prot
>Protein BTG3, Homo sapiens
>Monoamine oxidase, Homo sapiens
>RAR-related orphan receptor alpha, mouse
>C-X-C motif chemokine 10, Macaca nemestrina
>C4IU24|glyoxylate oxidase|EC 1.2.3.5|Brucella abortus str. 2308 A|TrEMBL
>ORAI calcium release-activated calcium modulator 2
>Disintegrin and metalloproteinase domain-containing protein 17, Homo sapiens
>P48449|Lanosterol synthase|EC 5.4.99.7|Homo sapiens|Swiss-Prot
>Tumor necrosis factor receptor superfamily member 13B, homo sapiens
>P25697|phosphoribulokinase|EC 2.7.1.19|Arabidopsis thaliana|Swiss-Prot
>Nuclear factor NF-kappa-B p105 subunit, Homo sapiens
>Histidine-containing phosphotransfer protein 5, Arabidopsis thaliana
>P17174|aspartate transaminase|EC 2.6.1.1|Homo sapiens|Swiss-Prot
>DNA helicase, Homo sapiens
>P55217|cystathionine gamma-synthase|EC 2.5.1.48|Arabidopsis thaliana|Swiss-Prot
>Mitogen-activated protein kinase kinase kinase 7, Homo sapiens
>Dopachrome isomerase, Homo sapiens
>Q03248|beta-ureidopropionase|EC 3.5.1.6|Rattus norvegicus|Swiss-Prot
>C6EF07|formate dehydrogenase-N|EC 1.1.5.6|Escherichia coli (strain B / BL21-DE3)|TrEMBL
>Mitogen-acivated protein kinase 1, Homo sapiens
>Signal recognition particle receptor FtsY, Escherichia coli (strain K12)
>B7J403|biotin synthase|EC 2.8.1.6|Acidithiobacillus ferrooxidans (strain ATCC 23270 / DSM 14882 / NCIB 8455)|Swiss-Prot

>Q9DBT5|AMP deaminase|EC 3.5.4.6|Mus musculus|Swiss-Prot
>Cryptochrome-1, Homo sapiens
>A3MHE3|arylformamidase|EC 3.5.1.9|Burkholderia mallei (strain NCTC 10247)|Swiss-Prot
>Bone morphogenetic protein 2, Homo sapiens
>P0CH37|alcohol dehydrogenase (NADP+)|EC 1.1.1.2|Mycobacterium smegmatis (strain ATCC 700084 / mc(2)155)|Swiss-Prot
>P0A2E2|glycine hydroxymethyltransferase|EC 2.1.2.1|Salmonella typhi|Swiss-Prot
>Bile acid-CoA:amino acid N-acyltransferase, Homo sapiens
>Mesaconyl-CoA hydratase OS=Methylobacterium extorquens

Testing Set:

Heparin Binding proteins:

>730057 | Suppressor protein STM1 (GU4 nucleic-binding protein 2) (G4p2 protein) (Triplex-binding protein 1) (3BP1) (Ribosomal subunits association factor) (AF) (POP2 multicopy suppressor protein 4) (TOM1 suppressor protein 1).
>461847 | Cleavage stimulation factor 64 kDa subunit (CSTF 64 kDa subunit) (CF-1 64 kDa subunit) (CstF-64).
>731705 | TPR repeat-containing protein YHR117W.
>17865670 | 26S protease regulatory subunit 4 (P26s4).
>68534974 | AQN-1 protein [Sus scrofa].
>84028283 | Probable ATP-dependent helicase YFR038W.
>135100 | Aspartyl-tRNA synthetase, cytoplasmic (Aspartate--tRNA ligase) (AspRS).
>116242826 | Tripartite motif-containing protein 41.
>134104968 | Chain A, Structure Of Cxcl12:heparin Disaccharide Complex.
>20141972 | WD-repeat protein 5 (WD-repeat protein BIG-3).
>6094485 | Rho guanine nucleotide exchange factor 5 (Guanine nucleotide regulatory protein TIM) (Oncogene TIM) (p60 TIM) (Transforming immortalized mammary oncogene).
>538354 | thrombospondin.
>74731080 | Uncharacterized protein C11orf52.
>114083 | Carbohydrate-binding protein AQN-3 (Zona pellucida-binding protein AQN-3) (Spermadhesin AQN-3).
>585379 | La protein homolog (La ribonucleoprotein) (La autoantigen homolog).
>2499599 | Mitogen-activated protein kinase 7 (Extracellular signal-regulated kinase 5) (ERK-5) (ERK4) (BMK1 kinase).
>74718607 | CDNA: FLJ21319 fis, clone COL02312.
>4506105 | prolactin [Homo sapiens].
>3127926 | collagen type VI, alpha 3 chain [Homo sapiens].
>546321 | protein C inhibitor; PCI [Homo sapiens].
>6649952 | bone morphogenetic protein 2 [Homo sapiens].
>4503413 | heparin-binding EGF-like growth factor [Homo sapiens].
>73858566 | heparin cofactor II precursor [Homo sapiens].
>1703371 | Sterol O-acyltransferase 2 (Sterol-ester synthase 2).
>113936 | Antithrombin-III precursor (ATIII).
>339741 | tumor necrosis factor.
>71152337 | Cdc42 effector protein 2 (Binder of Rho GTPases 1).
>730684 | E3 ubiquitin--protein ligase RSP5 (Reverses SPT-phenotype protein 5).
>4502461 | betacellulin [Homo sapiens].
>2498865 | rRNA biogenesis protein RRP5 (Ribosomal RNA-processing protein 5).
>74723962 | FLJ00195 protein.
>88853069 | vitronectin precursor [Homo sapiens].
>2497100 | Mitochondrial protein YML030W.
>74732911 | Within bgcn homolog (Drosophila).
>731815 | Hypothetical 19.2 kDa protein in SNP1-GPP1 intergenic region.
>1585498 | diamine oxidase.

>123509 | Heptapoietin A light chain (HPTA).
>1340171 | C4b-binding protein [Homo sapiens].
>6016557 | Melanoma-associated antigen B4 (MAGE-B4 antigen).
>116242759 | Rho GTPase-activating protein 4 (Rho-GAP hematopoietic protein C1) (p115).
>83721970 | sperm adhesion molecule 1 [Bos taurus].
>125259 | Casein kinase II subunit alpha (CK II alpha subunit).
>75428248 | Heparinase II protein precursor.
>730687 | 40S ribosomal protein S20.
>2500537 | ATP-dependent RNA helicase HAS1 (Helicase associated with SET1 protein 1).
>120549 | Follistatin precursor (FS) (Activin-binding protein).
>1723655 | Uncharacterized membrane protein YGR026W.
>1709190 | Sorting nexin MVP1.
>19856774 | Transcription factor EB.
>1168246 | Alpha-1A adrenergic receptor (Alpha 1A-adrenoceptor) (Alpha 1A-adrenoreceptor) (Alpha-1C adrenergic receptor) (Alpha adrenergic receptor 1c).
>4557727 | lipoprotein lipase precursor [Homo sapiens].
>8394103 | pleiotrophin [Rattus norvegicus].
>20455481 | Voltage-dependent L-type calcium channel subunit beta-1 (CAB1) (Calcium channel voltage-dependent subunit beta 1).
>118823 | Dolichol-phosphate mannosyltransferase (Dolichol-phosphate mannose synthase) (Dolichyl-phosphate beta-D-mannosyltransferase) (Mannose-P-dolichol synthase) (MPD synthase) (DPM synthase).
>134509 | Protein SIS1.
>145559466 | Probable ATP-dependent RNA helicase DDX43 (DEAD box protein 43) (DEAD box protein HAGE) (Helical antigen).
>62901400 | Vacuolar protein sorting-associated protein 66.
>116242631 | M-phase inducer phosphatase 3 (Dual specificity phosphatase Cdc25C).
>732161 | Hypothetical 55.9 kDa protein in MDS1-RPL13B intergenic region.
>15680217 | Cathepsin G [Homo sapiens].
>129719 | Platelet-derived growth factor A chain precursor (PDGF A-chain) (Platelet-derived growth factor alpha polypeptide) (PDGF-1).
>74761986 | HSD-4 protein.
>62288846 | PDZ domain-containing protein 4 (PDZ domain-containing RING finger protein 4-like protein).
>586408 | tRNA (cytosine-5-)-methyltransferase NCL1 (Trm4) (Multisite-specific tRNA:m5C-methyltransferase).
>1351928 | Methionine aminopeptidase 1 precursor (MetAP 1) (MAP 1) (Peptidase M 1).
>1711571 | RNA polymerase II transcriptional coactivator SUB1.
>4505849 | phospholipase A2, group IIA [Homo sapiens].
>113950 | Annexin A2 (Annexin II) (Lipocortin II) (Calpactin I heavy chain) (Chromobindin-8) (p36) (Protein I) (Placental anticoagulant protein IV) (PAP-IV).
>135807 | Prothrombin precursor (Coagulation factor II) [Contains: Activation peptide fragment 1; Activation peptide fragment 2; Thrombin light chain; Thrombin heavy chain].
>33959 | interleukin 8 [Homo sapiens].
>47523176 | porcine seminal protein I [Sus scrofa].
>133017 | 60S ribosomal protein L8-A (L4) (L4-2) (YL5) (RP6) (L7a-1) (Maintenance of killer protein 7).
>62243248 | insulin-like growth factor binding protein 3 isoform a precursor [Homo sapiens].
>307129 | hepatic lipase precursor.
>386853 | kininogen [Homo sapiens].
>1717874 | Ubiquitin carboxyl-terminal hydrolase 10 (Ubiquitin thioesterase 10) (Ubiquitin-specific-processing protease 10) (Deubiquitinating enzyme 10) (Disrupter of telomere silencing protein 4).
>119675632 | mast cell-restricted serine protease 7 [Mus musculus].
>120049 | Fibroblast growth factor receptor 2 precursor (FGFR-2) (Keratinocyte growth factor receptor 2) (CD332 antigen).
>11342670 | azurocidin 1 preproprotein [Homo sapiens].
>30172886 | Dynein light chain roadblock-type 2 (Dynein light chain 2B, cytoplasmic).

>2498960 | Ribosome biogenesis protein SSF2.
>2495255 | Non-histone protein 10 (High mobility group protein 2).
>731717 | Protein NSG1.
>1730165 | GU4 nucleic-binding protein 1 (G4p1 protein) (P42) (Protein ARC1).
>64654689 | Parathyroid hormone [Homo sapiens].
>84027749 | Uncharacterized protein YLR211C.
>462299 | Meiosis-specific protein HOP1.
>2280514 | histidine-rich glycoprotein [Homo sapiens].
>125932 | Tissue factor pathway inhibitor precursor (TFPI) (Lipoprotein-associated coagulation inhibitor) (LACI) (Extrinsic pathway inhibitor) (EPI).
>565136 | extracellular superoxide dismutase; EC-SOD [Homo sapiens].
>441174 | tissue plasminogen activator [Homo sapiens].
>153285461 | fibroblast growth factor 2 [Homo sapiens].
>547995 | Nucleosome assembly protein.
>122737 | Heparin-binding growth factor 1 precursor (HBGF-1) (Acidic fibroblast growth factor) (aFGF) (Beta-endothelial cell growth factor) (ECGF-beta).
>74732608 | Apoptosis-inducing factor 3 (Apoptosis-inducing factor-like protein).
>123116 | Hepatocyte growth factor precursor (Scatter factor) (SF) (Hepatopoeitin-A) [Contains: Hepatocyte growth factor alpha chain; Hepatocyte growth factor beta chain].
>74735535 | KIAA0731 protein.
>585880 | 60S ribosomal protein L4-A (L2) (YL2) (RP2).
>416746 | Azurocidin precursor (Cationic antimicrobial protein CAP37) (Heparin-binding protein) (HBP).
>50400820 | Methyl-CpG-binding domain protein 3 (Methyl-CpG-binding protein MBD3).
>136242 | N(2),N(2)-dimethylguanosine tRNA methyltransferase, mitochondrial precursor (tRNA(guanine-26,N(2)-N(2)) methyltransferase) (tRNA 2,2-dimethylguanosine-26 methyltransferase) (tRNA(m(2,2)G26)dimethyltransferase).
>119161 | Elongation factor 1-alpha (EF-1-alpha) (Translation elongation factor 1A) (Eukaryotic elongation factor 1A) (eEF1A).
>340146 | urokinase.
>1703084 | Acyl-CoA desaturase 1 (Stearoyl-CoA desaturase 1) (Fatty acid desaturase 1).
>416733 | C4b-binding protein alpha chain precursor (C4bp) (Proline-rich protein) (PRP).

Non-heparin binding:

>399508 | Peroxisomal hydratase-dehydrogenase-epimerase (HDE) (Multifunctional beta-oxidation protein) (MFP) [Includes: 2-enoyl-CoA hydratase ; D-3-hydroxyacyl CoA dehydrogenase].
>Protein Kinase C, Homo sapiens
>Tissue factor pathway inhibitor, Homo sapiens
>RIO-type serine/threonine-protein kinase Rio2, Archaeoglobus fulgidus
>Histatin-1, Homo sapiens
>Clathrin, light chain A, Homo sapiens
>Q7X9A6|plastoquinol-plastocyanin reductase|EC 1.10.9.1|Triticum aestivum|Swiss-Prot
>P53537|phosphorylase|EC 2.4.1.1|Vicia faba|Swiss-Prot
>Cyclin A-2, Homo sapiens
>Autophagy-related protein 3, Saccharomyces cerevisiae
>B0CEX1|Triose-phosphate isomerase|EC 5.3.1.1|Acaryochloris marina (strain MBIC 11017)|Swiss-Prot
>Mannose-binding lectin (protein C), Homo sapiens
>Q969G6|riboflavin kinase|EC 2.7.1.26|Homo sapiens|Swiss-Prot
>Q6UWM9|glucuronosyltransferase|EC 2.4.1.17|Homo sapiens|Swiss-Prot
>Q9BUT1|3-hydroxybutyrate dehydrogenase|EC 1.1.1.30|Homo sapiens|Swiss-Prot
>Caspase 8 (apoptosis-related cysteine peptidase), Danio rerio
>P20817|alkane 1-monooxygenase|EC 1.14.15.3|Rattus norvegicus|Swiss-Prot
>Proteasome maturation protein, Bos taurus
>Q96AT9|ribulose-phosphate 3-epimerase|EC 5.1.3.1|Homo sapiens|Swiss-Prot

>P10760|adenosylhomocysteinase|EC 3.3.1.1|Rattus norvegicus|Swiss-Prot
>Jun proto-oncogene (transcription factor AP-1), homo sapiens
>Cellular tumor antigen p53, Homo sapiens
>Vasopressin V1a receptor, Homo sapiens
>starch phosphorylase (Q9LIB2)
>Q9M643|Cycloeucalenol cycloisomerase|EC 5.5.1.9|Arabidopsis thaliana|Swiss-Prot
>Pancreatic lipase, Homo sapiens
>P22314|ubiquitin-activating enzyme|EC 6.3.2.B1|Homo sapiens|Swiss-Prot
>Q7Z5Q5|DNA-directed DNA polymerase|EC 2.7.7.7|Homo sapiens|Swiss-Prot
>Recoverin, Homo sapiens
>60S ribosomal protein L21, Homo sapiens
>Phospholipase C (P0C216)
>3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ
>P58966|(R)-citramalate synthase|EC 2.3.1.182|Methanosarcina mazei (strain ATCC BAA-159 / DSM 3647 / Goe1 / Go1 / JCM 11833 / OCM 88)|Swiss-Prot
>P14893|steroid DELTA-isomerase|EC 5.3.3.1|Bos taurus|Swiss-Prot

XB Patterns:

19 Most Occurring Patterns:

XBXBXXX
XXBXB
BXBXXXX
BXBXXX
XBXBX
XXBXBX
XBXB
XBXBXX
XBXX
BXBB
XBXBXXXX
XXBXBXXX
XXBXBXX
XXXBXB
BXXB
XXXXXXXXXXB
XXXXXXXXXXXB
XBXXX
XBXXXX

66 XB patterns:

XBXXBXBX
XBXBXXBX
XBBXBBX
XBBBXXBX
BBXXBBXX
XBBXXBBX
BXBXBBXB
XBBBXXBX
BBBBBXXXXX
XXBBXX
BXBXXX
XBBXXX

XBXBXX
BXXXXB
XXXXBB
XXXBXB
XXXBBX
BXXXBX
BBXXXX
XXBXXB
XBXXXB
XXBXBX
XBXXBX
BXXBXX
BBBXXX
XBBBXX
XXBBBX
XXXBBB
BXXXBB
XXBXBB
XBXXBB
BXXBBX
BXXBXB
BXBXXB
BBXXXB
BXBBXX
BXBXBX
XBXBXB
BBXXBX
BBXBXX
XBXBBX
XBXXBB
XXBBXB
XBBXBX
XBBBBX
XXBBBB
BBBBXX
BXBBBX
BXXBBB
BXBBXB
BBXXBB
XBXBBB
BBXBXB
BBXXBB
BBXBBX
BXBBBX
BBBXBX
BBBXXB
XBBXBB
BBBBBX
BBBBXB
BBBXBB
BBXBBB
BXBBBB
XBBBBB

BBBBBB

129 Biological Features:

Feature #	Feature Description
Feature 1	Charge of the Protein
Feature 2	Length of protein sequence
Feature 3	Molecular weight
Feature 4	Aromaticity
Feature 5	Instability Index
Feature 6	Isoeletric point
Feature 7	Helix fraction
Feature 8	Turn fraction
Feature 9	Sheet fraction
Features 10 – 29	Amino acid composition percentage
Features 30 – 49	Kyte & Doolittle index of hydrophobicity
Features 50 – 69	Flexibility
Features 70 – 89	Hopp & Wood of hydrophilicity
Features 90 – 109	Emini Surface fractional probability
Features 110 – 129	Janin Interior to surface transfer energy scale