Application of Support Vector Machines in Viral Biology

Sonal Modak, Swati Mehta, Deepak Sehgal, and Jayaraman Valadi

Abstract Novel experimental and sequencing techniques have led to an exponential explosion and spiraling of data in viral genomics. To analyse such data, rapidly gain information, and transform this information to knowledge, interdisciplinary approaches involving several different types of expertise are necessary. Machine learning has been in the forefront of providing models with increasing accuracy due to development of newer paradigms with strong fundamental bases. Support Vector Machines (SVM) is one such robust tool, based rigorously on statistical learning theory. SVM provides very high quality and robust solutions to classification and regression problems. Several studies in virology employ high performance tools including SVM for identification of potentially important gene and protein functions. This is mainly due to the highly beneficial aspects of SVM. In this chapter we briefly provide lucid and easy to understand details of SVM algorithms along with applications in virology.

Keywords Support vector machines · Supervised learning · Classification · Regression function identification · Epitope prediction · Quantitative structure activity relationships · Domain attributes · Attribute selection viral biology

S. Modak · S. Mehta · D. Sehgal
Life Sciences and Healthcare Unit, Persistent Systems Ltd., Pune, Maharashtra, India

J. Valadi (✉)
Life Sciences and Healthcare Unit, Persistent Systems Ltd., Pune, Maharashtra, India

Center for Development of Advanced Computing, Savitri Bai Phule Pune University, Pune, India
e-mail: jayaraman@cms.unipune.ac.in

P. Shapshak et al. (eds.), *Global Virology III: Virology in the 21st Century*,
https://doi.org/10.1007/978-3-030-29022-1_12

1 Introduction

Accurate annotation employing domain information extracted from sequence/structure and related attributes immensely enhances our current understanding of viral genomes. A major role is played by data driven modelling in recent advances made in vaccine development, epidemiology studies, pathogenicity determination, and drug design [1]. Introduction of NGS technology coupled with novel experimental techniques have provided very large volumes of data requiring accurate machine learning based modelling techniques. These techniques can be broadly categorized into supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning can be explained with a classic example of function annotation (see Fig. 1). In this task we have knowledge of certain number of sequences belonging to functional 'class1' from prior experimental annotation and knowledge of another set of sequences known not to be annotated as 'class2'. As shown in Fig. 1, a knowledge based model is built which separates data into two classes. This knowledge may be in terms of domain attributes extracted from sequences\structure, etc. The set of domain attributes are known as input data. Experimentally annotated class information is known as output data. The supervised learning model derives a functional relation between input and output. This model can be used to classify a query example to identify the functional class employing this model. This approach can be extended to classification into multiple functional classes.

In Unsupervised learning, we do not have prior knowledge about the classes. Unsupervised Learning is a class of Machine Learning techniques which enables us to discover patterns in the data. The data given to the unsupervised algorithm are not labelled, which means only the input variables (X) representing sequences\structure are presented to the algorithm with no corresponding output variables (Fig. 1). This type of learning is used extensively in viral biology to infer Phylogeny. The unsupervised learning method groups data without any prior knowledge of class labels. After the model is built one can derive knowledge about examples clustered in any

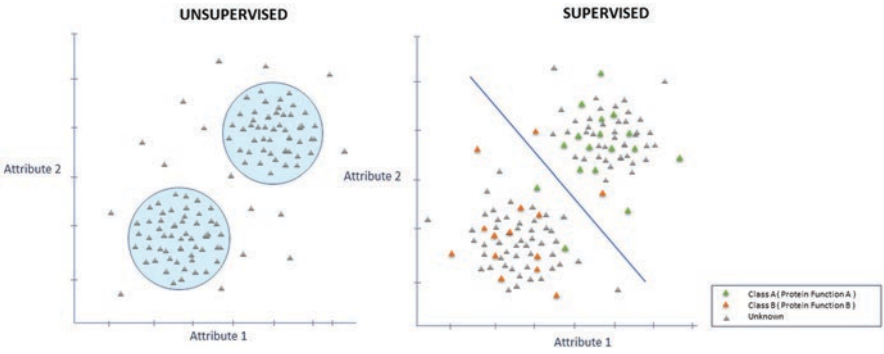

Fig. 1 Supervised vs. unsupervised learning

specific group. While supervised and unsupervised learning learn from data, the reinforcement learning paradigm learns from experience. In the following sections we provide details of SVM algorithms, a list of domain attributes presented to the algorithm, selection of informative attributes, and finally a discussion on some applications of SVM in viral biology.

2 Support Vector Machines for Classification

Support Vector Machines can be used both for supervised and unsupervised learning tasks. In viral biology, SVM is used mainly for supervised learning. SVM classifiers are a set of universal feed-forward network-based algorithms that have been rigorously formulated from statistical learning theory by Vapnik [2]. They are very popular machine learning paradigms which are routinely used in different branches of science and engineering.

2.1 SVM Binary Classifier for Linearly Separable Data

Let us take a simple case study to explain principle of SVM Linear Classification. The task is to build a model to separate a set of sequences belonging to functional class 1 from another set of sequences belonging to functional class 2. Class 1 examples can be peptides having antiviral activity while class 2 examples are not known to possess any antiviral activity. The input data vector for *ith* example is denoted by $\boldsymbol{x}_i$ and the corresponding class label is denoted y_i. The output of any example belonging to class 1 is represented by the subset $y_i = +1$ and those belonging to class 2 are represented by the subset $y_i = -1$. The hyperplane for the linearly separable data can be defined as:

$$\mathbf{w} \bullet \mathbf{x}_\mathbf{i} + b = 0$$

This hyperplane (Fig. 2) separates the data into two different classes. '**w**' refers to the weight vector with elements equal to the number of attributes. The problem here is to find out the best values of the elements of the weight vector, which maximize separation of the two classes with reference to a given performance measure (e.g. accuracy). This amounts to finding a hyperplane which maximizes the margin. This implies that at the training stage the examples belonging to class1 should be maximally separated from examples belonging to class 2. It can be shown that such a problem can be formulated as a Convex Quadratic Optimization problem [2]. The solution for such a convex optimization problem has only one global optimum as opposed to multiple local optimum solutions (algorithm can get stuck up in any of

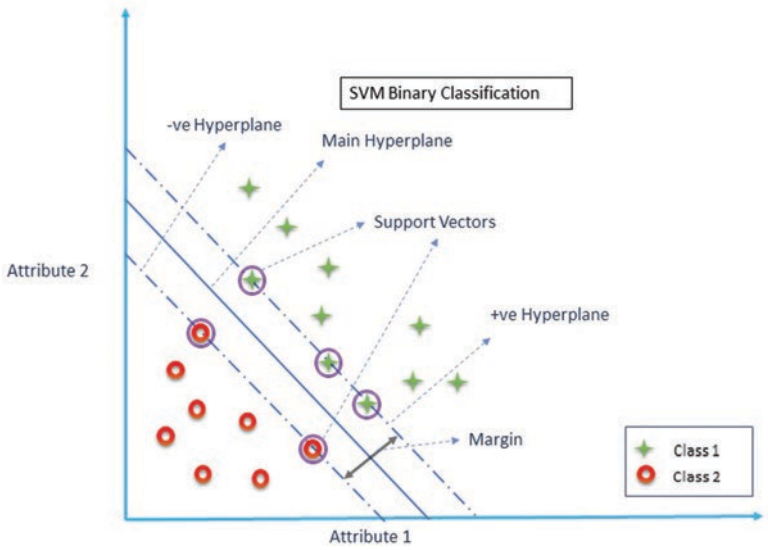

Fig. 2 Maximum margin-minimum norm classifier

the inferior local optima) like other candidate algorithms like neural network etc. have. It is this highly beneficial aspect coupled with superior performance has attracted researchers and practitioners from different fields to employ Support Vector Machines. After model building, the weight vectors can be obtained from only a subset of training examples. This subset is known as Support Vectors and hence the name Support Vector Machines. It must be noted here that SVM converts the original "**N**" dimensional problem into a one dimensional problem using dot products between the examples.

2.2 Non-linear Support Vector Machines

Biological data are inherently non-linear. A linear hyperplane cannot satisfactorily separate such non-linear data (Fig. 3). To handle these data SVM first transforms the data to a higher dimensional feature space and then employs a linear hyperplane. There are two inherent difficulties in the above approach: (i) It is difficult to find a suitable transformation by trial-and-error. (ii) We may have to employ a transformation to a very high dimensional space for reasonable classification accuracy which becomes computationally intractable. To solve these problems SVM employs appropriate kernel functions. Kernel functions are defined as a function of dot products in the original space and they are equivalent to the dot products in the higher dimensional feature space. SVM separating surface can now be defined as a linear

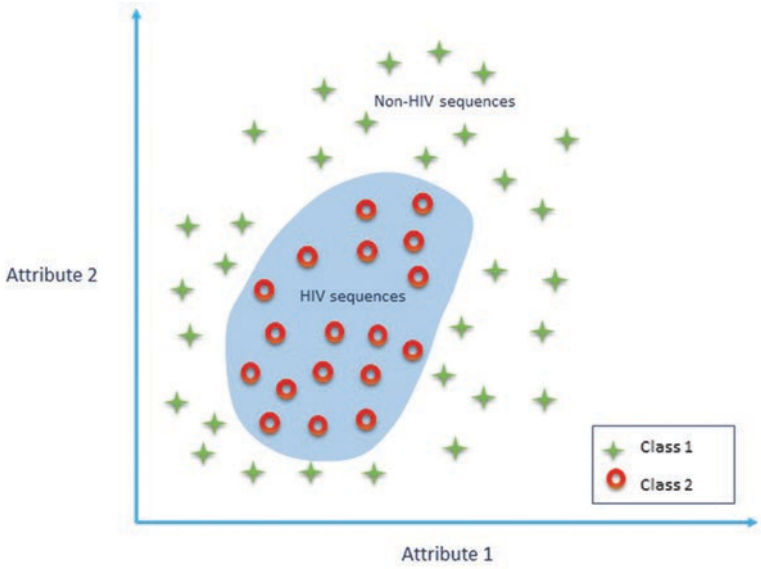

Fig. 3 Non-linearly separable data

hyperplane in the high dimensional feature space and introduction of appropriate kernel functions make it possible to do all the computations in the original space itself. Kernel functions have to satisfy Mercers Theorem; They have to satisfy the axioms of Hilbert space and have to be positive definite. The most popular kernel functions are Polynomial, Gaussian Radial Basis Function (RBF), and Multi-layer Perceptron kernel functions. Apart from these there are several domain dependent kernel functions. In computational biology, string kernels and Fisher kernels are very popular. Formulation as described above is known as Hard-margin SVM classification.

2.3 Soft Margin SVM

If we try to find a hyperplane which yields the maximum possible training accuracy, the margin obtained may become very narrow. Such a hyperplane while classifying the training set very well, over-fits the data and may fail miserably in unseen query test examples. It may be possible to increase the margin with slight loss of training accuracy (Fig. 4). This will generalize better than the one having a narrow margin and has more robust prediction capabilities. This trade-off between margin maximization and misclassification error in soft margin formulations can be obtained by optimizing a new parameter ‘C’.

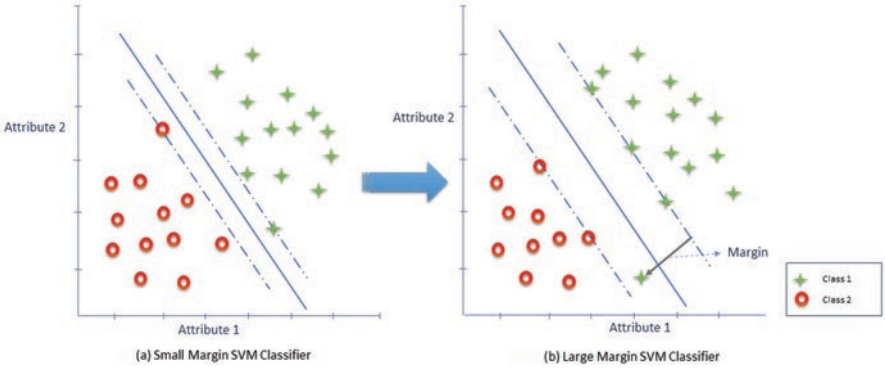

Fig. 4 Trade off: increasing margin/reducing misclassification

3 Brief Details of Classification of Real-Life Binary Datasets

Given a dataset we must first find the optimal hyperplane in the original dimension. In SVM terminology this is known as a linear kernel and after building the model we must estimate the required performance measure (e.g. accuracy). If it is not satisfactory, we must resort to nonlinear separation and employ conventional kernels like Polynomial, Gaussian Radial Basis Function (RBF), Exponential Radial Basis Function, Multi-layer Perceptron kernel functions etc. For every kernel, there are kernel parameters. With each kernel, apart from finding the best kernel parameters one must also tune the 'C' parameter as discussed in earlier section. If these kernels also are not satisfactory then we must resort to domain dependent kernels.

4 Support Vector Machines for Regression

In classification examples are grouped into discrete sets. In regression, a functional relationship is found between input data and output having continuous values. Many problems require a nonlinear model to adequately regress the data. The methodology described in the previous sections can be easily extended to employ SVM to handle nonlinear regression (Schölkopf et al. 1999) [2]. The methodology for linear regression is same as that of conventional models for regression; examples which are linearly classifiable can be done in the original dimension itself. What is different in SVM linear regression is that a novel epsilon-insensitive loss function is defined, which is robust against outliers in the data [3, 4]. For data that cannot be regressed linearly, a principle similar to the one implemented in classification problems can be extended easily; for such kind of problems, data needs to be taken to a higher dimensional feature space and subsequently regressed linearly. Appropriate kernel functions can again be defined to simplify computation.

5 Attributes Used in Viral Biology Problems

In viral biology we encounter a variety of attribute types, with each type providing huge magnitudes of domain attributes. Broadly, these attributes can be classified as sequence based, structure based, spectrum of light or radiation based (i.e. spectroscopic), microarray gene expression profiles etc. Protein sequence k-mer features range from amino acid (AA) (k = 1), dipeptide (k = 2), tripeptide (k = 3) to tetrapeptide (k = 4) and so on. It is possible to extract physiochemical properties like hydrophobicity, charge, hydrophilicity etc., from each of the AA alphabets. The simplest discrete set of features is the AA composition. Conversion of sequence information in terms of AA composition reduces the protein sequence into a 20 letter alphabet. While this is beneficial, we lose all sequence information. Recently Chou defined and introduced different types of pseudo- AA (PseAA) compositional attributes of protein sequences; these are a set of discrete numbers derived from AA sequences possessing some sort of sequence order or pattern information [5]. Ever since the first PseAA composition was formulated, these attributes have been successfully employed in several protein function identification tasks. Two classes of attributes frequently used in viral biology are listed below:

5.1 QSAR Descriptors

In quantitative structure activity relationship modelling, domain information about a molecule is provided in terms of different types of descriptors. The initially developed QSAR descriptors comprise hydrophobic, electric, and stearic parameters. Currently, descriptors of different dimensions ranging from 0 to 3 are routinely employed in modern QSAR analysis. Zero-dimensional descriptors comprise of atom counts, bond counts, molecular weight, sum of atomic properties; one dimensional descriptors two-dimensional descriptors deal with topological descriptors and three dimensional descriptors provide geometrical information. Originally QSAR is regression problem in which a functional relationship is obtained between activity of a molecule and the descriptors. This relationship can be linear or nonlinear so a regressor like SVM or random forest can be employed for this job. This is illustrated in Fig. 5.

5.2 PSSM Descriptors

Evolutionary information, one of the most important types of information in assessing functionality in biological analysis, has been successfully used to encode protein in many applications. PSIBLAST is used to repeatedly search specific databases, using a multiple alignment of high scoring sequences found in each search as input

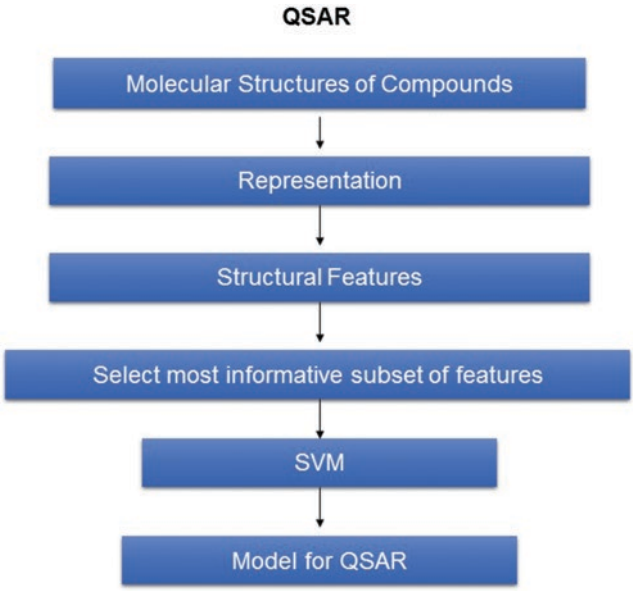

Fig. 5 QSAR regression using SVM

in the next round of searching. Normally iterations are continued until user specified number of iterations and at the end, the final Position Specific Scoring Matrix (PSSM) is generated. Such a matrix provides remote homology information and using PSSM attributes as descriptors in SVM would be useful if remotely connected sequences have similar functionalities. In the view of the fact that SVM requires the fixed length feature vectors, a vector of dimension 400 can be recovered from PSSM score matrix for use as input in SVM classifier.

Apart from the attributes described above, many different types of attributes are used, depending on the particular domain problem encountered.

5.3 Attribute Selection

Not all attributes are informative in data sets. Features which are non-informative will act as noise, do not have discriminative power & interfere with the classification process. Hence the model will have very little predictive accuracy. In Protein function identification in viral biology, several sequence and structural features can be extracted [6, 7]. For example the AA, dipeptide & tri-peptide compositional features put together amount to 8400 in number & not all of them will be important in a particular function annotation task. To select a subset of informative features by bruit force, we need to evaluate huge number of subsets of features which becomes

computationally time consuming. Various feature/attribute selection methods are available to simplify the process of subset selection. Feature selection techniques help us to avoid overfitting and improve model performance to provide faster and more cost-effective models; they also provide invaluable domain information. However, feature selection techniques have to employ appropriate search techniques, they bring in an additional level of complexity and computational cost. Feature selection techniques differ from each other in the way they incorporate this search in the added space of feature subsets in the model selection. Figure 6 illustrates the advantages of feature selection. These methods can be broadly classified as filter, wrapper and embedded methods.

5.3.1 Filter Ranking Methods

Filter ranking methods use some heuristics to score and rank the features (Fig. 7). In the example given above once the 8400 features are ranked by an appropriate filter method, the most informative subset of features can be selected, and the model can be built on this subset to maximize performance. Most popular filter methods include mutual information, student t-test, correlation-based feature selection (CFS) and several variants of the Markov blanket filter method, Minimum Redundancy-Maximum Relevance (mRmR) and Uncorrelated Shrunken Centroid (USC) algorithm. We give below some of the methods used in viral biology related problems:

5.3.1.1 Information Gain

Information gain score for any given attribute is calculated as the difference between entropy of the entire data set and the conditional entropy of for each possible value of the attribute. This can be done by binning each attribute and counting the fre-

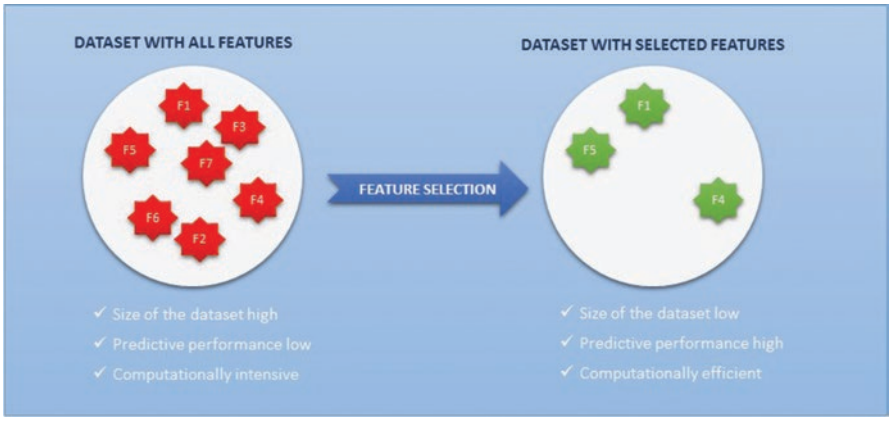

Fig. 6 Advantages of feature selection

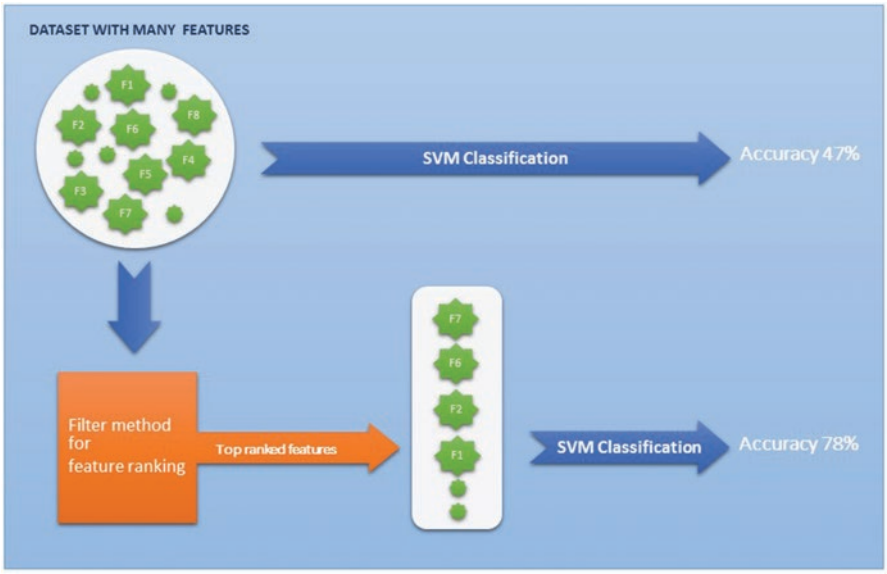

Fig. 7 Filter ranking & classification accuracy

quency of occurrence of different labels for the range of the attribute in each bin. Based on the score, top ranking attribute subset can be easily identified to build the model.

5.3.1.2 mRmR

The attributes are selected in such a way they are mutually dissimilar, non- redundant and maximally relevant simultaneously.

5.3.1.3 Mutual Information

Mutual information is a measure between random variables, that quantifies the information obtained about one of them, through the other. For the purpose of feature selection, mutual information between the subset of selected features and the target variable should be maximal.

5.3.1.4 Correlation Filter

The Correlation Feature Selection (CFS) selects subset of features that uncorrelated to each other but maximally correlated to the output variable.

5.3.1.5 Chi-Square

The chi-square test is a statistical test computes a score reflecting of independence to determine the dependency of two variables. We need to calculate chi-square statistics between every feature variable and the target variable and observe the existence of a relationship between the variables and the target. If the target variable is independent of the feature variable, we can discard that feature variable. If they are dependent, the feature variable is very important. For continuous variables, chi-square can be applied after "Binning" the variable.

5.3.2 Wrapper Methods

While filter methods are fast, they are not very accurate as they do not encode feature correlation. Wrapper methods employs a learning classifier for repeated evaluation of different subsets of features. These methods include forward selection & backward selection algorithms. In forward selection we start with an empty set and add features one by one which maximally improve accuracy until all features are added in the set. A subset can then be chosen which exhibits maximum accuracy. In backward selection we start with all features and remove least significant features one by one.

Recently recursive feature elimination wrappers have become very popular. In SVM recursive feature elimination algorithm, viz., SVM-RFE, the simulations start with all features and the algorithm weights are determined. Then features with least absolute value of weight are recursively removed until no feature is left out. Here again best performing subset can be easily identified which is used in the final model (see Figs. 8 and 9). Several wrapper based methods are population based and use Genetic algorithms, Ant Colony Optimization or other swarm intelligent methods. These methods mimic some nature inspired phenomena and evolve optimal solutions. Fie e.g. ACO is based on co-operative search behaviour of live ants. Biogeography is the study of distribution and dynamics of a large number of species geographically over a period of time. Biogeography based optimization (BBO)

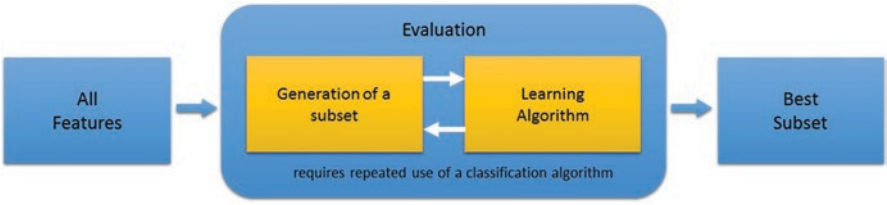

Fig. 8 Wrappers: schematic representation

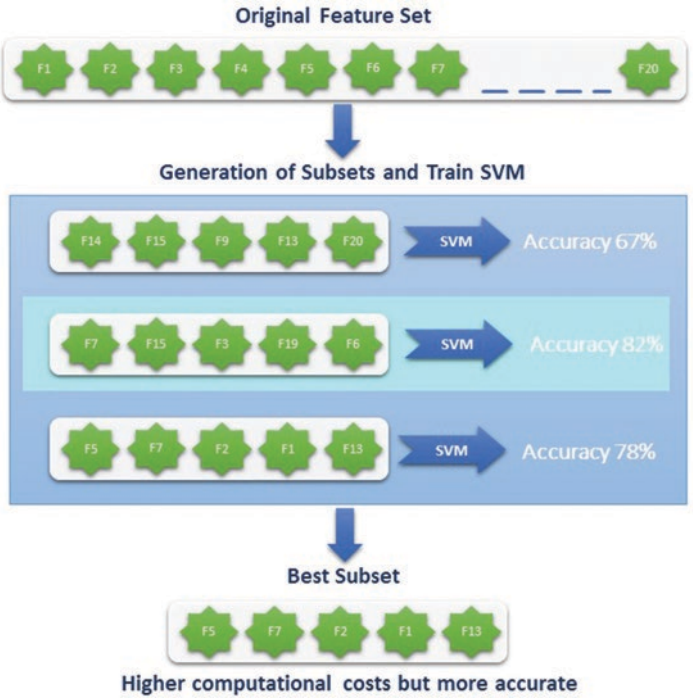

Fig. 9 Wrappers & classification accuracy

involves mimicking the natural processes of migration over a population in iterative generations, simulating discrete time. Atulji Srivatsava et al. employed BBO Simultaneous Feature Selection and MHC Class I Peptide Binding Prediction using Support Vector Machines and Random Forests [8].

5.3.3 Embedded Methods

In embedded class of feature selection techniques, optimal subset search is facilitated within the classification model. In random forest there are two inbuilt feature ranking methods, viz., Gini importance and variable importance. In SVM recursive feature elimination algorithm, viz., SVM-RFE, the simulations start with all features and the algorithm weights are determined. Then features with the least absolute value of weight are recursively removed until no feature is left out. Here again best performing subset can be easily identified which is used in the final model.

6 Performance Measures

While accuracy is the conventional performance measure, it may not be appropriate in all situations. In some examples we may require maximizing the positive accuracy while in some other situations negative accuracy may be the desired performance measure. Also, in imbalanced datasets, where we have more examples in one class than the other we have to optimize both positive and negative accuracies.

Referring to Fig. 10, true positives are the examples which are originally positive and are predicted positive by SVM. True negatives are the examples which are originally negative and predicted negative. False positives are the examples which are originally negative but predicted positive. False negatives are the examples which are originally positive, but predicted negative. With these definitions, we can define positive and negative accuracies. True positive rates or sensitivities are defined as;

$$TPR = \frac{\text{number of true positive examples}}{\text{total number of positive examples}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

True negative rate or specificity can be defined as:

$$TNR = \frac{\text{number of true negative examples}}{\text{total number of negative examples}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

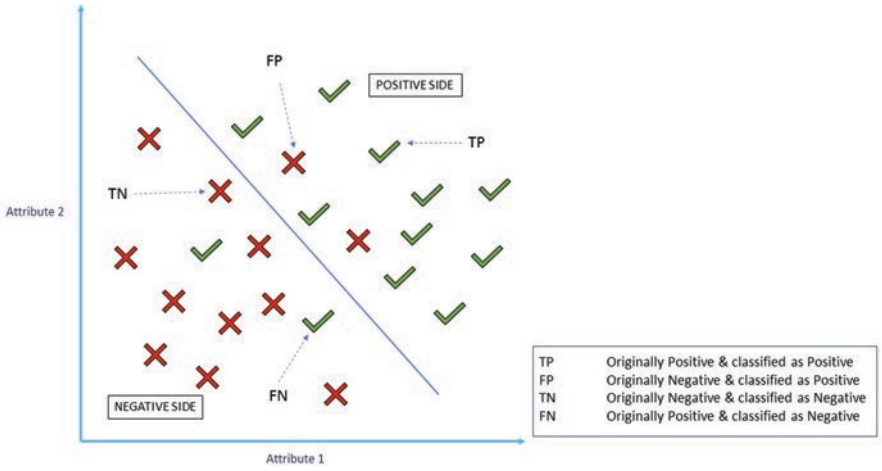

Fig. 10 Distribution of examples classified by the model

Precision or positive predictive value can be defined as:

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

F1 score is a harmonic mean of precision and sensitivity:

$$PPV = 2\left[\frac{\mathrm{PPV} * \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}\right] = \frac{2\mathrm{TP}}{2\left(\mathrm{TP} + \mathrm{FP} + \mathrm{FN}\right)}$$

Apart from these Matthew Correlation Coefficient is used as measure which provides optimal positive and negative accuracy and can be defined as:

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{\left(TP + FP\right)\left(TP + FN\right)\left(TN + FP\right)\left(TN + FN\right)}}$$

MCC score of -1 indiciates very poor classification and +1 indicates highest possible performance. In case of imbalance datasets it is customary to use MCC as the desired performance measure.

6.1 Cross Validation Measures

A simple way to test the performance is to split the data with 80% train and 20% test. The model id built on the 80% train data and tested on 20% test data. While this can be done for quickly estimating the performance of the model may not be fully adequate. To remove statistical bias two different cross validation measures are used to gauge the performance and obtain the best algorithm parameters. In K-fold cross validation, the training set is randomly divided into K-folds. To start with, the first fold is used as the test set and the remaining k $-$ 1 folds are used to build SVM hyperplane model. This model is evaluated by using the examples in the first fold. Similarly, each of the other k folds are used as test sets and the remaining k $-$ 1 folds are employed to build the models respectively. From these k experiments the cross validation accuracy is estimated as the average of k test accuracies. In leave one out cross validation procedure, each time one example is left out as a test example and the remaining n $-$ 1 examples are used to build the model. The built model is tested with the left out example. Conventionally fivefold or tenfold measures are used (k = 5 or k = 10). In k-fold cross validation, irrespective of the number of examples in the datasets, k different models are always built, whereas in leave-one-out cross validation, number of models is equal to number of examples in the training set.

7 SVM Extension to Solve Multi-class Type of Classification Problems

There are different algorithms, which address multi-class classification problem. Two well-known techniques include one-against-all method (Weston and Watkins 1999) and one-against-one technique [9]. One-against-all method considers the multi-class problem as a collection of binary classification problems. In general, k classifiers are needed to solve the k class problem. The kth classifier constructs a hyper-plane between class k and the $k - 1$ other classes. A majority vote across the classifiers is applied to classify the new test point. In one-against-one technique $k(k + 1)/2$ classifiers are needed. In each classifiers a model is built with examples of one class against examples of another class. Here again for a test example majority vote is needed to decide the class label.

8 Other SVM Types

8.1 Least Square SVM (LSSVM)

Least Square SVM classifier were proposed by Suykens and Vandewalle [10]. In their version of least square SVM they add a term in the objective function which penalizes square of error between prediction and actual class label. In this version, the problem is now formulated as a set of linear equations, instead of the convex quadratic problem for classical SVMs. Such a formulation makes computation simpler and faster. Several problems in bioinformatics has been solved using LSSVM. LSSVM formulation has also been extended for solving SVM problems.

8.2 One Class SVM

Several real-life datasets are highly imbalanced. Function annotation problems in viral biology have a small number of positive examples, while the negative examples can be very large. So such a distribution causes imbalance in the datasets and the minority class prediction accuracy will be very poor. One class SVM has been proposed in the literature to overcome this issue. In One class SVM only the data belonging to the majority class examples is used to build the model. There are two different version proposed in the literature for One class SVM. In the Tax and Duin's version [11], a model for the smallest hyper-sphere including all the majority class examples is formed. A new example is predicted as a majority class example if it falls inside the sphere. Otherwise it is predicted as a minority class example. For

non-linearly separable patterns appropriate kernel functions can be defined as in the case of binary SVMs. In the other version of the One class SVM, a hyperplane model is used instead of a hypersphere model [12]. One class SVM can also be used to detect anomalies and faults.

9 Applications of SVM in Virology

In this section, we outline a few important problems in viral biology where SVMs have been successfully applied on many case studies.

9.1 Quantitative Structure Activity Relationship (QSAR) Applications

Rapid assessment of desired activities of a large number of small-molecule compounds can be achieved by High throughput screening (HTS). QSAR analysis has been playing a key role in screening of compounds by building knowledge-based models [13]. This greatly reduces the experimental screening load. QSAR methodology focuses on finding a model, which allows for correlating the experimentally determined activity of a family of compounds with their molecular structure. Once a high performance model is built, it can be used to identify the activity of any new compound based on appropriate domain attributes extracted from their molecular structure. The set of atoms and covalent bonds between them can define a molecular structure. However, creation of structure-activity relationship models cannot be directly done from the structure of the molecule. Domain information has to be presented to the algorithm in the form of descriptors; molecular descriptors range from physicochemical and quantum-chemical to geometrical and topological features. The methodology of building QSAR models consists of four steps: (a) extracting descriptors from molecular structure (b) choosing most informative descriptors as per activity (c) building a model based on filtered molecular descriptors (d) screening molecule for activity in question. In Table 1, different example of descriptors as listed. These examples are categorised based on structural conformations [13].

Quantitative structure–activity relationship (QSAR) modelling with descriptor selection has become increasingly important because of a large number of descriptors of different types can be extracted in principle. Descriptor selection can improve the accuracy of QSAR classification studies and reduce their computation complexity by removing the irrelevant and redundant descriptors. Descriptor selection is an important pre-processing tool for QSAR studies. The sparse support vector machine (SSVM), one of the embedded methods, is of particular interest because it can perform descriptor selection and classification simultaneously.

Table 1 Examples of different descriptors based on structural conformation [15]

Category	Descriptors
2D QSAR descriptors	Constitutional descriptors
	Electrostatic and quantum-chemical descriptors
	Topological descriptors
	Geometrical descriptors
	Molecular fingerprints & fragment-based descriptors
3D QSAR descriptors	Comparative molecular similarity indices analysis
	Comparative molecular moment analysis
	Weighted holistic invariant molecular descriptors
	VolSurf approach
	Grid-independent descriptors

Further explanation is included in Sect. 5.3.

SVMs have been found to provide robust and accurate QSAR models for several problems encountered in viral biology. Two types of QSAR models can be build. First one is a regression problem in which a model is built against descriptors vs. experimentally annotated activities. This is a regression problem, schematically shown in Fig. 11. The second problem can be posed as classification problem. For this a threshold value for the experimental activities has to be defined. Compounds having activities less than these threshold activities are grouped into 'class1'. The other compounds are grouped into 'class2'. SVM classification model is built to separate compounds into two groups. A new query compound can then be classified as active or inactive as schematically represented in Fig. 12.

Human immunodeficiency virus (HIV) affects and destroys the immune system and causes acquired immunodeficiency syndrome (AIDS) disease. As per the **UNAIDS report** [14], 77.3 million [59.9 million–100 million] people have become infected with HIV since the start of the epidemic and 35.4 million [25.0 million–49.9 million] people have perished from AIDS-related illnesses since the start of the epidemic. Numerous molecular modelling approaches have been attempted to address the design of new anti-HIV compounds. Most of them are based on QSAR [15]. In an interesting and comprehensive study [15] QSAR based attributes were selected for predicting inhibiting activity of the compound against HIV proteins including protease (PR), reverse transcriptase (RT) and integrase (IN). Around 18,000 molecular descriptors which include geometric, electrostatic, structural, constitutional, path and graph fingerprints etc. were extracted utilizing the open source PaDEL software. To reduce the number of descriptors Attributes selection was carried out using 'Best-First' as the search method in Waikato Environment for Knowledge Analysis (Weka) suite. SMO regression algorithm in the Weka suite was

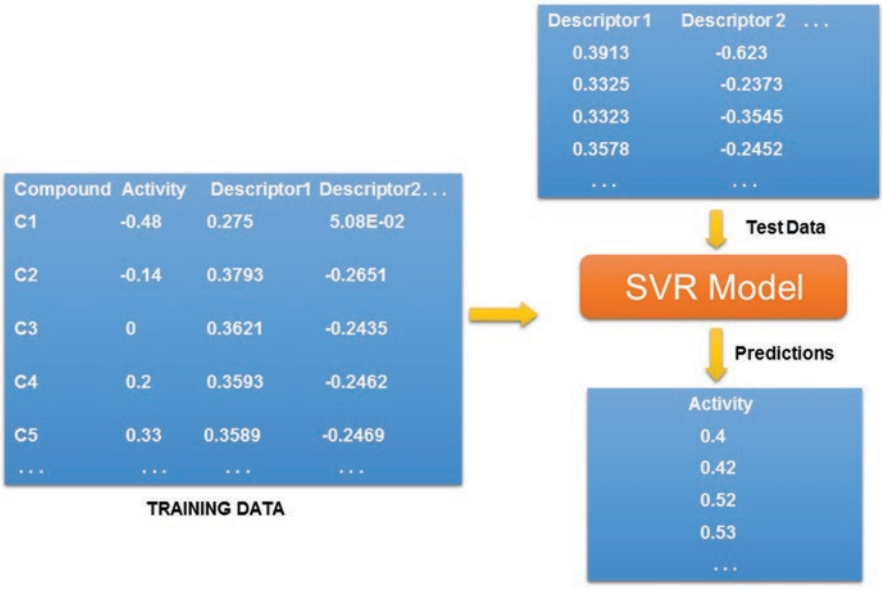

Fig. 11 SVM regression model

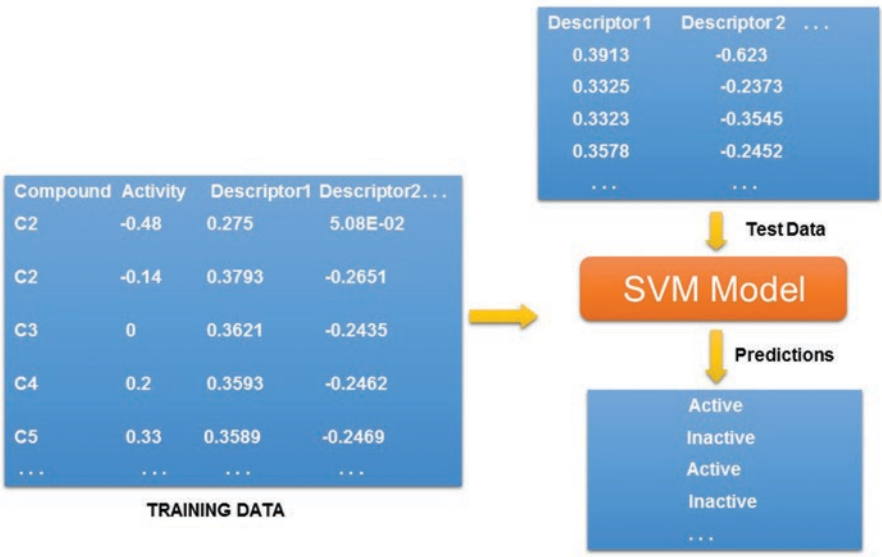

Fig. 12 SVM classification model

used to classify the data into active and inactive sets. The models were able to achieve excellent values of Pearson correlation coefficient for all the three data sets, *viz*, PR, RT, IN. An integrated web based Platform **HIVprotI** [16] was further developed using this model.

The tetra-hydro-imidazo[4,5,1-jk][1,4]-benzodiazepines (TIBOs), constitute a group of potent system inhibitors of HIV-1 reverse transcriptase. With a view to segregate TIBO compounds into high and low classes of inhibitors of HIV-1 reverse transcriptase, Hdoufane et al. carried out SAR studies on 89 TIBO derivatives using different classifiers, such as support vector machines, artificial neural networks, random forests, and decision trees [17]. They successfully employed seven molecular descriptors characterizing hydrophobic, electronic, and topological aspects of the molecules and obtained excellent training and test accuracies.

The successful identification of HIV proteins may have important significance in treatment since epidemiological and biological characteristics of HIV-1 and HIV-2 are quite different., Juan Mei et al. employed SVM along with other classifiers to predicted HIV-1 and HIV-2 proteins based on pseudo AA compositions and increment of diversity (ID) algorithm [18]. With jack knife tests, SVM models gave the highest prediction accuracy of 0.9909.

Both HBV and HCV are of immense significance as leading causes of liver cancer as well as co-infection with HIV. A potentially important study included 172 positives and 8998 negative cases and built a classification model of the HBV dataset; in the same study HCV dataset included 533 positives and 7287 negatives [19]. The data had obvious imbalance in the number of examples in the positive and negative data sets. Three different imbalance handling methods, viz., (i) Downsize, (ii) Multi downsize, and (iii) SMOTE were used. SMOTE provided the best performance; SVM prediction accuracies of 64% for HBV and 71% for HCV were reported for this model.

Influenza, a respiratory virus, is correlated with high morbidity and mortality rates. Neuraminidase (NA) and haemagglutinin (HA) are two major glycoproteins found on the surface of the influenza virus. Compounds that inhibit neuraminidase can protect host cells from viral infection and retard the spread of the virus among cells. A two staged approach has been used to build a QSAR classification model separating neuraminidase as active and inactive [20]. In the first stage minimum redundancy maximum relevance criteria was employed to select the most informative descriptors. The second stage employs the selected descriptors as input to SSVM L1-norm classifiers. The dataset consisted 479 neuraminidase inhibitors of H1N1 virus whose experimentally measured IC50 values were available. These set of training compounds were separated by thresholding the activity into two categories: active compounds with IC50 <20μM, while those with IC50 >20 μM were considered to be weakly active compounds. The 7 top descriptors selected gave an SVM classifier accuracy of 90.62% which is far higher than the earlier SVM approaches.

The classification of protein quaternary structure complex is of significant interest in computational biology research. Chi-Chou Huang at.al have developed a two-staged architecture for five class classification of grouping protein quaternary structure of a complex; the five classes are monomer, dimer, trimer, tetramer, and other subunit classes [21]. AA frequency, Shannon entropy and accessible surface areas were employed as domain attributes. One against all SVM classifiers were used of in which positive data consisted of examples of given class and negative data consisted of all the remaining classes. Due to this division number of examples in the positive side of the classifier was much less than the negative side. This created imbalance and reduced classification accuracy. To counter this, the author employs a bootstrap method for repeated sampling and generated different subsets of data. The majority class was further subjected to random sub-sampling. Mathews Correlation coefficient was used as performance measure. The bootstrapping method was able to produce an MCC of 0.696 and above. List of examples are given in Table 2.

Table 2 Illustrative examples for QSAR applications

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
1.	A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine [20]	To predict the neuraminidase inhibitors as active and inactive based on QSAR using 479 neuraminidase inhibitors of H1N1 virus with experimentally measured IC50 values.	Molecular structures of the compounds were sketched using Chem3D software. Dragon software (version 6.0) was used to generate 4885 molecular descriptors including all 29 blocks based on the optimized molecular structures, 2881 left after cleaning up
2.	In Silico SAR Studies of HIV-1 Inhibitors [17]	Classify TIBO compounds into two groups: High and low inhibitors of HIV-1reverse transcriptase based on QSAR studies.	500 molecular descriptors from five different classes (geometrical, topological, constitutional, electrostatic, and quantum-chemistry descriptors).
3.	HIVprotI: An integrated web based platform for prediction and design of HIV proteins inhibitors [16]	A web server to predict inhibition activity of a compound against HIV proteins namely protease (PR), reverse transcriptase (RT) and integrase (IN).	18,000 molecular descriptors extracted using PaDEL software which include geometric, electrostatic, structural, constitutional, path and graph fingerprints
4.	PClass: Protein quaternary structure classification by using bootstrapping strategy as model selection [21]	A web server for protein quaternary structure complex classification into 5 categories namely: Monomer, dimer, trimer, tetramer, and other subunit classes	AA freq. Shannon entropy and accessible surface areas

(continued)

Table 2 (continued)

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
5.	Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers [18]	To predict of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions	(i) 20 AA compositions (A1) (ii) 400 dipeptide compositions (A2) (iii) AA hydropathy compositions (H1) (iv) 36 hydropathy dipeptide compositions (H2)
6.	iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features [96]	Tool to discriminates between host located and non-host located phage proteins (PH & non-PH) and membrane and cytoplasm located host proteins (PHM & PHC).	PSSM, AA composition and structural features of the sequences Above features are generated using (1) PSSM file generated from PSI-BLAST and (2) SPD file generated from SPIDER2 software.
7.	Enhancement of hepatitis virus immunoassay outcome predictions in routine pathology data by data balancing and feature selection before the application of support vector machines [19]	Prediction of HBV and HCV for negative and positive using Balancing methods to counter negative samples.	25 variables from laboratory
8.	QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM) [97]	Predict bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors	MACCS fingerprint, 20 global molecular descriptors and 88 2D property-weighted autocorrelation descriptors calculated using CORINA Symphony
9	A computational model for predicting transmembrane regions of retroviruses [98]	Identify transmembrane regions in envelope glycoproteins of retroviruses (HERV, HIV, HTLV, SIV, MLV)	10 physicochemical and PSSM score features

9.2 SVM Applications Based on Next Generation Sequencing (NGS) Data

The term 'Next-Generation' Sequencing (NGS)' refers to the advancement in nucleic acid sequencing technologies. Numbers of sequence reads generated per run has progressively increased with time, due to improved understanding of molecular biology as well as technological advances. Current sequencing platforms are capable of generating enormous numbers of sequence reads in quick turnaround time, allowing researchers to explore all possible aspects of biomedical studies at molecular level and dig deeper in the genetic aspects. NGS has proven to be an efficient,

fast and reliable approach to solve problems in studies of evolution, ecology and genetics, overcoming the limitation of traditional molecular approaches [22]. Another great advantage of NGS approach over traditional molecular studies is that it is also cost efficient. End-to-end human genome can be sequenced in few hours using NGS technology, whereas, it took over a decade to sequence and assemble human genome using Sanger Sequencing. Based upon the chemistry, a number of NGS platforms have been developed since last decade. Bioinformatics knowledge plays an important role in assembling the fragments sequenced in parallel by mapping all the read sequences to the human genome reference. Depth of the sequencing, i.e. number of times the template has been sequenced, assures accuracy of sequencing, making sure that observed variation in sequenced data is the result of mutations, and not of sequencing errors. NGS can be used to sequence targeted regions identified in a genetic study, or entire genome including all coding genes (whole exome sequencing).

The variations in human genome can be a few nucleotide base changes (substitutions), insertions, and deletions of DNA, large genomic deletions of exons or whole genes and rearrangements such as inversions and translocations. All these anomalies are collectively termed 'mutations'. Traditional methods of sequencing were only able to discover handfuls of mutations including small insertions and deletions. This led to the development of dedicated assays, to discover additional types of variations. Some of the examples includes fluorescence in situ hybridization (FISH) for conventional karyotyping, or comparative genomic hybridization (CGH) microarrays to detect sub-microscopic chromosomal copy number changes such as microdeletions.

With recent advancements in NGS technologies and better understanding of life at genomic level, various questions have been answered using whole genome sequencing. Areas of applications includes genome diversity, metagenomics, epigenetics, discovery of non-coding RNAs and protein-binding sites, and gene-expression profiling by RNA sequencing [22–26]. Apart from high-throughput whole genome sequencing, typical applications of NGS methods in microbiology and virology are discovery of new microorganisms and viruses by using metagenomic approaches, investigation of microbial communities in the environment and in human body for understanding healthy and disease conditions, analysis of viral genome variability within the host, detection of antiviral drug-resistance mutations in patients with human immunodeficiency virus (HIV) infection or viral hepatitis, etc.

In the context of Microbial Analysis, the term metagenomics designates the analysis of all of the nucleic acid present in a given sample. Without isolating and culturing individual microbial species, entire communities of microorganisms can be explored. NGS applications in metagenomic studies include the discovery of novel viruses from clinical samples in human and animal diseases, e.g. the new Ebola virus Bundiubugyo [27], identification of a viral etiology of disease outbreak in honeybees [28], and involvement of a new arenavirus in transplant-associated disease clusters [29]. Scope of applications also include characterization of the viral community in the environment [30, 31], in animals [32], and viral community in

humans [33–36]. Due to high replication capacity and low fidelity of the replication enzyme, high intra-host variability is shown by reverse transcriptase-dependent viruses (e.g. hepatitis B virus, human immunodeficiency virus) and RNA viruses (e.g. hepatitis C virus, influenza virus). Such a set of closely related genomes within a given host allows a viral population to swiftly adapt to dynamic environments and evolve resistance to vaccines and antiviral drugs [37]. Significant work using NGS has been done for the characterization of intra-host variability of influenza virus [38, 39], HCV, HIV and HBV.

Jian'an Jia et al. designed an approach to distinguish between 2 disease groups caused by Hepatitis B Virus – Chronic Hepatitis B (CHB) and Hepatocellular Carcinoma (HCC) [40]. NGS was used to sequence the pre-S region of a large number of CHB and HCC individuals. The attributes used were word pattern frequency vector of various lengths ranging from k = 2 to k = 8. Maximum CV mean AUC of 0.93 k = 5. The prediction accuracy was found to be much higher than prediction results using KNN classifiers.

To investigate HBV genotypes and predict HCC status, Xin Bai et al. used NGS to sequence the pre-S region of the HBV sequence of 94 HCC patients and 45 chronic HBV (CHB) infected individuals [41]. Word pattern frequencies among the sequence data of all individuals were calculated and compared using the Manhattan distance. The individuals were grouped using principal coordinate analysis (PCoA) and hierarchical clustering. Word pattern frequencies were also used to build prediction models for HCC status using both K-nearest neighbours (KNN) and support vector machine (SVM). In the independent data set of 46 HCC patients and 31 CHB individuals, a good AUC score of 0.77 was obtained using SVM.

Apart from applications viral disease diagnosis, a recent study demonstrates usefulness of a hybrid approach in early assessment of the risk by predicting the host of influenza viruses using the Support Vector Machine (SVM) classifier based on the word vector, representation and feature extraction method for biological sequences [42]. Accuracies for host prediction in avian, human & swine influenza viruses were 99.7%, 96.9% & 90.6%, respectively. Table 3 contains some examples of SVM application using NGS data to address problems in virology studies.

9.3 SVM Applications Based on Spectroscopy Data

From array of several spectroscopic techniques, Raman spectroscopy and Infrared (IR) absorption spectroscopy have led to major breakthroughs in biological, pharmaceutical, and clinical research [43–45]. With use of visible-light laser beams, Raman spectroscopy can be used as a non-invasive characterization technique and achieve resolution same as fluorescence microscopy. The inelastic scattering of light photons by vibrating molecules in the samples is called as Raman scattering. Information about molecular vibrations produced due to change in frequencies of the photons are useful in diagnostic studies. Such change in frequencies are result of interactions of molecular bonds. Initial changes in almost all the types of diseases

Table 3 Illustrative examples for SVM applications based on NGS approach

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
1.	Next-generation sequencing revealed divergence in deletions of the preS region in the HBV genome between different HBV-related liver diseases [40]	Distinguish between 2 disease groups caused by Hepatitis B Virus – Chronic Hepatitis B (CHB) and Hepatocellular Carcinoma (HCC)	Nucleotide deletion % obtained from sequences. It is defined as 100 × (counts of reads with deletion in single nucleotide site)/(total number of reads including such a nucleotide site)
2.	Deep sequencing of HBV pre-S region reveals high heterogeneity of HBV genotypes and associations of word pattern frequencies with HCC [41]	Investigate HBV genotypes and to predict HCC status using sequences of pre-S region of the HBV sequence of HCC and HBV patients.	Word pattern frequency vector of various lengths ranging from k = 2 to k = 8
3.	Predicting the host of influenza viruses based on the word vector [42]	Predict the host(human, avian & swine) of influenza viruses based on the word vector	200-dimension vectors of all proteins and DNA sequences generated using "word2vec" To vectorize protein, the sequence is separated into overlapping words of size 2–4. The word vector of all the words are summed up and averaged that results in 200-dimension vector for each protein. Same done for DNA

(including cancer and viral infections) occur at molecular level. Laboratory tests are inadequate in identifying such changes due to some limitations. Raman spectroscopy has the potential to monitor these changes at molecular level at early stage of the disease [46]. Information about abnormalities can be retrieved from the spectral differences between normal and diseased samples, which is used for the purpose of diagnosis. With diverse areas of applications, Spectroscopy is a promising clinical tool for the real-time diagnosis of diseases and assessment of living healthy and cancerous tissue, cells and their subcellular compounds and structures. It can also be used to track the mode of action of drugs on a molecular level.

Due to its high sensitivity and selectivity Raman spectroscopy requires only a small sample volume and minimal preparation efforts. The high resolution, ease of sample preparation, and very short data collection time required make the technology ideal for use in the study of viruses and virally infected cells. As the acquisition can be fast, processes in real time can be studied. In different conditions and environments, informative molecular details can be extracted since water environment can disturb these spectra to a slight extent. Therefore, this technique is ideal for studies like viral protein assembly, dynamics, interactions and structural alterations, compared to other available methods. The stereochemistry and structures of pro-

teins and nucleic acid components of viruses, can be determined using spectroscopy [47, 48]. The conformational changes that leads to viral procapsid and capsid assembly was identified using Raman spectroscopy [49, 50]. Raman spectroscopy is effective also in distinguishing between even the homogenous viruses, thereby increasing its possible role even further in diagnostic medicine.

Dengue fever, Yellow fever, Japanese encephalitis, Murray Valley encephalitis, tick-borne encephalitis and West Nile encephalitis are diseases attributed to flavivirus infection. Early detection is important to prevent these diseases from progressing into the severe or terminal stages. Non-structural protein 1 (NS1) is acknowledged as one of the biomarkers for flavivirus related diseases. Radzol AR et al. defined a model for PCA-SVM with MLP kernel for classification of flavivirus biomarker, NS1 molecule, from Surface Enhanced Raman Spectroscopic (SERS) spectra of saliva [51]. Best PCA-SVM (MLP) model defined in this study yielded accuracy of 96.9%.

Another example of life-threatening viral infection is Hepatitis B, that attacks the liver. In a study analysing hepatitis B virus (HBV) infection in human blood serum using Raman spectroscopy combined with pattern recognition technique, SVM model with two different kernels i.e. polynomial function and Gaussian radial basis function (RBF) were investigated for the classification of normal blood sera from HBV infected sera based on Raman spectral features [52]. Best performance achieved for polynomial kernel of order-2 with accuracy of 98% using fivefold cross-validation.

In case of chronic hepatitis C, liver biopsy has been the reference for staging the degree of fibrosis until the last decade. For obvious reasons, non-invasive tests e.g. blood tests measuring the markers that are either involved in the synthesis or degradation of extracellular matrix, has to be the preferred alternatives for assessment of hepatic fibrosis. However, the performance of these non-invasive methods is limited in differentiating between mild and moderate stages of fibrosis and in evaluating the effect of treatments on liver fibrosis process. Use of Fourier transform infrared (FTIR) spectroscopy applied to the serum in the assessment of hepatic fibrosis, was demonstrated by Scaglia et al. [53]. Infrared spectral characteristics exhibited by serum from patients, were employed in differentiation of chronic hepatitis C patients with extensive hepatic fibrosis from those without fibrosis and thus predicting the degree of hepatic fibrosis. With leave-one-out cross-validation, the accuracy achieved was 97.7%.

A similar study was performed for the classification of dengue suspected in human sera. SVM models built on the basis of three different kernel functions including Gaussian radial basis function (RBF), polynomial function and linear function were employed to classify the human blood sera based on features obtained from Raman Spectra [54]. With the tenfold cross validation method, best results were obtained for the polynomial kernel of order 1 with diagnostic accuracy of about 85%.

The applications are not limited to only medicinal diagnosis. Viruses could infect over hundreds of different species of plants, including crops of tobacco, tomato,

pepper, cucumber, etc. Viruses can survive outside the plant, and remain in a dormant state to infect growing crops. Once the plant is infected, no chemical cure is effective, and usually all the infected crops should be removed. For detecting seeds infestation caused by cucumber green mottle mosaic virus (CGMMV), near-infrared (NIR) hyperspectral imaging system was used to discriminate virus-infected seeds from healthy seeds with partial least square discriminant analysis (PLS-DA) and least square support vector machine (LS-SVM) [55]. The classification accuracy for virus-infected watermelon seeds were 83.3% with the best model.

Whereas Jiyu Peng et al. proposed an approach to discriminate TMV-infected tobacco based on laser-induced breakdown spectroscopy (LIBS) [56]. Two different kinds of tobacco samples (fresh leaves and dried leaf pellets) were collected for spectral acquisition, and partial least squared discrimination analysis (PLS-DA) was used to establish classification models. In prediction set, 94.4% and 94.7% accuracies obtained for observed emission lines of dried & fresh leaves. Compared to PLS-DA, SVM was proved to be efficient to eliminate influences of moisture content. Some other examples are listed in Table 4.

9.4 SVM Applications for Epitope Prediction

An epitope is a specific target of a few AA residues on an antigen molecule that is recognized by B-cells or T-cells of the immune system [57, 58]. A B-cell epitope is the antigen portion that binds to B-cell Receptor (BCR) on B-cells, where BCR contains membrane-bound antibody. There are 2 types of B-cell epitopes based on their orientation. One is linear epitope that comprises of a continuous string of AA s. The second one consisting of most B-cell epitopes is conformational epitope which is made up of discontinuous AAs that comes close with protein folding [59, 60]. A T-cell epitope binds to the major histocompatibility complex (MHC) on surface of antigen-presenting cells (APCs) and MHC presents the antigen to the T-cell receptor (TCR) on T-cells [59]. The major histocompatibility complex (MHC) or human leukocyte antigen (HLA) is the gene family that helps the immune system to identify and destroy the foreign substance [61].

Vaccines have proven to be useful tools to control various viral diseases like influenza, smallpox, polio, hepatitis and rotavirus. The conventional methods of developing vaccines include attenuated or killed whole pathogen that improves immunity to a specific disease and involve only experimental methods of epitope identification. Vaccine development takes a long time with conventional methods because of the time consuming experimental screening of huge number of potential candidates [62]. The fact that only few AA residues are detected by B- and T-cells instead of whole pathogen is leveraged for vaccine development, understanding disease etiology, disease diagnosis and immune monitoring [58, 59]. Moreover, with advances in next-generation sequencing methods, proteomics, and transcriptomics

Table 4 Illustrative examples for SVM applications using spectroscopy

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
1.	Detection of cucumber green mottle mosaic virus-infected watermelon seeds using a near-infrared (NIR) hyperspectral imaging system: Application to seeds of the "Sambok Honey" Cultivar [55]	Classification of infected and healthy watermelon seeds using a near-infrared (NIR) hyperspectral imaging system. Hyperspectral imaging data 51 healthy & 45 infected samples were used.	Near infrared spectrum
2.	PCA criterion for SVM (MLP) classifier for flavivirus biomarker from salivary SERS spectra at febrile stage [51]	Classification of flavivirus biomarker, NS1 molecule, from Surface Enhanced Raman Spectroscopic (SERS) spectra of saliva. SERS spectra of 64 NS1 adulterated dataset and 64 control dataset were used.	Spectral data with 1801 features per spot per sample.
3.	Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning [52]	Analysis of hepatitis B virus (HBV) infection in human blood serum using Raman spectroscopy. Serum samples of 119 confirmed HBV infected patients and 84 healthy volunteers were used.	Raman spectral features
4.	Noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis C using serum Fourier transform infrared spectroscopy [53]	Non-invasive differentiation of chronic hepatitis C (CHC) patients with extensive hepatic fibrosis from those without fibrosis using Fourier transform infrared (FTIR) spectroscopy of serum. Serum samples of 12 patients with no hepatic fibrosis and 11 patients with extensive fibrosis were used.	Fourier transform infrared spectral profiles.
5.	Fast detection of tobacco mosaic virus infected tobacco using laser induced breakdown spectroscopy [56]	Detect TMV-infected tobacco based on laser-induced breakdown spectroscopy (LIBS).	Full spectrum and observed emission lines of laser-induced breakdown spectroscopy (LIBS) for fresh & dried leaves.
6.	Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM) [54]	Classification of dengue suspected human blood sera; use of Raman spectroscopy combined by deciphering spectral differences between dengue positive and normal sera. Raman spectra of 31 were dengue positive and 53 were negative were used.	Features obtained from Raman Spectra

as well as ever increasing immune system data and databases, epitopes can be identified in few years. Once the epitopes are predicted using computational methods, the peptides can be experimentally tested for its binding affinity and ability to elicit desired immune response. Immunoinformatics involves the development of bioinformatics tools that analyses data to predict B- and T-cell epitopes which can stimulate immune response. In-silico prediction methods of epitope prediction can be beneficial to decrease the number of potential epitopes for experimental confirmation, develop epitope-based vaccines for hypervariable viruses and develop chimeric vaccines [59, 62]. Epitope based-vaccines can be safer and less expensive than conventional methods [62].

Predicted epitopes should take into account the desirable features of epitopes such as they should be conserved in different parts of viral lifecycle, their binding affinity and efficacy, they should bind to more than one allele of immune system molecules and most of them are proteins [59, 62]. Most epitope prediction methods are based on proteins and their different descriptors including physicochemical properties related profiles of proteins, evolutionary data, sequence motifs and quantitative matrices (QM) [58, 59]. SVM has been one of the most popular methods used for both B-cell & T-cell epitope prediction.

9.4.1 T-Cell Epitope Prediction

T-cell epitopes are processed within a cell, linked with MHC & presented on T-cell surface to be recognized by T-cell receptor. Each of these steps decide the immunogenicity of T-cell epitopes. However, most of the T-cell epitopes focus on the step where a peptide is linked with MHC-I & MHC-II [59]. MHC-1 binds to peptides of length 9–11 AA s and its pockets prefers peptides with certain physicochemical properties. Hence, peptide-MHC-I binding prediction methods work on peptide sequences of 9 AA residues. On the other hand, MHC-II binds to longer peptides but the prediction methods focus on peptide part that binds to the MHC-II groove. Large number of databases like IEDB, EPIMHC and AntiJen, store epitopes verified through experimental approaches [59]. These have served as rich sources of positive examples for several prediction methods.

Different computational methods/models have been used to predict epitopes like use of Sequence Motif, motif matrix, quantitative affinity matrices (QAM) etc. . However, machine learning (ML) methods have proven to be the most robust method for prediction [63]. With high dimensionality of the data and the limited observations, SVM comes as a better method. In a study, 36 stimulatory peptides and 167 non-stimulatory peptides were gathered, and physical properties of 20 AA s were used to develop models from Artificial Neural Network, Decision Tree & SVM. SVM proved to outperform prediction of stimulatory peptides with maximum sensitivity of 0.76 [64].

MHC2Pred is one of the freely available tools based on SVM to predict MHC-II binding peptides [65]. To develop a model for MHC2Pred, binding & non-binding peptides, based on IC50, were collected from MHCBN and JenPep database.

Peptides with less than 9 AA residues were discarded and rest of the peptides were looked for 9 AA s that would bind the MHC-II groove using Matrix Optimization Techniques (MOT) package. A vector of length 20 was created for each AA in 9-mer peptide where binders were given +1 and non-binders a −1. Each peptide was thus represented by 180 (9 × 20) length vectors. This data was used to develop SVM model which was later validated using fivefold cross validation and got an overall accuracy of method is >78% [65].

SVMHC is another tool for prediction of both MHC class I and class II binding peptides [66]. For MHC-I prediction model, peptides of length 8–10 were represented by a binary sparse encoding. For MHC-II peptide binding prediction, matrices by Sturniolo et al. [67] were used. These matrices represent HLA-DR peptide binding specificity where HLA-DR is an MHC-II cell surface receptor [67] (see sr. no. 1 of Table 6).

Predicting immunogenicity of epitopes can help in vaccine design and POPISK is a tool that predicts reactivity of T-cells to peptides and identify positions that are recognized by TCR [68]. POPISK uses SVM model with a weighted degree string kernel (see sr. no. 2 of Table 6).

9.4.2 B-Cell Epitope Prediction

B-cell epitopes can be predicted based on physicochemical properties like hydrophilicity, flexibility, polarity, and exposed surface as well as secondary & 3D structures [62]. There are 566 AA indices that represents physicochemical properties of AA s listed in AAindex [69].

Linear epitopes can be predicted using antigen sequences by calculating AA propensity scales based on physicochemical properties. AA Propensities (AAP) calculation considers an overlapping window of length k AA s in a protein sequence and for each window, average propensity value of AA s is calculated, where propensity value can be hydrophilicity, accessibility, flexibility, polarity, antigenicity, beta-turn, surface exposed scale, etc. The average value is assigned to the AA in middle of the window. AA s residues that passes the threshold are considered as potential epitopes. A combination of different propensity values can be used with specific weights [70].

Due to poor performance of AA propensity scales, Machine learning (ML) methods were later adopted to distinguish B-cell epitopes from non-epitopes. BCPREDS and SVMtrip [71] are epitope prediction tools based on Support Vector Machine (SVM) [59]. More information on SVMtrip is provided in sr. no. 3 of Table 6.

Conformational B-cell epitopes can be predicted using features related to the structure of the proteins. One of the studies have used combination of physicochemical features, evolutionary PSSM features and structural features as protrusion index (PI), accessible surface area (ASA), relative accessible surface area (RSA) and B-factor [72] (see sr. no. 4 of Table 6). Physicochemical properties of AA s were derived from AAIndex. PSSM represents the attributes extracted from repeated multiple sequence alignment of sequences that can be generated using PSI-BLAST

with specific number of iterations. It is a scoring matrix where each position in the multiple sequence alignment is given an AA substitution scores. PSSM is used to incorporate evolutionary information of a peptide [73–75]. Another study by Ansari et al. [76] on conformational B-cell epitope uses 3 types of features namely binary profile of pattern (BPP), physiochemical profile of patterns (PPP) and composition profile of patterns (CPP) (see sr. no. 5 of Table 6). In this study, patterns of different lengths were created from the sequences. Then for each pattern 3 feature vectors were created, (1) BPP, a vector of length 21 based on binary number for occurrence and non-occurrence of AA, (2) PPP, a vector of length 5 based on 5 physicochemical properties named Hydrophobicity, Flexibility, Polarity_Grantham, Polarity_Ponnuswami, Antigenicity and (3) CPP based on composition of patterns. CBTope server uses this method for predicting B-cell epitopes [76].

Listed in Table 5 are some freely available T-cell & B-cell epitope prediction web servers based on SVM.

Information on some more SVM based epitope prediction studies have been provided in Table 6.

9.5 *Applications of SVM Involving Protein-Protein Interaction in Virology*

Proteins are the workhorses of a cell that carry out majority of the functions in a cell. Eighty percent of proteins are not functional in isolated forms but they operate in complexes by interacting with other molecules [77, 78]. Protein-protein interaction (PPI) is the physical & functional interactions of proteins that controls wide range of molecular processes in a cell, like signal transduction, cell-cell communication, transcription, replications etc. [79]. PPIs can be responsible for altering kinetic properties of enzyme, modifying proteins activity, changing specificity of protein binding, constructing new binding sites and regulatory function. Alteration or malfunction of these interactions can lead to diseases [79]. The collection of all the protein-protein interaction of cell or an organism is called interactome. The study of PPIs can help in predicting a biological process involving protein of unknown function, fasten the pace of understanding functional pathways or to know biochemistry

Table 5 List of freely available epitope prediction servers

Sr. no.	Server	Reference link	Epitope predicted
1.	MHC2Pred	http://www.imtech.res.in/raghava/mhc2pred/	T-cell epitope
2.	SVMHC	http://abi.inf.uni-tuebingen.de/Services/SVMHC	T-cell epitope
3.	SVRMHC	http://svrmhc.biolead.org/	T-cell epitope
4.	BCPRED	http://ailab.ist.psu.edu/bcpred/	B-cell epitope
5.	SVMTriP	http://sysbio.unl.edu/SVMTriP/prediction.php	B-cell epitope
6.	EPSVR	http://sysbio.unl.edu/EPSVR/	B-cell epitope
7.	CBTOPE	http://crdd.osdd.net/raghava/cbtope/	B-cell epitope

Table 6 Illustrative examples of epitope prediction based on SVM

Sr. no	Reference	Brief description of work	Attributes and attribute selection
1.	SVMHC: A server for prediction of MHC-binding peptides [66]	**Purpose:** Identification of MHC-I and MHC-II binding peptides **Dataset:** MHC-binding peptides of different lengths were extracted from the MHCPEP and SYFPEITHI databases.	For MHC-I – Binary sparse encoding of 8–10 k-mer length of AA s For MHC-II – Matrices representing HLA-DR peptide binding specificity
2.	POPISK: T-cell reactivity prediction using support vector machines and string kernels [68]	**Purpose:** Predict immunogenicity of peptides by predicting T-cell reactivity i.e. if a peptide is immunogenic or non-immunogenic using SVM with a weighted degree string kernel. **Dataset:** Extracted peptide binders of length 9 along with their associated human MHC class I alleles and immunogenicity from three databases, MHCPEP, SYFPEITHI and IEDB. Negatively annotated peptides were used as non-immunogenic peptides Final dataset – 558 immunogenic and 527 non-immunogenic peptides	Matched sub-sequences of length p at a position in 2 sequences
3.	SVMTriP A Method to Predict Antigenic Epitopes Using Support Vector Machine [71]	**Purpose:** Predict linear B-cell epitopes using SVM with RBF Kernel **Dataset:** Dataset constructed by extracting non-redundant linear B-cell epitopes (10AA, 12AA, 14AA, 16AA, 18AA, and 20AA) from IEDB. For negative dataset, non-epitope part of corresponding antigen used. Final dataset: 4925 non-redundant epitope sequences each for positive and negative dataset.	Tripeptide similarity using Blosum62 matrix and propensity scores
4.	Positive-unlabeled learning for the prediction of conformational B-cell epitopes [73]	**Purpose:** PUPre (Positive-Unlabeled Prediction) method to – (1) identify non-epitope residues using weighted SVM and (2) model to distinguish epitope and non-epitope residues **Dataset:** 2123 residues labelled as epitopes and 16,615 unlabeled residues by processing data from PDB	Feature vector of 239 features including 205 physico-chemical features collected from AAIndex 21 evolutionary PSSM features 13 structural features. **Attribute selection:** Wilcoxon rank-sum test was applied to select informative features and resulted in 89 selected features

(continued)

Table 6 (continued)

Sr. no	Reference	Brief description of work	Attributes and attribute selection
5.	Identification of conformational B-cell epitopes in an antigen from its primary sequence [76]	**Purpose:** Use SVM with RBF Kernel to identify conformational B-cell epitopes **Dataset:** 187 antigenic protein chains having 2261 amino acid residues that were antibody interacting and 107,414 amino acid residues as non-antibody interacting	Binary profile of patterns (BPP), Physico-chemical profile of patterns (PPP), composition profile of patterns (CPP) Explanation of features in Sect. 9.4.2
6.	SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction [100]	**Purpose:** Identification of linear B-cell epitopes using SVM string kernel prediction model **Dataset:** Linear B-cell epitopes of lengths 12- to 20-mers extracted from EL-Manzalawy dataset [101]	Sequences encoded in bi-profile manner where they have attributes from 2 pools – positive position-specific and negative position-specific profiles **Feature extraction:** Bayes Feature Extraction (BFE)
7.	Application of support vector machines for T-cell epitopes prediction [64]	**Purpose:** T-cell epitope prediction with an MHC I restricted T-cell clone. **Dataset:** 36 stimulatory peptides and 167 non-stimulatory peptides which were further divided into positive and negative set by random sampling	188 physical properties of 20 AA s **Attribute selection:** 'Ten factors extraction from 188 physical properties of 20 AA s

of a cell [77, 79, 80]. Knowledge of specific PPI can also help in identification of drug targets [79].

PPI data can be mapped to large scale networks where nodes represent proteins and edges represent their physical or functional interactions. These networks are known as PPI networks (PIN) [77, 79]. PPI networks can be used to extract various information like functionality of a protein based on its placement in the network as the closely linked proteins can have similar biological activity. PPI can also be used to decipher which complex a protein belongs to and the diseases related to a protein [79]. The knowledge that is encapsulated in the PPI can help improve the biological and biomedical applications [77].

Virus-host proteins interactions are key to viral infection and subsequent pathogenesis. Many PPIs are involved between virus and host during a viral infection where the virus proteins take over the host transcriptional machinery [78]. It has been believed that viral proteins bind to the host protein that are highly connected [81]. Endogenous interface, with respect to virus-host systems, are responsible for interactions in their own system i.e. host-host PPI and virus-virus PPI. On the other hand, exogenous interfaces are responsible of virus-host interactions. Both virus and host compete for endogenous and exogenous interfaces [81]. Mutations at protein interfaces can reduce or increase their binding affinities by changing protein electrostatics and structural properties. Virus and host proteins change their surface

resides through mutations as an evolutionary result to compete for binding partner. However, host tends to be less variable than viruses. Viruses diversify through various modes of molecular evolution, including conservation, horizontal gene transfer, gene duplication and molecular mimicry [81]. Viral proteins constantly inhibit host-host interactions and therefore, blocking such interactions between virus & host can aid in biomedical applications by identification of drug targets and developing antiviral therapies [81]. For e.g. a drug, Maraviroc, binds the cellular co-receptor CCR5, a receptor on white blood cells involved in immune system, preventing it from interacting with GP120 of HIV1 which is essential for entry of HIV-1 in host [82]. As viruses pose a global threat, understanding of virus & human PPIs can help in development of vaccines for treatment.

Comprehensive PPI networks have been generated using experimental methods. These experimental methods employ different techniques like tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display, X-ray crystallography, and NMR spectroscopy [79]. However, due to the huge PPI data and the time consuming experimental methods, computational methods are increasingly becoming popular to analyse the PPI networks and find out the functions of unexplored proteins. Computational methods of PPI detection are based on sequences, structure of molecules, gene fusion, phylogenetic tree and gene expression [79].

Detection of virus-host interactions using machine learning methods have proved to be very useful. Several SVM models have been developed for the same purpose; known PPIs as positive set, are used to train the models to predict whether two proteins interact or not. Positive set data can be extracted from experimental data available in the databases. Selecting negative dataset is complicated. Negatome, a database of negative interactions developed using text mining, can be used to gather negative data set [83, 84].

Emamjomeh et al. [85], developed SVM model to predict PPI interactions between human and hepatitis C virus (HCV) [D32]. In this study, SVM was combined with other learning methods like random forest (RF), Naïve Bayes (NB) and multilayer perceptron (MLP) Feature vectors were generated for HCV & human proteins which included six different AA composition (ACC), pseudo AA composition (PAC), PSSM as evolutionary information feature, network centrality measures, tissue information and post-translational modification (PTM) information [85]. AA composition is the simplest descriptor used to represent a protein sequence. However, with this descriptor the sequence order of AA s is lost and hence, pseudo AA is used which involves AA composition as well as sequence order-based features [5] (see sr. no. 1 of Table 7).

Cui et al. [86] developed an SVM model for prediction of virus-host PPI for 2 viruses, human papillomaviruses (HPV) and hepatitis C virus (HCV). This SVM model is based on relative frequency of AA triplets (RFAT) between virus & host AA sequences and GO annotations of protein. RFAT generates fixed length for variable length proteins and enables models to achieve a better accuracy. In this study, a vector based on AA triplets & biochemical similarity is generated. Based on biochemical properties of AA residues, 6 categories are defined as {IVLM}, {FYW}, {HKR}, {DE}, {QNTP}, and {ACGS}. Using this classification of AA s,

Table 7 Illustrative examples Protein-Protein interaction studies based on SVM

Sr. no	Reference	Brief description of work	Attributes and attribute selection
1.	Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method [85]	**Purpose:** Predicting PPI between human and hepatitis C virus using SVM combined with RF, NB and MLP. **Dataset:** 657 positive interactions from human-HCV PPI from IntAct database and 2910 negative interactions	AA composition, pseudo AA composition, PSSM, network centrality feature, tissue information feature, 31 PTM types
2.	Prediction of protein-protein interactions between viruses and human by an SVM model [86]	**Purpose:** Prediction of protein- protein interactions using SVM binary classifier. **Dataset:** Training dataset had 500 positive and negative interactions. Test set had 195positive and negative interactions. Positive dataset was extracted from the infection mapping project (I-MAP) whereas negative from HPRD by random selection of human proteins	Feature vector of relative frequency of 216 AA triplets Details of attribute Sect. 9.5
3.	An improved method of predicting interactions between virus [89]	**Purpose:** An improved method of predicting interactions between virus and proteins including human papillomaviruses (HPV) and hepatitis C virus (HCV), using SVM with RBF kernel.	Features used: Relative frequency of AA triplets (RFAT), the frequency difference of AA triplets (FDAT) between virus and host proteins, and AA composition (AC). RFAT feature generation-clustered 20 AA s into 4 groups based on chemical properties of side chain of the AA s yields 64 AA triplets
4.	A generalized approach to predicting PPI [78]	**Purpose:** Prediction of PPI between virus and host using SVM with RBF kernel. Additionally, a generic model to predict PPI of any virus & host **Dataset:** Multiple training and test datasets used	Features used – RFAT, FDAT, AC, normalized frequency of each AA group, transition & distribution. RFAT feature generation – 20 AAs into 7 groups based on dipoles & volumes of side chains of AA s yielding 343 possible AA triplets. A vector of 343 + 343 = 686 was generated for virus-host pair

(continued)

Table 7 (continued)

Sr. no	Reference	Brief description of work	Attributes and attribute selection
5.	Supervised learning and prediction of physical interactions between human and HIV proteins [88]	**Purpose:** Prediction of human-HIV PPI using SVM with a linear kernel. **Dataset:** 1028 human–HIV PPIs from four public databases, Biomolecular Interaction Network Database, the Database of Interacting Proteins, IntAct, and Reactome Negative dataset was generated by randomly pairing human and HIV proteins.	Four-mers sequence, protein domains responsible for interactions and PPI network information. Four-mers sequence and 7 categories of AA, were used to generate feature vector of 4802 (7^4 ∗ 2)

there are 6 × 6 × 6 = 216 possible AA triplets [86]. The protein sequence is converted into AA triplets and the vector of 216 length is created that contains the frequency of each category in sequences of variable length. LIBSVM [87] was used to generate model with the radial basis function (RBF) as a kernel function. For dataset, HCV & human interaction data was extracted from the infection mapping project (I-MAP) whereas for HPV, data was extracted from NCBI Bio Systems Database. For HCV accuracy of 85.1 was achieved whereas for HPV it was 87.5 [86] (see sr. no. 2 of Table 7).

RFAT has been used in many studies with different combinations of categories and k-mer. In a study of HIV and human PPI [88], four-mer sequences were used instead of triplet. With 7 categories and 4-mer sequences, RFAT vector of 4802 (7^4 ∗ 2) length was generated (see sr. no. 5 of Table 7).

Kim et al. [89] used 4 categories based on chemical properties of side chain of the AA s making 64 AA triplets combination (see sr. no. 3 of Table 7).

In another study of PPI by Zhou et al. [78] and Shen et al. [90], a similar feature vector of triplets is produced but 7 categories of AA residues are used instead of 6 and these categories are based on diploes and volumes of the side chains of AA s. With 7 categories 343 (7 × 7 × 7) AA triplets are possible. RFAT feature vector had 686 elements i.e. 343 for host and 343 for virus. Zhou et al. [78] uses more features as frequency difference of AA triplets (FDAT) between virus and host proteins, AA composition (AC) in each pair of host and virus proteins, normalized frequency of each AA group, transition and distribution of AA groups. As a result of these 6 features, a feature vector of length 1175 was created. Again, LIBSVM [87] with RBF was used to develop model. Best performance was obtained with combination of all these 6 features with accuracy of 85.64% (see sr. no. 4 of Table 7).

Most of the prediction methods are specific to a virus-host combination. However, there are SVM based methods that are generic enough to predict PPIs of virus and host that were not used for training set. The approach by Zhou et al. [78] is one of such methods i.e. it does not require model for each host-virus pair. Another method called DeNovo, is a generic method that can predict novel PPIs. This method is based on SVM that trains on different virus-host PPIs [91].

Table 7 shows some studies on SVM model that are used for protein-protein interaction of virus-host.

10 Miscellaneous Examples

Apart from above examples, there are some noticeable studies employing other approaches to address problems in virology. Microarray is a method that uses microscopic chip where each spot-on chip has a DNA/cDNA sequence attached. These sequences bind to the complementary unknown sequences & thereby detects gene expressions of thousands of genes. In Virology, Microarray is used to screen viruses for which genomes are available in GenBank by looking at the conserved viral sequences. Microarray gene expression profiles are also used to detect the immune response that can further help in classifying disease caused by viruses, that is conventionally done using quantitative real time PCR (qPCR). SVM can be used to detect immune response by using microarray gene expression data. Due to big size of microarray data, important features are extracted using feature selection methods.

In a study [92], the authors have reported that DNA microarray technology can be used as a high-throughput method to analyse polymorphisms within a short region of the FMDV genome encoding relevant functions in antigenicity and receptor recognition. Their SVM based methodology classifies the samples based on their hybridization signal. This prediction methodology has wide ranging applications to fine genotyping including studies of heterogeneous viral populations, genetic changes in virus, bacteria, and genes of rapidly evolving cells, such as tumor cells.

Predicting the hosts of newly discovered viruses is important for pandemic surveillance of infectious diseases. Li and SUN [93] investigated the use of alignment-based and alignment-free methods and support vector machine using mononucleotide frequency and dinucleotide bias to predict the hosts of viruses, and applied these approaches to three datasets: rabies virus, coronavirus, and influenza A virus [93] also showed that SVM predicts the hosts of viruses with a high degree of accuracy.

The phosphorylation of virus proteins by host kinases is linked to viral replication leading to an inhibition of normal host-cell functions. Unravelling of phosphorylation mechanisms in virus proteins can aid in drug design and treatment. In this study [94] a two-layered Support Vector Machines (SVMs) was applied to train a predictive model for identification of phosphorylation sites.

Replication of their DNA genomes is a central step in the reproduction of many viruses. [V4] proposes a novel least-squares support vector machines (LS-SVMs) model with viruses of herpes family along with data sets involving a collection of caudoviruses coming from three viral families under the order of caudovirales. The LS-SVM approach provides superior performance as compared to those given by the previous methods. Ensembled with previously proposed methods, the LS-SVM approach further improves the prediction accuracy for the herpesvirus replication origins. Recursive feature elimination was used to extract the most informative attri-

butes and provides important domain knowledge in terms of the most significant features of the data sets [95] further conclude LS-SVMs can potentially be a very reliable and robust tool for viral replication origin prediction.

11 Web Server

SVM has been used in a variety of studies on viruses across different data types. Some of the tools mentioned in these studies are available as standalone tools whereas others are used in the backend of freely available web-servers. Web servers are user friendly and more intuitive making it easy for user to input data and analyse the output. Table 8 shows some of the web servers based on SVM models that are used in virology.

Table 8 Examples of SVM based web servers SVM for Virology Studies

Sr. no	Reference	Brief description of work	Attributes and attribute selection
1.	A genotypic method for determining HIV-2 coreceptor usage enables epidemiological studies and clinical decision support [99]	**Purpose:** Geno2pheno is a web service to ensure that the virus can use only the CCR5 coreceptor (R5) and cannot evade the drug by using the CXCR4 coreceptor (X4-capable) using V3 loop of the HIV-2 glycoprotein **Link:** https://www.geno2pheno.org/ **Dataset:** To build model, 126 pairs of HIV-2 amino-acid sequences and phenotypic coreceptor usage as R5 or X4-capable	V3 loop region of the HIV-2 glycoprotein
2.	AVCpred: An integrated web server for prediction and design of antiviral compounds [102]	**Purpose:** AVCpred is a web server for prediction of antiviral compounds (AVC) for HIV, HCV, HBV, HHV & 26 other viruses with QSAR-based model **Link:** http://crdd.osdd.net/servers/avcpred **Dataset:** Antiviral compounds extracted from ChEMBL bioactivity database – 389 compounds for HIV, 467 for HCV, 124 for HHV, 112 for HBV, and 1391 for other 26 viruses	18,000 chemical descriptors (1D, 2D, and 3D) using PaDEL **Attribute selection:** Filter named 'RemoveUseless' followed by ClassifierSubsetEval (attribute evaluator) with BestFirst (search method) module available in Weka package

(continued)

Table 7 (continued)

Sr. no	Reference	Brief description of work	Attributes and attribute selection
3.	Prediction of linear B-cell epitopes of hepatitis C virus for vaccine development [75]	List of web servers based on SVM for B-cell epitope prediction: BCPred: http://ailab.ist.psu.edu/bcpred/ SVMTriP: http://sysbio.unl.edu/SVMTriP/prediction.php Bcell-HCV: http://e045.life.nctu.edu.tw/BcellHCV	Features used for prediction: BCPred: AAP propensity SVMTriP: Tri-peptide Bcell-HCV: Physicochemical properties
4.	SVMHC: a server for prediction of MHC-binding peptides [66]	**Purpose:** Identification of MHC-binding peptides **Link:** http://abi.inf.uni-tuebingen.de/Services/SVMHC **Dataset:** MHC-binding peptides extracted from the MHCPEP and SYFPEITHI databases of varying length.	Binary sparse encoding of 8–10 k-mer length of AAs

12 Concluding Remarks

In this review, we illustrated the use of Support Vector Machines as a tool for building learning models in viral biology. SVM plays a vital role in building Quantitative structure activity relationship models. The robustness and accuracy of SVM models based rigorously on statistical learning theory has paved the way for quicker, faster and reliable methods of identification of potent molecules in drug design. SVM models have also enabled development of tools for rational design of novel vaccines. Recent advances in NGS technology could also be easily incorporated with SVM for building models with increased performance. We have also listed large number of case studies and examples in different areas of viral biology where SVM has been deployed with productive results

References

1. Solomatine DP. Data-driven modelling: paradigm, methods, experiences. In: Proceedings of the 5th international conference on hydroinformatics; 2002 July 1. p. 1–5.
2. Mika S, Schölkopf B, Smola AJ, Müller KR, Scholz M, Rätsch G. Kernel PCA and de-noising in feature spaces. In: Advances in neural information processing systems; 1999. p. 536–542.
3. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988–99.

4. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other Kernel-based learning methods. Cambridge: Cambridge university press; 2000.
5. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics. 2009;6(4):262–74.
6. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
7. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Brief Bioinform. 2008;9(5):392–403.
8. Srivastava A, Ghosh S, Anantharaman N, Jayaraman VK. Hybrid biogeography based simultaneous feature selection and MHC class I peptide binding prediction using support vector machines and random forests. J Immunol Methods. 2013;387(1–2):284–92.
9. Weston J, Watkins C. Support vector machines for multi-class pattern recognition. In: Esann 1999 April 21, vol 99. p. 219–224.
10. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999;9(3):293–300.
11. Tax DM, Duin RP. Support vector domain description. Pattern Recogn Lett. 1999;20(11–13):1191–9.
12. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. Neural computation. 2000 May 1;12(5):1207–45.
13. Dudek AZ, Arodz T, Gálvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb Chem High Throughput Screen. 2006;9(3):213–28.
14. http://www.unaids.org/en/resources/campaigns/HowAIDSchangedeverything/factsheet.
15. Qureshi A, Rajput A, Kaur G, Kumar M. HIVprotI: an integrated web based platform for prediction and design of HIV proteins inhibitors. J Chem. 2018;10(1):12.
16. http://bioinfo.imtech.res.in/manojk/hivproti.
17. Hdoufane I, Bjij I, Soliman M, Tadjer A, Villemin D, Bogdanov J, Cherqaoui D. In silico SAR studies of HIV-1 inhibitors. Pharmaceuticals. 2018;11(3):69.
18. Mei J, Zhao J. Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. Sci Rep. 2018;8(1):2359.
19. Richardson AM, Lidbury BA. Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines. BMC Med Inform Decis Mak. 2017;17(1):121.
20. Qasim MK, Algamal ZY, Ali HM. A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine. SAR QSAR Environ Res. 2018;29(7):517–27.
21. Huang CC, Chang CC, Chen CW, Ho SY, Chang HP, Chu YW. PClass: protein quaternary structure classification by using bootstrapping strategy as model selection. Genes. 2018;9(2):91.
22. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31.
23. Ansorge WJ. Next-generation DNA sequencing techniques. New Biotechnol. 2009;25(4):195–203.
24. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. Microbiology in the post-genomic era. Nat Rev Microbiol. 2008;6(6):419.
25. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135.
26. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol. 2009;7(4):287.
27. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, Martinson V. A metagenomic survey of microbes in honey bee colony collapse disorder. Science. 2007;318(5848):283–7.
28. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF. A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med. 2008;358(10):991–8.

29. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, Quan PL, Lipkin WI, Downing R, Tappero JW, Okware S. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog. 2008;4(11):e1000212.
30. Wong K, Fong TT, Bibby K, Molina M. Application of enteric viruses for fecal pollution source tracking in environmental waters. Environ Int. 2012;45:151–64.
31. Bibby K, Viau E, Peccia J. Viral metagenome analysis to guide human pathogen monitoring in environmental samples. Lett Appl Microbiol. 2011;52(4):386–92.
32. Ge X, Li Y, Yang X, Zhang H, Zhou P, Zhang Y, Shi Z. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. J Virol. 2012;86(8):4620–30.
33. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One. 2009;4(1):e4219.
34. de Vries M, Deijs M, Canuti M, van Schaik BD, Faria NR, van de Garde MD, Jachimowski LC, Jebbink MF, Jakobs M, Luyf AC, Coenjaerts FE. A sensitive assay for virus discovery in respiratory clinical samples. PLoS One. 2011;6(1):e16118.
35. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis. 2012;6(2):e1485.
36. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe. 2010;7(6):509–15.
37. Woo HJ, Reifman J. A quantitative quasispecies theory-based model of virus escape mutation under immune selection. Proc Natl Acad Sci. 2012;109(32):12980–5.
38. Bartolini B, Chillemi G, Abbate I, Bruselles A, Rozera G, Castrignanò T, Paoletti D, Picardi E, Desideri A, Pesole G, Capobianchi MR. Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing. Microbiol Q J Microbiol Sci. 2011;34(4):391.
39. Selleri M, Piralla A, Rozera G, Giombini E, Bartolini B, Abbate I, Campanini G, Rovida F, Dossena L, Capobianchi MR, Baldanti F. Detection of haemagglutinin D222 polymorphisms in influenza A (H1N1) pdm09-infected patients by ultra-deep pyrosequencing. Clin Microbiol Infect. 2013;19(7):668–73.
40. Jia JA, Liang X, Chen S, Wang H, Li H, Fang M, Bai X, Wang Z, Wang M, Zhu S, Sun F. Next-generation sequencing revealed divergence in deletions of the preS region in the HBV genome between different HBV-related liver diseases. J Gen Virol. 2017;98(11):2748–58.
41. Bai X, Jia JA, Fang M, Chen S, Liang X, Zhu S, Zhang S, Feng J, Sun F, Gao C. Deep sequencing of HBV pre-S region reveals high heterogeneity of HBV genotypes and associations of word pattern frequencies with HCC. PLoS Genet. 2018;14(2):e1007206.
42. Xu B, Tan Z, Li K, Jiang T, Peng Y. Predicting the host of influenza viruses based on the word vector. PeerJ. 2017;5:e3579.
43. Wokaun A, Schrader B. Infrared and Raman spectroscopy-methods and applications. VCH, Weinheim; 1995, DM 298,-, ISBN 3-527-26446-9. Berichte der Bunsengesellschaft für physikalische Chemie. 1996;100(7):1268-.
44. Gremlich HU, Yan B. Infrared and Raman spectroscopy of biological materials. Boca Raton: CRC Press; 2000.
45. Wartewig S, Neubert RH. Pharmaceutical applications of Mid-IR and Raman spectroscopy. Adv Drug Deliv Rev. 2005;57(8):1144–70.
46. Vandenabeele P. Practical Raman spectroscopy: an introduction. Chichester, United Kingdom: Wiley; 2013 Jul 3.
47. Blanch EW, Hecht L, Barron LD. Vibrational Raman optical activity of proteins, nucleic acids, and viruses. Methods. 2003;29(2):196–209.

48. Tsuboi M, Kubo Y, Ikeda T, Overman SA, Osman O, Thomas GJ. Protein and DNA residue orientations in the filamentous virus Pf1 determined by polarized Raman and polarized FTIR spectroscopy. Biochemistry. 2003;42(4):940–50.
49. Benevides JM, Juuti JT, Tuma R, Bamford DH, Thomas GJ. Characterization of subunit-specific interactions in a double-stranded RNA virus: Raman difference spectroscopy of the φ6 procapsid. Biochemistry. 2002;41(40):11946–53.
50. Tuma R, Thomas GJ Jr. Mechanisms of virus assembly probed by Raman spectroscopy: the icosahedral bacteriophage P22. Biophys Chem. 1997;68(1–3):17–31.
51. Radzol AR, Lee KY, Mansor W, Omar IS. PCA criterion for SVM (MLP) classifier for flavivirus biomarker from salivary SERS spectra at febrile stage. In: 2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016 August 16, p. 6206–6209. IEEE.
52. Khan S, Ullah R, Khan A, Ashraf R, Ali H, Bilal M, Saleem M. Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. Photodiagn Photodyn Ther. 2018;23:89–93.
53. Scaglia E, Sockalingum GD, Schmitt J, Gobinet C, Schneider N, Manfait M, Thiéfin G. Noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis C using serum Fourier transform infrared spectroscopy. Anal Bioanal Chem. 2011;401(9):2919.
54. Khan S, Ullah R, Khan A, Wahab N, Bilal M, Ahmed M. Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). Biomed Opt Express. 2016;7(6):2249–56.
55. Lee H, Kim MS, Lim HS, Park E, Lee WH, Cho BK. Detection of cucumber green mottle mosaic virus-infected watermelon seeds using a near-infrared (NIR) hyperspectral imaging system: application to seeds of the "Sambok Honey" cultivar. Biosyst Eng. 2016;148:138–47.
56. Peng J, Song K, Zhu H, Kong W, Liu F, Shen T, He Y. Fast detection of tobacco mosaic virus infected tobacco using laser-induced breakdown spectroscopy. Sci Rep. 2017;7:44551.
57. Liang TC. Epitopes. https://www.sciencedirect.com/topics/immunology-and-microbiology/epitope.
58. Desai DV, Kulkarni-Kale U. T-cell epitope prediction methods: an overview. In: Immunoinformatics. New York: Humana Press; 2014. p. 333–64.
59. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T-and B-cell epitope prediction. J Immunol Res. 2017;2017:1–14.
60. Mukonyora M. A review of important discontinuous B-cell epitope prediction tools. J Clin Cell Immunol. 2015;6:358–62.
61. Genetics Home Reference. Human leukocyte antigens. https://ghr.nlm.nih.gov/primer/genefamily/hla.
62. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J Biomed Inform. 2015;53:405–14.
63. Backert L, Kohlbacher O. Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med. 2015;7(1):119.
64. Zhao Y, Pinilla C, Valmori D, Martin R, Simon R. Application of support vector machines for T-cell epitopes prediction. Bioinformatics. 2003;19(15):1978–84.
65. MHC2Pred: SVM based method for prediction of promiscuous MHC Class II binders. http://crdd.osdd.net/raghava/mhc2pred/info.html.
66. Dönnes P, Kohlbacher O. SVMHC: a server for prediction of MHC-binding peptides. Nucleic Acids Res. 2006;34(suppl_2):W194–7.
67. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol. 1999;17(6):555.
68. Tung CW, Ziehm M, Kämper A, Kohlbacher O, Ho SY. POPISK: T-cell reactivity prediction using support vector machines and string kernels. BMC Bioinf. 2011;12(1):446.

69. Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 1999;27(1):368–9.
70. Su CH, Pal NR, Lin KL, Chung IF. Identification of amino acid propensities that are strong determinants of linear B-cell epitope using neural networks. PLoS One. 2012;7(2):e30617.
71. Yao B, Zhang L, Liang S, Zhang C. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PLoS One. 2012;7(9):e45152.
72. Bhagwat M, Aravind L. Psi-blast tutorial. In: Comparative genomics. Totowa: Humana Press; 2007. p. 177–86.
73. Ren J, Liu Q, Ellis J, Li J. Positive-unlabeled learning for the prediction of conformational B-cell epitopes. BMC Bioinf. 2015;16(18):S12.
74. PSSM Viewer. https://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi.
75. Huang WL, Tsai MJ, Hsu KT, Wang JR, Chen YH, Ho SY. Prediction of linear B-cell epitopes of hepatitis C virus for vaccine development. BMC Med Genet. 2015;8(4):S3.
76. Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res. 2010;6(1):6.
77. Safari-Alighiarloo N, Taghizadeh M, Rezaei-Tavirani M, Goliaei B, Peyvandi AA. Protein-protein interaction networks (PPI) and complex diseases. Gastroenterol Hepatol Bed Bench. 2014;7(1):17.
78. Zhou X, Park B, Choi D, Han K. A generalized approach to predicting protein-protein interactions between virus and host. BMC Genomics. 2018;19(6):165.
79. Rao VS, Srinivas K, Sujini GN, Kumar GN. Protein-protein interaction detection: methods and analysis. Int J Proteomics. 2014;2014:147648.
80. Gonzalez MW, Kann MG. Protein interactions and disease. PLoS Comput Biol. 2012;8(12):e1002819.
81. Brito AF, Pinney JW. Protein–protein interactions in virus–host systems. Front Microbiol. 2017;8:1557.
82. MacArthur RD, Novak RM. Maraviroc: the first of a new class of antiretroviral agents. Clin Infect Dis. 2008;47(2):236–41.
83. Mei S, Zhu H. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. Sci Rep. 2015;5:8034.
84. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Res. 2013;42(D1):D396–400.
85. Emamjomeh A, Goliaei B, Zahiri J, Ebrahimpour R. Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method. Mol BioSyst. 2014;10(12):3147–54.
86. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. BMC Bioinf. 2012;13(7):S5. BioMed Central.
87. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST). 2011;2(3):27.
88. Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. Infect Genet Evol. 2011;11(5):917–23.
89. Kim B, Alguwaizani S, Zhou X, Huang DS, Park B, Han K. An improved method for predicting interactions between virus and human proteins. J Bioinforma Comput Biol. 2017;15(01):1650024.
90. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci. 2007;104(11):4337–41.
91. Eid FE, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics. 2015;32(8):1144–50.
92. Martín V, Perales C, Abia D, Ortíz AR, Domingo E, Briones C. Microarray-based identification of antigenic variants of foot-and-mouth disease virus: a bioinformatics quality assessment. BMC Genomics. 2006;7(1):117.

93. Li H, Sun F. Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. Sci Rep. 2018;8(1):10032.
94. Huang KY, Lu CT, Bretaña NA, Lee TY, Chang TH. ViralPhos: incorporating a recursively statistical method to predict phosphorylation sites on virus proteins. BMC Bioinf. 2013;14(16):S10.
95. Cruz-Cano R, Chew DS, Choi KP, Leung MY. Least-squares support vector machine approach to viral replication origin prediction. INFORMS J Comput. 2010;22(3):457–70.
96. Shatabda S, Saha S, Sharma A, Dehzangi A. iPHLoc-ES: identification of bacteriophage protein locations using evolutionary and structural features. J Theor Biol. 2017;435:229–37.
97. Qin Z, Wang M, Yan A. QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM). Bioorg Med Chem Lett. 2017;27(13):2931–8.
98. Liu Z, Lv H, Han J, Liu R. A computational model for predicting transmembrane regions of retroviruses. J Bioinforma Comput Biol. 2017;15(03):1750010.
99. Döring M, Borrego P, Büch J, Martins A, Friedrich G, Camacho RJ, Eberle J, Kaiser R, Lengauer T, Taveira N, Pfeifer N. A genotypic method for determining HIV-2 coreceptor usage enables epidemiological studies and clinical decision support. Retrovirology. 2016;13(1):85.
100. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. BMC Genomics. 2010;11(4):S21. BioMed Central.
101. EL-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. J Mol Recognit Interdiscip J. 2008;21(4):243–55.
102. Qureshi A, Kaur G, Kumar M. AVC pred: an integrated web server for prediction and design of antiviral compounds. Chem Biol Drug Des. 2017;89(1):74–83.