Published in final edited form as:
INFORMS J Comput. 2010 June 1; 22(3): 457–470.

Least-Squares Support Vector Machine Approach to Viral Replication Origin Prediction

Raul Cruz-Cano, **David S.H. Chew**, **Choi Kwok-Pui**, and **Leung Ming-Ying**
Department of Computer and Information Sciences, Texas A&M University-Texarkana, Texarkana, TX, 75501, USA, Raul.Cruz-Cano@tamut.edu

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore, david.chew@nus.edu.sg and Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA, 90089, USA

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore, stackp@nus.edu.sg

Bioinformatics Program and Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX, 79968, USA, mleung@utep.edu

Abstract

Replication of their DNA genomes is a central step in the reproduction of many viruses. Procedures to find replication origins, which are initiation sites of the DNA replication process, are therefore of great importance for controlling the growth and spread of such viruses. Existing computational methods for viral replication origin prediction have mostly been tested within the family of herpesviruses. This paper proposes a new approach by least-squares support vector machines (LS-SVMs) and tests its performance not only on the herpes family but also on a collection of caudoviruses coming from three viral families under the order of caudovirales. The LS-SVM approach provides sensitivities and positive predictive values superior or comparable to those given by the previous methods. When suitably combined with previous methods, the LS-SVM approach further improves the prediction accuracy for the herpesvirus replication origins. Furthermore, by recursive feature elimination, the LS-SVM has also helped find the most significant features of the data sets. The results suggest that the LS-SVMs will be a highly useful addition to the set of computational tools for viral replication origin prediction and illustrate the value of optimization-based computing techniques in biomedical applications.

Keywords

Replication origins; herpesviruses; caudoviruses; feature selection; least-squares support vector machines; *History:*

1. Introduction

In many viruses, replication of their DNA genomes is the central step of reproduction. Understanding the viral replication mechanism is of great importance in developing strategies to control the growth and spread of viruses (Delecluse and Hammerschmidt, 2000; Hartline et al., 2005; Villarreal, 2003) for various reasons related to health and economy. For example, viral DNA replication has been the target for a number of anti-herpesvirus drugs, including acyclovir (Wishart et al., 2006) which is used to treat patients infected with herpes simplex 1 and 2 (cold sores and genital herpes), varicella-zoster (shingles and

chicken pox), and the Epstein-Barr virus (mononucleosis). Bacteriophage (viruses living in bacteria are called bacteriophages) replication is also a big concern in dairy industries because viruses infecting lactic acid bacteria are a threat to industrial milk fermentation (Brussow, 2001).

Since replication origins are regarded as major sites for regulating genome replication, labor-intensive laboratory procedures have been used to search for replication origins (e.g., Deng et al., 2004; Newlon and Theis, 2002; Zhu et al., 1998). Early studies on the genome DNA sequences of herpesviruses have suggested that replication origins often lie around regions with unusually high concentration of palindromes (Masse et al., 1992; Reisman et al., 1985; Weller et al., 1985), where a palindrome is a stretch of DNA bases followed immediately by its reverse complement. Based on these observations, Leung et al. (2005) suggest a computational method using the scan statistics to locate significant clusters of palindromes and predict the likely locations of replication origins. Chew et al. (2005) have further developed palindrome-based scoring schemes for predicting replication origins in complete herpesvirus genomes. Their approach is to slide a window of size about 0.5% of the genome length over the sequence. As the window moves along, a score which reflects the concentration of palindromes in the window is calculated. The top scoring windows are then selected as predicted likely locations of replication origins. While the method has achieved a reasonable degree of accuracy, it can be further improved if the following issues can be addressed.

First, the method uses only a single sequence feature, namely the palindrome distribution, of the genome and does not offer any obvious generalization for multiple sequence features to be simultaneously taken into consideration. Second, when predicting replication origins for one herpesvirus, relevant information from other members of the viral family is not used. In (Cruz-Cano et al., 2007), we have proposed to address these issues by a machine learning approach, namely artificial neural networks (ANN). Like ANN, support vector machines (SVMs) can learn from characteristics of the known replication origins of those genomes in the training data set and then make predictions of where the replication origins of a new genome are likely to be. Moreover, the SVM approach has certain additional advantages over ANN. For example, for a two-class problem like the one in this study, an SVM can be trained so that the direct decision function maximizes the generalization ability (Abe, 2005). The SVM approach also allows recursive feature elimination to be conducted to suggest which sequence features are more important for identifying viral replication origins. Advantages of our SVM approach over other machine learning methods will be discussed later in greater details.

In this paper, motivated by the above considerations, we apply SVM technique for the replication origin prediction. Furthermore, we include more genome features, namely, family/subfamily information and the A+T content (recently shown in Chew et al. (2007) to be an important feature for prediction) as our input variables for SVMs. The palindrome scores, for the first time, are also included in our consideration. Consequently, we achieve higher accuracy in our prediction. The annexation of this information and the use of a feature selection technique lead us to conclusions beyond those reached in the ANN approach (Cruz-Cano et al., 2007). For example, we now believe that other information besides the one based on the content of a particular region of DNA can be used to determine if such region is a replication origin or not. Our findings in this work suggest that the best way to increase the accuracy of the predictions is not by creating or using more complex methods, but simply combining the predictions of several independent techniques. In extending this SVM prediction method to replication origin prediction in the Caudoviruses family, we also discovered that even if two families of viruses have many common characteristics (for example, the Herpesviruses and the Caudoviruses families are double-

stranded DNA with possibility of multiple replication origins) one cannot predict the replication origins in one family by using the information from another family.

The general construct of SVM allows any number of selected sequence characteristics to be included as input variables. In this study, we train the SVMs with different numbers of input variables containing information about the known replication origin locations and other characteristics of the genome sequence. The results indicate that SVMs with adequate prediction accuracy can be constructed. We shall briefly describe the SVM approach in the next section. The application of least-squares SVMs to predict replication origins in herpesviruses and caudoviruses and the prediction accuracy are presented in Section 3. A few concluding remarks are given in Section 4.

2. Least-Squares Support Vector Machines

2.1. Basic Concepts

SVMs (Vapnik, 1995) have already been used in various biological applications such as the identification of bacterial strains (Doran et al., 2007) and transition-metal-binding sites (Passerini et al., 2006), and the prediction of single nucleotide polymorphisms (Kong and Choo, 2007), protein stability changes (Cheng et al., 2006), and insurgence of human genetic diseases (Capriotti et al., 2006). Before describing its application to viral replication origin prediction, we first explain the basic concepts of SVMs briefly.

For an SVM, the available data is transformed into a numerical representation and then expressed as a matrix X and a vector *Y*. The rows of the matrix X store the vectors of inputs X_i's. The element y_i of the vector *Y* represents the desired output for the input vector X_i. X is an *N* by *n* matrix, where *n* is the dimension of the input vectors and *N* the number of input-output instances in the data set. Naturally, the length of the vector *Y* is *N*.

The modification of the parameters of the SVM in order to get the desired results is known as training. Usually, the behavior desired for an SVM is obtained by providing to it examples of inputs and the corresponding observed outputs. In most cases it is neither possible nor desirable to present all the possible instances to the SVM during training. Quite the contrary, it is expected that the SVM should be able to generalize the knowledge acquired from a reasonable number of examples. For this purpose, a portion of the data set is reserved and used to demonstrate the accuracy of the SVM for unseen cases. These instances are presented to the system during a procedure called testing which follows training.

For SVMs, the classification of the examples in both the training and test sets is performed by a decision function

$$D(X) = W^T \Phi(X) + b, \tag{1}$$

where $\varphi(X) = (\div_1(X), \div_2(X), \ldots, \div_L(X))^T$ is a vector of functions which transforms the classification problem from *n* dimensions to *L* dimensions, where usually $n \ll L$. This allows the classes to be separated by a hyperplane, even if they are not linearly separable in the original *n*-dimensional space of the input vectors *X*. The elements of the vector $W = (w_1, w_2, \ldots, w_L)$ and the bias term *b* are real numbers. It is considered that the decision function has correctly classified the input vector X_i if

$$D(X_i) \begin{cases} \geq 1 & \text{for} \quad y_i = 1 \\ \leq -1 & \text{for} \quad y_i = -1 \end{cases}. \tag{2}$$

In our case, it was preferable not to make Eq. (2) true for all the N examples in the original training set. Instead, we iteratively reduce the size of the training set and restrict our attention to only M input vectors called support vectors (SVs). How to determine which examples are dropped from the training set and why is explained later in this section.

2.2. Optimization of LS-SVMs

It should be pointed out that it is not necessary to perform computation in the high dimensional space, $\mathbb{R}^L$, in order to find the optimal SVM. Indeed, during the optimization and classification processes of SVM, the value of $\varphi(X)$ is not explicitly used, and all that are needed are the inner products $\varphi(X_i)^T \varphi(X_j)$. Under certain circumstances, these inner products can be represented by functions $K(X_j, X)$ which have the same form; just the values of their parameters are different depending on which SV X_j they are associated with. They are called kernel functions. In this case (1) can be represented as

$$D(X) = \sum_{j=1}^{M} \alpha_j y_j K(X_i, X) + b, \tag{3}$$

where M is the number of SVs and the parameters $A = (\alpha_1, \alpha_2, \ldots, \alpha_M)^T$ are real numbers to be determined. The most popular kernels are the Linear, Radial Basis Function (RBF) and Artificial Neural Networks kernels. We choose the RBF and Linear kernels for this application due to its documented ability to produce adequate results in many different fields of research (See, for example, Suykens et al. (2002)). The corresponding functions are: $K(X_j, X) = X_j^T X$ for linear kernel; and $K(X_j, X) = \exp(-\lambda \| x - X_j \|^2)$ for RBF. Here $\| \cdot \|$ denotes the norm function and λ is chosen to be 1. Since the algorithm proves to be robust for a wide range of λ values during preliminary tests, an exhaustive search for its most desirable value is not necessary. The invariability of the performance for the test set for a large range of values of λ has been pointed out also in (Abe, 2005) for a blood cell classification problem.

The generalization capabilities of the decision function can be maximized by increasing the margin, i.e. the distance between a decision function and the input vector nearest to it. It can be proved that the margin is optimized by minimizing:

$$Q(W) = \frac{1}{2}\|W\|^2, W = (w_1, w_2, \ldots, w_L)^T, \tag{4}$$

subject to the constraints in (2) for all input-output examples in the training set. Decision functions that provide maximum margins are known as hard-margin SVMs. Hard-margin SVMs may not always exist in real-life problems. One can get around this situation by introducing slack terms, represented by the ξ_i's, in (5) as follows:

Equation (2) is also modified by including the slack terms in the right-hand side of the inequality. The vector W has l elements, one for each function of the vector $\varphi()$, while only M slacks terms are needed, one for each support vector, i.e. one for each input-output vector which included in the training set. The resulting systems are known as soft-margin SVMs. The solution that minimizes (5) subject to the new constraints in (2) with equalities holding is called a least-squares support vector machines (LS-SVMs). LS-SVMs have performed well in various applications, including classification of calcium channel antagonists (Yao et al., 2005), quantification of bacteria (Borin et al., 2007) and proteomic studies (Caballero et al., 2007).

$$Q(W) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{M}\xi_i^2. \quad (5)$$

An attractive characteristic of LS-SVMs is that the optimal solutions for the vector W (or the vector A if using the kernel trick) and the bias term b can be found by solving a system of linear equations (Suykens and Vandewalle, 1999).

Besides vectors A or W and the parameter b, other parameters are needed to define the best LS-SVM. These include the M SVs X_i's in (3) and their corresponding y_i's. One approach is to consider all the vectors in the data set SVs, but this was found to limit the generalization capabilities of our LS-SVMs in our preliminary experiments. Moreover, there is no unique way to find out which vectors should be kept as SVs and which should not.

The strategy we adopt in this paper is to first use all the training data and then eliminate from it the SVs with little influence in the decision functions described in (1) and (3), i.e. SVs with small w_i's or α_j's. The reduced training set is then used to find a new set of values for A or W and b. This procedure is repeated until a predefined performance measure is no longer met. This strategy was proposed in (Suykens, 2000) and (Suykens et al., 2002) and has been widely used in SVM research due to its simplicity and effectiveness. Each time every one of the parameters is updated, an iteration is said to have occurred.

For positive cases (i.e., $y_i = 1$), we set C = number of negative cases/number of positive cases; while $C = 1$ for negative cases (i.e., $y_i = -1$). This formula is proposed for imbalance data in (Lee et al., 2001), where the different values of C are named C^+ and C^- for the classes of positive and negative cases respectively. An SVM with a severely imbalance set of SVs will tend to classify every example presented to it as a case of the over-represented class. It can be demonstrated that this choice of C values produces a more balanced set of SVs at the end of training. Hence it can improve the performance of the final SVM for the under-represented class.

2.3. Feature Selection using SVM

It is not always easy to determine which input variables should be included in the training and test data sets to obtain a suitable SVM. This issue is important because by including only useful information, a smaller classification system can be used to solve the problem. Simpler systems require less computational resources and usually lead to better performance for the test set. The process that deals with the reduction of the number of input variables is known as feature selection.

The change in the generalization capability of an SVM created by eliminating a variable can be accurately estimated (Abe, 2005). The process of iteratively eliminating from the data set the variables which produce the smallest change in the generalization capability of an SVM is called recursive feature elimination (RFE) (Guyon et al., 2002).

The Adaptive Scaling for Feature Selection (ASFE) method (Grandvalet and Canu, 2002) is an algorithm for automatic relevance determination of input variables in SVM. Relevance is measured by scale factors defining the input space and feature selection is performed by assigning smaller weights to irrelevant variables. The weights of the features are initialized to the same value; in our example this value is 1; and then they are updated by a descent algorithm while keeping W and b constant. Once the scaling factors have been updated they remain constant while the rest of the parameters (values and W, b, the number of SVs, etc.) are optimized. This process is stopped when a certain convergence criterion is met, e.g. the

difference between averages of the scaling factors from one iteration to the next is less than 0.001.

The algorithm has compared favorably to state-of-the-art feature selection procedures and demonstrated its effectiveness on tasks such as the demanding facial expression recognition problem (Grandvalet and Canu, 2002). These processes will be applied to our data sets in Subsection 3.5.

2.4. LS-SVMs versus Other Machine Learning Approaches

Other machine learning approaches were tested for this research before settling for LS-SVMs, for example, multilayer feedforward neural network based on multi-valued neurons (MLMVN) (Aizenberg and Moraga, 2006) and Fuzzy Inference Neural Networks (Rutkowska and Starczewski, 2000; Babuska, 2000). Preliminary tests show that these more complex approaches failed to achieve the degree of accuracy provided by LS-SVMs. The well-known feedforward back-propagation network provided by the MATLAB Neural Network Toolbox also failed to produce results of quality similar to those of the LS-SVMs.

Although ANN has been applied to the replication origin prediction problem before (Cruz-Cano et al., 2007), we have decided to explore the SVM approach because it offers several significant advantages:

1. SVMs are designed to provide the largest margin for the data at hand, while the ANN obtained after the application of a back-propagation algorithm, or any of its derivations, is only a locally optimal solution.
2. SVMs are less susceptible to overfitting than ANN. Overfitting occurs when a machine learning method approximates a set of data too closely, capturing the noise in it, and not its general characteristics. When overfitting occurs, there is good performance for the training set, but poor for the test set.
3. ANN are black boxes which can perform many tasks very effciently but cannot provide any further knowledge about the problem being investigated. In contrast, as mentioned in subsection 2.3, by studying how the margin changes when different input variables are deleted, one can obtain information about each variable's influence on the SVM. Such information may lend further insights into the relative importance of the variables for the real problem to be addressed. A few examples of this type of SVM applications are described in (Saeys et al., 2007).

We choose LS-SVMs over standard SVMs for the following considerations:

1. LS-SVMs has a lower computational complexity over SVMs. We only need to solve a system of linear equations in LS-SVM instead of a quadratic programming problem in SVM. This makes LS-SVMs easier to build from the ground up, and hence eliminating any dependency on software which may be diffcult to modify and to adapt to a particular application. This characteristic further facilitates the analysis and understanding of the feature selection and classification processes. Under these considerations, Abe (2008), Luo et al. (2005), Suykens et al. (2001) and Wei et al. (2007) chose LS-SVMs over SVMs in their studies.
2. The LS-SVMs have proved to be better classifiers than SVMs in numerous occasions. For example, in (Van Gestel et al., 2001), several classification methods are compared across 20 different classification problems. LS-SVMs turn out to be the best classifiers in 9 out of the 20 problems, whereas SVMs in 4 out of 20.
3. Preliminary tests in our problem showed that LS-SVMs were at least as adequate for our problem as SVMs. For the herpesviruses the best SVM correctly classified

97.46% of the testing set, while the LS-SVM got 98.10% of them right. For the Caudoviruses the best SVM correctly classified 96.92%, while the LS-SVM got 97.02% right.

For the SVMs we used the software provided by (Chang and Lin, 2001).

3. Application to Herpesviruses and Caudoviruses

3.1. Data Sets

Tables 1 and 2 show, respectively, the herpesviruses and caudoviruses that are used in this study and their known replication origins, documented in the annotations of the GenBank files.

The prediction strategy used for this research considers the viral genome as a set of equal size overlapping windows, with each window being a small DNA segment about 0.5% of the genome length. This window length is chosen because it is around the average length of the known replication origins reported in Tables 1 and 2 (0.448%). Locations of replication origins are predicted by computing the LS-SVM output for each window and selecting the top scoring windows. As in (Chew et al., 2005), we consider a prediction successful if the predicted location is within two map units of a known replication origin, where one map unit is equivalent to 1% of the genome length.

The features used for the construction of the LS-SVMs, for both herpesviruses and caudoviruses, are described below.

1. Subfamily/Family: All members of our herpesvirus data set belong to the Herpesviridae family and they are classified into the α, β and γ subfamilies according to their biological properties such as the range of hosts and types of infected cells. Our collection of caudoviruses belongs to the order Caudovirales and is divided into three families: Myoviridae, Siphoviridae and Podoviridae. The subfamilies of the herpesviruses and the families of the caudoviruses will be used as input variables to the LS-SVMs.
2. Palindrome scores: Because of the documented observation that replication origins often lie around regions with unusually high concentration of palindromes, two palindrome scores, namely the palindrome length score (PLS) and the base-weighted score (BWS1), described by Chew et al. (2005), are included as features of the LS-SVMs. Basically, PLS scores a palindrome proportional to its length whereas BWS1 scores a palindrome according to how rarely it is observed in a random nucleotide base sequence generated by a first order Markov chain. Regardless of the scoring scheme, the total score of a window is the sum of the scores of all palindromes whose centers lie within the window.
3. A+T content: The A+T content of a window refers to the percentage of A and T bases in the window. DNA replication typically requires the binding of an assembly of enzymes (e.g., helicases) to locally unwind the DNA helical structure, and pull apart the two complementary strands. Higher A+T content around the origins makes the two complementary DNA strands bond less strongly to each other. This facilitates the two strands to be pulled apart and initiate the replication process. As other studies (Chew et al., 2005,2007;Segurado et al., 2003) have reported the use of A+T content to locate DNA replication origins, we include it as a feature of our LS-SVMs.
4. Standardized window number: This is the feature that enables information about the location of the known replication origins to be fed into the LS-SVMs. First

described in (Cruz-Cano et al., 2007), the standardized window number is the window number divided by the total number of windows in the virus. Hence the window number will be normalized to a real number between 0 and 1. For example, if a virus has a total of 500 windows then the corresponding standardized window number for the 455th window is 455/500 = 0.91. The idea of including this variable as an input initially came from the observation that the replication origins are located in very similar parts of the genome in groups of viruses, especially for the herpesviruses family. Figure 1 gives a schematic representation of the genomes as vertical bars where the black colored regions are those windows close to known replication origins.

5. Dinucleotide scores: A dinucleotide is a word made up of any two nucleotide bases. The 16 possible dinucleotides in DNA are AA, AC, …, TT. In the past, measures of the dinucleotide content have been used as genomic signature for different bacteria (Jernigan and Baran, 2002; Karlin and Burge, 1995; van Passel et al., 2006). In our research the dinucleotide scores (Cruz-Cano et al., 2007) are 16 variables consisting of the natural logarithm of the proportion of each possible dinucleotide in each window divided by the product of the percentages (Pct.) of the two constituting single nucleotide bases in the whole DNA sequence of the virus. The score for a dinucleotide ab in window *w* of virus *v* is:

$$\text{score (ab)} = \log\left(\frac{\text{Times ab appears in window } w/\text{length of } w}{\text{Pct. of a in virus } v \times \text{Pct. of b in virus } v}\right).$$

The five features above are represented by a total of 23 input variables: 3 for family/subfamily classification, 2 for palindromes, 1 for A+T content, 1 for standardized window number, and 16 for dinucleotide scores. Other features were included during preliminary trials, e.g. mononucleotide and trinucleotide scores, but they do not seem to offer any significant advantage over the features mentioned above.

3.2. Tests for the LS-SVMs

As a preliminary assessment of the LS-SVM approach, with the above features, we randomly select windows from each of the herpesvirus and caudovirus data sets to form the training and test set using the procedure described in (Cruz-Cano et al., 2007).

After training, the LS-SVM is asked to classify windows randomly selected from the different genomes as close or not close to a replication origin. We intend to see how the LS-SVM compares with the ANN method presented in (Cruz-Cano et al., 2007). The preliminary test results indicate that the LS-SVM with the 23 input variables approach performs better than the ANN. For example, with the herpesvirus data the LS-SVM with 23 input variables has correctly classified 98.1% of over 7000 examples in the test set while the ANN only correctly classified 88.8%. The results for the Caudoviruses were again favorable for the LS-SVM with 23 variables with percentage of 82.9% of correct classifications compared to 79.3% for the ANN.

To assess the actual performance of the LS-SVMs in predicting replication origins for a new virus based on the information gathered from other related viruses, we carry out a test on each individual virus using all the other viruses in the same data set for training the LS-SVMs. For the herpesviruses, we implement 20 LS-SVMs. Each LS-SVM is trained with the information provided by 19 of the herpesviruses in Table 1 and then applied to predict the location of replication origins in the one remaining virus left out from the training set. The data is unbalanced since there are very few positive examples: 292 of the 5824 windows

for the caudoviruses and 365 of the 8637 windows for the herpesviruses. This problem was partially alleviated by using only a random 15% of the available negative examples. This circumstance also led us to use the formula described in the last paragraph of Section 2.2.

To avoid overfitting by the decision function, the support vectors with the lowest 5% weight in the decision function are discarded after each of the 15 iterations that is performed to create a final LS-SVM for each virus. Then, the few highest ranking windows, i.e., the windows with the highest output given by the LS-SVM, are selected as the predicted positions of the replication origins. The same test is also conducted on the caudoviruses in Table 2.

The post processing described in (Cruz-Cano et al., 2007) was applied. That is, first, a prediction is considered invalid if its position is too close to the two ends of the virus genome. Any window within the first three map units or the last two map units of the genome will not be considered as a valid prediction, where one map unit is equivalent to 1% of the genome length. These cut-off percentages are set according to the observed locations of the known origins for all the viruses of Figure 1. Second, a prediction is invalid if it lies within two map units from an already found valid prediction. This means that if two or more predictions are located within 2 map units only the prediction associated with the highest LS-SVM output among them is considered valid. Following these rules, we select the few valid windows with the highest output values from the LS-SVM to be the predicted locations of replication origins in the test sequence.

The performance of a prediction scheme is often quantified by two commonly accepted measures: sensitivity and positive predictive value (PPV). In our context, sensitivity is the percentage of known origins that are close to the regions suggested by the prediction scheme; and positive predictive value is the percentage of identified regions that are close to at least one known origin. Results from the performance tests are presented in the next section.

3.3. Performance Results

The sensitivity and PPV for the herpesvirus and caudovirus data sets are shown in Figure 2. The different prediction methods used are:

1. BWS1: using BWS1 (Chew et al., 2005) as the only predictor.
2. ANN: the artificial neural network approach as in (Cruz-Cano et al., 2007).
3. LSSVM23: using LS-SVMs with all 23 input variables described in Subsection 3.1.
4. LSSVM16: using LS-SVMs with only the 16 variables corresponding to the dinucleotide scores.
5. Art23: using LS-SVMs with all 23 variables, but with the replications origins scrambled randomly in each viral genome, therefore creating a set of artificial genomes. The purpose of these experiments will be explained in Subsection 3.5.

For the caudoviruses, where both BWS1 and ANN give rather low prediction accuracy, the LS-SVMs performs substantially better. On the other hand, in the herpesviruses where each of BWS1 and ANN already shows relatively good prediction accuracy, LSSVM23 cannot offer any improvement. However, with machine learning systems, it is often advantageous to use smaller sets of input variables that contain the most useful information. Our previous works with the herpesviruses have indicated that the 16 dinucleotide scores contain relevant information that contributes to the good prediction results when used as input variables for the ANN approach (Cruz-Cano et al., 2007). We have, therefore, also implemented LS-SVMs using only the 16 dinucleotides scores as input variables. This method (LSSVM16)

provides the best performance for herpesviruses so far. Notice that almost 90% of the herpesvirus origins are found using the top six predictions. The PPV also improves substantially. For the caudoviruses, however, the LS-SVMs with 23 variables remains superior.

It is interesting to note the contrast in prediction accuracy between the herpesviruses and caudoviruses. First, all the prediction methods used in this study perform better with the herpesviruses. Second, as seen in Figure 2, using only 16 dinucleotide scores as input variables works well for the herpesviruses but not for the caudoviruses. These observations may be explained by the difference in the degree of similarity among the genomes in the two data sets. Unlike the herpesvirus genomes which all come from the family of Herpesviridae, the caudoviruses come from three different families (Myoviridae, Podoviridae, and Siphoviridae) under the order of Caudovirales. As a consequence, the genome organizations of the herpesviruses are more similar to one another, making it possible to give reliable predictions on one herpesvirus based on the information of the other members in the data set. Furthermore, as the herpesviruses are closely related, the 16 dinucleotide scores seem to be sufficient to capture the important characteristics of the genomes and produce even more accurate predictions than the larger set of 23 input variables. The programs used to determine the performances of the LS-SVMs were run in a Dell Optiplex 745 Minitower, PentiumD 945/3.40GHz using MATLAB R2007 with Windows XP.

3.4. Joint Predictions

These preliminary results, the results presented in Section 2.4 (first paragraph) and Section 3 (Fig 2) and in (Chew et al., 2007) lead us to believe that there is not much room to improve the classification accuracy by creating or using more complex methods, but a better approach is to combine the predictions of several independent techniques. With several different methods for replication origin prediction available, it is possible to combine the results to look for “joint predictions” which refer to regions consistently predicted to be likely origins by all the methods. In the case of the herpesviruses, we take the top three predictions using LS-SVM along with the top three predictions made by BWS1 (Chew et al., 2005,Table 2) and ANN (Cruz-Cano et al., 2007, Table III) and list them in Table 3 and Table 4. The italic cells indicate the cases in which all three methods produce a similar prediction; i.e. less than 2 map units from one another. For the 20 viruses with known replication origins shown in Table 3, we find a total of 12 joint predictions all of which are close to one or more replication origins, giving a PPV of 100%. Even more remarkable is that these 12 predictions actually find 14 origins giving an average of 1.17 origins found per prediction. Although surprising, the average of 1.17 origins per prediction in this case is due to the fact that some of the CeHV1 (NC_004812) origins lie so close to one another that it is possible to find more than one origin by exploring the area around just one prediction. In contrast to herpesviruses, there are only two joint predictions found for the caudoviruses, neither of which is close to known replication origins. This is not unexpected because one can derive little reliable information for the caudoviruses from the BWS1 and ANN.

Table 4 presents prediction results for a number of other herpesviruses whose complete genome sequences are available but are not included in Table 1. These genomes are not used in the training or test sets for the LS-SVMs because no information on the locations of their replication origins is available. Given the good performance test results for the 20 herpesviruses with known replication origins, we also report the top three predicted locations from each prediction method and their joint predictions here. We expect that these predictions may help virologists experimentally determine the exact location of replication origins in these viral genomes.

3.5. Feature Selection

To assess the relative contributions of different input variables to the generalization capability of our LS-SVMs, we carry out the RFE process mentioned in Section 2.2 using 50% of the positive examples and 10% of the negative examples. The random selection leads to different rankings of the variables each time that the RFE algorithm is executed. Table 5 presents the averages and standard deviations of the rankings resulting from 12 executions of RFE (with the highest and lowest rankings for each variable excluded from the calculations).

We note that for both herpesviruses and caudoviruses, BWS1 and PLS are consistently considered the variables containing the least information. It is important to understand that this does not imply that these palindrome scores do not contain information useful for predicting replication origins. The reason for their early elimination is that during RFE, data are assumed to come from a deterministic objective function. In other words, due to their random nature and multiple contradictions, palindrome scores provide little information for LS-SVMs. Next, all the dinucleotides scores, and the A+T content are eliminated. Although the elimination of any one of these variables does not cause a significant decrease in the performance of the LS-SVMs, as a set they contain a great amount of information. Also, the order of their elimination changes from one run to another due to the random selection of the training set, this can be seen from Table 5 in the standard deviations of the rankings corresponding to the different features. These results are supported by experiments with RFE using SVMs, with both the linear and RBF kernels, and by another SVM-based feature selection technique called R^2W^2 (Weston et al., 2000) with the RBF and linear kernels. The software used for these experiments can be found in (Canu et al., 2005).

In both herpesviruses and caudoviruses, the standardized window number is considered the most important variable in all the RFE runs. Nevertheless, the prediction accuracy of the LS-SVMs on the artificial genomes with their replication origins placed at various random positions of the genome sequence has not been reduced tremendously (See Art23 in Figure 2). This suggests that the rest of the variables are informative enough to overcome the fact that the standardized window number no longer provide any useful information of the location of the real replication origins. This is further confirmed with a second experiment in which the standardized window number is removed from the set of input variables. Like Art23, the prediction accuracy is somewhat reduced but is not drastically worse than that of SVM23. The only catastrophic decrease in the performance of the LS-SVMs is observed when the original genomes were used as training set, but the artificial genomes as test set. In this situation, the LS-SVMs first receives a set of data in which the standardized window number is a good predictor and then it is asked to predict replication origins in data sets in which the same feature is basically a random number. The resulting performance on the herpesviruses and caudoviruses are respectively reduced to about one half and one third of that of the SVM23. The programs used to perform the RFE of the LS-SVMs were run in a Sun Blade 150 using MATLAB R2007 with operating system Unix.

The results of applying the ASFE to the viral data used for the RFE are presented in Tables 6 and 7. The ASFE software does not allow modification of the parameter C during running time, so it was fixed to a value of 1 instead of being adapted to compensate for the imbalance data. The parameter γ, as in Section 2.2, was set to 1. The results presented in Tables 6 and 7, like those presented in Table 5, are consistent for a wide range of values of γ and C besides those selected for our computational work. The programs used to perform ASFE were run in a Dell Optiplex 745 Minitower, PentiumD 945/3.40GHz using MATLAB R2008 with Windows XP.

The ASFE has a strong inclination for the A+T Content and the Family/Subfamily features, ranking them always as the best variables. The Standard Window Number is again considered among the most informative input variables, especially for the caudoviruses where it is the variable with the largest measure of contribution, on average. In the herpesviruses the dinucleotide AA is again considered the worst among the dinucleotides, where its average scaling value of almost zero indicates that it basically disappears from the SVM calculations. For both herpesviruses and caudoviruses, the repetitive structures variables are not considered completely irrelevant. Surprisingly, A+T Content is among the most relevant features for both the caudoviruses and herpesviruses. However, it is important to notice that the ASFE has a significantly larger average standard error (4.15) compared to the Recursive Feature Selection results (herpesviruses 2.5 and caudoviruses 2.7), suggesting that more detailed assessment of the validity of the results may be necessary.

To find out which feature selection method is the most appropriate for our problem, a comparison of the performance of an SVM trained using the top ASFE variables versus a LS-SVM created using the top RFE features would be of interest. This study can be done when the ASFE software is adapted to handle a larger training set and to compensate for the imbalance data.

4. Concluding Remarks

The potential of LS-SVMs for predicting viral replication origins has been demonstrated with the herpesvirus and caudovirus data sets. In the case of the caudoviruses where existing replication origin prediction methods do not perform well, the LS-SVM approach produces much better results. Alternatively, as in the case of herpesviruses where the other methods give reasonably successful predictions, LS-SVMs can be used in conjunction with these methods to further improve the prediction accuracy. In this paper, we have used the results common to three prediction methods as our joint predictions and obtain excellent PPV for the herpesviruses. We plan to explore other possible joint prediction schemes which may improve prediction accuracy in more general settings.

Finally, we note that the problem of determining the optimal set of sequence features for the LS-SVMs to give the best replication origin predictions still needs to be investigated in greater detail with more genome sequence data. The RFE process, which recursively maximizes the margin of the LS-SVM as the variables are being eliminated, will continue to be the key instrument for this pursuit.

Acknowledgments

We would like to thank the editor and two anonymous reviewers for helpful comments and suggestions. This research is supported in part by NIH grants S06GM26-1408, 5G12RR008124-11 and 3T34GM008048-20S1. K.P. Choi was supported by the National University of Singapore ARF research grant R-155-000-051-112. David Chew was supported by a Overseas Postdoctoral Fellowship from the National University of Singapore.

References

Abe, S. Support Vector Machines for Pattern Classification (Advances in Pattern Recognition). Springer-Verlag New York, Inc.; Secaucus, NJ, USA: 2005.

Abe, S. ANNPR ’08: Proceedings of the 3rd IAPR workshop on Artificial Neural Networks in Pattern Recognition. Springer-Verlag, Berlin, Heidelberg; 2008. Sparse least squares support vector machines by forward selection based on linear discriminant analysis; p. 54-65.

Aizenberg I, Moraga C. Multilayer feedforward neural network based on multi-valued neurons (mlmvn) and a backpropagation learning algorithm. Soft Comput 2006;11:169–183.

Babuska, R. Fuzzy Systems in Medicine, Studies in Fuzziness and Soft Computing. Vol. vol. 41. Physica-Verlag; 2000. Fuzzy clustering algorithms with applications to rule extraction; p. 139-173.

Borin A, Ferro MF, Mello C, Cordi L, Pataca LCM, Durn N, Poppi RJ. Quantification of lactobacillus in fermented milk by multivariate image analysis with least-squares support-vector machines. Anal Bioanal Chem 2007;387:1105–1112. [PubMed: 17171559]

Brussow H. Phages of dairy bacteria. Annu Rev Microbiol 2001;55:283–303. [PubMed: 11544357]

Caballero J, Fernndez L, Garriga M, Abreu JI, Collina S, Fernndez M. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. J Mol Graph Model 2007;26:166–178. [PubMed: 17229584]

Canu, S.; Grandvalet, Y.; Guigue, V.; Rakotomamonjy, A. Perception Systmes et Information. INSA de Rouen; Rouen, France: 2005. Svm and kernel methods matlab toolbox.

Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 2006;22:2729–2734. [PubMed: 16895930]

Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001

Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006;62:1125–1132. [PubMed: 16372356]

Chew DSH, Choi KP, Leung M-Y. Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. Nucleic Acids Res 2005;33:e134. [PubMed: 16141192]

Chew DSH, Leung M-Y, Choi KP. AT excursion: a new approach to predict replication origins in viral genomes by locating AT-rich regions. BMC Bioinformatics 2007;8:163. [PubMed: 17517140]

Cruz-Cano, R.; Chandran, D.; Leung, M-Y. Computational prediction of replication origins in herpesviruses. Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB ’07. IEEE Symposium on; 2007. p. 283-290.

Delecluse HJ, Hammerschmidt W. The genetic approach to the epstein-barr virus: from basic virology to gene therapy. Mol Pathol 2000;53:270–279. [PubMed: 11091851]

Deng H, Chu JT, Park N-H, Sun R. Identification of cis sequences required for lytic DNA replication and packaging of murine gammaherpesvirus 68. J Virol 2004;78:9123–9131. [PubMed: 15308708]

Doran M, Raicu DS, Furst JD, Settimi R, Schipma M, Chandler DP. Oligonucleotide microarray identification of bacillus anthracis strains using support vector machines. Bioinformatics 2007;23:487–492. [PubMed: 17204462]

Grandvalet, Y.; Canu, S. Adaptive scaling for feature selection in SVMs. In: Becker, S.; Thrun, S.; Obermayer, K.; Becker, S.; Thrun, S.; Obermayer, K., editors. Neural Information Processing Systems 15. MIT Press; 2002. p. 553-560.

Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach. Learn 2002;46:389–422.

Hartline CB, Harden EA, Williams-Aziz SL, Kushner NL, Brideau RJ, Kern ER. Inhibition of herpesvirus replication by a series of 4-oxo-dihydroquinolines with viral polymerase activity. Antiviral Res 2005;65:97–105. [PubMed: 15708636]

Jernigan RW, Baran RH. Pervasive properties of the genomic signature. BMC Genomics 2002;3:23. [PubMed: 12171605]

Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 1995;11:283–290. [PubMed: 7482779]

Kong W, Choo KW. Predicting single nucleotide polymorphisms (snp) from dna sequence by support vector machine. Front Biosci 2007;12:1610–1614. [PubMed: 17127407]

Lee, K.; Gunn, S.; Harris, C.; Reed, P. Classification of unbalanced data with transparent kernels. Neural Networks, 2001. Proceedings. IJCNN ’01. International Joint Conference on; 2001. vol.4

Leung M-Y, Choi KP, Xia A, Chen LHY. Nonrandom clusters of palindromes in herpesvirus genomes. J Comput Biol 2005;12:331–354. [PubMed: 15857246]

Luo F, Xu Y-G, Cao J-Z. Elevator traffic flow prediction with least squares support vector machines. Proc. International Conference on Machine Learning and Cybernetics 2005;vol. 7:4266–4270. Vol. 7.

Masse MJ, Karlin S, Schachtel GA, Mocarski ES. Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. Proc Natl Acad Sci U S A 1992;89:5246–5250. [PubMed: 1319057]

Newlon CS, Theis JF. DNA replication joins the revolution: whole-genome views of DNA replication in budding yeast. Bioessays 2002;24:300–304. [PubMed: 11948615]

van Passel MWJ, Kuramae EE, Luyf ACM, Bart A, Boekhout T. The reach of the genome signature in prokaryotes. BMC Evol Biol 2006;6:84. [PubMed: 17040564]

Passerini A, Punta M, Ceroni A, Rost B, Frasconi P. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins 2006;65:305–316. [PubMed: 16927295]

Reisman D, Yates J, Sugden B. A putative origin of replication of plasmids derived from Epstein-Barr virus is composed of two cis-acting components. Mol Cell Biol 1985;5:1822–1832. [PubMed: 3018528]

Rutkowska, D.; Starczewski, A. Fuzzy Systems in Medicine, Studies in Fuzziness and Soft Computing. Vol. vol. 41. Physica-Verlag; 2000. Fuzzy inference neural networks and their applications to medical diagnosis; p. 503-518.

Saeys Y, Inza I, Larraaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–2517. [PubMed: 17720704]

Segurado M, de Luis A, Antequera F. Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. EMBO Rep 2003;4:1048–1053. [PubMed: 14566325]

Suykens JAK. Least squares support vector machines for classification and nonlinear modeling. Neural Network World 2000;10:29–47.

Suykens, JAK.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. Least Squares Support Vector Machines. World Scientific Pub; Co, Singapore: 2002.

Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process. Lett 1999;9:293–300.

Suykens JAK, Vandewalle J, De Moor B. Optimal control by least squares support vector machines. Neural Networks 2001;14:23–35. [PubMed: 11213211]

Van Gestel T, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, Moor BD, Vandewalle J. Benchmarking least squares support vector classifiers. Machine Learning 2001;54:5–32.

Vapnik, VN. The nature of statistical learning theory. Springer-Verlag New York, Inc; New York, NY, USA: 1995.

Villarreal EC. Current and potential therapies for the treatment of herpes-virus infections. Prog Drug Res 2003;60:263–307. [PubMed: 12790345]

Wei L, Chen Z, Li J, Xu W. Sparse and robust least squares support vector machine: A linear programming formulation. Proc. IEEE International Conference on Grey Systems and Intelligent Services GSIS 2007 2007:1134–1138.

Weller SK, Spadaro A, Schaer JE, Murray AW, Maxam AM, Schaer PA. Cloning, sequencing, and functional analysis of oriL, a herpes simplex virus type 1 origin of DNA synthesis. Mol Cell Biol 1985;5:930–942. [PubMed: 2987682]

Weston, J.; Mukherjee, S.; Chapelle, O.; Pontil, M.; Poggio, T.; Vapnik, V. Advances in Neural Information Processing Systems 13. MIT Press; 2000. Feature selection for svms; p. 668-674.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucl. Acids Res 2006;34:D668–672. [PubMed: 16381955]

Yao X, Liu H, Zhang R, Liu M, Hu Z, Panaye A, Doucet JP, Fan B. Qsar and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines. Mol Pharm 2005;2:348–356. [PubMed: 16196487]

Zhu Y, Huang L, Anders DG. Human cytomegalovirus orilyt sequence requirements. J Virol 1998;72:4989–4996. [PubMed: 9573268]

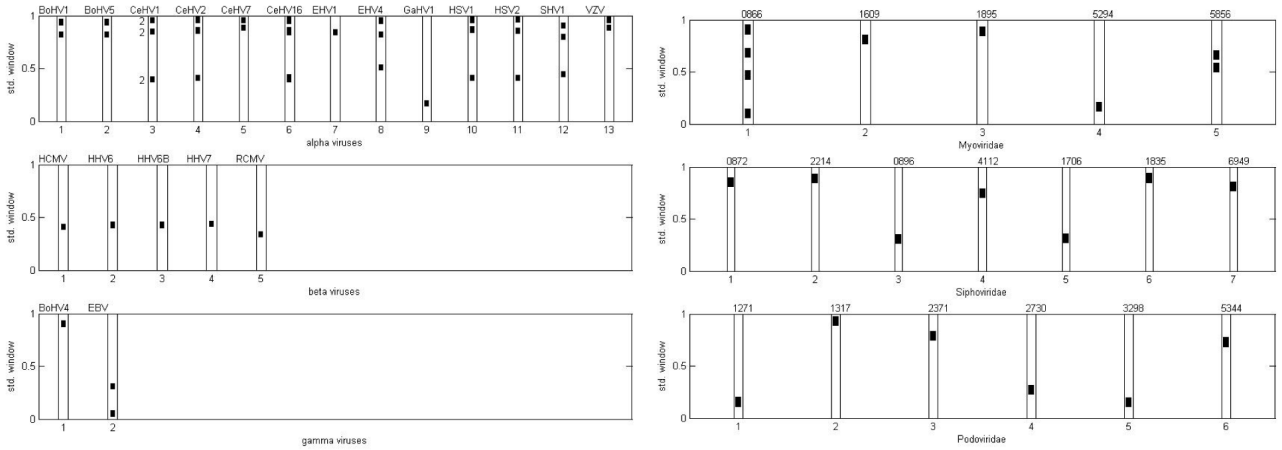

Figure 1.
Replication Origins of Herpesviruses (left) and Caudoviruses (right). The numbers on top of the Caudovirus DNA are the last 4 digits of their Accession Numbers.

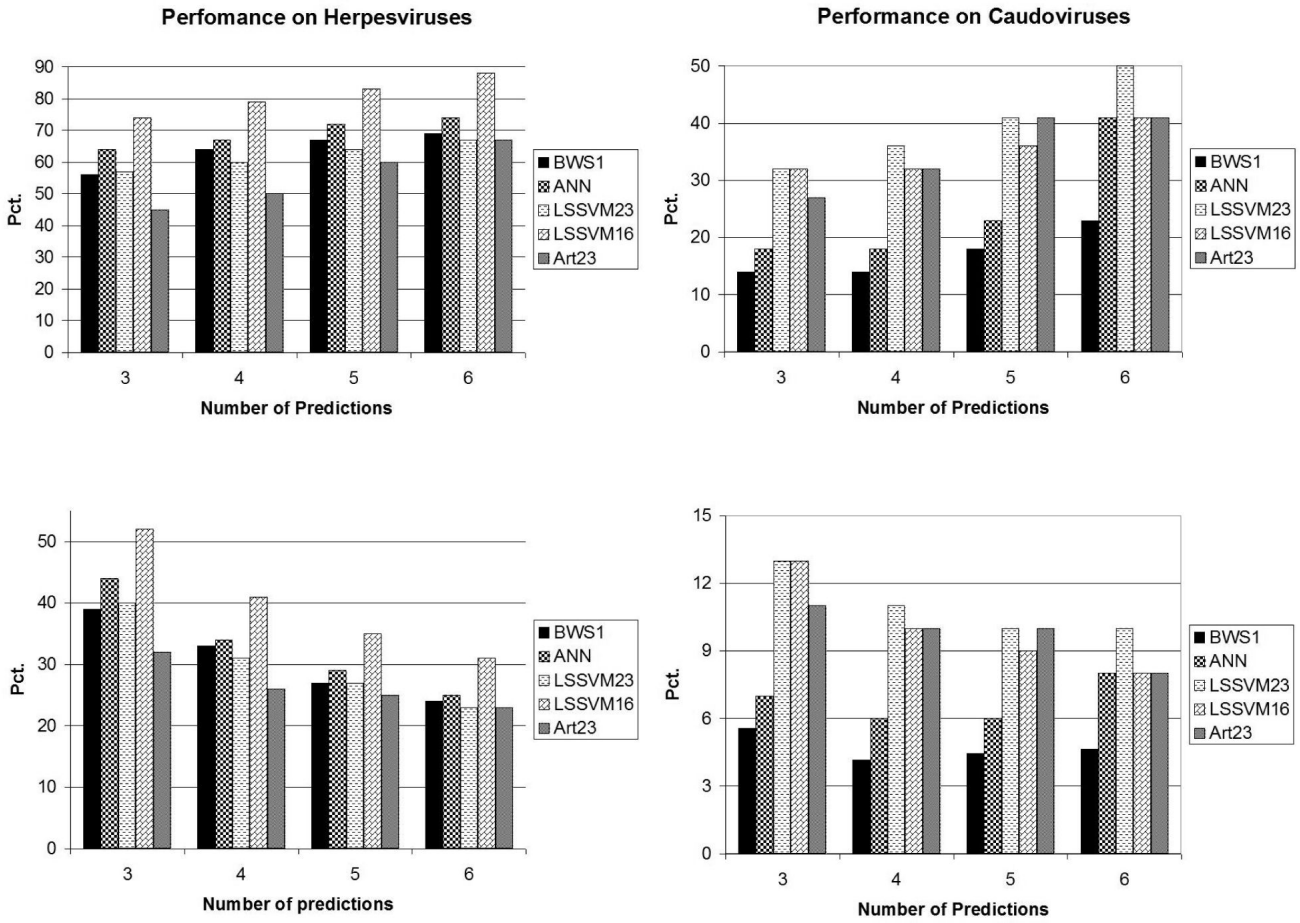

Figure 2.

The Sensitivity (top) and PPV (bottom) for the herpesviruses (left) and caudoviruses (right) are compared for the BWS1 method, the ANN method, LS-SVM with 23 variables, LS-SVM using the dinucleotides, and LS-SVM with 23 variables for the artificial genomes.

Table 1

Herpesviruses with complete genomes and known replication origins.

Virus	Abbre.	Accession	Known Origins	Genome Length	Window Length	Subfamily
Bovine herpesvirus 1	BoHV1	NC_001847	111080-111300 126918-127138	135301	300	α
Bovine herpesvirus 4	BoHV4	NC_002665	97143-98850	108873	250	γ
Bovine herpesvirus 5	BoHV5	NC_005261	113206-113418 129595-129807	138390	300	α
Cercopithecine herpesvirus 1	CeHV1	NC_004812	61592-61789 61795-61992 132795-132796 132998-132999 149425-149426 149628-149629	156789	350	α
Cercopithecine herpesvirus 2	CeHV2	NC_006560	61445-61542 129452-129623 144386-144557	150715	350	α
Cercopithecine herpesvirus 9	CeHV7	NC_002686	109627-109646 118613-118632	124138	300	α
Cercopithecine herpesvirus 16	CeHV16	NC_007653	62892-63070 133380-133578 149725-149923	156487	350	α
Human herpesvirus 4	EBV	NC_001345	7315-9312 52589-53581	172281	400	γ
Equid herpesvirus 1	EHV1	NC_001491	126187-126338	150224	350	α
Equid herpesvirus 4	EHV4	NC_001844	73900-73919 119462-119481 138568-138587	145597	350	α
Gallid herpesvirus 1	GaHV1	NC_006623	24738-25005	148687	350	α
Human herpesvirus 5 strain AD169	HCMV	NC_001347	93201-94646	230287	550	β
Human herpesvirus 6	HHV6	NC_001664	67617-67993	159321	350	β
Human herpesvirus 6B	HHV6B	NC_000898	68740-69581	162114	400	β
Human herpesvirus 7	HHV7	NC_001716	66685-67298	153080	350	β
Human herpesvirus 1	HSV1	NC_001806	62475 131999 146235	152261	350	α
Human herpesvirus 2	HSV2	NC_001798	62930 132760 148981	154746	350	α

Virus	Abbre.	Accession	Known Origins	Genome Length	Window Length	Subfamily
Murid herpesvirus 2	RCMV	NC_002512	75666-78970	230138	550	*β*
			63848-63908			
Suid herpesvirus 1	SHV1	NC_006151	114393-115009	143461	350	*α*
			129593-130209			
Human herpesvirus 3	VZV	NC_001348	110087-110350	124884	300	*α*
			119547-119810			

Table 2

Caudovirales with complete genomes and known replication origins.

Virus	Accession	Known Origins	Genome Length	Window Length	Family
Enterobacteria phage T4	NC_000866	115321 152947 16763 79355	168,903	450	Myoviridae
Streptococcus thermophilus bacteriophage Sfi21 Sfi21	NC_000872	34487-34784	40,739	150	Siphoviridae
Lactobacillus bacteriophage phi adh	NC_000896	13084-13469	43,785	150	Siphoviridae
Bacteriophage phiYeO3-12	NC_001271	5995-6136	39,600	100	Podoviridae
Enterobacteria phage 186	NC_001317	28576	30,624	100	Podoviridae
Bacteriophage P4	NC_001609	9086-9809	11,624	50	Myoviridae
Lactococcus phage c2	NC_001706	6635-7245	22,172	100	Siphoviridae
Bacteriophage sk1	NC_001835	24861-25861	28,451	100	Siphoviridae
Enterobacteria phage P2	NC_001895	29892	33,593	100	Myoviridae
Streptococcus thermophilus bacteriophage Sfi11	NC_002214	35254-35557	39,807	100	Siphoviridae
Enterobacteria phage P22	NC_002371	32840-32984	41,724	150	Podoviridae
Bacteriophage HK620	NC_002730	10369-10464	38,297	100	Podoviridae
Bacteriophage T3	NC_003298	5682-5758	38,208	100	Podoviridae
Lactobacillus casei bacteriophage A2 virus	NC_004112	32208-32717	43,411	150	Siphoviridae
Bacteriophage EJ-1 provirus	NC_005294	6884-6966	42,935	150	Myoviridae
Enterobacteria phage Sf6 virion	NC_005344	28385-28480	39,043	100	Podoviridae
Enterobacteria phage P1 virion	NC_005856	50726-51571 62688-62929	94,800	250	Myoviridae Siphoviridae
Salmonella typhimurium bacteriophage ES18 virion	NC_006949	37998-38082	46,900	150	Siphoviridae

INFORMS J Comput. Author manuscript; available in PMC 2011 June 1.

Table 3

First 3 Predictions for BWS1, ANN and LS-SVM methods.(Underline = one replication origin found; Bold = two replication origins found; Italics = joint prediction, i.e. all three methods produce predictions close to one another; 'N/A' = virus was not included in previous studies but has since been sequenced and available for analysis.)

Virus	BWS1 Rankings			ANN Rankings			LS-SVM Rankings		
	1	2	3	1	2	3	1	2	3
NC_001847	*113401*	124501	87301	105367	127219	*110755*	126919	*110755*	99979
NC_002665	54751	30215	72251	96637	22474	5993	104877	96138	66673
NC_005261	18901	*113401*	*129601*	*130002*	*112629*	68297	*112929*	*129703*	68297
NC_004812	**133001**	**149451**	**61601**	**150140**	**132640**	123892	124241	**149439**	**131240**
NC_006560	*129501*	144201	61601	*127986*	143722	123440	61544	121690	*129734*
NC_002686	18601	106201	121801	109445	118740	104048	119041	109445	101050
NC_007653	N/A	N/A	N/A	150899	122604	131687	62875	125050	132385
NC_001345	7601	*53201*	127601	*51164*	69951	150696	69951	7995	*51564*
NC_001491	116201	147001	47601	142539	121228	73715	120179	73715	109698
NC_001844	105351	143151	109901	128098	136847	44449	136847	120048	109898
NC_006623	68601	41651	99751	121748	137492	144488	123498	130144	137492
NC_001347	*94501*	174901	196351	214897	189616	*93982*	174776	139052	*93982*
NC_001664	8051	30101	110601	131370	68479	142200	68479	131370	113551
NC_000898	90801	132801	8801	69477	128973	157722	69477	132167	151732
NC_001716	9451	152251	133351	145740	9436	128615	48580	42988	143294
NC_001806	62301	*129851*	*148401*	*130259*	*147372*	9429	*146324*	*131308*	124323
NC_001798	74551	28001	12951	148109	133437	127150	125055	132740	147759
NC_002512	75901	110551	83601	134019	201028	213660	152144	75248	178509
NC_006151	38151	11551	93101	16445	127016	117568	129115	115469	63682
NC_001348	*119401*	*110101*	100501	20665	*120092*	*109311*	*120092*	61095	*109611*

Table 4

First 3 Predictions for BWS1, ANN and LS-SVM methods.(Italics= Joint Prediction, i.e. all three methods produce a prediction close to one another; 'N/A' = virus was not included in previous studies but has since being sequenced and available for analysis.)

Virus	BWS1 Rankings			ANN Rankings			LS-SVM Rankings		
	1	2	3	1	2	3	1	2	3
NC_002531	*113701*	123301	32701	119224	*114132*	127612	76387	*113534*	28757
NC_001987	99251	97001	54751	102165	98168	7244	63197	102165	62697
NC_004367	*116201*	133351	*23101*	*116818*	*22734*	56660	*115769*	*22734*	55961
NC_006146	*8001*	*34801*	*138801*	*136317*	*7995*	*34779*	*34779*	*8395*	*138716*
NC_006150	*161151*	147401	198001	*161557*	176944	88471	*161008*	87372	187383
NC_003521	*91201*	207001	*177001*	*176917*	*90557*	199107	*91157*	*176917*	98954
NC_001650	54001	6301	173251	12595	150690	139895	20243	53978	138995
NC_002229	160801	801	137601	137744	146928	52303	126564	12377	132953
NC_002577	158801	138401	*11201*	122703	131096	*10791*	*10791*	122703	69145
NC_006273	175451	*94051*	153451	*94477*	*89533*	175772	*93928*	28013	139518
NC_001493	55501	89701	9301	62320	123140	126136	7790	59923	123140
NC_007016	N/A	N/A	N/A	115639	24866	128221	115639	23068	111445
NC_002641	5601	*117951*	11551	152513	*119633*	38128	*116835*	155662	38128
NC_004065	92951	142451	200201	92880	184112	197302	184112	80240	169823
NC_001826	*99251*	*26251*	62001	*100708*	*26739*	70471	*26739*	70471	*101208*
NC_005881	21001	144001	*187501*	*184946*	196443	146456	*184446*	17995	202440
NC_007646	N/A	N/A	N/A	119854	6592	130940	118955	115359	35357
NC_005264	130401	*151601*	18801	*151437*	141049	87505	*151437*	129861	134256
NC_001350	103751	112501	81501	96190	67708	5746	68457	5996	96190
NC_002794	134101	10801	144901	107813	189121	169804	107813	129375	48965
NC_003401	132601	*117601*	3301	*116029*	26384	130421	*118129*	23985	64161
NC_003409	*23401*	*119701*	136501	124626	*118335*	*24266*	*24866*	*118635*	126124
NC_008211	N/A	N/A	N/A	N/A	N/A	N/A	172511	203829	15933
NC_008210	N/A	N/A	N/A	N/A	N/A	N/A	226857	172478	217520

Table 5

Results of the Recursive Feature Selection. Rank 1(resp. 23) indicates the most (resp. least) contribution to the LS-SVM.

Herpesviruses	Avr. Rank	Std. Dev.	Caudoviruses	Avr. Rank	Std. Dev.
Std. Win.	1	0	Std. Win.	1	0
GC	5.5	3.5	AC	3.9	2.2
TG	6	2.5	AA	5.8	2.5
GG	6.3	3.7	GT	6.4	3.2
TC	7.6	3.6	CT	6.8	2.9
AG	7.9	2.7	GG	8.3	2.3
GT	9.1	2.6	AT	8.8	4.8
GA	9.2	3.9	CC	9.5	3.8
CT	9.5	4	TA	9.8	3.2
α	9.7	4.1	TT	9.9	5
AT	9.9	3.7	CG	10.3	4.4
CG	10.6	5.3	TG	11	5.3
AC	10.8	3.3	GA	11.1	2.9
CA	11	2.3	GC	12	2.7
β	13	4.4	TC	12.3	4.4
CC	13.4	3.3	Myoviridae	14.3	2.6
A+T Content	14.6	2.8	CA	14.9	3.9
γ	18.3	0.5	Podoviridae	16.9	2.6
TT	18.7	1.5	AG	17.6	1.7
TA	19.8	0.4	Siphoviridae	20.1	1
AA	21	0	A+T Content	20.6	0.5
PLS	22	0	PLS	22	0
BWS1	23	0	BWS1	23	0

Table 6

Results of the Adaptive Scaling Feature Selection: Herpesviruses. The left block gives the average value and standard deviation of the scaling factor while the right block gives those of the rank.

Variable	Average (Scaling Factor)	Std. Dev. (Scaling Factor)	Variable	Average (Rank)	Std. Dev. (Rank)
A+T Content	2.5	1.0	A+T Content	1.6	0.8
β	1.2	0.1	α	3.4	1.4
α	1.2	0.1	γ	4.2	2.4
γ	0.9	0.4	β	4.3	1.7
GC	0.9	0.4	GC	6.4	3.2
Std. Win.	0.9	1.1	BWS1	7.5	2.8
BWS1	0.6	0.2	GG	8.2	2.7
GG	0.6	0.1	TG	9.3	4.0
TG	0.5	0.2	Std. Win.	11.4	6.9
CG	0.4	0.4	AT	11.9	4.9
AT	0.4	0.2	CG	11.9	7.2
CC	0.4	0.3	CA	12.3	4.2
CA	0.4	0.2	TA	12.7	5.7
TA	0.4	0.3	AC	13.1	4.1
AC	0.4	0.1	CC	14.3	6.1
TT	0.3	0.2	TT	14.7	5.2
PLS	0.2	0.2	CT	16.9	5.7
GT	0.2	0.1	PLS	16.9	5.9
CT	0.2	0.2	GT	17.8	5.8
GA	0.2	0.2	GA	18.8	6.4
AG	0.2	0.1	AG	19.2	6.4
TC	0.1	0.1	TC	20.1	6.5
AA	0.0	0.1	AA	20.9	6.6

INFORMS J Comput. Author manuscript; available in PMC 2011 June 1.

Table 7

Results of the Adaptive Scaling Feature Selection: Caudoviruses

Variable	Average (Scaling Factor)	Std. Dev. (Scaling Factor)	Variable	Average (Rank)	Std. Dev. (Rank)x
Std. Win.	5.2	3.3	Siphoviridae	4.0	2.3
A+T Content	2.3	2.1	Podoviridae	4.1	2.7
Podoviridae	2.2	0.5	Std. Win.	4.2	6.0
Siphoviridae	2.2	0.5	Myoviridae	4.9	1.7
Myoviridae	1.6	0.9	A+T Content	4.9	2.7
TA	1.1	0.7	TA	7.6	3.0
GG	1.0	1.1	GA	9.0	1.6
GA	0.8	0.3	TC	10.4	2.5
TC	0.7	0.3	CA	10.7	2.1
CA	0.7	0.2	AA	12.1	3.6
CG	0.7	0.2	GG	12.2	6.5
AA	0.6	0.3	CG	12.7	2.2
CT	0.6	0.5	CT	14.0	6.1
AC	0.6	0.8	PLS	15.2	5.5
PLS	0.5	0.4	TG	15.2	3.1
AT	0.5	0.2	AT	15.3	3.2
GT	0.5	0.5	TT	15.4	3.2
TG	0.5	0.3	CC	15.8	5.8
TT	0.5	0.2	GT	15.9	6.2
CC	0.5	0.4	AC	16.1	5.5
BWS1	0.3	0.4	GC	17.1	4.0
AG	0.3	0.2	BWS1	18.3	4.1
GC	0.3	0.2	AG	19.3	1.9