Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Computational Prediction of Replication Origins in Herpesviruses

Raul Cruz-Cano*[1, 2], Deepak Chandran[1] and Ming-Ying Leung [1, 3]
[1]Bioinformatics Program, University of Texas at El Paso, [2]Department of Electrical and Computer Engineering, University of Texas at El Paso, [3]Department of Mathematical Sciences, University of Texas at El Paso
500 West University Avenue
El Paso, Texas, 79968, U.S.A.

Abstract- Computational methods for replication origin prediction in individual herpesvirus genomes have been previously devised based on the locations of high concentrations of palindromes. In order to make use of similarities in genome composition and organization of related herpesviruses, an artificial neural network approach is explored. We implement feed-forward artificial neural networks trained by 17 input variables comprising the positions of known replication origins relative to the genome lengths and the dinucleotide scores. The overall prediction accuracy of the neural network approach for our data set is better than that of the palindrome based approach. Furthermore, suitable combinations of the prediction results given by the two approaches substantially increase the prediction accuracy achieved by either method applied individually.

I. INTRODUCTION

The herpesvirus family includes some of the well-known viruses such as herpes simplex, varicella-zoster (chicken pox), and cytomegalovirus. Some of these viruses are believed to pose major risks in patients with suppressed immune response after major surgeries like organ transplants, while others have been associated with life-threatening diseases such as AIDS and various cancers ([1], [2], [3] and [4]). As the central step in the reproduction of herpesviruses, viral DNA replication has been the target for a number of anti-herpesvirus drugs (e.g., acyclovir). To further develop strategies to control the growth and spread of viruses, it is important to understand the viral replication mechanism ([5], [6] and [7]).

Replication origins are places on DNA molecules where replication processes are initiated. As they are regarded as major sites for regulating genome replication, labor-intensive laboratory procedures have been used to search for replication origins ([8], [9] and [10]). Computational algorithms which predict likely replication origin locations can expedite the process by focusing the search to certain regions of the viral genome (see [11] and references therein).

Based on the observation that replication origins in herpesviruses often lie around regions of their DNA genome sequence with unusually high concentration of palindromes ([12], [13] and [14]), Leung et al. [11] suggest using statistically significant clusters of palindromes to computationally predict likely locations for replication origins prior to experimentation. DNA palindromes are words from the nucleotide base alphabet {A, C, G, T} that are symmetrical in the sense that they read exactly the same as their complementary sequences in the reverse direction. For example, the string GCAATATTGC is a DNA palindrome because its complementary sequence CGTTATAACG, when read in reverse, is exactly the same as itself. The high concentration of palindromes near replication origins is generally attributed to the fact that initiation of DNA replication typically requires an assembly of enzymes such as helicases to bind to the DNA, locally unwind the helical structure and pull apart the two complementary strands. The symmetry created by palindromes is advantageous for providing suitable binding sites for these DNA-binding proteins which are often dimeric in structure.

The statistically based palindrome prediction method is further improved in [15] and has achieved 80% sensitivity in detecting the known and documented replication origins on a set of 19 herpesvirus genomes. However, this method has a drawback as it does not make use of any information known about the replication origins locations in closely related members of the herpesvirus family. Since a number of herpesviruses are known to have similar overall genome organization, their replication origins are likely to be in similar positions. Knowledge about the locations of replication origins in one herpesvirus can be very relevant for predicting origins in other herpesvirus.

To address this issue, we introduce a prediction method based on artificial neural networks (ANN) which can learn from characteristics of the known replication origins of those genomes in the training data set and then predicts of where the replication origins of a new genome are likely to be. Our results indicate that the palindrome and ANN approaches complement each another very well. We find that the ANN method is able to predict those replication origins missed by the palindrome method, and that those locations predicted by both methods are highly likely to be true replication origins.

In this study, the ANN is trained with 17 input variables containing information about the known replication origin locations and the relative abundance of the 16 dinucleotides along the genomes sequences. Any two consecutive nucleotide bases in a DNA sequence are counted as a dinucleotide. For example, the sequence ACCCTG contains the dinucleotides AC, CC, CC, CT, and TG in that order. Among the 16 distinct dinucleotides which can be formed from the four letter nucleotide base alphabet, CC is observed twice, AC, CT, and TG are each observed once, and the other 12 dinucleotides are not observed at all. We shall describe later in this paper the scoring scheme used to calculate the relative abundance for the

16 dinucleotides along the genome sequences in our data sets.

We choose the dinucleotide scores as input variables for the following reason. With few exceptions, herpesviruses are classified into the α, β and γ subfamilies [16] according to their biological properties such as the range of hosts and types of infected cells. Generally, members within a subfamily have similar genome organization, their sequences are more conserved, and their replication origins are often found at similar locations. However, it would be desirable to have the replication origin predicted even before knowing the subfamily of the virus because the classification process may involve rather lengthy biological investigations. The ideal is to make use of certain sequence characteristics which can indicate subfamily information reasonably well and at the same time can easily be converted to numerical input variables for the ANN. Leung et al., in [17], have reported that the herpesviruses in the same subfamily tend to have more similar dinucleotide representation than those in different subfamilies. It is therefore conceivable that the 16 dinucleotide scores would capture useful sequence characteristics for each subfamily.

Section II is a brief review of a few concepts relevant to ANN. Details about the set of herpesvirus genome data used in this study is then described in Section III. How ANN is applied to the prediction of replication origins in these herpesviruses is explained in Section IV. In Section V the prediction accuracy of the ANN is evaluated and compared against the results presented in [15]. Finally, we give a few concluding remarks in Section VI.

II. ARTIFICIAL NEURAL NETWORKS

ANN is a mathematical model based on the human brain. Its behavior depends on the strengths of the connections, also called the weights of the network, among simple processing units or neurons. The modification of the weights of the network in order to get the desired results is known as training. Usually, the behavior desired for an ANN is obtained by providing to it examples of inputs and the corresponding observed outputs. These examples are also called instances or records. ANN can solve problems that are posed as classification, recognition, prediction or identification problems. Examples in bioinformatics include: prediction of promoter sites ([18] and [19]) on DNA sequences, protein secondary structure prediction ([20] and [21]), automatic classification of protein sequences ([22] and [23]), and prediction of glycosylation sites in amino acid sequences [24].

Inputs and outputs of ANN are represented as a series of real numbers. The conversion from the "natural" representation to a numerical representation can take a variety of forms; the specific process used for the research at hand is described in Section III. Usually the training data set for an ANN is represented as $(\mathbf{X}, \vec{\mathbf{y}})$ where the rows of the matrix $\mathbf{X}$ store the vectors of the inputs $\vec{\mathbf{x}}_i$'s. The element y_i of the vector $\vec{\mathbf{y}}$ represents the corresponding desired output for the input vector $\vec{\mathbf{x}}_i$. $\mathbf{X}$ is an N by n matrix, where N is the number of input-output instances in the data set and n is the dimension of the input vectors.

In most cases it is neither possible nor desirable to present all the possible instances to the ANN during training. Usually, a portion of the data set is reserved and used to demonstrate the accuracy of the ANN for unseen cases. These instances are presented to the system during a procedure called testing which follows training. The union of the training and test set is referred to, in this paper, as the overall data set. For this study, both training and test sets can be constructed from the 19 herpesvirus genomes with known replication origins presented later in Table I.

The analysis in the next sections is performed with feed-forward networks. The feed-forward topology selected for this study has n input units, h processing units and one output unit. More specifically, the function performed by an ANN with such a topology can be represented as a function f_{ANN} such that

$$f_{ANN}(\vec{\mathbf{x}}_i) = f_{net}\left(\sum_{k=1}^{h} b_k \cdot f_{net}\left(\sum_{j=1}^{n} w_{j,k} \cdot x_{i,j}\right)\right) \cdot \quad (1)$$

Here, i, j, and k are positive integers bounded above by N, n, and h respectively. All of $x_{i,j}$, $w_{j,k}$ and b_k are real numbers. The number h of hidden units, the weights $w_{j,k}$ and b_k are parameters that need to be determined. To find a desirable h might involve a trial and error process testing the performance of the networks with different values of h before settling on one. The values of $w_{j,k}$ and b_k can be initialized with algorithms which might save time during the training process, e.g. the Nguyen-Widrow initialization algorithm [25]. The Nguyen-Widrow algorithm is based on the expectation that picking appropriate weights to cover different regions in the input space $\mathbf{X}$ will substantially improve the learning speed. Finally, the activation function, denoted by f_{net}, also needs to be fixed. For the ANN presented in this paper, we use a sigmoidal activation function

$$f_{net}(x) = \frac{2}{1+\exp(-2x)} - 1 \quad (2)$$

which is one of the most common activation functions for ANN [26].

The performance of a network during the training process can be measured by the Mean Square Error (MSE) defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(f_{ANN}(\vec{\mathbf{x}}_i) - y_i)^2 \cdot \quad (3)$$

From (1) – (3), both $\partial \mathrm{MSE} / \partial w_{j,k}$ and $\partial \mathrm{MSE} / \partial b_k$ can be obtained for any j and k. One of the most successful training algorithms, the back-propagation algorithm, operates by modifying the values of the ANN parameters in the opposite direction to the gradient, therefore reducing the MSE.

Each time that every one of the weights is updated once is called an epoch. If a training algorithm is run for too many epochs, the network will experience "over-training", which is the phenomenon of reducing the MSE in the training set while increasing it in the test set. This happens because the ANN begins to capture the noise in the training examples instead of the general behavior of the system represented by the examples in the overall data set.

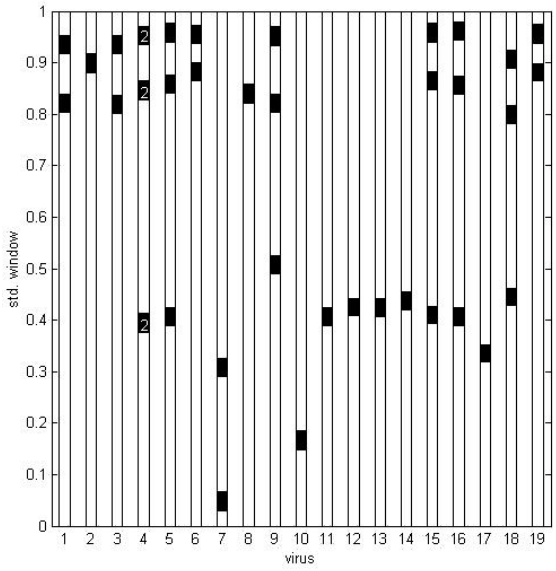

Fig. 1. Standardized window numbers for the known replication origin locations (darkened regions) in the 19 herpesvirus genomes numbered according to their ID numbers in Table I.

While the back-propagation algorithm is a steepest decent algorithm, there are other paradigms to train the networks, for example, the nonlinear least squares algorithms. This kind of methods require a larger amount of memory compared to the gradient techniques like back-propagation, but in many cases yield a smaller MSE and/or take fewer epochs to produce the same MSE [27]. The Marquart-Levenberg algorithm [27] belongs to this class of methods and is considered a modification of the Gauss-Newton method. Here the change in all the parameters is calculated simultaneously by solving the equation:

$$\Delta\overline{\omega} = [J^T(\overline{\omega})J(\overline{\omega}) + \mu I]^{-1} J^T(\overline{\omega})\overline{e}(\overline{\omega}) \qquad (4)$$

where $\overline{\omega}$ is the vector of all the weights $w_{j,k}$ and b_k ; i.e. $\overline{\omega} = [w_{1,1}\ w_{1,2}\ \ldots w_{1,h}\ w_{2,1}\ \ldots w_{2,h} \ldots w_{n,h}\ b_1\ b_2 \ldots b_h\ ]^{\mathrm{T}}$. The length of $\overline{\omega}$ is $(n \cdot h) + h$. $J(\overline{\omega})$ is a Jacobian matrix, size $N \times ((n \cdot h) + h)$, each of its entries is the gradient of an error with respect to a parameter of the ANN:

$$J_{i,j}(\overline{\omega}) = \frac{\partial(f_{ANN}(\overline{\mathbf{x}}_i) - y_i)}{\partial \omega_j},\ 1 \le i \le N,\ 1 \le j \le ((n \cdot h) + h).$$

The parameter μ in (4) is a real number selected in such a way that the MSE is reduced after the change of the values of the parameters of the ANN. The vector $\overline{e}(\overline{\omega})$ is the list of the N errors produced by an ANN with parameters $\overline{\omega}$.

Neural Networks trained with the Marquart-Levenberg algorithm have been successfully used in bioinformatics problems such as the data analysis and parameter determination of Protein-Lipid System [28] and the prediction of MHC Class 11-binding Peptides [29].

Another important concept associated with ANN is the preprocessing of the data. Basically, preprocessing is the procedure of converting the raw data provided by the experts on the problem at hand into the overall data set before presenting it to the ANN during training and testing. Preprocessing includes, among other transformations, the conversion from the "natural" to the numerical representation of the data and is not limited to input variables. Often times, complex preprocessing must be applied to the input data [30]. For example, in [31], the sequence information in DNA segments is compressed into 13 inputs before being presented to the ANN. In the current study, we use principal component analysis (PCA) as part of the preprocessing to reduce the number of variables used during the training and testing of the ANNs. PCA uses the idea that it is possible to eliminate, by an orthogonalization procedure, the features which provide the least variation in the data sets without degrading the performance of the classification process [32]. Since the orthogonalization procedure can be expressed as a matrix multiplication, the variables obtained after the PCA are just a linear combination of the original features presented to the PCA.

In many cases the format of the overall data needed to adequately train and test the ANN is not easy to interpret by humans. To facilitate the extraction of useful information from the trained ANN or from the outputs produced by it, a set of transformations might be applied to them. These operations are usually called post-processing.

The construction of the ANN for replication origin prediction in herpesviruses will be described in Section IV. To create and train these ANN, we use the facilities provided by the Matlab Neural Network Toolbox. Before discussing these implementation issues, we first present our data sets.

III. THE DATA SETS

We shall test the performance of the ANN approach to replication origin prediction by applying the method to the herpesvirus genomes listed in Table I. This is the same set of sequences in [15] where these viral genomes are considered as overlapping windows, with each window being a segment about 0.5% of the genome length. The number of overlapping windows contained in each of the viruses is listed in the column "Number of Windows" in Table I. Locations of replication origins are predicted by computing a palindrome score for each window and selecting the top scoring windows. Our ANN approach uses the same scoring idea. However, instead of the palindrome score, we compute for each window the standardized window number and the 16 dinucleotide scores.

The standardized window number is the window number divided by the total number of windows in the virus. 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. Fig. 1 gives a schematic representation of the 19 genomes as vertical bars where the shaded regions are those windows close to known replication origins.

The dinucleotide scores of each window reflect the relative abundance or rarity of the 16 dinucleotides in the window given the base composition of the genome. More precisely, the score for a dinucleotide XY in window w of virus v is the logarithm

of the ratio $w_{XY}/v_X v_Y$ where w_{XY} is the relative frequency of the dinucleotide XY in window w of the virus v and v_X, v_Y are respectively the relative frequencies of the bases X and Y in the viral genome.

TABLE I. HERPESVIRUSES WITH COMPLETE GENOMES AND KNOWN REPLICATION ORIGINS.

Id.	Virus name	Abbrev.	Accession	Known Replication Origins	Genome Length	Number of Windows
1	Bovine herpesvirus 1	BoHV1	NC_00184	111080-111300 (oriS)	135301	300
				126918-127138 (oriS)		
2	Bovine herpesvirus 4	BoHV4	NC_00266	97143-98850 (oriLyt)	108873	250
3	Bovine herpesvirus 5	BoHV5	NC_00526	113206-113418 (oriLyt)	138390	300
				129595-129807 (oriLyt)		
4	Cercopithecine herpesvirus 1	CeHV1	NC_00481	61592-61789 (oriL1)	156789	350
				61795-61992 (oriL2)		
				132795-132796 (oriS1)		
				132998-132999 (oriS2)		
				149425-149426 (oriS2)		
				149628-149629 (oriS1)		
5	Cercopithecine herpesvirus 2	CeHV2	NC_00656	61445-61542 (oriL)	150715	350
				129452-129623 (oriS)		
				144386-144557(oriS)		
6	Cercopithecine herpesvirus 9	CeHV7	NC_00268	109627-109646	124138	300
				118613-118632		
7	Human herpesvirus 4	EBV	NC_00134	7315-9312 (oriP)	172281	400
				52589-53581(oriLyt)		
8	Equid herpesvirus 1	EHV1	NC_00149	126187-126338	150224	350
9	Equid herpesvirus 4	EHV4	NC_00184	73900-73919 (oriL)	145597	350
				119462-119481 (oriS)		
				138568-138587(oriS)		
10	Gallid herpesvirus 1	GaHV1	NC_00662	24738-25005(oriL)	148687	350
11	Human herpesvirus 5 strain	HCMV	NC_00134	93201-94646 (oriLyt)	230287	550
12	Human herpesvirus 6	HHV6	NC_00166	67617-67993 (oriLyt)	159321	350
13	Human herpesvirus 6B	HHV6B	NC_00089	68740-69581(oriLyt)	162114	400
14	Human herpesvirus 7	HHV7	NC_00171	66685-67298	153080	350
15	Human herpesvirus 1	HSV1	NC_00180	62475 (oriL)	152261	350
				131999 (oriS)		
				146235 (oriS)		
16	Human herpesvirus 2	HSV2	NC_00179	62930 (oriL)	154746	350
				132760 (oriS)		
				148981 (oriS)		
17	Murid herpesvirus 2	RCMV	NC_00251	75666-78970 (oriLyt)	230138	550
18	Suid herpesvirus 1	SHV1	NC_00615	63848-63908 (oriL)	143461	350
				114393-115009 (oriS)		
				129593-130209 (oriS)		
19	Human herpesvirus 3	VZV	NC_00134	110087-110350	124884	300
				119547-119810		

The product $v_X v_Y$ in the denominator is the expected relative frequency of the dinucleotide XY in the genome assuming the genome sequence is generated as independent and identically distributed random variables taking values A, C, G, and T with probabilities equal to v_A, v_C, v_G, and v_T respectively.

If the dinucleotide XY occurs more abundantly in window w than expected from the random letter sequence model, the ratio $w_{XY}/v_X v_Y$ will be greater than one and its logarithm will be positive. Likewise, if XY occurs less abundantly than expected, the score will be negative.

IV. COMPUTATIONAL IMPLEMENTATION

In order to find the best number of hidden units, a series of neural networks are created with the Matlab Neural Network toolbox function 'newff', this function adds biases to all the hidden and output units. A bias is a real number which is added to the sum of the multiplications of the weights of a hidden or output unit by the values provided by the units in the previous layer. Since the inclusion of biases usually leads to a better performance for both the training and test set, it is very common to incorporate them in ANN. This addition is performed before applying the activation function 'fnet' described in (2). Matlab automatically includes the biases in any initialization and training algorithm applied to the ANN.

To initialize the networks with the Nguyen-Widrow method, the function 'init' is used on the networks. Then, the networks are trained with Levenberg-Marquardt back-propagation algorithm; preliminary tests have shown that this algorithm actually performed better than the back-propagation algorithms offered by the toolbox. The Matlab Neural Network toolbox provides a function 'train' that implements this algorithm, if the option 'trainlm' is selected.

The neural network with the best general performance with the data sets described above has 87 hidden units. With PCA preprocessing, also provided by the Matlab Neural Network Toolbox, the 17 input features of the data set are reduced to 13 "new" variables which account for over 98% of the total variance of the data. The results obtained after 600 epochs are shown in Table II. More information about all the algorithms mentioned in this section and other capabilities of the Matlab Neural Network Toolbox can be found in [33], which is freely accessible at www.mathworks.com.

A difficulty with this data set is the overwhelming presence of negative cases, 7587 or them versus only 602 positive cases. If the training set is constructed with the same proportions, there would be 92.64% negatives and only 7.36% positives. Given these circumstances the network might learn that by classifying everything presented to it as negative, it will achieve a very good MSE. This problem was solved by randomly selecting two thirds of the positive cases and only half of the negative cases for the training set. The rest of the records form the test set. The random selection is necessary to avoid training the network to recognize the replication origins of only certain viruses.

Notice that in this case, both training and test sets contain information, i.e. instances, from the same viruses.

TABLE II. PERFORMANCE MEASURED IN PERCENTAGE OF CORRECT CLASSIFICATION FOR THE OVERALL DATA SET AS WELL AS THOSE FOR THE POSITIVE AND NEGATIVE CASES.

Data Set	Overall	Positive	Negative
Training	99.6	96.5	99.9
Test	88.8	67.6	90.0

This is not a realistic representation of what a researcher with a new virus sequence would have available to train and test an ANN. In reality, the training set would exclusively be instances taken from the viruses with known origins, while the test set data would come from the new virus. In other words, training and test sets should contain information coming from different viruses.

To assess the performance of our ANN approach realistically, we implement 19 networks like the one described above. Each network is trained with 18 of the herpesviruses in Table I and then applied to predict the location of replication origins in the one remaining virus left out from the training set. To avoid overtraining, only 30 epochs are used during the training of these networks. In order to select appropriate regions as likely replication origins, the predictions are subject to the following post-processing steps.

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 the observed locations of the known origins for all the viruses of Table I.

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 ANN output among them is considered valid.

Following these rules, we select the few valid windows with the highest output values from the ANN to be the predicted locations of replication origins in the test sequence.

V. RESULTS AND DISCUSSION

Chew et al., in [15], have developed several palindrome scoring schemes for predicting replication origins in 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 then become predicted likely locations of replication origins. Among several palindrome scoring schemes considered in that paper, the based-pair weighted score BWS_1, which scores palindromes according to how unlikely they can be observed in a random nucleotide sequence generated by a first order Markov chain, gives the most accurate predictions. To compare the ANN approach with the palindrome method side by side, we present the top three ANN predicted replication origin locations for each of the herpesviruses listed in Table III along with the top three

predictions made by Chew et al. as reported in Table I of [15], using the BWS_1 palindrome scoring scheme.

Among the herpesviruses in Table III, each of the first 19 has at least one known replication origin [15]. We shall use these 19 genomes for assessing the prediction accuracy of the ANN approach. Although the remaining 20 viruses in Table III do not have any documented replication origin, we also report the predictions by ANN and BWS_1 as this information can assist biomedical researchers identify and confirm the replication origins of these viruses in their laboratories.

As in [15], we consider a prediction successful if the predicted location is within two map units of a known replication origin. 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; and positive predictive value is the percentage of predicted locations that are close to the known origins. A predicted location is considered close to a known origin if it is within two map units. The sensitivity and PPV from the 19 neural networks are displayed in Fig. 2, along with the same measures for the BWS_1 scheme.

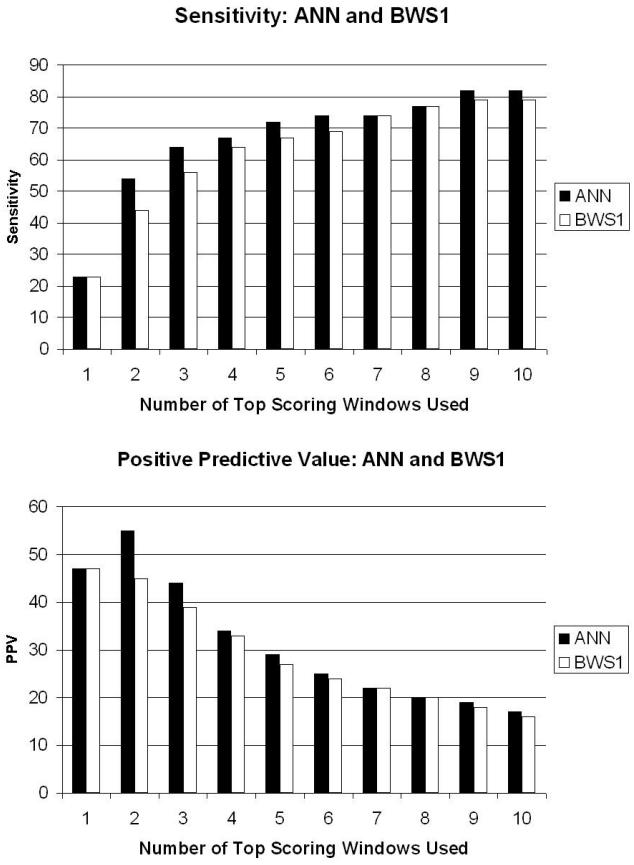

Fig. 2. Sensitivity and PPV (in percentage) using 1-10 top scoring windows

Usually there are three replication origins in a herpesvirus genome but in some exceptional cases, there are six.

TABLE III. FIRST 3 PREDICTIONS FOR BOTH BWS_1 AND ANN METHODS

Virus	BWS_1 Rankings			ANN Rankings		
	1	2	3	1	2	3
BoHV1	113401	124501	887301	105367	127219	110755
BoHV4	54751	30215	72251	96637	22474	5993
BoHV5	18901	113401	129601	130002	112629	68297
CeHV1	133001	149451	61601	150140	132640	123892
CeHV2	129501	144201	61601	127986	143722	123440
CeHV7	18601	106201	121801	109445	118740	104048
EBV	7601	53201	127601	51164	69951	150696
EHV1	116201	147001	47601	142539	121228	73715
EHV4	105351	143151	109901	128098	136847	44449
GaHV1	68601	41651	99751	121748	137492	144488
HCMV	94501	174901	196351	214897	189616	93982
HHV6	8051	30101	110601	131370	68479	142200
HHV6b	90801	132801	8801	69477	128973	157722
HHV7	9451	152251	133351	145740	9436	128615
HSV1	62301	129851	148401	130259	147372	9429
HSV2	74551	28001	12951	148109	133437	127150
RCMV	75901	110551	83601	134019	201028	213660
SHV1	38151	11551	93101	16445	127016	117568
VZV	119401	110101	100501	20665	120092	109311
AlHV1	113701	123301	32701	119224	114132	127612
AtHV3	99251	97001	54751	102165	98168	7244
CalHV3	116201	133351	23101	116818	22734	56660
CCMV	91201	207001	177001	176917	90557	199107
CeHv15	8001	34801	138801	136317	7995	34779
CeHV8	161151	147401	198001	161557	176944	88471
EHV2	54001	6301	173251	12595	150690	139895
GaHV2	160801	801	137601	137744	146928	52303
GaHV3	158801	138401	11201	122703	131096	10791
HCMV-M	175451	94051	153451	94477	89533	175772
HHV8	23401	119701	136501	124626	118335	24266
IcHV1	55501	89701	9301	62320	123140	126136
MCMV	92951	142451	200201	92880	184112	197302
MeHV1	5601	117951	11551	152513	119633	38128
MMRV	132601	117601	3301	116029	26384	130421
MuHV4	99251	26251	62001	100708	26739	70471
OsHV1	21001	144001	187501	184946	196443	146456
PSHV1	130401	151601	18801	151437	141049	87505
SaHV2	103751	112501	81501	96190	67708	5746
THV	134101	10801	144901	107813	189121	169804

We are therefore particularly interested in the prediction performance when using three to six top scoring windows.

In Fig. 2, the ANN predictions show slightly better sensitivity and PPV than the BWS_1. More interestingly, when examining the listed predictions by the two methods in Table III, one can see that the prediction performance can be substantially improved if we combine the two sets of prediction results appropriately. With three top-scoring windows, ANN and BWS_1 can only attain a sensitivity of 64% and 57% respectively.

However, when we combine the two sets of predicted locations, 30 out of the 39 known replication origins are located, giving a sensitivity of 77%, which surpasses the sensitivity of either one of the individual prediction schemes even using six top-scoring windows. Furthermore, we can see from Table III that there are 15 "joint" predictions (shaded) which refer to the same locations predicted by both the ANN and BWS_1. (Again, we consider two predicted locations to be the same when they are within two map units from each other.)

Among the 15 joint predictions, 13 are known origins, giving a PPV of 87%. This suggests that a jointly predicted origin from ANN and BWS_1 will highly likely be a true replication origin. It would also be of interest to see if the BWS_1 palindrome score can be a good input variable in the ANN approach. On examining the BWS_1 scores over a variety of windows in the herpesvirus genomes, we notice that this variable contains much inconsistency which can easily confuse the network during the training process. For example, among those windows with zero BWS_1 score, some are close to ORI and others are not. So, corresponding to the same value of this input variable, some of the outputs 1 and others 0. Since ANN training algorithms require that for the same input vector, the same output is always obtained, BWS_1, by itself, is therefore not a suitable input variable in the ANN approach. Moreover, our preliminary experiments show that the addition of the BWS_1 to the 17-variable overall data set described above does not improve the performance of the neural networks for a test set composed of randomly selected windows. The above observation leads to the question of what sequence characteristics can best serve as input variables to the ANN approach. This relates to the feature extraction issue which is quite a common problem in ANN applications.

VI. CONCLUDING REMARKS

This paper has demonstrated the contribution of ANN to the prediction of viral DNA replication origins. While we have only applied this approach to the herpesviruses, it should also work well for other viral families as long as a set of confirmed origins have been identified for some members of the family. However, when dealing with a new sequence from a viral family with no known replication origins, the ANN approach will not work because no training data is available. In such situations, we still have to rely on methods like the palindrome based schemes that use only features within the new sequence.

Apart from the 17 input variables we have used in this study, other sequence characteristics, such as distribution of close direct and inverted repeats, variations in percentages of A and T bases, DNA asymmetry, flanking sequence similarity, etc., have been reported to be relevant ([14], [34], [35] and [36]) to replication origin prediction in a variety of viral, bacterial, archaeal, and eukaryotic genomes. It is possible that some or all of these features can be incorporated to the ANN to further improve the prediction performance. The problem of how to select an optimal collection of sequence features to be included for replication origin prediction still remains to be further investigated.

ACKNOWLEDGEMENTS

This research is supported in part by NIH grants S06GM08012-35, 2G12RR008124-11 and 3T34GM008048-20S1.

REFERENCES

[1] J.J Bennett, J. Tjuvajev, P. Johnson, M. Doubrovin, T. Akhurst, S. Malholtra, T. Hackman, J. Balatoni, R. Finn, S.M. Larson, H. Federoff, R. Blasberg, and Y.Fong, "Positron emission tomography imaging for herpes virus infection: Implications for oncolytic viral treatments of cancer," *Nat. Med.*, 7(7), pp. 859–863, 2001..

[2] J. Biswas, S. Deka, S. Padmaja, H.N. Madhavan, N. Kumarasamy, and S. Solomon, "Central retinal vein occlusion due to herpes zoster as the initial presenting sign in a patient with acquired immunodeficiency syndrome (AIDS)," *Occl. Immunol. Inflamm.,* 9(2), pp. 103–109, 2001.

[3] L.G. Labrecque, D.M. Barnes, I.S. Fentiman and B.E. Griffin, "Epstein-Barr virus in epithelial cell tumors: A breast cancer study," *Cancer Res.*, 55(1), pp. 39–45, 1995.

[4] C. Vital, E. Monlun, A. Vital, M.L. Martin-Negrier, V. Cales, F. Leger, M. Longy-Boursier, M. Le Bras and B. Bloch, "Concurrent herpes simplex type 1 necrotizing encephalitis, cytomegalovirus ventriculoencephalitis and cerebral lymphoma in an AIDS patient," *Acta pathologica*, 89(1), pp. 105–108, 1995.

[5] H.J. Delecluse, and W. Hammerschmidt, "The genetic approach to the Epstein-Barr virus: From basic virology to gene therapy," *J. Clin. Pathol. Mol. Pathol.*, 53, pp. 270–279, 2000.

[6] C.B. Hartline, E.A. Harden, S.L. Williams-Aziz, N.L Kushner, R.J. Brideau and E.R. Kern, "Inhibition of herpesvirus replication by a series of 4-oxo-dihydroquinolines with viral polymerase activity," *Antiviral Res.*, 65, pp. 97-105, 2005.

[7] E.C. Villarreal, "Current and potential therapies for the treatment of herpes-virus infections," *Prog Drug Res.,* 60, pp. 263-307, 2003.

[8] Y. Zhu, L. Huang and D.G. Anders, "Human cytomegalovirus oriLyt sequence requirements," *J. Virol.,* 72, pp. 4989-4996, 1998.

[9] C.S. Newton, and J.F. Theis, "DNA replication joins the revolution: whole genome views of DNA replication in budding yeast," *BioEssays*, 24, pp. 300-304, 2002.

[10] H. Deng, J.T Chu, N. Park and R. Sun, "Identification of cis Sequences Required for Lytic DNA Replication and

Packaging of Murine Gammaherpesvirus 68," *J. Virol*., 78, pp. 9123-9131, 2004.

[11] M. Y. Leung., K.P. Choi, A. Xia, and L.H.Y Chen, "Nonrandom clusters of palindromes in herpesvirus genomes*," J. Computational Biol.,* 12, pp. 331-354, 2005.

[12] S.K. Weller, A. Spadaro, J.E. Schaffer, A.W. Murray, A.M. Maxam and P.A. Schaffer, "Cloning, sequencing, and functional analysis of oriL, a herpes simplex virus type 1 origin of DNA synthesis," *Mol. Cell. Biology*, 5, pp. 930-942, 1985.

[13] D. Reisman, J. Yates and B. Sugden, "A putative origin of Replication of plasmids derived from Epstein-Barr virus is composed of two cis-acting components," *Mol. Cell. Biology*, 5, pp. 1822-1832, 1985.

[14] M.J. Masse, S. Karlin, G.A. Schachtel, and E.S. Mocarski, "Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region," *Computational Chemistry*, 20, pp. 135-140, 1992.

[15] D.S.H. Chew, K.P. Choi, and M.Y. Leung, "Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpes viruses," *Nucleic Acids Research,* 33(15), e134, 2005.

[16] A. J. Cann, *Principles of Molecular Virology*, 4th Edition, Elsevier Academic Press, San Diego, 2005.

[17] M.Y. Leung, G.M. Marsh and T.P. Speed, "Over and under representation of short oligonucleotides in herpesvirus genomes," *J. Computational Biology*, 3(3), pp. 345-360, 1996.

[18] C. B. Chen and T. Li, "A hybrid neural network system for prediction and recognition of promoter regions in human genome," *J. Zhejiang Univ Sci B*., 6(5), pp. 401-407, 2005.

[19] S. Matis, Y. Xu, X. Guan, J.R. Einstein, R. Mural and E. Uberbacher, "Detection of RNA polymerase II promoters and polydenlation sites in human DNA," *Proc. Natl. Acad. Sci.* USA., 89, pp. 5246-5250, 1996.

[20] X. Huang, D.S. Huang, G.Z. Zhang, Y.P. Zhu and Y.X. Li, "Prediction of protein secondary structure using improved two-level neural network architecture," *Protein Pept Lett.*, 12(8), pp. 805-811, 2005.

[21] S. Qian and T.J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models", *Journal of Molecular Biology*, 202, pp. 865-884, 1988.

[22] W.R. Weinert and H.S. Lopes, "Neural networks for protein classification", *Appl Bioinformatics*, 3(1), pp. 41-48, 2004.

[23] C. Wu, G. Whitson, J. McLarty, A. Ermongkonchai and T.C. Chang, "Protein Classification artificial neural system", *Protein Science*, 1, pp. 667-677, 1992.

[24] K. Julenius, A. Mølgaard, R. Gupta, and S. Brunak, "Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites," *Glycobiology*, 15(2), pp. 153–164, 2004.

[25] B. Widrow and D. Nguyen, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," *Int. Joint Conf. on Neural Networks*, 21-26, 1990.

[26] W. Duch and N. Jankowski, "Survey of neural transfer functions," *Neural Computing Surveys*, 2, pp. 163–213, 1999.

[27] M.T. Hagan and M.B. Menhaj, "Training Feedforward neural networks with the Marquardt Algorithm," *IEEE Transaction on Neural Networks*, 5(6), pp. 989-993, 1994.

[28] P.V. Nazarov, V.V. Apanasovich, V.M. Lutkovski, M.M. Yatskou, R.B.M. Koehorst and M.A. Hemminga, "Artificial Neural Network Modification of Simulation-Based Fitting: Application to a Protein-Lipid System*," J. Chem. Inf. Comput. Sci*., 44, pp. 568-574, 2004.

[29] A. Zeng, Q.-L. Zheng, D.Pan and H.Peng, "Utilizing Modular Neural Networks to Predict MHC Class 11-binding Peptides," *2004 IEEE International Conference on Systems, Man and Cybernetics,* pp. 4588-4592, 2004

[30] R. Cruz-Cano, R. Cabeza, P. Nava and E. Martin del Campo, "Fuzzy Rule Extraction and Optimization for Rat Sleep-Stage Classification," *Proceedings of the 2005 International Conference on Machine Learning: Models, Technologies and Applications*, pp. 75-79, 2005.

[31] Y. Xu and E.C. Uberbacher, "Computational gene prediction using neural networks and similarity search," *Computational Methods in Molecular Biology*, (Eds. Salzberg, S.L., Searls, D.B. and Kasif,S.), pp. 109-128, Elsevier Science B. V., 1998.

[32] I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[33] H. Demuth and M. Beale, *Neural Network Toolbox (Version 4) For Use with MATLAB*, The Math Works, Inc., Natick, 2004.

[34] A.M. Breier, S. Chatterji and N.R. Cozzarelli, "Prediction of Saccharomyces cerevisiae replication origins", *Genome Biol*, 5, R22, 2004.

[35] P. Mackiewicz, J. Zakrzewska-Czerwinska, A. Zawilak, M.R. Dudek and S. Cebrat, "Where does bacterial replication start? Rules for predicting the oriC region", *Nucleic Acids Res.,* 16, pp. 3781–3791, 2004.

[36] S.L. Salzberg, A.J. Salzberg, A.R. Kerlavage and J-F. Tomb, "Skewed oligomers and origins of replication," Gene, 217, pp. 57–67, 1998.