AGRICULTURE AND NATURAL RESOURCES
Journal homepage: http://anres.kasetsart.org

Research article

Classifying DNA barcode sequences of four insects belonging to Orthoptera order using tensor network

Pradeep Bhadola[a,†], Yash Munnalal Gupta[b,†,*]

[a] *Centre for Theoretical Physics & Natural Philosophy, Nakhonsawan Studiorum for Advanced Studies, Mahidol University, Nakhon Sawan Campus, Phayuha Khiri, Nakhon Sawan 60130, Thailand*

[b] *Department of Biology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand*

Supplementary material

Tensor notations

In tensor network notation, as proposed by Roger Penrose in 1971, a tensor is represented by a geometric shape called a node and the indices are denoted by edges (legs) extending from these shapes. A scalar is represented as a node (circle) without any legs as it has no dimension (index), a vector V_j is a 1-dimensional object with only 1 index and is represented as a node with 1 edge (line). A matrix M_{ij} is 2-dimensional object with two indices *i* and *j* and is represented by a node with two edges. Similarly, a rank 3 tensor T_{ij} is denoted as a node with 3 edges. An n-rank tensor will be represented by a node with *n* edges. Tensor network notation makes it easier to follow the tensor operation. Matrix product or tensor contraction is the most common tensor operation and is inspired by the Einstein summation convention for tensor contractions. The general rule for contracting tensors is that the two edges (lines) connected with each other implies a contraction, or summation, over the connected indices. Fig. S1A shows the Penrose graphical representation of scalar, vector, matrix and a 3-rank tensor. Fig. S1B shows some examples of tensor contraction. The first example is the multiplication of a matrix *M* with a vector *V* along the direction of index *j*. The resultant is the vector *U* with dimension *i*. The multiplication of two matrices *A* and *B* results in a matrix *C* with dimension *i* being the same as the rows of *A* and the *k* columns of *B*. The summation sign is omitted, as in summation convention the repeated indexes are summed over. The third example is the trace of the product of two matrices, which results in a scalar S. The last example is the product of a matrix with a rank-3 tensor resulting in another rank 3 tensor.

Matrix product states (MPS) are the best understood tensor networks. MPS (also known as tensor train) are a family of tensors where a large tensor with N dimensions (N edge index) is factorized as a linear chain of small rank 3 tensors (3 index tensor) in the center and rank 2 tensors at both ends, as shown in Fig. S2. In Fig. S2, a 5-dimensional tensor is factorized into 5 small dimensional tensors (3 tensor of 3 dimension and 2 boundary tensors of 2 dimensions). This factorization substantially reduces the total number of parameters. The above factorization can be generalized for a tensor with any number of dimensions (indices). Mathematically, tensor factorization of a N dimensional tensor T with indices $i_1, i_2, \cdots, i_N$ can be approximated as the product of lower rank tensors *A's*, as shown in Equation 1:

$$T^{i_1,i_2,\ldots,i_N} = \sum_{\alpha_1,\alpha_2\ldots\alpha_N} A^{i_1}_{\alpha_1} A^{i_2}_{\alpha_1\alpha_2} \ldots A^{i_j}_{\alpha_{j-1}\alpha_j} \ldots A^{i_{N-1}}_{\alpha_{N-2}\alpha_{N-1}} A^{i_N}_{\alpha_{N-1}} \qquad (1)$$

† Equal contribution.

* Corresponding author.
E-mail address: yashmunnalalg@nu.ac.th (Y.M. Gupta)

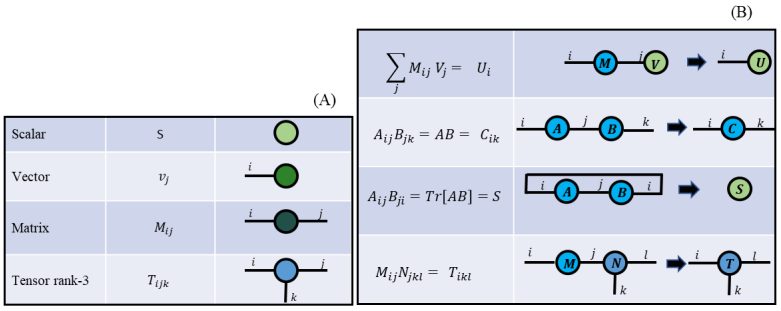

Fig. S1 (A) Tensor notation representing a scalar, vector, matrix and a third order tensor; (B) rules for tensor contraction used in tensor network *Matrix product states*

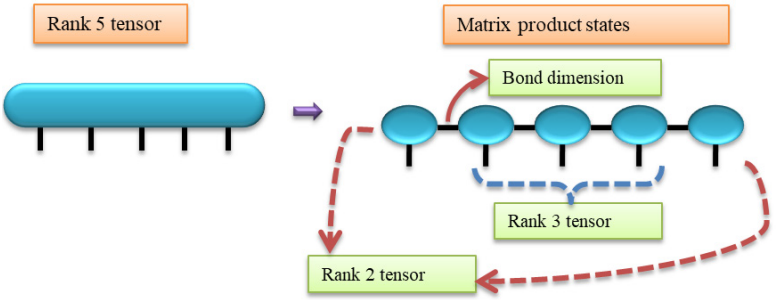

Fig. S2 Matrix product states factorization of rank 5 tensors

where α is known, as the bond indices are contracted or summed over. All the A's are different from each other and are distinguished from each other by the indices. Equation 1 can be represented in the MPS form, as shown in Fig S2. MPS are very successful in explaining the quantum system (especially one-dimensional systems) but recently have been very effective in fields, such as compressing high-dimensional data, machine-learning applications, such as a supervised kernel linear classifier, and in unsupervised generative modelling. MPS can be successfully applied to machine-learning task as they can effectively capture 1-dimensional correlations in the system but also, they can be modified for systems with higher dimensional correlations.

The tensor operation of contracting the bond indices α in Equation 1 will result in a good approximation of the tensor T for a sufficiently large value of α. The dimension of α is known as the bond index, sometimes called as tensor-train rank or virtual dimension. The bond dimension (m) can be considered as a parameter that controls the expressivity of the MPS. For a tensor $T^{i1,i2,\ldots,iN}$ with all N indices having the same dimension d, then, this tensor can be exactly represented as an MPS by choosing the bond dimension $m = d^{N/2}$. The advantage of replacing a big tensor T into an MPS is the reduction in the number of parameters required to specify the tensor. Given a tensor T with N indices each of dimension d, then d^N parameters are required to specify the tensor. Whereas if the tensor is approximated into an MPS with bond dimension m, then the number of parameters required is $2md$ from ends and $(N-2)md^2$ from the center. For large N, the parameters can be approximated as Ndm^2. The conversion of a tensor to MPS represents a great compression in the number of parameters as in the first case the parameters grow exponentially (d^N) with N, whereas in MPS, it grows linearly (Ndm^2) with N. The bond dimension is a parameter that can be chosen to best suit the need of the problem.

Plots for different number of units and k-mers

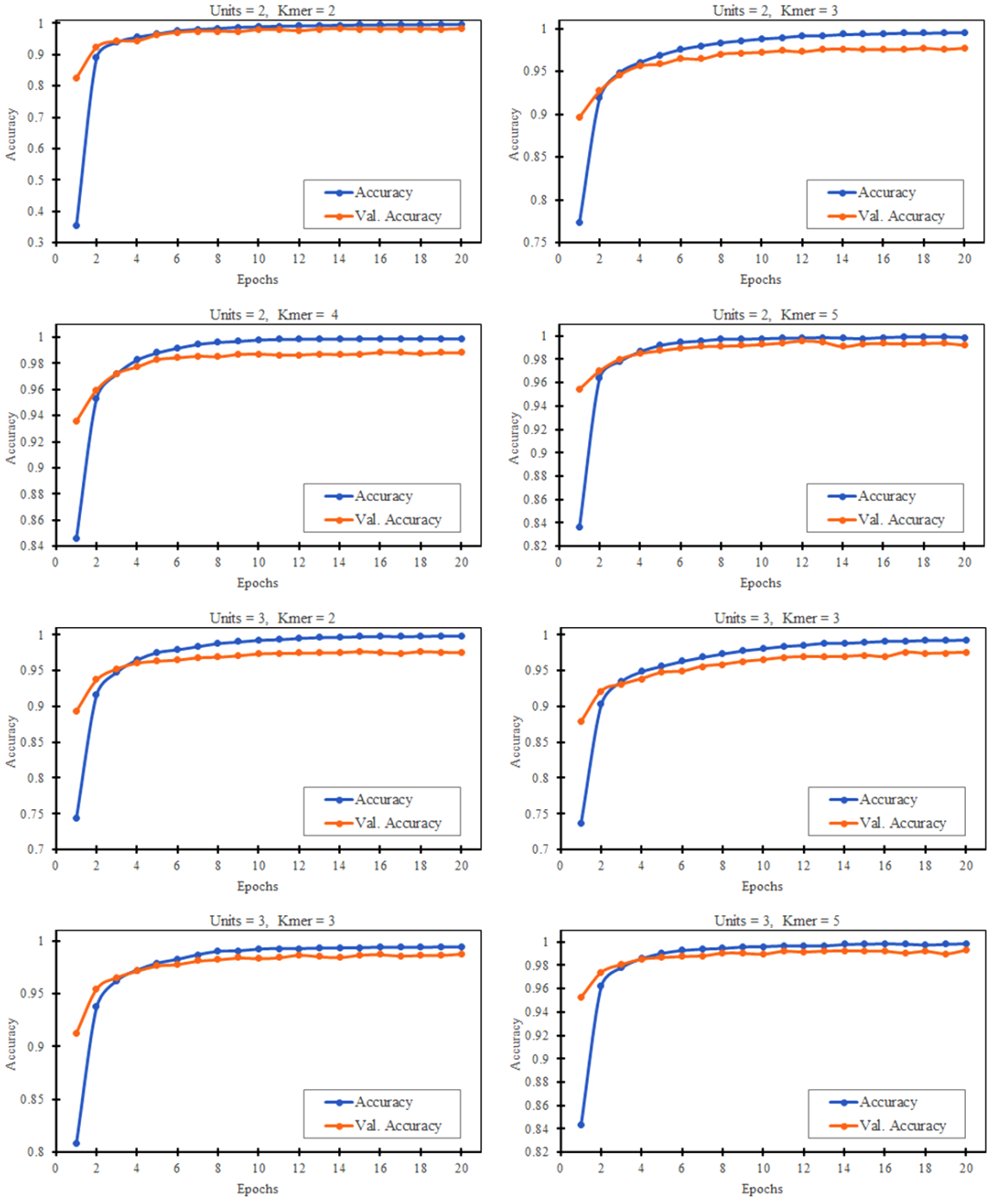

Fig. S3 Training and validation accuracy of the tensor network model with 2 and 3 units and *k*-mer in range 2–5

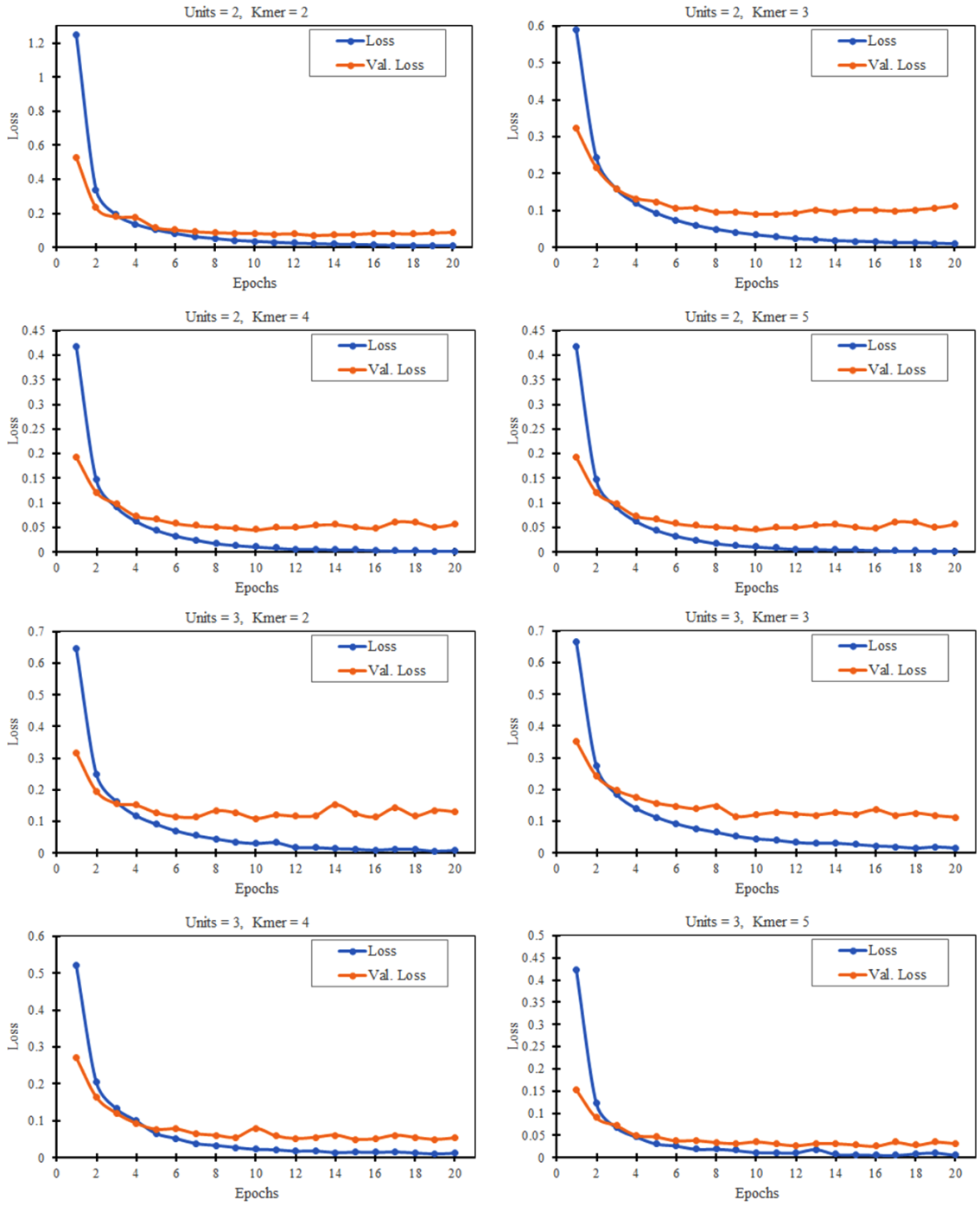

Fig. S4 Loss and validation loss of tensor network model for 2 and 3 units and *k*-mer in range 2–5

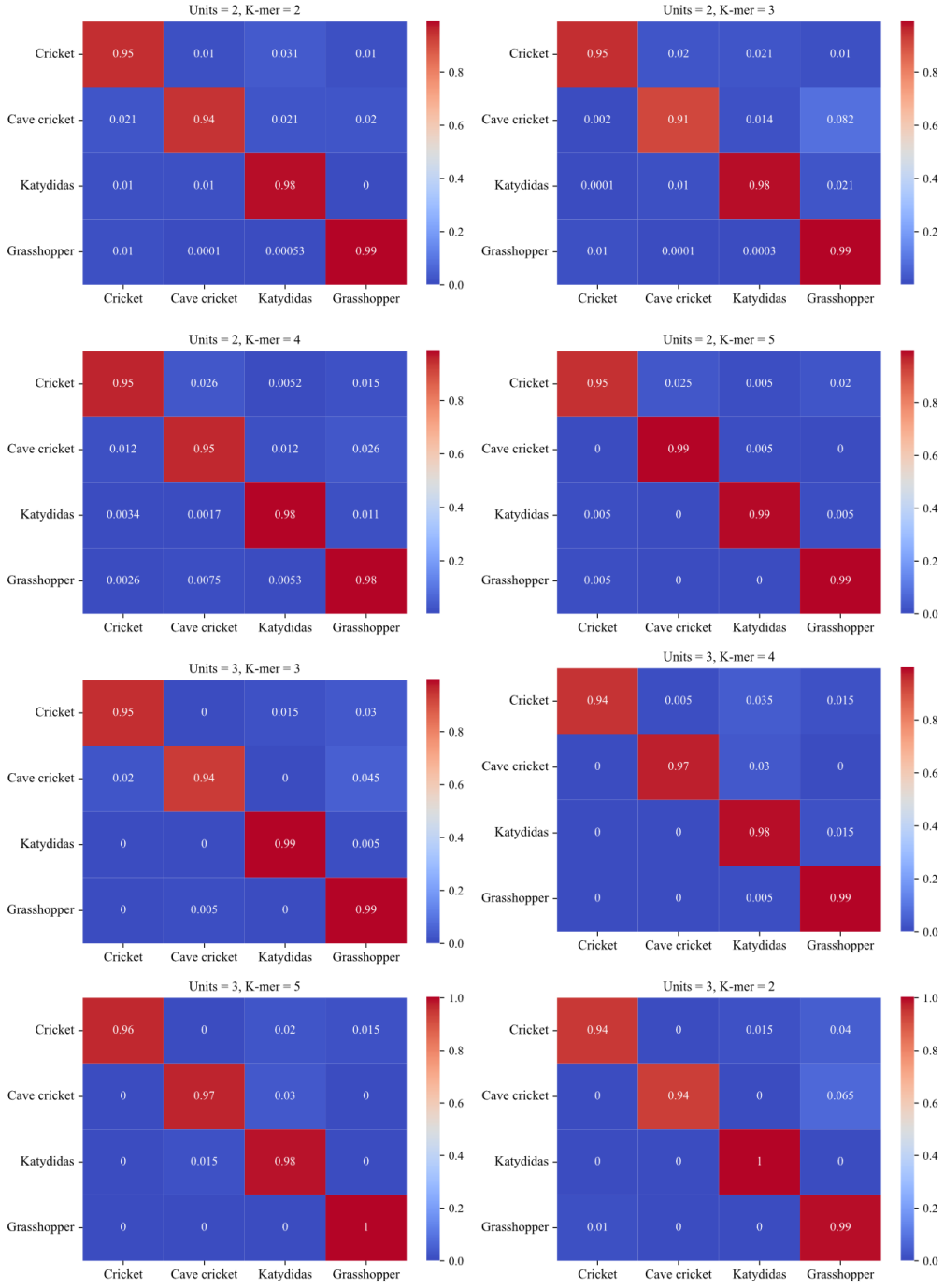

Fig. S5 Normalized confusion matrix for prediction on test set for tensor network for units 2 and 3