Bacteriocin detection with distributed biological sequence representation

Md Nafiz Hamid[1,2], Iddo Friedberg[2]

[1]Program in Bioinformatics and Computational Biology, [2]Department of Veterinary Microbiology and Preventive Medicine
Iowa State University, Ames, Iowa, USA

What are Bacteriocins?

Bacteriocins are peptide-derived molecules produced by bacteria that function as virulence factors, signaling molecules, and antimicrobials. They are surrounded by context genes that are responsible for the translation, modification, transport and self-immunity from the bacteriocin.

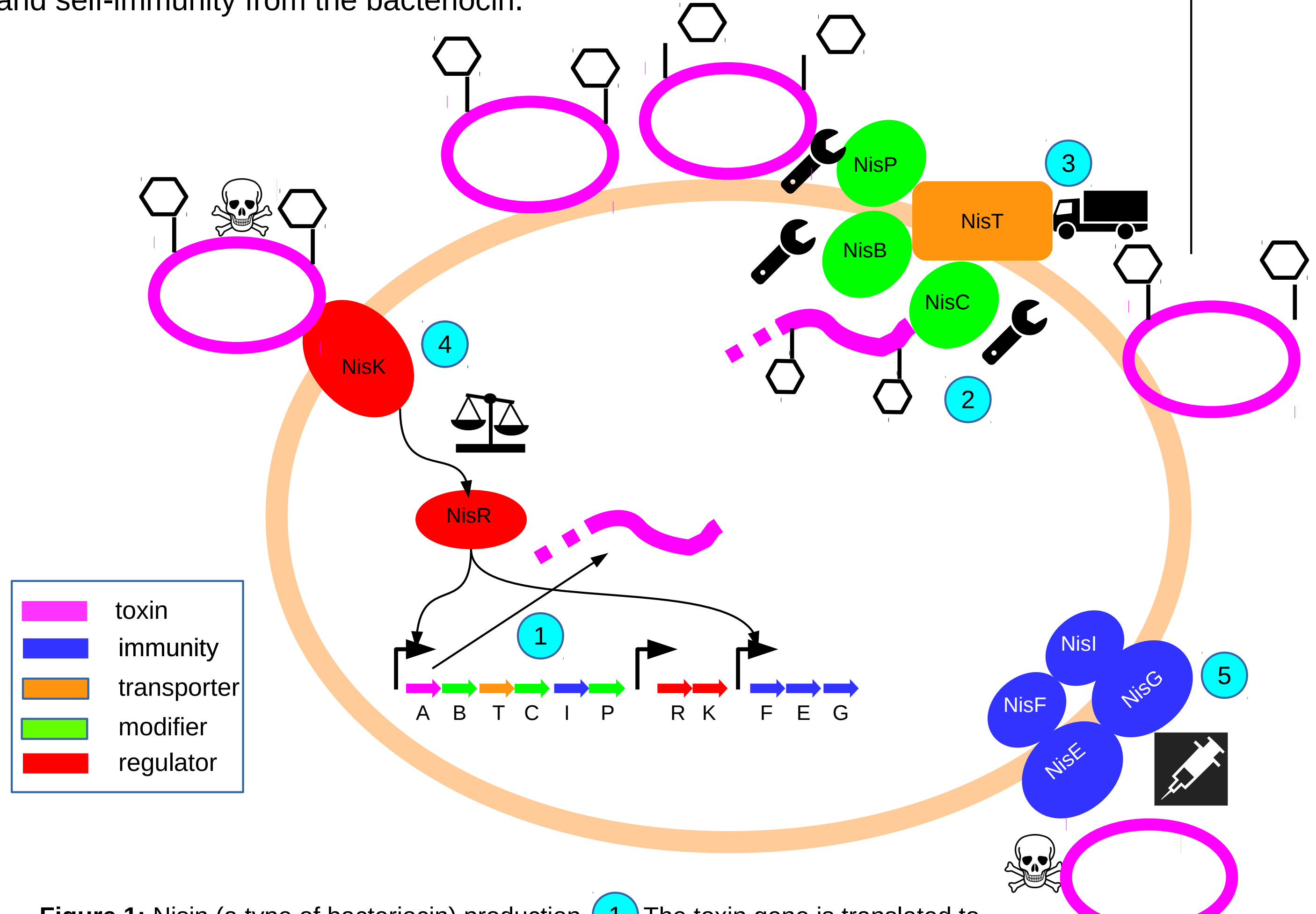
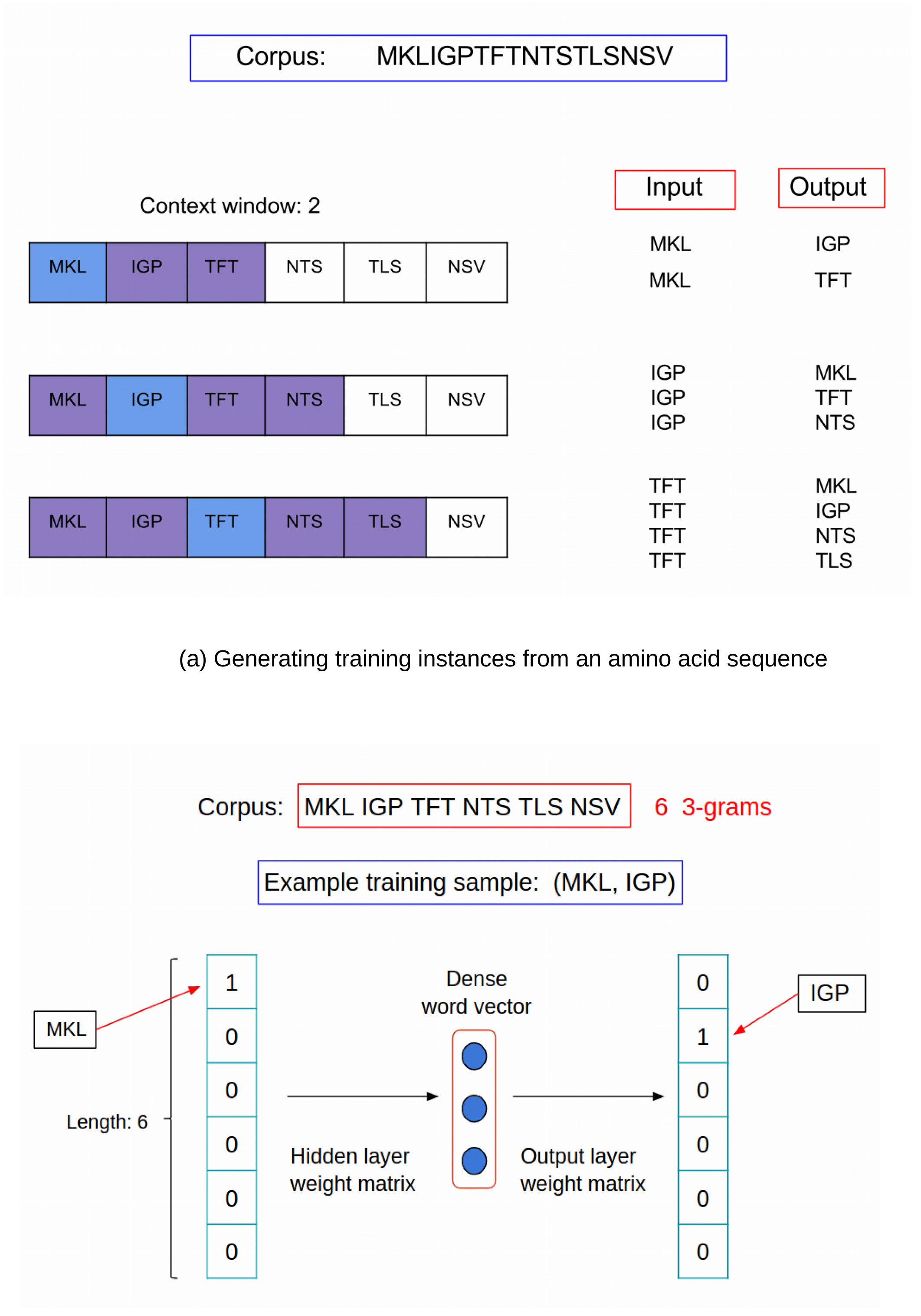


Figure 1: Nisin (a type of bacteriocin) production. 1 The toxin gene is translated to bacteriocin precursors 2 precursor bacteriocin is post-translationally modified by modifier genes, and turned into their biologically active forms 3 the precursor bacteriocin is exported by transporter genes 4 regulator genes control the production of bacteriocins 5 immunity genes protect the bacteria producing the bacteriocin from the toxin. These context genes have been shown to be largely conserved across unrelated species.

Can we find bacteriocins from amino acid sequences?

- **Problem**: homology-based methods are limited, as bacteriocins are diverse in sequence and contain many repeats.
- **Possible solution**: use word embedding rather than classic string representation
- The main challenge is a small training dataset (346 positive bacteriocin sequences from the BAGEL dataset)
- An informative representation of the sequences could take advantage of this small positive and a similar small negative dataset for supervised classification.

Representation learning with Word2vec



Corpus: MKLIGPTFTNTSTLSNSV

Context window: 2

| MKL | IGP | TFT | NTS | TLS | NSV |

Input	Output
MKL	IGP
MKL	TFT
IGP	MKL
IGP	TFT
IGP	NTS
TFT	MKL
TFT	IGP
TFT	NTS
TFT	TLS

(a) Generating training instances from an amino acid sequence

Corpus: MKL IGP TFT NTS TLS NSV 6 3-grams

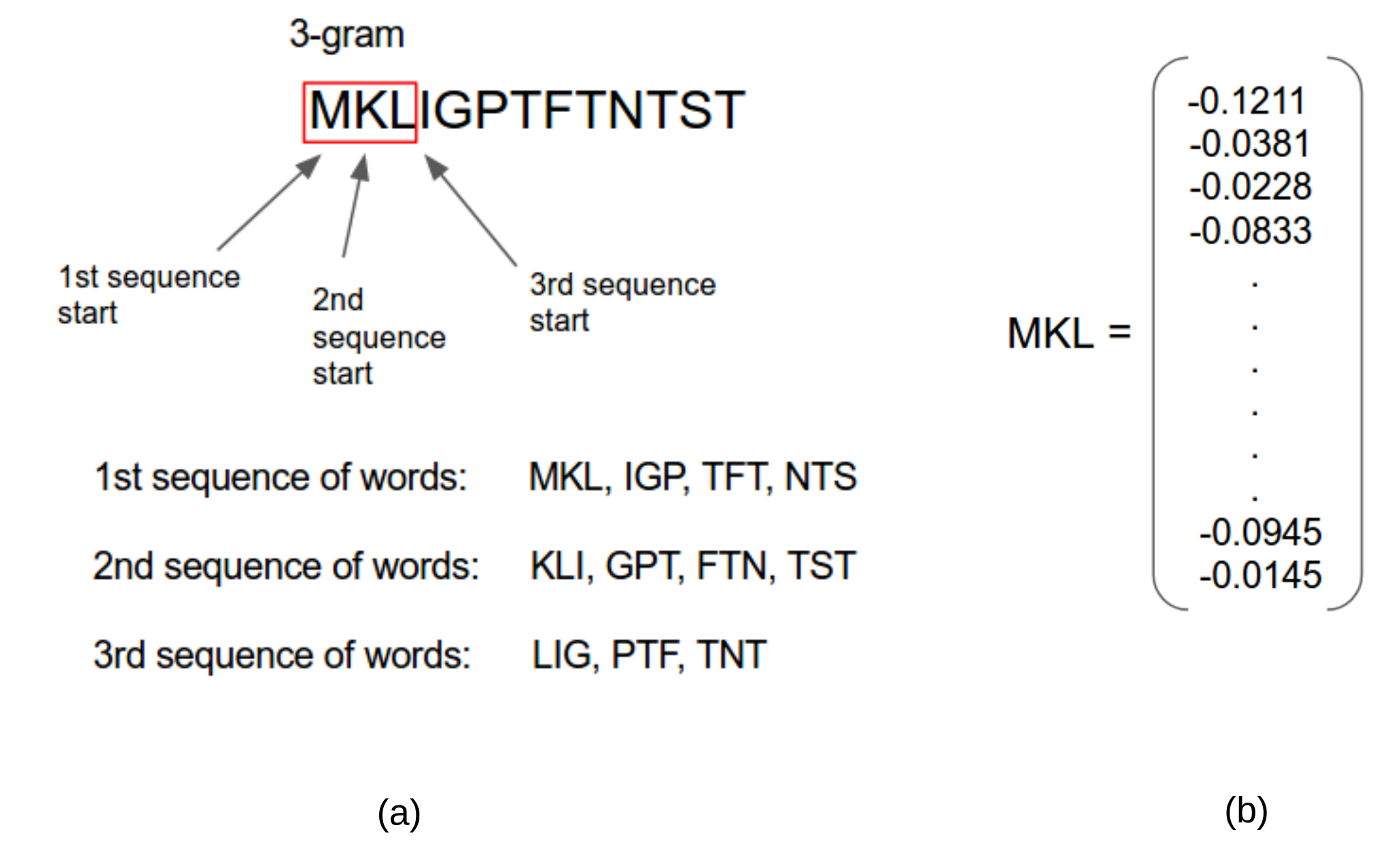
Example training sample: (MKL, IGP)

MKL
Length: 6
1 0 0 0 0 0
Dense word vector
Hidden layer weight matrix
Output layer weight matrix
0 1 0 0 0 0
IGP

(b) Neural network architecture for a single training instance

Figure 2: Representation learning for 3-grams with skip-gram training. (a) Training: a protein sequence (top left), is broken into 3-grams with each 3-gram associated with the previous and following two 3-grams. (b) the neural network architecture for MKL as input, and IGP as output.

Word2vec corpus - Uniprot TrEMBL bacteria database



3-gram
MKLIGPTFTNTST
1st sequence start
2nd sequence start
3rd sequence start

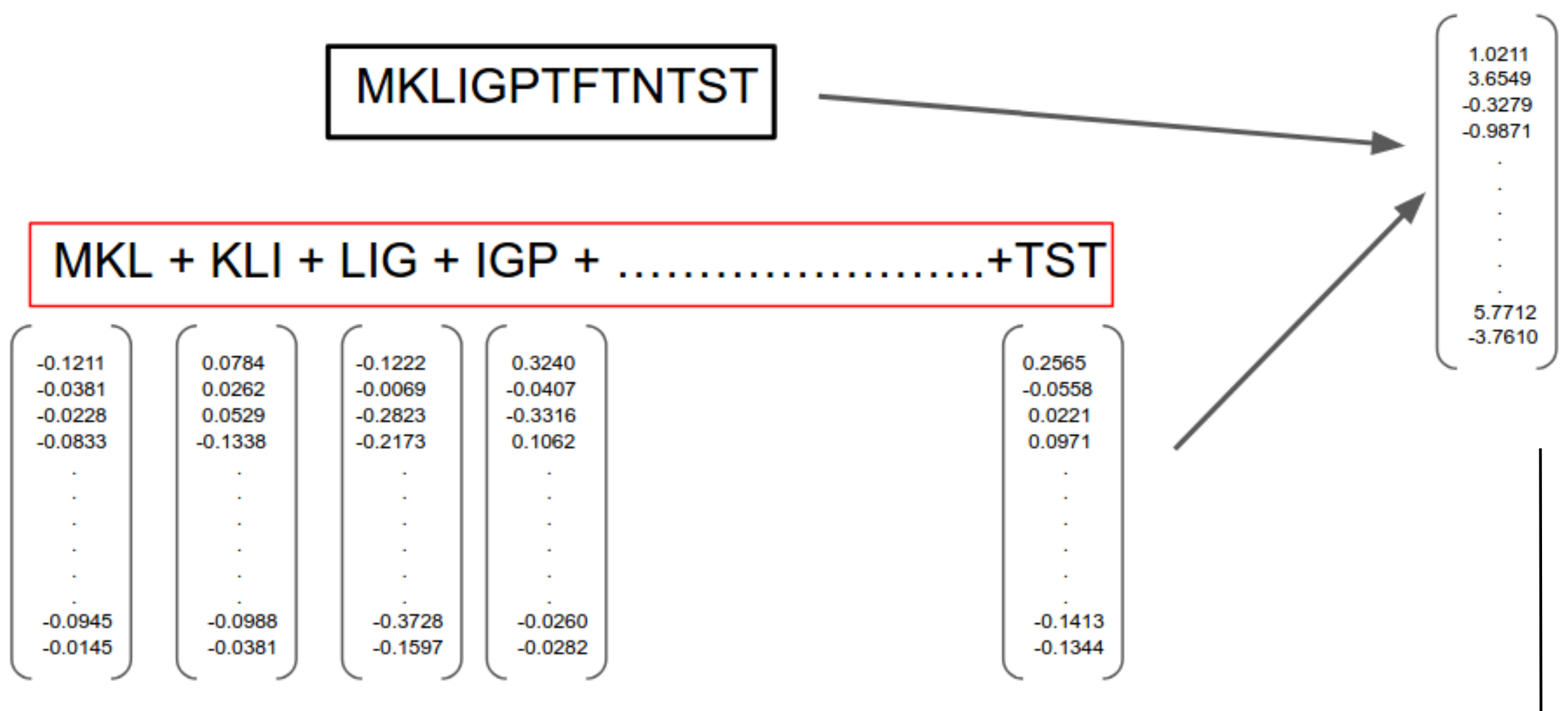
1st sequence of words:	MKL, IGP, TFT, NTS
2nd sequence of words:	KLI, GPT, FTN, TST
3rd sequence of words:	LIG, PTF, TNT

$$\text{MKL} = \begin{pmatrix} -0.1211 \\ -0.0381 \\ -0.0228 \\ -0.0833 \\ \vdots \\ -0.0945 \\ -0.0145 \end{pmatrix}$$

(a) (b)

Figure 3: Corpus is of 55,899,422 sequences. Vocabulary is of all 3-grams (20^3 = 8,000 words) (a) We generated 3 sequences from each sequence in TrEMBL bacteria database, and used all of them in training; (b) the final representation of a trigram as a size 200 dense vector.

Sequence representation with trigrams



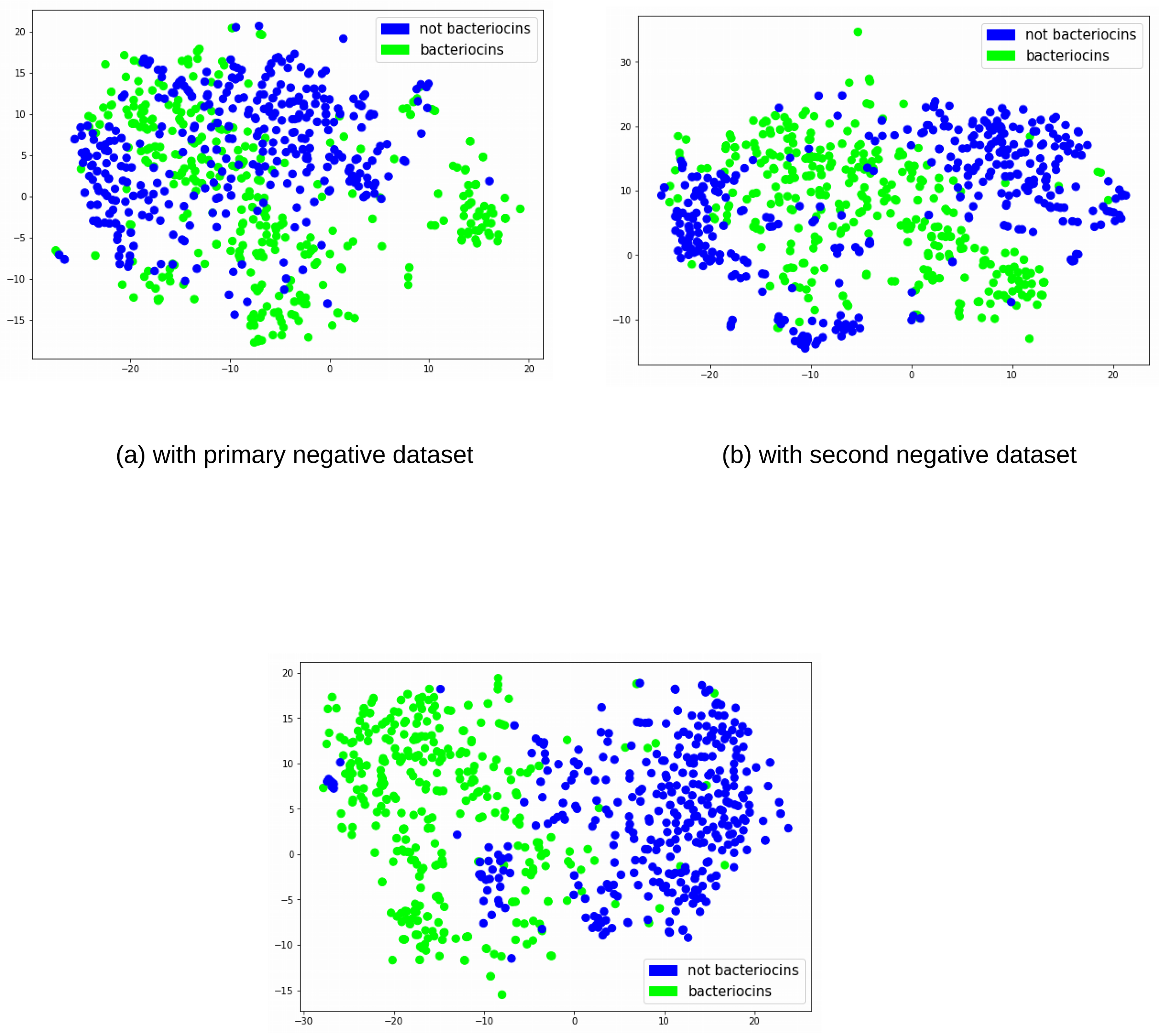
MKLIGPTFTNTST

MKL + KLI + LIG + IGP + ...+TST

$$\begin{pmatrix} -0.1211 \\ -0.0381 \\ -0.0228 \\ -0.0833 \\ \vdots \\ -0.0945 \\ -0.0145 \end{pmatrix} \begin{pmatrix} 0.0784 \\ 0.0262 \\ 0.0529 \\ -0.1338 \\ \vdots \\ -0.0988 \\ -0.0381 \end{pmatrix} \begin{pmatrix} -0.1222 \\ -0.0069 \\ -0.2823 \\ -0.2173 \\ \vdots \\ -0.3728 \\ -0.1597 \end{pmatrix} \begin{pmatrix} 0.3240 \\ -0.0407 \\ -0.3316 \\ 0.1062 \\ \vdots \\ -0.0260 \\ -0.0282 \end{pmatrix} \begin{pmatrix} 0.2565 \\ -0.0558 \\ 0.0221 \\ 0.0971 \\ \vdots \\ -0.1413 \\ -0.1344 \end{pmatrix} \rightarrow \begin{pmatrix} 1.0211 \\ 3.6549 \\ -0.3279 \\ -0.9871 \\ \vdots \\ 5.7712 \\ -3.7610 \end{pmatrix}$$

Figure 4: Each amino acid sequence is represented by the sum of the overlapping tri-grams.

Construction of a negative dataset



(a) with primary negative dataset

(b) with second negative dataset

(c) with third negative dataset

Figure 5: t-sne visualization of word vector representations for positive and negative bacteriocin amino acid sequences. We used the manually reviewed Uniprot Swissprot bacteria database to create 3 negative datasets. For the primary negative dataset, 346 sequences were taken that have the same length distribution as the positive bacteriocins. They were chosen such as not to be annotated as anti-microbial, or antibiotic.

Results

	Word2vec			*k*-mer		
Methods	Mean Precision	Mean Recall	Mean F1	Mean Precision	Mean Recall	Mean F1
SVM	**0.877** (±0.009)	**0.863** (±0.015)	**0.869** (±0.009)	**0.875** (±0.012)	0.808 (±0.017)	0.838 (±0.011)
Logistic Regression	0.850 (±0.012)	0.823 (±0.014)	0.834 (±0.009)	0.869 (±0.011)	**0.839** (±0.015)	**0.850** (±0.011)
Decision Tree	0.731 (±0.015)	0.747 (±0.020)	0.736 (±0.015)	0.755 (±0.015)	0.719 (±0.021)	0.733 (±0.014)
Random Forest	0.820 (±0.009)	0.813 (±0.013)	0.814 (±0.008)	0.833 (±0.052)	0.770 (±0.080)	0.797 (±0.053)

Table 1: Comparison between Word2vec and *k*-mer (*k* = 3) representation of reduced size of 200 by truncated SVD for 10-fold nested cross-validation with the primary negative dataset. Average precision, recall and F1 score are average of cross-validation done 50 times. Bold numbers are the best results for each representations. Overall SVM with the word2vec representation gives the best performance.

	Mean Precision	Mean Recall	Mean F1
Primary negative dataset	0.877	0.863	0.869
Second negative dataset	0.916	0.892	0.902
Third negative dataset	0.955	0.927	0.940

Table 2: SVM with Word2vec representation on three different negative sets

Discussion

- We applied the trained SVM model to potential areas of genomes from the Refseq database to predict novel bacteriocins.
- A total of 1186 putative bacteriocins in *Lactobacillus* with varying degrees of probability were found.
- We predicted 11 putative bacteriocins with a probability ≥ 0.95.
- Computational analysis of the surrounding region shows presence of context genes leading us to believe they might be Class II bacteriocins.

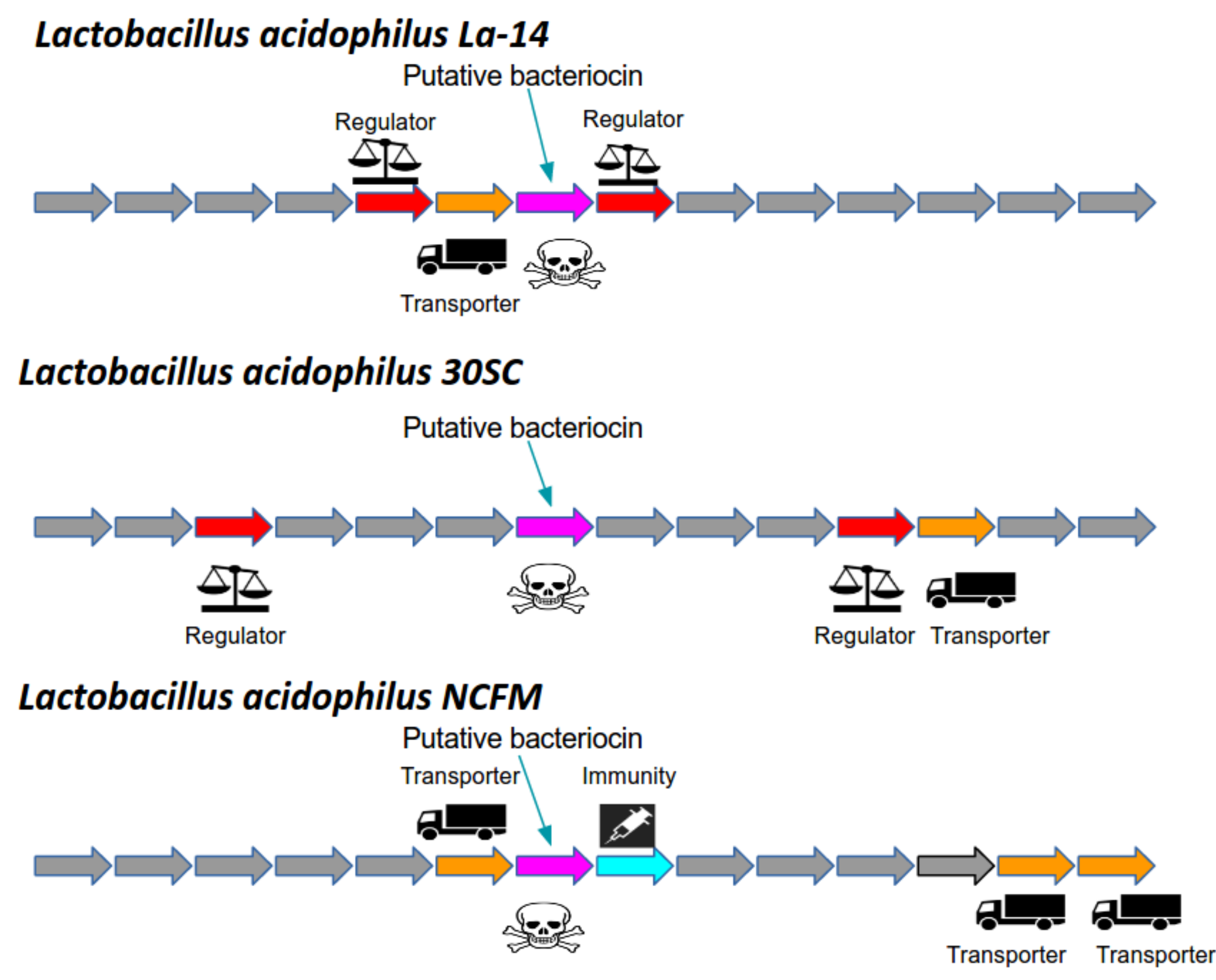


Figure 6: Context genes found surrounding the predicted bacteriocins within a ± 25kb range

Future Work

Experiments to verify some of the predictions are ongoing

References

- James T Morton, Stefan D Freed, Shaun W Lee and Iddo Friedberg. **A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins**. (2016) *BMC Bioinformatics* 201516:381
- Anne de Jong, Auke J. van Heel, Jan Kok and Oscar P. Kuiperds. **BAGEL2: mining for bacteriocins in genomic data** (2010) *Nucleic Acids Research, 38, W647–W651*

Contact

{nafizh, idoerg}@iastate.edu

This study was funded, in part by NSF award ABI Innovation 1551363 awarded to IF