PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme

Matheus Henrique Pimenta Zanon

Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio

19 de Novembro de 2021

PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme

Paper: **PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme**
Author: Li *et al.*
Journal: BMC Bioinformatics
Date: Sep. 2014
URL: `https://doi.org/10.1186/1471-2105-15-311`

Background

Note: This article is from 2014. From 2014 to the present day a lot of research has been done on the subject.

Background

Note: This article is from 2014. From 2014 to the present day a lot of research has been done on the subject.

- Long non-coding RNAs (lncRNAs, typically $>$200 nt), are of particular interest because they contribute to many important biological processes;

Background

Note: This article is from 2014. From 2014 to the present day a lot of research has been done on the subject.

- Long non-coding RNAs (lncRNAs, typically >200 nt), are of particular interest because they contribute to many important biological processes;
- It remains a challenge to distinguish mRNAs from lncRNAs;

Background

Note: This article is from 2014. From 2014 to the present day a lot of research has been done on the subject.

- Long non-coding RNAs (lncRNAs, typically >200 nt), are of particular interest because they contribute to many important biological processes;
- It remains a challenge to distinguish mRNAs from lncRNAs;
- LncRNAs show many features similar to mRNAs, such as poly(A) tails, splicing and approximate sequence length.

Background

- Several tools, such as CPC and PhyloCSF, have been developed based on known protein databases, intrinsic sequence features and sequence conservation properties.

Background

- Several tools, such as CPC and PhyloCSF, have been developed based on known protein databases, intrinsic sequence features and sequence conservation properties.
- A tool named Coding-Non-Coding Index (CNCI) was developed. It discriminates coding from non-coding transcripts using intrinsic sequence features.

Objective

- A characteristic k-mer based alignment-free tool named PLEK;

Objective

- A characteristic k-mer based alignment-free tool named PLEK;

PLEK takes calibrated k-mer frequencies of a transcript sequence as its computational features. With these features, the support vector machine (SVM) algorithm was used to build a binary classification model to separate lncRNAs from mRNAs.

Data description

Human protein-coding transcripts were downloaded from the **RefSeq** database (release 60) and human long non-coding transcripts were collected from **GENCODE v17**.

Data description

Human protein-coding transcripts were downloaded from the **RefSeq** database (release 60) and human long non-coding transcripts were collected from **GENCODE v17**.
There were $34,691$ protein-coding transcripts with the length of >200 nt in the human **RefSeq** dataset, and $22,389$ long (>200 nt) non-coding transcripts in the human **GENCODE** dataset.

Improved k-mer scheme

Approach: k-mer usage and sliding-windows with a one-nucleotide step-length to analyze each transcript.

Improved k-mer scheme

Approach: k-mer usage and sliding-windows with a one-nucleotide step-length to analyze each transcript.
A k-mer pattern is a specific string with k nucleotides, each can be A, C, G or T.

Improved k-mer scheme

Approach: k-mer usage and sliding-windows with a one-nucleotide step-length to analyze each transcript.

A k-mer pattern is a specific string with k nucleotides, each can be A, C, G or T.

For $k = 1$ to 5, had $4 + 16 + 64 + 256 + 1024 = 1,364$ patterns: 4 one-mer patterns, 16 two-mer patterns, 64 three-mer patterns, 256 four-mer patterns, and $1,024$ five-mer patterns.

Improved k-mer scheme

Approach: k-mer usage and sliding-windows with a one-nucleotide step-length to analyze each transcript.

A k-mer pattern is a specific string with k nucleotides, each can be A, C, G or T.

For $k = 1$ to 5, had $4 + 16 + 64 + 256 + 1024 = 1,364$ patterns: 4 one-mer patterns, 16 two-mer patterns, 64 three-mer patterns, 256 four-mer patterns, and $1,024$ five-mer patterns.

Sliding-window of length $k, k = 1, 2, ..., 5$, which slides along the transcript of length l by a step-length of one nucleotide

$$f_i = \frac{c_i}{s_k} w_k,\ \ k = 1, 2, 3, 4, 5.\ i = 1, 2, \ldots, 1364 \tag{1}$$

$$s_k = l - k + 1,\ \ k = 1, 2, 3, 4, 5 \tag{2}$$

$$w_k = \frac{1}{4^{5-k}},\ \ k = 1, 2, 3, 4, 5 \tag{3}$$

Construction of classification model

To produce a balanced training dataset, we collected all the 22,389 long non-coding transcripts from the **GENCODE** v17 dataset (labelled as the "negative" class) and randomly selected 22,389 protein-coding transcripts from the human **RefSeq** dataset (labelled as the "positive" class).

Construction of classification model

Features: The 1,364 calibrated k-mer usage frequencies of each transcript were regarded as computation features.

Construction of classification model

Features: The 1,364 calibrated k-mer usage frequencies of each transcript were regarded as computation features.
Scale: MinMax to range 0 to 1, using the $svm - scale$.

Construction of classification model

Features: The 1,364 calibrated k-mer usage frequencies of each transcript were regarded as computation features.
Scale: MinMax to range 0 to 1, using the $svm - scale$.
Classifier: Support vector machine (SVM) with a radial basis functional kernel, whose variance is gamma, was selected as the binary classifier.

Construction of classification model

Features: The 1,364 calibrated k-mer usage frequencies of each transcript were regarded as computation features.
Scale: MinMax to range 0 to 1, using the $svm - scale$.
Classifier: Support vector machine (SVM) with a radial basis functional kernel, whose variance is gamma, was selected as the binary classifier.
Hyper Parameters: Optimal C of the SVM and gamma of the kernel were obtained using the grid search.

Construction of classification model

Features: The 1,364 calibrated k-mer usage frequencies of each transcript were regarded as computation features.
Scale: MinMax to range 0 to 1, using the $svm - scale$.
Classifier: Support vector machine (SVM) with a radial basis functional kernel, whose variance is gamma, was selected as the binary classifier.
Hyper Parameters: Optimal C of the SVM and gamma of the kernel were obtained using the grid search.
Validation: 10-fold cross-validation.

Simulation of indel sequencing errors

PacBio and 454 platforms generate longer reads, which tend to be more easily assembled than short reads.

Simulation of indel sequencing errors

PacBio and 454 platforms generate longer reads, which tend to be more easily assembled than short reads.
A tool robust to such errors is desirable to distinguish lncRNAs and mRNAs, and facilitates annotation of lncRNAs and mRNAs of a species without whole-genome sequences.

Simulation of indel sequencing errors

PacBio and 454 platforms generate longer reads, which tend to be more easily assembled than short reads.
A tool robust to such errors is desirable to distinguish lncRNAs and mRNAs, and facilitates annotation of lncRNAs and mRNAs of a species without whole-genome sequences.
Simulated 0 to 3 single-base indel sequencing errors per 100 bases (the error rate p was 0% to 3%).

Construction of a real sequencing dataset

The first dataset was recently released by PacBio and, the second dataset, a HelaS3 cell line transcriptome, was sequenced by a 454 GS FLX Titanium platform.

Different usage frequencies of k-mer strings

Calculated the calibrated usage frequencies of all the 1,364 k-mer patterns in the positive training dataset (22,389 protein-coding transcripts) and negative training dataset (22,389 long non-coding transcripts)

Different usage frequencies of k-mer strings

Calculated the calibrated usage frequencies of all the 1,364 k-mer patterns in the positive training dataset (22,389 protein-coding transcripts) and negative training dataset (22,389 long non-coding transcripts)
Wilcox rank-sum test was used to determine which k-mer pattern usage was significantly different between mRNAs and lncRNAs.

Different usage frequencies of k-mer strings

Calculated the calibrated usage frequencies of all the 1,364 k-mer patterns in the positive training dataset (22,389 protein-coding transcripts) and negative training dataset (22,389 long non-coding transcripts)
Wilcox rank-sum test was used to determine which k-mer pattern usage was significantly different between mRNAs and lncRNAs.
With a significance level of 10^{-6}, was found that 1,278 patterns were significantly different in their usage

Performance in cross-species prediction

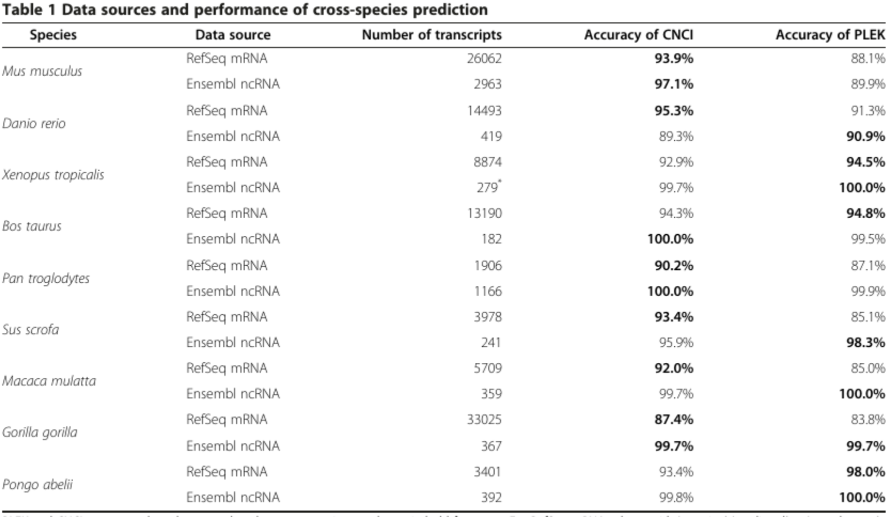


Table 1 Data sources and performance of cross-species prediction

Species	Data source	Number of transcripts	Accuracy of CNCI	Accuracy of PLEK
Mus musculus	RefSeq mRNA	26062	**93.9%**	88.1%
	Ensembl ncRNA	2963	**97.1%**	89.9%
Danio rerio	RefSeq mRNA	14493	**95.3%**	91.3%
	Ensembl ncRNA	419	89.3%	**90.9%**
Xenopus tropicalis	RefSeq mRNA	8874	92.9%	**94.5%**
	Ensembl ncRNA	279*	99.7%	**100.0%**
Bos taurus	RefSeq mRNA	13190	94.3%	**94.8%**
	Ensembl ncRNA	182	**100.0%**	99.5%
Pan troglodytes	RefSeq mRNA	1906	**90.2%**	87.1%
	Ensembl ncRNA	1166	**100.0%**	99.9%
Sus scrofa	RefSeq mRNA	3978	**93.4%**	85.1%
	Ensembl ncRNA	241	95.9%	**98.3%**
Macaca mulatta	RefSeq mRNA	5709	**92.0%**	85.0%
	Ensembl ncRNA	359	99.7%	**100.0%**
Gorilla gorilla	RefSeq mRNA	33025	**87.4%**	83.8%
	Ensembl ncRNA	367	**99.7%**	**99.7%**
Pongo abelii	RefSeq mRNA	3401	93.4%	**98.0%**
	Ensembl ncRNA	392	99.8%	**100.0%**

Figure: Data sources and performance of cross-species prediction

Robustness to indel sequencing errors

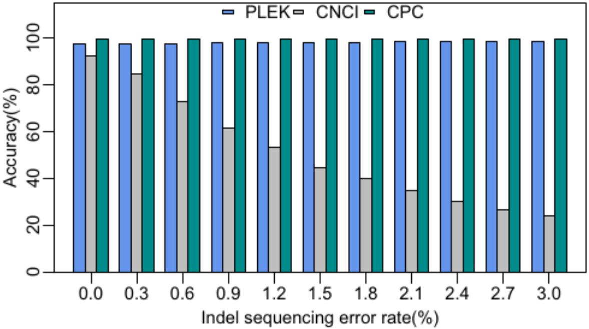
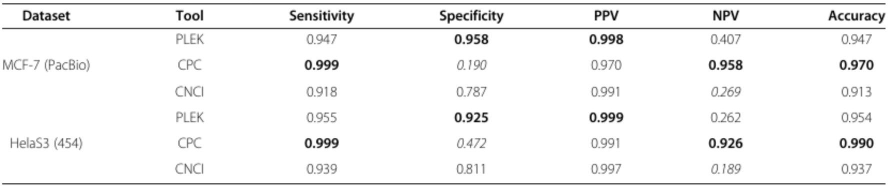


Figure: Comparison of robustness towards indel sequencing errors.

Robustness to indel sequencing errors



Dataset	Tool	Sensitivity	Specificity	PPV	NPV	Accuracy
MCF-7 (PacBio)	PLEK	0.947	**0.958**	**0.998**	0.407	0.947
	CPC	**0.999**	*0.190*	0.970	**0.958**	**0.970**
	CNCI	0.918	0.787	0.991	*0.269*	0.913
HelaS3 (454)	PLEK	0.955	**0.925**	**0.999**	0.262	0.954
	CPC	**0.999**	*0.472*	0.991	**0.926**	**0.990**
	CNCI	0.939	0.811	0.997	*0.189*	0.937

Figure: Performances on transcripts derived from PacBio and 454

Performance comparison on mouse datasets

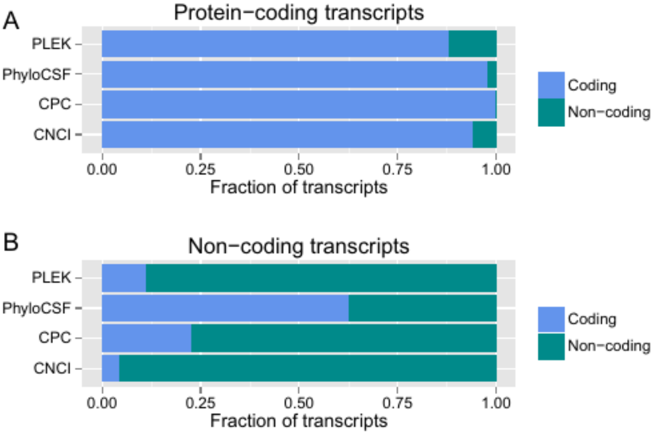
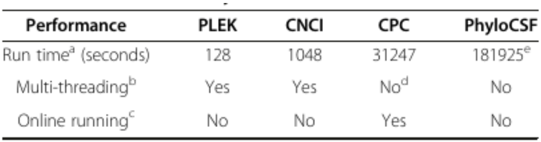


Figure: Results of PLEK, CPC, CNCI and PhyloCSF on mouse datasets.

Computational performance



Performance	PLEK	CNCI	CPC	PhyloCSF
Run time[a] (seconds)	128	1048	31247	181925[e]
Multi-threading[b]	Yes	Yes	No[d]	No
Online running[c]	No	No	Yes	No

Figure: Comparison of computational performances of PLEK, CNCI, CPC and PhyloCSF

Discussion

Prediction accuracy increases with the increasing k; however, this is accompanied by an increasing computation load.

Discussion

Prediction accuracy increases with the increasing *k*; however, this is accompanied by an increasing computation load.

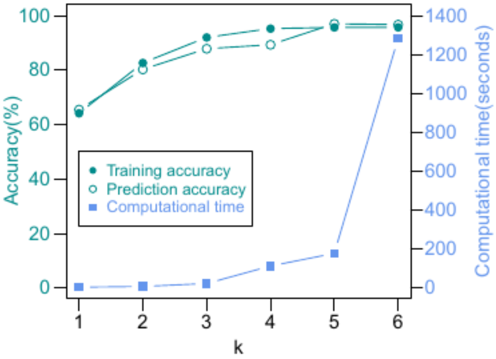


Figure: Performance comparison of various ranges of k.

Conclusion

PLEK is a useful tool for distinguishing protein-coding and non-coding sequences from high-throughput sequencing data of many species without reference genomes.

Note: This article is from 2014. From 2014 to the present day a lot of research has been done on the subject.

Conclusion

PLEK is a useful tool for distinguishing protein-coding and non-coding sequences from high-throughput sequencing data of many species without reference genomes.

Note: This article is from 2014. From 2014 to the present day a lot of research has been done on the subject.