

EMBO Practical Course, 12 – 16 May 2025 Didcot, United Kingdom

TEclass and TEclass2

TE classification using machine learning

Wojciech Makałowski

University of Münster, Münster, Germany

Adam Mickiewicz University, Poznań, Poland

Classification Problem

Typical Annotation Pipeline

RepeatModeler

Step 1: Start With a Good Genome Assembly

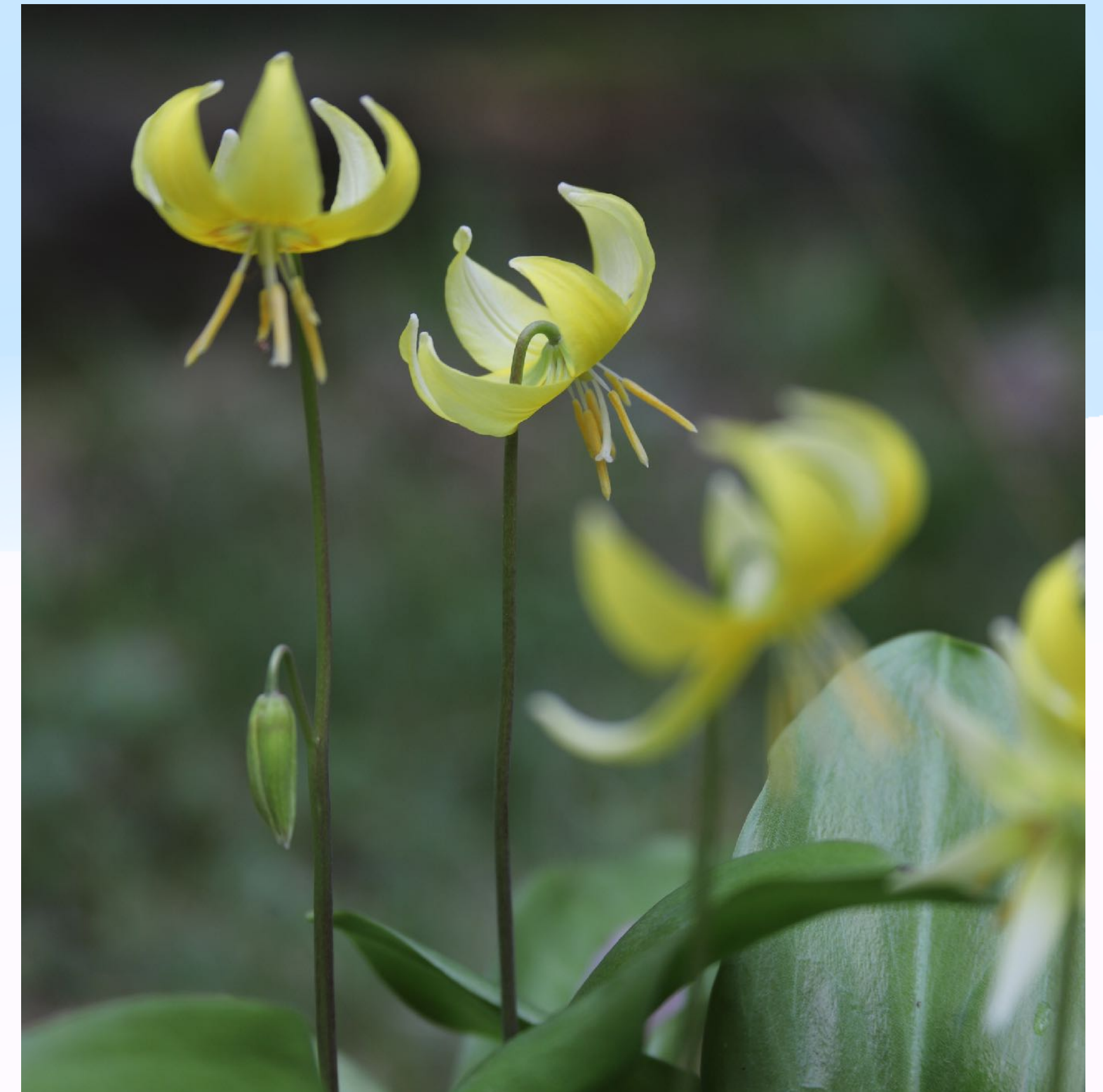
Step 2: Build a Repeat Database

Step 3: Discover Repeat Families

Step 4: Classify Repeat Families

Step 5: Create a Custom Repeat Library

Step 6: Annotate the Genome using the Library



Typical Annotation Pipeline

RepeatModeler

Step 1: Start With a Good Genome Assembly

Step 2: Build a Repeat Database

Step 3: Discover Repeat Families

Step 4: Classify Repeat Families

- Uses homology-based approaches against existing repeat databases (e.g. Dfam, Repbase if available).
- Assigns families to categories: DNA transposons, LTR retrotransposons, LINEs, SINEs, etc.
- **Unclassified repeats are labeled as Unknown.**

Over Eighty Per Cent Of Models Can Be Classified As “Unknown” by Repeat Modeler

Three different cultivars of *Medicago truncatula*

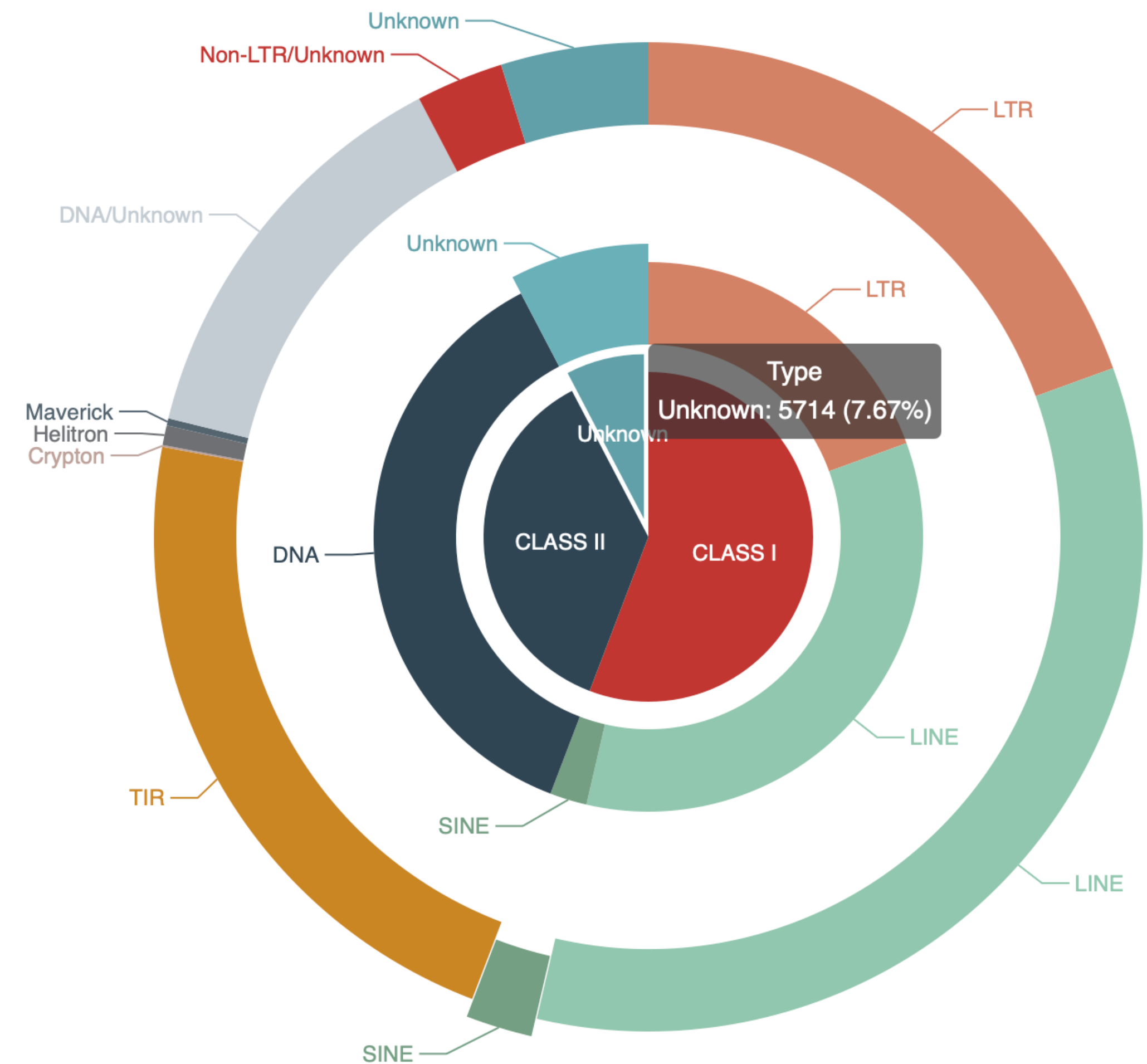
	<i>HM266</i>	<i>HM257</i>	<i>HM315</i>
Classified	403	447	398
Unknown	1912	2087	1917
Total	2315	2534	2315
Unknown fraction	0.83	0.82	0.83

Up To Eighty Per Cent Of Models Can Be Classified As “Unknown” by Repeat Modeler

FishTEDB example

<https://www.fishtedb.com/>

Type Unknown - 7.6%



Machine Learning To the Rescue



TEclass

The original - developed by György Abrusán around 2008

TEclass New Request 2025/05/01 18:08

Id

Title

File



Reverse Complement

☒

Text

Run

Fill Testcase

Reset



TEclass

<https://bioinformatics.uni-muenster.de/tools/teclass/index.hbi>

TEclass employs three distinct machine learning classifiers to categorize unknown eukaryotic transposable elements (TEs) into four primary functional categories: DNA transposons, LTR retrotransposons, LINEs, and SINEs. These are:

- Support Vector Machines (SVM): Utilizing the libsvm library with a Gaussian (RBF) kernel, SVMs are applied across various sequence length categories. (<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>)
- Random Forests: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. (<https://mtv.ece.ucsb.edu/benlee/librf.html>)
- Learning Vector Quantization (LVQ): A prototype-based supervised classification algorithm that represents each class by a set of prototypes. (http://www.cis.hut.fi/research/lvq_pak/README)
- Additionally, ORFs are being predicted as well.

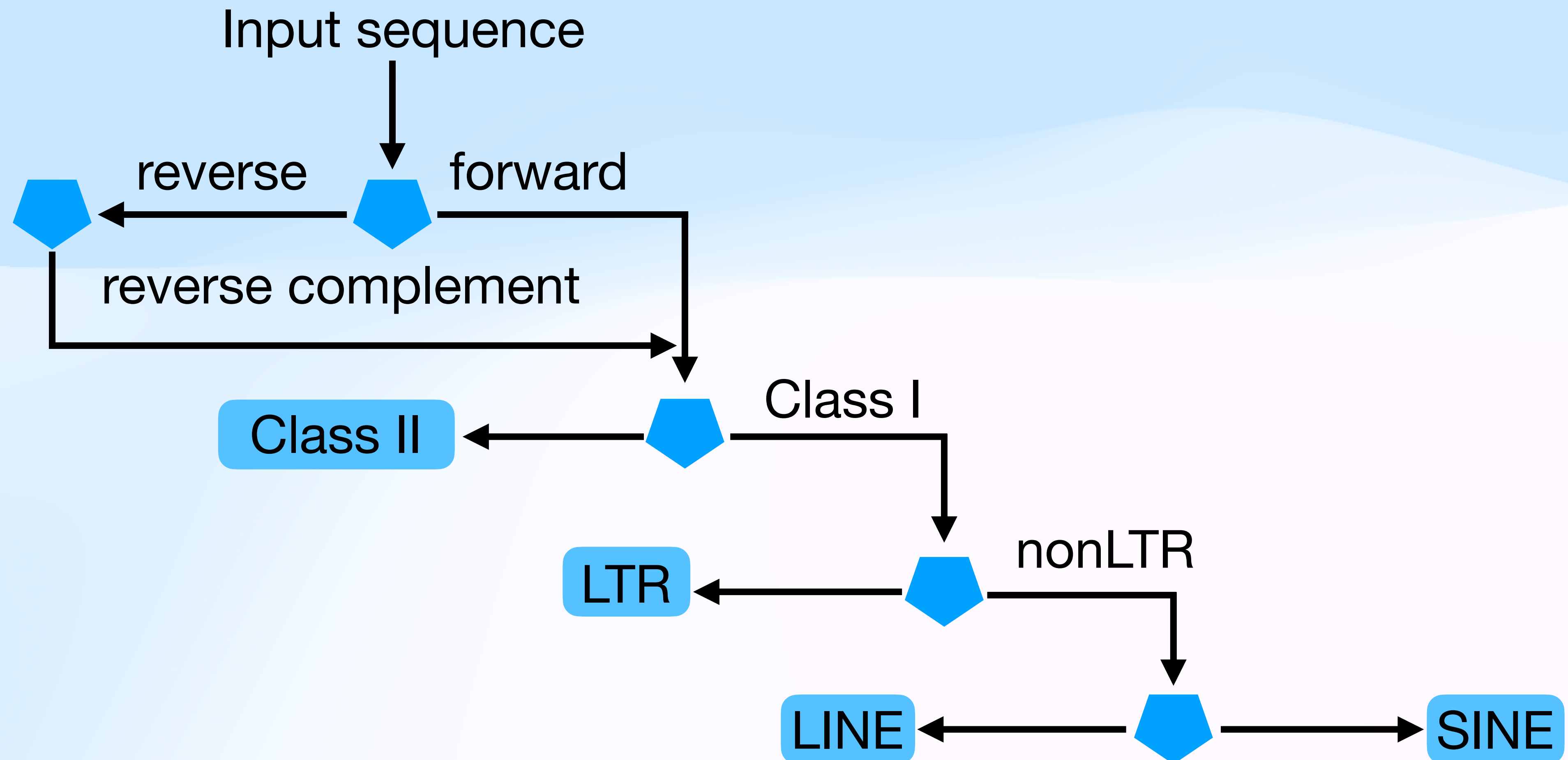
TEclass

<https://bioinformatics.uni-muenster.de/tools/teclass/index.hbi>

The classifiers are trained separately for different sequence length categories: 50 – 600 bp, 601–1800 bp, 1801–4000 bp, and over 4000 bp. The classification process follows a hierarchical approach:

- Orientation Determination: Forward versus reverse sequence orientation.
- Class Differentiation: DNA transposons versus retrotransposons.
- Retrotransposon Subclassification: LTR versus non-LTR elements.
- Non-LTR Subclassification: LINEs versus SINEs (applied only to sequences shorter than 1800 bp).

TEclass performs a hierarchical binary classification



TEclass - disadvantages

- Hierarchical classification causes accuracy drop with each step
- Last model update - May 2016
- Only major categories of TEs are predicted



TEclass - a side note 😊

Bioinformatics 25:1329-1330

One of the reviewers: The abilities of TEclass **to detect and annotate transposable elements are very limited** - it does not allow detection of multiple elements within the sequence (only one TE class is reported), and it does not report position of the element within the sequence. In my opinion, when a researcher already has a consensus sequence or a copy of TE extracted from a genomic sequence, it is usually quite easy to recognize the element type (e.g. based on its structure, coding regions, and other properties), so it is not necessary to use any program for this task. There might be a need for such classification in some unsupervised genome annotation pipelines, but **a web-based tool is of no use** for this purpose; a stand-alone program would be needed instead. Thus, **I could not see in which aspect the TEclass outperforms other methods of TE detection.**

Today: 356 citations (Google Scholar), 251 (Web of Science)

In the last 2755 days TEclass web site was accessed **568,678** times

Hands On!

You can access the program here:

<https://www.bioinformatics.uni-muenster.de/tools/teclass/generate/index.pl>

Input data can be taken from here:

https://www.bioinformatics.uni-muenster.de/share/EMBO_course/

TEclass2

Developed by Lucas Bickmann in 2008 (Master Thesis)

Improvements over TEclass

- Classification into sixteen TE superfamilies
- Modern machine learning architecture
- Much larger training dataset based on Dfam
- Probabilistic output - provides softmax scores for each TE category

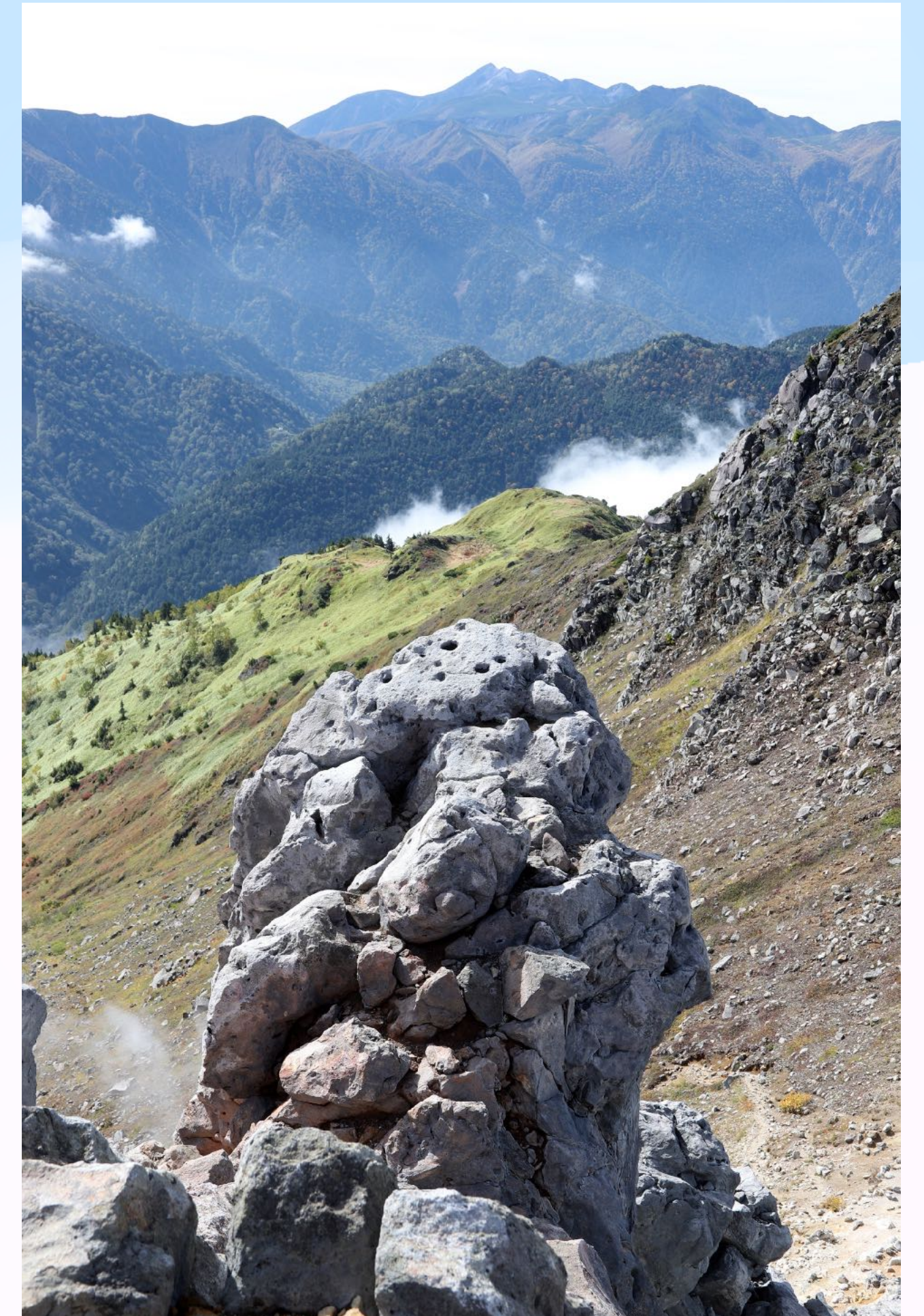


TEclass2 Is Using Linear Transformer

Vaswani et al. (Google Brain & University of Toronto) 2017 Attention Is All You Need

- Introduced the Transformer architecture.
- Replaced traditional RNNs (Recurrent Neural Networks) and CNNs (Convolutional Neural Networks) for sequence modeling.
- Based on a novel mechanism: self-attention, which lets the model weigh the importance of different words in a sequence, regardless of their position.

Result: Faster training, better performance on tasks like machine translation



Transformer - Basic Idea

According to ChatGPT

A Transformer is a type of computer program that's really good at handling sequences of stuff — like sentences, DNA, or protein sequences. It was first designed for translating languages but now it's used for everything from writing poetry to predicting protein shapes.

How Does It Work?

Imagine you're reading a sentence. As a human, you naturally keep track of the meaning of earlier words while you read new ones. A Transformer does something similar using a clever technique called ***attention***.

Transformer - Attention

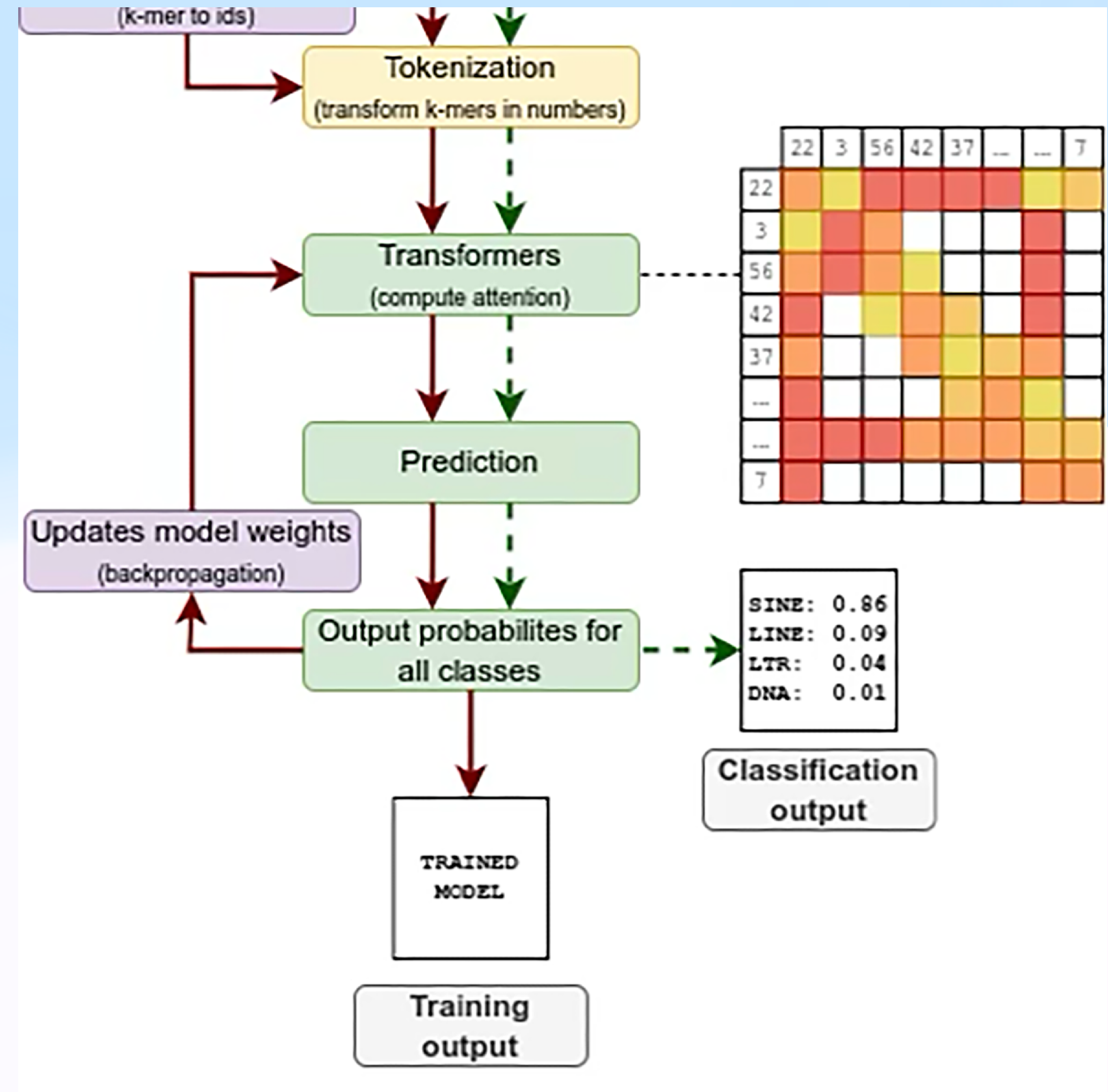
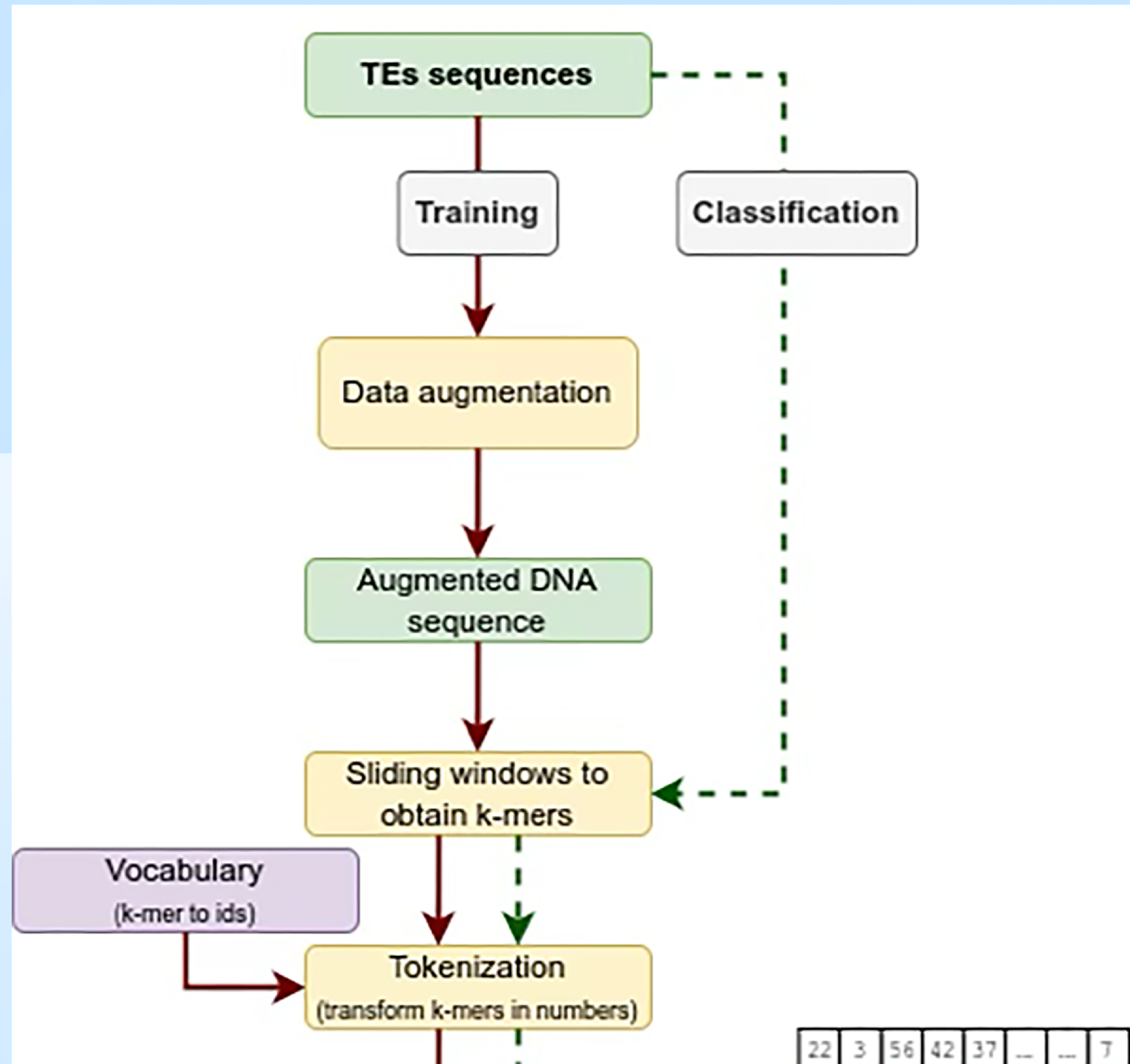
According to ChatGPT

Attention means:

👉 “When looking at one word (or letter, or amino acid), pay attention to other important parts of the sequence too.”

So instead of reading a sentence one word at a time like older programs did, a Transformer looks at the whole thing at once, figuring out which parts matter most to each other.

TEclass2 Flowchart



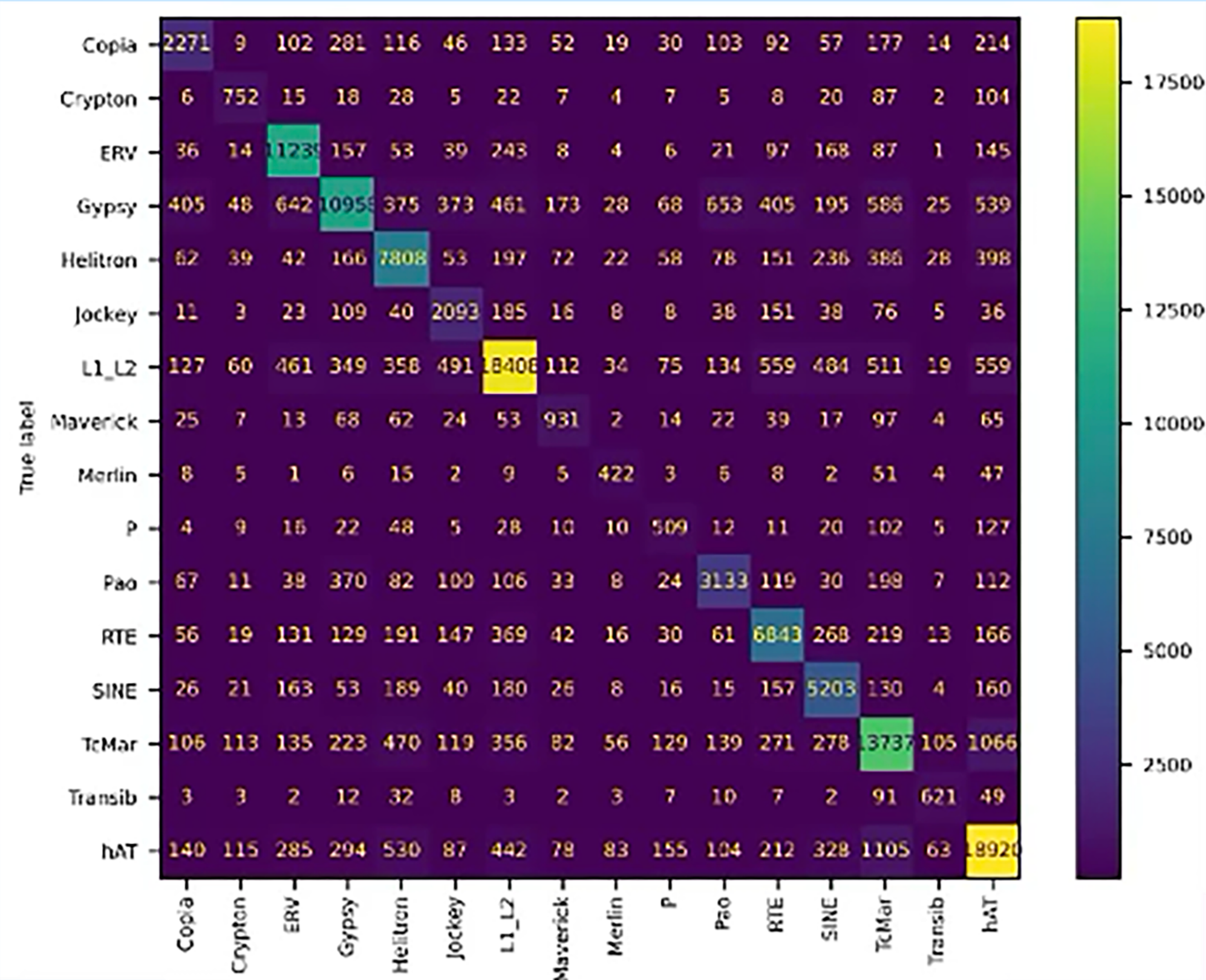
TEclass2 Training

Group	Train	Valid	Test	Total
Copia	18584	3717	2478	24779
Crypton	5453	1091	727	7270
ERV	61539	12319	8212	82124
Gypsy	79674	15935	10623	106232
hAT	114707	22941	15294	152943
Helitron	48980	9796	6531	65307
Jockey	14201	2840	1894	18935
L1/L2	113708	22742	15161	151610
Maverick	7218	1444	962	9624
Merlin	2971	594	396	3961
P	4692	938	626	6256
Pao	22192	4438	2959	29589
RTE	43500	8700	5800	58000
SINE	31960	6392	4261	42613
TcMar	86925	17385	11590	115900
Transib	4279	856	571	5705
Total:	660636	132127	88085	880848

Sequence Augmentation used by TEclass2

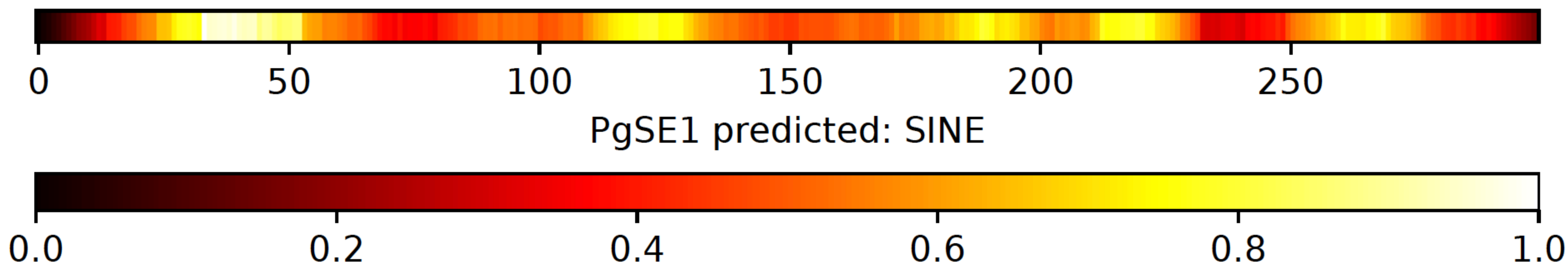
Augmentation	Description
SNP	Replace a single nucleotide with a random other nucleotide
Masking	Replaces bases with ambiguity (AGTC→N)
Insertion	Inserts a random sequence
Deletion	Deletes a random part of the sequence
Repeat	Repeats a random part of the sequence at another following random position
Reverse	Reverses the sequence
Complement	Computes the complement sequence
Reverse complement	Computes the opposing strand representation
Add tail	Adds a poly-A-tail to the end of the sequence
Remove tail	Removes a poly-A-tail at the end of the sequence if present

TEclass2 Performance



Flatted Local Attention Heat-Plot

Showing relative importance of each token in their neighboring regions



TEclass2 - Very Simple Interface

TEclass2

To run a TEclass2 request: please enter either a file or type/copy the sequence data in FASTA format into the "Sequence Data" field.

File
Sequence
Data



Run

Fill Testcase

Reset



TEclass2 output - probability

Select a probability for filtering results: 0.7

name	order	class	probability	Copia	Crypton	ERV	Gypsy	hAT
rnd-1_family-416 LTR/Gypsy	Class_I-LTR	Gypsy	0.972	0.020	0.000	0.000	0.972	0.001

Here probability is calculated as “softmax”

1. Compute Scores
2. Scale the Scores
3. Apply Softmax - scaled scores are converted into probabilities (values between 0 and 1). These probabilities always sum up to 1.

If the best probability is below the threshold (e.g. 0.7) the sequence is called “Unknown”

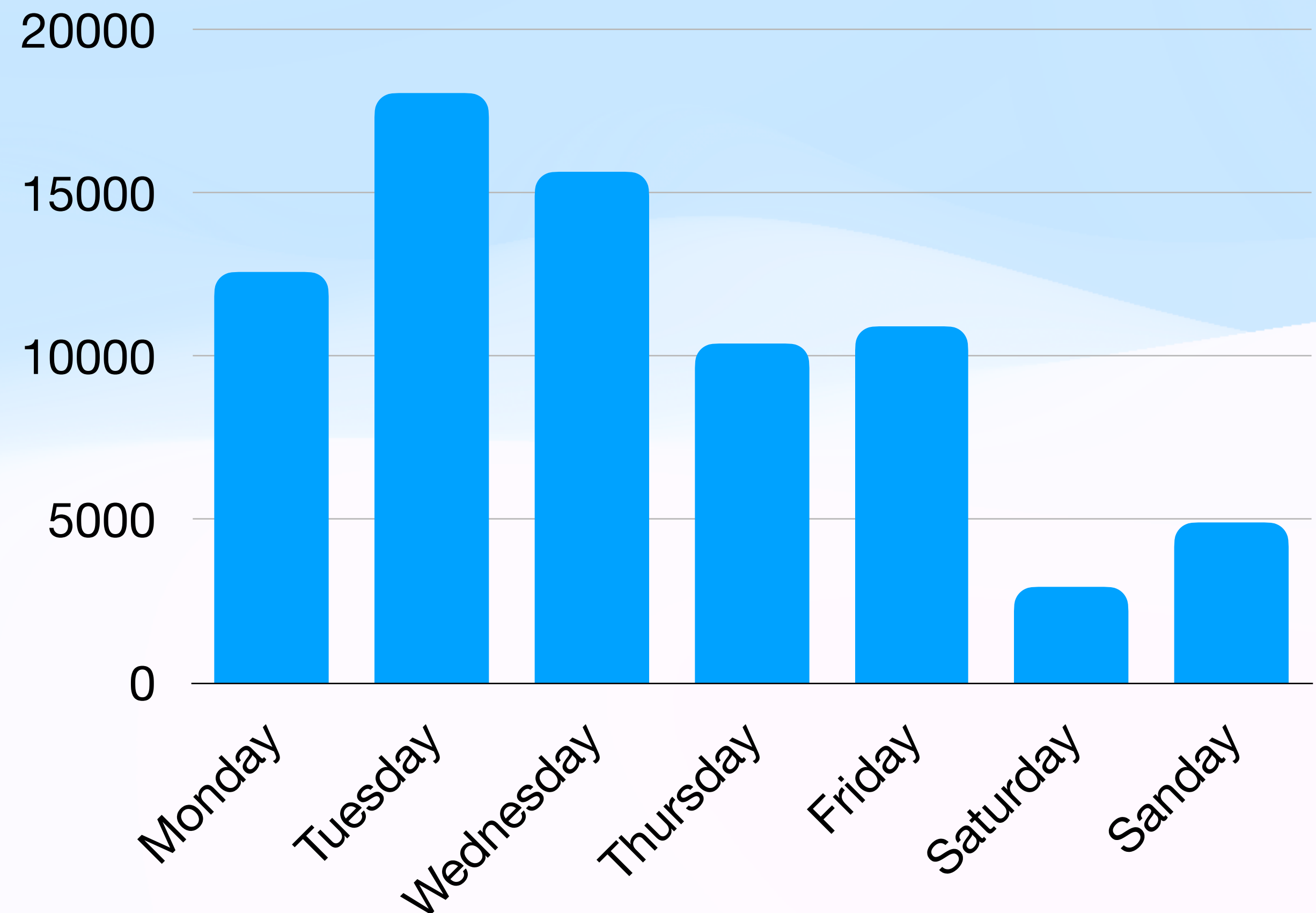
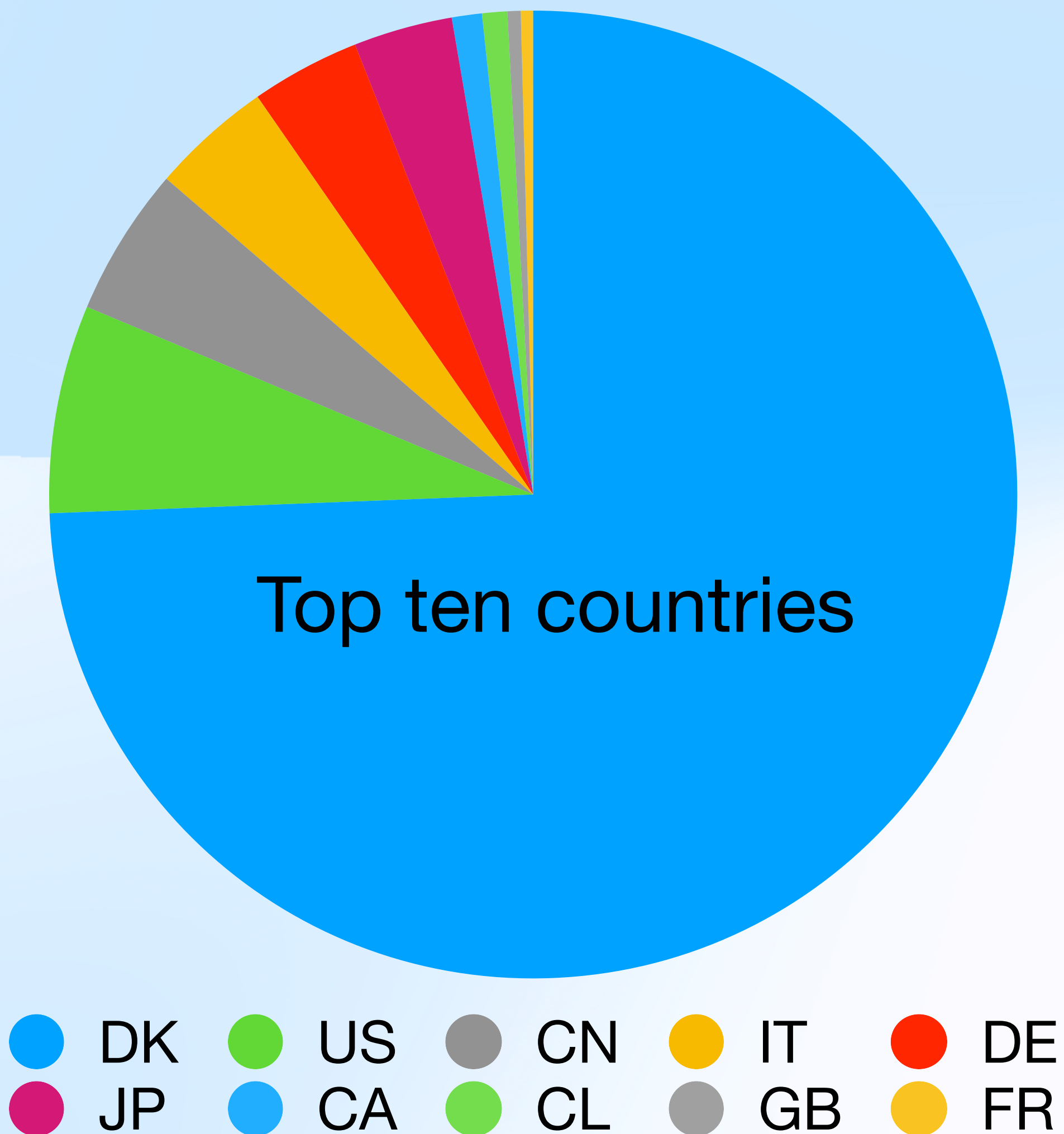
Back To *Medicago* Case

After manual curation

		<i>HM266</i>
Repeat Modeler Output	Classified	403
	Unknown	1912
After Manual Curation	Classified	814
	Unknown	123
TEclass2	Unknown	44



TEclass2 - Access Statistics - Last 322 Days



The Final Thought on Transformer

<https://arxiv.org/abs/2002.05202v1> by Noam Shazeer

4 Conclusions

We have extended the GLU family of layers and proposed their use in Transformer. In a transfer-learning setup, the new variants seem to produce better perplexities for the de-noising objective used in pre-training, as well as better results on many downstream language-understanding tasks. These architectures are simple to implement, and have no apparent computational drawbacks. We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence.

Hands On!

You can access the program here:

<https://bioinformatics.uni-muenster.de/tools/teclass2/index.pl>

Input data can be taken from here:

https://www.bioinformatics.uni-muenster.de/share/EMBO_course/

Expected Results

Dataset 1

Classification by	123	80	138	222	231	233	234	324	10321	282
RepeatModeler	Unkn	Unkn	Unkn	Unkn	Unkn	Unkn	Unkn	Unkn	Unkn	Unkn
Manual curation	Unkn	LTR/ Copia	SINE/ tRNA	DNA/hAT	Unkn	RC/ Helitron	Unkn	DNA/hAT	Unkn	SINE/ID
TEclass	LINE	LTR	SINE	DNA	LTR	Retro	LTR	Retro	LTR	SINE
TEclass2	LTR/ Gypsy	LTR/ Copia	SINE	DNA/hAT	LTR/ Gypsy	Unkn	LTR/ Copia	Unkn	LTR/ Copia	Unkn

Expected Results

Dataset 2

Classification by	Human Alu-Sx	Rattus L1	DF00000 1406.2	DF00028 9984.2	DF00028 9985.2	DF00028 9992.2	DF00029 0020.2	Tequ_3	Gnip_22	api_213
Dfam or GenBank	Alu-Sx	L1	Helitron	Unkn	Unkn	Unkn	Unkn	LINE	Gypsy	Gypsy
TEclass	SINE	LINE	DNA	SINE	Retro	LTR	LTR	LTR	LINE	LTR
TEclass2	SINE	L1_L2	Helitron	Unkn	Gypsy	ERV	Gypsy	Jockey	Gypsy	Gypsy