https://creationismeweersproken.blogspot.com/2023/01/de-rode-panda-en-cserhati-6-whole.html

THE RED PANDA AND CSERHATI (6): WHOLE GENOME K-MER SIGNATURE

Cserhati indicates in the BMC Genomics article that he doubts the classification of the red panda on the basis of morphology. Since he has talked more about DNA sequences than about morphology, he will mean that he (also) finds the classification based on DNA subject to doubt. All the classifications based on DNA that he has mentioned are based on relatively little DNA (for 2021 standards). For example, Flynn et al (2000) used the DNA sequence of four genes. That was the year 2000, of course, and then no more DNA sequences were available

Cserhati prefers to use the whole genome for red panda classification: a whole genome study. That is increasingly what is happening. For example, De Ferran et (2022) did not search for orthologous genes in their eleven species of otters, but used genome fragments as found during genome sequencing for DNA comparison.

Cserhati's preference for using the entire genome is therefore perfectly understandable. The method used by Cserhati to characterize the whole genome is *Whole Genome K-mer Signature*, abbreviated as WGKS.

There are two questions: what is WGKS? And how useful is WGKS

First: what is WGKS? This is addressed in this post. How useful WGKS is for species classification will be discussed in the next installment.

In the Methods section of the BMC Genomics article, Cserhati writes:

> *The WGKS algorithm that was used in the analysis is an alignment-free k-mer sequence comparison method. These methods involve the statistical comparison of k-mers between species.*
>
> *A k-mer is a segment of DNA k bp long,*

The k-mer signature is simply a list of all k-mers ordered in lexicographical order from AA … A to TT … T, together with their score values. For a given value k, there are 4^k possible k-mers. Thus, the k-mer signature also corresponds to a vector of 4^k numbers. Since octamers were analyzed, this corresponds to 65,536 possible octamers.

Cserhati says: count all k-mers eight bases long - octamers - , and then find their scores. I'll give an example in two parts: counting octamers and finding scores.

1 Counting octamers and the correlation between octamer numbers

DNA has four bases: ACGT. An octamer, a DNA sequence 8 base pairs long, can show all possible sequences from AAAAAAAA to TTTTTTTT. Four possibilities for place 1, four possibilities for place 2, and so on. That means 4^8= 65536 possibiities. A computer walks along the genome, and reads sequentially which sequence of 8 basepairs is found.

In a DNA sequence;

```
gagtgggcagcactccaaataccgttaagctggagcctcggt
```

the consecutive octamers are:

from base 1: `gagtgggc`

from base 2: `agtgggca`

from base 3: `gtgggcag`

and so on. The computer counts the number of times an sequence of 8 bases occurs. In this example of a short DNA sequence, each sequence of 8 bases occurs once. The count defines the k-mer signature; with 8 bases it is called the octamer signature.

In DNA from two related species one would expect to find approximately the same distribution of 8-base sequences, octamers.

However, species differ not only in important DNA but also in unimportant DNA. For example, in the length of a repeat 'ac' - acacacac or acacacacacacacacacaca, or a difference in the number of LINE1 elements. For example, a house mouse has hundreds of LINE1 elements (only a few of which are active as transposons (Jachowitz et al 2017)), and another species of mouse could have thousands of LINE1 elements. Such a repetition of the same

sequence as a plurality of LINE1 elements or a plurality of simple repeats will have a major influence on the distribution of octamers in a WGKS

Example of octamer counts:

A long example follows here, and I constructed it to show the influence of repetitive DNA.

Consider the following five sequences. In the first two sequences, a long or short ‘ac’ repeat is inserted relative to the third sequence. The fourth sequence differs from the first sequence in the part after the long ‘ac’ repeat. The fifth sequence has the the shorter ‘ac’ repeat, and is otherwise the same as the fourth sequence.

```
>seq_1
gagtgggcagcaacacacacacacacacacacacacacacacacacctccaaataccgttaagctggagcctcggt
>seq_2
gagtgggcagcaacacacacacacacacctccaaataccgttaagctggagcctcggt
>seq_3
gagtgggcagcactccaaataccgttaagctggagcctcggt
>seq_4
gagtgggcagcaacacacacacacacacacacacacacacacacactcttctggtccccacagactcagagaga
>seq_5
gagtgggcagcaacacacacacacacacactcttctggtccccacagactcagagaga
```

Inserting points to bring out the ‘ac’ repeat:

```
>seq_1
gagtgggcagca...acacacacacacacacacacacacacacac....ctccaaataccgttaagctggagcctcggt
>seq_2
gagtgggcagca...acacacacacacacac..................ctccaaataccgttaagctggagcctcggt
>seq_3
```

```
gagtgggcagca.....................................ctccaaataccgtta
agctggagcctcggt

>seq_4

gagtgggcagca...acacacacacacacacacacacacacacacacac...tcttctggtcccca
cagactcagagaga

>seq_5

gagtgggcagca...acacacacacacacacac.................tcttctggtcccca
cagactcagagaga
```

A simple program, CLUSTAL, alignes the sequences, and provides two types of trees:

```
CLUSTAL O(1.2.4) multiple sequence alignment


seq_4      gagtgggcagcaacacacacacacacacacacacacacacacactcttctggtccc--        58
seq_5      Gagtgggcagcaacacacaca----------------cacacacactcttctggtccc--        42
seq_1      gagtgggcagcaacacacacacacacacacacacacacacacctcc---aaataccgt        57
seq_2      gagtgggcagcaacacacac----------------acacacacctcc---aaataccgt        41
seq_3      gagtgggcagca--------------------------------ctcc---aaataccgt        25
      ************                                    *      * **


seq_4      -----cacagactcagagaga   74
seq_5      -----cacagactcagagaga   58
seq_1      taagctggagcctcggt----   74
seq_2      taagctggagcctcggt----   58
seq_3      taagctggagcctcggt----   42
           ** *** *
```

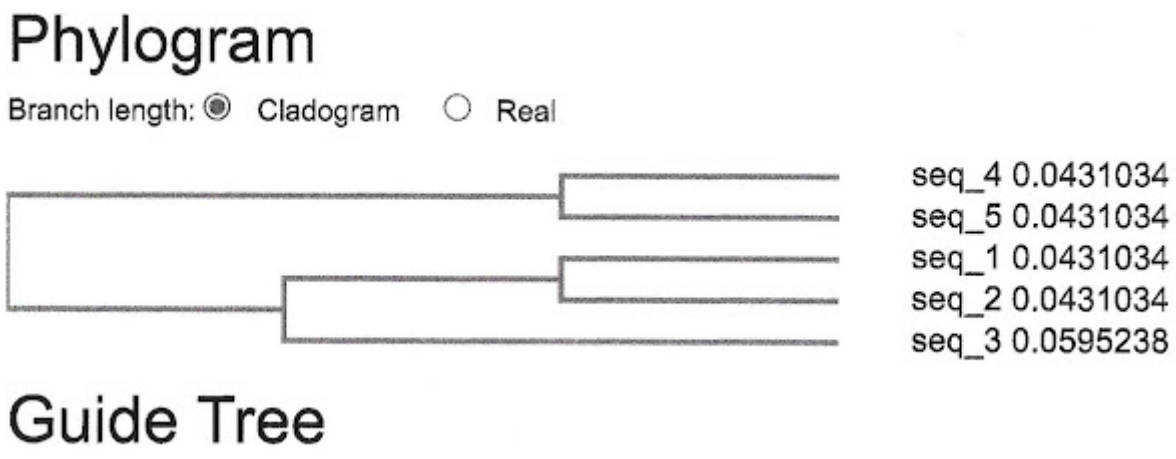

Figure 1. Graphic representation as Phylogram

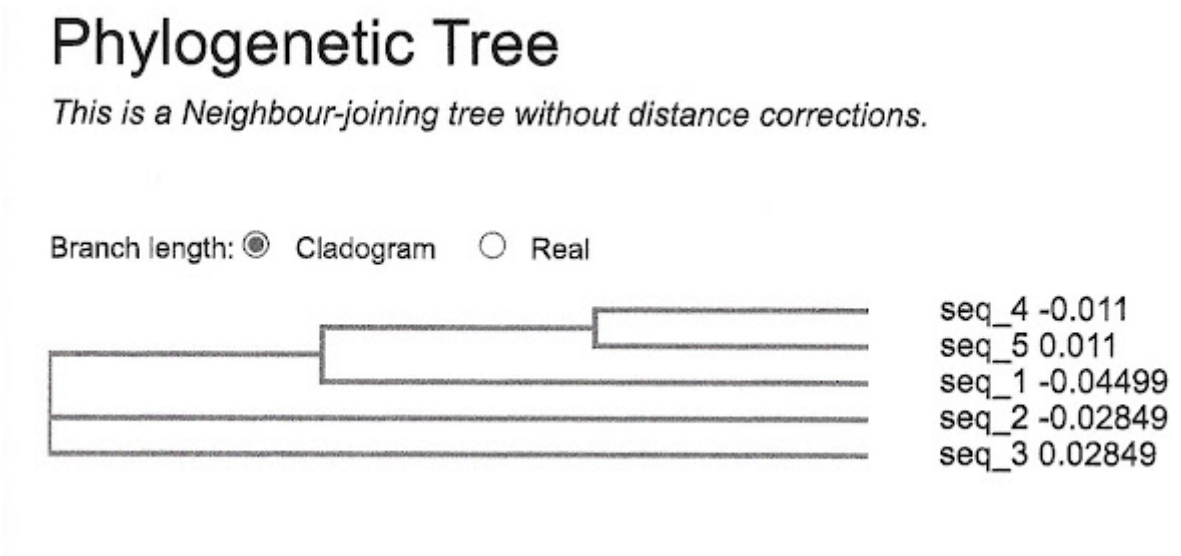

Figure 2. Graphic representation as Phylogenetic tree

In the phylogram the first three sequences are placed together, in the phylogenetic tree by Neighbor Joining sequences 4 and 5 are joined by the first sequence. In that case, the long repeat sends sequence 1 to sequences 4 and 5.

How does WGKS respond to repeats?

Octamer counts have been executed for these five sequences: 83 different octamers have been found. Of these, 81 occur zero or once in a sequence. The remaining two are *acacacac* and *cacacaca*, and occur multiple times. The octamer signatures begin:

string	seq 1	seq 2	seq 3	seq 4	seq 5
aaataccg	1	1	1	0	0
aacacaca	1	1	0	1	1
aagctgga	1	1	1	0	0
aataccgt	1	1	1	0	0
acacacac	13	5	0	14	6
acacacct	1	1	0	0	0
acacactc	0	0	0	1	1
acacctcc	1	1	0	0	0
acactctt	0	0	0	1	1
acagactc	0	0	0	1	1

The correlation matrix of the counts shows that sequence 3 deviates from the other four sequences:

	seq 1	seq 2	seq 3	seq 4	seq 5
seq 1	1	0.905	0.020338	0.915432	0.758678
seq 2	0.905	1	0.262668	0.693422	0.511118
seq 3	0.020338	0.262668	1	-0.27067	-0.43957
seq 4	0.915432	0.693422	-0.27067	1	0.935843
seq 5	0.758678	0.511118	-0.43957	0.935843	1

	seq 1	seq 2	seq 3	seq 4	seq 5
seq 1	1	0.992977	-0.07366	0.985514	0.964082
seq 2	0.992977	1	-0.00813	0.963535	0.936946
seq 3	-0.07366	-0.00813	1	-0.1939	-0.25839
seq 4	0.985514	0.963535	-0.1939	1	0.992437
seq 5	0.964082	0.936946	-0.25839	0.992437	1

2 Octamer scores

Cserhati does not use the octamer counts, but an 'octamer score'.

First point: how to calculate the score of any octamer. Second point: what would be the sensitivity of the score to repeats.

The octamer score is according to the Python Script motif program (github.com/csmatyi/motif_analysis):

Score Sc = $(O-E)/(O+E)$

The observed number of specific octamer equals O and the expected number of that octamer equals E. When observed number of the octamer equals the expected number the score Sc equals zero.When the observed number equals zero, $O=0$, the score equals -1.

I give two approaches to look at the sensitivity of the score to deviations from expected: the relative deviation from expected and the absolute deviation from expected.

i) The first approach gives the relative deviation from expected; genome size is not important in that case.

If $O=xE$, for $x \geq 0$, Sc = $(xE-E)/(xE+E) = (x-1)/(x+1)$.

At $x=0$ is Sc=-1; at $x=1$ is Sc =0; when x approaches infinity the limit becomes Sc = 1. The score Sc is strongly asymmetric.

Score Sc as a function of x plotted in two ways:

Lineair x-axis:

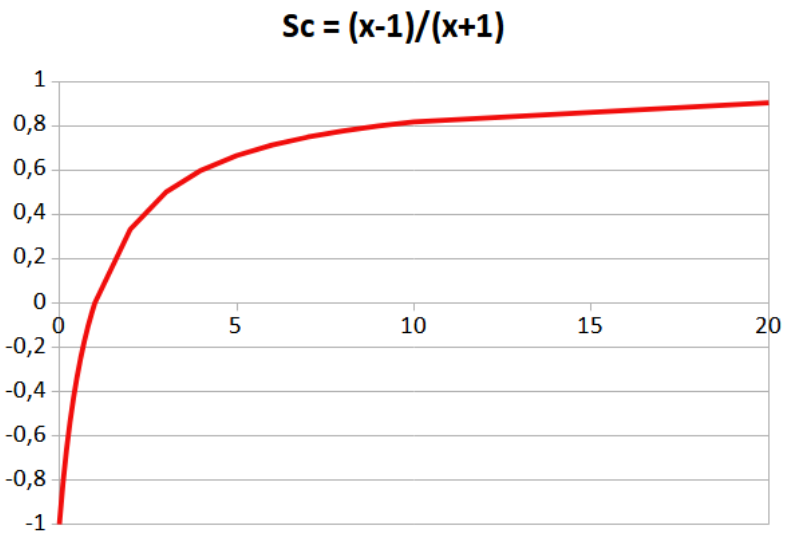

Figure 3 Score Sc = (x-1)/(x+1) as function of x

Logarithmic x-axis

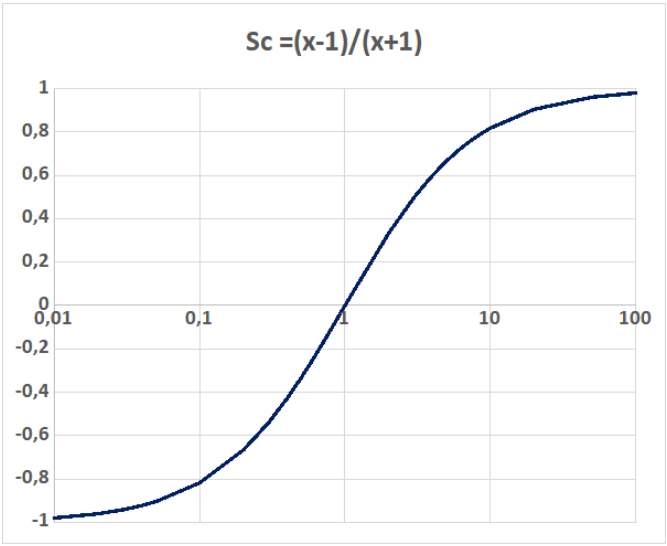

Figure 4 Score Sc = (x-1)/(x+1) as function of x

The score seems rather sensitive to deviations between observed number *O* and expected number *E* for fairly small deviations from expected, and insensitive to large deviations from expected.

ii) The second approach looks at the absolute deviation from expected: now genome size becomes an important factor.

There are 4^8= 65536 different octamers. Genome size is *N* base pair. A rough first approximation has expected number of an octamer at $E=N\text{x}4^{-8.}$ Observed number of that octamer is now $O=E+n$.

The score is now given by Sc= $(E+n\text{-E})/(E+n+E) = n/(2E+n)$. Genome size and absolute deviation from expected appear in this score.

The next figure has scores for genome sizes $N=10^6$, $N=10^7$, $N=10^8$, $N=10^9$ (separate lines); the absolute difference between observed and expected ranges from 10^1 tot 10^5 and is plotted at the x-axis.

With a large genome, in the order of 10^9 bp, the absolute deviation between observed and expected must be large for the score to change significantly. With a smaller genome, almost all changes lead to high scores.

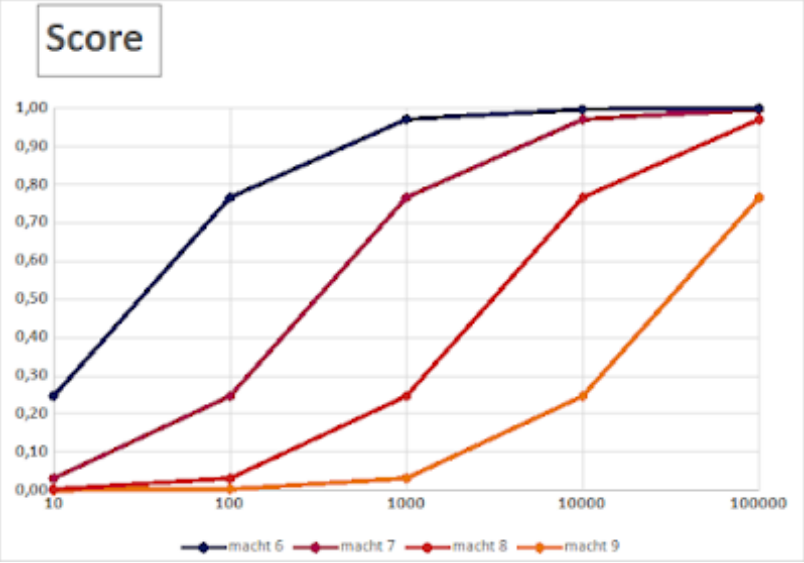

Figure 5. The score as a function of difference between observed and expected for different genome sizes, N = 10^6 (blue) to N=10^9 (orange)

Small duplications therefore have hardly any effect on the score, while transposon frequencies have major effects. Differences in frequency of transposons such as LINE1 with 15-20% of the genome will have a major

influence on the genome, because such transposons occur in such large numbers.

Cserhati writes:

> *Even if the genome is partially or completely duplicated, then the score value will not change. This is because both the Observed and Expected values will increase by the proportion that the duplicated genome is compared to the pre-duplication genome*

This is not correct. The expected number *E* of an octamer will depend less on the presence of duplications than the observed number *O,* if only part of the genome is duplicated. The influence of the non-duplicated part of the genome will predominate in *E*.

3 Take home message

The octamer score Sc = (*O*-*E*)/(*O*+*E*) is nonlinear making it doubtful whether this is a workable measure of any genome trait.

In a correlation matrix based on WGKS, differences in large amounts of repetitive DNA will have a major impact.

When working with octamer signatures from related species, we have a fairly similar octamer pattern derived from informative DNA; differences in repetitive DNA have a major impact on the correlations between the signatures of the species against that similar background of informative DNA.

When working with octamer signatures of species that are far apart in their phylogeny, we expect a higher influence of the difference in octamer pattern from informative DNA. Against this background of more difference due to informative DNA, the influence of repetitive DNA can diminish, disappear or, when using a large number of species from a group, average out.

A phylogeny on WGKS octamer patterns might perhaps approximate the main lines of a phylogeny, but it cannot be expected that WGKS octamer patters produce an accurate phylogeny on smaller scale comparisons – as with species in a family or even families within a superfamily.

***

Cserhati, M., 2021, A tail of two pandas – whole genome k-mer signature analysis of the red panda (Ailurus fulgens) and the Giant panda (Ailuropoda melanoleuca), BMC Genomics 22: 228

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07531-3

Jachowicz, J.W., Bing, X., Pontabry, J., Bošković,A., Rando, O.J., & Torres-Padilla, M-J (2017) LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nature Genetics 49: 1502

de Ferran, V., Figueiró, H., de Jesus Trindade, F, en 17 anderen , & Eizirik, E. (2022) Phylogenomics of the world's otters. Current Biology 32; 3650–3658,

The Python script motif_analysis_k-1.py at github.com/csmatyi/motif_analysis was used to generate WGKS profiles

https://github.com/csmatyi/motif_analysis