A Metric for Genomes Based on Unique Substrings

Hendrik Jan Hoogeboom
hoogeboo@liacs.nl

Walter A. Kosters
kosters@liacs.nl

Jeroen F. J. Laros
jlaros@liacs.nl

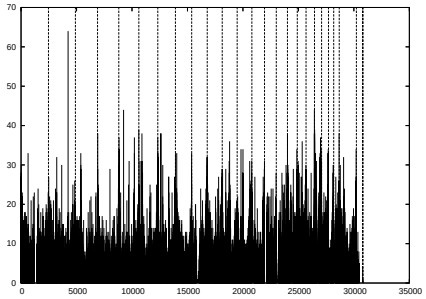
Universiteit Leiden

Unique substrings

Unique substrings are used for applications ranging from the development of primers, applied in PCR or MPLA experiments, to the fabrication of microarrays.
Occurrences that are few (or zero) in one species and many in the other have discriminating value. Just counting how often this happens can be used to find similarities between species.

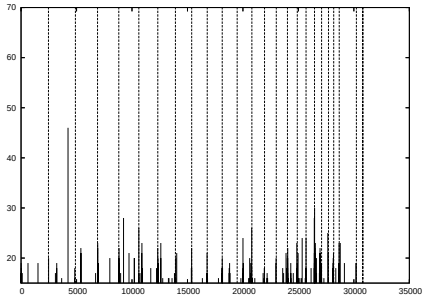
size	11	12	13	14	15	16
Human	0.2	47	1,335	15,412	85,793	346,600
Chimp	0.3	62	1,509	16,636	87,029	346,319

In this table we see the number ($\times 10^3$) of unique substrings of lengths 11 to 16 in the DNA of the Human and the Chimpanzee.



Here we see the occurrence of unique strings of length 12 in Human DNA, for each consecutive series of 100,000 basepairs. The vertical dotted lines denote the chromosome boundaries: 1 to 22, then X, Y, and finally the mitochondrial DNA. This mitochondrial DNA is so small that it does not show up in these graphs.

For each unique substring we can check whether it is present in a particular other species. In this case resulting in the following plot:



Here we see the occurrence of unique substrings of length 14 present in the Human genome, but not in the Chimpanzee.

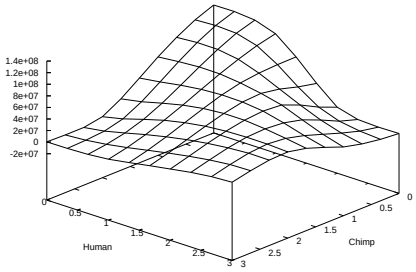
Pairwise comparison

We can take this idea one step further by counting not only unique strings, but all strings, and placing them in a matrix where position (x, y) denotes the number of times substrings were found x times in one species and y times in the other species. This results in a matrix of the following form ($\times 10^3$):

		Human			
		0	1	2	≥ 3
Chimp	0	150,783	4,486	1,216	498
	1	3,212	7,352	3,737	2,333
	2	602	2,621	4,011	4,907
	≥ 3	145	950	2,697	78,877

There are $3,212,000$ substrings that are present in the Chimpanzee once and do not occur in the Human genome.

The differences between the two species can be visualised by removing the main diagonal of the matrix and by plotting the remaining values, resulting in a picture like this:



Here we see all substrings of length 16 that occur more in the Human than in the Chimpanzee and vice versa.

The DALE project is financed in the ToKeN program of the Netherlands Organisation for Scientific Research (NWO) under grant number 634.000.430.

Data Assistance for Law Enforcement

http://dale.liacs.nl/

NWO ToKeN2000

Multiset metric

When a substring occurs three times in one genome and not in another genome, the substring is more defining than a substring that occurs once in one genome and twice in the other.

Therefore, we choose a metric d_f that accentuates larger differences in occurrences.

For a multiset X, let $S(X)$ denote its underlying set. For multisets $X, Y \subseteq \{1, 2, \ldots, n\}$ we define $d_f(\emptyset, \emptyset) = 0$ and

$$d_f(X, Y) = \frac{\sum_{i=1}^{n} f(x_i, y_i)}{|S(X) \cup S(Y)|}$$

with

$$f(x, y) = \frac{|x - y|}{(x + 1)(y + 1)}$$

With this metric, we can make a pairwise distance matrix of all the generated 4×4 matrices generated in the previous step:

	Y	S	H	Dm	D	Co	Ch	Ci	Ce	B
Y	.00									
S	.50	.00								
H	.61	.62	.00							
Dm	.51	.51	.58	.00						
D	.61	.61	.31	.57	.00					
Co	.61	.61	.32	.57	.32	.00				
Ch	.61	.61	.15	.57	.32	.32	.00			
Ci	.58	.58	.38	.54	.38	.39	.38	.00		
Ce	.51	.52	.59	.49	.58	.58	.58	.54	.00	
B	.53	.54	.57	.49	.56	.56	.56	.53	.49	.00

A distance matrix based upon substrings of length 16 for the species Yeast, SARS, Human, Drosophila Melanogaster, Dog, Cow, Chimpanzee, Chicken, C. Elegans and Bee. According to this metric, Human an Chicken have distance 0.38 to each other.

Using the matrix given above, we can make an unrooted phylogenetic tree that looks like this: