Compressing and Indexing Highly Similar Strings using LZW

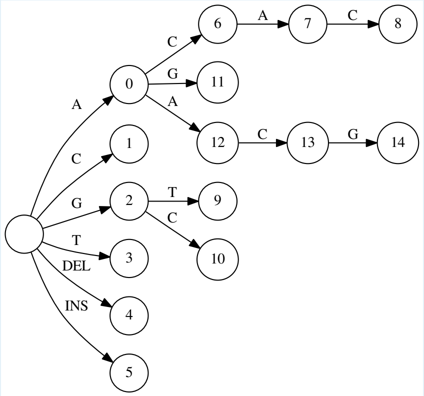
Problem

As the DNA sequencing methods have become relatively cheap and fast, a lot of DNA sequencing projects, such as *1000 Genomes* and *Genome 10K*, have emerged. These projects yield massive amounts of data. In order to minimize storage costs, an effective compression method is needed. Unfortunately, general-purpose compression algorithms are not feasible for this task.

It is known that similarity of any two human genomes is more than 99 %. Similarly, there is only a small difference between genomes within particular species. This redundancy can be used for compression.

Approach

Several modifications of the original LZW algorithm were studied in order to create an algorithm for compression of Highly Similar Strings. The final version of the presented algorithm, called ALZW (Alignment-based LZW), uses sequence alignment for identification of common phrases between multiple input sequences. These phrases are inserted into an LZW-like dictionary, which also contains special keywords for insertions and deletions. Each input sequence is then encoded relatively to a given reference sequence using keywords from the dictionary.



The image shows an example of the ALZW dictionary after compression of the following alignment:

```
A C A C G T C C G C - - A G G T A A C G
A C A C G T A C G C A C A G - - A A C G
```

Results

A new algorithm for compression of Highly Similar Strings was designed and implemented. The algorithm was tested on two datasets – 21 sequences of human chromosome 20 and 39 sequences of Saccharomyces cerevisiae. Despite having quite big memory requirements, the algorithm gives promising results in terms of compression speed and compression ratio (see the table) compared to similarly targeted RLZ and general-purpose GZip.

	Entropy [bpb]	
	H. chr. 20	S. cerevisiae
Original	2.21	2.12
ALZW	**0.36**	1.34
RLZ	0.38	**0.29**
GZip	2.29	2.40

Author: Ondřej Perutka | Supervisor: Jan Holub | CTU FIT 2015