UniqTag: Assign unique, reasonably stable, content-derived identifiers to genes

Shaun Jackman, Joerg Bohlmann, İnanç Birol

Abstract

Summary

UniqTag assigns unique identifiers to gene sequences, or even other arbitrary sequences of characters, that are derived from the k-mer composition of the sequence. Unlike serial or accession numbers, these identifiers are reasonably stable between different assemblies and annotations of the same data.

Availability and implementation

The implementation of UniqTag is available at `https://github.com/sjackman/uniqtag`

Supplementary data and the code used to generate it is available at `https://github.com/sjackman/uniqtag-paper`

Contact

Shaun Jackman <sjackman@gmail.com>

Introduction

Annotating genes, the process of identifying regions of a genome that code for genes, typically follows genome sequence assembly. These annotated genes are assigned unique identifiers by which they can be referenced. Assembly and annotation is often an iterative process, refining the method to produce better assemblies, or by the addition of the new sequencing data. Ideally these gene identifiers would be reasonably stable from one assembly to the next. Genes are typically assigned serial or accession numbers, which, although certainly unique, are not at all stable between assemblies. One small change in the assembly can result in a total renumbering of the annotated genes.

One solution to stabilize identifiers is to assign them based on the content of the gene sequence, rather than its position in the assembly. A cryptographic hash function such as SHA (Secure Hash Algorithm) (Dang, 2012) derives a message digest from the sequence, such that two sequences with the same content will

have the same message digest, and any two sequences with different content will have different message digests. When a message digest is used to identify a gene, the same gene in two assemblies with identical content will be assigned identical identifiers, but, by the design of a cryptographic hash function, even a slight change in the sequence, such as a single-character substitution, will result in a completely different digest and identifier.

A cryptography hash function is designed so that small changes in the message, even a single bit change, results in large changes to the message digest: an expected half the bits of the digest flip, called the avalanche effect (Feistel, 1973). In contrast, locality-sensitive hashing (LSH) aims to assign items that are similar to the same hash bucket. A hash function that, after a small perturbation of the sequence, assigns the same identifier to the sequence is desirable for identifying the genes of a genome sequence assembly project. One such locality-sensitive hash function, MinHash, was employed in identifying web pages with similar content (Broder, 1997). UniqTag implements MinHash, where the set of elements is the k-mer composition of the sequence, the hash function is the identity function and the minimal element is defined as the lexicographically minimal element, to assign reasonably stable identifiers to genes. These identifiers are intended for systematic identification, unique within an assembly, rather than as a biological name, which is typically assigned based on biological function or homology to orthologous genes.

Algorithm

The following symbols and terms are defined.

- Σ is an alphabet
- s and t are strings, a sequence of symbols from the alphabet Σ
- S is a set of strings over Σ
- ^{k} is the set of all strings over Σ of length k
- $\min S$ is the lexicographically minimal string of S
- $C(s)$ is the set of all substrings of s
- A k-mer of s is a substring of s with length k, also called an n-gram
- $C_k(s)$ is the set of all k-mers present in s
- $f(s, S)$ is the frequency of s in S, defined as the number of strings in S that contain s as a substring
- $f_k(s, S)$ is the frequency of the least frequent k-mers of s in S
- $F_k(s, S)$ is the set of the least frequent k-mers of s in S
- $u_k(s, S)$ is the UniqTag, the lexicographically minimal k-mer of the k-mers of s that are least frequent in S

The UniqTag $u_k(s, S)$ is defined as follows.

$$
\begin{aligned}
C_k(s) &= C(s)^k \\
f(s, S) &= |\{t \mid tSsC(t)\}| \\
f_k(s, S) &= \min_{tC_k(s)} f(t, S) \\
F_k(s, S) &= \{t \mid tC_k(s) f(t, S) = f_k(s, S)\} \\
u_k(s, S) &= \min F_k(s, S)
\end{aligned}
$$

Discussion

When iterating over multiple assemblies of the same data, it is rather inconvenient for gene identifiers to change completely from one assembly to the next. The gene identifier scheme described here attempts to address this common annoyance. By identifying the gene by its content, using the lexicographically minimal k-mer in the gene, the gene identifier is reasonably stable across assemblies.

A UniqTag will change due to either a difference in the locus of the UniqTag itself, or a difference that results in the creation of a unique k-mer that is lexicographically smaller than the previous UniqTag. Concatenating two gene models results in a gene whose UniqTag is the minimum of the two previous UniqTags, unless by chance one of the k-mer at the junction of the two sequences is lexicographically smaller. Likewise, splitting a gene model into two results in one gene that retains the previous UniqTag and a second gene that is assigned a new UniqTag, unless the previous UniqTag spanned the junction.

A UniqTag can be generated from the nucleotide sequence of a gene. Using instead the translated amino acid sequence of a protein-coding gene sequence results in a UniqTag that is stable across synonymous changes to the coding sequence, as well as to changes to the untranslated regions (UTR) and introns of the gene. Additionally since the amino acid alphabet is larger than the nucleotide alphabet, fewer characters are necessary for a k-mer to likely be unique, resulting in an aesthetically pleasing and shorter gene identifier.

Two gene models that have identical sequence will have the same UniqTag. It is possible that by chance two genes with different sequence that have no unique k-mer and similar k-mer composition are assigned the same UniqTag. This situation is most common for very short sequences. Genes that are assigned the same UniqTag are distinguished by adding a serial numerical suffix to the UniqTag.

Acknowledgements

Thanks to Nathaniel Street for his enthusiastic feedback, to the SMarTForests project, its funders and the organizers of the 2014 Conifer Genome Summit that made our conversation possible.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Genome British Columbia, Genome Alberta, Genome Québec and Genome Canada.

References

- Broder, A. Z. (1997) On the resemblance and containment of documents. *Compression and Complexity of Sequences*, 1997 Proceedings, 21-29.
- Dang, Q. H. (2012) Secure Hash Standard (SHS). *NIST FIPS*, 180(4), 1-35.
- Feistel, H. (1973) Cryptography and Computer Privacy. *Scientific American*, 228(5).