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

Shaun Jackman, Joerg Bohlmann, İnanç Birol

July 2, 2014

Summary

UniqTag assigns unique identifiers to gene sequences, or 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 code to reproduce it is available at
`https://github.com/sjackman/uniqtag-paper`

Contact

Shaun Jackman <sjackman@bcgsc.ca>

Introduction

The task of annotating the genes of a genome sequence often 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, by refining the method or by the addition of the more sequencing data. These gene identifiers would ideally be reasonably stable from one assembly

to the next. Genes are typically assigned serial or accession numbers, which, although certainly unique, are not stable between assemblies. A single 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. 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 two sequences that differ will have different message digests. If a cryptographic hash were used to identify a gene, the same gene in two assemblies with identical content would be assigned identical identifiers, but by design a slight change in the sequence, such as a single-character substitution, would result in a completely different digest and identifier.

A cryptographic hash function is designed so that small changes in the message, even a single bit change, results in large changes to the message digest: half of the bits of the digest are expected to flip, called the avalanche effect (Feistel, 1973). Locality-sensitive hashing (LSH) in contrast aims to assign items that are similar to the same hash bucket. A hash function that, after a small perturbation of the sequence, assigns an identical 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 of an item is the k-mer composition of the sequence, the hash function is the identity function and the minimal element is the lexicographically minimal sequence, 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 over Σ
- S is a set of strings over Σ
- 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

- $\arg\min_{t \in C_k(s)} f(t, S)$ is the set of the k-mers of s that are least frequent 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.

$$
\begin{aligned}
C_k(s) &= C(s) \cap \Sigma^k \\
f(s, S) &= |\{t \mid s \in C(t) \wedge t \in S\}| \\
u_k(s, S) &= \min \arg \min_{t \in C_k(s)} f(t, S)
\end{aligned}
$$

Results

UniqTag was used to assign identifiers to the protein sequences of six builds of the Ensembl human genome (Flicek, 2014) spanning five years. The overlap of UniqTag identifers between older builds and the current build 75 is shown in Table 1.

Build A	Build B	Only in A	In both	Only in B
55	75	24299	30997	17600
60	75	9365	34859	13738
65	75	1088	45931	2666
70	75	231	47955	642
74	75	0	48597	0

Table 1: The overlap of UniqTag identifers between older builds of the Ensembl human genome and the current build 75.

Discussion

When iterating over multiple assemblies of the same data, it is rather inconvenient when gene identifiers to change from one assembly to the next. UniqTag attempts to address this common annoyance. By identifying the gene by a feature of its content rather than an arbitrary serial number, the gene identifier is reasonably stable between assemblies.

A UniqTag will change due to a difference in the locus of the UniqTag itself, the creation of a least-frequent k-mer that is lexicographically smaller than the previous UniqTag, or the creation of a k-mer elsewhere resulting in the UniqTag no longer being a least-frequent k-mer. Concatenating two gene models results in a gene whose UniqTag is the minimum of the two previous UniqTags, unless one of the k-mer at the junction of the two sequences is lexicographically smaller. Similarly, splitting a gene model in 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 in the untranslated regions (UTR) and introns of the gene. Since the amino acid alphabet is larger than the nucleotide alphabet, fewer characters are required for a k-mer to be likely unique, resulting in an aesthetically pleasing shorter identifier.

Two gene models with identical sequence will have the same UniqTag. It is possible although uncommon that two non-identical genes with similar k-mer composition and no unique k-mer are assigned the same UniqTag. Genes that are assigned the same UniqTag are distinguished by adding a 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).
Flicek, P. (2014) Ensembl 2014. *Nucleic Acids Research*, 42(D1), D749-D755.