BIOINFORMATICS APPLICATIONS NOTE

Vol. 00 no. 00 2014
Pages 1–2

UniqTag: Content-derived unique and stable identifiers for gene annotation

Shaun Jackman[1,2,*], Joerg Bohlmann[3,4] and İnanç Birol[1,5 *]

[1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada
[2]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada
[3]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada
[4]Department of Forest Sciences, University of British Columbia, Vancouver, BC, Canada
[5]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

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 numbers, these identifiers are 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>
Inanc Birol <ibirol@bcgsc.ca>

1 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 more sequencing data. These gene identifiers would ideally be stable from one assembly and annotation to the next. Serial numbers are used for identifiers of genes annotated by software such as MAKER (Campbell, 2014), 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 unique 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 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 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.

2 DESCRIPTION

When iterating over multiple assemblies of the same data, it is rather inconvenient when gene identifiers change from one assembly to the next. UniqTag attempts to address this common challenge. By identifying the gene by a feature of its content rather than an arbitrary serial number, the gene identifier is 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 when a gene model is split in two, one gene is assigned a new UniqTag and the other retains

*to whom correspondence should be addressed

the previous 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 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 would be assigned the same UniqTag. It is possible that two genes that have no unique *k*-mer and similar *k*-mer composition are assigned the same UniqTag. Genes that have the same UniqTag are distinguished by adding a numerical suffix to the UniqTag.

2.1 Algorithm

The UniqTag $u_k(s, S)$, a substring of the string s with length k from the set of strings S, is defined as follows.

Σ is an alphabet. Σ^k is the set of all strings over Σ of length k. s and t are strings over Σ. $C(s)$ is the set of all substrings of s. A k-mer of s is a substring of s with length k. $C_k(s)$ is the set of all k-mers of s.

$$C_k(s) = C(s) \cap \Sigma^k$$

S is a set of strings over Σ. $f(s, S)$ is the frequency of s in S, defined as the number of strings in S that contain s as a substring.

$$f(s, S) = |\{t \mid s \in C(t) \wedge t \in S\}|$$

$\min S$ is the lexicographically minimal string of S. $u_k(s, S)$ is the UniqTag, the lexicographically minimal k-mer of those k-mers of s that are least frequent in S.

$$u_k(s, S) = \min \arg \min_{t \in C_k(s)} f(t, S)$$

3 RESULTS

UniqTag was used to assign nine-peptide identifiers to the first peptide sequence, that with the smallest ENSP number, of each gene of the Ensembl human genome. This analysis was repeated for nine builds of the Ensembl human genome (Flicek, 2014) spanning seven years and two major genome assemblies, NCBI36 up to build 54 and GRCh37 afterward. The number of common UniqTag identifers between older builds, from build 40 to build 74, and the current build 75 is shown in Figure 1. Also shown is the number of common gene and protein identifiers (ENSG and ENSP accession numbers) and the number of identical peptides sequences between builds. Although less stable than the gene ID, the UniqTag is more stable than the protein ID and the peptide sequence. Whereas the gene and protein identifiers can, with effort, be lifted over from older builds to the newest build, the UniqTag identifier can be generated without any knowledge of previous assemblies, making it a much simpler operation. The number of identical peptide sequences between builds shows the performance that would be expected of using a message digest, such as SHA-1, of the peptide sequence as the identifier. Supplementary figure S1 shows that the UniqTag identifiers are quite stable for values of k, the size of the UniqTag identifier, between 8 and 50 peptides.

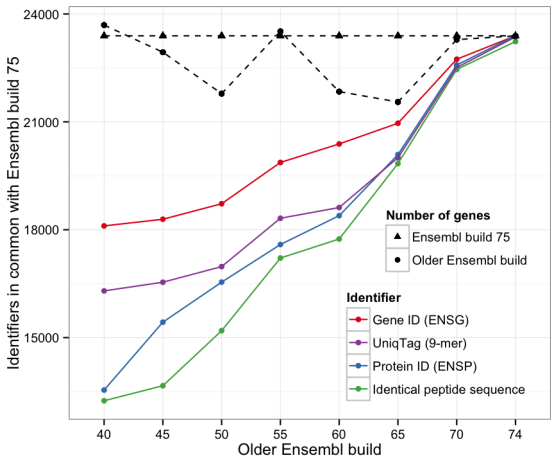

Fig. 1. The number of common UniqTag identifers between older builds of the Ensembl human genome and the current build 75, the number of common gene and protein identifiers, and the number of identical peptide sequences between builds.

ACKNOWLEDGEMENTS

The author thanks Nathaniel Street for his enthusiastic feedback, the SMarTForests project 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 Quebec and Genome Canada.

REFERENCES

Broder, A. Z. (1997) On the resemblance and containment of documents. *Compression and Complexity of Sequences*, 1997 Proceedings, 21-29.

Campbell, M. S. *et al.* (2014) MAKER-P: a tool-kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, 164(2), 513-524.

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. *et al.* (2014) Ensembl 2014. *Nucleic Acids Research*, 42(D1), D749-D755.