Independent Study Writeup

Nishant Jha

March 28, 2018

1 Private Edit Distance

The Human Genome contains around 3,200,000,000 base pairs, often denoted as 3,200 Mb [1]. Each base pair can have one of four values: (A, T, C, G) and so two bits are sufficient to represent all base pairs.

```
A => 00
T => 01
C => 10
G => 11
```

This also means a single byte can represent a sequence of 4 base-pairs. This means the Human Genome is about 800 megabytes long. Unfortunately, this is merely a theoretical lower bound. In practice, the average sequenced genome occupies around 200 gigabytes of disk space. This is a byproduct of how modern genetic sequencers work and the file format used to store the resulting data.

This unwieldy size necessitated an optimization that takes advantage of the large percentage of the genome sequence that is conserved between individual humans [2]. If a comparison needs to be made between two different genome sequences, a scientist can simply analyze how each genome sequence differs from a *reference genome sequence* rather than directly compare the base-pair sequences for each. This allows for a significant reduction in size. A fresh-off-the-sequencer genome is usually stored in a 200 GB FASTQ file, while the "diff" optimized storage method is stored in a variant (.vcf) file. This .vcf file usually only occupies around 125 MB which is more space efficient by orders of magnitude [3].

An analogous idea in computer science is that of *edit distance*. Put simply, the edit distance between two strings is the number of characters that must be substituted, inserted, or deleted to change one of the strings into another [4].

From both a practical and privacy standpoint, it is undesirable to send a patient's entire genome sequence whenever a comparison is required. Because each genome sequence is 200 gigabytes in size, bandwidth will quickly become a bottleneck. Additionally, genome

sequences are practically unparalleled in their ability to uniquely identify an individual. Furthermore, they contain sensitive medical data - not only the prescence of congenital diseases but predispositions to a wide swath of additional maladies.

The .vcf "diff" method only alleviates one of these problems - the one concerning bandwidth. Because the human reference genome sequence is publically available, reverse engineering an individual's genome sequence given a comprehensive variant file is trivial - simply start with the reference and make the changes described in the variant file. Similarily, if you are given a variant file that describes the set difference between your genome sequence and another person Y's sequence, the derivation of Y's sequence is trivial as well.

Clearly, we cannot allow this set difference to be calculated. However, if we take advantage of three features we can reduce the problem of calculating the set difference while maintaining privacy to calculating the *size* of the set difference while maintaining privacy.

1. Most differences between human genome sequences are *substitutions*, not insertions or deletions.

2. There exists a public reference genome and many variations that have already been computed.

3. Using probabilistic algorithms, the size of the set difference can be securely approximated while keeping the actual set difference private.

Talking about approximation protocol
Talking about approximation protocol
Talking about approximation protocol
Talking about approximation protocol
Talking about approximation protocol

This approximation process's accuracy is aided by the special distribution of human genome sequences.

For any two suitably unrelated individuals:

1. An overwhelming majority of their sequences are conserved, at least 99%

2. The places at which they differ, i.e. the location of their edits from a reference genome are not close together.

3. A majority of the differences between genome sequences are substitutions, around 80%

2 Honest-but-Curious Threat Model

3 Garbled Circuits

4 Differential Privacy

References

[1] https://www.nature.com/articles/35057062

[2] https://www.genome.gov/19016904/faq-about-genetic-and-genomic-science/

[3] http://www.internationalgenome.org/wiki/Analysis/vcf4.0

[4] http://repositorio.uchile.cl/bitstream/handle/2250/126168/Navarro_Gonzalo_Guided_tour.pdf