GENETIC INFORMATION

FOR PRIVACY AND SECURITY

GENEINFOSEC LLC

Sterling Sawaya, PhD

geneinfosec

http://www.geneinfosec.com

First printed, May 2017
Updated, September 2018

DOCUMENT BASED ON TUFTE LATEX BOOK TEMPLATE

Contents

Introduction

The purpose of this document is to provide the general understanding of genetic information needed to keep it secure. Information can be defined, in general, as data that reduces uncertainty. Here, the concept of *genetic information* will be defined as genetic data that reduces uncertainty about an organism or population. A patent pending method for concealing genetic information within genetic material is briefly described here*.

A basic understanding of genetic data is required to conceal genetic information. Genetic data is composed of sequences of nucleotide bases, abbreviated by A, C, T, and G. These sequences can be extremely long. The human genome, for example, has over 3×10^9 bases, and some organisms have genomes that are larger by many orders of magnitude. Perfectly concealing such large data-sets requires encryption methods that cannot be easily applied to genetic material itself. Specific types of genetic information, however, can be concealed within genetic material. For example, genetic material may be modified to conceal a single mutation. Data generated from such material might contain a large amount of genetic information, but sensitive genetic information could nevertheless be protected.

To protect sensitive genetic information present within genetic material a relatively simple approach can be employed. This approach has three general stages. First, the genetic material is divided into pieces. Second, separate pieces are randomly given their own unique molecular identifier. Third, the genetic material is combined with other pieces of genetic material that have also been given unique identifiers. The utility of this approach relies on an important property of genetic information, genetic data only provides useful information in aggregate.

*More details about this method can be found at geneinfosec.com.

For example, consider the sequence $ACGT$. This short sequence occurs multiple times in every genome of every species, and therefore cannot be used to identify a species, subspecies, family, or individual. Due to its ubiquity, it provides almost zero information. This sequence would only reduce uncertainly if one had never encountered earthly genetic material. Trivially short genetic sequences contain minimal genetic information, but as sequences increase in length, their information content increases rapidly. Consider this 20 base-pair sequence:

$GTGCCAGCAGCCGCGGTAAT$

This short sequence can be recognized as belonging to an RNA enzyme that has been highly conserved during evolution. However, although its function can be determined, the exact species from which it was obtained remains unclear. Expanding this sequence to 80 base-pairs in length can produce:

$GTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATA$
$TTAAAGTTGCTGCAGTTAAAAAGCTCGTAGTTGGATCTTG$

This longer sequence can now be recognized as belonging to a specific human gene, $RNA18S5$. Generally, as sequences increase in length their information content also increases. However, in many situations only genetic *variants* convey genetic information. Longer sequences can, but don't always, contain a larger number of informative genetic variants. Some genomic regions are more variant than others, and these variations provide information about the organism from which they originated.

To model privacy of genetic sequence data, one can build a model in which an adversary has a set of knowledge that is updated when sequence data is accessed. Though analytically tractable, these models rely on potentially erroneous assumptions about adversarial knowledge. Alternatively, the method of *differential privacy* can be employed[1]. Differential privacy compares two data-sets that differ by only one element, D and D'. It relies on a randomization function, K, to modify the data so that the addition of one element has a measurable impact, ϵ:

$$Pr[K(D) \in S] \leq exp(\epsilon)\ x\ Pr[K(D') \in S]$$

[1] *Differential Privacy* (2006) by Cynthia Dwork at Microsoft Research.

The randomization function is chosen so that ϵ is small enough to satisfy privacy concerns. Roughly speaking, this approach randomizes data so the addition of an individual's data results in a negligible change to the data-set, providing plausible deniability that an individual's data is present within that data-set.

The application of differential privacy here differs from previous approaches. Here we apply a randomization function to genetic material itself, generating a pool of genetic material to conceal sensitive genetic information. That is, K is a molecular modification of genetic material that produces a pool to be sequenced, generating genetic data-set, $K(D)$. The randomization methods can be designed so that the addition of an individuals genetic material results in a nearly undetectable change in the sequenced data.

With $S \subseteq Range(K)$, the potential output of the randomization function is given limits. For genetic material, $S \subseteq \mathbb{G}$, where $\mathbb{G}$ represents all possible genetic sequences. However, not all possible genetic sequences are relevant, and therefore $\mathbb{G}$ is often limited to a subset of potential genetic sequences, $\mathbb{G}_i$, where $\mathbb{G}_i \subset \mathbb{G}$.

Genetic information will be examined in three distinct forms:

- **Phylogenetic information** is a relative measure of the evolutionary relationship between genetic sequences. For example, uncertainty about the species from which a genetic sequence originated is reduced when that sequence is compared to known genome sequences. This information is often easy to extract from genetic data, and therefore is the most difficult to conceal.
- **Pedigree information** is similar in form, but not degree, to phylogenetic information. Pedigree information reduces uncertainty about how individual organisms are related. Pedigree information can be considered familial information when different genetic lineages are given family names. Pedigree information requires detail at a different scale than phylogenetic information, and is therefore easier to conceal.
- **Quantitative information** reduces uncertainty about traits present in an organism. Quantitative genetic information can be used to predict observable traits, such as height, weight and color, and thus can be used to identify the source of the genetic data. It can also be used to predict disease predisposition, behavioral tendencies, or other sensitive traits.

Some forms of genetic information are easier to conceal than others.

Concealment methods can favor concealment of one form of genetic information over others.

Phylogenetic Information

Phylogenetic information is obtained by comparing sequences. Most species have been evolving for millions of years since they last shared a common ancestor, resulting in large differences between their genome sequences. These differences can be used to measures the evolutionary distance between genetic sequences and construct a phylogeny (Fig. 1) representing the evolutionary relationship between species.

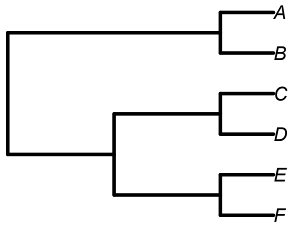


Figure 1: **A simple phylogeny** representing the evolutionary relationship between six species: A, B, C, D, E, F. Phylogenies are also sometimes called evolutionary trees.

The genome sequences from various plant and animal species are publicly available, allowing relevant genetic sequences to be attributed to those species. Even if the genome sequence of a species is unknown, its relationship with other known genomes can quickly be determined. The availability of genome sequences, as well as the ease by which sequences can be compared to make a phylogeny, hinders the concealment of phylogenetic information. Although species information is unlikely to warrant extreme security measures, understanding how phylogenetic information can be concealed can provide useful insight into concealing other forms of genetic information.

Consider the phylogeny in Figure 1. These distinct species each have their own unique genome sequences $\mathbb{G}_i$, and some genomes are more closely related than others. For example, $\mathbb{G}_A$ and $\mathbb{G}_B$ are more closely related to each other than the other species. We can utilize these relationships when seeking to conceal genetic information. If we seek to conceal genetic material from A, we start by breaking up the material and uniquely identifying the pieces. If these pieces are then mixed with a random set of material from this phylogeny, some concealment is possible. Using:

$$Pr[K(D) \in \mathbb{G}_i] \leq exp(\epsilon)\ x\ Pr[K(D') \in \mathbb{G}_i].$$

if $\mathbb{G}_i \subset \mathbb{G}_{A,B,C,D,E,F}$, the resulting pool can offer some privacy for genetic material from any species $A - F$.

Conceptually, phylogenetic information is concealed here because the pool of material to be sequenced contains a random amount of genetic material from each species. Genetic sequences generated from this pool reveal that organisms from this phylogeny are being analyzed, but the species of interest can be concealed.

Depending on the amount of privacy required, $S \subset \mathbb{G}_{A,B}$ may be sufficient when concealing A or B. Note that pooling distantly related species, such as $S \subset \mathbb{G}_{A,F}$, is weaker than using closely related species. Closely related species share more sequences in common, resulting in similar genomes that can be difficult to tell apart. In fact, when species are very closely related and their genetic sequences become similar enough, one species can be used to conceal the other. That is, sufficient privacy for A may be obtained with $S \subset \mathbb{G}_B$!

The concept of phylogenetic information is critical to concealment of any genetic information, even when species information has no need for concealment. Phylogenetic information can almost always be obtained from sequence data. This information reveals more than just the species from which it originated; it provides sub-species and lineage specific information that may indicate the regions or ethnicities from which the genetic material was obtained. To prevent a pool of material from changing when an individual's genetic material is added, the pool must contain phylogenetically relevant material.

For a trivial example, consider that the genetic material of a fish would be easily differentiated from a pool of mammalian genetic material. Pooling genetic material from one fish with another fish, on the other hand, could provide sufficient privacy if the other fish were closely related. For a more practical example, consider the genetic material from a distinct lineage of humans, such as that from an African Pygmy. Adding their genetic material to pool of closely related African tribes would provide greater privacy than if it were added to a pool of distantly related humans (or even worse, added to a pool composed of a fish's genetic material).

Concealing phylogenetic information comes at a cost, and ultimately some phylogenetic information must be disclosed. When concealing any form of genetic information, one must understand how phylogenetic information is disclosed or concealed during sequencing.

Pedigree Information

Much like phylogenetic information, pedigree information compares sequences to determine how they are related. The relationships examined in a pedigree, however, are relationships between individuals, not just individual pieces of DNA. To build a *phylogeny*, variations between sequences are used. Alternatively, to build a *pedigree*, variations between sequences can be used, but depending on the size of the pedigree, the co-occurrence of these variants within individual genomes is often more useful. For example, diploid genotypes* can be used to determine, or rule out, parentage. An individual with genotype A/A can only be formed by parents with genotypes of A/a or A/A, unless a mutation occurs in a parent with genotype a/a.

Groups of variants, and their co-occurrences, provide pedigree information. For example, the co-occurrence of specific variants found within haplotypes* can provide information about family lineages. Consider a situation in which the haplotype $A - B - C - D$ is very rare, and consequently this haplotype is sufficient to identify a specific family. A complete genotype from an individual, such as $A - B - C - D/a - b - c - d$, may therefore be sufficient to identify a specific individual in this family.

Genetic information can be concealed at the level of genetic material by physically disrupting the connections between genetic variants. Simply disrupting these connections can work to conceal haplotypes and diploid genotypes. If the disrupted genetic material is given unique random identifiers and combined with relevant genetic material, much of the pedigree information can be concealed. However, this general approach does not conceal all genetic information.

*A *diploid genotype* consists of two genetic regions, one inherited from the mother and one from the father.

*A *haplotype* is a group of variants inherited from a single parent, sometimes linked together on a chromosome.

Revealing some genetic information can allow other genetic information to be perfectly concealed. Consider the pedigree found in Figure 2, in which many family members have an unknown BRCA1 sequence (?/?). The presence of two family members that have tested positive for a harmful BRCA1 variant, individuals 202 and 203 (Carrie and John Jr.), indicates that this variant was inherited from one of this pedigree's grand-parents, 101 or 102.

Denote the harmful variant as "a", and a normal variant as "A" and assume that genotypes of the grandparents are A/a and A/A. Individuals 201 and 204 have been sequenced, and found they did not have harmful variant (their genotypes are A/A). Individuals 202 and 203 have been sequenced and are known to be carriers of the harmful variant (their genotypes are A/a).

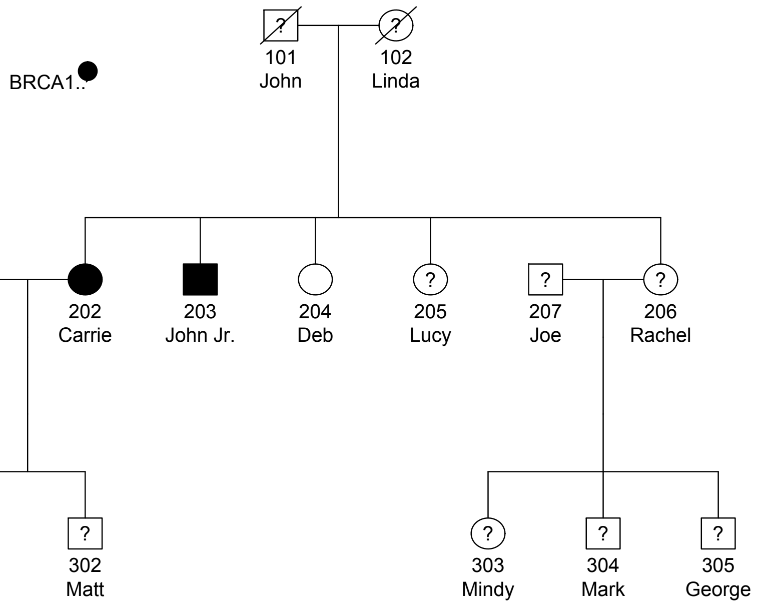


Figure 2: **A pedigree of a family** that carries a medically relevant genetic variant (here a BRCA1 mutation). Two family members have tested positive for the mutation, and two have tested negative. The rest of the family's genotypes are unknown.

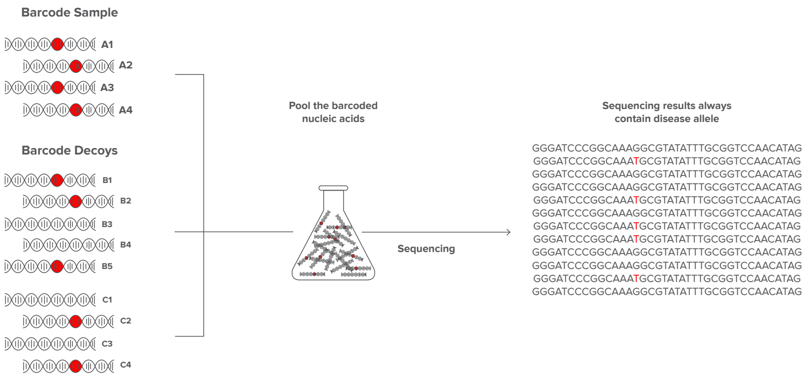
Now consider the origin of the genes found within individuals 202-206, the children of 101 and 102. The genes of the offspring originate from the genes of their parents (assuming there are no new mutations). That is, $\mathbb{G}_{202-206} \subset \mathbb{G}_{101,102}$. This rather obvious property of genetics can be utilized to conceal genetic information.

For an example, the children Mike and Matt (301, 302) desire to know their BRCA1 status. Their parents have been sequenced for this gene. Carrie (202) is a carrier of a harmful mutation (A/a), and her husband Jack (201) has been determined not to carry any harmful variant (A/A). These children have genotypes that are a combination of their parents, thus $\mathbb{G}_{301,302} \subset \mathbb{G}_{201,202}$. Using randomized barcodes, a sequencing protocol which pools the genetic material of these children with their parents genetic material may provide sufficient privacy. This pool would contain alleles A and a regardless of the genotypes of the offspring.

Even when the genotypes of the parents are unknown, e.g. for the offspring of Joe and Rachel (206, 207), this sequencing protocol can nevertheless offer differential privacy. The second generation of this family, 202-206, could obtain privacy if they had access to the genetic material of their parents, 101 and 102. If the genetic material of the parents is not available, a sufficiently large number of offspring can be pooled with a similar effect. That is, $\mathbb{G}_{202-206} \approx \mathbb{G}_{101,102}$.

Quantitative Information

Quantitative genetic information can be the most difficult to conceal. Like other forms of genetic information, quantitative information often involves multiple variants. However, unlike genetic information used to build phylogenies or pedigrees, quantitative information can be obtained from a single variant (e.g. a BRCA1 variant and cancer susceptibility). Therefore, simply breaking up the genetic variants and pooling them with phylogenetically similar variants may not provide sufficient security.


Barcode Sample
A1
A2
A3
A4
Barcode Decoys
B1
B2
B3
B4
B5
C1
C2
C3
C4
Pool the barcoded nucleic acids
Sequencing
Sequencing results always contain disease allele

```
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAATGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAATGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAATGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAATGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAATGCGTATATTTGCGGTCCAACATAG
GGGATCCCGGCAAAGGCGTATATTTGCGGTCCAACATAG
```

Figure 3: **Specific decoy nucleic acids can be used to conceal quantitative genetic information.** Decoys with a known disease allele are added to the pool to conceal whether the sample contains the disease allele.

To conceal a specific variant, a pool can be randomly generated to contain that variant in a random quantity. Consider the example in Figure 3. The data from the sequenced pool contains the disease variant at a random quantity, irrespective of the presence or absence of that variant within the individual being tested. Sufficient privacy can be obtained, provided that the pool is generated using phylogenetically relevant decoys. However, if an individual happens to have a disease variant that is phylogenetically distinct from the decoy allele variants, then the presence of their unique disease variant can be detected in the sequence data.

How can the decoys used in the pool be assured to be phylogenetically appropriate? Using the pedigree method discussed in the previous section, recent ancestors, such as the parents, can provide the necessary material. However, this approach requires access to the ancestral genetic material, and also requires that some of the genetic information of the ancestors is revealed. Furthermore, the exact genetic sequence found in a sample is usually unknown, otherwise the sequencing the sample would be unnecessary. Therefore, decoys must be carefully chosen to form appropriate pools for concealment.

In some cases, a trusted third party can generate decoys. This requires sufficient knowledge of the genetic material that is to be concealed. If the genetic sequences of relatives or ancestors is known, but perhaps their genetic material is not available, then these sequences can be used to design appropriate decoys. The trusted third party may have access to the relevant genetic material of the ancestors or relatives, and thus provide it for the consumer.

Importantly, an alternative approach can be taken to ensure that the decoy genetic material is appropriate. The genetic material to be concealed can be copied and mutated to become appropriate decoys! Using polymerase-chain-reactions, or site-directed mutagenesis, decoys with similar but not identical sequences can be generated by the consumer. The consumer can control how their genetic material is concealed. This approach can be tailored by each consumer, individually for each of their sequencing pools. Thus, the polymerase-chain-reaction itself can become a randomization procedure, mutating target nucleotides to conceal their information. The unpredictable nature of DNA polymerization offers extensive potential for cryptography of genetic material.

Summary

These methods allow genetic information to be concealed within genetic material. The process can be simple, the material from multiple samples can be pooled after random barcodes are attached. Advanced applications of these methods can offer greater concealment. However, to conceal genetic information one must first understand how information exists within genetic material. With this understanding, specific types of information can be concealed perfectly. Scientists utilizing this method can control which forms of genetic information they conceal and how deeply that information is hidden.

Differential privacy can be used to help determine the amount of privacy lost when adding an individual's genetic material to a pool. This method may be sufficient for many situations. However, modeling of adversarial knowledge may be necessary to obtain higher levels of genetic information security. If correctly designed, concealed genetic material will provide limited information to an adversary with access to the sequence data generated from that material.

The most advanced use of this method allows genetic material to be mutated to act as decoys for itself. Such advanced methods are the "secret sauce" that scientists can create to mitigate their own unique privacy concerns. The randomness of DNA polymerization allows for the cryptography of genetic material using molecules within a test-tube. The sequencing data generated from the modified genetic material can thus provide privacy guarantees. The genetic sequencing data generated by this method can then also have standard information security methods applied. Therefore, cryptography of genetic material offers an independent addition to information security, ensuring that sequencing data files do not have the potential to reveal sensitive genetic information.