Protecting DNA Sequence Anonymity with Generalization Lattices

Bradley Malin

October 2004

CMU-ISRI-04-134

Data Privacy Laboratory
Institute for Software Research International
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The increased collection, storage, and analysis of person-specific DNA sequences poses serious challenges to the protection of the identities to which such sequences correspond. Compromise of DNA privacy via re-identification, the inference of explicit identity of the individual from which the DNA was derived, is dependent on unique features that may be inferred from a DNA sequence. In this paper we introduce a computational method for anonymizing a collection of person-specific DNA database sequences. The method is termed DNA lattice anonymization (DNALA), and is based upon the privacy protection schema of k-anonymity. Under this model, it is impossible to observe or learn features that distinguish one genetic sequence record from $k-1$ other entries. We employ a concept generalization lattice to determine the distance between two residues in a single nucleotide region, which provides the most similar generalized concept for two residues (i.e. adenine and guanine are both purines). Each single nucleotide region is considered independent of each other region when determining the distance between sequences. The DNALA method chooses pairs of sequences to be anonymized to a sequence of minimal distance between the pair, and generalizes the pair accordingly. The method is tested and evaluated with several publicly available human population datasets.

Keywords: anonymity, confidentiality, privacy, genomics, genetic databases, k-anonymity

Contents

1 Introduction

Current statistics compiled by the National Center for Biotechnology Information demonstrate, that as of December 2002, approximately 14000 human genetic loci have been established, hundreds of which have been characterized as influencing genetic disease and polymorphic sites. [1]. The discovery and physical mapping of human genetic components have greatly benefited by recent technological developments in molecular biology, automated sequencing, and digital storage technology, thus allowing for an exponential increase in the discovery and differential analysis of genetic loci. [2] The rise in genetic data and resulting databases has a variety of uses including genetic and molecular biology basic research, clinical medical research, biopharmaceutical research and development [3], public health surveillance [4], and occupational safety [5]. Yet, despite the considerable benefits to research that will be produced with the collection of such data, scientists and society must consider the challenges to ensuring privacy that can occur when large amounts of information are collected on person-specific populations.

In recognition of the previous, the privacy of an individual's genetic information has been discussed at length in several communities, including those pertaining to law, public policy, molecular medicine, biopharmaceutical industry, and public health. [6] Discussions within and between such communities have, for the most part, focused on issues of 1) ownership of information, 2) the ethical duty of physicians and counsellors to protect their patient's rights, and 3) genetic discrimination. The previous arguments relate to the direct release and use of genetic information, however, genetic information is useful in realms beyond initial collection. Many groups harboring collections of genetic information share, or hope to do so in the future, databases for various endeavors, such as licensing to a private or academic research group, public use datasets, public health research, or various other groups. The protection of such data, due to its potential clinical, molecular, and pharmacological relations, has been labelled as one of the foremost challenges to the pharmacogenomics community. [7]

Recent research has demonstrated that DNA sequence data, devoid of any additional information beyond that of the originating institution, is vulnerable to attacks on privacy. [8, 9] Thus, it begs the question, "How should DNA sequences be anonymized to protect the identity of the individuals to which the sequences correspond?" More specifically, this research considers how to thwart re-identification of DNA through a trail attack. This attack exploits longitudinal medical information about an individual. Oftentimes an individual visits multiple data collecting institutions, and the individual may leave behind data corresponding to their medical or DNA records. As such, in a trail attack, DNA samples are matched to their identified personas through the use of unique distinguishing features in the set of institutions visited by the identified individual and the unidentified DNA data that has been left behind. A more detailed description of the trail attack can be found in. [10] Therefore, if we permit a data releasing institutions to share DNA that can not be correctly tracked, we can prevent re-identification.

Previous proposals to anonymize DNA have concentrated on single nucleotide polymorphism (SNP) regions for privacy protection. These protection methods are based on the notion that the majority of SNPs consist of two different residues only. However, when considering data collected for mutational analysis of a gene with variation beyond SNPs, variation beyond that of SNPs are possible, including single nucleotide variation, insertions, deletions. A robust anonymization schema must account for all possible variations in DNA sequence data. This paper presents a method called DNA Lattice Anonymization, or DNALA, that adheres to the privacy protection model of k-anonymity. The method protects privacy by guaranteeing that the DNA sequence of one individual will be exactly the same as the sequence of one other individual in the released data from a collecting institution. When a data collecting institution releases DNA sequence data under this method, the identity of every DNA sequence is guaranteed to be ambiguous to at least one other identity.

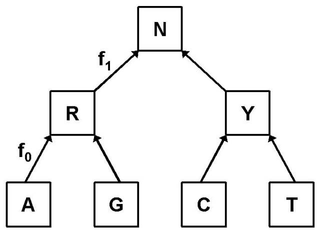


Figure 1: DNA generalization hierarchy for purines and pyrimadines.

2 Computational Disclosure Control

2.1 k-anonymity

This research builds upon previous work in the computational disclosure control for field-structured databases. The following definitions and concepts are derived from [11] in particular. The term data refers to entity-specific information, which is organized as a table of rows (records) and columns (fields). Each row of the table is referred to as a tuple and each column is referred to as an attribute. Each attribute can be thought of as a semantic category of information with a set of values. Since this research is concerned with the relationships between tuples in tables, let us define a table as $\tau(A_1^\tau, A_2^\tau, \ldots, A_m^\tau)$, where the set of attributes for table τ is $A^\tau = \{A_1^\tau, A_2^\tau, \ldots, A_m^\tau\}$. A tuple t of the table c for institution c is defined as $t[a_i^c, \ldots, a_j^c]$ and represents the sequence of values, $v_i \in A_i^c, \ldots, v_j \in A_j^c$. Fields in the table are referred to as attributes, each of which can be thought of as a semantic category consisting of a set of values. The value set available to an attribute A is organized into a domain generalization hierarchy DGH_A. Given an attribute A, a generalization for an attribute can be defined as a function, such that $f : A_i \rightarrow A_j$ is a generalization from level i to level j of a hierarchy. The following is defined as a generalization sequence:

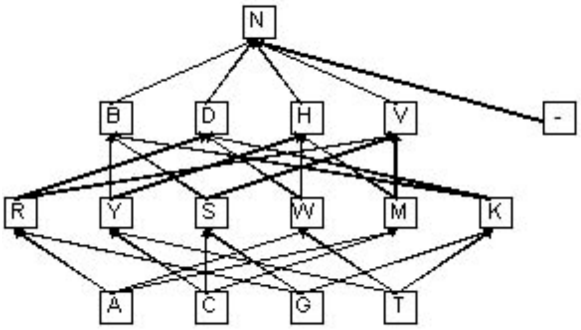
$$A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} \ldots \xrightarrow{f_{n-1}} A_n$$

In a DGH, the hierarchy is linear and unambiguous, such that a value can generalize to only one value per level in the hierarchy. The final level in the hierarchy consists of one value, which is suppression, or indeterminate. For example, consider the employment of the DNA generalization hierarchy in Figure 2.1. If we wish to generalize C and G together, we only need to generalize up one level, since f_0(A) = R and f_0(G) = R. To relate A and T, we must generalize to the indeterminate character N, since f_0(A)=R f_0(T)=Y, but a two level generalization does yield equality, since $f_1(f_0(\text{A}))$ = N equals $f_1(f_0(\text{T}))$ = N.

Not all attributes may be useful for re-identification purposes, and can be left in an ungeneralized state during the anonymization process. The set of attributes that are sensitive to re-identification in a table are termed the quasi-identifier, $QI^\tau = \{A_i^\tau, \ldots, A_j^\tau\}$. The privacy protection schema of k-anonymity for field structured databases is designed as follows. Given a table τ, with attribute set A^τ and quasi-identifier QI^τ, return a table τ', such that for every tuple $t \in \tau'$, there exists a minimum of $k - 1$ other tuples in the table that are indistinguishable from each other on the values of their quasi-identifier.

2.2 SNP Disclosure Control

Lin et al. [12] provide the first documented attempt to anonymize DNA sequence database entries. Their goal is the anonymization of single nucleotide polymorphisms (SNPs) in DNA sequences through the use of generalization hierarchies. Under their methodology, for every SNP position in a sequence, all sequences are generalized such that there exists a minimum number of other sequences with the same value in the position. This minimum value is referred to as a bin size. In addition, combinations of values from multiple SNP



A = Adenine	C = Cytosine
G = Guanine	T = Thymine
R = puRine	Y = pYrimadine
S = Strong hydrogen	W = Weak hydrogen
M = aMino group	K = Keto group
B = not A	D = not C
H = not G	V = not T
– = gap	N = iNdeterminate

Figure 2: left) DNA generalization lattice employed in the DNALA system. right) International Union of Biochemists (IUB) code for DNA and associated ambiguities.

regions are considered and the sequences are again generalized, such that for each combination of values there exist at minimum k-1 other sequences with the combination of values. There are several deficiencies to this model. First, the model does not scale well to regions with variation greater than 2 residues. Additional information loss will occur from generalization of single nucleotide positions without consideration for the relationship between all variant regions. For example, a ’C-G’ transversion value and a ’C-T’ transversion value will automatically be generalized to the genomic position of the SNP, instead of utilizing the union of the values, which would be C-G-T and still exclusionary on A. Second, there is no attempt to minimize the distance between sequences before data is generalized. Instead of anonymizing each polymorphism independently, it would be more beneficial to analyze the distance between sequences over all polymorphisms at the same time. As such, current SNP anonymization techniques will tend to overgeneralize released data. Such problems we attempt overcome in the methods provided below.

3 DNA Lattice Anonymization

The DNA Lattice Anonymization method, or DNALA, proposed below anonymizes DNA sequences by generalizing each sequence and its most similar sequence to a common sequence. Thus, the protection provides is k-anonymity with k equal to 2.

3.1 Domain Generalization Hierarchies

The similarity between sequences is based on a new type of generalization method, which we refer to as a domain generalization lattice (DGL). Conceptually, it is related domain generalization hierarchies, however, as will become apparent, generalization hierarchies are actually a special case of generalization lattices. In a DGL, a proper generalization function does not necessarily exist for each level of the graph. Rather, the generalization of a value may legally proceed to any value in a set of designated generalized values. Figure 2 demonstrates such a lattice, accompanied by concept/symbol definitions. This lattice is designed from the International Union of Biochemists nucleotide representation code [13] and is a representation of the union of all possible trees for single nucleotide generalization hierarchies, such as the one depicted in Figure 2.1.

It is through such a generalization lattice that the distance between two nucleotide values is determined. The leaf level is designated level 0, and is accessed as $level(residue)$, with each level above being one integer value greater. The distance function $gen(x, y)$ returns the distance in the lattice between concepts x

```
     1        10           20
S1 | actcactgaattactgactg
S2 | a---actgaatgactgactg
S3 | agagactgattcactgactg
S4 | agcaactgaatgactgactg
```

SNVR1	SNVR2	SNVR3	SNVR4	SNVR5
C	T	C	A	T
–	–	–	A	G
G	A	G	T	C
G	C	A	A	G

Figure 3: left) Sample sequences and right) resulting SNVRs

and y as follows. Let z be the value that x and y generalize to:

$$gen(x, y) = 2level(z) - level(x) - level(y) \tag{1}$$

Consider several examples: gen(A,C)=2 and the generalized value is M, gen(Y,S)=4 and the generalized value is N, and gen(A,-)=4 and the generalized character is N. The distance defined in this manner provides a measure of the number of residues of ambiguity that are added to both sequence positions by generalizing the values. The measure is both independent of the level of the original sequence values and the common generalized value.

3.2 DNALA

Here, we step through the core DNALA method, with the formal method provided in Algorithm 1.

Step 1: Identify variable regions. The first step of the DNALA algorithm is to identify single nucleotide variable regions ($SNVR$) in the DNA. We refer to these regions as $SNVRs$, since certain variants may not occur in a large enough proportion of the population to be considered as SNPs. Regardless, if there is any variation, then this region must be accounted for before data can be released and, therefore, will be useful for determining the distance between sequences. Each $SNVR$ will be an attribute for our DNA sequence table quasi-identifier, such that the quasi-identifier for a DNA sequence database will be $QI^{DNA} = \{SNVRDNA_1, \ldots, SNVRDNA_n\}$. Since raw DNA sequences for the same region of DNA can vary in length due to deletions and insertions in the sequence, we choose to make a global multiple sequence alignment (MSA) of the DNA sequences considered for anonymization. One can use any multiple sequence alignment (MSA) software or technique; for this study we employ CLUSTALW. [14] Once our MSA is determined, we identify SNVRs by simply determining when a position in the MSA has at least one sequence with a differing value than another sequence. For example, consider the sequence alignment shown in Figure 3. In the alignment, there exist five $SNVRs$ at positions 2, 3, 4, 10, and 12.

Step 2: Construct distance matrix. To find the most similar DNA sequence t for a particular sequence s, we measure the distance between sequences as follows. Each $SNVR$ is considered independent of every other $SNVR$. The distance between two sequences s and t is calculated as the sum of generalization distances between the residues in each variable region in the set of $SNVRs$ (V):

$$dist(s, t) = \sum_{v \in V} gen(s[v], t[v]) \tag{2}$$

where $gen(s[v], t[v])$ corresponds to equation 1. The gen function simply returns the minimum distance for x and y to generalize to a common concept as specified in the domain lattice. The gap value '-' is considered to be at the second level of the hierarchy.

Step 3: Pair off and generalize sequences. The pairing of sequences for anonymization proceeds in a greedy manner. For each sequence, which we call the referring sequence, the set of closest sequences, or proposed sequences, as designated by the distance matrix, is determined. This procedure is guaranteed to return a minimum of one proposed sequence for each referring sequence. Next, we iterate through the set of proposed sequences and, for each, it is determined if the referring sequence is in the proposed sequence's list of closest sequences. The occurrence of such an event we call reciprocity. If reciprocity exists, then we 1) cease our search for a sequence match, 2) pair off the referring and current proposed sequence, and 3) remove the two paired sequences from further consideration for any other sequence pairings. In the event that no reciprocity exists for the referring sequence, then we simply do nothing with the referring sequence and attempt to pair the next sequence in the set of unpaired sequences. After the final sequence in the set of unpaired sequences has been involved in an attempting pairing, there may still exist unpaired sequences. Thus, the above process is iterated until no sequences remain unpaired. In the event that there are an odd number of sequences, we simply pair the residual sequence, with the closest generalized sequence.

The above process is guaranteed to converge and can never run infinitely. The proof of this claim is simple and is based on the least element principle, with reference to the distances between sequences. Let D be the set of distances as defined by the distance matrix, such that each element corresponds to the distance between two different sequences. Every element within D is comparable, since the distances between sequences are considered as scalar values. Therefore, there must exist at least one element that is the minimum scalar value. This element corresponds to the occurrence of reciprocity and it can be removed from D. When this element is removed from the set D, as well as all other elements corresponding to either of the sequences for the removed element, then a new least element will exist. This process will continue until D is a null set and all sequences have been paired.

After all sequences are paired, the sequences are generalized according to the domain generalization lattice. The released set of sequences will be the generalized sequences with the gaps removed, which were inserted during the alignment process.

The DNALA system, as described, will provide a set of anonymized DNA sequences, however, the solution may not be optimal for all pairings due to the greedy aspect of the pairing process. This problem is derivative of the fact that as soon as reciprocity occurs, the sequences are paired off and not considered in for other sequences. While an optimal pairing may be found for two considered sequences, a later pairing may be forced into a non-optimal pairing. As such, one can derive a different set of anonymized sequences due to different pairings. Therefore, we introduce a probabilistic component to the anonymization technique to increase the number of optimal or near-optimal pairings. Steps 2-4 of the DNALA algorithm are repeated, which corresponds to all steps minus the identification of $SNVRs$ and the augmentation of the sequences, x times. For each repeat, we randomize the order in which 1) the unpaired sequences are searched and 2) the proposed set of sequences is searched. For each iteration, we keep a running total for the number of times each sequence i is paired with sequence j. The number x is chosen, such that we can identify the best matches for each sequence with high certainty. For evaluation purposes, we set x to a large value (i.e. 1000), beyond what would be the derived threshold. By this method, we will be able to match a majority of sequences to their best pair, but others will still be matched to their non-optimal pair.

3.3 Core DNALA Complexity

A brief overview of the complexity of DNALA is provided with respect to each major process of the system. The determination of variable regions is approximately O(mn), where m is the resulting length of the global MSA and n is the number of sequences considered for anonymization. Construction of the distance matrix will be approximately the O(n^2), where n is the number of input sequences. Now, as demonstrated above,

Algorithm 2 DNALA core system.

`Input:` $S = \{s_1, s_2, \ldots, s_n\}$`, a set of aligned sequences`
`Output:` $T = \{t_1, t_2, \ldots, t_n\}$`, a set of 2-anonymized ungapped sequences`
Identify variable regions
Construct distance matrix $D = S \times S$
Let P be a set of paired sequences, initially set to $\emptyset$
while $S \neq \emptyset$ **do**
  for each sequence $s \in S$ **do**
    Determine the set of closest sequences C_s to s
    for each $c \in C_s$ **do**
      Determine the set of closest sequences C_c to c
      if $s \in C_c$ **then**
        $P = P \cup < s, c >$
        $S = S - \{s, c\}$
        Break from internal loop
      end if
    end for
  end for
end while
Let $T = \emptyset$
for each pair $< s, c > \in P$ **do**
  Generalize $< s, c >$ to $< s', c' >$ such that $s' = c'$
  $T = T \cup \{s' - \text{“-”}, c' - \text{“-”}\}$
end for
return T

sequence pairing will converge, however, this convergence is bounded by the following equation:

$$\text{\# of comparisons} = \sum_{i=0}^{n/2} (n - 2i)^2 \approx n^3 \tag{3}$$

Thus the pairing process has a potential to be polynomial in approximately O(n^3). The generalization of pairs process will be O(nx) where x is the number of $SNVR$s. Since, the iteration of DNALA, to retrieve more probable paired sequences is simply a constant factor scaling to the DNALA algorithm, the overall complexity of DNALA is dominated by the pairing process and is O(n^3).

4 Evaluation

Several datasets are chosen to evaluate the DNALA system on. The first set of sequences consists of 54 human DNA sequences drawn from a 6.6kb region of the melanocortin gene promoter (MC1R) and is described in [15], where the sequences were used for analysis of general sequence variation. The second set of sequences consists of 30 sequences collected on a 4.2kb region of the pyruvate dehydrogenase E1 subunit locus (PDHA1), which is described in [16]. The third set, initially described in [17] is much larger in size, in that it is made up of 372 human mtDNA sequences of the hypervariable segment I control region (HVS1). Beyond the fact that all of these are human sequences, there are several additional reasons that help explain why such datasets are chosen for evaluation of the DNALA system. First, the range in the number of sequences helps to provide a relatively large sample of human DNA sequences for the same genomic region. Second, the data is a publicly available human population-based dataset and is a real example of the type of

Sample	Number of Sequences	Number of SNVR Regions	Increase in total lattice levels	Number of gaps anonymized	Number of SNVR Regions Used	Change in Average Distance
MC1R	54	95	0.88 (1.24)	0.44 (0.48)	4.03 (3.7)	2.88 (2.13)
PDHA1	30	266	0.5 (2.04)	0.25 (1.01)	4.30 (5.96)	5.71 (3.98)
HVS1	372	237	0.39 (3.88)	0.19 (1.92)	3.04(4.43)	3.59 (4.86)

Table 1: Generalization information for evaluation sequences. For cells with multiple numbers, the average score is followed by the standard deviation.

data that might be collected from a research institution. In fact, the NCBI is currently in the process of making such population-based data available to researchers and the general public through the PopSet[1] database at NCBI. The results of the following analysis are summarized in Table 1. The first question that we ask is "How much additional generalization is provided to the set of released DNA sequences resulting from DNALA?" We attempt to answer this by measuring the increase in how much generalization an anonymized sequence has with respect to the original sequence, in terms of the generalization lattice. For each sequence we sum the generalization hierarchy level score for each nucleotide. The increase in generalization seems to be dependent on the number of SNVRs available for anonymization. All of the sample sets had an average increase in generalization of less than 1. However, this number is a bit misleading, since the average number of regions used for generalization ranged from 3 to 4 SNVRs, with similar size standard deviations. This paradox is due to the fact that in each of the samples there are a certain number of sequences that require no generalization. This is expected, since certain sequences are more common in general populations than others. If we exclude the population that does need generalization, then the increase in total generalization level per sequence increases almost three to five fold, depending on the sample set. In addition, the generalization was analyzed with respect to the number of gaps added to a sequence from the MSA step. The number of gaps generalized for each sequence in all sets, is 0.5 or less, which suggests that multiple sequence alignments constructed from the released anonymous DNA sequences will be similar to those resulting from the anonymized sequences before the gaps are removed.

The second question that we ask is, "How similar are the sequences after anonymization?" In other words, how much does the total distance between all sequences decrease by generalization? We answer this by determining the change in average distance from each sequence to all other sequences. We find that the average distance between sequences increases with respect to the number of SNVRs available for generalization. This feature is most probably due to the fact that the anonymization process is converting the original sequences into clusters of size 2 and using the centroid of each cluster for the anonymized sequences. In such an event, the total distance will increase on average.

5 Discussion

The DNALA system provides a method of generating 2-anonymous DNA sequences from a given set of sequences. However there is still much evaluation necessary. First, the amount of generalization incurred in a set of sequences is dependent on the number of sequences available for population, as well as, the amount of variation between the sequences. However, when the number of sequences is low, such as 4, and the number of $SNVR$ regions is high, such as 250, then the released sequences may be overgeneralized

[1] Additional information about PopSet can be found at the NCBI website: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PopSet

in comparison to what is necessary for privacy protection. In such an event, it might be more useful to generalize sequences with pseudosequences constructed from known distributions of the population for $SNVRs$. Since the goal of generalization is to prevent the tracking of a sequence from one institution to the next, it is not necessary that a sequence must be the same as other sequences released from an institution. It is simply sufficient to make sure that the same sequence released from two institutions can not be uniquely matched. However, to afford such protection without anonymizing a sequence with another sequence from the same institution is currently an unsolved problem. More work is necessary in the area.

A second issue with this method is that when $k > 2$. the distance relationships are no longer metric. For instance consider three simple sequences for with one $SNVR$: seq_1=A, seq_2=C, and seq_3=T. The distance between sequence $dist(seq_1, seq_2) = 1$, $dist(seq_1, seq_3) = 1$, and $dist(seq_2, seq_3) = 1$. Each of the sequences claims that they are only one away from the other and thus are ideal for generalizing together. However, the generalization would push all sequences up two levels, not the expected one! This is due to the fact that while they differ by one, they differ at different subgraphs of the generalization lattice. Recent work in the field of data privacy has considered how to generalize with respect to orthogonal distances, similar to the type mentioned. As such, a second direction for this research would be to explore methods of anonymization with respect to non-scalar distances. However, there may exist ways that the DNALA method might be expanded to account for the higher levels of k-anonymity. Yet, such a method will be an approximation of the optimal generalization strategy and will scale exponentially in k.

An additional aspect to ponder is the usefulness of k-anonymized DNA data. How does one quantify the information loss in DNA, when it is not currently known what all of the applications will be? This is why information loss is characterized above as the amount of generalization induced and how distance between the released sequences compared to the original sequences. Yet, information loss is dependent on the use of the data. The single nucleotide generalization lattice model assumes that variation occurs in small sized regions (i.e. 1-3 residues). However, there are mutations that do lend themselves to such a model. For example, consider the growing number of trinucleotide repeat mutations, which are common for diseases such as Fragile-X, Huntington's disease, and myotonic dystonia. Such diseases have known phenotypes which can infer the repeat size, as has been demonstrated with Huntington's disease. [18] In such an event a single nucleotide generalization model will grossly overgeneralize the data and incur unnecessary loss in the semantic aspects of the sequences. Thus, an additional direction for future research would be to study the effect that anonymized DNA data has on different applications, such as evolutionary tree construction or pharmacogenomic correlations.

6 Conclusion

This research introduces a novel computational method for protecting the privacy of identities to which DNA sequences were derived from. The technique expands computational disclosure control theory for domain generalization hierarchies to generalization lattices. In the current model, DNA privacy is protected by generalizing pairs of sequences to a common sequence. Based on real world data, the anonymization schema appears feasible for anonymization of sequences from a relatively small database of 30 sequences to a larger database of approximately 400 sequences. Despite the fact that this technique learns which DNA sequences should be anonymized to a single sequence, and thus prevents the explicit identity of DNA sequences from being inferred, future research is still necessary to determine how such a privacy protection schema affects the ability to learn useful knowledge or data mine the sequences.

Acknowledgements

The author is indebted to the helpful discussions and insightful suggestions provided by Dannie Durand and Latanya Sweeney. This research was funded in part by the Data Privacy Laboatory at Carnegie Mellon University. This paper originally appeared as a Working Paper of Data Privacy Laboratory.

References

[1] Beroud C, Collod-Beroud G, Boileau C, Soussi T, and Junien C. UMD (Universal mutation database); a generic software to build and analyze locus-specific databases. *Human Mutation*. 2000; 15(1): 86-94.

[2] Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas S, and Cooper DN. Human gene mutation database - a biomedical information and research resource. *Human Mutation*. 2000; 15(1): 45-51.

[3] Altman RB. Bioinformatics in support of molecular medicine. In *Proceedings of the American Medical Informatics Association Annual Symposium*. Orlando, FL. 1998; pp. 53-61.

[4] Mendelsohn ML, Peeters JP, and Normandy MJ, editors. *Biomarkers and Occupational Health - Progress and Perspectives*. Washington, DC. Joseph Henry Press. 1995.

[5] Shulte PA and Perera FP, editors. *Molecular Epidemiology: Principles and Practices*. New York. Academic Press. 1993.

[6] Rothstein MA, editor. *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era*. New Haven. Yale University Press. 1997.

[7] Altman RB and Klein TE. Challenges for biomedical informatics and pharmacogenomics. *Annu Rev Pharmacol Toxicol*. 2002; 42: 113-33.

[8] Malin BA and Sweeney LA. Determining the identifiability of DNA database entries. In *Proceedings of the American Medical Informatics Association Annual Symposium*. Los Angeles, CA. 2000; 547-551.

[9] Malin BA and Sweeney LA. Re-Identification of DNA through an automated linkage process. In *Proceedings of the American Medical Informatics Association Annual Symposium*. Washington, DC. 2001; 423-7.

[10] Malin BA. Compromising privacy with trail re-identification: The REIDIT algorithms. Center for Automated Learning and Discovery Technical Report CMU-CALD-02-108, School of Computer Science, Carnegie Mellon University. Pittsburgh, PA: December 2002.

[11] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness, and Knowledge-based Systems*. 2002; 10 (7): 571-588.

[12] Lin Z, Hewett M, and Altman RB. Using binning to maintain confidentiality of medical data. In *Proceedings of the American Medical Informatics Association Annual Symposium*. San Antonio, TX. 2002; 454-458.

[13] Liebecq C, editor. *Biochemical Nomenclature: And Related Documents: A Compendium (International Union of Biochemistry and Molecular Biology), 2nd ed.* Chapel Hill, NC. Portland Press. 1992.

[14] Higgins DG, Thompson JD, and Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. 2000; 22: 4673-4680.

[15] Makova KD, Ramsay M, Jenkins T, and Li WH. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. *Genetics*. 2001; 158: 1253-1268.

[16] Harris EE and Hey J. X chromosome evidence for ancient human histories. *PNAS, USA*. 2000; 96: 3320-3324.

[17] Yao YG, Nie L, Harpending H, Fu YX, Yuan ZG, and Zhang YP. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am J Phys Anthropol*. 2002; 118: 63-76.

[18] Malin BA and Sweeney LA. Inferring genotype from phenotype through a knowledge-based algorithm. In *Pacific Symposium on Biocomputing*. 2002; 41-52.