

Query Protection in DNA Homology Search

Shogo Shimizu¹

shimizu-syogo@aait.ac.jp

Yeondae Kwon²

yekwon@lab.nig.ac.jp

¹ Advanced Institute of Industrial Technology, 1-10-40 Higashi-ooi, Shinagawa-ku, Tokyo 140-001, Japan

² Laboratory for Research and Development of Biological Databases, CIB-DDBJ Center, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Abstract

When DNA or protein databases are administrated outside queriers' organizations, queriers often do not want their query sequences to be exposed to database administrators. In this paper, we propose a secure homology search method without using encryption keys. Our method is based on q -gram and adopts a fuzzy matching method, called fuzzy vault, as a data encryption mechanism.

Keywords: query protection, homology search, database, q -gram, fuzzy vault

1 Introduction

There are a number of DNA or protein sequence databases available to the public. Biologists execute homology search against databases with their own query sequences and receive similar sequences to the queries as results. When databases are administrated outside queriers' organizations, queriers often do not want their query sequences to be exposed to database administrators. Thus, it is desirable that query sequences be protected from database administrators, while preserving the functionality of efficient homology search.

q -gram is a well-known method for efficiently filtering candidates of similar sequences in databases, which is based on the minimum number of common substrings that a query and a similar sequence should have [1]. One solution is to encrypt sequences in databases using encryption keys and match hashed (thus, irreversible) values of substrings. However, secure management of encryption keys becomes more cumbersome as the number of database users increases.

In this paper, we propose a secure homology search method without using encryption keys. Our method is based on q -gram and adopts a fuzzy matching method, called fuzzy vault [2], as a data encryption mechanism. A query sequence is protected unless its corresponding vault is unlocked.

2 Method

2.1 q -gram Filtering

Sequence similarity is measured by an edit distance. An edit distance of sequences s_1 and s_2 is defined as the total cost of edit operations such as insertion, deletion, and permutation needed to make s_1 and s_2 identical. Here we assume that every edit cost is 1 for simplicity. It is shown that if an edit distance of s_1 and s_2 of length n is d , then they have $n - (d + 1)q + 1$ common substrings of length q . Based on this lemma, q -gram can first select candidates for similar sequences efficiently, and then obtain final results by refining candidates according to the definition.

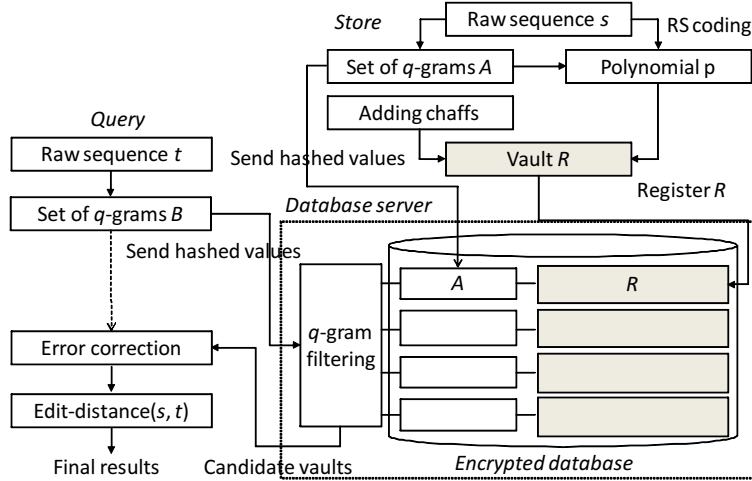


Figure 1: A procedure for enrollment and a query.

2.2 Encrypted Matching Using Fuzzy Vault

A fuzzy vault is a form of error-tolerant encryption operation where keys consist of sets. In a fuzzy vault scheme, a secret s is locked in a vault using a set A . If one tries to unlock the vault using another set B , unlock succeeds, therefore s is obtained only if A and B overlap substantially. In our method, s corresponds to a sequence in a database, and A and B correspond to the sets of q -grams generated from s and a user's query, respectively.

Given a sequence t and an integer d , a homology search method returns a set of sequences that contains all the sequences within edit distance d to t . Figure 1 shows the outline of our algorithm. When storing a sequence s in a database, construct a polynomial p of degree k from s using Reed-Solomon coding. Then construct a vault R that consists of right points generated from q -grams of s and p , and pseudo points, called chaffs. Chaffs are added to make the reconstruction of p from R impossible. A pair of R and a set of hashed values of q -grams of s , denoted as A , is registered in the database. When querying the database, a set of hashed values of q -grams of t , denoted as B , is sent to the server and matched against each A in the database. If the matched number of q -grams between A and B exceeds $n - (d + 1)q + 1$, then include R which corresponds to A as a candidate. The unlocking condition for a vault is determined so that s can be decoded from R by applying an error correction algorithm only if the edit distance of s and t is less than or equal to d .

3 Security Consideration

Even a database administrator cannot identify a query sequence unless he/she decodes R successfully. However, there is a risk of statistical analysis to the vaults in the database. For example, if an attacker knows frequent patterns in sequences, he/she can guess mappings from q -grams to x -coordinates using distributions of the points in the vaults. More detailed analysis is the future work.

References

- [1] Cao, X., Li, S., and Tung, A., Indexing DNA sequences using q -grams, *Proc. 10th International Conference on Database System for Advanced Applications (DASFAA 2005)*, 4–16, 2005.
- [2] Juels, A. and Sudan, M., A fuzzy vault scheme, *Des. Codes Cryptography*, 38(2):237–257, 2006.