Global Integration of Gene Databases by Eigen-ID Method

Tetsuro Toyoda **Akiko Itai**
toyoda@immd.co.jp itai@immd.co.jp
Institute of Medicinal Molecular Design
Kadokawa Hongo Bld. 4F, 5-24-5 Hongo, Bunkyo-ku, Tokyo, Japan

1 Introduction

Recently the amount of nucleic-acid and amino-acid sequences in databases is tremendously expanding. Databases which deal with genetic information are spread out in the world, and those data are accumulated independently in each database. Hereafter, considering that huge amount of sequence information will be registered in independent database sites, it is necessary to develop an efficient method of maintaining the relationships among the biological information registered in each database and a system of interoperable databases is required [1]. Each database uses different identifiers (IDs) such as ID of some letters used in SWISS-PROT or accession ID in GenBank. Since there are many cases in which IDs irrelevant to the sequences are assigned to them automatically or arbitrarily, the probability of assigning different IDs to the same sequence or the same ID to the different sequences seem considerable. It is often the case that same sequences are registered by different IDs in different databases. Up to this time, the correspondence between the data and ID is set by each database organization and is guaranteed by the dedicated maintenance of the database.

2 ID from Sequence itself

As one of the most effective solutions, we propose a new method, that we call "eigen-ID method [2]," which generates unique ID for the sequence by a one-to-one function from the data on nucleic-acid or amino-acid sequence. By using the same transfer function, it enables everyone, whenever and wherever, to assign one specific ID to one specific sequence and spontaneously to establish the consistency and correlation among the IDs in databases of the world. For this purpose, the formula of transfer function must be simple enough to be computationally implemented and fully utilized by many programmers. As an example, we present a simple function that transfers the sequence data to 32-letter IDs (eigen-IDs) as shown in Fig. 1. It employs a collision intractable hash function 'SHA' which is often applied to cryptography in the financial service industry [3]. SHA transfers a text on sequence to 160-bit binary data, which is then broken down to 32 blocks of 5-bit binary. Each block is denoted by one of 32 letters such as '0' to '9' and 'a' to 'v.' SHA very seldom gives the same binary data to different sequences, thus it can be considered as a one-to-one function from a practical point of view. We demonstrate here the utility of the eigen-ID method by reconstructing two databases containing sequences from SWISS-PROT and PDB using the eigen-IDs as primary keys of the records. In each reconstructed database it was found that the same IDs were given to the same sequences and different IDs were given to different sequences, which ensured the consistency and correlation between the databases

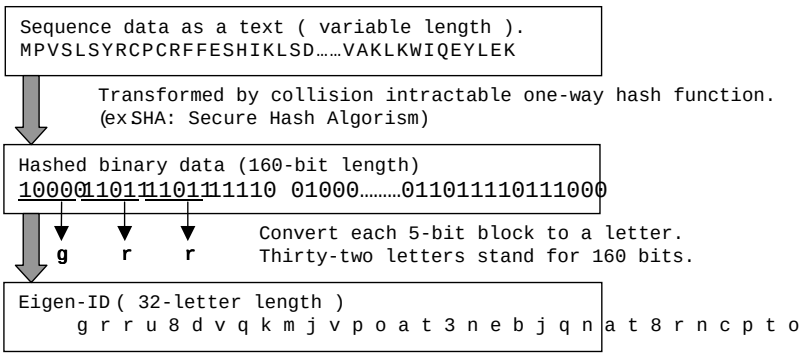

Figure 1: Scheme of creating eigen-ID from sequence data.

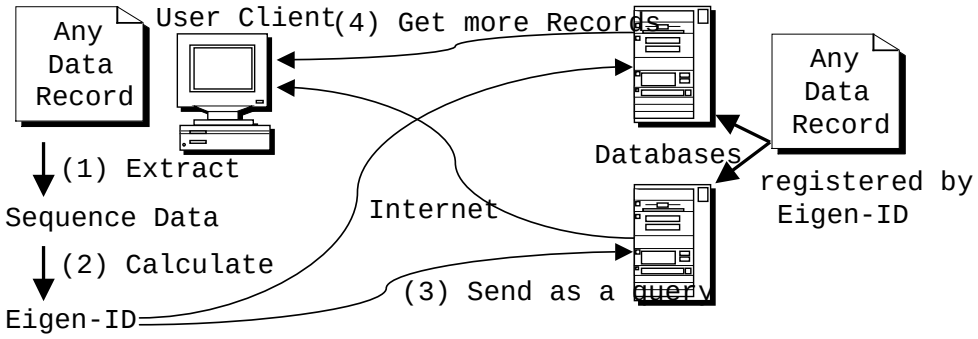

Figure 2: More efficient database environment globally united by eigen-ID.

3 Globally consistent IDs

If this sort of ID symbols are systematically assigned to the sequence databases throughout the world, more efficient database environment will be provided to the scientists. As shown in Fig. 2, eigen-ID method will give a great deal of utility to scientists who would be able to obtain the information relevant to a certain sequence from world-wide databases by sending the calculated eigen-ID as a query through the web. Eigen-ID system upholds the rule that "the same data must have the same ID" among independent databases. Since the method dissolves the labor of considering the correspondence and spontaneously grants the consistency among individual databases, it becomes easier for individuals to construct globally-consistent personal database and to release it on the world wide web.

References

[1] Malakoff, D., NIH urged to fund centers to merge computing and biology, *Science*, 284:1742, 1999.

[2] Toyoda, T. and Itai, A. Japanese Patent 11-227438.

[3] National Institute of Standards and Technology. Public Key cryptography using irreversible algorithms for the financial services industry, Part.2: The secure hash algorithm, X9.39-199x, 1991.