Abstract

The problem of data compression having specific security properties in order to guarantee user's privacy is a living matter. On the other hand, high-throughput systems in genomics (e.g. the so-called Next Generation Sequencers) generate massive amounts of genetic data at affordable costs.

As a consequence, huge DBMSs integrating many types of genomic information, clinical data and other (personal, environmental, historical, etc.) information types are on the way. This will allow for an unprecedented capability of doing large-scale, comprehensive and in-depth analysis of human beings and diseases; however, it will also constitute a formidable threat to user's privacy.

Whilst the confidential storage of clinical data can be done with well-known methods in the field of relational databases, it is not the same for genomic data; so the main goal of my research work was the design of new compressed indexing schemas for the management of genomic data with confidentiality protection.

For the effective processing of a huge amount of such data, a key point will be the possibility of doing high speed search operations in secondary storage, directly operating on the data in compressed and encrypted form; therefore, I spent a big effort to obtain algorithms and data structures enabling pattern search operations on compressed and encrypted data in secondary storage, so that there is no need to preload data in main memory before starting that operations.