

DNA-Miner: A System Prototype for Mining DNA Sequences*

Jiawei Han[†] Hasan Jamil[§] Ying Lu[†] Liangyou Chen[§] Yaqin Liao[†] Jian Pei[†]

[†] School of Computing Science, Simon Fraser University, Canada.

[§] Department of Computer Science, Mississippi State University, USA

With significant developments of computational biology and bioinformatics, the discovery of interesting patterns in biosequences, including DNA, RNA, and protein sequences, has become an important task in research and applications. An important goal of mining biosequences is to find sequence or repeating patterns hidden in DNA, or other bio-data in large databases. Previous studies have been using various techniques, including statistical analysis, machine learning, minimum description length (MDL) principle, etc. with fruitful results. However, most of these methods can only identify a proper subset of patterns that meet the specifications provided explicitly by users.

Recent studies in data mining have developed efficient and effective methods for mining sequential patterns in large databases. A distinct feature of these methods is that, given a specification of the patterns interested, the *complete* set of patterns satisfying the specification, instead of only a subset of them, can be found from the database. However, with the sophistication of bio-medical data, bio-sequence mining poses challenges to the current database-oriented sequential pattern mining methods in two aspects: (1) biosequences usually contain many long sequence patterns, which are rather different from shopping transactions where short patterns are dominant, and (2) bio-sequence patterns usually contain mutations, insertions and deletions, whereas business applications often ignore such cases or treat them as noises in studies.

Can we narrow down the gap between the two fields and extend sequential pattern mining methods to biosequences?

In this demo, a research system prototype, called **DNA-Miner**, is presented, which takes DNA data as input and performs sequential pattern mining for DNA analysis. Its system architecture consists of three components: **integrated DNA database**, **DNA mining module**, and **user interface**.

* The work was supported in part by the Natural Sciences and Engineering Research Council of Canada (grant NSERC-A3723), and the Networks of Centres of Excellence of Canada (grant NCE/IRIS-3).

The first component, **DNA database**, stores a large amount of DNA data for analysis. Although such data can be downloaded from the Web, integration of them from heterogeneous sources and preprocessing of the integrated data for data mining are challenging tasks. The integration of DNA data and construction of integrated DNA database will be part of the contents in this demonstration.

The second and also the key component of **DNA-Miner** is the **DNA mining module**, which contains the following three functions.

- **Mining repeating patterns.** Taking a DNA and the specification of repeating patterns as input, this function finds the complete set of (partial) repeating patterns in the DNA according to the specification.
- **Mining potential motifs in a DNA database.** Taking a set of DNA (in the form of a DNA database) and the specification of potential motifs (which provides the support threshold and features of potential motifs) as input, this function returns the complete set of patterns in the DNA database satisfying the specification.
- **Classification based on potential motifs.** After mining motifs in a DNA database, this function takes the mined motifs and the labeled DNA sequences to build classifier, summarizing the major and distinct features of each class, which can be used to test and classify new motifs, with certain accuracy.

The third component is the **user-interface** of **DNA-Miner**, which puts the system in a user-friendly, mining-query based, interactive environment. A graphical user interface will lead users to perform mining and watching the mining results. In particular, a user can obtain data from the Web (if an Internet connection is made available), submit their DNA data, and/or execute mining queries on-line. A mining query can be one of the above three types: *mining repeating patterns*, *mining potential motifs*, and *classification based on the mined potential motifs*.

In addition, we will compare the **DNA-Miner** system with some other DNA analysis systems by running both systems in parallel using the exact same DNA sequences as inputs and examine their outputs and performance. Statistics about such comparisons based on our system running history will also be provided.