

**Pattern-Matching Algorithms for Application in Bioinformatics**

Thesis proposal

for the degree of Master of Science in Computer Science

at Southern Connecticut State University

Julia Fainstein

December 2002

Thesis Advisor: Dr.Taraneh Seyed

**Major-Field Approval** – The advisor and the department chairperson

---

Advisor

---

Date

---

Chairperson

---

Date

### A. Title

#### Pattern-Matching Algorithms for Application in Bioinformatics

### B. Statement of Purpose

Search is one of the most important problems in computer science, present in almost all application for data and text processing. Although data are stored in various ways, text remains the main form of information exchange and applies well to computer science where large amounts of data are stored in linear files of letters. Such strings may correspond to any type of information. They are sequences of characters representing texts, sounds, images, DNA strains or strings of nucleotides or amino acids.

Pattern-Matching, or locating a given stream of information within a large pool of existing data, is therefore the most important subject in the domain of information processing. Pattern-Matching algorithms are basic components of software under most operating systems. They are fundamental to querying and information retrieval in many branches of science, among which are dictionary and text searches, development of computer anti-virus applications, sound retrieval and processing, image processing, molecular biology and genetics research (Crochemore, Lecroq, 2003, and Gusfield, 1997).

The goal of this project is to develop and implement a user-friendly, general-purpose platform for Pattern-Matching applications. It will focus on the application of Pattern-Matching algorithms in bioinformatics, although the platform and the programs developed in the context of this endeavor may be used in any other area of science directly related to Pattern-Matching. The platform will allow the user to examine the properties of genome and to compare various samples already stored in databases, in support of research and development in the area of

Bioinformatics. In order to utilize the Pattern-Matching algorithms for specific problems in the field of bioinformatics, adaptation or alteration of existing algorithms and methods is required to conform to requirements of the field.

This system will support Pattern-Matching algorithms and utilities required for information retrieval and processing. The platform should also support utilities for data analysis and performance measurement. Input data of various nature (including biological, textual, or pictorial) will be pre-processed into a uniform format on which the algorithms will operate.

### *C. Literature Review And Current State-Of-The-Art*

Strmat is an ongoing project published as source code on the world wide web. Strmat is a collection of C programs intended for the Unix environment, tied together with a menu system that implement a variety of string matching and pattern discovery algorithms. This project was initiated by Dan Gusfield (Strmat, 2000) at University of California, Davis, with support from Department of Energy and National Science Foundation. This collection does not run under Windows operating system without modifications and lacks support utilities and graphical front end.

Among the specialized tools are field-specific applications such as the recently developed “Matchsimile” (Navarro, Baeza-Yates, Arcoverde, 2002), which performs inexact matching (matching with allowance for errors, symbol insertions, deletions, substitutions, and transpositions) of short strings within long text. “Matchsimile” is a system to deal with multi-word expressions that can handle words formation rules using a mathematical model. It is utilized for word searches within large texts, where text is viewed as a sequence of words (as opposed to very long chains of characters) more suitable for dictionary-like searches.

Another recently developed field-specific tool is SEMEX (Search Engine for Melodic Excerpts), a prototype exploring and implementing string matching techniques for music retrieval. (Lemstrom, 2000). SEMEX is not in public domain at this time. It performs queries on pitch sequences, where music is viewed as a string of pitches (or pitch intervals). The existing version locates transposition invariant (sequence of notes may start from different notes) matches of monophonic (single melody) query patterns in music databases.

In addition, there must be a class of applications for large-scale biological and genetic querying which are employed by research laboratories on supercomputers requiring immense computational resources. These applications were used to sequence human genome but none is in public domain.

#### *D. Research Plan*

The proposed project will combine complex Pattern-Matching algorithms with user friendly front end on a readily-accessible PC platform. Combining algorithms into a unified platform will allow researchers in varying disciplines to access, process, and evaluate different algorithms in order to select the most suitable one for their field. The complete package will be used as a research tool in pattern recognition and Bioinformatics applications. The system will utilize existing technologies and generally available computing platforms commonly accessible to individual PC owner/user, offering a research platform to individuals or groups without large financial resources where resources of similar functionality may be overly costly, as well as to users without extensive computer knowledge.

#### *E. Methodology*

The categories of algorithms that will be represented in the proposed platform are exact match and inexact match, and may be applied to any source of information. A third subdivision

will represent two-dimensional Pattern-Matching algorithms. The majority of these algorithms are found in their pseudo code form in a pioneering work by Dan Gusfield (1997). The following algorithms will be implemented in this project:

- 1) Exact Match (searching for an exact copy of the Pattern (Crochemore, Lecroq, 2003)
  - a) Simple (Naïve) Method
  - b) Knuth Morris Pratt (KMP) Algorithm
  - c) Boyer Moore Algorithm
  - d) Semi-numerical String-Matching (Shift-AND method)
  - e) Dynamic Programming Algorithms
  - f) Suffix-Tree Methods with emphasis on applications suited for molecular biology, such as longest common substring and palindromes (McCreight, 1976)
- 2) Inexact Match (searching for Pattern in Text allowing for some degree of error or variations in the Pattern or in the Text (Baase, Van Gelder, 1999 )
  - a) K-mismatch
  - b) Shift-OR method
- 3) Two-dimensional Pattern-Matching algorithms well suited to image processing, such as detecting features in a picture or a bitmap (Crochemore, Lecroq, 2003)
  - a) Zhu-Takaoka (1989)
- 4) Matching algorithms suited for such areas as musical pattern search (Lemström, 2000)

A Windows-based graphical platform will be developed to allow users to apply the supported algorithms to their specific set of data, compare their performance and select the one that best suits their application. The user interface and algorithms described in this project will

be implemented using Microsoft Visual C++, and Visual Basic programming languages with browser access to the Web. The program will be developed as libraries which can also be used independently. The following utilities will be supported by the program to facilitate data acquisition and analysis:

- 1) Text Generator: To generate sample test data or to download it
- 2) Text Converter: To convert the file to a unified format
- 3) Pattern Locator: To locate a pattern among one or more texts using a selected algorithm from among algorithms described
- 4) Visualizer: To display the position of pattern within the text
- 5) Performance Analyzer: To measure and compare the performance of selected algorithms in terms of actual and theoretical running time

To implement and test the algorithms, sample test data will be collected, including text files, music information, picture and bitmap images, and protein/molecular data. Once the algorithms are implemented, the platform will be tested against sample data. Accuracy of results, runtime comparison, memory and disk space requirements will be compared and reported on different types of data. Results will be represented in a graphical viewer.

Documentation of the project will comprise a users' guide and detailed description of the different modules of the application. Complete listing of the program source code will be provided. The comparative performance of algorithms will be reported against different data sets. A detailed user's guide will consist of an overview of the tools' functionality, a tutorial, and system requirements. The program will be made available at the conclusion of the project.

*References*

- Baase, S., Van Gelder, A. (1999). *Computer Algorithms: Introduction to Design and Analysis*. (3rd ed.) Addison-Wesley.
- Crochemore, M., Lecroq, T. (2003). Pattern-Matching and Text Compression Algorithms. Chapter 8 from *The Computer Science and Engineering Handbook*, A.B. Tucker, Jr, ed., CRC Press, Boca Raton, 2003. (To appear). Section retrieved from the website of Department of Computer Science, King's College London: <http://www.dcs.kcl.ac.uk/teaching/units/csmtsp/B5.html>
- Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences*. (1st ed.) Cambridge University Press.
- Lemström, K.(2000). String Matching Techniques for Music Retrieval. *PhD Dissertation*. University of Helsinki, Finland, Department of Computer Science. (Advisor: Professor Esko Ukkonen). Department of Computer Science, Series of Publications A, Report A-2000-4. ISBN: 951-45-9573-4: <http://www.cs.helsinki.fi/u/klemstro/THESIS/#en>
- McCreight, E.(1976) A Space-Economical Suffix Tree Construction Algorithm. *Journal of the Association for Computing Machinery*, Vol. 23, No. 2, April 1976, pp. 262-272.
- Navarro, G., Baez-Yates, R., Arcoverde, J. (2002) Matchsimile: A Flexible Approximate Matching Tool for Personal Names Searching. Article Retrieved November, 2002 from: <http://citeseer.nj.nec.com/456711.html> and <http://www.matchsimile.com.br/>
- Strmat (2000). *Project initiated by Gusfield, D. at the University of California, Davis, with support from Department of Energy and National Science Foundation*. Source retrieved December, 2002 from <http://www.cs.ucdavis.edu/~gusfield/strmat.html>
- Zhu, F., Takaoka, T. (1989). A technique for two-dimensional Pattern-Matching. *Computing Practices, Communications of the ACM*, September 1989 Volume 32 No. 9.