# Application of Exact String Matching Algorithms towards SMILES Representation of Chemical Structure

Ahmad Fadel Klaib, Zurinahni Zainol, Nurul Hashimah Ahamed, Rosma Ahmad, and Wahidah Hussin

***Abstract*—**Bioinformatics and Cheminformatics use computer as disciplines providing tools for acquisition, storage, processing, analysis, integrate data and for the development of potential applications of biological and chemical data. A chemical database is one of the databases that exclusively designed to store chemical information. NMRShiftDB is one of the main databases that used to represent the chemical structures in 2D or 3D structures. SMILES format is one of many ways to write a chemical structure in a linear format. In this study we extracted Antimicrobial Structures in SMILES format from NMRShiftDB and stored it in our Local Data Warehouse with its corresponding information. Additionally, we developed a searching tool that would response to user's query using the JME Editor tool that allows user to draw or edit molecules and converts the drawn structure into SMILES format. We applied Quick Search algorithm to search for Antimicrobial Structures in our Local Data Ware House.

***Keywords*—**Exact String-matching Algorithms; NMRShiftDB; SMILES Format; Antimicrobial Structures.

## I. Introduction

BIOINFORMATICS has been regarded widely as a key discipline for the scientific development and it has a big amount of unlimited increasing and distributed information [1]. Bioinformatics has several categories included computational biology, computer science relating to biology, management of bioinformatics infrastructure and so forth [16]. Cheminformatics is the applying of computer and informational techniques with problems in the field of chemistry for the intended purpose of making better and fast decisions in the area of drug and chemical compounds [10]. Computational biology and chemistry that use computational methods handle large amount of data, so biologists can pinpoint genes, DNA or proteins and chemists can pinpoint chemical structures.

Manuscript received June 30, 2007. This work was supported by the Government of Malaysia - Ministry of Science, Technology and Innovation (MOSTI) under E-Science Grant no 01-01-05-5F0182.

Ahmad Fadel Klaib is Master Student at the School Computer Sciences, USM, Malaysia (e-mail: ahmadklaib@yahoo.com).

Zurinahni Zainol and Wahidah Hussin are Senior Lecturer at the School Computer Sciences, USM, Malaysia (e-mail: zuri@cs.usm.my, wahidah@cs.usm.my).

Nurul Hashimah Ahamed is Research Officer at the School Computer Sciences, USM, Malaysia (e-mail: nurul@cs.usm.my).

Rosma Ahmad is Senior Lecturer at the School Industrial Technology, USM, Malaysia (e-mail: rosmah@usm.my).

NMRShiftDB is a web database for organic structures and their nuclear magnetic resonance spectra [4, 13]. It is an open-source, open-access application and allows for spectrum prediction as well as for searching spectra, structures and other properties.

Chemical structures are represented using lines representing chemical bonds between atoms and shaped on 2D structural formulae. SMILES format is one of many ways to write a chemical structure in a linear format [18]. Linear formulas have an advantage over graphical structures because they are easily read through computer code and computational use, especially for search and storage [19]. SMILES format contains the same information as might be found in an extended connection table. The primary reason SMILES is more useful than a connection table is that it is a linguistic construct, rather than a computer data structure.

There are popular kinds of existing structures called Antimicrobial Structures [11] which have been shown to be effective in experimental infections with multi drugs as a substance that kills or inhibits the growth of microbes such as bacteria (antibacterial activity), fungi (antifungal activity) and viruses (antiviral activity) [20]. Thus our research aimed to help biologists and chemists by extracting Antimicrobial Structures from NMRShiftDB and allowing them to search for them in our Local Data Warehouse using Quick Search algorithm.

## II. Related Work

String-matching is a very significant subject in the wide domain of text processing. Nowadays, this problem received an enormous deal of attention due to various applications in computational biology. String matching algorithms play a key role in most of computer science problems, challenges and in implementation of computer software.

String-matching algorithms work as follows [17]: They compare the text with size $n$ with the pattern which size is equal to $m$. They first put the left ends of the pattern and the text, then compare text characters with pattern characters and after a mismatch among the comparison between the pattern and the text or a whole match between them they shift the pattern to the right and the same procedure is repeated until the the pattern reach to the right end of the text.

Our research included a survey of seven Exact String-matching algorithms, which they are Boyer-Moore algorithm

[5, 7, 14, 21, 24], Horspool algorithm [8, 22], Brute force algorithm [9, 23], Knuth-Morris-Pratt algorithm [2, 3, 6], Quick-Search algorithm [3, 12], Karp-Rabin algorithm [14] and Zhu-Takaoka algorithm [12]. Fig. 1 summarizes and compares these algorithms:

| Algorithm Name | Author | Year | Comparison Order | Preprocessing Phase | Preprocessing Time Complexity | Searching Time Complexity | Main Characteristics |
|---|---|---|---|---|---|---|---|
| Boyer-Moore algorithm | R. S. Boyer and J. S. Moore | 1977 | From right to left | yes | O(m+σ) | O(mn) | Uses both good-suffix shift and bad-character shift. Is not very efficient for small alphabets |
| Horspool algorithm | Nigel Horspool | 1980 | Is not relevant | yes | O(m+σ) | O(mn) | Uses only the bad-character shift with the rightmost character. It's more fast and easy to implement comparing to Boyer-Moore algorithm. |
| Brute force algorithm | - | Very Old | Is not relevant | No | No preprocessing time complexity | O(mn) | Only shift one by one character. It is not an optimal algorithm because it's not use the information that could be gained from the last comparison |
| Knuth-Morris-Pratt algorithm | Michael O. Rabin and Richard M. Karp | 1974 | From left to right | yes | O(m) | O(n+m) independent from the alphabet size | Uses the notion of the border of a string, so that increases the performance, decreases the delay, and decreases time of searching comparing with Brute force algorithm |
| Quick-Search algorithm | Sunday | 1990 | Is not relevant | yes | O(m+σ) | O(mn) | Uses only the bad-character shift. Is very fast especially for short pattern |
| Karp-Rabin algorithm | Michael O. Rabin and Richard M. Karp | 1984 | From left to right | yes | O(m) | O(mn) | Using hashing function. Is very effective for multiple pattern matching in one-dimensional string matching |
| Zhu-Takaoka algorithm | R. F. Zhu and T. Takaoka | 1989 | From right to left | yes | O(m+σ2) | O(mn) | Using hashing function. is very effective for multiple pattern matching in two-dimensional string matching |

Fig. 1 Summary of the algorithms that studied in this research

From the above figure, we concluded that Quick Search algorithm is very fast especially for short patterns, but for long patterns it is less efficient than other algorithms. So in this research we applied Quick Search algorithm to our domain due to the majority length of Antimicrobial Structures in our Local Data Warehouse are short. Noted that the maximum length of Antimicrobial Structures in our Local Data Warehouse is consists of 128 characters such as *"O=C1CCC8(C)(C(C1)C(OC3OC(C)C(O)C(OC2OC(C)C(O)C(O)C2(O))C3(O))CC7C8(CCC6(C)(C7(CC5OC(O)(CCC(C)COC4OC(CO)C(O)C(O)C4(O))C(C)C56))))"*

## III. Methodology

The general framework is divided into four phases as denoted by numbers 1, 2, 3 and 4 as shown in Fig. 2. The first phase is to mine Antimicrobial Structures in SMILES format from the NMRShiftDB. The second phase is the development of Local Data Warehouse to host the Antimicrobial Structures data that gained from the first phase. The third phase is the development of searching tool that would response to user's query. The final phase is to check weather the Antimicrobial Structures query is in our local data warehouse or not, by using Quick Search algorithm.

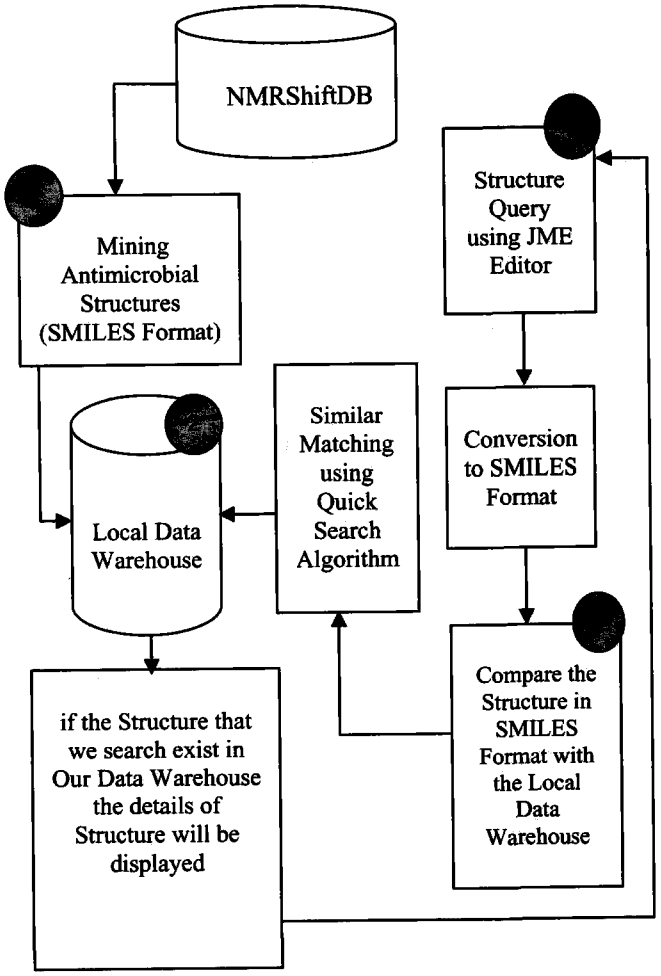


Fig. 2 Research Methodology

### A. First Phase: Mining Antimicrobial Structures

We extracted Antimicrobial Structures in SMILES format from NMRShiftDB which enabled us to perform 2D structure matching and comparison. The extraction process of Antimicrobial Structures is done via search by Keyword/Category such as "Antimicrobial", "Antibacterial", "Antifungal", "Antiviral" and etc.

### B. Second Phase: Development of Local Data Warehouse

After identifying the Antimicrobial Structures in the previous phase, the results stored in our Local Data

Warehouse. The Local Data Warehouse consists of 78 relational tables which denoted to the Antimicrobial molecules and its corresponding information.

### C. Third Phase: Development of Searching Tool

This phase included the development of searching tool that would response to user's query. User queries are captured using the JME Editor [15] which allows user to draw or edit molecules. Then it converted the drawn structure into SMILES format. Fig. 3 shows an example for user query using JME Editor Tool and Fig. 4 shows the conversion operation of the drawn structure into SMILES format:

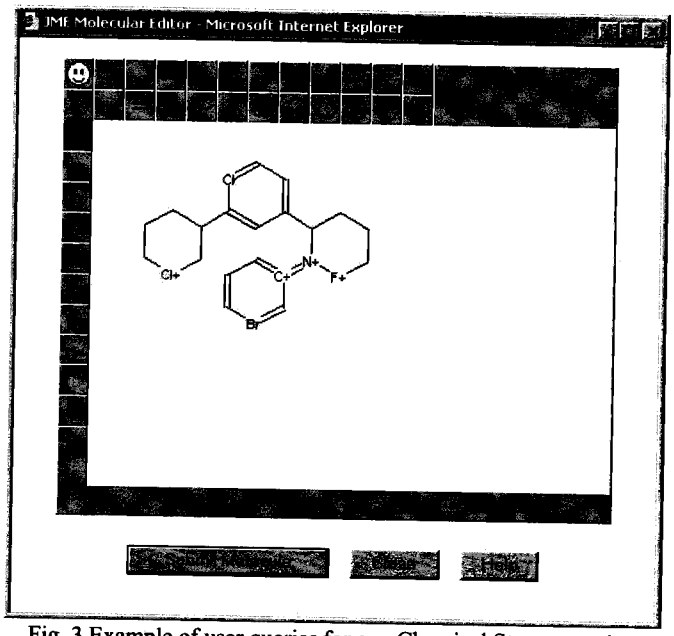


Fig. 3 Example of user queries for any Chemical Structure using JME [15]

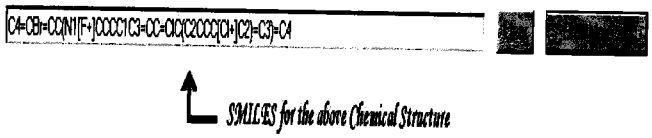


Fig. 4 Converts the above drawn structure into SMILES format

### D. Fourth Phase: Check the Foundation of the Queried Structure in our Local Data Warehouse

Using the queried structure in SMILES format as a pattern, searching is performed onto our Local Data Warehouse via Quick-Search algorithm to find all possible sub/structures occurrence of Antimicrobial Structures

Quick Search algorithm compares the pattern characters with the text characters from left to right. If the current text character does not matched with the pattern character, the pattern always shifted to the right by at least one character but not more than m+1 characters. Based on heuristic function called bad-character shift [12], the right character of the current text characters will be used. In order to access the similarity between two structures, we need to choose an appropriate measure to compare the two structures and calculate the similarity percentage. The similarity coefficients used in this study were Jaccard's coefficient and Jaccard's distance due to their efficiency in determining the similarity of chemical structures [25]. Fig. 5 below shows Quick Search algorithm flowchart:

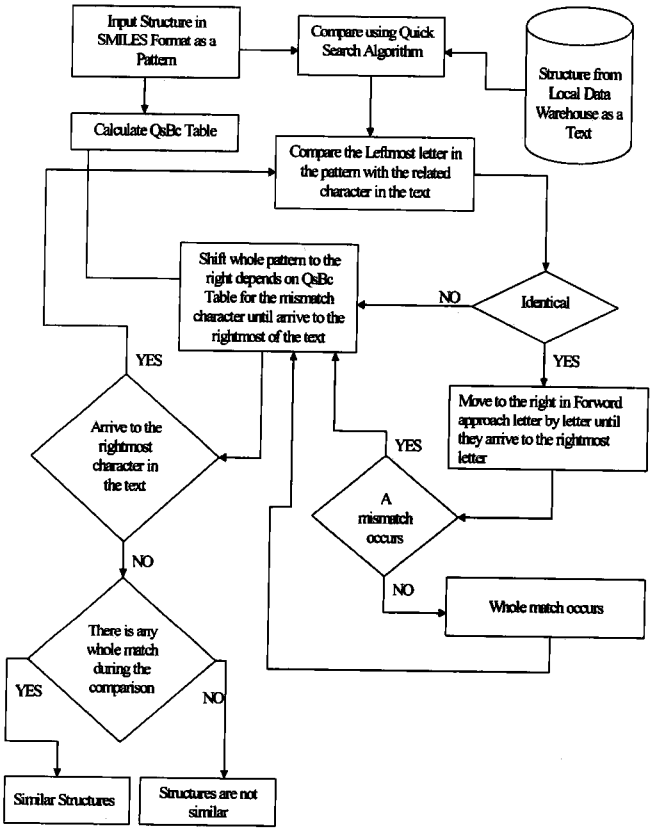


Fig. 5 Quick search Algorithm in Similar Structure Matching

## IV. Implementation

In this project, we used our local machine (Dell Intel(R) Core(TM) 2 CPU (1.80 GHz), 1014 MB RAM, Windows Vista 32-bit Operating System) with web based interface which runs on Microsoft IIS 7.0 platform. We used Macromedia Dream weaver MX version 6.0 to develop the web page which contain a combination of HTML, PHP 5.2.2 and JavaScript code. This web page retrieved data using PHP code through MySQL ODBC 3.51 driver which act as a middle-tier between MySQL Professional 5.1 database and front-end user. For creating and modifying NMRShiftDB on our local machine, we used MySQL-Front version 2.5 and as for graphics used, we created it by using Adobe Photoshop CS2 version 9.0.

Fig. 6 shows an example of screen shot for user input to draw the Antimicrobial Structure:

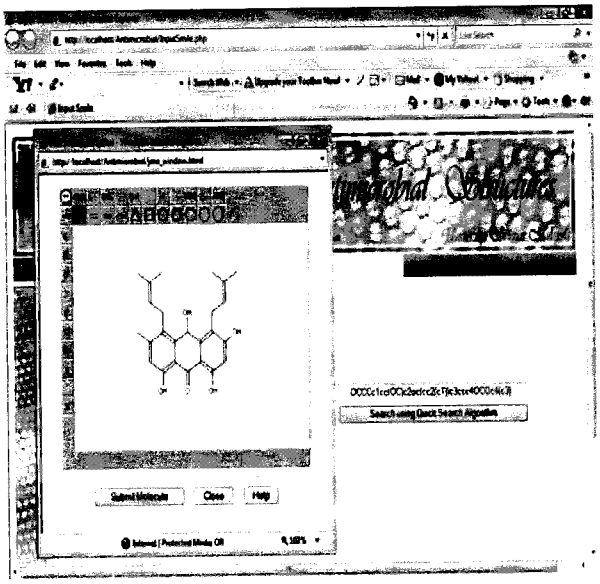

Fig. 6 Screen shot for user input and converting operation (1)

## V. Experimental Results

The performance of this search is evaluated using the accuracy and timing criteria's.

### A. Accuracy Criteria

The accuracy test determined whether the results that we gained from executing Quick Search algorithm are correct or not. Thus we tested the algorithm using input structure, and then we checked and compared the results with the existing structures in our Local Data Warehouse. For example if the input structure is "c3ccc", then the output will be shown as Fig. 7:

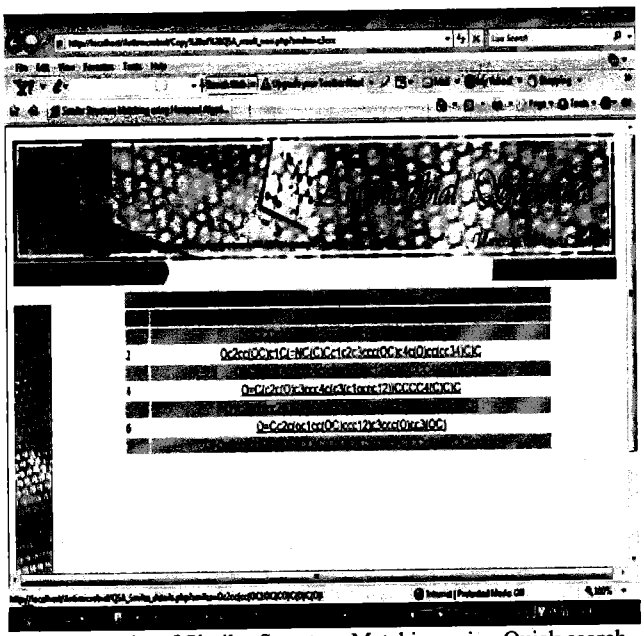

Fig. 7 Results of Similar Structure Matching using Quick search Algorithm

User can select any structure from the Similar Structures that appeared in Fig. 7. Then sample of corresponding information for the selected structure will be appeared as shown in Fig. 8:

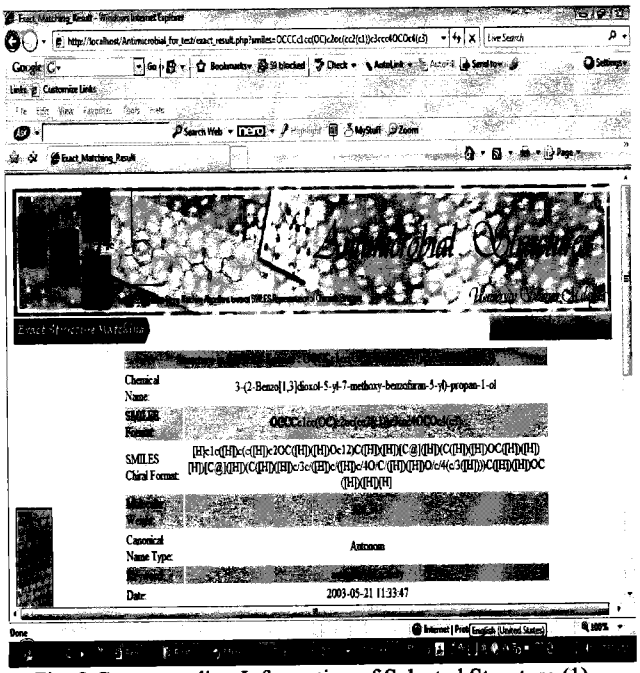


Fig. 8 Corresponding Information of Selected Structure (1)

Based on the results, we found that Quick Search algorithm is accurate algorithm and it gives us the expected results correctly.

### B. Timing

Fig. 9 below shows the time used to search a string using Quick Search algorithm in our Local Data Warehouse with different number of records from 60 - 20033 records.

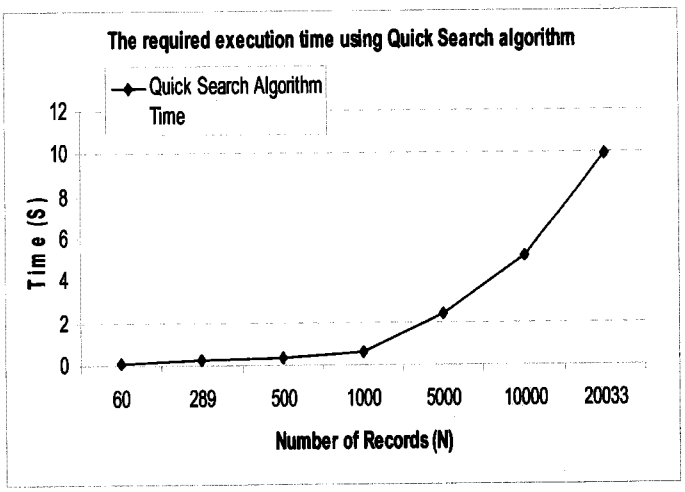


Fig. 9 The required time for executing structure using Quick Search algorithm

Based on the linear graph in Fig. 9, we concluded that the complexity of implementing Quick Search algorithm in our domain to search Antimicrobial Structure is in order N.

## VI. Conclusion and Future Work

In this paper we presented an approach to help biologists and chemists to search and manipulate Antimicrobial Structures using SMILES format by developing a searching tool that would response to user query using the JME editor and search it in our Local Data Warehouse using Quick Search Algorithm. The results shown that Quick search algorithm considered an accurate algorithm for our domain and it is implementation complexity is in order N. Further work, we are going to expand our Local Data Warehouse by extracting other Antimicrobial Structures from other databases.

## Acknowledgment

Our thanks to those who help us to develop the system; Belal Najjar and Siti Mazrishah, to Peter Ertle for providing the JME editor and to the Ministry of Science, Technology and Innovation for providing E-Science Grant .

## References

[1] Carlos Morel, "Bioinformatics for disease endemic countries: opportunities and challenges in science and technology development for health", Special Program for Research and Training in Tropical Diseases (TDR). Geneva, Switzerland, 2002, pp. 1-4.

[2] Chen Guang Li., "String Matching and the Knuth-Morris-Pratt Algorithm". Carleton University, Canada, 2006, pp. 1-8.

[3] Christian Charras and Thierry Lecroq, "Exact String Matching Algorithms". De Rouen University. France.

[4] Christoph Steinbeck and Stefan Kuhn. Open Content Databases and Open Source Libraries for Chemoinformatics. Cologne University Bioinformatics Center (CUBIC).

[5] Domenico Cantone and Simone Faro, "Forward-Fast-Search: Another Fast Variant of the Boyer-Moore String Matching Algorithm". Dipartimento di Matematica e informatica, Universita di Catania, Italy, 2003, pp. 10-24.

[6] Edward Reingold, Kenneth Urban and David Gries, "K-M-P string matching revisited". Department of Computer Science, Cornell University, USA, 1997, pp. 217-223.

[7] Greg Plaxton, "String Matching: Boyer-Moore Algorithm", Theory in Programming Practice. Department of Computer Science, University of Texas at Austin. 2005.

[8] Ireille Régnier and Wojciech Szpankowski, "Complexity of Sequential Pattern Matching *Algorithms*". Barcelona, Spain, 2004, pp.187-200.

[9] Jerome Mettetal and Ross Lippert, "Brute Force Algorithms: Motif Finding". 2004, pp. 1-7.

[10] Jun Xu and Arnold Hagler, "Chemoinformatics and Drug Discovery", Partners International. USA, 2002, pp. 566-600.

[11] Kanniah Rajasekaran, Gerald DeGray, Kanniah Rajasekaran, Franzine Smith, John Sanford, and Henry Daniell. "Expression of an Antimicrobial Peptide via the Chloroplast Genome to Control Phytopathogenic Bacteria and Fungi", Department of Molecular Biology and Microbiology and Center for Discovery of Drugs and Diagnostics, University of Central Florida, Florida, 2001, pp. 203-210.

[12] Maxime Crochemore and Thierry Lecroq, "Pattern matching and text compression algorithms". Chapter 2, pp. 12-14.

[13] NMRShiftDB. Available: http://nmrshiftdb.ice.mpg.de/nmrshiftdb. (Accessed February, 2007).

[14] Olivier Danvy and Henning Korsholm Rohde, "Obtaining the Boyer-Moore String-Matching Algorithm by Partial Evaluation". Department of Computer Science University of Aarhus, 2005, pp. 1-9.

[15] Peter Ertl, JME Editor. Available: http://www.molinspration.com. (Accessed February, 2007).

[16] Prasit Palittapongarnpim, "Thailand's bioinformatics initiatives", The National Center for Genetic Engineering and Biotechnology and Department of Microbiology. Faculty of Science, Mahidol University, Bangkok, Thailand, 2002, pp. 6-8.

[17] Rahul Thathoo, Ashish Virmani, S. Sai Lakshmi, N. Balakrishnan and K. Sekar1, "TVSBS: A fast exact pattern matching algorithm for biological sequences". India, 2006, pp. 47-53.

[18] Richard L. Rowley, R. Jeremy Rowley, John L. Oscarson and W. Vincent *Wilding*. "Development of an Automated SMILES Pattern Matching Program to Facilitate the Prediction of Thermo physical Properties by Group Contribution Methods", Department of Chemical Engineering, Brigham Young University. Provo, Utah, 2001, pp. 1110-1113.

[19] SMILES - A Simplified Chemical Language. Available: http://www.daylight.com. (Accessed March, 2007).

[20] Thomas E. Besser, Paul S. Morley, Michael D. Apley, Derek P. Burney, Paula J. Fedorka-Cray, Mark G. Papich, Josie L. Traub-Dargatz, and J. Scott Weese. "Antimicrobial Drug Use in Veterinary Medicine", 2005, pp. 617-629.

[21] Tim Bell, Matt Powell, Amar Mukherjee and Don Adjeroh, "Searching BWT compressed text with the Boyer-Moore algorithm and binary search". University of Central Florida, USA, 2001, pp. 1-10.

[22] TIMO RAITA, "Tuning the Boyer–Moore–Horspool String Searching Algorithm". University of Turku, Finland, 1992, pp. 879-884.

[23] Werner Arber, Daniel Nathans and Hamilton Smith, "DNA Mapping and Brute Force Algorithms". Berlin, Germany. pp. 1-29.

[24] Yusuke Shibata, Tetsuya Matsumoto, Masayuki Takeda, Ayumi Shinohara and Setsuo Arikawa, "A Boyer-Moore Type Algorithm for Compressed Pattern Matching". Montreal, Canada, 2004, pp.1-20.

[25] Peter Willet,John M Barnard and Geoffrey M. Down, 1998, Chemical Similarity Searching, Krebs Institute for bimolecular research and department of Information Studies,University of Sheffiled,UK, pp 983-996.