ELSEVIER

Computers in Biology and Medicine 35 (2005) 717–724

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Assessment of approximate string matching in a biomedical text retrieval problem

J.F. Wang[a], Z.R. Li[a, b], C.Z. Cai[a, c], Y.Z. Chen[a, *]

[a] *Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*
[b] *Department of Chemistry, Sichuan University, Chengdu 610064, P.R. China*
[c] *Department of Applied Physics, Chongqing University, Chongqing 400044, P.R. China*

Received 11 February 2004; accepted 2 June 2004

Abstract

Text-based search is widely used for biomedical data mining and knowledge discovery. Character errors in literatures affect the accuracy of data mining. Methods for solving this problem are being explored. This work tests the usefulness of the Smith–Waterman algorithm with affine gap penalty as a method for biomedical literature retrieval. Names of medicinal herbs collected from herbal medicine literatures are matched with those from medicinal chemistry literatures by using this algorithm at different string identity levels (80–100%). The optimum performance is at string identity of 88%, at which the recall and precision are 96.9% and 97.3%, respectively. Our study suggests that the Smith–Waterman algorithm is useful for improving the success rate of biomedical text retrieval.

Keywords: Bioinformatics; Biomedical; Data mining; Dynamic programming; Herb; Herbal medicine; Literature; Literature search; Medicine; Medicinal plant; Medinformatics; Plant; Smith–Waterman algorithm; Text; Text matching; Word; Word matching

1. Introduction

Text-based knowledge discovery and literature data mining tools are important for facilitating biomedical information extraction, fact finding, relationship search, and concept discovery [1–7]. Considerable

[*] Corresponding author. Tel.: 65-6874-6877; fax: 65-6774-6756
E-mail address: yzchen@cz3.nus.edu.sg (Y.Z. Chen).

0010-4825/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2004.06.002

interest has been directed at development of reliable text-based search methods for biomedical applications [1–3,7–16].

Text-based search tools generally rely on some form of text matching, which may find difficulty in cases of misspelled words [17] and morpheme or cross-lingual related problems [18]. In some cases, these problems occur at a non-negligible rate. For instance, it has been reported that text collections digitized via optical character recognition (OCR) contain 7–17% error [17]. Typographical and spelling errors have been found to be at the level of 1–3.2% and 1.5–2.5%, respectively [17]. The error rate for typing words or names from a foreign language can be as high as 38% [19]. This kind of error rate is of particular concern to biomedical fields with a larger percentage of words or names from Latin and other languages. These fields include medicine, microbiology, medicinal plants, herbal and traditional medicines. Therefore, search methods capable of dealing with these errors are useful for facilitating biomedical data mining.

Approximate string-matching (ASM) methods have been developed for literature search that allows mismatch, deletion and insertion errors in the text [17,19,20]. Most ASM methods are based on dynamic programming (DP). One such method, the Smith–Waterman algorithm, has been widely used for protein and DNA sequence alignment [21]. The advantage of this algorithm is its capability in matching texts that contain gaps of various lengths as well as mismatches. By modifying its parameters to conform to the problem of text matching, this algorithm may be used as a general ASM method for biomedical text retrieval as well as for protein and DNA sequence alignment.

This work examines the usefulness of the Smith–Waterman algorithm with affine gap penalty [22] for the retrieval of biomedical texts with a larger percentage of errors. The parameters for gap opening and extension in this algorithm are modified to suit text matching. The specific problem concerns with the literature search of active ingredients from medicinal herbs for certain therapeutic applications. Information about herbal active ingredients and that of therapeutically used herbs are from literatures of two different disciplines, medicinal chemistry and herbal medicine. Because of unfamiliarity with Latin words among some researchers, higher rates of grammatical, spelling, and format-related errors occur in some of these literatures. Moreover, there are a substantial number of herbs with a name highly similar to that of another herb. Hence, this problem is ideal for testing and adjusting ASM methods.

2. Methods

2.1. Data sources

Therapeutically used medicinal herbs were collected from herbal medicine literatures, from which a collection of 8000 medicinal herbs were generated (Collection I). Information about the chemical ingredients of medicinal herbs was collected from medicinal chemistry literatures and databases, from which a collection of 1900 medicinal herbs with known ingredients were generated (Collection II). These two collections are used in this work for analysis of the usefulness of ASM method in biomedical text retrieval. Manual inspection shows that various forms of errors occur in the herb names from both collections. The number of herbs with the same correct name in both databases is 480, while the number of those with erroneous names or multiple names in one or both source is 151. These 151 herbs are ideal for evaluating the text-matching algorithm and they are used in this work for a text-matching study.

All ASM methods allow a certain degree of mismatches and gaps in the searched text. This might result in the incorrect match of high-similarity texts. In this work, pairs of high-similarity medicinal herb names