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
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

are collected from Collections I and II. The criterion for selecting these pairs is that the text identity between the two herb names is between 80% and 100%. A total of 339 pairs of herbs were collected and used for text-matching study in this work.

2.2. Approximate string-matching algorithm

Our text-matching algorithm is based on the Smith–Waterman algorithm with affine gap penalty [22]. To optimally match a target string $\{a_i\}$ of l_1 characters ($i = 1, \ldots, l_1$) against a query string $\{b_j\}$ of l_2 characters ($j = 1, \ldots, l_2$), a matrix S of size $l_1 \times l_2$ can be constructed and the path that connects largest matrix elements defines the optimal match between the two strings of characters. The matrix element $S_{i,j}$ is calculated by using the following recursion formula:

$$S_{i,j} = \max \begin{cases} \max\limits_{1 \leqslant k \leqslant i} (S_{i-k,j} - W(k)) \\ S_{i-1,j-1} + S(a_i, b_j) \\ \max\limits_{1 \leqslant k \leqslant j} (S_{i,j-k} - W(k)) \\ 0, \end{cases} \tag{1}$$

where $0 \leqslant i \leqslant 1_1$ and $0 \leqslant j \leqslant 1_2$ and the initial conditions are $S_{0,j} = S_{i,0} = 0$ for all $0 \leqslant i \leqslant l_1$ and $0 \leqslant j \leqslant l_2$. $S(a_i, b_j)$ is the score for match or mismatch, and $W(k)$ is the penalty for a gap of length k

$$W(k) = h + gk, \quad k \geqslant 1, \tag{2}$$

where h and g are the gap opening penalty and gap extension penalty, respectively.

In this work, the following parameters are used for text matching:

$$h = 1,$$

$$g = 2,$$

$$S(a_i, b_j) = \begin{cases} 2, & a_i = b_j, \\ 0, & a_i \neq b_j. \end{cases}$$

These parameters are generated by a trial-and-error procedure that produces optimum accuracy for text matching. Different parameter values are scanned and the predicted text matching pairs are compared with the correct answers, from which the optimum parameters that give the highest percentage of correct answers can be selected.

2.3. Evaluation measures

In the study of information retrieval, prediction accuracy is routinely measured by three quantities: Recall (R), Precision (P) and F-measure (F) [23]. Recall is the percentage of correctly matched texts with respect to all the truly matched texts. Precision is the percentage of correctly matched texts with respect to the matched texts (including correctly and incorrectly matched texts). There is a trade-off between precision and recall. Improvement of recall is generally at the expense of precision. To achieve a sufficiently high recall with reasonable precision, a balance of recall and precision needs to be considered. A quantity, F-measure, is introduced to enable one to find the balance point [23] and it is given by

$$F = (\alpha P^{-1} + (1 - \alpha) R^{-1})^{-1}, \tag{3}$$

where α is a parameter in the region between 0 and 1 to control the relative importance of P and R. Earlier studies indicated that best text-matching results can be obtained when $\alpha = 0.5$ [23], which suggests that recall is half as important as precision. This value of α is used in this work, under which

$$F = 2 \times P \times R/(P + R). \tag{4}$$

3. Results, discussion and conclusion

Table 1 gives examples of the typical name errors or multiple names for some of the medicinal herbs studied in this work. There are several types of frequently occurring errors due to unfamiliarity with the grammars of herb names. For instance, some herbs are named after specific persons. If the specific epithet is named after a person whose name ends with a vowel or -er, the letter -i should be added. Otherwise if the person's name ends in a consonant, the letters -ii should be added. In the first situation, the correct name is like '*Boswellia carteri*'. In the second case, the correct name is like '*Elsholtzia oldhamii*'. If the genus name has the masculine ending of -us, the specific name might be spelled *albus*, but if the genus has a feminine spelling it would be *alba*. As shown in Table 1, the name of *Melilotus albus* is incorrectly given as '*Melilotus alba*' in Collection II. Also, specific epithets derived from geographical names usually end with *-ensis*, or *-nus*, *-inus*, *-ianus*, or *-icus*. Thus the name of '*Bupleurum chinense*' is incorrect and the correct name is "*Bupleurum chinensis*" as shown in Table 1.

There are also spelling errors in some of the herb names used in this study. As shown in Table 1, these generally involve mistyped letters, missing letters (deletions) or extra letters (insertions). Moreover, there are herbs with multiple names. For instance, '*Aconitum koreanum*' and '*Aconitum coreanum*' are the two names of the same herb. Sometimes, a hyphen "-" can be used or not used in a herb name, as in the case of '*Panax pseudo-ginseng*' and '*Panax pseudoginseng*'. These cases can be regarded as a special case of multiple herb names.

Table 1
Examples of name errors or multiple names of medicinal herbs

Name of a herb from Collection II	Name of a herb from Collection I	Source of problem	Type of problem	String identity between two names (%)
Elsholtzia oldhami	*Elsholtzia oldhamii*	Collection I	Grammatical error	94.7
Boswellia carteri	*Boswellia carterii*	Collection II	Grammatical error	94.4
Melilotus alba	*Melilotus albus*	Collection II	Grammatical error	86.6
Bupleurum chinense	*Bupleurum chinensis*	Collection II	Grammatical error	89.5
Tetragonia tetragonoides	*Tetragonia tetragonioides*	Collection I	Spelling error	96.0
Chrysanthemum morifolium	*Chrysanthemum mofifolium*	Collection I	Spelling error	95.8
Astragalus membranaceus	*Astragalus menbranaceus*	Collection I	Spelling error	95.7
Indigofera endecaphylla	*Indigofera hendecaphylla*	Collection II	Spelling error	95.8
Sparganium stoloniferum	*Sparaganium stoloniferum*	Collection I	Spelling error	95.8
Aconitum soongoricum	*Aconitum soongaricum*	Collection II	Spelling error	95.0
Spatholobus suberetus	*Spatholobus suberestus*	Both	Spelling error	95.4
Panax pseudo-ginseng	*Panax pseudoginseng*	Both	Multiple name	95.2
Aconitum koreanum	*Aconitum coreanum*	Both	Multiple name	94.1

Table 2
Statistics of text-matching performance of the Smith–Waterman algorithm with affine gap penalty

Minimum string identity in ASM (%)	TP	FP	FN	Recall (%)	Precision (%)	F-measure (%)
100	480	0	151	76.1	100.0	86.4
95	524	0	107	83.0	100.0	90.7
94	556	1	75	88.1	99.8	93.6
93	571	2	60	90.4	99.6	94.8
92	579	4	52	91.7	99.3	95.3
91	583	5	48	92.3	99.1	95.6
90	596	7	35	94.4	98.8	96.6
89	601	10	30	95.2	98.3	96.7
88	612	17	19	96.9	97.3	97.1
87	619	25	12	98.1	96.1	97.1
86	623	45	8	98.7	93.2	95.9
85	627	88	4	99.3	87.6	93.1
80	631	339	0	100.0	65.0	78.8

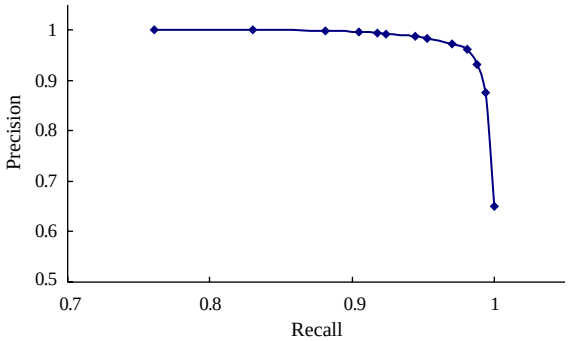

Fig. 1. Recall-precision curve.

Table 2 gives the results of text matching of several groups of herb names at different string-matching identities. These include 480 herbs with completely matched names in both Collection, 151 herbs with erroneous or multiple names, and 339 pairs of herbs of high-similarity names. It is found that, as the string identity decreases, the retrieved herbs are further increased. But at the same time a retrieval generates incorrect matched text which has high similarity with the text used to search. For example, when the minimum string identity (MSI) in the ASM algorithm is set at 95%, the recall is only 29.1%, but the precision is 100%. Fig. 1 presents the relationship between recall and precision. As MSI increases, so does the recall, but the precision is decreased. Therefore, a suitable minimum string identity needs to be selected when applying ASM algorithms to specific text-matching problems, at which we can get higher recall with relatively tolerable precision. From Table 2 one can find that, when the minimum string identity is at 88%, the F-measure is 97.14%, the recall is 96.99% and the precision is 97.30%, respectively. At this level, for the particular biomedical text retrieval problem studied in this work, one can find 139 more matched herbs, which are missed when using regular matching search methods.

Table 3
Examples of the correctly matched herb names by the Smith–Waterman algorithm with affine gap penalty

Herb name with error	Source of problem	No. of medline references by using herb name	No. of medline references by using matched name(s)
Rauvolfia serpentina	Collection I	25	61
Elsholtzia oldhamii	Collection I	0	0
Boswellia carteri	Collection II	2	9
Melilotus alba	Collection II	29	36
Bupleurum chinense	Collection II	23	28
Cinnamomum japonium	Collection II	0	1
Kalopanax septemlobu	Collection II	0	3
Artemisia dracunulus	Collection II	0	10
Fritillaria thumbergii	Collection II	3	33
Tetragonia tetragonioides	Collection I	1	3
Chrysanthemum mofifolium	Collection I	0	27
Astragalus menbranaceus	Collection I	0	137
Indigofera endecaphylla	Collection II	1	0
Sparaganium stoloniferum	Collection I	0	6
Aconitum soongoricum	Collection II	0	1
Orgza sativa	Collection I	0	5639

Examples of these matched medicinal herbs are given in Table 3. For instance, the name of the herb '*boswellia carterii*' in Collection I is different from that in Collection II. Its name in Collection I was obtained from a medical herbs dictionary, which is given as '*boswellia carterii*'. A search of Medline (http://www.ncbi.nlm.nih.gov/) by using '*boswellia carterii*' as keywords finds 7 references. The name of this herb in Collection II was obtained from the CCD database, which is given as '*boswellia carteri*'. A search of Medline by using '*boswellia carteri*' as keywords finds only 2 references. Only when using both of these names, one obtains relatively complete set of literature entries.

Spelling errors are almost inevitable, especially in the large volume of biomedical materials. Therefore, the use of an ASM-based search method such as that discussed in this work can be helpful for finding more references related to herbs, particularly if one is unfamiliar with herbs or the name of herbs contains spelling errors. For example, the herb '*Oryza sativa*' is misspelled as '*Orgza sativa*' in Collection I. A search of Medline by using '*Orgza sativa*' as keywords yields no references. By using our method, highly similar text '*Oryza sativa*' are identified and the use of this text as keywords enables the identification of a number of references in Medline.

As a by-product, ASM can also produce incorrect matches of names or words that are highly similar to another name or word. Examples of incorrectly matched herb names are given in Table 4. One way to optimize text-matching performance is to properly select the minimum string identity when applying the ASM algorithm so as to minimize the total number of errors. From Table 2, when minimum string identity is at 80%, we can find 151 more matched herbs, but the incorrectly matched herbs increased to 339. When the minimum string identity increased to 88%, there were also 17 mismatched herbs. Refinement of the mismatch and gap penalty functions of ASM may also be helpful in optimizing the rate of correct text matches while minimizing that of incorrect matches of high similarity texts. For instance, match/mismatch scores can be altered to favor common OCR, typographical and other mistakes, such as the y → g substitution that occurs in the case of Oryza → Orgza.

Table 4
Examples of incorrectly matched pairs of high-similarity names of different herbs

String identity between two herb names	Name of herb from Collection II	Name of herb from Collection I
94.4	*carpus macrophylla*	*Carnus macrophylla*
93.8	*alpinia japonica*	*Alpinia japonia*
92.9	*perus communis*	*Pyrus communis*
91.7	*rola odorata*	*Rosa odorata*
90.0	*Scutellaria discolor*	*Stellaria discolor*
89.5	*Dendrobium gibsonii*	*Dendrobium wilsonii*
88.9	*Laminaria japonica*	*Linaria japonica*
87.5	*Dioscorea batata*	*Dioscorea alata*
86.7	*Vicia amurensis*	*Vitis amurensis*
86.4	*Cymbidium aloifolium*	*Cymbidium longifolium*
85.7	*Cryptomeria japonica*	*Cryptotaenia japonica*
85.0	*Lygodium japonicum*	*Lycopodium japonicum*
84.6	*Phellodendron sachalinense*	*Phellodendron chinense*
84.2	*Salvia przewalskii*	*Sabina przewalskii*
83.3	*Artemisia mexicana*	*Artemisia keikana*
82.6	*Glycyrrhiza uralensis*	*Glycyrrhiza yunnanensis*
81.8	*Astragalus lusitanicus*	*Astragalus sinicus*
80.0	*Thalictrum aquilegifolium*	*Thalictrum acutifolium*

4. Conclusion

Our study suggests the capability of ASM in the retrieval of biomedical texts that contain various errors. Optimum text matching may be achieved by proper selection of minimum string identity when applying the ASM method for the retrieval of biomedical texts. Other strategies for improving match/mismatch scores, such as the use of specific scores that favor common OCR and typographical errors, can also be explored.

References

[1] M. Andrade, A. Valencia, Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics 14 (7) (1998) 600–607.
[2] I. Iliopoulos, A.J. Enright, C.A. Ouzounis, TEXTQUEST: document clustering of MEDLINE abstracts for concept discovery in molecular biology, Pac. Symp. Biocomput. (2001) 384–395.
[3] E.M. Marcotte, I. Xenarios, D. Eisenberg, Mining literature for protein–protein interactions, Bioinformatics 17 (4) (2001) 359–363.
[4] J. Ding, D. Berleant, D. Nettleton, E. Wurtele, Mining MEDLINE: abstracts, sentences, or phrases?, Pac. Symp. Biocomput. (2002) 326–337.
[5] L. Hirschman, J.C. Park, J. Tsujii, L. Wong, C.H. Wu, Accomplishments and challenges in literature data mining for biology, Bioinformatics 18 (2002) 1553–1561.
[6] R. Mack, M. Hehenberger, Text-based knowledge discovery: search and mining of life-sciences documents, Drug Discovery Today 7 (11) (2002) S89–S98.
[7] B.J. Stapley, L.A. Kelley, M.J. Sternberg, Predicting the subcellular location of proteins from text using support vector machines, Pac. Symp. Biocomput. (2002) 374–385.

[8] C. Blaschke, M.A. Andrade, C. Ouzounis, A. Valencia, Automatic extraction of biological information from scientific text: protein–protein interactions, Proc. Int. Conf. Intell. Syst. Mol. Biol. 7 (1999) 60–67.
[9] K. Humphreys, G. Demetriou, R. Gaizauskas, Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures, Pac. Symp. Biocomput. (2000) 505–516.
[10] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, L. Hunter, EDGAR: extraction of drugs, genes, and relations from biomedical literature, Pac. Symp. Biocomput. (2000) 517–528.
[11] B. Stapley, G. Benoit, Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts, Pac. Symp. Biocomput. (2000) 529–540.
[12] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, M. Carroll, Automatic extraction of protein interactions from scientific abstracts, Pac. Symp. Biocomput. (2000) 538–549.
[13] C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky, GENIES: a natural language processing system for the extraction of molecular pathways from journal articles, Bioinformatics 17 (2001) S74–S82.
[14] L. Wong, PIES, a protein interaction extraction system, Pac. Symp. Biocomput. (2001) 520–531.
[15] G. Leroy, H. Chen, Filling preposition-based templates to capture information from medical abstracts, Pac. Symp. Biocomput. (2002) 350–361.
[16] L. Wong, Gaps in text-based knowledge discovery for biology, Drug Discovery Today 7 (2002) 897–898.
[17] G. Navarro, A guided tour to approximate string matching, ACM Comput. Surv. 33 (1) (2001) 31–88.
[18] S. Schulz, U. Hahn, Morpheme-based, cross-lingual indexing for medical document retrieval, Int. J. Med. Inform. 58–59 (2000) 87–89.
[19] K. Kukich, Techniques for automatically correcting words in text, ACM Comput. Surv. 24 (4) (1992) 377–440.
[20] R.A. Baeza-Yates, G.H. Gonnet, Fast string matching with mismatches, Inf. Comput. 108 (2) (1994) 187–199.
[21] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[22] O. Gotoh, An improved algorithm for matching biological sequences, J. Mol. Biol. 162 (1982) 705–708.
[23] C.J. Rijsbergen, Information Retrieval, second ed., Butterworths, London, 1979.

Mr. Wang obtained his M.Sc. in Bioinformatics at the National University of Singapore in 2003. He has been involved in the development of herbal medicine information system and the study of the mechanism of medicinal herbs. He has also contributed to the development of two bioinformatics databases and in the study of pharmacokinetic properties of drugs. His research work has led to two publications in bioinformatics.

Dr. Li obtained his Ph.D. in Physical Chemistry at SiChuan University in 1994 and his M.Sc. in Physical Chemistry at SiChuan University in 1988. He has been involved in the development of string-matching algorithms and molecular modeling algorithms and their applications in data mining, sequence analysis, and drug design. His research work has led to three publications in bioinformatics and 18 publications in physical chemistry.

Dr. Cai obtained his Ph.D. in Biophysics at ChongQing University in 2003 and his M.Sc. in Condensed Matter Physics at ChongQing University in 1991. He has been involved in the development of herbal medicine information system and the study of the mechanism of medicinal herbs. Moreover, he is using machine-learning methods for studying protein function and for the analysis of traditional Chinese medicine recipes. His research work has led to seven publications in bioinformatics and 22 publications in material sciences.

Dr. Chen obtained his Ph.D. in Statistical Physics at UMIST UK in 1989 and his M.Sc. in Theoretical Physics at the Institute of Theoretical Physics Chinese Academy of Sciences China in 1985. He began his research in Biophysics and Computational Biology in Purdue University USA in 1989. He worked at ISIS pharmaceuticals and then National University of Singapore since 1997. He has published over 70 papers, obtained one and filed two US patents, developed three bioinformatics and drug design software packages and six bioinformatics databases.