거리반경기반 대표문자열 문제의 NP-완전

(The Consensus String Problem based on Radius is NP-complete)

나 중 채 † 심 정 섭 ††
(Joong Chae Na) (Jeong Seop Sim)

요 약 여러 문자열들을 비교하여 유사성 또는 거리(오차)를 계산하는 문제는 패턴매칭, 웹검색, 바이오인포매틱스, 컴퓨터 보안 등 다양한 응용 분야와의 연관성으로 인해 활발히 연구되어 왔다. 주어진 문자열 집합 내의 여러 문자열들의 거리를 비교하기 위해 주어진 집합 내의 모든 문자열들을 대표하는 한 문자열(대표문자열)을 찾는 방법이 있다. 대표문자열 방법은 주어진 문자열 집합과 가장 유사한 한 문자열을 찾는 방법으로 주로 이용되는 목적함수는 거리반경과 거리합이 있다. 거리반경은 집합 내의 문자열들과 특정 문자열과의 거리들의 최대값으로 정의되며, 모든 문자열들 중에서 최소의 거리반경을 만드는 문자열을 주어진 문자열 집합에 대한 거리반경기반 대표문자열이라 한다. 거리합은 집합 내의 문자열들과 특정 문자열과의 거리들의 합으로 정의되며, 모든 문자열들 중에서 최소의 거리합을 만드는 문자열을 주어진 문자열 집합에 대한 거리합기반 대표문자열이라 한다. 본 논문에서는 메트릭 거리함수에 대해 거리반경기반 대표문자열 문제가 NP-완전임을 증명한다.

키워드 : 대표문자열, 거리반경, NP-완전

Abstract *The problems to compute the distances or similarities of multiple strings have been vigorously studied in such diverse fields as pattern matching, web searching, bioinformatics, computer security, etc. One well-known method to compare multiple strings in the given set is finding a consensus string which is a representative of the given set. There are two objective functions that are frequently used to find a consensus string, one is the radius and the other is the consensus error. The radius of a string x with respect to a set S of strings is the smallest number r such that the distance between the string x and each string in S is at most r. A consensus string based on radius is a string that minimizes the radius with respect to a given set. The consensus error of a string x with respect to a given set S is the sum of the distances between x and all the strings in S. A consensus string of S based on consensus error is a string that minimizes the consensus error with respect to S.*

In this paper, we show that the problem of finding a consensus string based on radius is NP-complete when the distance function is a metric.

Key words : consensus string, radius, NP-completeness

· 이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2008-331-D00479)
· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 국제과학기술협력재단의 지원을 받아 수행된 연구임(No. K20717000000709B010000710)

† 종신회원 : 세종대학교 컴퓨터공학과 교수
jcna@sejong.ac.kr
†† 종신회원 : 인하대학교 컴퓨터정보공학부 교수
jssim@inha.ac.kr

논문접수 : 2009년 2월 25일
심사완료 : 2009년 3월 29일

정보과학회논문지 : 시스템 및 이론 제36권 제3호(2009.6)

1. 서 론

두 개 또는 그 이상의 문자열들을 비교하여 문자열들 사이의 유사성 또는 거리(오차)를 계산하는 문제는 패턴매칭, 웹검색, 바이오인포매틱스, 컴퓨터보안 등 다양한 응용 분야와의 연관성으로 인해 활발히 연구되어 왔다 [1-4]. 주어진 문자열 집합 내의 여러 문자열들의 거리를 비교하는 것을 다중문자열비교(multiple string comparison)라 하는데 여기에는 크게 두 가지 방법이 있다. 첫 번째 방법은 주어진 집합 내의 모든 문자열들의 적절한 배치(다중문자열배치)를 찾는 방법이며 두 번째 방

법은 주어진 집합 내의 모든 문자열들을 대표하는 한 문자열(대표문자열)을 찾는 방법이다. 각 방법들은 다시 응용되는 분야에 따라 적절한 목적함수(objective function)들이 적용된다.

다중문자열배치(multiple string alignment)를 위해서는 주어진 문자열 집합 내의 모든 문자열들의 길이를 같게 만들어야 하는데 이를 위해서 공백문자를 각 문자열들에 적절히 삽입한다. 다중문자열배치 방법에 사용되는 대표적인 목적함수로서 쌍합(sum-of-pairs)이 있다[5]. 쌍합은 공백문자가 적절히 삽입되어 같은 길이로 만들어진 각 문자열 쌍들의 거리의 합으로 정의된다. 주어진 문자열 집합에 대해 쌍합이 최소인 배치를 찾는 문제는 각 문자열들에 공백문자들을 어떻게 삽입하여 배치했느냐에 따라 거리가 달라진다. 특정 쌍합을 갖는 다중문자열배치를 찾는 문제는 비메트릭 거리함수에 대해서 Wang과 Jiang[6]에 의해 NP-완전(NP-complete)인 문제로 증명되었고 메트릭 거리함수들에 대해서도 NP-완전이 증명되었다[7-9].

대표문자열(consensus string) 방법은 주어진 문자열 집합 내의 모든 문자열들을 대표하는 한 문자열을 찾는 방법으로 대표문자열이 주어진 집합 내의 문자열일 필요는 없다. 대표문자열을 찾는 방법에 주로 이용되는 목적함수는 거리반경(radius)과 거리합(consensus error)이 있다. 먼저 거리반경은 집합 내의 문자열들과 특정 문자열의 거리들의 최대값으로 정의되며, 모든 문자열들 중에서 최소의 거리반경을 만드는 문자열을 주어진 문자열 집합에 대한 거리반경기반 대표문자열(consensus string based on radius)이라 한다. 거리반경은 두 문자열의 거리를 측정하는 거리함수(예: 해밍거리, 편집거리, 가중편집거리 등)에 따라 달라진다. Frances와 Litman[10]이 주어진 문자열들이 이진문자열이고 거리함수는 해밍거리일 때, 거리반경기반 대표문자열 문제(결정형)가 NP-완전임을 증명하였고 다항식시간 근사알고리즘들에 대한 연구 결과들이 제시되었다[11-14]. 본 논문에서는 일반 알파벳에 대한 문자열 집합 및 메트릭 거리함수에 대해 거리반경기반 대표문자열 문제가 NP-완전임을 증명한다.

거리합은 집합 내의 문자열들과 특정 문자열과의 거리들의 합으로 정의되며, 모든 문자열들 중에서 최소의 거리합을 만드는 문자열을 주어진 문자열 집합에 대한 거리합기반 대표문자열(consensus string based on consensus error)이라 한다. 거리합 역시 거리함수에 따라 달라지는데, 거리합기반 대표문자열 문제는 Wang과 Jiang[6]에 의해 비메트릭 거리함수에 대해 APX-완전임이 증명되었고 Sim과 Park[15]에 의해 특정 메트릭 거리함수에 대해 NP-완전임이 증명되었다. Elias는 모든 알파벳과 모든 메트릭 거리함수에 대해 NP-완전임을 증명하였다[9].

본 논문의 구성은 다음과 같다. 2장에서는 관련 용어 및 개념을 설명하고 3장에서는 NP-완전을 증명한다. 4장에서 결론을 맺는다.

2 관련 연구

문자열이란 문자집합(알파벳) Σ의 0개 이상의 문자가 연결된 형태를 말한다. Σ에 대한 모든 문자열의 집합을 Σ^*로 표기한다. 공백문자는 Δ로 표기하고 공백문자열은 ϵ으로 표기한다. 문자열 x의 길이를 $|x|$로 표기하고 $x[i]$는 x의 i번째 문자를 나타낸다. 문자열 x를 i번 반복해서 연결한 형태를 x^i로 나타낸다. 문자열 x에서 0개 이상의 문자들을 삭제하여 얻어지는 문자열 y를 x의 부분서열(subsequence)이라 하고 반대로 x를 y의 상위서열(supersequence)이라 한다.

두 문자열 x와 y가 주어졌을 때, x와 y의 거리(오차)를 나타내는 거리함수 $d(x,y)$는 x를 y로 변환하는데 필요한 최소 비용을 나타낸다. 잘 알려진 거리함수들 중 편집거리(edit distance)는 x를 y로 변환하기 위해 필요한 최소의 편집연산(삽입, 삭제, 교체)의 수로 정의된다. 해밍거리(Hamming distance)는 x와 y의 길이가 같을 때 정의되며, x를 y로 변환하는데 필요한 최소의 교체연산의 수로 정의된다. 편집거리는 비용행렬(cost matrix)을 이용하여 일반화할 수 있다. 비용행렬은 모든 문자쌍에 대한 교체 비용과 모든 문자들의 삽입, 삭제 비용을 정의한다. 가중편집거리(weighted edit distance)는 비용행렬을 이용하여 x를 y로 변환하기 위해 필요한 최소의 비용을 나타낸다. 이때, 비용행렬이 $a,b,c \in \Sigma \cup \{\Delta\}$에 대해, 다음 네 가지 조건을 만족하면 메트릭(metric)이라 한다. (i) $d(x,y) \geq 0$, (ii) $d(a,b) = d(b,a)$, (iii) $d(a,a) = 0$, (iv) $d(a,c) \leq d(a,b) + d(b,c)$.

문자열 집합 $S = \{s_1,...,s_m\}$와 거리함수 d가 주어졌을 때, S에 대한 문자열 x의 거리합과 거리반경을 각각 $E(x,S)$와 $R(x,S)$로 표기한다. 이때, $E(x,S) = \sum_{i=1}^{m} d(x,s_i)$, $R(x,S) = \max_{i=1}^{m} d(x,s_i)$이다. $E(x,S)$를 최소화하는 문자열 x를 S에 대한 거리합기반 대표문자열이라 하고 $R(x,S)$를 최소화하는 문자열 x를 S에 대한 거리반경기반 대표문자열이라 한다.

문자열 집합 S에 대한 다중배치(multiple alignment)는, 각 행에 0개 이상의 공백문자 Δ를 삽입하여 길이를 일치시킨 S의 문자열로 구성된 2차원 행렬로 표현될 수 있다. 예를 들어, $S = \{abcae, bcd, abde\}$일 때, 그림 1은 S에 대한 다중배치의 한 예이다. 그림 1에 대한 거리함

a	b	c	a	e
Δ	b	c	d	Δ
a	b	Δ	d	e

그림 1 $S=\{abcae, bcd, abde\}$에 대한 다중배치

수가 편집거리일 때, S에 대한 쌍합은 8이며, 거리반경기반 대표문자열은 $abcde$이고 거리반경은 2이다. 또한, 거리합기반 대표문자열 역시 $abcde$이며 거리합은 4이다.

3. 거리반경기반 대표문자열 문제의 NP-완전

본 장에서는 거리함수가 메트릭인 가중편집거리일 때, 거리반경기반 대표문자열 문제가 NP-완전임을 최단공통상위서열(shortest common supersequence) 문제를 이용하여 보인다. 여기에서 사용된 증명방법은 [16]에서 사용된 방법과 유사하다. 최단공통상위서열 문제와 거리반경기반 대표문자열 문제의 결정형은 다음과 같다.

정의 1. 최단공통상위서열 문제

유한알파벳 Σ, Σ^*로 구성된 유한 문자열 집합 $S=\{s_1,...,s_n\}$, 양의 정수 m이 주어졌을 때, 각 s_i $(1\le i\le n)$의 상위서열이며 길이가 m 이하인 문자열 $w(\in\Sigma^*)$가 존재하는가?

최단공통상위서열 문제는 이진알파벳 즉, $|\Sigma|=2$인 경우에도 NP-완전임이 알려져 있다[17-19]. 본 논문에서 $\Sigma=\{0,1\}$로 가정한다.

정의 2. 거리반경기반 대표문자열 문제

유한알파벳 Σ', $(\Sigma')^*$로 구성된 유한 문자열 집합 S', 메트릭인 비용행렬 M, 양의 정수 t가 주어졌을 때, $R(u,S')\le t$인 문자열 $u(\in(\Sigma')^*)$가 존재하는가?

먼저 주어진 최단공통상위서열 문제를 다음과 같이 거리반경기반 대표문자열 문제로 변환한다.

(i) $\Sigma'=\Sigma\cup\{a,b,\#,\$,*_1,*_2,\Delta\}$,

(ii) $S'=\{v_i \mid v_i=\#\, s_i\,\$, 1\le i\le n\}\cup$
$\cup\{v_{n+1}=\#*_1^m\$\}\cup\{v_{n+2}=\#*_2^m\$\}$,

(iii) $t=m$.

또한, 비용행렬 M은 그림 2와 같이 정의한다. 그림에서 음영된 부분은 M이 메트릭을 만족하면서 m보다 큰 값이면 충분하다. 여기에서는 모두 $m+1$로 가정한다. 이 변환 과정은 다항식시간 내에 수행됨을 쉽게 알 수 있다.

보조정리 1. $R(u,S')\le m$을 만족하는 문자열 u는 $\#\alpha\$$의 형태이다. 이때, $\alpha=\{a,b\}^m$이다.

증명. 먼저 u는 반드시 한 개의 #과 한 개의 \$를 포함해야 함을 보인다. S' 내의 모든 문자열 v_i $(1\le i\le n+2)$가 한 개의 #과 한 개의 \$를 포함하고 있으며 #과 \$는 각각 자신 이외의 문자와의 거리가 $m+1$이므로, u 내에 # 또는 \$가 없거나 각각 두 개 이상인 경우 u와 각 v_i와의 거리는 m보다 커지게 된다.

	0	1	a	b	$*_1$	$*_2$	#	\$	Δ
0	0	2	1	2	2	2			1
1	2	0	2	1	2	2			1
a	1	2	0	2	1	1			1
b	2	1	2	0	1	1			1
$*_1$	2	2	1	1	0	2			2
$*_2$	2	2	1	1	2	0			2
#							0		
\$								0	
Δ	1	1	1	1	2	2			0

그림 2 비용행렬

이제 $\alpha=\{a,b\}^m$일 때, $u=\#\alpha\$$임을 보인다. u가 반드시 한 개의 #과 한 개의 \$를 포함해야 하므로 $u=\beta\#\alpha\$\gamma$라 하자. 이때, $\beta,\alpha,\gamma\in\{0,1,a,b,*_1,*_2,\Delta\}^*$이며 특히 α는 v_{n+1}과 v_{n+2}에 의해 길이가 최소한 m임을 알 수 있다. 먼저 α는 $*_1$과 $*_2$를 포함할 수 없음을 보인다. 만약 α가 $i(i\ge 1)$ 개의 $*_1$을 포함하고 있다면, u와 v_{n+2}와의 거리가 m을 넘지 않도록 하기 위해 α는 i개의 $*_2$ 역시 포함해야 하며, 나머지(최소 $m-2i$개) 문자는 a 또는 b로만 구성되어야 한다. 하지만 이 경우, u와 $v_i(1\le i\le n)$와의 거리는 모두 m을 초과하게 된다. 따라서 α는 $*_1$과 $*_2$를 포함할 수 없으며 길이는 m이어야 한다. 한편, $0,1,\Delta$는 v_{n+1}과 v_{n+2}에 있는 $*_1$과 $*_2$와의 거리가 2이므로 α는 $0,1,\Delta$ 역시 포함할 수 없다. 따라서, $\alpha=\{a,b\}^m$이며 $\beta=\gamma=\epsilon$이다. □

정리 2. 거리반경기반 대표문자열 문제는 NP-완전이다.

증명. 거리반경기반 대표문자열 문제가 NP에 속함은 쉽게 알 수 있다. NP-완전임을 보이기 위해서 최단공통상위서열 문제가 거리반경기반 대표문자열 문제로 환원(reduction)됨을 보인다. 즉, S에 대해 길이 m 이내의 공통상위서열이 존재하면 S'에 대해 거리반경이 m 이내인 대표문자열이 존재하며, 그 역도 성립함을 보인다.
(1) S의 공통상위서열 w(단, $|w|\le m$)가 존재하면 $R(u,S')\le m$인 문자열 u가 존재: w의 0을 a로, 1을 b로 치환하여 생성된 문자열을 α라 하자. 만약 $|w|<m$

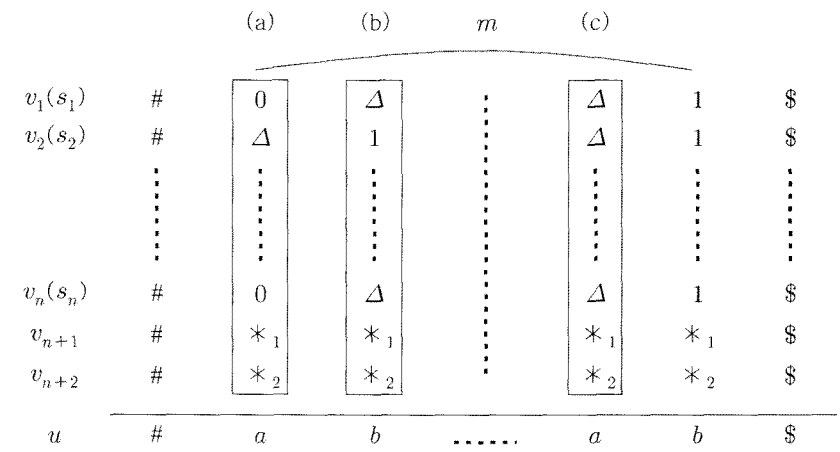

그림 3 $S' \cup \{u\}$에 대한 배치. (a) 모든 열이 $\{0, \Delta, *_1, *_2\}$들로만 구성된 경우, (b) 모든 열이 $\{1, \Delta, *_1, *_2\}$로만 구성된 경우, (c) 모든 열이 Δ인 경우

인 경우, a 또는 b를 α에 붙여서 길이를 m으로 만든다. $u = \#\alpha\$$라 하면 그림 3(a)와 같이 α의 각 a는 모두 $0, \Delta, *_1, *_2$과 배치(align)될 수 있으며 그림 3(b)와 같이 각 b는 모두 $1, \Delta, *_1, *_2$과 배치될 수 있다. 따라서 u와 각 $v_i (1 \le i \le n+2)$의 거리는 모두 m 이내이다. 즉, $R(u, S') \le m$이다.

(2) $R(u, S') \le m$인 문자열 u가 존재하면 S의 공통상위서열 w(단, $|w| \le m$)가 존재: 보조정리 1에 의해 $\alpha \in \{a, b\}^m$일 때, $u = \#\alpha\$$이다. α의 길이가 m이며 α와 각 $s_i (1 \le i \le n)$와의 거리는 최대 m이므로, α 내의 각 a는 s_i의 0 또는 Δ와 배치되어야 하며 α 내의 각 b는 1 또는 Δ와 배치되어야 한다. 이제 α의 각 a를 0으로, 각 b를 1로 치환한 문자열을 w라 하면 w는 길이 m인 $s_i (1 \le i \le n)$의 공통상위서열임을 알 수 있다. 만약 α의 a 또는 b가 모든 $s_i (1 \le i \le n)$의 Δ와 배치되면 그림 3(c)와 같이 α에서 그 문자를 삭제하여 길이가 m보다 짧은 S에 대한 공통상위서열을 얻을 수 있다. □

4. 결 론

본 논문에서는 거리반경기반 대표문자열 문제를 정의하고 메트릭인 가중편집거리에 대해 NP-완전임을 증명하였다. 다중문자열배치에 관한 연구는 바이오인포매틱스 등의 분야에서 매우 중요한 연구주제이다. 예를 들면 서로 다른 종들의 유전체에 같은 단백질이 결합하는 특정 서열 즉, 전사인자결합부위(transcription factor binding sites) 등을 찾기 위해 대표문자열 문제가 적용될 수 있다. 따라서 향후 다양한 크기의 알파벳, 그리고 다양한 거리함수에 대한 대표문자열 문제에 대한 연구가 필요할 것으로 생각된다.

참 고 문 헌

[1] Dan Gusfield, Algorithms on strings, trees, and sequences: computer science and computational biology, Cambridge University Press, 1997.

[2] S.F. Altschul and D.J. Lipman, Trees, Stars, and Multiple Biological Sequence Alignments, SIAM J. Appl. Math. 49(1), pp. 197-209, 1989.

[3] M.S. Waterman, Introduction to computational biology: Maps, sequences and genomes, CHAPMAN& HALL/CRC, 1995.

[4] X. Zha and S. Sahni, Highly compressed Aho-Corasick automata for efficient intrusion detection, ISCC 2008, pp. 298-303, 2008.

[5] H. Carrillo and D. Lipman, The multiple sequence alignment problem in biology, SIAM J. Appl. Math., 48(5), pp. 1073-1082, 1988.

[6] L. Wang and T. Jiang, On the Complexity of Multiple Sequence Alignment, Journal of Computational Biology, 1(4), pp. 337-348, 1994.

[7] P. Bonizzoni and G. Vedova, The complexity of multiple sequence alignment with SP-score that is a metric, Theoretical Computer Science, 259(1-2), pp. 63-79, 2001.

[8] W. Just, Computational complexity of multiple sequence alignment with sp-score, Journal of Computational Biology, 8, pp. 615-623, 2001.

[9] I. Elias, Settling the intractability of multiple alignment, Journal of Computational Biology, 13(7), pp. 1323-1339, 2006.

[10] M. Frances and A. Litman, On covering problems of codes, Theory of Computing Systems, 30(2), pp. 113-119, 1997.

[11] A. Ben-Dor, G. Lancia, J. Perone, and R. Ravi, Banishing bias from consensus sequences, In Proceedings of Symposium on Combinatorial Pattern Matching, pp. 247-261, 1997.
[12] L. Gasieniec, J. Jansson, and A. Lingas, Efficient approximation algorithms for the Hamming center problem, In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 905-906, 1999.
[13] K. Lanctot, M. Li, B. Ma, S. Wang, and L. Zhang, Distinguishing string selection problems, In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 633-642, 1999.
[14] M. Li, B. Ma, and L. Wang, Finding similar regions in many strings, In Proceedings of Annual ACM Symposium on Theory of Computing, pp. 473-482, 1999.
[15] J.S. Sim and K. Park, The consensus string problem for a metric is NP-complete, Journal of Discrete Algorithms, 1(1), pp. 111-117, 2003.
[16] J.S. Sim, C.S. Iliopoulos, K. Park, and W.F. Smyth, Approximate periods of strings, Theoretical Computer Science, 262(1-2), pp. 557-568, 2001.
[17] D. Maier, The complexity of some problems on subsequences and supersequences, J. ACM, 25, pp. 322-336, 1978.
[18] M. Middendorf, More on the complexity of common superstring and supersequence problems, Theoretical Computer Science, 125, pp. 205-228, 1994.
[19] K.J. Räihä, E. Ukkonen, The shortest common supersequence problem over binary alphabet is NP-complete, Theoretical Computer Science, 16, pp. 187-198, 1981.

나 중 채

1998년 서울대학교 컴퓨터공학과 학사 2000년 서울대학교 컴퓨터공학과 석사 2005년 서울대학교 컴퓨터공학부 박사 2006년 Department of Computer Science, University of Helsinki, Postdoctoral researcher. 2007년 건국대학교 u-science 기반 신기술 융합사업단, 연구교수. 2008년~현재 세종대학교 컴퓨터공학과 전임강사. 관심분야는 컴퓨터이론, 알고리즘, 생물정보학

심 정 섭

1995년 서울대학교 컴퓨터공학과 학사 1997년 서울대학교 컴퓨터공학과 석사 2002년 서울대학교 전기컴퓨터공학부 박사. 2002년~2004년 한국전자통신연구원 (선임연구원). 2004년 9월~현재 인하대학교 컴퓨터정보공학부(조교수). 관심분야는 알고리즘, 최적화이론, 바이오인포매틱스