数理モデル化と問題解決 20－10
（1998．7．24）

一般化suffix arrayによるタンパク質アミノ酸配列集合からの文字列出現特性の解析

川口俊朗 粕川雄也 松田秀雄 橋本昭洋
〒560-8531 豊中市待兼山町 1-3
大阪大学大学院 基礎工学研究科 情報数理系専攻

近年，いくつかの生物種では全DNA塩基配列が読みとられた．DNA配列の中で実際にタンパク質に翻訳されると予想される部分もアミノ酸配列の形で公開されている．それら配列は生物の持つ機能を考える上で重要な意味があると考えられている．アミノ酸配列が類似するタンパク質は同じ機能を持つことも分かっている．本論文では，一般化suffix arrayというデータ構造を用いて1つの生物のアミノ酸配列に含まれるすべての部分文字列から最も頻繁に出現する文字列を探索する．さらに，実際に公開されている11種類の生物種に対して本提案手法を実行した結果についても示している．

Analysis of charactristic strings in amino-acid sequences using generalized suffix array

Toshiro KAWAGUCHI, Takeya KASUKAWA, Hideo MATSUDA,
and Akihiro HASHIMOTO
Department of Informatics and Mathematical Science,
Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, 560-8531

Recently, complete genome sequences of several oraganisms are sequenced. Moreover amino-acid sequences are stored in public databases, and analysis of their sequences elucidates functions of oraganisms. It is known that sequences having similar substrings have the same function. This paper proposes to find the most frequently ocurred string in all amino-acid sequences in one organism by using generalized suffix array. Moreover, results of the analysis for 11 complete genomes are reported.

A	Alanine	L	Leucine
R	Arginine	K	Lysine
N	Asparagine	M	Methionine
D	Aspartic acid	F	Phenylalanine
C	Cysteine	P	Proline
Q	Glutamine	S	Serine
E	Glutamic acid	T	Thereonine
G	Glycine	W	Tryptophan
H	Histidine	Y	Tyrosine
I	Isoleucine	V	Vlaline

表 1: アミノ酸の表記法

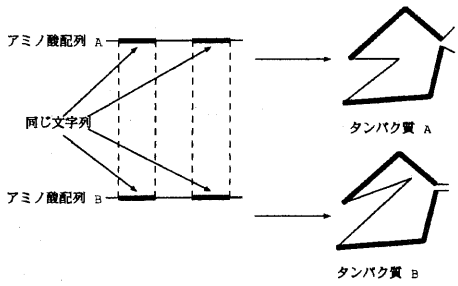

図 1: アミノ酸配列と立体構造

1 はじめに

分子生物学の分野では，生物の持つ DNA 塩基の並びを文字列として読み取り，解析する試みが数多くなされている．文字列として読み取られた DNA 塩基配列や塩基配列の翻訳であるアミノ酸配列は，WWW やデータベースなどの手段により公開され，コンピュータによる解析に用いられている．近年，読取り技術の向上により，生物種の全 DNA 配列（完全ゲノム配列）を読み取ることが可能になり，すでにいくつかの生物種の完全ゲノム配列が読み取られ GenBank[1] などのデータベースに登録されている．

DNA 塩基配列は 3 文字ずつ 20 種あるアミノ酸の 1 つに翻訳されタンパク質を構成していく．アミノ酸は 20 種のアルファベット (表 1) で表記される．

タンパク質はさまざまな立体構造をとるが，共通した文字列をもつアミノ酸配列のタンパク質は立体構造もよく似ており (図 1) 同じ機能をもつことが分かっている．

ある生物において同じ機能のタンパク質を見つけだし，出現の特徴を解析をすることは重要な意味をもつと考えられる．そこで，文字列の出現頻度を調べる方法として一般化 suffix array を用いる．この方法は，後述するように空間使用量が非常に少なくてすむので，長大となる完全ゲノムのアミノ酸配列の解析をすることが可能である．本論文では，実際に公開されている 11 の生物種に対する実行結果も示す．

2 一般化 suffix array

2.1 文字列

S を長さ n（$n \geq 1$）の文字列 $a_1 a_2 \ldots a_n$ とする（なお，$a_1, a_2, \cdots, a_n$ はそれぞれ 1 つの文字を表すものとする）．S に対して，$1 \leq i \leq j \leq n$ のとき，$a_i a_{i+1} \ldots a_j$ と表される文字列を S の部分文字列といい，$S[i,j]$ と表す．特に，$S[1,j]$，$S[i,n]$ をそれぞれ S の接頭語 (prefix)，接尾語 (suffix) という．S に含まれるのすべての接尾語は，$X_i = S[i,n](1 \leq i \leq n)$ と表される．

そして，2 つの接尾語 X_i と X_j について，$S[i,k] = S[j,k]$ かつ $a_{i+k} < a_{j+k}$ であるような $k(0 \leq k \leq n)$ が存在する時，X_i が辞書式に X_j よりも小さいという．この時，k は X_i と X_j の共通する部分文字列の長さである．以降この k を $lcp(X_i, X_j)$ と表す．

2.2 suffix array

配列 $I[1], I[2], \ldots, I[n]$ は，$X_i(1 \leq i \leq n)$ と対応するインデックスであり，i の値を要素とする．suffix array とは，すべての接尾語を辞書式順序にならべかえた時の配列 I である [2]．つまり，$X_{I[1]} < X_{I[2]} < \cdots < X_{I[n]}$ となるように配列 I を定める．さらに後で述べる文字列検索の効率化のための配列 $H[i] = lcp(X_{I[i-1]}, X_{I[i]})(2 \leq i \leq n)$ も定める (図 2)．

	H	I					
1	-	4	A	R	R	A	Y
2	1	1	A	Y			
3	0	3	R	A	Y		
4	1	2	R	R	A	Y	
5	0	5	Y				

$S = ARRAY$

図 2: suffix array の例

2.3 一般化 suffix array

一般化 suffix array とは複数の文字列を扱えるように拡張された suffix array のことである [3]. 長さ $n_1, n_2, \cdots, n_m$ の m 個の文字列の集合 $\mathcal{S}$ を $S_1, S_2, \cdots, Sm$ とする. この時, S_j の接尾語を, $X_{(i,j)} = S_j[i, n_j](1 \leq i \leq n_j, 1 \leq j \leq m)$ と表す. 配列 $Seq[1], Seq[2], \cdots, Seq[m]$ は, $S_j(1 \leq j \leq m)$ と対応するインデックスであり, j の値を要素とする.

すべての文字列のすべての接尾語を辞書式順序にならびかえた時の配列 I, Seq が一般化 suffix array である. つまり, $N = n_1 + n_2 + \cdots + n_m$ とすると, $X_{(I[1],Seq[1])} < X_{(I[2],Seq[2])} < \cdots < X_{(I[N],Seq[N])}$ となるように配列 I, Seq を定める. suffix array と同様に配列 $H[i] = lcp(X_{(I[i-1],Seq[i-1])}, X_{(I[i],Seq[i])})(2 \leq i \leq N)$ も定める.

一般化 suffix array は整数を要素とする 3 つの配列 H, Seq, I (図 3) で表現され, 整数が 4bytes で表されるとすると, 空間使用量は $12N$bytes と非常に少なくてすむ.

2.4 作成アルゴリズム

一般化 suffix array を作成するアルゴリズムはいくつか研究されているが, 本研究では, 空間使用量が最も少ないクイックソートに基づいた方法 [4] を使用した. つぎにアルゴリズムの概略を説明する.

まず分割語として接尾語の 1 つを選ぶ. すべての接尾語を分割語と最初の 1 文字を比べ辞書的に, より小さい, 等しい, より大きいの 3 つのグループに分ける. そして, 3 つのグループでそれぞれ分割語を選び再帰的に並べかえていく. ここで, 例えば "等しい" グループでは最初の 1 文字が等しいものが集まっている事がわかっているので, 2 文字めを比べる. あとの 2 つは同じように最初の 1 文字で各グループに分ける. 配列 H の値もグループの境界の要素から順次求まっていく.

	H	Seq	I					
1	-	2	3	A	B			
2	1	2	1	A	R	A	B	
3	2	1	1	A	R	R	A	Y
4	1	1	4	A	Y			
5	0	2	4	B				
6	0	2	2	R	A	B		
7	2	1	3	R	A	Y		
8	1	1	2	R	R	A	Y	
9	0	1	5	Y				

$S_1 = ARRAY, S_2 = ARAB$

図 3: 一般化 suffix array の例

3 文字列出現の解析アルゴリズム

同じ機能をもつタンパク質のアミノ酸配列は同じ文字の並びを多数持っている. 例えば図 4 のようになる. 破線で囲った部分は全く同じ文字列を表す. ここで保存領域と書かれた部分はタンパク質の機能を決定する上で特に重要な部分であるために保存されていると考えられる.

そこで, 保存領域を求めるために 一般化 suffix array を用いて, 長さ l 以上の最も出現頻度の大きい文字列を求める.

$$lcp(X_{I[i]}, X_{I[j]}) = \min_{k \in [i, j-1]} (lcp(X_{I[k]}, X_{I[k+1]})) \quad (1)$$

(文献 [2]) であるので, 一般化 suffix array で配列 H の値が l 以上で連続する部分を求めればそれが, 長さ l 以上の共通部分文字列をもつ接尾語集合である. 出現頻度は集合に含まれる接尾語の数と等しく, 最大となるような接尾語集合の共通部分文字列が求める長さ l 以上の最も出現頻度の大きい文字列であり, 保存領域であると考えられる.

Strings", In Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, pp.360-369, 1997.

[5] H.P.Klenk, et al.: "The complete genome sequence of the hyperthermopholic, sulphate-reducing archaeon *Archaeoglobus fulgidus*", Nature, vol.390, no.6658, pp.364-370, 1997.

[6] C.M.Fraser , et al.: "Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*", Nature, vol.390, no.6660, pp.580-586, 1997.

[7] F.Kunst, et al.: "The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*", Nature, vol.390, no.66657, pp.249-256, 1997.

[8] F.R. Blattner, et al.: "The complete genome sequence of *Escherichia coli* K–12", Science, vol.277, no.5331, pp.1453-1462, 1997.

[9] R.D. Fleischmann , et al.: "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd", Science, vol.269, no.5223, pp.496-512, 1995.

[10] Jean-F. Tomb , et al.: "The complete genomes sequence of the gastric pathogen *Helicobacter pylori*", Nature, vol.388, no.6642, pp.539-547, 1997.

[11] D.R. Smith , et al.: "Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomic", Journal of Bacteriol, vol.179, pp.7135-7155, 1997.

[12] C.J. Bult , et al.: "Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*", Science, vol.273, no.5278, pp.1058-1073, 1996.

[13] C.M. Fraser , et al.: "The minimal gene complement of *Mycoplasma genitalium*", Science, vol.270, no.5235, pp.397-403, 1995.

[14] R. Himmelreich , et al.: "Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*", Nucleic Acids Research, vol.24, no.22, pp.4420-4449, 1996.

[15] T.Kaneko , et al.: "Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire gen ome and assignment of potential protein-coding regions", DNA research, vol.3, no.3, pp.109-136, 1996.