情 報 学 基 礎 36－2
（1995．2．7）

高速遺伝子配列検索

平岡進、西川哲夫、笠原直子、永井啓一
(株) 日立製作所中央研究所

検索配列に類似している配列をデータベースから高速に検索する技術を提案する。遺伝子配列を比較しスコアを求めるダイナミックプログラミング法は、計算量が大きく検索に用いることは困難である。一般的に用いられている遺伝子配列検索プログラムはあらかじめ当りをつけた類似部分周辺だけを調べるため、ダイナミックプログラミング法によれば類似している配列を見逃す可能性がある。本稿ではあらかじめ作成したスコア表を用いて、データベース中のすべての配列に対してスコアの上限値を高速に求める技術を提案する。スコアの上限値に基づいてダイナミックプログラミング法を行う配列を選択することで漏れのない高速遺伝子配列検索が実現できる。

Rapid similarity search algorithm of nucleic acid database

S.Hiraoka, T.Nishikawa, N.Kasahara, K.Nagai
Central Research Laboratory, Hitachi, Ltd

Smith and Waterman dynamic programming algorithm is too computer intensive to be used for database search. Most of the common database search programs use approximate scores to reduce search time instead of Smith and Waterman score. They may miss some sequence similarities. We present an algorithm for database search which uses score table of k-tuples. The method calculates the upper limit of Smith and Waterman score for all database sequences in nearly interactive time. The upper limit enables us to eliminate less similar sequences without missing any sequence similarities.

1 まえがき

検索配列に類似している配列をデータベースから検索する相同性検索について述べる。

近年、遺伝子工学の発展により遺伝子配列データが急増している。DNA配列を蓄積したデータベースであるGenBankには1994年12月現在、エントリー数237,775、総文字数230,485,928が収録されている。

実験によって新しく決定された配列の機能を推定する方法の一つとして、データベースから類似した配列を検索する相同性検索の重要性が増しつつある。相同性検索を行う最も基本的な方法として、検索配列とデータベース内の各配列との間でSmithとWatermanによって提案されたDP(Dynamic Programming)法によるアライメントを行い、高いスコア順に各配列を表示する方法がある[1)]。

DNA配列同士のアライメントにおいて一般的に用いられているスコアシステムは、n文字の挿入・欠失に対して-8n-4点、一致した1文字に対して4点、置換している1文字に対して-3点である。DP法では考えられる全てのアライメントの中で最大のスコアを与えるアライメントを求める。

急速に増大しているデータベースのサイズを考えるとデータベース内の全配列に対してDP法によるアライメントを行うことは困難である。そのため一般的には計算量が少ないFASTAまたはBLASTと呼ばれるプログラムが用いられている[2,3)]。DP法, FASTA,BLASTによる検索時間は検索配列と検索条件によって異なるが、ワークステーション(SUN SPARC station10)を用いてそれぞれほぼ10時間、10分、10秒程度である。

FASTAではDP法によるスコアよりも少ない計算量で求められるinitnと呼ばれるスコアを求めている。このスコアは検索配列とデータベースの配列で完全に一致する部分配列を探し出し、それらを繋ぎ合わせた経路でのスコアである。FASTAではこのスコアが高い順に配列を表示しており、上位の配列に対しては計算範囲を限定したDP法によるスコアを計算している。

BLASTも基本的にはFASTAと同じアルゴリズムを用いている。ただしFASTAよりも長い部分配列で一致を探し、部分配列の繋ぎ合わせの際に挿入・欠失を考慮していない。またデータベースを圧縮することでより高速化を図っている。BLASTではスコアに加えて、配列をポアソン分布で近似し検索配列と偶然一致する確率を求めている。そして確率の低い順に配列を表示しており、DP法によるスコアの計算は行っていない。

FASTA,BLASTは検索配列とデータベースの各配列で完全に一致する部分配列の付近だけを調べるため、DP法におけるスコアの高い配列を見落とす可能性がある。そこで検索もれがなく高速なDNA配列検索プログラムが必要と考えられる。

我々はこれまでに配列成分表[4)]を提案し、置換、挿入、欠失が少なくスコアの高い配列に限定し高速配列検索を実現した。配列成分表は文書の全文検索に用いられている文字成分表[5)]を遺伝子配列におけるあいまい検索に応用したものである。あらかじめ一定長のあらゆる配列に対して、データベース中の各配列に含まれるかどうかを調べて配列成分表を作成する。配列成分表を用いて検索配列の部分配列を多数含む配列を選択することで高速検索が実現できる。

本報告ではより一般的なDP法の高速化として、高速なスコア計算方法とそれをふるいとして用いることによるDP法の高速化について述べる。

2 スコア表

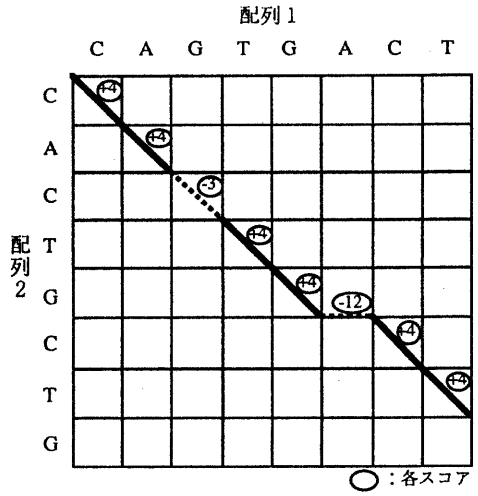
Smith-Watermanにより提案されたDP法は配列の部分同士を比較する方法だが、ここでは説明を容易にするため配列の全体同士を比較する場合について説明する。

DP法では考えられる全てのアライメントの中で最大のスコアを与えるアライメントを求める。全てのアライメントは比較する2配列をx,y方向に置いたマトリクス中の一経路で表すことができる。

図1は配列1 (CAGTGACT)と配列2 (CACTGCTG)間のアライメントの一例をマトリクス中の経路と共に示したものである。マトリクスにおける縦と横の線分はそれぞれ挿入と欠失に相当し、斜めの線分は一致または置換に相当する。各線分のスコアを経路に添って総計したスコアがこのアライメントのスコアである。

ここで例えば図2に示すように配列1を4文字づつCAGTとGACTに分割する。CAGTと配列2に対する経路と、GACTと配列2に対する経路を接続した経路と、それぞれの経路のスコアの和の集合を考える。これらは配列1と配列2に対する経路とスコアの集合を含む。即ちCAGTと配列2のDP法によるスコアとGACTと配列2のDP法によるスコアの和によっ

て、配列1全体と配列2のDP法によるスコアの上限値を求めることができる。



配列1 C A G T G A C T
配列2 C A C T G — C T G

図1　DP法によるアライメント

図2で分割したスコアの和が実際のDP法によるスコアよりも大きめの値となる理由は、分割部分で不連続な経路も計算に含まれるためである。実際、図2に示した経路に相当するアライメントは存在しない。

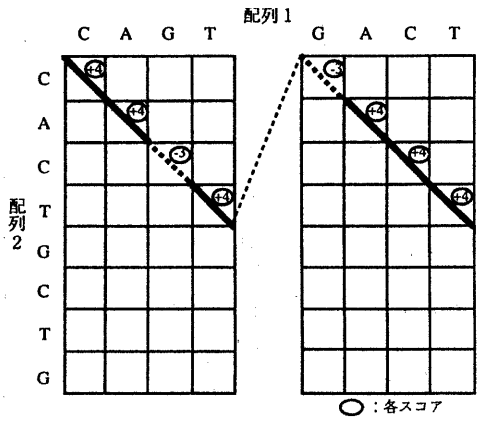


図2　配列分割アライメント

ここで、あらかじめ配列1を分割した部分配列CAGT、GACTと配列2のDP法によるスコアが求められていた場合には、それらの和の計算だけでDP法によるスコアの上限値を求めることができる。実際には図3に示すようにデータベース中の各配列に対して一定長のあらゆる配列とのDP計算を行い、その結果をスコア表として記憶しておく。

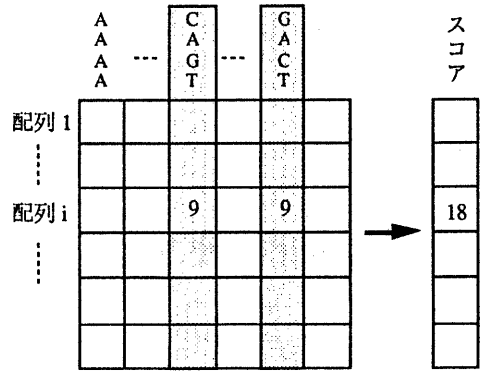


図3　スコア表

図2,3では検索配列の分割長ktupを4文字として説明した。実際には、十分な検索速度と性能を得るためにより大きいktupを用いた。スコア表の大きさのktup依存性を図4に示す。スコア表の各スコアを1byteで表したとしても、データベース中の一配列当りのスコア表の大きさは4^{ktup}となりGenBankの平均配列長1000文字よりも大きい。

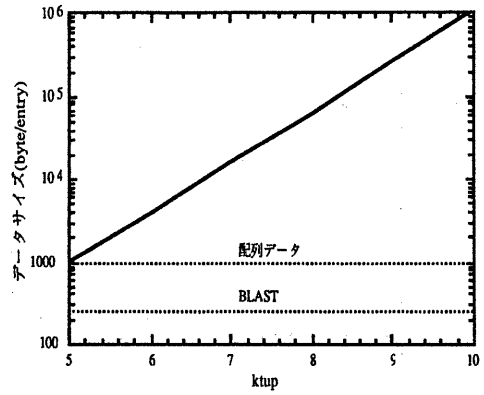


図4　データサイズ

スコア表は大きく、しかも作成にはかなりの計算量が必要となるが、出来上がったスコア表を基に分割したスコアの和を求める計算は極めて高速に行うことができる。また従来のDNA配列検索プ

ログラムと異なり、本方式ではDP法によるスコアの上限値が求められる。即ち、スコア表から求めたスコアの和が一定値以下の配列はDP法を行うまでもなくDP法によるスコアも一定値以下であることがわかる。本方式をふるいとして用いてDP法を行うことで、検索漏れがなく高速なDNA配列検索が可能である。

3 検索時間

検索配列の一定長の全部分配列を、FASTAはハッシュ表、BLASTはオートマトンに登録しデータベースの全配列をチェックする。ハッシュ表、オートマトン作成時間は検索時間全体のなかで無視できる程度である。そしてハッシュ表、オートマトンによるデータベース各配列のチェックと本方式のスコア表の和をとる処理はどれも計算量が大きくなく、それぞれの検索時間は検索の際にアクセスするデータの量で評価できると考えられる。そこで検索の際アクセスするデータ量を比較することで検索時間の比較の代用とした。

図5はBLASTと本方式のアクセスデータ量である。横軸は本方式の分割配列長ktup、縦軸はデータベースの一配列当たりのアクセスデータ量である。検索配列とデータベース中の配列の長さはGenBankの平均配列長に近い1000文字とした。

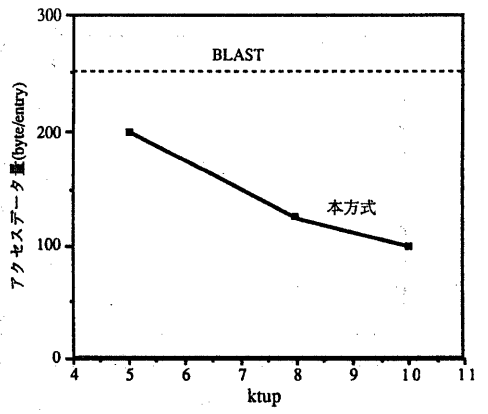


図5 アクセスデータ量

FASTAはデータベースをそのままアクセスしているため、一配列当たりのアクセスデータ量はデータベース中の配列の長さである1000バイトとなる。BLASTはあらかじめデータベース中の配列の各文字をACGTだけの4種類の文字に変換し、1バイトではなく2bitで表すことによって1/4に圧縮している。これによりBLASTのアクセスデータ量はデータベース中の配列の長さの1/4の250バイトとなっている。

本方式ではデータベース中の各配列に対して検索配列の分割した数だけスコア表にアクセスする。そこでスコア表の各値が1バイトとすると本方式のアクセスデータ量は1000/ktupバイトとなる。本方式のスコア表は大きいが、アクセスするのはごく一部でありアクセスデータ量は小さい。

図5に示すように本方式のアクセスデータ量はBLASTのアクセスデータ量よりも小さい。ただし本方式はBLASTと異なり完全なシーケンシャルアクセスではない。そこで本方式の検索時間はBLASTと同程度と考えられる。実際、ktupが8において本方式の検索時間はワークステーション(SUN SPARC station10)を用いてBLASTと同程度であった。

FASTAとBLASTではあらかじめ作成したハッシュ表とオートマトンを用いてデータベース全体をチェックするためアクセスデータ量はデータベース総配列長に依存し検索配列長にはあまり依存しない。一方、本方式でのアクセスデータ量は検索配列長に比例する。そのため検索配列長が図5で仮定した1000文字でなく、現在の配列決定装置の一般的な値である400文字の場合にはアクセスデータ量は減り、より検索時間は短くなると予想される。

4 スコア分布

GenBankから長さ1000の配列を選びだしスコア表を作成し、スコア表中の各値の分布を調べた結果、図6の様になった。

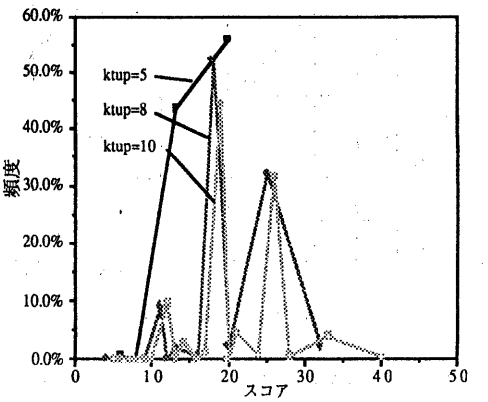


図6 スコア表の値の分布

挿入・欠失を含むアライメントが行われることは少なく、置換だけによる7点単位のピークの間に挿入・欠失の影響による低いピークが見られる。またスコア表の値の分布の中心は約20でありktupにあまり依存しないことがわかった。

次にスコア表の値の分布から本方式によるスコアの分布を求める。図7に長さ1000文字の配列の全体同士を比較した場合の本方式のスコアとDP法によるスコアの分布を示した。

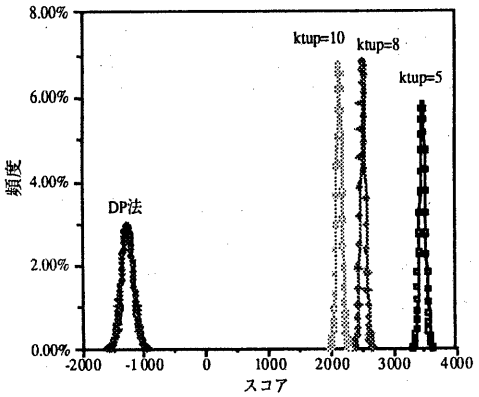


図7　スコア分布

配列の全体同士の比較のため完全に一致すれば4000点、完全に異なっていれば-3000点となる。DP法のスコアは置換によるスコアだけを考慮して二項分布で近似できる。そして本方式のスコアはスコア表の値を検索配列を分割した回数だけ加えたものであり，例えばktup=5の場合，図6の分布を200回加えた分布となる。

遺伝子配列比較で一般的な配列の部分同士の比較の場合、完全に一致すれば4000点だが、完全に異なっていてもスコアは負になることはない。部分同士の比較ではDP法と本方式のスコア分布も図7とは異なるが、図7を本方式の性能を見積るための目安とすることはできる。

本方式のスコアはDP法によるスコアの上限を与えるものであり，出来る限り低くDP法によるスコアに近い値が得られることが望ましい。図7からそのためには大きいktupを用いればよいことが分かる。これは大きいktupを用いると図6の分布の加算回数が減少し、それだけ低いスコアを得ることが出来るためである。また図2における不連続部分が少なくなり、それだけDP法のスコアに近づくためでもある。ktupは大きいほうがよいが、スコア表の大きさはktupに対して指数関数で増加するため、あまり大きいktupを用いることはできない。

図7から本方法で必要なktupの値を見積ることができる。例えば1000文字の中で20%の置換があった場合、DP法のスコアは2600である。ktupとして5を用いた場合、ほとんどの配列のスコアは3000以上であり、ふるい落とすことのできるスコアが2600以下の配列はほとんど存在しない。即ち図7から20%の置換に対して有効なふるいとするためには8以上のktupを用いなければならないことがわかる。

5　スコア比較

次に実際の遺伝子データベースにおいてDP法のスコアとの比較を行う。始めにBLASTのスコアとDP法のスコアを比較する。図8はGenBankから配列長500文字の配列を取り出して対を作成し、各対においてDP法のスコアとBLASTによるスコアをそれぞれx,y軸にプロットしたものである。完全に一致した場合、DP法のスコアは2000点、BLASTのスコアは2500点である。両スコアには相関関係があり、右下に分布が制限されているが、これではDP法のふるいとして用いることはできない。

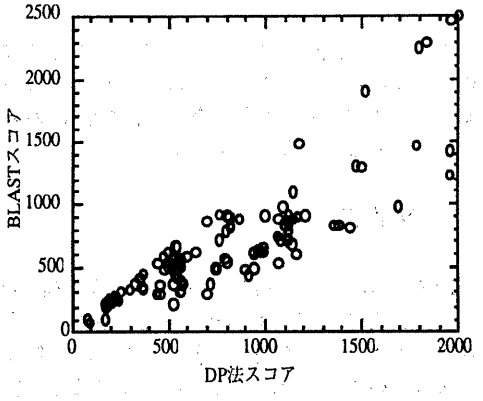


図8　BLASTとDP法のスコア比較

図8からDP法のスコアは高いがBLASTのスコアが低い対が存在することが分かる。特にDP法のスコアが高い場合に、ある特定のDP法のスコアに対するBLASTのスコアの範囲を特定することができない。このためBLASTをDP法の前段のふるいとして用いた場合には検索漏れが生じてしまう。FASTAのinitn、そしてBLASTでスコアと同様に用いられている確率も、DP法のスコアとの関係は同様であ

り、DP法のためのふるいとして用いた場合には検索漏れが生じてしまう。

図9は図8と同様に本方式のスコアとDP法のスコアを比較したものである。ktupには8を用いている。本方式のスコアは常にDP法のスコア以上の値となっており、検索漏れの生じないふるいとして用いることができることが分かる。例えばDP法で1800点以上の配列が必要な場合、本方式で1800点以上の配列のみDP法の計算を行えばよい。これによりDP法の計算量を減少させることができる。

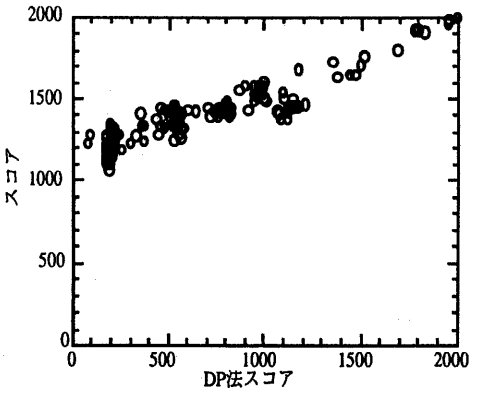


図9　本方式とDP法のスコア比較

さらに本方式のスコアはDP法のスコアとほぼ直線関係にある。そこで本方式のスコアはDP法のふるいとして用いるばかりでなく、直接DP法のスコアを推定することも可能である。

6　検索例

DP法、本方式、FASTA、BLASTによる検索結果の一例を表1に示す。検索配列としてグロビンの一種であるCHPAAAを用い、各方式のスコアをDP法のスコア順に示した。FASTAのスコアはinitnを用いた。また括弧内は各方式のスコアによる順位である。本方式ではktupとして8を用いている。

本方式、FASTA、BLAST共にDP法と順位は若干異なるが、DP法で上位の配列を漏れなく検索することに成功している。この例ではFASTA、BLASTに比べて本方式ではDP法による順位との違いは少ない。

表1　各方式による検索結果

エントリー名	DP法	本方式	FASTA	BLAST
CHPAAA	7272	7272(1)	7272(1)	9090(1)
CHPAAB	7211	7251(2)	5655(3)	7074(2)
GORAAC	7155	7188(3)	3976(11)	4920(4)
HSEGL1	7076	7146(4)	5365(7)	4841(5)
HUMHBB	7069	7139(5)	5358(8)	4841(5)
GIBCATAT	6800	6922(6)	5374(6)	4793(7)
ORAHBBE	6799	6859(7)	5391(4)	3143(8)
PPEGLOG	6799	6859(7)	5391(4)	3143(8)
MACAAD	6714	6852(9)	6061(2)	5659(3)
CEBGLOBIN	5643	6375(10)	4558(9)	1332(22)
LEMHBE	4747	5809(11)	4233(10)	1848(12)
GCREGLOB	3895	5480(15)	3213(12)	1217(27)
GCRHBEGEB	3888	5514(14)	2300(15)	1217(27)
TARHBEGPS	3770	5592(12)	2475(14)	1159(30)
TARHBE	3746	5571(13)	2549(13)	1159(30)

7　おわりに

スコア表を用いた高速なスコア計算方法を提案した。そしてそれをふるいとして用いることにより、計算量が極めて大きいDynamic Programming法による遺伝子配列検索が検索漏れなく、しかも高速に実現できる可能性を示した。

本方式実用化にはスコア表の圧縮が不可欠であり、今後検討を進めていく。また本報告ではDNAへの応用例を示したが、アミノ酸への応用も検討する予定である。

参考文献

1) T.F.Smith and M.S.Waterman;
Identification of common molecular subsequences
J.Mol.Biod., 147, 195(1981)

2) William R. Pearson;
Rapid and Sensitive Sequence Comparison with FASTP and FASTA
Methods in ENZYMOLOGY vol.183 p.63

3) Stephen F.Altschul et al.;
Basic Local Alignment Search Tool
J.Mol.Biod., 215, 403(1990)

4) T.Nishikawa et al.;
Rapid identity searching program for DNA sequences and its applications to cDNA grouping
Proc. Genome Informatics Workshop 1994, 194(1994)

5) 加藤、他;
大規模文書ﾃﾞｰﾀﾍﾞｰｽ用ﾃｷｽﾄｻｰﾁﾏｼﾝの開発
1991年情報学ｼﾝﾎﾟｼﾞｳﾑ予稿集, p.97(1991)