Vol.2013-MPS-96 No.8
Vol.2013-BIO-36 No.8
2013/12/11

Contigのk-mer coverage値の分布特徴量を用いたDouble assembly methodの提案

大城 絢子[†1]　岡崎 威生[†2]　名嘉村 盛和[†2]

概要：高速処理機能を搭載したギガシークエンサより獲得した読み取り配列は、多くの読み取りミスを含んでいる。読み取りミスの部位を除去し信頼性の高い結合配列を生成するために、Velvet、ABySS、SSAKEといった、k-mer を用いた de novo assembly method が多く提案されてきた。しかし用いるアセンブリ手法や k 値にアセンブリの結果が依存している事から、本研究では異なる従来手法、k 値により生成したcontig に対し、contig を生成する k-mer の coverage 値の分布情報をもとに生成した配列結合ルールを用いた double assembly を提案した。提案手法の検証には、評価値として出力 contig の正解率、被覆率を用いた。

キーワード：k-mer，p-value，Decision tree、ハイブリッドアセンブリ

Double assembly method with waveform characteristics of contig's p-value for k-mer's coverage

Abstract: Reads sequences from giga sequencer with parallel processing include many read errors. Various de novo assembly methods by use of k-mer has been proposed in order to remove read error region and derive accurate contigs, such as Velvet, ABySS, SSAKE and so on. But assembly result depends on assembly algorithm and k-value. We designed a double assembly method merging different k-mers and assembly methods with waveform characteristics of contig's p-value for k-mer's coverage. We evaluated it's performance with coverage ratio, correct ratio of output contigs.

1. 背景

高速処理を用いたギガシークエンサー [1] の開発により膨大なゲノム DNA 配列の獲得が可能になったことで、DNA 配列解析の研究がさらに盛んに行われるようになった。しかしその反面、シークエンサーによる読み取りミスが頻繁に発生し、特に未知ゲノムの再構築である de novo アセンブリが困難になっている。この問題を解決するために、k-mer を用いた de novo アセンブリ手法 [2] が多く提案されてきた。k 値は読み取り配列長より短く設けられる部分配列であり、読み取り配列に対して 1base ずつずらすことで生成され、読み取り配列データにおける出現頻度値を示す coverage 値とともにハッシュテーブルに格納される。coverage 値が著しく小さい k-mer については、シークエンサーによる読みとりミスと見なされハッシュテーブルから除去される。Velvet[3] や ABySS[4] では、各 k-mer をノードとした de Bruijn graph を生成し、$(k-1)$base の重複長をもつ k-mer 間にエッジを設け、Rockband[5] などの経路決定法を適用し contig を生成する。SSAKE[6] や VCAKE[7] では k-mer をノードとした prefix tree を隣接グラフとして適用し contig を生成している。IDBA[8] では用いる k 値を増加させることで重複部位の信頼性を維持しながら contig を生成している。SGA[9] においては k-mer の coverage 値情報を用いて、シークエンサーによる読み取りミスを含むと考えられる読み取り配列を除去し、読み取り配列をノードとした隣接グラフを適用することで配列結合の精度向上を試みている。また近年、複数の従来アセンブリ手法を組み合わせた、ハイブリッドアセンブリも多く提案されている。MAIA[10] や GAA[11] では従来のアセンブリ手法によって生成した contig らをノードとした隣接グラフを生成し、グラフ内における node 間の重複長の信頼

[†1] 現在，琉球大学大学院理工学研究科
[†2] 現在，琉球大学工学部情報工学科

性やノード間、つまり contig 間のアラインメントスコアを重みとすることでさらに長い contig を生成している。このように従来手法では、配列間または k-mer 間の重複情報を用いたアセンブリ手法がほとんどであるが、de novo アセンブリの実行において信頼性の高い contig を生成するのは困難である。

本研究では事前実験として、重複長と配列結合の信頼性の関連性を、5base 以上の重複部位を持つ配列による結合配列が元配列に含まれる (inconsistent) か否 (inconsistent) かの比較により図 1 のように示した。

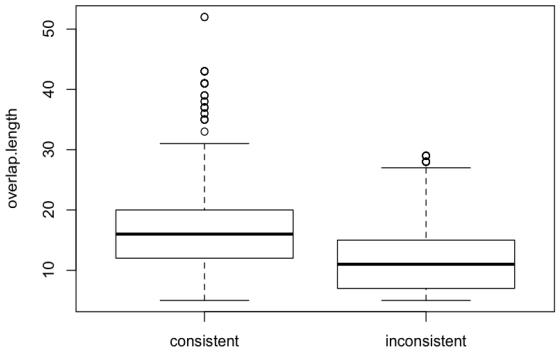

図 1 5base 以上の重複部位を有する結合配列と信頼性の関連性

図 1 より、重複長が 35base を超えると結合配列の信頼性はある程度保証されるが、重複長が 5-30base の範囲である場合、誤結合配列と正結合配列が多く混在していることがわかる。よって、配列間の重複長のみでは、配列結合の信頼性評価としては不十分であると言うことができる。

また用いるアセンブリ手法や k 値に、アセンブリの結果が依存していることも事前実験からわかっており、それらを統合することでさらに頑健なアセンブリの実行が可能であることもわかっている [12]。ここでは複数の従来手法や k 値の結果を統合する DAwH(Double Assembly method with Heuristic) を提案しているが、誤結合配列が多く発生している。それを防ぐために、De Bruijn より生成された contig が k-mer により構成されている点に着目し、各 contig を k-mer の coverage 値の集合と見なすことで、それらの特徴量を配列結合の決定に用いている。

本研究では、異なるアセンブリ手法や k 値を用いて生成した contig の統合に、contig を構成している、k-mer の coverage 値の分布特徴量より獲得した配列結合ルールを適用した、double assembly method を提案する。

2. k-mer の coverage 値の分布と contig の結合精度の関連性

前節でも述べたように、k-mer を用いたアセンブリ手法で生成された contig は、各 k-mer のもつ coverage 値の集合とみなせる。そこで本研究では、従来手法により生成された contig を構成する k-mer の coverage 値の分布と配列結合の精度の関連性について観察した。まず k-mer を用いた従来手法である Velvet や ABySS を用いて複数の k 値についてアセンブリを行い、生成した contig に対し、5base 以上の重複部位を持つような組み合わせの contig について、これらによる結合配列の正誤と各 contig を生成する k-mer の coverage 値の分布状況を観察した。実験には DNA 配列データベースである NCBI[13] の E.coli K-12 substr. MG1655 より 30000base の DNA 配列から読み取り配列をランダムに生成したものを用いた。

本研究では、シークエンサーによる読み取りエラーを含む配列に対して k-mer を用いた従来手法より生成した contig を double assembly として用いる。そのため、contig にはリードエラーを含む k-mer は含まれていないと仮定し、リードエラー無しの読み取り配列を、元配列よりランダムに生成した。表 1 にデータの特徴を示す。

表 1 検証用データの特性

Species	Length	read length	number of reads
Escherichia coli	30000	50	30000

配列データに対して k=15-30 について ABySS、Velvet で生成した contig を用いた。k-mer の coverage 値の分布と配列結合の正誤の関連性を観測するため、実験には正しい contig(元配列に含まれる) のみ使用した。また、k-mer の coverage 値は k 値に大きく依存するため、複数の k 値による contig の double assembly を行う際には coverage 値の正規化が必要であることから、(1) 式を用いて各 contig を構成する k-mer の coverage 値 ($c_{i,i=1,\ldots,n} \in C$) を p-value に変換した。

$$p_{c_i} = \frac{|\{c_i \in C | c^C \geq c_i|}{|C|} \quad (1)$$

本稿ではある 1 本の contig について 5base 以上の重複部位を有する複数の contig の、k-mer の coverage 値の p-value の分布状況と contig 結合の正誤について観測した結果を図 2〜図 5 に示す。

正結合配列または誤結合配列を生成するような組み合わせの contig の k-mer の p-value の分布状況の比較結果から、p-value の分布状況のパターンが類似しているがわかる。よって配列結合の正誤と contig のを構成する k-mer の coverage 値の分布には関連性があると言えることから、結合配列の特徴を学習した結合ルールを適用することで、配列結合の信頼性の改善が期待できる。そこで本研究では、従来手法によって生成した contig を構成する k-mer の coverage 値の p-value の分布情報をもとに配列結合ルールを獲得し double assembly に適用することで配列結合の信頼性の改善を試みる。

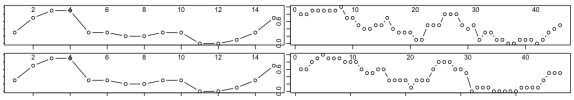

図 2 正結合配列を生成する, 各 contig の k-mer の coverge 値の p-value の変動状況 (前方 contig を固定)

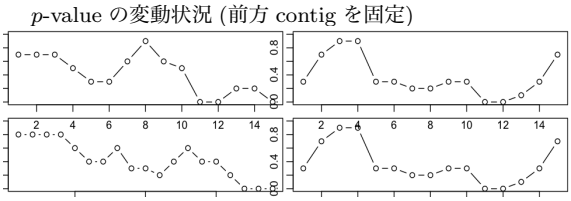

図 3 正結合配列を生成する, 各 contig の k-mer の coverge 値の p-value の変動状況 (後方 contig を固定)

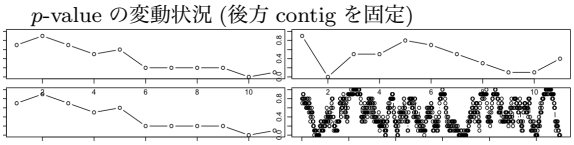

図 4 誤結合配列を生成する, 各 contig の k-mer の coverge 値の p-value の変動状況 (前方 contig を固定)

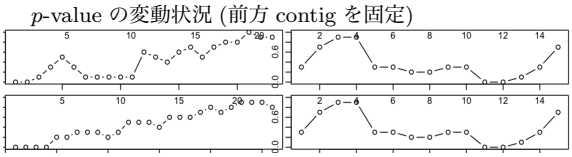

図 5 誤結合配列を生成する, 各 contig の k-mer の coverge 値の p-value の変動状況 (後方 contig を固定)

3. Contig の k-mer coverage 値の分布特徴量による配列結合ルールの獲得と、結合ルールを用いた Double assembly method

本研究の提案手法で用いる配列結合ルール獲得までの流れと、配列結合ルールを用いた提案手法の手続きについて述べる。所属が既知である過去のデータからある特徴や知識を抽出し、新たに観測されたデータ特徴から所属を推測する機械学習アルゴリズム [14] を取り入れた DNA 配列のアセンブリ手法はこれまでにいくつか提案されている。例えば Jeong ら [15] は k-mer の coverage 値の p-value を用いてシークエンサーによる読み取りミスを含む配列や contig を予測した。あらかじめ結合配列の正誤を識別できるルールを用いることで出力 contig の正解率を上げることが可能になる。C4.5[16][17] は教師つき学習アルゴリズムの一つで、あらかじめ入力された学習データの属性の特徴をもとに、属性を判別するための判別ルールを生成する。また判別ルールの生成に有効な特徴パラメータも重みつきで出力され、アセンブルの評価関数の定義に役立つことが期待されることから、本研究では既知配列から学習データを生成し特徴パラメータを定義し C4.5 に適用することで配列結合ルールを獲得した。

次に本研究の提案手法で用いた、配列結合ルール獲得のための特徴量について述べる。contig を構成する k-mer の coverage 値の p-value の分布情報を用いるために、本研究では、重複を有する組み合わせ contig のそれぞれ (former,latter) の「p-value の変動情報」、「p-value の頻度情報」、そして p-value の分布情報の「相関 (類似度)」に着目した。

はじめに本研究では、図 2-5 のように p-value の分布情報の波形表現が可能であることから、波形の周波数解析として一般的に多く利用されている、フーリエ変換 [18] を用いた。波形の急激な変位点を示す、フーリエ変換後の高周波成分、低周波成分、またフーリエ変換後の成分の総和を特徴パラメータに利用した。

また contig の各位置における p-value の増減状況は、p-value の変動状況を把握するうえで重要な情報となることから、p-value の変動情報を式 (2) のように 3 値に変換し、変換後の配列内の和を波形の勾配 $Coef_{wav}$ として式 (3) のように定義した。

$$q_{i(i=1,\ldots,n)} = \begin{cases} -1 & (i > i-1) \\ 0 & (i = i-1) \\ 1 & (i < i-1) \end{cases} \tag{2}$$

$$Coef_{wav} = \sum_{i=1}^{n} q_i \tag{3}$$

さらに、各 p-value 集合配列の要素数における増加値の割合も式 (2) を用いて (4) のように定義し特徴パラメータに追加した。

$$R_{inc} = \frac{|\{q_i \in Q | q_i = 1\}|}{|Q|} \tag{4}$$

次に、各 contig の p-value の集合配列における p-value の頻度情報も分布特徴量の定義として有効であると考え、特徴パラメータとして用いた。ここでは頻度値=0 である p-value の値に加え、p-value の度数分布図も波形データとして見なせることから、p-value の度数分布 ($range = 0.1$) のフーリエ変換後の要素の総和も用いた。

最後に、重複部位を有する contig 間の p-value の分布情報の相関 (類似度) は contig 結合の正誤に関連性があると考え、p-value の分布情報を波形として表現した場合の類似度を特徴量として用いた。類似度には、各組み合わせ contig の p-value の集合配列の相互相関関数の最大値 $M_{crossCorfun}$、p-value の度数分布の相関係数 $CorCoef$、ハミング距離 D_{HAM} を用いた。また前方 contig の末端と後方 contig の先端、つまり contig の重複部位の p-value のノルムも用いた。

これまで説明した特徴パラメータをまとめ、表 2 に示す。

以上より、既知配列より生成した学習データから上記で説明した特徴パラメータを定義し配列結合ルールを獲得し、試験データに適用することで配列結合の信頼性の改善を試みる。contig の k-mer coverage 値の分布特徴量による結合ルールを用いた、double assembly method の手続きを以下に述べる。

(1) 既知配列を用意する。

(2) 指定した read coverage、配列長の条件を満たすような

表 2 配列結合ルール獲得のための特徴パラメータ

変動	$Coef_{wav}^{f,l}$ 前後 contig の p-value 集合配列の波形の勾配
	$R_{inc}^{f,l}$ 前後 contig の増加率
	$F_{high}^{f,l}$ 前後 contig のフーリエ変換の高周波成分
	$Sum_F^{f,l}$ 前後 contig のフーリエ変換の総和
	$F_{low}^{f,l}$ 前後 contig のフーリエ変換の低周波成分
分布	$p_{null}^{f,l}$ 前後 contig の頻度値=0 である p-value
	$Sum_{Ffreq}^{f,l}$ 前後 contig の p-value 度数分布のフーリエ変換後の総和
	CC 前後 contig の p-value 集合配列の相関係数
相関	CC_{freq} 前後 contig の p-value 度数分布集合配列の相関係数
	CCF_{freq} 前後 contig の p-value 度数分布配列のフーリエ変換後の相関係数
	M_{ccf}^{F} 前後 contig の p-value 集合配列のフーリエ変換後の相互相関関数最大値
	M_{ccf} 前後 contig の p-value 集合配列の相互相関関数の最大値
	D_{ham} 前後 contig の p-value 度数分布配列のハミング距離
	M_{ccf}^{freq} 前後 contig の p-value 度数分布の相互相関関数の最大値
	$\|p_{f,l}\|$ 前方 contig の末端と後方 contig の先端の p-value のノルム

読み取り配列データセットをランダムに生成する。

(3) (2) で生成した読み取り配列データを用いて、複数の従来手法、k 値についてアセンブリを実行し contig を生成する。

(4) 5base 以上の重複部位を持つような組み合わせの contig を列挙する。

(5) 列挙した組み合わせ contig による結合配列を元配列と比較することで、正誤に分類する。

(6) 各組み合わせ contig について特徴パラメータを定義し学習データを生成する

(7) (6) で生成した学習データを C4.5 に適用し、結合ルールを獲得する。

(8) 試験用データである読み取り配列を用いて (3)(4) を実行する。

(9) (8) に結合ルールを適用し、ルールの条件を満たす結合配列を出力とする。

4. 有効性検証実験

前節で述べた学習データの特徴パラメータを decision tree である C4.5 に与え、獲得した配列結合ルールの有効性を確認するための検証実験を行った。まず結合ルールの学習データ自身への学習精度 $Le-R$ を表 3 から式のように定義した。

$$Le-R=\frac{C/C+W/W}{C/C+C/W+W/C+W/W} \quad (5)$$

C4.5 の出力結果においては correct ルール、wrong ルールが獲得されるため、correct ルールを満たす contig 組み合わせのみ出力した場合と、全組み合わせ contig より wrong ルールを満たす contig 組み合わせを取り除いた場合についての結果を示す。

5base 以上の重複部位を有する contig より結合配列を生成する場合と、それに対し結合ルールを適用した場合の出力 contig の変化を観察した。ルールの有効性を評価するために評価値として正解率 $Cor-R$ を、出力 contig 数 N_{output} と正結合配列数 $N_{correct}$ を用いて $N_{correct}/N_{output}$、被覆率 $Cov-R$ を、元配列長 L_{ref}、contig によって再現できた元配列の総数 $B_{correct}$ を用いて、$B_{correct}/L_{ref}$ のように定義した。

検証実験には表 1 と同様のデータを利用し、double assembly には 4 種類の k 値、2 種類の従来手法を用いた。結合ルール獲得のための学習データの生成、ルール適用のための試験データ生成には表 4 のような複数の k 値、従来アセンブリ手法の組み合わせによる double assembly を行った。

	学習データ		試験用データ	
手法	ABySS	Velvet	ABySS	Velvet
k 値	16,18	17,19	15,19	17,21

表 4 学習データ生成に用いた k 値と従来アセンブリ手法

学習データの decisopn tree への適用により、4 つの correct ルール、18 の wrong ルールを獲得した。ルールの内容の一部とルール生成に多く用いられた特徴パラメータを、引用率の高い順から表 5, 表 6 に示す。

表 5 獲得した結合ルール (一部)

rule1	F_{high}^{f} <= 3.3 -> class correct [0.990]
rule2	p_{null}^{f} <= 0.7 and p_{null}^{l} <= 0.2 F_{high}^{f} > 12 and F_{high}^{l} > 4.8 -> class correct [0.929]
rule5	Sum_F^{l} > 1983.868 and F_{high}^{f} > 3.3 F_{high}^{f} <= 3752.9 and F_{low}^{f} <= 0.7 -> class wrong [0.976]
rule6	M_{ccf}^{F} > 120142 and CCF_{freq} <= 911099.5 -> class wrong [0.969]

表 6 ルール獲得に引用された特徴パラメータ

99.29 % Sum_F^{l}	39.50 % F_{high}^{f}	22.05 % F_{high}^{l}	20.64 % p_{null}^{l}
18.75 % CCF_{freq}	14.15 % M_{ccf}^{F}	11.91 % $\|p_{f,l}\|$	5.54 % $F_{low}^{f,l}$

表 6 より、後方の contig の p-value 集合配列のフーリエ変換の総和、前方の contig の p-value 集合配列のフーリエ変換後の高周波成分がルールの生成に用いられたことがわかった。

次に、獲得した結合ルールの、学習データに対する判別結果と $Le-R$ を表 7 に示す。

	有効	無効
正結合	C/C 正結合配列に対して有効と判定した配列数	C/W 正結合配列に対して無効と判定した配列数
誤結合	W/C 誤結合配列に対して有効と判定した配列数	W/W 誤結合配列に対して無効と判定した配列数

表 3 識別結果のフォーム

	有効	無効	Le-R
正結合	531	6	0.971
誤結合	19	292	

表 7 結合ルールの学習データへの学習効果

表 7 より、一部の判別ミスはあるものの、ほとんどの結合配列に対して正しく分類できたことがわかる。そこで次に、実際ルールを適用し double assembly を行い、ルール適用前と比較した結果を、$Cor-R$、$Cov-R$ を用いて観測した結果を表 8 に示す。

手法	適用前	適用後 (correct)	適用後 (wrong)
出力数	848	586	392
正結合配列	537	376	370
誤結合配列	313	210	22
$Cor-R$	0.63	0.64	0.94
$Cov-R$	1.0	0.999	0.999

表 8 結合ルールの学習データへの学習効果

学習データに correct ルールを適用した場合、$Cov-R$ は 1 %減少したが $Cor-R$ は 1 %改善したのに対し、wrong ルールを適用した場合は $Cov-R$ は 1 %減少したものの、$Cor-R$ は 30 %の改善が確認できた。次に結合ルールを試験用データに適用した結果を、従来手法である Velvet、ABySS において各 k 値を用いた場合、結合ルールを適用しない double assembly(DAwH) と、ルールを適用した場合について比較した結果を表 9 に示す。

表 9 より、ABySS や Velvet の結果と比較した場合、$Cor-R$ は減少したが、$Cov-R$ が改善したことから、特定の k 値、手法に依存しないアセンブリが実行できたことが確認できる。また DAwH の結果と比較した場合、correct ルールを適用した際、$Cov-R$ を維持したまま $Cor-R$ は 3 %改善した。さらに wrong ルールを適用した場合は $Cov-R$ を維持しながら $Cor-R$ は 16 %の改善が確認できた。よって結合ルールの試験用データへの有効性が確認できた。また correct ルールを用いるより wrong ルールを適用し誤結合配列を排除した方が、$Cor-R$ の改善率が高いことも確認した。

5. まとめと課題

本研究では複数の従来アセンブリ手法、k-mer による double assembly に対し、配列結合ルールを適用した double assembly method を提案した。検証実験より、contig を構成する k-mer の coverage 値の分布特徴量を用いた結合ルールを適用し誤結合配列の除去することで配列結合の信頼性が改善したことを確認した。結合ルールの獲得には、contig を構成している k-mer の coverage 値の p-value 変換値集合配列のフーリエ変換後の総和、高周波成分が有効な特徴パラメータであることもわかった。

今後の課題として、用いるデータの元配列を変更して試験データを生成し結合ルールを適用した場合の有効性の検証を行い提案手法の頑健性を確認する必要があると考えている。さらにリードエラーを含んだデータについてもルールの有効性を確認する必要がある。

参考文献

[1] Lincoln D Stein: The case for cloud computing in genome informatics. Genome Biology 2010, 11:207.
[2] Miller JR, Koren S, Sutton G: Assembly algorithms for next-generation sequencing data. Genomics 2010, 95:315-327.
[3] Daniel R. Zerbino and Ewan Birney : Algorithms for de novo short read assembly using de Bruijn graphs, Genome Research, vol.18, pp.821- 829, (2008)
[4] ”Jared T. Simpson, kim Wong, Shaun D. Jackman, et al”ABySS: A parallel assembler for short read sequence data, Genome Research, vol.19, pp.1117- 1123, (2009)
[5] Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read de Novo Assembler
[6] Rene L. Warren , Granger G. Sutton , Steve J. M. Jones and Robert A. Holt : Assembling millions of short DNA sequences using SSAKE, Bioinformatics, vol.23 no.4, pp.500-501, (2007)
[7] William R. Jeck, Josephine A. ReinhardtDavid A. Baltrus, Matthew T. Hickenbotham, Vincent Magrini, Elaine R. Mardis, Jeffery L. Dangl and Corbin D. Jones:Extending assembly of short DNA sequences to handle error,BIOINFORMATICS APPLICATIONS NOTE,Vol. 23 no. 21, pp.2942—2944(2007)
[8] Yu Peng, Henry Leung, S.M. Yiu, Francis Y.L. Chin : ”IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler”, Research in Computational Molecular Biology,14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Volume 6044/2010, 426-440
[9] Jared T. Simpson and Richard Durbin : ”Efficient de novo assembly of large genomes using com- pressed data structures” Genome Res. 2012 22: 549-556 originally published online December 7, 2011
[10] “Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek,Jean-Marc Daran, Marcel Reinders, and Dick de Ridder” Integrating genome assemblies with MAIA” Vol.26 ECCB 2010, pp433—439
[11] “ Guohui Yao. Liang Ye, Hongyu Gao, Patrick Minx, Wesley C. Warren, George M. Weinstock ” Graph accordance of next- generation sequence assemblies, Vol. 28 no. 1, pp13-16.(2012)
[12] ”Ayako Ohshiro, Takeo OKAZAKI, Hitoshi AFUSO, Morikazu NAKAMURA”:A study of double assembly

手法	ABySS (k=15)	ABySS (k=19)	Velvet (k=17)	Velvet (k=21)	DAwH (A15,19+V17,21)	提案手法 (correct ルール)	提案手法 (wrong ルール)
出力数	66	38	13	9	597	558	402
正結合配列	66	38	13	9	387	374	325
誤結合配列	0	0	1.0	0	210	184	77
$Cor-R$	1.0	1.0	1.0	1.0	0.64	0.67	0.80
$Cov-R$	0.93	0.76	0.98	0.98	1.0	1.0	1.0

表 9 結合ルールの試験データへの学習効果

method for DNA sequences,IPSJ SIG- 33, 2013
[13] "Richard.C. Singleton":On computing the fast Forier Transform,Communications of the ACM Vol.10 pp647-654(1967)
[14] National Center for Biotechnology Information,http://www.ncbi.nlm.nih.gov/
[15] Thomas G. Dietterich:'Machine-Learning Research Four Current Directions' (AAAI) AI Magazine Volume 18 Number 4 (1997)
[16] Jeong-Hyeon Choi, Sun Kim, Haixu Tang, Justen Andrews, Don G. Gilbert and John K. Colbourne:' A machine- learning approach to combined evidence validation of genome assem- blies' Vol. 24 no. 6 , pp. 744-750,(2008)
[17] J. Ross Quinlan Morgan Kaufmann, San Mateo,CA: "C4.5:Programs for Machine Learning"(January 1993)
[18] Thomas G. Dietterich:'Machine-Learning Research Four Current Directions' (AAAI) AI Magazine Volume 18 Number 4 (1997)