DNA配列アセンブリの信頼性向上に関する研究

博士 (工学) 学位論文
Doctoral thesis of Engineering

DNA 配列アセンブリの信頼性向上に関する研究

A Study on accuracy improvement for DNA sequence assembly

2017 年 9 月 (September 2017)

大城 絢子 (**Ayako OHSHIRO**)

琉球大学大学院 理工学研究科
総合知能工学専攻

University of the Ryukyus
Graduate School of Engineering and Science
Interdisciplinary Intelligent Systems Engineering Course

本論文は、博士 (工学) の学位論文として適切であると認める。

論文審査会

岡崎 威生 印

(主　査)　岡崎　威生

名嘉村 盛和 印

(副　査)　名嘉村 盛和

遠藤 聡志 印

(副　査)　遠藤 聡志

– 要旨 –

高速並列読み取り処理を用いたギガシークエンサーの性能の飛躍的な向上と低コスト化により、膨大な DNA 配列データの高速な獲得が可能になったことで、読み取り配列であるリードの結合過程である DNA アセンブリの研究が世界中で盛んにされている。アセンブリにより生成された結合配列である contig はその後 DNA 配列解析に用いられるが、正しい配列解析には正しい配列データが前提であるため、アセンブリの精度は配列解析の分野においても最も重要な位置づけにある。

ギガシークエンサーの登場により 1 日あたりに獲得できる配列数は大幅に増えたものの、シークエンサーによる読み取りエラー部位への対応、正しいアセンブリの実行は現在でも大きな課題となっている。代表的な読み取りエラー除去法として k-mer があるが、用いる k の値や経路探索といった、手法によりアセンブリの結果が大きく異なるため、獲得した contig が本当に正しいのか、そして最適なアセンブリ実行の条件の決定が困難といった、頑健性と信頼性の改善が課題になっている。

本研究では、DNA 配列アセンブリにおける信頼性向上のために、以下の提案を行った。

- 複数の k-mer と複数の手法を組み合わせ k-mer の特徴量を用いたダブルアセンブリ
- 複合決定木による判別ルールを用いたダブルアセンブリ

有効性を検証するために、上記の提案手法と従来のアセンブリ手法との比較・従来の機械学習アルゴリズムとの性能比較を行った。検証の結果、第 1 の提案手法を用いることで従来手法では困難であった、長く正しい結合配列の生成、高い被覆率の獲得が可能になった。さらに第 2 の提案手法を用いることで、判別ルールの学習能力が向上し、従来の機械学習アルゴリズムによる判別ルールで対応できなかった結合配列への対応が可能になり、アセンブリとしての信頼性が向上する可能性が高いことがわかった。

– Abstract –

By drastically improving the performance of giga sequencer with the high-speed parallel read processing and cost reduction, it has became possible to derive massive DNA sequence data at high speed, so DNA sequence assembly studies are actively being around the world. Contigs, which is binded sequence by the assembly, is used for DNA sequence analysis. Because the correctness of the sequence analysis depends on the accurate sequence data, the accuracy of the assembly is the most important position in the field of sequence analysis.

The number of acquired read sequences per day has increased drastically with giga sequencer. But removal of read error region by the sequencer and correct assembly is difficult problem even now. K-mer has been applied much as a read error elimination method. Because assembly results depend on the k value and path search algorithm, it is difficult to determine k value and path search algorithm for the optimal assembly result. Therefore, improvement of robustness and accuracy are needed.

In this study, two method had been proposed for accuracy improvement of DNA sequence assembly.

- Double Assembly method with contig for k-mer's coverage
- Complex decision tree for getting the contig binding rules in double assembly

In order to verify the effectiveness of proposal, we had comparative experiments between traditional assembly method and traditional machine learning algorithm. From the verification results, first proposed method was shown that it became possible to generate longer and more high coverage which was difficult with the traditional assembly method. Furthermore, second proposed method was shown that the ability of discrimination rules were improver by using them. Therefore, it becomes possible to be deal with contigs which traditional machine learning could not, and it was found that there is a high possibility that accuracy of assembly is improved.

—研究業績—

学術雑誌など

1. Ayako OHSHIRO, Takeo OKAZAKI and Morikazu NAKAMURA, “Rule-based Assembly for Short-read Datasets Obtained with Multiple Assemblers and k-mer Sizes”, *IPSJ Transactions on Bioinformatics*, Vol.10, pp.9-15, 2017

2. Ayako OHSHIRO, Takeo OKAZAKI and Morikazu NAKAMURA, “Double assembly method with characteristics of k-mer’s coverage for contig”, *International Journal of Computer Science and Netowrk Security*, Vol.14, No.2, pp.54-61, 2014

国際学会における発表

1. Ayako OHSHIRO, Takeo OKAZAKI, Hitoshi AFUSO, and Morikazu NAKAMURA “Accuracy improvement for DNA assembly by multiple overlap processing”, Proceedings of International Technical Conference on Circuits/Systems Computers and Communications:F-M1-05, 2012

国内会議における発表

1. Ayako OHSHIRO, Hitoshi Afuso, Takeo OKAZAKI and Morikazu NAKAMURA, “Rule-based Assembly for Short Read Data Set obtained with Multiple Assemblers and k-mer Sizes”, 情報処理学会, バイオ情報学研究会研究報告, SIG-BIO-42, pp.1-8, 2015

2. Ayako OHSHIRO, Takeo OKAZAKI and Morikazu NAKAMURA, “A study on complex decision tree construction for getting the rules of contig binding in DNA double assembly”, 情報処理学会, バイオ情報学研究会研究報告, SIG-BIO-39, pp.1-7, 2014

3. 大城 絢子, 岡崎 威生、名嘉村 盛和, “Contig の k-mer coverage 値の分布特徴量を用いた Double assembly method の提案 “ 情報処理学会, バイオ情報学研究会研究報告, SIG-BIO-36, pp.1-6, 2013

4. Ayako OHSHIRO, Takeo OKAZAKI, Hitoshi AFUSO, and Morikazu NAKAMURA “A study of double assembly method for DNA sequences”, 情報処理学会, バイオ情報学研究会研究報告, SIG-BIO-33, pp.1-5, 2015

5. 大城 絢子, 岡崎 威生、安富祖 仁、名嘉村 盛和, “Short read シーケンサーデータに対する複次重複処理による結合信頼性向上の検討”, 情報処理学会, バイオ情報学研究会研究報告, SIG-BIO-27, pp.1-2, 2011

6. Ayako OHSHIRO, Takeo OKAZAKI, Hitoshi AFUSO, and Morikazu NAKAMURA “A study of DNA assembly algorithm using shortest common superstring problem”, 情報処理学会, バイオ情報学研究会研究報告, SIG-BIO-22, pp.1-5, 2010

謝辞

本論文は筆者が琉球大学大学院理工学研究科 総合知能工学専攻 博士後期課程に在籍中の研究成果をまとめたものです。本研究を遂行するにあたり、日頃より非常に多くの方々にお世話になりました。ここに改めて感謝の意を示します。

活動全般にわたり、研究に向かう姿勢、熱意、そして本テーマの実施の機会を与えていただき、常日頃から厳しくも温かい励ましと御指導を賜りました琉球大学工学部工学科 岡﨑 威生准教授に甚大なる敬意と感謝の意を示します。

学位論文審査会にて、本研究について貴重な御指導とご助言を頂きました琉球大学工学部 工学科 名嘉村 盛和教授、遠藤 聡志教授、常日頃より研究の進捗について声をかけてくださり、親しみをこめて接してくださいました琉球大学工学部 工学科 玉城史朗教授、学部時代より多大なるご指導をいただきました當間 愛晃准教授に厚く敬意と感謝の意を申し上げます。

本研究の遂行にあたり、研究室でともに研究を重ねた安富祖仁さん、Sonam Thsering さん、伊佐英寿さん、沼田翔平さん、青木史林さんには研究室でのセミナーにて貴重なご意見、ご協力をいただけたことが多々ありました。心からの感謝の意を表します。

博士課程の在籍中より研究に加え教育の機会を与えてくださり、研究・教育の両面においてご指導、ご教授くださった沖縄大学法経学科の金城秀樹准教授、八幡幸司准教授に心から感謝の意を示します。

勤務を継続しながら本研究の継続を快く許可くださった琉球大学大学院 医学研究科 臨床薬理学講座/臨床研究教育管理学講座 植田 真一郎 教授、日頃より研究の進捗について声をかけてくださった同講座の松下 明子助教、徳重 明央助教、医学部附属病院 臨床研究教育管理センター 池原 由美特命助教、亀田 美保特命助教、同講座、センターの職員のみなさまに心より感謝申し上げます。

最後に、在籍中にどのような状況においてもいつも温かく見守り応援してくれた家族、周りの方々、知花 健吾氏に対して深い感謝の意を表して謝辞とさせていただきます。

目 次

研究業績 **5**

謝辞 **7**

第 **1** 章　序論 **1**
1.1　研究背景 1
1.2　DNA の配列獲得 3
1.3　研究目的と提案 6
1.4　論文の構成 6

第 **2** 章　**DNA** 読み取り配列 **9**
2.1　DNA シークエンシング 9
2.2　シークエンサーによるリードの生成 11

第 **3** 章　アセンブリによる全配列の生成 **14**
3.1　アセンブリにおける隣接グラフ 14
3.1.1　Overlap Layout Graph 14
3.1.2　De Bruijn Graph 14
3.1.3　Prefix tree 16
3.2　経路探索法 17
3.2.1　ダイクストラ法 17
3.2.2　ウォーシャルフロイド法 17
3.2.3　二分探索木 18
3.2.4　ランダムウォーク 18
3.2.5　重み優先探索法 19
3.3　重み優先探索と機械学習アルゴリズムによる DNA アセンブリ 20
3.3.1　機械学習アルゴリズム 20
3.3.2　判別ルール獲得のための学習データの生成 21
3.3.3　各 Support Vector Machine の学習能力比較 22
3.3.4　重み優先探索と機械学習アルゴリズムによる DNA アセンブリ . . . 23
3.3.5　重み優先探索と機械学習アルゴリズムによる DNA アセンブリの性能評価 24
3.4　アセンブリの性能比較 24
3.4.1　k 値や手法のアセンブリへの影響 27

3.4.2 複数の k 値適用によるアセンブリ改善の可能性 32
3.4.3 複数の手法の適用によるアセンブリ改善の可能性 33
3.5 まとめ 33

第 4 章 k-mer の分布特徴量を用いた DNA ダブルアセンブリ **35**
4.1 複数の k-mer と手法を統合した DNA ダブルアセンブリ 36
4.2 ダブルアセンブリの有効性検証のための性能比較実験 36
4.3 決定木アルゴリズム 37
4.4 DNA ダブルアセンブリにおける判別ルールの獲得 40
4.4.1 contig 上の k-mer の coverage 値の分布と配列結合の正誤の関係 . . . 40
4.4.2 k-mer の coverage 値の分布特徴量 41
4.5 Contig の k-mer の分布特徴量を用いた DNA ダブルアセンブリ 44
4.6 DAwCC の有効性検証のための性能比較実験 44
4.6.1 性能比較に用いた評価指標 45
4.6.2 実験データと判別ルール 45
4.6.3 学習データを用いたアセンブリ結果へのルールの有用性 48
4.6.4 ターゲットデータを用いたアセンブリ結果へのルールの有用性 . . . 48
4.7 まとめ 49

第 5 章 複合決定木によるルールを用いた DNA ダブルアセンブリ **51**
5.1 複合決定木の評価指標の選択 51
5.1.1 複数の目的変数の適用 51
5.1.2 被覆最小値・重複長と結合正誤の関係 51
5.1.3 複数判別器の生成 53
5.1.4 複数の目的変数のルールによる正判別分布 56
5.1.5 複合決定木の生成 57
5.1.6 誤結合ルール適用による二段階判別の検討 58
5.2 複合決定木による判別ルールを用いた DNA ダブルアセンブリ 60
5.3 従来手法との性能比較実験 61
5.3.1 実験に用いたデータと k 値・手法 61
5.3.2 性能評価に用いる指標 61
5.3.3 誤結合ルールの選択 62
5.4 性能比較結果 65
5.5 まとめ 67

第 6 章 総括 **68**

参考文献 **70**

図目次

1.1 真核生物のゲノム DNA の構造の図 1
1.2 セントラルドグマ 2
1.3 生命情報工学における配列解析 4
1.4 読み取り配列から結合配列の獲得まで 5
1.5 読み取りミスを含むリードデータが引き起こすアセンブリの困難さ 5
1.6 本論文の概要 8

2.1 DNA が PCR による複製・断片化されるまでの流れ 9
2.2 PCR 法による複製 10
2.3 蛍光バンドの検出による配列決定 11
2.4 ショットガン法による配列決定 11

3.1 アセンブリにおける Overlap Layout Graph 15
3.2 アセンブリにおける De Bruijn Graph 16
3.3 アセンブリにおける Prefix tree 16
3.4 k 値や手法の影響 (正解率 corR) 27
3.5 k 値や手法の影響 (被覆率 covR) 27
3.6 k 値や手法の影響 (最長正結合配列長) 28
3.7 k 値や手法の影響 (最長結合配列長) 28
3.8 Velvet にて各 k 値による contig の復元領域 30
3.9 ABySS にて各 k 値による contig の復元領域 31

4.1 重複長と結合の信頼性の分布 38
4.2 k-mer より構成される Contig 40
4.3 正結合における各 contig の k-mer の変動状況 (前方固定) 42
4.4 正結合における各 contig の k-mer の変動状況 (後方固定) 42
4.5 誤結合における各 contig の k-mer の変動状況 (前方固定) 42
4.6 誤結合における各 contig の k-mer の変動状況 (後方固定) 42
4.7 判別ルールを適用したダブルアセンブリの流れ 45

5.1 被覆最小値と結合正誤の分布 52
5.2 重複長と結合正誤の分布 52
5.3 正結合ルール による正判別分布 57
5.4 誤結合ルール による正判別分布 57

5.5 複合決定木による判別ルールを用いた DNA ダブルアセンブリの手続き . . 61
5.6 ダブルアセンブリにより生成された contig 長と正誤の分布 62

表 目 次

2.1 これまで開発されてきたシークエンサー 12

3.1 識別結果のフォーム 22
3.2 各 SVM algorithm の学習能力 23
3.3 断片配列数が 300 本の時の判別精度 25
3.4 断片配列数が 900 本の時の判別精度 26
3.5 k-value がアセンブリに与える影響 (VELVET) 28
3.6 k-value がアセンブリに与える影響 (ABySS) 28
3.7 複数の k 値利用時:Velvet 32
3.8 複数の k 値利用時:ABySS 33
3.9 Velvet, ABySS の複数の k 値利用時のアセンブリの性能 33

4.1 検証用データの特性 36
4.2 k 値と手法の組合せ 37
4.3 DAwH と従来手法の性能比較 37
4.4 前後 contig の k-mer の特徴量を用いた説明変数 43
4.5 学習データ生成に用いた k 値と従来アセンブリ手法 46
4.6 判別ルールの学習データへの学習能力 46
4.7 獲得した結合ルール 47
4.8 ルール獲得に引用された特徴パラメータ 48
4.9 学習データのダブルアセンブリへのルールの効果 48
4.10 結合ルールの試験データへの学習効果 49

5.1 各説明変数間のケンドールの順位相関係数 54
5.2 被覆最小値 、 重複長の判別式と判別能力 55
5.3 重複長、被覆最小値による判別ルールと判別能力 56
5.4 被覆最小値の判別能力 56
5.5 重複長の判別能力 56
5.6 各誤結合ルールを追加した場合のアセンブリ性能の変化 59
5.7 学習データ及び試験データ生成に用いた k 値と従来アセンブリ手法 62
5.8 複合決定木より獲得された正結合ルール一覧 63
5.9 複合決定木より獲得された誤結合ルール一覧 64
5.10 長い誤結合配列を除去できた判別ルール 65

5.11 従来手法と複合決定木による判別ルールを用いたダブルアセンブリの性能比較 66

第1章　序論

1.1　研究背景

生命は遺伝子によって決定されており、生命の維持をはじめとした生命活動には遺伝子が前提となっている。つまり遺伝子の情報を明らかにすることであらゆる生命情報を明らかにすることが可能になる。

遺伝子は、化学的にみるとDNAと呼ばれるデオキシリボ核酸(Deoxyribose Nucleic Acid)という物質により構成されており、遺伝情報を担うDNAとその情報に基づいて生成されるタンパク質によって維持される。DNAはアデニン(Adenine)、チミン(Thymin)、グアニン(Guanin)、シトシン(Cytosin)から成るヌクレオチドとよばれる生体物質が集まってできたもので、生物の細胞や組織を作り上げるのに必要なすべての情報が蓄えられており、親から子へと世代を通じて伝えられていく。

1953年にJ.D.Watsonら[1]により、DNAは右巻きの二重らせん構造であること、図1.1のように一つ一つの核の中に納められており、らせん構造はアデニンとチミン、シトシンとグアニンの対によって構成されていることが明らかになった。

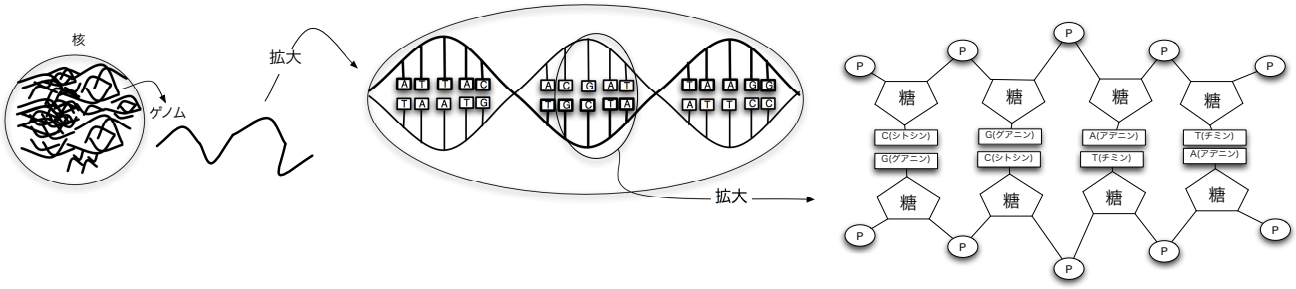

図 1.1: 真核生物のゲノムDNAの構造の図

タンパク質生成にむけて、らせん構造は核内にあるタンパク質によりほどかれ、次に複製機能をもつ、タンパク質集合の酵素であるRNAポリメラーゼにより塩基の転写が行われる。図1.2のように、転写されたDNAはメッセンジャーRNAと呼ばれる。メッセンジャーRNAはリボソームによりアミノ酸に変換され、最後に立体構造に折り畳まれることによりタンパク質となる。

DNAの情報からmRNAを経てタンパク質が生成されるまでの一連の流れのことをセントラルドグマと呼ぶ。

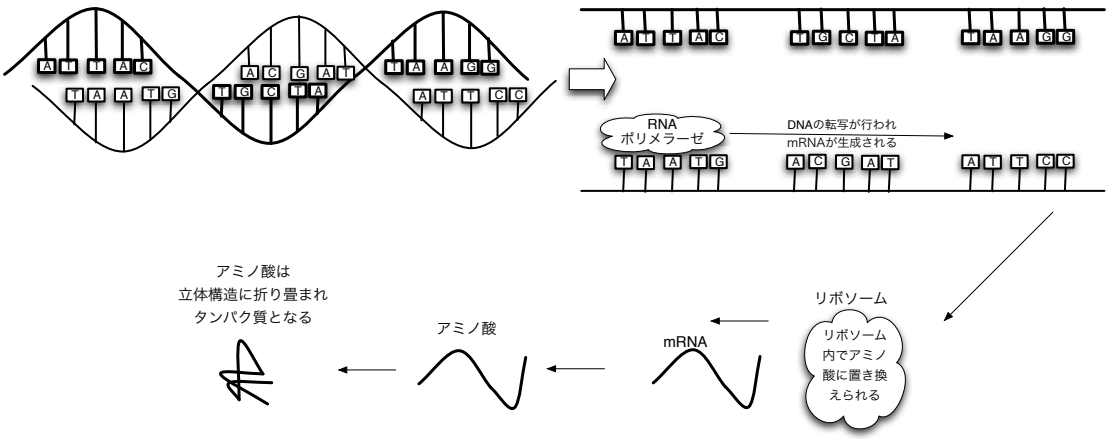

図 1.2: セントラルドグマ

Watson らにより DNA の二重らせん構造が発見されたことで、分子生物学の分野が確立されはじめた。1975 年にフレデリック・サンガーによるサンガー法、ウォルター・ギルバートによるマクサムギルバート法による、DNA 塩基配列の解読手法が発表されたことで、ヒトの遺伝子を明らかにする「ヒトゲノム計画」の概要が 1985 年に発表され、1990 年に正式にスタートした。15 年計画であったヒトゲノム計画では、当初の予定より 2 年早い 2003 年にはヒトの全遺伝子の 99 %の配列情報が 99 %の精度で明らかにされた [2]。この成果は世界中で大きな話題を呼び、遺伝による病気の診断といった研究への貢献が期待された。当時ゲノムの解読にかかった費用は 3000 億円だったといわれていた。実はこのような形での解読の成功にはコンピュータ技術の大きな進歩が貢献しており、ヒトゲノム解読の計画と同時に、DNA の読み取り技術も確実に発展を遂げている。解読技術が開発された当時は、読み取りは手動で行われており、1980 年代には 1000 塩基の配列の読み取りに 3 日を費やすというスピードであった。ヒトゲノム解読の成功にむけ、蛍光させた塩基をレーザーで高速に読み取りサンガー法を自動化させた DNA 蛍光シークエンサーが 1998 年に開発されたことで、DNA 配列のデジタルデータの獲得が可能になった。

このように DNA 塩基配列の読み取りが自動化され、テキストによる配列の獲得が可能になってから、ゲノム領域における、コンピュータを利用した配列解析の研究が盛んになった。複数の遺伝子配列データ比較による相同性の発見、アミノ酸配列間の類似関係を算出する配列アラインメント [3] や特定の遺伝子の体内における機能を決定する遺伝子機能推定、他生物種との類縁関係をまとめる進化系藤樹、大規模なアミノ酸配列に共通する部位を特定するための、アミノ酸モチーフ検出 [8] などが挙げられる。これらの配列解析の結果は、今日のがん遺伝子の特定や個々人の遺伝子に合った治療薬の適量を知ることを可能にし、テーラーメイド治療の発展に貢献している。このようにして、解析技術により生命のもつ情報を明らかにする研究分野が誕生した。医学や薬学工学、統計学などの分野の統合による生物情報工学の分野は「バイオインフォマティクス」と呼ばれており、数学者、計算科学者、統計学者が、医療分野などの情報管理をする為の技術を発達させ、農業、薬理学、その他の商業的応用の観点から、ゲノムに関する生物学的データを組織化す

るのに役立っている。

一方で、2003 年に終了したヒトゲノム計画に続き、ゲノムの解読技術つまり DNA のシークエンス技術も大きな発展を遂げている。2007 年にはアメリカのジェームス・ワトソン博士らにより、1 億 2000 万という費用で、2 ヶ月でヒト一人のゲノム解読を完了させた。ヒトゲノム計画と比べると、読み取りにかかる時間と費用が大幅に改善されたことがわかる。現在では 1 ヶ月の期間と数百万の費用でヒトゲノムの解読が可能になり、今後数年のうちで、個人ゲノムを 1 時間程の時間と数万円の費用で解読することが可能な時代が到来すると言われている。ゲノム解読技術の誕生時に比べると、現在の技術は 1000 倍程の性能を持っていると言える。ゲノム解読の技術の高速・低コスト化を目指した開発により、ヒトゲノムの解読を皮切りにマウスやチンパンジー、さらに酵母菌や古細菌まで、幅広い生物種のゲノム解読がなされ、あらゆる生物情報がデジタル化され、蓄積されてきた。

このような、わずか数年間のゲノム読み取り技術の急速な発展の背景には、シークエンサーの性能向上が大きく関係している。2005 年には、高速処理を搭載したゲノム配列の読み取り装置であるギガシークエンサーが開発された。従来のシークエンサーとは異なり、読み取り長を短くし、読み取りの並列化を可能にしたことで、短期間での大規模なゲノム配列のデータ蓄積が可能になり、配列解析の研究は従来にも増してさらに盛んに行われるようになった 。例えば 2007 年にはアカゲザルの全ゲノム解読が終了し、ゲノム解析の結果、ヒトやチンパンジーとの遺伝情報の違いが 1-2.5 %であったことが発表された。2010 年にはキンカチョウのゲノム解読が完了し言語の起源解明への大きなヒントの獲得に繋がった。現在までに 6366 種類のゲノム解読が完了していると報告されている。ただしこれらの成果はゲノム解読つまり生命の設計図が明らかにされたということであり、各部位の働きについてはほとんど知られておらず、それらの機能を明らかにすることが課題である。

1.2 DNA の配列獲得

計算機を用いた配列解析には配列データが必要であるが、本節では DNA の配列がデータとして獲得されるまでの過程と現在抱えている課題について概説する。

読み取り装置である DNA シークエンサーは読み取り可能な長さに上限があり、一度に全配列を読み取ることはできない。そこでシークエンシングの前処理として、DNA 配列を複製しシークエンサーの読み取りが可能な長さになるまで断片化がおこなわれる。シークエンサーへは断片化された配列が入力され、読み取りがおこなわれる。シークエンサーからは読み取り配列がリードデータとしてテキスト形式で出力され、リード間の重複情報にもとづいて結合される。リードをノード、重複情報は重みつきエッジとして隣接グラフで表現され、グラフ上で最適経路の探索がおこなわれる。決定した経路に従い結合配列が生成される。獲得された結合配列は contig と呼ばれており、リードから contig 獲得までの過程はアセンブリ [41][42] と呼ばれている。アセンブリの過程を経て生成された配列データは、図1.3のように配列解析などの分野で利用されていく。

高速並列分散処理機能を搭載したギガシークエンサーの登場により、ゲノム配列のシー

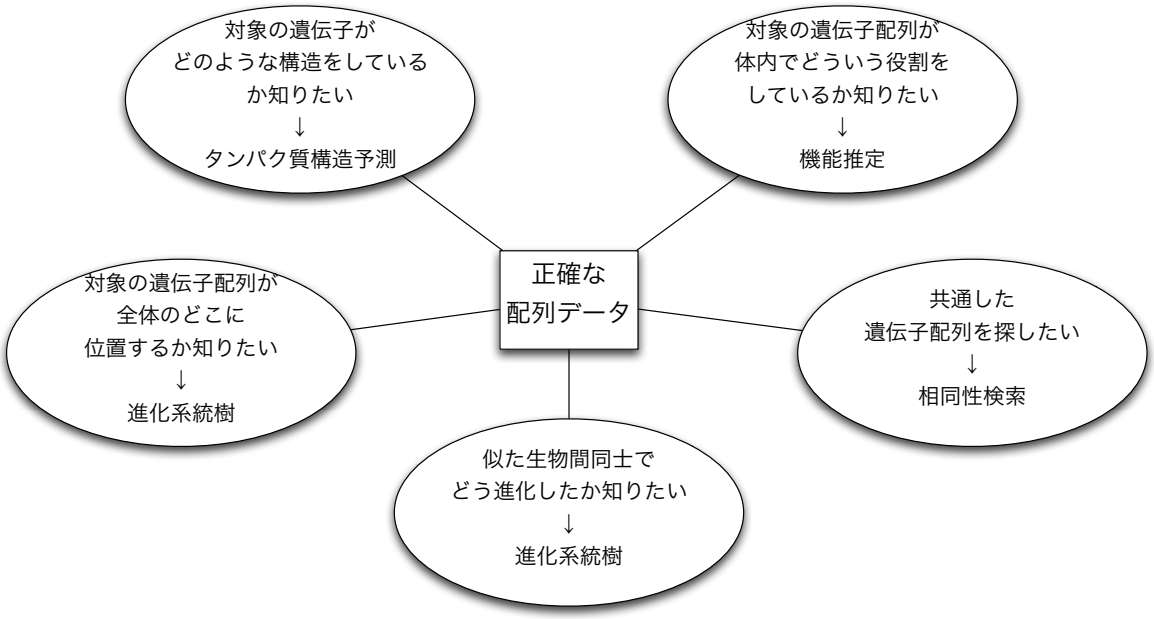

図 1.3: 生命情報工学における配列解析

クエンシング技術は飛躍的に発展した。従来のシークエンシング技術と異なり、複製回数を増やし断片化したリードの並列処理をおこなうため、短時間で膨大な DNA 配列データの獲得が可能になった。しかし高速での並列読み取りが可能な反面、読み取りエラーが多くリード長が短いため、獲得されるリードの精度が低いという特徴がある。そのため図 1.5のように、リードを直接扱うような従来のアセンブリ手法を用いた場合の信頼性の高い contig の獲得は困難になっている。

ギガシークエンサーを用いたリード獲得が主流になって以来、リード上における読み取りエラー除去の処理が必要になってきた。その代表的な手法が、k-mer を用いたエラー除去法である。k-mer は k base の部分配列であり、k の値は読み取り配列長より短く設定され、k-mer はリードに対して 1base ずつシフトさせることで生成される。リードデータ中における出現頻度の値とともに、ハッシュテーブルに格納される。ギガシークエンサーによるリードは、元配列の複製回数が特に大きいことが特徴である。そのため、読み取りエラーが生じた場合、生成された k-mer の出現頻度の値は著しく小さくなることが予想されることから、読み取りエラー部位の除去は、k-mer の出現頻度の値に基づいて行われる。k の値によってハッシュテーブルの構造や出現頻度の値の分布も変化するため、アセンブリの結果は k の値に依存するという特徴も持つ。k-mer を用いた読み取りエラーの除去の手続き・アセンブリについては 3 章で詳しく述べる。

k-mer に限らずこれまで多くの読み取りエラー除去法・隣接グラフの生成法・グラフ上における経路探索法が提案され、様々な観点によるアセンブリ性能評価指標も定義され比較されてきた。それに伴い手法やアセンブリ実行の初期値にアセンブリ結果が左右されるため、最適なアセンブリの実行が困難になるという問題も生じてきている。正しく解析す

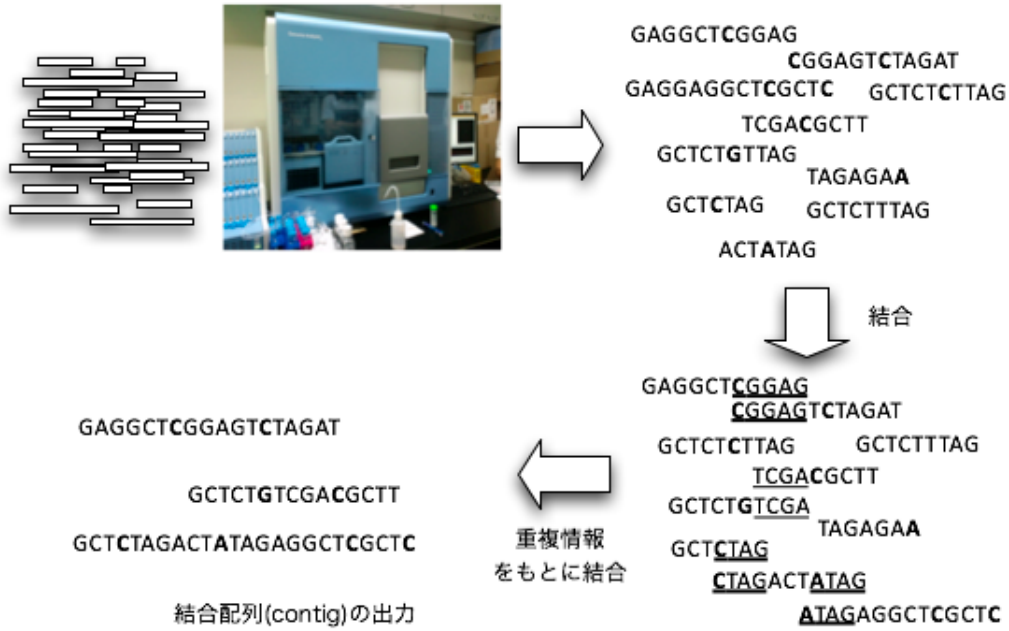

図 1.4: 読み取り配列から結合配列の獲得まで

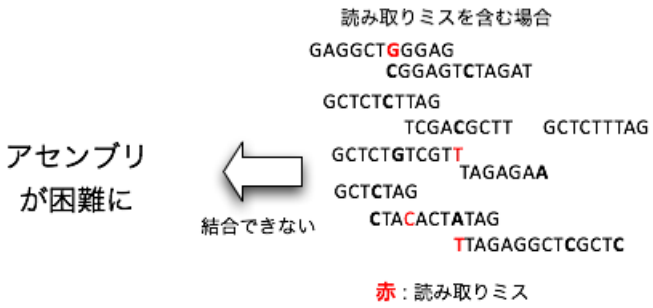

図 1.5: 読み取りミスを含むリードデータが引き起こすアセンブリの困難さ

るには正しい配列データが前提であるとともに、信頼性の高いアセンブリによる DNA 配列の生成は、バイオインフォマティクスの分野において最も重要な位置づけにある。

1.3 研究目的と提案

本研究では、精度の高い DNA 配列を生成するための「信頼性が高い DNA 配列アセンブリ」のために以下の手法を提案した。

1. k-mer の特徴量を用いて獲得した判別ルールを適用し、複数の k-mer と手法を組み合わせた DNA 配列ダブルアセンブリ
2. 複合決定木による判別ルールを用いた DNA 配列ダブルアセンブリ

提案手法の有効性を検証するために、1 の手法では元配列の再現率・contig の正解率について従来の複数のアセンブリ手法と比較した。2 の手法では判別ルールの性能つまり学習能力の向上という目的のもと判別ルールの生成法を提案し、従来のアセンブリのみならず、従来の機械学習アルゴリズムにより生成される判別ルールとの学習能力の比較をおこなった。

1.4 論文の構成

本論文は、6 つの章から構成されている。本章では研究背景と、シークエンサーより獲得した読み取り配列であるリードから、DNA アセンブリによって生成される結合配列である contig の獲得までについて簡単に述べた。

第 2 章では従来の代表的な DNA シークエンシング技術について述べ、従来のシークエンサーの性能とシークエンサーによるリード獲得までの流れについて説明した。

第 3 章ではリードから結合配列である contig 生成までの過程と、従来のアセンブリ手法の性能比較、それらの特徴と課題について述べた。従来のシークエンサーデータと、読み取りエラーを含むギガシークエンサーデータに対しての各隣接グラフの構成を述べ、主に読み取りエラーを除去する k-mer について紹介した。さらに各データの特徴に応じたグラフ上にて実施される経路探索を紹介した。次に、過去に提案した重み優先探索と、機械学習アルゴリズムより獲得した判別ルールを適用したアセンブリ手法について述べた。獲得した判別器と配列結合に設けた結合下限値といった複数のアプローチにより、アセンブリの信頼性を維持した。検証実験では結合下限値や読み取り配列数といったパラメータのチューニングをおこなった。最後に k-mer を用いた従来のアセンブリに対して、k 値の与え方や手法がアセンブリに与える影響を観察し、その特徴について議論した。

第 4 章では contig の特徴量を用いた、複数の k-mer と手法の統合による DNA アセンブリ手法について述べた。先述のとおり k-mer を用いた従来のアセンブリ手法には、用いる k の値や手法といった実行の初期値にアセンブリ結果が依存するという問題、さらに各 k の値や手法のアセンブリ実行結果が相補的になっているという第 3 章での検証実験の結果

を踏まえ、はじめに複数の従来のアセンブリ手法の統合によるヒューリスティックなDNAダブルアセンブリにより頑健性の向上を目指した。さらに配列結合の信頼性維持のために、結合配列であるcontigの特徴量を用いて結合正誤の判別ルールを獲得した。k-merを用いる従来手法によって生成されたcontig上の、k-merの分布に着目した場合、分布の特徴と配列結合の信頼性に関連があることを仮説検証実験にて明らかにした。そこで周波数解析を用いた説明変数を定義し学習データ生成に用いた。判別ルールはC4.5により獲得しダブルアセンブリに適用した。検証実験では従来手法との性能比較をおこない、元配列の被覆率や正しいcontig長、正解率といったアセンブリ性能の改善の可能性について議論した。

第5章では、第4章にて提案したダブルアセンブリ手法に適用した判別ルールの性能改善のための、複合決定木の生成法について述べた。従来のアセンブリ手法にて、配列結合の指標として用いられることの多かった特徴量である「リード間の重複長」「contig上のk-merの被覆値であるcoverageの最小値」と結合正誤との分布関係を観察し、これらの目的変数としての利用を検討した。目的関数を1つとする従来の決定木アルゴリズムに対し、結合配列の特徴量を複数用いて目的変数とし追加することで、配列結合に対する判別基準の要素を増加させ判別能力を改善する可能性を検証した。また量的変数であるこれらの特徴量を目的変数として扱うため、決定木アルゴリズムを用いて結合の正誤へ質的変換をおこなった。さらに第4章では配列結合の「正誤」の各ルールをアセンブリへ適用したのに対し、正誤2値の両ルールの2段階適用による判別能力の可能性について検討した。以上の結果を踏まえ、結合配列の判別ルールの性能改善というアプローチよりアセンブリの性能改善を目的とし、複合決定木による判別ルールを用いたDNAアセンブリアルゴリズムを提案した。従来のアセンブリ・従来の決定木アルゴリズムとの性能比較実験の結果より、複数の目的変数より決定木を構成する「複合決定木」からは、従来の決定木に比べ多くのルールが獲得され、これまで正しく判別できなかった結合配列に対しても対応が可能になったことが示された。さらに正誤についてのルールを2段階に用いることで、両ルールの正判別不可であった結合配列へも対応が可能になり、判別ルールとしての学習能力が改善された。これらを適用したダブルアセンブリに関しても、従来のアセンブリに比べ長い結合配列群における被覆率や正解率といった性能が改善されたことが示された。

最後に本論文における提案の成果と結論、期待される効果について述べた。

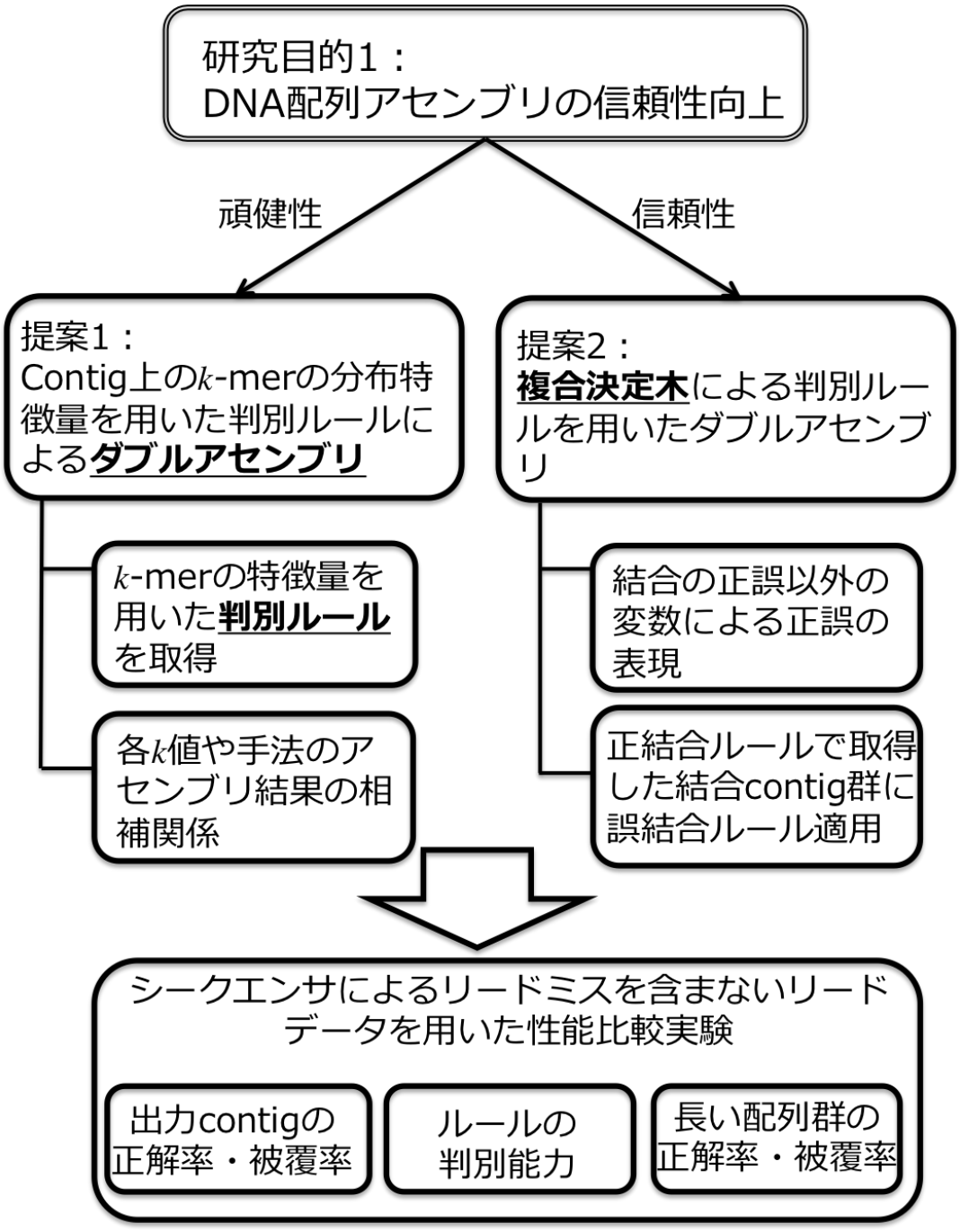
研究目的1：
DNA配列アセンブリの信頼性向上
頑健性
信頼性
提案1：
Contig上のk-merの分布特徴量を用いた判別ルールによるダブルアセンブリ
提案2：
複合決定木による判別ルールを用いたダブルアセンブリ
k-merの特徴量を用いた判別ルールを取得
各k値や手法のアセンブリ結果の相補関係
結合の正誤以外の変数による正誤の表現
正結合ルールで取得した結合contig群に誤結合ルール適用
シークエンサによるリードミスを含まないリードデータを用いた性能比較実験
出力contigの正解率・被覆率
ルールの判別能力
長い配列群の正解率・被覆率

図 1.6: 本論文の概要

第2章　DNA読み取り配列

ゲノム解読に必須のDNAシークエンシングから、DNA配列アセンブリの入力に必要な読み取り配列データであるリード獲得までの過程について、従来手法の推移とともに概説する。

2.1　DNAシークエンシング

DNA配列はDNAシークエンシングと呼ばれる、読み取り処理による塩基配列の決定が基本手段となっている。DNAシークエンシングは一般的に、特殊な薬品による対象個体の細胞採取により開始される。図2.1のように加熱によりDNAの二重らせん構造が解かれたDNAはその後、人工的に合成した短いプライマーを用いてDNAポリメラーゼにより大量複製される。

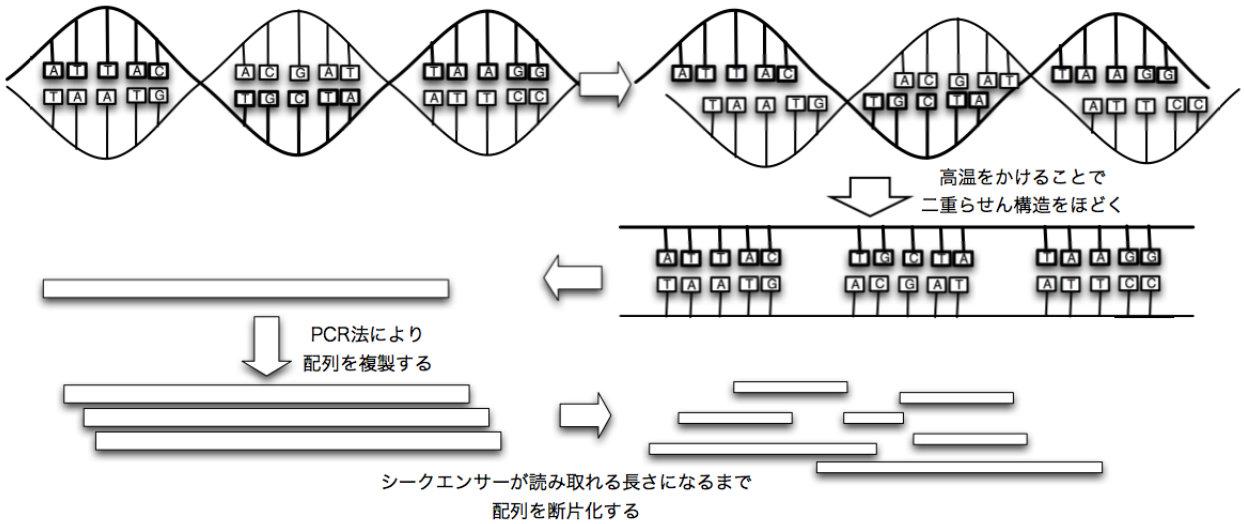

図 2.1: DNAがPCRによる複製・断片化されるまでの流れ

1. DNAに高温をかけ二重らせん構造をほどく。
2. PCR法により増幅する。
3. シークエンサーが読み取り可能な長さになるまで断片化する。

この複製技術は PCR(polymerase chain reaction) 法と呼ばれ 1983 年に開発されたものであるが、塩基配列決定の他に短い DNA 配列の多重複製などにも用いられる。図2.2に簡単な流れを示す。

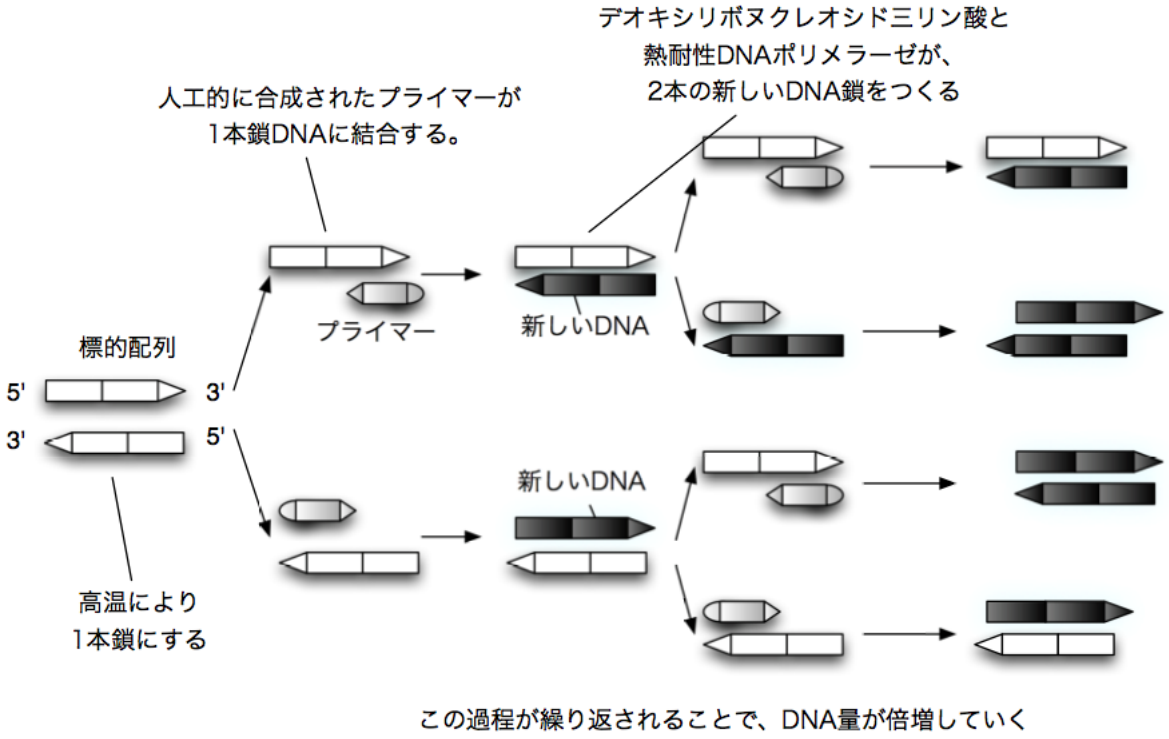

図 2.2: PCR 法による複製

多くの遺伝情報を明らかにするため、つまり多くの DNA をより速く読み取るために、これまで多くのシークエンシング技術が開発されてきた。はじめて DNA 配列の読み取りを可能にしたのは、ウォルター・ギルバートとフレデリック・サンガーが中心になって 1975 年に開発された Sanger 法 [5] であった。サンガー法は PCR 法と電気泳動法を組み合わせた手法で、はじめに配列を特定したい鋳型 DNA とプライマー、既知配列であるデオキシヌクレオチド (dNTP) と、あらじめ蛍光標識された既知配列であるジデオキシヌクレオチド (ddNTP) を用意する。PCR 法を用いて伸長された対象の配列は、ゲルを用いた電気泳動法により決定される。電気泳動法は荷電粒子が反対極に移動する原理を用いたもので、ジデオキシヌクレオチドとの反応をもとに配列を決定する。

1977 年にウォルター・ギルバートが開発した Maxm-Gilbert 法 [6] は DNA の化学分解を用いた手法である。対象の DNA 配列中の特定の塩基を ^{32}P や ^{33}P で放射標識し任意の塩基で科学的に切断し、サンガー法同様電気泳動法により塩基配列を検出する。

ショットガン法は、制限酵素により断片化された DNA 配列であるリード間の重複部位をアラインメントにより結合し結合配列を生成する手法であり、ヒトのような巨大な配列決定の高速化に貢献した。図2.4のように、次世代シークエンス法では画像処理技術により高密度でショットガン法をおこなうようになった。

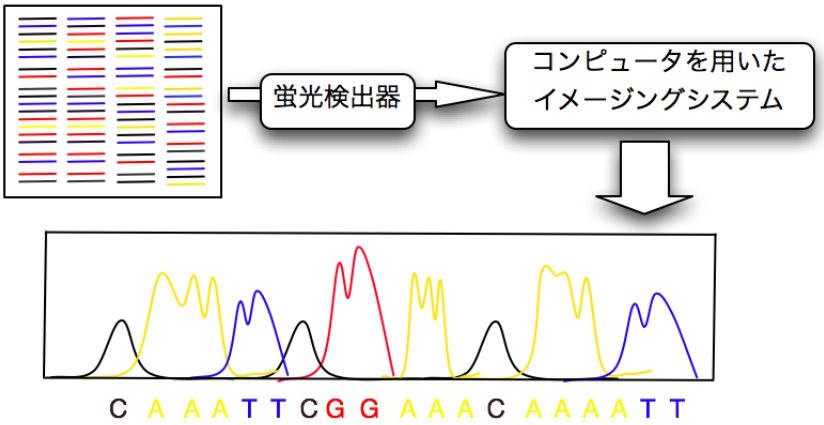

図 2.3: 蛍光バンドの検出による配列決定

波長の異なる蛍光発色による標識をおこない、レーザを当てた蛍光バンドの検出により、図2.3のように波形のピークに対してコンピュータによる配列決定の自動化をおこなう。

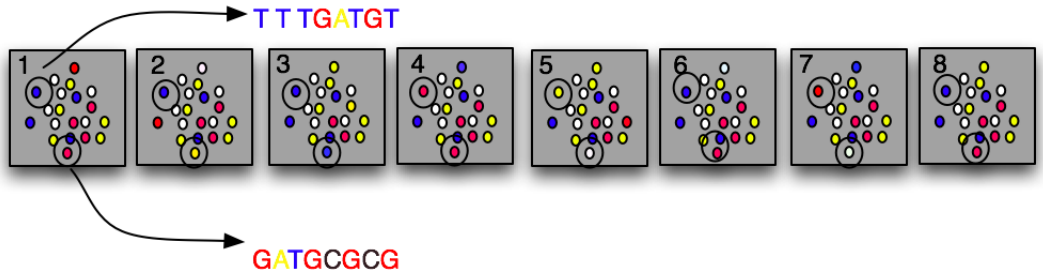

図 2.4: ショットガン法による配列決定

遺伝子地図を用いて配列を結合する階層化ショットガン法と、計算機を用いて各リードを結合していく全ゲノムショットガン法がある。精度は高いが読み取りに時間を要する階層化ショットガン法に比べ、全ゲノムショットガン法は短時間に大量にリード獲得ができるため、現在でも多くのシークエンサーで取り入れられている。

2.2　シークエンサーによるリードの生成

1977 年にサンガー法が発表されて以来 1983 年には PCR 法の開発と、DNA シークエンシング技術は著しく発展し、1986 年には読み取りを自動でおこなう蛍光 DNA 自動シークエンサーが開発された。1995 年には全ゲノムショットガン法が開発され、1998 年にはそれを搭載したシークエンサー ABI3700 が発売された。そして近年、並列分散処理により読み取り処理をおこなうギガシークエンサーが発展したことで、大規模な DNA 配列の入

手が可能になってきている。DNA ポリメラーゼやリガーゼを用いて逐次 DNA 合成反応を行い，蛍光・発光などの方法により塩基配列を決定する Genome Analyzer 、454 FLX , SOLiD などのシステムは第 2 世代シークエンサーと呼ばれている。これまでのシークエンス機能の推移を、シークエンサーと合わせて表2.1に示す。1 日あたりの、一度に読み取れる長さをリード長、読み取れるリードの本数、取得できる総塩基数を比較する。

表 2.1: これまで開発されてきたシークエンサー

Platform	機種名	発売年	リード数	リード長 (bp)	取得塩基数 (GBp)
ABI Sanger	3730xl	1998	96	800	0.00000768
454	GS20	2005	200,000	100	0.02
illumina	GA	2006	28,000,000	25	0.7
454	GS FLX	2007	400,000	250	1
SOLiD	1	2007	40,000,000	25	1
illumina	GA	2008	28,000,000	35	1
454	GS FLX Titanium	2009	1,000,000	500	0.45
illumina	Hiseq	2011	2,000,000,000	100	200
SOLiD	5500xl	2011	3,000,000,000	60	180
454	GS FLX+	2011	1,000,000	700	0.5
IonTorrent	Proton	2012	5,000,000	200	10
illumina	Hiseq2500	2012	3,000,000,000	100	600
PacBio	RSC2XL	2012	36,000	4300	0.155

初期のシークエンサーの特徴として、1 回あたりのリードの読み取り長が長く、リード数が少ないため 1 日あたりに獲得できる塩基数が少ないのに対し、第 2 世代以降のシークエンサー (ギガシークエンサー) は短時間に大量に読み取る点やリード長が短い点が挙げられる。1 回あたりのリード長が短く本数が多いため 1 日で大規模なデータの獲得が可能になったことがわかる。

2012 年頃には、DNA の増幅を行わず単一の DNA 分子を鋳型とする第 3 世代シーケンサが開発され（PacBIO 社製 RS)、短い運転時間で連続的に数千塩基の読み取りができるようになった。PCR 法による増幅の必要が無く、アセンブリの必要性も減少すると言われている。しかしこの 1 分子リアルタイムシークエンサーの読み取り精度には改善の余地があると言われており、第 2 世代シークエンサーによる配列アセンブリに頼らざるを得ないというのが現状である [4]。配列の決定に DNA 合成酵素による DNA 鎖伸長反応を利用することが共通の特徴であり、配列を読み取る精度も速度も、使われる DNA 合成酵素の生化学的特性に依存することになり、反応条件、反応時間、精度に制約が生じる。この弱点を軽減するため、近年では化学反応ではなく、微細な穴を DNA1 分子が通過する際の電荷や水素イオン，表面温度といった物理的特徴計測して配列を読み取る post-light シー

クエンサーは第 4 世代シークエンサーとして注目を浴びている。

また前章では、複数のアセンブリを組み合わせたハイブリッドアセンブリについて述べたが、シークエンシング技術の分野においても、用いるプラットフォームによって、塩基配列の読み取りエラーにパターンがあることも報告 [64] されていることから、近年では複数のシークエンサーを用いて配列読み取りをおこなうことで各シークエンサーによる読み取りエラーの補正を目的としたシークエンシング法も提案されている [67][66]。

第3章　アセンブリによる全配列の生成

配列解析の分野にて前提とされるDNA配列データには、シークエンサーより出力された読み取り配列データである「リード」を正しく結合する過程である「DNA配列アセンブリ」が必要とされ、リードを用いて表現する「隣接グラフ」の生成と、隣接グラフ上における「経路探索」による結合配列決定と大きく2つの過程にわけられる。読み取り長や読み取りエラー率といったシーケンサーの特性、つまりリードの特性に応じて、これまで多くの隣接グラフや経路探索法の組み合わせとしてアセンブリ手法が提案され、その性能や特徴について議論されてきた。本章では代表的な従来の隣接グラフおよび経路探索法について説明し、後半ではk-merを適用した複数のアセンブリ手法を用いて、同一のリードデータのアセンブリを実行した結果を比較し、課題とその改善案について検討した。

3.1　アセンブリにおける隣接グラフ

アセンブリのリード結合において表現される隣接グラフの構造は、リード間の重複情報に基づいたOverlap Layout Graph、k-merを用いたde Bruijn graph、k-merをデータ構造で表現するprefix treeの3つに分類される。

3.1.1　Overlap Layout Graph

読み取り長が比較的長いリードデータのアセンブリには、図3.1のようにリードを直接ノード(頂点)として表現し、ノード間つまりリード間の重複情報をエッジとした隣接グラフが用いられることが多い。生成された隣接グラフは、重複部位であるoverlap部分により構成されていることからoverlap layout graph(OLG)と呼ばれている。

ノードを繋ぐエッジ上にはノード間の重複長が重みとして設けられ、グラフ上で全てのノードを通過するようなハミルトンパス問題により経路探索をおこない、決定したパスに従いcontigを生成する。

3.1.2　De Bruijn Graph

ギガシークエンサーの登場により、読み取り長が短く、リードエラーが多発するようになって以来、k-merの出現頻度値を用いた読み取りミス部位の推定法[67]などの、リード

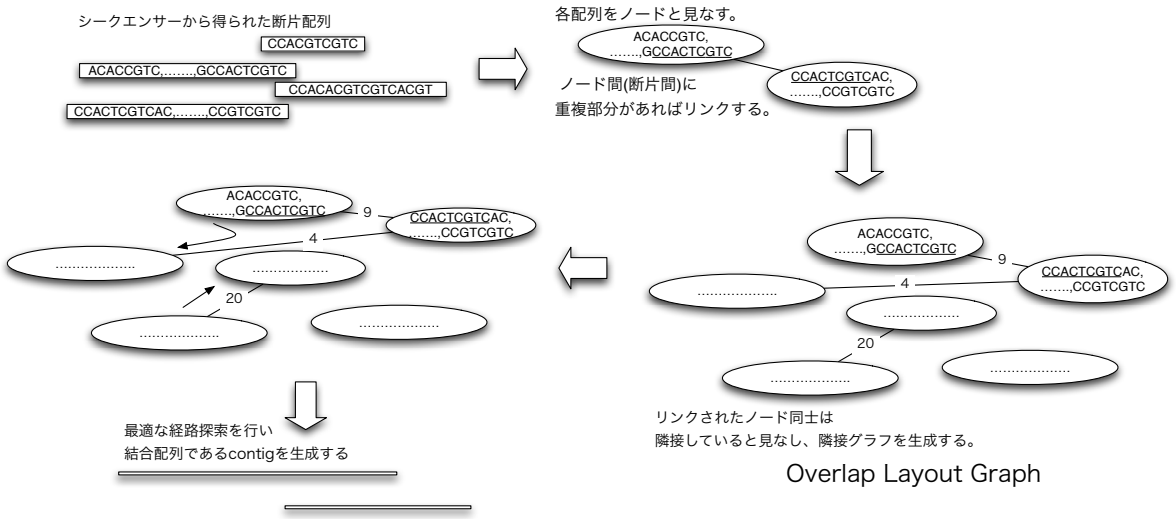

図 3.1: アセンブリにおける Overlap Layout Graph

上におけるエラー除去の処理が必要になった。k-mer を用いた代表的な隣接グラフとして De Bruijn Graph が挙げられる。

リードの信頼性を維持しながら隣接グラフを生成するためにまず、各リードに対して、k-mer と呼ばれる k base の定数長の部分配列を生成する。k の値はアセンブリの実行時にユーザが決定し入力する値である。k-mer はリード長より短く設定され、リードに対して 1 base シフトし生成するため、l base のリードからは $(l-k+1)$ 個の k-mer が獲得できる。各 k-mer は、シークエンサーデータにおける出現頻度の値を示す coverage value とともにハッシュテーブルに格納される。ここで coverage value の値が著しく小さい k-mer については、シークエンサーによる読み取りエラー部位と見なされ除去の対象になる場合がある。次に k-mer をノード、(k-1) base の重複を持つような k-mer 間に設定されるエッジにより、隣接グラフを生成する。k-mer をノードとした隣接グラフは De Bruijn Graph(DBG) と呼ばれており、最後に DBG 上の経路探索により決定した経路に従って contig を生成する。リードから DBG を経て contig を獲得するまでの流れを図3.2に示す。

グラフ上における経路探索は、全てのエッジを通過するような経路を決定するオイラニアン問題 [37] とされ決定パスに従って contig が生成される。

Zhenyu ら [81] のように、OLG、DBG 両グラフにおけるアセンブリの性能を比較した報告もある。k-mer を用いた代表的な従来アセンブリである Velvet[43] では、生成した DBG に、ハッシュテーブルにおけるリードの位置情報を用いた Pebble や Rockband[22] といった手法を用いて contig を生成し、ABySS[23] ではこれに加え k の値を 1 ずつ増やし一定の条件を満たした場合の結果を出力とする手法を提案した。

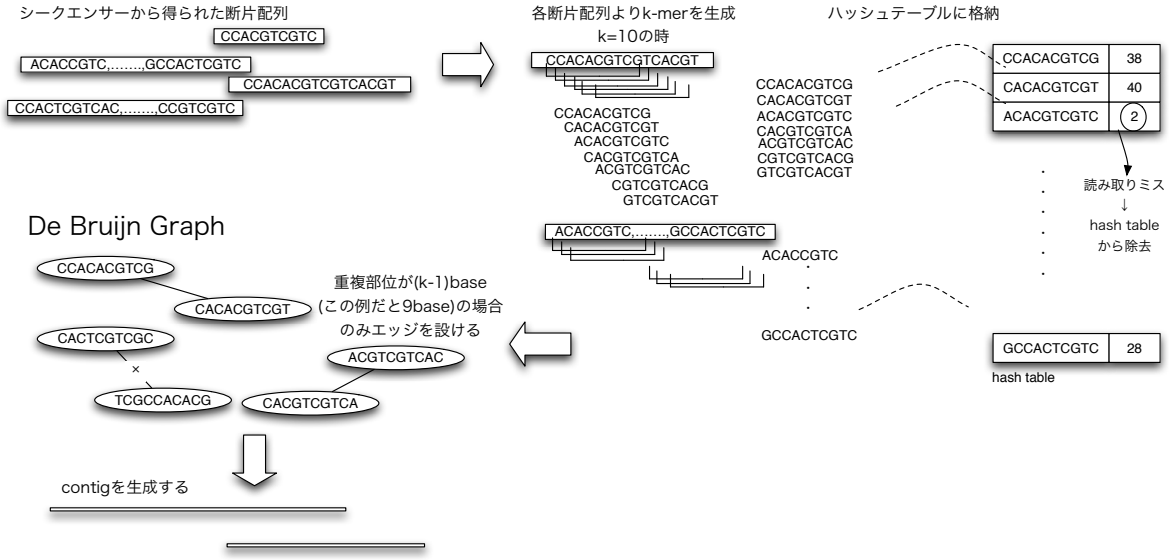

図 3.2: アセンブリにおける De Bruijn Graph

3.1.3　Prefix tree

複数の文字列の集合を木構造で表現するものでありトライ木と呼ばれる。各ノードは、共通の接頭辞により構成されている。

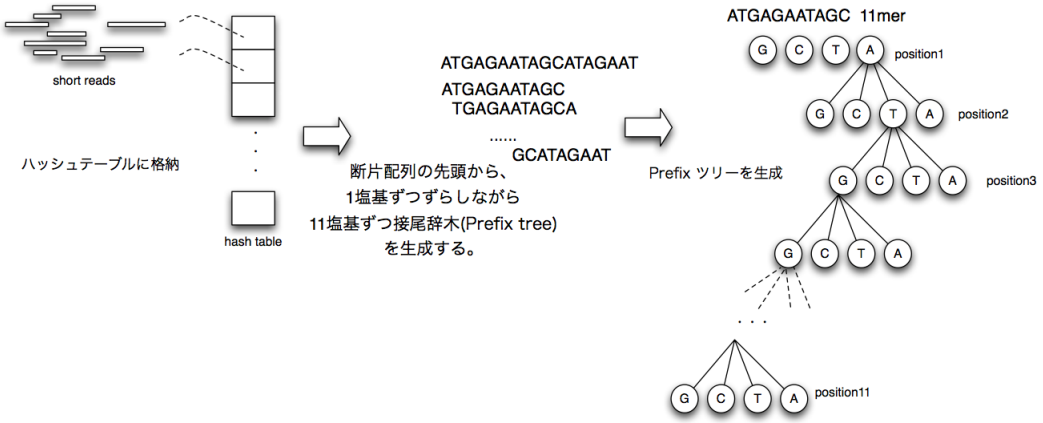

図 3.3: アセンブリにおける Prefix tree

SHARCGS[44]、SSAKE[20] や VCAKE[20] のように、k-mer をノードとした prefix tree を隣接グラフとして適用し contig を生成している手法もある。(k-1) base という高い制約条件を満たしたノード間にしかエッジを設けないという graph の構造が特徴的であり、結合の信頼性も高いと予想される。近年では SGA[27] のように、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リードを除去し、OLG を生成することで配列結合の精度向上を試みているアセンブリもある。

一方で、先述の通り k-mer を用いたアセンブリの特徴として、k の値をアセンブリの実行時に決定する必要があるため、初期値である k 値にアセンブリの結果が依存してしまうという問題がある。

3.2　経路探索法

グラフ理論において、ノードと重みつきエッジより構成される隣接グラフ上で利用される一般的な経路探索法にダイクストラ法 (Dijkstra's algorithm)[84] やウォーシャルフロイド法 (WarshallFloyd Algorithm)[83]、ランダムウォーク (Random Walk Algorithm)[85] があり、最短経路の探索に用いられることが多い。

3.2.1　ダイクストラ法

ダイクストラ法は、与えられた 2 頂点間の最短経路探索を目的とした問題で、各ノードにおいてエッジの重みが最小であるノードを選択する探索法である。以下に手順を示す。

ダイクストラ法

[入力] 重み付き隣接グラフ $G = (V, E)$、d_{ij}:ノード i と j の距離

[出力] ダイクストラ法により決定した経路

step1 始点および終点 node を選択

step2 終点 node から始点 node に向かって重みが最小化されるように node を選択

step3 始点 node に到着したところで終了

step4 経路決定

3.2.2　ウォーシャルフロイド法

ウォーシャルフロイド法では 1 本のエッジの通過を 1 ステップとし、各ステップ数において重みの和が最小となるような経路を決定する。ステップ数を増加させ、終了条件を満たした場合に経路探索を終了する。ウォーシャルフロイド法を利用した経路探索ではステップ数毎に、2 頂点間の全組み合わせについての最短経路、最短距離行列を更新する必要がある。手順を以下に示す。

ウォーシャルフロイド法

[入力] 重み付き隣接グラフ $G = (V, E)$、d_{ij}:ノード i と j の距離、p_{ij}:ノード i から j までの経路

[出力] ウォーシャルフロイド法により決定した経路

step1 $d_{ij} > d_{kj}$ なら $d_{ij} = d_{ik} + d_{kj}$,$p_{ij} = p_{ik}$ に更新

step2 $h = h + 1$ とし step1 に戻る。$2^h \geq n - 1$ の場合終了

3.2.3　二分探索木

検索効率を向上させるための木構造を成すデータ構造であり、要素の挿入や削除を高速におこなうことができる。木の最上部にある根 (ルート) から最下部にある葉 (ノード) までの経路を探索する。左の子¡親の値¡右の子の値というルールが定められており、根から葉にたどるように経路が探索される。経路探索の手順を以下に示す。

二分探索木

[入力] 二分木、探索する値 x

[出力] 二分探索木法により決定した、ルートから x までの経路

step1 ノードの値を n とする

step2 根 (ルート) より検索開始、$x = n$ であれば検索終了

step3 検索ノードとルートのノードの値を比較し「x<n」であれば左のノード、逆であれば右のノードへ移動

step4 葉へたどり着くまで step3 を繰り返す

3.2.4　ランダムウォーク

乱歩または酔歩ともよばれる。経路探索の際、次に訪れるノードをランダムに、つまり不規則に選択する。次のステップを 0.5 の確率で選択する場合や 4 方向に 0.25 の確率で選択する高次元ランダムウォークもモデル化されている。ブラウン運動の表現やインターネットのような巨大なネットワーク探索に有効な手段とされており、グラフ全体ではなく局所的な情報をもとに検索をおこなうという特徴を持つ。経路探索の簡単な手順を以下に示す。$X_n(n = 1, 2, ...)$ を独立かつ同分布な R^d 値確率変数の列とするとき、

$$S_0 = s, S_n = s + \sum_{i=1}^{n} X_i \tag{3.1}$$

によって定義される確率過程 $S_n(n = 1, 2, ...)$ をランダムウォークと呼ぶ。

3.2.5　重み優先探索法

従来のシークエンサーによる、比較的長いリードのアセンブリには、リード間の重複長と経路長を優先させることで信頼性の高く長い contig の獲得が可能であった。それに対し、配列処理をおこなうギガシーケンサーによる比較的短いリードのアセンブリに同様のアプローチでアセンブリを実行し、リード間の重複長やパス長を優先させると、リード間の重複長が短いにも関わらず長い contig、つまり信頼性の低い contig が生成されるという問題が生じる。ギガシーケンサリードに対応した信頼性の高いアセンブリを実行するためには、まず「長く正しい」contig の生成に向け、出力 contig の元配列の復元率である「被覆率」を改善する必要がある。そのために、従来の隣接グラフにおける node つまりリード間の重複情報を用いたアイディアと、従来の経路探索手法における全 node 間の経路行列を格納するアイディアを組み合わせ、リード間の重複長の合計値が最大化されるような重み経路探索法を提案した。はじめに、各ノードと、ノード間つまりリード間の重複長を重みとしたエッジより隣接グラフを生成する。エッジを 1 度通過することを 1 ステップとし、各ステップ数における経路の重みの総和を格納する重み行列、訪問重複を避けるために訪問済みノードの情報を格納するための経路行列を生成する。さらに、各ノードでの信頼性の高いエッジ選択を可能にするためにエッジの重み、つまりリード間の重複長に下限値を設定することで結合の信頼性を維持し被覆率の改善に向けた経路探索をおこなう。以下に手順を示す。

重み優先探索によるアセンブリ**: ALG3.1**

[入力] リードデータ $R = \{r_1, r_2, \cdots, r_n\}$

結合下限値 lim

[出力] ALG3.1 により生成された contig 群 $C = \{c_1, c_2, \cdots, c_n\}$

step1 各リードをノードとし、ノード間に設定した lim 以上の重複部分があればエッジを与える。エッジを 1 本経由することを 1 ステップとし、全てのノード間の重み d を格納した距離行列 A とノード自身の次に訪問が可能なノードを格納した経路行列 P、訪問済みノードを記憶する行列 PS を用意する

step2 探索を開始する。通過可能なエッジ数 e を 1 ずつ増やし、d が最大化されるよう経路を選択し P、PS を更新し A には新たな経路による d を格納する

step3 e の追加により訪問可能になったノードを経由する方が d が大きくなる場合は、P に新たにノードを追加し、A には新たな経路による d を格納する。その際には訪問情報を格納した PS も更新する。

step4 $e = e + 1$ とし Step2、3 の手続きを繰り返し、$e = \frac{n}{2}$ に達した時探索を終了する

step5 全ノードに対して最大の d を生成するような経路を最適経路とし、P に従い contig 群 C を出力

3.3 重み優先探索と機械学習アルゴリズムによるDNAアセンブリ

前述のとおり、一般的な DNA 配列アセンブリ手法では、各リードをノード、重複情報をエッジとした隣接グラフを生成し、グラフ上における経路探索をおこなう。前節では、配列間の重複長が結合の信頼性に影響することに着目し、重複長を重みとした隣接グラフ上における重み優先探索法を提案したが、結合の指標として重複長のみでは結合の信頼性を維持するには不十分である。近年、大規模データの獲得が可能になったことにより、データにおける何らかの特徴抽出を目的とした機械学習アルゴリズムの利用が注目されている。これらの背景より、本節では機械学習アルゴリズムを用いた重み優先探索によるアセンブリについて説明する。

3.3.1 機械学習アルゴリズム

計算機技術の発展による大規模データの容易な獲得が可能になって以来、膨大なデータにおける特徴を用いて未知データにおける特徴を推定したり、再利用が可能な知識を獲得するための機械学習アルゴリズム [34] の分野の研究が盛んになっている。現在ではメールのフィルタリングや地震の予想、人工知能を搭載したロボットの開発や医療分野での画像診断など、幅広い領域で実用化されている。大規模なデータを複数のグループに分類する場合にはクラスタリング手法が、未知データに対して特徴量をもとに 2 値または 3 値の属性に分類する場合には、線形判別器である判別分析や回帰分析が用いられる。さらに線形表現が困難なモデルに対応した、非線形判別器である Support Vector Machine や決定木アルゴリズムも提案されている。Support Vector Machine[86] は 2 値の分類問題を解くことを目的としており、本来線形の判別器であったが、kernel 法の利用により非線形の識別器に拡張された。それにより線形表現が困難な複雑な学習を、高次元の特徴空間を設けることにより線形判別が可能になった。Kernel 関数には画像の分類でよく用いられる polynomial kernel、汎用的な gaussian kernel、大規模なテキストデータの学習でよく利用される linear kernel、ニューラルネットワークと代用される hyperbolic tangent kernel がある。

神経細胞であるニューロンにおいて結合部分であるシナプスの信号の発火の有無を学習し、ネットワーク表現したニューラルネットワーク [38] も代表的な機械学習アルゴリズムの一つである。複雑な構造を持つデータであるほど、獲得した判別器に偏りが生じるという従来の手法の課題を解決するために、20 世紀に入ってから、過学習を避け汎用性の高い判別器を生成するアンサンブル学習アルゴリズムも多く提案されている。個々に学習した複数の判別器を融合させて汎化性能を向上させる方法である。Bagging 法 [9] で

ははじめに 1 つの学習データを、ランダムサンプリングにより複数の学習データへの分割することで複数の判別器を生成し、最終的には複数の判別器の多数決判定によって行われた。それに対し Wagging 法では、ランダムサンプリングではなく、重み付けをランダムに行い、複数の判別器を生成している。Random forest 法 [11] では、説明変数をランダムに選択し、独立性の高い判別器群を作成している。Boosting 法 [10] では、重みを均等に割り振った学習データを用い、判別器生成の過程で誤判別したサンプルに対して重みを大きくしていくことで分類精度の向上を図っており、最終的には複数の判別器の投票で行われる。Bagging 法より判別精度が高いことが知られている。また Wagging 法と、Boosting 法を組み合わせることで、ランダムに重みをつけた学習データに対しランダムサンプリングをおこなう MultiBoost 法 [12]、その他にも計算量軽減のために Boosting 法を並列化した Parallelized Boosting 法 [30] も開発されている。近年では Bagging 法と Boosting 法の統合により、学習データのランダムサンプリング後に誤判別サンプルへの重みを更新する IBB アルゴリズム [13] も開発されている。このように多くのアンサンブル学習アルゴリズムが提案されていることから、従来の決定木アルゴリズムである C4.5[32]、アンサンブル学習の性能比較に関する成果 [16][15] も多く報告されている。

しかし学習データの加工、数回の判別器の更新の過程を経る必要があることや用いるパラメータをランダムに選択していることから、アンサンブル学習は汎用性をもつ反面、弱判別器であるというデメリットもある。また判別器の内容がブラックボックスになっているため、未知データへの適用が困難という弱点もある。アンサンブル学習をおこなう過程において、有用なパラメータを優先的に選択することで、汎用性のある、信頼性の高い判別器生成をおこなう必要があることも機械学習の分野において課題の一つとなっている。

3.3.2　判別ルール獲得のための学習データの生成

アセンブリにて生成された contig に対して何らかの特徴量を用いることであらかじめ結合の正誤が予測できれば、獲得する contig の正解率を改善することが可能である。重み優先探索法ではノード間の重みをもとに経路を決定していることから、リード間の重複長の情報をもとに機械学習アルゴリズムを用いて contig の正誤を判別するルールを獲得することが可能である。そのためには、あらかじめ contig の正誤を目的変数、各結合配列の重複情報を説明変数とした学習データが必要である。本研究ではギガシークエンサーを参考に、既知配列を元配列とし、リードをランダムに生成し重み優先探索を用いて contig を生成した。contig を元配列と比較し正結合配列、誤結合配列に分類し目的変数を定義した。説明変数には重複長 $ovlp$、結合配列長 $contig$、$ovlp$*$contig$、$ovlp/contig$、log_{ovlp}、log_{contig} の 6 つの変数を説明変数と定義し、contig の正誤を示す 正結合, 誤結合 を目的変数とした学習データの作成手順を以下に示す。

アセンブリに向けた学習データ生成と判別ルール獲得の手順: **ALG3.2**

[入力] アセンブリ対象のリードデータと生物種が同様の既知配列 G_{sim}

機械学習アルゴリズム ML

[出力] ALG3.2 により生成された判別ルール $DR = \{dr_1, dr_2, \cdots, dr_n\}$

step1 ギガシークエンサーにならい、G_{sim} を複製しランダムに断片化しリードを生成

step2 Step1 で生成したリードに、対象リードデータと同様のアセンブリ手法を適用して生成した contig 群 $C_{sim} = \{c_{sim1}, c_{sim2}, \cdots, c_{simn}\}$ を元配列との比較により $c_{simn} \subset G_{sim}$ の場合正結合、そうでない場合誤結合に分類する

step3 各 contig の持つ特徴量 $F = \{f_1, f_2, \cdots, f_n\}$ を定義する

step4 正結合または誤結合を目的関数、F を説明変数とし学習データ D_{tr} を生成

step5 D_{tr} に ML を適用し DR を獲得

3.3.3　各 Support Vector Machine の学習能力比較

判別能力の高い判別ルールを獲得するには、学習能力の高い機械学習アルゴリズムを用いる必要がある。アセンブリへ適用する手法決定のために、同一の試験データにおける各 Support Vector Machine の学習能力を評価した。学習能力の評価のために、識別結果を表 3.1のように分類した。

表 3.1: 識別結果のフォーム

	有効	無効
正結合	C/C 正結合配列に対して 有効と判定した数	C/W 正結合配列に対して 無効と判定した数
誤結合	W/C in 正結合配列に対して 有効と判定した数	W/W in 正結合配列に対して 無効と判定した数
	実際出力される配列数	

(C/C+W/C) がアセンブリ適用時の実際の出力数となる。C/W と W/C は誤判別の数になるので、学習能力 LeR を式3.2のように定義した。

$$LeR = \frac{C/C + W/W}{C/C + C/W + W/C + W/W} \tag{3.2}$$

検証用データとしての既知配列には RIKEN[71] のデータベースサイトより獲得した酵母菌の塩基配列データを一部 (2325base) 用い ALG3.2 に適用したところ正結合例 346 本、誤結合例 346 本を獲得した。検証した各 support vector machine の学習能力を表3.2に示す。

表 3.2: 各 SVM algorithm の学習能力

Polynomial kernel SVM				Gaussian kernel SVM			
	有効	無効	学習能力		有効	無効	学習能力
正結合	344	2	99.5 %	正結合	345	1	99.7 %
誤結合	1	345		誤結合	1	345	
Linear kernel SVM				Hyperbolic tangent SVM			
	有効	無効	学習能力		有効	無効	学習能力
正結合	344	2	98.4 %	正結合	328	18	96.9 %
誤結合	9	337		誤結合	3	343	

アセンブリへの適用を考慮すると、正結合配列が最も多く正しく出力できる SVM が望ましい。Polynomial kernel は 2 本の正結合配列と 1 本の誤結合配列に対して、Gaussian kernel は各 1 本の正結合配列および誤結合配列のみ誤判別が発生しており、学習能力も 99 %を上回り特に学習能力の高さが確認できた。Linear kernel や Hyperbolic tangent に関しても 11 本または 21 本の誤判別に留まっており学習能力の高さが伺えたが前者には及ばなかった。学習能力が最も高く正結合配列を多く出力している Gaussian kernel SVM を重み優先探索アルゴリズムを用いた DNA アセンブリへ適用する機械学習アルゴリズムとして採択した。

3.3.4　重み優先探索と機械学習アルゴリズムによる DNA アセンブリ

ギガシーケンサリードに対する被覆率の改善に向けた、重み優先探索と機械学習アルゴリズムを用いた判別ルールと結合下限値適用によるアセンブリを提案した。手順を以下に示す。

重み優先探索と機械学習アルゴリズムによるアセンブリ**: ALG3.3**

[入力] リードデータ $R = \{r_1, r_2, \cdots, r_n\}$

結合下限値 lim

[出力] ALG3.3 により生成された contig 群 $C = \{c_1, c_2, \cdots, c_n\}$

step1 R に対して重み優先探索によるアセンブリ ALG3.1 を行い、contig を獲得する

step2 アセンブリ適用のための判別ルール獲得に向けた学習データ生成 ALG3.2 を実行し判別ルール DR を獲得する

step3 Step1 で獲得した contig 群に Step2 で獲得した DR を適用し、選別された contig 群を出力結果 C とする。

リード間の重複長を考慮しながらグラフ上にて経路決定をし、誤結合が発生した場合でも機械学習アルゴリズムにて獲得した判別ルールにより誤結合配列の出力を抑え、信頼性の高い contig の出力および被覆率が改善されることが期待できる。

3.3.5　重み優先探索と機械学習アルゴリズムによる DNA アセンブリの性能評価

重み優先探索と機械学習アルゴリズムによる DNA アセンブリの有効性を検証するためにおこなった実験について述べる。性能評価指標として、生成された contig が正結合、つまり元配列と完全一致である場合には c、誤結合である場合には inc としてその出力数を観測し、正しい contig が元配列を復元している割合を示す被覆率を $covR$ で定義した。

検証用データとしては RIKEN[71] のデータベースサイトより獲得した酵母菌の塩基配列データを一部 (2400base) 用い、学習用・試験用データとして用いた。ギガシークエンサーの特徴の一つとして大規模で比較的短いリードデータの産出が挙げられる。そこで読み取り配列数、つまりリードデータ量の変化による下限値や機械学習アルゴリズムの効果を検証するためにリード長を 50-70base、リード数を 300、900 本、結合下限値を 3、5、7、11 と変化させアセンブリ性能の変化を観察した。各アセンブリの結果を表**??**-表**??**に示す。

リード数が少ない (300 本) 場合に下限値を過度に上げると、特に 5 以上の場合に正結合配列を獲得しやすくなる一方で、被覆率が急激に減少していることがわかる。低下限値の場合においては機械学習の適用により誤結合配列が大幅に減少しており、機械学習の効果が確認できる。それに対してリード数が多い (900 本) 場合、リード数が十分にあるので結合下限値を過度に上げても被覆率を維持したまま正結合配列を多く獲得できる。一方で低結合下限値の適用時には機械学習の効果が確認できるが結合下限値の適用時には誤結合の発生がほとんど起こらないため、機械学習の効果が見られないこともわかる。

以上より、リード数が多くない場合、機械学習アルゴリズムと 5 base 以下の結合下限値の設定で結合の信頼性の高いアセンブリが実行できる。

3.4　アセンブリの性能比較

前述のとおり、用いる手法や k の値によって、アセンブリの結果は大きく異なる。そのため配列が完全に未知である de novo アセンブリにおいては、配列獲得に適した手法や手法内における適切な k 値の決定が非常に困難である。今後ますます活発化が予想される de novo アセンブリの信頼性改善に向け、本研究では複数の手法や k 値の適用に着目した。はじめに、手法や k 値の違いが与えるアセンブリ性能の変化について観察をし、これらの統合によるアセンブリ性能改善の可能性について検討する。

下限値	機械学習無し			機械学習あり					
0	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	311 (42.8 %)	98.9 %	正結合	310	1	311	99.3 %	98.9 %
	誤結合	415 (57.2 %)		誤結合	4	411	415		
	合計	**726**		合計	**314**				
3	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	309 (83.5 %)	**99.9** %	正結合	306	3	309	97.5 %	**98.9** %
	誤結合	61 (16.5 %)		誤結合	6	55	61		
	合計	370		合計	312		370		
5	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	316 (95.4 %)	98.9 %	正結合	309	7	316	96.6 %	98.99 %
	誤結合	15 (4.6 %)		誤結合	4	11	15		
	合計	331		合計	313		331		
7	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	318 (99.3 %)	92.7 %	正結合	309	9	318	97.1 %	92.7 %
	誤結合	2 (0.7 %)		誤結合	0	2	2		
	合計	320		合計	309		320		
11	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	318 (100 %)	92.7 %	正結合	310	8	318	97.4 %	92.7 %
	誤結合	0 (%)		誤結合	0	0	0		
	合計	**318**		合計	**310**		318		

表 3.3: 断片配列数が 300 本の時の判別精度

下限値	機械学習無し			機械学習あり					
0	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	1109 (48 %)	99.5 %	正結合	1105	4	1109	99.6 %	99.5 %
	誤結合	1182(52 %)		誤結合	5	1177	1182		
	合計	**2290**		合計	**1110**		2291		
3	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	1103 (95 %)	99.4 %	正結合	1100	3	1103	99.2 %	99.4 %
	誤結合	57 (5 %)		誤結合	6	51	57		
	合計	1160		合計	1106		1160		
5	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	1117 (98.8 %)	99.4 %	正結合	1111	6	1117	98.8 %	99.4 %
	誤結合	13 (1.2 %)		誤結合	7	6	13		
	合計	1130		合計	1118		1130		
7	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	1140 (99.9 %)	99.4 %	正結合	1140	6	1146	99.3 %	99.4 %
	誤結合	1 (0.01 %)		誤結合	1	0	1		
	合計	1141		合計	1141		1147		
11	結果		被覆率		有効	無効	計	判別精度	被覆率
	正結合	1144 (100 %)	**99.4** %	正結合	1138	6	1144	99 %	**99.4** %
	誤結合	0 (%)		誤結合	0	0	0		
	合計	**1144**		合計	**1138**		1144		

表 3.4: 断片配列数が 900 本の時の判別精度

3.4.1　k 値や手法のアセンブリへの影響

DNA ギガシークエンサーによるリードから全配列を獲得するには、結合段階の前に、k-mer によるハッシュテーブルの生成時に、シークエンスエラーと考えられる箇所を除去し、配列を結合する。k 値を用いたアセンブリを実行する際には、あらかじめ k 値を決定する必要があるが、リードデータに適切な k の値が未知である場合、最良のアセンブリ結果を獲得するのは困難である。従来のアセンブリ手法では長い contig の出力が可能な k 値を適切な条件としてきたが、獲得された contig を元配列と比較した場合、不一致であるケースが多い。用いる k 値や手法がアセンブリ結果に与える影響を明確にするために、同一のデータに対して複数の手法・k 値によりアセンブリを実行し各結果とその関係を比較した。

実験データには、従来手法である SSAKE の有効性検証のためのデータとして用いられた、Herpesvirus data(6 万 base) を用いた。生成された contig が正結合配列、つまり元配列と完全一致である場合には c、誤結合配列である場合には inc に分類し、出力 contig 数を $Number$、出力中における最長 contig 長を max、正しい contig が元配列を復元している割合を示す被覆率を $covR$ で表した。アセンブリの従来手法として Velvet と ABySS を利用し k の値を 15 から 27 に変化させ実行した結果を表3.5-表3.6に、Velvet は回文配列の発生による計算量の爆発を防ぐため、用いる k の値は奇数のみであるという制約がある。

さらに、各結果の評価値ごとの推移を図3.4-図3.5に示す。アセンブリ結果の出力中における正しい contig の割合は、アセンブリの信頼性を示す指標として $corR$ の推移を示した。また信頼性の高く長い contig の生成が最も望ましく、信頼性の低い長い contig は除去するべきであるため、出力中における最長正結合配列を $MLcor$、最長誤結合配列を $MLincor$ とし、これらの推移を図3.6-図3.7に示す。

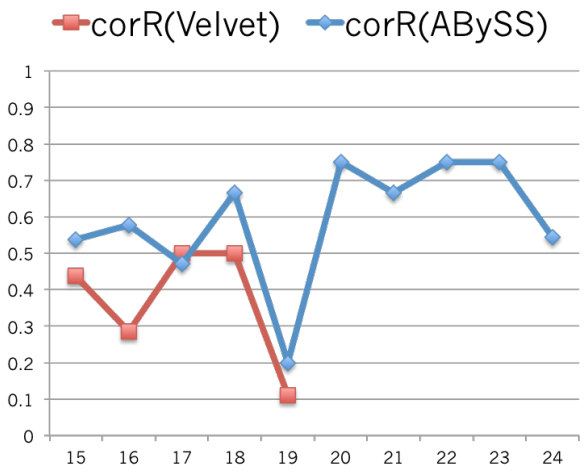

図 3.4: k 値や手法の影響 (正解率 corR)

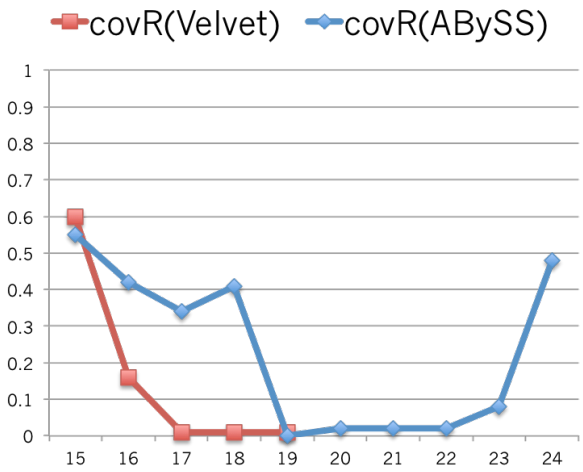

図 3.5: k 値や手法の影響 (被覆率 covR)

表3.5より velvet を用いて k 値を 15 から 27 へ変化させアセンブリを実行した場合、正解率が最も高いのは Velvet の場合 k=17、ABySS の場合 k=20 の実行時であり、他の k 値

表 3.5: k-value がアセンブリに与える影響 (VELVET)

k	15		17		19		21		23		25		27	
Result	c	inc	c	inc	c	inc	c	inc	c	inc	c	inc	c	inc
Contig	7	9	2	5	1	1	1	1	1	8	29	51	177	245
Max	17176	5123	8660	12180	1196	58807	1198	58809	1200	14202	2831	3922	422	385
covR	0.6		0.16		0.01		0.01		0.01		0.39		0.37	

表 3.6: k-value がアセンブリに与える影響 (ABySS)

k	15		16		17		18		19		20		21	
Result	c	inc	c	inc	c	inc	c	inc	c	inc	c	inc	c	inc
Contig	28	24	15	11	8	9	8	4	1	4	3	1	2	1
Max	6123	6766	11150	17178	7179	17180	11785	17182	25	28825	1195	58806	1197	58806
covR	0.55		0.42		0.34		0.41		0.0004		0.02		0.02	
k	22		23		24		25		26		27			
Result	c	inc	c	inc	c	inc	c	inc	c	inc	c	inc		
Contig	3	1	3	1	6	5	17	12	38	45	120	126		
Max	1197	58808	3694	55131	10031	14194	6324	3331	3292	3920	1508	873		
covR	0.02		0.08		0.48		0.68		0.49		0.52			

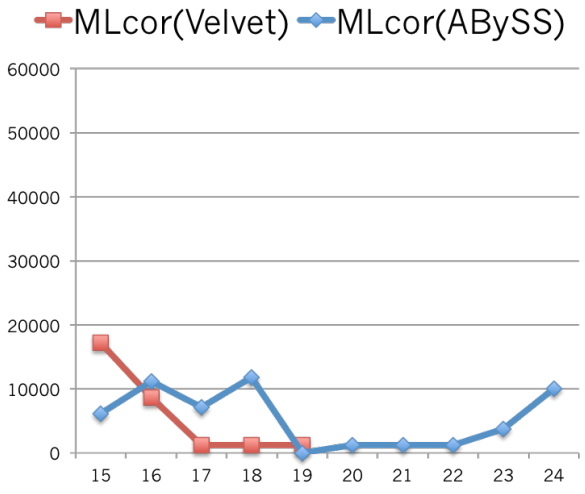

図 3.6: k 値や手法の影響 (最長正結合配列長)

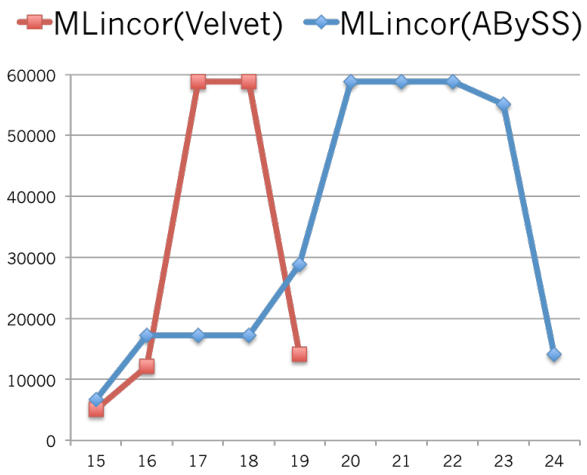

図 3.7: k 値や手法の影響 (最長結合配列長)

の結果と比較してもばらつきが見られる。Contig の長さに関しては k=19,21,23 の実行時に長い contig が生成されることがわかる。しかしこれらは誤結合配列であるため covR が大きく落ち込んでおり、図3.5、図3.6より、結果として k=15 で実行した際の covR が最大となっていることがわかる。表3.6の ABySS についても k=20,21,22,23 での実行時に長い contig を獲得できたものの、誤結合配列であるため、covR は他の k 値の実行時に比べ最小となっている。

表3.5-表3.6および図3.4-図3.7の結果より、用いる k 値によってアセンブリの contig 長、被覆率といった性能が大きく変化し、これらに関連性は無くばらつきが見られること、長い誤結合配列も生成されることがわかった。

次に復元の特徴を詳しく観察するために、各手法の k 値による contig の、元配列の復元領域について観察した結果を図3.9に示す。横軸は正しい、つまり元配列と完全一致した contig が復元した位置を、縦軸は一致した回数を示す。

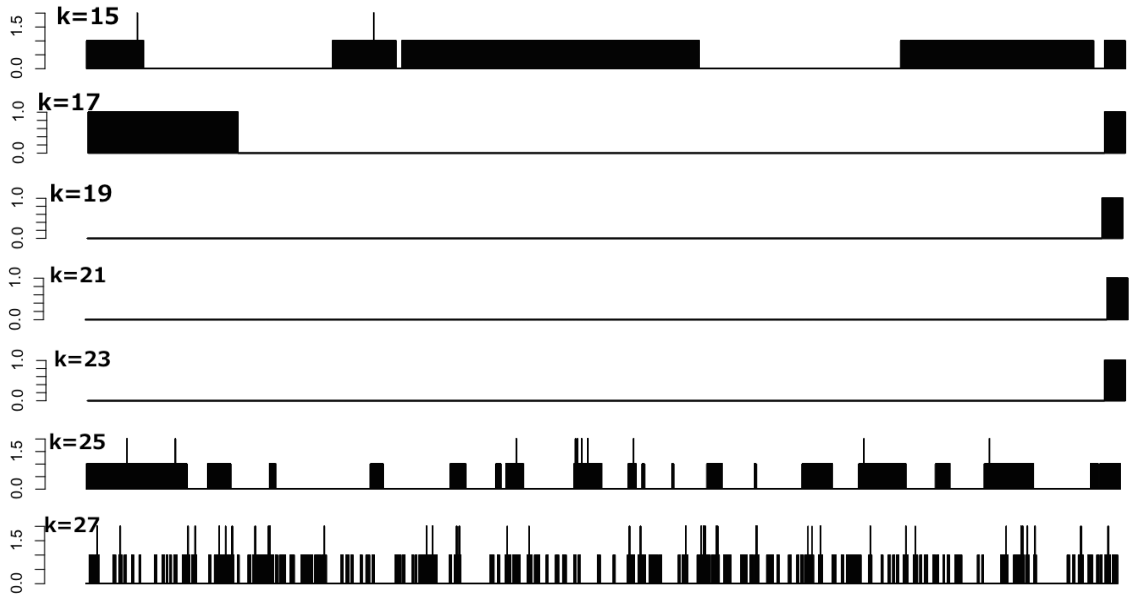

図 3.8: Velvet にて各 k 値による contig の復元領域

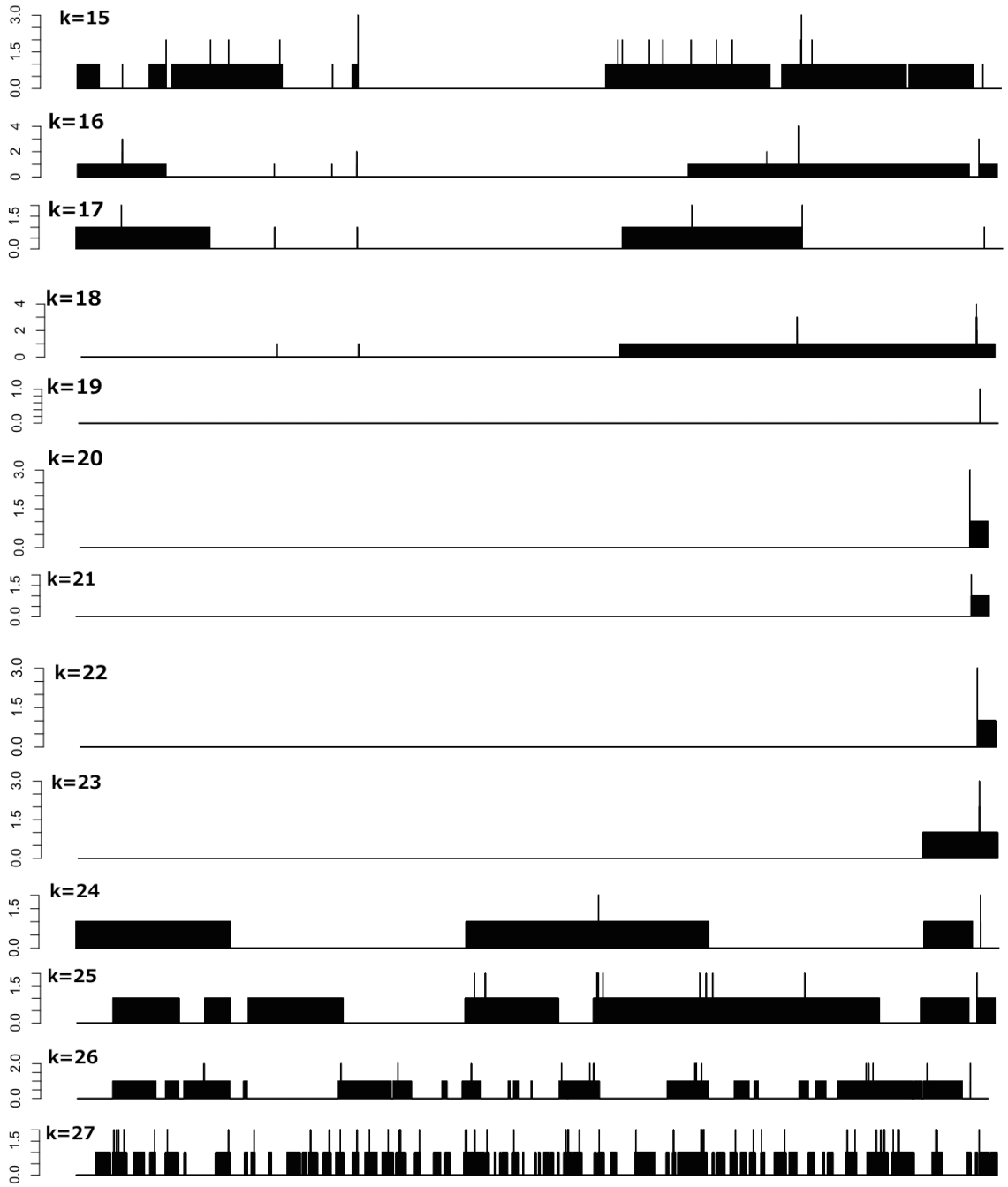

図 3.9: ABySS にて各 k 値による contig の復元領域

ABySS の k=19-23、Velvet の k=19-23 による contig は長い誤結合配列であったため、復元領域が他の実行結果に比べて狭いことがわかる。また ABySS の k=15、Velvet の k=15 による contig を正しく結合することで、被覆率が 1.0 に限りなく近づくことも期待できる。

表3.5、表3.6、図3.9より、用いる k の値や手法が、出力される contig 数や信頼性、長さに影響を与えていること、異なる k や手法のアセンブリの結果を統合することでさらにアセンブリの性能が改善されることが期待できる。特に de novo アセンブリの際には、適切な k が未知であるため複数の実行の条件の統合は合理的であると言える。

3.4.2　複数の k 値適用によるアセンブリ改善の可能性

複数の k 値や手法による contig の結合による、アセンブリの性能改善の可能性の有無を確かめるための実験をおこなう。はじめに、同一の手法内における複数の k 値適用時について検証した。図3.9をもとに、被覆領域が相補的になっている結果をもつ k 値の組み合わせを決定し、contig の全組み合わせに対して一定の重複部位をもつ場合に結合配列とする。既述の通り、「重み優先探索と機械学習アルゴリズムによる DNA アセンブリ」の性能評価にて、結合対象の配列数が十分に多くない場合は結合下限値を過度に高く設けるべきでないとの知見が得られた。k-mer を用いた従来手法により獲得した contig は本数が多くないため、本実験における結合下限値を 5 base とした。

単数の k 値および複数の k 値を用いた場合の出力 contig 数 ($contig$)・被覆率 ($covR$)・最長 contig 長 (Max) といった性能を比較する。実験データには、3.4.1 と同一のデータを適用した。

図3.9より、Velvet については k=15,25、k=19,21、k=21,23、k=25,27 の被覆領域が相補的であるため、これらの contig 結合をおこなうことで $covR$ や Max の改善が見込める。結合結果を表3.7に示す。

k	19,21	21,23	25,27
$contig$	1	3	250
$covR$	<u>0.02</u>	<u>0.02</u>	0.39
Max	1199	1201	<u>2921</u>

表 3.7: 複数の k 値利用時:Velvet

表3.5と表3.7の比較結果より、k=19,21 また k=21,23 について、被覆率 $covR$ については単数 k の適用時 (0.01) に比べ、複数同時に用いた結果が 0.01 と改善されることがわかった。また k=25,27 を単数で用いた場合、被覆率 $covR$ に変化はないものの、Max については単数 k の適用時 (2831, 422) に比べ、複数同時に用いた結果が 2921 と改善されることがわかった。

次に ABySS についても同様に、被覆領域が相補的である k=15,16、k=16,17、k=17,18、k=24,25、k=25,26 の 2 つの値の組み合わせを用いた結果を表3.8に示す。

k	15,16	16,17	17,18	24,25	25,26
$contig$	63	35	24	25	74
$covR$	0.63	0.54	0.55	0.67	0.76
Max	13195	18316	23427	10888	10888

表 3.8: 複数の k 値利用時:ABySS

表3.6と表3.8の比較結果より、特に k=25,26 を単数で用いた場合の $covR$(0.68, 0.49) に比べ、複数同時に用いた結果が 0.76 と、9–27 %改善された。これは全組み合わせについても言えることである。また Max についても全組み合わせについて改善されているが、特に k=17,18 を単数で用いた場合 (7179, 11785) に比べ、複数同時に用いた結果が 23427 と、大幅に改善されることがわかった。

以上より、単数の k 値に比べ、複数の k 値利用によりアセンブリ性能の改善の可能性があることがわかった。

3.4.3　複数の手法の適用によるアセンブリ改善の可能性

次に、複数の手法によるアセンブリ性能の改善の可能性の有無を確かめるための実験をおこなった。復元領域が相補的である ABySS k=15+Velvet k=25、ABySS k=18+Velvet k=15 を組み合わせとして結合した結果を表3.9に示す。

k	a15v25	a18v15
$contig$	27	20
$covR$	0.91	0.81
Max	22469	28820

表 3.9: Velvet, ABySS の複数の k 値利用時のアセンブリの性能

表3.5, 表3.6, 表3.9より、ABySS(k=15)Velvet(k=25) の組み合わせについて、単数の k 値を用いた場合の $covR$=0.55, 0.39 と比べ、複数同時に用いた結果が 0.91 と、60 %を上回る改善が見られ、Max=17176, 11785 と比べ複数用いた結果が 22469 と改善が見られた。ABySS(k=18)Velvet(k=15) の組み合わせについても $covR$=0.41, 0.6 から 40 %の改善、Max=17176, 11785 から 28820 と改善されたことより、複数のアセンブリ手法の結果を統合した場合にもアセンブリ性能の改善の可能性があることがわかった。

3.5　まとめ

本章では、代表的な従来のアセンブリとして、リードの特性に応じた隣接グラフやグラフ上における経路探索アルゴリズムについて述べた。「重み優先探索と機械学習アルゴリ

ズムによる DNA アセンブリ」の性能評価実験においては、リード数が十分に多い場合には低結合下限値と機械学習アルゴリズムの設定が好ましいことがわかった。そして k-mer を用いた従来のアセンブリが k 値や手法によって結果を左右することについて、同一のデータに対する複数の手法と k 値のアセンブリ実行により検証し、各結果が相補的になっていることも明らかになった。これらの結果をうけ、複数の k-mer、複数のアセンブリ手法により出力された contig を再度統合するという過程を加えることで、正しい最長 contig 長、被覆率といったアセンブリの性能が改善される可能性があることを示した。

次章では、本章で述べた結合下限値と機械学習アルゴリズムの組み合わせや、k 値や手法によるアセンブリ性能の変化を踏まえ、複数の k 値および手法を組み合わせたダブルアセンブリ、さらに contig 上における k-mer の coverage 値の分布特徴量を用いた DNA ダブルアセンブリ手法の提案をおこなう。

第4章　k-merの分布特徴量を用いたDNAダブルアセンブリ

ギガシークエンサーより獲得したリードデータのアセンブリによる全配列に向けてこれまで多くのアプローチによるアセンブリ手法が提案され、その特徴について議論され比較されてきた。近年ではこれらを組み合わせ、二段階に渡ったアセンブリ手法が増えてきている。例えばIDBA[24]では用いるk値を増加させることで、一定の条件を満たした場合を最適なk値とし、重複部位の信頼性を維持しながらcontigを生成した。k値が極端に小さい場合は隣接グラフ上にエッジが多発し経路探索を困難にする一方で、k値が大きい場合はリードエラーの除去が困難になると、改めて最適なk値の決定の困難さを示していた。これらを受けIDBA-UD[57]では、あるk値を用いてDBGにて生成したcontigを直接結合する手法を提案した。k-merのcoverage値の分布情報をもとに、ギガシークエンサーの読み取りエラー部位を除去する手法[60]の開発が盛んになり、Davidら[56]のようにあらかじめ読み取りエラー部位を除去しIDBA-UDを用いたアセンブリ手法も提案されており、複数の手法の組み合わせによるアセンブリ手法も注目されている。例えばMAIA[27]やGAA[26]では、相補的である複数の従来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間の信頼性についてcontig間のアラインメントスコアを用いながら、更に長いcontigを生成している。CISA[58]ではcontigの結合下限値とアラインメントスコアにより二段階目の結合をおこなっている。

その他にもベイジアンネットワークを利用したゲノムアセンブリ手法で、複数のアセンブラによるde Bruijn graph上における経路の決定にマルコフ連鎖モンテカルロ法を用いたアセンブリ手法[74][74]も提案されている。Rauschら[80]はリード間のアラインメントスコアを算出したアラインメントグラフを生成しcontigを生成する手法を提案した。BLESS[72]では、k-merを格納したハッシュテーブルの利用には大量のメモリが必要であることを指摘しBloom filter[73]を用いることで、ハッシュテーブルの構造を工夫し、使用メモリの軽減を試みている。

本章では、前章でのk値や手法の与えるアセンブリへの影響の検証結果をうけ、はじめに複数のアセンブリ結果の統合によるアセンブリを提案した。次に、信頼性を維持させるために誤結合配列の発生を防ぐための結合ルールの必要性について述べた。従来手法による contig 上の k-mer の coverage value の分布の特徴と結合配列の正誤に見られる関連性から、これらの分布特徴量を結合配列ルール獲得に用い、最後に従来手法との性能比較をおこなった。

4.1　複数のk-merと手法を統合したDNAダブルアセンブリ

複数のk値と手法の組合せによるヒューリスティックなダブルアセンブリ DAwH(Double Assembly with Heuristic) について述べる。はじめに従来手法において複数のk-mer と手法を用いて生成した contig をノードとした隣接グラフを生成し、5 base 以上の重複部位を有するノード同士を結合する。以下に具体的な手順について述べる。

複数のk-mer と手法の統合によるダブルアセンブリ **DAwH: ALG4.1**

[入力] リードデータ $R = \{r_1, r_2, \cdots, r_n\}$

[出力] ALG4.1 により生成された contig 群 $C = \{c_1, c_2, \cdots, c_n\}$

step1 R に対し、k-mer を用いた従来手法に k の値を 15-27 まで変化させアセンブリを実行し contig を獲得する

step2 k 値、手法の組み合わせを決定する

step3 決定した k 値、手法による contig の全組み合わせに対し、5 base 以上の重複を持つ場合に結合する。

従来手法において生成された contig らを、重複部位を有する場合さらに結合することで、従来手法では生成できなかった長い contig の獲得が期待できる。

4.2　ダブルアセンブリの有効性検証のための性能比較実験

DAwH の有効性を検証するために、単一の k 値を用いた従来手法との性能比較実験をおこなった。アセンブリの信頼性を評価するためには contig の長さだけでなく結合の「正しさ」を評価することが重要である。性能評価値として獲得 contig の元配列に対する被覆率 $CovR$、結合の信頼性を示すための、出力 contig における正しい結合配列の割合を示す正解率 $CorR$、獲得 contig における最長正結合 contig 長 $Longest$ を用いた。

実験には DNA 配列データベースである NCBI[82] に登録されている E.coli K-12 substr. MG1655 を用いた。大腸菌データはヒトなどのホ乳類と比べて塩基配列の構造が複雑でないため、アセンブリの性能比較で用いられることが多い。計算機環境の制約により、full genome の一部を用い 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。表4.1にデータの特徴を示す。

表 4.1: 検証用データの特性

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

従来手法としてABySSとVelvetを用い、表4.2のようにk値、従来手法を組み合わせてダブルアセンブリをおこなった。単一のk-merを用いた従来のアセンブリ手法と、それらをヒューリスティックに組み合わせた提案手法DAwHの性能比較結果を表4.3に示す。

表 4.2: k値と手法の組合せ

ABySS(k=17)+ABySS(k=18)
ABySS(k=25)+Velvet(k=15)
ABySS(k=18)+Velvet(k=15)

表 4.3: DAwHと従来手法の性能比較

Method	従来手法		DAwH
k value	ABySS(k=17)	ABySS(k=18)	ABySS(k=17)+ABySS(k=18)
$covR$	0.34	0.41	0.55
$CorR$	0.47	0.66	0.29
$longest$	7179	11785	23427
k value	ABySS(k=25)	Velvet(k=15)	ABySS(k=25)+Velvet(k=15)
$covR$	0.68	0.6	0.91
$CorR$	0.58	0.43	0.5
$longest$	6324	17176	22469
k value	ABySS(k=18)	Velvet(k=15)	ABySS(k=18)+Velvet(k=15)
$covR$	0.41	0.6	0.8
$CorR$	0.66	0.43	0.3
$longest$	11785	17176	28820

どの組合せの統合結果においても単一のk-merを用いた従来手法と比べて正しく長いcontigが生成できたことがわかる。また被覆率$covR$についても20-40 %の改善が確認できた。しかしその反面、誤結合配列を多く出力し$corR$が減少していることがわかる。またcontig間の重複長と結合contigの信頼性の分布を観測した図4.1からも明らかなように、重複長の値が大きいほど結合の信頼性は高いが、一方で重複長が5–30 baseの区間においては重複部位の正結合または誤結合が混在していることもわかる。

重複長に何らかの評価指標を加え正しくcontigを結合するために、次に学習データによる判別ルールの獲得について検討した。次節では判別ルール獲得のための特徴量の定義について述べる。

4.3　決定木アルゴリズム

データにおける、目的変数の分類を予測する際に用いられる代表的な手法として決定木アルゴリズムが挙げられる。決定木は学習データの特徴を条件分岐によるルールで表現し

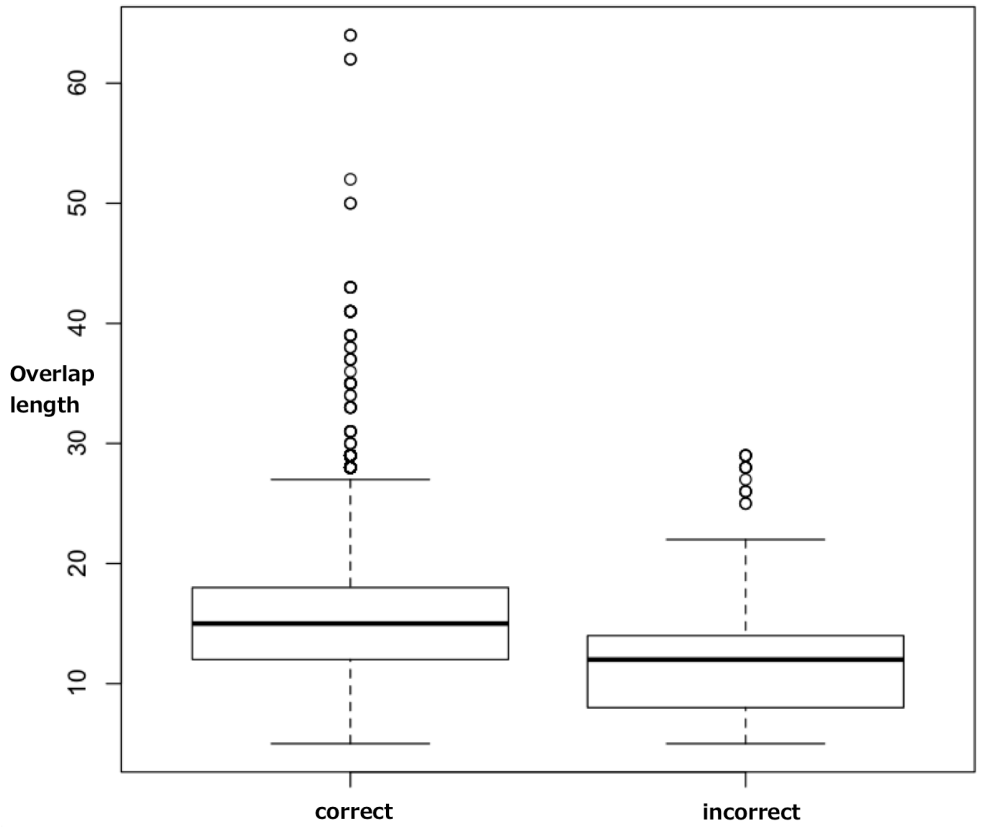

図 4.1: 重複長と結合の信頼性の分布

たもので、説明変数とその閾値により構成される。決定木作成の手順はおもに「データを分割する指標を計算」、「データの分類」に分かれている。データ分割の指標には、手法によって用いられている指標が異なり、ID3 (Iterative Dichotomiser 3)[33] や ID3 を改良した C4.5 では情報利得、CART(Classification And Regression Tree)[35] ではジニ係数が用いられている。本節では、本研究における提案手法に適用した C4.5 について説明する。

C4.5 は教師つき学習アルゴリズムの一つで、情報エントロピーの概念を用いて、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。学習データのサンプル $S_i = \{ s_1, s_2, \dots \}$ は、各サンプルの所属 $C = \{c_1, c_2, \dots \}$ とその特徴を表すベクトル $s_i = \{ x_1, x_2, \dots \}$ より構成されている。入力された学習データにおける全ての説明変数の情報利得比 (Information Gain Ratio:IGR) を求め、その中で最も大きい情報利得比をもつ説明変数を、決定木における第一分岐点とする。情報利得比を求めるにははじめに情報利得 (Information Gain)、分割情報量 (Split Information:SI) を求める。目的変数つまり所属 $C = c_1, c_2, ..., c_t$ における各 $c_i \in C$ の全体におけるサンプル数の割合を $p(c_i)$ とすると、エントロピー $H(C)$ を式4.1のように求めることができる。

$$H(C) = -\sum_{t=1}^{t} p(c_i) log_2 p(c_i) \tag{4.1}$$

各説明変数 $V = \{v_1, v_2, ..., v_i\}$ の全体に対する割合を $p(F_i)$、v_i、クラス c_i のサンプル数の全体に対する割合を $p(v_i, c_i)$ とし、条件付きエントロピー $H(C|F)$ を式4.2のように定める。

$$H(C) = -\sum_{j=1}^{u} \sum_{t=1}^{t} p(v_i, c_i) log_2 \frac{p(v_i, ci)}{p(v_i)} \tag{4.2}$$

これらを用いて、情報利得 $IG(C;V)$ は、

$$IG(C;V) = H(C) - H(C|V) \tag{4.3}$$

と定義され、さらに V における分割情報量 $SI(V)$ は、

$$SI(V) = -\sum_{j=1}^{u} p(v_j) log \tag{4.4}$$

と表すことができる。これらを用いて V における情報利得比 $IGR(C;V)$ は、

$$IGR(C;V) = \frac{IG(C;V)}{SI(V)} \tag{4.5}$$

と定義することができる。C4.5 では全説明変数に対し最大情報利得比をもつ説明変数を、決定木の構成における第 1 分岐点の候補とし、参照率が高い説明変数とされる。

4.4　DNA ダブルアセンブリにおける判別ルールの獲得

従来の k 値や手法を統合するヒューリスティックなダブルアセンブリにより、被覆率や正しい最長 contig 長の改善の可能性を示し、一方で誤結合配列の生成が多く生じ正解率の減少が課題になった。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 [48] 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。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 に適用することで判別ルールの獲得を試みた。本節では判別ルール獲得のための特徴量の定義について述べる。

4.4.1　contig 上の k-mer の coverage 値の分布と配列結合の正誤の関係

k-mer を用いたアセンブリ手法で生成された contig は、k-mer の集合とみなすことができる。各 k-mer は断片配列データ中における出現頻度値である coverage 値を所持していることから、各 contig は図4.2のように数値の集合と見なすこともできる。

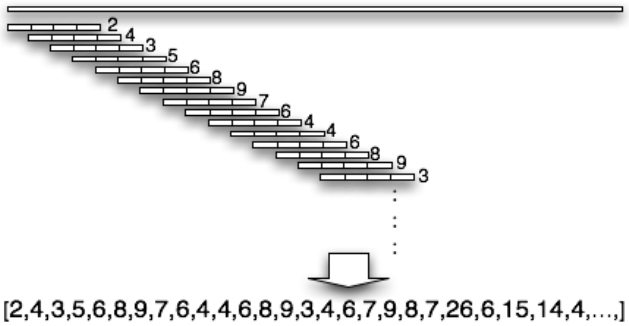

図 4.2: k-mer より構成される Contig

また k-mer の coverage 値は断片配列データにおけるデータの信頼性を示す指標とも考えられることから、本研究では contig 上における k-mer の coverage 値の分布特徴量と結合配列の正誤の関係について調べるために、従来手法により生成された contig 上における 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 と 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、5 base 以上の重複部位を持つような contig の組み合わせを列挙し結合した。次に結合配列の正誤と各 contig を生成する k-mer の coverage 値の分布を観察した。contig 上における k-mer の coverage 値の変動情報を波形と見なし、波形の特徴と contig による結合配列の正誤に関連があるかについて調べた。実験には表4.1と同じ DNA

配列データを用い、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、正しい contig(元配列に含まれる) のみ使用した。k-mer の coverage 値は k 値に大きく依存する (k 値が大きい程 coverage 値は小さく、小さい程 coverage 値は大きくなる) ため、複数の k 値による contig のダブルアセンブリをおこなう際には coverage 値の正規化が必要であることから、式4.6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。

$$p_{c_i} = \frac{|\{c_i \in C | c^C \geq c_i|}{|C|} \tag{4.6}$$

k-mer の coverage 値の変動情報と配列結合の正誤の関連の有無を確かめるために、ある 1 本の contig に着目し、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。各前方・後方 contig を固定した場合についての正結合配列を生成するパターンを図4.3-図4.4に、誤結合配列を生成するパターンを図4.5-図4.6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 の分布状況は波形表現した際に類似していることがわかる。配列結合の正誤と contig 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、結合元の各 contig 上における k-mer の coverage 値の分布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。従来手法による contig 上の k-mer の coverage 値の p-value の分布特徴量をもとに判別ルールを獲得し、ダブルアセンブリに適用することで配列結合の信頼性の改善を試みる。

4.4.2　k-mer の coverage 値の分布特徴量

判別ルール獲得に用いる分布特徴量について説明する。contig 上の k-mer の coverage 値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 の前方・後方の p-value の変動情報、頻度情報、相関 (類似度) に着目した。

はじめに、波形の変動情報の利用法については、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 [46] を用いた。フーリエ変換では、1 つの波形を、周波数の異なる正弦波と余弦波の成分組み合わせで表現する手法であり、音声データや心電図データといった波形の周波数解析に多く用いられている。本研究では波形の特徴量の一つであり、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 $U^{f,l}_{freq}$、低周波成分 $L^{f,l}_{freq}$、またフーリエ変換成分のパワーバリューである総和 $W^{f,l}$ を特徴パラメータに利用した。さらに 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4.8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 $D^{f,l}$ として式4.9のように定義した。各 p-value 集合配列の要素数における増加値の割合である増加率も式 eq:II のように定義し説明変数として定義した。

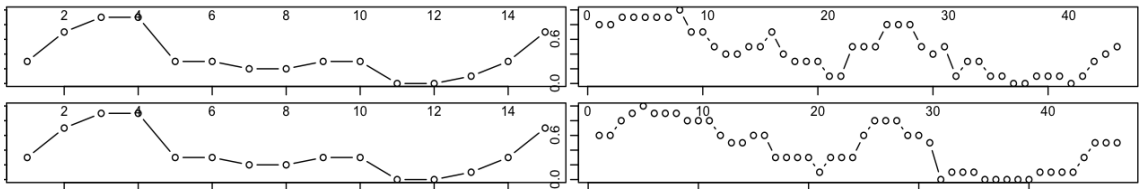

図 4.3: 正結合における各 contig の k-mer の変動状況 (前方固定)

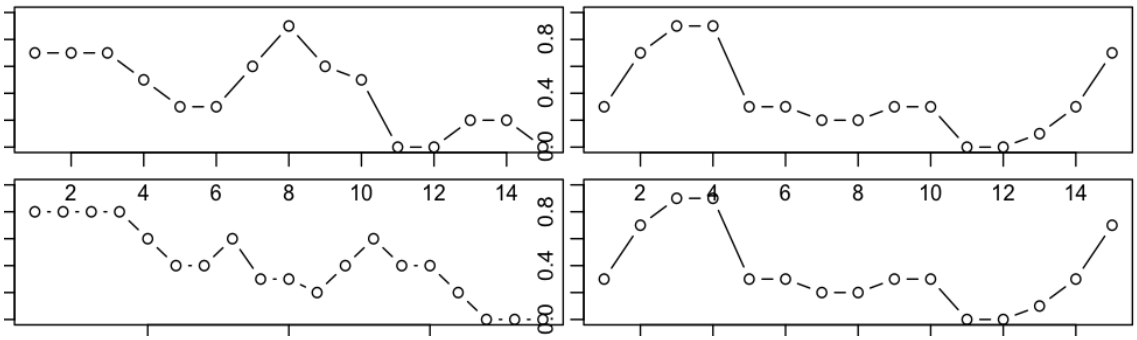

図 4.4: 正結合における各 contig の k-mer の変動状況 (後方固定)

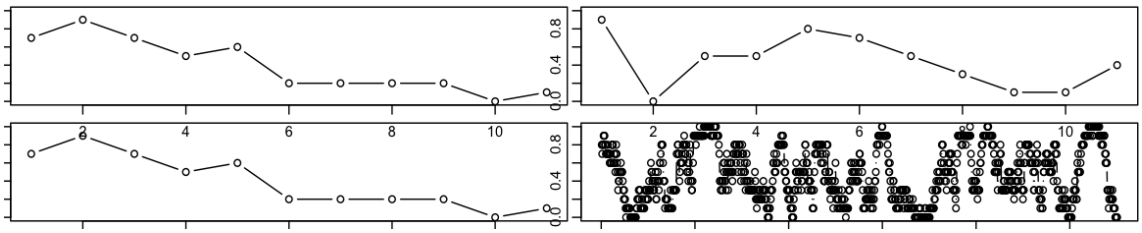

図 4.5: 誤結合における各 contig の k-mer の変動状況 (前方固定)

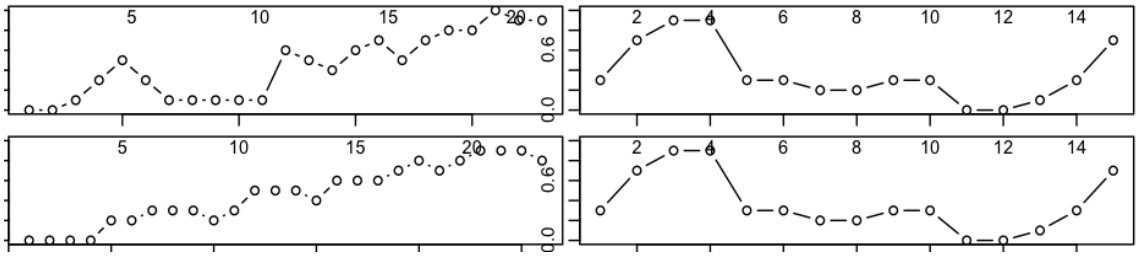

図 4.6: 誤結合における各 contig の k-mer の変動状況 (後方固定)

$$d_i = \begin{cases} -1 & (p_i < p_{i-1}) \\ 0 & (p_i = p_{i-1}) \\ 1 & (p_i > p_{i-1}) \end{cases} \tag{4.7}$$

$$D = \sum_{i=1}^{n} d_i \tag{4.8}$$

$$I = \frac{|\{d_i | d_i = 1\}|}{|n|} \tag{4.9}$$

次に、波形の頻度情報については、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、説明変数として用いた。出現頻度値が特に小さい場合は配列の信頼性に関わってくるため、頻度値=0であるp-valueの値$Q^{f,l}$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$W_{freq}^{f,l}$も用いた。

最後に、波形の相関(類似度)については、各組み合わせcontigのp-valueの集合配列のフーリエ変換後の相互相関関数の最大値を$\Phi_{\max}^{f,l}$、p-valueの度数分布の集合配列の相関係数をρ_{freq}、ハミング距離をHと定義し、説明変数として定義した。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のノルムも説明変数に用いた。これまで説明した説明変数とその定義を表4.4にまとめる。

表 4.4: 前後contigのk-merの特徴量を用いた説明変数

変動	$D^{f,l}$	p-value集合配列の勾配
	$I^{f,l}$	増加率
	$U_{freq}^{f,l}$	フーリエ変換の高周波成分
	$L_{freq}^{f,l}$	フーリエ変換の低周波成分
分布	$W^{f,l}$	フーリエ変換のpower value
	$Q^{f,l}$	頻度値=0であるp-value
	$W_{freq}^{f,l}$	p-value度数分布のフーリエ変換後のpower value
相関	ρ	p-value集合配列の相関係数
	ρ_{freq}	p-value度数分布集合配列の相関係数
	CCF_{freq}	p-value度数分布配列のフーリエ変換後の相関係数
	$\Phi_{\max}^{f,l}$	p-value集合配列のフーリエ変換後の相互相関関数最大値
	H	p-value度数分布配列のハミング距離
	M_{ccf}^{freq}	p-value度数分布の相互相関関数の最大値
	$R^{f,l}$	前方contigの末端と後方contigの先端のp-valueのノルム

以上より、既知配列より生成した学習データに対して上記で説明した説明変数の定義により判別ルールを獲得し、アセンブリ対象のリード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。学習データにC4.5を適用した場合、正結合配列の特徴を表す正結

合ルールと誤結合配列の特徴を表す誤結合ルールの両方が獲得できる。本研究では、ダブルアセンブリに対して「正結合ルールを満たすような結合 contig を抽出した場合」と「全出力において誤結合ルールに合致する結合 contig を削除した場合」の 2 通りの適用時の結果について検証する。

4.5 Contig の k-mer の分布特徴量を用いた DNA ダブルアセンブリ

contig の k-mer coverage 値の分布特徴量による結合ルールを用いたダブルアセンブリ DAwCC(Double Assembly method with Contig for k-mer's coverage) を提案した。手続きを以下に述べ、流れを図4.5に示す。

contig の分布特徴量を用いたダブルアセンブリ DAwCC: ALG4.2

[入力] リードデータ $R = \{r_1, r_2, \cdots, r_n\}$

機械学習アルゴリズム ML

[出力] ALG4.2 により生成された contig 群 $C = \{c_1, c_2, \cdots, c_n\}$

step1 R に対して DAwH(ALG4.2) を実行し contig 群を生成する

step2 Step1 で生成した contig 群に、ML=C4.5 としてアセンブリに向けた学習データ生成と判別ルール獲得の手順 ALG3.2 を適用して判別ルール DR を獲得する。

step3 Step1 で獲得した contig 群に Step2 で獲得した DR を適用し、選別された contig 群を出力結果 C とする。

DAwH に判別ルールを適用した DAwCC を実行することで、従来のアセンブリを比べ被覆率に加え信頼性の高い結合配列の獲得が期待できる。

4.6 DAwCC の有効性検証のための性能比較実験

DAwCC のアセンブリの有効性を検証するための性能比較実験の結果について述べる。はじめに判別ルールの学習データ自身に対する学習能力を観察し、次にルール獲得に用いた学習データのアセンブリ結果へ直接判別ルールを適用し、判別ルールとしての有効性を検証する。最後に判別ルールを、試験データを用いたダブルアセンブリに適用し、従来の k-mer を用いたアセンブリの結果と性能比較をおこなう。

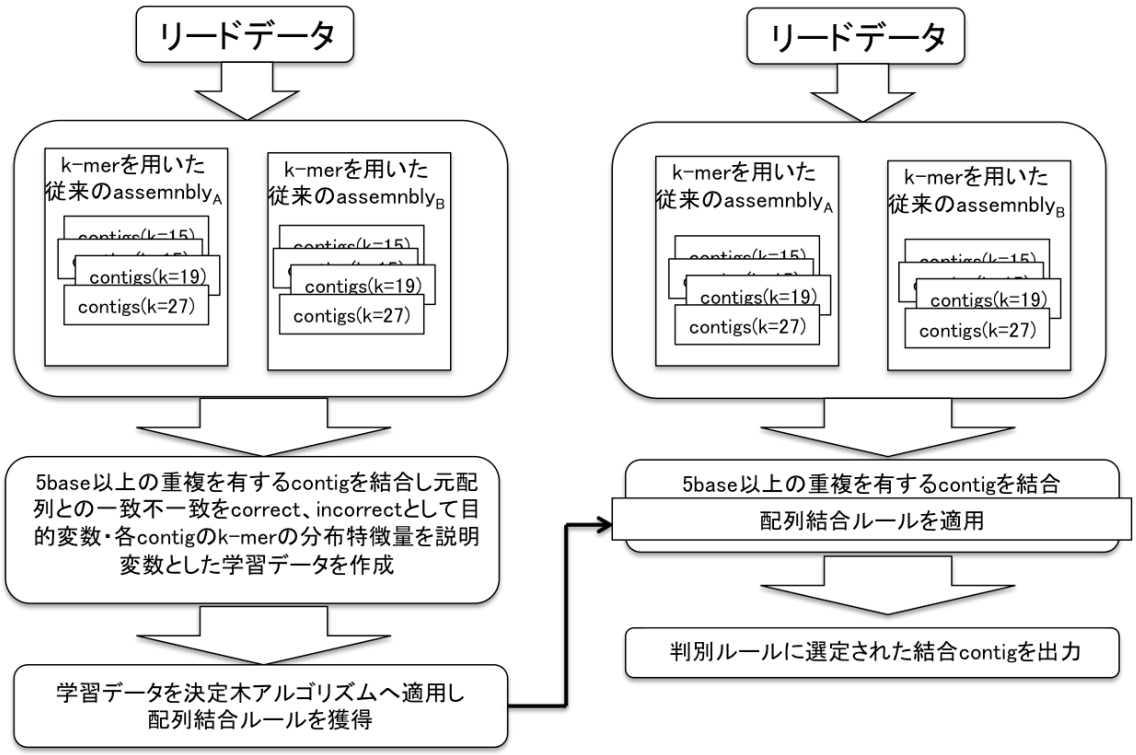

図 4.7: 判別ルールを適用したダブルアセンブリの流れ

4.6.1　性能比較に用いた評価指標

判別ルールの学習能力の検証には表3.1と式3.2を、判別ルールのアセンブリへの判別能力の検証には 4.2 と同様の評価値を用いた。判別ルールの学習データに対する有用性や試験データに対する有用性、従来のアセンブリ手法との比較結果について述べる。はじめに判別ルールの獲得元となった、学習データを用いたダブルアセンブリへの、ルールの効果について検証する。

4.6.2　実験データと判別ルール

実験データとして、元配列には表4.1と同様のデータを利用し、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4.5のような複数の k 値、従来アセンブリ手法の組み合わせとした。

ダブルアセンブリにおける判別ルールの有用性を検証するためには、試験データのみならず、ルールの獲得元である学習データにおける判別能力を検証する必要がある。獲得した結合ルールの、学習データに対する判別結果と LeR を表4.6に示す。

6 本の正結合配列と 19 本の誤結合配列に誤判別が発生してており、これらの contig は本来の決定木アルゴリズムによるルールにて判定が困難な contig であると考えられる。判

表 4.5: 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		試験用データ	
手法	ABySS	Velvet	ABySS	Velvet
k 値	15,17	17,19	16,18	15,17

表 4.6: 判別ルールの学習データへの学習能力

	有効	無効	LeR
正結合	531	6	0.971
誤結合	19	292	

別ルールの生成アルゴリズムの拡張による性能改善が課題である。しかし一部の判別ミスはあるものの、97 %以上の結合配列に対して正しく分類できたことがわかる。

学習データに対する decision tree の適用により、4 つの正結合ルール、18 の誤結合ルールを獲得した。ルールの内容一覧を表 4.7 に示す。

C4.5 は、決定木生成に用いられた説明変数の情報を参照することもできる。判別ルール生成の過程で多く用いられた説明変数を、引用率の高い順から表4.8に示す。

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主にルールの生成に用いられていることから、配列結合に有効な評価指標であることがわかった。

表 4.7: 獲得した結合ルール

正結合 ルール	
rule1	U^f_{freq} <= 3.3
rule2	W^f <= 0.7, Q^l <= 0.2 , U^f_{freq} > 12 and U^l_{freq} > 4.8
rule3	$W^{f,l}$ > 1983.868, U^f_{freq} > 3752.9
rule4	W^l <=1983.868
誤結合 ルール	
rule5	W^l_{freq} > 1983.868 , U^f_{freq} > 3.3 , U^f_{freq} <= 3752.9 , L^l_{freq} <= 0.7
rule6	$\Phi^{f,l}_{\max}$ > 120142 , CCF_{freq} <= 911099.5
rule7	CCF_{freq} =< 911099.5, H> 4 , Q^l>0.2 , I^l =<0.2285485, U^f_{freq}>3.3,L^l_{freq} >− 5.5
rule8	CCF_{freq} =<6.56175e + 07, W^l> 5.823744 , W^l =<1983.868 , U^f_{freq}> 2275.2
rule9	$R^{f,l}$>− 0.7 , $R^{f,l}$ =<0.2 , Q^l=<0.2 , U^f_{freq} >3.3 , U^f_{freq}=<12 ,U^l_{freq} >6, U^l_{freq}=<9.5
rule10	$R^{f,l}$ >− 0.7 , $R^{f,l}$=<0.8 , Q^l=<0.2, W^l=<5.235331 , U^f_{freq}>3.3, U^f_{freq}=<12, L^f_{freq}> 4.8
rule11	CCF_{freq}>99, I^f=<0.218525, U^f_{freq}=<4.8 , L^f_{freq}=<0.3179367
rule12	D^f=<0 , Q^l>0.1, ρ=<2.15, W^f=<6.59273 , U^f_{freq}>3.3
rule13	CCF_{freq}=<87884.5 , Q^l=<0.2 , W^l>15.59678 , U^f_{freq}> 3.3
rule14	$\Phi^{f,l}_{\max}$=<56.88169 , Q^l >0.2, Q^l =<0.5, U^f_{freq}>3.3 , L^f_{freq}> 6.3
rule15	$R^{f,l}$> 0.2 , I^l=<0.1666667, L^f_{freq}> 4.8
rule16	$\Phi^{f,l}_{\max}$=<133.7969 , Q^l=<0.2 , U^f_{freq}>3.3 , L^f_{freq} >15.5
rule17	CCF_{freq} =<400365 , D^f> 10 , U^f_{freq}> 3.3
rule18	$\Phi^{f,l}_{\max}$>19.95666 , CCF_{freq}>99 , ρ=<2.15 , W^f=<17.03178 , W^l=<4.8
rule19	CCF_{freq}=<87884.5 , Q^f> 0.7 , Q^l=<0.2 , ρ_{freq}=<0.6083328, U^l_{freq}>4.8
rule20	$\Phi^{f,l}_{\max}$=<133.7969 , Q^l=<0.2, I^f=<0.1764706, W^l> 7.45922 , U^f_{freq} > 3.3
rule21	$R^{f,l}$=<− 0.5, $\Phi^{f,l}_{\max}$> 56.88169, ρ_{freq}=<0.4219435 , U^f_{freq}>3.3 , U^l_{freq} > 4.8
rule22	$\Phi^{f,l}_{\max}$>5085.468, U^l_{freq}=<4.8

表 4.8: ルール獲得に引用された特徴パラメータ

99.29 %	W^{l}_{freq}
39.50 %	U^{f}_{freq}
22.05 %	U^{l}_{freq}
20.64 %	Q^{l}
18.75 %	CCF_{freq}
14.15 %	$\Phi^{f,l}_{\max}$
11.91 %	$R^{f,l}$
5.54 %	$L^{f,l}_{freq}$

4.6.3　学習データを用いたアセンブリ結果へのルールの有用性

判別ルールの有用性を検証するために、学習データへ直接ルールを適用し、その影響を観察する。

次にルール獲得元となった学習データに対してダブルアセンブリを行い、判別ルールを適用した結果について、結果前と後を比較することで検証する。判別ルールを適用したダブルアセンブリ、ルール適用前のダブルアセンブリの結果を、$CorR$、$CovR$ を用いて観測した結果を表 4.9 に示す。学習データに正結合ルールを適用した場合、$covR$ は 1 %減

表 4.9: 学習データのダブルアセンブリへのルールの効果

手法	$DAwH$	$DAwCC_{\text{正結合}}$	$DAwCC_{\text{誤結合}}$
出力数	848	586	392
正結合配列	537	376	370
誤結合配列	313	210	22
$CorR$	0.63	0.64	0.94
$covR$	1.0	0.999	0.999

少したが $CorR$ は 1 %改善したのに対し、誤結合ルールを適用した場合は $covR$ は 1 %減少したものの、$CorR$ は 30 %の改善が確認できた。

4.6.4　ターゲットデータを用いたアセンブリ結果へのルールの有用性

学習データへのルールの効果を検証できたので、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ダブルアセンブリ (DAwH) と、ルールを適用した場合について比較した結果を表 4.10に示す。

表 4.10: 結合ルールの試験データへの学習効果

手法	ABySS (k=15)	ABySS (k=19)	Velvet (k=17)	Velvet (k=21)	$DAwH$	$DAwCC_{\text{正結合}}$	$DAwCC_{\text{誤結合}}$
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
$CorR$	1.0	1.0	1.0	1.0	0.64	0.67	0.80
$covR$	0.93	0.76	0.98	0.98	1.0	1.0	1.0

単一の k-mer を用いる ABySS や Velvet の結果と比較した場合、$covR$ は k 値や手法によって異なっており、バラツキが見られるのに対し、ダブルアセンブリにおいては改善されていることがわかる。判別ルールを適用しない場合、正解率である $corR$ は 30 %以上減少しているのに対し、特に誤結合ルール適用後には 16 %改善されていることがわかる。しかし $DAwCC_{\text{正結合}}$ の $corR$ は 0.67 であることから、正結合ルールで誤って取得してしまった誤結合配列や取得できなかった正結合配列が残されていることがわかる。同様に $DAwCC_{\text{誤結合}}$ による $corR$ は 0.8 であることから、誤結合ルールで除去できなかった誤結合配列が残されていることがわかる。これらの課題は、配列結合の信頼性を示す評価指標を用いて正結合ルールまたは誤結合ルールを増やすことによる判別ルールの性能向上によって解決の余地がある。よって判別ルールを用いたダブルアセンブリにより、正解率を維持しつつ、特定の k 値、手法に依存しない頑健なアセンブリを実行できる可能性について確認することができた。

4.7　まとめ

用いる k 値や手法の影響を受けず、かつ被覆率を改善できるようなアセンブリを実行するために、複数の k-mer と手法の結果を統合した DAwH(Double Assembly with Heuristic) を提案した。k-mer を用いた従来手法との性能比較の結果、被覆率や結合の信頼性を示す正解率・最長正結合配列の改善の可能性を確認した。一方で誤結合配列の発生による正解率の減少が見られた。結合の信頼性を維持しつつダブルアセンブリをおこなうために、決定木アルゴリズムである C4.5 による判別ルールの獲得を検討した。k-mer を用いたアセンブリ手法により生成した contig を k-mer の coverage value の集合と見なし波形表現した場合、波形と結合の正誤に関連性が見られた。この情報を利用し、判別ルールの生成に contig 上における k-mer の coverage value の分布特徴量を用いた DAwCC(Double Assembly method with Contig for k-mer's coverage) を提案した。

提案手法との性能比較の対象として、単一の k 値を用いた従来手法である Velvet, ABySS を用いた。性能比較実験の結果より、正解率を維持しつつ、被覆率の改善・最長正結合配列を改善できるアセンブリの可能性を確認した。さらに判別ルールによる全ての正結合配列の取得・全ての誤結合配列の除去が完全に行われていないことから、判別ルールの性能改善により、さらなる正解率の改善が期待できる。

第5章　複合決定木によるルールを用いたDNAダブルアセンブリ

4 章にて提案した DAwCC では、結合の正誤を用いて学習データを生成し、決定木である C4.5 を用いて判別ルールを獲得してダブルアセンブリに適用した。従来のアセンブリ手法である Velvet や ABySS と比べ、長い正結合配列の生成が可能になり、判別ルールの適用により正解率も改善されたが、完全に取得できなかった正結合 contig や、除去できなかった誤結合配列が残されていた。本章では、結合の正誤と関連した別の特徴量を目的変数とした、複数の決定木により構成される複合決定木を提案した。さらに複合決定木より正結合ルールと、長い誤結合配列の除去が可能な誤結合ルールを抽出し適用したダブルアセンブリを提案した。検証実験の結果、複合決定木より取得される判別ルールは、従来の決定木によるルールに比べ判別能力が改善し、それによりダブルアセンブリの精度もさらに改善される可能性が高いことを確認した。

5.1　複合決定木の評価指標の選択

5.1.1　複数の目的変数の適用

本節では、複数の特徴量の目的変数としての利用による判別ルールの精度改善の可能性について述べる。前章では判別ルールの獲得の目的変数として結合配列の正誤を用いた。本節では、結合配列の正誤に加え複数の目的変数を用いることで決定木を増やし、正結合配列の取得率・誤結合配列の除去率の改善を目指す。はじめに結合正誤との関連性が強く、結合の精度の指標として用いる変数を列挙し、それらの目的変数としての適性を検証した。次にこれらを用いた判別器生成のための手法を述べた。最後に結合正誤のみ用いた場合と比べ配列結合決定の精度の改善の可能性があるかについて検証した。

5.1.2　被覆最小値・重複長と結合正誤の関係

2 章でも述べたように、k-mer を用いた従来のアセンブリ 手法である ABySS や Velvet では、coverage value の値が著しく小さい k-mer を「シークエンサーによる読み取りミスを含む」と見なし hash table から除去するという手続きを設けていた。またシークエンサーデータへの読み取りエラー補正のアプローチとして ECHO や EDAR では k-mer の coverage value の出現頻度の分布を用いてエラー部位の推定を行っていた。配列の信頼性を維持する

ための手続きである。このようなアプローチに従って考慮した場合、k-mer coverage value の最小値 (被覆最小値) が著しく小さい contig をダブルアセンブリに用いると結合配列は誤結合である可能性が高いと考えられる。また SGA などの読み取り配列間の重複部位の長さ優先探索や GAA などの配列間のアラインメントスコアの重み付けなどの観点を取り入れ、本研究では配列間の重複長が短いほど、誤結合を発生させる可能性が高いと考えた。それらの指標としての可能性の検証のため、被覆最小値、重複長の値と結合配列の正誤の分布の関連性について調べた。図5.1は被覆最小値の値と結合配列の正誤の分布を、図5.2は重複長の値と結合配列の正誤の分布を示し、横軸は被覆最小値、重複長の値を、縦軸はそれらのカーネル密度を表す。

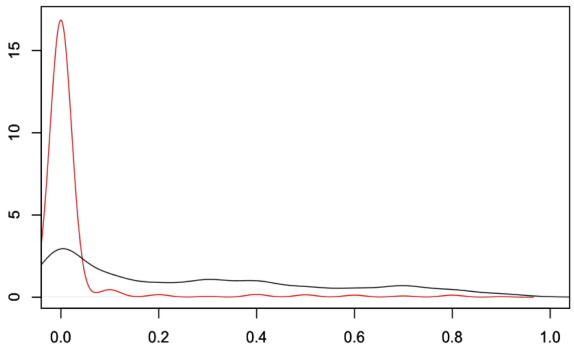

図 5.1: 被覆最小値と結合正誤の分布

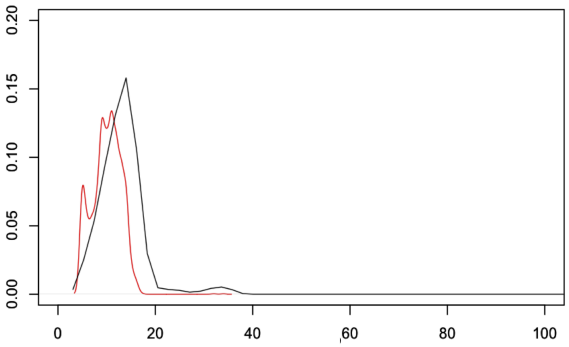

図 5.2: 重複長と結合正誤の分布

被覆最小値については 0∼0.1 の範囲で誤結合 contigs が偏って発生していること、図5.2より重複長については 5∼8base の範囲に僅かではあるが誤結合配列の偏りが発生していることが確認できる。被覆最小値、重複長の正結合群における分散は大きいが、誤結合配列群に偏りの特徴が見られることから、誤結合配列群における被覆最小値、重複長の特徴を利用することで、正結合配列群に混入した誤結合配列の除去が期待できる。よって被覆最小値、重複長の正結合配列の判別ルールへの誤結合配列判別ルールの追加による、誤結合配列の除去を試みる。

精選された判別ルール獲得のため、学習データを生成する各説明変数は独立している必要がある。そのため一般的に、機械学習により判別器を獲得する場合には、各変数間の依存関係を調べ、相関関係のある変数を除去する。各変数間の依存関係を測定するために、相関関係を調べた。本研究における学習データを構成している説明変数は k-mer の coverage value の分布特徴量であるため、これらは正規分布に従わないという特徴をもつ。よって説明変数間の相関係数の算出には式5.1のように定義される、Kendall の順位相関係数を用いた。

P:あるペアの順序関係が 2 つのリストで一致している数

n:アイテムの個数

$$\tau = \frac{2P}{1/2n(n-1)} - 1 \tag{5.1}$$

表5.1に各説明変数間の相関係数を示す。

学習データ上における説明変数間の相関係数が0.9を超えた場合に、これらの変数は依存していると考えられる。表5.1において、CCF_{freq} と $\Phi_{\max}^{f,l}$ の相関係数が0.95であり、これらの相関関係は独立したルール生成の際にノイズとなってしまう可能性が高いため、一方を学習データから削除する必要がある。本研究では決定木生成における引用率の低い CCF_{freq} を除去した。

5.1.3　複数判別器の生成

被覆最小値、重複長の特徴を利用、つまり被覆最小値、重複長を推定するための手法を選択するためには、学習データの構成に対応した解析法を選択する必要がある。目的変数の候補である被覆最小値、重複長は量的変数であり、用いる説明変数がcontig上における k-merのcoverageの分布特徴量であることを考慮した場合、多変量解析手法の1つである重回帰分析への適用が妥当であると考えられる。重回帰分析では、複数個の説明変数 $\{x_1, x_2, \ldots x_i\}$ と目的変数 y の間に線形関係があることを仮定し式5.2のような回帰式の生成を目的とする。

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_i x_i + e \tag{5.2}$$

ここで a は切片、b は偏回帰係数、e は予測値 $\hat{y}$ と実測値 y との残差を表す。偏回帰係数は、各説明変数間の相互の影響を取り除いた際の、目的変数への影響を示す。また重回帰式のモデルへのあてはまりの良さを示す予測精度として、$\hat{y}$ と y の相関係数である重相関係数 R、表現した際の偏差平方和における予測値の平方和の割合を示す決定係数 R^2 が定義されており、1.0に近いほど予測精度が高いことを示す。

目的変数の予測に適切な説明変数を選択することで、予測値と実測値の誤差の少ない回帰式を生成することができる。各説明変数のばらつきによる変数間の影響を取り除いた偏相関係数を用いることで目的変数を予測する。また複数の回帰式が生成された場合、回帰モデルの選択基準値であるAIC(Akaike's Information Criterion)[79]によって回帰式を選択する。回帰式獲得のため、はじめに従来手法により獲得したcontigの被覆最小値、重複長を目的変数、contigにおける k-merの分布特徴量を説明変数とした学習データを生成し、重回帰分析への適用により回帰式を獲得した。決定係数のとる最大値は1.0であり、1.0に近い値であるほど回帰式としての予測能力が高いといえる。被覆最小値、重複長それぞれについて生成した回帰式と獲得した決定係数・AICを表5.2に示す。

被覆最小値の予測式の決定係数が0.01、重複長の予測式の決定係数が0.2と適合性が低い値をとっていることから、被覆最小値や重複長の線形表現が困難であることがわかる。図5.1～図5.2に示した被覆最小値・重複長と結合配列の正結合、誤結合の分布より、分散が大きいことが原因であると考えられる。目的変数には分散の大きい量的変数ではなく、分散の小さい質的変数の適用により頑健な判別が可能と考えられる。そこで被覆最小値・重複長の質的変換により質的変数として取り扱うことについて考えた。

前節においてルールの獲得のために目的変数として用いた正結合・誤結合の2値表現である結合正誤に従い被覆最小値・重複長の2値で表現するために、本研究ではC4.5を用

表 5.1: 各説明変数間のケンドールの順位相関係数

	D^f	D^l	I^f	I^l	U^f_{freq}	U^l_{freq}	L^f_{freq}	L^l_{freq}	Q^f	Q^l	W^f	W^l	ρ	ρ_{freq}	CCF_{freq}	$\Phi^{f,l}_{\max}$	H	$R^{f,l}$
D^f	1.00	-0.09	0.22	-0.02	0.00	0.02	-0.04	0.03	0.02	0.02	-0.04	0.05	-0.07	0.03	0.02	0.03	0.06	0.05
D^l		1.00	-0.03	0.25	0.00	0.04	0.01	-0.01	0.02	0.07	0.01	-0.04	0.05	0.08	0.05	0.05	-0.07	0.01
I^f			1.00	0.05	-0.37	-0.08	0.07	-0.02	-0.33	-0.11	-0.39	-0.07	-0.13	-0.21	-0.28	-0.27	0.12	0.14
I^l				1.00	-0.11	-0.27	0.07	0.02	-0.09	-0.24	-0.12	-0.27	-0.10	-0.18	-0.25	-0.24	0.06	0.04
U^f_{freq}					1.00	0.12	-0.16	-0.02	0.69	0.16	0.86	0.14	0.15	0.51	0.54	0.55	-0.20	-0.15
U^l_{freq}						1.00	-0.03	-0.20	0.13	0.72	0.13	0.80	0.16	0.59	0.57	0.58	-0.22	-0.16
L^f_{freq}							1.00	0.02	-0.11	-0.04	-0.18	-0.03	-0.04	-0.11	-0.14	-0.13	0.07	-0.01
L^l_{freq}								1.00	0.03	-0.17	-0.02	-0.22	-0.01	-0.16	-0.12	-0.13	0.07	0.02
Q^f									1.00	0.19	0.65	0.15	0.17	0.45	0.49	0.50	-0.18	-0.16
Q^l										1.00	0.17	0.63	0.17	0.57	0.54	0.55	-0.23	-0.17
W^f											1.00	0.14	0.15	0.49	0.53	0.54	-0.20	-0.17
W^l												1.00	0.14	0.53	0.52	0.52	-0.20	-0.18
ρ													1.00	0.22	0.23	0.19	-0.63	-0.12
ρ_{freq}														1.00	0.74	0.77	-0.30	-0.21
CCF_{freq}															1.00	0.95	-0.29	-0.20
$\Phi^{f,l}_{\max}$																1.00	-0.26	-0.20
H																	1.00	0.14
$R^{f,l}$																		1.00

表 5.2: 被覆最小値 、 重複長の判別式と判別能力

	$variables$	$Coef_{reg}$	決定係数	$Coef_{det}$
被覆最小値	$\Phi^{f,l}_{\max}$	-6.230×10^{-6}	0.012	0.009
	ρ	5.141×10^{-6}		
	Intercept	1.191×10^{1}		
重複長	$\Phi^{f,l}_{\max}$	5.558×10^{-8}		
	D^l	-4.299×10^{3}		
	I^f	-3.893×10^{-1}		
	I^l	-3.242×10^{-1}	0.214	0.203
	W^f	6.061×10^{-5}		
	W^l	7.211×10^{-5}		
	U^f_{freq}	-1.409×10^{-4}		
	U^l_{freq}	-1.700×10^{-4}		
	Intercept	4.461×10^{-1}		

いた。結合正誤を目的変数、被覆最小値と重複長を説明変数とした学習データの C4.5 への適用により、量的変数である被覆最小値と重複長の正結合・誤結合への質的変換が可能である。質的変換の手順を以下に示す。

被覆最小値・重複長の結合正誤変換による判別ルール生成の手順

[入力] アセンブリ対象のリードデータと生物種が同様の既知配列 G_{sim}

[出力] 結合正誤、被覆最小値、重複長を用いて生成した判別ルール $DT_{cov}, DT_{ovlp}, DT_{accu}$

step1 ALG3.2:アセンブリに向けた学習データ生成と判別ルール獲得の手順 に従い被覆最小値, 重複長を説明変数、結合正誤を目的変数とした学習データ D_{obj} 生成

step2 D_{obj} に決定木アルゴリズム C4.5 を適用し、獲得した結合正誤のルールに従い被覆最小値, 重複長の値を結語正誤へ変換

step3 結合正誤、被覆最小値、重複長を目的変数、k-mer の特徴量を説明変数とした学習データ D_{cov}、D_{ovlp}、D_{accu} より判別ルール DT_{cov}、DT_{ovlp}、DT_{accu} 獲得

被覆最小値、重複長を結合正誤へ質的変換し目的変数とすることで、重回帰式に比べ精度の高い判別器の生成が期待できる。上記の手順に従い獲得した判別ルールを表5.3に示す。

表 5.3: 重複長、被覆最小値による判別ルールと判別能力

ruleID	ルール	判別制度
rule1	被覆最小値 > 0　−>　正結合	0.996
rule2	重複長 > 14　−>　正結合	0.933
rule3	重複長 <= 14 and 被覆最小値 <= 0　−>　誤結合	0.728

表5.3に従い、0.1 以上の値を持つ被覆最小値、15 以上の値をもつ重複長を正結合、0 以下の値をもつ被覆最小値、14 以下の値をもつ重複長を誤結合に変換した。このようにして正結合、誤結合に質的変換された被覆最小値・重複長を目的変数、contig における k-mer の分布特徴量を説明変数とした学習データを生成し、C4.5 への適用により判別ルールを生成した。また回帰式における決定係数同様、獲得した判別ルールの判別能力を評価した。被覆最小値・重複長によるルールの判別能力を表5.4∼ 表5.5に示す。

表 5.4: 被覆最小値の判別能力

	有効	無効	LeR
正結合	394	3	0.937
誤結合	14	236	

表 5.5: 重複長の判別能力

	有効	無効	LeR
正結合	400	8	0.951
誤結合	24	215	

被覆最小値についての判別能力は 93.7 %、重複長についての判別能力は 95.1 % と、重回帰式を大きく上回る高い判別能力を持つことがわかった。以上より、本研究では結合正誤への被覆最小値・重複長についての決定木の統合をおこなう。

5.1.4　複数の目的変数のルールによる正判別分布

前節では、従来の判別ルール取得に用いた特徴量である結合正誤に加え、被覆最小値、重複長を用いた判別ルールの獲得の可能性について検証した。本節では、被覆最小値、重複長についての判別ルールの追加により、結合正誤についての判別ルールのみ用いた場合と比較し正判別できる結合配列の数がどれほど増えるかについて検証した結果について述べる。各結合正誤、被覆最小値、重複長を目的変数として獲得した判別ルールを同一の試験データに適用したとき、正しく判別できた結合配列と該当する目的変数へ分類し、分布を観察した。結合正誤、被覆最小値、重複長の各決定木より獲得した正結合、誤結合ルールの正判別分布を図5.3∼ 図5.4 に示す。

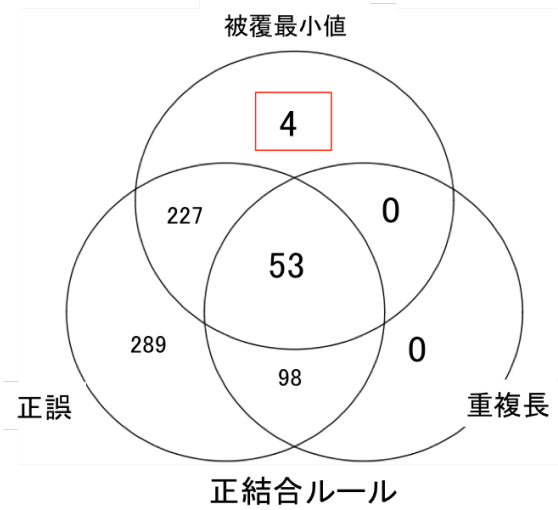

図 5.3: 正結合ルール による正判別分布

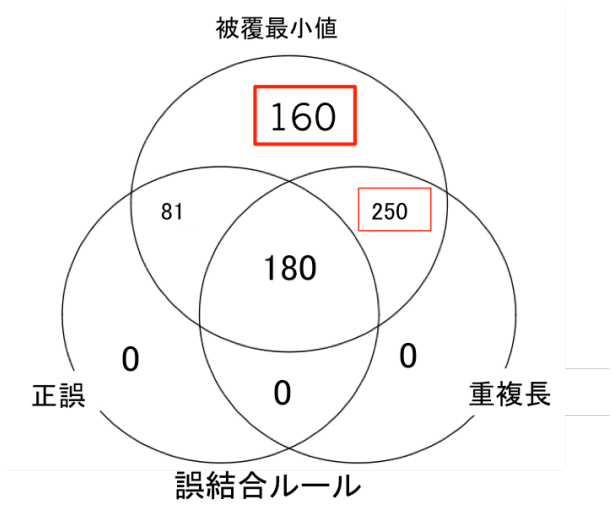

図 5.4: 誤結合ルール による正判別分布

図5.3において、例えば 53 本の正結合配列は結合正誤、被覆最小値、重複長全ての正結合ルールにより正しく正結合と判別され、4 本は被覆最小値によるルールのみ正しく判別でき、227 本は結合正誤、被覆最小値、の正結合ルールにより正しく判別できた結合配列の数を意味する。同様に図5.4において、160 本の誤結合配列は被覆最小値による誤結合ルールのみにより正しく誤結合と判別され、250 本は被覆最小値、重複長により正しく判別されたが結合正誤によるルールでは判別できなかった数を示す。よって図5.3～ 図5.4より、従来用いた結合正誤のみによるルールでは正しく判別できなかったが被覆最小値、重複長により正しく判別できる結合配列が増える、つまり複数の目的変数を用いて結合ルールとして用いることによる、判別能力の改善の可能性が確認できた。

5.1.5　複合決定木の生成

アセンブリ適用に向けた、複数の目的変数を用いた複合決定木生成法を提案する。

ALG5.1:アセンブリ適用に向けた複合決定木生成と判別ルール獲得の手順

[入力] アセンブリ対象のリードデータと生物種が同様の既知配列 G_{sim}

特徴量 $F = f_1, f_2, \cdots, f_n$

機械学習アルゴリズム ML

[出力] ALG5.1 により生成された決定木 $DT_{cov}, DT_{ovlp}, DT_{accu}$

Step1 ALG3.2 アセンブリに向けた学習データ生成と判別ルール獲得の手順に従い DAwH を用い、ML=C4.5、F=k-mer の coverage value の分布特徴量として結合正誤のルールを獲得

Step2 被覆最小値、重複長 を説明変数、結合正誤を目的変数とした学習データに C4.5 を適用し結合正誤のルールを獲得

Step3 Step2 で獲得したルールに従い重複長、被覆最小値を正誤へ質的変換

Step4 正誤、重複長、被覆最小値を目的変数、F を説明変数とした学習データに C4.5 を適用し 3 つの決定木を獲得

共通の説明変数を持つ目的変数、つまり決定木が 1 本から 3 本に増えることでルールが追加され、従来の決定木からは獲得できなかった判別ルールの生成が期待できる。

5.1.6　誤結合ルール適用による二段階判別の検討

正結合ルール適用結果への誤結合ルール追加による、二段階判別による精度改善の可能性について述べる。重複長や被覆最小値と配列の正結合、誤結合の分布図5.1～図5.2に示したように、誤結合ルールの追加により、正結合配列群に混入した誤結合配列を除去できる可能性がある。そこで、各誤結合ルールを正結合ルールへ追加し実際にアセンブリを実行し結合配列を判別することで、正解率などの性能が向上するかについて調べた。DNA 配列のアセンブリにおいて、長い誤結合配列は、獲得 contig において最も大きなノイズとして扱われる。また従来の判別ルールの課題として、長い正結合配列のみでなく長い誤結合配列も生成したことが課題となっていることから、本研究では特に長い誤結合配列を正しく除去できるような誤結合ルールを有用であると見なす。

結合正誤、重複長、被覆最小値について獲得した正結合ルールに各誤結合ルールを正誤#、重複長#、被覆最小値#として追加しアセンブリに適用し、評価指標の変化を観察した。今回用いた実験データからは 14 本の正誤、9 本の重複長、6 本の被覆最小値についての誤結合ルールを獲得した。これらを正結合ルールの適用結果に 1 本ずつ追加し、誤結合ルール追加前後の結合配列の出力数#.Output、正結合配列数#.Correct、正解率 $corR$、最長誤結合配列長 ML.Incorr の変化を観察した結果を表5.6に示す。

表 5.6: 各誤結合ルールを追加した場合のアセンブリ性能の変化

誤結合ルール	#.Output	#.Correct	$corR$	N50	$covR$	ML.Incorr
正結合ルールのみ	512	**350**	0.68	7356	0.96	15767
正誤 1	508	**350**	0.68	7356	0.96	15767
正誤 2	508	**350**	0.68	7356	0.96	15767
正誤 3	502	349	0.69	7356	0.96	15767
正誤 4	481	331	0.68	7357	0.96	15767
正誤 5	509	**350**	0.68	7357	0.96	15767
正誤 6	506	345	0.68	7357	0.96	15767
正誤 7	508	348	0.68	6585	0.96	15767
正誤 8	500	343	0.686	7356	0.96	15767
正誤 9	510	349	0.68	7356	0.96	15767
正誤 10	499	340	0.68	7356	0.96	15767
正誤 11	501	342	0.68	7356	0.96	15767
正誤 12	512	350	0.68	7356	0.96	15767
正誤 13	509	349	0.68	7356	0.96	15767
正誤 14	502	343	0.683	7356	0.96	15767
重複長 1	430	282	0.65	7357	0.96	15767
重複長 2	491	339	0.69	5657	0.96	15767
重複長 3	442	305	0.69	7356	0.96	15767
重複長 4	468	326	0.696	5678	0.96	15767
重複長 5	236	194	**0.82**	7891	0.96	**10860**
重複長 6	487	333	0.68	6814	0.96	15767
重複長 7	498	**350**	0.70	5658	0.96	15767
重複長 8	478	330	0.69	6585	0.96	15767
重複長 9	102	62	0.60	3989	0.762	10872
被覆最小値 1	453	300	0.66	6316	0.96	15767
被覆最小値 2	494	**350**	0.708	7357	0.96	15767
被覆最小値 3	507	348	0.68	6585	0.96	15767
被覆最小値 4	500	342	0.684	7356	0.96	15767
被覆最小値 5	512	350	0.68	7356	0.96	15767
被覆最小値 6	59	50	**0.847**	7899	0.81	**10860**

誤結合ルールとして正誤1、正誤2、正誤5および重複長7、被覆最小値2を適用した場合、#.Outputの数は減少したが#.Correctの値は350本と変化していないことより、正結合配列を残しながら誤結合配列のみ削除できたことがわかる。それに対し正誤12や被覆最小値5は正結合ルールのみ用いた場合と#.Output、#.Correctに変化がないことから、誤結合配列を削除できていない。また重複長5や被覆最小値6の適用時には$corR$が0.68から0.82または0.847と20％の改善が見られ、ML.Incorrについても15767から10860と長い誤結合contigが除去されていることより、誤結合ルール適用によるアセンブリの性能向上の可能性が確認できた。4章ではアセンブリに正結合ルールと誤結合ルールを別途適用したが、正結合ルールの適用結果に対し誤結合ルールの組合せを工夫し適用することで、アセンブリの性能がさらに改善されることが期待できる。

5.2　複合決定木による判別ルールを用いたDNAダブルアセンブリ

複合決定木による判別ルールを用いたダブルアセンブリDAwCDT(Double Assembly with Comparative Decision Tree)を提案する。以下に手順を、図5.2にその流れを示す。

ALG5.2:複合決定木による判別ルールを用いたダブルアセンブリ DAwCDT

[入力] リードデータ $R = \{r_1, r_2, \cdots, r_n\}$

[出力] ALG5.2により生成されたcontig群 $C = \{c_1, c_2, \cdots, c_n\}$

Step1 ALG5.1:アセンブリ適用に向けた複合決定木生成と判別ルール獲得の手順に従いDAwHを用い、ML=C4.5、F=k-merのcoverage valueの分布特徴量として結合正誤のルールDR^{CDT}を獲得

Step2 DR^{CDT}より長い誤結合contigを除去できる誤結合ルールDR^{CDT}_{inco}を抽出

Step3 Step2で獲得した誤結合ルールと全ての正結合ルールDR^{CDT}_{cor}を、複合決定木より獲得した判別ルールDR_{CDT}とする

Step4 複数のk-merと手法の統合によるダブルアセンブリDAwH: ALG4.1に従い獲得したcontigにDR^{CDT}_{cor}を適用しルールを満たすcontigを獲得

Step5 Step4で獲得したcontigにDR^{CDT}_{inco}を適用し誤結合配列を除去し出力

目的変数を1変数から3変数に増やすことで、これまで判別できなかった結合contigへの対応が可能になり判別能力が改善され、結果としてそれを適用したアセンブリの性能も改善されることが期待できる。

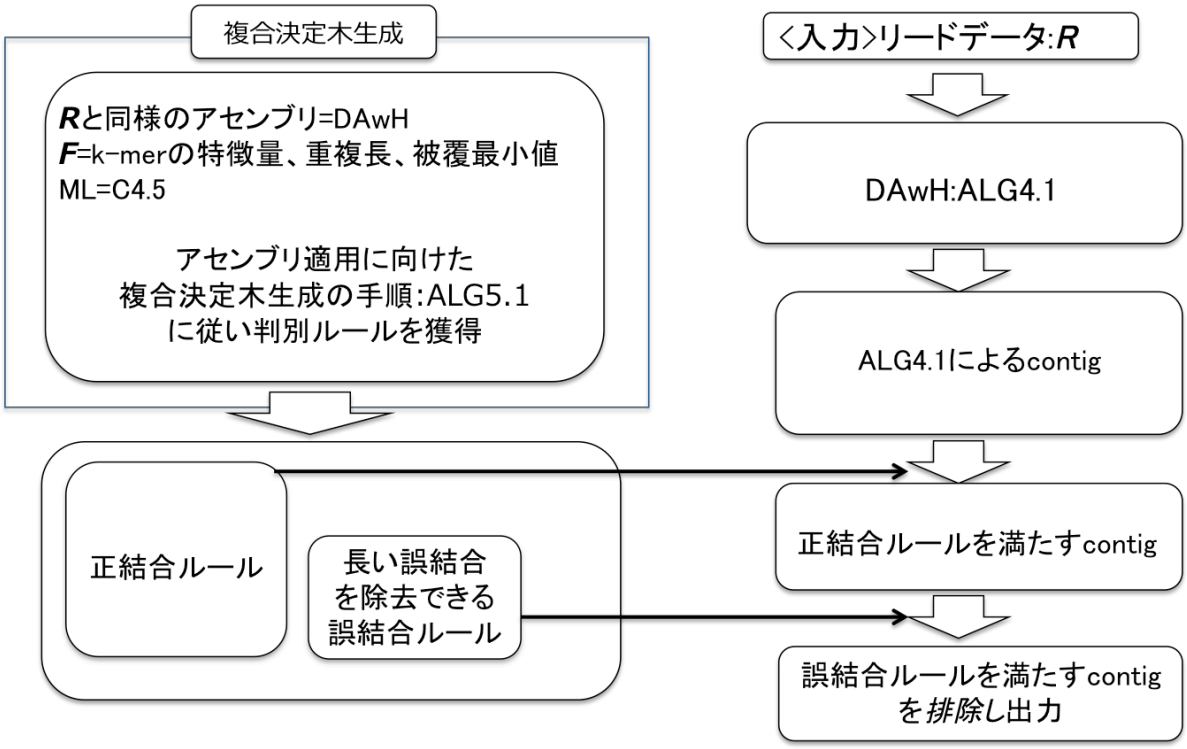

図 5.5: 複合決定木による判別ルールを用いた DNA ダブルアセンブリの手続き

5.3　従来手法との性能比較実験

複合決定木による判別ルールを用いたダブルアセンブリの有効性を検証するため、従来のアセンブリ手法、従来の決定木アルゴリズムによる判別ルールとの性能比較を、同一の実験データを用いておこなった。本節では比較実験に用いた評価指標、配列データについて述べ、比較結果を示す。

5.3.1　実験に用いたデータと k 値・手法

実験データとして、元配列には表4.1と同様のデータを利用し、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5.7のような複数の k 値、従来アセンブリ手法の組み合わせとし、異なるリードを用いておこなった。

5.3.2　性能評価に用いる指標

複合決定木による判別ルールの判別能力としての性能、判別ルールを適用したダブルアセンブリより出力された結合 contig の、被覆率や正解率、長さといった、アセンブリの性能を検証するため、他手法との比較実験をおこなった。決定木の性能比較の指標として、

表 5.7: 学習データ及び試験データ生成に用いた k 値と従来アセンブリ手法

	Train and test data	
Combination of Method	ABySS	Velvet
Combination of k-mer	16,18	15,17

4 章にて定義した評価指標に加え、獲得できた正結合配列の数である#.Correct、最長誤結合配列長である ML.Incorr、N50 を利用した。また従来手法による contig を統合するダブルアセンブリには、従来手法と比べ長い contig の獲得が期待される。また、図 5.6 に示すように、ダブルアセンブリによる contig は出力の約 60 %が 1500base 以上である。

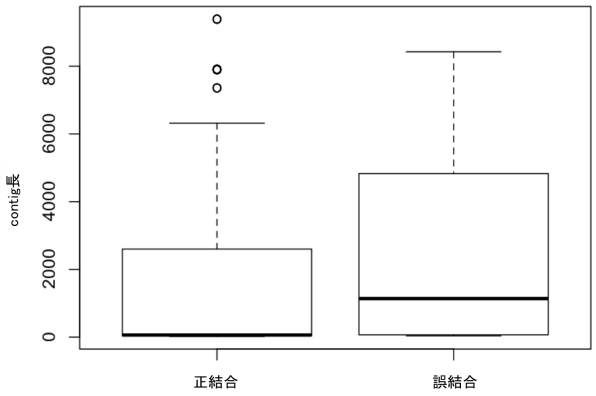

図 5.6: ダブルアセンブリにより生成された contig 長と正誤の分布

よって長い contig 群つまり 1500base 以上の contig 群における正解率を $corR_{1500}$、$covR_{1500}$ とし評価指標として追加した。

5.3.3　誤結合ルールの選択

今回利用した実験データにより 36 の正結合ルール、32 の誤結合ルールを獲得した。従来の決定木より結合正誤のみを用いて判別ルールを獲得した場合に比べ、正結合ルールについては 27 本、誤結合ルールは 15 本多く獲得することができた。正結合ルール一覧を表 5.8に、誤結合ルール一覧を表5.9に示す。

アセンブリに適用する誤結合ルールについては、今回用いたリファレンス配列長が 30000base であることを考慮し、10000base 以上の誤結合配列を除去できた誤結合ルールを有用と見なした。有用な誤結合ルールと削除された誤結合配列長の一覧を表5.10に示す。

特に、複数の長い誤結合配列の削除が可能な誤結合ルールを有効な誤結合ルールと考えることができる。重複長 5 は 4 本、重複長 9 は 6 本、被覆最小値 1 は 2 本削除できていることから、これらを組み合わせて複数の誤結合ルールを複合決定木として適用することとし

表 5.8: 複合決定木より獲得された正結合ルール一覧

Rule ID	ルールの内容
重複長 1	$D^l \leq -10, \rho \leq 2.91$
重複長 2	$33 < \Phi_{\text{freq}}, D^f \leq 2, 0.6 < U^l_{\text{freq}}, U^l_{\text{freq}} \leq 5.1, L^l_{\text{freq}} \leq -0.10$
重複長 3	$\rho \leq 2.91, 1488.1 < U^l_{\text{freq}}, -0.10 < L^l_{\text{freq}}$
重複長 4	$D^f \leq 2, Q^f \leq 0.2, 0.63 < \rho, \rho \leq 0.96$, $-0.10 < L^l_{\text{freq}}$
重複長 5	$8 < H, 0.96 < \rho, \rho \leq 2.91, L^l_{\text{freq}} \leq 0.503$
重複長 6	$\rho \leq 0.28, -0.10 < L^l_{\text{freq}}$
重複長 7	$D^f \leq 2, Q^f \leq 0.2, \rho \leq 0.96, I^l \leq 0.17, L^l_{\text{freq}} \leq -0.10$
重複長 8	$Q^l < 0.5, U^f_{\text{freq}} \leq 1.3, 5.1 < U^l_{\text{freq}}, L^l_{\text{freq}} \leq -0.10$
重複長 9	$Q^l \leq 0.5, \rho \leq 2.91, L^l_{\text{freq}} \leq -0.10$
重複長 10	$D^f \leq 2, \rho \leq 0.96, -0.10 < L^l_{\text{freq}}$
被覆最小値 1	$0.1 < Q^l, \rho \leq 5.52, I^l \leq 0.38, 8.003 < W^f_F, L^f_{\text{freq}} \leq 9.5$
被覆最小値 2	$R \leq 0.1, 643.86 < W^f_F >, 708.6 < U^f_{\text{freq}}$
被覆最小値 3	$-2 < D^f, 0.1 < Q^l, \rho \leq 5.52, I^l \leq 0.384, 8.00 < W_F$
被覆最小値 4	$\rho_{\text{freq}} \leq 380631, 2.83 < W^f_F, 566.37 < W^l_F, L^l_{\text{freq}} \leq 1490.2$
被覆最小値 5	$R \leq 0.3, D^f \leq 1, 0.1 < Q^l, Q^l \leq 0.3, 1.12 < \rho, 5.52 < \rho, I^l \leq 0.38$
被覆最小値 6	$4.20 < \rho_{\text{freq}}, D^f \leq 1, \rho \leq 1.12, I^f \leq 0.5, I^l \leq 0.5, U^f_{\text{freq}} \leq 401.4$
被覆最小値 7	$1.22 < \rho, I^l \leq 0.38, 9.6 < U^f_{\text{freq}}, U^l_{\text{freq}} \leq 4.2$
被覆最小値 8	$D^f \leq 1, H \leq 5, \rho \leq 1.12, I^f \leq 0.5, I^l \leq 0.5$
被覆最小値 9	$0.3 < R, R \leq 0.6, 0.01 < Q^l, I^l \leq 0.38, 2.83 < W^f, W^f \leq 8.00$
被覆最小値 10	$0.1 < Q^l, \rho \leq 5.528.003 < W^f, 6.6 < U^l_{\text{freq}}$
被覆最小値 11	$\Phi_{\text{freq}} < 898574, 566.37 < W^l, U^f_{\text{freq}} \leq 708.6, U^l_{\text{freq}}\ leq 1490.2$
被覆最小値 12	$7 < H, 0.1 < Q^l, \rho \leq 5.52, I^l \leq 0.38, 2.83 < W^f$
被覆最小値 13	$0.8 < Q^l, 5.52 < \rho, W^f \leq 643.86, U^l_{\text{freq}} \leq 130$
被覆最小値 14	$D^l \leq 1, \rho \leq 6.55, 0.38 < I^l, I^l \leq 0.5, 2.83 < W^f$
被覆最小値 15	$I^f \leq 0.125, I^l \leq 0.5, W^f \leq 2.83, W^l \leq 25.3$
被覆最小値 16	$H \leq 2, 0.38 < I^l$
被覆最小値 17	$0.2 < Q^l, I^l \leq 0.38, 643.9 < W^f$
正誤 1	$R \leq 0.1, 708.6 < U^f_{\text{freq}}$
正誤 2	$\leq 0.1, 898574 < \Phi_{\text{freq}}, U^l_{\text{freq}} \leq 1905.7$
正誤 3	$D^f \leq 1, D^l \leq 1, Q^l \leq 0.3, \rho \leq 2.91, U^f_{\text{freq}} \leq 401.4$
正誤 4	$\rho \leq 2.91, L^l_{\text{freq}} \leq -0.10$
正誤 5	$R \leq 0.1, 182.32 < \rho, I^f < 0.23$
正誤 6	$5 < H, 182.32 < \rho$
正誤 7	$0.1 < R, 2908.5 < \Phi_{\text{freq}}, \Phi_{\text{freq}} \leq 6787.5, 2.91 < \rho$
正誤 8	$0.6 < Q^f, 0.24 < I^l, W^l \leq 5.33$
正誤 9	$W^f \leq 331.4$

表 5.9: 複合決定木より獲得された誤結合ルール一覧

Rule ID	ルールの内容
重複長 1	$H \leq 8, 0.96 < \rho, \rho \leq 2.91, U^l_{\mathrm{freq}} \leq 1488.1, -0.10 < L^l_{\mathrm{freq}} > -0.10$
重複長 2	$2 < D^f, L^l_{\mathrm{freq}} \leq -0.7$
重複長 3	$1.3 < U^f_{\mathrm{freq}}, 5.1 < U^l_{\mathrm{freq}}, U^l_{\mathrm{freq}} \leq 14.9, L^l_{\mathrm{freq}} \leq -0.10$
重複長 4	$2 < D^f, -0.10 < L^l_{\mathrm{freq}}$
重複長 5	$2.91 < \rho$
重複長 6	$2 < D^f, 1 < D^l$
重複長 7	$7 < D^f$
重複長 8	$-10 < D^l, 0.5 < Q^l, L^l_{\mathrm{freq}} \leq 0.10$
重複長 9	$R \leq 0.2$
被覆最小値 1	$0.5 < I^l$
被覆最小値 2	$116387.2 < \Phi_{\mathrm{freq}}, \Phi_{\mathrm{freq}} \leq 898574$
被覆最小値 3	$0.1 < R, H \leq 5, 0.6 < Q^l, I^f \leq 0.227$
被覆最小値 4	$R \leq 0.6, D^f \leq -2, Q^f \leq 0.8, 9.5 < U^f_{\mathrm{freq}}, 4.2 < U^l_{\mathrm{freq}}$
被覆最小値 5	$\Phi_{\mathrm{freq}} \leq 4.2, 5 < H$
被覆最小値 6	$I^l \leq$0.5
正誤 1	$0.1 < R, 6787.5 < \Phi_{\mathrm{freq}}, 0.291 < \rho, \rho \leq 0.321, I^l \leq 0.24$
正誤 2	$0.1 < R, 6787.5 < \Phi_{\mathrm{freq}}, 0.6 < Q^f, \rho_{\mathrm{freq}} \leq 0.58, \rho \leq 0.182, 5.33 < W^l$
正誤 3	$0.1 < R, 6787.5 < \Phi_{\mathrm{freq}}, 0.6 < Q^f, -0.031 < \rho_{\mathrm{freq}}, \rho_{\mathrm{freq}} \leq 0.56, 0.291\rho, \rho \leq 0.182$
正誤 4	$R \leq 0.1, 489 < \Phi_{\mathrm{freq}}, -1 < D^f, 0.3 < I^f, W^l \leq 5464.053$
正誤 5	$56.78 < \Phi_{\mathrm{freq}}, D^l \leq 1, Q^l \leq 0.1, 0.291 < \rho$
正誤 6	$R \leq 0.1, W^l \leq 5464.053, U^f_{\mathrm{freq}} \leq 708.6, 1905.7 < U^l_{\mathrm{freq}}$
正誤 7	$R \leq 0.1, H \leq 5, 0.182 < \rho, I^f < 0.23$
正誤 8	$R \leq 0.1, \Phi_{\mathrm{freq}} \leq 898574, 331.414 < W^{f,l}, U^f_{\mathrm{freq}} \leq 708.6$
正誤 9	$-5 < D^l, 3 < H, 0.291 < \rho, 10.1 < U^f_{\mathrm{freq}}, U^f_{\mathrm{freq}} \leq 18.2$
正誤 10	$1 < D^f, \rho \leq 0.291, 4.21 < W^f, I^l_{\mathrm{freq}} \leq 7.6, -0.10 < L^l_{\mathrm{freq}}$
正誤 11	$1 < D^f, 3 < H, 0.1 < Q^l, 0.291 < \rho, U^f_{\mathrm{freq}} \leq 18.2$
正誤 12	$0.1 < R, \Phi_{\mathrm{freq}} \leq 2908.5, 0.291 < \rho, 19.155 < W^l$
正誤 13	$0.1 < R, \Phi_{\mathrm{freq}} \leq 2908.5, 1 < D^l, 0.291 < \rho$
正誤 14	$D^l \leq 1, 5 < H, Q^l \leq 0.3, 401.4 < U^f_{\mathrm{freq}}, -0.10 < L^l_{\mathrm{freq}}$
正誤 15	$1 < D^l, Q^f \leq 0.1, \rho \leq 0.291, 4.21 < W^f$
正誤 16	$0.3 < Q^l, \rho \leq 0.291, 4.214 < W^f$
正誤 17	$1 < D^l, \rho_{\mathrm{freq}} \leq -0.17, 4.21 < W^f, -0.101 < L^l_{\mathrm{freq}}$

表 5.10: 長い誤結合配列を除去できた判別ルール

除去された配列長	Rule ID
15767	重複長 5, 重複長 9, 被覆最小値 6
15767	重複長 4, 重複長 5, 重複長 6, 重複長 7, 重複長 9, 被覆最小値 6
12695	重複長 4, 重複長 5, 重複長 6, 重複長 7, 重複長 8, 重複長 9, 被覆最小値 6
10872	重複長 2, 重複長 4, 重複長 5, 重複長 7, 重複長 9, 被覆最小値 6
10860	重複長 2, 重複長 7, 重複長 9, 被覆最小値 1
10859	重複長 9, 被覆最小値 1

た。また長い誤結合配列を除去できた全ての誤結合ルールが重複長、被覆最小値の決定木によるものであることからも、複合決定木によるルールの判別能力の改善が確認できる。

5.4　性能比較結果

複合決定木によるルールの判別能力、さらにルールを適用したアセンブリの性能を検証するために、従来のアセンブリ、従来の決定木との性能比較をおこなった。比較の対象として従来のアセンブリである Velvet、ABySS、CISA、従来の決定木アルゴリズムを用いた判別ルールを適用したダブルアセンブリとの比較をおこなった。比較結果を表5.11に示す。

はじめにルールの判別能力について比較する。 #.Correct について、従来の決定木により結合正誤の情報のみを用いて判別ルールを適用したダブルアセンブリである $DAwCC_{cor}$ で取得した正結合配列の数が 234 本であるのに対し、複数の目的変数より構成される複合決定木による正結合ルールを適用したダブルアセンブリ $DAwCDT_{correct}$ の結果が 350 本であり、従来の決定木である C4.5 に比べ、抽出した正結合配列数が 116 本改善されたことがわかる。しかし、正解率である $corR$ は 0.68 まで減少したことから有用な誤結合ルール (重複長 5 + 重複長 2) を適用したところ、0.827 まで改善され、誤結合ルールの効果が確認された。ML.Incorr については誤結合ルール (重複長 5 + 重複長 9) を適用したダブルアセンブリ、また誤結合ルール (重複長 5 + 被覆最小値 9) を適用したダブルアセンブリにおいて、これらのルールを用いない場合に比べて長い誤結合配列を削除できていることから、有効な誤結合ルールを適用した効果が確認できる。長い contig 群における正解率である $corR_{1500}$ は 1.0 となっており、正結合配列のみ生成する複合決定木であることもわかる。よって決定木としての判別能力が C4.5 に比べ改善されたことが示された。しかしその反面、長い正結合配列を誤って削除した複合決定木もあり、ML.Corr が減少している結果も見受けられる。誤結合ルール (重複長 5 + 重複長 9) を適用したダブルアセンブリの結

表 5.11: 従来手法と複合決定木による判別ルールを用いたダブルアセンブリの性能比較

Method	#.Output	#.Correct	$corR$	N50	$covR$	ML.Incorr	ML.Corr	$corR_{1500}$	$covR_{1500}$
Velvet (k=15)	20	19	0.950	2963	0.980	34	4815	1.000	0.810
Velvet (k=17)	12	12	1.000	7889	0.980	-	10850	1.000	0.900
ABySS (k=16)	54	54	1.000	3048	0.870	-	4817	1.000	0.720
ABySS (k=18)	40	40	1.000	7891	0.770	-	10852	1.000	0.720
CISA	38	38	1.000	4044	0.980	-	10852	0.465	0.939
DAwH	671	412	0.614	7849	0.990	15767	18729	0.465	0.939
$DAwCC_{correct}$	338	234	0.690	7849	0.960	15767	18729	0.570	0.938
$DAwCC_{Incorrect}$	444	315	0.710	7906	0.990	15767	18729	0.610	0.940
$DAwCDT_{correct}$	512	315	0.680	7356	0.960	15767	18729	0.570	0.938
$DAwCDT_{correct}$+ 重複長 5 + 重複長 2	231	191	0.827	5631	0.960	10859	10854	0.890	0.908
$DAwCDT_{correct}$ + 重複長 5 + 重複長 9	43	33	0.767	10854	0.628	63	10855	1.000	0.618
$DAwCDT_{correct}$ + 重複長 5 + 被覆最小値 1	180	147	0.817	3148	0.603	1122	7892	1.000	0.550

果は 10855base と、従来のアセンブリ手法である Velvet や ABySS、CISA と比較すると改善されているが、従来の提案手法である $DAwCC$ と比較すると改悪の結果となっている。

次に、複合決定木による判別ルールを適用したダブルアセンブリの性能を比較した。従来のアセンブリである Velvet や ABySS は k の値や手法によって、被覆率である $covR$、$covR_{1500}$ にバラツキが見られるのに対し、従来の決定木による判別ルールを用いたダブルアセンブリの $covR$ は安定した値を維持している。Velvet や ABySS との比較結果より、長い contig 群における被覆率である $covR_{1500}$ は、複合決定木による正結合ルールを用いた場合 $DA^{CDT_{correct}}$ や誤結合ルール (重複長 5 + 重複長 2) 適用のダブルアセンブリにより改善された。よって複合決定木による正結合ルールと誤結合ルールの適用により、長い正結合配列を多く残せたことがわかった。本実験にて新たな評価指標被覆率 $covR_{1500}$ を用いたことで、Velvet や ABySS は長い contig の生成が困難であることもわかった。

5.5　まとめ

本章では判別ルールの性能改善という観点より、複数の目的変数の適用に着眼点をおいた。4 章にて目的変数として適用した結合の正誤を示す結合正誤に加え、従来のアセンブリにてリード結合の評価指標として用いられてきた k-mer coverage value の最小値 (被覆最小値)、重複長の適用について検討した。これらは量的変数であるため、目的変数として扱うため重回帰分析の適用を検討したが、量的変数の分散が大きく判別式の性能が低くなることがわかった。そこでこれら 2 変数を従来の決定木 C4.5 を用いて質的変換をおこない、結合正誤と同様に決定木を生成することとした。各 2 変数を目的変数として contig の判別をおこなう検証実験からは、結合正誤を用いた決定木による判別ルールでは対応できなかった contig を判別できていることから有効性も確認できた。さらに正結合ルールの判別結果に誤結合ルールを適用するという 2 段階判別により、正結合ルールにて誤って出力される誤結合配列を誤結合ルールにより削除できることも明らかになった。

以上より、1 つの学習データから複数の決定木を生成し、正結合ルール、誤結合ルールを組み合わせた判別ルールを用いたダブルアセンブリ DAwCDT(Double Assembly with Comparative Decision Tree) を提案した。

従来の決定木アルゴリズム、および従来のアセンブリとの性能比較の結果、従来の決定木 C4.5 に比べ正結合配列を多く出力し、アセンブリにおいて最も生成を避けたい長い誤結合配列の出力を防ぐことができた。さらにそれらを適用したアセンブリの性能については、従来のアセンブリに比べ正解率、長い contig 群における被覆率が改善されたことが確認され、判別ルール生成のための決定木として、そしてアセンブリ性能としての 2 点の見地において性能の改善を確認できた。

第6章　総括

本研究では、ギガシークエンサーによるDNA配列のアセンブリアルゴリズムの「頑健化・信頼性」改善による信頼性向上を目的とし、2つの手法を提案した。

1. 頑健化のために、複数のk-merと手法を統合し、k-merの特徴量利用の判別ルールを適用したダブルアセンブリ手法

2. 信頼性改善のために、複合決定木による判別ルールを適用したダブルアセンブリ手法

さらに上記の提案手法の有効性検証を、従来のアセンブリ手法および従来の機械学習アルゴリズムとの比較によりおこなった。

一般的にDNA配列は特殊な細胞採取作業により開始し、DNAシークエンシングと呼ばれる作業ののちシークエンサーに読み取られ、塩基配列データ(リード)として出力される。出力されたリードを結合し結合配列であるcontigを生成するまでの過程はアセンブリと呼ばれる。従来型と比べ大規模データの高速読み取りが可能なギガシークエンサーが開発されたが、読み取り長が短く読み取りエラーが多いため、アセンブリを困難にし信頼性を低下させるという課題が生じていた。

これらの課題を解決するために本研究では、はじめにギガシーケンサーの読み取りエラー部位への対応として利用されるk-merを用いたアセンブリが、用いるk値や手法によって結果がバラつき、各結果が相補的であることを示した。それらを踏まえ、複数のk-merと手法によるcontigを統合するダブルアセンブリを提案した。k-merを用いた手法によるcontigにおける各k-merは出現頻度を示すcoverage(被覆値)の情報を有することから、contig上におけるk-merのcoverage valueの分布を波形として表現した。その結果波形の特徴とcontig結合の正誤に関連性がみられたので、周波数解析による特徴量を用いて結合の正誤を決定することとした。以上より、k-merの分布特徴量のフーリエ変換成分や勾配、結合対象であるcontigの波形の類似度など18の変数を説明変数とし、結合の正誤を目的変数とした学習データを用いて、決定木アルゴリズムであるC4.5への適用により判別ルールを獲得したダブルアセンブリDAwCC(Double Assembly method with Contig for k-mer's coverage)を第1に提案した。正結合ルール、誤結合ルールの獲得が可能であるため、正結合ルールを満たすcontigを取得した場合と全結合より誤結合ルールを満たす配列を除去した場合ついて検証した。検証実験の結果、ダブルアセンブリをおこなうことで、従来のアセンブリで生成が困難であった長い正結合配列の生成、それによる被覆率の改善が可能になった。さらに判別ルールの適用によりcontigの結合をおこないながら誤結合の出力を未然に防ぐことができ、従来手法と比べ長く正しいcontigの生成および被覆率の改善が可能になった。

第 2 の提案手法においては、判別ルールの判別能力の改善という観点からアセンブリの性能改善を試みた。Support Vector Machine や決定木アルゴリズムといった従来の機械学習アルゴリズムによる判別ルールでは対応が困難な contig が残されていることから、結合の信頼性を示す結合正誤に加え、従来のアセンブリにて結合精度の指標として取り入れられてきた k-mer の coverage value の最小値を示す被覆最小値、重複長の目的変数としての適用について検討した。この新たな 2 変数は結合正誤 と異なり量的変数であるため重回帰分析による判別式の生成を試みたが、2 変数そのものがバラツキを持ち、判別精度が著しく低いことがわかった。そこで被覆最小値、重複長の値の分布と結合正誤の情報を利用し質的変換をおこない、C4.5 を用いて正結合、誤結合ルールを拡張させた。さらに正結合、誤結合ルールを 2 段階に同時適用することで、両ルールの対応が困難な contig の取得または除去 (特に長い誤結合配列) を試みた。これらの発想を導入した複合決定木より獲得した判別ルールを適用したダブルアセンブリ DAwCDT(Double Assembly with Comparative Decision Tree) を提案した。検証実験の結果、従来の決定木による判別ルールでは対応できなかった contig への判別が可能になり、より幅広く contig 結合の正誤の判別が可能になった。その結果アセンブリ性能としても被覆率、正解率共に改善させる可能性を示すことができた。

本研究にて提案した複合決定木、ダブルアセンブリのアイディアは Velvet や ABySS に限らず、k-mer を用いた全てのアセンブリへの適用が可能であり、多くのアセンブリの信頼性向上への貢献が期待できる。

参考文献

[1] Watson,J.D.andCrick,F.H.C.“MolecularStructureofNucleicAcids”,Nature,Vol.171, pp.733-738, 1953

[2] 古川俊治：ヒトゲノム・遺伝子解析研究の現状と課題, 慶應法学第 18 号, pp1-44, (2011)

[3] Altschul, S.F., Gish, W., Miller, W. and Myers, E.M. “Basic Local Alignment Search Tool”, Journal of Molecular Biology, Vol.215, Issue.3, pp.403-410, 1990

[4] 城口克之：RNA シークエンシング : 生物物理 53(6), pp290-294, (2013)

[5] Sanger, F, Coulson, A. R. : J. Mol. Biol., 94, pp.441-448, (1975)

[6] Allan M. Maxam and Walter Gilbert.: “A New Method for Sequencing DNA”. Proceedings of the National Academy of Sciences, USA 74 (2): pp560-564, (1977)

[7] 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 assemblies, Bioinformatics, vol.24 no.6, pp.744-750, (2007)

[8] Jun S. Liu : The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem, Journal of the American Statistical Association, Vol.89, Issue 427, (1994)

[9] Breiman, L. : Bagging predictors, Machine Learning, Vol. 24, pp123-140 (1996).

[10] Freud, Y. and Schapire, R. E. : A decision theoretic generalization of on-line learing and an application to boosting, Journal of Computer and System Sciences, Vol.55, pp.119-139 (1995).

[11] LEO BREIMAN : Random Forests, Machine Learning, 45, 532, 2001

[12] Webb, G. I. : MultiBoosting: A Technique for Combining Boosting and Wagging, Machine Learning Vol. 40, pp.159-196 (2000)

[13] Yoshiaki YASUMURA, Kuniaki UEHARA: An Ensemble Learning Method Integrating Bagging and Boosting The 19th Annual Conference of the Japanese Society for Artificial Intelligence, 3F1-01 (2005)

[14] Yu, C. and Skillicorn, D.B. : Parallelizing Boosting and Bagging, Technical Report 2001-442,Depertment of Computing and Information Science, Queen’ s University (2001)

[15] Quinlan, J. R. : Bagging, boosting, and C4.5, Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp.725-730, (1996)

[16] Bauer, E. and Kohavi, R : An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants, Machine Learning, Vol.36, pp.105-142 (1999)

[17] Jason Pella, Arend Hintzea, Rosangela Canino-Koninga, Adina Howeb, James M. Tiedjeb,c, and C. Titus Browna : Scaling metagenome sequence assembly with probabilistic de Bruijn graphs, PNAS , vol. 109 ,no. 33.pp1327213277, August 14, 2012

[18] These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure : Qingpeng Zhang, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, C. Titus Brown, Genomics,12 May 2014

[19] Xiaohong zhao, Lance E. Palmer : EDAR(An Efficient Error Detection and Removal Algorithm for Next Generation Sequencing Data) , (2010)

[20] Rene L. Warren , Granger G. Sutton , Steve J. M. Jones and Robert A. Holt : Assembling millions of short DNA sequences using SSAKE, Bioinformatics, 23 (2007), 500-501

[21] 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)

[22] Daniel R. Zerbino, Gayle K. McEwen, Elliott H. Margulies, Ewan Birney : Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read de Novo Assembler, PLoS one Vol. 4, Issue12, e8407 (2009).

[23] Jared T. Simpson, kim Wong, Shaun D. Jackman, et al ABySS : A parallel assembler for short read sequence data, Genome Research, 19(2009), 1117-1123.

[24] 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 6044 (2010), 426-440.

[25] Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek, Jean-Marc Daran, Marcel Reinders, and Dick de Ridder : Integrating genome assemblies with MAIA, ECCB, 26 (2010), 433439.

[26] Guohui Yao. Liang Ye, Hongyu Gao, Patrick Minx, Wesley C. Warren, George M. Weinstock : Graph accordance of next-generation sequence assemblies, 28(2012),13-16.

[27] Jared T. Simpson and Richard Durbin : ”Efficient de novo assembly of large genomes using compressed data structures” Genome Res. 22: pp.549-556 (2012)

[28] Alesky V. Zimin, Douglas R. Smith, Granger Sutton : Assembly reconciliation, Bioinformatics, 24 (2008),42-45

[29] Jared T. Simpson and Richard Durbin : Efficient de novo assembly of large genomes using compressed data structures Genome Research 22 (2012) 549-556

[30] Yongchao Liu, Bertil Schmidt and Douglas L Maskell:Parallelized short read assembly of large genomes using de Bruijn graphs, BMC Bioinformatics (2011)

[31] Ayako Ohshiro, Takeo Okazaki, Hitoshi Afuso, Morikazu Nakamura : A study of double assembly method for DNA sequences, IPSJ SIG 33(2013)

[32] Quinlan, J. R : C4.5 Programs for Machine Learning (1993)

[33] Quinlan, J. R:”Induction of decision trees.” Machine learning 1.1 pp81-106.(1986)

[34] Thomas G. Dietterich : Machine-Learning Research Four Current Directions (AAAI) AI Magazine, 18 (1997)

[35] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J.:Classification and regression trees. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software.(1984)

[36] Wei-Chun Kao, Andrew H. Chan and Yun S. Song : ECHO: A reference-free short-read error correction algorithm Genome Res. 2011 21: 1181-1192 originally published online April 11, (2011)

[37] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman:An Eulerian path approach to DNA fragment assembly, PNAS, August 14 ,vol. 98 no. 17 (2001)

[38] Angeleri E, Apolloni B, de Falco D, Grandi L., (1999) DNA fragment assembly using neural prediction techniques, *Int. J. Neural. Syst*, Vol9, Issue.6, pp.523-44.

[39] Martin Hunt, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman and Thomas D Otto: REAPR: a universal tool for genome assembly evaluation

[40] Xiaohu Shen, Manohar Shamaiah, and Haris Vikalo : Iterative Learning for Reference-Guided DNA Sequence Assembly from Short Reads: Algorithms and Limits of Performance ,22 March 2014

[41] Lincoln D Stein: The case for cloud computing in genome informatics. Genome Biology 2010, 11:207.

[42] Miller JR, Koren S, Sutton G: Assembly algorithms for next-generation sequencing data. Genomics 2010, 95:315-327.

[43] 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)

[44] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, et al.：SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing, Genome Research, vol.17, pp.1697-1706,(2007)

[45] 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.29422944(2007)

[46] ”Richard.C. Singleton”:On computing the fast Fourier Transform,Communications of the ACM Vol.10 pp647-654(1967)

[47] Pandey V, Nutter RC, Prediger E (2008) Applied Biosystems SOLiDTM System: Ligation-Based Sequencing. In: Next Generation Genome Sequencing: Towards Personalized Medicine, Wiley, pp 2941.

[48] 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 assemblies’ Vol. 24 no. 6 , pp. 744-750,(2008)

[49] Qingpeng Zhang, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe C. Titus Brown : These are not the k-mers you are looking for: BIOicient online k-mer counting using a probabilistic data structure

[50] kmacs: the k-Mismatch Average Common Substring Approach to alignment-free sequence comparison: Chris-Andre Leimeister, and Burkhard Morgenstern, 30 (10), Bioinformatics, May 15, (2014)

[51] Bertil Schmidt, Ranjan Sinha, Bryan Beresford-Smith and Simon J. Puglisi: A fast hybrid short read fragment assembly algorithm, BIOINFORMATICS APPLICATIONS NOTE, pp.22792280,Vol. 25 no. 17 (2009)

[52] Martin Hunt, Chris Newbold, Matthew Berriman and Thomas D Otto: A comprehensive evaluation of assembly scaffolding tools, Genome Biology (2014)

[53] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron and Ewan Birney : Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, BIOINFORMATICS, Vol.28, no.8, (2012).

[54] Juliane D. Klein, Stephan Ossowski, Korbinian Schneeberger, Detlef Weigel, Daniel H. Huson : LOCAS A Low Coverage Assembly Tool for Resequencing Projects,PLoS one Vol. 6,Issue8, 123-140 (2011).

[55] Heng Li and Richard Durbin: Fast and accurate short read alignment with BurrowsWheeler transform, BIOINFORMATICS, Vol. 25 no. 14 2009, pages 17541760(2009)

[56] David Coil, Guillaume Jospin, and Aaron E. Darling: A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data, (2014)

[57] Yu Peng, Henry C. M. Leung, S. M. Yiu and Francis Y. L. Chin : IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth ,BIOINFORMATICS, Vol. 28 no. 11 2012, pages 14201428 ,(2012)

[58] Lin S.H, and Liao Y.C.: CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes, Vol.8, No.3, e60843 (2013).

[59] Anthony M. Bolger, Marc Lohse and Bjoern Usadel : Trimmomatic: A flexible trimmer for Illumina Sequence Data, Bioinformatics Advance Access published April 1, (2014)

[60] M Stanley Fujimoto, Paul M Bodily, Nozomu Okuda, Mark J Clement, Quinn Snell: Effects of error-correction of heterozygous next-generation sequencing data,BMC Bioinformatics, (2014)

[61] Aarti Desai, Veer Singh Marwah, Akshay Yadav, Vineet Jha, Kishor Dhaygude, Ujwala Bangar, Vivek Kulkarni, Abhay Jere: Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data, PLOS ONE, Volume 8,April, (2013)

[62] Riccardo Vicedomini, Francesco Vezzi, Simone Scalabrin, Lars Arvestad, Alberto Policriti: GAM-NGS: genomic assemblies merger for next generation sequencing,BMC Bioinformatics ,vol14,2013

[63] Alkan C, Sajjadian S, Eichler EE: Limitations of next-generation genome sequence assembly. Nat Methods 2010, 8(1):61-65.

[64] Yang X, Chockalingam SP, Aluru S: A survey of error-correction methods for next-generation sequencing. Brief bioinform. 2013, 14(1):56-66.

[65] Kelley DR, Schatz MC, Salzberg SL, et al: Quake: quality-aware detection and correction of sequencing errors. Genome Biol , 11(11):116. (2010)

[66] Paul Medvedev, Eric Scott, Boyko Kakaradov and Pavel Pevzner : Error correction of high-throughput sequencing datasets with non-uniform coverage, ISMB, Vol. 27, pages 137141,(2011)

[67] Boisvert S, Laviolette F, Corbeil J : Ray simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol. 17 (11): 15191533,(2010).

[68] 水島-菅野純子:次世代シークエンサーの医療への応用と課題, モダンメディア 57 巻 8 号 2011［その他］ p225

[69] Paired de Bruijn Graphs: A Novel Approach for Incorporating Mate Pair Information into Genome Assemblers,Paul Medvedev, Son Pham, Mark Chaisson, Glenn Tesler, and Pavel Pevzner. Journal of Computational Biology. November 2011, 18(11): 1625-1634

[70] Mick Watson:Quality assessment and control of high-throughput sequencing data, Bioinformatics and Computational Biology, 00235 (2014)

[71] RIKEN BRC DNA BANK (ja), http://dna.brc.riken.jp/ja/, (2011)

[72] BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads

[73] Bloom,B. : Space/time trade-offs in hash coding with allowable errors. Commun. ACM, 13, 422— 426.(1970)

[74] Mark Howison, Felipe Zapata, Erika J. Edwards, Casey W. Dunn ,Bayesian Genome Assembly and Assessment by Markov Chain Monte Carlo Sampling,PLOS ONE, Volume 9 , Issue 6 , e99497(2014)

[75] Atif Rahman and Lior Pachter ,CGAL: computing genome assembly likelihoods,Genome biology , 14:R8,(2013)

[76] Omega: an Overlap-graph de novo Assembler for Metagenomics,Bahlul Haider, Tae-Hyuk Ahn, Brian Bushnell, Juanjuan Chai, Alex Copeland, and Chon- gle Pan,Bioinformatics,July 1, 2014 30 (13)

[77] Ergude Bao, Tao Jiang and Thomas Girke: Align- Graph: algorithm for secondary de novo genome as sembly guided by closely related references, Vol. 30 ISMB , p319— 328,(2014)

[78] Rayan Chikhi and Paul Medvedev:Informed and automated k-mer size selection for genome assembly, BIOINFORMATICS, Vol. 30 no. 1, pages 3137,(2014)

[79] Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, Budapest, pp.267-281.

[80] Tobias Rausch, Sergey Koren, Gennady Denisov, David Weese, Anne-Katrin Emde, Andreas Doring and Knut Reinert: A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads, BIOINFORMATICS, Vol. 25 no. 9, pp11181124, (2009)

[81] Zhenyu Li, Yanxiang Chen,DeshengMu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, Bicheng Yang and Wei Fan:Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph BRIEFINGS IN FUNCTIONAL GENOMICS, VOL 11. NO 1. pp25-37 (2012)

[82] National Center for Biotechnology Information,http://www.ncbi.nlm.nih.gov/

[83] Cormen, Thomas H.; Leiserson, Charles E., Rivest, Ronald L. : The Floyd-Warshall algorithm, pp.558-565 MIT Press and McGraw-Hill. (1990)

[84] Dijkstra, E.W.:A note on two problems in connexion with graphs. In Numerische Mathematik, 1, pp269-pp271 (1959)

[85] F. Gobel., A.A. Jagers:Stochastic Processes and their Applications, Vol. 2, Issue 4, pp311-336(1974)

[86] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. Automation and Remote Control, 24, (1963)