RNA-Seqデータを用いた
転写解析手法に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2014年1月

大野　朋重

関連研究論文

- 学術論文

  1-1 Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: A method for isoform prediction from RNA-Seq data by iterative mapping, *IPSJ Transactions on Bioinformatics*, Vol. 5, pp.27–33, 2012.

- 国際会議

  2-1 Tomoshige Ohno, Motokazu Ishikawa, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: An improved RNA-Seq analysis method for isoform prediction from RNA-Seq data by iterative mapping. *The 21st International Conference on Genome Informatics (GIW 2010)*, Hangzhou, China, December 16–18, 2010.

  2-2 Tomoshige Ohno, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda: Integrative Prediction of miRNA-mRNA interactions from high-throughput sequencing data, *RECOMB/ISCB Conference on Regulatory and Systems Genomics, with DREAM Challenges 2013*, Toronto, Canada, November 8–12, 2013.

内容梗概

生物の遺伝情報は DNA 塩基配列としてゲノムに保存されている. DNA が RNA へと転写され, その RNA がタンパク質へ翻訳されたり他の DNA/RNA/タンパク質と結合したりすることにより生物学的に機能するようになる. つまり RNA 転写産物が生物学的形質を発現させたり別の RNA の転写を調整したりしており疾患との関連も指摘されているため, 転写全体像 (トランスクリプトーム) とその過程で機能する転写制御関係を解明することは生物学のみならず, 医学・薬学的見地からも重要である.

近年トランスクリプトーム観測の新技術として RNA-Seq が広く用いられるようになりつつある. RNA-Seq とは高速シーケンサを用いて RNA 配列断片 (リード) を大量に読み取る技術であり, これにより原理的には未知の転写産物の配列や転写量を同時に求めることができる. データ量の大きさのため計算機による解析が必要であるが, 解析手法が確立されておらず, その確立が望まれる.

本研究ではまず遺伝子がどのような転写産物を生成 しているかを調べるためにアイソフォーム推定手法に関する研究を行った. RNA-Seq を用いたアイソフォーム推定ではまず各リードのゲノム上での由来箇所を求めるために, リードとゲノム配列との対応を調べるマッピングと呼ばれる操作を行う. しかし, リードの短さやシーケンスエラー, ゲノム配列の相同性などの原因のために由来箇所を 1 ヶ所に特定できないリードが多数存在する. そこでマッピングと発現量推定を反復的に行う反復マッピングを提案し, 従来手法よりも高い精度でアイソフォームを推定できるようになった.

次に, miRNA による遺伝子の転写調整を解明するために, miRNA-遺伝子間相互作用予測に関する研究を行った. 従来の相互作用予測は, miRNA とターゲット遺伝子における結合領域の配列相補性に基づき行われていたが, この予測結果は偽陽性を大量に含む. 発現量情報をそこで本研究では, 遺伝子と miRNA の発現量を入力として正準相関分析を行うことにより, 偽陽性を大幅に減らすことができた.

目次

第 1 章　序論 1

1.1　本論文の背景 1

1.2　本論文の目的 3

1.3　本論文の構成 4

第 2 章　反復マッピングによるアイソフォーム推定 5

2.1　緒言 5

2.1.1　遺伝子の発現 5

2.1.2　選択的スプライシングとアイソフォーム 6

2.1.3　遺伝子発現解析手法 8

2.1.4　完全長 cDNA 8

2.1.5　高速シーケンサ 13

2.2　RNA-Seq 発現解析 16

2.2.1　RNA-Seq とは 16

2.2.2　RNA-Seq 解析の流れ 17

2.2.3　従来手法の問題点 25

2.3　提案手法 25

2.3.1　提案手法の概要 25

2.3.2　提案手法の詳細 27

2.4　実験 39

2.4.1　使用データ・実験条件 39

2.4.2　実験結果 40

2.4.3　考察 42

2.5　結言 44

第 3 章　正準相関分析を用いたマイクロ RNA-遺伝子間相互作用予測手法 45
3.1　緒言 45
3.2　miRNA の機能 45
3.3　miRNA-遺伝子間相互作用予測 47
3.3.1　従来の miRNA-遺伝子間相互作用予測手法 47
3.3.2　従来の予測手法の問題点 48
3.4　提案手法 49
3.4.1　提案手法のアイデア 50
3.4.2　正準相関分析 50
3.4.3　CCA を用いた相互作用予測 52
3.5　実験 52
3.5.1　データ 53
3.5.2　結果 53
3.5.3　考察 55
3.6　結言 56

第 4 章　結論 57

謝辞 59

参考文献 61

図目次

2.1 DNA から成熟 mRNA への転写 6
2.2 選択的スプライシング 7
2.3 各トランスクリプトーム解析手法の対象範囲 11
2.4 マイクロアレイ実験の手順 12
2.5 サンガー法と高速シーケンサを用いたシーケンシング 15
2.6 循環行列の作成 18
2.7 BW 変換 19
2.8 逆 BW 変換 20
2.9 逆 BW 変換による S の構築 21
2.10 BW 変換を用いた文字列 ACT のパターン検索 22
2.11 RNA-Seq を用いたアイソフォーム推定 23
2.12 提案手法の流れ 26
2.13 ジャンクション検出 28
2.14 推定遺伝子モデルの構築 29
2.15 推定遺伝子モデル 31
2.16 推定アイソフォーム 33
2.17 発現量定量化 34
2.18 再マッピング 38
2.19 ACTB と他の遺伝子の関連 41
2.20 提案手法と Cufflinks の比較 42
3.1 miRNA の生合成 46
3.2 miRNA と mRNA の配列相補性 48
3.3 データベース内の miRNA-遺伝子間相互作用 49

3.4 miRNA-遺伝子の発現量相関係数分布 51
3.5 CCA を用いた miRNA-遺伝子間相互作用推定モデルの構築 53
3.6 提案手法の結果 54

表目次

2.1 Ensembl 上のヒト遺伝子のアイソフォーム数 8

2.2 高速シーケンサの比較 16

2.3 BW(S$) と SORT(S$) の比較による $\gamma(i)$ 20

2.4 アイソフォーム推定結果と Ensembl の比較 40

2.5 ユニークリード/マルチリード数 40

2.6 エキソンとイントロンの推定精度 40

3.1 提案手法による予測結果分類 54

3.2 Precision/Recall と F-measure 55

第 1 章　序論

1.1　本論文の背景

生物の遺伝情報は DNA (デオキシリボ核酸: deoxyribonucleic acid) に保存されており, この遺伝情報全体をゲノムと呼ぶ. ゲノムはアミン (amin), チミン (thymine), グアニン (guanine), シトシン (cytosine) という 4 種類の塩基の配列という形として保存されており, それぞれの頭文字をとり A, T, G, C の並びで表現され, 生体を主に構成するタンパク質の設計図となる. DNA は複製された後, pre-mRNA (messenger RNA) に写し取られ (転写; transcription), スプライシングと呼ばれる修飾を受け成熟 mRNA になった後, その成熟 mRNA を鋳型として対応するアミノ酸がペプチド鎖として並べられる (翻訳; translation) ことによりタンパク質がつくられる. この一連の過程をセントラル・ドグマ (central dogma) と呼ぶ. ゲノム配列が生命を規定するとして, 1990 年あらゆる生物のゲノム配列を解読し, タンパク質コード領域である遺伝子を決定する ゲノムプロジェクトが開始された. ヒトゲノムについては 2000 年にドラフト配列 [1] が, 2003 年には完全配列 [2] が発表された.

このゲノム配列は全ての細胞の染色体に保存されているが, 細胞が属する組織の違いや, 環境や疾患に応じて DNA から転写される転写産物やその転写量が異なる. 特定の組織や特定状況下における転写像全体をトランスクリプトーム (transcriptome) と呼び, 生物学のみならず医療や薬学などにおいても重要な研究分野として位置づけられている.

トランスクリプトーム中で最も重要であるとされてきたのがタンパク質配列を伝達する mRNA である. ヒトは約 10 万種類のタンパク質で構成されるため, 当初は同数程度の遺伝子がヒトゲノム上に存在すると考えられていた. しかし, ヒトゲノム解読が進むにつれて, ヒトの遺伝子数は予想よりも遥かに少ないことが明らかになっていった. 当初 10 万程度と考えられていたヒト遺伝子数は, 2001 年のヒトゲノムドラフト配列では約 3

万個であるとの推定が報告され, 2004 年には約 30 億塩基のヒトゲノム配列に基づいた推定によると 21,787 個であるとの研究成果が発表された. 最新の研究には, ヒト遺伝子は 19,042 個とするものもある [3].

この遺伝子数とタンパク質の多様性の差は選択的スプライシング (alternative splicing) によるものである. 真核生物の遺伝子内には, タンパク質に翻訳されるエクソン (exon) とスプライシングの際に除去されるイントロン (intoron) が存在しているが, スプライシングの際にはエクソンが選択的に組み合わされることにより, アイソフォーム (isoform) と呼ばれる複数の異なる mRNA が 1 つの遺伝子から転写され, 異なるタンパク質へと翻訳される. アイソフォームは, 1 つの遺伝子から異なる機能をもつ複数のタンパク質を生成する役割を果たし, また疾患との関連も報告されているため [4], トランスクリプトームにおけるアイソフォームの転写の解明は生命システムを理解する上において重要であると言える.

タンパク質コーディング配列を規定する mRNA の他に, ノンコーディング RNA (non-coding RNA; ncRNA) と呼ばれる, 転写はされるもののタンパク質に翻訳されない RNA の存在が近年明らかになって来た. ヒトが生成する全ての RNA にアノテーションを付けることを目指した ENCODE プロジェクト [5] によると, ヒトのタンパク質コーディング遺伝子領域はヒトゲノムの 2% 程度に留まる一方で, ヒトゲノムの 80% 以上は ncRNA に転写され機能する. つまり, ヒトゲノムの大半は ncRNA 領域であることが明らかになった.

ncRNA は生成過程や機能によりさらに細かなクラスに分けられ, それぞれ主に DNA, RNA, タンパク質などに対して特有の作用をすることが明らかにされつつある. ncRNA の各クラスの機能として大きなものの 1 つは遺伝子発現の各プロセスを行うことである [6]. 転写された pre-mRNA は snRNA (small nuclear RNA) によりイントロンが除去される. その後 3 つの塩基配列の並びであるコドンに対応したアミノ酸を tRNA (transfer RNA) が運搬し, rRNA (ribosomal RNA) が構成するリボゾームにおいてそのアミノ酸を結合することにより翻訳が行われる. ncRNA の大きな機能のもう 1 つは,

遺伝子の転写調整である [7, 8, 9]. この転写調整に関わる ncRNA のクラスには lncRNA (long RNA), siRNA (small interfering RNA), miRNA (micro RNA) などがあるが, ターゲットとする遺伝子の多様性や疾患との関連などから特に miRNA が注目されている.

遺伝子発現を観測するのに従来はマイクロアレイが用いられていた. しかし, マイクロアレイを用いた発現観測では既知の遺伝子配列をプローブとしてハイブリダイゼーションをさせ蛍光強度を測定することにより発現量を測定するため, アイソフォームレベルの発現解析ができない. そこで近年トランスクリプトーム観測の新技術として RNA-Seq が広く用いられるようになりつつある. RNA-Seq とは, 高速シーケンサを用いて大量の RNA 塩基配列断片を読み取る技術であり, その大量データを解析することにより, 転写された RNA の配列及び転写量をゲノムワイドに観測することが可能となる. しかし, RNA-Seq データを用いた転写産物配列の再構築や転写量の定量化, 転写産物間相互作用の推定など, 解析手法の構築において課題がある.

1.2 本論文の目的

本研究では, RNA-Seq データを用いた転写解析手法を提案することを目的とする. そのために, アイソフォーム推定手法と, miRNA-遺伝子間相互作用予測手法を提案する.

まずアイソフォーム推定に関して取り組む. 生体を構成する上で最も重要であると考えられる mRNA について, RNA-Seq データからアイソフォームを推定する手法を提案する.

次に miRNA-遺伝子間相互作用予測に関して取り組む. それぞれの遺伝子の発現を制御する因子として miRNA に着目し, どの miRNA がどの遺伝子を制御しているのかを予測する手法を提案する.

1.3 本論文の構成

本論文は 4 章構成である．第 2 章では mRNA に着目し, RNA-Seq データからアイソフォームを推定する手法を提案する．第 3 章では miRNA に着目し, miRNA の標的遺伝子予測手法を提案する．

第 2 章では, mRNA 発現におけるアイソフォームについて述べた後, トランスクリプトームを観測するための技術である RNA-Seq とそれにより生成されるデータからアイソフォームを推定する手法について述べる．実際の RNA-Seq データに対して提案手法を適用し, 推定されたアイソフォームを既知のアイソフォームと比較する．また, 提案手法の性能を既存手法と比較し, そこで得られた結果を基に考察を行う．

第 3 章では, miRNA の機能について述べた後, miRNA-遺伝子間相互作用予測手法を提案する．遺伝子発現制御に大きく関与していると考えられている miRNA に着目し, RNA シーケンスデータからつくられた発現プロファイルを用いることにより miRNA-遺伝子間相互作用予測を行った結果について述べる．また, 既存手法を用いて miRNA-遺伝子間相互作用を予測したデータベースと比較を行い, 提案手法の性能を評価し, 考察を行う．

最後に第 4 章では本研究の結論を述べるとともに, 今後の課題や展望について触れる．

第 2 章　反復マッピングによるアイソフォーム推定

2.1　緒言

本章では, 遺伝子発現について触れたのち, 選択的スプライシングアイソフォームについて説明する. その後アイソフォーム推定に主として用いられてきた手法としてシーケンスベースのトランスクリプトーム解析手法について説明し, 最後に近年急速に普及しつつある高速シーケンサについて述べる.

2.1.1　遺伝子の発現

遺伝子 (gene) は生物の遺伝的な形質を規定する因子であり, 遺伝情報の単位とされる. 遺伝子からタンパク質が合成され生命機能を果たすようになるまでの過程を発現と呼ぶが, 発現には転写と翻訳というプロセスが含まれる. 転写とは DNA 配列を基に mRNA が合成される過程であり, 翻訳とはリボソーム内で mRNA の情報に基づき決定されたアミノ酸配列からペプチド結合によりタンパク質が合成される過程である. 発現の過程において作られた転写産物 (mRNA) の量をその遺伝子の発現量と呼ぶ. 遺伝子の転写制御により, それぞれの mRNA が必要とされる量だけ存在するとされる. 遺伝子の発現では最終的にタンパク質が生成されるため, 発現量はタンパク質の量で表されるべきだが, タンパク質の量を測定することは困難であることと, タンパク質の量は mRNA に依存していることにより, mRNA の量を発現量として測定することが広く行われている. 遺伝子発現に関する情報を蓄積するために様々なデータベースが提供されている. 主なものとして各生物種の塩基配列に関する情報を蓄積する GenBank[10], タンパク質立体構造データベースの Protein Data Bank(PDB)[11, 12], などがある. これらのデータベースに蓄積された情報はトランスクリプトーム (transcriptome:転写産物の全体) 解析をはじめ, 医学・薬学・生物学において広く利用される.

2.1.2　選択的スプライシングとアイソフォーム

真核生物の遺伝子発現においては, 転写の際に DNA が一旦 mRNA 前駆体 (pre-mRNA) になり, 翻訳に不要な領域を取り除くスプライシング (splicing) という現象を経て成熟 mRNA(mature mRNA) が生成される（図 2.1）[13]. このとき成熟 mRNA に残る領域をエキソン (exon), スプライシングにより除去される領域をイントロン (intron) と呼ぶ. またエキソンの接続部分をスプライスジャンクション, または単にジャンクションと言う. ほぼ全てのイントロンは “GT (mRNA 中では GU)” という塩基配列で始まり “AG” で終わるという GT-AG ルールが存在し, ヒトの場合は 98% のイントロンについてこれが当てはまっている [14].

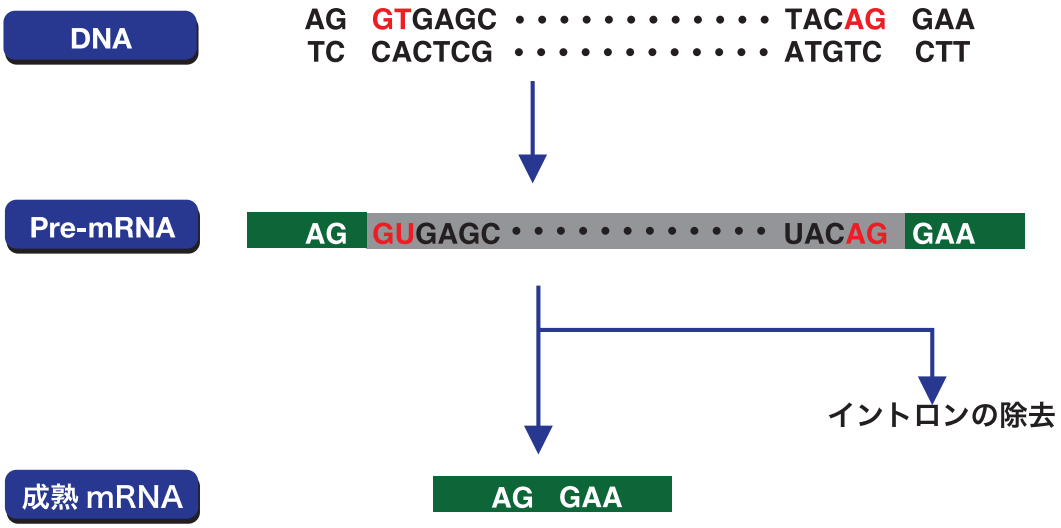

図 2.1　DNA から成熟 mRNA への転写

真核生物の遺伝子発現においては, 選択的スプライシング (alternative splicing) と呼ばれる現象により 1 つの遺伝子から複数のタンパク質が生成される. このとき生成されるそれぞれの成熟 mRNA をアイソフォーム (isoform) と呼ぶ（図 2.2）.

各生物の遺伝子数はゲノムサイズに比較してそれほど大きくなく, 各生物の遺伝子数はマウスが 2 万個強, ヒトに至っては 2 万個足らずであるとの報告もある [3]. それにも

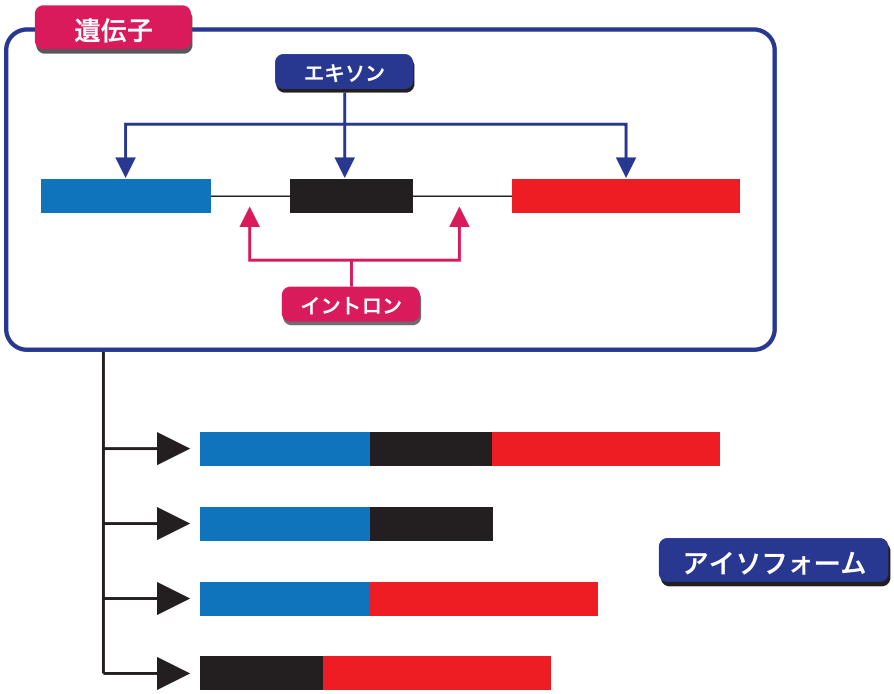

図 2.2　選択的スプライシング

関わらず, 生体内で多種多様な細胞が作られ生命活動が維持されている. これは選択的スプライシングにより, 生命としての機能を果たすタンパク質に多様性が生まれているためであり, 特にヒトに関してはヒト遺伝子のうち 92〜94% で選択的スプライシングが発生しているという報告がある [15].

このように生物の生命活動は選択的スプライシングにより高度に制御されており, スプライシングのエラーは病気につながる [4, 16]. そのため選択的スプライシングを解明することは生命の成り立ちを知る上で非常に有用である. そこで Ensembl というデータベースには各生物種ごとにアイソフォームの情報が格納されている. そのうち表 2.1 はヒト遺伝子のアイソフォームの登録エントリ数の推移をまとめたものであるが, 更新の度にエントリ数がかなり増加しており, 未知のアイソフォームも大量にあると考えられる. 数千通りもの転写パターンを生物実験により手作業で確かめるのは大変困難でありコストもかかるため, 遺伝子発現に関するデータから情報科学的なアプローチで選択的

スプライシングを解析する手法の確立が求められている.

表 2.1　Ensembl 上のヒト遺伝子のアイソフォーム数

バージョン	リリース日	アイソフォーム数
GRCh37.59	2010/8/3	151225
GRCh37.60	2010/11/3	157480
GRCh37.61	2011/2/2	167074

2.1.3　遺伝子発現解析手法

転写領域特定や発現量測定を行うために様々な手法が提案されてきた. 本節では主な遺伝子発現解析手法について説明する.

2.1.4　完全長 cDNA

RNA を鋳型として DNA 鎖を合成するレトロウイルスの逆転写酵素を成熟 mRNA に加えることで各 mRNA に対する DNA のコピーとなる相補的 DNA (complementary DNA; cDNA) が得られる. この反応によって合成された一本鎖 DNA 分子を DNA ポリメラーゼで二本鎖 DNA 分子にし, これをプラスミドかウイルスベクターに挿入してクローニングすることで得られるクローンを cDNA クローンといい, 1 回の調整で得られた全 mRNA から作ったクローンの集団を cDNA ライブラリという. 特に転写された RNA と全く同じ長さの cDNA は完全長 cDNA (full-length complementary DNA) と呼ばれる. cDNA クローンはゲノム内で mRNA に転写される領域, すなわちエキソン領域のみを含み, スプライシングにより除去されるイントロン領域を含まない. そのためゲノム配列と完全長 cDNA を比較することでゲノムの転写領域を特定することが可能である（図 2.3）. cDNA ライブラリには転写産物の mRNA が高発現領域ほど大量に存在するため, ライブラリ中の mRNA の量を測定することにより発現量が得られる. しかし, 一本鎖 cDNA から第二の DNA 鎖を合成する際に働く DNA ポリメラーゼは, 結合して

いる RNA 分子を押しのけながら合成反応を進めるが, 最初の DNA 鎖の 3' 末端に塩基対形成した RNA だけは残りその後のクローニングの段階で分解されるため, cDNA ライブラリでは元の mRNA 分子の 5' 末端の塩基配列が欠けていることが多い.

2.1.4.1 シーケンスベースの発現解析

完全長 cDNA がゲノム配列全体から生成される mRNA をクローニングするのに対して, タグと呼ばれる数十 bp 程度の短い mRNA の断片を利用したシーケンスベースの発現解析手法も存在する. これらの手法では, タグをゲノム配列に対してマッピングし, タグがマップされた位置と量から転写領域と発現量を求め, それらの情報を用いてトランスクリプトーム解析を行う. 以下で代表的な シーケンスベースの発現解析手法を説明する.

EST シーケンスベースのトランスクリプトーム解析の初期に用いられた手法として EST (expressed sequence tag) がある [17]. これは発現している mRNA をランダムに cDNA クローン化し, 5' または 3' 末端からシークエンスし塩基配列を決定することによって解析を行う方法である. このアプローチはトランスクリプトームの複雑さを世に知らしめた最初の試みであった. 現在, EST は完全長 cDNA ライブラリー作成技術やランダムプライマーを使った ORESTES（open reading frame EST; タンパク質をコードする可能性のある領域の EST）[18] という方法により, 末端シーケンスとは異なる領域のシーケンスへと可能性が広がっている. このアプローチはゲノム注釈付け, 遺伝子発見, 遺伝子発現プロファイルに大いに貢献しているが, 反面, 得られるシーケンス配列が長すぎたり短すぎたりすること, すべての mRNA を解析するためにかかる高いシーケンス経費のため非効率である.

SAGE EST の非効率性を解決するために開発されたのが, 短いシーケンスタグの効率的なシーケンスを使った SAGE(serial analysis of gene expression) である [19]. SAGE は, 短いヌクレオチドシーケンスタグ（SAGE タグ）といわれる転写物

の 3’ 末端の決まった領域からはじめの 10bp のタグを切り出し, 1 本鎖 DNA 内でタグを 10 個以上連鎖させたものをシーケンスすることにより, 1 シーケンスで 10 倍以上のトランスクリプトーム情報を提供できる. ただ 10bp という短さのためゲノムにマッピングしたときの精度がよくなかったが, 後に 20bp 程度の SAGE タグを切り出しクローニングすることを可能にした “LongSAGE[20, 21]” や, 26bp の “SuperSAGE[22]” が開発され, マッピング精度も向上した.

CAGE 転写物の 5’ 末端の大規模な解析により, 予測ではない実際の転写開始点が確定するばかりでなく, その上流にあるであろうプロモーター領域の予測 (promoter usage) および新規遺伝子の発見が大いに期待される. 5’ 末端配列解読は EST でも行われてきたが, 非効率および全遺伝子の 5’ 末端を網羅しているか定かでないこと, また完全長 cDNA ライブラリーによる解読でさえ長い転写物に対して偏りがあるという問題があった. 一方でコンピュータによるプロモーター予測・探索が進んできたものの, その限定された信頼性から実験的なアプローチによる確認が必要とされていた [23]. これらを解決する方法として転写物の 5’ 末端からの短いタグを複製かつ効率よくシーケンスする CAGE 法 [24] が開発された.

SAGE から応用されてできた CAGE 法は, それまで mRNA の高次構造により難しかったとされる 5’ 末端までの cDNA 合成を可能にした耐熱性逆転写酵素の発見や, トレハロース添加による高温かでの逆転写反応, および mRNA の 5’ 末端構造を選別してくる技術 (cap-trapper 法やオリゴキャッピング法など) などの完全長 cDNA 技術を使い, 遺伝子の 5’ 末端から 20bp のタグを切り出し, それを連結することにより効率的にシーケンスする方法である.

図 2.3 に EST, SAGE, CAGE, および完全長 cDNA が対象とする範囲を示す. EST はゲノム全体をカバーするが, タグの量が不十分であるため発現量の定量化が難しいといった短所を持つ. 逆に SAGE 及び CAGE はそれぞれ 5’ 末端, 3’ 末端領域のみを対象とするため網羅性に欠ける. 完全長 cDNA は 3’ 末端が欠けていたり, 複数種のアイソ

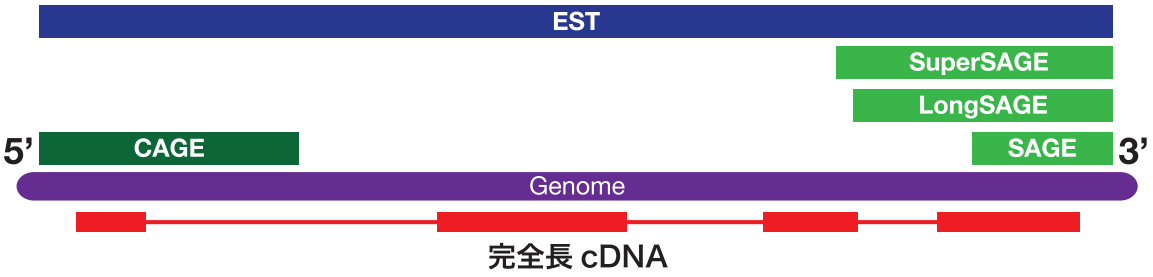

図 2.3　各トランスクリプトーム解析手法の対象範囲

フォームが存在するときスプライスジャンクションを検出するのが困難であるなどの問題がある.

2.1.4.2　マイクロアレイ

遺伝子発現量を測定するための手法としてマイクロアレイと呼ばれる, ハイブリダイゼーションを利用し数千個から数万個の遺伝子の発現量を測定する手法が広く用いられている. ハイブリダイゼーションとは, DNA 鎖が塩基間の特異的な水素結合を介して二本鎖を形成する反応であり, 厳密に反応条件を整えることで完全に相補的な場合のみ安定な二本鎖を形成させることができる. ハイブリダイゼーションを利用した発現量測定手法として RT-PCR 法 [25, 26] やノーザンハイブリダイゼーション法 [27] などがあるが, これらは一回の実験によって数個の発現量を測定するのに適しており主に特定遺伝子を特定とした解析に用いられてきた. それに対してマイクロアレイは数千個から数万個の遺伝子の発現量をシステマティックに測定することが可能で, 網羅的な発現解析に広範に利用されている. マイクロアレイの原理は, 古くから分子生物学において用いられてきた相補的な DNA–DNA 間, DNA–RNA 間のハイブリダイゼーションに基づいている.

マイクロアレイ実験は, 図 2.4 に示すように, DNA アレイの作成, 対象細胞サンプルの調整, ハイブリダイゼーションという 3 つの手順からなる. DNA アレイとしては米国の Affymetrix 社が作成した GeneChip がよく利用される. GeneChip は数万の遺伝子

それぞれについて選別した, 特徴的な 25 塩基程度の DNA 配列解析（遺伝子プローブ）が基盤上に高密度に配置されたものである. また, ハイブリダイゼーション時のノイズを除去するため, 非特徴的な遺伝子プローブも配置されている. 遺伝子プローブとして用いる短い DNA 配列（オリゴヌクレオチド）の人工的な合成には, 半導体製造技術である光リソグラフィー法が応用されている. この技術によりオリゴヌクレオチド合成時のコントロールが容易となり, 他の DNA アレイよりも品質が安定し, 実験の再現性の向上等につながっている. DNA アレイを用意した後, 対象細胞のサンプルから得た mRNA の cDNA への逆転写, cDNA のビオチン標識を加えた上での増幅, さらに cRNA 断片化等の調整作業を行う. 調整後, 標識済 cRNA 断片を DNA アレイにハイブリダイズし, スキャナで各プローブのシグナル強度を読み取ることで, 遺伝子発現を解析する. 遺伝子の発現量は遺伝子プローブ（PM 配列）のシグナル強度から非特徴的な配列（MM 配列）のシグナル配列を取り除いた上で数値化される.

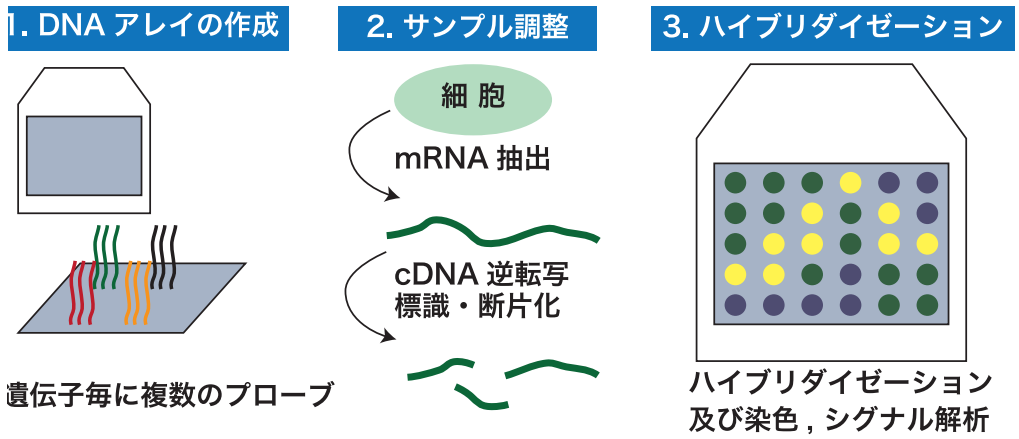

図 2.4　マイクロアレイ実験の手順

このようにマイクロアレイにおける発現量は蛍光強度によって表されるが, 測定される蛍光強度の絶対値には誤差が存在する. マイクロアレイによる測定誤差の要因としては, プローブ DNA を基盤上に配置する装置の精度, 測定に用いる組織の摘出から RNA 抽出までの処理と時間経過, ハイブリダイゼーション時の溶液のイオン強度と温度, 蛍光

強度検出時の背景の状態などがある．また, ダイナミックレンジの広さもマイクロアレイによる発現量測定の際に問題となる．ダイナミックレンジとは測定可能な最大値と最小値の比であり, マイクロアレイにおいては一定以上発現している遺伝子はすべて発現量が等しく測定されてしまい, 遺伝子発現解析に対する障害となっている．

2.1.5 高速シーケンサ

従来のシーケンシングには Sanger 法 [28] と電気泳動をベースとしたキャピラリー式シーケンサが用いられてきた．しかしゲノムワイドのシーケンスには時間がかかり, 改良も限界に達しつつあった．そこで近年ゲル電気泳動の必要のないシーケンサが開発され, スループットが数百倍へと飛躍的に伸びた．次世代シーケンサは画像解析と並列化によりそのスループットを向上させている [29].

従来のサンガー法による DNA シーケンシングと, 次世代シーケンサまたは第 2 世代シーケンサとも呼ばれる高速シーケンサによる DNA シーケンシングの原理を説明する（図 2.5）[30]．最初のステップが DNA の断片化である点は両者に共通している．サンガー法ではその後 2 つの異なるアプローチがある．ゲノムショットガンによる *de novo* シーケンスでは, ランダムに断片化された DNA はプラスミドベクターへとクローニングされ, 大腸菌 (*Escherichia coli*) を変異させるために使われる．それに対して予め定められたターゲットに対する再シーケンスでは, ターゲット側のプライマーを PCR により増幅する．いずれのアプローチでもシーケンシングの際のテンプレートが得られることになる．それからテンプレートの変質, プライマーのアニーリング, プライマー伸長を循環して行うサイクルシーケンシングが行われる．伸長されたプライマーは, シーケンスを行いたい領域の配列に対して相補的であり, 蛍光ジデオキシヌクレオチド (dideoxynucleotides; ddNTPs) と反応するが, その反応は確率的に停止する．そのとき加えられた ddNTP の種類により反応が止まった位置の塩基が分かる．キャピラリーを用いた高解像度のゲル電気泳動により配列が決定される．96 ないし 384 のキャピラリーで並列的に電気泳動を行い, ソフトウェアによりベースコールの際のエラーの確率分布

を考慮して最終的な配列が出力される．サンガー法は改良が続けられ, 今日ではリード長 1,000bp, 塩基当りのシーケンス精度 99.999% を達成するまでに至っている．

一方, 高速シーケンサによるシーケンシングでは DNA 断片化の後, 試験管内で共通のアダプタ配列が付加される．アダプタ付加の代わりに mate-paired タグのライブラリを生成することもある．PCR などによる増幅産物はクローン化・クラスタ化される．このようにして生成されたライブラリは, 酵素反応と画像解析にサイクルによりシーケンスされる．今日普及している多くのプラットフォームではプライマーのテンプレートを逐次的に伸長することにより配列を解読するが, 合成反応を起こす酵素はポリメラーゼとリガーゼのどちらを用いても可能である．その後, 各塩基を合成反応させることで蛍光化し, アレイ全体を画像解析にかけることによりデータを得る．サンガー法に対する高速シーケンサの長所には以下のようなものがある．

- *in vitro* でシーケンスライブラリを作成しクローンの増幅を行うため, 大腸菌の変異やクローン生成などのような, 並列化のためのボトルネックを避けることができる．
- アレイベースのシーケンシングにより, 既存手法よりも高度な並列化が可能となる．$1\mu m$ のオーダーで配列解読が可能であるため, 並行走査により数億本のリードが得られるようになる．
- アレイは平面に固定されるため一様な量の試薬で酵素反応を起こすことができる．通常 μl 単位の試薬が使用されるが, アレイ全体をシーケンスしてもアレイ上に残り還元されるため, 実質上の試薬の使用量はピコリットルもしくはフェムトリットルというレベルになる．

これらの特性から, 高速シーケンサはハイスループットであるばかりでなく, シーケンシングにかかるコストを劇的に削減することも期待される [31]．サンガー法のコストは 1Kb あたり$0.50 程度と言われており, ヒトゲノムの解読に$70,000,000 かかったのに対して, 高速シーケンサによるヒトゲノム解読のコストは$48,000 ～ $1,600,000 程度と言

われている [32].

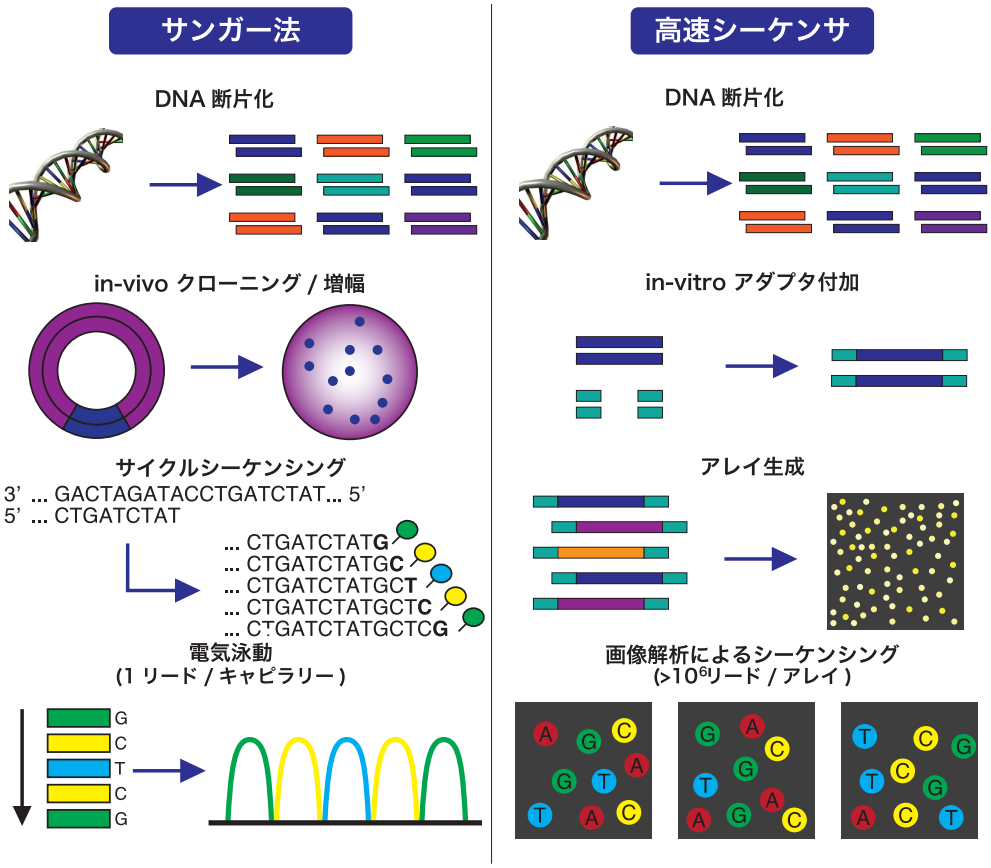

図 2.5　サンガー法と高速シーケンサを用いたシーケンシング

高速シーケンサが出力するのは数十〜数百塩基からなる塩基配列の断片であり, リード (read) と呼ばれる. 1 回の運転で得られる数千万個のリードにより数億 〜数十億塩基の解読が可能となり (表 2.2[32]), 得られたデータはゲノム配列決定, 転写解析やメタゲノムエピゲノムの研究などへと使用される [35, 36]. また, 高速シーケンサは DNA だけでなく RNA もシーケンスすることができる. 本節で述べたように高速シーケンサによる配列解読はハイスループットであり費用対効果が高いため, 高速シーケンサを用いることにより効率的なトランスクリプトーム解析が行えるようになることが期待される. 次章で高速シーケンサを用いた RNA シーケンス技術とその解析手法について述べる.

表 2.2　高速シーケンサの比較

プラットフォーム	リード長 (塩基)	運転時間 (日)	Gb per run	本体価格 (US$)
Roche/454 GS FLX Titanium[33]	330*	0.35	0.45	500,000
Illumina/Solexa GA II	75 or 100	4‡, 9§	18‡, 35§	540,000
Life/APG SOLiD 3	50	7‡, 14§	30‡, 50§	595,000
Polonator G.007	26	5§	12§	170,000
Helicos BioSciences HeliScope[34]	32*	8‡	37‡	999,000
Pacific Biosciences	964*	N/A	N/A	N/A

* 平均リード長

‡ Fragment run

§ Mate-pair run

GA: Genome Analyzer, GS: Genome Sequencer, N/A: Not Available,
SOLiD: support oligonucleotide ligation detection

2.2　RNA-Seq 発現解析

本章では, RNA-Seq データを用いた発現解析について説明する.

2.2.1　RNA-Seq とは

第 2 章で述べたように, 高速シーケンサは近年 RNA-Seq と呼ばれる技術が広く用いられるようになりつつある [37]. これは高速シーケンサを用いて mRNA をシーケンスし, 得られた大量の mRNA リードを用いて発現解析を行うというものである [38]. 前述の通り高速シーケンサは大量のリードを出力するため, 各リードがゲノムのどの位置からどれだけ得られたかという情報から, 各遺伝子の発現量を求めたり, 新規発現領域を発見したりできる可能性を持っている. また発現箇所の塩基配列が 1 塩基単位のレベルで

得られるため, 原理的には選択的スプライシングの推定が可能であり, そのための解析手法が求められている.

2.2.2 RNA-Seq 解析の流れ

2.2.2.1 マッピング

マッピングとは, ゲノム配列を参照して各リードの発現由来箇所を特定することである. このとき参照されるゲノム配列をリファレンス配列と呼ぶ. このステップは EST をはじめとしたシーケンスベースの解析手法とも共通しており, 当初は BLAST[39] や BLAT[40] のようなアライメントアルゴリズムが使用されていた. しかし, 高速シーケンサのスループットの向上に伴うリード数増加のためにより高速にマッピングを行うためのアルゴリズムが求められるようになった. そこで, ハッシュテーブルを用いてリードのインデックスを作成して配列が一致する箇所を探すという手法が提案された. 代表的な実装には MAQ[41], SeqMap[42] などがある. 今日では BW 変換 (Burrows-Wheeler Transform)[43] という操作によりリファレンスとなるゲノム配列をインデックス化する Bowtie[44], BWA[45], SOAP[46, 47] といったソフトウェアが広く利用されている.

Burrows-Wheeler 法は当初, もとのデータ文字列を容易に圧縮できる形に変換する (順序を変える) という考え [48] を用いて, 長いデータ文字列に対する効率的, 無損失性圧縮ツールを作成することを目的としていた [49]. ところが, BW 変換自体が任意の長さの部分文字列に対して非常に高速で省メモリ容量の探索ツールであることがすぐに認識されるようになった [50, 51]. ここでは以下の文字列

$$S = \mathrm{CACTAACTGA} \tag{2.1}$$

を用いて BW 変換と逆 BW 変換の構築, ならびに BW 変換を用いた検索について説明する.

BW 変換 文字列 S (この例では長さ $n = 10$) の BW 変換は以下のように構築される. 図

2.6 に示すように, 0 から $n-1$ まで番号を振られた各行が, 左から右に連続して一文字ずつ循環回転（シフト）するように行列 $Z(S)$ をまず作成する. 次のステップでは辞書式順序で $Z(S)$ の行をソートする. これにより図 2.7 に示す行列 $Z_1(S)$ が得られる. 行列 $Z_1(S)$ の最後の列が S の Burrows-Wheeler 変換 BW(S) である. (2.1) に対して BW 変換したものは

$$\mathrm{BW}(S) = \mathrm{TGCAAAATCC} \tag{2.2}$$

となる.

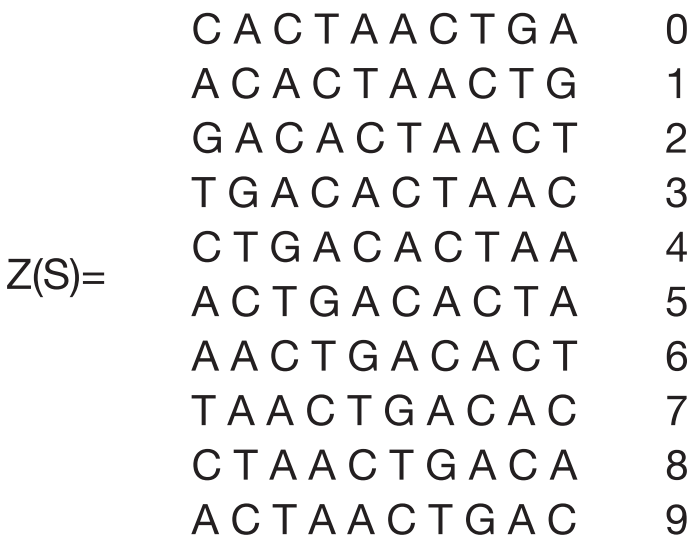

図 2.6　循環行列の作成

逆変換可能性 S のすべての循環回転の BW 変換は上で与えられた文字列と同じである. この意味で図 2.7 に示される変換は逆変換可能ではない. しかし, S の正しい位相を見つけることを可能にするデータを持っていると仮定すれば逆変換が可能となる. 文字列 S の位相を決めるために, $Z_1(S)$ の行を調べる. $Z_1(S)$ の行は S のすべての循環シフトであるので, 少なくともそれらの 1 つは S と等しくなければならない. r で最初のそのような行の索引を表す. 図 2.7 では $r = 4$ である. こ

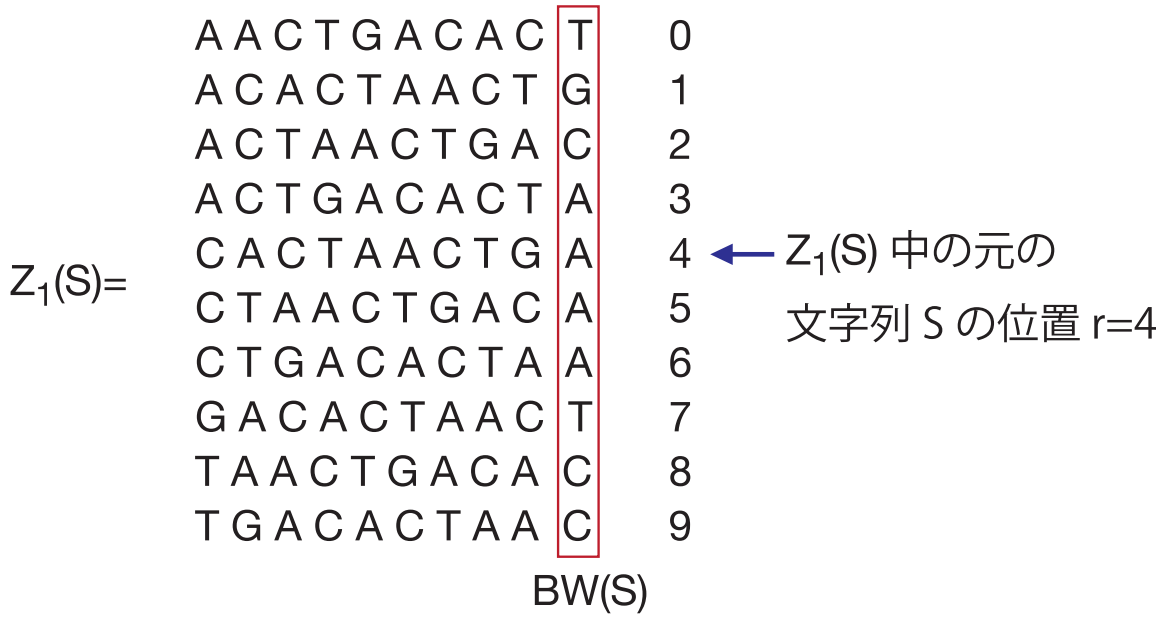

図 2.7　BW 変換

の情報は BW(S) の正しい逆変換を可能にする．位相を決めるために考えられるもう一つの方法は, 終端文字$を付け加えることである．文字列に$は一度だけ現れ, 辞書式順序において最も低いと仮定する．(2.1) で与えられる S の代わりに CACTAACTGA$となり, $は S の最後の文字として現れることを知っているので, BW(S) 中の$の位置は行索引 r に相当する．

逆 BW 変換　もう一つの行列 $Z_2(S)$ を作る．$Z_2(S)$ は行列 $Z_1(S)$ の最後の列を先頭に移すことで得られる．図 2.8 に $Z_1(S)$ と $Z_2(S)$ の両方を示す．S の BW 変換 BW(S) は $Z_1(S)$ の最後の列であり, 同時に $Z_2(S)$ の最初の列である．$Z_1(S)$ の最初の列（$Z_2(S)$ の 2 番目の列）はアルファベット順に並んでいることにより, BW(S) から導くことができ, SORT(S) と表される．逆変換を始めるとき $Z_1(S)$ と $Z_2(S)$ の残りの列はすぐには分からないが, ともに S のすべての循環シフトを含んでいるはずである．それらは 2 つの行列では異なる順番で現れるが, 1 対 1 の対応関係がある（図 2.8 中の矢印）．この対応関係から生じる行番号 i の変換を

$\gamma(i)$ で示す. 行列 $Z_1(S)$ の i 行を一文字だけ右から左へ循環シフトすれば, 行列 $Z_1(S)$ の $j = \gamma(i)$ 行へと変化する.

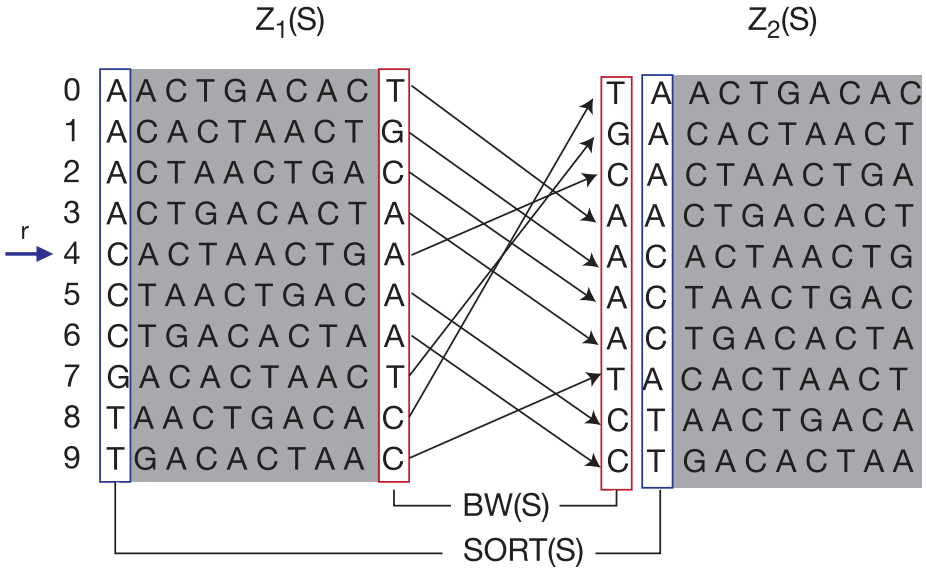

図 2.8　逆 BW 変換

S は 2 通りのやり方で再構築できる（図 2.9）. 左から右へ向かって再構築する場合は, 位置 $i = 4$ から初めて, $SORT(S)$ の文字に対し $\gamma(i)$, $\gamma(\gamma(i))$, ... を次々に適用する. 一方, 右から左へは, 再び位置 $i = 4$ から始めて $\gamma^{-1}(i)$, $\gamma^{-1}(\gamma^{-1}(i))$, ... を次々に BW(S) の文字に適用することで可能である.

BW 変換を用いたパターン検索　$S\$ = \text{CACTAACTGA\$}$ のように, (2.1) の最後に\$を付け加え, 前述のように文字列の終わりのラベルとしてこれを用いる. S\$の BW 変換は $\text{BW}(S\$) = \text{TCAG\$AATCCA}$ となり, BW(S\$) と SORT(S\$) を比較することにより得られる変換 $\gamma(i)$ は表 2.3 のようになる.

表 2.3　BW(S\$) と SORT(S\$) の比較による $\gamma(i)$

i	0	1	2	3	4	5	6	7	8	9	10
$\gamma(i)$	2	5	6	10	1	8	9	3	0	7	4

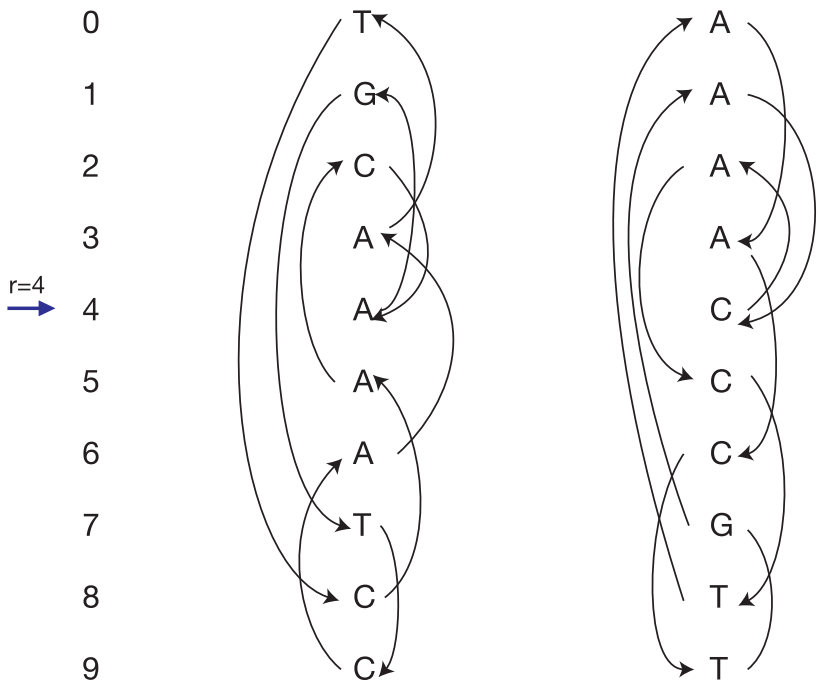

図 2.9　逆 BW 変換による S の構築

あるパターンが S 内に何回出現するかを数えるために BW 変換を応用する例としてこれを用い, S 内にパターン ACT が何回現れるかを調べる．SORT(S$) の要素に $\gamma(i)$ を適用することによって逆 BW 変換を実行する．SORT(S$) で出現するすべての文字 A に対して $\gamma(i)$ を適用し, $\gamma(A)$ の行き先に文字 C を持つ SORT(S$) の要素がいくつあるかを見る．さらに, 次のステップとして AC の次の文字として T を持つものがいくつあるかを見る．この様子は変換 $\gamma(i)$ を矢印によって図式的に表した図 2.10 に示されている．「矢印をたどる」ことによって S$における部分文字列 ACT の出現回数は 2 であることが分かる．このように与えられた文字列 S に対して BW 変換の圧縮された形が与えられると, パターン検索を実行することができ, S 内の P の出現の回数を P の長さに比例した時間で計算することができる [52].

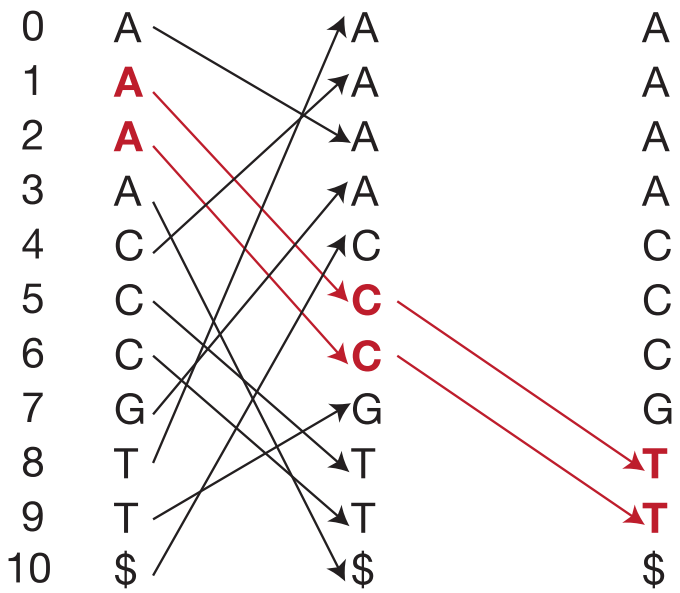

図 2.10　BW 変換を用いた文字列 ACT のパターン検索

2.2.2.2　スプライシングの検出

スプライシングの検出とは, RNA-Seq データからスプライスジャンクションを見つけ, スプライシングが どのように起きているかを解析することである. 成熟 mRNA をシーケンスする RNA-Seq においては, リードとゲノム配列を単純に比較するだけではスプライスジャンクションにまたがるリードはマッピングできない.

そこで分割したリードをマッピングすることによりスプライシングを検出し, ジャンクションを特定する. 代表的な手法として TopHat[53], SpliceMap[54], MapSplice[55] などがある.

2.2.2.3　アイソフォーム推定

アイソフォーム推定とは, RNA-Seq データのマッピング結果とスプライシングの検出において得られた情報から, 選択的スプライシングの発生とそれに伴い生成されるアイソフォームを推定することである. 図 2.11 のように, マッピングの結果から推定する手

法と, リードをアセンブルしてゲノム配列と比較することにより転写パターンを推定する手法が存在する. アセンブルとはリードをつなぎ合わせて元の配列（RNA-Seq においては mRNA 配列）を再構成することである. マッピングからの推定は EST, アセンブルによる推定は完全長 cDNA による推定にそれぞれ類似していると言える. 現在のところリード長が十分でないこと, 配列の読み取り誤差があることなどからアセンブルは非常に困難であり, Cufflinks[56] をはじめとしてマッピング結果からアイソフォームを推定する手法が広く用いられている.

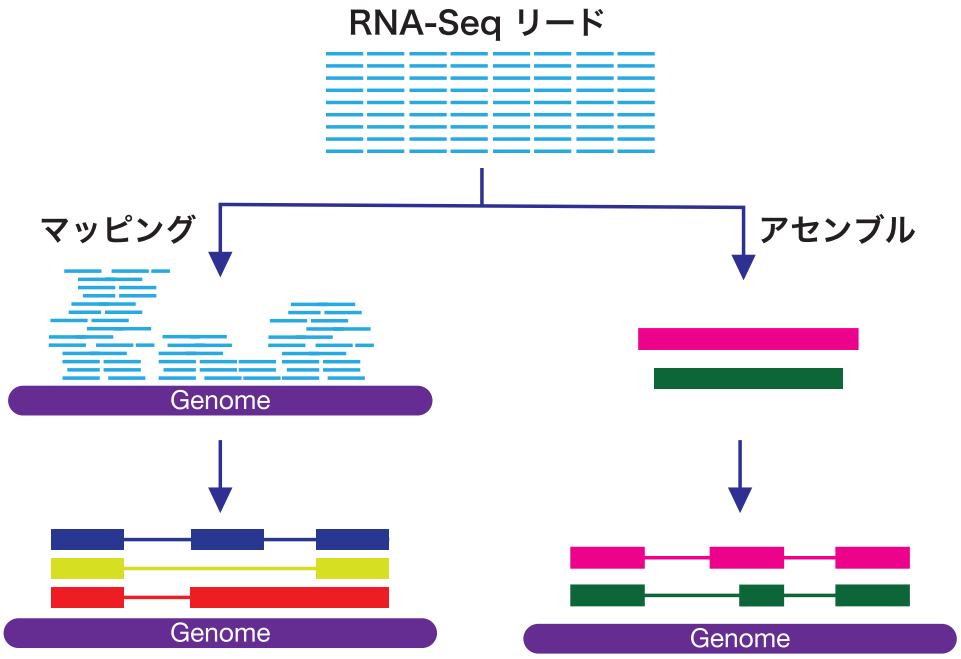

図 2.11　RNA-Seq を用いたアイソフォーム推定

2.2.2.4　発現量定量化

発現量定量化とは, マップされたリード数を正規化し各遺伝子または各アイソフォームの発現量を比較可能にすることである. リードの分布は独立同分布 (independent identical distribution; i.i.d) であると考えられるので RNA-Seq における発現量は, 各

箇所にマップされたリード数に対応する．そこで既知もしくは推定遺伝子モデルが与えられたときに, 各遺伝子もしくはアイソフォームに由来するリードを数え上げ, それを配列長で正規化することにより発現量を得る．

既知の遺伝子モデルを用いるものとしては ERANGE[57] が代表的である．ERANGE においては発現量の単位として RPKM(Reads Per Kilobase of exon model per Million mapped reads) が使用される．これは遺伝子ごとに発現量を正規化するものであり, 以下のように定義される．

$$\mathrm{RPKM} = 10^9 \times \frac{c}{Nl} \tag{2.3}$$

ここで, c は遺伝子にマップされたリード数, N はマップされた全リード数, l は遺伝子内でのエキソン長であり, 与えられた遺伝子モデルから求められる．

一方, Cufflinks は遺伝子モデルを使用せず推定された各アイソフォームに対してリードを配分することでアイソフォームレベルの発現量を推定する．FPKM (Fragments Per Kilobase of exon per Million fragments mapped) という単位が用いられ, その定義は

$$\mathrm{FPKM} = 10^9 \times \frac{f}{Nl} \tag{2.4}$$

で与えられる．ここで f は各アイソフォームにおいてエキソンにマップされたリード断片の数である．単なるリード数でないのは, ジャンクションにまたがるリードも考慮される ためである．FPKM は式 (2.3) で定められる RPKM と類似しているが, RPKM は既知の遺伝子モデルを用いてエキソンにマップされたリードを数え上げ, 遺伝子ごとの発現量を計算するのに対して, FPKM は推定されたアイソフォームごとにフラグメントを数え上げ, アイソフォームレベルの発現量を計算している．

また, シーケンスベースの遺伝子発現解析手法である SAGE の解析においては

TPM(Transcripts Per Million)[58] が用いられた. TPM は以下のように定義される.

$$\tau_i = \frac{\nu_i}{l_i} \left(\sum_j \frac{\nu_j}{l_j} \right)^{-1} \tag{2.5}$$

ここで l_i は各アイソフォームのエキソン長, ν_i はトランスクリプトームにおいて各アイソフォームに由来するフラグメントの塩基数である. 異なる実験間での発現量の比較のために TPM が用いられることもある.

2.2.3 従来手法の問題点

高速シーケンサによるシーケンシングで出力されたリードは数十〜数百 bp と短いため, マッピングの際に発現箇所が必ずしも一箇所に特定できるとは限らない. マッピングの結果, 発現箇所が一意に特定することができたリードをユニークリード (unique read), 複数箇所にマップされたリードをマルチリード (multiread) と呼ぶ.

マルチリードの発現箇所特定には正確な発現量が必要であるが, 発現量定量化にはマルチリードの発現箇所を正確に特定する必要がある. つまりこれらの 情報は不可分であり, マッピング後に発現量を定量化してその情報を基にマッピングをし直すという繰り返しが必要である. しかし, 従来手法ではマルチリードの配分は, ユニークリードで計算した発現量にのみ基づいており, マッピングの繰り返しは行われていない. これにより推定発現量と実際の発現量に差が生じ, 更にアイソフォーム推定の際に誤予測の原因になると考えられる.

2.3 提案手法

2.3.1 提案手法の概要

本研究ではアイソフォームの推定精度を向上させることが目的である. 3.3 節で述べたように, 従来手法はマッピングが繰り返されておらず, 推定発現量が実際とは異なり誤っ

たアイソフォーム推定へとつながっていると考えられる．そこでマッピングを繰り返し行うことで精度の向上を目指す．

本手法には入力として以下を与える．

- RNA-Seq データセット
- リファレンス配列（ゲノム）
- 転写開始点・終了点情報

提案手法は以下の 4 つのステップからなる（図 2.12）．

1. マッピング・ジャンクション検出
2. アイソフォーム推定
3. 発現量定量化
4. 再マッピング

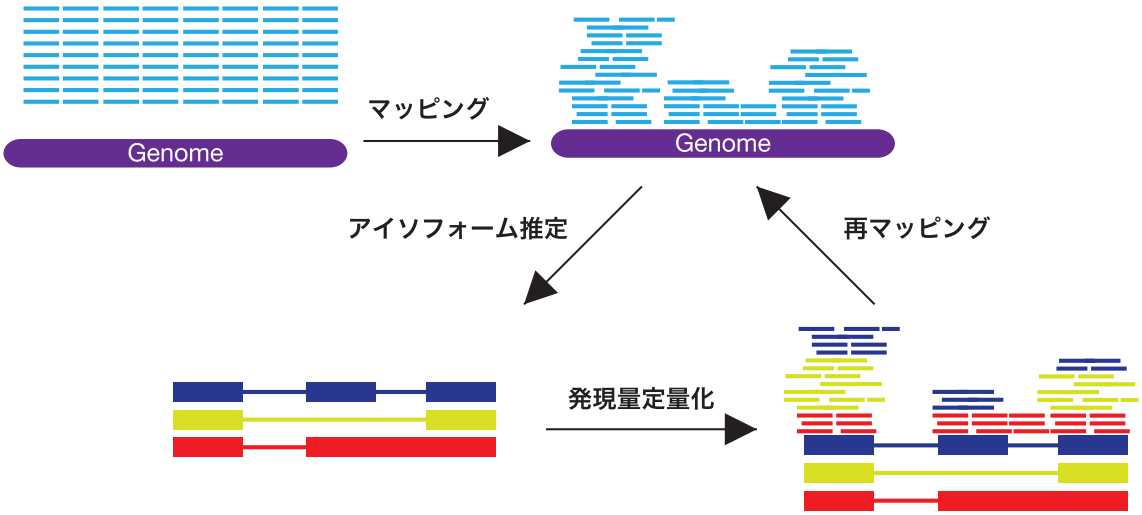

図 2.12　提案手法の流れ

2.3.2 提案手法の詳細

前節で述べた提案手法の 4 ステップのそれぞれについて詳細を説明する.

2.3.2.1 マッピング・ジャンクション検出

このステップでは高速シーケンサから得られたリードのマッピングを行い, 同時にスプライスジャンクションを検出する.

まず Bowtie を用いてマッピングを行う. ミスマッチの上限 2 つとして挿入・欠失なしでリファレンスとなるゲノム配列に対してアライメントする. マッピングの際の誤りとして

RNA-Seq データには, mRNA のスプライスジャンクションにまたがる領域から発現したリードも多数含まれるが, 単純なマッピングではそれらのリードは正しくマップされない. そこでジャンクションにまたがる領域から得られたリードを見つけ出し, そのリードがマップされた位置からジャンクションを検出する. リードを 25 塩基ずつに分割し一方に対してマッピングを行い, マップされたら 一定長以内でもう一方がマップされる箇所を探す. その中で GT-AG ルールを満たすような箇所をジャンクションとして検出する. (図 2.13).

これによりユニークリード・マルチリードが区別され, ジャンクション候補と各領域にどれだけリードがマップされたかという情報が得られる.

以上をまとめると, 本ステップのアルゴリズムは以下のようになる.

入力 RNA-Seq データセット, リファレンス配列

出力 マッピング位置, ジャンクション

1. Bowtie を用いたマッピング
2. ジャンクション検出
   (a) リードの分割

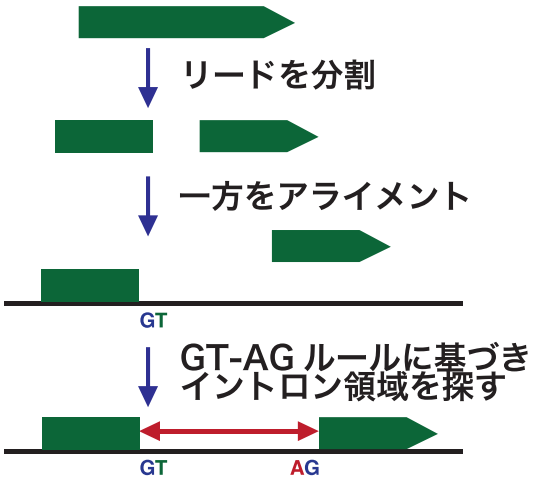

図 2.13　ジャンクション検出

(b) 一方をマップ

(c) GT-AG ルールに基づき他方をマップ

3. マッピング・ジャンクション検出結果の出力

2.3.2.2　アイソフォーム推定

以下の手順で 1 つの遺伝子のアイソフォームを推定する.

1. 推定遺伝子モデルの構築
2. アイソフォームの有無の判定 [59]
3. 推定アイソフォームの出力

推定遺伝子モデルの構築

アイソフォームを推定するために, まずゲノム上の転写領域を特定し, 遺伝子領域内でどこがエキソンに当たるかという推定遺伝子モデルを構築する. これはデータベースから取得した転写開始点・終了点情報により, 遺伝子領域を特定する. その遺伝子領域情報と, 前ステップで検出されたスプライスジャンクションの位置情報を組み合わせること

で実現される（図 2.14）.

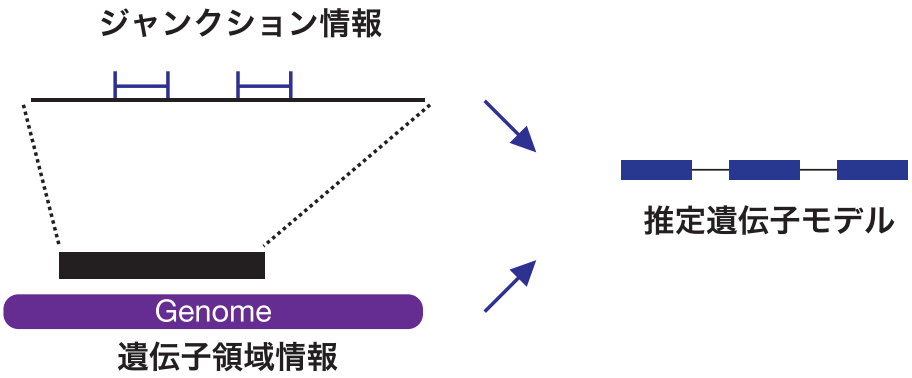

図 2.14　推定遺伝子モデルの構築

アイソフォームの有無の判定

複数のエキソンを持つと推定された遺伝子に対して, 各遺伝子内で選択的スプライシングが発生しているか, つまり複数のアイソフォームを持つかどうかを判定する. 各エキソンにリードがマップされる確率は多項分布に従うと仮定し, カイ二乗検定を行うことで判定する.

多項分布　多項分布 (multinomial distribution) とは, 二項分布を一般化した確率分布である. k 種類の事象が発生する確率をそれぞれ $p_1, \ldots, p_k(\sum p_i = 1, p_i \geq 0, \forall i \in (1, \ldots, k))$ としたとき, n 回の試行を行った結果それぞれの事象がそれぞれ x_i 回起こる確率は以下の式で表される.

$$f(x_1, \ldots, x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \tag{2.6}$$

パラメータは p と n であり, このような多項分布は $M(p, n)$ と表される. $k = 2$ のときは二項分布となる. 多項分布の平均 $E(x_i)$, 分散 $V(x_i)$ はそれぞれ

$$E(x_i) = np_i \tag{2.7}$$

$$V(x_i) = np_i(1-p_i) \tag{2.8}$$

となることが知られている. 観測値を n_i, 期待値を $m_i = np_i$ $(i = 1, \ldots, k)$ とし, m_i, n_i が十分大きいとき, 変量 $\chi^2 = \sum \frac{(n_i - m_i)^2}{m_i}$ は近似的に自由度 $k-1$ の χ^2 分布に従う.

カイ二乗検定 カイ二乗検定とは, 観測されたデータが期待したものとどれだけずれているか, もしくはどれだけ合っているかを検定するものである. 検定は, 観測された事象における相対頻度（観測度数）を, 期待する理想の分布（期待度数）と比較することで行われる. 帰無仮説 H_0 を「観測度数は期待度数に従う」とし, 観測度数と期待度数のずれ（χ^2:カイ二乗値）を観測値 O と期待値の差の二乗を期待値 E で割って合計したものとして算出する.

$$\chi^2 = \sum \frac{(E-O)^2}{E} \tag{2.9}$$

算出されたカイ二乗値から, 自由度 k に基づくカイ二乗分布の確率密度関数 $f(x, k)$, $x > 0$ を用いて, 求めたカイ二乗値をとる場合に帰無仮説 H_0 が成立する確率 $P(\chi^2)$（P 値）を求める.

$$f(x,k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} e^{-\frac{x}{2}} x^{\frac{k}{2}-1} \left(\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \right) \tag{2.10}$$

$$P(\chi^2) = 1 - \int_0^{\chi^2} f(x,\, k)\, dx \tag{2.11}$$

このようにして得られた P 値が有意水準より小さければ帰無仮説は棄却され, 対立仮説が採用される. 逆に P 値が有意水準より大きければ, 帰無仮説が採用される.

カイ二乗検定を用いた選択的スプライシングの発生の有無の判定 カイ二乗検定を用いて, 選択的スプライシングが発生しているか, つまりその遺伝子に複数のアイソフォー

ムが存在するかを判定する.

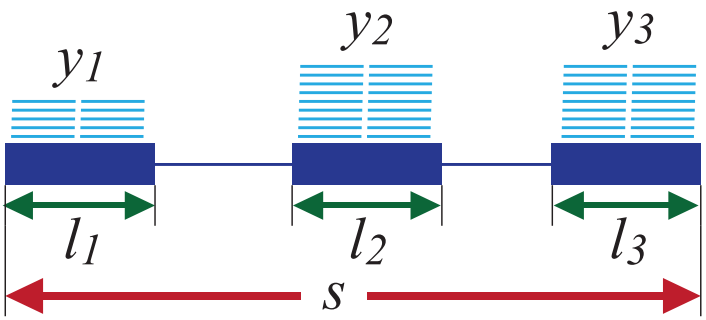

図 2.15　推定遺伝子モデル

図 2.15 のような長さ s の遺伝子内の各エキソン $e_1, \ldots, e_n$ (長さ $l_1, \ldots, l_n$, マップされたリード数 $y_1, \ldots, y_n$) に対して, マップされたリード数から選択的スプライシングの有無を判定する. $Y = (y_{e_k})_{k=1,\ldots,n}$ は, 選択的スプライシングが起きていない場合は多項分布 $M((p_e)_{e=1}^n, T)$ に従うと考えられる. なお, $T = \sum y_k$ であり, 遺伝子全体のリード数の観測値である. また, p_e は以下の式で表される.

$$p_e = \frac{l_e}{\sum_{i=1}^n l_i} \tag{2.12}$$

そこで以下の前提で自由度 $n-1$ のカイ二乗検定を行う.

- 帰無仮説: 選択的スプライシングが起こっていない
- 対立仮説: 選択的スプライシングが起こっている

p 値が有意水準（ここでは 0.05）より小さければ帰無仮説が棄却され, その遺伝子では選択的スプライシングが起きていると判定される.

例として, 図 2.15 で $(l_1, l_2, l_3) = (100, 100, 100)$, $(y_1, y_2, y_3) = (30, 60, 60)$ となったときのアイソフォームの有無をカイ二乗検定により判定する. ここではエキソン長はすべて同一であるため, リード数の期待値はそれぞれ $(30+60+60)/3 = 50$ となるので, カイ二乗値は以下のように計算できる.

$$\begin{aligned}\chi^2 &= \frac{(50-30)^2}{50} + \frac{(50-60)^2}{50} + \frac{(50-60)^2}{50} \\ &= 12\end{aligned}$$

この場合は自由度が 2 であるため, 自由度 2 の確率密度関数 $f(x) = \frac{1}{2}e^{-\frac{x}{2}}$ を用いて P 値は次のように求められる.

$$\begin{aligned}P(12) &= 1 - \frac{1}{2}\int_0^{12} e^{-\frac{x}{2}}dx \\ &= 1 + \left[e^{-\frac{x}{2}}\right]_0^{12} \\ &= e^{-6} \approx 0.00248\end{aligned}$$

このようにして求められた P 値は有意水準 0.05 よりも小さいことから帰無仮説は棄却され, 対立仮説が採用される. つまり, この遺伝子では選択的スプライシングが発生し, 複数のアイソフォームが存在していると判定されたことになる.

推定アイソフォームの出力

選択的スプライシングが起きていると判定された遺伝子で, どのエキソンが選択的スプライシングに関わっているかを特定するために, 各遺伝子内のエキソンのうち, 発現量が有意に大きい, または小さいものを選び出す.

各エキソンに対して以下のように発現量 R_e を計算する.

$$R_e = \log(\tilde{y}_e) = \log \frac{y_e}{l_e \cdot s} \tag{2.13}$$

各エキソンに対して以下の式に従い Z-Score を計算する.

$$z_e = \frac{R_e - \mathrm{median}(R)}{\mathrm{MAD}(R)} \tag{2.14}$$

ここで MAD とは Maximum Absolute Deviation のことであり, 以下のようにして求められる.

$$\mathrm{MAD} = \max |R_e - \mathrm{median}(R)| \tag{2.15}$$

Z-Score の絶対値が閾値以上の場合はそのエキソンが選択的スプライシングに関わっていると判定する. 選択的スプライシングに関わるエキソンが発現している転写パターン, 発現していない転写パターンを推定アイソフォームとして出力する（図 2.16）.

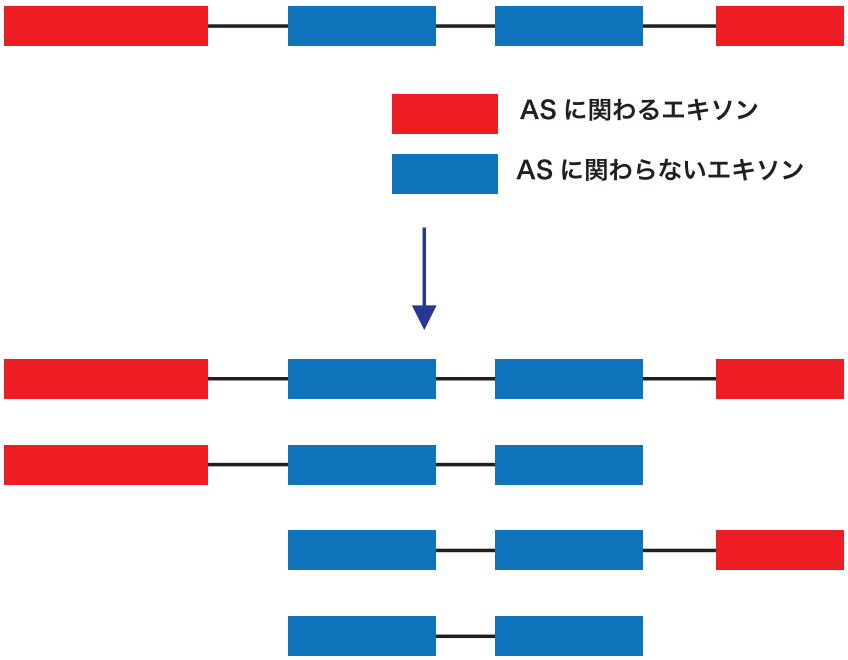

図 2.16　推定アイソフォーム

2.3.2.3　発現量定量化

これまでの過程で各遺伝子について推定アイソフォームと各エキソンのリード数が得られたことになる. しかし各リードがどのアイソフォームから発現したものなのかはこの時点では分からない. そこで本ステップでは各アイソフォームにリードを配分することにより 1 つの遺伝子内で各アイソフォームがどれだけ発現しているかを定量化する

(図 2.17 において発現由来アイソフォーム毎に各リードを色付けする作業に相当).

図 2.17　発現量定量化

エキソン e がアイソフォーム j において発現していれば $I_{e,j} = 1$, そうでなければ $I_{e,j} = 0$ となるようなバイナリ行列を I とし, T_j をアイソフォーム j に由来するリード数とすると, 以下の関係が成り立つ.

$$Y_e = \sum_{j \in \text{isoforms}} \frac{p_e}{\sum_i p_i \cdot I_{i,j}} \cdot I_{e,j} \cdot T_j \tag{2.16}$$

式 (2.16) において T_j は未知であるが, これが分かれば各アイソフォームごとの発現量が得られることになる. しかし一般に式 (2.16) においては T_j は一意に定めることができないため, 左辺と右辺の差が最小になるように T_j を推定する必要がある.

EM アルゴリズム EM(Expectation-Maximization) アルゴリズムは, 観測できないデータがある場合の最尤推定 (maximum likelihood estimation) のためのアルゴリズムである [60]. 反復法の一種で, 期待値 (expectation; E) ステップと最大化 (maximization; M) ステップを交互に繰り返すことにより確率モデルのパラメータを推定する手法であり, 期待値最大化法とも呼ばれる. E ステップでは, 現在推定されている潜在変数の分布に基づいてモデルの尤度を計算する. M ステップでは E ステップで求まった尤度の期待値を最大化するようなパラメータを求める. M ステップで求められたパラメータは, 次の E ステップで使われる潜在変数の分布を決定する為に用いられる.

x を観測データ, y を観測できないデータ, θ をパラメータとするとき, 次の式で定

義される対数尤度 (log likelihood) を最大化する θ を計算することが目標となる.

$$\log P(x|\theta) = \log \sum_y P(x, y|\theta) \tag{2.17}$$

ただし, 実際には最大化が困難であるため, 反復によって対数尤度を単調に増加させ, その極大化を図る. そこで, ランダムもしくは他の何らかの方法で定めた初期パラメータを θ^0 とし, t 回反復させたあとのパラメータを θ^t とする. ここで, 式の変形を行うことによって EM アルゴリズムを導く. まず, 上で定義された対数尤度の式の左辺を次のように変形する.

$$\log P(x|\theta) = \log P(x, y|\theta) - \log P(y|x, \theta) \tag{2.18}$$

この式の両辺に $P(y|x, \theta^t)$ を掛けて y についての和をとり, 次の式を得る.

$$[\log P(x|\theta) = \sum_y P(y|x, \theta^t) \log P(x, y|\theta) - \sum_y P(y|x, \theta^t) \log P(y|x, \theta) \tag{2.19}$$

右辺の第 1 項を $Q(\theta|\theta^t)$ とおくと, 次の式が得られる.

$$\log P(x|\theta) - \log P(x|\theta^t) = Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \sum_y P(y|x, \theta^t) \log \frac{P(y|x, \theta^t)}{P(y|x, \theta)} \tag{2.20}$$

最後の項は相対エントロピーであるので常に非負であるため, 次の式が成立する.

$$\log P(x|\theta) - \log P(x|\theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t) \tag{2.21}$$

よって, $\theta = \theta^t$ とおけば右辺は 0 になるので, $\theta^{t+1} = \arg\max_\theta Q(\theta|\theta^t)$ とすることにより右辺が非負となり, 尤度は増大するか変化しないままとなる.
以上より, EM アルゴリズムは次のとおりとなる.

1. 初期パラメータ θ^0 を決定し, $t = 0$ とする.
2. $Q(\theta|\theta^t) = \sum_y P(y|x, \theta^t) \log P(x, y|\theta)$ を計算する（E ステップ）.
3. $Q(\theta|\theta^t)$ を最大化する θ^* を計算し, $\theta^{t+1} = \theta^*$ とし, 更に $t = t + 1$ とする（M ステップ）.
4. $Q(\theta|\theta^t)$ が増加しなくなるまでステップ 2 と 3 を繰り返す.

EM アルゴリズムを用いた発現量定量化 前述のとおり, 式 (2.16) において, EM アルゴリズムを用いて各アイソフォームの発現量を推定する.

遺伝子全体のリード数 T はポアソン過程 Poisson($\lambda \times s \times p$) に従うと考えられる. ここで s は遺伝子長, p は相対発現量, λ は正規化要素である. ポアソン過程は, 正規分布に当てはまらないような低発現データを処理することができ [61], EST や SAGE においても効果を発揮した [62, 63].

アイソフォーム j におけるエキソン e のリード数を Y_e^j とし, $Y = (Y_e)_{e=1\ldots n}$ とすると, Y_e^j は以下の階層モデルに従う.

$$T_j \approx \text{Poisson}(\lambda_j) \text{ with } \lambda_j = \lambda \cdot \frac{1}{\sum_i p_i \cdot I_{i,j}} \cdot q_j$$
$$(Y_1^j, \ldots, Y_n^j)|T_j = m_j \approx M\left(\left(\frac{p_e}{\sum_i p_i \cdot I_{i,j}} \cdot I_{e,j}\right)_{e=1}^{n}, m_j\right)$$
$$\forall i \in (1, \ldots, k)$$

ここで q_j はアイソフォーム j の相対発現比であり, λ は sequencing depth と transcript の長さに対する正規化要素である. $I_{i,j}$, y_i, p_i が与えられたとき, λ と q_j を求めるために EM アルゴリズムを適用する. 各アイソフォームの発現比 P_j を推定することで. 隠れ変数 T_j が求められる. ここでアイソフォーム j のエキソン i に由来するリード数を $y_{i,j}$, アイソフォーム j に由来するリード数を m_j とすると, 完全データの尤度は以下のように表される.

$$P(y_{1,1}, \ldots, y_{n,k}) = \prod_{j=1}^{k} \left(\frac{\exp \lambda_j \lambda_j^{m_j}}{m_j!} \cdot \begin{pmatrix} m_j \\ y_{1,j} \cdots y_{n,j} \end{pmatrix} \cdot \prod_{i=1}^{n} \left(\frac{p_i}{\sum p_l I_{i,j}} \cdot I_{i,j} \right) \right)$$

$$\log P(y_1, \ldots, y_n) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i!$$

$$\text{with } \mu_i = \lambda \cdot p_i \sum_{j=1}^{k} q_j I_{i,j}$$

E ステップ, M ステップはそれぞれ以下のように表される. そこで m_j にランダムな初期値を与えてこれらのステップを繰り返す.

E-step

パラメータが既知であると仮定すると, ステップ v におけるアイソフォーム j からの観測リード数は以下のように表される.

$$\hat{m}_j = \mathbb{E}_{q_j}(v)(Y_j | c_1, \ldots, c_n) = \sum_{i=1}^{n} \frac{p_i q_i^{(v)} I_{i,j}}{\sum_{l=1}^{k} p_l q_l^{(v)} I_{i,l}} \cdot c_i \tag{2.22}$$

M-step

以下の式に従い尤度を最大化する.

$$\hat{\lambda} = \sum_j \hat{m}_j = c$$

$$\hat{q}_j = \sum_i p_i I_{i,j} \cdot \frac{\hat{m}_j}{c}$$

対数尤度の相対変化量が閾値 ϵ より小さくなれば収束したと見なし繰り返しを終了する.

Quality score

EM アルゴリズムは時として局所解に陥ることがある. そこで, 得られた解がどれだけ尤もらしいかを調べる. エキソン e の推定リード数 $Y_e^{\exp}$ に対して以下の値を計算する. $Y_e^{\exp}$ は T_j と q_j から計算できる.

$$\chi_G^2 = \sum_{e=1}^{n} \frac{(y_e - Y_e^{\exp})^2}{Y_e^{\exp}} \tag{2.23}$$

χ_G^2 は自由度 $n-1$ のカイ二乗値である．quality score $= \log_{10}$ p-value として Quality を計算し，これが一定の値より小さければ EM アルゴリズムを用いた発現量推定をやり直す．

上記のようにして遺伝子内の各アイソフォームの相対発現量を得る．この発現量定量化ステップは選択的スプライシングが起こっていると判定された遺伝子のみに対して行われる．

2.3.2.4 再マッピング

従来手法においては，マルチリードの発現由来箇所が特定されず，各箇所に由来するリード数のカウントに誤りが生じ，結果としてアイソフォーム推定の際に誤りが生じることが問題であった．本ステップでは，マルチリードの発現由来箇所特定のために再マッピングを行う．

ある箇所の発現量が大きいほど，その箇所由来のマルチリードも多いと考えられる．そこで，マルチリードがマップされたエキソンについて発現量を計算し，その推定発現量に比例してマルチリードを配分する．

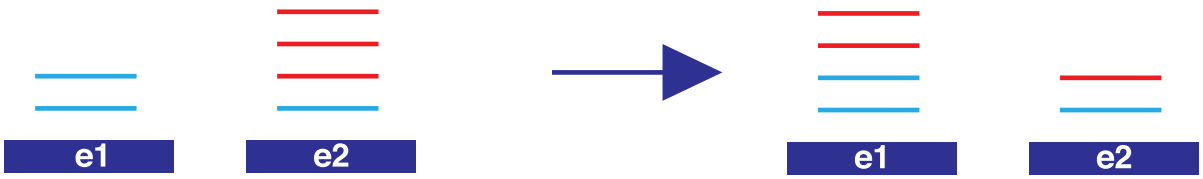

図 2.18　再マッピング

図 2.18 では青線がユニークリード，赤線がマルチリードを表す．ユニークリードはエキソン e_1 が 2 本，エキソン e_2 が 1 本なのに対して，e_2 には 3 本のマルチリードがマップされている（図 2.18 左：再マッピング前）．これは実際の発現の様子を再現していな

いと考えられる. そこでマルチリードの 2 本を e_2 から e_1 へマップし直すことで実際の発現量と対応させる. この後, 再度選択的スプライシング推定・発現量定量化を行う,

2.4 実験

提案手法の有効性を評価するために以下のような実験を行った.

2.4.1 使用データ・実験条件

以下のデータを提案手法に適用した.

RNA-Seq **データ** ヒトの脳の RNA-Seq データ

- 75 塩基
- 16,748,521 リード

リファレンス配列 UCSC hg19 （ヒトゲノム配列）

転写開始点・終了点 Ensembl GRCh37.59 (hg19 に対応)

Cufflinks にも同データを適用して, それぞれの結果を既知のアイソフォームデータベースと比較し, その正解率を比較する. ここでは既知のアイソフォームのデータベースとして Ensembl を使用した. Ensembl には生物学的実験により発現が確認された多数のアイソフォームが登録してあり, それらの推定精度により提案手法の有用性を検証する.

遺伝子発現においては, すべての遺伝子のすべてのアイソフォームが同時に発現するというわけではない. そのため, 推定されたアイソフォームのうちどれだけが “正解” であるかが重要となる. そこで, 推定されたアイソフォームが Ensembl に登録してあれば “正解” とし, どれだけ正確にアイソフォームが推定されたかを算出し, 従来手法である Cufflinks と比較する.

2.4.2 実験結果

2.4.2.1 ゲノム全体での解析

ゲノム全体にわたり提案手法, 従来手法 (Cufflinks) それぞれでアイソフォームの推定をし, その正解率を比較した. なお, 提案手法ではデータベースを参照することにより転写開始点・終了点を決定するが, Cufflinks はリードのマッピング結果から推定する. そのため, Cufflinks の評価の際には Ensembl 上で遺伝子領域とされている領域との共通集合を抽出しその精度を調査した.

表 2.4　アイソフォーム推定結果と Ensembl の比較

	出力アイソフォーム数	正解率
提案手法	29282	66.7
Cufflinks	26393	40.7

表 2.4 は推定されたアイソフォームと Ensembl の正解率の比較結果である.

表 2.5　ユニークリード/マルチリード数

ユニークリード	マルチリード
2,732,046	8,326,834

表 2.5 はマッピング後のユニークリードとマルチリードの数である.

表 2.6　エキソンとイントロンの推定精度

(%)	エキソン	イントロン
提案手法	92.6	96.4
Cufflinks	70.0	95.0

表 2.6 はエキソンとイントロンの推定精度を表にしたものである. 推定されたアイソフォームに含まれるエキソンとイントロンがそれぞれ Ensembl に登録されたものと

どれだけ一致したかという割合を計算したものであり, エキソン/イントロンレベルの precision に相当する. エキソン, イントロンともに提案手法が高い推定精度を示した.

2.4.2.2 特定遺伝子におけるアイソフォーム推定結果

図 2.20 に提案手法と Cufflinks で推定された結果を示す. これは遺伝子名 ACTB の領域での転写パターンに相当する. 遺伝子 ACTB は, アクチンというタンパク質を生成する遺伝子であり, 複数種のアイソフォームを持つことが知られている. またこの遺伝子は様々な生物種に共通して存在しており, それぞれの生物で多様な遺伝子ネットワークと関連があるとされており（図 2.19）, この遺伝子のアイソフォームの推定は遺伝子発現解析において有用であると考えられる.

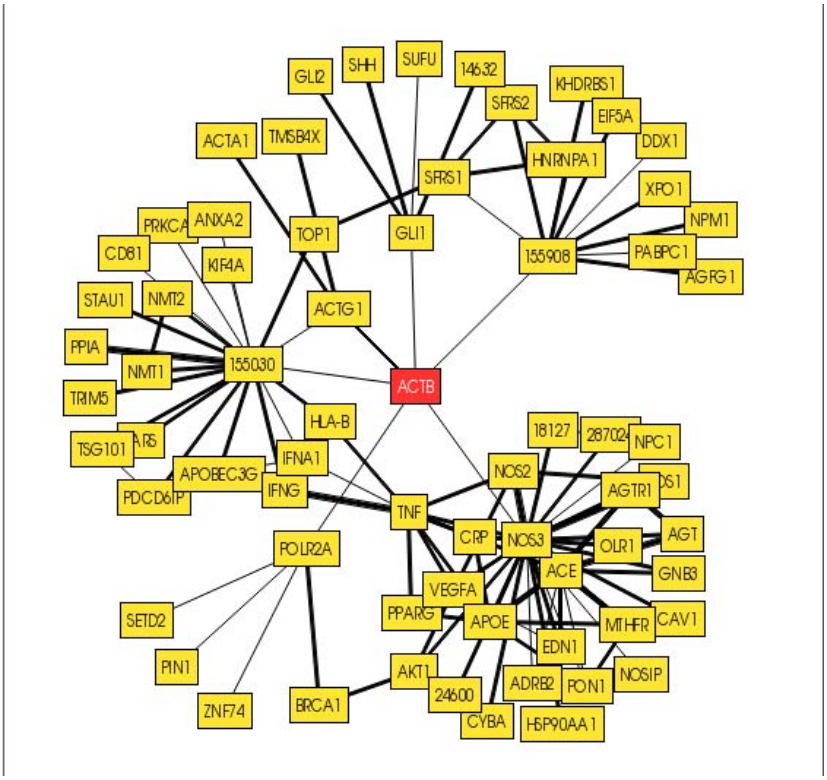

図 2.19　ACTB と他の遺伝子の関連

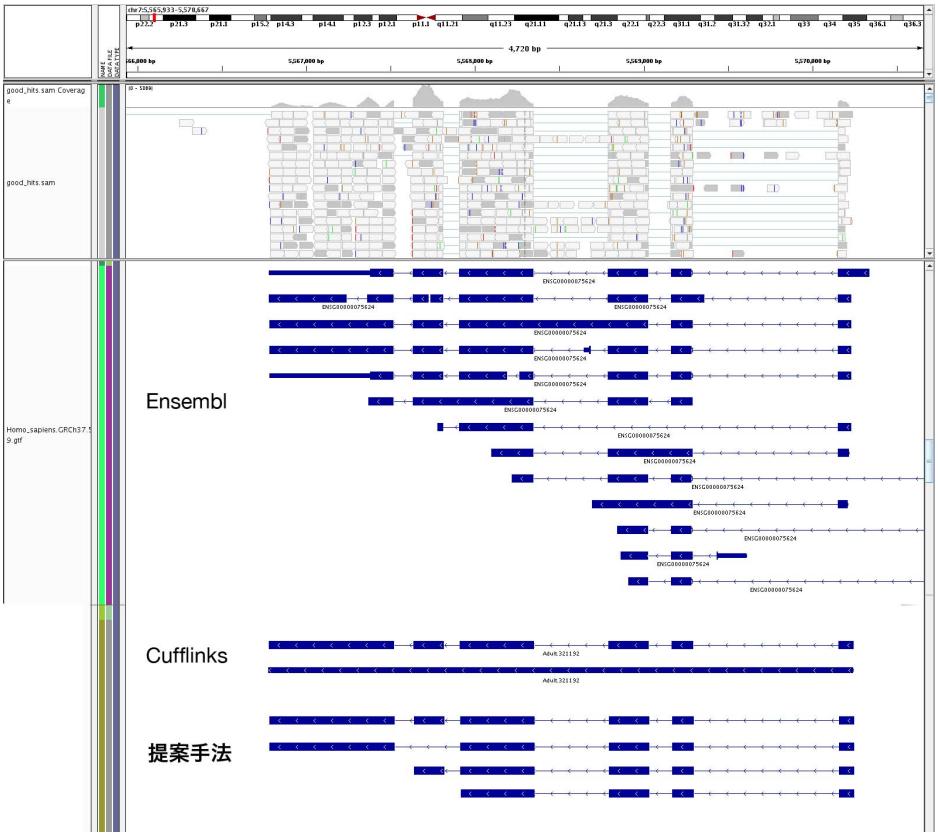

図 2.20　提案手法と Cufflinks の比較

アイソフォームを推定した結果, 提案手法では 4 つのアイソフォームが推定され, うち 2 つは Ensembl に登録されているものと一致したのに対して, Cufflinks では選択的スプライシングが推定されなかった. 図 2.20 が示すように, Cufflinks で推定できなかったアイソフォームも正確に推定できており, 改善を示した.

2.4.3　考察

Cufflinks に比べ, 高いアイソフォーム推定精度を示し, 目的であるアイソフォーム推定精度向上は達成できた. 第 2 章で述べたように, Ensembl の登録アイソフォーム数は短期間で大量に増加しており（表 2.1）, 今回使用した Ensembl のバージョン

(GRCh37.59) では “不正解” と判定されたアイソフォームの中にも, 今後発現が確認され Ensembl の将来のバージョンに追加される新規アイソフォームが含まれる可能性もある. そこで新規アイソフォームが含まれる可能性の 1 つの指標として, エキソン・イントロンがどれだけ正しく検出されているかを調査した（表 2.6）. その結果提案手法では高い精度でエキソン・イントロンを検出できており, 不正解と判定されたアイソフォームの中にも新規アイソフォームが含まれる可能性を示した. これは再マッピングにより, 従来手法では発現箇所と推定されたものの, 実際には発現していないような, False-Positive のエキソンを減らすことができた結果であると考えられる. 実際に図 2.20 では, Cufflinks はイントロン部分にマップされたマルチリードの存在から, 遺伝子全体を single exon と見なしたアイソフォーム推定をしている. しかしマッピング結果から候補アイソフォームを出力する際には予め形状が指定されたエキソンごとに 選択的スプライシングに関わっているかどうかを推定するため, アイソフォーム同士で異なる形状のエキソンが重なり合っているときにはそれらを識別することが出来ない.

提案手法では発現量比に基づきマルチリードを配分するため, 発現量の低いアイソフォームから得られたマルチリードが, 同様の配列を持つ高発現のアイソフォームに配分され, 結果として低発現のアイソフォームの一部が推定されなくなる可能性がある. しかし一方で，従来手法の Cufflinks では，発現量比で考えると本来は高発現領域に配分されるべきマルチリードまで低発現領域に配分されることにより，高発現領域のアイソフォームの一部が検出されなくなる可能性もある. 提案手法と Cufflinks でそれぞれ出力されたアイソフォーム数を比べると，表 2.4 に示すように提案手法の方が出力アイソフォーム数が多く Ensembl と比較した時の正解率も提案手法の方が高い. このことから，再マッピングによるマルチリードの配分はアイソフォームを正しく検出するのに寄与していると考えられる.

表 2.5 に示すように, マップされたリードのうち約 75% がマルチリードであり, アイソフォーム推定および発現量推定の精度を向上させるためには提案手法のような反復マッピングが不可欠であると考えられる. ただしこの例のようにユニークリードがある

程度の量存在することが前提となる. 例えば *maize*（トウモロコシ）のように繰り返し配列が大量に存在するような場合には, ユニークリードがほとんど存在せず反復マッピングを実行してもリードの分布は収束せず, 正確な発現解析は行えない可能性がある. 逆にマルチリードが少ない場合には従来手法でも十分な推定精度で発現解析を行うことができると考えられる. しかし, ヒトをはじめとしたアイソフォームが多数存在するような哺乳類では一般にユニークリード・マルチリードが大量に存在し, 提案手法によるアイソフォーム推定は有用であると考えられる.

2.5 結言

本研究ではアイソフォーム推定のための RNA-Seq 解析手法を提案した. 発現量を推定後, 発現レベルに応じてマルチリードを配分しなおす再マッピングを行うことにより, マッピングミスを減らしアイソフォーム推定精度の向上を目指した.

従来手法とアイソフォームの推定結果を比較して高い正解率を示した. さらにエキソン/イントロンの推定精度も高く, 新規アイソフォーム検出の可能性も示した.

今後の課題はマッピング結果からの遺伝子領域特定を可能にすることである. 本手法では転写開始/終了点データベースを用いたが, データが登録されていない生物種に対しての解析や新規発現領域などに対しても適用可能にするように改良が求められる. その他, アイソフォームごとの発現量推定の際の計算時間の短縮等が今後の課題として挙げられる.

第 3 章　正準相関分析を用いたマイクロ RNA-遺伝子間相互作用予測手法

3.1　緒言

マイクロ RNA (miRNA) は 18〜25 塩基程度の一本鎖スモール RNA であり, タンパク質への翻訳はされないノンコーディング RNA の一種である [64]. miRNA は遺伝子発現調節機能を持ち, 細胞の発生や分化などのさまざまな生物学的現象において重要な役割を果たしていると考えられている.

ヒトゲノム中に存在する 1000 種類を超える miRNA はヒト遺伝子の約 60% に対して作用すると考えられており, がんや心臓病, 肥満など様々な疾患との関連も報告されている [65]. 近年では, miRNA は創薬ターゲットとしても注目を浴びつつあり, miRNA の機能解明を目指した研究や医薬品開発が進められている. [66, 67, 68].

現在のところ miRNA-遺伝子間相互作用を実験的に高速に検証する技術は確立されていない. そのため計算学的な手法を用いて miRNA-遺伝子間相互作用を予測する必要がある.

本章では正準相関分析を用いて miRNA-遺伝子間相互作用を予測するための手法を提案する.

3.2　miRNA の機能

miRNA は真核生物の遺伝子発現を転写後レベルで抑制する機能を持つ [69]. 植物や線虫ではゲノムからの遺伝子転写自体を抑制することも報告されている [70]. miRNA は通常ゲノム上の非遺伝子領域に存在するが, タンパク質コーディング遺伝子のエクソン内に存在する miRNA も報告されている [71, 72]. miRNA は一旦 primary miRNA

(pri-miRNA) と呼ばれる単鎖 RNA に切り出された後, miRNA 前駆体 (pre-miRNA) を経由し, 最終的な 22 塩基程度の成熟 miRNA (mature miRNA) を形成してターゲット遺伝子に結合することにより遺伝子の翻訳を阻害する. ゲノム上にある miRNA の生合成過程を図 3.1 に示す [73].

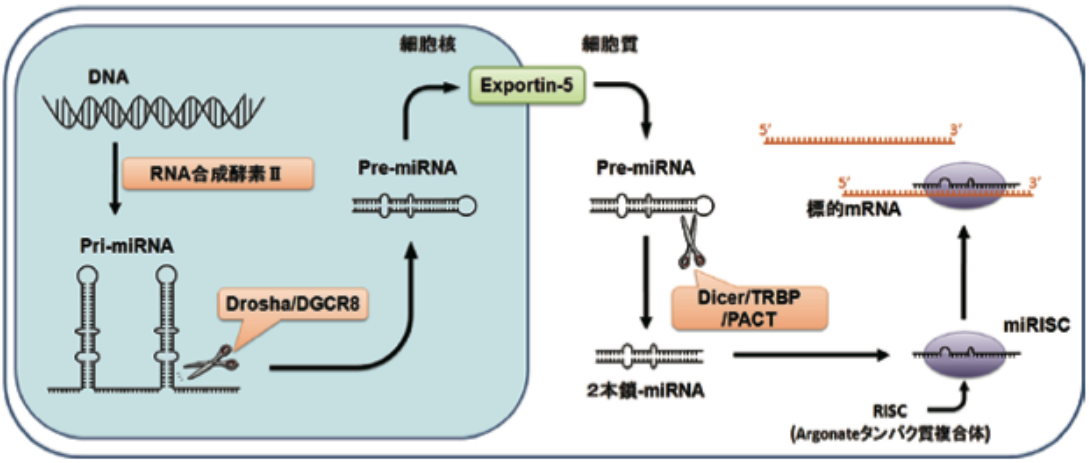

図 3.1　miRNA の生合成

核内でゲノムから転写された長い単鎖 RNA から, RNA ポリメラーゼ II によって miRNA の配列部分を含む数百〜数千塩基長の pri-miRNA が転写される [74]. pri-miRNA には, 配列依存的に複数のステムループ構造が生じるが, Drosha と呼ばれる酵素により切断され, ヘアピン状の小さな単鎖 RNA を構成する. これが miRNA 前駆体となり, Exportin-5 という核外輸送タンパク質によって核内から細胞質に輸送される. pre-miRNA は一般には Dicer と呼ばれる酵素とその補因子 (TRBP) からなる複合体によって両端が切断され, 22 塩基対程度の短い二本鎖 RNA を形成した後, 成熟 miRNA になると考えられているが, 遺伝子のイントロン領域に存在する mirtron と呼ばれるクラスの miRNA は Drosha の代わりに Ldbr (Lariatdebranching enzyme) という酵素による修飾を受けることが明らかにされたり, Dicer を必要としない miRNA 産生経路が発見されるなど, pre-miRNA 生成過程は現在盛んに研究が進められている [75, 76, 77].

二本鎖 pre-miRNA はその後乖離して, ガイド鎖と呼ばれる側だけが選択的に残されて成熟し, パッセンジャー鎖と呼ばれるもう一方の鎖は分解される. 完成した成熟 miRNA は RISC (RNA-induced silencing complex) と呼ばれるリボ核酸と Argonature タンパク質の複合体に取り込まれ, 機能的 miRNA-RISC 複合体 (miRISC) を形成する.

こうしてつくられた miRISC は標的とする mRNA に接近して, 3'-非翻訳領域 (3'-UTR) にある相補的配列部位に miRNA を結合させ, 遺伝子翻訳を中断させるか, その遺伝子を分解して発現を阻害すると考えられている. miRNA が mRNA の完全な相補的配列部位に結合した場合には mRNA は分解されるが, miRNA の 5' 末端から数えて 2〜8 塩基目のシード配列を除き, 完全な相補的配列でなくとも遺伝子発現調整機能を果たす. このような場合には配列依存的に阻害効果の強弱が決定される.

1 つの miRNA は単一の遺伝子を標的とするわけではなく, 複数の異なる遺伝子に対して作用し得る. 1 つの miRNA が標的とする遺伝子数は平均 200 個と推定されている [78]. 逆に 1 つの mRNA は複数の miRNA による影響を受ける. miRNA のデータベースである miRBase には Release 20 の時点で 30,424 種類の miRNA が登録されており, そのうち 2,555 種類がヒトの miRNA である [79, 80, 81, 82]. miRNA-遺伝子間相互作用を検証するために全ての組合せについて生物学的実験を行うのは時間的にもコスト的にも現実的でないため, 計算学的手法を用いた予測が用いられている.

3.3 miRNA-遺伝子間相互作用予測

本節ではまず従来の miRNA-遺伝子間相互作用予測手法に関して述べる. その後, 従来の予測手法の問題点について述べる.

3.3.1 従来の miRNA-遺伝子間相互作用予測手法

miRNA のターゲット推定にはシードペアリングが用いられる [83, 84]. これは, miRNA のシード配列の相補性を用いて相互作用の有無を予測するものである. 前述

の通り, miRNA がある mRNA に対して作用するためには, miRNA のシード配列は mRNA の結合部位に対して完全に相補的でなければならない (図 3.2). ループ部分を挟んで 13 塩基目以降は基本的には相補的配列である必要があるが, 塩基置換が起こっていても相互作用を持ち得る. そこで mRNA と miRNA で局所アライメントを行うことにより相互作用を予測する. このときシード配列は完全一致とし, その他の領域ではある程度のミスマッチを許容する. クエリとする mRNA 配列は, 遺伝子の 3'UTR に限る手法もあれば, 5'UTR やイントロンなども含める手法もある.

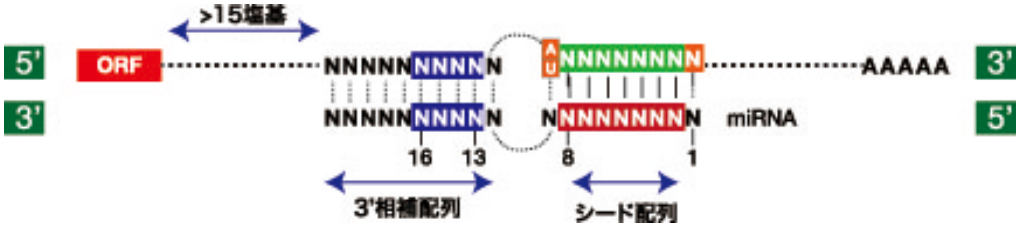

図 3.2　miRNA と mRNA の配列相補性

3.3.2　従来の予測手法の問題点

従来手法で miRNA-遺伝子間相互作用を予測したデータベースとして microRNA.org[85] および mirDIP[86] がある. 図 3.3 にそれらの推定相互作用と生物学的に検証された相互作用 (mirTarBase[87], miRWalk[88], miRecords[89] の和集合. 図中の "Experimentally validated") を示す.

これらのデータベースに含まれる配列相補性に基づく推定 miRNA-遺伝子間相互作用は false-positive を大量に含んでいる. これは相互作用の予測を数塩基程度のシードマッチングに依存しているためであると考えられる.

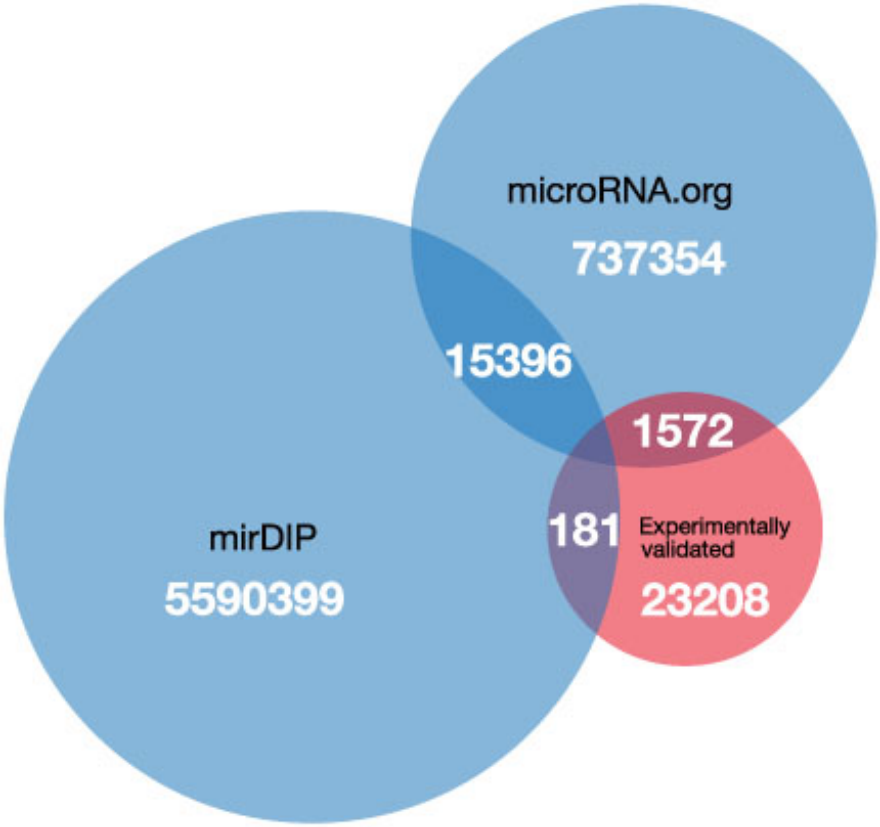

図 3.3　データベース内の miRNA-遺伝子間相互作用

3.4　提案手法

本節では, 提案手法のアイデアについて述べた後, 正準相関分析について説明し, 正準相関分析を用いた miRNA-遺伝子間相互作用予測手法について述べる.

3.4.1 提案手法のアイデア

miRNA と相互作用予測においては数塩基程度のシードマッチングに依存していることが問題であると述べた. そこで, 情報量を増やすために発現量情報を用いることが考えられる [90, 91]. Wang *et al.* 2009 [92] では miRNA とターゲット遺伝子の発現量の相関について調査している. 図 3.4 は, miRNA 303 個と遺伝子 2,217 個からなる 17,777 組の miRNA-遺伝子の組について発現量の相関係数の分布を図にしたものである. 横軸が相関係数, 縦軸が密度を表す.

図 3.4 では相関係数が -0.5 および 0.5 の付近にピークが見られ, それぞれの miRNA と遺伝子の発現量の相関は低い. これは, miRNA と遺伝子の相互作用が 1 対 1 ではなく多対多の関係であるためであると考えられる. 1 つの miRNA がターゲットとする遺伝子は約 200 個と推定されており, それぞれの miRNA は複数のターゲット遺伝子を持つ [78]. 一方, 各遺伝子も複数の miRNA と相互作用を持つ [73].

このように miRNA-遺伝子間相互作用は多対多の関係であると考えられ, 単純に miRNA と遺伝子の発現量の相関係数を見るだけでは不十分であり, miRNA-遺伝子間相互作用を説明するモデルの構築が必要であると考えられる.

3.4.2 正準相関分析

正準相関分析 (canonical correlation analysis; CCA) とは, 2 群の変量間の関係を調べるための統計解析手法である. 正準相関分析は重回帰分析の一般形である. 重回帰分析では 1 個の従属変数と複数個の独立変数の線形合成変数の相関が最大となるような独立変数の重みを求めていく. それに対して正準相関分析は, 従属変数, 独立変数という区別ではなく, それぞれ複数の変数からなる変数群それぞれについて線形合成変数を求め, 2 つの合成変数の相関 (正準相関) が最も大きくなるような重み係数を求める. 合成変数は複数個合成変数間の相関が最大のものから順次求めていく.

以下に正準相関分析の手順について説明する. それぞれ p 個, q 個の変数を含む 2 つの

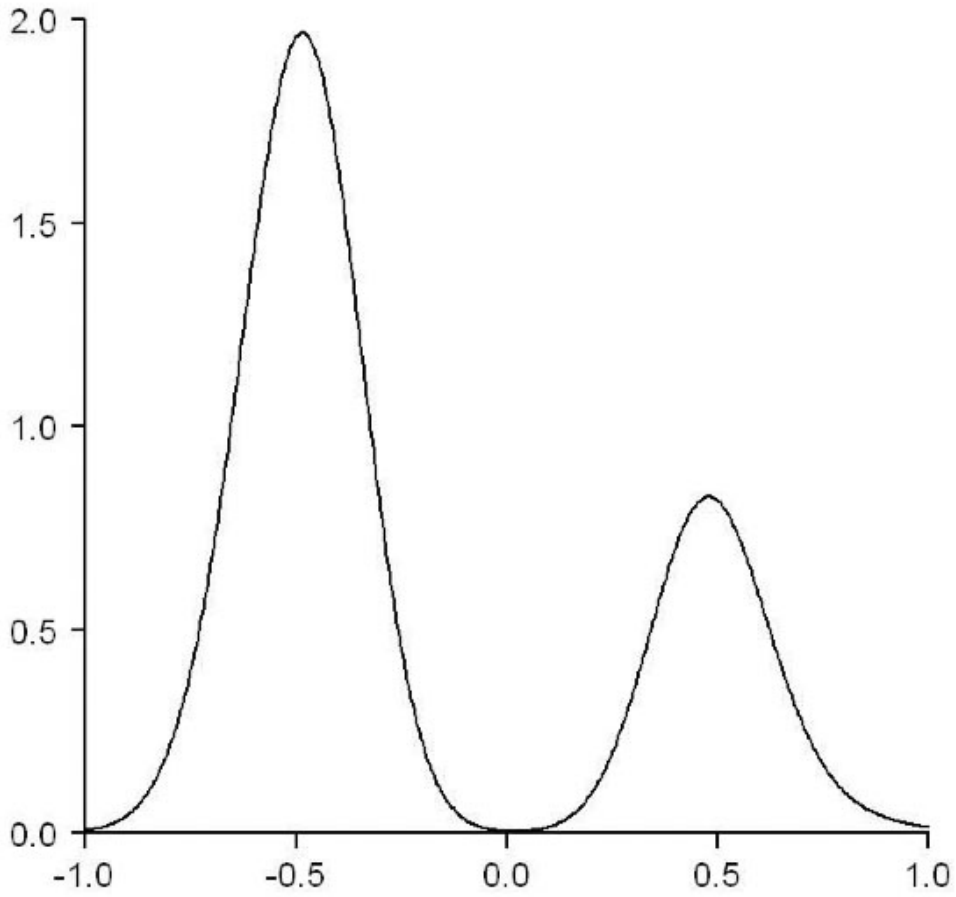

図 3.4　miRNA-遺伝子の発現量相関係数分布

変数群を $(x_1, x_2, \cdots, x_p)$, $(y_1, y_2, \cdots, y_q)$ とする．また, $\min r = (p, q)$ 種類の重み係数を $(a_{i1}, a_{i2}, \cdots, a_{ip})$, $(b_{i1}, b_{i2}, \cdots, b_{iq})$, $(i = 1, 2, \cdots, r)$ とする．この重み係数により新たに作成される合成変数を f_i, g_i とする．

CCA では

$$f_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p g_j = b_{i1}y_1 + b_{i2}y_2 + \cdots + b_{iq}y_q \qquad (3.1)$$

に対して, 合成変数 f_i と g_i の相関が最も高くなるような重み係数をそれぞれ求める．こ

のような重み係数は r 個存在する．f_i と g_i は, 第 i 正準変量という．また, f_i と g_i の間の相関係数 ρ_i は第 i 正準相関係数と呼ばれ相関係数の高い順に $\rho_1 \leq \rho_2 \leq \cdots \rho_r$ となる．

このように正準相関分析は 1 対 1 や 1 対多の変数間の関係のみに注目するのではなく, 多対多の関係を持つ変量間の相関を最もよく説明するモデルを構築する．前述のように miRNA-遺伝子間相互作用も多対多の関係であり, 正準相関分析を用いることにより予測精度の向上が期待される．

3.4.3 CCA を用いた相互作用予測

ここでは, RNA-Seq データと miRNA-Seq データを用いる． miRNA-Seq とは, miRNA をシーケンスするためのプロトコルであり, トータル RNA 分離の後に．ゲル電気泳動をかけることにより小さな RNA 分子のみを選択的に抽出しシーケンスを行う．このようにしてシーケンスした miRNA-Seq データを複数集め, miRNA プロファイルを作成する [93, 94, 95]．得られたリードを参照ゲノム配列に対してマッピングし, 既知の miRNA に対してヒットしたリード数を数え上げる．同様に遺伝子に対してもリード数の数え上げを行う．

そのようにして得られた miRNA および遺伝子のプロファイルを CCA にかけ, 得られた平面上で距離が近い miRNA と遺伝子の組合せを推定相互作用として選び (図 3.5)．データベースとの積集合を出力する．

3.5 実験

提案手法の有用性を検証するために, 実データを用いた実験を行った．

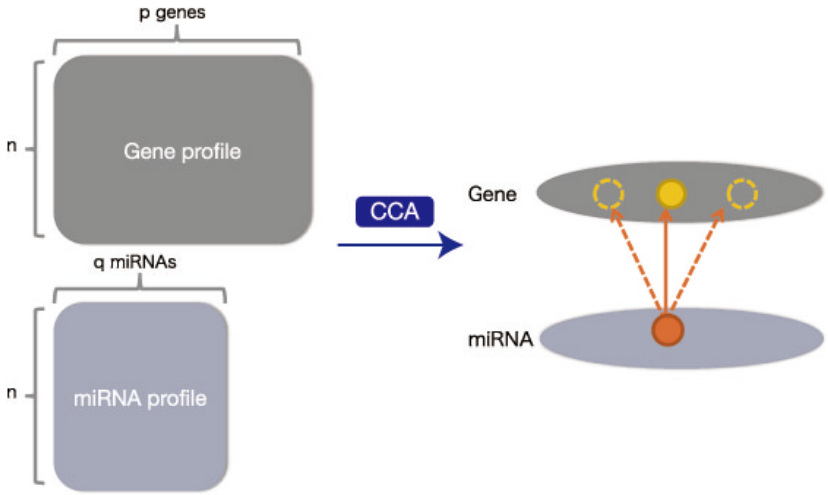

図 3.5　CCA を用いた miRNA-遺伝子間相互作用推定モデルの構築

3.5.1　データ

本実験では GEUVADIS プロジェクトのデータを用いる [96]. これは 449 サンプルのヒトリンパ細胞株から作成したライブラリに対して RNA-Seq 及び miRNA-Seq を行ったデータである.

miRNA-遺伝子間推定相互作用データベースとして microRNA.org[85] 及び mirDIP[86] を使用し, 正解データとして生物学的実験により検証された相互作用を格納する mirTarBase[87], miRWalk[88], miRecords[89] を用いた.

3.5.2　結果

提案手法を適用した結果, 図 3.6 のような結果が得られ, false-positive は microRNA.org については 737354 個から 2626 個へ, mirDIP については 5590399 個から 1619 個へと大幅に減らすことができた.

次に提案手法による true-positive (TP), false-positive (FP), false-negative (FN) の変化を調べたところ, 表 3.1 のようになった.

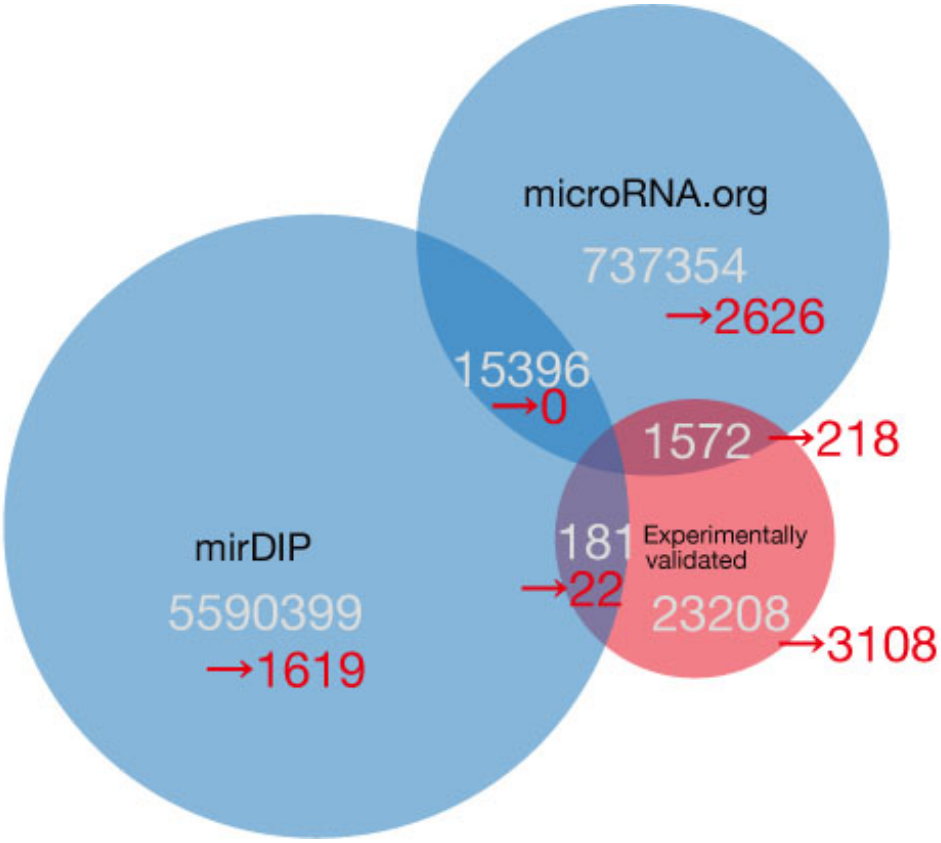

図 3.6　提案手法の結果

表 3.1　提案手法による予測結果分類

	TP	FP	FN
mirDIP	181	5590399	23027
mirDIP + 提案手法	22	1597	3086
microRNA.org	1572	735782	21636
microRNA.org + 提案手法	218	2407	2890

これらの結果から Precision/Recall 及び F-measure を計算すると表 3.2 のようになった.

相互作用の総予測数に対する正解予測数は, mirDIP では 5590399 組中 181 組

表 3.2　Precision/Recall と F-measure

	Precision (%)	Recall (%)	F-measure (%)
mirDIP	0.00324	0.780	0.00645
mirDIP + 提案手法	1.36	0.707	0.411
microRNA.org	0.213	1.36	2.28
microRNA.org + 提案手法	8.30	0.701	7.60
mirDIP + microRNA.org + 提案手法	5.30	13.7	7.64

(0.00324%) から 1619 組中 22 組 (1.36%) へ, microRNA.org では 737354 組中 1572 組 (0.213%) から 2625 組中 218 組 (8.30%) へとそれぞれ改善し，false-positive を大きく削減することで正解の予測率が向上した.

3.5.3　考察

microRNA.org における配列相補性に基づく miRNA-遺伝子間相互作用は, 737354 組であったが, そのうち真の相互作用は 1572 組に留まり, その割合は 0.213% と, 約 500 組に 1 組の真の相互作用を持つ計算になる. それに対し提案手法適用後は, 予測された 2626 組のうち真に相互作用を持つものは 218 組と, その比は 8.30% と約 12 組に 1 組の精度まで改善された. 真の相互作用を 1 組発見するまでの実験回数の期待値は, 提案手法適用前は 250 回であったのに対し, 提案手法適用後は 6 回となる. qPCR による生物学的実験を用いた miRNA-遺伝子間相互作用の検証には 1 組の相互作用当たり 24 〜 48 時間必要であるため [97], 従来は平均して 250 〜 500 日程度と膨大な時間がかかり, それに伴うコストも莫大なものになるため生物学的実験による検証は非現実的であった. 一方, 提案手法適用後は 1 組の真の相互作用を発見するまでに 6 〜 12 日と, 生物学的実験による検証が現実的なものとなる数字にまで予測精度が改善した.

3.6 結言

本研究では正準相関分析を用いた miRNA-遺伝子間相互作用予測手法を提案した. 従来の配列相補性に基づく標的遺伝子探索では大量の false-positive の相互作用が推定されていた. その問題に対し, mRNA と miRNA の発現プロファイルも併せて入力として与え, 最も尤もらしい相互作用を推定しフィルタとして用いることにより配列相補性に基づく miRNA-遺伝子間相互作用データベースの false-positive をふるい落とすことができた. これにより miRNA-遺伝子間相互作用をより正確に予測することができ, 創薬などにおいて有用であると期待される.

第 4 章　結論

本研究では, RNA-Seq データを用いてアイソフォームを推定する手法と, miRNA-遺伝子間相互作用予測手法を提案した.

第 2 章では, 反復マッピングを行いアイソフォームの推定精度を改良する手法を提案した. RNA-Seq データには大量のノイズが含まれる. そのようなデータからアイソフォームを正確に推定するためには, ノイズの影響をなるべく軽減しなければならない. しかし従来手法では, 複数の箇所にマップされるマルチリードの影響によりアイソフォームがうまく推定できないという問題があった. このような問題に対し, 反復マッピングを用いたアイソフォーム推定手法を提案した. マルチリードの由来箇所はゲノムワイドな発現量情報が必要であるのに一方, マッピングを正確に行わないと発現量情報が正しく得られない. そこで発現量推定とマルチリードを振り分け直す再マッピングを反復的に行うことによりノイズの影響を軽減し, アイソフォームを推定する手法を提案した. 実データを用いて検証したところ, 既知のアイソフォームを従来手法よりも高精度に推定できていることが確認された. このことから, 提案手法を用いることによりアイソフォームを高精度に推定することができると言える.

第 3 章では, miRNA-遺伝子間相互作用を予測する手法を提案した. miRNA は予想されていたよりも大量に存在しており, miRNA と遺伝子の組合せ数は膨大なものになる. そのような状況下で, 配列相補性に基づく相互作用予測のみでは大量の false-positive が誤って推定されてしまっていた. このような問題に対し, miRNA-Seq および RNA-Seq データから得られた発現プロファイルを用い, 正準相関分析を行うことにより推定相互作用をフィルタリングする手法を提案した. 提案手法では, 発現量の変化を最もよく説明する miRNA-遺伝子間相互作用を選択する. これにより, Recall を悪化させることなく Precision を改善することができた.

ヒトの生体内には約 10 万種類のタンパク質が存在すると言われているが, 遺伝子数

は約 2 万個しかないとされている．アイソフォームはこの差を埋める役割を果たしていると考えられており，ヒトに存在するアイソフォームをより正確に推定する手法の開発は，遺伝子の機能とそれの働きとして現れる種々の生命現象の解明に貢献できると考えられる．

また, ヒトの生体内で転写される RNA にはノンコーディング RNA が多く含まれているが，mRNA と比べてその機能には未解明の部分が多い．ノンコーディング RNA の中でも，miRNA による遺伝子の転写制御が注目を集めている．本研究で提案した miRNA-遺伝子間相互作用予測手法は，miRNA が制御するターゲット遺伝子の予測精度の向上を達成しており，これにより miRNA が遺伝子に及ぼす影響をより正確に予測できることから，miRNA が関与する疾患や生命現象の解明に貢献するものと考えられる．

ヒトをはじめとした様々な生物についてゲノム解読が完了したポストゲノム時代と呼ばれる現在, 疾患や生命現象のメカニズムを解明するためにトランスクリプトームの重要性が高まっている. そのような中で, 新規のものを含めた転写産物をゲノムワイドにアイソフォームレベルで調べることのできる RNA-Seq データを用いて転写解析を行うことの重要性もますます高まっていくと考えられる. RNA-Seq データを用いた, 転写産物やその相互作用のより正確な推定・予測を可能にすることで, 種々の生命現象や疾患についての知見がより効率的に得られるようになり, 生物学のみならず医学・薬学にも貢献できると考えられる.

謝辞

本論文は, 著者が 2009 年から 2011 年まで大阪大学大学院情報科学研究科博士前期課程在学中, 2011 年から現在まで大阪大学大学院情報科学研究科博士後期課程在学中に行ってきた, 生物情報科学に関する研究成果をまとめたものです.

本研究の全過程の遂行ならびに本論文をまとめるにあたり, 懇切なるご指導, ご鞭撻を賜わりました大阪大学大学院情報科学研究科バイオ情報工学専攻 松田秀雄 教授には, ここに厚く御礼申し上げます. 貴重なお時間を割いていただき丁寧なるご教示を賜りました大阪大学大学院情報科学研究科バイオ情報工学専攻 若宮直紀 教授, 清水浩 教授, 前田太郎 教授, 四方哲也 教授に厚く御礼申し上げます.

また本論文をまとめるにあたり, 研究の進め方に対する様々なご指導, ご助言を頂きました大阪大学大学院情報科学研究科 竹中要一 准教授には, 厚く御礼申し上げます.

大阪大学情報科学研究科 大安裕美 特任研究員には, 本研究の全過程において数多くの御指導, 御助言を頂きました. 心より御礼申し上げます.

本論文をまとめるにあたり, 様々な御支援, 御指導を頂きました大阪大学大学院情報科学研究科バイオ情報工学専攻 瀬尾茂人 助教には, 厚く御礼申し上げます.

大阪大学サイバーメディアセンター 木戸善之 特任講師には, 本研究の全過程において数多くの御指導, 御助言を頂きました. 心より御礼申し上げます.

大阪大学大学院情報科学研究科バイオ情報工学専攻松田研究室の皆様には, 本研究をまとめるに際して, 様々な御支援, 御指導いただきましたことを感謝いたします.

最後に, 本研究の遂行に際し, 著者を御激励, 御支援下さいました方々へ心より感謝致します.

参考文献

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, Vol. 409, No. 6822, pp. 860–921, 2001.

[2] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, Vol. 431, No. 7011, pp. 931–945, 2004.

[3] W. G. Feero, A. E. Guttmacher, and F. S. Collins. Genomic Medicine — An Updated Primer. *The New England Journal of Medicine*, Vol. 362, pp. 2001–2011, 2010.

[4] N. A. Faustino and T. A. Cooper. Pre-mRNA splicing and human disease. *Genes & Development*, Vol. 17, pp. 419–437, 2003.

[5] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, Vol. 489, No. 7414, pp. 57–74, 2012.

[6] J. S. Mattick and I. V. Makunin. Non-coding RNA. *Human Molecular Genetics*, Vol. 15, No. Suppl 1, pp. R17–R29, 2006.

[7] T. Phillips. Small Non-coding RNA and Gene Expression. *Nature Education*, Vol. 1, No. 1, p. 115, 2008.

[8] J. H. Yoon and K. Abdelmohsen and M. Gorospe. Posttranscriptional gene regulation by long noncoding RNA. *Journal of Molecular Biology*, Vol. 425, No. 19, pp. 3723–3730, 2013.

[9] J. E. Wilusz and H. Sunwoo and D. L. Spector. Long noncoding RNAs: functional surprises from the RNA world. *Genes Development*, Vol. 23, No. 13, pp. 1494–1504, 2009.

[10] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers.

GenBank. *Nucleic Acids Research*, Vol. 39, No. Database issue, pp. D32–D37, 2011.

[11] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, Vol. 28, No. 1, pp. 235–242, 2000.

[12] H. M. Berman. The Protein Data Bank: a historical perspective. *Acta Crystallographica Section A*, Vol. A64, No. 1, pp. 88–95, 2008.

[13] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. ニュートンプレス, 2010.

[14] M. Burset, I. A. Seledtsov, and V. V. Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, Vol. 28, No. 21, pp. 4364–4375, 2000.

[15] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, Vol. 456, No. 7221, pp. 470–476, 2008.

[16] M. A. Garcia-Blanco, A. P. Baraniak, and E. L. Lasda. Alternative splicing in disease and therapy. *Nature Biotechnology*, Vol. 22, No. 5, pp. 535–546, 2004.

[17] M. D. Adams, J. M. Kelly, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, Vol. 252, No. 5013, pp. 1651–1656, 1991.

[18] E. D. Neto, R. G. Correa, S. Verjovski-Almeida, M. R. S. Briones, M. A. Nagai, J. W. d. Silva, M. A. Zago, S. Bordin, F. F. Costa, G. H. Goldman, A. F. Carvalho, A. Matsukuma, G. S. Baia, D. H. Simpson, A. Brunstein, P. S. L. d. Oliveira, P. Bucher, C. V. Jongeneel, M. J. O'Hare, F. Soares, R. R. Brentani, L. F. L. Reis, S. J. d. Souza, and A. J. G. Simpson. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings of the*

National Academy of Sciences of the United States of America, Vol. 97, No. 7, pp. 3491–3496, 2000.

[19] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial Analysis of Gene Expression. *Science*, Vol. 270, No. 5235, pp. 484–487, 1995.

[20] S. Saha, A. B. Sparks, C. Rago, V. Akmae, C. J. Wang, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu. Using the transcriptome to annotate the genome. *Nature Biotechnology*, Vol. 20, No. 5, pp. 508–512, 2002.

[21] C.-H. Wei, P. Ng, K. P. Chiu, C. H. Wong, C. C. Ang, L. Lipovich, E. T. Liu, and Y. Ruan. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 32, pp. 11701–11706, 2004.

[22] H. Matsumura, S. Reich, A. Ito, H. Saitoh, S. Kamoun, P. Winter, G. Kahl, M. Reuter, D. H. Krüger, and R. Terauchi. Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proceeding of the National Academy of Sciences of the United States of America*, Vol. 100, No. 26, pp. 15718–15723, 2003.

[23] 中村真理. CAGE シークエンスベースの発現解析. ゲノムネットワーク, Vol. 49, No. 17, pp. 2688–2693, 2004.

[24] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sakai, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Science of the United States of America*, Vol. 100, No. 26, pp. 15776–15781, 2003.

[25] R. K. Saiki, S. Scarf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, Vol. 230, No. 4732, pp. 1350–1354, 1985.

[26] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. PPrimer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, Vol. 239, No. 4938, pp. 487–491, 1988.

[27] J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Science of the United States of America*, Vol. 74, No. 12, pp. 5350–5354, 1977.

[28] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, Vol. 94, pp. 441–448, 1975.

[29] M. Baker. Next-generation sequencing: adjusting to data overload. *Nature Methods*, Vol. 7, No. 7, pp. 495–499, 2010.

[30] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, Vol. 26, No. 10, pp. 1135–1145, 2008.

[31] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church. Advanced Sequencing Technologies: Methods and Goals. *Nature Reviews Genetics*, Vol. 5, No. 5, pp. 335–344, 2004.

[32] M. L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, Vol. 11, No. 1, pp. 31–46, 2009.

[33] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, Vol. 55, No. 4,

pp. 641–658, 2009.

[34] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. DiMeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss, and Z. Xie. Single-Molecule DNA Sequencing of a Viral Genome. *Science*, Vol. 320, No. 5872, pp. 106–109, 2008.

[35] O. Morozova and M. A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, Vol. 92, No. 5, pp. 255–264, 2008.

[36] 菅野純夫. 超高速シークエンス時代の遺伝子転写研究. 実験医学, Vol. 27, No. 1, pp. 8–13, 2009.

[37] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, Vol. 10, pp. 57–63, 2009.

[38] B. J. Haas and M. C. Zody. Advancing RNA-Seq analysis. *Nature Biotechnology*, Vol. 28, No. 5, pp. 421–423, 2010.

[39] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, Vol. 215, No. 3, pp. 403–410, 1990.

[40] W. J. Kent. BLAT — The BLAST-Like Alignment Tool. *Genome Research*, Vol. 12, No. 4, pp. 656–664, 2002.

[41] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, Vol. 18, No. 11, pp. 1851–1858, 2008.

[42] H. Jiang and W. H. Wong. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, Vol. 24, No. 20, pp. 2395–2396, 2008.

[43] H. Li and R. Durbin. Fast and Accurate Short Read Alignment with Burrows-

Wheeler Transform. *Bioinformatics*, Vol. 25, No. 14, pp. 1754–1760, 2009.

[44] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, Vol. 10, p. R25, 2009.

[45] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, Vol. 25, No. 14, pp. 1754–1760, 2009.

[46] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, Vol. 24, No. 5, pp. 713–714, 2008.

[47] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, Vol. 25, No. 15, pp. 1966–1967, 2009.

[48] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. *Technical Report*, Vol. 124, pp. 1–24, 1994.

[49] A. Polanski and M. Kimmel. バイオインフォマティクス. シュプリンガー・ジャパン, 2010. 後藤修 訳.

[50] P. Ferragina and G. Manzini. Opportunistic data structures with applications. *41st Symposium on Foundations of Computer Science*, pp. 398–398, 2000.

[51] R. Lippert. Space-efficient whole genome comparisons with Burrows-Wheeler transforms. *Journal of Computational Biology*, Vol. 12, No. 4, pp. 407–415, 2005.

[52] J. Healy, E. E. Thomas, J. T. Schwartz, and M. Wigler. Annotating large genomes with exact word mathes. *Genome Research*, Vol. 13, pp. 2306–2315, 2003.

[53] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, Vol. 25, No. 9, pp. 1105–1111, 2009.

[54] K. F. Au, H. Jiang, L. Lin, Y. Xing, and W. H. Wong. Detection of splice junc-

tions from paired-end RNA-Seq data by SpliceMap. *Nucleic Acids Research*, Vol. 38, No. 14, pp. 4570–4578, 2010.

[55] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. Macleod, D. Y. Chiang, J. F. Prins, and J. Liu. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, Vol. 38, No. 18, p. e178, 2010.

[56] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. v. Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, Vol. 28, No. 5, pp. 511–515, 2010.

[57] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, Vol. 5, No. 7, pp. 621–628, 2008.

[58] Y. Zhao, Q. Li, C. Yao, Z. Wang, Y. Zhou, Y. Wang, L. Liu, Y. Wang, L. Wang, and Z. Qiao. Characterization and quantification of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis of gene expression. *Human Reproduction*, Vol. 21, No. 6, pp. 1583–1590, 2006.

[59] H. Richard, M. H. Schulz, M. Sultan, A. Nürnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S. A. Haas, and M.-L. Yaspo. Prediction of alternative isoforms from expression levels in RNA-Seq experiments. *Nucleic Acids Research*, Vol. 38, No. 10, p. e112, 2010.

[60] 阿久津達也. バイオインフォマティクスの数理とアルゴリズム. 共立出版, 2007.

[61] A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 1998.

[62] S. Audic and J. M. Claverie. The significance of digital gene expression profiles. *Genome Research*, Vol. 7, No. 10, pp. 986–995, 1997.

[63] T. Beissbarth, L. Hyde, G. Smyth, C. Job, W. M. Boon, S. S. Tan, H. S. Scott, and T. P. Speed. Stastical modeling of sequencing errors in SAGE libraries. *Bioinformatics*, Vol. 20, No. Suppl. 1, pp. i31–i39, 2004.

[64] C. K and R. N. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, Vol. 8, No. 2, pp. 93–103, 2007.

[65] H. S. Soifer, J. J. Rossi, and P. Saetrom. MicroRNAs in Disease and Potential Therapeutic Applications. *Molecular Therapy*, Vol. 15, No. 12, pp. 2070–2079, 2007.

[66] W. Wu. MicroRNA: potential targets for the development of novel drugs? *Drugs in R&D*, Vol. 10, No. 1, p. 1, 2010.

[67] M. Lindow and S. Kauppinen. Discovering the first microRNA-targeted drug. *Journal of Cell Biology*, Vol. 199, No. 3, pp. 407–412, 2012.

[68] E. v. Rooij, A. L. Purcell, and A. A. Levin. Developing MicroRNA Therapeutics. *Circulation Research*, Vol. 110, No. 3, pp. 496–507, 2012.

[69] B. R. Cullen. Transcription and processing of human microRNA precursors. *Molecular Cell*, Vol. 16, No. 6, pp. 861–865, 2004.

[70] D. Baulcombe. RNA silencing in plants. *Nature*, Vol. 435, No. 7006, pp. 356–363, 2004.

[71] A. Rodriguez, S. Griffiths-Jones, J. Ashurst, and A. Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Research*, Vol. 14, No. 10A, pp. 1902–1910, 2004.

[72] V. N. Kim, J. Han, and M. C. Siomi. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, Vol. 10, No. 2, pp. 126–139, 2009.

[73] 新飯田俊平. 新たな核酸創薬への期待 – マイクロ RNA 研究の最近の動向 –. 科学技術動向, Vol. 124, pp. 24–33, 2011.

[74] Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim.

MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, Vol. 23, No. 20, pp. 4051–4060, 2004.

[75] G. Ruby, C. H. Jan, and D. P. Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, Vol. 448, No. 7149, pp. 83–86, 2007.

[76] K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell*, Vol. 130, No. 1, pp. 89–100, 2007.

[77] S. Cheloufi, C. D. Santos, M. M. Chong, and G. J. Hannon. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, Vol. 465, No. 7298, pp. 584–589, 2010.

[78] D. Didiano and O. Hobert. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural & Molecular Biology*, Vol. 13, No. 9, pp. 849–851, 2006.

[79] S. Griffiths-Jones. The microRNA Registry. *Nucleic Acids Research*, Vol. 32, No. Database Issue, pp. D109–D111, 2004.

[80] S. Griffiths-Jones, H. K. Saini, S. v. Dongen, and A. J. Enright. miRBase: tools for microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, Vol. 34, No. Database Issue, pp. D140–D144, 2006.

[81] S. Griffiths-Jones, H. K. Saini, S. v. Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, Vol. 36, No. Database Issue, pp. D154–D158, 2008.

[82] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, Vol. 39, No. Database Issue, pp. D152–D157, 2011.

[83] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosine's, indicates that thousands of human genes are microRNA

targets. *Cell*, Vol. 120, No. 1, pp. 15–20, 2012.

[84] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, Vol. 136, No. 2, pp. 215–233, 2009.

[85] D. Bettel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Research*, Vol. 36, No. Database issue, pp. D149–D153, 2008.

[86] E. A. Shirdel, W. Xie, T. W. Mak, and I. Jurisica. NAViGaTing the micronome – using multiple microRNA prediction databases to identify signaling pathway-associated microRNAs. *PLoS One*, Vol. 6, No. 2, p. e17429, 2011.

[87] S. D. Hsu, F. M. Lin, W. Y. Wu, C. Liang, W. C. Huang, W. L. Chan, W. T. Tsai, G. Z. Chen, C. J. Lee, C. M. Chiu, C. H. Chien, M. C. Wu, C. Y. Huang, A. P. Tsou, and H. D. Huang. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, Vol. 39, No. Database Issue, pp. D163–D169, 2011.

[88] H. Dweep, C. Sticht, P. Pandey, and N. Gretz. miRWalk – database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of Biomedical Information*, Vol. 44, No. 5, pp. 839–847, 2011.

[89] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, Vol. 37, No. Database issue, pp. D163–D169, 2009.

[90] A. Muniategui, R. Nogales-Cadenas, M. Vázquez, X. L. Aranguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano, and A. Rubio. Quantification of miRNA-mRNA Interactions. *PLoS One*, Vol. 7, No. 2, p. e30766, 2012.

[91] A. Muniategui, J. Pey, F. Planes, and A. Rubio. Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics*, Vol. 14, No. 3, pp. 263–278, 2013.

[92] Y.-P. Wang and K.-B. Li. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*, Vol. 10, p. 218, 2009.

[93] C. J. Creighton, J. G. Reid, and P. H. Gunaratne. Expression profiling of microRNAs by deep sequencing. *Briefings in Bioinformatics*, Vol. 10, No. 5, pp. 490–497, 2009.

[94] M. R. Friedläunder, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, Vol. 26, No. 4, pp. 407–415, 2008.

[95] E. Zhu, F. Zhao, G. Xu, H. Hou, L. Zhou, X. Li, Z. Sun, and J. Wu. miRTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Research*, Vol. 38, No. Web Server issue, pp. W392–W397, 2010.

[96] T. Lappalainen, M. Sammeth, M. R. Friedlander, P. A. C. t. Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. v. Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, T. G. Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Hasler, A.-C. Syvanen, G.-J. v. Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, Vol. 501, No. 7468, pp. 506–511, 2013.

[97] D. E. Kuhn, M. M. Martin, D. S. Feldman, A. V. T. Jr., G. J. Nuovo, and T. S. Elton. Experimental validation of miRNA targets. *Methods*, Vol. 44, No. 1, pp. 47–54, 2008.