

学位論文

ノードの行列状配置構造による
文字列間の高速類似度判定回路

金沢大学大学院自然科学研究科
電子情報システム専攻

平成 10 年度入学

藤井 直樹

主査	佐々木 公洋	助教授
副査	畑 朋延	教授
主任指導	畑 朋延	教授

平成 12 年 1 月 31 日提出

目次

第 1 章	序論	7
1-1	本研究の背景	7
1-2	DNA 解析 ^{1),2),3),4)}	7
1-2-1	ゲノム	8
1-2-2	知識情報処理	9
1-2-3	DNA が表す機能の解明	10
1-2-4	コンピュータによる遺伝子解析	11
1-3	本研究の目的	11
第 2 章	遺伝子解析におけるホモロジー検索	13
2-1	遺伝子に起こる突然変異 ³⁾	13
2-1-1	遺伝子に起こる突然変異	13
2-1-2	遺伝子突然変異により塩基配列に起こる変化	14
2-2	表を用いた塩基配列の一致比較	15
2-2-1	完全一致	15
2-2-2	塩基置換における一致比較	16
2-2-3	塩基欠落における一致比較	17
2-2-4	塩基混入における一致比較	18
2-3	ノードの行列状配置構造による類似判定	18
2-3-1	表を用いた塩基配列の類似判定	18
2-3-2	ノードの行列状配置構造による類似判定	20
2-3-3	ノードで行う点数処理	22
2-3-4	ノードの行列状配置構造による並列処理	23
第 3 章	ノードの行列状配置構造による 類似度判定法の評価	25
3-1	シミュレータの基本仕様	25
3-1-1	シミュレータの基本動作	25
3-1-2	パラメータ設定	26
3-2	塩基配列生成方法	31
3-2-1	塩基置換	32

3-2-2	塩基欠落	33
3-2-3	塩基混入	33
3-2-4	塩基変化(3種混合).....	34
3-2-5	連続塩基置換	34
3-2-6	連続塩基欠落	35
3-2-7	連続塩基混入	36
3-3	シミュレーション結果.....	37
3-3-1	各ノードが出力する点数	37
3-3-2	全体に対する変化した塩基数の割合による類似度の傾向.....	40
3-3-3	塩基配列の塩基数の変化による類似度の傾向	48
3-4	考察	50
3-4-1	塩基置換と連続塩基置換	50
3-4-2	塩基欠落と連続塩基欠落の比較	51
3-4-3	塩基混入と連続塩基混入の比較	53
3-4-4	7種類の塩基配列生成方法による傾向の違い.....	54
第4章	高速類似度判定回路の構成.....	56
4-1	各ノードの回路構成.....	56
4-2	回路全体.....	59
4-3	H-SPICEによるシミュレーション.....	60
4-3-1	bufferの特性.....	60
4-3-2	高速類似度判定回路の回路シミュレーション.....	62
4-3-3	bufferを考慮したことによるC言語でのシミュレーション結果の変化	65
4-3-4	遅延時間	67
第5章	結論	70
5-1	本研究のまとめ	70
5-1-1	遺伝子解析におけるホモロジー検索.....	70
5-1-2	ノードの行列状配置構造による類似度判定法の評価	70
5-1-3	高速類似度判定回路の構成	71
5-2	今後の展望	71
	謝辞.....	73
	参考文献.....	75
	口頭発表.....	76

目次

図 2-1	突然変異の種類	14
図 2-2	遺伝子突然変異による塩基配列の変化	15
図 2-3	完全一致	16
図 2-4	塩基置換	17
図 2-5	塩基欠落	17
図 2-6	塩基混入	18
図 2-7	塩基の変化による一致度の点数化	19
図 2-8	6×6 のノードの行列状配置構造	20
図 2-9	ノードの入出力	22
図 2-10	ノードの一致度による組み合わせ	22
図 2-11	ノードの並列処理	24
図 3-1	塩基変化数による類似度の傾向(塩基置換)	26
図 3-2	文字列の欠落数による認識度の違い	27
図 3-3	しきい値によるパラメータの変更	28
図 3-4	ずれが生じたときに点数が引き継がれるノード	30
図 3-5	塩基置換による塩基配列の変化	32
図 3-6	塩基欠落による塩基配列の変化	33
図 3-7	塩基混入による塩基配列の変化	34
図 3-8	連続塩基置換による塩基配列の変化	35
図 3-9	連続塩基欠落による塩基配列の変化	35
図 3-10	連続塩基混入による塩基配列の変化	36
図 3-11	各ノードの出力(塩基置換)	38
図 3-12	各ノードの出力(塩基欠落)	38
図 3-13	各ノードの出力(塩基混入)	39
図 3-14	各ノードの出力(塩基変化(3種混合))	39
図 3-15	変化した塩基数の割合による類似度の傾向(塩基置換)	41
図 3-16	変化した塩基数の割合による類似度の傾向(塩基欠落)	42
図 3-17	変化した塩基数の割合による類似度の傾向(塩基混入)	43

図 3-18	変化した塩基数の割合による類似度の傾向(塩基変化(3種混合))....	44
図 3-19	変化した塩基数の割合による類似度の傾向(連続塩基置換)	45
図 3-20	変化した塩基数の割合による類似度の傾向(連続塩基欠落)	46
図 3-21	変化した塩基数の割合による類似度の傾向(連続塩基混入)	47
図 3-22	塩基配列の塩基数を変化による類似度の傾向	48
図 3-23	任意の箇所での塩基置換における類似度	49
図 3-24	連続した箇所での塩基置換における類似度	50
図 3-25	塩基置換と連続塩基置換の比較	51
図 3-26	塩基欠落と連続塩基欠落の比較	52
図 3-27	塩基混入と連続塩基混入の比較	53
図 3-28	7種の塩基配列生成方法による平均値の傾向	54
図 4-1	ノードのブロック図	56
図 4-2	類似度判定回路ノード部(a)	58
図 4-3	類似度判定回路ノード部(b)	59
図 4-4	ノードの行列状配置構造による類似度判定回路	60
図 4-5	bufferの入出力特性	61
図 4-6	類似度判定回路のシミュレーション結果	62
図 4-7	C言語でのシミュレーション結果	63
図 4-8	回路シミュレーションとC言語でのシミュレーションとの差	63
図 4-9	bufferの特性を考慮に入れたC言語でのシミュレーション	64
図 4-10	回路シミュレーションとの差	64
図 4-11	bufferを考慮することによる類似度の傾向の変化(9塩基)	65
図 4-12	bufferを考慮することによる類似度の傾向の変化(20塩基)	66
図 4-13	対角線上にあるノードの出力電圧(完全一致)	68
図 4-14	対角線上にあるノードの出力電圧(塩基置換)	68
図 4-15	対角線上にあるノードの出力電圧(塩基変化(3種混合))	69

表目次

表 1-1	DNA の塩基配列数	9
表 3-1	パラメータの分類	29
表 3-2	パラメータ (図 3-11、12、13、14)	37
表 3-3	パラメータ(図 3-15、16、17、18、19、20、21)	40
表 3-4	パラメータ(図 3-22).....	48
表 4-1	パラメータ(回路シミュレーション).....	62
表 4-2	パラメータ(図 4-14).....	67

第1章 序論

1-1 本研究の背景

昨今のインターネットブームにより、今までコンピュータと縁のなかった家庭にもインターネット端末として、パソコンが入り込むようになった。このように比較的身近な存在となりつつあるコンピュータであるが、最高性能を持った新機種が数ヶ月後には旧機種となるように、性能向上にはめざましいものがある。この性能向上をもたらす理由としては、「処理速度の高速化」が挙げられる。これは計算する速度を速くするということであり、コンピュータ内部で行う処理は基本的には変わっていないということである。

コンピュータは、扱う全てのデータを電気のオン、オフ、つまり 2 進数として表現し、用いることで、処理を行っている。そのため、コンピュータが行う処理には曖昧さがなく、同じ条件であれば、結果は常に同じものとなる。この特徴は正確さが必要な場合には大変重要なものと言えるが、逆に曖昧な情報を扱うのはあまり得意ではないと考えることもできる。例えば、コンピュータは二つのデータを比較する際に「一致している」「一致していない」のどちらかとして判断するので、人が見ればすぐ「似ている」と分かるような、ほんの一部分が変化している 2 種類のものでも「一致していない」、つまり別のものとして判断してしまう。

この類似判断はソフトウェアを使ってプログラムを組めば、コンピュータで実現することは可能であるが、一致、不一致を判断する場合に比べると格段に時間がかかってしまう。そこで、ハードウェアの機能として類似判断を処理することが出来れば、簡単に、そして高速に似ているものを探し出すことが出来るため、いろいろな場面で利用が期待されている。

1-2 DNA 解析 ^{1),2),3),4)}

近年の医学の発展はめざましく、我が国日本は長寿大国としても知られている。しかし、全ての病気が治療可能なわけではなく、悪性新腫瘍、すなわち「ガン」や、原因

不明とされるアルツハイマー病など、不治の病とされる病気も多数存在している。この我々の健康を脅かす病気のほとんどは、何らかの意味で遺伝子と関係していることが分かっている。全ての生物が持っている固有の遺伝子を解析する細胞が持っている DNA（デオキシリボ核酸）と、それに書き込まれたすべての遺伝情報はゲノムと呼ばれており、ある病気はそれにかかっている人自身のゲノムの変化によって引き起こされる。また、他の生物（例えば微生物）のゲノムが作り出す毒素などが原因となって病気になることもある。いずれにしても、何らかの生物のゲノムが原因となって、私たちヒトの病気が起こることが非常に多い。このゲノムはただ病気と深い関係があるばかりではない。私たちヒトを始めとしたすべての生物は固有のゲノムを持っており、生物の姿・形を作る情報がゲノムの中に書き込まれている。その解明を強力に推進しようとしているのがヒトゲノム計画である。

1-2-1 ゲノム

人間の体はおよそ 5 万～10 万の遺伝子によって作り出されているということが分かっている。遺伝子の数は違うが、他のすべての生物も多くの遺伝子の働きによって身体ができており、それぞれの生物が持つすべての遺伝情報を指して、ゲノムと呼んでいる。

「遺伝子」とか「ゲノム」という言葉は単なる概念ではなく、それを実現している物質がある。それが染色体であり、DNA 分子であり、DNA 塩基配列である。ヒトの体は、およそ 30 兆個の細胞から構成されているが、その細胞一個一個に核があり、その核の中に両親から由来する二組の染色体が存在している。すべての遺伝情報を含んだ染色体は、巨大な DNA 分子とヒストンなど何種類かのタンパク質を主成分とした巨大な構造体からなっており、遺伝情報は DNA 分子の中に書き込まれている。その DNA 分子はアデニン（A）、チミン（T）、グアニン（G）、シトシン（C）という 4 種類の核酸塩基がつながったものであり、ヒトの染色体上の DNA には約 30 億対の塩基がひも状の二重らせん構造で配列されている。この中には、受精して発生する過程から日常生きていく過程で必要な、ほとんどすべての遺伝情報が盛り込まれている。言い換えると、遺伝情報はこれらの 4 種類の文字で書かれた文章といえ、23 対の染色体（22 対の常染色体と X、Y の性染色体）に分けられている。それらの DNA の 30 億の文字からなる文字配列には、10 万ほどの遺伝子に対応する文字配列が含まれており、その文字配列に応じたタンパク質のアミノ酸配列が作られ、それらの約 10 万種類のタンパク質が身体のすべての

働きを担っている。

このように DNA は、われわれの生存に不可欠のものであるが、少しずつ変化しながら親から子に連綿として伝えられてきたものであることから、進化の歴史が刻まれている分子化石と見ることもできる。地球上のすべての生物が三五億年にわたる進化の産物であることは明らかであるが、ではどうすれば DNA で進化がわかるのであろうか。それは現在生きている生物の DNA の塩基配列を知り、比べることによって可能となる。

たとえば、ヒトの DNA の塩基配列だけを見ても過去に何が起こったかはわからないが、別の種であるチンパンジーの DNA の塩基配列をもってきて比較すると、二つの DNA の間で塩基が違っている場所があり、その違いを二つの種の進化の道筋をたどる手がかりとすることができる。この違いは霊長類の進化過程で、どちらかの種で突然変異によって塩基の置き換わり（塩基置換）が生じ、それが種の内部で定着したものである。すなわち DNA には、進化の結果が爪痕として残されているのである。

1-2-2 知識情報処理

コンピュータによる遺伝子解析が進めば進む程、扱うデータの量が膨大なものになるという、非常に大きな問題があることが分かっていた。コンピュータの威力は大変なものではあるが、ヒトゲノム計画の場合はデータは非常に膨大で複雑である。まず、配列

	DNA 塩基配列数	遺伝子数
ヒト	30 億	5~10 万
マウス	30 億	5~10 万
イネ	4.5 億	4 万
ショウジョウバエ	2 億	2 万
シロイヌナズナ	1.3 億	2~2.5 万
線虫	1 億	1.5 万
酵母	1340 万	6000
大腸菌	470 万	4500

表 1-1 DNA の塩基配列数

のデータだけを考えてみても、二つの配列が完全に一致することはなく類似度が問題となり、どのくらい似ているか、またどこが似ているかということなどを定量的に解析することは、コンピュータでも非常に時間がかかる。しかし、類似度の解析からは、生物種同士の進化的な近さを評価することができ、たんぱく質の機能の推定も可能となる。それだけではなく、データベースの種類は一つではなく、遺伝子の地図、DNA の文字配列、アミノ酸配列、タンパク質の立体構造、遺伝子と遺伝子の関係、遺伝病、これらに関係する文献など様々なデータベースがあり、しかもデータがそれぞれお互いに非常に関係が深いと、単に一つのデータ

ベースの中だけで検索するのではなく、データベースからデータベースへわたり歩いて総合的なデータを収集することが求められる。このように、複雑な検索ができるような統合化されたデータベースが、ゲノム研究の大きな流れとなっている。

しかし、ゲノム研究における知識情報処理は、データを蓄え、検索するということにとどまらない。ゲノム研究の情報処理には、まったく質的に違う側面があり、一つは、生物のシステムがあまりにも複雑だということ、もう一つはシステムを構成する機能素子（すなわちタンパク質）の機能のメカニズムをまだ理解出来ていないということである。ある生物種のゲノムを完全に解読したとき、数千から数万の遺伝子が提示されるが、その何割かは全く機能の分からないたんぱく質に対応している。現在は、まだそのようなデータに対する情報処理が十分確立していない。

1-2-3 DNA が表す機能の解明

大きなゲノムの解析は、ゲノムの地図作りとそれに基づく配列解析の二段階で行われている。現在の技術で、一度に配列の解読ができるのは、高々DNAの数100文字分程度の断片である。そこで、現実の巨大なDNAを小さく切断し、それらの断片の解読結果から再構成することになるが、これはあたかも数10年分の新聞の切り抜きにしたあと、もとの新聞を完全に再構成するような問題に相当する。すぐに分かるように、これは非常に難しい問題である。しかし、もしある切り抜きのどこかに「明日は東京オリンピック開幕……」というフレーズが含まれていたとすれば、その切り抜きがどの日の新聞の中にあっただかということはすぐに調べることが出来る。それと同じように、30億のDNA文字配列の中に、他にはないユニークなフレーズがたくさんあれば、解析が非常に楽になるであろう。実際、ヒトゲノム計画でも、まずできるだけ多くのユニークな配列を染色体の中に見つけることが計画され、この段階は「ゲノムの地図作り」と呼ばれている。最近ではSTSと呼ばれるDNA中のユニークなフレーズが1万個以上個見つけられており、十分細かいDNAの地図が提供されている。

地図ができると、今度はDNAを小さな断片に切断して配列を解析していくことになる。しかし、ヒトゲノムはあまりにも大きいため、何段階かに切断しそれぞれの断片を増幅して解析していく。100万文字くらいの大きなDNA断片に使われるベクター（一種の入れ物）としてYAC（酵母人工染色体）が用いられ、より小

小さな断片ではコスミド、プラスミドなどのベクターが用いられる。最後に数 100 文字くらいの小さな DNA 断片になって始めて、完全な DNA の文字配列が解読されることになる。これらの解析には、ゲル電気泳動や、PCR などの技術が多用されるが、大量配列解析のための装置は、今では大半が自動化されている。

このようにして構造解析で地図とシーケンス（文字配列）が得られると、次にそこに書かれた生物学的な意味を解明する機能解析の段階に進む。モデル生物を用いたり、遺伝子の破壊実験を行ったりして、病気、発生、分化、免疫応答と遺伝子の関連を明らかにするような研究が盛んに行われている。

1-2-4 コンピュータによる遺伝子解析

現在の DNA 解析能力では、ヒトゲノムの文字配列にそのまま取り組んで成果を上げるには、ことさら大きな組織や多額の設備投資が必要である。近年、DNA の解析能力が画期的に上がったと言っても、それはまだ 30 億文字を処理するには低すぎる。生物ゲノムの構造解析技術が急速な進歩をとげ、DNA 塩基配列およびそれに付随したデータが大量に蓄積されつつあり、この傾向は今後益々増大するものと思われる。また自動シーケンサーの進歩にともない、大量の素配列データが短時間に蓄積されるようになったため、その後の処理に大変な時間と労力がかかる。これからますます DNA の解析が活用される時代が来ると考えると、DNA の構造解析、及び機能解析を進行させていくうちに、技術開発を計って、いまの数百倍に解析能力をあげる努力をしておく必要がある。当面の国際協力は、アメリカの初代責任者だったワトソンが言ったような文字配列分析の作業の分担よりも、技術開発の努力の持ち寄りのほうが、成果が上がると思われる。

しかし 10 億規模の文字配列を解析するにふさわしい方法論と、大量の試料を析できる技術の開発なしではゲノム解析は次のステップに進めない。更に言えば、技術開発があればこそ、ヒトゲノム解析がすすんだあとに来るさらに大規模な DNA 解析時代に対処できるといえる。

1-3 本研究の目的

このように遺伝子解析において DNA の機能解析を行う上で類似判定をより高速に行うことが求められる。従来のソフトウェアによる比較では、逐次処理を行っ

ているため処理時間が非常に長くなってしまふ。それを改善しようと各種の高速化アルゴリズムが提案されているが、根本的な解決にはなっていない。

ハードウェアで高速に検索を行えるものとして CAM (Content Addressable Memory: 連想メモリ)があり、この検索に要する時間は検索するデータの長さに依らず一定であるが、これは完全に一致するデータしか検索することが出来ない。これを利用して、検索したい塩基配列を小さい塩基列に分割し一致するデータの検索を行い、そのデータよりどの程度の塩基列が一致していたかを調べることで類似した塩基配列を探すことも無理ではないが、塩基配列の分割方法、分割する長さなどにより結果が変化するため、一つの塩基配列に対して複数回検索を試みなければ正確な類似性が得られないため、結局、時間がかかることになってしまう。

そこで、本研究では比較処理の並列性を生かしたハードウェアによる高速類似度判定法を提案し、それに基づいた回路を用いることで高速な処理を行うことを目的としている。

最後に本論文の概要を以下に示す。

- 第 2 章では、遺伝子に起こる突然変異について紹介し、突然変異からおこる塩基配列の変化を簡単に調べるための表を用いた塩基配列の一致比較、及びその表を元にしたノードの行列状配置構造について述べる。
- 第 3 章では、ノードの行列状配置構造による類似度判定回路を C 言語によりソフトウェアで構築し、そのシミュレーション結果について述べる。
- 第 4 章では、C 言語のシミュレーション結果により構成した回路、その回路による出力結果について述べる。
- 第 5 章では、本研究で提案するノードの行列状配置構造による文字列間の類似度判定回路をまとめ、今後の展望について述べる。

第2章 遺伝子解析におけるホモロジー検索

2-1 遺伝子に起こる突然変異³⁾

2-1-1 遺伝子に起こる突然変異

生物の遺伝情報を担う染色体上の遺伝子の本体は DNA である。DNA は上に述べたように、自分と同じ分子を正確に、忠実に複製する一方で、体を作る体細胞から生殖細胞が形成されるときには、染色体数は半減するが、受精により両親から受け継いだ遺伝子は確実に受け継がれ、さらに子孫へとその遺伝子は伝えられる。したがって、DNA 上にたくわえられた遺伝情報は、世代を経ても安定に子孫に受け継がれていくものである。

しかし、地球上に生息するすべての生物は、太陽からの紫外線や生物を取り巻く自然環境、または人間が作り出した数多くの有害物質、放射線などにさらされている。このような外的要因の他にも、生物細胞自体の分裂、増殖、DNA 複製の際の誤りが長い年月には少なからず起こる可能性が考えられる。このような内外の要因によって、遺伝情報に生じた変化を突然変異という。突然変異には、染色体に起こる染色体突然変異と、遺伝子に生じる遺伝子突然変異がある。(図 2-1)

生物には人間が手を加えない自然状態でも、低い頻度ではあるが突然変異が起こり、これを自然突然変異、または偶発突然変異という。これに対して人間が放射線や化学物質などを作用させて高率に突然変異を起させることもできる。このように人間が手を加えて誘発させた突然変異を人為突然変異、または誘発突然変異という。突然変異が生物の体を構成する体細胞に生じたものを体細胞突然変異といい、がんの初期段階になったり、種々の奇形や疾病の原因になったりするが、子孫に伝わることはない。しかし、これが生殖細胞に起こると生殖細胞突然変異と呼ばれて、その突然変異は子孫に伝わる。

染色体突然変異	数量的変化	倍数性	コルヒチン、高低温
		異数性(不分離)	加齢、有機水銀
	形態的変化	切断	X線、ナイトロジェンマスタード
		欠失	X線、ナイトロジェンマスタード
		逆位	X線、ナイトロジェンマスタード
		転座	X線、ナイトロジェンマスタード
重複	X線、ナイトロジェンマスタード		
遺伝子突然変異	塩基の置換	5-プロモウラシル、2-アミノプリン、エチルメタンスルホン酸 &	
	削除	プロフラビン、アクリジンオレンジ、ICR-170	
	挿入		
	主鎖切断	エチルメタンスルホン酸、X線	

図 2-1 突然変異の種類

2-1-2 遺伝子突然変異により塩基配列に起こる変化

遺伝子 DNA のレベルで突然変異が起こると、DNA を構成する塩基に種々の変化が生じる。DNA の二重鎖の 1 本の鎖の一つの塩基が、他の塩基に置換した塩基置換が起こると、その DNA が複製することにより、置換された塩基をもとに複製した二重鎖の塩基対がもとの塩基対とは別のものとなり、その後この DNA の複製によって生じた DNA はすべて同じ塩基対が置換されたものになる。このような塩基対の置換によって、その塩基対を含むコドンが変化し、翻訳によって生じたアミノ酸も変化する。このようなアミノ酸の変化によってタンパク質の性質が変化したり酵素の活性が失われたりすることもある。

また、一つの塩基が欠失したり、付加されたりすると、DNA の塩基の数自体が変化し、タンパク質合成の際には変化した塩基以下のすべての三つずつの塩基(コドン)の枠組みが変わる塩基枠変化となり、タンパク質合成の際には mRNA (messenger RNA) のかなり広い範囲にわたってコドンが変化し、翻訳されるアミノ酸が変化して、タンパク質の性質に大きな変化をもたらす。(図 2-2)

このように、遺伝子突然変異による塩基配列の変化は 3 種類あり、本研究では「塩基置換」、「塩基欠落」、「塩基混入」として表現する。

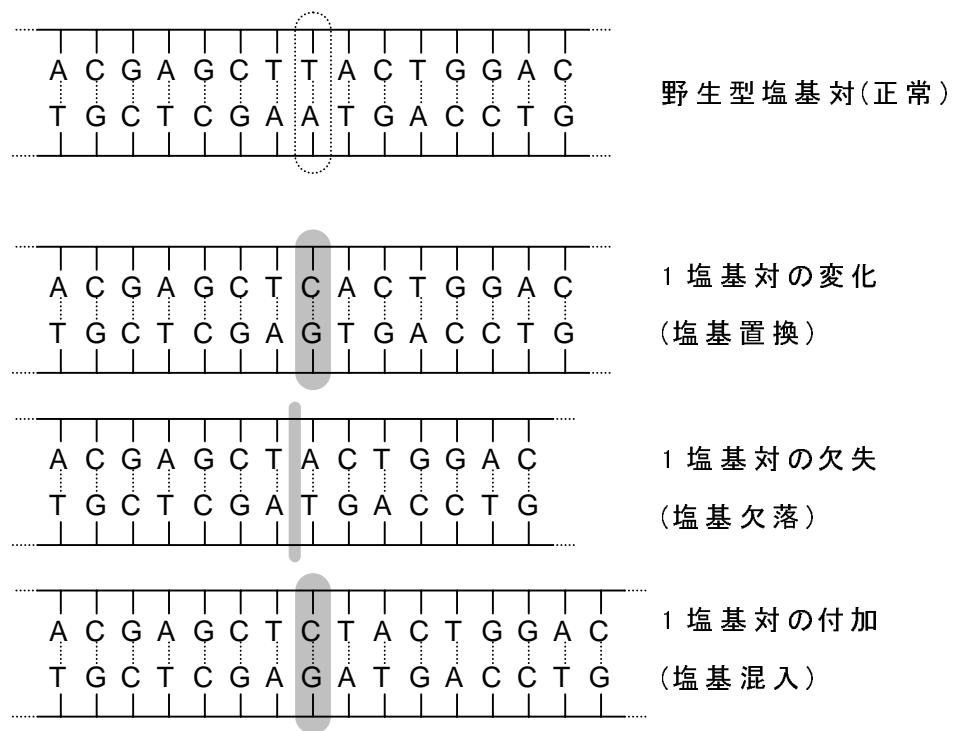


図 2-2 遺伝子突然変異による塩基配列の変化

2-2 表を用いた塩基配列の一致比較

遺伝子突然変異による塩基配列の変化には 3 種類あることを前節で示した。この塩基配列の変化が起きたとき、塩基配列は塩基置換により全体の一部分のみが変化している、もしくは塩基欠落、塩基混入により一部分が欠落、混入してそれ以降の塩基がずれていると考えることが出来る。そのため、元の塩基配列と変化後の塩基配列を人間が視覚により比較した場合、変化が少ない二つの塩基配列はかなり類似しているものと認識することが出来、数 10% の変化であれば少し類似している、元の塩基配列からほとんど変化してしまっているものは、ほとんど類似していないというふうに認識することが出来る。しかし、コンピュータを用いて普通に比較した場合は、「一致」「不一致」で判断し類似を考慮しないため、変化箇所の多少に関わらず、「不一致」という判断しかできない。そこで、次に示すアルゴリズムを用いることによって、ハードウェア的に類似判断できるのではないかと考えた。

2-2-1 完全一致

二つの塩基配列を比較するにあたり、図 2-3 のような表を用いた。比較する塩基配列を、表内の最上行、及び最左列に書いてあり、ここでは、10 塩基で構成された

塩基配列を例として用いている。また、二つの塩基配列を比較し、1つ1つのセルごとに、対応する二つの塩基が一致している場合は「○」、不一致の場合は「×」として表内のセルに示してある。なお、DNA 分子は、決まった塩基の対(A-T と G-C)を作る性質(「ハイブリダイゼーション」と呼ばれる)があり、片方の塩基が何か分かれば、自ずと対となっている塩基が何であるかが分かるため、本論文で示す例では塩基対の片方だけを記述してある。

図 2-3 は、二つの塩基配列が完全に一致する場合の例となっている。図を見ると分かるように、左上から右下に向かった対角線上に一致を示す「○」がならんでいることが分かる。

∖	A	C	G	A	G	C	T	T	A	C
A	○	×	×	○	×	×	×	×	○	×
C	×	○	×	×	×	○	×	×	×	○
G	×	×	○	×	○	×	×	×	×	×
A	○	×	×	○	×	×	×	×	○	×
G	×	×	○	×	○	×	×	×	×	×
C	×	○	×	×	×	○	×	×	×	○
T	×	×	×	×	×	×	○	○	×	×
T	×	×	×	×	×	×	○	○	×	×
A	○	×	×	○	×	×	×	×	○	×
C	×	○	×	×	×	○	×	×	×	○

図 2-3 完全一致

2-2-2 塩基置換における一致比較

次に、左にならんだ塩基配列に対し、上に並んだ塩基配列を一箇所変化させた時、つまり塩基置換の場合の一致比較を図 2-4(a)に示す。この図を見ると、対角線上の塩基置換している箇所が不一致を示す「×」となり、そのほかの対角線上のセルには変化がないことが分かる。さらに、塩基置換が起きた箇所を 3 箇所に増やした場合の例を図 2-4(b)に示す。これを見ると分かるように、対角線上の塩基置換が起きた 3 箇所が「×」に変化している。このように塩基置換している箇所を増やした場合でも、塩基置換した数だけ、塩基置換が起こった箇所に「×」が並ぶことになる。

	A	C	G	A	T	C	T	T	A	C
A	○	×	×	○	×	×	×	×	○	×
C	×	○	×	×	×	○	×	×	×	○
G	×	×	○	×	×	×	×	×	×	×
A	○	×	×	○	×	×	×	×	○	×
G	×	×	○	×	×	×	×	×	×	×
C	×	○	×	×	×	○	×	×	×	○
T	×	×	×	×	○	×	○	○	×	×
T	×	×	×	×	○	×	○	○	×	×
A	○	×	×	○	×	×	×	×	○	×
C	×	○	×	×	×	○	×	×	×	○

(a) 1 箇所の塩基置換

	A	C	G	C	A	C	G	T	T	C
A	○	×	×	×	○	×	×	×	×	×
C	×	○	×	○	×	○	×	×	×	○
G	×	×	○	×	×	×	○	×	×	×
A	○	×	×	×	○	×	×	×	×	×
G	×	×	○	×	×	×	○	×	×	×
C	×	×	×	○	×	○	×	×	×	○
T	×	○	×	×	×	×	○	○	○	×
T	×	○	×	×	×	×	○	○	○	×
A	○	×	×	×	○	×	×	×	×	×
C	×	×	×	○	×	○	×	×	×	○

(b) 3 箇所の塩基置換

図 2-4 塩基置換

2-2-3 塩基欠落における一致比較

次に、左に並んだ塩基配列に対し、上に並んだ塩基配列の一箇所、塩基を欠失させた時、つまり塩基欠落場合の一致比較を図 2-5(a)に示す。なお、塩基欠落により変化した方の塩基配列数が減少するため、最後尾に任意の塩基を付加してある。これを見ると塩基欠落した箇所(5 文字目の”G”)以降の塩基が前にずれるため、一致箇所を示す「○」も塩基欠落した箇所から対角線と平行に一列分ずれていることが分かる。さらに塩基欠落が起きた箇所を 3 箇所に増やした場合の例を図 2-5(b)に示す。これを見ると分かるように、塩基欠落が起きた箇所で「○」の列が対角線と平行にずれていることが分かる。このように塩基欠落している箇所を増やした場合は、「○」が対角線と平行に並ぶ箇所が、塩基欠落した数だけ対角線から離れていくことになる。

	A	C	G	A	C	T	T	A	C	G
A	○	×	×	○	×	×	×	○	×	×
C	×	○	×	×	○	×	×	×	○	×
G	×	×	○	×	×	×	×	×	×	○
A	○	×	×	○	×	×	×	○	×	×
G	×	×	○	×	×	×	×	×	×	○
C	×	○	×	×	○	×	×	×	○	×
T	×	×	×	×	×	○	○	×	×	×
T	×	×	×	×	×	○	○	×	×	×
A	○	×	×	○	×	×	×	○	×	×
C	×	○	×	×	○	×	×	×	○	×

(a) 1 箇所の塩基欠落

	A	G	A	T	T	A	C	G	A	T
A	○	×	○	×	×	○	×	×	○	×
C	×	×	×	×	×	×	○	×	×	×
G	×	○	×	×	×	×	×	○	×	×
A	○	×	○	×	×	○	×	×	○	×
G	×	○	×	×	×	×	×	○	×	×
C	×	×	×	×	×	×	○	×	×	×
T	×	×	×	○	○	×	×	×	×	○
T	×	×	×	○	○	×	×	×	×	○
A	○	×	○	×	×	○	×	×	○	×
C	×	×	×	×	×	×	○	×	×	×

(b) 3 箇所の塩基欠落

図 2-5 塩基欠落

2-2-4 塩基混入における一致比較

最後に、左に並んだ塩基配列に対し、上に並んだ塩基配列の一箇所、塩基を混入させた時、つまり塩基混入の場合の一致比較を図 2-6(a)に示す。なお、塩基混入により塩基配列に混入する塩基は任意とし、また、変化した方の塩基配列数が増加するため、最後尾の塩基を削除してある。これを見ると塩基欠落した時と同様、塩基混入した箇所(5文字目の”G”)以降の塩基が後ろにずれるため、一致箇所を示す「○」も塩基混入した箇所から対角線と平行に一列分ずれていることが分かる。さらに塩基混入が起きた箇所を3箇所に増やした場合の例を図 2-6(b)に示す。こちらの場合も同様に、塩基混入が起きた箇所ですべて「○」の列が対角線と平行にずれていることが分かる。このように塩基混入している箇所を増やした場合も、「○」が対角線と平行に並ぶ箇所が、塩基混入した数だけ対角線から離れていくことになる。

	A	C	G	T	A	G	C	T	T	A
A	○	×	×	×	○	×	×	×	×	○
C	×	○	×	×	×	×	○	×	×	×
G	×	×	○	×	×	○	×	×	×	×
A	○	×	×	×	○	×	×	×	×	○
G	×	×	○	×	×	○	×	×	×	×
C	×	○	×	×	×	×	○	×	×	×
T	×	×	×	○	×	×	×	○	○	×
T	×	×	×	○	×	×	×	○	○	×
A	○	×	×	×	○	×	×	×	×	○
C	×	○	×	×	×	×	○	×	×	×

(a) 1箇所の塩基混入

	A	C	G	T	A	C	T	G	C	T
A	○	×	×	×	○	×	×	×	×	×
C	×	○	×	×	×	○	×	×	○	×
G	×	×	○	×	×	×	×	○	×	×
A	○	×	×	×	○	×	×	×	×	×
G	×	×	○	×	×	×	×	○	×	×
C	×	○	×	×	×	○	×	×	○	×
T	×	×	×	○	×	×	○	×	×	○
T	×	×	×	○	×	×	○	×	×	○
A	○	×	×	×	○	×	×	×	×	×
C	×	○	×	×	×	○	×	×	○	×

(b) 3箇所の塩基混入

図 2-6 塩基混入

2-3 ノードの行列状配置構造による類似判定

2-3-1 表を用いた塩基配列の類似判定

表を用いた塩基配列の一致比較により、完全一致する二つの塩基配列が3種類の塩基変化に対して、変化した箇所が不一致となる、また一致する塩基が元の塩基配列からずれるという特徴があった。この特徴から、

- 一致箇所が続く場合は、一致を示す「○」が対角線と平行に並ぶ
- 対角線と平行に「○」が並ぶ列に入る「×」の数だけ、塩基置換が起きたと

考えられる

- 対角線と平行に「 \times 」が並ぶ列が対角線から離れた距離だけ、塩基欠落、塩基混入が起きたと考えられる

という三つのことが言える。そのため、表から塩基置換、塩基欠落、塩基混入が起きた箇所、回数が分かれば、二つの塩基配列の類似度の度合いが分かるのではないかと考えた。そこで、前節で示した塩基配列の一致比較の表をもとに、最初に与えた点数を塩基置換、塩基欠落、塩基混入が起きた箇所で減点していくことで、二つの類似の度合いを表す点数とする方法を考えた。その例を図 2-7 に示す。

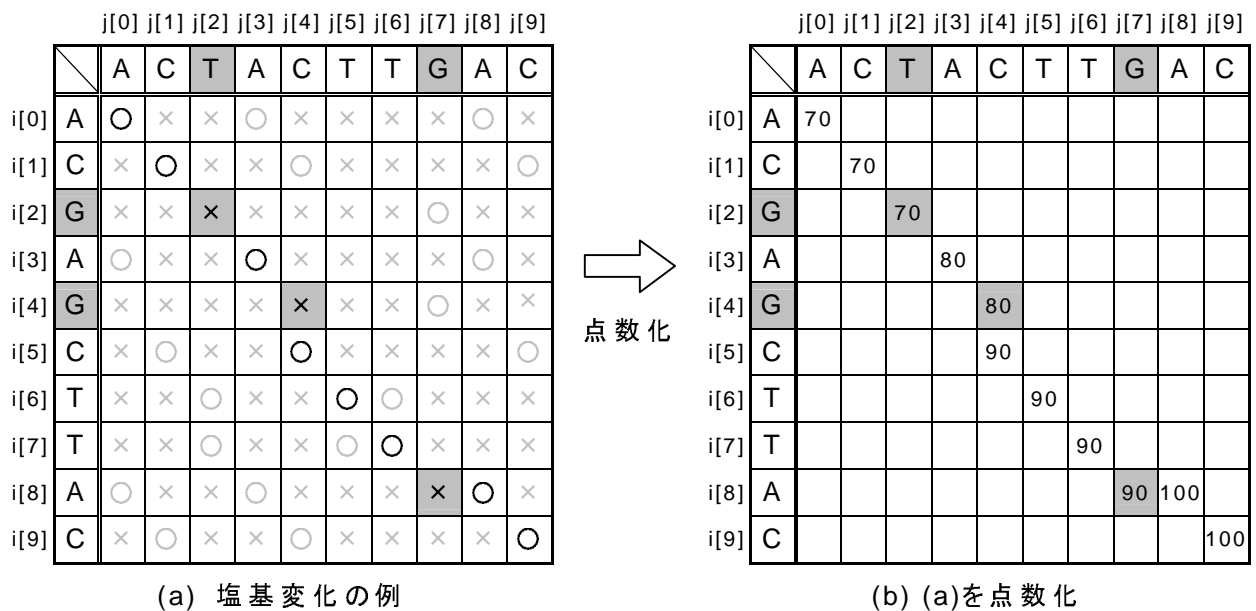


図 2-7 塩基の変化による一致度の点数化

まず、図 2-7(a)は最左列(以降 i 列と表記)の塩基配列に対して、最上行(以降 j 行と表記)の塩基配列が塩基変化した例が示してあり、 $j[2]$ で $i[2]$ が「T」に塩基置換、 $j[3]$ と $j[4]$ の間で $i[4]$ の「G」が塩基欠落、 $j[7]$ で「G」が塩基混入している。そしてこの一致比較の表より、右下のセルに 100 点を与え、塩基変化した箇所で 10 点ずつ減点していったものが図 2-7(b)である。なお、この例ではわかりやすくするため、図 2-7(a)の色の薄い印で示したセルの点数は省略し、濃い印で示したセルにあたる部分の点数のみしか示していない。

最初に与えられた 100 点という点数は、塩基置換、塩基欠落、塩基混入している箇

所でそれぞれ 10 点ずつ減点され、左上のセルは 70 点という点数になっている。この方法では、一致していれば点数をそのまま引き継ぎ、塩基が変化していれば減点した点数を引き継ぐため、それぞれのセルでの点数は、それ以前のセルの一致情報が含まれたものであると考えられ、セルの点数よりそのセルに対応する塩基以降の類似の度合いが推し量られる。そのため、最終的に到達する左上のセルの点数は二つの塩基配列全体の一致情報を含んだ点数といえ、我々はこの点数を「二つの塩基配列の類似度」と定義した。

2-3-2 ノードの行列状配置構造による類似判定

表を用いた塩基配列の一致比較を点数化することによって二つの塩基配列の類似度を定義した。これを回路として実現するにおいて、表のセル1つ1つを、それぞれ一致比較を行いそれに応じて入力点数を減点する回路に置き換え、さらにその回路を表のように行列状に配置し、塩基変化の特徴を生かして接続することにより、配列全体の類似度を求める回路を構成できるのではと考えた。我々はこのセルと置き換える1つ1つの回路をノードと名付けた。このノードの行列状配置構造を図 2-8 に示す。

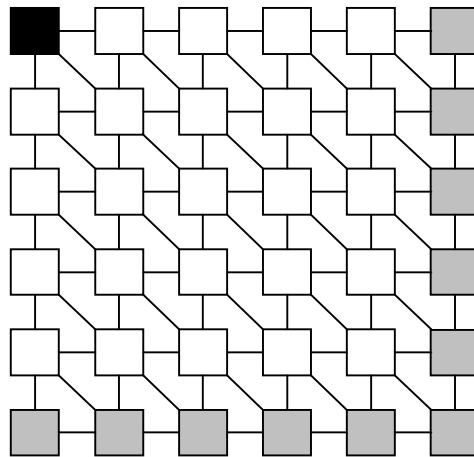


図 2-8 6×6 のノードの行列状配置構造

図 2-7 の例で示した減点法では、「 」が並ぶ部分のみの点数を示し、また最初に点数を与えるセルを右下のセルとしたが、これは塩基配列のどの部分が一致しているかが分かっているためであり、回路に塩基配列のデータを入力して類似判定を行う場合、もちろんこのようなことは分からない。そのため、全てのノードに

において一致情報を元にして適切な減点を行い、最終的に出力される点数が二つの塩基配列の類似の度合いを含んだ点数となるために、この減点法をさらに拡張し、以下のようなアルゴリズムを用いることとした。

- 最右列、最下行のノード(図 2-8 の灰色のノード)に、一致度に応じてある決まった点数を与える。
- それぞれのノードにおいて、以下の規則により下、右下、右方向の 3 方向のノードから点数を受け取り、一番高い点数のものを自身の点数とする。
 - 自身のノードが「」の場合には、右下のノードからの点数はそのまま受け取り、下、右のノードからの点数は減点して受け取る。
 - 自身のノードが「`x`」の場合、下、右下、右のノードからの点数は、減点して受け取る。下、右のノードからの点数は、「」の場合に比べて減点量を増やし、右下のノードからの点数の減点量は、下、右のノードからの点数の減点量よりも少なくすることにより優先順位を与え、対角線方向からの点数を優先させる。
- これを回路全体のノードで行うことにより、最右列、最下行のノード(図 2-8 の灰色のノード)から左上のノード(図 2-8 の黒のノード)に向かって順次点数が決定されていき、左上のノードの点数を最終的に二つの塩基配列の一致度に応じた点数としてこれを類似度とする。

ここで、3 方向からのデータをそれぞれそのまま、もしくは減点した後に受け取り、一番高い点数のものを自身の点数とするとした。これは、前で説明したとおり、入力される点数はそこに到達するまでのノードの一致情報による減点がなされており、高い点数というのはあまり減点されていない、つまり一致しているノードの数が多いたということが言えるため、ほかの 2 方向のノードからの点数より高い点数であったということは、そのノードに到達するルートのほうが、より塩基が一致していたと考えられるからである。

また、下、右のノードからの点数よりも、対角線方向である右下のノードの点数を優先させる理由として、表を用いた塩基配列の一致比較で示したとおり、一致が続く場合には対角線と平行に「」が並ぶことが分かっており、塩基欠落、塩基混入して塩基配列がずれた場合にこの「」の列が対角線に平行にずれる。このことを言い換えると、点数のルートが斜め方向から、上方向、もしくは左方向へと変わ

とも言えるわけで、逆に考えるとこの方向から来る点数は明らかに塩基変化が起きたと考えられるためである。

2-3-3 ノードで行う点数処理

図 2-8 でノードの行列状配置構造を示したが、この中からノードを一つ取り出し、そのノードに点数が入出力されるノードと共に示したものが図 2-9 となる。

この図のように、ノード N には、右下、下、右に位置するノードである A、B₁、B₂ の 3 方向から点数データが入力され、その点数データをノードの一致情報を元にそれぞれ減点し、それを元にノード N の点数が計算され、左上、上、左に位置する C のノードへと出力される。

これからさらに入力部を抜き出して考える。入力は 3 方向から来るわけだが、下、右のノードからの点数に比べて、右下のノードからの点数を優先させるとしたが、下、右のノードからの点数は同じように扱うため、ここではそれぞれ B₁、B₂ として点数計算の点で対等に扱う。そのため、3 つの入力は減点方法により「斜め方向」からの入力と、「縦、横方向」からの入力の 2 つにわけられる。

まず、点数計算において自身の一致度と入力元のノードの一致度の考えられる組み合わせとして、図 2-10 に示すように 8 通りが考えられる。この 8 通りの組み合わせを、2-4-2 で示した減点法のアルゴリズムに基づいて分けると、図 2-10 の「減点量」の欄のように 6 つのパターンにわけられる。

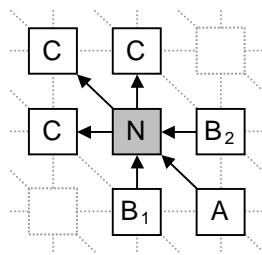


図 2-9 ノードの入出力

入力元	ノード N	減点量
A : ○	○	P-1(減点なし)
A : ×	○	P-1(減点なし)
A : ○	×	P-2
A : ×	×	P-3
B ₁ , B ₂ : ○	○	P-4
B ₁ , B ₂ : ×	○	P-4
B ₁ , B ₂ : ○	×	P-5
B ₁ , B ₂ : ×	×	P-6

図 2-10 ノードの一致度による組み合わせ

実際に減点量には「減点なし」を除いた 5 つのパターンがあるが、減点の仕方としては、ある塩基で不一致となった場合、パターンに応じて決まった点数を引くという、元の点数から減算する方法(以下減算法と表記)と、1 以下の数字を掛けることによって元の点数を減少させるという、元の点数を乗算する方法(以下乗算法と

表記)の 2 種類が考えられる。減算法を用いた場合、元の点数に関わらず減点量は常に一定となる。それに対し、乗算法を用いた場合は元の点数が小さくなるにつれて次第に減点量が小さくなるという傾向となる。逆に言えば高い点数の時は大きく減点されると言えるわけで、この点は類似を考えた上では望ましいとは言えない。しかし、本研究では最終的に高速類似度判定回路を設計することが目的となっているため、仕様を決定する際に回路設計を念頭に置いて考えなければならない。

詳細については「第 4 章 高速類似度判定回路の構成」で述べるが、本研究で設計した回路は電圧の大小で点数を表現するため、減算する際には抵抗を用いることによりノードに入力される電圧を減少させるという構成となっている。これはノードを行列状に配置するという回路構成より、1つ1つのノードをなるべく簡単に小さいものとなるように目指したために、このような回路の仕様とした。この電圧減少の仕方は上記で述べた乗算法と同じであるため、本高速類似度判定法ではこちらの減点法を用いることとした。

2-3-4 ノードの行列状配置構造による並列処理

本研究で提案するノードの行列状配置構造では、比較する塩基配列の塩基数に対して、必要となるノードの数は塩基数の 2 乗個となる。このノードがそれぞれ逐次的に点数処理を行うものであれば、塩基配列の長さが増えるに従い、膨大な処理時間を必要としてしまう。

しかし、このノードの点数処理は、右、右下、下のノードの点数が決定してさえいれば、ほかのノードの点数処理と関係なく行うことが出来る。図 2-11 で例を示す。この図中の灰色で示したノードの点数が確定していれば、黒で示したノードの点数はそれぞれ独立して求めることが出来、これを並列処理と考えることも出来る。そのため、点数処理に必要な時間は、左上から右下の対角線と直行する対角線に平行な方向に並ぶノードの列数で決まり、 n 個からなる二つの塩基配列の点数処理を行う場合のこのノードの列数は $(2n-1)$ 列となり、ほぼ塩基数の 2 倍の列数だとみなせる。

よって、実際にかかる計算時間は、比較する塩基配列の塩基数に比例した時間で済むことになる。

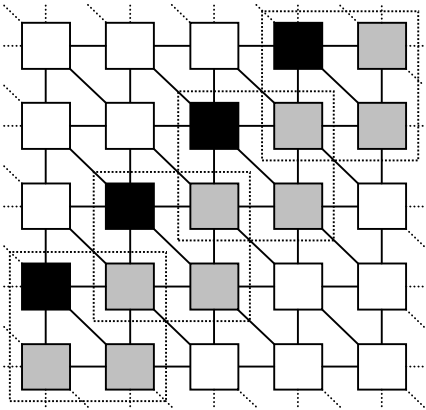


図 2-11 ノードの並列処理

第3章 ノードの行列状配置構造による

類似度判定法の評価

本章では、2章で提案したノードの行列状配置構造による類似度判定回路を、C言語を用いてソフトウェア的にシミュレーションを行ったので、そのシミュレーション方法、結果、及び考察について述べる。

3-1 シミュレータの基本仕様

3-1-1 シミュレータの基本動作

作成した塩基配列類似判定シミュレータは、まず、任意の二つの全く同じ配列を自動生成し、片方の配列の任意の位置で塩基置換、塩基欠落、塩基混入によって塩基を変化させて、類似した配列を生成している。また、塩基変化する箇所数は指定可能で、任意の数の塩基を変化させることが可能となっている。

次に、こうして作成した二つの塩基配列データを元に、最右列、最下行のノードにおいて、それぞれのノードに対応する二つの塩基配列の塩基を比較し、一致していればそのノードの点数を100点とし、一致していなければ図2-10の「ノードN: x、A: x」の時の減点量にあたる「P-4」だけ100点から減点して、そのノードの点数とする。この具体的な減点量については後で述べる。

このようにして初期値を与えた後、右下から左上のノードへと順次それぞれのノードの点数を決定していくが、ソフトウェアでのシミュレータでは「2-4-4 ノードの行列状配置構造による並列処理」で述べたように、並列処理が行えない。そのため、同時に点数処理を行うことが出来る左下から右上の対角線と平行に並ぶノードを順次処理を行っていき、その列にある全てのノードの点数処理が終了した後に、次のノード列の点数処理へと進むようになっている。最終的に左上のノードの点数を二つの塩基配列の類似度として出力するものとなっている。

また、本シミュレータでは塩基配列の長さに応じて、必要となる全てのノードを仮想的に作り、そのデータを保持することによって、シミュレーションを行ったときに各ノードにおける点数がどんな値であったかを必要に応じて確認することが出来るようになっている。

3-1-2 パラメータ設定

前章で減点量には 5 種類あると述べた。この減点量を変化させることにより、出力される類似度の傾向が変化する。本研究では、類似度を決定する上で重要な値となるこの減点量を「パラメータ」と表現する。

任意の塩基配列と、その塩基配列に塩基変化を行ったものとの間の類似度について調べた。20 塩基からなる塩基配列に対し、0～20 箇所の塩基が置換したときに出力される類似度の傾向を図 3-1 に示す。

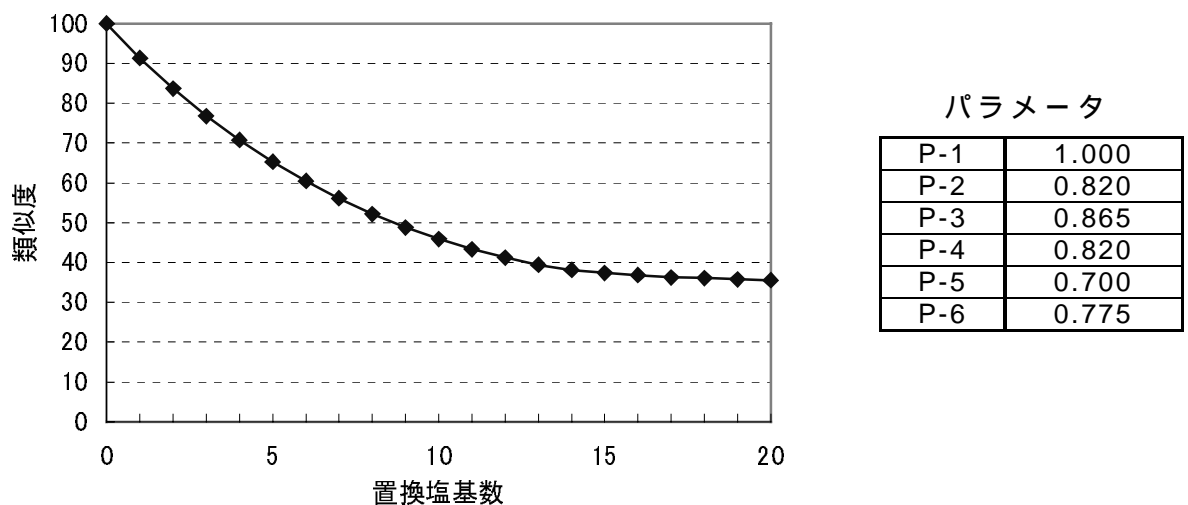


図 3-1 塩基変化数による類似度の傾向(塩基置換)

このように、出力される類似度は指数関数的な傾向となることが分かる。これは減点の方法として 1 以下の数を掛けるという乗算法で減点を行っているため、当然の傾向であると言える。そのため、塩基置換だけではなく他の塩基変化方法でも同様な結果となる。しかし、この減点法では、塩基の変化数が少ないときの点数の減少量が多く、塩基の変化数が多くなるにつれて点数の減少量が少なくなるという傾向となる。乗算法を用いている以上、このような傾向になることは避けられない。

本来、人間がある二つの物を見比べ、それらが似ていると判断する場合、ごく一部のみの変化の場合は「ほとんど同じもの」と判断する。その変化が多少増減しても似ているという認識はあまり変わらない。また、ほとんど一致していない二つのものを見比べた場合、「全然違うもの」と判断する。こちらの場合もほんの一部しかない似ている部分が多少増減しても、似ていないという認識はほとんど変わら

ない。一方、かなり似ているものと全然似ていないもの間にあたる、似ている部分と似ていない部分のふたつの部分を持ったものを見比べた場合、似ている部分の多少の増減により類似性の判断が変わりやすい。

この人間が感じる類似度の違いについて、ある文字列から一部分が欠落したときに元の文字列が推測できるかという例で図 3-2 に示す。(a)の文字列を元に、一部分を欠落させたものが(b)、(c)、(d)となり、欠落文字数がそれぞれ異なる。まず、欠落文字数が少ない(b)の場合では、ほぼ元の文字列に書いてある内容が推測できる。また、この文字列からさらに一文字が欠落した場合、例えば「遺」の文字が欠落したとしても「伝子配列」の部分から、おそらく「遺」が欠落したであろうことは容易推測でき、一文字が欠落したことによる推測度の違いはあまりない。また、ほとんどの文字が欠落してしまっている(d)の場合では、元の文字列を推測することはほとんど不可能であり、仮に欠落している一文字が分かったとしても、元の文字列を推測するのはかなり難しい。それに対し、適度な文字が欠落している(c)では、さらに一文字欠落する場合、「伝」や「複」のように、文字によってはその周辺の文字列を推測する手がかりがなくなってしまうため、推測できなくなる可能性がある。そのため、(c)のような適度な文字列が欠落している場合は、欠落している文字数により推測度が変化する可能性が高い。これらを言い換えると、元の文字列を推測するにおいて、全体のうちのごく一部が欠落している場合や、ほとんど欠落してしまっている場合は、一文字の重要度はそれほど高くなく、適度に欠落している場合は、一文字の重要度が高まるというように、欠落している文字数により一文字一文字の重要度が変化するといえる。

- (a) 膨大で複雑なデータの集合である遺伝子配列
- (b) 膨 で複雑なデ タの集合 ある遺伝子配列
- (c) 膨 複 なデ タの集合 ある 伝 配列
- (d) 複 タ る 配

図 3-2 文字列の欠落数による認識度の違い

この考え方を元に出力される類似度の傾向を考えると、変化した塩基数が少ない、または多い場合は、その塩基数による類似度の変化が少ない、つまり減少度が

少ないという傾向になり、半数程度の塩基が変化している場合は類似度の変化が少ない、つまり減少度が多いというような傾向となるのが適当ではないかと考えた。しかし、図 3-1 のように変化した塩基数が少ないときに大幅に減少し、変化した塩基数が増えるとともに減少量が少なくなるという傾向は、人間の類似物の認識性とは違うものと言える。

そのため、乗算による減点を行い、なおかつ出力される類似度の傾向ができるだけ上記のような傾向となるように、ある点数を基準として、各ノードの点数処理において入力される点数が基準の点数以下だった場合に、パラメータを変化させるという方法を用いることにした。このしきい値を用いたパラメータ変更の例を図 3-3 に示す。

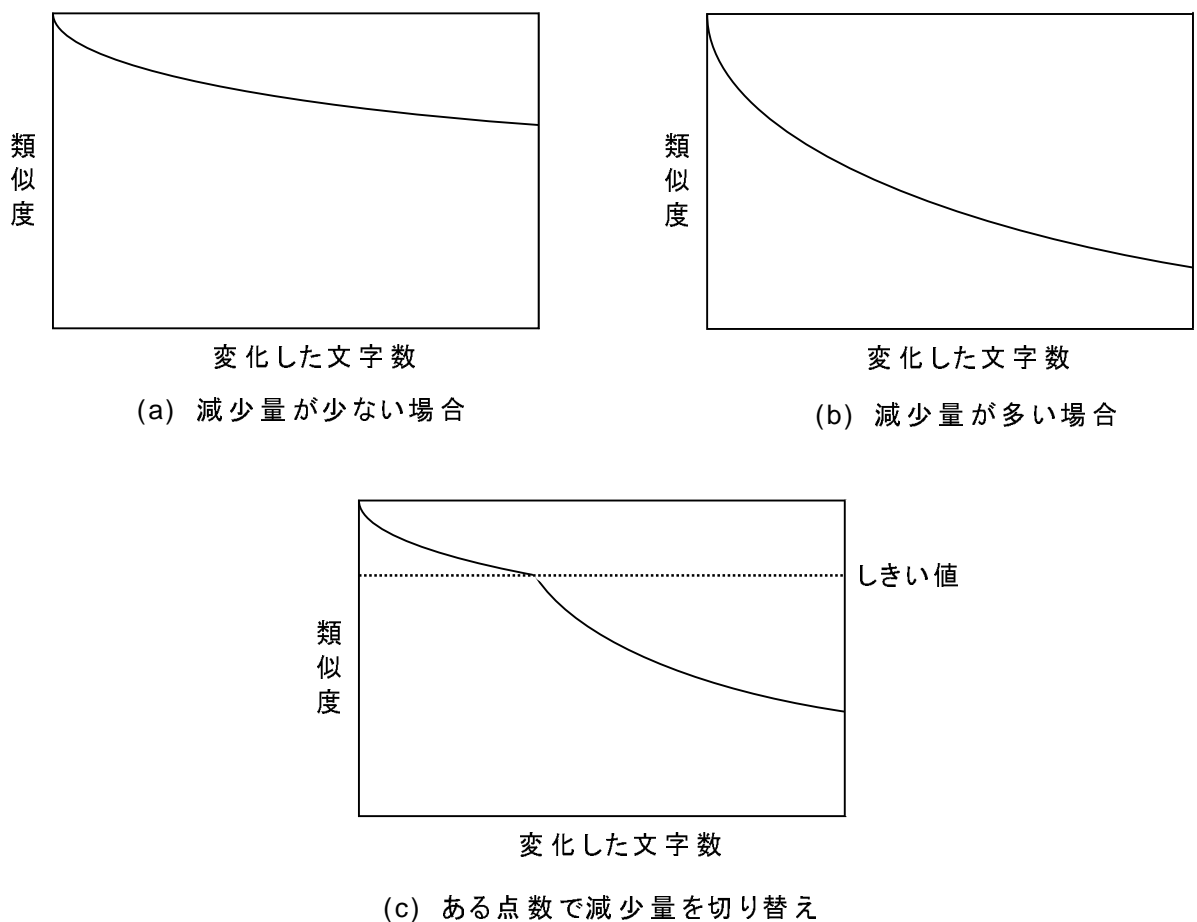


図 3-3 しきい値によるパラメータの変更

図 3-3(a)は減少量を少なくした場合、(b)は減少量を多くした場合の類似度の傾向となっており、ただ減少量を変化させただけでは指数関数のような傾向という

のは変わらない。そこで、ある点数をしきい値とし、そのしきい値より高い点数の場合は減少量を少なく、しきい値を下回った場合に減少量を大きくするようにパラメータを切り替えると、図 3-3(c)のような傾向となる。これは変化した塩基数が少ない、または多い場合は減少度が少なく、半数程度の塩基が変化している場合は減少度が多いという、先ほど述べた人間が感じる類似判定とほぼ同じような傾向となる。

いかに回路の複雑化を避け、なおかつ出力が満足のいく傾向に近づかせるかという二つの点を同時に満たすことは無理であり、両方がそれなりに満足するという条件を選び、それぞれの点で妥協するということになる。そのため、しきい値を用いたパラメータの変更で出力されるグラフの傾向は多少いびつなため、人間の類似度の感じ方と完全に一致した傾向とは言えないが、それなりに満足行く傾向であると言える。そのため、本研究ではこのような傾向でも問題ないと判断した。

具体的なパラメータについてだが、「図 2-10 ノードの一致度による組み合わせ」をしきい値による減点量の変化などを含めて拡張し、表 3-1 のように分類し、パラメータ設定において以下のような条件を設定した。

1. P-1 は減点しない。
2. P-1 より P-2、P-4 より P-5 のほうが減点量が多い。
3. 同じ一致度のパターンの場合、斜め方向である A からの点数の減少量を、縦、横方向である B₁、B₂ からの点数の減少量よりも少なくする。
4. P-3、P-6 の「×」が連続した場合は P-2、P-5 より減点量を減らす。
5. P-2 と P-3、P-5 と P-6 の比率を同じにする。
6. (P-2×P-4)よりも P-5 の減点量を小さくする。
7. しきい値以下のパラメータ P' は、しきい値以上の時のパラメータ P をそれぞれ何倍かすることで求める。

入力元	ノード N	減点量	
		しきい値以上	しきい値以下
A : ○ or ×	○	P-1(減点なし)	P'-1(減点なし)
A : ○	×	P-2	P'-2
A : ×	×	P-3	P'-3
B ₁ , B ₂ : ○ or ×	○	P-4	P'-4
B ₁ , B ₂ : ○	×	P-5	P'-5
B ₁ , B ₂ : ×	×	P-6	P'-6

表 3-1 パラメータの分類

以上のような条件を設定した理由をそれぞれ挙げる。

1. は先に述べたとおり、一致が続いているとみなすため減点しない。

2. は P-2、P-5 は不一致であると判断されるため、当然である。

3. は一致が続く場合には斜め方向に一致を示す「 \circ 」が続くため、斜め方向から来る点数を縦、横方向からくる点数よりも優先させるということである。

4. については、実際の遺伝子突然変異において、一度の突然変異で複数の塩基が塩基置換、塩基欠落、塩基混入するということがある。当然その場合はその複数の塩基数だけ「 \times 」がつながることとなるが、これは一度の突然変異による変化なため、同じ数の塩基が複数の箇所に変化したときより類似性が高いといえる。そのため、「 \times 」が連続した場合は減点を減らすこととした。

5. は、斜め方向と、縦、横方向という面で条件は違うが、「 \circ 」「 \times 」、「 \times 」「 \times 」という条件は同じであるため、4. で述べた「 \times 」が連続した場合の減点量の優遇度を同じにするという意味で、P-2 と P-3、P-5 と P-6 の比率を同じにした。

6. については、「2-3 表を用いた塩基配列の一致比較」で述べた、塩基欠落及び塩基混入の時の黒い「 \circ 」、「 \times 」で示したセルの点数が引き継がれるようにするためにこの条件が必要となる。図 3-4 にずれが生じた際の一致度の様子を示す。

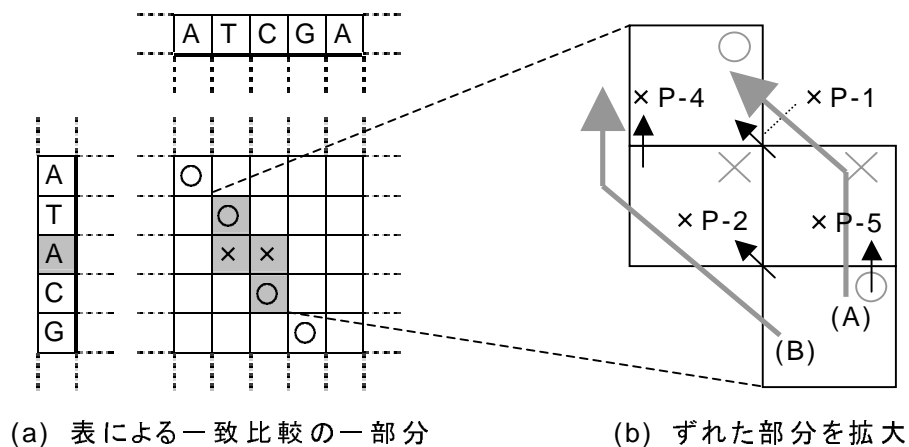


図 3-4 ずれが生じたときに点数が引き継がれるノード

このノードの行列状配置による類似度判定回路のアルゴリズムでは、図 3-5(b) の(A)のルートを通ることとなっているが、点数によっては(B)のルートを通ってしまう可能性がある。また、(A)のルートを通れば1回のずれで減点される回数は1回であるが、(B)のルートを通ると1つのずれで2回減点されることとなるため、

ずれの数と減点回数が一致しないという点からも好ましくない。そのため、必ず(A)のルートを通るようにするためには、(B)のルートを通るよりも(A)のルートを通った方が減点量が少なくなるようにしなければならない。

出発点のノードの点数を 100 点とした場合、それぞれのルートは以下のような点数となる。

$$(A) : 100 \times P - 5 \times P - 1 = 100 \times P - 5$$

$$(B) : 100 \times P - 2 \times P - 4$$

よって $P - 5 > P - 2 \times P - 4$ となれば、(A)のルートを通る方が常に高い点数となるため、この条件が必要となる。

7 は、P' でも 1 ~ 6 の条件を満たすようにするため、それぞれの減点量の比率を維持できるように、パラメータ P のそれぞれの値を何倍かして求めることとした。

以上のような 7 つの条件を満たすパラメータを設定してシミュレーションを行った。実際のシミュレーションでのパラメータの指定は、塩基配列の塩基数が変化しても対応できるように、全塩基数に対する 1 塩基の割合を考慮した減少量を指定しており、実際に掛ける数字を指定しているわけではない。例えば、図 3-1 の P-2 の指定方法でいうと、次のようになる。

$$P-2 = 1 - (\underline{3.6} \div \text{塩基数}) = 1 - (\underline{3.6} \div 20) = 0.820$$

このように実際には上式の下線の付いた斜体の部分を指定している。この章でのシミュレーション結果を示す際にパラメータについても表記しているが、今後はこの斜体の部分を表記する。

3-2 塩基配列生成方法

「2-2-2 遺伝子突然変異により塩基配列に起こる変化」の項で、塩基の変化には塩基置換、塩基欠落、塩基混入の 3 種類あることを述べた。そのため、塩基配列類似判定シミュレータは、任意の複数箇所で塩基置換が起きた塩基配列、塩基欠落が起きた塩基配列、塩基混入が起きた塩基配列を生成するという、3 種類のモードがある。そして

一つの塩基配列中に塩基置換、塩基欠落、塩基混入がランダムに複数起こるという 3 種混合モードも用意した。

さらに、塩基配列が遺伝子突然変異により変化する際に、一度の突然変異で変化する塩基数は一つとは限らず、一回の変化で連続した箇所がまとめて置換、欠落、混入することもある。そのため、前述の 4 つのモードのほかに、連続した箇所が一度に塩基置換、塩基欠落、塩基混入が起きる塩基配列を生成する 3 つのモードもさらに用意した。この 7 種類の塩基配列生成方法について詳述する。

3-2-1 塩基置換

塩基置換とは、塩基配列中の塩基がほかの塩基と置き換わることである。塩基置換の例を図 3-5 に示す。

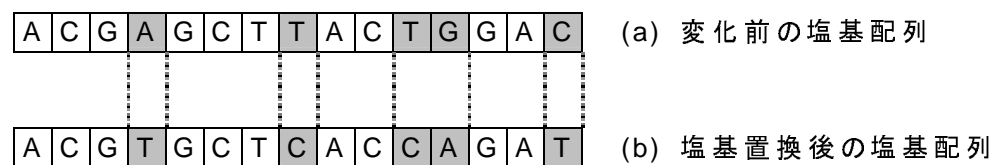


図 3-5 塩基置換による塩基配列の変化

塩基置換の場合は、図 3-5 のように変化させる塩基を指定された箇所数だけあらかじめ決め、そのあとその塩基を違う塩基に変更させている。それにより同じ箇所の塩基が置換されることはなく、例えば 5 箇所の塩基を塩基置換した場合は、図 3-5 のようにならず違う場所の塩基 5 箇所が塩基置換される。

塩基がほかの塩基と置き換わる際に、遺伝子突然変異の場合再び同じ塩基となる可能性もある。その場合塩基変化後の塩基配列は、塩基変化していないものと同じと言えるが、これは結果的に塩基変化していないのと同じである。シミュレーションでは塩基配列が変化したときの類似度を求めることが目的のため、そのようなパターンは考慮せず、本シミュレータが生成する塩基置換した塩基配列では同じ塩基となることはないものとし、必ずほかの塩基と置き換わるようにした。この仕様により、全塩基を塩基置換した場合は、1 箇所も塩基が一致しない塩基配列が生成される。なお、この塩基置換では図 3-5 の例のように連続した塩基が置換されることもある。

3-2-2 塩基欠落

塩基欠落とは、塩基配列中の塩基が欠失し、塩基配列全体が短くなってしまうことである。図 3-6 に塩基欠落の例を示す。

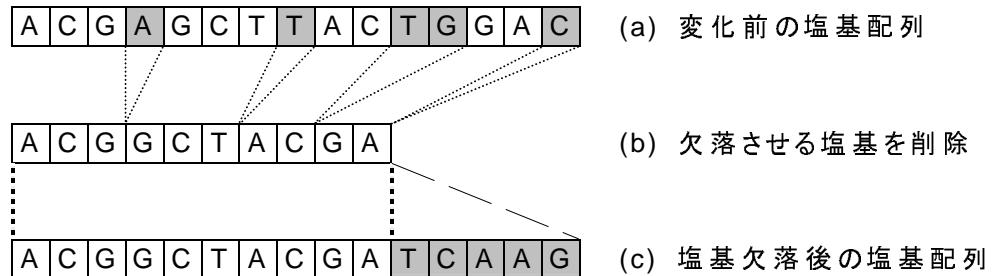


図 3-6 塩基欠落による塩基配列の変化

塩基欠落の場合は、欠失させる塩基をあらかじめ決めておき、まずその塩基を全て削除して、残った塩基を前方向に詰める。こうして出来た塩基配列は当然元の塩基配列に比べ短いものとなっている。しかし、現実には図 3-6(b)のように最後の塩基の後ろには何も無いわけではなく、その次の塩基配列がつながっている。そのため、そして後ろの空いた部分には任意の塩基列を追加している。このような方法で塩基欠落の塩基配列を生成しているため、一度の塩基配列生成で「後ろに追加した塩基をもう一度欠失させる」という事は起こらない。また、全塩基に対して塩基欠落させた場合は、元の塩基配列を任意の塩基配列に置き換えるということになるため、同じ塩基配列となる可能性がないわけではない。この塩基欠落も図 3-6 の例のように連続した塩基列が欠失するということもある。

3-2-3 塩基混入

塩基混入とは、塩基配列中に塩基が混入し、塩基配列全体が長くなってしまうことである。図 3-7 に塩基混入の例を示す。

塩基混入の場合は、塩基置換、塩基欠落の場合と異なり、複数の塩基を混入させる場合も 1 塩基ずつ挿入している。塩基混入が起こる場合、塩基混入が起きた塩基配列は必ず元の塩基配列よりも長くなる。そのため、図 3-7(b)のように最後の塩基がはみ出す形となるので、この塩基を削除して塩基数を合わせる。これを混入させる塩基の数だけ繰り返すことにより、塩基混入した塩基配列を生成している。この塩基混入の性格上、塩基混入した塩基配列生成過程に挿入された塩基は、必ず最終的に生成される塩基配列中に残っているというものではなく、途中で削除されて

しまう可能性がある。つまり元の塩基配列の塩基が必ず削除されるわけではなく、挿入された塩基が削除されるかもしれないということである。そのため、塩基配列の塩基数分、塩基混入を行っても全塩基が入れ替わっているわけではなく、前の方にある塩基は残っている可能性が高く、後ろの方にある塩基は削除されている可能性が高いというものとなっている。この塩基混入もほかの 2 つの塩基変更方法と同様、連続した箇所にも塩基が混入されることもある。

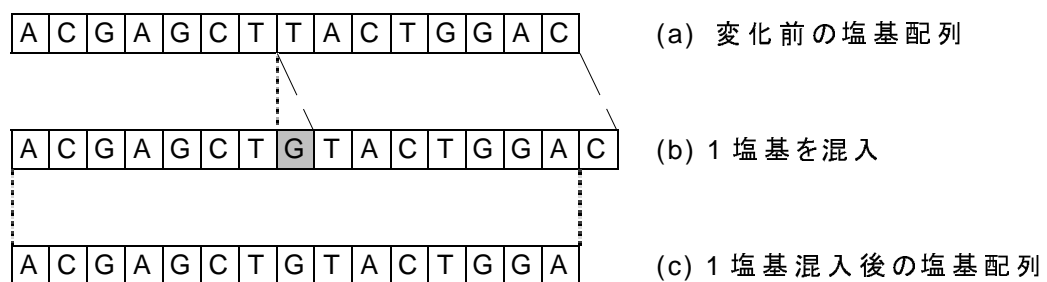


図 3-7 塩基混入による塩基配列の変化

3-2-4 塩基変化(3種混合)

上記で述べた塩基置換、塩基欠落、塩基混入をランダムに行うのが、この塩基変化(3種混合)である。塩基変化の方法がそれぞれ異なるため、1回ごとにどの塩基変化方法でどの塩基を変化させるかを決めている。それを指定回数行うことで、最終的な塩基配列を生成している。この塩基配列生成過程において、どの部分に変更が加えられたかについては調べず、純粹にランダムに選択した塩基について変更を行っている。そのため、塩基置換が行われた塩基について再び塩基置換が行われ、元の塩基に戻ることや、塩基混入で挿入された塩基が塩基欠落で削除され、元の塩基配列に戻ることも十分にあり得る。このことより塩基配列の塩基数分、塩基変化(3種混合)を行っても全塩基が入れ替わるわけではなく、元の塩基配列と同じ配列が出来ている可能性も少なからずあるということになる。

3-2-5 連続塩基置換

連続塩基置換とは、塩基配列中の連続した塩基列がそれぞれ異なった塩基に置換されるというものである。図 3-8 に連続塩基置換の例を示す。

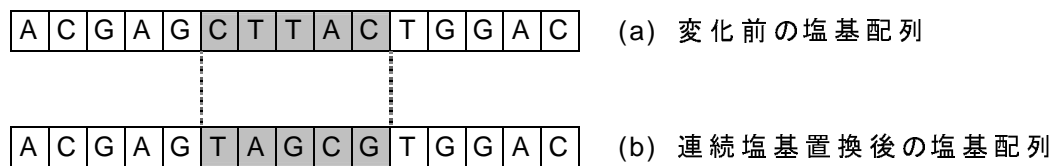


図 3-8 連続塩基置換による塩基配列の変化

図 3-8 のように連続した箇所の塩基がほかの塩基に置き換えられるのが、この連続塩基置換である。この塩基配列生成方法は連続した塩基列が置換したときにランダムな箇所にある塩基の置換と比べて、どのような類似度となるかを確認するために作成したため、連続して塩基置換される部分が複数になることはなく、必ず連続した部分は 1 箇所となる。そのため、塩基置換の時と同様、全塩基数分連続塩基置換が行われた場合は、1 箇所も塩基が一致しない塩基配列が生成されることになる。

3-2-6 連続塩基欠落

連続塩基欠落とは、塩基配列中の連続した塩基列が欠失するというものである。

図 3-9 に連続塩基欠落の例を示す。

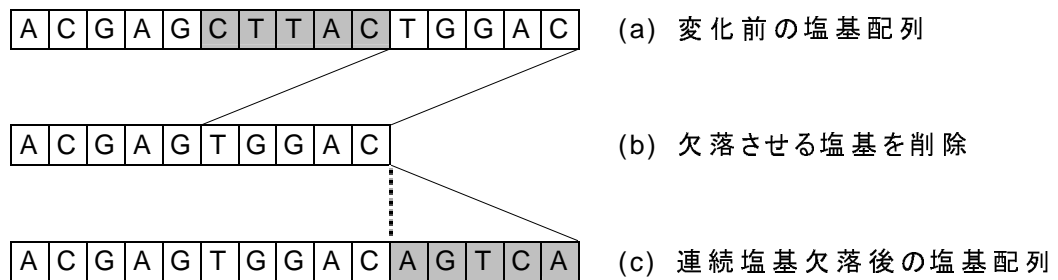


図 3-9 連続塩基欠落による塩基配列の変化

連続塩基置換と同様、この連続塩基欠落も連続した箇所の塩基が欠失し、残った塩基を前方向に詰めた後、後ろの空いた部分に任意の塩基列を追加することで塩基配列を生成する。この場合も連続して塩基欠落する部分が複数になることはなく、必ず連続した部分は 1 箇所となる。また、塩基欠落の時と同様、塩基配列の塩基数分連続塩基欠落を行った場合は、任意の塩基からなる塩基配列に置き換わることとなる。

3-2-7 連続塩基混入

連続塩基欠落とは、塩基配列中に連続した塩基列が挿入されるというものである。図 3-10 に連続塩基混入の例を示す。

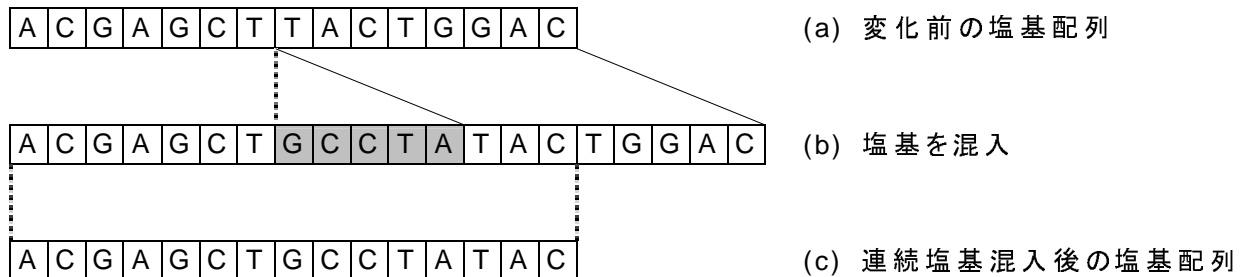


図 3-10 連続塩基混入による塩基配列の変化

この連続塩基置換は、ある 1 箇所に塩基列が挿入され、後ろにはみ出した塩基を削除して塩基数を合わせることにより、塩基配列を生成している。このとき、挿入した塩基列が削除されるということはなく、必ず最終的な塩基配列に残るようにしてある。また、塩基配列の塩基数分連続塩基混入を行った場合、上記の 2 つの連続した箇所の変化による塩基配列生成方法と同じく、任意の塩基からなる塩基配列に置き換わることとなるが、これはランダムな箇所での塩基混入の場合と異なるため、塩基混入と連続塩基混入でのシミュレーション結果を比較するときは、このことに注意しなければならない。

本章で示すシミュレーション結果は、基本的にこの節で述べた塩基置換、塩基欠落、塩基混入、塩基変化(3 種混合)、連続塩基置換、連続塩基欠落、連続塩基混入の 7 種類の塩基配列生成方法で作成した塩基配列を元にそれぞれ類似度を求めたものとなっている。

なお、任意の塩基配列を自動生成するのではなく、ある特定の塩基配列の場合にどのような類似度となるかを確かめるために、二つの塩基配列を指定できるモードも用意した。

3-3 シミュレーション結果

3-3-1 各ノードが出力する点数

実際にソフトウェアでのシミュレータ内の仮想ノードがどのような点数を出力しているのかを、意図的に配列を操作した塩基配列を用いて確認した。10 塩基からなる塩基配列に塩基置換、塩基欠落、塩基混入、塩基変化(3 種混合)が起きた場合の、表 3-2 のパラメータを用いたときの各ノードの出力例を図 3-11、12、13、14 に示す。この表で上端、左端に示された文字列は比較対象である塩基配列、表内の数字は各ノードが出力した点数となっており、小数点第 1 位を四捨五入して表示している。また灰色で示した塩基が塩基変化箇所、太い枠で囲ってあるセルは点数が引き継がれたノードの流れ、灰色で示したセルが減点箇所となっている。

	しきい値以上	しきい値以下	しきい値
P-1	0.00	0.00	80 点
P-2	0.60	2.40	
P-3	0.45	1.80	
P-4	0.60	2.40	
P-5	1.00	4.00	
P-6	0.75	3.00	

表 3-2 パラメータ (図 3-11、12、13、14)

この図を見ると分かるように、第 2 章で提案した文字列間の類似度判定法のアルゴリズムどおりに、各ノードの出力が決定されていることが分かる。

	A	C	T	G	A	C	T	T	G	C
A	79	62	59	80	88	82	83	90	91	96
G	62	79	76	78	82	88	86	87	96	96
T	76	76	84	80	78	85	94	91	91	96
A	85	78	80	84	85	80	87	94	91	96
G	88	85	78	85	88	85	83	90	100	96
C	85	94	85	80	85	94	86	87	91	100
T	83	86	94	87	83	86	94	91	91	96
T	90	87	91	94	90	87	91	94	91	96
A	100	91	91	91	100	91	91	91	94	96
C	96	100	96	96	96	100	96	96	96	100

図 3-11 各ノードの出力(塩基置換)

	A	T	C	T	T	C	G	A	C	T
A	52	50	65	78	78	82	85	94	91	96
G	59	52	60	80	82	82	87	87	94	96
T	77	78	64	80	85	86	86	87	91	100
A	86	79	78	78	78	85	90	91	91	96
G	81	86	82	82	82	81	90	94	91	96
C	82	84	91	82	86	87	83	90	100	96
T	86	87	88	91	87	90	87	87	94	100
T	90	91	87	94	91	87	94	90	91	100
A	96	94	91	91	94	91	91	100	91	96
C	96	96	100	96	96	100	96	96	100	96

図 3-12 各ノードの出力(塩基欠落)

	A	T	G	G	C	T	A	G	T	C
A	76	59	77	82	80	86	96	87	91	96
G	79	76	77	82	86	82	86	96	91	96
T	77	84	78	76	82	91	86	87	96	96
A	77	81	84	81	78	82	91	90	91	96
G	78	77	86	90	81	82	86	91	94	96
C	84	81	79	86	90	82	86	90	91	100
T	86	90	83	83	90	90	86	90	96	96
T	87	91	90	87	87	96	87	90	96	96
A	96	91	91	94	91	91	96	91	94	96
C	96	96	96	96	100	96	96	96	96	100

図 3-13 各ノードの出力(塩基混入)

	A	C	T	C	A	G	C	T	G	C
A	77	78	76	81	90	82	83	90	91	96
G	61	77	82	79	81	90	86	87	96	96
T	73	74	82	86	82	83	90	91	91	96
A	81	75	76	82	91	82	87	94	91	96
G	84	81	78	76	82	91	83	90	100	96
C	82	90	82	83	79	84	91	87	91	100
T	86	86	90	86	83	83	88	91	91	96
T	90	90	91	90	90	87	87	94	91	96
A	100	91	94	91	96	94	91	91	94	96
C	96	100	96	100	96	96	100	96	96	100

図 3-14 各ノードの出力(塩基変化(3種混合))

3-3-2 全体に対する変化した塩基数の割合による類似度の傾向

ある任意の塩基配列に対して、塩基操作する箇所を変化させたときに出力される類似度の傾向を調べた。塩基操作対象の塩基配列は 20 塩基からなり、任意または連続した 0 箇所～20 箇所の塩基を 7 種類の塩基操作により変化させた。パラメータは表 3-3 のようになっている。このパラメータは、回路設計において誤差が少なくなるように最低値が 25 点を下回らないようにして与えた。これについては次章で詳しく述べる。その変化塩基数ごとに 10000 回データを取り、そのデータの平均値、最大値、最小値を出したものを図 3-15、16、17、18、19、20、21 の上部に示す。また、出力された類似度がどのような分布となっているかを出力された類似度の多さにより色分けしたものを同図下部に示す。この図は出力された点数を 5 点おきに集計し、その数に応じて数が多いほど濃い色で示している。

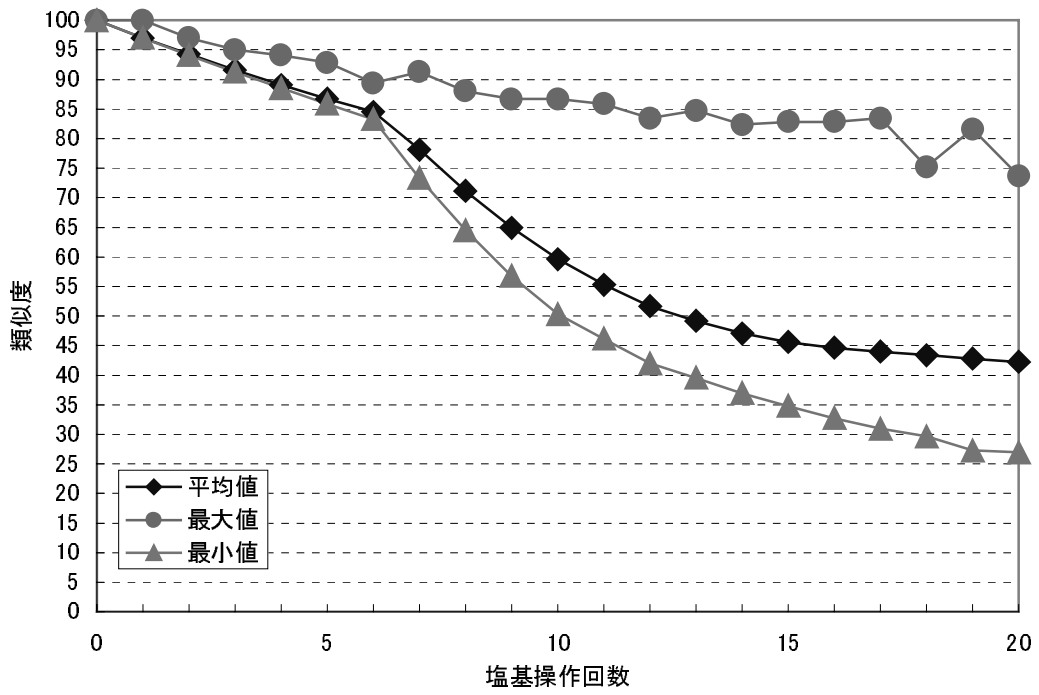
	しきい値以上	しきい値以下	しきい値
P-1	0.00	0.00	85 点
P-2	0.60	2.40	
P-3	0.45	1.80	
P-4	0.60	2.40	
P-5	1.00	4.00	
P-6	0.75	3.00	

表 3-3 パラメータ(図 3-15、16、17、18、19、20、21)

平均値の傾向を見ると、塩基変化数が少ないときの減少量を少なくするようにするという、意図したとおりの結果となっている。これらの傾向については、ランダムな箇所が変化した場合と連続した箇所が変化した場合のグラフを比較して次節で詳しく考察する。

次に分布図の傾向を見ると、塩基置換の場合の出力点数のばらつきが少なく、それに対して塩基欠落、塩基混入の場合は多少ばらつきが多くなっている。そして、ランダムな箇所が変化した場合と連続した箇所が変化した場合を比べると、連続した箇所が変化した場合のほうがさらに出力点数がばらついている。また、分布図を見ると、75 点付近、60 点付近に分布の谷が出来ている。これは、85 点をしきい値とし、その点数を下回った場合は減少量が増えるために、75～80 点の範囲に入る類似度になりにくいということであり、それに伴い 60～65 点の範囲に入る類似度も算出されにくいため、このような分布となると言える。こちらについても次節で詳しく考察する。

塩基置換



出力点数 分布図

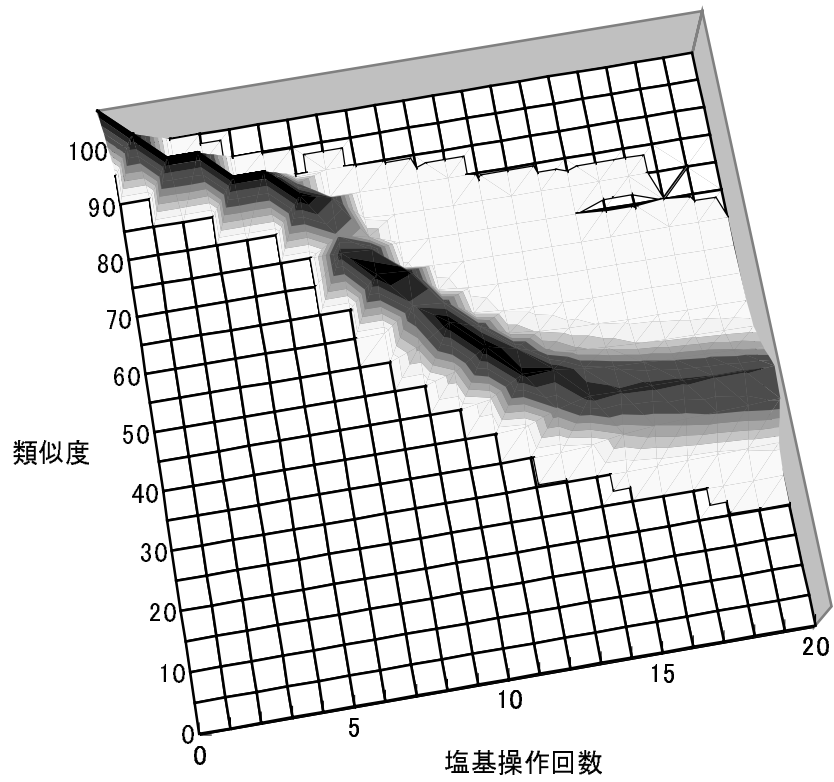
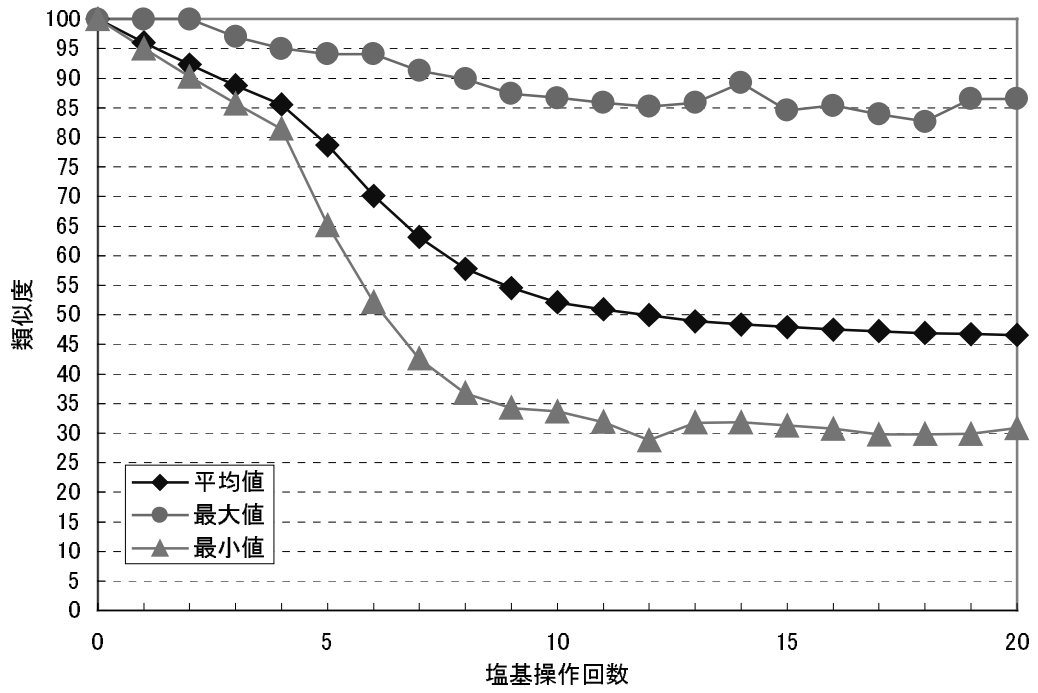


図 3-15 変化した塩基数の割合による類似度の傾向(塩基置換)

塩基欠落



出力点数 分布図

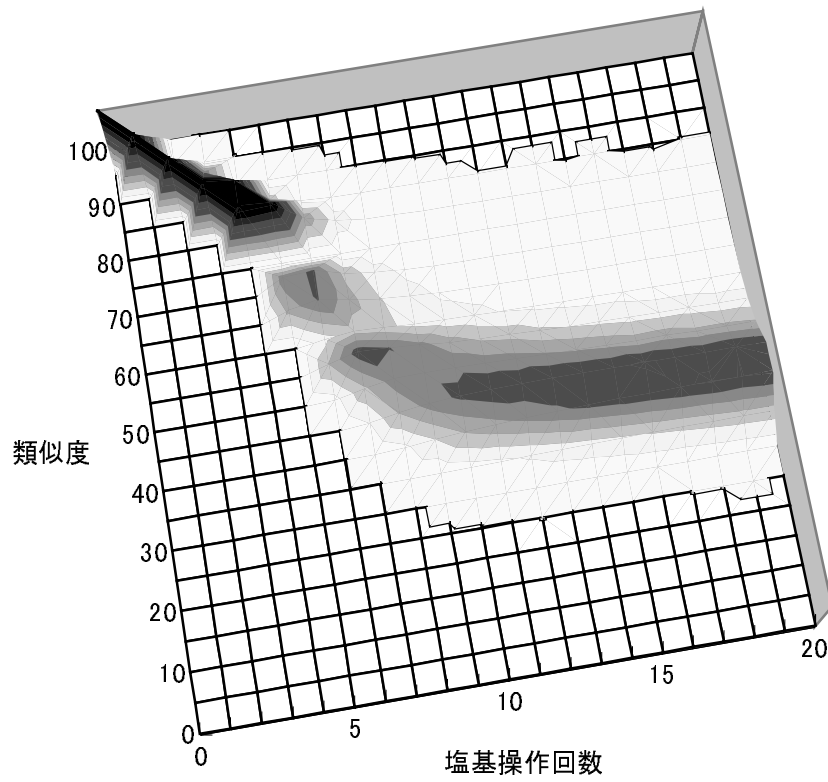


図 3-16 変化した塩基数の割合による類似度の傾向(塩基欠落)

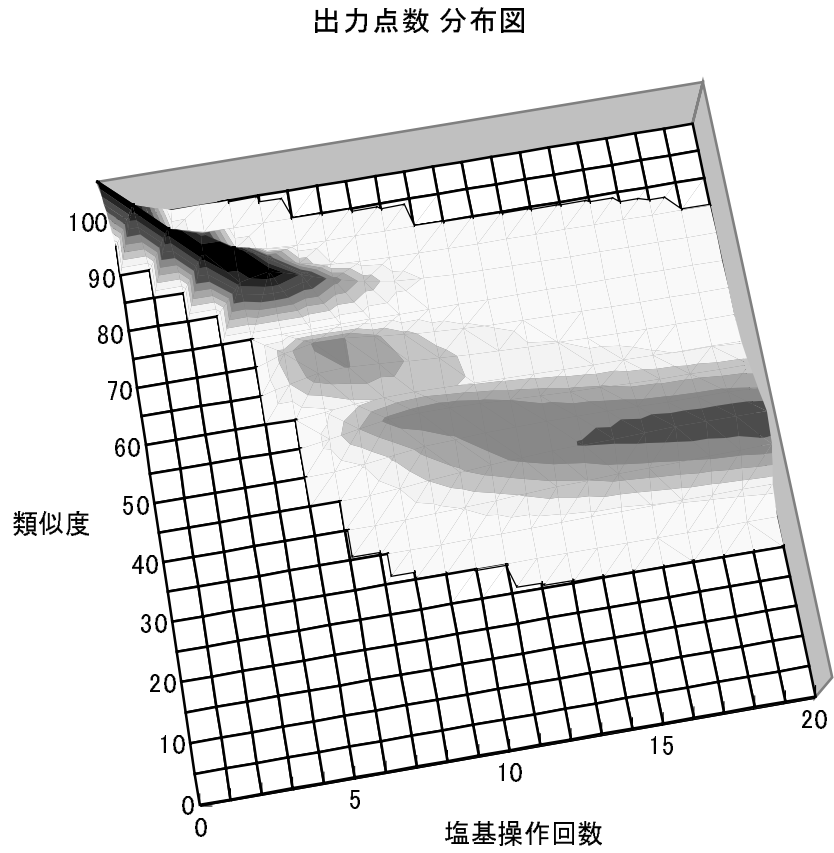
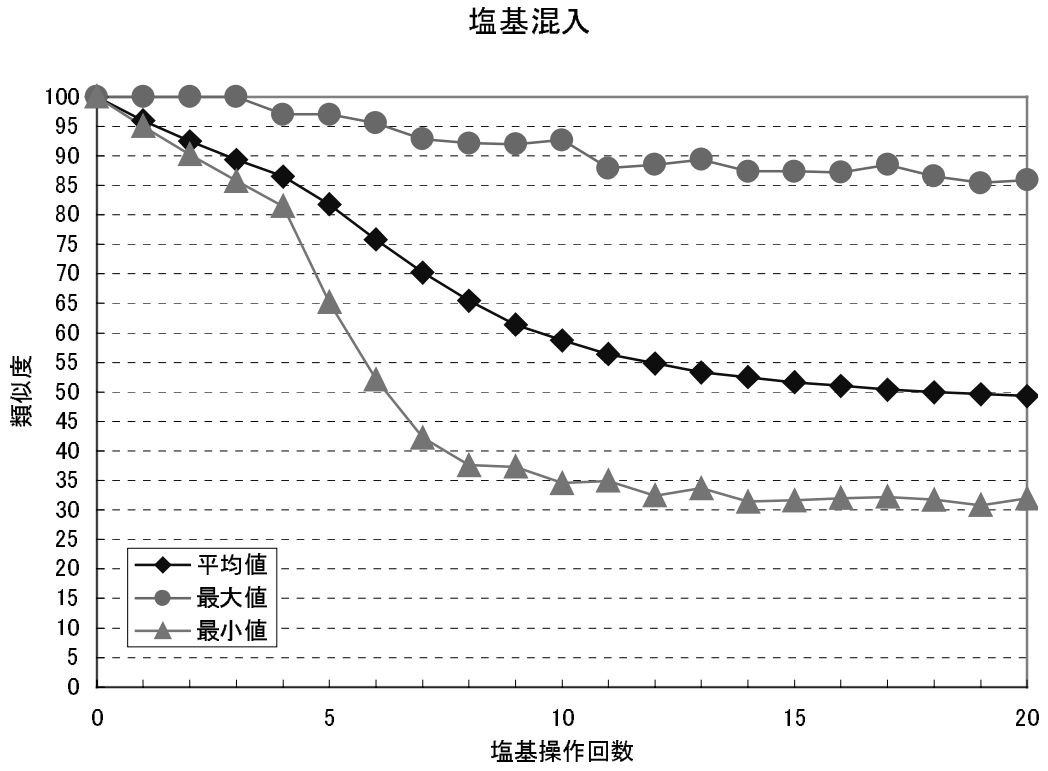


図 3-17 変化した塩基数の割合による類似度の傾向(塩基混入)

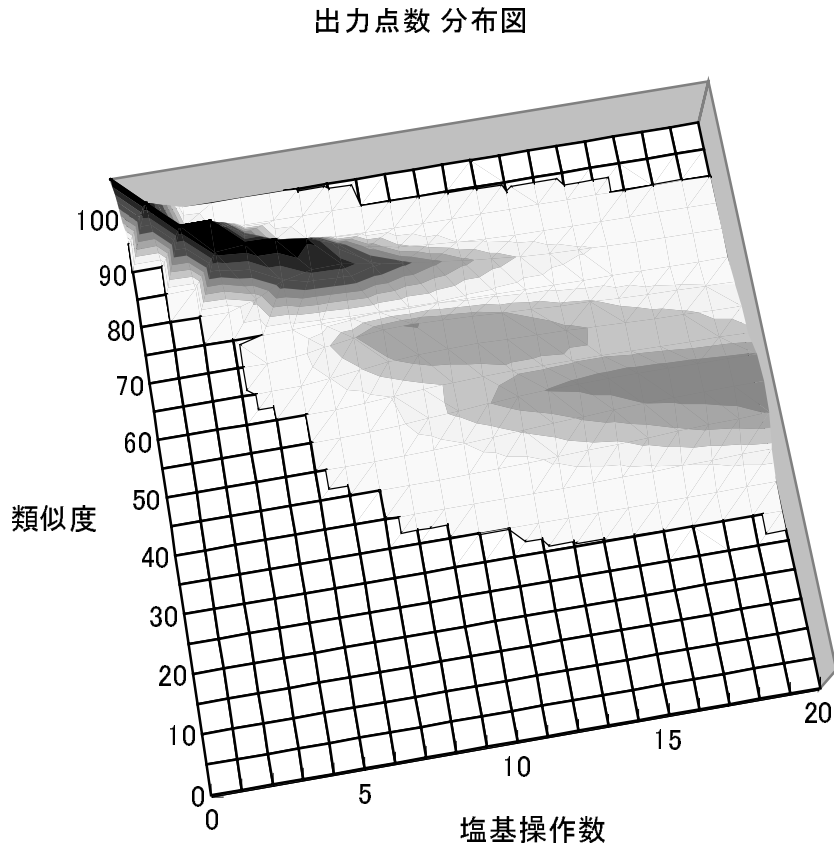
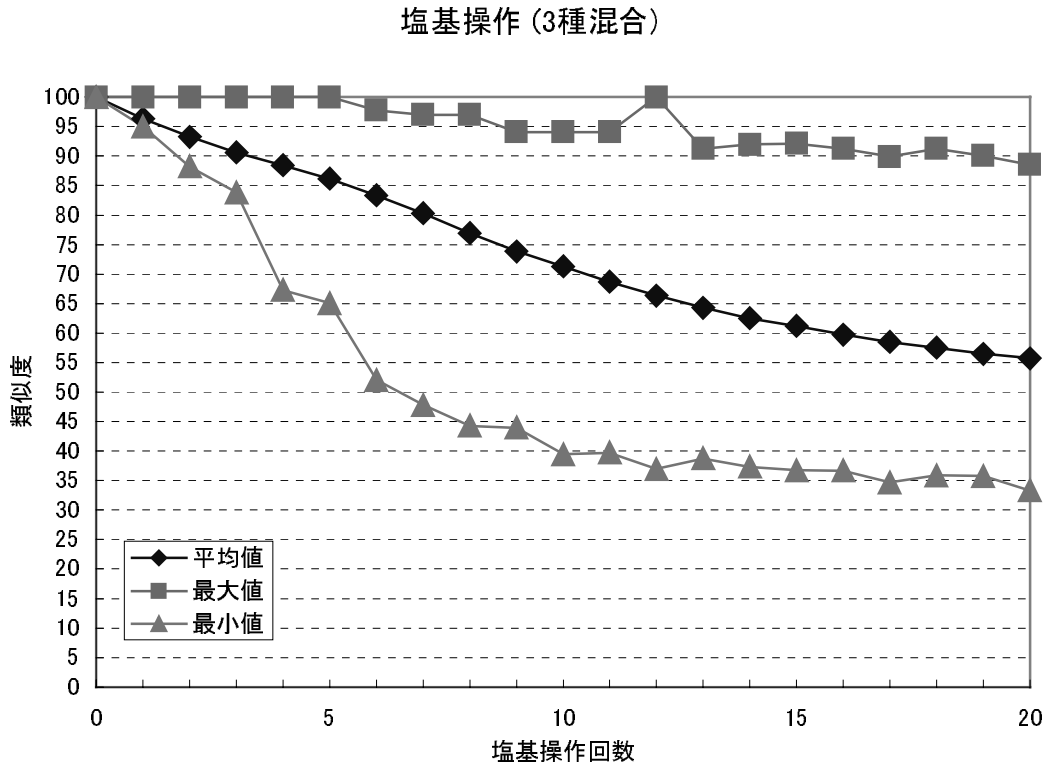
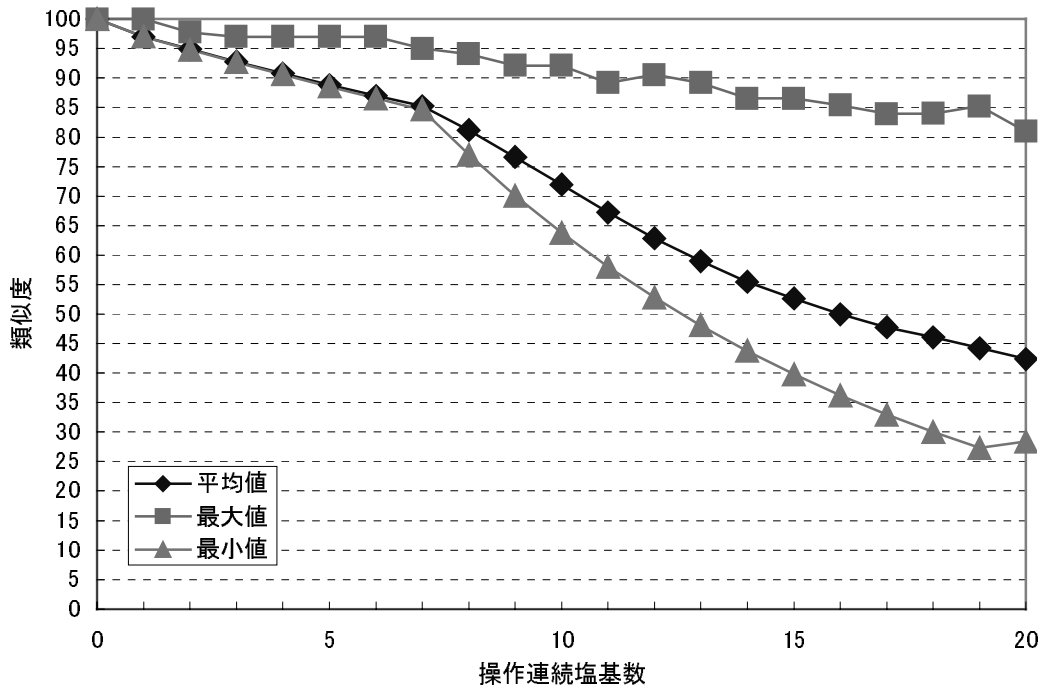


図 3-18 変化した塩基数の割合による類似度の傾向(塩基変化(3種混合))

連続塩基置換



出力点数 分布図

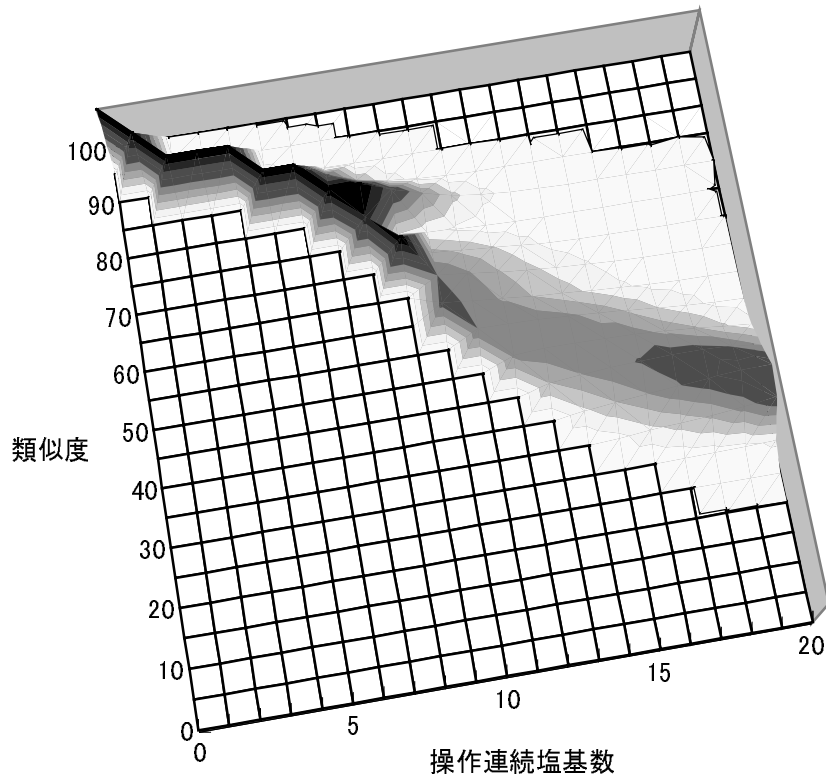


図 3-19 変化した塩基数の割合による類似度の傾向(連続塩基置換)

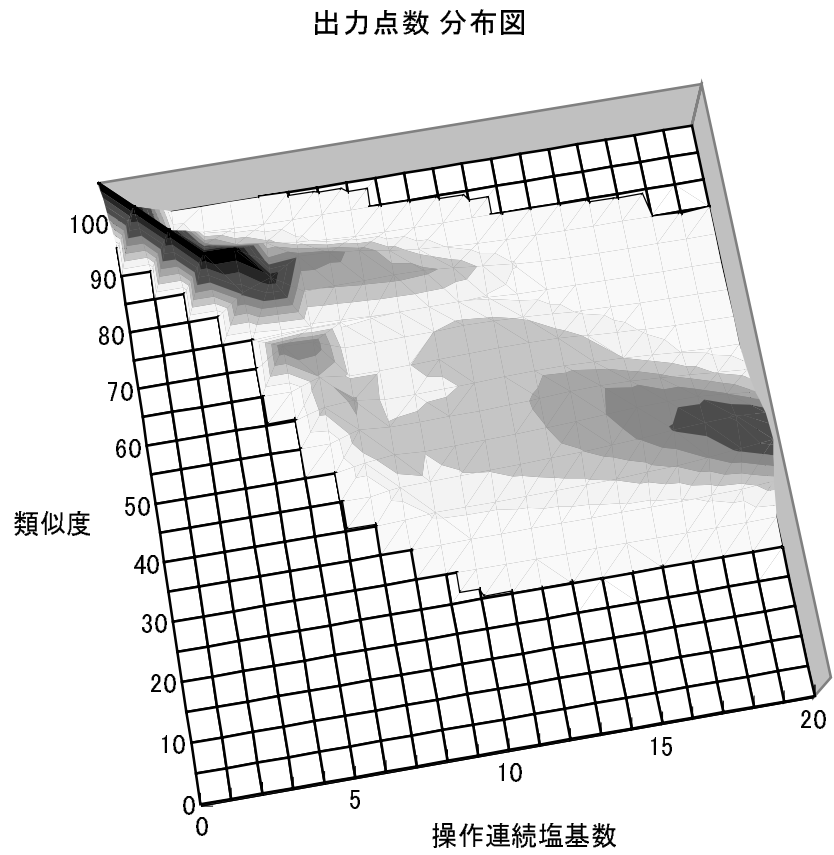
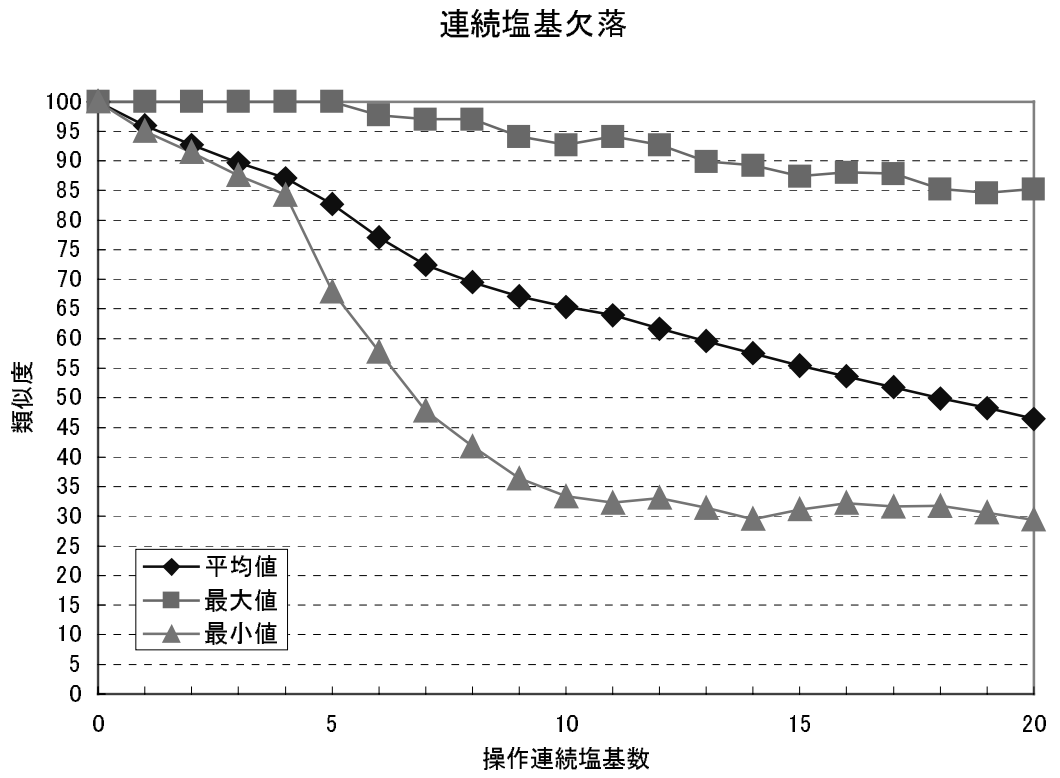


図 3-20 変化した塩基数の割合による類似度の傾向(連続塩基欠落)

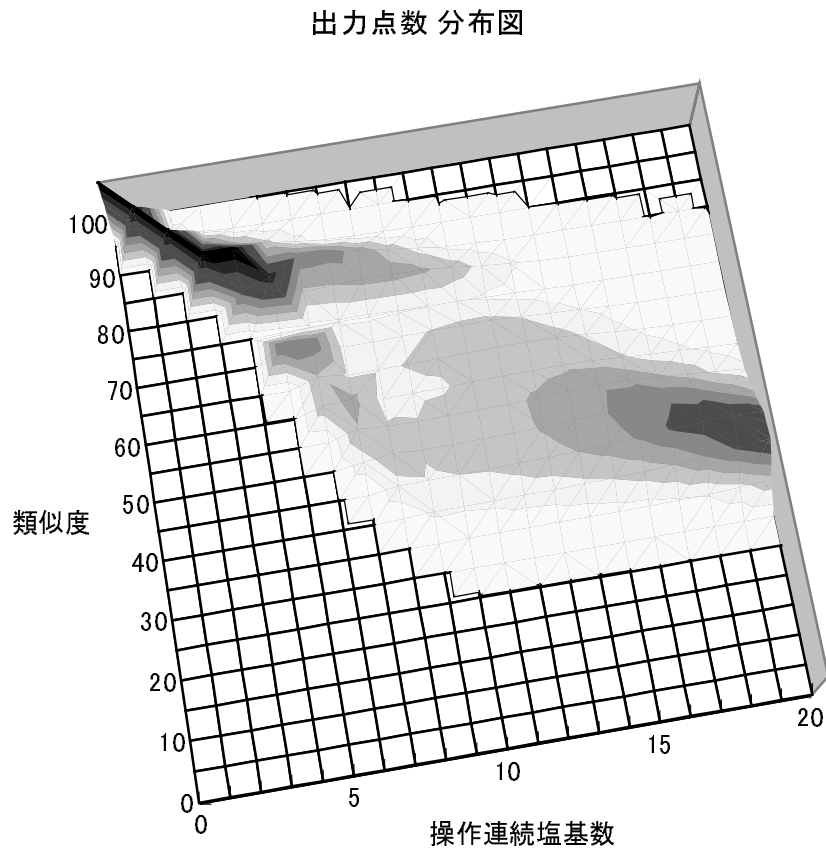
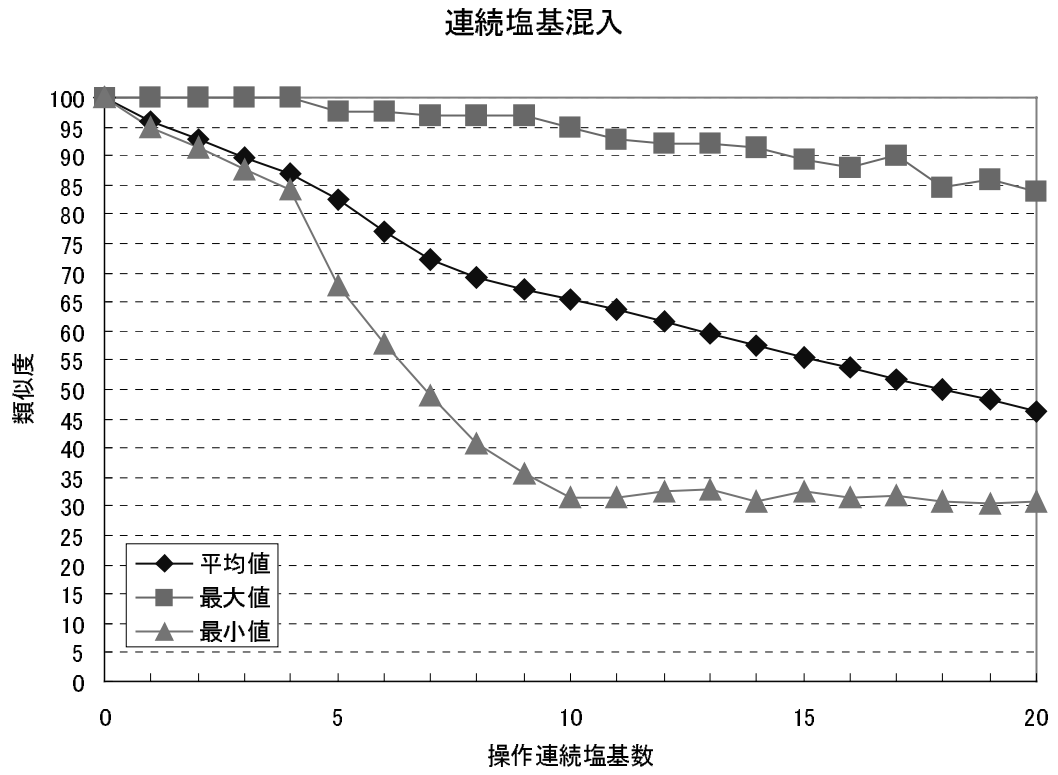


図 3-21 変化した塩基数の割合による類似度の傾向(連続塩基混入)

3-3-3 塩基配列の塩基数の変化による類似度の傾向

この塩基操作する箇所を変化させたときに出力される類似度の傾向をさらに塩基配列の塩基数を変化させて調べた。下に示す図 3-22 は、10、20、40、60、80、100、200、400 塩基からなる塩基配列に対し、全体の塩基配列の 0～100%に当たる塩基を 5% 間隔(10 塩基は 10% 間隔)で塩基置換し、表 3-4 のパラメータでそれぞれの割合において 1000 回シミュレーションを行ったときに出力される類似度の平均値を示している。

	しきい値以上	しきい値以下	しきい値
P-1	0.00	0.00	80 点
P-2	0.80	2.00	
P-3	0.64	1.60	
P-4	0.80	2.00	
P-5	2.00	5.00	
P-6	1.60	4.00	

表 3-4 パラメータ(図 3-22)

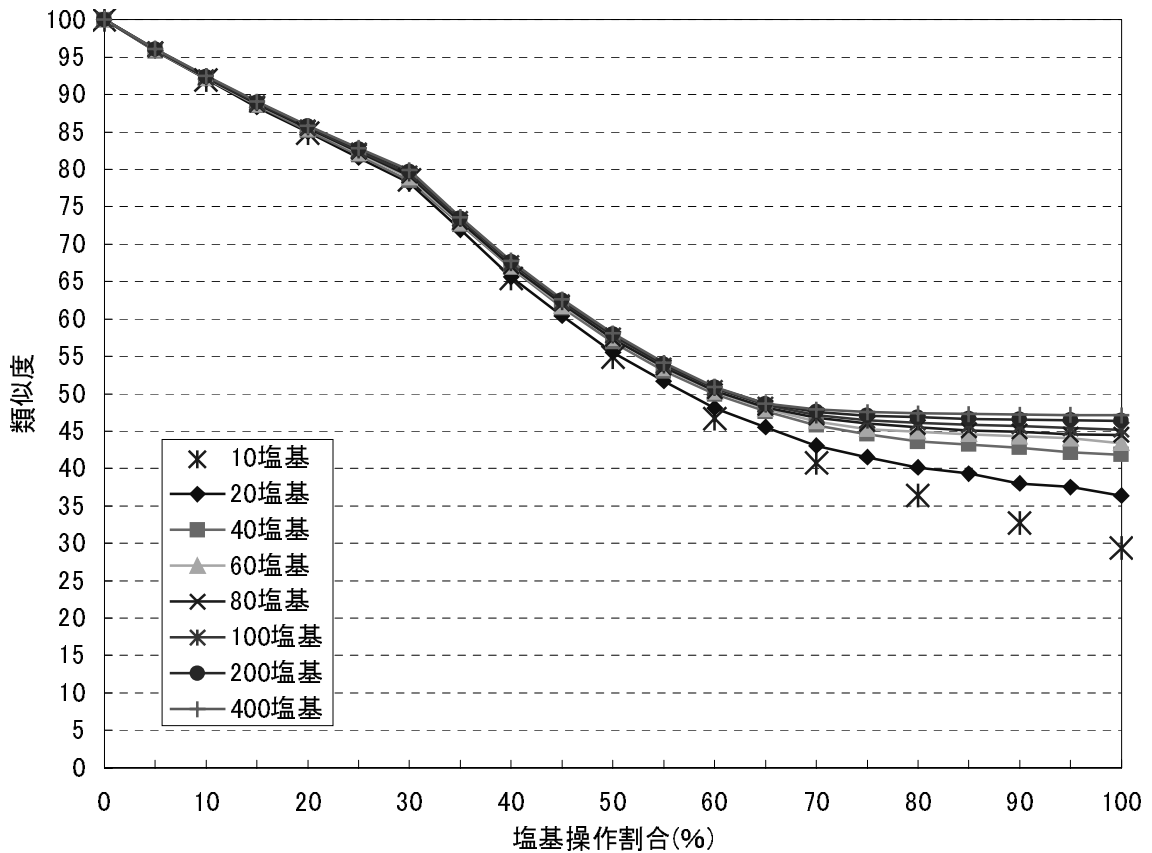


図 3-22 塩基配列の塩基数を変化による類似度の傾向

これを見ると大まかな傾向は変わらないものの、塩基配列の塩基数が多くなるにつれて徐々に出力される点数が上がっていくという傾向となった。これは以下のような理由が考えられる。

まず、本類似度判定シミュレータでは、パラメータを元に塩基数に応じて減点量を変化させている。この時点で、全体に対する変化している文字数の割合が同じ場合でも、図 3-23 のように全体の塩基数が違った場合は違った値となり、塩基数の多い方が高い値となる。

A	T	T	C	G	A	C	T	G	G
A	T	T	A	G	A	C	C	G	G

10 文字中 2 文字の変化 → 全体の 20%
 実際につく点数 : $100 \times 0.92 \times 0.92 = 84.64$

G	T	A	C	G	A	C	A	A	T	G	C	T	A	G	C	G	T	G	A
G	C	A	C	G	A	G	A	A	A	G	C	T	A	G	A	G	T	G	A

20 文字中 4 文字の変化 → 全体の 20%
 実際につく点数 : $100 \times 0.96 \times 0.96 \times 0.96 \times 0.96 \doteq 84.93$

図 3-23 任意の箇所での塩基置換における類似度

さらに、連続した箇所で変化が起きた場合は、2 つ目以降の変化した箇所の減点量が一つ目よりも少なくなる。同じ割合の連続した箇所が変化した場合、減点量が少なくなならない箇所は 1 箇所のみなので、そのため全体の塩基数が多くなるにつれて減点量が少なくなる箇所が相対的に増える。そのため、図 3-24 のように任意の点で変化した場合より連続した箇所で変化した場合のほうが、出力される類似度の差はさらに開く結果となる。

以上のような理由により、今のシミュレータによるパラメータの変更方法では文字数が増えるに従い徐々にグラフの傾向が上がるのは当然といえる。仮に文字数が増えなくても減点量が同じになるようにパラメータを設定できた場合、グラフの傾向が一致する可能性が高い。

しかし、シミュレータで文字数が増えた場合にその文字数に応じて減点量が同じになるような計算式を作らなければならない、さらに 10 種類あるパラメータはそれぞれパラメータの条件を満たすものでなければならない。そのようなプログラムを組むことは困難なため、塩基数を変化させるたびに、その都度塩基数に応じて適切な出力となるパラメータを見つけることとした。

A	T	T	C	G	A	C	T	G	G
A	T	T	A	C	A	C	T	G	G

10文字中2文字の変化 → 全体の20%

実際につく点数： $100 \times 0.92 \times 0.936 = 86.11$

G	T	A	C	G	A	C	A	A	T	G	C	T	A	G	C	G	T	G	A
G	T	A	C	G	A	G	G	C	A	G	C	T	A	G	C	G	T	G	A

20文字中4文字の変化 → 全体の20%

実際につく点数： $100 \times 0.96 \times 0.968 \times 0.968 \times 0.968 \doteq 87.08$

(参考： $100 \times 0.96 \times 0.96 \times 0.968 \times 0.968 \doteq 86.36$)

図 3-24 連続した箇所での塩基置換における類似度

3-4 考察

3-4-1 塩基置換と連続塩基置換

図 3-15、図 3-19 の上図を一つのグラフとして図 3-25 に示す。

どちらもほぼ同じような傾向となっているが、平均値、最大値、最小値ともに連続塩基置換のほうが高い値を示している。これは同じ変化塩基数の場合、ランダムな箇所よりも連続した箇所の塩基が変化していた時に高い点数となるようなアルゴリズムとなっているため、妥当な結果であると言える。なお、20 塩基を変化させた場合はどちらも全ての塩基を置換することとなるため、平均値はほぼ同じとなっている。

最大値に関しては、連続塩基置換での塩基配列は塩基置換のほうでも生成される可能性があるため、理論上同じような値となるはずであり、塩基置換の最大値が連続塩基置換の最大値と重なるはずである。しかし、塩基置換で連続塩基置換での塩基配列を生成するためには、まず連続した塩基のみが変化されることが必要で、さらに元の塩基配列に類似した塩基列へと変化されなければならない。そのためその確率はかなり低く、このような差がでたと思われる。繰り返し回数をさらに増やせば、この最大値はほぼ一致すると考えられる。

図 3-15、19 の点数の分布を見ると、ほぼ同じ傾向であるが、若干塩基置換のほうが点数のばらつきが少ない。塩基置換の点数を求める際に重要となることは、単独の箇所で変化した塩基の個数と、連続した箇所で変化した塩基の個数である。その

個数の組み合わせにおいて出現する確率を考えると、例えば塩基変化が少ない場合は連続した箇所に変化する確率はかなり低く、ほとんどが単独の箇所に変化するというように、もともと組み合わせにばらつきが少ないため、このように塩基置換の方はばらつきが少ないと言える。それに対し、連続塩基置換のほうは、変化する連続した箇所は同じであるため、基本的に点数は変わらない。そこで、点数を左右するのは、元の塩基配列とどれだけ類似している塩基列に置換したかということである。これによってかなり似ている塩基列に置換した場合はかなり高い点数となり、全然似ていない塩基列に置換した場合は低い点数となる。そのため、連続塩基置換はこのようにばらつきが生じると考えられる。

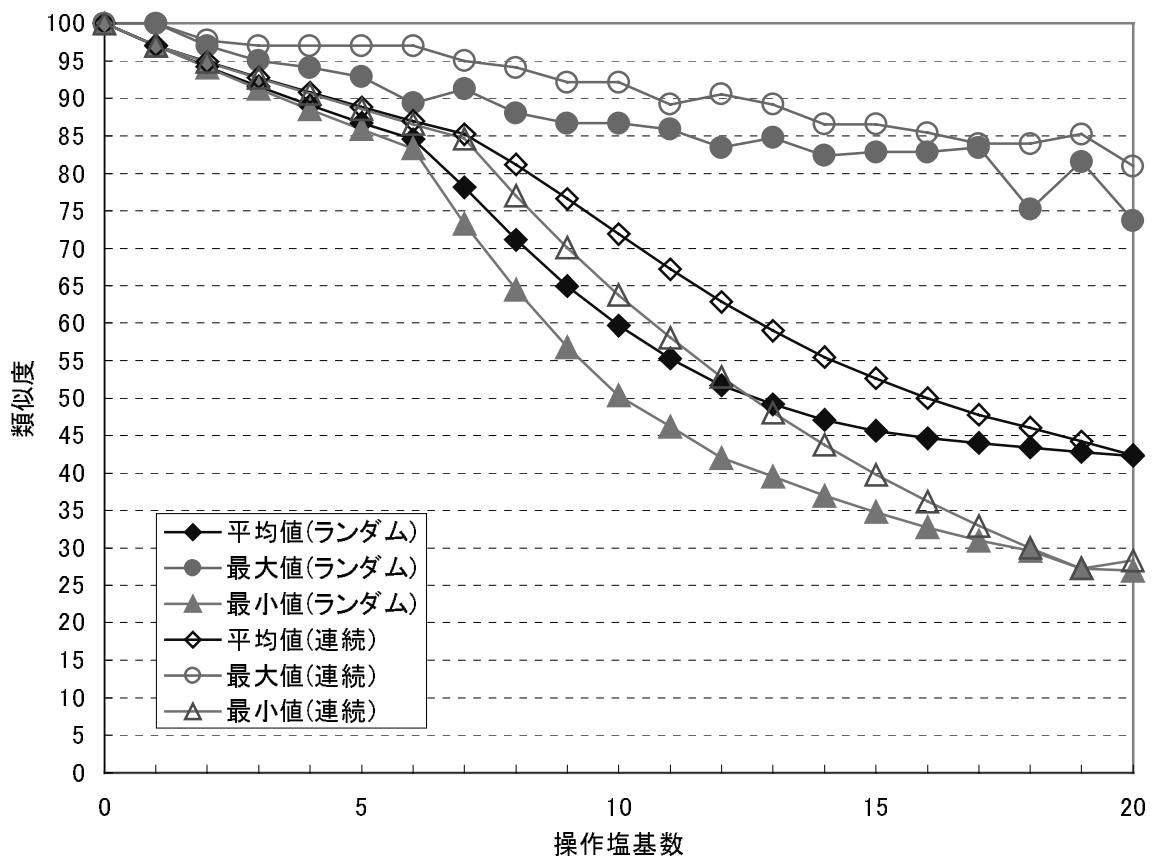


図 3-25 塩基置換と連続塩基置換の比較

3-4-2 塩基欠落と連続塩基欠落の比較

図 3-16、図 3-20 の上図を一つのグラフとして図 3-26 に示す。

こちらについても塩基置換と同様、平均値、最大値、最小値ともに連続塩基欠落

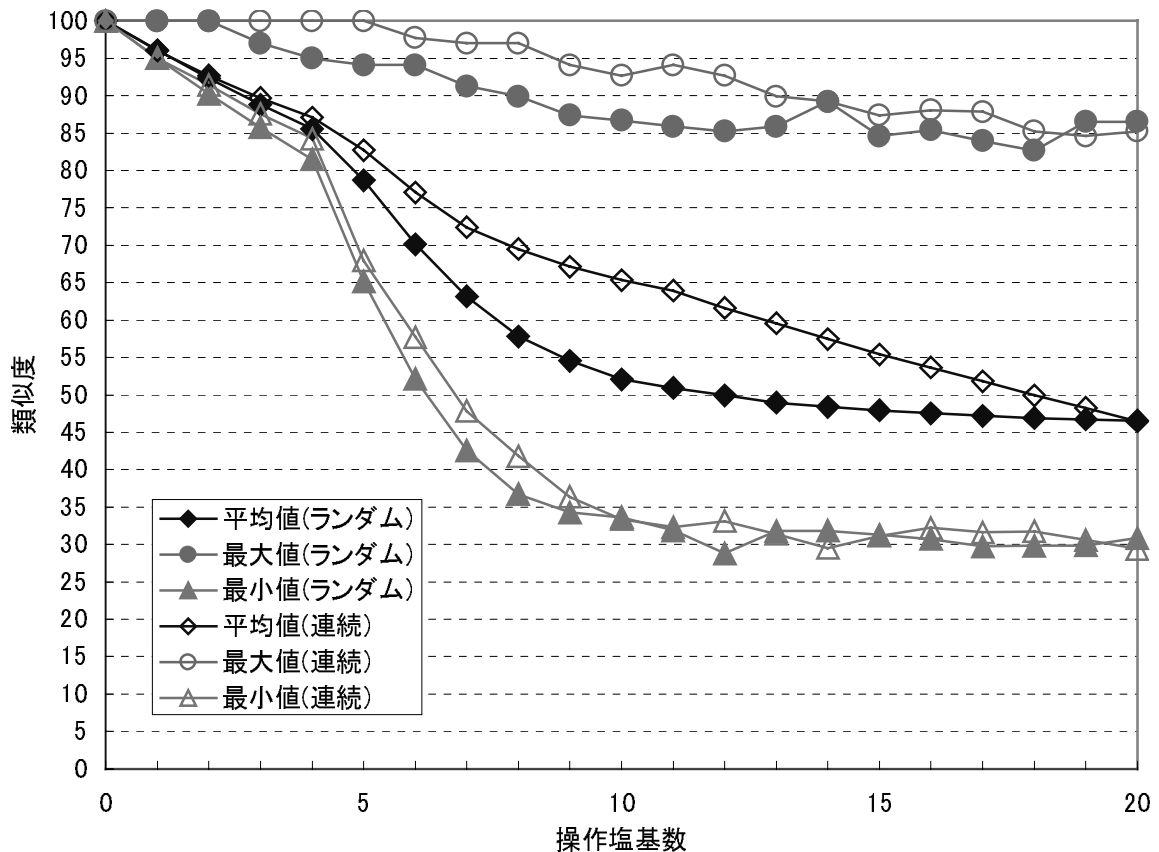


図 3-26 塩基欠落と連続塩基欠落の比較

のほうが高い値を示している。また、平均値を比較すると塩基置換の二つと比較してもかなり差がある。これには同じ変化塩基数の場合連続した箇所の塩基が変化していた時に高い点数となるようなアルゴリズムとなっているという理由のほかに、塩基欠落の場合は欠落させた塩基数だけ、後ろに任意の塩基列を追加することが理由として挙げられる。塩基置換の場合は必ず違う塩基と置き換わっているが、塩基欠落の場合追加する塩基は任意であるため、例えば一番後ろの塩基が欠落したときに同じ塩基が追加されると、元の塩基と同じものになるなど、塩基置換に比べて類似性の高い塩基列になりやすいという傾向がある。そして、連続塩基置換の時と同様、点数を左右するのは後ろに追加された塩基列が、元の塩基配列とどれだけ類似しているかということであり、類似性の高い塩基列になりやすいということは、それだけ生成された類似度の分布の幅が高い点数のほうへと広がるということにつながる。そのため、結果的に平均値が上がると考えられる。なお、20塩基を変化させた場合は、どちらも全ての塩基をランダムな塩基列となるため、平均値はほぼ同じとなっている。

最大値に関しては、塩基置換の時と同様、理論上同じような値となるはずであり、繰り返し回数をさらに増やせば、この最大値はほぼ一致すると考えられる。

3-4-3 塩基混入と連続塩基混入の比較

図 3-17、図 3-21 の上図を一つのグラフとして図 3-27 に示す。

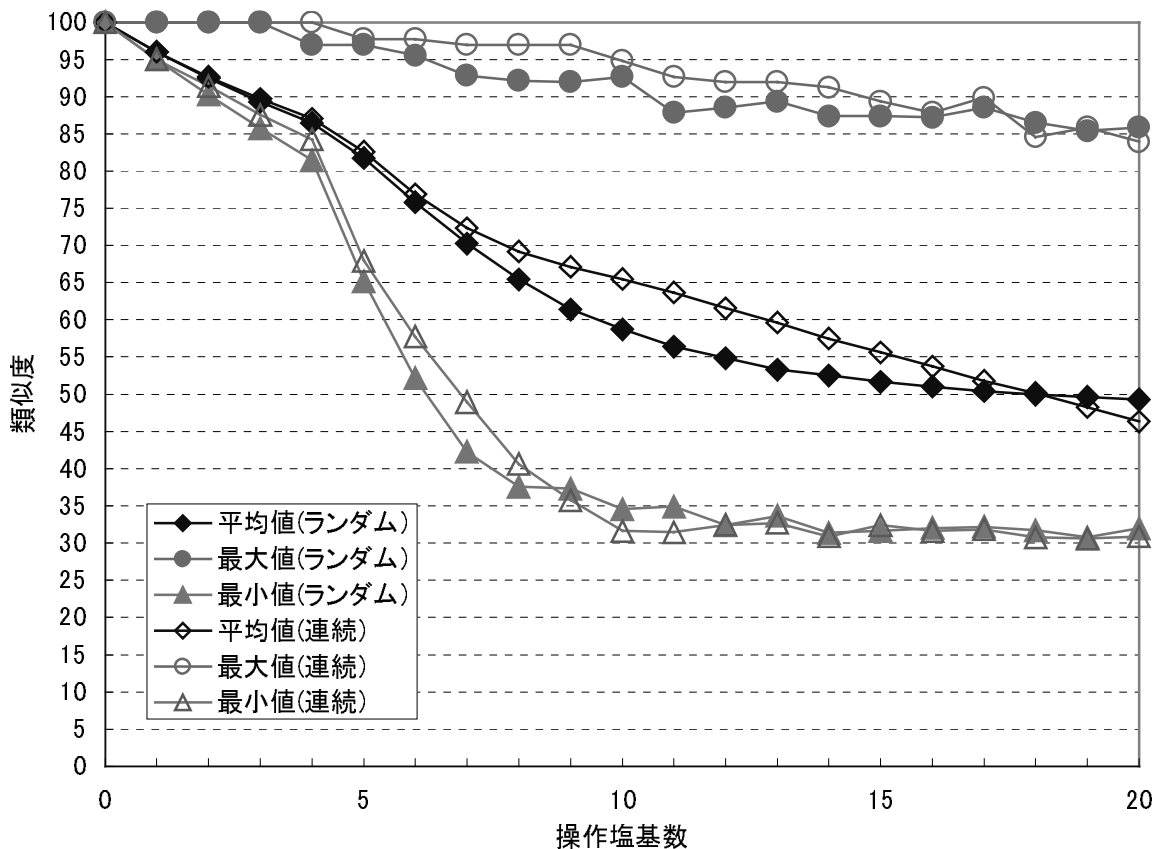


図 3-27 塩基混入と連続塩基混入の比較

このグラフを見ると、一見塩基混入と連続塩基混入のグラフがほとんど一致しているように見えるが、これは二つの塩基列生成方法によって生成される塩基列にあまり差がないからではない。「3-2 塩基配列生成方法」で述べたとおり、塩基混入により生成される塩基配列は、塩基配列の塩基数分、塩基混入を行っても全塩基が入れ替わっているわけではない。つまり、元の塩基列が残っている可能性が高い。それに対し、連続塩基混入の場合は挿入する塩基が必ず最終的な塩基配列に残るようにしてあるため、塩基配列の塩基数分、塩基混入を行えば全塩基が入れ替わり、任意の塩基配列となる。そのため、同じ回数塩基操作を行った時に生成される塩基配列に含まれる、元の塩基配列の塩基数は異なっていることになる、つまり塩

塩基混入は元の塩基が残っている可能性が高いため、全体的に高い点数となるといえる。本来、塩基欠落と塩基混入は一致している箇所がどちらにずれかの違いであって、基本的には点数の傾向は同じとなるものである。そのため、本来の塩基混入のグラフは塩基欠落のグラフとほぼ一致する傾向となると考えられる。それが塩基配列生成方法の違いのために、全体的に高い点数となり連続塩基混入の傾向に近づく結果となったと考えられる。

3-4-4 7種類の塩基配列生成方法による傾向の違い

以上の7種類の塩基配列生成方法による類似度の傾向のグラフのうち、平均値を示すグラフを一つにまとめた。それを図 3-28 に示す。

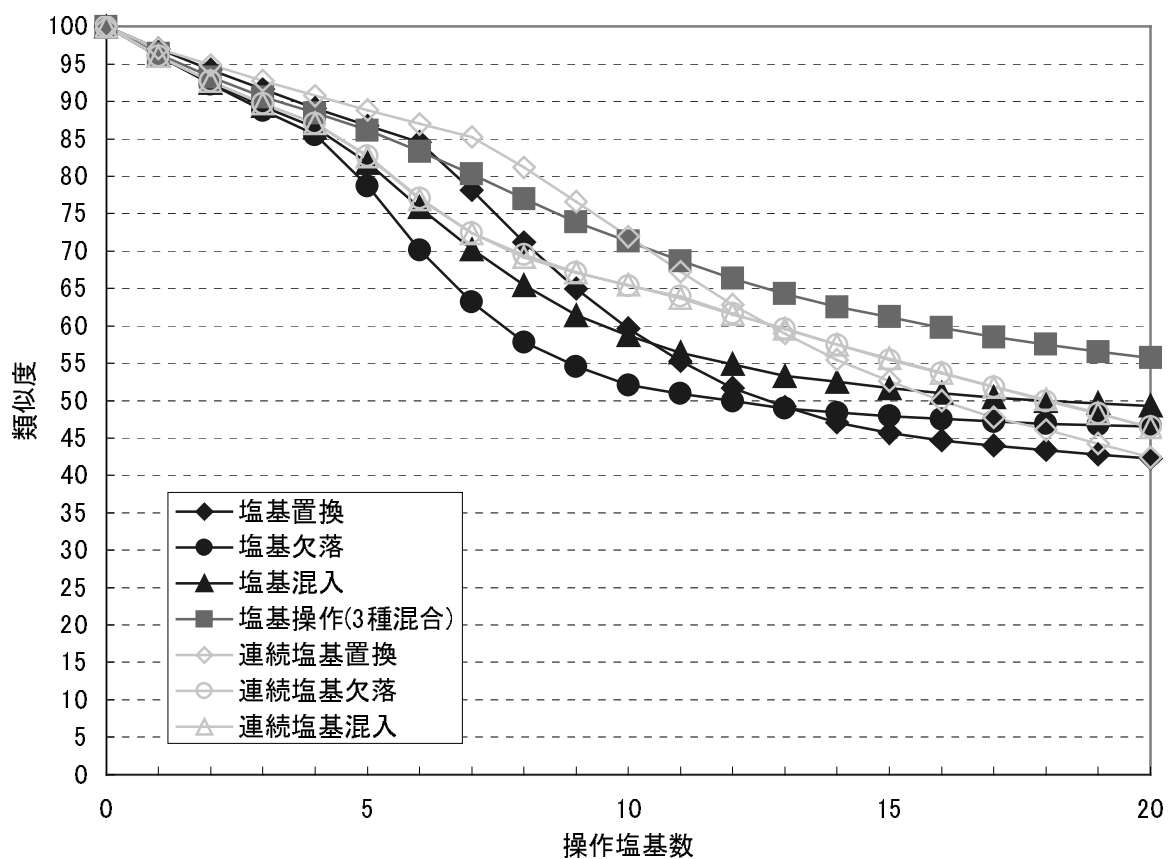


図 3-28 7種の塩基配列生成方法による平均値の傾向

この平均値を示すグラフのうち、まずランダムな箇所塩基置換、塩基欠落、塩基混入した場合の類似度の傾向について考察する。これをみると塩基置換、塩基混入、塩基欠落の順に高い値となっている。塩基混入については前項で述べたように、本来塩基欠落と同じ傾向となるはずなので、塩基置換と塩基欠落について比較す

る。塩基変化が起きた塩基が 13 塩基までは塩基置換のほうが高い値を示し、14 塩基以降は塩基欠落のほうが高い値となっている。これは図 3-12 のように、変化した塩基数が少ない場合は、点数の決定に関して後ろに追加された塩基列の類似性にはあまり関係がない。それに対し塩基置換の場合は同じ塩基列に変化することはないものの、前後の塩基と同じ塩基となる可能性はあるため、点数の決定には置換された塩基にも影響を受ける。このため、変化した塩基数が少ない範囲では塩基置換のほうが高い値となったと思われる。また、変化した塩基数が多い場合では、塩基欠落の場合、似ているといえるほどの塩基列が残るわけではないので、実際には図 3-12 のように残った塩基に対応したノードの点数が引き継がれているとは限らない。そうなると、元の塩基配列に対する後ろに追加された塩基列の類似性が重要となるわけで、同じ塩基に置換されることはない塩基置換に比べて、任意の塩基列が追加される塩基欠落のほうが似た塩基列になる可能性が高いと言える。そのため、このような違いが現れたと考えられる。

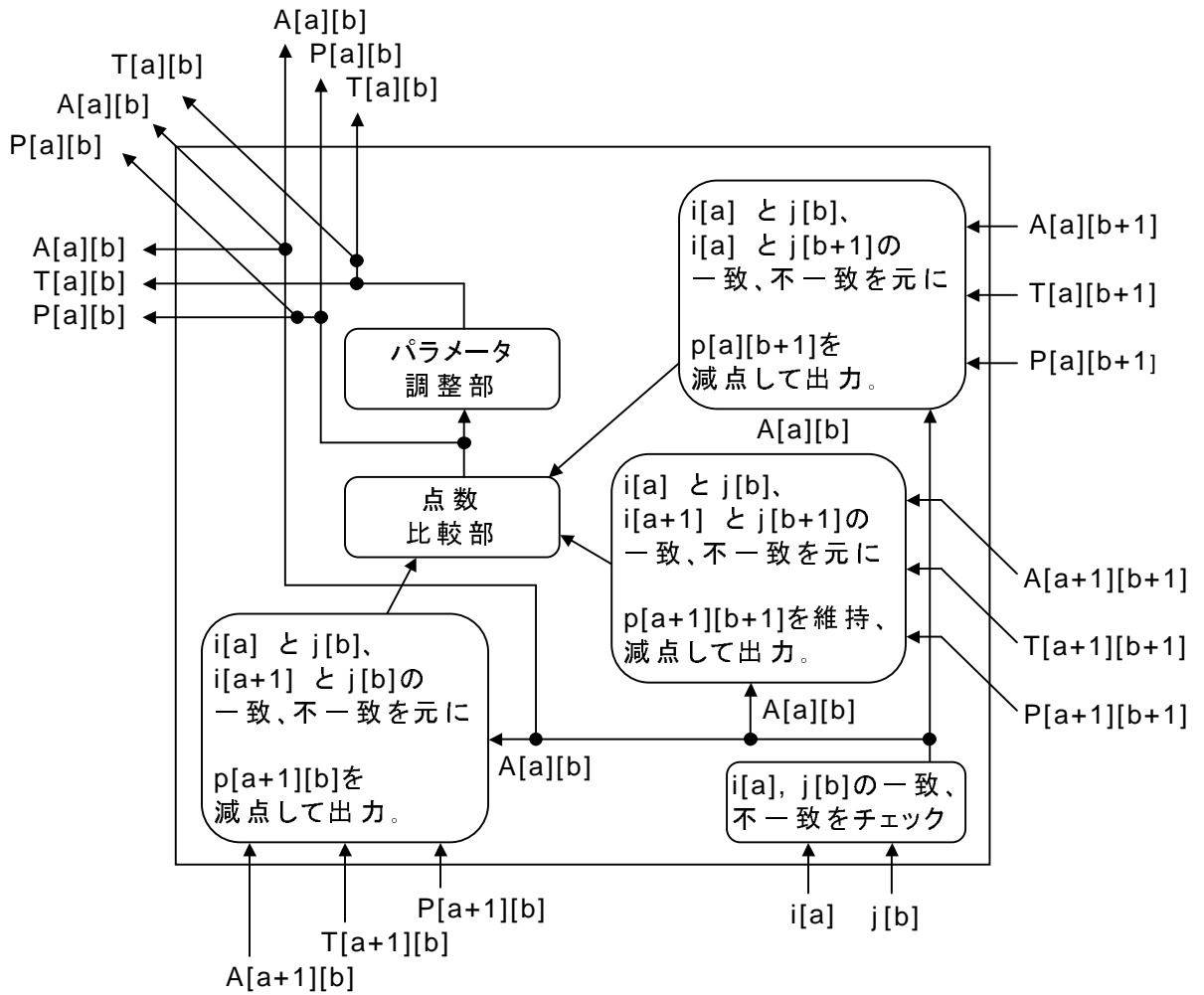
次に、連続塩基置換、連続塩基欠落、連続塩基混入した場合の類似度の傾向を見てみる。こちらに関しては連続塩基欠落、連続塩基混入はきれいに一致しており、出力された点数の分布もほぼ同じとなっている。これはすでに述べていたとおりの結果と言える。また、連続塩基置換と、連続塩基欠落、連続塩基混入の関係についてもランダムな箇所でのそれらとの関係と同じと言える。

最後に塩基変化(3 種混合)について考察する。この塩基変化(3 種混合)のグラフは操作塩基数が多くなるにつれて、他の 6 種類の塩基配列生成方法よりも高い値となっている。これは「3-2-4 塩基変化(3 種混合)」で述べたとおり、20 回塩基変化を行っても全ての塩基が入れ替わるわけではないということがまず挙げられ、さらにこの入れ替わる塩基は、塩基混入の場合よりも少ないため、より高い値を示していると言える。また、出力された点数の分布を見ると、かなり広範囲に点数が散らばっており、様々な塩基配列が生成されていることが分かる。この塩基変化(3 種混合)では、ほかの 6 種類の塩基配列生成方法による塩基配列を生成することが理論的には可能であり、さらに全く同じ塩基配列になる可能性もあるため、高い類似性を持つ塩基配列からほとんど似ていない塩基配列まで、多種多様な塩基配列が生成された結果だと言える。

第4章 高速類似度判定回路の構成

4-1 各ノードの回路構成

以上のようなシミュレーション結果を元に、我々が提案する高速類似度判定回路の各ノードを設計した。図 4-1 にノードのブロック図を示す。



- $i[x], j[y]$: 塩基情報
- $A[x][y]$: $i[x], j[y]$ の一致情報
- $P[x][y]$: $i[x], j[y]$ のノードにおける点数
- $T[x][y]$: $P[x][y]$ がしきい値以上かどうかの情報

図 4-1 ノードのブロック図

このブロック図は、隣接したノードの一致情報である $A[x][y]$ 、しきい値より高い点数かどうかを示す $T[x][y]$ を受け取り、その情報を元に $P[x][y]$ の点数を減点する 3 つの減点部、減点部から出力されたうち最も高い点数を選択する点数比較部、出力する点数がしきい値以上かどうか調べ、次のノードでのパラメータを決定するパラメータ調整部からなっている。

このブロック図を元に電子回路を作成した。図 4-2(a)、(b)に示す。なお、NOR、E-OR、NOT、コンパレータに関しては論理回路で示してある。この電子回路図で赤で示した部分が入力、青で示した部分が出力となっており、また(a)と(b)は紙面の都合上分割して示してあるが、「Out 1、2、3」でつながっている。

この回路はブロック図同様、大まかに減点部、点数比較部、パラメータ調整部により構成されており、減点部は 3 つそれぞれに二つのパラメータによる回路が用意されており、出力する点数を $T[x][y]$ で切り替えるという仕組みになっている。点数は電圧を用いることで表現し、シミュレーションと同じパラメータを回路で実現するために、抵抗の比を利用して入力された電圧を降下させている。また、 $P[a][b]$ の出力部にはほかのノードからの影響を取り除くために buffer がつけてある。この buffer については次節で詳しく述べる。なお、一つのノードあたりのトランジスタ数は 206 個となっている。

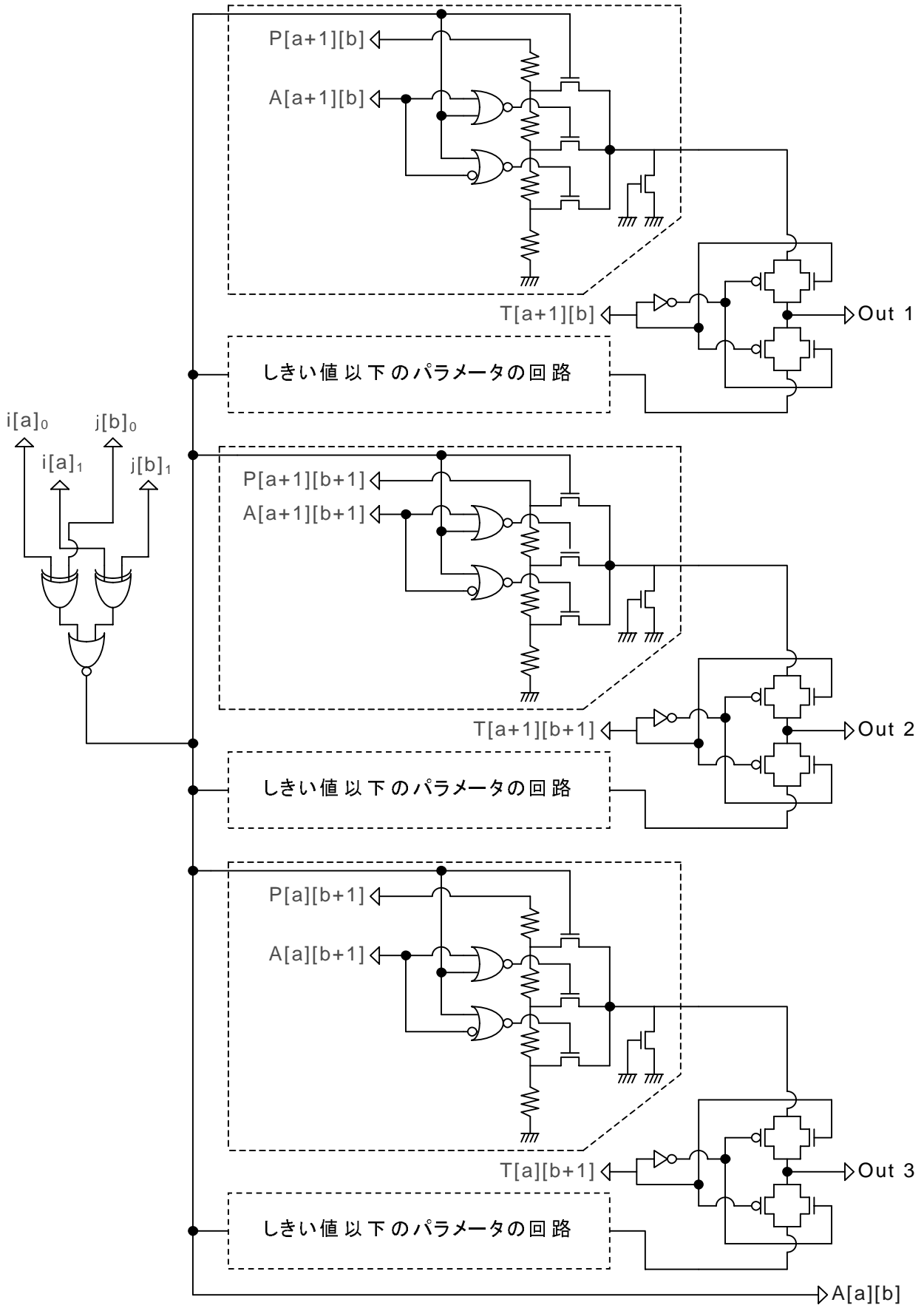


図 4-2 類似度判定回路ノード部(a)

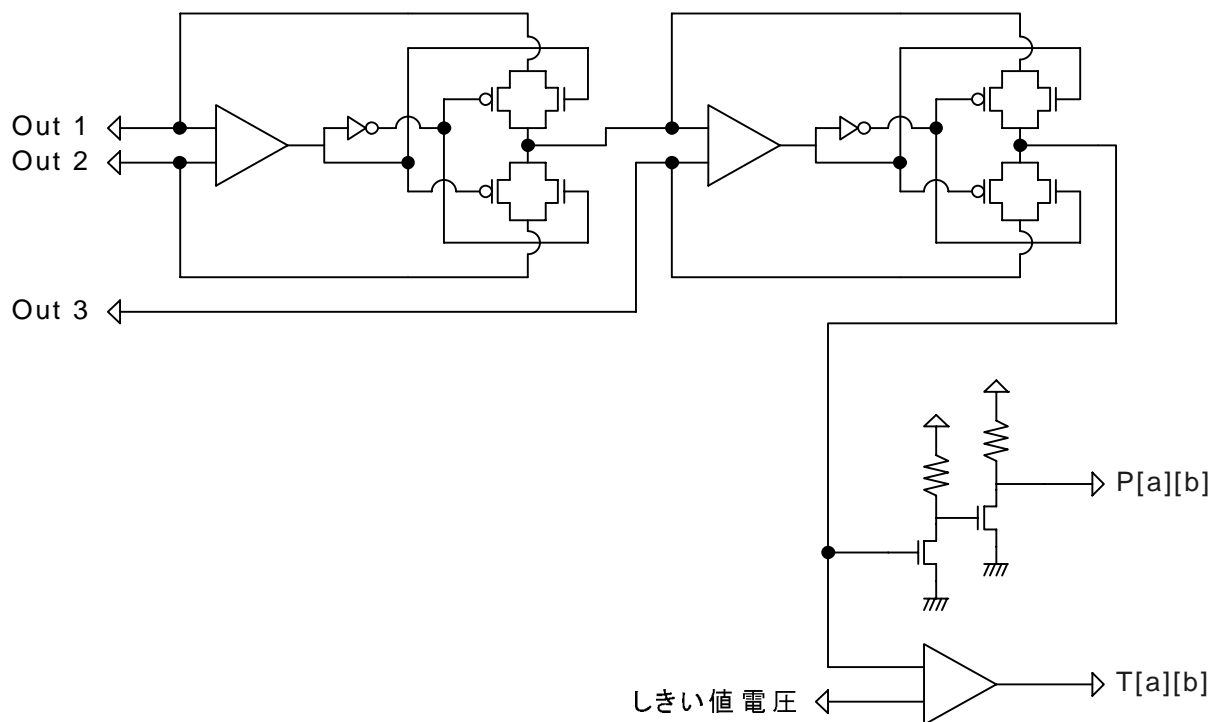


図 4-3 類似度判定回路ノード部(b)

4-2 回路全体

この電子回路で示したノードを行列状に配置、接続し、塩基配列のデータを入力する i_0, i_1, j_0, j_1 を接続した、ノードの行列状配置構造による類似度判定回路が図 4-4 となる。図中のノード間の接続は 3 本の配線を 1 本の太線で示してある。

この類似度判定回路では、塩基配列データを保持するレジスタにシフトレジスタを用いる。そのため、各ノードが塩基配列データ、隣接したノードからの情報を元にデータを出力し、回路全体のノードの出力電圧が安定するのに十分な時間間隔のクロックにより、塩基配列データが 1 塩基分ずつずれて、各ノードに同時に入力される。遺伝子解析では膨大な塩基配列から、ある塩基配列を探し出すということを行うため、片方の塩基配列は一定である。そのため、最初にまず CLK1、CLK2 により二つの塩基配列 i, j を塩基数分のクロック入力し、その後 CLK2 により塩基配列 j が 1 塩基ずつシフトして入力される。つまりクロックが入力されるごとに片方の塩基配列をずらし、二つの塩基配列の類似度を計算、 $P[a][b]$ から出力されることとなるため、高速類似度判定が可能となる。

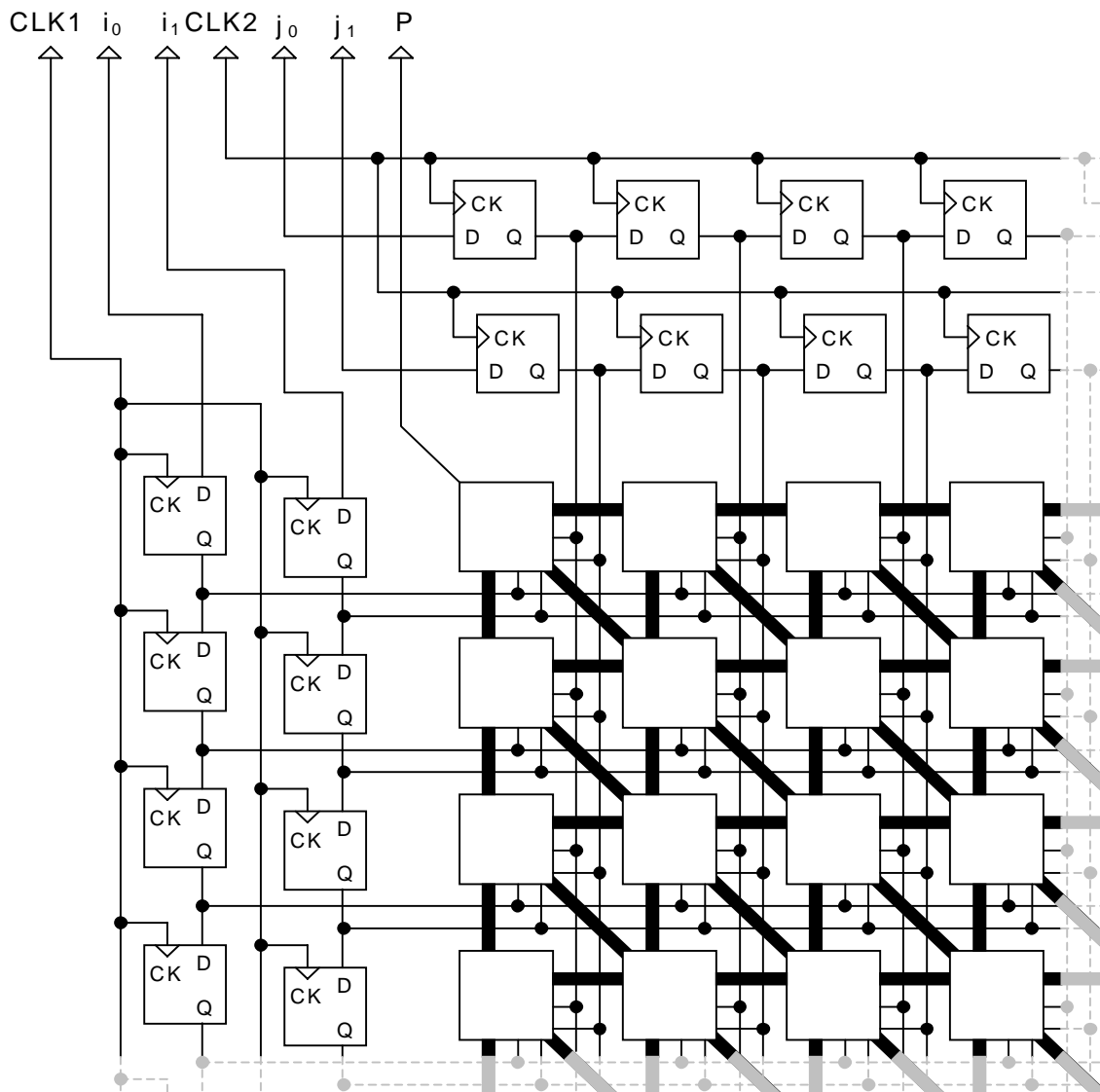


図 4-4 ノードの行列状配置構造による類似度判定回路

4-3 H-SPICE によるシミュレーション

4-3-1 buffer の特性

この回路には、ノードの出力部に他のノードとの相互の影響を取り除くように buffer(緩衝増幅器)がつけてある。この buffer の出力が入力に対して線形でないと、当然ノードが出力する点数も違ったものとなるため、まずこの buffer が入力に対して線形に出力する範囲を調べた。入力に対する buffer の出力結果を図 4-5 に示す。これより buffer による電圧利得がほぼ 1 となる範囲は 2.2V ~ 3.8V であることが分かる。この類似度判定回路では、ノードから出力される電圧が非常に重要であるた

め、出力電圧の誤差を小さくするためにはこの回路が動作する電圧幅が出来るだけ大きくなければならない。そのため 3.8V を利用可能な最大電圧、2.2V を利用可能な最低電圧とした。また、図 4-5 下図のように、3.75V ~ 3.0V では入力電圧に対して出力電圧が若干上昇し、3.0V ~ 2.25V では若干低下している。入力電圧との差は最大で 0.019V となっており、高電圧部と低電圧部とで実際以上に少し差が開くことになるため、類似度の高いものはより高めの点数に、低いものはより低めの点数になる。しかし、元々類似性の評価方法として、類似性の高いものと低いものとの点数の差を大きくするようにしているため、類似度がよりはっきりした傾向になるという影響ですむと思われるため、この誤差は差し支えないと判断した。

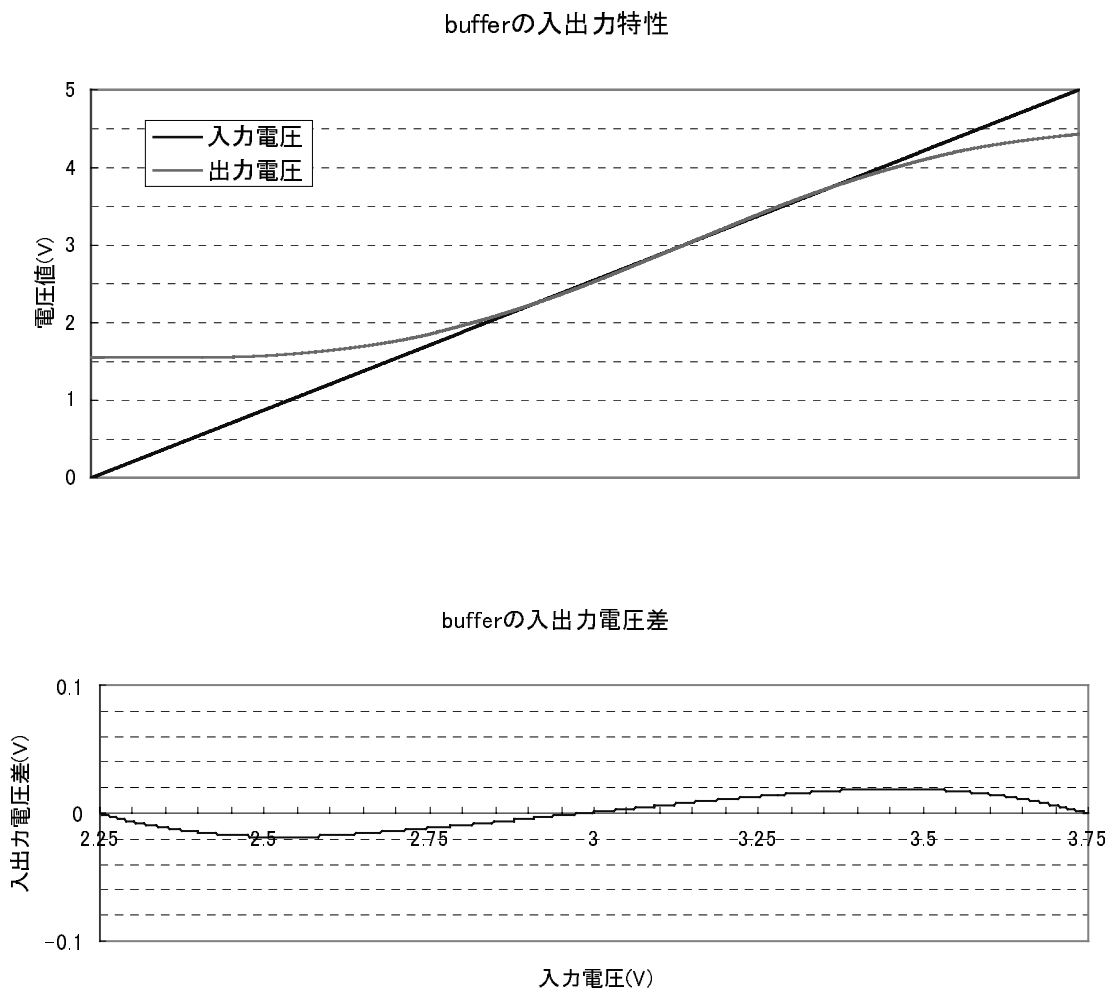


図 4-5 buffer の入出力特性

出力電圧を 3.75V を類似度の 100 点、1.75V を 0 点として評価すると、25 点に当たる電圧は 2.25V となる。前章で述べた C 言語によるシミュレーションでは、最低

値はどれも 25 点以上であったので、実際に出力される電圧は buffer の正常動作範囲内となり、出力が入力に対して大幅に変化することはない。なお、出力電圧の 0.02V が類似度の 1 点に当たる。

4-3-2 高速類似度判定回路の回路シミュレーション

このような特性を持つ buffer を使い、9×9 のノードを接続した類似度判定回路の表 4-1 のパラメータを用いてシミュレーションを行った。図 4-6 に回路シミュレーションの結果を、図 4-7 に同じ塩基配列、同じパラメータにより C 言語でシミュレーションを行ったときの結果を示す。さらに、図 4-8 に二つの結果の差を示す。なお、この例での数字は各ノードの点数を示し、小数点以下第 3 位で四捨五入してある。

	しきい値以上	しきい値以下	しきい値
P-1	0.00	0.00	80 点
P-2	0.30	1.20	
P-3	0.225	0.90	
P-4	0.30	1.20	
P-5	0.50	2.00	
P-6	0.375	1.50	

表 4-1 パラメータ(回路シミュレーション)

	A	G	T	A	G	C	T	A	C
A	85.84	79.16	81.09	86.34	89.11	88.10	91.22	99.79	96.20
C	85.82	84.88	81.28	84.15	85.43	93.69	91.24	92.57	99.83
T	83.15	88.90	89.40	82.80	87.30	88.28	93.15	94.66	96.10
C	84.50	86.25	92.18	88.64	83.71	91.89	91.53	92.57	99.83
A	92.62	83.71	88.64	97.09	88.45	86.09	91.24	96.39	96.10
G	88.38	91.89	86.09	88.45	96.88	88.23	89.27	94.66	96.10
C	89.37	91.53	91.24	89.27	89.27	96.65	89.27	92.57	99.83
T	94.76	92.57	96.39	94.66	92.57	92.57	96.39	92.57	96.10
G	96.20	99.83	96.10	96.10	99.83	96.10	96.10	96.10	96.10

図 4-6 類似度判定回路のシミュレーション結果

	A	G	T	A	G	C	T	A	C
A	79.40	59.26	62.40	80.84	84.80	84.80	89.59	100.00	95.32
C	80.50	79.40	75.57	79.06	80.84	90.74	89.22	90.74	100.00
T	77.62	84.80	85.16	77.96	83.33	84.80	90.74	93.75	95.32
C	79.40	81.86	89.22	85.16	79.40	89.22	89.22	90.74	100.00
A	89.22	79.40	85.16	95.32	85.16	81.94	89.22	95.32	95.32
G	84.80	89.22	81.94	85.16	95.32	85.16	86.29	93.75	95.32
C	86.29	89.22	89.22	86.29	86.29	95.32	86.29	90.74	100.00
T	93.75	90.74	95.32	93.75	90.74	90.74	95.32	90.74	95.32
G	95.32	100.00	95.32	95.32	100.00	95.32	95.32	95.32	95.32

図 4-7 C 言語でのシミュレーション結果

	A	G	T	A	G	C	T	A	C
A	6.44	19.90	18.69	5.50	4.31	3.29	1.63	-0.22	0.88
C	5.32	5.47	5.72	5.09	4.59	2.94	2.02	1.83	-0.17
T	5.53	4.09	4.24	4.84	3.97	3.47	2.40	0.91	0.78
C	5.10	4.39	2.96	3.49	4.31	2.67	2.31	1.83	-0.17
A	3.40	4.31	3.49	1.78	3.29	4.15	2.02	1.07	0.78
G	3.57	2.67	4.15	3.29	1.57	3.07	2.99	0.91	0.78
C	3.09	2.31	2.02	2.99	2.99	1.33	2.99	1.83	-0.17
T	1.01	1.83	1.07	0.91	1.83	1.83	1.07	1.83	0.78
G	0.88	-0.17	0.78	0.78	-0.17	0.78	0.78	0.78	0.78

図 4-8 回路シミュレーションと C 言語でのシミュレーションとの差

これを見ると、ノードの計算が進むにつれて二つのシミュレーションとの差が開いていき、出力される類似度は 6.44 点の開きとなった。また、このシミュレーション結果での最大の差は、19.90 点となっており、とても同じ動作をする回路とは言えない結果となった。

この二つの点数を比べると、大半が回路シミュレーションのほうが高い点数となっていることが分かる。類似度判定回路は C 言語でのシミュレーションに用い

たパラメータから、抵抗値を調整することで同じパラメータとなるように設計しているため、基本的には同じような減点となるはずである。しかしC言語でのシミュレーションと結果が一致せず、このように回路シミュレーションでの出力が上昇してしまうのはbufferの影響ではないかと考え、C言語のシミュレータプログラムをバッファの特性も考慮に入れてもう一度C言語でシミュレーションを行った。そのシミュレーション結果を図4-9に、図4-6の回路シミュレーションとの差を図4-10に示す。

	A	G	T	A	G	C	T	A	C
A	84.92	77.60	79.89	85.11	88.19	87.32	90.64	100.18	95.62
C	84.93	83.98	79.99	83.28	84.16	93.26	90.82	91.78	99.99
T	81.94	88.38	88.91	81.58	86.69	87.39	92.52	94.48	95.62
C	83.40	85.38	91.91	88.02	82.63	91.71	90.90	91.78	99.99
A	92.45	82.63	88.02	97.34	87.85	84.81	90.82	96.13	95.62
G	87.39	91.71	84.81	87.85	97.15	87.37	88.26	94.48	95.62
C	88.26	90.90	90.82	88.26	88.26	96.64	88.26	91.78	99.99
T	94.48	91.78	96.13	94.48	91.78	91.78	96.13	91.78	95.62
G	95.62	99.99	95.62	95.62	99.99	95.62	95.62	95.62	95.62

図 4-9 buffer の特性を考慮に入れた C 言語でのシミュレーション

	A	G	T	A	G	C	T	A	C
A	0.92	1.57	1.20	1.23	0.92	0.78	0.57	-0.39	0.58
C	0.89	0.90	1.30	0.87	1.27	0.42	0.42	0.79	-0.16
T	1.21	0.51	0.49	1.22	0.60	0.89	0.62	0.17	0.47
C	1.10	0.87	0.27	0.62	1.08	0.18	0.64	0.79	-0.16
A	0.16	1.08	0.62	-0.24	0.60	1.29	0.42	0.26	0.47
G	0.99	0.18	1.29	0.60	-0.27	0.86	1.02	0.17	0.47
C	1.12	0.64	0.42	1.02	1.02	0.01	1.02	0.79	-0.16
T	0.28	0.79	0.26	0.17	0.79	0.79	0.26	0.79	0.47
G	0.58	-0.16	0.47	0.47	-0.16	0.47	0.47	0.47	0.47

図 4-10 回路シミュレーションとの差

この buffer の特性を考慮に入れた C 言語でのシミュレーションではほぼ回路シミュレーションと同じような結果となり、出力結果の差は 0.92 点、二つのシミュレーションによる最大の差も 1.57 点と比較的小さくなった。

回路シミュレーションと C 言語でのシミュレーションでの各ノードが出力する点数の差を調べたところ、buffer の特性を考慮しない場合は平均 2.80 点の差であったのに対し、buffer の特性を考慮に入れた場合は平均 0.58 点の差であった。そのため、C 言語でのシミュレーションで buffer の特性を考慮に入れるとより回路シミュレーションと一致する結果となることが分かる。

4-3-3 buffer を考慮したことによる C 言語でのシミュレーション結果の変化

類似度判定回路を設計し、回路シミュレーションを行うことで、C 言語でのシミュレーション結果と食い違う結果がでることが分かった。そのため、回路に組み込んである buffer の特性を、C 言語でのシミュレータにも組み込むことで、出力される差を 1 点以下に抑えることが出来た。これにより、C 言語で行うシミュレーションの結果が点数が上昇する方向へと変化したが、この変化についてさらに考察する。図 4-13 に 9 塩基からなる塩基配列に対し塩基置換が起きた場合の類似度の傾向を示す。

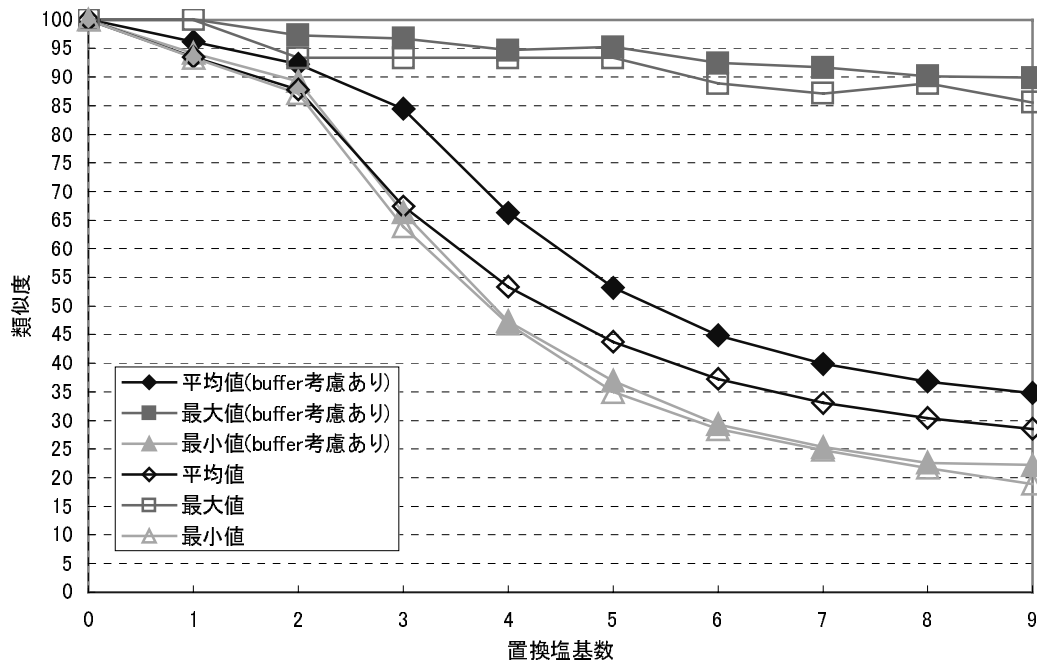
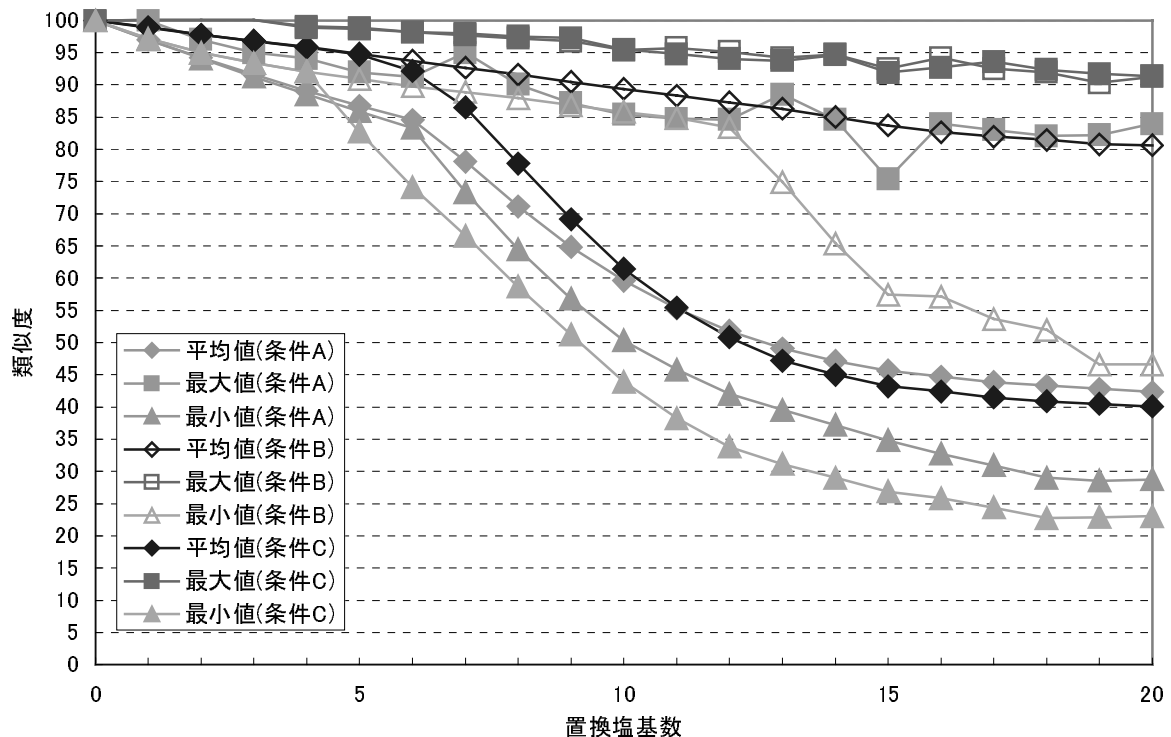


図 4-11 buffer を考慮することによる類似度の傾向の変化(9 塩基)

これを見ると最大値、最小値はそれほど変わらないが、平均値は上昇する方へと変化している。これだけを見ると、傾向としてはそれほど大きな変化はないように思われる。次に、20塩基からなる塩基配列に対し塩基置換が起きた場合の類似度の傾向を3種類の条件で行い、その傾向を見た。図 4-14 に示す。

この3つの条件ではパラメータは表 4-2 の同じものを使っている。まず、条件 A だが、これは3章で示した塩基置換のグラフと同一のものである。この条件 A でのシミュレーションに、buffer を考慮するように変更を加えたものが条件 B となる。条件 B では、高電圧部では若干出力電圧が上昇傾向となる buffer の影響によりあまり電圧が下がらなくなるため、結果として高得点となる。その結果、20塩基全てを置換した場合でもほとんど点数は下がらないため、今までの傾向とは全然違うものとなってしまっている。そこで、条件 B のしきい値 85 点から 93 点へと変更したものが条件 C となる。グラフを見れば分かる通り、しきい値を上げることによって減点量が多くなり、今までのグラフの傾向と近いものとなった。ただし、buffer の影響は受けているため、高得点部ではより高い値となり、低得点部ではより低い値となるというふうになっている。



条件 A: buffer の考慮なし、しきい値 85 点
 条件 B: buffer の考慮あり、しきい値 85 点
 条件 C: buffer の考慮あり、しきい値 93 点

図 4-12 buffer を考慮することによる類似度の傾向の変化(20 塩基)

	しきい値以上	しきい値以下	しきい値
P-1	0.00	0.00	85 点 or 93 点
P-2	0.60	2.40	
P-3	0.45	1.80	
P-4	0.60	2.40	
P-5	1.00	4.00	
P-6	0.75	3.00	

表 4-2 パラメータ(図 4-14)

本研究では、buffer を考慮したときの C 言語でのシミュレーションに合わせて、さらに類似度判定回路の修正をしたときの回路シミュレーションは行わなかったが、すでに C 言語でのシミュレーションと、回路シミュレーションとの間での差はあまりなかったことが分かっているため、おそらくパラメータを変更しても、ほぼ一致する結果となると思われる。

4-3-4 遅延時間

また、この類似度判定回路の動作時間を調べるため、塩基変化が起きた場合の対角線上にあるノードの出力電圧を調べた。まず、塩基配列が全て一致したときのノードの出力電圧を図 4-13 に示す。

これを見ると、一番右下のノードである node99 で多少出力電圧に揺れが見られるほかは、一部に小さな振動が見られるだけで、おおかた安定した出力電圧であることが分かる。この小さな振動も、振動の幅は大きいところで 0.002V 程(0.1 点)しかなく、buffer による誤差等を考えるとほとんど影響がないと思われる。次に、塩基置換により減点された場合の対角線上のノードの出力電圧を図 4-14 に示す。

この図より、全てのノードで電圧が降下していく速さ、つまり電圧降下の傾きはほとんど同じであることが分かる。図 4-15 に示すように、塩基変化(3 種混合)が起きた塩基配列で回路シミュレーションしてみても、傾きはほとんど同じであった。また、傾きがほとんど同じであるため、node11 の出力が安定する時間はほぼ低下した電圧に比例している。おそらく、電圧降下にかかる時間よりもノード間のデータ伝搬遅延が小さいのではないかと考えられる。これにより、一つの塩基配列の類似度を計算し終わるまでにかかる時間は、電圧降下にかかる時間にノード間のデータの伝搬遅延を合わせたものと言える。

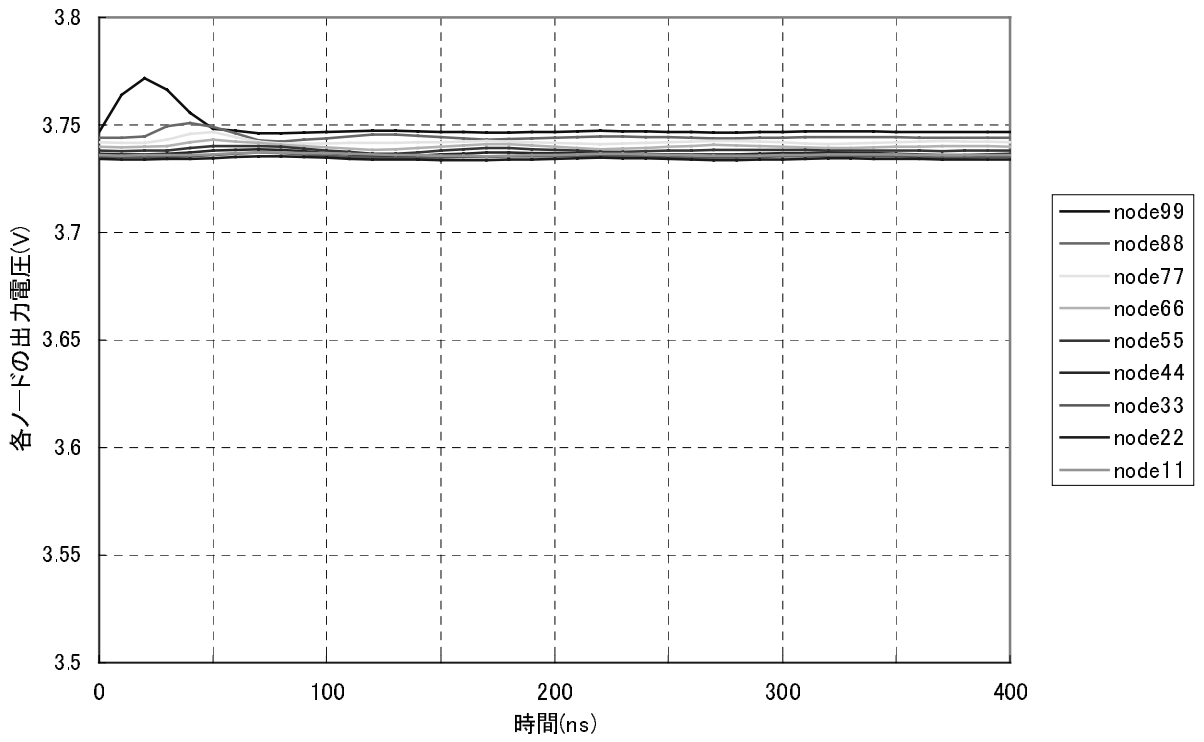


図 4-13 対角線上にあるノードの出力電圧(完全一致)

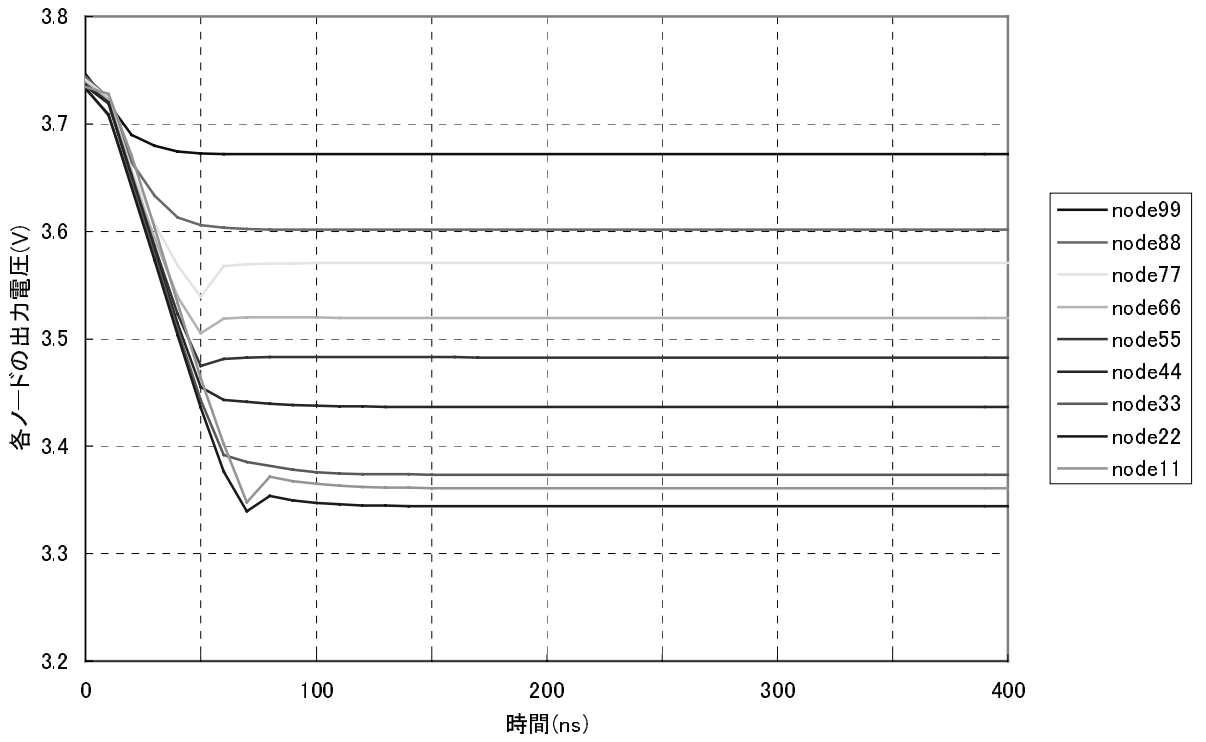


図 4-14 対角線上にあるノードの出力電圧(塩基置換)

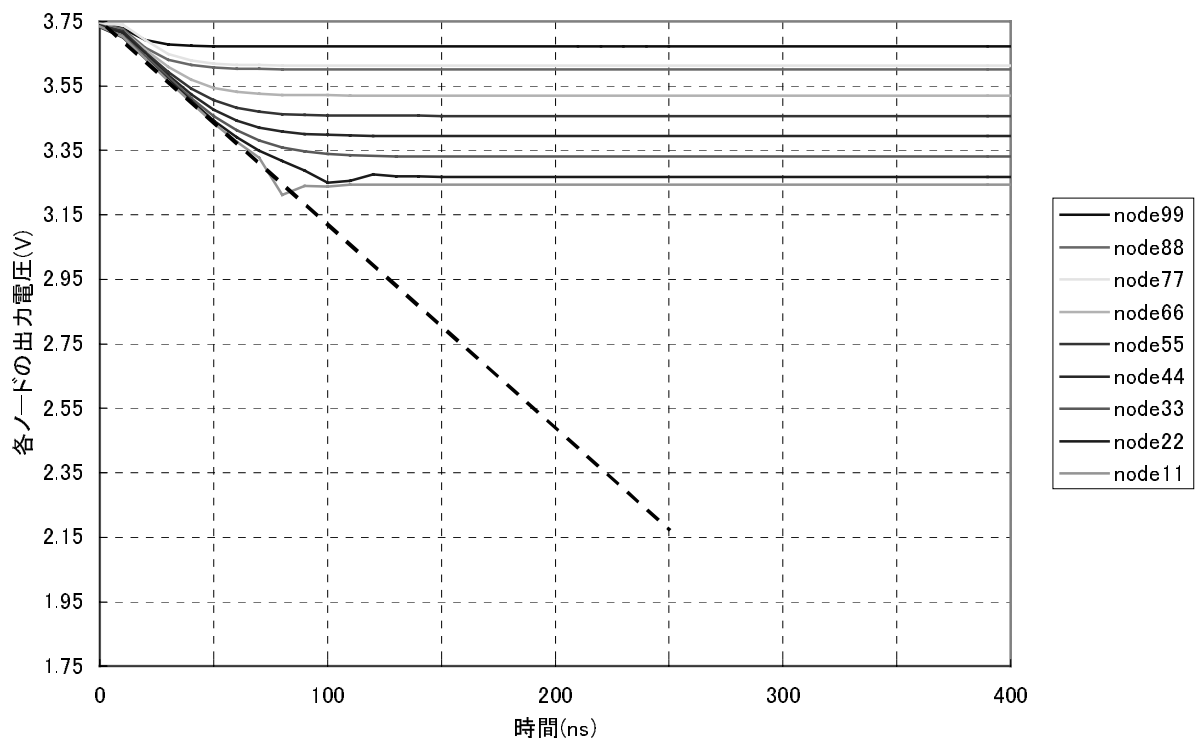


図 4-15 対角線上にあるノードの出力電圧(塩基変化(3種混合))

塩基数が n 塩基の時の演算時間は、「2-3-4 ノードの行列状配置構造による並列処理」で示したように、並列なノードの列数($2n-1$)に比例するため、

$$\text{処理時間} = \text{電圧降下時間} + \text{ノード間の伝搬遅延} \times (2n-1)$$

となる。各ノード間の伝搬遅延を調べたところ、約 5ns であったため、 9×9 の場合にかかるノード間の伝搬遅延は約 85ns だと推定される。電圧降下の傾きが図 4-15 に点線で示すような傾きだとすると、利用可能な最低電圧である 2.25V まで低下するには約 230ns かかる。ノード間の伝搬遅延が約 85ns であることから、実際の電圧降下にかかる時間は約 145ns であると考えられる。

100 塩基からなる塩基配列を比較した場合、これらの数字から 1 回の類似判定に約 $1.15\ \mu\text{s}$ かかる。これは類似度判定回路が約 0.87MHz で動作させることが出来るということで、1 秒間に約 87 万塩基対の類似度判定が出来る計算となる。

第5章 結論

本章では、ノードの行列状配置構造による文字列間の高速類似度判定回路についてまとめ、今後の展望について述べる。

5-1 本研究のまとめ

本研究では、遺伝子の突然変異による塩基列が変化したものの類似検索というところにヒントを得、文字列間の類似性をある一定の点数で評価し、高速で出力できる回路を設計することを目的とした。

5-1-1 遺伝子解析におけるホモロジー検索

- 遺伝子の突然変異には塩基置換、塩基欠落、塩基混入の 3 つのパターンがあり、変化する前の塩基配列と遺伝子の突然変異が起きたものとの表を用いた塩基配列の一致比較を行うと、3 つのパターンとも、対角線と平行なセルに一致を示す「」が並ぶことが分かった。
- その表を用いた塩基配列の一致比較を元に、「」が並ぶ位置が対角線から離れていくと類似性も下がっていくという事を利用し、3 方向からの入力のある一定の規則に従って計算し、そのうち最大のものを選択するという「ノード」を行列状に配置することによって、二つの塩基列の類似性を類似度という点数として判定する、ノードの行列状配置構造による文字列間の類似度判定法を提案した。

5-1-2 ノードの行列状配置構造による類似度判定法の評価

- C 言語により提案した類似度判定法を実現できるプログラムを作成し、塩基配列を自動で生成させることにより、膨大な塩基配列パターンでのシミュレーションが行えるようになった。
- 塩基置換、塩基欠落、塩基混入、塩基変化(3 種混合)、連続塩基置換、連続塩基欠落、連続塩基混入の 7 つの場合において、塩基配列生成方法を述べ、どのような塩基配列の場合にどのような類似度となるかを調べた。

- しきい値という適当な電圧を境に減点量を変化させることによって、出力される類似度は、全体に対して変化した塩基が少ない、または多い場合は、1塩基あたりの類似度の減少量が少なく、全体に対して半数ほどの塩基が変化している場合は減少量が多いといった傾向となった。
- 塩基配列に含まれる塩基数を変化させ手シミュレーションを行うことにより、多少違いがあるものの、どのような塩基数でも同様な結果となることを確認した。
- 7種類の塩基配列生成パターンでの違いを比較することで、それぞれの違いには根拠があることが分かり、それを踏まえると出力される類似度は、ほぼ同じような傾向を示していることが分かった。

5-1-3 高速類似度判定回路の構成

- C言語でのシミュレーション結果を基に、文字列間の類似判定を行う高速類似度判定回路の設計を行った。
- 回路の設計にあたり、各ノード間の相互の影響を防ぐため、bufferを各ノードの出力端につけ、その特性を回路シミュレーションで調べた。
- 設計した高速類似度判定回路に電圧を与えて回路シミュレーションを行うことにより、出力される類似度の傾向を調べた。その結果、bufferの影響を受けるため、C言語によるシミュレーションよりも全体的に高い値となった。
- 設計した高速類似度判定回路のbufferが与える影響をC言語でのシミュレーションにも組み込むことにより、ほぼ一致する出力となることが分かった。
- 類似度判定回路のノード全てが安定した出力電圧となるまでの時間は、低下する電圧量と、接続したノード数に比例していることが分かった。 n 塩基からなる塩基配列の場合の演算時間 T は、

$$T = 145 + 5 \times (2n-1) \quad [\text{ns}]$$

となる。

5-2 今後の展望

本研究の今後の展望として、以下のようなことが挙げられる。

- C言語でのシミュレーションと、回路シミュレーションでの誤差をさらに小

さいものになるように修正を加える。

- 今回の回路設計では 9×9 の大きさの類似度判定回路を設計したが、さらに大きな類似度判定回路を設計することで、提案した類似度判定法が大きな回路でも問題なく動作することを確認する。
- buffer を考慮した場合の C 言語でのシミュレーションで適切なパラメータを探し、そのパラメータに応じた類似度判定回路を設計し、シミュレーションを行う。
- レイアウト設計を行い、実際の類似度判定回路を作成する。

我々が提案したノードの行列状配置構造による文字列間の高速類似度判定回路は、DNA 解析における類似塩基配列検索にヒントを得て、研究を始めた。

しかし、この類似度判定回路は DNA 解析だけにしか利用できないわけではなく、入りに 4 値のデータ、つまり 2bit のデータを入力して類似度判定を行っているため、その入力ビット数を変えることによって、様々なデータ列の類似度判定が出来ると考えられる。

そのため、パラメータなどをうまく設定することで、似た文章を探し出すといった単語検索ではない文字列検索などに利用できる可能性がある。

謝辞

本研究を行うにあたり、多くの方々の御助言、御協力を頂きました。この場を借りて感謝の意を表したいと思います。

修士からこの研究室で新たに研究することとなり、二年間にわたりさまざまな面での御助言、御指導をして頂いた金沢大学教授 畑朋延先生に心から感謝致します。

また、様々な面での御助言、御指導をして頂き、集積回路分野の自由な研究をさせて下さった佐々木公洋助教授に深く感謝いたします。

WSの管理、生活面、研究室行事などでお世話になったなどをして頂いた吉田行男助手に感謝いたします。

大学4年生の時に大変お世話になり、他研究室へと移ったにもかかわらず、集積回路分野での御助言、御指導をしていただいた金沢大学集積回路工学研究室教授 故鈴木正國先生に深く感謝すると共に、心から御冥福をお祈りします。

大学4年生の時に大変お世話になり、大学院進学後も御助言、御指導、VDEC関連等で大変お世話になった集積回路工学研究室 北川章夫 助教授に感謝します。

本研究及び学生生活において御指導、御助言、御協力下さった集積回路工学研究室 秋田純一助手に深く感謝いたします。

金沢大学に入学して以来、幅広い専門分野について基礎から教えて頂いた電気・情報工学科の先生方に深く感謝致します。

時には優しく、時には厳しく様々なことを教えて頂いた博士課程後期の一番ヶ瀬剛氏、辻口達也氏、土岐和之氏、長嶋満宏氏、宮下均氏、吉野幸夫氏、金濟徳氏に感謝します。

この2年間、共に学び、共に笑い、様々なことを教えてくれ、苦楽を共にした修士課程二年の清水直樹君、二ツ寺政友君、集積回路工学研究室 小川明宏君、高瀬信二君、中橋憲彦君、早川史人君に感謝いたします。

本研究に共に取り組み、苦楽を共にしてくれた集積回路工学研究室 佐々木勝光君に深く感謝します。

同じ研究室で過ごし、研究、日常生活面において楽しい思い出を作ってくれた博士課程前期一年、佐々木健次君、高橋幸大君、1999年度卒業研究生 上田香織さん、皆森

雅文君、金田亮君、上明戸順一郎君、近田晴彦君、丹後智之君、中尾政史君、森田晴紀君、和田芳信君、1998年度博士課程前期卒業生の川越進也さん、鍋谷定孝さん、増田豊君、集積回路工学研究室 田口和彦君、夏目雅弘君、前多和洋君、水橋嘉章君 1998年度卒業研究生 スクリーンハナさん、四柳毅君にも感謝いたします。

集積回路工学研究室の研究生である、博士課程前期一年今井豊君、数馬晋吾君、藤田隼人君、水野浩樹君、渡辺晃君、1999年度卒業研究生 笠井稔彦君、大門慎治君、高松直樹君、辻川隆俊君、遠山治君、中村公亮君、蓮達弘君、水木誠君、通信工学研究室博士課程前期一年村上崇君に感謝します。

最後になりましたが、私の学生生活をあらゆる面で支えてくださった両親、兄弟、そして、そのほかたくさんの方々に心から感謝いたします。

参考文献

- 1) 黒田行昭著「基礎遺伝学」(近代遺伝学の流れ) 裳華房(1995)
- 2) 宝来聡著「DNA 人類進化学」岩波科学ライブラリー52
- 3) 国立遺伝子研究所「電子博物館」第1版(1999)
- 4) 文部省科学研究費補助金重点領域研究「ゲノムサイエンス」
「21世紀のキーワード「ゲノムサイエンス」を知っていますか」

口頭発表

藤井直樹、中村真樹、前多和洋、鈴木正國、北川章夫

EDA による HDL を用いた磁気浮上制御チップの設計と試作

平成 10 年度電気関係学会北陸支部連合大会

会場 福井工業大学