論　文

ヒトゲノム解析アクセラレータLSIアーキテクチャ

佐々木勝光[†]　秋田　純一[††]　深山　正幸[†]　吉本　雅彦[†]

A Fast Homology Search LSI for Human Genome Analysis

Masamitu SASAKI[†], Junichi AKITA[††], Masayuki MIYAMA[†], and Masahiko YOSHIMOTO[†]

あらまし　ゲノム情報解析の分野において比較処理が重要な要素となっているが，現在行われているソフトウェアを用いた並列計算機による並列処理では非常にコストがかかる．本研究はこの問題を改善するために，比較処理の並列性を生かした高速類似度判定アルゴリズムを提案し，それに基づいたLSIアーキテクチャを提案する．本アーキテクチャの特徴は，類似を判定し類似度を計算する演算要素をアナログ回路で構成し，高速化のため，この演算要素をマトリックス状に配置して並列処理を行う点である．VDEC0.6 μm 3LMプロセスを用いることで，24行24列のマトリックス演算回路（55kTr）を4.5×4.5 mm^{2}のLSI上に実装した．その結果，1秒間に最大約480万塩基列を対象とした類似度判定が実現できる見通しを得た．

キーワード　ゲノム，ホモロジー検索，高速類似度判定法，ダイナミックプログラミング（DP）法，マトリックスアーキテクチャ

1. まえがき

近年ヒトゲノムの塩基配列が決定され，次のステップとしてポストシーケンスと称して特定のタンパク質を発生させる塩基配列の解析が重要視されている．ゲノム情報解析が進むことにより癌（がん）の克服や先天性異常の回避等，画期的な医学の進歩が期待される．このような異常をきたす原因は，遺伝子を構成する塩基の突然変異による塩基配列の変化である．突然変異とは何らかの原因で塩基が突然他の塩基に置き換わったり，消えたり，現れたりする変化のことである．ゆえに類似した塩基配列が生成する遺伝情報の解析や，間違った遺伝情報を生成する塩基配列の特定が重要視されている．このような解析や特定を行う際には，類似性を求めるための比較処理（ホモロジー検索）が頻繁に行われ，重要な処理となっている．

† 金沢大学工学部電気・情報工学科，金沢市
Department of Electrical and Computer Engineering, Kanazawa University, 2-40-20 Kodatsuno, Kanazawa-shi, 920-8667 Japan
†† 公立はこだて未来大学システム情報科学部，函館市
Department of Media Architecture, Future University - Hakodate, 116-2 Kametanakanomachi, Hakodate-shi, 041-8655 Japan

通常はシーケンサと呼ばれる解読器で解読した数百塩基程度のDNA断片の塩基配列を大型並列コンピュータに与え，ソフトウェアによって機能解析などの様々な情報処理を行うが，一般にゲノムの塩基配列の情報量は膨大であるため，計算処理時間が問題となることが多い．ゲノムの塩基配列の情報量は，例えばヒトの全ゲノムの場合で約30億塩基対，すなわち6Gビットと膨大なものとなる．各種高速アルゴリズムが検討されているが劇的な高速化は望みにくく，また並列計算機の導入は処理速度向上にはいくらか有効であるが非常に高価であるために普及しにくいという問題点がある．

本研究では，ゲノム情報解析に重要なホモロジー検索を主な対象とし，その処理速度の劇的な向上のために高度な2次元並列処理構成をとる専用の大規模集積回路（VLSI）のアーキテクチャを検討し，スーパコンピュータよりもはるかに小型で低コストな検索システムを実現する見通しが得られたので，その結果を報告する．

2. アルゴリズム

2.1 塩基配列の比較処理（ホモロジー検索）

基本的にゲノムはA（アラニン）T（チミン）C（シトシン）G（グアニン）の4塩基からなっている．ま

た相補性によりAとT，CとGが必ず対の関係にある．ここで，両親から子へと親情報が遺伝する際や細胞分裂の際の外的要因により，ある塩基が他の塩基に突然置き換わる（置換），欠失する（欠落），付加する（混入），といった突然変異が発生することがある．これにより塩基配列に変化が生じて遺伝情報が変化し，奇形などの先天性異常や癌などの後天性異常を引き起こす．先述のとおり，塩基の突然変異には置換，欠落，混入の3種類がある．図1には一例として(a)正常配列，(b)正常配列から塩基対A-TがC-Gに置き換わった塩基置換，(c)塩基対A-Tが欠失した塩基欠落，(d)塩基対C-Gが付け加わった塩基混入を示した．ある患者の疾患原因が塩基配列の突然変異によるものかどうかを検査するときに，正常配列と比較を行い配列異常箇所を探し出すことができるし，その配列異常がどのような経緯で発症したのかを生物種間で比較することで調査することができる．更にその疾患がどのような配列異常で発症するか解明されれば遺伝子レベルで予防に努めることができる．

2.2 表を用いた比較

そこで先に述べたように比較処理が重要視されるわけである．我々は比較処理を高速に行うために，図2のように表を用いて二つの塩基配列A，Bを配置することにより，各塩基の一致，不一致を並列に比較することができるようにした．ここで，各セルを演算部とする．図2には(a)完全に一致する配列Bとの比較，(b)塩基置換が発生している配列Cとの比較，(c)塩基欠落が発生している配列Dとの比較，(d)塩基混入が発生している配列Eとの比較の四つの例を示した．図2(a)より，完全に一致している場合は対角線上に一致を示す○が並ぶことがわかる．図2(b)より，塩基置換が発生している場合は対角線上の不一致な部分が×となって，対角線上に並んでいた○の列が対角線と平行に左上へとずれていることがわかる．同様に図2(c),(d)より塩基欠落が起きた場合は，対角線上に並んでいた○の列が対角線と平行に下へずれ，塩基混入が起きた場合は右へとずれることがわかる．このことから一致が続いた場合は対角線と平行に，つまり左上へと信頼度（類似度）が伝達され，不一致な部分では左上，上，左へとずれが生じ，信頼度（類似度）が低下することがわかる．ゆえに下辺と右辺に初期点を与え，一致が続く場合は減点せずに左上へ点数を伝達し，塩基変化（ずれ）が起きて信頼度（類似度）が低下する箇所で減点していくことで，配列全体の一致度に応じた点数を求めることができると考えた．

また，塩基置換が発生した場合は全体の一部分のみが変化しており，塩基欠落，塩基混入が発生した場合は一部分が変化して，それ以降の塩基がずれていると考えることができる．そのため，もと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できる．また，数十%程度の変化であれば少し類似している，もと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ように認識することができる．そこで本研究ではその認識を類似度として表現することにより，次に示すアルゴリズムを用いることによって，ハードウェア的に比較処理（ホモロジー検索）を実行させる．

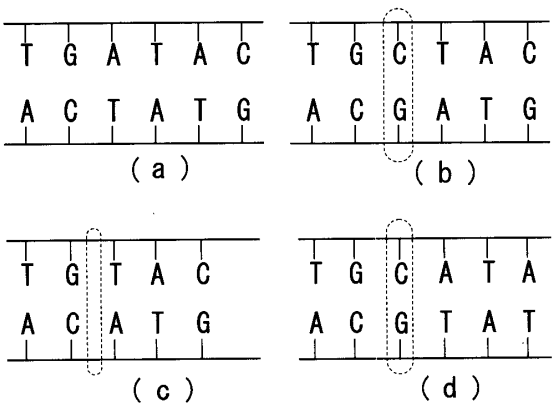


図1 突然変異
Fig. 1 Mutation.

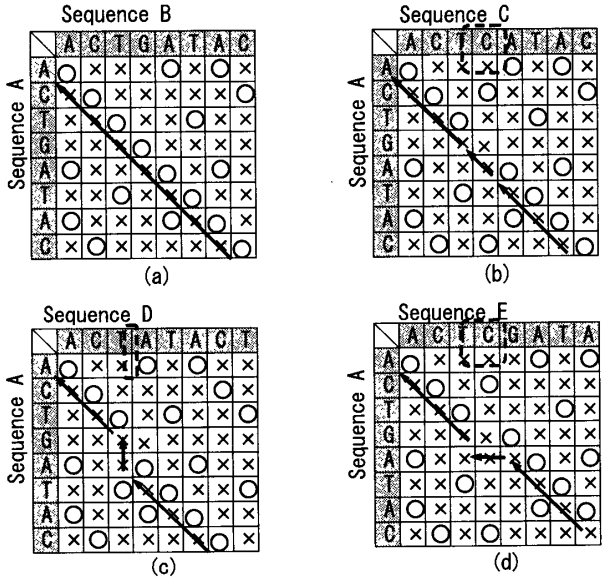


図2 類似度判定アルゴリズム
Fig. 2 Fast homology search algorithm.

2.3 点数処理

詳細な一致度を求めるために，減点の際に以下のようなアルゴリズムを用いることとした．

次の減点法のアルゴリズムにより下，右下，右方向の3方向から点数を受け取り，減点処理を行い，一番高い点数をその演算部の点数とし上，左上，左へと出力する．

・類似度を算出しようとしている演算部が「○」の場合には，入力元（右，右下，下）の○×にかかわらず，右下からの点数はそのまま受け取り，下，右からの点数は減点して受け取る．

・類似度を算出しようとしている演算部が「×」の場合には，入力元の○×に応じた減点量を設定し，右，右下，下からの点数を減点して受け取る．自身が「○」の場合と比べて減点量を増やす．

右下からの点数の減点量は，右，下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．

点数計算において演算部自身の一致，不一致と入力元の一致，不一致の組合せとして図3に示すように8通りが考えられ，この8通りの組合せを先に示した減点法のアルゴリズムに基づいて分けると，P-1～P-6の六つのパターンに分けられる．

ここで，3方向からのデータをそれぞれそのまま，若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理由として，入力される点数はそこに到達するまでの一致情報により減点がなされており，高い点数というのは一致しているノードの数が多いということを示している．そこで，ほかの2方向のノードからの点数より高い点数であるということは，そのルートの方がより一致している，つまり信頼度が高いと考えられるからである．

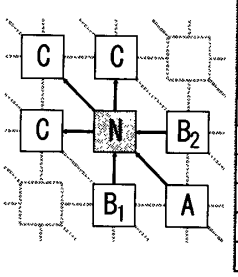
また右，下からの点数よりも右下からの点数を優先し，減点量を比較的少なくする理由として，塩基欠落，塩基混入して塩基配列がずれた場合には，この「○」の列が対角線に対して平行にずれる．これは，上方向若しくは左方向からくる点数は，明らかに塩基変化が起きたルートを通った点数であると考えられるからである．

上記減点アルゴリズムを考慮して，最右列と最下行に初期点を与え，塩基変化（ずれ）が起きた箇所で減点していくことで，配列全体の一致度に応じた点数を求めることができる．

2.4 塩基の変化による一致度の点数化

図4に類似度を求める際の具体例を示す．ある塩基配列とそれから置換，欠落，混入が1塩基分ずつ生じた塩基配列とを比較した例を示す．我々が提案する判定法をもとに類似度を求めたものが図4(b)である．図4では簡単のために点数の流れのみを表示している．まず最初に右下のセルに初期点として100を与える．このセルが不一致ならば100から減点した点数を初期点として与える．点数の流れを見ていくと，混入が発生した部分で左へとずれて100から90へと減点されている．しかし次の段では左上へと一致が続いているので点数はそのまま伝達されている．同様に欠落，混入が発生した部分では81，77へと減点され，77が最終的に出力されている．

このように，提案する判定法は前のセルの一致情報を考慮に入れて点数を決定していくため，図4(b)で最終的に左上端に出力される「77」という点数は，二つの塩基配列全体の一致情報を含んだ点数であると考えることができる．我々はこの点数を二つの塩基配列の類似度とした．



入力元	N	parameter
A :○	○	P-1
A :×	○	P-1
A :○	×	P-2
A :×	×	P-3
B1, B2:○	○	P-4
B1, B2:×	○	P-4
B1, B2:○	×	P-5
B1, B2:×	×	P-6

図3 減点パラメータ
Fig. 3 Deduction parameter.

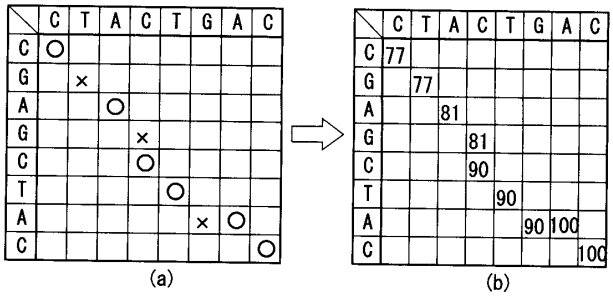


図4 点数化
Fig. 4 Turn into score.

3. ヒトゲノム解析アクセラレータLSIアーキテクチャ

3.1 演算要素PE

上記のアルゴリズムに基づいて図5のようにLSIアーキテクチャを考案した．特徴の一つ目としては，「○」「×」の類似判定及び点数処理（減点処理）を行う部分を演算要素PEに置き換え，アナログ回路で構成した点である．類似度（点数）は電圧を用いて表現することとした．演算要素PEの構成を示したブロック図を図6，論理回路図を図7に示す．図6に示されているようにPEの入力は，そのPEに対応する塩基データ（A，T，C，Gを2bitで示したもの）と，右，右下，下のPEでの点数（電圧値）と一致情報（1bit）を入力としている．その理由としては先のアルゴリズムで説明したように，塩基の置換，欠落，混入の突然変異により一致を示す○の列が3方向へずれることから，各PEでの点数を決定する際の要素として右，右下，下のPEでの点数と一致情報を用いることでパラメータによる点数処理が可能となり，より詳細な類似度判定を行うことができると考えたからである．

処理の流れとしては，まず一致比較部でそのPEに対応する塩基データ（inx1，inx2，iny1，iny2）の一致情報を算出する．得られた一致情報と右，右下，下のPEから入力された一致情報（ina from R，ina from LowerR，ina from Lower）とを用いて，三つの減点部において，入力された三つの点数（inp from R，inp from LowerR，inp from Lower）それぞれに対して減点するかどうかを決定し，点数処理を行う．更に減点部で処理された点数のうちで最高点を点数比較部で選択し，そのPEでの一致情報と最高点を上，左上，左の次のPEへと出力する．

図7に示されているように，一致比較部は二つのEXORと一つのNORで構成されている．減点部は二つのNORと抵抗素子，及びNMOSスイッチで構成されており，どの割合で減点処理を行うかを一致情報を用いて選択できるようになっている．ここで抵抗比により電圧を下げることで減点処理を行っている．点数比較部はコンパレータとCMOSスイッチからなる回路の2段構成となっている．

3.2 マトリックス状配置

二つ目の特徴は高速化を目的に，演算要素PEをマトリックス状に配置し，その際に3方向のずれを考慮した接続となっている点である．図8のようにPEをマトリックス状に配置することにより並列に比較処理

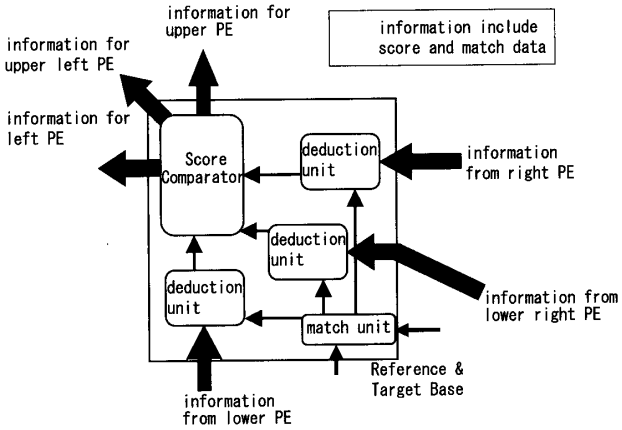


図6 演算要素PEのブロック図
Fig. 6 Block diagram of processor element.

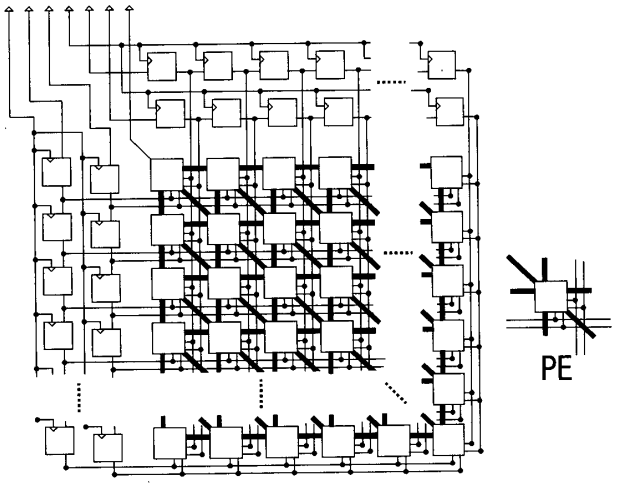


図5 ヒトゲノム解析アクセラレータLSIアーキテクチャ
Fig. 5 A fast homology search LSI architecture for human genome analysis.

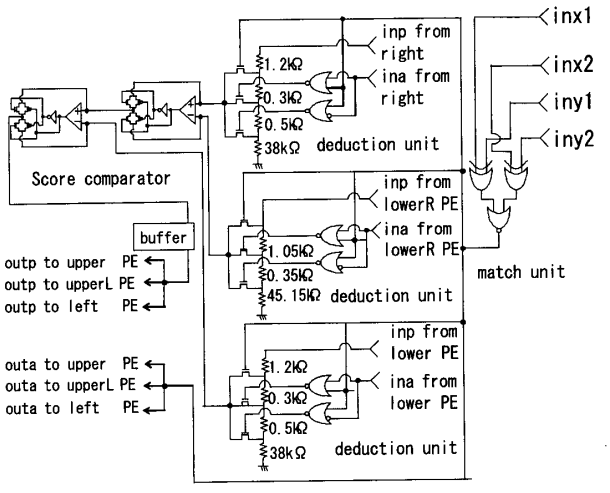


図7 演算要素PEの論理回路図
Fig. 7 Logic circuit diagram of processor element.

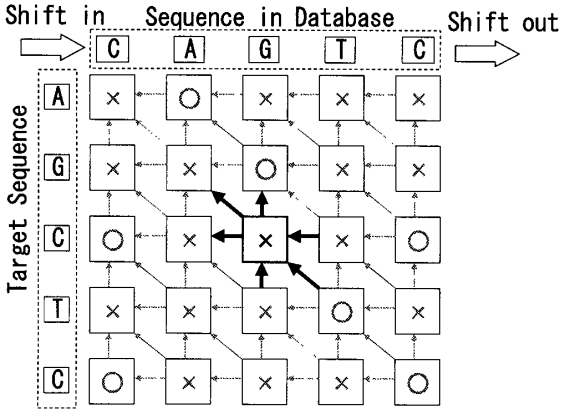


図8　5行5列配置
Fig. 8　5×5 matrix circuit.

を行うことができるため，高速に類似度を求めることができる．更に各PEを3入力3出力として接続することで，アルゴリズムにある3方向のずれに対応している．ここで，塩基配列データベースを「Sequence in Database」，比較対象配列を「Target Sequence」と呼ぶことにした．図8では「Sequence in Database」だけがシフトレジスタを用いてシフトインするように示してあるが，実際は図5に示されているように「Target Sequence」側にもシフトレジスタが用意されており，配列の長さに応じて順次シフト入力し，その都度類似度を算出できる構成となっている．

3.3　信号線の不使用

三つ目の特徴は，信号伝搬時間の制約となる大域的な信号線を使用していない点である．ゆえにいったん塩基データが入力されれば，即座に各PEで処理が並列に行われるため，高速に全体の類似度が出力される構成となっている．

4.　処理能力評価

4.1　シミュレーション条件

本研究で提案するヒトゲノム解析アクセラレータLSIアーキテクチャの処理能力を評価した．演算要素PEを24行24列に回路を構成し，シミュレーションを行った．図9に示されるような「Target Sequenceと比較してSequence in Databaseの最右端の塩基のみが一致している状態から，Tがシフトイン，Aがシフトアウトして完全に不一致な状態になる」という条件のもと，対角線上にあるPEの出力が安定するまでの時間をSynopsys社製回路シミュレータ「PowerMill」を用いたシミュレーションにより求めた．

すべてが不一致であるため，最初に右下端のPEで

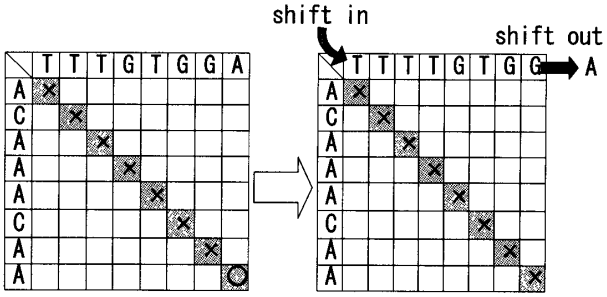


図9　シミュレーション条件
Fig. 9　Simulation condition.

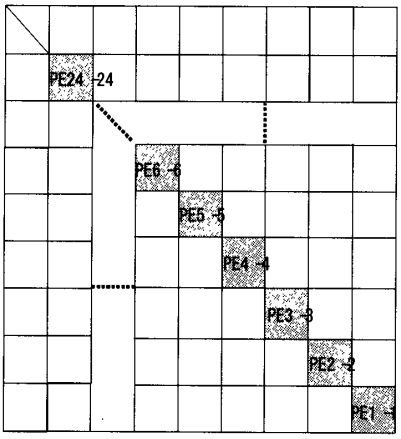


図10　ラ　ベ　ル
Fig. 10　Label.

減点処理された点数が左上のPEへと減点されながら伝達され，最終的に左上端のPEに類似度が出力されるという流れでシミュレーションを行った．点数を電圧で表現しているという本回路の特性上，減点には電圧降下を伴うので，処理に最も時間がかかるケースである．最右列，最下行PEに与える初期点となる電圧は3.7V，電源電圧は5Vである．また，対角線上のPEに図10のようなラベル付けを行った．

4.2　シミュレーション結果

図11は上記のシミュレーション条件でのシミュレーション結果をもとに対角線上の各PEが収束するまでの時間をプロットしたものである．PE2-2が収束してからPE24-24が収束するまでの時間を最小2乗法で求めて，それをPE間数22で割ったものを，PEとPEの間を伝達する時間Tとし，定数をCとすると演算処理時間Aは式(1)のように考えられる．ここで，塩基データの入力がいったん終了した状態が初期状態であり，次の塩基データのシフトインが完了してから各PEでの電圧が収束するまでの時間が演算処理時間である．

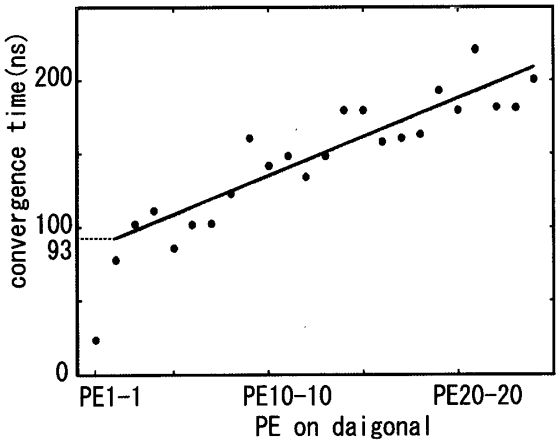


図11 シミュレーション結果
Fig. 11 Simulation result.

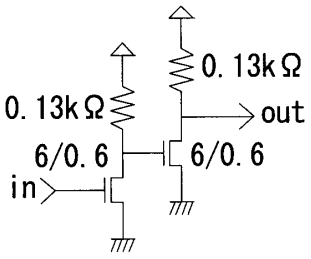


図12 バッファの回路図
Fig. 12 Circuit diagram of buffer.

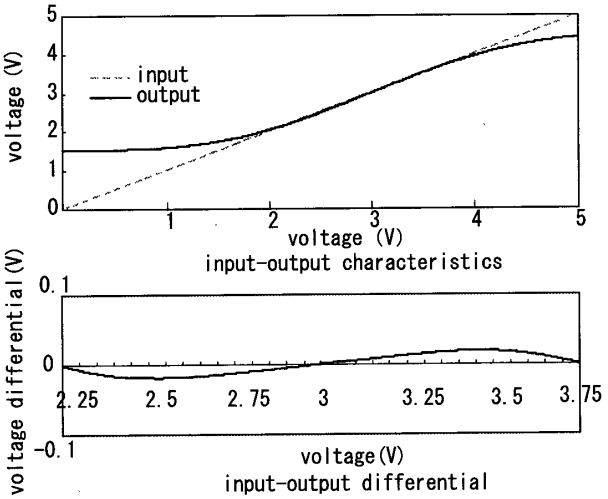


図13 バッファの入出力特性
Fig. 13 Input-output characteristics.

$$A = C + T(n-2) \tag{1}$$

$(n-2)$ はPE2-2から最左上端PEn-nまでのPE間数である．図11から式(2)が得られる．

$$A = 93 + 5.3(n-2) \tag{2}$$

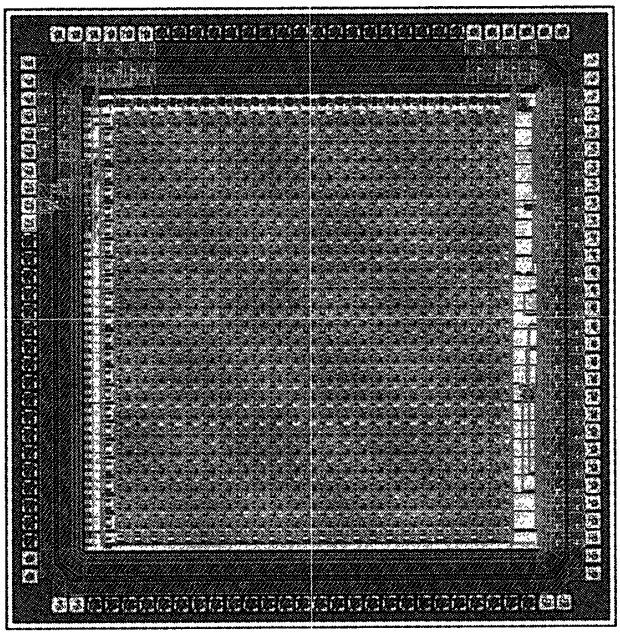


図14 プロットイメージ
Fig. 14 Plot image.

□プロセス	0.6 μm 3層金属配線
□トランジスタ数	55104個
□チップサイズ	4.5×4.5[mm^2]
□規模	24行24列

図15 チップ諸元
Fig. 15 Chip parameter.

ここでシミュレーション結果のばらつきについて考察する．

このばらつきの原因として，バッファが挙げられる．バッファ回路を図12に，バッファの入出力特性を図13に示す．この入出力特性による最大誤差は0.02 Vとなり，これは類似度で1点に相当する．このバッファの特性により各PEでの電圧収束時間にばらつきが生じており，これはリニアリティの良いバッファに置き換えることで改善できる．

ここで式(2)にn=24を代入すると，210 ns (n=24)という値を得ることができる．

以上の結果から，本回路は24塩基からなる塩基配列を比較した場合，約0.21 μsで演算を行うため，塩基列をシフト入力した場合は1秒間に約480万塩基からなる配列を比較する事が可能であることを見積もった．

CMOS0.6 μm・3層金属配線のプロセスを用いて，4.5 mm角のチップ上で本回路のレイアウトを行った．回路規模としては約5万5千トランジスタを集積し，演算要素PEを24行24列に配置した．使用した抵抗は2層目のポリシリコンを用いた高抵抗を用いて作成

した．消費電力は約5Wと見積もられる．図14と図15にプロットイメージとチップ諸元を示す．

今回設計したバッファ回路の特性上，処理が1段進むごとに，一致していても20mV程度の電圧効果が発生することと，バッファの正常動作電圧帯が約1.7V（図13）であることから，最大で85段の並列処理まで正常に行うことができる．ゆえに，（並列処理段数）＝（$2n$列（行）-1）から，現在のバッファを用いれば最大43行43列の規模まで拡大することができる．

ここでリニアリティの良いバッファを用いてチップサイズを拡大して，PEを100行100列に配置した場合を考える．式(2)にn=100を代入すると，612ns（n=100）という値を得ることができる．この結果から，チップサイズを拡大しPEを100行100列に配置した場合に約0.6μsで演算が可能であることがわかる．ゆえに，1秒間に約160万塩基からなる配列を比較することが可能であると見積もることができる．

5. むすび

我々は，ポストシーケンスの分野で重要視されている比較処理に有効である高速類似度判定アルゴリズムを提案し，ヒトゲノム解析アクセラレータLSIアーキテクチャを考案した．更に，それに基づいてレイアウトを行った．シミュレーション結果から回路規模を24行24列とした場合，1秒間に約480万塩基からなる配列を比較する事が可能であることを見積もり，高速・安価・小型のホモロジー検索システム実現の見通しを得た．今回設計を行った回路では24箇所の塩基変化まで精度を保った測定が可能である．現段階では，ゲノム上の塩基配列の中で個人（例えば健康な人と病気の人）間で異なっている塩基「Single Nucleotide Polymorphism」（1塩基多型）を探す場合には十分実用的である．更に，リニアリティの良いバッファを用いることができれば，回路規模を100行100列とした場合に1秒間に約160万塩基列を比較することが可能であることを見積もることができた．VLSIの集積度が進むとPEの集積度の向上や低消費電力化が期待でき，家庭でも解析を実行できるほどになるため，将来的には高速で低コストな遺伝子解析機器や携帯医療端末などへの応用が考えられる．

謝辞 本研究でのチップ試作は東京大学大規模集積システム設計教育研究センターを通しローム(株)の協力で行われたものである．本チップの設計はAvant!ツールを用いて行われたものである．

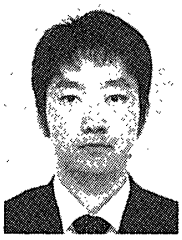
文　献

[1] J. Akita and J. Sese, "Fast homology search architecture for recognizing GenomeFunction," Proc. 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI2001), vol.VI, Part I, pp.6–9, July 2001.

[2] 佐々木勝光, 秋田純一, 深山正幸, 吉本正彦, "ヒトゲノム解析アクセラレータLSIアーキテクチャ," システムLSI北九州ワークショップ, pp.311–314, Nov. 2001.

[3] 佐々木勝光, 秋田純一, 深山正幸, 吉本正彦, "ヒトゲノム解析アクセラレータLSIの設計," 2002信学総大, C-12-23, 2002.

（平成14年10月4日受付，15年2月12日再受付）

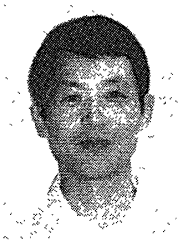


佐々木勝光（学生員）

平15金沢大学大学院自然科学研究科博士前期課程了．工修．ヒトゲノム解析用アーキテクチャ研究に従事．

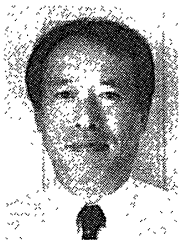
秋田　純一（正員）

平5東大・電子卒．平10同大大学院工学系研究科電子情報工学専攻博士課程了．博士（工学）．平10から金沢大学工学部電気情報工学科助手．平12から公立はこだて未来大学システム情報科学部情報アーキテクチャ学科講師．視覚系の機能をもつ画像処理系などの高速並列処理系の集積回路アーキテクチャ，及び人間中心の実世界志向インタフェースとそのハードウェアに関する研究に従事．情報処理学会，日本ロボット学会各会員．



深山　正幸（正員）

1988筑波大・第三学群情報学類卒．1995北陸先端科学技術大学院大学情報システム研究科博士前期課程了，工修．2000金沢大学電気電子システム工学科助手（現職）．現在，マルチメディア集積システムの研究に従事．



吉本　雅彦（正員）

昭52名古屋大大学院工学研究科前期博士課程了．博士（工学）．昭52三菱電機(株)入社．高性能MOSスタティックRAM，画像処理システムLSIなどのVLSI設計研究に従事．平12から金沢大学工学部電気電子システム工学科教授マルチメディア応用システムVLSIのアーキテクチャ研究に従事．情報処理学会会員．