－論文－

●グリッドコンピューティングによるヒト染色体DNA塩基配列の周期性探索

新製品開発研究所　久米 勝嘉

ヒトのDNAに潜むであろう長大な周期性を解析し、遺伝子病との関連を明らかにするNTTデータ・Cell Computing大規模実証実験・遺伝子病治療研究プロジェクト（2002年12月から2003年4月末まで行われた㈱NTTデータによるCellComputing大規模実証実験研究テーマの一つとして東亞合成㈱ 新製品開発研究所　吉田徹彦のプロジェクトテーマが採用[1]）の成果の一つとして、ヒト染色体DNA塩基配列中に10,000塩基以上の長大な周期性を有するタンデムリピート領域を見出したのでここに報告する。

1　緒　　言

2003年4月にヒトゲノム配列の完全解読（全ゲノム配列の99%の部分に対して精度99.99%）が終了した[2]。　今後はこのゲノム配列情報を解読し、その機能を明らかにしてゆくことで、最終的には医療分野、医薬品開発などに利用することが期待されている。ヒトゲノム塩基配列に対する研究の多くは遺伝子に着目したものであり、ゲノム配列中に存在する遺伝子の発見、遺伝子機能の解明を行うものである。しかし、ヒトゲノム配列中の遺伝子領域の占める割合は25%程度であり、さらにこの遺伝子領域中でタンパク質に翻訳されるエクソン領域は約1%と全体に占める比率は小さい。また、ヒトゲノム配列中には塩基配列比率で35%以上にもなる様々な繰り返し配列が含まれていることが知られている[3]。この繰り返し配列はハンチントン病の原因となる3塩基配列（CAG）の繰り返しのように疾病との関係が明らかにされたものも存在するが、その多くは機能的な意味が判明していないものが大半である。

このようにヒトゲノム配列には、まだ十分に解析が進んでいない領域や機能の判明していない塩基配列が多数存在しており、ここには未知のまだ解読できていない情報やルールが何らかの形でゲノム配列中に記述されていることが予想される。しかし、ヒトゲノム配列のデータ量は膨大であり、新規なルールの発見のための解析を実用的な期間内で実施することは困難である。

今回、㈱NTTデータによるGridComputing技術のデモンストレーションであるCellComputing大規模実証実験[4]に参画することで、数万台規模のコンピュータを利用することにより得られる膨大な計算能力を用いて大規模な解析を実施することが可能となった。GridComputingとは、ネットワーク上の多数のコンピュータを利用するための技術の総称であり、今後のコンピュータの利用方法を大きく変える技術として高い関心が寄せられている。その一つの応用方法として多数のコンピュータに計算作業を分散させることで超高速のコンピュータとして膨大な計算処理を行うことが期待されており、地球外知的生命体の探査を目的としたSETI@Homeがその応用例として有名である[5]。

今回、配列の一部分が非常に長い周期で繰り返し出現するタンデムリピートの探索をおこなった。ヒトゲノム中に数多く存在することが知られている繰り返し配列には、一定長の配列が連続して繰り返すタンデムリピートと呼ばれるタイプのものと、ゲノム中に数多くの類似配列が存在している分散型反復配列と呼ばれるタイプのものが存在する。このうちタンデムリピートについては、先に述べたように周期の短いものについては、各種の遺伝子病と関係しているものがあることが知られているが、周期の長いものについてはまだ検討が行なわれていない。また、塩基配列中に含まれるタンデムリピート配列を探索する作業を実施するためのソフトウェアが多数提案されているが、既存の探索ソフトウェアには必要な計算量の問題から検出可能な繰り返し周期の長さや、取り扱える塩基の全長等に制約が存在している。たとえばBenson[6]らのTandem Repeat Finderは、Fig.1に示すようなタンデムリピートを配列と周期について一定の基準内での変動を許容しながら検出するプログラムであるが検出可能な周期は2,000以下に限られる。

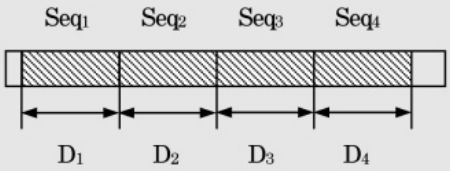

Fig.1　Tandem repeat

長周期のタンデムリピートの探索が困難であるのは、探索対象内のすべての可能性のある繰返し周期について検討する必要があるため計算量が周期の大きさとデータ量に比例して大きくなることと、さらに繰り返す塩基配列が同一ではない場合を考慮したり、繰り返しの周期が完全に一致しない場合を考慮すると急速に計算量が増加することにある。今回、長周期のタンデムリピートを探索するにあたり、1周期の長さが1万塩基以上のものを対象とすることとした。

一方、繰り返しの周期が変動することは今回は考慮しないこととした。また、繰り返し周期が長い場合、配列全体が一様に類似している可能性は非常に低いと考えられるため、Fig.2に示したようなタンデムリピート繰り返し単位の配列のうち、その一部分のみが類似しているタイプの繰り返し配列を探索することとした。これは一定長の塩基配列Seq_{org}に類似した塩基配列Seq_{sim}が一定の周期間隔Dで出現するタイプの繰り返し領域と表現することもできる。このような部分配列が一定の周期で繰り返す領域を部分タンデムリピートと呼ぶこととし、繰り返しの回数nに応じてn回繰り返し部分タンデムリピートと呼ぶことにする。例えばFig.2のケースは4回繰り返し部分タンデムリピートである。

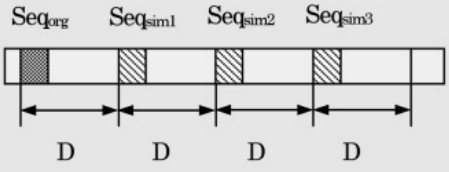

Fig.2　Partly repeated tandem repeat

今回の探索では部分タンデムリピートの周期Dが1万以上、部分タンデムリピート中の繰り返す部分配列の長さを100塩基、部分配列の類似度が70%以上のものを探索することとした。

2　実　　験

2.1　GridComputing用プログラムLateen

2.1.1　プログラム

㈱NTTデータのCellComputingは、GridComputingを構成するための技術としてすでに実績のあるUnited Devices社で開発されたMetaProcessorPlatform技術を採用している。CellComputing用のアプリケーションはプログラムの配布先のオペレーションシステム環境（今回はMicrosoft Windows™のみ）で動作するプログラムであり、かつMetaProcessorPlatformの機能を利用するためのUD関数を追加するためにC言語のソースコードが用意できるものである必要がある[7]。このため既存のプログラムの流用は断念し、探索プログラムの基本部分は新規に作成することとした。この基本部分に㈱NTTデータによりUD関数の付加によるセキュリティ機能、動作状況の把握機能の追加、動作時に表示されるスクリーンセーバープログラムの追加作業が行われ、CellComputing大規模実証実験で利用するプログラムLateen（Large-sequence Analytical Tool Energized by an Extensive Network）を作成した。

Lateenは、まず探索対象の染色体配列から一定長の塩基配列Seq_{org}を取り出し,Seq_{org}に一定以上の類似度で一致する同じ長さの配列Seq_{sim}の位置を探索する。この作業をSeq_{org}の位置を変化させて実施し、染色体内の一定の距離範囲の類似する2つの配列の組み合わせの位置情報をすべてリストアップする。入力塩基配列データはfasta形式の塩基配列データを使用する。結果出力ファイルは、比較元配列の開始位置、比較元配列と比較先配列の塩基配列距離、2配列の一致確率が記述される。また、比較する配列の長さ、出力する一致確率の下限値、探索する2配列の間隔の範囲のパラメータが制御ファイルで指定可能であり、このパラメータを調節することでCellComputing大規模実証実験における各端末での計算分担量の調整をおこなった。

2.1.2　使用データ

解析対象となる塩基配列データは、CellComputing大規模実証実験の開始時点で塩基配列の配列解析が終了し、配列データが公開されていた第13,14,20,21及び22番染色体の塩基配列データをNCBIのwebサイト内[8]のHumanGenomeResourcesより、その時点（2002年10月）での最新の配列データをダウンロードして利用した。入手した塩基配列データは、一部にまだ塩基配列が決定されていないギャップ部分が含まれているため、ギャップ部分に関してはギャップ長さの分の記号Nとして表現し、配列決定されている部分と結合した。類似配列の探索計算はCellComputing実験参加者のPCで実施されるが、Lateenプログラムは解析対象の塩基配列をすべてメモリ上に展開するためPCのメモリが少ない場合、動作に不具合が生じる可能性がある。今回、実験参加者の平均的なメモリ搭載量を考慮し、いったん結合した染色体配列データを2,500万塩基単位で分割した。このとき分割による探索漏れをなくすため1,200万塩基分をオーバーラップさせて分割をおこなった。このため探索する類似配列の間隔の上限は1,200万塩基とした。

2.1.3　動作パラメータ

Lateenの動作を制御するパラメータのうち、比較する塩基配列の単位は100塩基、類似配列として採用する配列の一致確率の下限は0.7とした。並列化のための計算対象の分割は、計算対象データ自体の分割、及び分割データ内の計算担当部分の分割により実施した。計算対象データとして、第13,14,20,21及び22番の染色体配列をそれぞれ2,500万塩基単位で分割した21種の塩基配列データを利用した。この分割された各塩基配列データについて、検討を行う部分タンデムリピートの周期の範囲を109種類に分割した。これにより合計1,537種類に計算処理作業を分割した。

2.2　CellComputing大規模実証実験

CellComputing大規模実証実験は、2002年12月20日から2003年4月30日まで実施された。㈱NTTデータ社内に設置されたCellComputingサーバーより、Lateenプログラムと分割された染色体塩基配列データ及び計算範囲を指定したパラメータファイルがCellComputing実験参加者のPCに配信される。Lateenプログラムは配信されたCellComputing参加者の各PCで動作し類似領域の探索を実施する。なお、動作中はFig.3のようなスクリーンセーバー画面が表示される。

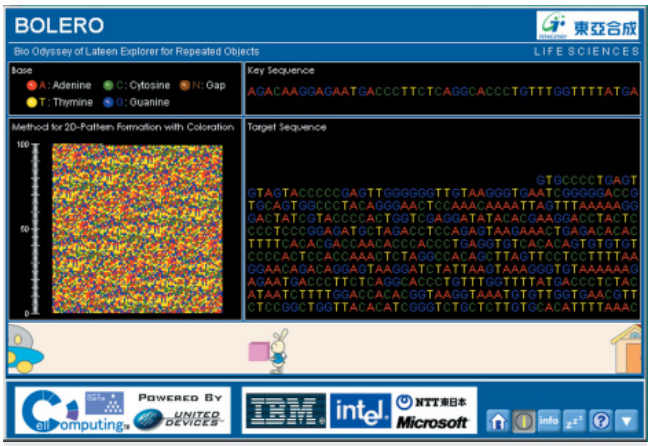

Fig.3　CellComputing screen saver

各PCでの計算終了後、結果出力である類似配列の組み合わせデータリストは暗号化され、インターネット経由でCellComputingサーバーに返送される。データの返送を確認した後、再度新しい配列データと計算範囲の指定ファイルが送信され異なるエリアの探索が開始される。また、今回の実験では各PCで正常に計算が行なわれたかどうかを検証するために同じ設定のパラメータファイルの組み合わせを複数のPCに送付しその結果が一致するかどうかで結果の信頼性を確認した。返送されたデータについては結果の検証後、一括してオフラインで受け取り解析作業の実施をおこなった。

2.3　データ処理

Lateenにより収集された類似配列の組み合わせデータから、長周期部分タンデムリピートの探索を実施した。探索作業は比較元配列Seq_{org}に対する類似配列Seq_{sim}の先頭塩基位置とSeq_{sim}の先頭塩基の位置の差で表される類似配列間の距離についてリストを作成し、このリストから配列間距離が整数比となるSeq_{org}とSeq_{sim}の組み合わせを探索することで実施した。

3　結果と考察

3.1　CellComputingでの計算処理結果

第13,14,20,21及び22番染色体の塩基配列データ中の類似領域の探索作業を計算対象を1,537分割して実施した。分割した各job 1つあたりの実行時間は、参加者の標準的なPC（PentiumIII 1GHz相当）で計算を行った場合、約24時間であった。CellComputing大規模実証実験に参加したPCの台数は最終的に12,206台であり、これらのPCで分担して冗長度10（同じ計算を10台で独立に実行）で計算を実施した。対象範囲の検討に必要な計算は4ヶ月間のCellComputimng大規模実証実験期間中にすべて完了した。Lateenプログラムにより収集された類似配列の組み合わせは、総計293,472,764組であり、出力データサイズはgzip形式で圧縮した状態で15.8GBであった。

3.2　長周期リピートの探索結果

Lateenの実行により得られた類似配列情報データを解析し、繰り返し回数が多く、周期の長い部分タンデムリピートの探索を行なった。繰り返す部分配列の一致確率が70%以上で類似配列の繰り返し回数が3回以上、かつ一定間隔で繰り返す周期が1万以上という条件で探索した結果、発見された部分タンデムリピートの総数は11,826個であった。繰り返しの回数については4回繰り返すものが最大であり5回以上繰り返すものは発見できなかった。各染色体毎の3回、4回繰り返しの長周期部分タンデムリピートの数についてはそれぞれTable1,2に示した。条件に適合する長周期のn回繰り返しの部分タンデムリピートの数は、各染色体とも繰り返しの回数、配列の一致確率の設定により大きく変化した。

Table1　Number of 3times repeat region

Chromosome	DNA sequence similarity				
	100%	>95%	>90%	>80%	>70%
Chr13	0	2	20	290	1,503
Chr14	0	0	3	481	2,864
Chr20	0	0	3	438	3,167
Chr21	0	0	6	125	691
Chr22	0	0	7	699	3,601

Table2　Number of 4times repeat region

Chromosome	DNA sequence similarity				
	100%	>95%	>90%	>80%	>70%
Chr13	0	0	0	0	7
Chr14	0	0	0	0	0
Chr20	0	0	0	0	1
Chr21	0	0	0	0	1
Chr22	0	0	0	0	0

部分タンデムリピートの周期について最大のものは、3回繰り返しの場合、探索範囲の上限である600万であり、4回繰り返しの場合は約193万であった。3回繰り返しの部分タンデムリピートの周期について各染色体別にその分布状況をFig.4に示した。部分タンデムリピートの数はいずれの染色体においても20万以下の周期のものが多い傾向が見られた。

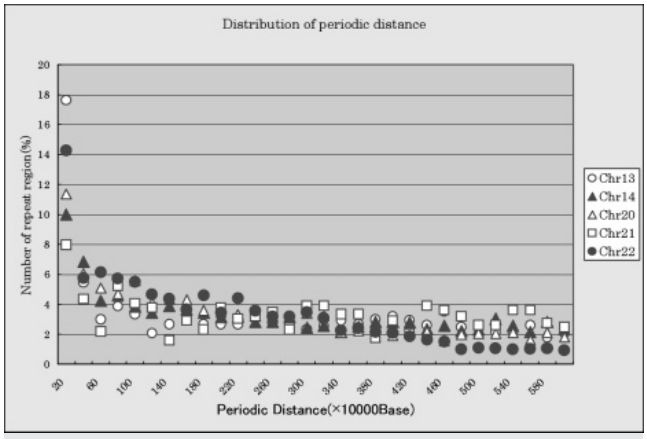

Fig.4　Distribution of periodic distance

さらに今回探索された部分タンデムリピートの染色体上での分布をみるために開始位置と周期の分布についてプロットした結果をFig.5-9に示す。染色体塩基配列中の部分タンデムリピートの開始位置の出現場所の分布には偏りがあり特定の領域に集中する傾向が見られた。しかし、部分タンデムリピートが多く出現する場所は今回検討した染色体毎に異なっており共通する傾向は見られなかった。また、部分タンデムリピートの周期と開始位置の間の関係についても各染色体間で共通の傾向は見られなかった。

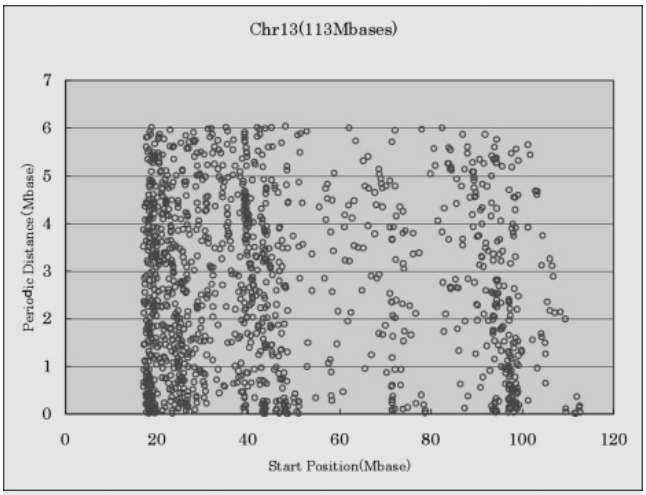

Fig.5　3times repeat distribution in Chromosome13

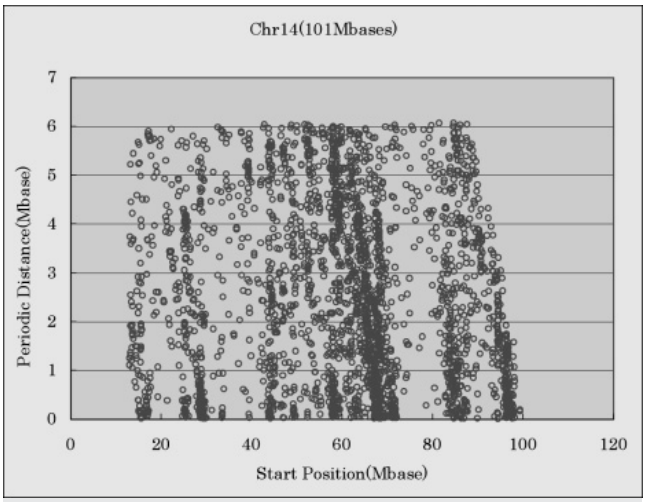

Fig.6　3times repeat distribution in Chromosome14

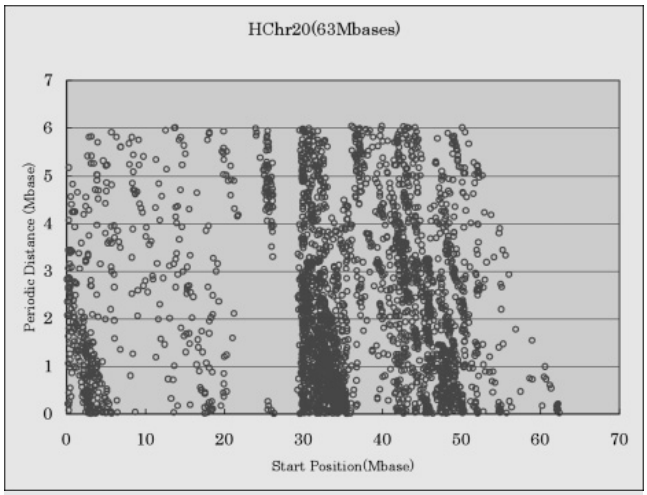

Fig.7　3times repeat distribution in Chromosome20

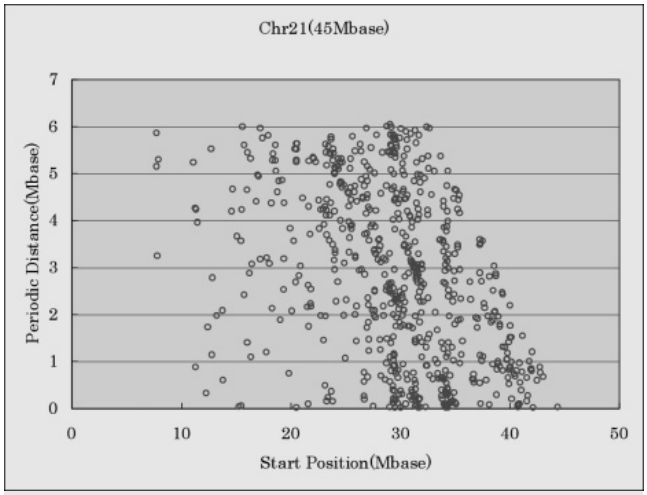

Fig.8　3times repeat distribution in Chromosome21

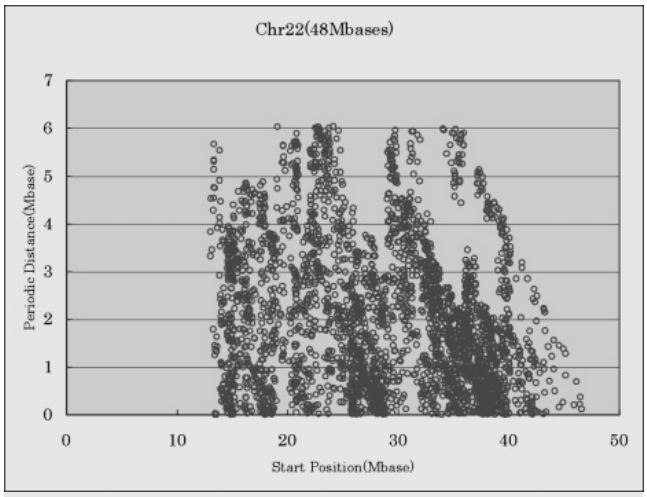

Fig.9　3times repeat distribution in Chromosome22

今回検出された長周期部分タンデムリピートのうち、4回繰返しの部分タンデムリピートは偶然に出現する可能性は非常に低く、なんらかのルールを反映した領域が検出されていると考え、その領域の検討を行なった。今回検出された4回繰返しの部分タンデムリピートの染色体塩基配列分割データ上の位置、周期に関するデータをTable3に示した。染色体塩基配列分割データは1,200万塩基を重複させ先頭から2,500万塩基長単位で分割したものであり、例えばChr13_7は13番染色体の7番めの分割データを示している。

Table3　Start position of each repeated similar sequence

chromosome data	Start Base adress Seq_{org}	Seq_{sim1}	Seq_{sim2}	Seq_{sim3}	Distance
Chr13_7	16,548,100	16,558,944	16,569,788	16,580,632	10,844
Chr13_7	16,548,200	16,559,010	16,569,820	16,580,630	10,810
Chr13_7	16,548,300	16,559,076	16,569,852	16,580,628	10,776
Chr13_7	16,548,400	16,559,142	16,569,884	16,580,626	10,742
Chr13_7	16,548,500	16,559,208	16,569,916	16,580,624	10,708
Chr13_7	16,548,600	16,559,274	16,569,948	16,580,622	10,674
Chr13_7	16,548,700	16,559,340	16,569,980	16,580,620	10,640
Chr20_4	8,300,600	8,361,894	8,423,188	8,484,482	61,294
Chr21_2	7,037,400	8,971,457	10,905,514	12,839,571	1,934,057

また、4回繰り返す部分配列をTable4に、この配列情報を2D色彩法[9]で表示したものをFig.10,11,12に示す。ここで各塩基はAを赤、Tを黄、Cを緑、Gを青として表現した。また、4回繰り返している部分配列は図中の白線に挟まれた部分である。2D画像で表現することにより、検出された領域は類似性の高い部分配列が同じ周期で4回繰り返して出現したものであることが確認された。

Table4　Sequences of repeated region

>Chr13_7_1 Seq_{org}
TCTTCATGATGGGCCCAGGTAAGCATGGACCATTCCTTCAGGATGGAGCCAGGTAAGTGTGGACCATTTCTTCAGAATACAGCCAGGTAATCATGGACTG
>Chr13_7_2 Seq_{org}
TTTCTTCAGGATGGGCCCAGGTAAGCATGGACCATTCCTTCAGGATGGAGTCACGTAAGTGTGGACCTTTCCTTCAGGATGGAATCAGGTAAGCAGGGAC
>Chr13_7_3 Seq_{org}
CATTCCTTCAGAATACAGCCAGGTAAGCATGCACAGTTTCTTCAGGATGGGCCCAGTTAATCATGGACCATTCCTTCAGGATGTAGTCAGGTAAGTGGGG
>Chr13_7_4 Seq_{org}
ACCATTCCTTCAGGATGGATAAAGGTAACCATGGACTGTTTCTTCAGGATGGGCCCAGGTAAGCATGGACCATTCCTTCAGGATGGAGCCAGGTAAGTGT
>Chr13_7_5 Seq_{org}
GGACCATTCCTTCAGAATACAGCCAGGTAAGCATGGAGTGTTTCTTCAGGATGGGCCCAGGTAAGCATGGAACATTCCTTCAGGATGGAGTCAGGTAATT
>Chr13_7_6 Seq_{org}
GTGGACCATTCCTTCAGAATACAGCCAGGTAAGCATGGACTGTTTCTTCAGGATGGGCCCAGGTAAGCATGGACCATTCCTTCAGGATGGAGTCACATAA
>Chr13_7_7 Seq_{org}
GTGTGGACCATTCCTTCACGATGGAATCAGGTAGGCAGGGACCATTCCTTCAGCATACAGCCAGGTAAGCATGGACTGTTTCTTCAGGATGGGCCCAGGT
>Chr20_4 Seq_{org}
AAAATTAGCCGGGCATGGTGGCGGGCGCCTATAGGTCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGCGTGAACTCAGGAGGCAGAGCTTGCAGTGAG
>Chr21_2 Seq_{org}
GCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCGGGGAGGCAGAGCTTGCAGTGAGCCGAGATGGTGCCACTGCACTCCAGCCTGGGCAACAGAG

Fig.10　4times repeated region of Chr13_7

Fig.11　4times repeated region of Chr20_4

Fig.12　4times repeated region of Chr21_2

この4回繰り返している100塩基長の配列の由来をTable5に示す。

Table5　Source of repeat region

Chromosome data	Seqorg	Seqsim1	Seqsim2	Seqsim3
Chr13_7	34mer repeat	34mer repeat	34mer repeat	34mer repeat
Chr20_4	AluY	AluSx	AluJb	AluSx
Chr21_4	AluY	AluSx	AluSq	AluSg

第13番染色体で発見された7箇所の繰り返し部分配列は、いずれも34塩基を一単位とする繰り返し配列であった。4回繰り返しの部分タンデムリピートが発見された領域は34塩基単位の類似した配列が少々の変異を伴いながら約50,000塩基ほど繰り返すミニサテライト領域に存在し、発見された4回繰り返し部分タンデムリピートは、この長く続くミニサテライト領域の規則性を反映したものであることが判明した。このミニサテライト領域について、同じくAを赤、Tを黄、Cを緑、Gを青として表現し、幅を170塩基（34塩基×5）とした2D色彩法で表現した結果をFig.13に示す。

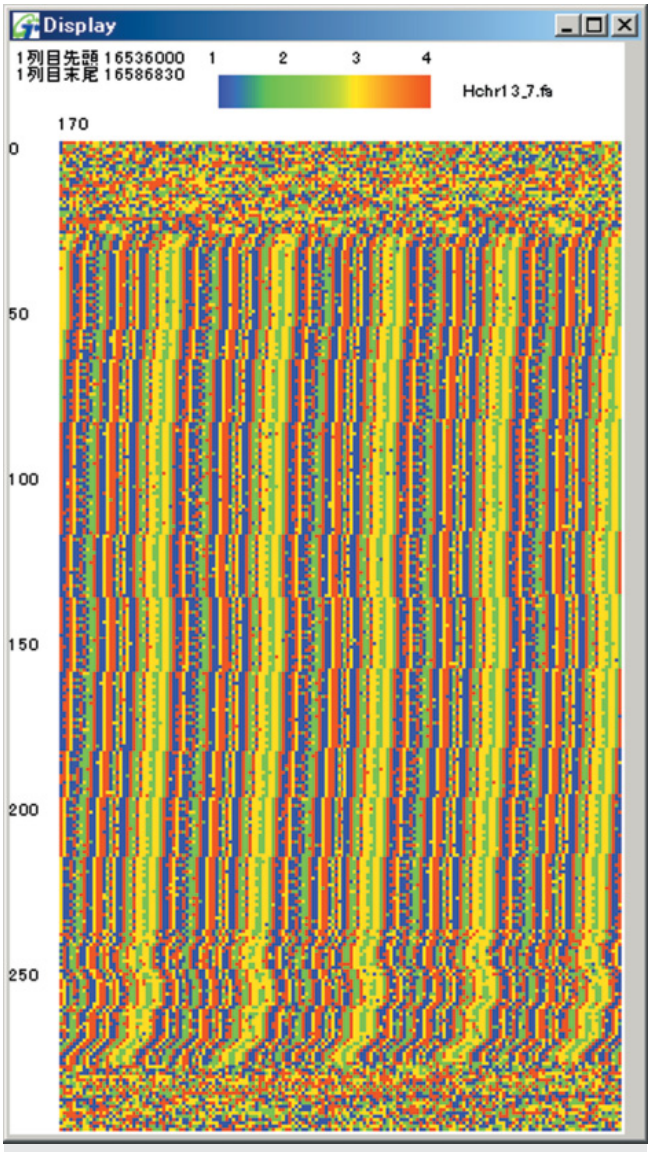

Fig.13　Minisatellite region of Chrmosome 13

また、第20,21番染色体での繰り返し部分配列はいずれもAluファミリーに属する配列であった。Aluはヒトゲノム中に大量に存在する分散型反復配列であり、30億塩基のヒトゲノム全体に100万個以上が存在するとされている。このため平均では約3,000塩基あたり1回出現することになる。Aluがランダムに分布すると仮定した場合の3回繰返しの部分タンデムリピート総数の推定値と測定値の比較を行い、その結果をTable6に示した。

Table6　Calculated .vs Observed Repeat region number

	Number of repeat region 3times	4times	3times/4times
calc.	600,000,000	200,000	3,000
obs.	11,826	7	1,689
calc./obs	50,736	28,571	

Calc. = (Number of total possible sequence pair) / (Alu distance)

実際に発見された3回繰返し、4回繰返しの部分タンデムリピートの総数はAlu配列がランダムに分布していた場合に予想されるリピート領域数に比較して数万分の1と少ない。これはFig.4-9に見られる長周期部分タンデムリピートの存在位置の分布の大きな偏りを反映しているものと考えられる。この偏りの原因はAluの存在密度が染色体中の場所により異なること[10]が考えられるが、他にも部分タンデムリピートを構成する制約条件が存在している可能性も考えられる。この制約条件の探索する試みの一つとして4回繰返しの部分タンデムリピート中の繰り返し出現している部分領域を含む遺伝子を調査した。その結果をTable7に示す。Aluに関係した部分配列が繰り返している20,21番染色体中の領域の場合、計8箇所の繰り返し出現する部分配列のうち4箇所が何らかの遺伝子中に含まれていた。ヒトゲノム中の遺伝子領域はゲノム中の25%程度であることから遺伝子が高密度に存在する領域であることが長周期の部分タンデムリピートが存在する条件である可能性がある。

Table7　Genes related with repeat region

sequence	repeat reagion containing gene
Chr13_7	
Seq_{org}	none
Seq_{sim1}	none
Seq_{sim2}	none
Seq_{sim3}	none
Chr20_4	
Seq_{org}	BIG2 (BREFELDIN A-INHIBITED GUANINE NUCLEOTIDE EXCHANGE PROTEIN 2)
Seq_{sim1}	CSE1L (CHROMOSOME SEGREGATION 1-LIKE)
Seq_{sim2}	STAU (STAUFEN, DROSOPHILA, HOMOLOG OF)
Seq_{sim3}	none
Chr21_2	
Seq_{org}	none
Seq_{sim1}	none
Seq_{sim2}	CRYZL1 (CRYSTALLIN, ZETA-LIKE 1)
Seq_{sim3}	none

3.3　まとめ

ヒトDNA塩基配列について、従来検討されたことのない長周期の繰り返し構造（長周期部分タンデムリピート）の探索をおこなった。具体的には、第13,14,20,21及び22番染色体について、100塩基長塩基配列が類似度70%以上で10,000塩基以上の一定の周期で3回以上繰り返す条件を満たす領域を探索した。5回以上一定の長周期で部分配列が繰り返す領域は存在しなかったが、4回繰り返す領域を7箇所、3回繰り返す領域を11,826箇所で発見した。これら部分タンデムリピートの染色体中での出現位置は一様ではなく偏りが存在し、ゲノム配列中の何らかの特徴、規則性を反映していると考えられる。

謝　辞

CellComputing大規模実証実験に参加する機会を与えて頂き、プログラム開発、計算の実施にあたって多大なご支援をいただいた㈱NTTデータ技術開発本部の鑓水リーダー、副田様、CellComputingグループメンバーの皆様に深く感謝いたします。またCellComputing大規模実証実験に参加頂いた参加者の皆様に深く感謝いたします。

引用文献等

1) CellComputing「遺伝子病治療研究プロジェクト」, http://www.cellcomputing.jp/project/index_b.html
2) International Consortium Completes Human Genome Project press release, http://www.ornl.gov/TechResources/Human_Genome/project/50yr/press4_2003.htm
3) J.C.Venter, et al Science 291,1304 (2001).
4) CellComputing 大規模実証実験, http://www.cellcomputing.jp/test/index2.html
5) SETI@home, http://setiathome.ssl.berkeley.edu
6) G.Benson, *Nucleic Acids Res*., 27, 573 (1999).
7) UNITED DEVICES™ MetaProcessor Platform Version2.2 Application Developer 's Guide (2002).
8) NCBI Human Genome Resources, http://www.ncbi.nlm.nih.gov
9) T.Yoshida,N.Obata,K.Oosawa, *J.Mol.Biol*, 298,343 (2000).
10) The Genome International Sequencing Consortium, *Nature* **409**, 860 (2001).