PC クラスタを利用したバイオインフォデータマイニング

島本 真衣　(指導教員：小口 正人)

1　はじめに

近年、発達する情報化社会では、データの蓄積と運用が非常に重要になっており、情報システムにおいて処理されるデータ量が膨大になってきている。特にバイオインフォマティクス分野においては、将来的に役立つことが予想される貴重なデータが蓄積され続けており、それらを効率良く解析し利用することが求められている。これには既存のデータマイニング技術を応用できるが、データによっては属性値の種類が極めて多いなど、バイオ分野独特のデータ構造も考慮していく必要がある。

そこで本研究では、バイオデータを PC クラスタシステム上でマイニングし、トランザクションデータを用いてマイニングを行った際とのシステムの振舞の違いを考察する。

2　相関関係抽出とその並列アルゴリズム

相関関係抽出では、巨大なデータからあるパターンが現れる頻度 (サポート値) を調べる。その頻度が多ければ、そのパターンは有意義なデータとなり、販売戦略などに活用出来る。

相関関係抽出で扱うデータはしばしば巨大であるため、データベースを分散し計算処理を並列化して、多数台のコンピュータをネットワークで接続した PC クラスタなどの環境でマイニング処理を実行する並列相関関係抽出の研究が行われている。相関関係抽出の代表的な 2 つのアルゴリズムの概要を説明し、本研究で用いる並列化アルゴリズムを紹介する。

2.1　Apriori アルゴリズム

1994 年に Agrawal らによって提案されたもので、発見された頻出アイテムセットから候補アイテムセットを生成し、繰り返し数え上げを行っていくアルゴリズムである。1 回目のデータベーススキャンで頻出 1 アイテムセットを抽出し、それらを元に候補 2 アイテムセットを生成する。2 回目のデータベーススキャンで候補 2 アイテムセットから頻出 2 アイテムセットを抽出する。これを候補アイテムセットが生成されなくなるまで繰り返していくことで、頻出パターンをすべて発見していくアルゴリズムである。

Apriori アルゴリズムには、候補アイテムセットを格納するために大容量のメモリが必要となる、何度も繰り返しデータベースをスキャンする可能性があるといった問題点がある。

Apriori をベースにした並列相関関係抽出のアルゴリズムはいくつか提案されているが、本研究ではハッシュ関数を使用して Apriori を並列化する HPA (Hash Partitioned Apriori)[1] を用いる。

2.2　FP-growth アルゴリズム

2000 年に Han らによって提案されたもので、巨大なトランザクションデータベースから相関関係抽出に必要な情報をコンパクトに圧縮したデータ構造である FP-tree を利用している。候補パターンを生成せずに頻出パターンを抽出することで、Apriori アルゴリズムの問題点を改善したアルゴリズムである。

FP-tree は次のように構築される。1 回目のデータベーススキャンで各アイテムのサポート値を求め、頻出アイテムを抽出し、抽出された頻出アイテムをサポート値により頻度が降順になるように並び替え (これを F-list とする)、空 (null) のラベルを持つ木のルートを作成 (これを T とする) する。2 回目のデータベーススキャンで、F-list に従ってトランザクションから頻出アイテムを抽出し、ソートする。T が F-list の要素である子を持っていれば、その子のカウントを 1 増やし、持っていなければ、新しくカウント 1 を持つ子を作る。F-list の最後までこの操作を繰り返す。FP-tree の構築例を図 1 に示す。

FP-growth はこのように構築された FP-tree の性質を利用することにより、頻出パターンを発見していくアルゴリズムである。FP-growth の並列相関関係抽出のアルゴリズムは、本研究では HPA を元に行われた既存研究 [2] で提案されたものを用いる。

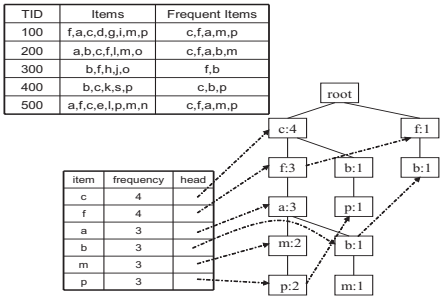

図 1: FP-tree の構造例

3　バイオインフォデータ

3.1　DNA

遺伝情報をコーディングする生体物質を DNA(Deoxyribo Nucleic Acid) と呼ぶ。DNA はデオキシリボース（糖）とリン酸、塩基 (アデニン (A)、グアニン (G)、シトシン (C)、チミン (T)) から構成されていて、アデニン (A) はチミン (T) と、グアニン (G) はシトシン (C) とだけ結びつき、対を成す鎖は相補的な構造になっている。

3.2　SNP

図 2 で示すように、DNA 上で、それぞれの染色体が 1 塩基異なっている部位を SNP(Single Nucleotide Polymorphism:一塩基多型) と呼ぶ。人によって薬の効き方や副作用の出方が違うのはこの SNP が原因ではないかと考えられている。SNP データは HapMap プロジェクト [3] の HP 上で公開されている。

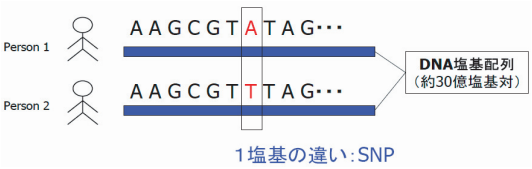

図 2: SNP(一塩基多型)

4 実験内容と実験環境

本研究はトランザクション数 (被験者数)45 人、アイテム種類数 (塩基番号) 約 50000 個のデータを用いて実験を行う。まず公開されている SNP データを相関関係抽出処理できるよう、出力方法、ハッシュ関数処理、ソーティング、フォーマット変換を行ったのち、変換したデータを 4 ノード分に分割し、Apriori と FP-growth という 2 つの異なるアルゴリズムに基づいたプログラムを PC クラスタ上で実行する。

実験には 18 台の PC を Gigabit Ethernet で接続した PC クラスタを用いる。各 PC は CPU が Intel Pentium4 1.5GHz、メインメモリが 384MB、OS が Linux2.6.9-1667(Fedora Core 3) である。

5 実行結果と考察

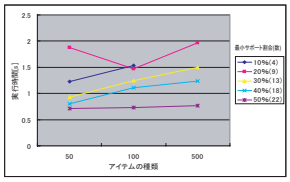

図 3: HPA 実行時間

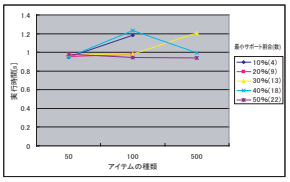

図 4: PFP 実行時間

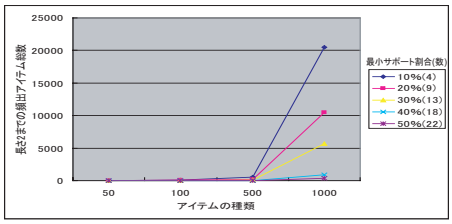

図 5: 長さ 2 までの頻出アイテム総数

データに手を加えアイテムの種類を変化させた時の実行時間を図 3 と図 4 に、その際の長さ 2 の頻出アイテム総数を図 5 に示す。

図 3、4 からもわかるように HPA・PFP のアルゴリズム共にアイテムの種類が 500 までは問題なく実行できた。アイテムの種類を 1000 にすると、PFP は実行できず、HPA も図 5 のように頻出アイテムの数が急激に多くなってしまい、最小サポート値を高めに設定しないと、実行が途中で止まってしまう。また頻出アイテム数が多くなることから、候補アイテムの数も膨大になり、最小サポート値を高く設定しても実行時間が非常に長くなってしまうことがわかった。

そこで CPU が Intel Xeon 3.8GHz、メインメモリが 4GB、OS が Linux2.6.9-55(CentOS4.5) の PC クラスタ上で同じデータを用いて実行したところ、頻出アイテム数が多い場合でも短い時間で実行が終了したが、最後まで実行できるデータは限られていた。図 6 はアイテムの種類が 50 のときの HPA の実行結果の比較である。Xeon & 4GB のクラスタでは実行時間が約半分になっていることがわかる。HPA・PFP 共に Xeon & 4GB のクラスタでの実行時間はほぼ半分程度の短縮であったが、アイテムの種類が 1000・最小サポート値が 50 %の HPA での実行時間は、Pentium & 384MB のクラスタでは 715［s］だったのに対し、Pentium & 4GB のクラスタでは 11［s］と大幅に短縮された。

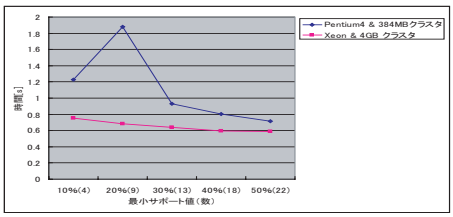

図 6: 各クラスタでの実行結果の比較

6 まとめと今後の課題

ハッシュ関数によって一定範囲に数値を出力する等、データの出力方法や HPA と PFP のプログラムの修正を行うことで、アイテムが 1000 種類のデータまで実行することが出来た。しかしアイテムの種類が 1000 を超えると、HPA では、頻出アイテムの数が膨大になり、メモリ上に確保したデータ領域の容量を超え実行が途中で止まってしまう。バイオインフォデータは通常のトランザクションデータと比較してアイテム種類数が極めて多いなどの特徴を持ち、マイニング処理には工夫が必要であることがわかった。

そこで一台あたりのメインメモリが 4GB の PC クラスタ上でも同じデータを用いて実行したところ、実行時間が短縮されたが、最後まで実行できるデータのアイテムの数と最小サポート値には変化がなかった。このことから実行が完了しない原因は全メモリ容量だけでなく、プログラム中におけるデータ領域の確保の仕方にもあることがわかった。今後は実行が止まってしまうプログラムの箇所を特定し、修正を行う。

謝辞

本研究を進めるにあたり、データ変換プログラムの提供および、大変有用なアドバイスを頂いた本学准教授の瀬々先生に深く感謝いたします。

参考文献

[1] 小口正人、喜連川優:”ATM 結合 PC クラスタにおける動的リモートメモリ利用方式を用いた並列データマイニングの実行”, 電子情報通信学会論文誌,Vol.J84-D-I, No.9, pp.1336-1349, 2001 年 9 月

[2] Iko Pramudiono and Masaru Ksuregawa: ”Tree structure based Parallel Frequent Pattern Mining on PC cluster”, DEXA2003,September 2003

[3] HapMap プロジェクト, http://www.hapmap.org/

[4] 島本真衣、小口正人:”PC クラスタを利用したバイオインフォデータマイニング”, 第 70 回情報処理学会全国大会,2ZK-5,2008 年 3 月発表予定