ICOT Technical Memorandom: TM-1062

TM-1062

並列推論マシンによるゲノム解析

新田　克己

June, 1991

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

並列推論マシンによるゲノム解析

新田 克己
(財) 新世代コンピュータ技術開発機構

1 はじめに

新世代コンピュータ技術開発機構 (ICOT: Institute for New Generation Computer Technology) は、第5世代コンピュータ (FGCS) プロジェクトの推進のために1982年に設立された財団法人である。ICOTは、今までに並列推論マシン実験機 Multi-PSI と並列論理型言語 KL1 を開発し, 並列の知識処理を行う環境を整えてきた。ICOTでは並列推論マシンの応用として、遺伝子情報の解析プログラムを開発している。ここでは、並列推論マシンと遺伝子への応用研究の概要を紹介する。

2 並列推論マシン

2.1 並列推論マシンの構成

FGCS プロジェクトは1990年代の知識情報処理の基盤技術を開発するための10年のプロジェクトである。このプロジェクトの特徴は、論理型言語をベースにしてハードウェア、基本ソフトウェア、応用ソフトウェア技術を開発することである。

論理型言語で書かれたプログラムの実行は、論理式の三段論法の繰り返しによる定理証明に相当する。従って、ICOTでは開発した計算機を推論マシンと呼んでいる。プロジェクトの前期には、逐次型推論マシン PSI を開発し、中期には、並列推論マシン実験機 Multi-PSI を開発した (図1)。Multi-PSI は64台の要素プロセッサ (PE) が2次元のメッシュ構造に結合されたもので、並列プログラミング技術の実験のために開発されたものである。さらに、本年度は 並列推論マシン PIM の開発が完成する。PIM には異なる技術を用いた PIM/p, PIM/m, PIM/c, PIM/k, PIM/i という5つの種類がある。例えば、PIM/p は、8台の PE が1つのクラスタを構成し、各クラスタごとに共有メモリを持ち、64のクラスタがハイパーキューブで結合されている。

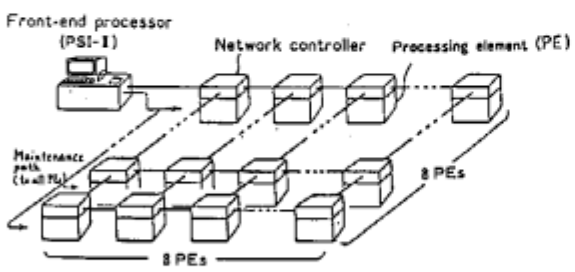

図 1: Multi-PSI の構成

Multi-PSI および PIM の上では、並列論理型言語 KL1 で書かれたプログラムが並列に実行される。

2.2 並列推論マシンの特徴

並列推論マシンを知識処理に使うことの特徴を以下に示す。

1. 論理型言語によるプログラミング
   並列推論マシンでは論理型言語でプログラムを書く。論理型言語は、記号処理を得意とする言語であり、人間の知識や推論過程を自然に記述できる利点がある。ICOTでは、論理型言語をベースにした知識処理技術やデータベース技術を開発してきている。

2. 並列処理による高速化
   プログラムは並列に実行される。膨大なデータや探索を必要とするような大規模問題においては、並列処理を行うことにより、実行時間の大幅な削減が期待できる。論理型言語 KL1 では、与えられた問題を、複数のプロセスがデータを交換しながら解決する、という形にモデル化してプログラミングする。並列処理にともなうプロセッサ割り付けや同期処理を容易に実現できる。

3 ICOT のゲノム解析

現在の分子生物学のデータベース技術・データ解析技術の問題点は以下のように集約される。

- 機能情報の利用が困難である。
- データ量が爆発的に増えている。
- 生物学知識の利用技術が不足している。

機能情報の充実と有効利用には、生物学のデータベースの統合化と、利用ノウハウの知識ベース化とが要求される。データ量爆発への対処には、並列処理などを利用した配列の検索／解析の高速化技術が有効である。生物

学知識の有効利用には、生データから知識を抽出する知識獲得技術や、生物学知識ベースを利用した配列の解析技術などが必要である。

ICOTでは、並列処理技術、知識処理技術、データベース技術などの応用として、以下の活動を行っている。

3.1 分子生物学の統合的知識ベース

既存のデータベースの問題点は、前述のように機能記述の利用が不十分なことにある。機能記述の充実は、機能のデータベースとの併用で図ることができる。データベースの併用を容易化するには、データベースを何らかの形で統合する必要がある。統合には3つの方法が考えられる。

1. データベースの標準化

2. アクセスソフトウェアの開発

   従来から試みられている方法としては、データベースの併用をサポートするアクセスソフトウェアの開発がある。その際データベース全体をDBMS下に統合管理することは当然有効だが、膨大なデータ量を蓄える環境を用意する必要に加えて、一般的なデータベースモデルである関係モデルには載りにくいデータであること(フォーマットの複雑さ、フォーマットの頻繁な変更など)が問題となる。

   ICOTでは現在、PSI上の非正規関係データベース管理システムKappaにGenBank、PIRと酵素DBが容易に格納できることを確認し、統合環境を試作中である。

3. 知識ベースアプローチ

   機能の記述は推論に利用できるように、意味論のしっかりした言語で記述されていることが望ましい。関係モデルはひとつの有力な候補である。しかし、機能記述の典型である化学反応式もまた、関係モデルに載りにくいデータ構造をしている。

   ICOTでは現在、*QUIXOTE* というDOOD言語(Deductive and Object-Oriented Database language)を開発し、分子生物学の知識を、データベースを含めて統一的に表現することを考えている。

3.2 蛋白質配列の並列解析プログラム

3.2.1 マルチプルアライメント

蛋白質配列の解析の1つの方法として、マルチプルアライメントがある。マルチプルアライメントのアルゴリズムは、多く発表されているが、その中でもDP(Dynamic Programming)の方法が良く知られている。DPの手法は2つの配列をアライメントする代表的な手法であるが、3つ以上の配列を同時にアライメントすることはノードの数が指数的に増加するので一般には困難である。そこで通常は、与えられた配列を似ているものから2つずつ選んでアライメントしていく方法が多く用いられているが、実行時間がかかり、アライメントの品質も満足できるものではなかった。

ICOTでは、以下の3つの並列アライメント・プログラムを開発した。

1. 3次元DPによるアライメント
   DPのアライメント手法を3次元に拡張し、3つの配列を同時にアライメントする並列プログラムである。入力された配列(4本以上)は、3本ずつ予備的にアライメントされ、これらがマージされて全体のアライメントが作られる。

2. トーナメント方式によるアライメント
   まず入力された複数(N本)の配列の中ですべての2本ずつの配列の組を比較する。その中で最も類似する2本の配列をアライメントして1本の抽象的な配列にまとめる。この操作を繰り返して、入力のすべての配列を最終的に1本の配列にまとめ、その結果を利用して、すべてのアライメントを再構築する。

3. シミュレーテッド・アニーリングによるアライメント
   ランダムに1つの配列の1つのアミノ酸を選び、そこにギャップを挿入してアライメントを微小変形し、Scoreを計算する、という操作を繰り返すことにより、アライメントの質を高めていくシミュレーテッドアニーリング(Simulated Annealing)手法を金久實教授(京都大学)は開発した。ICOTは、その手法に基づき、温度スケジュール不要の並列シミュレーテッドアニーリングによるアライメントプログラムを実現した。

以上の3つの手法を組み合わせることによるトータルシステムを実現中である。

3.2.2 蛋白質の折れ畳みシミュレーション

個々のアミノ酸の疎水性/親水性の性質を利用した蛋白質の折れ畳みのシミュレーションの並列プログラムを、Multi-PSIの上で実現中である。

4 おわりに

並列推論マシンによるゲノム解析について紹介した。並列プログラミングは世界的にまだ開発途上の技術であるが、ICOTでは この1～2年にその技術は急速に蓄積してきている。今後、並列推論マシンによる有効性が認識され、多くの解析ツールが作られるものと思われる。