

TM-1292

ゲノム情報との出会い

内田 俊一

February, 1994

© Copyright 1994-2-25 ICOT, JAPAN ALL RIGHTS RESERVED

ICOT

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

(03)3456-3191 ~ 5

Institute for New Generation Computer Technology

内田俊一（新世代コンピュータ技術開発機構）

今から6年前の1988年頃、私は第五世代コンピュータに適した応用問題を探していました。第五世代コンピュータのプロジェクト開始から7年がたち、64台のプロセッサから成る並列推論マシンの実験機が完成し、その上で世界初の本格的な並列OSも動きだした頃のことです。

まだ、最終目標の1000台規模の第五世代コンピュータプロトタイプの開発は始まっていませんでしたが、私たち担当研究者は、その成功に自信を持ち始めていたのです。しかし、同時にいくつかの心配事が頭に浮かんでいました。その一つは、このような大規模な知識処理向きのコンピュータをうまくデモできる応用問題が見つけれられるだろうかということでした。

現在の主流であるノイマン型コンピュータは、科学技術計算の類は得意ですが、知識処理の類は得意ではありません。従って、ノイマン型コンピュータの上に作られている知識処理の応用問題は、規模が小さく、第五世代コンピュータプロトタイプシステムを評価し、デモする目的には不十分なもののしかありませんでした。

それでは、自分たちで応用問題を見つけ、応用プログラムを作ってしまうという方針を固めて、何がよからうかと物色を始めました。ちょうどその頃、アルゴンヌ国立研究所（ANL）のRoss Overbeekさんから、その年の11月に、論理型言語と知識処理技術を分子生物学的解析に利用したらおもしろいと思うので、その議論をするワークショップを開くから来ないかという招待を受けました。

また、その背景として、米国ではHuman Genome Projectという大きなプロジェクトが走りだしたこと、そこでは分子生物学者がヒトのゲノムをどんどん読みとり、遺伝子データベースに集積していること、ゲノムに記述されている遺伝情報の解読がむずかしく新しいコンピュータ技術が求められていることなどを伝えてきました。Overbeekさんは、このワークショップの議論をもとに、ANLで一つのプロジェクトを起こすことを狙っていたのでした。

このANLのワークショップには、吉田かおるさん（現在、ローレンスバークレー研究所）と参加することとし、Overbeekさんが送ってくれた分子生物学の大量の入門書や論文を読みました。最初に分子生物学の用語に悩まされたことはいうまでもありません。最初から英語で勉強するのは辛すぎるということで日本語の入門書を揃えて、約1ヶ月、多忙な日常業務の合間に付焼き刃ながら、DNAの構造、RNAへのコピー、蛋白質の合成などについての知識を詰込みました。

この時点で、ゲノムに記述されている遺伝情報を解読しようという問題が知識処理の応用問題として魅力に富むものであることを確信しました。また、それ以上に、生物の遺伝子に関する情報処理やそれに基づいて蛋白質を合成し、生体を作りだしていくメカニズムそのものに対する興味のとりこになってしまいました。

ゲノム情報の解析は、いろいろな「謎解き」を含んでいます。遺伝子の機能や発現の機構、蛋白質の機能と構造、酵素反応、細胞の振舞などの解明をとおして、われわれは、“超”複雑な生物のメカニズムの一端をのぞき見ることができます。

たった4種の核酸から組み立てられるDNA、20種のアミノ酸からできている蛋白質、その簡素な構成要素に比べて、見事なまでに複雑でかつ効率良い生体機構、コピーミスをはとんどしない遺伝情報の複製機構、それに、自己と非自己をギリギリの線で識別する芸術的ともいえる免疫機構など。これらは、コンピュータ研究者の知識欲と好奇心をかきたてずにはおきません。コンピュータの構造との比較論、新しい情報処理機械への夢、生物である自分自身の構造を知ることによる興奮、いくらでも興味は広がります。

ANLのワークショップは、米国の分子生物学とコンピュータ科学の双方の先端分野で活躍している人が30人ほどが集まりました。私たちは分子生物学者から、ゲノムの解析と遺伝情報の解読における問題点の解説を受け、それを解決するためにどのようなコンピュータ技術を開発しなければならないかを議論し、レポートにまとめました。議論はバラエティに富み、おもしろかったものの、まとまりを欠き、そのためかOvebeekさんの意図したようなプロジェクトは成立しなかったようです。

我々は、帰国後、ICOTに遺伝子情報処理の研究グループを作る努力を始めました。といっても、ICOTにこのような分野の研究者はいませんから、誰か適当な先生を探して、教えてもらう必要があります。ANLやNIHの研究者とは共同研究をすることとしましたが、新人育成のためには国内の専門家の協力が不可欠でした。

そこで、私と吉田さん、それに米国留学し、やはり遺伝子情報処理に感化された小長谷明彦さん(NEC)の3人で、大学の先生とICOT研究員などからなるワーキンググループ(WG)づくりを始めました。

この遺伝子情報処理WGは、分子生物学の分野から、金久先生をリーダーに迎え、宮澤三造さん(当時、遺伝研、現在、群馬大)、八尾徹さん(当時、蛋白工学研、現在、三菱化成)などにメンバーなってもらい、1989年7月にスタートしました。

このWGの最初の頃は、コンピュータ科学と分子生物学の接点となるような研究テーマの存在を求めた勉強会であり、双方の語る言葉が通じないことも多々ありました。

しかし、これらの方々の協力により、ICOT内には遺伝子情報処理に興味を持つ研究者があらわれ、1990年4月から、新田克己さん、石川幹人さんをリーダーとする遺伝子情報処理の研究チームを発足させることができました。

このころ金久先生は、文部省の重点領域研究「ゲノム情報」の準備をしていて、分子生物学とコンピュータ科学の両分野の研究者の集め、これらの二つの分野にまたがる新しい研究分野をつくらうとしていました。米国ならば、研究者の流動性は高く、ANLやNIHにみるごとく、研究資金さえあれば、分子生物学者がコンピュータ科学者を必要に応じて雇うことで、新分野の研究者を作り出すことが容易です。しかし、日本では、通常は、関連分野の研究者や学生の興味を喚起して、その新分野に移行してもらい研究者を確保するか、新たに育成するかしていかなければなりません。

このような興味喚起の場として、この重点領域研究は、ICOTにとっても渡りに舟でした。その年の12月には、金久先生がゲノム情報ワークショップの第1回目を企画し、ICOTと共催で開催しました。このワークショップでゲノムの話をはじめて聞いたコンピュータ研究者も多かったことと思います。その後、双方の分野の研究者は、当初のお互の専門用語や研究のフィロソフィーの違いを乗り越え、共通の興味を見出し、本格的なゲノムの解析や情報処理の研究を開始しました。

ICOTの研究チームも、この重点領域研究で知りあった分子生物学の研究者との共同研究を通じて、多くのことを学び、研究内容を充実させています。ゲノム情報のワークショップにおいても、並列推論マシンの上での配列解析や蛋白質の構造解析の研究成果の発表やデモも行ない、この分野の発展に貢献できるまじになりました。

ICOTの研究グループの発足の前後では、私は、コンピュータ関係者には、遺伝子情報処理がいかに面白いが、また、第五世代コンピュータの応用分野としていかに有望かを説いてまわりました。また、分子生物学の関係者には、コンピュータの利用、特に、データベースやネットワーク、パーソナルコンピュータ、それに、第五世代コンピュータの並列処理や知識処理の技術の利用が、分子生物学の発展に不可欠であることを説いたものです。

最近のゲノム情報の研究会やワークショップの発表を見ると、もはや、私が説いてまわる必要はなくなったことがわかります。東大医科研 ヒトゲノム解析センターの高木先生などの活躍により、新しいコンピュータ技術の重要性は分子生物学の研究者にもよく認識してもらえるようになったと思います。

ICOTの研究グループの成果も、当初の第五世代コンピュータプロトタイプの評価とデモを目的とするレベルを優に越えて、分子生物学の研究に直接的に貢献できるような成果もだせるようになりました。まさに、重点領域研究と歩調をあわせて発展してきたといえます。

現在、ICOTは、第五世代コンピュータ技術の普及を目的とする後継プロジェクトを実施しています。これにより、汎用のワークステーションや汎用並列マシン上でも第五世代コンピュータの言語やソフトウェアが使えるようになり、その結果、ICOTの遺伝子情報処理関連のソフトウェアや、その基盤となっている知識処理の技術が、広くゲノムの情報の解読に役立つ段階となります。

これは、第五世代コンピュータのソフトウェアが、評価、およびデモ用の段階を越えて、ゲノムの情報処理そのものを切り開く武器としての役割を担う段階に入ったことを意味しています。

新しい武器を使う時、その使い手はその武器に習熟することが求められ、また、武器の製作者は、使い手の意見を聞いて、絶えまなく武器を改良していくことが求められます。これからも、両方の分野の研究者が協力して、この新しい武器を磨き、ゲノム情報処理の分野をさらに発展させていって欲しいものです。

== Appendix =====

「ゲノム解析に伴う大量知識情報処理の研究」 第1回総括班会議 平成3年6月17日

第五世代コンピュータ技術の応用として見たゲノム情報処理

=====

内田 俊一（新世代コンピュータ技術開発機構、ICOT）

1. ICOTにおけるゲノム解析へのアプローチ

- 88/11 アルゴンヌ国立研究所で開催された「論理プログラミング(知識処理技術)を生物学的な解析に利用することによる新しい展開の可能性」に関するワークショップがきっかけ。
- 89/4 ICOTにゲノム関係の研究グループをつくることを計画。
(平成1)
- 89/7 ETLより、新田克己氏を迎え、GIP研究担当者を作り、勉強を始める。
同時に知識ベース管理ソフトウェアの研究グループ内に遺伝子DBの研究テーマを加える。
- 89/7 金久先生をリーダーに迎え、GIP-WGを作り、計算機科学と生物学を結びつけるための勉強を開始。
- 90/4 ICOT研究所組織変更 第7研究室にGIPの研究グループを作る。また、
(平成2) 第3研究室に、PSI上のDB管理システムKappaにGenBankなどを格納し評価。
- 90/6 海外研究機関との共同研究の立ち上げ:ANLへPSI設置(6月)、NIHへ設置(10月)、LBLへ設置(12月)その後、ANL、NIH、LBLの研究者が協力して、論理プログラミングを用いた実用的な解析システムの開発に着手。
21番染色体のDBと大腸菌のDB。これと並行して、蛋白質のフォールディングのシミュレーションも実施(NIH)
ICOTは、この箱庭的システムを見ながら、その大規模化したときの、第五世代コンピュータ技術を適用する方法や必要な機能の研究を実施。
- 90/12 知識情報処理とヒトゲノム計画WS開催
分子生物学者と計算機研究者との間の橋渡しを意図。
- 91/4 ICOTの研究内容が、模索の段階を通過し、明確化してきた。
(平成3) 1)「並列処理を用いた蛋白質の並列解析」-金久研、PERI
2)「並列処理によるフォールディングのシミュレーション」-NIH
3)「Kappaの非正規形DBの生物学的DBへの適合性の評価実験」
4)「知識表現言語、Quixoteを用いた化学反応を記述実験」

91/5 7研GIPグループ(6名)NIHにてMini-WS

開催テーマ:蛋白質のフォールディングシミュレーション

91/6 56シンポジウム(九段会館 6/19-20)にて、上記の1)、3)のデモと発表を実施。

2. 分子生物学的研究と計算機研究との橋渡しと問題点

1) 論理プログラミング言語による並列配列解析プログラム

記号処理が中心で、知識処理としては、レベルが高くない。並列推論マシンによる高速処理で、生物学の側に貢献できる。将来、より高度の知識をルールとして見出せたときに、そのルールをプログラムに取り込むことが、従来型の言語より、遥かに容易。

ICOTのマシンも、このレベルだと、生物学者にツールとして使ってもらえる完成度、信頼性に到達している。計算機研究者がプログラムを書けるくらい配列解析のアルゴリズムなどが確立している。知識(ルール)の作り込みは、共同でやる必要がある。

2) 並列処理によるフォールディングのシミュレーション

アミノ酸の疎水性/親水性の性質を利用しており、記号処理と比較的シンプルな知識を用いている。ICOTマシンの高速処理が貢献できる。しかし、アルゴリズムなど、生物学的な条件など未解明な部分多く、生物学者と密な共同研究が必要。計算機研究者も、生物側にのめりこまないとうまくいかない。生物学者のほうも、論理プログラミングや並列処理を勉強する必要有り。

3) 知識ベース、知識表現の利用による生物学的データの統合DB化

知識ベースや、知識表現自身、計算機分野でも、まだ十分枯れていない技術。また、化学反応など生物学的にみても、知識として他の化学反応や分子的な構造との関連が十分定着していない。しかし、生物学者は、この辺でも使えるツールが欲しいと希望する。気持ちはわかるが、この辺は、双方の先端分野を結ぶことになり、短期で実質的な成果やツールは得にくい。

既存のGenbanやPIRを、Kappaに格納して、情報検索的につかうのは知識のレベルが高くないところでの処理なので、実質的な成果は、出せそう。

分子生物学者と計算機研究者の間を結ぶときは、双方にとって先端的となる、うまいテーマ設定が必要。米国の研究者の場合は、既存の汎用機やワークステーション、ネットワークで結合された大データベースがあり、生物学者同しも、パブリックドメインソフトウェアやデータベースを使って少しでもよいツールを作ろうとしている。よいツールを作り公開して利用者が増えると、その開発者は評価され予算が増える。そして、また優れたツールを作るという拡大再生産の機構ができています。

ICOTもこの機構に加わることであり、まず、並列推論マシンという強力な記号処理、知識処理ツールを国際ネットワーク経由で提供し、汎用機に比べ有利な条件を

出したうえで、パートナーをさがし、このマシンとネットワークを用いた共同研究を行い、成果として出てくるソフトウェアをパブリックドメインのツールとして広めたいと考えている。

この場合、並列推論マシンはその上のソフトウェアと組み合わせられ、高速の配列解析サーバや、生物学的なDBのサーバとして機能するのが、当面の形態である。その後は、知的なソフトウェアが他のマシンより、容易に作成できる特徴をいかして、ネットワーク上に分散した解析システムの中核としての機能を果たすことを目指したいと考えている。