CICSJ Bulletin Vol.16, No.1, February 1998

目次

CICSJ INDEX にもどる

バイオインフォーマティックスの進展

－　特集にあたって　－

三菱化学・　横浜総合研究所　八尾　徹

yao@rc.m-kagaku.co.jp

先日、日本化学会の情報化学部会誌の編集委員の三戸さん（三井化学）から、「ゲノム情報解析」の特集をしたいので編成案を考えてほしいとのご依頼があった。化学系の研究者の方々にバイオ分野の研究動向やアプローチを知って頂く良い機会と考え、下記のような構成を作った。「化学」との結びつきを特に強調するのではなく、むしろあるがままにこのゲノム解析分野の最近のトピックスを各著者に紹介してもらう方が良いと考えた。

一つの生物種のゲノム情報には、その生命体を司る全遺伝子が含まれるという意味で、「生命の設計図」が書き込まれていると考えられる。これは従来のような個々のタンパク質や核酸の研究とは次元の違う研究を誘発する。長年かかってきた各生物種のゲノム情報解読が、一昨年のインフルエンザ菌(180万ﾍﾞｰｽ)の完了を初めとして、現在までに微生物で10数種類、古細菌で２種類、真核生物で１種類（酵母1250万ﾍﾞｰｽ)が終了している。引き続いて線虫、ショウジョウバエなども近く完了の予定であり、ヒト(30億ﾍﾞｰｽ)は2003年完了を目指して全世界で分担作業が進んでいる。

このような膨大なデータの蓄積と解析にはコンピュータが必須であり、今やバイオインフォーマティックスの重要性は日に日に高まって行っている。これまでの配列解析技術・立体構造解析技術を拡大して適用するのみならず、遺伝子間のネットワーク解析や、生物種間の比較解析など新しいアプローチも必要になって来ている。これらの状況を生々しく紹介して頂く。

先ず、東京大学医科学研究所のヒトゲノム解析センターの高木利久さんに「進化するゲノムデータベース」と題して、ゲノムデータベース開発の経緯と今後の動向を書いて頂く。特に、ゲノム計画を契機として開発が活発化したゲノム関連データベースが、統合化・再編成・インターネット化を経て大きな進展を遂げ、更に広範な生物医学データベースに変貌しつつある様子をご紹介頂く。

次に、大阪大学細胞生体工学センターの中井謙太さんには「ゲノム配列データ解析の最近の進歩」を書いて頂く。ゲノム配列にコードされた遺伝子の位置とその翻訳配列を予測する遺伝子同定法の最近の進歩と、転写制御情報の解読研究、ＤＮＡチップ技術によるタンパク質発現データの解析研究などにも触れて頂く。

次に、ヘリックス研究所のバイオインフォーマティックスグループの古谷さんとスウィンデルスさんに「配列情報からタンパク質立体構造・機能へ」と題して、ＤＮＡ塩基配列あるいはタンパク質アミノ酸配列から機能を予測するための種々の方法、特に立体構造予測を経由して機能予測へ至る過程を、成功事例を含めてご紹介頂く。

最後に、京都大学化学研究所の金久さんと坊農さんで「ゲノム情報から生命の設計図へ」と題して、ゲノム情報に含まれる遺伝子間の相互関連ネットワークを理解する研究について書いて頂く。金久グループで進められている代謝ネットワークデータベースの構築は生命の設計図解明のための重要なステップであろう。

追加として、欧米のこの分野での最近の顕著な動きを八尾がご紹介する。特に、ゲノム情報解析とタンパク質解析・計算構造生物学との融合について強調する。

やお とおる　Ｔｏｒｕ　ＹＡＯ
Tel.045-963-3123, Fax.045-963-3249

進化するゲノムデータベース

東京大学医科学研究所ヒトゲノム解析センター　高木　利久
E-mail: takagi@ims.u-tokyo.ac.jp
http://www.genome.ad.jp

ゲノム計画を契機として開発が活発化したゲノムデータベースは、ゲノム計画から産生されるデータはもちろんのこと、生物学、医学、薬学、農学などの急速に膨張しつつある生命科学諸分野から生み出される膨大な生命情報をも取り込みながら、統合化、再編成、インターネット化を経て大きく発展を遂げ、現在は広範な生命情報データベースに変貌しつつある。本稿ではゲノムデータベース開発のこれまでの経緯と今後の動向について簡単に紹介する。

1 はじめに

1980年代の終わりに端を発するゲノム計画の大幅な進展は、生物学医学を始めとした生命科学全般に大きな変革をもたらしつつある。この変革の波は創薬を始めとしたバイオ産業にも急速に及びつつあり、バイオ系企業が近年一斉に「ゲノムインフォマティックス」を担当する部門の整備・拡充に乗り出すというような状況も生まれている。

この変革の源の一つは、ヒトを始めとした各種生物のゲノム計画が産み出す膨大なゲノム情報にある。ここではそのゲノム情報を中核とした生命情報データベースについて、その開発の経緯と今後の動向について簡単に紹介する。

2 ゲノム計画とデータベース

生命科学関連のデータベースは、文献情報やアミノ酸配列情報など早くから収集が図られたものは、その歴史を1960年代までさかのぼることができる。その後、たんぱく質立体構造や染色体遺伝的地図に関するデータベースが1970年代に作られるようになった。核酸配列については1980年代初めにはデータベースが確立した。しかしながら、1990年代始めまでは、つまり、ヒトを中心とした各種のゲノム計画がスタートするまでは、データベースに納められているデータ量も少なく、また、データベースと照らし合わせて解析すべきデータの発生量も限られており、さらに、データの情報解析手段も限られていたこともあり、生命科学分野においてデータベースの利用は一部の利用者に限られていた。また、その当時はインターネットやデータベースの技術が未成熟であったこともあり、現在のようにデータベースが簡便に検索できるようにはなっていなかった。このことも、一般に計算機を苦手とする生物学者が積極的にデータベースを利用しなかった要因であった。

ところが、ゲノム計画が始まり、しかも、それが予想を上回るスピードで進展してくると、塩基配列、染色体地図、遺伝子発現プロファイルなどのゲノム情報が洪水のごとく産出され、いや

でもデータベースを利用しなければ研究できないような状況が生まれてきた。それと同時に、より使いやすい高度で統合化されたデータベースへの需要も大きくなり、それに呼応する形でデータベース開発が世界中で活発に進められるようになった。なお、データベース開発のニーズの高まりは、ゲノム計画の進展の他にも、近年の構造生物学を始めとした生命科学全般の急激な拡大とそれに伴う情報爆発による面も大きい。

3 データベースの進化

上に述べたように、ここ数年のゲノム計画を中核とした生命科学の進展と情報化が、データベースの利用と開発とを促し、それにより生命科学分野のデータベースは大幅な進展を遂げた。データベースがどのように変わったかを以下に簡単にまとめてみる。

3.1 統合化と再編成

ヒトゲノム計画が始まった頃のデータベースは、その開発の歴史的経緯や政治的背景などによりばらばらに管理運営されていたため、複数のデータベースにまたがるような検索は困難であった；データベースに格納されるデータのひとかたまり（エントリと呼ぶ）がおもに文献単位であったため、遺伝子単位やゲノム単位の解析をするには利用者自身がデータの編集処理を行わなければならなかった；などの問題を抱えていた。その上、データベースには精度の低い、あるいは、根拠の薄いデータが含まれていたりしたため、生物学的仮説の生成や検証に使うには不十分なものであった。

ゲノム計画を遂行するには、塩基配列データや地図データなどの狭義のゲノムデータを整理しデータベース化するだけでは不十分である。文献データはもちろんのこと、遺伝子発現プロファイルデータ、タンパク質の配列・構造・機能データ、家系データ、疾患データなどもあわせて整理し、それをゲノムデータと統合化してその中からゲノムの構造や機能に関する知見を得ることが不可欠である。さらに、統合化に際しては各エントリを再編成・再精製してゲノムや遺伝子といった生物学的に意味のある単位毎にまとめることが必要である。

そこで、ここ数年、データ量の爆発的な増大に対処する一方で、統合化と再編成に関する研究開発が世界中で精力的に進められてきた。オブジェクト指向ＤＢや演繹ＤＢなどの先進的なデータベース技術からのアプローチ、クライアント・サーバ方式による分散データベースによって統合化を図ろうとするインターネット技術からのアプローチなどいろいろな観点からの研究開発が行われてきたが、現状ではインターネット技術からのアプローチが主流となっている。これはインターネットの代表的なツールであるWWW のホームページ作成に使用されるHTML のリンク機能を使って同一データベース内あるいは異種データベース間の各エントリの間にリンクをはることによって統合化を図るものである。

現在このようなリンク機能を使って、種々の統合化の観点から多くのデータベースが構築されている。もし、このような統合データベースにご興味のある方はゲノムネットのホームページ（図1、http://www.genome.ad.jp）にアクセスして見ていただきたい。上で述べた統合化や再編成の例が、とくに日本で開発されたデータベースについて、ご覧いただけよう。

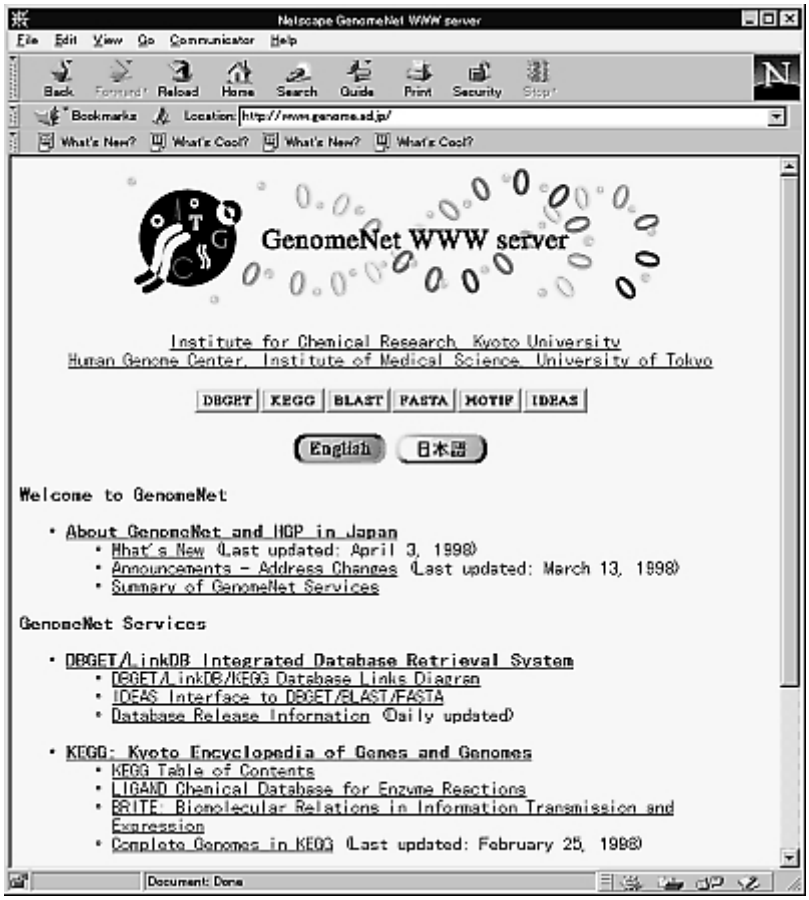


図１　ゲノムネットのホームページ

3.2 生命情報データベースへの変貌

ヒトゲノム計画は染色体地図作成の段階を終え、現在大規模シークエンシングの時代を迎えつつあり、2003年から2005年には全ゲノム配列が決定される見込みである。枯草菌や大腸菌などの微生物についてはすでに全ゲノム配列が決まったものも少なくない。このような研究の急激な進展を受けて、ゲノム研究の関心の一部は、個々の遺伝子の機能や遺伝子の発現調節ネットワークに、つまり、いわゆる機能解析研究に移りつつある。

このような状況に呼応してデータベースの方も新しいタイプのものが作られるようになってきた。従来、個々の遺伝子や個々のタンパク質についてはその構造や機能に関する分子レベルの情報がデータベース化されてきたが、最近では、以下に示すような新しい種類のデータベースが出現してきた。これらのデータベースはもはやゲノム解析のためのデータベースというより、生命科学研究全般のためのデータベースと呼ぶ方が相応しいものに進化しつつある。

1) 遺伝子ネットワークDB：細胞内の代謝系やシグナル伝達系などについて記述したもの。これは通常遺伝子や酵素およびそれらの相互作用などのデータベースと統合されるため、一般に遺伝子百科事典の名前で呼ばれている。

2) 反応／相互作用DB：酵素反応や転写因子などをまとめたもの。上の遺伝子ネットワークDBがその名のごとく分子反応のカスケードを納めているのに対して、こちらはおもに二つの分子の単一の反応を2項関係として記述している。

3) 変異DB：遺伝子やタンパク質の変異を集めたもの。疾患単位にまとめられたものと遺伝子やタンパク質を単位にまとめられたものとがある。

4) 比較解析DB：全ゲノム配列が決定された生物種について、それらの生物種間や同一生物種内のORF間の関係やそれをクラスタリングした結果を記述したもの。ORF間の関係としてはオーソログ（種分化の結果生じた相同遺伝子）やパラログ（遺伝子重複による同一ゲノム中に生じた相同遺伝子）などがある。

5)その他：上の4種類以外にも、タンパク質構造分類DB、ゲノムやタンパク質の2次元電気泳動DB、異常／選択的スプライスDB、免疫データベースなど多様なデータベースが最近開発されてきている。

これらの新しいデータベースについてはその数が多く、ここでとてもその内容を紹介することはできない。ご興味のある方は筆者らが開発したゲノム関連WWWサーバ検索システムCLUE（図2、http://clue.genome.ad.jp）を使ってどのようなデータベースが開発されているか調べてみていただきたい。

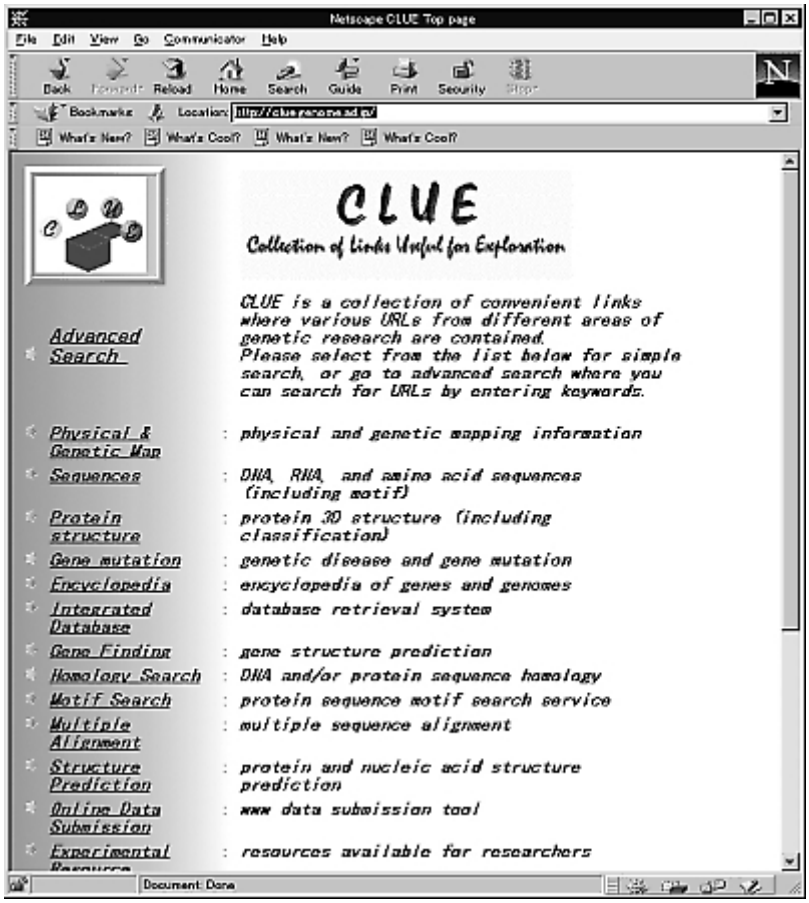


図２　ゲノム関連WWWサーバー検索システムCLUE

4 おわりに

これまで述べてきたようにゲノムデータベースは大きな進展を遂げ、多様で広範な生命情報データベースに変貌しつつある。それに伴い、データベースの利用者もゲノム研究者から生命科学全般の研究者へと急速に拡大してきている。このようなデータベース開発の流れは今後ますます

強まることが予想され、生命科学研究およびバイオ産業に不可欠の存在になって行くものと思われる。

参考文献

１）金久實編. シリーズ・ニューバイオフィジックス「ヒューマンゲノム計画」. 東京、共立出版、 1997.
２）高木利久、金久實編. ゲノムネットのデータベース利用法. 東京、共立出版、1996.
３）内山郁夫、高木利久. 「微生物ゲノム比較解析用データベースの開発」、蛋白質 核酸 酵素、 Vol.42、No.17、3046-3051 (1997).
４）高木利久. 「ゲノムデータベースから生物学医学データベースへ」、実験医学、Vol.16、No.1、(1998).

たかぎ　としひさ　TAKAGI, Toshihisa
ゲノムデータベース、ゲノム情報解析の研究に従事。
連絡先　〒108-8639　東京都港区白金台4-6-1　東京大学医科学研究所ヒトゲノム解析センター
電話　03-5449-5613

ゲノム配列データ解析における最近の進歩

大阪大学細胞生体工学センター　中井謙太
nakai@imcb.osaka-u.ac.jp

微生物のゲノム配列が次々と決定され、線虫やシロイヌナズナ、 そしてヒトゲノムの全塩基配列決定がいよいよ視野に入ってきた。 ゲノム配列にコードされた遺伝子の位置とその翻訳配列を予測す る、いわゆる遺伝子発見問題は、特にヒトなどの場合には当てに ならないと言われてきたが、最近はかなり精度の良いものも登場 してきた。また、転写制御情報解読を目指した研究も徐々に始ま りつつあり、DNA チップ技術などに基づいた大量データ産生と 相まって、今後の進展が期待される。

1　はじめに

最近のゲノムシークエンシングの進捗状況にはめざましいものがある。特に微 生物ゲノム解析の分野では、大腸菌、枯草菌、パン酵母など、古くから研究の進 んでいる種の配列決定が終了し、そこにコードされていると思われる遺伝子の組 織的な機能解析が進められている。それ以上に、これまであまり研究の進んでい なかった微生物種に対しても、まずゲノムの塩基配列を決定して、それをもとに 研究戦略を考える、という発想の転換をもたらしつつある。これはゲノム情報の コンピュータ解析の分野からみても、とてもチャレンジングな状況である。すな わち、「新しい生物のゲノム配列を与えられたとき、そこからどれだけの情報を 読み取ることができるか」という問題が、いよいよ現実のものとなってきたわけ である。このとき、最も強力な解析手段となるのは、大腸菌 K-12 株に代表され る、レファレンスデータベースとの比較であろう。しかし、技術的に見れば、そ の基礎になる配列比較法／相同性検索法の研究はすでにかなり成熟していると考 えられる（もっとも、最近の gapped BLAST の登場(1)のように、検索速度と検出 感度をバランス良く向上させていく試みは重要であるし、解析結果にも質的な向 上をもたらすことであろうが）。従って、私見によれば、配列解析の本領は、や はりデータベース中に類似配列が存在しない遺伝子について、第一にその位置を 正確に予測し、第二にその機能を予測するための材料を提供することにある。そ こで、本稿ではこれらの問題に関連する最近の話題を、筆者らの研究を中心に紹 介する。

2　遺伝子発見問題

ゲノム配列中にある遺伝子を探す "gene finding" の問題を考えるとき、最も生 物学的に妥当と思われる方法は、細胞内での認識方法にならって、転写開始点と 終結点を探す方法であろう。しかし、この方法は現在のところはあまり現実的で はなく、ふつう gene finding といった場合は、暗黙のうちに（アミノ酸配列の） コード領域を発見する問題と仮定されている。従って、RNA のままで機能する RNA 遺伝子は、既知のデータとの相同性がない限り、検出できないことになる。 もっとも、一般に細菌のゲノムは遺伝子間領域が非常に短いことが多いので、コー ド領

域を予測した後、十分に広い遺伝子間領域が残れば、RNA 遺伝子を含む候補 と考えることもできる。

細菌ゲノムの場合は、コード領域がイントロンで分断されていないため、同一 フレームで 100 アミノ酸程度以上終止コドンが入らない領域は、コード領域であ る可能性が高い。しかし、それだけでは短い遺伝子を見落としてしまうため、ア ミノ酸配列情報をコードしている領域は、塩基配列の統計的性質が遺伝子間領域 とは有意に異なっている（たとえば 3 文字ごとに周期性が現れるなど）という性 質を利用して予測を行う。典型的なプログラムとして、Borodovsky らの GeneMark が知られている(2)。また、矢田らによる GeneHacker は、統計の取り 方を改良したり、翻訳開始シグナル（SD 配列）の情報を取り入れるなどして、よ り高い性能を実現している(3)。

ヒトなどの高等生物では、遺伝子領域のほとんどをイントロンが占めることも めずらしくないため、遺伝子発見問題が非常に難しくなる。これまでにたくさん のプログラムが発表されてきたが、いずれも実用レベルに達しているとは言い難 かった。そのため、一時期 cDNA の断片配列を大量に収集した EST データベー スなどに対するホモロジー検索を組み合わせるアプローチが流行した。しかし、 昨年発表された GENSCAN というプログラムは、ホモロジー検索を使わなくても、 どのプログラムよりも高い性能を示したことで一躍有名になった(4)。GENSCAN がなぜそれほどの成功を収めたのかについては諸説があるが、後述の HMM（隠れ マルコフモデル）の考え方をベースにして、丁寧に遺伝子構造のいろいろな特徴 をモデル化したことが大きいと言われている。逆に言えば、これまでの限界を打 ち破るような革新的なアイディアが盛られているわけではないので、今後の発展 に対しての楽観は禁物である。

3　遺伝子の機能予測

ゲノムにコードされたアミノ酸配列情報を読み取ることができれば、そのポリ ペプチドがどのような立体構造をとるかを予測し、それに基づいて機能を予測す ることが、次の論理的な目標になる。しかし、いわゆる立体構造予測が非常に難 しいことは広く知られているし、たとえおおまかな立体構造が予測できたとして も、それをもとに、相互作用する相手を予測したり、自分が触媒する酵素反応を 予測したりすることは、現時点では夢のまた夢である。それではデータベースに 相同なタンパク質のデータがない限り、タンパク質の機能予測は不可能なのであ ろうか。筆者らは、直接機能そのものがわからないまでも、機能推定の手がかり になるような情報を引き出すことは可能と考えて、以下に紹介するように、アミ ノ酸配列情報と、その発現を制御する塩基配列情報の両面から研究を行っている(5)。

3．1　タンパク質の細胞内局在部位予測

真核細胞内部には、さまざまなオルガネラが存在し、それぞれ細胞内の機能を 分業している。従って、タンパク質が細胞内のどのオルガネラに局在しているか を知ることができれば、そのタンパク質の機能を推定する役に立つ。一方、タン パク質は、ほとんどの場合、細胞質内で生合成された後、自身のアミノ酸配列中 に書き込まれた局在化シグナルの情報に従って細胞内で仕分けされ、最終目的地 へ輸送される。従って、少なくとも原理的には、アミノ酸配列中の局在化シグナ ルを検出することで、タンパク質の細胞内局在部位

を予測できるはずである。筆 者らは、さまざまな局在化シグナルの配列上の特徴を知識ベース化して、それを もとに任意のアミノ酸配列の細胞内局在部位を予測するプログラム PSORT（ピー ソート）を構築し、インターネットを通して公開してきた(6)。PSORT は、その予 測能力自体にはまだまだ改善の余地があるものの、世界的に見てもユニークなシ ステムとして、多くの利用者に支持されてきた。 しかし、PSORT プログラムも、最初の発表以来年月を経るにつれて、解決され なければならない問題を無視できなくなってきた。問題は多岐にわたるが、技術 的に超えなければならない最大の障壁は、システムの最適化の問題であった。 PSORT プログラムの動作は大きく二つに分けることができる。前半は、与えられ た配列に対して、いろいろな配列上の特徴をチェックする小プログラムを走らせて、その配列を特徴づけるスコアの集合を用意する部分で、後半は、そのスコア の集合をもとに、それぞれの候補部位に局在する確率を評価する部分である。 「システムの最適化」とは、後半部分の計算法を、与えられた局在部位既知のア ミノ酸配列の集合から導くことを意味する。従来のプログラムは、古典的な "if-then" 型のエキスパートシステムとして構築されていたこともあって、この部 分が弱く、前半のスコア計算プログラム群（いわゆる知識ベース）を改良したり、 局在部位既知の学習データを拡張したりしたいときに、手作業でパラメータを調 節しなくてはならなかった。カリフォルニア大学バークレー校の Paul Horton 氏 がいろいろな機械学習のテクニックを検討した結果、スコアの集合をベクトルと みなして、ベクトル間のユークリッド距離によって、k-nearest neighbor 法を適 用した場合に一番高い予測性能が得られることが示された(7)。現在、この方法を用 いて、perl 言語で書き直したプログラムを PSORT II と称して、試験的に WWW で公開している。近いうちに、細部をブラッシュアップして、予測性能を客観的 に測定した後、正式公開する予定である。

３．２　遺伝子の発現情報による機能予測

遺伝情報の具体的内容は、詰まるところ、どんなタンパク質をどういう条件で 合成するか、ということに尽き、後者の条件は、主に転写のレベルで制御されて いる。しかし、従来のゲノム情報解析は、この後者の制御情報の解読に関してほ とんど無力であった。その最大の原因は、制御配列のデータと対応する制御内容 のデータの双方が不足していたためであったが、今やその状況は変わりつつある。 特に、DNA チップやマイクロアレイと呼ばれる、１センチ四方程度の個体表面上 に高密度に任意のオリゴヌクレオチドプローブを固定する技術の進歩によって、 遺伝子の発現情報の組織的な解析が、これから数年のうちに飛躍的に進歩するこ とが期待される。

そこで筆者らは、科学技術振興事業団の矢田哲士氏らと共同で、実験的研究が 進んでいて、比較的メカニズムが単純な、大腸菌と枯草菌の転写制御情報の解釈 を試みた。原核生物の RNA ポリメラーゼには、プロモータ配列を認識する部分が シグマ因子というサブユニットとして存在する。このシグマ因子は一般に複数種 類存在し、それぞれが独自のプロモータ配列特異性を示す。細菌類はこれを転写 制御に利用している。たとえば、細菌に熱ショックを与えると、特別なシグマ因 子が合成され、それが RNA ポリメラーゼに結合することにより、それまでとは異 なる一群の遺伝子が活性化される。従って、ある遺伝子がどのシグマ因子に支配 されているかを知ることができれば、その遺伝子がおおざっぱにどのような条件 で発現するかを知ることができ、そのことはその遺伝子の機能を推定する上で、有力な手がかりになり得る。都合の良いことに、すべての遺伝子は、少なくとも １つのシ

グマ因子に支配されているはずであるし、この予測はコード領域部分と は独立に行えるので、データベースに類似配列が存在するかどうかには関係なく ものが言えるわけである（もっとも、いわゆる主要因子に支配されている場合に は、あまり大した情報は得られないが）。また、シグマ因子依存性は、実験によっ て検証可能な性質であり、今後の機能解析プロジェクトでも取り上げられるべき テーマの一つであることも、この先予測精度を向上させていく上で心強い。具体 的な予測方法としては、まず文献からそれぞれのシグマ因子が認識するプロモー タ配列を収集して、多重アラインメントし、それをもとに、各々のシグマ因子の プロモータ認識能力をシミュレートする HMM を構築した。詳細は、枯草菌につ いての我々の報告を参照されたい(8)。なお、大腸菌については、国立遺伝学研究所 の石浜明教授の協力を得て、大学院生の石井崇洋氏が in vitro の検証実験を行っ ていることも付記しておく。

一方、真核生物の RNA ポリメラーゼは、シグマ因子をもたず、転写反応は複雑 な転写因子の組み合わせで制御されていると考えられている。筆者らは、とりあ えず矢田らが開発した、YEBIS という高性能モチーフ検出プログラム(9)を使って、 類似メカニズムで制御されていると思われる酵母のプロモーター群に頻出するモ チーフの検出に取り組んでいる。この試みが実を結べば、1998 年には全配列が決 定される予定の線虫ゲノムにも挑戦して、DNA 配列から遺伝子の発生時期特異的・ 組織特異的発現を予測する問題にも取り組んでみたいものである。

4　おわりに

与えられたゲノム配列にコードされた遺伝子をすべて発見して、その発現条件 やタンパク質としての機能を知ることができたとしても、もちろんそれでその生 物のことがすべてわかるわけではない。次のステップは遺伝子間の相互作用を通 して、さまざまな表現形を理解していくことであろう。これはつまり、生命のシ ミュレーションであるが、微生物や線虫・ショウジョウバエなどについては、近 い将来、そういうものが必ず必要になる。そのとき、配列解析はどの程度貢献で きるだろうか。転写制御のカスケードであれば、プロモータ解析が相応の貢献を する可能性はあるだろう。しかし、タンパク質間相互作用の予測は、それが立体 構造予測を間にはさむために極めて難しくなる。おそらく、類似種のゲノムをど んどん配列決定することで、comparative genomics をやるのが一番手っ取り早い かもしれない。それ以外にも、将来は、工業的に望ましい制御配列を設計する問 題などもおもしろいテーマになるものと思われる。

参考文献

1) Altschul, S. F. ; Madden, T. L. ; Schaffer, A. A. ; Zhang, J. ; Zhang, Z. ; Miller, W. ; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402 (1997).
2) Borodovsky, M. ; McIninch J. GeneMark: Parallel Gene Recognition for both DNA Strands. Computers & Chemistry, 17, 123-133 (1993) .
3) Yada, T. ; Hirosawa, M. Gene Recognition in Cyanobacterium genomic sequence data using the hidden Markov model. Intellig. Syst. Mol. Biol., 4, 252-260 (1996). Yada, T. ; Hirosawa, M. Detection of short protein coding regions within Cyanobacterium genome:

Application of the hidden Markov model. DNA Res., 3, 355-361 (1996). http://www-scc.jst.go.jp:8080/sankichi/GeneHacker/
4) Burge, C. ; Karlin, S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268, 78-94 (1997).
5) 中井謙太. ゲノム機能予測における知識ベースアプローチ, 榊、金久、中村、 大木、小原、高木編　ゲノムサイエンス：生命の全体像の解明をめざして, 蛋白 質核酸酵素 増刊号, 42, 3001-3007 (1997).
6) Nakai, K. ; Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics, 14, 897-911 (1992).
7) Horton, P. ; Nakai, K. Better prediction of protein cellular localization sites with the k nearest neighbor classifier. Intellig. Syst. Mol. Biol., 5, 147-152 (1997). http://www.imcb.osaka-u.ac.jp/nakai/psort.html
8) Yada, T. ; Totoki, Y. ; Ishii, T. ; Nakai, K. Functional prediction of B. subtilis genes from their regulatory sequences. Intellig. Syst. Mol. Biol., 5, 354-357 (1997).
9) Yada, T. ; Totoki, Y. ; Ishikawa, M. ; Asai, K. ; Nakai, K. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. Bioinformatics, in press (1998).

なかい　けんた　NAKAI, Kenta
好きな小説家は John le Carre と高村薫。たまに研究室でリコーダーを吹き鳴ら してひんしゅくをかっています。
〒565-0871 大阪府吹田市山田丘 1-3 大阪大学 細胞生体工学センター

ゲノム情報から生命の設計図へ

京都大学化学研究所　坊農秀雅、金久實
{bono,kanehisa}@kuicr.kyoto-u.ac.jp

１　はじめに

1995年に単独で生きることのできる生物として始めてヘモフィルスのゲノムが決定された。それからわずか2年余りの間に約10生物種程度の全ゲノムが決定されてきた。ゲノム解析は生命系を構成する部品としての遺伝子、または分子を個別に調べる手法から、分子間相互作用やそのつながりである分子間ネットワークをシステマティックに解析する方向へと移行しつつある。この分子間相互作用、すなわち生命系を構成する部品と部品のつながりを系統的に調べることができれば、そこから生命の設計図を読み解くことが可能である。

２.１ゲノム解析の現状

ゲノム解析は、ゲノムの中にはすべての遺伝子が含まれているのだからゲノムの塩基配列を端から端まで決めてしまってそこから遺伝子を同定して機能を推定するというストラテジーをとる。これは遺伝子個々の機能を知るためには、いずれにせよ個々（遺伝子）の「かたち」を調べる必要があるのだから、先にすべて（ゲノム）の「かたち」を決めておいたほうが効率がよいという考え方に基づいている。ともすると忘れられることであるが、遺伝子の「かたち」を調べるのは生物の表現形つまり「はたらき」を調べるためである。

表１に示された生物種のゲノムが現時点で決定されている。生物のゲノム配列を決定すると言っても何もヒトのゲノムのみに限ったことではない。ゲノムサイズが小さいこと（例えば *M.genitalium*）、病原菌であり医学的に価値があること（例えば *H.pylori*)、商業的な応用に役に立つこと（例えば *S.cerevisiae*）などの理由により、様々な生物種のゲノムシーケンシングが行われている。シーケンス技術に関してはここでは述べないが、ゲノムスケールのシーケンシングに向けた大量シーケンス技術が確立されつつある。文献1などを参考にされたい。ゲノムシーケンシングの技術革新により、ヒトのゲノムは西暦2005年までにはすべてシーケンスされると言われている。

このようにゲノムの「かたち」を決めるほうはある程度のめどがついているが、その「はたらき」を調べることは始まったばかりである。機械的に生産されてくる塩基配列情報（ゲノムの「かたち」）は大量で人間の処理能力をはるかに越えるものとなっている。そこで、コンピュータを用いた解析に期待が寄せられている。

２.２ゲノム情報解析

現在、新たな塩基配列が決定されると常套手段として用いられているコンピュータを用いた解析方法は主に以下の３つである。

1. 遺伝子（コード領域）を見つける (Gene Finding)
2. 遺伝子（主にタンパク質）の機能を割り当てる (Gene Assignment)

もしある生物種のゲノムすべてが決定されたときには

3． 複数の遺伝子セットから生命系（たとえば代謝系）を再構築する (Functional Reconstruction)

ことも最近行われるようになってきた。現在のところ、まだ精度に欠けるものであるが、急ピッチで精度向上の研究が進められている。世界各地（おもにアメリカ合衆国であるが）のゲノムセンターではオートシーケンサーから読まれた塩基配列に対してすぐさまコード領域予測プログラムをかけ、さらに予測されたコード領域（エクソン）に対して FASTA/BLAST といったホモロジーサーチプログラムを用いてデータベース中に似た配列がないか探し、新しく見つかった配列の機能を割り当てている。もちろん、データベース中に新しく見つかった配列と似た配列の情報がないと機能を割り当てることは不可能だし、たとえ似た配列が見つかっても機能未知であったりと、そううまく機能を割り当てることはできないのが現実である。大腸菌のゲノムが昨年すべて読み解かれた[2]が、モデル生物として実験室でよく用いられよく調べられているはずの大腸菌でさえも約4000の遺伝子中、約4割はまだ機能が全く推定できない[3]遺伝子であった。生物はまだまだわからないことだらけということである。

Archaea	*Archaeoglobus fulgidus*	2,178,400	TIGR
	Methanobacterium thermoautotrophicum	1,751,298	GTC
	Methanococcus jannaschii （メタン生成菌）	1,664,977	TIGR
Eubacteria	*Borrelia burgdorferi*	910,724	TIGR
	Escherichia coli （大腸菌）	4,639,221	Wisconsin, Japan
	Haemophilus influenzae （ヘモフィルス）	1,830,135	TIGR
	Helicobacter pylori （ピロリ菌）	1,667,867	TIGR
	Bacillus subtilis （枯草菌）	4,214,814	BSORF, SubtiList
	Mycoplasma genitalium	580,073	TIGR
	Mycoplasma pneumoniae （肺炎マイコプラズマ）	816,394	ZMBH
	Synechocystis sp.（ラン藻）	3,573,470	Kazusa
Eukaryote	*Saccharomyces cerevisiae* （出芽酵母）	12,069,313	SGD/MIPS

表１　ゲノムの決まった生物種 （最新の情報は http://www.genome.ad.jp/kegg/java/org_list.html にて公開）

3　部品の解析からシステムの解析へ

ゲノムシーケンシングの次のステップとして遺伝子の発現情報も解析されている。しかし、遺伝子が持つ配列情報と発現情報がすべてわかればゲノム解析は終了するのだろうか。個々の部品

（遺伝子）がわかっても生命のシステム全体がわかるわけではない。部品がどのように組み合わさってシステムが構成されているかがわからなければ、生命現象を理解したことにはならないのである。

ゲノム解析からポスト・ゲノム解析へ向かう重要なステップとして部品間の結線図に相当する遺伝子相互作用やタンパク質間相互作用をシステマティックに解析し理解していくことが必要である。すでにゲノムの決まった出芽酵母では 6000 の遺伝子の系統的な遺伝子破壊による解析や異なる2つのタンパク質間相互作用を系統的に調べる研究が開始されている。つまり「かたち」に重点を置いた分子生物学から、生物の「はたらき」に向けたマクロな分子間生物学へと移行しつつあるとも言えるである。

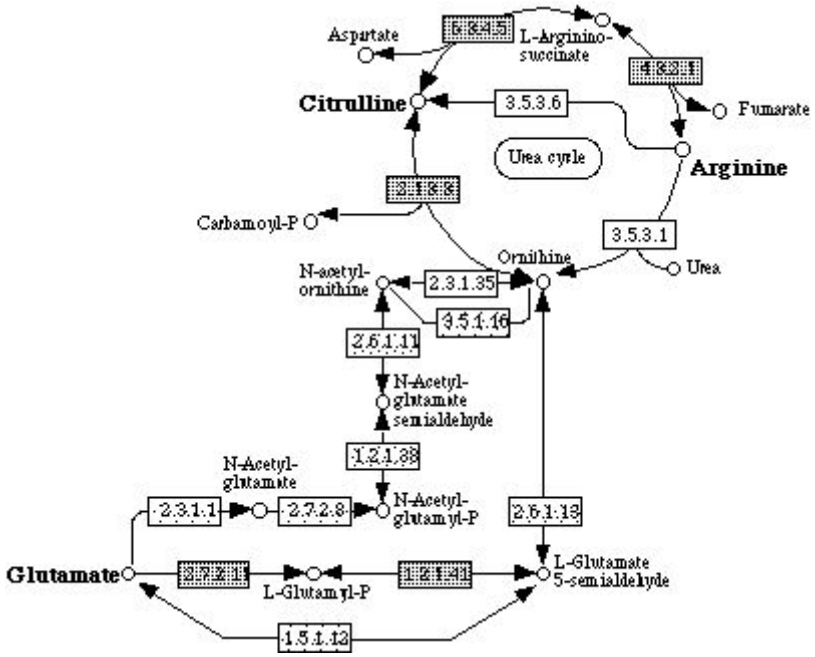


図1　カタログ化された Arginine 合成パスウェイ（大腸菌とヘモフィルスの両方で見いだされた遺伝子は濃い影、大腸菌でのみ見いだせた遺伝子は薄い影で示した）

４　遺伝子・ゲノム百科事典(KEGG)

既存の分子生物学関連データベースはDNAの塩基配列(GenBank, EMBL)やタンパク質のアミノ酸配列(PIR, SwissProt)や立体構造(PDB)といった個々の分子の「かたち」の情報にすぎない。これに対し、これまで個々の生物、組織、細胞で明らかにされてきた「はたらき」の情報のデータベースはない。我々はこの「はたらき」をデータベース化すべく、KEGG (Kyoto Encyclopedia of Genes and Genomes) プロジェクト[4]を開始した。KEGG プロジェクトの主な内容は以下の3点である。

１．生命系の機能カタログを分子生物学、細胞生物学の広範な知識を分子のパスウェイ情報としてコンピュータ化し、生命系の機能カタログを構築すること。
２．すべての生物種で、ゲノム解析がもたらす遺伝子カタログの各部品（遺伝子産物）を構築した機能カタログ上に対応づけること。

3． データベース化した相互作用データから可能なパスウェイを計算したり、ポスト・ゲノム解析に伴う新しい情報技術を開発すること。

現在、KEGG は完全にゲノムが決まった12種類の生物を含めて約20種類の生物種の代謝系や一部の制御系（シグナル伝達や細胞周期など）を公開している。図1は Arginine 合成経路に関する生化学の知識をコンピュータ化したものである。四角の箱は酵素（すなわち遺伝子産物）を示し、中に書かれた番号は酵素番号（EC 番号）である。また影のついた四角は大腸菌とヘモフィルスのゲノムから見いだされた遺伝子に対応している（図の脚注参照）。パスがつながってきちんと合成系が構成されているかによって、逆に各遺伝子の機能予測が正確になされているかがわかる。この例の場合、大腸菌は Glutamate から Arginine を合成することができるとわかるが、ヘモフィルスは Glutamate から Arginine を合成できるかこの図からだけでは不明である。ヘモフィルスの場合のようにパスが切れていれば、ゲノムにある遺伝子を見落としている（つまり遺伝子予測が不適切）か、あるいは別の反応経路が存在するかの可能性が考えられ、それぞれ計算で調べることができる。図2は後者の別の反応経路が存在するかどうか、 KEGG のホームページ (http://www.genome.ad.jp/kegg/) から WWW でアクセスし調べている例である。ここで使われている計算の基礎にあるものが二項関係という概念である[5]。図1の代謝経路を例にとって説明すると、酵素反応を基質と生成物の二項関係で表し、複数の基質や生成物があるときは二項関係のリストで表現する。2分子の関係のみでシンプルに表現するところが普通の酵素反応式と異なる点である。図1ではヘモフィルスの Arginine 合成パスウェイは見つけられなかったわけだが、ヘモフィルスに存在する酵素反応の二項関係をもとにパス計算を実行するといくつか可能なパスウェイの候補が見つかる。しかし、この場合 Glutamate から Arginine に至るパスウェイの中間代謝物の Citrulline がヘモフィルスの成長に必要であると指摘されており[6]、残念ながら計算されたパスウェイはメインに使われているパスウェイではないらしい。

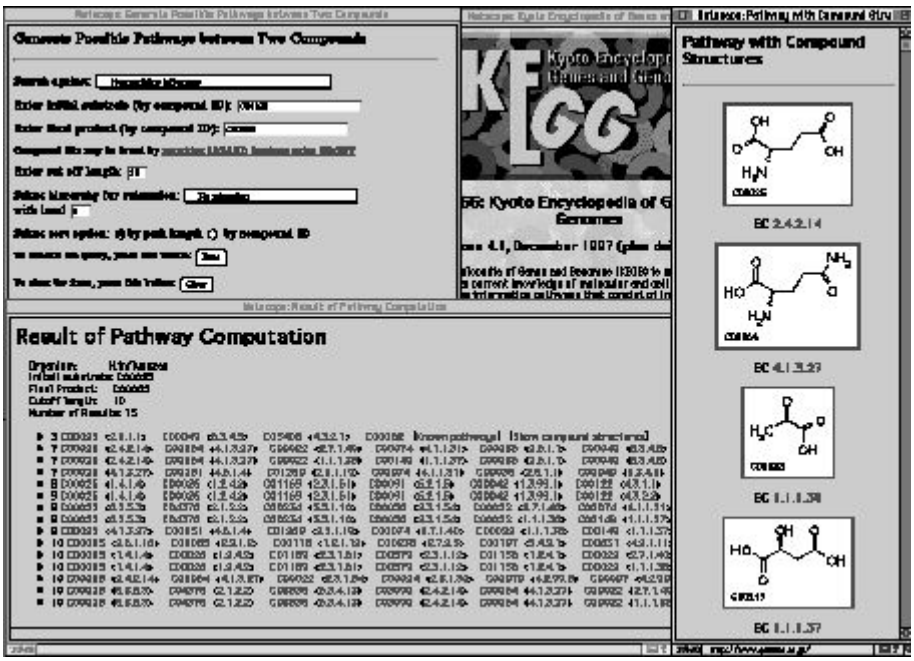


図2　Glutamate(C00025) から Arginine(C00062) に至るパスウェイをヘモフィルスで検索している画面（左側で可能なパスウェイを計算、右側でその一つの酵素番号と中間代謝物の連なりを表示している）

5　パスウェイから生命の設計図へ

　図2でやっていることは新しく決定されたゲノムの酵素遺伝子を集めて既知の代謝パスウェイが再構成できるかどうか、つまり蓄えられた「関係」のデータから、それらを「演繹」して新しい関係を導きだしていることにほかならない。しかし、現段階では既知の生物学的知識はこのようにコンピュータ上で利用できる形にはなっていない。知識同志を組み合わせて推論することがコンピュータ上でできるように知識のコンピュータ化を計っていく必要がある。この生物学的知識のコンピュータ化の作業があって始めて、部品からシステムの構築、あるいはシステム同士の比較といったポスト・ゲノム時代の情報処理が可能となるであろう。

生命の設計図が書けるぐらい、生物学は物理学や化学のように原理的な面からの解析ができていないのが現状である。しかしヒューマンゲノム計画により、生物学はその歴史が始まって以来初めてシステマティックにデータが得られるようになってきた。今後はさらに劇的な速さでデータ量が増していくことが見込まれている。このような状況下で大量のデータを解析し、さらにそこから生まれた大量の知識をもとに論理的な計算をコンピュータに行わせて、生命の設計図を読み取っていくことが要求されている。「はたらき」の情報を組み合わせて、生命の設計図をヒトが描けるようになるのもそう遠い未来のことではないのかもしれない。

参考文献

1. 榊 佳之 他編：「ゲノムサイエンス」蛋白質核酸酵素1997年12月号増刊
2. Blattner, F.R. et al：The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1474 (1997).
3. Moxon, E.R. and Higgins, C.F.：A blueprint for life. *Nature*, **389**, 120-121 (1997).
4. Kanehisa, M.：A database for post-genome analysis. *Trends Genet.*, **13**, 375-376 (1997).
5. Ogata, H. et al：Analysis of binary relations and hierarchies of enzymes in the metabolic pathways. *Genome Informatics 1996*, 128-136, (1996).
6. Tatusov, R.L. et al Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Current Biol*. **6**, 279-291 (1996).

ぼうのう　ひでまさ　BONO, Hidemasa;　かねひさ　みのる　KANEHISA, Minoru
連絡先 〒611　京都府宇治市五ケ庄　京都大学化学研究所
電話 0774-38-3270 FAX 0774-38-3269
WWW http://www.kuicr.kyoto-u.ac.jp/kanehisa/

計算構造生物学とゲノム情報解析の融合

－－欧米の動向－－

三菱化学・　横浜総合研究所
（兼）ＭＳＩ／ＴＭＳＩ顧問　八尾　徹
yao@rc.m-kagaku.co.jp, yao:msi.com

１．はじめに

一昨年のインフルエンザ菌を初めとして、最近次々と各生物種ゲノムの全塩基配列が解 読されてきた。この膨大な配列データベースの中から意味のある機能部位を同定し、生命 現象の鍵を見出そうとする、いわゆる「バイオインフォーマティックス」が急速に重要に なってきた。

一方、タンパク質の機能をその立体構造を介して深く理解しようとする「構造生物学」 は、近年ますます重要視されて来ており、欧米ではここ数年その推進策が図られて来てい る。構造生物学の具体的な手段としては、Ｘ線結晶構造解析、ＮＭＲ溶液構造解析、更に は極低温電子顕微鏡、ＡＦＭなどがあるが、これらに加えてコンピュータによる立体構造 予測、モデリング、フォールディングシミュレーション等も有力な手段になりつつある。

上記ゲノム情報の解析に、この立体構造を介した機能予測が大きな貢献をすることは明 らかであり、ここ１～２年にタンパク質の立体構造データベース解析をしていた研究者が 相次いでゲノム情報の解析分野に殺到している。正に、計算構造生物学とゲノム情報解析 の融合が始まっている。

更に、これまで開発された各種のタンパク質解析手法を自動化しようとする動きが活発 である。膨大なゲノム情報を解析するには、従来のように個々に慎重に解析するのでは、 とても追いつかないからである。コンピュ・タによる第一次スクリ・ニングという性格を 帯びている。今後、データベースの増大と共にその割合は高まるだろう。一方、新たな課 題として「比較ゲノミックス」と「ゲノムネットワ・ク」が重要になって来ている。　これらの動向を、欧米の学会・シンポジウムや研究機関の例を取り上げてご紹介する。

２．立体構造解析の進歩

最近の学会や研究機関の動向を述べる前に、この１０年位のタンパク質立体構造解析の 進歩をかいつまんで述べておく。（表１、表２参照）

1）タンパク質の立体構造データは、１９８６年当時 300件位であったが、 現在は6000件を越している。一方、アミノ酸配列データは、当時の約3000件から 現在１０万件を越す勢いにある。従って、その格差は開く一方である。

2）タンパク質の立体構造決定法の中心はＸ線結晶構造解析法であり、そ の解析の対象・精度・速度は非常に向上した。しかし、結晶化のボトルネックは 依然解消されていない。１０年ほど前から可能になった核磁気共鳴（ＮＭＲ）に よる溶液構造解析法はその後格段の進歩を遂げ、現在では毎年決定される 新しい構造の約２０％を占めるようになった。しかし、大きいタンパク質は解析出 来ない。又、極低温電子顕微鏡で二次元結晶を用いて膜タンパク質の構造決 定が出来るようになったが、まだ数個決定された程度である。

3）上記立体構造データの増大に伴い、そのデータベース解析から立体 構造形成の原理を経験的に把握しようとする研究が台頭してきた。即ち、部分 構造や全体構造の特徴把握や分類が進み、更に、予測に使うためのパラメー タ抽出なども行われている。

4）タンパク質のフォールディング機構とその中間状態を実験的に調べる 技術が進み、その描像についての詳細な議論がなされるようになって来た(1)。

5）タンパク質のフォールディング過程をコンピュータシミュレーションしよ うとする研究も行われているが(2)、現在は部分構造位までしか成功していない。

6）アミノ酸配列から立体構造予測をすることはまだ出来ていないが、実用 的な意味では下記の二つに見るべき進歩があった。

- ホモロジーモデリング法(3)－－配列ホモロジーのあるタンパク質の既知 立体構造を基に推定構造を構築する方法で、この１０年間に大幅な進歩があ り、基礎研究にも医薬分子設計などの応用面にも大いに役立っている。
- 構造認識法(4)－－数年前に新たに提唱された方法で、配列ホモロジー がなくても立体構造パターンの同じものがある場合に有効であり、成功事例が 出て来た(5)。これらはいずれも今後立体構造データ数が増えるに従って適用 範囲が増大する。

7）数年前タンパク質立体構造パターンの有限説が提唱され(6)、かなりのコン センサスをもって受け入れられている。このことは、構造解析・構造予測・設計な どに多大のインパクトを与えている。Ｃｈｏｔｈｉａによれば立体構造パターンは1000 タイプ位だろうとの予測であり、現在の立体構造データベースＰＤＢ6000個を分 類すれば 300タイプ位とみなされる。従って今後構造解析が進みタイプ数が充 実するにつれて、上記6)の二つの方法の有効性はますます上がってくると予想 される。

表１　構造生物学／計算構造生物学

構造生物学	生体分子の立体構造を通して、生命現象の深い理解に迫ろうとする研究分野	
・測定機器による立体構造解析 Ｘ線結晶構造解析 ＮＭＲ溶液構造解析 電子線結晶構造解析 中性子線解析 ＡＦＭ解析ほか		
・コンピュータによるデータベース解析－－データ数の増大 特に、立体構造データベース解析 二次構造能の特徴、ターン構造、内外性、コア構造 超二次構造、側鎖構造 全体構造、立体構造の分類、相互作用の特徴 ・分子シミュレーション　－－構造・機能・フォールディング 原子レベルのシミュレーション　　－－分子動力学法 アミノ酸レベルのシミュレーション　－－格子モデル ・立体構造予測 二次構造予測、内外性予測 ホモロジーモデリング 構造認識法 Ａｂ－Ｉｎｉｔｉｏ予測		**計算構造生物学**

表２　タンパク質立体構造解析とゲノム情報解析

１９８６年	１９９７年
アミノ酸配列ＤＢ　３０００　------>　１００００ 立体構造ＤＢ　２５０　------>　６０００ 立体構造データベース解析 ----> 二次構造、ル・プ構造 ----> 部分構造、側鎖構造、相互作用 「ホモロジー３０％以上 なら立体構造類似」　ホモロジーモデリング法	

「立体構造パターン　----〉　立体構造の分類
有限説」

--------〉　構造認識法

立体構造予測コンテスト
ＣＡＳＰ１
ＣＡＳＰ２

ゲノム情報の解読　------〉
バイオインフォーマティックス

３．国際学会・シンポジウムでの話題

最近、欧米でのタンパク質関連の国際学会・シンポジウムでは、急速にゲノム情報解析とのつながりに関する話題が増えて来た。この背景には、１９９５年のインフルエンザ菌の全ゲノム解読(7) に対するタンパク質データベース解析側からの大きな貢献の実績(8) が影響していると考えている。タンパク質解析の技術が、膨大なゲノム配列データに隠されている情報を引き出すことに、非常に役立つとの認識が高まっている。 以下に、最近のいくつかの事例をご紹介する。

1) 欧州蛋白工学国際シンポジウム　　(1996年 3月 3-6日、仏国モンペリエ 450名, 245件)

ヨーロッパの蛋白工学関連の研究者の集まりで、研究内容の発表と研究協力・施策などが議論された。主題は「タンパク質の構造から機能へ」(From Folds to Functions) で、その中のトップのテーマが「バイオインフォーマティックス」であった。

酵母ゲノムの全塩基配列が５年間10ケ国で分担して解読された過程の紹介(A.Golfeau)と、それに対するバイオインフォーマティックス解析の内容紹介(C.Sander)があった。

Ｓａｎｄｅｒは、長年タンパク質の立体構造データベース解析・立体構造予測・設計などに携わり、多くの実績を持つ研究者である。彼が最近ゲノム情報解析分野に転身したことは、非常に大きな意味がある。

ここで彼は、ホモロジーサーチ、アラインメント、系統樹作成などの配列データ解析に加えて、立体構造予測の各種手法（二次構造予測、内外性予測、ホモロジーモデリング、構造認識法など）及び機能予測法(PROSITEなど) を駆使して、実験のみでは得られないコンピュータ解析の威力を示した(9)。

2) 米国生物物理学会シンポジウム (1997年 3月 2-6日、米国ニューオリンズ 3,600名, 2,500件)

生物物理学のあらゆる分野を、９つの分野、約１００のカテゴリーに分類して講演・ポスター発表が行われた。ここでも今回「ゲノム配列情報の立体構造ベースの解析」という特別シンポジウムが開かれたことは特筆すべきであろう。

タンパク質静電ポテンシャル計算(DELPHI)で有名なＢ．Ｈｏｎｉｇ教授の司会で、アルゴンヌ国立研究所で開発されている総合的ゲノムデータベース解析システムＭＡＧＰＩＥの紹介に続いて、Ｓ．Ｂｒｙａｎｔ(NCBI/NIH)から構造認識法の最近の進歩と配列データ解

析とのリンクについて、更にＡ．Ｓａｌｉ (Lockefeller Univ.)から自動ホモロジーモデリングシステムＭＯＤＥＬＥＲをゲノム情報に適用することについて話があった。

Ｓａｌｉは昨年全塩基配列が解読された酵母ゲノムにこの手法を適用している。３０％以上ホモロジーがあれば自動モデリングが出来、ＮＭＲ構造以上の精度が得られること、ゲノム配列の約２０％が対象となることを強調した。

タンパク質の理論計算屋であるＨｏｎｉｇ自身も、自分のグループの約半分は、今後ゲノム情報解析に当てると言っていた。

尚、この学会ではＮＩＨ，ＮＳＦ，ＤＯＥなどが構造生物学推進の施策を取っていることも紹介された。

3) 欧州蛋白工学国際シンポジウム (1997年 6月29日-7月 1日、英国ノルビッチ　300名, 110件)

今回の主題は、「ゲノム情報からタンパク質構造・機能へ」と「タンパク質の設計・改良」であった。

冒頭の基調講演で、ＴＩＧＲのＣ．Ｖｅｎｔｅｒが、独自に開発したいくつかの新しい解析法を説明したあと、特に「比較ゲノミックス」と「ゲノム工学」のアプローチを強調した。既に全遺伝子情報が解読されているいくつかの細菌群と古細菌を比較解析することによって、種間の系統的な差や代謝能力の差が理解された。また遺伝子ノックアウトなどによるＧｅｎｏｍｅ　Ｅｎｇｉｎｅｅｒｉｎｇによって、代謝のメカニズムや必要最少遺伝子セットの解明が可能になりつつある。

次に、Ａ．Ｂａｉｒｏｃｈが急増するゲノムデータに対する基盤データベース整備の重要性とイメージ情報の必要性を、Ｍ．Ｂｅｖａｎが植物ゲノムの解析状況を、更にはＭ．Ａｓｈｂｕｒｎｅｒが記述的知識を含めた階層的なデータベースの構築状況を述べた。 更に、ここでは欧州の構造生物学の展開について二つの話題があった。一つは、ＥＵ全１５ヶ国の構造生物学の現状と今後の課題について調査報告書が７月に刊行されること、そしてこれに基づきＥＵ全体としても構造生物学を推進する施策が取られることが表明された。もう一つは、この場で構造生物学の産業移転プログラム「ＳＢＩＰ」(Structural Biology Industrial Platform)の結成に向けた会議が行われた。ＩＰ制度は、基礎研究の産業移転を促進するために、研究者と企業が交流できる場を提供するのが主たる活動であり、ＳＢＩＰはその後１０月に正式にスタートした。

4) 米国タンパク質学会シンポジウム　(1997年 7月 12-15日、ボストン 1400名, 600件)

タンパク質研究のあらゆる側面（理論・実験・測定・計算）をカバ・する学会であるが下記のような注目すべき特別セッションがあった。

・タンパク質構造変化と病気との関係（プリオンなど）
・タンパク質のフォールディング
・タンパク質のコンピュータシミュレーション
・ゲノム情報の比較解析・ネットワーク解析
・タンパク質の設計

特に遺伝子ネットワークについての図式化(R.Brent) とシミュレーション(A.Arkin) の重要性が実際の解析事例で示された。単純なゲノム遺伝子の並びが、条件によって複雑な生命挙動の原因になるか、時間的・場所的変動の原因が、遺伝子ネットワークシミュレーションによって示された。まだ始まったばかりであるが、今後の大きな研究領域が示唆された。

４．欧米研究機関の動向

ここ２～３年に訪問した研究機関の中から、ゲノム情報解析に関係あるいくつかの研究機関の状況をご報告する。特に、タンパク質の立体構造データベース解析に詳しい研究者が多くこのゲノム情報解析分野に関係し始めていることが大きな傾向である。

1) ＥＢＩ／ＥＭＢＬ (欧州バイオインフォーマティックス研究所、ケンブリッジ）(10)ＥＭＢＬ（欧州分子生物学研究機構）の分室として、ゲノム解析のサンガーセンターに隣接して設立された。総勢９１名で、下記のように構成されている。

・データベースサービス（５３名）

ＤＢの構築・維持・提供・技術開発

ＳｗｉｓｓＰｒｏｔ，ＴｒＥＭＢＬ，ＭＳＤ

ＤＢの共同開発　１０件 FlyBase, IMGT DB, Radiation Hybrid DB など

・データベース解析研究（２８名）

ｄｅｒ，Ｓ．Ｗｏｄａｋなど計算構造生物学に強い研究者が参画－－ゲノム情報から立体構造を通して機能へ

立体構造分類の自動化（ＤＡＬＩ）
各種生物種の全塩基配列解析（例．酵母，らん藻）

・産業プログラム（１３名）

１９社がメンバー。

毎月セミナー（データベース、解析法）
データベース開示、ソフトウエア開示。相談・指導。

Ｓａｎｄｅｒが最近このＥＢＩ専任になり、Ｗｏｄａｋ教授(Brussel大) がＥＢＩ兼務になったことは、一つの傾向を示している。

2) ＢＳＭＵ／ＵＣＬ（生体分子の構造とモデリングユニット、ロンドン）(11)

ＵＣＬ（ロンドンカレッジ）の中に、Ｊ．Ｔｈｏｒｎｔｏｎ教授が率いる構造モデリンググループがある。立体構造データベース解析の第一人者であるＴｈｏｒｎｔｏｎは、上記ＥＢＩも兼務している。ここでは、下記のような解析やシステム作りがなされている。

・タンパク質立体構造分類データベース（ＣＡＴＨ）
・立体構造とゲノム情報の橋渡し－ＳＡＳ(Seq. Annotated by Str.)
・立体構造予測－Ｔｈｒｅａｄｅｒ（Ｄｒ．Ｄ．Ｊｏｎｅｓ）
・タンパク質の分子認識機構－－相互作用解析システム

Ｔｈｏｒｎｔｏｎは、通常ゲノム情報の配列のみの解析では機能は４０%以下しか分からないが、立体構造データの解析技術を利用すれば更に２０～４０%向上出来ると断言した。Ｊｏｎｅｓは、構造認識法プログラムＴｈｒｅａｄｅｒで、1994年と1996年に行われた立体構造予測コンテストで優秀な成績を収めたが、最近この改良版を公開し、更に現在ゲノム情報解析用に自動化を進めている。(Warwick Colledge に移籍) (12)

3) ＮＣＢＩ／ＮＩＨ（国立バイオテクノロジー情報センター，メリーランド）(13)

配列データベース解析の第一人者であるＤｒ．Ｄ．Ｌｉｐｍａｎの率いる計算生物学（Computational Biology)のでは、下記のような研究が進んでいる。

・タンパク質の立体構造分類（ＶＡＳＴ）の自動化（Ｓ．Ｂｒｙａｎｔ）
・モチ-フの抽出－－配列および立体構造
・Ｃａｎｃｅｒ　Ｇｅｎｅ　Ａｎａｔｏｍｙ　Ｐｒｏｊｅｃｔ

異なった細胞毎の遺伝子発現レベルを調べる。

・第２回立体構造予測コンテストＣＡＳＰ２の評価（Ｓ．Ｂｒｙａｎｔ）－－配列ホモロジー１０%位のものでもかなり良い結果が得られた。
・ゲノム情報からのタンパク質ファミリー分類（Ｅ．Ｋｏｏｎｉｎほか）(14)

4) ＭＣＳ／ＡＮＬ（数学・コンピュータシステムグループ、シカゴ）(15)

アルゴンヌ国立研究所の中のＭＣＳは、米国ＨＰＣＣ計画の中心人物の一人であるＲ．Ｓｔｅｖｅｎｓが率いている。ＤＯＥのグランドチャレンジプログラムの下、最先端コンピュータを生命科学分野へ応用することにも力を入れている。

・大規模系統樹の作成
・代謝パスウエイのデータベース構築
・ゲノム全配列データの自動解析システム（ＭＡＧＰＩＥ）

Similiarity, Motif, ORF, 2nd.Str., Repeat Seq., Pathway

５．終わりに

以上、学会・研究機関でのいくつかの動きをご紹介した。繰り返しになるが、全体としてまとめると下記の通りとなろう。

1) タンパク質の解析技術、特に立体構造データベース解析技術がゲノム情報解析に非常に重要であるとの認識が高まっている。

2) 多くの優秀な立体構造データベース解析研究者あるいは立体構造理論解析研究者が、ゲノム情報解析に参画し始めている。

3) タンパク質解析用の各種プログラムやシステムは、ゲノム情報解析用に改良されつつある。特に、大量の配列データの解析に対処するため、各種解析ソフトウエアの自動化計画が進んでいる。

4) 次々と解読が完了してくる各生物種のゲノムデータベースが公開されると同時に、あるいはその前に、網羅的な解析が行われている。

5) 各生物種のゲノムデータベースの比較解析が、急速に始められている。その成果も次々と発表され始めている。

日本においても、構造生物学およびゲノム情報解析について、多くの省庁で計画が進んでいるが、ここでは触れない。欧米の動きがご参考になれば幸いである。これらの調査をさせて頂いてきた、三菱化学・およびＭＳＩ（米国サンジェゴ）に感謝申し上げる。

参考文献

1) Dill,K.and Chan,H.: From Levinthal to pathways to funnels. Nature Structural Biology,4,10-19(1997)
2) Shakhnovich,E.: Theoretical studies of protein-folding thermodynamics and kinetics. Current Opinion in Structural Bilogy,7,29-40(1997)
3) Lee,R.: Protein Model Building using Structural Homology. Nature,356,543-544(1992)
Sanchez,R.and Sali,A.: Advances in comparative protein-structure modelling. Current Opinion in Structural Biology,7,206-214(1997)
4) Bowie,J.Luthy,R.& Eisenberg,D.: A method to identify protein sequences that fold

into a known three-dimensional structure. Science,253,164-170(1991)
Jones,D.,Taylor,W.& Thornton,J.: A new approach to protein fold recognition. Nature,358,86-89(1992)
5) Matsuo,Y.and Nishikawa,K.: Protein Structural Similiarities Predicted by a Sequence-Structure Compatibility Method. Protein Science,3,2055-2063(1994)
Jones,D.: Progress in Protein Structure Prediction. Current Opinion in Structural Biology,7,377-387(1997) 6) Chothia,C.: One Thousand Families for Molecular Biologists. Nature,357,543-544(1992) 7) Fleischmann,R.et al(40 persons): Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. Science,269,496-(1995)
8) Sander,C.et al(10 persons): Challenging Times for Bioinformatics. Nature,367,647-648(1995)
9) Bork,P.,Ouzounis,C.and Sander,C.: From Genome Sequences to Protein Function. Current Opinion in Structural Bilogy,4,393-403(1994)
10) http://www.ebi.ac.uk 11) http://www.biochem.ucl.ac.uk 12) http://globin.bio.warwick.ac.uk 13) http://www.ncbi.nlm.nih/structure 14) Tatusov,R.,Koonin,E.and Lipman,D.: A Genomic Perspective on Protein Families, Science 278,631-637(1997)
15) http://www.mcs.anl.gov

やお　とおる　　Ｔｏｒｕ　ＹＡＯ

　情報科学と生命科学の橋渡しをしたい。タンパク質やゲノム情報のコンピュータ解析は その良い例であると思っている。更に、分子シミュレーション技術が化学・バイオ研究 や材料開発に大いに役立つように普及活動をしたい。趣味は、テニスと囲碁。

・連絡先：〒227-8502 横浜市青葉区鴨志田町1000 三菱化学・横浜総合研究所
Tel.045-963-3123, Fax.045-963-3249, http://www.msi.com

CICSJ Bulletin Vol.16, No.2, April 1998

目次

本文の表示には、Adobe Acrobat Reader version 3.0 以上が必要です。

・特集：専用計算機

・部会記事

・部会行事

・関連行事

・編集後記

CICSJ INDEX にもどる

大規模科学技術計算の夢：　『専用計算機』特集の序にかえて

国立がんセンター研究所生物物理部　　相田　美砂子
maida@ncc.go.jp

化学者にとっての大きな夢物語の一つは「生命現象の解明」です。『化学と工業』1998年４月号の特集は「生命化学」でした。「生命化学」という領域があったとは、私は知りませんでしたが、名称はともかくとして、「生命活動がどのように営まれているのかを分子のレベルで解明する」ことへの興味と関心は、最近非常に高まっています。そして、計算機を使った研究手法の「生命化学」への適用は、計算機能力の増大と共に、その重要性を増しています。「生命化学」を視野に入れた量子化学や計算化学の分野の一つの方向として、CICSJ Bulletin では、vol.15, no.2 (1997) に分子動力学法を使った生体高分子への取り組みについて、vol.15, no.5 (1997) に、生体高分子の量子化学計算を特集いたしました。これらの記事においては現在適用されている量子化学や計算化学における方法やその限界等について解説していただきました。専門外の方々にも、この分野は今「動いている」という雰囲気が伝わっていれば幸いです。

最近のコンピュータの進歩には目をみはるものがあります。１０年ほど前には実行可能ではあってもかなり困難であったことは、今では「朝飯前」で実行可能です。その頃「夢物語」であったもののうちのいくつかは、今では「夢」ではなくなっています。しかし、研究者の欲というものは際限のないものです。ここまでできるなら、もうちょっとそこまで、と、さらに先を期待してしまいます。ところが、汎用の計算機を使っているかぎり、現在および近い将来に実現可能なことには限度があります。そのため、専用の計算機を開発することによる一層の進歩に期待が高まっています。今号においては、非経験的分子軌道法と分子動力学法を対象とした専用計算機を開発しておられる方々に、現状における機能と実現可能な夢について御執筆をお願いいたしました。専用計算機は、ある目的を達するために必要な計算だけが高速で実行可能になるよう、専用に設計された計算機です。専用計算機を使って、より大きな量子化学計算あるいは分子動力学法による計算が、より容易に実行できるようになることが期待されます。

「大規模科学技術計算」という表現が生まれてからどのくらい時間がたったでしょうか。時がすすむにつれて、同じことばでもその定義は変わっていきます。私が非経験的分子軌道法計算をはじめた頃は、かなり小さめな基底関数を使った核酸塩基の Hartree-Fock 計算でも「大規模科学技術計算」の範疇に入りました。ところがそのような計算は、今ではふつうにあるワークステーションを使っても、数分で終わってしまいます。ハードウェアの進歩とともに、研究の幅は広がり、奥行きは深くなり、新たなアイディアが生まれてきます。また一方、その進歩したハードウェアを使いこなすソフトウェアの進歩も重要です。ただやみくもに大規模計算をすればそれなりの結果が得られるわけではありません。「大規模科学技術計算」のめざす方向の一つが「生命現象の解明」ですが、そのためにまず実行可能でなくてはならないことは、現実の系のシミュレーションです。ただ分子一つの計算ではなく、その環境すべてを考慮に入れた計算です。このようなシミュレーションは、ハードウェアの進歩だけを期待していたのでは実行可能になりません。理論化学、量子化学、計算化学、計算機化学、そして化学に限らず物理学や生物学におけるいろいろな分野の研究者の叡智を集めることによって、大規模科学技術計算から実のある結果が得られるようになり、そしてますます計算機の果たす役割が大きくなっていき、「夢物語」が夢ではなくなっていくのです。「化学者の夢」が「夢物語」ではなくなり、実際のジョブの実行結果として語ることのできる日も近い、と楽しみにしています。

あいだ　みさこ　　AIDA, Misako
連絡先　　〒104-0045　東京都中央区築地５－１－１
国立がんセンター研究所　生物物理部
TEL: 03-3542-2511 (ext.4601)　　FAX: 03-3546-1369

非経験的分子軌道計算専用システムMOE のソフトウェア

北海道教育大学教育学部釧路校　小原　繁
九州大学大学院システム情報科学研究科　村上和彰
お茶の水女子大学理学部　長嶋雲兵
島根大学総合理工学部　網崎孝志
工業技術院物質工学工業技術研究所　田辺和俊・北尾　修
大正製薬株式会社 創薬研究所　北村一泰・高島　一
富士ゼロックス株式会社 EC 技術開発部宮川宣明・稲畑深二郎・山田　想
obara@kus.hokkyodai.ac.jp

本報告では、非経験的分子軌道計算専用システム MOE のソフトウェア構成および機能について概説する。本システムは、主に非経験的分子軌道計算を高速に行うことを目的としている。この計算で最も時間を要するのは、２電子積分と呼ばれる量の計算である。現在最もポピュラーな２電子積分計算方法である小原の漸化表式を用いた計算は、全ての演算が乗算と加算で構成される。特に（係数）×（中間積分）+（中間積分）という形式の積和演算が、２電子積分計算のための演算の大多数を占める。本システムは、積和演算を高速に行える浮動小数点演算部を有する専用LSI を多数備えることで、２電子積分の計算を並列に行い、非経験的分子軌道計算を高速化することを1つの特徴とする。

1. はじめに

計算機と情報ネットワークの飛躍的な発展に伴い、計算機シミュレーションによる機能性材料や医薬の設計は1 つの研究分野を形成しつつある程に進展している。計算機シミュレーションの強みは理論や科学的知識に立脚して分子や物質の設計を可能にできる点である。このため、化学や医薬品産業の研究者も計算機シミュレーションの普及とさらなる発展を切望している。数ある分子シミュレーション手法の中でも非経験的分子軌道法は、分子個々の物理的化学的性質と分子間相互作用という物質科学の基礎的知見を提供する。そして、分子動力学シミュレーション等で用いるポテンシャル関数に関する知見を与えて分子集合系の巨視的物性算出を可能にする。従って、非経験的分子軌道法は分子シミュレーション各手法の基盤といっても過言ではない。そのため、国内外において数多くの非経験的分子軌道法プログラムが開発され、海外のいくつかはすでに商品化されて実際の分子設計に利用されるに至っている。非経験的分子軌道法が格段に実行しやすくなったため、分子の電子状態の精密な波動関数を誰も容易に得ることができるようになった。この波動関数を解析し種々の物理量の期待値を求めることを通して分子系の物理的・化学的な性質や現象について理解を一層深めていくことが広く行われる状況になっている。研究者が独自の解析を行なうことは理論化学者ばかりではなく実験化学者にも可能になって来ている。しかしながら、非経験的分子軌道計算は、基底関数の数 N の 4 乗に比例する演算量と補助記憶量を必要とする。そのため、現在のスーパーコンピュータシステムを使っても、演算量およびディスク容量の点からたかだか l00 原子程度の分子系に適用されているにすぎず、まだまだ生命あるいは化学現象の素過程の分子論的解明とそれに立脚した材料や医薬の設計には不十分といわざるを得ない。物質科学のさらなる発展を考えると、非経験的分子軌道計算の飛躍的な高速化と補助記憶容量等の計算コストの低減化は不可欠である。これの解決策は、カスタム LSI の開発を含む専用計算システムの開発が最も現実的であると考えられる。非経験的分子軌道計算の飛躍的な高速化と補助記憶容量等の計算コストの低減化の実現は、物質科学の各分野へ大きなインパクトをあたえる。本研究の目的は、このような非経験的分子軌道計算の超高速化専用デバイスと計算機システム MOE の開発に加えて、MOE を利用することを念頭においた分子軌道計算プログラムの開発である。

2. 非経験的分子軌道計算とMOE

MOE の作成にはハードウェア上とソフトウェア上の解決すべき多くの問題がありこれらを総て解決するには

長期間を要する。そこで、第1次計画においては、非経験的分子軌道計算法としては最も基本的であり世界的に最も広く利用されているハートリーフォック(HF)法を高速化することに力点を置いている。この方法は、1)データ入力、2)分子積分計算、3)フォック行列の作成、4)フォック行列の対角化、5)全エネルギー計算からなる。解くべき方程式が非線形方程式であるため、ステップ2)で計算された分子積分をDisk等の補助記憶に蓄積し、答えが一定になるまでステップ3)、4)を反復する。このうち、ステップ2)の演算量は基底関数の数Nの4乗に、ステップ3)、4)はNの3乗に比例する。最も計算時間を要するのはステップ2)であり、現状では全実行時間の90%以上を占める。ワークステーションクラスタによる並列計算の経験から、ステップ4)の対角化はホストで行なうこととし、ホストとMOE間の通信量をNの2乗のオーダにおさえるため、MOEではステップ2)と3)を高速化することを目的とした。つまり、Fock行列の計算式(r,s,t,u=1~N)は、

$$F_{rs} = T_{rs} + V_{rs} + \sum_t \sum_u P_{tu} \left\{ (rs,tu) - \frac{1}{2}(rt,su) \right\}$$

であり、ここでT_{rs}は、電子の運動エネルギー積分、V_{rs}は核と電子の引力積分と呼ばれ、第3項は、電子間の反発を表わす2電子反発項である。この第3項を以後F'_{rs}と表わす。

$$F'_{rs} = \sum_t \sum_u P_{tu} \left\{ (rs,tu) - \frac{1}{2}(rt,su) \right\}$$

F'_{rs}は密度行列の要素Nと2電子積分(rs,tu)、(rt,su)の和の積となる。MOEで計算するのは上式のF'_{rs}である。F'_{rs}の(rs,tu)が分子積分と呼ばれ、N^4個の計算量と補助記憶量が必要となる。次に分子軌道計算の並列化について簡単に説明する。

3. 2電子積分

2電子積分(rs,tu)には4個のガウス関数が含まれている。各ガウス関数の軌道量子数を使用して2電子積分を(**rs**,**tu**)の様に表現すると、軌道量子数の大きな2電子積分（例えば([**r**+**1**]**s**,**tu**)）は軌道量子数の小さな積分（(**rs**,**tu**)や([**r**-**1**]**s**,**tu**)等）により表現できる。[1)]

$$\begin{aligned}
([\mathbf{r}+\mathbf{1}_i]\mathbf{s},\mathbf{tu})^{(m)} &= (\mathbf{P}-\mathbf{R})_i(\mathbf{rs},\mathbf{tu})^{(m)} + (\mathbf{W}-\mathbf{P})_i(\mathbf{rs},\mathbf{tu})^{(m+1)} \\
&+\left\{\frac{N_i(\mathbf{r})}{2z}\right\}([\mathbf{r}-\mathbf{1}_i]\mathbf{s},\mathbf{tu})^{(m)} + \left\{-\frac{r\,N_i(\mathbf{r})}{2z^2}\right\}([\mathbf{r}-\mathbf{1}_i]\mathbf{s},\mathbf{tu})^{(m+1)} \\
&+\left\{\frac{N_i(\mathbf{s})}{2z}\right\}(\mathbf{r}[\mathbf{s}-\mathbf{1}_i],\mathbf{tu})^{(m)} + \left\{-\frac{r\,N_i(\mathbf{s})}{2z^2}\right\}(\mathbf{r}[\mathbf{s}-\mathbf{1}_i],\mathbf{tu})^{(m+1)} \\
&+\left\{\frac{N_i(\mathbf{t})}{2z+2h}\right\}(\mathbf{rs},[\mathbf{t}-\mathbf{1}_i]\mathbf{u})^{(m+1)} \\
&+\left\{\frac{N_i(\mathbf{u})}{2z+2h}\right\}(\mathbf{rs},\mathbf{t}[\mathbf{u}-\mathbf{1}_i])^{(m+1)}
\end{aligned}$$

この関係式を使用すれば、任意の軌道量子数の2電子積分を軌道量子数が零（s型関数）の積分$(ss,ss)^{(m)}$から計算できることになる。関係式は、斉一次漸化関係式であるので、2電子積分の計算は（係数）×（中間積分）+（中間積分）という積和演算を必要回数繰り返すことにより出来ることになる。コンピュータに適したシンプルは演算により2電子積分を計算できることがMOEの重要な要素になっている。

4. MOE ハードウエアの構成

図 1.1 は、MOE 専用ハードウエア・システムの構成を概念的に示すものである。LAN（Local Area Network）などのネットワークに接続されたホスト・ワークステーション（あるいはパーソナル・コンピュータ）にダイレクトに（図 1.1 の上側の構成）、あるいはホスト・ワークステーションとの間にワークステーションを挟んで（図 1.1 の下側の構成）、複数の MOE 専用ボードが接続される。通常は前者の構成を採るが、後者の構成を採ることも可能である。ワークステーションと MOE 専用ボードとは 100Mbps（bit per second）程度以上のデータ転送性能を有するシリアルバスで接続される。

前者の構成で接続する MOE 専用ボード数を増加させた場合には、シリアル・バスの通信性能がボトルネックとなってボード数の増加に見合った計算性能の向上が得られない場合がありうる。後者の構成を採れば、ボード数に見合った計算性能を維持してボード数を増加させることができ、より大規模な計算を短時間で行うことが可能となる。

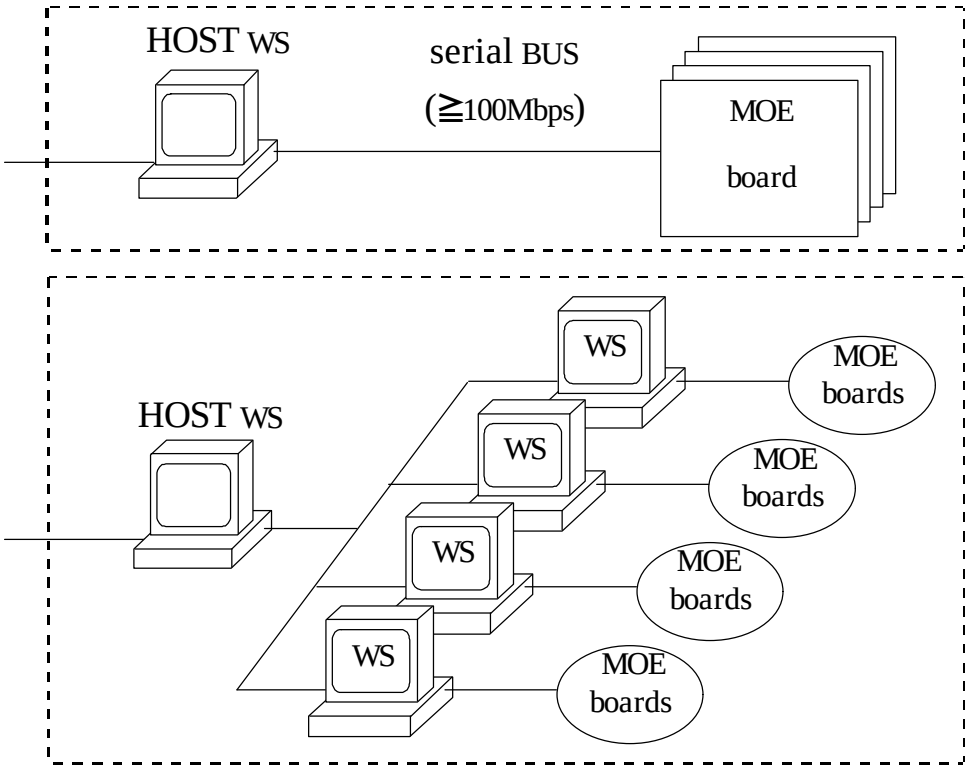


図 1.1. MOE **専用ハードウエア・システムの概念構成。**

5. システムにおける処理の流れ

MOE専用システムにおける非経験的分子軌道計算の処理の主要な手順を、図1.2に沿って説明する。

ホスト計算機における主な処理手順は、以下の通りである。まず、係数行列の試行値から密度行列を計算する。次に、インデックス R を MOE 専用 LSI に割り当てる。このインデックスは、単一の原子核を中心として同じ軌道量子数を有する複数の基底関数からなる1つの縮約シェルに対応する。ここでは計算の対象とする系を構成する縮約シェルの総数が Nshell 個であるとし、インデックス R は 1 から Nshell までの範囲の値を取る。システム内に複数ある専用 LSI には、別々のインデックスが割り当てられ、各専用 LSI はフォック行列要素 F(I,*)（*は 1 から N ; N は、系に含まれる縮約基底関数の数）の計算を行う。フォック行列の計算の間に専用 LSI から必要な密度行列が要求されるため、ホスト計算機はリクエストに応じて密度行列を送信する。全ての R に関する計算が終了してフォック行列が揃うと、ホスト計算機はフォック行列の対角化計算を行い、新たな係数行列を計算する。

MOE 専用 LSI における処理の手順は以下の通りである。インデックス R の割り当てを受けると、専用 LSI はフォック行列 F(I,*)の計算を開始する。この計算は、インデックス S,T,U に関する 1 から Nshell までのループを

この順に回し、その最も深い位置で2電子積分を計算し、それに適当な密度行列要素を乗算したものをフォック行列要素に足し込むことによって行われる。用いられる密度行列要素は P(K,*)および P(J,*)（*は 1 から N）であり、これらは T および S に関するループの開始時にホスト計算機から受信する。専用 LSI 上にローカルに保持しなければならない密度行列要素は、N×N 要素を有する密度行列のうちの高々数行分であり、専用 LSI に外付けする高速 SRAM で十分な容量を確保できる。

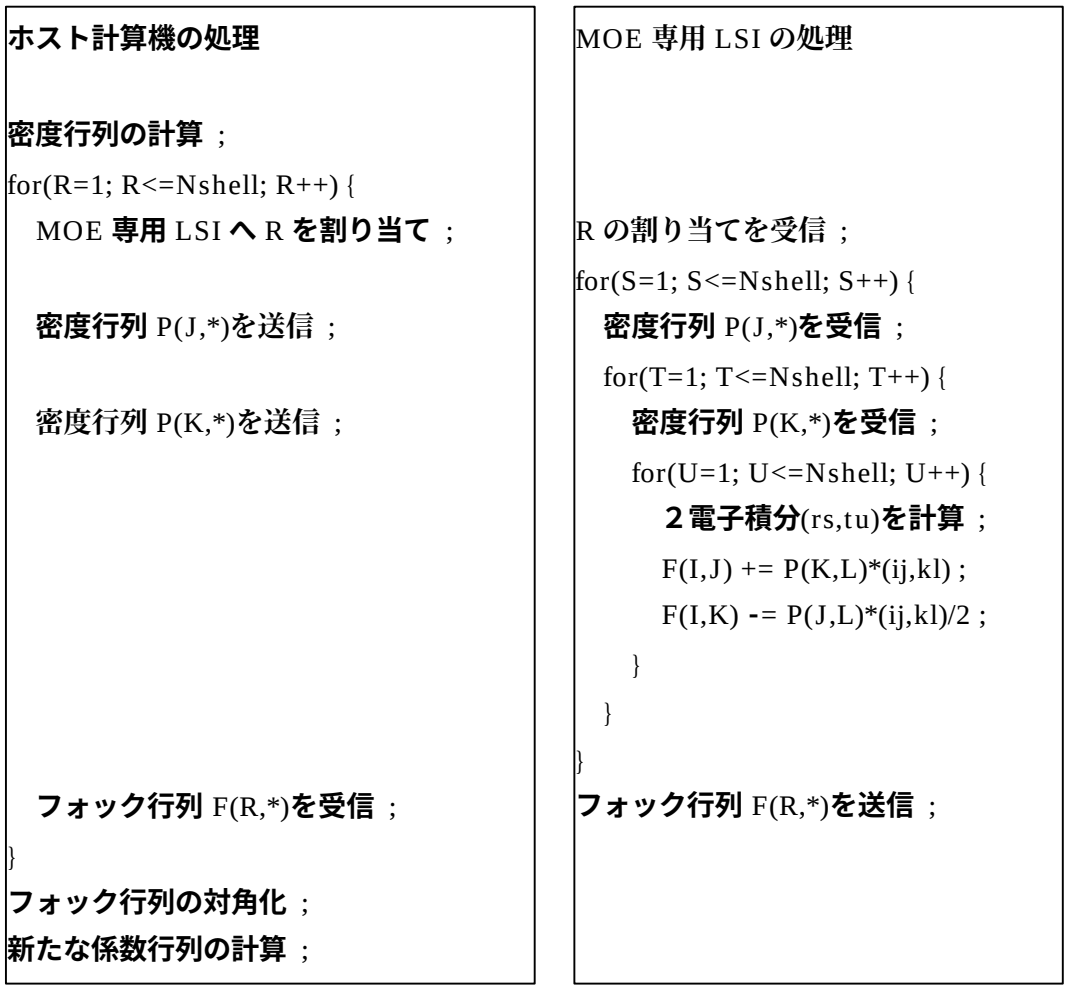


図 1.2. MOE **専用システムにおける処理の主要な手順。**

参考文献

1) Obara, S., Saika A.; Efficient recursive computation of moelcular integtals over Cartesian Gaussian Function, *J. Chem. Phys*, **84**, 3963-3974(1986); Obara, S., Saika A.; General recurrence formulas for molecular integrals over Cartesian Gaussian Functions, *J. Chem. Phys*, **89**, 1540-1559(1988); Honda, M., Sato, K., and Obara, S.; Formulation of molecular integrals over Gaussian functions treatable by both the Laplace and Fourier transforms of spatial operators by using derivative of Fourier-kernel multiplied Gaussian, *J. Chem. Phys*, **91**, 3790-3804(1991).

――

おばら　しげる　OBARA, Shigeru

ここ数年は、非経験的分子軌道計算の高速化とそのための理論の研究に従事しています。この他に、教育大学の本務である教員養成のための勉強会やセミナーにも力を注いでいます（これはかなりの激務です）。趣味、音楽発声。

085-8580 北海道釧路市城山 1-15-55 北海道教育大学教育学部釧路校化学教室 Tel:0154-41-6161

非経験的分子軌道計算専用システムMOEのハードウェア

九州大学大学院システム情報科学研究科	村上和彰
北海道教育大学教育学部釧路校	小原　繁
お茶の水女子大学理学部	長嶋雲兵
島根大学総合理工学部	網崎孝志
工業技術院物質工学工業技術研究所	田辺和俊
大正製薬株式会社創薬研究所	北村一泰・高島　一
富士ゼロックス株式会社 EC 技術開発部	宮川宣明・稲畑深二郎・山田　想

moe@c.csce.kyushu-u.ac.jp

本稿では、前掲の「非経験的分子軌道計算専用システム MOE のソフトウェア」に続いて、そのハードウェア、特に現在開発中の MOE 専用 LSI の構成および機能について説明する。本 LSI は小原の漸化形式に基づく2電子積分計算の高速化を目的とし、（係数）×（中間積分）＋（中間積分）という形式の積和演算に特化したメモリーメモリ型の浮動小数点積和演算器を有する。本演算器はクロック周波数 100MHz でパイプライン動作し、乗算とそれに引き続く加算の結果を 10nsec 毎に生成する。すなわち、200MFLOPS の浮動小数点演算処理性能を有する。まず、MOE 専用ボードの構成と機能を説明した後、上記の特徴を実現する MOE 専用 LSI の構成と機能について詳細に説明する。

1 MOE 専用ボードの構成と機能

1.1 ボード構成の概要

図 1 に、MOE 専用ボードの構成を示す。MOE 専用ボード上には、IEEE1394 物理層 LSI（図中の 1394PHY）、IEEE1394 リンク層 LSI（図中の 1394LLC）、ブリッジ LSI（図中の bridge）、MOE 専用 LSI（図中の MOELSI）、SRAM（図中の LM）の、計5種類の LSI が配置される。

最初の2種類はホストーボード間あるいはボードーボード間を接続する IEEE1394 インタフェイス部であり、またブリッジ LSI はボード内部の結合に用いる PPRAM-Link と IEEE1394 との間でのデータの変換を行う。これら3種類の LSI によりボード内部とボード外部との間の通信インターフェイスが構成され、ホストと末端の専用 LSI との間でのデータの転送が実現される。

1ボード上では、5個の MOE 専用 LSI が PPRAM-Link を通してブリッジに接続されている。さらに MOE 専用 LSI の各々には、外部メモリ（LM）として5個の 4M ビット SRAM チップが接続されている。

PPRAM-Link は、17 ビット（データ 16 ビットおよびフラグ 1 ビット）の信号線から成るパラレル・リンクである。各ビットに対応する信号線では 50MHz の速度で 1 ビットのデータ転送が行われ、結合網としては 100M バイト/秒のデータ転送速度を有する。

1.2 ボードにおける処理の流れ

MOE 専用ボードにおける処理の流れは次の通り。

まず、ホストから IEEE1394 を介して送信された共通データを受信すると、通信インターフェイス部（以下通信

I/F 部）はそれを PPRAM-Link フォーマットのデータに変換し、ボード上の通信網を介して全ての MOE 専用 LSI（以下 PE: Processor Element）へ送信する。これらのデータは、各 PE や PE に付属するローカルメモリ（LM）上に保持される。通信 I/F 部は、共通データの送信を終了すると、データ受信待ちの状態となる。

ホストから IEEE1394 を介してある特定の PE へ密度行列要素データが送信されると、ターゲットとなる PE が含まれるボード上の通信 I/F 部はそのデータを受信し、PPRAM-Link フォーマットのデータに変換する。変換されたデータはボード上の結合網を介してターゲットの PE へ送信される。通信 I/F 部は、その時点で処理すべきデータがなければデータ受信待ちの状態となる。

ターゲットとなった PE は、受信した密度行列要素データを用いて Fock 行列要素を計算する。PE における計算の詳細については後述する。計算が終了すると、Fock 行列要素データを（ターゲットをホストとして）ボード上の結合網を介して送信する。

ボード上の PE が送信した Fock 行列要素データを受信すると、通信 I/F 部はそれを IEEE1394 フォーマットに変換し、IEEE1394 を介してホストへ送信する。

通信インターフェイス部で処理される主なデータを表 1 にまとめる。

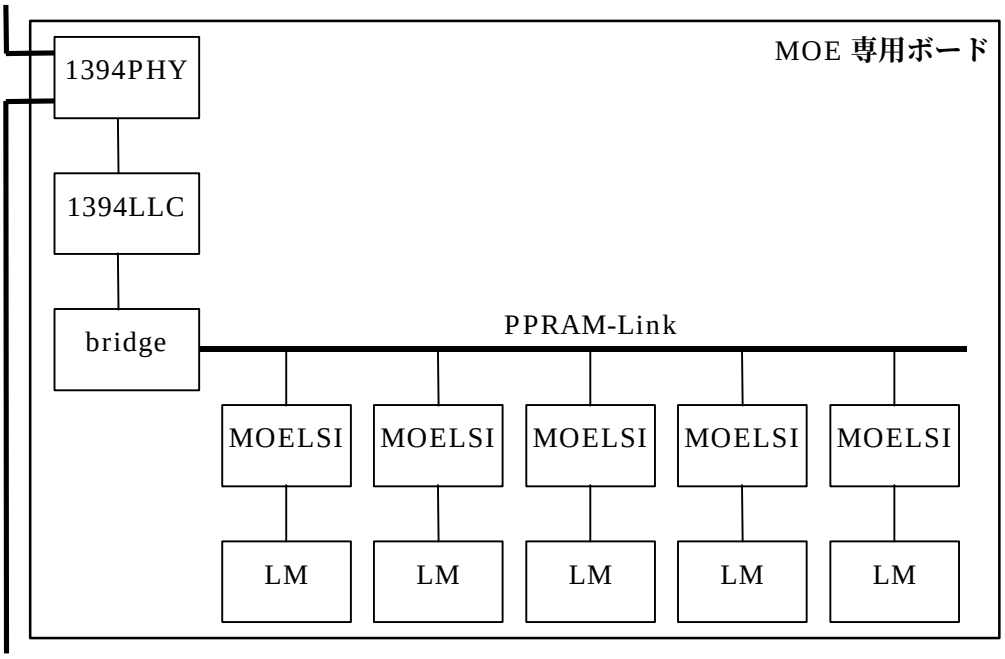


図 1: MOE 専用ボードの構成

表 1: ボード上の通信インターフェイス部で処理される主なデータ

データの種類	ソース	ターゲット
共通データ	ホスト	全ての PE
密度行列要素	ホスト	特定の PE
Fock 行列要素	PE	ホスト

1.3 IEEE1394－PPRAM-Link インターフェイス

ボード上の通信インターフェイス部に関して説明を付け加える。

まず、IEEE1394 物理層 LSI（以下 1394PHY）は IEEE1394 シリアル・バスを直接ドライブするチップであり、そのデータ転送を制御するのが IEEE1394 リンク層 LSI（以下 1394LLC）である。ブリッジ LSI は、1394LLC 内のレジスタを読み書きすることによって IEEE1394 を用いた転送を実行する。

1394PHY は、IEEE1394 シリアル・バスのアービトレーション機能、IEEE1394 データのエンコード/デコード機能などを有している。1394LLC は、送受信データを一時的に保持しておく機能、1394PHY より受け取ったデータのターゲット・アドレスが自ノードのものであるかどうかを判定する機能、送出するデータに付加するヘッダを生成する機能などを有する。ブリッジ LSI は、1394LLC が外部より受信したデータを取り出して PPRAM-Link のフォーマットに変換する機能、逆にボード上の PE から受信したデータを IEEE1394 のフォーマットに変換して 1394LLC へ書き込む機能を有する。

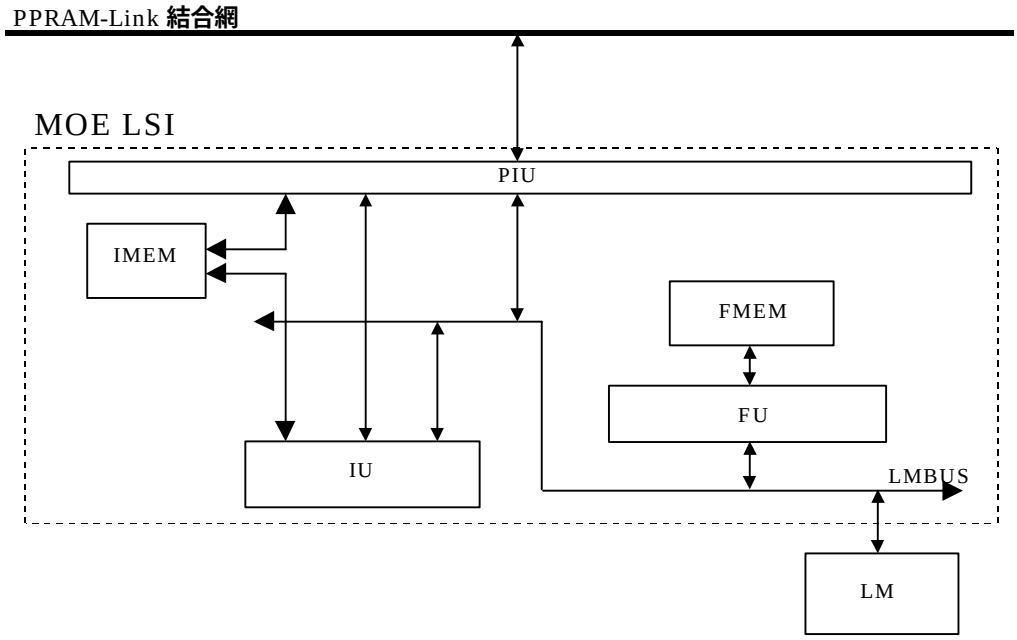


図 2: MOE 専用 LSI の構成

2 MOE 専用 LSI の構成と機能

2.1 専用 LSI の構成と機能の概要

MOE 専用 LSI は、内部に RISC 型プロセッサ・ユニット（IU）、浮動小数点演算ユニット（FU）、PPRAM-Link インターフェイス・ユニット（PIU）、整数メモリ（IMEM）、浮動小数点メモリ（FMEM）を備えた、PPRAM（Parallel Processing Random Access Memory）型の LSI である。100MHz のクロック周波数で動作する。外部にはローカル・メモリ（LM）が接続される。

図 2 に MOE 専用 LSI のブロック図を示す。IU、FU、および PIU は LSI 内部のバス（LMBUS）を介して相互にデータの入出力を行うことができる。また LMBUS は、LM にも接続されており、IU、FU、PIU から LM にアクセ

スすることも可能となっている。PIU は、ボード上の PPRAM-Link にも接続されている。さらに、IMEM は IU および PIU に、FMEM は FU に接続されている。

次に、MOE 専用 LSI の機能の概要を説明する。MOE 専用 LSI は、PPRAM-Link を介してホストから受信した初期データを、データの種類別に IMEM、FMEM、LM に格納する。これらのデータは、IU あるいは FU で随時用いられる。同じようにホストから受信した密度行列要素は、LM に格納される。IU における処理により計算すべき2電子積分が決められ、FU は IU の制御のもとで初期積分および係数の計算を、また自律的に計算手順を読み出して漸化計算を行う。得られた Fock 行列要素は一旦 LM に格納された後、ホストへ転送される。これらの処理に関して、以下の節で詳細に説明する。

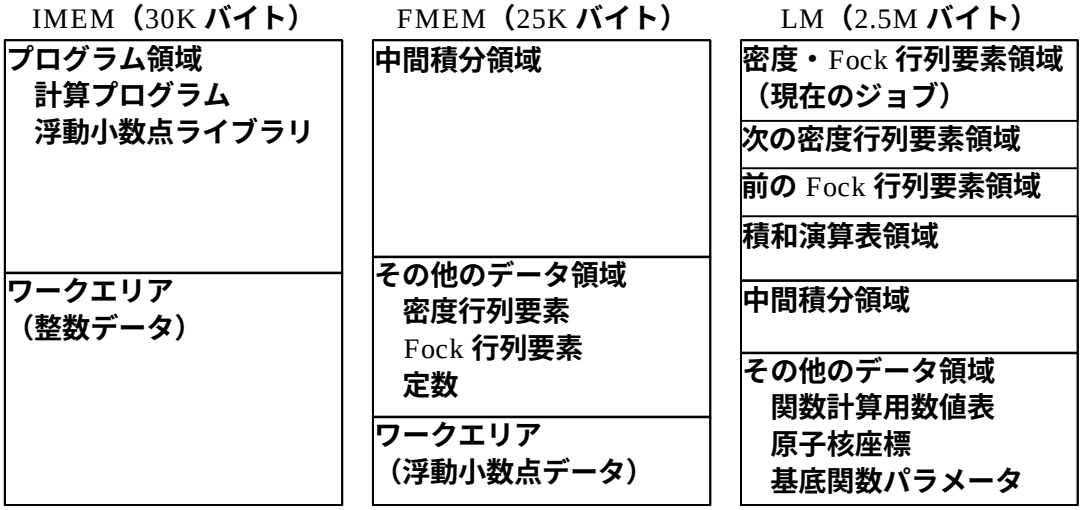


図 3: 各メモリに用意されたデータ領域

2.2 メモリ構成

MOE 専用 LSI における処理で用いられるメモリの構成、および各メモリに格納されるデータについて説明する（図3参照）。

MOE 専用 LSI は、30K バイトの整数データを格納できる IMEM と、25KB バイトの浮動小数点データを格納できる FMEM を内部に備える。また、外部には 2.5M バイトのデータを格納できる LM が接続されている。いずれもアクセス時間の短い SRAM（Static Random Access Memory）で構成されている。

IMEM には、IU で実行されるプログラム（計算プログラム、浮動小数点演算用ライブラリ）、およびプログラムで使用される整数データが格納される。

FMEM には、FU が取り扱う浮動小数点データが格納される。その大部分は2電子積分の漸化計算の途中で出現する中間積分を格納するための領域であるが、その他に密度行列要素や Fock 行列要素の一部および定数データ、プログラムで使用される浮動小数点データもここに格納される。

LM には主に、密度行列要素、Fock 行列要素、２電子積分の漸化計算のための計算手順を記した積和演算表が格納される。ここに格納されている密度行列要素および Fock 行列要素は、１つのジョブを終了して次のジョブの割り当てを待つ時間をなくすため、あらかじめ次のジョブのために必要な密度行列要素、および前のジョブで計算された Fock 行列要素も格納されている。また、2電子積分を構成する基底関数の軌道量子数が大き

大きい場合には、その漸化計算の途中で出現する中間積分が FMEM の容量を超える場合があり、FMEM に格納できない分を一時的に格納する場所としても LM が用いられる。その他に逆数、平方根の逆数、指数関数、誤差関数を計算する際に用いる数値テーブル、原子核座標や原子軌道を記述する基底関数に関連するパラメータもここに格納される。

```
ホスト計算機の処理

密度行列の計算 ;
for(R=1; R<=Nshell; R++) {
   MOE 専用 LSI へ R を割り当て ;

   密度行列 P(J,*)を送信 ;

   密度行列 P(K,*)を送信 ;









   フォック行列 F(R,*)を受信 ;
}
フォック行列の対角化 ;
新たな係数行列の計算 ;
```

```
MOE 専用 LSI の処理



R の割り当てを受信 ;
for(S=1; S<=Nshell; S++) {
   密度行列 P(J,*)を受信 ;
   for(T=1; T<=Nshell; T++) {
      密度行列 P(K,*)を受信 ;
      for(U=1; U<=Nshell; U++) {
         ２電子積分(rs,tu)を計算 ;
         F(I,J) += P(K,L)*(ij,kl) ;
         F(I,K) -= P(J,L)*(ij,kl)/2 ;
      }
   }
}
フォック行列 F(R,*)を送信 ;
```

図 4: MOE 専用システムにおける処理の主要な手順

2.3 専用 LSI における処理の流れ

MOE 専用 LSI で行う主要な処理の流れを、図 4 に沿って説明する。

計算が開始されると、まず MOE 専用 LSI はホストから送信されたインデックス R を受信する。その受信をトリガーとして、IU は Fock 行列要素 F(I,*)の計算を開始する。

IU はまず、インデックス S に関するループを回し、その中で密度行列要素 P(J,*)の転送要求をホストに送信し、受信した密度行列要素を LM に格納する。また、S のループの内部で T のループを回し、その中で密度行列要素 P(K,*)の転送要求をホストに送信し、受信した密度行列要素を LM に格納する。さらに T のループの内部で U のループを回す。

U に関するループの中で、２電子積分の計算が行われる。2電子積分計算のために必要な係数および初期

積分の計算は、IU が FU を制御しながら行う。この時、計算に必要なデータ（原子核座標や基底関数パラメータなど）は LM から随時読み出して用いる。係数および初期積分の計算が終了すると、FU は自律的に2電子積分の漸化計算を行う。これは、LM から積和演算表を読み出し、そこに記述されている積和演算を実行する、という繰り返しで行われる。積和演算表に記述されている演算処理が終了して2電子積分の数値が求まると、IU は FU を制御して、得られた2電子積分を Fock 行列要素へ足し込む。この時に用いる密度行列要素は、LM から読み出して用いる。ここまでの処理が終了すると、IU は実行すべき全ての２電子積分計算が終了したかどうかを判断し、終了していなければ次に計算する2電子積分を決める。全ての2電子積分計算が終了すると、IU は計算した Fock 行列要素を LM に格納し、ジョブが終了したことを PIU に伝える。PIU は LM に格納された Fock 行列要素をホストへ転送する。

専用 LSI は、Fock 行列要素を送出した時点で次のインデックス R が既に割り当てられていれば、それに対応する Fock 行列要素計算を開始する。

2.4 IU の機能

IU は RISC 型プロセッサであり、内部に整数レジスタを含んでいる。IU を制御する命令は IMEM に格納されている。IU は 100MHz の周波数のクロックで動作し、パイプライン動作により理想的には1クロック毎に命令を読み出すことが可能である。通常の RISC プロセッサと同様に整数の算術演算、論理演算や分岐を行える他、整数レジスタと IMEM/LM とのロード/ストア、浮動小数点演算ユニット（FU）に含まれるレジスタと FMEM とのロード/ストア、FU を用いた浮動小数点演算の制御などを行うことができる。

Fock 計算を行う際には、計算すべき2電子積分の決定を主に整数の算術演算および論理演算を用いて行い、係数および初期積分の計算や Fock 行列要素の計算には浮動小数点演算の制御も行う。２電子積分の漸化計算の際には、内部の特定のステータス・レジスタを監視し続け、FU による積和演算表の実行の終了を検出して Fock 行列要素の計算を実行する。

2.5 FU の機能

２電子積分の漸化計算時に浮動小数点数の積和演算を高速に行うことが、FU の最も重要な機能である。FU は 100MHz の周波数のクロックで動作し、パイプライン動作により1クロック毎に積和演算命令を実行する。浮動小数点の乗算およびその結果を用いた浮動小数点の加算の結果が 10nsec 毎に生成されるので、200MFLOPS 相当の性能を有する。

FU には2つの状態がある。１つは IU における浮動小数点命令に従って演算を行うスレーブ状態、もう1つは LM に格納された積和演算表から積和演算命令を読み出して自律的に浮動小数点数の積和演算を実行するマスター状態である。２電子積分の漸化計算時には、FU は後者の状態となる。

IU が LM に格納されている積和演算表の開始アドレスおよび終了アドレスを FU に通知すると、FU はマスター状態に移行して、指定された開始アドレスから順番に積和演算命令を取り出して実行を始める。終了アドレスに達すると、FU はスレーブ状態に戻る。前述のように、FU の状態はステータス・レジスタとして IU から監視ができるので、IU は漸化計算の終了を検知できる。

FU 内部には、上記の動作を実現するため、LM のアドレスを生成する機能、漸化計算の終了を判定するためのアドレス比較機能、LM から読み出した積和演算命令をデコードする機能、演算に関係する中間積分の格

納されている FMEM（あるいは LM）のアドレスをデコードされた積和演算命令から生成する機能も有している。

なお、LM は MOE 専用 LSI の外部にあるため、内部に存在するメモリに比べてアクセスに要する時間が長くなる。そのため、1クロックサイクルで１つの積和演算命令を読み出すのは困難である。積和演算命令を1クロックサイクル毎に1つ実行させて漸化計算を高速に行うために、本 LSI は1回の LM へのアクセスで2つの連続した積和演算命令を読み出すようにしている。そのため、2つの命令のうち一方を1クロックサイクルの間は実行せずに保持しておく機能をも、FU は有している。

FU は、積和演算以外にも、乗算や加算の単独での演算機能をも有している。

2.6 PIU の機能

PIU は、MOE 専用 LSI 内部にあって、専用 LSI 内部とボード上の PPRAM-Link との間のインターフェイスを行う。内部レジスタを用いた IU との I/O および LM、IMEM に対する自動メモリ・アクセスの機能を有している。これにより、IU からの指示に基づいた送出パケットの自動生成およびその PPRAM-Link への送出、PPRAM-Link から取り込んだ自ノード宛てのパケットに関わるメモリ・アクセスおよび受領したパケットに対応する応答パケットの生成・送出を自動的に行うことを可能にしている。

PIU は 100MHz のクロックで動作するが、ボード上の PPRAM-Link では 50MHz クロックに同期してデータ転送が行われるため、PIU 内部で 50MHz クロックを生成して用いる。また、PPRAM-Link のデータを確実に内部に取り込むため、取り込んだデータの位相を自動調節する機能、および並列に入力される 17 ビットのデータ間のスキューをなくすための自動デスキュー機能も有している。

PPRAM-Link から取り込まれたデータは、一旦 PIU 内部のメモリ（FIFO）に格納された後、パケットに記述されたターゲット・アドレスに格納される。IMEM および LM が、ターゲットとして指定可能である。逆に送出のために専用 LSI 内部のメモリから送られたデータは、PPRAM-Link のパケットに変換された後に一旦 FIFO に格納される。FIFO に格納されたパケットは、PIU が PPRAM-Link 結合網の使用権を獲得した時に送出される。

3 おわりに

以上、MOE のハードウェア構成について述べた。MOE 専用 LSI のテープアウトを本年7月末、システム全体の完成を9月末に予定している。本プロジェクトに関する情報は、

http://kasuga.csce.kyushu-u.ac.jp/~moe

から入手可能である。また、本プロジェクトに関する問い合わせは、

moe@c.csce.kyushu-u.ac.jp

宛にお願いしたい。

分子動力学法への応用が可能な専用計算機　ITL-MD One

株式会社　画像技研　　高田　亮
ldd00553@niftyserve.or.jp

１．MD One開発の経緯

専用計算機を開発する事になったのは、５年前筆者と杉本先生の著書「手作りスーパーコンピュータへの挑戦」[1)]との出会いがきっかけであった。偶然書店で見かけたこの本を社に持ち返り話題にしたところ、面白そうだと言うことになり、検討を重ねた結果、自社のノウハウが生かせる分野であり、真剣に取り組めば、トップランナーになりうると考え、実現化を進める事になった。

今から考えてみると随分無謀な話だが、一面識もない杉本先生の研究室にお伺いしたところ、意外にも、杉本先生は共同研究の話を快く引き受けて下さり、資金面では東京都の「産・学・公助成事業」の助成金を頂けることになり、LSIの設計まではとんとん拍子に話が進んだ。一方一号機は、電総研の古明地様（以前からGrape systemに興味を持たれ、実際に使用もされていた）が、採用して下さる事になり、製品名も "ITL-MD One"と決まり、1995年の３月に何とか出荷にこぎつける事が出来た。古明地様には、その後も、実際のシミュレーションに応用した場合の評価をいろいろ行っていただき[2)]、伝道師的な役割を果たして頂いている。この裏付けがなければ、単に高速なハードウェアと言うだけで、一般の分子動力学法に使用できるシステムとして認知されるまでには、随分時間がかかったであろうと想像される。その他の様々な分野でも、実際のユーザの方々から、様々なご指摘、アドバイスを頂いたり、評価を頂きながら少しずつ応用範囲を広げて現在に至っている。

２．MD One System

以下に簡単に専用計算機MD Oneの構成と動作原理につき説明する。
MD One システムの構成は図１の通りである。

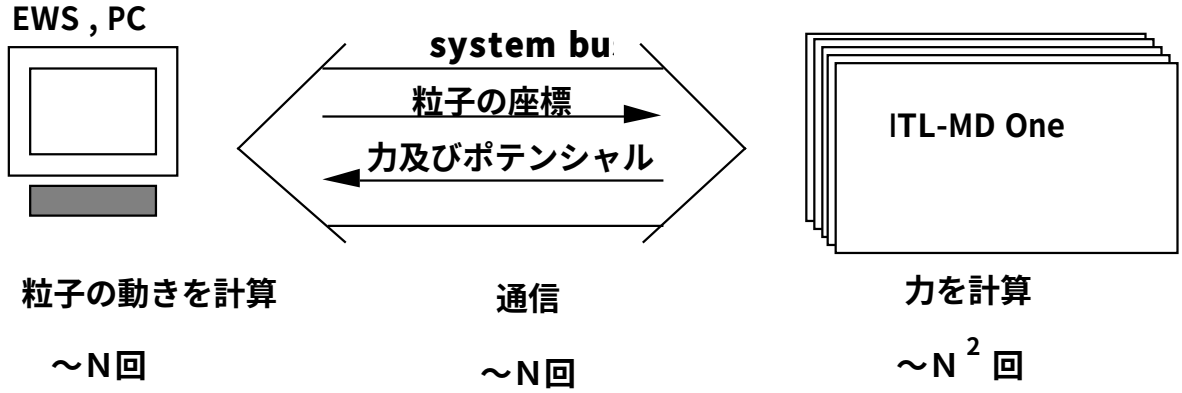


図１　MD One システムの構成

分子動力学法等の多体問題の計算は、個々の粒子に働く力を計算し、その計算に基づいて粒子の座標を更新して行くプロセスとして捉える事が出来る。これを計算量に応じて二つの部分に分ける。複雑だが計算量の少ない（Nのオーダー、例えば一万）粒子の動きを計算したりする部分は、市販のワークステーション等のホストコンピュータにより行い、単純だが計算量が極めて多くなる（Nの二乗のオーダー、例えば一億）力の計算等の部分は専用ハードウェアで処理する。ある粒子が他の全ての粒子から受ける力は、ハード的に加算されるので、本装置とホストの間の通信量はNのオーダーにしかならず、通信のボトルネックは起こらない。この様に高速でありながら、ある程度の柔軟性をもった計算機システムを構築出来る事がITL-MD Oneの特徴と言える。

以下に図２を参照しながら、Forceを計算する場合のLSIの動作につき説明する。

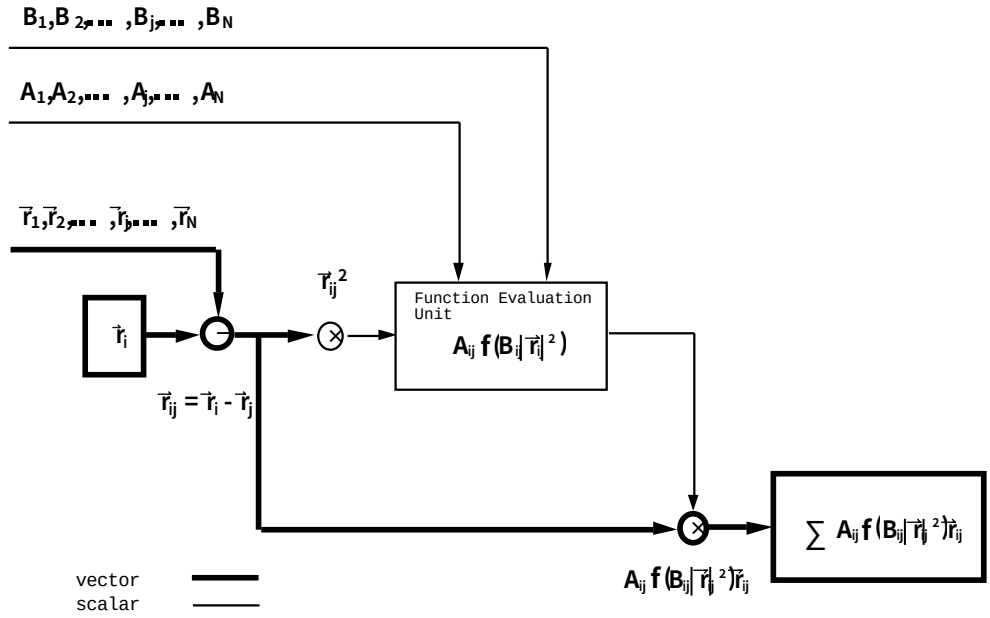


図２　Force計算時のLSIの動き

このモードでは、対象となる粒子はLSIのレジスタにセットされ、相互作用の対象となる粒子の座標が、メモリから順次供給される（①）。これらの粒子の座標は、減算器で相対距離とされ（②）、距離の自乗を計算するunitに送られる（③）。計算された相対距離の自乗は、関数発生器に送られ、メモリから供給された係数を使用して、これらの粒子間に働く二体力が計算される（④）。計算された二体力は、①で計算された相対距離ベクトルと掛け合わせられ、ここまでで、２つの粒子の間に働く力が計算されたことになる。最後に、これらの力は 積算器で積算され（⑤）、その結果注目している粒子に働く力が全て加算されて出力される。

MD One の性能は、例えば、LSIが4個搭載されたPCIバス用のもので、4Gflopsである。この値は、昨今の最高速のCPU（例えば、300MHzのPentium IIや、600MHzのAlpha 21164A）の性能から考えてそれほど高くないようにも思えるが、実際は、これらのCPUを搭載したPC、EWSと比較しても100倍程度のスピードがあるようである。MD Oneは多体問題専用に作られているため、実効性能とPeak性能に大きな差がないのに比べて、汎用のCPUは、普通にプログラムをコンパイルしただけでは、Peakの性能を引き出すことが難しいためと考えられる。

MD Oneのユーザはリピートで使用して頂けるケースが多いので、コスト／パフォーマンスについては高く評価していただいているようである。ただ最初にプログラムを移植して動かすまでに、多少時間がかかることもあり、実際計算している問題に、MD Oneをうまく応用出来るかどうかの判断もなかなか難しいので、気軽に使用する訳には行かないという面もある。今後は、この点を考慮し、応用面でのコンサルティングや、プログラムの移植まで含めたサービスを展開して行く必要があるだろう。また、電総研の古明地様、生命研の上林様のところのPEACH[2)]のように、MD One上で稼働するsimulation programを一般に公開していただける例も増えて来たので、普及にはずみがつくのではないかと期待している。

３．今後の専用計算機

専用計算機の未来像として、ここでは、MD Oneの様な専用計算機にまとをしぼって今後十年間程度を視野に置き、既存の技術でどのような展開が考えられるかを述べてみたい。

３．１　性能向上

計算機の性能を向上する要素として、いろいろ考えられるが、以下の３点に注目しよう（１）並列度（集積可能なゲート数）、（２）動作クロック、（３）データ転送速度。この十年間位で（１）は二桁以上向上しているのに対し、（２）、（３）は一桁プラスアルファ位のオーダーの向上であることから、今後も（１）＞（２）＞（３）の順で性能向上が見込まれると考えてもよいのではないだろうか。これまでの汎用CPUの性能向上は（１）と（２）の進歩の効果を併用して行われて来たが、ここに来て、その向上速度に陰りが見えてきた事が指摘されている[3)]。

専用計算機の側から見ると半導体の集積度が向上して、並列性が増して行ったと考えて、どのような事が起こるであろうか。並列度が極限まで上がって、１粒子当り一個の演算Unitを割り当てる事が出来たとしよう。３万体程度の計算を行う場合、３万個の演算Unitが並列で動作することになる。この事は、すでに天文学の並列計算機で、東大の杉本先生のグループが1000個のオーダーの並列計算機を1995年に稼働させておられる事から、決して荒唐無稽な話ではないと言える。

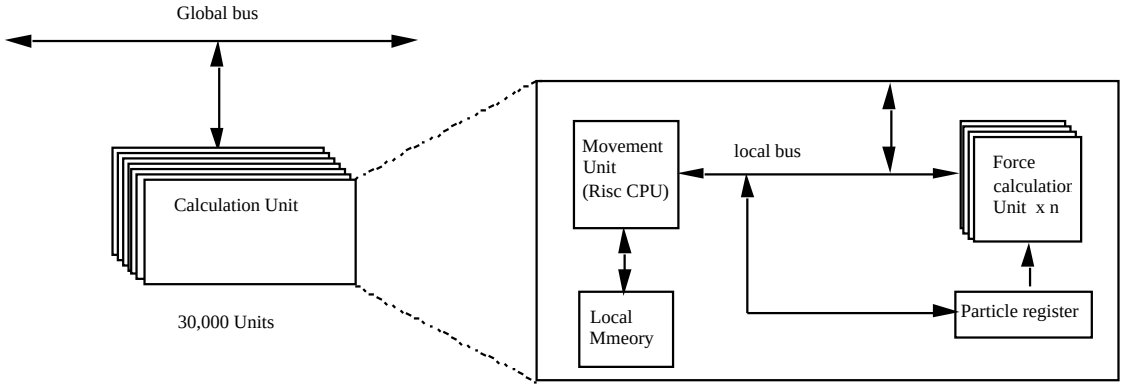


図３　３万個の演算Unitを持つ並列計算機

一つの演算Unitには、何種類かの相互作用を計算する演算機があり、また、相互作用の計算結果に応じて粒子のmovementを計算する機能があるとする。Movementの計算には汎用性を保つため、通常のmicroproccessorをcoreとして搭載すればよい。この事により、計算結果（相互作用等）をステップ毎に転送する必要がなくなり、データ転送速度のネックを低減することが出来る。この様な演算Unitでは、すべての粒子座標のデータが注目しているUnitに到達したとほぼ同時に、対象粒子の座標の更新が完了する。演算Unitをつなぐバスは常にフル稼働していて、そこには、更新された粒子座標が逐次現われて来る。ホストはこの粒子座標を随時取り込んでシミュレーションの進行状況をモニタすれば良い（実際のシミュレーションでは、境界条件による制約等があり、単純には考えられないと思われるが）。見方を変えれば、この様なシステム構成は、現在EWS等のホストと、MD Oneで構成しているシステムを一つの演算Unitの中に閉じ込めた様なイメージになっている。自分の担当分の粒子座標のみの更新に専念している、複数のMD Oneシステムが、協調して大規模な系を計算していると考えれば良い。

最終的にネックになるのは、データの転送速度であろう。簡単のため、クーロン力の計算のみを行う場合を考えてみよう。３万体程度の計算を行った場合、粒子間相互作用の計算には、約30演算が必要なので、1stepあたり、30000 x 30000 x 30 = 9G operations の計算が必要である。粒子の座標データを倍精度浮動小数点であると仮定すると、３万体では4 x 3 x 30000 = 360Kbyteとなる。これを1setp毎に更新して全ての演算Unitに送信しなければならない。仮に1Gbyte/sec の転送速度のバスを使用したとして、送信時間は360μsecを要する事になる。この場合9G operations /360 μsec = 25テラflops位までは、可能性がある事になる。一方分子軌道計算の研究者の方々が話題にされておられるような、ペタflops相当の計算能力を持つマシンを実現するためには最低でも40Gbyte/secの転送能力を持つバスを使用する必要があることになる。

実際問題としては、すべてのマシンで数万演算Unit を集積するわけでは無いだろう。ただ、この様なゴールを設定しておけば、あとは、より小規模なシステムへの応用は比較的容易である（並列度に応じて、各演算Unitを複数回使用すれば良いので）。

データ転送と言う側面から見ると、現在の並列計算機の構成では、隣接した演算Unit（例えば、同一LSI内）との通信は比較的高速に行う事が出来るが、物理的に離れたUnit（例えば、異なるボード上に搭載されたUnit）との通信にはボトルネックとタイムラグが生ずる。多体問題のシミュレーションにおいては、本質的に離れた所にある粒子からの寄与は少ないと考えられるので、階層的な構造のプログラムを使用する事などにより、データ転送速度のネックをある程度緩和出来る可能性はあるように思われる。

３．２　応用分野の開拓

またもう一つの大きなベクトルとしては、応用分野を広げる事が考えられる。MD Oneを発表して以来、さまざまな分野の方からお問い合わせを頂いた。ご検討頂いたものの、現在のハード／ソフトの構成では、必ずしも十分な性能が得られないため、見送りとなったケースが多数存在するのも事実で、このことはMD Oneの様な並列専用計算機に対する潜在的需要が高い事の証明でもある。これらのケースは、貴重な財産と捉えており、今後開発するシステムでは、こういった分野にも対応出来るようにしたいと考えている。MD Oneの構造は、複数の演算回路をパイプライン的につないだものなので、このパイプラインの構造をある程度フレキシブルに入れ替えられれば、様々な計算で使用出来ると考えられるからである。用途が広がることにより、ユーザが増え、コストも下がるので、分子動力学法による計算にもメリットが出てくる。

いずれにせよ、これからの専用計算機の開発には、ハードソフトともに、広い範囲の方々からいろいろなご協力を頂かないと、良いものは出来ない。次世代機の開発に関しては、今現在準備を進めているところであり、東京理科大の山登先生を始めとする様々な分野の方々のご協力が頂ける予定である。次世代機で上に述べた未来像のどこまで実現可能かはわからないが、現在の半導体技術で達成できるベストのシステムを構築する意気込みで取り組みたい。

４．後書に代えて

先日、生命工学研究所上林様のご紹介で、CBI研究会に参加させていただいた際、国立がんセンター研究所の相田様から、CICJS BulletinにMD Oneや専用計算機について夢物語でも良いから、何か書いてみないかとお誘いを受けました。その時は、気軽にご返事してしまったのですが、今になってこの様な拙文を載せていただいて本当に良いのか不安になっているところです。ただ、この記事が読者の方々が専用計算機に興味を持っていただくきっかけになれば幸いです。

参考文献

１）杉本　大一郎、　手作りスーパーコンピュータへの挑戦、講談社、(1993)
２）komeiji et.al., *J. Comp. Chem.,* **18**, pp1546-1563(1997)
３）Lineley Gwennap, ***日経エレクトロニクス*****,** **710**, pp55-65(1998)

たかた　りょう　TAKATA,Ryo
連絡先　〒182-0025　東京都調布市多摩川3-36-19
株式会社　画像技研
Tel. 0424-87-5003　Fax. 0424-87-5004
URL: http://www.cnet-tc.ne.jp/i/itl

副部会長就任にあたって

三戸 邦郎
Kunio.Sannohe@mitsui-chem.co.jp

吉田部会長を筆頭とする新年度役員体制がこの３月よりスタート致しました。吉田部会長にはご多忙なところ、３年目の慰留を無理にお願いした経緯がありましたため、不肖ながら私も副部会長として部会長をサポートすることになりました。

私は、1993年に編集幹事として情報化学部会に加えて頂いて以来、既に５年が経ち、当然、御役御免と考えておりましたところ、このような大役を仰せつかり、本当に勤まるものかと大いに不安を抱いているところです。副部会長の役割については、現・吉田部会長の３年前（副部会長退任時）の「具体的任務を担当すべき」との提案に基づき、それぞれ企画担当（時田先生）と編集担当（三戸）を主として勤めさせて頂く予定です。

さて、ここで、最近の部会役員会活動を、私が関わって参りました限られた視点からではありますが、ご紹介し、今後の活動に対する部会員の皆様のご意見・ご批判を伺いたいと思います。

私が、この部会と接点を持つ切っ掛けとなりました、部会誌・*CICSJ Bulletin*の編集という活動は、現状のようなインターネットの活用も余り盛んではなかったその当時においては部会員の皆様との直接接点のある貴重な活動の場であり、その都度、様々な課題が浮き彫りにされておりましたが、それは編集役員のみの議論の場であり、多くの場合、課題解決に至らないケースが殆どでした。細矢先生が部会長の時（約３年前）に、この２ヵ月に一度開催される編集会議のタイミングで他の部会役員の方々にもご出席願い、継続的な議論・検討が可能な場（編集会議を拡大した常任幹事会）を設けて頂きました（細矢治夫，「バトンタッチ」,*CICSJ Bull.*, 14(2), 12 (1996)）。 その後、この場を通じて、例えば、当部会の最大行事である情報化学討論会の活性化に向けての討論主題の決定や、会員名簿の見直し、部会Home Pageの開設等々を継続的に議論し、僅かながらも種々の改良施策が実行に移されてきております。昨年熊本で開催されました討論会も、当然、松岡先生（熊本大学）を始めとする世話役の諸先生方の大きなご尽力があってのことではありますが、関東・関西以外の開催としては参加者も多く、活発な討議が展開され、今後の更なる発展を予感させる結果となりました。なお、本年度の討論会は、北里大学の米田茂隆先生が世話人となり、11月10～11日に東京「こまばエミナース」にて、例年同様に構造活性相関シンポジウム（世話役：北里大学・梅山先生）と共同開催する予定です。

また、今年度は、新たに事業企画委員会を発足させ、部会員の皆様にとって、より有益な講習会の開催企画なども積極的に展開する予定を立てております。また、これまで何人かの役員の先生により、ボランティア的に進められて来ました、「部会Home Page」も環境を整え本格運用への移行が可能になりつつあります。

一方、吉田部会長が、本誌で以前に「現在、検討が進められておりますコンピュータ・ケミストリ関連の通商産業省のプロジェクトが実現の運びになれば、大きな推進力となるのではないかと期待をしております。」（吉田元二,「副部会長退任に当たって」， *CICSJ Bull.*, 13(2), 3 (1995)）と語っておられた、プロジェクトが大学連携産技制度の下、「高機能材料設計プラットフォーム開発」としてこの４月より4ヶ年計画で進められることになり、吉田部会長が本プロジェクトの推進母体である（財）化学技術戦略推進機構にて中心的なお立場で、ご活躍されます。これを期に国内に於ける当該分野の発展も更に加速されると期待されます。

情報化学部会としても更に活躍の場を広げ、皆様の期待に応え魅力のある部会活動を展開して行きたいと考えます。今後とも部会員の方々の率直なご意見・ご批判・ご提案を頂きますよう、宜しくお願い申し上げます。

第９回情報化学講習会
「２１世紀の研究開発と特許～情報発信と権利取得～」を終えて

平成１０年1月１２日（月）、少し開催日を間違えていれば、大雪にぶつかって大変だったことも予想されますが、まだ所々雪は残っていたものの、幸いにも晴天に恵まれ、日本化学会情報化学部会主催による第９回情報化学講習会・２１世紀の研究開発と特許～情報発信と権利取得～」を無事、盛会のうちに終えることができました。ご多忙な中、今回講師をお願いいたしました國枝興先生、平井保先生、岩澤まり子先生に、深く感謝いたします。特に、図書館情報大学の岩澤まり子先生には、企画から設定まで、ほとんどすべての面にわたってお世話頂きました。この場をお借りしまして、誠に失礼かとは存じますが、厚くお礼申し上げます。そして、年初のご多忙中、寒い中、講習会にご参加いただきました皆様に、深く感謝いたします。

最初にご講演頂きました、國枝 興 先生に、プロローグでお話頂きましたように、企業のみならず、大学、研究所等も含めて、いや、科学技術に関連する個人にとっても、２１世紀に生き残るためのキーワードは、ハイテクノロジー、国際化と並んで、特許だと言われて久しいように思われます。特に、不況が長引く昨今、いかにして効率的に有効な科学技術の研究開発を行い、有効な特許を獲得するかは、多くの科学技術関連企業にとって、極論するならば、死活問題と言えるかもしれません。しかしながら、インターネット時代の特許情報の取得方法、近年の特許の動向、特に国際的な動向、現代の特許制度が抱える問題点など、ここ数年の特許に関する急速な動きに関しては、あるいは文責者の勝手な思いこみかもしれませんが、意外と、その知識は、広く普及してるとは言い難いように思われます。この講習会では、特にこうした点を含めてお話しして頂くと共に、十分な質疑応答の時間を設け、ご講演いただきました先生方と参加して頂きました皆様との直接の対話により、より現場での問題に対応した情報をお持ち帰りいただこうと企画しましたが、これらの点については、やや贔屓目があるかもしれませんが、ある程度、当初の目的を達成できたのではないかと考えています。

今後、どういった形式の講習会が望まれるのか～もちろん、どんな内容の講習会の企画が望まれるのかも含みますが～について、当日ご参加いただきました皆様はもちろん、情報化学に興味を持って頂いています皆様のご意見をお聞かせいただけましたら幸いです。日本化学会情報化学部会では、今後とも、様々な試みを行っていきたいと考えておりますので、どうかよろしくご参加のほど、お願いいたします。

（文責・高木達也）

第 10 回情報化学講習会
「生体分子シミュレーション～基礎から２１世紀への展望まで～」を終えて

平成１０年1月２２日（木)、第９回情報化学講習会の１０日後、同じく日本化学会情報化学部会主催による第１０回情報化学講習会、「生体分子シミュレーション～基礎から２１世紀への展望まで～」を無事、開催することができました。他組織における同様な企画が重なったこともあり、やや、参加者数が寂しかった感じはありますが、そのぶん、より突っ込んだ質疑応答ができたのではないかと考えております。ご多忙な中、ご講演いただきました電子技術総合研究所の古明地勇人先生、分子科学研究所の岡本祐幸先生、味の素中央研究所の鈴木榮一郎先生に深く感謝いたしますと同時に、ご参加、ご討論いただきました皆様に、心よりお礼申し上げます。また、本講習会の企画担当を致しました本文の文責者としましては、日本化学会情報化学部会幹事の先生方のご助言、ご助力に対し、また、事務的な面で大変お世話になりました情報化学部会事務局の方々に、この場をお借りいたしまして、厚く御礼申し上げます。

分子動力学法、モンテカルロ法を中心とする、生体分子シミュレーションは、近年、急速に普及してきたように思います。本講習会は、より多くの方に生体分子シミュレーションに興味を持って頂くこと、そして、興味を持って頂くだけでなく、実際に計算を行うのに必要な（しかし、意外と教科書では得られない）基礎情報を取得していただくこと、さらに、今後どのような発展が予想されるか～いわば夢を持っていただくことを目的としました。そこで、古明地先生には、特に、生体分子シミュレーションの実際的な方法論について、詳細にわたり解説していただきました。先述のように、このような情報は、意外と書物や論文からは得られないことが多く、大変貴重な情報ではなかったかと思います。また、岡本先生には、徐冷モンテカルロ法、拡張アンサンブル法等を中心に、蛋白質の folding simulation について解説いただき、１つの蛋白質のアミノ酸配列情報のみから蛋白質の立体構造を予測しうる可能性を示していただきました。最後に、鈴木先生には、核磁気共鳴法を用いて、蛋白質の動的構造に関する情報を得るために分子シミュレーションを導入する手法について、解説頂きました。核磁気共鳴法は、元来、動的な構造情報の取得を得意としているはずで、今後より一層の発展が期待される分野かと存じます。

「講習会の究極の目的は、講習会が必要なくなることだ」とは、どなたの言葉だったか失念しましたが、なかなか妙味のある言葉のように思います。残念ながらというべきかどうか、幸いにというべきか、微妙なところではありますが、生体分子シミュレーション、いえ、分子シミュレーションに関しては、まだまだ講習会が必要な状況が続くように思われます。日本化学会情報化学部会では、今後とも、時代に即した、充実した講習会を企画し、情報化学の発展と普及の一翼を担っていきたいと考えております。皆様のご意見、ご参加を心よりお待ち申し上げております。

（文責・高木達也）

第９回ケモメトリックス・ワークショップ開催報告

キッコーマン（株）研究本部　相島鉄郎
LDH00052@niftyserve.or.jp

シカゴに本社を置くSPSSはStatistical Packages for Social Sciences　という名前からも分かるように、60年代から汎用統計ソフトウェアとして社会科学分野で広く用いられてきた。しかし従来、もう一つの汎用ソフトウェアであるSASがケモメトリックスを初めとする化学分野で多用されてきたにもかかわらず、実験計画法などの研究を効率的に遂行するために必要な手法が備わっていないことなどから、SPSSの自然科学分野での利用例は必ずしも多くはなかった。ところが近年、ニューラル・ネットワーク、さらに実験計画法もオプションとして製品ラインに加え、自然科学分野の研究者に対する利便性を急速に増している。そこで11月20日、日本化学会会館６階会議室において10時から、日本たばこ産業・榊 武志博士による開会の挨拶に続き、シカゴから招いたAnthony J. Babinecが「汎用統計ソフトウェアによる実験計画法と他変量データ解析の基礎と応用」について、SPSSのバージョン7.5とニューラル・ネットワークおよび実験計画法のソフトウェアを駆使し下記の内容で講演した。

１　多変量解析法
- 一変量データの取り扱いとｔ-検定及び分散分析
- 試料数が少ないデータの検定
- 探索的な多変量解析-主成分分析、因子分析
- パターン認識法-判別分析
- 検量法-重回帰分析、非線回帰分析、データ変換

２　実験計画法
- 要因計画法
- 部分要因計画法
- D-最適化法

３　ニューラルネットワーク
- パターン認識への利用
- 検量への利用
- 適用に際しての留意事項

約５０人の聴衆中には、外国からの留学生も２人含まれており、海外におけるケモメトリックスの認知度合いの高さを示した。今までのワークショップ同様、今回も全体的には企業からの参加者が多く、ケモメトリックスの手法の応用面への関心の高さをうかがわせた。今回の講演は、実験計画のソフトウェアを用いた、D-最適化法など最新手法の紹介や美しい応答曲面の図示はあったものの、多変量解析については従来手法の基礎に関する丁寧な説明と応用例の紹介が多かった。しかし、ややもすれば基礎なしに一挙に応用に走る傾向も少なくない、我が国におけるケモメトリックスの現状を考えれば、重回帰分野における残差検定などの基礎的な話題は、本ワークシップの本旨に適うものである。また、さまざまな分布を示す２群データのパターン認識に判別分析とニューラル・ネットワークを適用し、試料分布に応じて両手法の分類結果がいかに異なるかの三次元図を用いた比較は、統計的手法とニューラル・ネットワークの違いを理解するためには非常に有用であった。

お茶の水女子大学・藤枝修子教授の閉会の挨拶により講演会を終了した。講演終了後、約３０名が参加した日本化学会会館５階会議室における約２時間の懇親会は、参加者どうしの意見交換はもととより、演者に対して様々な質問も発せられ非常に有意義なものであった。

あいしま　てつお　AISHIMSA, Tetsuo
連絡先　〒278　千葉県野田市野田399　キッコーマン（株）研究本部　電話 0471 23 5980

CICSJ Bulletin Vol.16, No.4, August 1998

目次

CICSJ INDEX にもどる

コンビナトリアルケミストリー　特集の序に変えて

科学技術振興事業団　　有國　尚

arikuni@yo.rim.or.jp

コンビナトリアル・ケミストリー（コンビケム）の登場は，指数関数的な化合物の合成を可能にした。しかしながら，多数の化合物群を効率よく合成しても，その生物評価に時間がかかってしまえぱ効果は半減してしまう。この課題は，ハイスループットスクリーニング（HTS）法と呼ばれる自動評価システムによって解決され，多数の化合物群を自動的にスクリーニングし，生物活性を調べることが可能にしている。

しかしながら、コンビケムといえども，無限にある組み合わせのすべてを合成することは，理論的には可能でも非現実的である。そこで，いかに優れたライブラリーを効率よくデザインできるかが最も重要なポイントになってきている。ターゲットとなる化合物に焦点を絞って，ライブラリー（化合物群）をデザインできないか一一。そういった方向に欧米でも視線が向いている。コンピューター上で化合物のデザインを効率よく行うことができれば，時間とコストが大幅に低減できるはずである。

最近、立体構造ライブラリーからターゲットとなる化合物を絞り込んでいく「ストラクチャー・ベース・フォーカシング」という考え方が，話題を集めている。

目的とするところは、ライブラリーの多様性を維持したまま、いかに合成すべき化合物の数を減らすかと言うことである。

創薬の世界では、ターゲットとなるタンパク質や受容体の構造に基づいて、薬をデザインする手法（ラショナル　ドラック　デザイン）が存在し、成果を挙げている。

こうした手法を用いて、特定の機能を発揮するために必要不可欠な構造の因子が明らかにし、分子の立体的構造情報を含むさまざまなデータベースから分子を選抜することによって，ターゲットを志向した多様性のある化合物ライブラリーを得ようとする試みが為されている。

今回、『コンビナトリアル　ケミストリー』特集を組むにあたり、こうした最近の動向を含め、大学や企業において実際にコンビケムに直面している方々に原稿をお願いした。今の状況を生々しく紹介して頂く。

先ず、吉富製薬大阪研究所の井上佳久さんには、『Combinatorial Chemistry & High-Throughput Screening　最近のトピックス』と題して、コンビケム全般に渡る最近動向の解説とコンビケムの応用の広がりについて文献を紹介して頂く。

次に、科研製薬の井口氏には、計算化学の立場から、製薬企業内での取組、計算化学からのアプローチについて、書いて頂いた。さらに、大阪大学の深瀬浩一氏には有機合成化学のから見た現状について書いて頂き、ライブラリー合成の際の問題点等にも触れて頂く。

最後に、Structural Bioinformatics Inc. のK. Ramnarayan氏に『Unification of Bioinformatics, Combinatorial Chemistry and Chemoinformatics with a Structure Based Approach』と題して、「ストラクチャー・ベース・フォーカシング」について、

解説して頂く。この特集が皆様に一層の関心を持って頂く機会になれば幸いである。

Combinatorial Chemistry & High-Throughput Screening

最近のトピックス

吉富製薬株式会社大阪研究所　　井上　佳久

inouey@yoshitomi.co.jp

１．はじめに

コンビナトリアルケミストリーとHigh-Throughput Screening（以下ＨＴＳ）のブームに火がついてはや４年が経過し、製薬各社は基礎技術の導入、自動合成装置の導入、固相合成の確立、ロボットの導入などが一段落していると思われる。しかし最近は製薬だけでなく、農薬、触媒、先端材料開発などに応用されてきている。本論では最近のトピックスをまとめた。

２．コンビナトリアルケミストリー最近の話題

２．１　パラレル精製

スプリット＆ミックスでは不純物などの問題、パラレルでは精製などの問題がある。ほとんどの会社がパラレル合成を行っているようだが、そのうち精製を高速化する方法の開発が急務である。その中には、Glaxo-Wellcome社でされている－２０℃に凍結して液―液抽出するlolly-pop法など、Axys社でされている固相抽出、毎時40試料を処理するBiotage社・山善などのカラム、イオン交換樹脂、リリー社・Parke-Davis社・Signal社などのpolymer-bound quenching reagent等がある。システムとしては、コンビケム社でされているようなＬＣ／ＭＳを用いた分析・分取システムが構造解析を考えた上では有利だと思われる。アボット社もHPLC/ MS/ UV/ ELSD (evaporative, light scattering detection)でシステム構築をしている。これにバリアン社やブルカー社のＬＣ／ＮＭＲを接続するとcharacterizationも同時にできることになる。例としてパンラブ社ではBiotageを用いて月に６０００反応を精製している。長期間の安定性の問題と、純度は９５%以上ならよいのか？という問題とアッセイの精度の問題がまだ残っている。

２．２　反応の追跡

精製に時間とお金をかけるなら、当然反応条件を最適で行う方が目的物は多くとれるし精製の手間も少なくなると思われる。そういう意味から反応をＮＭＲやＩＲで追跡する方法も開発されている。アルゴノート社ではＩＲを用いてビーズ１粒で状態を測定できるようにしている。アフィマックス社やNovartis社もＩＲやＮＭＲで行っている。小林修先生は新規のＣＰ－ＭＡＳ用ＮＭＲチューブを開発された。

２．３　合成に関して

traceless linkerの開発が盛んである。珪素を用いたEllmanらの方法、SK&Bの方法、Oxford Diversity社の方法は興味深い。他にもSignal の方法やoxime樹脂を用いたメルクの方法もある。ビーズとしては、ArgoGelはポリスチレンにＰＥＧ部を２つ導入したもので、TentaGel よりも安定で、またローディングも多く、膨潤性も異なる。ArgoPoreは非常に固く、膨潤性があまりなく、アニオン化学に有利である。液相合成、固相合成に次ぐ第三の波としてFluorous systemが

出てきた。フルオロカーボンが通常の溶媒に溶けないことを利用し、固相合成の代わりに使用するものである。

また、スプリット法だけでなくパラレル法でも使えるIRORIのraidofrequency memory tagsに対抗したChironのTranSortも興味深い。機械としてはアルゴノートのQuest210、CharybdisのCalypso、RobbinsのFlexChem、ＨＰのHP7686液相合成装置、Tecan/Applied Biosystemsの固相合成装置、モリテックスのL-COS、大日本精機のCC Factory、国産化学のコンビワークス、東京理化のCCS-600Vなど多種開発されてきている。Orchid Biocomputerはマイクロチップ上に反応装置を組み込んだシステムを開発している。３センチ角のチップに１４４反応槽がある。SK&Bでオーファンレセプターを１００個近く見つけているが、それのリガンドをこの手法を用いて探索するそうである。

２．４　試薬など

参考になる試薬やビーズ関連情報のリンクは以下にあるので、参考にしてほしい。

http://www.orgsyn.riken.go.jp/CombiChemJ.html

http://chemlibrary.shimadzu.co.jp

２．５　コンビナトリアルバイオロジー

以前から日本では抗生物質の開発で使われていたこの技術が最近また注目されている。Kosan社はacyl transferase, keto synthase, keto reductaseなどの酵素反応を行うものと, acylcarrier proteinがクラスターとして存在していて，配列を変えたりしてpolyketidesを合成している。最近は別の菌へ特定の酵素を導入して新規な構造をつくる技術へと展開されている。成功例としてChromaXome 社の耐性菌へ有効な新規構造の発見などがある。

２．６　コンビナトリアルケミストリーは他の分野にも

医薬・生化学以外にも適用されて成果をあげている。触媒が一番多いが、薄膜などにも応用されている。文献のみ以下に示す。

JACS,118,8983('96),

JOC,61,8940('96),

Mol.Div.1,266('96),

TL,8313 ('96),

Acc.Chem.Res.,29,169 ('96),

Chem.&Brit.,33,18('97),

Anal.Chem.,69,3413(1997),

Angew.Chem.Int.Ed.Engl.,36,1610& 36,1703('97),

Nature, 389,944('97),

Science,279,837('98),

現代化学,1998年2号,p12,

化学４月号,pp70-71('98)

中でも小林修先生の開発されている触媒のカプセル化(JACS,)や魚住泰広先生の開発されている固定化ＭＯＰ触媒(TL,)はコンビナトリアル分野だけでなく，広く工業化学の上で重要だと思われる。

３． ＨＴＳ最近の話題

３．１　ミニチュア化

９６穴プレートでのアッセイから最近は384穴，864穴へ移行しつつあるが，欧米では1536穴，6144穴へと移行している。オーロラ・コーニング・EVOTEC・Imaging Research・Greiner Lab.などが開発しており，り，PharmacopeiaやNovartisなどの大手が導入している。これらの技術を確立するには１ナノリットル以下の精密なdispensing が必要で，EVOTEC・Cartesian・Packardなどが開発している。インクジェット式がよいようである。検出器も高精度のものが要求される。これらに対応すべく，Imaging ResやSIBIAなどはCCDカメラを用いたものを開発中であり，EVOTECはfluorescence correlation spectroscopy(FCS)を開発している。昨年のAnal.Chem.にFCSを用いて１分子を検出したという報告もあるので，高精度なのであろう。他にLJL Biosci.社もある。こちらへ移行すると，試料の量が少なくてすむことや，１日に１０万検体処理できる速度やコストのメリットがある。しかし測定ミスや溶媒の揮発性も考慮に入れる必要がある。

３．２　代謝関連

今年の日本薬学会では代謝を考慮したＨＴＳのシンポジウムが開催されたが，欧米の先進性が印象的だった。ここではその内容は省略するが，薬物開発の初期の段階でＡＤＭＥ(absorption, distribution, metabolism, excretion) を見ておくことが有益なのは当然である。GENTEST 社は96穴プレートでCYP1A2,CYP2C9,CYP2C19,CYP2CD6,CYP3A4 などとのアッセイを既知の薬物を用いて行う方法を開発している。高速で薬物候補のスクリーニングも可能だし，薬物相互作用の参考データにもなる。Xenometrix社は自動gene profiling system を開発しており，候補薬剤の遺伝子発現への作用，つまり効率や毒性をDNAや蛋白質の損傷，酸化的ストレス，炎症性応答などで検出できるようにしている。

昨年のファインケミカルズに書いたAcacia 社と同様な方法をとる会社も増えている。Rosetta Inpharmatics 社はイースト菌でのノックアウトや種々の化合物への作用をみてターゲットとしての評価を行ったり，リード化合物の解析も行っている。BMS では種々のP450酵素をHepG2 やSV40T に移植して毒性をみている。Merck やChiron では Caco-2を用いて吸収性をみているが問題点が多いようだ。SKB ではHPLCのカラムの支持体にimmobilized artificial membranes (IAM)を用いて吸収をみている。エンドセリン拮抗薬SB-209670 を最低ラインとしており，１０倍以上よいSB-217242を見出している。

３．３　マスを用いた方法

今夏のアメリカ化学会ではマスを用いた方法の特別セッションがあるので，今後いくつか報告されるだろうが，Chiron では混合物の状態でマスをとり，受容体との結合実験の後に再度結合したもののマスをとり，何が結合していたかを見出す方法を報告している。これだと簡単な構造活性相関まで可能である。スプリット＆ミックス法で合成していようが，パラレル法で合成していようが，混合物でアッセイする場合に使用可能である。事前に合成した構造がわかっているからというのが特徴である。

３．４　その他の方法

昨年アボットのFesik博士らの提唱したSAR by NMR は非常にインパクトのあるものだった。蛋白質の分子量の上限はあるものの15Nでラベルされた数百mgあればmMの活性のものが見出されたらそれを最適化し，さらにそれが結合した状態で次の分子を結合させ，それを最適化し，それらの分子同士を結合させるとnMオーダーの阻害剤になるというものだった。混合物でもアッセイができるので１日に数万可能だという。未確認情報だが，ラベルしなくても蛋白質の構造解析がすんでいなくても，低分子化合物のみ分けてNMRを測定できて，１日千個だがスクリーニングできる手法が開発されたらしい。また，Ｘ線結晶構造解析を行うための結晶に低分子化合物をソーキングして，リード化合物を見出しながら，結合位置が確認できる方法も開発されているようだ。

一昨年のファージを用いたエリスロポエチンのアゴニストペプチドを見出す方法は見事だったが，それに続いてG-CSF の非ペプチドのアゴニストを見い出したというScienceの報告が飛び込んできた。

４．　インフォーマティクス最近の話題

４．１　ライブラリ解析

既存の薬物のフラグメント(drug-like fragment)を知識ベース化してライブラリ構築に役立てようという試みが数社でなされている。また薬物の活性解析用にChemical Design社の方法を用いて4-point pharmacophore を用いた例として，12個の既知のfibrinogen 拮抗薬は10万化合物のランキングでトップ202に入ったとRPR 社から報告されている。BMSでは遺伝アルゴリズムを使用して400億化合物をクラス分けしているが，他の方法よりも時間的にも内容的にもよかったらしい。KnollのPerry博士はmost descriptive compound (MDC) method はsphere exclusion よりもよく，Ghose & Crippen atom-type記述子は166個のMACCS keyのセットよりもよかったと報告している。一方アボットのBures博士は２次元記述子よりも３次元記述子の方が最近はよい結果を与えていること，K-Meansクラスタリング法とsphere exclusion selection method は Ward解析法とほぼ同程度の結果でかつ時間は短かったと報告している。

４．２　スクリーニング

ＤＯＣＫを用いて合成前にスクリーニングするという手法がある。Kuntz教授らはカテプシンＤの阻害剤を開発するにあたり，可能性のある300億化合物を1000個に絞って合成し，１回目は70nMだったが，さらに展開してnM以下の強い阻害剤を見出した。

５．　参考文献

コンビナトリアルケミストリー（化学同人）

末永俊朗、J.Mass Spectrom.Soc.Jpn., 45(3), 265-288 (1997)

井上佳久、ファインケミカル、26(16)、９月１５日号、pp5-35 (1997)

「雑感：計算化学者から見たCC&HTSの取り組み」

科研製薬（株）創薬研究所　　井口　潔

inoguchi@lab.kaken.co.jp

コンビケムの特集記事が組まれるとのことで、恐らく大手企業の方々の壮大なＣＣ＆ＨＴＳのシステムや成功例等も紹介されることであろうと容易に想像が出来き、赤面の思いは禁じ得ないが、この拙文は、製薬企業における取り組み」と言うよりは、むしろ遅れてＣＣを始めようとする我々の苦労話に過ぎないことを、予めご容赦戴きたい。

さて、私自身、遅ればせながら「コンビナトリアルケミストリー」という戦略を具体的に耳にしたのが９５年半ばのことで、当時の出向先（Molecular Research　Institute、在Palo Alto）のボスであったGildaLoew所長と議論したのが、そもそものきっかけであった。

その当時私は、計算化学技術のトレーニングも兼ねて、あるオリゴペプチドのミメティクス化研究に従事していた。極めて具体的な話であるが、設計したファーマコファーを用いて化合物データーベース検索を行い、得られてくるヒット化合物について購入できるなら購入し、手に入らなければそのものを合成するという手法が通常採られている。私は丁度その過程でコンビケムの事を知ったが故に、出来るならばヒットから選ばれた幾つかの提案化合物について、それぞれの母核を基にした数十個規模づつのライブラリーとして用意し、実際のアッセイ系を用いたファーマコファーバリデーション＆リファインメントを一気に、しかも効率的に行いたい、という漠然とした欲望を持ち始めていた。一方、「非合理的」という「枕詞」付きで、しかも大規模スクリーニングと結びつきながら台頭を始めていたコンビナトリアルケミストリーに対する所長の当時の評価は低く、「既に多くの化合物ライブラリーが手元にあるなら話は別だが、我々は最適化をやっているのではなくリードファインディングを＜合理的に＞やっているのだから、数を期待してはいけない。現状ではまだ多品種大規模ライブラリーを用意するのも容易ではないし、むしろ数はペプチドリードに任せておいて、計算化学者であるべき我々はあくまでも＜合理的に＞用意する数を絞って提案することこそ本領だ。」、という結論であった。

ArQuleのHogan氏（特徴ある新規母核ライブラリーを設計、合成するコンビベンチャー創始者の一人）とも旧知の仲であった所長は、その時既に、コンビケムに対して確たる考え方を持っていたようである。　無論、計算化学者のスタンスとしてはそれが当たり前なのかも知れないが、医薬品化学者の一人として、はばかりながらリード発掘の難しさ（他人様の「ミーツー」を見つけて育てるのでさえ苦労が多い）を知る私にとっては、ターゲット情報に忠実になりつつも、出来る限り「緩い縛りの」バイアスドコンビケムライブラリーが持つ可能性の方に密かに惹かれるようになった。「計算化学者への心身の改造が不十分」、という反省も少しは持つべきなのは重々承知しているが、現実問題としてリードファインディングの過程にあるこの時点においては、コンビナトリアルな合成化学的フレキシビリティーを考慮しつつ選抜化合物を料理しない限り、その母核のポテンシャルが発揮されなままにチャンスを失うばかりか、苦労して設計したファーマコファーや分子生物学的構造要件を正当に評価出来ないままにテーマの終焉を迎えてしまうことにもなりかねない。やはり短期間に構造要件の確認ができ、構造活性相関データ（ファーマコファーリファインメント）として利用できる価値を考えれば、ある程度の規模のコンビケム

であっても、それはドラッグデザインに欠くことの出来ないツールたり得るのではないだろうか。

私が出向先から戻って来た９６年春は、丁度日本でコンビケム研究会が発足し、一部大手製薬メーカーとコンビケムベンチャー間の提携話も幾つか持ち上がるなど、コンビケム関連技術が一気にクローズアップされるようになった時期でもあった。

願ってもないタイミングであり、お陰で関連情報収集もかなり進んだが、現場の状況はといえば、かろうじてペプチドミメティクスの流れから、コンフォメーション固定を目的とする「異常アミノ酸ライブラリー」構築が一部で立ち上げられたばかりで、依然として戦力の多くは旧来の医薬品化学、即ち通常の改良新薬創製につぎ込まれていた。

最近になり、ようやく現場でもペプチドミメティクスや固相合成に関する理解が広がり、ホットしたのも束の間、今度は生物評価を担当する研究員から不満の声が聞かれるようになってきた。

即ち、「よそでコンビケムと言えば、ＣＣ＆ＨＴＳと同義であり、やるなら徹底的に化合物数を増やしてくれないと、中途半端な数では、サンプル処理能力向上を目的に新しいハードを導入する訳には行かない。」、といった意見である。分母、即ちスクリーニングサンプル数が潤沢でない我々にとって、これは昨年頃から浮かび上がってきた大きな問題の一つである。我々の側は、出来るだけ合理的にバイアス（絞り込み）しながら、出来るだけ多種母核の化合物ライブラリーを供給しようと努力している最中ではあるが、それでも合成可能な化合物総数の規模は限られており、例えば、「万に一つヒットするのが当たり前というのがＨＴＳの原則で、それ以上のヒットがある場合は、アッセイ系の感度に問題があるか、サンプル化合物の顔が極めて偏っている」といったご意見や、数百、数千穴のマイクロプレートを使った「ウルトラＨＴＳ」等という極めて「大陸型農業的（広大な土地にセスナ機から種、農薬、肥料をばらまいて、巨大な機械で収穫）」な方法論等が情報として先行している現状では、プレート一枚分にも満たないサンプル数を苦労して用意していること自体、非常に説明し辛い。問題の解決には程遠く、奇妙な言い訳にな

るかもしれないが、生物評価側の人々には、わが国が元々得意とする「集約型米作農業的」な評価系（少し説明を加えるならば少ない土地から最大限の収穫量を期待するべく、整然と苗を植え、的確に給水、肥料を与える戦術）を考案してもらうよう提案している。

即ち、少ない化合物数から出来る限り多くの情報を得ることを主眼とし、必ずしも「High Throughput」ではなくとも、「High Informative」なアッセイ系、次の段階の化合物設計に際し確たる方向性を示す事が出来るようなシステムを構築してもらうようお願いしている。

現実に、「all or nothing」の評価結果では構造活性相関把握が難しい事が多く、少ない化合物数でも色々なプロファイルを併せて観測頂ければ（具体的なアイディアは社外秘ということで割愛する）、非常にやりがいも頼りがいもある情報が得られてくるものと確信している。

何れにしても、多様性評価やファーマコファーベースのバーチャルスクリーニング等、それなりの根拠をもって設計化合物が選抜でき、且つ、効率よく意図したライブラリーが設計でき、そして何よりも現実に合成可能か、といった一連のプロシージャー、これこそが我々の側から解決

すべき問題であることは既に周知のことであり、エキスパートシステムや、データーベース(検索技術も含めて）の果たす役割も益々大きくなって来ている。

　そうした状況であるが故に、先達でもあるCICSJ読者諸氏からも、悩めるこの子羊に、何かアドバイスや叱咤、激励等を戴ければ幸甚の至りである。最後に、お目汚しも甚だしい文章にお付合い戴いた読者諸氏に心より感謝すると共に、このような機会を与えて下さったCICSJ Bulletin編集委員諸氏にもこの場を借り、深謝したい。

以上。

有機合成化学者の立場から見たコンビナトリアルケミストリー

大阪大学大学院　理学研究科　　深瀬　浩一

koichi@chem.sci.osaka-u.ac.jp

コンビナトリアルケミストリーのブームが続いている。これとハイスループットスクリーニング（HTS：ロボットを大々的に使用した自動アッセイシステム），とバイオインフォーマティクス（生命情報科学：遺伝情報などの生命情報を，蛋白質・核酸などの生物分子の立体構造をもとに理解する）を合わせた３点セットは創薬のトレンドの決まり文句になってしまったかのようである。

コンビナトリアルケミストリー自身の変化も速く、また報告数も増えているだけでなく、筆者は医薬品開発の素人であるので、現状を把握するのは難しい。しかしながらはっきりとした傾向は見えてきたので、それについて手短に述べさせていただきたい。

すでに多くの方はコンビナトリアルケミストリーの原理は御存知であると思うので、ここでは述べないが、当初は枯れ草の中から針を見つけだすように、多数の化合物をランダムに合成しその中から目的化合物を見出す手法のように考えられていた。実際この手法はビーズ状の固相担体に用いて、固相上でbinding assayを行う場合は非常に有効である。溶液中でアッセイを行う場合でも、デコンボルーションと呼ばれる手法に代表されるように、混合物のままアッセイを行ってもその中からヒット化合物を見出すことのできる手法が数多く報告された。このことが過大に評価され、コンビナトリアルケミストリーといえば，すぐにでも数万～数百万にもおよぶ化合物が合成できて，絨毯爆撃的にアッセイを行うことで，たちまち医薬品候補が見つかるかのようなイメージができあがってしまった。

現実はそう簡単ではなく、多数の化合物を効率的に合成するためには、高効率かつ高選択的な合成手法が必要となる。そのような合成法を確立するためには相当な努力と時間が必要とされる。もう半年以上も前のことではあるがSIBIAでは今や一人で週に２００化合物を合成している。Arrisでは昨年３月には月２万４千化合物が合成されたそうである。Tregaでは今年中に100万化合物の合成が計画されている。これらのベンチャーでは合成にそれなりの人的資源が割かれている。アッセイについても混合物のアッセイから単一化合物のアッセイが主流になってきた。絨毯爆撃ははずれが多いのでそれなりのデザインしたライブラリーが用いられている（筆者には当初から絨毯爆撃的ではなかったように思われる。）。

イメージが先行していたのが、落ち着いてきたということであろうか。コンビナトリアルケミストリーは通常の研究に無理なく組み込める形になってきた。例えばコンビナトリアルケミストリーでは数が重要ではない（量より質）というのが最近の傾向だ。ある程度の数は合成するにしても、目的化合物を絞って化合物の総数を減らす方向にある。例えばCombiChem社ではバーチャルでは500億個の合成可能なライブラリーを準備しているが、実際に保有しているライブラリーは１万個である。ただしアッセイに用いる化合物の数が１万個ということで、新しいより適当な化合物が合成されれば、ライブラリーのメンバーを入れ替える。（彼等の用意しているアッセイでのヒット率がだいたい1%くらいになるようにライブラリーメンバーを選ぶ。ヒットの基準は未公開。）グラクソ-ウェルカムでは20万化合物くらいのCompound Libraryを用意している。

ライブラリーの合成手法でいえばスプリット&プール合成による混合物合成からパラレル合成による単一化合物の合成が主流となっている。現在はさらに合成化合物の精製システムまで考慮して、最も速くライブラリーを構築するためにはどのような戦略で合成を進めていくのかが、重要となってきている。例えばライブラリー構築のためには短段階合成ではあっても、各段階は高収率であることが必要であった。そのための合成のプロトコールを確立するのに最も時間を要し、実際のライブラリー合成は律速ではなかった。適当な自動精製システムを組み入れることで、ある程度収率が低くてもライブラリー構築に用いることが可能となる。

当初は固相合成が中心であったが、現在は固相合成でも液相合成でもとにかく化合物を迅速に作っていこうということである。液相合成も普通に用いられるとなると，コンビナトリアルケミストリーが従来技術とどのような差があるのかという疑問が生じるだろう。実際のところコンビナトリアルケミストリーでそれ程特別なことが行われているわけではない。重要なことは効率合成（いかに短時間に多くの化合物を合成するか）に向けて多大な努力が続けられていることである。迅速合成を達成するためには合成法の開発，分離精製法の工夫等，様々な有機合成上の工夫等、地道な努力によって始めて可能になる。Stillのフラッシュクロマトグラフィーが有機合成に与えたインパクトははかりしれないが，これに類似した地道な努力とアイデアの積み重ねが要求されている。例えば液相合成での迅速合成を行うためには液-液抽出が問題となるが，バリアン社のEXTUBE（水層を吸着する）に代表される固相抽出カラムを用いると一挙に解決できる。96穴反応槽に棒を96本立てて，冷却して水を凍結させた後，棒を抜き取れば反応槽に有機層が残るという手法もある。液相合成でも固相合成のように基質に過剰の反応剤を作用させて反応を完了させた後，過剰の反応剤を官能基を有する樹脂でクエンチするというポリマー・クエンチ法も一般的となった。固相合成では固相に結合させるリンカー部に用いる官能基の存在が問題となったこともあったが、リンカー部が跡形もなくなるような手法も開発された。固相合成で用いられる反応も増えてきている。筆者は固相上でのグリコシデーション反応に多孔質ポリスチレンを用いることで、種々の溶媒が使用可能であることを示した。溶媒の制限がなくなったことから多孔質ポリスチレンを用いることで均一反応であれば何でも固相合成に適用可能であることを示唆している。どの手法を用いるにせよ，副生成物が生じないように合成のプロトコールを確立することが（自動精製システムを用いるにせよ）重要であり，この段階に最も時間を要する。

コンビナトリアルケミストリーに必要とされるのは，高収率，高選択性，しかも短時間で終了する反応であり，このような反応あるいは手法を開発するための基礎的な有機合成研究の役割は非常に大きく，アカデミックの支援が是非とも必要とされる所以である。

生産性を向上させるためには常に新しいテクノロジーを導入していくことが重要だが，欧米の製薬企業はコンビナトリアルケミストリーとHTSを，R&Dの効率化を行うための重要なテクノロジーとして捉えている。コンビナトリアルケミストリーとHTSは創薬研究全体に必要な期間をそれ程短縮するものではないが，リード創製とリード最適化には圧倒的な武器となることはいうまでもないだろう。現在，ヒトゲノムプロジェクトに限らず様々な生物のゲノム解析が怒涛の勢いで進められており，これに伴い今後は構造の明らかとなった転写産物（蛋白質）の数が爆発的に増加していく。これらの蛋白質の中には新しいタイプの薬剤開発の重要な標的が含まれていることは確実であり，増加する研究対象に対処するためにもコンビナトリアルケミストリーとHTSはキーとなるテクノロジーである。

コンビナトリアルケミストリーの最近の動向について簡単に述べさせていただいた。"コンビナトリアルケミストリーで見つかった医薬品"について質問を受けることがあるが，まだこの手法が広まってから５年程しかたっておらず，現状では多数の候補が見出されている段階である。しかしコンビナトリアルケミストリーは創薬の現場に無理なく組み込めるように進化してきており，欧米ではすでに定着したと考えてよい。そこでこのような設問はすでに意味がなくなっているともいえる。多数の誘導体を効率的に合成する経路を開発する，精製法を工夫するなどの地道な取り組みによって，今そこにある化合物の多くにコンビナトリアルケミストリーを適用できるのである。

機能性分子の探索・開発がリード創製とリード最適化によってなされることを考慮するならば，コンビナトリアルケミストリーが創薬のための方法論にとどまらないことはいうまでもない。有機金属錯体における触媒開発，種々のセンサー，有機材料の開発には簡単に適用できそうである（既に開発研究は始まっている。）。一方で無機化学分野への適用はより容易であろう。ここでもすでに触媒開発，超伝導物質開発，耐磁性物質の開発にも適用され始めている。創薬と同様に rational design と randum screening によって展開されてきたこれらの分野で，"rational and randum"を巧みに効率化するコンビナトリアルケミストリーが有効に利用されつつあることはむしろ当然と言える。

ある目的地に到達するために使う手段は多様である．「コンビナトリアルケミストリー」は一つの化学分野というよりは，機能分子に到達するための新たな手段・方法論である。現時点では，この新たな「ハイ・ウェイ」を整備・実用化していくための周辺領域を包括して「コンビナトリアルケミストリー」というキーワードが用いられているようだ。近い将来この方法論が各研究分野で当然のアプローチの一つとして定着した時には，むしろ発展的な意味で"現行の"「コンビナトリアルケミストリー」という "フィーバー"は消え，基盤技術として根づくのではなかろうか。

現在欧米との差は非常に大きいように思われますが，米国でも高々数年先行しているにすぎません。この分野の重要性が十分に認識され，現状を変革するための具体的な対策を講ずれば，追いつき追い越して，さらには新しい視点からその先につながる新技術を見出していくことは可能でしょう。

このような目標を立てて，我々は近畿化学協会の支援を得て，１９９６年４月にコンピューター化学部会の下にコンビナトリアルケミストリー研究会を旗揚げしました。昨年からはより独立した形での運営を行っております。主な活動としては２日間のセミナーを年２回開催しています。ここでは内外の研究者から最新の成果が報告されており，質疑応答も極めて活発で，最近では200名を越える参加者があるなど好評を得ています。コンビナトリアルケミストリーは合成と分析・検定，計算科学，生物試験の３つが組合わさって有効に機能するので，他分野間での協力が不可欠であるだけでなく，研究者も幅広い視野をもってこれに取り組むことが要求されています。したがって、合成化学，コンピューター化学，生物化学など多様な研究領域間での情報・議論の接点となるべく，プログラムを配慮しています。

コンビナトリアルケミストリーは決して単なるブームで終わるテクノロジーではありません。研究活動全般における効率化，迅速化の大きな波とともに押し寄せてきているように思えます。さらには経済活動の大きな波の一部であるとも云えるでしょう。

しかしながらその基盤にあるのは基礎的な化学です。このような観点からは，繰り返しますが是非ともアカデミックの支援が必要です。

次回は９月8日，9日の両日、かながわサイエンスパーク（川崎市）で第７回研究会を開催する予定にしております。興味のもたれる方は是非参加いただけますようお願い申し上げます。

Unification of Bioinformatics, Combinatorial Chemistry and Chemoinformatics with a Structure Based Approach

K. Ramnarayan, D. Rideout, J. Wang H. Zhu and E. Maggio

Structural Bioinformatics Inc.

San Diego, CA 92127. USA.

Introduction:

Traditionally, drug discovery was practiced by screening natural and synthetic products, to find molecules having the desired biological activity. Although this technique has yielded some of the most popular drugs in use today, it is an expensive and laborious process. Advances in Molecular Biology have enabled access to a flood of novel protein drug targets and resulted in a revolution in understanding the human genome. Computational Chemistry techniques have rapidly evolved so that now one can make intelligent guesses as to 3D shapes of the targets against

which drugs are designed. With advances in combinatorial chemistry instrumentation, it is now possible to make thousands of compounds in a reasonable amount of time provided that the chemistry is first adapted to the particular combinatorial chemistry machine or robot. In spite of this ever increasing access to large numbers of chemicals to screen, the costs associated with both synthesis, purification and screening are very high. However, when protein structure information is combined with combinatorial chemistry, screening efforts can be made dramatically more cost effective and the drug discovery cycle can be shortened considerably.

Protein Modeling Will Shape the Future of Combinatorial Chemistry:

The two most important dynamic changes shaping the pharmaceutical industry today are:

(1)the availability of thousands of new protein targets as a result of the human genome project and other gene sequencing efforts around the world.

(2)the availability of molecular diversity on a grand scale through the advent of combinatorial chemistry,.

A newly emerging third dynamic is also at work -- the growing recognition that structural pharmaco-genomics will play a critical role in determining which preclinical candidate molecules are advanced into clinical trials. The element which ties all of these

dynamics together is structure -- specifically the 3D structure of protein targets to which combinatorial chemicals are designed either to bind or to mimic.

Over the last ten years, a quiet revolution in 3D protein modeling has been underway, which culminated earlier this year in a landmark paper in Science Magazine [1]. In it, Professor Kim of the University of California, Berkeley, an expert and thought leader in the protein structure field, asserted that the slow but relentless improvement in the quality of 3D protein modeling has finally led to models which are directly useful in structure based drug discovery. Advanced techniques include: ab initio augmented homology modeling; local tertiary structure prediction augmented with proprietary ab initio loop generation algorithms; and direct ab initio modeling of small proteins [2] (e.g., certain growth factors, hormones, etc.) up to 20 - 30 residues, among others. Other advanced structure prediction techniques on the horizon employ wring-mode and distance matrix approaches [3]. As a result, it is no longer necessary to consider determining the X-ray crystal structure of each and every protein in the human genome in order to access the information content of the corresponding protein structures. In addition, molecular modeling permits practical access to genetic polymorphisms (i.e., genetic structural variants - estimated to average roughly 150 per protein across the human population). Computationally modeled structural variants of drug targets may be used to rank potential drug candidates before selection as clinical candidates by successive docking experiments. This realization unlocks a major portion of the human genome (and non-human genomes) and permits the protein-structure-focusing of combinatorial chemistry.

The ability of new combinatorial chemistries to produce literally millions of potential compounds often outstrips the ability of powerful high-throughput screening technology to keep pace. Additionally, the general commercial availability of combinatorial chemistry synthesizers makes it possible for any pharmaceutical company to start producing their own molecules. But the question remains what to produce?

Virtual combinatorial libraries can describe literally tens of millions, even billions, of structures theoretically accessible by combinatorial chemistry. But by themselves these virtual libraries are useful only to point to synthetic possibilities. It is impractical to synthesize even a small percentage of all of the potential molecular diversity available through combinatorial chemistry. As more and more organic chemists focus on new combinatorial reactions, this problem will only get worse. The principal way combinatorial chemists have attempted to deal with this growing problem is by synthesis of limited libraries (10,000 - 20,000 compounds) that purport to sample pharmacophore space. Since pharmacophore space is n-dimensional, (with hundreds of descriptions to choose from and the relative importance of such descriptors is very difficult to determine), such synthetic sampling schemes are inherently flawed, and in the end the results still depend largely on the scientific intuition of the persons selecting the sample molecules intuition is very difficult to measure.

In any event, this sampling approach becomes increasingly more impractical as the body of virtual chemistries continues to grow daily. (i.e., while testing a 10% sample of 100,000 compounds (i.e., 10,000 compounds) may be possible, sampling and in vitro testing 10% of a 1 million or 100 million compound library is impractical in terms of costs, time and data management.

Chemoinformatics and Chemical Diversity:

Chemoinformatics has been defined as the integration of information resources, automation, and drug design* (Frank Brown, Oxford Molecular Group). Chemoinformatics includes chemical databases suitable for 2-dimensional and 3-dimensional structure searching, QSAR, pharmacophore identification, molecular modeling, and diversity calculations. The purpose of chemoinformatics is to answer the question of which compounds to make and test in a drug-discovery cycle, in order to find active compounds while keeping library size down to a workable number. This field is becoming the rate-determining step in modern high-throughput drug discovery because chemistry and biology are both so high throughput. Structural Bioinformatics has developed a unique approach to chemoinformatics that involves the application of protein surface dynamics-based pharmacophore templates which can be derived even when x-ray and NMR-based structural data is unavailable.

Generally, biologically active molecules fall into active clusters or islands in pharmacophore space. A library that covers pharmacophore space evenly (with molecules arrayed in a grid-like pattern) will more often contain molecules within every active cluster, when compared to an uneven library of the same size (with molecules clumped together and spread unevenly). Strategies for the effective spanning of pharmacophore space with a minimal number of compounds are one of the key goals of chemoinformatics. One approach to optimizing coverage involves filling diversity holes in a given library, that is, finding molecules that occupy regions of pharmacophore space that are unoccupied in the current library. The cell-based algorithms divide pharmacophore space into a grid of multiple cells. The cells are designed so that biological activity tends to be similar for molecules within a given cell and different from one cell to the next. A library with optimal coverage of chemistry space will include molecules in most or all of these cells.

The term sparse arrays library has been used to refer to individual compounds chosen through chemoinformatics that do not form a complete array, yet are all synthesized through the same set of chemical reactions. Sparse arrays have the advantage of being smaller and thus less expensive to synthesize and screen than complete arrays. If chosen on the basis of pharmacophore diversity, sparse arrays have the disadvantage that they provide less SAR data and that they do not allow the discovery of unexpected hits found in complete. If chosen on the basis of virtual screening with pharmacophore templates, however, sparse arrays can be quite valuable in terms of SAR data because the number of hits per compound tested is much higher.

Analysis of scaffolds, as opposed to side chains, shows very poor diversity for synthetic organic molecules in general and compounds which have been chosen as drug leads in particular. Combinatorial chemists are beginning to address this issue with greater creativity by employing novel polycyclic heterocycles, chiral scaffolds, and unusual metal-catalyzed chemistry including olefin metathesis, samarium-promoted coupling, and a wide variety of palladium-catalyzed reactions.

Multivariation refers to a situation in which two or more precursors in a combinatorial library must have a particular structure in order for biological activity to occur, although the rest of the molecules containing these pieces may all be inactive. The existence of multivariation provides an important reason for making array libraries: in the absence of a lead or virtual screen, the only way to find novel multivariate leads is to synthesize and screen large arrays. A versatile, robust virtual screen such as the Dynapharmtm template can rapidly identify multivariate lead compounds without resorting to the synthesis of large libraries even when there are no good leads.

A system for the management of data and the management of experimental flow are crucial part of contemporary drug discovery. Ideally, the collection, storage, reporting, and analysis of chemical and biological data is automated in a manner that allows all scientists in the organization rapid, straightforward access to the data that they need and minimizes redundant activities and bottlenecks that can slow the drug design process. This requires a fully integrated, web-based system that is optimized for data mining. Self-consistent terminology, a user-friendly interface, tables customizable to different scientists, and ability to handle human error is some of the requirements for such a system. This latter factor is important because combinatorial chemistry is accompanied by combinatorial amplifications of errors: in a large array of 30 x 30 x30, for example, a single mislabeled precursor will ruin 900 of the products. Most library syntheses have at least one error. Automation of combinatorial chemistry leads to reduced errors and to some extent to increased efficiency.

In a typical Study, only about 10 nverified leads with IC50 < 10 μM were identified by a diversified library of 30,000 compounds. Michael Grifith of Trega has noted that we need to screeen very large numbers of compounds to find activity. The usefulness of screening large, diverse libraries for hits varies dramatically with the nature of the target. Scientists from Combichem have noted that they could identify 391 molecules from a large, diverse library that inhibited dopamine uptake by 50% at 30μM , but only 19 that inhibited inducible nitric oxide synthetase and only 1 that inhibited tumor necrosis factor activity. The Dynapharmtm approach makes it possible to identify leads with minimal synthesis and screening (on the order of 100compounds), even for very difficult targets.

Application of Protein Structure Based Library Design to Tumor Necrosis Factor Receptor Antagonist Design:

The use of protein structural information to computationally prescreen virtual combinatorial libraries overcomes all of these limitations. In the example described below, 3D structural information derived from the surface of TNF was used to computationally screen a virtual library array consisting of 136,000 compounds with a common scaffold structure. The virtual library itself was generated based on the common pharmacophoric elements of a lead molecule discovered by searching a database of Available Chemicals using a pharmacophoric template * Dynapharmtm. Molecules mimicking a putative pharmacophore region of TNF surface were selected from the virtual library by molecular similarity. A total of 15 molecules were synthesized with the results summarized below. All 15 showed some level of activity; three showed adequate binding activity to qualify as valid candidates for optimization to drug leads. It was not necessary to pre-select and synthesize a sample of molecules representing an imaginary sampling of so-called pharmacophore space, since one million molecules may be pre-screened computationally in less than an hour; ten million in less than a day. All 15 of these molecules inhibited TNF binding to its receptor by more than 35% at 50 micromolar. One of these molecules exhibited a Ki below 10 micromolar. The acquisition and in vitro testing cost (cost of synthesis) for 15 molecules is insignificant compared to that for a 10,000 molecule sampling of so-called pharmacophore space. Finally, additional virtual libraries can be generated and computationally prescreened as quickly as new combinatorial chemistries can be designed.

Conclusion:

In Structural Bioinformatics Inc., we adopted a topological approach for similarity description, which is analogous to the atom pair approach, however, reformulated to take into account the electrostatic features of molecules. We developed a nonhierarchical clustering method allowing fast assembling of molecules. The sizes of clusters are controlled by specifying a threshold for the degrees of similarity within a cluster. Our approach is advantageous to the others in that it incorporates the binding-related features of a molecule into topological descriptors. Two molecules can be compared in terms of their binding-related features while avoiding to be involved in the complicated multi-conformation problems. Our clustering method is fast, with extent of clustering controlled based on intra-cluster similarity. The results of clustering are not history-dependent, unlike some other nonhierarchical methods.

As the repertoire of combinatorial chemistry continues to expand, it is clear that the use of computational technologies (e.g., molecular similarity or computational high throughput automated docking) to focus and reduce the scale of combinatorial synthetic efforts, offers the only practical means to access the full range of molecular diversity available now and in the future from combinatorial chemistry.

REFERENCES:

E. Pennisi, (1998) Science 279, 978.

Dudek, M.J., Ramnarayan, K. and J.W. Ponder (1998) J. Comp. Chem. 19, 548.

J. Bohr, H.G. Bohr and S. Brunak (1996) Europhysics News 27, 50.

P. Willett. (1987) "Similarity and clustering in chemical information

systems." Research Studies Press: Letchworth.

P. Willett, V. Winterman. (1986) Quant. Sruct. Activ. Relat. 5, 18-25.

R.D. Brown, Y.C. Martin. (1996) J. Chem. Inf. Comput. Sci. 36, 572-584.

J.D. Holliday, S.S. Ranade, P. Willett. (1996) QSAR 14, 501-506.

R.E. Carhart, D.H. Smith, R. Venkataraghavan. (1985) J. Chem. Inf. Comput. Sci. 25, 64-73.

C.H. Reynolds, R. Druker, L.B. Pfahler. (1998) J. Chem. Inf. Comput. Sci. 38, 305-312.

L.B. Kier, L.H. Hall. (1976) Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, NY, U.S.A..

A.R. Katritzky, E.V. Gordeeva. (1993) J. Chem. Inf. Comput. Sci. 33, 835.

M. Hassan, J.P. Bielawski, J.C. Hempel, M. Waldman. (1996) Mol. Divers. 2, 64-74.

D. Liu, H. Jiang, K. Chen, R. Ji (1998) J. Chem. Inf. Comput. Sci. 38, 233-242.

R.D. Cramer III, D.E. Patterson, J.D. Bunce. (1988 J. Am. Chem. Soc. 110, 5959-5967.

S.D. Pickett, J.S. Mason, I.M. McLay. (1996) J. Chem. Inf. Comput. Sci. 36, 1214-1223.

K. Davies. (1996) In Molecular Diversity and Combinatorial Chemistry.

Libraries and Drug Discovery. I.M. Chaiken and K.D. Janda Eds.; American Chemical

Society: Washington, D.C., pp 309-316.

P. Willet. (1990) In Concepts in Applied Molecular Similarity; Wiley: New York, pp. 43-65.

R.A. Jarvis, E.A. Patrick. (1973) IEEE Trans. Comput. C-22, 1025-1034.

CICSJ Bulletin Vol.16, No.6, November 1998

目次

図書館経営とビブリオメトリックス

駿河台大学文化情報学部 岸田和明
kishida@surugadai.ac.jp

本稿では、図書館経営における重要な問題である、購入雑誌の選択や文献の廃棄・別置に対して、ビブリオメトリックスの手法を応用する試みを概観する。具体的には、Bradfordの法則やObsolescenceに関する規則性についての数学的モデル、および図書の貸出に関する確率論的モデルを紹介する。

1 はじめに

図書館経営における意思決定を支援するために、実証的なデータに基づく数量的な手法を導入する試みは、1920年代から30年代にかけて展開され、その後1960年代から70年代ごろに、1つの領域として確立したと考えられる。その貢献者は、雑誌数とその掲載関連論文数の規則性を発見したS.C.Bradfordであり、文献利用の経年的な減少を指数関数を用いてモデル化したR.E.BurtonとR.W.Keblerであり、さらには、オペレーションズリサーチの手法を導入したP. M.Morseらであろう。本稿では、このうち、Bradfordにより発見された規則性と、文献の経年的な減少に焦点を当て、そのモデルを概観する。また、図書の貸出に関する確率論的なモデルに関しても簡単に触れる。

2 Bradfordの法則

ある図書館が化学に関する雑誌論文を網羅的に収集することを考えたとする。この際、化学分野の中心的な雑誌（コアジャーナル）を識別し、それを定期購入することは比較的たやすいが、直接的に化学分野に関連しない学術雑誌に掲載された化学文献を発見・収集することは一般に難しい。この問題はほとんどの図書館が直面している切実な問題であると言えよう。
この問題に関して、次のような経験則が幅広く成立することが知られている。まず、各学術雑誌をその主題（ここでは「化学」）に関連する論文の掲載件数の多い順序で並べ、その順位をrとかく（r=1,2,...,N ）。次に、r位までの雑誌に掲載されたその主題分野の論文総数（累積総数）を$X(r)$とする。そして、ある標本を抽出し、 rと$X(r)$をプロットすると、一般に、図1のようなグラフを描くことができる（なお、横軸は対数変換してある）。このグラフは、様々な分野において非常に普遍的に得られるので、その発見者の名をとって、Bradfordの法則（経験則）[1)]と呼ばれている。このグラフに対する数学的モデルとしては、その形状から帰納的に、

$$X(r) = a\log(1+br) \quad (1)$$

のような式が導かれるが（ここで、 aとbはパラメータ）、さらに、これを微分すると、

$$\frac{dX(r)}{dr} = \frac{a}{(1/b)+r} = \frac{a}{r+B} \quad (2)$$

を得る（B=1/b ）。ここで、B=0 ならば、(2)式は有名なZipfの法則に形式的に等しい。Zipfの法則は、単語の出現頻度や都市の人口など、ある少数の特定の個体に度数が集中する一方、度数の少ない数多くの個体が存在するという「集中と分散の現象」を記述したものであり、Bradfordの法則は結局、数多くの関連文献を集中的に掲載する少数のコアジャーナルと、ごく少数の関連

論文を掲載する多数の周辺的雑誌が存在することを示していることになる。この法則は雑誌購入の意思決定に対して理論的な基盤を提供するとともに、その成立原理の探究を通じて、科学コミュニケーションの解明にも貢献している。

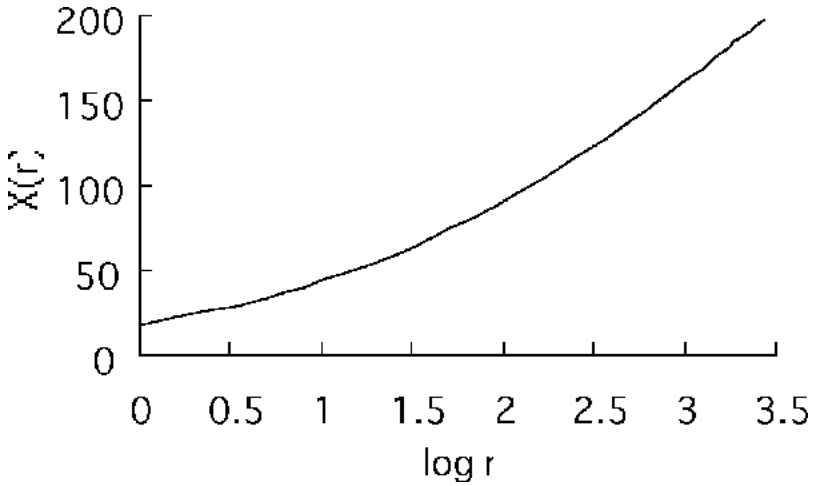


図1雑誌の順位と累積関連論文数とのプロット（例）

3 Obsolescenceに関する規則性

現代的な図書館経営におけるもう1つの現実的かつ切実な問題として書庫の狭隘化がある。様々な統計が示すように、第2次世界大戦以降、出版量は年々増加しており、多くの図書館が書庫スペースの不足に悩んでいる。1つの解決策は、古くなったものや利用されなくなったものを保存書庫へ移したり、廃棄することであるが、この処置は当然、それによって生じる利用者の不便が最小になるようになされるべきであり、そのためには文献利用の将来予測のモデルが必要になる。
一般に、文献利用の経年変化は、放射性物質の崩壊を模倣した、指数関数的な減少モデル、

$$u(t) = Ae^{-ct} \quad (3)$$

で近似できることが経験的に知られている[2)]。ここで、$u(t)$ はその出版からt 年が経過した文献の利用回数（または被引用回数）であり、 Aとc はパラメータである。文献利用の経年的減少は、一般に、obsolescence（老化）と呼ばれ、(3)式はその規則性を表わしている。

もし(3)式が正しければ、その図書館における、その文献の「一生」の間の利用総数U_0 は、
$e^{-c} = \alpha$として、

$$U_0 = A(1 + \alpha + \alpha^2 + \alpha^3 + \ldots) = A/(1 - \alpha) \quad (4)$$

であるが、出版後 年でその文献を廃棄するとすれば、

$$A(\alpha^T + \alpha^{T+1} + \alpha^{T+2} + \ldots) = A\alpha^T(1 + \alpha + \alpha^2 + \ldots) = U_0\alpha^T \quad (5)$$

だけの利用がその「犠牲」となる[3)]。そこで、このような利用者の不便を最小にするような別置・廃棄計画が必要となる。

4 雑誌の購入・廃棄計画の最適化

図書館はより利用されるものを優先しつつ、できるだけ数多くの雑誌を定期購入したいが、それには予算・書庫スペースの限界がある。ここで問題を単純化して、購入予算は無限に存在するが、書庫スペースには限界があると仮定してみる。これをS と表記すると、利用の多い順にr 誌をt 年間保存した場合、それに要する書架スペースは$S=r\times t$ で表わされる。この状況では、雑誌を数多く購入すればするほど、書架スペースの制限からそれらを早い時期に廃棄・別置せねばならず、逆に、廃棄・別置の時期を遅らせようとすれば、購入する雑誌の数を減らさなければならない。

ここでもし、(4)式のobsolescenceの係数 $\alpha = e^{-c}$および1論文あたりの利用総数U_0 が、「すべての雑誌のすべての関連論文で等しい」と仮定できれば、それらの上位r 誌をt 年間保存した場合の利用総数D は、(1)式と(5)式を使って、

$$D = a\log(1+br)(U_0 - U_0 e^{-cs}) = aU_0\log(1+br)(1-e^{-cSr}) = C\log(1+br)(1-e^{-Mr}) \quad (6)$$

となる（簡単のため、$C=aU_0$ 、$M=cS$ と置いた）。この関数は図2のように最大値を持つので、それに対応するr 誌を購入し、それを$t=S/r$ 年間保存すれば、利用の総数を最大にするという観点からの最適な経営方針となる[4)]。

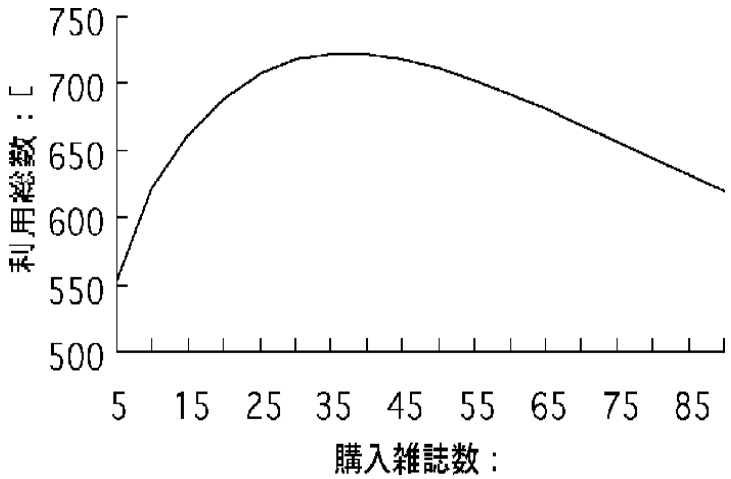


図2 書架スペース一定の条件の下での購入雑誌数と利用総数との関係（例）

(6)式の最右辺の微分は簡単な計算により、

$$\frac{dD}{dr} = C\left[\frac{1}{(1/b)+r} - \frac{e^{-Mr}}{(1/b)+r} - \log(1+br)e^{-Mr}Mr^{-2}\right] \quad (7)$$

となるので、この$dD/dr=0$ を満たすr が最適購入雑誌数となる。この方程式は、実際のデータからパラメータb 、M （すなわちc ）が決まれば、

$$\frac{d^2D}{dr^2} = C\left\{-\frac{1}{\left[(1/b)+r\right]^2}\left(1+e^{-Mr}\right) - 2\frac{e^{-Mr}Mr^{-2}}{(1/b)+r} - \log(1+br)e^{-Mr}Mr^{-3}\left(Mr^{-1}-2\right)\right\} \quad (8)$$

を使って、Newton-Raphson法（すなわち、$r^{i+1} = r^i - D'/D''$ ）などで解くことができる。

5 図書の貸出に関する確率論的モデル

ここまでは雑誌についてのモデルを議論してきたが、一方、図書の貸出に関しては、確率論的なモデルが用いられることが多い。ある1冊の図書の1年間の貸出を想定すると、その貸出という事象が 回生じる確率は、貸出の発生がポアソン過程であると仮定して、平均 λのポアソン分布、$g(x) = e^{-\lambda}\lambda^x/x!$で記述できる。さらに長年にわたっての貸出を考えるならば、(3)式で表わされるobsolescenceの効果を考慮に入れなければならず、貸出の発生を非定常ポアソン過程、

$$g(x,t) = e^{-m(\lambda,t)}m(\lambda,t)^x/x!, \quad m(\lambda,t) = \int_0^t \lambda e^{-cs}ds \quad (9)$$

とする必要がある。さらに、パラメータλ が蔵書全体でガンマ分布

$$f(\lambda) = \frac{\beta^k}{\Gamma(k)}\lambda^{k-1}e^{-\beta\lambda} \quad (10)$$

に従うと仮定すれば、複合確率分布の理論を使って、

$$P(x,t) = \int_0^\infty g(x,t)f(\lambda)d\lambda = \binom{x+k-1}{x}p(t)^k\left[1-p(t)\right]^x, \quad x = 0,1,2,\ldots,$$

ここで、$p(t) = \left[1+\frac{1}{c\beta}(1-e^{-ct})\right]^{-1}$ (11)

のように計算でき[5)]、結局、蔵書全体における図書の貸出回数による分布は、特殊なパラメータを持つ負の2項分布となる（ *p(x,t)*は*t* 年間に*x* 回貸し出される図書の相対度数分布を意味する）。このことは、いくつかの実際の貸出データによって確かめられているが、それらによって、雑誌に関するBradfordの法則と同様に、集中的に貸し出されるごく少数の図書と、ほとんど貸し出されない数多くの図書が存在する「集中と分散の現象」が、図書の場合にも一般的に観察されることが知られている。
なお、(11)式は、図書の貸出の経年変化を記述しているので、このモデルによる未貸出図書の将来予測が可能であり、図書の廃棄・別置の問題に応用できる。

6 おわりに

以上、図書館経営に応用されるビブリオメトリックスのモデルを概観したが、現状では、これらは実際に活用される道具というよりは、理論上の示唆にとどまっている。しかし、これらは実際的な図書館経営の意思決定における理論的基盤として、重要な役割を担っているといえよう。

参考文献

1) S.C.Bradford. 上田修一訳. 〝特定主題についての情報源〟. 情報学基本論文集I. 上田修一編. 勁草書房. 1989. p.159-168.
2) 岸田和明. 蔵書管理のための数量的アプローチ：文献レビュー. Library and Information Science. No.33, p.39-69 (1995)
3) Brookes, B.C. 〝The growth, utility, and obsolescence of scientific periodical literature〟. Journal of Documentation. Vol.26, No.4, p.283-294 (1970)
4) Buckland, Michael K. Book Availability and the Library User. New York, Pergamon, 1975. 196p.
5) Burrell, Quentin L. 〝A note on a ageing in a library circulation model〟. Journal of Documentation. Vol.41, No.2, p.100-115 (1985)

きしだ かずあき KISHIDA, Kazuaki
現在、California大学Berkeley校に訪問研究員として滞在し、情報検索の研究に従事しています。
連絡先 〒357-8555埼玉県飯能市阿須698駿河台大学文化情報学部 0429(74)7124

引用分析によるリサーチ・オン・リサーチ

窪田国際事務所　窪田輝蔵

TerzoK@aol.com

リサーチ・オン・リサーチを引用分析を使って行い科学や研究の構造に迫る試みがある．情報化学の領域で最も引用の多かったホットペーパーは何であったか，研究と研究がどのように結びついているかを表す「科学の地図」や，どのテーマに研究が集中しているかを示す「研究の最前線」などが紹介される．

1 はじめに

雑誌ネイチャーの前編集長ジョン・マドックスは，説得力のある科学論文は「すでになされていることから出発するものでなければならず」また「それ以前に理解されていたことと，その論文が伝えようとすることとの関連が，読者に容易に理解されるものでなければならない」と言っている．そして，日本からの論文は，この点でいささか欠けるところがあり，日本人科学者は，誰が何をやったか，或いはやっているか，世界の先端で何が起こっているかということに，比較的無頓着で独り善がりなところがあるとも言っている [1)]．「すでになされていること」といい，「以前に理解されていること」といい，それは，具体的には，引用文献に現われてくるはずである．

研究者が自分の研究成果を発表するとき，必ず関連する他の論文を引用する．引用は，著者が論文の中で述べる新しい知見を獲得する過程で，拠り所にしたり参考にした「すでになされていること」や，その分野ですでに「以前に理解され」共通の認識になっている科学的知の遺産を明示し，自分のそれらの遺産に対する「知的債務」を確認する作業である．

一方で，引用は債務だけではなく，「知的債権」をも示す．科学者は新知見を得たからこそ論文を発表するのであって，「すでになされていること」や「以前に理解されていること」をただ追認するのではない．先行する研究業績とは異なる別の新しい発見を報告することで、自らの業績の新しさと正しさを世界の科学全体に対して主張するのである．科学史家ロバート・マートンは，引用を「知的所有権を表示するラベル」だと言っている [2)]．「知的債務」といい「知的債権」といい，引用・被引用の関係は，何と何とがどう関わっているかということを明らかにするものであって，実は，科学を構造化するキイになっている．したがって，引用行動を分析することは，科学そのものを科学すること，すなわちサイエンス・オブ・サイエンスにつながるのであり，研究を研究する，リサーチ・オン・リサーチのための一つの重要な方法になっている．

2 「研究」を研究する

リサーチ・オン・リサーチ，或いはサイエンス・オブ・サイエンスの必要性が注目されるようになったのは，科学が，聖なる好奇心に発する個人的な探究心に依存していた嘗ての牧歌的科学から，二つの世界大戦を経て，科学が個人の手から切り離され，国家と社会にとって不可欠の資源になったという歴史的社会的経緯がある．科学とその研究は今や政治問題になったのであり，科学そのもの，研究そのものが，科学的研究の対象になったのである．

研究を研究するためにはいくつかの方法がある．例えば，研究従事者や学位取得者の数，学会や研究機関の分布，公的私的研究投資額などを取り上げて，いわば科学の人口比率や科学研究資源から科学に迫るというやり方がある．これは，いわば，科学に対するインプットを調べて，国や社会の科学力の指標を探ろうとするものである．一方，科学のアウトプットを見る方法もある．新発見、技術開発，特許や論文数，被引用の度数，国内・国際会議の数等を指標にして，科学の生産性を計ろうというものである．

引用分析は，したがって，アウトプットを見るリサーチ・オン・リサーチの一つの分野であるに過ぎない．それは科学という巨大な営みの特定の局面に光を当てた研究であって，科学研究という人類の営みの全貌を一挙に明らかにするという類の，いわば伝家の宝刀のようなものではない．しかし，前述したように，引用分析には，他の方法とは異なった，大きな特徴がある．それは，科学の内部構造に迫ることができるということである．引用分析によって，論文と論文の間の主題関連性を指示することが出来るから，ある論文のまわりにどれだけの論文が集まっているか，その論文群に共通する主題は何であるか，研究者たちの間での，その時点でのホット・トピックスは何であるか等のことが見えてくる．それは，科学というものを個々の研究者の単発的な仕事の集合から，研究者自身も気がつかない，ある構造をもった塊として提示する．引用分析によって，科学には構造があるということが，分かってくる．

「科学は構造を持つ」と言ったのはユージーン・ガーフィールドであった[3)]．彼は，引用索引 (Science Citation Index) を作ることによって，論文と論文を結び付け，研究の最前線を掴み出し，科学の構造を世に示したのであった．科学が構造を持っているということが，個々の研究者に対して意味することは，研究は孤立しては成り立たないということである．意識するとしないにかかわらず，科学的研究は何らかの形で，どこかで，互いに関連しているということである．それは，世界に知られているの第一線の国際的科学者だけでなく，計算化学専用の計算機開発のために，目前のLSI設計に集中して取り組んでいるエンジニアでも同じである．自分の仕事が，その分野の知的技術的遺産全体のどの部分を担い，それにどのような価値を付加するのかを自覚的に問うことを求められるということであり、また，それは可能なことだということである．

ジョン・マドックスは，「科学は，あたかも，まだ作られていないピースを組み合わせて，無限のジクソーパズルを解く過程に似ている」とも言っている．彼の科学観によれば，研究とはジクソーパズルの一片を作り出すことであり，それを，全体の絵を完成するために，正しい場所に嵌め込んでいく作業ということになる．その考え方の当否は別として，世界をリードする科学ジャーナルが，このような科学観によって編集されていることは心に留めておかねばならない．冒頭に引用した彼の言葉と併せて考えると，日本の研究者は，多く，ジクソーパズルの全体図に対する関心が薄く，自分の研究に閉じこもりがちで，孤立気味ということになる．孤立無援或いは唯我独尊で行くという道もあろう．しかし、世界を目指す人であれば，世界で何が起こっているのか，研究者たちは何処に行こうとしているのかについて無関心ではいられないであろう．そのような研究者に引用分析によるリサーチ・オン・リサーチは一つの回答を与えている．

3 ホットペーパー

ユージン・ガーフィールドが創立したISI (Institute for Scientific Information) 社は，世界の主要科学ジャーナルを集めて，引用・被引用関係のデータベースを作っている．そのデータを使って，ISIは最も多く引用された科学者は誰かを示す「最多被引用科学者」 (most cited authors) とか，最も研究者の注目を集めているテーマ「最多被引用研究領域」 (most cited subject areas) などの興味あるレポートを発表している．最多被引用数を時間のパラメーターで処理すると，ある論文の発表後の時系列的被引用パターンが得られる．発表後直ちに関心を呼び引用を集める論文とか，反対にじわじわと注目されて行くものといった，いわば論文のタイプとその引用パターンが分かるのである．

ISIでは，短時日のうちに引用を多く集める論文をホット・ペーパーと呼んで，隔週刊のレポート誌 Science Watchに発表している．また，ユージン・ガーフィールドは，自ら発行人になって， The Scientist 誌を出版し，ここで Science Watch のレポートから注目すべき科学的成果をピックアップし，その更に詳しい解説を試みている．

以下， Science Watch と The Scientist から情報化学 (computational chemistry) に関係すると思われる記事の中から，先ずホット・ペーパーのいくつかを紹介する．

1990年9月号の Science Watch は次世代の新発見を予測する意味で，当時の最新実験技術方法の展望を試みた[4]．化学の分野で，1987年に最も多く引用された方法論文が5件リストアップされているが，そのうちの一つが，時間依存量子力学的手法を分子力学（ＭＭ）に提示したイスラエルの化学者 Ronnie Kosloff が1988年にJournal of Physical Chemistry に発表した論文であった[5]．この論文は発表後約一年で37回引用され，ISIのホットペーパーにランクされた．その後も引き続き引用され続け，1998年の今日までに計686回に及んでいる．引用論文の年代別，国別分布（上位10位まで）を参考までにあげてみる．

年代別分布		国別分布	
1989	37	USA	302 (44.02)
1990	55	Germany	120 (17.49)
1991	65	Israel	83 (12.10)
1992	74	France	66 (9.62)
1993	60	England	40 (5.83)
1994	74	India	36 (5.25)
1995	79	Spain	26 (3.79)
1996	102	Denmark	25 (3.64)
1997	89	Russia	23 (3.35)
1998	51	Japan	21 (3.06)

#Anchor177383

この表を見て，1996年に引用が増えたのは何故か，日本からの引用が世界の3%であるのは妥当か，というような疑問や問題に答える知識も余裕もないが，年代別分布から方法に関する論文の寿命の長さを感じたり，また，国別分布から MD の世界地図を見て取ったりすることが許される

とすれば，これが MD 研究の一分野の歴史と地図を表しているということだけは言い得るであろう．

もう一つ，ホットペーパーを紹介する．分子軌道 (MO) の計算機実験の論文である．1991年4月の Science Watch は，化学分野のホットペーパー・ランキングのトップに J. J. P. Stewart の電算機実験法の論文が踊り出たと報じた[6]．著者である Stewart は，Science watch のインタビューに答えて，「情報化学には現在三つの分野がある．第一は ab initio で，純粋数学を使う．正確で精度が高いが金を食う．次ぎは，半実験的方法とでも呼ぶべきもので，正確さはまあまあ (modest) だがずっと早い．三番目は分子力学法だが，エネルギー最小構造を探索するわけで，早いが限定的である」と言って，自分の論文は，第二の半実験法を，新しいパラメータを使うことによって，速度を損なうことなく精度を30-40%上げることが出来ることを述べたものだと解説した．

この論文は，発表後徐々に引用が増加し，10年経った今日でも盛んに引用され，実に総計2347回に及んでいる．前述の Kosloff の場合と同様，年代別，国別の分布を示す．

年代別		国別	
1989	7	USA	573 (24.41)
1990	39	Japan	468 (19.94)
1991	80	Germany	345 (14.70)
1992	162	Spain	151 (6.43)
1993	258	Russia	139 (5.92)
1994	306	France	119 (5.07)
1995	353	England	106 (4.52)
1996	388	Poland	97 (4.13)
1997	476	Italy	68 (2.90
1998	278	Switzerland	50 (2.13)

年々引用が増加してきた（1989年は約半年間の集計）ということは，この論文が古典的な位置を占め始めたことを示すと同時に，多方面で利用される可能性を持った汎用性の立証になっている．事実，この論文を引用している文献は，日本のものを一覧しただけでも，フラーレンから薬物設計まで多岐にわたっている．一方国別を見ると，日本が二位にあって，実に全体の20% を占めていることが目を引く．日本は，この分野で，世界のトップに肩を並べているということが，引用分析からも見て取れる．

リサーチ・オン・リサーチは，このように，近過去の科学史であり，同時に国際関係図，或いはむしろ，地球規模の「見えない大学」見取り図になり得るのである．すなわち，自分が世界の何処に位置しているのか，歴史的展開のどの時期に生きているのかを知る手立てになるのであり，科学というジクソーパズルのどの一片を作っているのかを自覚的に確認する助けになるのである．

多くの研究者たちは目前の実験や設計，そして計算に集中しているから，えてして「木を見て森を見ない」類の陥穽に陥ることがある．時に世界の科学地図を広げたり，自分のテーマの直近の研究歴史に関心を寄せることは，決して無駄なことではない．

4 研究の地図と最前線

4. 1. 研究の地図

科学は生きていて，常に生成発展し，新しい分野が生まれ出てくる．今日それは，研究分野の境界領域で多く起きている．情報化学もその一つである．CICSJ Bulletinの特集を見てみると，薬物設計，計算化学，専用計算機等々があり，1998年2月には生命情報学が取り上げられた．計算構造生物学も論じられていて，この分野の学際性が良く現われている．

今から4年前，ガーフィールドは，当時ようやく一つの学問分野として形をなしてきた構造生物学を対象にして，学として成立してきた過程、その前駆的な研究を調べて，一つの研究地図を描いて見せた[7)]．彼は先ずコア·ペーパーをISIのデータベースから拾い出した．その方法は，構造生物学をキイワードにして検索を行い26件の論文を拾い出し，更にその26が引用している文献を集めて539論文を選ぶことから始めた．この539の内500回以上引用されている論文を拾ったところ，17件の論文が残り，これを構造生物のコア・ペーパーとした．すなわち，これら17論文は構造生物学の形成に最も深く関わった中心的論文群であると仮定したのである．

彼は更に，この17件のコア・ペーパーが同じ論文に一緒に引用されている頻度数を調べた．すなわち「共引用 co-citation」[8)]の方法である．理論的に，17論文全部が全て共引用されていれば，そのペア数は17 x 16の二分の一になるはずであるが，実際に共引用が行われて出来たペアは62だけであった．どれがどれと実際に何回共引用されたかをマトリックスにしたのが図1である．

論文 (0024)は論文 (0027)と42回共引用された．同様，論文 (0034)と (0030)とは277回の共引用があった等々の関係を示す．

共引用の頻度はそれぞれの論文の扱っている主題の相互関連性を強く示唆するから，頻度の多さは両者の間の距離と反比例する．ガーフィールドはクルスカルの多元的比例法[9)]を使って図2を作った．

図2では論文をアルファベットで表記している．すなわち，0024=A, 0025=B, …….の関係である．紙面に限りがあり，ここで17論文を全部リストアップできないので，地図上の混み合って密度の高い塊から二·三の例を見る．例えば，GとKは，以下に示すように，共にタンパク質構造の金属（錫）結合に関する論文であり，J, N, O, QはメッセンジャーRNAにまつわるDNA制御という遺伝学に属し，EとLはインターロイキン-1という免疫学が主題になっている．しかし，いずれもそのアプローチにおいて，後に構造生物学という手法を導き出しているという点で，その先駆けになったことを示唆している．

図1

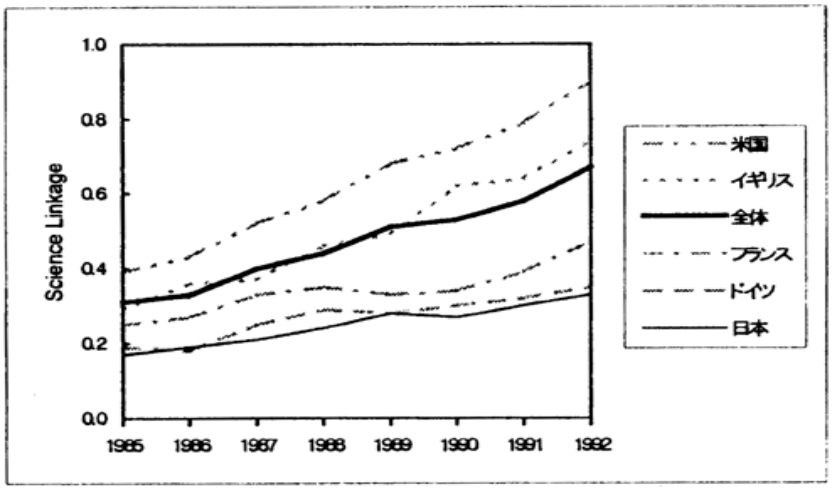
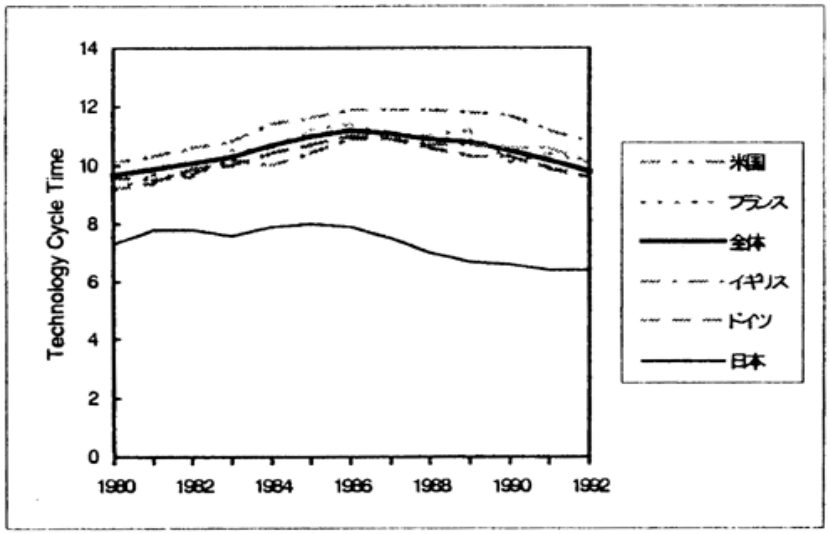


図2



金属結合

G ― Miller, J. et al.: Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. Embo Journal 4 (1985)

K ― Berg, J. M.; Potential metal-binding domains in nucleic acid-binding proteins. Science 232, 485 (1986)

RNA転写

J― Caput, D. et al.; Identification of a common nucleotide sequence in the 3-translated region of messenger-RNA molecules specifying inflammatory mediators. Proceedings of National Academy of Science US, 83, 1670 (1986)

N ― Shaw. G et al.; A conserved A-U sequence from the 3’ untranslated region of GM-CSF messenger-RNA mediates selective messenger-RNA degradation. Cell, 46, 659 (1986)

O ― Lee. W, et al.; Purified transcription factor AP-1 interacts with TPA-inducible enhance elements. Cell, 49, 741 (1987)

Q ― Lenardo, M. J. et al.; NF-k-B: a pleiotropic mediator of inducible and tissue-specific gene control. Cell, 58, 227 (1989)

免疫

E ― Bevilacqua, M. P. et al.; Interleukin-1 induces biosynthesis and cell surface expression of procoagulant activity in human vascular endothelial cells. Journal of Experimental Medicine, 160, 618 (1984)

L - Bevilacqua, M. P. et al.; Recombinant tumor necrosis factor induces procoagulant activity in cultured human vascular endothelium: Characterization and comparison with the actions of interleukin-1. Proceedings of National Academy of Science US, 83, 4533 (1986)

これらは今日の計算構造生物学の前々駆的土台になった論文群と言うことができるのであり，それぞれがジクソーパズルの欠かせない一片になっている研究である．更に興味深いのは，この17件の論文の中に，コンピュータソフトが含まれていることである．それは以下のタイトルのB論文であった．

Brooks. B, et al.; CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry, 4, 187 (1983)

地図上では左端に現われている．関連しているM論文が抗原･抗体組織の三次元構造を2.8オングストロームの解像度で表したものであると言えば，その関連度は理解できるであろう．仮に，同じような調査研究を情報化学について行ったとしたら，間違いなくコンピュータソフトが，その場合はもっと地図の中心部分に，現われてくるはずである．前述した Stewartの論文が，ある塊の中心に位置することは，大いにあり得るのである．

コンピュータソフトが生物学かという議論が嘗てはあった．計算機は道具であって研究そのものではないという議論である．好むと好まざるとにかかわらず，実験法や分析法などの方法論を提示した論文に異常に多く引用が集まる傾向がある．コンピュータソフトも，分野によっては，引用データを歪める因子になりうることは否定できない．しかし，情報化学の場合は計算機実験が重要な研究課題の一つであるから，ソフトウェアに関する論文がコアになることは，歪みでも何でもなく，むしろ妥当であると評価しなければならない．ついでながら，リサーチ・オン・リサーチにおいて，研究のための新しいインプット要素として，コンピュータの利用時間を加えねばならなくなってきている．どれだけ多くコンピュータを使えるかが研究のアウトプットに大きく影響するからである．

ガーフィールドの方法を使って情報化学の地図を作れば，すでに先端を走っている研究者も，これから新しく参入しようとする若い人々も，構造生物学の場合のように，先駆者たちの関心と業績を参照でき，より広い視野を獲得できるであろう．

4. 2. 研究の最前線

ISIは，共引用によってコア・ペーパーを選び出す手法を更に発展させて，共引用されているペアを引用している論文群から，一定のルールに従ってクラスターを作り，そのタイトル語の中か

ら，出現頻度による重み付けを行い，その分野の専門家による表題を付けて，研究の最前線（リサーチ・フロント）として発表した．

「科学の地図」によって，科学の構造を明らかにしたガーフィールドは，「研究の最前線」によって，科学の先端分野を引用分析によって分類し索引付けをしたことになる．それは在来の図書館学的な分類･索引とは違った，科学の内的な構造から作り出された，研究者自身のための道具であった．

研究の地図や最前線が，一見迂遠に見えて，実は研究の推進になくてならないものだ，ということは，それが研究のレビューと展望を文献によって裏打ちしているからである．八尾徹は構造生物学／計算構造生物学の動向をCICSJ Bulletinに概念図で示している[10)]．ガーフィールドの地図は八尾の概念図を文献のクラスターを標識にして地図化したものであり，「分子シミュレーション」とか「立体構造予測」とかの見出語を，更に詳細な論文タイトルからの用語にして示したものなのである．概念図が，ある研究領域の全体像を，一瞥だけで捉えるのに有効であることに，異論を挟む者はいない．科学の地図や最前線は，それに加えて，それぞれに関連するコア・ペーパーを，瞬時に提供できるという強力な機能を併せ持っている．仮に，概念図を料理のメニューだとすれば，地図や最前線は実際に食べられる料理そのものだと言うことが出来る．引用分析によるリサーチ・オン・リサーチの効用である．

「研究の最前線」は，嘗てSTNを通して提供されていたが，現在はOn Demand ベースで商品化されている．

5 おわりに

引用は引用論文と被引用論文の主題関連性を示す．そしてその関連性を指示しているのは，他でもない，引用論文の著者自身である。そこには分類をするとか索引を施すといった第三者の専門家が入り込む余地はない．それらの仕事は論文の著者によって，別の形ではあるが，すでになされてしまっている．であるから，主題の専門家でなくても，A引用論文とZ被引用論文は互いに関わっているという事実を，著者の指示の妥当性を信ずる限り，確認できる．すなわち，主題について素人であっても，その関連性を捉え，頻度を測定し密度を計ることができるという特徴がある．少し飛躍した言い方だが，科学の素人でも科学を計ることが出来，また研究を研究することが出来るのである．

科学研究の「透明性」が言われるようになった．素人にも分かる言葉で科学を語り，研究を開示する必要性が叫ばれている．その責任を科学者にのみ負わせるのではなく，科学の外にいる者が，科学に迫り研究を分析することに挑戦する必要もある．それこそが健全な科学ジャーナリズムであって，引用分析は新しい科学ジャーナリズムに格好の材料を提供するであろう．ユージン・ガーフィールドが The Scientist を発行しているのは，その試みの一つである．

(謝辞: 「2: ホットペーパー」に掲げた Koslofと Stewartの二論文の被引用パターン_年代･国別分布_は図書館情報大学小野寺夏生先生に調べていただき，その結果を提供していただいたものである．感謝して確認させていただく)

参考文献

1) ジョン・マドックス; Nature編集長からのメッセージ．日経サイエンス，10月，44-45 (1989)

2) Merton, R. ; Priorities in Scientific Discovery. American Sociological Review, 22 (6), 635-659 (1957)

3) Garfield, E. ; Mapping the precursors of modern structural biology. Current Contents, December 5 (1994)

4) To Anticipate Tomorrow’s Discoveries, Look to Today’s Hot Techniques, Tools. Science Watch, Vol.1, No. 8, September 1990

5) Kosloff, R. ; Time-Dependent Quantum Mechanical Methods for Molecular Dynamics. Journal of Physical Chemistry, 92, 2087-2100 (1988)

6) Stewart, J. J. P. ; Optimization of Parameters for Semiempirical Methods. I. Method, II. Applications. Journal of Computational Chemistry, 10 (2), 209-220 (1989)

7) Garfield, E. ; Mapping the precursors of modern structural biology. Current Contents, December 5 (1994)

8) Small, H. ; Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents, Jounral of the American Society for Information Science, 24 (4), 28-31 (1973)

9) Kruskal, J. ; Multidimentional Scaling. Sage Publications, p. 5 (1978)

10) 八尾徹：計算構造生物学とゲノム情報解析の融合_欧米の動向_，CICSJ Bulletin, 16 (1), 22 (1998)

くぼた　てるぞう　KUBOTA , Teruzo

1985年まで紀伊国屋書店で洋雑誌の輸入販売に携わっておりました．その時，この小論で触れましたISI社のユージン・ガーフィールド博士を知ることになり，引用分析に興味を持つことになりました．情報化学の門外漢の挑戦です．素人ゆえの誤りが多々あろうかと思います．叱正をお願します．

連絡先　〒245-0051 横浜市戸塚区名瀬町290-2　窪田国際事務所

特許解析で見る技術開発動向

―引用分析を中心として―

科学技術庁科学技術政策研究所　富澤宏之
tomizawa@nistep.go.jp

1. はじめに

本特集号の一記事として表題のテーマが筆者に与えられたが、本稿では、特許解析全般についてではなく、主として特許の引用に関する定量的分析について述べることとしたい。技術開発動向を把握するための特許解析は我が国の多くの企業において日常的に行われているが、引用分析はほとんど活用されていないと思われるためである。また、本特集号の記事としては、ビブリオメトリクスの手法の広がりという観点から特許解析について述べるのが適切であろう。さらに、特許のデータからは、技術に関する情報だけでなく、「科学」に関する情報を得ることもできることを指摘し、特許の引用分析には大きな可能性があることを示したい。

2. 特許情報の意義

特許制度は、発明者の権利を保護する制度であり、またそれを通じた発明の奨励を目的としていることは広く理解されている。しかし、それだけではなく技術情報の普及も特許制度の機能として重要である。特許制度は発明者に新しい技術を公開させる代償として独占権を与えるという論理に基づいているためである。その意味で、本質的に特許は技術情報源として利用されるべく制度化されているのである。

具体的には、特許情報は主として「特許公報」や「公開特許公報」によって公開される。また学術文献の場合と同様に、索引等が付与され二次情報化されてデータベースとして利用される。これらの情報は、技術水準の把握、技術知識の向上、新たな技術開発のヒント、研究開発テーマの選定、などに役立つ。また、他社の技術開発動向や特許戦略、さらには企業戦略の把握にも利用される。さらに、技術情報だけではなく権利情報としての価値も有している。つまり、発明者の所属、権利者（出願人）、専有技術の範囲などが公開されるため、他社の保有する権利を知ることができ、重複研究や無駄な投資を防ぐことや、自社技術の権利を守るための情報として重要である。さらには、研究者や企業名を参照することにより、ライセンシングや共同研究などにも活用される。

特許情報は、技術開発を行う企業のみに重要なのではなく、科学技術論や経済学の分析データとしても用いられている[1],[2],[3]。例えば、技術開発の生産性の指標、あるいは競争力（ないしその要因）の指標のソースとしてよく用いられており、また、技術開発のプロセスを示すデータや技術開発組織に関するデータとして、あるいは技術動向調査や技術予測にも用いられる。

3. 特許における引用とは何か

引用とは「自分の説のよりどころとして他の文章や事例または古人の語を引くこと」（『広辞苑』第四版）であり、それが重要な役割を果たすのは学術論文に限らない。法律の判例の引用がその代表的な例であり、特に、判例法主義に基づく英米法では引用が重要な機能を果たす。しばしば指摘されるように、SCI（Science Citation Index）は、このような英米法のシステムがモデルとなっている[4]。

さて、特許における引用とはどのようなものであろうか。具体的には、特許の申請書（あるい

は明細書）のなかで先行特許や文献の引用が行われる。また、米国では、特許審査の結果に関するレポートが審査官によって作成されており、そのレポートにおいても先行特許や科学技術文献が引用されている。その目的は、出願者による引用の場合、他の先行特許との差異を主張し、あるいは先行特許や科学技術文献との関連性を示し、そのことにより特許権を主張することである。一方、審査官による引用は、審査対象の発明が特許の要件を満たすかどうかを明確にすることが目的である。なお、米国の特許制度は、世界のなかでも例外的に先願主義ではなく先発明主義を採用しているため、特許だけでなく多様な文献の引用が特に重要となる。

4. CHI Research社の特許引用データ

特許の引用を索引化して活用しようというアイディアは古くからあるにもかわらず[4]、また、現在、様々な特許データベースが作成されているにもかかわらず、引用情報を含む特許データベースは少ない。このような数少ないデータベースの作成機関が米国のCHI Research社（注1）（以下、CHIと略記）である。CHIは、米国の政府機関であるNSF（National Science Foundation, 全米科学財団）の委託により、科学技術文献や特許に基づく分析を行っており、またそのためのデータベースを作成している。そのような分析の代表例は、米国の科学技術の状況を包括的に示す"Science and Engineering Indicators"のなかに見ることができる[5]。NSFが1972年以降、２年ごとに発行してきたこの報告書がビブリオメトリクスの発展に果たした役割は特筆すべきものであり、特にマクロ・レベルの指標については常に先がけ的存在であった。なお、CHIの会長であるF. Narin氏は、ビブリオメトリクス（あるいはサイエントメトリクス）の研究者としても有名で、このような報告書の執筆に携わる他、多数の研究論文を発表している[6]。

CHIの特許データベースには、1975年以降の米国特許と1978年以降の欧州特許（欧州特許庁の交付した特許）の書誌情報や引用文献・引用特許の情報が収録されている。原理的には、この引用データを用いればSCIと同様に様々な分析が可能となる。多様な分析の可能性については最後に簡単に触れることとし、次節では主として、CHI自身が開発したいくつかの興味深い指標について述べる。

5. 主要な特許引用データの概要

CHIデータベースから得られる最も基本的な指標は、特許件数および特許の被引用回数である。特許件数を技術開発の生産性や技術力の指標として用いる場合、全ての発明が特許出願されるわけでないこと、個々の発明の質や価値の違いが全く無視されていること、および、産業や技術分野によって特許の価値が大きく異なること、などの問題がある。しかし、適当な条件のもとで、技術開発の生産性や技術力を反映した統計的数値を得ることは可能と考えられる。このような問題は、科学論文件数の場合とほぼ同列に論ずることができる。

特許の被引用回数については、頻繁に引用される特許は技術的な質が高いのか、という問題があり、学術論文の被引用回数と共通する議論が成り立つとともに特許に特有の問題がある。学術論文の場合、引用を行う目的はいくつかあるものの、基本的には、引用相手の論文の持つ知的価値を利用しているのであり、頻繁に引用される論文はそうでない論文より高い知的価値を認められた場合が多いと考えられる。一方、特許の場合、引用を行う目的は、引用する側の特許発明の新規性や先進性などを主張ないし確証することであり、多くの場合、引用された特許の価値については中立的な捉え方をしていると考えられる。しかし、学術論文の場合もそうであるように、ほとんど引用されない特許が全体の大部分を占めており、ごく一部が特に頻繁に引用されていることから、引用された特許は、相対的には重要性が高いと考えることはできよう。

特許の被引用回数が技術の価値を反映しているかどうかについては、実証的な検証も行われている。典型的な検証方法は、特許の被引用回数に基づくランキングが、技術内容の面からの専門

家による判断と整合的かどうかを調べるというものである。いくつかの研究によって、頻繁に引用される特許は専門家によって重要と見なされているものと一致することが示されている[7], [8]。

これら以外の主要指標のひとつは、米国特許の審査報告書における科学論文の引用回数であり、"Science Linkage"と命名されている。審査報告書における科学論文の引用は、技術とそれが依拠する科学とを関係付けるものと考えられ、したがって、その回数は科学との関係性の強さを示すと解釈できる[9]。もちろん、科学と技術の双方向的な関係のうち、科学から技術への一方向しか捉えることができないという限界はある。

もうひとつの指標は、"Technology Cycle Time"と命名され、「引用対象の公表時点から引用時点までの経過年数の中間値」と定義されている。すなわち、審査報告書における引用について、引用対象が「古いか／新しいか」を示す量であり、技術進展の速度に関連している指標であると考えられる。例えば、ある組織が取得した特許の"Technology Cycle Time"が小さい場合、すなわち、より新しい特許・文献を引用している場合、その組織の行う技術開発の速度が大きいと解釈する。このような解釈が適切かどうかを直接検証するには、技術開発のリードタイムと"Technology Cycle Time"の値の関係を調べることが必要であろう。間接的な検証としては、例えば、技術開発をブレークスルー型とインクリメンタル型に区別し、両者の間で"Technology Cycle Time"の値を比較する、などの方法が考えられる。後者の方が先行技術の改良が多く概して技術開発のリードタイムが短いと考えることができるためである。いずれの検証も今後の課題であるが、次節で述べるように、この指標が技術進展の速さを表すという解釈を支持する分析結果がいくつか得られている。

6. 分析例

具体的な例として、科学技術の相互連関性に関わると考えられる二つの量、"Science Linkage"と"Technology Cycle Time"を中心としたいくつかの分析結果を示す。

まず、マクロレベルでの"Science Linkage"の状況を、出願者の国別の集計値によって見てみよう（注2）。やや古いデータであるが1985年から92年における主要国の"Science Linkage"の値の推移を図１に示した。各国の値および全体の値はほぼ一貫して増加している。日本の値は、図に示した国の中でも特に小さい。

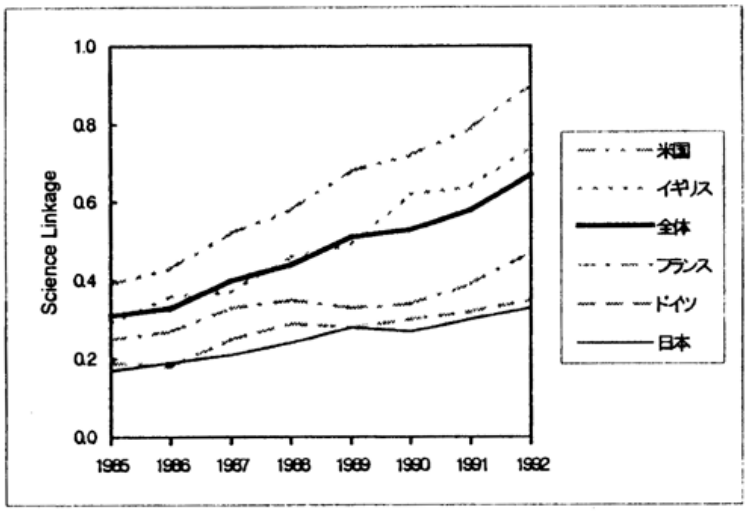


図１　主要国のScience Linkageの推移

このような国別の値の違いは、各国の特許件数の分野別の分布によってある程度、説明できる[10]。例えば、医薬品分野のように明らかに科学との連関が強い分野で多くの特許を取得してい

る国は、"Science Linkage"の値が大きくなる。

"Science Linkage"の値が、技術分野によって大きく異なることを実例によって見てみよう。ここでは、他の要素による影響をできるだけ排除するために、米国特許を一定数以上取得した企業に限定して調べた（注3）。全分野では特許１件あたり1.5件の科学論文を引用している。技術分野別にみるとバイオテクノロジー分野（14.4）が突出して大きく、製薬（7.3）、農業（3.3）、化学（2.7）、医療電子工学（2.2）、半導体エレクトロニクス（1.3）がそれに続いている。一方、値が小さい分野は、自動車・自動車部品（0.1）、自動車以外の輸送機械（0.1）、一般機械（0.1）、などである。

次に、"Technology Cycle Time"の1980年から92年の出願者の国別の値を図２に示した。各国の値および全体の値は1986年頃から減少傾向にある。国別では、日本の値が特に小さい。

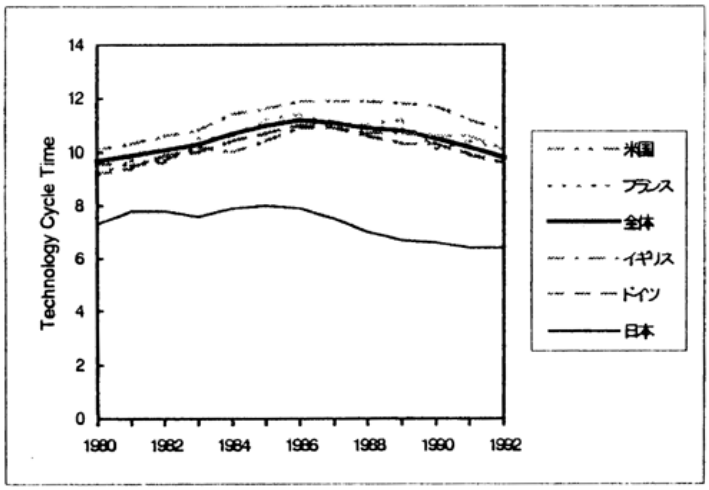


図２　主要国のTechnology Cycle Timeの推移

このような違いは、各国の特許件数の分野の分布による以上に、各国の技術の性格の違いが現れた結果と考えられる。なぜなら、同一技術分野内においてさえ日本の値が例外的に小さい分野がいくつかあるためである[10]。特に、自動車関連技術では、日本の値の小さいことが特徴的である。

もうひとつの分析例として、複数の指標間の関係を見るために "Technology Cycle Time"と特許被引用回数の相関を調べた。ここで用いた特許被引用回数は、"Current Impact Index"と呼ばれるもので、「ある１年間に、企業が前年までの５年間に取得した特許が引用された回数を、同じ期間における米国特許の平均被引用回数で除した値」と定義されている。図３に、日本企業260社の値をプロットした。"Technology Cycle Time"については1990－95年の値、"Current Impact Index"については1995年の値を用いた。"Technology Cycle Time"の値の小さい企業のほうが特許の被引用回数が多い傾向が見られる。これを解釈するには更なる分析が必要だが、技術進展の速い領域の研究開発を行う企業は影響力のある技術を生み出していることを示しているように思われる。

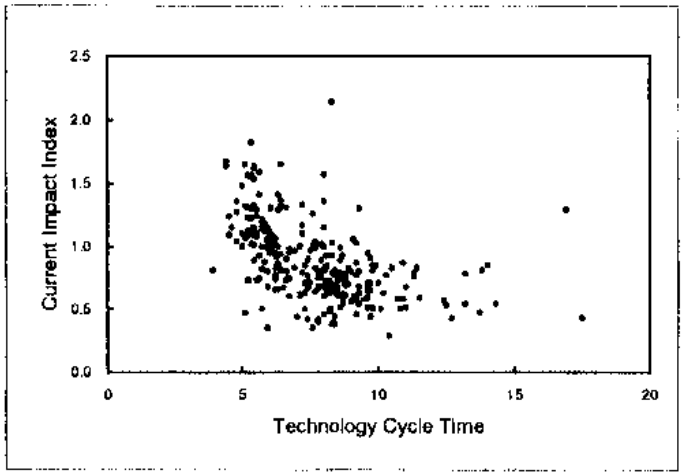


図３　日本企業260社のTechnology Cycle Time
と特許被引用回数の関係（1990－1995）

7. 展望

限られた紙幅のなかでは論じることができなかったが、特許データに適用しうるビブリオメトリクス手法は多い。例えば、co-author analysis（共著分析）の諸手法を応用し、co-inventor analysisとでも呼ぶべき手法を用いれば、技術開発のヒューマン・ダイナミックスの分析が可能となろう。また、共引用分析（co-citation analysis）をはじめとする引用分析の諸手法は、技術と技術あるいは科学と技術の連関を明らかにする可能性を持つ点が魅力である。さらに、共語分析（co-word analysis）のように書誌的情報だけでなく技術内容に踏み込んだ分析も可能である。

特許データからは技術についての情報しか得られないわけではなく、“Science Linkage"の指標がそうであるように、科学に関する情報を得ることもできる。最近、我が国では研究評価への取り組みが本格化しており、定量的手法が注目されている。しかし科学論文の発表件数を評価に直結させるような単純な方法には問題が多い。むしろ、科学研究の成果の評価であっても、“Science Linkage"を用いて科学論文が技術にどのようなインパクトを与えたかを調べる、など多様な視点からの評価を採り入れるべきではないだろうか。その意味でも、特許データのこれまで以上の活用が望まれる。

注

1. http://www.chiresearch.com/
2. 米国特許データに基づく値であるため、米国とその他の国では条件が異なるという問題はある。なお、特許制度は国によって異なるので、特許データの国際比較は原理的に困難である。
3. 米国特許を取得した世界中の企業1100社の1993年から97年までの５年間の値を用いた。ここで用いた技術分類は、特許に付与されたIPC（International Patent Classification）に基づいている。

参考文献

[1] Pavitt, K., "Patent Statistics as Indicators of Innovative Activities: Possibilities and Problems", Scientometrics, Vol. 7, Nos. 1/2, 1985, pp. 77-100
[2] Griliches, Zvi, "Patent Statistics as Economic Indicators: A Survey", The Journal of Economic Literature, Vol. 28, No. 4, December, 1990, pp. 1661-1707
[3] OECD, "The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technology Indicators", Paris: Organisation for Economic Co-operation

and Development, 1994. <http://www.oecd.org/dsti/sti/stat-ana/prod/EAS_MAN.HTM>
[4] Garfield, E., "Citation Indexing: Its Theory and Application in Science, Technology and Humanities", New York: John Wiley and Sons, 1979
[5] National Science Board, "Science & Engineering Indicators -1998". Arlington, VA: National Science Foundation, 1998 (NSB 98-1).
<http://www.nsf.gov/sbe/srs/seind98/start.htm>
[6] Narin, F. and Olivastro, D., "Technology Indicators Based on Patents and Patent Citations", in Handbook of Quantitative Studies in Science and Technology, A.F.J. van Raan, ed., Amsterdam: Elsevier Science Publishers, 1988, pp. 465-507
[7] Worcester Polytechnic Institute, "Analysis of Highly Cited Patents: Are They Important?", Report prepared for the U.S. Patent Office, 16 December 1988.
[8] Albert, M.B., Avery, D., Narin, F. and McAllister, P., "Direct Validation of Citation Counts as Indicators of Industrially Important Patents", Research Policy, Vol. 20, No. 3, June, 1991, pp. 251-259
[9] Collins, Peter and Wyatt, Suzanne, "Citations in Patents to the Basic Research Literature", Research Policy, Vol. 17, No. 2, April, 1988, pp. 65-74
[10] 富澤宏之, 丹羽冨士雄,「計量文献学的方法に基づく科学技術情報の動学的分析」, 研究・技術計画学会第8回年次学術大会講演要旨集, pp.188-194，1993年10月24日, 東京.

とみざわ ひろゆき　　TOMIZAWA, Hiroyuki
専門は、科学技術政策論および科学技術活動の定量的分析。
連絡先　〒100 東京都千代田区永田町1-11-39　科学技術庁科学技術政策研究所　Tel:03-3581-0968

研究活動評価とビブリオメトリックス

学術情報センター　根岸正光
negishi@rd.nacsis.ac.jp

1　はじめに

筆者は1987年に文献抄録データベースを用いて国別の採録論文数調査を行った1)。これは筆者にとってはじめてのビブリオメトリックス(bibliometrics)的調査研究となったが、これは同時にこの種の調査としてはわが国初の大規模なものでもあった。それ以来、各種のデータベースを調査対象として論文数や引用数の統計調査を行ってきている。本稿では、こうした経験に即して、研究活動の評価とビブリオメトリックスの関わりについて検討してみることにしたい。

2　研究評価の枠組みとビブリオメトリックス

筆者も研究分担者として参加した平成5･6年度科学研究費補助金「我が国における研究評価手法の総合的研究」報告書では、研究評価の方法を検討している2)。これを踏まえて、一般的な研究評価の枠組みあるいは構成要素を掲げてみると、およそ次のようなものが考えられる。

評価対象：個人、機関、国、研究プロジェクト等の単位がある。
評価目的：予算配分、研究継続の可否、研究の軌道修正など。
評価時期：事前、中間、事後評価。
評価者：自己評価、他者評価（同一機関内、外部専門家、外部非専門家）。
評価指標・方法：主観評価、客観評価（ビブリオメトリックスなど）。
評価基準：学術性、実用性。

主観評価と客観評価 —— 個々の研究評価とは、これらの組み合わせになるから、研究評価なるものが一概に律することのできない、難しい作業になることが想像できるであろう。学力試験では○×式の客観テストのような、評価者（採点者）の恣意の入る余地のないものを設計できる場合もあろうが、研究評価の場合にこうした客観テストあるいは客観データによる評価だけで済ますことができる例はほとんどあるまい。結局、評価というものは最終的には評価者の主観的価値判断に頼らざるを得ないことになる。とはいえ、こうした主観評価をなるべく公平、不偏のものにするために、種々の客観的データが評価者に供給される必要がある。ビブリオメトリックス的な諸指標は、こうした意味合いにおいて、研究活動評価にとって重要なデータとなる。

3　研究活動の指標としてのビブリオメトリックス

研究活動の出力としての論文 —— 研究活動を評価するために、研究活動を一般の生産過程と同様にモデル化して、それへの入力と出力を測定して、研究効率などの評価に結びつけようとすることが多い。入力としてはヒト・モノ・カネ、すなわち研究者数、研究設備、研究費などであり、出力は「研究成果」であるが、その成果を実際に何で計測するかは大いに議論のあるところである。一般には「論文」が研究成果の代名詞的に扱われる場合が多いが、その当否は研究分野の性格に依存する。研究の結果が新たな発見であれば、その発見された物・現象自体が本来の研究成果であり、発明ならば、その発明品そのものが研究成果であるはずである。論文とは、こうした発見・発明の内容を研究者間に広く伝えるためのもの、すなわち科学コミュニケーションの

手段として普及したものである。なお、発明に関しては、成果を公開しつつ発明者の利益を確保して、社会的に発明を促進するために、特許制度が考案されている。

従って、成果公開の動機を持たないもの（軍事機密研究など）や、論文が成果伝達の手段として適さないもの（コンピュータ・プログラムなど）では、論文が研究成果の代替物とはならないであろう。先にふれた1987年の文献抄録データベースによる国別の論文数調査の際、その結果の記者発表の席で、ある新聞記者から、「すべからく科学研究の最先端は軍事研究にあり、これは論文などで公表されるはずがない。従って、このような論文数の調査では研究水準の測定はできない。」というコメントがあったことを記憶しているが、これも一面の真理には違いない。この点は何も軍事機密に限ったことではなく、民生分野での企業秘密にも該当する。機密保持を優先して、新技術を論文や特許で公表せず、隠密裏に自社新製品に応用して、販売利益を確保するということは実際にありうる。企業にしてみれば研究自体が目的ではないから、こうした行動はごく自然なことであろう。

論文数調査／引用数調査／引用関係分析／用語分析 —— このように、研究成果の測定手段として論文が適当でない分野もあるのであるが、論文が研究成果の代名詞になっているような分野も多いことは確かで、これらの分野に対してはビブリオメトリックス的なデータが大いに活用されうる。すなわち、論文を大量観察の対象として、これに統計的分析を加えて新たな知見を得ようとする方法である。

これには、論文を、国別、年次別、機関別、主題別、掲載雑誌別等々の属性に区分して統計するという「論文数」統計の系統と、論文中の引用に着目して、被引用数を国別、雑誌別等々に統計する「引用数」統計の系統がある。さらに、引用は論文と論文を結合するので、こうした論文間の結合グラフを作成して、これを分析の対象とする「引用関係分析」の系統がある。また、論文に使われる用語に着目して、用語の共通性や共存性などから、論文や用語のクラスター化を試み、何らかの知見を得ようとする「用語分析」の方法もあるが、これはビブリオメトリックスというよりも、情報検索や自然言語処理の研究に属することが多くなろう。

ビブリオメトリックス／サイエントメトリックス／科学社会学／図書館情報学 —— 論文関連の統計ばかりはなく、研究者数、研究投資額など、研究活動の入力側の諸係数を含めて、研究活動を定量的データに基づいて分析しようとする分野はサイエントメトリックス(scientometrics)とよばれ、ビブリオメトリックスはその一分野ということになる。筆者らは、先に電気・電子系大学院を評価対象として、その教員や出身者の学会での活動と企業の評価、それに発表論文数統計を加えて総合的分析を行ったことがあるが、これはサイエントメトリックスといってよかろう[3)]。サイエントメトリックスは科学社会学(sociology of science)における一方法であるから、ビブリオメトリックスも科学社会学の一部という位置付けになる。一方、ビブリオメトリックスにより、図書館運営論やメディア論的分析を展開すれば、これは図書館情報学の一手法ともなる。

4　論文数統計調査による研究活動評価

論文数調査の方法 —— 研究活動の「評価」という観点では、論文数と引用数の分析が一般に用いられる。論文数調査では、大量の雑誌論文を属性別にカウントする必要がある。そのためには、文献抄録型のデータベースに対して、オンライン情報検索システムを用いて、属性を組み合わせた検索を行い、システムから応答される該当論文数を記録して行くという方法を用いる。例えば国別・年次別の収録論文数を得るには、単純には「COUNTRY=Japan & YEAR=1998」といった検索を組織的に繰返すことになる。このような大量のヒットを生成する検索は、検索システムにとっては過酷で、検索時間もかかり、検索料金もかさむことになるから、個人的に気軽にできるような調査ではない。

研究規模指標としての論文数 —— 論文数は研究活動の総体的規模を表すデータと考えられる。ここには、研究活動の規模に比例して論文も多く書かれることになるであろうという仮定がある。もっとも、論文生産の慣習、あるいは生産性は分野によって大いに異なるから、分野間の比較は難しい。人文系にみられるような、研究成果を畢生の大著をもって世に問うといった対極的事例を持ち出すまでもなく、自然科学分野でも、発表される論文数には分野間で相当の開きがある。論文数をその分野の研究者数で割って、分野別の研究者一人当たりの年間論文数といったものが出せると面白いが、分野別研究者数の算定が難しい。

筆者らは、1996年にCA(Chemical Abstracts)ほか、自然科学系の大型の文献抄録データベースを対象として、国別・分野別・年次別の論文数調査を行って結果を公表した4)。これは前述の1987年調査の続編でもある。これによれば、近年わが国の研究者による論文数は、大部分の分野で米国に次いで第2位の地位を獲得するに至っている。これは、科学に対するわが国の絶対規模的な国際貢献の度合いを表すものとして重要であろう。

データベース間格差 —— 次に、この論文数を各国の研究者数や研究投資額などで割引いて比較するとどうなるかは興味ある点である。しかし、研究者数や研究投資額などの国際統計には多くの問題があるので、とりあえず論文数を国別の人口で除して比較してみた。すると、わが国の国際的な順位は多くの分野で5、6位に後退する。しかし化学を対象とするCAデータベースでの統計では、米国を上回る分野も見られるという状況で、データベース間の格差が顕在化する。CAは和文誌も含めて、わが国発行の雑誌をよくカバーして論文を採録しているが、他のデータベースは欧米誌中心の採録であり、このため格差が生じていると考えられる。欧米のデータベース作成機関、ひいては、そうしたデータベースの多数派利用者である欧米の研究者における、欧米誌中心の考え方をそのまま受け入れるかどうかは、国際派と国粋派で議論の分かれるところであろうが、少なくとも工学系の諸分野のように、和文論文を含めて、日本の雑誌への論文発表が盛んな分野では、日本の雑誌を十分にカバーしたデータベースによって、統計が採取される必要があると考えられる。

機関別統計 —— 国別統計からさらに立ち入って、大学別など、機関別統計も大いに興味をひくものである。この場合、データベースに入力されている著者の所属に基づいて分類統計することになるが、わが国の機関の場合、とくに英訳時の記述が不安定のため集計は大変難しい（わかりやすい例では、"Tokyo Univ"と "Univ of Tokyo"のような変動であるが、学部レベルまで含めると記法は非常に多様化する)。結局、該当簡を印刷して、調査員が逐一判定しながら集計するという膨大な作業になる。この辺りについては、著者段階での記述の標準化とあいまい検索技術方面での進展に期待するところ大である。

5　引用数統計調査による研究活動調査

論文の品質指標としての被引用数 —— 引用とは、その研究が前提とし、また批判の対象とする既往の知識を明示して、読者の理解を助けるというのが本来の役割であろう。従って、引用される論文が必ずしもよい論文とはかぎらず、引用統計の問題点も種々指摘されている5)。とはいうものの、被引用数は論文の質に関するほとんど唯一の客観的指標として盛んに利用されている。

引用索引とその効用 —— 引用数調査に関しては、利用できるデータベースはISI(Institute for Scientific Information)社のSCI(Science Citation Index)に限られる。SCIは自然科学分野を対象にしたもので、他に社会科学のSSCI(Social Science Citation Index)、人文系分野のA&HCI(Arts and Humanities Citation Index)がある。それぞれに印刷版、CD-ROM版、オンライン・データベースがあり、さらに最近ではWeb of Scienceと称するオンライン検索ソフトも提供している6,7)。これら引用索引データベースは、各雑誌論文について、著者、標題等の書誌的

データに加えて、論文末尾にある引用文献リストをも入力したデータベースである。これにより、ある論文から出発して、その論文を引用している新しい論文を、さらにその論文を引用している論文をというように、研究の展開過程を時間軸方向に検索できる。ある論文を読み、その引用文献リストで過去の関連論文を知ってその論文をみ、その引用文献リストからさらに以前の関連論文を探るという、時間を遡る検索は普通に可能であるが、逆方向の検索は、引用索引を用いなければ不可能である。引用索引の本来的意義、効用はまさにこの点にある。

しかしながら、わが国でSCIというと、個人別に被引用数を勘定して評価に使うものとの認識の方が一般的のように思われる。これは、引用索引本来の使い方がなされていないことの現れで、研究計画の策定や研究遂行、結果のとりまとめなど、研究の各段階において、引用索引による調査があまり行われていないということであろう。これはわが国の研究態様の特質を反映したものであろうか。そもそもSCIの前書きには、個人評価には使うなという戒めが以前から書かれていた。

インパクト・ファクター —— ISIはインパクト・ファクターで有名なJCR(Journal Citation Report)をCD-ROMで毎年刊行している。JCRは、当該年のSCI収録論文から引用された前年、前々年発表の論文について、その引用回数を雑誌別に集計したもので、これから算出される論文当たりの平均引用数は、インパクト・ファクター(IF)と命名され、雑誌の格付け、ひいてはその掲載論文に対する格付けとして、頻繁に言及される指標となっている。

IFは元来はSCIやCurrent Contents誌に収録するべき雑誌の選定指標として考案されたものである。SCIは自然科学全体について年間5,600誌から90万件の論文を採録しているが（データベース版。印刷版では3,500誌）、CAは化学関係のみで年間8,000種の雑誌から45万件の論文を採録している。つまり、SCIへの収録誌はそれだけ「厳選」されているわけで、この選定のための手段としてIFが考えられたというのは納得できる。こうした点からすれば、IFは図書館における購入雑誌選定の資料として用いるのが本来であろう。とはいえ、IFは雑誌の評価、そして論文の評価指標に用いられ、個人の業績審査の際、各論文にその掲載誌のIFを掛けて集計するといった方法もとられることがあるようである。

IFについては、専門誌よりもnatureなどの総合誌の方が高い値が出るとか、レビュー論文の比率が高い雑誌の方が有利になる、過去2年間の論文への引用だけを集計するので、短期決戦型のテーマが有利になる等々、種々の欠陥、限界が既に指摘されていて、補正の算式の提案などもある8)。そもそも、分野間では論文引用の慣習、つまり論文スタイルに相当の差があるので、分野をこえたIFの比較はあまり意味がない。また、著者所属機関の記述と同様に、原論文における引用文献（雑誌名）の記述法のばらつきから、雑誌別の集計は不安定にならざるを得ない。ISIでは、SCI収録誌については、雑誌名記法のユレをデータベース化して名寄せの精度を上げているようであるが、非収録誌については充分な名寄せは困難であろう。

国別引用統計 —— ISIでは、NSI(National Science Indicators)と称する国別・年次別・分野別引用統計ファイルを不定期に刊行している。これによれば、各国の論文に対する平均引用数を国別に得ることができ、筆者もこれを用いて分析を行ったことがある2)。なお、国別と雑誌別を組み合わせた統計はJCRやNSIでは得られない。こうした統計のため、筆者らは学術情報センター所在のSCIデータベースそのものを使って調査を行った9)。ISIはSCIデータベースについて、検索システムを用いたこうした組織的な引用数調査を禁じているので、この調査はISIと別途契約を交わして実施したものである。これによって、雑誌別の国別論文数すなわち雑誌の国際性などが明らかになった。

国別の引用数統計は、およそその国の当該分野における国際的な研究水準を表すもの、研究水準の評価指標と考えられる。しかしながら、こうした引用数統計の国際比較が有意である分野は

限られる。これは、ガーフィールドとともにIFの創始者として著名なISIのスモール博士も指摘していたことだが、数学、天文学など研究者社会が世界的に一体化している少数の分野を除いては、国別の引用数統計をただちに各国の研究水準と解釈するには無理がある。確かに、科学研究は各国の置かれた地理的、経済的、社会的等々の状況を反映して展開されている。つまり、研究テーマに対する関心は国情を反映して国ごとに異なるのが自然であり、従って引用数の国別比較には注意が必要である。

6　日本版SCIデータベースの構築

既述のとおり、SCIは収録誌を「厳選」することから欧米誌中心の指向が強く現れている。一方、文献抄録型データベースに関しても述べたように、わが国では工学系など和文論文を含む研究発表が活発な分野が多いが、これらの国内誌の多くはSCIの収録対象になっていない。わが国の研究活動を振興するためには、こうした国内誌の論文を活用しやすくすることも重要である。こうした考え方から、学術情報センターでは平成7年度から「引用索引データベース」の編集を進めており、近々公開する予定である。これはまさにSCIデータベースの日本版である。当面は、予算の関係などから、理工系分野の200誌程を「厳選」して収録を進めているが、将来的には学術会議登録学会のすべての学会誌4000誌ほどに拡充してゆきたいと考えている。近頃、わが国では科学技術立国があらためて唱導されるような状況であるが、そうであればなおさら、わが国日本独自の科学振興が必要である。こうした状況において、日本版SCIデータベースの必要性は一層高まっていると思われる。

研究評価データベース ―― ところで、引用索引データベースから得られる引用数統計による研究評価は、データベースの副産物的な利用法である。一方、研究評価を本来の目的にしたデータベースの構築も行われている。英国のWellcome財団のROD(Research Outputs Database)がそれである10)。Wellcome財団は医学・生命科学研究に対して大規模な助成を行っているが、こうした助成研究がしかるべき成果をあげているか、助成研究の審査・選定が適切に行われているかという観点から、助成研究に関するデータベースを構築して、それらの評価を試みている。わが国において研究評価をより組織的に展開するためには、わが国でもこうした研究評価用のデータベースの構築が検討される必要があろう。

7　数字の一人歩き現象：おわりにかえて

最近、研究評価に対する議論が盛んになっている。文部省学術審議会は平成9年に「学術研究における評価の在り方について」と題する文部大臣あての建議を行った。また文部省大学審議会では先頃「21世紀の大学像と今後の改革方策について ― 競争的環境の中で個性が輝く大学」と題する答申をまとめ、この中で大学評価のための独立機関の設置を提案している。こうした中で、とくにインパクト・ファクターを中心に、ビブリオメトリックス的指標を安直に引用する言説も増えているように感じられる。この種の用例は、単に自説の補強材料として、たまたま目にした都合のいい数字を持ち出しただけという体裁のもので、この点でいわば数字の一人歩きが起こっている。既述の通り、これら指標の欠陥、限界はすでに指摘されており、これをわきまえた慎重な扱いが必要である。

もっとも、数字の一人歩き現象、我田引水的な引用はあらゆる統計データに共通の性質ともいえる。しかし、例えば経済統計では、各々の専門家が多数いて、啓蒙的解説本も普及しているから、一人歩きにもおのずと歯止めが掛かる。この点、ビブリオメトリックスに関して、わが国では専門的な研究者層が極めて手薄で、素人談義が横行する危うさがある。この際、科学技術立国

のため、科学振興のため、さらにそのための研究活動評価の研究深化のために、わが国におけるデータベースの形成とビブリオメトリックス研究の振興を訴えて、本稿の結びとしたい。

参考文献

1)根岸正光．学術研究論文数の国際比較調査 ― 結果と考察．学術月報，41(7)，40-47(1988).
2)根岸正光．引用度数分析による学術研究の国際比較．学術研究と評価 ― 我が国における研究評価手法の総合的研究（平成5・6年度科学研究費補助金総合Ａ，研究代表者：中井浩二），1995，II･49-II･72.
3)市川惇信；植之原道行；根岸正光．大学の研究及び研究者養成機能を評価する ― 電気電子系分野における評価指標の間の相関．電子情報通信学会誌，78(6)，552-559(1995).
4)根岸正光．学術論文数の国際比較調査 ― 結果の概要と分析視点．情報管理，39(4)，245-257(1996).
5)山崎茂明．生命科学論文投稿ガイド．東京，中外医学社，1996，153p.
6)窪田輝蔵．科学を計る ― ガーフィールドとインパクト・ファクター．東京，インターメディカル，1996，220p.
7) http://www.isinet.com
8)山崎茂明．インパクトファクターをめぐる議論：正しい理解と研究への生かし方．情報管理，41(3)，173-182(1998).
9)Negishi, M.; Adachi, J. Overseas acceptance of Japanese scientific papers as seen in a citation analysis. Preprints of the 3rd International Conference on Japanese Information in Science, Technology and Commerce. Nancy, INIST-CNRS, 1991, 243-259. (ISBN 2-904975-72-1)
10)山崎茂明；根岸正光．研究評価のためのデータベース ― Wellcome財団のResearch Outputs Databaseを中心として．情報管理，41(6)，436-444(1998).