数理解析研究所講究録 1288 巻 2002 年 20-31

バイオインフォーマティクスにおける非線型問題

(株)日立製作所 ライフサイエンス推進事業部　井原 茂男(Sigeo Ihara)
Life Science Group,
Hitachi, Ltd.

I. はじめに

バイオインフォーマティクスを俯瞰し、ゲノムのシークエンスが完了した後を想定したいわゆるポストゲノムシークエンス時代へのその方向性、ビジネスの可能性についても触れ、急速に発展しつつあるこの分野における応用数学の必要性について述べてみたい。

II. バイオインフォーマティクスの現状

2.1 バイオインフォーマティクスの必要性

ヒトを含む様々な種のゲノムデータの急速な増加と、それにともなって加速された遺伝子解析情報、遺伝子発現解析情報、蛋白質構造情報、臨床データなど関連する生物情報の急速な増加は目をみはるものがある[1]。毎年、2.5倍づつデータの量は増加しているともいわれている。図1を参照されたい。

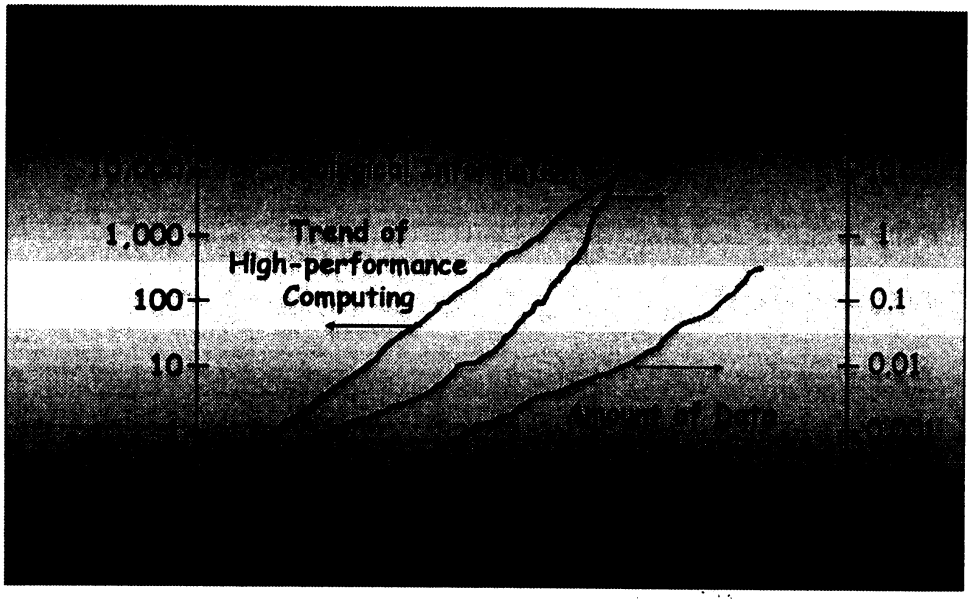

図 1　バイオインフォーマティクスのデータ、計算量の伸び

特に様々なデータベースの中でも、ヒトをはじめとした種々の生物に対する DNA 分子における4つの塩基A(アデニン)G(グアニン)T(チミン)C(シトシン)の並び、すなわち DNA の配列データが大量にデータベースに蓄積されつつある。DNA の配列解析技術の進歩、ベンチャービジネスの台頭による国際プロジェクトの活性化などにより、データの蓄積の急峻さには目をみはるものがある[2]。通信ネットワーク急速な整備と普及およびその高速化を背景に、IT(Information Technology)環境が整えば、研究者は、空気や水と同じように好きなだけ大量のデータにアクセス可能になった。ところがデータがあまりに大量であることから、世界中に散在する大量のデータを処理し、データの良し悪しを判定し、データとデータの相互の関係を調べ、単なるAGTCの文字列情報から DNA の生物学的な機能を推定しさらなる生物学的な意味を抽出するには、人間の能力の限界をはるかに超えた情報処理能力が要求され、計算機に依存せざるをえない状況になった[3]。バイオインフォーマティクスの必要性はまさにここある。

一方、データの伸びに較べて計算量は圧倒的に多くなることが予想される。単純なホモロジー解析から複雑なホモロジー解析に移り変わり、さらに計算量が要求されるタンパク質の構造解析やフォールディング解析、さらにはバーチャルリガンドスクリーニングの精度を上げていくと計算量のオーダーが1－2次から3次、4次と増加する。一方、計算機の性能の伸びは時間に関してほぼ線形で増加することがしられているため、必要な計算を行うためにはギガフロップスからテラフロップス級の計算機投資、さらには高速のネットワーク環境、解析したデータを再利用するためのデータベース(DB)環境が必要になる。応用数学への期待としては、これらの問題のおける計算アルゴリズムによる劇的な効率化が挙げられる。それではどのような問題がどういうところで解決されるのを期待され、またどのような効果を期待されているのだろうか？以下全体像をとらえつつ対象ごとに分けて調べてみたい。

2.2 バイオインフォーマティクスの適用対象

バイオインフォーマティクスはもともとの狭い意味では、DNA あるいはゲノム情報を扱うための実験室のデータ管理システム、DNA の文字配列からその構造の特徴を解析するツール、およびそのデータベース環境のことであった。しかし最近ではバイオインフォーマティクスを広義に解釈し、分子生物学、情報処理、統計学、数学を基礎とし、DNA の配列解析はもとより、文献データ処理などのデータの解析ソフトウェア、さらには広い意味では蛋白質の構造解析から化合物探索まで、生物、医学、および薬学関連の情報解析にかかわる情報処理全てを指し示すことが多い。従ってその目的も膨大で多種多用な生物情報を効率よく整理し、解析を通じてデータの生物学的あるいは医学的意味を明らかにすることに移りつつある。

従って、ゲノム配列のデータを始めとして多種多用な生命情報をどのように整理し統合するかが大きな課題である。ゲノム配列のデータは、無料のデータベースとして

NCBI (National Center for Biotechnology Information) 、EBI(The European Bioinformatics Institute)、およびDDBJ(The DNA Data Bank of Japan)で人類共通の財産として管理・運営されているが、それ以外の多種多用な生命情報は、公共あるいは民間企業を問わず多世界中の各地に散在し、そのデータ形式もまちまちである。そのようなデータを整理・統合できる、あるいは少なくともデータベース利用者にとっては統一化されて使い良いように見える情報の利用環境の構築が急がれている。最近ではXML(The Extensible Markup Language)をインターフェースに用いてデータの共有化を図ろうという試みが行われつつある。

2.3 バイオインフォーマティクスの特徴と利点

昔の生物のやり方では、細胞から何とかRNAを採取し、cDNAを採取してクローニングし、DNAをシークエンシングしていくという流れである。シークエンスのデータが全部わかっている、すなわちゲノム情報すなわちDNA情報がわかっているため、ゲノムの情報から情報処理技術によってどのような細胞にはどのような遺伝子が発現し、機能をしているのかという情報を抽出していくのかが課題になる。このとき、情報処理技術から得られた情報をもとに、実験も行う必要があろう。現在では、ゲノムの大量のデータから、科学とか医学に有用な情報を何らかの形で出すことが重要になっている。そのとき、まずは、特定の遺伝子をホモロジー解析により当たりをつけて、遺伝子の特徴的を抽出し、その機能をタンパク質の発現あるいはタンパク質の構造から類推する。そのためのタンパク質の構造予測においては、2次、3次、4次の構造を予測していくという手順を踏むことが多い。現実にはタンパク質のフォールディング、立体構造の決定は困難であるとされており、精度についての国際コンテストが開催されている。
バイオインフォーマティクスの利点として以下が考えられる:

・網羅的な解析結果
・他の比較との比較が容易
・かなり情報は公開されている
　ただし一般には未公開の私企業の最新のデータも存在する
・統合化されたアノテーション付加
　単なる作業員のコメントから種々のDBのリンクした総合情報
・サービスビジネス形態が様々
　ソフト、ハード、オンサイト&インターネット

網羅的な解析結果が得られることについてまず考えてみたい。バイオインフォマティックスは技術的には未成熟であるにしても、実験に比べて、情報を一遍に処理でき、何らかの結果が網羅的に入手できるところに特徴がある。すなわち、ヒトの遺伝子の数は数万から以上ともいわれており、その一塩基多型(Single Nucleotide Polymorphism)

の変化を考慮するとかなり大きい数にはなるであろうが、遺伝子の数としては有限である。かなり大きい数であるが、有限の対象を扱うことがバイオインフォーマティクスのキーコンセプトである。実際、有限の数の情報であるので、有限数の情報を全て網羅することは可能で、それぞれに対して網羅的に解析してしまえば、ある仮定のもとでの全部の答えは出てしまう。どこまで自分の欲しい情報がはいっているかは別にして網羅的情報のなかに埋もれているにせよ答えははいっているであろう。まずそう考えてデータを解析する。どしても欲しい答えが見つからないときには、計算のパラメータを調節し再計算を行うか、全く別のアプローチを考える。こういう状況が従来の解析手法と大きく異なる点である。

他のデータとの比較が容易であるのは、データが非常に整理されており、データベース間でデータ同士のリンクを張ってあることが多いためである。データベースの情報処理技術の進歩によりデータの相互参照がより容易になり、インターネット上で多くの情報が無料で公開されているため、ある生物のゲノム情報とほかの生物のゲノム情報との比較(望まれているのは、ヒトとラットとのDNAの比較)、あるいはヒトのミューテーションのデータベース、臨床データベースなどほかのいろいろな情報との比較参照がインターネット経由で容易にできるようになった。一般には非公開で、特別のユーザにのみ公開されているベンチャー企業の最新のデータも存在している。そのため、データおよびソフトウェアの無料試用の範囲と、有償試用の範囲をユーザは意識してデータおよびツールを使う必要がある。

当然ながら電子化された情報であるため、データのハンドリングの利便性があげられる。ちょっとしたコメント、ちょっとしたアノテーションをデータにつけることができる。例えば、データの取得の一次作業員が、この領域のデータちょっと信頼できなかったとか、本当はシークエンスのときの精度が充分でていないのではないかといったコメントの付加である。一方、DNA配列データと3次元のタンパク質の立体構造の構造データをリンクさせ、複雑な計算を要するシミュレーションによってタンパク質のエネルギーの最小原子配置から次の安定構造までの変形する様子のアニメーションの付加することも考えられる。またどの生物種の特有な機能に関係しているかといった生物学的な意味付けの情報も必要とされている。このように、いろいろの形式、様々なレベルでアノテーションをつけることができる。

ビジネス面から考えると、バイオインフォーマティクスは発展途上であることもあり、いろいろな形態のビジネスが考えられる点に特徴がある。サービスビジネスの仕方を次に見てみたい。

2.4 バイオインフォーマティクスのビジネス

バイオインフォーマティクスのビジネスとしては以下の形態が考えられる:

1. データベースコンテンツビジネス

2. 解析ソフトウェア販売—単体販売
3. 統合パッケージ(DB、ツール、環境)
4. サービスビジネス
   a. オンサイトでの環境構築サービス
   b. インターネットによる情報提供サービス(ASP)。

公開されている公共のデータを取得するのとは異なった実験方法によるデータを蓄積し、データベース化し、そのデータベースを販売するデータベースのコンテンツサービスがある。例として、セレラ社のデータベースがある。国際チームとは異なったシークエンスの方法で、データを取得し、データベースを構築する。

次に、情報処理のためのソフトウェアを販売する。単体のソフトウェアを売る場合と統合パッケージ、ある程度まとまったデータベースとツールなどの環境とを売るというサービスがある。製薬メーカーの実験室、あるいは研究所全体の環境構築なども有りうる。

遺伝子あるいは疾患のデータを権利関係上、あるいは倫理上どうしても外に出せないということがあるため、ある程度オンサイトのサービスは必要になっている。一方、あるデータベースおよび計算機環境全体あるいはその一部をインターネットで提供するASP。いろいろなサービスがある。ユーザーニーズの多様性を考えると、これらオンサイトサービスとオフサイトサービスとを混在化したようなするような方向が重要になってくるであろう。

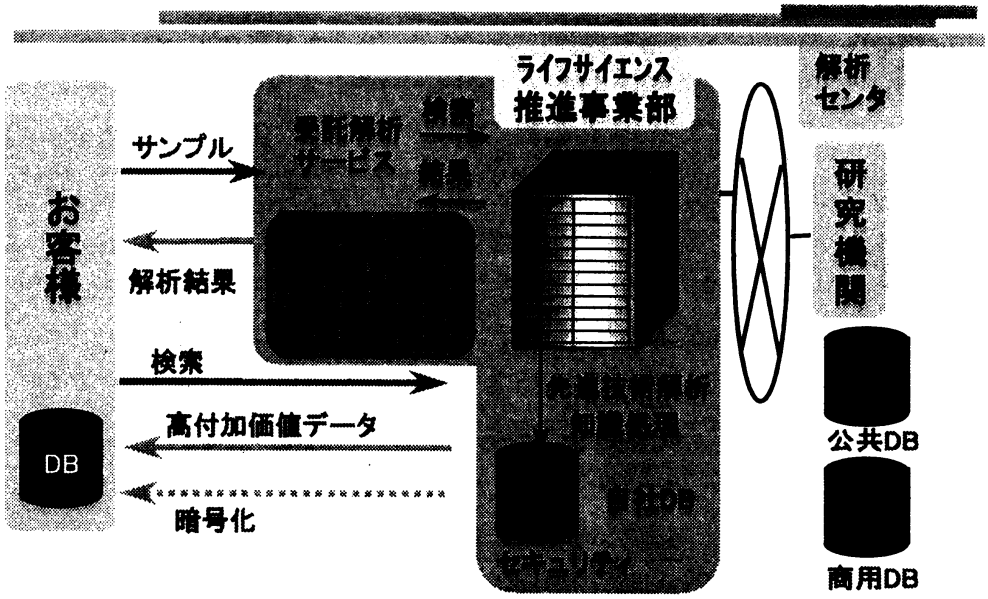

図2　バイオインフォーマティクスの日立のサービスビジネスモデル

バイオインフォーマティクスのサービスビジネスのモデルとして図2を参照されたい。この場合、ユーザからサンプルをいただい、解析結果を返すことになる。情報例えば配列だけの場合では、高い付加価値データを添付することでサービスを行う。結果を配信するのときの暗号化がキーになる。一方で実際の実験の結果、例えば配列のシークエス解析の結果、SNPのディスカバリーおよびタイピングの結果、チップによる発現解析、タンパク質の機能解析、タンパク質間相互作用の結果など総合的な解析結果をどこまで迅速かつ正確にユーザに返答できるかが競争力になろう。実際の解析のときには、一般に公共データベース、商用のデータベース、および独自の二、三のデータベースを使った解析サービスを行う。

以上、バイオインフォーマティクス全般について述べてきた。ほかの企業に較べてより高速性、高性能、高信頼性のあるソフトウェアの開発は必須であり、そのための新しいアルゴリズム開発は極めて重要である。統合パッケージでは、利便性、簡便性が課題のとなる。サービスビジネスでは、新しいビジネスモデルを考えれば、ビジネス特許になっていく。応用数学への期待としては、これらの問題解決のための方法論の改良、処理計算における計算アルゴリズムによる劇的な効率化が期待されている。それではどのような具体的な課題があるのだろうか？以下概要を述べてみたい。

III バイオインフォーマティクスの要素技術

バイオインフォマティックスを要素技術の観点から俯瞰してみる。

・実際の解析では、ゲノムすなわち全DNAのなかで、遺伝子領域の予測であろう。また遺伝子をコントロールしている転写制御領域を予測することも重要である。

・いろいろな生物のゲノムが解析されることによる種を比較して遺伝子の生物学的意味を検討する比較ゲノム解析が重要性を増すであろう。モデル生物、実験動物の遺伝子の機能解明は今後大いに進むであろう。

・また同じ生物種であってもその個体差を反映した、DNAの一塩基多型（SNP）による集団遺伝学的な解析などゲノム解析としては重要になりつつある。

・DNAチップによる遺伝子のRNAレベルでの発現から、遺伝子のネットワーク、遺伝子の探索を行うことも重要である。

・ゲノムからのタンパク質構造の予測、リガンドとタンパク質構造との関係はいうまでもなく重要である。タンパク質間の相互作用解析、さらに物質と生命機能と関連をみる総合的なパスウェイ解析など重要になってくる。様々なタンパク質の量の時間変化をみる細胞の機能シミュレーションも重要になりつつある。

今後は、それぞれ別個に発展してきた情報解析と実験解析とをいかに融合させるかという技術である。実験と数理とが乖離していては問題が解決できない。

一方、バイオ関連技術とナノテクノロジーとの融合も重要である。単に生体だけの機能に関してではなく、機械あるいは生体と機械を融合させたような細胞レベルでのサイ

ボーグ、米国の人気SFテレビ番組であるスタートレックボイジャーのボーグにあたるようなナノメータスケールでの機械生命の可能性までもが見え出してきている。

3.2 DNA配列解析

ある未知のDNA配列を手に入れたとき、そのDNA配列がすでにデータベースに入っているかどうか、データベース中の既存のどのDNA配列とどのくらい類似しているのかについてまず調べる必要がある。いま手にしているDNAのどの部分が、どこの染色体のどのDNA配列と類似しているのかを調べることにより、DNA配列の機能を調べようとするわけである。最近では高速でかつ高感度のBLAST(Basic Local Alignment Search Tool)ソフトウェアがまずは使われる。詳細な情報が得られるFASTAがある。さらに詳細に配列を調べたい場合には、ダイナミックプログラミングのアルゴリズムを用いたSmith-Waterman法が用いられることが多い。ただし、詳細な比較をするほど計算が膨大になるので注意が必要である。さらに、多くの配列を比較し共通部分を抜き出したり、配列の変化を調べたりできるクラスター・アラインメント解析がある。このほか最近では、データベースのDNA配列の構造に構造の生物学的な意味付けを行う目的でDNAの配列の中から遺伝子を構成するエクソン部分を予測するツール、転写制御領域の予測ツールなどを活用することが多い。いずれの場合でも解析の効率化のキーポイントは、計算アルゴリズムの高速化、および計算精度であり、かなりいろいろの解析ツールがある現在においても、計算アルゴリズムの今後の改良が期待されている。

3.3 遺伝子発現解析:DNAチップ・マイクロアレイ解析／パスウェイ解析

微細加工技術により数千から数万の遺伝子を一枚のチップ上に配置し、それぞれの遺伝子の発現状況をモニターできるDNAチップ／マイクロアレイの台頭で、多数の遺伝子が関与する遺伝子の発現過程、すなわち遺伝子がいつ、どこで、どのように、発現するかを調べることができる遺伝子発現プロファイル解析が重要になりつつある。この発現プロファイル解析では、機能が近い遺伝子は細胞内で同じような制御を受け、同じような時間変化をすると仮定し、遺伝子、あるいは機能をパターンに分類する。この仮定からどの遺伝子とどの遺伝子に相関があるかを調べていくわけである。これらの解析のためには、統計学上実験データの測定誤差、結果のばらつきをできるだけ小さくおさえ、できるだけ意味のあるデータを抽出するためのデータマイニングの技術が今後益々重要になるであろう。

このように、mRNAレベルの実験でいかに有用な遺伝子のネットワーク情報を導くかがひとつのキーである。蛋白質の場合では、酵母ツーハイブリッド法[4]などによって、ある蛋白質とある蛋白質の二つのの相互作用も実験的に見出すことができる。これらの情報に対して計算機を利用し、整理し分かりやすいネットワークの形に描画することは、その生物学的な理解をする上で重要である。蛋白質のネットワークの例として図3

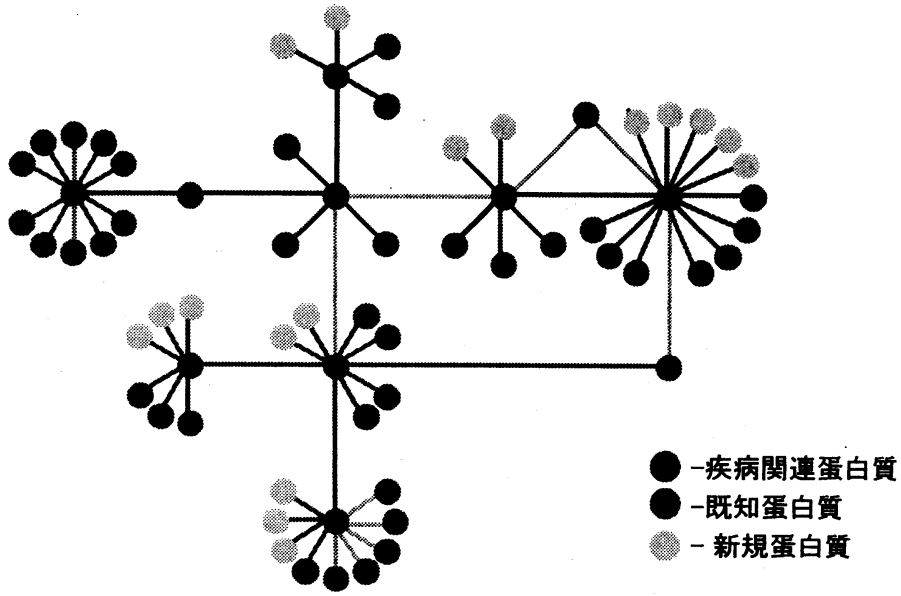

図3 タンパク質—タンパク質相互作用ネットワーク

を参照されたい。環境物質と生体との係りや生体内部のシグナル伝達[5,6]を調べるには、蛋白質だけでなく低分子と蛋白質との相互作用も調べていかねばならない。

これらの問題を解決するためには、既存の文献から必要な情報を整理した形でユーザに提供するようなテキストマイニングが今後は重要性を増すであろう。より使いよいシステムにするためには、あいまいな検索をしたときに、より高い精度で、より広範囲なデータを提供するための自然言語処理手法の開発が必須であろう。

生成される蛋白質の量を時間の変化としてとらえ、連立の常微分方程式を解くシミュレーションによって細胞の変化を予測することは 1970 年代に定性的解析として流行していたが、蛋白質の性質が明確になってきた現在では定量的観点から再び注目されるようになってきた。さらには細胞と細胞との関係、組織全体のシミュレーションを行おうとする試みも新局面を迎えつつある。最近の結果の例を図4に示す。応用数理の観点からみると、計算誤差を抑え、しかも系がスティッフな場合でも安心して使える数値計算のアルゴリズム開発という古くて新しい課題がある。

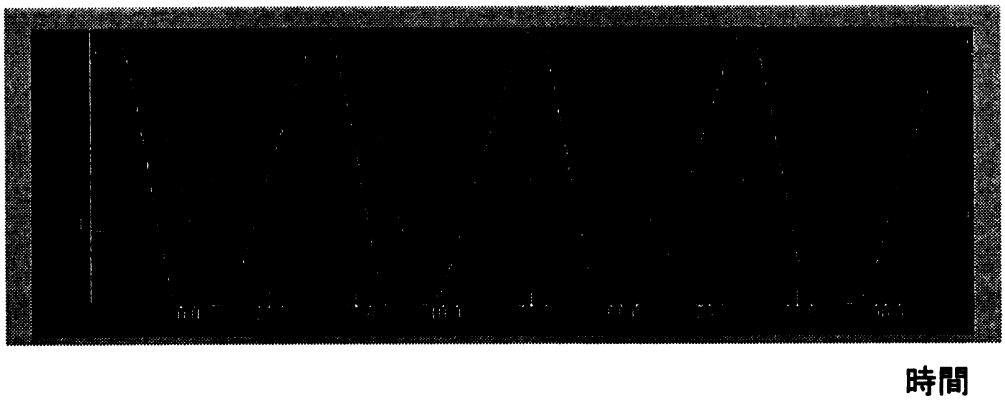

図４　細胞シミュレーション

3.4　蛋白質構造解析・リガンド解析そのほか

一次元情報である DNA 配列から蛋白質の機能を予測することは、現在では限られた場合に対してのみ成功しており、実用に供するレベルには達していないとも言われているが、極めて重要な課題である。ある DNA 配列の立体構造が分かっていれば、その配列との類似度から新規の DNA 配列に対する蛋白質の構造が予測できる。この類似度が 30%以上であれば高い精度で構造が予測できる。20－30%ではかなり困難になる。この領域はでモチーフなどのプロファイリングが重要になる。10%以下ではデータベースを用いた経験的、演繹的な手法だけでは構造の予測は困難で、構造予測を物理化学的に行わざるを得ず、構造予測を行うことは理論的、計算機資源上からも極めて困難な状況になる。とはいえ、今後は分子・原子レベルの配置構造や動力学のシュミュレーション、例えば分子力場法や分子動力学が重要になってくることは間違いないであろう。その後では、電子の安定状態や化学反応を扱える分子軌道法などが重要になる。これらの応用に対して、計算手法の改良、物理モデルの改良など応用数学的問題を解くことが、今後の大きな発展のために必要になっている。従来は経験的手法を使っていたが、これからだんだんと物理化学的手法もだんだんと導入されてくるであろう。物理化学的手法では、計算時間はテラフロップス級の計算機で 100-1000 時間ぐらいかかる。今後は単なる原子状態を計算するだけでなく、分子軌道法によって化学反応を取り扱えるようになるであろう。空間領域は原子の数で100—10 万。計算時間領域がで、ピコセカンドからナノセカンドぐらいを扱えるようになる。生物の素過程

の現象が起こっているマイクロセカンドに向けてこれらの手法を使いある部分は可能になるであろう。

図５　蛋白質と低分子化合物とのドッキング解析

バーチャルリガンドスクリーニングへの応用を考えると、現在は、リセプターを剛体の物体と考え、リガンド全体をゴムの塊のモデルで剛体のどの部分に塊が入るかの計算をして、リガンドのスクリーニングの計算を行っている。最近では、図5に結果の例として示すように、リセプターを原子の集団と考え、リガンド全体も原子の集団と考えたモデルで計算を行い。リガンドが結合部位に入るときにリセプターの原子を緩和させる計算が行われるようになってきた。今後は、電荷の移動による誘電率の変化をとりあつかって、結合部位の計算を行なうという、化学反応を電子状態から取り扱うような計算が行われだしている。これにより、スクリーニングの精度が向上するであろう。これらの計算は計算機インテンシブであり、理論的な計算アルゴリズムの発展による計算の高速化がキーとなるであろう。計算のオーダーを飛躍的に減少させることができる新しい計算アルゴリズム開発がチャレン仁グな課題として横たわっている。

3.5 遺伝統計学的解析

分子レベルのミクロな解析とは異なり、実際の遺伝形質や疾患が生物集団でどう遺伝していくかを調べるマクロスコピックな現象論的手法である集団遺伝学的手法も益々重要になっている。集団遺伝学的手法では、疾患に関係のある遺伝子を探索するとき、臨床データを参照しつつ、疾患と遺伝子との相関関係を求めることになる。そ

のとき遺伝子あるいは DNA 全体にある区切りをもたらす特徴的な配列をマーカーとして導入する。従来の遺伝子マーカー以外に、最近では、マーカー間の距離が短く、ゲノムにおいて密度の高い SNP(Single Nucleotide Polymorphism;一塩基多型)もマーカーとして使われ、さらに従来からのコンセプトの見直しも進み、効率的に遺伝子を見出せる下地が整いつある。解析の対象も、従来から解析が行われて来たメンデル遺伝性疾患遺伝子に起因する疾患から、多くの感受性遺伝子を持ち、環境、習慣、文化といった非遺伝的要因も関与しているような疾患へと移りつつある。そのための解析手法も、大規模な家系解析から大規模な関連解析へと移行しつつある。ある程度絞り込んだマーカーに対してさらに連鎖解析も行われる。また罹患同胞対解析など、感受性遺伝子と多型マーカーとの連鎖を解析する方法も発展しつつある。これらの解析を行う場合、臨床データを扱うことが不可欠である。取り扱えるデータが多くないこともあるので、少ないデータから効率よく精度の高い結果を得るための統計学的なマイニング手法の導入も必須である。また、遺伝子の世代にわたる移動をランダムな事象と考えるのではなく、他の遺伝子の移動と相関を持たせることが可能なような新しい計算の枠組みの提案が待たれている。

3.6 バイオインフォマティックスの今後

ハードウェアの価格が安くなると同時に、今度ソフトウェアもただで配るというオープンソース化の動きが盛んになるであろう。応用としては、データと現象との関連を発見するマイニング手法がデータの蓄積とともに大きく飛躍するに違いない。また、データの統合化も起こって、3次元データのアノテーション、統合データベースなどデータが高付加価値化していくであろう。

従来はどちらかというとウエット実験とバイオインフォーマティクス解析を行う者同士がお互い参考にする程度であったが、両方の側の研究者の意識改革も必要ではあるだろうが、両者を融合させた解析を行うための計算機利用環境の整備、より進んだ解析ツールの研究が重要である。

臨床データを扱う上で個人の遺伝子情報の機密を保つことは、人間の尊厳や人権を尊重するための必要条件になりつつある。そのための匿名化システムの構築、さらにはデータの保全といったことだけでなく、広い意味での運用、教育システムを構築しなければならない

今後バイオインフォーマティクスには自分の現在の理解を超えたところで新しい革新技術が次々と起こってくるに違いない。全く新しい技術を開発する革新性、いろいろな新しい波を捉えることができる感性、さらにはすばやく応用数理の問題として定式化し解決するという機敏さがこの分野で今最も必要とされている。いろいろな分野の方々、特に読者の方々の参加を期待しつつ筆をおきたい。

[1] 榊 佳之 ヒトゲノム ー解読から応用・人間理解へー、岩波新書 728
(岩波書店、2001、東京)
[2] J.D. Watson, *A Passion for DNA, Genes, Genomes and Society*, (Cold Spring Harbor Laboratory Press, 2000, Cold Spring Harbor, New York)
[3] T.K. Attwood and D.J. Parry-Smith, *Introduction to Bioinformatics*, (Addison Welsley Longman, Edinburgh Gate, 1999).
[4] 坂本 健 生体分子間相互作用データの収集と情報解析、実験医学 Vol. 19 No.11 (増刊) 2001.
[5] KEGG
[6] SPAD