インターネット利用による
生命情報統合解析システム

génie

- GENome Information analysis systEm -

Advanced Technology Institute, Inc.

Summary

The Advanced Technology Institute, Inc. (abbr. ATI) provides an excellent scientific environment of computer software for genetic and protein engineering studies through the Internet. In this article we present an overview of a computer system for DNA-protein sequence analysis *génie* that is developed by ATI in collaboration with RIKEN (The Institute of Physical and Chemical Research) . Since the system is constructed on the UNIX workstation and designed with an efficient man-machine interface in mind, users can easily operate it by using a pointing device (mouse) and gain insight into the relationship between structure, function and properties of biomolecules. This article surveys the system design and the major applicatoin programs included in the system such as the prediction of the tertiary structure of protein.

生命情報統合解析システム*génie*について

◎はじめに－ポストゲノムを目指して－

２００３年には約３０億対と言われるヒトの全遺伝情報の内容が解明されると言われる中、１９９９年１２月にはヒトの２２番染色体の全遺伝子配列が決定されました。しかしながら、ロゼッタストーンの如くその意味を"解読"できなければ単に情報の収集のみにしか過ぎません。この膨大な情報処理のためには超並列マシンのような非常に計算パフォーマンスの高いコンピュータが要求されるのみならず、遺伝情報を解読するための解析ソフトウエアの開発が本質的なキーを握っていると言っても過言ではありません。然るに我が国では解析アルゴリズムの研究開発のようなソフトウエアの重要性に対する認識が甘く、米国に比べて大幅に立ち遅れているのが現状であります。
このような状況において、情報通信のインフラストラクチュアが充実しつつある今、インターネット利用による我が国独自のコンピュータ解析を開発することは極めて時期を得たものであり、且つ社会貢献にも寄与するものと言わなければなりません。
このような認識の下、アドバンスド　テクノロジー　インスティテュート（略称　ＡＴＩ，Ｉｎｃ．）では理化学研究所の全面的な協力を得て独自の高度な生命情報統合解析システム、例えば遺伝子配列から蛋白質コーディング領域やプロモータ領域などのような機能部位を予測あるいは蛋白質高次構造を予測するシステム *génie* の構築を行っています。
世界は遺伝子の同定・塩基配列の決定から、所謂ポストゲノム解読に向けて、生体高分子の立体構造と機能の解明を目指す方向に流れています。そして*génie*　はその世界的な研究潮流であるゲノムインフォルマティックスから構造生物学へ向けた２１世紀生命情報統合解析システムの先駆けとも言えましょう［*Fig. 1*］。

◎バイオテクノロジーにおける情報利用

バイオテクノロジーの研究に携わる研究者にとってさまざまな情報が必要となりますが、とりわけ、遺伝子工学や蛋白質工学においては遺伝子配列や蛋白質のアミノ酸配列が第一義的な情報であることは異論のないところであります。*Fig. 1*に見られるように、このような配列情報を手にした時、誰もが最初に考えるのはこの配列がこれまでに分かっているものの中で類似のものがあるかどうかでしょう。もし類似のものが見つかった場合この配列のホモロジー解析によって、手にした遺伝子や蛋白質の機能をある程度類推することができます。このホモロジー検索は、現在行われている配列情報利用技術の中でもっとも利用されているものの一つであります。ついで、遺伝子配列の場合蛋白質コーディング領域やスプライス部位あるいはプロモータ部位の推定へと進み、一方遺伝子配列の情報あるいは実験的に直接蛋白質の配列が得られた場合は二次構造などの高次構造を予測し、最終的には蛋白質分子あるいはこれらが作る分子集合の機能を予測することです。通常は高次構造の予測からこの最終段階が解析アルゴリズムの技術において最も難しいと言われています。即ち、

一次情報　→　生物的な意味を見い出す　→　生物機能に結び付ける

と言う作業が必要で、例えば言語解析においてアルファベットの並びからその意味を求

め価値ある情報を引き出すのと同じことです。このように機能がある程度解明されて初めて得てその知識を応用することにより、異なる機能を持った蛋白質分子などを人工的に設計して作ることができるものと考えられます。バイオテクノロジーの真価はこの部分で発揮されると思われますが、生命情報の利用技術としては最も難しい部分であり、将来のポストゲノムに向けた進展の中で最も重要なキーを握っていると言っても良いでしょう。

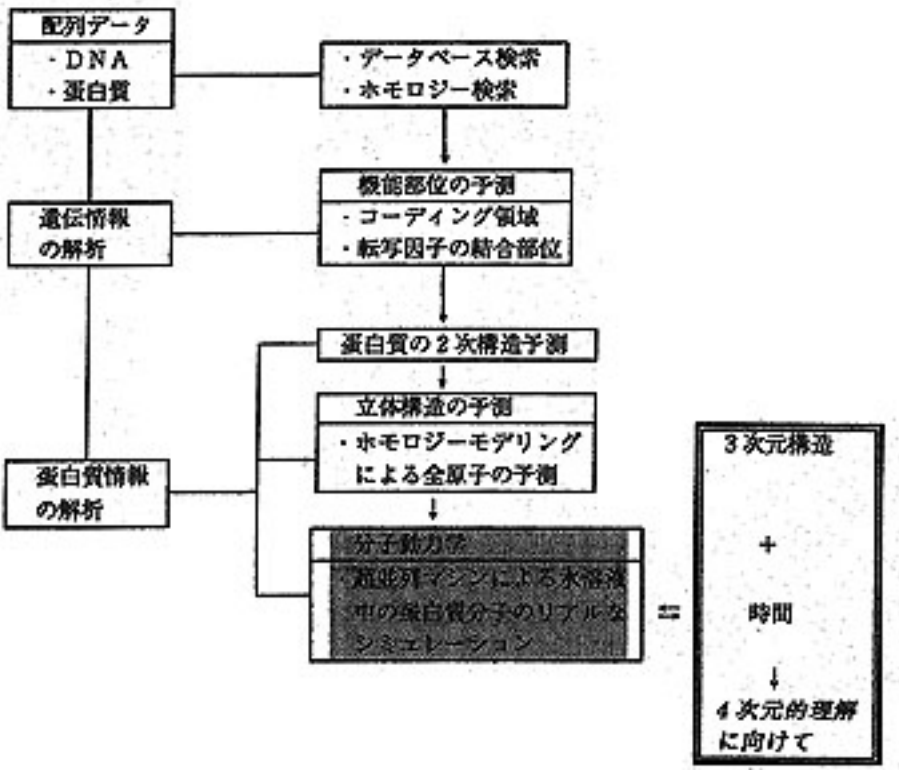


Fig.1 ヒトゲノム計画の流れ－ポストゲノムへ向けて－。
時空構造（４次元構造）：如何なるメカニズムで１次元（配列情報）に埋め込まれるのか？

○*génie* のデザイン

システムをデザインする際のコンセプトとして、World Wide Web (WWW) インターフェイスによりユーザーフレンドリーなものを意識しつつ、コンピュータに不慣れな初心者のみならずプロの使用にも耐える高度なシステムを目指しています。システムデザインは*Fig.2*に示してあります。

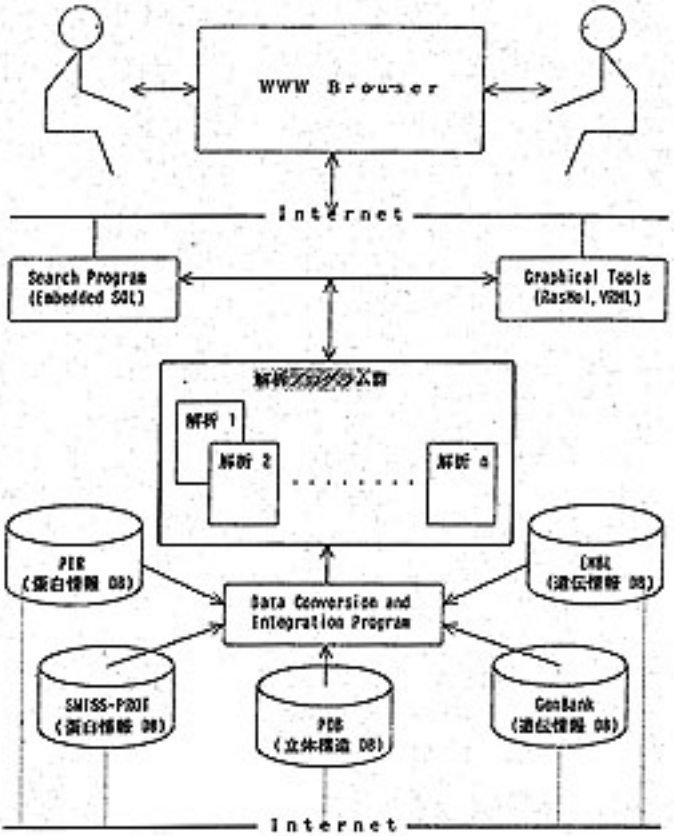


Fig.2 システムデザイン。最新のデータベースに豊富な解析プログラム群がインターネットを介して利用できる。

データベースとしては；
・ＰＩＲ（蛋白質配列Ｄ．Ｂ．）
・Ｓｗｉｓｓ－Ｐｒｏｔ（蛋白質配列Ｄ．Ｂ．）
・ＰＤＢ（蛋白質立体構造Ｄ．Ｂ．）
・ＥＭＢＬ（遺伝子配列Ｄ．Ｂ．）
・ＧｅｎＢａｎｋ（遺伝子配列Ｄ．Ｂ．）
が用意されており、インターネットによりいずれも最新のリリースされたデータを利用することができます。これらデータベースの個々の内容は遺伝子や蛋白質名あるいはＥＮＴＲＹ　ＣＯＤＥなどにより高速検索した後その内容を表示することができ、更に豊富な解析プログラム群が搭載されています。解析結果はさまざまな形で可能な限りグラフィックス表示され、とりわけ蛋白質の立体構造はＲａｓＭｏｌやＶＲＭＬなどにより、回転、移動、拡大が可能となっています。

以下に*génie* の特徴をまとめますと；

●インターネットによる利用が可能

・手元にデータベースを持つ必要がなく、更新などの煩わしい管理などから解放される

●優れたユーザーインターフェイス

●独自の解析アルゴリズム

・Ｎ１４法による露出領域の予測

・Ｊｏｉｎｔ法による二次構造予測

・超高感度ホモロジー検索

・世界初のDistance-Constraint法を採用した蛋白質立体構造予測

のようになります。

特筆すべきことは、最近蛋白質立体構造は２件の特許；

１．発明の名称：蛋白質の立体構造の予測演算方法及び予測演算装置
（特許番号第2856306号）

２．発明の名称：蛋白質の立体構造の予測精度演算方法及び予測精度演算装置
（特許番号第2930851号）

が成立し、その道の専門家により注目を集めていることです。
ではここで、この立体構造予測の数学的手法の概要について簡単に述べて見ますと以下のようになります；

＜Wako-Scheraga-Kubota法＞

●全C_α原子間の相対距離情報 d_{ij} → 全C_α原子のデカルト座標（X，Y，Z）

１．初期構造として目的蛋白質の全C_α原子のデカルト座標

$\{(X_i, Y_i, Z_i) ; i=1, \cdots, N$（残基数）$\}$

を乱数により求める

↓

２．全C_α原子間の相対距離

$$d_{ij} = \{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2\}^{1/2}$$

を求める

↓

3．同様に参照蛋白質側のデカルト座標から相対距離$\langle d_{ij} \rangle$を求める

↓

4．$F = \Sigma\Sigma w_{ij} (d_{ij} - \langle d_{ij} \rangle)^2 / m$をminimizeするようなデカルト座標$(X_i, Y_i, Z_i)_{min}$が予測構造である

［ここで、w_{ij}は統計的重みでmは全残基ペアーの数N(N-1)/2である］

その結果；

●計算手法がとてもシンプルにして且つ迅速な解析が可能

・参照蛋白質とホモロジーのない領域の座標も自動計算するので煩雑な手続きが不要

●予測精度の向上

・参照蛋白質とのホモロジーが３０％以下のものでも～３Å台の精度が得られる

●予測精度の推定が可能

・X線結晶解析による構造が未知のままでもアミノ酸残基ごとの予測精度の推定が可能

と言った他には類を見ない特徴を備えています。

なお、参考のために*Fig.3* に参照蛋白質として採用した立体構造が既知の大腸菌の酸化型チオレドキシンとのホモロジーを示す併置配列を示し、これに基づいて本予測手法によって予測されたチオレドキシンの立体構造のステレオ図と予測制度推定プロフィールを*Fig.4* に掲載します。

```
         10        20        30        40        50        60        70        80        90        100       110       120       130
KQIE-SKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKY-SNVIFLEVDVDDCQDVASECEVKCTPTFQFF
:::: :::: :::: :::: :::: :::: :::: :::: :::: :::: :::: :::: :::: :::: :::: :::: ::::
KIIHLTDDSFDTDLVKADGA-ILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYIERGIPTLLLF
KKGQKVGEFSGA-NKEKLEATIN
:::: :::: :::: :::: ::::
KNGEVAATKVGALSKGQLKEFLD
```

Fig.3 ヒトの酸化型チオレドキシン（上）と大腸菌の酸化型チオレドキシン（下）との併置配列。２つのアミノ酸配列において同一のアミノ酸残基（：印で示した）の割合、つまりホモロジーの程度は２９．１％である。ハイフン" ー" はアミノ酸残基の欠失を示す。

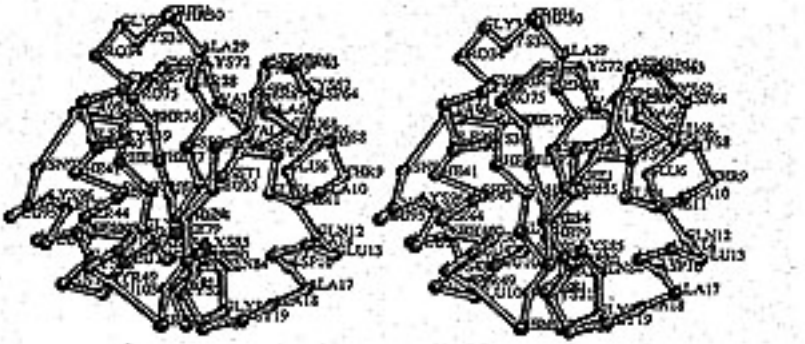


Fig.4 上図：予測されたヒトの酸化型チオレドキシンの立体構造のステレオ図。
下図：予測精度推定プロフィール。縦軸の値が高い程予測精度が良いことが推定される。

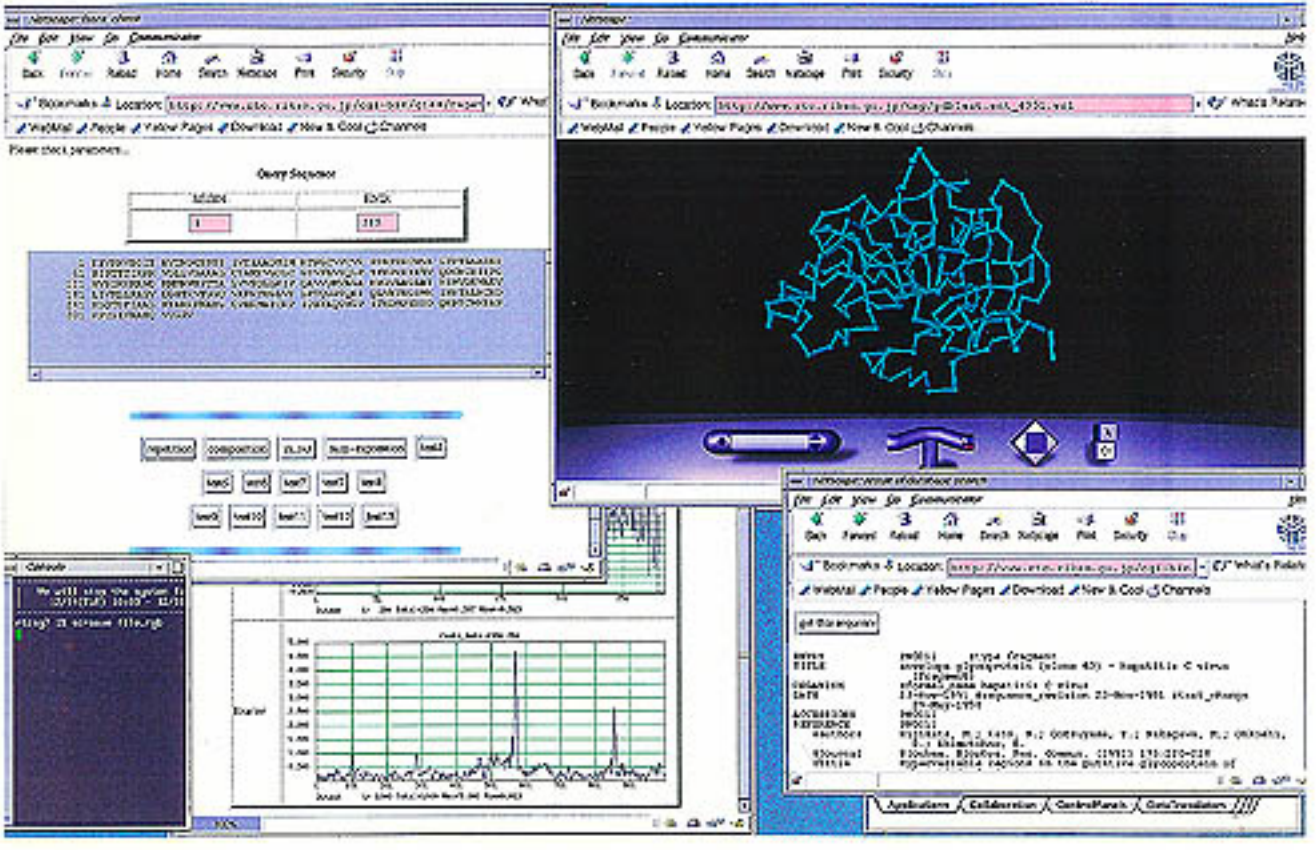
次ページに本システムにおいて利用可能な豊富な解析プログラム群を表１にまとめておきます（現在開発中のものも含まれます）。

表1 利用可能な解析プログラム群

Application programs
Databases
Retrieval:
· GenBank
· EMBL
· NBRF-PIR
· SWISS-PROT
· PDB
DNA
DNA → amino acid sequence （DNA配列からアミノ酸配列への変換）
Frequency of codon usage （コドンの使用頻度）
Restriction enzyme sites （制限酵素の切断部位）
Homology search （DNAのホモロジー検索）
Free energey calculation of basepairs*)（塩基対の自由エネルギー計算）
Prediction*)（人工知能による遺伝情報の機能部位の予測）:
· Coding region
· Splicing site
· Promotor region
Protein
Amino acid composition （アミノ酸配列からアミノ酸組成を求める）
Internal sequence repetition（自己相関関数による配列の繰り返し性の解析）
Prediction of exposed region（N_{14}法による露出領域の予測）
Prediction of folding type
Search for functional site（相互相関関数による機能部位の予測）
Homology:
· Correlation （相関法によるホモロジー計算）
· Dot-matrix
Homology search （蛋白質のホモロジー検索）
2nd structure prediction （二次構造予測）:
· Chou-Fasman
· Robson
· Joint
3D structure prediction （立体構造予測）

*）印は現在開発中のもの

[Appendix]
最終ページの図は*génie* の画面表示例。左側ウインドウに表示されているアミノ酸配列はC型肝炎ウイルスの外膜蛋白質のもの。右ウインドウに描かれた立体構造は膵臓の消化酵素であるエラスターゼ。ＶＲＭＬにより表示したものです。
因みに膵臓の消化酵素を大量に投与するガンの治療法がある（但し、正当医学ではない）。消化酵素が腫瘍の部分を消化分解すると言う単純な前提に基づいていると言われています。


Query Sequence
Bookmarks
Location:
WebMail
People
Yellow Pages
Download
New & Cool
Channels
Applications
Collaboration
Control Panels
Communicator