情報学基礎 10－4
（1988. 9. 16）

蛋白工学支援システムにおける解析アルゴリズム

藤本哲知、高橋裕信、窪田 綾[1)]、中島広志[2)]、大井龍夫[3)]
[1)]三洋電機筑波研究所、[2)]金沢大学医療短大、[3)]京都女子大学

ワークステーションSUN3/260上に、蛋白質一次構造（アミノ酸配列）データベースと一次構造解析プログラムとを結合させた蛋白工学支援システムを構築した。このシステムに従えば、データベースの任意の蛋白質を検索することができ、そのアミノ酸配列あるいは組成からその蛋白質の２次構造や folding typeの予測など様々な有用な情報を得ることができる。さらに、マウスの使用やマルチウインドウによる複数の解析結果の表示など、マン－マシン インター フェースの向上に努めており、その操作性は単純明快なものとなっている。本報告では、システムの概要並びに一次構造解析のアルゴリズムについて述べたい。

なお、本システムにおける一次構造データベース及び解析プログラムは我々独自によって構築且つ開発されてきたものである。

On the algorithm of nemerical analysis of protein sequence based on the correlation function

Tetsunori FUJIMOTO、Hironobu TAKAHASHI、Yasushi KUBOTA[1)]、Hiroshi NAKASHIMA[2)]、Tatsuo OOI[3)]
1)Laboratory of Intelligent Systems、Tsukuba Research Center、SANYO Elctoric Co.,Ltd.、2-1 Koyadai、Tukuba、Ibaraki、305
2)The School of Allied Medical Professions、Kanazawa Univ.、Kanazawa Ishikawa 920
3)Kyoto Women's University、Higashiyama、Kyoto 605

We are developing a system for protein sequence analysis, coupling the database of amino acid sequences of proteins with application programs on workstation SUN3/260. By the system we can retrieve any sequence data stored in the database and analyse them to get available information on folding type or secondary structure of proteins from their amino acid compositions and sequences. The fundamental algorithm of the analysis is based on the correlation function which is widely used for the analysis of random data.

Here, we will describe the algorithm of numerical analysis of protein used in the system.

１．緒言

蛋白質は核酸と共に生命の最も基本的な担い手である。中でも特に蛋白質は生化学反応、病気に対する免疫、分子識別などその機能は生命の維持に本質的且つ不可欠な役割を演じている。この機能は立体構造と相関があり、そして立体構造はそのアミノ酸配列(一次構造)によって一意的に決定される。蛋白質研究の最も重要な課題は与えられたアミノ酸配列に従って生体内で形成される立体構造を予測することである。この問題の完全な解答は現在、得られていない。このように極めて明快でありながら解決が困難な問題では他にはフェルマー の大定理があげられるであろうか？ さて、蛋白質はアミノ酸が線形に重合したものなので、アミノ酸残基をある適当な物理化学的なパラメータで置き替えれば蛋白質の有限な数列表現が得られる。この線形性のために我々は種々の統計学や情報学の手法を駆使して様々な立体構造に関する情報を得ることができる。また、１９７０年代にはSangerやMaxam-Gilbert法により DNA塩基配列決定に大きな進展を見る至った。このため従来はアミノ酸分析によって蛋白質の一次構造の決定を行っていたものが、DNA の塩基配列から翻訳することにより行われるようになった。ここに至って、蛋白質の一次構造に関するデータは飛躍的に蓄積された。このように、今や蛋白質の一次構造の数学解析やデータ処理にコンピュータの利用は必要不可欠のものとなっている。

他方、 microelectronics 技術の急速な発展によりいまやその能力においてmainframe に匹敵する EWS (Engineering Workstation)が広く普及するに至った。ＵＮＩＸを operating systemとして採用された EWS は計算能力のみならず bitmap display や multiwindow system にサポートされた優れたman-machine interface を実現している。さらには、RPC(Remote Procedure Call) や NFS (Network File System)に代表される分散処理システムは、様々のタイプの分散されたソフトウエア、ハードウエア資源へのアクセスを可能にしている。こうした高度なソフトウエア環境のもとで、EWS は蛋白質構造の解析研究に新たなる可能性を与えるであろう。このような理由のために我々はSUN 3/260 上に蛋白質一次構造の解析システムとデータベースを結合した統合システムを構築しつつある。このような統合システムは、九州大学大型計算機センターにおいてmainframe(FACOM M-382 OS IV/F4)上で久原らによっても試みられている[1,2]。

本報告では、我々が独自に構築し且つ開発して来たデータベースおよび蛋白質一次構造解析のアルゴリズムについて述べたい。

２．データベースと検索

蛋白質一次構造はその立体構造と機能に関する情報を秘めており、最も本質的な情報と考えられるので、我々はこの一次構造データを収集し一つのデータベースを構築した。各データは蛋白質の名前とアミノ酸の個数が記述される ”name ”で始まり”／／”で終わる。アミノ酸残基は一文字コードで表されている：例えば ｍはメチオニン、 ｅはグルタミン酸を表す等々。なお nbrf とあるのは 米国ＮＢＲＦ (National Biomedical Research Foundation) のデータベースと同じ蛋白質の entry name である（図１）。全

```
Protein sequence database.
name   759    p3 protein   human influenza a virus    updated   12/07/84
source        human influenza a virus (strain a/pr/8/34)
nbrf          p3iv34
reference     1   (sequence translated from the genomic rna sequence)
  authors     fields, s., and winter, g.;
  journal     cell 28, 303-313, 1982
comment       this protein is probably one of the three rna-dependent
              rna polymerases.
sites            from        to      description
    matp            1       759      mature protein
sequence      759 aa
     merikelrnlmsqsrtreiltkttvdhmaiikkytsgrqeknpalrmkwmmamkypitad     60
     kritemiperneqgqtlwskmndagsdrvmvsplavtwwnrngpmtntvhypkiyktyfe    120
     rverlkhgtfgpvhfrnqvkirrrvdinpghadlsakeaqdvimevvfpnevgariltse    180
     sqltitkekkeelqdckisplmvaymlerelvrktrflpvaggtssvyievlhltqgtcw    240
     eqmytpggevknddvdqsliiaarnivrraavsadplasllemchstqiggirmvdilkq    300
     npteeqavgickaamglrisssfsfggftfkrtsgssvkreeevltgnlqtlkirvhegy    360
     eeftmvgrratailrkatrrliqlivsgrdeqsiaeaiivamvfsqedcmikavrgdlnf    420
     vnranqrlnpmhqllrhfqkdakvlfqnwgvepidnvmgmigilpdmtpsiemsmrgvri    480
     skmgvdeysstervvvsidrflrvrdqrgnvllspeevsetqgtekltitysssmmwein    540
     gpesvlvntyqwiirnwetvkiqwsqnptmlynkmefepfqslvpkairgqysgfvrtlf    600
     qqmrdvlgtfdtaqiiklllpfaaappkqsrmqfssftvnvrgsgmrilvrgnspvfnynk    660
     atkrltvlgkdagtltedpdegtagvesavlrgflilgkedrrygpalsinelsnlakge    720
     kanvligqgdvvlvmkrkrdssiltdsqtatkrirmain                         759
//
```

図１　アミノ酸配列データのformat例

ての蛋白質はその生物的観点から分類され４１個のファイルに収められている。また、効果的な検索のためにアルファベット順に並べた蛋白質名とそれが記載されているファイル名からなるリスト・ファイルを用意した。キーボードから与えられる蛋白質名は、その蛋白質データが記載されているファイル名の情報を与えるこのリスト・ファイル上で探索される。このようにして、実際の蛋白質データへのアクセスはリスト・ファイル上で得られた情報をもとにしてなされる。なお、立体構造に関するデータは米国 Brookhaven 国立研究所のものを利用した[3]。

物理化学的パラメータ

我々は疎水性や、αらせんβ構造をとる傾向度などアミノ酸に固有な５３個の物理化学的パラメターを用意している。

検索アルゴリズム

２つの文字列 $s_1s_2\cdots s_m$ 及び $t_1t_2\cdots t_n$ が与えられた時、２つの文字 s_i と t_j の間のズレの量として；

$$f(i,j)=\min\{f(i-1,j)+1,\ f(i,j-1)+1,\ f(i-1,j-1)+d(s_i,t_j)\}$$

を導入した[4]。ここで、$d(s_i,t_j)$ は $s_i=t_j$ の時 0、$s_i\neq t_j$ の時 1の値を採る。この量をもとに、予め設定されたズレの許容範囲内の文字列を探しだすことで misspellingを許す検索を行う。なお、このアルゴリズムはＣ－言語で記述した。

３．一次構造の数学解析

本システムには次のような解析プログラムが組み込まれている；

(1) 配列の繰り返し性の検出[5、6]

(2) folding type（α，β，α／β，α＋βなどの折れたたみ方のタイプ）や細胞内外分泌予測[7、8]

(3) homology法による２次構造予測[9]

(4) 相関法によるsequence homologyの検出[6、10]

以下にこれらのアルゴリズムについて述べたい。

相関法による配列の繰り返し性[5,6]や sequence homology の検出[6,10]

２０個のアミノ酸は疎水性のような物理化学的な量で表現できるので、与えられたアミノ酸配列はこのような値の数列に変換できる。従って、配列の繰り返し性の程度を計るために自己相関関数が定義できる。即ち、n 残基の長さの蛋白質 X の数列に対して、τ残基のラグの自己相関関数を；

$$A(\tau)=\frac{\sum_{i=1}^{n-\tau}(x(i)-\bar{x})(x(i+\tau)-\bar{\bar{x}})}{[\{\sum_{i=1}^{n-\tau}(x(i)-\bar{x})^2\}\{\sum_{i=1}^{n-\tau}(x(i+\tau)-\bar{\bar{x}})^2\}]^{1/2}}$$

但し、$\bar{x}=\frac{1}{n-\tau}(\sum_{i=1}^{n-\tau}x(i))$、$\bar{\bar{x}}=\frac{1}{n-\tau}(\sum_{i=1}^{n-\tau}x(i+\tau))$

と定義する。ここで、{x(i)}はある適当な物理化学的な量で変換された数である。同様に、２つのアミノ酸配列、X（通常は X の部分列）と Y の間の homology の程度を計るために相互相関関数；

$$C(j)=\frac{\sum_{i=1}^{n}(x(u+i-1)-\bar{x})(y(j+i-1)-\bar{y})}{[\{\sum_{i=1}^{n}(x(u+i-1)-\bar{x})^2\}\{\sum_{i=1}^{n}(y(j+i-1)-\bar{y})^2\}]^{1/2}}\quad、$$

但し、$\bar{x}=\frac{1}{n}(\sum_{i=1}^{n}x(u+i-1))$、$\bar{y}=\frac{1}{n}(\sum_{i=1}^{n}y(j+i-1))$

を導入する。この式は、X のある固定した長さ n の部分列（u番目の残基で始まり、u+n-1 番目で終わる）を配列 Y の１番目の残基（j=1)から順次走査しながら相関値を計算して行くことに相当する。更にこの手法は２次元的に拡張できる[6,10]。即ち、

$$C_p(i,j)=\frac{\sum_{i=-k}^{k}(x_p(i+1)-\bar{c}_p)(y_p(j+1)-\bar{c}_p)}{[\{\sum_{i=-k}^{k}(x_p(i+1)-\bar{c}_p)^2\}\{\sum_{i=-k}^{k}(y_p(j+1)-\bar{c}_p)^2\}]^{1/2}}$$

ここで、(2k+1) は比較すべきsegment（アミノ酸配列の断片）の長さ――(2k+1) の長さの "frame"或いは、 window"――であり、$\bar{c}_p$はパラメータ p の２０個のアミノ酸上での平均値である。ここで、Ｓ／Ｎ比を良くするため相関関数の平均を採る；

$$\overline{A}(\tau)=\frac{1}{n}\sum_{p=1}^{n}A_p(\tau),\quad \overline{C}(j)=\frac{1}{n}\sum_{p=1}^{n}C_p(j),\quad \overline{C}(i,j)=\frac{1}{n}\sum_{p=1}^{n}C_p(i,j)$$

ここで、nはパラメターの個数である。上式は算術平均なのでn種のパラメータは互いに独立であることが望ましい。そうでなければ、たがいに相関するパラメータで求められた相関関数にweightがかかるからである。この問題を解決するために収集されたパラメータ群に因子分析の手法を用いて以下のようなパラメータ・セットを選び出した[6,10]；

⑴偏比容[11]、⑵折り返しを形成する傾向度[12]、⑶α－アミノグループのpK値[13]、⑷極性度[14]、⑸相対的な突然変異度[15]、⑹α－カルボキシルグループのpK値[13]

従って、n=6である。なお、$C_p(i,j)$を計算するための"window"の長さは11が適切である、即ち、(2k+1)=11。

これら6つのパラメータからなるセットを使えば良好な構造homologyが得られる（即ち、高い相関――0.4以上――が得られる領域に高い立体構造上のhomology が認められる）。

folding type や細胞内外分泌予測[7,8]

蛋白質のアミノ酸組成は２０次元空間の１点として表される。従って、蛋白質の集団はこの２０次元の組成空間にある分布を呈するであろう。そこで細胞の内外など生体内での位置が既知のものや、folding typeの既知の蛋白質を収集しその組成空間内においていくつかのグループに分けることを試みた。即ち、全データのアミノ酸組成の平均を原点にとり、各座標軸を適当に規格化しておく。まず、第一ステップとしてある半径内の蛋白質を中心グループとして集める。次に、この中心グループ外の蛋白質をreference として一つ採り、このreference point の方向から６０度の立体角に入る蛋白質を求め、そして最も高い密度の方向を、reference とする蛋白質をシフトして行くことで探し求める。この手順を集められる蛋白質が１０個以下になるまで次々と行う。こうして１３個に分類されたグループにはそれぞれ生体内の位置（細胞の内外）や生物学的機能（enzymeやnonenzyme）や folding type に強い相関があることが判明した。

Homology法による２次構造の予測[9]

この手法の基本的アイデアは、配列の似た（homologousな）蛋白質は類似のコンフォメーションを採るであろうと言う単純な発想から出ている。配列が既知の或る蛋白質を立体構造が既知のいくつかの蛋白質と比較し、これとhomologousな配列の２次構造（α、β、Coil）を調べ、その個数（n_α、n_β、n_c）を数え上げる。即ち $\overline{C}(i,j)$が0.3以上のものが８残基以上続いた時、その残基のペア（i，j）をhomologousとする ここで、２種類のweighting factor νとW を導入する。即ち、$0.3\leqq\overline{C}<0.4$の時ν＝１、$0.4\leqq\overline{C}<0.5$の時ν＝２、等々。Wは３つのコンフォメーション状態α、β、coilの存在比率を考慮したfactor である。こうして予測に使われるパラメターqとして

$$q_k=W_k\sum_{j=1}^{n_k}\nu_j\quad(k=\alpha,\beta\ \text{or coil})$$

但し、W_α=1.3、W_β=1.4、W_c=1.0

が採用される。

この予測法による精度は約60%程度であり[9]Chou-Fasman[17]やRobsonら[18]による方法よりも良い値となっている[16]。

4．解析例

ここでは、本システムによる3つの代表的な蛋白質の解析例を示したい。まず、自己相関関数を羊のケラチンB2A[19]に適用し配列の繰り返し性の有無を見てみる。次に、ヒトのインターロイキンil-2[20]に対してホモロジー法による2次構造の予測を行う。最後に、ペニシロペプシン[21]、エンドチアペプシン[22]及び血圧調節に関与すると言われているヒトのレニン[23]についてホモロジーの有無を検出する。レニンを除いてこの2つの立体構造は知られている。図2はケラチンが5残基ごとのはっきりした周期性を有していることがわかる（左下のウインドウ）。50残基のラグ付近から不規則な領域が続き、73残基から再び周期性が表れる。この蛋白質のパワースペクトルを右下のウインドウに示す。

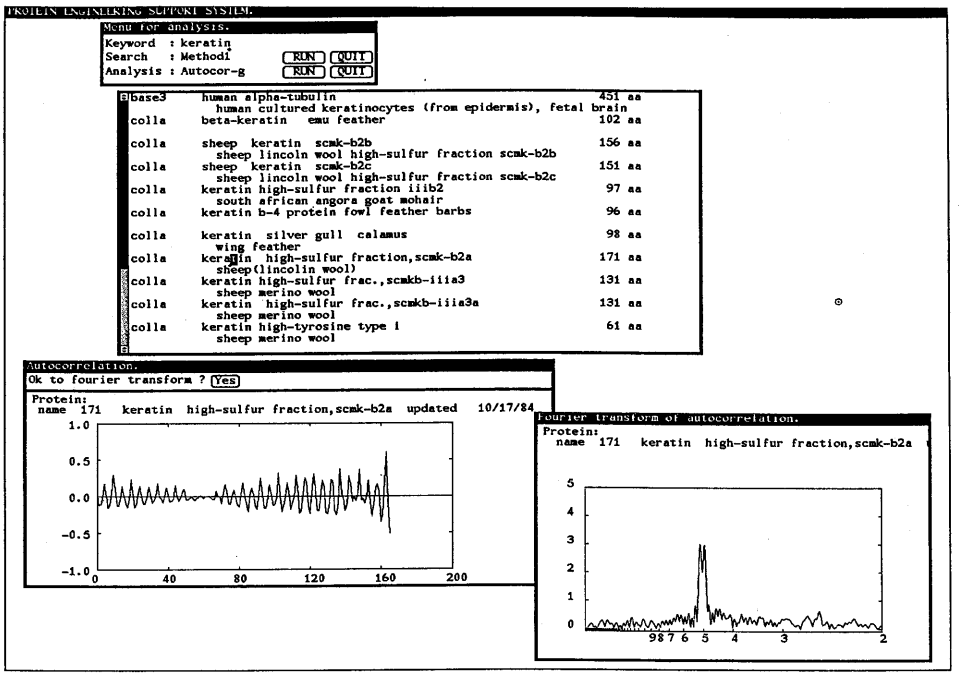

図2　自己相関関数によるケラチンの解析

図3はヒトのインターロイキンに対する結果である；この蛋白質のfolding typeはαタイプであることが予測されている（左下のウインドウ）.この上のウインドウにはこの予測の信頼率が表にされている。左上のウインドウはホモロジー法による2次構造予測の結果をシンボライズして表現されている。参考のために右側2枚のウインドウにRobson及びChou-Fasman法による予測も表示した。

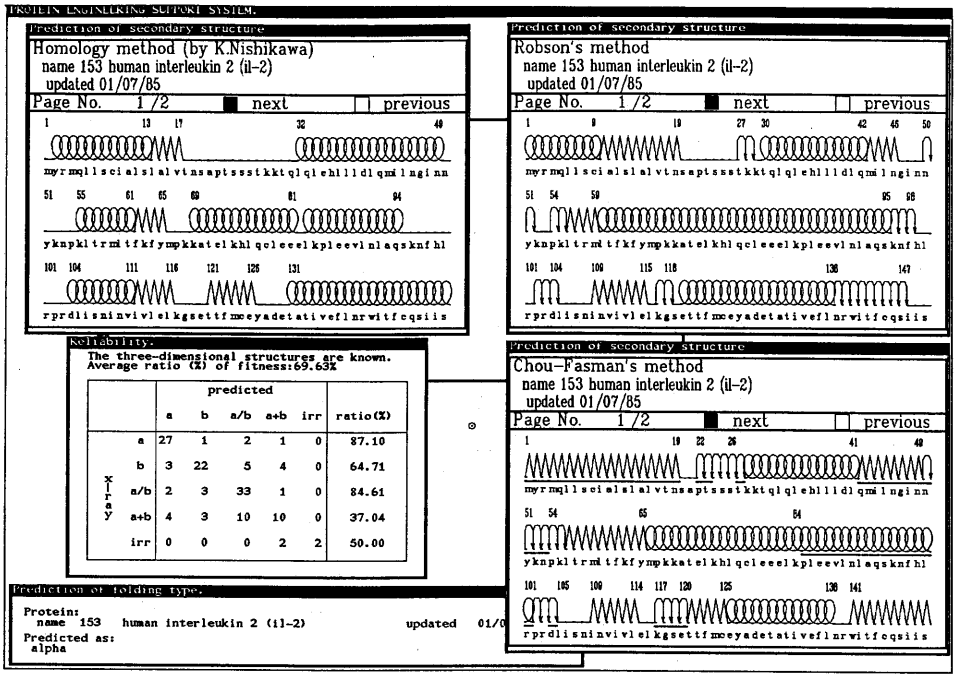

図３　インターロイキンil-2の folding typeと２次構造の予測

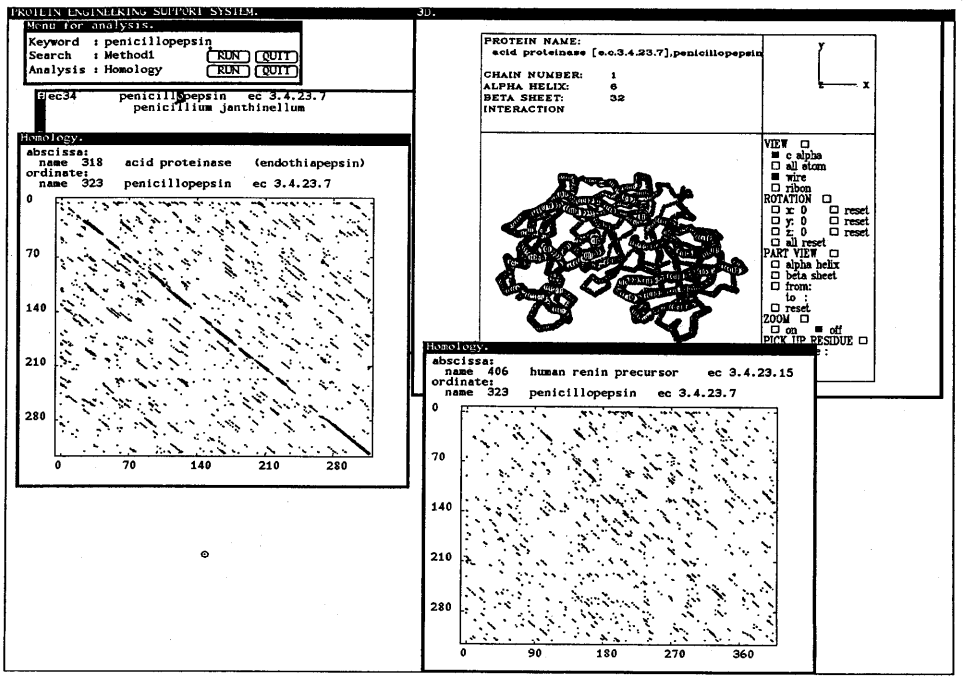

図４　penicillopepsinの、endothiapepsinと reninに対するhomology

図４ではペニシロペプシンに対して、エンドチアペプシンとレニンとのホモロジーを見た。エンドチアペプシンに対しては対角線上に 0.6以上の$\overline{C}$(i,

j)が連続して見られ（左上のウインドウ）これらの蛋白質の立体構造が似ていることが推測される。一方レニンに対してはそれ程のホモロジーは認められない（右下のウインドウ）。右上のウインドウにはペニシロペプシンの立体構造が３Ｄグラフィクスにより示されている。

５．結語

本システムによる一次構造解析の基本アルゴリズムは相関関数にもとずいている。この関数はいくつかの物理化学的パラメータで計算された相関値の算術平均をとることでＳ／Ｎ比を高めることに成功している。

更に、本システムは良好なマン－マシン　インターフェイスが得られるように、マルチ　ウインドウによる複数の解析結果を画面に表示するなど蛋白工学研究者の新たなる発想を促すのに役立てられると思われる。またマウスによる操作性の向上もあわせて計られている。

なお、本システムは蛋白質の立体構造予測に向けて引き続き発展中である。

参考文献

(1) S.Kuhara,et al:Nucelic Acids Research,12,89 (1984)
(2) 久原　哲、他：情報学基礎研資料８６－３、情報処理学会　(1986)
(3) F.C.Bernstein,et al: J.Mol.Biol.,112,535 (1977)
(4) 電子通信学会編："パターン情報処理", pp134-135,コロナ社 (1983)
(5) Y.Kubota,et al: J.Thoer.Biol.,91,347 (1981)
(6) Y.Kubota: Bull.Inst.Chem.Res.,Kyoto Univ.,60,309 (1982)
(7) K.Nishikawa,Y.Kubota and T.Ooi: J.Biochem.,94,981 (1983)
(8) K.Nishikawa,Y.Kubota and T.Ooi: J.Biochem.,94,997 (1983)
(9) K.Nishikawa and T.Ooi: Biochim.Biophys.Acta,871,45 (1986)
(10)Y.Kubota,et al: Biochim.Biophys.Acta, 701,242 (1982)
(11)E.J.Cohn and J.T.Edsall: "Proteins,Amino Acids,and Peptides" Van Nostrand-Reinhold,Princeton,New Jersey (1943)
(12)M.Levitt: Biochemistry,17,4277 (1978)
(13)H.A.Sober Ed.:"Handbook of Biochemistry,Selected Data for Molecular Biology" 2nd ed., The Chemical Robber Co.,Cleveland, Ohio (1970)
(14)R.Grantham: Science,185,862 (1974)
(15)M.O.Dayhoff Ed.:"Atlas of Protein Sequence and Structure" Vol.5, Suppl.3,National Biomedical Research Foundation,Washington,D.C.(1978)
(16)K.Nishikawa: Biochim.Biophys.Acta,748,285 (1983)
(17)P.Y.Chou and G.D.Fasman: Adv.Enzymol.,47,45 (1978)
(18)J.Garnier,D.J.Osguthorpe and B.Robson: J.Mol.Biol.,120,97 (1978)
(19)T.C.Elleman: Biochem.J.,130,833 (1972)
(20)T.Taniguti,et al:Nature,302,305 (1983)
(21)J.Tang,et al:Nature,271,618 (1978)
(22)T.L.Blundell,B.L.Sibanda and L.Pearl:Nature,304,273 (1983)
(23)T.Imai,et al:Proc.Natl.Acad.Sci.USA,80,7405(1983)