修　士　論　文

蛋白質類似部分構造の構造分布に基づいて
配列-構造相関を提示するシステムに関する研究

指導教官　佐藤　賢二　助教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

辰本　将司

2000 年 3 月

目　次

図　目　次

表　目　次

第 1 章
序論

1.1　研究の背景

代謝、運動、免疫、遺伝などのさまざまな生命現象は、生体中の蛋白質によって担われている。蛋白質はその特殊な構造により、高度でかつ多様な機能を有している。自然から得られる蛋白質の機能は非常に複雑であり、立体構造を決定する為には多くの時間と人手を必要とする。これらを解決する為に、蛋白質のアミノ酸配列から構造、機能へと導く相関関係を探索する研究が続けられてきた。現在では、解析技術の進展により蛋白質の構造-機能の相関に関する知見が徐々に得られつつある。しかし、これまでの研究では、配列-構造の相関は部分的にしか解明されていない問題がある。

人間の染色体には約 10 万の遺伝子が存在する。生体内では、遺伝子の塩基配列がリボソームにより蛋白質のアミノ酸配列に変換される。世界レベルのゲノム解析プロジェクトにより、ヒトの全塩基配列を同定する作業が急速に行われ、2003 年には、すべての遺伝子の塩基配列が決定される予定である。蛋白質のアミノ酸配列は、その情報を持つ遺伝子の塩基配列を、そっくりそのまま反映しており、3 個の連続した塩基配列の一組で、一個のアミノ酸を記述する。塩基配列からアミノ酸配列に変換する遺伝暗号が解明され、現在では、塩基配列からアミノ酸配列へと半自動的に変換できる。配列の決定と比較して立体構造の決定には、X 線結晶解析法や NMR 溶液構造解析法等の最新の技術を駆使しても多くの時間と人手が必要であり、これまでに構造が決定された蛋白質は、配列が決定されたものの 8%程度にすぎない。

「蛋白質の立体構造はアミノ酸配列により決定されている」という Anfinsen の仮説[1]に基づいて、アミノ酸配列から立体構造を予測する方法が研究されてきた。しかし、配列-構造の相関関係について決め手となる法則はまだ発見されていない。Chothia により、蛋白質の立体構造は、エネルギー的に安定するものは無限の種類ではなく有限個であり、その基本構造はたかだか千種類程度しかないことが推定された[2]。これまでに立体構造が決定された蛋白質は 1 万以上あり、立体構造分類データベース(SCOP)によるとこれまでに解明された基本構造は 400 を超える[3]。これまでの予測法で、有望とされている方法に蛋白質立体構造データベース(PDB)[4]の構造データを用いた折りたたみ(フォールド)予測法がある。しかし、立体構造既知の蛋白質との配列類似性が高い蛋白質についてはある程度立体構造を予測できるが、配列類似性のない蛋白質についての予測は非常に難しい。さらに、これまでの予測法では、配列全体の類似した蛋白質が PDB に登録されていないと予測精度が上がらないという問題がある。

1.2　研究の目的と効果

本研究の特色は、PDB に登録されている構造既知のアミノ酸配列を部分配列(フラグメント)単位に分解し、そのレベルで配列-構造相関を解析する点にある。これにより、蛋白質全体を比較してもわからなかった局所的な配列-構造相関を指摘できる可能性がある。さらに、分解した部分構造集合に対して網羅的に解析を行うことで、予測し易いと思われる部分とそうでない部分を分かりやすく提示できることが期待される。また、この配列-構造相関提示システムでは、ユーザに対する利便性を考慮してオプションの設定を自由に設定できる。これにより、精度の高い情報を厳選して提示するか、精度が低くても浅く広く情報を提示するか使い分けることができる。さらに、蛋白質の構造を調べるにはその形を「視る」ことが重要である。これに対し本システムでは、ブラウザ上でアミノ酸配列、二次構造、立体構造をそれぞれのレベルで可視化表示できる。

1.3　本論文の構成

本論文の構成は以下のようになっている。

第 1 章において、本研究の背景と目的および効果について述べる。

第 2 章において、研究内容を理解するのに必要な基本的概念や関連研究についての説明を行う。

第 3 章において、システムの設計について説明を行う。

第 4 章において、構築したシステムの検証を行う。

第 5 章において、残された課題について述べる。

第 2 章
蛋白質モデリングの現状

2.1　アミノ酸の性質

アミノ酸は炭素原子を中心に側鎖と水素原子とアミノ基とカルボキシル基が接続した分子構造を持っている。蛋白質は、20 種類のアミノ酸がペプチド結合で分岐なく鎖状につながった高分子である。短いもので数十個、長いものでは数千個ものアミノ酸が連り、巨大で複雑な分子を構成する。ポリペプチド鎖におけるアミノ酸の順序(配列)を一次構造という。この鎖の長さとアミノ酸の並び方の多様性が、必要とされるさまざまな蛋白質を構成する素となる。蛋白質分子の立体構造、物理化学的諸性質、および機能は、アミノ酸の配列と側鎖の性質に大きく左右される。アミノ酸の違いは、側鎖の違いである。それぞれの側鎖は性質が異なっていて、他の側鎖で置き換えることができない。

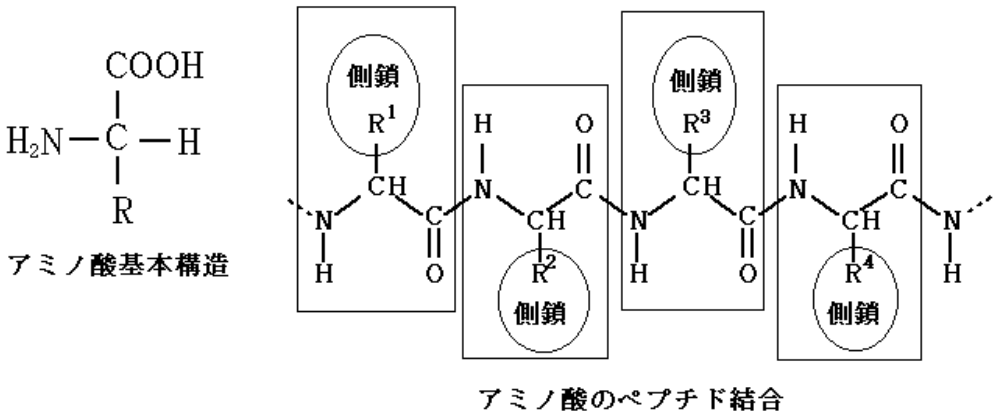


図 2.1　アミノ酸の基本構造とペプチド結合

3文字表記	1文字表記	アミノ酸
Ala A Alanine	His H Histidine	Thr T Threonine
Arg R Arginine	Ile I Isoleucine	Trp W Tryptophan
Asn N Asparagine	Leu L Leucine	Tyr Y Tyrosine
Asp D Aspartic acid	Lys K Lysine	Val V Valine
Cys C Cysteine	Met M Methionine	Asx B Asn or Asp
Gln Q Glutamine	Phe F Phenylalanine	Glx Z Gln or Glu
Glu E Glutamic acid	Pro P Proline	Unk X Unknown
Gly G Glycine	Ser S Serine	

表 2.1　アミノ酸種類と表記

アミノ酸は、枝分かれした側鎖の性質によって、極性アミノ酸と非極性アミノ酸とに分けることができる。極性のアミノ酸はさらに中性、塩基性、酸性と分類することができる。また、アミノ酸を大きさで分類することもでき、一般に蛋白質の構造にとって側鎖の大きさは、化学的性質と同じくらい重要である。

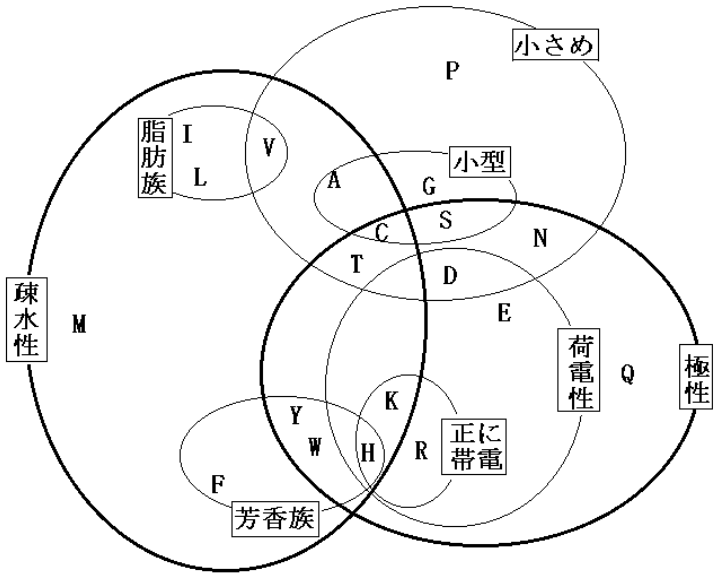


図 2.2　アミノ酸の分類

このようなアミノ酸の各性質に関してまとめたものに、京都大学化学研究所で作成されているアミノ酸指標データベース(AAindex)[5]があげられる。AAindex では、文献に基づいたアミノ酸の指標を集積し、クラスター分析することによりアミノ酸相互の類似性が解析されている。解析結果より、側鎖の物理化学的性質は大きく 5 つに分類することができる。また、2 つのアミノ酸同士の類似度を定義したアミノ酸置換行列(Matrix:後述)の解析が行われている。下図は、アミノ酸指標、アミノ酸の置換行列、さらに場所依存の置換行列とみなすことができるプロファイルの関連を示している。

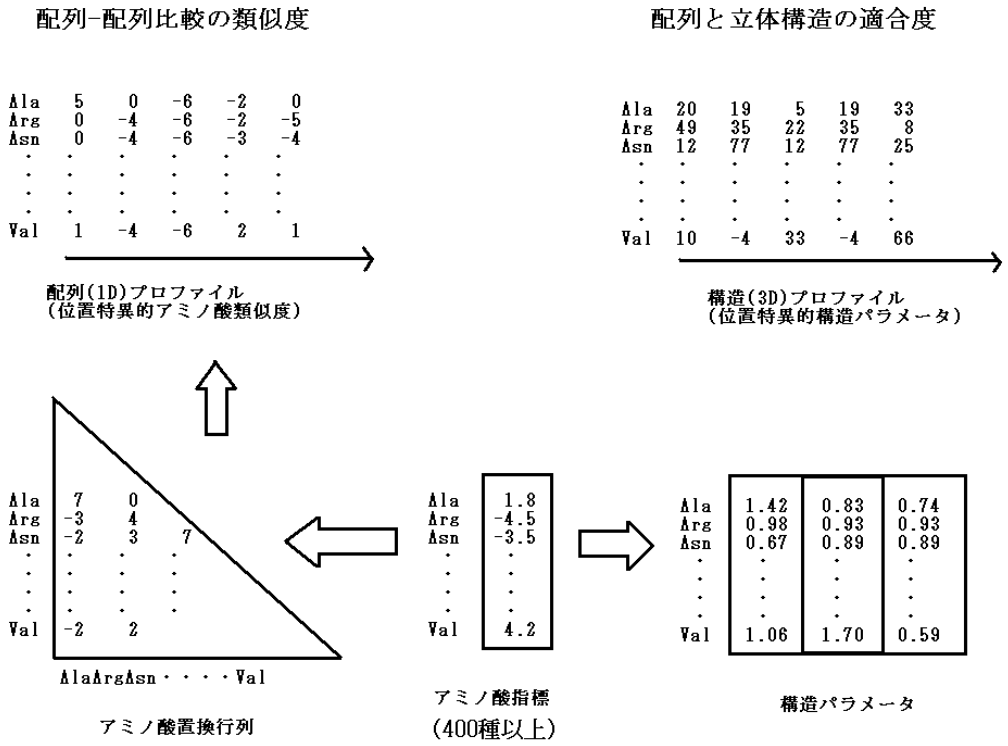


図 2.3　アミノ酸指標とプロファイルの関連

2.2　蛋白質のモデリング

2.2.1　蛋白質の代表的な構造と配座的性質

蛋白質の複雑な立体構造は、歴史的背景から、一次構造、二次構造、三次構造および四次構造の 4 つのレベルにわけて考えられている。

- 一次構造：蛋白質を構成する個々のアミノ酸の配列を表す。
- 二次構造：ポリペプチド鎖の部分部分でみられる局所的な構造(αヘリックスやβシート：後述)。側鎖の相互作用による二次構造の会合様式を記述する為に、超二次構造とよばれる構造単位も導入された。
- 三次構造：折りたたまれたポリペプチド鎖の全体的なトポロジーを示す。
- 四次構造：複数のサブユニット(単量体)が会合してできた蛋白質の立体構造。

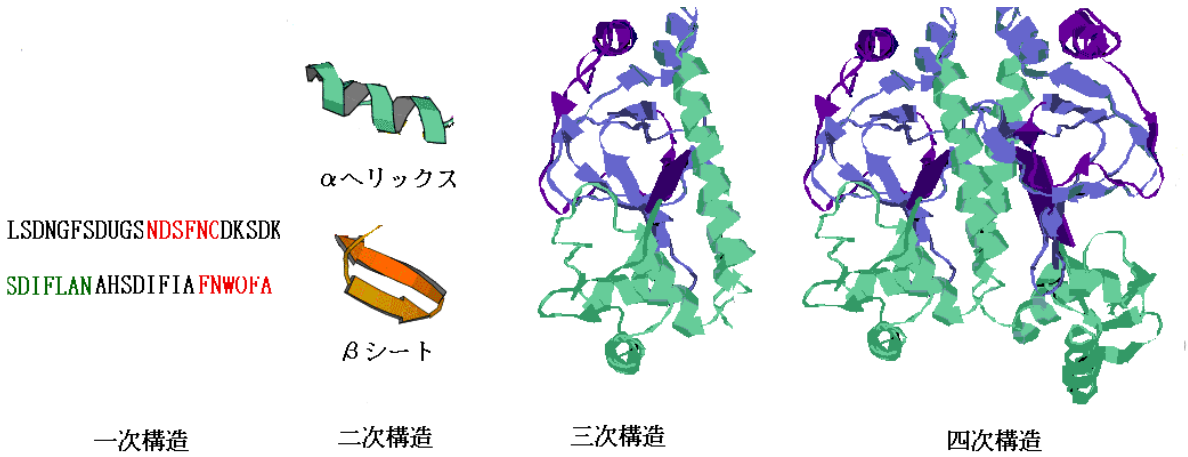
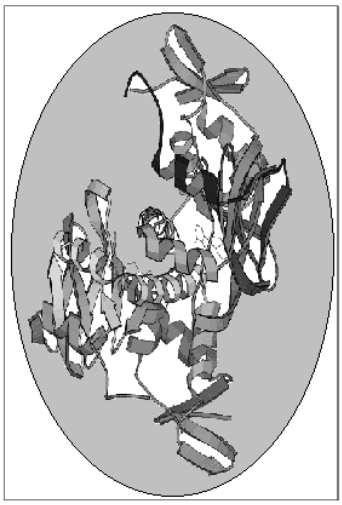
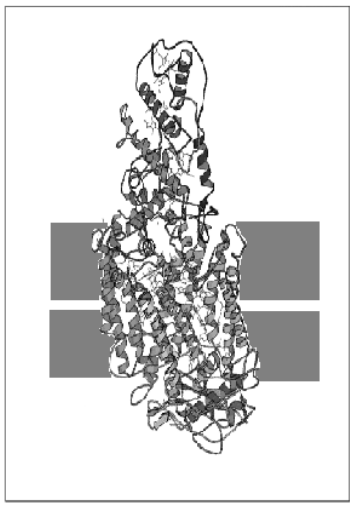


図 2.4　蛋白質の構造

蛋白質は、その存在する環境により、水溶性蛋白質と膜蛋白質とに分類できる。水溶性蛋白質は、生体内の細胞質に溶けて他の分子と結合したり解離したりしながら、代謝や免疫などの働きをする。膜蛋白質は生体膜に埋もれており、細胞内外での低分子やイオンを輸送する機能がある。この 2 つは相互作用および構造という観点から比較すると明らかに異なっている。水溶性蛋白質は水中での存在を可能にする為に、内部では疎水性、外部では水溶性の構造をつくり出す必要があり、蛋白質の構造、分子内部の相互作用は複雑になる。膜蛋白質の場合は、脂質という疎水性の環境に存在するので、膜内および水中での相互作用を明確に分離できる。この為、構造、相互作用に関する問題を単純化できる。その他に生体内では、爪、毛髪、皮膚、腱、軟骨や細胞マトリックスなどを構成しているコラーゲンやケラチンといった繊維状蛋白質などがある。





	水溶性蛋白質	膜蛋白質
環境	自分で作る。	外部から与えられている。
相互作用	(相互作用の寄与の分離が難しい。)	(水中と膜内でおもに働く相互作用を分離できる。)
構造	比較的複雑	比較的単純
立体構造情報	数千個	数十個
適した構造予測法	エネルギー依存型予測法	知識依存型予測法

図 2.5　水溶性蛋白質と膜蛋白質の比較

個々のアミノ酸側鎖が骨格や他の側鎖とどのように相互作用し、二次構造や三次構造の中でどのような役割をしているかは、蛋白質の立体構造を理解する上で特に重要である。蛋白質の構造は主鎖の三次元的な形状と各側鎖の回転角で表すことができる。蛋白質の主鎖のねじれ角(二面角)は、アミノ酸残基の中心炭素原子からみると、ペプチド結合の炭素との回転角 ϕ、と隣のペプチド結合の窒素との回転角 ψ、さらに、側鎖の回転角 ω で記述される。ϕ と ψ の変化は、隣接した非結合原子間の立体的な衝突により幾何学的に制限を受ける。また、ϕ と ψ によって蛋白質の骨格はほとんど完全に記述できるので、立体構造の妥当性を判定するのに重要な値である。

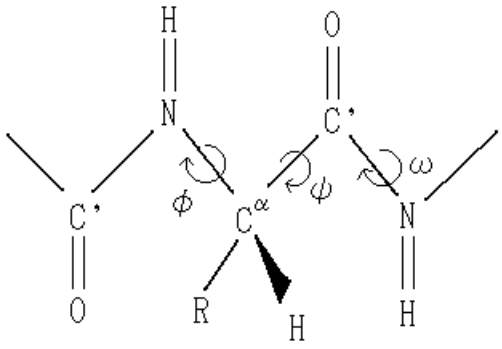


図 2.6　二面角

2.2.2　二次構造の類型

蛋白質の代表的な二次構造には、α ヘリックスと β シートとターンがある。二次構造は蛋白質の立体構造で特徴的な部分構造を指す。なお、それらの二次構造に属さない部分のことをコイルとすることで、蛋白質の全体がいずれかの二次構造で表せるようにしている。しかし実際の二次構造の定義は、蛋白質立体構造データベース(PDB)をみても明らかなように、研究者の判断に任されており、ある程度の曖昧性が存在する。

- αヘリックス：　αヘリックスのうち右巻きのものは、蛋白質内部で最も容易に識別できる二次構造として知られている。既知の球状蛋白質に含まれる残基の約28～38%はαヘリックスの状態で存在している[6]。αヘリックスは、一種の反復構造である。すなわち、らせん軸に沿って5.4Å毎に構造が繰り返され、1周期あたり3.6個のアミノ酸残基を含む。αヘリックスの構造は、残基nのカルボニル基と残基(n+4)のNH間で繰り返される水素結合により安定な構造になっている。αヘリックスは、その環境に依存して変化することが知られており、理想的なαヘリックス構造(φ=-57°、ψ=-47°)は、類似したいくつかの構造の中の1つにすぎない[7]。実際の蛋白質では、少し異なる構造(φ=-62、ψ=-41°)が観測される[8]。この配座は、水溶媒とも水素結合を形成できる点で、理想的な構造よりも有利である。

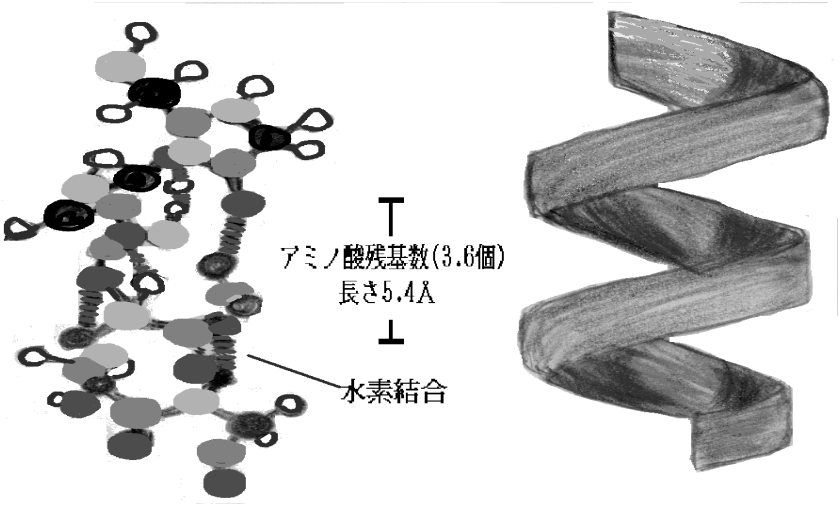


図2.7　αヘリックス

- βシート：　球状蛋白質では構成残基の約20～28%をβシートが占めている[6]。βシートはαヘリックスの次に高い規則性を持ち、また、周期性を備えている。ポリペプチド骨格が線状の伸張配座をとった状態をβストランドとよび、これが集合したものをβシートとよぶ[9]。同一鎖内の残基間で相互作用は可能ではないので、βストランドはより複雑な構造であるβシートの構成要素としてのみ安定である。ポリペプチド骨格にある水素結合の供与基と受容基はすべて水素結合の形成に関与している。しかし、これらの結合は鎖間で形成されるので、エネルギー的にαヘリックスより不利である。βシートは配列上の広範囲に分散したβストランドから形成される為、蛋白質の立体構造に与える影響は大きい。

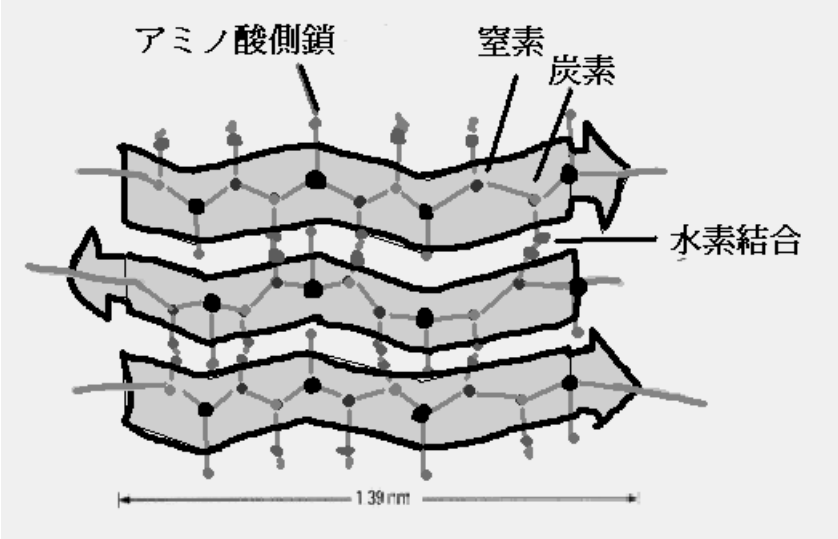


図 2.8　β シート

- ターン：　球状蛋白質を構成する全残基の約三分の一はターン(ループ)領域に存在している。代表的なターンとしては、逆平行なβストランドをつないだβターン(ヘアピンループ)がある[10]。蛋白質の表面には、帯電したアミノ酸や極性アミノ酸を含むターンが良くみられる。βターンの多くは 7 残基よりも短く、一般に大きなループの配座は明確に定まらない。逆転ターンでは、ペプチド基は規則的な水素結合による対は結合しない為、蛋白質の表面に現れることが多い。

αヘリックスやβシートは、蛋白質の分子中に非常に多くみられる構造である、この観点から、蛋白質を 5 種類に分類することができる。

- α蛋白質：　αヘリックスのみを含む構造。最も少ない。
- β蛋白質：　最も機能的に富んだ構造で、主としてβシートを含む構造。2 番目に多い。
- α+β蛋白質：　αヘリックスの領域とβシートの領域が離れてランダムに存在する構造。最も多い。
- α/β蛋白質：　一次構造上でαヘリックスととβシートが交互に規則的に存在する構造。3 番目に多い。
- ランダム(コイル)蛋白質：　規則的な二次構造を持たない構造。

20 種の天然アミノ酸の多くが、統計的にみて特定の二次構造をとり易い傾向をもっている、たとえば、Ala、Arg、Gln、Glu、Met、Leu および Lys はαヘリックスに多く見出されるのに対し、Cys、Ile、Phe、Thr、Trp、Tyr および Val はβシート中に多く現れる傾向がある。二次構造はあまり長くなく、αヘリックスで 10～15 残基、βストランドは 3～10 残基程度である。

蛋白質の構成要素である二次構造を図的に表現するときには、αヘリックスを円筒やらせん状リボンで表し、βストランドを骨格のアミノ酸末端からカルボキシ末端へと向かう幅広の矢印によって表すことが一般的である。これにより、蛋白質の全体像を明確に浮かび上がらせることができる。

2.2.3　相同蛋白質

共通の祖先から進化した蛋白質は相同であるといわれる。2 つの相同蛋白質の配列はほぼ同一のこともあり、また大規模な変異により大きく異なることもある。相同蛋白質における配列類似性の保存状態は立体構造上の類似性(構造類似性)のそれに比べるとあまり良くはない。蛋白質の立体構造は、進化の間に配列上の変化はあった場合でも良く保存されている。このことから、蛋白質が特異的な機能を発現する為には、共通の構造が不可欠であると予想されている。

相同蛋白質でも、配列相同性が低く構造類似性が高い蛋白質では、一般に蛋白質の表面に近い領域(ループ)に配列上の違いが顕著に現れることが経験的に見出されている。これらの領域では、側鎖の物理化学的性質さえも異なっている。しかし、蛋白質内部では、残基の変化は表面に比べて少なく、中心を形成する二次構造は、相同蛋白質間では良く保存されている。相同蛋白質では、二次構造は位置や長さを変えたり、完全に欠落することもあるが、αヘリックスがβシートに置き換えられたりすることはない。蛋白質の配列データの比較解析は、生物の進化史および進化機構の解明、蛋白質の機能および機能部位の推定、蛋白質の三次構造あるいは部分構造の推定に用いられる。

2.2.4　蛋白質のフォールディング

リボソーム上で塩基配列がアミノ酸配列に翻訳されると、細胞内でアミノ酸配列が折りたたまれ、蛋白質としての活性を持つようになる。この過程はフォールディングとよばれている。フォールディングを経てポリペプチド鎖が特異的に折りたたまれた状態ではじめて、蛋白質はその機能を発現することができる。

蛋白質のフォールディング機構については、ポリペプチド鎖に沿って比較的近距離の相互作用で安定化される二次構造が形成され、その後側鎖の特異的な三次構造が形成されるとする階層モデルがある。

アンフィンセンの実験により、蛋白質のネイティブ構造は、アミノ酸配列と溶媒環境によって決まる自由エネルギー最小の状態であることが示された。初期のフォールディングの研究では、蛋白質の立体構造の形成反応が自由エネルギー最小へと向かう熱力学過程を探る為に、中間状態(折り紙モデル)の研究が盛んに行われた[11-13]。

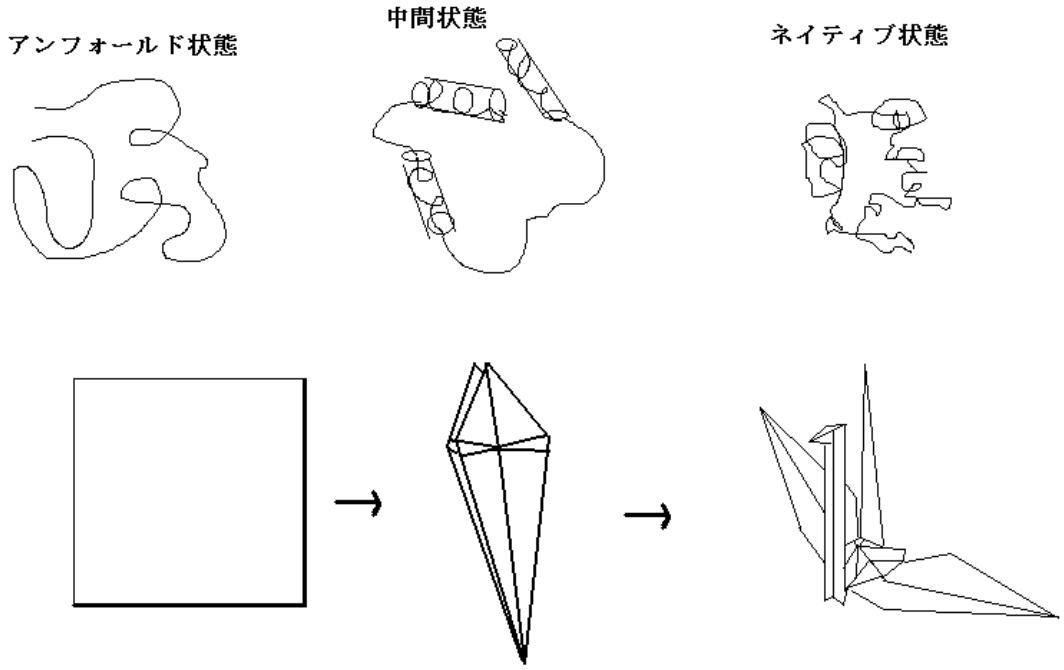


図 2.9　蛋白質の折りたたみ反応と折り紙モデル

しかしその後、球状蛋白質の変性・再生反応は協同的に起こり、自然状態とアンフォールドした状態の二状態転移でうまく近似されることが実験的に示された。この実験結果により、蛋白質の折りたたみの中間状態が不安定であることが明らかになった。

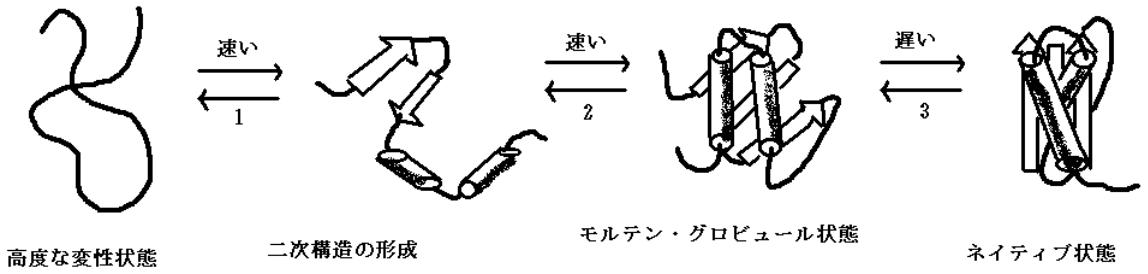


図 2.10　モルテン・グロビュールと構造形成モデル

ところが、酸や熱などによって引き起こされる変性状態で、例外的な中間状態構造が発見された。このときの中間状態を、分子はコンパクトな球状(グロビュール)であるが、特異的な三次構造が溶解(モルテン)しているという意味をこめて、モルテン・グロビュールとよんだ。モルテン・グロビュールは典型的な階層モデルで、はじめに水素結合によりネイティブと同様な二次構造が形成される(反応 1)。次に二次構造セグメントが疎水的相互作用により集合してコンパクトな中間状態が形成される(反応 2)。この状態がモルテン・グロビュール状態であり、特異的な三次構造はまだ形成されてない。構造形成反応の最もゆっくりした段階(律速段階)はモルテン・グロビュールとネイティブ状態の間(反応 3)にある[14]。

フォールディング・ファネル(漏斗)・モデルでは、蛋白質の構造に対して、その構造の持つ自由エネルギー値が与えられている。時間とともに、蛋白質のアミノ酸の結合角が変化し、蛋白質が持つ自由エネルギーがより少なくなる構造に折れ曲がっていく。蛋白質の初期状態の位置から、漏斗の穴に落ちるまでに通った構造の経路がフォールディング経路となる。初期状態は無数に存在するが、ネイティブ構造にたどり着

く経路は 1 つしかない。

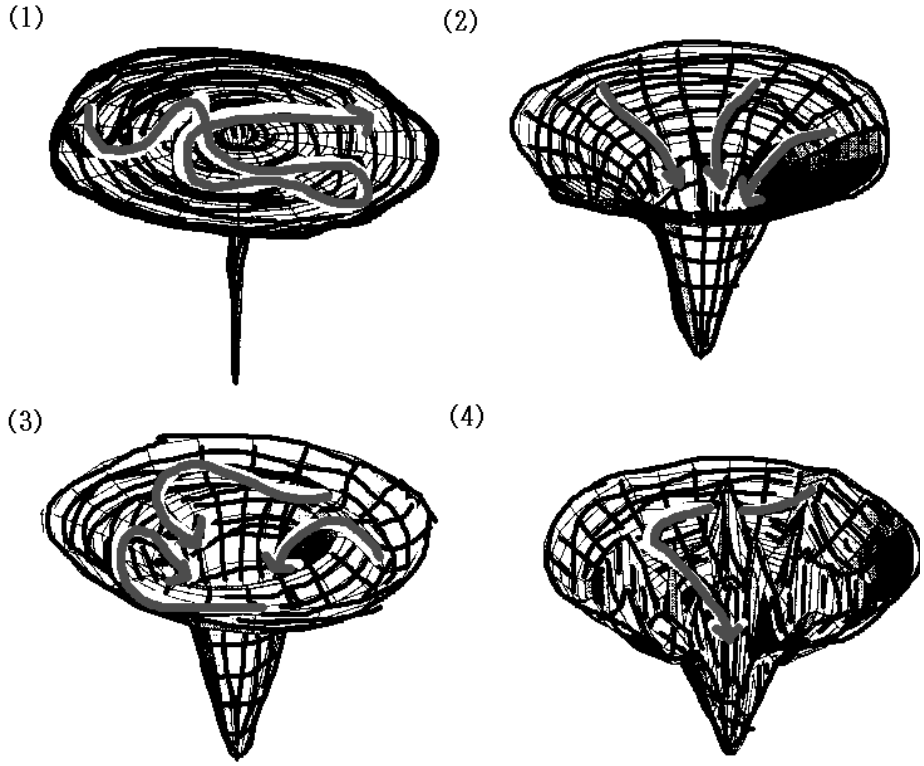


図 2.11　フォールディング・ファネル反応のエネルギーモデル

1　ランダムなフォールディング。
2　エネルギーの障壁のまったくないスムースなフォールディング。
3　いくつものエネルギー障壁を持つフォールディング。
4　多くのエネルギー障壁を持つフォールディング。

蛋白質のネイティブ構造を安定化する力は、配列上で近くに位置するアミノ酸間の相互作用(局所的相互作用)と、離れたアミノ酸間の相互作用(非局所的相互作用)に分けることができる。ネイティブ構造はこれらの総和によって決まる全体的な自由エネルギーの最小状態である。

2.3　代表的な蛋白質モデリング法

2.3.1　知識依拠型蛋白質モデリング法

知識依拠型蛋白質モデリングは以下の段階から成り立っている。

1　目的蛋白質と関連性を持つ構造既知の参照蛋白質の類似検索。
2　構造不変領域(SCR, structurally conserved region)と構造可変領域(SVR, structurally variable region)の同定。
3　目的蛋白質と参照蛋白質の間での SCR 配列の並置。
4　参照蛋白質の鋳型構造の座標を利用した目的蛋白質の SCR の構築。
5　SVR の構築。
6　側鎖のモデリング。
7　エネルギー極小化法と分子動力学を利用した構造の精密化。

1　目的の蛋白質の配列を立体構造既知の蛋白質の配列と比較して相同配列を同定する。相同性配列の検索プログラムとして FASTA[15]、BLAST[16]があげられる。
2 - 5.　互いに強い関連のある同族タンパク質では、二次構造、とりわけαヘリックスとβストランドは配列全体を通じて相対的に同じ配向で存在することが知られている。その為、これらの領域は同族の他の蛋白質に対して原子座標を割り当てる際に、その基本的な枠組として利用される。同族蛋白質では、一般に構造の大部分は非常に類似しており、構造的に不変な領域(SCR)が大部分を占めると考えられるが、構造がかなり異なる領域も存在している。可変領域（SVR）はほとんど配列相同性を示さず、残基の挿入や欠落が起こり易い部位であるのに対し、SCR は高い配列相同性を示すことが、相同蛋白質に関する多くの研究から明らかにされている。蛋白質構造の間にみられる重大な差異はループ領域に現れることが多い。この領域では、アミノ酸の挿入と欠損により配列の長さが変化する。構造が不明な領域のモデリングを行う上で良い指針となるのは、相同蛋白質における長さの等しい部分配列(セグメント)の構造で

ある。

6. 側鎖の正確な配座は、蛋白質の構造内部でそのアミノ酸が置かれた局所環境に基本的に依存する。蛋白質の内部では疎水相互作用が支配的であり、アミノ酸残基の隙間のない充填はこの相互作用により実現されている。二次構造や他の残基との三次構造的な接触もまた、側鎖の配座に影響を及ぼすことが知られている。

7. SCR と SVR をつなげている領域では、立体的に大きな歪みが生じているので、エネルギー極小化を行ない、構造を緩和させる必要がある。エネルギー極小化と分子動力学の計算は、配座空間の局所領域を探索し、さらに磨きのかかった構造をつくり出す上で有効である。

蛋白質の立体構造形成の予測モデルでは、アミノ酸のどの部分がαヘリックスを形成し、どの部分がβストランドを形成するかを予測する二次構造予測から、二次構造がどのような空間配置で詰め合せているか(パッキング)を予測する二次構造のパッキング予測へと立体構造予測のアプローチが考えられた。二次構造予測法には、20 種の天然アミノ酸の多くが、統計的にみて特定の二次構造をとり易い傾向をもつという事実を利用した統計的方法、側鎖の疎水的、親水的および静電的性質の分析を通じ、タンパク質の折りたたまれ方の規則を見つけ出そうする立体化学的方法および相同性／神経回路網依拠型の方法といった 3 種類の方法が利用されている。代表的な二次構造予測法には、局所的な類似配列を用いたホモロジーモデリング法と、局所的な類似構造を用いた Salamov-Solovyev 法、ニューラルネットによる予測法などがある。現在の最新の二次構造予測法では、配列情報をニューラルネットで学習させた予測法で約 72%の予測精度が得られている[17]。

ホモロジーモデリング法は、すでに立体構造がわかっている蛋白質との配列相同性を利用して、蛋白質の立体構造を予測しようとする方法である。ミオグロビンとヘモグロビンの構造が決定されたことにより、両蛋白質の間に進化的類縁関係が構造上、機能上で類似性が認められた。しかし、配列上の類似性を比べてみると、同一残基率は 25%しかなかった。配列上の相違にもかかわらず、構造、機能においてホモロジーな蛋白質がその他にも多く見つかり、共通した祖先から進化した蛋白質間ではアミノ酸配列が大きく変化しても立体構造は保存されることが見出された。この経験則からホモロジーモデリングが生まれた。配列間の有意な類似性(一般に同一残基率 25%

以上)は、配列と構造の相関関係の証拠になる。

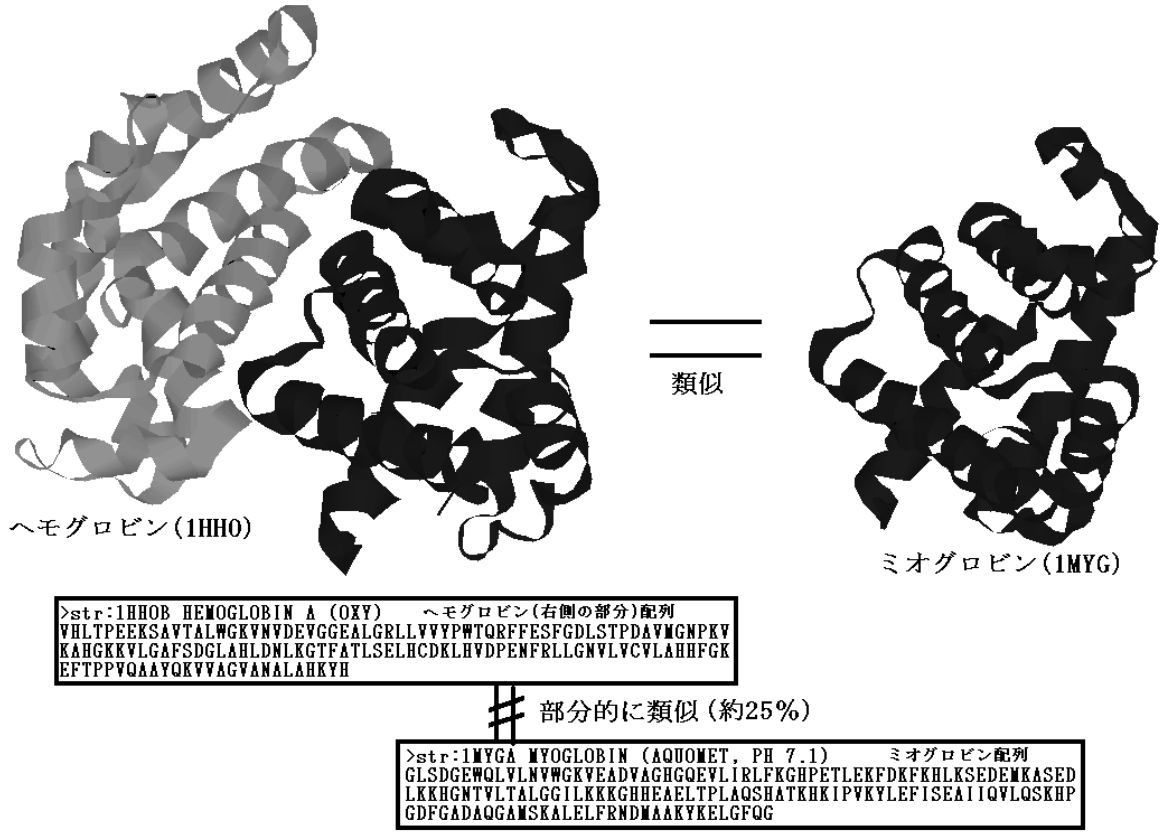


図 2.12　ミオグロビンとヘモグロビンの配列と構造

3D-1D 法は、アミノ酸配列はわかっているが立体構造は未知の蛋白質に対して、その配列を予め一定の評価で選び出した立体構造のライブラリと直接比較し、配列に適合する既知の立体構造があるかどうかを捜して構造を予測する方法である。立体構造のライブラリを作ることと、配列-構造適合性評価関数を作ることが基本になる。与えられた構造に配列を”乗せ”てみて、矛盾がないかどうかを定量的に評価する方法である。ここで、配列を構造に乗せるとは、構造上で配列を前後にずらしたり、挿入、

欠損を考慮したギャップを入れるアライメントにより、最適な適合性を調べることである。3D-1D 法では立体構造と配列の適合性は評価関数によって定量的に評価される。3D-1D 法にはアライメントの問題点がある。アライメントとは対応関係のことであり、このとき、構造に対して配列をどのように乗せるかが問題である。最適なアライメントをいかにして求めるかは動的計画法(ダイナミックプログラミング、DP)とよばれる数学的手法を用いることによって、最適解が一意的に求められることが、Bowie らによって示された[18]。3D-1D 法は PDB の登録増加と CASP(蛋白質立体構造予測コンペティション)によって急速に発展した。各アミノ酸種が立体構造中の特定の環境あるいは状態下にどの程度の頻度で出現するかを示すパラメータを、構造既知データの統計的解析に基づいて定義し、それらを用いて配列と構造との間の全体的な適合性を示す評価値を計算する。このような基本的な方法論はほぼ確立されているが、配列-構造適合性の検出能力(予測精度)は十分ではない。

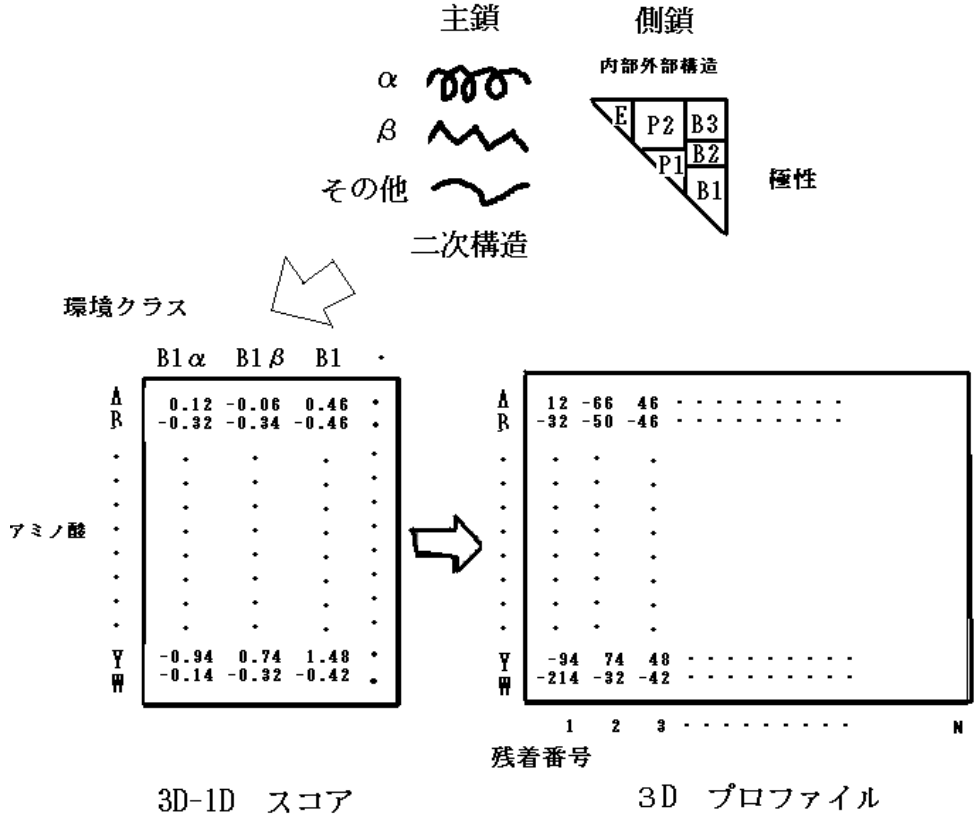


図 2.13　3D-1D 法の手順

立体構造中でアミノ酸が置かれている環境を、3 つの二次構造状態と 6 つの側鎖の状態からなる合計 18 のクラスに分類する。これにより 1 つの立体構造は 18 の記号からなる 1 つの記号列に変換される。立体構造既知の蛋白質で各アミノ酸が各クラスに存在する頻度を調べ、3D-1D スコアのマトリックスを定義する。環境クラスの記号をこのマトリックスに対応するカラムに変換することで、1 つの立体構造の記号列は、1 つの 3D プロファイルとして表現される。

スレッディング法は、立体構造上のアミノ酸間の相対距離に着目し、特定の相対距離に出現する 2 残基または 3 残基の組合せの出現頻度から部分構造とと配列の適合性を評価する方法である。配列情報を立体構造に直接マップしてその整合性の度合いを統計ポテンシャルで測ることにより、離れたアミノ酸間の非局所的相互作用を考慮している。

2.3.2　エネルギー依拠型モデリング

エネルギー依拠型モデリングは、指定領域において可能なアミノ酸配座をサンプリングし、その中から最低エネルギーの配座を見つけ出す為に、幾何学的な判定基準を使用する。エネルギー依拠型モデリングは、情報科学で発展してきた探索手法が本質的には用いられる。大きく分類すると 3 つに分けることができる

(A)　一般的に用いられるモンテ・カルロ(MC)法あるいは原子間力のシミュレーションを行う分子動力学(MD)法。
(B)　ヒューリスティックな手法としてのシミュレーティッド・アニーリング(SA)法あるいは遺伝的アルゴリズム(GA)計算による手法。
(C)　完全に全てを探索する手法(Exhaustive sampling)。

（A）の手法についての長所は、いろいろな物理量の統計平均を得ることができ、実験値に対応させることが可能であることがあげられる。ただし、ネイティブの蛋白質構造はあまりに協同性が高い為に、内部回転角や原子座標に独立の自由度を与えて

しまうので、通常の MC 手法によって構造を探索しても、効果的なシミュレーションは達成できない。その為、生体高分子のシミュレーションにおいては、MC 計算よりは MD 計算が主流である。しかし、MD の構造探索における欠点として、よほど長いステップ数のシミュレーションをしない限り、初期構造から離れた領域は探索することが困難であることがあげられる。一方、（B）の手法では、ある程度この問題は解決されており、常温の MD では運動エネルギーが低くすぎて越えられないようなエネルギー障壁を、人為的な工夫によって乗り越えることができる為、ずっと広い空間を高速に探索できる。問題点としては、統計力学的に正しい調和を作らないので、自由エネルギーやエントロピーという物理量を正しく翻訳できないことがあげられる。（C）の手法では、(B)の問題点を、分配関数を計算して系の自由エネルギーやエントロピーを算出し、実験値と対応させることで解決できる。しかし、完全に全てを探索するには、相手の数が天文学的な数となり、一般には不可能である。エネルギー依拠型モデリング法の精度は、計算に使用される基底関数系（basis set）と計算方法に依存している。基底関数系の選択が悪いと、幾ら長い時間をかけて計算を行っても無意味な結果しか得られない。どの基底系を使用するかの判断は、計算の目的や研究される分子の性質により異なってくる。一般的な基底関数として、共有結合エネルギー、水素結合エネルギー、静電相互作用およびファンデルワールス相互作用があげられる。

自己無撞着静電場法[19]では、静電相互作用が蛋白質の三次構造を決定していると考える。この方法では、各アミノ酸は静電エネルギー的に安定でなければならない、すなわち、各残基の双極子モーメントは、蛋白質全体がつくり出す静電場の中でエネルギー的に最適な配向をとらなければならない。初期近似として最初に静電エネルギーだけが計算される。そして、各アミノ酸の双極子モーメントの配向を変化させ、その静電エネルギーが最適な状態に到達するまで調整を繰り返す。蛋白質全体のエネルギーは、次にすべてのエネルギー項を含めた条件で極小化され、この操作は自己無撞着が達成されるまで繰り返される。この方法は、小さな蛋白質の構造では良好であるが、大きな蛋白質の構造では組合せの数が指数関数的に増加し、実用的な計算時間にならないという問題がある。

2.4　蛋白質モデルの吟味

蛋白質モデルの吟味には、立体化学的な精度、充填特性、折りたたみ構造の信頼性が必要である。構築されたモデルの立体化学的な性質は、結合長、結合角、ねじれ角といったパラメータの精度やアミノ酸キラリティーに基づいて評価される。モデル構造の立体化学的な質を判定する上で重要な指標の１つは、主鎖ねじれ角φとψの分布である。また、蛋白質における疎水性残基と親水性残基の分布は蛋白質モデルの信頼性を判断する目安となる。

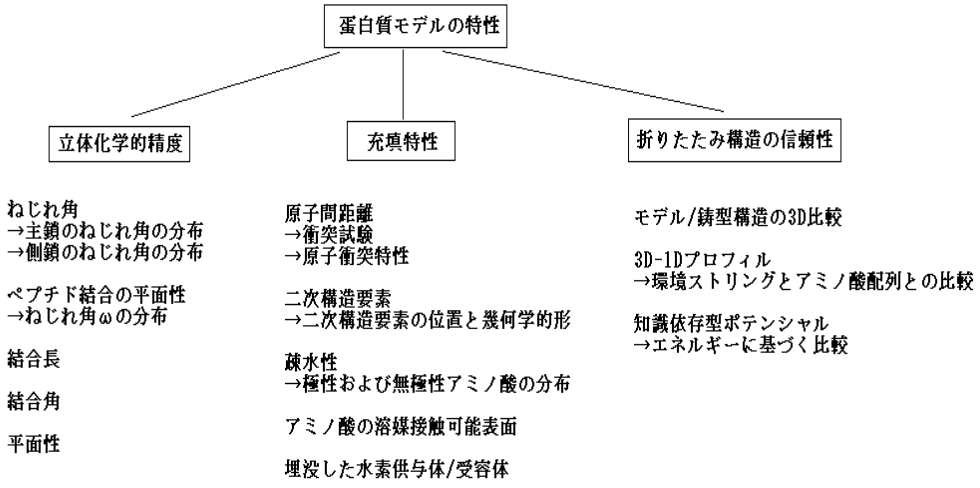


図 2.14　蛋白質モデルの評価

第 3 章

配列-構造相関提示システム

3.1　本研究で用いた蛋白質のデータ

ゲノムネット(http://www.jaist.genome.ad.jp/)で扱うデータベースは、エントリとよぶ単位が集まったテキストファイル(フラットファイル)として扱われ、各エントリにはエントリ名(またはアクセッション番号)というデータベース内でユニークな名前がつけられている。その中で、蛋白質の立体構造に関連するデータベースとして、PDB と PDBSTR[20]がある。X 線結晶解析や NMR などの実験により決定された蛋白質の立体構造データを集めたものが PDB である。蛋白質の多くは複数のアミノ酸配列から構成されているため、そのような蛋白質を登録する場合は、複数のアミノ酸配列を 1 つのエントリ内に記述する。これは配列解析を行うときの障害となるので、アミノ酸配列毎に立体構造情報をデータベース化した PDBSTR が別途用意されている。複数ある各アミノ酸配列をアミノ酸鎖、もしくは単に鎖とよぶ。PDB に蛋白質が登録されると同時に蛋白質をアミノ酸鎖毎に分割した PDBSTR エントリが作られる。PDB と PDBSTR の関係は 1 対多の関係にあり、5 本のアミノ酸鎖から構成される 1 つの PDB ファイルに対して、5 つの PDBSTR ファイルが対応している。PDB ファイルのエントリ名は、4 文字で表記し、PDBSTR ファイルのエントリ名は、対応する PDB ファイルのエントリ名(4 文字)とアミノ酸鎖名(1 文字)の合計 5 文字となる。

```
HEADER    HYDROLASE (ACID PROTEINASE)             04-MAY-95   1FIV       1FIV   2
COMPND    MOLECULE: FELINE IMMUNODEFICIENCY VIRUS PROTEASE;              1FIV   3
COMPND   2 EC: 3.4.23.16                                                 1FIV   4
SOURCE    ORGANISM_SCIENTIFIC: FELIS CATUS;                              1FIV   5
SOURCE   2 ORGANISM_COMMON: CAT;                                         1FIV   6
SOURCE   3 EXPRESSION_SYSTEM: ESCHERICHIA COLI                           1FIV   7
EXPDTA    X-RAY DIFFRACTION                                              1FIV   8
AUTHOR    A.WLODAWER,A.GUSTCHINA,L.RESHETNIKOVA,J.LUBKOWSKI,A.ZDANOV     1FIV   9
REVDAT   2   10-JUN-99 1FIV    1       JRNL                             1FIV A10
REVDAT   1   31-JUL-95 1FIV    0                                        1FIV  10
SEQRES   1 A  113  VAL GLY THR THR THR THR LEU GLU LYS ARG PRO GLU ILE  1FIV  94
SEQRES   2 A  113  LEU ILE PHE VAL ASN GLY TYR PRO ILE LYS PHE LEU LEU  1FIV  95
SEQRES   3 A  113  ASP THR GLY ALA ASP ILE THR ILE LEU ASN ARG ARG ASP  1FIV  96
SEQRES   4 A  113  PHE GLN VAL LYS ASN SER ILE GLU ASN GLY ARG GLN ASN  1FIV  97
SEQRES   5 A  113  MET ILE GLY VAL GLY GLY GLY LYS ARG GLY THR ASN TYR  1FIV  98
SEQRES   6 A  113  ILE ASN VAL HIS LEU GLU ILE ARG ASP GLU ASN TYR LYS  1FIV  99
SEQRES   7 A  113  THR GLN CYS ILE PHE GLY ASN VAL CYS VAL LEU GLU ASP  1FIV 100
SEQRES   8 A  113  ASN SER LEU ILE GLN PRO LEU LEU GLY ARG ASP ASN MET  1FIV 101
SEQRES   9 A  113  ILE LYS PHE ASN ILE ARG LEU VAL MET                  1FIV 102
SEQRES   1 B    7  ACE ALA VAL STA GLU ALA NH2                         1FIV 103
REMARK   3                                                           1FIV  21
HELIX    1   1 ARG A   40  ASP A   42  5                               1FIV 112
HELIX    2   2 ARG A  104  LYS A  109  1                               1FIV 113
SHEET    1   A 4 TYR A  23  LEU A  28  0                               1FIV 114
SHEET    2   A 4 GLU A  15  VAL A  20 -1  N  VAL A  20   O  TYR A  23   1FIV 115
SHEET    3   A 4 VAL A  71  ILE A  75 -1  N  GLU A  74   O  PHE A  19   1FIV 116
ATOM      6  CG1 VAL A   4      -4.863   4.974  23.099  1.00 62.74      1FIV 134
ATOM      7  CG2 VAL A   4      -4.545   2.777  21.972  1.00 64.09      1FIV 135
ATOM      8  N   GLY A   5      -2.648   5.919  21.005  1.00 50.85      1FIV 136
ATOM      9  CA  GLY A   5      -2.678   6.581  19.720  1.00 47.60      1FIV 137
```

```
MEMBER      1FIVA         113    PROTEIN      1FIV  95/05/04   95/07/31
DEFINITION  MOLECULE: FELINE IMMUNODEFICIENCY VIRUS PROTEASE;
            EC: 3.4.23.16
SOURCE      ORGANISM_SCIENTIFIC: FELIS CATUS;
            ORGANISM_COMMON: CAT;
            EXPRESSION_SYSTEM: ESCHERICHIA COLI
SEGMENT      1 of  2  ( A of A,B )
DEPOSITOR   A.WLODAWER,A.GUSTCHINA,L.RESHETNIKOVA,J.LUBKOWSKI,A.ZDANOV
RESOLUTION  RANGE      10.0 - 2.0  ANGSTROMS
FEATURES        FROM       TO       DESCRIPTION
  HELIX    5     40       42    1
  HELIX    1    104      109    2
  SHEET    0     23       28    A
  SHEET   -1     15       20    A  N  VAL A  20   O  TYR A  23
SEQUENCE
  VGTTTTLEKR PEILIFVNGY PIKFLLDTGA DITILNRRDF QVKNSIENGR QNMIGVGGGK
  RGTNYINVHL EIRDENYKTQ CIFGNVCVLE DNSLIQPLLG RDNMIKFNIR LVM
STRUCTURE        X-CA    Y-CA    Z-CA   PHI    PSI    X-CB    Y-CB    Z-CB
    4 VAL   4    -2.51    4.17      22.67 -999 -27   -3.96    3.74   23.01    117
    5 GLY   5    -2.68    6.58      19.72  -37 -25   -2.43    8.08   19.90     57
    6 THR   6     0.37    4.77      18.31 -148 122    1.61    5.58   17.96     83
    7 THR   7     1.35    1.15      18.93 -110 121    0.31    0.03   18.80 S  128
    8 THR   8     4.84    0.29      17.66 -114 115    5.92    1.37   17.85 S  134
    9 THR   9     5.87   -3.34      17.19  -80 159    5.19   -4.38   16.30 S  137
   10 LEU  10     9.57   -4.32      17.43 -100  26   10.02   -4.73   18.89    156
    5 GLY   5    -2.68    6.58      19.72  -37 -25   -2.43    8.08   19.90     57
    6 THR   6     0.37    4.77      18.31 -148 122    1.61    5.58   17.96     83
    7 THR   7     1.35    1.15      18.93 -110 121    0.31    0.03   18.80 S  128
    8 THR   8     4.84    0.29      17.66 -114 115    5.92    1.37   17.85 S  134
    9 THR   9     5.87   -3.34      17.19  -80 159    5.19   -4.38   16.30 S  137
   10 LEU  10     9.57   -4.32      17.43 -100  26   10.02   -4.73   18.89    156
```

図 3.1　PDB(1FIV:上図)と PDBSTR(1FIVA:下図)

PDB エントリでは実験手法などの記述に加えて、蛋白質を構成する原子単位の三次元座標情報が大部分を占めている。一方、PDBSTR エントリには、アミノ酸単位の三次元座標情報が記述されている。一般に、PDB エントリの結晶構造の分解能は 2.5Åと 1.5Åの間か、それよりも良くなければ、精密な立体構造データとして構造予測に用いないことが多い。

PDB エントリ内に記述されている二次構造データは、登録者(実験者)の主観的な判断に委ねられ、エントリ毎に異なった規準で決定した二次構造が表記されている。この曖昧性を排除する為に、DSSP(Dictionary Secondary Structure of Protein)プログラム[6]が二次構造の特定に用いられることが多い。このプログラムは、個々の PDB エントリの原子座標から一定の規準に従い正規化した各アミノ酸の構造データ(DSSP ファイル)を作成する。DSSP ファイルは PDB エントリと 1 対 1 で対応し、エントリ名は 4 文字で表記されている。DSSP ファイルには、各アミノ酸に対する、座標情報、二次構造情報、内外位置情報、二面角情報などが記述されている。

3.2 配列類似度検索

3.2.1 検索プログラムの種類

配列類似度を比較する代表的なプログラムとして BLAST（Basic Local Alignment Search Tool）と FASTA(Sequence Similarity Search)がある。これらのプログラムはホモロジー検索プログラムとよばれる。BLAST は NCBI(National Center for Biotechnology Information)で開発されたプログラムで、ギャップを入れない部分配列の HSP(高スコア断片)を複数集めて評価する方式をとっており、高速に検索できる。FASTA は W.Pearson によって開発されたプログラムで、はじめに文字の良く一致する領域を高速に検索し、最終的にはギャップを入れた完全なアライメントを行う方式をとっており、高い精度で検索できる。一般にホモロジー検索をするときには、はじめに BLAST を用い満足する結果が得られないときに FASTA を用いる。また、ホモロジー検索にかける配列を、問い合わせ配列(query sequence)とよぶ。それぞれのプログラムには、WWW と電子メールとコマンドの 3 種類の使用法がある。

- WWW: 探索結果から、リンクをたどってさらに情報を収集することが可能であるが、時間のかかる検索には向かない。
- 電子メール: 検索要求をまとめて送信できるので、時間のかかる計算をするのに向いている。
- コマンド: 自分専用のライブラリ(後述)に対して検索したり、検索結果を加工するなど高度な使い方ができる。

3.2.2 BLAST

BLAST のオプションとしては、H＝ヒストグラム表示数、V＝スコア表示数、B＝アライメント表示数、E＝ランダムなヒットが起こる配列数の期待値などがある。V や B は類似配列が多数ヒットし、出力数が多くなりすぎるときに数を制限する為の閾値パラメータであり、出力数が少ないときには働かない。また、これらの閾値パラメータは、検索の速度と正確さの兼ね合いを決める働きをもっている。通常 BLAST では、平均的にうまくいくような値がデフォルトで設定されているが、ユーザの要求により、出力個数を増やしたい(類似性が低いものまで見たい)場合は E を上げ、逆に類似度が高い配列だけを検索するには E を下げることができる。その他に Matrix とよばれるオプションが選択できる。これはアミノ酸間の置換のし易さを 20×20 の行列で表したもので、アミノ酸置換行列(Matrix)とよばれる。代表的なものには、PAM や BLOSUM とよばれるものがある。Matrix を用いることで、蛋白質配列の類似性を評価する際に、単なる文字の一致ではなく、アミノ酸の性質に基づいた評価が行える。

Ala	2																			
Arg	-2	6																		
Asn	0	0	2																	
Asp	0	-1	2	4																
Cys	-2	-4	-4	-5	12															
Gln	0	1	1	2	-5	4														
Glu	0	-1	1	3	-5	2	4													
Gly	1	-3	0	1	-3	-1	0	5												
His	-1	2	2	1	-3	3	1	-2	6											
Ile	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
Leu	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
Lys	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
Met	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
Phe	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
Pro	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
Ser	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
Thr	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
Trp	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Tyr	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
Val	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

図 3.2　PAM250　Matrix

BLAST では自分専用配列データベース(ライブラリ)を作成し、これを探索空間としてホモロジー検索を行うことができる。その場合に必要な BLAST 用インデックスファイルは、FASTA 形式のファイル(libraryfile)を setdb コマンドに与えることにより、libraryfile.atb、libraryfile.bsq、libraryfile.ahd という名前で生成できる。

FASTA 形式(または Pearson 形式)

```
>名称およびコメント
配列

例
>STR:1FIVA MOLECULE: FELINE IMMUNODEFICIENCY VIRUS PROTEASE; EC: 3.4.23.16
VGTTTTLEKRPEILIFVNGYPIKFLLDTGADITILNRRDFQVKNSIENGRQNMIGVGGGK
```

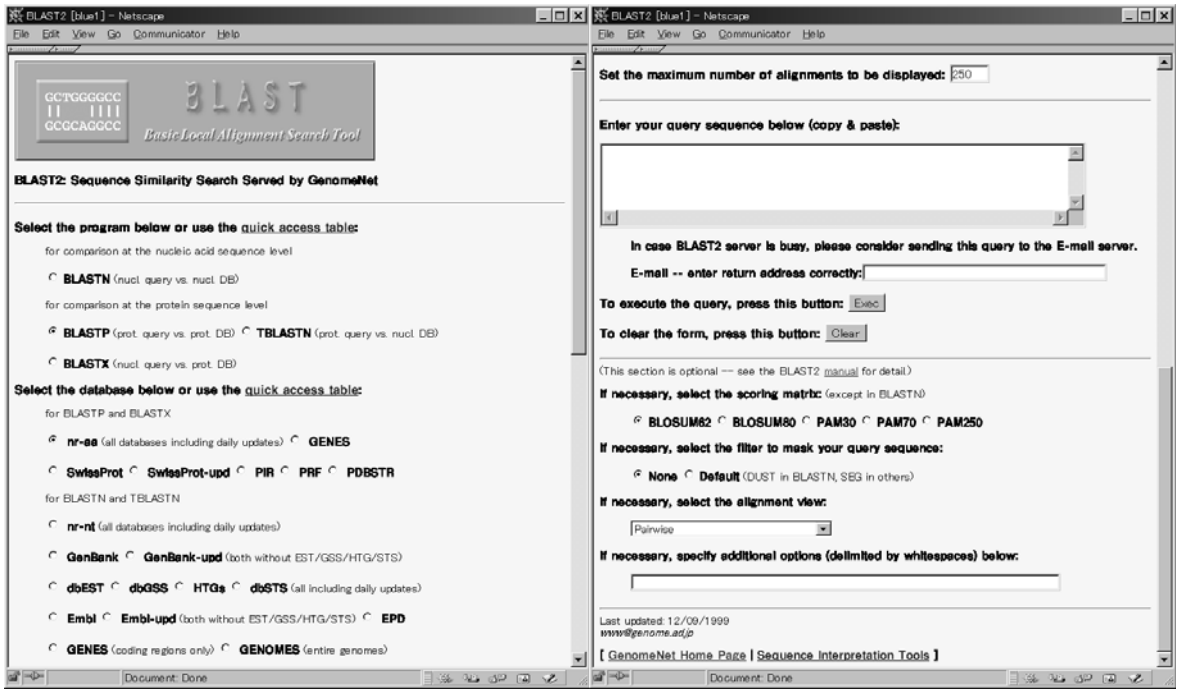


図 3.3　BALST の実行例

3.3 構造の可視化

蛋白質の構造を視覚的に見るために、本研究では東京大学医科学研究所で開発された PDB highlight を用いる。PDB highlight は、蛋白質の類似構造について検索・解析するシステム(PACADE)[21]と連動したシステムで、蛋白質のアミノ酸配列や二次構造や立体構造などを、それぞれのレベルで可視化表示することができる。また、単に表示するだけでなく、外部のデータベースを参照したり解析サービスを呼び出したりすることも可能である。PDB highlight は Chemscape Chime というプラグインを立体構造の可視化に用いている。PACADE は PDB のデータを独自に解析し、蛋白質のアミノ酸をヘリックス、ストランド、その他の 3 種類の二次構造で分類している。

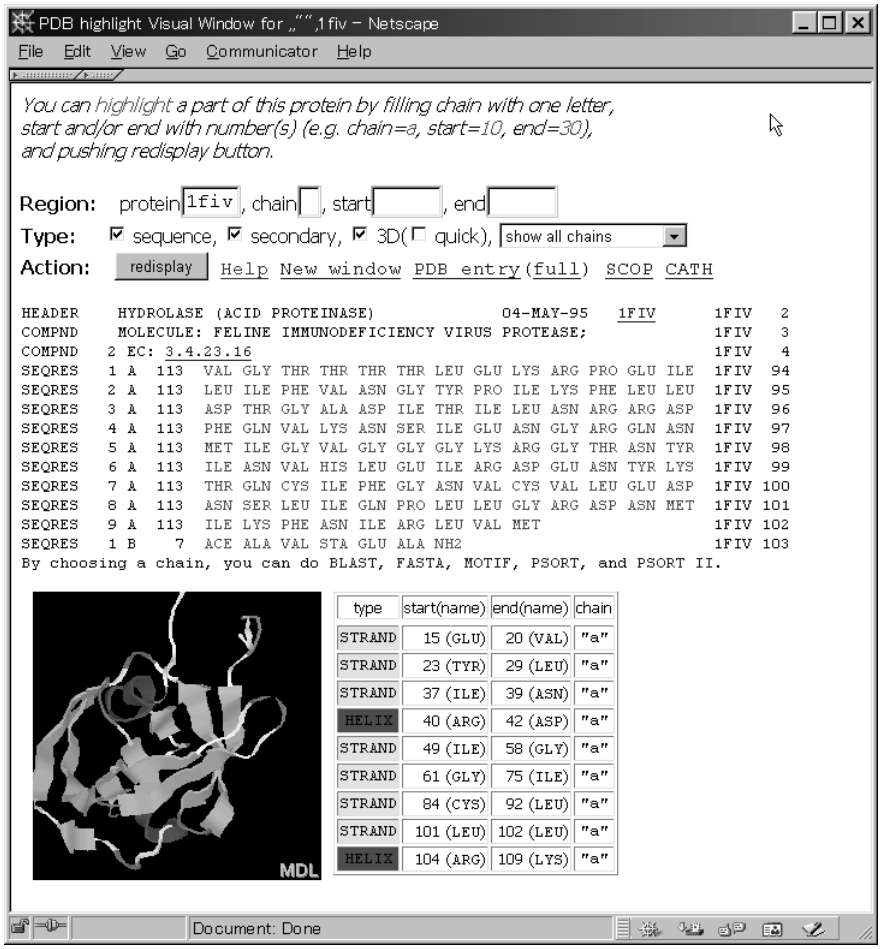


図 3.4　PDB highlight

3.4　1FIVA の類似部分配列の構造分布

はじめに、予備実験として猫科の免疫不全ウイルスの蛋白質分解酵素である 1FIV(PDB エントリ名)の A 鎖である 1FIVA(PDBSTR エントリ名)について、類似部分配列がどれだけ検索されるか以下の基準で調べた。

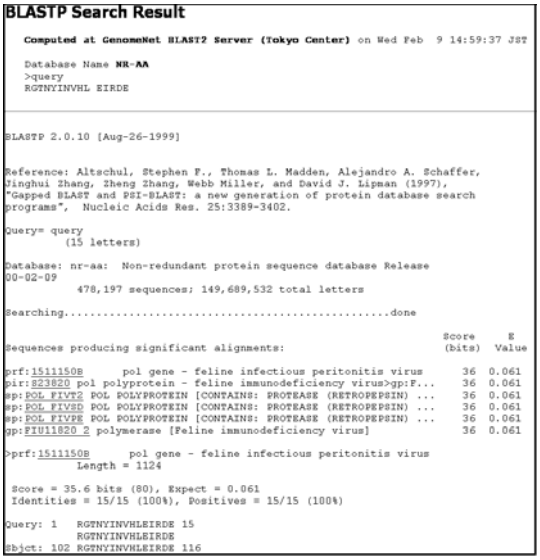
- ホモロジー検索プログラム(www 上の BALST と FASTA)
- 閾値の設定(デフォルト)
- フラグメント長(6 残基～13 残基)
- 検索上限(20 位)
- ライブラリ(PDBSTR 全体)

二次構造	残基番号	残基数	ターン位置	アミノ酸情報	前 後 2 アミノ酸
Random	1-14	14	10-12	VGTTTTLEKRP EIL	IF
Strand	15-20	6		IFVNGY	IL PI
Random	21-22	2		PI	GY KF
Strand	23-29	7		KFLLDTG	IK AD
Random	30-36	7	31-32,34-35	ADITILN	TG RR
Strand	37-39	3		RRD	LN FQ
Helix	40-42	3		FQV	RD KN
Random	43-48	6	46-67	KNSIEN	QV GR
Strand	49-58	10		GRQNMIGVGG	EN GK
Random	59-60	2		GK	GG RG
Strand	61-75	15		RGTNYINVHLEI RDE	GK NY
Random	76-83	8	78-80	NYKTQCIF	DE GN
Strand	84-92	9		GNVCVLEDN	IF SL
Random	93-100	11	94,98-99	SLIQPLLG	DN RD
Strand	101-102	2		RD	LG NM
Random	103	1		N	RD MI
Helix	104-109	6		MIKFNI	DN RL
Random	110-113	4	110	RLVM	NI

表 3.1　1FIVA の二次構造データ

BLAST では、検索結果の上位に類似配列が表示されたが、FASTA では、部分配列に対応できないのか検索結果に類似配列が表示されなかった。BLAST の検索結果は、同一鎖とみられる配列が多く検索された。このことは、PDB に偏りが存在することを反映している。例えば、これまでに立体構造が判明した lysozyme という蛋白質は、配列上の類似性が 95%以上であるものについての PDB 登録件数は 200 を超えている。これらの立体構造が本質的には同じものであり、データ同士の関係は一次配列のみに基づいて識別できる。ある蛋白質の立体構造が実験により解析されると、それと類縁関係を持つ蛋白質が同じ手法で解析できるようになることが多いので、ほとんど同一

の配列･構造を持つ蛋白質が PDB に数多く登録されている。図 3.6 は、各 7 残基のフラグメントに対してホモロジー検索をした結果から、同一鎖と見られるアミノ酸配列を取り除き、残りのアミノ酸配列の類似部分から導き出した各二次構造の分布を示している。これにより、正しい二次構造が大多数を占めることが確認された。(図 3.6 では、最下列の二次構造は DSSP から得た正しい二次構造で、山なりの形状の二次構造は類似部分配列から得られた二次構造を表している。)



BLASTP Search Result

Computed at GenomeNet BLAST2 Server (Tokyo Center) on Wed Feb 9 14:59:37 JST

Database Name **NR-AA**
>query
RGTNYINVHL EIRDE

BLASTP 2.0.10 [Aug-26-1999]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= query
(15 letters)

Database: nr-aa: Non-redundant protein sequence database Release 00-02-09
478,197 sequences; 149,689,532 total letters

Searching.....done

Sequences producing significant alignments:	Score (bits)	E Value
prf:1511150B pol gene - feline infectious peritonitis virus	36	0.061
pir:S23820 pol polyprotein - feline immunodeficiency virus>gp:F...	36	0.061
sp:POL_FIVT2 POL POLYPROTEIN [CONTAINS: PROTEASE (RETROPEPSIN) ...	36	0.061
sp:POL_FIVSD POL POLYPROTEIN [CONTAINS: PROTEASE (RETROPEPSIN) ...	36	0.061
sp:POL_FIVPE POL POLYPROTEIN [CONTAINS: PROTEASE (RETROPEPSIN) ...	36	0.061
gp:FIU11820_2 polymerase [Feline immunodeficiency virus]	36	0.061

>prf:1511150B pol gene - feline infectious peritonitis virus
Length = 1124

Score = 35.6 bits (80), Expect = 0.061
Identities = 15/15 (100%), Positives = 15/15 (100%)

Query: 1 RGTNYINVHLEIRDE 15
RGTNYINVHLEIRDE
Sbjct: 102 RGTNYINVHLEIRDE 116

図 3.5　BLAST の実行結果

図 3.6　1FIVA の類似部分配列の二次構造分布

(図 3.6: 略称として α ヘリックス=h, H、β ストランド=s, S、コイル=c, C と表記する。)

3.5　二次構造予測法との比較

さらに予備実験として二次構造予測法(SSThread Program)と比較した。構造未知の蛋白質(大腸菌の複製開始蛋白)について 7 残基毎のフラグメントに分け、WWW 上の BLAST をデフォルトの値を用いて上位 10 位までの類似配列を検索した。次に、PDB highlight を用いて類似部分配列の二次構造を特定し、多数決でアミノ酸配列と二次構造を対応付けた。構造未知の配列に対して既存の二次構造予測法と比較することで、どのような違いが出るのか調べた。一般に、これまでの二次構造予測法は、α ヘリックスの構造を高い精度で予測することができるといわれている。実際に比較した結果では、α ヘリックスに関して一部で顕著な違いが見られた。

構造未知配列(大腸菌の複製開始蛋白)

>sp:REPY_ECOLI [P22308] REPLICATION INITIATION PROTEIN.
MSELVVFKANELAISRYDLTEHETKLILCCVALLNPTIENPTRKERTVSFTYNQYAQMMNISRE
NAYGVLAKATRELMTRTVEIRNPLVKGFEIFQWTNYAKFSSEKLELVFSEEILPYLFQLKKFIKY
NLEHVKSFENKYSMRIYEWLLKELTQKKTHKANIEISLDEFKFMLMLENNYHEFKRLNQWVL
KPISKDLNTYSNMKLVVDKRGRPTDTLIFQVELDRQMDLVTELENNQIKMNGDKIPTTITSDSY
LHNGLRKTLHDALTAKIQLTSFEAKFLSDMQSKYDLNGSFSWLTQKQRTTLENILAKYGRI

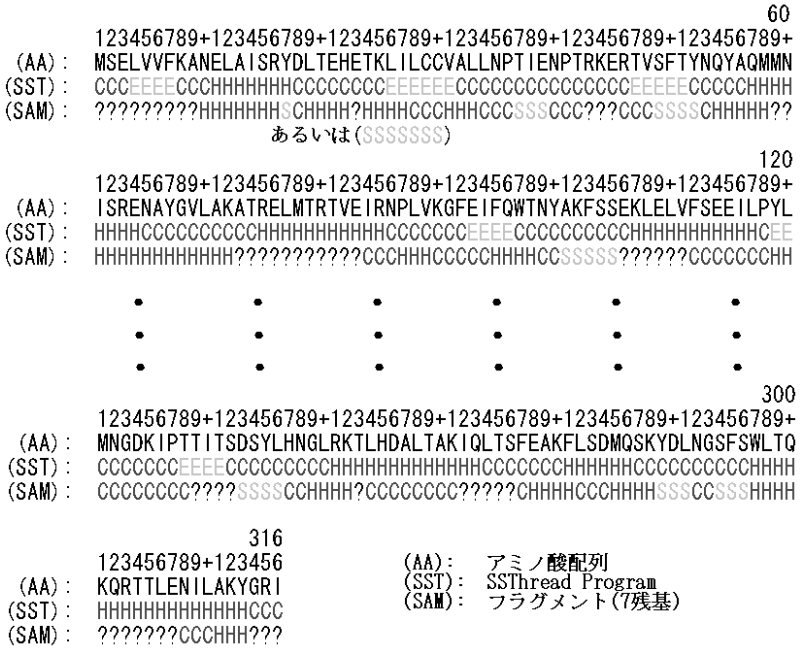


図 3.7　二次構造予測比較

（図 3.7: SSThread Program では、β ストランドを E として表記する。）

3.6　配列-構造相関提示システムの構築

以上の予備実験を経て、部分配列の類似性に基づいて配列-構造相関を示すシステムを下図の流れで構築した。

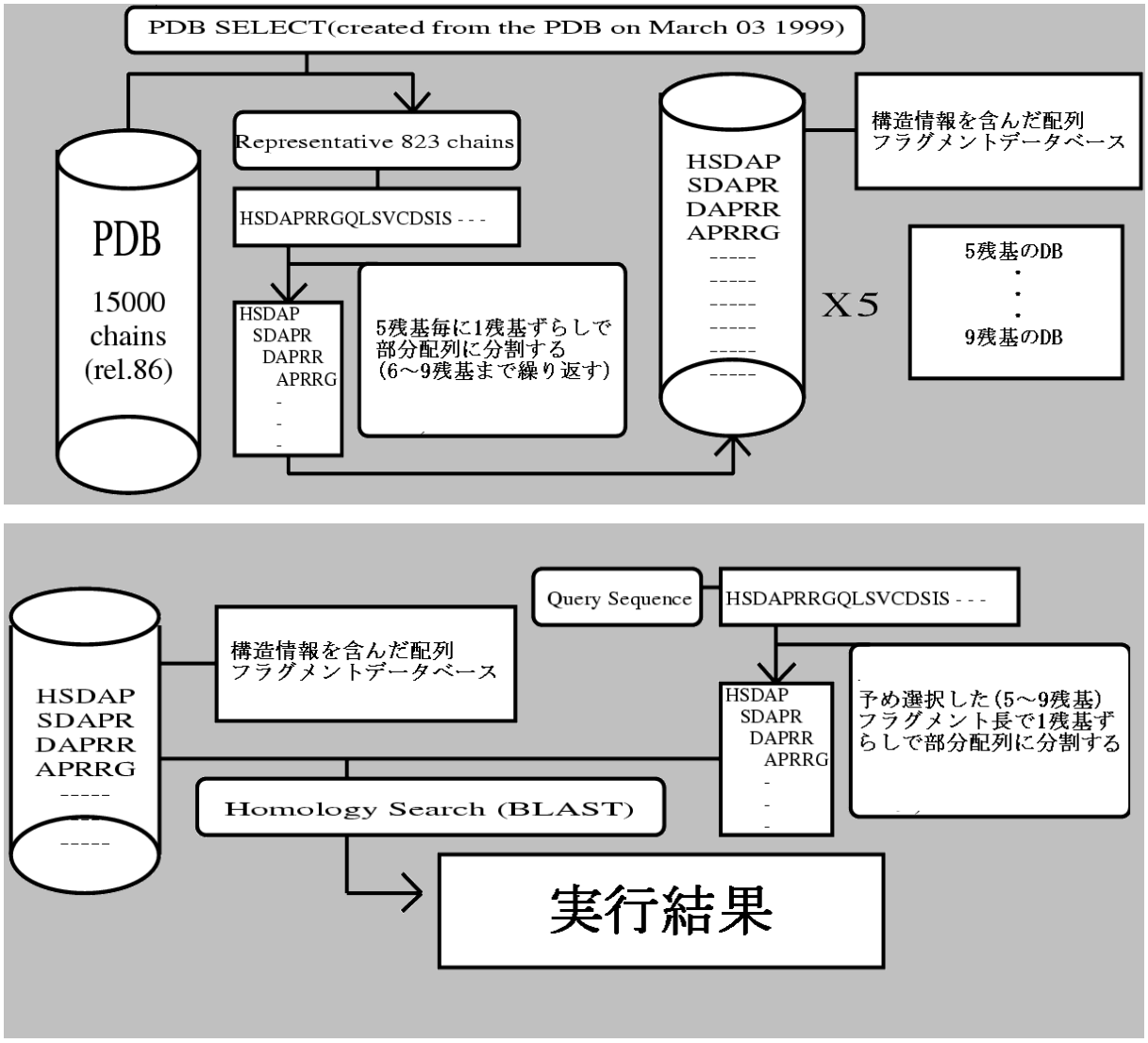


図 3.8　システム構成図

- このシステムでは、部分配列に対する類似配列を検索(ホモロジー検索)する必要がある。ホモロジー検索のプログラムには BLAST と FASTA があるが、WWW 上で 1FIVA の二次構造分布について実験した結果より、FASTA は短い部分配列を検索するには向かないことが判明しているので、BLAST を用いた。

- 配列に対応する構造情報を得る為に PDB のエントリを用いた。部分配列に対する類似部分配列を BLAST で検索する為には、BLAST 用の部分配列ライブラリを作成する必要がある。これまでに PDB のリリース番号 86 に登録された蛋白質のアミノ酸配列は 15500 鎖あり、すべてアミノ酸配列を部分配列毎に分割して、オリジナルのライブラリの FASTA 形式ファイルに変換すると 300 万行以上のファイルになってしまう。しかし、BLAST 用のオリジナルのライブラリを作成するには限界があり、60 万行以上の FASTA 形式ファイルを登録することはできない。この問題を解決する為には、PDB の 15500 鎖を何らかの形で要約する必要がある。ここでは PDB_SELECT データベースを用いることでこの問題を解決した。PDB_SELECT データベース[23,24]では、PDB の中で配列と構造が類似している蛋白質をひとまとめにした。代表蛋白質で PDB 全体を表現している。さらに、配列と構造が同一とみられるエントリ名の異なる複数の蛋白質がもたらすデータの偏りを取り除くことができた。PDB_SELECT データベースにより、15500 鎖を 888 鎖まで要約することができた。
- PDB には、アミノ酸配列の番号と対応する二次構造の位置が異なったり、アミノ酸が所々抜けたりするファイルが存在する。また、ファイル毎で各アミノ酸に対する二次構造の区別が曖昧になっているので、整合性を図る為に二次構造情報を原子座標から正しく特定したデータを用いることにより曖昧性を取り除いた。各アミノ酸に対する二次構造情報の同定には、DSSP ファイルのデータと PACADE を基に作成したデータの両方に適合したデータを用いた。これにより各アミノ酸毎に正しく構造情報を確定できたアミノ酸配列は 823 鎖になった。
- アミノ酸配列と二次構造を適合させた 823 鎖に対して、5 残基毎の部分配列に 1 残基ずらしで分割した部分配列データ集合を作成した。また同様に 6～9 残基で分割した部分配列データ集合も作成した。作成した部分配列データ集合を FASTA 形式に変換して BLAST 用の部分配列ライブラリを作成した。
- 問い合わせ配列(query sequence)に対し、ライブラリ作成時と同様に、予め選択したフラグメント長の部分配列に分割して、同じ長さの部分配列ライブラリと比較したホモロジー検索を行う。

3.7　システムの実行例

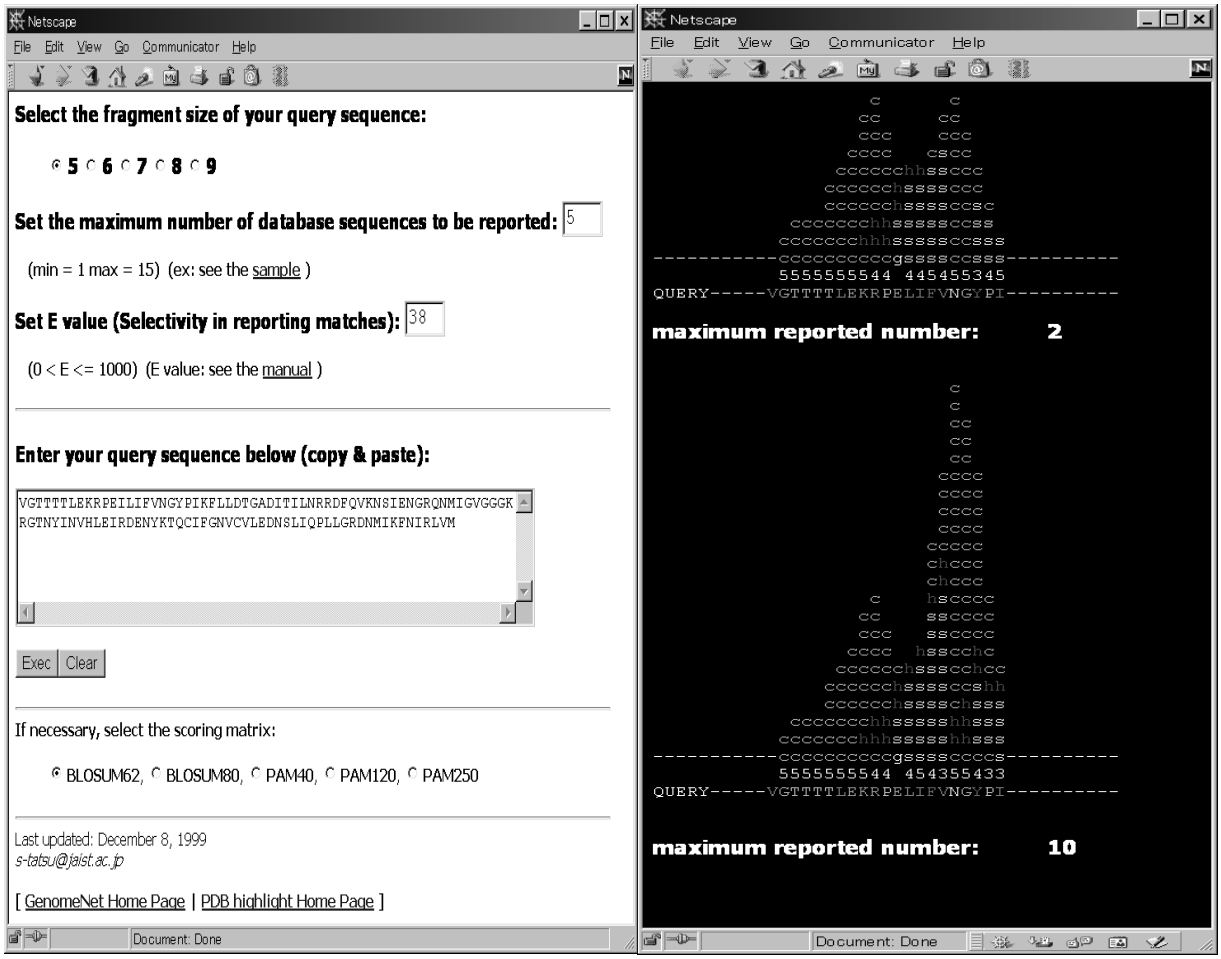


図 3.9　システムの実行例

- Select the fragment size of …

  部分配列の長さの指定。

- Set the maximum number of …

  採用する検索結果数。数を少なくすると右図上のように山の高さが低くなる。数を多くすると右図下のように山の高さが高くなる。検索結果の上位から指定した数の類似部分配列を基に構造情報を構成する。

- Set E value　…

  ランダムなヒットが起こる配列数の期待値。数を少なくすると類似度が高い配列だけを検索する。逆に、数を多くすると検索条件が緩くなる(出力個数が増える)。

- Enter your query sequence　…

  問い合わせ配列の入力フォーム。

- If necessary, select　…

  Matrix の選択。

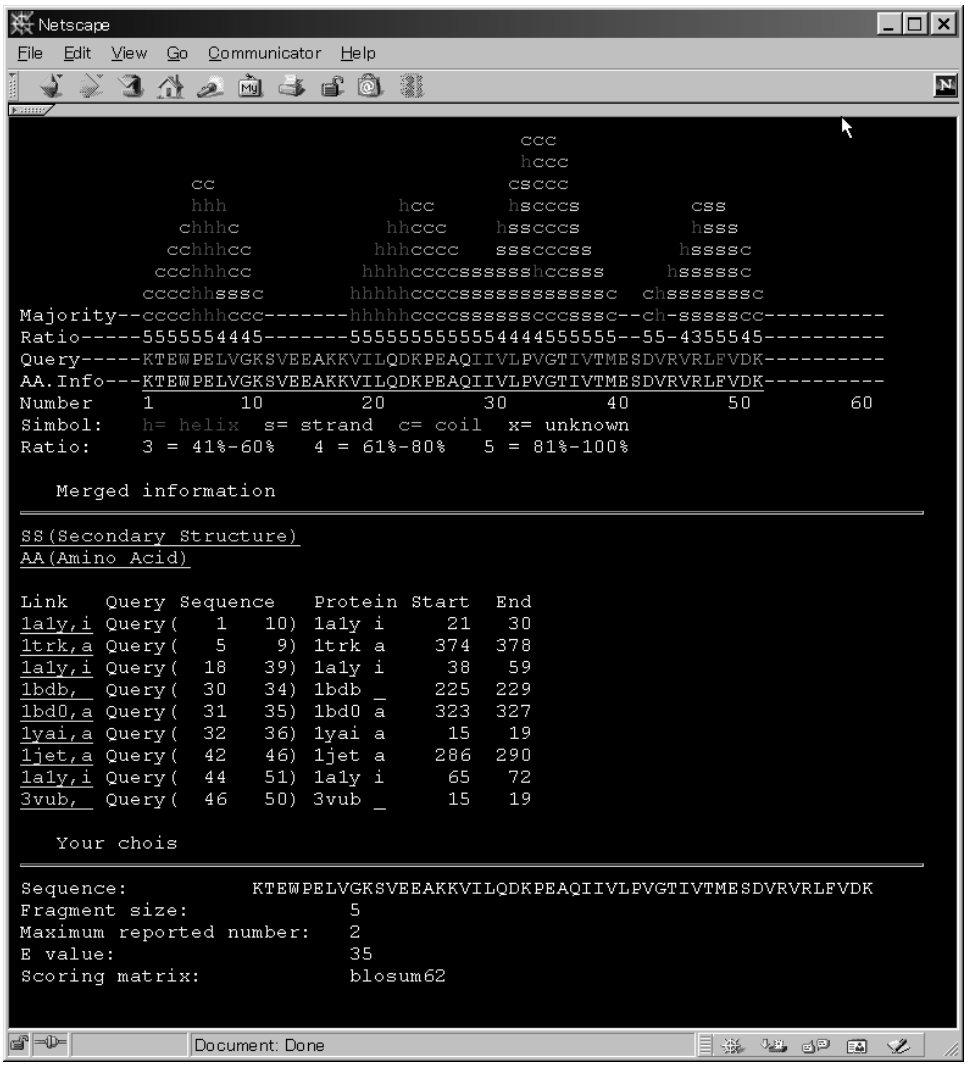


図 3.10　システムの実行結果 1

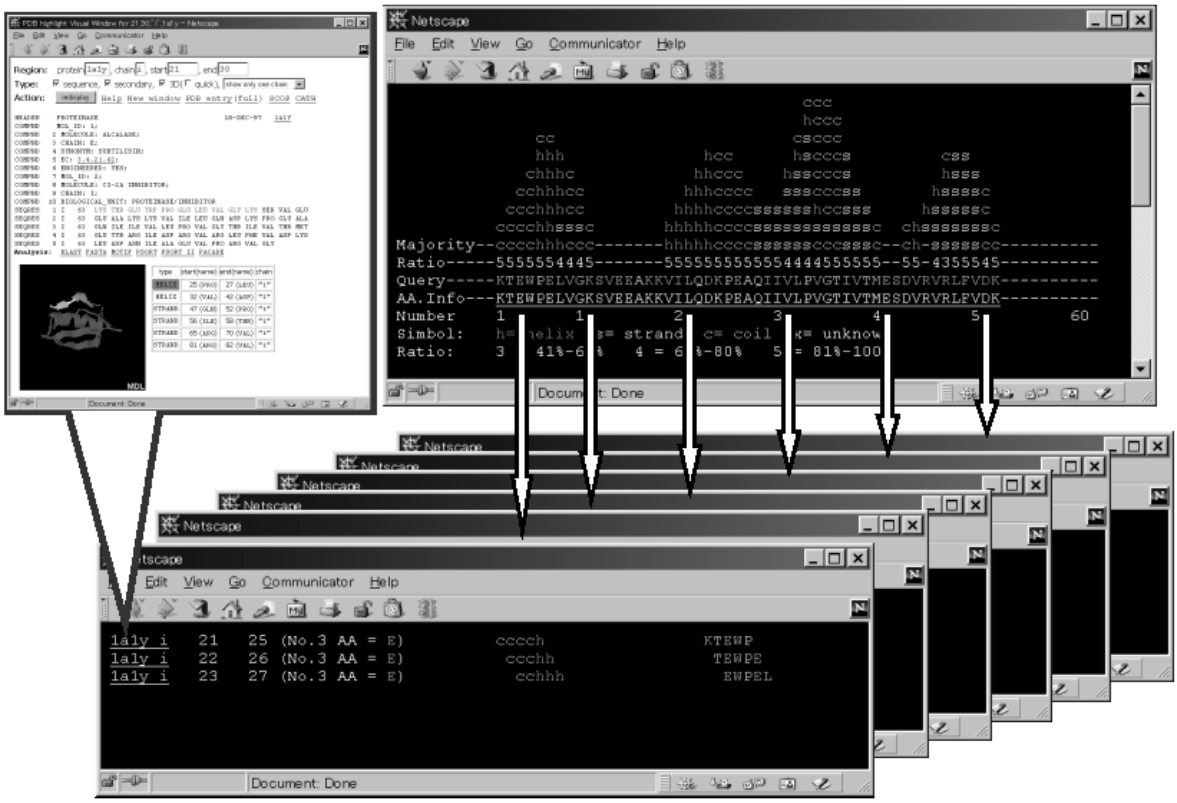


図 3.11　システムの実行結果 2

- Majority(二次構造)

  類似部分配列の二次構造情報を基に、問い合わせ配列の各アミノ酸毎に二次構造を決定した結果。

- Ratio(数字)

  類似部分配列の二次構造の中で多数決で決定した二次構造の占める割合。

- AAInfo(アミノ酸)

  問い合わせ配列の各アミノ酸毎に検索された類似部分配列を記述したファイルにリンクしている。リンクされたファイルには、部分配列の PDBSTR 名、配列の位置番号、アミノ酸配列および二次構造情報が記入されている。また、部分配列の PDBSTR 名から PDB highlight にリンクしており、部分配列の二次構造や立体構造などを、それぞれのレベルでみることができる。

- Number(数字)

  問い合わせ配列のアミノ酸番号。

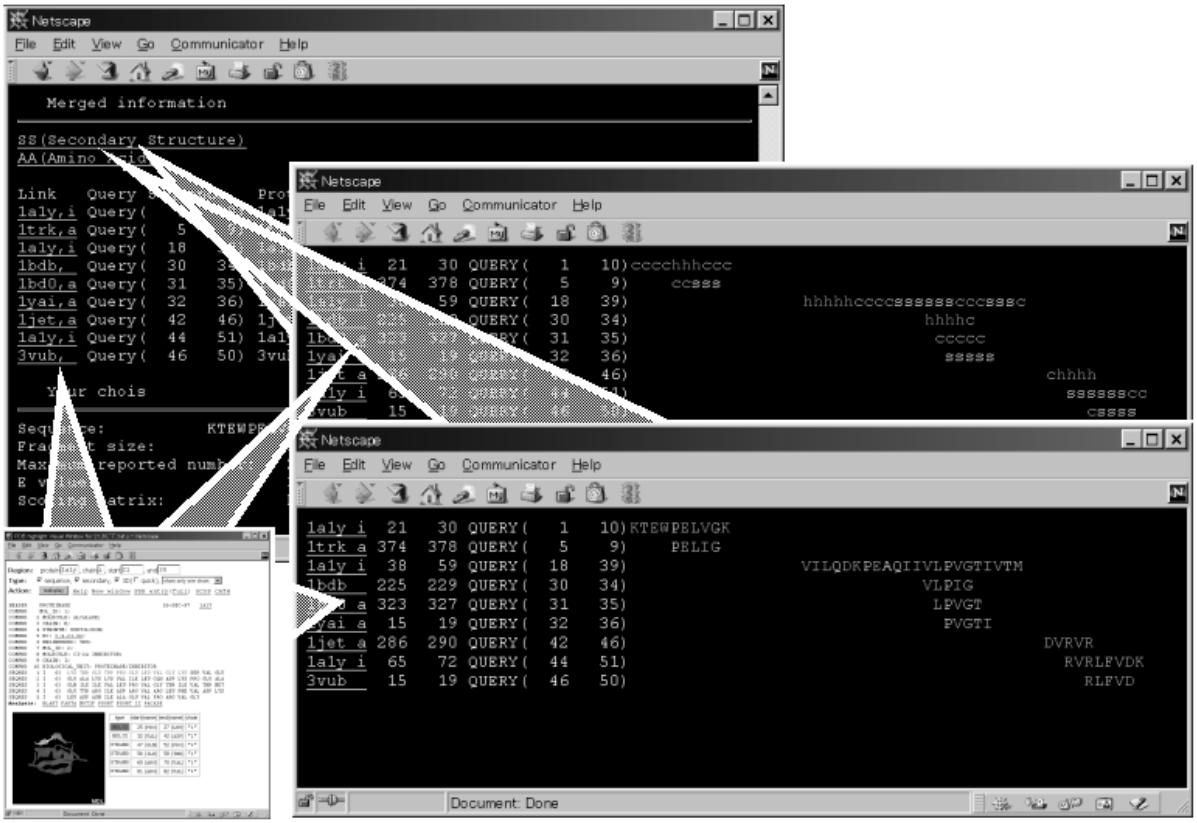


図 3.12　システムの実行結果 3

このシステムの部分配列ライブラリでは、アミノ酸配列が 1 残基ずらしで各フラグメント長に分割されている。これにより、類似部分配列の検索結果の上位に同一アミノ酸鎖の 1 残基だけずれた部分配列もヒットしてしまう。これを解決するため、検索された部分配列の中に同一アミノ酸鎖の複数の部分配列が存在したとき、近傍の部分配列をひとまとめにして表示するようにしている(上図左)。上図右では、ひとまとめにした部分配列のアミノ酸配列情報と二次構造情報について、問い合わせ配列とアライメントした状態で表示している。また、ひとまとめにした PDBSTR 名から PDB highlight にリンクしており、二次構造や立体構造などを、それぞれのレベルで見ることができる。

第 4 章

配列-構造提示システムの検証

4.1 検証データ 1

構築したシステムに対し、以下図の条件の DATA SET を用いてどれだけの配列-構造相関を示せるか検証した。ここでは、配列-構造相関の評価の 1 つとして二次構造情報について以下の評価尺度で判断した。

評価尺度 A:　二次構造を示せて一意に特定できる割合(多数決が成立する)。

評価尺度 B:　二次構造を示せるが一意に特定できない割合(多数決が成立しない)。

評価尺度 C:　二次構造を示せない割合(類似配列が存在しない)。

評価尺度 D:　全体(A+B+C)に対する二次構造の正当率。

評価尺度 E:　二次構造が示せ一意に特定できる割合(A)に対する二次構造の正当率。

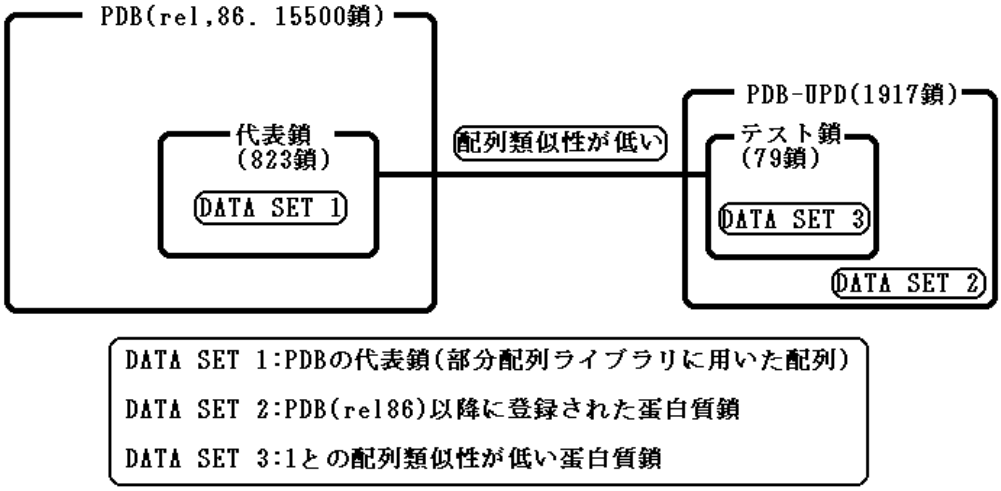


図 4.1　検証データ 1 の DATA SET

4.1.1　フラグメント長変化による二次構造予測精度変化 1

評価尺度 / フラグメント長	A	B	C	D	E
5 残基	93.5%	2.8%	3.7%	90.2%	96.5%
6 残基	97.0%	2.3%	0.7%	94.4%	97.3%
7 残基	99.2%	0.3%	0.5%	94.9%	95.9%
8 残基	99.1%	0.5%	0.4%	98.6%	99.5%
9 残基	99.5%	0.2%	0.3%	99.9%	99.9%

条件:　DATA SET 1、採用した検索結果数=5、E value=38、Matrix=BLOSUM62

表 4.1　DATA SET 1 のフラグメント長変化による二次構造予測精度

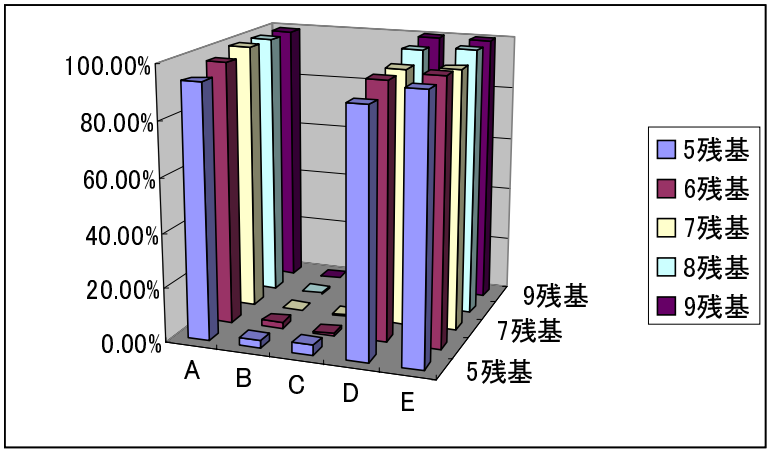


図 4.2　DATA SET 1 のフラグメント長変化による二次構造予測精度

DATA SET 1 では、探索空間である部分配列ライブラリに問い合わせ配列自身のフラグメントが必ず含まれているので、検索結果の上位に正しい二次構造を持った部分配列が得られる。また、フラグメント長が長いほど正しい構造が導き出されることがわかる。

デフォルトの E=10 は長い配列同士の類似配列の検索結果は良好であるが、短い配列に対しては検索ヒット数が少なく向かない。ここで E=38 に設定することで、類似部分配列が検索される量のバランスをとることにした。

4.1.2　配列類似度の違いによる二次構造予測精度変化

評価尺度 / DATA SET	A	B	C	D	E
1	93.5%	2.8%	3.7%	90.2%	96.5%
2	76.0%	8.3%	15.6%	50.6%	66.6%
3	68.1%	10.5%	21.4%	36.8%	54.0%

条件: フラグメント長=5、採用した検索結果数=5、E value=38、Matrix=BLOSUM62

表 4.2　配列類似度の違いによる二次構造予測精度

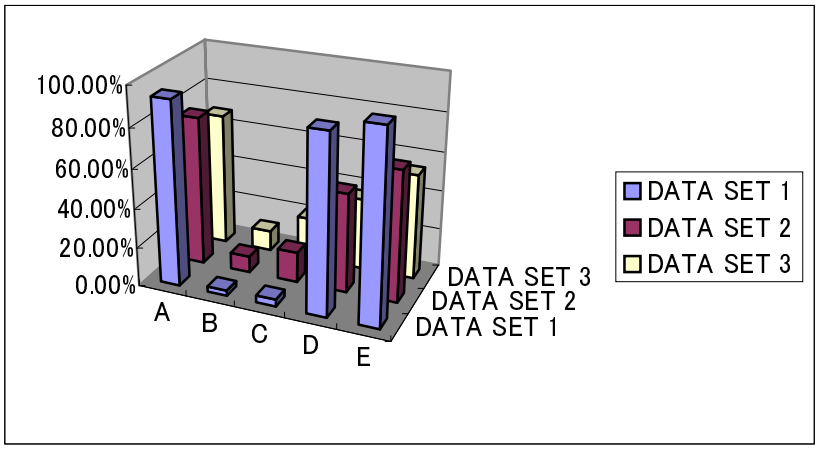


図 4.3　配列類似度の違いによる二次構造予測精度

DATA SET 1 では部分配列ライブラリに問い合わせ配列自身(同一配列)が含まれているので、必ず類似の部分配列が存在する。逆に DATA SET 3 では類似の部分配列が少なくなるよう予め調節している。このことから二次構造情報の正当率は以下のように変化することが予想できる。

DATA SET 1　＞＞ DATA SET 3

実際の結果では、DATA SET 3 のアミノ酸配列に対して類似配列は平均 2 割以上で検索されなかった。また、二次構造の精度も 4 割以下と正しい二次構造を得ることができなかった。配列の類似度に比較して予測精度が下がることがわかる。

4.1.3　E value 変化による二次構造予測精度変化

評価尺度 / E value	A	B	C	D	E
2	20.3%	2.5%	77.4%	10.0%	49.3%
38	68.1%	10.5%	21.4%	36.8%	54.0%
75	86.3%	10.1%	3.6%	48.5%	56.2%
99	90.1%	8.6%	1.3%	51.3%	56.9%

条件: DATA SET 3、フラグメント長=5、採用した検索結果数=5、Matrix=BLOSUM62

表 4.3　DATA SET 3 の E value 変化と二次構造予測精度

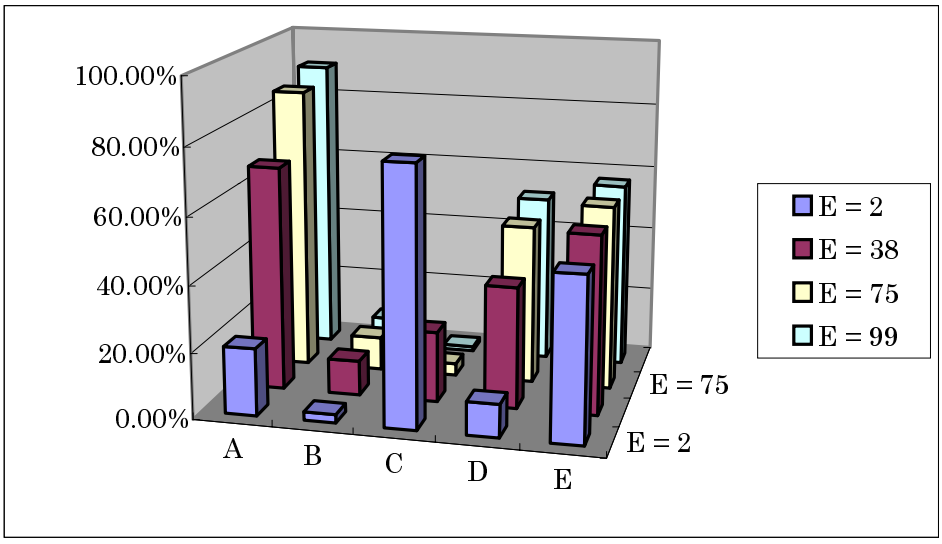


図 4.4　DATA SET 3 の E value 変化と二次構造予測精度

最も厳しい条件である DATA SET 3 の E value 変化による二次構造の正当率の変化について検証した。E=2 では、アミノ酸配列に対して類似配列は平均 2 割強しか得られなかった。E=99 では、アミノ酸配列に対して平均 9 割以上の類似配列が得られた。しかし、実際に得られた二次構造の精度については、どちらも 5 割前後の正当率である。これにより、類似度の高い配列が存在しないと配列-構造相関を示せないことがわかる。

4.2　検証データ 2

検証データ 1 と同様に以下の評価尺度で検証した。

DATA SET 4:　DATA SET 1 と同じ部分配列ライブラリの 823 鎖。(ただし自分自身を含まないように、新たに部分配列ライブラリを作成した)

DATA SET 5:　配列と構造が同定できた PDB 全体(13228 鎖)

評価尺度 A:　二次構造を示せて一意に特定できる割合(多数決が成立する)。
評価尺度 B:　二次構造を示せない割合(類似配列が存在しない)。
評価尺度 C:　全体(A+B)に対する二次構造の正当率。
評価尺度 D:　二次構造が示せ一意に特定できる割合(A)に対する二次構造の正当率。

4.2.1　フラグメント長変化による二次構造予測精度変化 2

4.2.2

評価尺度 / フラグメント長	A	B	C	D
5 残基	76.6%	23.4%	41.6%	54.3%
6 残基	84.9%	15.1%	48.4%	57.0%
7 残基	86.6%	13.4%	50.3%	58.1%
8 残基	86.5%	13.5%	50.3%	58.2%
9 残基	86.3%	13.7%	50.0%	58.0%

条件:　DATA SET 4、採用した検索結果数=5、E value=38、Matrix=BLOSUM62

表 4.4　DATA SET 4 のフラグメント長変化による二次構造予測精度

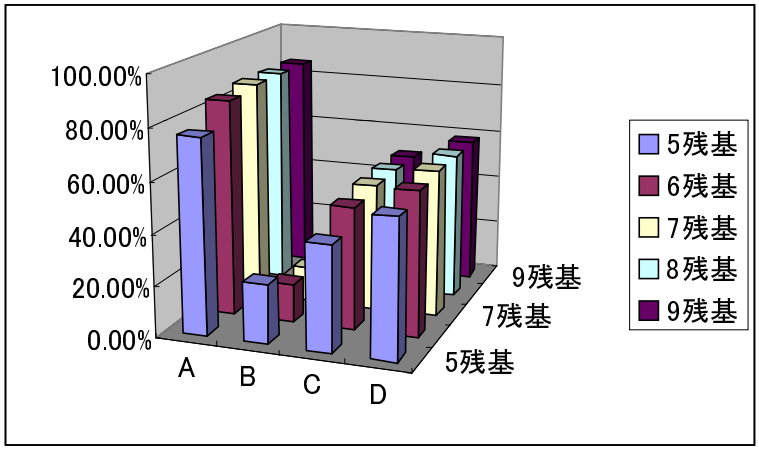


図 4.5　DATA SET 4 のフラグメント長変化による二次構造予測精度

探索空間に用いたこれまでの部分配列ライブラリでは、立体構造が特定された蛋白質に対してできる限り少ない数の代表蛋白質を使って作成している。これに対し、問い合わせ配列自身(同一配列)を含まない部分配列ライブラリをもとにした検索では、類似度の高い配列が減少する。これにより、検索結果の水準が低くなり二次構造情報の正当率は低くなる。

4.2.2　構造既知蛋白質全体での二次構造予測精度変化

評価尺度 / フラグメント長	A	B	C	D
5 残基	80.6%	19.4%	60.4%	74.9%

条件:　DATA SET 5、採用した検索結果数=5、E value=38、Matrix=BLOSUM62

表 4.5　DATA SET 5 の二次構造予測精度

PDB 全体に対する二次構造予測は、二次構造が示せ一意に特定できる配列部分に限って 74.9%の高い数値が得られた。これは、現在最も有力な二次構造予測法とほぼ同一の精度であることを示している。DATA SET 3 の E value 変化と二次構造予測精度より、E value の調節を最適化することにより高い精度で予測することができる可能性がある。

評価尺度 / 残基数	A	B	C	D	アミノ酸鎖数
0 - 50 残基	72.3%	27.7%	47.5%	65.7%	719鎖
51-100残基	80%	20%	60.5%	75.6%	1427鎖
101 - 150残基	83.3%	16.7%	66%	79.2%	2806鎖
151 - 200残基	85.2%	14.8%	71.7%	84.2%	1917鎖
201 - 250残基	78%	22%	53.1%	68.1%	1912鎖
251 - 300残基	79.3%	20.7%	60.4%	76.2%	1079鎖
301 - 350残基	82.9%	17.1%	62.2%	75%	1186鎖
351 - 400残基	84.3%	15.7%	68.6%	81.4%	649鎖
401 - 450残基	79.7%	20.3%	58.3%	73.1%	485鎖
451 - 500残基	80.2%	19.8%	59.5%	74.2%	372鎖
501 - 550残基	73.4%	26.6%	61.1%	83.2%	197鎖
551 - 1200残基	74.4%	25.6%	47.2%	63.4%	479鎖

表 4.6　DATA SET 5 の残基数による二次構造予測精度

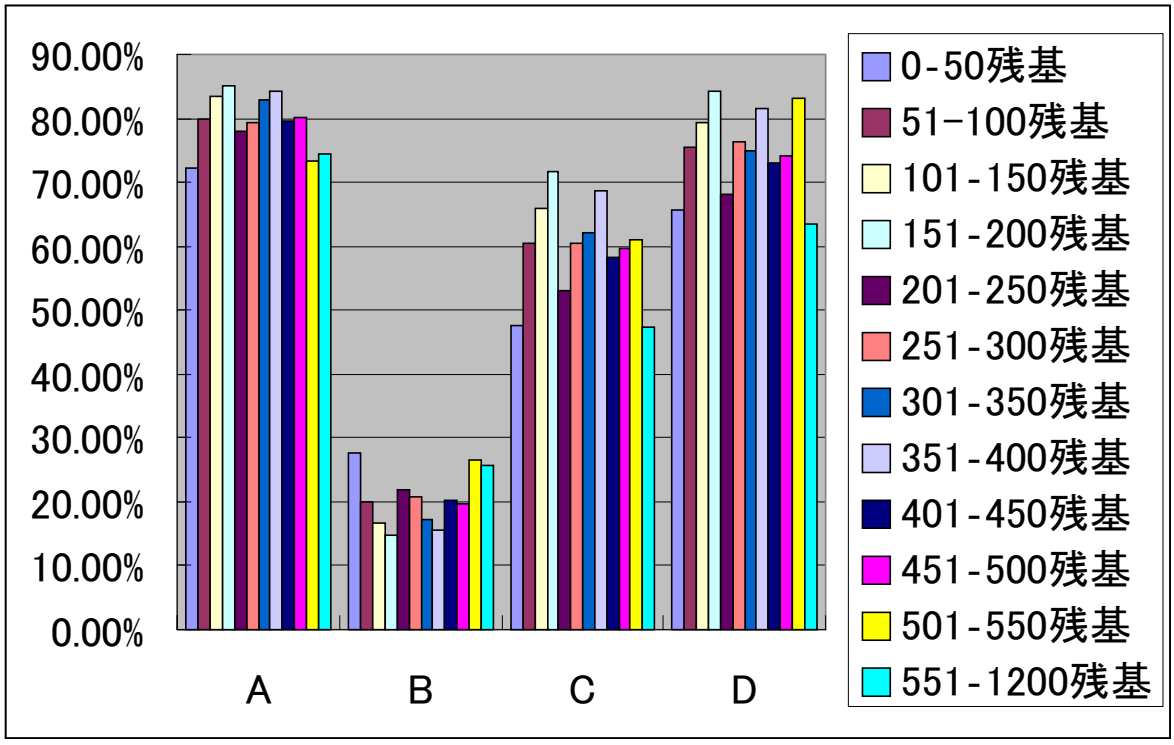


図 4.6　DATA SET 5 の残基数による二次構造予測精度

各アミノ酸配列について配列長(残基数)毎に検証した結果、50 残基以下のアミノ酸鎖と 551 残基以上のアミノ酸鎖では、配列全体に対する二次構造予測の精度は 50%を切っていた。特に、短い配列(30 残基以下の配列)の予測精度は極めて低かった。また長い配列では、750 残基以上の配列の予測制度が低かった。これは、極端に短い配列や長い配列が、代表蛋白を用いた部分配列ライブラリを作る過程でふるい落とされてしまったことが原因と考えられる。201-250 残基の間で予測精度が落ちるのは、検索部分配列ライブラリに用いた配列の長さの平均値が 233 残基であり、E=38 の条件では、類似度の検索精度が低く厳密な類似配列だけを用いることが出来なかったことが原因と考えられる。

第 5 章

まとめ

本研究では、既知の立体構造予測手法の欠点を補うために、部分配列の類似性に基いてデータベースから配列-構造相関を提示するシステムを構築した。また、システムの性能を調べるため、各種のパラメータやデータセットおよび PDB 全体について網羅的な検証を行った。構築したシステムについて部分配列の類似性から構造情報として二次構造をもとにどれだけ蛋白質の配列-構造相関の情報が示せるのか検証を行った結果、極端に短いまたは長い配列では二次構造の予測精度が低いことが判明した。残基数 50 未満の蛋白質は、ペプチドと呼ばれ、球状蛋白質とは違う形成を行う為、二次構造を特定できないことが予測精度の低下と考えられる。PDB_SELECT では、888 鎖のアミノ酸鎖で PDB 全体を代替しているが、部分配列ライブラリに用いたアミノ酸鎖は、823 鎖しかない。この 2 つの差である 65 鎖は、配列の途中でアミノ酸が欠損していることが原因で、部分配列ライブラリとして用いなかった。一般に、巨大分子である蛋白質を構成するアミノ酸の数が多いと、X 線結晶解析の結果得られる原子の数が多くなりすぎ全てに正確な構造マップを形成することができず立体構造としてのアミノ酸の欠損が起こる。以上により長い配列の二次構造予測精度が低いのは、長い配列に対して配列-構造相関を持った配列が部分配列ライブラリに用いられなかったことが原因と考えられる。解決方法として、アミノ酸が決損した配列に対して、どの位置で欠損したのかを特定し、新たに部分配列ライブラリに取り入れることで精度が上がる可能性がある。また今回は一意に E の値を定めたが、機械学習を用いて配列の長さ毎に適切な値を調節することでも精度があがる可能性がある。

このシステムでは、類似配列を検索するプログラムとして BLAST を用いている。

300 残基のアミノ酸配列に対して、配列-構造相関の提示に約 5 分の時間がかかる。これは 300 残基をフラグメントに分割して、各フラグメント毎に BLAST 検索を行うことが原因である。検索時間の短縮の為にいろいろな類似配列を検索するプログラムを試したが、短い部分配列に対応できたプログラムは BLAST しか見付けることができなかった。問い合わせ配列を部分配列に分割して類似部分配列を検索するのでなく、長い配列のままで類似部分配列を検索することにより、相当な時間短縮が見込める。しかし、既存の類似配列検索プログラムでは、長い配列に対して短い類似配列を検索することはできなかった。これを解決するするためには、短い配列のデータベースに対し、長い問い合わせ配列を使って高速に類似性を検索できるプログラムを開発することが必要である。

構築したシステムでは、二次構造の他に、構造情報として二面角情報、アミノ酸の内外位置情報などを得ることができる。しかし、このシステムはまだプロトタイプでどのように二次構造以外の構造情報を提示するのか決定していない。全ての構造情報について提示することができれば配列-構造相関がより精度高く提示できることになる。

システムの改良としては、PDB 全体の解析を行った結果を用いて、アミノ酸の残基数毎に配列-構造相関を示せた部分配列を特定し、部分配列ライブラリ化することで、問い合わせ配列の長さに適したシステムを構築できることになる。また、データベース中から高い精度で得られる部分はそのままにして、そうでない部分だけに既存の立体構造予測法を適用することで適切な予測を行う手法へと発展させることができるであろう。

謝辞

本研究を進めるにあたり、貴重な御助言、多大な御指導を頂いた佐藤賢二助教授に感謝の意を表し、心から御礼申し上げます。

また、本研究を行うにあたって有益な御助言、御示唆を頂けた小長谷明彦教授に心から感謝致します。

さらに佐藤研・小長谷研の皆様には、日頃から御討論、御協力を頂き、意義深い研究生活を送ることができました。深く感謝致します。

最後に、全面的な援助協力をしてくれました父母へ深く感謝の意を表しつつ、本論文の結びと致します。

参考文献

[1] Anfinsen, C.B. "*Principles that govern the folding of protein chains*", *Science*, **181**, 223-230(1973).

[2] Chothia, C. "*One thousand families for the molecular biologist*", *Nature*, **357**, 543-544(1992).

[3] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. "*SCOP: a structural classification of proteins database for the investigation of sequences and structures*", J Mol Biol, **247**, 536-540(1995).

[4] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, D., Shindyalov, L.N., and Bourne. PE. "*The Protein Data Bank*", *Nucleic Acids Res*, **28**, 235-242 (2000).

[5] Kawashima, S., Ogata, H., and Kanehisa, M. "*AAindex: amino acid index database*", *Nucleic Acids Res*, **27**, 368-369(1999).

[6] Kabsch, W. and Sander, C. "*Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features*", *Biopolymers*, **22**, 2577-2637(1983).

[7] Pauling, L., Corey, R.B., and Branson, H.R. "*The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain*", *Proc Natl Acad Sci* USA, **37**, 205-211(1951).

[8] Barlow, D.J. and Thornton, J.M. "*Helix geometry in proteins*", *J Mol Biol*, **201**, 601-619(1988).

[9] Chou, K.C., Pottle, M., Nemethy, G., Veda, Y., and Scheraga, H.A. "*On the use of distance constraints to folds a protein*", *J Mol Biol*, **162**, 89-112(1981).

[10] Sibanda, B.L. and Thornton, J.M. "*Knowledge-based prediction of protein structures and the design of novel molecules*", *Nature*, **316**, 170-174(1985).

[11] Creighton, T.E. "*The protein folding problem. In Mechanisms of Protein Folding (R. Pain, ed.)*", *IRL Press Oxford*, 1-25(1994).

[12] Creighton, T.E. "*The energetic ups and downs of protein folding (News & Views)*", *Nature Struct Biol*. **1**, 135-138(1994).

[13] 後藤祐児. "*立体構造はどのようにしてできるか？, in 蛋白質 - この微妙なる設計物*(赤坂一之 編)", *吉岡書店*, 54-78(1995).

[14] Kuwajima, K. "*The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure*", *Proteins: Struct. Funct. Gen*, **6**, 87-103(1989).

[15] Pearson, W. R. and Lipman, D.J. "*Improved tools for biological sequence comparison*", *Proc Natl Acad Sci USA*. **85**, 2444-2448(1988).

[16] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. "*Basic local alignment search tool*", *J Mol Biol*, **215**, 403-410(1990).

[17] Rost, B. *"PHD: predicting one-dimensional protein structure by profile based neural networks*", *Meth Enzymol*, **266,** 525-539(1996).

[18] Bowie, J.U., Luthy, R., and Eisenberg, D. "*A Method to Identify Protein Sequences That Fold into a Known Three-dimensional Structure*", *Science*, **253**, 164-170(1991).

[19] Piela, C. and Scheraga, H.A. "*On the multiple-minima problem in the conformational analysis of polypeptides. I. Backbone degrees of freedom for a perturbed alpha-helix*", *Biopolymers*, **26**, S33-S58(1987).

[20] http://www.genome.ad.jp/dbget-bin/show_man?PDBSTR

[21] Tsukamoto, Y., Takiguchi, K., Satou, K., Furuichi, E., Takagi, T., and Kuhara, S. "*Application of a deductive database system to search for topological and similar three-dimensional structures in protein*", *Computer Applications in the Biosciences (CABIOS),* **3**,183-190(1997.Apr. No2).

[22] Hobohm, U., Scharf, M., Schneider, M., and Sander, C. "*Selection of a representative set of structures from the Brookhaven Protein Data Bank*", *Protein Science*, **1**,409-417(1992).

[23] Hobohm, U. and Sander, C. "*Enlarged representative set of protein structures*" *Protein Science*, **3**, 522(1994).

研究業績

Tatsumoto, S. and Satou, K. "*A System for Displaying Protein Structure Information Based on Sequence Fragment Similarity*", *Genome Informatics*, **10**, 237-238(1999)