平成 13 年度知能システム科学専攻修士論文

# 郷モジュールに基づくフォールディングシミュレーション

**名波 剛**

# Folding simulation of protein structures based on GO module

Takeshi Nanami

提出年月日 平成 14 年 2 月 12 日 (火)
修正再提出年月日 平成 14 年 2 月 26 日 (火)

主査教官：山村 雅幸 助教授
審査教官：小林 重信　 教授
審査教官：樺島 洋介 助教授

# 郷モジュールに基づくフォールディングシミュレーション

名波 剛

# Folding simulation of protein structures based on GO module

Takeshi Nanami

**abstract :** There is no doubt that biotechnology has been playing a crucial role in not only itself but also other related scientific studies. Protein structure prediction, one of indispensable issue in the study, has been discussed in various ways. No one state, however, that the issue can derive unknown structures with enough accuracy. In this paper, we apply GO module idea and exon-shuffling hypothesis for prediction and detection of protein structures. GO module is a set of C-alpha atoms in a certain protein and its boundary is defined using average square distances, while exon-shuffling contributes to protein evolution. With the assumption of strong relationship between GO module and molecular evolution, first and foremost, we confirm that the boundaries of GO modules correspond to them of exons. And then the new approach using those ideas is described. As the result, this novel method derived some correspondent structures to well known a protein in aaRS family.

## 1 はじめに

近年、ゲノム解析やバイオテクノロジーの急速な発展と計算機の飛躍的な能力上昇に伴い、遺伝情報などを計算機を用いて解析する遺伝子情報処理の分野が脚光を浴びている.

そのバイオテクノロジーの一分野にタンパク質の立体構造を求める分野があり, 大変重要なテーマとなっている. それは, タンパク質の立体構造はその機能と密接に関係しているため，立体構造が似ていれば機能のメカニズムが似ている場合が多く, 立体構造から機能のメカニズムが推定できることが多い．そして，この構造と機能を知ることは製薬などの幅広い応用の場となっている.

今日，立体構造を生物化学的実験で求める方法に X 線回折や核磁気共鳴 (NMR) などがある．しかし，いずれの手法においてもタンパク質の立体構造を決定するのは容易ではなく，一つのタンパク質を決定するのに数ヶ月以上の期間を要することが普通である.

そこで，計算機を用いてタンパク質の立体構造を予測することが非常に重要になってきている. タンパク質の立体構造予測とは 1 次元のアミノ酸配列から複雑に折り畳まれた 3 次元の立体構造を求めることである. タンパク質はそのアミノ酸配列に応じて自発的に一定の立体構造へと折りたたまれるため, 立体構造形成に必要な情報はすべてアミノ酸に与えられていると考えられている. 原理的にはアミノ酸配列さえ与えられれば, それをもとに立体構造を予測できるはずである. そして, 折りたたみの過程を知る上で物理化学的観点から理論的に理解する必要があり, この問題は「折りたたみ (フォールディング) 問題」として知られ長く追求されてきたが, 構造を予測するような理論的理解は得られていない.

そのため, 経験的な方法による予測によって得ようとする動きがある. それには配列の類似性を用いたホモロジーモデリングや立体構造と配列間の適応度を用いたスレッディング法などである．しかし，いずれのアプローチにおいても立体構造を予測するのに決定的な方法やそのフォールディングのプロセスを完全に解明されたものはいまだに見つかっておらず, 多くの研究が今もなお行われている.

そこで, 本研究では, タンパク質の構造を分割してその機能を探ろうとしている郷モジュール [1, 2] と様々なタンパク質が発生した理由として，真核生物の DNA 配列のうちタンパク質にコード されるエキソン部分が混ぜ合わされたとされるエキソンシャッフリング [3, 4] に着目する．そして, この 2 つにより興味深い仮説が立てられる.

1. モジュールがフォールディングのシード (種) として働くのではないか

2. モジュール単位での組換えを行えるように生命は進化してきたのではないか

という 2 点である.

本研究では, この 2 つの仮説を用いてタンパク質の立体構造予測法の提案もしくはタンパク質に対して新たな知見を得ることを目的とする. 今回, 提案手法の実装にあたっては, Protein Data Bank(PDB) からモジュールを抽出して, それを求めるアミノ酸配列に割り当てる. 割り当てられた結果に対して, モジュール単位でのフォールディングを行う. その手法には SA(Simulated Annealing) を採用する.SA を用いての最適化は, 目的関数の定義域が十分大きく, 全探索が事実上不可能のような問題に適している. 本手法を aaRS(アミノアシル化 tRNA 合成酵素) に適用し, その得られた構造に対しての評価と分析を行う.

本論文の構成は以下のとおりである.2 章ではタンパク質の立体構造とフォールディング及び本研究に取り入れる郷モジュールとエキソンシャッフリングについて詳細に説明すると共に, その性質を述べることで本研究の着眼点を明らかにする. 3 章では提案手法の概念と提案手法の実装を行う. 4 章では計算機実験の結果と考察をする. 5 章において本研究のまとめと今後の課題について述べる.

# 2 準備

本章では, タンパク質の立体構造とフォールディング及び提案手法で用いられる郷モジュールとエキソンシャッフリングを紹介し, 本研究において特に重要となる郷モジュールの境界とエキソン境界の関係について述べる. そして, 最後にこの郷モジュールとエキソンシャッフリングから導き出された 2 つの仮説について述べる.

## 2.1 タンパク質の立体構造

タンパク質はアミノ酸が数個から数千個単位でペプチド結合してできた鎖のような構造を持つ分子である.(図 1)

そして, アミノ酸は中心炭素原子 ($C_\alpha$) を持ち, その

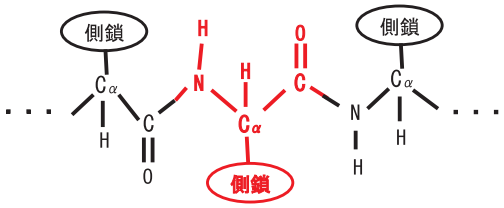


図 1: タンパク質の立体構造

炭素原子に, 水素原子, アミノ基, カルボキシル基が結合しているという共通の性質を持っている. 1 つのアミノ酸と他のアミノ酸を区別するのは,$C_\alpha$ と結合している側鎖である.

このアミノ酸による鎖が複雑に折り畳まれエネルギー的に最も安定な構造をとる.

## 2.2 フォールディング

フォールディング [5] とはアミノ酸配列としてコードされた 1 次元的な遺伝情報が, タンパク質として機能を持った天然の特異的な 3 次元立体構造に変換される過程のことである. そのフォールディング過程においてはアミノ酸残基同士の局所的相互作用からスタートする. つまり, 空間的に近いアミノ酸同士が相互作用して部分的構造が形成されると, 次に, これらの構造の間の相互作用が起きて, 最終的な構造にいたるとされている.

そして,「タンパク質の天然立体構造はそのアミノ酸の 1 次配列 (=遺伝情報) により一意的に決定付けられる」(Anfinsen の仮説). これは, タンパク質のフォールディングが純粋に物理化学的な過程であり, フォールディングの分子機構が物理, 化学の原理に基づいていることを意味している. つまり,「天然条件下においては, タンパク質の天然立体構造は全系におけるギブス自由エネルギーが最小の構造に対応する.」とされている.

しかし, タンパク質の鎖は多様な立体構造をとることができ, それに対応した多数のエネルギー極小値をとる. そのため, エネルギー曲面は大変複雑で天然状態の極小値を探すのが困難となっており, その構造を求めることができずにいる.

## 2.3 郷モジュール

ここで，郷モジュールの考え方を示す. 郷モジュールとは立体構造既知のタンパク質から決まる空間的にコンパクトな単位で, アミノ酸残基が平均約 15 残基程度の長さを持つ.

モジュールの境界は下記の式で求められる $F_i$ の極小値である.

$$F_i = \sum_{j_1 \le j \le j_2} r_{ij}^2/(j_2 - j_1)$$

$F_i$ は $i$ 番目の残基の $C_\alpha$ と他の残基の $C_\alpha$ との平均二乗距離で求められる. 但し, 比べる残基の数は $i$ 番目の残基から $k$ 以内の残基である. $j_1$ は $\max(1, i-k)$, $j_2$ は $\min(n, i+k)$, $r$ は残基間の距離, $n$ はそのタンパク質の全体の残基数である.

この式で表しているものは $i$ 番目の残基がどれだけ周囲のアミノ酸残基のローカルな中心になっているかということである. つまり, $i$ 番目の残基が極小値を取るときは, その周りにアミノ酸残基が固まっているという指標でもある.

そして, この郷モジュールは以下のような性質を持っている.

- 空間的にコンパクト
- モジュールの境界はエキソンの境界と一致する傾向にある
- エネルギー準位として極小値の近くにある
- 2 次構造 ($\alpha$ ヘリックスや $\beta$ ストランド) の途中から途中までという構造を持ちやすい

郷らはこのようなモジュールがタンパク質の機能の進化のプロセスでどのようにデザインされてきたか. また, 新機能を持った全く人工のタンパク質をデザインする指標としてモジュールを利用するなどの研究を進めている.

## 2.4 エキソンシャッフリング

エキソンシャッフリングを知るには, イントロンとエキソンの関係及び, 真核生物におけるタンパク質の合成過程 (図 2) を知ることが非常に重要である.

図 2 が示すように, タンパク質の合成過程はイントロンとエキソンに分かれた DNA 配列から RNA に転写され, 真核生物においては RNA スプライシングと

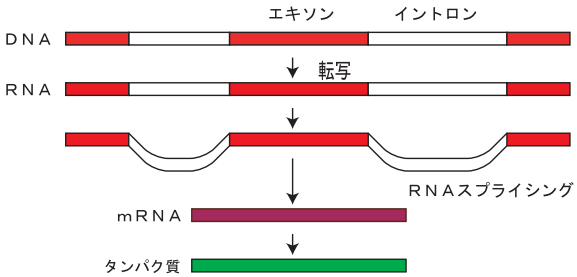


図 2: 真核生物のタンパク質合成過程

呼ばれる動作によってイントロンが取り除かれる. そして, RNA は mRNA として細胞質に移動しアミノ酸に翻訳されタンパク質へとフォールディングされる. エキソンとは, 最終的にタンパク質にコード化される部位である. 逆に, イントロンとは RNA スプライシングでで取り除かれてコード化されない部位である. このような流れの中で, 一見無駄に思えるイントロンは様々なタンパク質を発生させる上で非常に重要な役割を果たしてきたと考えられている. 一つ目には, スプライシングのされ方によって一つの遺伝子からいくつかの違ったタンパク質を合成する. 二つ目として, 遺伝子に数多くのイントロンが含まれている際に, その部分を用いての組換えが容易となる. つまり, このイントロンの部位で組換えが起こりエキソン間をまぜ合わせる. このことをエキソンシャッフリングと呼ぶ. (図 3)

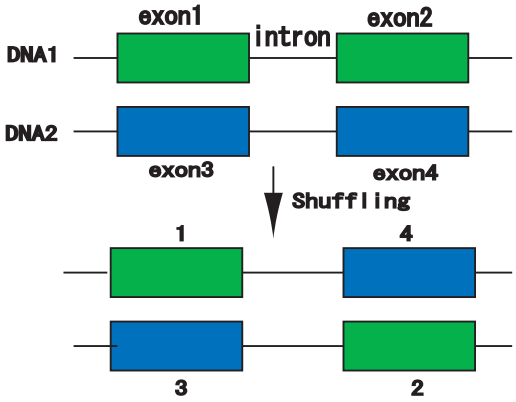


図 3: エキソンシャッフリング

実際に, 大きな遺伝子ではエキソンの数百倍以上の長さを持つイントロンが多い. そのため, イントロン上で組換えのオペレーションが行われ, エキソン部位の情報は保存されたままになっていることが非常に多い.

### 2.5 エキソンの境界とモジュールの境界

モジュールの境界とエキソンの境界は一致する傾向にあるということを構造既知のタンパク質とそのエキソンの位置の分布を比べたところが表 1 である. これは,Arabidopsis thaliana(シロイヌナズナ) の aaRS タンパク質におけるものである. この他にもいくつかのものを調べた結果, 数残基ずれることも見られたがほぼ一致する傾向にあった. これらのことから何が考えられるかの考察については次節で述べる.

| modul | Residue number | exon |
|---|---|---|
| M1-M2 | 20 | - |
| M2-M3 | 37 | - |
| M3-M4 | 46 | - |
| M4-M5 | 60 | - |
| M5-M6 | 74 | 76-77 |
| M6-M7 | 95 | - |
| M7-M8 | 104 | - |
| M8-M9 | 134 | - |
| M9-M10 | 156 | 155 |
| M11-M12 | 173 | - |
| M12-M13 | 180 | - |
| M13-M14 | 198 | - |
| M14-M15 | 217 | 210 |
| M15-M16 | 243 | - |

表 1: エキソンとモジュールの境界

### 2.6 郷モジュールとエキソンシャッフリング

上記に示した郷モジュールとエキソンシャッフリングから 2 つ仮説が導き出される.

1. モジュールがフォールディングのシード (種)
2. モジュール単位での組換えを行えるように生物は進化

**フォールディングのシード とは** タンパク質がフォールディングするに際しモジュールが最初にアミノ酸残基の局所的な相互作用で部分構造を形成している部分であると考えられる. つまり, 最初にモジュールが形成され, 次にモジュール同士の相対位置がいろいろと変わって全体構造の形成に至るということである. これは, モジュールがエネルギー極小値の近くにあり, 空間的にコンパクトであるという性質をもっているためにシードとなる可能性がある. さらに, モジュールはモジュール内部での相関が強く, 外との相関が比較的弱い構造をとるので, 他の領域との相互作用で自分自身の構造が壊されにくいためでもある.

**モジュール単位での組換えとは** エキソンは多種多様なタンパク質を発生させるのに組換えを行ってきたとされている. ここで, エキソンの境界とモジュールの境界が一致するのは, モジュールという単位で組換えが可能なように生命は進化してきたのではないかと考えられる

本研究においては, 上記の 2 つの仮説に基づいてタンパク質の立体構造を求める手法の提案を図る.

## 3 提案及び実装

本章では, タンパク質の立体構造予測の手法の提案を行う. モジュールとエキソンシャッフリングから導き出された 2 つの仮説により, 従来にはない新しい立体構造予測である. 最初に, 提案手法の概念を説明し, 次に実装方法の説明をする.

### 3.1 提案手法の概念

上記に示したように本研究では, 郷モジュールとエキソンシャッフリングを元に従来にはない立体構造予測法を提案する. 提案手法の主な流れを図 4 に示す. 本手法は

1. モジュールの抽出
2. モジュール割り当て
3. SA による構造最適化
4. 構造に対して GA による組換え

という 4 つの事柄から成り立っている. モジュールの抽出では立体構造既知のタンパク質からモジュールを抽出し, データベースを作成する. アミノ酸配列へのモジュールの割り当てでは, アライメントを利用してモジュールを求めるアミノ酸に割り当てていく. モジュー

ルが割り当てられた配列に対して SA を用いてエネルギーが最小化になるようにフォールディングさせる. そして, 複数の構造最適化されたものを初期個体として GA による組換えのオペレーションを行う.

本稿では，この 4 つの事柄のうち 3. の SA による構造最適化までを実装した.4. の GA による組換えは今後の課題とする.

## 3.2 モジュールの抽出

モジュールの抽出には, 立体構造既知のタンパク質からそのタンパク質の各アミノ酸残基の $C_\alpha$ 間の平均二乗距離を求め, その極小値を求めることによりモジュールの境界を求める. 図 5 がアミノ酸残基の平均二乗距離を求めたグラフになるが, 現段階ではグラフを利用して手動でモジュールの境界を定めて抽出を行っている.

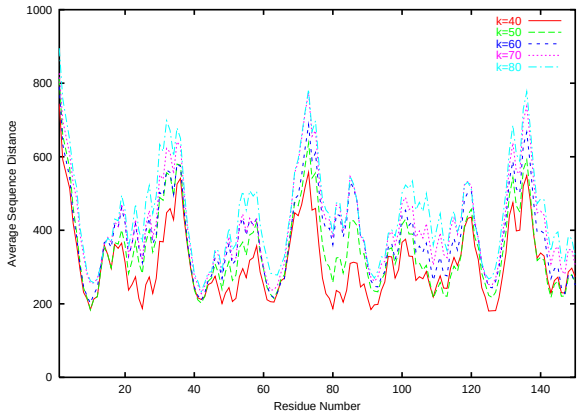


図 5: アミノ酸残基の平均二乗距離

実際に,PDB(Protein Data Bank) を用いて aaRS(アミノアシル化 tRNA 合成酵素) の約 100 個前後のタンパク質に対してモジュールを抽出した. その結果, モジュールのデータベースとしては,1500 個以上のデータセットが集まった.

そのアミノ酸残基の長さは平均約 16 残基であり, それを長さによる頻度グラフで表したものが図 6 である.

その構造は,$\alpha$ ヘリックスや $\beta$ ストランドといった 2 次構造の途中から始まり, 途中までという構造をとることが多く, モジュールの境界はエキソンの境界と一致するというデータが得られた.(図 7) これは，郷によって求められた平均約 15 残基, モジュール構造の特徴, モジュールの境界とエキソンの境界がほぼ一致するという知見と同じであった.

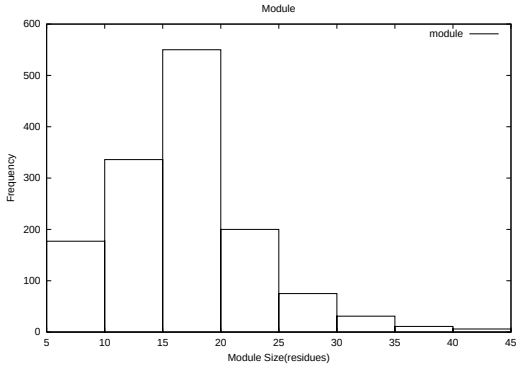


図 6: モジュールの頻度グラフ

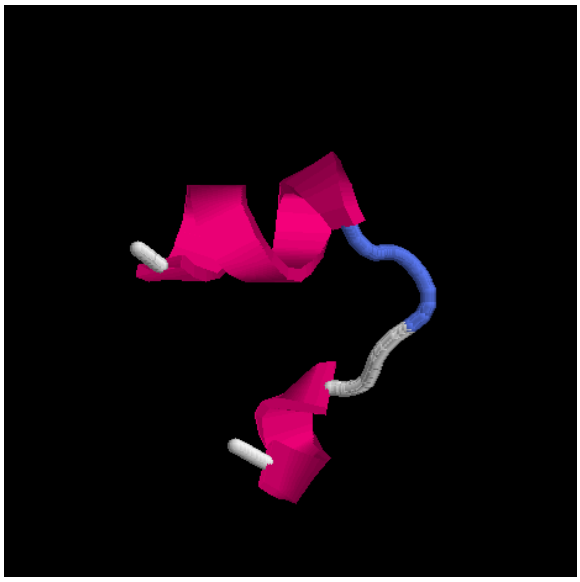

図 7: モジュールの構造

## 3.3 モジュールの割り当て

モジュールの割り当てでは, 立体構造既知のタンパク質から求められたモジュールを求めるアミノ酸配列に割り当てを行う. その処理は以下のようである.

1. アライメント [9] スコアの高いもの順にモジュールを割り当てる
2. モジュールがないところにリンクの生成

リンクとはモジュールとモジュールの間が埋まらなかったところである. つまり, 割り当て時にはリンクは 1 次元のアミノ酸配列情報しかもっていない.

また, アミノ酸配列に割り当てる際に, 一意の割り当て方だけではなく複数の割り当てができる. これは, 立

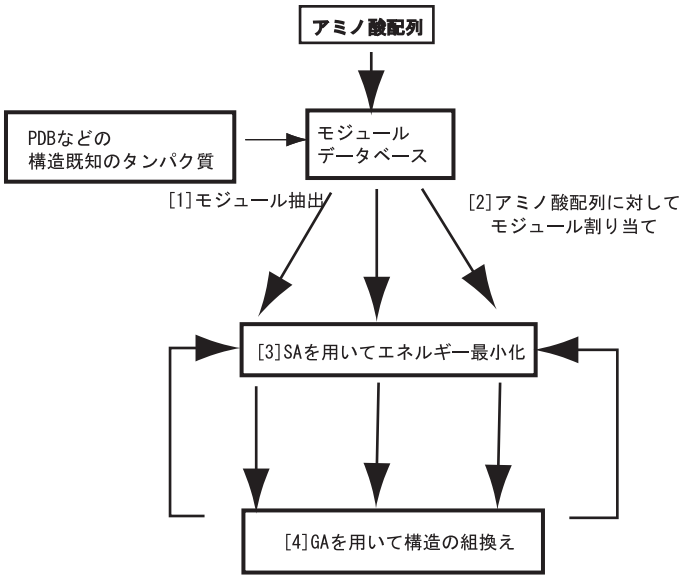


図 4: 提案手法の全体図

体構造を求めるアミノ酸配列上に重複して, モジュールが割りてられるために起こることである.(図 8)

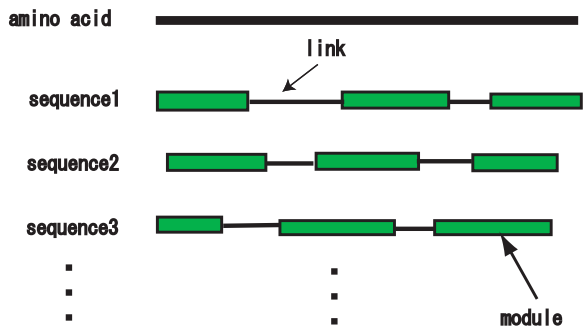


図 8: モジュールの割り当て

アライメントのスコアによる割り当てをすることで, アミノ酸間で置換が起こりやすい残基においてはほぼ一致しているとみなしている. このアライメントはギャップ無しによるもので, そのアミノ酸間の置換行列として BLOSUM50 置換行列 [9] を採用した.

また, モジュールの抽出を行ってみて得られた知見にエキソンの境界とモジュールの境界が数残基程度ずれることが時折見られた. これは, アミノ酸の欠損や挿入などで起こったのではないかと考えられる. そのため, これを解決するために, 各モジュールの割り当てのオペレーションの中に両端の 1,2 残基程度のオーバーラップを許すように割り当てを行う.

## 3.4 SA による構造最適化の実装

SA を用いたフォールディングアルゴリズムを実装する. その際の評価関数としてエネルギーを用いる. ランダムに動かすものとして, モジュールの重心である. まず, 実装方法のアルゴリズムを説明し, 次にエネルギーの評価関数のことについて説明する.

### 3.4.1 アルゴリズム

実装方法のアルゴリズムについての全体の流れを図 9 に示す.

最初にモジュールが割り当てられたアミノ酸配列に対して初期位置を決定する. その際に, リンクの部分はアミノ酸配列の 1 次元の情報なので, 初期は位置は結合長や結合角などの簡単な制限により座標を与える.

その初期配列に対して, モジュールの重心をランダムに移動させる. この際に, 移動させるのは X,Y,Z の変位のほかに回転による移動も付け加える.

リンクがあることで, モジュールだけを動かしたとしても最適な構造とはならないので, リンク内の最適化を行う. このリンクの部分の最適化には, ローカルな形で SA を用いる. このリンク内の SA では, 動かす対象をすべての原子にするのではなく, アミノ酸残基の $C_\alpha$ をランダムに動かす. 動かした結果を構造全体ではなく, リンク内においてのみのエネルギーで評価する. リ

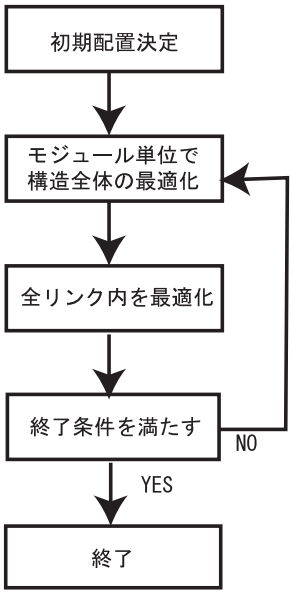


図 9: アルゴリズムのフローチャート

ンク内を最適化した後, 終了条件を満たさなければ, モジュールごとの SA に戻り, 構造の最適化をしていく.

具体的な SA の様子を図 10 に示す.

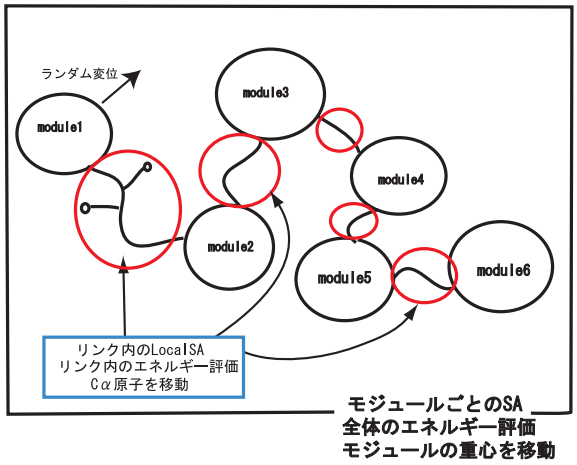


図 10: 階層的 SA

### 3.4.2 エネルギー評価関数

タンパク質のエネルギー評価をする関数として様々な力場 (Force) パラメータがあるが, 本手法ではそのうち AMBER[6, 7] の力場パラメータを用いてエネルギーの計算を行う. 下記にそのエネルギーの関数示す.

1. 共有結合 (bond)

$$U_{bonds} = \sum_{bonds} K_r(r - r_{eq})^2$$

2. 結合角 (angle)

$$U_{angles} = \sum_{angles} K_\theta(\theta - \theta_{eq})^2$$

3. ねじれ角 (torsion)

$$U_{tortions} = \sum_{tortions} \frac{V_n}{2}(1 + \cos(n\phi - \gamma)$$

4. ファンデルワールス相互作用 (VDW)

$$U_{vdw} = \sum_{i<j} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}]$$

5. 静電気相互作用 (coulomb)

$$U_{coulomb} = \sum_{i<j} \frac{q_i q_j}{\epsilon R_{ij}}$$

この式のうち, それぞれのパラメータはエネルギーが作用している原子同士によって一意に決まるパラメータである. 共有結合 (bond) とは 2 原子間の距離で得られるエネルギーで原子同士がバネのようなものでつながっていると近似したものである. 結合角 (angle) とは, 共有結合をしている 3 原子間で角度によるエネルギーである. ねじれ角 (torsion) のポテンシャルエネルギーとは,4 つの原子で作られる 2 つの平面の角度によるエネルギーである. この 3 つのエネルギーは結合に関するもので $N$ 体系の場合は $O(N)$ の計算時間ですむ. 一方, ファンデルワールス相互作用 (vdw) や静電気相互作用 (coulomb) はすべての原子間で働くエネルギーである. ファンデルワールス力は近い距離だと強く反発し, 遠くになると小さな力で引き合う. この力は Lennard-Jones が提唱したポテンシャルで近似されている. この 2 つの項の計算は工夫も無しにすると $O(N^2)$ の計算時間がかかり, 特に本手法のように大きなタンパク質に対して計算する上では工夫が必要となる.

### 3.4.3 エネルギー評価の実装

SA において求められる点としてエネルギーの評価の高速化がある. 本手法では, エネルギー評価を高速化するためにいくつかの工夫を盛り込んである.

- 距離によるカットオフの導入
- 各原子ごとの距離によるリスト作成 [8]

まず, 一つ目のカットオフの導入とは通常このような生体分子の計算を行う際に, 分子間の相互作用の項 (VDW 力とクーロン力) は距離によるカットオフを導入する事が多い. その距離は 8Å から 15Å 程度で切断することが一般的であり, 本手法においてもこの距離の間によるカットオフを導入する. このカットオフの距離のことを $r_{off}$ と定義する.

次に,SA では 1 ステップの変位が少ないことを利用して, 毎回すべての原子の距離を求めるのではなく, 何ステップかに一度, 自分が分子間相互作用している原子のリストを作成する. そのリストは, カットオフの距離よりも長い距離で作っておかなければならない. なぜならば,SA の 1 ステップ前にカットオフの外にあった原子が変位を行ったときに, カットオフの距離の中に存在してしまうからである. さらに, リストの距離はリストの更新回数にも比例してくる値である. 変位させる距離を $R_{max}$, リストを更新するステップ数を $S_{step}$ とすると, リスト距離 $r_{list}$ は $r_{list} > r_{cutoff} + 2 * S_{step} * R_{max}$ とする.(図 11)

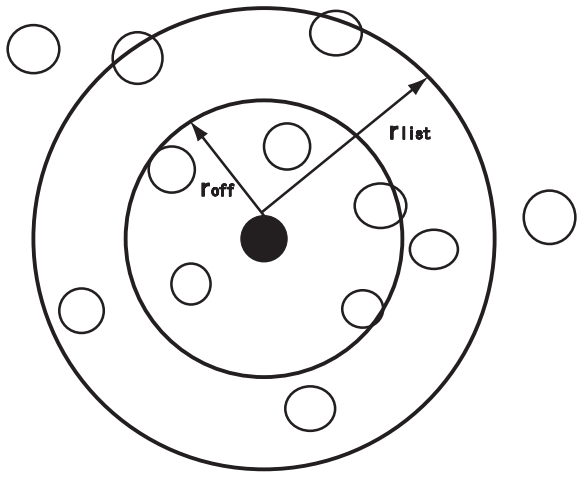


図 11: リスト作成

# 4 計算機実験の結果と考察

本章では，前章で述べた実装した方法をを立体構造既知のを aaRS タンパク質 (1JII) のアミノ酸配列に適用し, その結果を比較して考察をする.

## 4.1 実験目的

実験目的は, 提案手法を用いた際にモジュールの部分が核となり求めるタンパク質に似た構造へとフォールディングすることができるかの比較をするための実験を行った.

## 4.2 実験条件

この求めるタンパク質 1JII は既に構造が求められているタンパク質であり, その構造が知られている. そのため,1JII と全く同じような配列を持つようなタンパク質をモジュールのデータベースからは取り除いてある.

本手法の中でのパラメータとして

- カットオフ距離:8Å
- 相互作用のリストの更新:10 ステップ
- モジュールの最大変位:0.5Å
- $C_{alpha}$ の最大変位:0.1Å
- 初期温度:1000

とした.

## 4.3 評価法

立体構造を評価する方法として,

- 立体構造の形
- RMSD(Root Mean Square Distance)

を用いて評価する. RMSD とは,

$$RMSD = \frac{1}{N}\sum_{i=1}^{N}(R_i - r_i)$$

で求められる構造間の距離の差を表したもので, アミノ酸残基などの単位でその構造のずれを評価するものである. 但し,$N$ は比べるものの全体数,$R_i$ と $r_i$ は 2 つの構造の対応する位置のそれぞれの座標などである.

## 4.4 結果

モジュールの割り当てられた様子を図 15 に示す. ターゲットのタンパク質の構造を図 12 に示す. 計算の結果のタンパク質を図 13 に示す.

この得られた結果に対して,RMSD において $C_\alpha$ 原子の差をとり, その距離を測定した結果,20.04Å と大変大きな差になった.

図 12: ターゲットのタンパク質:1JII

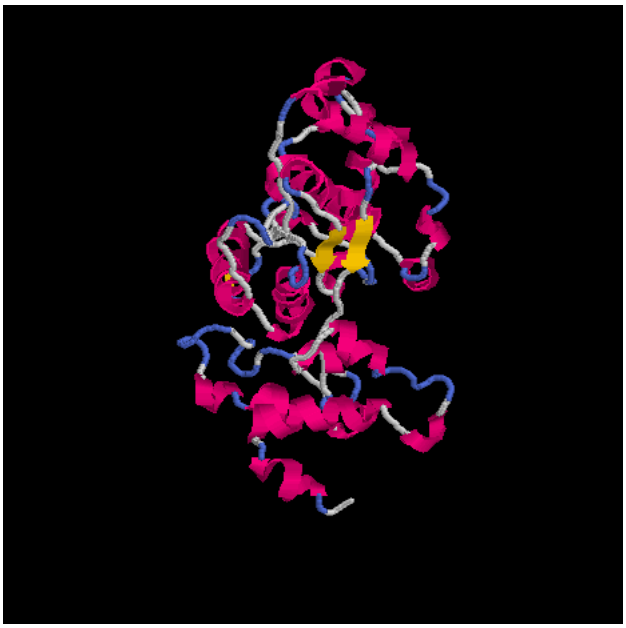

図 13: 計算結果のタンパク質

### 4.5 考察

2 つの構造を比べると全く同じようなタンパク質の部品を使わなかったにもかかわらず, 部分的に類似しているところが数多く見られた. これは, モジュールがこの立体構造を形成するための各となる可能性は十分あると考えられる.

しかし,RMSD 距離では 20.04Å という大きな値になり, 全体的な構造の評価としてはよくない. その原因として

1. リンクの近似で大きなずれが生じている
2. モジュールの割り当てがうまくいってない
3. モジュールの絶対数が足りない

といったことが考えられる.
本稿での実装では, リンクの部分をリンク内での SA を動かして構造を最適化させることを行った. しかし, リンクの部分の配列が非常に長いところもあり, ローカルな情報しか見ていないリンクでのずれが大きなものとなっていると考えられる.
本手法では, モジュール内部は動かさないのでモジュールの割り当てが大変重要である. 本稿のモジュールの割り当てでは, アライメントのアミノ酸置換行列として BLOSUM50 を採用したが, この置換行列が最も本手法で適切な置換行列かが問題である. アライメントの質は置換行列によって定まるので適切な置換行列の検討が必要になると考えられる.
モジュールの数が 1500 個程度だと割り当てれるモジュールの数がかなり限られてきてしまっている. 今回はモジュールの抽出を手動でやったため, モジュール数が少なくなってしまった. そのため, 求めるアミノ酸配列に対してモジュールの割り当てが少なくなり, モジュール間に大きな隙間ができることになってしまった. そこで, 本手法を現実的なものとするには, この抽出を自動的にやることによって, モジュールの数を増やし, モジュールの割り当てられる数を増やすことでリンク部分を減らすことができると考えられる.

また, 構造最適化のための SA は, モジュールごと動かすのとリンク内での最適化と 2 段ぐみになっており, 一つの結果を得るために大変時間がかかるものとなっている. そのため,GA などの組換えを行う際には, フォールディング方法の見直しなどが必要になってくると考えられる.

## 5 まとめ

本研究では, 郷モジュールとエキソンシャッフリングの考え方に基づき新たな立体構造予測手法を提案し, そ

の手法を用いての実験を行った.

本手法を aaRS タンパク質に適用し比較をした結果, 求める立体構造と部分的に一致するところが見られたのでモジュールが立体構造を形成する可能性はあるが, 全体的な構造としては RMSD 距離による比較でわかるようにかなりの差が生じてしまい, 予測手法としては現段階では実用的ではなく, 研究としては完成途上である.

そこで, 本研究で達成できたことと今後の課題を解決法の考察をいれて示す. まず, 現段階で達成できたこととして以下のものがある.

1. モジュールの性質の確認
2. 1500 個程度のモジュールデータベース作成
3. ギャップ無しのアライメントによるモジュールの割り当て
4. エネルギー計算の高速化
5. 階層的に SA を用いての構造最適化

今後の課題としては, 以下のようなものが考えれる.

**モジュールの自動抽出化** 現実的に,20 残基以上のモジュールを求めるアミノ酸配列に割り当てるのはかなり難しくなっており, 割り当てられることはほとんどない. そのため,20 残基以上のモジュールは抽出しなくてもよいのではないかと考えられる. 具体的な手法として, まず極小値を見つけてから, その極小値から 20 残基程度の幅に再び極小値が表れるかを調べる. もしあるようならば, それをモジュールとして抽出する. もしないようならば, 次の極小値からその次の極小値までの残基数を調べる. これを繰り返すことで,20 残基程度以内のモジュールを自動的に抽出することができると考えられる.(図 14) 自動抽出によるモジュールデータの増加は, リンク部分を減らすことや割り当て方の種類を増やすことができる. そのため, すぐに本手法を改善できる方法として期待できる.

**アライメントのアミノ酸置換行列の検討** 今回用いた BLOSUM50 の置換行列以外にも PAM 行列や別の BLOSUM 行列等の様々な置換行列が提案されている.[9] 本手法では, 求めるアミノ酸配列と各モジュールとのペアワイズアライメントで比較をしている. モジュールの配列は平均 15 残基と大変短いので, このような短い配列と長い配列を比較する場合に適しており, さらにギャップ無しで高い精度が上げられるような置換行列が望ましい. そこで、それぞれの置換行列の導出の経緯を考える必要がある.PAM 行列は非常によく似たタンパク質のアライメントから置換データを得たため, 進化的な関連性の強いような行列である.BLOSUM 行列は BLOCKS データベース [10] と呼ばれるタンパク質ファミリーのアライメントされたギャップ無しの領域から導出された経緯がある. このような経緯を考えると, 本手法では BLOSUM 行列で, さらに BLOSUM80 などのように後ろの数字が高いものがよいと考えられる. この数字は小さな値ほど より進化的に離れた配列の検索が可能となっているので, 本手法ではできるだけ大きな値のものを選び, 進化的にも近いもののみを割り当てていくことで, 適切な割り当てが可能となるのではないかと考えられる.

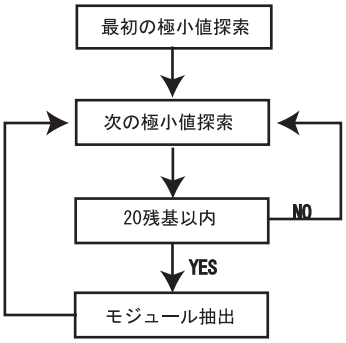


図 14: モジュールの自動抽出化アルゴリズム

**リンク内の近似の見直し** 現段階では, リンク内に SA を走らせることで 2 段階の SA となっているため, 大変時間がかかるので, 根本的な見直しが必要となる. そこで, リンクの部分はたまたまモジュールが割り当てることができなかったとみなして, 全リンク内を最初に構造最適化をしてしまい, リンクをモジュールとみなして構造最適化をしてしまう方法が考えられる. このようにすれば,2 段階に SA を動かす必要がなくなる. しかし, これにはリンクの部分が非常に多いと構造の核となるものがなくなってしまうので, たくさんのモジュールが割り当てられていないとうまくいかないと思われる.

**SA 計算の並列化** エネルギー計算がある程度速くなっているが, まだ SA での計算に時間がかかり GA による組換えまではできない. そこで,SA による構造最適化を並列化計算させることで GA の導入が現段階よりは現実的になると期待される.

**GA による組換えの導入** 組換えでは,GA を用いて組換えをさせる. 交叉手法として一点交叉を用いるのが妥当である. それは, モジュールの割り当てが求めるアミノ酸配列をすべてをカバーするようにはほとんど割り当てることができないため, 複数個作られた割り当て配列のうちモジュールが割り当てられていないところが出てくる. その部分を交叉させることで、モジュールの部位が保存されたまま組換えを起こすことが可能になる. さらに, 割り当てられたモジュールが全く違うような構造を持っていたときにも, そのような構造は組換えによって淘汰されていき, 最終的には残らないのではないかと思われる. そして, このようなオペレーションにより, エキソンシャッフリングとモジュールの性質から導き出されたモジュール単位での組換えを行えるように生物が進化してきたという仮説を取り入れることができ, タンパク質の多様性へのアプローチが期待できる.

今後, 上記にあげたような課題を解決することで現段階よりも精度の高い予測が実現できることが期待される.

## 謝辞

本研究を行なうにあたり, 何もわからないところから終始多大なる御指導ならびに御教示を頂きました山村 雅幸助教授に深く感謝の意を表します. また、本研究を進める上で, 色々と相談に乗って頂いた山村研究室の皆様に深く感謝致します. 最後に学生生活を暖かく見守ってくれた両親に心から感謝します.

## 参考文献

[1] M. Go and M. Nosaka., Protein Architecture and the Origin of Introns Cold Spring Harbor Symp.Ouat.Biol.,52,915-924,1987.

[2] K. Yura,S. Tomoda and M. Go., Repeat of a helix-turn-helix module in DNA-binding proteins Protein Engineering Vol.6 no.6 621-628,1993.

[3] Walter Gilbert, Why genes in pieces, Nature Vol.271,501,1978.

[4] W.Gilbert, The Exon Theory of Genes, Cold Spring Harbor Symp.Ouat.Biol.,52,901-905,1987.

[5] 中村春木, 有坂文雄, タンパク質のかたちと物性, 共立出版.

[6] Peter A. Kollman,Scott J. Weiner,David A. Case, A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins, J.Am.Chem.Soc.,106,765-784,1984.

[7] Peter A. Kollman,Wendy D. Cornell,Piotr Cieplak, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, J.Am.Chem.Soc.,117,5179-5197,1995.

[8] L.Verlet, Coumputer Experiments on Classical Fluids.I.Thermodynamical Properties of Lenanrd-Jones Molecules, Phys.Rev.,159,98,1967.

[9] 阿久津達也, 浅井 潔, 矢田 哲士, バイオインフォマティクス, 医学出版 2000.

[10] Henikoff,J.G. and Henikoff,S. Automated assmbly of protein blocks for database searching, Nucleic Acids Research,19,6565-6572,1991.

# A 付録

モジュールの割り当て

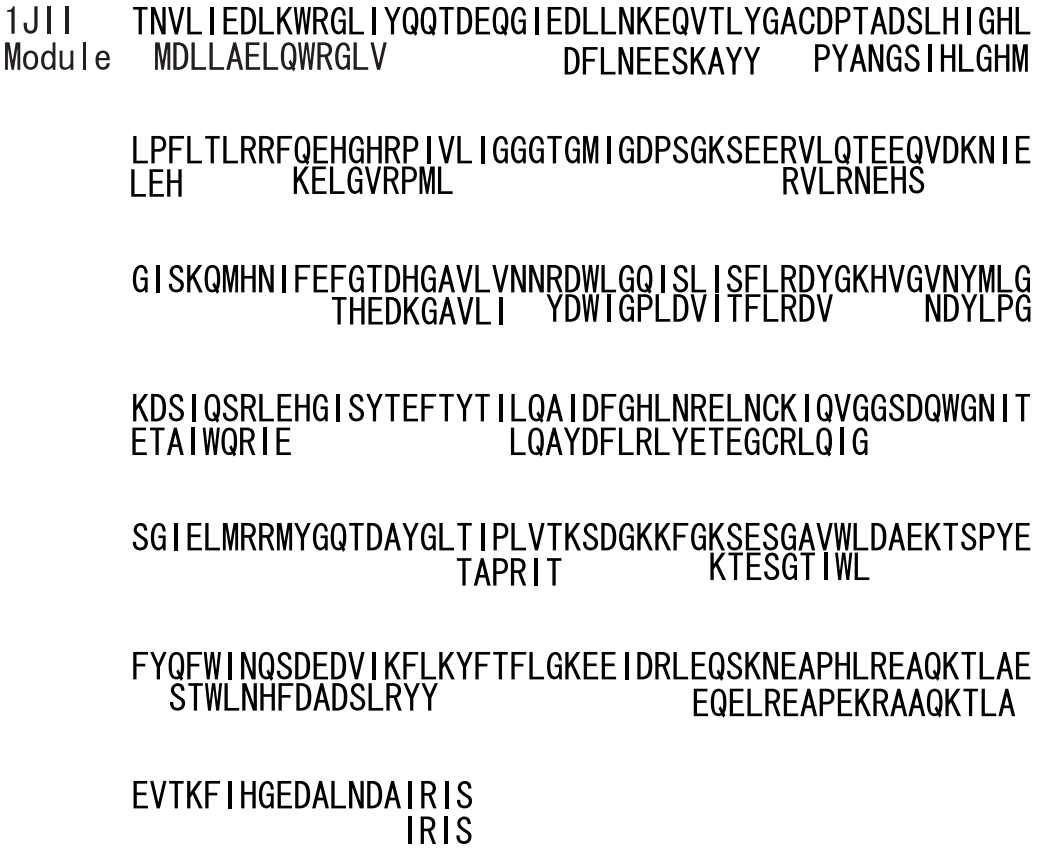


図 15: モジュールの割り当てた結果