

博士論文

# タンパク質の動的構造解析に関する研究

応用生命工学専攻  
生物情報工学研究室

西村麻彦

## 目次

序論	1
第1部 二次構造予測に基づいた拘束 MD シミュレーション	3
第1章 背景	3
第2章 二次構造拘束の方法	5
2.1 二次構造予測プログラム：PSIPRED	5
2.2 拘束エネルギー	5
2.2.1 主鎖二面角空間における予測結果の分布	5
2.2.2 確率から拘束エネルギーの導出	9
2.2.3 拘束エネルギーマップの設計	9
2.3 考察	14
第3章 拘束エネルギーの評価	15
3.1 目的	15
3.2 材料・方法	15
3.3 結果	17
3.3.1 各拘束における主鎖二面角の分布の比較	17
3.4 考察	21
第4章 拘束による folding の評価	22
4.1 目的	22
4.2 材料・方法	22
4.3 結果	25
4.3.1 各拘束におけるエネルギー分布とエネルギー遷移	25
4.3.2 各拘束における RMSD 遷移	28
4.3.3 各拘束におけるクラスタリング結果	32
4.3.4 各拘束における Gly7 の主鎖二面角	37
4.4 考察	40
第5章 All- $\alpha$ タンパク質	41
5.1 目的	41
5.2 材料・方法	41
5.3 結果	42
5.3.1 エネルギー分布と遷移	42
5.3.2 RMSD	43
5.3.3 クラスタリング	45
5.4 考察	46

第 6 章	All- $\beta$ タンパク質.....	47
6.1	目的.....	47
6.2	材料・方法.....	47
6.3	結果.....	48
6.3.1	エネルギー分布と遷移.....	48
6.3.2	RMSD.....	50
6.3.3	クラスタリング.....	52
6.4	考察.....	53
第 7 章	$\alpha$ + $\beta$ タンパク質.....	54
7.1	目的.....	54
7.2	材料・方法.....	54
7.3	結果.....	55
7.3.1	エネルギー分布と遷移.....	55
7.3.2	RMSD.....	57
7.3.3	クラスタリング.....	58
7.4	考察.....	61
第 8 章	結論.....	62
<b>第 2 部</b>	<b>Disorder の定量化.....</b>	<b>63</b>
第 1 章	背景.....	63
第 2 章	定量化の方法.....	66
2.1	電子密度マップの作成.....	66
2.2	Order Degree の計算.....	66
2.2.1	Raw Order Degree.....	67
2.2.2	B-factor による補正.....	69
2.2.3	総電子数による補正.....	71
第 3 章	解析.....	72
3.1	各アミノ酸の Order Degree.....	72
3.2	各二次構造の Order Degree.....	77
3.3	Accessible Surface Area と Order Degree.....	81
3.4	リガンド周囲の Order Degree.....	90
3.5	低 Order Degree 領域の配列解析.....	95
第 4 章	結論.....	103
<b>総括</b>	<b>104</b>	
<b>謝辞</b>	<b>105</b>	
<b>参考文献</b>	<b>106</b>	

# 序論

本論文「タンパク質の動的構造解析に関する研究」は、タンパク質の「動き」に焦点を当てて行ってきた研究について書かれたものである。

タンパク質の主な働きは、ある特定の構造になり、その構造特有の機能を果たすことである。そのため、これまでに多くのタンパク質の構造が、X線結晶構造解析やNMRなどの実験手法を用いて決定されてきた。そして、決定された構造を観察することにより、そのタンパク質の機能が解明されてきた。

しかし、タンパク質が機能を発揮する時の構造は明らかになってきたが、その構造が出来上がるまでの過程は依然分からないままである。タンパク質は、20種類のアミノ酸が連なってできた一本の鎖として合成され、そのアミノ酸配列によって定められた天然構造へとfoldする。そして、そのアミノ酸配列が似ているタンパク質は同じような構造になることが一般的であるが、中には環境のちょっとした変化や、少しのアミノ酸残基の違いで全く違う構造になる事もある。また近年になって、そのアミノ酸鎖単体では特定の構造を持たず、ただ揺らいでいるだけなのだが、ある特定のターゲットが近付いてきた時のみにある構造へと変化して、その機能を発揮するというdisorderタンパク質も発見された。これらの事は、タンパク質が動いている過程のどこかに何らかの要因が働いた結果であると考えられ、その事を理解するにはタンパク質の動きに関する情報が不可欠である。

その動きに関する情報収集の際、役に立つのが計算機である。これまで、実験では、Φ値解析による遷移状態における構造を解明する研究や、光学装置と急速混合装置の組み合わせによってミリ秒以下の単位で吸光度の変化を記録し、その変化からタンパク質の動きを解明する研究が行われてきた。しかし、これらの実験手法ではタンパク質の大まかな動きしか観測できない。そこで、原子レベルで動きを観測することができる分子動力学(MD)シミュレーションを用いて、タンパク質の動きを解析する実験が行われるようになってきた。また計算機の別の使い方として、これまでに行われてきたタンパク質実験の結果から動きに関する大量のデータを統計解析することで、動きに関する理解を深めることも可能である。

このような背景を受けて、筆者はMDシミュレーションによる天然構造の探索(第1部)とdisorderに関する統計解析(第2部)の2つの実験を行った。

第1部では、二次構造予測のデータを使ってMDシミュレーションを行い、天然構造へとfoldさせることを試みた。現在のところ、タンパク質が伸びきった鎖の状態(伸展構造)から、何の工夫もなくMDシミュレーションを走らせただけでは、天然構造へとfoldすることは殆ど無い。この問題を克服するためには何らかの情報が必要で、本論文では二次構

造予測の情報による拘束を加えて MD シミュレーションをすることでタンパク質の天然構造を見つけ出すことを試みた。

第 2 部では、**disorder** に関する研究を行った。**disorder** もタンパク質の動きに起因している現象であり、近年研究が盛んになってきているが、現在の **disorder** の定義には問題がある。そこで、X 線結晶構造解析データから得た電子密度に基づいて **disorder** を再定義することを試みた。その後、新たに定義した **disorder** のデータから **disorder** に関する知見を得るために、様々な視点から統計解析を行った。

# 第1部 二次構造予測に基づいた拘束 MD シミュレーション

## 第1章 背景

MD シミュレーションとは、エネルギー $E$ という力場を与え、各原子の時間発展を計算する手法である。力場とは、結合長、結合角、二面角、静電相互作用などの強さを規定するパラメータである。そして、そのエネルギーを小さくしようと原子は動いていく。それがMD シミュレーションの基本原理である。そして、タンパク質分子を使ってMD シミュレーションを行えば、原理的には自由エネルギー最小構造が天然構造となるはずである。

しかし、タンパク質の天然構造をMD シミュレーションで発見することは困難である。困難である理由として、主に2つの理由が挙げられる。1つ目の理由は、力場の精度が良くないということ、そして2つ目の理由は、タンパク質がとり得る構造数が膨大であるということである。

第1の理由である「力場の精度が良くない」ことにより、現実では最も安定であるとされているタンパク質の天然構造が、シミュレーションでは最安定構造として評価されないという問題がある。これまでに、タンパク質用の多くの力場が開発されてきた。本研究で使用しているMD プログラムパッケージであるAMBER9[1]で利用されている主な力場はff94[2]、ff96[3,4]、ff99[5]、ff03[6,7]といったものである。これらの力場にはそれぞれ特徴があり、ff94、ff99そしてff03は $\alpha$ -helixが形成されやすく、ff96は $\beta$ -sheetが形成されやすい傾向がある[1]。このように用いた力場によって、最安定となる構造が異なっていることがある。

第2の理由である「タンパク質がとり得る構造数が膨大である」ことにより、計算時間に限りがあるシミュレーション中に、目的としているタンパク質の天然構造を発見できないという問題がある。タンパク質を構成しているアミノ酸残基の数が多くなっていくにつれて、とり得る構造の数は増えていき、シミュレーションでそれらの構造を網羅することは現在の計算機性能では不可能である。

そこで本研究では、タンパク質の主鎖二面角( $\phi, \psi$ )に拘束エネルギー $E_{rst}$ を加えてMD シミュレーションを行うことを試みた。主鎖二面角はタンパク質の形を決定する主な要因であり、二次構造を形成することや、多くの構造をとり得るということも、個々のアミノ酸残基の主鎖二面角の組み合わせによるものであると考えられる。そのため、本来の力場

で規定されているエネルギー  $E$  に主鎖二面角に拘束エネルギー  $E_{rst}$  を加えることで、「力場が正しくない」ことによる誤った二次構造形成が補正されるとともに、主鎖二面角がとる値に偏りが生じることで主鎖二面角の組み合わせの数が減り、とり得る構造数を減らすことができる考えた。

そして、その主鎖二面角の拘束エネルギー  $E_{rst}$  の値を決めるために、二次構造予測プログラム PSIPRED[78,79]を使用した。各アミノ酸残基の主鎖二面角が最終的にどのような値をとるのか?ということを前もって知ることは出来ないので、PSIPRED の二次構造予測の結果をもとに、そのアミノ酸残基がとると予測される主鎖二面角の値の確率  $P(A|S)$  を求め、その  $P(A|S)$  の自然対数をとる ( $\ln P(A|S)$ ) ことで拘束エネルギー  $E_{rst}$  を見積もった。この  $P(A|S)$  をベイズの定理を用いて考えてみると、式は

$$P(A|S) = \frac{P(S|A)}{P(S)} P(A)$$

で、 $A$  を主鎖二面角、 $S$  を二次構造予測の結果とした場合、 $P(A)$  はあるアミノ酸残基の主鎖二面角の事前分布で、何の情報も無い状態では、シミュレーションに用いる力場に従うと考える。 $P(S|A)$  は、ある主鎖二面角の値をとっている時、二次構造予測  $S$  と返す確率である。 $P(S)$  は二次構造予測の結果  $S$  と返す確率で、この値は構造に依存しない、つまり、シミュレーション中に変化しないので、この項をエネルギーとして加えても、微分して得られる力が常に 0 となることから、無視して構わない。そして、左辺の  $P(A|S)$  は、二次構造予測  $S$  とされた時の主鎖二面角  $A$  である確率、いわゆる事後分布である。この  $P(A|S)$  が拘束エネルギー  $E_{rst}$  の基になる。

まとめると、二次構造予測の結果を受けて、ある主鎖二面角の値が出現する確率を変化させるということである。または、予測された二次構造を形成しようとする協調的な力が加わると考えても良い。

まずは、 $P(A|S)$  を求めるため、正解の分かっているタンパク質のデータセットを用いて、PSIPRED の二次構造予測分布  $P(S|A)$  を調べることから始めた。

## 第2章 二次構造拘束の方法

### 2.1 二次構造予測プログラム：PSIPRED

PSIPRED[78,79]は、アミノ酸配列を入力すると、各アミノ酸残基がとると予測される二次構造とその予測の信頼度を出力する (図 2.1)。

```
Key
Conf: Confidence (0=low, 9=high)
Pred: Predicted secondary structure (H=helix, E=strand, C=coil)
AA: Target sequence

# PSIPRED HFORMAT (PSIPRED V3.0)

Conf: 975561349999999999973338937998617654089999999520036758983279
Pred: CCCCCCCHHHHHHHHHHHHCCCCCEEEEECEEECCCCCCCCCHHEEEEEEECCC
AA:  ALGSNLNTPVEQLHAALKAIQLSNTHLVTSSFYKSKPLGPQDQPDYVNAVAKIETELS
      10      20      30      40      50      60
```

図 2.1 psipred の出力結果の例

予測される構造は、ヘリックス (H)、シート (E)、コイル (C) の3状態である。そして、その予測の信頼度が 0~9 の 10 段階で示される。以後、予測された構造(C, H, E)と信頼度(0~9)を組み合わせせて表記する (例えば、予測が C で信頼度が 5 なら、「C5」のように)。

### 2.2 拘束エネルギー

#### 2.2.1 主鎖二面角空間における予測結果の分布

まず、主鎖二面角を拘束するためのエネルギーの値を決定するため、主鎖二面角空間において、PSIPRED の各予測がどのように分布しているか調べた。調査のために使用したタンパク質のデータセットは、SCOP 1.71[82]の fold representative から膜タンパク質、NMR 構造、RESSEQ と ATOM の不一致があるものを除外した 587 個のタンパク質であり、そ



の総残基数は 137640 である。これらのタンパク質のアミノ酸配列を入力として PSIPRED から得られた二次構造予測と信頼度の結果と実際の主鎖二面角の値 ( $\phi, \psi$ ) を使って、主鎖二面角空間を  $60^\circ$  刻みに区切ってできた 36 区画における各予測結果の出現確率を計算した。

その結果、図 2.2 のようになった。このグラフについて説明すると、タイトルの  $\phi, \psi$  はその区画の範囲を表しており、括弧の中のデータ数はその区画内の主鎖二面角の値を持ったアミノ酸残基数である。そして、その区画における予測された二次構造とその信頼度の組み合わせの出現割合を積み重ね棒グラフで示してある。区画  $-180^\circ < \phi \leq -120^\circ, 120^\circ < \psi \leq 180^\circ$  を例にとると、E9 の割合は 0.25 とその区画の 4 分の 1 を占めており、また、H は殆ど存在していないことが分かる。そして、この割合から拘束エネルギー  $E_{rst}$  が計算される。

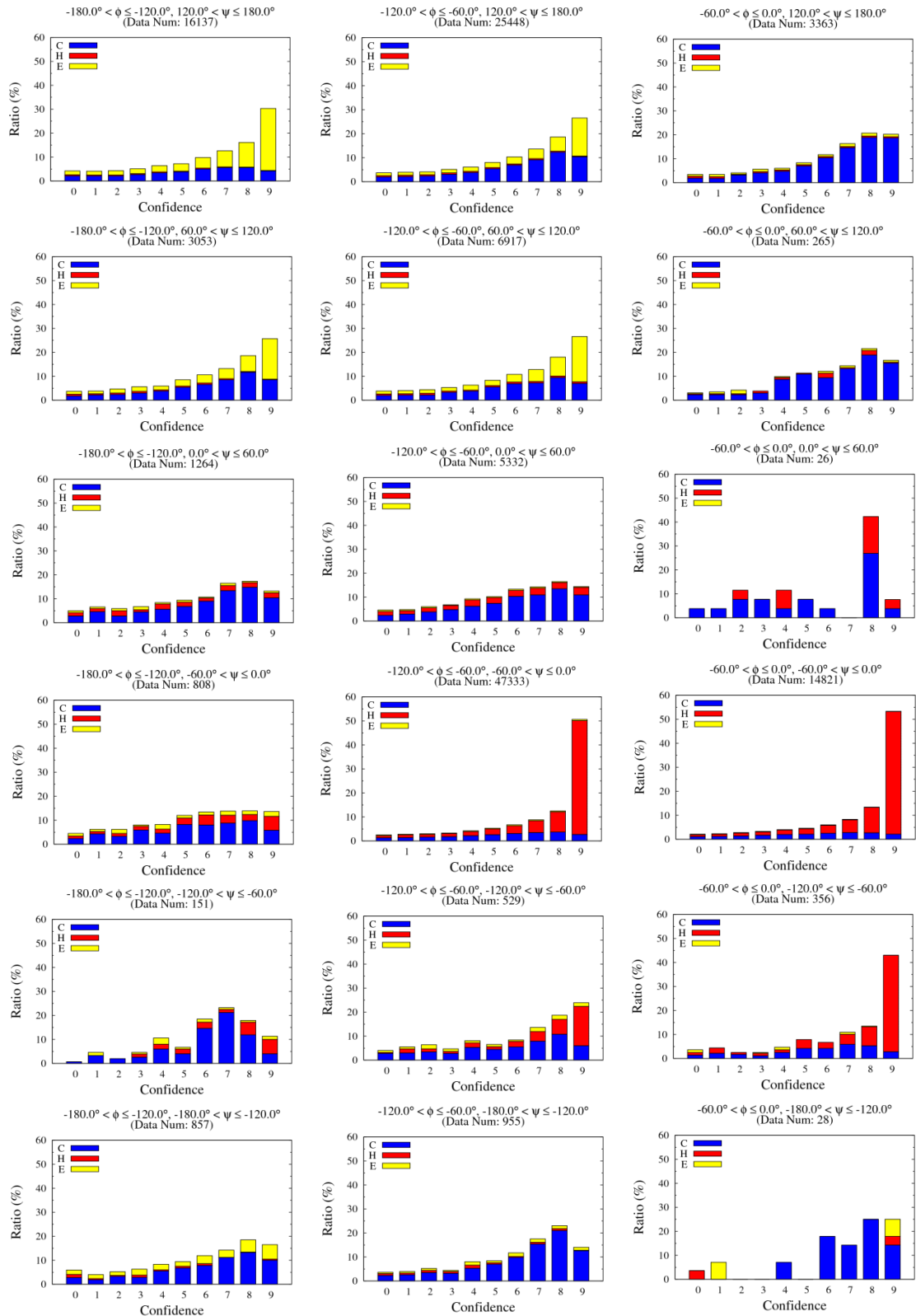


図 2.2 各区画の予測出現確率 ( $\phi \leq 0^\circ$ )

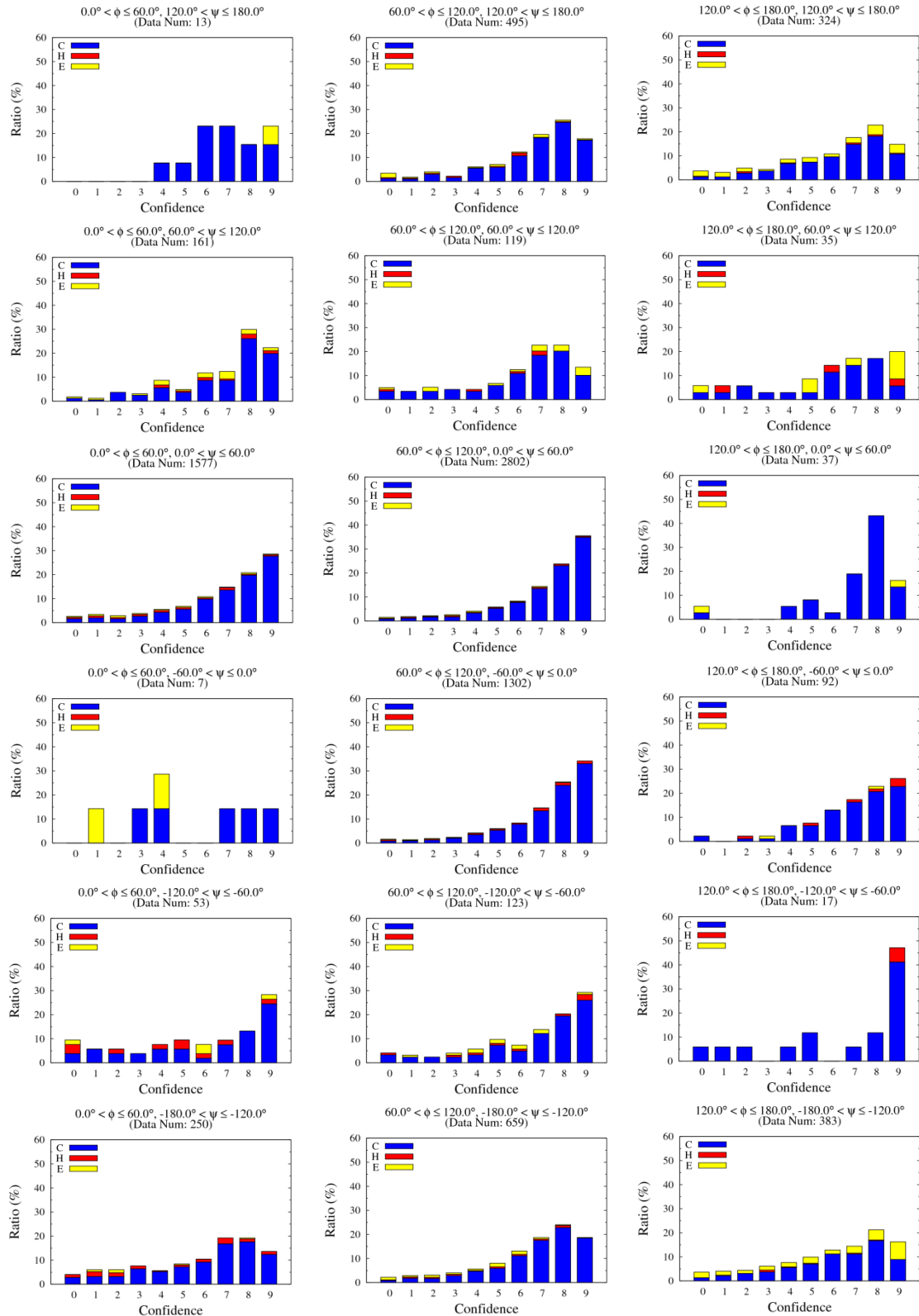


図 2.2 各区画の予測出現確率 ( $\phi > 0^\circ$ )

## 2.2.2 確率から拘束エネルギーの導出

得られた出現確率  $P(i,j)$  を用いて、各区画における、各予測の拘束エネルギー  $E_{rst}(i,j)$  を以下の式で求めた。 $i$  は  $\phi$  の区画インデックス ( $i=1\sim6$ )、 $j$  は  $\psi$  の区画インデックス ( $j=1\sim6$ ) とする。

$$E_{rst}(i,j) = -RT \ln P(i,j)$$

$R$  は気体定数で  $1.9876 \times 10^{-3}$  kcal/mol、 $T$  は絶対温度(K)である。今回はシミュレーションを 400 K で行うため  $T=400$  として計算している。なお、 $P=0$  である区画において  $E_{rst}$  が  $\infty$  になってしまうことを避けるため、その区画における  $P$  を、その予測において  $P > 0$  である他の区画の最小値とした。

## 2.2.3 拘束エネルギーマップの設計

各区画における拘束エネルギーは求めたが、そのまま適用しようとすると、各区画間で拘束エネルギーが不連続であり、MD シミュレーションを行う上で不具合が生じる。シミュレーションで使用できるようにするためには、各区画の  $E_{rst}(i,j)$  を滑らかに繋ぐ必要がある。

そこで、 $E_{rst}(i,j)$  を滑らかに繋ぐためコサイン関数を使った以下のような式を用いた。

$$h \times 0.5 \times [1 + \cos(\eta \times d\theta + 180)]$$

$0.5 \times [1 + \cos(\eta \times d\theta + \pi)]$  の部分が  $0\sim1$  へと滑らかに変化させる役割を持つ。 $h$  は隣接する  $E_{rst}$  との差を表す。 $\eta$  は周期、 $d\theta$  は変化を始める起点からの距離を表す。今回は、 $30^\circ$  の間隔で変化することを想定しているため、 $\eta = 180/30 = 6$  で、 $0^\circ < d\theta < 30^\circ$  となる。

なお、区画の中心を表す数字として  $C(n) = -180 + (n-0.5) \times 60$  を定義する。また、 $E_{rst}(i,j)$  の  $i,j$  は  $1\sim6$  の値をとると定義したが、二面角は  $-180^\circ$  と  $180^\circ$  を境に循環しているため、 $i,j$  が 1 の時  $-1$  された場合は  $0$  ではなく  $6$  となり、 $i,j$  が 6 の時  $+1$  された場合は  $7$  ではなく  $1$  に戻る、というように循環しているとする。最終的に、以下のように拘束エネルギー  $E_{rst}(\phi,\psi)$  マップを設計した。

1. 区画の中心から±15°の範囲内 ( $\alpha(i)-15 \leq \phi \leq \alpha(i)+15$  かつ  $\alpha(j)-15 \leq \psi \leq \alpha(j)+15$ ) なら、

$$E_{rst}(\phi, \psi) = E_{rst}(i, j)$$

2.  $\phi$ のみが区画の中心から±15°の範囲外 ( $\alpha(i)+15 < \phi < \alpha(i+1)-15$  かつ  $\alpha(j)-15 \leq \psi \leq \alpha(j)+15$ ) なら、

$$E_{rst}(\phi, \psi) = E_{rst}(i, j) + \{E_{rst}(i+1, j) - E_{rst}(i, j)\} \times 0.5 \times [1 + \cos(6 \times \{\phi - (\alpha(i)+15)\} + 180)]$$

3.  $\psi$ のみが区画の中心から±15°の範囲外 ( $\alpha(i)-15 \leq \phi \leq \alpha(i)+15$  かつ  $\alpha(j)+15 \leq \psi \leq \alpha(j+1)-15$ ) なら、

$$E_{rst}(\phi, \psi) = E_{rst}(i, j) + \{E_{rst}(i, j+1) - E_{rst}(i, j)\} \times 0.5 \times [1 + \cos(6 \times \{\psi - (\alpha(j)+15)\} + 180)]$$

4.  $\phi, \psi$ ともに区画の中心から±15°の範囲外 ( $\alpha(i)+15 < \phi < \alpha(i+1)-15$  かつ  $\alpha(j)+15 \leq \psi \leq \alpha(j+1)-15$ ) なら、

$$E_1 = E_{rst}(i, j) + \{E_{rst}(i+1, j) - E_{rst}(i, j)\} \times 0.5 \times [1 + \cos(6 \times \{\phi - (\alpha(i)+15)\} + 180)]$$

$$E_2 = E_{rst}(i, j+1) + \{E_{rst}(i+1, j+1) - E_{rst}(i, j+1)\} \times 0.5 \times [1 + \cos(6 \times \{\phi - (\alpha(i)+15)\} + 180)]$$

$$E_{rst}(\phi, \psi) = E_1 + (E_2 - E_1) \times 0.5 \times [1 + \cos(6 \times \{\psi - (\alpha(j)+15)\} + 180)]$$

以上の式によって作成された拘束エネルギーマップを図 2.3、2.4、2.5 に示す。

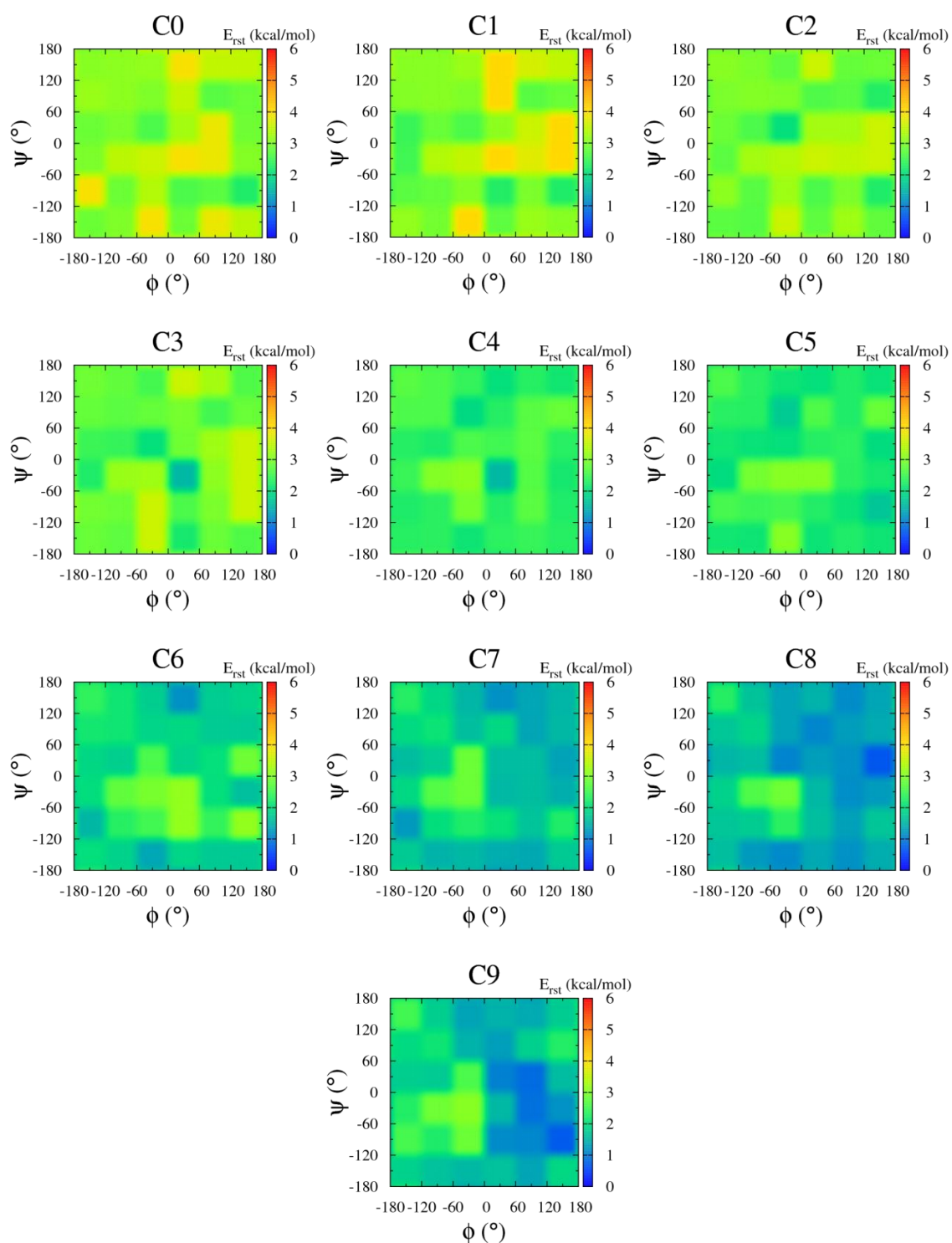


図 2.3 予測 C0~9 の拘束エネルギーマップ

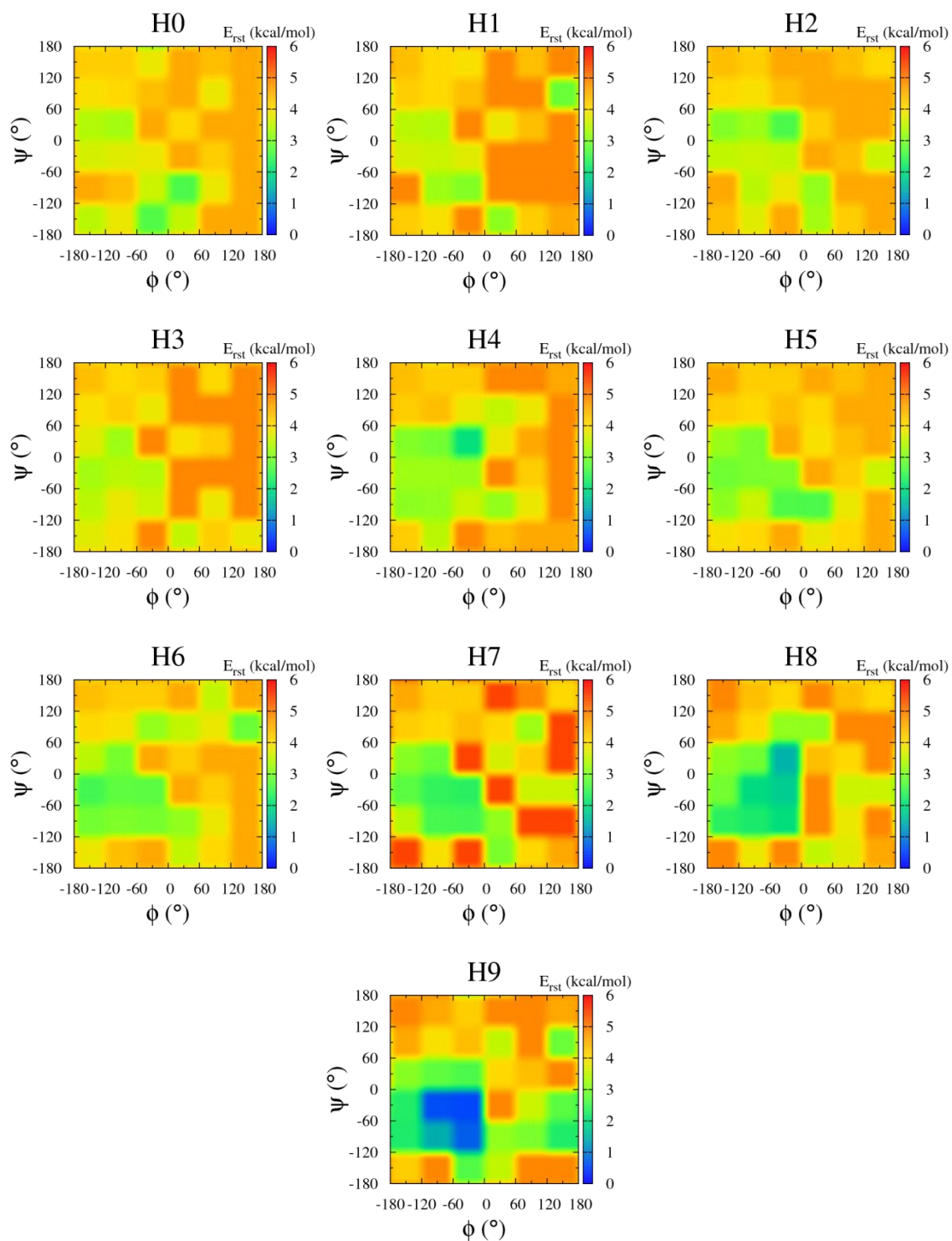


図 2.4 予測 H0~9 の拘束エネルギーマップ

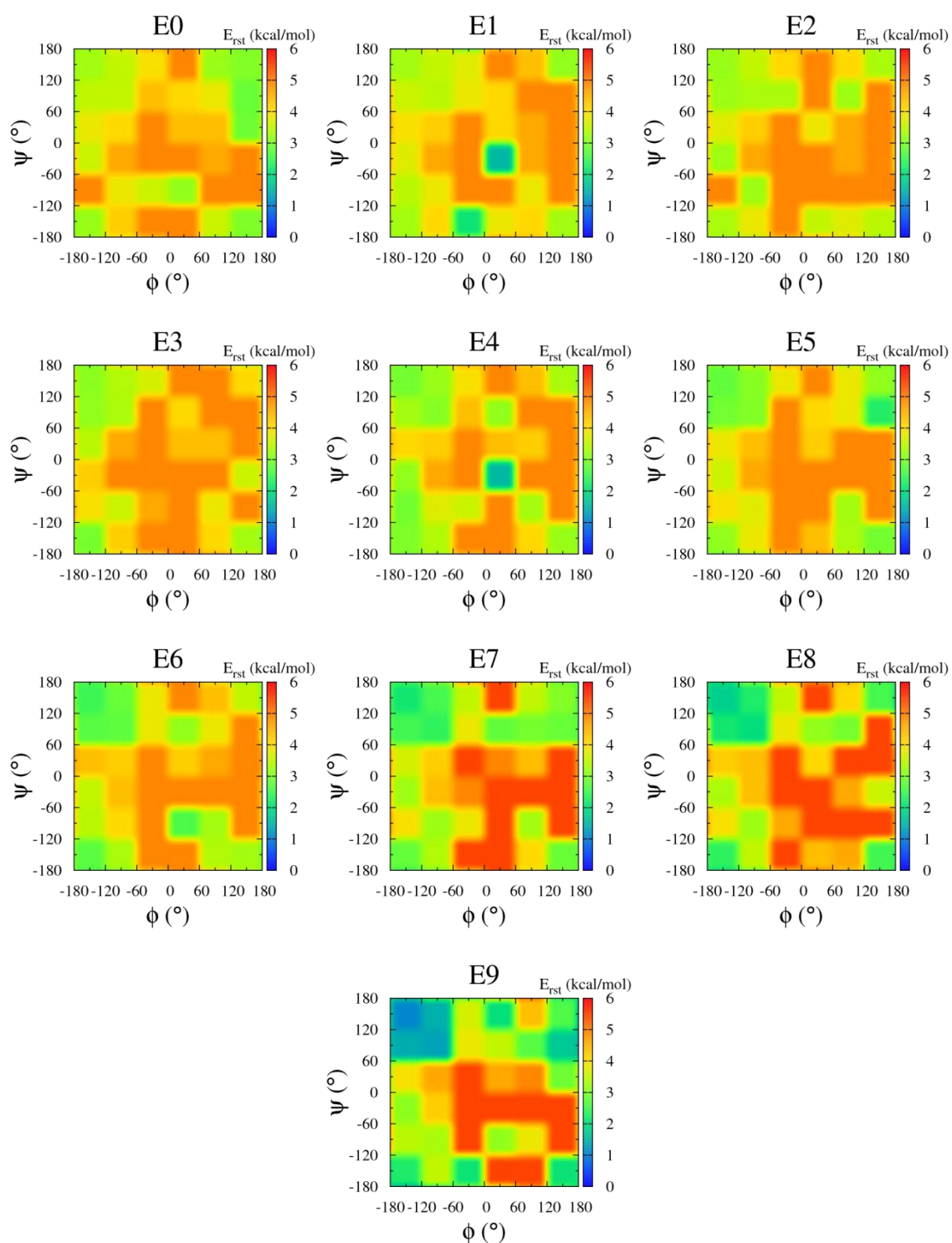


図 2.5 予測 E0~9 の拘束エネルギーマップ



## 2.3 考察

各二次構造予測における完成した拘束エネルギーマップについて考察する。

まず、Cに関するマップを見ると、信頼度が上がるにつれて $\phi > 0^\circ$ に低エネルギー領域が出来ていることが分かる。この結果は、その領域の二面角の値をとっている限り二次構造は形成できないという背景に起因していると考えられる。また、Cと予測された場合、全信頼度において、シート領域に比べてヘリックス領域の方が少し高いエネルギーとなっている。この理由は、ヘリックス領域の二面角の値である時にヘリックスを形成する割合が、シート領域の二面角の値である時にシートを形成する割合より高いという背景がある、もしくはPSIPREDにとってシートと予測する確率がヘリックスより低いためにシート領域の二面角の値である残基がCと予測される傾向にあるという可能性が考えられる。総じてみると、Cとして拘束がかけられると、ヘリックス領域に留まる確率が少し下がることになると言える。

次に、Hに関するマップを見ると、予想通りヘリックス領域が低エネルギー領域となっており、かつ信頼度が上がるにつれてよりエネルギーが低くなっていることが分かる。また、エネルギーが高い領域は $\phi > 0^\circ$ の部分で、シート領域はそれに比べればやや低いのでシート領域にも幾分か留まりやすいのではないかと考えられる。

最後に、Eに関するマップを見ると、シート領域に低エネルギー領域を持ち、それ以外は概ね高くなっている。また、E1とE4において、区画 $0^\circ < \phi \leq 60^\circ$ ,  $-60^\circ < \psi \leq 0^\circ$ の拘束エネルギーが低い値をとっている。図2.2を見ると分かるように、この区画はデータ数が7つしかないため、1つのデータが占める割合が大きくなり、その結果として拘束エネルギーの値は小さくなっていると考えられる。しかし、データ数が少ない区画は立体障害によってエネルギーが非常に高くなるため、2~3 kcal/mol程度の安定化ではその効果を打ち消すことはできない。そのため、この区画がシミュレーション中に多く見られるようになることは無いと考える。

全体的に見て、概ね予想通りの拘束エネルギーマップが出来上がった。この拘束エネルギーを用いることで力場が補正され、正しい二次構造分布が得られることが期待される。

## 第3章 拘束エネルギーの評価

### 3.1 目的

前章で作成した拘束エネルギーを実際にタンパク質分子の MD シミュレーションに適用し、どのように主鎖二面角が分布するか調査、そして拘束エネルギーを評価する。

### 3.2 材料・方法

シミュレーション対象は、アラニン残基をアセチル基と N-メチル基でキャップした、Ace-Ala-NMe (Alanine-dipeptide) というペプチド鎖である (図 3.1)。

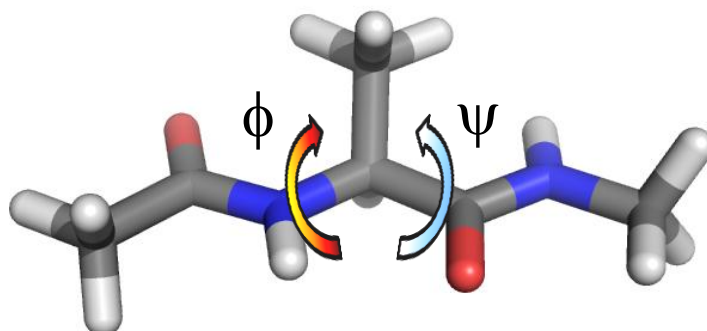


図 3.1 Alanine-dipeptide (stick 表示)

シミュレーション条件は、力場は ff03 に各二次構造予測に対する拘束エネルギーを適用して補正されたもの、温度は 400 K、溶媒は陰的溶媒モデルの GB/SA OBC type II[8,9]、水素を含む結合を剛体として扱い[10,11]、時間刻みは 2 fs で 100 ns になるまでシミュレーションを行った。500 step (1 ps)毎に構造の座標を記録し、最終的に 10 万の構造を得た。

以後の解析は、これらの構造を用いて行った。なお、初期構造は伸展構造から始め、5000 step の最小化を行った後、1 ns の平衡化を行い、その後、先述した 100 ns の解析対象となるシミュレーションを行った。

なお、本シミュレーションで使用している陰的溶媒モデルとは、水分子を陽に配置することなく溶媒効果を得るための手法である。溶媒である水分子を溶質の周囲に陽に配置すると、計算量が増え、シミュレーション時間が大幅に減ってしまう。そこで、溶媒を連続体として近似することで計算量を減らし、かつ溶媒効果を取り入れる方法として陰的溶媒モデルが開発された。その結果、長時間のシミュレーションが可能となった。本研究では天然構造の探索を目標としており、その目標達成のためには出来るだけ長くシミュレーションを行う必要がある。そのため、以後のシミュレーションにおいても陰的溶媒モデルを使用していく。

### 3.3 結果

#### 3.3.1 各拘束における主鎖二面角の分布の比較

得られた座標から主鎖二面角 ( $\phi, \psi$ ) の値を計算し、各拘束パターンにおける主鎖二面角の分布を求めた。まずは、無拘束の場合の主鎖二面角の分布を図 3.2 に示す。

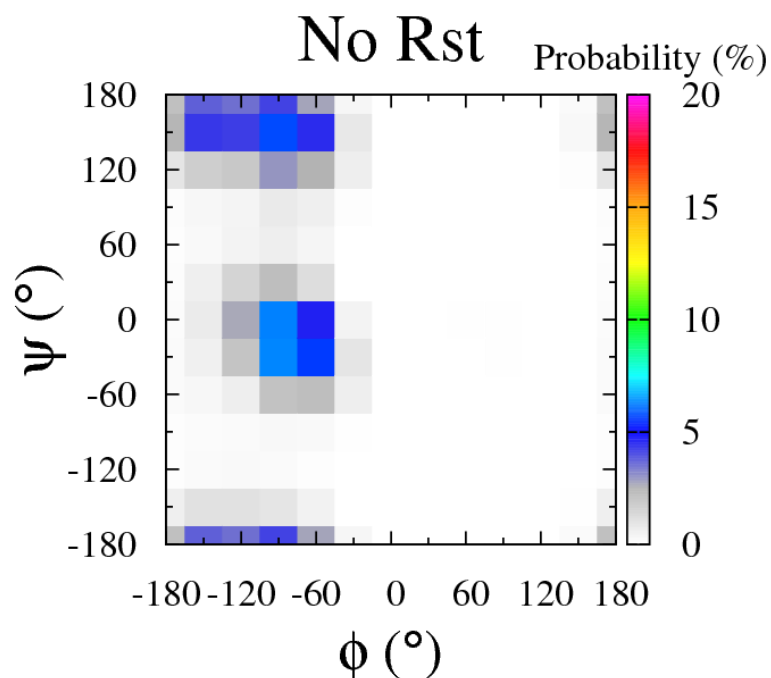


図 3.2 拘束無しの主鎖二面角分布

この図は、シミュレーション中に見られた主鎖二面角 ( $\phi, \psi$ ) の割合を、 $30^\circ \cdot 30^\circ$ の区画に区切られた空間上に示している。拘束無しの場合、主にヘリックス領域とシート領域を中心として分布しており、ヘリックス領域の方が存在割合は高めであることが分かる。

続いて、各拘束の結果を図 3.3、3.4、3.5 に示す。

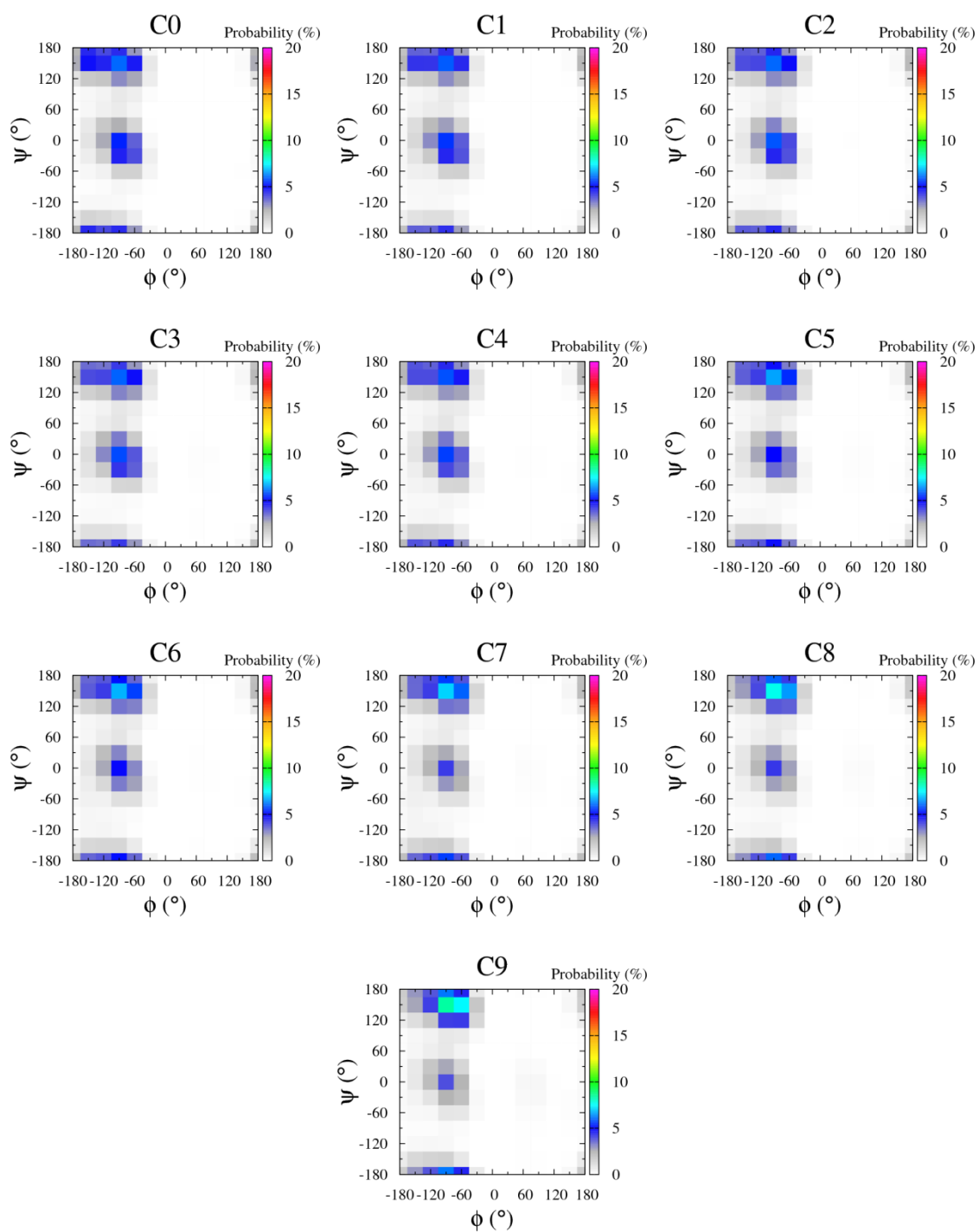


図 3.3 C0~9 の主鎖二面角分布

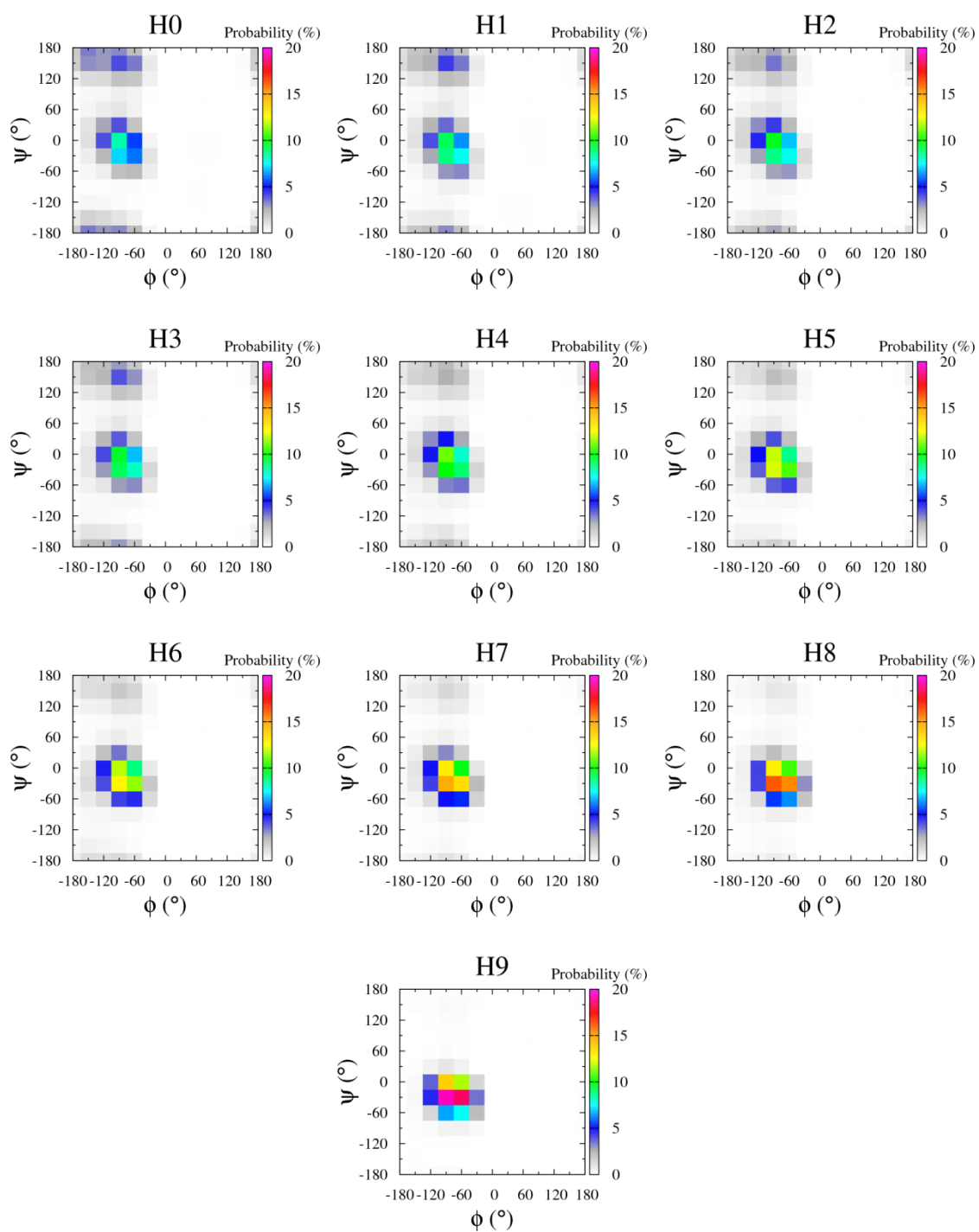


図 3.4 H0~9 の主鎖二面角分布

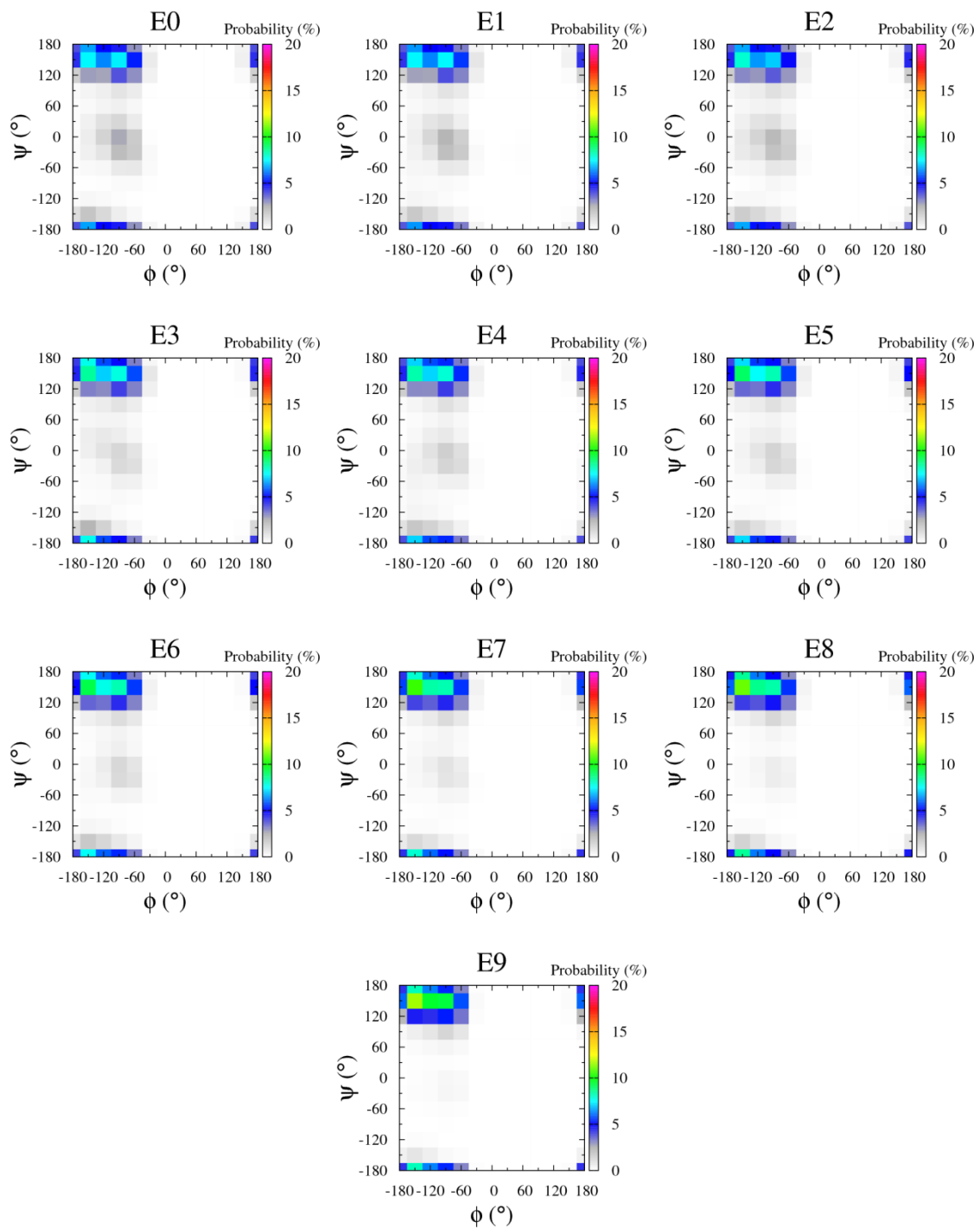


図 3.5 E0~9 の主鎖二面角分布

### 3.4 考察

各予測における主鎖二面角分布について考察する。

まず、Cに拘束された Alanine-dipeptide について見てみると、信頼度が上がるにつれて、拘束無しの場合と比べてヘリックス領域に存在する割合が減り、その分シート領域に存在する割合が増えていくことが分かる。これは、拘束エネルギーがヘリックス領域において少し高いことが影響を与えているためであると考えられる。また、C8,9においては、 $\phi > 0^\circ$  の領域における割合が少し増えていることが分かる。増える理由は拘束エネルギーマップを見れば分かるように、その領域が他の領域に比べて低エネルギーであるためである。しかし、その存在割合がヘリックス領域やシート領域より低い理由は、もともとこの領域は他の領域に比べてそれほど安定な構造でないため、拘束エネルギーによる多少の安定化効果では大きく存在確率が上げることができなかつたためと考える。

次は、Hの拘束について見てみると、信頼度が上がるにつれてヘリックス領域に存在する割合が増えていることが分かり、適切に拘束エネルギーが作用していることが確認された。また、信頼度9以外ではシート領域にも多少の存在が認められる。これは、先の拘束エネルギーマップで見たように、シート領域における拘束エネルギーがそれ程高くないことによる影響と考えられる。本来、力場 ff03+GB/SA という組み合わせはヘリックス領域を強めに安定化してしまうので、このようなヘリックスに有利な拘束を加えることでヘリックス領域しか観測できなくなるのではという懸念があった。それを回避することができたことによって、二次構造予測が外れている場合でも天然構造が形成される可能性が残ると推測する。

最後に、Eの拘束について見てみると、信頼度が上がるにつれてシート領域に存在する割合が増え、代わりにヘリックス領域の割合が減っていることが分かる。E9におけるヘリックス領域の拘束エネルギーはかなり高いが、それでもヘリックス領域に微かに存在していることから、やはりヘリックス領域がかなりの安定であることが分かる。

総じて、予想通りの結果が得られた。そして、H9やE9といった強い拘束でも他の領域に移ることができる可能性があることが、PSIPREDの予測結果が外れた場合でも天然構造へと fold する可能性を残すことになる。



## 第4章 拘束による folding の評価

### 4.1 目的

小さなタンパク質に対して、天然構造における二次構造を基に拘束を加えて MD シミュレーションを行い、正しく天然構造へと folding するか調査する。また、信頼度の違いによって folding にどのような影響を及ぼすか調べる。

### 4.2 材料・方法

シミュレーション対象は<sup>シニョリン</sup>chignolin (PDB ID: 1uao)[53]を用いた。chignolin は $\beta$ -sheet を持つ 10 残基のタンパク質 (図 4.1) である。

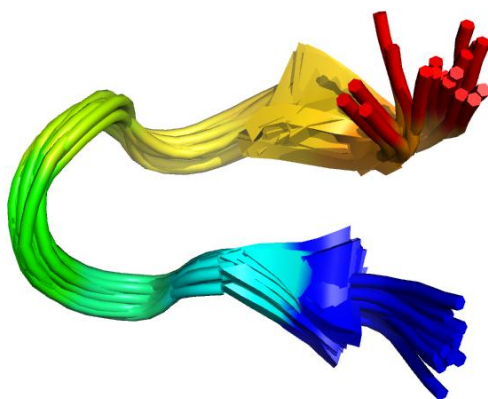


図 4.1 chignolin の天然構造

本研究では、11 パターンの状況で MD シミュレーションを行った。1 つは対照実験として、拘束無しの MD シミュレーション、残りの 10 パターンは DSSP[57]を用いて判定した天然構造における二次構造に従って拘束し (両端の残基は除く)、かつタンパク質全体を信頼度 0~9 それぞれに設定した MD シミュレーションである。

1uao のアミノ酸配列 (AA)、DSSP による判定 (DSSP) は以下の通りである。CONF は信頼度で、今回は全体を 0~9 に設定した 10 パターンでシミュレーションを行う。

1uao  
 AA: GYDPETGTWG  
 DSSP: CEECCCEEC  
 CONF: -NNNNNNNN- (N = 0~9)

ここで、DSSP の判定について説明しておく。DSSP は各残基ごとに形成している構造に応じて、表 4.1 にある 8 種類の構造に分類する。

表 4.1 DSSP による構造判定の種類

G	$3_{10}$ helix	H	$\alpha$ helix
I	$\pi$ helix	T	hydrogen bonded turn
B	isolated $\beta$ -bridge	E	extended strand
S	bend	C (空白)	coil

1uao を例にとれば、Tyr2, Asp3, Thr8, Trp9 が E と判定されている。つまり、これらの残基はシート構造を形成しているということになる。

二次構造拘束以外のシミュレーション条件は、力場として ff03、マルチカノニカル法 (探索温度: 300K ~ 400 K)、溶媒は陰的溶媒モデルの GB/SA OBC type II、水素を含む結合を剛体として扱い、時間刻みは 2 fs で 100 ns になるまでシミュレーションを行った。500 step (1 ps) 毎に構造の座標を記録し、最終的に 10 万の構造を得た。以後の解析は、これらの構造を用いて行った。なお、初期構造は伸展構造から始め、5000 step の最小化を行ってから、2 ns の平衡化を行い、その後、先述した 100 ns の解析対象となるシミュレーションを行った。

ここで、マルチカノニカル法について簡潔に説明しておく。マルチカノニカル法は、エネルギー空間を等確率で分布するようにさせる手法である。なぜ等確率で分布するようにさせる必要があるのか? という点、以下のような問題があるためである。定温条件で MD シミュレーションを行うと、エネルギーの出現分布があるエネルギー値を頂点とするベル型の分布となる。そして、温度が下がるにつれて、頂点となるエネルギーの値も下がっていく。つまり、温度が高いと高エネルギー状態をとり続け、温度が低いと低エネルギー状態をとり続ける。この性質によって問題が生じる。温度が低い (300 K 程度) 条件でシミュレーションを行っているとき、天然構造以外のエネルギーが低い準安定構造にトラップされてしまうことがある (ローカルミニマム問題)。そして、大抵の場合、この準安定構造が解かれる事なく、その形を維持したままシミュレーションが続いてしまうため、天然構造は

発見できないことになる。一方、温度が高い（400 K 程度）条件でシミュレーションを行うと、水素結合などの相互作用によって安定化しようとしても、高エネルギー状態を維持しようとするため、すぐに離されてしまう。よって、安定な構造ができる確率が少なくなる。結果として、天然構造は発見できない。そして、高すぎず、低すぎず、丁度良い温度を見つけようとするのは長時間に渡る試行錯誤を要する。これらの問題を解決するための方法の1つがマルチカノニカル法である。マルチカノニカル法は、その時々にとってのエネルギーの値によって、擬似的なエネルギーを加えることで、全てのエネルギーの出現確率を同じになるよう調節する。このようにすれば、準安定構造にトラップされ続けることや、相互作用の無い不安定な構造をとり続けることも無く、様々な安定構造へと遷移し続け、結果として天然構造を見つけることが可能になる。また、加えていた擬似的なエネルギーを取り除くことで、任意の温度での存在確率を求めること（**reweighting**）が出来ることもマルチカノニカル法のメリットである。このような利点から、これまでに多くのマルチカノニカル法を使った研究が行われている[12-52]。

## 4.3 結果

### 4.3.1 各拘束におけるエネルギー分布とエネルギー遷移

まずは、マルチカノニカル法によって、温度 300 K~400 K の範囲で適切にサンプリングが行われているかどうかを調べるため、探索されたエネルギー分布とエネルギー遷移をそれぞれ図 4.2 に示す。なお、 $\ln P$  は、そのエネルギーにおける存在確率  $P$  の自然対数をとった値である。

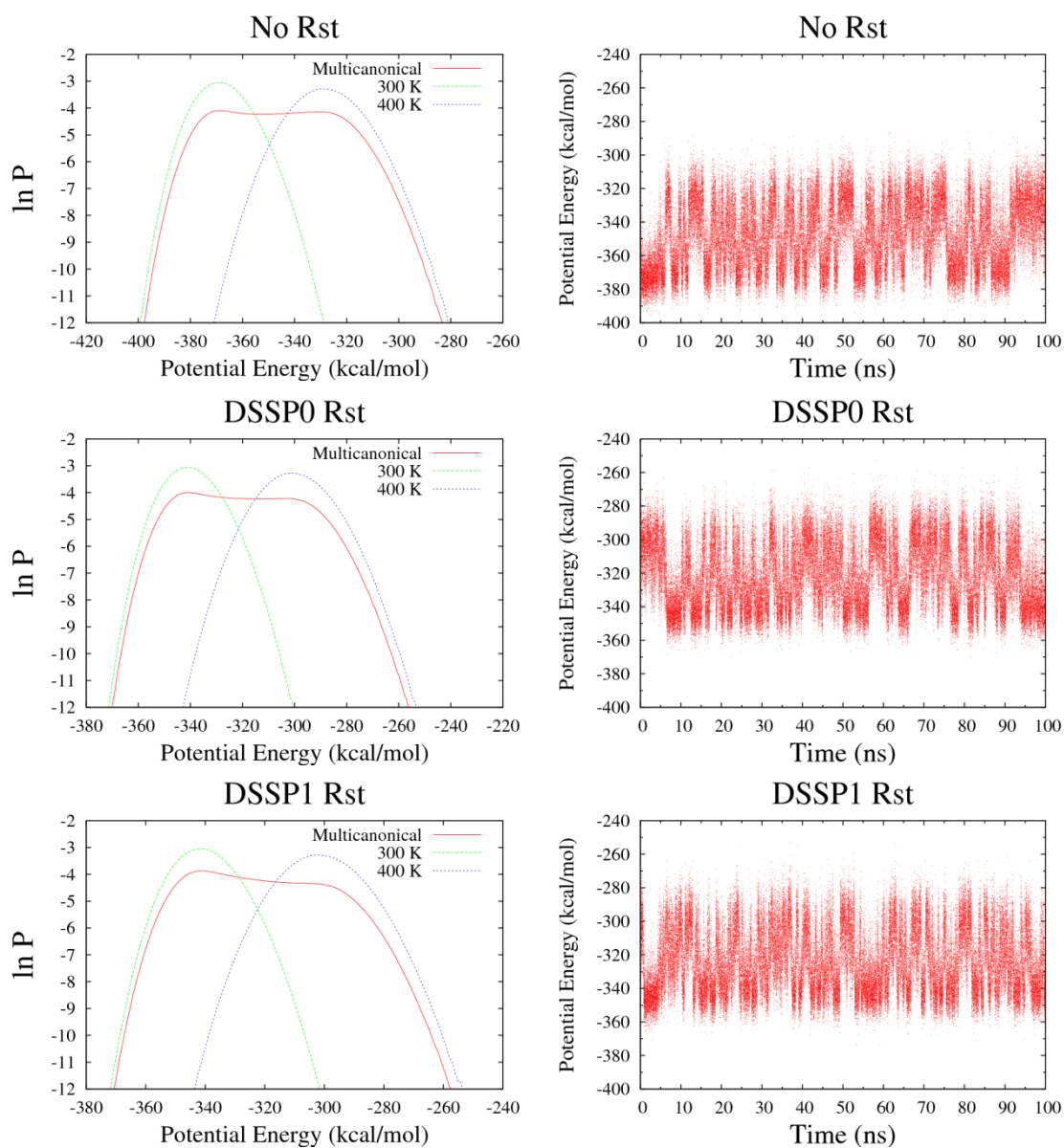


図 4.2 エネルギー分布とエネルギー遷移 (拘束無し、信頼度 0,1)

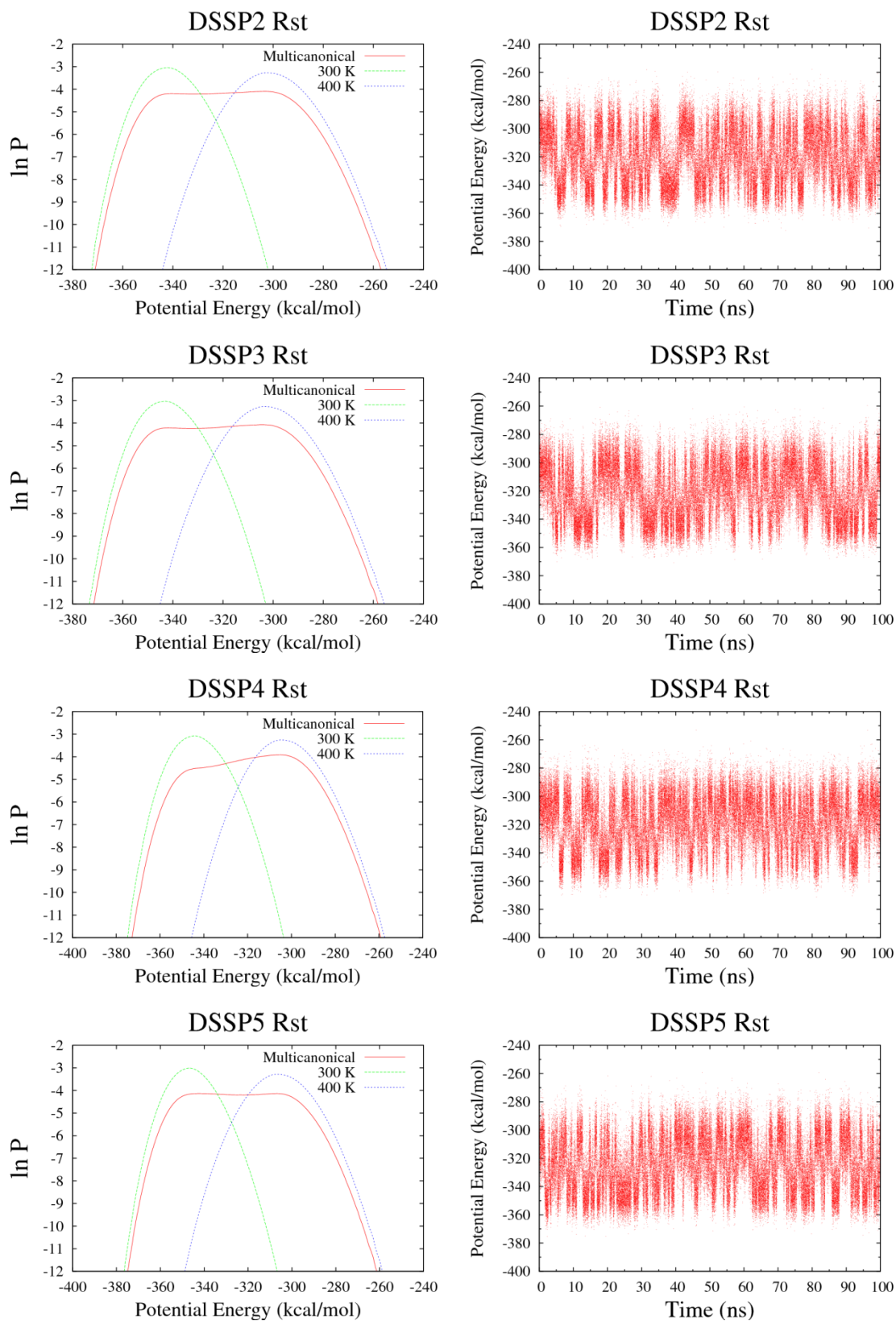


図 4.2 エネルギー分布とエネルギー遷移 (信頼度 2~5)

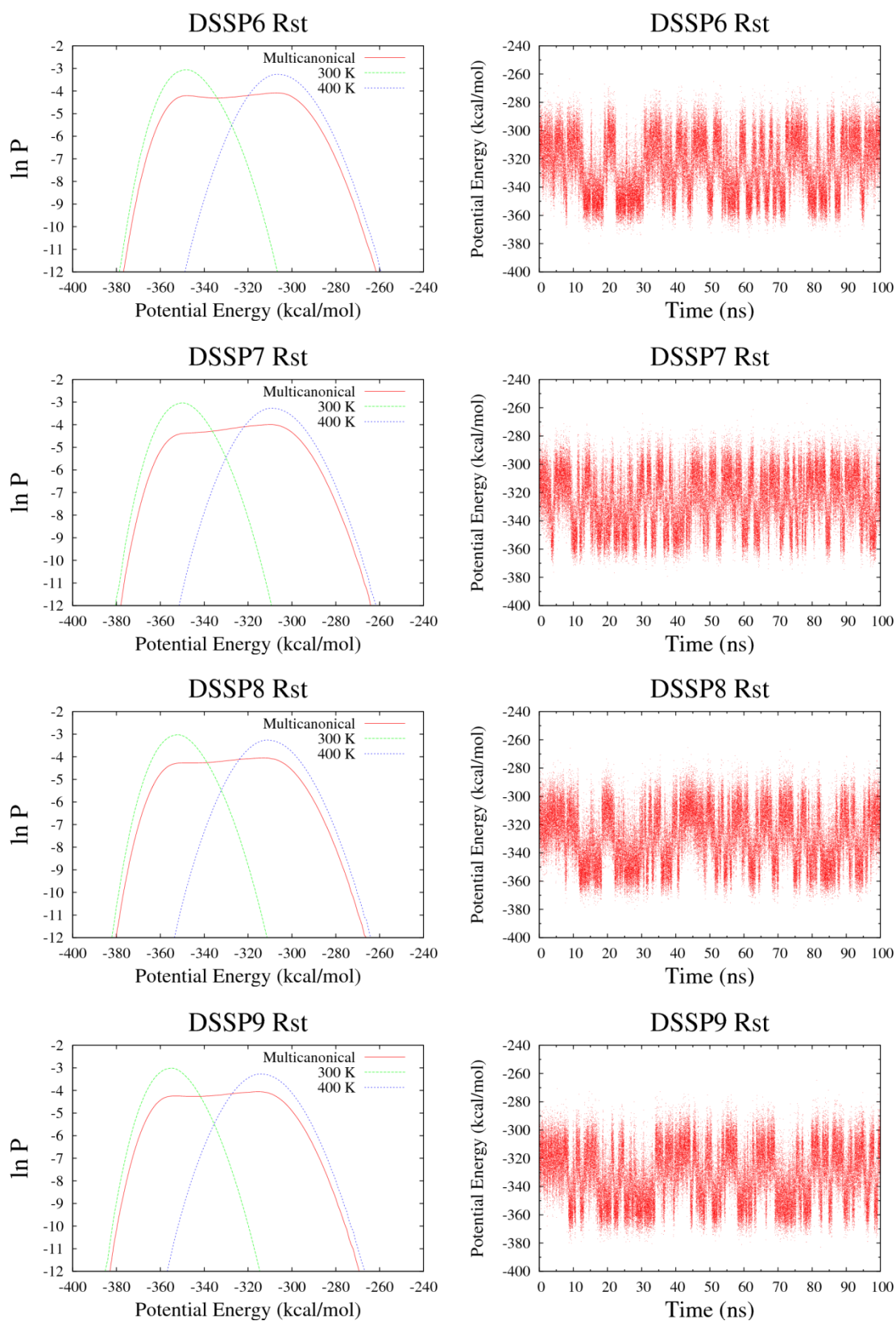


図 4.2 エネルギー分布とエネルギー遷移 (信頼度 6~9)

図 4.2 から、全パターンでのシミュレーションにおいて、エネルギー分布は 300 K~400 K をフラットに覆っており、かつエネルギー遷移は頻繁に上下していることから、マルチカノニカル法は上手く調整できたことが確認できた。

### 4.3.2 各拘束における RMSD 遷移

次に、シミュレーション中に天然構造が現れたかどうかを調べるために RMSD を計算した。この節における RMSD(Root Mean Square Deviation)とは、天然構造をリファレンス構造として、シミュレーション中に現れた構造を対応する原子間の距離が最小になるよう重ね合わせた時のそれら距離の平均値で表す。つまり、天然構造に似た構造をとっていると RMSD の値は小さくなり、逆に天然構造とはかけ離れた構造をとっていると大きな値となる。今回は、各アミノ酸残基の C $\alpha$ 原子を対象として RMSD を計算した。

以上の条件で計算した RMSD 遷移を図 4.3 に示す。

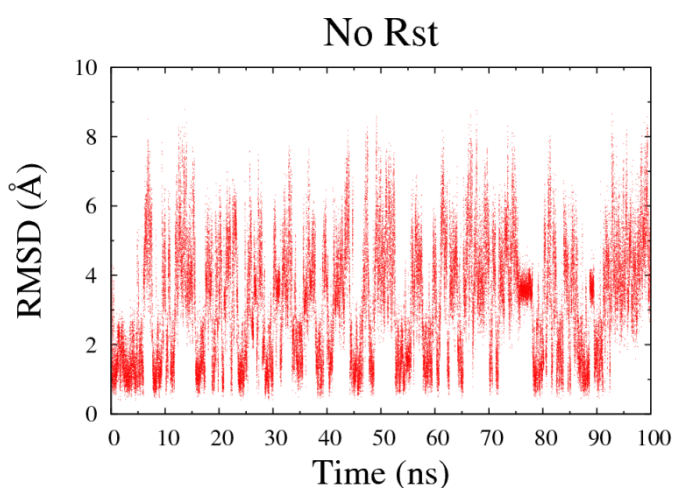


図 4.3 RMSD 遷移 (拘束無し)

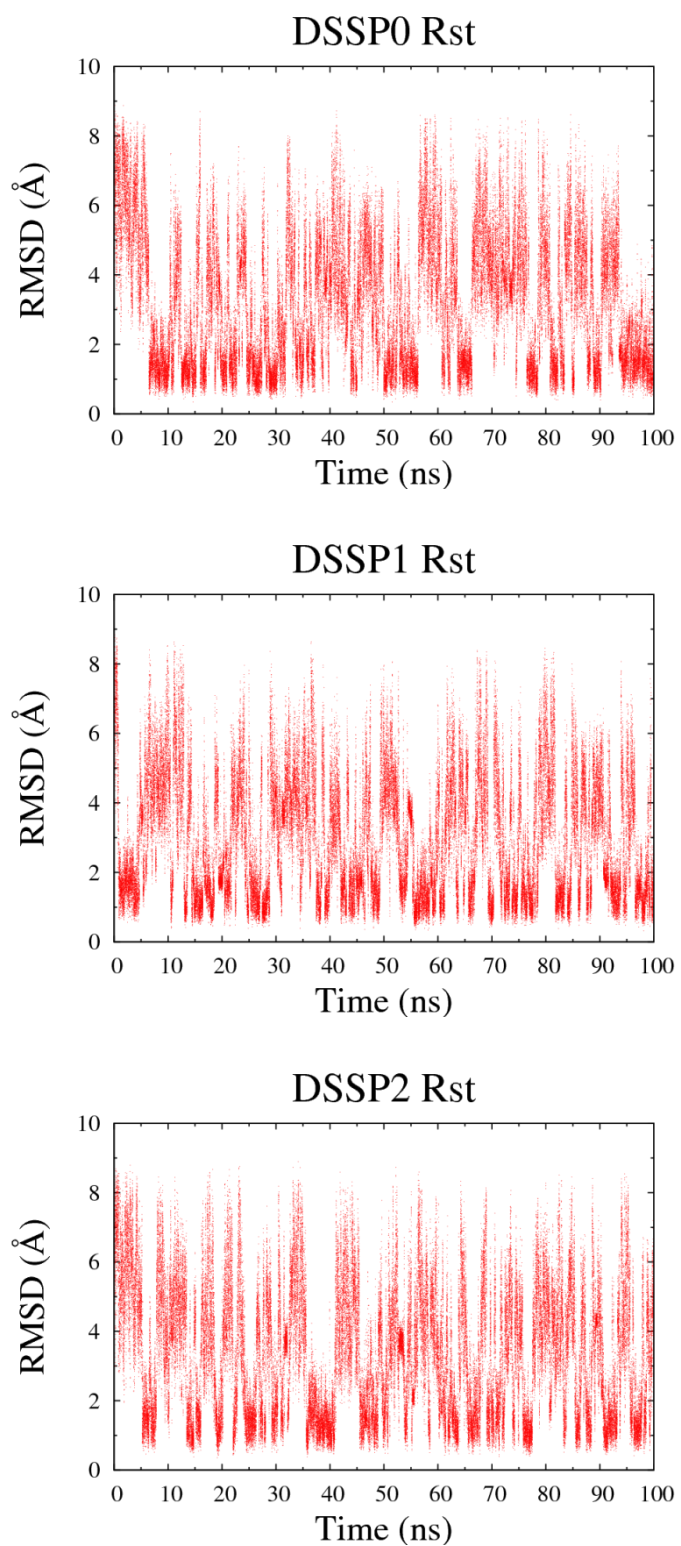


図 4.3 RMSD 遷移 (信頼度 0~2)



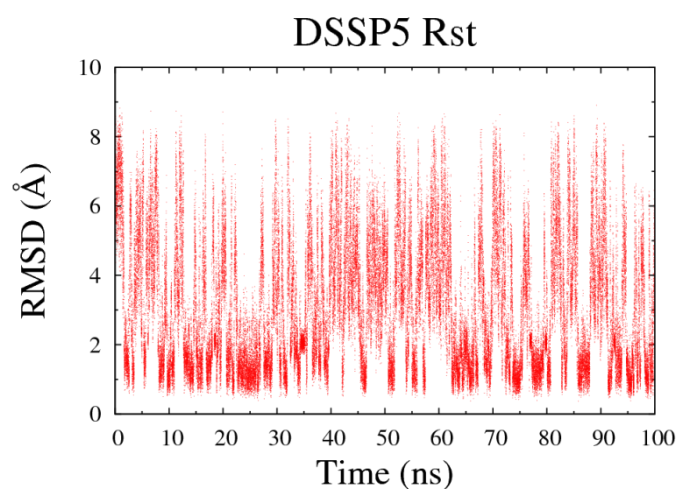
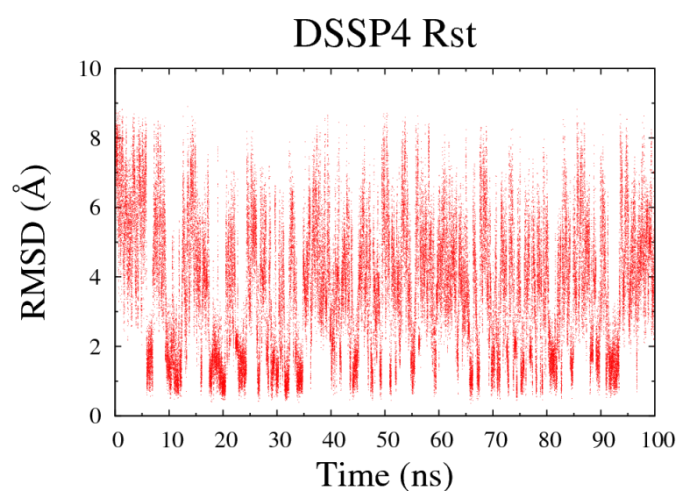
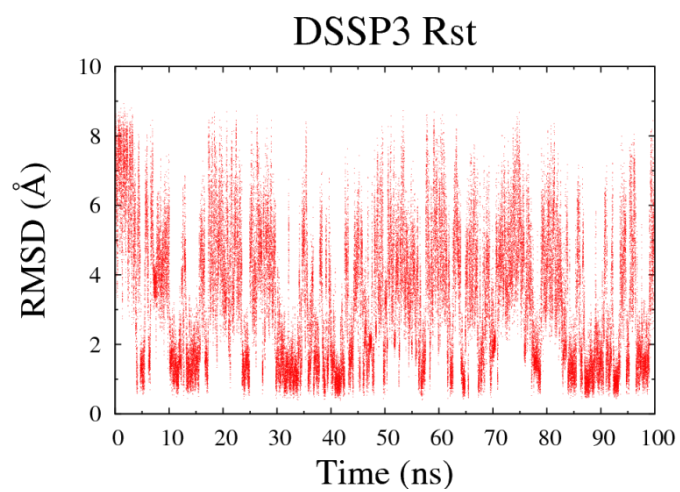


図 4.3 RMSD 遷移 (信頼度 3~5)

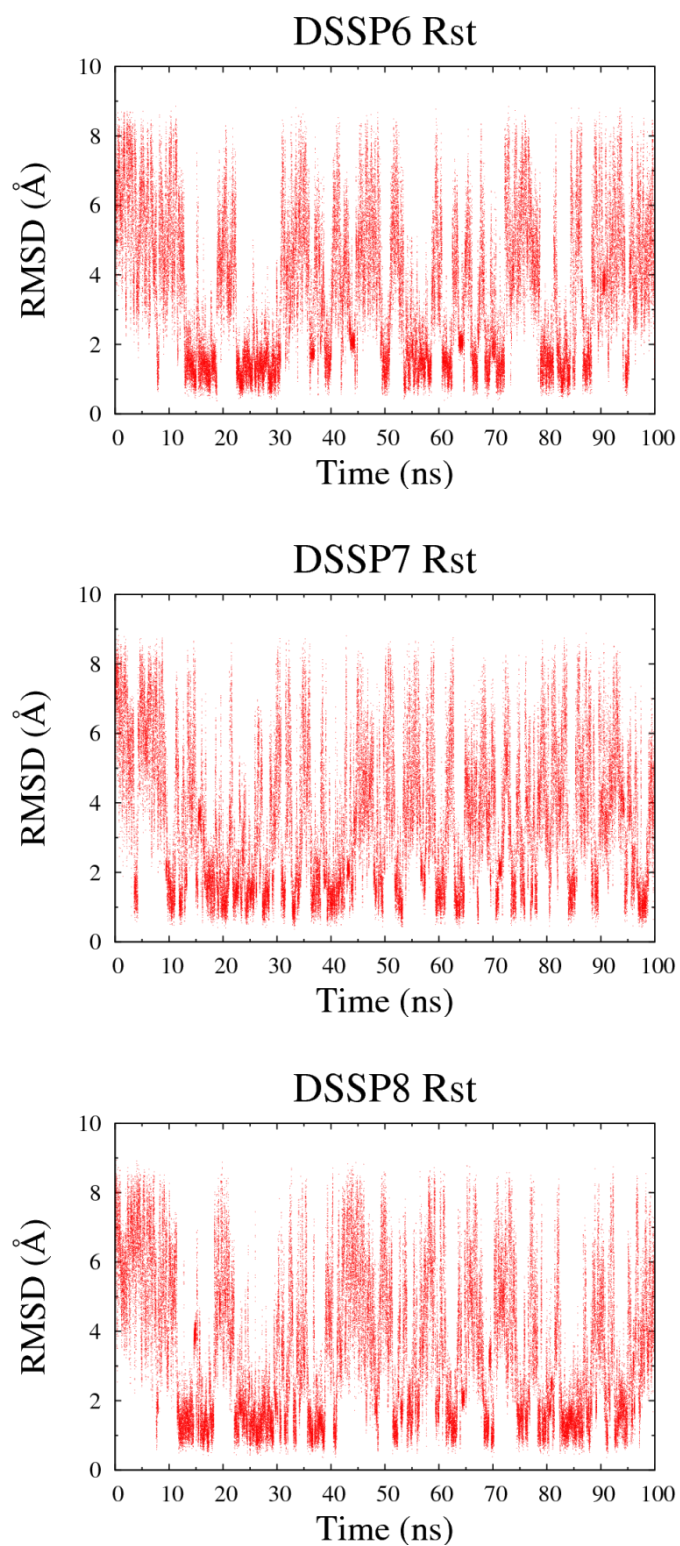


図 4.3 RMSD 遷移 (信頼度 6~8)

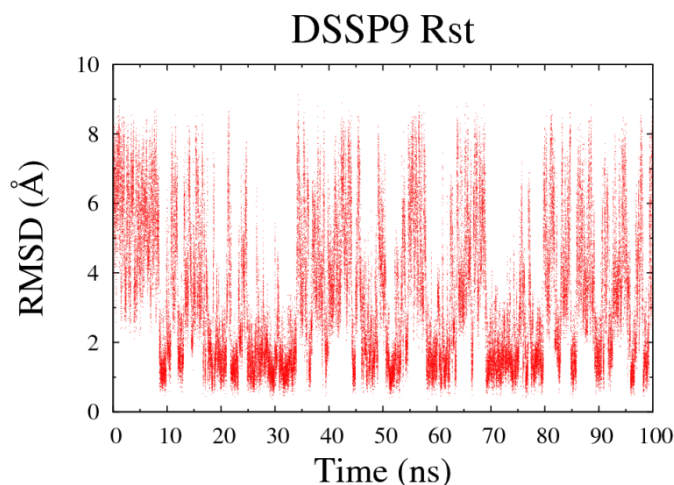


図 4.3 RMSD 遷移 (信頼度 9)

図 4.3 を見て分かるように、全パターンにおいて  $\text{RMSD} < 1.0 \text{ \AA}$  に何度もなっており、この事から全パターンで天然構造へ fold しているということが確認された。

### 4.3.3 各拘束におけるクラスタリング結果

最後に、どの程度の割合で天然構造が形成されていたかを調べるためクラスタリングを行った。

最初に、クラスタリングに用いる構造を選び出す。選ぶ基準は、温度 300 K で reweighting した後の存在確率が  $1.0 \times 10^{-8}$  以上の構造とした。

ここで、マルチカノニカル法における reweighting について説明する。温度 300 K で reweighting(再重み付け)とは、ある構造のポテンシャルエネルギーを基にして温度 300 K という条件下で存在する確率を求めるということである。図 4.2 の拘束無しの場合を例にとると、このシミュレーションはマルチカノニカル法によって、ポテンシャルエネルギーの出現確率が等しくなるようサンプリングされている(赤)が、それを 300 K における出現確率(緑)になるように、つまりポテンシャルエネルギーが  $-380 \text{ kcal/mol}$  付近の構造は存在確率が高くなり、 $-300 \text{ kcal/mol}$  付近では存在確率は低くなる、というように構造ごとに存在確率を変えることになる。

また、存在確率が  $1.0 \times 10^{-8}$  以上という条件を付けた理由は、構造を減らしてクラスタリングにかかる計算時間を減らすためである。このようにしてもクラスタリングの結果が殆ど変わらないのは、存在確率の高い構造のみでクラスタリングの結果が殆ど決定するためである。

以上のようにして選ばれた構造群を使ってクラスタリングを行う。

次に、クラスタリングの手順について説明する。手順は以下の通りである。

1. 構造群の中から1つの構造をリファレンスとして選択して、残りの構造群に対して RMSD を計算する。計算した RMSD がある値(=クラスタ半径)以下の構造でクラスタを作る。
2. クラスタ内にある構造の平均構造を求める。平均構造をリファレンスとして、再度構造群に対して RMSD を計算する。
3. 2.で計算した RMSD がクラスタ半径以下の構造でクラスタを作る。
4. 3.で作られたクラスタと 2.で受けたクラスタが同等の内容であれば、それらの構造を1つのクラスタとして抜き出し、まだ構造群に構造が残っている場合は 1.へ戻る。もし、2つのクラスタが異なっている場合は 3.のクラスタを入力として 2.へ戻る。

今回のクラスタリングの条件は、全 C $\alpha$ 原子を対象とし、クラスタ半径は 1.0 Å とした。以上の条件でクラスタリングを行い、最も占有率の高いクラスタの代表構造 (rainbow color) と天然構造 (gray) を重ね合わせた図とそのクラスタの存在割合をそれぞれ図 4.4 に示す。

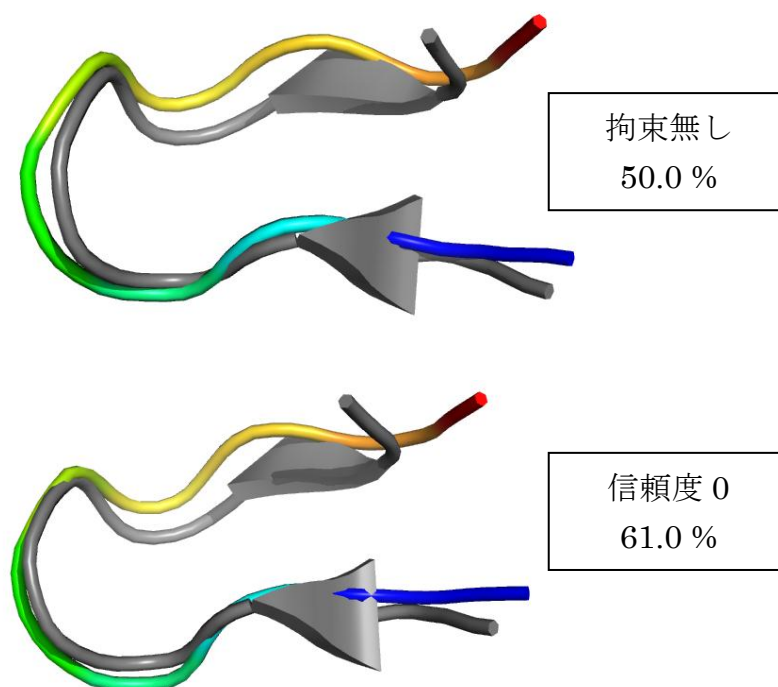


図 4.4 最も占有率の高いクラスタの代表構造 (rainbow color) と天然構造 (gray) を重ね合わせた図とそのクラスタの存在割合 (拘束無し、信頼度 0)

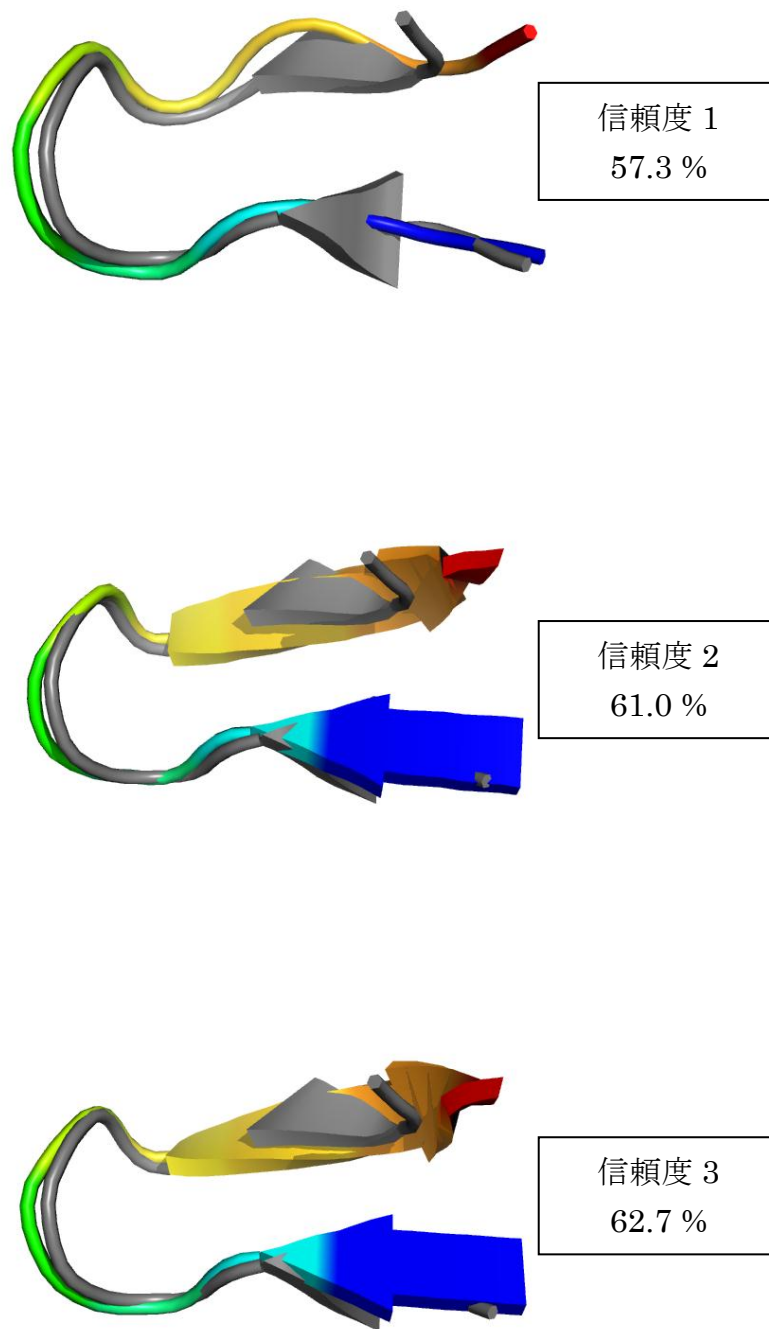


図 4.4 最も占有率の高いクラスターの代表構造 (rainbow color) と天然構造 (gray) を重ね合わせた図とそのクラスターの存在割合 (信頼度 1~3)

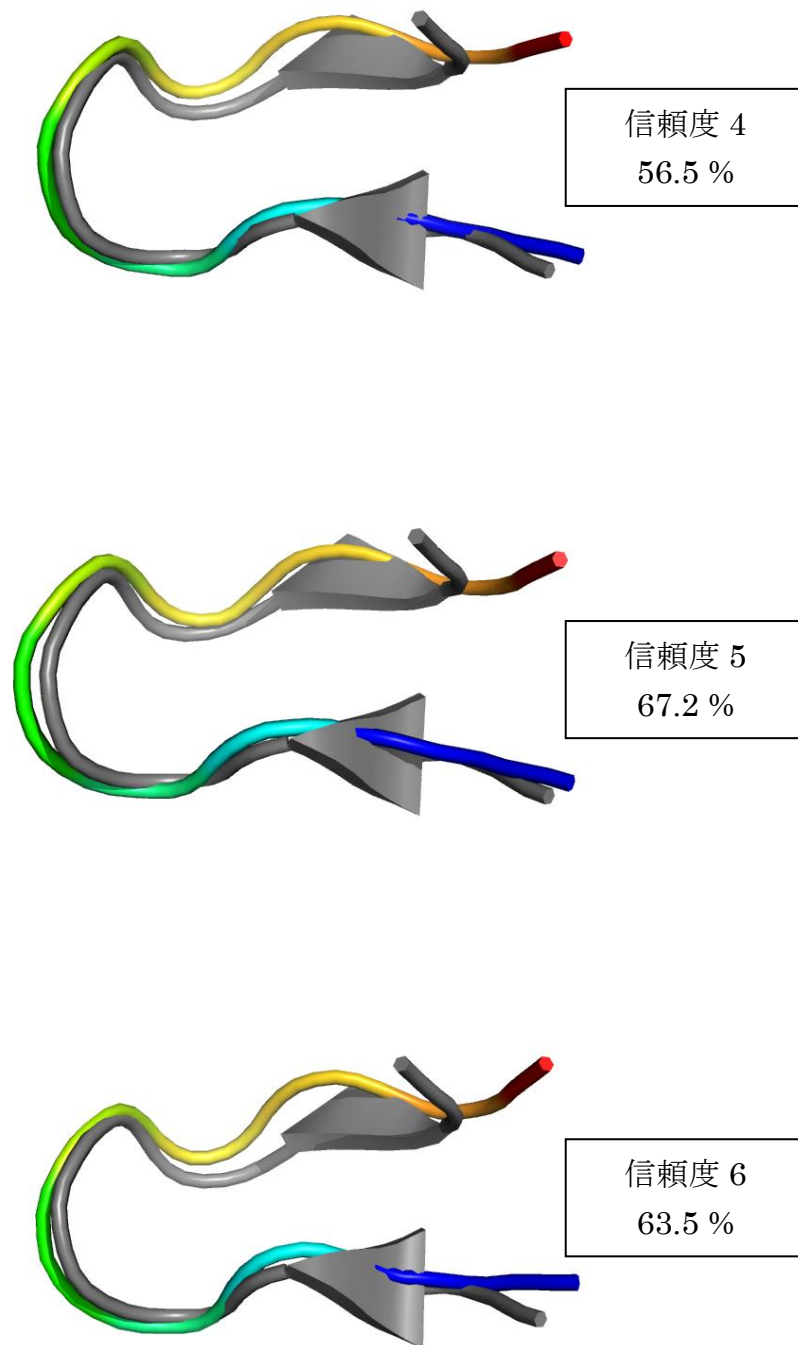


図 4.4 最も占有率の高いクラスターの代表構造 (rainbow color) と天然構造 (gray) を重ね合わせた図とそのクラスターの存在割合 (信頼度 4~6)

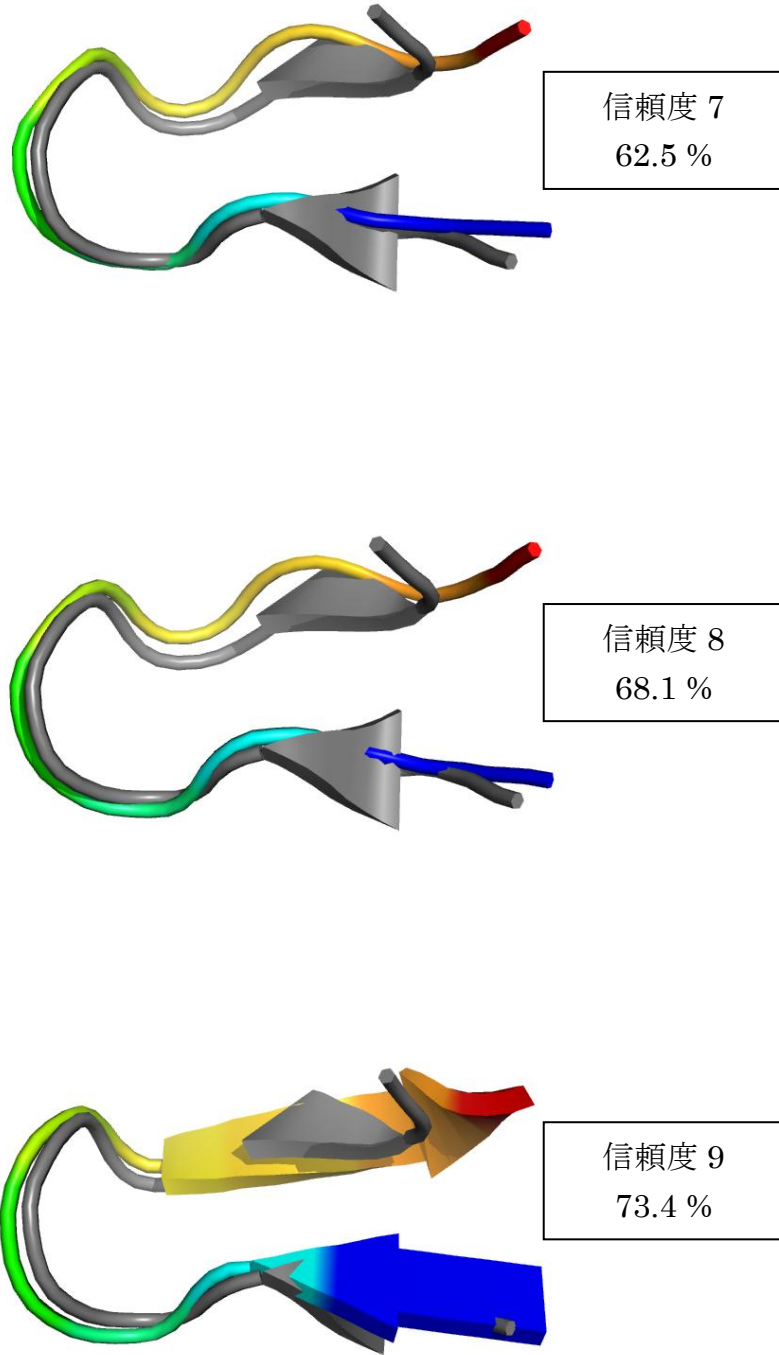


図 4.4 最も占有率の高いクラスターの代表構造 (rainbow color) と天然構造 (gray) を重ね合わせた図とそのクラスターの存在割合 (信頼度 7~9)

図 4.4 から、全拘束パターンにおいて、クラスタリングによって最も占有されているクラスターの構造は天然構造であることが分かる。また、どの信頼度においても、拘束無しのものより占有率が上がっており、天然構造を安定化して見つけやすくしている。

#### 4.3.4 各拘束における Gly7 の主鎖二面角

chignolin の 7 番目の残基であるグリシンは、天然構造では  $(\phi, \psi) = (90^\circ, 0^\circ)$  付近の値をとっている。今回、この残基に適用した C 拘束には、信頼度が高くなるにつれて  $\phi > 0^\circ$  の領域にある程度安定化する能力を持っている。そのため、各 C 拘束がどの程度 fold に作用したかを調べるため、シミュレーション中の Gly7 の主鎖二面角の出現確率(reweighting 無し)の分布を調べた。その結果を図 4.5 に示した。

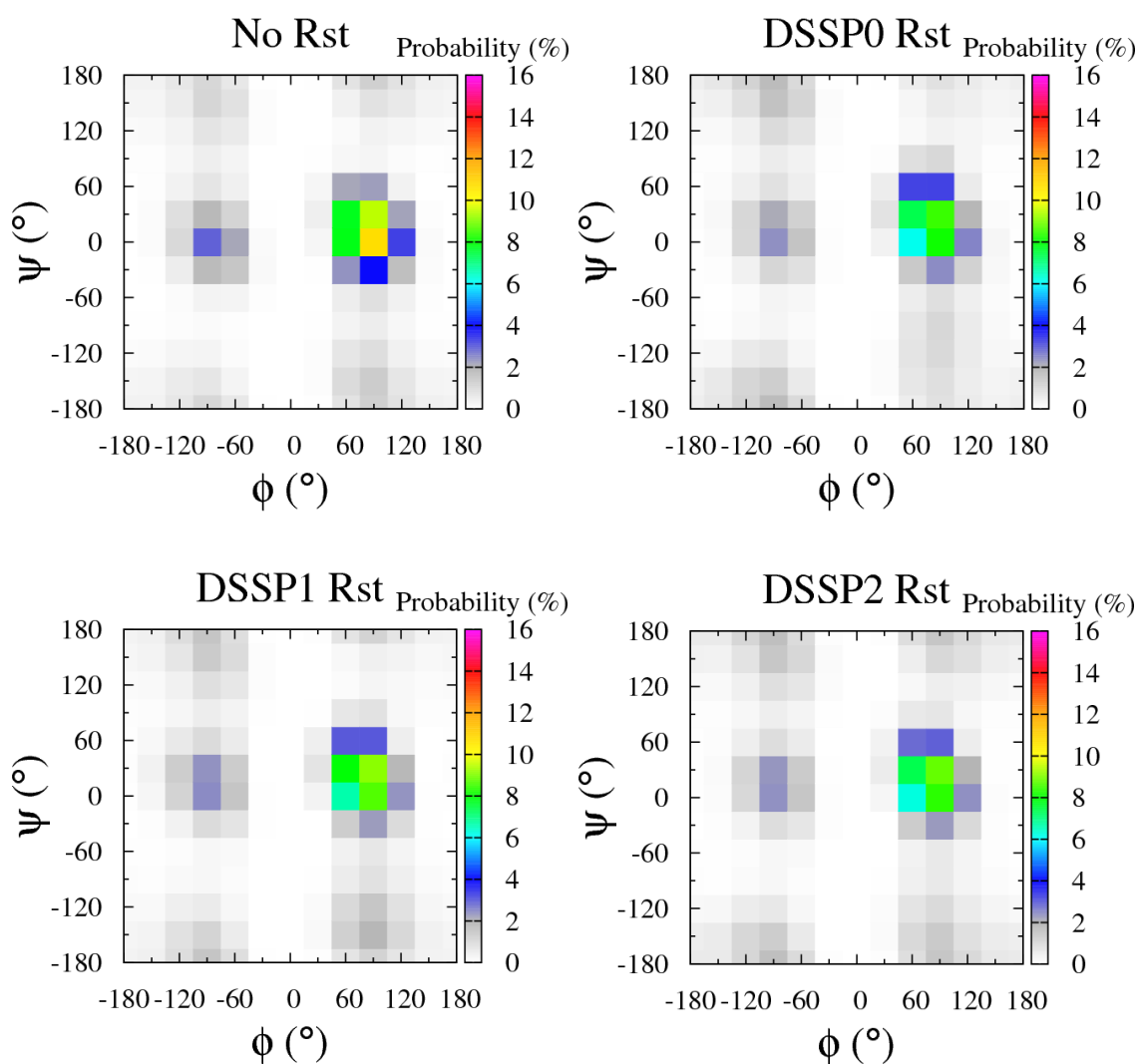


図 4.5 Gly7 の主鎖二面角の分布 (拘束無し、信頼度 0~2)



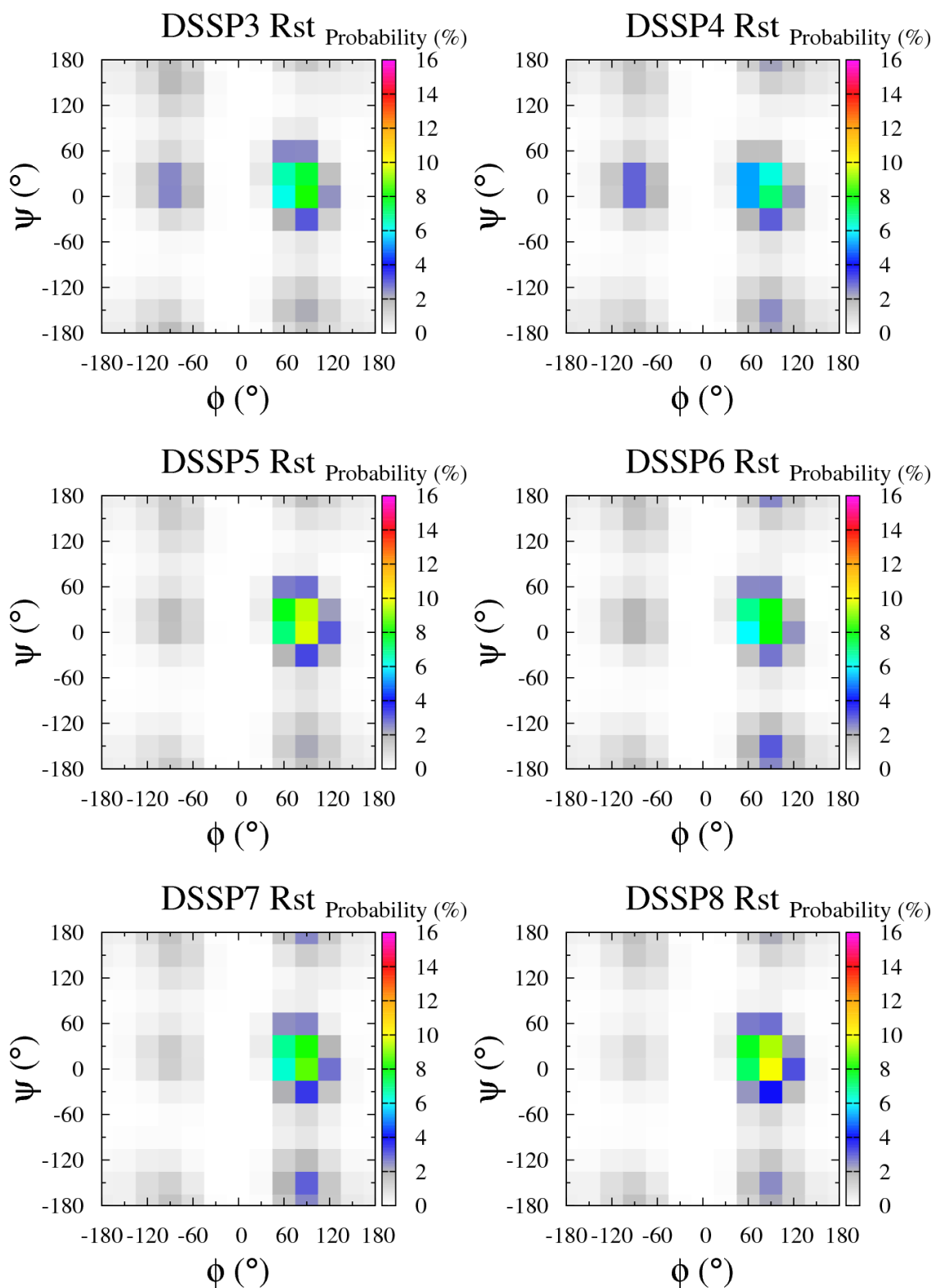


図 4.5 Gly7 の主鎖二面角の分布 (信頼度 3~8)

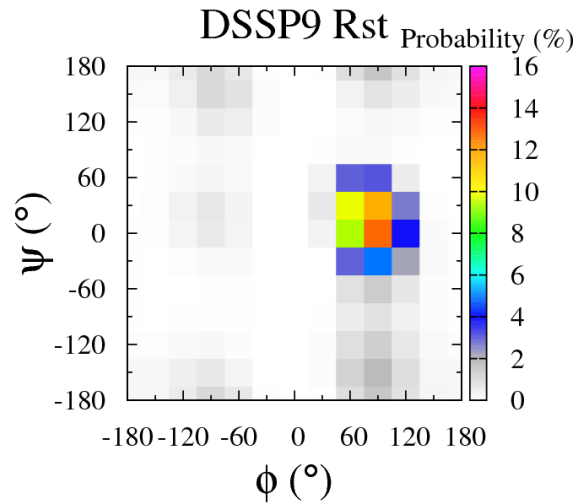


図 4.5 Gly7 の主鎖二面角の分布 (信頼度 9)

図 4.5 から、信頼度 9 ではかなりの確率で  $\phi > 0^\circ$  の領域に存在していることが分かる。また、拘束無しの場合と信頼度 8 が同じくらいの確率で  $(\phi, \psi) = (90^\circ, 0^\circ)$  付近に存在しており、それより低い信頼度では拘束無しに比べて正解領域に存在している確率が低い。信頼度 0~4 くらいまでは、そもそも拘束エネルギーマップから見て、 $(\phi, \psi) = (90^\circ, 0^\circ)$  付近の領域は他の領域に比べて安定という訳では無いので、拘束無しの場合より存在確率が低くなったと推測される。また、信頼度 5~7 ではヘリックス領域は排斥されるが、代わりに  $\phi > 0^\circ$  で  $\psi$  が  $-150^\circ$  や  $180^\circ$  付近に存在するようになるため、 $(\phi, \psi) = (90^\circ, 0^\circ)$  付近の存在確率が減ってしまっている。

## 4.4 考察

chignolin によるシミュレーションによって、天然構造と同じ二次構造の拘束を加えることができれば、信頼度が低い場合でも拘束無しの場合と比べて、天然構造へと fold させる確率が上がることが確認された。

信頼度が低くても天然構造へと fold できる理由としては、E による拘束がどの信頼度においてもシート領域をよく安定化するので、その拘束されたアミノ酸残基領域は真直ぐに伸びている機会が多く、結果としてその伸びた部分同士が重なり合い、シート構造ができる確率が増えていると考えられる。

また、その伸びた部分が重なるには適切なターンの形成もまた必要である。今回、そのために必要なものの1つとして Gly7 が  $(\phi, \psi) = (90^\circ, 0^\circ)$  付近の値になる必要があった。そして、C9 拘束はこの領域に最安定点を持っており、このことも天然構造への fold させる確率を上昇させた要因であると考えられる。グリシンには側鎖がないため、 $\phi > 0^\circ$  の領域の値をとることに何の障害も無いが、基本的に側鎖を持つアミノ酸がシミュレーション中にこの領域に入っている確率は低いので、天然構造を形成するためにグリシン以外のアミノ酸残基が  $\phi > 0^\circ$  の領域に入る必要がある場合、このことが fold の律速となり得る可能性が考えられる。そのような残基に C9 拘束が加えられることによって、天然構造へと fold する確率を上げることが可能になる。

## 第5章 All- $\alpha$ タンパク質

### 5.1 目的

All- $\alpha$ タンパク質に対して二次構造拘束を適用して、実際に天然構造まで fold するかどうか調べる。

### 5.2 材料・方法

シミュレーション対象の All- $\alpha$ タンパク質は PDB ID: 1pv0[54]である。1pv0 は 2つのヘリックスを持つ、46 残基のタンパク質である (図 5.1)。

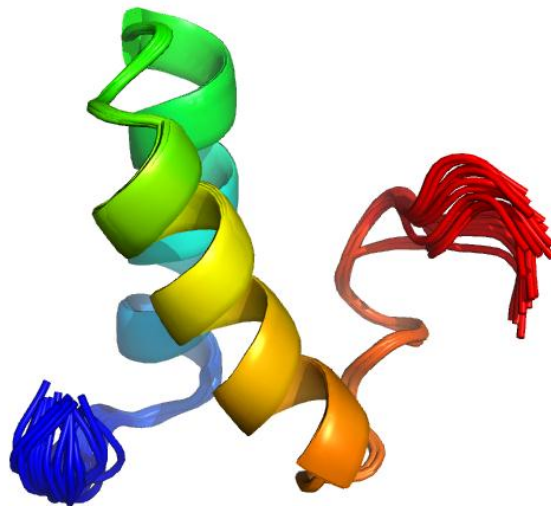


図 5.1 1pv0 の天然構造

1pv0 のアミノ酸配列 (AA)、DSSP による二次構造の割り当て (DSSP)、PSIPRED の二次構造予測結果 (PRED) と信頼度 (CONF) は以下ようになった。

1pv0

```
AA: MRKLSDELLI ESYFKATEMN LNRDFIELIE NEIKRRSLGH IISVSS
DSSP: CCSCCHHHHH HHHHHHHHHT CCHHHHHHHH HHHHHTTCTT TTTCCC
PRED: CCCCCHHHHH HHHHHHHHHC CCHHHHHHHH HHHHHHHHHH HHHCCC
CONF: 9754499999 9999988606 7999999999 9999970534 230269
```

PSIPRED の予測は、第 1 ヘリックス (残基番号 6-19) は完全に正解しているが、第 2 ヘリックス (残基番号 23-35) は C 末端の方へ伸びている。しかし、予測が誤っている残基の信頼度は低くなっていることが分かる。また、DSSP で T と判定されている残基があるが、これは水素結合を介してターンを形成していることを意味している (表 4.1 参照)。

シミュレーション条件は、二次構造拘束を適用、力場は ff03、マルチカノニカル法 (探索温度: 300K ~ 400 K)、溶媒は陰的溶媒モデルの GB/SA OBC type II、水素を含む結合を剛体として扱い、時間刻みは 2 fs である。

シミュレーション手順は、伸展構造を出発点とし、5000 step の最小化、その後 2 ns の平衡化を行った後、実際に解析対象となるプロダクションランを走らせた。

プロダクションランはトータルで 500 ns を行った。その際、500 step(1 ps)毎に原子座標を記録しておいたので、合計 50 万個の構造を得たことになる。解析には、これらの構造を用いて行った。

## 5.3 結果

### 5.3.1 エネルギー分布と遷移

シミュレーション中に現れたエネルギーの分布とそのエネルギー遷移をそれぞれ図 5.2、図 5.3 に示した。

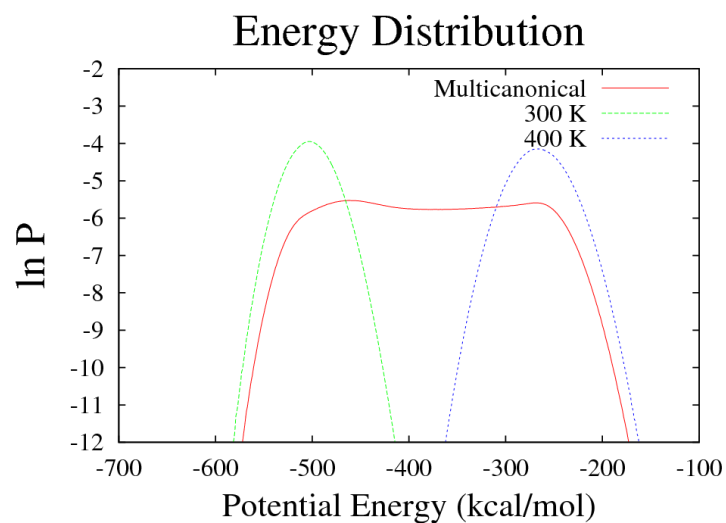


図 5.2 1pv0 のエネルギー分布

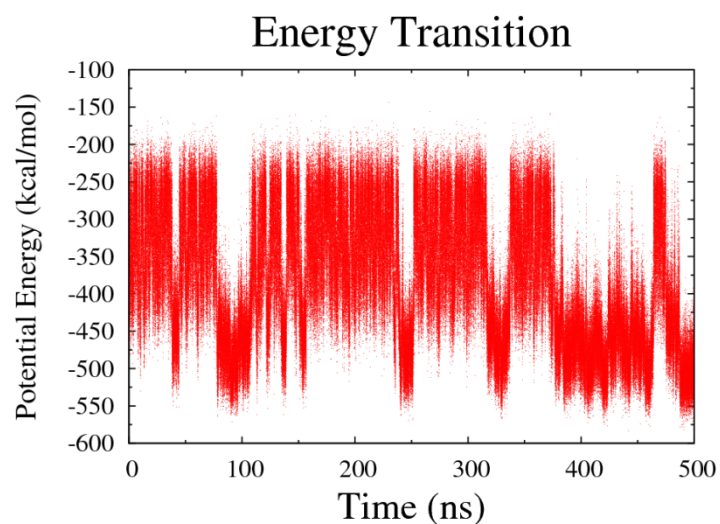


図 5.3 1pv0 のエネルギー遷移

図 5.2、図 5.3 を見ると分かるように、マルチカノニカルのエネルギー分布は 300 K~400 K までのエネルギー範囲を覆っており、かつエネルギー遷移も上下に何度も往復しているので特定の構造にトラップされていない。これらのことから、マルチカノニカル法のポテンシャルエネルギーが適切に求まったことが確認された。

### 5.3.2 RMSD

次に、天然構造が観察されているかどうか確認するために 2 種類の RMSD を計算した。1 つは全 C $\alpha$  (残基番号 1-46) を対象として計算した RMSD、もう 1 つは天然構造中でへ

リックスを作っている残基の  $C\alpha$  (残基番号 6-19,23-35) を対象とした RMSD である。それぞれ図 5.4、図 5.5 に示す。

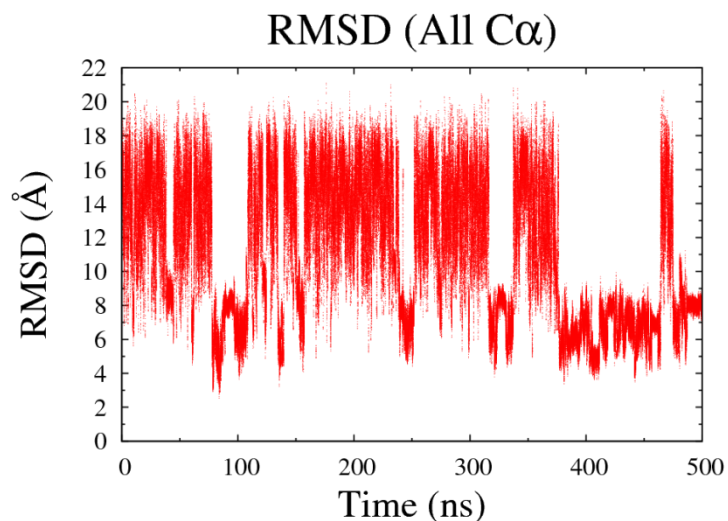


図 5.4 1pv0 の全  $C\alpha$ 原子 RMSD

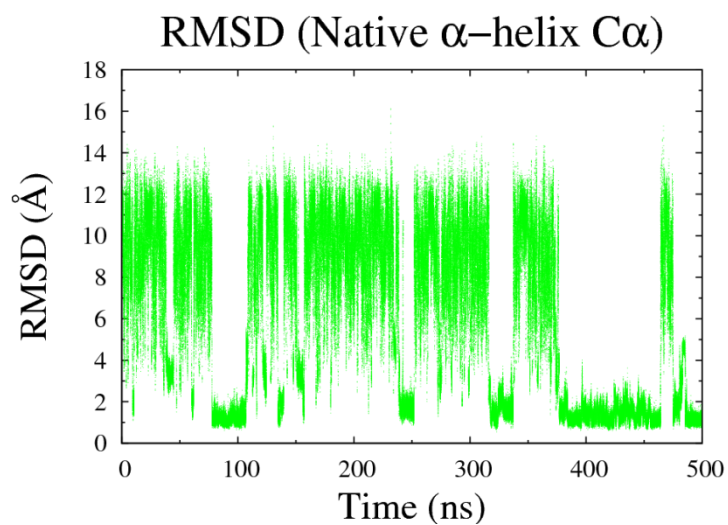


図 5.5 1pv0 の天然構造ヘリックス  $C\alpha$ 原子 RMSD

図 5.4 と図 5.5 から、100, 250, 330, 400-450 ns で天然構造に近い構造をとっていることが分かる。また、図 5.3 のエネルギー遷移と比較してみると、RMSD が低くなっている部分はエネルギーも低い値を示しており、安定な構造であることが分かる。

### 5.3.3 クラスタリング

クラスタリングの方法は、4.3.3 で示したとおりである。クラスタリングの条件は、全  $C\alpha$  原子を対象として、クラスタ半径を  $2.5 \text{ \AA}$  とした。存在確率は  $300 \text{ K}$  に reweighting したものを使用している。以上の条件でクラスタリングした結果、最も存在確率が高かったクラスタの代表構造に最も近いシミュレーション中に現れた構造（最占有クラスタ構造）と天然構造との重ね合わせた図を図 5.6 に示す。

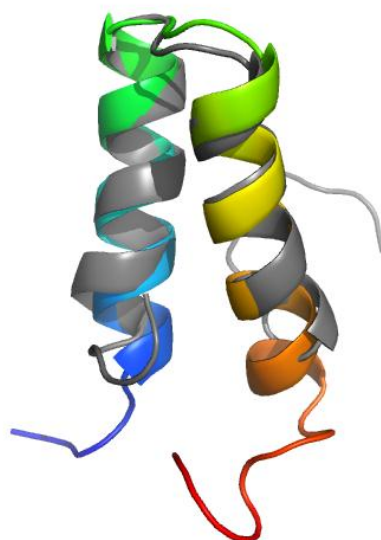


図 5.6 最占有クラスタ構造(color)と天然構造(gray)を重ね合わせた図

図 5.6 を見ると、2 本のヘリックスの配置がほぼ天然構造と同じであることが分かる。N,C 末端の向きは天然構造とは異なるが、これら末端部分はシミュレーション中에서도揺らいでおり、定まった構造をとっていなかったので問題はないとする。



## 5.4 考察

今回、All- $\alpha$ のタンパク質に二次構造予測拘束をかけてMDシミュレーションを行ったところ、天然構造を観測することに成功し、かつ、その構造が最も存在確率の高い構造であったことから、このシミュレーションは成功したと言える。

また、二次構造予測の結果で第2のヘリックスがC末端側に余計に伸びるように予測されていたが、その誤った予測部分の信頼度が低かったことによって、その部分に誤ったヘリックスが形成されなかったことも良い結果であった。

最後に、2本のヘリックス領域の間にある3残基のループ部分があるが、天然構造においてこれらの残基は、 $\phi > 0^\circ$ やシート領域の値をとっている。先に示したように、Cと予測された場合、ヘリックス領域を排斥するような作用が多少あるので、このこともうまく作用して適切なループの形成を促進したのではないかと考えられる。

## 第6章 All- $\beta$ タンパク質

### 6.1 目的

All- $\beta$ タンパク質に対して二次構造拘束を適用して、実際に天然構造まで fold するかどうか調べる。

### 6.2 材料・方法

シミュレーション対象の All- $\beta$ タンパク質は PDB ID: 1i6c[55]である。1i6c は 3つの  $\beta$ -strand を持つ、39 残基のタンパク質である (図 6.1)。

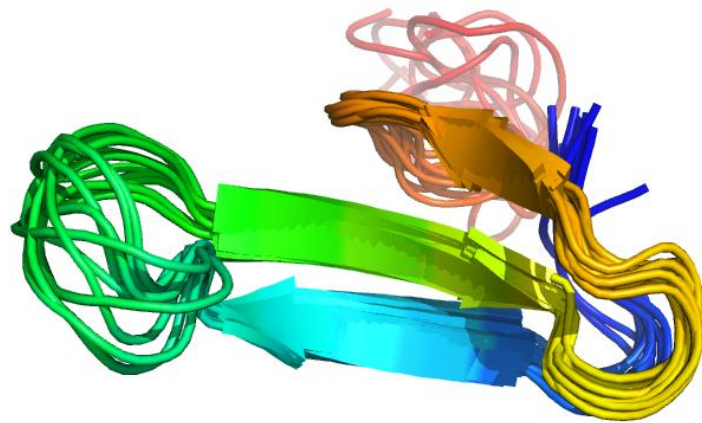


図 6.1 1i6c の天然構造

1i6c のアミノ酸配列 (AA)、DSSP による二次構造の割り当て (DSSP)、PSIPRED の二次構造予測結果 (PRED) と信頼度 (CONF) は以下ようになった。

1i6c

```
AA: KLPPGWEKRM SRSSGRVYF NHITNASQWE RPSGNSSSG  
DSSP: CCCSSEEEEE CSSSSCEEEEE ETTTTEEESC CCCSSSCCC  
PRED: CCCCCEEEEE ECCCCCEEEEE ECCCCCEEC CCCCCCCCC  
CONF: 9789828879 6699958989 7887648467 898888898
```

PSIPRED の予測は、第 1 (残基番号 6-10)、第 2  $\beta$ -strand (残基番号 17-21) は全てカバーできているが、第 3  $\beta$ -strand (残基番号 26-28) では一か所だけ予測できていない。また、DSSP の判定で各 strand の間の残基は S や T と判定されているが (表 4.1 参照)、これはその部分が曲がっている、またはターンを形成していることを示している。

シミュレーション条件は、二次構造拘束を適用、力場は ff03、マルチカノニカル法 (探索温度: 300K ~ 400 K)、溶媒は陰的溶媒モデルの GB/SA OBC type II、水素を含む結合を剛体として扱い、時間刻みは 2 fs である。

シミュレーション手順は、伸展構造を出発点とし、5000 step の最小化、その後 2 ns の平衡化を行った後、実際に解析対象となるプロダクションランを走らせた。

プロダクションランはトータルで 500 ns を行った。その際、500 step(1 ps)毎に原子座標を記録しておいたので、合計 50 万個の構造を得たことになる。解析には、これらの構造を用いて行った。

## 6.3 結果

### 6.3.1 エネルギー分布と遷移

シミュレーション中に現れたエネルギーの分布とそのエネルギー遷移をそれぞれ図 6.2、図 6.3 に示した。

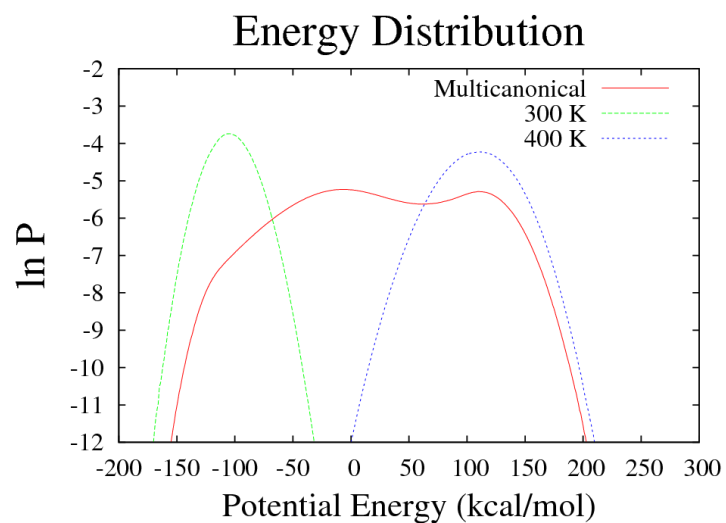


図 6.2 li6c のエネルギー分布

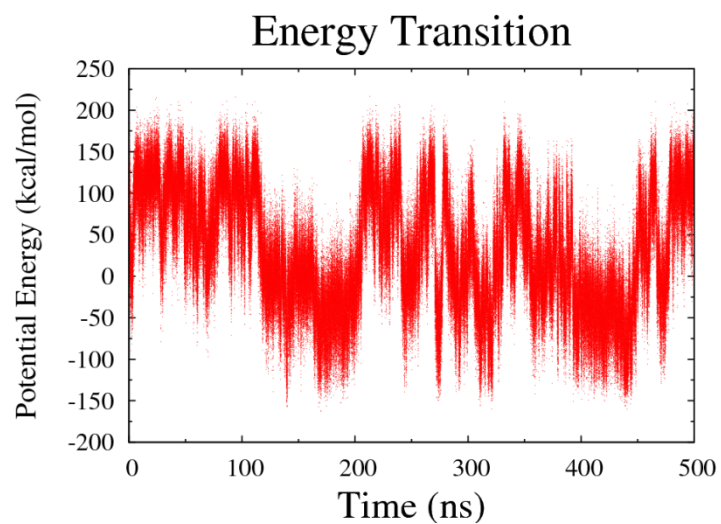


図 6.3 li6c のエネルギー遷移

図 6.2 を見ると、低エネルギー側の分布確率が下がってしまっており、まだ少しマルチカノニカルのパテンシャルエネルギーが適切ではないが、一応 300 K~400 K の範囲を覆っており、かつ、図 6.3 のようにトラップされている様子もないため、このまま解析することとした。

### 6.3.2 RMSD

次に、天然構造が観察されているかどうか確認するために2種類のRMSDを計算した。1つは全C $\alpha$ （残基番号1-39）を対象として計算したRMSD、もう1つは天然構造中でシートを作っている残基のC $\alpha$ （残基番号6-10,17-21,26-28）を対象としたRMSDである。それぞれ図6.4、図6.5に示す。

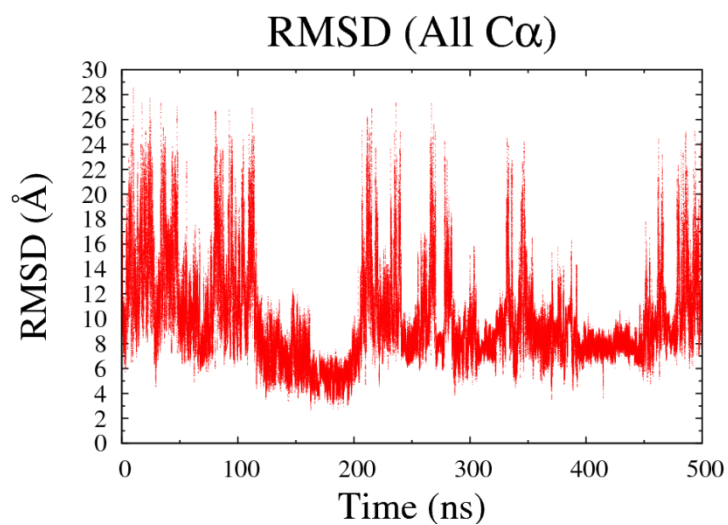


図 6.4 1i6c の全 C $\alpha$ 原子 RMSD

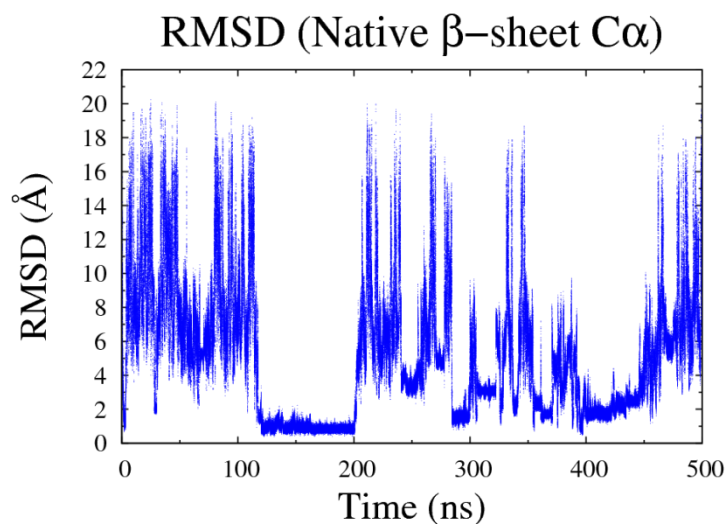


図 6.5 1i6c の天然構造シート C $\alpha$ 原子 RMSD

図 6.5 から、天然構造と同じシート構造が形成されていることが確認できる。また、シート構造が出来ている時にエネルギーも低くなっており、安定な構造であると言える。

### 6.3.3 クラスタリング

クラスタリングの方法は、4.3.3 で示したとおりである。クラスタリングの条件は、全  $C\alpha$  原子を対象として、クラスタ半径を  $2.5 \text{ \AA}$  とした。存在確率は  $300 \text{ K}$  に reweighting したものを使用している。以上の条件でクラスタリングした結果、最も存在確率が高かったクラスタの代表構造に最も近いシミュレーション中に現れた構造（最占有クラスタ構造）と天然構造との重ね合わせた図を図 6.6 に示す。

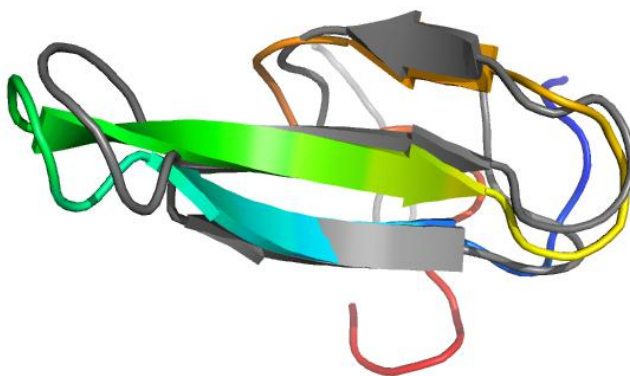


図 6.6 最占有クラスタ構造(color)と天然構造(gray)を重ね合わせた図

図 6.6 から、天然構造と同じ配置でシート構造を形成していることが分かる。また、第 1  $\beta$ -strand と第 2  $\beta$ -strand の間のループの部分は、天然構造とは異なる形をとっているが、図 6.1 を見ると分かるように、この部分は非常に構造が揺らいでいるので特に問題は無いと判断した。

## 6.4 考察

今回、All- $\beta$ のタンパク質に二次構造予測拘束をかけてMDシミュレーションを行ったところ、天然構造を観測することに成功し、かつ、その構造が最も存在確率の高い構造であったことから、このシミュレーションもAll- $\alpha$ 同様に成功したと言える。

また、二次構造予測において、第3  $\beta$ -strandの開始部分がCと判定されていたが、この部分も適切にシート構造が形成されていた。この理由は、Cの拘束がヘリックスを排斥して、その結果、シート領域の主鎖二面角の値をとる確率が高まったためだと考えられる。

最後に、ループ部分について考察する。 $\beta$ シートを形成するには、真っ直ぐに伸びた $\beta$ -strandに加えて、適切なループを形成してstrand同士が近付かなければならない。そのため、天然構造と同じ形のループを形成させることもまた大事なことである。そして、そのループには、しばしば $\phi > 0^\circ$ の値をとる残基が存在しており、そのことがシート形成のネックとなることがよくある。今回の1i6cでいえば、第1  $\beta$ -strandと第2  $\beta$ -strandではGly15、第2  $\beta$ -strandと第3  $\beta$ -strandでは、Asn25が天然構造で $\phi > 0^\circ$ の値をとっている。特に、側鎖を持っているアミノ酸は $\phi > 0^\circ$ の値をとりにくいことは、先に行ったAlanine-dipeptideの実験からも明らかである(3.3.1参照)。しかし、本実験では、これらの残基はそれぞれC9、C6の拘束がされており、そのため幾分か $\phi > 0^\circ$ の値をとりに易くなっていたはずである。この事もまた、シート構造形成の役に立っていたと考えられる。



## 第7章 $\alpha+\beta$ タンパク質

### 7.1 目的

$\alpha+\beta$ タンパク質に対して二次構造拘束を適用して、実際に天然構造まで fold するかどうか調べる。

### 7.2 材料・方法

シミュレーション対象の $\alpha+\beta$ タンパク質は PDB ID: 1e0g[56]である。1e0g は 3つのヘリックス (3<sub>10</sub>ヘリックスが1つ、 $\alpha$ ヘリックスが2つ) と 2つの $\beta$ -strand を持つ、48 残基のタンパク質である (図 7.1)。

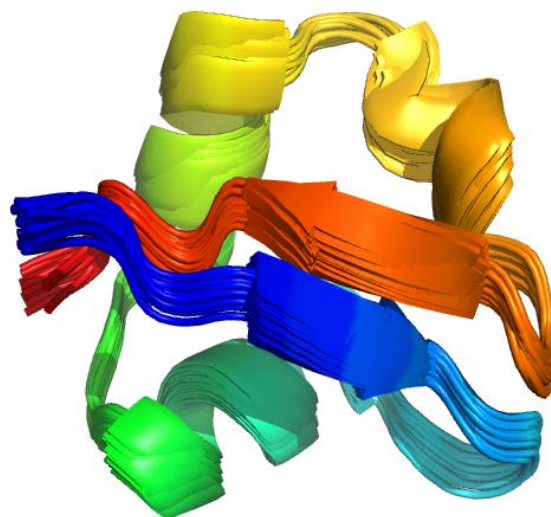


図 7.1 1e0g の天然構造

1e0g のアミノ酸配列 (AA)、DSSP による二次構造の割り当て (DSSP)、PSIPRED の二次構造予測結果 (PRED) と信頼度 (CONF) は以下ようになった。

1e0g

```
AA: DSITYRVRKG DSLSSIAKRH GVNIKDVMRW NSDTANLQPG DKLTLFVK
DSSP: CCCEEEECTT CCHHHHHHHH TCCHHHHHHH CSCGGGCCTT EEECCCCC
PRED: CCEEEEECCC CCHHHHHHHH CCCHHHHHHH CCCCCCCCCC GEEEEEEC
CONF: 9649998789 9889999986 8988999996 4986656899 88999849
```

PSIPRED の予測は、第 1、第 2 ヘリックスと第 1  $\beta$ -strand は全てカバーできているが、第 2  $\beta$ -strand では一か所だけ予測できていないのと、 $3_{10}$ ヘリックス(DSSP で「G」)が全く予測されていない。

シミュレーション条件は、二次構造拘束を適用、力場は ff03、マルチカノニカル法 (探索温度 : 300K ~ 400 K)、溶媒は陰的溶媒モデルの GB/SA OBC type II、水素を含む結合を剛体として扱い、時間刻みは 2 fs である。

シミュレーション手順は、伸展構造を出発点とし、5000 step の最小化、その後 2 ns の平衡化を行った後、実際に解析対象となるプロダクションランを走らせた。

プロダクションランはトータルで 500 ns を行った。その際、500 step(1 ps)毎に原子座標を記録しておいたので、合計 50 万個の構造を得たことになる。解析には、これらの構造を用いて行った。

## 7.3 結果

### 7.3.1 エネルギー分布と遷移

シミュレーション中に現れたエネルギーの分布とそのエネルギー遷移をそれぞれ図 7.2、図 7.3 に示した。

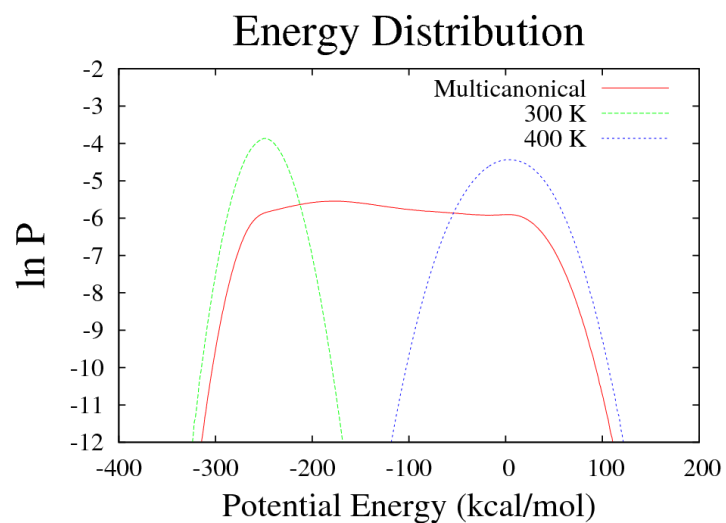


図 7.2 1e0g のエネルギー分布

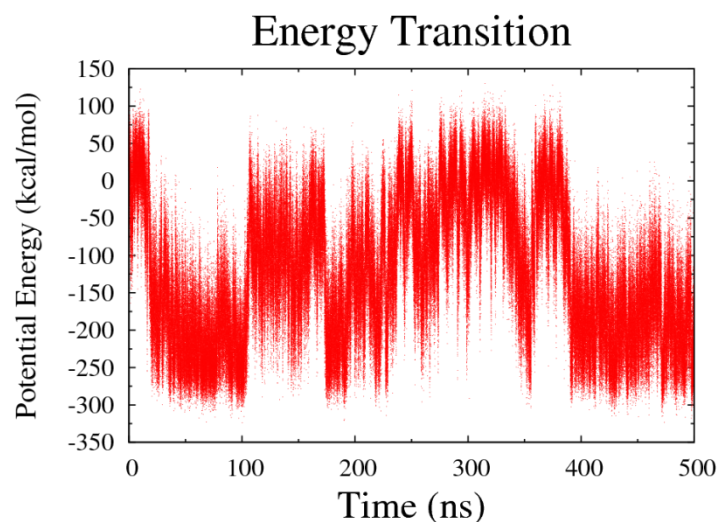


図 7.3 1e0g のエネルギー遷移

図 7.2、図 7.3 を見ると分かるように、マルチカノニカルエネルギー分布は 300 K~400 K までのエネルギー範囲を覆っており、かつエネルギー遷移も上下に何度も往復しているため特定の構造にトラップされていない。これらのことから、マルチカノニカル法のポテンシャルエネルギーが適切に求まったことが確認された。

### 7.3.2 RMSD

次に、天然構造が観察されているかどうか確認するために3種類のRMSDを計算した。1つ目は全C $\alpha$  (残基番号1-48)を対象として計算したRMSD、2つ目は天然構造中でヘリックスを作っている残基のC $\alpha$  (残基番号13-20,24-30,34-36)を対象としたRMSD、3つ目は天然構造中でシートを作っている残基のC $\alpha$  (残基番号4-7,41-44)を対象としたRMSDである。それぞれの結果を図7.4、図7.5、図7.6に示す。

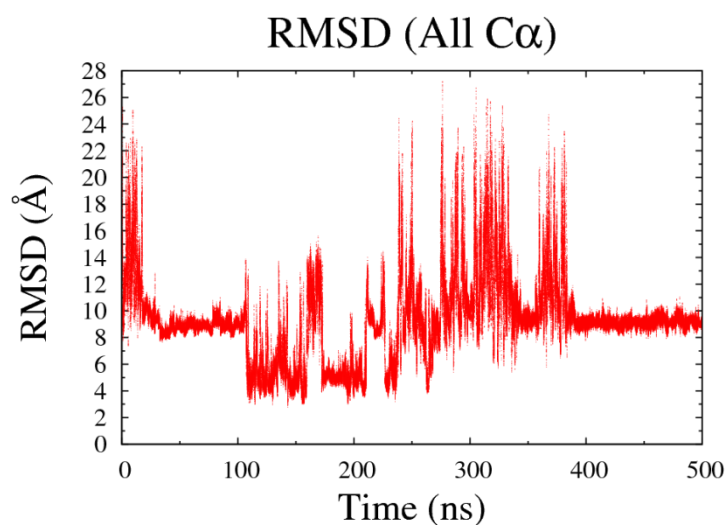


図 7.4 1e0g の全 C $\alpha$ 原子 RMSD

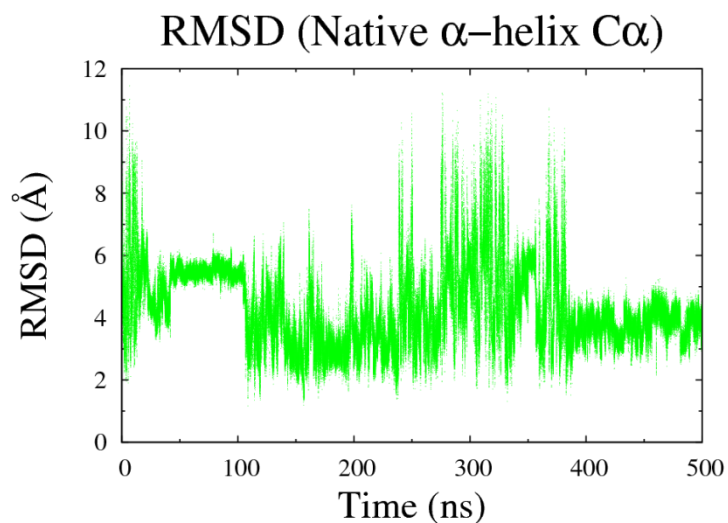


図 7.5 1e0g の天然構造ヘリックス C $\alpha$ 原子 RMSD

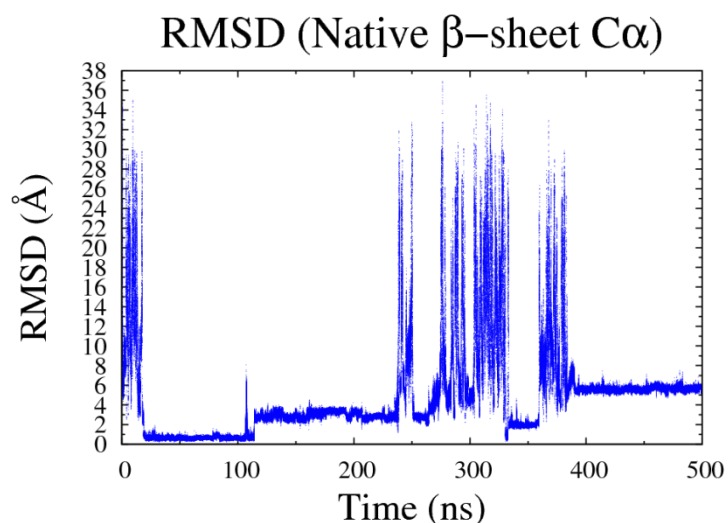


図 7.6 1e0g の天然構造シート C $\alpha$ 原子 RMSD

図 7.5 と図 7.6 から、各二次構造の形成について見ると、ヘリックスでは RMSD が 2.0 Å 付近が最低値であり、天然構造での配置を再現しているとは言い難い値である。シートでは、20-100 ns の間で天然構造のシートを形成していることが分かる。また、全体の RMSD (図 7.4) とエネルギー遷移 (図 7.3) を比べると、RMSD が低いときはエネルギーが高く、RMSD が 10.0 Å 付近に停滞している時にエネルギーが低くなっている。これは、天然構造が適切に探索できていない可能性がある。

### 7.3.3 クラスタリング

クラスタリングの方法は、4.3.3 で示したとおりである。クラスタリングの条件は、全 C $\alpha$  原子を対象として、クラスタ半径を 2.5 Å とした。存在確率は 300 K に reweighting したものを使用している。以上の条件でクラスタリングした結果、最も存在確率が高かったクラスタの代表構造に最も近いシミュレーション中に現れた構造 (最占有クラスタ構造) と天然構造との重ね合わせた図を図 7.7 に示す。

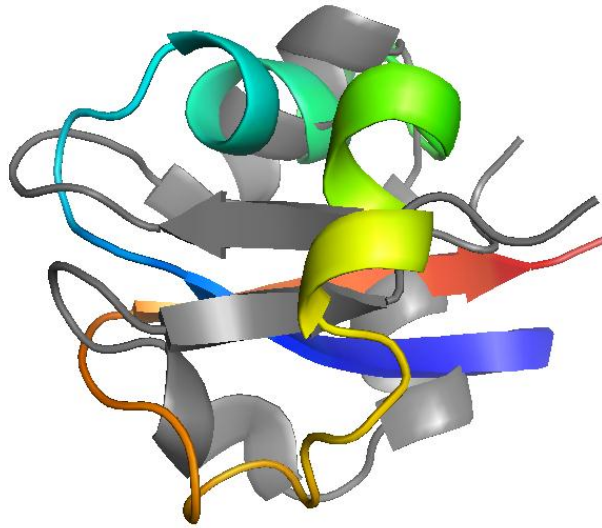


図 7.7 最占有クラスター構造(color)と天然構造(gray)を重ね合わせた図

図 7.7 を見ると分かるように、最も存在確率の高い構造は天然構造では無かった。より詳細に解析するため、この構造を天然構造のヘリックス部分（残基番号 12-36 の  $C\alpha$ ）、シート部分（残基番号 1-11,37-48 の  $C\alpha$ ）のみで重ね合わせを行った。その結果を図 7.8 と図 7.9 に示す。

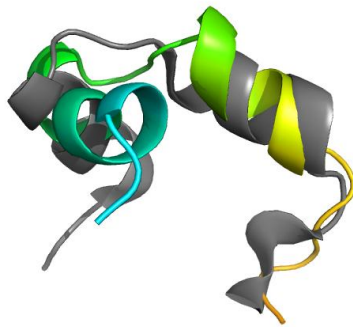


図 7.8 ヘリックス部分の重ね合わせ

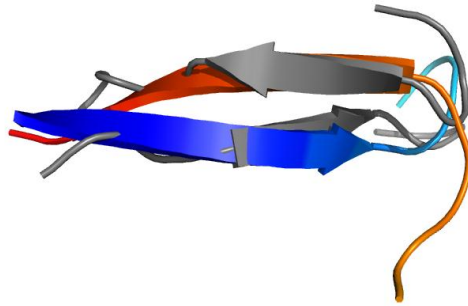


図 7.9 シート部分の重ね合わせ

図 7.8、図 7.9 を見ると、二次構造の形成部分はそれ程外れていないことが分かる。しかし、それ以外の部分の主鎖の形が異なっていることから、天然構造から大きく外れた構造になっている。この事から、二次構造以外の部分に問題があることが明らかになった。

## 7.4 考察

1e0g では、最占有クラスタで天然構造を得ることが出来なかった。その理由を考察する。

最初に、RMSD が下がって天然構造に近付いても、なぜエネルギーが下がらないか考える。まず、二次構造予測で、H,E と予測された残基は、天然構造でもヘリックスとシートを形成しているので、この事は天然構造のエネルギーを下げる要因になり得ない。残っているのは C と予測された残基である。まず、第 1  $\beta$ -strand と第 1 ヘリックスの間の残基を見ると、これらの残基の天然構造における主鎖二面角は、Gly が  $\phi > 0^\circ$  の値で他の残基はシート領域の値なので、C による拘束はエネルギーを下げる方向に働くはずである。次に、第 1 ヘリックスと第 2 ヘリックスの間の残基だが、ここも Gly が  $\phi > 0^\circ$  の値で他の残基はシート領域の値なので、拘束はプラスに働く。そして最後に、第 2 ヘリックスから第 2  $\beta$ -strand の間の残基であるが、この部分で  $3_{10}$ ヘリックスの予測が外れていることに加えて、そのヘリックスの両端側の残基は、天然構造でヘリックス領域の値をとっている。つまり、この領域にある多くの残基は、天然構造ではヘリックス領域の値をとっており、それら残基に C による拘束をかけることは大きなエネルギー増加を促す。特に、ヘリックス構造を作るといった大きな安定化効果を得ることなく、ただヘリックス領域の値になる残基にとってこの拘束は大きな不利となる。この結果として、天然構造に近付いてもそれ程エネルギーが下がらないということが起こったと考えられる。

そして、クラスタリングの結果得られた第 1 クラスタの構造と天然構造を比べてみると、 $3_{10}$ ヘリックスの領域付近の残基が天然構造より真っ直ぐに伸びている傾向が分かる。この事もまた、C による拘束が天然構造を遠ざけている証拠の 1 つである。

これらの実験結果から、C による拘束をもう少し見直す必要があるということが明らかとなった。C という状態は、二次構造を形成していないという事であり、それ以外の情報は無い。そこで、C と予測された場合には、主鎖二面角  $\phi, \psi$  の値を予測するプログラムが使うことが可能であると考えられる。主鎖二面角値予測の予測精度は、二次構造予測に比べると劣るが、二次構造予測における C という情報よりは役に立つと考えられる。C と予測される部分の中で、特にターンを形成する部分においては、二次構造程では無いとしても、主鎖二面角の値には幾分かのパターンがあると予測される。そして、主鎖二面角値予測はそのようなパターンを考慮した結果を返すことが期待され、その結果をもとにエネルギーを加えれば、二次構造予測の C によるエネルギーよりは天然構造へと近づく可能性がある。また、二次構造予測が外れて C と判定された場合においても、この主鎖二面角値予測によってヘリックス寄り、もしくはシート寄りの結果が返されれば、適切な二次構造を形成することが可能になると考えられる。



## 第8章 結論

今回行った、二次構造予測に基づいた拘束 MD シミュレーションによる天然構造の探索は概ね成功した。この事からも、二次構造予測のデータを用いてタンパク質の主鎖二面角を制御するというやり方が有効であることが示された。

しかし、課題も残されている。それは C と判定された時の拘束についてである。二次構造予測の精度が向上して、二次構造の場所をほぼ完璧に予測出来るようになれば、次に予測される構造はループやターンの形であると予測される。これらの構造は、ヘリックスやシート構造に比べて、規則正しい構造では無いが、それでも何らかの特徴はあるはずである。今回の C 拘束は、C と予測されたものを全て混ぜて作った結果、ヘリックス領域を排斥するというエネルギーマップになった。しかし、C と予測された中で更に分類していけば、また違った拘束エネルギーマップが作られ、より天然構造を再現できるようになると考える。

今後は、拘束エネルギーの値を調整したり、その区画を変更したり、拘束エネルギーをアミノ酸別に分けたりして、主鎖二面角の拘束の方法をより洗練していくと共に、二次構造予測以外の情報を用いること検討していきたいと考えている。

## 第2部 Disorder の定量化

### 第1章 背景

タンパク質の **disorder** 領域とは、定まった構造を持たずに揺らいでいる柔軟性の高い領域のことである (図 1.1)。

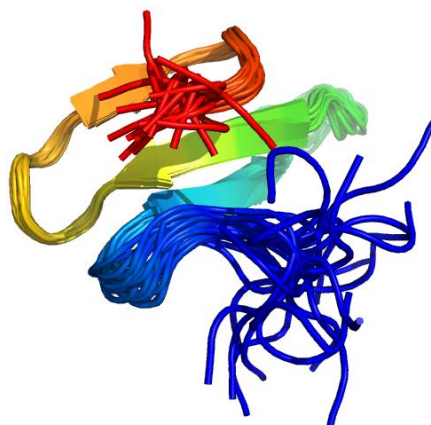


図 1.1 disorder 領域の例

N,C 末端 (青、赤) 部分がモデルによって構造が異なる

これまでのところ、このような定まった構造を持たない領域はタンパク質の機能に何の役にも立っていないと考えられてきたが、近年になって **disorder** 領域の中にはタンパク質の機能に関与しているものが少なくないことが分かり、**disorder** に関する研究が盛んになってきた[83-116]。それらの研究から分かってきたことを以下に紹介する。

機能を持った **disorder** 領域の特徴として、普段は揺らいでいるが、特定のターゲットが近づくとそのターゲット特異的な構造へと変化してその機能を果たす、という性質がある。そのため、ターゲットとの相互作用は、普段揺らいでいるという高いエントロピー状態を相殺して余りある強い相互作用である必要がある。このことが、機能性 **disorder** タンパク質のリガンド結合能の親和性は低い (高エントロピーによる結合エネルギーの相殺) が、特異性は高い (特定の強い相互作用を要求) 原因であるとされている[111]。また、様々な構造へと変化して、構造が大きく異なる複数のターゲットに対して機能をするとする **hub** タンパク質というものも存在する[106]。

さらに、disorder 領域はその他の領域より速い進化をする[115]。本来、disorder 領域はそのタンパク質が持つ機能には大して重要でない部分なので、基本的にアミノ酸残基の変異に寛容である。つまり、定まった構造を成している配列部分とは異なり、disorder 領域の変異は多様で、荷電性アミノ酸残基が非極性アミノ酸残基に変化したり、欠失したりすることは比較的頻繁に起きる、言い換えれば、独自のアミノ酸置換行列を持っているとも言える。そのため、進化速度が速く、新たな機能を持つ確率が高くなる。真核生物に disorder タンパク質が多く見られるのも、このことが原因であると考えられている[114,116]。

このように disorder に関する様々な研究が行われているが、その disorder の定義は「disorder' or 'not disorder'」の2値で決められている。より具体的に述べると、X線結晶構造解析において座標が定まらなかったアミノ酸残基である。このような残基は'missing residue'と呼ばれ、それら missing residue が連続している部分を disorder 領域と呼ぶのが一般的である。missing residue は PDB ファイル中の REMARK 465 に記載されている (図 1.2)。

```
REMARK 465
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465  M RES C SSSEQI
REMARK 465  GLU A 134
REMARK 465  ASN A 135
REMARK 465  ARG A 136
REMARK 465  ASP A 218
REMARK 465  GLN A 219
REMARK 465  SER A 220
REMARK 465  ILE A 221
REMARK 465  HIS A 222
REMARK 465  PHE A 223
REMARK 465  GLN A 224
REMARK 465  LEU A 225
REMARK 465  PHE A 226
REMARK 465  GLU A 227
REMARK 465  LEU A 228
REMARK 465  GLU A 229
REMARK 465  LYS A 230
REMARK 465  ASN A 231
```

図 1.2 REMARK 465 の例

このように、missing residue によって disorder 領域は決められているが、disorder はその鎖の柔軟性を反映していると考えられるなら、その disorder を示す値は「0 か 1」のような2値ではなく、0~1 のような連続値で表現された方が自然である。例として、構造因子ファイルから構成した電子密度マップとタンパク質分子を重ね合わせたものを図 1.3 に示す。

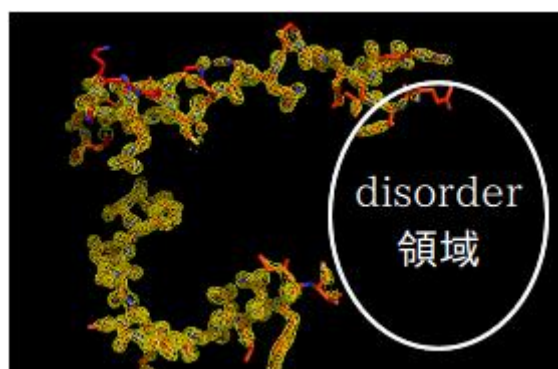


図 1.3 電子密度マップと分子の重ね合わせの例  
黄色の網で囲まれている部分は電子密度が高いことを表す。

図 1.3 を見ると、**disorder 領域** (**missing residue** であるため見えない) の開始点と終了点にある残基の電子密度は低いことが分かる。電子密度が低いのは、結晶中において、その座標に原子が留まらず動いていることを表す。つまり、この残基もある程度 **disorder** しているということである。

そこで、本研究では **missing residue** でない、一般的に **order** とされているアミノ酸残基周りの電子密度に基づいて **disorder** を連続値で表すことを試みた。

## 第2章 定量化の方法

### 2.1 電子密度マップの作成

CCP4[58]というプログラムパッケージを用いて、PDB ファイルと構造因子ファイルから電子密度マップを作成した。作成手順は以下の通りである。

1. cif2mtz で、構造因子ファイルを CCP4 用のファイルフォーマットである mtz ファイルへと変換する。
2. sfall[62-67]で、PDB ファイルから位相を求める。
3. rstats で、Fc を Fo に合わせる。
4. sigmaa[68-77]で、次の fft で使うフーリエ係数を計算する。
5. fft[59-61]で、電子密度マップを計算する。

### 2.2 Order Degree の計算

前節で求めた電子密度マップを用いて、disorder の程度を表す尺度 (Order Degree) を計算する。手順は以下の通りである。

1. 各残基の周辺の電子密度を足し合わせ、その残基の理想電子数で割る。
2. 1.で求めた値を B-factor によって補正する。
3. 2.で求めた値を PDB 中の総電子数で補正する。

データセットは、RCSB PDB サイト[80]において、配列一致度 30%以上のタンパク質を除き、X 線結晶解析における分解能が 2.0 Å 以上で、R<sub>free</sub> 因子が 0.20 以上という条件で集まったタンパク質で、かつ電子密度マップを求めることに成功した 766 個のタンパク質を使った。このデータセットの総残基数は 260871 個である。

## 2.2.1 Raw Order Degree

まず、各残基において、PDB ファイルに記録されている原子の座標を中心として、<sup>ファン</sup>van der Waals (vdW)半径以内にある電子密度を足し合わせていく。用いた各原子の vdW 半径の具体的な値を表 2.1 に示す。

表 2.1 各原子の vdW 半径

原子	vdW 半径 (Å)
H	1.20
C	1.70
N	1.55
O	1.52
S	1.80

CCP4 で作成された電子密度のデータは、解像度の 3 分の 1 の長さで区切られた空間を単位として、電子密度の値を格納している。そのため、そのまま電子密度を足し合わせてしまうと、タンパク質によって異なった大きさの空間で積分することになる。この違いを無くすため、各空間の電子密度の値（負の値は 0 として扱った）を線形補間して 0.2 Å 間隔の空間における電子密度の値を計算することで、空間の大きさを揃えた。

そして、vdW 半径内の空間における電子密度の値を足し合わせた後、その体積 ( $0.2 \text{ Å} \times 0.2 \text{ Å} \times 0.2 \text{ Å}$ ) をかけて周辺電子数を計算する。本研究ではタンパク質の主鎖のみの揺らぎについてのデータも得るために 2 種類の周辺電子数を計算した。1 つはその残基の全原子を対象として vdW 半径内の電子密度を足し合わせて求めた周辺電子数(All)、もう 1 つは主鎖原子 (PDB ファイル中の Atom Record で C, CA, N, O と記述されている原子) を対象として vdW 半径内の電子密度を足し合わせて求めた周辺電子数(Backbone)である。それぞれの周辺電子数を求めた後、周辺電子数(All)をそのアミノ酸残基の理想電子数で割った値を Raw Order Degree (All)、周辺電子数(Backbone)を主鎖原子の理想電子数で割った値 Raw Order Degree (Backbone)として算出した。なお、構成原子から計算した結果、理想電子数は表 2.2 のようになった。

表 2.2 理想電子数

種類	理想電子数	種類	理想電子数
Ala	38	Met	70
Cys	54	Asn	60
Asp	60	Pro	52
Glu	68	Gln	68
Phe	78	Arg	84
Gly	30	Ser	46
His	72	Thr	54
Ile	62	Val	54
Lys	70	Trp	98
Leu	62	Tyr	86
Backbone	27		

上記のようにして求めた全 Raw Order Degree の分布を調べた結果を図 2.1 に示した。

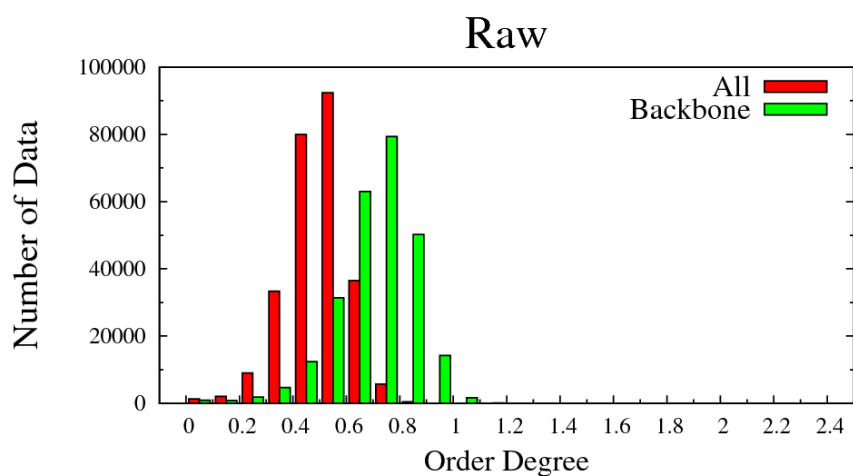


図 2.1 Raw Order Degree の分布

図 2.1 を見ると、All の Order Degree のピークが 0.5 付近に存在していることが分かる。つまり、この値が order している残基の基準となる。しかし、タンパク質によって全体的に 0.5 より低い値しか存在しない場合が見られた。このことから、タンパク質によって何らかの違いが存在しており、それを補正する必要があることが明らかとなった。

## 2.2.2 B-factor による補正

Raw Order Degree を計算する時、vdW 半径内の電子密度を足し合わせて値を求めたが、結晶を作る際の温度、結晶の出来の良さ、X 線の解像度等の実験条件によって電子密度の像は幾らか不鮮明になり、それが原因で電子密度が過小評価されてしまうものもあると考えられる。その不鮮明の度合いは **B-factor** という値で示されるが、この値は各タンパク質によって異なる。そのようなバックグラウンドの不鮮明さを考慮するために、PDB ファイル中に載っているそのタンパク質分子に属する原子の最小 **B-factor** による補正を行った。

**B-factor** がある値  $B$  の時、距離  $r$  における電子密度  $\rho$  は以下の式で近似される [64,67]。

$$\rho(r) = \sum_i^5 C_i \times \left\{ \frac{4\pi}{(D_i + B)} \right\}^{1.5} \times e^{-4\pi^2 r^2 / (D_i + B)}$$

なお、各原子における  $C_i$  と  $D_i$  の値を表 2.3 に示す。



表 2.3 各原子の  $C_i$ ,  $D_i$

原子	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$
H	0.493002	0.322912	0.140191	0.040810	0.003038
C	2.310000	1.020000	1.588600	0.865000	0.215600
N	12.21260	3.132200	2.012500	1.166300	-11.52900
O	3.048500	2.286800	1.546300	0.867000	0.250800
S	6.905300	5.203400	1.437900	1.586300	0.866900

原子	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
H	10.510900	26.125700	3.142360	57.799698	0.0
C	20.843899	10.207500	0.568700	51.651199	0.0
N	0.005700	9.893300	28.997499	0.582600	0.0
O	13.277100	5.701100	0.323900	32.908897	0.0
S	1.467900	22.215099	0.253600	56.172001	0.0

以上の式で求めた電子密度を使い、vdW 半径を 10000 分割して積分をすることで、その B-factor における理想電子数を求めた。その B-factor 補正された理想電子数で、2.2.1 で求めた周辺電子数を割ることで B Corrected Order Degree を求めた。その値の分布を図 2.2 に示す。

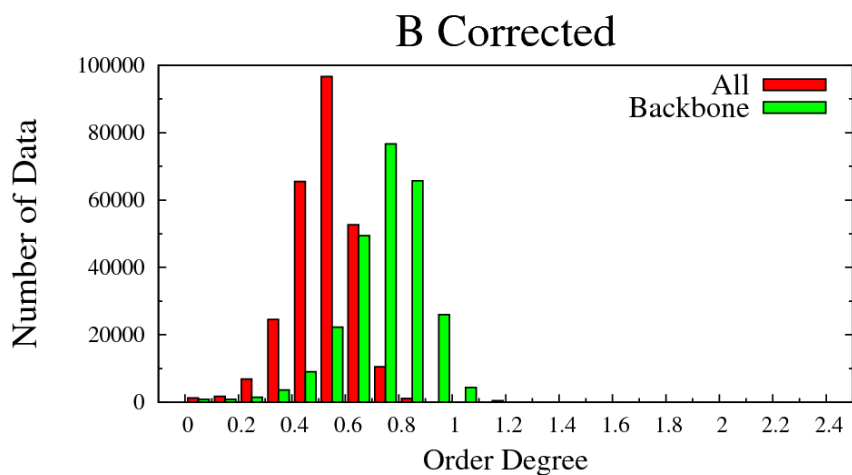


図 2.2 B-factor 補正した Order Degree の分布

B-factor による補正を加えると、図 2.1 の Raw に比べて 1.1~1.2 倍の値になっていた。

### 2.2.3 総電子数による補正

CCP4によって再構成された電子密度マップの単位胞全体を積分して求めた総電子数と、PDBファイルから求めた総電子数を比べるとその値が大きく異なっているものがある事が分かった。この違いを補正するために、PDBファイルの総電子数を電子密度マップの総電子数で割った値を乗じることで、PDBファイルの総電子数に合わせるようにした。この総電子数による補正とB-factorによる補正両方を加えたB-E Corrected Order Degreeを計算した結果、その分布は図2.3のようになった。

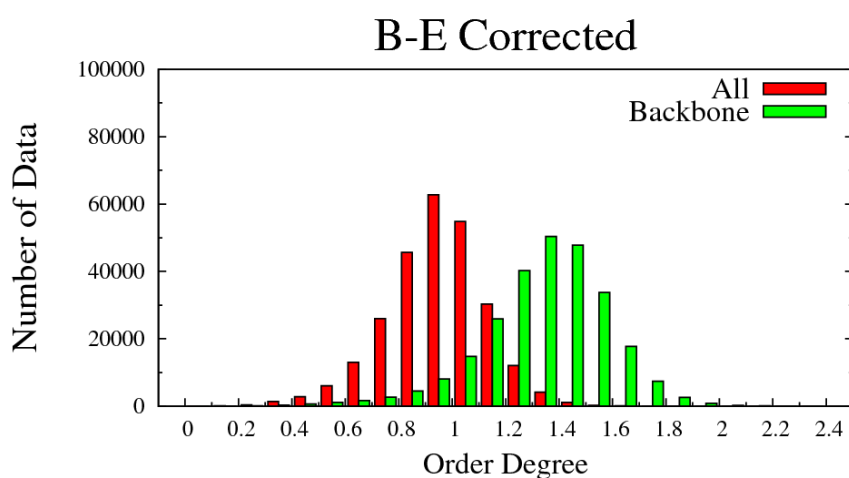


図 2.3 B-factor と総電子数補正した Order Degree の分布

総電子数補正を加えることで、All は 0.9~1.0、Backbone は 1.3~1.4 をピークとする正規分布のような形となった。また、All で 0.8 以上、Backbone で 1.2 以上であるデータ数の割合を調べたところ、それぞれ 80.96%、77.08% という結果となった。

以降、B-factor と総電子数による補正を加えた B-E Corrected Order Degree を Order Degree と呼び、この値を用いて解析を行う。

## 第3章 解析

### 3.1 各アミノ酸の Order Degree

まず、各アミノ酸の Order Degree がどのように分布しているかを調べた。Order Degree を 0.1 ずつに区切った区画において、そのアミノ酸のデータ数の何%がその区画を占めているかを図 3.1 は示している。なお、今回用いている 766 個のタンパク質のデータセットにおける各アミノ酸残基データ数と全体に占める割合を表 3.1 に示す。

表 3.1 各アミノ酸のデータ数

アミノ酸	データ数 (割合)	アミノ酸	データ数 (割合)
Ala	21858 (8.38 %)	Met	3977 (1.52 %)
Cys	3025 (1.16 %)	Asn	12506 (4.79 %)
Asp	16140 (6.19 %)	Pro	12394 (4.75 %)
Glu	15800 (6.06 %)	Gln	9923 (3.80 %)
Phe	10607 (4.07 %)	Arg	12324 (4.72 %)
Gly	20932 (8.02 %)	Ser	16295 (6.25 %)
His	6139 (2.35 %)	Thr	15255 (5.85 %)
Ile	14347 (5.50 %)	Val	17822 (6.83 %)
Lys	14273 (5.47 %)	Trp	4627 (1.77 %)
Leu	22504 (8.63 %)	Tyr	10072 (3.86 %)

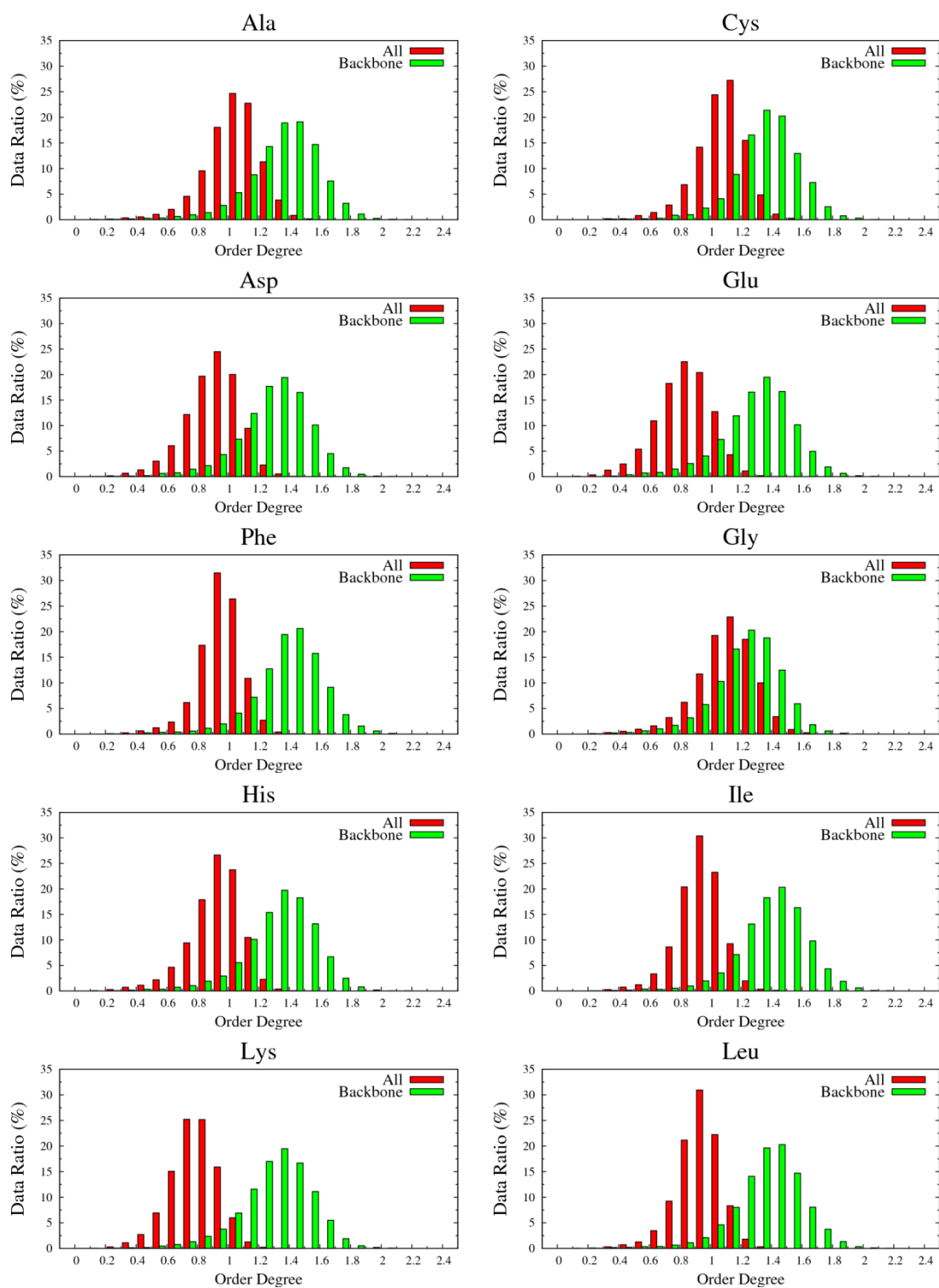


図 3.1 各アミノ酸の Order Degree (Ala~Leu)

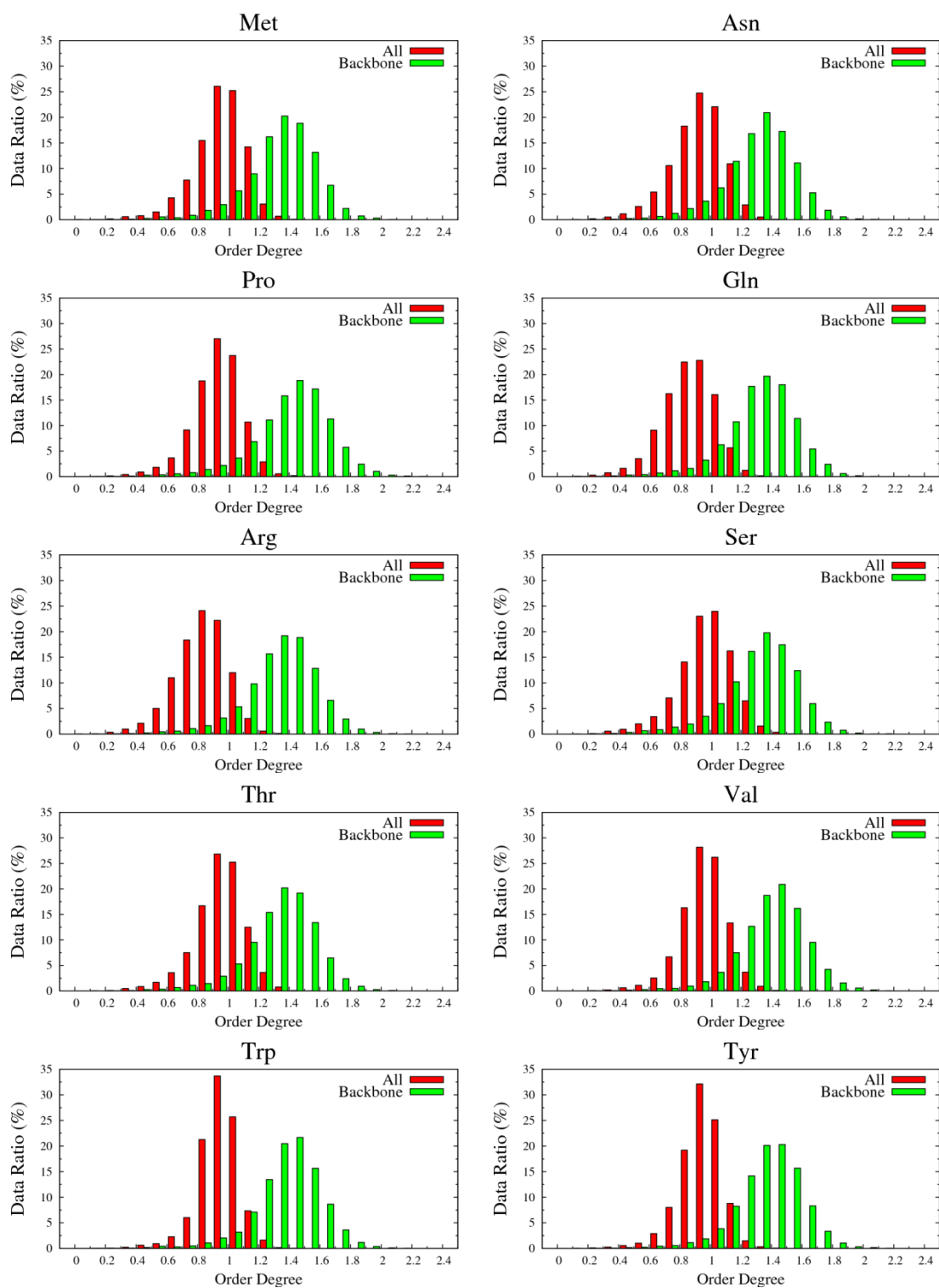


図 3.1 各アミノ酸の Order Degree (Met~Tyr)

図 3.1 より分かることを挙げていく。まず、All では、Glu, Lys, Arg における分布のピークが 0.8~0.9 と他に比べてやや低い区画にある。この事は、これらが荷電性残基であるため、タンパク質表面に位置することが多いので、結果として溶媒に露出して揺らいでしまうということであると考えられる。そのように考えると、親水性残基は Order Degree が小さい方へ分布しがちであり、一方で疎水性残基は高い方へ分布しがちであることが図 3.1 から読み取れる。

一方で、Backbone では、Gly を除く全残基において 1.4 付近をピークとする分布となっており、ピークの高さも 20 %程度であり、それ程顕著な違いが見られない。この事から、主鎖の disorder 具合はタンパク質の表面に存在していることには大きく影響されないということが示唆される。そして、唯一側鎖が無いアミノ酸である Gly は、その動きやすさのため Backbone における Order Degree のピーク位置が他より低くなっている。

より分かりやすく詳細に解析するため、各アミノ酸残基の All と Backbone において、Order Degree が  $\geq 0.8$ (All)、 $\geq 1.2$ (Backbone)である割合を求めた。その結果をそれぞれ表 3.2、表 3.3 に示した。

表 3.2 Order Degree(All)  $\geq 0.8$  の割合

Ala	91.25 %	Cys	94.47 %	Asp	76.53 %	Glu	61.24 %
Phe	89.29 %	Gly	93.25 %	His	81.56 %	Ile	85.78 %
Lys	48.63 %	Leu	84.87 %	Met	84.85 %	Asn	79.54 %
Pro	83.90 %	Gln	68.36 %	Arg	62.10 %	Ser	85.83 %
Thr	85.78 %	Val	88.76 %	Trp	89.80 %	Tyr	87.13 %

セルの色：赤 >90%、黄 80~90%、青 <80%

表 3.3 Order Degree(Backbone)  $\geq 1.2$  の割合

Ala	79.26 %	Cys	82.11 %	Asp	70.60 %	Glu	70.64 %
Phe	83.87 %	Gly	60.17 %	His	76.82 %	Ile	84.88 %
Lys	72.48 %	Leu	82.36 %	Met	78.47 %	Asn	73.96 %
Pro	83.85 %	Gln	75.38 %	Arg	77.54 %	Ser	75.10 %
Thr	78.34 %	Val	84.59 %	Trp	85.14 %	Tyr	83.44 %

セルの色：赤 >80%、黄 70~80%、青 <60%

まず、表 3.2 (All) を見ると、親水性残基の割合が低いことが分かる。溶媒に露出することが多いために側鎖が揺らぎやすいと考えられる。また、割合が最も高いのが Cys であるが、この残基は S-S 結合によって側鎖を繋ぐことができるため、揺らぎが少なる傾向にあると推測される。

一方、表 3.3 (Backbone) を見ると、Gly が突出して低い割合になっている。側鎖が無いことによって主鎖が動きやすくなっているためと考えられる。また、親水性残基も、疎水性残基に比べると割合が低くなる傾向があることが分かる。このことから、親水性残基は、溶媒に露出していることによって幾分か主鎖の動きやすさに影響していると考えられる。

## 3.2 各二次構造の Order Degree

次に、各二次構造において、Order Degree がどのように分布しているかを調べた。なお、二次構造情報は DSSP を用いて判定した。ここで再度、DSSP による構造判定の種類を表 3.4 に示しておく。

表 3.4 DSSP による構造判定の種類

G	$3_{10}$ helix	H	$\alpha$ helix
I	$\pi$ helix	T	hydrogen bonded turn
B	isolated $\beta$ -bridge	E	extended strand
S	bend	C (空白)	coil

そして、前節と同様に、Order Degree を 0.1 ずつに区切った区画において、その二次構造データ数の何%を占めているか調べた。その結果を図 3.2 に示す。なお、今回用いている 766 個のタンパク質のデータセットの各二次構造データ数を表 3.5 に示す。

表 3.5 各二次構造のデータ数

G	11151	H	77004
I	103	T	32095
B	3291	E	60193
S	23351	C (空白)	53632



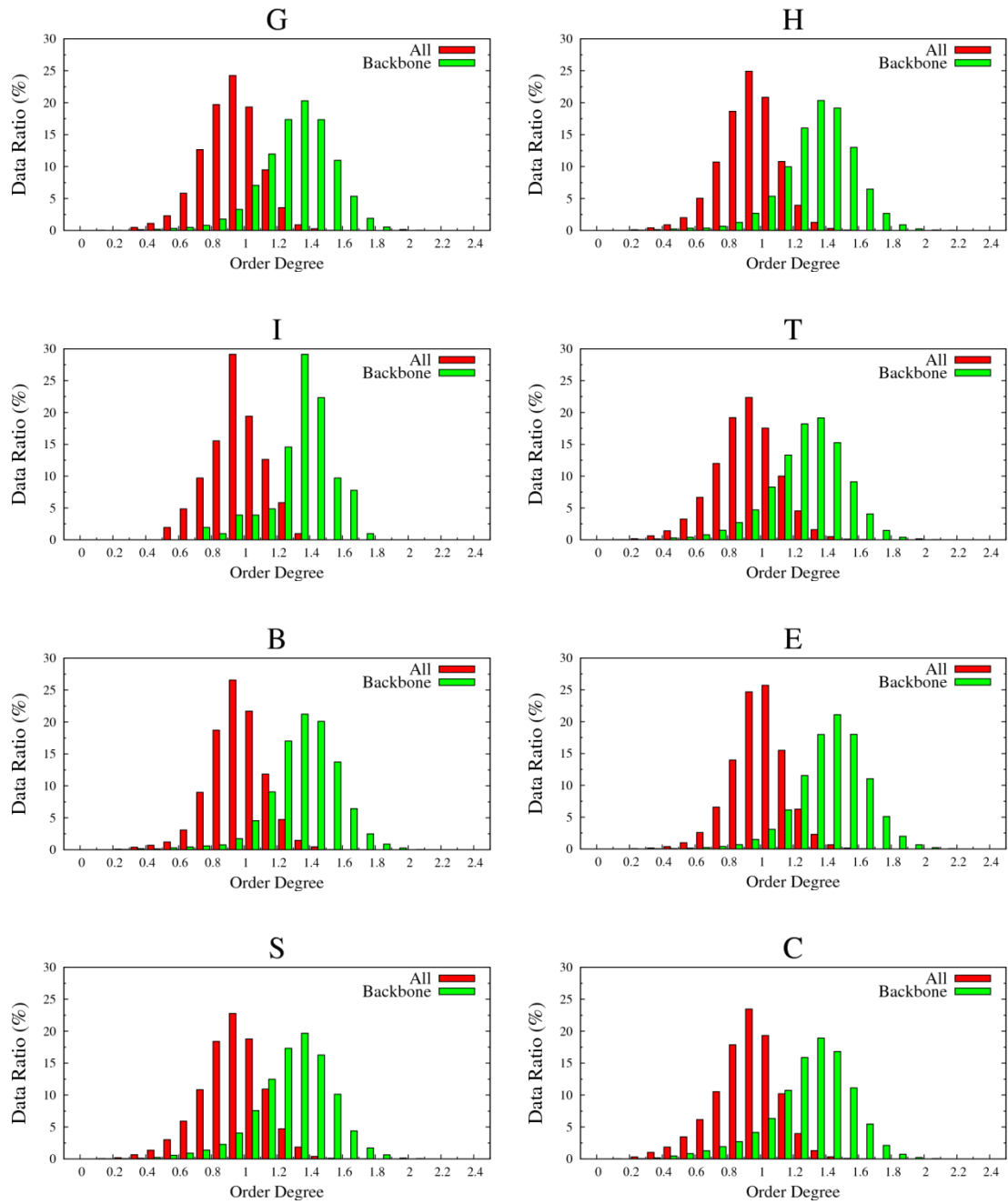


図 3.2 各二次構造の Order Degree

図 3.2 を見ると、E だけが All のピークが 1.1~1.2 で、Backbone のピークが 1.4~1.5 と、他の構造に比べてピークが 1 つ右へシフトしていることが分かる。このことから、シート構造は他の構造に比べて堅牢であることが分かる。

また、Backbone に着目してみると、S と C の Order Degree はかなり低い部分まで分布していることが分かる。このことから、ただ曲がっているだけ、もしくは何の構造も形成していない残基の主鎖は安定化する要因が少ない、もしくは無いので揺らぐ傾向にあることが分かる。

より詳細に調べるため、前節で行ったように、各二次構造の All と Backbone において、Order Degree が  $\geq 0.8$ (All)、 $\geq 1.2$ (Backbone)である割合を求めた。その結果をそれぞれ表 3.6、表 3.7 に示した。

表 3.6 Order Degree(All)  $\geq 0.8$  の割合

G	77.58 %	H	80.84 %
I	83.50 %	T	75.85 %
B	85.52 %	E	89.28 %
S	77.97 %	C (空白)	76.62 %

セルの色：赤 >90%、黄 80~90%、青 <80%

表 3.7 Order Degree(Backbone)  $\geq 1.2$  の割合

G	73.96 %	H	78.94 %
I	84.47 %	T	67.85 %
B	82.19 %	E	87.70 %
S	70.30 %	C (空白)	71.34 %

セルの色：赤 >80%、黄 70~80%、青 <60%

まず、All について分析すると、G,T,S,C が他に比べて割合が低いことが分かる。この理由についてそれぞれ考えてみる。T はターンという構造で、ターンは大抵タンパク質が丸くなるための折り返しであり、そのため、タンパク表面に位置することが多いと推測される。その結果、主に側鎖が溶媒に曝されて揺らぐと予測される。また、S と C というのは特定の安定した構造を持っている訳では無いという状態なので、揺らぐ割合が高くなっていると推測される。最後に G であるが、これは 3 残基で作る螺旋構造を意味していて、その構造が揺らぐ傾向にあるということは、この構造にそれ程安定化させる力が無いことが原因であると考えられる。4 残基螺旋の H、5 残基螺旋の I、と螺旋に関わる残基が増えるほど割合が増えていくのは、安定化させる能力の向上によるものであると見なせる。また、B,E、いわゆるシート構造においては割合が高く、他の構造に比べて堅牢な構造であるといえる。

次に、Backbone について見ると、T,S,C が低い傾向にあり、続いて G,H といった螺旋構造が続き、1 番割合が高いのは E という結果になっている。主鎖になると一段と T が低くなっている。T が S,C という何の構造を持たない状態より低い理由を考察すると、T はタンパク質の表面に位置する確率が高いためであると考えられる。そして、S,C という状態は必ずしもタンパク表面に位置しているとは限らないので、結果として、その割合が T を多少上回っていると予測される。また、主鎖においても螺旋構造は関わる残基数が多い程、揺らぐ割合が少ないという結果になっていることから、関わる残基数が多い程螺旋構造が堅牢になるといえる。最後に、主鎖においてもシート構造は堅牢である。

### 3.3 Accessible Surface Area と Order Degree

これまでの解析で Order Degree と溶媒の露出度との間に関係がありそうだと分かった。そこで、DSSP プログラムによって出力される Accessible Surface Area (ASA: 溶媒露出表面積)の値を使って、Order Degree と ASA の関係を調べてみた。各アミノ酸残基の All と Backbone それぞれで関係を調べた結果を図 3.3 に示す。なお、Order Degree は 0.1、ASA は  $5.0\text{\AA}^2$  の間隔で区切った区画において確率を計算している。

参考として、DSSP プログラムによって出力される ASA の分布を調べたところ、図 3.4 のようになった。

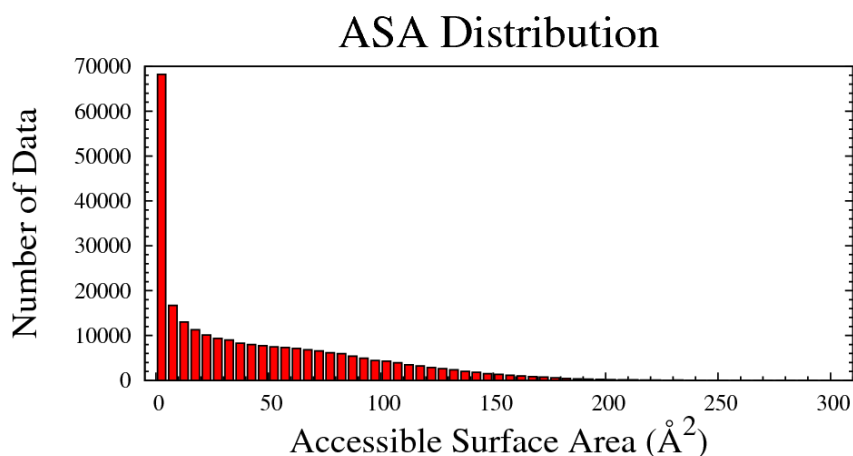


図 3.4 DSSP による ASA の分布

図 3.4 から分かるように、ASA が  $0\sim 5\text{\AA}^2$  である残基がほとんどである。これらの残基はタンパク質の内部に存在しているということになる。なお、今回のデータセットにおける ASA の最大値は  $307\text{\AA}^2$  であった。

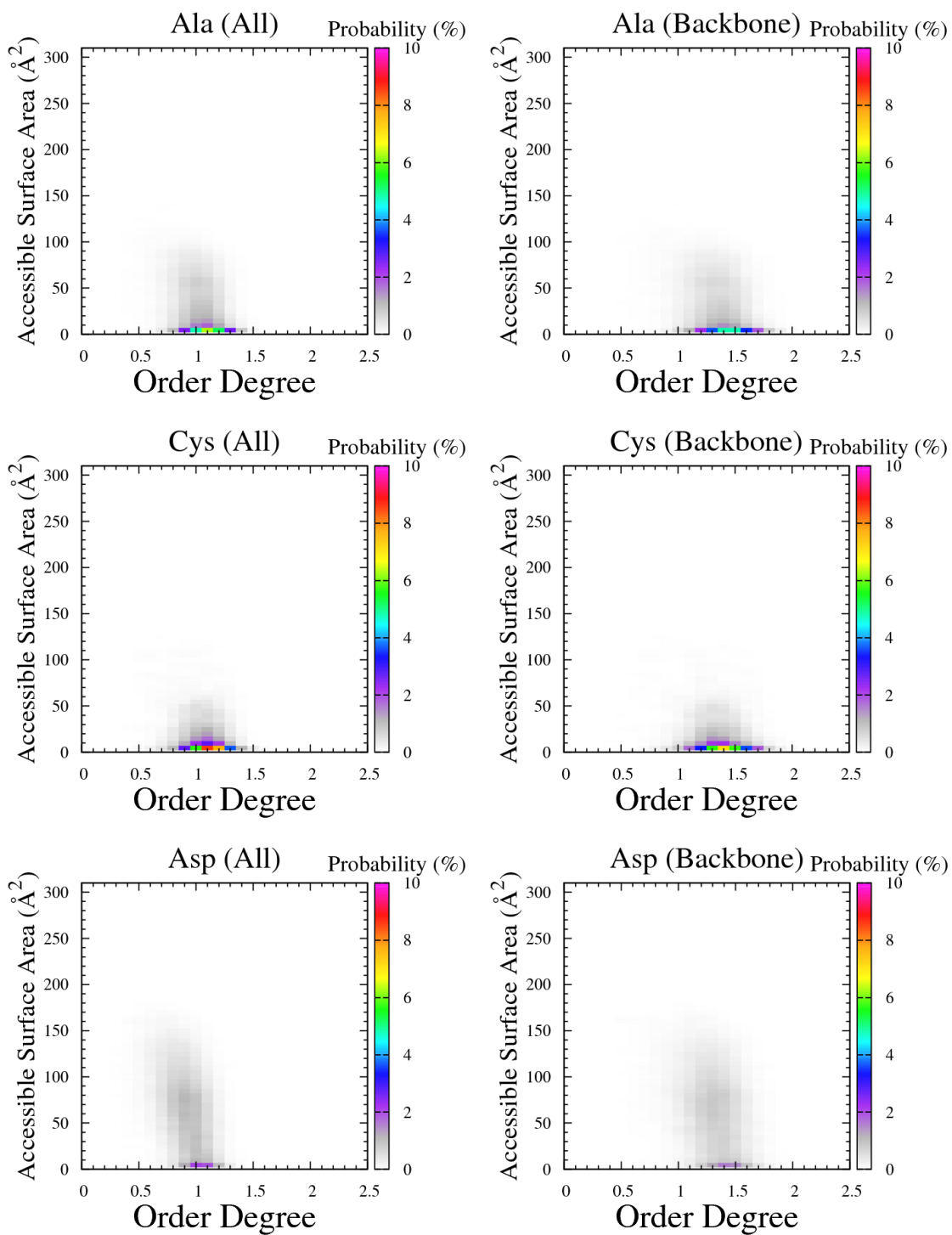


図 3.3 各アミノ酸の Order Degree と ASA (Ala, Cys, Asp)

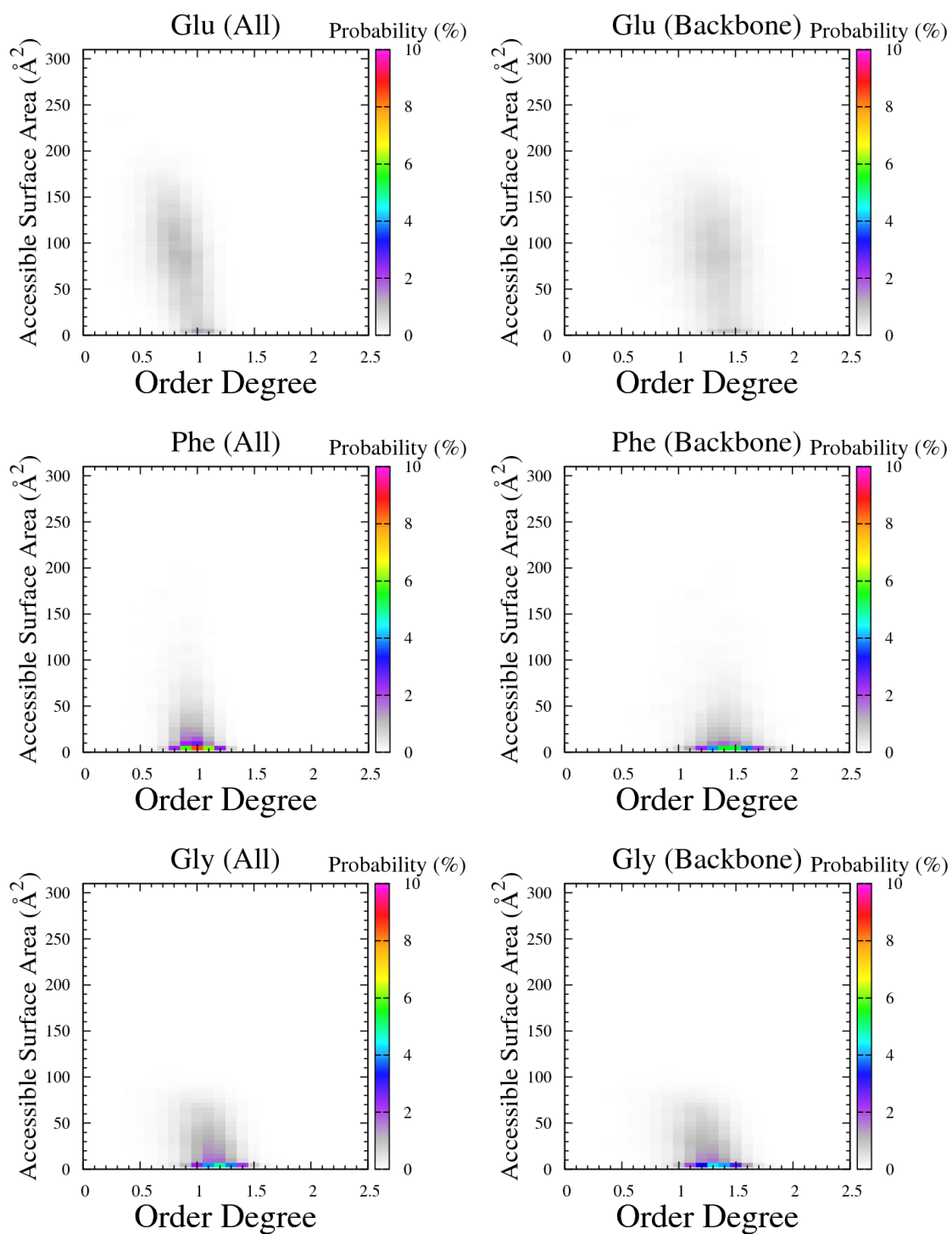


図 3.3 各アミノ酸の Order Degree と ASA (Glu, Phe, Gly)

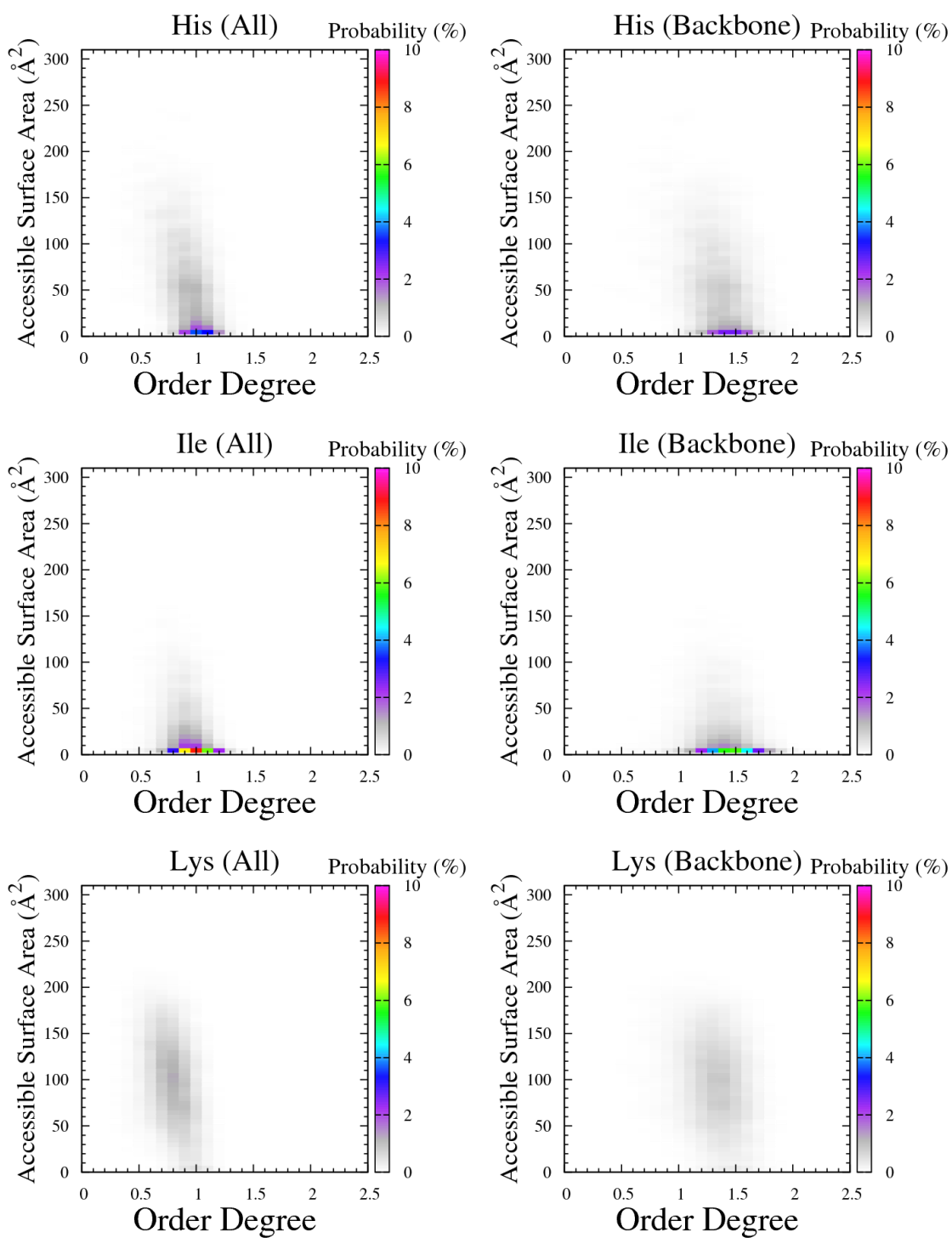


図 3.3 各アミノ酸の Order Degree と ASA (His, Ile, Lys)

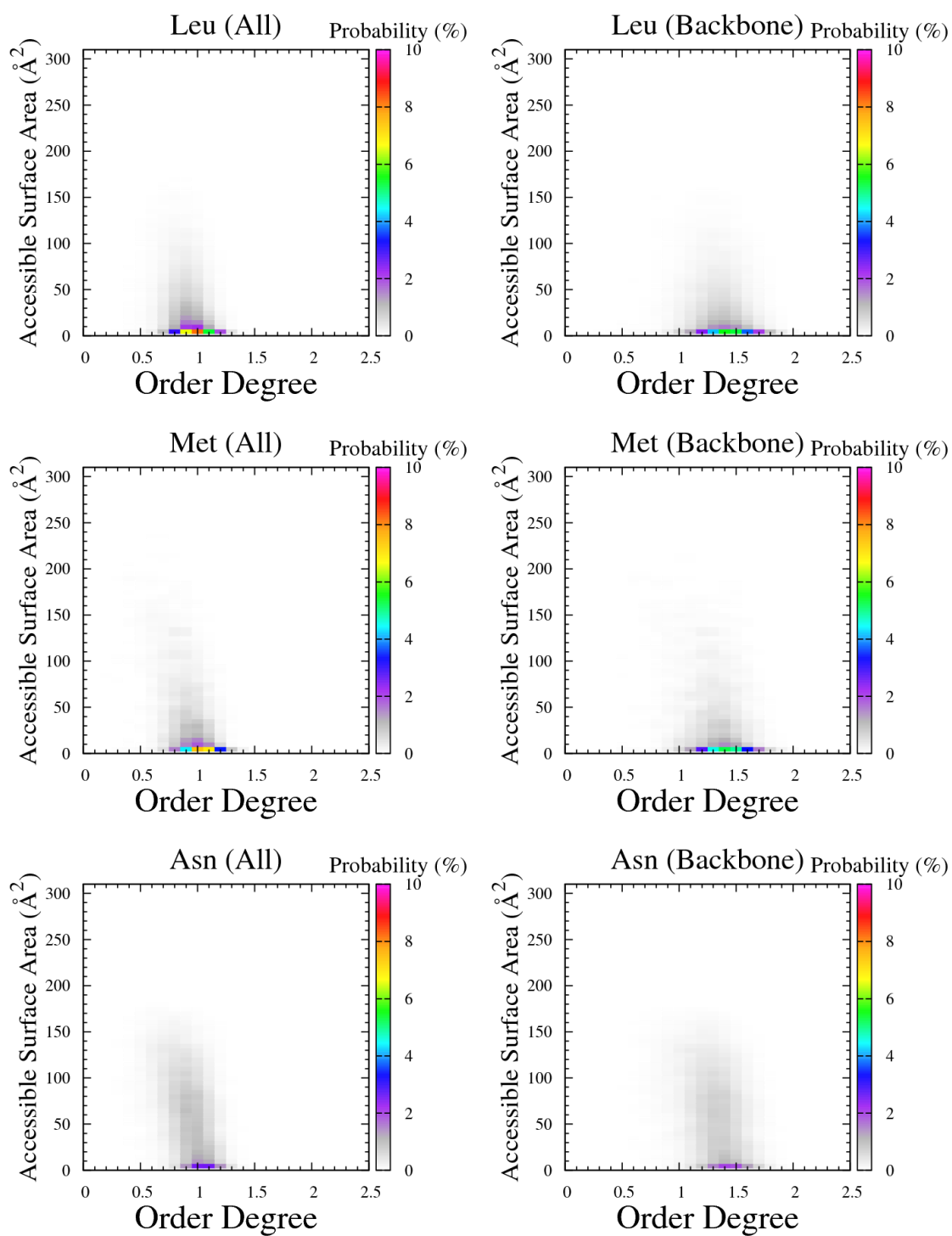


図 3.3 各アミノ酸の Order Degree と ASA (Leu, Met, Asn)



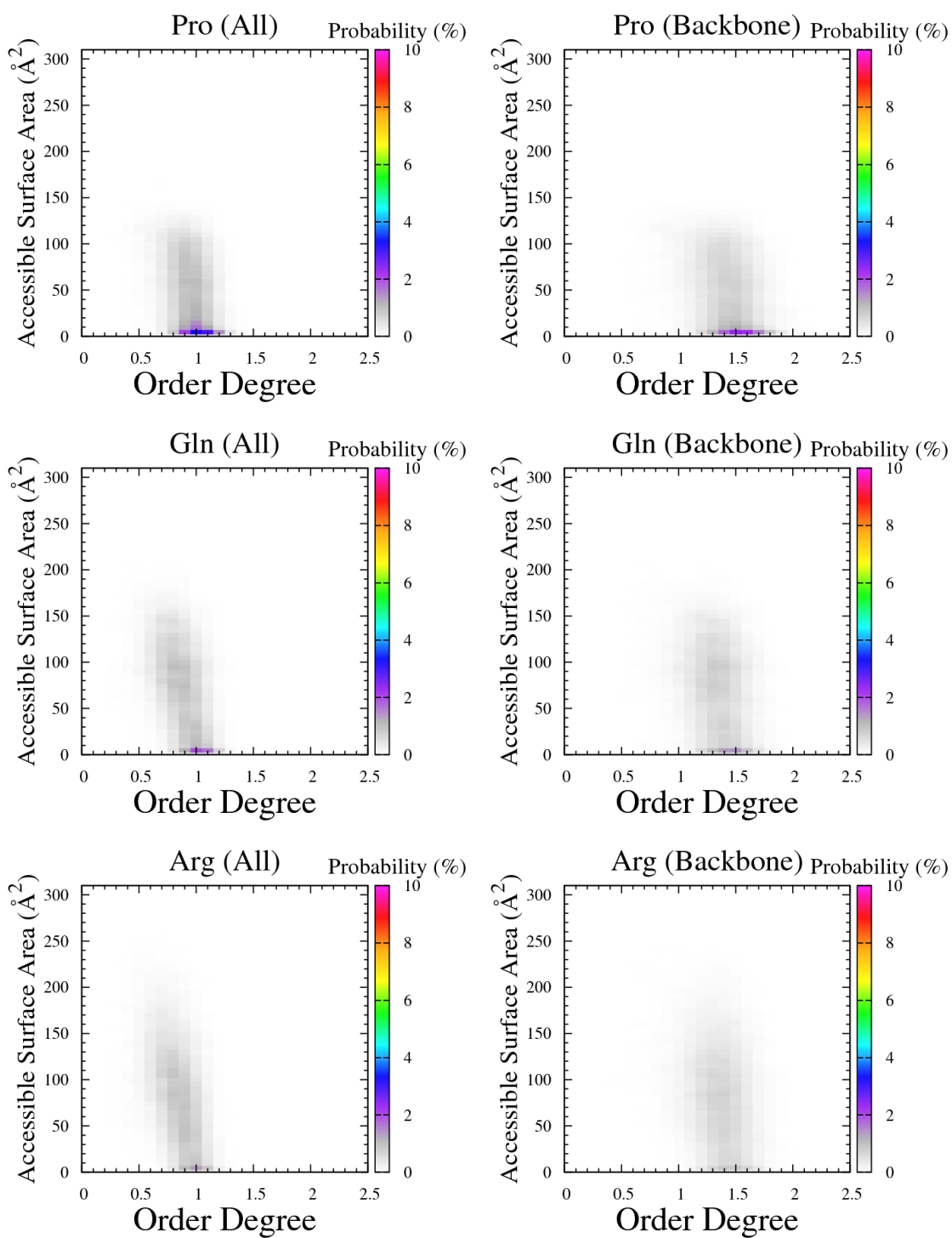


図 3.3 各アミノ酸の Order Degree と ASA (Pro, Gln, Arg)

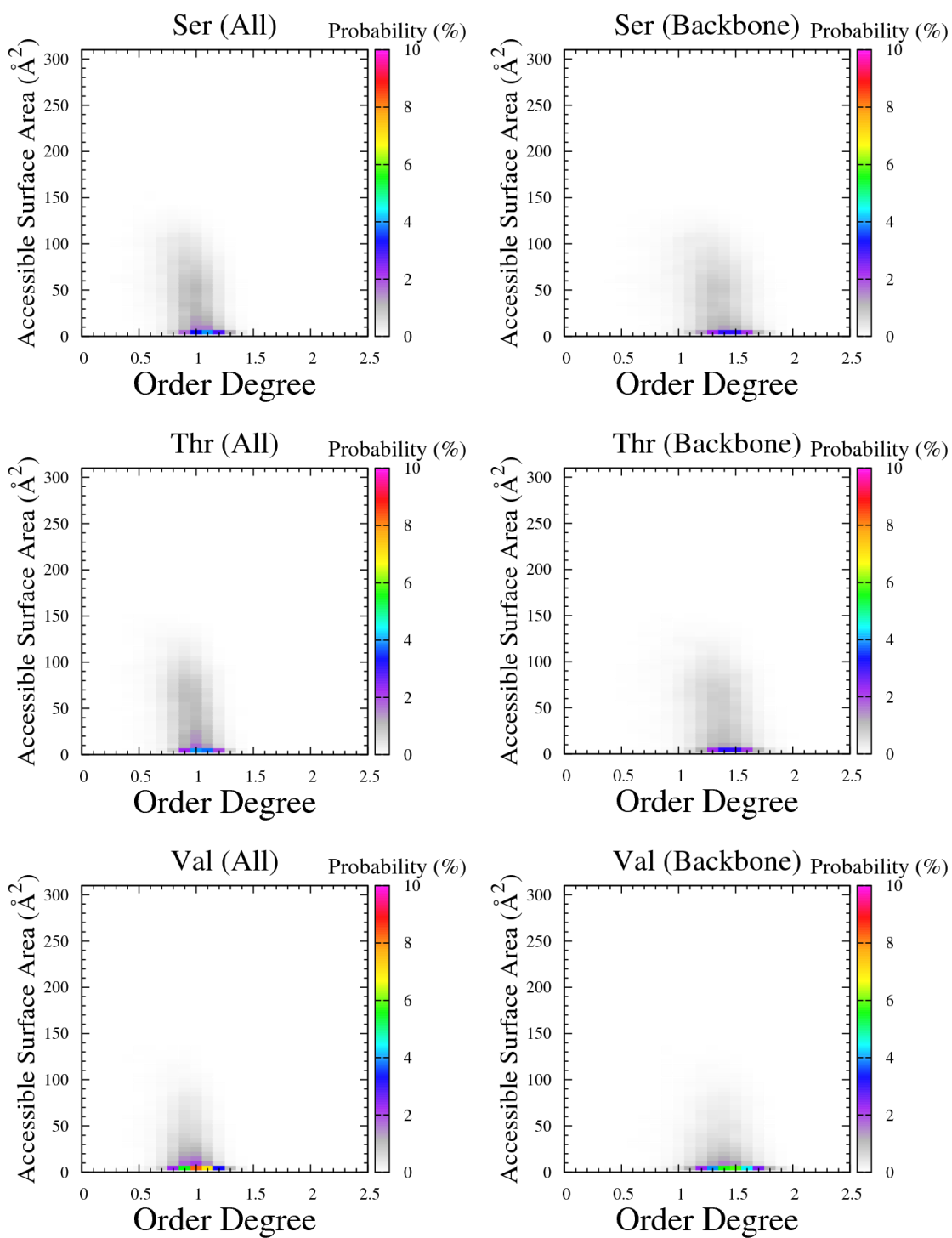


図 3.3 各アミノ酸の Order Degree と ASA (Ser, Thr, Val)

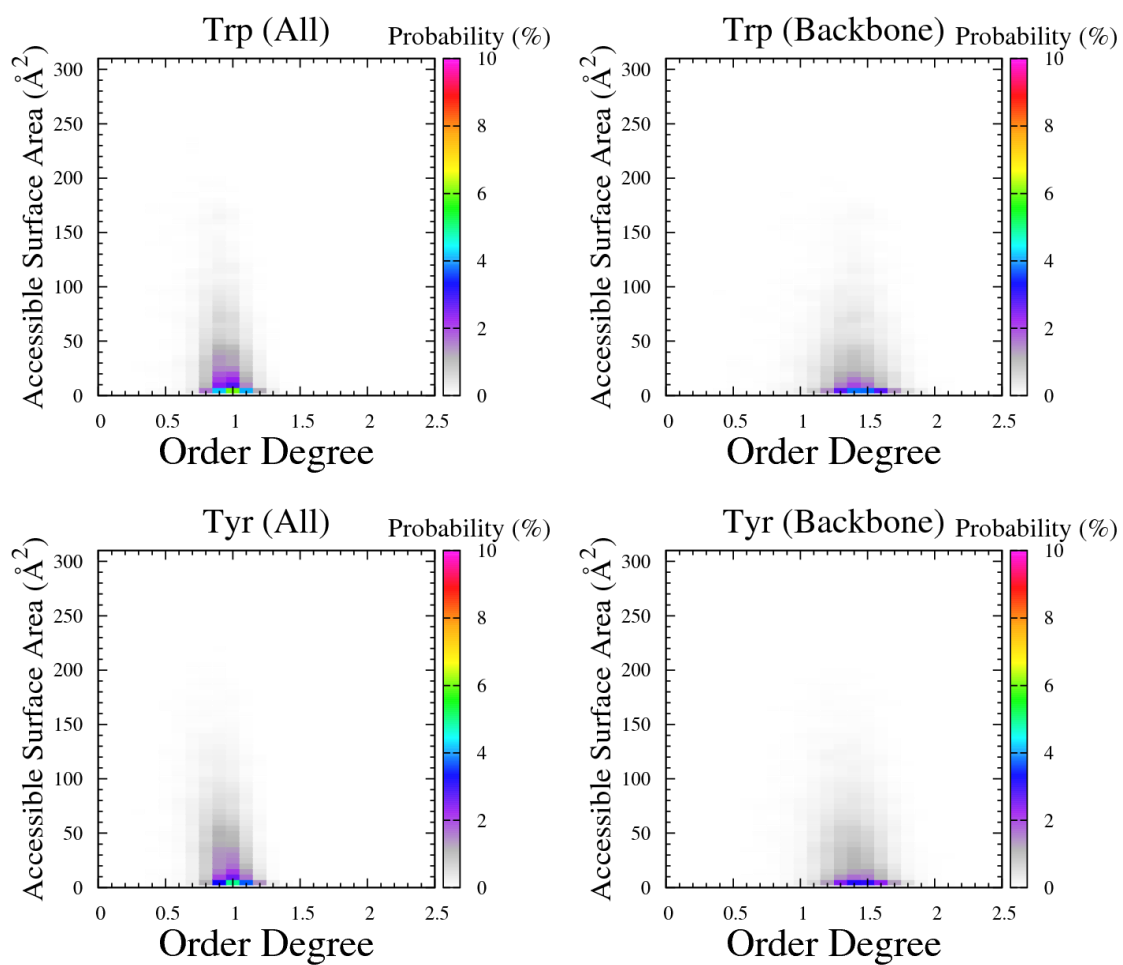


図 3.3 各アミノ酸の Order Degree と ASA (Trp, Tyr)

図 3.3 から分かることは、まず、Asp, Glu, Lys, Asn, Gln, Arg といった極性が強い残基の溶媒露出面積は広く、そして、その面積が広がるほど Order Degree(All)が低くなる確率が大きくなっている。一方で、Phe, Ile, Leu, Val, Trp, Tyr といった疎水性残基は露出面積が大きくなっても Order Degree は低くなりにくい。その理由は、疎水性側鎖の周りを水分子が囲むことによって動きを制限していると考えられる。

また、All と Backbone における最大確率の場所はどちらも露出表面積が 0~10 Å<sup>2</sup> の領域にあるが、Backbone における最大確率は All より必ず低くなっている。つまり、Backbone は All に比べて、それ程露出表面積が低いことを要求しない。言い換えれば、溶媒露出による揺らぎへ影響は、主鎖より側鎖の方が受けやすいということになる。

### 3.4 リガンド周囲の Order Degree

次に、リガンド周辺残基の Order Degree を調べた。なお、リガンド周辺残基の定義は、リガンドに属する原子から 5.0 Å 以内に原子が 1 個以上入っているアミノ酸残基とした (データは Gul Saad により提供された)。リガンドの種類は非常に多く全てを解析することはできないので、今回は周辺残基数の多い上位 10 個のリガンドについて解析を行った。その上位 10 個のリガンドとその周辺残基数を表 3.9 に示す。そして、各リガンドの周辺残基の各アミノ酸の割合をグラフにしたものを図 3.4 に示す。

表 3.9 解析対象リガンドと周辺残基数

リガンド名	周辺残基数
Ethylene Glycol	3798
Glycerol	3649
Sulfate Ion	2227
Calcium Ion	1601
Chloride Ion	1125
Heme	983
Acetate Ion	946
Zinc Ion	832
Sodium Ion	666
FAD	586

これらのリガンドに対してそれぞれ、周辺残基の All と Backbone の Order Degree の分布を 0.1 刻みでカウントした結果を図 3.5 に示す。

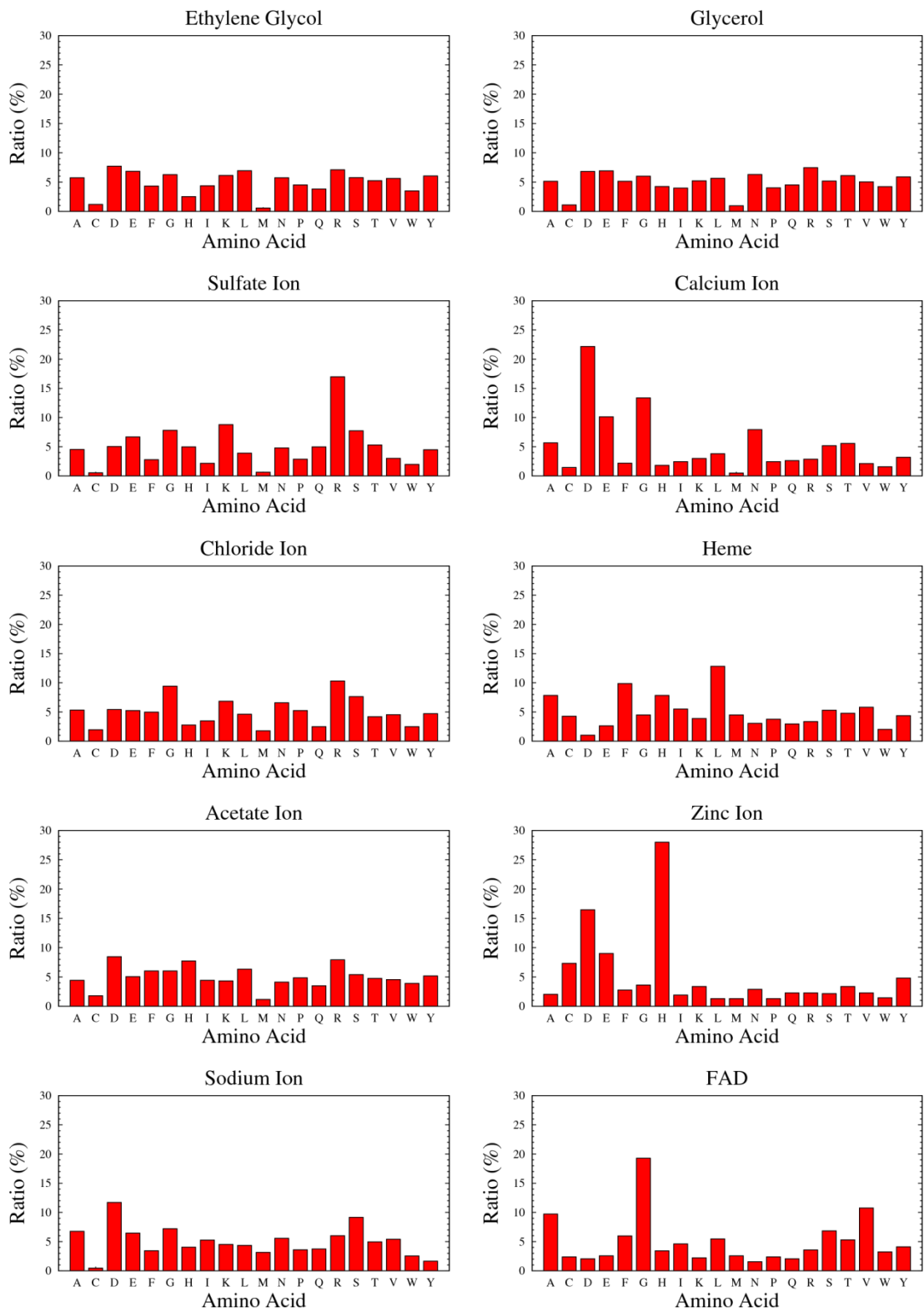


図 3.4 各リガンドの周辺残基の各アミノ酸の割合

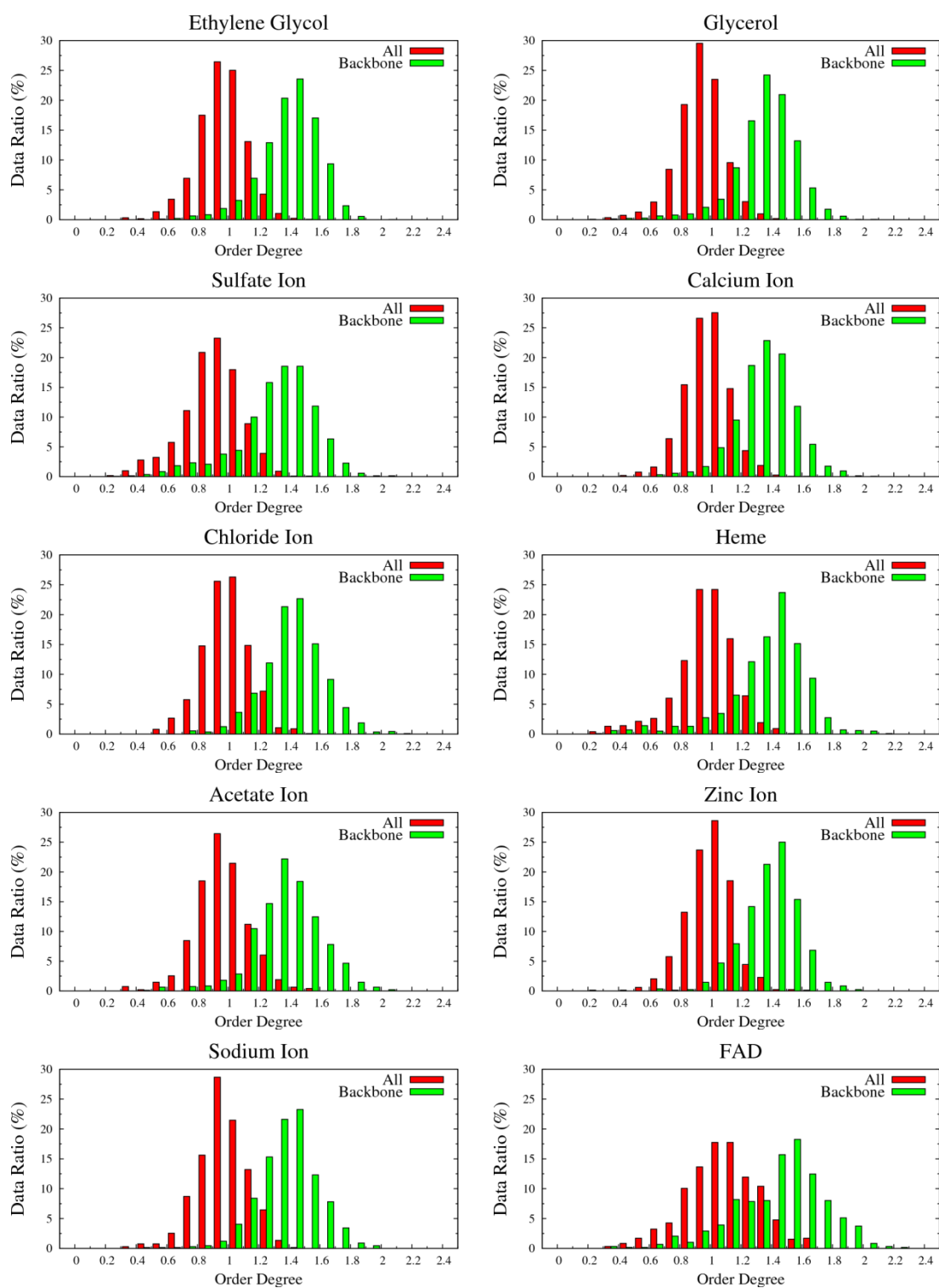


図 3.5 各リガンドの周辺残基 Order Degree

続いて、各リガンドの All と Backbone において、Order Degree が  $\geq 0.8$ (All)、 $\geq 1.2$ (Backbone)である割合を求めた。その結果をそれぞれ表 3.10、表 3.11 に示した。

表 3.10 Order Degree(All)  $\geq 0.8$  の割合

Ethylene Glycol	87.72 %
Glycerol	86.12 %
Sulfate Ion	75.98 %
Calcium Ion	91.06 %
Chloride Ion	90.76 %
Heme	86.06 %
Acetate Ion	86.58 %
Zinc Ion	91.35 %
Sodium Ion	86.94 %
FAD	89.61 %

セルの色 : 赤 >90%、黄 80~90%、青 <80%

表 3.11 Order Degree(Backbone)  $\geq 1.2$  の割合

Ethylene Glycol	86.14 %
Glycerol	82.75 %
Sulfate Ion	74.18 %
Calcium Ion	82.26 %
Chloride Ion	87.38 %
Heme	81.29 %
Acetate Ion	82.54 %
Zinc Ion	85.20 %
Sodium Ion	85.13 %
FAD	80.54 %

セルの色 : 赤 >80%、黄 70~80%、青 <60%



図 3.4、図 3.5 そして表 3.10 と表 3.11 を総合して分かることは、まず、リガンドの周辺残基の **Order Degree** は、平均的な **Order Degree** (2.2.3 参照) より高い値となる傾向があるということである。リガンドはその周辺残基と相互作用をして安定化する要因があるため、その場に存在しており、そのため、その周辺残基の **Order Degree** が平均より高くなると考える。

次に気付くことは、硫酸イオンだけが他のリガンドに比べて **Order Degree**( $\geq 0.8(\text{All})$ 、 $\geq 1.2(\text{Backbone})$ )の割合が低い。この理由として考えられることは、硫酸イオンの周辺残基の割合(図 3.4)で **Arg(R)**と **Lys(K)**が高い割合を占めていることである。3.1 節で見たように、この 2 つの残基は **Order Degree** が低くなる傾向がある。このことは、荷電性と側鎖の大きさに関係があると推測した。硫酸イオンの電荷はマイナスで、これら残基の側鎖の先端の電荷はプラスなので、その部分で相互作用して安定化するが、その側鎖は長いため、その側鎖の途中の部分、ひいては主鎖の部分においてはリガンドによる安定化の効果が薄れて、揺らいでしまうという可能性が考えられる。

最後に、図 3.5 の **FAD** における **Backbone** の分布が 2 つの山が重なったような分布をしているが、図 3.4 から分かるように **FAD** の周辺残基で最も多くの割合を占めている **Gly** によるものであると考えられる。**Gly** は側鎖にある残基に比べて **Backbone** の **Order Degree** が低い方へ分布する (3.1 節参照) ので、図 3.5 では 2 つの分布が重なったような形になっていると推測した。

### 3.5 低 Order Degree 領域の配列解析

最後の解析として、低 Order Degree が連続して続く配列がどのような組成をしているかを調べた。今回は、低 Order Degree 残基を Backbone で 0.7 以下の残基として定義した。このように定義して低 Order Degree 領域の配列とその出現回数を調べた結果、表 3.12 のようになった。

連続残基数	配列・出現回数
1	A 95
	C 7
	D 83
	E 87
	F 28
	G 204
	H 34
	I 26
	K 75
	L 54
	M 14
	N 60
	P 31
	Q 52
	R 49
	S 110
	T 48
	V 40
	W 9
	Y 22

2	AA 2	AC 1	AD 1	AF 2	AG 1	AH 1	AK 3	AN 1	AP 2	AR 1	AS 1
	AT 3	AV 1	AW 1	AY 1							
	CH 1	CT 2									
	DA 2	DD 1	DE 5	DF 2	DG 2	DK 1	DP 1	DQ 1	DR 2	DS 1	DT 1
	DV 2										
	EA 5	ED 2	EE 1	EF 1	EG 3	EI 1	EN 1	EQ 2	ES 3	EV 1	EW 1
	FA 1	FC 1	FE 1	FK 1	FL 1	FQ 1	FY 1				
	GA 1	GC 1	GD 1	GE 3	GG 9	GI 2	GK 4	GN 4	GP 1	GQ 1	GR 1
	GS 3	GT 4	GV 1	GW 2							
	HG 1	HH 7	HI 1	HK 2	HT 2						
	IK 1	IN 2	IS 1	IV 1							
	KA 1	KD 2	KE 3	KG 3	KH 1	KK 1	KN 1	KP 2	KR 1		
	LD 1	LE 2	LF 1	LG 2	LI 2	LK 1	LN 1	LP 1	LQ 1	LR 1	
	MA 2	ME 1	MG 1	MK 1	ML 1	MP 1	MQ 1	MR 1	MT 1	MV 2	
	NA 1	NG 1	NI 1	NK 1	NL 1	NP 2	NR 4	NS 3	NV 1		
	PA 3	PD 2	PE 3	PF 1	PR 1	PS 2	PV 1	PY 1			
	QD 1	QF 1	QG 3	QL 1	QN 1	QQ 1	QT 1	QV 1	QY 1		
	RA 1	RD 2	RE 2	RG 1	RR 1	RT 2	RV 1	RW 1			
	SA 5	SD 4	SE 3	SG 5	SK 2	SL 1	SM 1	SN 3	SP 1	SQ 1	
	SR 1	SS 4	ST 4	SV 1	SY 1						
	TA 1	TD 1	TE 1	TG 4	TH 1	TK 1	TL 1	TN 1	TQ 1	TR 1	TW 1
	VD 1	VE 1	VG 4	VK 1	VL 1	VN 1	VQ 2	VR 1	VS 1	VT 2	VV 1
	WE 2	WQ 2									
	YA 1	YL 1	YM 2	YP 1							

3	AAS 1 ADA 1 ADF 1 ADG 1 ADY 2 AFE 1 AGV 1 AKD 1 ANS 1 ARY 1 ATY 1 CGT 1 DAF 1 DAS 1 DEA 1 DEK 1 DGT 1 DSE 1 DWD 1 EAH 1 EEA 1 EEI 1 EEL 1 EGL 1 EIA 1 EIN 1 EKF 2 ETP 1 EVE 1 FEI 1 FGS 1 FHD 1 FLT 1 FVQ 1 FYE 1 GGL 1 GGV 1 GGW 1 GSV 1 GVE 1 GVG 1 GYG 1 HAE 1 HGN 1 HHH 1 HSF 1 IED 1 IKD 1 ILD 1 IRD 1 IRG 1 KAE 2 KDL 1 KEA 1 KGE 1 KKI 1 KSI 1 KTA 1 LDN 1 LGW 1 LHK 1 LKA 1 LNL 1 LRD 1 LST 1 MNK 1 MTA 1 NDD 1 NDG 1 NFG 1 NGS 1 NLY 1 NQH 1 NQQ 1 NRK 1 NSN 1 PNP 1 PQE 1 PRV 1 PSP 1 PVR 1 QDE 1 QEE 1 QKQ 1 QWG 1 REQ 1 RKA 1 RRA 1 RRP 1 SAL 1 SEF 1 SEW 1 SGG 1 SKT 1 SLG 1 SPD 1 SSL 1 STG 1 TAD 2 TFI 2 TSG 1 VDL 1 VFK 2 VGH 1 VNY 1 VRG 1 VVL 1 YHK 1 YIT 1 YQG 1
4	AAGE 1 AGDS 1 AGEP 1 AKAG 1 APAV 1 APGG 1 AQAA 1 AQKQ 1 AVGS 1 CNQR 1 DDFI 2 DKTM 1 DLF I 1 EADK 1 EEGK 1 EEGS 1 EEIA 1 EFLT 1 EKTF 1 ENLS 1 GAWK 1 GDGI 1 GENW 1 GGSH 1 GGSP 1 GGVS 1 GGVT 1 GRFG 1 GRLS 1 GRVE 1 GVDG 1 HRLQ 1 IDGK 1 KDGS 1 LHEL 1 LNRI 1 LQDG 1 LSND 1 MATG 1 NDAT 1 NKEI 1 NPDA 1 PAAL 1 PIIK 1 QADG 1 QTHT 1

	RVPL 1 RYVE 1 SASK 1 SEAQ 1 SSEF 1 TAGS 1 TGGP 1 TLDE 1 TPLR 1 TQPN 1 TVTR 1 TVTS 1 VEQP 1 VEVL 1 VVLS 1 YQDF 1
5	APRGG 1 CVREG 1 DLAHQ 1 DSQGV 1 DSSRR 1 EDGKK 1 ERQEF 1 ETANG 1 FYWAH 1 GKDRL 1 IVERV 1 KESLQ 2 KGSQA 1 KKGCA 1 KNEEE 1 LGGVV 1 LKGKL 1 LVRSP 1 LYSSA 1 NEPGS 1 NTAGL 1 PDGSY 1 PGYTG 1 QAQSS 1 QFRGR 1 RGNVS 1 RKKEF 1 RYQDF 1 SEDWV 1 SGWPP 1 SNETL 1 SQTPD 1 SWTAL 1 TAGGK 1 TPNSR 1 TTTGD 1 VGSPG 1 VISHK 1 VLRGT 1 YEVSL 1 YSNVY 1
6	DDRHLL 1 DPDTEH 1 DPLTRT 1 EEETRE 1 EQINWI 1 ETIAAE 1 GGREDS 1 GKSGRT 1 HAPDYY 1 HESAGT 1 IKQLIE 1 LALGRA 1 LEQLLE 2 LGEMVP 1 LIIPPL 1 MAYTDG 1 MDEPDL 1 QFHQEI 1 QLPTPV 1 TGFGFA 1 TKTSRG 1 VPIIYE 1 WLKGKL 1 YNEPGS 1

7	ADLPNHF 1 ALDSLVL 1 ANSIKKI 1 APVNYKA 1 ATAEAAD 1 DGSTGEF 1 EQINWIY 1 FTKKDSV 1 GIKMEGL 1 IGIHGQN 1 KDHSFFG 1 KLEEGS 1 LGWVASK 1 RCGLDLI 1 RFSSAWG 1 SLPASKQ 1 SMSNCAA 1 TLIIPPL 1 TRDQLTE 1 TYPEFLA 1 YLRPSGR 1 YNAGEGD 1
8	DEVSPRML 1 DYLNPGSI 1 EDGIRLVS 1 EREQKTA 1 FSNREGKL 1 GIKMEGLE 1 IITKTSRG 1 LAPNVLAD 1 LDMQISID 1 LNGKAVTT 1 LWAKLDAD 1 PINALSNS 1 TRTEVEKK 2 WLDPLTDS 1
9	AIYGDVLNL 1 ELAGSVYDG 1 GSSPAYLKE 1 GYLVAECT 1 PPHNRWAF 1 SAGDLASFQ 1 VQIHSVITA 1
10	AMTDKGFTNT 1 LARVGRMLD 1 LDSAGELPTL 1 LNGKAVTTED 1 LPASKQMLAV 1 MDEPDLKDLF 1 YMRDPTFVRL 1 YSIGNLKPDT 1
11	KYMRDPTFVRL 1 RFQRPDYIQDA 1 RFWNNFGANAL 1 RGQIESYLVAN 1 TEGSSPAYLKE 1 VVAEYYRAGFA 1

12	ADFESLLSRPV 1 DAPSQIEVKDVT 1 EIAVQIHSVITA 1 MPPSQPVDGFEL 1 VDTGNGGESHSE 1 YQGASLHNQTDR 1
13	GILLDVSRALDAV 1
14	ADFESLLSRPVLE 1 APDAVAQQWAIFRA 1 EDYVDPASEVQTHD 1 GDMSSNPAKETFTT 1 KVQYPDYYAIIKEP 1 QAKSGTGKTCVFST 1 SAGWRTRRHGDPAD 1 VMTNLGEKLTDEEV 1 WARTNLQHTNDKLF 1
15	VFIGGTPLSQDKTRL 1
16	AALTLLVYIARAEETI 1 DHPTLTFNEDFKIFLT 1 GWVDLTTGARSPYNAP 1 LNRIADHPTLTFNEDF 1 VFIGGTPLSQDKTRLK 1
17	LVDEGRIDPAQPVTEYV 1
18	GRIKQLIELDYLNPGSIR 1 PSLLQAFDLTADPRWRQV 1 WRFLKGHPVPDPKNNAT 1
19	IVVATNPSGRLISELFQKL 1 LRAAGFERPSPVQLKA IPL 1 WRLPPDWLAVNRATGALS I 1 YGIKDVPGDRTTIDLTEDE 1
20	IVVATNPSGRLISELFQKLP 1
25	GHALARLGATAAQLPDLEQRLEQTY 1
29	DLKTIAQRIQNGSYKSIHAMAKDIDLLAK 1
45	SKVQYPDYYAIIKEPIDLKTIAQRIQNGSYKSIHAMAKDIDLLAK 1
49	GLRAAGFERPSPVQLKA IPLGRCGLDLIVQAKSGTGKTCVFSTIALDSL 1

表 3.12 低 Order Degree 領域の配列

続いて、この低 Order Degree 領域の配列におけるアミノ酸の出現傾向を調べた。図 3.6 は、各アミノ酸について出現する割合をグラフにしたものである。図 3.7 は、各アミノ酸の全体に占める割合（表 3.1 参照）を 1 とした場合に、どれ程この領域特異的に出現しているかを表している。

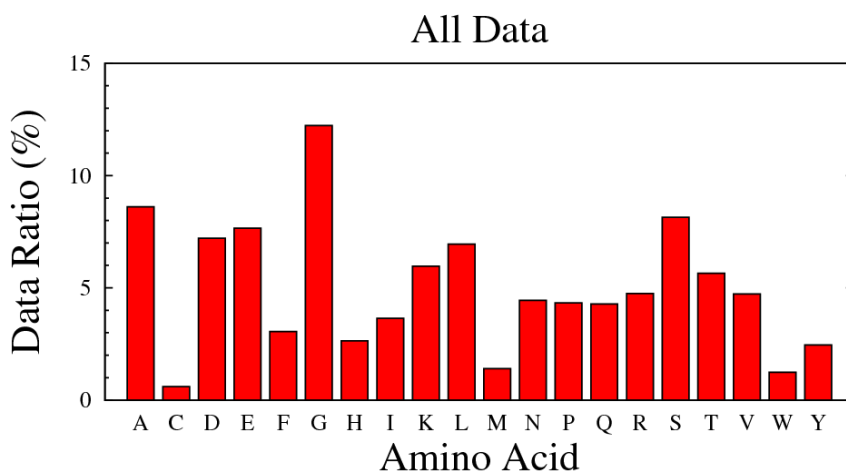


図 3.6 アミノ酸の出現割合

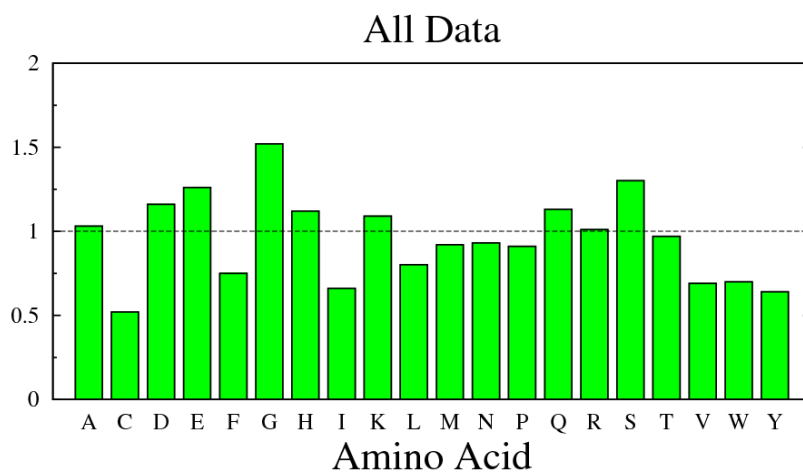


図 3.7 平均的な出現頻度との比較



表 3.12、図 3.6、図 3.7 から分かることは、まず Gly が多いということである。Gly は先のアミノ酸ごとに調べた Order Degree の分布で見たように、主鎖原子におけるその値が他の残基に比べて低かったことから予想通りである。その理由も、Gly には側鎖が無いので、主鎖二面角の回転におけるエネルギー障壁が小さく、動きやすいということに起因していると考えられる。

図 3.7 において、その他の残基を分析すると、Asp, Glu, His, Lys, Gln, Ser などの極性残基が平均よりも多く現れ、疎水性残基である Phe, Ile, Val, Trp, Tyr や架橋する Cys は平均よりも出現頻度は少ない。基本的に disorder しているところはループやターンを形成している、もしくはタンパク質の末端であることが多いと予想される。つまり、タンパク質の表面で溶媒に曝露していることになるので極性残基が多く見られ、逆に疎水性残基は見られなくなる傾向が現れていると推測する。

まとめると、低 Order Degree (Backbone) 領域において、よく見られる残基の傾向は側鎖が小さい、親水性であるといえる。

## 第4章 結論

本研究では、**disorder** の指標として、電子密度に基づいた **Order Degree** という尺度を計算し、解析してその傾向を見てきた。これまでの **missing residue** に基づいた **disorder** 領域に関する実験から、**disorder** 領域の配列傾向は **order** 領域に比べて、芳香族アミノ酸の割合が少ない、正味荷電が高い、親水性に関する指標が高い、配列の複雑度が低い（同じアミノ酸が続く等）ということが分かっている。3.5 節では、**order** 領域とされている中から **disorder** 領域のように揺らいでいると推測される配列を探し出した。探し出された配列の傾向は、先に示した **disorder** 領域の研究結果とほぼ同じである。このことから、**order** とされている配列でも、**disorder** 領域における配列傾向に似た部分配列は揺らぐ傾向にあると結論付けられる。

今後の展望として、この **Order Degree** のデータを学習して、アミノ酸配列から各残基の **disorder** 度合いを予測するプログラムを作ることや、立体構造予測の際の補助的なパラメータとして活用していくことを考えている。

# 総括

本論文では、第1部で二次構造予測に基づく拘束 MD シミュレーションによる天然構造への fold に関する研究、そして第2部では電子密度データに基づいた disorder に関する研究について述べた。

第1部の主な目的は、MD シミュレーションで対象タンパク質を伸展構造から天然構造へと fold させることにあった。しかし、それを成し遂げるために克服しなければならない問題として、天然構造を最安定構造としなければならないということがあった。そこで本研究では、二次構造予測の情報を付加することにより、天然構造をより安定させることができるかどうかを検証した。実験は概ね成功したが、予測自体が大きく外れている場合には逆効果となることも考えられるので、今後は、更に他の観点からの予測情報を付加してより天然構造を探索しやすくできるよう改善をしていきたいと考えている。

第2部では、disorder の尺度を自身で新たに定義して、それを解析することで何か新たな知見が得られないか調べた。その結果、order とされている領域の中にも disorder 寄りの領域があり、その領域は従来の定義による disorder 領域の傾向と似ていることが分かった。今回の研究では、どのような配列が disorder となり易いかという傾向まで解析したが、今後はその disorder 領域がどのように動いていくのかということ調べていきたいと考えている。

# 謝辞

本研究を行う環境を提供して頂き、そして指導して頂いた清水謙多郎教授に感謝します。アグリバイオインフォマティクス教育研究ユニットの寺田透特任准教授には、この論文に記されている研究全てにおいて、何度も議論を交わし、研究の道筋を示して頂いたことを感謝します。中村周吾准教授には **PSIPRED** の予測結果に関するデータを解析・提供して頂くと共に、アドバイスを頂いたことに感謝します。このおかげで、二次構造拘束のエネルギーマップを完成させることが出来ました。角越和也助教には、セミナー発表の際に有益な指摘を多くして頂き、有難うございました。また、本研究室所属修士課程1年の **Gul Saad** さんには、リガンド結合に関する情報を提供して頂きました。この事により、**disorder** に関する知見をより深めることができました、有難うございました。最後に、セミナーにおける私の研究報告を聞いていただき、かつ多くの意見を出して頂いた生物情報工学研究室の先輩・後輩に感謝します。

## 参考文献

- [1] Case TAD DA, Cheatham TE, Simmerling CL, Wang J, Duke RE, R, Luo KMM, Pearlman DA, *et al.* "AMBER 9." San Francisco, CA. (2006).
- [2] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, Jr., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell and P.A. Kollman. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules." *J. Am. Chem. Soc.* **117**, 5179-5197 (1995).
- [3] P.A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot and A. Pohorille. "The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data." *In Computer Simulation of Biomolecular Systems*. **3**, A. Wilkinson, P. Weiner and W.F. van Gunsteren, Ed. Elsevier, 83-96 (1997).
- [4] M.D. Beachy and R.A. Friesner. "Accurate ab initio quantum chemical determination of the relative energies of peptide conformations and assessment of empirical force fields." *J. Am. Chem. Soc.* **119**, 5908-5920 (1997).
- [5] J. Wang, P. Cieplak and P.A. Kollman. "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?." *J. Comput. Chem.* **21**, 1049-1074 (2000).
- [6] Y. Duan, C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo and T. Lee. "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations." *J. Comput. Chem.* **24**, 1999-2012 (2003).
- [7] M.C. Lee and Y. Duan. "Distinguish protein decoys by using a scoring function based on a new Amber force field, short molecular dynamics simulations, and the generalized Born solvent model." *Proteins* **55**, 620-634 (2004).
- [8] A. Onufriev, D. Bashford and D.A. Case. "Exploring protein native states and large-scale conformational changes with a modified generalized Born model." *Proteins* **55**, 383-394 (2004).
- [9] M. Feig, A. Onufriev, M. Lee, W. Im, D. A. Case and C. L. Brooks, III. "Performance comparison of the generalized Born and Poisson methods in the calculation of the electrostatic solvation energies for protein structures." *J. Comput. Chem.* **25**, 265-284 (2004).
- [10] Tohru Terada and Akinori Kidera "Generalized form of the conserved quantity in

- constant-temperature molecular dynamics" *J. Chem. Phys.* **116**, 33 (2002).
- [11] Tohru Terada, Yo Matsuo, and Akinori Kidera "A method for evaluating multicanonical potential function without iterative refinement: Application to conformational sampling of a globular protein in water" *J. Chem. Phys.* **118**, 4306 (2003).
- [12] Bernd A. Berg and Thomas Neuhaus "Multicanonical algorithms for first order phase transitions" *Physics Letters B* **267**, 249-253 (1991).
- [13] Bernd A. Berg and Thomas Neuhaus "Multicanonical ensemble: A new approach to simulate first-order phase transitions" *Phys. Rev. Lett.* **68**, 9-12 (1992).
- [14] Ulrich H. E. Hansmann and Yuko Okamoto "Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem" *Journal of Computational Chemistry* **14**, 1333-1338 (1993).
- [15] Ulrich H.E. Hansmann and Yuko Okamoto "Comparative study of multicanonical and simulated annealing algorithms in the protein folding problem" *Physica A: Statistical Mechanics and its Applications* **212**, 415-437 (1994).
- [16] Yuko Okamoto and Ulrich H. E. Hansmann "Thermodynamics of Helix-Coil Transitions Studied by Multicanonical Algorithms" *J. Phys. Chem.* **99**, 11276-11287 (1995).
- [17] G. R. Smith and A. D. Bruce "A study of the multi-canonical Monte Carlo method " *J. Phys. A: Math. Gen.* **28**, 6623 (1995).
- [18] Ulrich H.E. Hansmann, Yuko Okamoto and Frank Eisenmenger "Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble" *Chemical Physics Letters* **259**, 321-330 (1996).
- [19] Nobuyuki Nakajima, Haruki Nakamura, and Akinori Kidera "Multicanonical Ensemble Generated by Molecular Dynamics Simulation for Enhanced Conformational Sampling of Peptides" *J. Phys. Chem. B* **101**, 817-824 (1997).
- [20] Nobuyuki Nakajima, Junichi Higo, Akinori Kidera and Haruki Nakamura "Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation" *Chemical Physics Letters* **278**, 297-301 (1997).
- [21] Ulrich H. E. Hansmann and Yuko Okamoto "Finite-size scaling of helix-coil transitions in poly-alanine studied by multicanonical simulations" *J. Chem. Phys.* **110**, 1267 (1999).
- [22] Satoshi Ono, Nobuyuki Nakajima, Junichi Higo and Haruki Nakamura "The multicanonical weighted histogram analysis method for the free-energy landscape along structural transition paths" *Chemical Physics Letters* **312**, 247-254 (1999).
- [23] Yuji Sugita and Yuko Okamoto "Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape" *Chemical Physics Letters* **329**, 261-270 (2000).
- [24] Ayori Mitsutake and Yuko Okamoto "Helix-coil transitions of amino-acid

- homo-oligomers in aqueous solution studied by multicanonical simulations" *J. Chem. Phys.* **112**, 10638 (2000).
- [25] Yuko Okamoto "Protein folding simulations and structure predictions" *Computer Physics Communications* **142**, 55-63 (2001).
- [26] Jucichi Higo, Oxana V. Galzitskaya, Satoshi Ono and Haruki Nakamura "Energy landscape of a  $\beta$ -hairpin peptide in explicit water studied by multicanonical molecular dynamics" *Chemical Physics Letters* **337**, 169-175 (2001).
- [27] Shura Hayryan, Chin-Kun Hu, Shun-Yun Hu and Rung-Ji Shang "Multicanonical parallel simulations of proteins with continuous potentials" *Journal of Computational Chemistry* **22**, 1287-1296 (2001).
- [28] M. Isobe, H. Shimizu and Y. Hiwatari "A multicanonical molecular dynamics study for a model protein-g" *Computer Physics Communications* **142**, 144-147 (2001).
- [29] Ulrich H. E. Hansmann "Protein-folding simulations in generalized ensembles" *International Journal of Quantum Chemistry* **90**, 1265-1554 (2002).
- [30] A. Baumketner and Y. Hiwatari "Running Multicanonical Simulations on Deformed Energy Surface: Application to a Model Protein" *J. Phys. Soc. Jpn.* **71**, 1001-1002 (2002).
- [31] Soonmin Jang, Youngshang Pak and Seokmin Shin "Multicanonical ensemble with Nosé-Hoover molecular dynamics simulation" *J. Chem. Phys.* **116**, 4782 (2002).
- [32] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto "Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test" *J. Chem. Phys.* **118**, 6664 (2003).
- [33] Yukihisa S. Watanabe, Yoshifumi Fukunishi and Haruki Nakamura "Modelling of third cytoplasmic loop of bovine rhodopsin by multicanonical molecular dynamics" *Journal of Molecular Graphics and Modelling* **23**, 59-68 (2004).
- [34] George Chikenji, Yoshimi Fujitsuka and Shoji Takada "Protein folding mechanisms and energy landscape of src SH3 domain studied by a structure prediction toolbox" *Chemical Physics* **307**, 157-162 (2004).
- [35] Vasyi Aleksenko, Wooseop Kwak and Ulrich H.E. Hansmann "Generalized-ensemble simulations of all-atom protein models" *Physica A: Statistical Mechanics and its Applications* **350**, 28-37 (2005).
- [36] Michael Bachmann, Handan Arkin and Wolfhard Janke "Multicanonical study of coarse-grained off-lattice models for folding heteropolymers" *Phys. Rev. E* **71**, 031906 (2005).
- [37] Daisuke Mitomo, Yukihisa S. Watanabe, Narutoshi Kamiya and Junichi Higo "Explicit and GB/SA solvents: Each with two different force fields in multicanonical conformational sampling of a 25-residue polypeptide" *Chemical Physics Letters* **427**,

- 399-403 (2006).
- [38] Daisuke Satoh, Kentaro Shimizu, Shugo Nakamura and Tohru Terada "Folding free-energy landscape of a 10-residue mini-protein, chignolin" *FEBS Letters* **580**, 3422-3426 (2006).
- [39] S. G. Itoh and Y. Okamoto "A new generalized-ensemble algorithm: multicanonical-multioverlap algorithm" *Molecular Simulation* **33**, 83-89 (2007).
- [40] Takao Yoda, Yuji Sugita, Yuko Okamoto "Cooperative folding mechanism of a  $\beta$ -hairpin peptide studied by a multicanonical replica-exchange molecular dynamics simulation" *Proteins: Structure, Function, and Bioinformatics* **66**, 846-859 (2007).
- [41] Satoru G. Itoh and Yuko Okamoto "Effective sampling in the configurational space of a small peptide by the multicanonical-multioverlap algorithm" *Phys. Rev. E* **76**, 026705 (2007).
- [42] Olav Zimmermann and Ulrich H.E. Hansmann "Understanding protein folding: Small proteins in silico" *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1784**, 252-258 (2008).
- [43] Narutoshi Kamiya, Yasushige Yonezawa, Haruki Nakamura and Junichi Higo "Protein-inhibitor flexible docking by a multicanonical sampling: Native complex structure with the lowest free energy and a free-energy barrier distinguishing the native complex from the others" *Proteins: Structure, Function, and Bioinformatics* **70**, 41-53 (2008).
- [44] J. Higo, J. Ikebe, N. Kamiya, H. Nakamura and H. Shindo "An Enhanced Conformational Sampling of a 40-Residue Protein Consisting of Alpha and Beta Secondary Structures in Explicit Solvent" *Journal of Proteomics & Bioinformatics* **S2**, 255 (2008).
- [45] N. Kamiya, Y. Yonezawa, H. Nakamura and J. Higo "Assembly Simulation of Four Peptide Chains in Explicit Water by Multicanonical Molecular Dynamics" *Journal of Proteomics & Bioinformatics* **S2**, 265 (2008).
- [46] Ryuichiro Ishitani, Tohru Terada and Kentaro Shimizu "Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations" *Molecular Simulation* **34**, 327-336 (2008).
- [47] Tohru Terada and Kentaro Shimizu "A comparison of generalized Born methods in folding simulations" *Chemical Physics Letters* **460**, 295-299 (2008).
- [48] Tohru Terada, Daisuke Satoh, Tsutomu Mikawa, Yutaka Ito and Kentaro Shimizu "Understanding the roles of amino acid residues in tertiary structure formation of chignolin by using molecular dynamics simulation" *Proteins: Structure, Function, and Bioinformatics* **73**, 621-631 (2008).



- [49] Tadaomi Furuta, Kentaro Shimizu and Tohru Terada "Accurate prediction of native tertiary structure of protein using molecular dynamics simulation with the aid of the knowledge of secondary structures" *Chemical Physics Letters* **472**, 134-139 (2009).
- [50] Ryota Jono, Yuusuke Watanabe, Kentaro Shimizu and Tohru Terada "Multicanonical ab initio QM/MM molecular dynamics simulation of a peptide in an aqueous environment" *Journal of Computational Chemistry* **31**, 1168-1175 (2010).
- [51] Jinzen Ikebe, Daron M. Standley, Haruki Nakamura and Junichi Higo "Ab initio simulation of a 57-residue protein in explicit solvent reproduces the native conformation in the lowest free-energy cluster" *Protein Science* **20**, 187-196 (2011).
- [52] Jinzen Ikebe, Koji Umezawa, Narutoshi Kamiya, Takanori Sugihara, Yasushige Yonezawa, Yu Takano, Haruki Nakamura and Junichi Higo "Theory for trivial trajectory parallelization of multicanonical molecular dynamics and application to a polypeptide in water" *Journal of Computational Chemistry* **32**, 1286-1297 (2011).
- [53] Shinya Honda, Kazuhiko Yamasaki, Yoshito Sawada and Hisayuki Morii "10 Residue Folded Peptide Designed by Segment Statistics" *Structure* **12**, 1507-1518 (2004).
- [54] Susan L. Rowland, William F. Burkholder, Katherine A. Cunningham, Mark W. Maciejewski, Alan D. Grossman and Glenn F. King "Structure and Mechanism of Action of Sda, an Inhibitor of the Histidine Kinases that Regulate Initiation of Sporulation in *Bacillus subtilis*" *Molecular Cell* **13**, 689-701 (2004).
- [55] René Wintjens, Jean-Michel Wieruszkeski, Hervé Drobecq, Pierre Rousselot-Pailley, Luc Buée, Guy Lippens and Isabelle Landrieu "1H NMR Study on the Binding of Pin1 Trp-Trp Domain with Phosphothreonine Peptides" *The Journal of Biological Chemistry* **276**, 25150-25156 (2001).
- [56] Alex Bateman and Mark Bycroft "The structure of a LysM domain from *E. coli* membrane-bound lytic murein transglycosylase D (MltD)" *Journal of Molecular Biology* **299**, 1113-1119 (2000).
- [57] Wolfgang Kabsch and Christian Sander "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features" *Biopolymers* **22**, 2577-2637 (1983).
- [58] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin and K. S. Wilson "Overview of the CCP4 suite and current developments" *Acta Cryst.* **D67**, 235-242 (2011).
- [59] A. Immirzi, *Crystallographic Computing Techniques*, ed. F.R.Ahmed, Munksgaard, **399** (1966).

- [60] L. F. Ten Eyck "Crystallographic fast Fourier transforms" *Acta Cryst.* **A29**, 183-191 (1973).
- [61] R. J. Read and A. J. Schierbeek "A phased translation function" *J. Appl. Cryst.* **21**, 490-495 (1988).
- [62] International Tables for X-ray Crystallography Vol.IV Kynoch Press, (1974).
- [63] L. F. Ten Eyck "Efficient structure-factor calculation for large molecules by the fast Fourier transform" *Acta Cryst.* **A33**, 486-492 (1977).
- [64] R. C. Agarwal "A new least-squares refinement technique based on the fast Fourier transform algorithm" *Acta Cryst.* **A34**, 791-809 (1978).
- [65] "Refinement of protein structures", Proceedings of the Daresbury Study Weekend 15-16 November, 1980 (Compiled by P.S. Machin, J.W. Campbell and M. Elder).
- [66] Brünger A. T. "Free R value: a novel statistical quantity for assessing the accuracy of crystal structures." *Nature* **355**, 472-475 (1992).
- [67] International Tables for Crystallography vol. C Kluwer, (1995).
- [68] V. Luzzati "Resolution d'un structure cristalline lorsque les positions d'une partie des atoms sont connues: traitement statistique" *Acta Cryst.* **6**, 142-152 (1953).
- [69] G. A. Sim "The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures" *Acta Cryst.* **12**, 813-815 (1959).
- [70] G. A. Sim "A note on the heavy-atom method" *Acta Cryst.* **13**, 511-512 (1960).
- [71] R. Srinivasan "Weighting functions for use in the early stage of structure analysis when a part of the structure is known" *Acta Cryst.* **20**, 143-144 (1966).
- [72] W. A. Hendrickson and E. E. Lattman "Representation of phase probability distributions for simplified combination of independent phase information" *Acta Cryst.* **B26**, 136-143 (1970).
- [73] G. Bricogne "Methods and programs for direct-space exploitation of geometric redundancies" *Acta Cryst.* **A32**, 832-847 (1976).
- [74] H. Hauptman "On integrating the techniques of direct methods and isomorphous replacement. I. The theoretical basis" *Acta Cryst.* **A38**, 289-294 (1982).
- [75] Rogers, D. in Computing Methods in Crystallography (Rollett, J.S.,ed.) Pergamon Press, 126-127 (1985).
- [76] R. J. Read "Improved Fourier coefficients for maps using phases from partial structures with errors" *Acta Cryst.* **A42**, 140-149 (1986).
- [77] R. J. Read "Structure-factor probabilities for related structures" *Acta Cryst.* **A46**, 900-912 (1990).
- [78] David T Jones "Protein secondary structure prediction based on position-specific scoring

- matrices" *Journal of Molecular Biology* **292**, 195-202 (1999).
- [79] Buchan, D.W., Ward, S.M., Lobley, A.E., Nugent, T.C., Bryson, K. and Jones, D.T. "Protein annotation and modelling servers at University College London." *Nucl. Acids Res.* **38 Suppl**, W563-W568. (2010).
- [80] <http://www.rcsb.org/pdb>
- [81] <http://www.disprot.org/>
- [82] Alexey G. Murzina, Steven E. Brennera, Tim Hubbarda and Cyrus Chothia "SCOP: A structural classification of proteins database for the investigation of sequences and structures" *Journal of Molecular Biology* **247**, 536-540 (1995).
- [83] Romero P, Obradovic Z, Kissinger CR, Villafranca JE and Dunker AK "Intelligent Data Analysis for Protein Disorder Prediction" *Proc. IEEE Int. Conf. on Neural Networks* **1**, 90-95 (1997).
- [84] Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E and Dunker AK "The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder" *Proc. Genome Informatics* **9**, 193-200 (1998).
- [85] Romero P, Obradovic Z and Dunker AK "Folding minimal sequences: the lower bound for sequence complexity of globular proteins." *FEBS Letters.* **462**, 363-367 (1999).
- [86] Li X, Obradovic, Z, Brown CJ, Garner EC and Dunker AK "Comparing predictors of disordered protein" *Proc. Genome Informatics* **11**, 172-184 (2000).
- [87] Williams RM, Obradovic Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ and Dunker AK "The protein non-folding problem: amino acid determinants of intrinsic order and disorder." *Proc. 6th Pacific Symposium on Biocomputing*, 89-100 (2000).
- [88] Romero P, Obradovic Z, Li X, Garner EC, Brown CJ and Dunker AK "Sequence complexity of disordered protein." *Proteins: Structure, Function and Genetics* **42**, 38-48 (2001).
- [89] Tompa P. "Intrinsically unstructured proteins." *Trends Biochem Sci.* **27**, 527-533 (2002).
- [90] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM and Obradović Z "Intrinsic disorder and protein function." *Biochemistry.* **41**, 6573-6582 (2002).
- [91] Uversky VN "What does it mean to be natively unfolded?" *Eur J Biochem* **269**, 2-12 (2002).
- [92] Uversky VN "Natively unfolded proteins: a point where biology waits for physics." *Protein Sci.* **11**, 739-756 (2002).
- [93] Slobodan Vucetic, Celeste J. Brown, A. Keith Dunker and Zoran Obradovic "Flavors of protein disorder" *Proteins: Structure, Function, and Bioinformatics* **52**, 573-584 (2003).
- [94] Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, et al. "Predicting intrinsic

- disorder from amino acid sequence." *Proteins: Structure, Function, and Genetics* **53**, 566-572 (2003).
- [95] Clay Bracken, Lilia M Iakoucheva, Pedro R Romero and A Keith Dunker "Combining prediction, computation and experiment for the characterization of protein disorder" *Current Opinion in Structural Biology* **14**, 570-576 (2004).
- [96] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF and Jones DT "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life." *J Mol Biol.* **337**, 635-645 (2004).
- [97] Garbuzynskiy SO, Lobanov MY and Galzitskaya OV "To be folded or to be unfolded?" *Protein Sci.* **13**, 2871-2877 (2004).
- [98] Anthony L Fink "Natively unfolded proteins" *Current Opinion in Structural Biology* **15**, 35-41 (2005).
- [99] Lise S and Jones DT "Sequence patterns associated with disordered regions in proteins." *Proteins.* **58**, 144-150 (2005).
- [100] Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, et al. "Comparing and combining predictors of mostly disordered proteins." *Biochemistry* **44**, 1989-2000 (2005).
- [101] Tompa P, Szasz C and Buday L "Structural disorder throws new light on moonlighting." *Trends Biochem Sci.* **30**, 484-489 (2005).
- [102] Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S "Assessing protein disorder and induced folding." *Proteins* **62**, 24-45 (2006).
- [103] Tanja Mittag and Julie D Forman-Kay "Atomic-level characterization of disordered protein ensembles" *Current Opinion in Structural Biology* **17**, 3-14 (2007).
- [104] Fuxreiter M, Tompa P and Simon I. "Local structural disorder imparts plasticity on linear motifs." *Bioinformatics.* **23**, 950-956 (2007).
- [105] Schlessinger A, Liu J and Rost B "Natively Unstructured Loops Differ from Other Loops." *PLoS Comput. Biol.* **3**, e140 (2007).
- [106] Singh GP, Ganapathi M and Dash D "Role of intrinsic disorder in transient interactions of hub proteins." *Proteins* **66**, 761-765 (2007).
- [107] A Keith Dunker, Israel Silman, Vladimir N Uversky and Joel L Sussman "Function and structure of inherently disordered proteins" *Current Opinion in Structural Biology* **18**, 756-764 (2008).
- [108] Russell RB and Gibson TJ "A careful disorderliness in the proteome: sites for interaction and targets for future therapies." *FEBS Lett.* **582**, 1271-1275 (2008).
- [109] David Eliezer "Biophysical characterization of intrinsically disordered proteins" *Current Opinion in Structural Biology* **19**, 23-30 (2009).

- [110] Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK and Uversky VN. "Close encounters of the third kind: disordered domains and the interactions of proteins." *Bioessays*. **31**, 328-335 (2009).
- [111] Tanja Mittag, Lewis E. Kay and Julie D. Forman-Kay "Protein dynamics and conformational disorder in molecular recognition" *J. Mol. Recognition* **23**, 105-116 (2010).
- [112] Schaefer C, Schlessinger A and Rost B. "Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be." *Bioinformatics*. **26**, 625-631 (2010).
- [113] Lobanov MY, Furletova EI, Bogatyreva NS, Roytberg MA and Galzitskaya OV. "Library of disordered patterns in 3D protein structures." *PLoS Comput. Biol.* **6**, e1000958 (2010).
- [114] Avner Schlessinger, Christian Schaefer, Esmeralda Vicedo, Markus Schmidberger, Marco Punta and Burkhard Rost "Protein disorder — a breakthrough invention of evolution?" *Current Opinion in Structural Biology* **21**, 412-418 (2011).
- [115] Celeste J Brown, Audra K Johnson, A Keith Dunker and Gary W Daughdrill "Evolution and disorder" *Current Opinion in Structural Biology* **21**, 441-446 (2011).
- [116] Siltberg-Liberles J. "Evolution of structurally disordered proteins promotes neostructuralization." *Mol Biol Evol.* **28**, 59-62 (2011).