

配列 - 構造比較の分解能は、配列比較を越えられるか？

2001年4月21日

バイオインフォマティクス談話会  
京都大学理学部 宮田研究室

宮澤 三造

miyazawa@smlab.sci.gunma-u.ac.jp

## 背 景

真空中における蛋白質分子の conformational energy を通常 of 半経験的エネルギーポテンシャルを用いて評価すると、必ずしも、native structure は non-native structure より低いエネルギー状態として評価されない。(Novotny, J. et al., 1984)

- 溶媒効果の重要性を示唆
- (計算の精度は、溶媒効果の評価に左右。)

## 背 景

真空中における蛋白質分子の conformational energy を通常 of 半経験的エネルギーポテンシャルを用いて評価すると、必ずしも、native structure は non-native structure より低いエネルギー状態として評価されない。(Novotny, J. et al., 1984)

→ 溶媒効果の重要性を示唆

→ (計算の精度は、溶媒効果の評価に左右。)

疎水相互作用を考慮した非常に単純な経験的ポテンシャルでも、native structure を non-native structure から区別できる。(Eisenberg & McLachlan, 1986)

→ **Folding(立体構造予測) より Inverse folding (構造に最適な配列を探索) は容易か？**

⇒ 構造予測に対する一般的な要請も背景として Inverse folding の研究が隆盛。

## 背景

真空中における蛋白質分子の conformational energy を通常半経験的エネルギーポテンシャルを用いて評価すると、必ずしも、native structure は non-native structure より低いエネルギー状態として評価されない。(Novotny, J. et al., 1984)

→ 溶媒効果の重要性を示唆

→ (計算の精度は、溶媒効果の評価に左右。)

疎水相互作用を考慮した非常に単純な経験的ポテンシャルでも、native structure を non-native structure から区別できる。(Eisenberg & McLachlan, 1986)

→ Folding(立体構造予測) より Inverse folding (構造に最適な配列を探索) は容易か？

⇒ 構造予測に対する一般的な要請も背景として Inverse folding の研究が隆盛。

実際には Inverse folding にも、Folding を考える際の問題点と類似の問題が存在。

## 構造と配列の適合性とは？

定義: 構造と配列の適合性  $\equiv$  構造 - 配列の安定性

## 構造と配列の適合性とは？

定義: 構造と配列の適合性  $\equiv$  構造 - 配列の安定性

この定義が適切であれば、根本的な意味で、問題が容易になったわけではない。

## 構造と配列の適合性とは？

定義: 構造と配列の適合性  $\equiv$  構造 - 配列の安定性

この定義が適切であれば、根本的な意味で、問題が容易になったわけではない。

安定性の評価は、極めて困難。ある構造のエネルギーの評価は容易だが、分配関数の評価 / 多数の状態からなる変性状態の自由エネルギーの評価は極めて困難。

例: 一アミノ酸置換による安定性変化 ( $\Delta\Delta G$ ) の予測

## 構造と配列の適合性とは？

定義: 構造と配列の適合性  $\equiv$  構造 - 配列の安定性

この定義が適切であれば、根本的な意味で、問題が容易になったわけではない。

安定性の評価は、極めて困難。ある構造のエネルギーの評価は容易だが、分配関数の評価 / 多数の状態からなる変性状態の自由エネルギーの評価は極めて困難。

例: 一アミノ酸置換による安定性変化 ( $\Delta\Delta G$ ) の予測

違いは探索空間の大小; 配列 - 構造比較においては、構造データベースが探索空間。



## 配列 - 構造比較の問題点 I

配列  $i$  が構造  $s$  をとる確率:

$$-\log(\mathcal{P}(s|i)) = \beta E^{conf}(s|i) + \log(\sum_s \exp(-\beta E^{conf}(s|i)))$$

## 配列 - 構造比較の問題点 I

配列  $i$  が構造  $s$  をとる確率:

$$-\log(\mathcal{P}(s|i)) = \beta E^{conf}(s|i) + \log(\sum_s \exp(-\beta E^{conf}(s|i)))$$

1. 適合性評価関数としては第 1 項 (エネルギー項) のみを考慮すればよい場合; 第 2 項 (分配関数) は一定
  - 残基の欠失 / 挿入を許さない threading による配列  $i$  に適合する構造の検索
2. 第 2 項 (分配関数) が無視できない場合
  - 構造  $s$  に適合する配列の検索
  - 残基の欠失 / 挿入を許す配列・構造アライメントによる配列  $i$  に適合する構造の検索

配列 - 構造比較と蛋白質構造に最適なアミノ酸配列デザインは類似。

配列 / アラメント  $i$  が 構造  $s$  状態にある確率:

$$-\log(\mathcal{P}(s|i)) = \beta E^{conf}(s|i) + \log(\sum_s \exp(-\beta E^{conf}(s|i)))$$

配列 / アラメント  $i$  と 構造  $s$  の適合確率:

$$\mathcal{P}(i|s) = \frac{\mathcal{P}(s|i)\mathcal{P}(i)}{\sum_i \mathcal{P}(s|i)\mathcal{P}(i)}$$

蛋白質構造に最適なアミノ酸配列 / アラメント  $i^{opt}$ :

$$\begin{aligned} \mathcal{P}(i^{opt}|s) &= \max_i \mathcal{P}(i|s) \\ &= \max_i \mathcal{P}(s|i) \quad \text{if} \quad \mathcal{P}(i) = \text{constant} \end{aligned}$$

配列 - 構造比較と蛋白質構造に最適なアミノ酸配列デザインは類似。

配列 / アラメント  $i$  が構造  $s$  状態にある確率:

$$-\log(\mathcal{P}(s|i)) = \beta E^{conf}(s|i) + \log(\sum_s \exp(-\beta E^{conf}(s|i)))$$

配列 / アラメント  $i$  と構造  $s$  の適合確率:

$$\mathcal{P}(i|s) = \frac{\mathcal{P}(s|i)\mathcal{P}(i)}{\sum_i \mathcal{P}(s|i)\mathcal{P}(i)}$$

蛋白質構造に最適なアミノ酸配列 / アラメント  $i^{opt}$ :

$$\begin{aligned} \mathcal{P}(i^{opt}|s) &= \max_i \mathcal{P}(i|s) \\ &= \max_i \mathcal{P}(s|i) \quad \text{if} \quad \mathcal{P}(i) = \text{constant} \end{aligned}$$

分配関数の評価については、配列デザインの研究の分野で、いくつかの試みがなされているが、配列-構造比較に利用できる近似は限られよう。

→ 分配関数はアミノ酸配列依存性を無視しアミノ酸組成のみに依存すると近似し、評価関数を定義。(言い替えると、アミノ酸組成一定の場合は、第二項を無視する。)

配列 - 構造比較と蛋白質構造に最適なアミノ酸配列デザインは類似。

配列 / アラメント  $i$  が構造  $s$  状態にある確率:

$$-\log(\mathcal{P}(s|i)) = \beta E^{conf}(s|i) + \log(\sum_s \exp(-\beta E^{conf}(s|i)))$$

配列 / アラメント  $i$  と構造  $s$  の適合確率:

$$\mathcal{P}(i|s) = \frac{\mathcal{P}(s|i)\mathcal{P}(i)}{\sum_i \mathcal{P}(s|i)\mathcal{P}(i)}$$

蛋白質構造に最適なアミノ酸配列 / アラメント  $i^{opt}$ :

$$\begin{aligned} \mathcal{P}(i^{opt}|s) &= \max_i \mathcal{P}(i|s) \\ &= \max_i \mathcal{P}(s|i) \quad \text{if} \quad \mathcal{P}(i) = \text{constant} \end{aligned}$$

分配関数の評価については、配列デザインの研究の分野で、いくつかの試みがなされているが、配列-構造比較に利用できる近似は限られよう。

→ 分配関数はアミノ酸配列依存性を無視しアミノ酸組成のみに依存すると近似し、評価関数を定義。(言い替えると、アミノ酸組成一定の場合は、第二項を無視する。)

⇒ 分解能に限界。

## エネルギーポテンシャルによる配列 - 構造比較の問題点 II

例: myoglobin sequence と hemoglobin structure( $\alpha^2 \beta^2$  tetramer) の比較

$$E^{\text{intra}}(\text{myoglobin threaded into myoglobin}) \\ > E^{\text{intra}}(\text{myoglobin threaded into } \alpha \text{ hemoglobin}) + E^{\text{between subunit}}(\text{myoglobin } \alpha\beta^2)$$

myoglobin,  $\alpha$ ,  $\beta^2$  tetramer の安定性を論ずるには、多量体形成によるエントロピーロスを考慮にいれないといけない。

## エネルギーポテンシャルによる配列 - 構造比較の問題点 II

例: myoglobin sequence と hemoglobin structure( $\alpha^2 \beta^2$  tetramer) の比較

$$E^{\text{intra}}(\text{myoglobin threaded into myoglobin}) \\ > E^{\text{intra}}(\text{myoglobin threaded into } \alpha \text{ hemoglobin}) + E^{\text{between subunit}}(\text{myoglobin } \alpha\beta^2)$$

myoglobin,  $\alpha$ ,  $\beta^2$  tetramer の安定性を論ずるには、多量体形成によるエントロピーロスを考慮にいれないといけない。

多量体形成によるエントロピーロスを考慮せずエネルギーだけでは、配列 - 構造間の適合度を評価できない。

→ 何らかの近似的取り扱いが必要。

## エネルギーポテンシャルによる配列 - 構造比較の問題点 II

例: myoglobin sequence と hemoglobin structure( $\alpha^2 \beta^2$  tetramer) の比較

$$E^{\text{intra}}(\text{myoglobin threaded into myoglobin}) \\ > E^{\text{intra}}(\text{myoglobin threaded into } \alpha \text{ hemoglobin}) + E^{\text{between subunit}}(\text{myoglobin } \alpha\beta^2)$$

myoglobin,  $\alpha$ ,  $\beta^2$  tetramer の安定性を論ずるには、多量体形成によるエントロピーロスを考慮にいれないといけない。

多量体形成によるエントロピーロスを考慮せずエネルギーだけでは、配列 - 構造間の適合度を評価できない。

→ 何らかの近似的取り扱いが必要。

⇒ 分解能に限界。



## アミノ酸置換と配列 - 構造の適合性

## アミノ酸置換と配列 - 構造の適合性

1. 適合確率から Mutation matrix を予測。

- Dayhoff's PAM matrix と相関

## アミノ酸置換と配列 - 構造の適合性

1. 適合確率から Mutation matrix を予測。
  - Dayhoff's PAM matrix と相関
2. 適合確率から各アミノ酸座位におけるアミノ酸変異度 (mutability) を予測;  $-\sum_i p(i, q) \log p(i, q)$  (Shakhnovich et al., 1996)
  - 類似配列で観測される変異度と高い相関。

## アミノ酸置換と配列 - 構造の適合性

1. 適合確率から Mutation matrix を予測。
  - Dayhoff's PAM matrix と相関
2. 適合確率から各アミノ酸座位におけるアミノ酸変異度 (mutability) を予測;  $-\sum_i p(i, q) \log p(i, q)$  (Shakhnovich et al., 1996)
  - 類似配列で観測される変異度と高い相関。
3. アミノ酸置換による安定性変化 ( $\Delta\Delta G$ ) の低分解能予測

## 統計ポテンシャルは、エネルギーポテンシャルか？

蛋白質の立体構造において観測される、 $(\phi, \psi)$  分布, 残基間の統計分布から評価した統計ポテンシャルは、knowledge-base potential としてだけでなくエネルギーポテンシャル と見なせるか？

## 統計ポテンシャルは、エネルギーポテンシャルか？

蛋白質の立体構造において観測される、 $(\phi, \psi)$  分布, 残基間の統計分布から評価した統計ポテンシャルは、knowledge-base potential としてだけでなくエネルギーポテンシャル と見なせるか？

十分条件

1. 蛋白質の native state は 各種の相互作用に関し”**frustrate**”していない。
  - ”frustrate” していないことは折り畳み可能性の必要条件であるらしい。
2. 蛋白質配列は、(配列空間に関し) 熱平衡状態である。

蛋白質配列は、最適配列と見なせるか？

蛋白質配列は、(配列空間に関し) 熱平衡状態と見なせるか？

蛋白質配列は、最適配列と見なせるか？

蛋白質配列は、(配列空間に関し) 熱平衡状態と見なせるか？

1. 残基の  $(\phi, \psi)$  分布は、 $(\phi, \psi)$  ポテンシャルに関するボルツマン分布に類似。

→ 直接的には、 $(\phi, \psi)$  ポテンシャルに関し、"frustrate" していないことを意味するが、熱平衡状態を示唆する。



蛋白質配列は、最適配列と見なせるか？

蛋白質配列は、(配列空間に関し) 熱平衡状態と見なせるか？

1. 残基の  $(\phi, \psi)$  分布は、 $(\phi, \psi)$  ポテンシャルに関するボルツマン分布に類似。

→ 直接的には、 $(\phi, \psi)$  ポテンシャルに関し、"frustrate" していないことを意味するが、熱平衡状態を示唆する。

2. 多数の蛋白質における隣接残基頻度分布は、相対誤差 10% をもって隣接エネルギーに関し熱平衡状態にあるとみなせる。

蛋白質配列は、最適配列と見なせるか？

蛋白質配列は、(配列空間に関し) 熱平衡状態と見なせるか？

1. 残基の  $(\phi, \psi)$  分布は、 $(\phi, \psi)$  ポテンシャルに関するボルツマン分布に類似。

→ 直接的には、 $(\phi, \psi)$  ポテンシャルに関し、"frustrate" していないことを意味するが、熱平衡状態を示唆する。

2. 多数の蛋白質における隣接残基頻度分布は、相対誤差 10% をもって隣接エネルギーに関し熱平衡状態にあるとみなせる。

3. ほとんどの蛋白質の  $(\phi, \psi)$  エネルギーおよび隣接残基間相互作用エネルギーはともに、

$$\frac{(E_{native} - \langle E \rangle_{\text{in equilibrium}})}{\sigma_E} < 2 \quad (1)$$

4. 蛋白質における  $(\phi, \psi)$  エネルギーと隣接残基間相互作用エネルギー  $((E_{native} - \langle E \rangle)/\sigma)$  は、互いに無相関。