タンパク質モデリングの幾何学

児玉 大樹 (こだま ひろき)

東京大学大学院 数理科学研究科
数理科学連携基盤センター
生物医学と数学の融合拠点 (iBMath)

数理科学講演会
2013 年 6 月 28 日

最近の悩み

生物屋さんの前で話をするときには
「数学の話を 10 分でしてくれ」と言われる。

数学屋さんの前で話をするときには
「生物の話を 60 分でしてくれ」と言われる。

iBMath について 1

生物医学と数学の融合拠点 (iBMath)
Institute for Biology and Mathematics of Dynamical Cell Processes
本拠地: アネックス棟

iBMath について 2

生物医学と数学の融合拠点 (iBMath)
Institute for Biology and Mathematics of Dynamical Cell Processes
本拠地: アネックス棟

メンバー:
井原茂男 (拠点長)
大田佳宏
児玉大樹
中田庸一
鮑園園
藤博之
松家敬介

学内の主な共同研究者:
栗原裕基 (医学系研究科)
坪井俊 (数理)
時弘哲治 (数理)
和田洋一郎 (アイソトープ総合センター)

五十音順

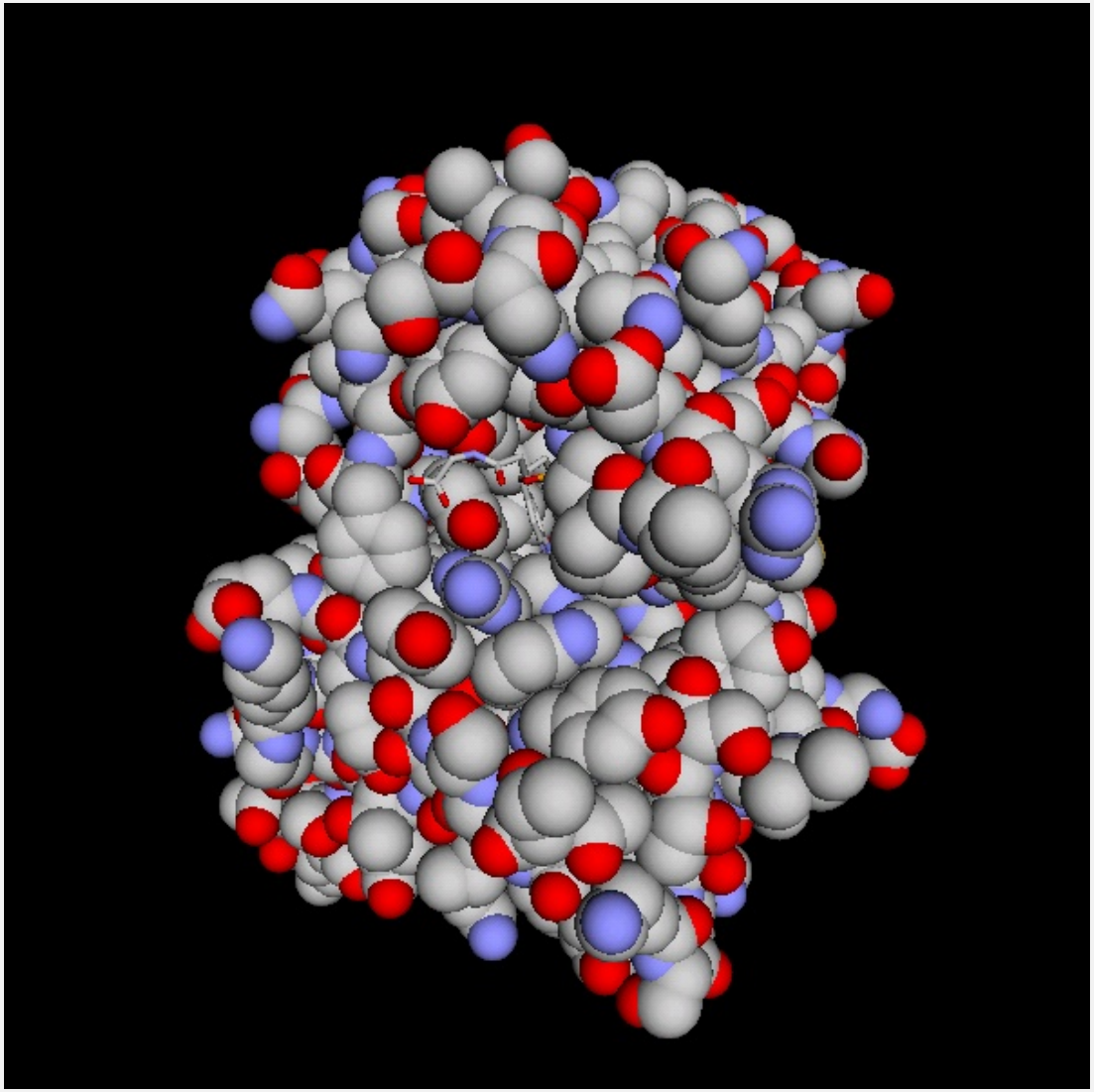
iBMath について 3

iBMath には
「転写の機構解明のための動態システム生物医学数理解析拠点」
という名前もある。

転写: DNA の情報を元に RNA を合成する反応。

以下の三点からアプローチ。
(I) 転写過程の高分解能時系列実験と数理モデリングとシミュレーション
(II) 細胞の集団運動の実験とシミュレーション
(III) 転写因子の蛋白質構造数理解析

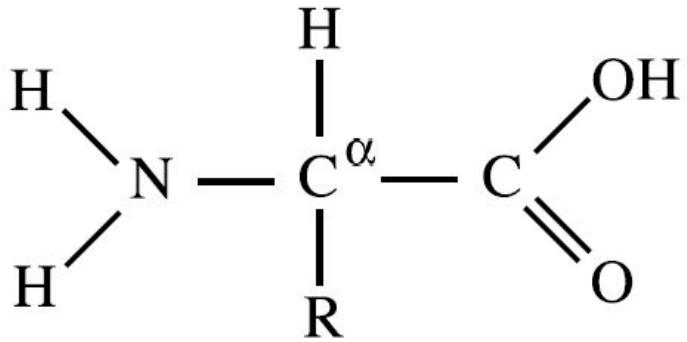
タンパク質とは 1



画像: PDBjViewer

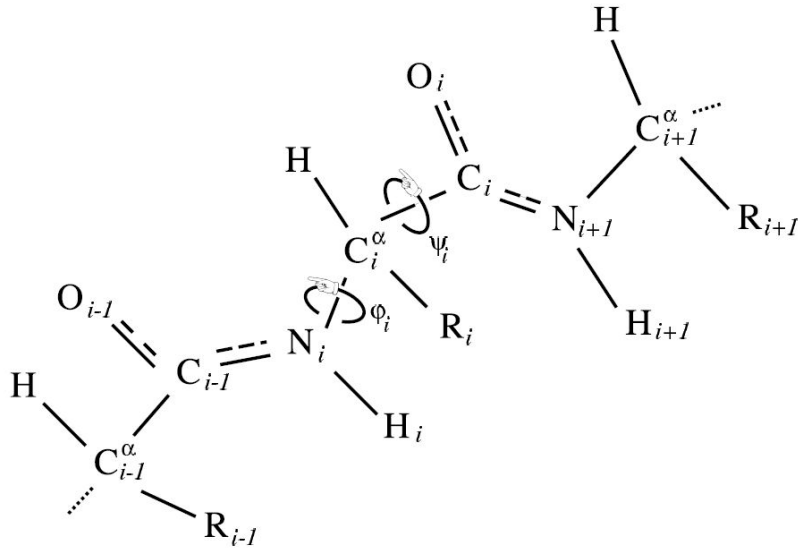
タンパク質とは 2

タンパク質を構成するアミノ酸 (H_2N-$C^{\alpha}HR$-COOH) は 20 種類ある。



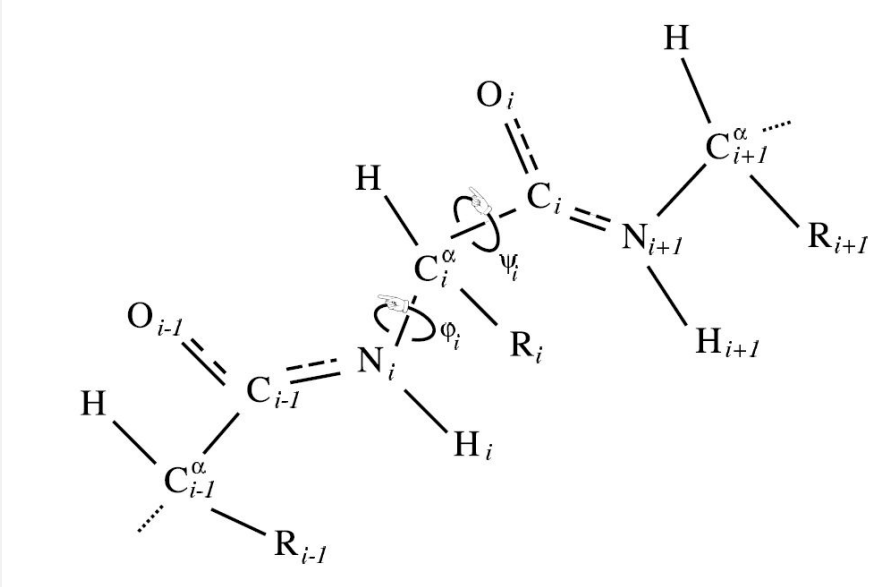
a) Typical amino acid

アミノ酸のカルボキシル基 (-COOH) と、別のアミノ酸のアミノ基 (H_2N-) が脱水縮合して (-CO-NH-) を形成することにより、アミノ酸のポリマーであるポリペプチドが出来る。この (-CO-NH-) の部分をペプチドユニットと呼ぶ。



タンパク質とは 3

アミノ酸のカルボキシル基 (-COOH) と、別のアミノ酸のアミノ基 (H_2N-) が脱水縮合して (-CO-NH-) を形成することにより、アミノ酸のポリマーであるポリペプチドが出来る。この (-CO-NH-) の部分をペプチドユニットと呼ぶ。



一列の鎖であるポリペプチドはこのままでは安定ではないが、ポリペプチドが折りたたまって、ペプチドユニットの酸素原子 O と、別のペプチドユニットの水素原子 H が水素結合をつくり、安定した立体構造を得る。これがタンパク質である。

Gromov の問題

ポリペプチドを構成するアミノ酸の列を (DNA にある設計図通りにではなく) ランダムにとると、ほとんどの場合ポリペプチドは折りたたまらない。

問.[Gromov] アミノ酸の数が $N = 50$ 個からなるポリペプチドは全部で 20^{50} 通り考えられるが、そのうち何個が折りたたまって安定した立体構造を得るか？

Protein Folding 問題

一方で、実際にタンパク質を構成するポリペプチドは、アミノ酸配列が同じであれば同じ形に折りたたまることが経験的に知られている。

問.(Protein Folding 問題) 与えられたアミノ酸の列に対して、それがどのように折りたたまるのか決定せよ。(cf. [Smale])

このように、ポリペプチドの折りたたみに関しては興味深い (難しい) 問題がいろいろある。

タンパク質データバンク

実は、既知のタンパク質については Protein Folding 問題を考える必要はあまりない。多くのタンパク質の立体構造は具体的に各原子の (x,y,z) 座標のデータとして測定され、タンパク質データバンク (PDB) として公開されている。

http://www.wwpdb.org/

タンパク質データバンク (PDB) のデータ例

```
425 ATOM   17  HE1 MET A   1      -14.035 -11.506 -24.526  1.00  1.00           H
426 ATOM   18  HE2 MET A   1      -14.827  -9.956 -24.796  1.00  1.00           H
427 ATOM   19  HE3 MET A   1      -15.441 -11.047 -23.556  1.00  1.00           H
428 ATOM   20  N   ALA A   2      -15.711  -4.605 -20.374  1.00  1.00           N
429 ATOM   21  CA  ALA A   2      -16.979  -4.205 -19.687  1.00  1.00           C
430 ATOM   22  C   ALA A   2      -17.091  -2.676 -19.662  1.00  1.00           C
431 ATOM   23  O   ALA A   2      -16.258  -1.993 -19.092  1.00  1.00           O
432 ATOM   24  CB  ALA A   2      -16.975  -4.742 -18.252  1.00  1.00           C
433 ATOM   25  H   ALA A   2      -15.015  -3.934 -20.534  1.00  1.00           H
434 ATOM   26  HA  ALA A   2      -17.822  -4.618 -20.222  1.00  1.00           H
435 ATOM   27  HB1 ALA A   2      -16.190  -4.259 -17.689  1.00  1.00           H
436 ATOM   28  HB2 ALA A   2      -16.802  -5.808 -18.267  1.00  1.00           H
437 ATOM   29  HB3 ALA A   2      -17.928  -4.538 -17.789  1.00  1.00           H
438 ATOM   30  N   GLU A   3      -18.119  -2.138 -20.277  1.00  1.00           N
439 ATOM   31  CA  GLU A   3      -18.303  -0.654 -20.295  1.00  1.00           C
440 ATOM   32  C   GLU A   3      -19.788  -0.325 -20.086  1.00  1.00           C
441 ATOM   33  O   GLU A   3      -20.511  -0.025 -21.022  1.00  1.00           O
442 ATOM   34  CB  GLU A   3      -17.824  -0.091 -21.641  1.00  1.00           C
443 ATOM   35  CG  GLU A   3      -16.292  -0.031 -21.661  1.00  1.00           C
444 ATOM   36  CD  GLU A   3      -15.805   1.133 -20.789  1.00  1.00           C
445 ATOM   37  OE1 GLU A   3      -15.760   2.245 -21.290  1.00  1.00           O
446 ATOM   38  OE2 GLU A   3      -15.484   0.892 -19.636  1.00  1.00           O
447 ATOM   39  H   GLU A   3      -18.774  -2.714 -20.726  1.00  1.00           H
448 ATOM   40  HA  GLU A   3      -17.727  -0.212 -19.496  1.00  1.00           H
449 ATOM   41  HB2 GLU A   3      -18.170  -0.729 -22.441  1.00  1.00           H
450 ATOM   42  HB3 GLU A   3      -18.222   0.904 -21.777  1.00  1.00           H
451 ATOM   43  HG2 GLU A   3      -15.891  -0.959 -21.279  1.00  1.00           H
452 ATOM   44  HG3 GLU A   3      -15.952   0.115 -22.675  1.00  1.00           H
453 ATOM   45  N   GLN A   4      -20.243  -0.389 -18.858  1.00  1.00           N
```

pdb1jt8.ent

タンパク質モデリング

例えば、新しい薬を開発するときに、従来は薬の候補物質一つごとに実験動物の体内や試験管の中で実験を行なっていた。
これには大量の費用と時間が掛かるし、動物実験には倫理的問題も発生するので、コンピュータの中のシミュレーションで代替したい。

タンパク質データバンク (PDB) のような各原子の座標は立体構造を完全に与えるデータだが、これは複雑すぎてシミュレーションには向かない。
(スパコンで計算するより実験したほうが費用も時間も掛からない)

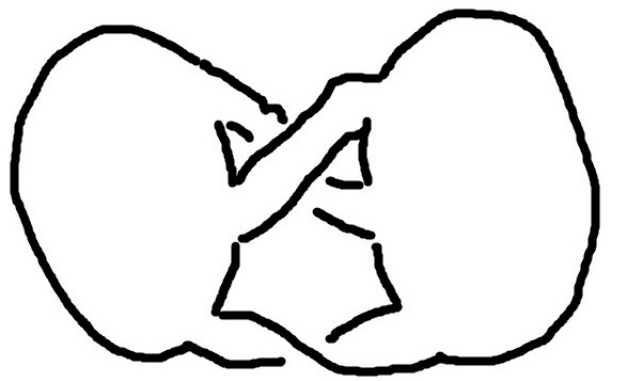
コンピュータシミュレーションのためには、タンパク質を
「コンピュータに扱える程度に単純な形で」
「タンパク質の持つ特徴を失わないように」
モデリングすることが必要である。

ファットグラフ

先行研究として、Penner らによって提唱されたタンパク質のファットグラフモデルについて説明する。

定義: 有限グラフ $\Gamma = (V, E)$ に次の二つの情報を付加したものをファットグラフと呼ぶ。
(i) 各頂点 $v \in V$ に対し、v を始点とする辺たちの cyclic order が与えられる。
(ii) 各辺 $e \in E$ に対し、ラベル 'n'(non-flipped) またはラベル 'f'(flipped) が与えられる。

ファットグラフ F の頂点を円板 (disc) で、辺を帯 (band) で置き換えることにより、ファットグラフを図示することができる。



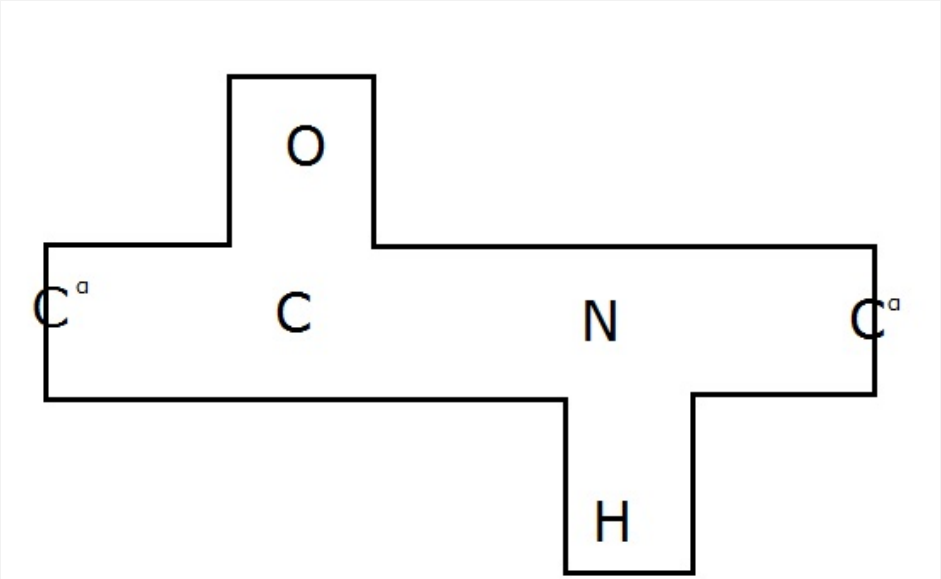
ファットグラフモデル [Penner 他] 1

(タンパク質データバンクなどで) 立体構造が完全にわかっているタンパク質 P に対し、次のルールでファットグラフ $F(P)$ を構成する。
$F(P)$ を P のファットグラフモデルと呼ぶ。

Step1. まずグラフ $\Gamma = \Gamma(P)$ を作る。
頂点の個数はペプチドユニットの個数に等しい。つまり、タンパク質を構成するアミノ酸の数を N とすると $|V| = N - 1$ である。
隣り合うペプチドユニットを辺で結ぶ。さらに、水素結合しているペプチドユニットも辺で結ぶ。このようにして得られたグラフが $\Gamma(P)$ である。

ファットグラフモデル [Penner 他] 2

Step2. $\Gamma(P)$ の頂点と辺に構造を付加する。
各頂点を図のような (位相的) 円板に置き換えることで頂点上の cyclic order を与える。



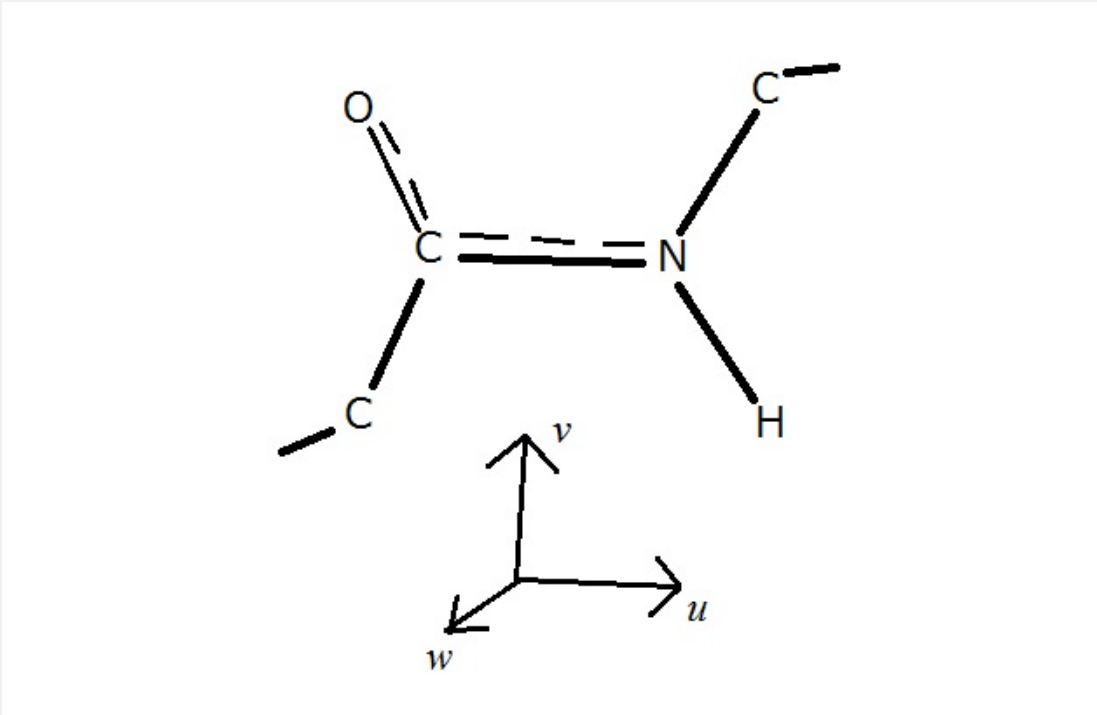
各辺には、その両端点に対応するペプチドユニットの向きに合わせて、non-flipped か flipped かを与える。

準備: ペプチドユニットの枠 (frame) 1

前提知識:

ペプチドユニットを構成する六つの原子は (ほぼ) 同一平面上にある。

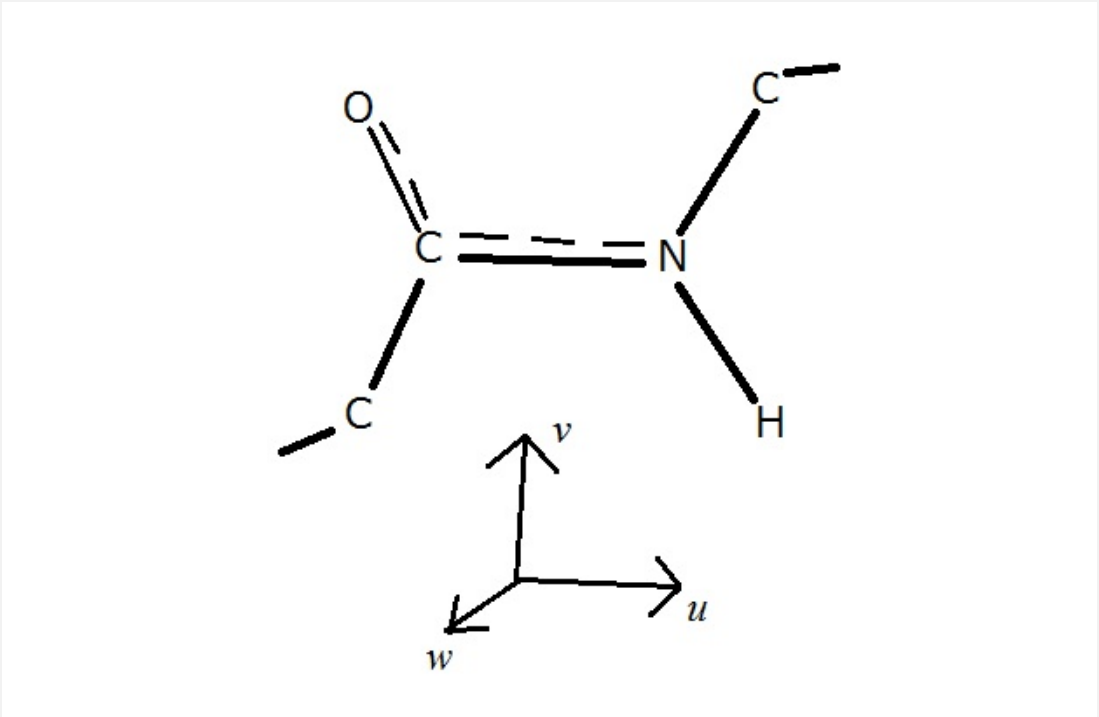
ペプチドユニットに対し、互いに直交する三本の単位ベクトル $\vec{u}, \vec{v}, \vec{w}$ を図のように定める。



$\vec{u}$ は C_i から N_{i+1} の方向の単位ベクトル。
$\vec{v}$ はペプチドユニットの平面に含まれ、$\vec{u}$ に垂直で C_i^{α} から O_i の方向に近い単位ベクトル。
$\vec{w} = \vec{u} \times \vec{v}$.

Figure : ペプチドユニットの枠

準備: ペプチドユニットの枠 (frame) 2



$\vec{u}$ は C_i から N_{i+1} の方向の単位ベクトル。
$\vec{v}$ はペプチドユニットの平面に含まれ、$\vec{u}$ に垂直で C_i^{α} から O_i の方向に近い単位ベクトル。
$\vec{w} = \vec{u} \times \vec{v}$.

Figure : ペプチドユニットの枠

$U = (\vec{u}, \vec{v}, \vec{w})$ をペプチドユニットの枠 (frame) と呼ぶ。
枠は 3×3 行列とみなせ、$SO(3)$ の元になる。

回転行列

i 番目のペプチドユニットと
j 番目のペプチドユニットが辺で結ばれているとする。

i 番目のペプチドユニットから j 番目のペプチドユニットへの
回転を表す行列を $\hat{R}^i{}_j$ と書くことにすると、

$$\hat{R}^i{}_j\vec{u}_i = \vec{u}_j,\ \hat{R}^i{}_j\vec{v}_i = \vec{v}_j,\ \hat{R}^i{}_j\vec{w}_i = \vec{w}_j.$$

まとめて書くと $\hat{R}^i{}_j U_i = U_j$.
これを解いて $\hat{R}^i{}_j = U_j U_i^{-1}$.

これが固定座標系から見たペプチドユニットの回転行列である。

座標変換

固定座標系から見たペプチドユニットの回転行列

$$\hat{R}^i{}_j = U_j U_i^{-1}$$

を i 番目のペプチドユニットから見た回転行列に変換する。

$$R^i{}_j = U_i^{-1} \hat{R}^i{}_j U_i = U_i^{-1} U_j U_i^{-1} U_i = U_i^{-1} U_j.$$

これが二つのペプチドユニットの
相対的な位置関係を表す回転行列である。

non-flipped か flipped か

$\mathbf{SO(3)}$ の元として、non-flipped を表す元 $I = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix}$ と

flipped を表す元 $J = \begin{pmatrix} 1 & & \\ & -1 & \\ & & -1 \end{pmatrix}$ を考える。

d を $\mathbf{SO(3)}$ 上の自然な距離、
つまり $d(A, B) = \arccos(\frac{\mathrm{Tr}\, AB^{-1} - 1}{2})$ とする。

$R = R^i{}_j$ に対し、$d(R, I) < d(R, J)$ のときに辺にラベル 'n' を、
そうでないときにラベル 'f' を与える。

注: $d(R, I) < d(R, J) \iff r_{22} + r_{33} > 0.$

「結果」

結果.[Penner 他](2010) 二つのタンパク質 P と P' に対して、このようにして構成されたファットグラフが一致していれば、元のタンパク質 P と P' は「ほぼ同じ」立体構造を持つ。
(注 1. 元論文では仮定はもうちょっと弱く、ファットグラフ $F(P)$ から計算できる数 (頂点数・種数・境界の成分数など) を 10 個定義して、それらが全て一致すれば、というものである。)
(注 2. 元論文の結論は、CATH 3.2.0(1997) というデータベースにある 114,215 種類のタンパク質のうち 19 組の例外を除けば、データベースの分類の深さ 4 までは一致する、というものである。)

つまり、ファットグラフモデルはかなり情報量を落としているにもかかわらず、元のタンパク質の立体構造の情報を (ある意味で) 失っていないので、
タンパク質モデリングの良い例となっている。

紙テープ模型

ファットグラフモデルのもう一つの特徴: ファットグラフは紙テープなどを使って模型として実際に作ってみせることが容易である。

今日は紙テープ模型を用いて、タンパク質の代表的な二次構造である α ヘリックスと β シートを紹介する。

αヘリックス

ここにαヘリックスの絵

β シート

ここに β シートの絵

ファットグラフモデルの限界

ファットグラフモデルはタンパク質の大まかな立体構造を把握するには極めて強力だが、反面、実際に製薬の場面で使おうと思うと、情報量が落ちすぎだということがわかってくる。

製薬の世界では、今現在ある薬の候補のタンパク質のほんの一部を取り替えて、似たような立体構造のタンパク質を作って性能を比較することが多いからである。

そもそも、ファットグラフモデルは、回転行列 $R^i{}_j$ という $\mathbf{SO(3)}$ の元をわざわざ $\{\text{nonflipped}, \text{flipped}\}$ という $\mathbb{Z}/2\mathbb{Z}$ の元で近似して作ったものだった。各辺に $\mathbf{SO(3)}$ の情報をそのまま与えたらどうなるだろう？

タンパク質のモデュライ空間

Γ を連結な有限グラフとする。
(あるタンパク質 P に対する $\Gamma(P)$ が念頭にある。)

$\Gamma = (V, E)$, $E \subset V \times V$, $(u, v) \in E \Longleftrightarrow (v, u) \in E$ と書くことにする。
E から $\mathbf{SO(3)}$ への写像 $R\colon (u, v) \mapsto R^u{}_v$ で、
以下の性質をみたすもののなす集合を $M(\Gamma)$ と書く。
(1) $(u, v) \in E$ ならば $R^v{}_u = (R^u{}_v)^{-1}$.
(2) $(v_0, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_0) \in E$ ならば $R^{v_0}{}_{v_1} R^{v_1}{}_{v_2} \cdots R^{v_{n-1}}{}_{v_0} = I$.

特に、$\Gamma = \Gamma(P)$ のときに $M(\Gamma(P))$ をタンパク質 P のモデュライ空間と呼ぶ。$M(\Gamma(P))$ のある一点 (タンパク質の実際の回転行列 $R^i{}_j$ を反映されて得られる点) がタンパク質のモデルと見なせる。

モデュライ空間についてのコメント

実は、$|V| = N - 1$ ならば $M(\Gamma) \cong \mathrm{SO}(3) \times \cdots \times \mathrm{SO}(3)$ ($N - 2$-times) なので、モデュライ空間自体は位相幾何学的には面白い空間ではないが、数学的に面白いかどうかはあまり重要ではない。

結果

[Penner Andersen 他](2013)
水素結合に現れる $R^i{}_j \in \mathbf{SO(3)}$ 達の分布をまとめた。
注: 隣り合うペプチドユニットの結合角の分布については
[Ramachandran 他](1963) で報告されているので、50 年ぶりの進歩。

[大田 児玉 他](投稿中)
投稿中なので具体的なところは話せない！

今後の展望

転写に関わる高分子の構造のモデル化をしたい。
RNA: これはすでに研究が進んでいる [Rydes] [Bonn]
DNA: こっちが狙い目

個人的な意見:
転写に関わる高分子は大きくて複雑なものがおおいので、用いるモデルは $\mathbf{SO(3)}$ を用いた精密なものでなく、ファットグラフや単なるグラフのほうが良いのでは？

ご清聴ありがとうございました。