7ZB-01

正規化ハミング距離を用いた三次元点集合マッチングの高速化とインフルエンザウイルス解析への応用

佐々木 耀一† 渋谷 哲朗‡ 大森 亮介§ 伊藤 公人§ 有村 博紀†

†北海道大学大学院情報科学研究科 ‡東京大学医科学研究所ヒトゲノム解析センター

§北海道大学人獣共通感染症リサーチセンター

1 はじめに

1.1 背景. 近年，タンパク質や高分子の三次元構造の高速な検索は重要な問題になっている．例えば，あるタンパク質の構造が与えられているが，その性質がわかっていないときに，既に構造と性質が知られているタンパク質との構造的に類似しているかを調べることにより，性質を予想することが行われている [1].

我々は，タンパク質の構造を三次元空間に分布する点集合ととらえて，二つの点集合 S と T の間の三次元点集合マッチング問題を考察し，高速なアルゴリズムを提案した [5][7]．このアルゴリズムは，点集合の構造の類似距離には平行移動と回転に関する最小二乗和閉包距離（RMSD）を採用し，その下限推定を用いた枝刈りにより，高速な照合を実現する．

1.2 本研究の目的. 本稿では，点の座標以外の情報を利用して，RMSD の三次元点集合マッチングを高速化することを考える．最近，分子生物学分野では，二つのタンパク質の間で，それらの RMSD 値と DNA 配列の文字列距離に密接な関係があることが指摘されている [6]．そこで，我々は，この RMSD と文字列距離の関係を用いて，三次元点集合マッチングの高速化を考える．初めに，実際のインフルエンザウイルスのタンパク質データベースを用いて，上記の r と d の相関関係を実験的に調べる．次に，観察した相関関係に基づいて経験的枝刈り関数 f を定める．最後に，我々の照合アルゴリズムに枝刈り関数を組み込み，提案手法による高速化の有用性を調べる．

1.3 関連研究. 本論文の関連研究として，Shibuya [2] は，データベースとパターンを共に点列の入力とし，RMSD 指標によるマッチングを考察し，データベースサイズに対して線形時間で実行するアルゴリズムを提案している．その他に，Akutsu [3] らは，ハッシュを用いた手法を提案しており，$O(N)$ 期待計算時間を実現している．また，Shibuya [1] は，幾何接尾辞木と呼ばれる，三次元構造に対する索引データ構造を示している．これは，文字列に対する索引構造である接尾辞木 [4] を基に，タンパク質のような複雑な三次元構造に対して適応させたデータ構造である．

Faster Approximate 3-Dimensional Point Set Matching Using Relationship between RMSD-Score and Normalized Hamming Distance and Application to Analysis of Influenza Viruses; †Yoichi SASAKI, ‡Tetsuo SHIBUYA, §Ryosuke OMORI, §Kikihito ITO, †Hiroki ARIMURA; †Graduate School of Information Science and Technology, Hokkaido University; ‡Laboratory of Sequence Analysis, Human Genome Center, University of Tokyo; §Hokkaido University Research Center for Zoonosis Control

2 準備

整数 $a \le b$ に対して，$[a..b]$ で閉区間 $\{a, a+1, \ldots, b\}$ を表す．三次元空間 $\mathbb{R}^3$ における要素数 m の点集合 $P = P[1..m] = P[1] \cdots P[m]$ を考える．ここに，各 $P[i] = (x_i, y_i, z_i) \in \mathbb{R}^3$ $(i \in [1..m])$ は $\mathbb{R}^3$ の点であり，点の順序は任意に決めておく．以下のように，点列間の RMSD スコア [1] を，点集合間の RMSD に拡張する [5, 7].

点列として表されたサイズ n の点集合 $P = P[1..n]$ と $Q = Q[1..n]$ 間の**最小平方平均二乗和距離** (minimum root mean square distance, **最小** *RMSD*) を，すべての順列 $\pi : [1..n] \to [1..n]$ と，回転行列 $R \in \mathbb{R}^{3\times3}$，移動ベクトル $v \in \mathbb{R}^3$ に関する最小の RMSD 値 [1]

$$MinRMS\,D(P, Q) \tag{1}$$

$$= \min_{\pi} \min_{R,v} RMSD_{R,v}(\pi(P), Q) \tag{2}$$

$$= \min_{\pi} \min_{R,v} \left\{ \sum_{i=1}^{n} (R(P) + v - Q)^2 \right\}^{1/2} \tag{3}$$

と定義する．ここに，π による点列 P の並べ替えを $\pi(P[1..m]) := (P[\pi(1)], \ldots, P[\pi(n)])$ とした．

本論文で考察する最小 RMSD スコアに関する三次元点集合マッチング問題 [7, 5] を次のように定義する．

定義 2.1 最小 *RMSD* スコアに関する三次元点集合マッチング問題は，データ点集合 $T = \{t_1, \ldots, t_n\}$ と，パターン点集合 $P = \{p_1, \ldots, p_k\}$，非負実数 $r > 0$ を受け取り，T 上の最小 RMSD スコア r でのパターン P の出現位置 $Q \subseteq T$ を全て見つける問題である．

3 提案手法

本節では，RMSD と NHD の間の関係 [6] を利用した枝刈り手法を定式化し，これを用いて前節で説明した点集合マッチング問題の高速化を考える．提案手法では，次の文字列距離を用いる．

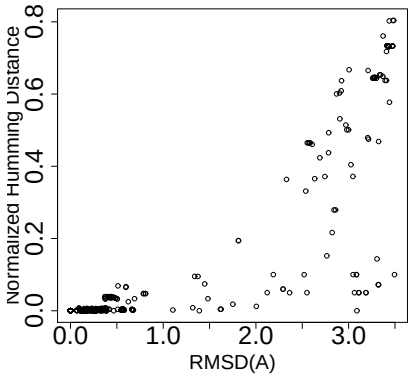

図 1: 実験：RMSD と NHD の散布図

定義 3.1 点集合 P と Q の間の順列 π の下での**正規化ハミング距離** *(NHD)* を，対応する点のラベルの不一致数を長さで正規化したもの

$$NHD_\pi(P,Q) = NHD(\pi(P),Q) \tag{4}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\,\pi(P)[i] \neq Q[i]\,] \in [0,1] \tag{5}$$

と定める．ここに，$\mathbb{I}[\cdot]$ は特性関数である．

RMSD **による** *NHD* **の上限関数**（または**上限関数**）とは，任意の実数関数 $f : [0,\infty) \to [0,1]$ である．多くの点集合の対 $\tau = (P,Q)$ に対して，関係

$$RMSD(\pi(P),Q) \leq r \Rightarrow NHD(\pi(P),Q) \leq d := f(r) \tag{6}$$

が成立するならば，与えられた RMSD 値の上限 r に対して NHD 値の上限 $d = f(r)$ を用いて，再帰的なマッチングアルゴリズムの繰り返しにおいて，第 i 番目の点対に対して，制約違反 $NHD(P[\pi(1)]\cdots P[\pi(i)]]), Q[1..i]) > d$ が成立したとき，これ以上の探索を行わないという枝刈りが行える．一般には，対上のある同時確率分布 D に従う点集合の対 $(P,Q) \sim D$ に対して高い確率で式 (6) が成立するならば，一定の照合ミスを許した上で，照合の高速化を期待できる．

4 実験

実験 1. 実験では，PDB database 内にある 342 種類のインフルエンザウイルスのヘマグルチニンと呼ばれるタンパク質データに対して，それらの RMSD が 3.5（Å）以下となる組み合わせに対して，RMSD と HND の値を計測した．これより，r の上限と d の上限の間にわれわれが主張するような上記の関係があるかどうかを調べた．

実験 2. 実験 1 の結果のグラフから，経験的な枝刈り関数 f を求めた (表 1)．また，この関数 f を用いて，

r (A)	0.0~0.1	0.1~0.3	0.3~0.5	0.5~2.0	2.0~3.0	3.0~
d	0.003	0.006	0.04	0.2	0.8	0.1

表 1: 今回用いた経験的な上限関数 $d = f(r)$

RMSD r	0.1		0.15		0.2		0.25	
手法	照合	時間	照合	時間	照合	時間	照合	時間
提案手法	0	25	4	166	4	368	4	1532
全解出力	0	67	4	301	4	1463	4	9341

表 2: 実験 2．各 RMSD 値毎に，見つけた照合数（照合）と要した計算時間（時間，単位 sec) を示す．

PDB database* の一組のインフルエンザウイルスのヘマグルチニン (4bgw.pdb(点の数=486) VS 4bgx.pdb(点の数=485)) に対して，f による枝刈りの正しさを確認した．

5 結論

本稿では，３次元空間での点集合マッチングを考察し，RMSD と文字列距離である NHD の関係から考えた関数を用いて枝刈りによる高速化を行うアルゴリズムについて議論した．実験結果からは，RMSD と NHD に我々の指摘した関係があることを観察した．

今後の課題として，実験 2 で与えた枝刈り関数の損失率を定式化し，枝刈り関数のクラスを設計して，要求する損失率に対して，適切な枝刈り関数を与える．

参考文献

[1] Shibuya Tetsuo, ”Geometric Suffix Tree : Indexing Protein 3-D Structures,” Jounal of the ACM, Vol.57, No.3, Article 15, 2010.

[2] Shibuya Tetsuo, ”Searching Protein 3-D Structures in Linear Time,” RECOMB 2009, LNCS 5541, 115, 2009.

[3] Akutsu Tatsuya, Onizuka Kentaro, Ishikawa Masato, ”New Hashing Techniques and Their Application to a Protein Strucure Database System,” Proc. Hawaii Int. Conf. System Sciences 5, 197-206, 1995.

[4] Esko Ukkonen, ”On-line construction of suffix-trees,” Algorithmica, Vol. 13, No. 3, 249-260, 1995.

[5] Yoichi Sasaki, Tetsuo Shibuya, Kimihito Ito, and Hiroki Arimura, ”Efficient Approximate 3-Dimensional Point Set Matching Using Root-Mean-Square Deviation Score”, In Proc. SISAP2015,2015

[6] Aneerban Bhattacharya, et al. ”Assessing model accuracy using the homology modeling automatically software”, Proteins: Structure, Function, and Bioinformatics, Volume 70, Issue 1, Pages 105118, 2008

[7] 佐々木 耀一，渋谷 哲朗，伊藤 公人，有村 博紀，”三次元空間における効率良い近似点集合マッチングと分子パターン照合への応用”，第 42 回バイオ情報学 (BIO) 研究会, 2015

*http://www.rcsb.org/pdb/home/home.do