大規模なタンパク質データ解析のための高速な局所配列特徴抽出法の開発

蝦名　鉄平

独立行政法人　理化学研究所
脳科学総合研究センター　大脳皮質回路可塑性研究チーム

「統合データ解析トライアル」 研究終了報告会

2014年3月2日

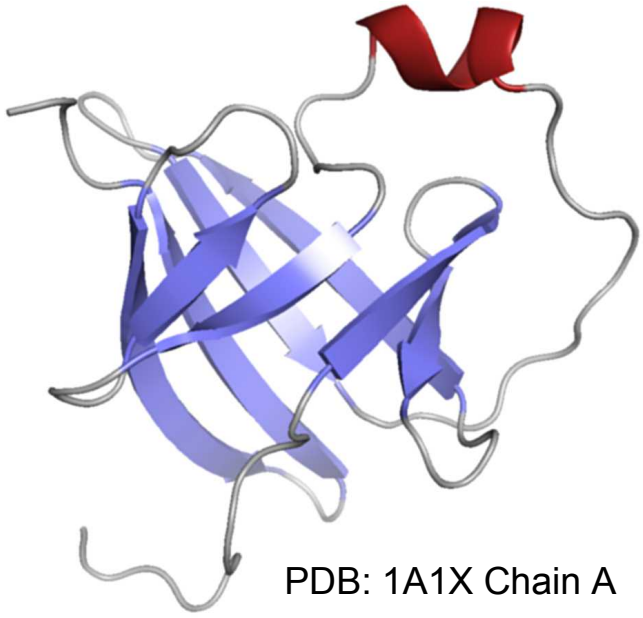
研究開発の目的

「高速な」タンパク質の配列特徴抽出法を開発する

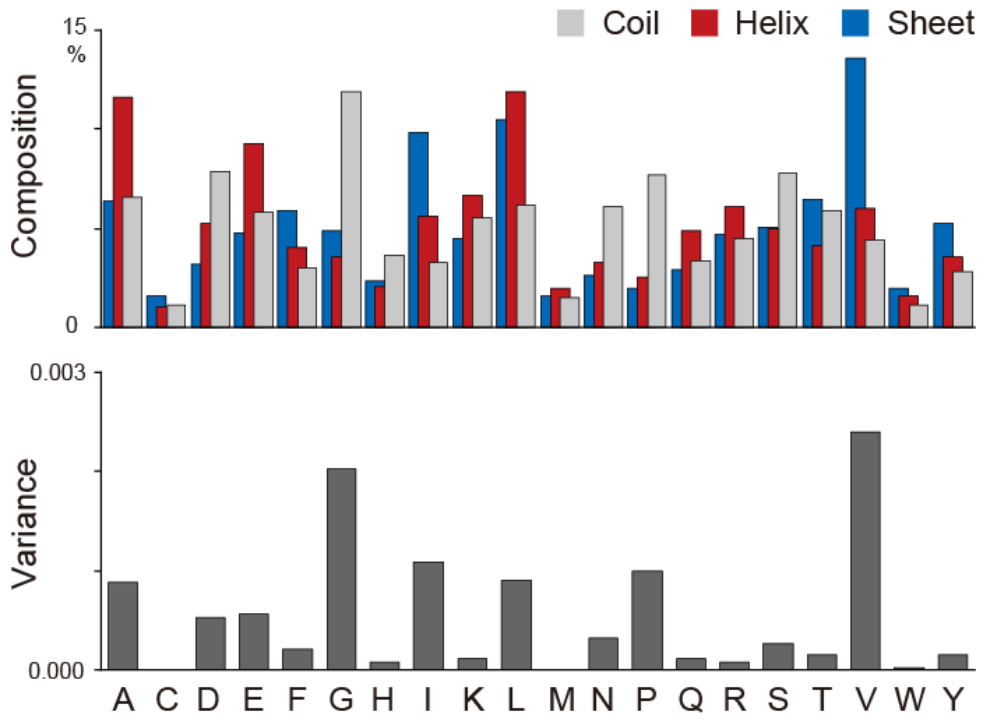
- 「配列特徴抽出法」とは？

  ある構造や機能領域に対応するアミノ酸配列の規則性や特徴を調べる方法

二次構造を形成する領域のアミノ酸組成



PDB: 1A1X Chain A

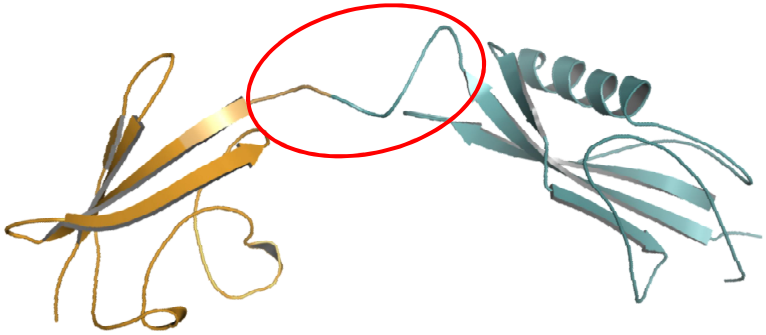


研究開発の目的

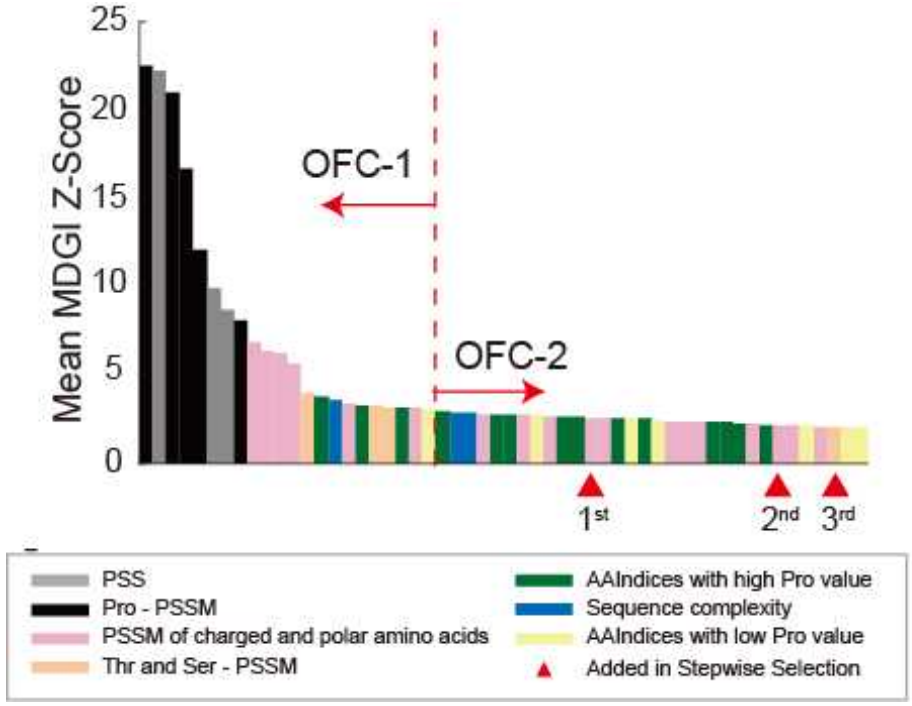
機械学習法を用いた特徴抽出

多くの「特徴」候補の中から、重要な特徴を抽出できる

ドメインリンカー領域の配列特徴



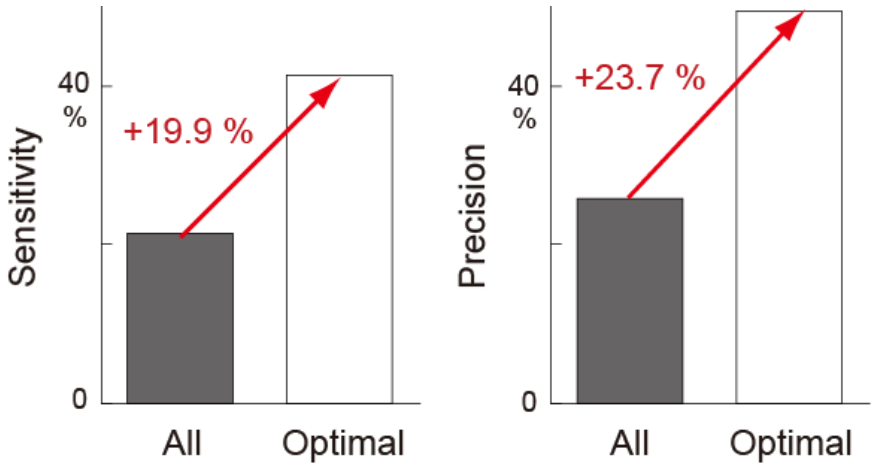
Ebina et al. *Bioinformatics* (2011) Vol. 27, p.487-494

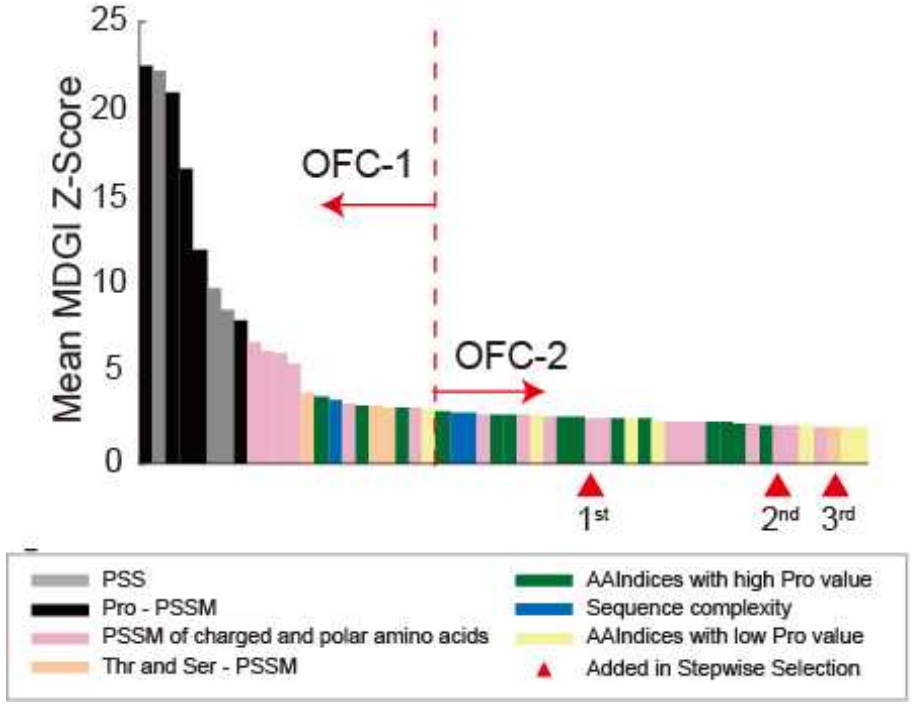


研究開発の目的

機械学習法を用いた特徴抽出

多くの「特徴」候補の中から、重要な特徴を抽出できる



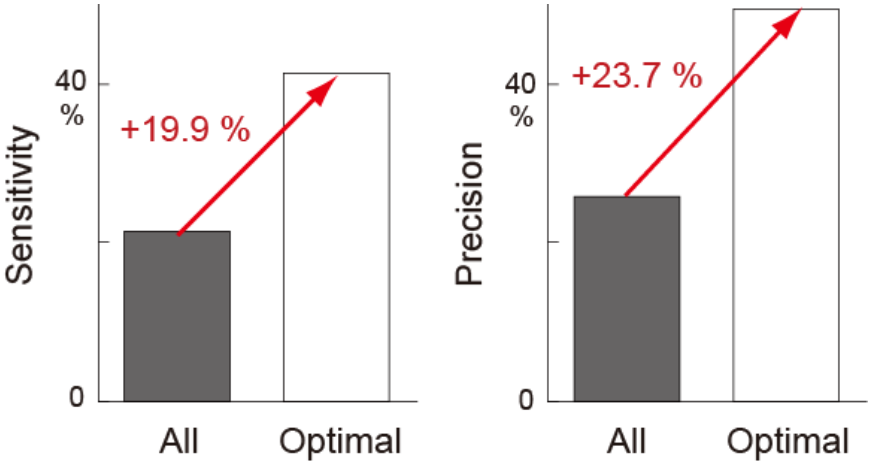


Ebina et al. *Bioinformatics* (2011) Vol. 27, p.487-494

研究開発の目的

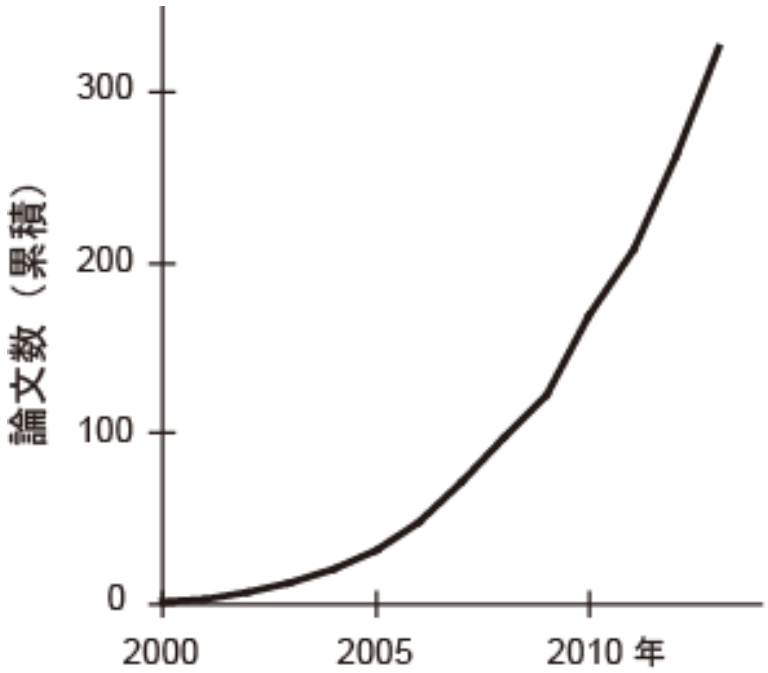
機械学習法を用いた特徴抽出

多くの「特徴」候補の中から、重要な特徴を抽出できる



Ebina et al. *Bioinformatics* (2011) Vol. 27, p.487-494

PubMed検索の結果: “feature selection” + “protein”

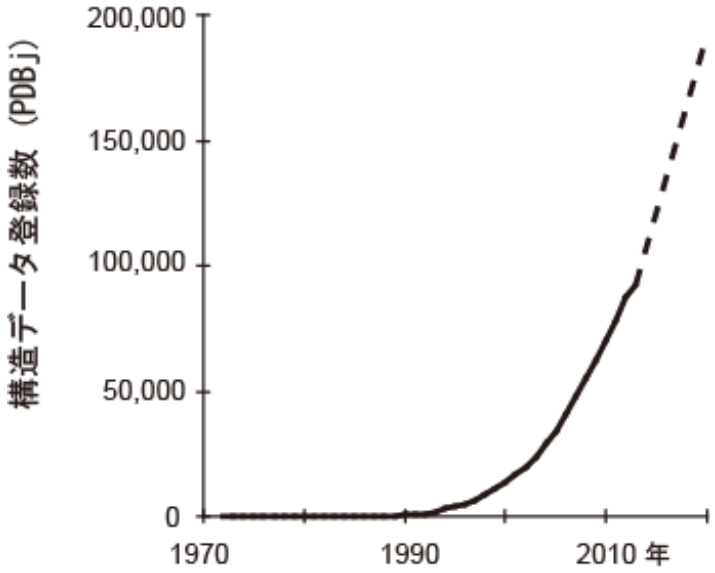


研究開発の目的

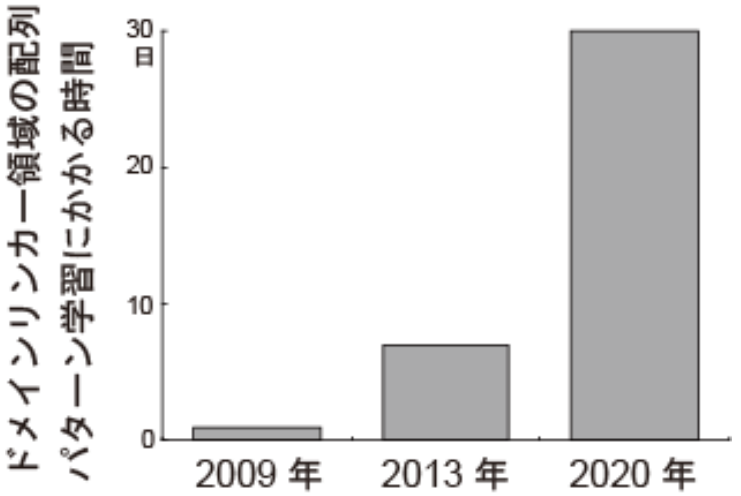
機械学習法を用いた特徴抽出

- 従来の方法の問題点

  計算量がデータ数に対して非線形に増加する



個々のユーザが、PCを使って配列の特徴を抽出する事が難しくなる

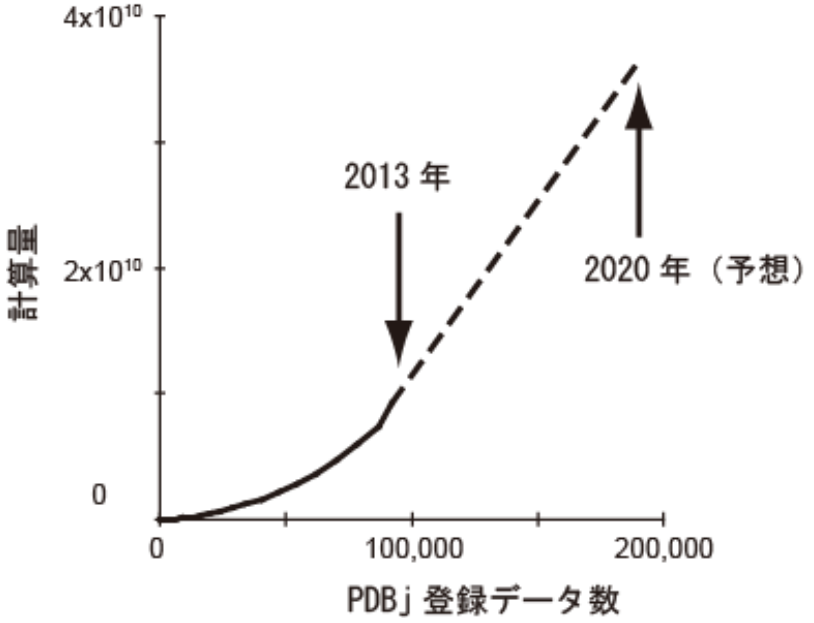


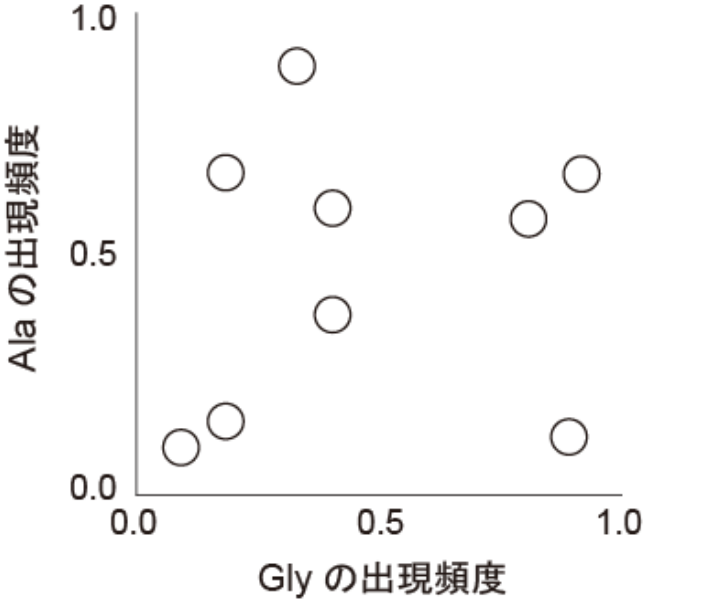
研究開発の目的

「高速な」タンパク質の配列特徴抽出法を開発する

- 提案する解決策

  「クラスタリング」を利用して、特徴抽出に不要なデータを排除する
  & 代表データのみを利用する





○ 各アミノ酸残基に対応するベクトルデータ
ベクトル要素：アミノ酸の出現頻度など

研究開発の目的

「高速な」タンパク質の配列特徴抽出法を開発する

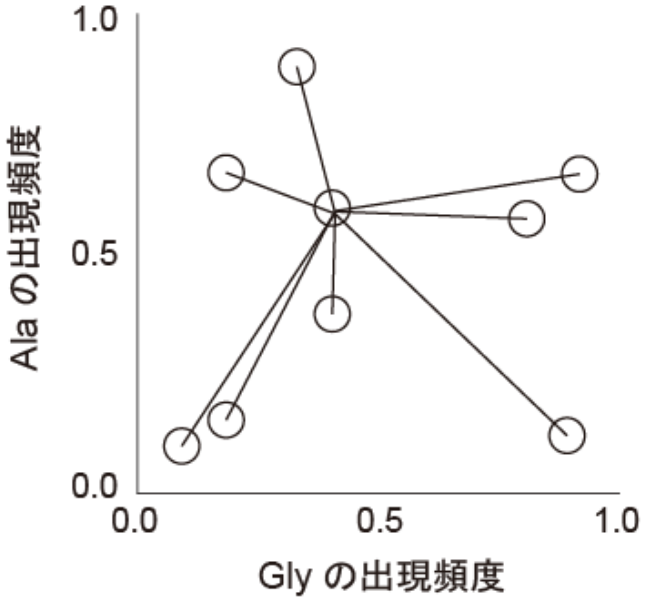
- 提案する解決策

  「クラスタリング」を利用して、特徴抽出に不要なデータを排除する
  & 代表データのみを利用する

一般的な階層的クラスタリング法でデータ間の類似度計算にかかる時間（計算量）は$N^2 / 2$　以上

↓

類似度計算の「速い」アルゴリズムを利用する



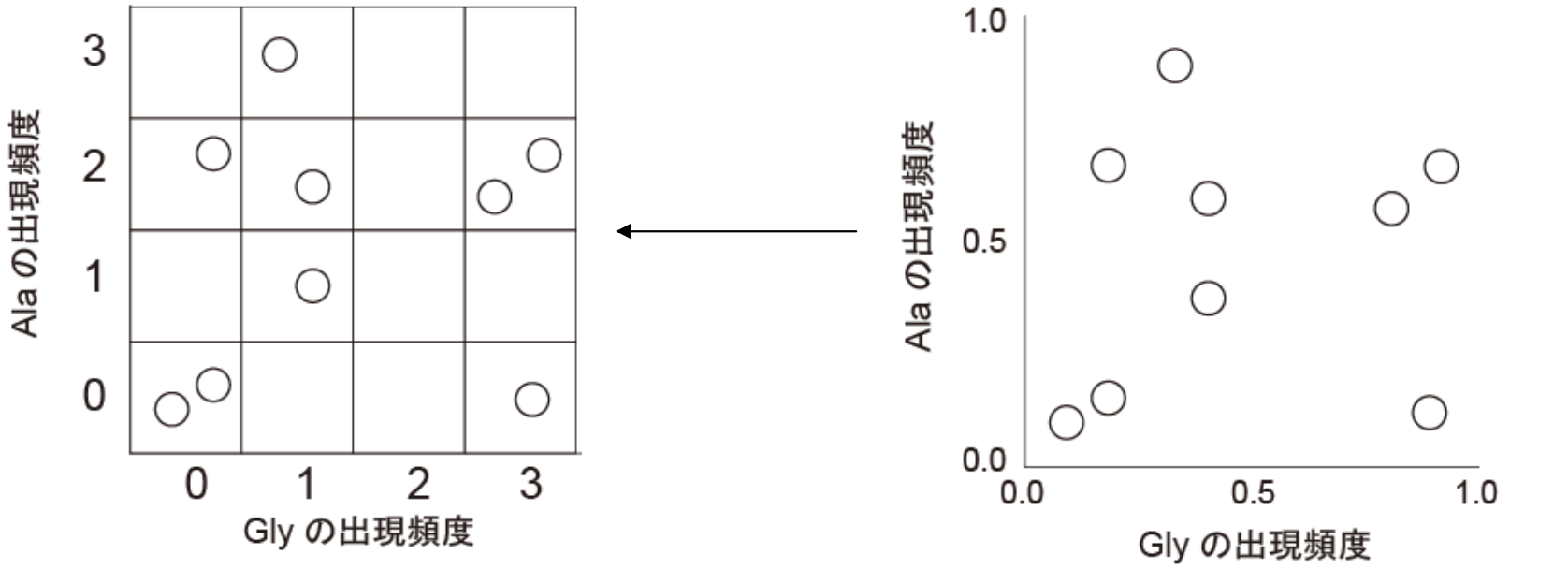
○ 各アミノ酸残基に対応するベクトルデータ
ベクトル要素：アミノ酸の出現頻度など

研究方法：クラスタリング法の検討

BOOLによるクラスタリング

- BOOL[1]：Binary cOding Oriented cLustering

1. 各ベクトル要素を k 段階に離散化 （例は $k = 4$）



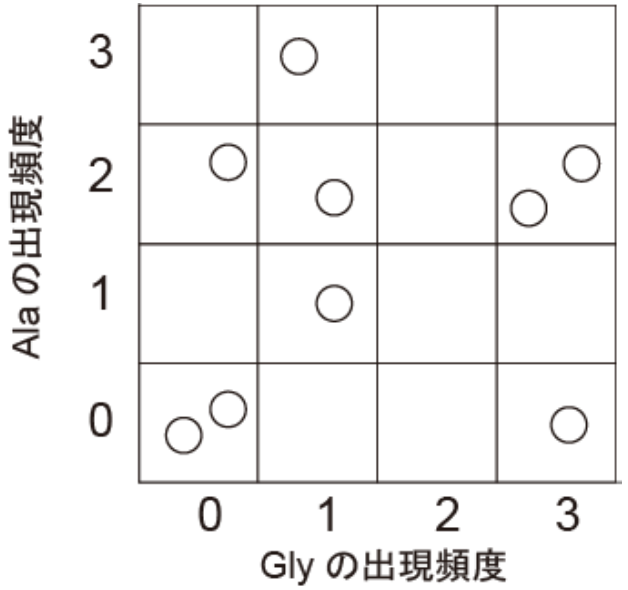
[1] 杉山麿人・山本章博（2011）2進符号化を活用した高速かつ柔軟なクラスタリング人工知能学会全国大会（第25回）抄録集、1P2-Ib-3in

研究方法：クラスタリング法の検討

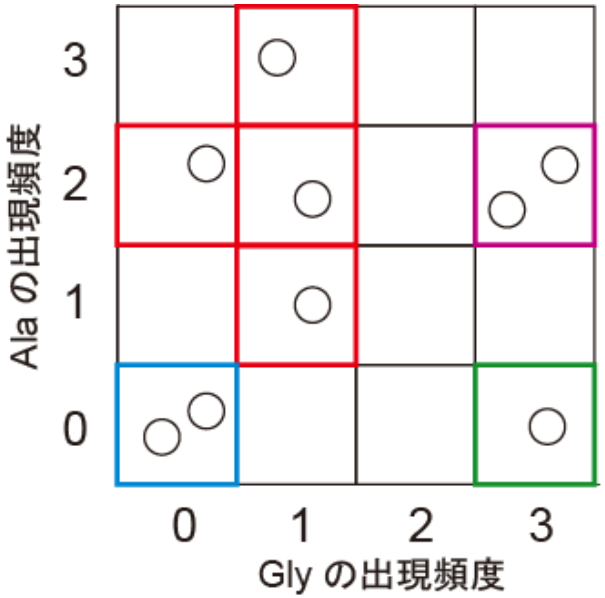
BOOLによるクラスタリング

- BOOL：Binary cOding Oriented cLustering

1. 各ベクトル要素を *k* 段階に離散化 （例は *k* = 4）



2. 枠間の距離が*L*以下のグループを同一のクラスに分類する（例：*L* = 1）

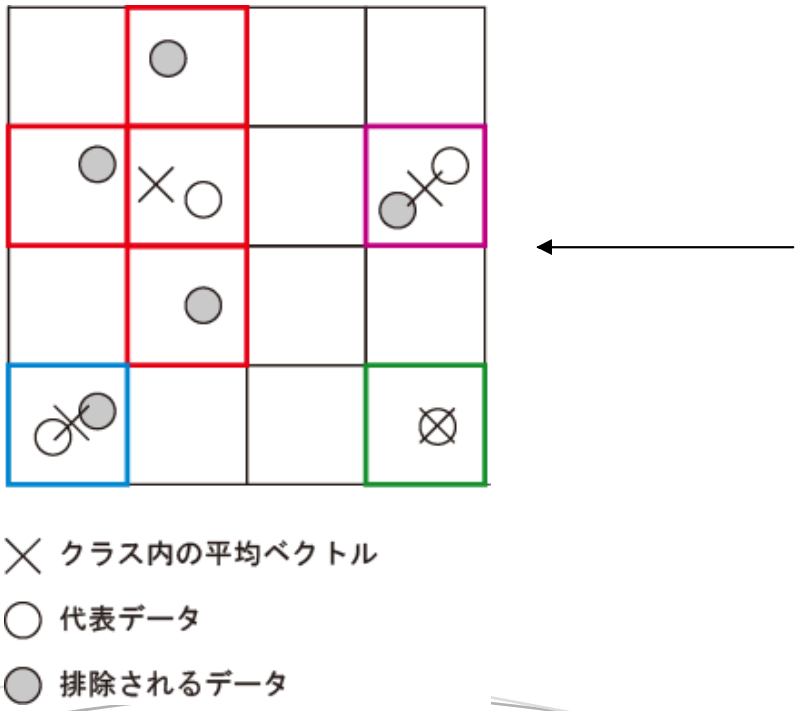


研究方法：クラスタリング法の検討

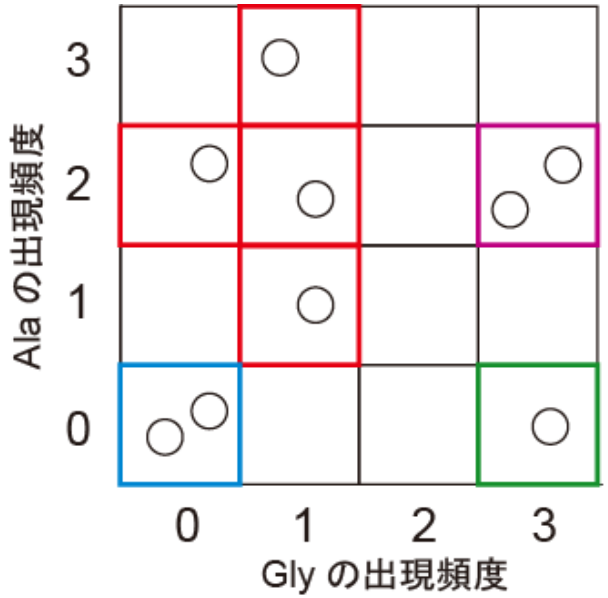
BOOLによるクラスタリング

- BOOL：Binary cOding Oriented cLustering

3. 各クラス内で、ベクトル重心にもっとも近いベクトルを代表として選出する



2. 枠間の距離が*L*以下のグループを同一のクラスに分類する（例：*L* = 1）



研究方法：クラスタリング法の検討

BOOLによるクラスタリング

- モデルデータ（タンパク質の二次構造 – 配列データ）

PDBj

結晶構造
X線回折の解像度が2.0Å以上
DNAなど、他の高分子を含まないデータ

2013年7月現在
14,831構造
代表配列数：2,907
753,349残基（代表配列中）

ベクトルデータ化

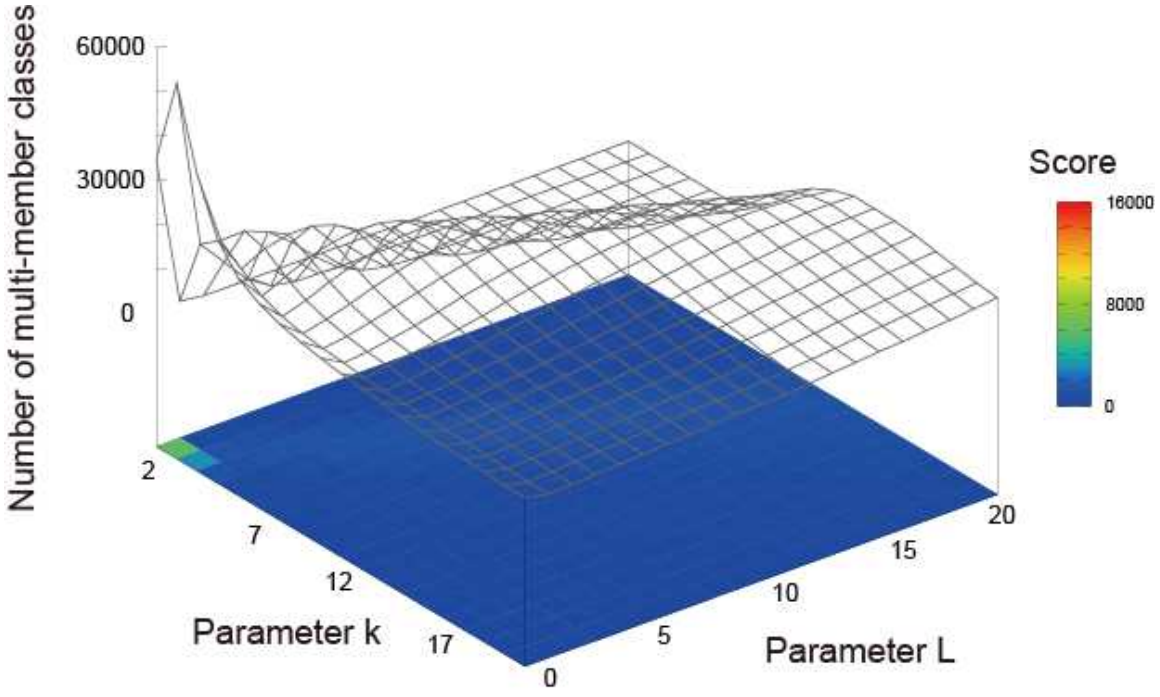
- 位置特異的スコア行列（PSSM)

ベクトルデータのクラスタリング
代表ベクトルを用いてループ構造のアミノ酸配列特徴抽出

結果：クラスタリング法の検討

計算アルゴリズムの開発： BOOLによるクラスタリング

- BOOLのパラメータを最適化する



$$Score = N_{multi} \times (1 - {}^{N_{single}}\!/_{N_{all}})$$

N_{multi}: number of groups consisting of > 1 vectors

N_{single}: number of groups consisting of 1 vectors

1パラメータあたりの実行時間：
＜ 1 min for 750,000 vectors

(Single-linkage clustering: ~12 h)

結果：クラスタリング法の検討

BOOLによるクラスタリング

- 実行例: k=2, L=0 (x_i = 0.0 ~ 0.5:a, ~ 1.0:b）

83,154 classes (34,533 multi and 48,621 single data classes).

ベクトルデータ

```
1:0.817797 2:0.000000 3:0.556497
1:0.635280 2:0.051251 3:0.889154
1:0.407159 2:0.387025 3:0.588367
1:0.339585 2:0.345508 3:1.000000
1:0.446991 2:0.189112 3:0.733524
1:0.299632 2:0.121324 3:0.636949
1:0.398687 2:0.006567 3:0.769231
1:0.385633 2:0.493384 3:0.517958
1:0.554455 2:0.000000 3:0.909241
1:0.423326 2:0.066955 3:0.749460
1:0.300679 2:0.018429 3:0.923375
1:0.348877 2:0.000000 3:0.538860
1:0.425723 2:0.380857 3:0.666002
1:0.910569 2:0.000000 3:0.882114
1:0.421225 2:0.041575 3:0.598468
1:0.849412 2:0.040000 3:0.672941
```

Alaの保存度　Cys　Asp

ベクトル要素

BOOL →

BOOLの結果

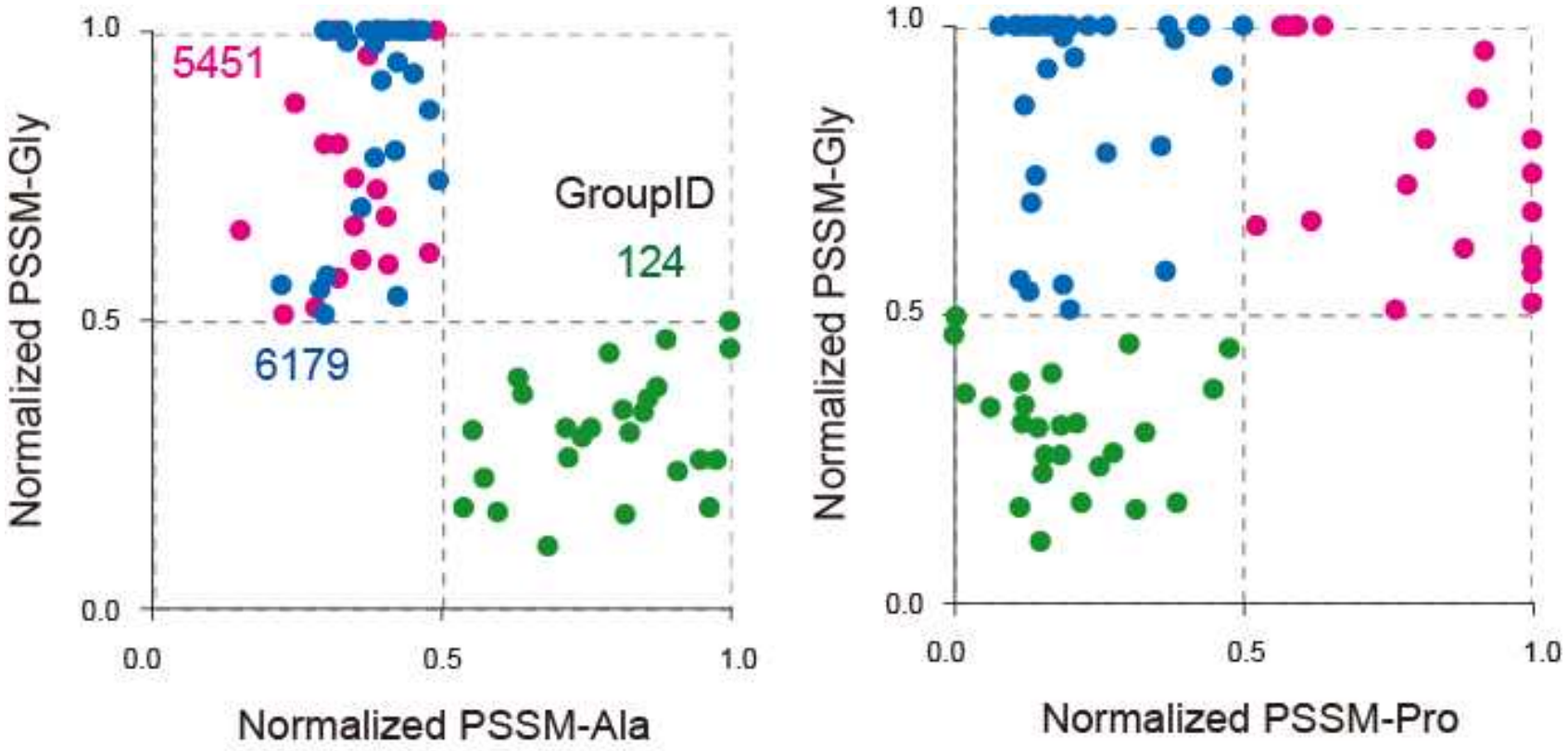
```
babbaababaaaabbbbaba :124
babbaababaaaabbbbaba :124
aabaabaaaaaabaaaaaba :5451
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
babbaababaaaabbbbaba :124
aabaabbabaababababaaa :6179
aabaabaaaaaabaaaaaba :5451
aabaabaaaaaabaaaaaba :5451
aabaabbabaababababaaa :6179
babbaababaaaabbbbaba :124
aabaabbabaababababaaa :6179
babbaababaaaabbbbaba :124
```

離散化した要素のラベル　グループ番号

結果：クラスタリング法の検討

BOOLによるクラスタリング

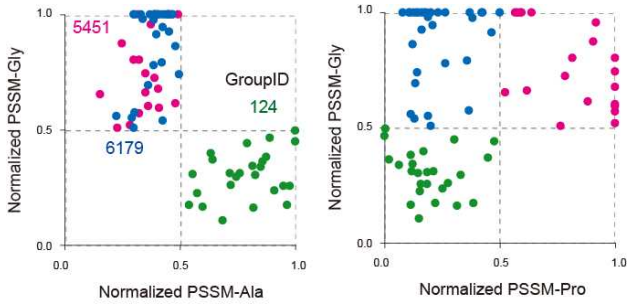
- 実行例: k=2, L=0 (x_i = 0.0 ~ 0.5:a, ~ 1.0:b)



結果：クラスタリング法の検討

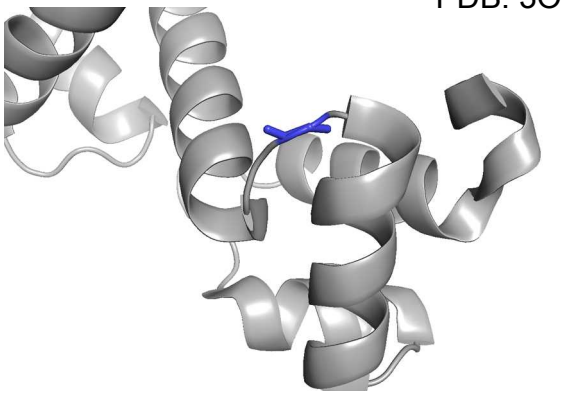
BOOLによるクラスタリング

- 実行例: k=2, L=0 (x_i = 0.0 ~ 0.5:a, ~ 1.0:b)



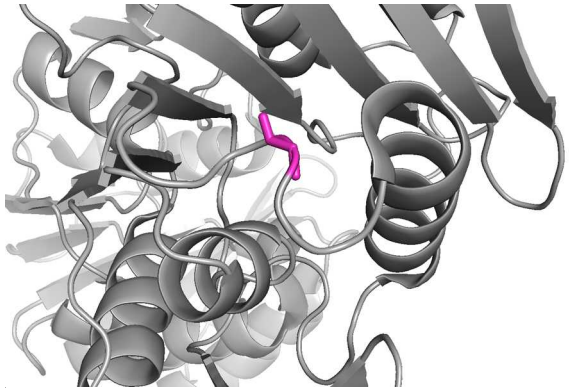
6179.

PDB: 3OIO



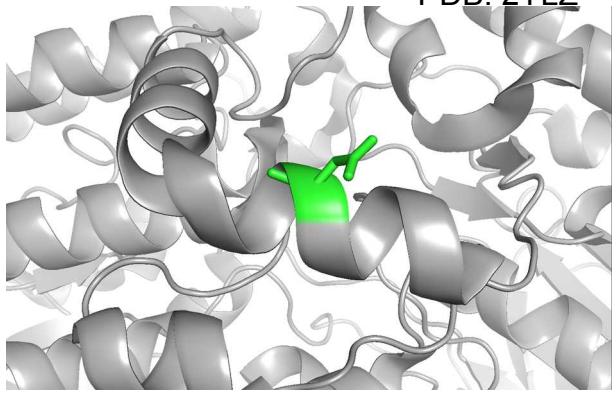
5451.

PDB:2ZUL

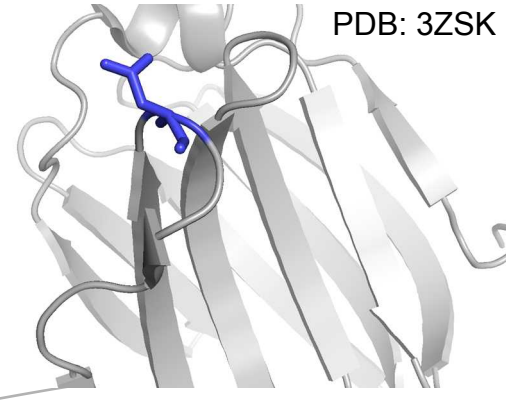


124.

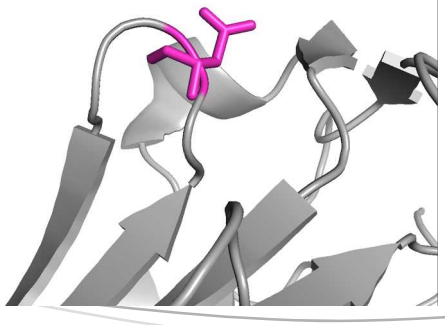
PDB: 2YLZ



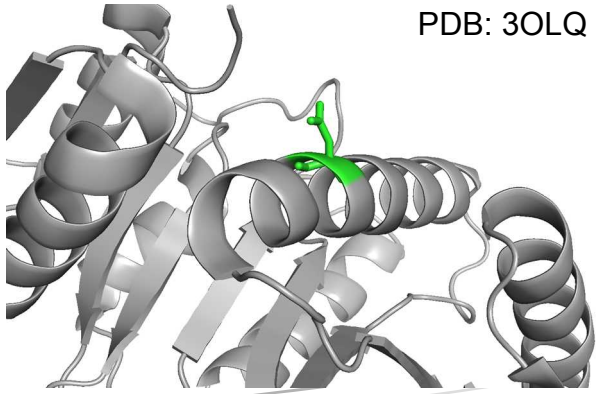
PDB: 3ZSK



PDB: 1PMH

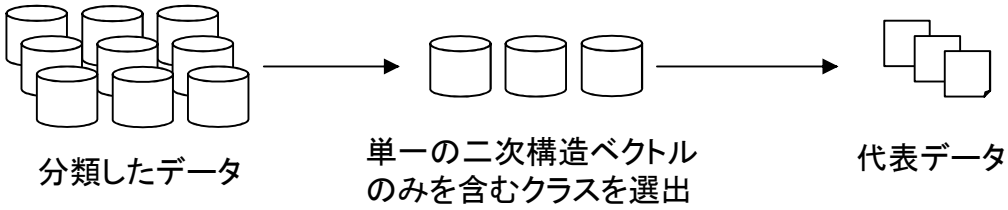


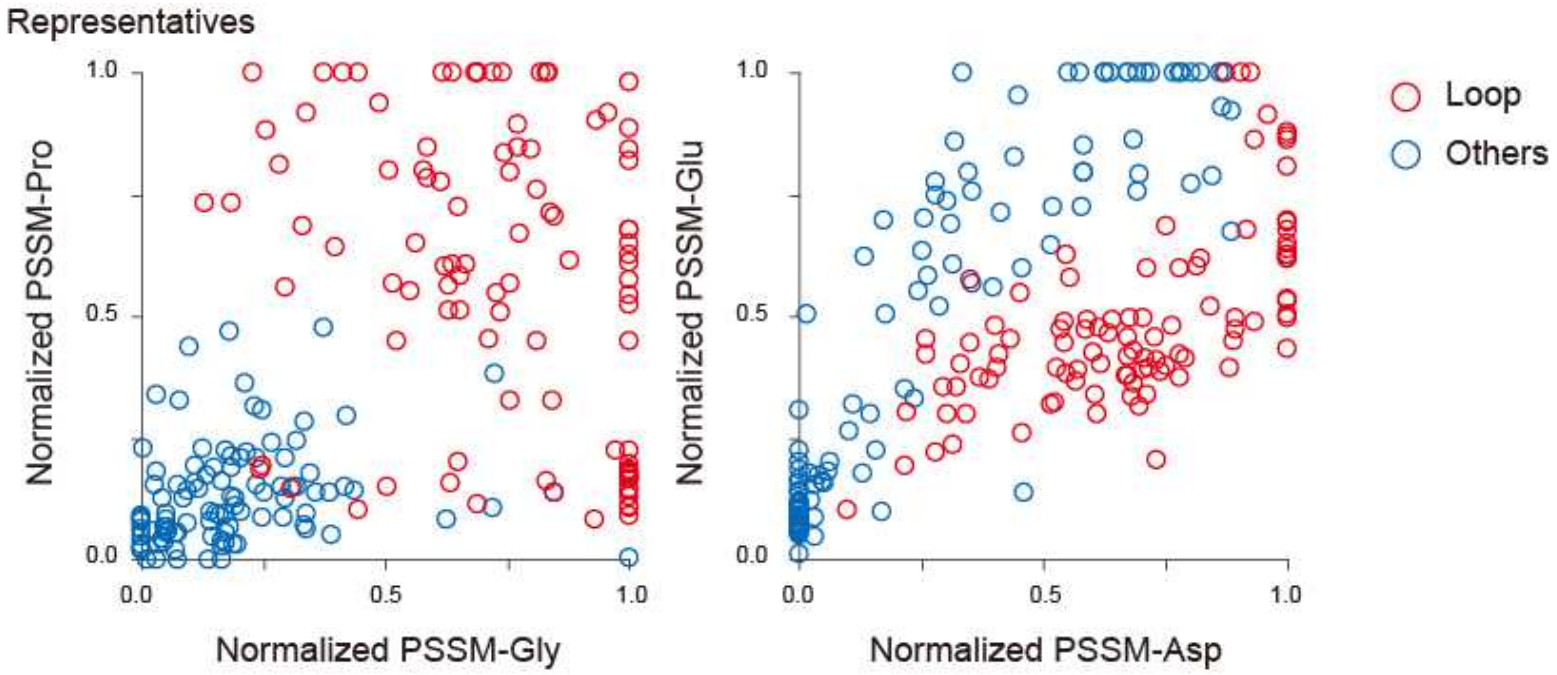
PDB: 3OLQ



結果：代表ベクトルの選出　＆　特徴抽出

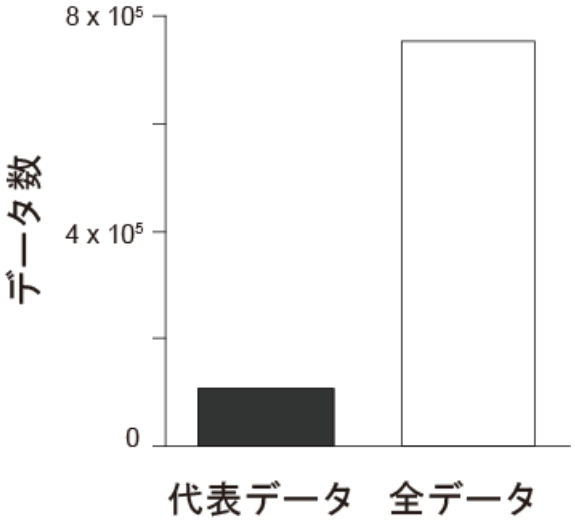
代表データのアミノ酸保存度特徴

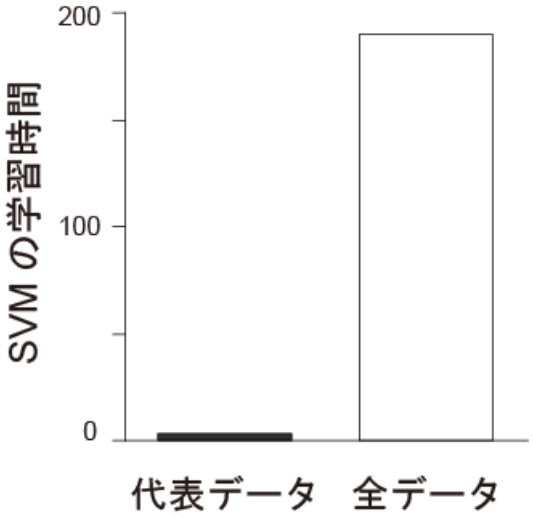


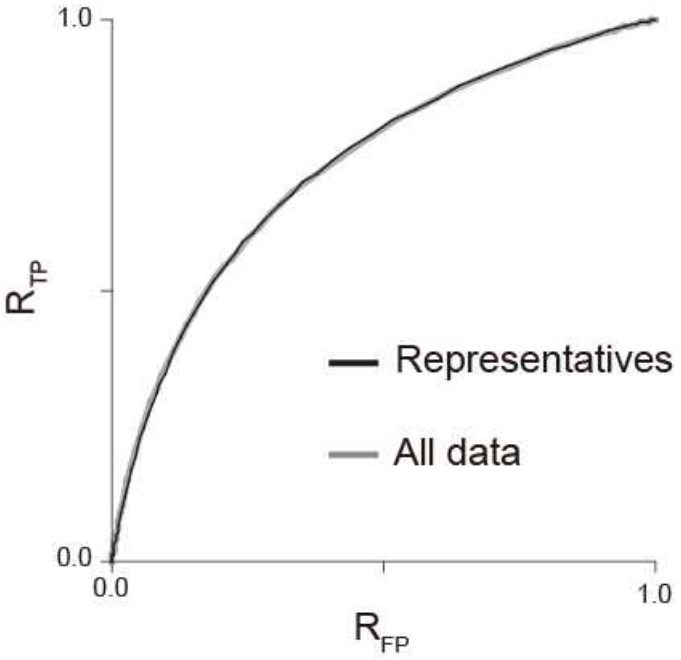


結果：代表ベクトルの選出　＆　特徴抽出

SVMによる特徴検出効率の比較

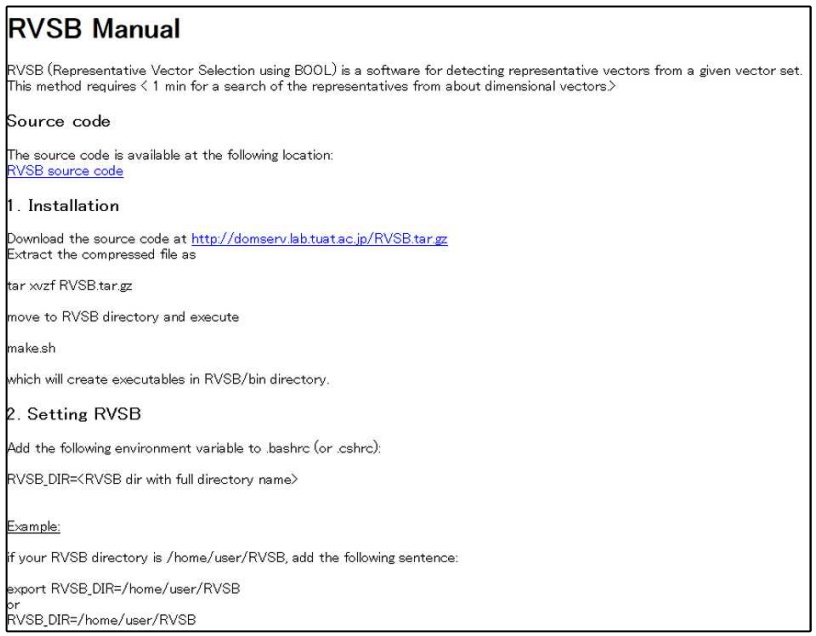






開発したツールについて

- ソースコードを公開しています（http://domserv.lab.tuat.ac.jp/ebina.html）



RVSB Manual

RVSB (Representative Vector Selection using BOOL) is a software for detecting representative vectors from a given vector set.
This method requires < 1 min for a search of the representatives from about dimensional vectors.>

Source code

The source code is available at the following location:
RVSB source code

1. Installation

Download the source code at http://domserv.lab.tuat.ac.jp/RVSB.tar.gz
Extract the compressed file as

tar xvzf RVSB.tar.gz

move to RVSB directory and execute

make.sh

which will create executables in RVSB/bin directory.

2. Setting RVSB

Add the following environment variable to .bashrc (or .cshrc):

RVSB_DIR=<RVSB dir with full directory name>

Example:

if your RVSB directory is /home/user/RVSB, add the following sentence:

export RVSB_DIR=/home/user/RVSB
or
RVSB_DIR=/home/user/RVSB

RVSB: Representative Vector Selection using BOOL

公開中のツール：

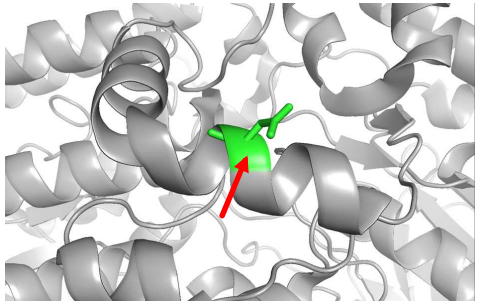
- BOOLの最適パラメータ探索プログラム
- BOOLによるクラスタリング　＆　代表ベクトル選出プログラム
- 代表ベクトルのランク付け、「重要」なデータの選出プログラム

ユーザが作成した任意のベクトルデータについて、クラスタリングのための最適パラメータの決定～代表データの解析が非常に短い時間（～1日）で可能になります。

開発したツールについて

- 応用例

  タンパク質構造・配列の解析：対象残基の機能や構造を推定する
  既存の構造に新規のアノテーションを加える



構造/配列上の任意の残基を選択する

例えば、ユーザが新規に発見した「機能部位」や「構造モチーフ」、新規タンパク質の各残基など

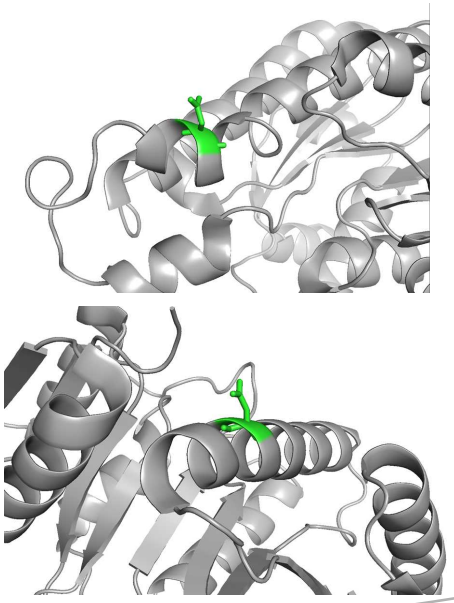
対象の残基をベクトルデータ化＆離散化する

```
1:0.817797 2:0.000000 3:0.556497
babbaababaaaabbbbaba
```

離散化後のベクトルデータベース

```
babbaababaaaabbbbaba :124
babbaababaaaabbbbaba :124
aabaabaaaaaabaaaaaba :5451
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
aabaabbabaababababaaa :6179
```

同じラベルを持つ残基をデータベースから検索
(~ 1 min)



開発したツールについて

- 今後の展望など

1. BOOLを他のクラスタリング法と組み合わせる事によってより「正確」、かつ高速な方法を提案する
2. 解析プログラムの追加（「特徴」の表示など）
3. 特徴抽出に適した代表データ選出アルゴリズムの開発など、ツールの改良を進める