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

JST バイオサイエンスデータベースセンター
「統合データ解析トライアル」 中間激励会
2013年11月29日

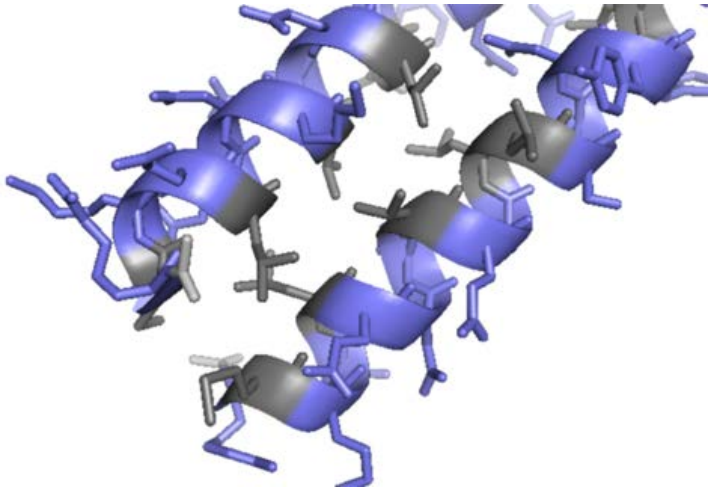
研究開発の目的

「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

Coiled Coilモチーフの配列パターン



内部に側鎖が埋もれているアミノ酸
疎水性

側鎖が露出しているアミノ酸
親水性

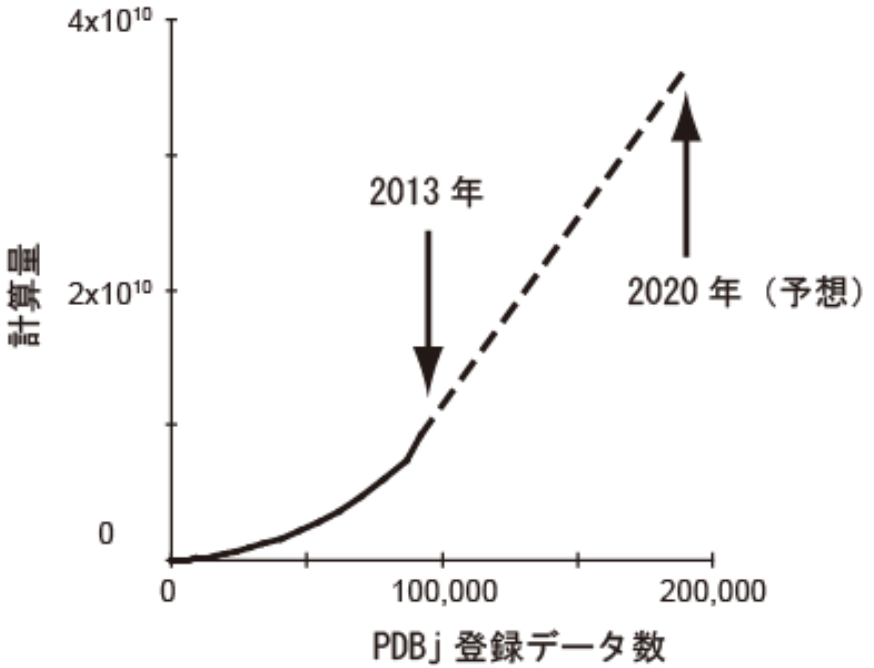
RMKQLEDKVEELLSKNY
HLENEVARLKKLVGER

2~4残基毎に疎水性残基が現れる

研究開発の目的

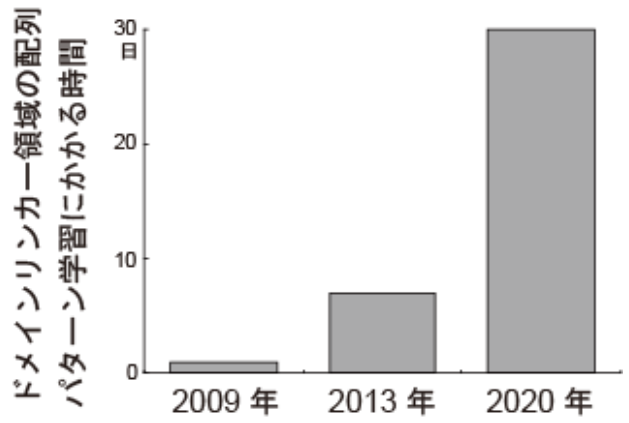
「高速な」タンパク質の配列特徴抽出法を開発する

- 従来の方法の問題点



階層的クラスタリング法の例
（計算量はデータ数の二乗として算出）

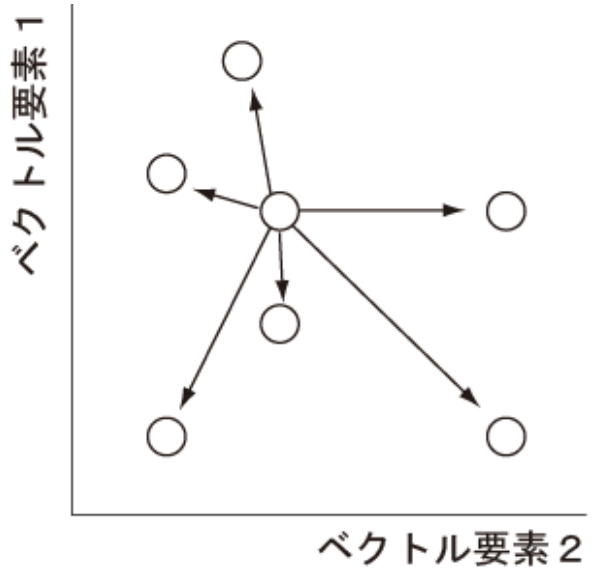
- 個々のユーザが、配列の特徴を検出する事が難しくなる
- タンパク質の構造・機能予測法の開発に膨大な時間が必要になる



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

- 従来の方法の問題点



○ 局所配列に対応するベクトル
ベクトル要素：アミノ酸の出現頻度など

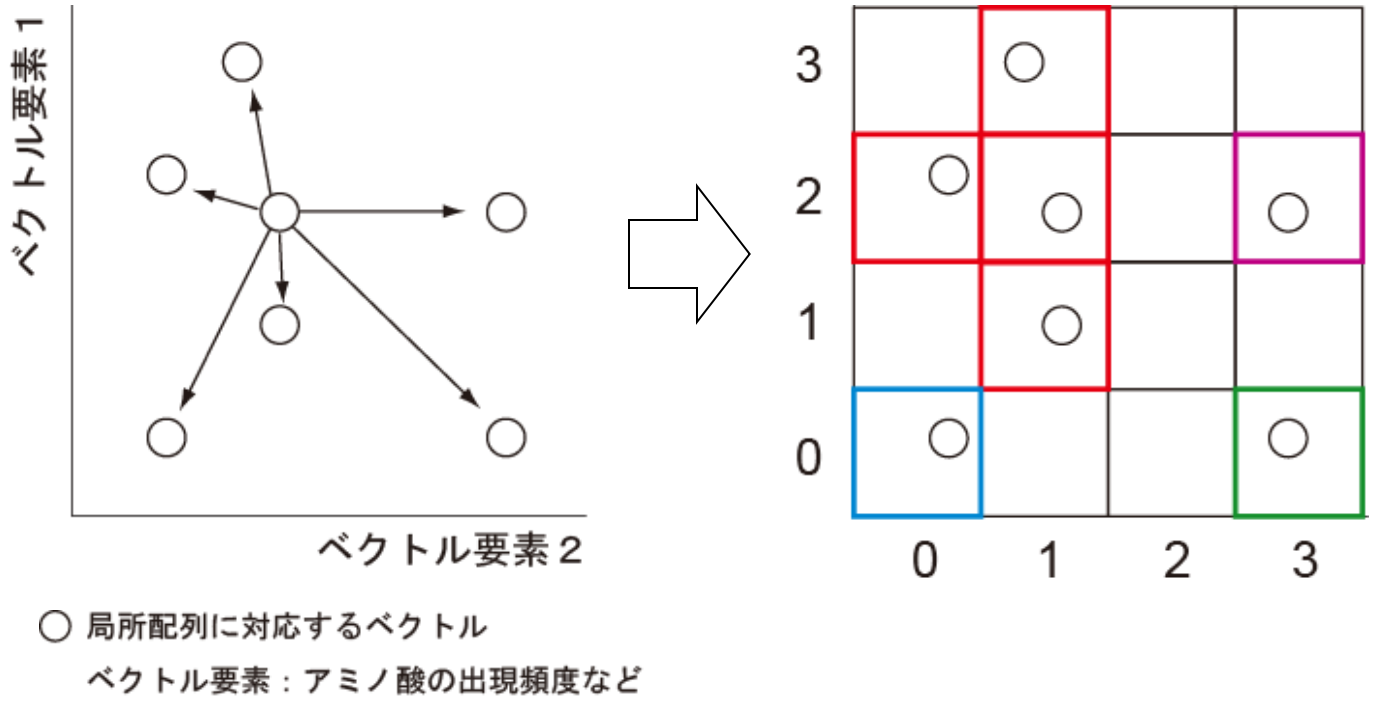
データ間の類似度計算に時間がかかる

計算量は　$N^2 / 2$　以上

研究方法

類似度計算の高速化：BOOLによるクラスタリングを利用する

- BOOL[1]：Binary cOding Oriented cLustering



1. ベクトル要素をk段階に離散化（k=4）
2. 同じ枠か隣接する枠に入っているベクトルを同じクラスとして分類する

[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

研究開発のスケジュール

研究開発項目	平成２５年 １０月	平成２５年 １１月	平成２５年 １２月	平成２６年 １月
１．BOOLによる局所配列クラスタリング法 の最適化				
２．提案手法による配列特徴抽出法の開発				
３．開発したプログラムを提供するためのWebページ作成				

進捗状況

1. 配列特徴抽出のためのテストデータセットの構築

2. BOOLを用いたクラスタリング法の評価

3. クラスタリング法最適化に向けた条件検討

進捗状況

1. **配列特徴抽出のためのテストデータセットの構築**

2. BOOLを用いたクラスタリング法の評価

3. クラスタリング法最適化に向けた条件検討

研究方法：テスト用データセットの構築

- タンパク質配列 – 二次構造データセット

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
724,338残基（代表配列中）

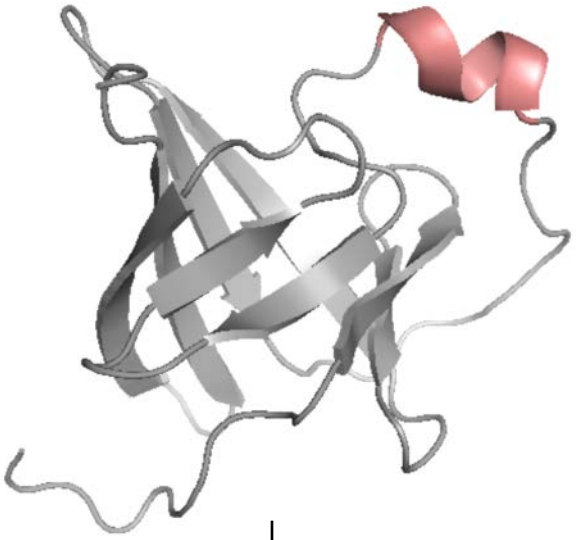
ベクトルデータ化

- 位置特異的スコア行列（PSSM)
- 各残基について±0 or 5 残基分のWindow

ベクトルデータのクラスタリング

研究方法：テスト用データセットの構築

- ベクトルデータの作成方法



DEYQRTWVAVVEEETSFLRARVQQIQV
PLGDAARPSHLLTSQLPLMWQLY・・・

対象の残基±0 or 5残基の断片配列

LGDAARPSHLL

GDAARPSHLLT

DAARPSHLLTS

PSSMの算出
(PSI-BLAST with NR database)

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S
−149	69	−157	−231	−400	429	−129	504	−213	−459	−412	−120	−188	−407	−319	−11
−199	−77	−49	341	−416	48	463	−293	−144	−327	−222	137	−47	−336	−237	−7
320	−288	−274	−266	−271	−219	−214	−109	−301	−315	−271	−204	−278	−391	645	−1
29	24	−240	−308	−207	−50	−214	−293	−254	12	181	−94	465	−150	−58	−1
−88	170	−132	−206	334	11	−145	−241	−211	−25	−181	−13	−173	−84	287	22
−215	−314	−280	−15	−406	44	−206	−350	−307	−375	−193	−224	−324	−407	744	−2
−52	−97	168	−127	−249	177	−22	−23	−163	−197	−334	−111	−240	−325	−224	45
−217	−37	−28	−105	−402	584	50	−323	399	−331	−42	−63	−170	−363	237	−8
−145	−270	−379	−415	−228	−264	−336	−421	−357	234	422	−97	283	−84	−360	−21
11	−92	96	−164	−235	−87	−70	87	−214	−47	5	−137	−117	−259	382	−9
−118	−282	14	−249	−291	−227	−234	−300	−249	−158	14	−221	−155	202	564	1
−88	−138	250	−163	309	−163	−184	−116	−44	−292	−245	−173	−229	176	−268	37

N（= 20 x 断片配列長）次元のベクトル
今回は20 or 220次元

中心残基の二次構造をもとに、Loop、Helix、Sheetのベクトルデータとして利用する。

進捗状況

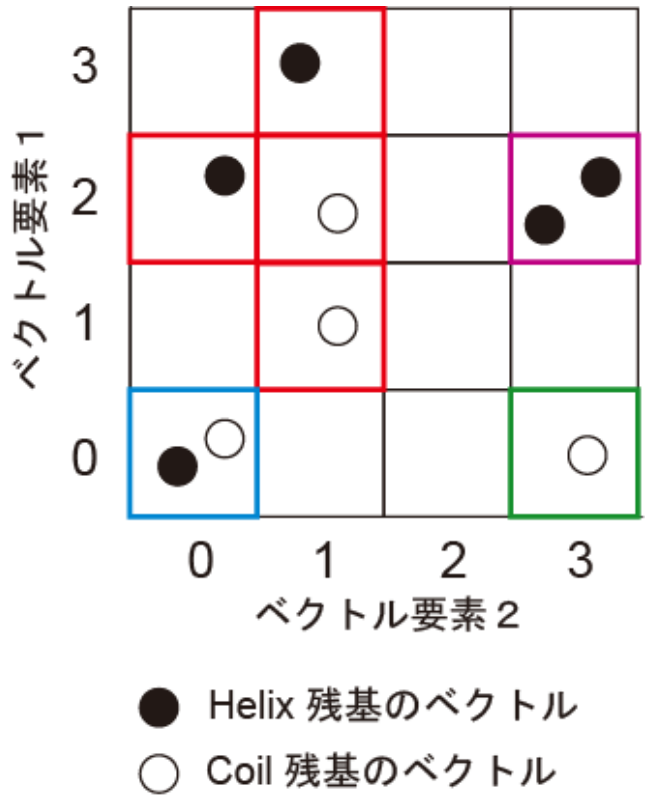
1. 配列特徴抽出のためのテストデータセットの構築

2. BOOLを用いたクラスタリング法の評価

3. クラスタリング法の最適化

研究方法：BOOLによるベクトルデータのクラスタリング

- どのようなクラスタリングの結果が得られればよいのか？



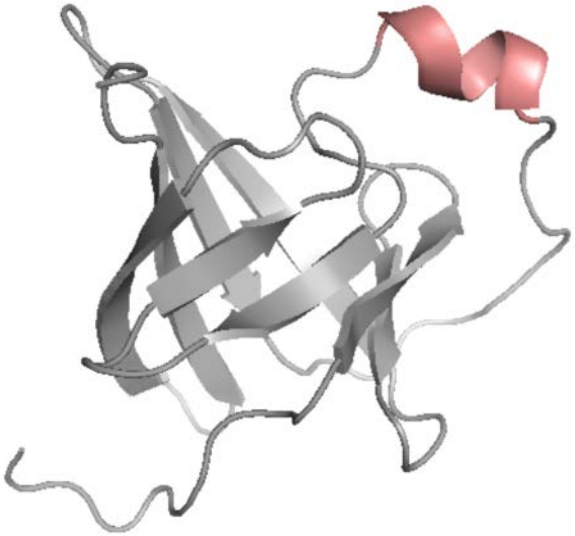
1. 複数のデータを含むクラスタが多く作られる

代表データのみを利用する事でデータ数（≒特徴抽出時間）の削減が期待できる

2. 単一のクラスには同じ二次構造を持ったベクトルのみが含まれる

特徴抽出の効率向上（正確性向上など）が期待できる

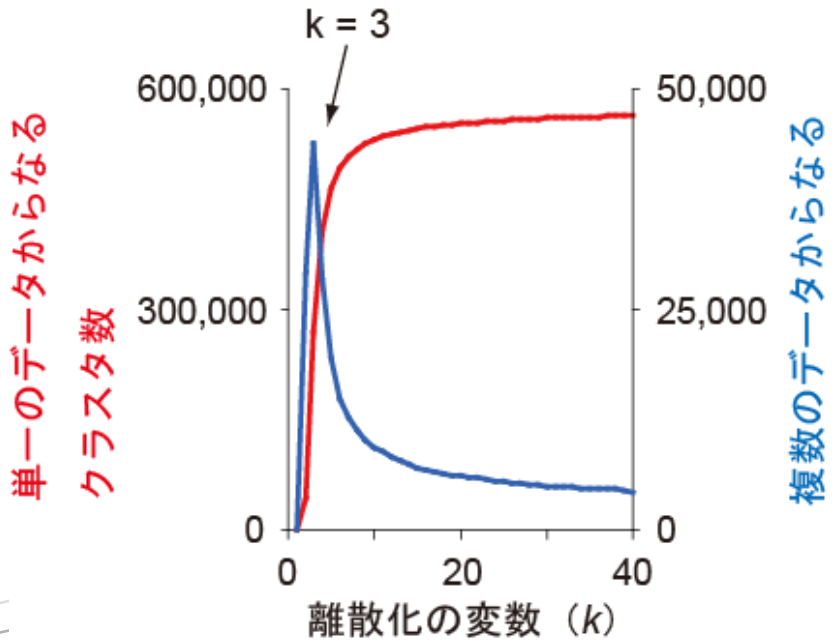
結果：20次元でクラスタリング（WS = 1）

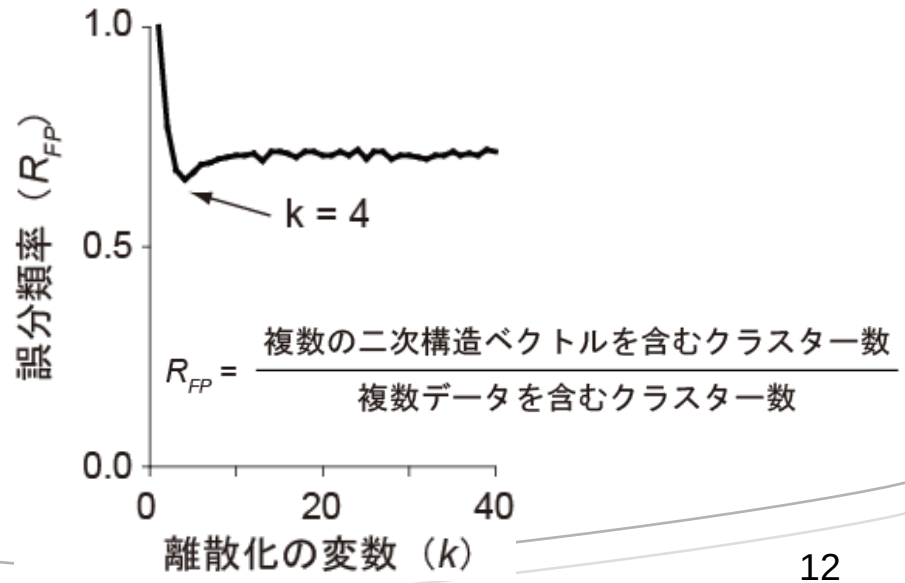


DEYQRTWVAVVEEETSFLRARVQQIQV
PLGDAARPSHLLTSQLPLMWQLY・・・

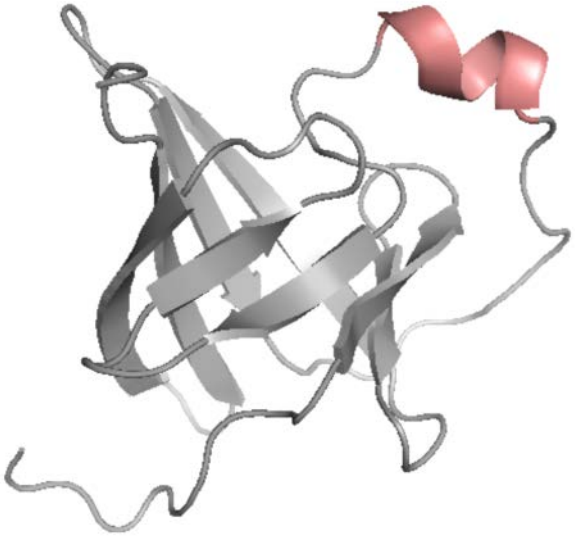
R
P
S

各残基のPSSM





結果：220次元でクラスタリング（WS = 11）



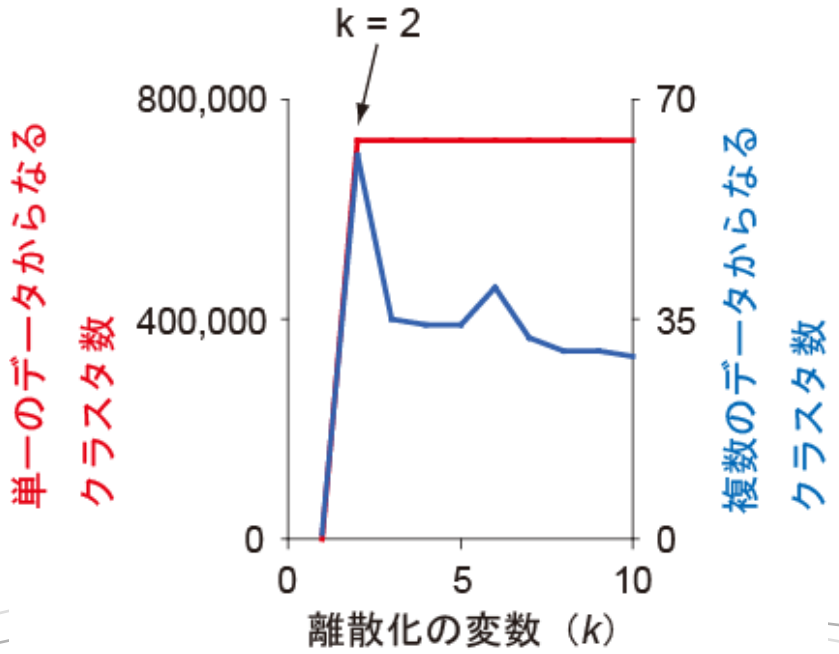
DEYQRTWVAVVEEETSFLRARVQQIQV
PLGDAARPSHLLTSQLPLMWQLY・・・

LGDAARPSHLL

GDAARPSHLLT

DAARPSHLLTS

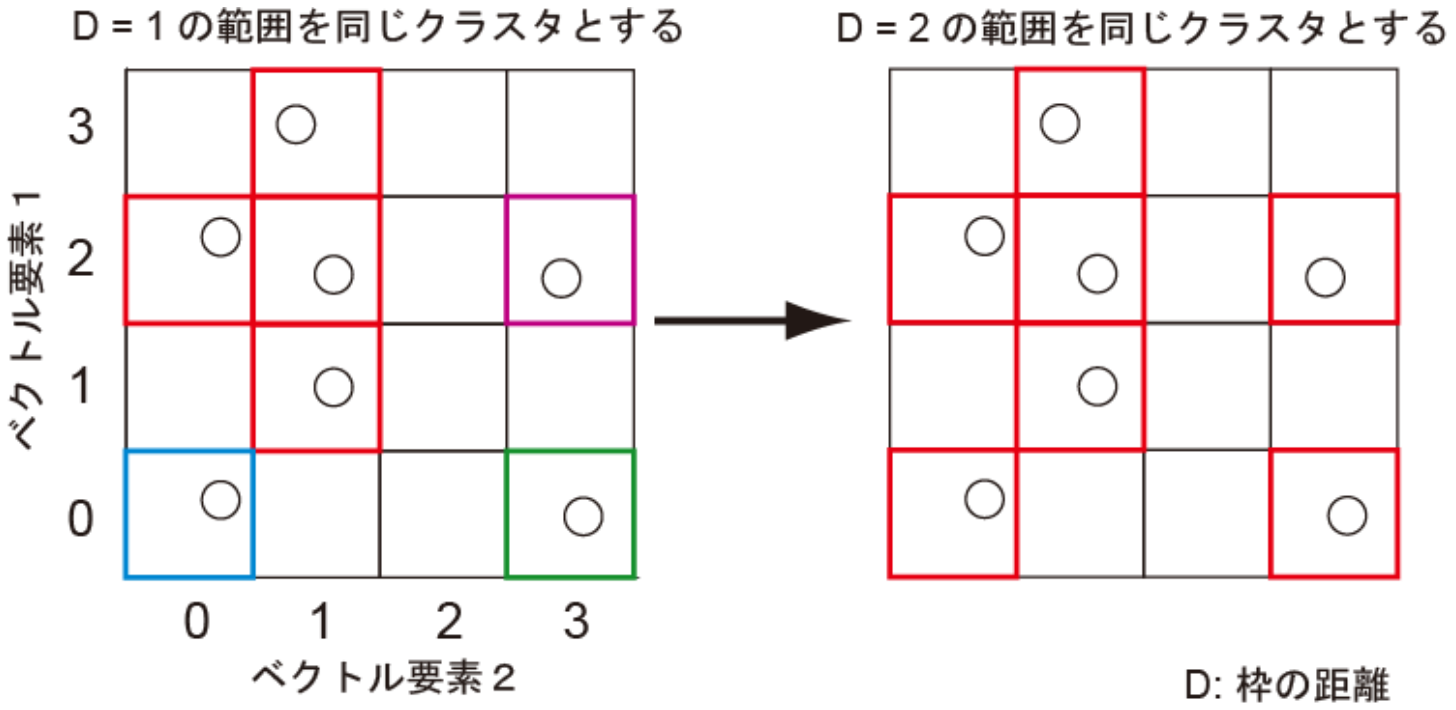
対象の残基±5残基の断片配列



ほぼ全てのベクトルが別々のクラスタに分類されてしまう

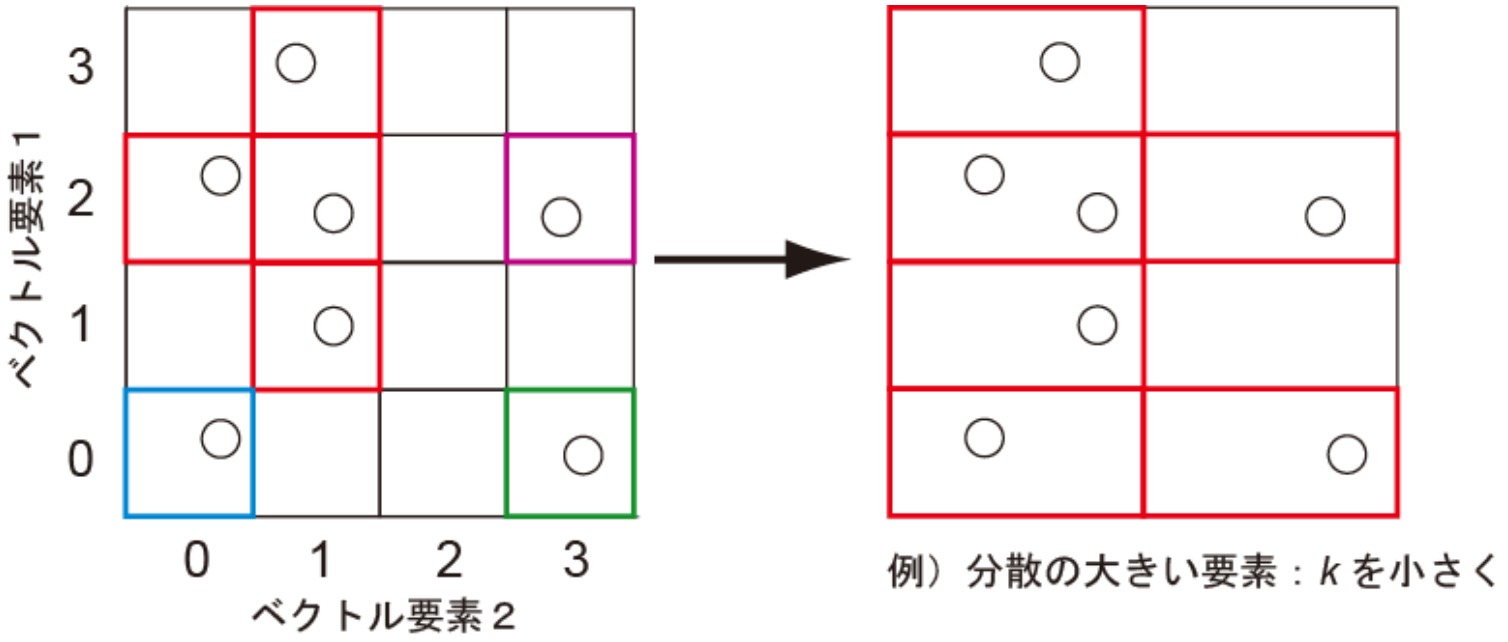
結果：220次元でクラスタリング（WS = 11）

- 解決方法1：同じクラスタとして定義する枠の範囲を拡大する（検討中）



結果：220次元でクラスタリング（WS = 11）

- 解決方法2：要素ごとに離散化パラメータを変化させる（検討中）

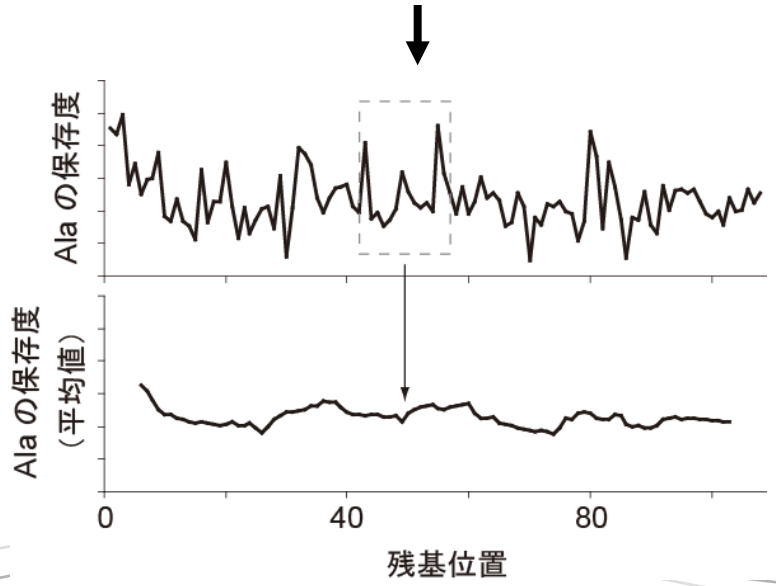


結果：220次元でクラスタリング（WS = 11）

- 解決方法3：ベクトル作成方法を変更する （次数を落とす）

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S
−149	69	−157	−231	−400	429	−129	504	−213	−459	−412	−120	−188	−407	−319	−11
−199	−77	−49	341	−416	48	463	−293	−144	−327	−222	137	−47	−336	−237	−7
320	−288	−274	−266	−271	−219	−214	−109	−301	−315	−271	−204	−278	−391	645	−1
29	24	−240	−308	−207	−50	−214	−293	−254	12	181	−94	465	−150	−58	−1
−88	170	−132	−206	334	11	−145	−241	−211	−25	−181	−13	−173	−84	287	22
−215	−314	−280	−15	−406	44	−206	−350	−307	−375	−193	−224	−324	−407	744	−2
−52	−97	168	−127	−249	177	−22	−23	−163	−197	−334	−111	−240	−325	−224	45
−217	−37	−28	−105	−402	584	50	−323	399	−331	−42	−63	−170	−363	237	−8
−145	−270	−379	−415	−228	−264	−336	−421	−357	234	422	−97	283	−84	−360	−21
11	−92	96	−164	−235	−87	−70	87	−214	−47	5	−137	−117	−259	382	−9
−118	−282	14	−249	−291	−227	−234	−300	−249	−158	14	−221	−155	202	564	10
−88	−138	250	−163	309	−163	−184	−116	−44	−292	−245	−173	−229	176	−268	37

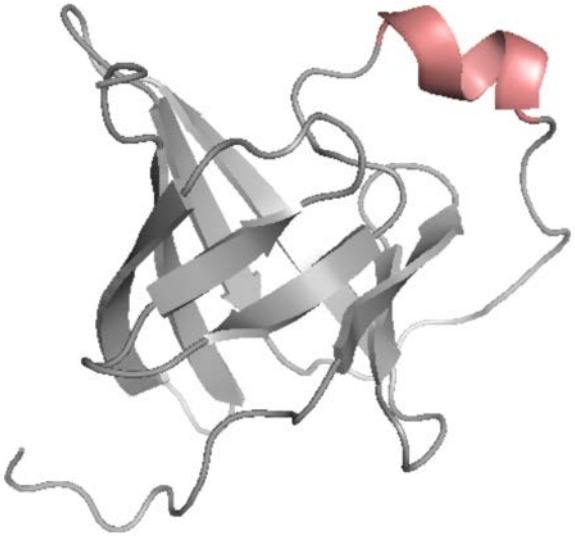
N (= 20 x 断片配列長）次元のベクトル
対象残基＋周辺残基の情報を組み込むため



要素を並べるのではなく、平均値を使う

→ 11残基分の情報を平均する事で
次数を落とす（*N* = 20）

結果：20次元でクラスタリング（WS = 11の平均値を利用）



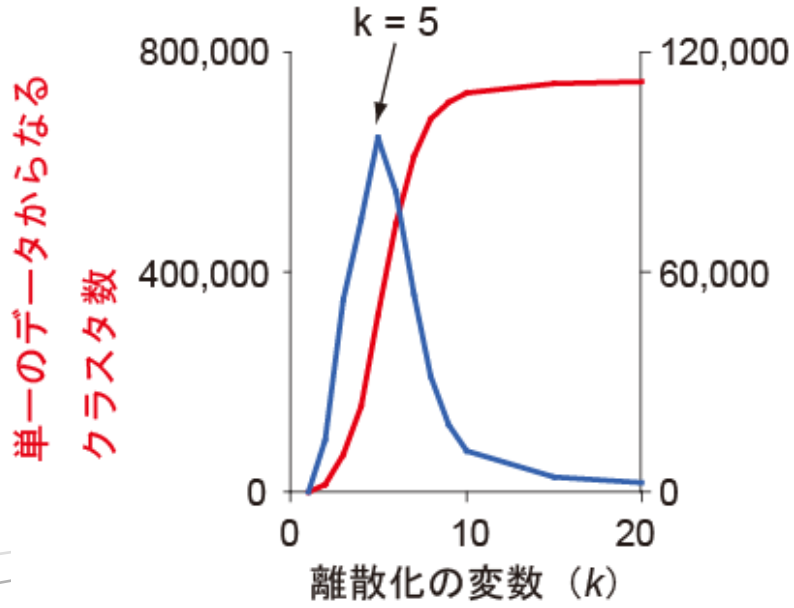
DEYQRTWVAVVEEETSFLRARVQQIQV
PLGDAARPSHLLTSQLPLMWQLY・・・

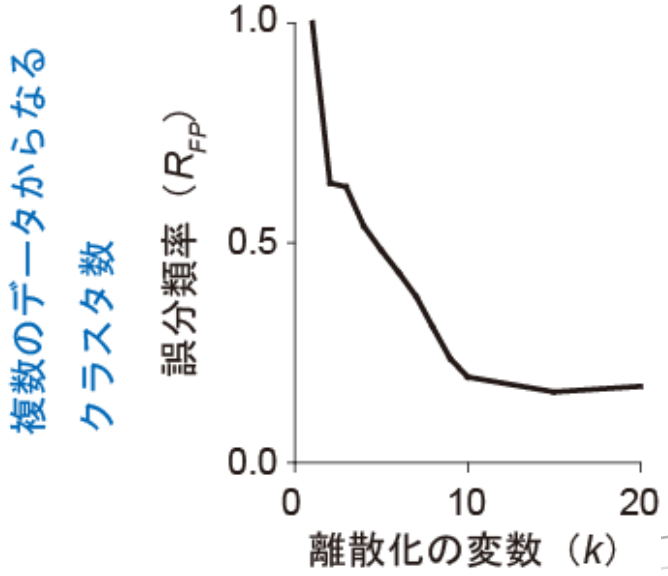
LGDAARPSHLL

GDAARPSHLLT

DAARPSHLLTS

対象の残基±5残基の断片配列





応用例：機械学習法によるパターン認識への応用

Test data

配列＋二次構造情報データセット

ベクトルデータ作成 (Window Size: 11 ; *N* = PSSM 20d x 1 ws = 20)

~ 605,000 vec. (4/5 of the test data)

All

全ベクトルデータ

~ 290,000 vec.

単一データ・クラスタ
のベクトルデータ

Single + Rep

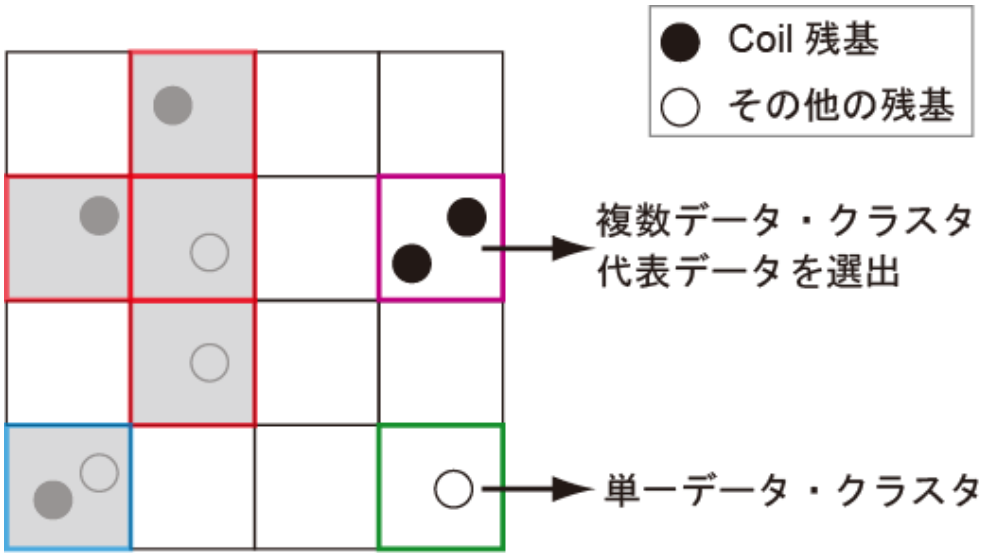
＋

複数データ・クラスタから
抽出した代表ベクトル

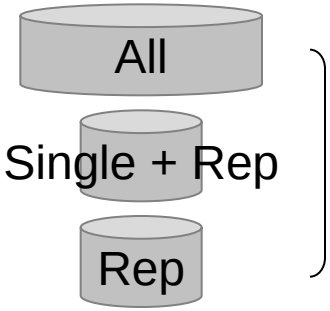
~ 19,000 vec.

Rep

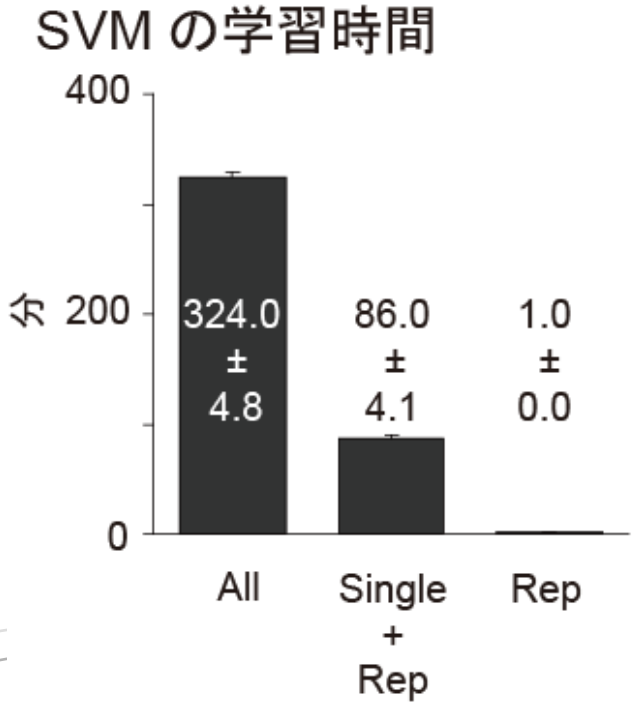
複数データ・クラスタから
抽出した代表ベクトル

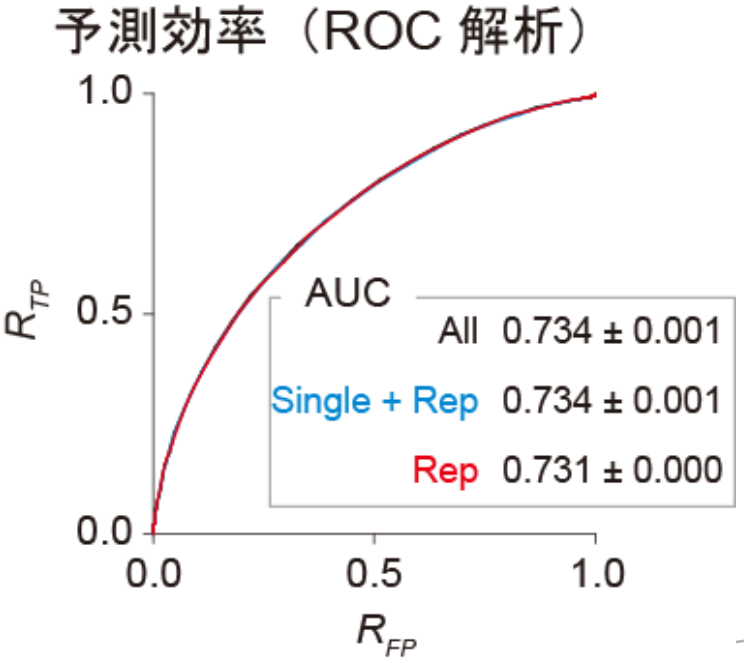


応用例：機械学習法によるパターン認識への応用



SVMによるパターン学習・予測効率の算出
5-fold cross validation test,
SVM^{light}, Linear Kernel,パラメータはデフォルト





今後の予定

- ～2014年　1月
  1. クラスタリングのパラメータ　D、kの最適化
  2. ベクトル作成に用いる特徴の数値化方法の検討
     → 平均値では表せない周期的なパターンのスコア化など
     （自己相関関数などで算出）
  3. 既存の手法との比較
     → 計算時間、分類効率など
- 2014年　1月～
  1. 配列特徴抽出法の開発
  2. 開発したツールのWeb公開