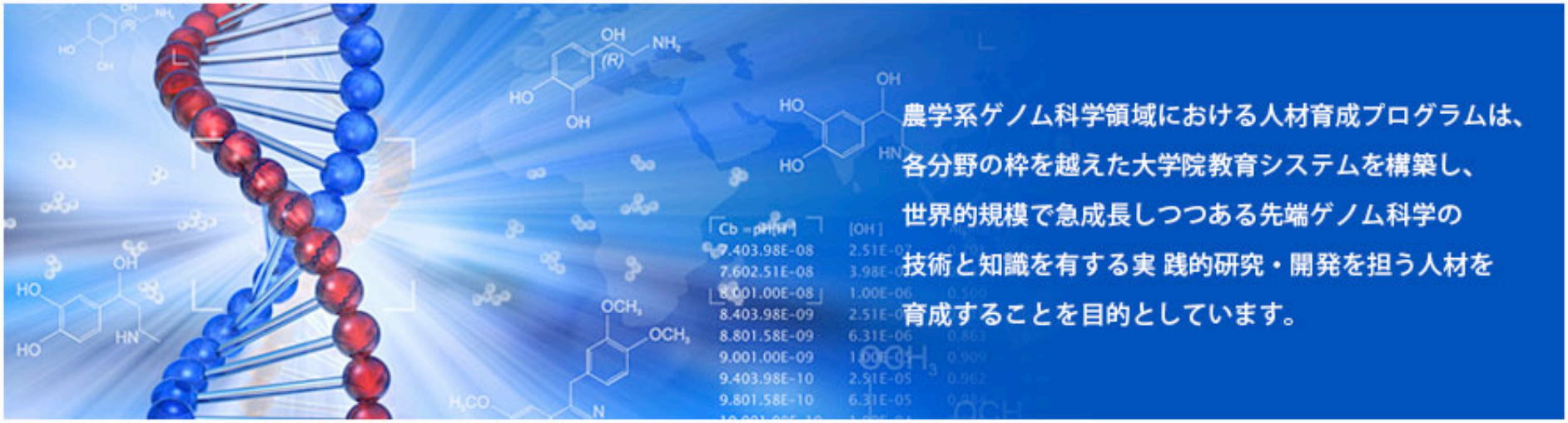
スパコンを用いたS-PLUSのパフォーマンス検証

東京農工大学農学府農学部
石井一夫



内容

1、導入
2、本解析法の3要素
3、本解析法の事例
（大腸ガンマーカー探索）

PMC Machine Learning (並列化モンテカルロ組み合わせ最適化法)

モンテカルロ法、並列分散処理、データマイニングを組合わせた数理モデルの最適化法

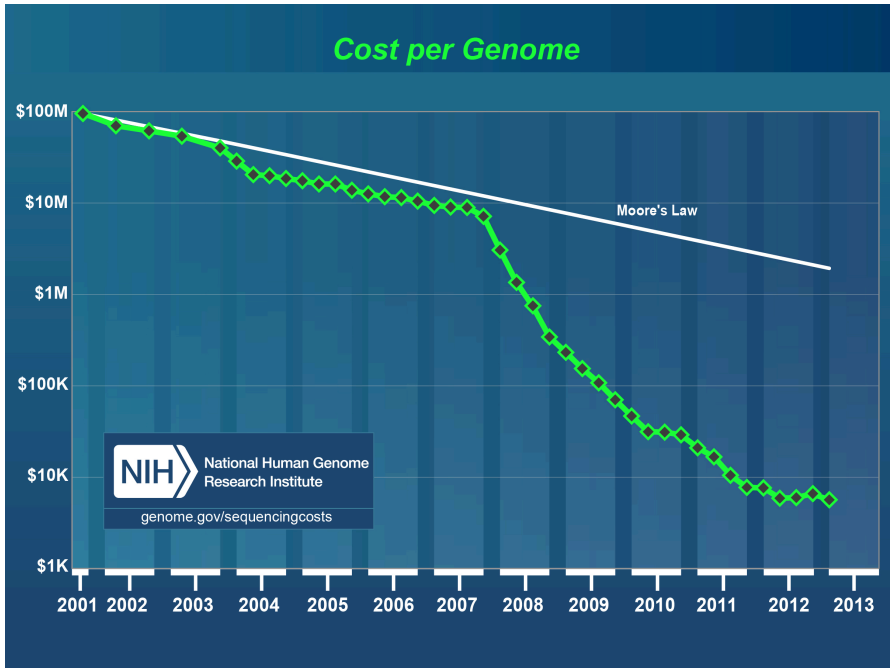
PMC Machine Learning

生物学的な表現型予測（臨床診断やゲノム編集の効果判定など）のための説明変数の探索と数理モデル作成を行うデータマイニング法

背景

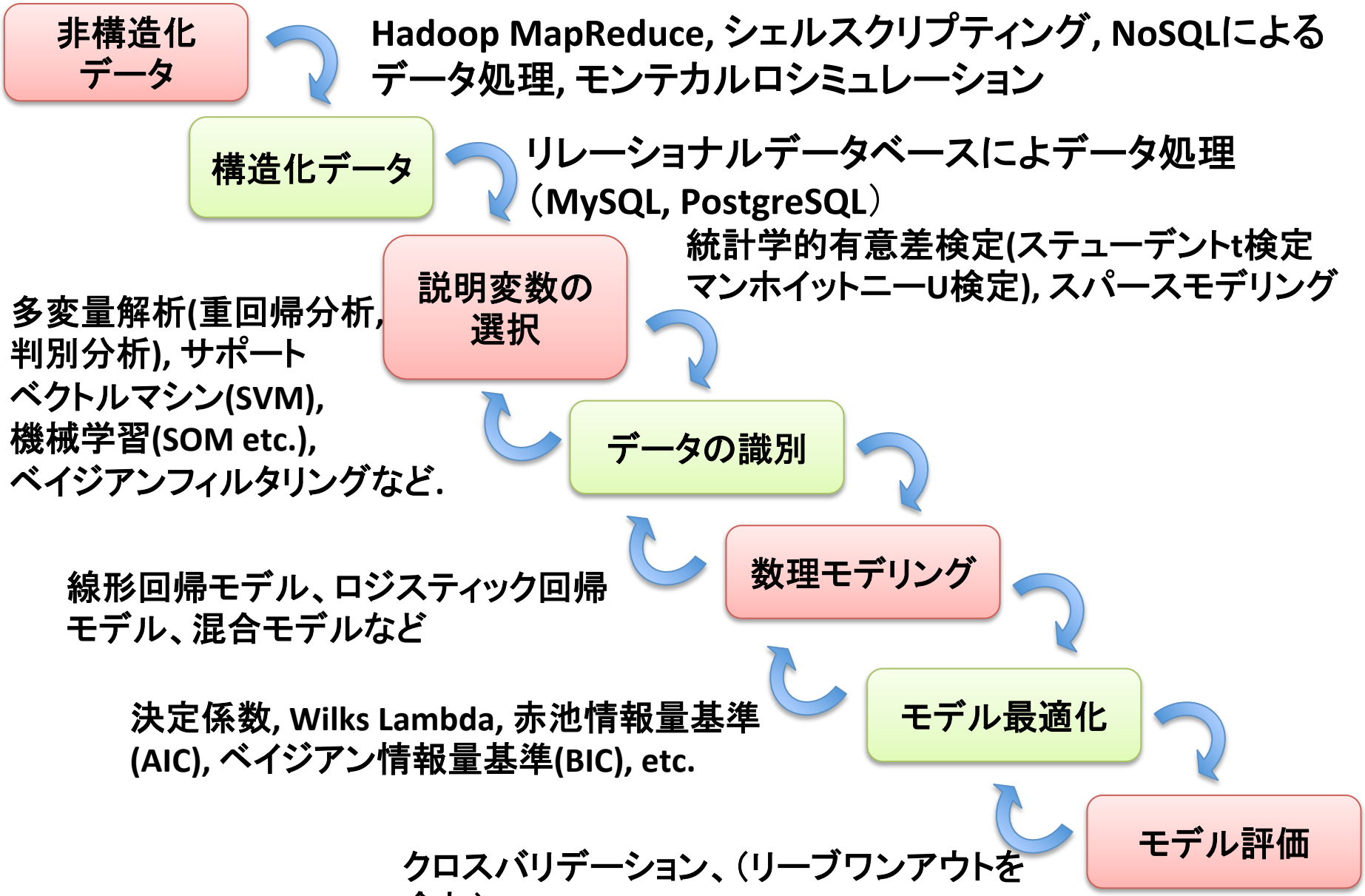
ゲノムビッグデータの時代

- 次世代シーケンサーによるゲノムビッグデータのデータ産生が年々増加.

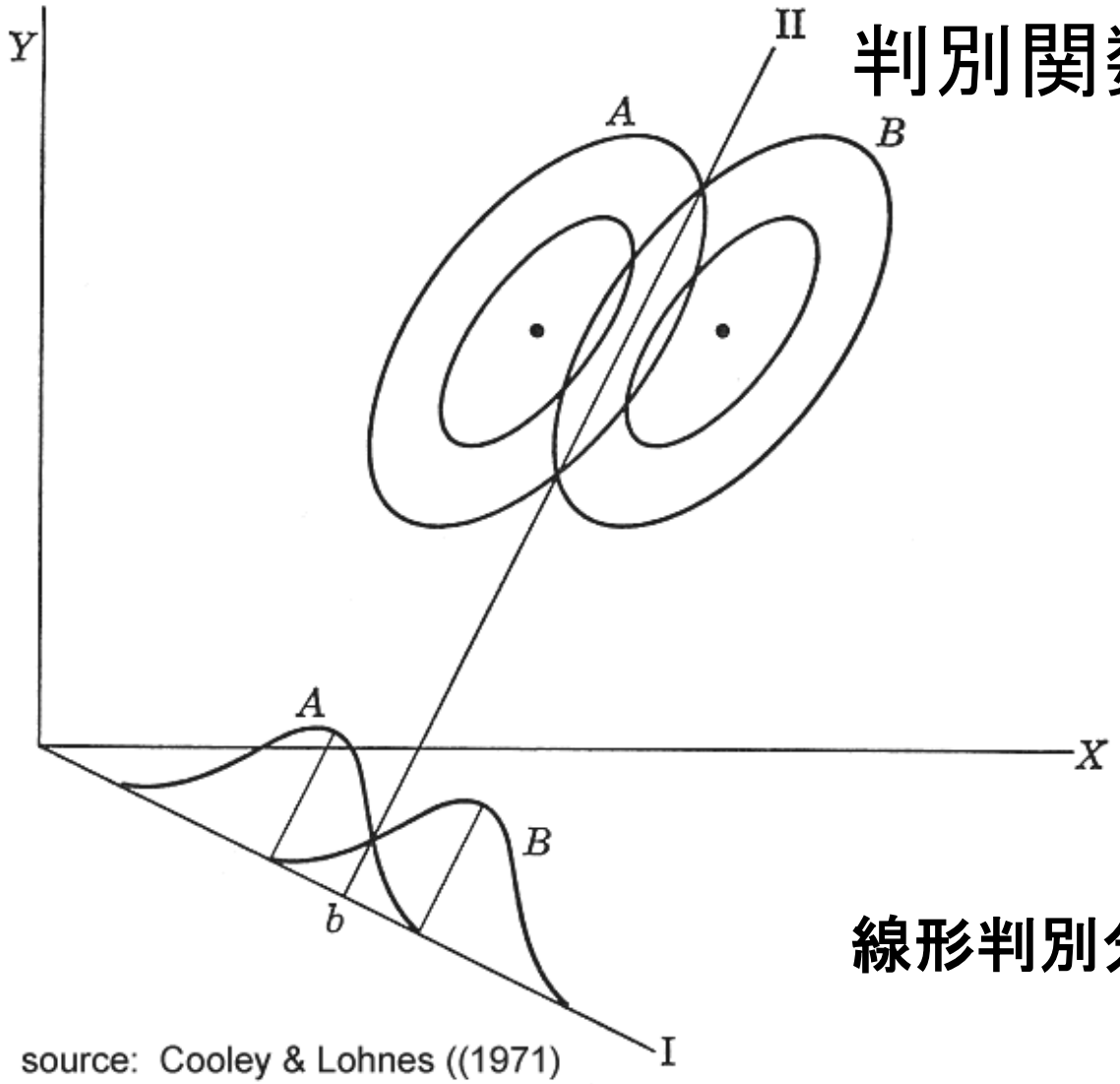


次世代シーケンサー

分析工程の概要
ビッグデータの数理モデリング



データマイニング例



source: Cooley & Lohnes ((1971)

Figure 9.1.

判別関数

線形判別分析(LDA)

線形判別分析による判別関数

判別スコア

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \cdots + u_p X_{pkm}$$

$f_{km} = k$ 群中の事象mを判別する判別関数の判別スコア

$X_{ikm} = k$群中の事象mを判別するための変数Xi

u_i = 関数の特性を形成する係数

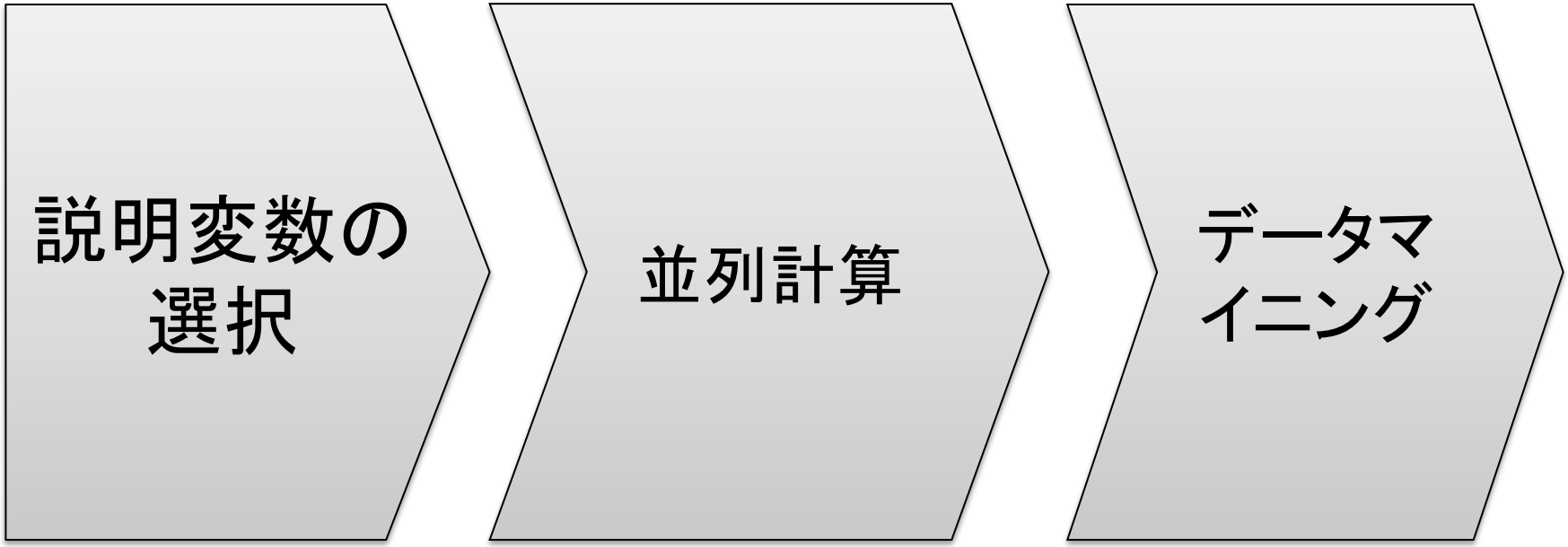
数理モデリング

説明変数の選択に無作為抽出法（モンテカルロ法）を用いる。

大量計算のために、コンピュータクラスタ、HPC（スパコン）のメニーコアCPU、GPGPUなどを使用。

本発明の構成

PMC Machine Learning



無作為抽出
ブートストラップ
ジャックナイフ
マルコフ連鎖
　モンテカルロ法

PC クラスタ
メニーコアCPUによる
HPC
GPGPU

多変量解析
最尤推定法
機械学習
ベイズ推定法

1. ランダムサンプリング

組み合わせ最適化

無作為抽出法; モンテカルロ法

最適化問題の近似的解決法

3つのサンプル抽出方法

1. ブートストラップ – 無作為復元抽出
2. ジャックナイフ – 無作為非復元抽出
3. マルコフ連鎖モンテカルロ法 -

特殊な無作為パターン作成法(ブラウン運動など)

サンプルの性質によりどの方法を使うかが異なる

生物学的ビッグデータの説明変数の選択は通常のサーバでは
しばしば困難となる。

3 データマイニング

- 多変量解析(重回帰分析, 判別分析, 主成分分析, クラスター分析)
- 機械学習(サポートベクトルマシン、ディープラーニング)
- 最尤推定法
- ベイズ推定法

RNA-Seqによる大腸ガン診断マーカーの探索

ゲノムデータを用いた数理モデリング法の開発

臨床データ（大腸ガン18検体
大腸ガン転移巣18検体、
健常者18検体）

PMC-Machine Learning 法と
数理モデリングによる
予測分析

RNA-Seq データ

前処置：
① 0 を 0.00001 に変換
②対数変換
③z 変換 (正規化)
④ステューデント t 検定
⑤欠失値を含むデータの
削除

判別分析による識別：

感度、特異度、ウィルクス
ラムダによる近似的最適化
数理モデルの選択

スーパーコンピュータ
12 TB メモリ, 240 core CPUs

→ 39,254個 および59,710個の候補マーカー

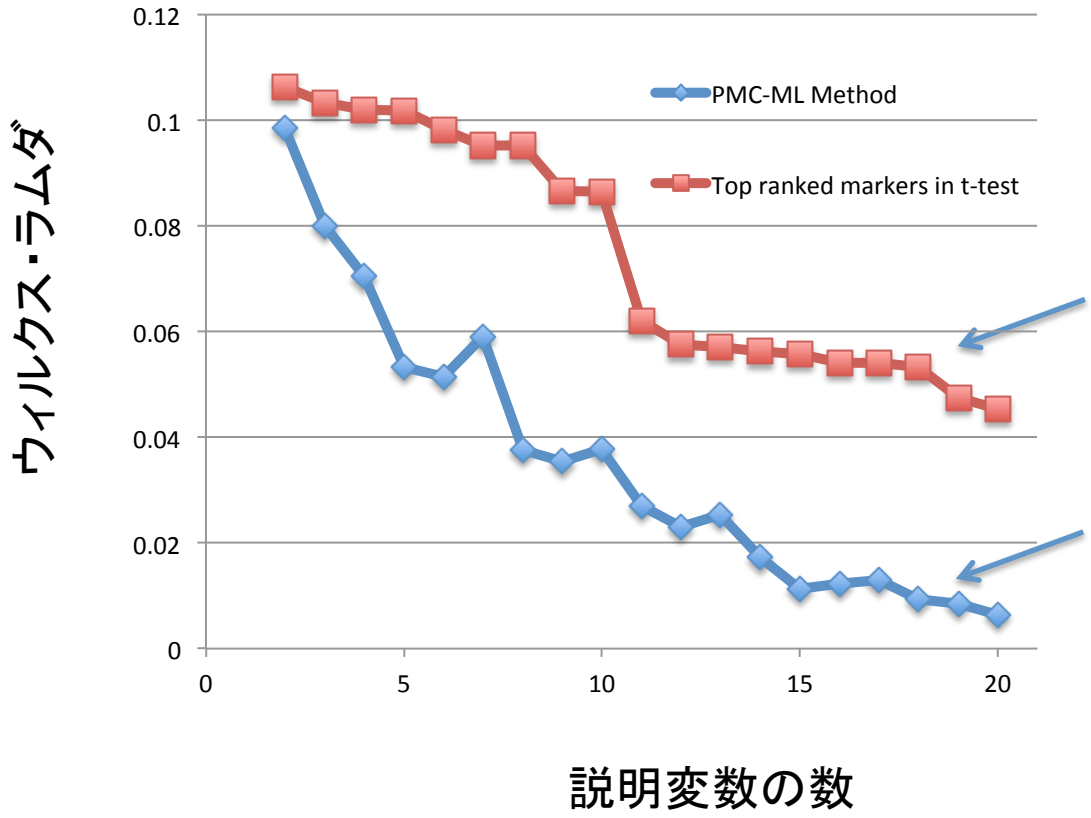
判別分析のための説明変数の選択:
臨床診断のための説明変数の最適化

59710 変数	PMC-Macine Learning 法 (提案方法)					t検定でトップにランクされたマーカーを順に使用		
変数の数	最適化感度	最適化特異度	最適化ウィルクスラムダ	組合わせの数	計算数	感度	特異度	ウィルクスラムダ
2	100	100	0.098494412	1782612195	25000000	94.44444	100	0.1063116
3	100	100	0.080003376	3.55E+13	25000000	94.44444	100	0.1032141
4	100	100	0.070364949	5.30E+17	25000000	94.44444	100	0.1018654
5	100	100	0.053173885	6.32E+21	25000000	94.44444	100	0.1017744
6	100	100	0.051402475	6.29E+25	25000000	94.44444	100	0.09822851
7	100	100	0.058983354	5.37E+29	25000000	94.44444	100	0.09524777
8	100	100	0.037535903	4.01E+33	25000000	94.44444	100	0.09524777
9	100	100	0.03545374	2.66E+37	25000000	94.44444	100	0.08657079
10	100	100	0.037736265	1.59E+41	25000000	94.44444	100	0.08637157
11	100	100	0.026923708	8.61E+44	25000000	100	100	0.06201575
12	100	100	0.023048371	4.28E+48	25000000	100	100	0.05755275
13	100	100	0.025404411	1.97E+52	25000000	100	100	0.05699506
14	100	100	0.017240702	8.39E+55	25000000	100	100	0.05622788
15	100	100	0.011310486	3.34E+59	25000000	100	100	0.05579949
16	100	100	0.012255157	1.25E+63	25000000	100	100	0.05405629
17	100	100	0.012905971	4.37E+66	25000000	100	100	0.0540155
18	100	100	0.009266327	1.45E+70	25000000	100	100	0.05330141
19	100	100	0.008479639	4.56E+73	25000000	100	100	0.04749363
20	100	100	0.006378994	1.36E+77	25000000	100	100	0.04531512

250x100000

各説明変数につき、25,000,000回乱数に基づく説明変数の組み合わせを選択。感度、特異度および、ウィルクスラムダで、評価。

線形判別分析による識別マーカの最適化
(転移性大腸ガン: 健常者= 18 : 18)



t検定で上位にランクされたマーカーを順に使用

今回提案された新規最適化方法

PMC-Machine Learning 法により選択された近似的最適化組み合わせマーカによる Wilks' Lambda は、t検定で上位にランクされたマーカーを順に使用した場合に比べて著しく改善された。

まとめ

無作為抽出、並列計算、データマイニングを組合わせた新規データマイニング法PMC Machine Learningを開発

PMC Machine Learning により生物学的な表現型予測のための近似的最適化数の作成が可能である.

PMC Machine Learningは、ゲノムビッグデータ分析に有効である。

謝辞

東京農工大学
古崎利紀、小林拓嗣、山形洋平

References

特許出願
石井一夫, 古崎利紀, 大森哲郎, 沼田周助,
並列処理装置、並列処理方法、および、並列化処理用プログラム,
出願番号: 2015-5315

ドイツの会社から商品化予定.

完