ビッグデータ、人工知能の
ゲノム医療への貢献
～第２段階の展開へ向けて～

東北大学　東北メディカル・メガバンク機構
東京医科歯科大学　生命医療情報学
田中　博

医療ビッグデータ時代の到来

(1) 次世代シーケンサなどによる「ゲノム/オミックス医療」による網羅的分子情報蓄積

(2)モバイルヘルス(mHealth）によるWearable センサ情報の継続的蓄積（unobstructed monitoring)

(3) Biobankによるゲノム・コホート情報

大量データの急激な
コストレス化かつ高精度化 ← !!

ゲノム：13年→1日(1/5000) 3500億→10万円(1/350万)

個別化医療・予測医療
健康・医療の適確性の飛躍的な増大

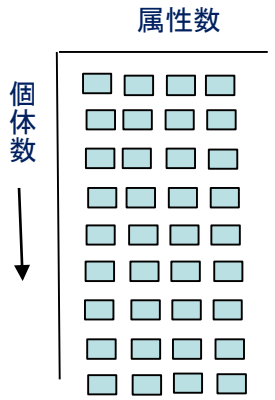
医療の「ビッグデータ革命」

～ゲノム・オミックスデータの基軸的な特徴～

＜目的もデータ特性も従来型と違う＞

従来の医療情報の「ビッグデータ」

Big “Small Data”　(n≫p)



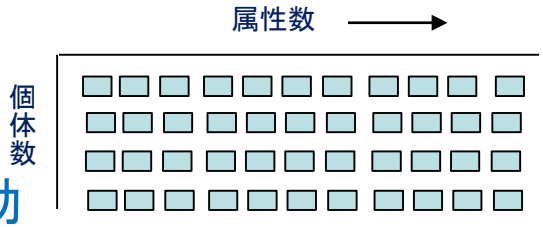
医療情報・疫学調査では 属性数：10項目程度

– 目的：Population MedicineのBig Data

⇒個別を集めて「集合的法則」を見る

網羅的分子情報などのビッグデータ

Small “Big Data”(p≫n)



ゲノム医療　個体データ属性種類数が膨大

属性に比べて個体数 少数:従来の統計学が無効

「P大N小問題」：多変量解析困難:GWASで単変量解析の羅列

– 目的：例えば医療の場合Personalized Medicine

⇒大量データを集めて「個別化パターン」の多様度を摘出

新しいデータ科学の必要性

TMDU

医療ビッグデータ時代の到来（米国）

ゲノム医療の実践

第１段階 ゲノム医療の発展

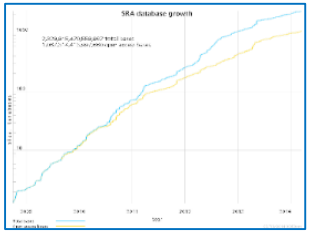
次世代シーケンシングの臨床普及(2010～)

全ゲノム（x30：100Gb)・エキソーム解析（x100：6Gb）

米国では数十の著名病院で実施

ゲノム・オミックス情報の蓄積

2000兆塩基 (2 Pb)
が登録(NCBI:SRA)



医療ビッグデータ

第２段階 医療ビッグデータ時代

医療情報との統合

電子カルテからの
臨床フェノタイプ

医療ビッグデータ

学習アルゴリズム

ゲノム医療知識

人工知能AI

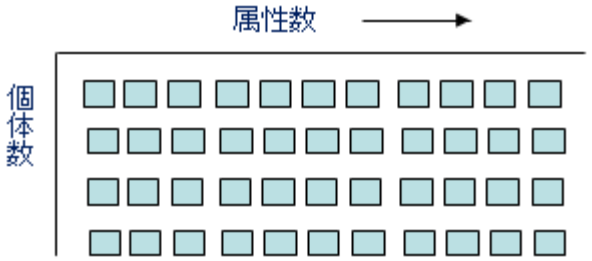
MayoClinicでは
10万人患者WGS

TMDU

「ビッグデータ」のData 原理

問題点　　属性値数(p)≫サンプル数(n)

p: 数億になる場合あり　　n: 多くても数万、通常数千



これら膨大な属性変数がすべて独立ならばビッグデータの構造解析は不可能。単変量解析の羅列（GWASのManhattan Plot)しか可能でない

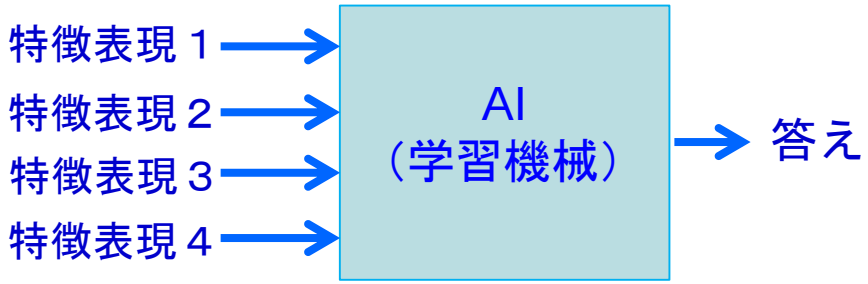
ビッグデータ・スパース仮説

ビッグデータは、多数であるが属性値数より少ない独立成分が基底となって、相互にModificationして構成されている。
（独立成分の推定は、サンプル数とともに増加する）

データ構成性の原理（principle of compositionality)

Deep Learning による人工知能革命

- 機械学習のこれまでの限界
  - 分類・判別する学習機械（システム）
  - 対象の特徴表現ベクトルを与えて分類
  - 与え方に関して細かな技法にとらわれる
- 「教師あり学習」人間を越えられない
  - 分類対象の特徴と正解を与え学習機械（AI）を構築
  - 対象の表現(画像等)と概念を結合できない



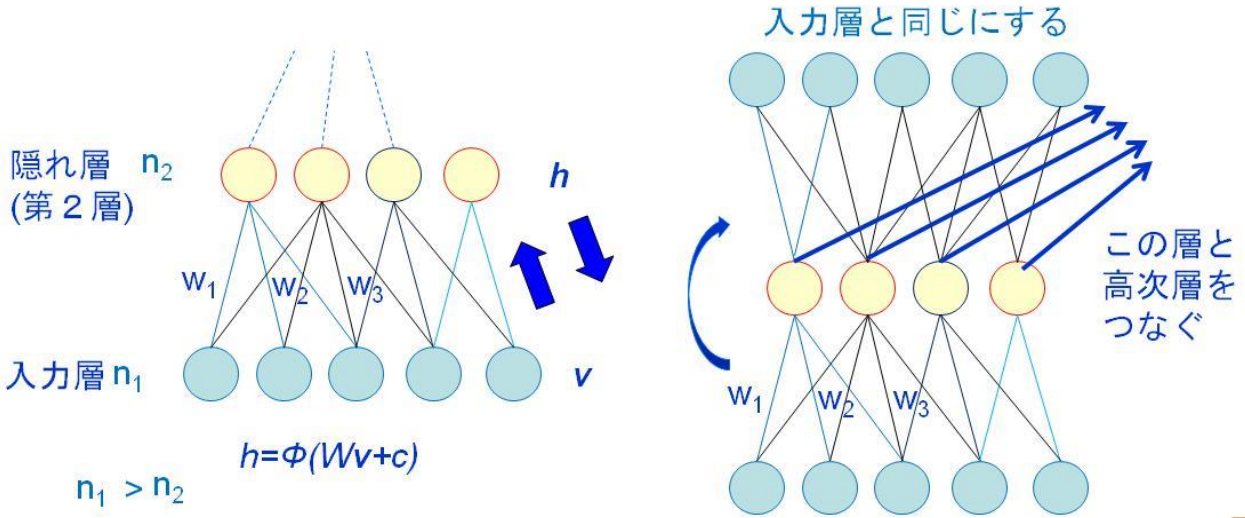
TMDU

Deep Learningの革命性1

- DLは対象の固有の構造を記述する特徴表現や対象の高次特徴量を自ら学ぶ「教師なし学習」を行う
- 「内在的な特徴表現の学習」を自動的に行う
  - 自己符号化（Autoencoder）
  - 制限ボルツマンマシン
- 多層ネットの各層ごとに次元の少ない中間層を介して復元できるか

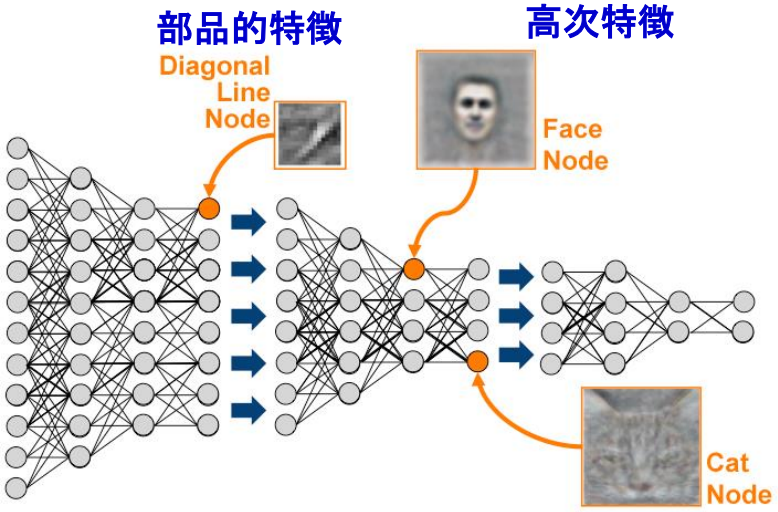
  →できるだけ復元に効果的な特徴量を探索する

  →内在的な特徴量を見出す

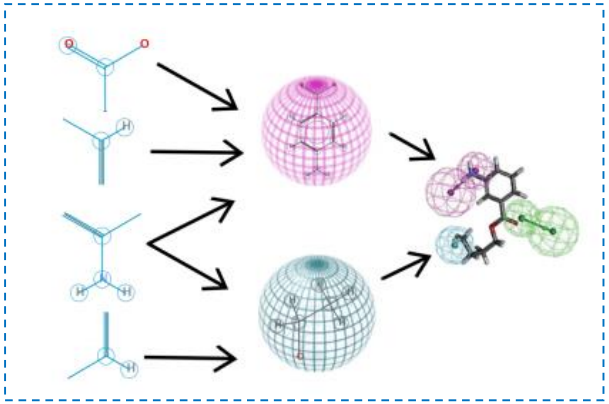


Deep Learningの革命性２

- 各層ごとに自己符号化を行うので何層でもネットを組める →Deep Learning
- 第一層で学習した特徴量を使って、つぎの階層を作るので高次の特徴量が作られる
- 特徴的表現と概念を結びつけるため「教師あり学習」が最後に必要である
- 自動特徴抽出によってこれまでの学習手法の限界を克服した（構造的理解）



AI創薬への応用（構造活性相関学習）
Pharmacophoreの摘出



参照→ http://ngbrc.com/concept2.html（NPO次世代生命医学研究所：講演スライド集）

次の段階へ：現在のゲノム医療の致命的限界

成功した臨床実装（米国では2010年から今日までこれだけ）

1. **希少先天遺伝疾患の原因遺伝子**を**病院の現場で**シーケンサにより同定
2. **がんのドライバー遺伝子変異**を同定、適切な**分子標的薬**を処方
3. 患者の**薬剤の代謝酵素の多型性**を**先制的に同定**し、副作用を防ぐ

しかし

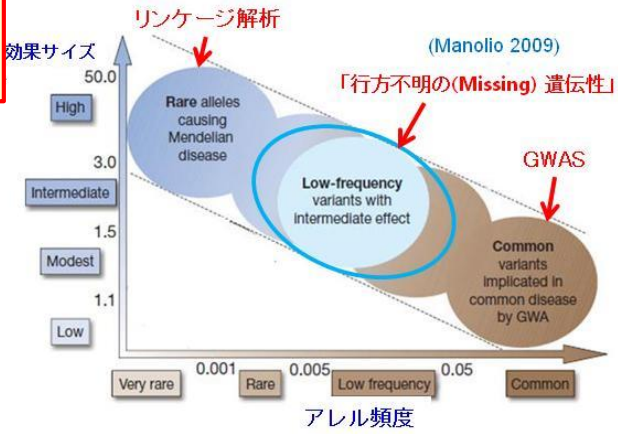
多因子疾患の機序/発症予測などには全く無力である

- 「単一遺伝的原因」帰着アプローチの限界
- 「行方不明の遺伝力」の主要な原因
  - 複数の疾患関連遺伝子間の相互作用: GxG
  - 環境と遺伝子の相互作用が：G x E

SNPの相対リスク低い(1.1～1.3)理由
GxE組合せ特異的効果を環境要因の平均

多因子疾患は個人の＜遺伝的体質と環境要因＞の＜相互作用の結果。シーケンスだけでは解明不能

疾患発症の遺伝要因と環境要因の相互作用は
加算的 $(G \oplus E)$でもなく乗算的 $(G \otimes E)$ でもない
＜(G,E) 組合せ特異的な効果＞である
例　大腸がんの遺伝要因と環境（生活習慣）要因



多因子疾患機序解明の地平

＜遺伝子要因と環境との相互作用の基底＞はどんな機序で行われているか

エピゲノム

環境によるエピゲネティク修飾

オランダ
飢饉 (1944)

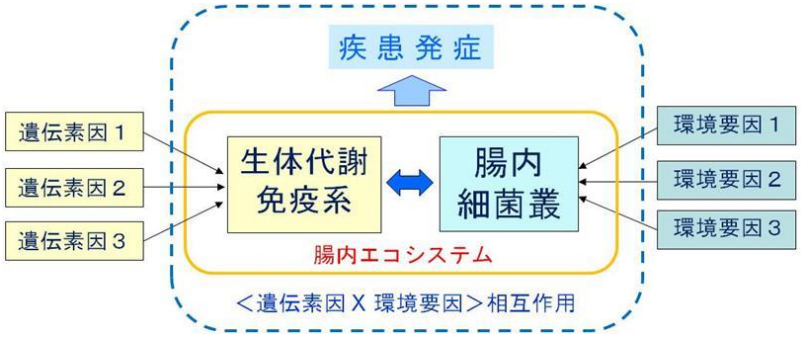
DOHaD(Developed Origin of Health and Diseas)学説
オランダ飢饉のとき、母親の胎内にいた人々
出生30年後、肥満、糖尿病、心疾患、高罹患率
過度な低栄養：肝臓のPPARα/γ（倹約遺伝子）メチル化低下・遺伝子発現がオン
エピジェネティック変化は**可変**：短期的変化、**長期的「記憶」**次の世代も

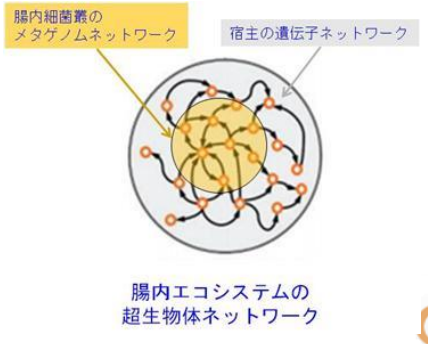
環境因子 → Epigenome**変化** ⇒ **遺伝子発現調節** → **疾病発症**

メタゲノム

Microbiomeにおける生体細菌叢相互作用

- 食事などの栄養物質による環境要因は、**腸内細菌叢の代謝物を介して**、宿主の生体機構に相互作用
- 心筋梗塞や糖尿病、**腸内細菌が産出する代謝物**（短鎖脂肪酸やTMAOなど）が**生体シグナル物質**や**生体活性物質**となって**受容体や転写因子の活性化**して生体側の**遺伝子ネットワーク**に働きかける。
- 腸内細菌叢と生体の＜超生物系（supra-organization）＞において＜環境要因x遺伝素因＞の相互作用





メタゲノム

TMDU

今後のゲノム医療における バイオインフォマティクス

第２段階のゲノム・オミックス医療

多因子疾患の機序・発症予測のBioinformatics

- ゲノム情報だけでなく環境情報（exposome）との相互作用：GxE=T（phenome）
- Clinical Sequencingだけでは解明できない

Bioinformaticsと医療情報学との融合の推進

- 従来のゲノム情報学の延長線上にゲノム医療はない
- これまでのBioinformaticsと異質な医療の世界
  - 電子カルテ、臨床情報処理、病院情報システム、健康情報
- 国レベルのゲノム医療“IT”センターの必要性
- 国規模の臨床オミックス統合データベース
  - データベースのアーキテクチュア研究・開発の予算措置
  - 医療AI・医学knowledge discoveryの推進

TMDU

ご清聴ありがとうございました

TMDU