

人工知能（AI）創薬の現状と将来

東京医科歯科大学 名誉教授

東北大学 東北メディカル・メガバンク機構

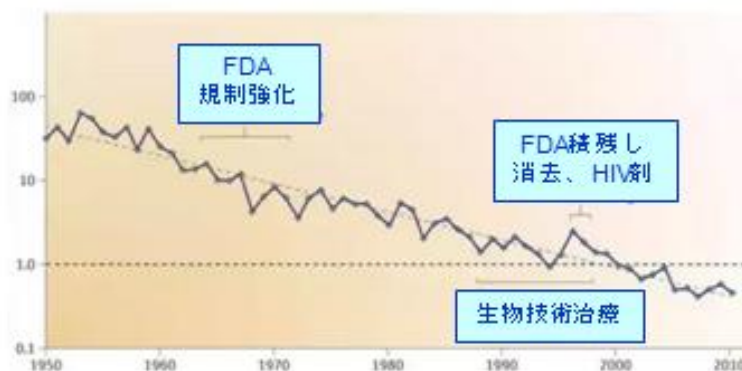
田中 博



創薬をめぐる状況と解決の方向

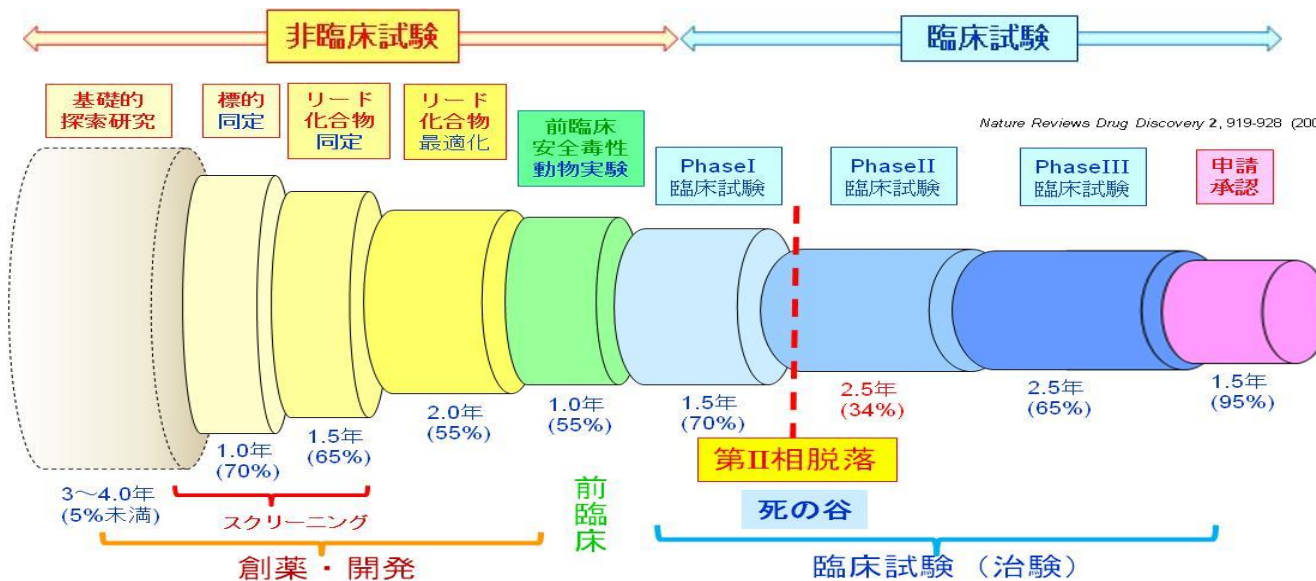
- 医薬品の開発費の増大
 - 1 医薬品を上市するのに約1000億円以上
- 開発成功率の減少
 - 2万~3万分の1の成功率
 - とくに**非臨床試験**から**臨床試験**への間隙
 - **phase II attrition** (第2相脱落)
- 臨床的予測性
 - 医薬品開発過程の**できるだけ早い段階**での**有効性・毒性の予測**
- **臨床予測性の早期での実施**
 - 罹患者のiPS細胞を使う

10億ドル開発費で薬剤数



Nature Reviews Drug Discovery (2012)

ヒトの<薬剤-疾患-生体系>のビッグデータを早期R&D段階で使う



Nature Reviews Drug Discovery 2, 919-928 (2003)

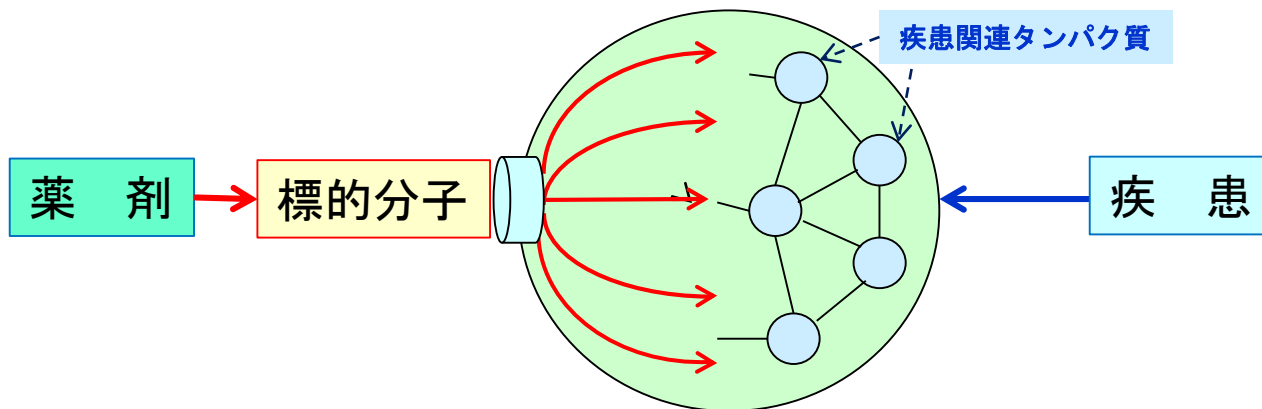
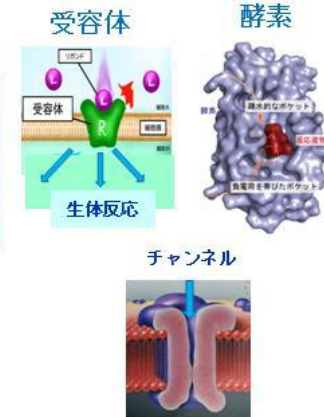
疾患・薬剤・標的の関係

病気の主要な要因

疾患関連タンパク質（複数）

薬：疾患関連タンパク質に影響を示す
標的タンパク質に作用し阻害する

薬剤の標的分子
受容体・酵素・チャンネルなど



生体システム/ネットワーク

ビッグデータ計算創薬 1

計算創薬(computational drug discovery)の新しい方向

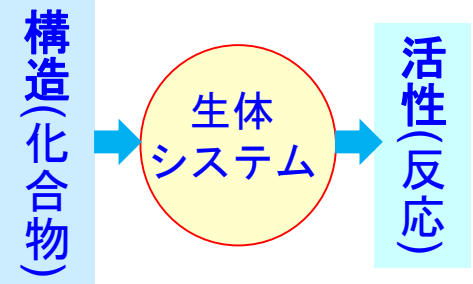
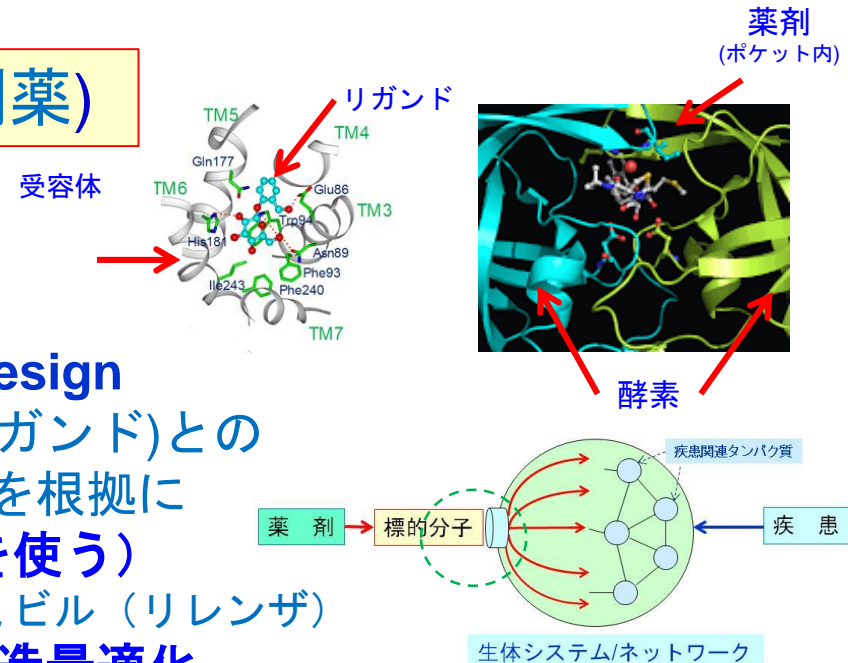
これまでの計算創薬 (*in silico* 創薬)

分子(結合構造)中心

- 分子構造解析・分子設計
- Structure-based rational drug design
- 標的分子(受容体・酵素)と薬剤(リガンド)との結合構造(ポケット)の分子構造を根拠に
- リガンドの分子設計(量子化学等を使う)
 - 成功例: インフルエンザ薬 ザナミビル(リレンザ)
- 標的に結合するリード化合物・構造最適化
- 結合後の生体システムの反応・振舞い
 - 明確な取扱いがない

定量的構造活性相関(QSAR)

- 化合物の分子構造と生体活性の関係
- しかし両者の間には生体システムがある

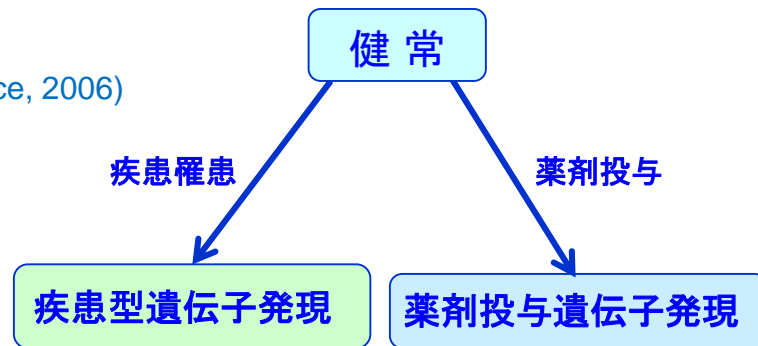


ビッグデータ計算創薬 発現プロファイル比較型創薬・DR

● 薬剤特異的遺伝子発現

— CMAP(Connectivity Map) (Lamb J, science, 2006)

- 薬剤投与による遺伝子発現プロファイル変化
- 米国 ブロード研究所,1309化合物,
5種類のがんの培養細胞
約7000 遺伝子発現プロファイル
- シグネチャ (“刻印”) 差別的発現遺伝子代表群
- DB利用：シグネチャを「問合せ」として投入
類似性の高い順に化合物を提示
- 最近はLINCSデータベース：100万種の薬剤特異的発現DBが存在



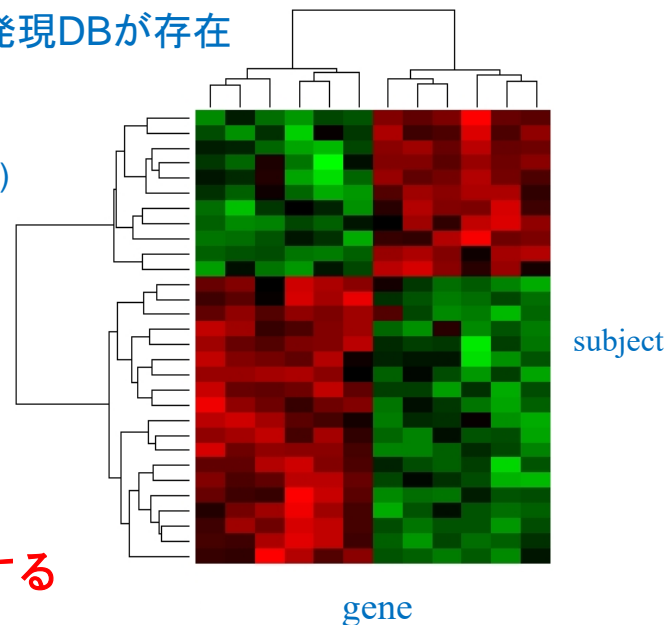
● 疾病特異的遺伝子発現

— GEO (Gene Expression Omnibus) (Barrett T, 2007)

- 疾病罹患時の遺伝子発現プロファイルの変化
- 米国NCBI作成・運用 2万5千実験,
70万プロファイル (欧州 ArrayExpress)
- EBIが作成、サンプル数同程度

基礎には分子ネットワークの疾病/薬剤特異的变化
遺伝子発現プロファイル変化

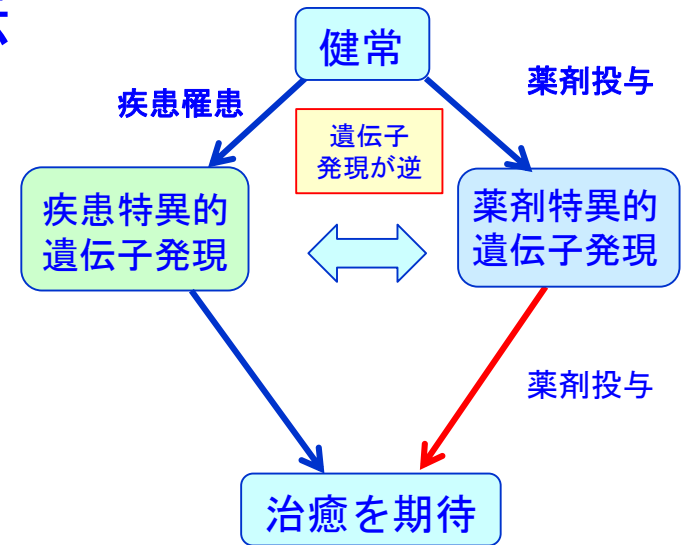
≈ 分子ネットワーク活動構造変化を反映する



遺伝子発現プロファイルによる有効性予測

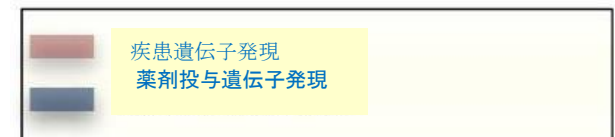
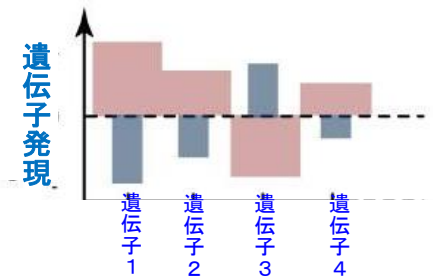
- 遺伝子発現シグネチャ逆位法

- Signature revision法 (Irio, 2012)
- 疾患によって**健常状態から変異**
「疾患特異的遺伝子発現プロファイル」
- これに**薬剤投与の変化を起こす**
「薬剤特異的遺伝子発現プロファイル」
- **両者のパターンが負に相関する**
- ノンパラメトリックな相関尺度で評価



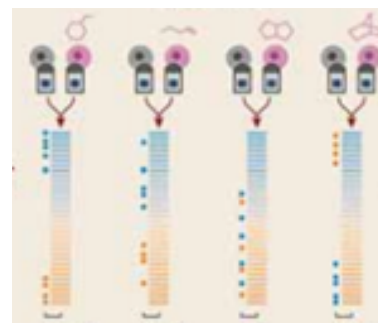
- 効果がお互いに打ち消すなら**有効性が期待される**

- 例：炎症性腸疾患に抗痙攣剤(topiramate),
- 骨格筋委縮にウルソール酸



リストは疾患遺伝子発現の差別的発現順序
横の点は薬剤遺伝子発現のシグネチャ

青は発現が**上昇**した遺伝子
赤は発現が**下降**した遺伝子



強正 弱正 弱負 強負

遺伝子発現プロファイルによる毒性予測

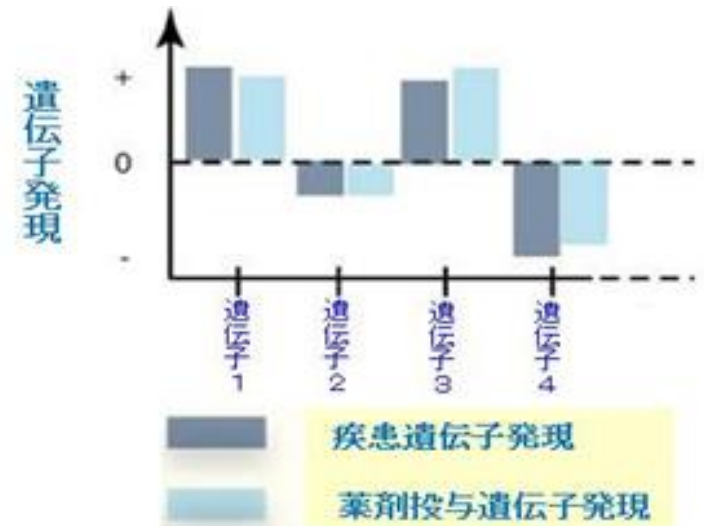
- 連座法 guilt-by-association :

- **薬剤－疾患間 副作用予測**

- 薬剤特異的シグネチャと
- 疾患特異的シグネチャが
- ノンパラメトリック相関 正
- **毒性・副作用の予測**

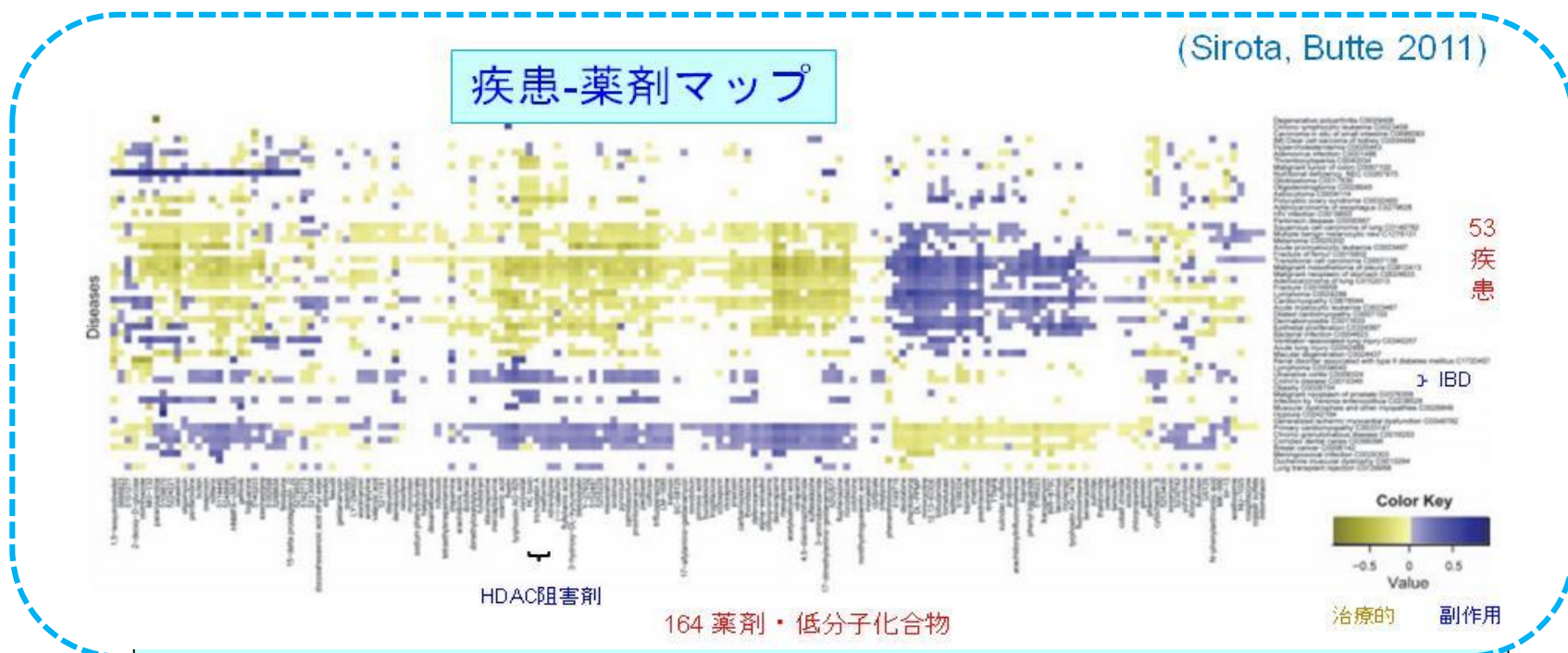
- **薬剤－薬剤間**

- 薬剤ネットワークからのDR
- Connectivity map から薬剤特異的遺伝子発現の薬剤間の親近性をノンパラメトリック親近性尺度 (GSEA)で評価
- この類似性のもとに薬剤ネットワーク構築
- 近隣解析によりDR
- 例：抗マラリア剤をクローン病に適応



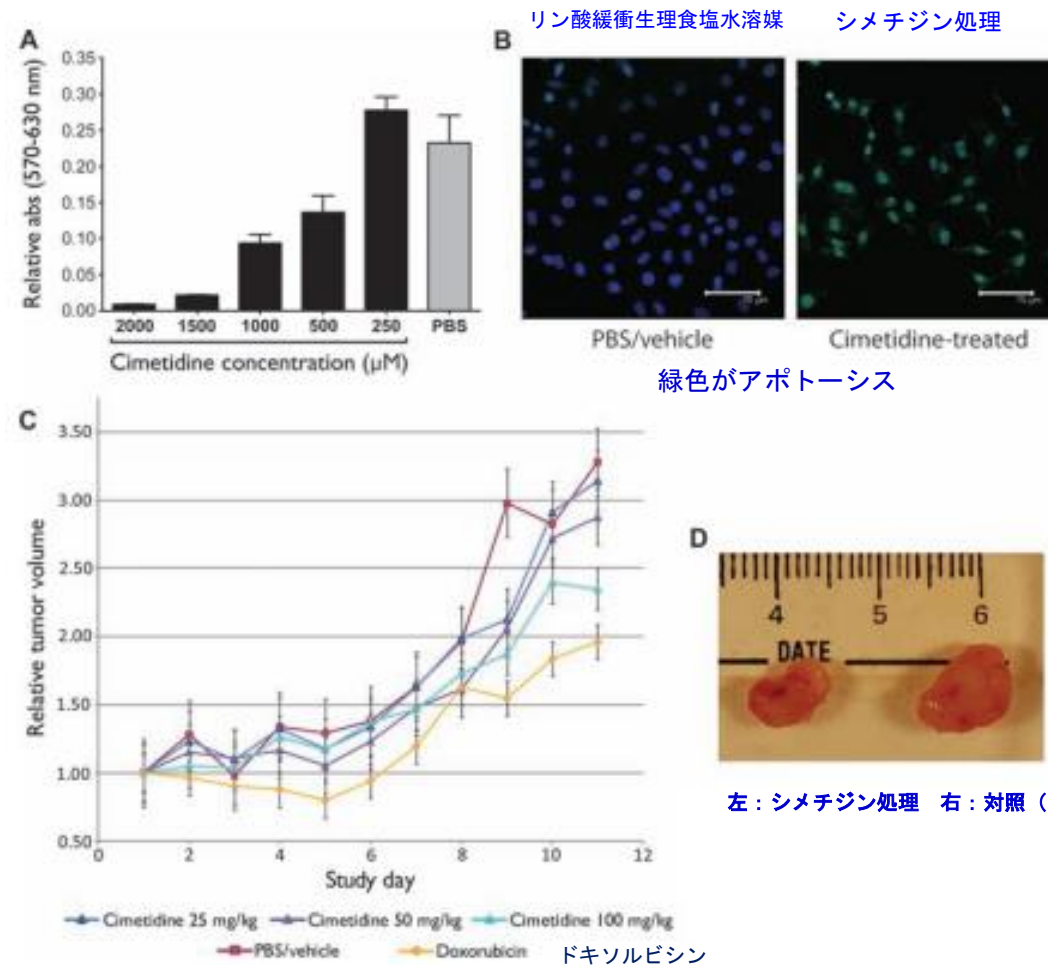
遺伝子発現プロファイルによる 薬剤の有効性・毒性の俯瞰

- 疾患罹患時と薬剤投与時の遺伝子発現プロファイルの
正・負のパターン相関性を基準
- 〈疾患－薬剤〉の「有効性・毒性」予測俯瞰図



動物実験での実証

シメチジン(cimetidine:ヒスタミンH2受容体拮抗薬) →肺腺癌(LA)に有効か
 予測スコア -0.088 であったが gefinitib (イレッサ) の-0.075より高い



遺伝子発現プロファイルによる疾患-薬剤ネットワーク

遺伝子発現プロファイルの類似性を相関係数、ESによってリンク (Hu, Agarwal, 2009)

疾患-疾患、薬剤-薬剤、疾患-薬剤のネットワークを発現プロファイルより構成

疾患 (disease-disease) 645 組
 疾患-薬 (disease-drug) 5008 組
 薬 - 薬 (drug-drug) 164,374 組

結果

①疾患-疾患NWの60%はMeSH (既知体系)

その他は分子レベル疾患分類学
 遺伝子発現の類似性による疾患体系

②主な発見

<疾患 - 疾患>

HSP (Hereditary Spastic Paraplegia

(遺伝性痙攣性対麻痺)

⇒bipolar 双極性障害

Solar keratosis 日光性角化症

⇒ cancer(squamous)

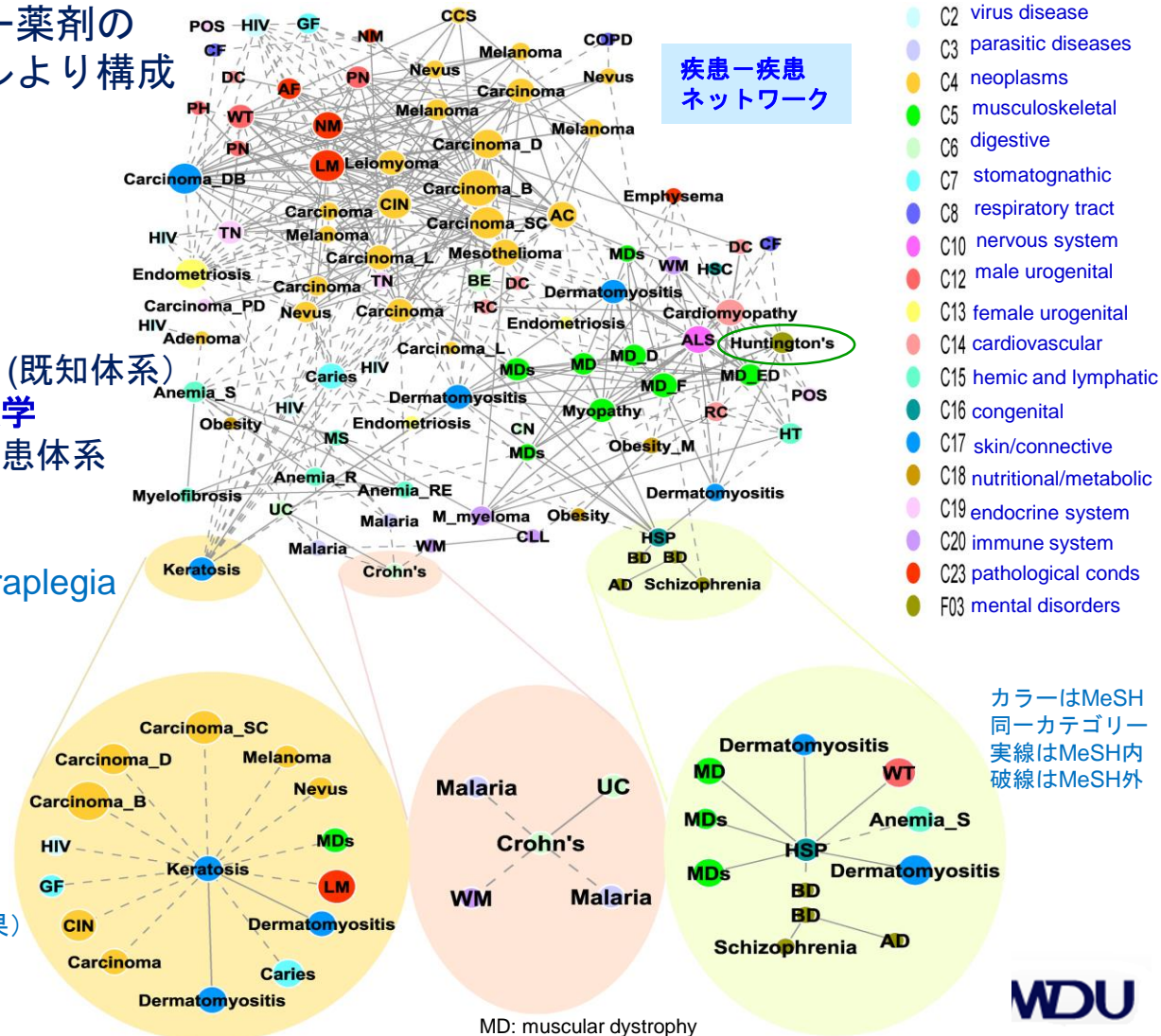
<疾患 - 薬>

有効性: マラリア治療薬

⇒ Crohn's disease

(ベトナム経験: クロウン病罹患保護効果)

ハンチントン病に種々の薬剤



遺伝子発現プロファイルによる 疾患-薬剤ネットワーク

疾患-薬剤ネットワーク

(Disease-drug network: 右図)

橙色 49 疾患

緑色 216 薬剤

(全体で906対 疾患-薬剤結合)

Tamoxifen (乳がんのホルモン療法薬)

有効性 (破線: 負の値をもっている)

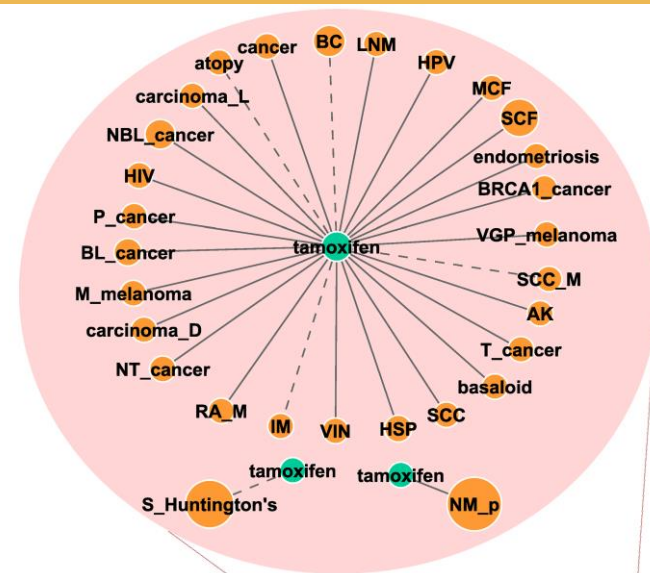
⇒ アトピー,
マスト細胞分泌抑制、
アレルギー抑制

⇒ Hunting病に多数のDR薬

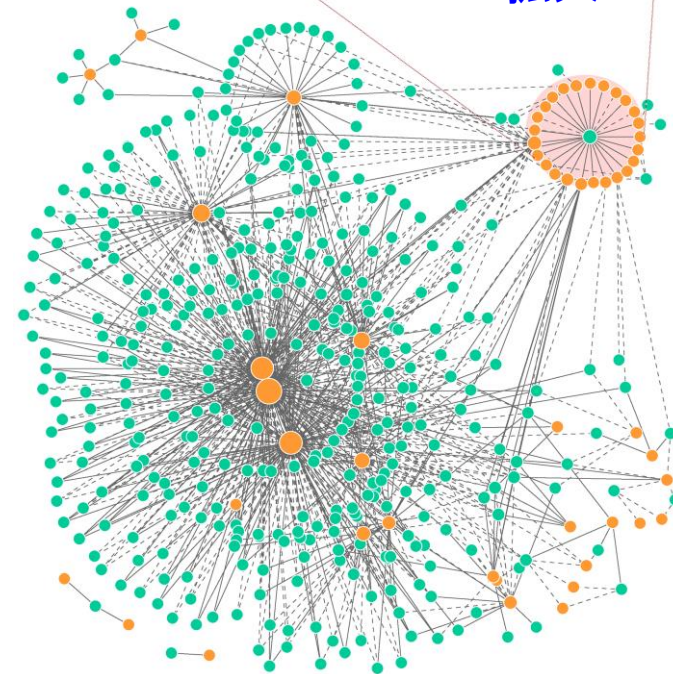
副作用 (実線: 正の値をもっている)

副作用の予測 ⇒ 多くのがん

⇒ 発癌性



拡大

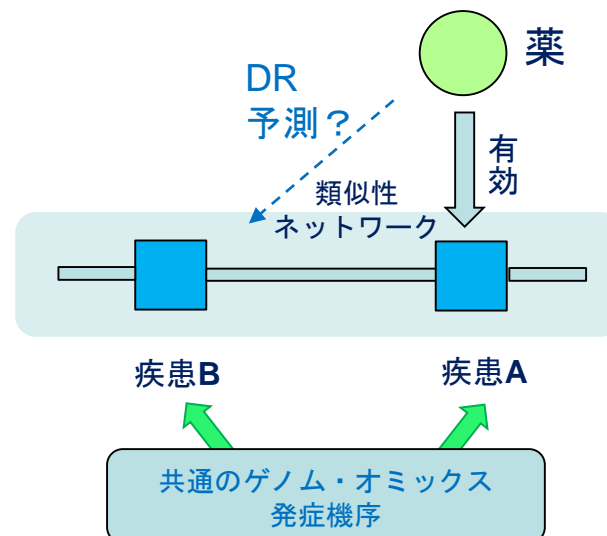


疾患-薬剤ネットワーク

ビッグデータ創薬/DRの基本原理2

疾患ネットワーク準拠創薬/DR

- 従来の疾患体系 nosology
 - Linne以降300年に亘って表現型による疾病分類
 - 臓器別・病理形態学別の疾患分類学
- ゲノム・オミックスレベルでの発症機構での疾患分類
 - 発症の**内在的 (intrinsic)機構の類似性**を**基準に**疾患ネットワーク（疾患マップ）をつくる
 - ゲノム・オミックスによる内在的疾患機序の概念が基礎

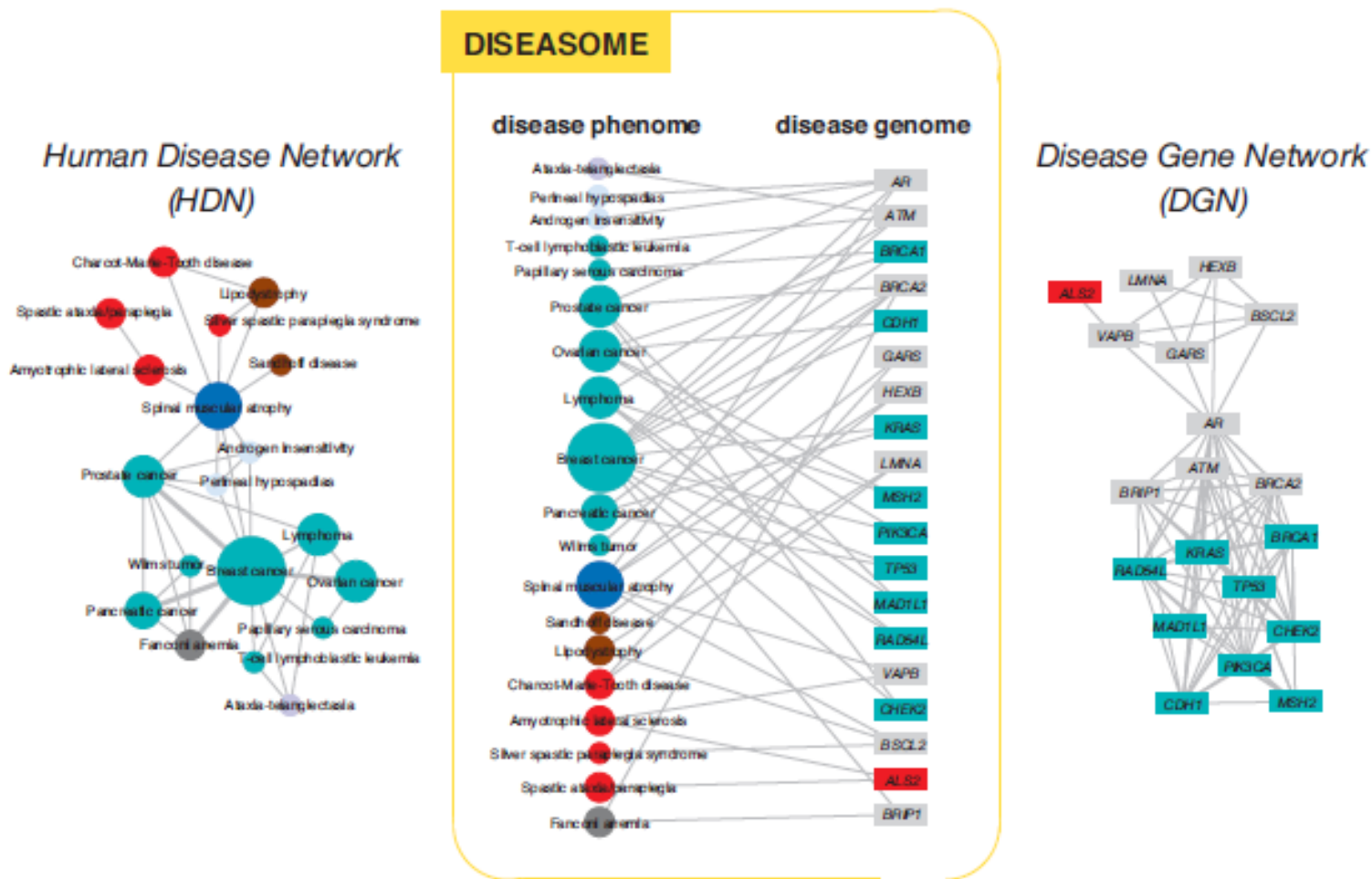


第1世代型

Diseasomeと疾患遺伝子

- **OMIM**から 1,284 疾患と 1,777 疾患遺伝子を抽出
- **ヒト疾患ネットワーク (HDN)**
 - 867疾患は他疾患へリンクを持つ 細胞型や器官に非依存
 - 516疾患が巨大クラスターを形成
 - 大腸がん、乳がんがハブ形成
 - がんはP53 やPTENなどにより最結合疾患 がんなどは後天的変異
 - 疾患を網羅的に見る見方：臓器や病理形態学に非依存
 - リンネ（12疾患群分類）以来300年続いた分類学を越える
- **疾患遺伝子ネットワーク (DGN)**
 - 1377遺伝子は他の遺伝子へ結合
 - 903遺伝子が巨大クラスター
 - P53がハブ
- ランダム化した疾患/遺伝子ネットワークに比べ
 - 巨大クラスターのサイズが有意に小さい
- **疾患遺伝子は機能的なモジュール構造**
 - 同じモジュールに属する遺伝子は相互作用し
 - 同一の組織で共発現し、同じ**GO**（遺伝子オントロジー）を持つ

疾患ネットワーク Diseasome



1つ以上の疾患関連遺伝子を共有する疾患

1つ以上の疾患を共有する疾患関連遺伝子

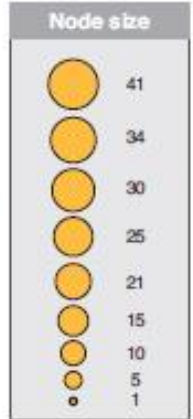
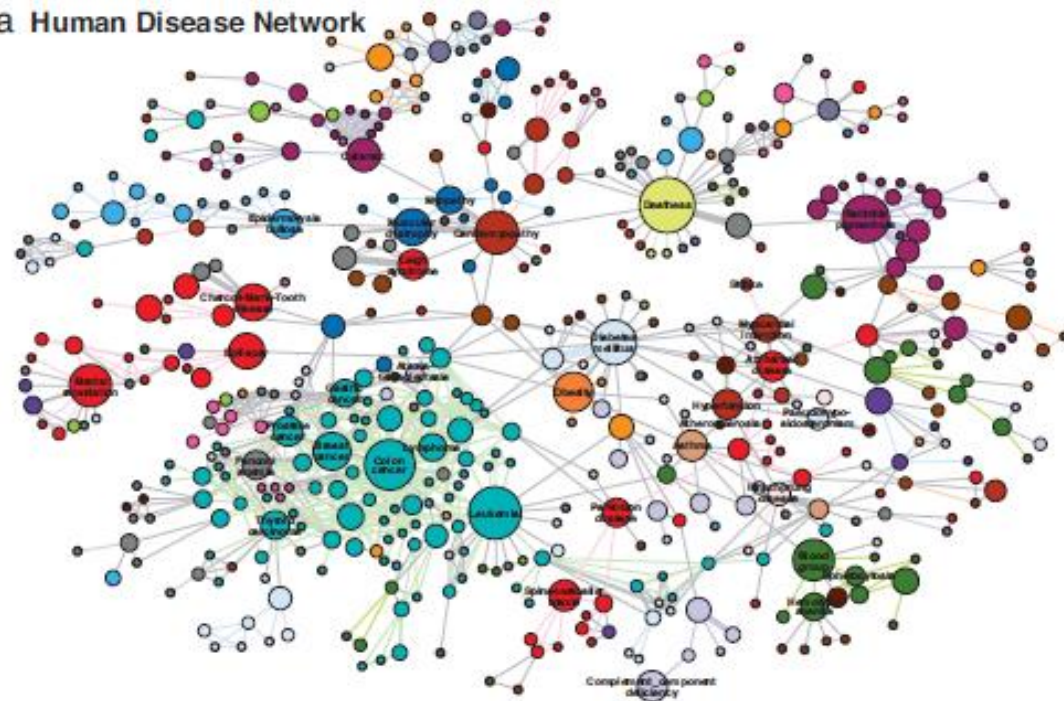
Kwang-Il Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi The human disease network PNAS, 2007

疾患 ネットワーク (HDN)

Nodeの直径
疾患に関与している原因
遺伝子の数に比例

リンクの太さ
疾患間で共有している
原因遺伝子の数

a Human Disease Network

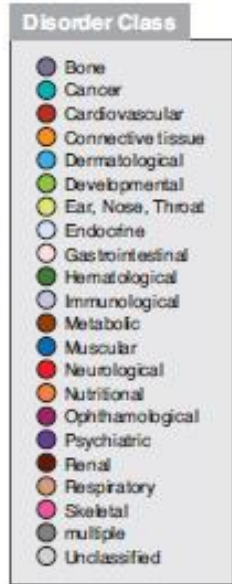
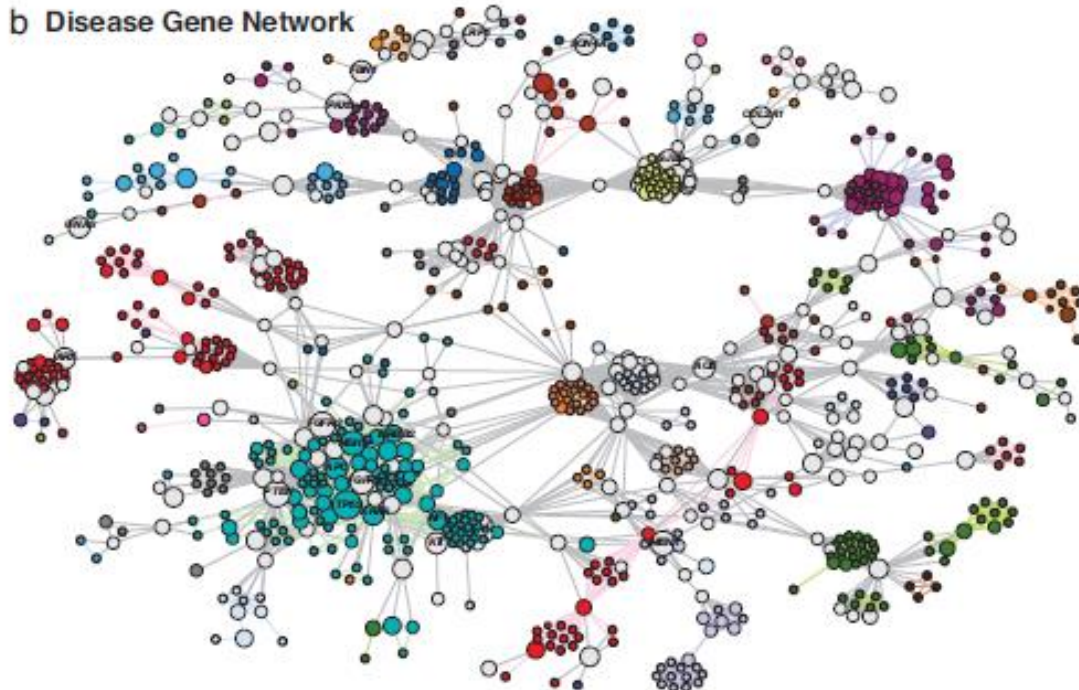


疾患遺伝子 ネットワーク (DGN)

Nodeの直径
その遺伝子を原因にして
いる疾患の数に比例

2つ以上の疾患に関与し
ていると明灰色の遺伝子
ノード

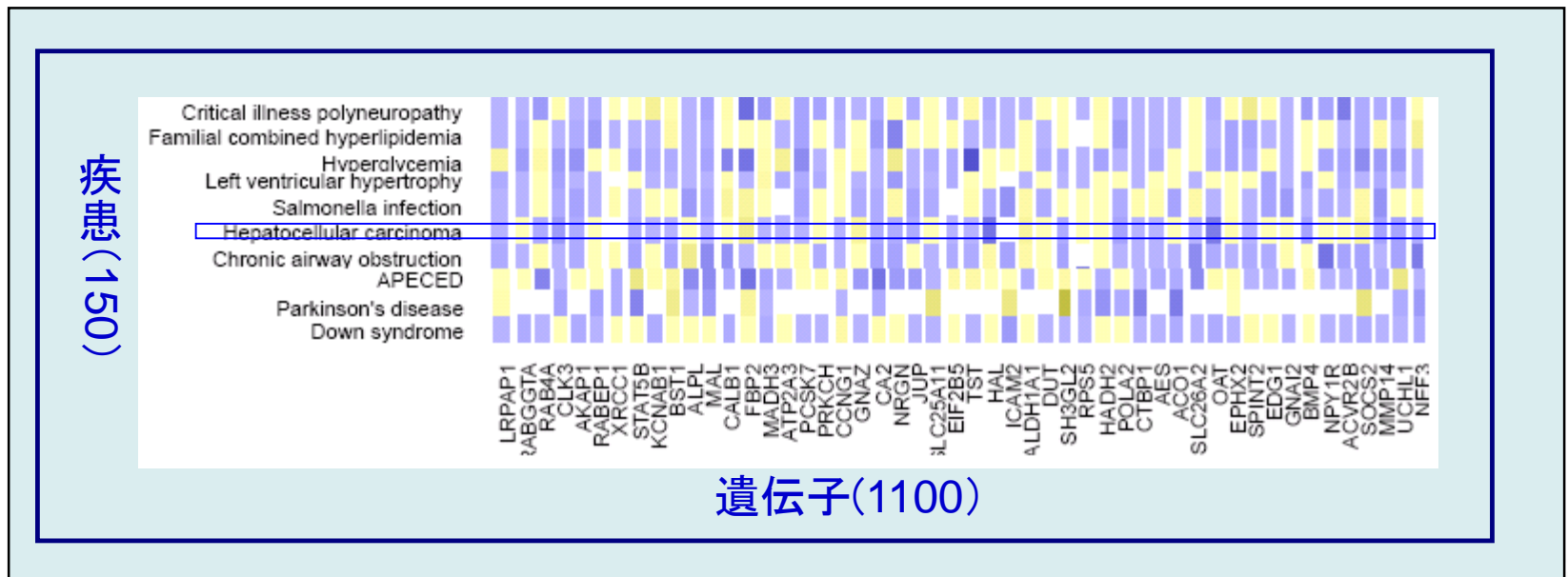
b Disease Gene Network



第2世代型

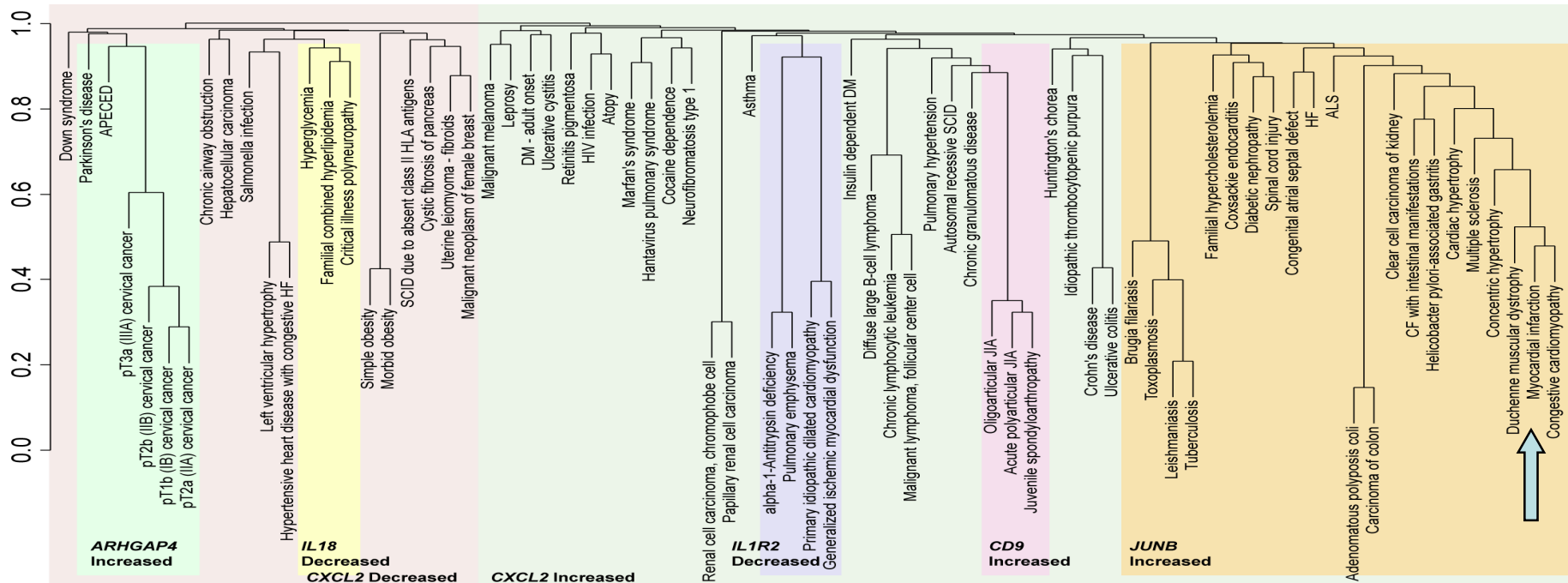
GENOMED (A. Butte et al)

- 遺伝子発現DBのGEO (Gene Expression Omnibus) 利用
 - 約20万のサンプル
- 疾患名は注釈文より用語集UMLSを用いて抽出
- 疾患ごとに多数の遺伝子発現パターンを平均化




Gene-Expression Nosology of Medicine

- 疾患を平均遺伝子発現パターンよりクラスター分類
 - 臓器別疾患分類では予想できない疾患間の親近性
 - 分類項目はサイトカインの遺伝子発現と相関
 - 疾患の再体系化に基づいた医薬の repositioning
- さらに656種類の臨床検査を結合した分析
- 心筋梗塞・デュシャンヌ型筋ジストロフィーに近い



疾患ネットワークから 生体分子ネットワークを基盤とする創薬/DRへ

疾患ネットワークに準拠した創薬/DR

- 
- 疾患のゲノム・オミックス機序に基づいている→内因的機序を考慮した点で評価
 - しかし「疾患関連遺伝子」と「薬剤」の相互作用の関係が明示的ではない

集成的な創薬/DRのフレームワーク

- 生命分子ネットワークを作用の<場>とする
- 薬剤の足場である<標的分子>と
疾患の足場である<疾患関連分子>との
<相互作用>を基礎とする枠組み

3層の生体・薬剤のネットワーク間の関係図式

疾患ネットワーク

プロファイル比較型
創薬/DR

薬剤ネットワーク

薬剤Cは疾患Dに薬効

疾患D

薬剤C

現象

機構

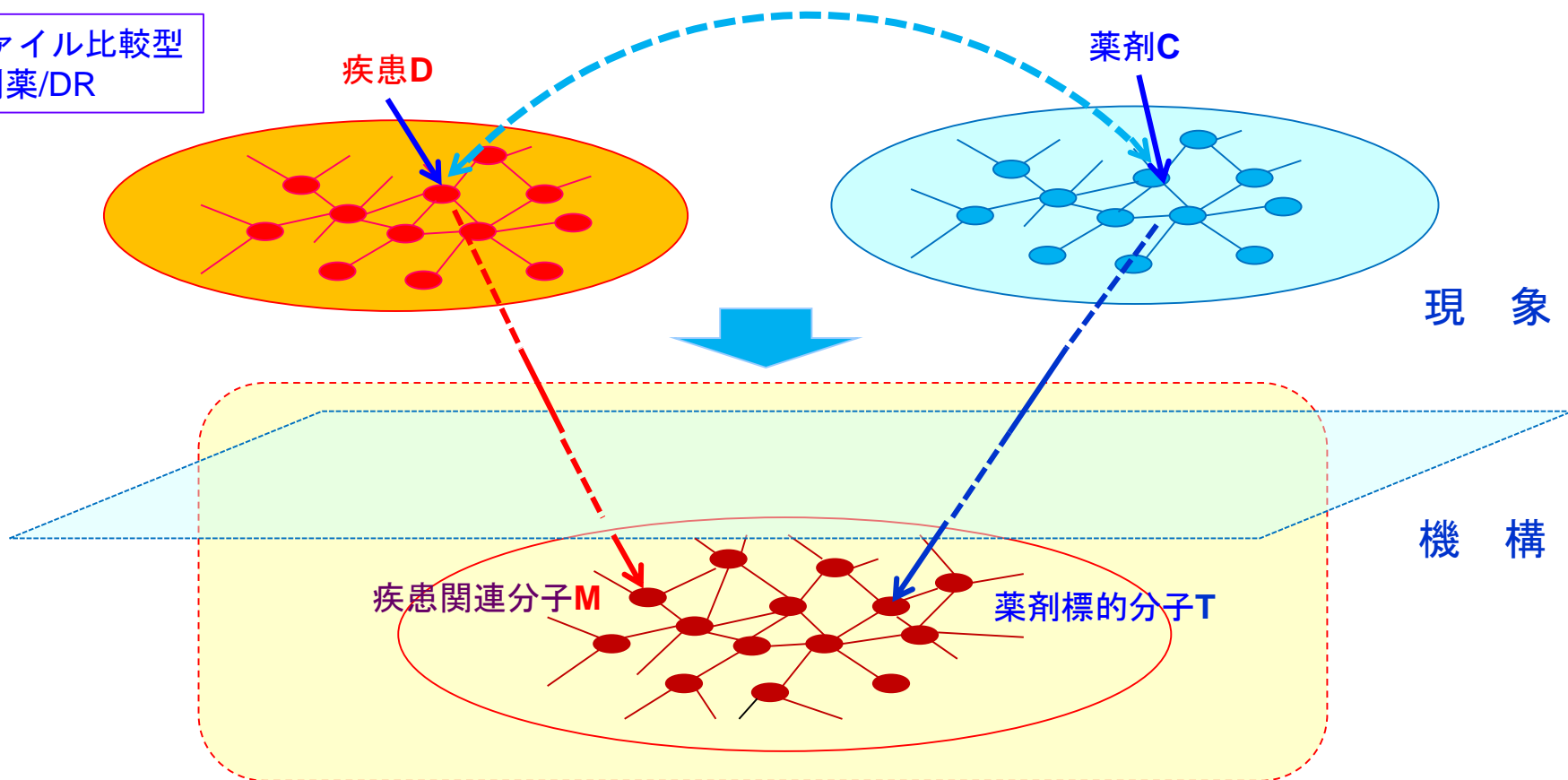
疾患関連分子M

薬剤標的分子T

分子ネットワーク型
創薬/DR

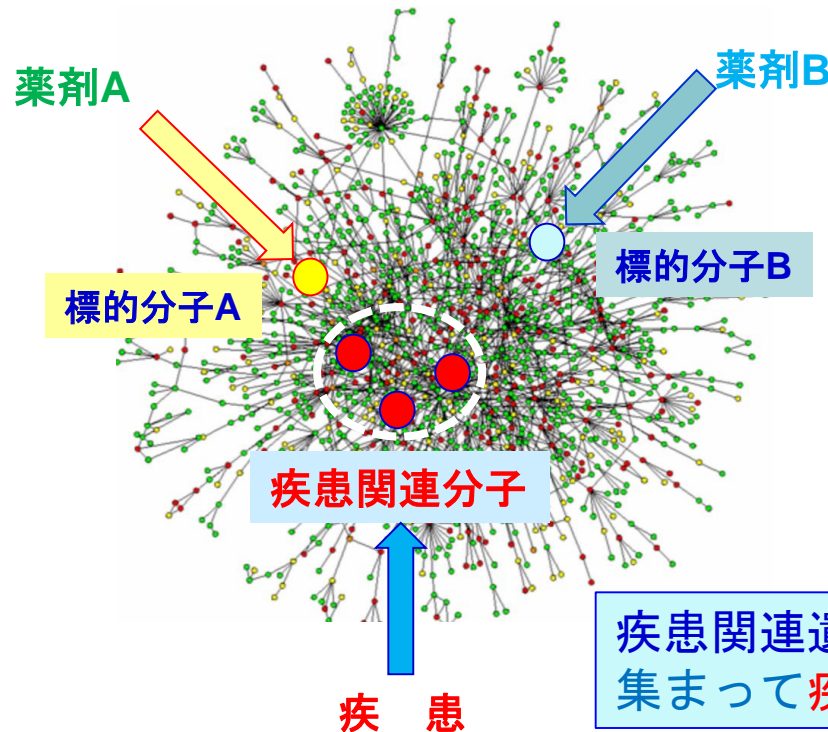
生体分子ネットワーク的マイクロ対応

生命システム



標的分子や疾患関連分子の タンパク質相互作用ネットワーク (PPIN)

- 薬剤ネットワークと疾患ネットワークの基盤：生体分子ネットワーク
- タンパク質相互作用ネットワーク (PPIN) での創薬/DR戦略
- PPIネットワーク場を基礎にして距離 (近接性) を検討
- 薬 剤：薬剤の標的分子 (タンパク質) によって PPI場と繋がる
- 疾 患：疾患特異的発現遺伝子を疾患関連分子 (タンパク質) へ翻訳、
- PPIN場内での薬剤 (標的分子) と疾患 (疾患関連遺伝子) の「代理人」の距離・近接性を基準に、薬理作用のインパクト力を評価

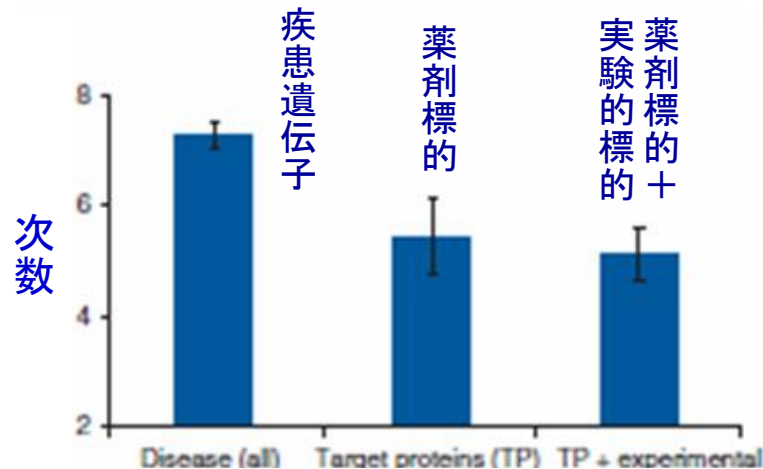
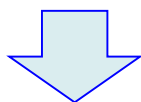


タンパク質相互作用
ネットワーク (PPIN)

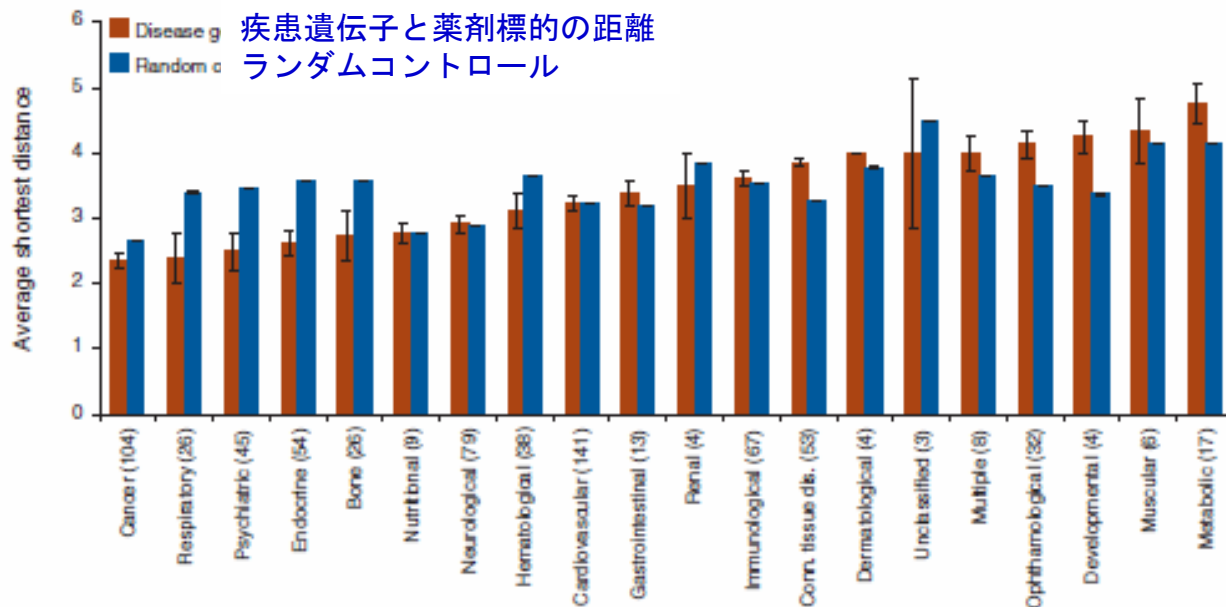
疾患関連遺伝子はネットワーク上の近傍に
集まって疾患モジュールを形成する

標的タンパク質と疾患遺伝子の距離

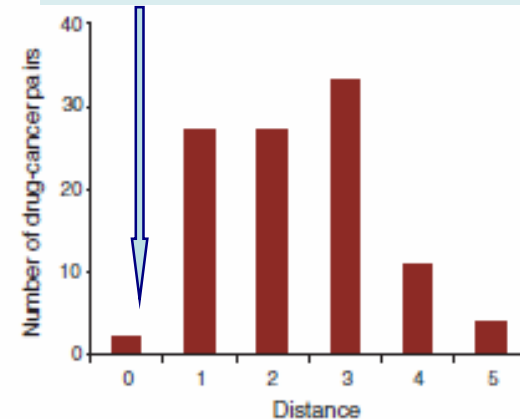
薬剤標的タンパク質と疾患関連タンパク質の間の距離：**2~4リンク**



Yildirim M A, et al, NATURE Biotechnology 2009



抗がん剤の場合
疾患遺伝子と距離0の標的

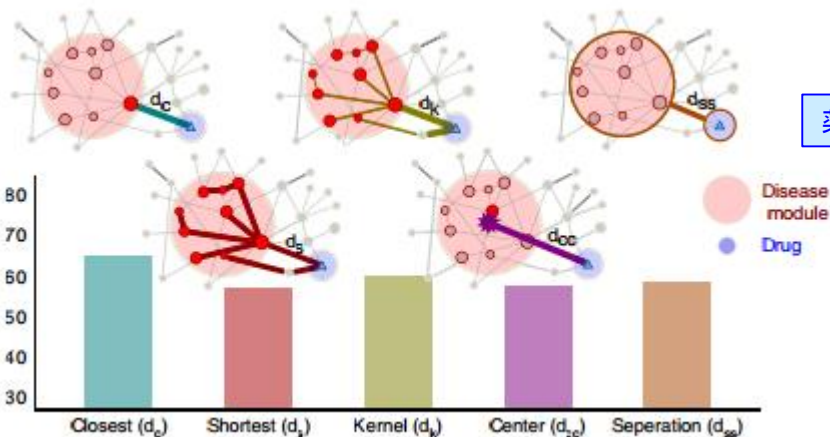
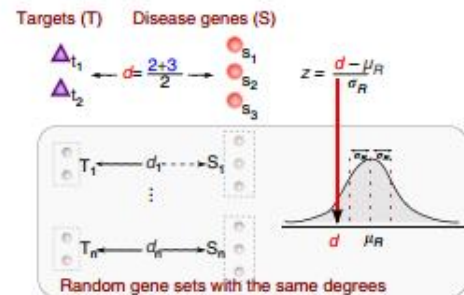
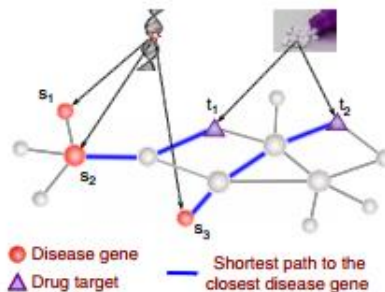


抗がん剤の標的分子と疾患遺伝子の間に距離

タンパク質相互作用ネットワークでの 近接性によるDR

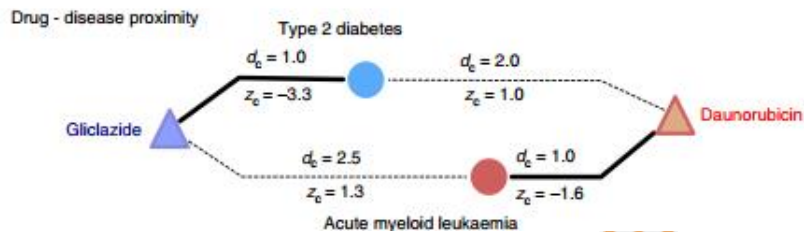
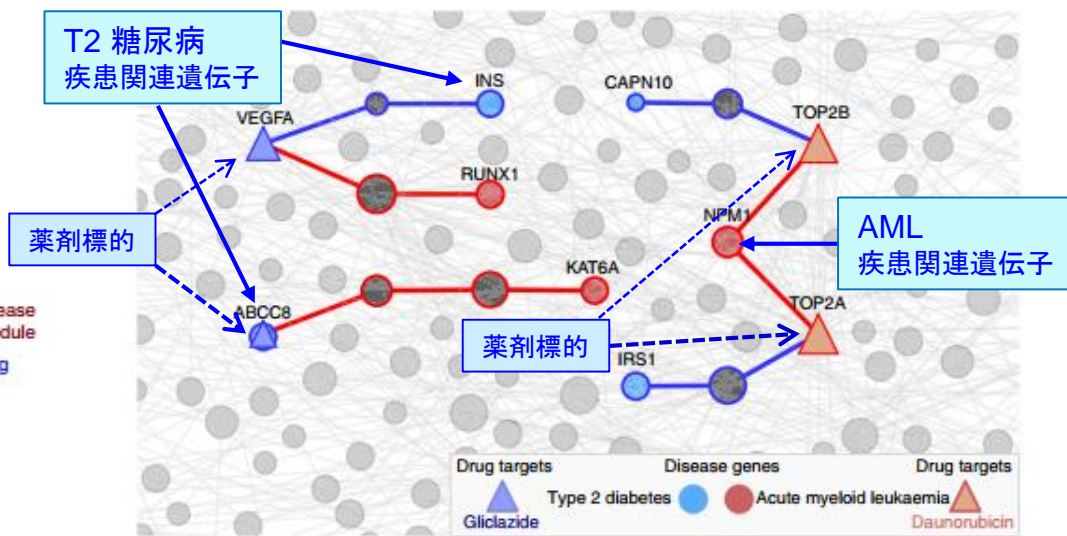
相対近接指標 d_c :

- ①最近接の疾病関連分子との最短経路長の平均
- ②同じサイズで度数の分布より近接指標を計算して規格化⇒zスコア
($z < -0.15 \Rightarrow$ 近接)
- ②様々な近接指標の中ではclosest measure d_c が一番薬効を予測する



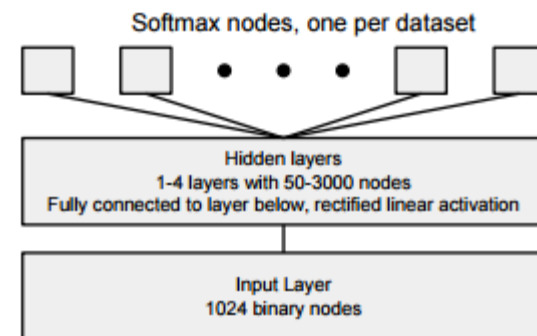
大半の薬剤は標的と疾患関連分子
2リンク離れている

(Gunev, Barabasi, 2016, Nat. Com)



Deep learning : 創薬からの注目

- 創薬を巡る状況
 - 平均14年、約2000億円 (\$1.7 B) の費用
 - 市場化された新薬の減少
 - 創薬に費やす期間・コストを低減したい
- Kaggle (データサイエンス競技会)にMerck社が出題
Molecular Activity Challenge (2012).
 - 15データセットから異なった分子の生物学的活動を予測するモデルの開発コンテスト
 - 勝利したモデルは深層学習 deep learning を用いたモデル
- Google in collaboration with Stanford (2015)
 - Stanford 大学の Pande 研究室と共同研究
バーチャルドラッグスクリーニングに対する
deep learningによるツール開発
"Massively Multitask Networks for Drug
Discovery"



Artificial Intelligence (AI) と創薬

- 標的分子選択と妥当性検証
 - 適切な分子標的の選択
- Virtual screening と選択 ←
 - 適切な化合物に対するクラス判定
 - 研究例：ChEMBLに対するdeep learning
 - 13 M 化合物特徴量 (ECFP12), 1.3M 化合物, 5k 薬剤標的
 - Ligand-based 標的予測, 7種の予測法とAUC比較
 - Deep learningは、SVM, k-最近隣法, logistic回帰より優位
 - DLで構造活性相関を学習する
 - 特徴量の抽出、薬理機序への理解
 - リード最適化
- システム薬理学
 - ネットワーク病態学よりの創薬戦略
 - 他のシステムへの影響(毒性, 副作用)

Pharmacophoreの抽出

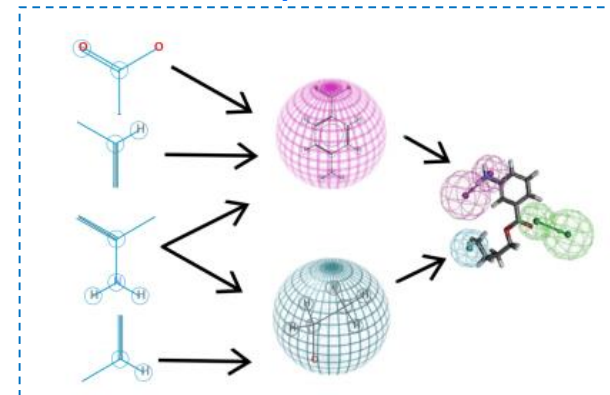


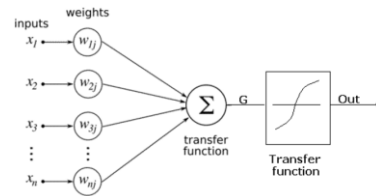
Figure . Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

Deep Learning による 人工知能革命

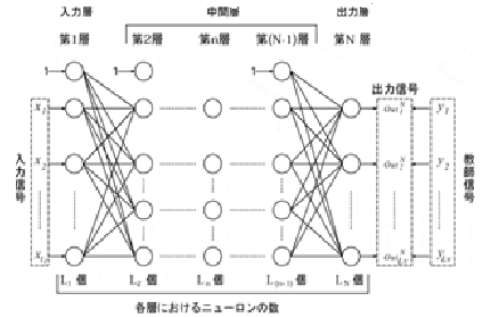
- 機械学習のこれまでの限界

- 「教師あり学習」

- 分類対象の特徴と正解を与え学習機械 (AI) を構築



神経情報素子



多層ニューロネットワーク

- Deep Learningの革命性

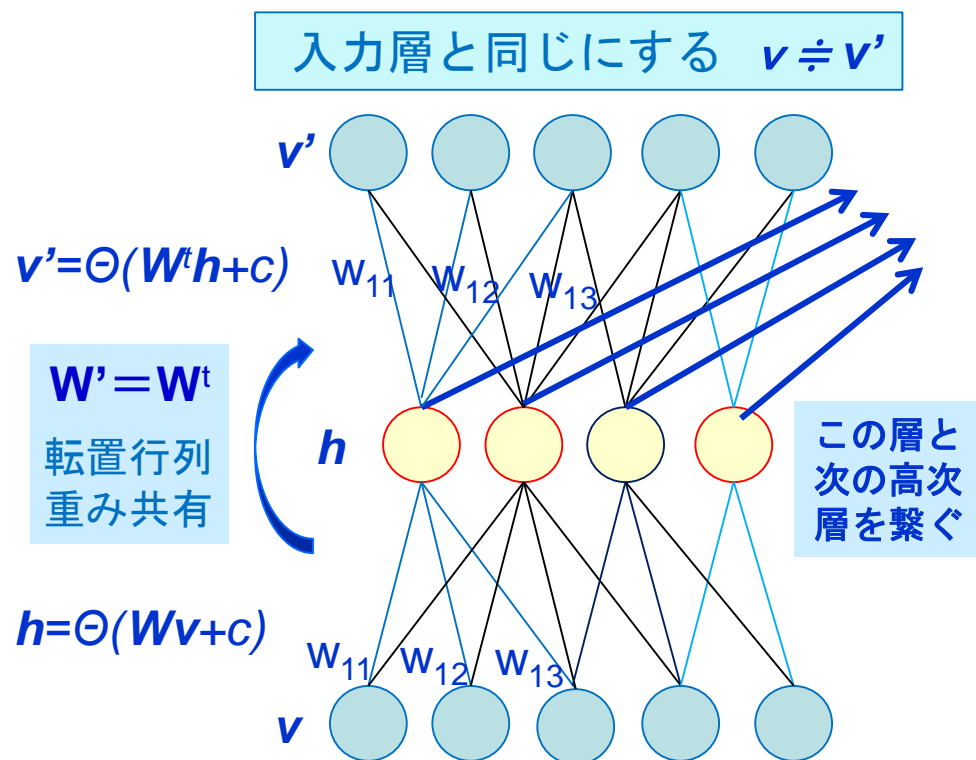
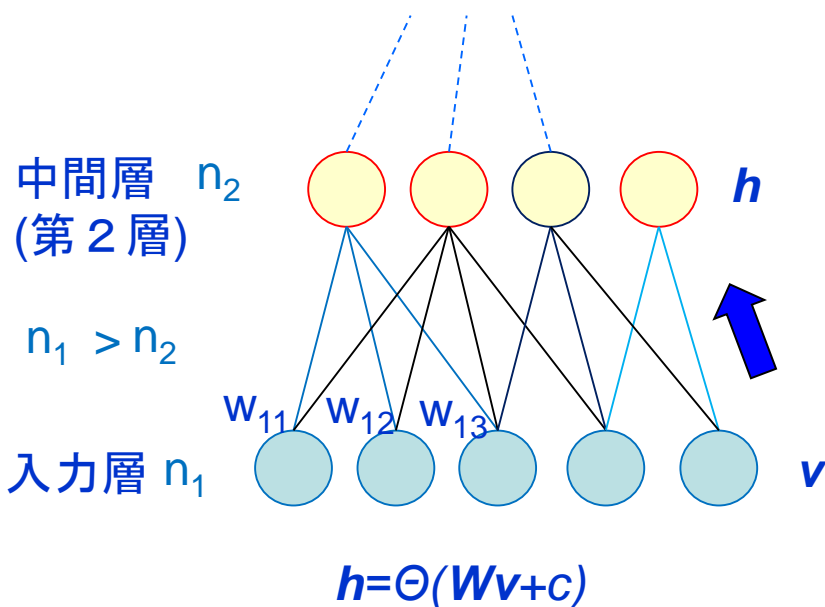
- 「教師なし学習」

- 対象の特徴表現や対象の高次特徴量を自ら学ぶ



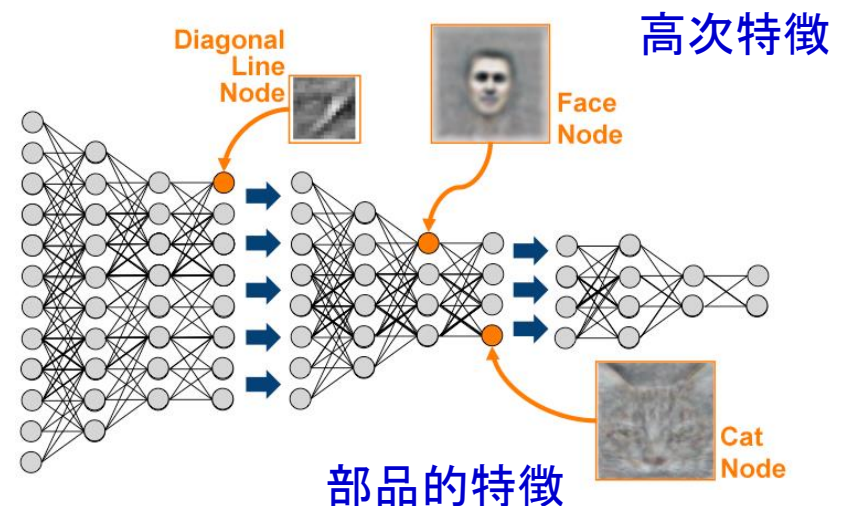
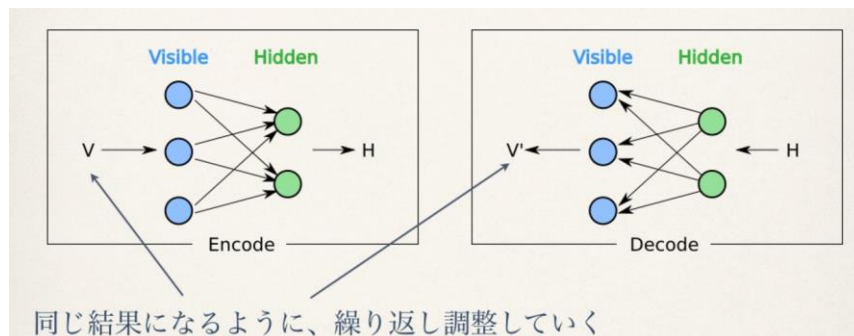
DLの革命点 Autoencoder 1

- 対象に固有な**内在的特徴**を学ぶ**自己符号化の原理**
- 格段ごとに入力の少ない中間層を入力へ逆投影して復元できるか
- 次元を圧縮され可及的に復元する ($1000_{\text{nodes}} \Rightarrow 100_{\text{nodes}} = ? \Rightarrow 1000_{\text{nodes}}$)
 - 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



DLの革命点 Autoencoder 3

- 各層ごとに自己符号化を行うので**何層でも組める**
 - 各層間で「自己符号化」の積上げ (autoencoder stack)
- 第一層で学習した特徴量を使って次の階層を作るので**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。
- 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
 - 内在的な特徴量による構造的な理解
- 人間の「思考の枠組み」を超えた正解の低次
 - 「**アルファGo**」が定石にない手で碁の名人に勝つ



Deep Learningによる創薬・DR

1) 生体ネットワーク (PPIN) 特徴量の抽出

- タンパク質相互作用ネットワーク(PPIN)のNW結合を学習し**特徴表現** (特徴NW基底) を出力。
- 学習集合を部分ネットワークの集合から決める
- ノードを起点とした素NWでPPIN全体を覆う集合

2) 多層Deep Auto-encoderのDLで学習.

- 特徴的NW基底の「教師無し」学習
- 次元縮約による特徴的NW基底の抽出

3) DL特徴NW基底空間における正例補完

- DrugBankからの正例とその増加 (SMOTE法)

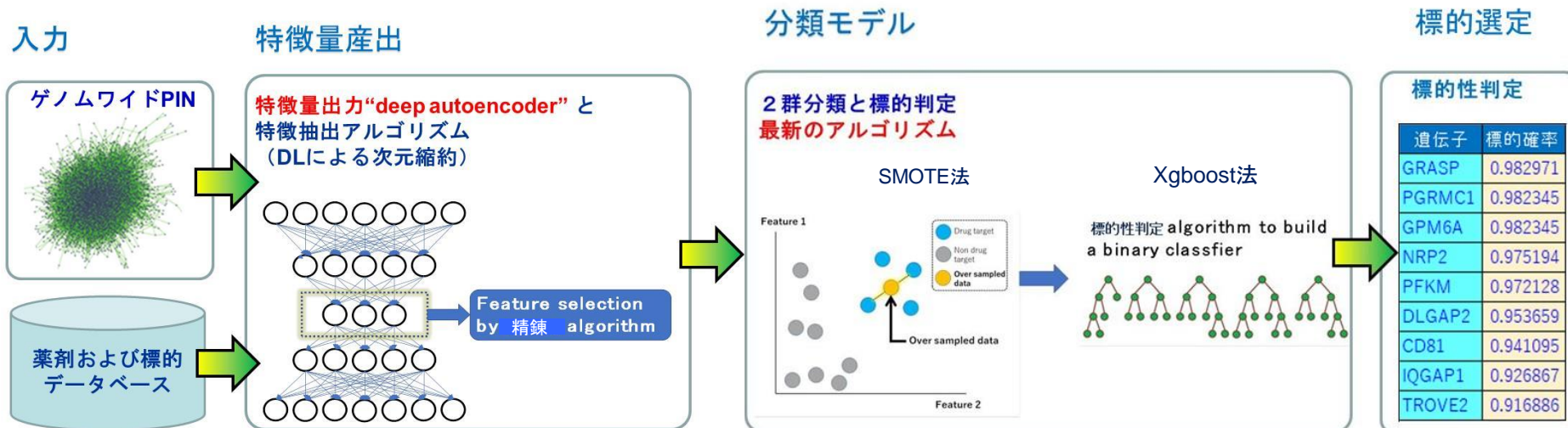
4) DL特徴NW基底量を用いた機械学習分類

- Xgboot法などを用いたDL特徴量からの判別ネットワーク・タンパク質の標的性の判定

Deep Learningによる創薬・DR

分類部 DrugBankを利用した 当該分子を標的とする既製薬剤の探索

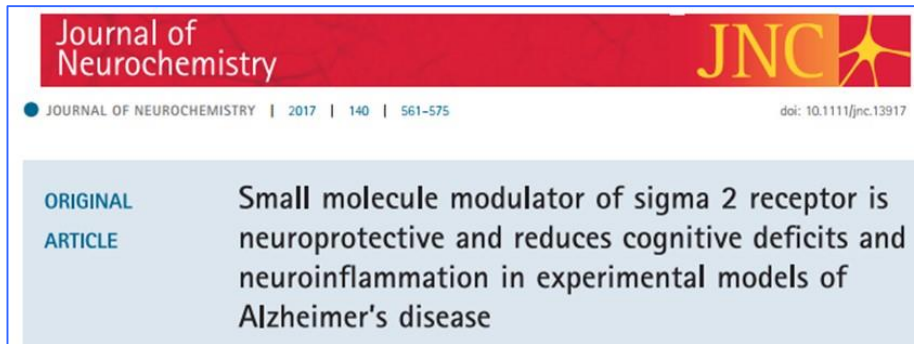
既製薬剤がない→新規薬剤探求（創薬）
既製薬剤がある→DRの検討



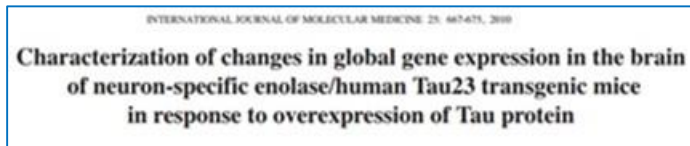
従来の機械学習（Random Forrest）と同じ成果は得られている

実験的研究との付合 1

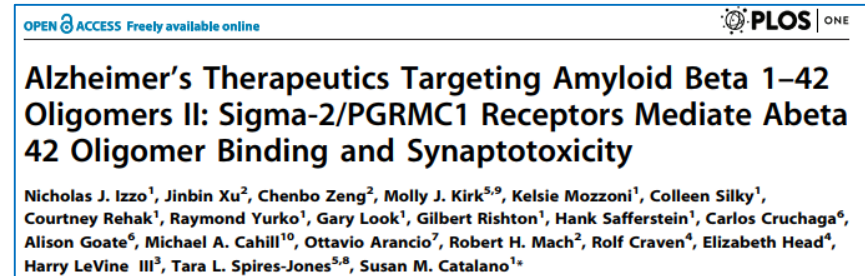
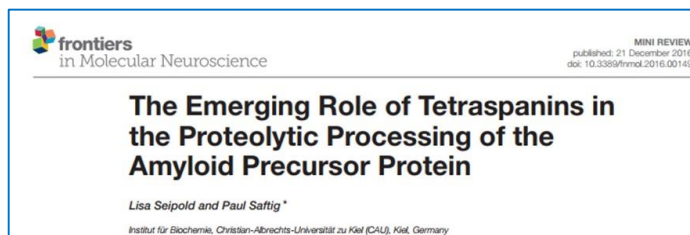
PGCM1 : progesterone receptor membrane 1



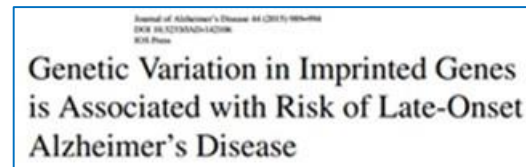
GPM6A : Glycoprotein M6A



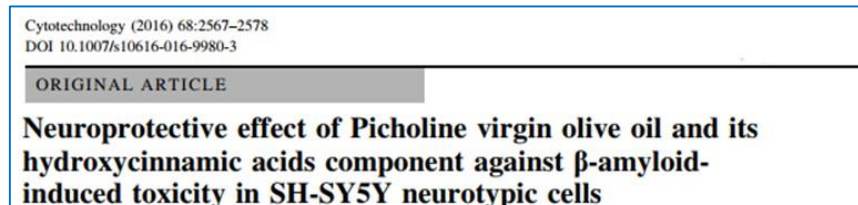
CD81:Tetraspanins family



DLGAP2 : DLG-Associated Protein 2



PFKM: Phosphofruktokinase

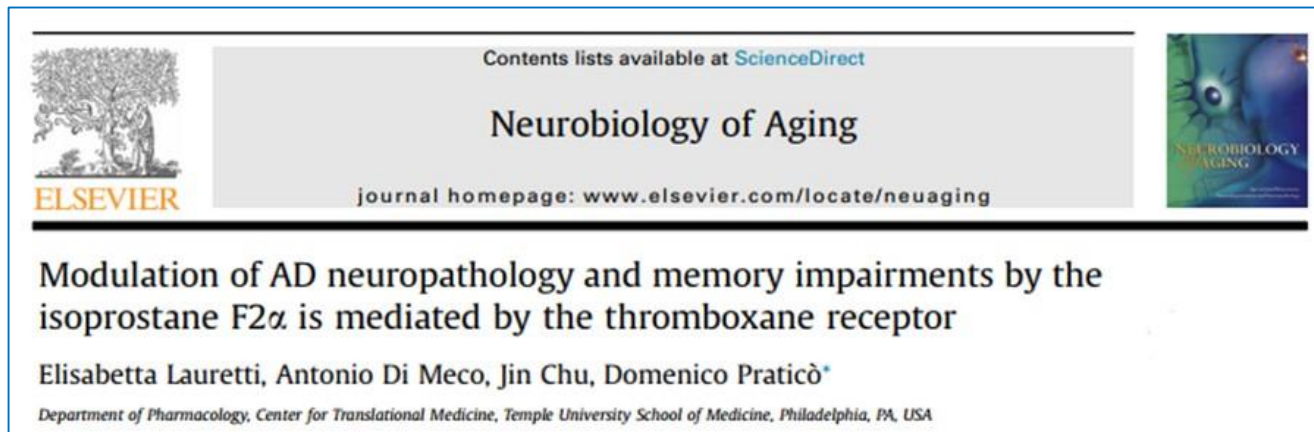


実験的研究との付合 2

WISP-2/CCN5 : WNT1 inducible signaling pathway protein 2



TBXA2R: thromboxane A2 receptor



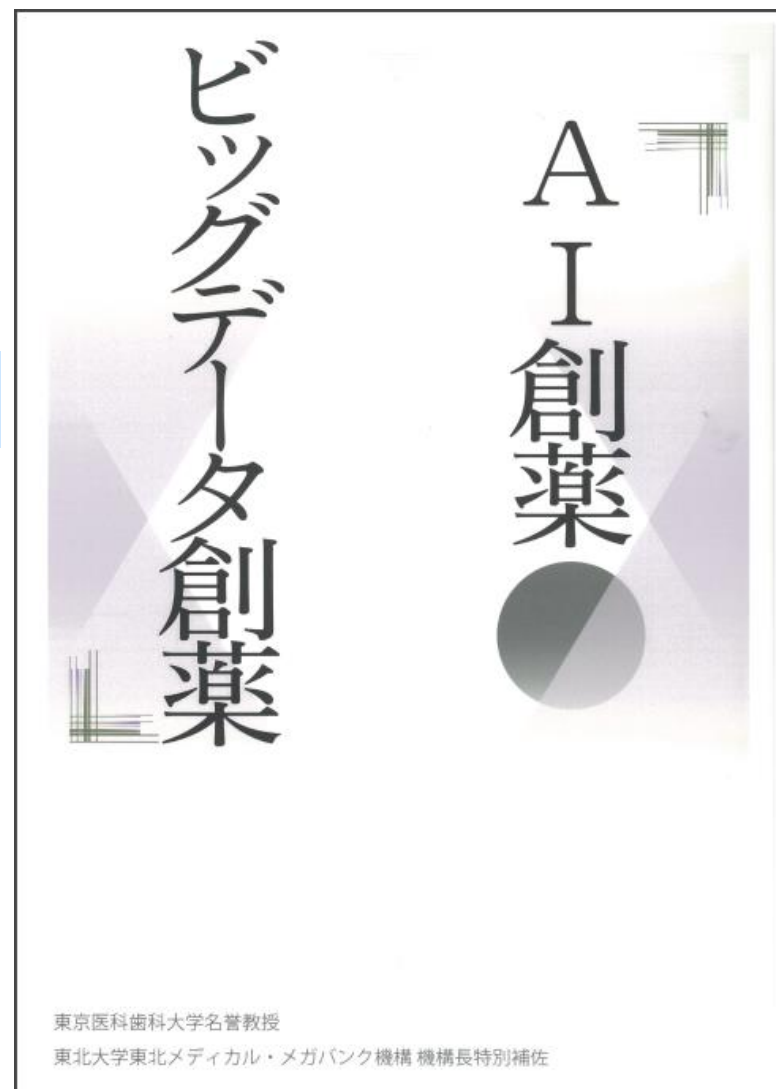
DL型NNへの期待と困難点

- 医療・創薬の応用は大きく期待される
 - 本質的に「教師なし学習」:人間が思いつかない解を提示
 - 現状では、画像分類・解釈と文章理解が優れているので、遺伝子発現プロファイル解析や病態推移の理解への応用
 - 例: ヒトmicrobiomeの分類・階層的表現を得た
 - 6つのがんで遺伝子発現をmiRNAとともに分類した。
 - 異なったMicroarrayを含むがん発現を分類の特徴表現を導き分類した。
 - Convolution ネットワークを使用して画像としての遺伝子発現を分類した。
 - 遺伝子発現プロファイルの自動アノテーション
 - 期待される本質的な寄与
 - 超多次元（生命医学）ネットワークから革新的知の発見
- DL型ニューラルネットは困難点もある
 - 特徴表現を自己学習するが基本的にはBlack Boxで解析が必要
 - 大量のデータを必要とする
 - DL型NNには、ハイパーパラメータが多種類があり、使用に関して選択問題が残る
 - 計算時間が長くコストが大きい。

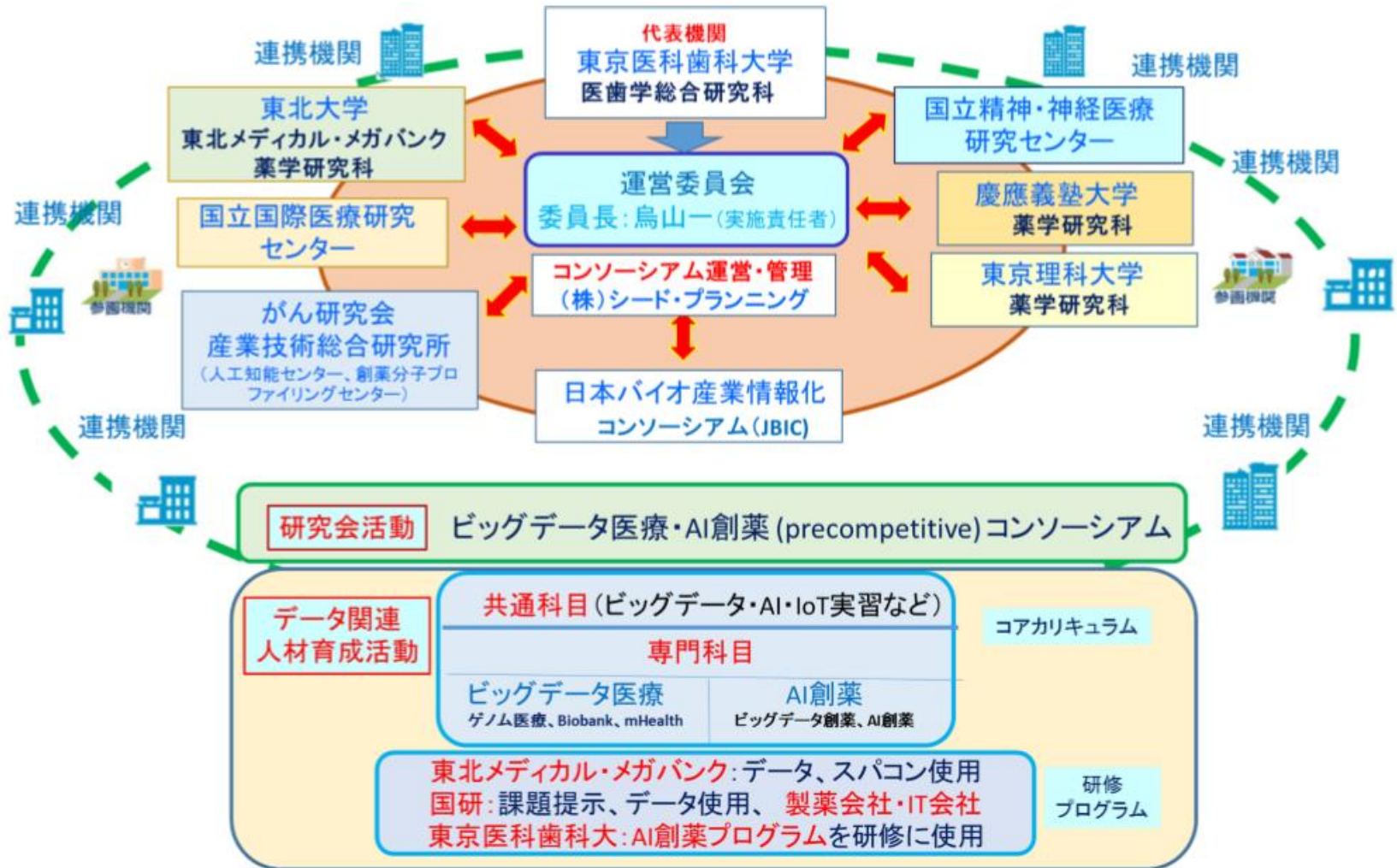
田中 博 著

「AI創薬・ビッグデータ創薬」

薬事日報社 6月19日刊行



ビッグデータ医療・AI創薬コンソーシアム



ご清聴ありがとうございました

