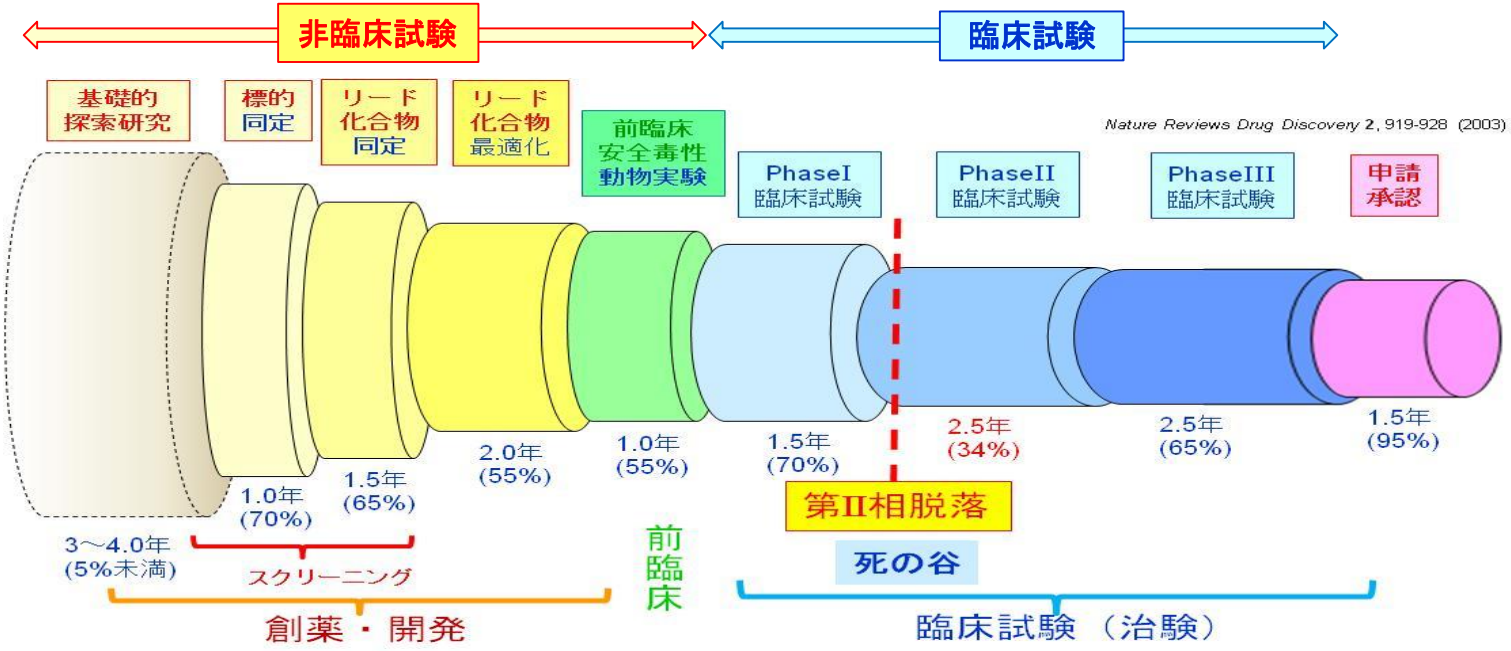
ビッグデータから人工知能を用いて創薬・DRを行う

東京医科歯科大学
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

- 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**
  - **ヒトの薬剤 - 生体関連のビッグデータを使う**

Nature Reviews Drug Discovery **11**, 191-200 (2012)



第一部
ビッグデータ計算創薬・DR

ドラッグ・リポジショニング

薬剤適応拡大

ヒトでの安全性と体内動態が十分に分かっている
既承認薬の標的分子や作用パスウェイなどを、体系的・論理的・網羅的に解析することにより**新しい薬理効果**を発見し、その薬を**別の疾患治療薬として開発する**創薬戦略

利　点

（1）既承認薬なので、ヒトでの安全性や体内動態などが既知で臨床試験で予想外の副作用や体内動態の問題により開発が失敗するリスクが少なく**開発の成功確率が高い**

（2）既にあるデータや技術（動物での安全性データや製剤のGMP製造技術など）を再利用することで、**開発にかかる時間とコストを大幅に削減できる**

（3）**DR候補探索**に**疾患生命情報ビッグデータ知識DB**を使用できる。

Tanaka K…,**Tanaka H**, ..,Mizushima T, *Nature Comm*, 4:2686, 2013

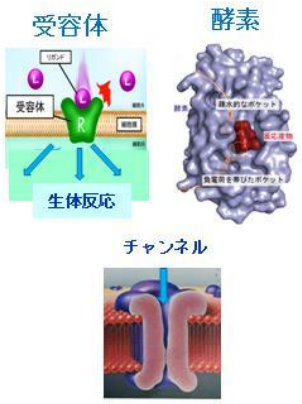
疾患・薬剤・標的の関係

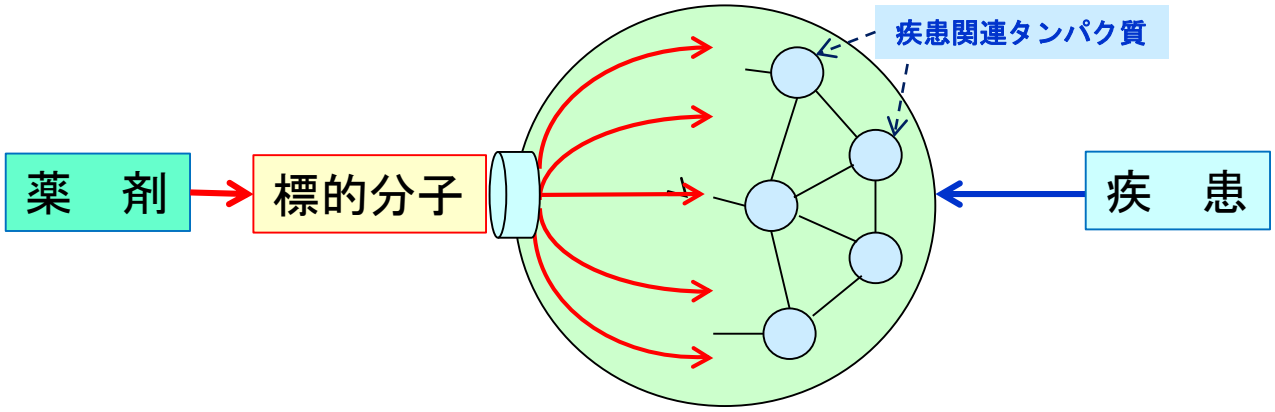
病気の主要な要因

疾患要因タンパク質（複数）

薬：疾患要因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





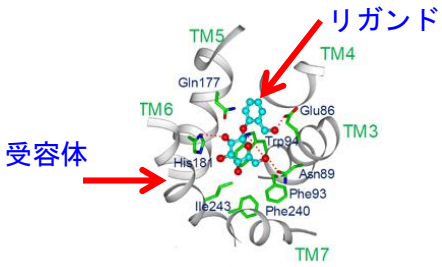
生体システム/ネットワーク

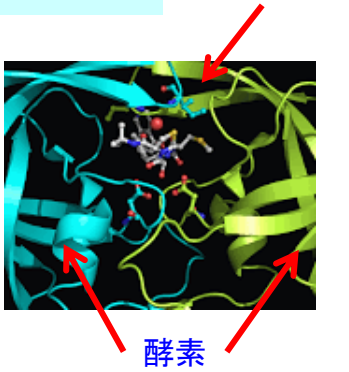
TMDU

ビッグデータ計算創薬１

計算創薬(*in silico*創薬)の新しい方向

これまでの計算創薬

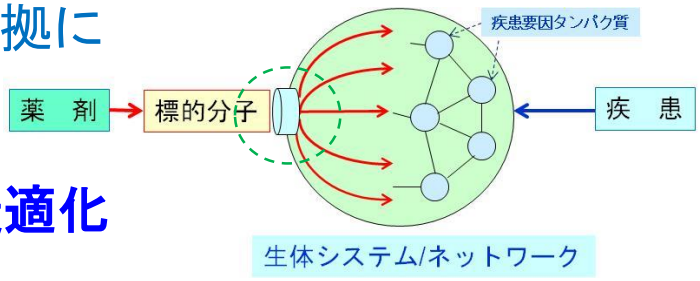




分子(結合構造)中心

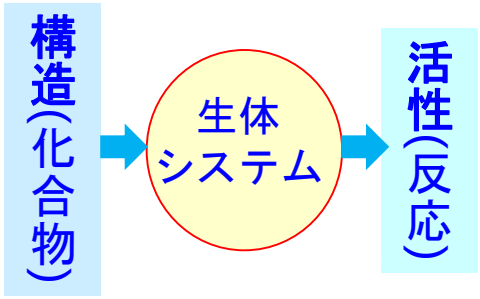
- **分子構造解析・分子設計**
- **標的分子**(受容体・酵素)**と薬剤**(リガンド)との**結合構造**（ポケット）の分子構造を根拠に
- **リガンドの分子設計(量子化学等)**
  - 成功例：インフルエンザ薬　タミフル
- 標的に結合する**リード化合物・構造最適化**
- **結合後の生体システム**の反応/振舞い

明確な取扱いがない



定量的構造活性相関(QSAR)

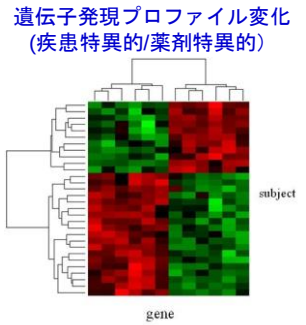
- 化合物の分子**構造**と生体**活性**の**関係**
- **両者の間**には生体システムがある



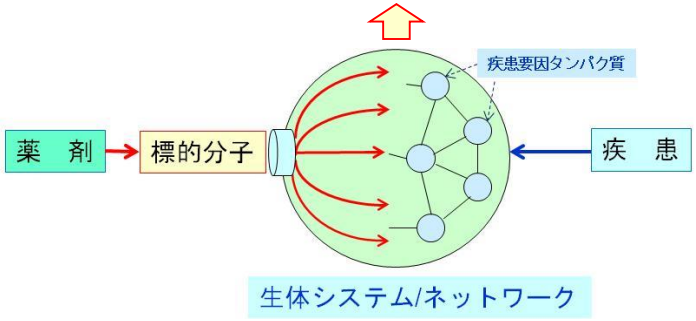
ビッグデータ計算創薬２

新しい計算論的創薬のアプローチ(網羅的分子プロファイル創薬)

疾患罹患状態における
疾患関連遺伝子（タンパク質）に起因し決定される
疾患時の**生体のゲノムワイドな特異状態**
疾患特異的な網羅的分子プロファイル変化



薬剤投与による
標的分子と薬剤分子の結合に起因し起こる
投与時の生体のゲノムワイドな反応/振舞い
薬剤特異的な網羅的分子プロファイル変化



網羅的分子プロファイル⇒分子ネットワーク全体変化

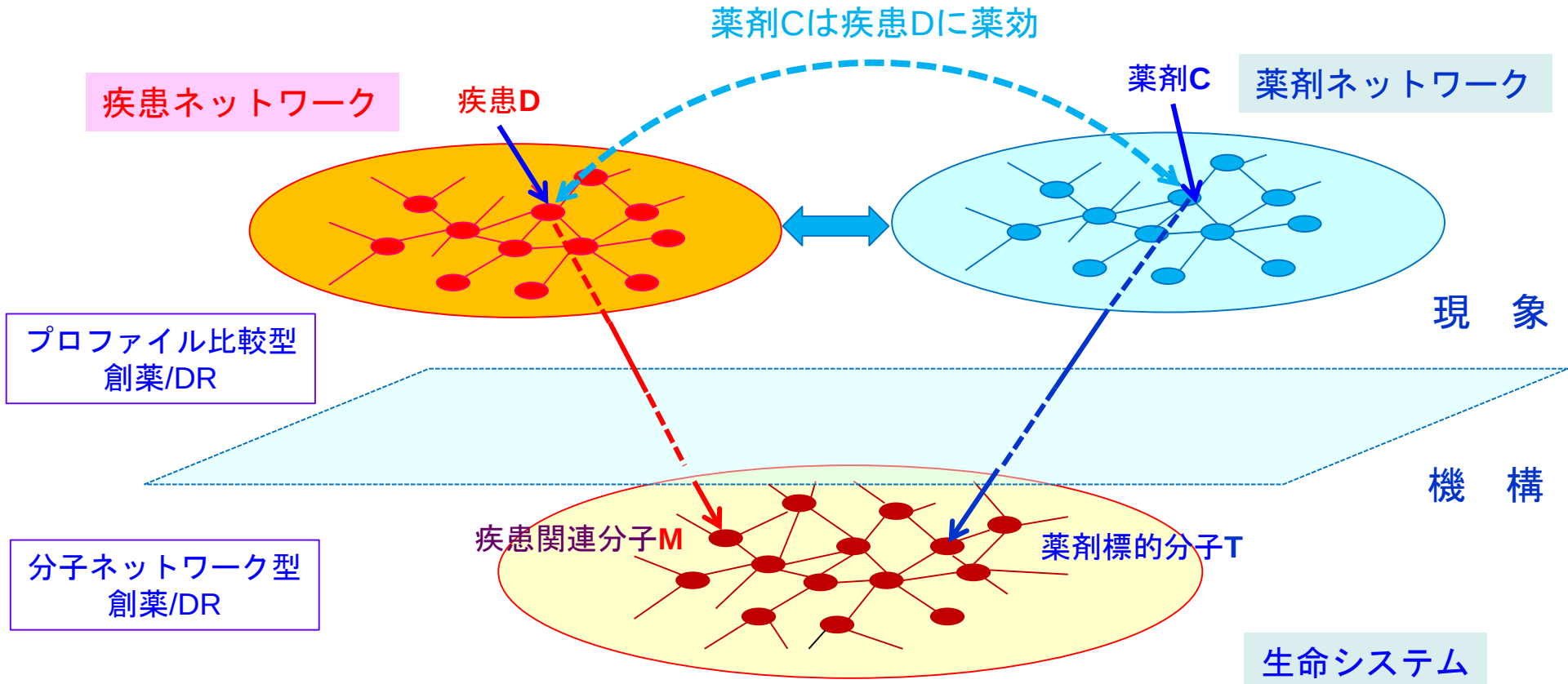
＜疾患状態の生体＞に**＜薬剤ー標的分子の結合＞**を引き起す作用によって
ゲノムワイドな 生体分子環境がどう変化するか「生命システム観点からの理解」

化合物, 標的分子, 疾患間の関係の「ビッグデータ」DBを利用

TMDU

Network型計算創薬・DRの基本的枠組み

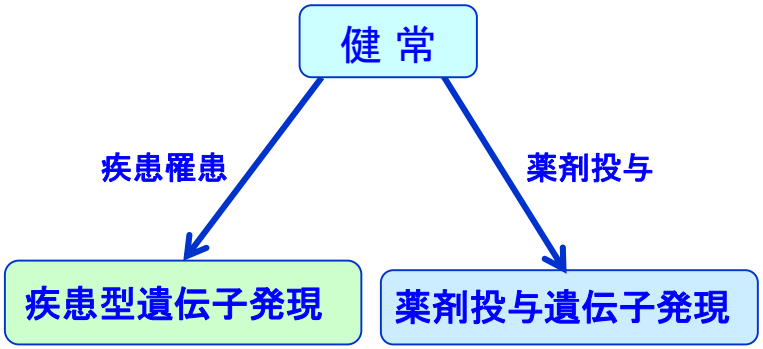
3層の生体・薬剤のネットワーク間の関係図式



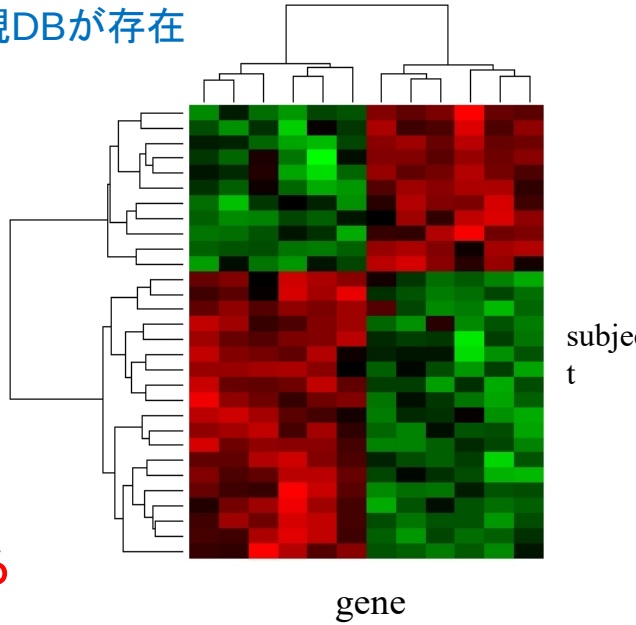
１．遺伝子発現プロファイル比較型創薬・DR

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)
    - 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
    - シグネチュア (署名)差異的発現遺伝子代表群
    - DB利用：シグネチャアを「問合せ」：
      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



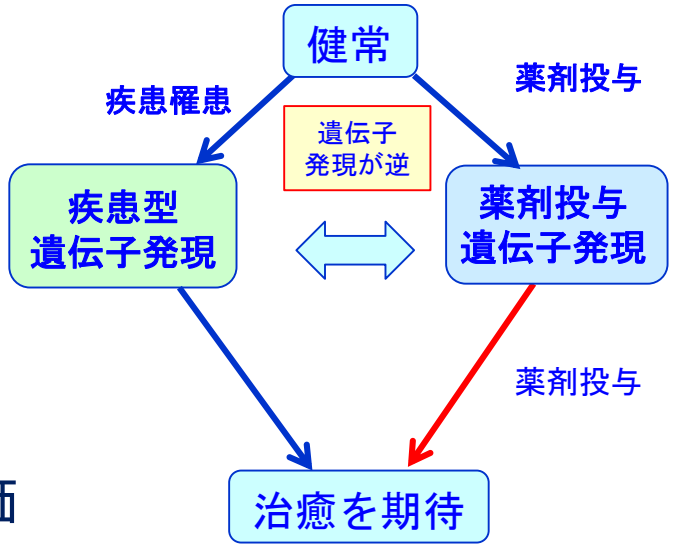
- 疾病特異的遺伝子発現
  - GEO（Gene Expression Omnibus),
    - 疾病罹患時の遺伝子発現プロファイルの変化
    - 米国NCBI作成・運用　2万5千実験,
      70万プロファイル（欧州　ArrayExpress）
    - もEBIが作成、サンプル数同程度

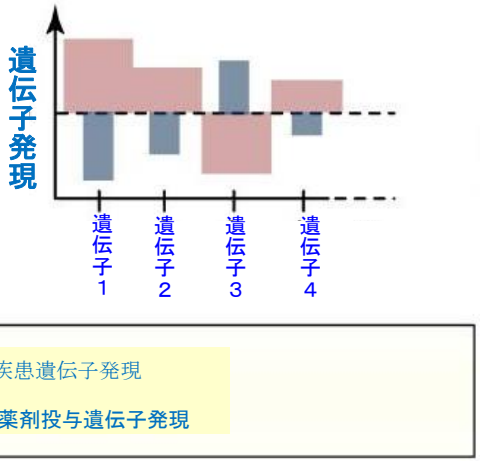


基礎には分子ネットワークの疾病/薬剤特異的変化
遺伝子発現プロファイル変化
≈　分子ネットワーク活性構造変化を反映する

遺伝子発現プロファイルによる有効性予測

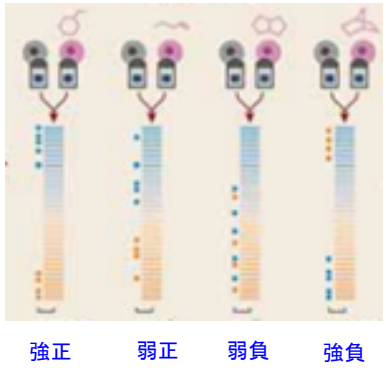
- 遺伝子発現シグネチャア逆位法
  - 疾患によって健常状態から変異
    「疾患特異的遺伝子発現プロファイル」
  - これに薬剤投与の変化を起こす
    「薬剤特異的遺伝子発現プロファイル」
  - 両者のパターンが負に相関する
  - ノンパラメトリックな相関尺度で評価
- 効果が相加的なら**有効性**が期待される
  - 例：炎症性腸疾患に 抗痙攣剤(topiramate),
    骨格筋委縮にウルソール酸





ベースは疾患遺伝子発現
横の点は薬剤遺伝子発現

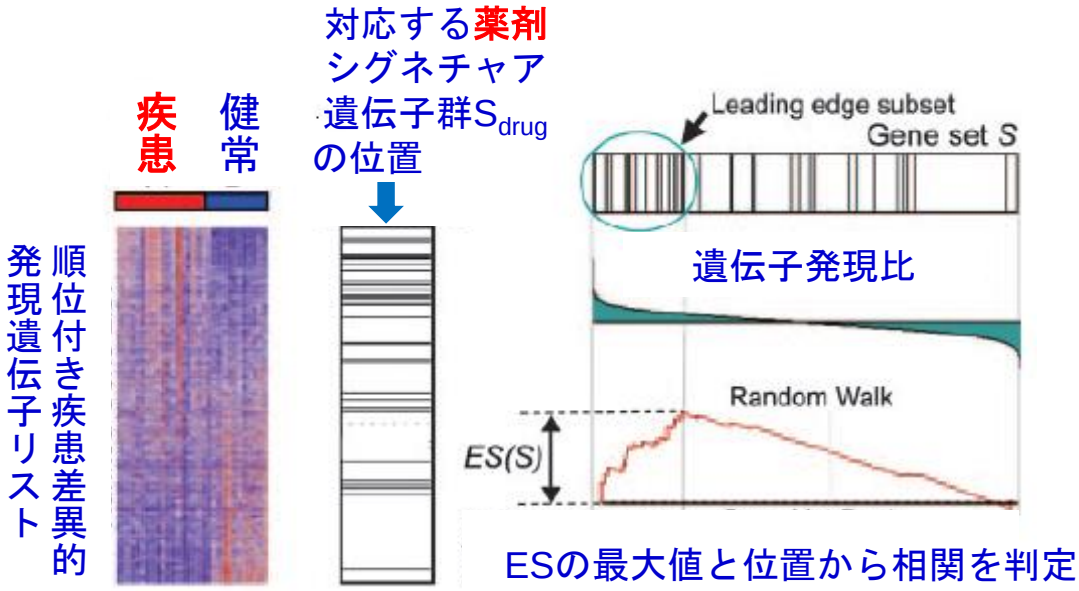
青は発現が**上昇**した遺伝子
赤は発現が**下降**した遺伝子

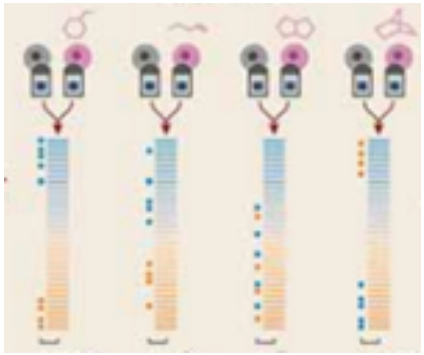


Non-parametric な相関尺度で評価

Gene Set Enrichment Analysis（GSEA)

– 対照と比較して順位づけられた遺伝子リストの上位に密集しているかの尺度

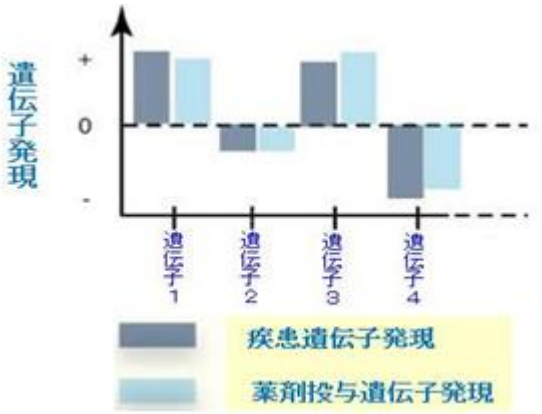




発現比ランクの高い順から遺伝子を調べ
遺伝子リストS_{drug}中に**該当する遺伝子が存在したらES (Enrich Score)を加算**、無ければ減算

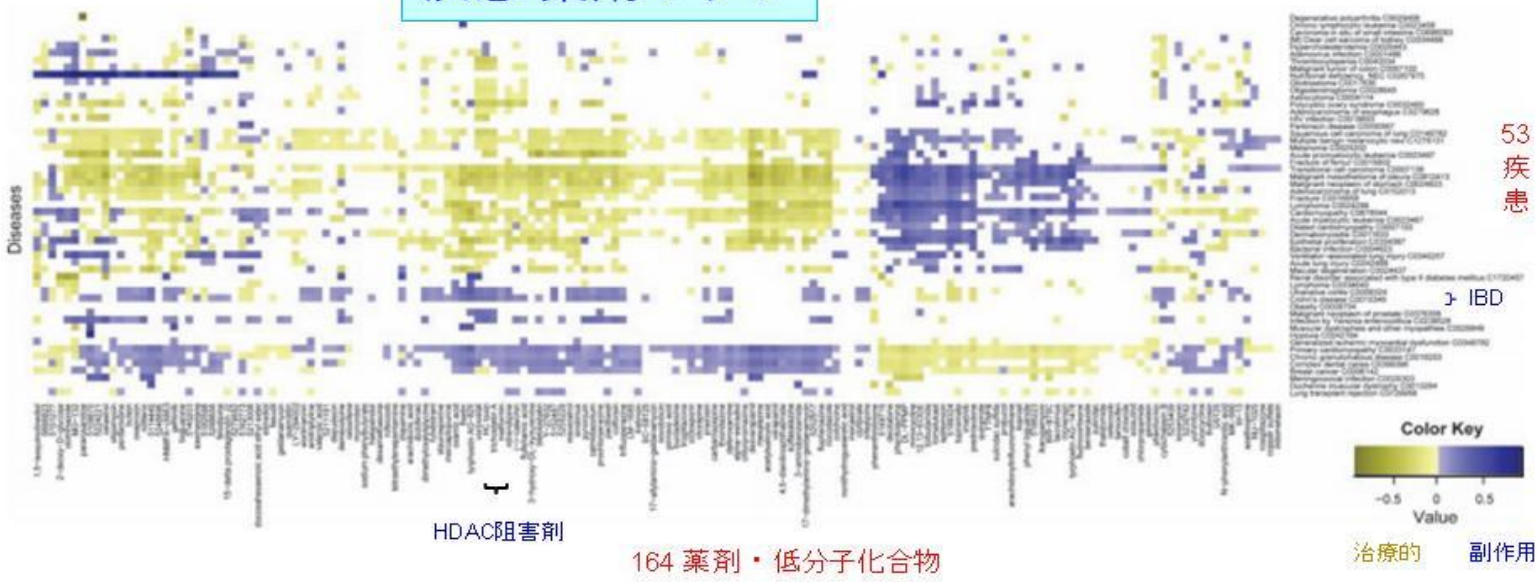
遺伝子発現プロファイルによる毒性予測

- 連座法 guilt-by-association ：
- **薬剤ー疾患間　副作用予測**
  - 薬剤特異遺伝子発現プロファイルと
  - 疾患特異的遺伝子発現プロファイルが
  - ノンパラメトリック**正に**相関
  - 毒性・副作用の予測



(Sirota, Butte 2011)

疾患-薬剤マップ



発現プロファイル原理による
＜疾患－薬剤 Map＞に基づく計算DR

- NCBI・**GEO**から**100疾患**のシグネチャアを取得
- **c-Map**より得た**164の薬剤・化合物**の
  薬剤特異的遺伝子発現プロファイル
  疾患-薬剤間で類似性スコアを計算
- 約16000組の疾患-薬剤間の2664組が
  有意、半数以上が治療的関連(負)あり
- 100疾患内, 53疾患有意に164薬剤と関連

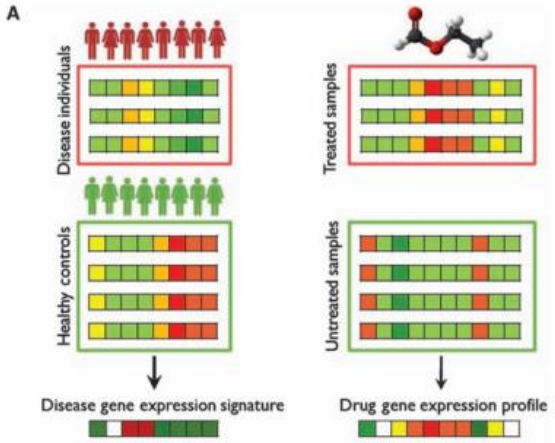


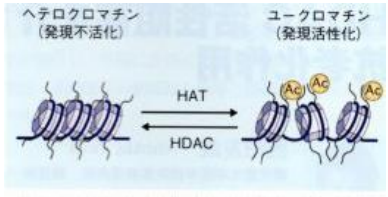
Table 1. Drugs and diseases with the most indications.

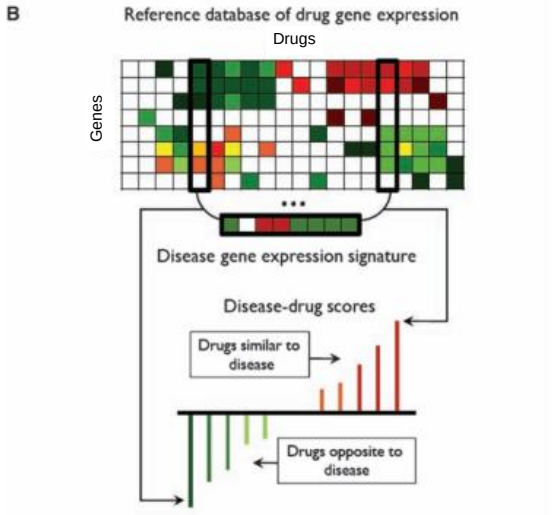
Drugs with most indications		Diseases with most indications	
Vorinostat HDAC阻害剤	21	Transitional cell carcinoma	95
Gefitinib	18	Melanoma	79
HC toxin	18	Cardiomyopathy	73
Colforsin	17	Adenocarcinoma of lung	72
17-Dimethylamino-geldanamycin	16	Multiple benign melanocytic nevi	68
Trichostatin A	16	Squamous cell carcinoma of lung	67
3-Hydroxy-DL-kynurenine	15	Malignant neoplasm of stomach	66
5114445	15	Dermatomyositis	63
Dexverapamil	15	Malignant mesothelioma of pleura	53
Prochlorperazine	15	Primary cardiomyopathy	48

(Sirota, Butte 2011)

Drug group	Drugs
PI3K inhibitors	LY-294002 and wortmannin
HSP90 inhibitors	Geldanamycin, raloxifene, monorden, and sodium phenylbutyrate
HDAC inhibitors	Vorinostat, HC toxin, and trichostatin A
Salicylate anti-inflammatory agents	Sulfasalazine, mesalazine, and acetylsalicylic acid

Canonical	Noncanonical
Cancers	Crohn's disease and lung transplant
Ulcerative colitis and Crohn's disease	Polycystic ovary and glioblastoma
	Cardiomyopathy and cancer

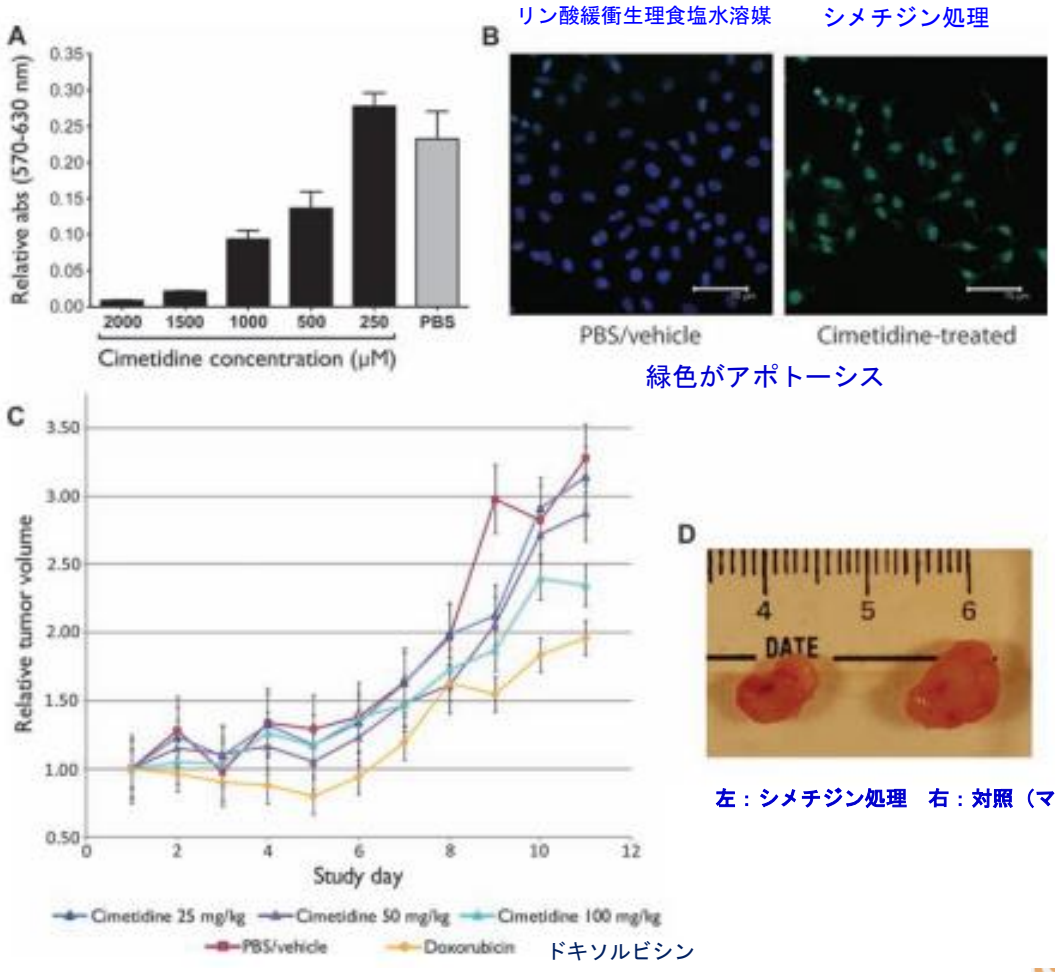




TMDU

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 の-0.075より高い



遺伝子発現Profiling による疾患ー薬剤ネットワーク（Hu, Agarwal）

遺伝子発現プロファイル(c-Map)での相関係数、ES指標によりネットワーク表示

疾患ー疾患、薬剤ー薬剤、疾患ー薬剤の
ネットワークを発現プロファイルより構成

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

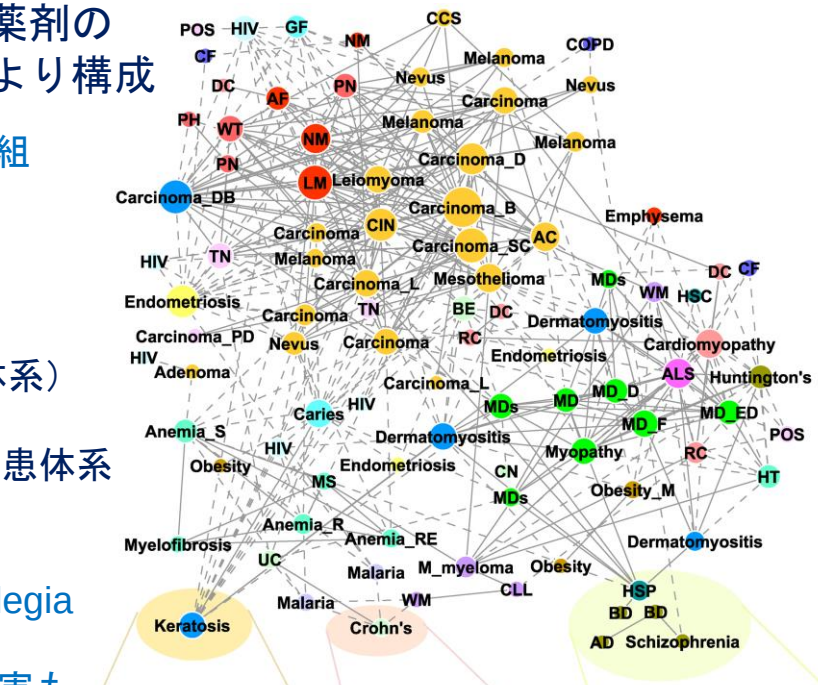
結果

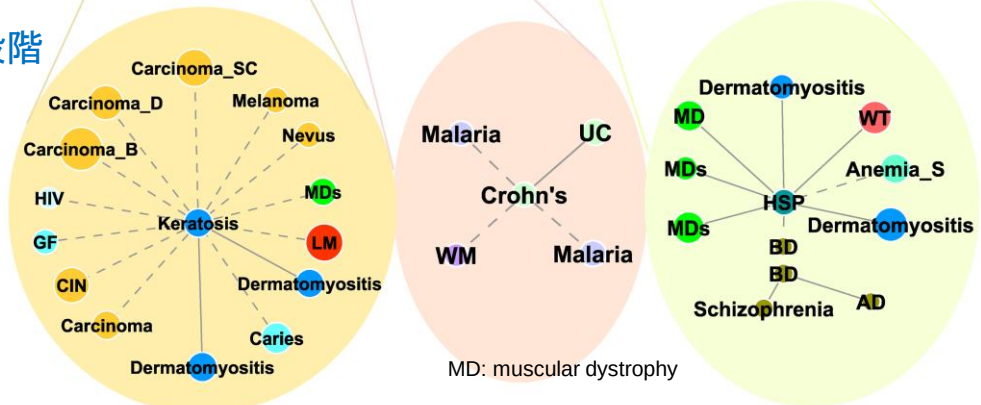
①疾患関連の60%はMeSH (既知体系）
その他は**分子レベル疾患分類学**
Transcriptomeの類似性による疾患体系

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 --精神障害も
Solar keratosis 日光性角化症
⇒ cancer(squamous) --前癌段階

＜疾患 - 薬＞
有効性：**マラリア治療薬**
⇒ Crohn’s disease
ハンチントン病に種々の薬剤





MD: muscular dystrophy

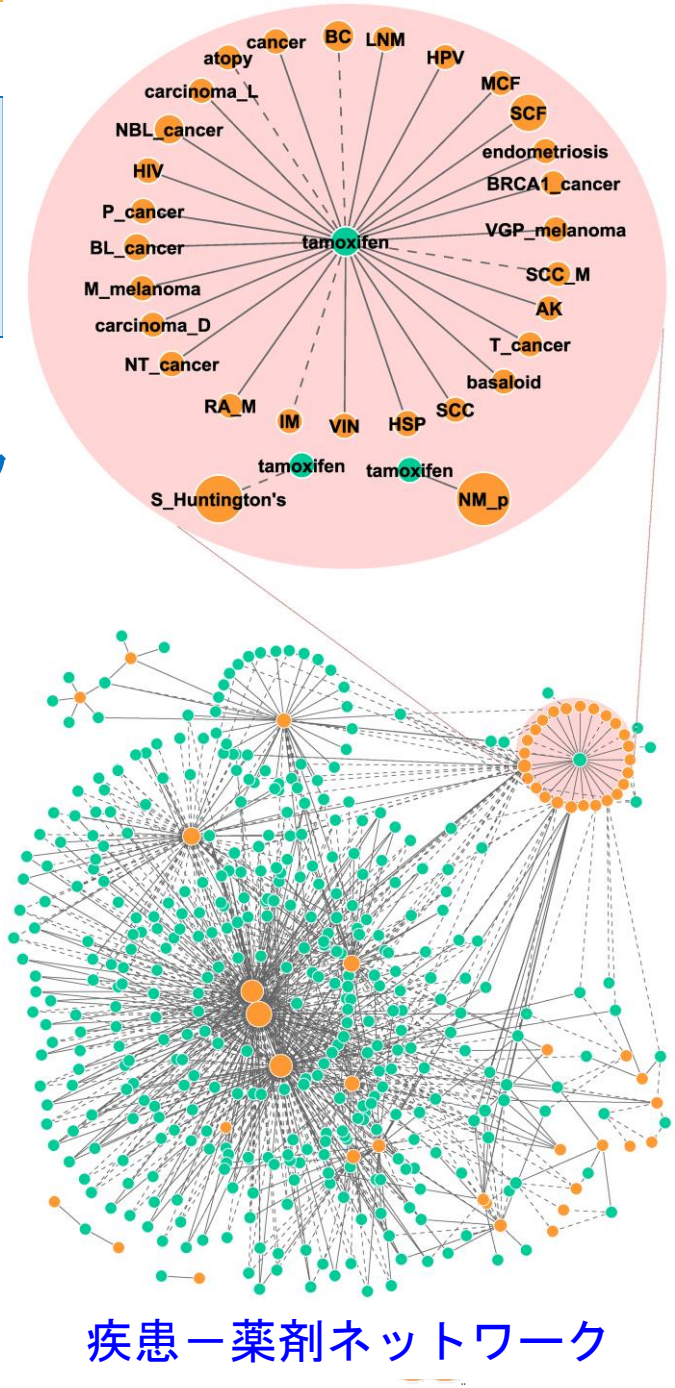
C2 virus disease
C3 parasitic diseases
C4 neoplasms
C5 musculoskeletal
C6 digestive
C7 stomatognathic
C8 respiratory tract
C10 nervous system
C12 male urogenital
C13 female urogenital
C14 cardiovascular
C15 hemic and lymphatic
C16 congenital
C17 skin/connective
C18 nutritional/metabolic
C19 endocrine system
C20 immune system
C23 pathological conds
F03 mental disorders

カラーはMeSH
同一カテゴリー
実線はMeSH内
破線はMeSH外

遺伝子発現 Profiling による Drug-Diseaseネットワーク

疾患－薬剤および薬剤－薬剤ネットワーク
（Disease-drug network:右図）
橙色節　49 疾患,　緑色節　213 薬剤
906　疾患－薬剤結合
実線　正値　破線　負値

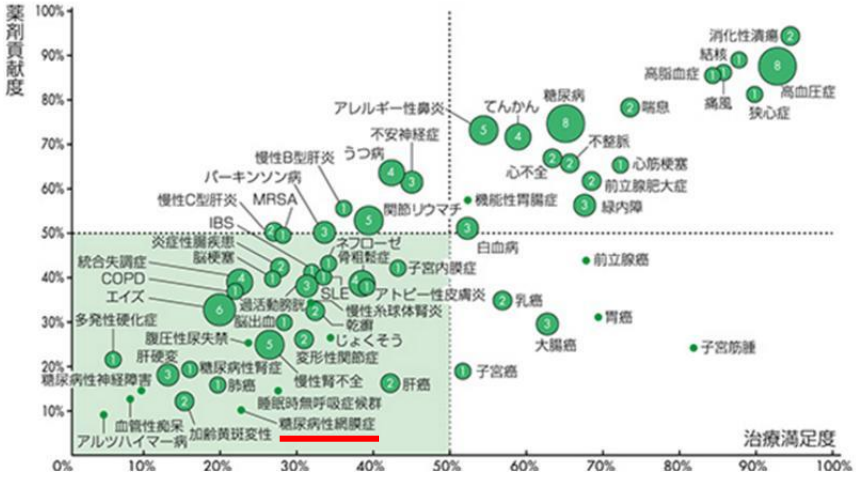
Tamoxifen (breast cancer)
有効性　負の値をもっている
⇒アトピー,
⇒マスト細胞分泌抑制、
アレルギー抑制
Hunting病に多数のDR薬
副作用　正の値をもっている
副作用の予測
⇒ 発癌性



疾患－薬剤ネットワーク

我々の研究室での成果

- 対象疾患　(Sibata et al. 2015)
  - 薬剤貢献度と治療満足度がともに低い糖尿病性網膜症 (diabetic retinopathy) の薬剤探索
- 方法
  - Signature revision法を適用
  - **疾患特異的遺伝子発現**
    - GEOから糖尿病性網膜症の遺伝子発現プロファイルを収集 (GSE53257)
    - 対照：16サンプルの健常例
    - 206遺伝子 疾患signature を確定
      - 130 up-regulated
      - 76 down-regulated genes
  - cMAPより疾患と負値ESの薬剤特異的発現を提示する有意な薬剤を探索
- 結果
  - 1600組のなかで37組の＜疾患 - 薬剤＞が有意、その中でも**11剤が負値のES**
  - FDR（*q*値）＜0.005
  - thapsigargin (score -0.983, p-value 0.00002), alprenolol (score -0.892, p-value 0.00026), ionomycin (score -0.896, p-value 0.00208), phenylpropanolamine (score -0.814, p-value 0.00219) など
- 考察
  - thapsigargin ：endoplasmic reticulum (ER) ストレスに関与。ER stress は NF-kB を活性化
  - 糖尿病性網膜症は本質的には炎症反応
  - NF-kB は the unfolded protein response (UPR) で制御されている .
  - ER stress がこの炎症の制御に役立つ可能性がある



ライフサイエンス振興財団

糖尿病性網膜症のDR候補化合物

	薬剤	SCORE	p 値
1	thapsigargin	-0.983	0.00002
2	alprenolol	-0.892	0.00026
3	ionomycin	-0.896	0.00208
4	phenylpropanolamine	-0.814	0.00219
5	etiocholanolone	-0.621	0.00961
6	kinetin	-0.72	0.01249
7	triflupromazine	-0.706	0.0155
8	vanoxerine	-0.681	0.02274
9	cicloheximide	-0.657	0.03185
10	khellin	-0.579	0.03975
11	rotenone	-0.625	0.04852

TMDU

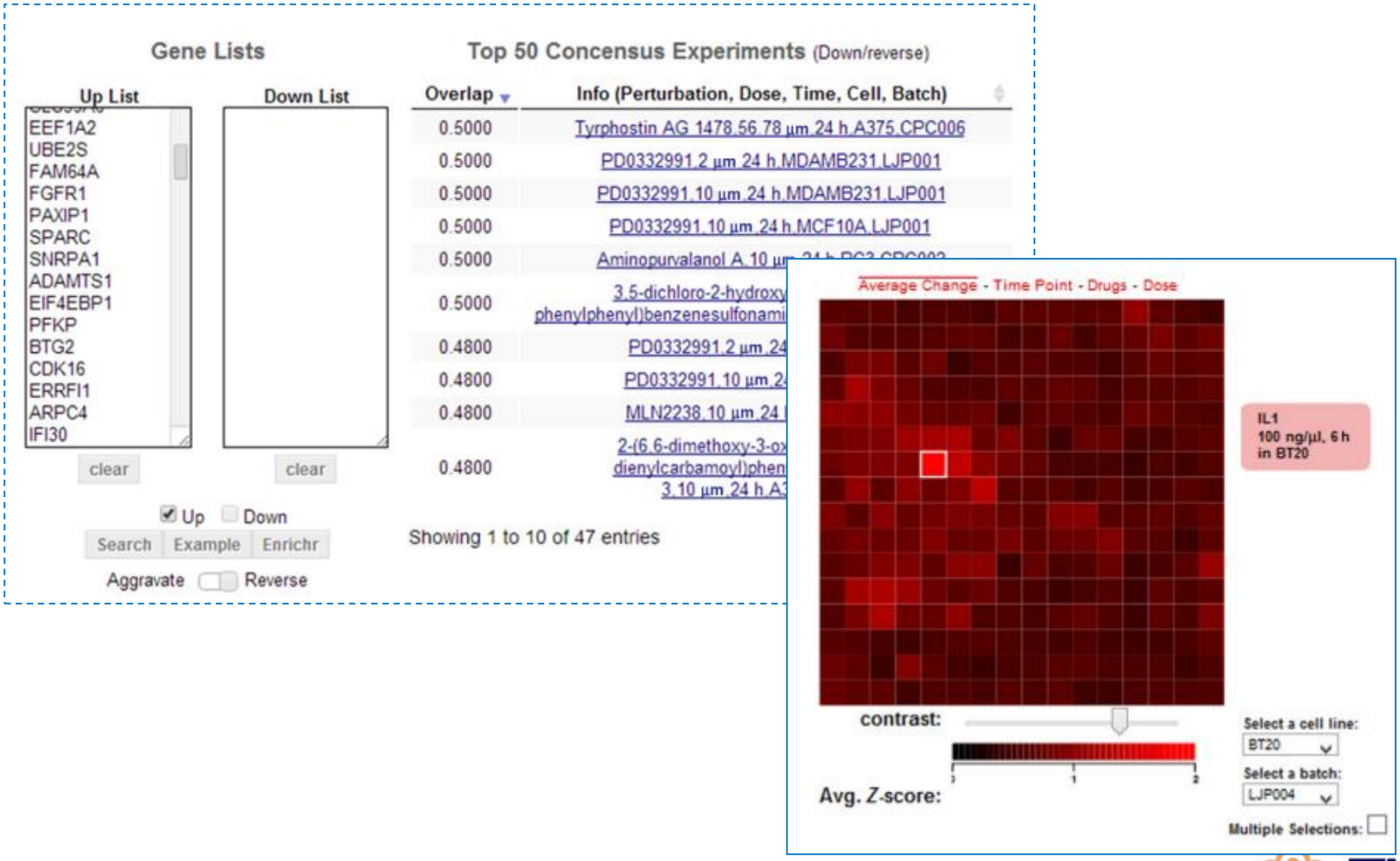
近年のビッグデータ化
LINCS

- **LINCS** (library of Integrated network-based cellular signatures)
  - GE-HTS(gene expression high throughput screening)の１つ
  - 摂動（化合物添加）を与え調節系を介して、細胞表現型を観察する
  - 遺伝子発現変化⇒差異的発現　**signature**
  - cMAP (2006, Lamb)に比べてスケール拡大
    - cMAPは、４つの細胞系列～　1300化合物　FDA認可薬剤
      Micro array (mRNA) Affymetrix U 113で遺伝子発現測定
- NIHから助成, **100万の遺伝子発現プロファイル**を **L1000 技術**で測る
  - Broad Institute cMAPと同じメンバーが考案
  - 1000遺伝子の発現しか測定しない　ゲノムワイドな遺伝子発現プロファイル（～全遺伝子22000 genesの発現）をGEOから作ったモデルで推定する
  - 相互依存性高い⇒1000遺伝子にすべて情報が含まれている
- **L1000技術**
  - 細胞溶液からリガンド媒介増幅によってmRNA増幅
  - 遺伝子特異的なProbeはcDNA（mRNA) にtaqリガーゼでアニールする
  - ProbeはPCRで増幅され、ルミネックスビーズと遺伝子特異的部分で対形成する
  - 対形成した差異染色ビーズはレーザを用いて検出され定量化される
  - ビーズの上の対形成したprobeの密度を測る　80の恒常的発現校正遺伝子
- **22412 摂動遺伝子発現**
  - 　　56　細胞コンテキスト（ヒト初代培養細胞、がん培養細胞）について
  - 16425　化合物、薬剤
  - 　5806　遺伝子ノックアウト(RNAi, miRNA)、過剰発現
  - 総計で100万ぐらい遺伝子発現プロファイルがある
- **Genometry がL1000™ Expression Profiling技術でヤンセンと契約**
  - 25万種類の化合物

TMDU

LINCSの問合せ画面

--- LINCS Canvas Browser ---

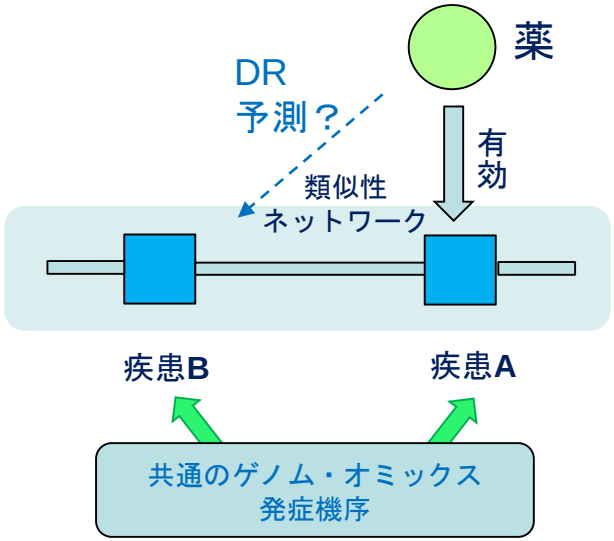


2. 疾患ネットワーク創薬/DR

疾患ネットワーク空間を基礎にした
ビッグデータ創薬/DR
＜疾患ネットワークでの近接性＞

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年に亙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病機序の概念が基礎



疾患形成のゲノム・オミックス機序

- 疾患関連遺伝子型（第１世代型）
  - 原因遺伝子、疾患感受性遺伝子の変異・多型が主要発症機序
- 疾患オミックス型（第２世代型）
  - 疾患オミックスプロファイルの変容が主要発症機序
  - Transdisease omics
- 疾患分子ネットワーク型（第３世代型）
  - 分子ネットワークの歪みが主要発症機序
  - がんなどで遺伝子型（肺腺がん等）でない通常のがん

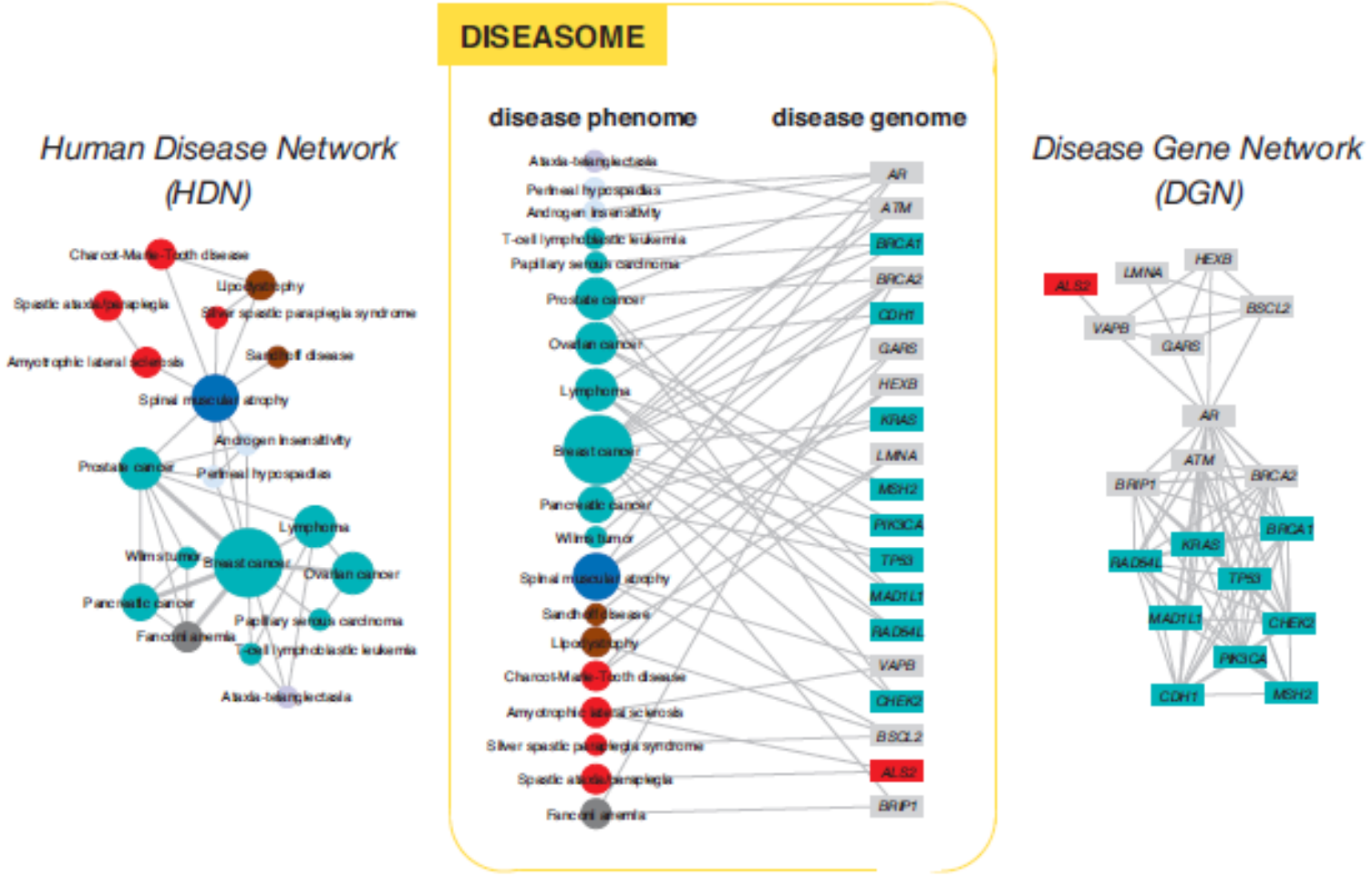
第１世代型
Diseasomeと疾患遺伝子

- **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

TMDU

疾患ネットワーク Diseasome

(Goh,Barabasi et al.)



１つ以上の疾患関連遺伝子を共有する疾患

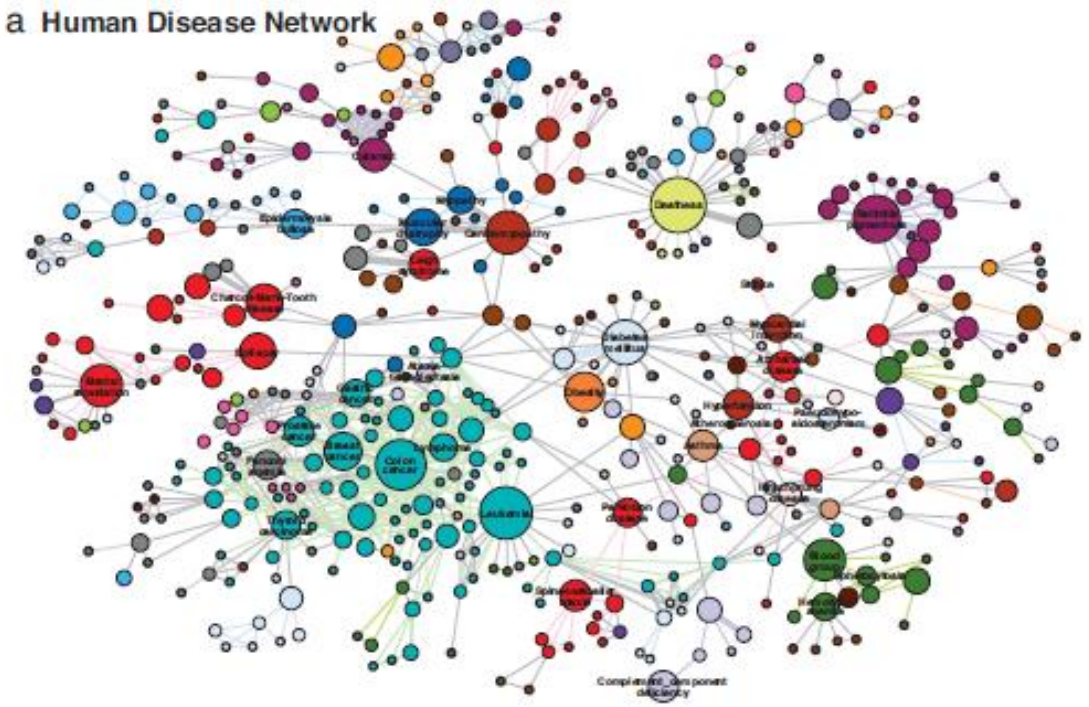
１つ以上の疾患を共有する疾患関連遺伝子

Kwang-Il Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi The human disease network　PNAS2007

TMDU

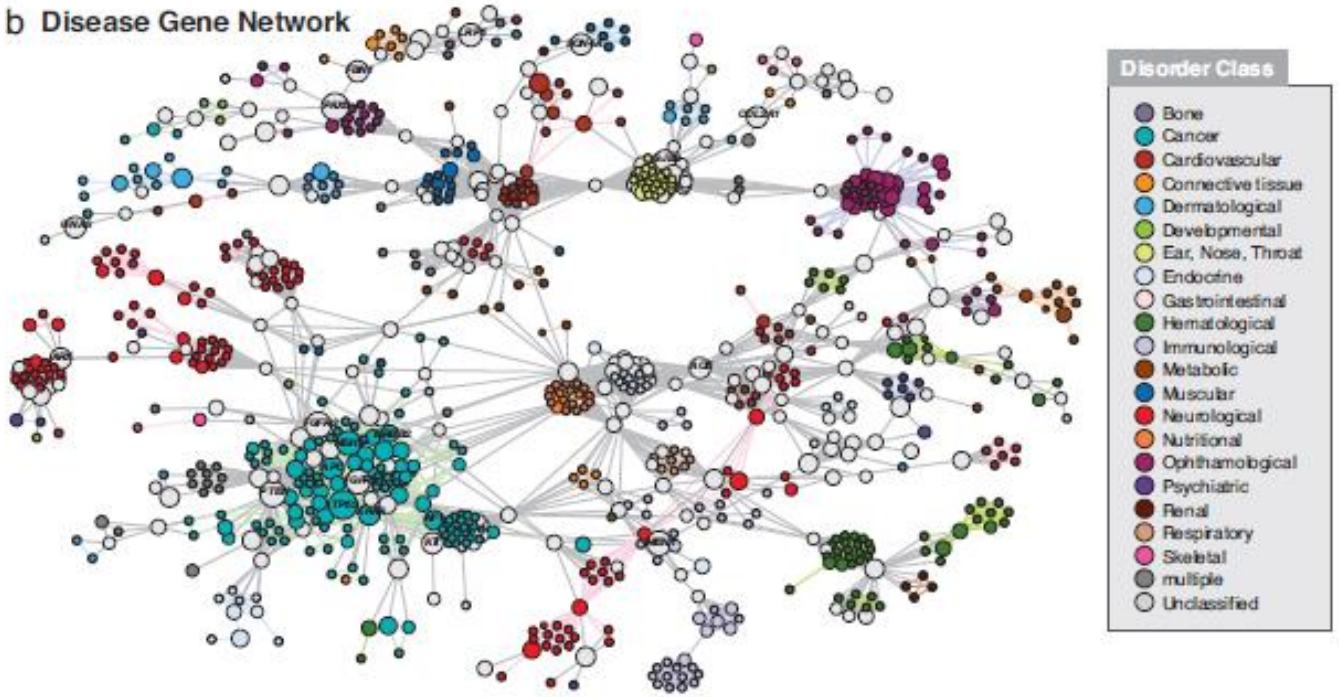
疾患
ネットワーク
(HDN)

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



疾患遺伝子
ネットワーク
(DGN)

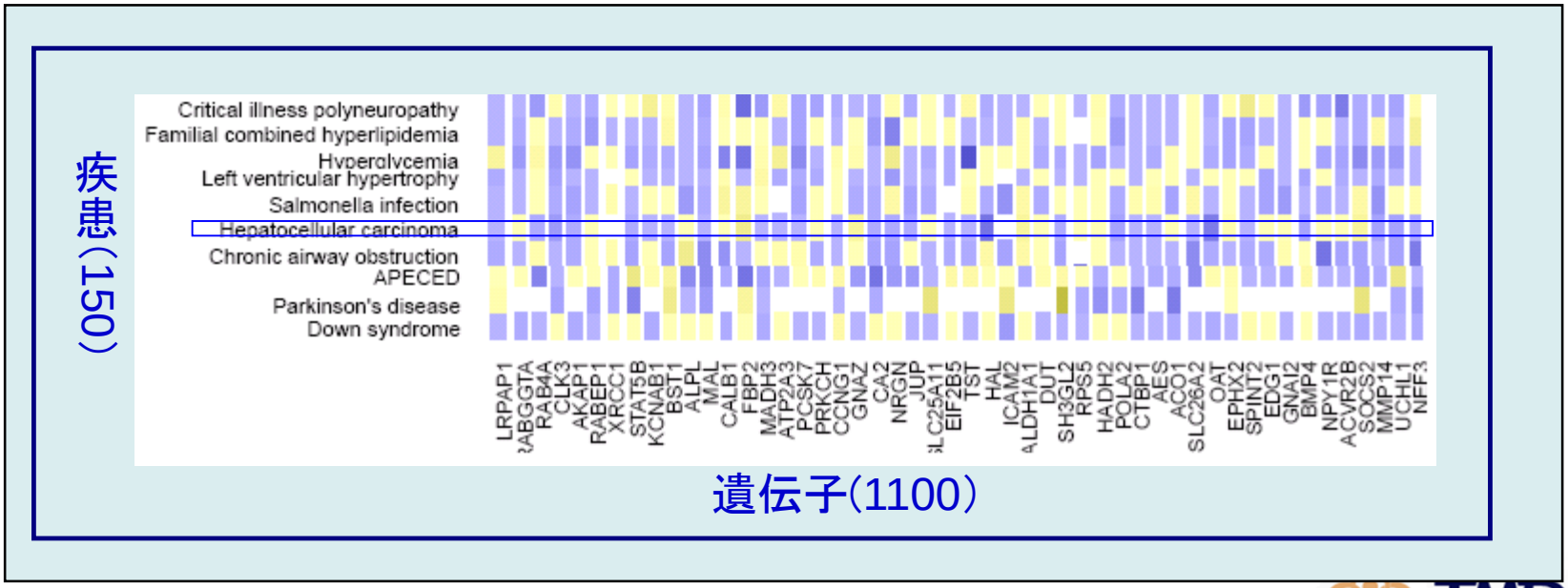
Nodeの径
その遺伝子を原因にしている疾患の数に比例
2つ以上の疾患に関与していると明灰色の遺伝子ノード



第２世代型

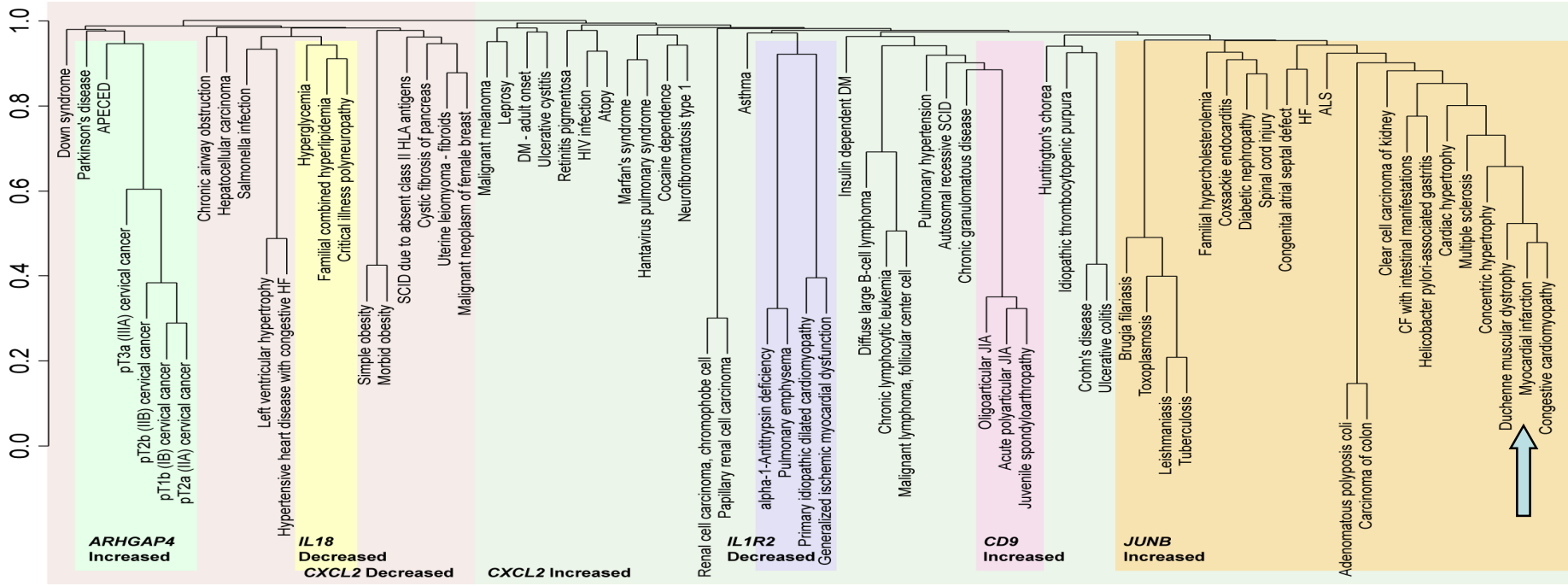
GENOMED (A.Butte et al)

- 遺伝子発現DBのGEO（Gene Expression Ominibus)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



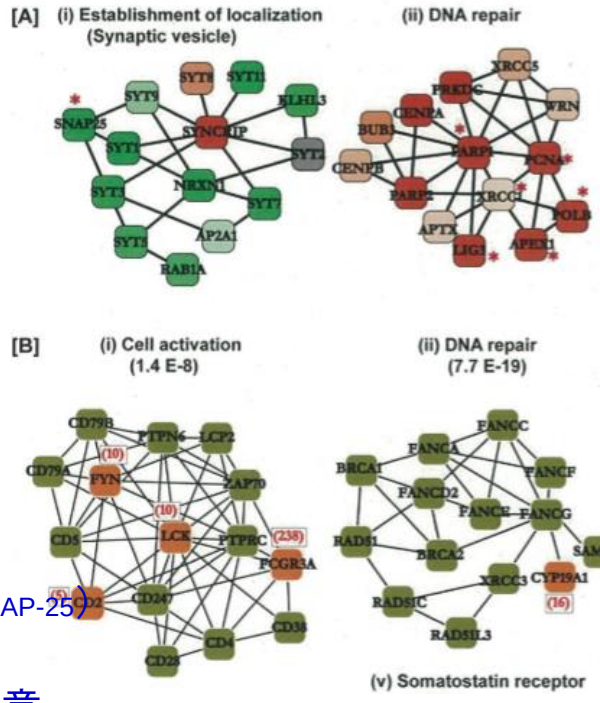
Gene-Expression Nosology of Medicine

- 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が近い



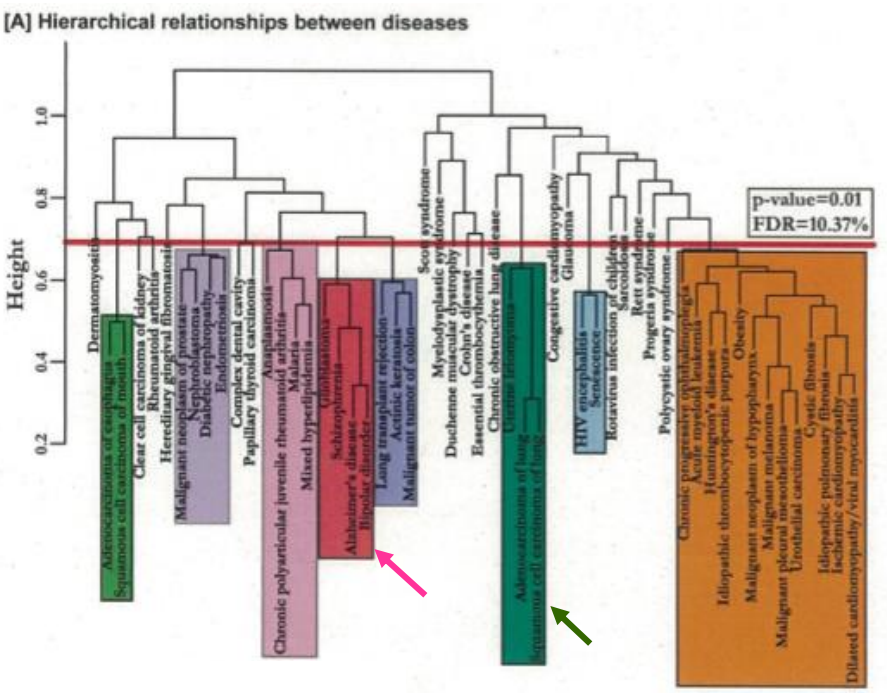
Transcriptomeの変化をPPIに投影した疾患ネットワーク（Butte)

- ネットワークモジュール
  遺伝子発現プロファイルではなくPPIを機能4620モジュール に分解
  ＜**機能moduleごとの疾病罹患時の平均発現変化**＞をもとに
  疾患ネットワーク構築
- **基本方法**
  - GEOから信頼性などより54の疾患を選択
  - 各疾患について各moduleに含まれる遺伝子群の
    疾患時と健常時の発現差の*t* 統計量の平均
  - **MRS:** Molecular Response Score
    各疾患に各モジュールで定義（ベクトル量）
  - 疾患間の相関は、両疾患の健常時発現を制約とした
  - MRSの偏相関係数
- **疾患ネットワークの性質**
  - **138の有意な類似性**：ランダム化ネットに対し有意
  - p<0.01,FDR=0.1
  - 疾患類似性: 肺がん群(修復pasM), 精神疾患(synapsM:SNAP-25)
- **138の有意な疾患相関**
  - 17は少なくとも1つの共通薬：14疾患は共通の薬剤に有意
  - Flourarcil（日光性角化）⇒大腸がん、ほかDoxorubicin
- 疾患の大半を占める**59モジュール**：「共通"疾患状態"モジュール」
  - 「共通疾患状態シグネチャア」薬剤標的分子に富んでいる
  - この遺伝子群を標的にする薬剤は有意に多くの他の疾患の薬剤にもなっている

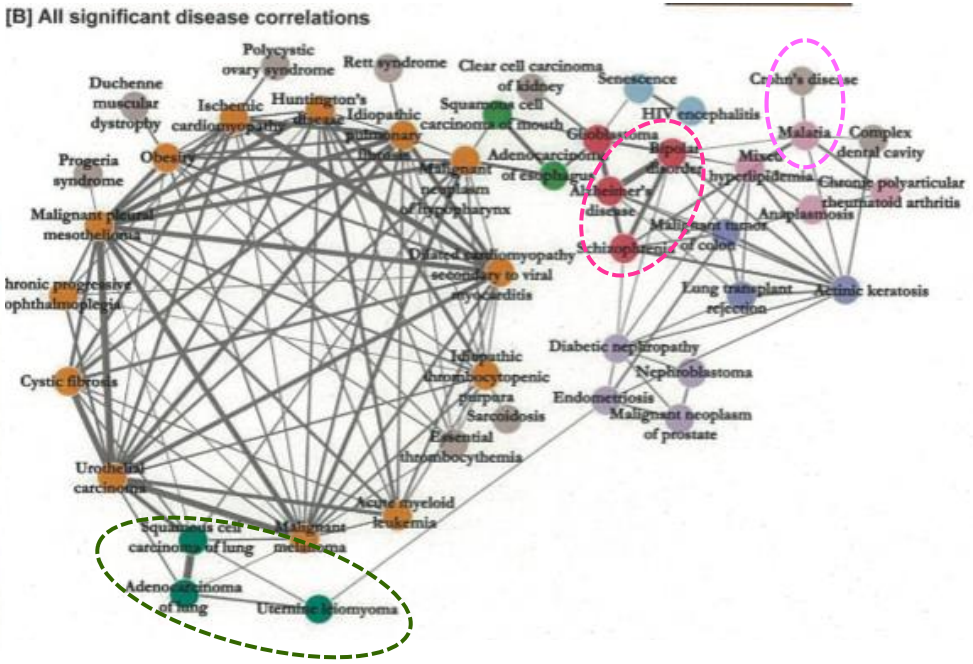


Transcriptomeの変化をPPIに投影した疾患ネットワーク（Butte)

- アルツハイマー症、統合失調症、双極性障害がグループ化
- 子宮筋腫と肺がん、マラリアとクローン病
- 17のがんが１つの群ではない。がんの異質性
- 疾患ネットワーク間の遺伝子共有は高くない（遺伝子外効果）



階層的クラスタリング

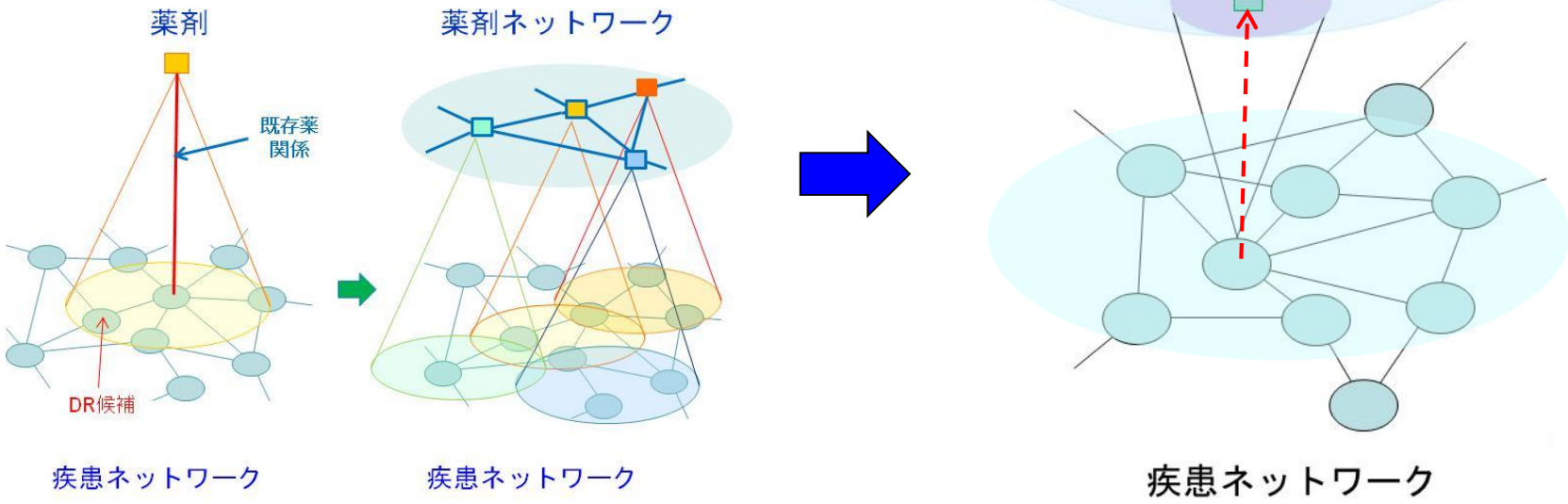


疾患ネットワーク

DRの方法論から創薬方法論へ

- 疾患ネットワークの十全な形成
  - 内在的機序の近親性から**疾患ネットワーク**
  - 医薬品の有効性・毒性の近傍 Projection
    ⇒ DRにおける**有効性**はすでに確立
- 創薬への展開
  - 薬剤階層のネットワークは既に確立
  - 投与時生体反応の近親性だけではなく
  - 化合物の構造的近親性(finnger print) からも作成
  - 疾患から逆投影。創薬の可能性探索
- 疾患ネットワークと薬剤ネットワーク間写像
  - **双方向性・対等性**

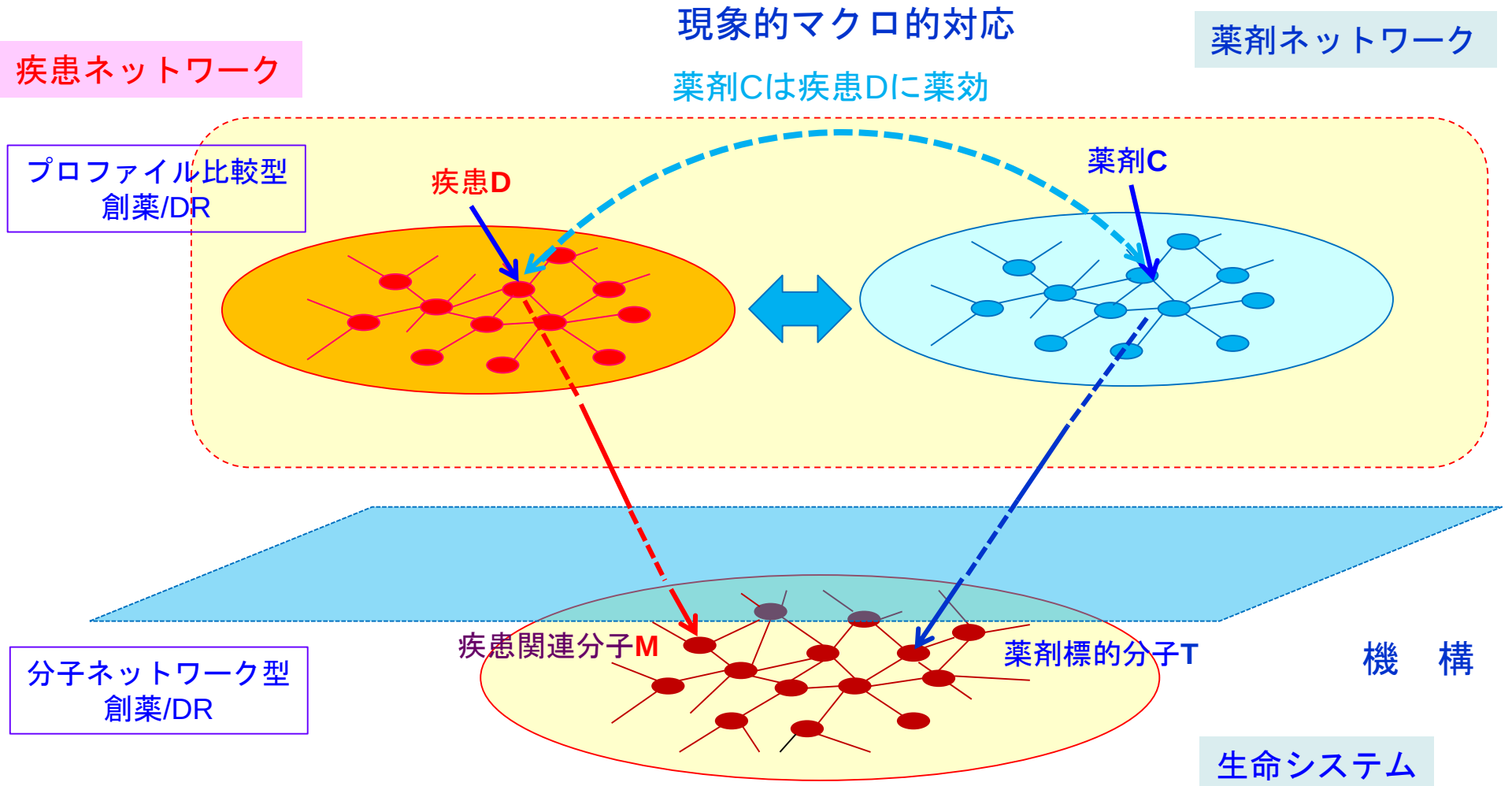
疾患から薬剤ネットワークへの逆投影
Multi-Topology **双対写像　創薬方法論**



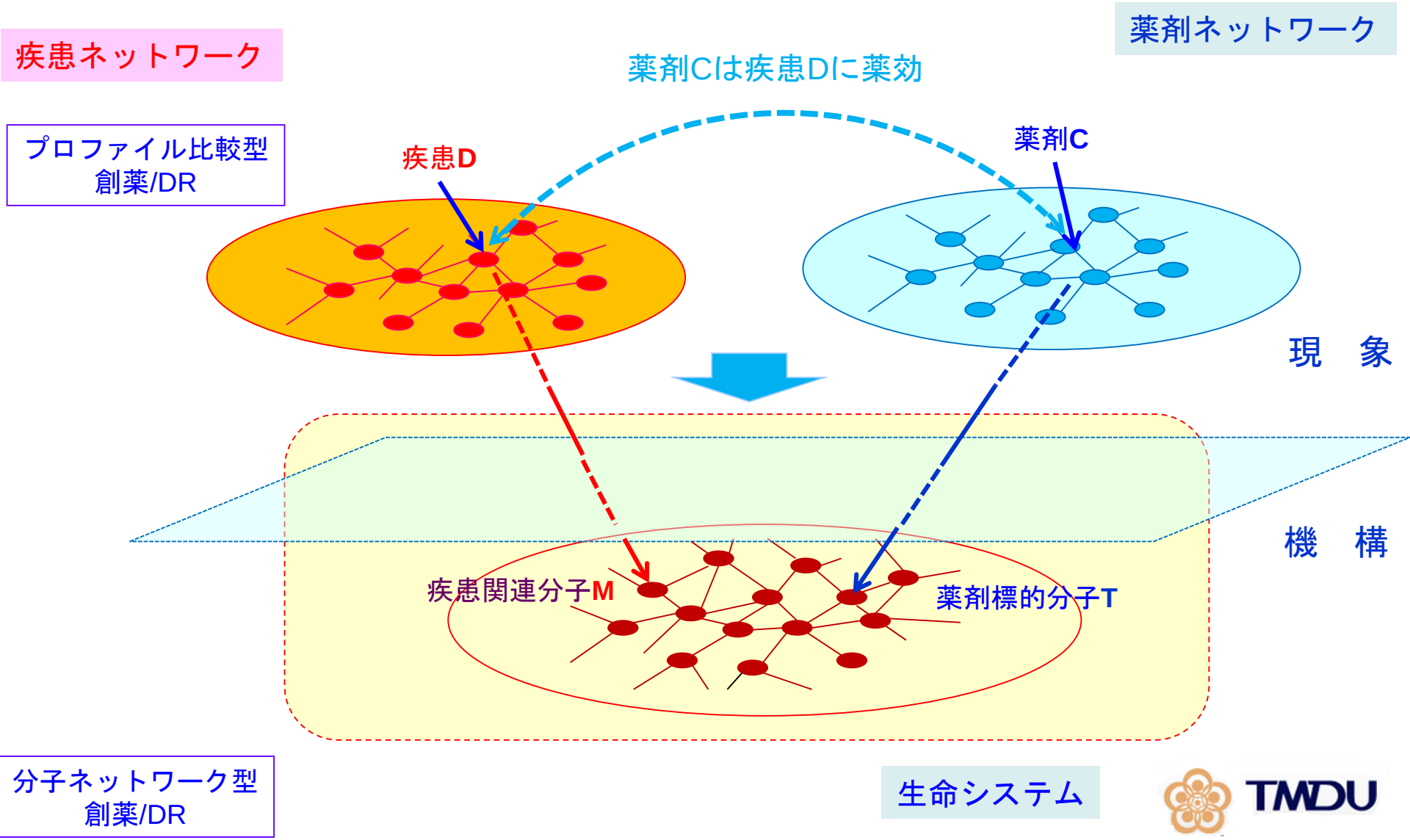
＜疾患-薬剤-標的分子＞の多階層ネットワークによる

ビッグデータ創薬/DR

3層の生体・薬剤のネットワーク間の関係図式

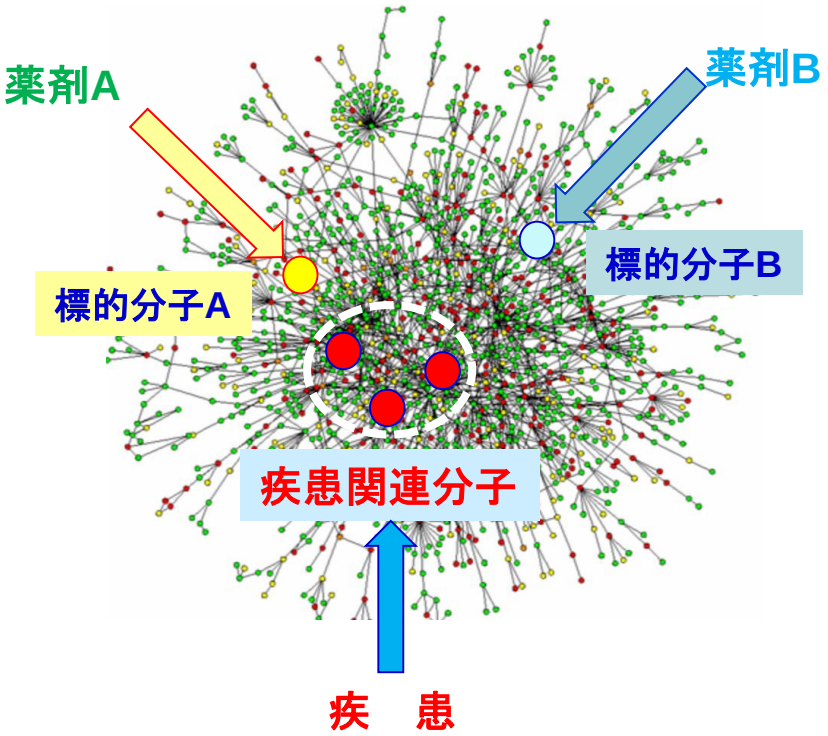


3層の生体・薬剤のネットワーク間の関係図式



標的分子や疾患要因分子の
タンパク質相互作用ネットワーク（PPIN）

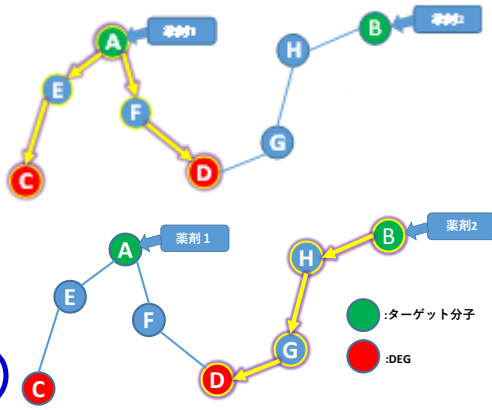
- 薬剤ネットワークと疾患ネットワークを媒介する第３の生体ネットワーク
- タンパク質相互作用ネットワーク（PPIN）での創薬/DR戦略
- PPIネットワーク場を基礎にして距離（類似性）を検討
- **薬　剤**：**薬剤の標的分子**（タンパク質）によって　PPI場と繋がる
- **疾　患**：**疾患特異的発現遺伝子**を疾患要因分子（**タンパク質**）へ翻訳、
- PPIN場内での**薬剤標的分子と疾患の「代理人(疾患遺伝子）」**の**距離・親近性**を基準に、薬理作用のインパクト力を評価



タンパク質相互作用
ネットワーク（PPIN）

PPIの基づくDR（肺腺癌の例）

- **Interactome**(タンパク質相互作用)ネットワーク（Sun, 2016）
  - HPRD（Human Protein Reference Database）
    - 37,070 PPI, 9465 タンパク質
  - STRINGS（Search Tool for the Retrieval of INteracting Genes/proteins)
    - 184 M PPI, 9,643,763タンパク質 --- 個々に計算



- **薬剤⇒標的分子**：DrugBank
  - 7,759 薬剤、4300タンパク質
  - 12,604 の薬剤ー標的分子組 （4,452薬剤, 1,617タンパク質）
- **疾患遺伝子の差異的遺伝子発現データ（DEG）**
  - TCGA (The Cancer Genome Atlas)より差異的発現遺伝子を同定
    - 445 肺腺癌例, 19 正常例, 疾患遺伝子 FC >2.0 or <0.5,FDR<0.01, **927 差異的発現遺伝子**
- **薬剤の疾患遺伝子への影響力 評価IPS** (Impact power score)
  - **薬剤の標的分子**と**疾患遺伝子**の間の**ネットワーク距離**の**総合評価**
    「再出発ありランダム歩行RWR**」**でネットワーク距離を評価
  - **標的分子からランダム歩行を繰り返す**（出発点から再出発あり）
    s時点後, 疾患遺伝子のノードにどれだけの確率で滞在しているかを**IPS**とする
  - 一定の時間が過ぎると、**定常状態**になり、歩行で滞在確率分布は変化しない。
    定常状態での**疾患遺伝子ノードに滞在している確率**の総和が薬剤の評価になる

$$\mathbf{P}^{s+1} = (1-\gamma)\mathbf{M}\mathbf{P}^{s} + \gamma\mathbf{P}^{0}$$

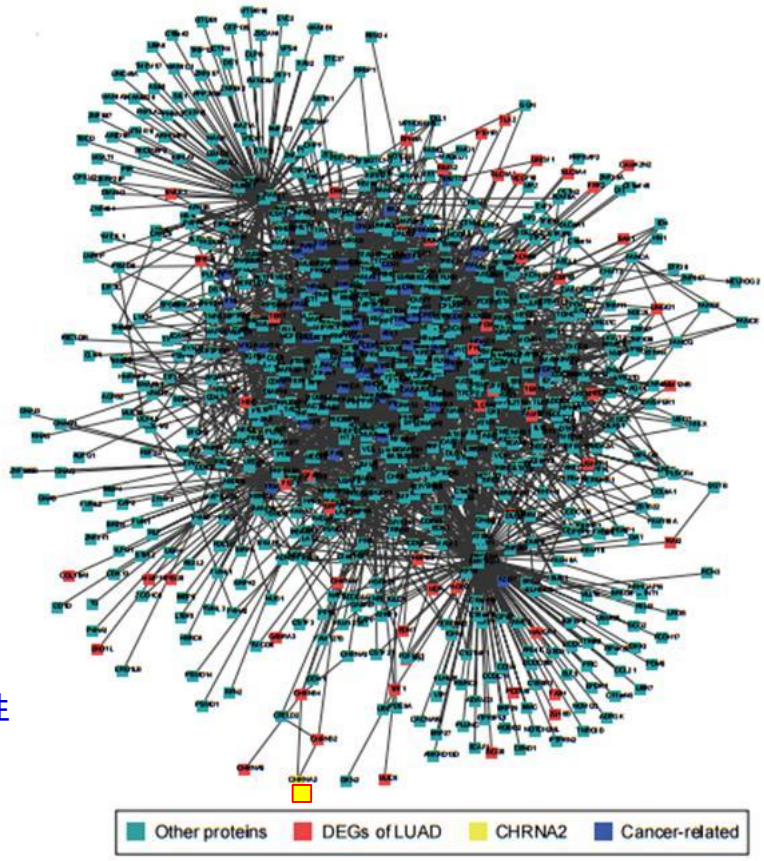
$\mathbf{P}^{s}$：時点sでの各ノードでの滞在確率　$\mathbf{M}$：各ノードへの遷移確率　γ：再出発確率

TMDU

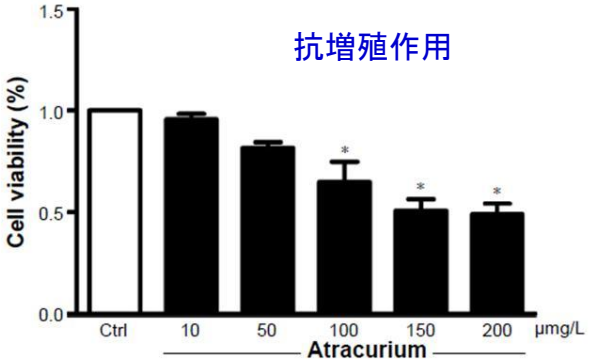
Interactome DR 結果の検証

Drug ID	Drug name	Target	Score	Rank
DB00416	Metocurine Iodide	CHRNA2	0.966581	1
DB00565	Cisatracurium besylate	CHRNA2	0.966581	1
DB00732	Atracurium	CHRNA2	0.966581	1
DB00657	Mecamylamine	CHRNA2	0.966581	1
DB02457	Undecyl-phosphinic acid butyl ester	LIPF	0.953846	5

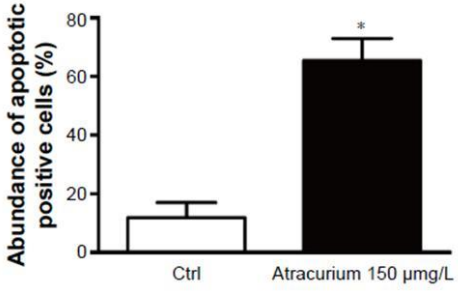
- **HPRD**と**STRINGS**の両方のPPINのランダム歩行でtop5%で共通な**145薬剤**を同定
- 最高スコアを挙げた**Atractium**を選択
- 薬剤標的は**CHRNA2**(Cholinergic Receptor Nicotinic Alpha 2）**でアポトーシス経路である**
- **培養細胞A549**（ヒト肺胞基底上皮腺癌）胞）の抗増殖作用を確認



抗増殖作用



アポトーシス活動性



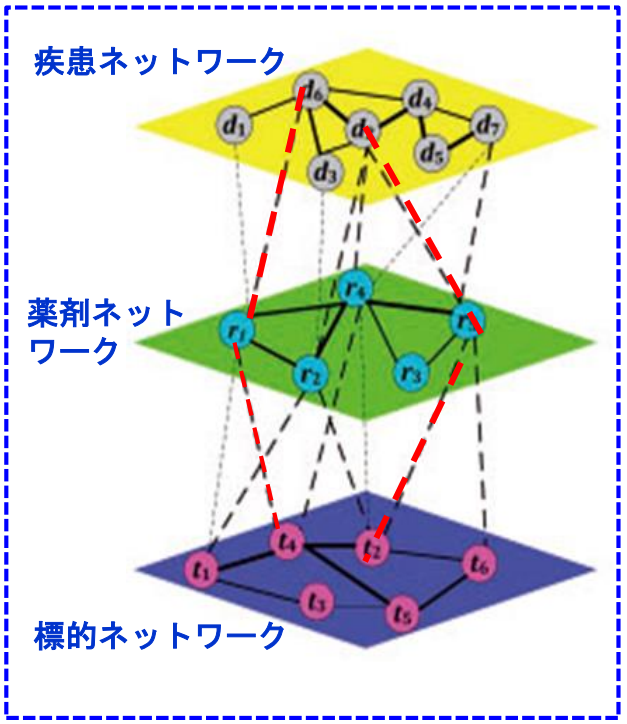
TMDU

3階層生命ネットワークでの創薬/DR

- 3階層の生体ネットワーク
  - 疾患ネットワーク：網羅的分子による内在的機序
  - 薬剤ネットワーク：化学構造によってネットワーク
  - 標的ネットワーク：薬剤と標的（DrugBank参照）
- 各層のネットワーク内結合
  - 稠密に自己完結的に構築可能
- 各層ネットワーク間のリンク
  - 成功した＜疾患－薬剤＞の事実の根拠のみ
  - 階層間はスパースな結合である

創薬/DRとは

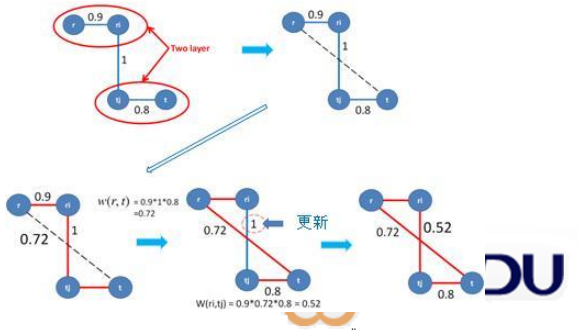
未発見の階層間リンクを
既存の階層間リンクの事実と
各層のネットワークから推測



(Wang et al. 2014)

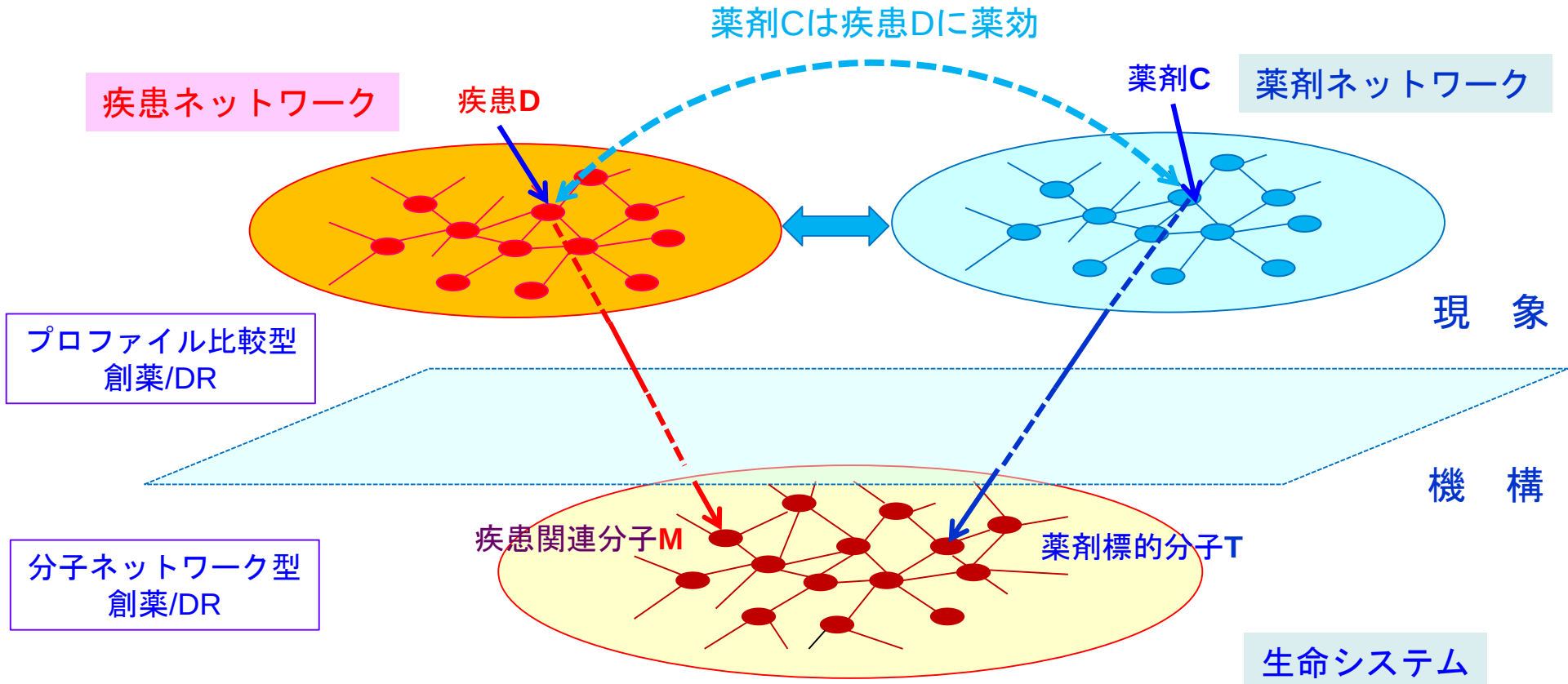
Wang et al. 2014は

- 階層間リンク（事実）と各階層内のリンクより階層間のリンクの強さを計算する方法を提案している



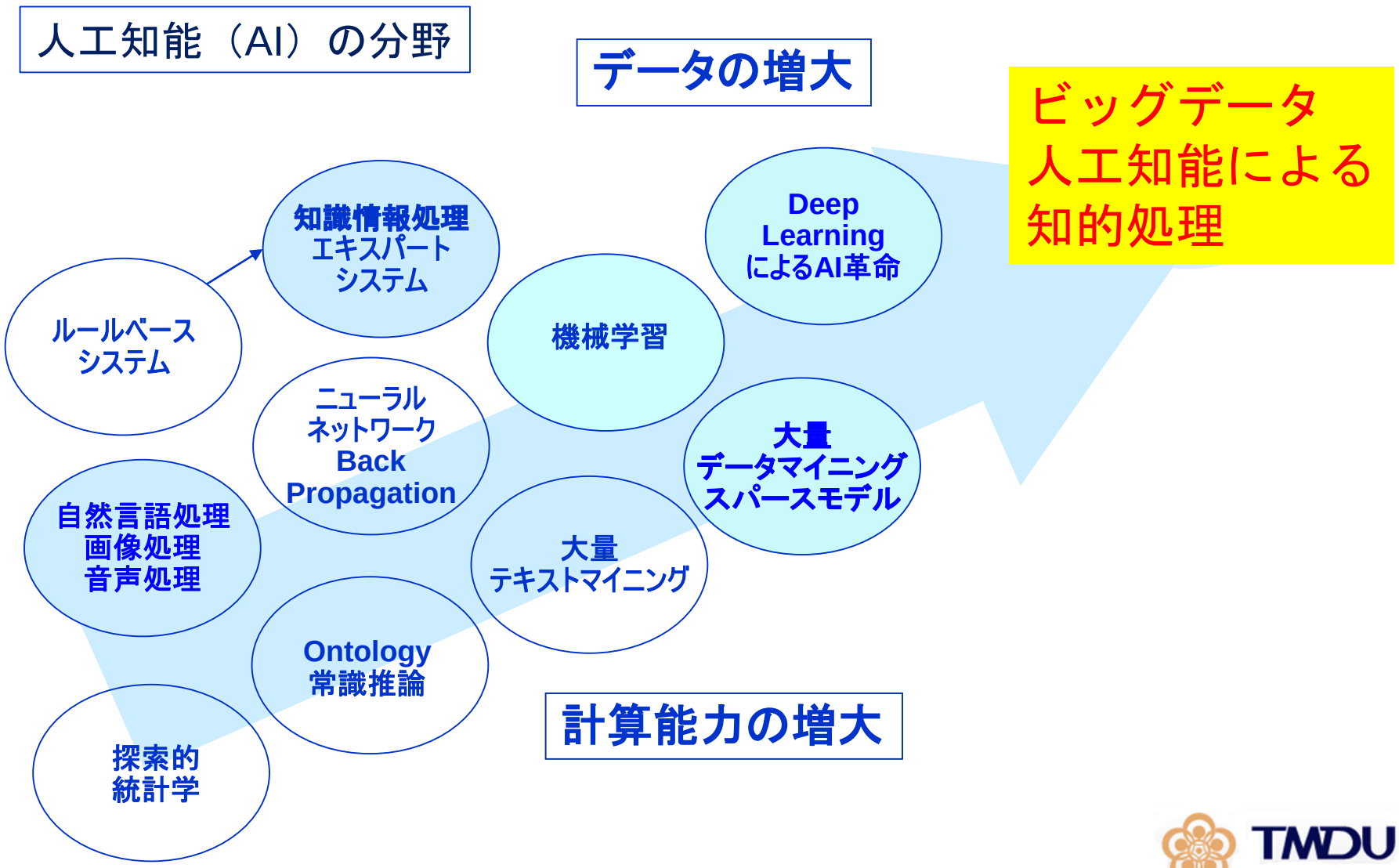
プロファイル型計算創薬の原理

3層生体・薬剤ネットワークのFramework



第２部
人工知能（AI）創薬・DR

人工知能への期待


人工知能（AI）の分野
データの増大
ビッグデータ
人工知能による
知的処理
知識情報処理
エキスパート
システム
ルールベース
システム
Deep
Learning
によるAI革命
機械学習
ニューラル
ネットワーク
Back
Propagation
大量
データマイニング
スパースモデル
自然言語処理
画像処理
音声処理
大量
テキストマイニング
Ontology
常識推論
探索的
統計学
計算能力の増大

医療分野の人工知能の歴史

記号（シンボル）的知識処理

ニューロネットワーク処理

1970

問題解決の一般探索手法 GPS
解決木の高速探索（ゲーム）

ニューロネットワーク
3層の学習機械 Perceptron
入力層、隠れ層、出力層

1980

推論システム（if-thenルールシステム）
知識の表現と利用（専門家システム）
医療診断システム（Mycin, Internist-I)
大ブーム　医療から産業応用の期待波及

多層型ニューロネット
後方伝播 Back Propagation
結合係数修正アルゴリズム

期待消滅！

1990

しばらく停滞！

知識発見 機械学習
Machine Learning, KDD
診断知識のDBからの学習

ニューロネットワーク型
多層型ニューロネット
深層学習 Deep Learning
結合係数修正アルゴリズム
画像処理から創薬まで

2000

知識準拠診療支援（DSS)
医療ターミノロジー
医療オントロジー

自己紹介と医療人工知能の歴史

人工知能(AI)を医療・創薬へ応用

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

1980から1995
第１期の
AIブームの時
医療AI研究に従事

私の人生

東大医学部

医療AI研究

東京医科歯科大

ゲノム医療

BigData医学
AI創薬

記号知識処理

問題解決の探索法（GPS）

医学「知識」を計算機に格納
医療診断システム（MYCIN）
知識工学：大ブーム
政府：第５世代コンピュータ
知識の移植問題

ブーム消滅！

医療機械学習
診断知識のDBからの学習

診療支援
医学の用語や
概念体系の基礎理論

ニューロネット(NN)

単純NN
パーセプトロン
判別能力の限界

1970
以前

1980

多層NN
バックプロ
パゲーション
重み修正の限界

1990

ブーム消滅！

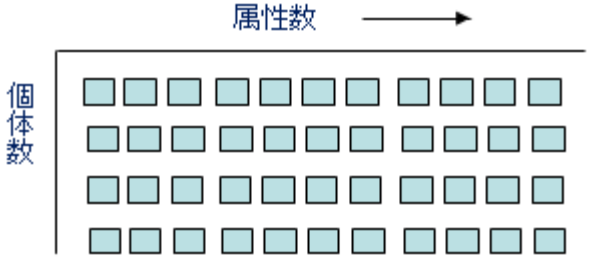
2000

Deep Learning
多層NN
「教師なし」特徴学習

「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

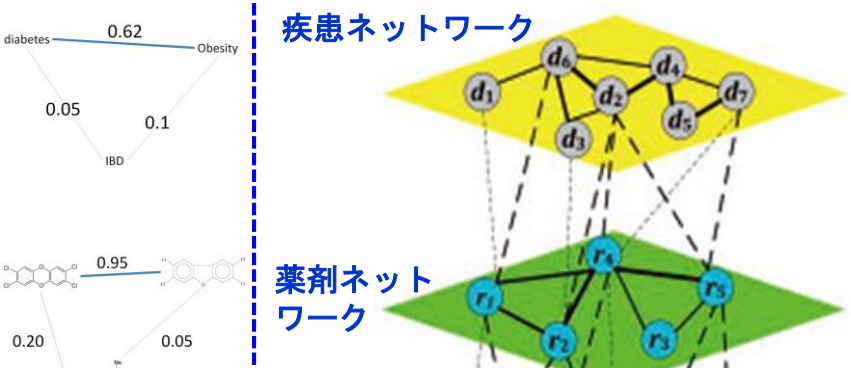
ビッグデータ・スパース仮説

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）

データ次元縮約の原理（**principle of compositionality**)

Wangの異質ネットワークの計算アルゴリズム

- **３層ネットワーク構成**
  - 疾患-疾患 (d_i), 薬剤-薬剤(r_i), 標的-標的 (t_i)
  - 疾患-薬剤間の距離を2階層のレベル内での距離から計算
- 各ネットワークで距離定義
  - 疾患：表現型⇒MeSHの共通項数
  - 薬剤：化学構造⇒Tanimotoスコア
  - 標的：Protein配列類似性⇒Smith-Waterman法
  - 疾患-薬剤：過去研究、薬剤-標的ネット：Drugbankより
  - 失われた疾患-病気 Edge復元
- **結合係数W(*i*,*j*)更新法**



$$w(d,r)=\sum_{d_i\in D}\sum_{r_j\in R} w(d,d_i)\times w(d_i,r_j)\times w(r,r_j)$$

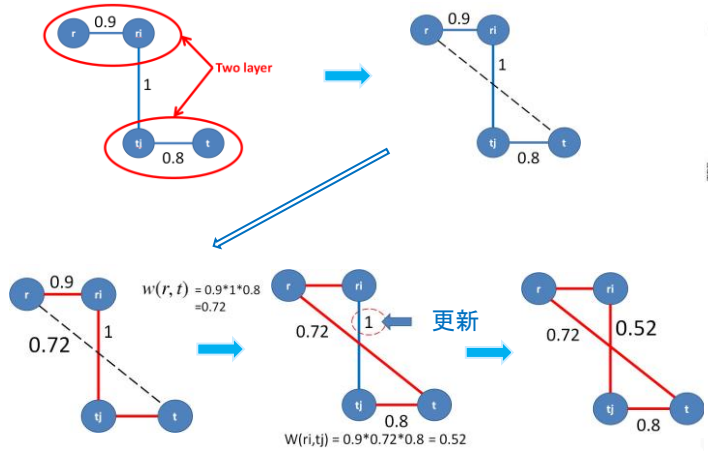
$$w(d,t)=\sum_{r_i\in R}\sum_{r_j\in R} w(d,r_i)\times w(r_i,r_j)\times w(r_j,t)$$

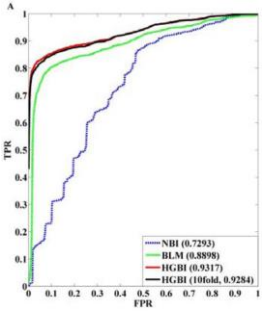
$$w(d,r)=\sum_{t_i\in T}\sum_{t_j\in T} w(d,t_i)\times w(t_i,t_j)\times w(t_j,r)$$

結合係数更新のマトリックス表示

$$W_{dr}^{k+1}=\alpha W_{dr}^{k}\times(W_{rr}\times W_{rt}^{k}\times W_{tt}\times W_{rt}^{kT})+(1-\alpha)W_{dr}^{0}$$

$$W_{rt}^{k+1}=\alpha(W_{dr}^{kT}\times W_{dd}\times W_{dr}^{k}\times W_{rr})\times W_{rt}^{k}+(1-\alpha)W_{rt}^{0}$$



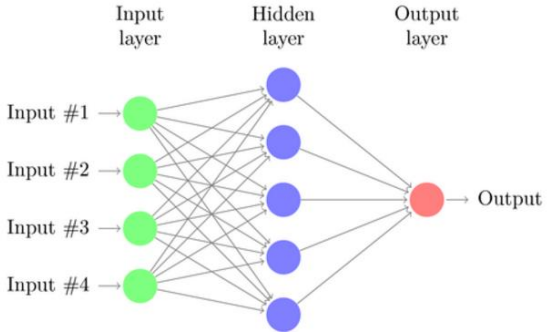


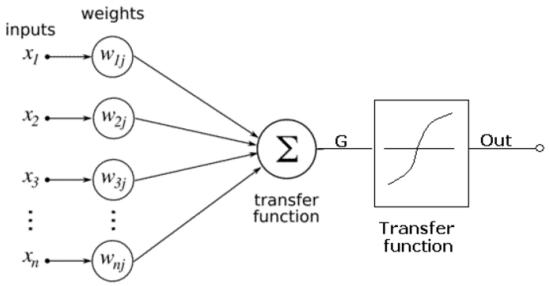
従来の方法より
DR推定精度高い
ROC曲線

Deep Learning 型人工知能の革命性

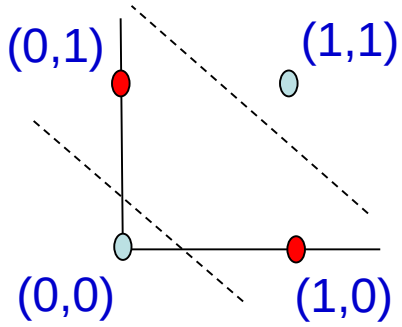
従来のニューロネットワーク

古典的Neural Network・パーセプトロン(1970年代）



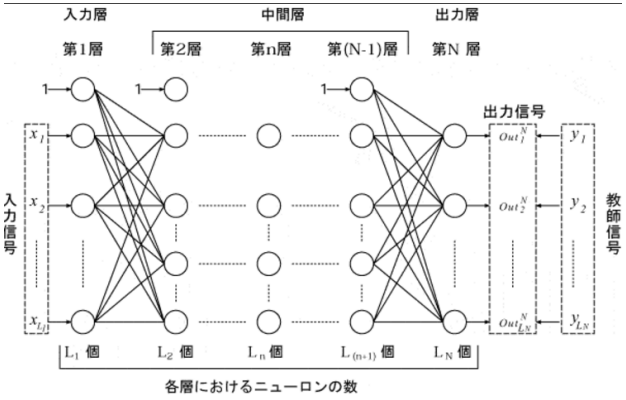


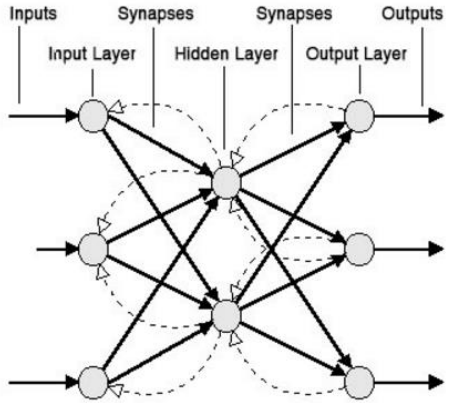
XOR



線形分離できない

多層Neural NetworkとBack projection (1980年代）





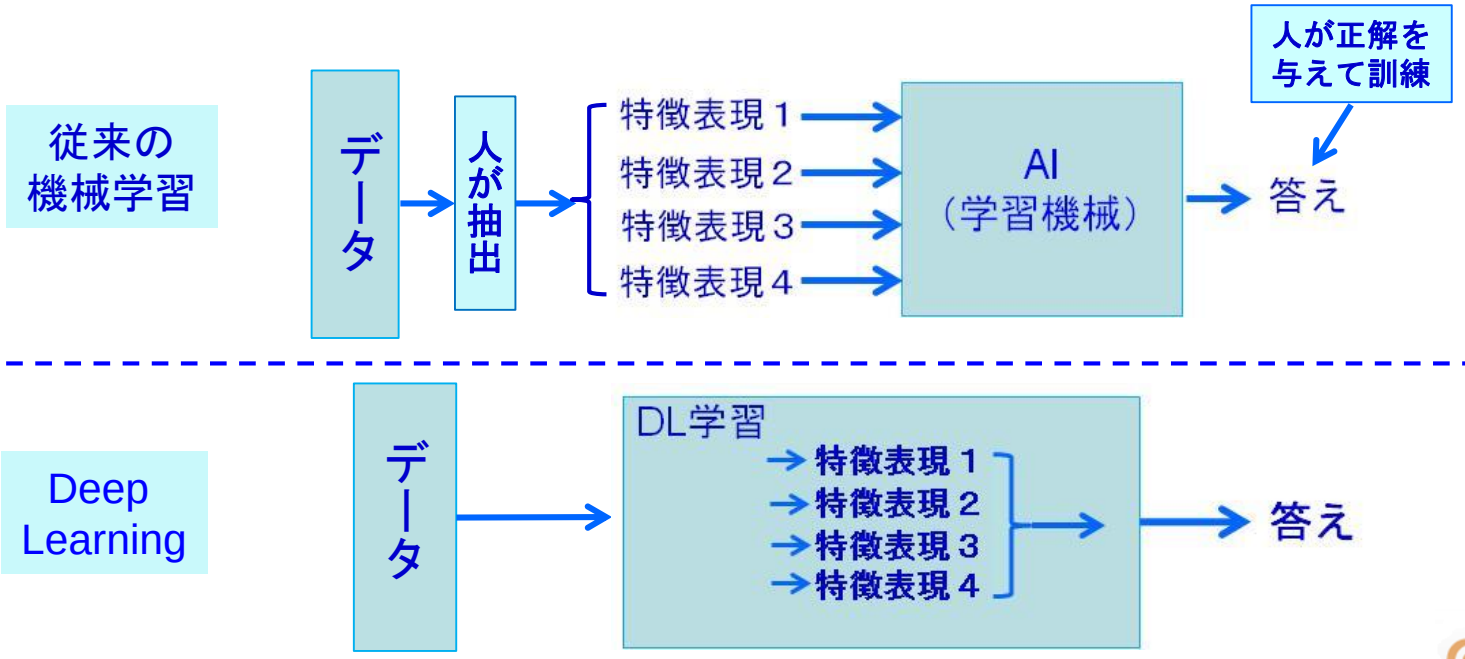
Back Propagation (1986　Rumelhart)
望ましい出力との誤差を教師信号として与える事により、次第に結合係数を変化させ、最終的に正しい出力が得られるようにする。
結合係数を変える事を学習と呼ぶ。
この学習方法には、最急降下法（勾配法）が使われる。出力層へ寄与の高いノードの重みが変更。

多層にわたる逆伝搬で修正感度減衰

TMDU

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
- 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

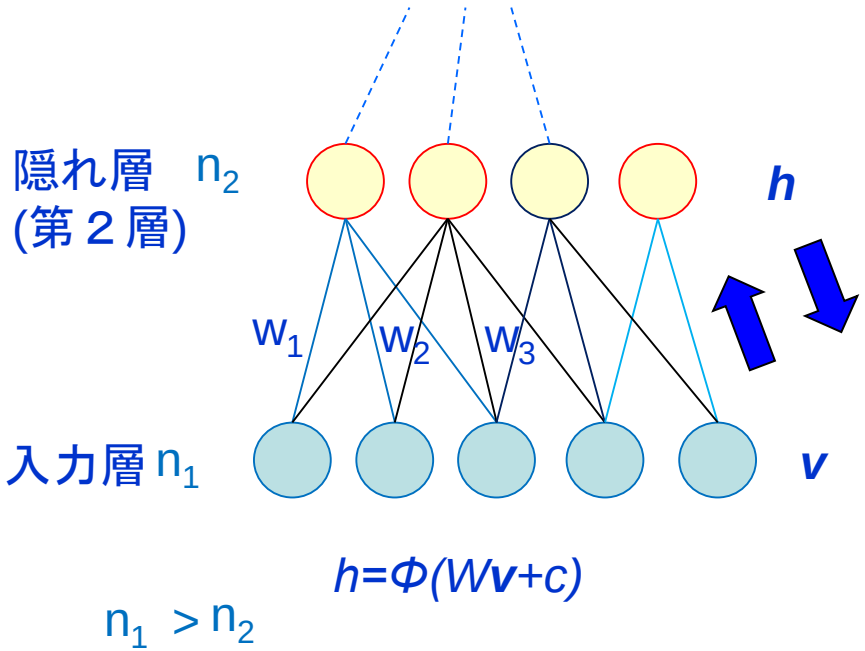


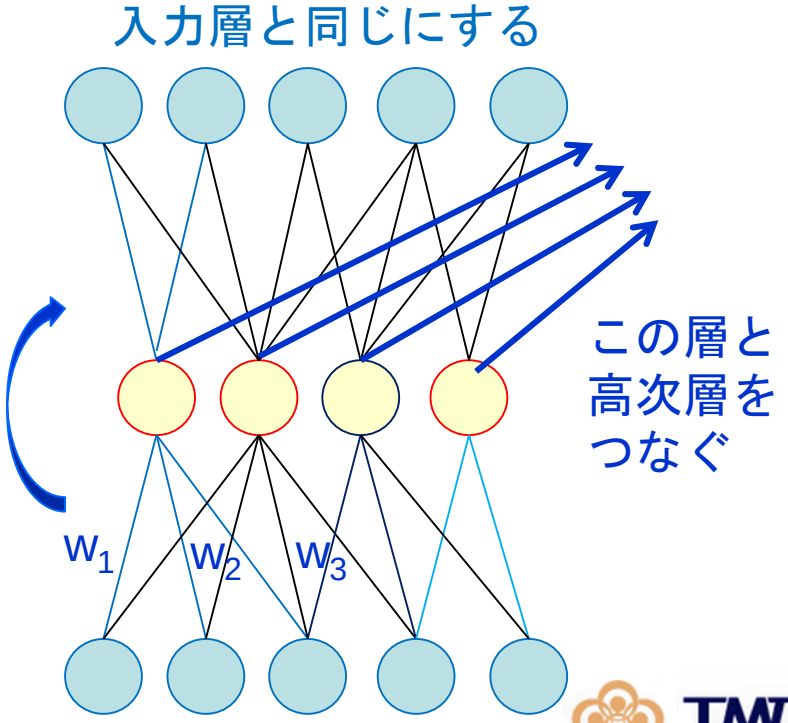
DLの革命点 Autoencode 1

- 対象に固有な内在的特徴を学ぶ自己符号化の原理
- 格段ごとに入力を少ない中間層を介して復元できるかを行なう
- 次元を圧縮されて可及的に復元する

→できるだけ復元に効果的な特徴量を探索する

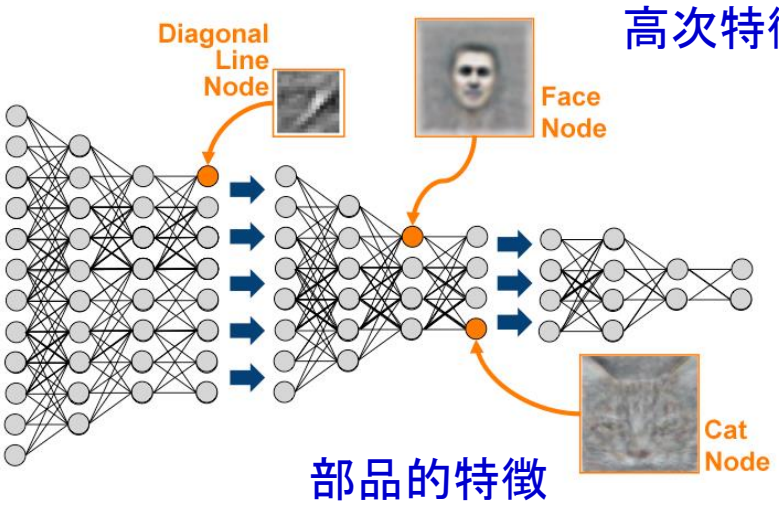
→内在的な特徴量を見出す

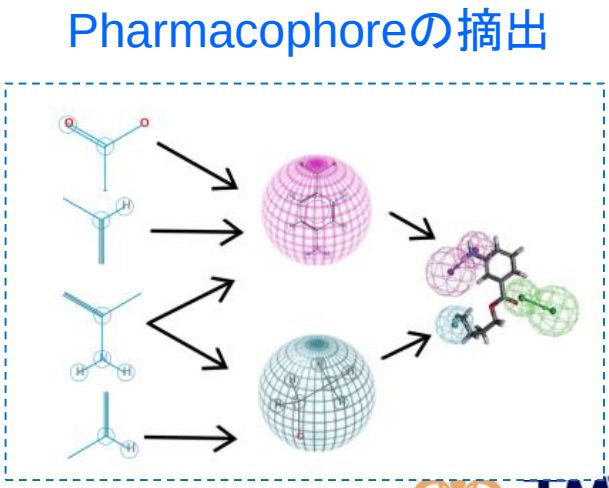




DLの革命点 Autoencode２

- 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**



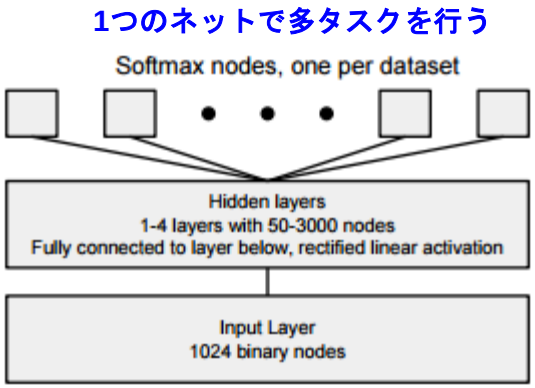


Deep Learningの創薬へ応用

Deep learning: 創薬からの注目

- **創薬を巡る状況**
  - 平均14年、約１000億円を超える費用
  - 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構造活性相関のデータを学習して構造から分子の生物学的活性を予測するモデルの開発コンテスト
  - 勝利したモデルは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"



Massively Multitask Networks

Artificial Intelligence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: SVM, k-nearest nb, 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の摘出

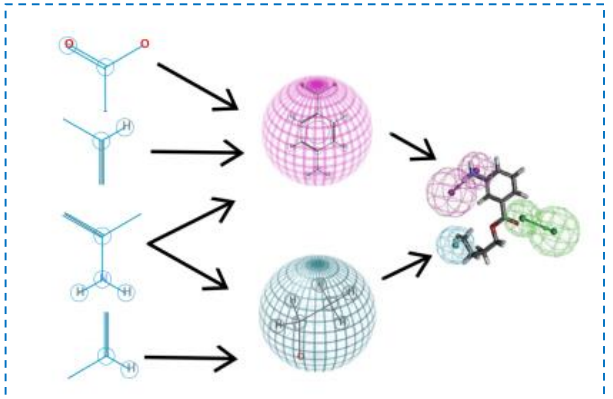


Figure Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

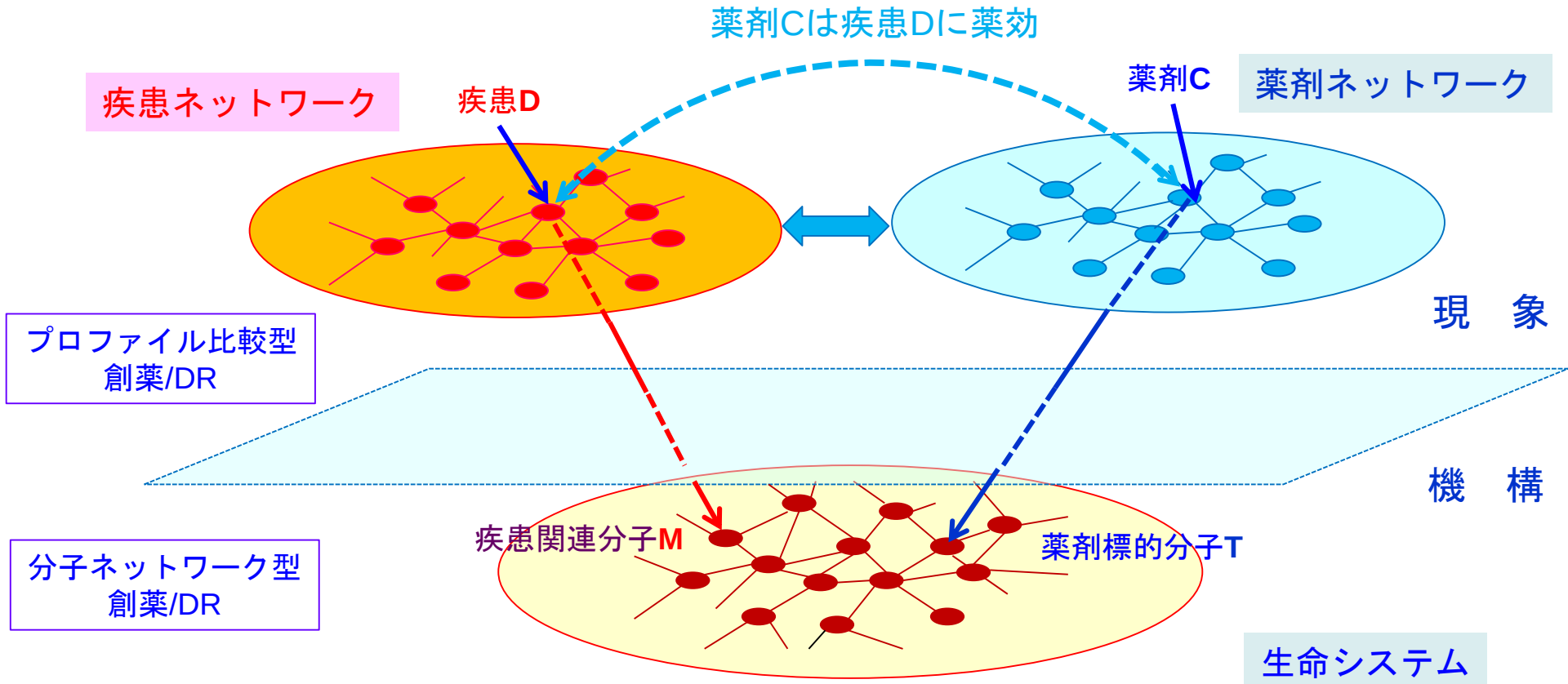
ECFP(chemical substructure: Enhanced cyan fluorescent)

我々の研究室での Deep Learningを用いた 創薬・DR 研究

長谷武志･辻真吾･田中博
東京医科歯科大学
次世代生命医学研究所

プロファイル型計算創薬の原理

3層生体・薬剤ネットワークのFramework



Feasibility Studyとしての従来の機械学習での試み

Deep Learningによる
多次元ネットワーク縮約法
(Hase,Tanaka 2017)

- 医療・創薬ビッグデータへの応用性高い
- 超多次元ネットワーク情報構造の急増
  - ゲノム医療＜網羅的分子情報ー臨床表現型情報＞
  - ゲノムコホートにおける＜遺伝子情報ー環境（生活様式）情報＞
- Deep Learning-based Network Contraction
  「DLネットワーク縮約法」

  超多次元ネットワーク情報構造⇒
  少数の特徴的ネットワーク基底に分解
- 線形分解ではない。非線形分解で基底への射影

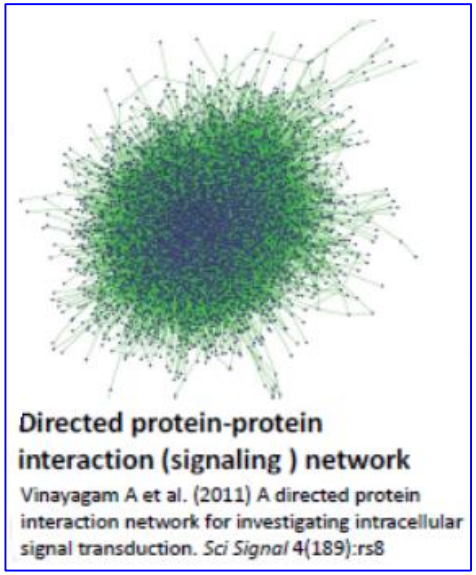
TMDU

特徴的ネットワーク基底への分解

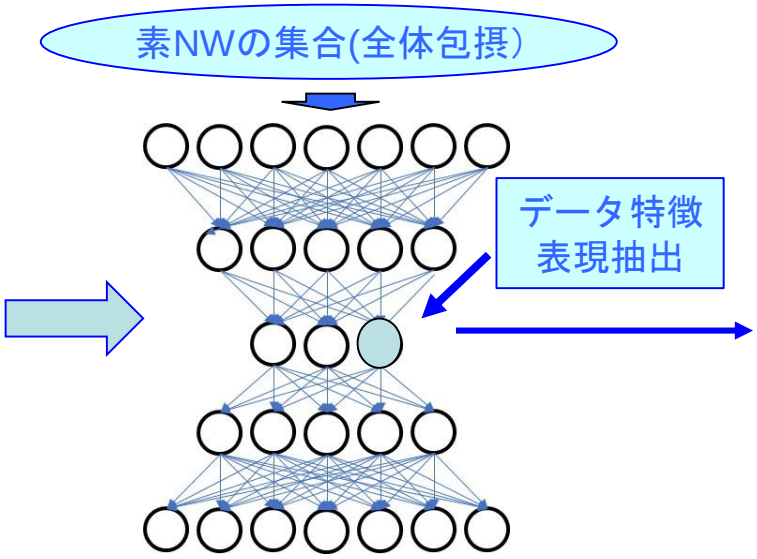
特徴的ネットワーク基底の和に縮約

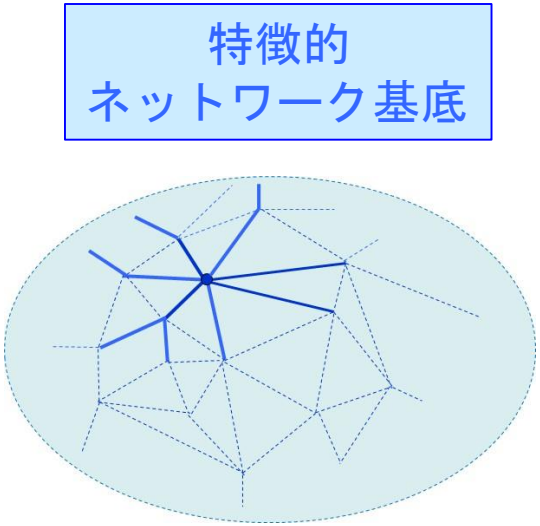
特定のノードを起点とした素NW（部分NW）の集合
全体NWを包摂する集合にDL反復自己学習

特徴的ネットワーク基底：トポロジーのみの構造/頻度構造



PPIネットワーク





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Stacked 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

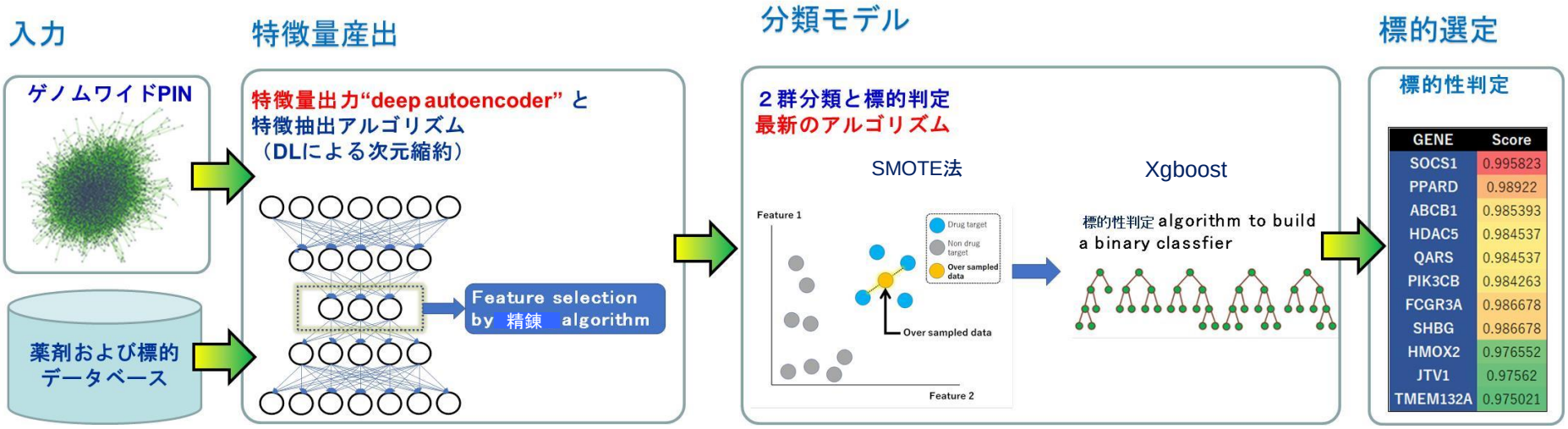
ネットワーク・タンパク質の標的性の判定

TMDU

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

＜疾患- 標的分子＞予測結果

Target disease	Top 25 genes for potential novel drug target for a target disease
Anti-Alzheimer's disease	***SOCS1***; …1A; HSPG2; PARD6A; BLNK; PRKCI; YES1; PPP3CA; MMP11; SPTAN1; PTPRC; FBLN1; RPLP1; FAM107A; TRADD; NR4A1; LAT; NF2; PRKCE; KIT; NID1
Anti-Anxiety	***S1PR1; CNR1; MTNR1A; CCL4; F2; TEC; IL8; CRHR1; AGTR2; OPRD1; IL8RA; RNF43; RHO***; SP6; RAB13; DRD4; IL8RB; MMP9; MMP2; OPR…M1; IL1B; GNAS; S1PR3; KIT; GRM2
Anti-Rheumatoid	***SLC22A5; GRASP; KIT; SLC22A4; CFH; …CO…; HSP90AA1; UBB; DHRS3; SCTR; ADORA…1; N…***; C6orf47; NR4A1
Anti-Breast Cancer	***SHC1; NFKB1; RELA; ID2; RAC1; SRC; RUNX…; HDAC2; I…***
Anti-Colorectal	…***HC1; GHR; HRAS; PLCG1; …CK; LYN***
Anti-Pancreatic	…***H2; DSCC1; TAF9; AIF1; …L2; NSMCE4A; ASPH***;
Anti-Melanoma	…RAP; HMGB2; FEN1; HNRNPA…; …NA; MUC1; LMNB1; GDE1; PSEN1; P…

SOCS1はJAK/STAT pathwayを介してサイトカインの応答を変動させ、中枢神経系の炎症を制御

Review

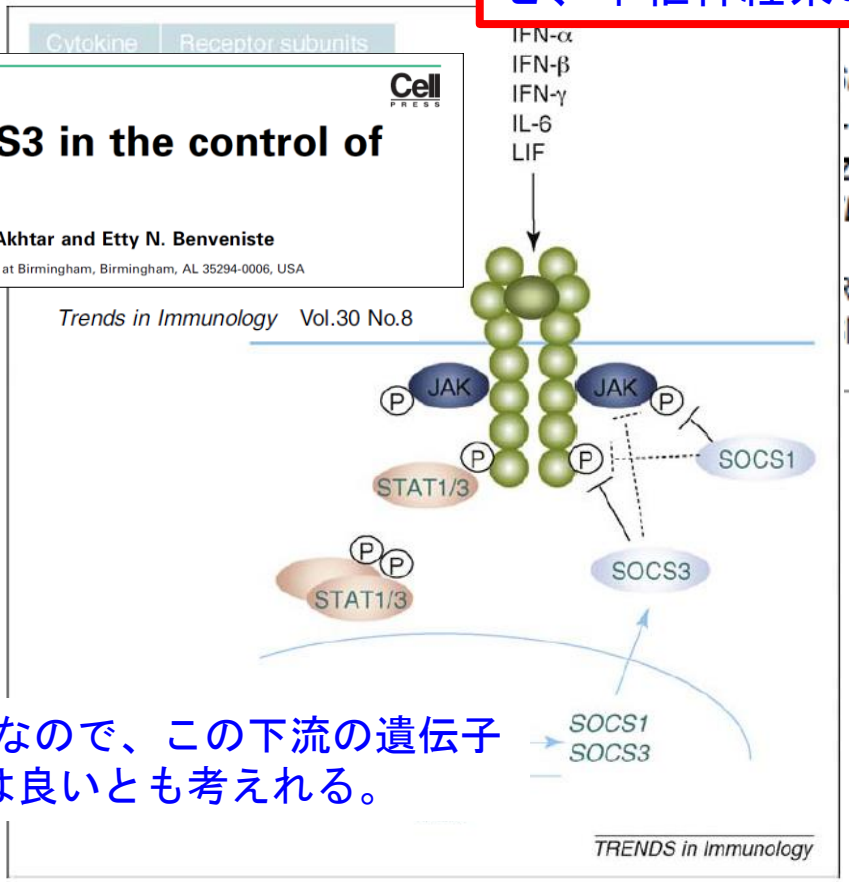
Cell PRESS

SOCS1 and SOCS3 in the control of CNS immunity

Brandi J. Baker, Lisa Nowoslawski Akhtar and Etty N. Benveniste

Department of Cell Biology, The University of Alabama at Birmingham, Birmingham, AL 35294-0006, USA

Trends in Immunology Vol.30 No.8



しかし、SOCS1は上流の遺伝子なので、この下流の遺伝子を標的にした方が、長期投与には良いとも考えれる。

TMDU

疾患-標的分子リンクの同定よりDRへ

機械学習で予測された、新規標的の情報(disease A とtargetの情報,標的がdisease Aの新規標的分子、青いリンク)を、既知のdrug-target-disease interaction network をマップし薬剤の新しい適用疾患（赤リンク）を予測

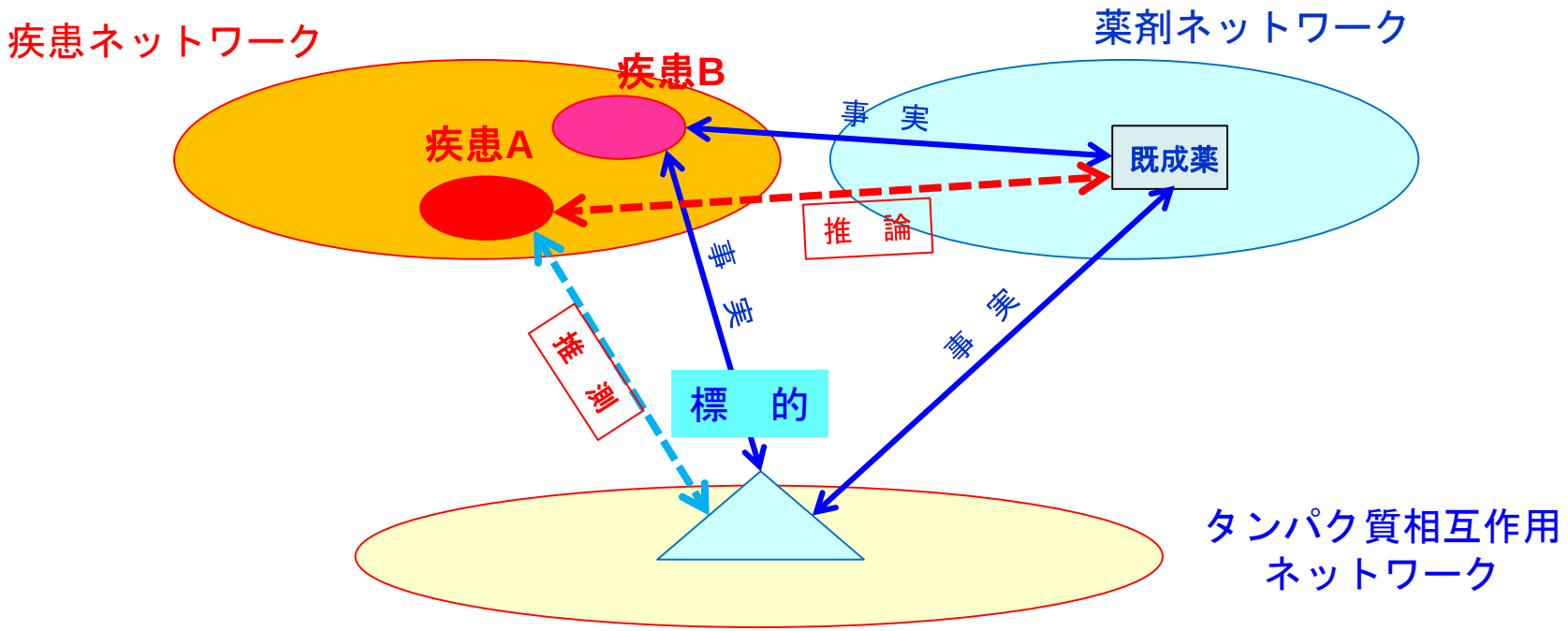


Table 3. Predicted repositionable drug candidates

Target disease	Candidate repositionable drug
Anti-Alzheimer's disease	Imatinib; Marimasta[t]; Nilotinib; Regorafenib; Sorafenib; Tamoxifen; Urokinase
Anti-Anxiety	3-Methylfentanyl; Agome[p]atine; Amitriptyline; Amoxapine; Antihemophilic Factor; Apomorphine; Aripiprazole; Bromocripti[n]e; Buprenorphine; Butorphanol; Cabergoline; Canakinumab; Captopril; Chlorpromazine; Coagula[t]ion Factor IX; Codeine; Dextromethorphan; Dextropropoxyphene; Dopamine; Drotrecogin alfa; Ethylmo[r]phine; Etorphine; Fentanyl; Halothane; Hirulog; Hydrocodone; Hydromorphone; Imatinib; Ketamine; Ketobemidone; L-DOPA; Lepirudin; Levallorphan; Levorphanol; Lisuride; Loperamide; Lo[x]apine; Marimastat; Melatonin; Menadione; Methadone; Methadyl Acetate; Methotrimeprazine; Mino[c]ycline; Morphine; Naloxone; Naltrexone; Nilotinib; Olanzapine; Ondansetron; Oxycodone; Oxymorpho[n]e; Paliperidone; Pergolide; Pethidine; Pramipexole; Promazine; Propiomazine; Quetiapine; Regorafenib; Remifentanil; Remoxipride; Risperidone; Ropinirole; Rotigotine; Sorafenib; Sufentanil; Suramin; Thi[e]thylperazine; Ziprasidone
Anti-Rheumatoid	Acetylcholine; Adenosine; Amiloride; Aminohippurate; Aminophylline; Amphetamine; Ampicillin; Azidocillin; Benzylpenicil[l]in; Cefalotin; Cefdinir; Cefixime; Cephalexin; Choline; Cimetidine; Clonidine; Cyclacillin; Desipramine; Diphenhydramine; Dopamine; Dyphylline; Enprofylline; Epinephrine; Furosemide; Grepafloxa[c]in; Histamine Phosphate; Imatinib; Imipramine; Insulin, isophane; Ipratropium bromide; L-Arginine; L-C[a]rnitine; Levofloxacin; Lidocaine; Liothyronine; Lomefloxacin; Mepyramine; Methamphetamine; Niac[i]n; Nicotine; Nilotinib; Norepinephrine; Norfloxacin; Ofloxacin; Oxtriphylline; Pentoxifylline; Probene[c]id; Procainamide; Quinidine; Quinine; Regorafenib; Rifabutin; Secretin; Sorafenib; Spermine; Te[o]adenoson; Testosterone; Tetraethylammonium; Theophylline; Thiamine;

Neuroscience 304 (2015) 316–327

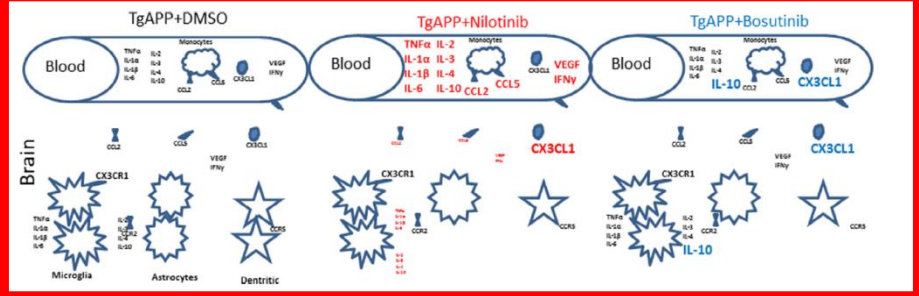
慢性白血病の抗がん剤であるニロチニブがアルツハイマーのDR薬剤として選定

NILOTINIB AND BOSUTINIB MODULATE PRE-PLAQUE ALTERATIONS OF BLOOD IMMUNE MARKERS AND NEURO-INFLAMMATION IN ALZHEIMER'S DISEASE MODELS

I. LONSKAYA,[a] M. L. HEBRON,[a] S. T. SELBY,[a] R. S. TURNER[b] AND C. E.-H. MOUSSA[a]*

[a] *Department of Neurology, Laboratory for Dementia and Parkinsonism, Georgetown University Medical Center, Washington D.C. 20007, USA*

[b] *Department of Neurology, Memory Disorders Program, Georgetown University Medical Center, Washington D.C. 20007, USA*



DL型NNへの期待と困難点

- 医療・創薬の応用は開始段階で応用成功例は少ない
  - 本質的に「教師なし学習」:人間が思いつかない解を提示
  - 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が期待される
    - 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
- DL型ニューラルネットの困難点
  - 特徴表現を自己学習するが基本的にはBlack Box
  - 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。

今後の戦略・方向

- 第２世代のゲノム医療・創薬
- Deep Learningによる〈多次元ネットワーク情報構造〉の縮約
  - 創薬だけでなく、ビッグデータ医療への適応可能
  - ゲノム医療の〈網羅的分子情報ー臨床表現型〉の相関ネットワーク構造
  - バイオバンクの〈遺伝素因ー環境要因〉と発症
- AI創薬の「枠組み」実行方向は「見えてきた」
- 本年中に、いよいよAI創薬の実装に着手しなければならない。米国に持って行かれる。
  - 製薬企業、IT企業、医療機関を束ねた集中的プロジェクトを推進するために「ビッグデータ医療・AI創薬コンソーシアム」を設立する

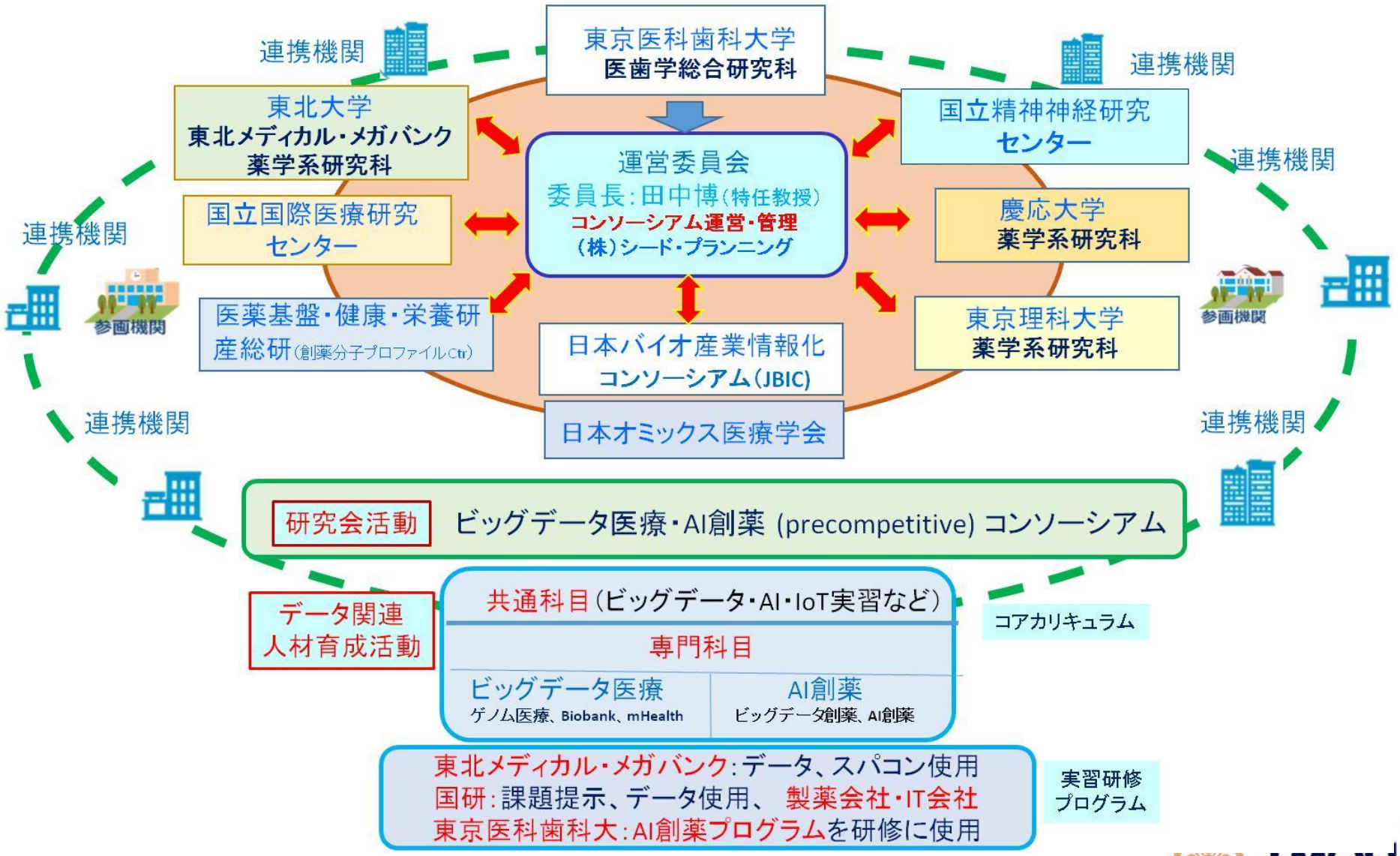
TMDU

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



ご清聴有難う御座いました

ビッグデータ解析に向けた2つの人工知能（AI）方法の適用

- 数理的知識処理：データマイニング、探索
  統計学の数理的枠内で次元縮約

  ⇒　スパース推定による従来手法の
  次元落ちの正則化
- ニューロネットワーク：Deep Learningによる特徴量抽出を用いた次元縮約

  ⇒Deep LearningのAutoEncode機能
  を用いた実質的な独立次元抽出に
  基いた解析・予測

TMDU

数理的知識処理
スパース推定による次元落ちの正則化

従来の重回帰分析

$\mathbf{x} = (x_1, \ldots, x_p)$ と目的変数 y に関して n 組のデータ $\{(y_i, \mathbf{x}_i); i = 1, \ldots, n\}$

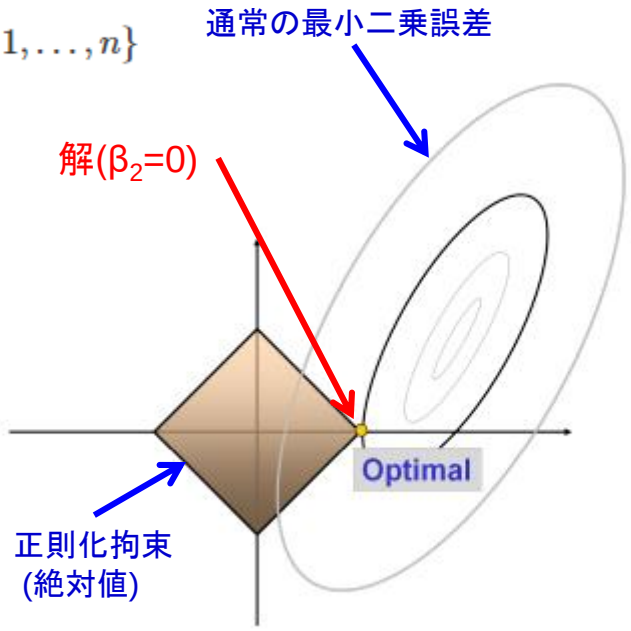
$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Lasso(L_1型正則化重回帰分析)

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2, \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le t.$$

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

通常の最小二乗　　正則化項 (絶対値)



寄与の低いβj は0になる⇒変数選択と次元落ち正則化が同時に達成できる

様々な変法：Larsアルゴリズム(λを∞から減少), elastic net, adaptive lasso, grouped lasso

TMDU

様々なスパース正則化の利用

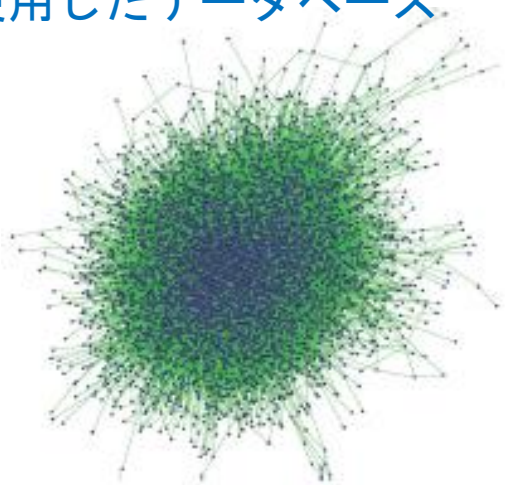
- GWASへの応用

  GWASにおけるgene-gene interactionの取り込み(主効果と相互作用）
  - Correlated SNPs (Ayers and Cordell, 2010)
  - More power while having a lower false-discovery rate (FDR) （He and Lin, 2011）
  - Pathwayに含まれているSNP間だけ相互作用を認める（Lu, Latourelle、2013)
- 遺伝子発現プロファイルへの応用
  - Biomarker（差異的発現遺伝子）が明確化
- 主成分分析にスパース正則化
  - 主成分の解釈が容易になる
  - 次を最小化 $$Q_\lambda(v_1, X) = \frac{1}{2}\mathrm{trace}[(X - z_1 v_1^T)^T (X - z_1 v_1^T)] + \sum_{j=1}^{p} p_\lambda(|v_{1j}|),$$
- 判別分析でも正則化により次元縮約

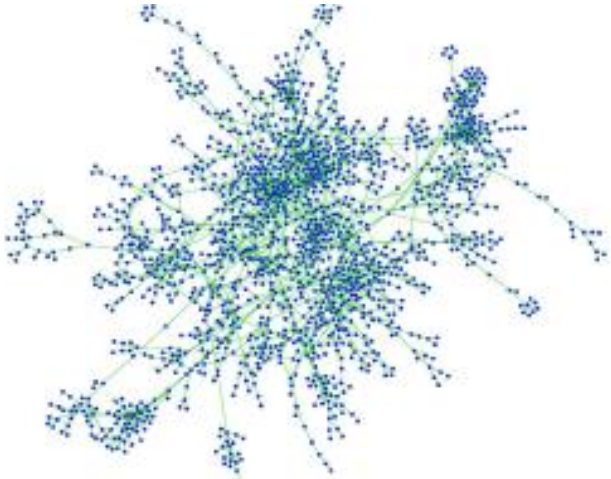
TMDU

生体ネットワークの3種類のタンパク質ネットワーク

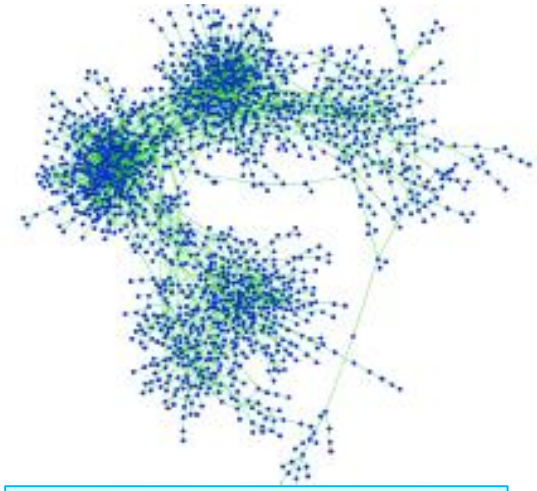
使用したデータベース



タンパク質相互作用ネットワーク



合成用量致死ネットワーク



合成致死ネットワーク

３種のネットワークと、薬剤と標的分子の情報を、訓練データ・試験データの構築に用いた。

Step 1：ネットワーク解析を行い、機械学習モデルを生成するための特徴量を抽出する。

Step 2：機械学習を用いて対象疾患（疾患Ａ）に対する新規標的分子を予測する

Step 3：予測された新規標的分子をベースにして、この疾患Ａに対する新しいDRの薬剤を推測する

TMDU

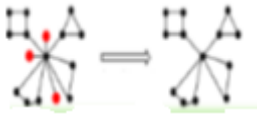
Step 1：Network解析による各遺伝子の特徴量の抽出

そのノードのネットワーク特徴量を表す21種のネットワーク統計量を、各遺伝子に対して、各ネットワークを解析して算出した合計63種のネットワーク統計量を各遺伝子の特徴量として予測モデルの構築に用いた。

1．**次数中心性**（他のタンパク質との結合次数が多い）

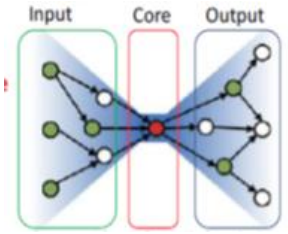
2．**近接中心性**

他の全ノードとの距離の平均、ネットワークの中心に位置する程小さい、すなわち他の位置ノードと全般的に「近い」



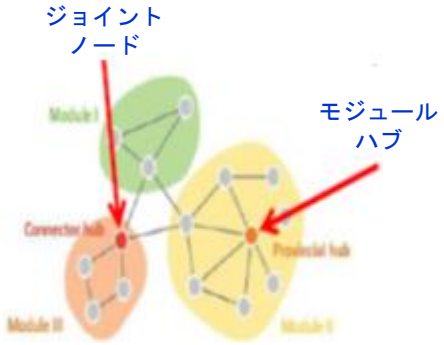
3．**媒介中心性**

ネットワーク内の各ノード間を繋ぐ様々なパスが対象となるタンパク質をどれだけの頻度で通過するか、通過する頻度の割合が高い程、中心性がある



4．**蝶ネクタイ構造指標**：

ネットワークに様々なノードから入力信号が入ってくると、中心的なコアで一旦集約されて様々な遺伝子へと拡散する構造（bow-tie)においてコア部にいるか、両末端部にいるかを表す指標である。



5. モジュール指標

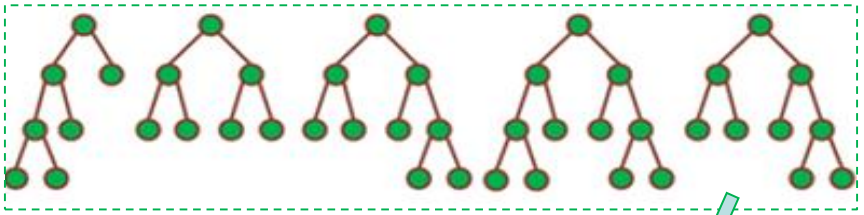
ネットワークが幾つかのモジュールに分かれる（積み木方式）ときモジュール間を繋ぐノードか、各モジュールのハブかなどの指標である。

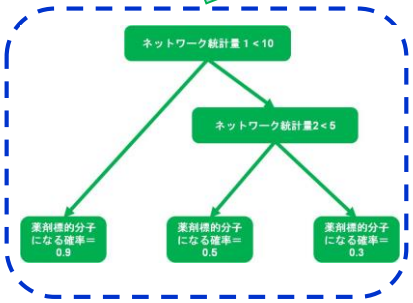
TMDU

Step 2 機械学習モデルによる決定木

Random Forest

**63 個の特徴量からランダムに
いくつかの特徴量を選んで判別
ルールの決定木を1000個作成**





遺伝子A
ネットワーク特徴量1=0.5
ネットワーク特徴量2=1.3
. . .

63種の特徴量

遺伝子B
ネットワーク特徴量1=2.3
ネットワーク特徴量2=4.0
. . .

1000個
決定木

薬剤標的でない
遺伝子A

薬剤標的である
遺伝子B

ビッグデータと機械学習

- **The ASCO (米国臨床癌学) CancerLinQ initiative**
  - 診療の現場(EHR)から大量の診療データを集め分析
  - 新しい臨床治験へのガイドライン作成
  - 17万人のがん症例データベースを構築。各がん１～２万人の症例を集める
  - 学習システムを構築し治療知識を統計学習、ニューロネットを駆使して学習。
    BigDataにおけるLearning systemの不可欠性
  - 2013年に、CancerLinQのプロトタイプを完成、10万人以上の乳がんを蓄積、完全規模へ継続構築中
- **IBM Watsonのがんセンターへの普及**
  - 質問・応答（QA)システム、知識探索、ライト・オントロジー
  - Memorial Sloan-Kettering Cancer Center（MSKCC）などと共同
  - Watsonを母体に**The Oncology Expert Adviser software** (OEA)開発
  - 他にNew York Genome Centerとglioblastoma (グリア芽細胞腫）知識生成
- **Cancer Commons initiative**
  - Rapid learningのインフラ整備
  - 目的：患者の個別症例と最新の知識を更新
  - 個々の患者の”Donate Your Data”(DYD)登録
- Google X project, “Human Longevity Inc.”

IBM Watson Learning Big Data

TMDU