

AI創薬に期待すること

東京医科歯科大学 医療データ科学
東北大学 東北メディカル・メガバンク機構

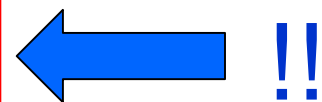
田中 博



医療ビッグデータ時代

- (1) 次世代シーケンサ (Clinical Sequencing)による「ゲノム/オミックス医療」における網羅的分子情報収集/蓄積
- (2) Biobank/ゲノムコホート普及による分子・環境情報の蓄積
- (3) モバイルヘルス(mHealth) によるWearable センサの連続計測による生理データの蓄積 (unobstructed monitoring)

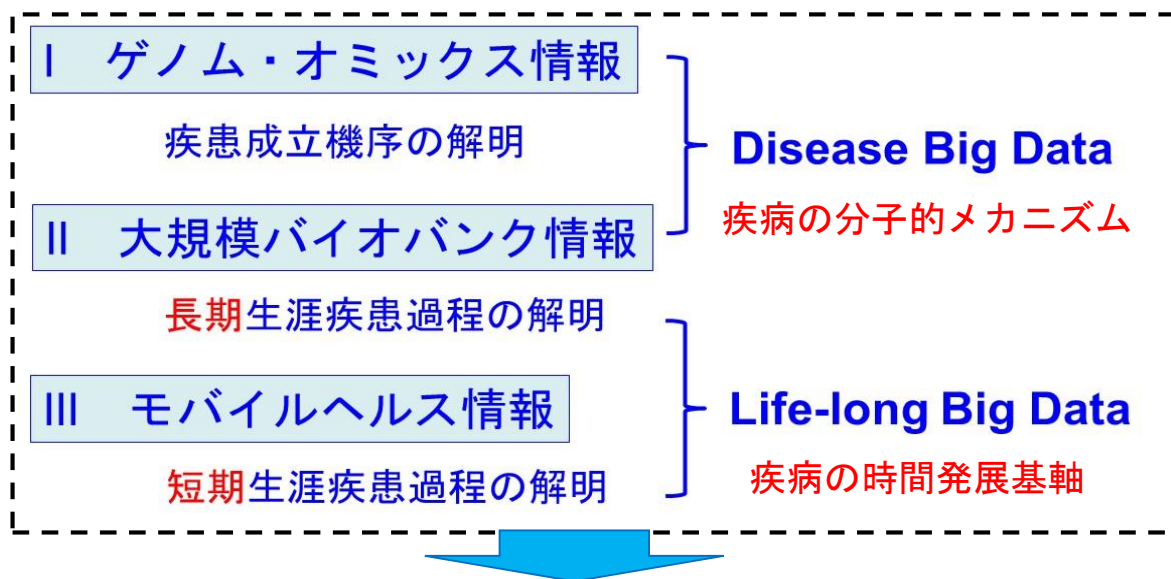
急激な大量データの出現
コストレス化かつ高精度化



ゲノム : 13年→1日(1/5000) 3500億→10万円(1/350万)

医学・医療のパラダイム変革
医学/創薬へ利用可能なデータの飛躍的な増大

ビッグデータが医療のパラダイム変換を可能にする



従来の医学・医療

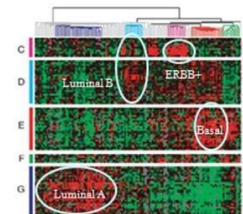
新時代の医学・医療

Population医療



個別化医療

内在的サブタイプ
(乳がん)

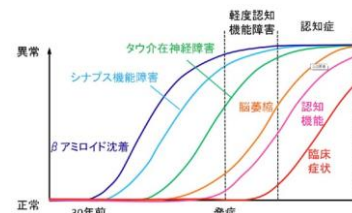


Reactive 医療



Proactive 医療

先制医療
(アルツハイマー)



Occasional 医療



Life-long 医療

計算創薬・DRの基本枠組みと
生体分子プロファイル型
創薬・DR

創薬をめぐる状況と解決の方向

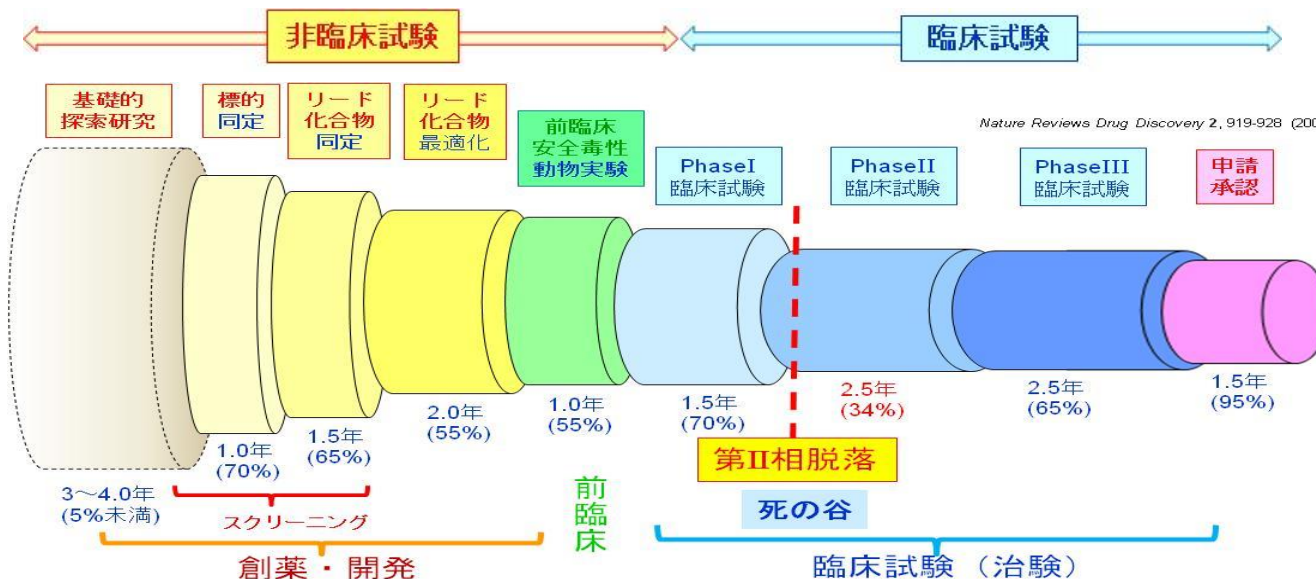
- 医薬品の開発費の増大
 - 1 医薬品を上市するのに約1000億円以上
- 開発成功率の減少
 - 2万~3万分の1の成功率
 - とくに**非臨床試験**から**臨床試験**への間隙
 - **phase II attrition** (第2相脱落)
- 臨床的予測性
 - 医薬品開発過程の**できるだけ早い段階**での**有効性・毒性の予測**
- **臨床予測性の早期での実施**
 - 罹患者のiPS細胞を使う

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



ドラッグ・リポジショニング (DR)

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。

生体プロファイル型計算創薬・DR

計算創薬(*in silico*創薬)の新しい方向

薬剤
(ポケット内)

これまでの計算創薬

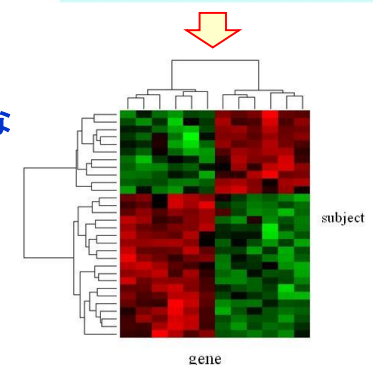
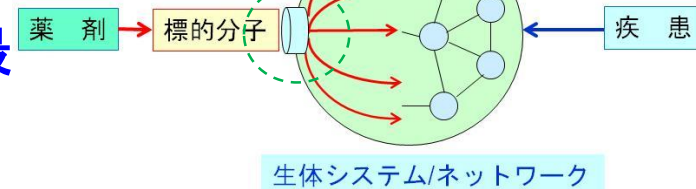
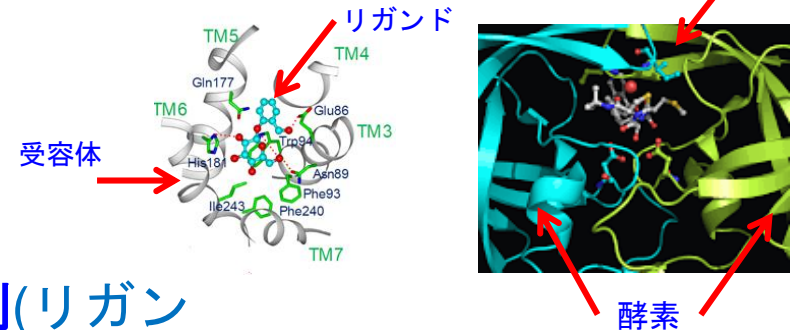
分子(結合構造)中心

- 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
- 標的に結合するリード化合物・構造最適化

— 成功例：インフルエンザ薬 タミフル

生体プロファイル型創薬

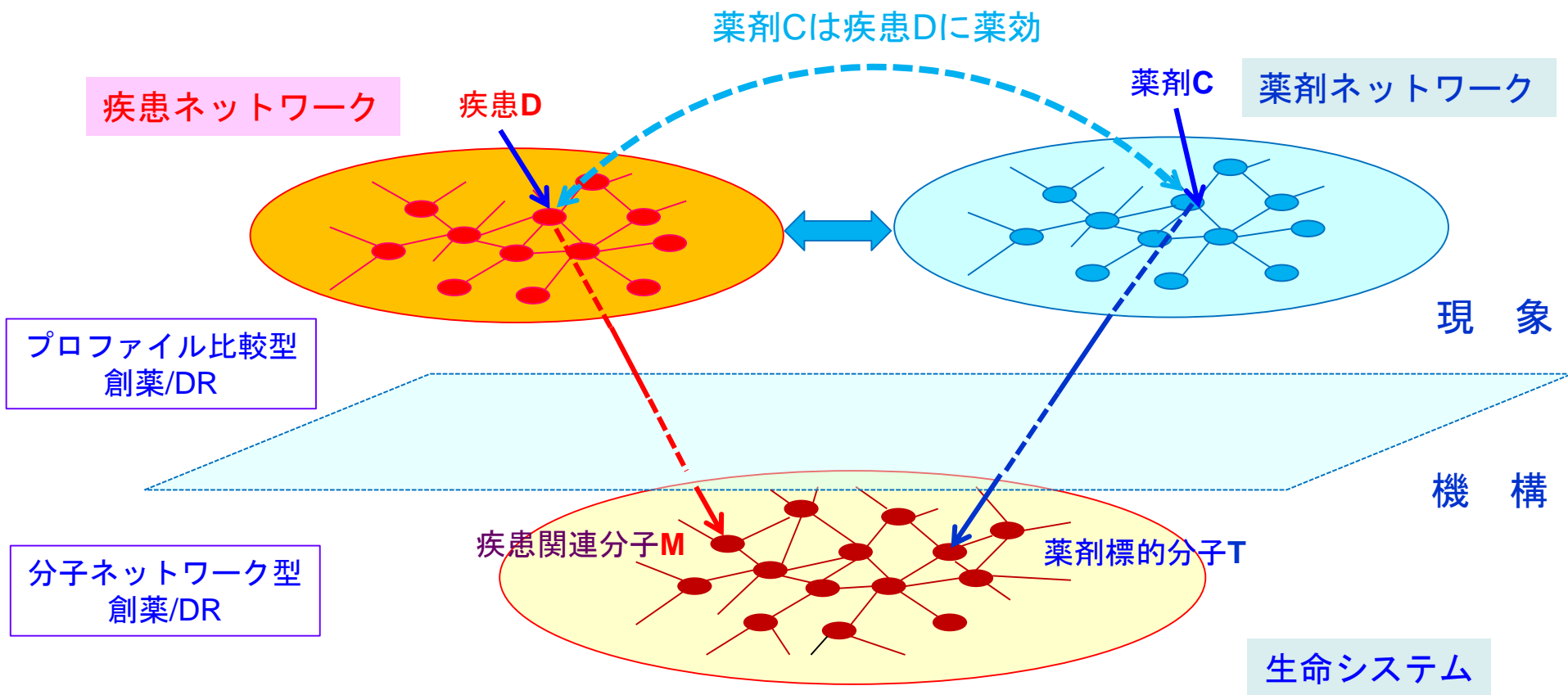
- 結合後の生体システムのゲノムワイドな反応/振舞いに注目する
- 化合物の分子構造と生体活性の関係には両者の間には生体システムがある



遺伝子発現プロファイル変化
(疾患特異的/薬剤特異的)

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



計算創薬・DRの体系

計算創薬・DRの「非学習的」アプローチ

「ビッグデータ創薬・DR」

— 疾患－薬剤プロファイル直接比較

- 「現象論的」アプローチ
- 疾患罹患時と薬剤投与時の遺伝子発現プロファイル比較

— 生体分子ネットワーク準拠近接解析

- 「機構論的」アプローチ
- 疾患ネットワーク上の比較
- タンパク質相互作用ネットワーク疾患遺伝子と標的分子
- 信号伝達・転写制御ネットワークによるシステム薬理学

計算創薬・DRの「学習的」アプローチ

「AI創薬」

— Virtual Screeningへの人工知能・機械学習の応用

- Ligand-based AIバーチャルス・クリーニング
- Structure-based AIバーチャル・スクリーニング

— 標的分子探索に人工知能を用いた方法

- Hase－Tanakaの多層Deep AutoEncoderを用いた標的分子探索法

— 化合物の人工知能を用いた自動設計

ビッグデータ創薬・DR (非学習型アプローチ)

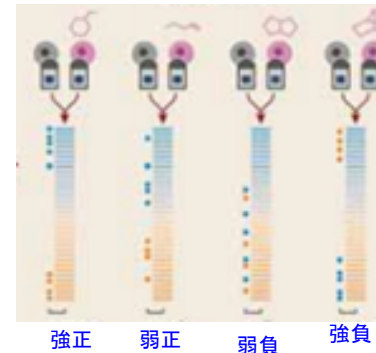
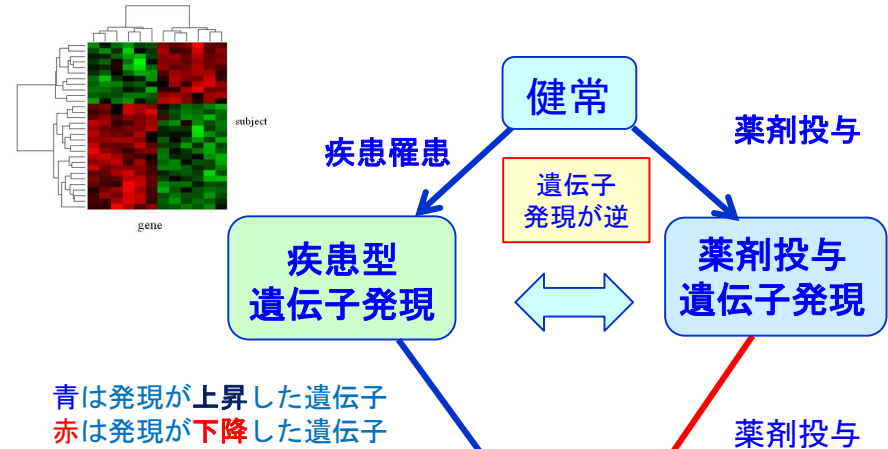
「疾患－薬剤ネットワーク」相互関連型

1. 遺伝子発現プロファイル 直接比較型

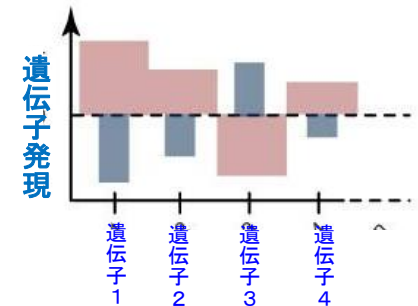
遺伝子発現プロファイルによる有効性予測

遺伝子発現シグネチャ逆位法

- 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」
- 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)
- 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**」
- 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
- 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



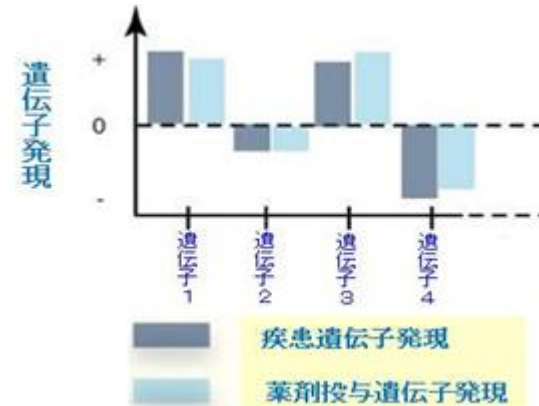
DBは薬剤遺伝子発現
横の点は疾患遺伝子発現



■	Disease signature	疾患遺伝子発現
■	Drug response sign	薬剤投与遺伝子発現

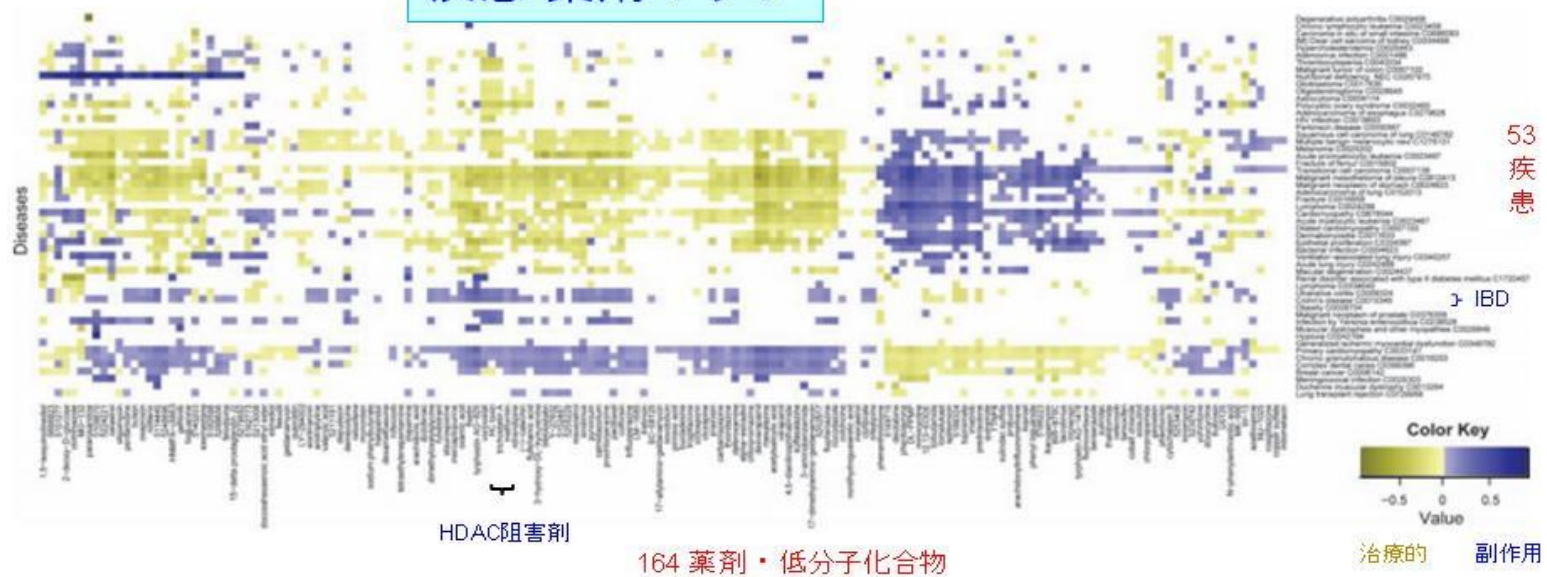
遺伝子発現プロファイルによる毒性予測

- 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



遺伝子発現Profiling による疾患－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－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 - 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

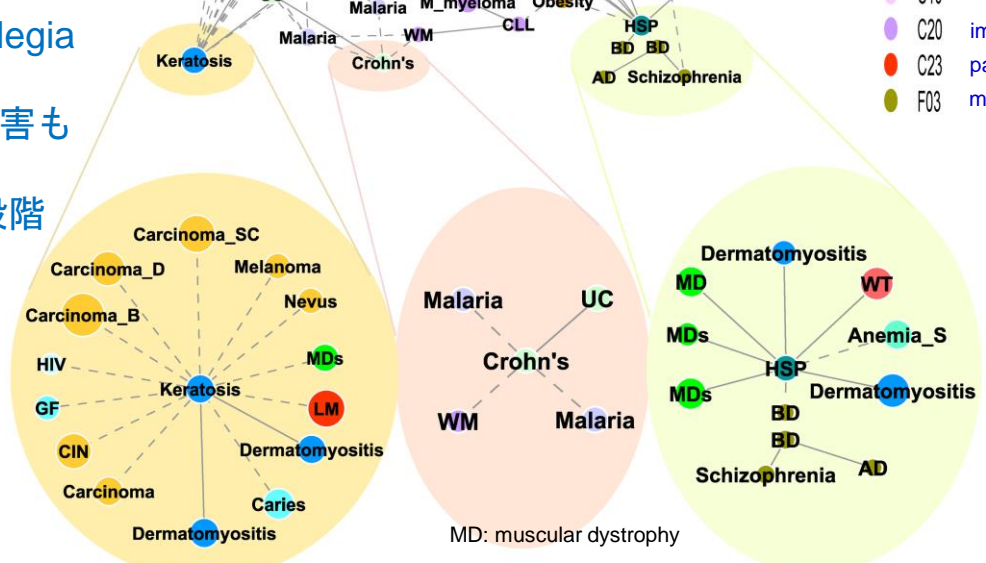
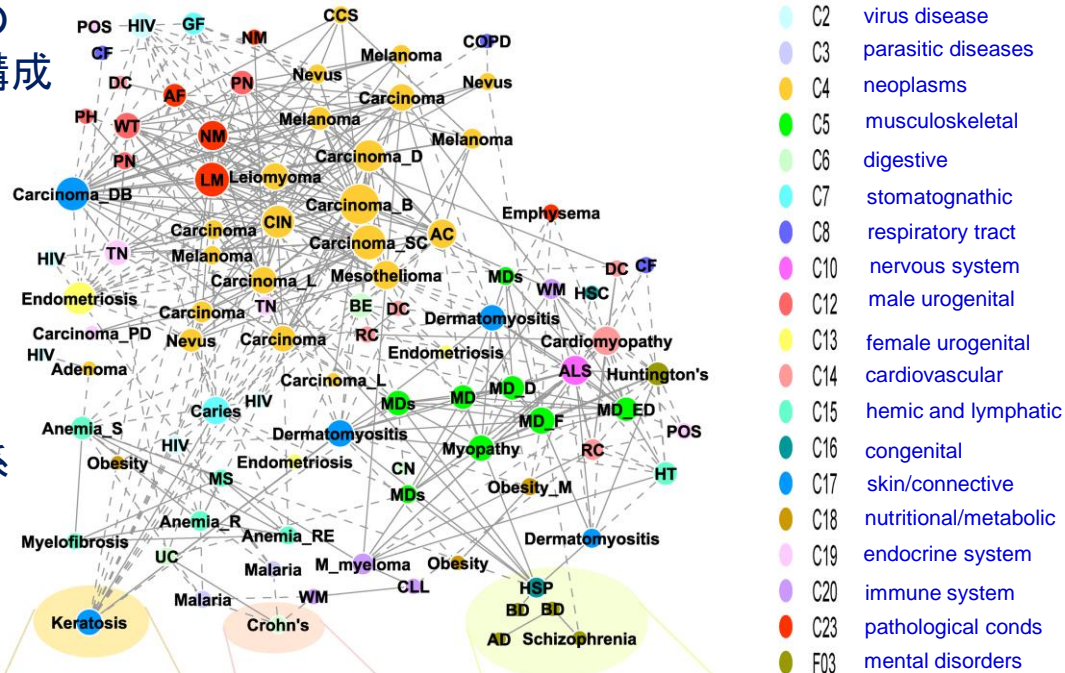
⇒ cancer(squamous) --前癌段階

<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



カラーはMeSH
 同一カテゴリー
 実線はMeSH内
 破線はMeSH外

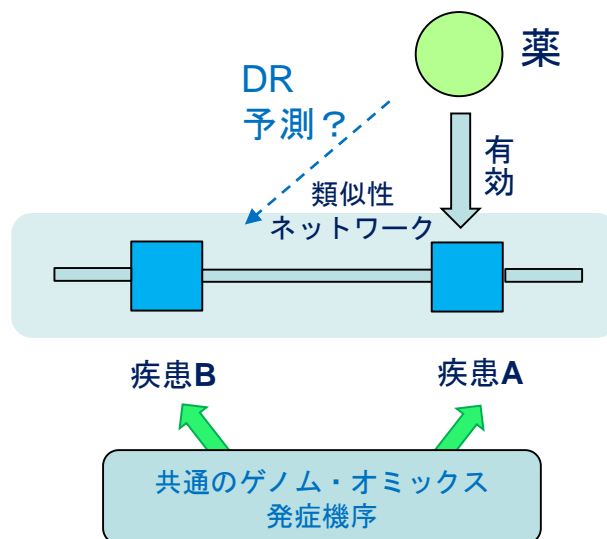
MD: muscular dystrophy

2. 疾患・薬剤ネットワーク 近接解析

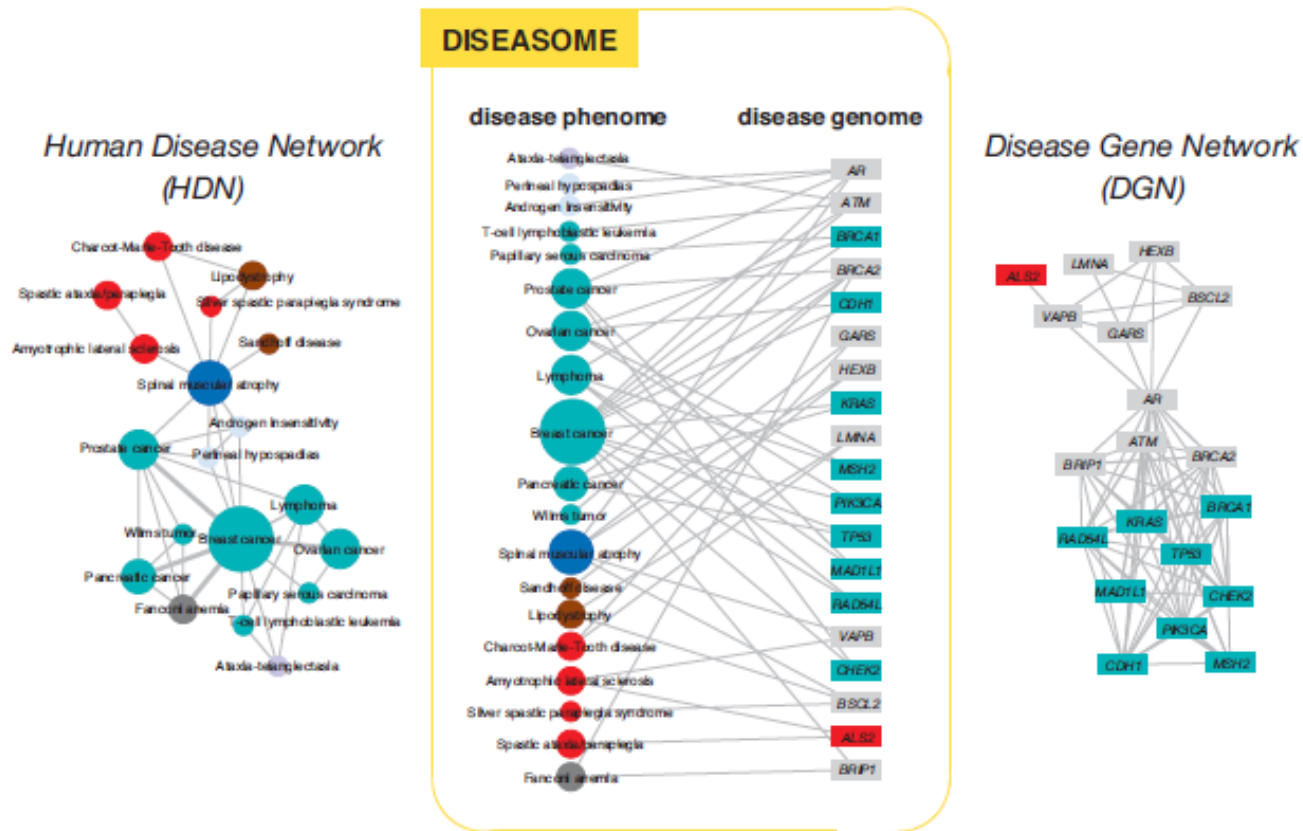
ビッグデータ創薬/DR

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 疾患原因遺伝子準拠 Diseasome(Goh,Barabasi et al.)



1つ以上の疾患関連遺伝子を共有する疾患

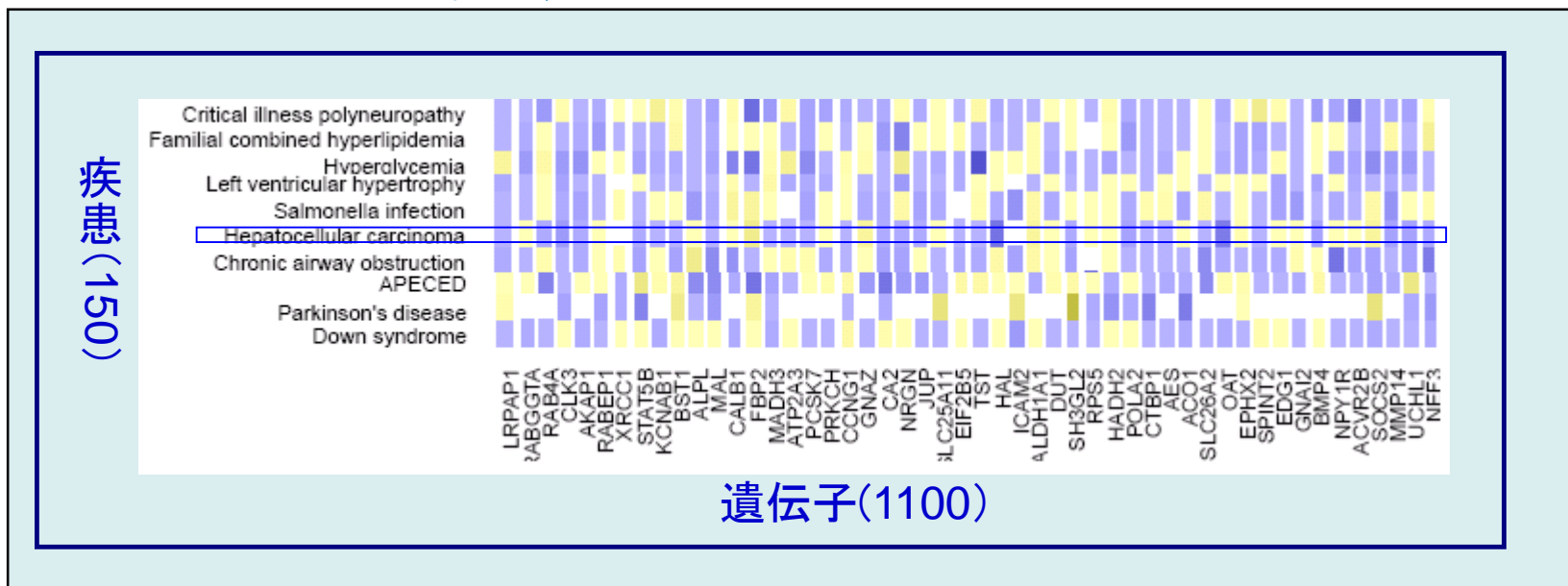
1つ以上の疾患を共有する疾患関連遺伝子

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



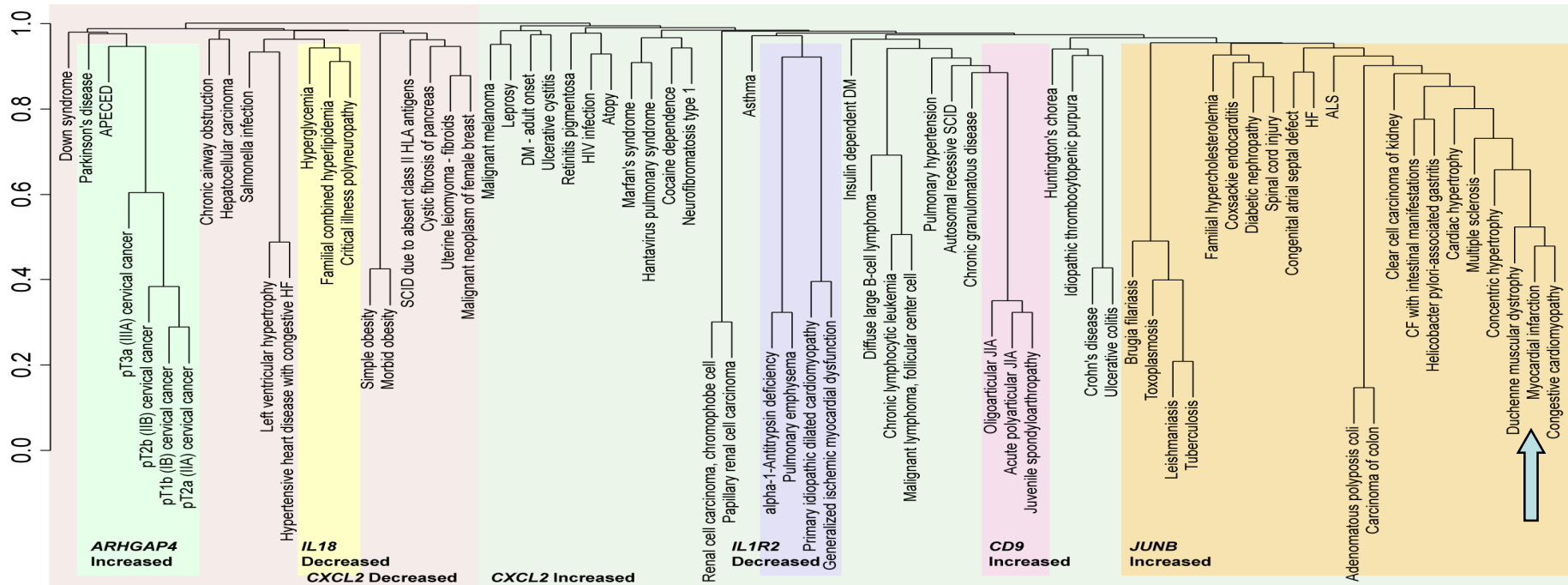
第2世代型 遺伝子発現プロファイ型 GENOMED (A.Butte et al)

- 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



ビッグデータ創薬・DR (非学習型アプローチ)

生体分子ネットワーク準拠型

3層の生体・薬剤のネットワーク間の関係図式

薬剤ネットワーク

薬剤Cは疾患Dに薬効

現象

機構

生命システム



疾患ネットワーク

プロファイル比較型
創薬/DR

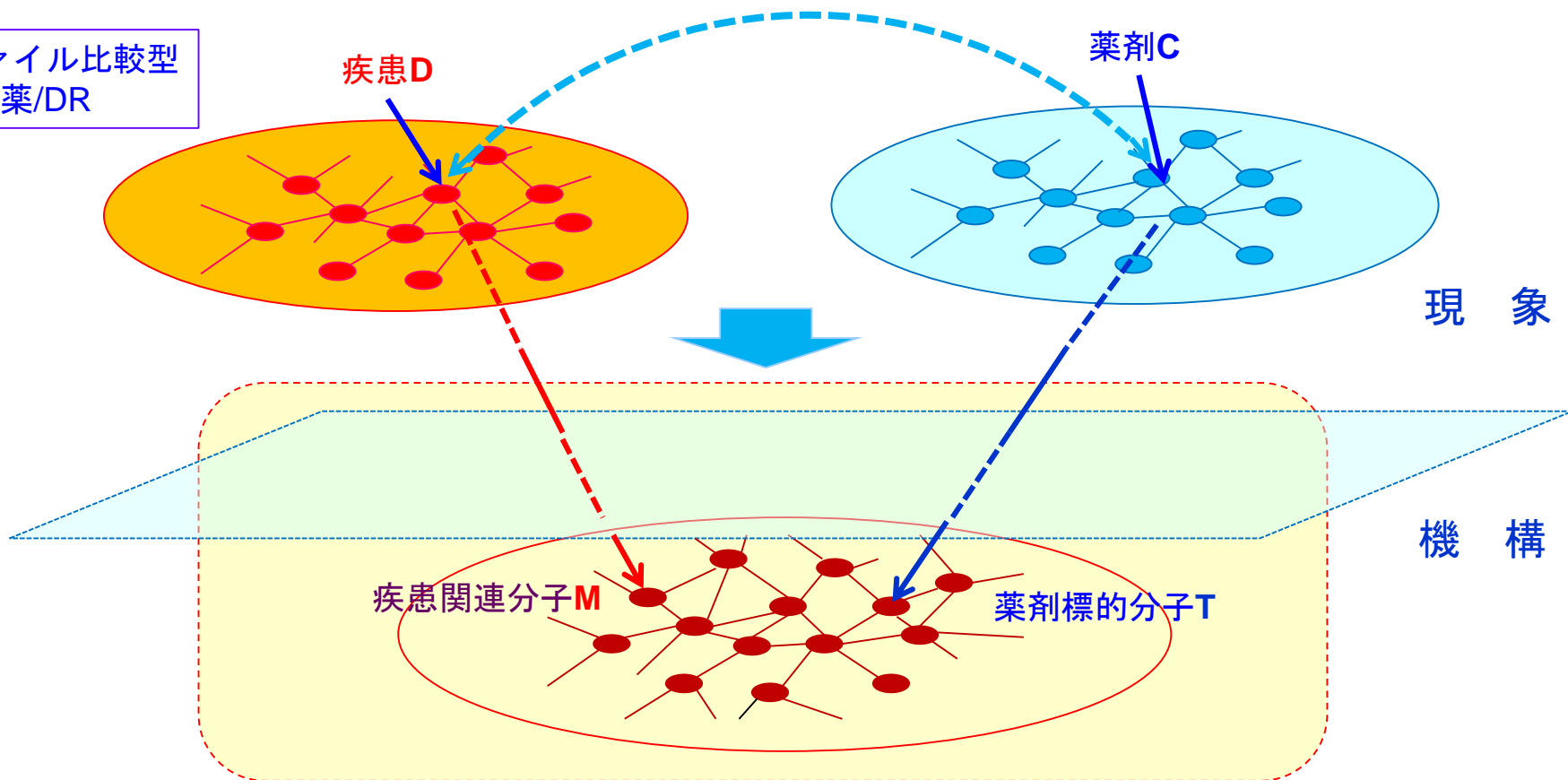
疾患D

薬剤C

疾患関連分子M

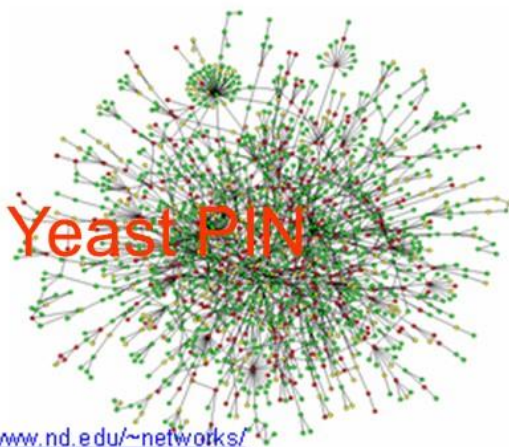
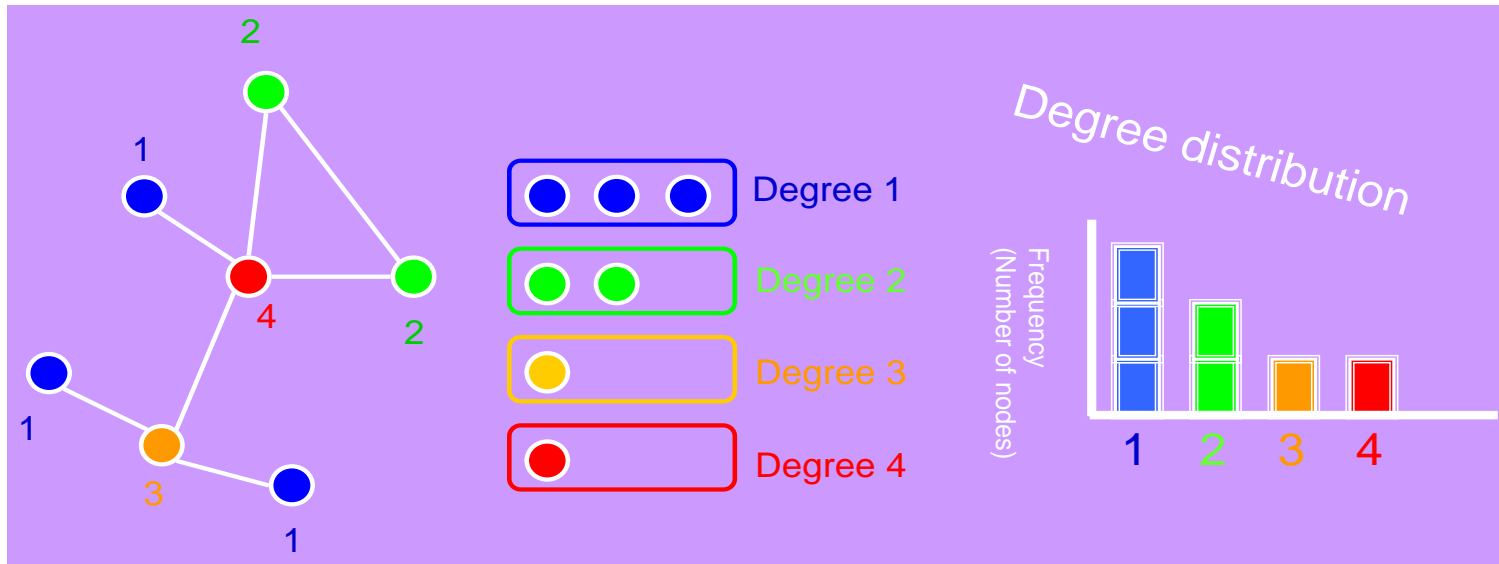
薬剤標的分子T

分子ネットワーク型
創薬/DR

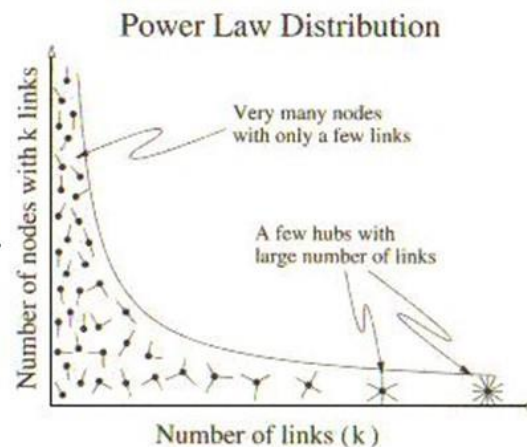


タンパク質相互作用 ネットワークの構造と薬剤標的分子

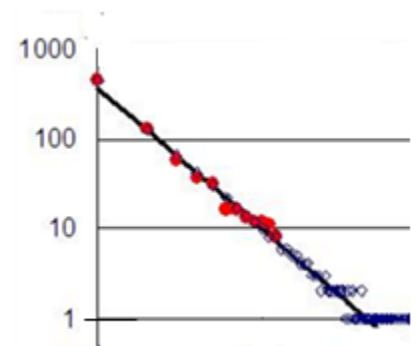
タンパク質相互作用ネットワーク(PIN)では数少ない相互作用が集中したタンパク質(hub)と相互作用が1や2の多数の末端タンパク質(branch)が存在する



<http://www.nd.edu/~networks/>

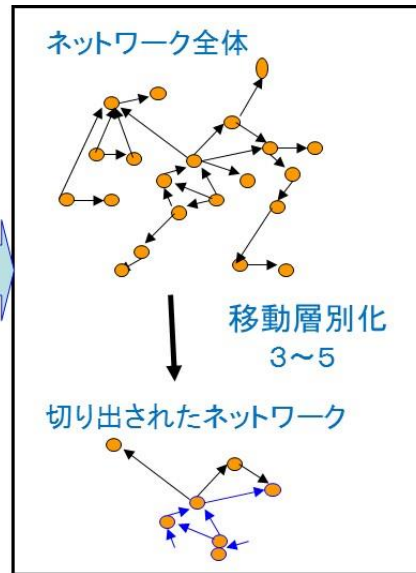
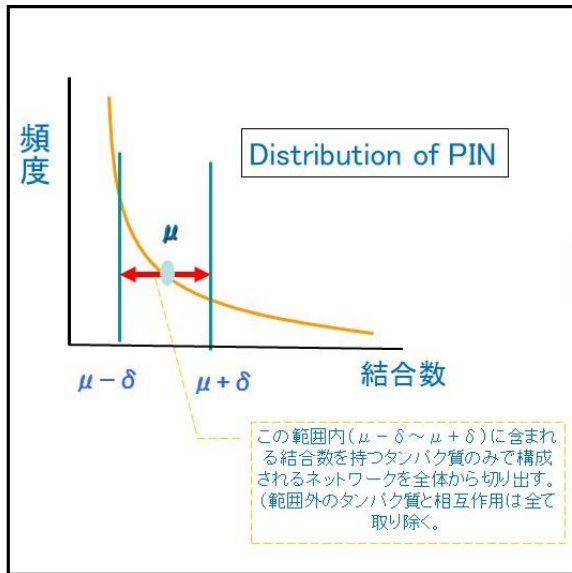


<http://www.macs.hw.ac.uk/~pdw/topology/ScaleFree.html>

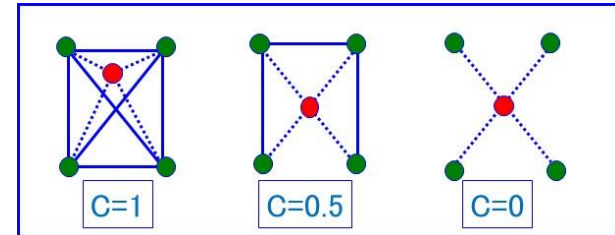


Log-log変換で直線

結合次数ごとの部分ネットワーク構造の結合密度の解析

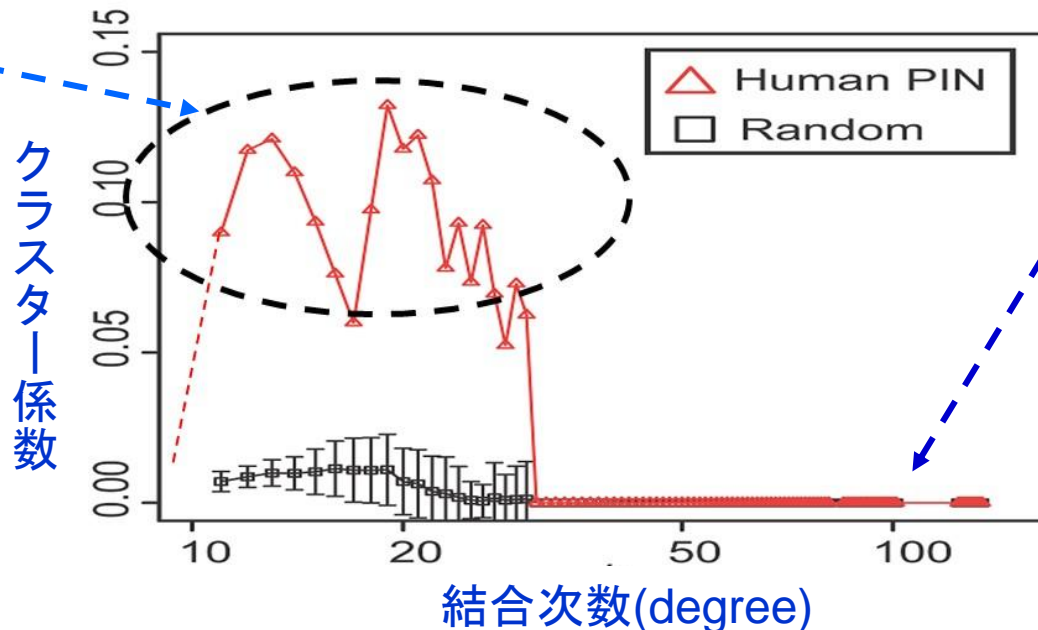


クラスター係数



Hase, T., Tanaka, H et.al (2009)
Structures of protein protein interaction network and their implications on drug design. *PLoS Compt Biol.* 5(10):

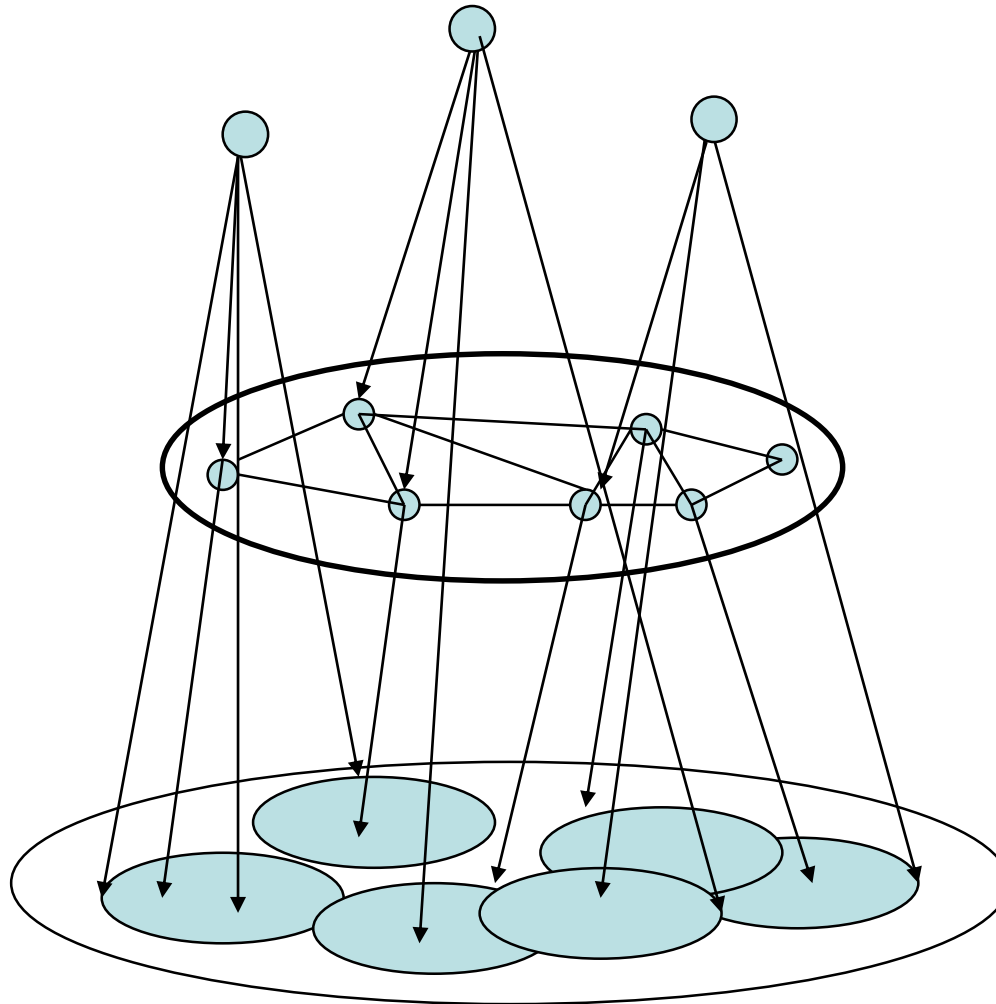
中程度の次数 (7~42) を持つタンパク質は多数の密なモジュールを構成



高い次数を持つノード(スーパーハブ)はお互いに密に結合しない

タンパク質相互作用から見られる

生命情報ネットワークの構造



高層
高次数 ハブ
次数
> 31 ヒト
> 39 酵母

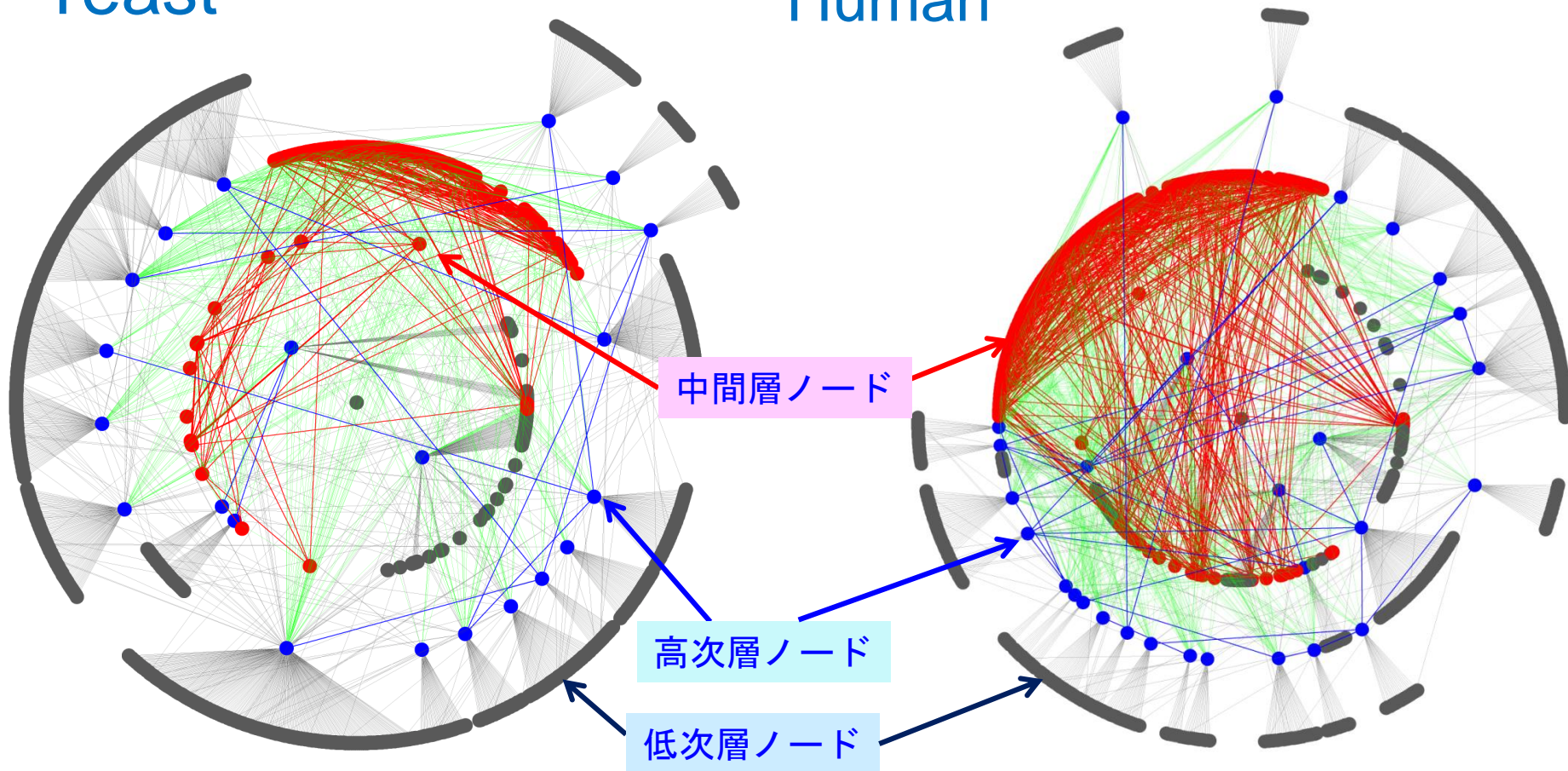
中間層
中程度次数
次数
6 ~ 30 ヒト
6 ~ 38 酵母

低層
低次数 ブランチ
次数 < 6

タンパク質相互作用ネットワークの Cloud Topology (3環トポロジー)

Yeast

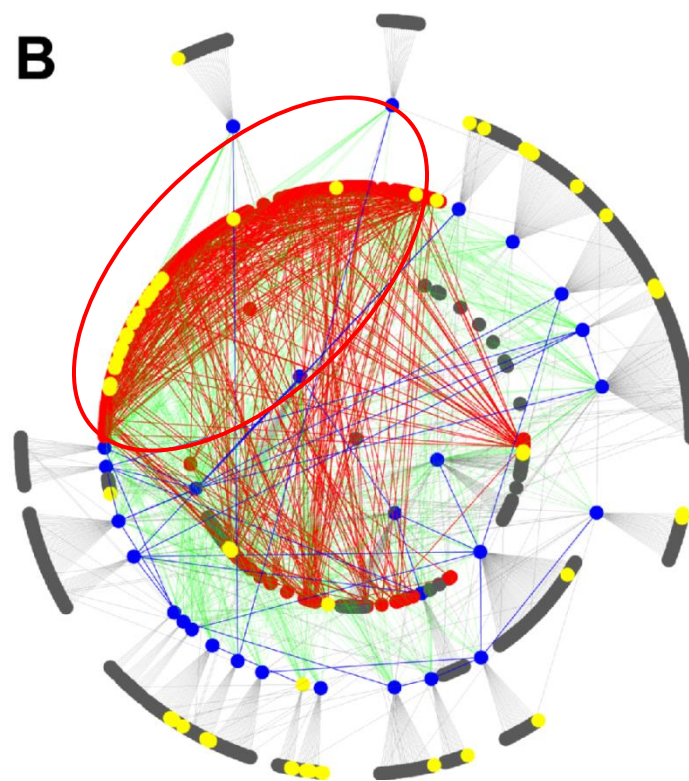
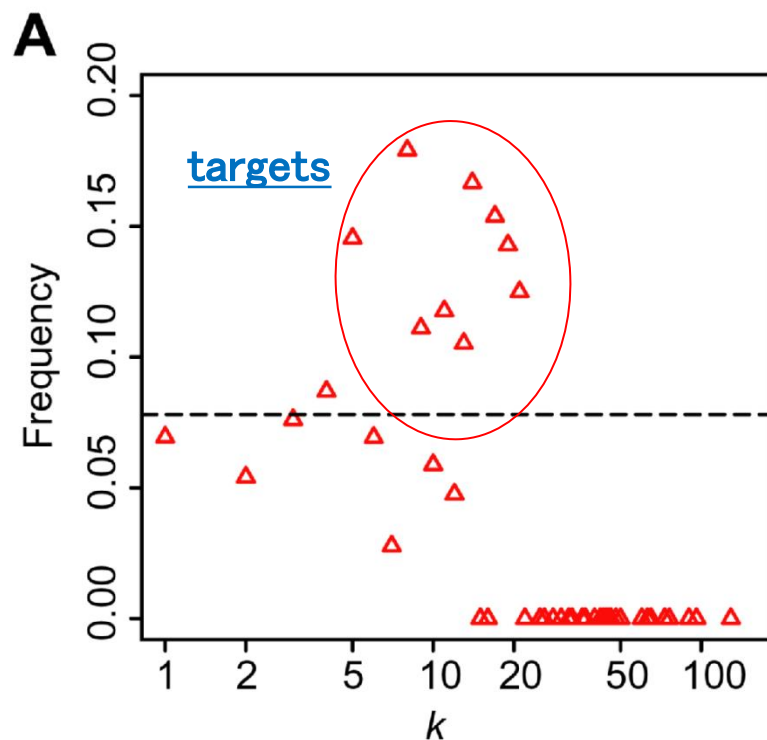
Human



中間層の次数ノードは PPI バックボーンを形成する

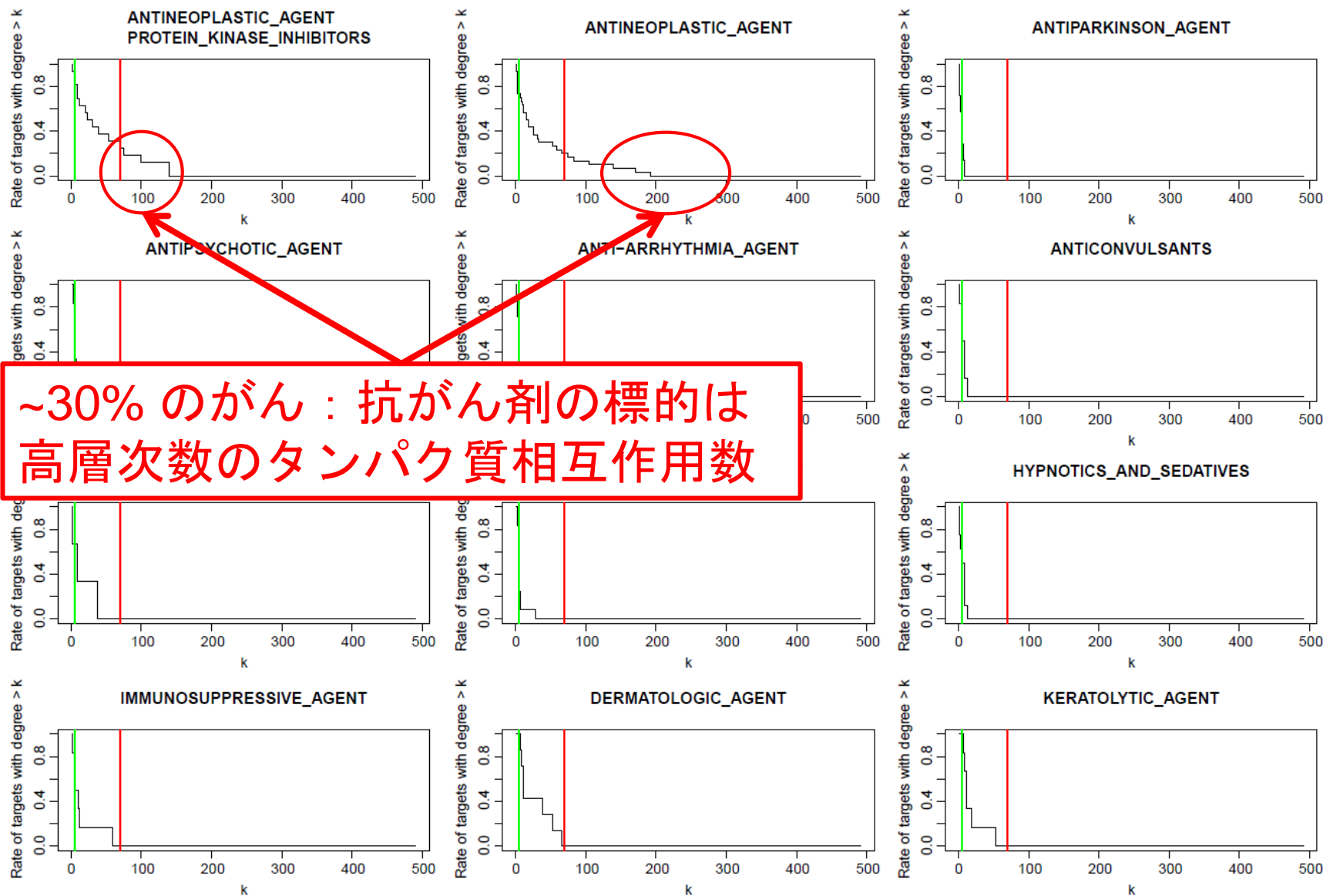
灰色, 赤, 青は、それぞれ低層、中層、高層の次数のノードをそれぞれ表す。

薬剤標的分子と結合度数



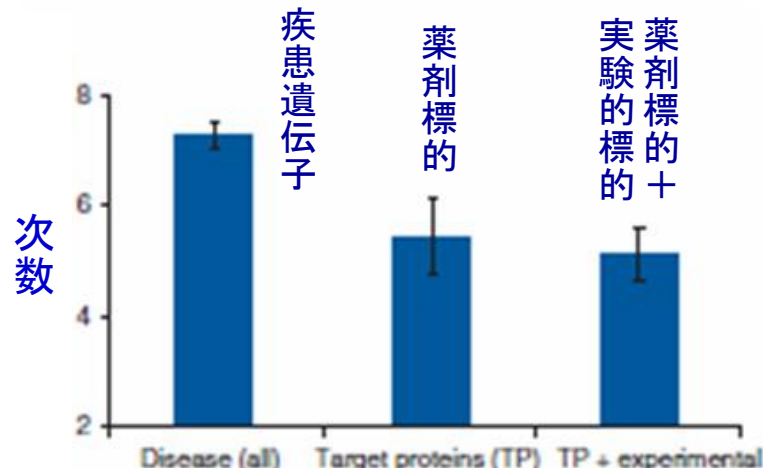
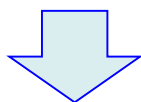
中層レベルのノードは治療薬として最適な標的である。それゆえ、多くの市場にある薬剤標的は、ヒトのバックボーンタンパク質に集中している

がん疾患遺伝子は高層次数ハブのタンパク質が多い

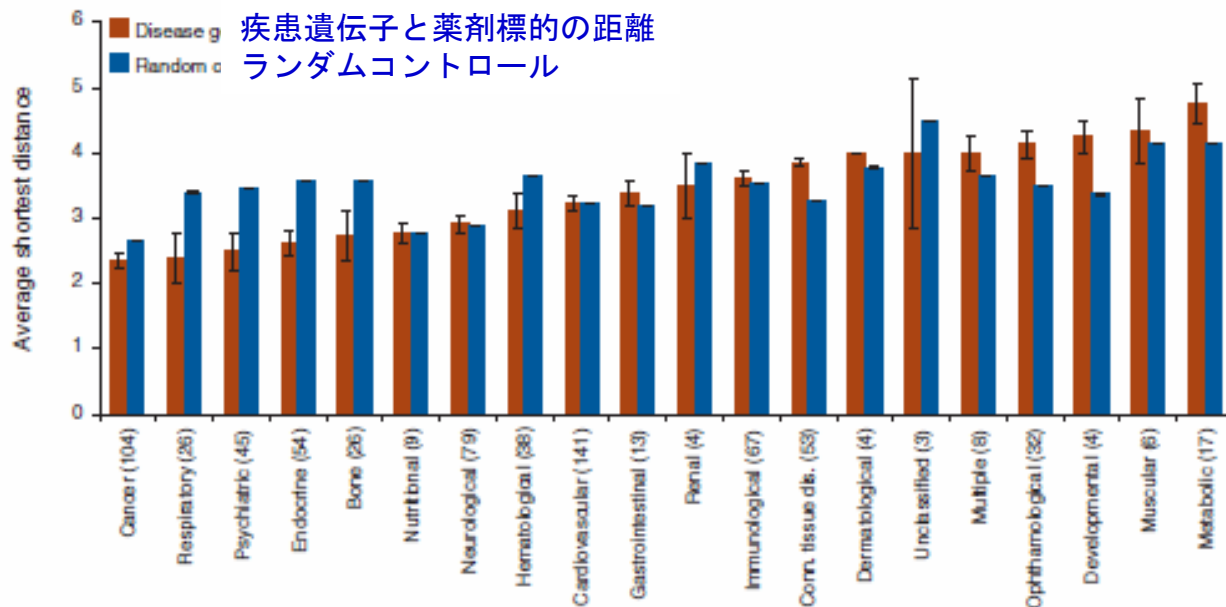


標的タンパク質と疾患遺伝子の距離

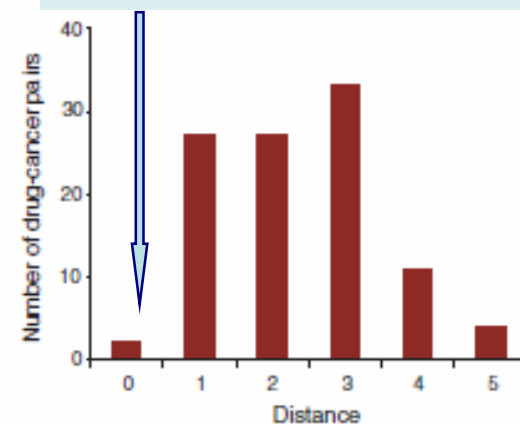
薬剤標的タンパク質と疾患関連タンパク質の間の距離：2~4リンク



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

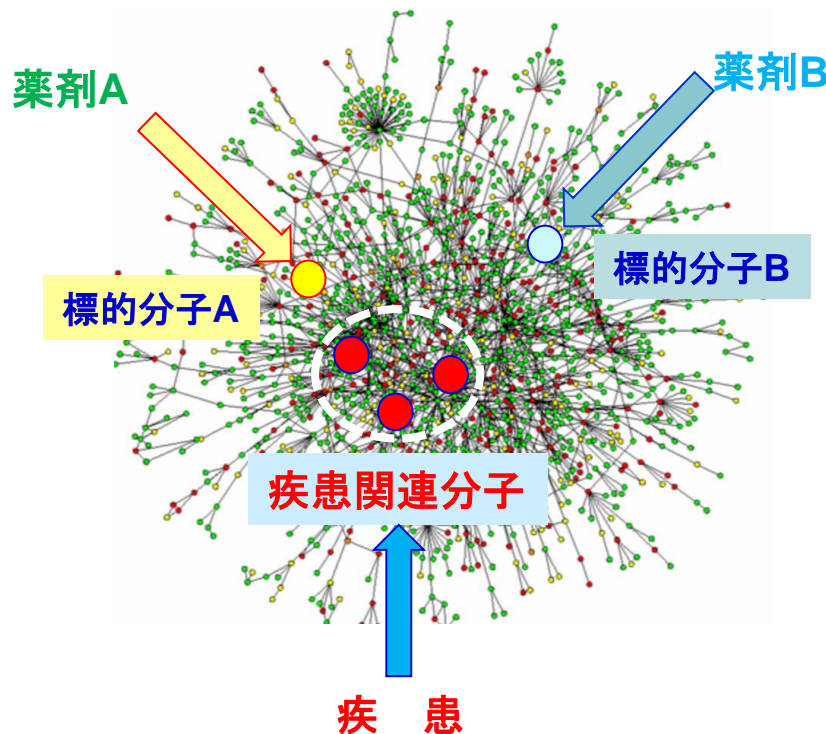


抗がん剤の標的分子と疾患遺伝子の間に距離

タンパク質相互作用ネットワークを 基盤にした計算創薬/DR

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）

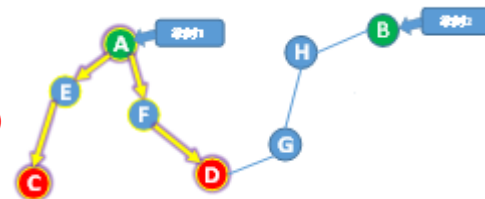
- **Interactome**(タンパク質相互作用)ネットワーク (Sun, 2016)

- **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-標的分子組 (4,452薬剤, 1,617タンパク質)

- **疾患遺伝子の差別的遺伝子発現データ (DEG)**

- **TCGA** (The Cancer Genome Atlas)より差別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** 差別的発現遺伝子

- **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

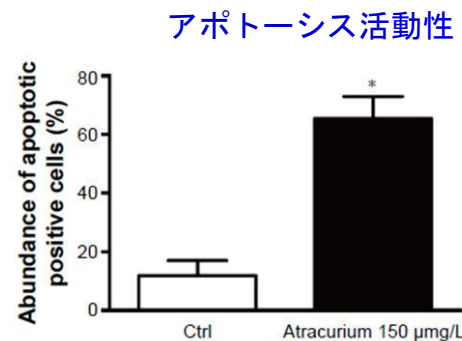
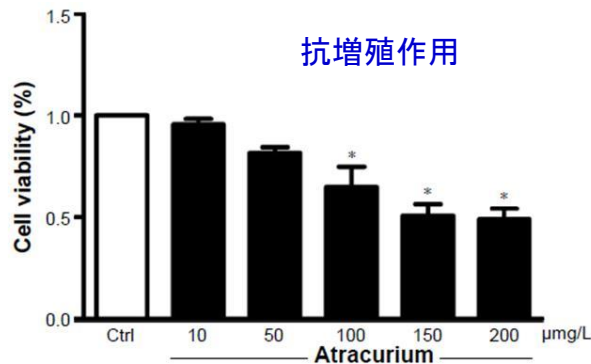
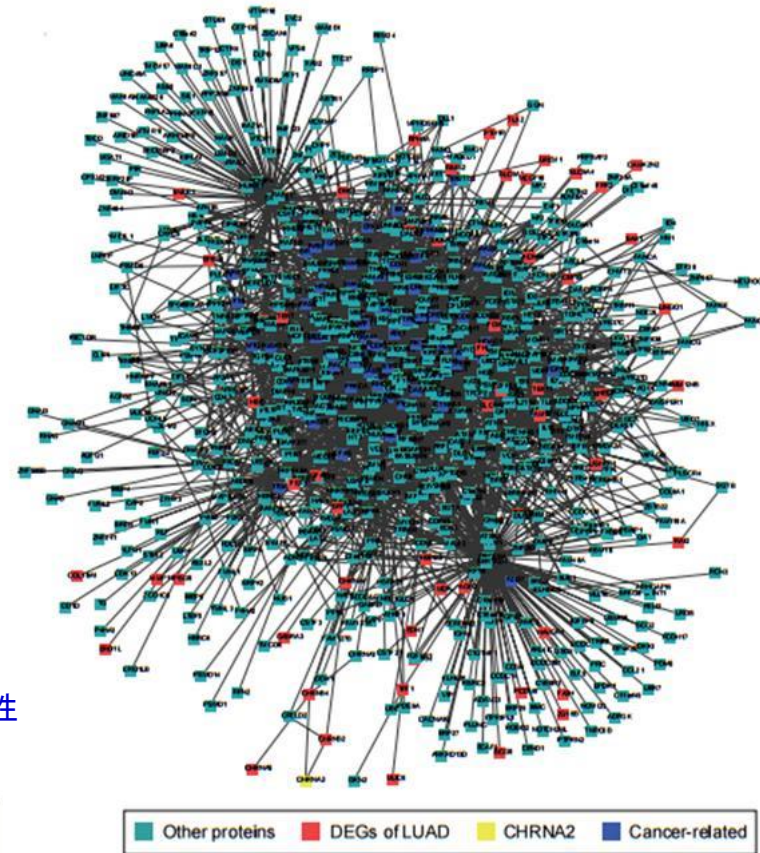
$$\mathbf{P}^{s+1} = (1-\gamma)\mathbf{M}\mathbf{P}^s + \gamma\mathbf{P}^0$$

\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

タンパク質相互作用ネットワーク 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

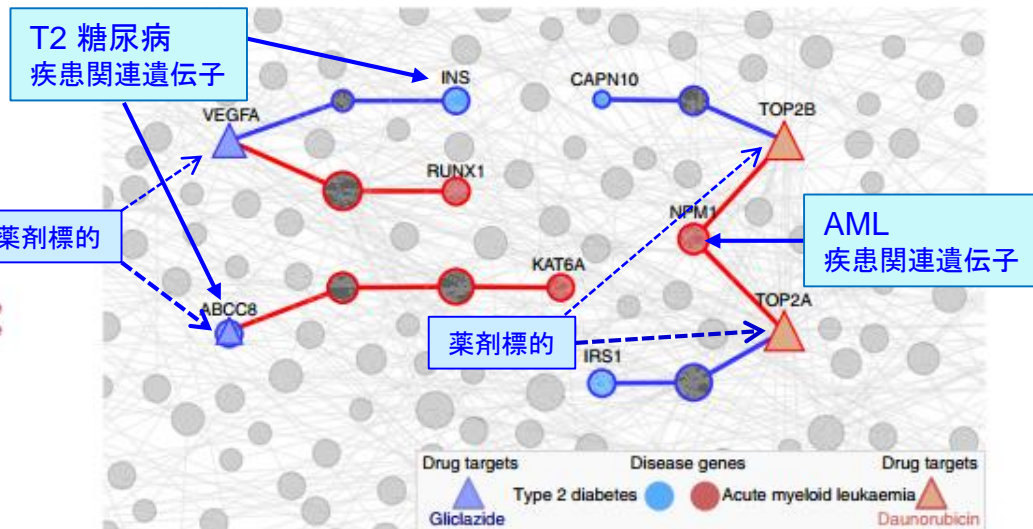
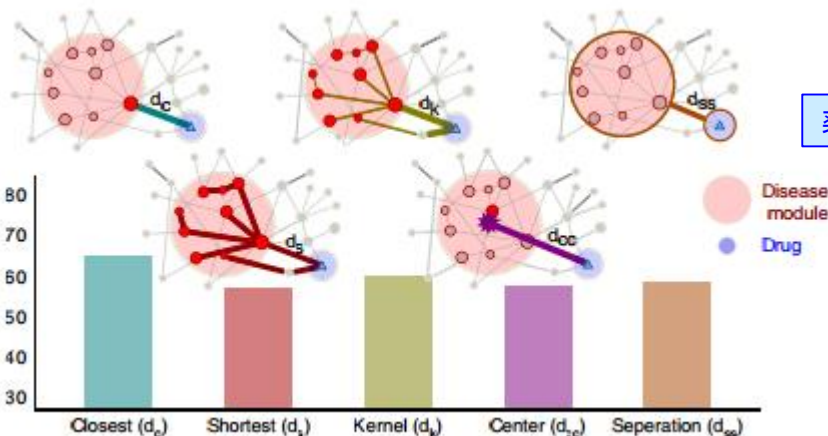
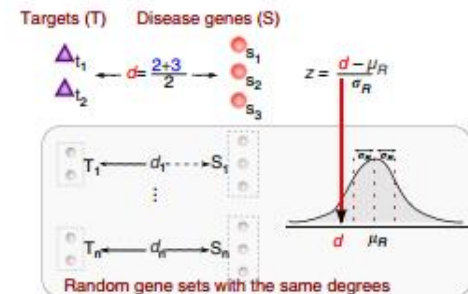
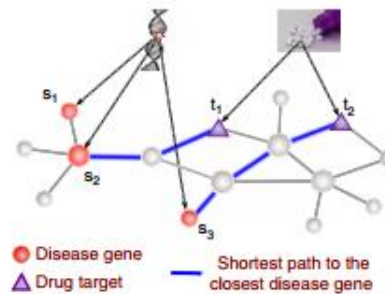
- **HPRDとSTRINGSの両方のランダム歩行で145薬剤・化合物が共通**
- **最高スコアを挙げたAtractiumを選択**
- **標的はCHRNA2(Cholinergic Receptor Nicotinic Alpha 2) でアポトーシス経路である**
- **培養細胞A549 (ヒト肺胞基底上皮腺癌細胞) の抗増殖作用を確認**



タンパク質相互作用ネットワークでの 近接性によるDR

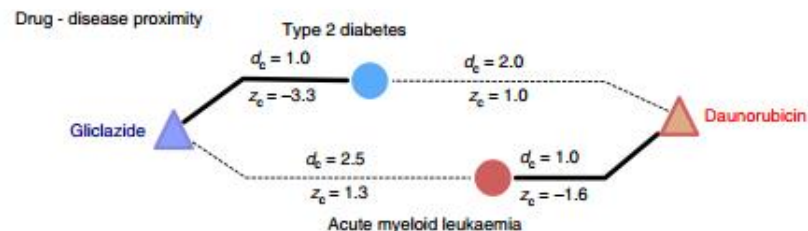
相対近接指標 d_c :

- 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 $\Rightarrow z$ 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



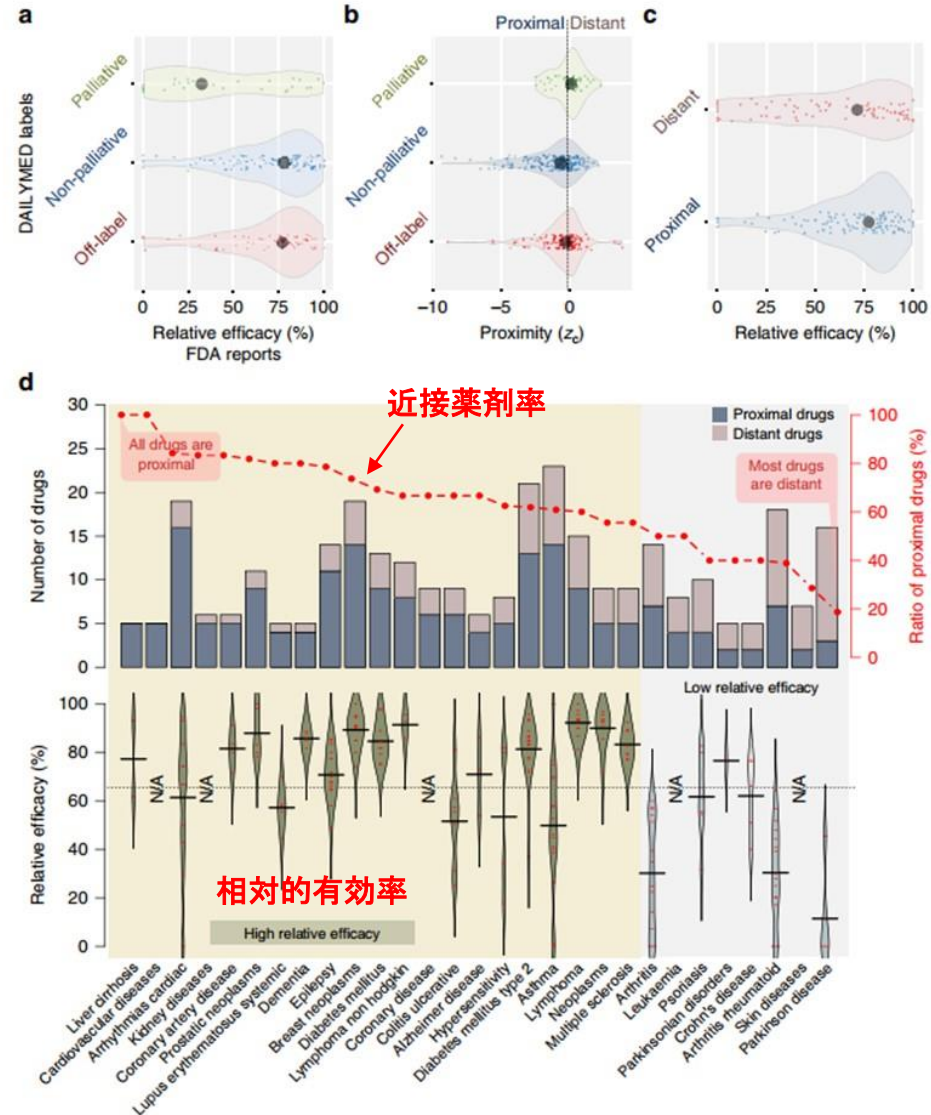
大半の薬剤は標的と疾患関連分子
2リンク離れている

(Gunev, Barabasi, 2016, Nat. Com)



相対近接性による薬効予測

- 疾患モジュールの内部/近接に標的分子を持つ必要がある
- これまでの研究では疾患関連分子と標的分子の距離が大
 - 対症療法・緩和療法：疾患原因ではなく症状を標的としている
 - 標的分子が疾患関連分子の数は少ない (402対のうち62)
- 既成の薬は疾患と近接的である
- 緩和療法は遠隔的である
- Off-labelは緩和より近接的である
- 近接薬剤の治験の頻度は高い
- 薬剤は選択的であるが排他的ではない
- 相対的有効性と近接指標は相関する
- 平均の標的分子の数は3.5個である



AI創薬・DR (学習型アプローチ)

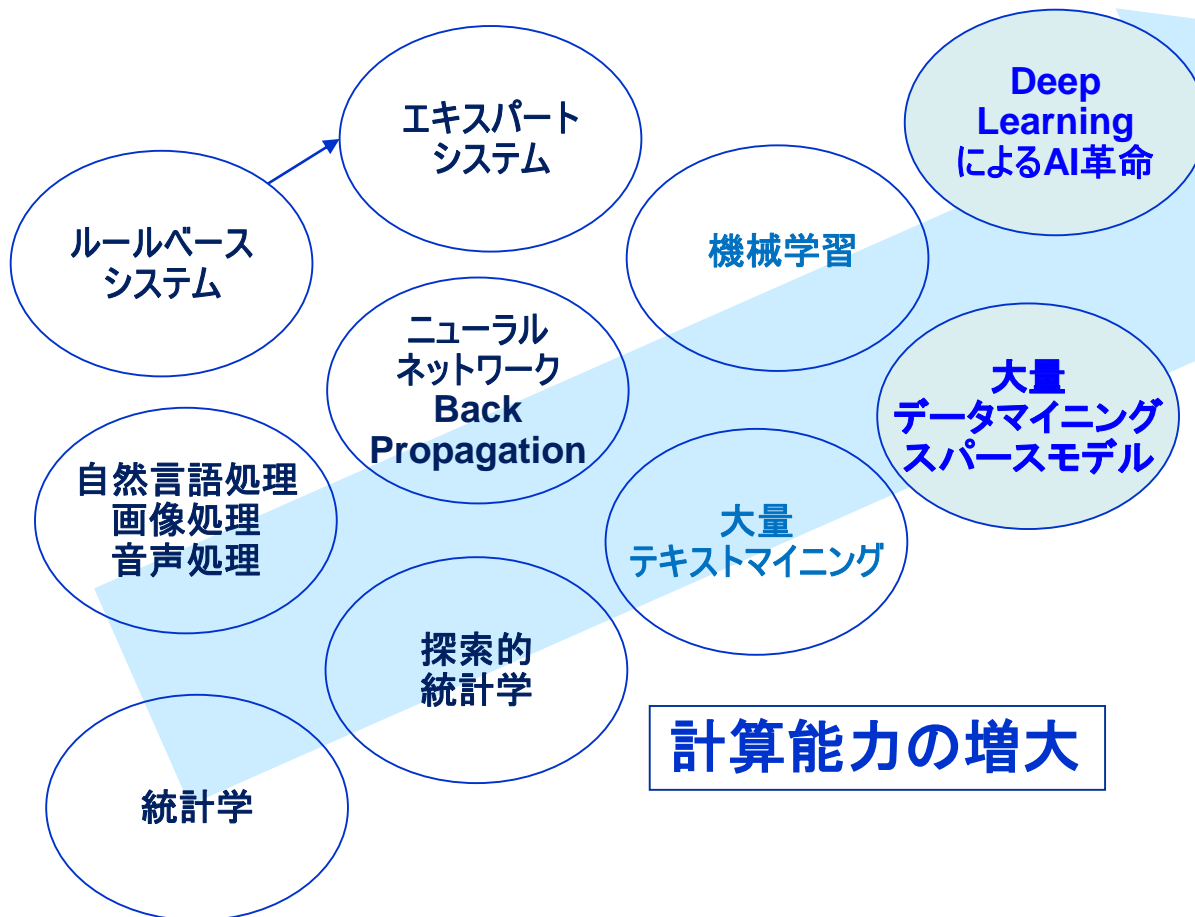
人工知能とDeep Learningの革命性

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



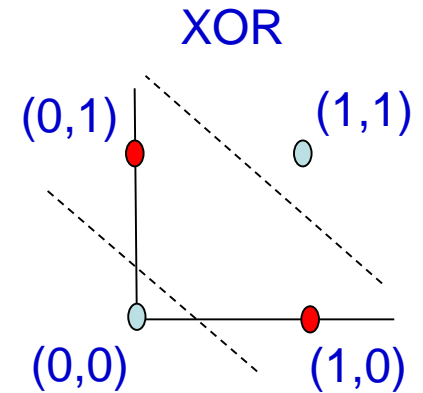
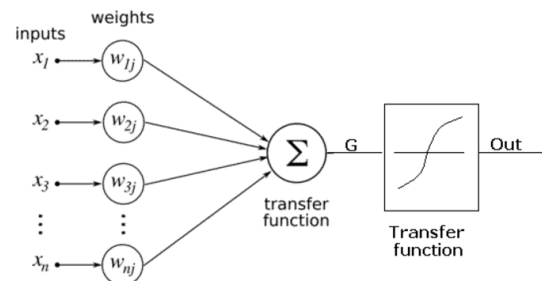
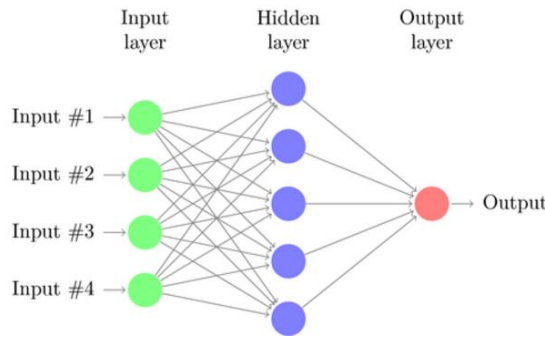
人工知能の最近の話題

- 「**アルファ碁**」 (Google DeepMindによるコンピュータ囲碁プログラム) が2016年3月に数多くの世界戦優勝経験のあるプロ棋士李世石 (Lee Sedol : 九段) に挑戦し、**4勝1敗と勝ち越した**
 - チェス : IBM 「Deep Blue」 が1997年に当時の世界champion, カスパロフ氏 (ロシア) に勝利
 - 将棋 : ボンクラーズ, 2012年米長永世棋聖に勝利
 - 「アルファ碁」にはニューラルネットワーク (Deep Learning) が使われた。評価経験則が人間によってコードされていない
 - 最初、棋譜に記録された熟練した棋士の手と合致する手をさすように訓練され、次に、ある程度の能力に達すると、強化学習を用いて**自身と多数の対戦 (3000万回) を行う**ことで上達した。
- 人工知能が1000万枚の画像を与えて「猫」を認識するニューロンをできたと2012年に発表



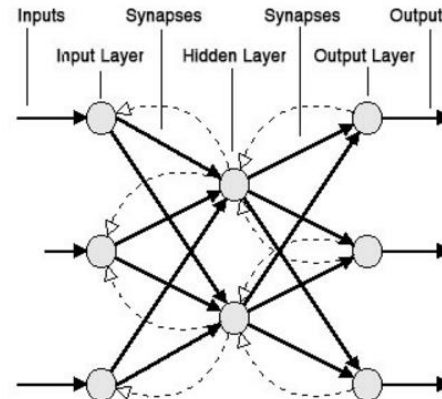
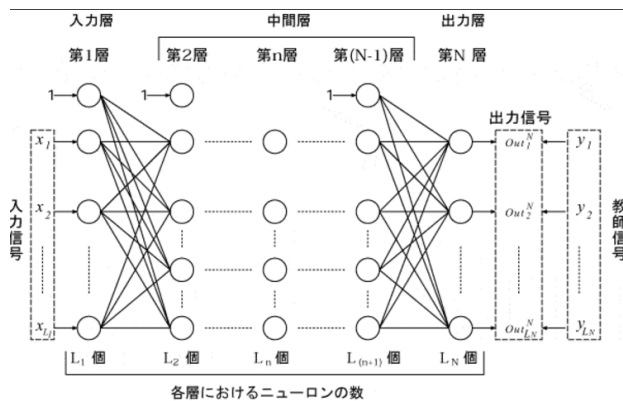
従来のニューロネットワーク

古典的Neural Network・パーセプトロン(1970年代)



多層Neural NetworkとBack projection (1980年代)

線形分離できない

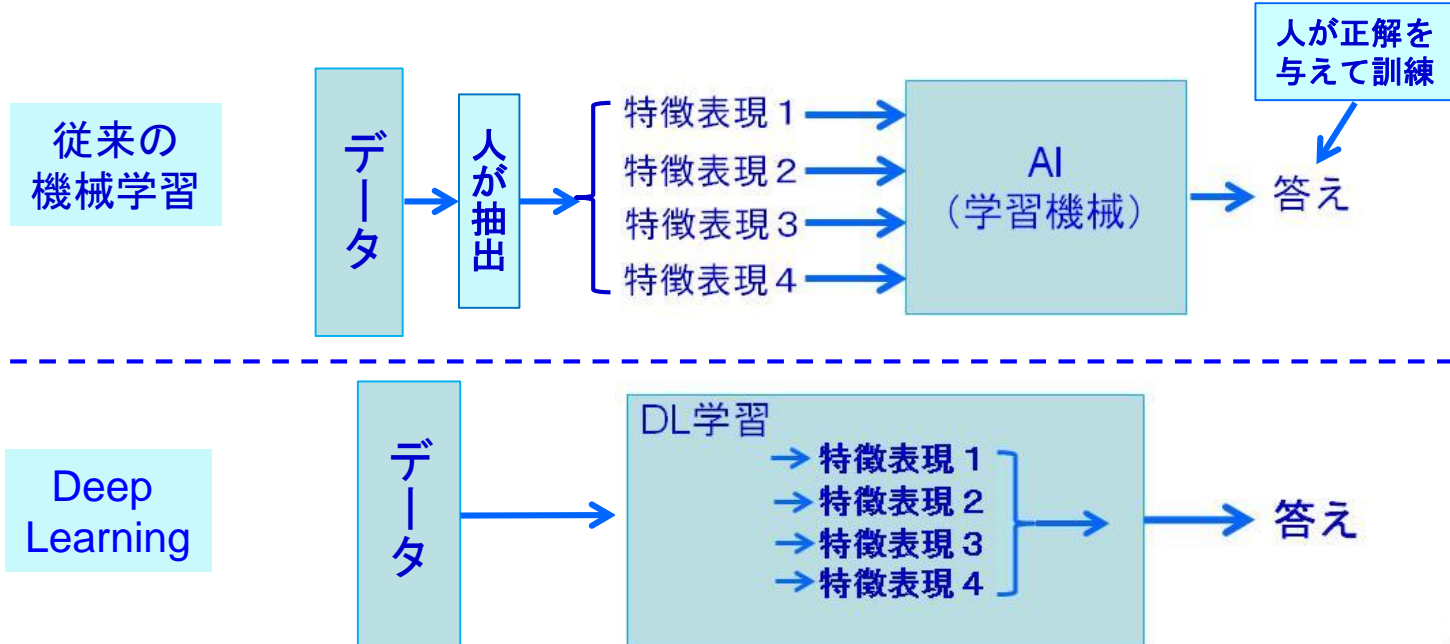


Back Propagation (1986 Rumelhart)
 望ましい出力との誤差を教師信号として与える事により、次第に結合係数を変化させ、最終的に正しい出力が得られるようにする。結合係数を変える事を学習と呼ぶ。この学習方法には、最急降下法(勾配法)が使われる。出力層へ寄与の高いノードの重みの変更。

多層にわたる逆伝搬で修正感度減衰

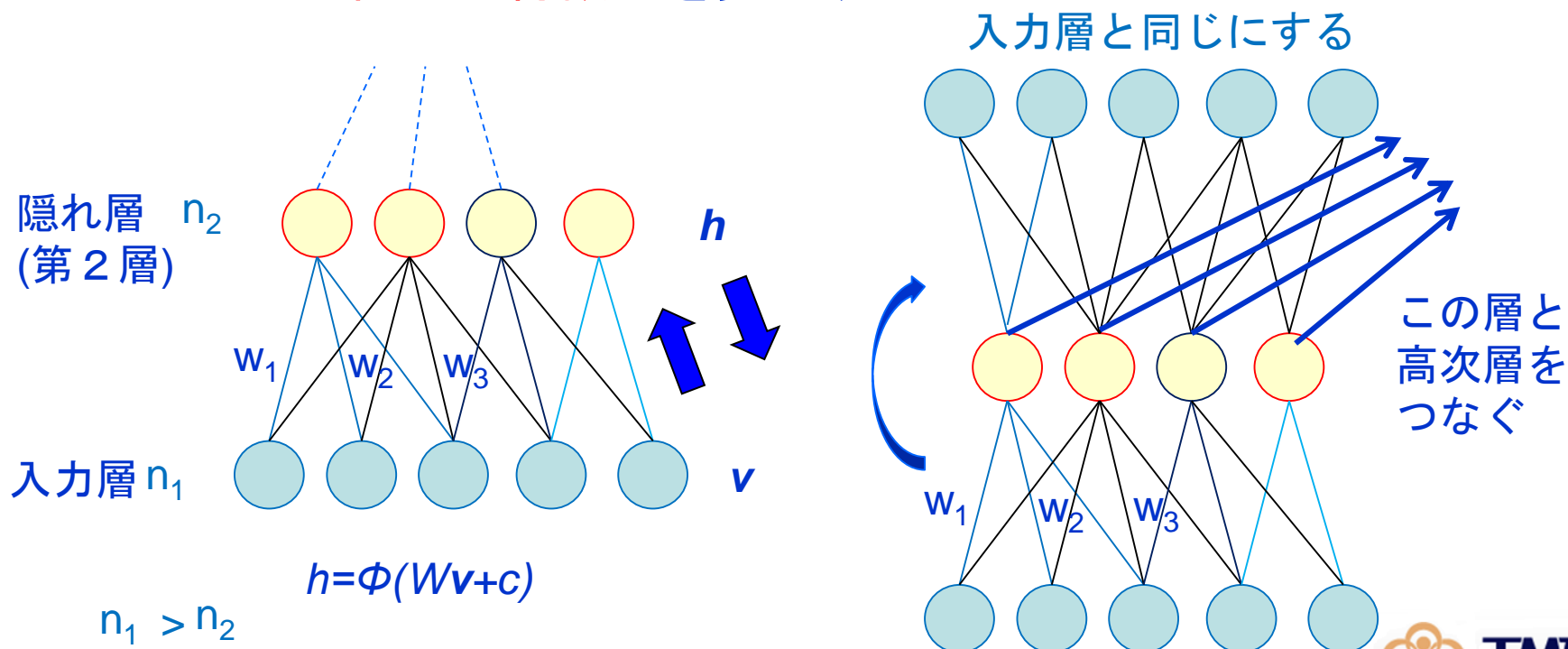
Deep Learning による 人工知能革命

- 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



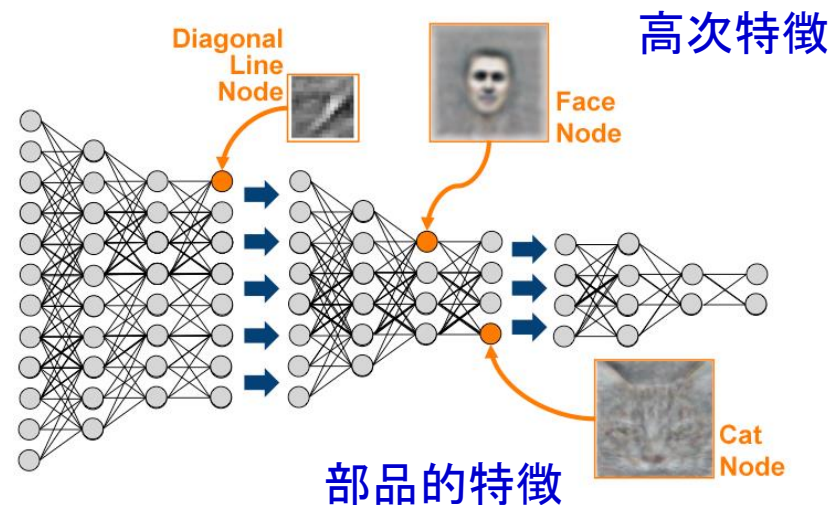
DLの革命点 Autoencode 1

- 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 Autoencoder 2

- 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



AI創薬・DR (学習型アプローチ)

Deep Learningの創薬への応用

Deep learning: 創薬からの注目

- **Kaggle** (データサイエンス競技会)に**Merck社**が出題
Molecular Activity Challenge (2012).

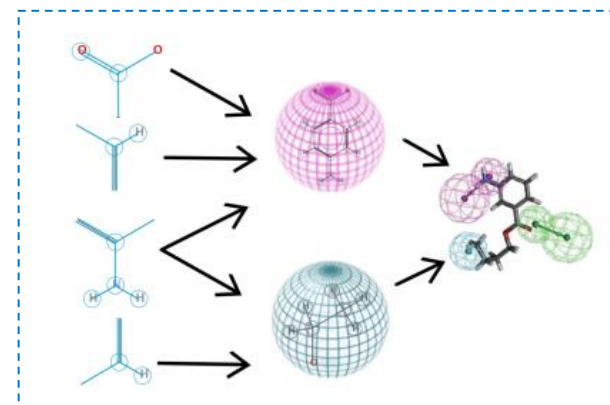
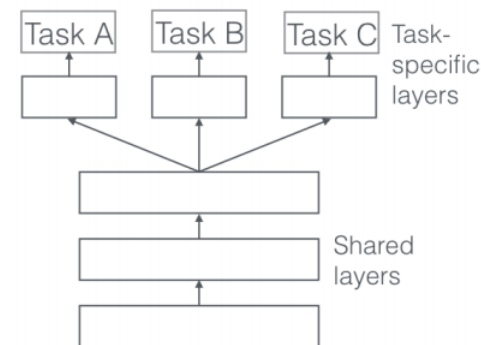
- 15種類の標的分子に対する化合物データセットから異なった**構造活性相関のデータ**を学習して**構造から分子の生物学的活性**を予測するモデルの開発コンテスト
勝利したモデルはdeep learning を用いた

- **Unterthiner**の大規模な構造活性相関 (QSAR)研究

- 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がSVM, k-最近隣法, logistic回帰より有効
- 特徴量の抽出、薬理機序への理解

- **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"

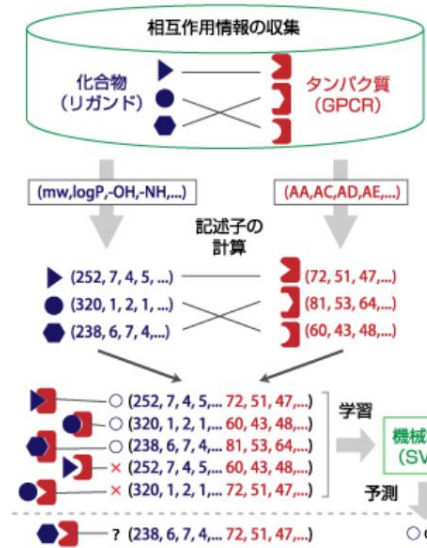


AI創薬の方法

- **Virtual Screeningへの人工知能・機械学習の応用**
 - **Ligand-based** AIバーチャルスクリーニング
 - **Structure-based** AIバーチャルスクリーニング
- **標的分子探索に人工知能を用いた方法**
 - Hase-Tanakaの多層Deep AutoEncoderを用いた標的分子探索法
- **その他**
 - 化合物の人工知能を用いた自動設計
 - 合成経路設計
 - AI毒性学

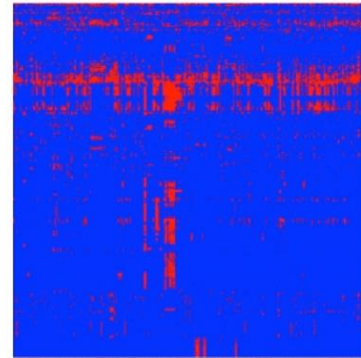
奥野らのCGBVS法

- 膨大な化合物候補と多数の標的タンパク質候補との組合せの相互作用評価。
- 既知のタンパク質（標的）と化合物の相互作用を機械学習, 相互作用の有・無を判定
- タンパク質と薬剤候補化合物学習の記述子
 - 標的タンパク質：2アミノ酸や3アミノ酸の出現頻度、構成アミノ酸の特性など
 - 薬剤候補化合物：分子量、炭素などの構成原子数、部分構造の有無、疎水性度など、化合物の通常2次構造の特徴と物性
 - 両者の間の相互作用情報



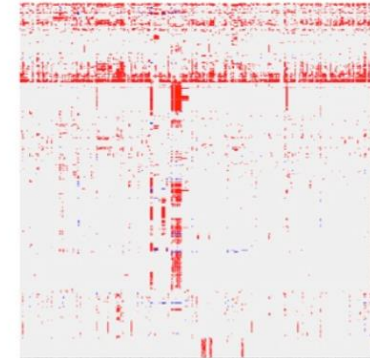
化合物 500 個 × Kinaseターゲット 388 個

CGBVS予測結果



赤: スコア > 0.8
青: スコア < 0.8

アッセイデータ



赤: 活性有り < 30 μM
青: 活性なし > 80 μM
灰: データ無し

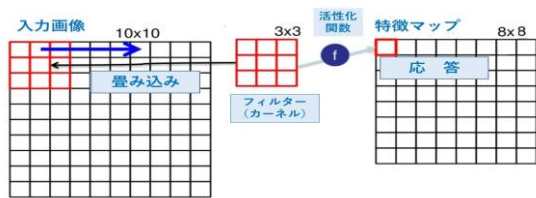
これまでの相互作用の有無が既知である標的タンパク質と化合物の組を選び、相互作用がある場合を「正例」として、相互作用がない場合を「負例」としてする

バーチャルスクリーニングへの AIの応用 (structure-based)

AtomNet (Wallch, 2015)

- トポロジー(近接)情報を学習できるDeep Learning法
 - コンボリューション型Deep Learning
(convolutional deep neural network)
- 入力データ
 - 薬剤標的分子：1 Å 3Dグリッドでの原子座標
 - 標的の結合サイト内の低分子
- ネットワーク構造
 - 3D コンボリューション層構成
 - 3D convolutional filter
- 学習法：確率的勾配降下法、逆伝播法
- 結果：活性あるいは非活性クラス的确率を推定
- 既存法と比較評価
 - ドッキング法 (Smina, Autodock Vina) を上回る予測精度 (AUC)

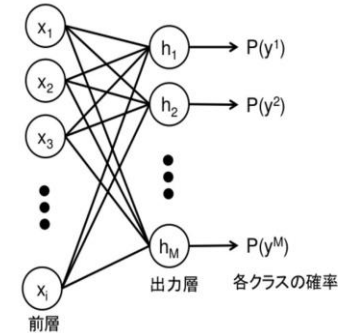
バーチャルスクリーニングへのAIの応用 (structure-based)



0	3	1	5
1	6	2	4
1	2	1	3
7	1	1	2



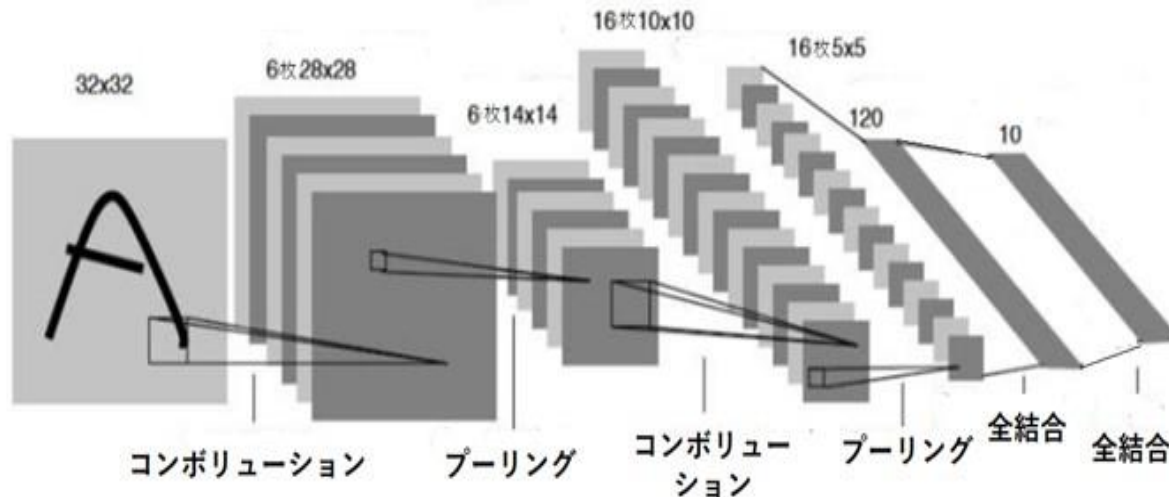
6	5
7	3



Softmax

$$P(y^j) = \frac{\exp(h_j)}{\sum_{j=1}^M \exp(h_j)}$$

各クラスの確率を算出して
最大値を認識クラスとする



コンボリューション型DNNの全体のアーキテクチャ

Deep Learningに基づく 標的分子探索

薬剤標的分子探索

対象疾患決定後、治療に有効な生体側の標的分子を探索
標的分子探索の範囲を限定

⇒ ヒト・タンパク質相互作用ネットワーク (PPIN)

学習的アプローチ

その疾患にこれまで有効な標的分子が既知
既知の標的分子がPPIN上でどのような位置にあるのか
帰納学習する

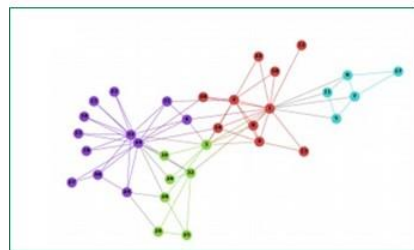
PPINはHPRDでは<1万タンパク質 × 一万タンパク質>の
超多次元ネットワークで通常の機械学習では困難



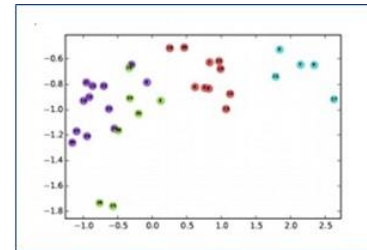
Deep learning による
<ネットワーク埋め込み **Network Embedding**>

Deep Learningによる Network Embedding

超多次元ネットワークをそれより遥かに低次元のLatent Spaceに写像



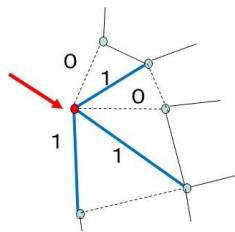
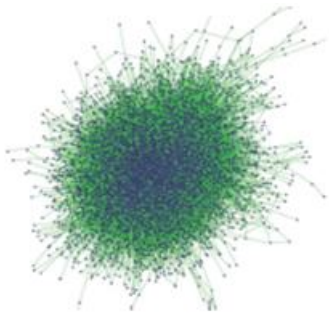
Original network



Latent space

Structural Deep Network Embedding

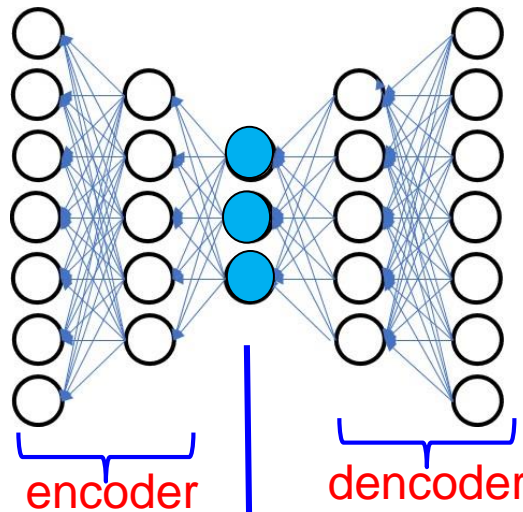
PPIN



$$v_i = (0, 0, 0, 1, 0, 1, 0, \dots)$$

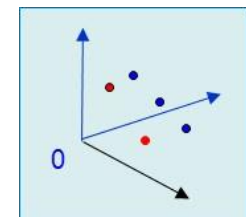
全節点の近接ベクトル

入力層



対称出力層

潜在空間
Latent space



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) Structural Deep Network Embedding 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) Latent Space における正例補完

- DrugBank からの正例とその増加 (SMOTE 法)

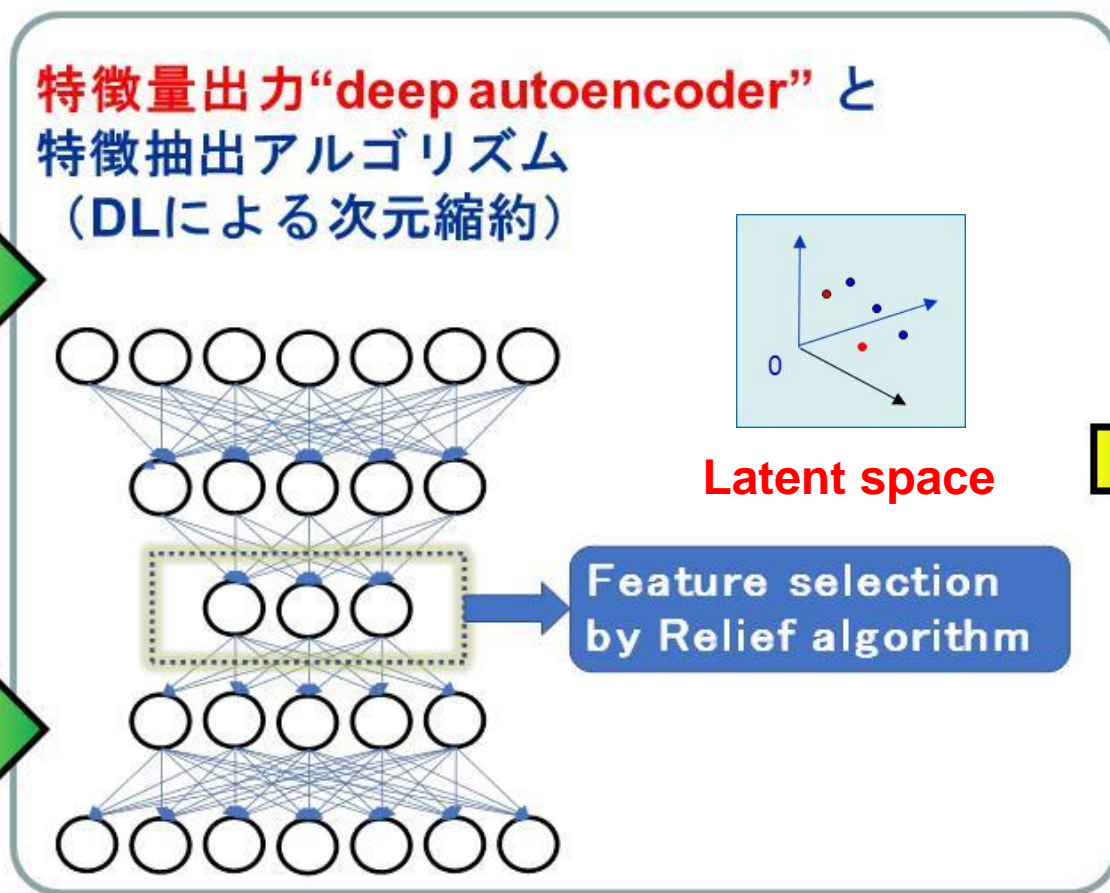
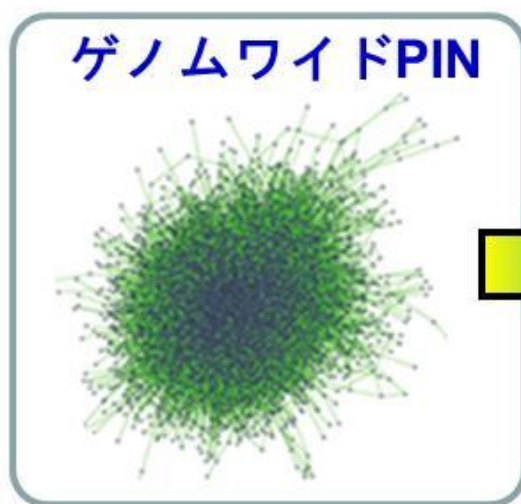
4) Latent Space での正例を用いた機械学習分類

- Xgboost 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

入力

特徴量産出



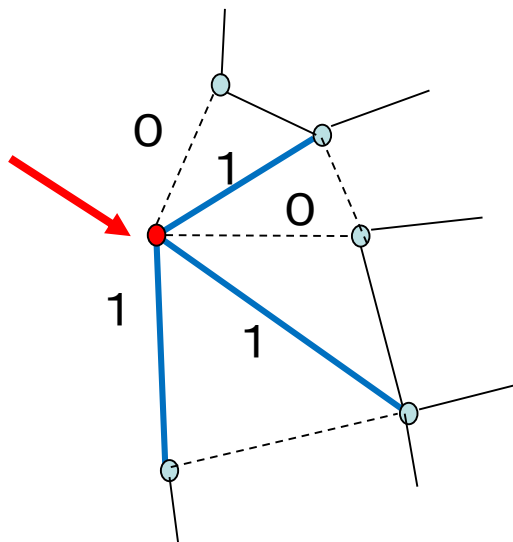
Deep Learning と SVD (singular value decomposition)の精度の違い

あるタンパク質ノードに
注目する

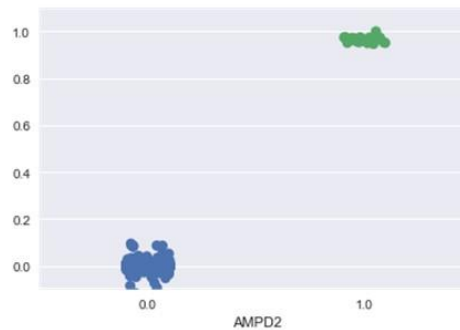
周りのノードで

結合しているノードは 1
結合していないノードは 0
とすると0, 1の近接ベクトル
で結合を表現できる。

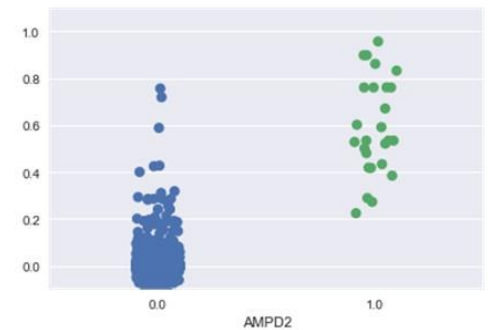
$$v_i = (0, 0, 0, 1, 0, 1, 0, \dots)$$



AMPD2 (adenosine monophosphate deaminase 2)
degree=26

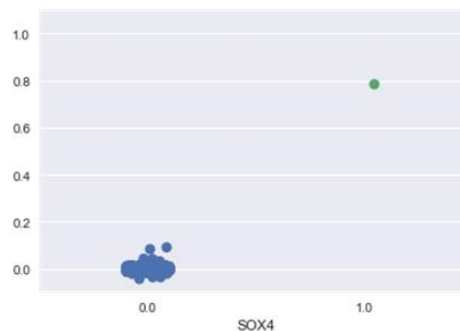


Autoencoder

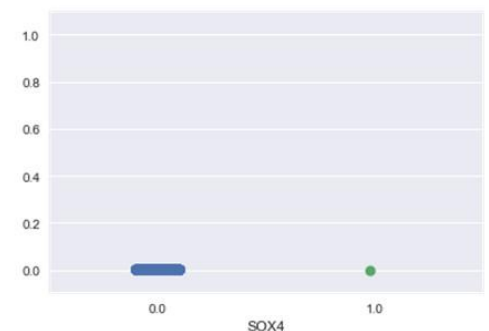


SVD

SOX4 (SRY-box 4)
degree=1



Autoencoder



SVD

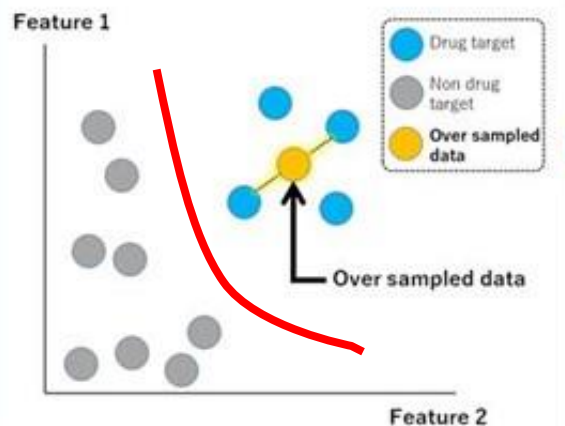
N=8,502

Deep Learningによる創薬・DR

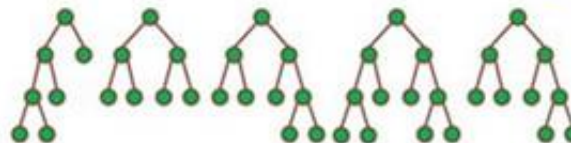
分類モデル

標的選定

2群分類と標的判定 最新のアルゴリズム



標的性判定 algorithm to build
a binary classifier

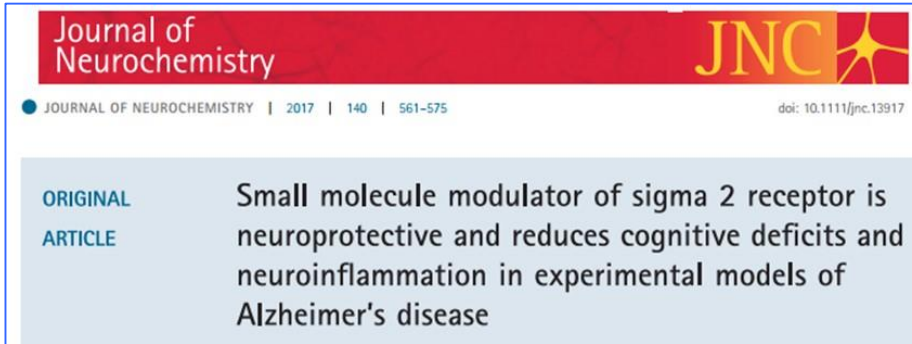


標的性判定

遺伝子	標的確率
GRASP	0.982971
PGRMC1	0.982345
GPM6A	0.982345
NRP2	0.975194
PFKM	0.972128
DLGAP2	0.953659
CD81	0.941095
IQGAP1	0.926867
TROVE2	0.916886

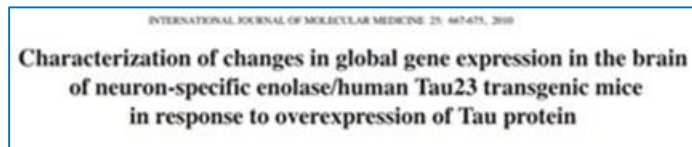
実験的研究との付合 1

PGCM1 : progesterone receptor membrane 1

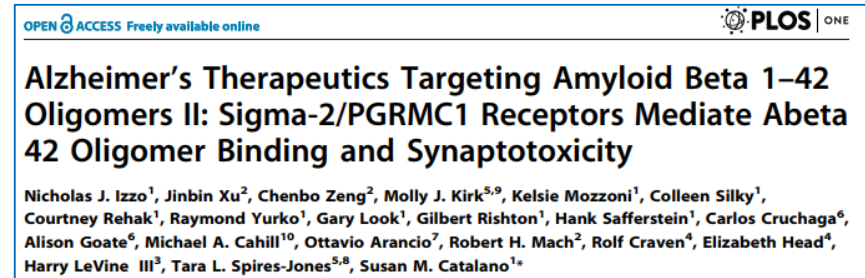
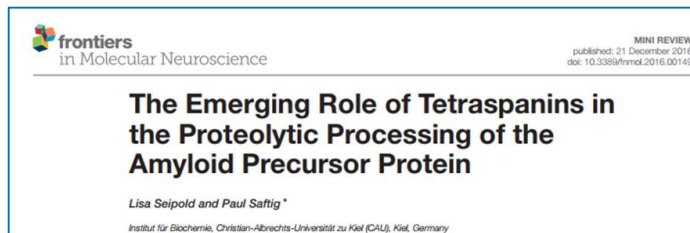


神経保護の効果 (neuroprotective) 認知不全・炎症に治療効果

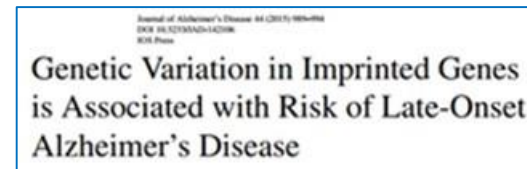
GPM6A : Glycoprotein M6A



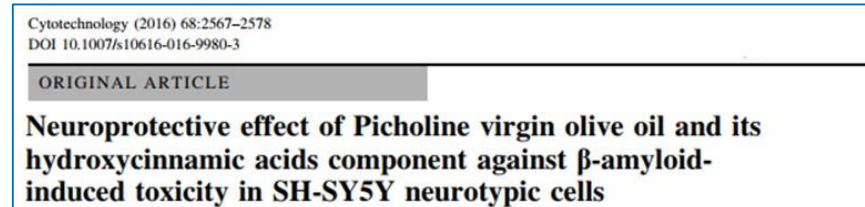
CD81: Tetraspanins family



DLGAP2 : DLG-Associated Protein 2



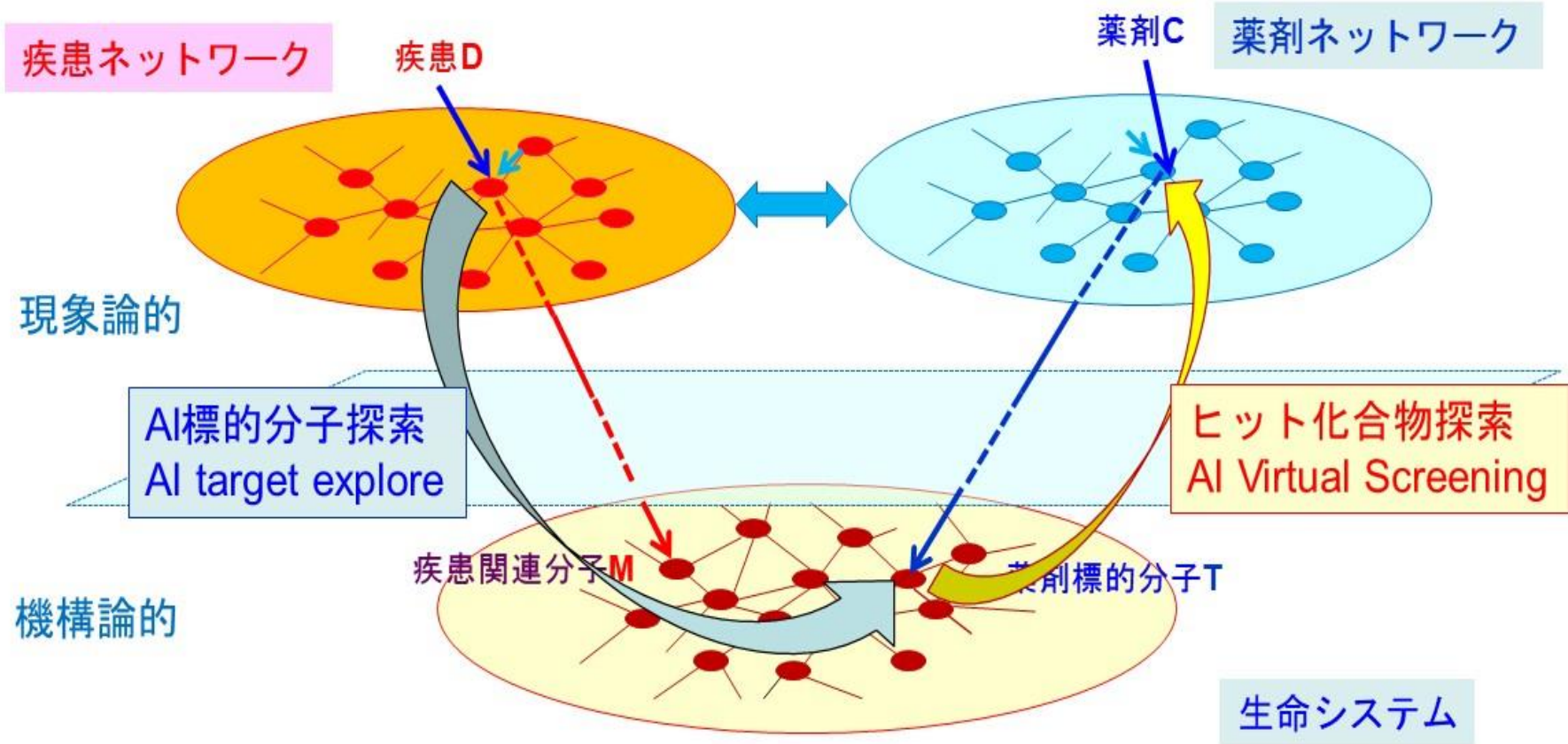
PFKM: Phosphofruktokinase



AI創薬の実現

3層生体・薬剤ネットワークによるAI創薬の過程

薬剤Cは疾患Dに薬効



人工知能による医学〈知〉革命

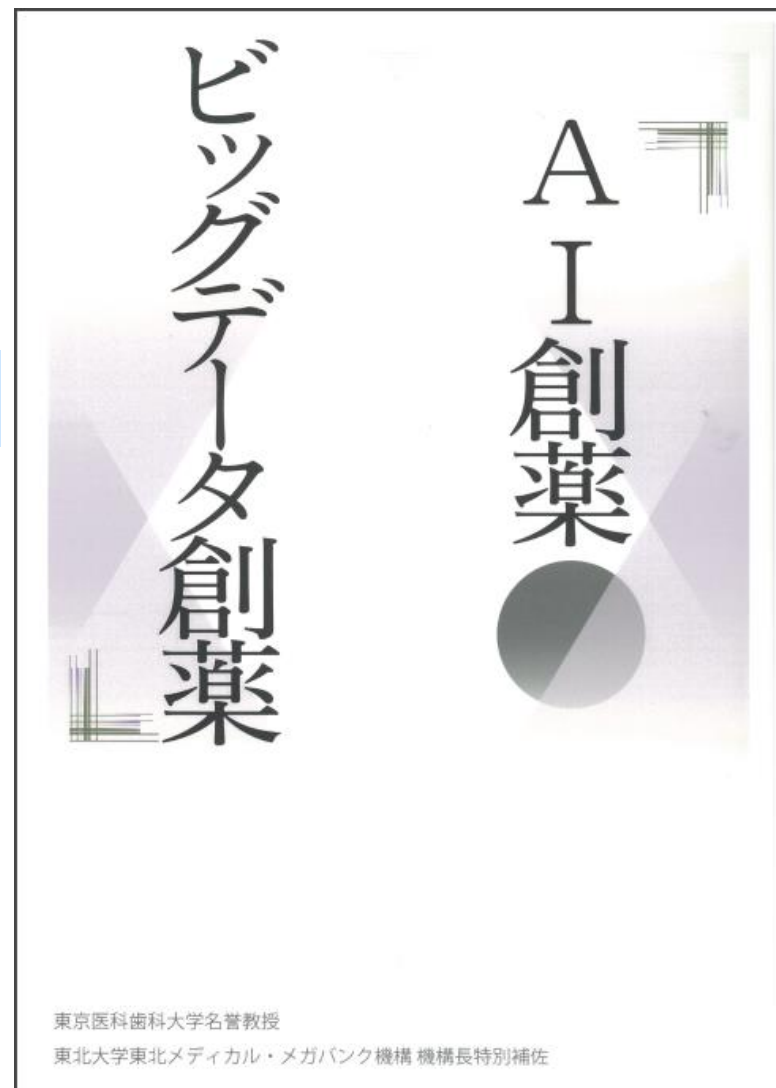
- ビッグデータ医療・創薬時代：次元縮約
- Deep Learningによる〈多次元ネットワーク情報構造〉の縮約
 - 創薬だけでなく、ビッグデータ医療への適応可能
 - ゲノム医療の〈網羅的分子情報－臨床表現型〉の相関ネットワーク構造
 - バイオバンクの〈遺伝素因－環境要因〉と発症
- 人工知能「データからの教師なし学習」はこれまでの医学知識のDOXAを根本から洗い直す

ヒトの**仮設駆動的な〈知〉**とAIの**データ駆動的な〈知〉**との
「**共創的cocreativeな〈知〉**」が
これからの**人類の未来の進むべき途の探索**を可能にする

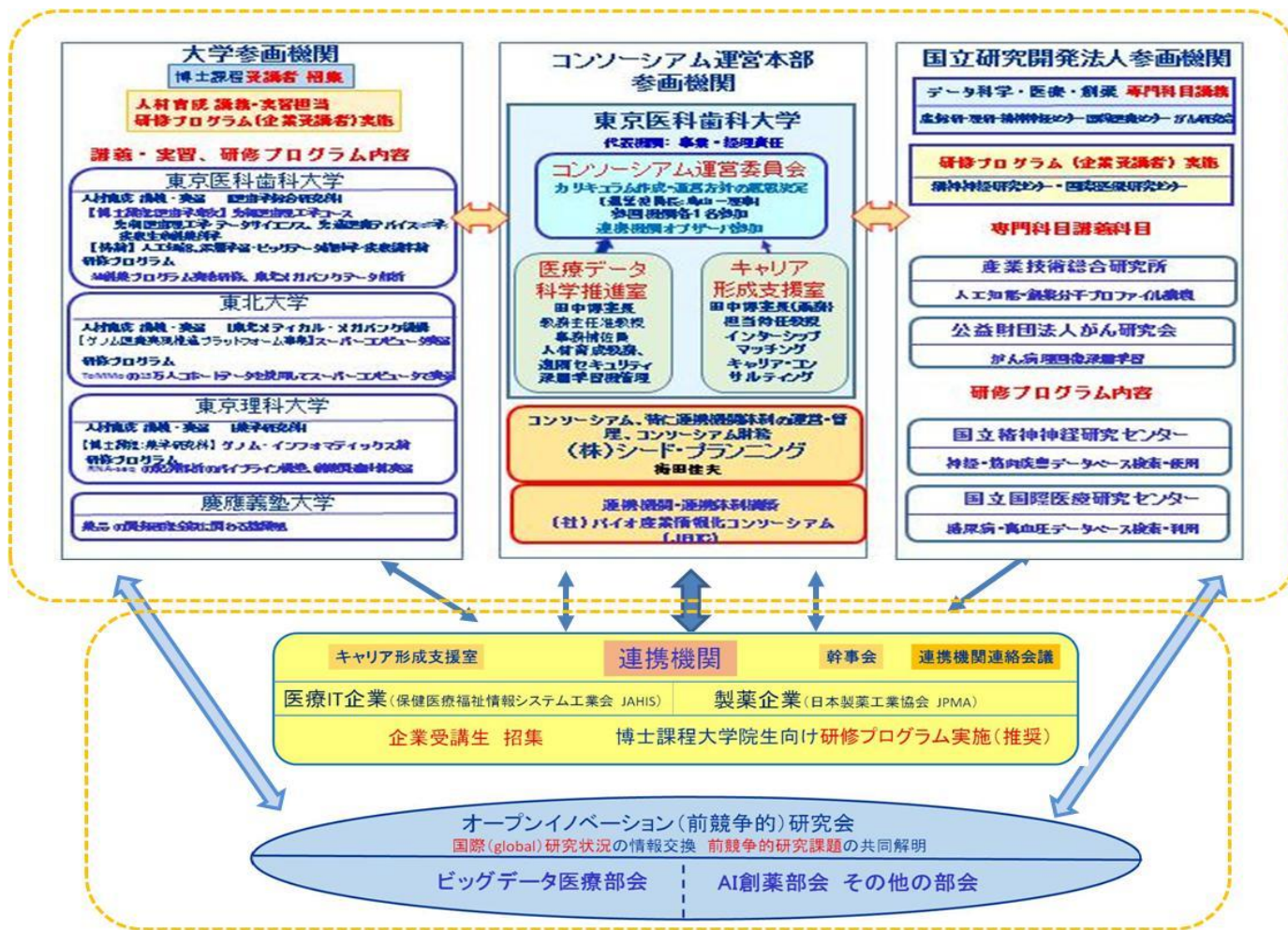
田中 博 著

「AI創薬・ビッグデータ創薬」

薬事日報社 6月23日刊行



医療・創薬 データサイエンスコンソーシアム(meddic)



武田薬品、アステラス、大塚、第一三共、中外製薬、田辺三菱、ファイザー、協和発酵キリン、小野薬品
旭化成、帝人ファーマ、大鵬薬品、日本たばこ、科研製薬、富士通、日本電気、味の素など

ご清聴有難う御座いました

