

ディープラーニングと生体ビッグデータを を活用した医療・創薬技術とその応用

東京医科歯科大学 医療データ科学推進室
東北大学 東北メディカル・メガバンク機構

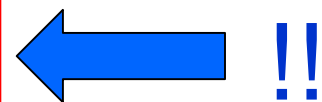
田中 博



医療・創薬分野への ビッグデータ時代の到来

- (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)

個別化医療・医療の国民レベルの向上
医療・生命情報の適確性の飛躍的な増大



医療におけるビッグデータ

I 次世代シーケンサによるゲノム情報

— 網羅的分子情報の急速な蓄積

II 大規模バイオバンクによる情報蓄積

— ゲノム情報・環境生活関連情報

III モバイルヘルスによる生理変量

— 連続的生理モニターによる情報蓄積

新しい
タイプの
医療ビッグ
データ

IV 表現型医療情報の蓄積

— 電子化の普及による医療情報の蓄積

旧来のタイプの
医療データの
大容量化

医療の「ビッグデータ革命」

～ゲノム・オミックスデータの基軸的な特徴～

＜目的もデータ特性も従来型と違う＞

従来の医療情報(IV)の「ビッグデータ」

N -Big Data ($n \gg p$)

医療情報・疫学調査では属性数：10項目程度

- 目的：Population MedicineのBig Data
⇒個別を集めて「集合的法則」を見る

網羅的分子情報(I~III)などビッグデータ

P -Big Data ($p \gg n$)

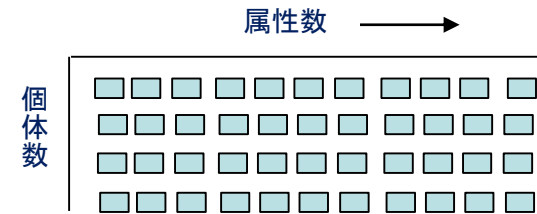
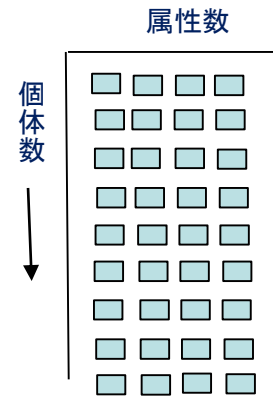
1個体に関するデータ属性種類数が膨大

属性に比べて個体数 少数:従来の統計学が無効

「新NP問題」：多変量解析:GWASで単変量解析の羅列

- 目的：例えば医療の場合Personalized Medicine

⇒大量データを集めて「個別化パターン」の多様性を抽出



データ科学の必要性

ビッグデータは医療のパラダイムを変革する

- 医療は近年大きくパラダイム変換しつつある。
- 2000年頃から、「ビッグデータ医療」の概念出現の前に、パラダイム変換の概念として、次の2つの概念が提示されてきた

個別化医療

Personalized Medicine

先制医療

Preemptive Medicine

個別化医療 personalized medicine

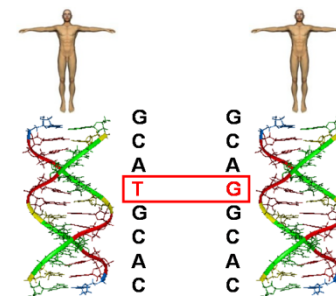
- 層別化医療(stratified),
- 2つの個別化医療の概念

— 先天的遺伝的素因

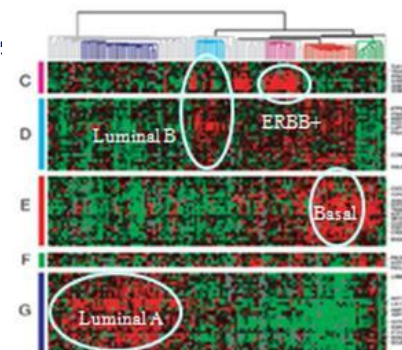
- 疾患原因遺伝子,
 - 家系調査/Linkage解析、先天的疾患
- 疾患感受性遺伝子
 - 多型性：一塩基多型(SNP),数千万種類
- 薬剤代謝酵素の多型性
 - 副作用・薬効無効の個人存在

— 後天的要因（疾患の個性）

- 後天的遺伝子変異（がん等）
 - 内在的サブタイプ



1塩基の変異が疾患や薬剤への応答性に関与する場合があります

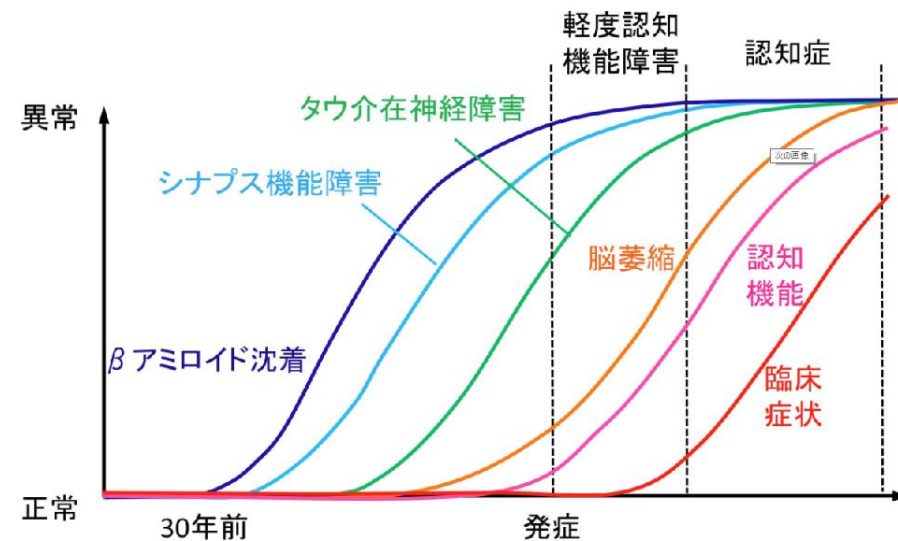


個別化医療による 医療のパラダイム変換

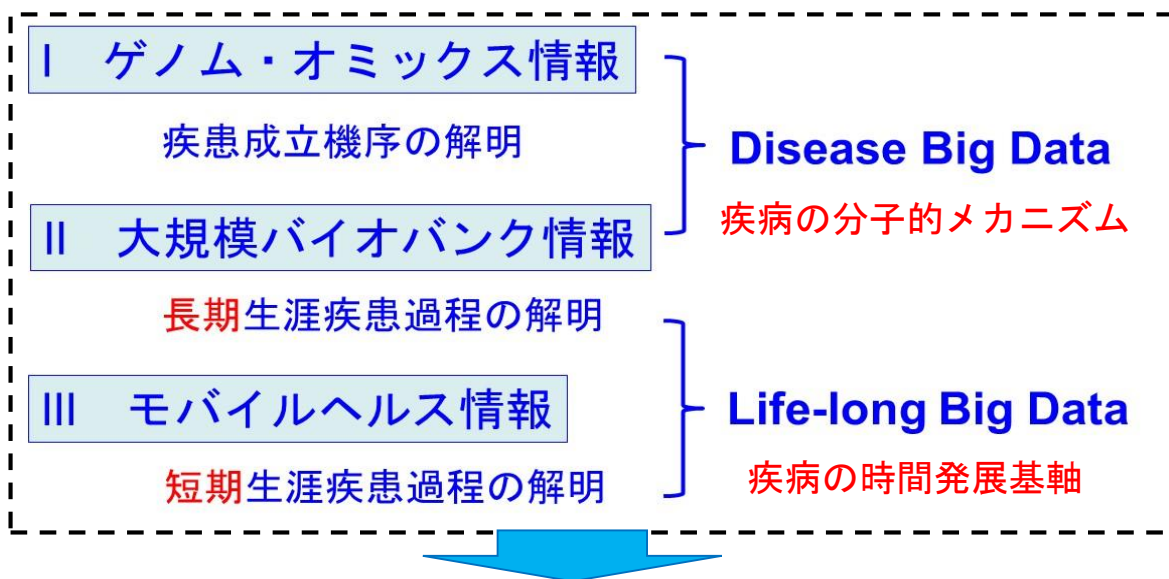
- 従来のpopulation医学<One size fits for all>はもはや成り立たない。
 - 同一の病名で括られているが、内在的亜型 (intrinsic subtype) が多数存在する
 - 第1種：薬剤代謝酵素の多型性
 - 第2種：がんの亜型解析と分子標的薬の選択
- 医療の隅々に浸透するポピュレーション医学の桎梏を克服

先制医療 preemptive medicine

- Preemptive medicine (2005)
 - NIH長官Zerhouni が2005年NIH長期計画で
 - “By making use of precise molecular knowledge (分子情報を使用) to detect disease before symptoms are manifest, and intervening before disease can strike.” (発症前に直撃)
- 従来の医療：疾病罹患後の医療
 - 対応的 (reactive)
 - 機会主義的 (occasional)
- 疾患罹患後の治療
 - 治癒の困難性
 - 例：アルツハイマー治療
 - 医療経済的圧迫
- 早期診断・早期治療
- 予測医療・予防医療(proactive)
- 疾患に関する見方の変化
 - 生涯にわたる疾患把握
- Life-course-oriented healthcare



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

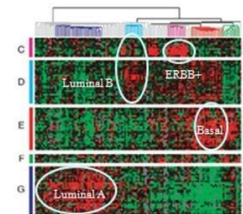
新時代の医学・医療

Population医療



個別化医療

内在的サブタイプ
(乳がん)

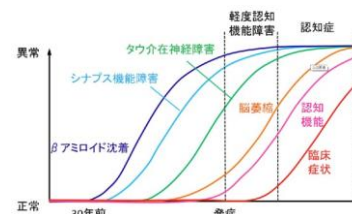


Reactive 医療



Proactive 医療

先制医療
(アルツハイマー)



Occasional 医療



Life-long 医療

ビッグデータ 医療時代へ

ゲノム・オミックス医療の2つの流れ

米国でのゲノム医療

Precision Medicine (精密医療)

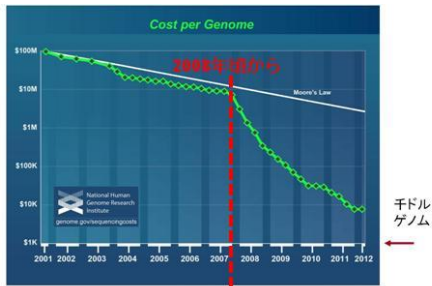
- 「シーケンス革命」(2007)からの怒濤の展開(2010から)
- 個々の患者の「治療医学」レベル質的向上:臨床実装の推進
 - 稀少疾患の原因遺伝子変異の同定
 - がんのドライバー遺伝子変異の同定と分子標的薬の選択
 - 薬剤代謝酵素の多型性の同定と個別化投与

欧州でのゲノム医療

Genomic Biobank (バイオバンク)

- 「集合的遺伝情報」の価値⇒ゲノム・バイオバンクへ流れ
- 国民医療(医療の国民レベル)の向上:社会福祉国家の理念
- 「予防医学」レベル質的向上のためにゲノム情報導入
 - 大規模前向きpopulation型バイオバンク/ゲノム・コホートの確立
 - 遺伝的素因と環境要因(生活習慣)との相互作用に基づいた「多因子疾患」の発症予測を通じた「国民医療の向上」
 - 生涯的健康/疾病管理へ

米国ゲノム・オミックス医療の流れ



DNA Sequencing Cost: the National Human Genome Research Institute

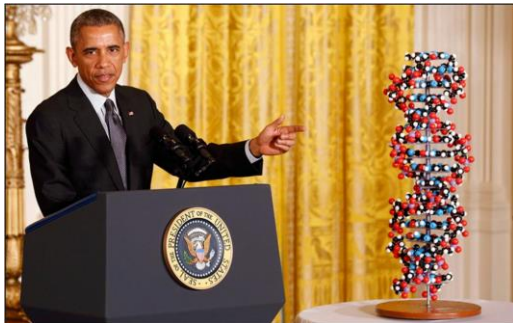
シーケンス革命 2007/8

2005~ NGS 454 (LS,Roche)
2007/8~454, Solexa (Illumina),
SOLiD (LT,TF)
シーケンス革命



	HiSeq2500	Ion Proton
本体価格	約1億円	約3500万円
モード / チップ	ハイアウトプット	ラビッドラン
解析時間	11日	27時間
リード長 (bp)	2 x 100	2 x 150
データ産出量 (Gb)	約600	約120
試薬コスト (ヒト1人全ゲノム)	数十万円	不可 エクソームのみ

急速な高速化と廉価化
ヒトゲノム解読計画13年,3500億円
⇒1日,10万円



オバマ大前統領 Precision Medicine Initiativeを
開始、2015年1月 大統領一般年頭教書演説

先陣争いの時代

第一期

ゲノム多型性の認識
Hapmap 計画(2002)
GWAS研究など

2005~ NGS 登場
(454,Solexa,SOLID)
2007/8~
シーケンス革命

国際がんコンソーシアム開始
ICCG (2008年)
2011頃からがん
変異成果報告

薬剤代謝酵素の多型性の判別・電子カルテで警告・Preemptive PGx
Vanderbilt大病院

Undiagnosed Genetic Diseaseの原因遺伝子POC同定
MCW小児病院

Cancer Driver Geneの同定と抗がん剤治験
Dana Faber CC

2007年
2009年
2010年
2011年
2012年
2013年
2014年
2015年
2016年
2017年

ゲノム・オミックス医療の臨床実装の普及
ゲノム・オミックス情報のビッグデータの出現

ゲノム医療の国家的取組み
NIH "BD2K" 計画・各種ゲノムコンソーシアム開始

オバマ大統領 年頭教書
Precision Medicine initiative 政策の発表

国家政策の時代

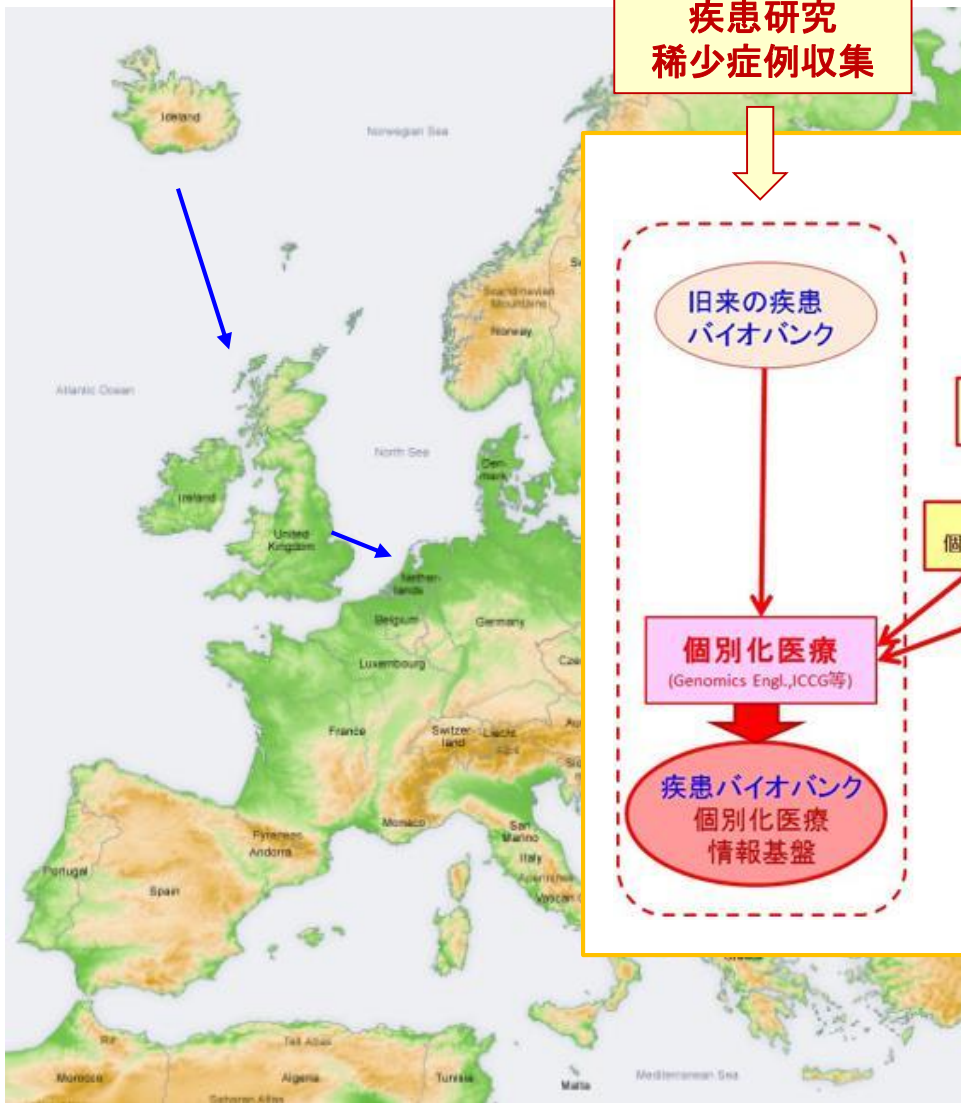
第二期

精密医療普及期

第三期

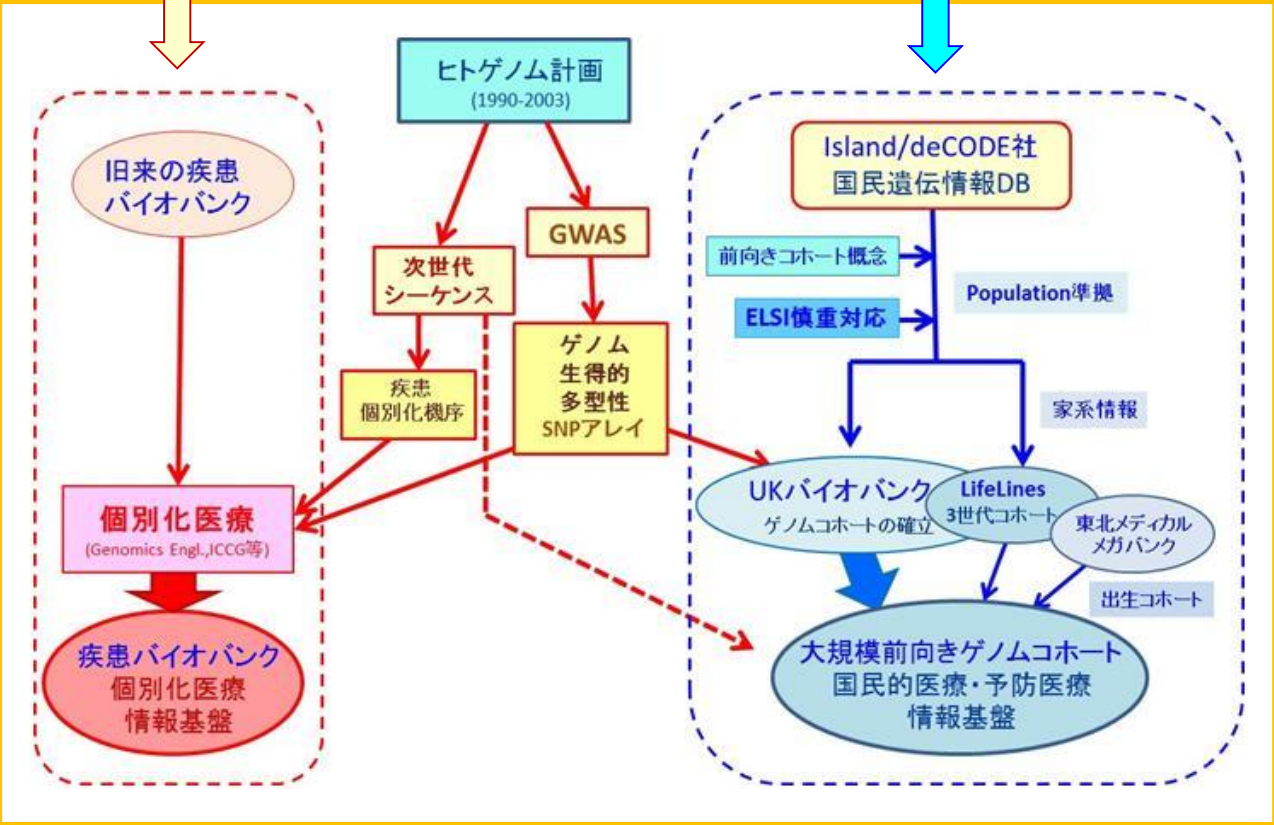
100万人コホート:バンダービルト大学設開始(2016-2020)
NCI "National Cancer MoonShot" 10年計画開始
各州でのプレジジョン医療計画開始(カリフォルニア、ペンシルバニア)

第2の流れ 欧州のバイオバンクの普及



疾患研究
稀少症例収集

「集合的遺伝情報」による
国民レベルでの医療向上



ビッグデータ医学/医療の2つの流れに起因する 大規模な生命情報DB/KBの出現と利用

- ヒトゲノム解読計画以降急速に進展
 - Hapmapプロジェクト, 1000 genome, がんICGC, TCGA, TopMED
 - ゲノム変異・多様体
 - dbSNP, HGMD, **Clinvar**, **Clingen**, OMIM, GWAS catalog
 - 表現型との対応: dbGaP, EGA
 - 遺伝子発現プロファイル
 - 疾患特異的transcriptome: **GEO**, **ArrayExpress**,
 - 薬剤特異的transcriptome: **c-Map**, **LINCS**
 - タンパク質
 - 3次元構造: PDB, Swiss-Prot,
 - タンパク質間相互作用: **HPRD**, **STRING**, BIND
 - 分子ネットワーク、パスウェイ
 - KEGG, TRANSFAC, BioCyc, Reactome
- 各種バイオバンク症例ベース（制限アクセス）
 - UK biobank, BMBRI, 東北メディカル・メガバンク
- これらの大規模DB/KBを組合せてゲノム医療/創薬を推進

医学/医療へのビッグデータの衝撃

	HiSeq 2500	HiSeq Proton
本体価格	約1億円	約2500万円
モード / チップ	ハイブリッド / フロント	カブリッド / フロント
解読速度	118	2798
リード長 (bp)	2 x 100	2 x 150
データ量 (G)	8960	8320
設置コスト (ヒト1人ゲノム)	約1万円	約1万円

次世代シーケンサの登場 シーケンス革命 (2007)



コストレスで高精度な網羅的分子情報の出現

1. ゲノム・オミックス医学/医療の進展

- Clinical Sequencingによるゲノム・オミックス医療の臨床実装の急速な進展

2. Biobank/ゲノムコホートの世界的普及

- 個別化医療/予防の情報基盤として普及

3. 大規模な生命情報DB/KBの出現

- ゲノム・オミックスによるDB/KBの膨大化

医療の「ビッグデータ」革命は どんな既存のパラダイムに変革しているか

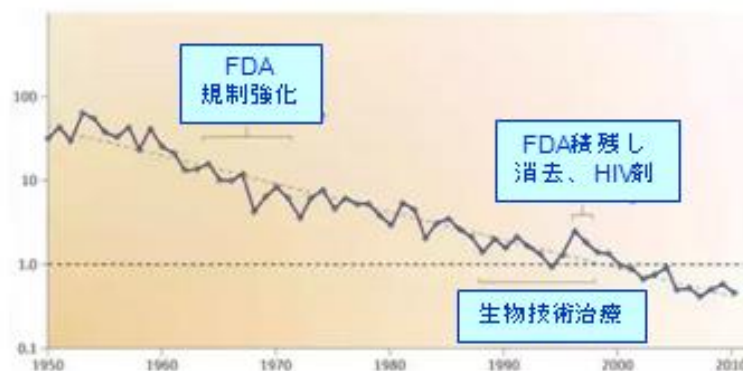
- Population（集会的）医学からパラダイム転換
 - <One size fits for all>の集会的医療はもはや成り立たない
 - 個別化医療“Personalized medicine”の概念
 - 個別化医療実現のために<個別化・層別化パターン>がどれだけ有るか
網羅的に調べる：どこまでの粒度で個別化・層別化すればよいか
- Clinical research（臨床研究）のパラダイム転換
 - 臨床研究の基礎：従来の範型RCTは、個別化概念を取扱えない
 - <EBM: (statistical) evidence based>の呪縛からの解放
 - 「標本」統計・「推測」統計学に制約されない臨床研究
 - Real World Data・ビッグリアルワールドデータからの知識生成
 - Learning Health System: 学習的医療実践

生命情報ビッグデータを利用した 創薬戦略の進展へ

創薬をめぐる状況と解決の方向

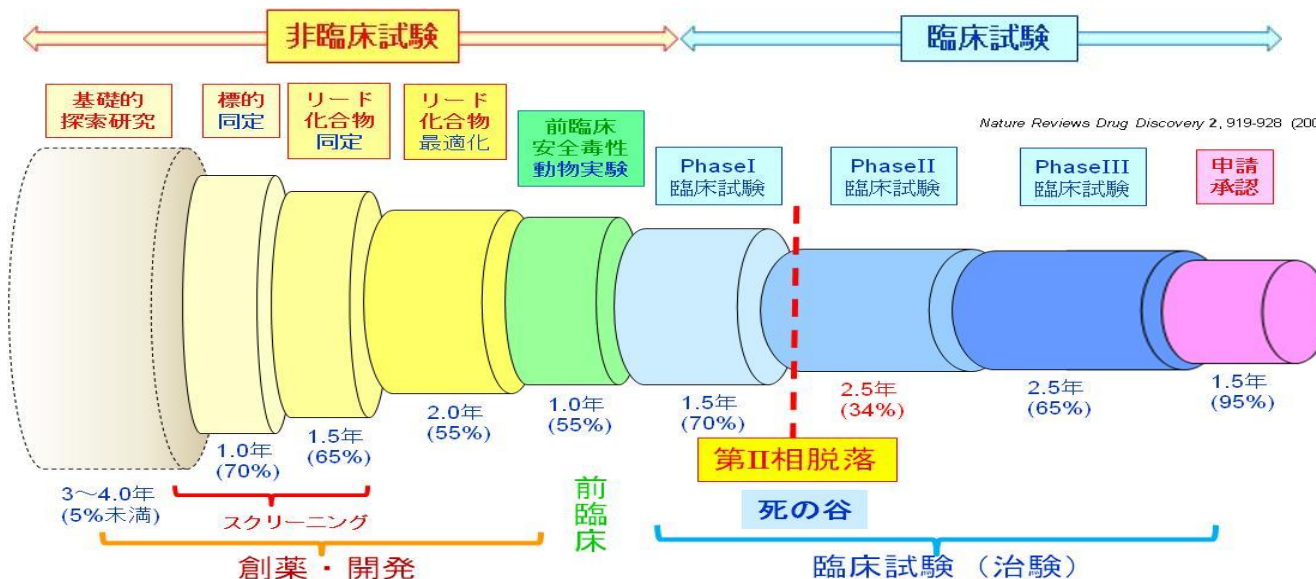
- 医薬品の開発費の増大
 - 1 医薬品を上市するのに約1000億円以上
- 開発成功率の減少
 - 2万~3万分の1の成功率
 - とくに**非臨床試験**から**臨床試験**への間隙
 - **phase II attrition** (第2相脱落)
- 臨床的予測性
 - 医薬品開発過程の**できるだけ早い段階**での**有効性・毒性の予測**
- **臨床予測性の早期での実施**
 - 罹患者のiPS細胞を使う

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)

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。

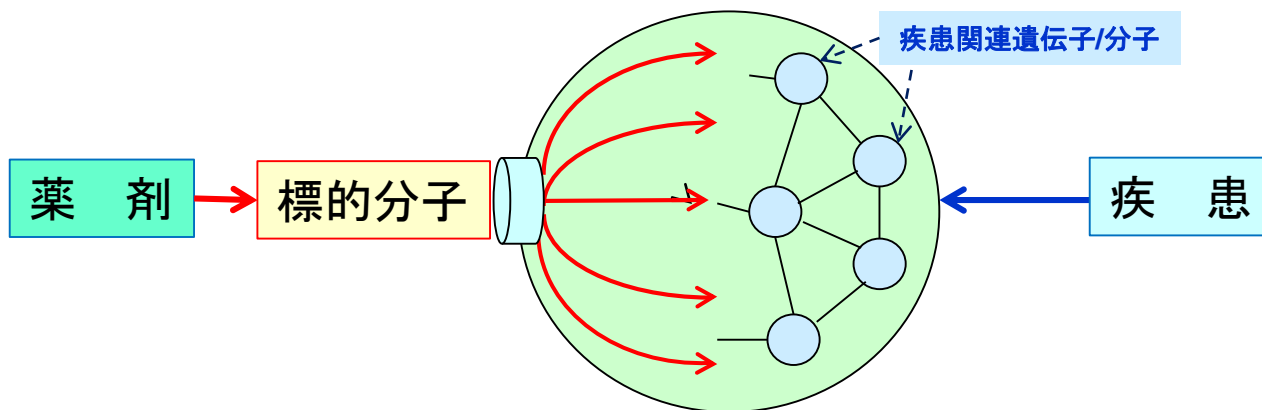
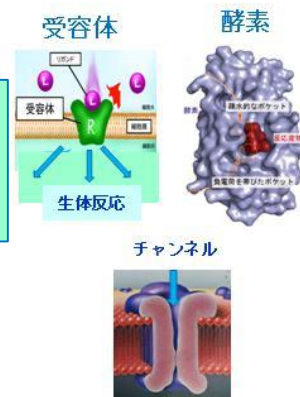
疾患・薬剤・標的の関係

病気の主要な要因

疾患関連遺伝子/分子(複数)

薬：薬剤標的分子を通して作用する
疾患関連遺伝子/分子に影響を持つ

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

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

生体プロファイル型計算創薬・DR

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

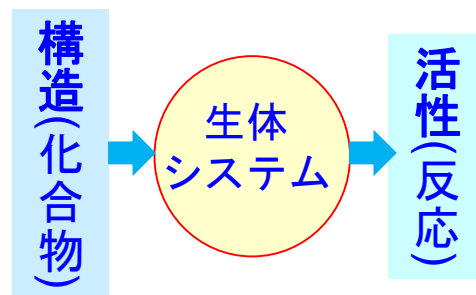
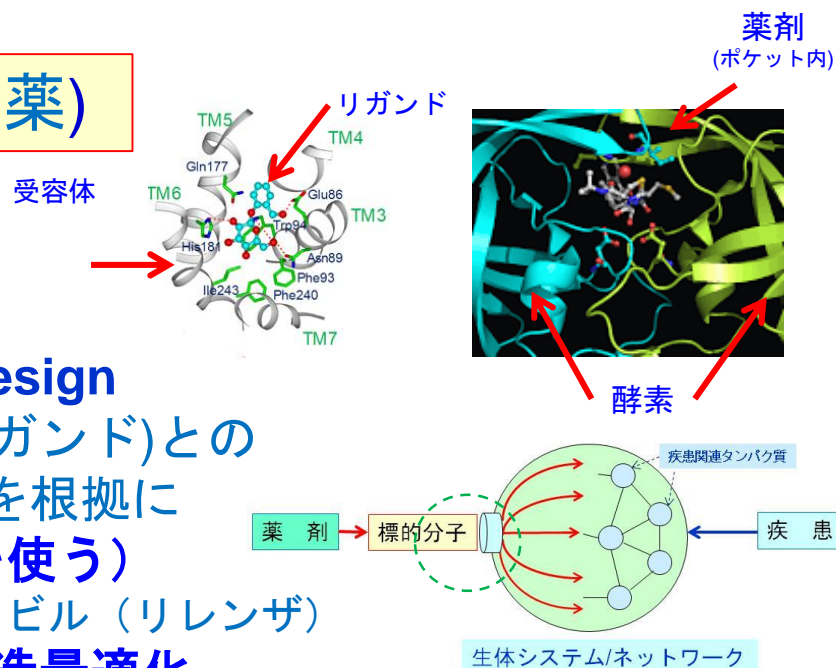
これまでの計算創薬 (*in silico* 創薬)

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



生体分子プロファイル型計算創薬

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

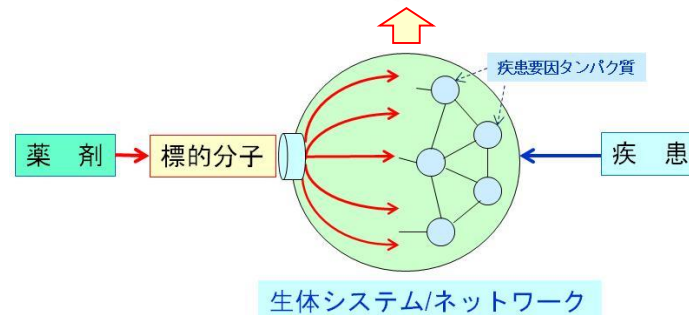
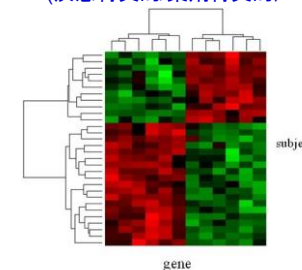


薬剤投与による

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

遺伝子発現プロファイル変化
(疾患特異的/薬剤特異的)



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

創薬・DRの基本的枠組み

疾患・薬剤・生体系をネットワーク間の関係として認識

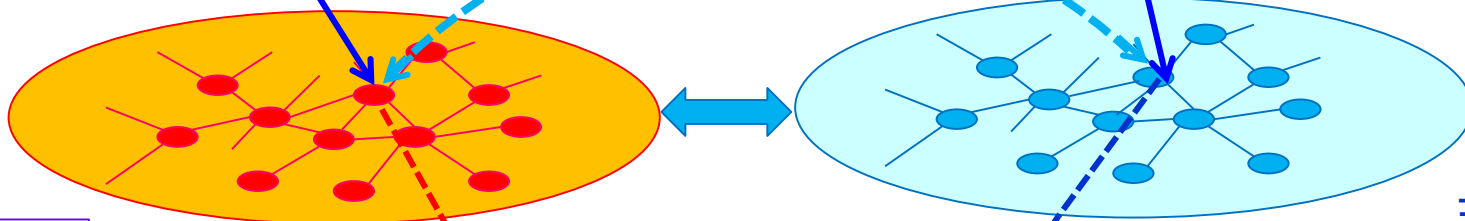
薬剤Cは疾患Dに薬効

疾患ネットワーク

疾患D

薬剤C

薬剤ネットワーク



現象

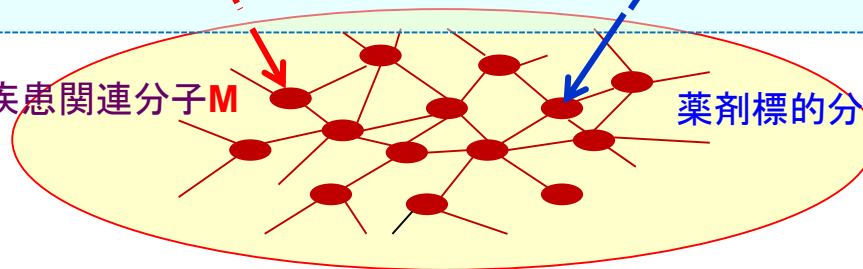
プロファイル比較型
創薬/DR

機構

分子ネットワーク型
創薬/DR

疾患関連分子M

薬剤標的分子T



生命システム

生体分子プロフィール型の 計算創薬・DRの 体系

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
 - 疾患ネットワークと薬剤（化合物）ネットワークの比較
- 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

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
- その他
 - 化合物自動設計、合成経路探索、毒性予測

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

3層の生体・薬剤のネットワーク間の関係図式

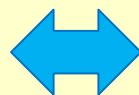
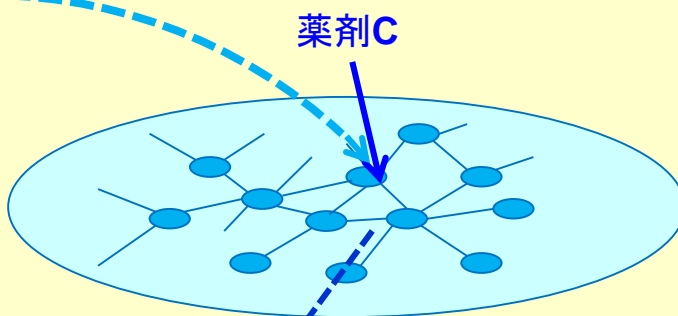
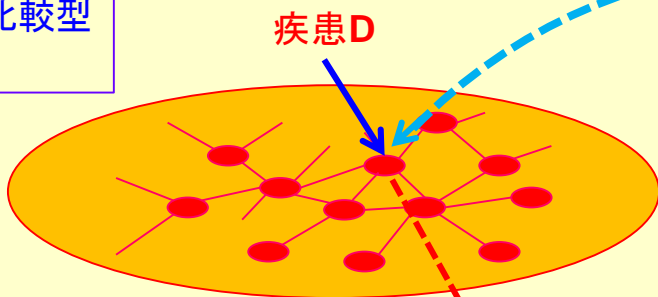
現象的マクロ的対応

薬剤ネットワーク

薬剤Cは疾患Dに薬効

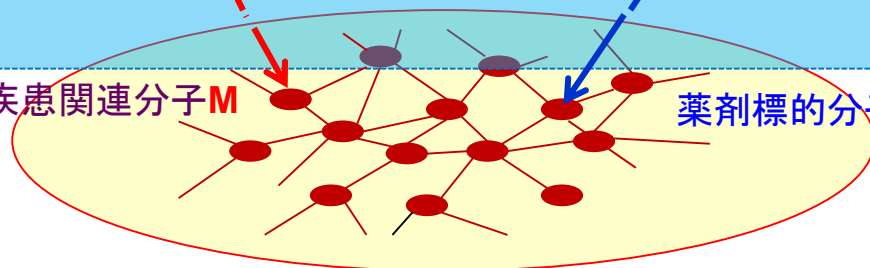
疾患ネットワーク

プロフィール比較型
創薬/DR



分子ネットワーク型
創薬/DR

疾患関連分子M



薬剤標的分子T

機構

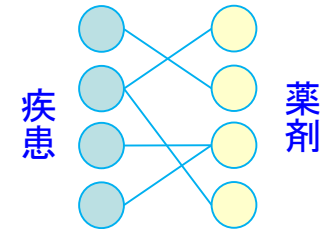
生命システム

生体分子プロファイル型創薬/DR 非学習アプローチの深化

生体分子プロファイル比較

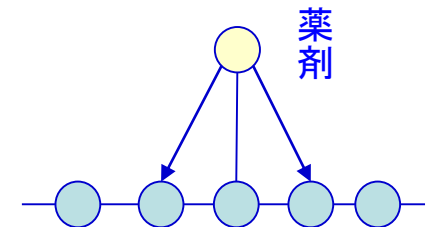
第1段階：疾患・薬剤プロファイル直接比較

- 疾患罹患時と薬剤投与時の生体反応の遺伝子発現プロファイルを比較。
- パターン正負相関性に基づく有効性毒性予測



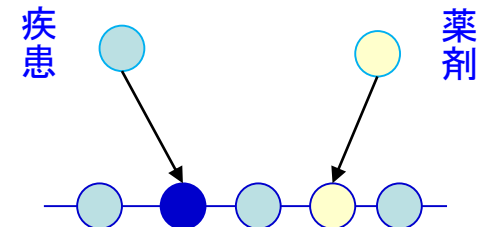
第2段階：疾患・薬剤ネットワーク近接解析

- 疾患あるいは薬剤の集合をネットワーク表現
- ネットワーク近接性に基き有効性・毒性予測



第3段階：生体ネットワーク媒介型比較

- 生体分子ネットワークを<場>として、疾患・薬剤の作用の足場分子を同定
- 足場分子間の相互作用（総合的距離）の評価に基づき有効性・毒性予測



生体分子ネットワーク

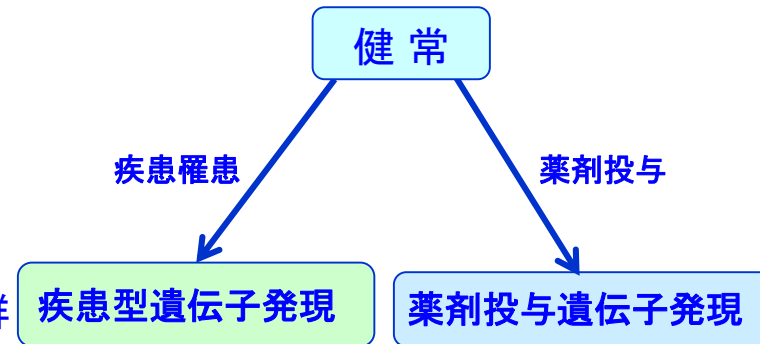
1. 遺伝子発現プロファイル 直接比較型

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
- シグネチャ (署名) : 差別的発現遺伝子代表群
- DB利用 : シグネチャを「問合せ」 :
類似性の高い順に化合物のプロファイルを示す
- 最近はLINCSデータベース : 100万種の薬剤特異的発現DBを搭載



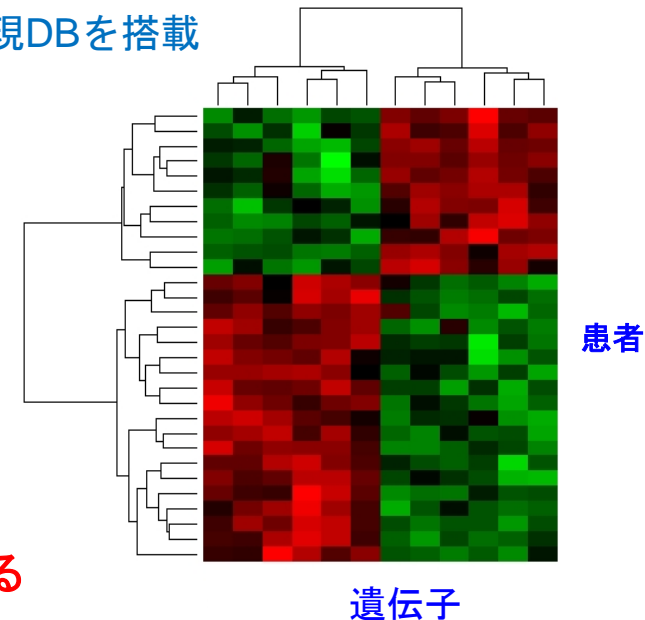
● 疾病特異的遺伝子発現

— 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)
- 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

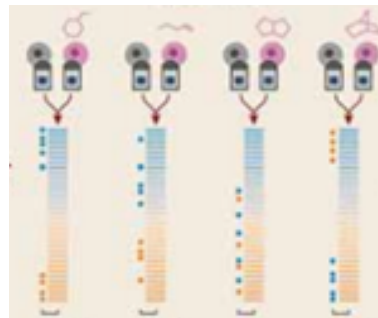
- 疾患によって**健常状態から変異**
「疾患特異的遺伝子発現プロファイル」
- これに**薬剤投与の変化を起こす**
「薬剤特異的遺伝子発現プロファイル」
- **両者のパターンが負に相関する**
- **ノンパラメトリックな相関尺度で評価**

● 効果が相加的なら**有効性**が期待される

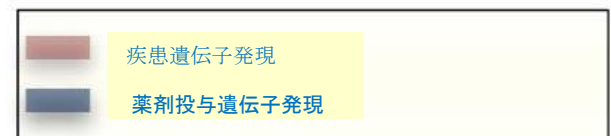
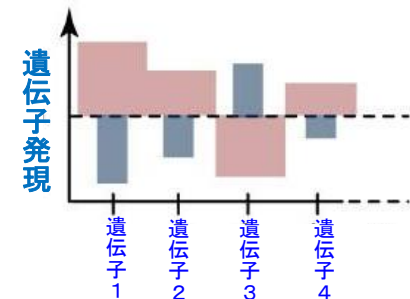
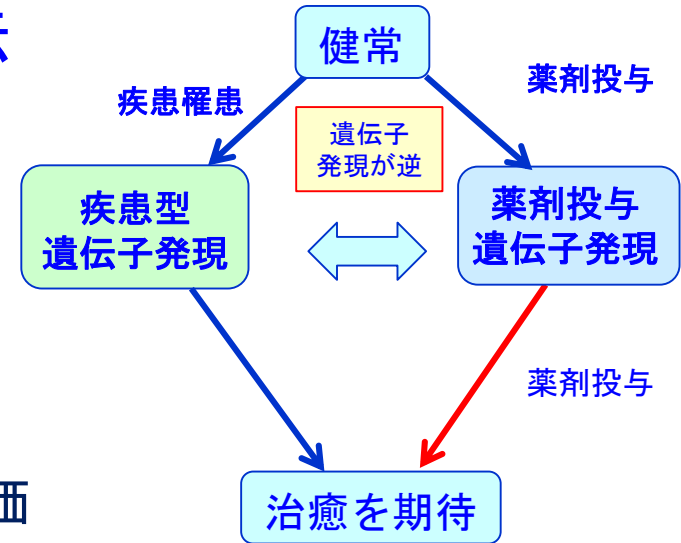
- 例：炎症性腸疾患に抗痙攣剤(topiramate), 骨格筋委縮にウルソール酸

DB格納プロファイルは**薬剤**遺伝子発現
問合せ（横点）署名は**疾患**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



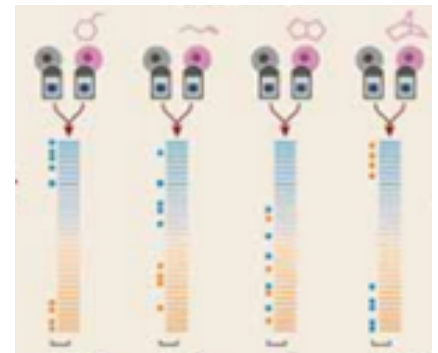
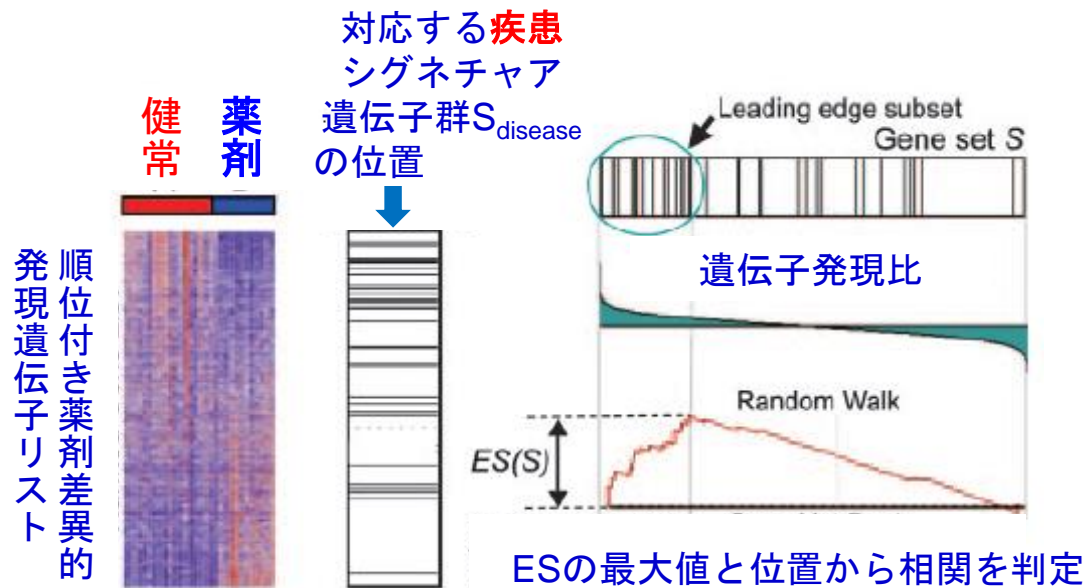
強正 弱正 弱負 強負



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_{disease} 中に該当する遺伝子が存在したらES (Enrich Score)を加算、無ければ一定値を減算

遺伝子発現プロファイルによる毒性予測

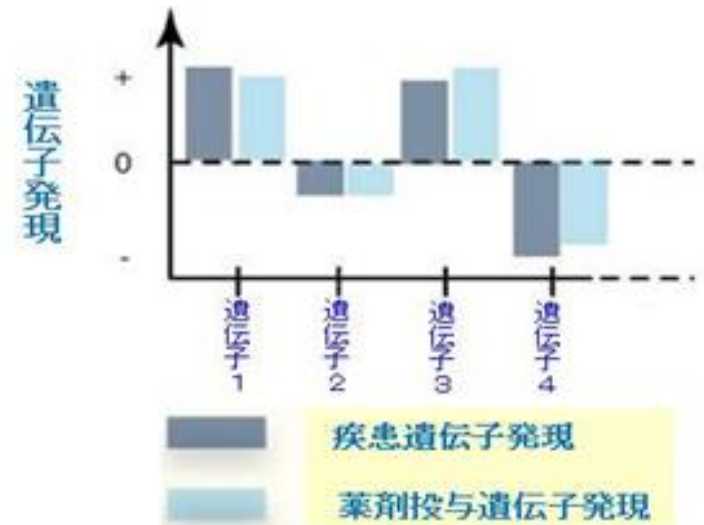
- 連座法 guilt-by-association :

- **薬剤－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の 正・負のパターン相関性に基づく計算DR

(Sirota, Butte 2011)

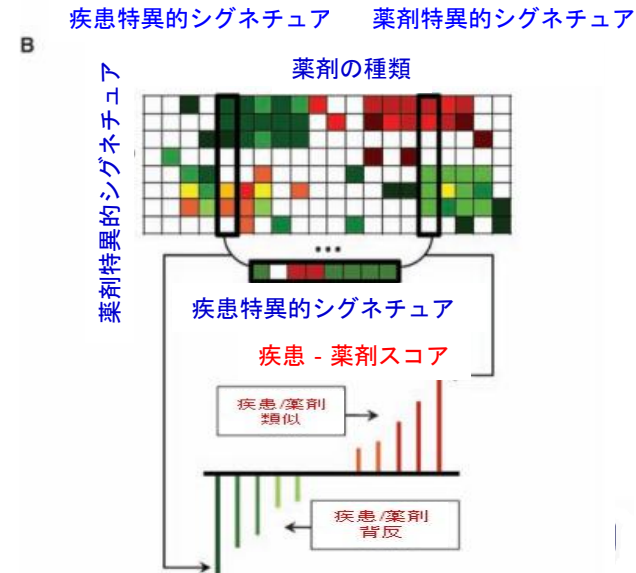
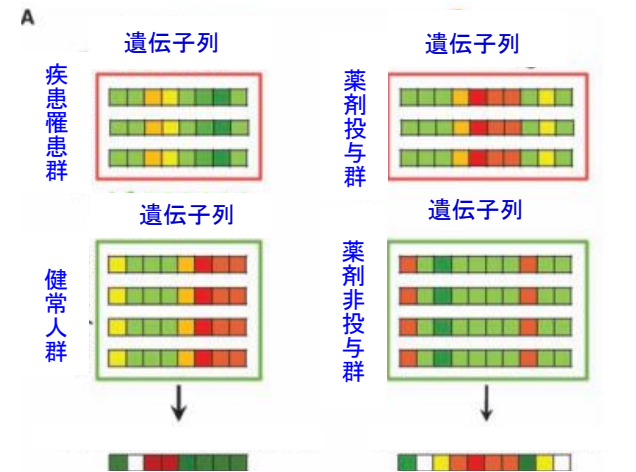
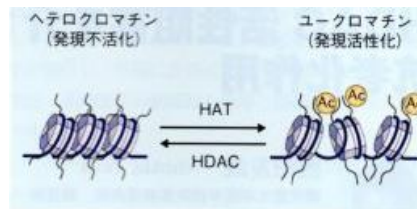
- NCBI・GEOから100疾患のシグネチャを取得
- c-Mapより得た164の薬剤・化合物の
薬剤特異的遺伝子発現プロファイル
疾患-薬剤間で類似性スコアを計算
- 約16000組の疾患-薬剤間の2664組が
有意、半数以上が治療的関連(負)あり
- 100疾患内, 53疾患が有意に164薬剤と関連

適応の多い薬剤と疾患

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- α -kynurenine		15	Malignant neoplasm of stomach	66
5114445		15	Dermatomyositis	63
Dexverapamil		15	Malignant mesothelioma of pleura	53
Prochlorperazine		15	Primary cardiomyopathy	48

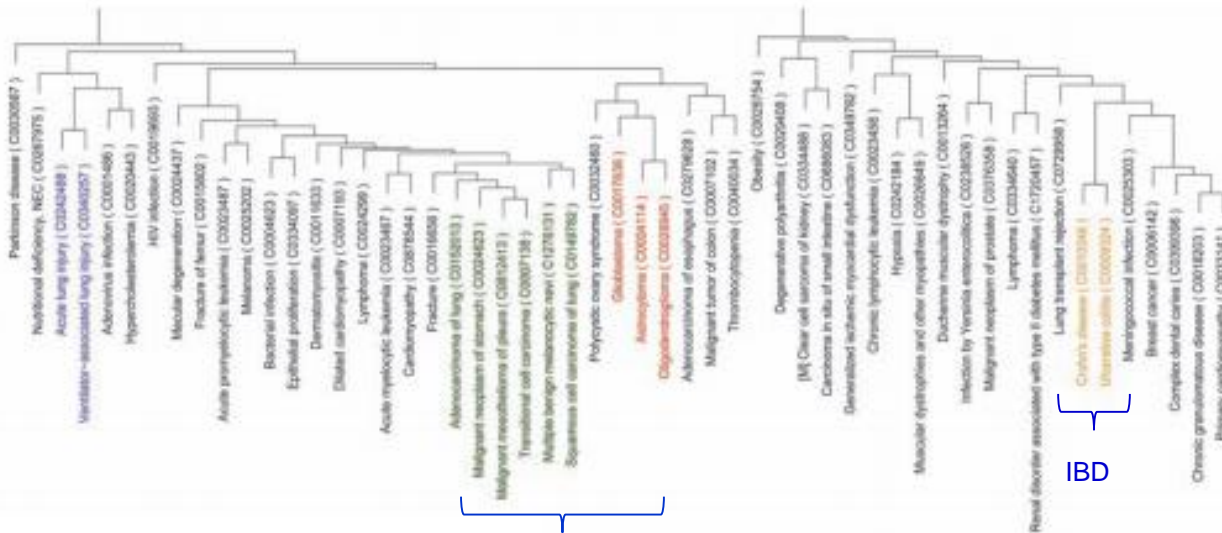
Drug group	Drugs
PI3K inhibitors	LY-294002 and wortmannin
HSP90 inhibitors	Geldanamycin, raloxifene, monorden, and sodium phenylbutyrate
HDAC inhibitors	Vorinostat, HC toxin, and trichostatin A
Salicylate anti-inflammatory agents	Sulfasalazine, mesalazine, and acetylsalicylic acid

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



遺伝子発現プロファイル比較による 疾患－薬剤関係に基づく計算/DR

疾患群の階層的クラスター



疾患のクラスター解析

- ・がんの大クラスター
- ・IBD: 潰瘍性大腸炎、クローン病のクラスター

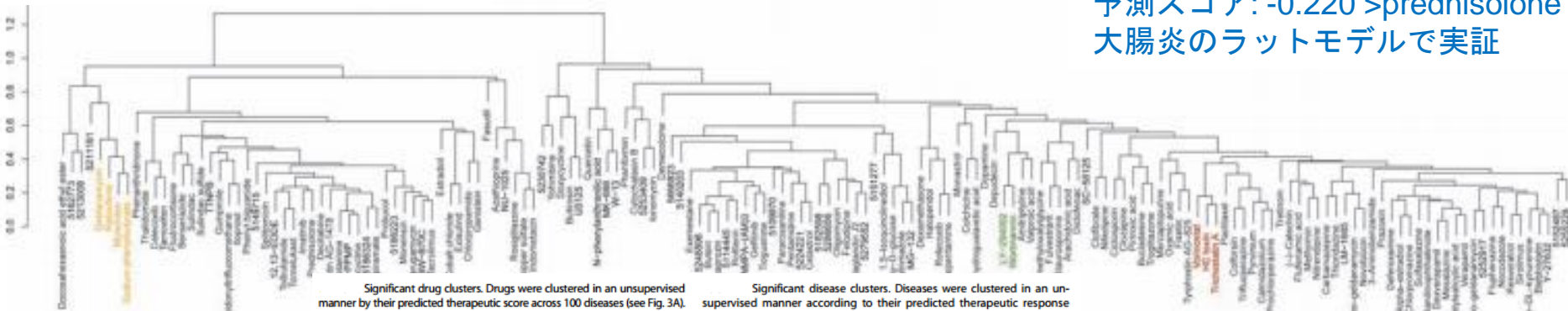
薬剤のクラスター解析

- ・HDAC阻害剤
HC toxin, trichostatin
- ・PI3K, 抗炎症剤クラスター
- ・HSP90関連薬剤

DR候補:

Topiramate(抗痙攣剤) IBDに有効
予測スコア: -0.220 > prednisolone
大腸炎のラットモデルで実証

薬剤群の階層的クラスター



Significant drug clusters. Drugs were clustered in an unsupervised manner by their predicted therapeutic score across 100 diseases (see Fig. 3A).

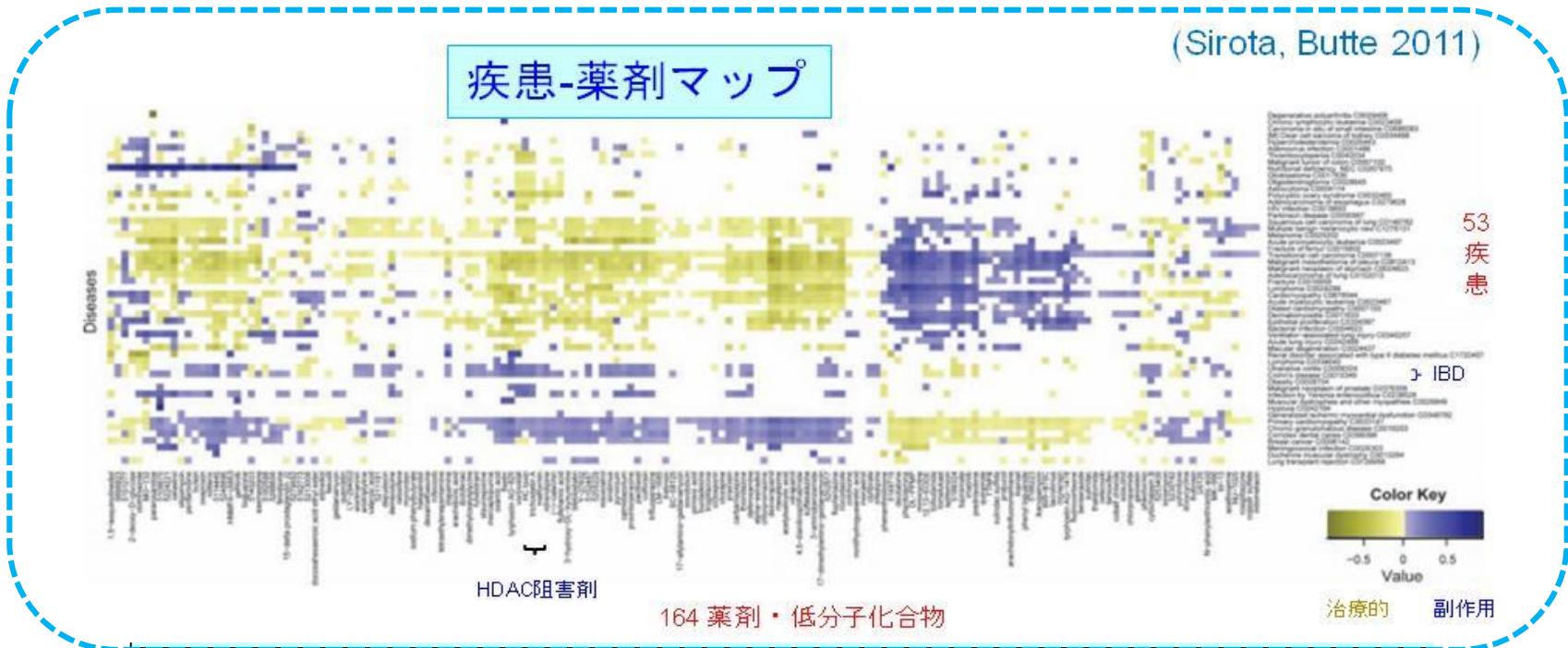
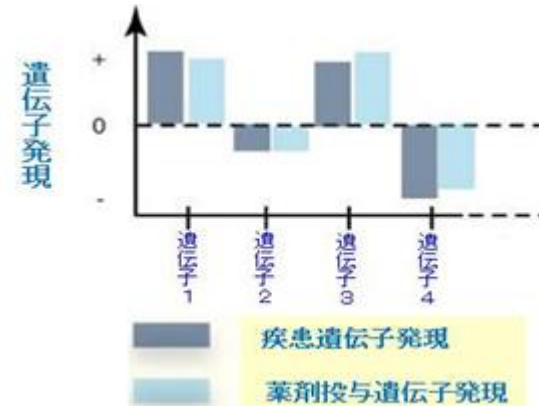
Drug group	Drugs
<u>PI3K inhibitors</u>	LY-294002 and wortmannin
<u>HSP90 inhibitors</u>	Geldanamycin, raloxifene, monorden, and sodium phenylbutyrate
<u>HDAC inhibitors</u>	Vorinostat, HC toxin, and trichostatin A
Salicylate anti-inflammatory agents	Sulfasalazine, mesalazine, and acetylsalicylic acid

Significant disease clusters. Diseases were clustered in an unsupervised manner according to their predicted therapeutic response across 164 drug compounds (see Fig. 3B).

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

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**



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

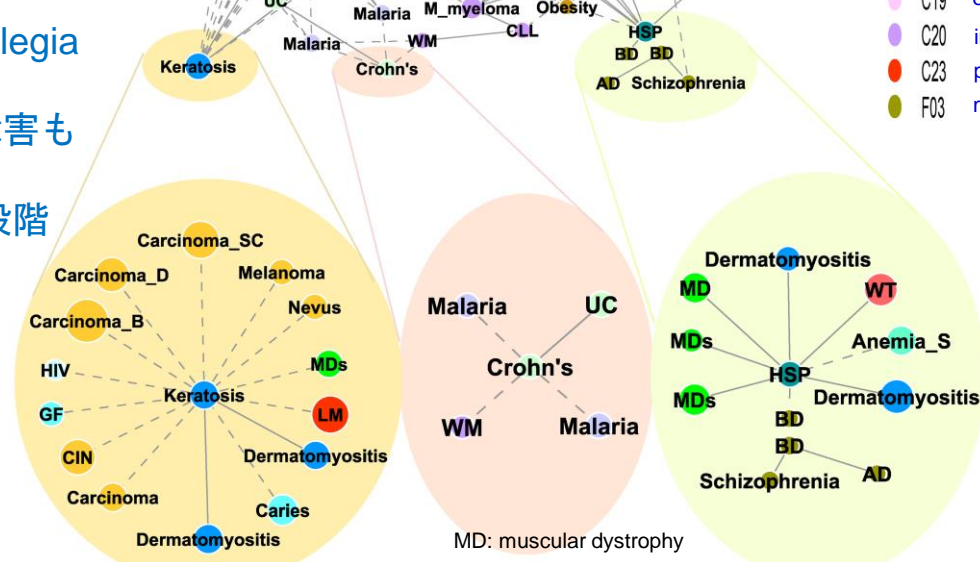
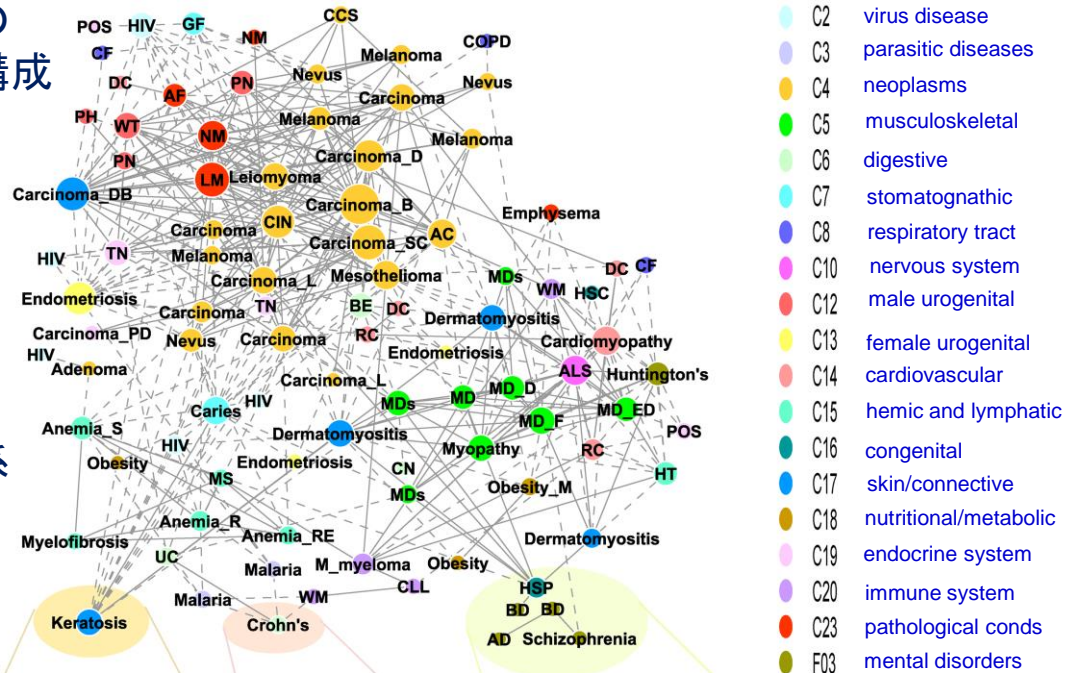
⇒ cancer(squamous) --前癌段階

<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



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 正值 破線 負値

Tamoxifen (乳がん医薬)

有効性 負の値をもっている

⇒アトピー,

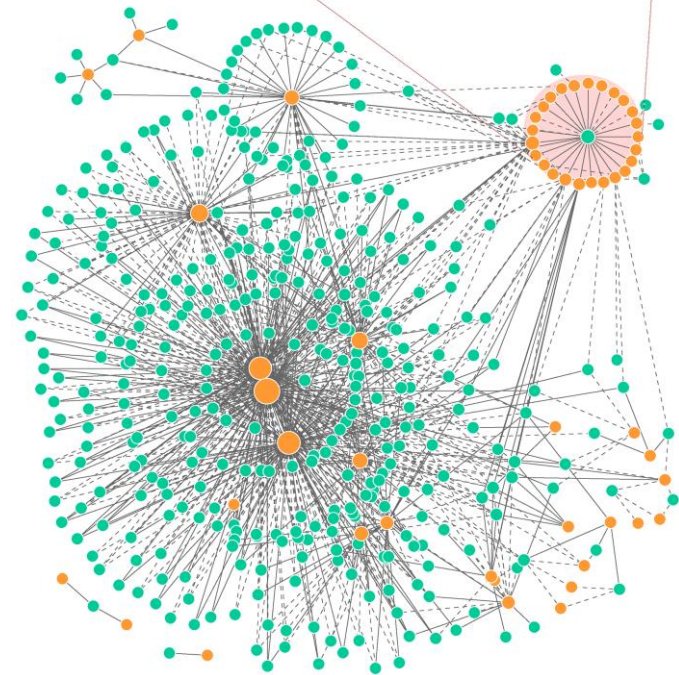
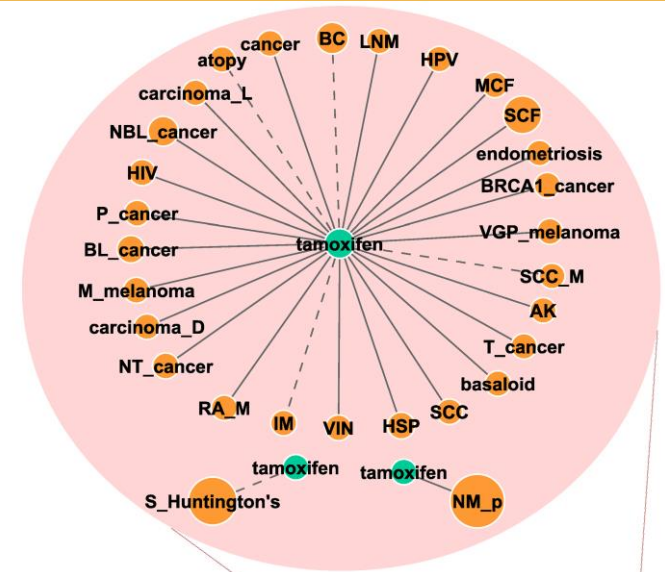
⇒マスト細胞分泌抑制、
アレルギー抑制

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

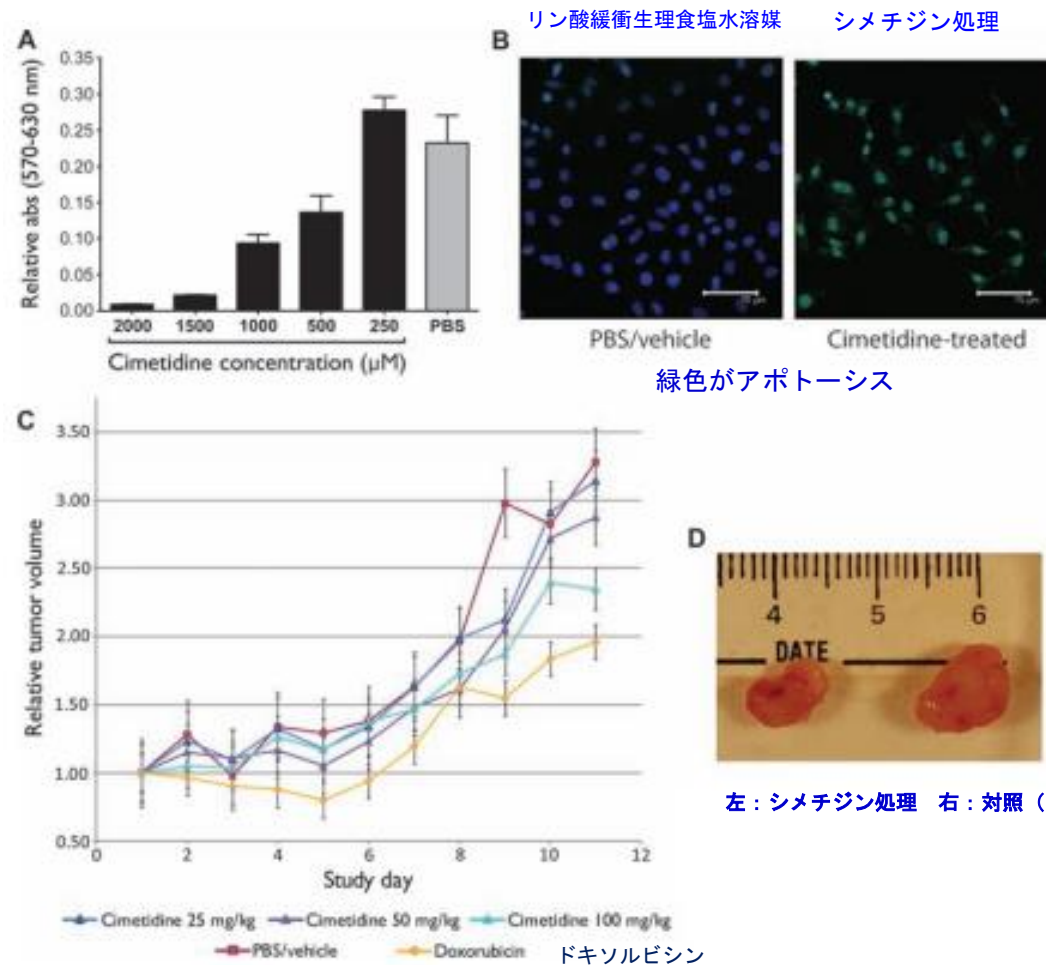
⇒ 発癌性



疾患－薬剤ネットワーク

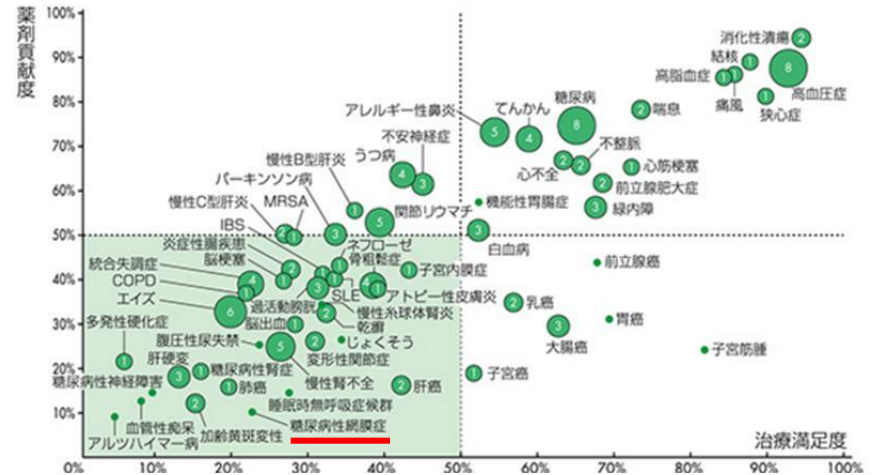
動物実験での実証

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



我々の研究室での成果

- 対象疾患 (Shibata et al. 2016)
 - 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



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

- 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), **タブシガルギン (小胞体ストレス誘導)**
 - 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

	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

- 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

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⇒差別的発現 **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ー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

LINCSの問合せ画面

--- LINCS Canvas Browser ---

Gene Lists

Up List

- EEF1A2
- UBE2S
- FAM64A
- FGFR1
- PAXIP1
- SPARC
- SNRPA1
- ADAMTS1
- EIF4EBP1
- PFKP
- BTG2
- CDK16
- ERRFI1
- ARPC4
- IFI30

clear

Down List

clear

Up Down

Search Example Enrich

Aggravate Reverse

Top 50 Consensus Experiments (Down/reverse)

Overlap	Info (Perturbation, Dose, Time, Cell, Batch)
0.5000	Tyrphostin AG 1478.56.78 μm 24 h A375 CPC006
0.5000	PD0332991.2 μm 24 h MDAMB231.LJP001
0.5000	PD0332991.10 μm 24 h MDAMB231.LJP001
0.5000	PD0332991.10 μm 24 h MCF10A.LJP001
0.5000	Aminopurvalanol A.10 μm 24 h PC9 CPC002
0.5000	3,5-dichloro-2-hydroxyphenylphenyl)benzenesulfonami
0.4800	PD0332991.2 μm 24 h BT20
0.4800	PD0332991.10 μm 24 h BT20
0.4800	MLN2238.10 μm 24 h BT20
0.4800	2-(6,6-dimethoxy-3-oxo-1,2,3,4-tetrahydro-1H-benzodiazepin-5-yl)carbamoyl)phenyl)propan-1-amine 3.10 μm 24 h A375

Showing 1 to 10 of 47 entries

Average Change - Time Point - Drugs - Dose

IL1
100 ng/μl, 6 h
in BT20

contrast:

Avg. Z-score:

Select a cell line: BT20

Select a batch: LJP004

Multiple Selections:

2. 疾患・薬剤ネットワーク 近接解析

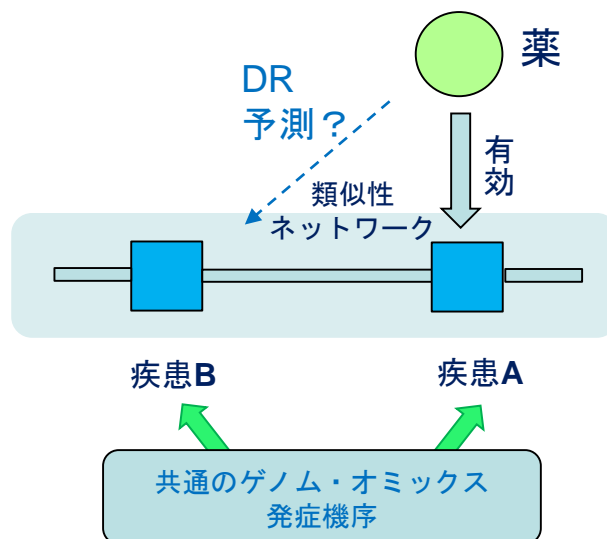
遺伝子発現プロファイルから 疾患・薬剤ネットワークへ

- 遺伝子発現プロファイルを基礎にした「シグネチャア逆位法」
 - 広くDR候補を枚挙する方法としては有効
 - もう少し対象を絞り込む方法はないか
 - 疾患の内因的機序(ゲノム・オミックス機序)によるDRは可能か
- 疾患のゲノム・オミックス機序による**疾患間のネットワーク化**
 - 同様に薬剤間のネットワーク化も可能

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



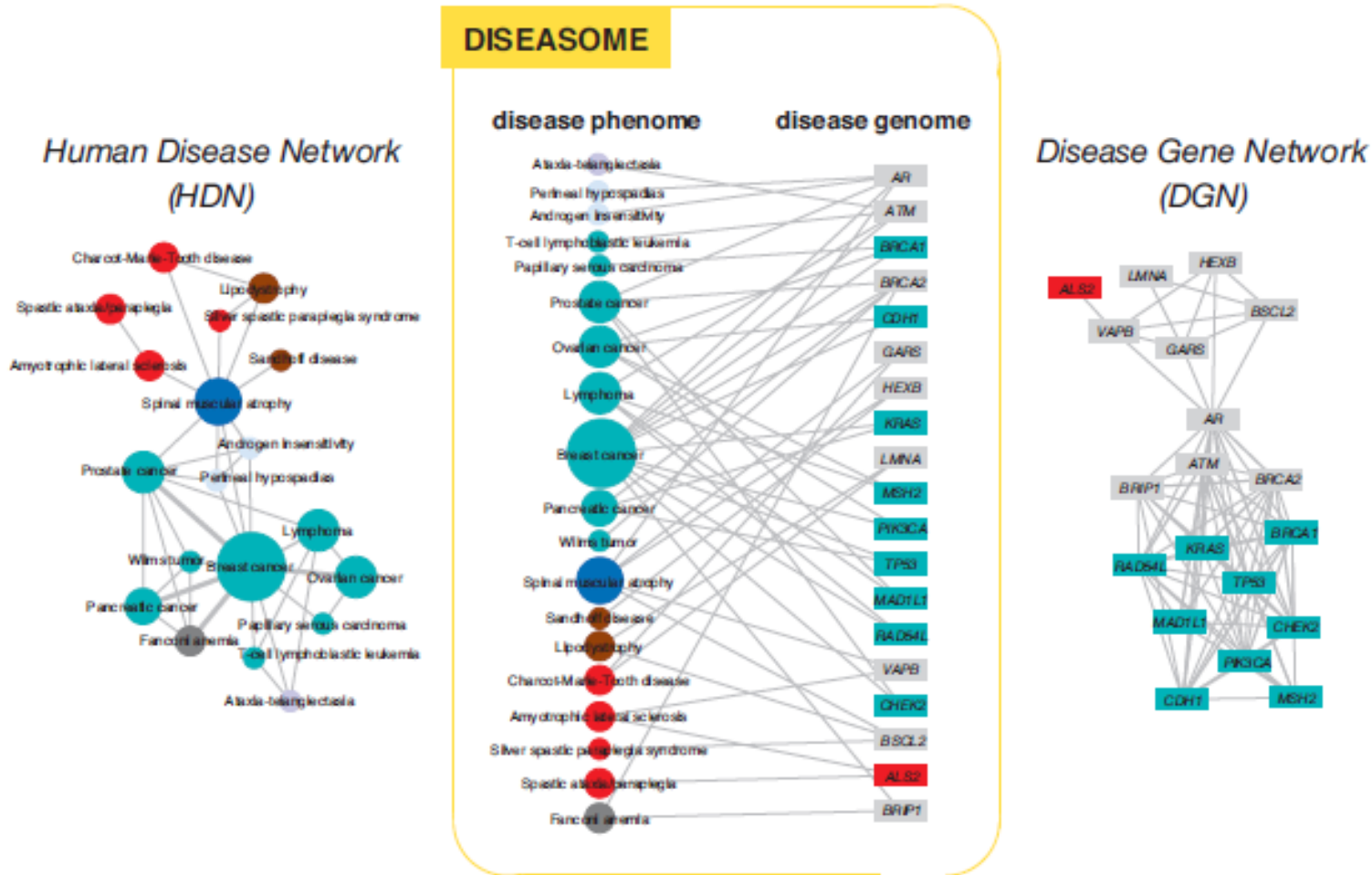
疾患形成のゲノム・オミックス機序

- 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、「分子表現型 molecular phenome」
 - Trans-disease 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
 - 大半の疾患（先天的稀少疾患を除く），“common disease”はネットワークの歪み

第1世代型 疾患原因遺伝子準拠 Diseasome

- **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 (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



疾患 ネットワーク (HDN)

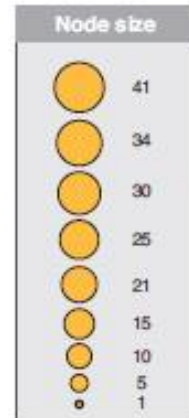
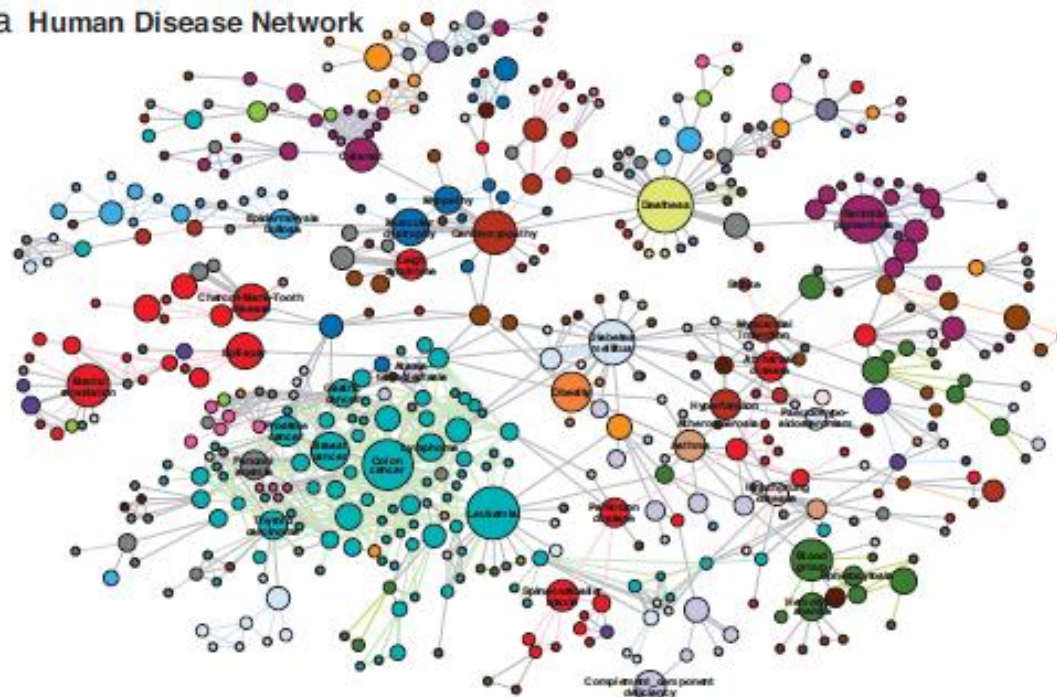
Nodeの直径

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



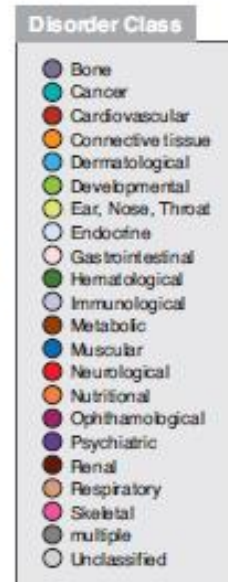
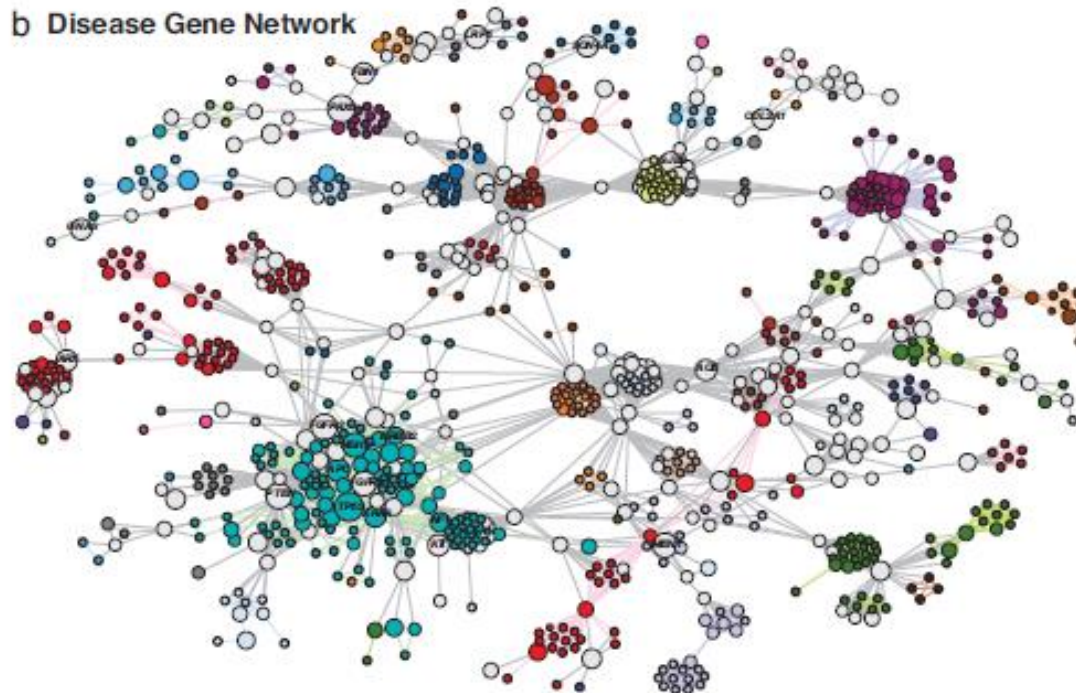
疾患遺伝子 ネットワーク (DGN)

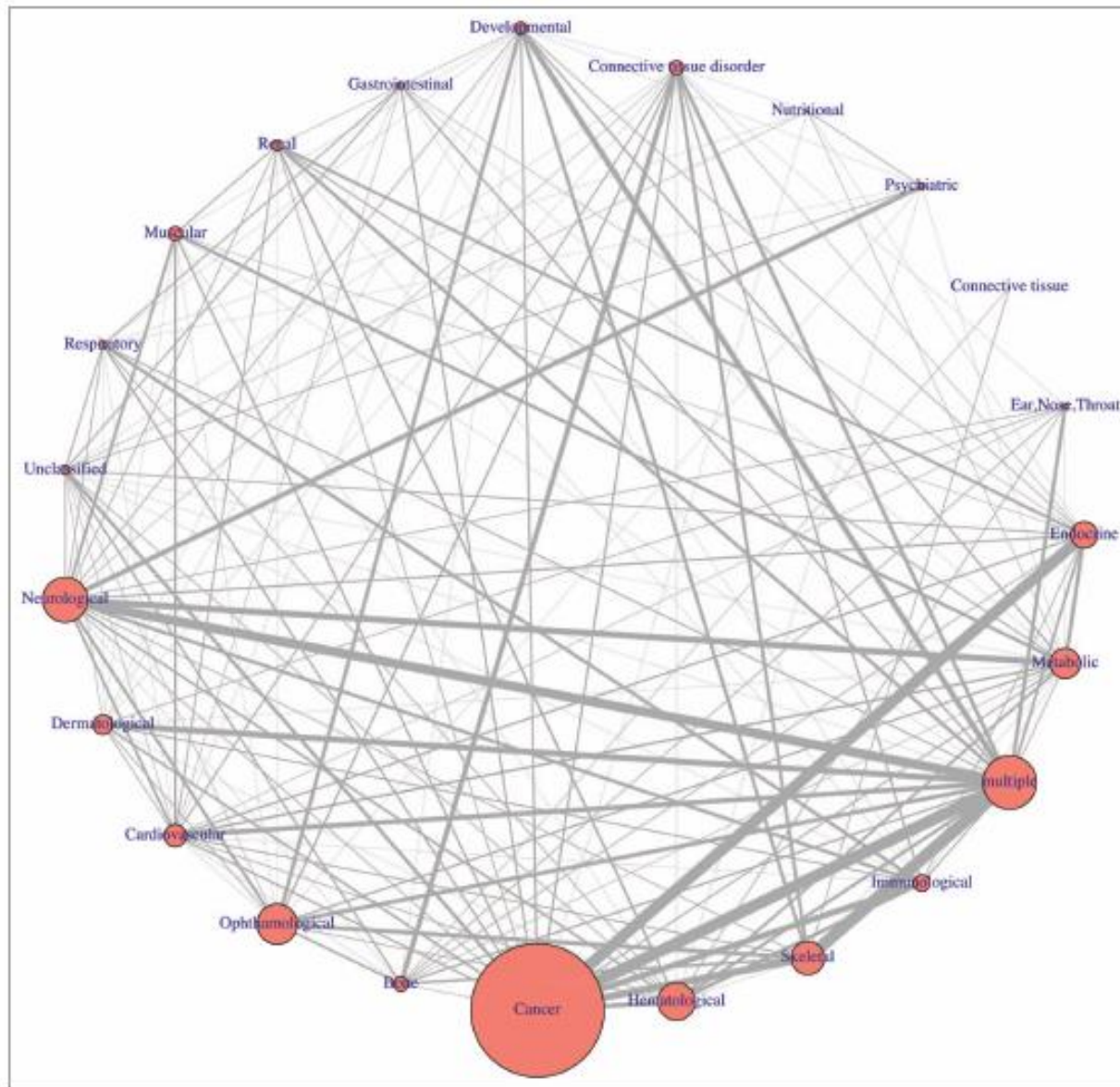
Nodeの直径

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





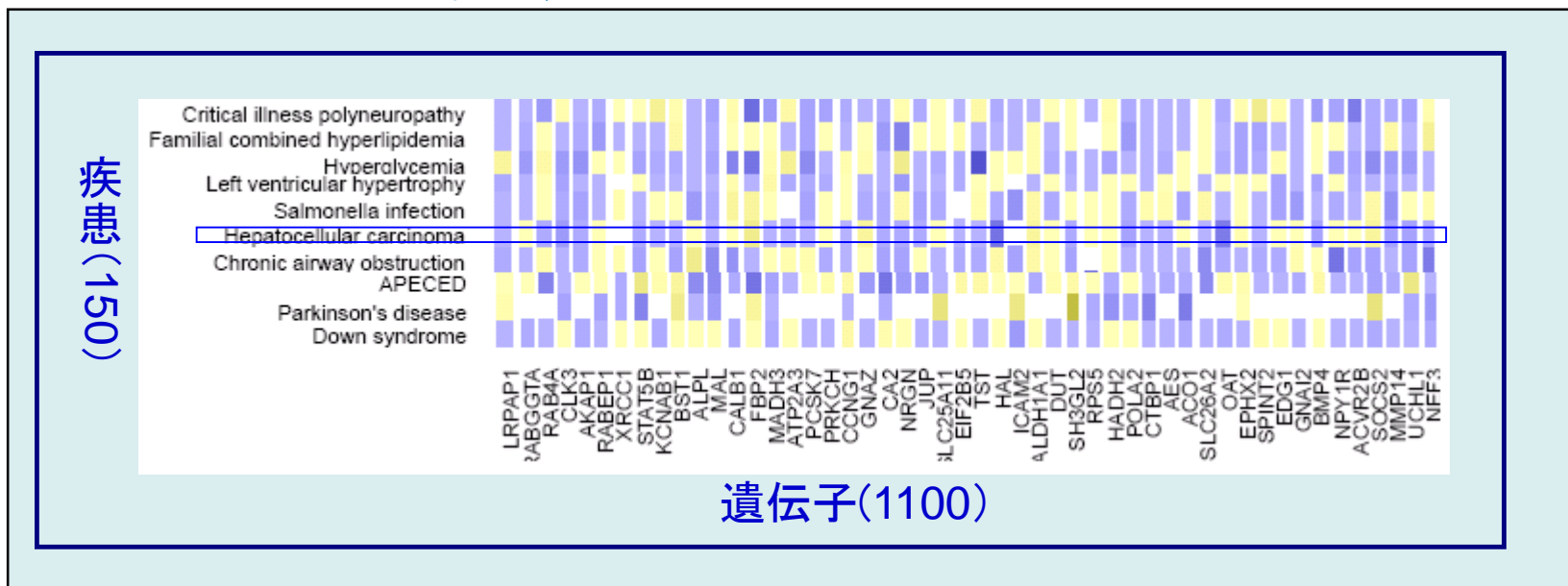
23グループのOMIM疾患のネットワーク
 ノードの径：共通リンクの合計、リンクの太さ：共通遺伝子数

Diseasomeを巡る状況

- Mendel疾患、複雑疾患、環境疾患へと発展
- **他のネットワークと融合**
 - タンパク質相互ネットワーク、代謝ネットワーク
 - PPIの近傍(Vanunu)、代謝網での酵素の基質の共有
 - GWAS (WTCCC, NIH-GAD)のSNPの共有
 - すべてがつながり偽陽性のネットワークで有効性低い
 - miRNA, 環境因子 (annotation MEDLINE)
 - 電子カルテから時系列病歴収集
 - 進化的直系的表現型性 (他の動物も利用)
 - パスウェイ準拠型の疾患ネットワークも
- **表現型疾患ネットワークも存在する**
 - Phenotype : MeSHの頻度をベクトルとする(van Driel)
- **Diseasomeは、臨床表現型NWと分子NWを繋げる機構**
 - 遺伝子を通して疾患間を移動できる
 - Systems pathobiology, nosology, personalized medicine

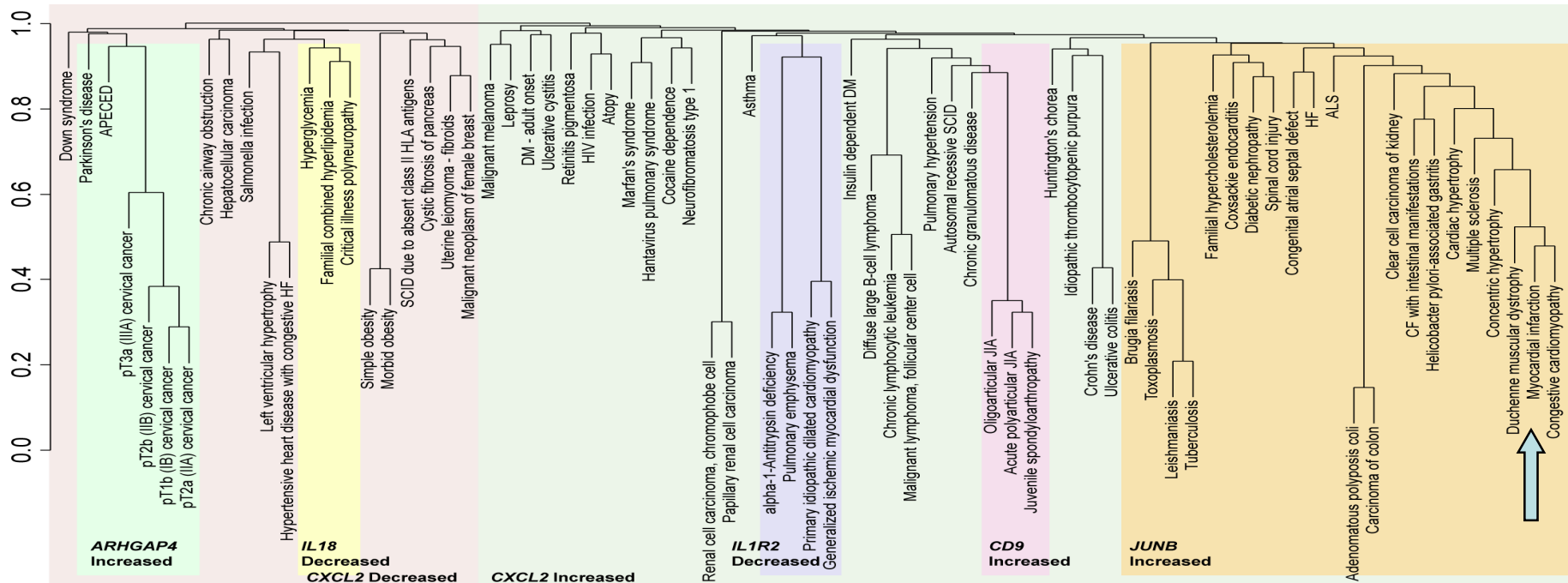
第2世代型 遺伝子発現プロファイ型 GENOMED (A.Butte et al)

- 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



第3世代 疾患分子ネットワーク準拠型 (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
疾患ネットワークを構築

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

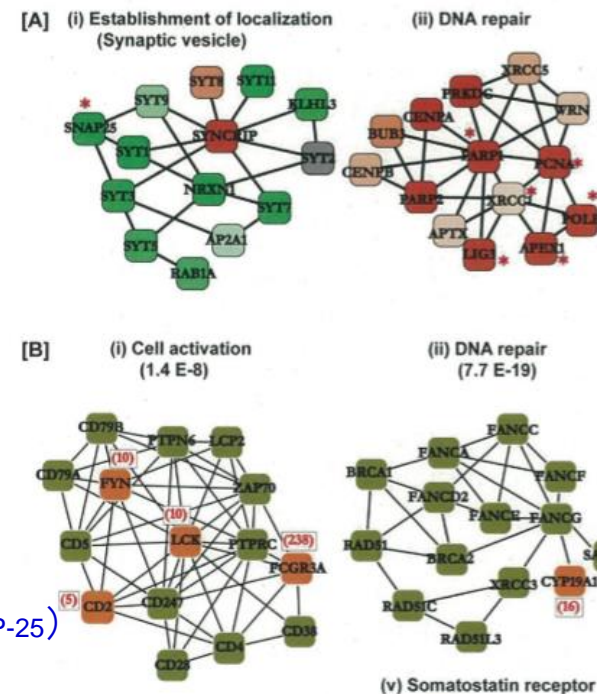
- 138の有意な類似性: ランダム化ネットに対し有意
- $p < 0.01$, FDR=0.1
- 疾患類似性: 肺がん群(修復pasM), 精神疾患(synapsM:SNAP-25)

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」

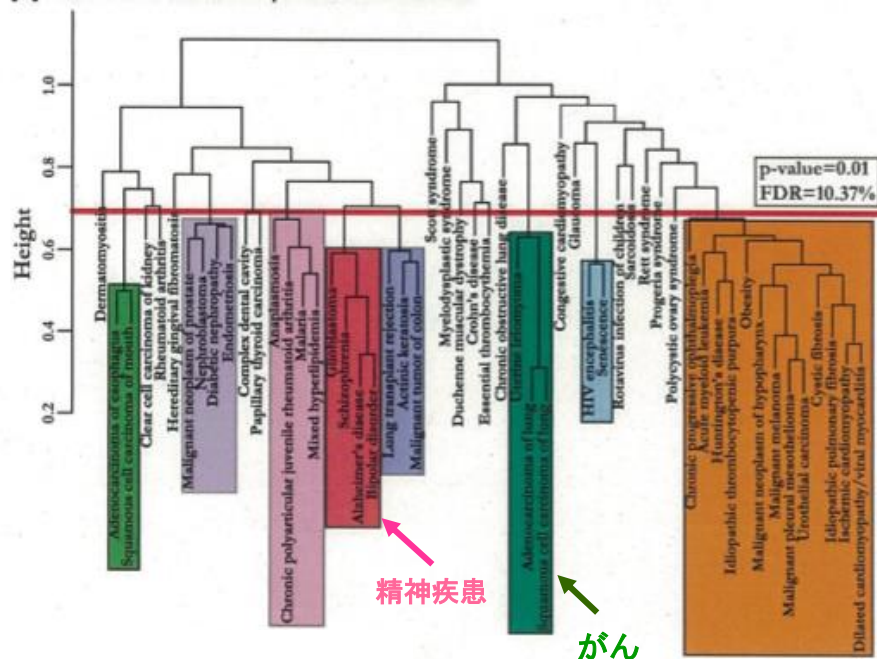
- 「共通疾患状態シグネチャ」薬剤標的分子に富んでいる
- この遺伝子群を標的にする薬剤は有意に多くの他の疾患の薬剤にもなっている



遺伝子発現の変化をPPIに投影した疾患ネットワーク

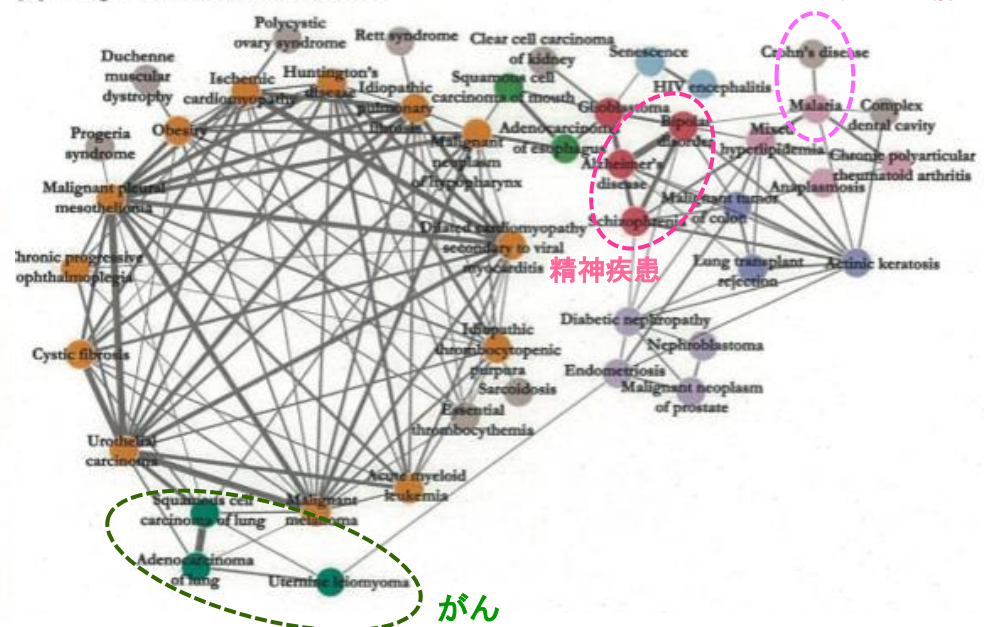
- アルツハイマー症、統合失調症、双極性障害がグループ化
- 子宮筋腫と肺がん、マラリアとクローン病
- 17のがんが1つの群ではない。がんの異質性
- 疾患ネットワーク間の遺伝子共有は高くない（遺伝子外効果）

[A] Hierarchical relationships between diseases



階層的クラスタリング

[B] All significant disease correlations

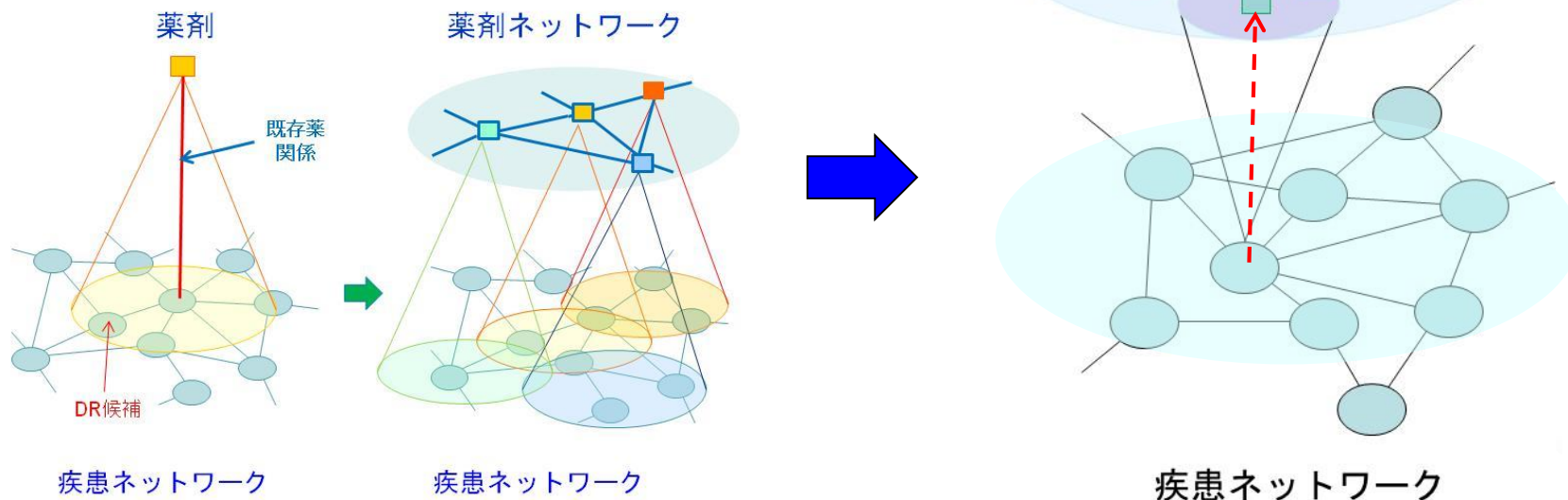


疾患ネットワーク

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(finger 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




AI創薬 第II部

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

疾患-薬剤ネットワーク対応から 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「標的分子」の相互作用の関係が明示的ではない

MMOAによる創薬/DRの3層ネット枠組み

- 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

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

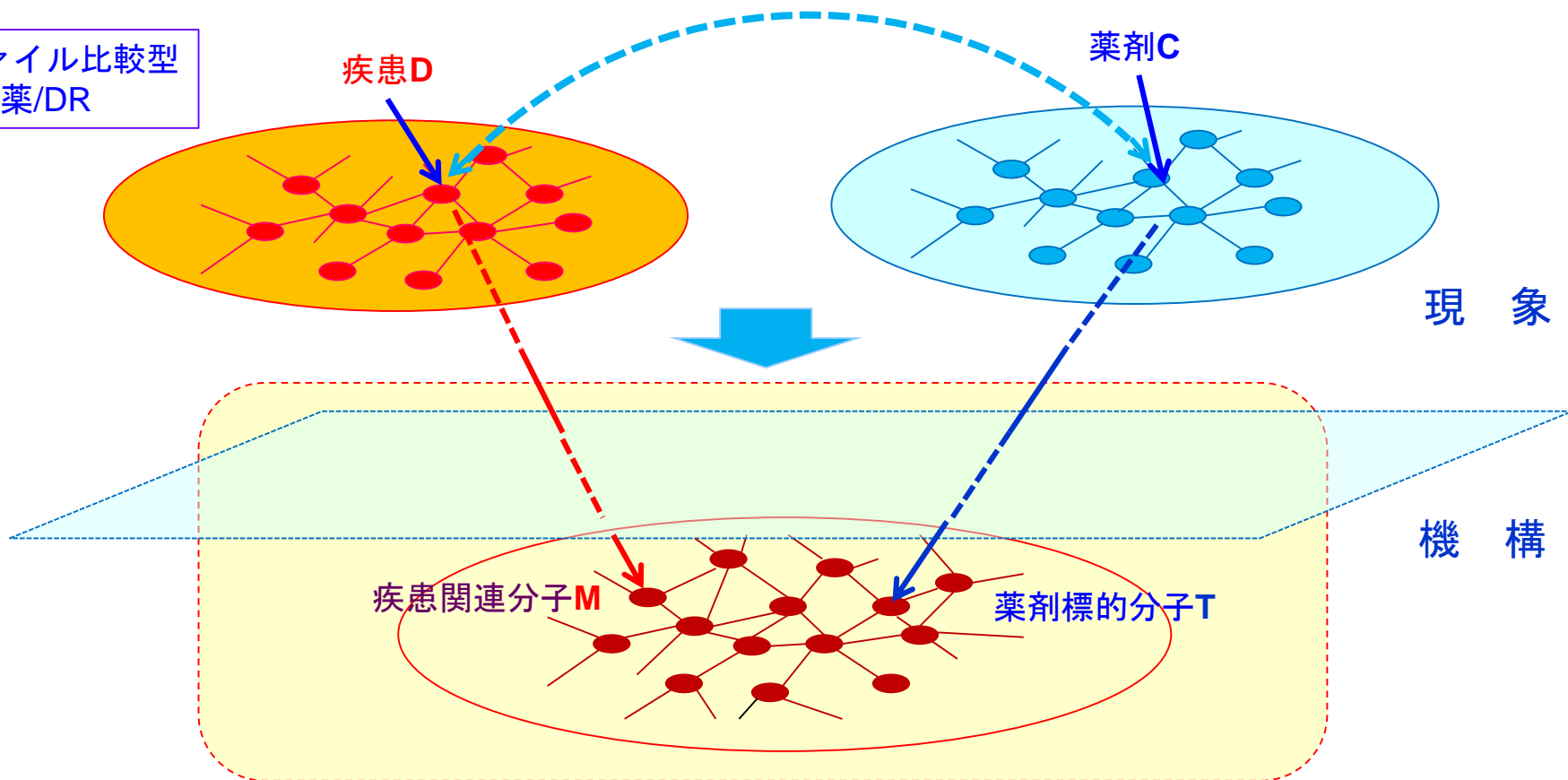
疾患D

薬剤C

疾患関連分子M

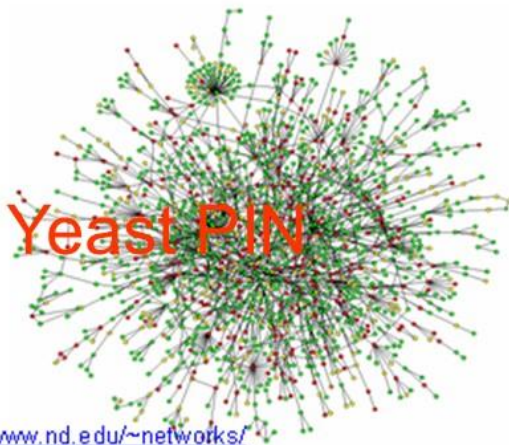
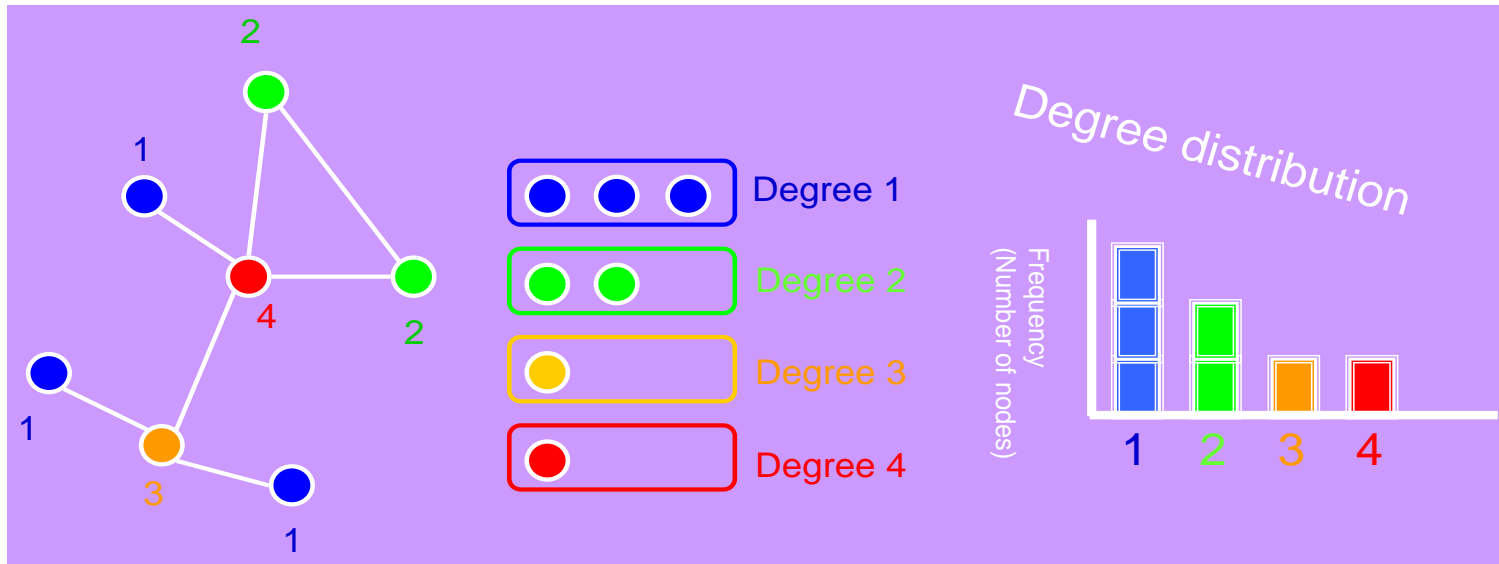
薬剤標的分子T

分子ネットワーク型
創薬/DR

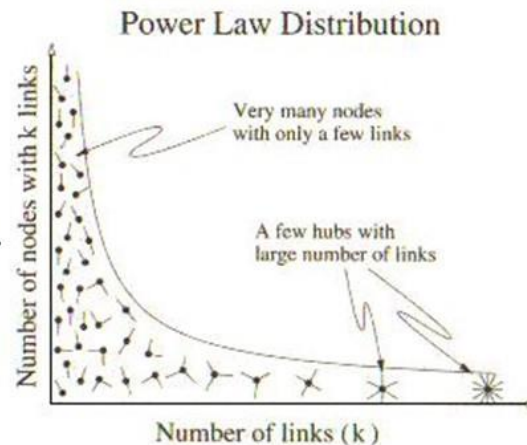


タンパク質相互作用 ネットワークの構造と薬剤標的分子

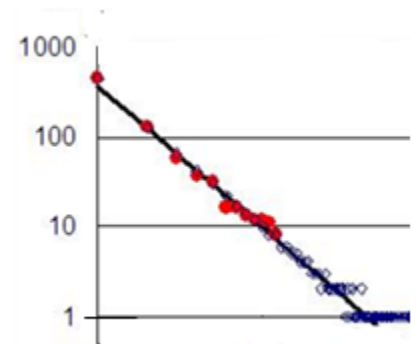
タンパク質相互作用ネットワーク(PIN)では数少ない相互作用が集中したタンパク質(hub)と相互作用が1や2の多数の末端タンパク質(branch)が存在する



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

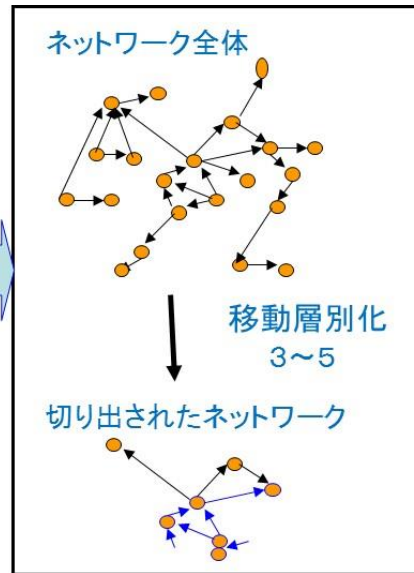
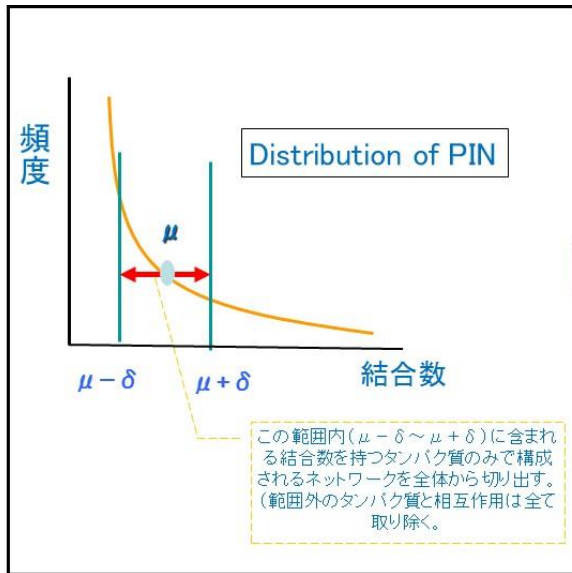


<http://www.macs.hw.ac.uk/~pdw/topology/ScaleFree.html>

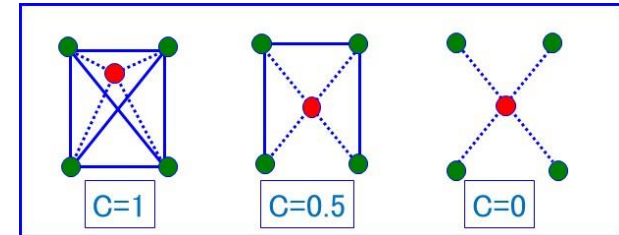


Log-log変換で直線

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

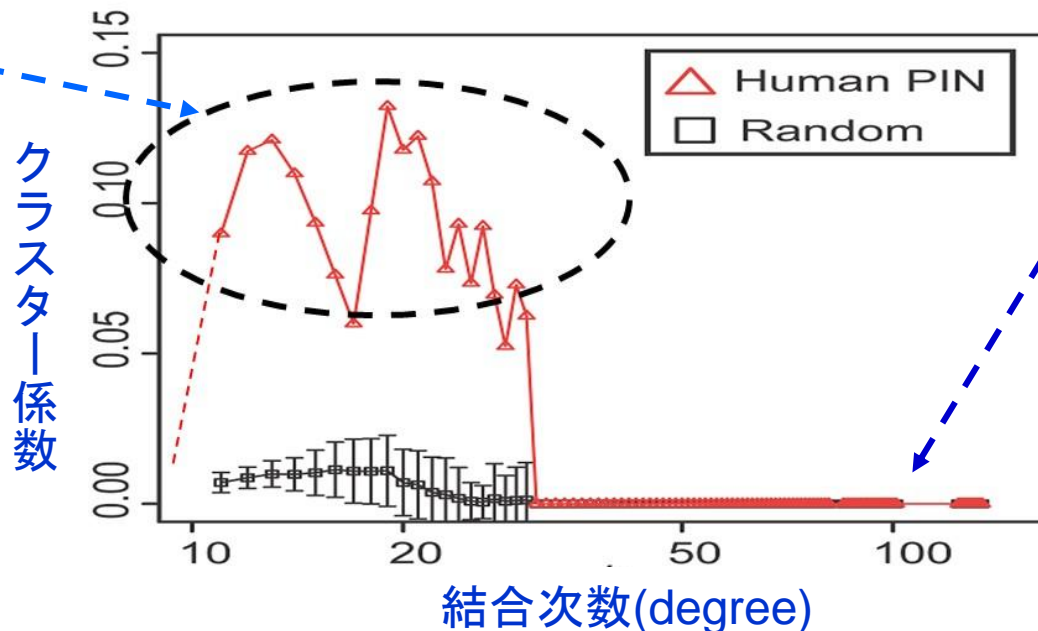


クラスター係数



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

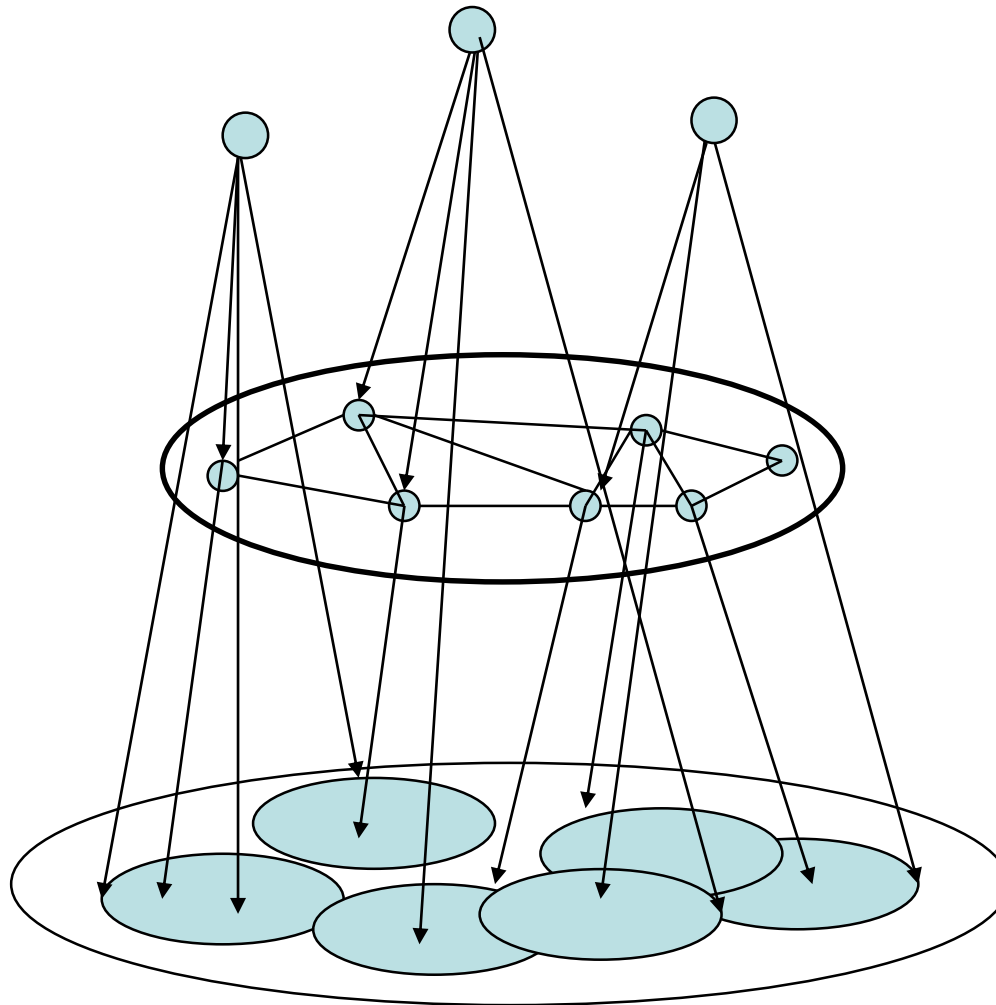
中程度の次数 (7~42) を持つタンパク質は多数の密なモジュールを構成



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

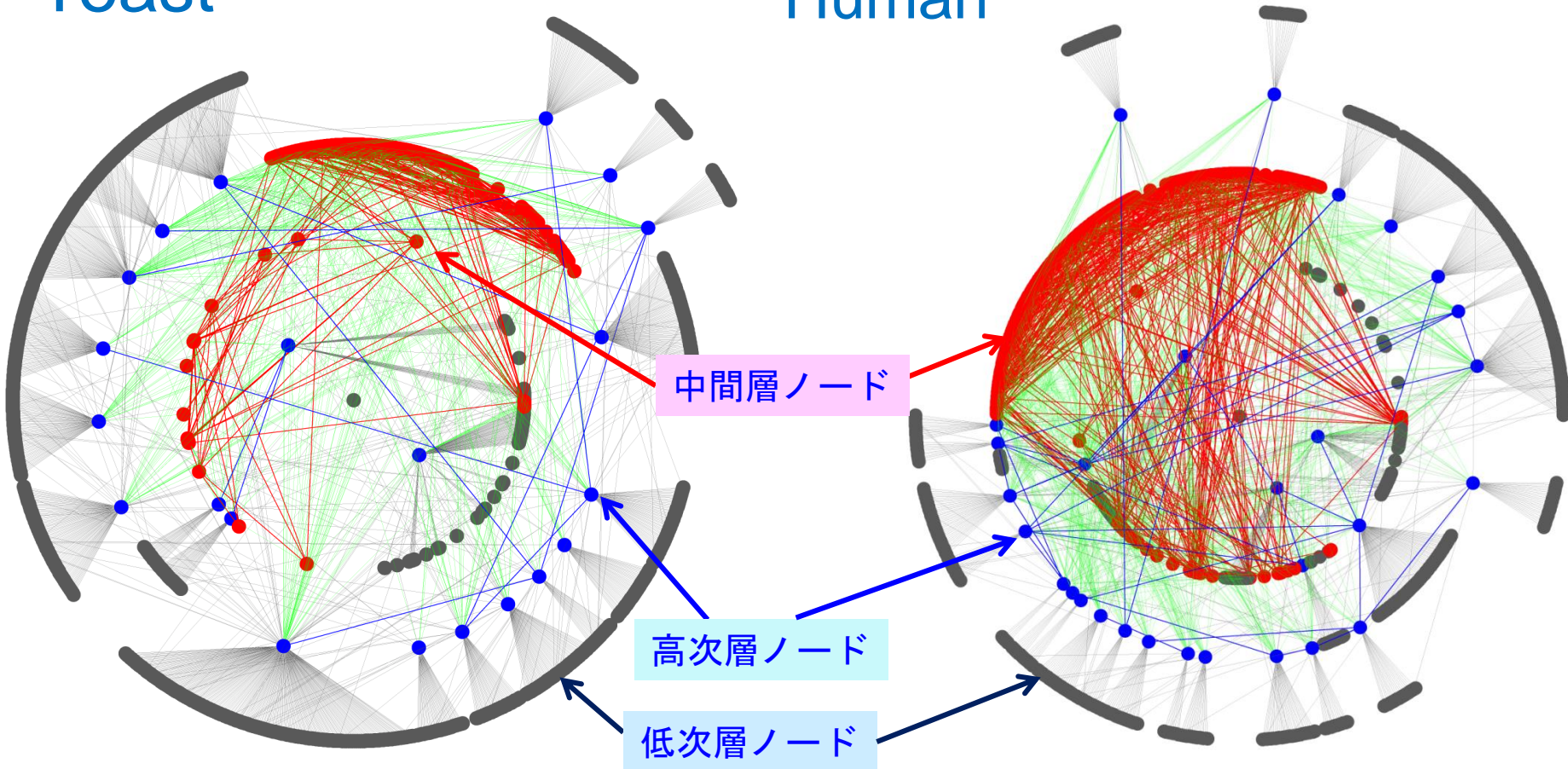
中間層
中程度次数
次数
6 ~ 30 ヒト
6 ~ 38 酵母

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

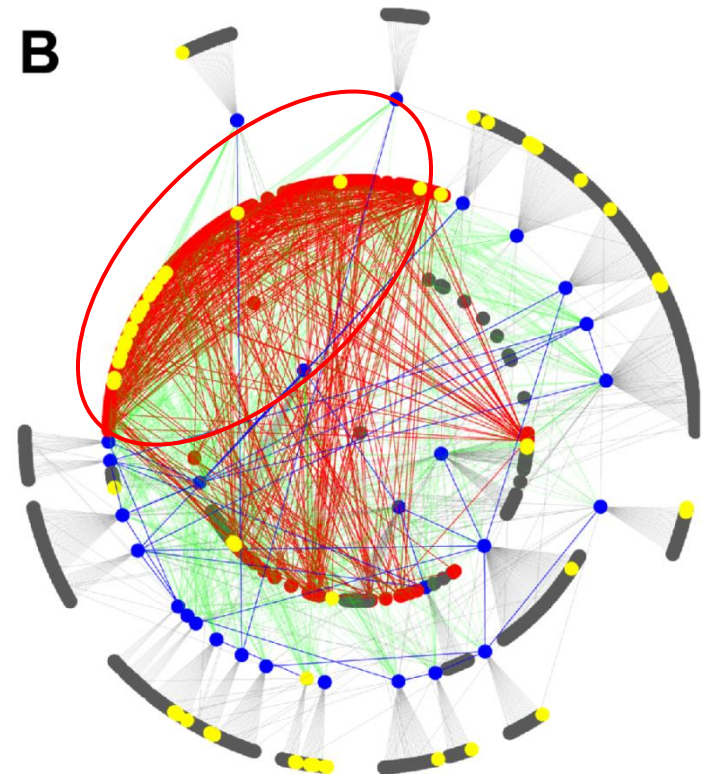
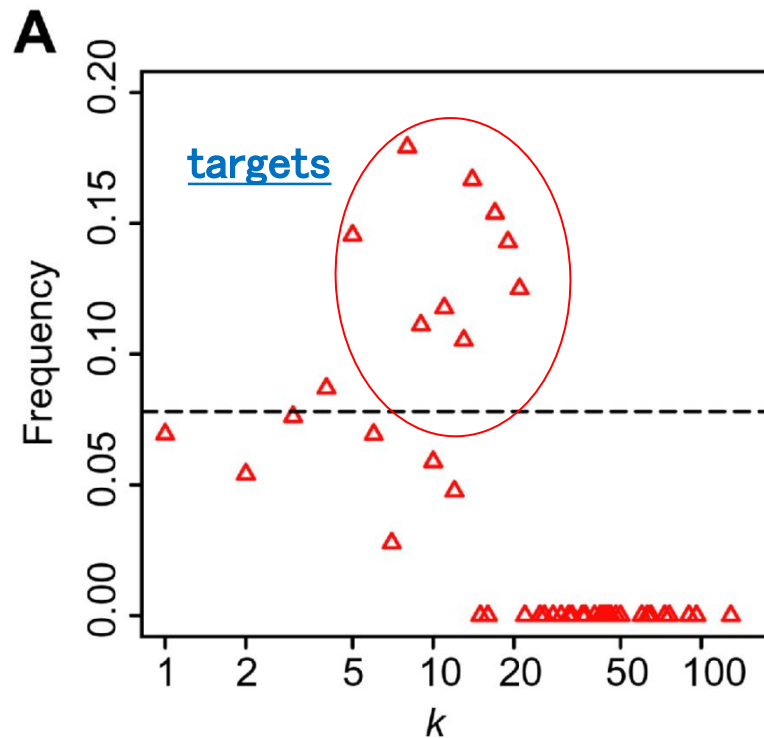
Human



中間層の次数ノードは PPI バックボーンを形成する

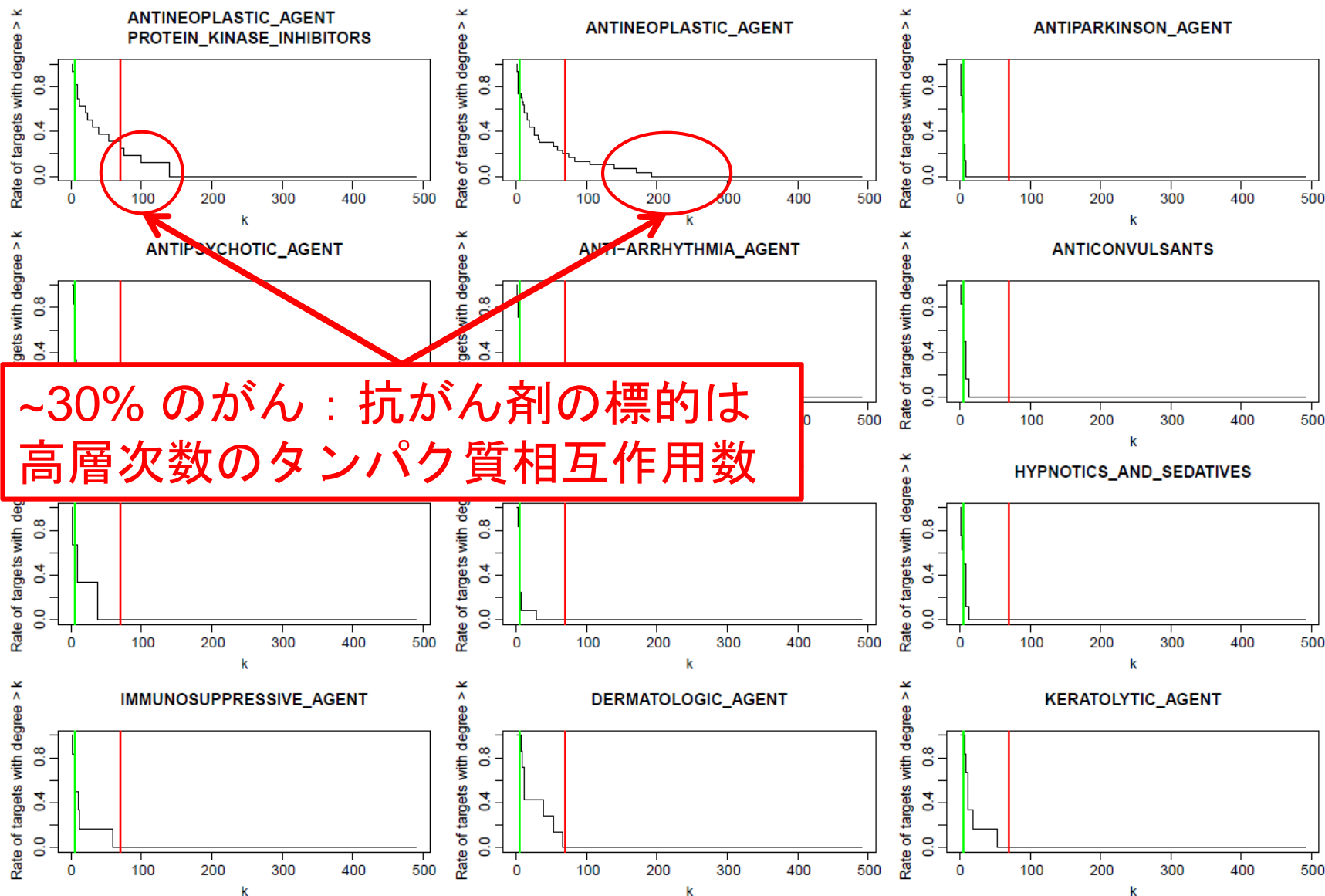
灰色, 赤, 青は、それぞれ低層、中層、高層の次数のノードをそれぞれ表す。

薬剤標的分子と結合度数

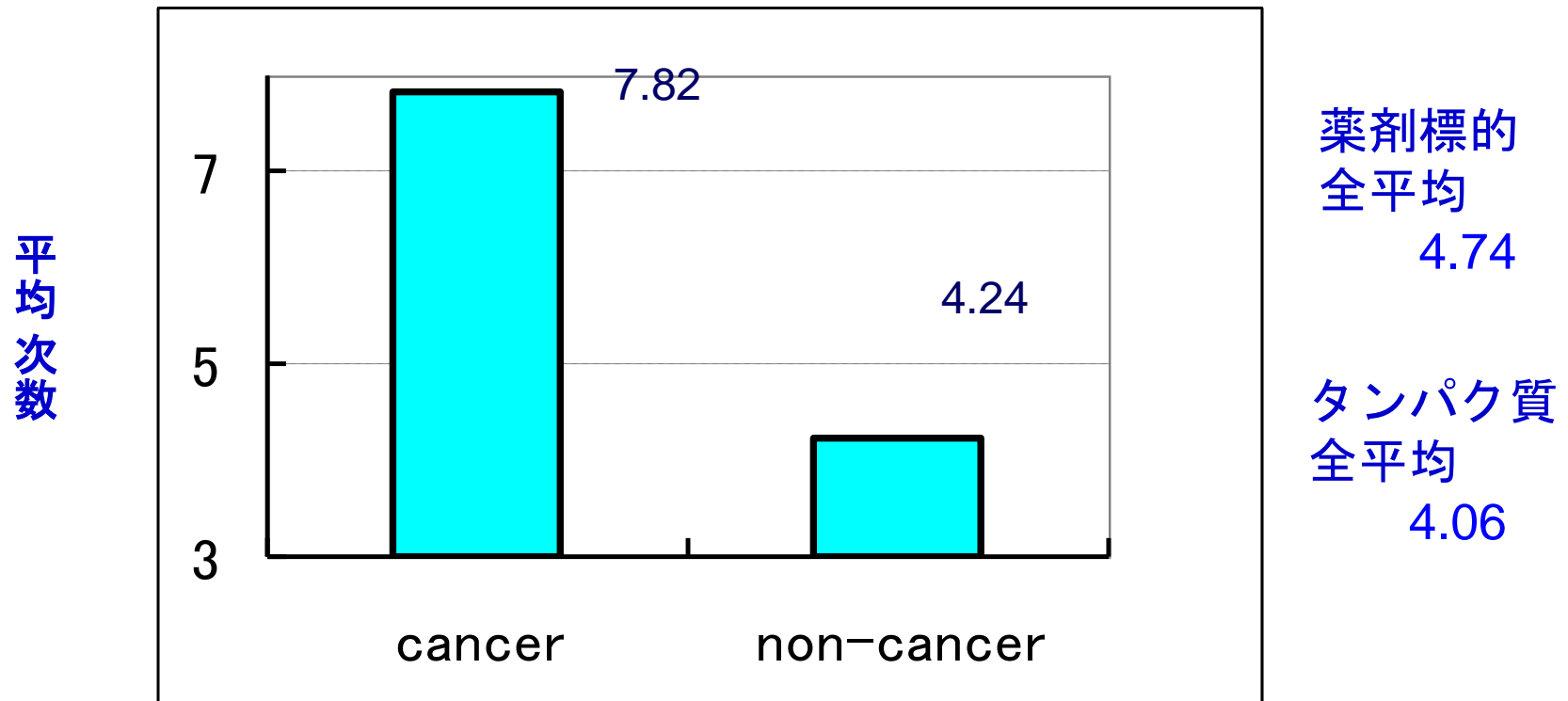


中層レベルのノードは治療薬として最適な標的である。それゆえ、多くの市場にある薬剤標的は、ヒトのバックボーンタンパク質に集中している

がん疾患遺伝子は高層次数ハブのタンパク質が多い



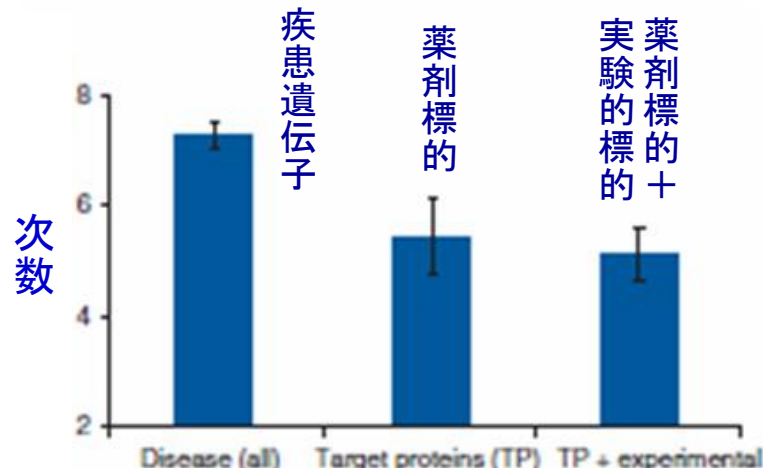
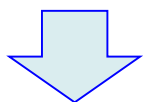
薬剤標的分子と結合次数



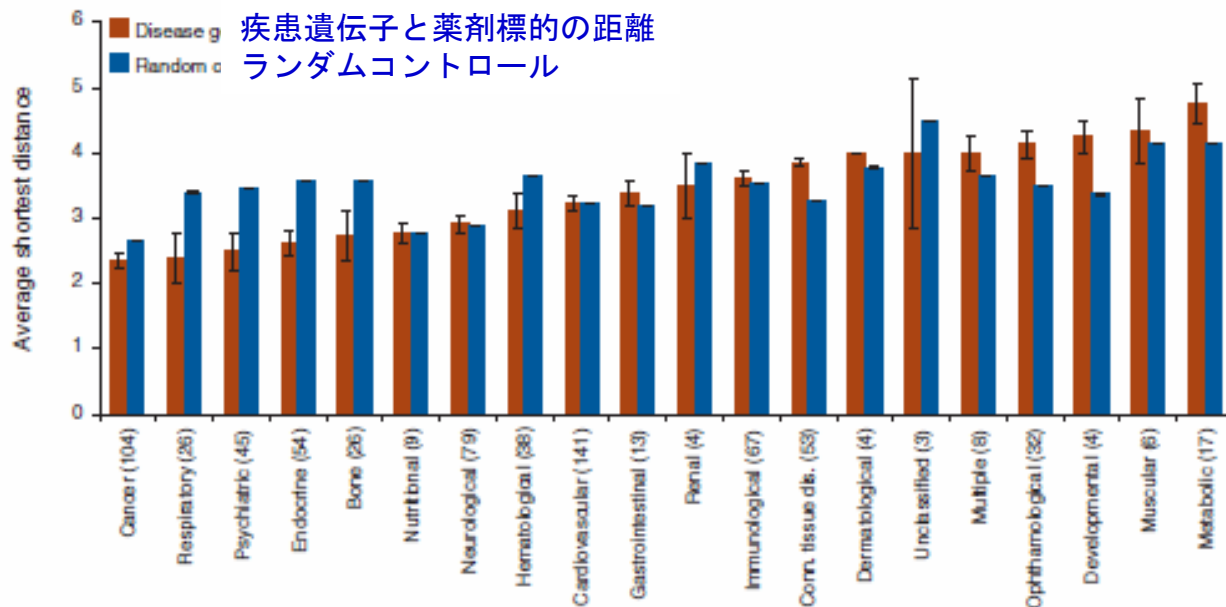
抗がん剤 ($P=0.01$)の標的分子は平均的に次数が高い。
抗がん剤がより厳しい副作用を起こす理由である。

標的タンパク質と疾患遺伝子の距離

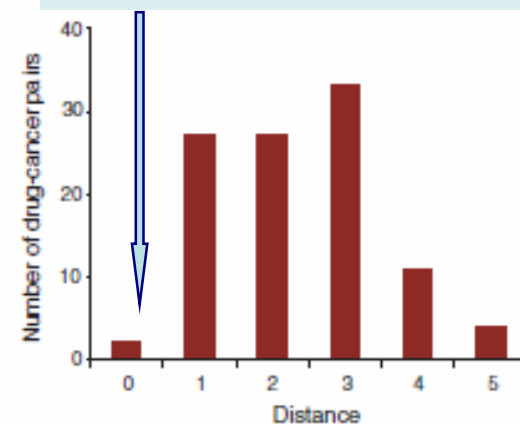
薬剤標的タンパク質と疾患関連タンパク質の間の距離：2~4リンク



Yildirim M A, et al, NATURE Biotechnology 2009

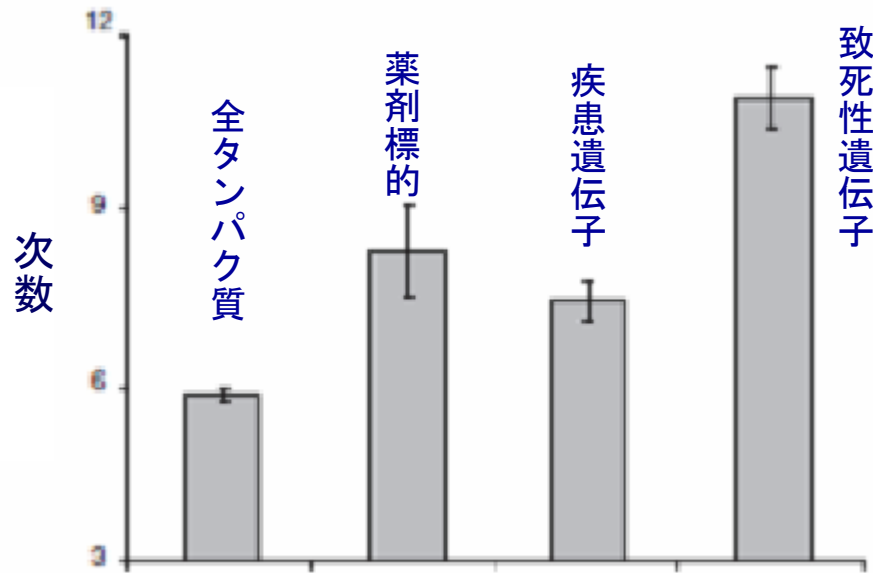


抗がん剤の場合
疾患遺伝子と距離0の標的



抗がん剤の標的分子と疾患遺伝子の間に距離

薬剂標的分子と結合次数

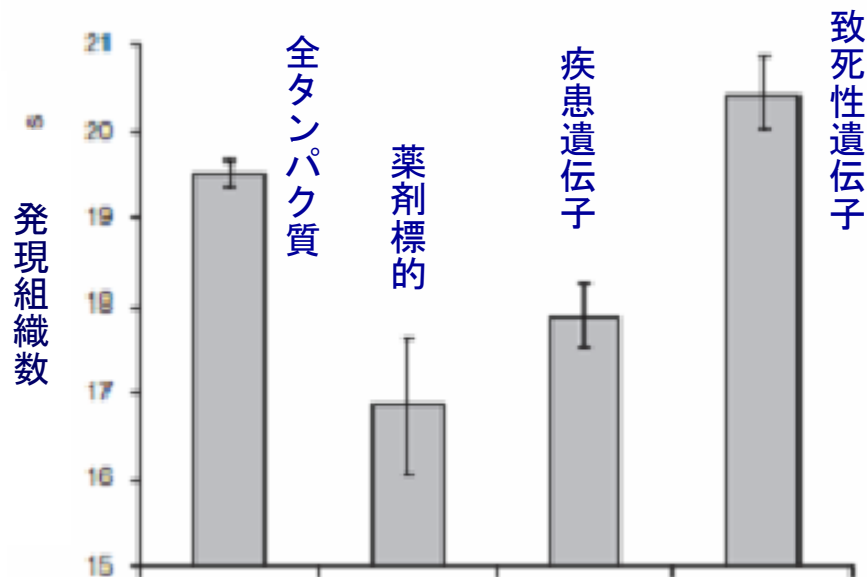


結合次数

薬剂標的タンパク質

致死のタンパク質

疾患関連遺伝子タンパク質



発現組織数

薬剂標的タンパク質

致死のタンパク質

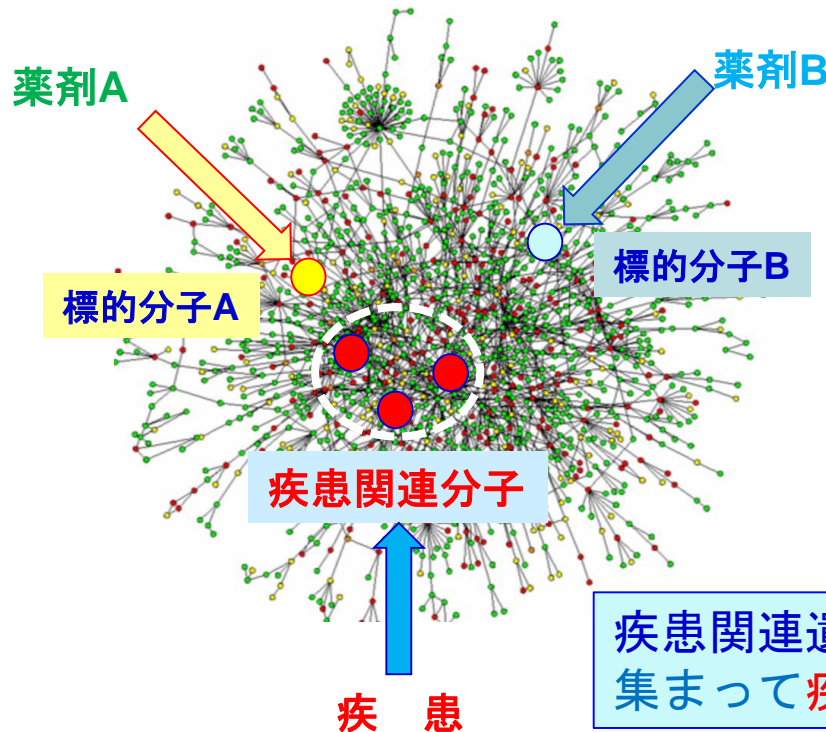
疾患関連遺伝子タンパク質

(Yıldırım M A, et al, NATURE Biotechnology 2007)

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

PPIの基づくDR（肺腺癌の例）

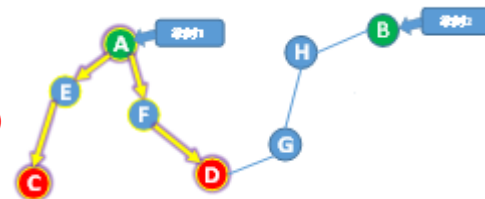
- **Interactome**(タンパク質相互作用)ネットワーク (Sun, 2016)

- **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

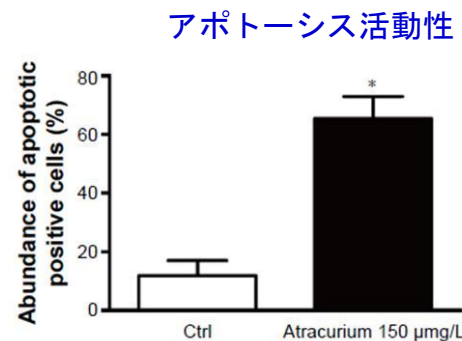
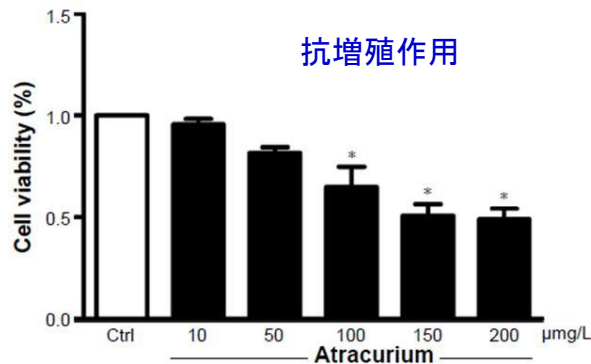
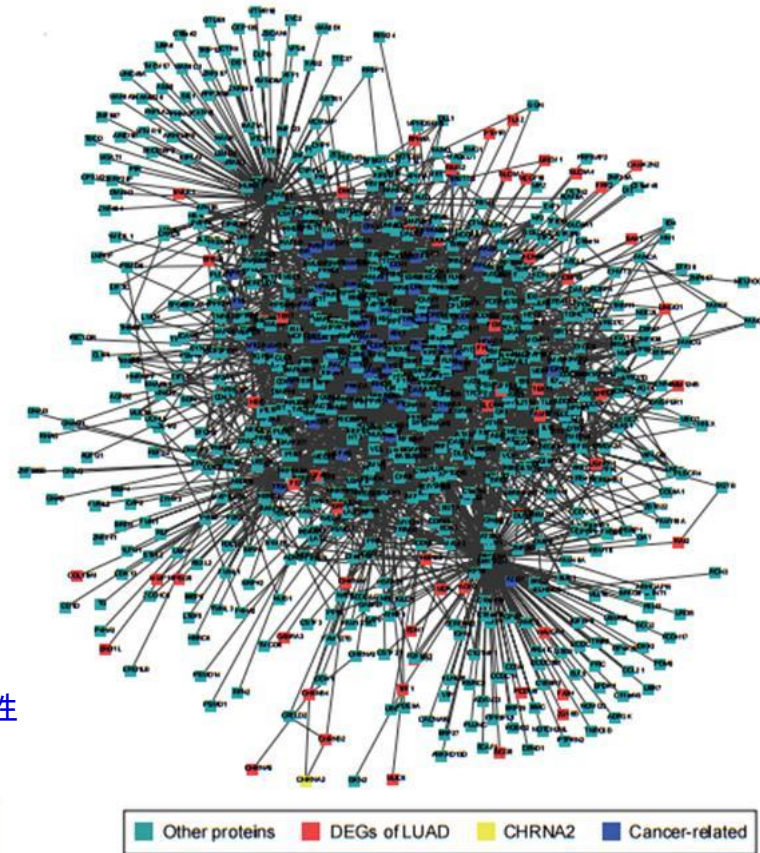
$$\mathbf{P}^{s+1} = (1-\gamma)\mathbf{M}\mathbf{P}^s + \gamma\mathbf{P}^0$$

\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

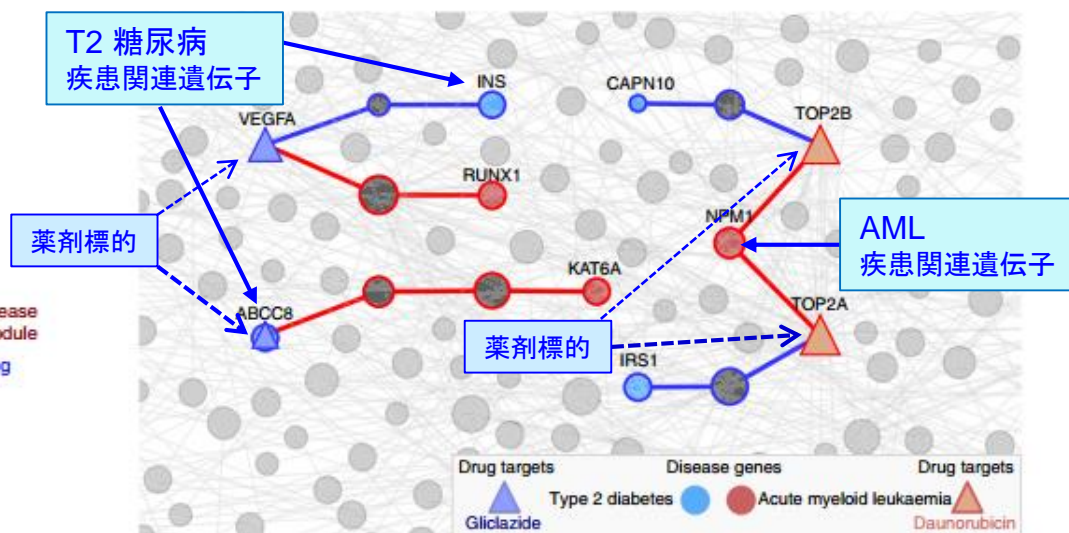
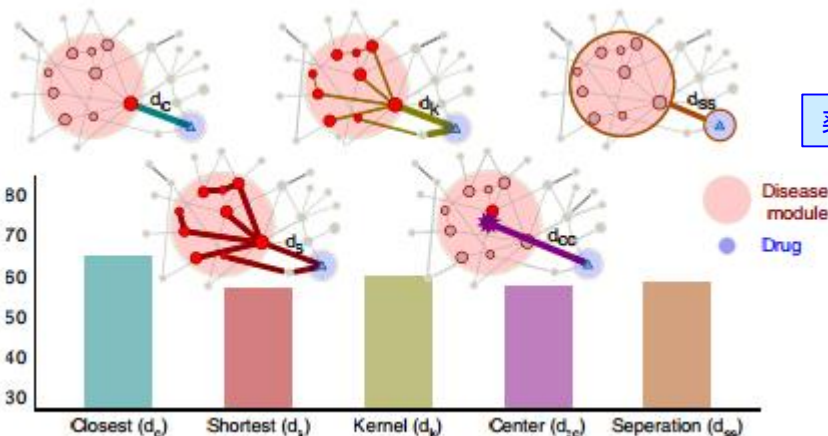
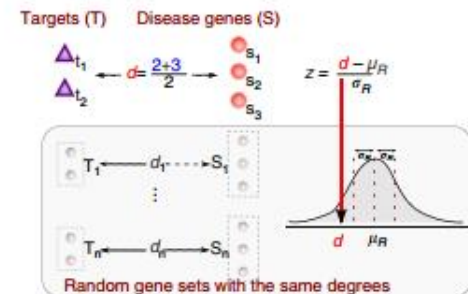
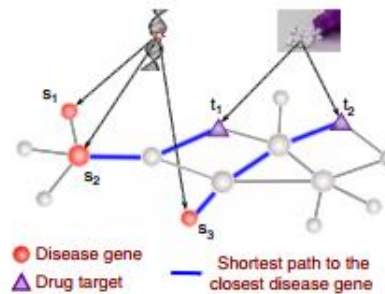
- **HPRDとSTRINGSの両方のランダム歩行で145薬剤・化合物が共通**
- **最高スコアを挙げたAtractiumを選択**
- **標的はCHRNA2(Cholinergic Receptor Nicotinic Alpha 2) でアポトーシス経路である**
- **培養細胞A549 (ヒト肺胞基底上皮腺癌細胞) の抗増殖作用を確認**



タンパク質相互作用ネットワークでの 近接性によるDR

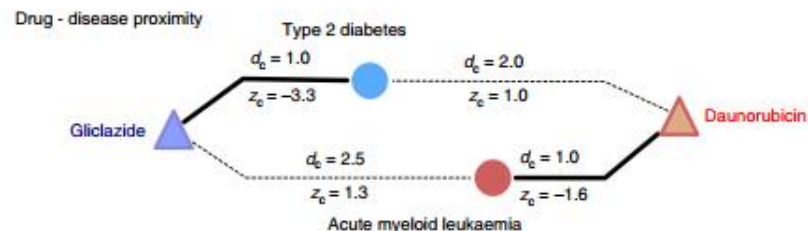
相対近接指標 d_c :

- 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⇒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



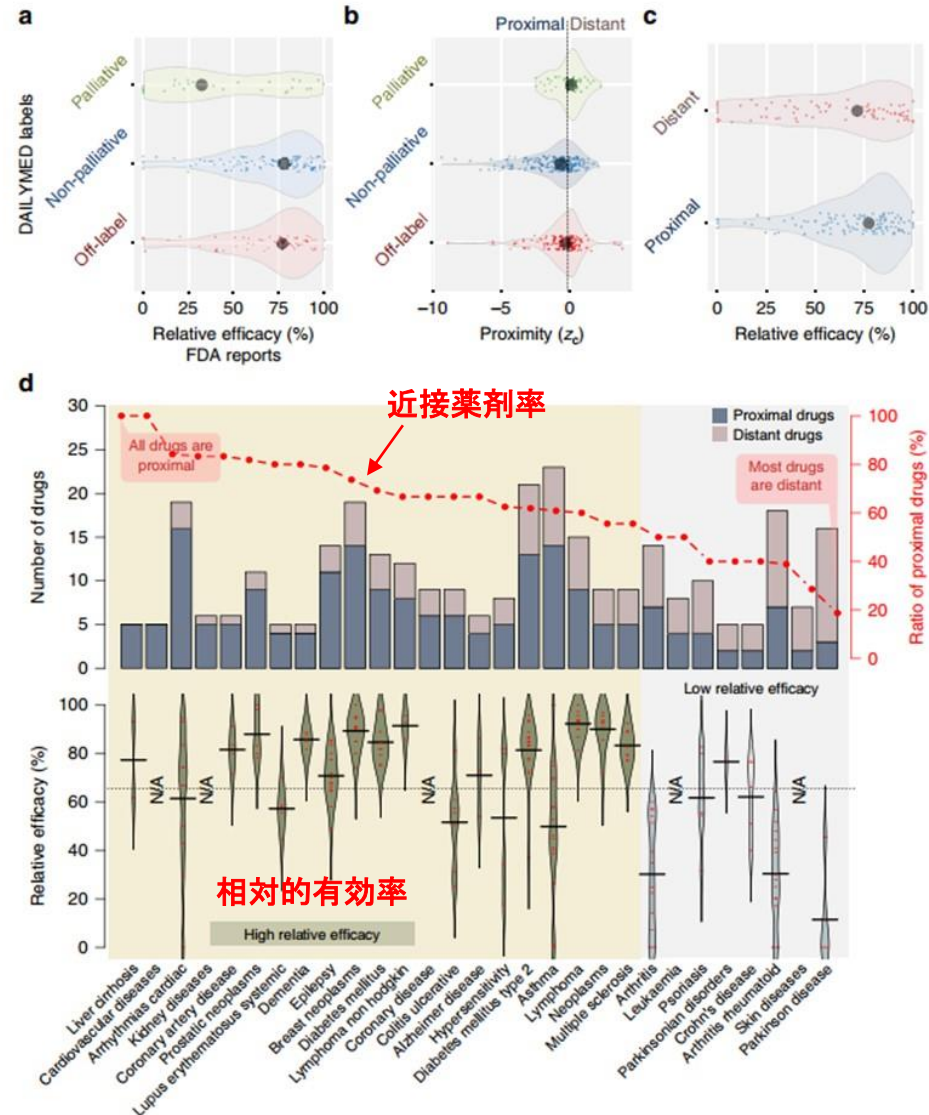
大半の薬剤は標的と疾患関連分子
2リンク離れている

(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



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

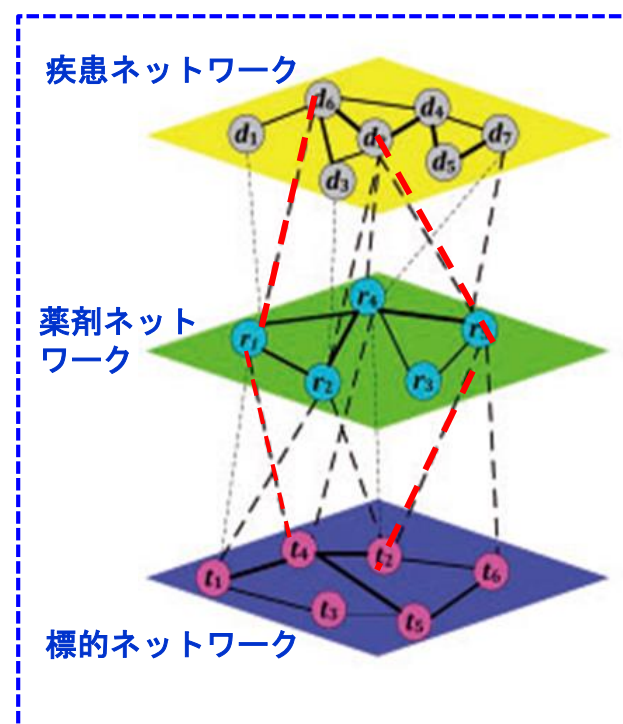
創薬/DRとは

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

(Wang et al. 2014)



異質 ネットワーク 創薬/DR

(Wang et al. 2014)

3層ネットワーク構成

- 疾患ネットワーク (d_i)
- 薬剤ネットワーク (r_i)
- 標的ネットワーク (t_i)

各ネットワークで距離定義

- 疾患ネット：MeSHの共通項数
- 薬剤ネット：Tanimotoスコア
- 標的ネット：Smith-Waterman法

結合係数 $w(i,j)$ 更新法

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

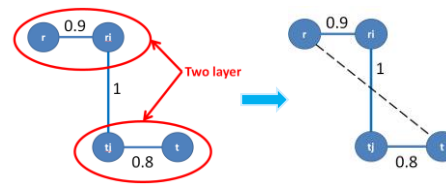
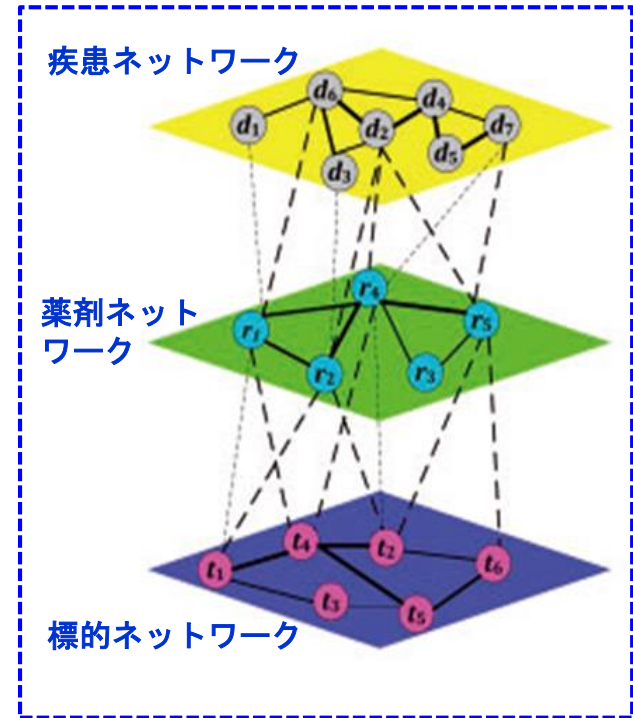
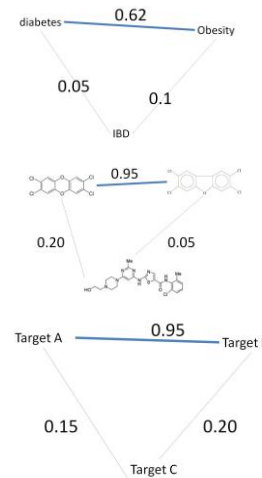
$$w(d, t) = \sum_{r_i \in R} \sum_{r_j \in R} w(d, r_i) \times w(r_i, r_j) \times w(r_j, t)$$

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

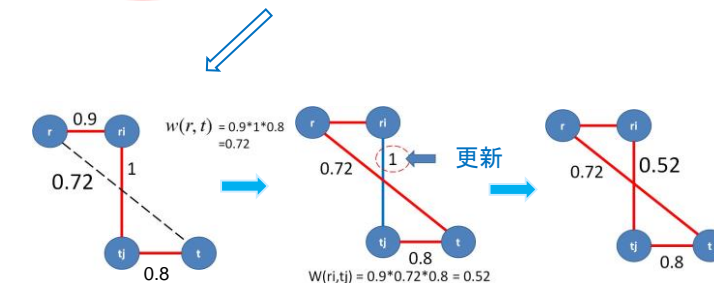
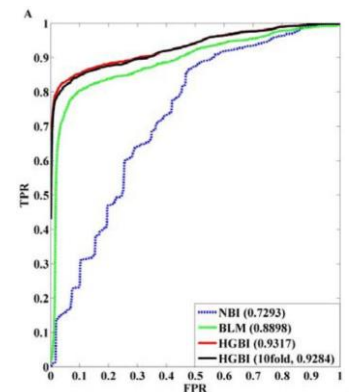
結合係数更新のマトリックス表示

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

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



従来の方法より
DR推定精度高い
ROC曲線



AI創薬・DR (学習型アプローチ)

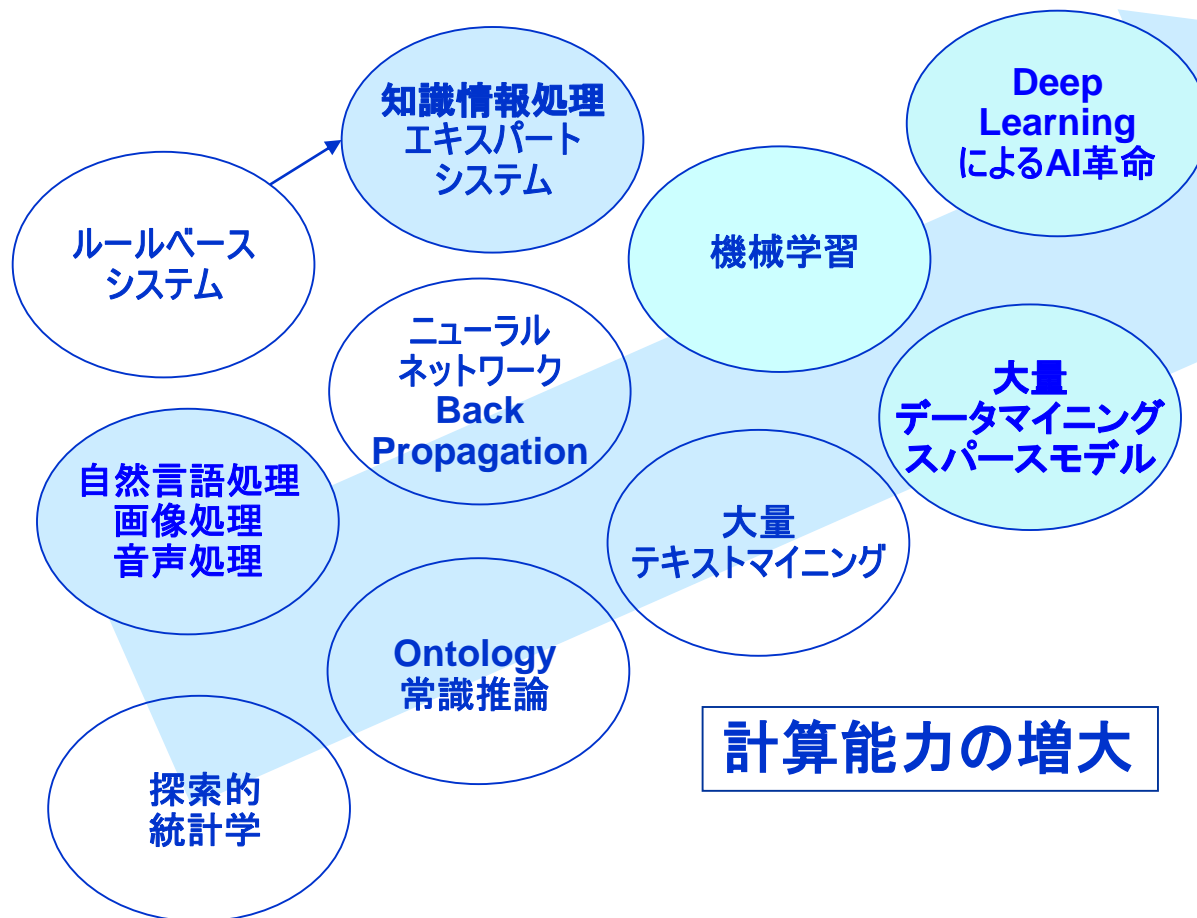
人工知能とDeep Learningの革命性

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
 - **評価経験則が人間が与えていない。強化学習を用いて自分自身と多数の対戦 (3000万回) を行う**
- 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



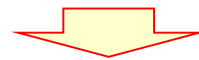
「ビッグデータ」のData 原理

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**)

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は高い
- 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–臨床表現型情報>
 - ゲノムコホート<遺伝素因–環境要因(生活習慣)>
- Deep Learning-based Network Contraction
「DLネットワーク縮約法」

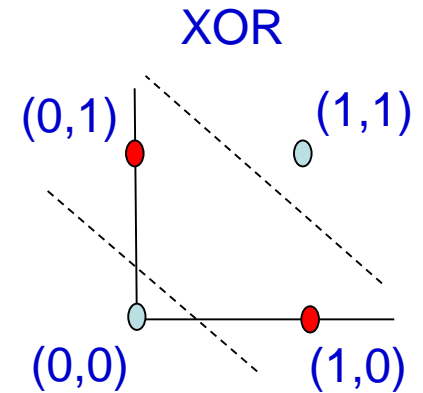
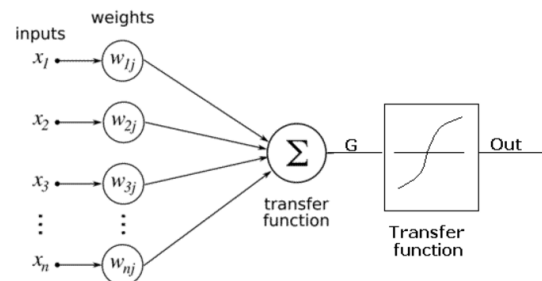
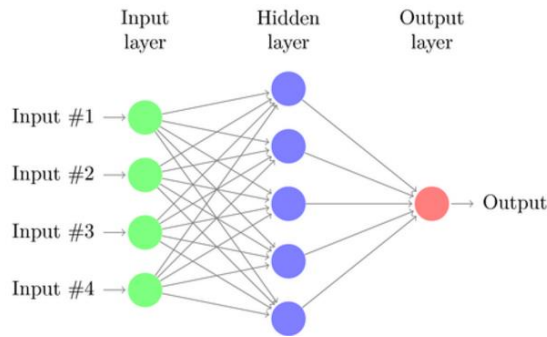
超多次元ネットワーク情報構造⇒

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
 - 線形分解（特異値分解：SVD）との比較

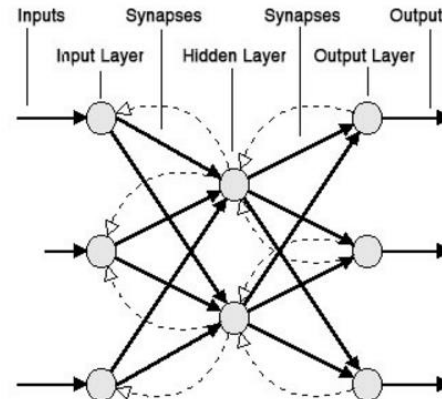
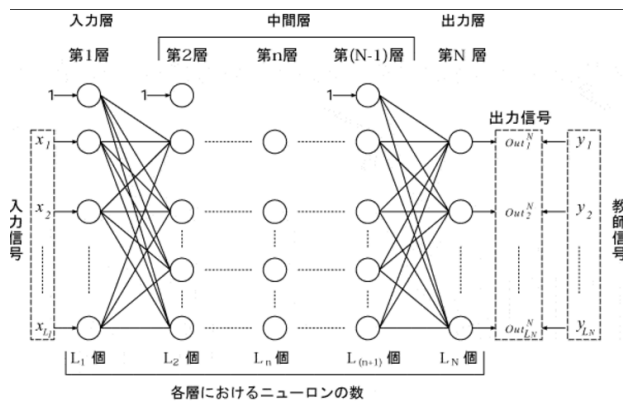
従来のニューロネットワーク

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

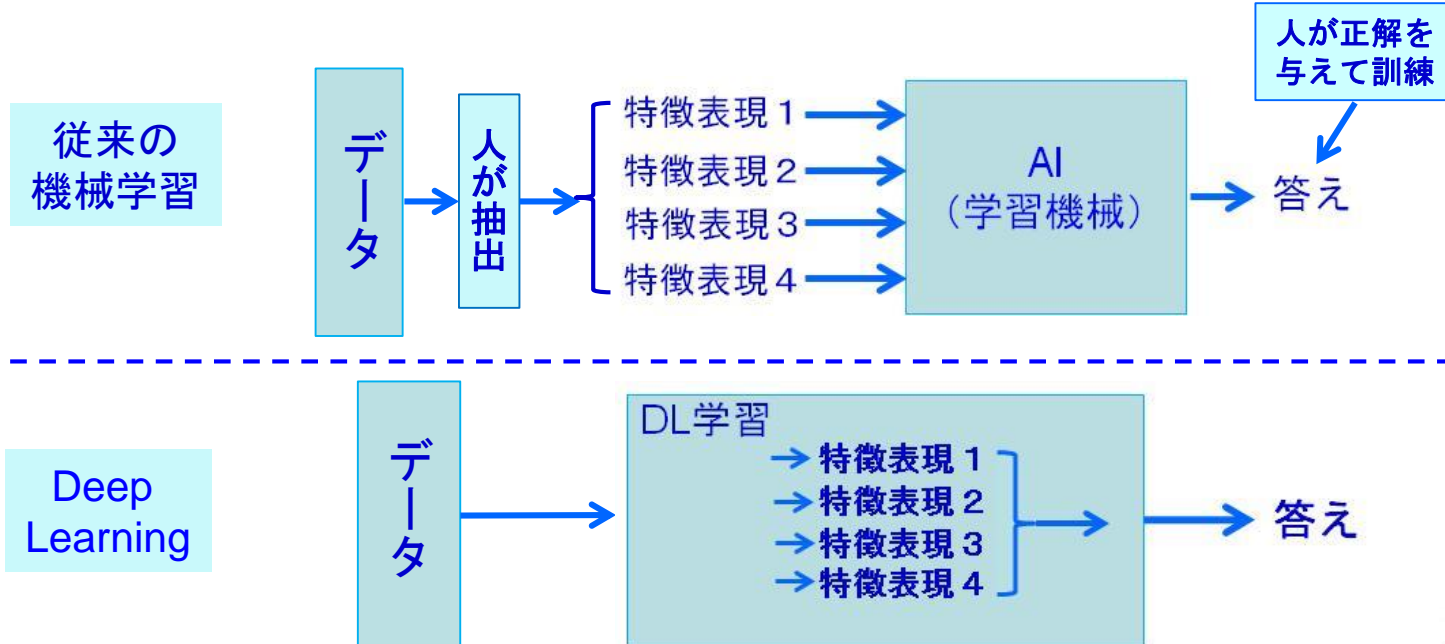


Back Propagation (1986 Rumelhart)
 望ましい出力との誤差を教師信号として与える事により、次第に結合係数を変化させ、最終的に正しい出力が得られるようにする。結合係数を変える事を学習と呼ぶ。この学習方法には、最急降下法(勾配法)が使われる。出力層へ寄与の高いノードの重みの変更。

多層にわたる逆伝搬で修正感度減衰

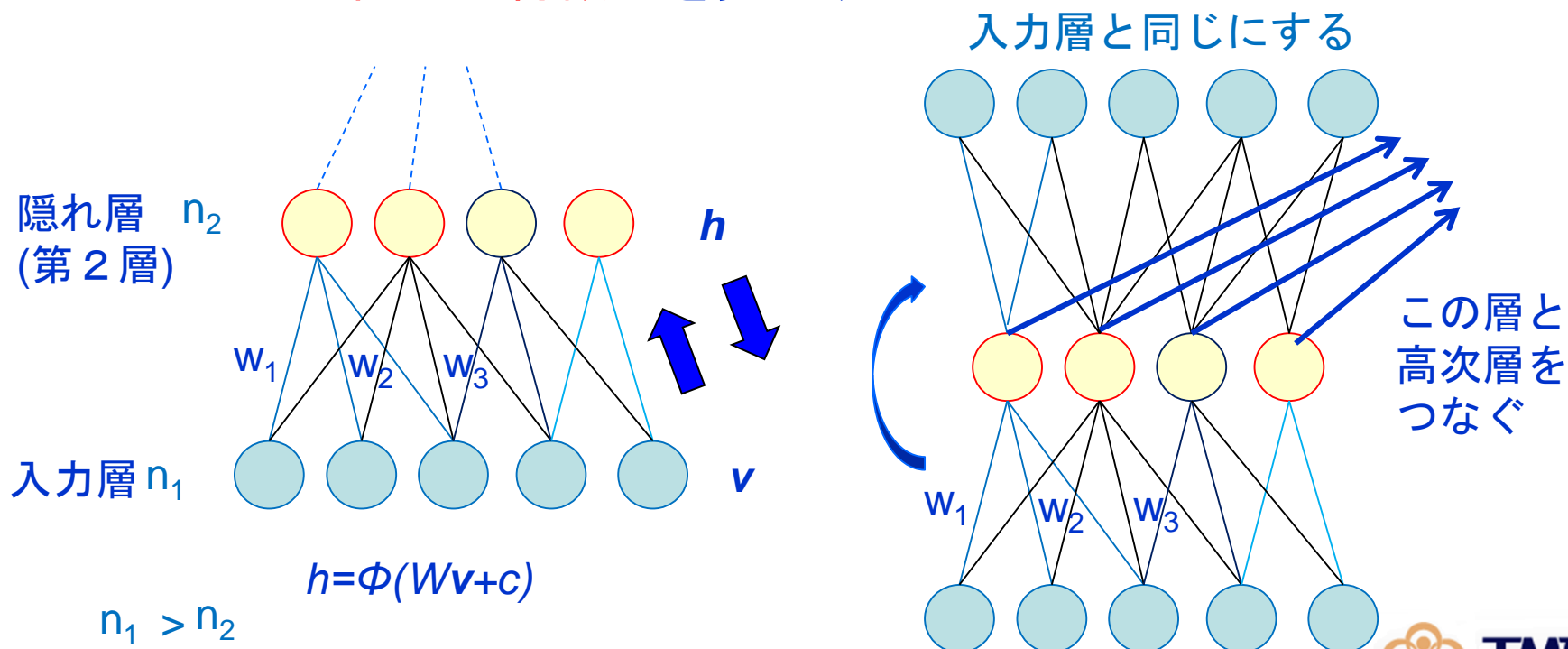
Deep Learning による 人工知能革命

- 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



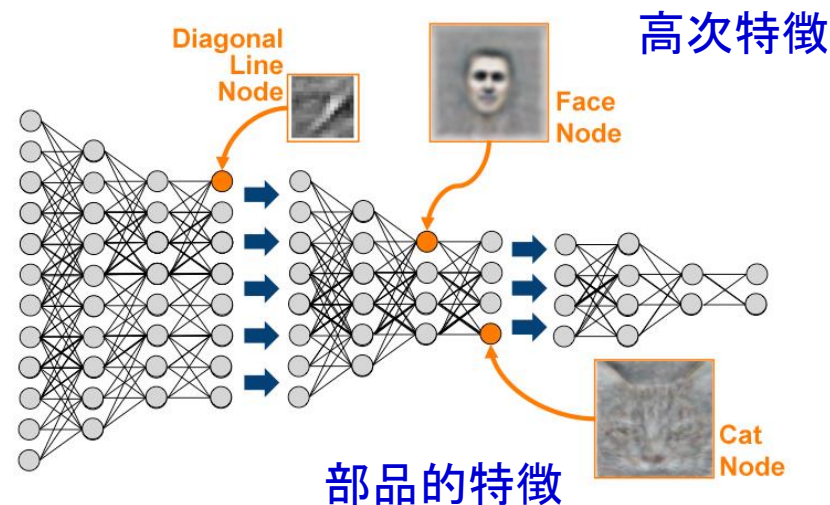
DLの革命点 Autoencode 1

- 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 Autoencode 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



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).

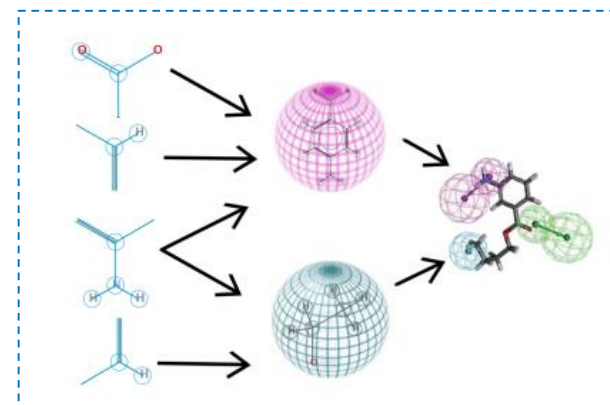
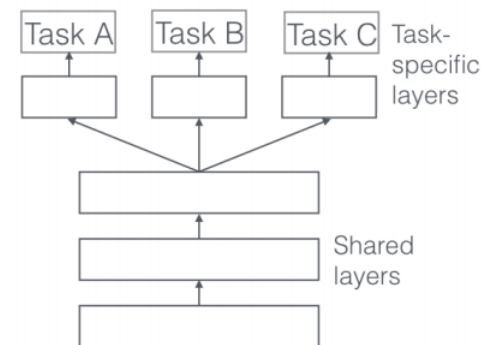
- 15種類の標的分子に対する化合物データセットから異なった**構造活性相関のデータ**を学習して**構造から分子の生物学的活性**を予測するモデルの開発コンテスト
勝利したモデルはdeep learning を用いた

- **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」の現在の研究

- **標的分子探索に人工知能を用いた方法**
 - 疾患に有効性のある標的分子の探索
 - Hase-Tanakaの多層Deep AutoEncoderを用いた標的分子探索法
- **Virtual Screening (QSAR)への人工知能・機械学習の応用**
 - 標的分子に相互作用（ヒット）する化合物の探索
 - Ligand-based AIバーチャルスクリーニング
 - Kaggle, Unterthinerなど
 - Structure-based AIバーチャルスクリーニング
 - Wallachなど
- **その他**
 - 人工知能を用いた**化合物自動設計**
 - Gomez-Bombarelliなど
 - **合成経路自動探索**
 - **AI毒性学**

Artificial Intelligenceと創薬

- 標的分子選択と妥当性検証
 - 疾患に対する適切な分子標的の選択
- Virtual screening とヒット化合物の判定
 - 適切な化合物に対する相互作用の有無の判定
 - 研究例：ChEMBLに対するdeep learning
 - 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
 - Pande研 Massive Multitask Network
 - 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

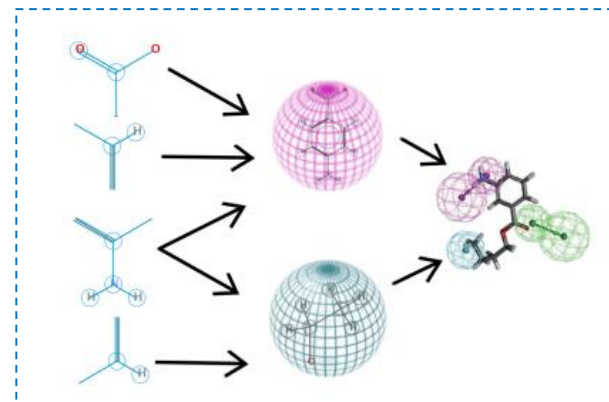


Figure . Hierarchical nature of fingerprint features: by combining the ECFP features we can build reactive centers. By pooling specific reactive centers together we obtain a pharmacophore that encodes a specific pharmacological effect.

AI創薬のDL以外の方法

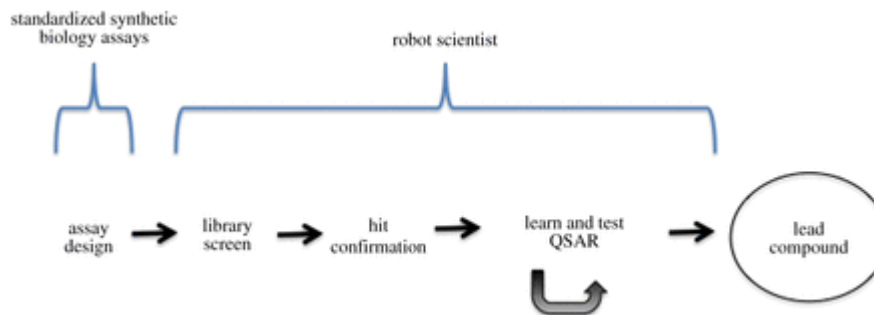
- Berg社のAI創薬
 - AIを方法として膵臓がんの抗がん剤を開発中
 - 膵臓がんと非患者の14兆のゲノム・オミックス情報を比較。
 - 調節不全パスウェイのシステム推定
 - システム薬理学的AIによる創薬
- マンチェスター大学（Cambridgeとも共同）

Artificially-intelligent Robot Scientist for new drugs

- ライブラリースクリーニング, ヒット化合物の確証, リード化合物などの自動化
- 構造活性相関 (Quantitative Structure Activity Relationship) (QSAR) を反復学習する
- 熱帯病、寄生体のDHFR (ジヒドロ葉酸還元酵素：薬剤耐性) を標的にして学習、細胞を合成生物学操作
- 血管新生阻害因子 (抗がん剤) をDR候補を探索
- 最上位にコンセプト木 (“root: assay triple screen”など)



Robot scientist Eve at work

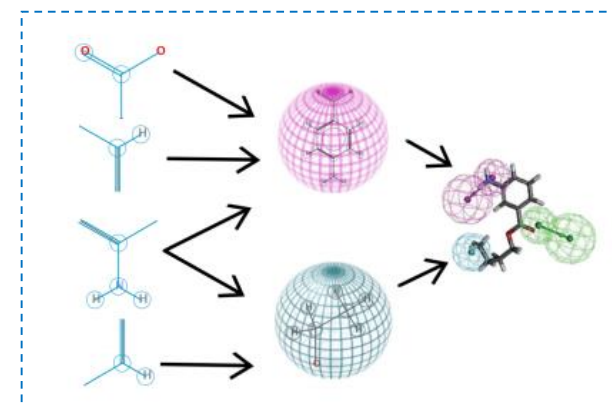
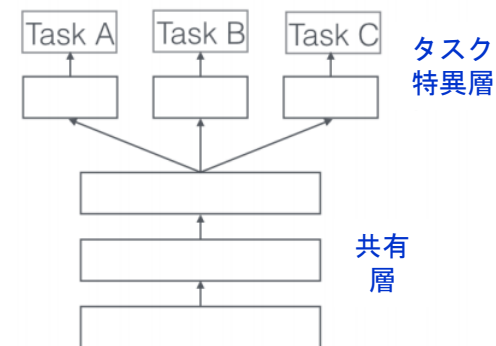


Virtual Screening (QSAR)への AIの応用

バーチャルスクリーニングへの AIの応用 (ligand-based)

- Kaggleでの課題：QSARモデル
 - マルチタスク・ディープラーニング
 - 15種類のassayのデータに1つのMT-DNN
 - 各課題に共通な層 (shared layers)を用意
 - 関連性のある他の課題のデータを利用できる
 - データ数の増大により雑音を除去できる
 - 個々の課題に単一NNを用意するより高精度 (Dahl2014)
- 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

マルチタスクDNN



ECFP(chemical substructure: Enhanced cyan fluorescent)

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.

AI創薬 第III部

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

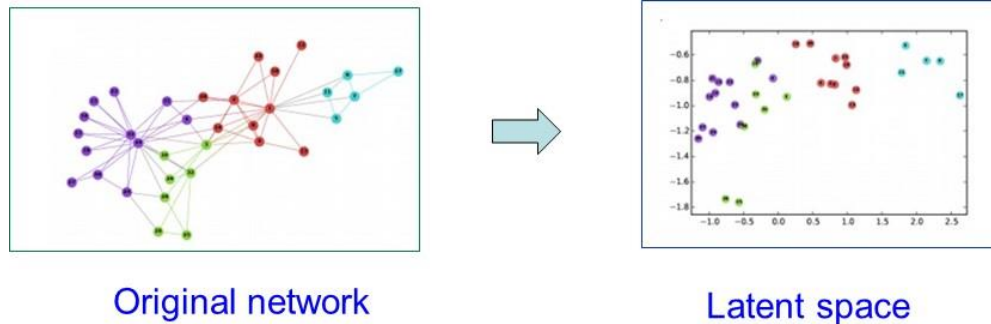
PPINはHPRDでは<1万タンパク質 × 一万タンパク質>の
超多次元ネットワークで通常の機械学習では困難



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

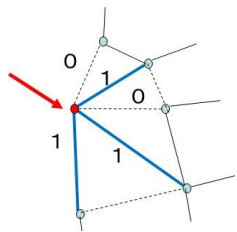
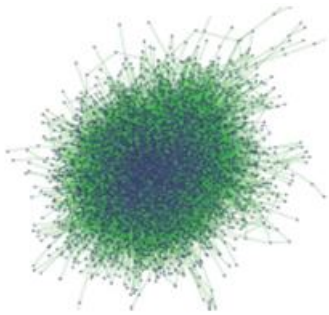
Deep Learningによる Network Embedding

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



Structural Deep Network Embedding

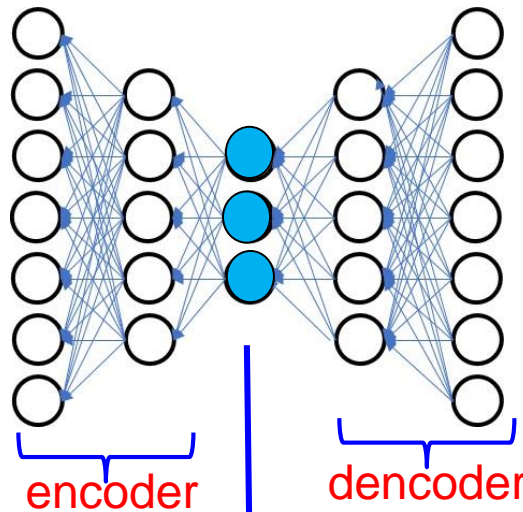
PPIN



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

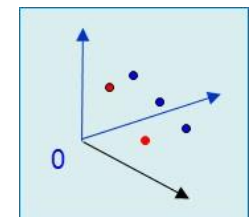
全節点の近接ベクトル

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) 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)

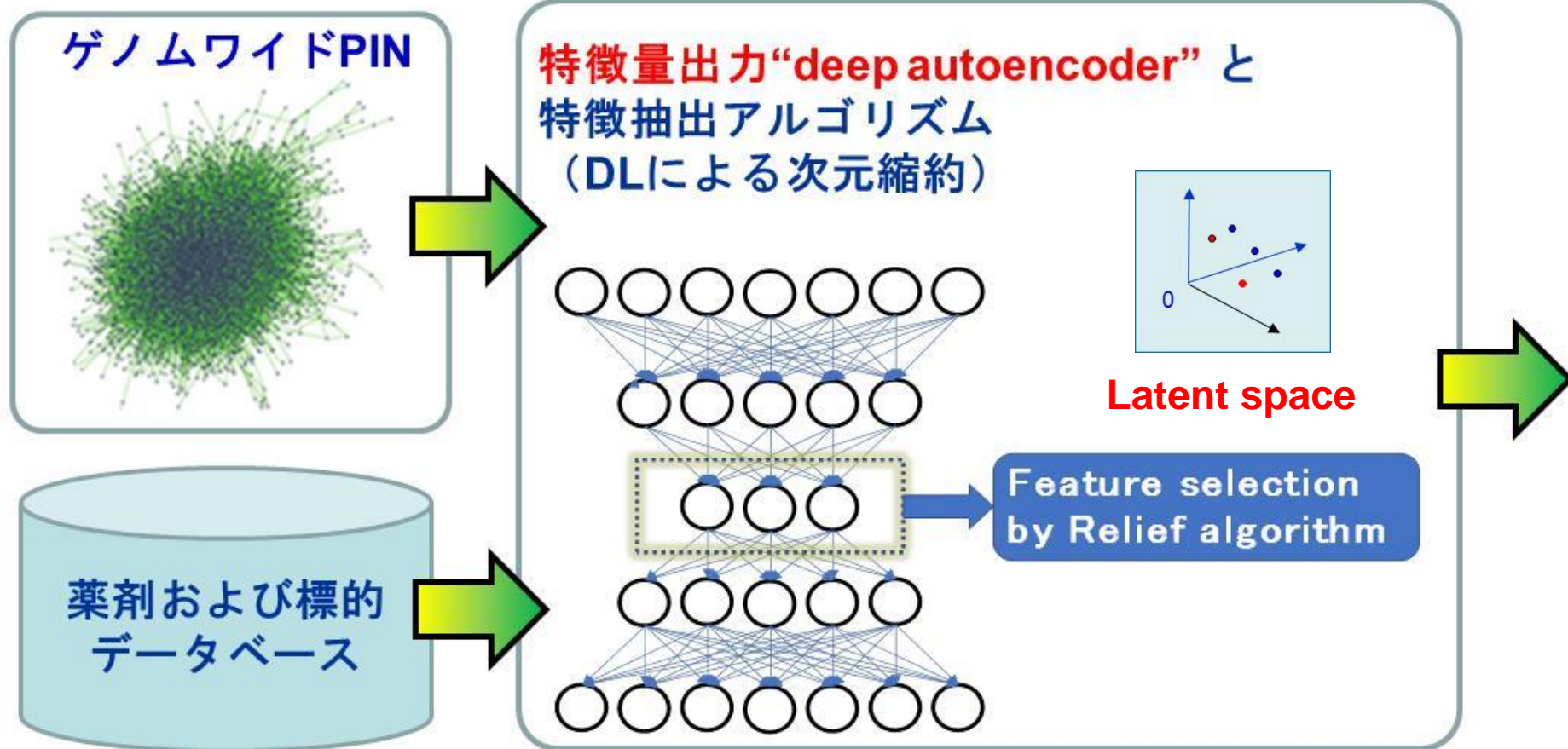
4) DL特徴NW基底量を用いた機械学習分類

- 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

Deep Learningによる創薬・DR

入力

特徴量産出



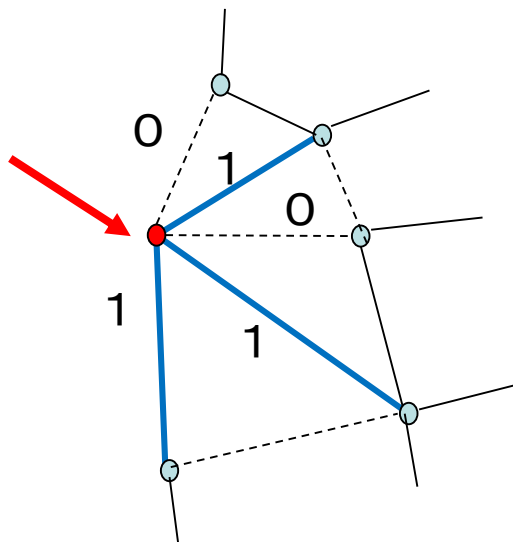
Deep Learning と SVD (singular value decomposition)の精度の違い

あるタンパク質ノードに
注目する

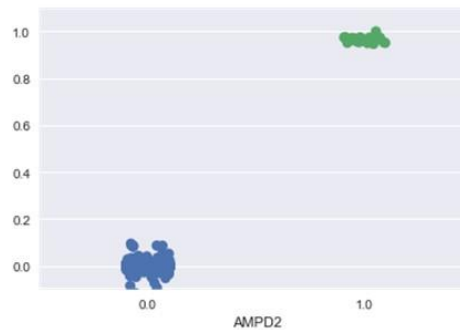
周りのノードで

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。

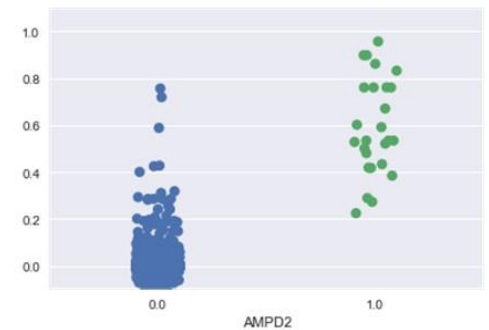
$$v_i = (0, 0, 0, 1, 0, 1, 0, \dots)$$



AMPD2 (adenosine monophosphate deaminase 2)
degree=26

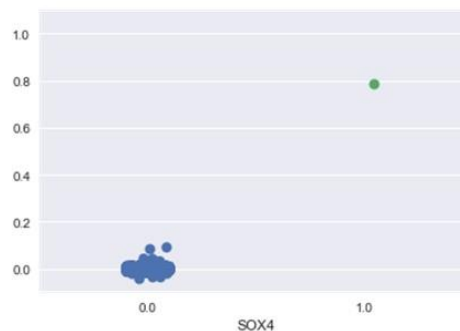


Autoencoder

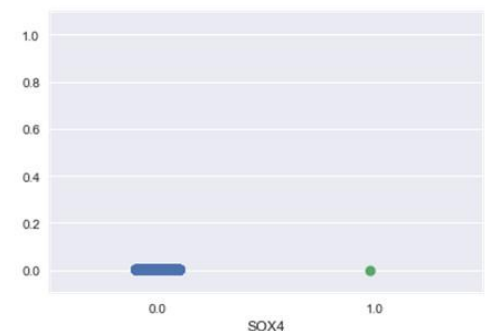


SVD

SOX4 (SRY-box 4)
degree=1



Autoencoder



SVD

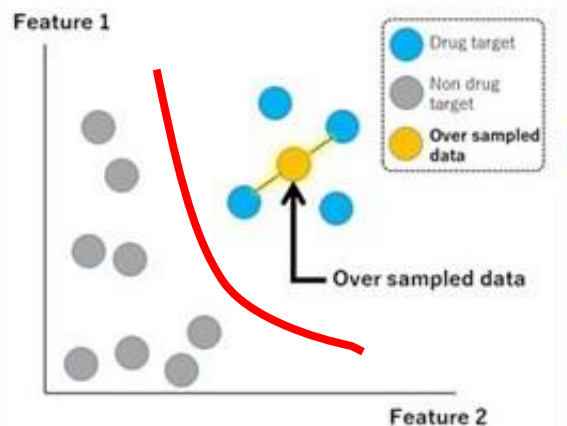
N=8,502

Deep Learningによる創薬・DR

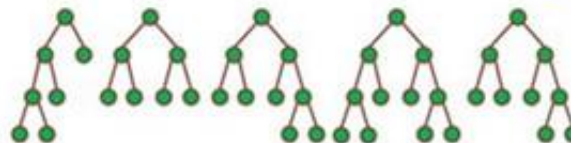
分類モデル

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

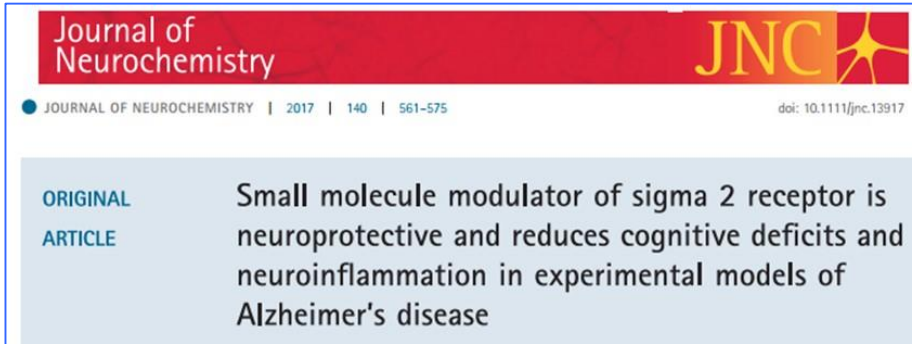


標的性判定

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

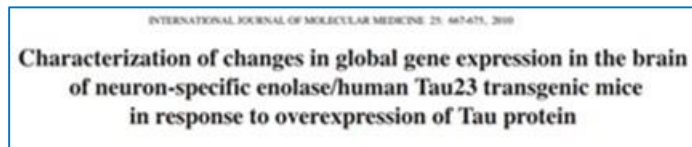
実験的研究との付合 1

PGCM1 : progesterone receptor membrane 1

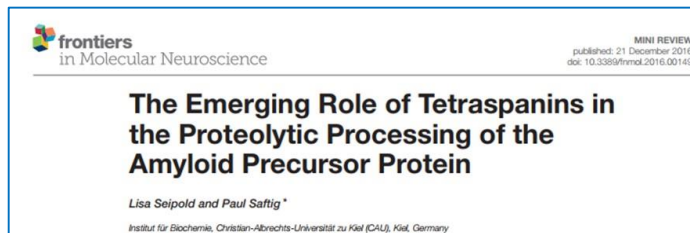


神経保護の効果 (neuroprotective) 認知不全・炎症に治療効果

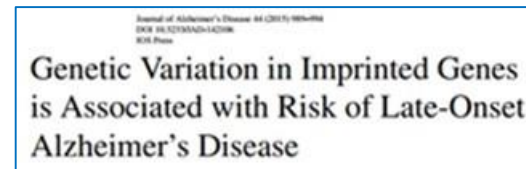
GPM6A : Glycoprotein M6A



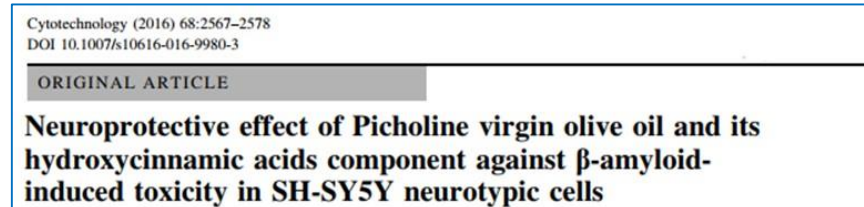
CD81: Tetraspanins family



DLGAP2 : DLG-Associated Protein 2



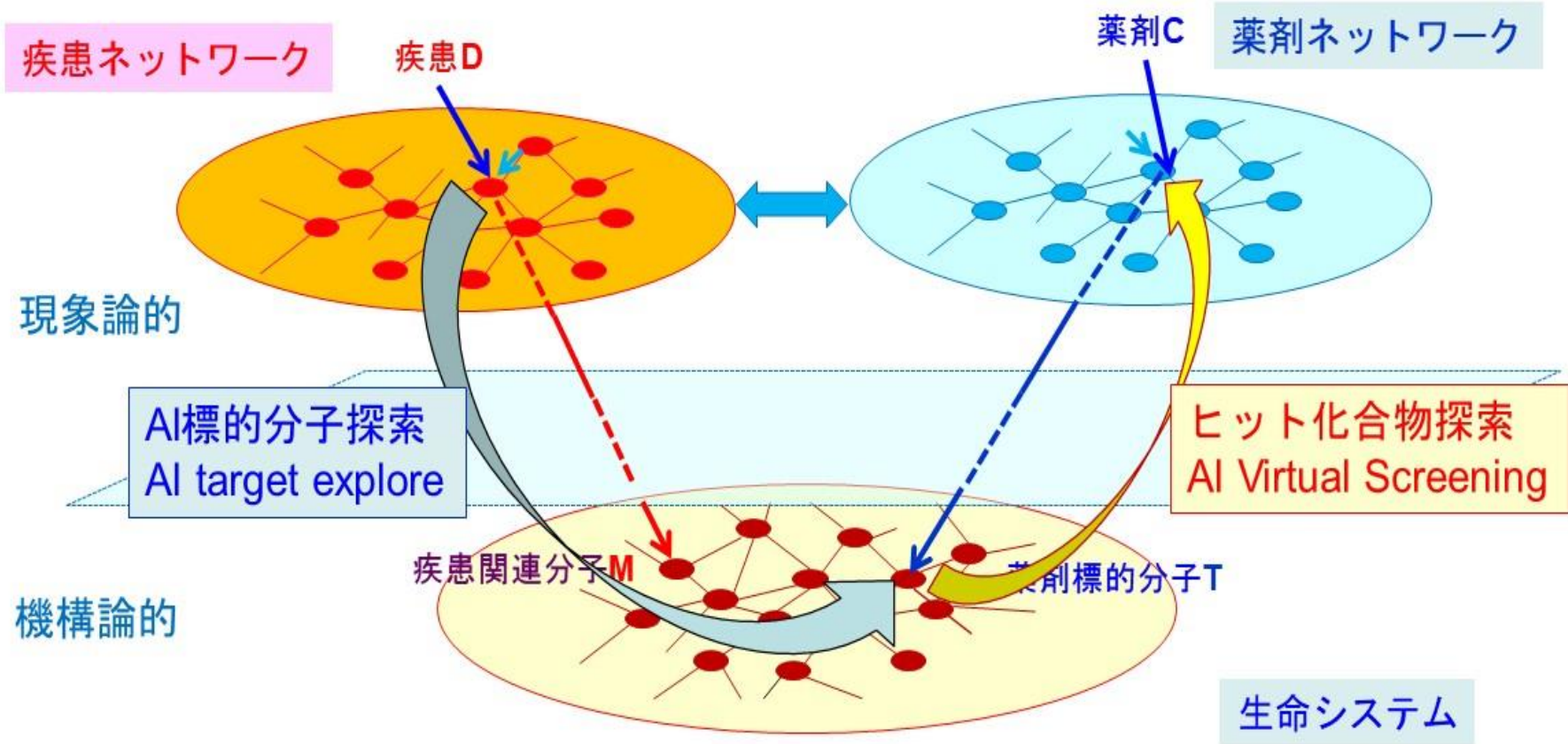
PFKM: Phosphofruktokinase



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

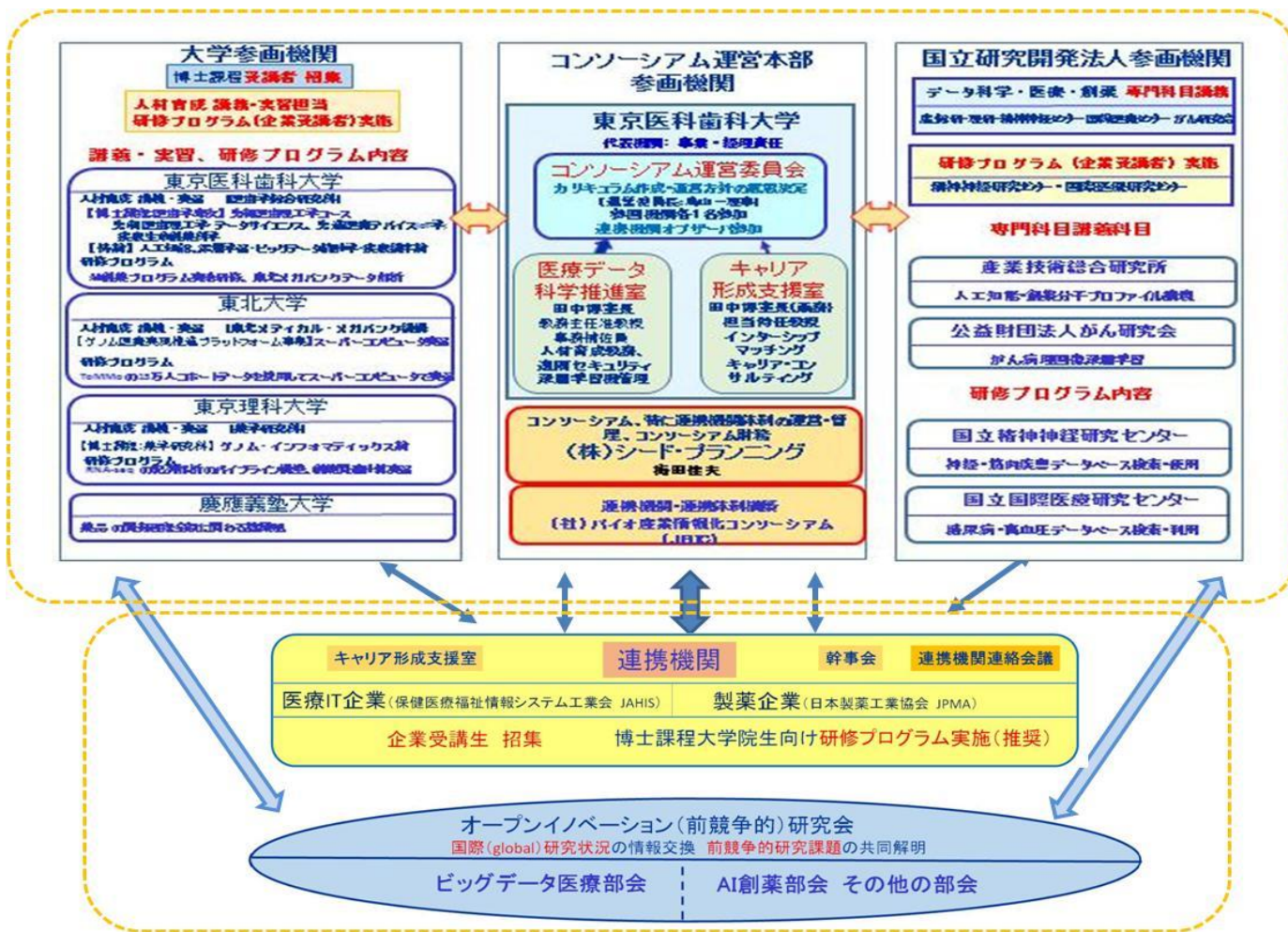
田中 博 著

「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

