遠隔インタラクティブ講義「計算生命科学の基礎V」
「計算科学・データサイエンスと生命科学の融合 基礎から医療・創薬への応用まで」

はじめに 計算生命科学の概要

受講生の皆さんへ

現代の生命科学は、急速な変革を遂げつつあります。その変革の原動力は、生物の大規模データ（ビッグデータ）の蓄積と、それに促された計算機科学・シミュレーション科学・人工知能学・データサイエンスなどの研究分野の緊密な連携、すなわちコンピュータを活用した計算生命科学の進歩です。

計算生命科学は、ゲノムの遺伝情報・生体分子の立体構造と相互作用・細胞レベルの代謝・生理や疾患までの高次生命活動の多階層のビッグデータを定量的かつシステマティックに解析し、シミュレーションにより予測して、それらの統合により生命を理解することを目指します。 その急速な発展は農学や医学の分野にも大きな影響を及ぼし、ゲノム医療などの応用も実現しつつあります。

計算生命科学は、現代の生命科学の推進に不可欠な知識を提供します。この遠隔講義では、CBI学会・日本バイオインフォマティクス学会の企画協力を得て、生命科学と理工学の学際研究領域である計算生命科学に興味を持たれる方々に、その基礎と将来の展望を学んでいただき、基礎から応用までの研究開発を支える人材の育成を目指しています。

講義の目的・範囲・対象者

講義の目的

- 「計算生命科学」の基礎と応用を学ぶ機会を遠隔講義で全国に提供
- 「計算生命科学」のさまざまな分野で活躍する人材の育成を目指す

講義内容の範囲

- 遺伝情報
  →
- 計算機科学・シミュレーション科学・人工知能学・データサイエンス
  →
- 医療応用

受講者対象

- 計算生命科学に興味をお持ちの方全て(高校生・大学生・大学院生~企業・アカデミアの研究者)

計算科学・データサイエンスと生命科学の融合
基礎から医療・創薬への応用まで

計算生命科学の基礎V

第1編 ゲノムから分子構造までの計算生命科学の基礎と実践

企画　白井　剛(長浜バイオ大学)

10月10日 臨床シークエンスの実際―情報解析を中心に―	加藤 護 (国立がん研究センター)
10月17日 機械学習によるバイオビッグデータの実践的利用	山西 芳裕 (九州工業大学)
10月24日 X線結晶解析・NMR・電子顕微鏡・AFMを統合した相関構造解析	神田 大輔 (九州大学)
10月31日 二次代謝物のデータサイエンス	金谷 重彦 (奈良先端科学技術大学院大学)

第2編 構造生命科学のための分子シミュレーション

企画　田中 成典 (神戸大学大学院)

11月　7日 生命系の分子動力学シミュレーション	池口 満徳 (横浜市立大学大学院)
11月14日 フラグメント分子軌道法に基づく構造生命科学	福澤 薫 (星薬科大学)
11月21日 溶液中における生体関連分子複合系の自由エネルギー解析	松林 伸幸 (大阪大学)
11月28日 分子シミュレーションを活用したインシリコ創薬支援	広川 貴次 (産業技術総合研究所・筑波大学)
12月　5日 QM/MM法の概略と応用 茨城大学大学院理工学研究科	森 聖治 (茨城大学)

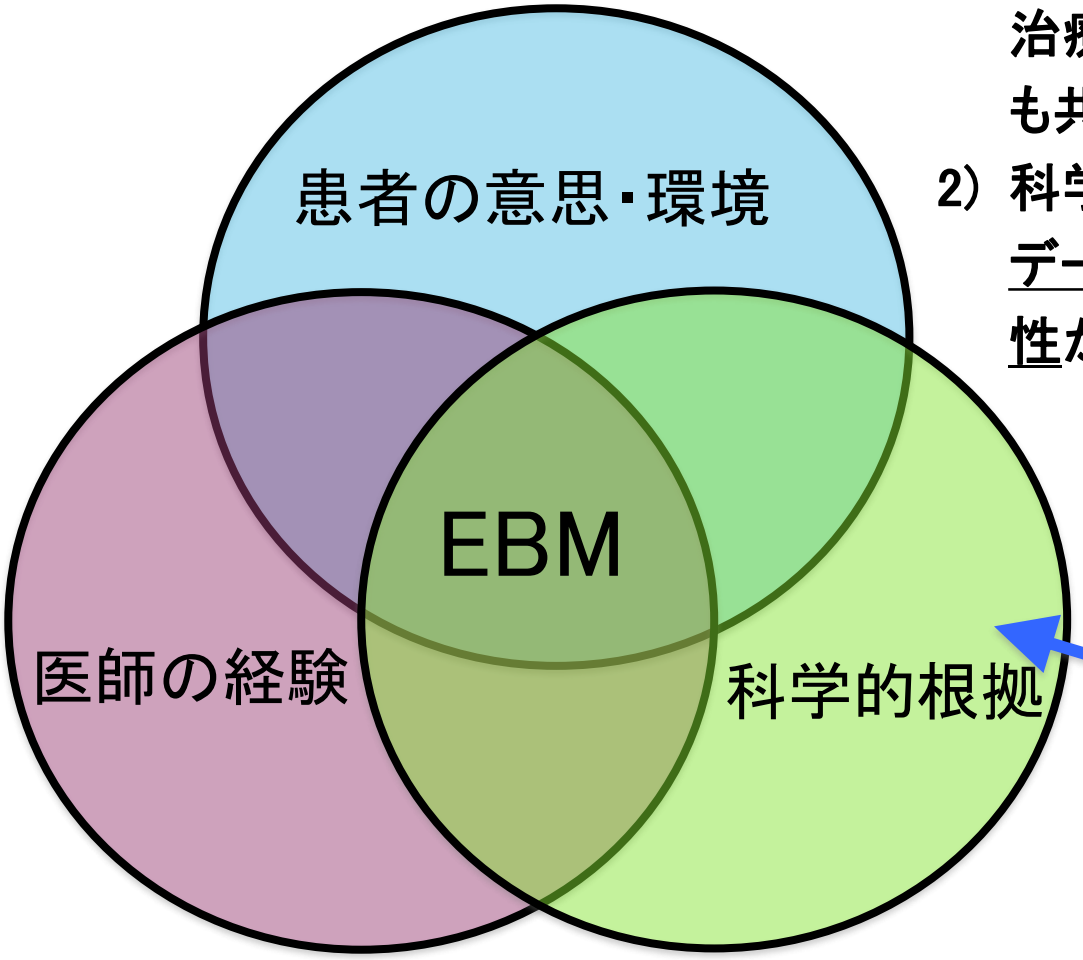
第3編 健康科学・医療・創薬への応用

企画　森 一郎 (神戸大学大学院)

12月12日 医薬品業界におけるデータサイエンティスト	都地 昭夫・北西 由武(塩野義製薬株式会社)
12月19日 さまざまな生命科学データの接続で見える新たな知見	由良 敬(お茶の水女子大学・早稲田大学)
1月　9日 シミュレーションとAIの融合による創薬	本間 光貴 (理化学研究所)
1月16日 ビッグデータを活用した健康科学への挑戦	國澤 純 (医薬基盤・健康・栄養研究所)
1月23日 脳情報の可視化とその応用	西田 知史 (情報通信研究機構)

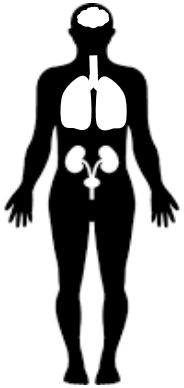
企画協力: CBI学会、日本バイオインフォマティクス学会

Evidence Based Medicine (EBM)
科学的根拠に基づく医療



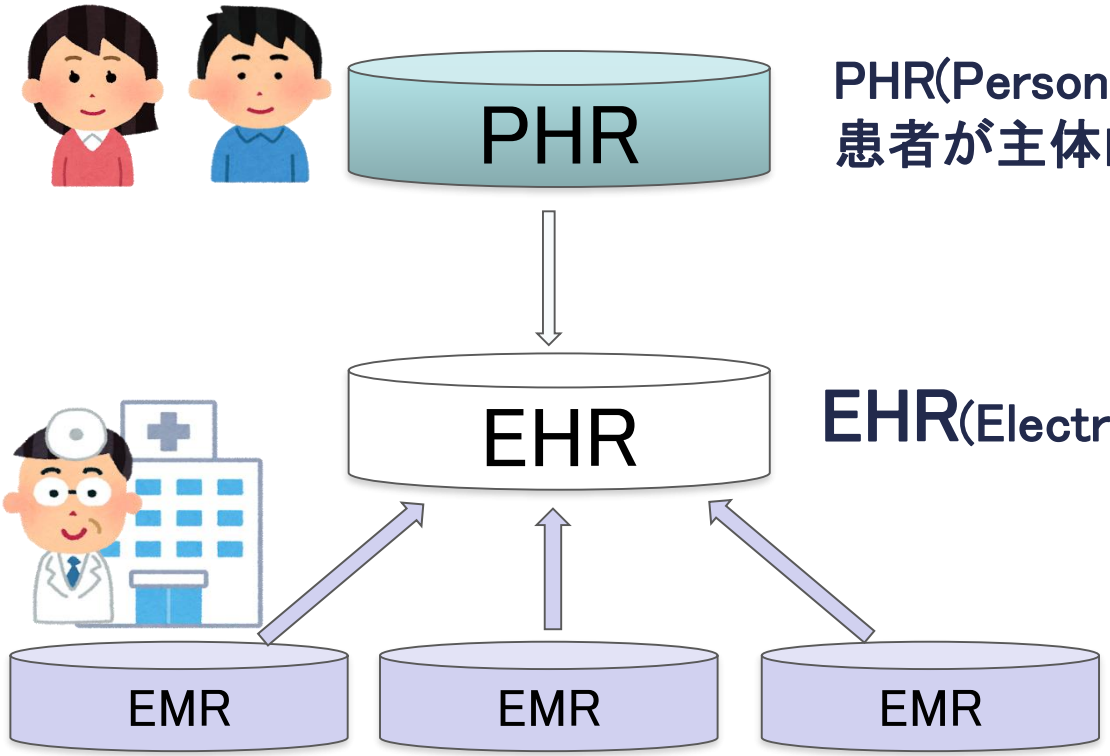
1) なるべく客観的な疫学的観察や統計学による治療結果の比較に根拠を求めながら、患者とも共に方針を決める医療。
2) 科学的根拠として、従来基礎生物学に属するデータ(ゲノミクス、構造ゲノミクスなど)の重要性が高い。

EHR(Electronic Health Record)
ゲノミクスデータ
構造ゲノミクスデータ
など



医療情報の電子化

1) 体重・体温・血圧など個人(患者)が主体的に記録した健康情報をPHR(Personal Health Record)という。最近ではスマホなど携帯機器(ウェアラブル・ヘルスケア機器)を使って記録する場合もある。
2) 医療機関(医師など)が記録する診療・治療の電子情報をEMR(Electronic Medical Record)という。主にカルテやレセプトのデータを指す。
3) PHRとEMRを統合したものがEHR(Electronic Health Record)である。



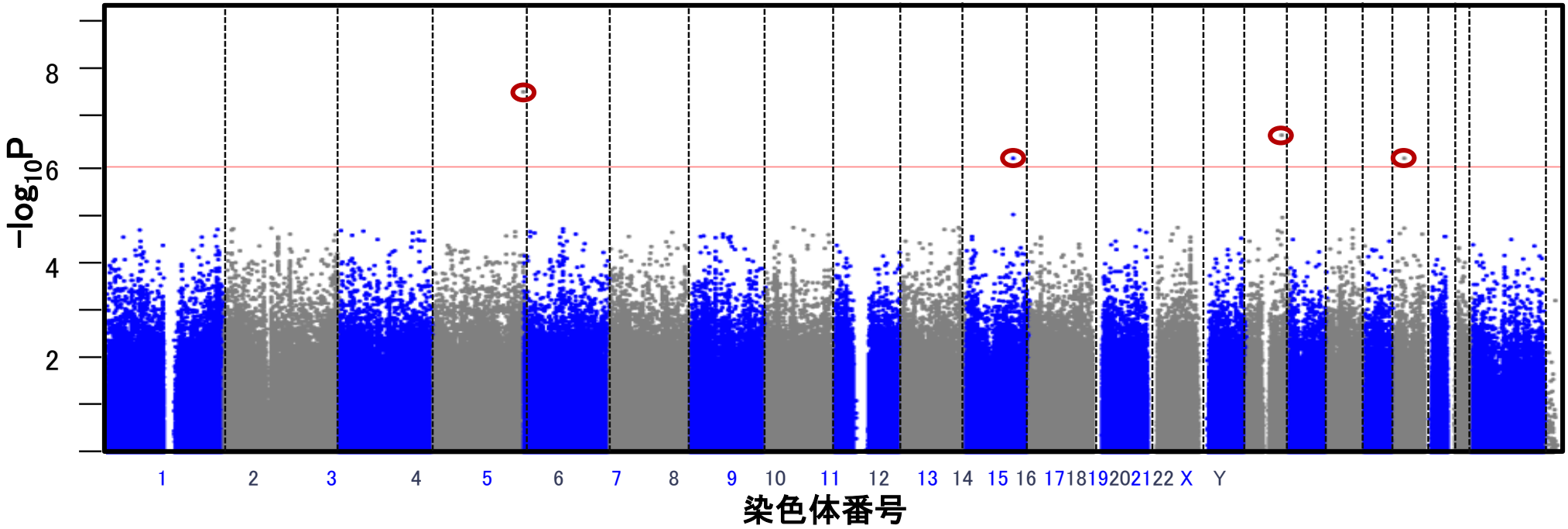
PHR(Personal Health Record)
患者が主体的に記録する健康情報

EHR(Electronic Health Record)=PHR+EMR

EMR(Electronic Medical Record)
医療機関の診療・治療情報
・電子カルテ・看護記録
・レセプト(保健請求)

全ゲノム相関解析(GWAS = Genome Wide Association Study)

1) GWASは、ゲノム全体をほぼカバーするような多数の１塩基多型（SNP=変異）や繰り返し配列多型の遺伝子型を、多人数の被験者(コントロールを含む)について決定し、主に多型の頻度と疾患や量的形質との関連を統計的に調べる方法。後ろ向きコホート研究(症例対照研究、ケースコントロール研究)の一種であるので、オッズ比やχ^2検定(P値)による評価を行う。



がんパネル検査

1）がんゲノム解析により、がんは幾つかの契機となる変異(ドライバー変異)から始まって、いくつかの変異を蓄積して段階的に悪性腫瘍化することがわかってきた。

正常		良性腫瘍		悪性腫瘍
First-Hit がん抑制遺伝子変異 APC遺伝子	⇨	Second-Hit 正常がん抑制遺伝子の不活性化 APC遺伝子 B-catenin遺伝子	⇨	がん遺伝子活性化変異 がん抑制遺伝子の欠失 K-RAS遺伝子 p53遺伝子 SMAD2/4遺伝子	⇨	遺伝子変異大量蓄積 染色体異常 Telomerase遺伝子 その他多くの遺伝子

2）ドライバー変異の種類とその後蓄積する変異によって、がんの予後と有効な治療薬は大きく異なる。このことから、治療方針を決定する際に患者の特定の遺伝子(がんパネル)を配列決定し、変異を特定するクリニカルシークエンスが提唱され、保険適用を目指していくつかの拠点病院で先進医療として実施されている。

	変異サイト												
遺伝子	乳がん				大腸がん				肺がん				有効な薬
PIK3CA	*			*	*					*	*		mTOR阻害薬
BRCA1/2		*		*				*					PARP阻害薬
ERBB2				*	*		*				*		
EGFR	*									*			EGFR阻害薬
ALK									*				ALK阻害薬

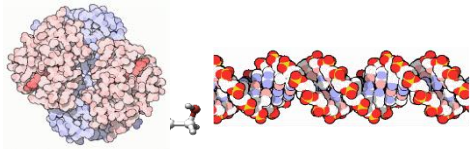
がんパネルの例

http://www.mss.co.jp/business/life-science/plession/about.html

第1編 ゲノムから分子構造までの計算生命科学の基礎と実践

10月10日 臨床シークエンスの実際―情報解析を中心に―
加藤 護（国立がん研究センター）

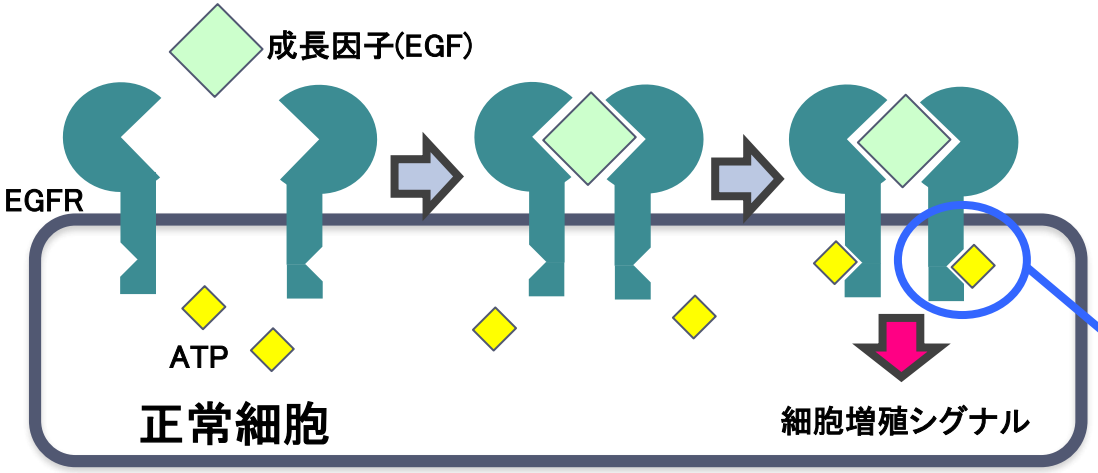
10月17日 機械学習によるバイオビッグデータの実践的利用	山西 芳裕（九州工業大学）
10月24日 X線結晶解析・NMR・電子顕微鏡・AFMを統合した相関構造解析	神田 大輔（九州大学）
10月31日 二次代謝物のデータサイエンス	金谷 重彦（奈良先端科学技術大学院大学）

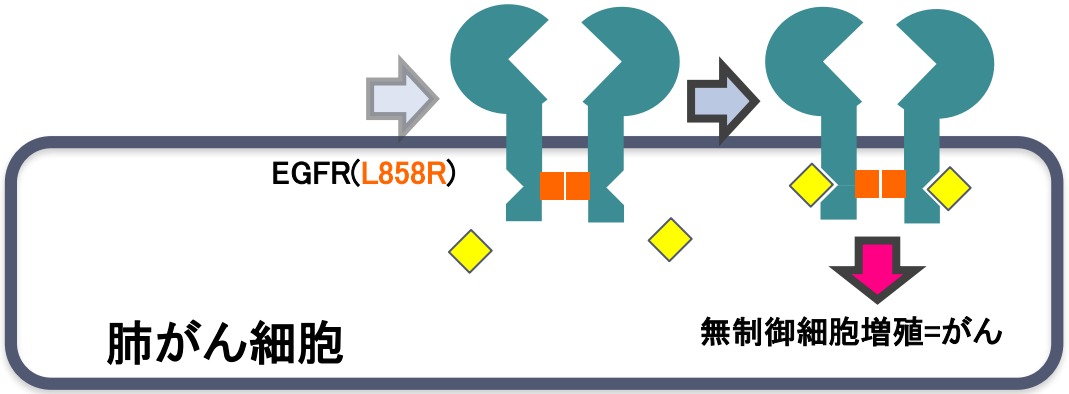


なぜ病気になるのか？

1） 細胞の増殖は制御されている。成長因子(ホルモン)が分泌される→細胞表面の上皮成長因子受容体(EGFR：タンパク質)に結合→EGFRが2量体化→細胞内でATPを結合しリン酸化される過程を経て、細胞核に増殖(分裂)シグナルが伝達される。

2） 肺がん患者の多くで、EGFRの858番目のアミノ酸Leu(ロイシン)がArg(アルギニン)に変異している(L858R)。L858Rは成長因子がなくてもEGFRが2量体化し、無制限に増殖シグナルを発生する。つまり、窒素原子が差し引き3原子増えるだけで病気になる！

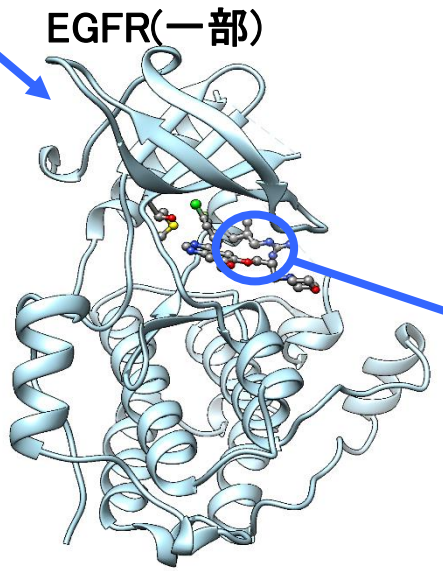


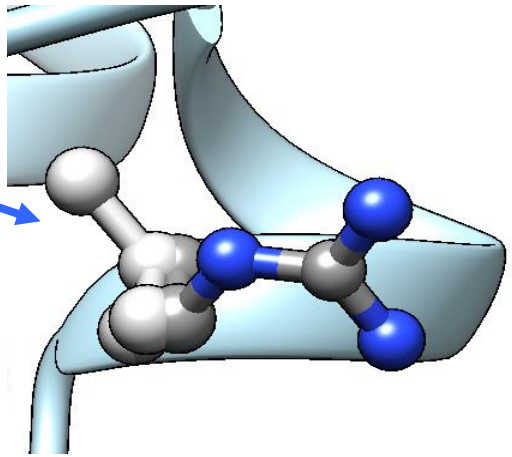


正常なEGFR遺伝子 ...GGGCTGGCC... (Leu)

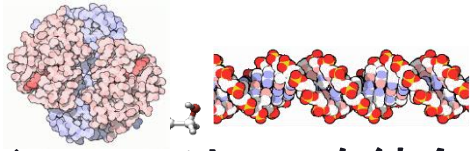
x

患者のEGFR遺伝子 ...GGGCGGGCC... (Arg)





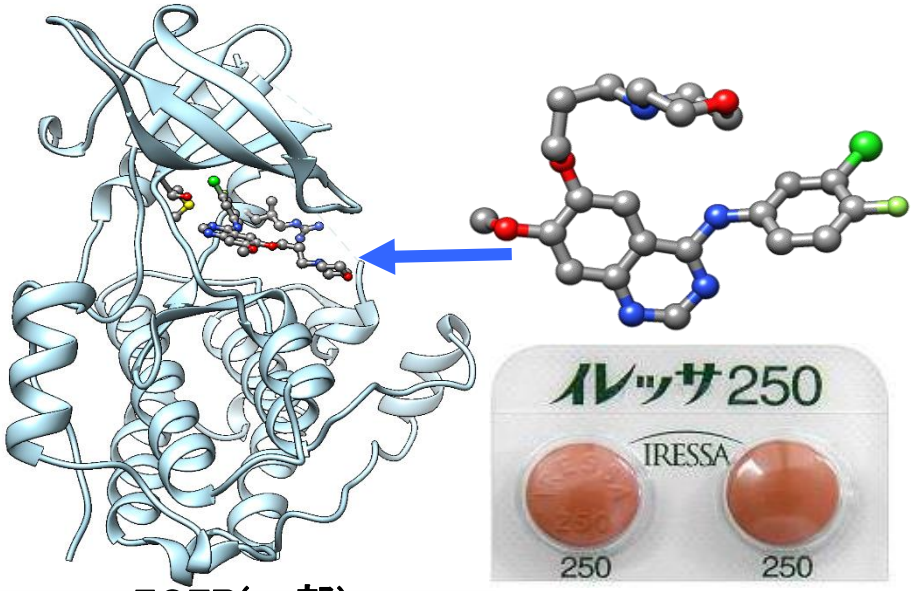
858番目のLeu→Arg (L858R)



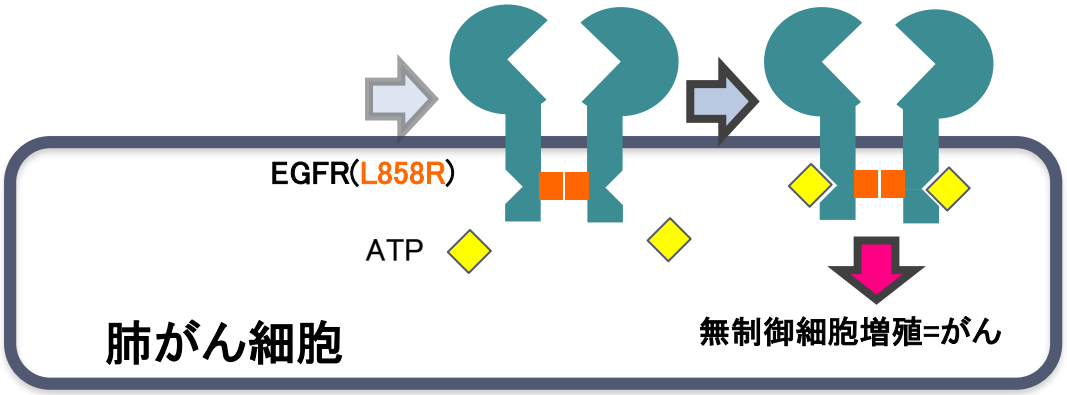
なぜ薬で病気が治るのか？

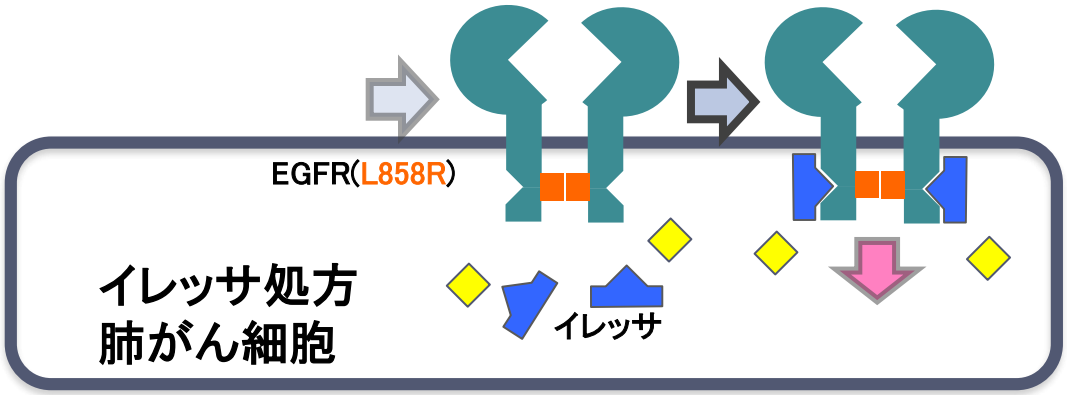
1) EGFRはATPを結合・分解しないと増殖シグナルを送れない。これを特異的にブロック(阻害)すればがん細胞の増殖を止められる。

2) 抗肺がん薬イレッサは、タンパク質の構造に基づく薬剤設計（Structure-Based Drug Design, SBDD）によって設計され L858RのEGFRのATP結合部位に特異的に結合し、ATPの結合を阻害する。

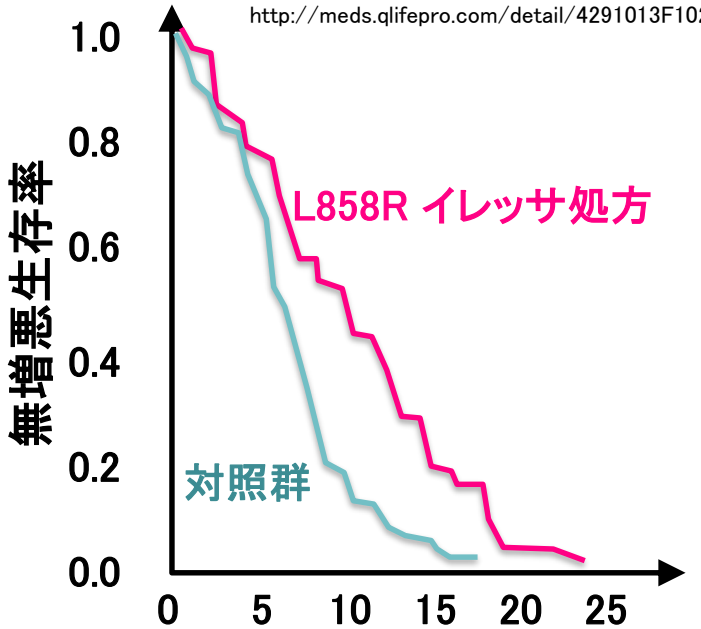


EGFR(一部)　　イレッサ



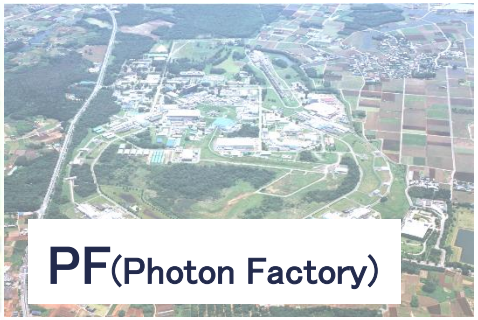


http://meds.qlifepro.com/detail/4291013F1027/



Mok *et al. New Engl J Med.* 361, 947(2009)

X線結晶解析(XP)・核磁気共鳴法(NMR)・電子顕微鏡法(EM)



X線結晶解析(XP)

核磁気共鳴法(NMR)

90%

8%

電子顕微鏡法(EM)

< 2%

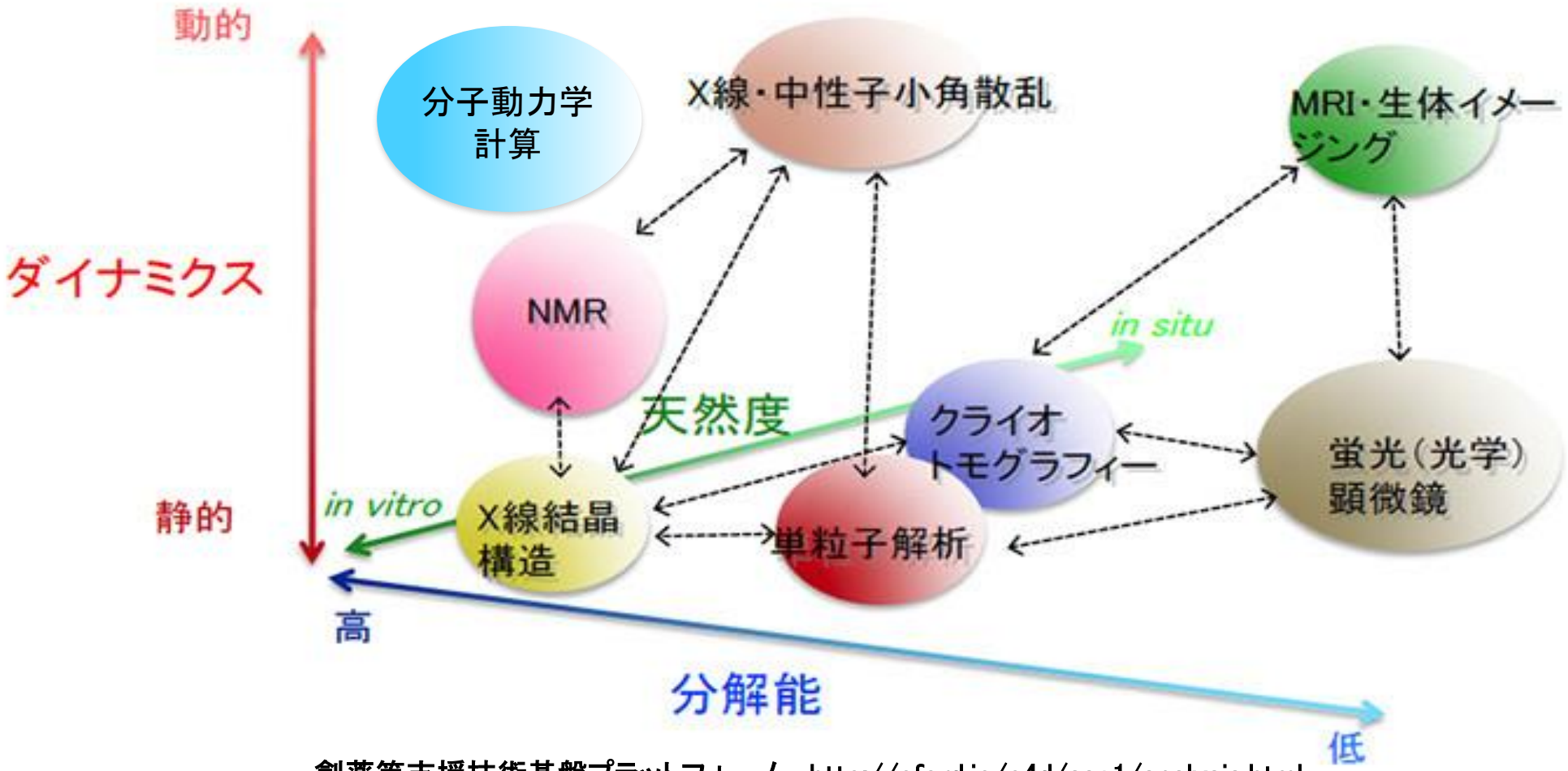
WORLDWIDE PDB
PROTEIN DATA BANK

~145,000 entries

https://www.rcsb.org/stats/growth

相関構造解析(correlated structure analysis)

1) 様々な実験＋理論手法からのデータを総合して生体超分子の構造・ダイナミクス・機能を解析する研究を、相関構造解析(correlated structure analysis または Hybrid/Integrated structure analysis)という。



創薬等支援技術基盤プラットフォーム http://pford.jp/p4d/sac1/analysis.html

第1編 ゲノムから分子構造までの計算生命科学の基礎と実践

10月10日 臨床シークエンスの実際―情報解析を中心に―　加藤 護 (国立がん研究センター)
10月17日 機械学習によるバイオビッグデータの実践的利用　山西 芳裕 (九州工業大学)

10月24日

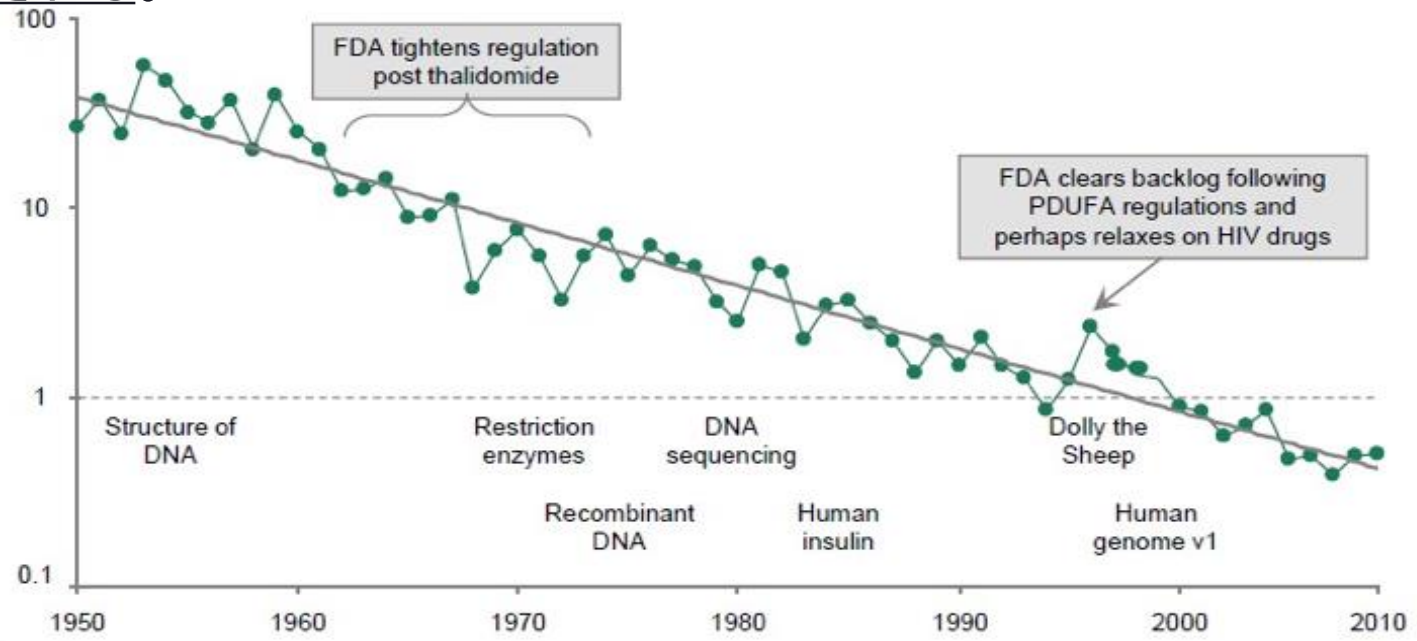
X線結晶解析・NMR・電子顕微鏡・AFMを統合した相関構造解析

神田 大輔 (九州大学)

10月31日 二次代謝物のデータサイエンス　金谷 重彦 (奈良先端科学技術大学院大学)

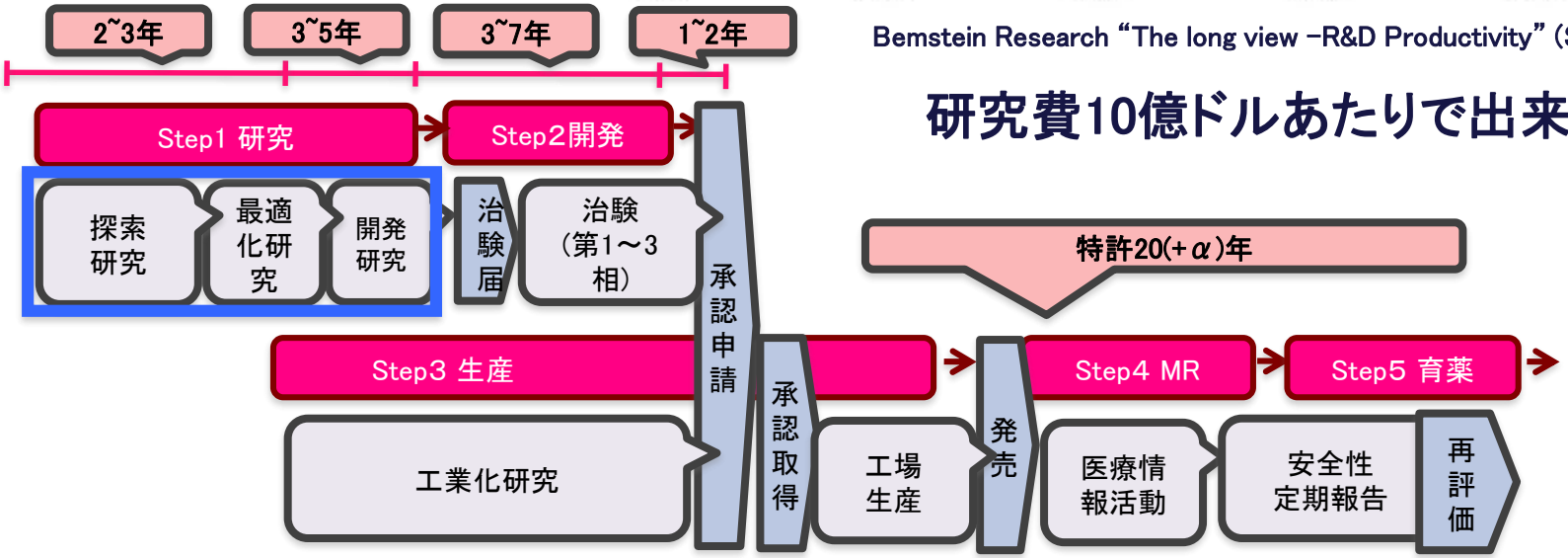
容易なドラッグターゲットの枯渇

1) 新薬研究開発費は高騰し続けており、<u>データサイエンスを活用した探索研究～開発研究の効率化が必要とされる</u>。



Bemstein Research “The long view –R&D Productivity” (Sep. 30 2010)

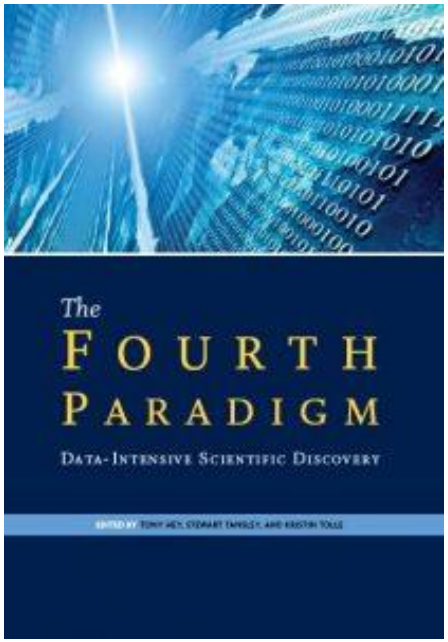
研究費10億ドルあたりで出来る新薬の数



創薬プロセス

The Fourth Paradigm: Data-Intensive Scientific Discovery (2009)

https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/



1st : 数学的手法と経験的手法(実験)による科学
（アリストテレス、BC1世紀～）

2nd : 論理構築による科学
（ライプニッツ、AD18世紀～）

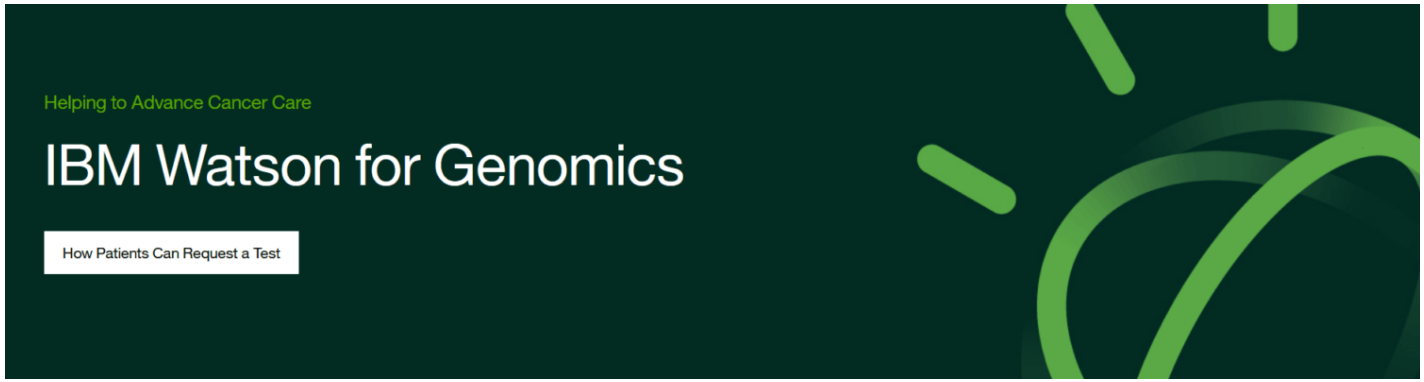
3rd : コンピュータシミュレーションによる科学
（ジョン・フォン・ノイマン、AD20世紀～）

4th : データ集約型科学(1st~3rdを内包する)
（ジム・グレイ、AD21世紀～）

存在するが「知らないデータ」の抽出と活用

「AI、がん治療法助言　白血病のタイプ見抜く」

日本経済新聞　2016/8/4



膨大な医学論文を学習した人工知能(AI)が、診断が難しい60代の女性患者の白血病を10分ほどで見抜いて、東京大医科学研究所に適切な治療法を助言、女性の回復に貢献していたことが4日、分かった。女性患者は昨年、血液がんの一種である「急性骨髄性白血病」と診断されて医科研に入院。2種類の抗がん剤治療を半年続けたが回復が遅く、敗血症などの危険も出た。そこでがんに関係する女性の遺伝子情報をワトソンに入力すると、急性骨髄性白血病のうち「二次性白血病」というタイプであるとの分析結果が出た。ワトソンは抗がん剤を別のものに変えるよう提案。女性は数カ月で回復して退院し、現在は通院治療を続けているという。

東大とIBMは昨年から、がん研究に関連する約2千万件の論文をワトソンに学習させ、診断に役立てる臨床研究を行っている。〔共同〕

データサイエンスによる医療・創薬

ビッグデータ

- ゲノム・遺伝子発現プロファイルGWASなど(構造化データ)
- 文献・画像など(非構造化データ)
- スパースデータ　属性値数(p) >> サンプル数(n)

⇩ ディープラーニングなどの機械学習 ⇩

ドラッグリポジショニング（薬剤適用拡大）

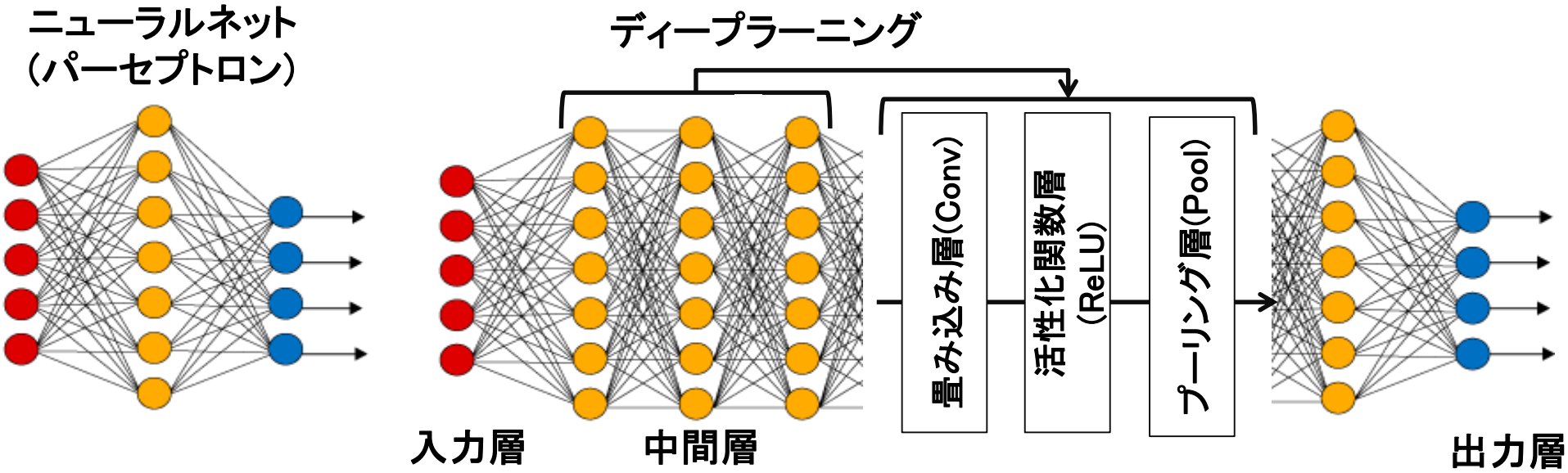
- **臨床試験(フェーズ3)で脱落した医薬品の有効活用**
- **安全性試験などの省力化が可能**

天然物(生薬・二次代謝物)の利用

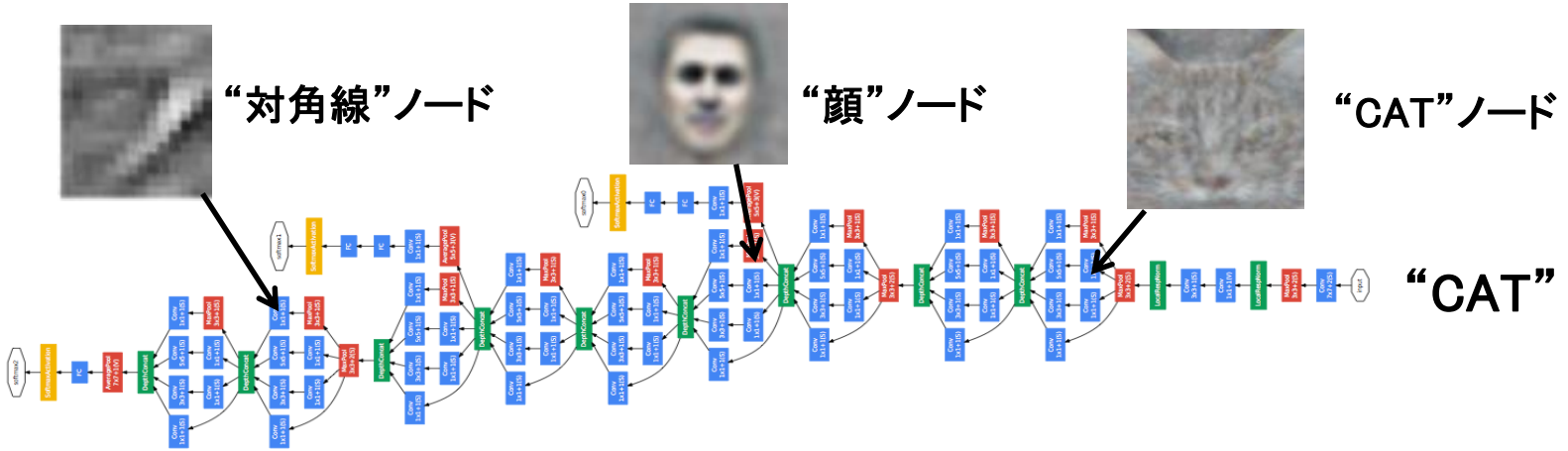
- **医薬品未利用の生薬などの有効成分の有効活用**
- **新規薬剤骨格の抽出**

ディープラーニング

1) 従来のニューラルネットと比べ高度に多階層化されている。各層で自己符号化を行い、これを順次積上げる(autoencoder stack) ので高次の特徴量が作られる。
2) スパースデータの「次元圧縮」と「内在的特徴量の抽出」に優れているとされる。



Google cat



https://googleblog.blogspot.com/2012/06/using-large-scale-brain-simulations-for.html

第1編 ゲノムから分子構造までの計算生命科学の基礎と実践

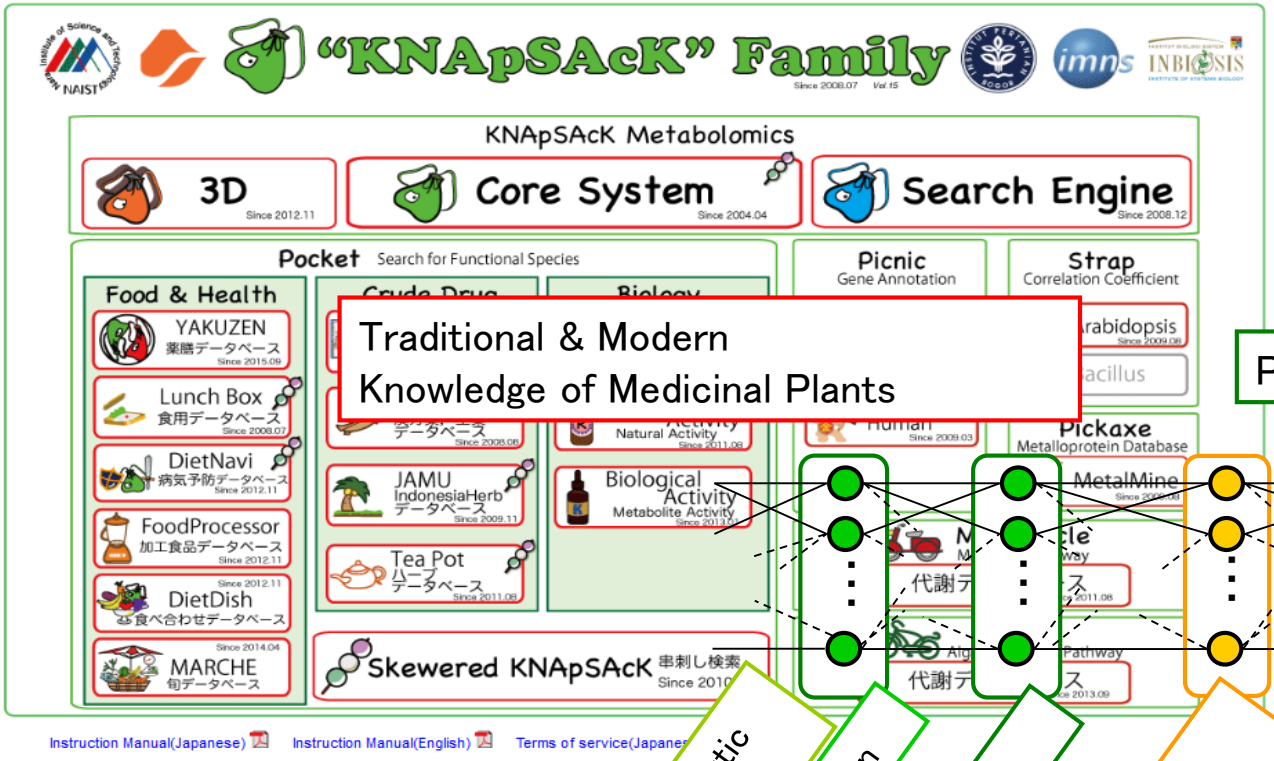
10月10日 臨床シークエンスの実際―情報解析を中心に―　　加藤 護（国立がん研究センター）

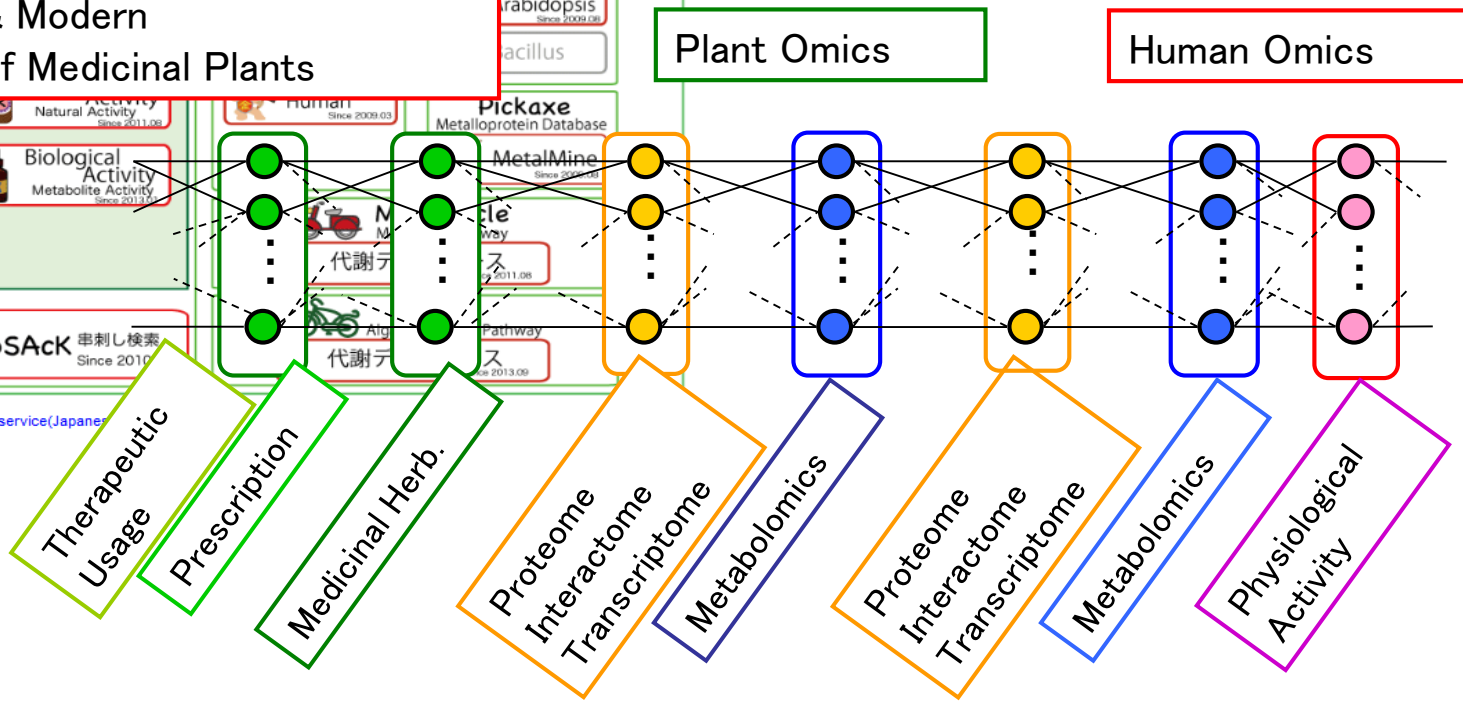
10月17日 機械学習によるバイオビッグデータの実践的利用
山西 芳裕（九州工業大学）

10月24日 X線結晶解析・NMR・電子顕微鏡・AFMを統合した相関構造解析　　神田 大輔（九州大学）
10月31日 二次代謝物のデータサイエンス　　金谷 重彦（奈良先端科学技術大学院大学）

天然物(生薬・二次代謝物)の利用のためのデータサイエンス

1)　天然物の構造–機能相関DB(薬用生物2万種、生物-天然物11万組、生薬-効能1万組、漢方処方810種) KNApSAcKのデータから新規薬剤骨格や効能の探索。





KNApSAcK

http://kanaya.naist.jp/KNApSAcK/

第1編 ゲノムから分子構造までの計算生命科学の基礎と実践

10月10日 臨床シークエンスの実際―情報解析を中心に―	加藤 護 (国立がん研究センター)
10月17日 機械学習によるバイオビッグデータの実践的利用	山西 芳裕 (九州工業大学)
10月24日 X線結晶解析・NMR・電子顕微鏡・AFMを統合した相関構造解析	神田 大輔 (九州大学)

10月31日 二次代謝物のデータサイエンス

金谷 重彦 (奈良先端科学技術大学院大学)