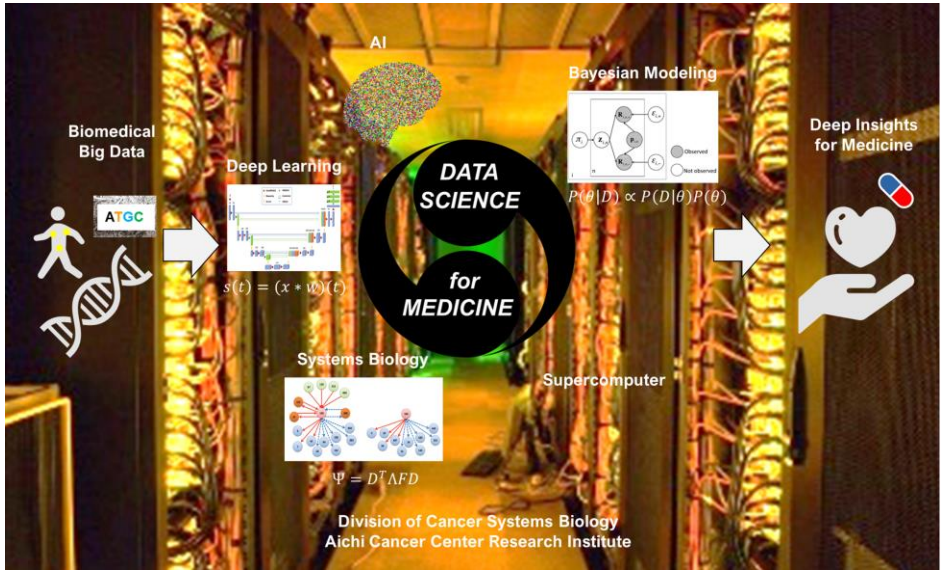
AIとスパコンを活用したがんビッグデータ解析手法の開発

愛知県がんセンター
システム解析学分野

システム解析学分野について

データ科学の力で医療へ貢献！



- がん細胞から得られた**ゲノムデータ**などの**生体ビッグデータ**を**AI**と**スーパーコンピュータ**（スパコン）を使って**解析する方法**の研究を行っています
- 患者さんのデータから、**がん細胞の複雑なシステム**に関わる情報を抽出し、**一人ひとりに合わせた医療**へつなげることを目指しています

研究室メンバー

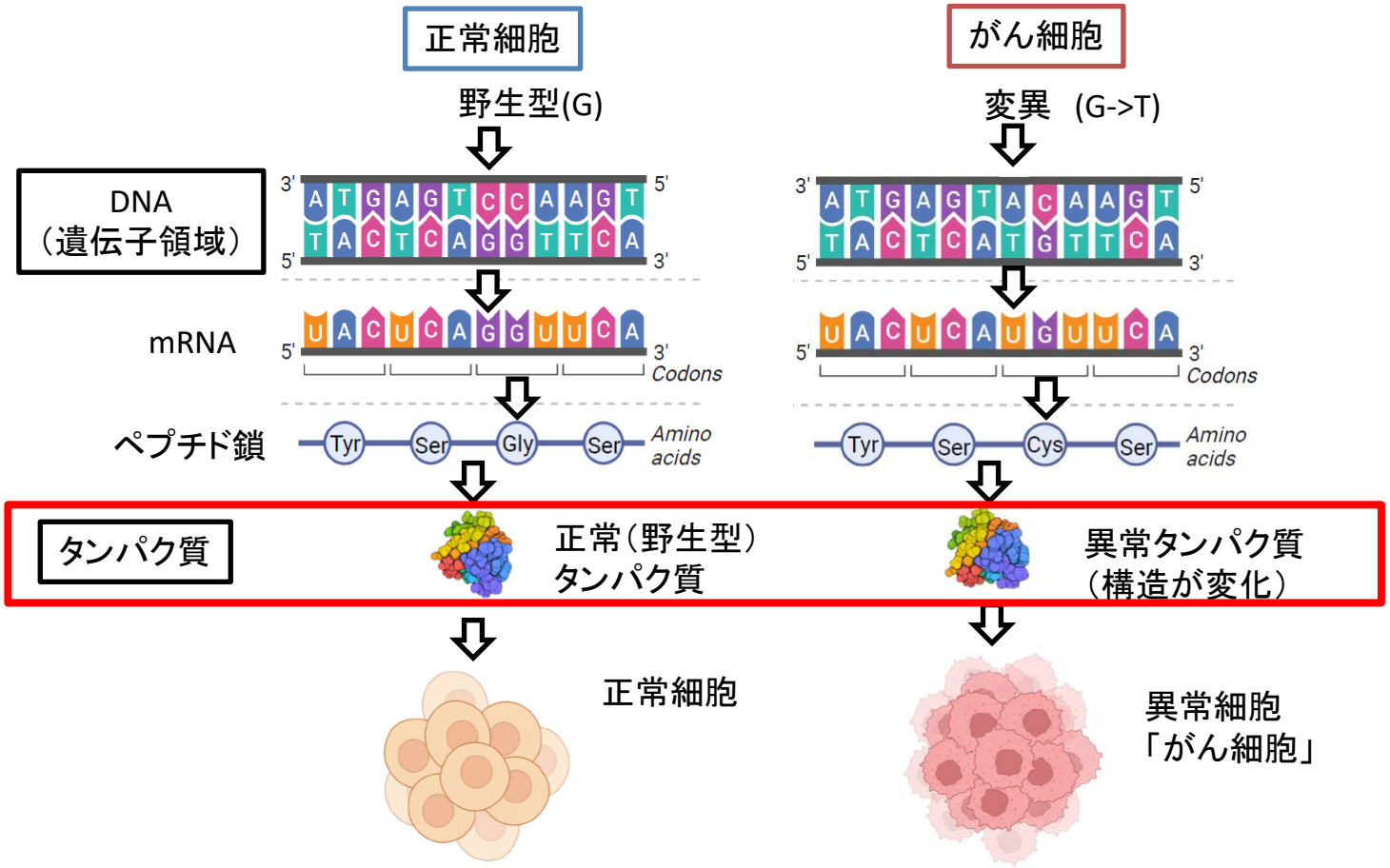
分野長：	山口　類
研究員：	郭　中梁
リサーチレジデント：	武藤　理
	木曽田 暁
任意研修生：	末廣 智也
連携大学院生：	袁　一凡
研究補助：	鈴木 一基
事務員：	竹中 亜由美

研究室HP
https://cancer-c.pref.aichi.jp/site/folder5/1154.html

はじめに

- がん細胞は、正常な細胞中の**DNA（ゲノム）配列**に**変異**が起きることにより、**タンパク質の構造**が変化するなどして、細胞が異常に増殖する能力を獲得しています
- 計測技術の発展により、DNA等の生体分子を計測した大量のデータを得ることができるようになっています
- がん細胞中のDNAに変異がおきる場所やパターンは多様であり、がん細胞の性質を知ることや薬を選択するために、コンピュータを使って計測データからDNAの塩基配列を決定し、またどこにどのような変異が生じているかを正確に検出する必要があります（**研究１**）
- さらに変異の結果タンパク質の性質がどのように変化するかを予測することも重要です（**研究２**）
- 研究のサイクルを加速し成果を現場に還元するには、得られた大量のデータを迅速かつ簡単に解析できるシステムも必要です（**研究３**）
- ここでは、近年発展している、がんのビッグデータ解析の状況をシステム解析学分野での研究を交えて概説します

DNAと変異



細胞中の**DNA**には**遺伝子**と呼ばれる領域（約２万か所）が含まれています。
遺伝子領域には**タンパク質の設計図**となる情報が含まれています。
多くの場合がん細胞では**遺伝子に変異**が入り、**構造が変化した異常なタンパク質**が作られることにより、がんの性質に影響を与えています。またその変異は多様です。そのため**遺伝子変異を正確に検出**し**タンパク質の性質を推測**することが重要です。

背景：DNA等の生体分子が大量かつ高速に計測可能に

次世代シークエンサー（NGS）: DNA配列等を読む機械

2021年8月時点で約512ドル

近い将来100ドルを切る予想
（解析費用は含まない）

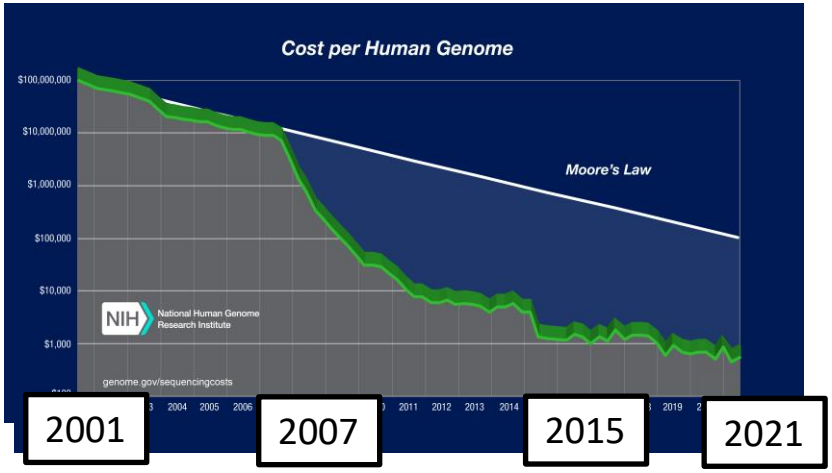
大規模ゲノム解析プロジェクト
Genomics England
All of Us
ToMMo

厚労省・ゲノム解析実行計画
AMED 大規模臨床がん全ゲノム
データ解析プロジェクト

NHGRIの予測：2025年までに
40ExaB (10^9GB)のシークエンスデータが全世界で産生
全ゲノムに換算すると
約2億人分

100億円

一人分のゲノムを決定するためのシークエンスコスト



10万分の1

10万円

https://www.genome.gov/about-genomics/fact-sheets

次世代シークエンサー(**NGS**)と呼ばれる計測機械の性能向上により、一人分の全ゲノム情報を得るのにかかるコストは20年で10万分の1になりました。

これらのシークエンスデータからがん細胞特有の変異・変化を見つけることで、がんの原因や治療法に関する情報を得ることができます。

現在、保険診療で行われている**がんゲノム医療**でも役立てられています。

シークエンスデータって、どんなデータ？

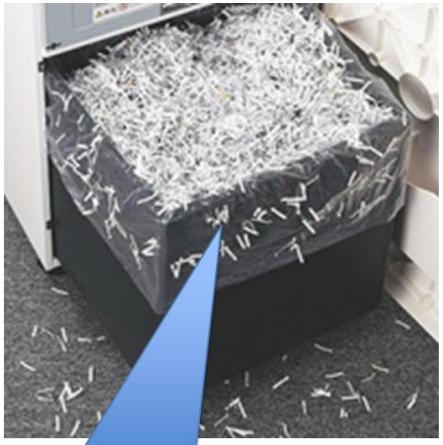
生のサンプル

DNAとして抽出
ATCGの4種の文字

次世代シークエンサー

得られるデータ ゲノムシュレッダー

100文字ぐらいの断片になった
21億ピース
の文字列断片がコンピュータに
吐き出される



ATCCGGTAAAT.....TTCA
100～150塩基

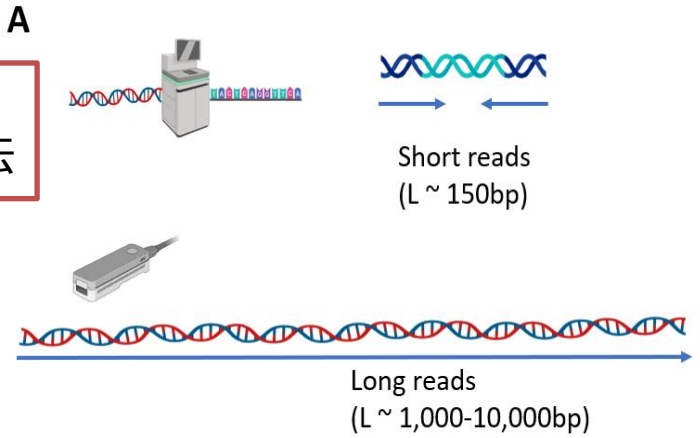
※全ゲノムシークエンスを想定

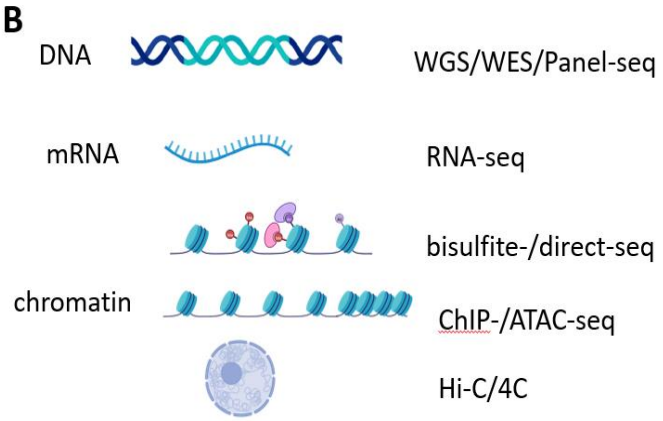
一人分のゲノムは、**4種類の塩基**（A、T、G、C）からなる塩基対、**約30億個分**からなります。現在主流のシークエンサーでは、一度にはDNA中の**約100塩基対分**しか読めません。これは**ATGCの4文字からなる30億文字分の文章が書かれた書類**の束が、シュレッダーにかけられ**100文字ずつの断片**になって出てくるようなものです。通常、がん細胞特有の変異を見つける場合、**がん細胞由来の書類を40コピー分、正常細胞由来の書類を30コピー分**、シークエンサーで読み取ります。つまり**21億ピースの文字列断片**が得られます。**ここから各ピースがゲノム上のどこにあったかを正確に決定し、**
がん細胞と正常細胞の違いを見つけ出す（変異を検出する）必要があります。

生体分子データの多様化

NGSでは、DNAに限らずRNA等他の分子の情報も得ることができます。
またこれらの分子は、多数の細胞からなるがん組織だけでなく、
単一の細胞ごとや腸内細菌叢から得ることもできるようになっています。
このような複雑なデータの解析にはAIや機械学習（ML）の手法が必要です。

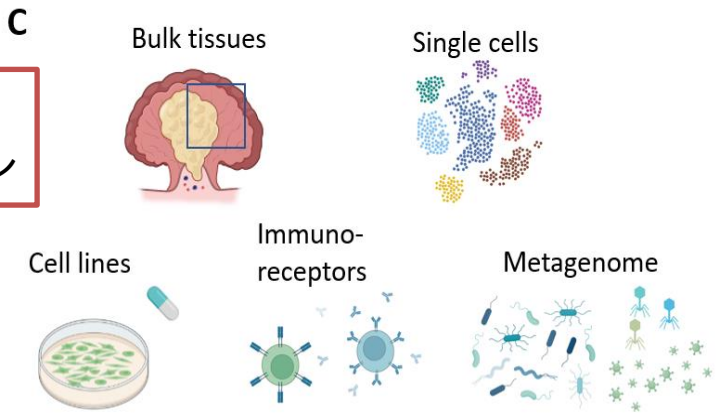
多様な
計測方法

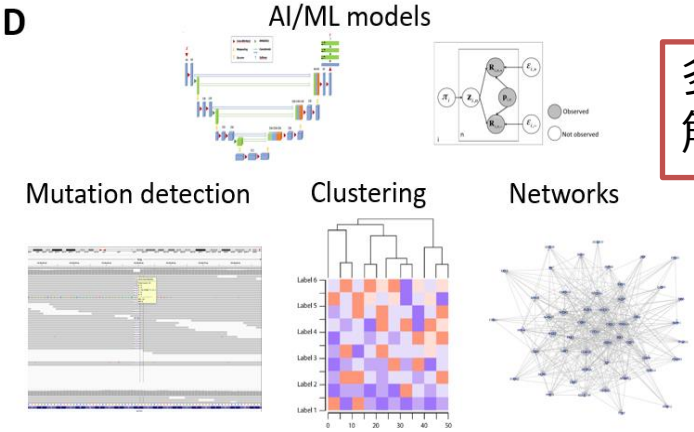




多様な
分子

多様な
サンプル

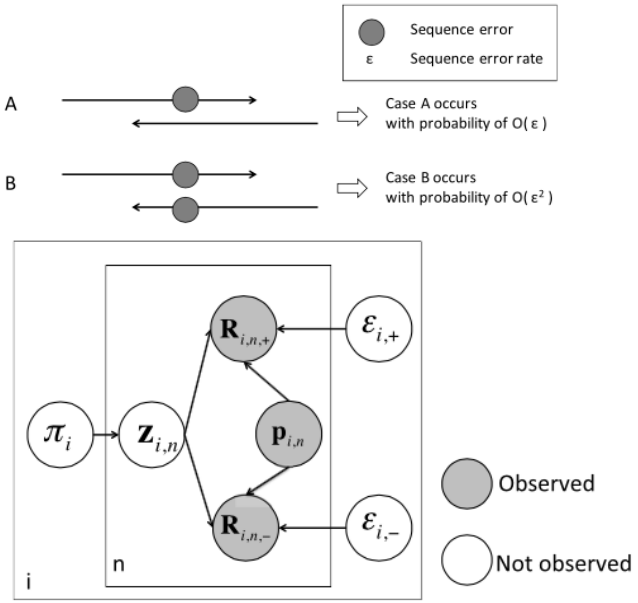




多様な
解析方法

生体分子データから有用な情報を抽出する手法の開発を行っています。

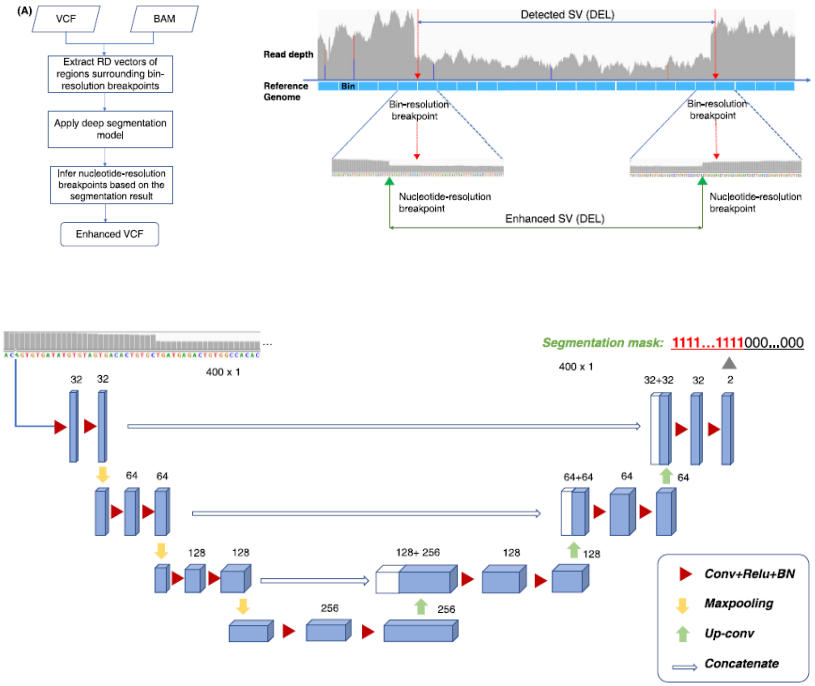
ベイズ統計モデル × 深層学習モデル (AI)



$$p(\mathcal{R}_i, \mathcal{Z}_i|\boldsymbol{\gamma}_i, \boldsymbol{\alpha}_{i,+}, \boldsymbol{\alpha}_{i,-}) = p(\pi_i|\boldsymbol{\gamma}_i)p(\epsilon_{i,+}|\boldsymbol{\alpha}_{i,+})p(\epsilon_{i,-}|\boldsymbol{\alpha}_{i,-}) \cdot \prod_n p(\boldsymbol{R}_{i,n,+}, \boldsymbol{R}_{i,n,-}|z_{i,n}, \epsilon_{\pm,i}, \boldsymbol{\pi}_{i,n}, \boldsymbol{p}_{i,n})p(z_{i,n}|\pi_i)$$

短い塩基変異の検出法

Moriyama et al., Bioinformatics, 2019

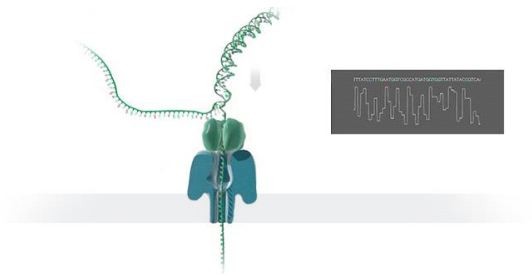


長い塩基変異の検出法

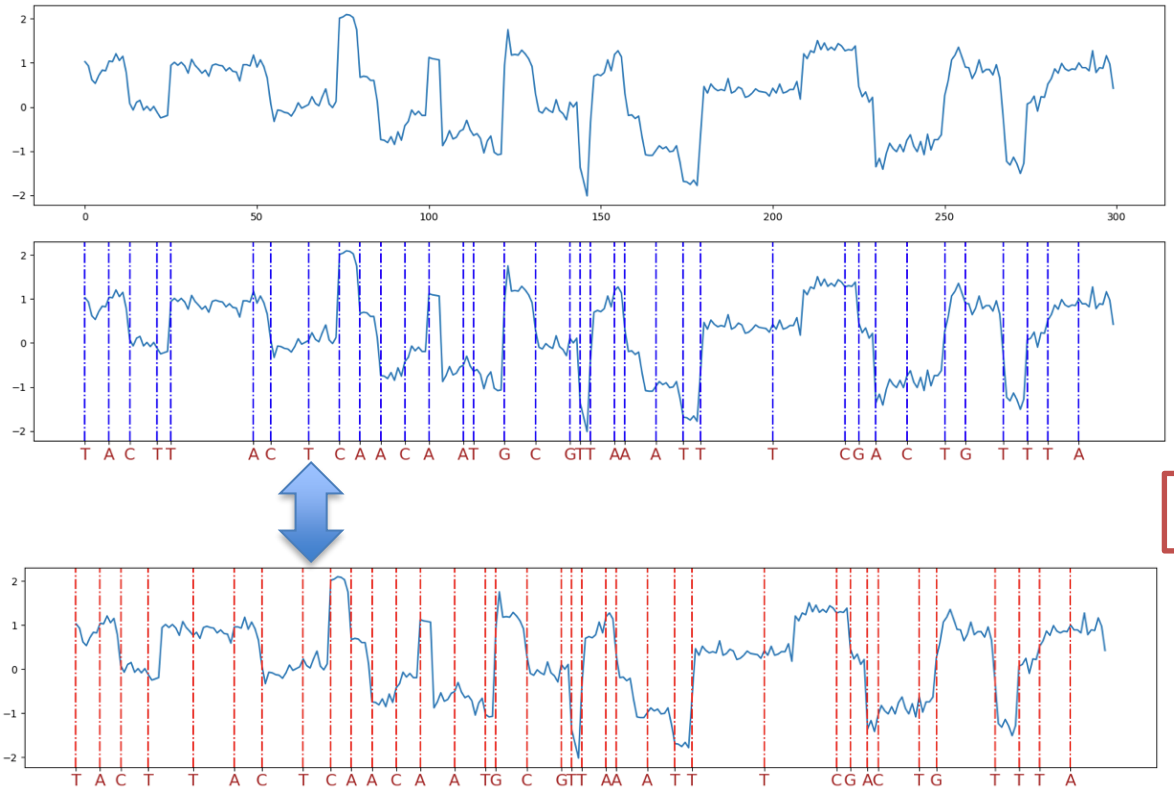
Zhang et al., PLoS Compt Biol, 2021

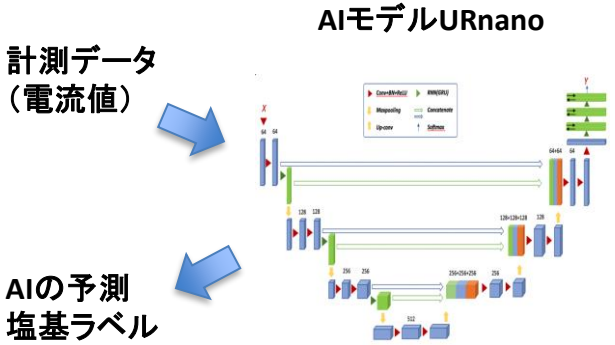
我々が開発したDNAシークエンスデータから変異を検出する手法の例を示します。
図で示されていますが、それぞれ数式で記述される数理モデルとなっています。

研究（１）：深層学習モデル(AI)によるDNA配列推定



- ナノポアシークエンサーと呼ばれる、計測装置から得られるデータの解析手法の紹介をします
- 従来の装置に比べて、長いDNA断片(1000塩基以上)を計測できる反面、まだエラーが大きい問題があります
- 計測データ（電流値）の複雑なパターンからDNAの塩基を推定するのが難しいからです
- 我々は新しい深層学習モデル(URnano)を開発しました
- AIは大量のデータを学習することで、複雑な問題を解決することができます





AIによる高精度の予測を達成

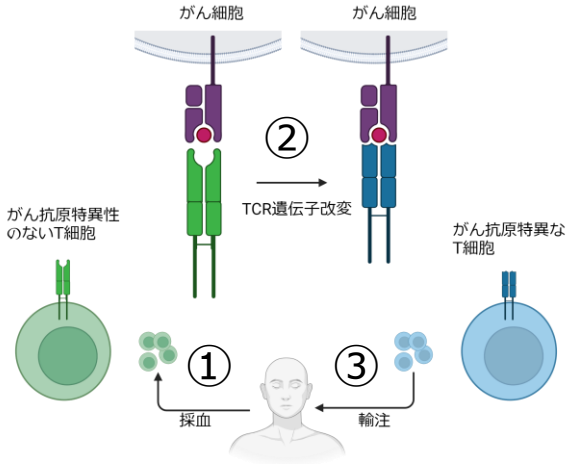
正解
塩基ラベル

Zhang et al., BMC Bioinformatics, 2019

研究（２）：タンパク質結合能予測AIモデル

改変型T細胞輸注療法

結合能を高めた受容体タンパクをAIで生成したい：
ベイズ統計を応用　結果から原因を推測



Bayesian Method for TCR Design

Aim: Search for TCR structures with higher binding affinity to pMHC

受容体タンパク質

TCR Structure

S_{TCR}

pMHC

Forward Model

$A_{TCR} = f(S_{TCR})$

$p(A_{TCR}|S_{TCR})$

結合能予測

Binding Affinity

A_{TCR}

Template-based
Modeling
Software: Rosetta

Reverse Model

$S_{TCR} = f^{-1}(A_{TCR})$

Bayes' Law

$p(S_{TCR}|A_{TCR}) \propto p(A_{TCR}|S_{TCR}) \times p(S_{TCR})$

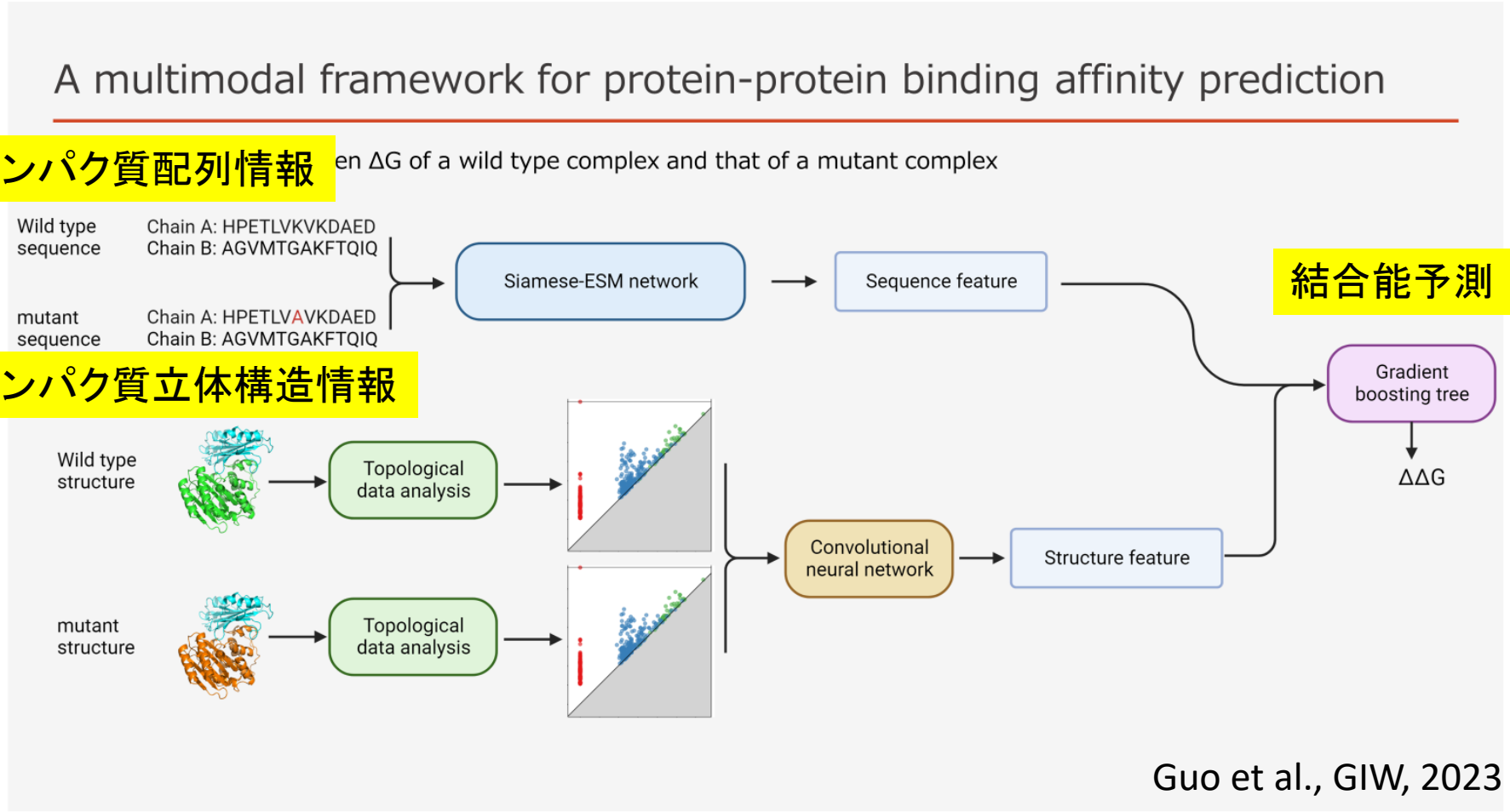
Generate and search TCR structures by Monte Carlo sampling

免疫細胞は受容体タンパクで、がん細胞表面の目印となるタンパクを認識して攻撃します。
AIを使って、目印となるタンパク質への結合能を高めたタンパク質をデザインする手法の開発を目指しています。

腫瘍免疫制御TR分野（松下ラボ）との共同研究

研究（２）：タンパク質結合能予測AIモデル

前のページで述べた結合能を高めた受容体タンパクの設計に向けて、
まずタンパク同士の結合能力を、アミノ酸配列情報と立体構造の情報から
正確に予測するAIの開発に成功しています。



Guo et al., GIW, 2023

研究（３）:情報管理・解析システムの開発

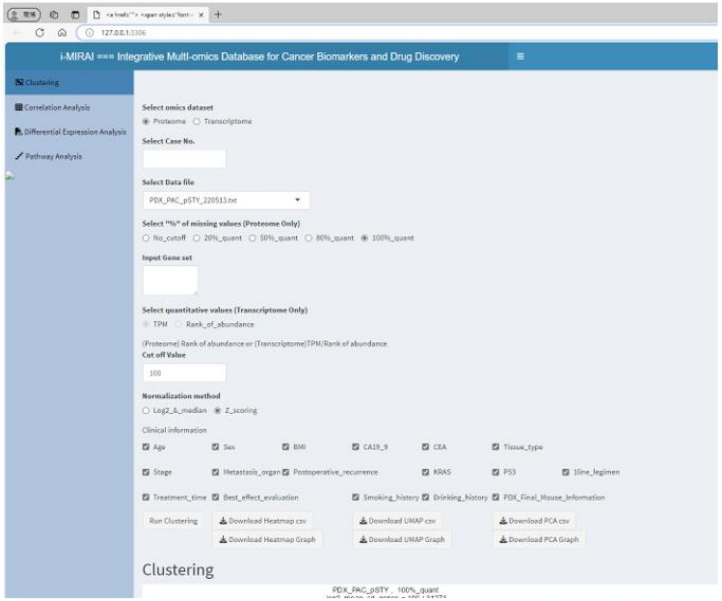
DNA/RNA/Protein

重点プロジェクトで産生される大量かつ複数種類のデータを、特別なプログラミングなどを必要とせずに統合的に解析できるシステムの開発を行っています。

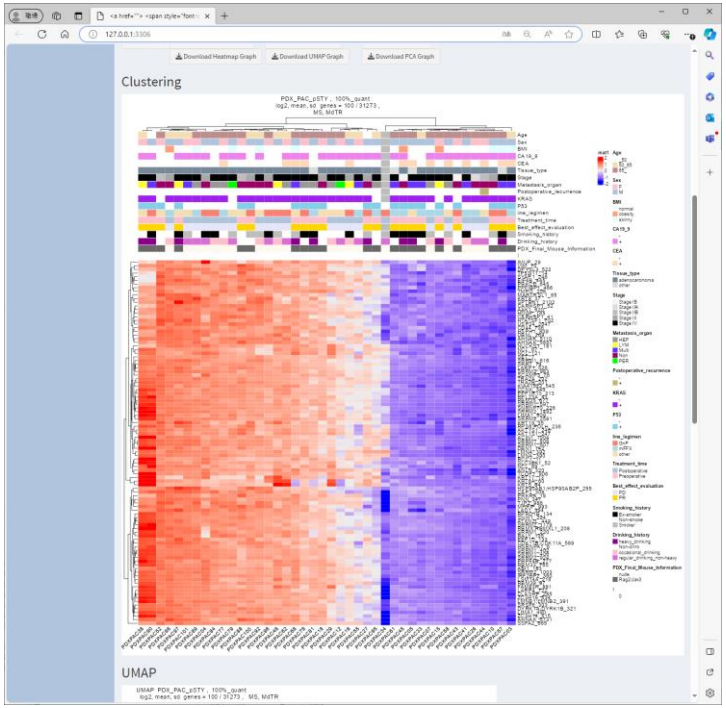
データ管理・共有・統合・再活用
Information management system

i-MIRAI

INPUT 項目↵



クラスタリング/Heatmap



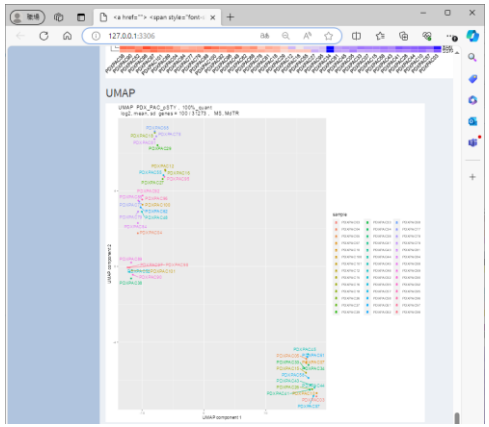
分子診断TR分野（田口ラボ）との共同研究

研究（３）:情報管理・解析システムの開発

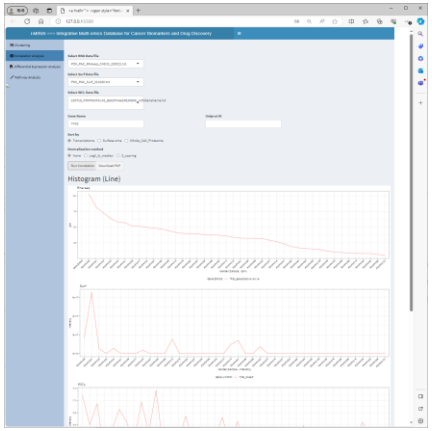
i-MIRAI

実際の研究現場では複雑なデータに対して、条件を変えながら複数種の解析を実施し、得られた多角的な情報を基に仮説を立て、次のステップへ進むことを繰り返します。
このシステムはそのプロセスを加速します。

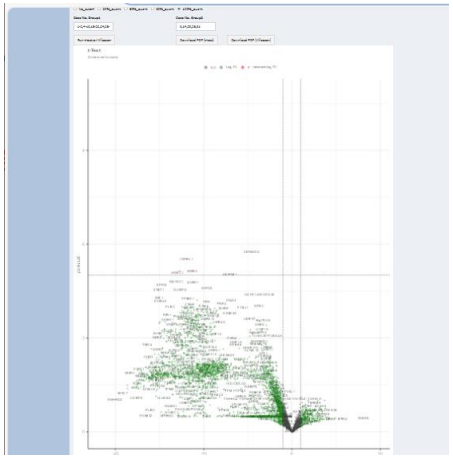
次元削減（UMAP/PCA）



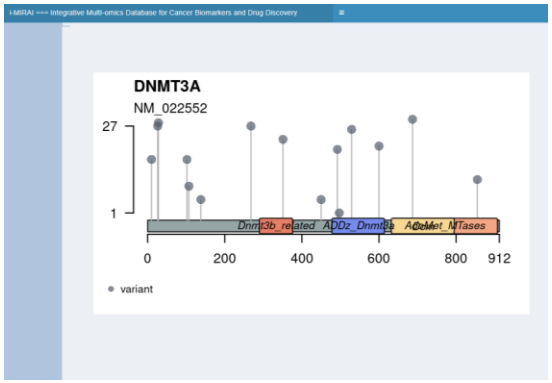
相関解析



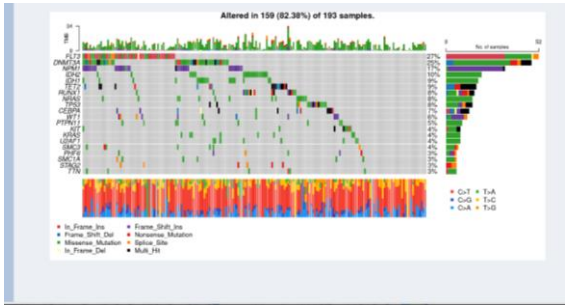
発現差解析



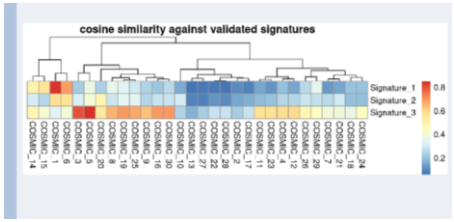
Lollipop plot



Oncoplot



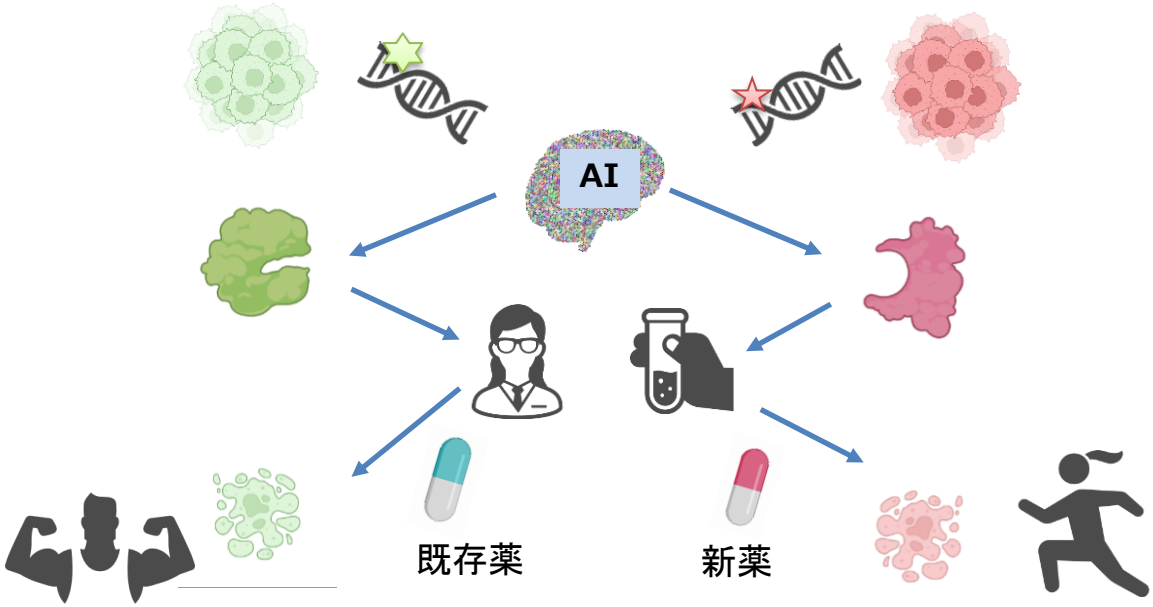
Mutational signature



関連した話題：タンパク質の立体構造がAIで予測できると

実験的に決定困難なタンパク立体構造を
これまでにない速度と労力で推測することができる

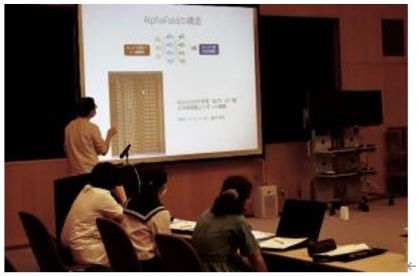
がんのメカニズムの究明や薬の開発が加速度的に進むことが期待



　タンパク質の立体構造と性質は密接に関わっています。しかし多くのタンパク質の立体構造を実験的に決めることが難しいことが課題でした。
　その課題に対して近年、タンパク質のアミノ酸配列（1次元情報）から立体構造（３次元情報）を予測するAIの開発が、世界中で急速に進んでいます。

高校生向け基礎実験体験講座

高校生がいち早く最新のAIを体験！



講義の様子１：
AlphaFold のアルゴリズムについての講義

実習の様子１：
スパコンへのログイン

実習の様子３：
AlphaFold から予測された EGFR タンパク立体構造の可視化。
野生型と変異型の構造を重ね合わせて比較。

愛知県がんセンターでは毎年8月に高校生を対象とした体験講座を実施しています。
一昨年は、将来の研究者の卵である高校生達に最先端のタンパク質立体構造予測AI（AlphaFold*）を体験してもらいました。

*: Jumper et al., Nature, 2021

おわりに

次世代シークエンサーによりがん細胞から取得される大規模計測データ（がんビッグデータ）とその多様化について紹介しました

個々のがんの性質を知り治療戦略の立案につなげるには、それらのデータから有用な情報を抽出するための手法およびシステムの開発が重要です

ここではシステム解析学分野で開発を進めている解析手法および解析システムの一端をを紹介しました

医療AIの開発と発展が進むなか、当分野では愛知県がんセンターの病院と研究所の方々と協力して、データ科学の力で医療に貢献できるよう研究を進めてまいります