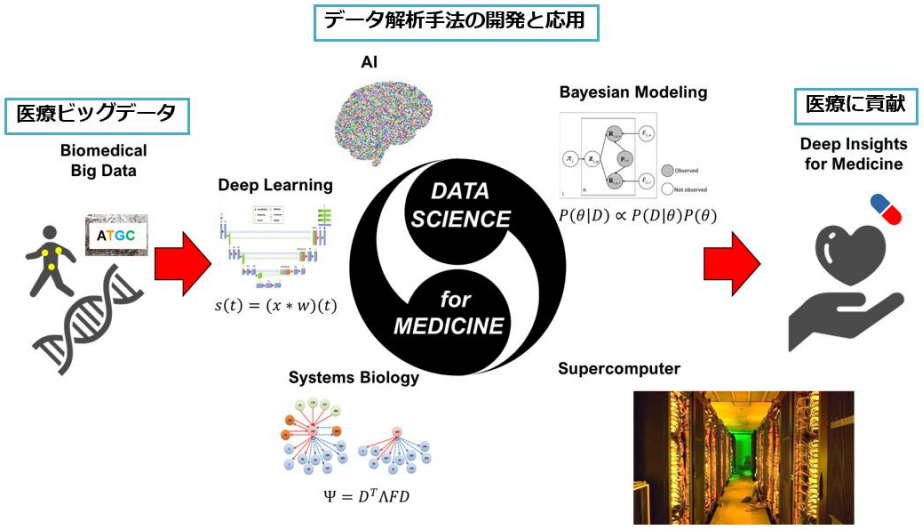
AIとスパコンを活用したがんビッグデータ解析手法の開発

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！



- 愛知県がんセンター研究所の研究室です
- **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
- 複雑なデータから**がん細胞のシステム**の特性をあぶりだし**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

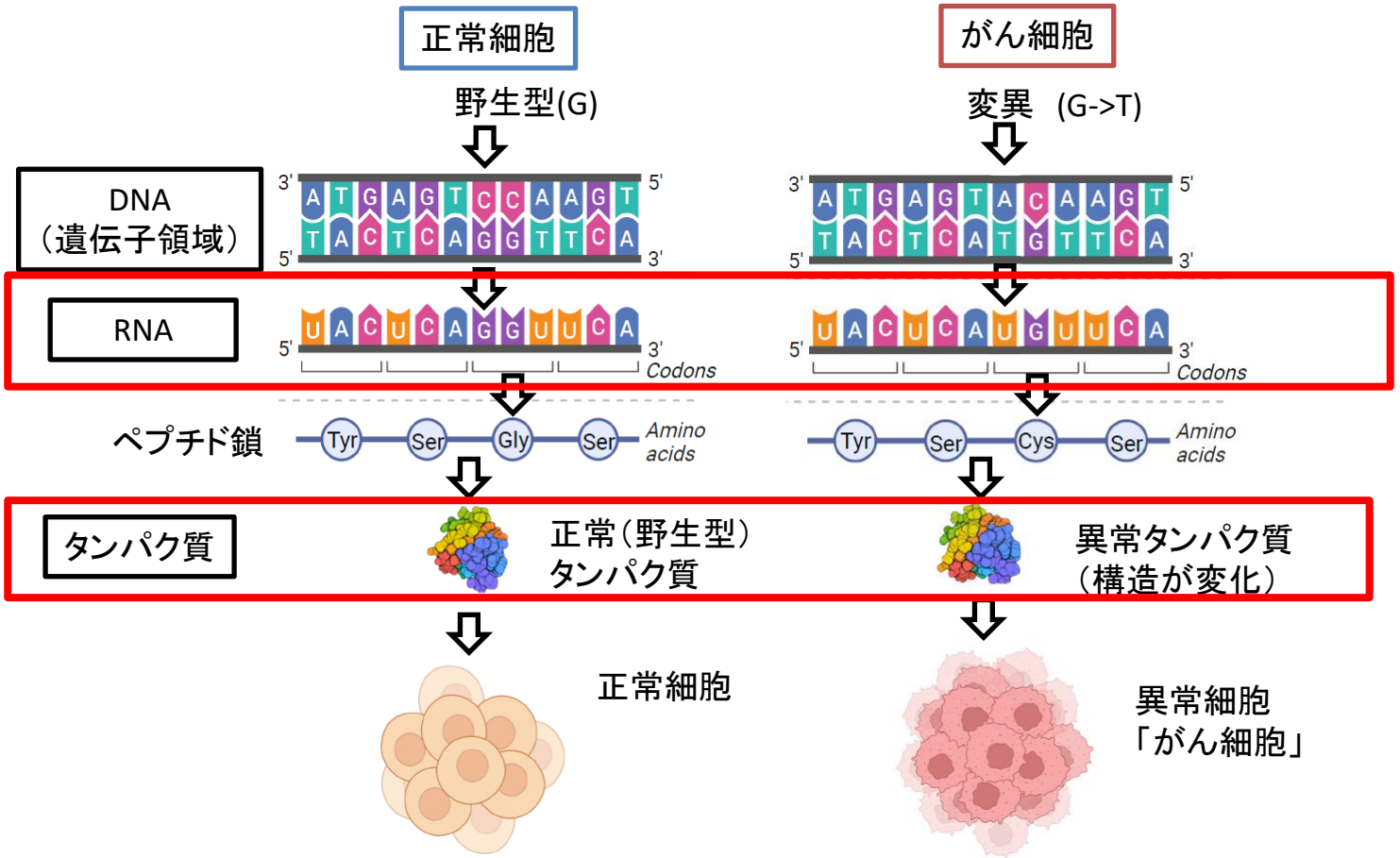
研究室HP

分野紹介YouTube

はじめに

- 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
- 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**が変化することなどより、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、がん細胞から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
- **研究１：**DNA分子から得られた生データをコンピュータと数理モデルを使い解析することで、**DNAの塩基配列を決定**し、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ことが必要です
- **研究２：**またDNA中の遺伝子領域から**RNAが作られ分解されるまでの動的なプロセス**の様子を知ることは、がん細胞の薬剤への応答の振る舞い等を理解するうえで重要です
- **研究３：**さらに変異による**タンパク質の結合能力等の性質の変化を予測**することも重要です
- **研究４：**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**簡便に複雑な解析を実行できるシステム**も必要です

遺伝子からRNAへ、RNAからタンパク質へ



細胞中の**DNA**には**遺伝子**と呼ばれる**タンパク質の設計図**が入った領域（約２万か所）があります。タンパク質が作られる時、遺伝子から**設計図のコピー**が必要な数だけ**RNA**に写し取られ（**転写**）、それを基にタンパク質が組み立てられます（**翻訳**）。

がん細胞では遺伝子の設計図が書き換わる（変異する）ことで、異常なタンパク質が作られ細胞の性質を変化させます。そのため遺伝子の変異箇所、転写されたRNAの量および、翻訳された異常タンパク質の構造や機能を知ることが重要です。

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）

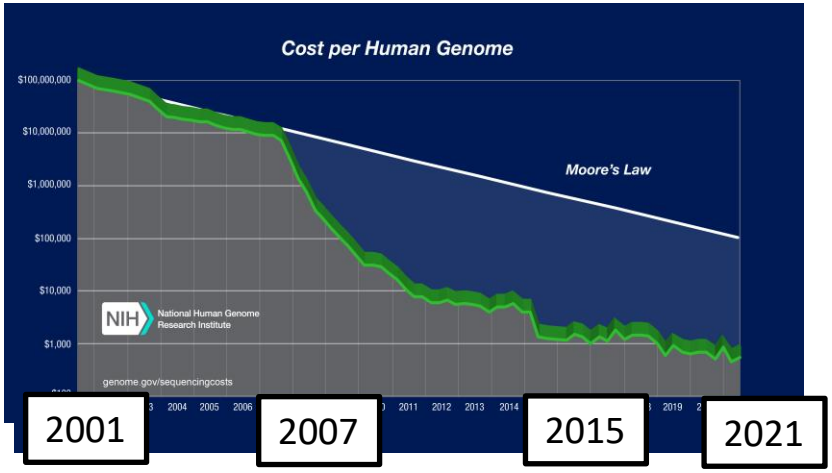
大規模ゲノム解析プロジェクト
Genomics England
All of Us
ToMMo

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？

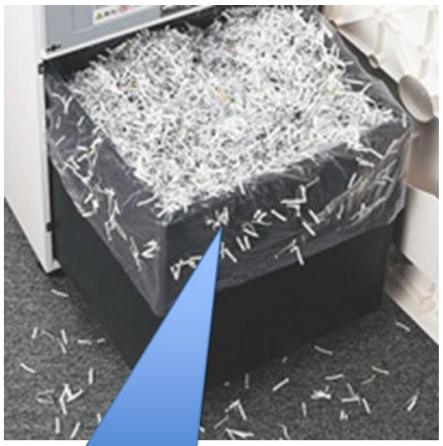
生のサンプル

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

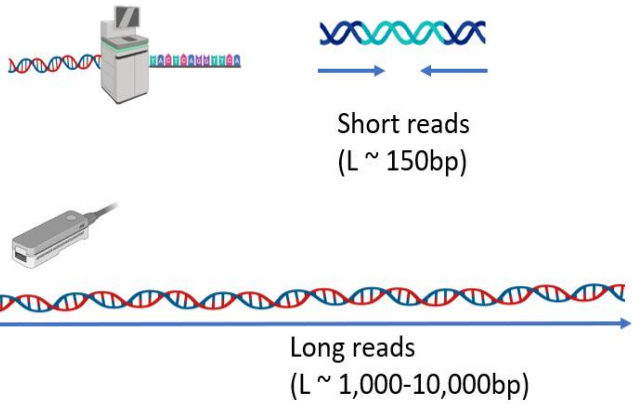
一人分のゲノムは、**4種類の塩基**（A、T、G、C）からなる塩基対、**約30億個分**からなります。現在主流のシークエンサーでは、一度にはDNA中の**約100塩基対分**しか読めません。これは**ATGCの4文字からなる30億文字分の文章が書かれた書類**の束が、シュレッダーにかけられ**100文字ずつの断片**になって出てくるようなものです。通常、がん細胞特有の変異を見つける場合、**がん細胞由来の書類を40コピー分**、**正常細胞由来の書類を30コピー分**、シークエンサーで読み取ります。つまり**21億ピースの文字列断片**が得られます。**ここから各ピースがゲノム上のどこにあ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

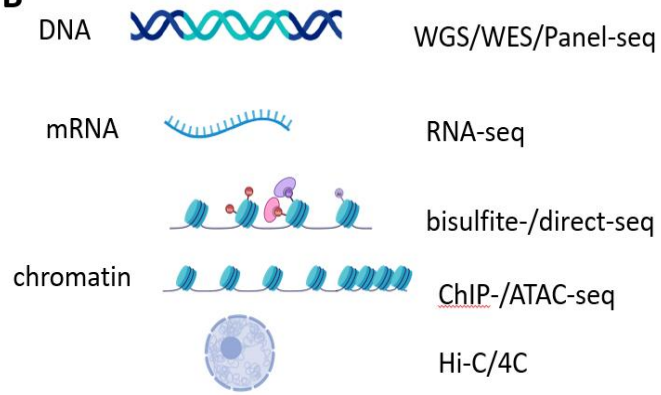
多様な
計測方法

A



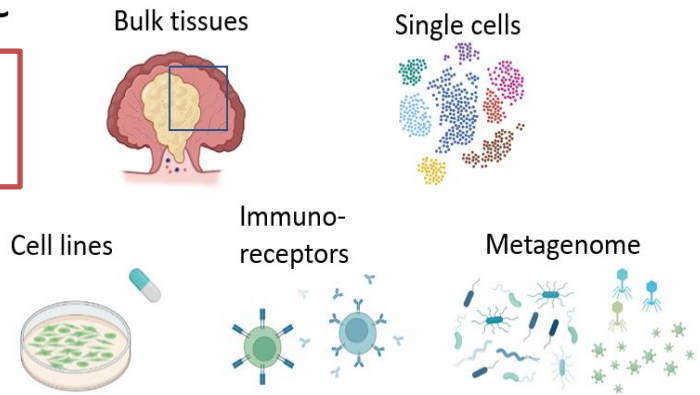
多様な
分子

B



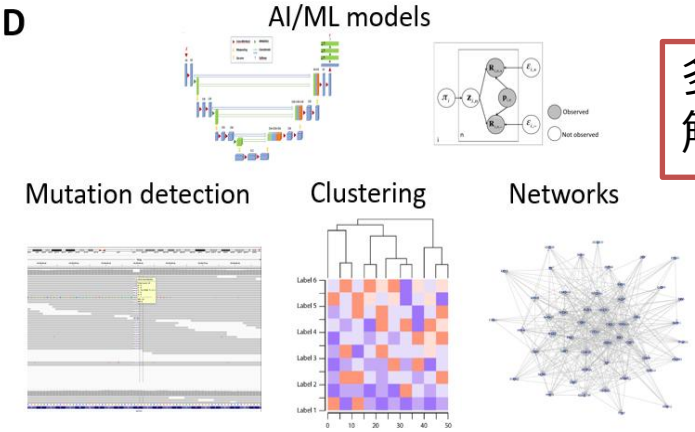
多様な
サンプル

C



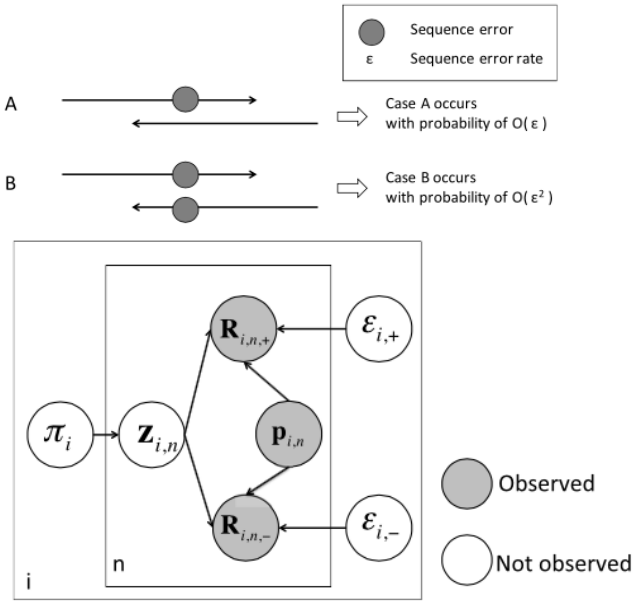
多様な
解析方法

D



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。

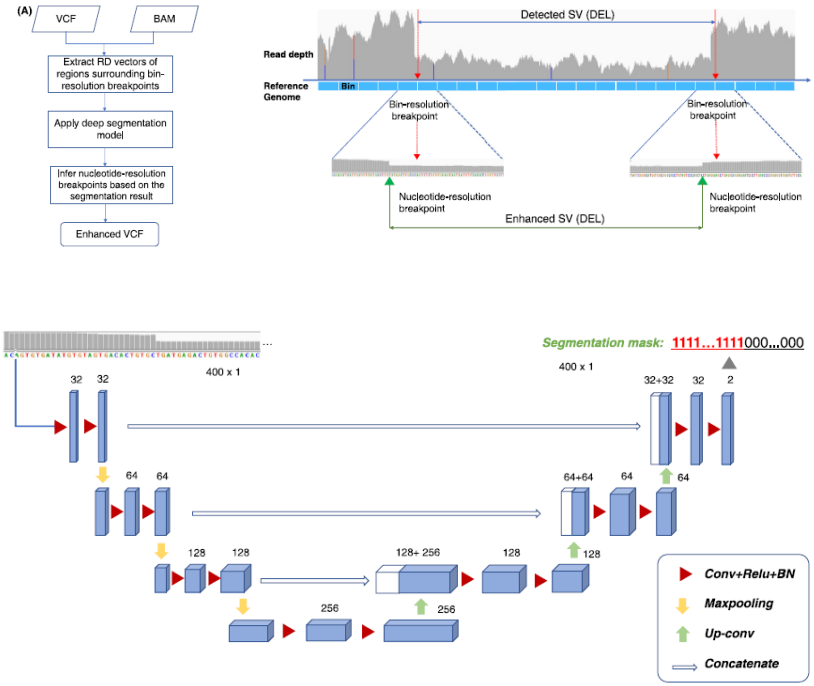
ベイズ統計モデル × 深層学習モデル (AI)



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

短い塩基変異の検出法

Moriyama et al., Bioinformatics, 2019

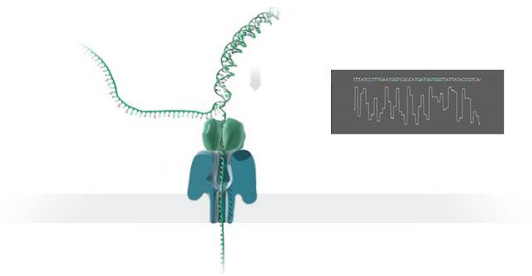


長い塩基変異の検出法

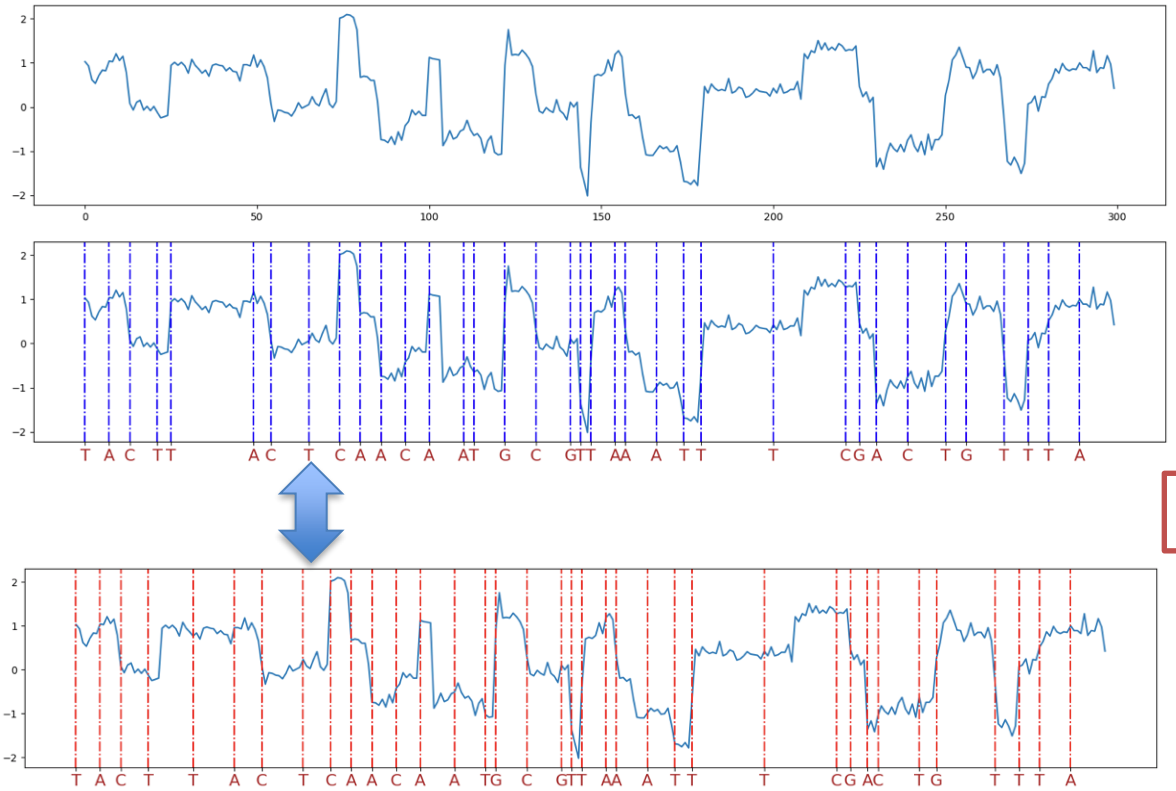
Zhang et al., PLoS Compt Biol, 2021

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）：AIによるDNA塩基配列推定

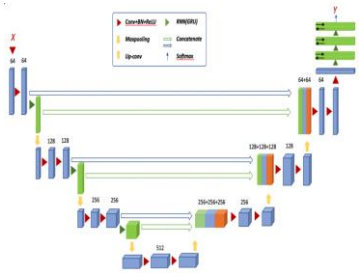


- ナノポアシークエンサーと呼ばれる、DNA計測装置から得られるデータの解析手法の紹介をします
- 従来の装置に比べて、長いDNA断片(1000塩基以上)を計測できる反面、まだエラーが大きい問題があります
- 計測データ（電流値）の複雑なパターンからDNAの塩基を推定するのが難しいからです
- 我々は新しい深層学習モデル(URnano)を開発しました
- AIは大量のデータを学習することで、複雑な問題を解決することができます



計測データ（電流値）

AIモデルURnano



AIの予測塩基ラベル

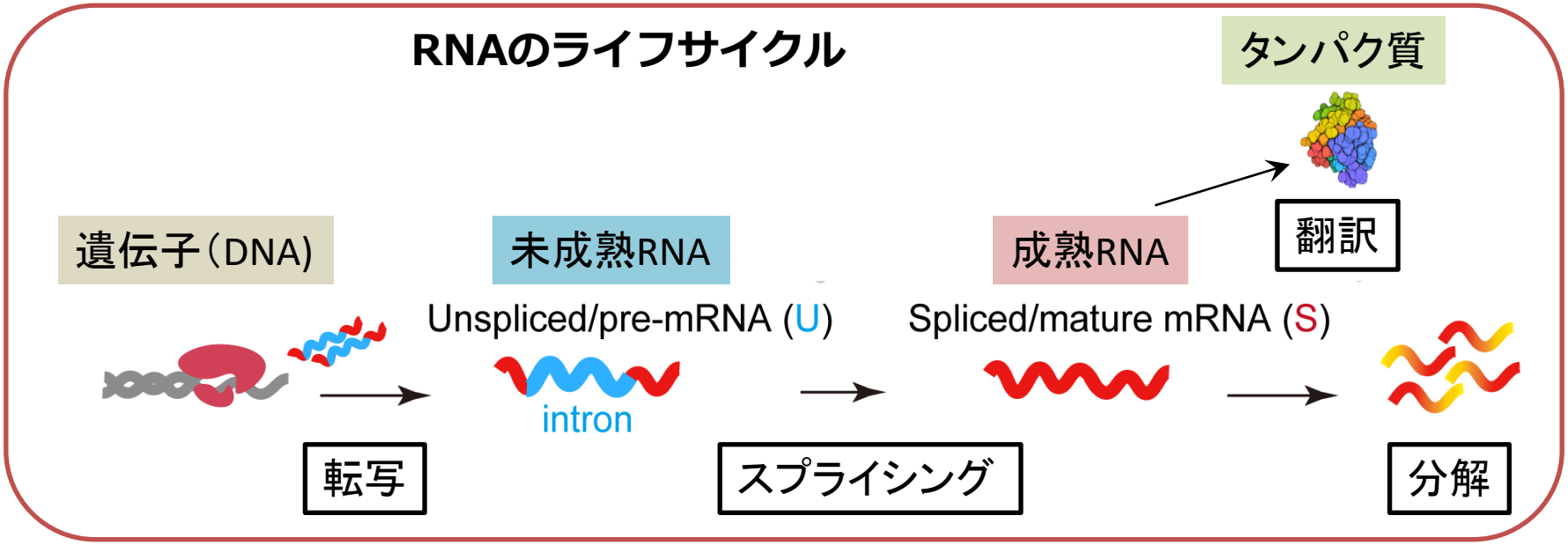
AIによる高精度の予測を達成

正解塩基ラベル

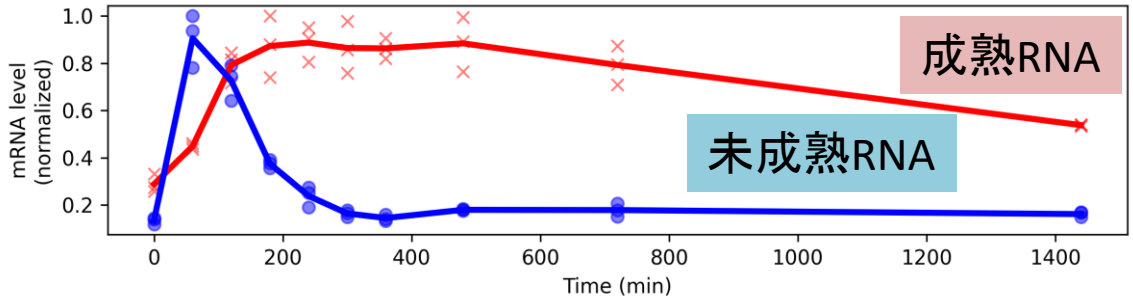
Zhang et al., BMC Bioinformatics, 2019

研究（２）：RNA量の時間変化の要因を探るAIの開発

RNAは遺伝子から転写後、スプライシングを受け、タンパク質に翻訳可能な成熟RNAになり、分解されるという複雑なライフサイクルを持っています。計測されるRNAの量は、転写、スプライシングおよび分解といったプロセスのバランスで決まりますが、それらのプロセス自体の時間変化を計測するのは容易ではありません。



計測データ：ANXA9遺伝子のRNA量の時間変化



直接計測が難しい

?

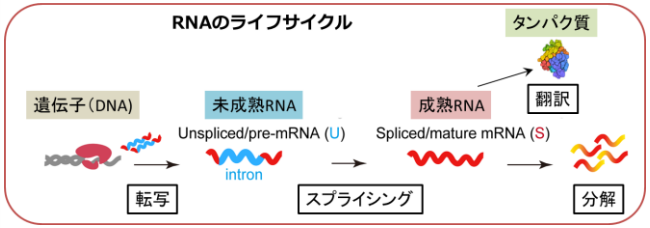
転写　分解

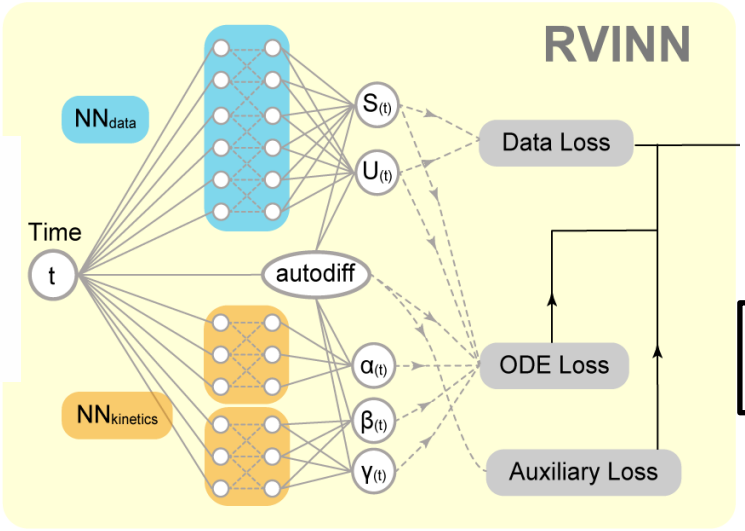
スプライシング

研究（2）：RNA量の時間変化の要因を探るAIの開発

そこでRNA量の計測データから、転写や分解といったプロセスの時間変化を推論できる、新しいAIモデル(**RVINN**)を開発しました。ここではAIにRNAのライフサイクルに関する背景知識を教えこむことで、生物学的に合理的な推論を可能としています。この方法を使って、がん細胞の薬剤への反応等の特徴を明らかにできることが期待されます。

生物学的背景知識をAIに教育

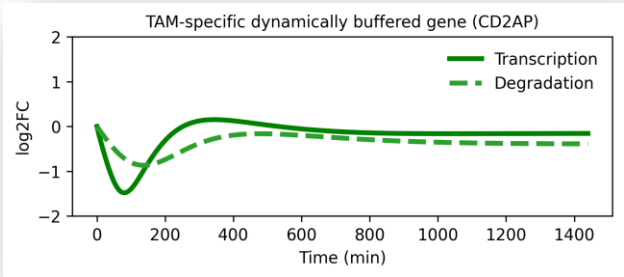




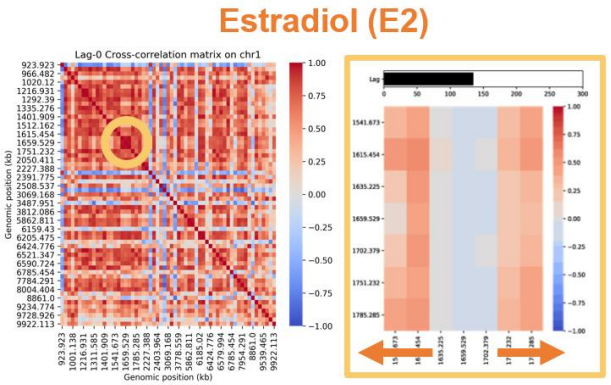
推論

RNA量変化データ

転写率・分解率の時間変化



がん細胞の薬剤応答の特徴

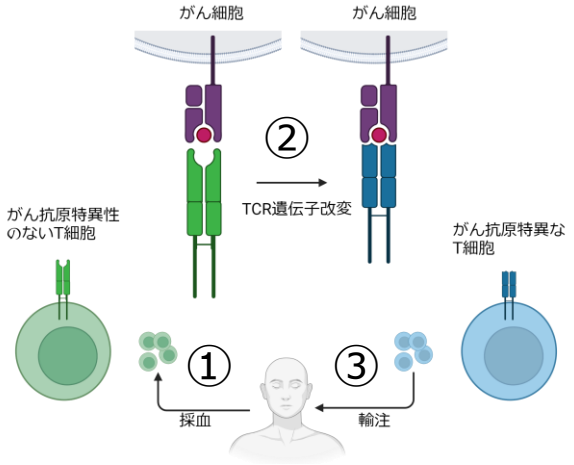


Muto et al., Bioinformatics, 2025

研究（３）：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作りたい：
AIとベイズ統計を応用　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

ward 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

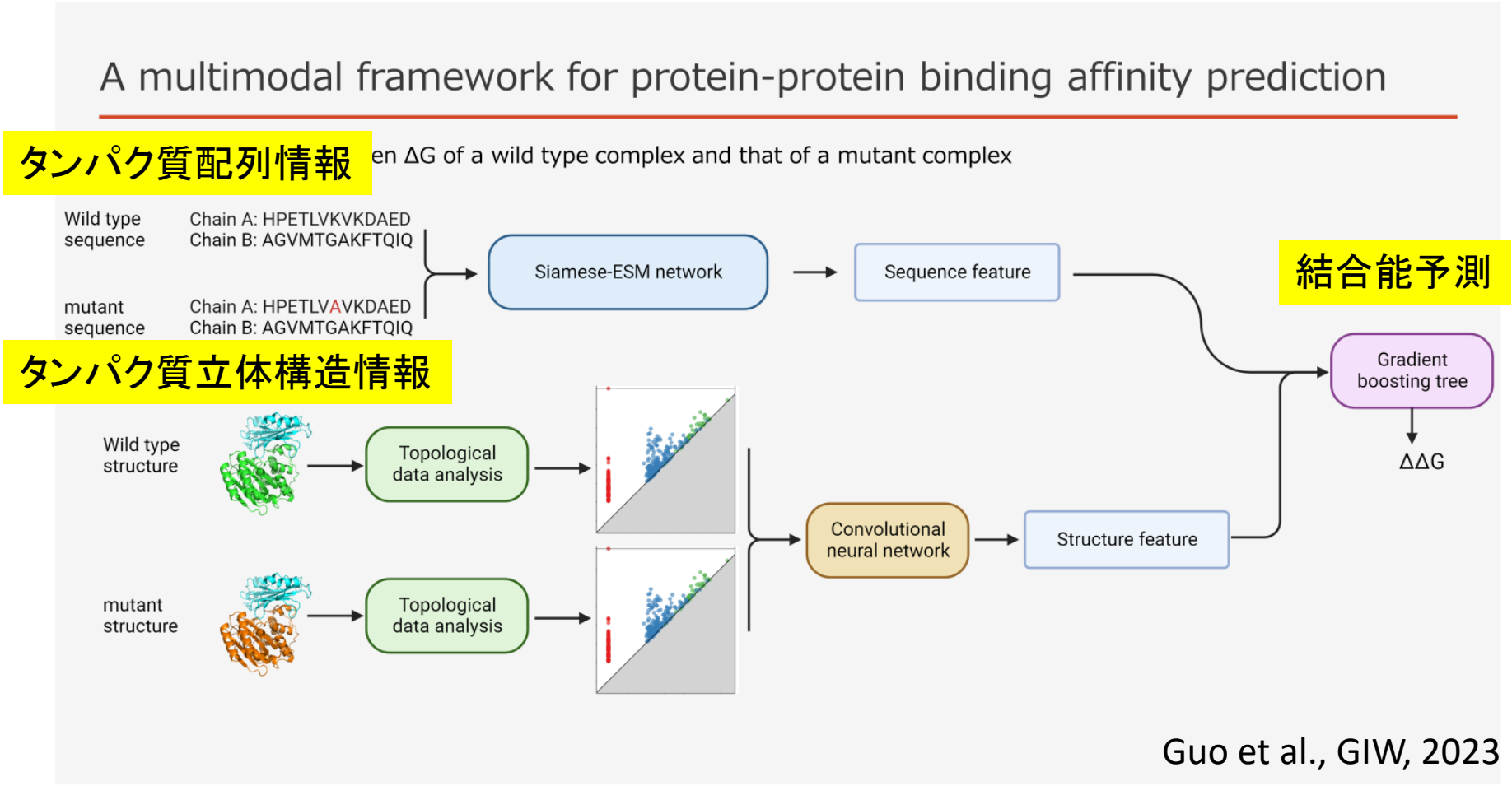
$p(S_{TCR}|A_{TCR}) \propto p(A_{TCR}|S_{TCR}) \times p(S_{TCR})$

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。

研究（３）：タンパク質結合能予測AIモデル

当分野では、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（４）:情報管理・解析システムの開発

DNA/RNA/Protein

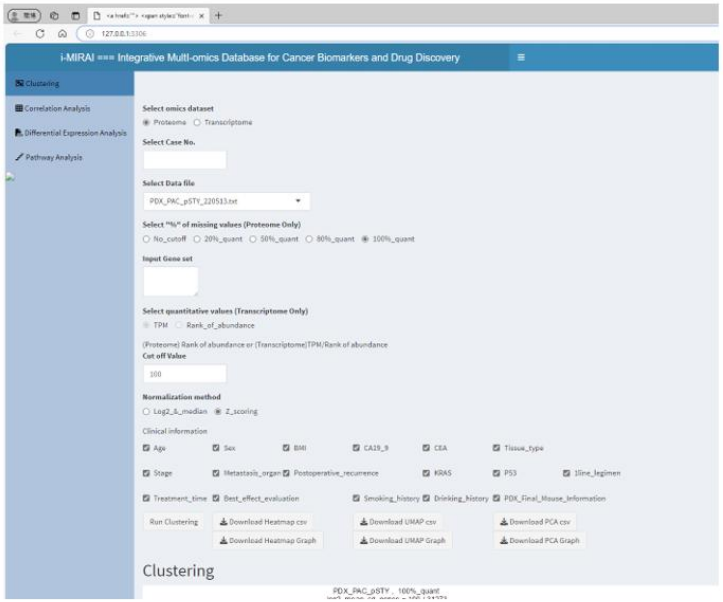
重点プロジェクトで産生される大量かつ複数種類のデータを、特別なプログラミングなどを必要とせずに統合的に解析できるシステムの開発を行っています。

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

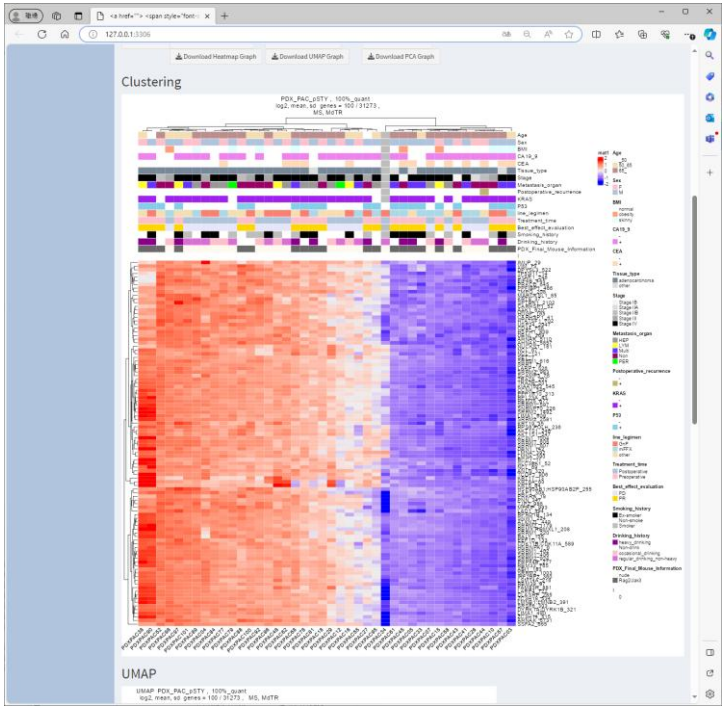
i-MIRAI

R Shiny

INPUT 項目



クラスタリング/Heatmap

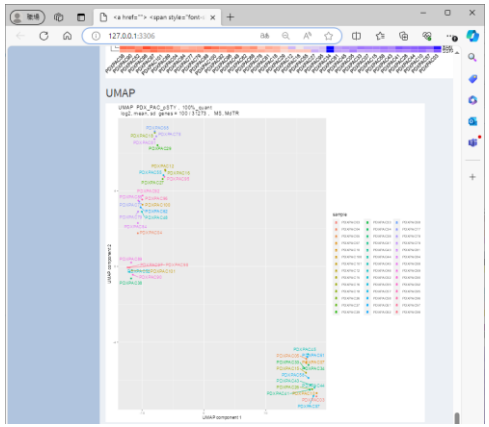


研究（４）:情報管理・解析システムの開発

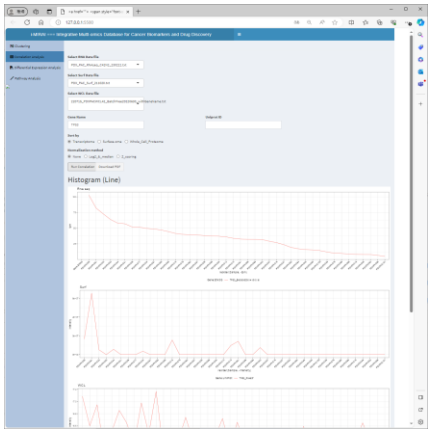
i-MIRAI

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。

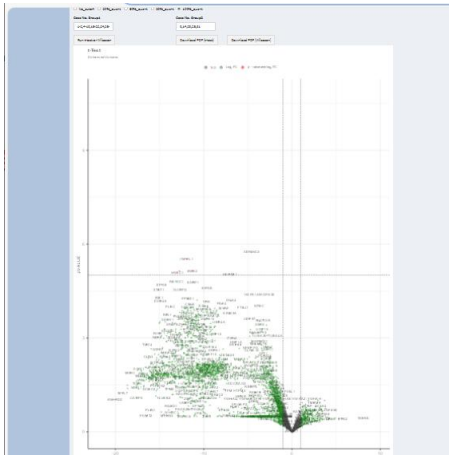
次元削減（UMAP/PCA）



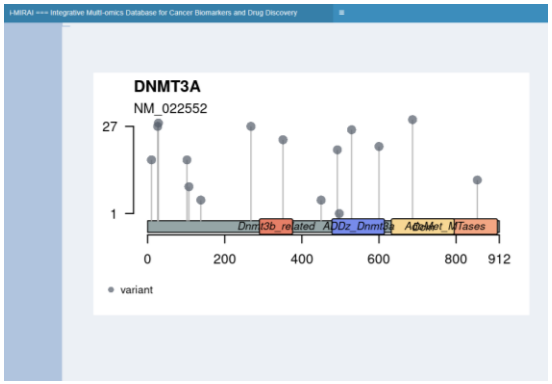
相関解析



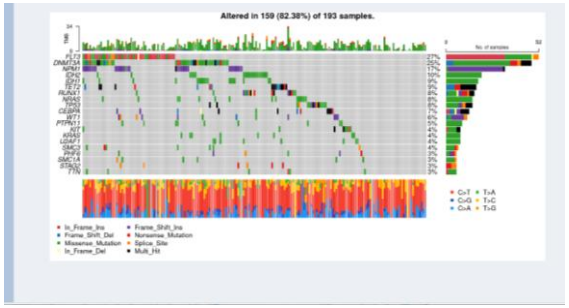
発現差解析



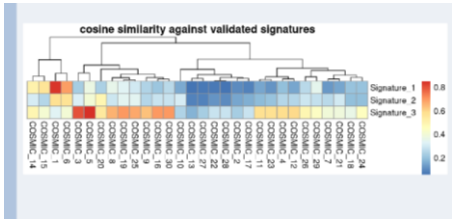
Lollipop plot



Oncoplot



Mutational signature



おわりに

次世代シークエンサーによりがん細胞から多様で大規模な計測データ（がんビッグデータ）が得られるようになっています

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、応援をいただければ幸いです