第10回ゲノム医療協議会	資料 1-3
令和４年12月23日	

第10回ゲノム医療協議会資料

AI活用の可能性とゲノムデータ基盤に対する提案

Hiroyuki Kobayashi, MD, PhD
Preferred Networks, Inc.

AI開発に特化したプリファードネットワークス(PFN)の特長

Kaggle Grandmasterなど
多数のトップ人材*1

国内トップクラス
のエンジニア集団

最適なAIモデルの構築

Green500*2世界1位
(2020年、2021年)

省電力世界No.1
のスパコン

モデル精度の向上

シェアNo.1企業とのBtoBで
国内最大のユニコーン企業*3に

業界トップ企業と
の豊富な協業経験

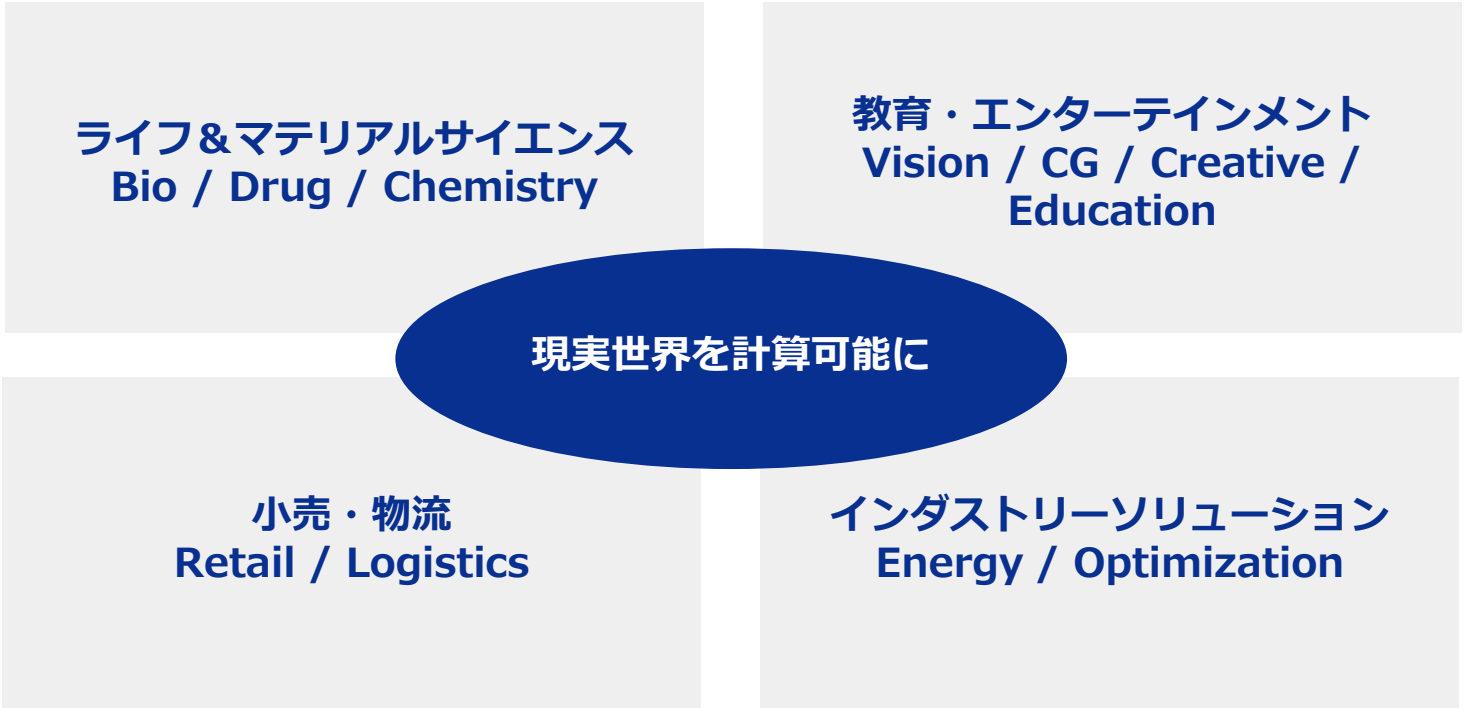
適切な作業仮説の立案

＊1 日本に十数名しかいないKaggle Garndmasterが3名在籍、同様に数十名しかいないKaggle Masterが11名在籍（2021年時点）。
＊2 TOP500ランキングの各ベンチマーク値をその消費電力で割った値による、電力効率の良い高性能計算の実現を評価するランキング。
＊3 STARTUP DB「国内スタートアップ評価額ランキング最新版（2022年3月）」で時価総額第一位（3,549億円）。

Preferred Networks

ライフサイエンス事業は注力領域の一つ

PFNが持つAIに関する豊富なケイパビリティを組み合わせることで幅広い事業領域に対応



Preferred Networks

ゲノムデータ基盤活用におけるAI開発の観点

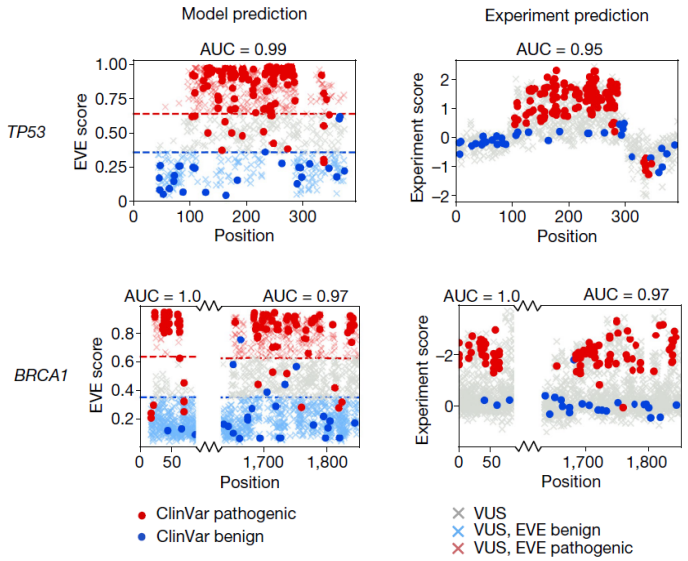
AI企業としての貢献	● 中長期的に将来の医療への応用を目指す「医学的貢献」 ● 短期的に現時点で技術を必要としている患者・家族・医療者に対する「医療的貢献」
具体的なユースケース	● 新規リスク遺伝子など探索的な発見 ● ゲノムドリブンの創薬プロセスへの応用 ● 診断精度・効率性の向上（早期診断・個別化予防等） ● 治療予後の予測精度の向上（薬剤反応性等）
現状の課題	● 予測モデルの精度 ● 一般化の可能性 ● リスク遺伝子の信頼性 ● データセキュリティ

バリアントの病原性を連続的なスコアで定量化する

進化に伴う配列分布をモデル化することで、遺伝子変異の病原性（疾病発生リスク）の定量化に成功

- Frazer J, et al. Nature 2021

- これまでに同定されてきた疾患関連遺伝子におけるタンパク質変異の98%以上は、その病原性（疾患を引き起こす特性）が不明だった
- 種間で繰り返されるアミノ酸配列は生物学的重要性のマーカーであり、そのような高度に保存された配列の変更は、問題を引き起こす可能性が高く、病原性に関連していると考えられる
- 従来の「既知疾患ラベルによる機械学習モデルの学習」を行うのではなく、生物間の膨大な時間の中での配列変異分布をモデル化することで、遺伝子変異から生じる疾病リスクについて連続性を考慮した高精度な予測が可能となった



「EVE」（evolutionary model of variant effect）と名付けられたアルゴリズムによる病原性のスコアリング（左図）と実際の実験データに基づく遺伝子変異の病原性（右図）の比較。TP53やBRCA1といった良く知られるがん抑制遺伝子における各単一アミノ酸変異体の病原性をEVEはきわめて高精度に予測している。

遺伝子変異の病原性を定量的に評価することによって、疾病リスクをより正確に捉えることが可能になる

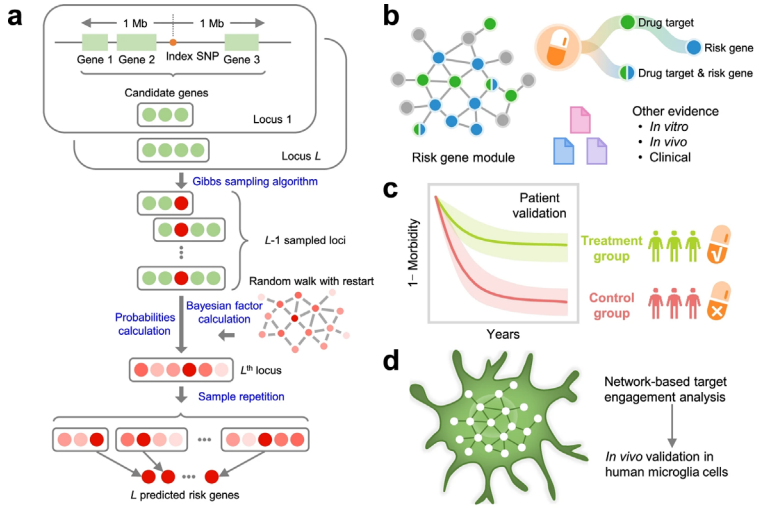
Preferred Networks

機械学習の活用による新たな創薬ターゲットの発見

機械学習を活用したゲノム解析手法によってアルツハイマーリスク遺伝子を特定し、新たな創薬ターゲットを発見

- Fang J et al. Alzheimer's Res & Ther 2022

- 従来のGWASの知見から生み出された疾患遺伝子の多くに治療標的としての意義は見いだせなかった
- 一方、最近GWAS遺伝子座のリスク遺伝子を推定するベイズ型フレームワーク、integrative risk gene selector (iRIGS)が開発されいくつかの疾患で従来より高精度のリスク遺伝子が同定されている
- アルツハイマー型認知症でiRIGSを用いた解析によって得られたリスク遺伝子の多くが既知のドラッガブルタンパク質をコードしており、既存薬剤の再利用や新たな創薬候補に対する示唆をもたらすものであった



aアルツハイマー病 (AD) リスク遺伝子を特定するための、ネットワークベースのベイジアンアルゴリズムのフレームワーク 。マルチオミクス データと遺伝子ネットワークを統合してGWAS遺伝子座からリスク遺伝子を推測。b アルツハイマーリスク遺伝子とタンパク質間相互作用をベースとした創薬標的ネットワークの構築。c標的薬剤使用者と AD の結果との関係をテストするための集団ベースの検証。dヒトミクログリア細胞におけるネットワーク予測薬の提案された作用機序の実験的検証。

機械学習を用いてリスク遺伝子を絞り込むことで、新たな治療標的を見出すことが可能

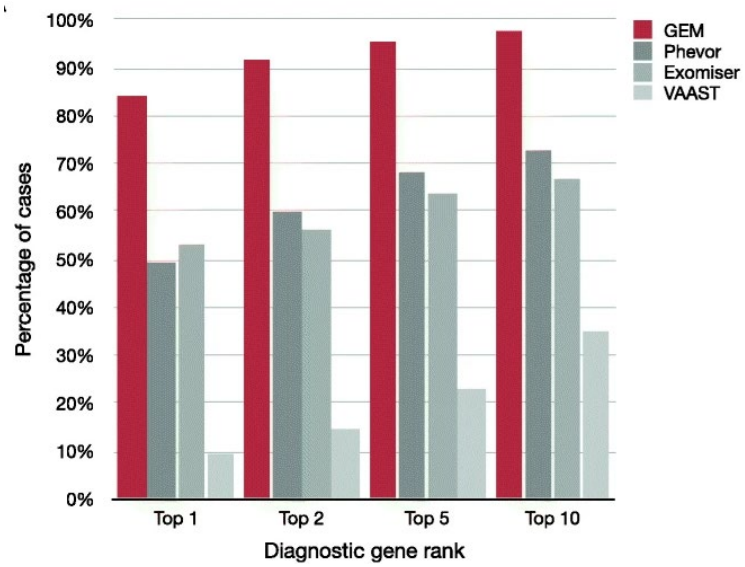
Preferred Networks

全ゲノム配列から高精度かつ迅速な診断を可能にする

大規模なゲノムデータベースを参照しAIを活用することで、新生児の遺伝性疾患を迅速に特定することが可能

- De La Vega FM, et al. Genome Medicine 2021

- 深刻な遺伝性疾患を抱えて生まれてくる新生児は全世界で毎年約700万人にのぼる
- 生後48時間以内の診断が転帰を左右するが、現状の遺伝子配列の解析には数時間、さらに特定の疾患診断のための手動分析に数日から数週間かかる
- 多様な集団のゲノム配列に関する大規模なデータベースと臨床情報などから、米Fabric Genomics社のAIアルゴリズムを用いることで、疾患の原因となる遺伝子エラーの上位2種類のいずれかを92%の精度で迅速に特定することができた（右図）



ベンチマークコホート119例のうち、上位1位、2位、5位、10位の遺伝子候補の中から真の原因遺伝子が特定された割合。米Fabric Genomics社のAIアルゴリズムであるGEMの診断感度は、既存のバリアント優先順位決定法であるPhevor、Exomiser、VAASTよりも明らかに優れていた。

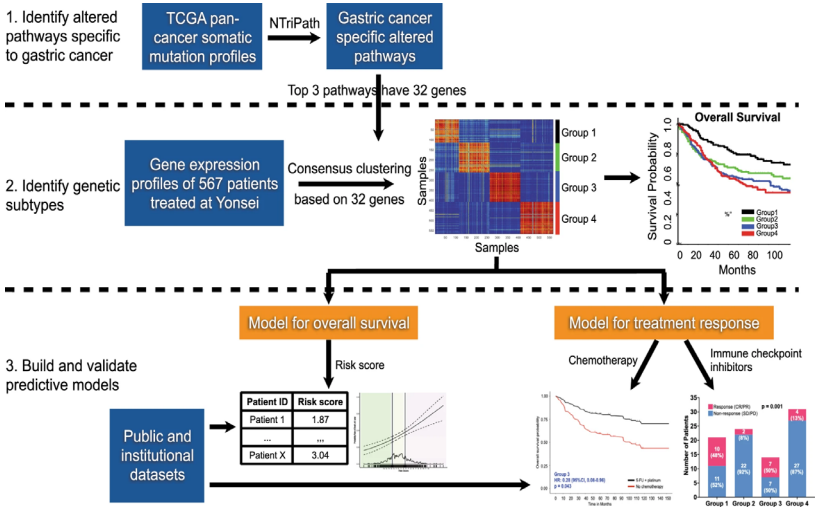
希少バリアントに関する十分な情報を随時参照することができれば、 AIを活用することで迅速かつ正確な診断が可能になる

Preferred Networks

遺伝子発現シグネチャーのパターンから治療反応性を予測する

ゲノムシーケンスとAIによって、胃がん患者が治療に対してどのように反応するかを高精度に予測できる
- Cheong JH, et al. Nature Communications 2022

- 胃がん治療において化学療法や免疫チェックポイント阻害薬に対する患者の治療反応を正確に予測するバイオマーカーは十分でない
- 米メイヨークリニックの研究チームは、独自の機械学習アルゴリズムNTriPathを用い、胃がんに特有となる32の遺伝子からなるシグネチャーを同定した（右図１）
- その後567人の患者において、これら32遺伝子の発現レベルに対する教師なしクラスタリングによって、予後を決定する4つの分子サブタイプを導出した（右図２）



3左．線形カーネルを用いたサポートベクターマシンを構築し、5年全生存の予後を示すリスクスコアを生成し、3つの独立したデータセットを用いてこのリスクスコアを検証している。
３右．得られた分子サブタイプは、胃切除後の5-FUとプラチナ製剤の補助療法や、転移・再発した患者における免疫チェックポイント阻害剤への反応を予測することも明らかにされている。

治療転帰の情報を備えたゲノムシーケンスデータがあれば、予後予測をより正確かつ効率的に行うことができる

Preferred Networks

ゲノムデータ基盤活用におけるAI開発の観点

AI企業としての貢献	● 中長期的に将来の医療への応用を目指す「医学的貢献」 ● 短期的に現時点で技術を必要としている患者・家族・医療者に対する「医療的貢献」
具体的なユースケース	● 新規リスク遺伝子など探索的な発見 ● ゲノムドリブンの創薬プロセスへの応用 ● 診断精度・効率性の向上（早期診断・個別化予防等） ● 治療予後の予測精度の向上（薬剤反応性等）
現状の課題	● 予測モデルの精度 ● 一般化の可能性 ● リスク遺伝子の信頼性 ● データセキュリティ

ユーザビリティの観点からの提案

- □ **サンプルの多様性が一定レベル担保されている**
  - ➢ データの偏りはモデルの精度や一般化可能性に大きく影響を与えるため、なるべく年齢・性別・社会経済的因子*1の面などで多様性が担保されていることが望ましい
- □ **人種間の違いが可視化されている**
  - ➢ 現在利用可能な多くのバイオバンクコホートと日本人データの相違や相同を明瞭化することで、人種非依存的なモデルの開発が可能になる*2
- □ **検索性が高いUIが実装されている**
  - ➢ データサイエンティストやAIエンジニアに与えられた多くのタスクはターゲットが明確であり、UIの機能として検索性が高いことが優先される*3
- □ **チュートリアルが充実している**
  - ➢ 現状のゲノムデータベースに対してはどのように扱ってよいかわからないという声が多く、活用方法などをわかりやすく示したチュートリアルなどが揃っていると活用しやすい

＊1 教育歴や年収、婚姻歴や同居家族の有無など疾病の罹患に一定程度の影響を及ぼすと考えられる背景因子。
＊2 多くの疾患では人種や民族によって有病率が異なるため、人種の違いによる影響を考慮することで、バリアントが疾病の罹患リスクとどの程度関連しているのかをより正確に評価することができると考えられる。
＊3 アクセッションナンバーや遺伝子名などのキーワードを用いた、主にメタデータの検索を想定。

他業種間連携によるゲノムデータベース活用の拡大

AI企業としてこれまで経験した課題

- バイオバンクとの共同研究ではデータベース基盤開発が主で、実用的な目的が十分に明確ではなかった
- 製薬企業との創薬プロジェクトでも、新たなゲノムデータの活用は限定的で、従来の疾患メカニズムに依拠した分子生成モデルにとどまっていた

異業種タスクフォースの提案

- 業種を超えたタスクフォースを組成し、AI×ゲノムの先進事例に実験的に取り組む
- 新たなリスク遺伝子の探索とそれに基づく創薬プロジェクトを立ち上げ、具体的な課題を可視化する
- これらのプロセスを見える化し、アーカイブ/チュートリアルとしてオープンにすることで、同様の取り組みへのチャレンジを広く促すことができる

Preferred Networks

Preferred Networks

Making the real world computable