

アレルギー予測データベースの構築に向けたタンパク質言語モデルの開発と予測根拠の可視化

近藤一成（昭和女子大, 国立衛研）, 橋本莉奈, 白石萌花, 立石ゆう, 永瀬優菜, 成田桜, 今井良南, 春山しおり（昭和女子大）

Objectives

ゲノム編集技術や細胞性食品などの新しい技術は持続可能な食料供給に大きく寄与できる一方で、食経験が乏しい生物種に由来するタンパク質、改変されたアミノ酸配列、異種宿主で生産されたタンパク質などが新たに食品としてヒトが摂取する機会が増大する。したがって、新規・改変タンパク質の安全性、特にアレルギー性を科学的に評価する手法の確立は、これら技術の社会実装に不可欠なテーマである。

現在、タンパク質言語モデルを用いたアレルギー予測研究はこれまでの性能追求から、その判断根拠の明確化、説明可能性にシフトしている。しかしながら、タンパク質言語モデルが何を理解してどう判断しているかをヒトが理解できる形で可視化することは用意ではなく、確立した方法は存在しない。

そこで、汎用されている可視化の方法やタンパク質言語モデルの内部特徴の抽出などを用いて検討を行い、アレルギー予測の根拠を明確化し、その成果を最終的にデータベース構築に繋げていく。

Methods

1. 食物アレルギー性の予測
- アレルギータンパク質アミノ酸配列は、COMPARE、AllergenOnlineより入手して、配列確認、短い配列や重複配列除去、シグナル配列除去などを行い作成した。非アレルギーのアミノ酸配列は、Uniprotから各食品種（小麦など）のアミノ酸配列を取得後アレルギーを削除して作成した。予測判定には、protT5とESM2をconcat融合させたモデルを作製したものを利用した。
2. 食物アレルギー性予測の可視化
- 可視化は3つの方法で行った。
1. IG (Integrated Gradient、統合勾配)
2. ISM (in silico mutagenesis、インシリコでのアミノ酸置換)
3. Occlusion

Results

① アレルギー予測システムの構成

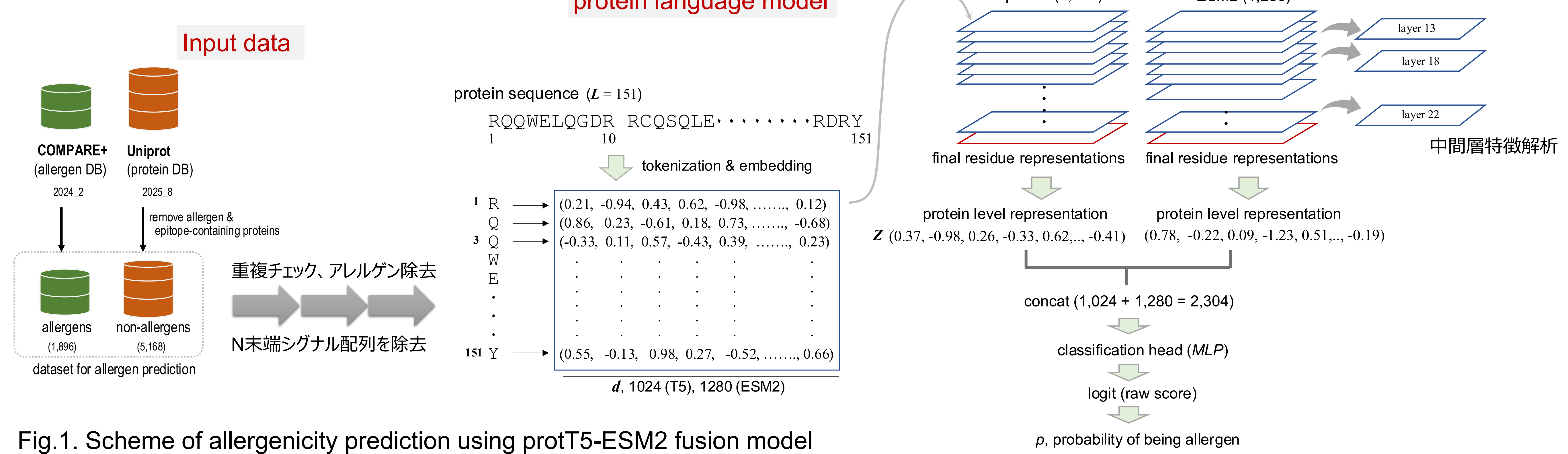


Fig.1. Scheme of allergenicity prediction using protT5-ESM2 fusion model

② 汎化性能の評価

タンパク質言語モデルは、protBERT, protT5、protT5-ESM2モデルを比較検討した結果、protT5-ESM2が最も汎化性能に優れていた。

汎化性能評価は、評価する食品種（ex.リンゴ）の全データを学習から除外して、リンゴデータでアレルギー性を予測してMCC, F1で性能評価した。

論文等で汎用されるrandom 5-fold validationでは学習側と評価側に類似したデータが分散されるため、見かけ上の性能がMCC 0.980とかになるが単なるデータリークである。そのため、LOOCVを行った。

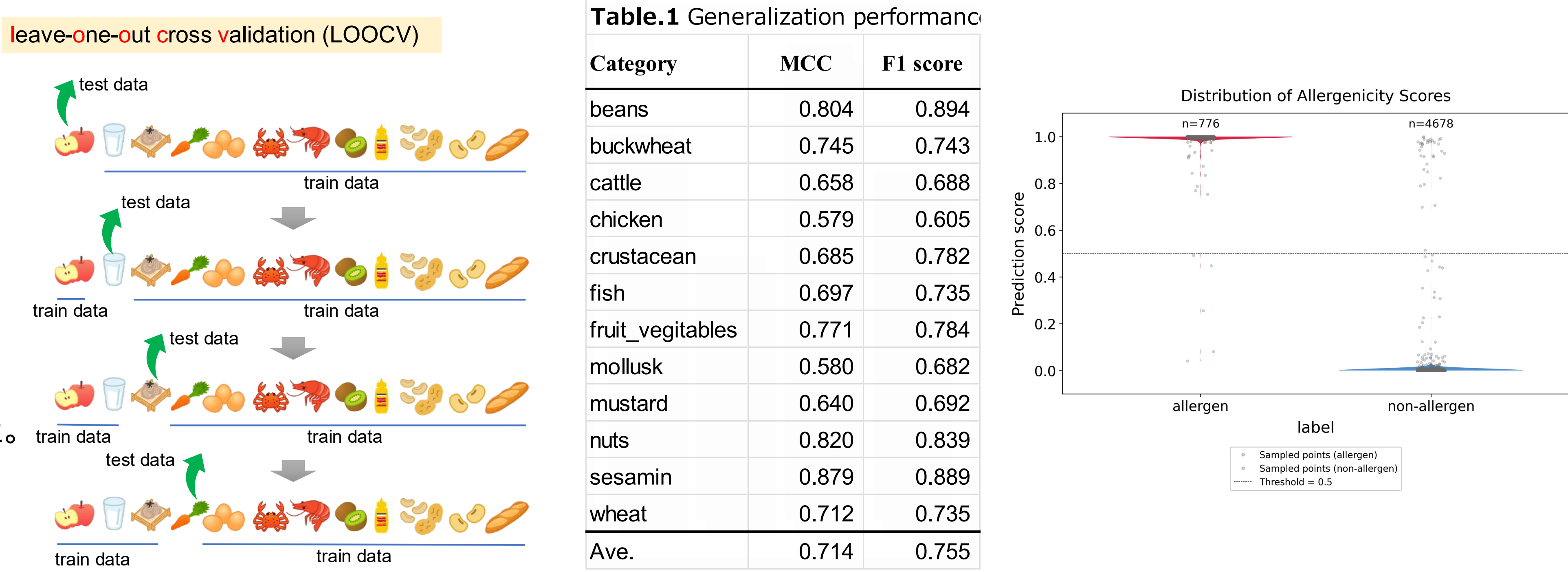
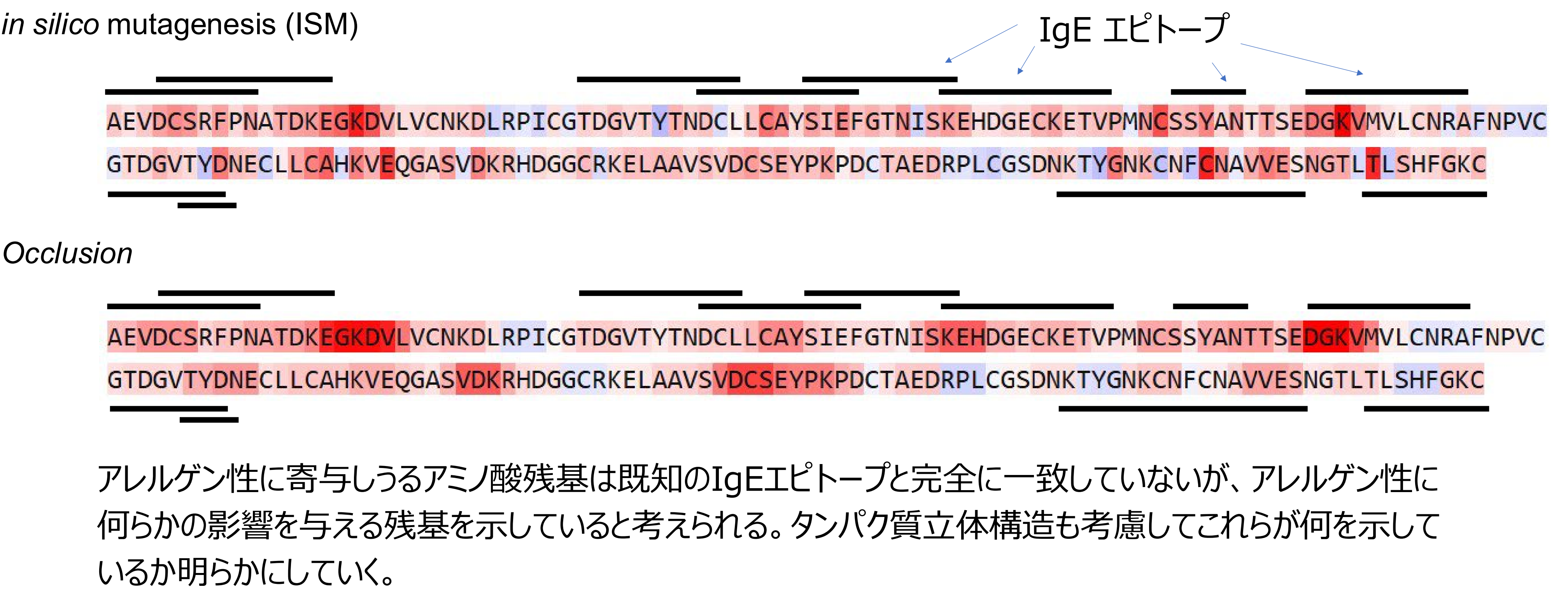


Fig.2. Allergenicity prediction performance of protT5-ESM2 fusion model

② 判断根拠の可視化

鶏卵の代表的アレルギー（Ovomucoid, Gal d 1）のアレルギー予測に影響するアミノ酸残基の可視化を以下のISM, Occlusion法で示した。

赤色アミノ酸残基はアレルギー性にプラスに影響する残基、青色はマイナスに影響する残基を示す。2手法で同様の傾向を示している。



Conclusion

アレルギー性予測の判断根拠の可視化を複数の手法で行い、重要なアミノ酸残基の位置、領域を明らかにできた。今後は、それらの残基がどういう意味を持つのかを、タンパク質言語モデルの中間層からの特徴抽出も用いながら構造的にも考察していく。最終的には、判断根拠を明示できるアレルギー予測システム（データベース）の構築を目標としていく。

