論文 / 著書情報
Article / Book Information

題目(和文)	高精度なホモロジーモデリングのための配列アライメント手法の開発
Title(English)	Development of a protein sequence alignment method for accurate homology modeling
著者(和文)	牧垣秀一朗
Author(English)	Shuichiro Makigaki
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11681号, 授与年月日:2020年12月31日, 学位の種別:課程博士, 審査員:石田 貴士,秋山 泰,岡﨑 直観,村田 剛志,関嶋 政和
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11681号, Conferred date:2020/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 系 知能情報 コース	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of
学生氏名： Student's Name	牧垣 秀一朗	指導教員（主）： Academic Supervisor(main)	石田 貴士
		指導教員（副）： Academic Supervisor(sub)	

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

タンパク質は生命活動に重要な働きをしている生体高分子であり、その立体構造情報は生物学、生化学、薬学において広く用いられている。しかし、タンパク質の立体構造を決定するための実験手法の改善にもかかわらず、アミノ酸配列データベースと比べると立体構造データベースは小さいままとなっているのが現状である。よって、与えられたアミノ酸配列の立体構造モデルを生成するタンパク質立体構造予測は、依然として重要な役割を果たしている。相同タンパク質の構造を用いて構造未知のタンパク質の立体構造を予測するホモロジーモデリングは、良いテンプレートと配列アラインメントを見つけることができれば、予測立体構造モデルがより正確になるため、現在、最も実用的な手法である。実際、相同性検出手法はよく研究され、遠縁の相同タンパク質を高感度に発見することができる。しかし、相同性検出性能の向上を目的とした配列アライメントを用いたホモロジーモデリングによる立体構造予測の精度は、理想的な配列アラインメントから生成される立体構造予測の精度よりも低いことが多い。より高精度なホモロジーモデリングを実現するためには、配列アライメントの改善が重要であり、またこれは未解決の問題である。したがって、ホモロジーモデリングによる正確な構造予測に適切なアライメントを生成する自動化手法が必要とされている。本論文の貢献は、配列類似度の低い立体構造予測対象タンパク質とテンプレートタンパク質のペアを用いた、正確なホモロジーモデリングのための、新しい配列アライメント手法を開発したことである。一般に、配列アライメント生成は相同性検出ツールと統合されており、相同性検出ツールは検索結果と配列アラインメントを出力するが、本研究では、この配列アライメント生成に焦点を当てた。本手法で生成されたアライメントの精度を予測立体構造モデルの精度に基づいて検証し、我々の手法は、特に遠縁相同タンパク質に対して、より適切な配列アライメントを生成することを目的とする。

まず、既知の相同タンパク質の構造アライメントを用いた機械学習モデルに基づく配列アライメント生成法を提案した。構造アライメントにおいては、タンパク質立体構造間の構造的な差異が最小化されるため、構造アライメントによって生成された配列アライメントはホモロジーモデリングに適している。機械学習を用いて配列アラインメントを直接予測することは困難であるが、その代わりに、既存の静的なアミノ酸置換行列やプロファイル比較ではなく、学習したモデルから動的に置換スコアを予測し、配列アライメント生成時に利用する。予測立体構造の精度を最先端の手法と比較することで手法を評価し、我々の手法は、既存手法で得られたアライメントから生成された立体構造予測よりも正確な立体構造モデルを生成することができた。また、提案手法の機械学習モデルが、入力から何を学習したことで精度の向上が得られたかについて解析を行い、配列プロファイル上のアミノ酸出現頻度における多様性がその判定に影響を与えていることを示唆した。

次に、先に述べた手法では精度が不十分であった遠縁タンパク質との配列アライメントについて、Intermediate Sequence Search（ISS）で得られる中間的な相同配列を用いることで、ホモロジーモデリングに対してより高品質な配列アライメントを生成する手法を提案した。ISS は遠縁相同タンパク質の検出において有用であることが知られているが、配列アライメントを生成することはできない。また、その検索結果をホモロジーモデリングに用いるには別の手法で配列アライメントを生成しなければならないが、クエリ配列とターゲット配列の類似度が低いため、既存手法では配列アライメント生成が難しい。そこで、本研究では中間配列検索結果を用いて配列アライメントの生成を行う手法を提案し、相同性検出の感度と選択性を比較して評価し、さらに予測立体構造モデルの精度に基づいて、生成された配列アライメントの精度を検証した。その結果、特に遠縁相同タンパク質ペアに対して、本手法がより適切な配列アライメントを生成できることを示した。また、中間的な相同配列のうちのどれを配列アライメント生成に利用するかの組み合わせについても検討し、可能な全ての組み合わせから配列アライメント長が最長のものを選択することで、より良い配列アライメントが得られることを示した。

これらの提案手法によって生成された配列アライメントを用いたホモロジーモデリングでは、既存の手法による配列アライメントを用いたものに比べて、より高品質なタンパク質立体構造モデルが予測できることが確認され、その改善はタンパク質間相互作用予測などのさらなる生物学的な応用においても十分に影響があるものであることが示された。

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 系 知能情報 コース	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of
学生氏名： Student's Name	牧垣 秀一朗	指導教員（主）： Academic Supervisor(main)	石田 貴士
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

Homology modeling is currently the most practical method because the predicted models are accurate if a good template and sequence alignment can be found. The accuracy of structure prediction generated from the sequence alignment based on recent highly sensitive homology detection methods is often less accurate than the prediction generated from ideal sequence alignment. Therefore, there is a need for automated methods to generate appropriate alignments for accurate structure prediction by homology modeling. This paper's contribution is developing a new sequence alignment method for accurate homology modeling using query and template protein pairs with low sequence similarity.

First, we proposed a machine learning-based method to generate sequence alignments using known homologous proteins' structural alignments. Sequence alignments generated by structural alignments are suitable for homology modeling because, in the structural alignment, the structural differences between the query and the template protein structure are minimized. It is difficult to predict sequence alignments directly using machine learning. Instead, we dynamically predicted substitution scores from the trained model, rather than existing static amino acid substitution matrices and profile comparisons and used them during sequence alignment generation.

Next, we proposed a method to generate higher quality alignments with distantly related proteins using intermediate homologous sequences obtained by Intermediate Sequence Search (ISS) to improve the insufficient accuracy of results by the previously described method. This study is the first to demonstrate the generation and evaluation of sequence alignments using ISS results in the context of homology modeling. ISS's basic idea is to find sequences of distantly related proteins by intermediate related proteins' sequences between a query and hit sequences. ISS is useful in detecting distantly related proteins but cannot generate sequence alignments. It is also difficult to generate sequence alignments using existing methods due to the low similarity between the query and target sequences. This study proposed a sequence alignment method using the intermediate sequence information.

Based on the proposed methods, we could generate more appropriate alignments, especially for distantly related homologous proteins, and succeeded in predicting better protein tertiary structure models.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).