JSEC2020　アブストラクト

CB001

研究タイトル	自然言語処理と機械学習を用いたタンパク質の高発現塩基配列の創製
研究カテゴリー	計算生物学・バイオインフォマティクス
学校名	甲南高等高校
都道府県	兵庫県
研究者氏名	南　慧
研究者(代表者)学年	2 年(高校・高専)

研究の要約

本研究では、核酸の塩基配列を自然言語として扱い、大規模コーパスを作成し、自然言語処理用の機械学習ソフトを用いてタンパク質の発現量を向上させる遺伝子配列を特定した。

大規模コーパスは、遺伝子非翻訳領域の塩基配列とその２次構造に着目し、２次構造の構成要素であるステムとループを形態素として扱う。次に製作したコーパスと、遺伝子発現効率の情報を紐づけ、これを教師データとして、機械学習を行う。具体的には、遺伝子部分からなる単語の分散表現と、この分散表現に基づくニューラルネットワークモデルを構築し、作成されたモデルから発現効率を最大化させる単語のクラスター分析を行った。更にこのクラスタ重心から、発現効率を向上させる遺伝子断片の同定を行った。

この過程で、発現効率の数値をラベル化し、分類予測が行えるかを検証した。

次に構築された単語ベクターのクラスタ分析を行い、発現効率の高い遺伝子断片（単語）を特定し、遺伝子断片から新しい高発現が期待できる塩基配列（文章）を合成する方法を提案した。

本研究では、fastText 組み込みのニューラルネットワークによる分類予測を用いたが、塩基配列の部分配列を分散表現できることが検証できたので、今後は更にサポートベクターマシン等を用いて回帰学習による定量分析を通し、更なるタンパク質発現効率を持つ塩基配列を創製したい。

●確認事項

研究に用いているもの (人間、脊椎動物、微生物、組み換えDNA、細胞組織、どれも用いていない)	どれも用いていない
大学・研究機関などでの実験や装置使用があるか	はい:奈良先端科学技術大学院大学
昨年までの研究からの継続研究か	いいえ(継続研究ではない)