

Identyfikowanie eksonów w sekwencjach nukleotydowych

Identyfikowanie eksonów

Budowanie modelu rozpoczęto od przygotowania danych: sekwencje przedstawiające eksony i introny zostały zaetykietowane, a następnie przetasowane. Następnie z powstałej listy stworzono zbiory treningowe i testowe. Klasy były zbalansowane w następujący sposób:

	Liczba intronów	Liczba eksonów
Zbiór treningowy	211	317
Zbiór testowy	50	83

Następnie poszczególne sekwencje podzielono na k-mery o długości 5, co pozwoliło uchwycić lokalne motywy biologiczne oraz osiągnąć najlepsze rezultaty. Do wektoryzacji wykorzystano *TfidfVectorizer()*, który automatycznie wprowadza ważenie cech oraz normalizację (L2). Dalsza standaryzacja nie była konieczna, a mogłaby zaburzyć biologiczne znaczenie k-merów. Model stworzona za pomocą *LinearSVC* z parametrem *dual = False*, ze względu na charakter danych (duża liczba cech, mało próbek) oraz aby umożliwić na bezpośredni dostęp do parametru *coef_*, potrzebny w kolejnych wersjach modelu. Osiągnięta dokładność wyniosła **97%**, z miarą $F1 \geq 96\%$.

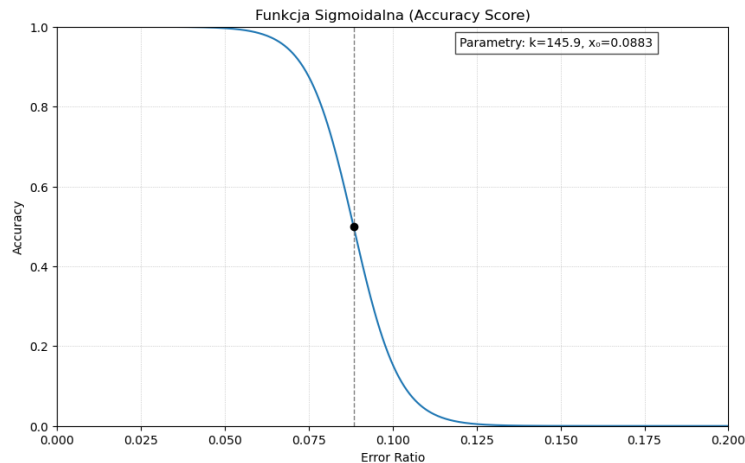
Kolejnym krokiem było zastosowanie *Recursive Feature Elimination (RFE)* w celu zidentyfikowania najistotniejszych k-merów. Funkcja iteracyjnie usuwała po 10% cech z najmniejszymi wagami do osiągnięcia 200 k-merów, które miały największy wpływ na klasyfikację eksonów i intronów. Zwiększyło to dokładność i miare F1 ok **1%**.

Identyfikacja eksonów w sekwencji

Druga część opierała się na stworzeniu algorytmu znajdującego eksony w sekwencjach eksonów i intronów. Sekwencje przygotowano kolejno łącząc elementy z danych testowych w jedną długą sekwencję, gdzie prawdopodobieństwo wystąpienia eksonu wynosiło 50%.

Do wykrywania granic ekson/intron zastosowano algorytm „*sliding window*”. Polega on na klasyfikacji fragmentów sekwencji o ustalonej długości (*window*) przy użyciu wcześniej wytrenowanego modelu. Następnie okno przesuwają się o określoną wartość (*step*) i proces klasyfikacji jest powtarzany. W końcowym etapie każdemu nukleotydowi przyporządkowuje etykietę na podstawie uśrednionych predykcji wszystkich okien, w których się pojawił.

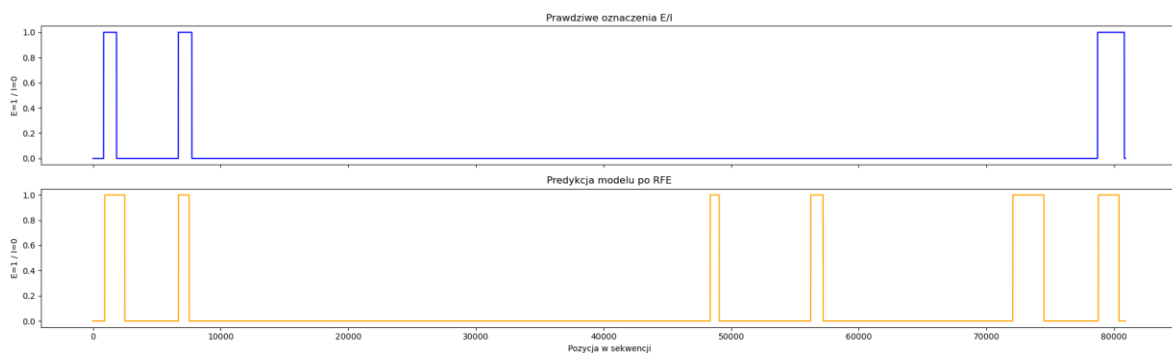
Metrykę dokładności oparto na przekształceniu znormalizowanej liczby błędnie zaklasyfikowanych nukleotydów (*error_ratio*) przez funkcję sigmoid. Parametry *k* i *x_0* zostały dobrane arbitralnie w oparciu o obserwowany rozkład wartości *error_ratio*. Parametr *k* kontroluje stromość spadku funkcji, a *x_0* określa próg, od którego ocena zaczyna maleć.



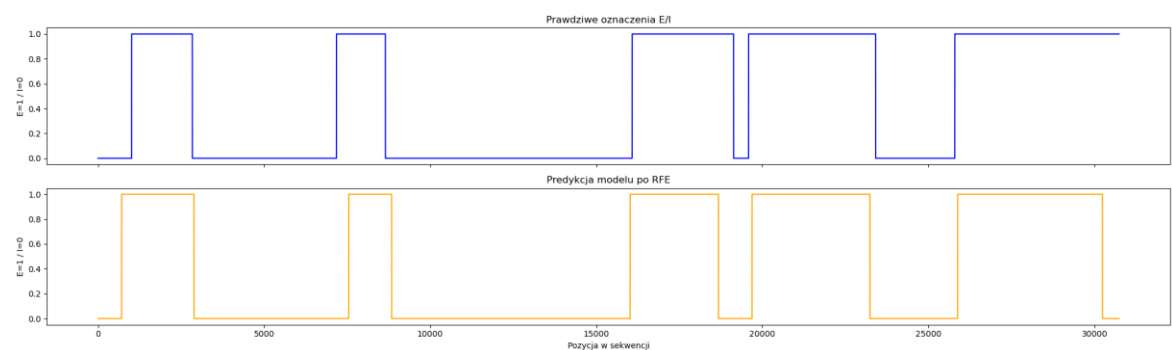
Ostatnim krokiem było dobranie parametrów (*window*, *step*) za pomocą *grid search*.

Wyniki

Najlepsza predykcja osiągnęła 95% dokładności. Wybrana metryka wysoko nagradza precyzyjne określenie granic ekson/intron, jednak skutecznie nie penalizuje krótkich fragmentów błędnej klasyfikacji. Potencjalnym usprawnieniem mogłoby być uwzględnienie k-merów zamiast pojedynczych nukleotydów lub zastosowanie modeli głębokiego uczenia, takich jak sieci neuronowe, które lepiej wychwytyją zależności kontekstowe. Dodatkowo, integracja informacji biologicznych, np. sygnałów splicingowych, mogłaby zwiększyć trafność predykcji. W przyszłości warto również rozważyć użycie metryk uwzględniających zarówno precyzję, jak i ciągłość przewidywanych eksonów.



Rysunek 1: Dokładność 95%



Rysunek 2: Dokładność 88%