PredGenIA:
Transformers para Predicción Genómica

Graciana Castro - Romina Hoffman - Mateo Musitelli
Tutores: María Inés Fariello - Federico Lecumberry
Instituto de Ingeniería Eléctrica

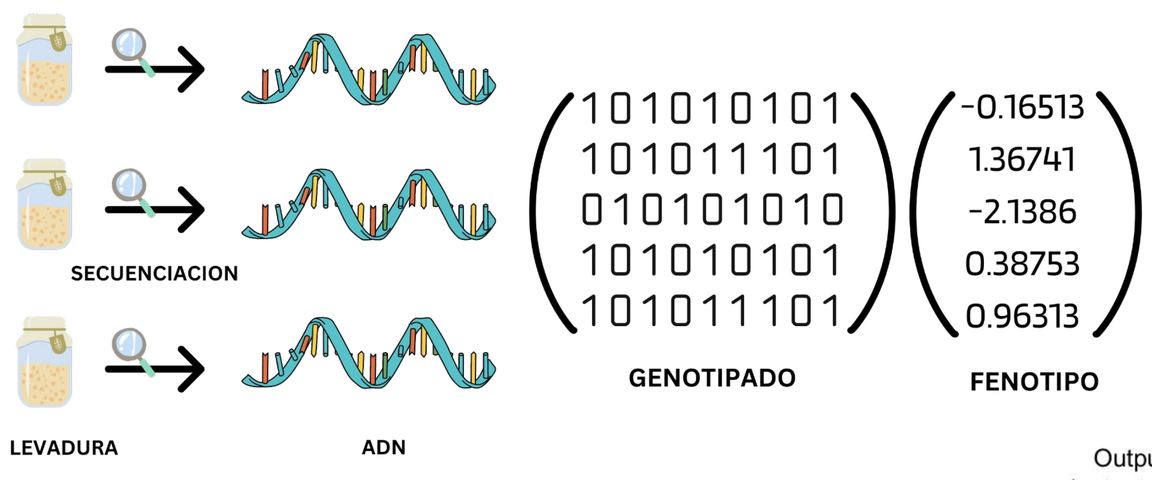
Ingeniería deMuestra
Inteligencia Artificial

Introducción

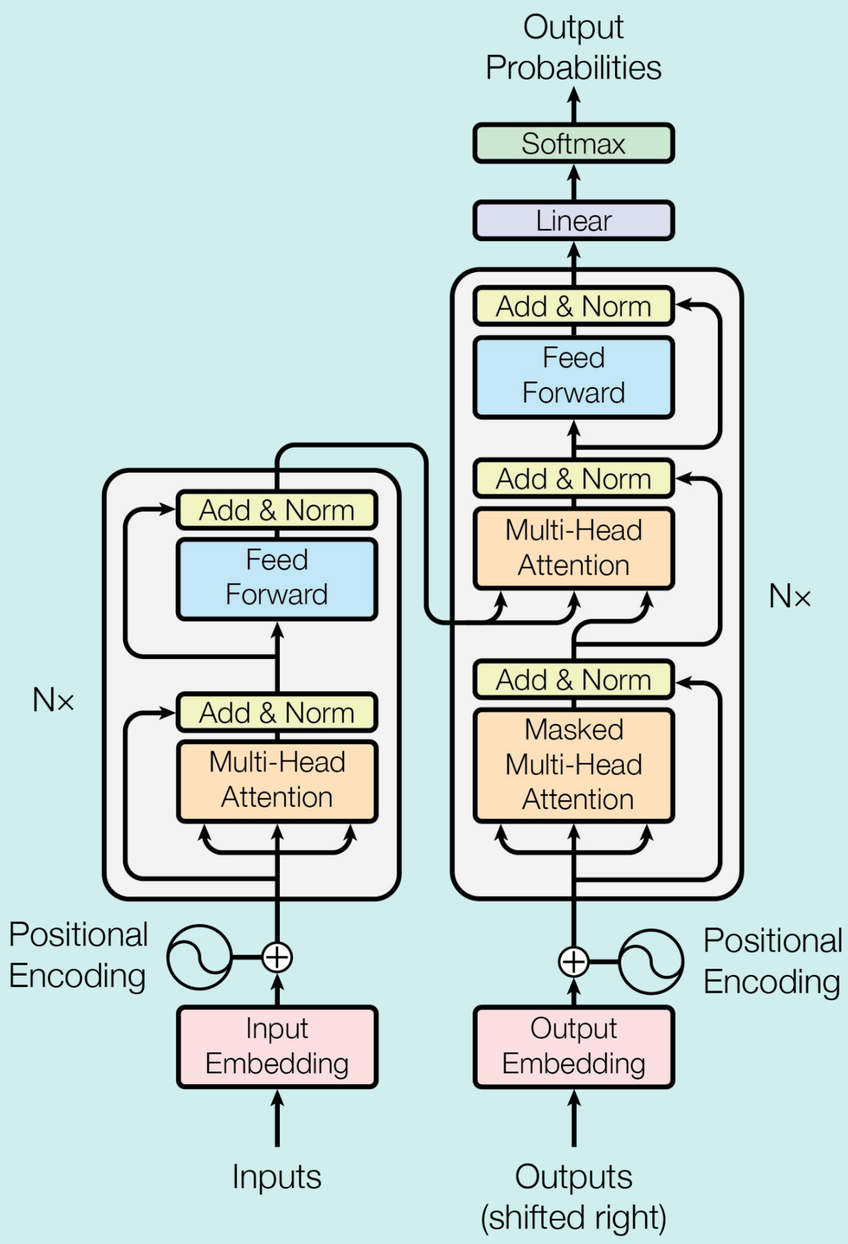
Nuestro proyecto busca aplicar el algoritmo de los Transformers, en la rama de la predicción genómica. Debido a los buenos resultados obtenidos de aplicar este algoritmo en el campo del NLP, y los posibles paralelismos entre los datos genómicos y los datos lingüísticos, es que nos planteamos analizar si los mismos buenos resultados se obtienen en este nuevo campo.

Base de Datos

- 1008 cepas de levadura.
- Genotipo: 11623 SNPs, codificados con valores 0 o 1.
- Fenotipo: crecimiento de cada cepa en 48 ambientes, cuantificado numéricamente.
- Trabajamos con dos ambientes: Lactato y Lactosa.



Transformers

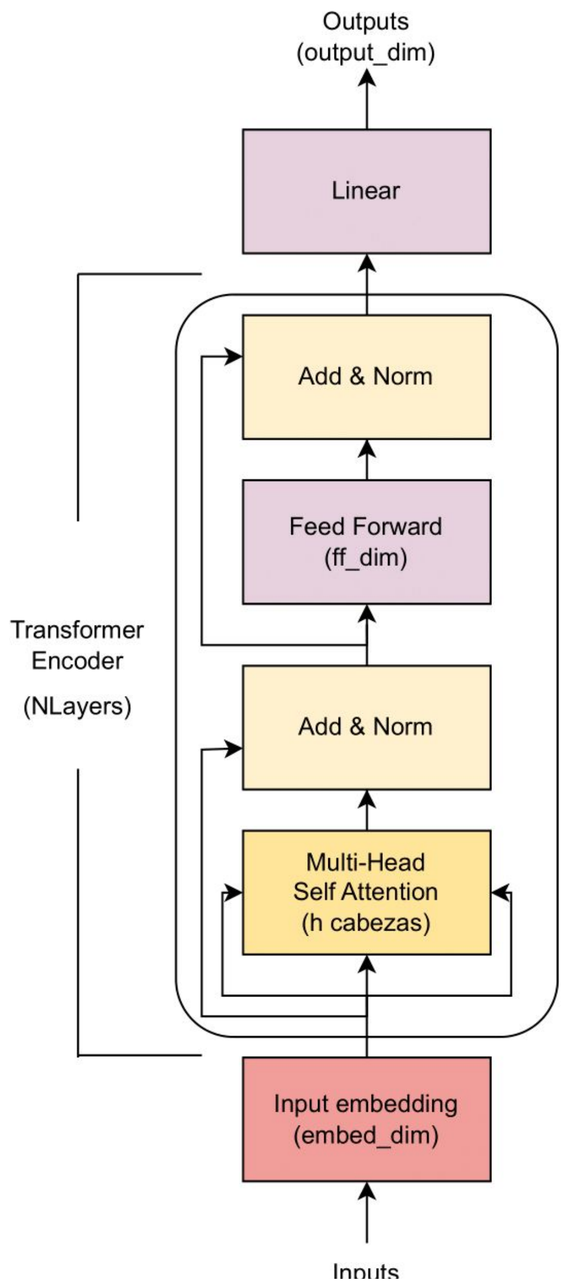


Transformer presentado en “Attention is all you need” (Vaswani et al., 2017).

- Modelo de estructura bidireccional Encoder-Decoder, basado principalmente en el mecanismo “Self-Attention”.
- “Self-Attention”: mecanismo que logra establecer dependencias globales entre secuencias de datos, logrando identificar qué posiciones de una secuencia tienen estrecha relación sobre determinada posición de la otra.
- Encoder: estructura que capta la información y estructura de los datos de entrada.
- Decoder: estructura que logra generar nuevos datos a partir de la información obtenida en el encoder.

Nuestro modelo

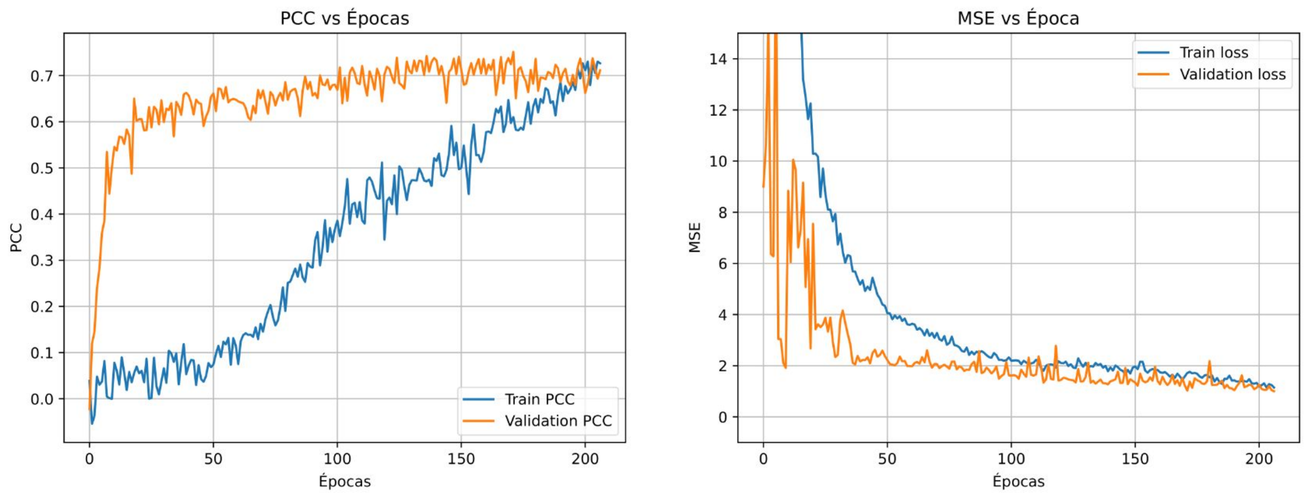
- Basado en el modelo GPTransformer (Jubair et al.,2021), propuesto para trabajar con una base de datos de cebada.
- Cuenta únicamente con estructura Encoder.
- Una capa lineal cumple la función de la capa de Embeddings y Positional Encoding.
- La salida es una capa lineal.



Resultados y conclusiones

Se logró entrenar el modelo para datos genómicos, obteniendo resultados satisfactorios. Sin embargo, los modelos lineales siguen teniendo mejores resultados con menor costo computacional.

Como trabajo futuro se podría explorar con otras bases de datos genómicas. Otros métodos de inicialización y/o regularización de parámetros. Búsquedas más exhaustivas de hiperparámetros óptimos.



Curvas de aprendizaje de nuestro modelo, obtenidas para la predicción de dos fenotipos

Comparación del coeficiente de determinación de nuestro modelo (One trait y Multitrait), con otros resultados expuestos por Grinberg et al., 2019 y Elenter et al., 2021.

Ambiente	Grinberg	GBM	One trait	Multitrait
Lactato	0.568	0.830	-0.543	0.504
Lactosa	0.582	0.860	0.349	0.553

Fundación Julio Ricaldoni
INGENIERÍA EN EL URUGUAY
FACULTAD DE INGENIERÍA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY