Variant effect scores from DNA language models

Edoardo Calderoni, Ruochen Li

Computational Modelling in System Genetics
Technische Universität München

Mentor: Pedro Tomaz da Silva

Abstract

The task of estimating the deleteriousness of a variant is pivotal in human genetics. We exploit the flexibility and power of DNA language models to extract a wide variety of effect scores based on nucleotide dependencies and model embeddings. We benchmark our scores by computing their correlation with nine promoter mutagenesis screens. Finally, we compare our measures to determine the best way of obtaining effect scores from DNA LMs. Our approaches exhibit a marginal improvement over the old influence score.

1 Introduction

Approximating variant effects is a long-standing issue in genetics. Traditional approaches, based on sequence alignment, assign a score to each variant according to how evolutionarily conserved it is, pertaining to the assumption that conserved nucleotides tend to be functionally important. However, this criterion is not sufficient to characterize variant effects. For instance, a variant in a highly conserved non-coding repeat region is less deleterious than one that introduces a stop codon in a similarly conserved region.

DNA language models, which are trained to predict the probability of a nucleotide appearing at a certain position given the context sequence, provide a flexible option for estimating variant effects. Previous studies [1] have attempted to exploit DNA LMs to extract a deleteriousness score by considering

the log ratio of the probabilities assigned by the model ($\hat{p}$) to the alternative and reference nucleotides

$$score(i) := \log_2 \left(\frac{\hat{p}(n_i = ALT_i|\text{context})}{\hat{p}(n_i = REF_i|\text{context})} \right).$$

A new promising approach [4] uses DNA LMs to obtain nucleotide dependencies, which reveal the impact of a query variant on a target nucleotide located elsewhere in the sequence. This dependencies are shown to capture functional elements and complex regulatory phenomena. Furthermore, the outgoing dependencies of a variant can be summed to obtain a measure of deleteriousness, called influence score, relying on the idea that variants that influence the rest of the nucleotides in the sequence have a large impact.

Nevertheless, the influence score is only an initial attempt at capturing nucleotide importance through dependencies, and several interesting ideas remain unexplored. Additionally, scores can also be obtained from DNA LMs by considering the embeddings generated by the model. Our goal is to examine these ideas in detail, and to investigate the best way of leveraging DNA LMs to estimate variant effects.

2 Methods

2.1 Nucleotide dependencies

The dependency between variants i and j is a measure of how a variant i disrupts the probability distribution of the nucleotides at position j. These dependencies can be extracted from DNA language models by computing the probability distribution of nucleotides at position j, while changing the nucleotide $k \in \{A, C, G, T\}$ at i. We consider two definitions of nucleotide dependency. *Maximal dependencies* are given by

$$e_{ij}^{max} := \max \left\{ \left| \log_2 \left(\frac{\hat{p}(n_j = k|n_i = ALT_i)}{\hat{p}(n_j = k|n_i = REF_i)} \right) \right| \right\}_{k \in \{A,C,G,T\}},$$

whereas *entropic dependencies* are defined as

$$e_{ij}^{ent} := \sum_{k \in \{A,C,G,T\}} \hat{p}(n_j = k|n_i = REF_i) \left| \log_2 \left(\frac{\hat{p}(n_j = k|n_i = ALT_i)}{\hat{p}(n_j = k|n_i = REF_i)} \right) \right|.$$

Instead of taking the maximum, entropic dependencies consider a weighted sum of the absolute log odds ratios, under the supposition that the nucleotides with the highest probability under the reference distribution may also have the highest relevance. Removing the absolute value would yield the formula for the KL-divergence between the alternative and reference distribution. Entropic dependencies measure the expected log odds change if we were to sample nucleotides at random according to the reference distribution.

2.2 Influence scores

The influence score estimates the effect of a variant by summing over the outgoing dependencies, under the assumption that variants that disrupt the distribution of the nucleotides elsewhere in the sequence are the most deleterious. It is defined as

$$\inf(i) := \sum_j e_{ij}^{max}.$$

An entropic version of the influence score can be defined as

$$\inf^{ent}(i) := \sum_j e_{ij}^{enf}.$$

2.3 Weighted-sum scores

The idea that a variant should be important if it affects the rest of the sequence can be developed further. Intuitively, if variant i drastically affects position j, but position j is not deleterious with regards to anything else, the score of i should not be largely affected. While computing the score of i, we should weight the dependency between i and j according to the importance of j. This naturally gives rise to a class of scores that can be represented as

$$\text{score}(i) = \sum_j \alpha_j e_{ij}.$$

We consider two different types of weights.

2.3.1 Probabilistic weights

Probabilistic weights value variants based on their probability under the reference and alternative distributions. Ignoring the absolute value, terms in the weighted sum with probabilistic weights look like symmetric KL divergence terms. They are defined as

$$\alpha_j := \max\left\{\hat{p}(n_j = k|n_i = ALT_i) + \hat{p}(n_j = k|n_i = REF_i)\right\}_{k\in\{A,C,G,T\}}.$$

2.3.2 Log odds ratio weights

The idea of log odds ratio weights is to approximate the deleteriousness of a variant using the log odds ratio (as done in [1]), which can be computed easily at every position in the sequence. They are given by

$$\alpha_j := \max\left\{\left|\log_2\left(\frac{\hat{p}(n_j = k|n_i = REF_i)}{\hat{p}(n_j = REF_j|n_i = REF_i)}\right)\right|\right\}_{k\in\{A,C,G,T\}}.$$

2.3.3 Term-wise maximum

Since both maximal dependencies and (probabilistic and log odds ratio) weights involve taking a maximum, it could be beneficial to maximize their product altogether rather than separately, in order to select one nucleotide that is both highly influenced and important:

$$\sum_j \max_{k\in\{A,C,G,T\}} \{\alpha_j(k)\} \max_{l\in\{A,C,G,T\}} \{e_{ij}(l)\} \to \sum_j \max_{k\in\{A,C,G,T\}} \{\alpha_j(k)e_{ij}(k)\}.$$

2.4 PageRank scores

Another natural way to implement the idea of weighting dependencies proportionally to the importance of target variants is to directly use their scores as weights. This can be done by implicitly defining all the scores as the solution of a system of linear equations. Indeed, taking n variants into consideration and given dependencies e_{ij} for $i, j \in \{1 \dots n\}$ we can set the score π_i of i to satisfy

$$\pi_i = \sum_{j\neq i} e_{ij}\pi_j.$$

and solving for $\pi_1 \dots \pi_n$. However, this can be computationally expensive for large values of n. Instead of solving the linear system in $\mathcal{O}(n^3)$, PageRank [2] considers the limiting distribution of a Markov chain on the weighted graph induced by the dependencies, which can be approximated in $\mathcal{O}(tn^2)$, with t being the number of steps of random walks on the graph. In our case, with $n \simeq 2000, t = 10$, the approximation is 200 times faster to compute.

However, given transition probabilities P_{ij}, PageRank assigns scores to the nodes that satisfy the condition

$$\pi_i = \sum_j P_{ji}\pi_j,$$

rather than

$$\pi_i = \sum_j P_{ij}\pi_j$$

as desired. Indeed, a node will be visited often if many important nodes point to it. For this reason, to achieve the second condition, we define the transition probabilities of the Markov chain as the transpose of the dependency matrix

$$P_{ij} := e_{ji}.$$

Moreover, to ensure that the Markov chain has a unique limiting distribution, the authors introduce the possibility, at each step, of visiting a random node with small uniform probability $(1 - \eta)$, instead of continuing the random walk and visiting a neighbor of the current node:

$$P_{ij} := \eta P^{RW}_{ij} + (1 - \eta)\frac{1}{n}$$

or equivalently

$$P = \eta P^{RW} + (1-\eta)\frac{1}{n}\mathbb{1}\mathbb{1}^T.$$

We consider two choices of η, 0.85 and $1 - 10^{-5}$, as our dependency matrix has no off-diagonal zero entries, and it is thus Ergodic.

2.4.1 Variant-specific transition probabilities

PageRank needs a graph with one node for each position in the sequence. But given a certain position in the sequence, we would like to have separate scores for the three variants corresponding to the non-reference nucleotides. This is can be done by enforcing e_{ij} and e_{ji} to consider ALT_i instead of taking the maximum. For a specific variant (i^*, ALT_{i^*}) we set

$$\tilde{e}_{i^*j} = \max\left\{\left|\log_2\left(\frac{\hat{p}(n_j = k|n_{i^*} = ALT_{i^*})}{\hat{p}(n_j = k|n_{i^*} = REF_{i^*})}\right)\right|\right\}_{k\in\{A,C,G,T\}}$$

$$\tilde{e}_{ji^*} = \max\left\{\left|\log_2\left(\frac{\hat{p}(n_{i^*} = ALT_{i^*}|n_j = k)}{\hat{p}(n_{i^*} = ALT_{i^*}|n_j = REF_j)}\right)\right|\right\}_{k\in\{A,C,G,T\}}$$

and $\tilde{e}_{ij} = e_{ij}$ for $i, j \neq i^*$.

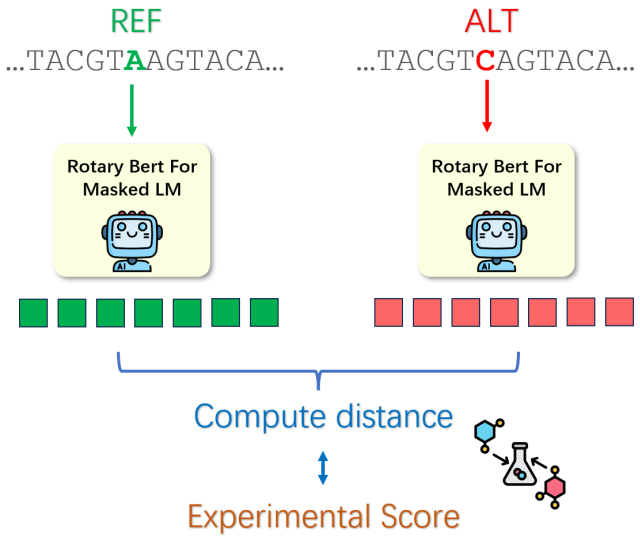

Figure 1: Embeddings-based Method.

2.5 Embedding-based scores

This method utilizes the Rotary BertFor Masked LM[3] to generate embeddings. The embeddings are used to calculate the distance between the reference

sequence and the alternative sequence, which is then employed to estimate the variant effect score. To enhance the robustness and accuracy of our results, we explored various methods for generating embeddings and different algorithms for calculating distances, conducting numerous experiments to optimize our approach.

2.5.1 Rotary Bert for masked LM

The RotaryBertForMaskedLM is a specialized adaptation of the BERT (Bidirectional Encoder Representations from Transformers) model that integrates the Rotary Position Embedding (RoPE) into its architecture. The integration of RoPE within the attention mechanism is pivotal as it allows each token to retain a sense of its original position in the sequence relative to other tokens. This positional awareness is critical in tasks where the relationship and order among tokens determine the meaning of the input. Therefore, this model is suitable for our tasks.

2.5.2 Generating embeddings

Embedding representations encode not only the sequential order of base pairs (A, T, C, G) but also capture contextual information within the sequence. During the pretraining phase, BERT employs the masked language model (MLM) task, which enables the model to learn local contextual information by predicting missing segments. Given the limited nature of our dataset and the infeasibility of full training, we opted to use a pretrained Rotary BERT for Masked LM, which provides a suitable and efficient solution.

As shown in Figure 1, we independently generated embeddings for both the reference sequence and the alternative sequence. The distance between the two embeddings was then computed and analyzed for correlation with experimental scores, using both Pearson's r and Spearman's ρ coefficients. Finally, we evaluated whether this approach outperformed the current state-of-the-art method, the influence score.

We summarize our experiments in the following table:

Context Size	**Integration Techniques**	**Model Depth**
Whole Sequence	Concatenate	Last 1, 2, 4 layers
	Average	Last 1, 2, 3, 4, 5 layers
Variant Position	Concatenate	Last 1, 2, 4 layers
	Average	Last 1, 2, 4 layers

Table 1: Summary of Embedding Generation Experiments

2.5.3 Distance measures

We employed five distinct methods for calculating distances:

1. Euclidean Distance (EC): this measure represents the straight line distance between two points in Euclidean space. It is commonly used for its simplicity and effectiveness in many machine learning applications. The formula for Euclidean distance between two points $\mathbf{p}$ and $\mathbf{q}$ in an n-dimensional space is given by:
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

2. Cosine Similarity (CD): often used to measure cosine of the angle between two vectors, which reflects their orientation rather than magnitude. It is especially useful in text analysis and genomics. Cosine similarity is computed as:
$$\text{cosine}(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\| \|\mathbf{q}\|}$$
The distance can then be calculated as $1 - \text{cosine}(\mathbf{p}, \mathbf{q})$.

3. Manhattan Distance (MHT): also known as the taxicab or city block distance, this metric measures the sum of the absolute differences of their Cartesian coordinates. It is useful in grid-like path planning and urban settings. The formula is:
$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|$$

4. Mahalanobis Distance (MHB): a measure of distance considering the correlation between variables, which differentiates it from Euclidean distance. It is particularly effective in identifying outliers in multivariate data. The formula for Mahalanobis distance is:
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q})^T \mathbf{S}^{-1} (\mathbf{p} - \mathbf{q})}$$
where $\mathbf{S}$ is the covariance matrix of the data points.

5. Jensen-Shannon Divergence (JSD): this method provides a way of measuring the similarity between two probability distributions. It is symmetric and bounded, making it a reliable measure for statistical analysis. The Jensen-Shannon divergence is computed as:
$$JSD(P||Q) = \frac{1}{2} D(P||M) + \frac{1}{2} D(Q||M)$$
where P and Q are the distributions, $M = \frac{1}{2}(P + Q)$, and D is the Kullback-Leibler divergence.

2.6 Benchmark

Our scores are benchmarked against the promoter mutagenesis screen of nine different promoters: F9, GP1BB, HBB, HBG1, HNF4A, LDLR, MSMB, PKLR, TERT-GBM. We obtained confidence intervals by performing 100 bootstrap samples for each promoter, computing the correlation for each sample, and adding/subtracting two standard deviations to the mean correlation. We considered both Pearson's and Spearman's correlation coefficients.

3 Results

3.1 Weighted sum scores

Weighted sum methods showcase a similar performance to that of the influence score, with a slightly higher average correlation.

Remarkably, Figure 2 shows that probabilistic weights only seem to work well when paired with term-wise maximums.

Finally, entropic dependencies don't lead to significant improvements with respect the their maximal counterpart.

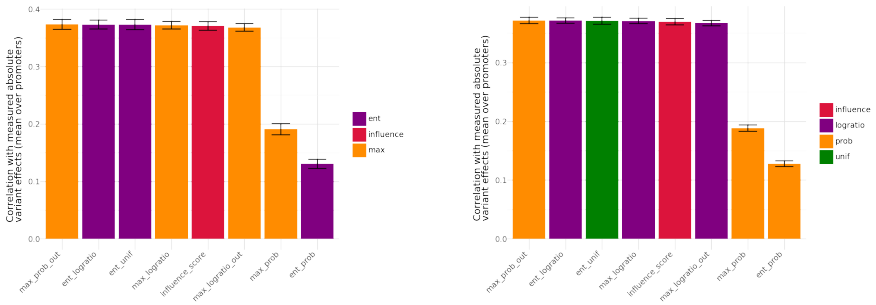

Figure 2: Plot of the performance of weighted sum scores, colored by dependency type (left) and weight type (right).

3.2 PageRank scores

The PageRank scores that we defined all display similar performances. An attempt to define PageRank scores on sparsified graphs with a uniform random walk failed due to the graph being too similar to a path graph. This made it difficult to propagate information on it in a small number of iterations.

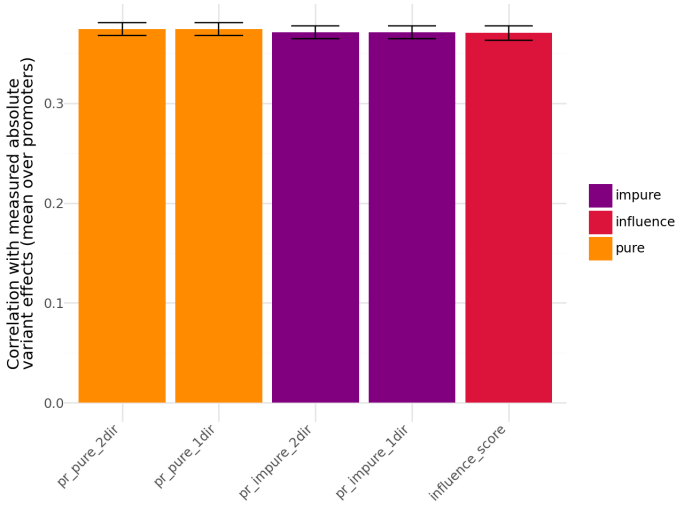

Figure 3: Performance of PageRank scores.

3.3 Embedding-based scores

As it is shown on Figure 4 , most of the subplots exhibit a slight positive correlation, while certain genomic loci, such as HBG1 demonstrates a stronger positive correlation. This suggests that embeddings generated through the final four layers of the whole sequence, combined with MHB distance calculations, can effectively capture gene expression changes at specific loci. However, for some gene locations, such as F9, the correlation is low, indicating limited sensitivity in reflecting expression variations at these sites.

Figure 5 highlights that the best performance is achieved by whole-sequence embeddings. Moreover, the Mahalanobis distance seems to yield slightly better results than other distance metrics.

Finally, Figure 6 showcases that changing the number of embedding layers yields similar results, no matter if they are aggregated by averaging or concatenating.

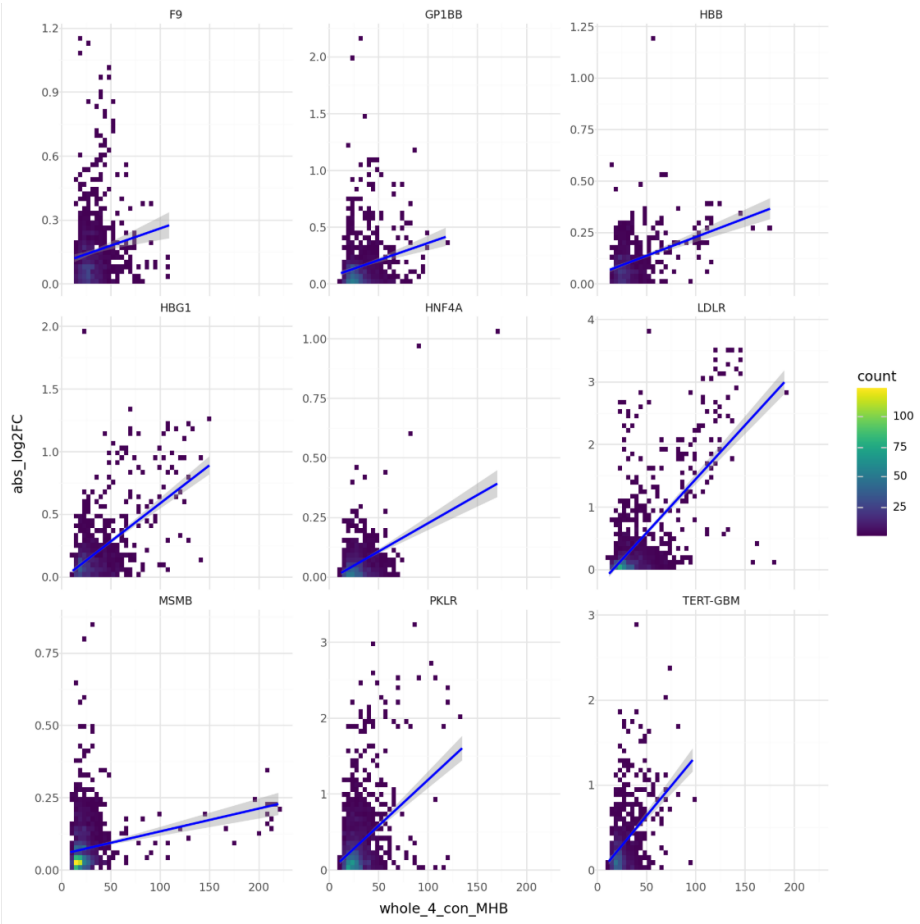

Figure 4: Correlation analysis between concatenated embeddings from the last four layers, computed using MHB distance, and gene expression changes across different loci.

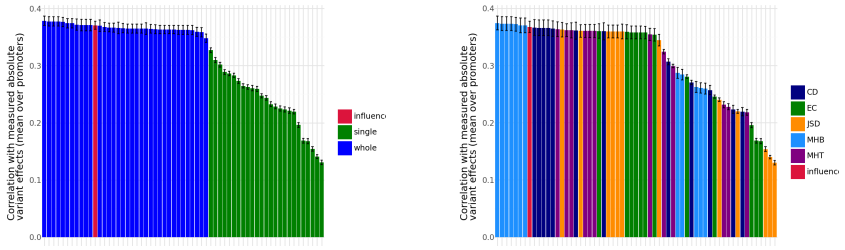

Figure 5: Performance of embedding scores, colored by whether they considered single or whole sequence embedding (left), and distance metric (right).

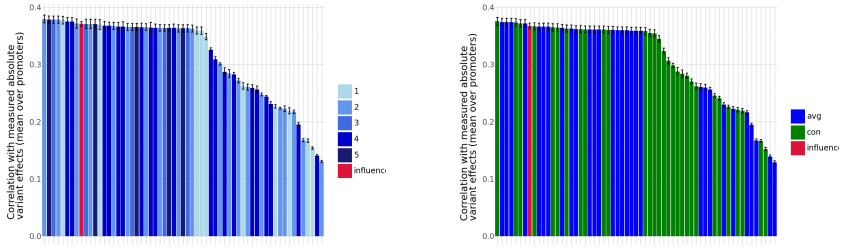

Figure 6: Performance of embedding scores, colored by number of layers considered (left) and aggregation function (right).

4 Discussion

In summary, we defined and compared a wide range of variant effect scores, without obtaining significant improvements. While it's possible that the influence score is already able to extract all the functional information learned by DNA LMs about the deleteriousness of variants, it is also plausible that the benchmarks that we considered were simplified settings that weren't able to highlight the difference between the influence score and more advanced measures. Future research could address more thoroughly whether the scores that we defined in this work could find great applicability in learning the deleteriousness of variants in complex non-coding regions of the genome.

Other approaches to constructing effect scores could be investigated, for example considering different context lengths for computing dependencies, or using other node centrality metrics on the graph induced by dependencies.

By exploring these avenues, DNA LMs have the potential to become essential tools in unsupervised variant effect prediction and other areas of computational biology.

References

[1] Gonzalo Benegas et al. "GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction". In: *bioRxiv* (2024). DOI: 10.1101/2023.10.10.561776. eprint: https://www.biorxiv.org/content/early/2024/04/06/2023.10.10.561776.full.pdf. URL: https://www.biorxiv.org/content/early/2024/04/06/2023.10.10.561776.

[2] Lawrence Page et al. "The PageRank Citation Ranking : Bringing Order to the Web". In: *The Web Conference*. 1999. URL: https://api.semanticscholar.org/CorpusID:1508503.

[3] Jiaming Su, Mosharaf Ahmed, Yuntao Lu, et al. "Roformer: Enhanced transformer with rotary position embedding". In: *Neurocomputing* 568 (2024), p. 127063.

[4] Pedro Tomaz da Silva et al. "Nucleotide dependency analysis of DNA language models reveals genomic functional elements". In: *bioRxiv* (2024). DOI: 10.1101/2024.07.27.605418. eprint: https://www.biorxiv.org/content/early/2024/07/27/2024.07.27.605418.full.pdf. URL: https://www.biorxiv.org/content/early/2024/07/27/2024.07.27.605418.