AALBORG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Revisiting K-mer Profile for Effective and Scalable Genome Representation Learning

Abdulkadir Celikkanat Andres R. Masegosa and Thomas D. Nielsen

VILLUM FONDEN

This work was supported by a research grant (VIL50093) from VILLUM FONDEN

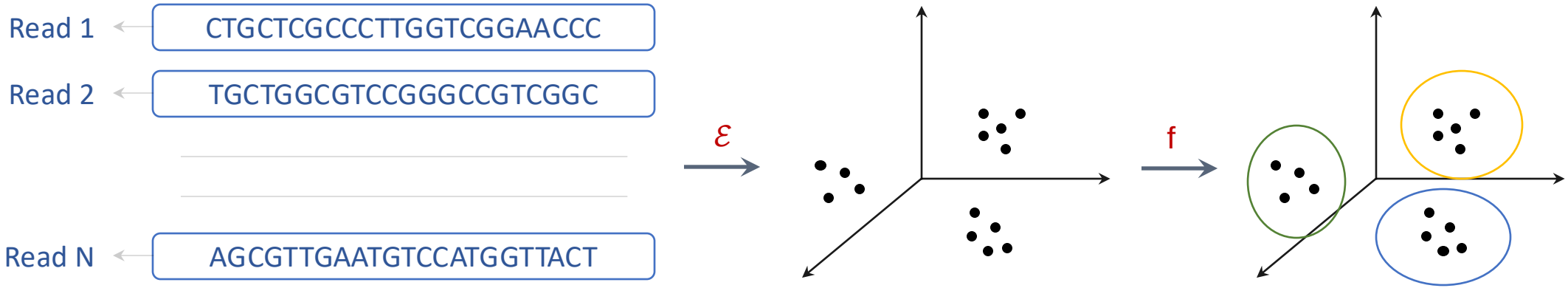
Metagenomics Binning Problem

Genomes

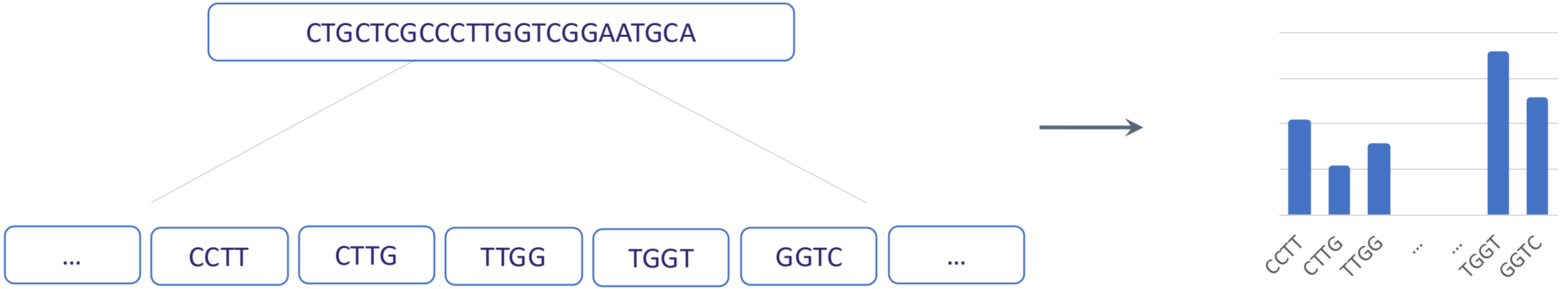
Contigs

Reads

Metagenomics Binning Problem



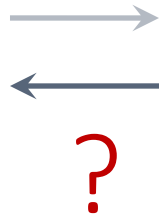
k-mers / k-grams

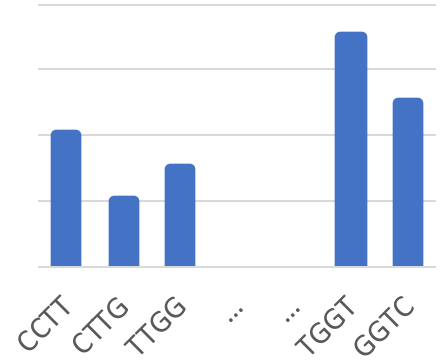


- *k-mers* are used to address several challenges:
  - Variable-length sequences.
  - The ambiguity in read direction
  - Complementary strands

Identifiable reads

CTGCTCGCCCTTGGTCGGAATGCA





Theorem. Let $\mathbf{r}$ be a read of length ℓ. There exists no other distinct read having the same k-mer profile if and only if it does not satisfy any of the following conditions:

1. $r_1 \cdots r_{k-1} = r_{\ell-k-2} \cdots r_\ell$ and $r_i \neq r_1$ for some $1 < i < \ell - k - 2$.
2. $r_i \cdots r_{i+k-2} = r_j \cdots r_{j+k-2}$ and $r_g \cdots r_{g+k-2} = r_h \cdots r_{h+k-2}$ for some indices $1 \leq i < g < j < h \leq \ell - k + 2$ where $r_{i+k-1} \cdots r_{g-1} \neq r_{j+k-1} \cdots r_{h-1}$.
3. $r_i \cdots r_{i+k-2} = r_j \cdots r_{j+k-2} = r_h \cdots r_{h+k-2}$ for some indices $1 \leq i < j < h \leq \ell - k + 2$ where $r_{i+k-1} \cdots r_{j-1} \neq r_{j+k-1} \cdots r_{h-1}$.

- ***Identifiable reads*** can be uniquely reconstructed from their given k-mer profile.

Identifiable reads

- ***Identifiable reads*** can be uniquely reconstructed from their given k-mer profile.
  - But using large values of k is ***impractical***.

- Lipschitz equivalent spaces.

Proposition. Let $M_1 = (\aleph_\ell, d_{\mathcal{H}})$ and $M_2 = (\mathbb{N}^{|\Sigma^k|}, \|\cdot\|_1)$ be the metric spaces denoting the set of identifiable reads and their corresponding k-mer profiles equipped with edit and ℓ_1 distances, respectively. The k-mer profile function, $c : M_1 \to M_2$, mapping given any read, $\mathbf{r}$, to its corresponding k-mer profile, $c_{\mathbf{r}} := c(\mathbf{r})$, is a Lipschitz equivalence, i.e. it satisfies

$$\forall \mathbf{r}, \mathbf{q} \in \Sigma^\ell \quad \alpha_l d_{\mathcal{H}}(\mathbf{r}, \mathbf{q}) \leq \|c_{\mathbf{r}} - c_{\mathbf{q}}\|_1 \leq \alpha_u d_{\mathcal{H}}(\mathbf{r}, \mathbf{q}) \tag{1}$$

for $\alpha_l = 1/\ell$ and $\alpha_u = k|\Sigma|^k$, so M_1 and M_2 are Lipschitz equivalent.

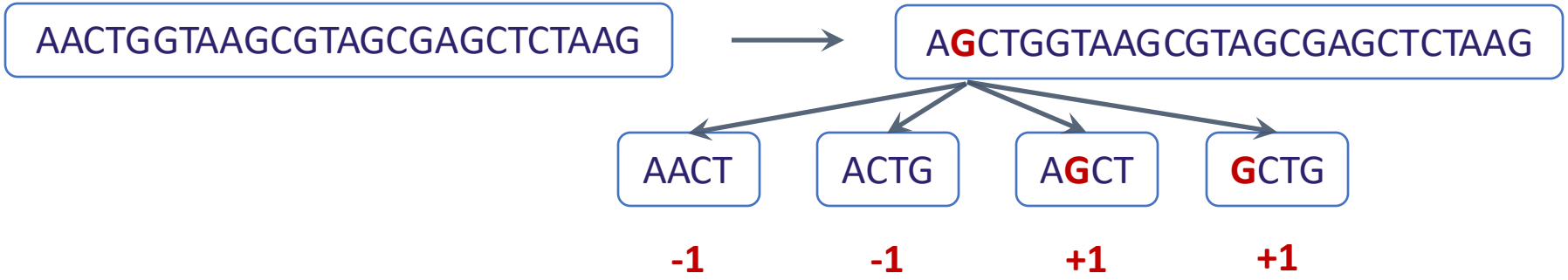
Linear read embeddings

k-mer profile: First, consider the definition of *k-mer profiles*:

$$\mathcal{E}_{kmer}(\mathbf{r}) \coloneqq \sum_{\mathbf{x}\in\Sigma^{\mathrm{k}}} c_{\mathbf{r}}(\mathbf{x})\mathbf{z}_{\mathbf{x}}$$

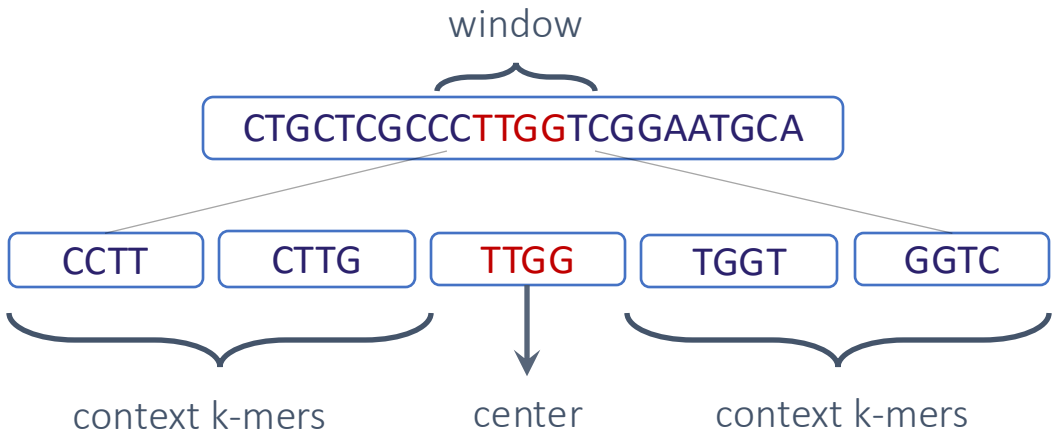
where $\mathbf{z}_x$ represents the canonical basis vector for the k-mer $\mathbf{x} \in \Sigma^{\mathrm{k}}$, i.e. ($\mathbf{z}_{\mathbf{x}} \in \{(u_1, \ldots, u_{|\Sigma|^{\mathrm{k}}}) \in \{0,1\}^{|\Sigma^k|} : \sum_i u_i = 1\}$).

- k-mers are not independent!



Linear read embeddings

Poisson model:



- $o_{\mathbf{x},\mathbf{y}}$ indicates the number of average co-appearances of k-mers i $\mathbf{x}$ and $\boldsymbol{y}$ per read within a window size ω

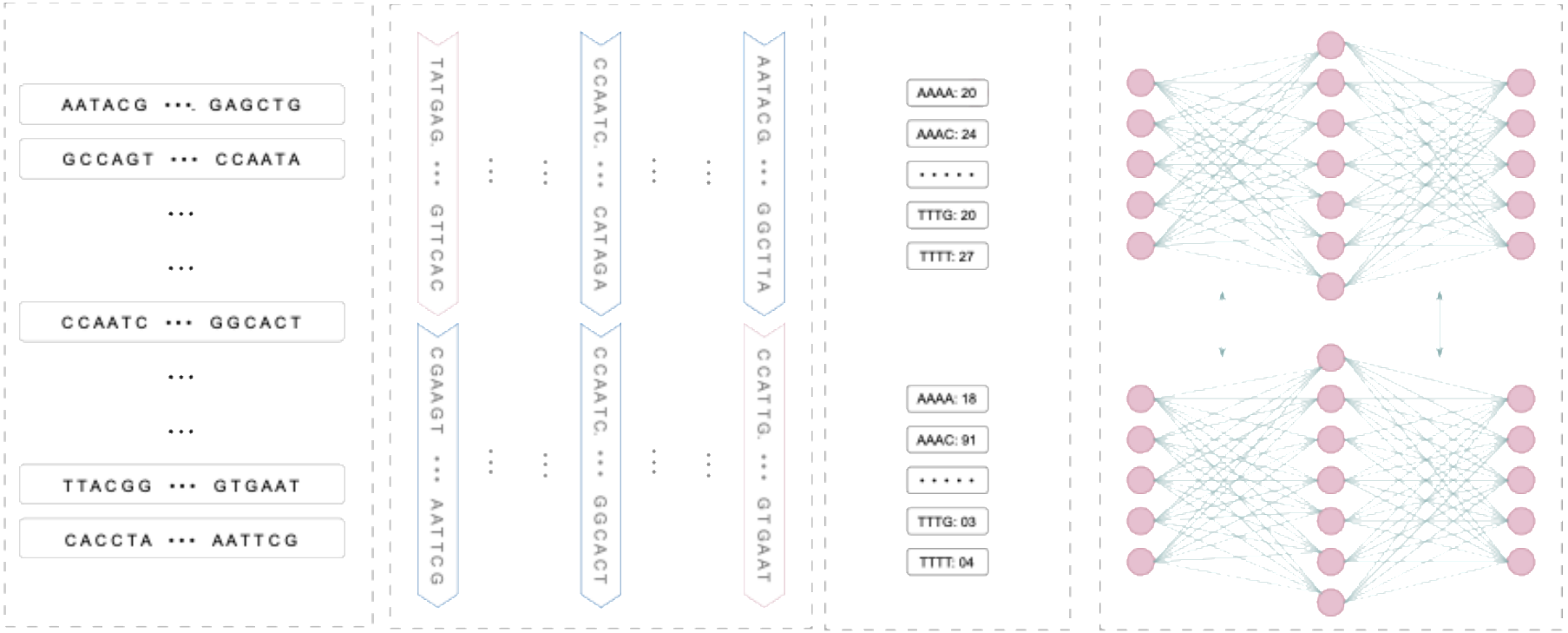
$$o_{\mathbf{x},\mathbf{y}} \sim Pois(\lambda_{\mathbf{x},\mathbf{y}}) \qquad \lambda_{\mathbf{x},\mathbf{y}} \coloneqq \exp(-\|\mathbf{z}_{\mathbf{x}} - \mathbf{z}_{y}\|)$$

- The embedding of read, $\mathbf{r}$, is given by

$$\mathcal{E}_{Pois}(\mathbf{r}) \coloneqq \frac{1}{\sum_{\mathbf{x}\in\Sigma^{\mathrm{k}}} c_{\mathbf{r}}(\mathbf{x})\mathbf{z}_{\mathbf{x}}} \sum_{\mathbf{x}\in\Sigma^{\mathrm{k}}} c_{\mathbf{r}}(\mathbf{x})\mathbf{z}_{\mathbf{x}}$$

Non-linear read embeddings

Non-linear model:

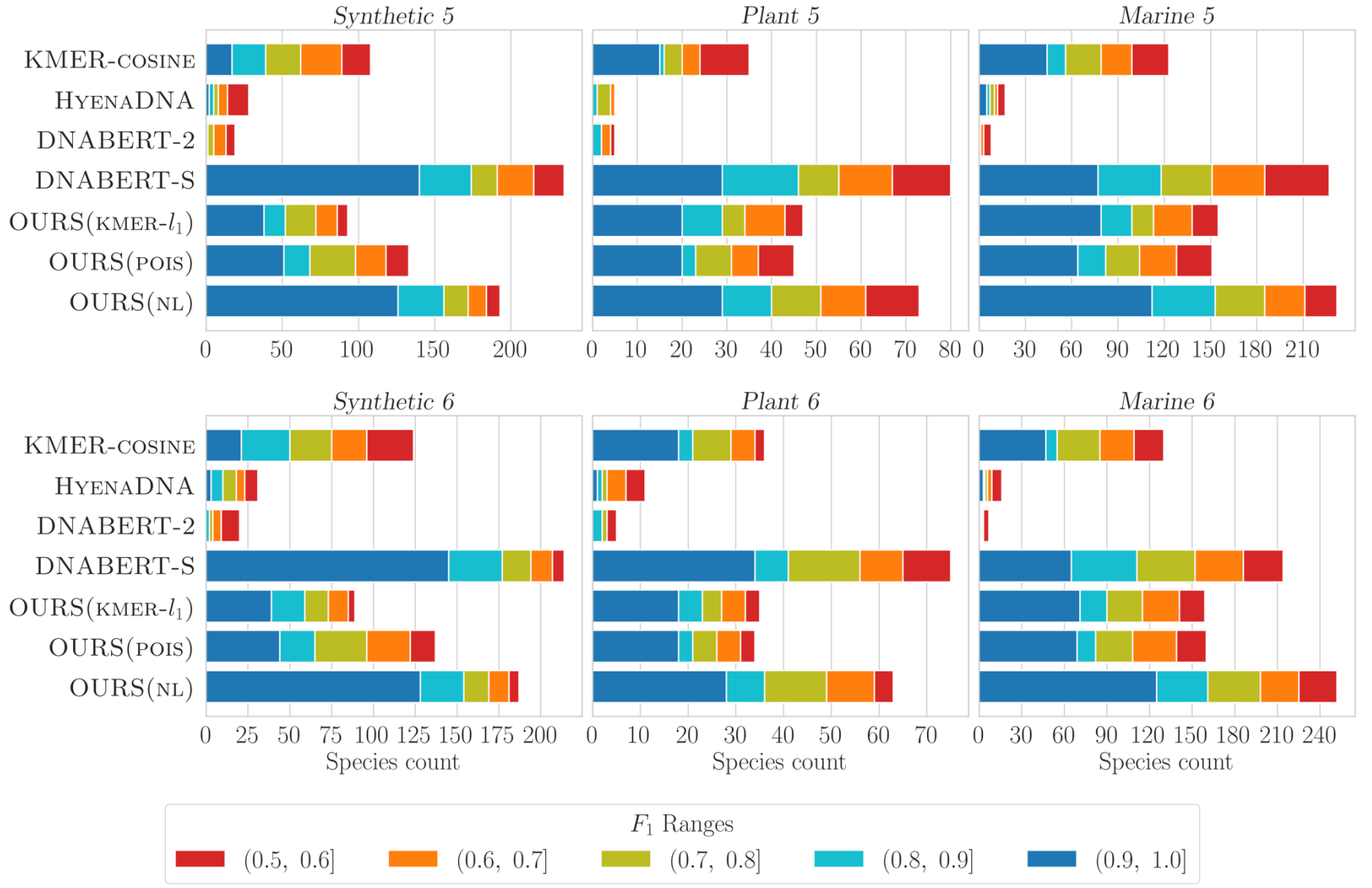


$$\mathcal{L}_{\mathrm{NL}}\left(\{y_{ij}\}_{(i,j)\in\mathcal{I}}|\Omega\right) := -\frac{1}{|\mathcal{I}|}\sum_{(i,j)\in\mathcal{I}} y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij})$$

$$p_{ij} = \exp\left(-\|\mathcal{E}_{\mathrm{NL}}(\mathbf{r}_i) - \mathcal{E}_{\mathrm{NL}}(\mathbf{r}_j)\|^2\right)$$

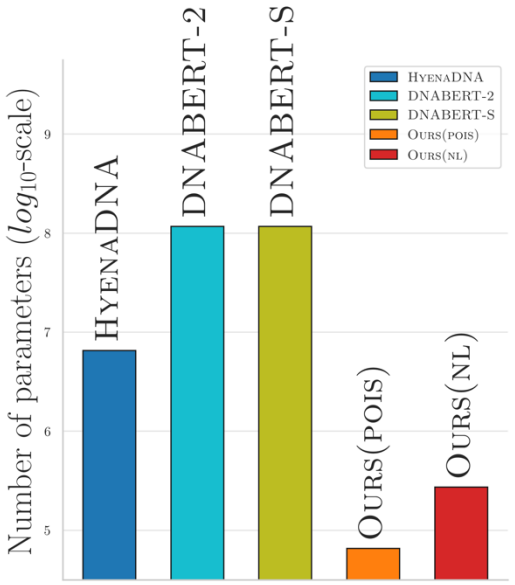
[1] Pan, Shaojun, Xing-Ming Zhao, and Luis Pedro Coelho. "SemiBin2: self-supervised contrastive learning leads to better MAGs for short-and long-read sequencing." Bioinformatics 39.Supplement_1 (2023): i21-i29.

Experiments



Conclusion

- We provide a theoretical analysis of the k-mer space, offering insights into why k-mers serve as powerful and informative features for genomic tasks.
- We show that scalable, lightweight models can provide competitive performance in the metagenomic binning task, highlighting their efficiency in handling complex datasets.
- We demonstrate that models based on k-mers remain viable alternatives to large-scale genome foundation models.



AALBORG UNIVERSITY

Thank you!

For the implementation, datasets, and more details, please visit the address:
https://github.com/abdcelikkanat/revisitingkmers