MetaBAT

Qiulin Zhang, Yu Zheng
ECES-T480

What is MetaBAT

MetaBAT stands for

Metagenome Binning with Abundance and Tetra-nucleotide frequencies

A software tool for binning contigs from samples

More efficient than other binning tools on large metagenomic datasets

Tetranucleotide Frequencies (TNF)

ATCGATTCAATGAC...ATCC

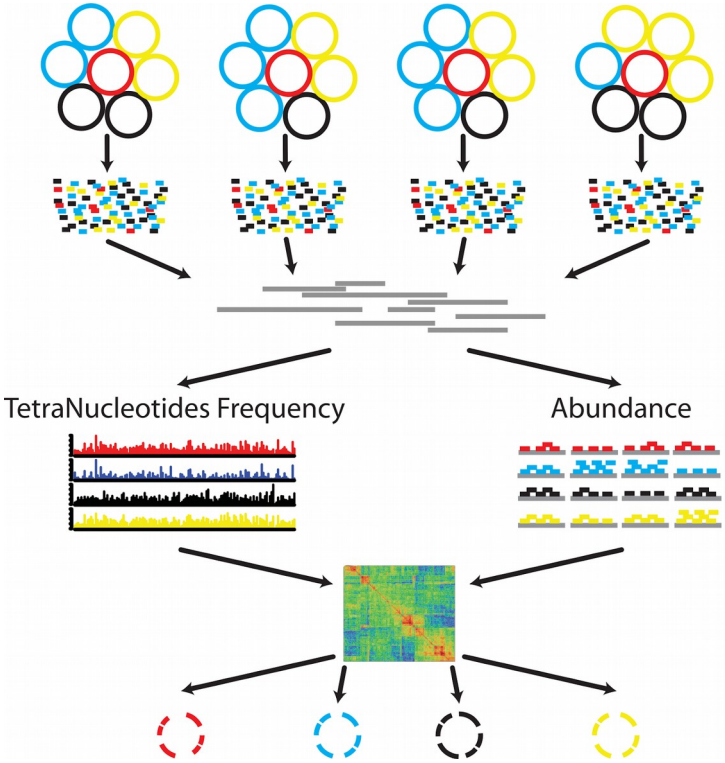
ATCG

4-mers

AAAA	**150,124**
AAAC	**103,010**
AAAG	**255,331**
AAAT	**9,182**
...	
TTTT	**270,463**

256 types of 4-mers

Process



https://doi.org/10.7717/peerj.1165/fig-1

Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Tetranucleotide Frequency Probability Distance (TDP)

The empirical posterior probability that two contigs are from different genomes :

$$P(T|D) = \frac{P(T)P(D|T)}{P(T)P(D|T) + P(R)P(D|R)}$$

T = inter , R = intra, P(T) = 10 * P(R)
D is the Euclidean TNF distance between two contigs

TDP(logistic regression):

$$P(D_{ij}; b_{ij}, c_{ij}) = \frac{1}{1 + e^{-(b_{ij} + c_{ij} * D_{ij})}}$$

b, c are derived from empirical data

Abundance distance probability (ADP)

The abundance distance as the non-shared area of two normal distributions:

$$P\left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) = \frac{1}{2} \int \left| \phi_{\mu_1, \sigma_1^2} - \phi_{\mu_2, \sigma_2^2} \right|$$

μ is the mean of the contig,
σ is the variance of the contig,
Φ represents a normal distribution having two parameters μ and σ^2.

Abundance distance probability (ADP)

Simplified version,
Φ is cumulative normal distribution

$$P\left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) = \begin{cases} \Phi_{\mu_1,\sigma_1^2}(k_0) - \Phi_{\mu_2,\sigma_2^2}(k_0), & \text{if } \sigma_1^2 = \sigma_2^2 \\ \Phi_{\mu_1,\sigma_1^2}(k_2) - \Phi_{\mu_1,\sigma_1^2}(k_1) + \Phi_{\mu_2,\sigma_2^2}(k_1) - \Phi_{\mu_2,\sigma_2^2}(k_2), & \text{otherwise} \end{cases}$$

$$k_0 = \frac{\mu_1 + \mu_2}{2}$$

$$k_1^* = \frac{\sqrt{\sigma_1^2 \cdot \sigma_2^2 \cdot \left((\mu_1 - \mu_2)^2 - 2 \cdot (\sigma_1^2 - \sigma_2^2) \cdot \log(\sigma_2/\sigma_1)\right)} - \mu_1 \cdot \sigma_2^2 + \mu_2 \cdot \sigma_1^2}{\sigma_1^2 - \sigma_2^2}$$

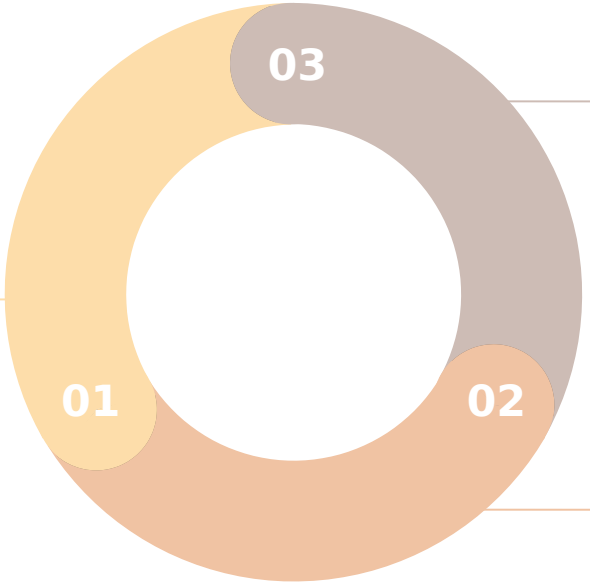
$$k_2^* = \frac{\sqrt{\sigma_1^2 \cdot \sigma_2^2 \cdot \left((\mu_1 - \mu_2)^2 - 2 \cdot (\sigma_1^2 - \sigma_2^2) \cdot \log(\sigma_2/\sigma_1)\right)} + \mu_1 \cdot \sigma_2^2 - \mu_2 \cdot \sigma_1^2}{\sigma_1^2 - \sigma_2^2}$$

$$k_1 = \min(k_1^*, k_2^*) \quad \text{and} \quad k_2 = \max(k_1^*, k_2^*).$$

Integrate TDP and ADP of each contig pair

$$P\left(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) = \begin{cases} \max\left(\mathrm{TDP}, \mathrm{ADP}\right), & \text{if } \mathrm{TDP} > 0.05 \\ \mathrm{ADP} \cdot w + \mathrm{TDP} \cdot (1 - w), & \text{otherwise} \end{cases}$$

Clustering Algorithm



Find a seed contig

(e.g., having the greatest coverage),

and set it as the initial medoid.

Find a new medoid out of all member contigs.

Recruit all other contigs within a cutoff distance (i.e., parameters p1 and p2) to the seed.

Keep large bins (e.g., >200 kb), and dissolve all other bins into free contigs.