Alignment-free tools for metagenomics-data analysis

Robert Deibel

Eberhard-Karls Universität Tübingen

robert.deibel@student.uni-tuebingen.de

November 27, 2017

Overview

1. Metagenomics
   - Metagenome
   - NGS and alignment
2. Alignment-based approach
3. Alignment-free methods
   - Statistics as similarity measurement
   - CVTree
   - D_2^S, D_2^* and their normalization
   - Consideration of mismatches
   - Consideration of mismatches
   - Machine learning – BH-SNE

Metagenomics

Metagenome

- A metagenome is the whole set of transcripts found in a sample.
- Metagenomics is the study of those
- $>$ 90% unculturable microorganisms
- design of antibiotics, analysis of microorganismal life

NGS and alignment

- Advances in sequencing made metagenomics possible
- NGS generates comparable reads

Metagenomics

Goals

- insight in microorganismal life
- first evidence of origin and function
- independent from databases and coding regions

Alignment-based approach

Advantages

- Align sequences against database
- Profiles can be analyzed
- BLAST $>$ 80% accuracy

Alignment-based approach

Advantages

- Align sequences against database
- Profiles can be analyzed
- BLAST $>$ 80% accuracy

Disadvantages

- Low speed
- Dependent of databases
- Unsequenced transcripts cannot be matched
- Databases mostly consist of coding sequences

Different approaches

Statistics

- Utilizes statistics differing in power
- based on k-tuple counts
- Have to be applied through (dis)similarity matrix
- Further analysis follows

Machine learning

- Optimization of a function
- based on k-mer signature
- applies BH-SNE
- Visualizes data in scatter plots

Measuring similarity

$$D_2 = \sum_{w \in \mathcal{A}^k} X_w Y_w$$

where:

- X_w, Y_w number of occurrences in A, B
- $\mathcal{A}$ alphabet
- k is length of w

Measuring similarity

$$D_2 = \sum_{w \in \mathcal{A}^k} X_w Y_w$$

where:

- X_w, Y_w number of occurrences in A, B
- $\mathcal{A}$ alphabet
- k is length of w

Problem

D_2 is not normalized $\Rightarrow$ results vary on different factors

Measuring similarity

Assume sequences are generated through Markov chain

Measuring similarity

Assume sequences are generated through Markov chain

$$D2z(A, B) = \frac{D_2(A, B) - E(D_2)}{\sqrt{Var(D_2)}}$$

Measuring similarity

Assume sequences are generated through Markov chain

$$D2z(A, B) = \frac{D_2(A, B) - E(D_2)}{\sqrt{Var(D_2)}}$$

- compared to five other measures $D2z$ outperformed them

Measuring similarity

Assume sequences are generated through Markov chain

$$D2z(A, B) = \frac{D_2(A, B) - E(D_2)}{\sqrt{Var(D_2)}}$$

- compared to five other measures $D2z$ outperformed them
- needs two parameters, k and r

Measuring similarity

Assume sequences are generated through Markov chain

$$D2z(A, B) = \frac{D_2(A, B) - E(D_2)}{\sqrt{Var(D_2)}}$$

- compared to five other measures $D2z$ outperformed them
- needs two parameters, k and r
- $E(D_2)$ and $Var(D_2)$ calculated with Markov chain in mind

Calculating the expected value

MC of order zero

Probability of $A = B$ or...

Calculating the expected value

MC of order zero

Probability of $A = B$ or...

Sum of background probabilities f_a^A, f_a^B to the power of k

Calculating the expected value

MC of order zero

Probability of $A = B$ or...

Sum of background probabilities f_a^A, f_a^B to the power of k

$$E(D_2) = \left(\sum_{a \in \mathcal{A}} f_a^A f_a^B \right)^k$$

Calculating the expected value

MC of order zero

Probability of $A = B$ or...

Sum of background probabilities f_a^A, f_a^B to the power of k

$$E(D_2) = \left(\sum_{a \in \mathcal{A}} f_a^A f_a^B \right)^k$$

MC of order 1

Sum of probabilities for $|w| = k$, under consideration of MC

Calculating the expected value

MC of order zero

Probability of $A = B$ or...

Sum of background probabilities f_a^A, f_a^B to the power of k

$$E(D_2) = \left(\sum_{a \in \mathcal{A}} f_a^A f_a^B \right)^k$$

MC of order 1

Sum of probabilities for $|w| = k$, under consideration of MC

$$\sum_{|w|=k} p^A(w_1) p^A(w|w_1) p^B(w_1) p^B(w|w_1)$$

Constructing phylogenetic trees – CVTree

- considers (k-2)-th order MC estimated by E_w^X

Constructing phylogenetic trees – CVTree

- considers (k-2)-th order MC estimated by E_w^X

$$Hao = \frac{1}{2}(1 - C)$$

Constructing phylogenetic trees – CVTree

- considers (k-2)-th order MC estimated by E_w^X

$$Hao = \frac{1}{2}(1 - C)$$

$$C = \frac{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)\left(\frac{Y_w - E_w^Y}{E_w^Y}\right)}{\sqrt{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)^2 \sum_w \left(\frac{Y_w - E_w^Y}{E_w^Y}\right)^2}}$$

Constructing phylogenetic trees – CVTree

- considers (k-2)-th order MC estimated by E_w^X

$$Hao = \frac{1}{2}(1 - C)$$

$$C = \frac{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)\left(\frac{Y_w - E_w^Y}{E_w^Y}\right)}{\sqrt{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)^2 \sum_w \left(\frac{Y_w - E_w^Y}{E_w^Y}\right)^2}}$$

- Observations in composition vector
- subtraction of background "noise" through MC
- C is cosine between vectors

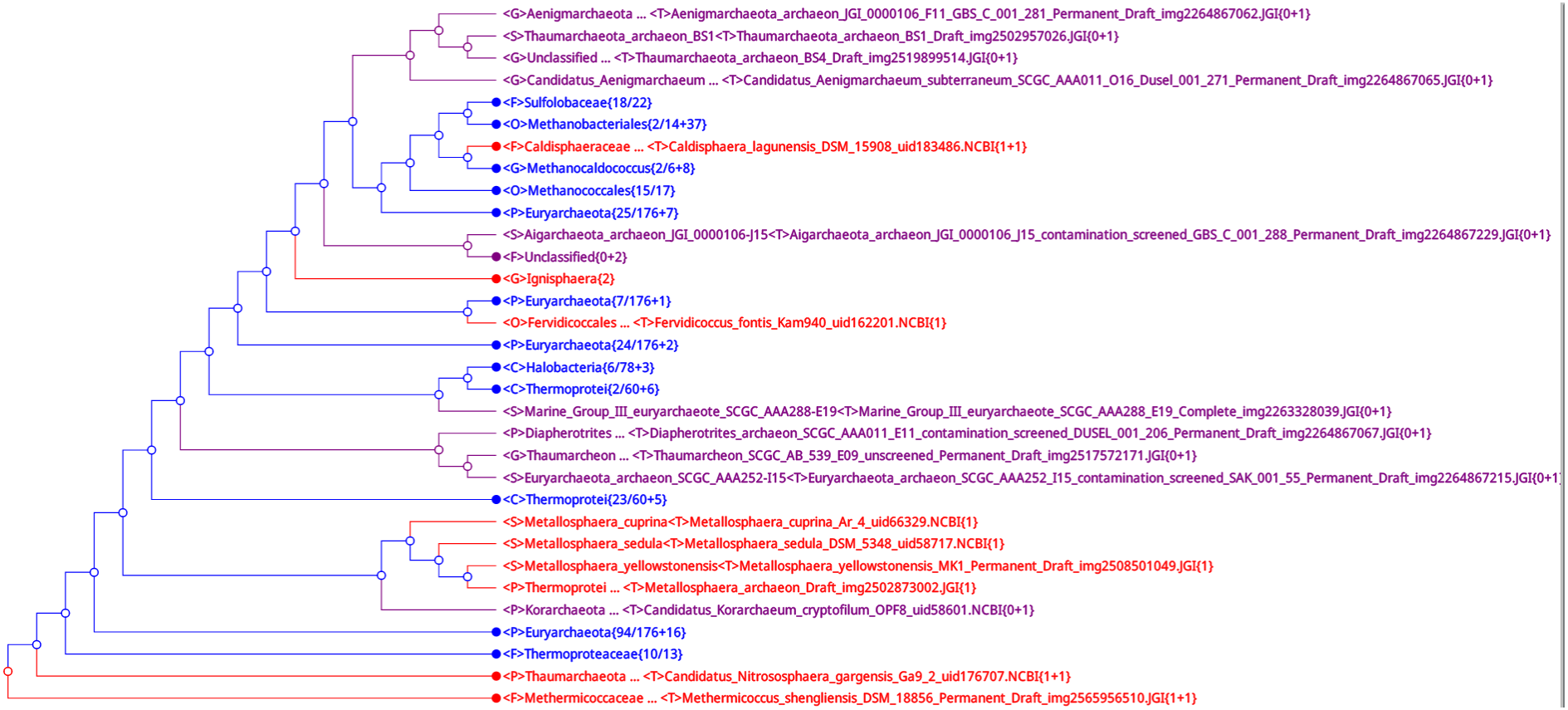

Figure: Computed phylogenetic tree through application of neighbor joining dissimilarity matrix

Nucleotide frequency

- Related approach to *Hao*
- Consider di-nucleotide frequency

$$\rho_{ab}(A) = \frac{f_{ab}}{f_a f_b}$$

- Can be extended to tri- and tetra nucleotides
- l_p norm as dissimilarity measure

$$\delta(A, B) = \sum_{ab \in A} |\rho_{ab}(A) - \rho_{ab}(B)|$$

D_2^S, D_2^* and their normalization

- For two normal random variables $XY/\sqrt{X^2+Y^2}$ is also normally distributed

D_2^S, D_2^* and their normalization

- For two normal random variables $XY/\sqrt{X^2+Y^2}$ is also normally distributed

$$D_2^S = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\widetilde{X}_w^2 + \widetilde{Y}_w^2}}$$

D_2^S, D_2^* and their normalization

- For two normal random variables $XY/\sqrt{X^2+Y^2}$ is also normally distributed

$$D_2^S = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\widetilde{X}_w^2 + \widetilde{Y}_w^2}}$$

- D_2^* utilizes that number of occurrences is approximately Poisson; mean and variance are the same

D_2^S, D_2^* and their normalization

- For two normal random variables $XY/\sqrt{X^2+Y^2}$ is also normally distributed

$$D_2^S = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\widetilde{X}_w^2 + \widetilde{Y}_w^2}}$$

- D_2^* utilizes that number of occurrences is approximately Poisson; mean and variance are the same

Conclusions

1. D_2^S and D_2^* have higher power than D_2
2. D_2^* has highest power when k equals *motif* length
3. D_2^* has higher power for short sequences

but again both not normalized

Normalization and Neighborhood

Normalization

- Normalization to d_2^S and d_2^*
- 0 when sequences are the same and close to one if anti-correlated
- Now applicable to metagenomic data or varying types of sequences

Normalization and Neighborhood

Normalization

- Normalization to d_2^S and d_2^*
- 0 when sequences are the same and close to one if anti-correlated
- Now applicable to metagenomic data or varying types of sequences

Consideration of mismatches

- instead of w the statistics should consider the neighborhood $\varsigma(w)$
- If $w' \in \varsigma(w)$ w' has a certain number of mismatches with w
- Reverse complement can be included similarly
- statistics can then be modified

Performance under the mismatch model

Test parameters

- sequences from mouse embryo
- positive and negative set maximum of 30% repetitions
- Dissimilarity was calculated
- Threshold was applied
- prediction of dissimilarity lower than threshold resulted in positives
- Predictions were compared to real data
- Testing with different parameters for k, r and mismatch weight

Performance under the mismatch model

Test parameters

- sequences from mouse embryo
- positive and negative set maximum of 30% repetitions
- Dissimilarity was calculated
- Threshold was applied
- prediction of dissimilarity lower than threshold resulted in positives
- Predictions were compared to real data
- Testing with different parameters for k, r and mismatch weight

Test conclusions

- *Hao* performed worse than d_2^S and d_2^*
- d_2^S and d_2^* performed best with mismatch weight of 0.05 and $k = 4$
- Overall d_2^S achieved best results in testing

Machine learning – BH-SNE

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)

Title of the publication

Journal Name 12(3), 45 – 678.

The End