Inférence des modèles SBM par des tenseurs trains pour l'identification des taxons pour le métabarcoding

Mohamed Anwar ABOUABDALLAH
Directed by : Nathalie Peyrard[1] Alain Franc[2] Olivier Coulaud[3]

[1] INRAE, Unité MIAT, Toulouse, France

[2] INRAE, UMR BioGeCo, Pierroton & EPC INRAE / Inria Talence, France

[3] Inria, HiePACS, Talence, France

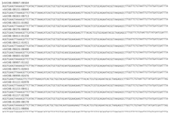
Séminaire des doctorants

Scientific context

Plant diversity

⇒

Sequences



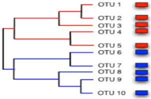
⇒

Dissimilarity matrix

$$\begin{pmatrix} 0 & 2 & \cdots & 3 \\ \vdots & 0 & \ddots & 11 \\ \vdots & \ddots & \ddots & 0 \\ 3 & \cdots & \cdots & 0 \end{pmatrix}$$

⇒

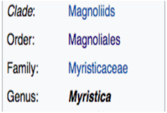
Clusters



⇓

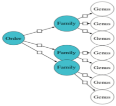
⇒

⇒



Clade:	Magnoliids
Order:	Magnoliales
Family:	Myristicaceae
Genus:	***Myristica***

⇒



- **Is there an adequacy between botanical and molecular classifications on a coarse taxonomic level ?** (first year)
- **How to scale up the SBM model to large datasets ?** (second and third years)

Summary

1 Introduction

- Data set
- High taxonomics levels
- General approach

2 Adequacy between botanical and molecular classifications

- 30 replicates
  - Presentation of the results
  - Which factors explain NMI ?
- Whole dataset

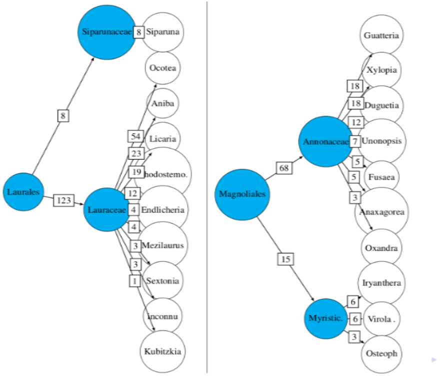
3 Tensor trains approximation for SBM inference

- Model parameters
- SBM possible inferences methods
- Putting SBM on a tensor train

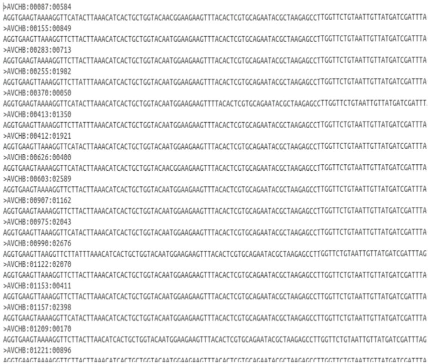
Introduction

Data set

- **1458 trees from an experimental plot in Guyana.**

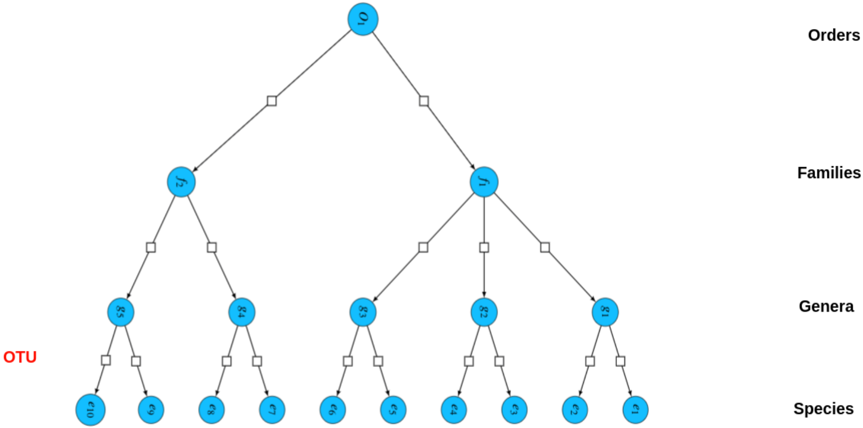


Order, family, genus and species of each individual.



```
>AVCHB:00087:00584
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAACGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00155:00849
AGGTGAAGTTAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00283:00713
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00255:01982
AGGTGAAGTTAAAGGTTCTTATTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00370:00050
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTT.
>AVCHB:00413:01350
AGGTGAAGTTAAAGGTTCTTATTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00412:01921
AGGTGAAGTTAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00626:00400
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAACGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00603:02589
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00907:01162
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00975:02043
AGGTGAAGTTAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00990:02676
AGGTGAAGTTAAGGTTCTTATTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTAG
>AVCHB:01122:02070
AGGTGAAGTTAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:01153:00411
AGGTGAAGTTAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:01157:02398
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:01209:00170
AGGTGAAGTAAAAGGTTCTTACTTAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTAG
>AVCHB:01221:00896
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
```

DNA sequence of each individual.



Orders

Families

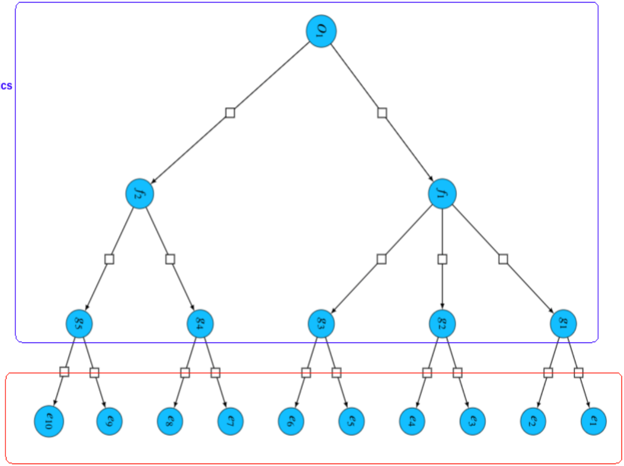
Genera

OTU

Species

High taxonomics levels

OTU



Orders

Families

Genera

Species

General approach

The four steps of the approach

- **Step 1 :** Choice of sub-samples to study :
  - First experiment: 30 replicates : 10 orders and 20 families

Taxonomic level	Total	Selected
Orders	20	10
Families	56	20

  - Our work :
    - Find families in orders;
    - And genera in families;

General approach

The four steps of the approach

- **Step 1 :** Choice of sub-samples to study :
  - Second experiment: Whole dataset

    Selection :

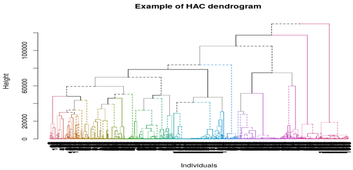
Taxonomic level	sequences	Number of taxa	Minimal size
Species	313	55	5
Genera	845	36	10
Families	1349	30	10
Orders	1357	11	15

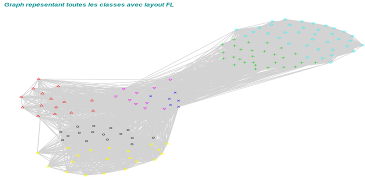
General approach

The four steps of the approach

- **Step 2 :** Building partitions with four methods for each sub-sample and with Smith Waterman and kmer dissimilarities :

  - **M_1 : Agglomerative Hierarchical Clustering (AHC)**
  - **M_2 : Stochastic Block Model (SBM).**





General approach

The four steps of the approach

- **Step 3 :** Comparing the classifications two by two
  - Using visual tools
  - Using NMI to characterize the adequacy/independence
- **Step 4 :** Analyse the different indices and visualise them
  - Using histogram representation
  - Computing statistics on the distribution (mean, median, ...)

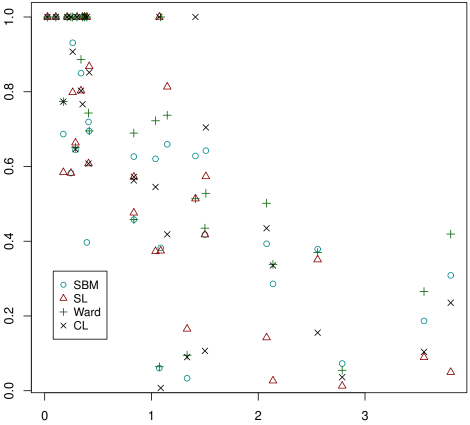
Adequacy between botanical and molecular classifications

Results for the 30 replicates

		Families		Genera		Pooled	
Method		SW	kmers	SW	kmers	SW	kmers
AHC	Ward	1	0.61	0.83	0.73	0.87	0.71
	SL	0.88	0.54	0.75	0.59	0.76	0.58
	CL	0.85	0.63	0.75	0.71	0.75	0.67
SBM		0.57	0.52	0.82	0.66	0.68	0.63

Is there a correlation between r_{mean} value and NMI index ?

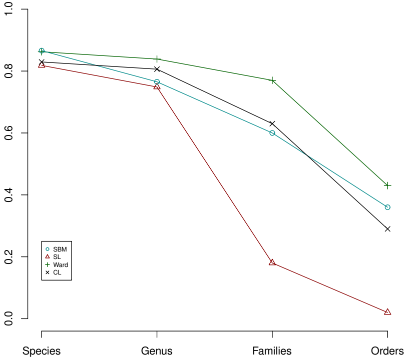
Smith Waterman dissimilarities



Results as a function of Taxonomic level

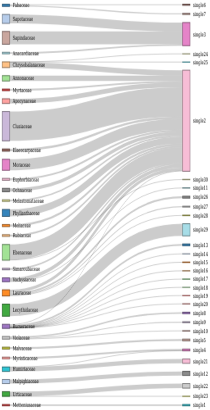
Evolution of NMI index as a function of Taxonomic levels:

kmer based distances

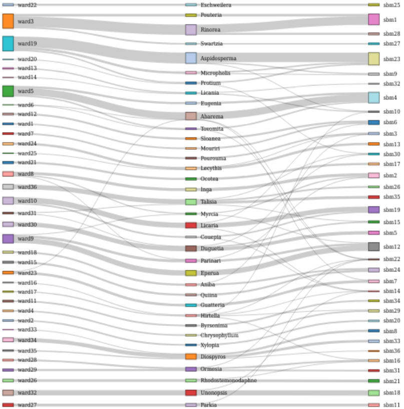


Sankey plots for genera

SL vs Botanics



Ward vs SBM vs Botanics



Interest of SBM models

The main differences between AHC and SBM :

- AHC produces community vs SBM produces classes (not necessary communities)
- Outputs of SBM are : Z and Λ.

Let's talk about Λ :

Case 1 :

$$\Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 1 \end{pmatrix}$$

There are 3 communities $\Longrightarrow$ SBM $\simeq$ CAH

Interest of SBM models

Case 2 :

$$\Lambda = \begin{pmatrix} 22 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 1 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 23 & 7 \\ 8 & 5 & 1 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 19 \end{pmatrix}$$

There are 2 communities $\Longrightarrow$ SBM (warning) $\neq$ CAH

Case 3 :

$$\Lambda = \begin{pmatrix} 22 & 9 & 11 \\ 6 & 23 & 7 \\ 8 & 5 & 19 \end{pmatrix}$$

There are no communities $\Longrightarrow$ SBM (warnings) $\neq$ CAH

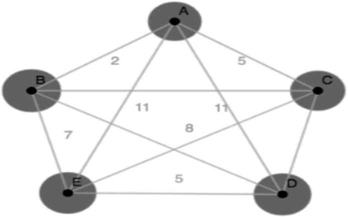
Tensor trains approximation for SBM inference

Model intuition

Data set

$$\begin{pmatrix} 0 & 2 & 5 & 11 & 11 \\ \vdots & 0 & \ddots & \ddots & 7 \\ \vdots & \vdots & \ddots & \ddots & 8 \\ \vdots & \ddots & \ddots & 0 & 5 \\ \vdots & \dots & \dots & \dots & 0 \end{pmatrix}$$

Representation as graph



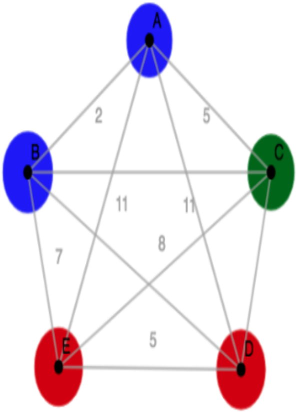
Model parameters

Statistical model : SBM

$$\boxed{\begin{matrix}\lambda \\ \alpha\end{matrix}}$$

$\Downarrow$

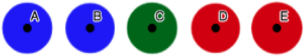
Z



Clusters inference

Let B be the number of classes

$$Z = (Z_1, \dots Z_n)^T \in \mathbb{M}_{n,B}([0,1])$$



Model Hypothesis

- H_1 Knowing Z, distances are independent
- H_2 The latent variables $Z_{i,i=1,\dots,n}$ are iid in $\{1, \dots B\}$.

J.-J. Daudin et al, A mixture model for random graphs, Stat Comput (2008)

Parameters estimation

- $\Lambda \in \mathbb{M}_{B,B}, \lambda_{b,b'}$: The parameter of Poisson probability to have a distance d between a vertex of class b and a vertex of class $b^{'}$.

$$\forall b, b' = 1, \ldots, B, \lambda_{b,b'} = Z_b^T \Lambda Z_{b'}$$

- $D_{i,j} | Z_{i,b} = 1, Z_{j,b'} = 1 \sim \mathrm{Pois}(\lambda_{b,b'})$
- $\alpha \in [0,1]^B, \alpha_i$: The probability of belonging to each cluster.
- Infering Z requires first to obtain $\hat{\theta} = (\hat{\alpha}, \hat{\Lambda})$. We proceed by $\hat{\theta}_{mv} = \mathrm{argmax}(P(D|\theta))$
- Having Λ, α and D, we chose the most probable configuration for Z.

EM algorithm

$$Z = (Z_1, \dots Z_n)^T \in \mathbb{M}_{n,B}([0,1])$$

- Infering Z needs to obtain $\hat{\theta} = (\hat{\alpha}, \hat{\Lambda})$ we proceed by $\hat{\theta}_{mv} = \text{argmax}(P(D|\theta))$
- The most naturel way is the EM algorithm. Each iteration involves two steps :
  - **E-step :** Compute : $Q(\theta, \theta^t) = \mathbb{E}_Z[\log P_\theta(D|Z)|\theta^t, D]$
  - **M-step :** $\theta^{(t+1)} = \text{argmax}_\theta Q(\theta|A, \theta^t)$

⚠ The main difficulty of the EM algorithm is to compute the marginals: It needs B^{n-1} sum.

Dempster et al, "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society (1977)

Inference approach

There are two main classes of methods :

Monte-Carlo methods : characterize a distribution by randomly sampling values out of the distribution.

- + Precision : Accurate
- — Computation time : Slow

Inference approach

There are two main classes of methods :

Monte-Carlo methods : characterize a distribution by randomly sampling values out of the distribution.

+ Precision : Accurate

— Computation time : Slow

Variational methods: we assume the independence of the nodes knowing the graph to approximate marginal by mean field.

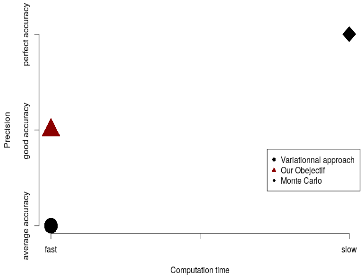
— Precision : Average accuracy

+ Computation time : Fast

Inference approach

- Our approach consists of using tensor trains to compute the marginals. It has already been adopted for Markov Random Field by Novikov.

Expectation of this approach :



+ Precision : Good to perfect accuracy

+ Computation time : Fast

Novikov et al, Putting MRFs on a Tensor Train (2014)

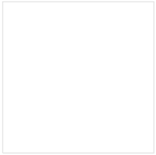
Tensor and tensors trains

Definition of a tensor

0 mode Tensor

1 mode Tensor

2 mode Tensor

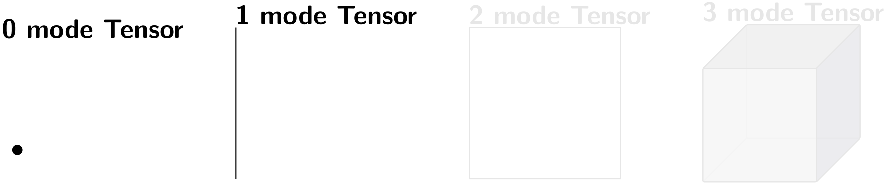


3 mode Tensor

A d-mode tensor is a multidimensional array $T \in \mathbb{R}^{n_1,\dots,n_d}$ where $n_i \in \mathbb{N}, i = 1, \dots, d$, with $T_{x_1,\dots,x_d} \in \mathbb{R}, 1 \leq x_i \leq n_i$

Tensor and tensors trains

Definition of a tensor



0 mode Tensor

1 mode Tensor

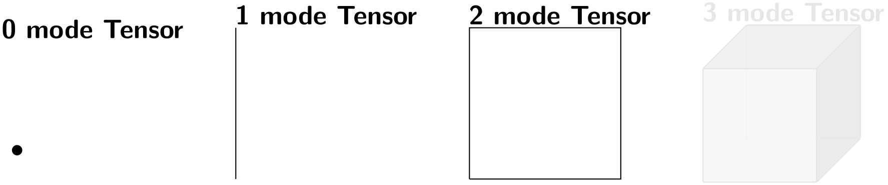
2 mode Tensor

3 mode Tensor

A d-mode tensor is a multidimensional array $T \in \mathbb{R}^{n_1,\ldots,n_d}$ where $n_i \in \mathbb{N}, i = 1, \ldots, d$, with $T_{x_1,\ldots,x_d} \in \mathbb{R}, 1 \leq x_i \leq n_i$

Tensor and tensors trains

Definition of a tensor



0 mode Tensor

1 mode Tensor

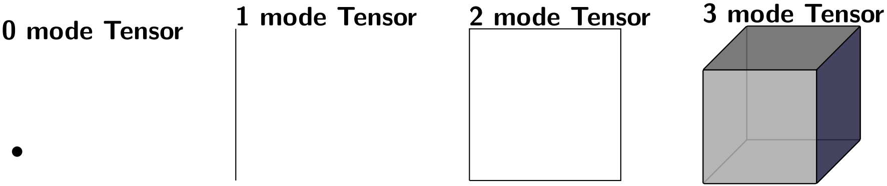
2 mode Tensor

3 mode Tensor

A d-mode tensor is a multidimensional array $T \in \mathbb{R}^{n_1,\dots,n_d}$ where $n_i \in \mathbb{N}, i = 1,\dots,d$, with $T_{x_1,\dots,x_d} \in \mathbb{R}, 1 \leq x_i \leq n_i$

Tensor and tensors trains

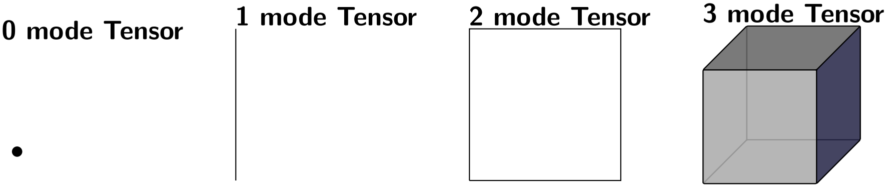
Definition of a tensor



A d-mode tensor is a multidimensional array $T \in \mathbb{R}^{n_1,\dots,n_d}$ where $n_i \in \mathbb{N}, i = 1, \dots, d$, with $T_{x_1,\dots,x_d} \in \mathbb{R}, 1 \leq x_i \leq n_i$

Tensor and tensors trains

Definition of a tensor



0 mode Tensor

1 mode Tensor

2 mode Tensor

3 mode Tensor

A d-mode tensor is a multidimensional array $T \in \mathbb{R}^{n_1,\ldots,n_d}$ where $n_i \in \mathbb{N}, i = 1, \ldots, d$, with $T_{x_1,\ldots,x_d} \in \mathbb{R}, 1 \leq x_i \leq n_i$

Tensor train

Tensor Train

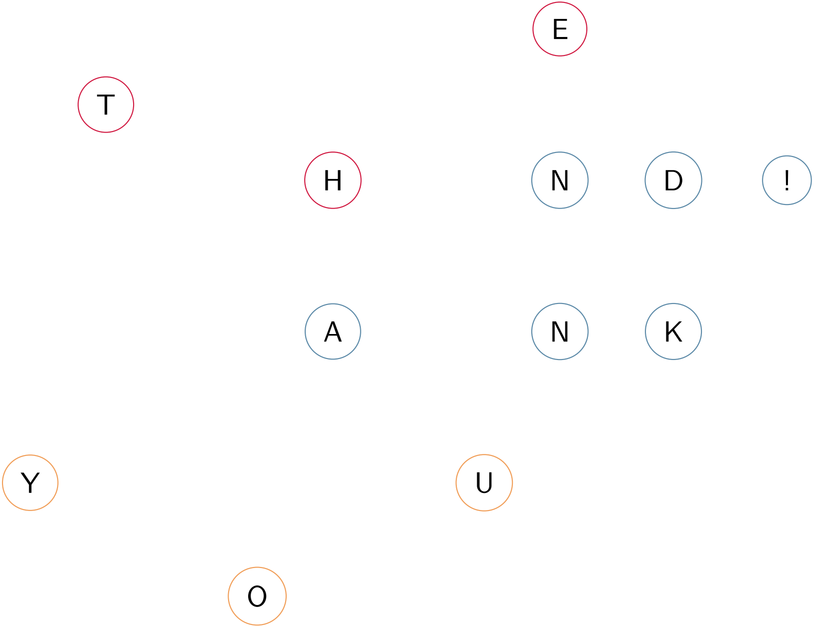
Apparoximation of the tensor T by another D such that $T \approx D$

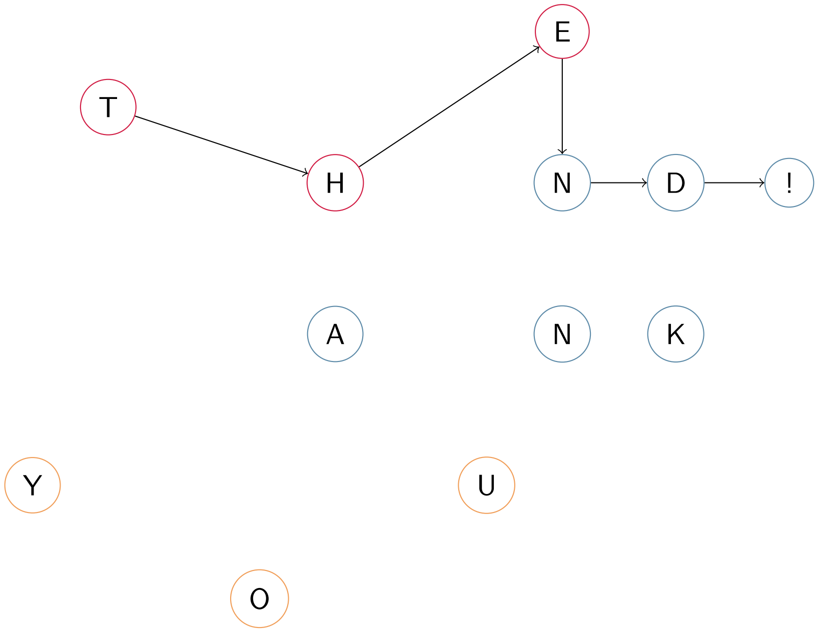
$$T(x_1, \dots, x_d) = G_1^T[x_1] G_2^T[x_2] \dots G_d^T[x_d]$$

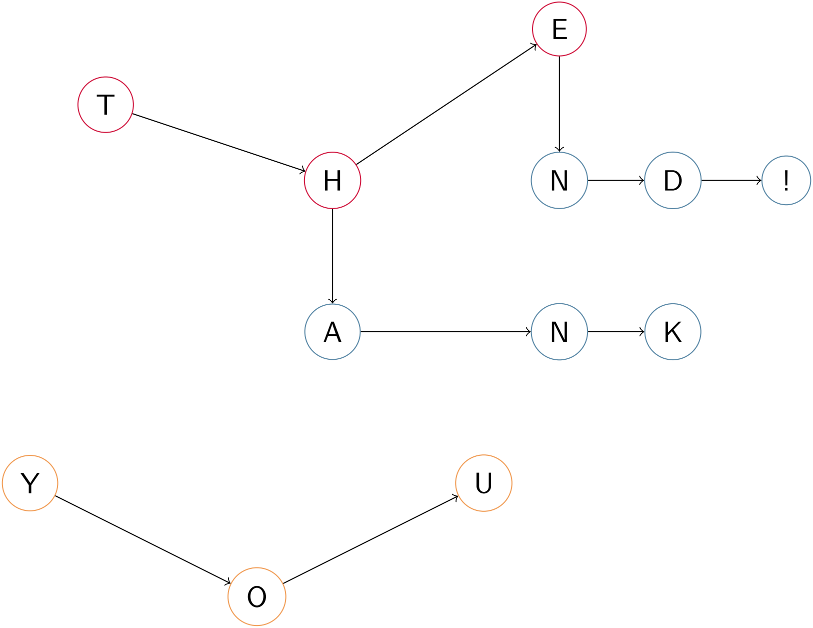
with $G_i^T[x_i] \in \mathbb{M}_{r_{i-1}, r_i}(\mathbb{R})$ and $r_d = r_0 = 1$

- $+$ Storage requires much less memory space
- $+$ Can be used for matrices
- $+$ Efficients operations

Oseledets, Tensor-Train Decomposition SIAM Journal on Scientific Computing (2011)


E
T
H
N
D
!
A
N
K
Y
U
O


E
T
H
N
D
!
A
N
K
Y
U
O


E
T
H
N
D
!
A
N
K
Y
U
O