

# MNHN-Tree-Tools

## Manual



Laboratoire de Biophysique  
Structure et Instabilité des Génomes  
Sorbonne Universités  
CNRS UMR 7196 / INSERM U\_II54  
43 rue du Cuvier,  
75231 Paris Cedex 05

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	License . . . . .	4
1.2	General Overview . . . . .	5
<b>2</b>	<b>Installation</b>	<b>6</b>
2.1	Generic Install . . . . .	6
2.2	Quick Install on Debian 10 . . . . .	7
<b>3</b>	<b>Tutorial</b>	<b>10</b>
3.1	Preparing your data . . . . .	10
3.2	Investigating k-mers and Principal Components . . . . .	11
3.3	Adaptive Clustering . . . . .	13
3.4	Tree Building, Tree Visualization and Tree Investigation . . . . .	17
3.5	Comparison to Ground Truth . . . . .	22
<b>4</b>	<b>Tools</b>	<b>28</b>
4.1	fasta2kmer . . . . .	28
4.2	kmer2pca . . . . .	30
4.3	pca_visual_extract . . . . .	33
4.4	cluster_dbSCAN_X . . . . .	34
4.5	compareSW . . . . .	39
4.6	find_satellite . . . . .	40
4.7	silhouette . . . . .	41
4.8	consens . . . . .	43
4.9	sequence_multiplicity . . . . .	44
4.10	lengths_from_fasta . . . . .	45
4.11	compare_norms . . . . .	46
4.12	adaptive_clustering_X . . . . .	47
4.13	split_set_to_fasta . . . . .	51
4.14	split_set_to_projections . . . . .	52

## CONTENTS

4.15split_sets_to_newick . . . . .	53
4.16split_set_to_matrix_X . . . . .	55
4.17split_set_from_annotation . . . . .	57
4.18split_set_from_swarm . . . . .	58
4.19filter_split_sets_by_min . . . . .	60
4.20find_sequence_in_split_sets . . . . .	61
4.21print_connections . . . . .	62
4.22tree_map_for_sequence . . . . .	63
4.23tree_map_for_split_set . . . . .	65
4.24virtual_evolution . . . . .	68
4.25simulation_verification . . . . .	74
4.26pca2densitymap . . . . .	75
4.27pca2densityfile . . . . .	77
4.28reverse_with_mask . . . . .	82
4.29reverse_complement_with_mask . . . . .	84
4.30digest_X . . . . .	85
4.31replace_N_sequence . . . . .	86
4.32tree_pureness . . . . .	88

<b>A Bibliography</b>	<b>94</b>
-----------------------	-----------

CONTENTS

# 1 | Introduction

This manual is written as a guide for scientists using and continuing to develop *MNHN-Tree-Tools*. As such this guide provides you with a installation procedure, detailed usage instructions and technical details for every tool contained within this suite and a tutorial to get the reader started with a typical clustering and density distance based phylogenetics as well as tree building.

## 1.1 License

Copyright (C) 2019-2020 Thomas Haschka

This software is provided 'as-is', without any express or implied warranty. In no event will the authors be held liable for any damages arising from the use of this software.

Permission is granted to anyone to use this software for any purpose, including commercial applications, and to alter it and redistribute it freely, subject to the following restrictions:

1. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.
2. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.
3. This notice may not be removed or altered from any source, binary distribution and this manual.

## 1.2 General Overview

*MNHN-Tree-Tools* was developed in order to cluster and infer phylogenetic evolution into repeated sequences in a single genome, namely:  $\alpha$ -satellites in primate genomes. During the development the versatility of the tools presented herein was both discovered and enforced. As such this tool is applicable to all kinds of data discovery tasks on datasets of sequences. Notebly this tool has been used for experiances around Deoxyribonucleic Acid (DNA) barcoding tasks Operational Taxonomic Unit (OTU) detection.

*MNHN-Tree-Tools* is a suite that allows you to gain insights into a dataset composed of nucleic or protein sequences provided as FASTA [1] files. The tool has the ability to cluster such a dataset and to built phylogenetic trees based on a density-distance based criteria. This is achieved by a repeated application of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [2] algorithm. Clustering and gaining such sequence-sequence dependency or phylogentic insights are at the heart of this software. Nevertheless this software provides you with all kinds of tools that might be of use, such as, but not limited to: The generation of fasta datasets using simulated evolution, the projection of FASTA files into a Principal Component Analyses (PCA) based subspace, the evaluation of k-mer vectors from fasta datasets or the calulation of clustering parameters such as the Silhouette [3] index on a dataset.

## 2 | Installation

### 2.1 Generic Install

In this chapter we will provide you with the details that allow you to install *MNHN-Tree-Tools*. The primary distribution method is in the form of source code on github:  
<https://gitlab.in2p3.fr/mnhn-tools/mnhn-tree-tools>

For a full installation a machine with a UNIX or UNIX-like system is required. The machine currently has to be of *x86\_64* architecture and support the *AVX* instruction set. In order to compile the full suite of tools you need a toolchain, consisting of:

- *make*: Which is required to build the suite and various subsets
- *compiler/linker*: A working C compiler, the suite was found to compile with *gcc* and *clang*.
- *BLAS/LAPACK* An implementation of the BLAS/LAPACK libraries.
- *OpenCL*: A working implementation of the OpenCL language. The suite was tested to compile with the *nvidia-opencl-toolkit* or the *intel-opencl-sdk*.
- *SDL2*: An implementation of the SDL2 library.
- *MPI*: An implementation of the Message Passing Interface (MPI). The suite was tested with *OpenMPI*.
- *PNG*: An implementation of the PNG graphics library.

You may make sure that you not only have the binary versions of the supporting libraries installed but also the header files, otherwise the software will not be able to compile. As such you may need to install the developer *-dev* versions of the libraries. In the next step you may clone the *MNHN-Tree-Tools* source code from github. You can do this by issuing the following command

in your UNIX terminal:

```
git clone https://gitlab.in2p3.fr/mnhn-tools/mnhn-tree-tools.git
cd mnhn-tree-tools
```

Listing 2.1 – Cloning the source from github.

Cloned and entered the mnhn directory you need to edit the head of the make file to suite your UNIX systems needs, select the correct compiler flags and linking commands to the libraries. You may consult the forums and documentation of your System/Distribution to find the correct values. In general *CC* is the command of your compiler, *MPICC* the MPI wrapper of your compiler. *CFLAGS* the options that you want to pass to your compiler. The other options are the flags required by the linker to find the appropriate libraries.

```
CC=gcc
MPICC=mpicc
CFLAGS= -O2 -march=native -ftree-vectorize -fomit-frame-pointer
LAPACK=-llapack
MATH=-lm
PTHREAD=-pthread
OPENCL=-lOpenCL
PNG=-lpng
MPI=-lmpi
SDL=$(shell sdl2-config --cflags) $(shell sdl2-config --libs)
```

Listing 2.2 – Head of the Makefile

With your flags adjusted you may compile the code issuing

```
mkdir bin
make all
```

Listing 2.3 – Building all tools

If you do not need the full functionality of *MNHN-Tree-Tools* you can install each tool individually by calling:

```
mkdir bin
make name-of-the-tool
```

Listing 2.4 – Building individual tools

This might especially be convenient if you are not able to install all the libraries in question on your system as the tool that you might really need might not depend on the library that you might not be able to install.

## 2.2 Quick Install on Debian 10

If you just want to get all tools working quickly we provide you with the following quick install script for *Debian 10* and derived

distributions:

```
sudo apt-get install git build-essential libpng-dev libsdl2-dev \
liblapack-dev libopenmpi-dev libpocl-dev ocl-icd-opencl-dev pocl-opencl-icd

git clone https://gitlab.in2p3.fr/mnhn-tools/mnhn-tree-tools.git

cd mnhn-tree-tools
mkdir bin
make all
cd bin

# make MNHN-TREE-TOOLS available from any folder ( this is temporary, you
# may modify your .bashrc and similar files to make this permanent
export PATH=$PATH:$PWD
```

Listing 2.5 – Installing MNHN-Tree-Tools on Debian 10

## CHAPTER 2. INSTALLATION

# 3 | Tutorial

In this chapter we will walk you through basic manipulations of nucleic datasets, tree building, naming conventions and overall usage to get you started with *MNHN-Tree-Tools*.

## 3.1 Preparing your data

At first let us get started with a dataset that you might have. For this tutorial we recommend you to have a dataset of reasonable size at hand. While these tools work and have been tested for millions of sequences we recommend you that you use a dataset of about 5000 to 10000 sequences for this tutorial to progress fast and learn how to use these tools quickly. If you have no dataset available you can generate one with our built in virtual biosphere generator the *virtual\_evolution* tools shown in section 4.24. To generate such a dataset you might run:

```
virtual_evolution_controlled 15 100 10000 2 -1 50000 5 test.fasta
```

Listing 3.1 – Generate a test dataset using the *virtual\_evolution* tools

which should yield a decent dataset to discover *MNHN-tools*.

The tools presented herein are built with FASTA [1] files in mind. If you have your sequence data available in a different file format you may need to convert it beforehand. If you bring your own FASTA file for this tutorial please note that the *MNHN-tools* are only able to treat human readable non single line FASTA files. As such a sequence may not contain more than 1000 characters per line. If your file contains lines that are longer then the 1000 character limit you may convert it to a conforming FASTA file using the *seqret* tool from EMBOSS [4]. Several application domains of *MNHN-Tree-Tools* are used to use FASTQ files. If

you have datasets in this format you may use the following *bash* script to convert your FASTQ files to FASTA files.

```
#!/bin/bash
sed -n '1~4s/^@/>/p;2~4p' $1 > $2
```

Listing 3.2 – Script to convert FASTQ to FASTA files

In such a case it might also be necessary to apply *seqret* from EMBOSS on the FASTA files resulting from this conversion in order to ensure appropriate line lengths.

Before using a dataset with *MNHN-Tree-Tools* you further may make sure that it is suited for the task that you would like to achieve. The tools presented herein are in general made to compare sequences of at least a reasonable similarity. As calculating the consensus of unaligned datasets consisting of millions of sequences might be cumbersome, we suggest to examine the variation in sequence lengths first. The *lengths\_from\_fasta* tool is perfect for this task (c.f. section 4.10) and may be run like:

```
lengths_from_fasta your.fasta > lengths
```

Listing 3.3 – *lengths\_from\_fasta* tutorial

The tool yields all the sequence lengths for the sequences of your dataset on separated lines in the *lengths* file. We leave it to the reader to analyze with the diverse tools available to find out whether the lengths do not have a variance that is reasonable in respect to the average sequence length.

## 3.2 Investigating k-mers and Principal Components

A convenient way to transform a set of sequences into an alignment free representation, hence a representation where we do not need to align sequences in order to calculate their distance or compare them is the *k-mer* representation. In the *k-mer* representation we compare all possibilities of sequences of certain *k* length with a sequence in question and count how often such a sequence appears in our sequence of question. Let us imagine that *k* equals four. In such a case we obtain 256 possibilities to form a sequence (out of nucleotides) or  $4^4$  sequences. As we count how often any of these sequences occurs in our target sequence, we can represent the target sequence as a vector of 256

k-mer frequencies, a vector holding the counts how often each sequence occurs in the target sequence for the 256 k-mers. In order to perform this transformation we run the *fasta2kmer* tool (c.f. section 4.1). In our case we are not generating 4-mers, but 5-mers for our sequences and thus vectors of  $5^4 = 1024$  values in length:

```
fasta2kmer test.fasta 5 2 0 > test.kmer
```

Listing 3.4 – Generating 5-mers with *fasta2kmer*

on a supposedly dual core machine, and hence we are selecting 2 threads. To choose the right *k-mer* size you may be advised to use a length that captures the information in your sequences. Hence the longer your sequences are, the higher the  $k$  of the sequences shall be. But beware every increase in  $k$  also generates an increase of  $k^4$  data and therefore we strongly discourage using  $k > 7$  if the length of your sequences permits such a choice. To encode enough information it is suggested that you have as much k-mers at hand as your sequence is in length. Thus armed with a k-mer length of  $k = 4$ , for instance we shall be fine for sequences of up to 256 nucleotides. Typing

```
head -n 2 test.kmer
```

Listing 3.5 – Investigating 5-mers with head

allows us to investigate the first two lines of our calculated *k-mers* file and as such the representation in k-mer form of the first two sequences of a supplied FASTA dataset *test.fasta*. Having the k-mer representation of our sequences at hand we can perform a dimensionality reduction step, selecting a representation in a subspace that still covers the variance of dataset, a method called principal component analysis (PCA). For in depth details we refer you to your favorite *linear algebra* textbook. To perform PCA on the dataset we have the *kmer2pca* tool (c.f. section 4.2) at hand. Applying the tool on our *test.fasta* dataset together with its k-mer representation *test.kmer*, we are able to perform a PCA like:

```
kmer2pca test.kmer test.pca test.ev 7 2
```

Listing 3.6 – Performing PCA on the k-mer representation of our sequences

and are using our supposed dual core machine, calculating the projections onto the seven first principal components with the

largest corresponding eigenvalues. The eigenvalues printed in `test.ev` allow us to get an idea of the information variance that we capture in our 7 dimensional PCA subspace. A good measure is to perform the sum of all eigenvalues, and then to sum up, in this case the first 7 and calculate the percentage of that the sum of the first seven represents over the sum of all eigenvalues. Even if this value might seem low, subspaces with a dimensionality that is larger than 7 can lead to the *curse of dimensionality* in the subsequent clustering parts of this tutorial, and we might note that even with a capture below 30% our algorithm still yielded interesting results.

But first let us investigate the PCA subspace, and the sequences in it. In order to do this we can use the *gnuplot* program [5]: In the gnuplot shell one can call:

```
plot 'test.pca' using 1:2
plot 'test.pca' using 1:3
splot 'test.pca' using 1:2:3
```

*Listing 3.7 – Plotting PCA with gnuplot*

where the `1:2` represents the first and second component. As our subspace is 7 dimensional we can plot 2 or 3 dimensional projections. The principal components represent our axes and as it is a linear projection and rotation the units are *k-mer* frequencies. Generating images like the ones shown in figure 3.1: As outlined by the figure it might not be evident to detect different clusters ( in this case 2 as the projection shows a dataset like the one generated with the command shown in listing 3.1 ). For a deeper inspection you might refer to the *pca2densityfile* and *pca2densitymap* tools shown in sections 4.27 and 4.26 respectively. Besides this tool we further have a PCA sequence selector, that allows us to extract and investigate sequences, from dense ensembles, in literature also cited as comets [6], and directly access the sequences. This tool is highlighted in section 4.3.

### 3.3 Adaptive Clustering

One core feature of the *MNHN-Tree-Tools* is an adaptive density based clustering algorithm that can be used to study phylogenetic features and the evolution of species by investigating typical DNA barcoding targets such as the 16S or 18S ribosomal RNA or Cyclooxygenase 1 (COX1) coding regions. Further

### 3.3. ADAPTIVE CLUSTERING

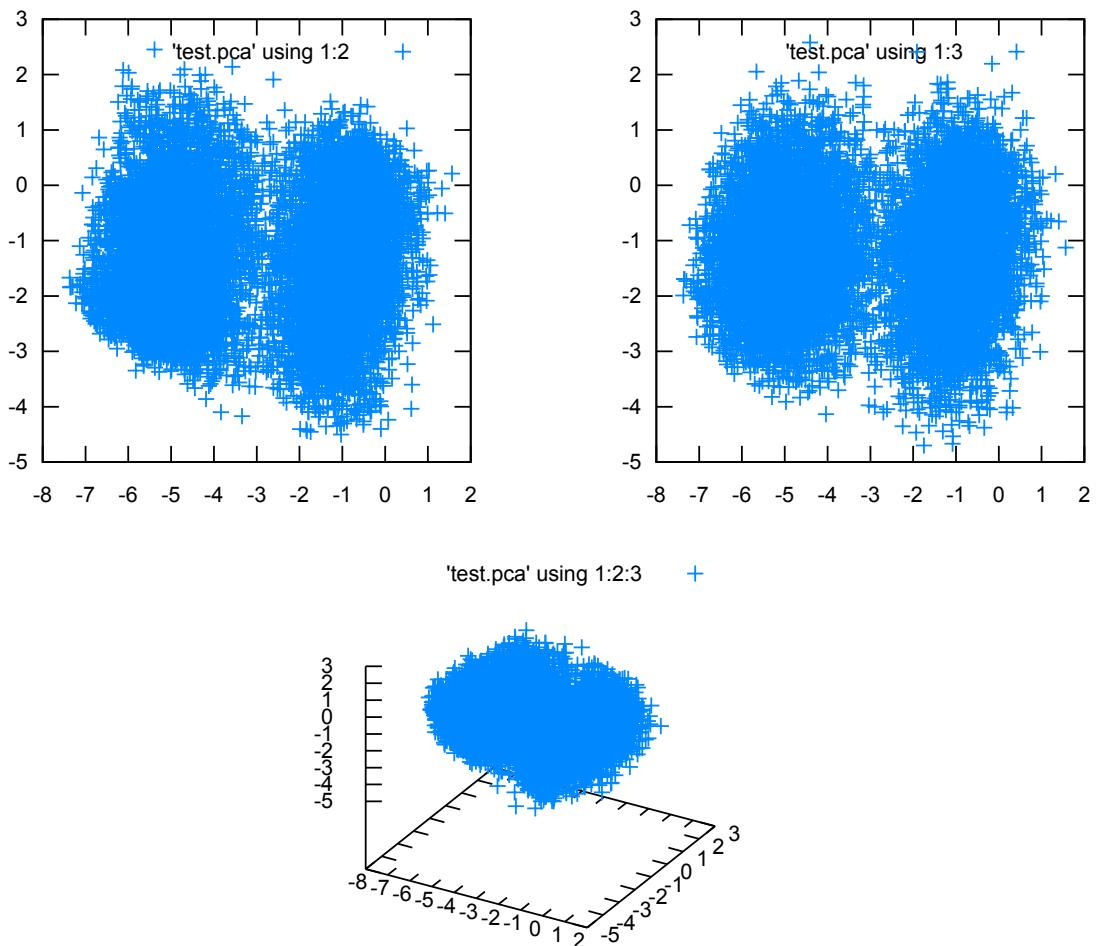


Figure 3.1 – Example plots of PCA projections, upper left components 1 and 2, upper right 1 and 3 and lower 1,2 and 3

the tool might be of use for all kinds of purposes where similar sequence clusters are to be investigated. As such *MNHN-Tree-Tools* was successfully applied in the quest to discover families of  $\alpha$ -satellites, repeated sequences, in the human genome.

Adaptive clustering is implemented in the MNHN tools with different underlying sequence distance measures and optimized for different computing architectures. As we have already highlighted the PCA tools during the tutorial and for the sake of the rapidity of the computation we show how we can use the PCA based adaptive clustering method in order to investigate our dataset of sequences:

```
mkdir clusters
adaptive_clustering_PCA test.fasta 1 0.001 3 clusters/L 2 7 test.pca > cl-out
```

Listing 3.8 – Performing PCA based adaptive clustering

Here we iteratively use the DBSCAN algorithm [2] from a starting value of  $\epsilon_{int} = 1$  and increasing in it each step by  $\Delta\epsilon = 0.001$ . Together with the constant *minPoints* of 3 value this forms different density limits. The scanning algorithm will find us sequences above these density limits, and as the density decreases in each step regions that house connected sequences above a certain density are growing and will merge, and hence, denser regions are found to be in less denser regions. This denser regions in less denser regions allow us the interpret dependencies, build dendograms and investigate evolution. A simple illustration of the algorithm is shown in figure 3.2. For a more detailed overview the reader is referred to section 4.12. The command in listing 3.8 yields us with an output file *cl-out* and a bunch of binary *split-sets* in the *cluster* folder. A binary *split-set* is a file that holds a partitioning of our dataset. As we have run an adaptive clustering run, several different partitions for different densities have been returned by the algorithm. These will later on be the layers of our tree. Hence we named them L, and the tool returned us files that hold the name like *L0000*, *L0001* etc. indicating the layer of the tree. From the most outer leaf node *L0000* to the most inner root *LNNNN* with *NNNN* being the highest number available. For each of the clusterings we also know, thanks to the data that we stored in *cl-out* which binary *split-set* belongs to which  $\epsilon$  and hence to which density limit.

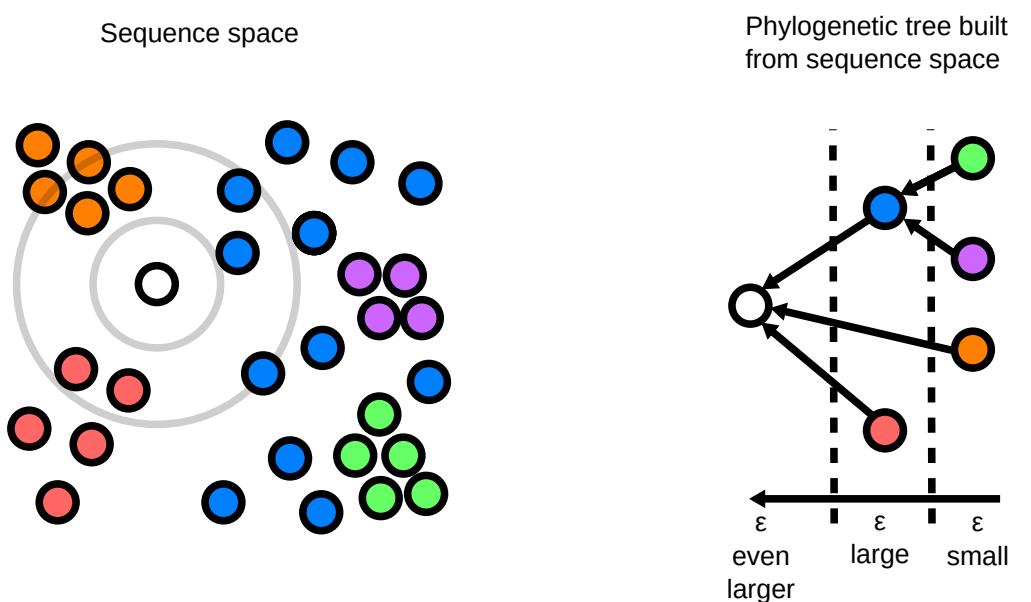


Figure 3.2 – The adaptive clustering algorithm finds dense (new) clusters in less dense (old) more dispersed clusters. Finally the *MNHN-Tree-Tools* collection allows us to build a tree from like the one on the right, from a sequence space as the one on the left.

## 3.4 Tree Building, Tree Visualization and Tree Investigation

Having our *split-sets* from an adaptive clustering run at hand, and having looked already at the *cl-out* file we already have guessed what is next. In the *cl-out* file different there is at the lower and a list of connections between different clusters at different layers. The next step is to build a tree, or a dendrogram representing the density structure of our dataset and hence obtain the left side of figure 3.2. *MNHN-Tree-Tools* comes with a tool to infer a Newick tree [7] from the clusters obtained during an adaptive clustering run. Here we run the tool *split\_sets\_to\_newick* outlined in section 4.15 using the results from our adaptive clustering run:

```
split_sets_to_newick 0 0 clusters/L* > tree.dnd
```

Listing 3.9 – Generating a Newick tree with *split\_sets\_to\_newick*

and obtain a Newick tree in the file *tree.dnd*. You can use a tool of choice to plot the Newick tree, but herein we show how to do it with the Newick tools by the University of Geneva [8].

```
nw_display -s -b 'opacity:0' -w 600 tree.dnd > tree.svg
```

Listing 3.10 – Using the Newick tools [8] to plot a dendrogram

where the options are *-s* in order to generate a *scalable vector graphics* (SVG) file, *-b 'opacity:0'* is here to hide the number indicating the distance between the clusters. *-w* the size, width of the image. The reader shall be referred to the Newick tools manual for all the vast possibilities of options. If you use your own dataset you might have to adjust the width so that all the labels are visible. In general you should be presented with a tree like the one shown in figure 3.3. Here in the image which is the tree of clusters obtained from our virtual biophere, generated by the command in listing 3.1 where two partitions with a distance of at least 5 nucleotides was requested. One of these underwent 10000 single nucleotide mutations and the other 20000, and hence both partitions diffused, as seen in the figures of the PCA 3.1 and as it can be seen in the figures of the tree 3.3.

Looking at figure 3.3 Lets say we would like to identify the clusters *L5C0* containing 3216 sequences and *L5C6* containing 3784 clusters. Both at the forefront of two large subtrees. In order to access the sequences in these clusters we have to convert layer 5 to FASTA files. Layer 5 exists by *MNHN-Tree-Tools*

### 3.4. TREE BUILDING, TREE VISUALIZATION AND TREE INVESTIGATION

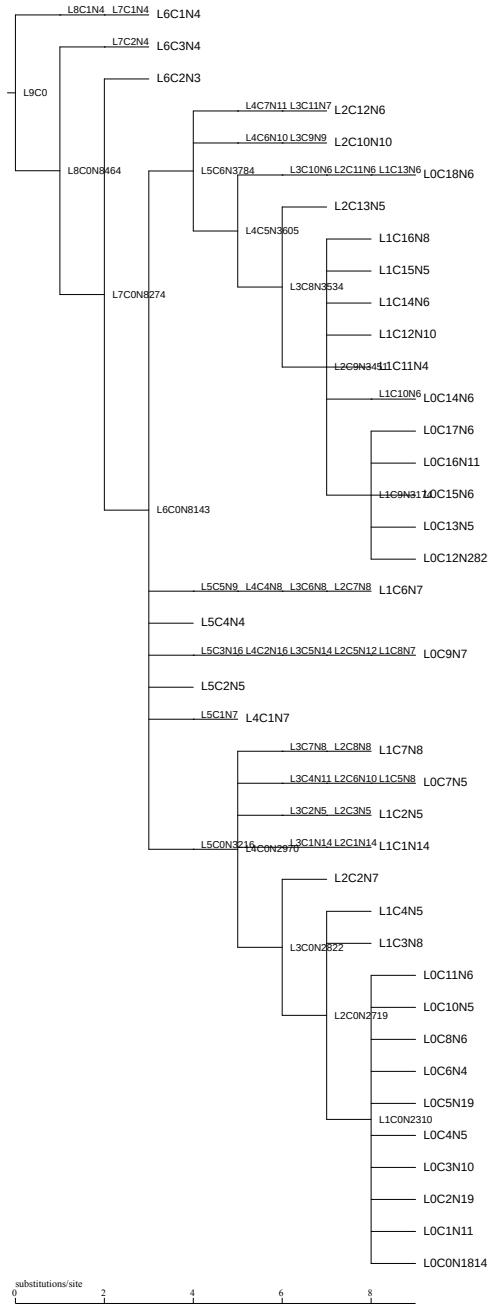


Figure 3.3 – A Newick tree with the cluster labels visible, each cluster is labeled by L<sub>X</sub>C<sub>Y</sub>N<sub>Z</sub> where X is the layer number, Y the cluster number and Z the number of sequences in that cluster. The substitutions/site are automatically added by the Newick tools and have in our case no significance.

convention in a binary *split-set* file named with the suffix number *0005*, hence in our case, as demanded by the command in listing 3.8 named in the folder *clusters* named *L0005*. In order to convert the sequences of layer 5 to *FASTA* files we use the tool *split\_set\_to\_fasta* as described in section 4.13. We note here to the reader, that his personal tree, even using our commands might slightly differ due to different random number generator mechanisms on different systems, and that he might have to adapt cluster and layer numbers to the tree he has obtained.

The *split\_set\_to\_fasta* tool is called here like:

```
mkdir fasta-layer-5
split_set_to_fasta test.fasta clusters/L0005 fasta-layer-5/C
```

### Listing 3.11 – Converting a binary *split-set* to *FASTA* sequences

and hence we obtain the *FASTA* files for layer 5. The files for our clusters *L5C0* and *L5C6* are respectively stored in the files *fasta-layer-5/C-000* and *fasta-layer-5/C-006* in *FASTA* format. Using the same mechanism we can identify any sequence anywhere in the tree. Now one could open these files for instance with *SEAVIEW* and throw all kind of statistical tools, or our *consens* tool highlighted in section 4.8 on these files.

But let us do something more interesting let us identify the two clusters in our PCA diagram, such as the one shown in figure 3.1. In order to do this we can use a tool called *split\_set\_to\_projections*, further described in section 4.14. We run this tool likewise:

```
mkdir projections-layer-5
split_set_to_projections test.fasta test.pca 7 clusters/L0005 \
    projections-layer-5/p
```

### Listing 3.12 – Converting a binary *split-set* to PCA projections

Analogous to the *FASTA* files we obtain the files containing the PCA projections of these clusters. Again in our case the projections for *L5C0* and *L5C6* reside in files *p-000* and *p-006* in folder *projections-layer-5*. We can right now use the *gnuplot* [5] software to locate them:

```
plot 'test.pca' using 1:2 lt rgb '#333333', \
    'projections-layer-5/p-000' using 1:2 lt rgb '#FF8800', \
    'projections-layer-5/p-006' using 1:2 lt rgb '#0088FF'
```

### Listing 3.13 – Highlighting cluster projections with *gnuplot*

which generates in our case the figure 3.4. Note that as the algorithm is written in order to infer the tree structure and consensus (center of gravity) of clusters / sequence families, sometimes the

### 3.4. TREE BUILDING, TREE VISUALIZATION AND TREE INVESTIGATION

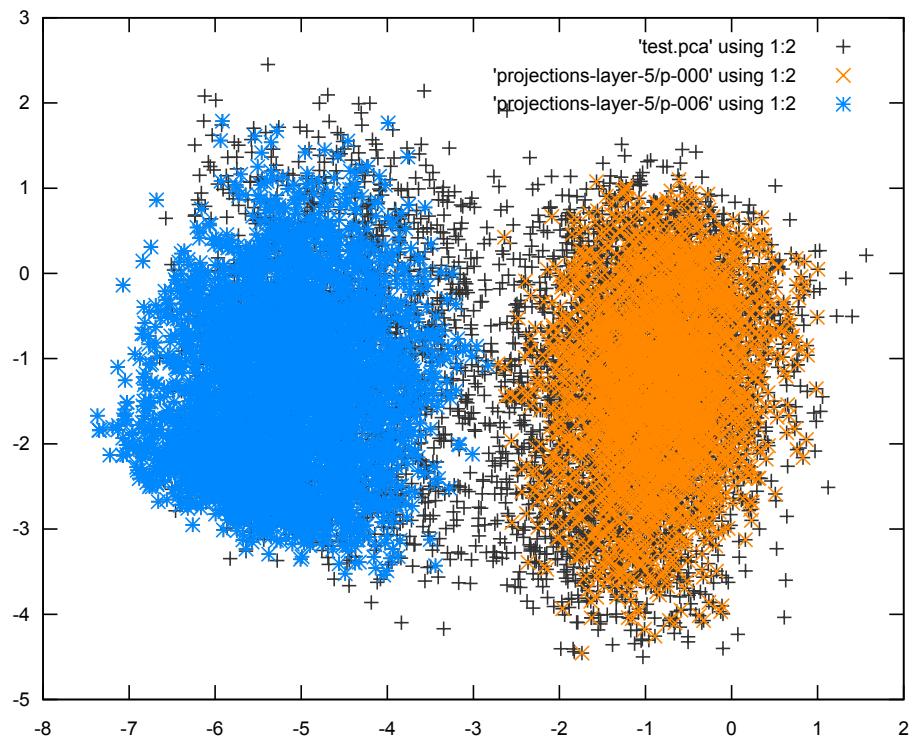


Figure 3.4 – The projections of cluster  $L5C0$  in orange, and the projections of cluster  $L5C6$  in blue projected onto the first two principal components

clusters might appear small when projected. In this case you can search for a more optimal cluster size varing the  $\epsilon$  value around the layer in question. You can find this value in the output file, in our case *cl-out* as requested in listing 3.8. You may find lines like the following:

```
Layer 5 has 7 clusters
Coverage at layer 5: 0.704100
Epsilon at layer 5: 1.100000
```

**Listing 3.14 – Layer 5 description in our adaptive cluster run output file**

As the algorithm is tree preserving. If you have the feeling that the clusters are to small you can run a single *DBSCAN* run [2] with the tools *cluster\_dbscan\_X* tools shown in section 4.4, choosing an epsilon value just below the one of the following layer, which in this case is described to be:

```
Layer 6 has 4 clusters
Coverage at layer 6: 0.815400
Epsilon at layer 6: 1.180000
```

**Listing 3.15 – Layer 6 description in our adaptive cluster run output file**

This way more defuse sequences might be taken into account, just before the clusters merge due to the  $\epsilon = 1.18$  density critera. You can hence if you like oversample the tree.

Let us further investigate the structure of the tree, and let us say we would like to see how a single sequence flows through the tree, and where it appears in the tree. In order to achieve such a task we have the *tree\_map\_for\_sequence* tool shown in section 4.22. Lets say we are interesed to find the following sequence that resides in our initial dataset in our tree:

```
>sequence_6773
CGGCTTACGAACTGCCGTTGTTCACCTGGGCACAGTGTACCAACGCAA
AGATGGAGTTACTCGCGAACAAATAGGGAGCGTTGTCCGGTGAGTTAG
```

**Listing 3.16 – A sequence of interest**

We hence save this sequence into a separated FASTA file called *target.fasta*. We can now use the the *tree\_map\_for\_sequence* tool to create a map for the Newick tools by issuing the following command:

```
tree_map_for_sequence target.fasta test.fasta '#FF8800' clusters/L* > target.map
```

**Listing 3.17 – Mapping the *target.fasta* sequence in the Newick tree**

This allows us to color the clusters in the tree that contains the target sequence. To generate a SVG file you can use the Newick tools command:

```
nw_display -s -b 'opacity:0' -w 600 tree.dnd -c target.map > tree-seq6773.svg
```

Listing 3.18 – Drawing the dendrogram highlighting the sequence in *target.fasta*

which results in a dendrogram as shown in figure 3.5.

## 3.5 Comparison to Ground Truth

In this section we show you how you can investigate not just the location of one single sequences but of a whole ensemble of sequences within the tree obtained by an adaptive clustering run. If you know about certain sequences that form a group in your own dataset you can follow us on this route, generating your own clusters that you might want to investigate in your own trees. The dataset generated with the command highlighted in listing 3.1 is partitioned as we could also suggest from the tree into two subgroups. The *virtual\_evolution* (c.f. section 4.24) tool has generated two partitions of 5000 sequences. As the partitions are written out sequential by the *virtual\_evolution* tool the first 5000 sequences belong to the first partition, the next 5000 sequences belong to the second partition. We can thus generate an artificial clustering or binary *split-set* in order to further investigate the two known partitions with the *MNHN-Tree-Tools*. In order to generate such a *split-set* we will use the *split\_set\_from\_annotation* tool outlined in section 4.17. In order to do this we first have to generate the annotation file. On the *bash* [9] shell on a Unix/Linux system you can do this in the following way:

```
for ((i=0;i<5000;i++)); do echo "one" >> gt-annotation; done
for ((i=0;i<5000;i++)); do echo "two" >> gt-annotation; done
echo "one" >> unique-gt-annotation
echo "two" >> unique-gt-annotation
```

Listing 3.19 – Generationg an annotation file for *split\_set\_from\_annotation*

which creates us the *gt-annotation* file that contains 5000 lines with just the single word *one* on each line followed by another 5000 lines with the single word *two* on each line. We further create the *unique-gt-annotation* file containing just two lines with

### CHAPTER 3. TUTORIAL

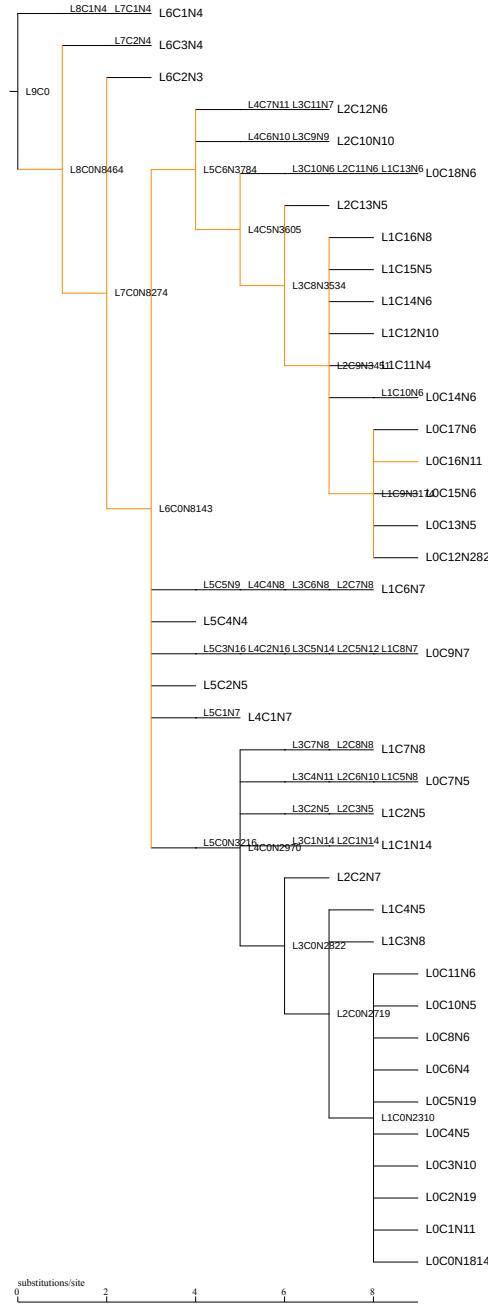


Figure 3.5 – The tree, tracing down a target sequence highlighted by the *tree\_map\_for\_sequence* tool.

the statements *one* and *two*. The *unique-gt-annotation* file is used, first to describe that statements that indicate the clusters to be formed, in our case *one* and *two* and second to order the clusters in the resulting *split-set* file. As *one* fills the first line, and *two* the seconds in the *unique-gt-annotation* file we know that sequences marked as *one* in the *gt-annotation* file will be part of the first clusters while those marked to be *two* will be part of the second cluster. With these files at hand we can execute the *split\_set\_from\_annotation* command:

```
split_set_from_annotation test.fasta gt-annotation unique-gt-annotation \
ground-truth
```

**Listing 3.20** – Creation of a *split-set* cluster file by manual annotation

which generates us the ground truth binary *split-set* indicating the two clusters.

A question that one might ask, knowing that the first 5000 sequences form a partition and the second 5000 sequences form an other partition as created by using our *virtual\_evolution* tools: Where are we going to find these sequences in the tree. In order to investigate such a matter we can use the *tree\_map\_for\_split\_set* tool as shown in section 4.23:

```
tree_map_for_split_set ground-truth test.fasta 1 2 '1,#000000' \
clusters/L* > tree-one.map
tree_map_for_split_set ground-truth test.fasta 1 2 '2,#000000' \
clusters/L* > tree-two.map
```

**Listing 3.21** – Creating a Newick tools [8] map file for a *split-set*

which yields us the *tree-maps* that we can use with the Newick tools [8] to highlight different the different datasets in the trees. In order to do this we generate two SVG files for each tree executing:

```
nw_display -s -b 'opacity:0' -w 600 tree.dnd -c tree-one.map > tree-one.svg
nw_display -s -b 'opacity:0' -w 600 tree.dnd -c tree-two.map > tree-two.svg
```

**Listing 3.22** – Using the Newick tools [8] to visualize partitions

together with the dendogram created in listing 3.9. The two resulting trees are shown in figure 3.6. Finally we can quantize if our tree correctly maps the two clusters by applying our *Pureness* calculation. This allows us to check weather we obtain the same clusters as the ground truth and weather clusters are a mixture of the ground truth or not. For the details of the Pureness, and other indices to evaluate the quality of a tree against ground

### CHAPTER 3. TUTORIAL

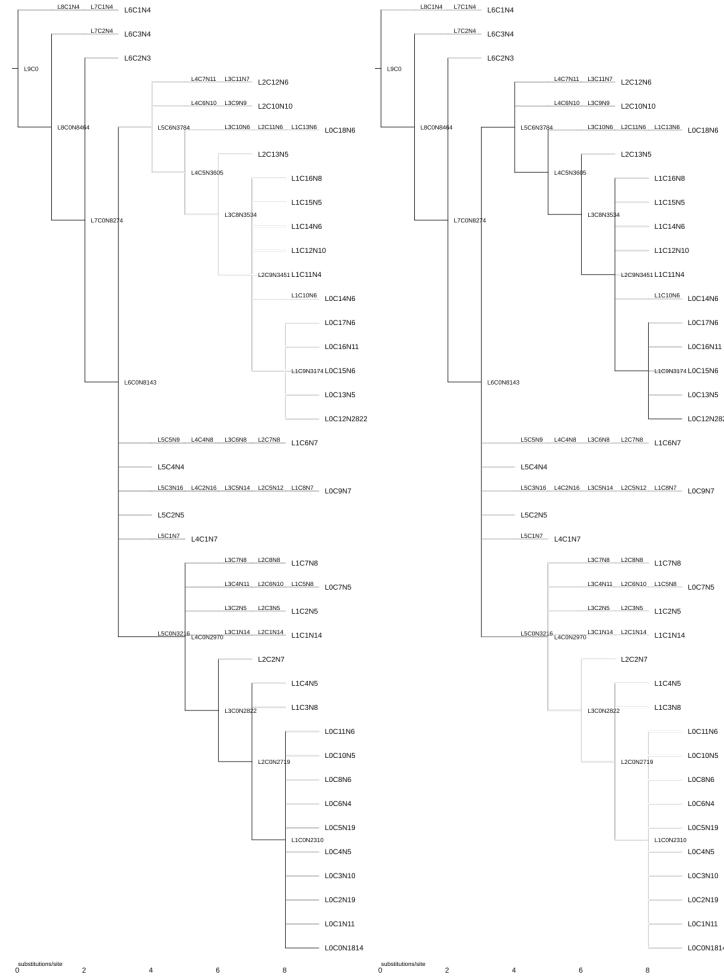


Figure 3.6 – Tracing down clusters using the *tree\_map\_for\_split\_set* tool. On the left we have the first partition while on the right the second partition is shown. Using the logarithmic representation we can also easily track down the core clusters *LOC0* and *LOC12* that have generated these two families, forming these two subtrees (lower on the right, upper on the left).

truth the reader is referred to section 4.32. In order to perform this quality measurement we execute the *tree\_pureness* tool:

```
tree_pureness test.fasta ground-truth clusters/L* > pureness
```

**Listing 3.23 – Evaluating a tree against ground truth with the *tree\_pureness* Tool**

from where we obtain the following table:

0.001103	8.500000	0.052632
0.250433	7.500000	0.117647
0.134946	6.000000	0.214286
0.002904	5.000000	0.166667
0.002793	3.000000	0.250000
0.002923	2.500000	0.285714
0.978448	1.000000	0.500000
0.981508	0.500000	0.666667
0.969282	0.000000	0.500000
0.973321	0.500000	1.000000

**Listing 3.24 – The output of the *Pureness* table**

where the first column is the Purness as in equation 4.23, the second the Cluster Correspondance according to equation 4.24 and finally the Impure over Pures as outlined in equation 4.25. The table shows the values for each layer from the leaf nodes on top to the root of the three on the bottom. We see that in Layer 5 (we count from the top, and note that the first line corresponds to Layer 0) we have a very good Pureness of 0.003 which counts only for the already impure clusters, and that only about quater 28% of the clusters ( as shown in the third column ) are impure. The middle column shows that at this moment the tree, nevertheless is not only composed of two Datasets, but contains more clusters in this Layer. A quick look at the Tree in figure 3.3 shows that these clusters have very few elements and have formed due to the fluctuation of diffusion. They would have been filtered away by a more agressive *minPoints* value in the adaptive clustering run.

CHAPTER 3. TUTORIAL

# 4 | Tools

In this chapter we describe each tool in detail, how to use it, the algorithm used to implement it, and algorithmic specificities.

## 4.1 *fasta2kmer*

### 4.1.1 General

The *fasta2kmer* tool implements the calculation of k-mer frequencies for each sequence to be found in a FASTA [1] file. The tool can calculate both k-mer frequencies for nucleic and protein sequences. k-mers are all possible sequences for k letters. The frequencies are how often each of these k-mers is found in the entire sequence.

### 4.1.2 Usage

```
fasta2kmer [fasta-file] [kmer-length] [number-of-threads] [protein?] > kmerbase
```

Listing 4.1 - Calling the *fasta2kmer* tool

The *fasta2kmer* tool has the following arguments.

1. *fasta-file* The fasta file containing the sequences to calculate k-mer frequencies of.
2. *kmer-length* The length of the k-mers to calculate frequencies from.
3. *number-of-threads* The number of individual threads to be available for this computation. A good choice is the number of logical cores available on the machine in question.
4. *protein?* Set 0 for nucleic sequences, 1 for protein sequences.

5. *kmerbase* The output file where to write kmer frequencies for each sequence to. The format of this file is tablulator seperated. The first entry in each line is writes *sequence\_N* where N is the sequence number attributed by counting from 0 to N by occurance of sequences in the FASTA file. The following entries the line, purely numeric are the frequencies of the k-mer in questions.

### 4.1.3 Example

```
fasta2kmer test.fasta 5 8 0 > test.kmer
```

Listing 4.2 - Example call *fasta2kmer* tool

In this case one calculate 5-mer frequencies for the file test.fasta using 8 threads. test.fasta contains nucleic sequences. The output ist stored in test.kmer.

### 4.1.4 Algorithm

The nucleic k-mers are calculated using a binary scheme. As the nucleic letters contain only four members A, C, G, T they can be represented by 00, 01, 10, 11 in binary form. To generate all possible k-mers we generate a binary number of  $2k$  in length containing only ones. I.e. for  $k = 3$  we generate 111111 = 63 and count down to 0 generating all 64 3-mers by comparing the binary represenation with the letter code.

The protein code uses a different scheme generate proteic k-mers. A recursive function called  $k$  times shifting position counting to from 0 to 19 is called generating all possible protein k-mers for the length  $k$ .

The frequencies are calculated by comparing the code at each overlay position of the sequence to calculate frequencies for. An early stop condition for a mismatching letter is implemented. The comparison is performed multithreaded distributing the number sequences to calcualte k-mer frequencies from accross the number of threads requested.

### 4.1.5 Implementation

The algorithms for the tool are implemented in *kmers.c* The interface to the algorithms is implemented in *fasta2kmer.c*

## 4.2 kmer2pca

### 4.2.1 General

The *kmer2pca* tool implements the calculation of and projection onto principal components. The tools yields the projections onto the principal components as well as the eigenvalues of covariance matrix.

### 4.2.2 usage

```
kmer2pca [kmer-file] [projections] [eigenvalues] [dimensions] [n-threads]
```

Listing 4.3 – Calling the *kmer2pca* tool

The *kmer2pca* tool has the following arguments.

1. *kmer-file* A file containing k-mer frequencies as calculated by the *fasta2kmer* tool presented in section 4.1
2. *projections* The file where projections onto the principal components will be written to. This file contains tabulator seperated values, each line corresponds to the line in k-mer frequencies input file.
3. *eigenvalues* The file where the eigenvalues of covariance matrix are written to. The eigenvalues are sorted from smallest to largest.
4. *dimensions* The number of principal components to project on. The number of dimensions of the resulting subspace. The k-mers are projected onto the [dimension] principal components corresponding to [dimension] highest eigenvalues.
5. *n-threads* The number of threads to be made available for this computation. A good choice is the number of logical cores available on the machine.

### 4.2.3 example

```
fasta2kmer test.kmer test.pca test.ev 7 8
```

Listing 4.4 – Example call *kmer2pca* tool

In this example we calculate the projections into a seven dimensional suspace spun by the 7 principal components with the highest variance. The projections of the k-mers stored in test.kmer onto these 7 principal components are written to test.pca. The eigenvalues of the covariance matrix built from the kmers in test.kmer are written to the test.ev file. 8 threads are used to perfromed to built the covariance matrix.

#### 4.2.4 algorithm

The PCA is straight forward implemented in its most common form. We have the vectors of k-mer frequencies for each sequence. We hence can call each k-mer a features, of which we have the number of sequences samples. In order perform a PCA we first need to obtain a feature covariance matrix, and hence calculate the covariance entre k-mers for the dataset of sequences that have. We first calculate the mean of all k-mers, let us call a single k-mer  $X$ :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (4.1)$$

where  $n$  is the number of sequences. For two different k-mers, let us call them  $X$  and  $Y$  we can than calculate the covariance:

$$\sigma_X \sigma_Y = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}). \quad (4.2)$$

and further built the covariance matrix for all available covariances from all available k-mers.

$$C = \begin{bmatrix} \sigma_1 \sigma_1 & \cdots & \sigma_1 \sigma_n \\ \vdots & \ddots & \vdots \\ \sigma_n \sigma_1 & \cdots & \sigma_n \sigma_n \end{bmatrix} \quad (4.3)$$

From this covariance matrix solve the eigenvalue equation:

$$Cv_i = \lambda_i v_i \quad (4.4)$$

where  $v_i$  are the eigenvectors or in the sense of PCA the principal components while  $\lambda_i$  are the associated eigenvalues.

Once the principal components are obtained they are sorted together with the associated eigenvalues. In order to calculate

the projections a sample  $S$  and hence a sequence in its k-mer representation has to be projected onto the principal components  $v_i$  with the highest eigenvalues. The number of projections is given by the *dimension* parameter of the tool. The projections  $p_i$  are the simple inner product of a k-mer representation of a sequence  $S$  with the eigenvector  $v_i$ ,

$$p_i = \langle v_i | S \rangle . \quad (4.5)$$

As such we have effectively removed the dimensionality to the number of *dimensions* as selected.

The algorithm is implemented by parallelization on different levels. The algorithm spends for a great number of samples most of its time calculating the matrix elements of equation (4.2). The matrix itself rests in most cases of a small rank, the number of individual k-mers. As such solving the eigenvalues and eigenvectors (principal components) in equation (4.4) is especially for k-mers of small  $k$  fast. Hence we focused our efforts on the calculation of the matrix elements in (4.2). As the covariance matrix is symmetric we calculate only the half of the elements transposing them into the other half. Each row of the matrix is calculated in its own thread, allowing us to calculate several rows on several cores in parallel. Besides this we used Single Instruction Multiple Data (SIMD) instructions in evaluating equation (4.2). We do this in calculating partial sums:

$$\sigma_{X\sigma Y}[1] = \sum_{i=1}^{n/4} (X_{[(i+4)n/4]} - \bar{X})(Y_{[(i+4)n/4]} - \bar{Y}) \quad (4.6)$$

$$\sigma_{X\sigma Y}[2] = \sum_{i=2}^{n/4} (X_{[(i+4)n/4]} - \bar{X})(Y_{[(i+4)n/4]} - \bar{Y}) \quad (4.7)$$

$$\sigma_{X\sigma Y}[3] = \sum_{i=3}^{n/4} (X_{[(i+4)n/4]} - \bar{X})(Y_{[(i+4)n/4]} - \bar{Y}) \quad (4.8)$$

$$\sigma_{X\sigma Y}[4] = \sum_{i=4}^{n/4} (X_{[(i+4)n/4]} - \bar{X})(Y_{[(i+4)n/4]} - \bar{Y}) \quad (4.9)$$

$$\sigma_{X\sigma Y} = \sum_{i=1}^4 \sigma_{X\sigma Y}[i]. \quad (4.10)$$

Further the partial sums are implemented by the Kahan summation algorithm [10], avoiding floating point errors in summing a

large number of values.

Calculating the eigenvectors and eigenvalues is performed by calling the optimized *dsyev* routine from the LAPACK libraries [11].

### 4.2.5 Implementation

Both the algorithm and the interface are implemented in *kmer2pca.c*.

## 4.3 pca\_visual\_extract

### 4.3.1 General

The PCA visual extract tool allows you to visualize a sequence dataset, that has been transformed to a k-mer representation on which PCA was performed on, and select sequences using a graphical user interface. In order to calculate the necessary representations the reader might look at the tools *fasta2kmer* and *kmer2pca* highlighted in sections 4.1 and 4.2 respectively.

### 4.3.2 Usage

The Interface for the *pca\_visual\_extract* tool can be brought up by calling:

```
pca_visual_extract [fasta] [pca] [dimensions] [first-dim] [second-dim] \
[out-fasta]
```

Listing 4.5 – Calling the *pca\_visual\_extract* tool

with the arguments being:

1. *fasta* A FASTA file that has been processed by the *fasta2kmer* and *kmer2pca* tools (c.f. sections 4.1 and 4.2) to create a representation of the dataset in a PCA subspace.
2. *pca* The projections onto the principal components as generated by *kmer2pca*.
3. *dimensions* The number of principal components the sequences have been projected onto. The number of dimensions of the PCA subspace.
4. *first-dim* The first principal component to visualize.

5. *second-dim* The second principal component to visualize.
6. *out-fasta* A FASTA file where the selected sequences will be written to.

Here *first-dim* and *second-dim* allow you to choose an arbitrary plane spun by two principal components to select your sequences in a visual manner. Once the call to the program was issued on the shell and all the arguments are coherent a graphical user interface as shown in figure 4.1 is presented on the screen. On the top of figure 4.1 you see a window in which you can draw a rectangle using your mouse cursor. The terminal as shown below shows you how many sequences you have selected. Once you are fine with your selection you might hit the Enter key in order to write out your selection to a FASTA file.

### 4.3.3 Example

```
pca_visual_extract test.fasta test.pca 7 2 3 /tmp/out.fasta
```

Listing 4.6 – Example of the *pca\_visual\_extract* tool

Will bring up an interface such as the one shown in figure 4.1, representing the sequences in *test.fasta* in their 7 dimensional PCA subspace as stored in *test.pca*. A two dimensional image representing the second and third principal component is drawn. Using the mouse one can draw a rectangle selecting sequences. Once the right sequences are selected hitting the Enter key will store them to */tmp/out.fasta*.

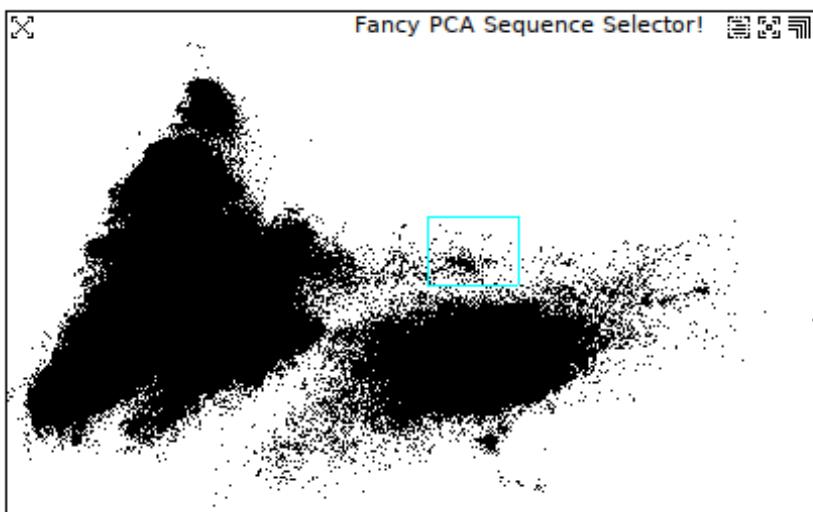
### 4.3.4 Implementation

The tool is implemented in *pca\_visual\_extract.c* and makes use of the Simple Direct Layer 2 (SDL2) library for its drawing routines and portability across systems.

## 4.4 cluster\_dbscan\_X

### 4.4.1 General

The *cluster\_dbscan\_PCA*, *cluster\_dbscan\_kmerL1*, *cluster\_dbscan\_kmerL2*, *cluster\_dbscan\_SW* and *cluster\_dbscan\_SW\_GPU* implement clustering of a dataset of sequences using the DBSCAN [2] algorithm.



```

haschka@epr:/mnt/data/uni/mnhn/mnhn-tools-manual
F50_tree_from_annotation.pdf
F50_tree_from_annotation.svg
SILVA_138_SSURef_NR99_tax_silva_trunc.fasta
SILVA_138_SSURef_tax_silva_trunc-dna-F200.split_set_from_swarm
SILVA_138_SSURef_tax_silva_trunc-dna-for-swarm.fasta
SILVA_138_SSURef_tax_silva_trunc-dna-for-swarm.swarm
SILVA_138_SSURef_tax_silva_trunc-dna-for-swarm.uniq
SILVA_138_SSURef_tax_silva_trunc-dna.fasta
SILVA_138_SSURef_tax_silva_trunc-dna.pca-7v
SILVA_138_SSURef_tax_silva_trunc-dna.split_set_from_swarm
SILVA_138_SSURef_tax_silva_trunc-dna.uniq
SILVA_138_SSURef_tax_silva_trunc.5mer
SILVA_138_SSURef_tax_silva_trunc.fasta
SILVA_138_SSURef_tax_silva_trunc.uniq
full_group_label
full_group_label_uniq
full_group_split_set
haschka@epr /mnt/data/uni/mnhn/mnhn-tools-manual $ ./code/bin/pca_visual_extractor ../silva/SILVA_138_SSURef_tax_silva_trunc-dna.fasta ../silva/SILVA_138_SSURef_tax_silva_trunc-dna.pca-7v 7 1 2 /tmp/out.fasta
Calculating...
Selected 510 Sequences !
Hit enter to save to fasta file: /tmp/out.fasta

```

Figure 4.1 - The color inverted interface of the *pca\_visual\_sequence\_selector*.

The different variants implement DBSCAN clustering using different sequence-sequence distances and platforms.

#### 4.4.2 Usage

```
cluster_dbscan_pca [fasta] [pcaproj] [dimensions] [epsilon] [minPoints] \
[outfiles-fasta-prefix] [outfiles-values-prefix]

cluster_dbscan_kmerL1 [fasta] [kmers] [epsilon] [minPoints] \
[outfiles-fasta-prefix]

cluster_dbscan_kmerL2 [fasta] [kmers] [epsilon] [minPoints] \
[outfiles-fasta-prefix]

cluster_dbscan_SW [fasta] [epsilon] [minPoints] \
[outfiles-fasta-prefix]

cluster_dbscan_SW_GPU [fasta] [epsilon] [minPoints] \
[outfiles-fasta-prefix]
```

Listing 4.7 – Calling the *cluster\_dbscan* tools

Where the different tools perform as disguised by their name:

1. *cluster\_dbscan\_pca*: Clustering of sequences in a PCA subspace such as the one obtained using the *kmer2pca* tool described in section 4.2.
2. *cluster\_dbscan\_kmerL1*: Clustering of sequences in a k-mer representation applying the  $L_1$ -norm also known as the Manhattan distance between the k-mer frequency vectors that represent each sequence as a distance between them.
3. *cluster\_dbscan\_kmerL2*: Clustering of sequences in a k-mer representation applying the  $L_2$ -norm also known as the Manhattan distance between the k-mer frequency vectors that represent each sequence as a distance between them.
4. *cluster\_dbscan\_SW*: Clustering of sequences using the Smith-Waterman distance measure.
5. *cluster\_dbscan\_SW\_GPU*: Clustering of sequences using the Smith-Waterman distance measure on an OpenCL supported device, for instance a GPU.

The tools have the following arguments:

1. *fasta*: A FASTA file containing the sequences to be clustered. In the case of Smith-Waterman clustering the FASTA file has to contain nucleic sequences as clustering of proteinic sequences under a Smith-Waterman distance is not im-

plemented. In case the Smith-Waterman distance is evaluated on GPUs the maximum sequence length is limited to 300 nucleotide bases.

2. *pcaproj*: The projections onto the principal components as calculated by the *kmer2pca* tool.
3. *dimensions*: The number of dimensions, principal components found in the file.
4. *epsilon*: The epsilon neighbourhood radius as used by the DBSCAN [2] algorithm.
5. *minPoints*: The minimum number of sequences to be found within an epsilon neighbourhood in order to form or expand a cluster according to the DBSCAN algorithm.
6. *outfiles-fasta-prefix*: The DBSCAN based tools all generate a FASTA file per cluster found. With this argument you can tell the program where to store those clusters. The number of clusters to be generated is capped at 500 clusters. The resulting files will be named according to your prefix with a number at the end for each different cluster.
7. *outfiles-values-prefix*: If this parameter is added to the PCA version of the DBSCAN implementation the clusters are not only stored as individual fasta files but also as files containing the projections onto the principal components for the clustering in question. This for instance allows one to draw the different clusters into visual representations of the PCA projections.

#### 4.4.3 Algorithm

The utility implements the DBSCAN algorithm straight forward as described in its original article [2]. The DBSCAN algorithm is perfect in finding *density connected* subsets of a dataset. The algorithm as such find all connected ensembles in a dataset that exceed a certain density:

$$\rho = \frac{n}{V(\epsilon)} \tag{4.11}$$

where  $V$  is the Volume within an epsilon neighborhood defined by the radius  $\epsilon$ .  $n$  corresponds to the number of sequences within the aforementioned volume, and is defined using the *minPoints*

parameter. The formula for the volume depends on the metric used, but is hypercubic for  $L_1$  measurements, and hyperspheric for  $L_2$  measurements. Sequences scattered in sequence space with a density below  $\rho$  are found to be outliers and will not be taken into account and are hence not to be found in the resulting clusters.

The different distance measures in the *region expand method* as described in the original paper use various means of optimisation and parallelization. In the case of PCA, kmerL1, kmerL2 distance methods up to eight distances between sequences are calculated in parallel making use of the AVX SIMD instructions. In the case of the Smith Waterman distance measure an OpenCL (GPU) method is available allowing to efficiently calculate multiple distances in parallel on OpenCL devices such as GPUs.

The Smith Waterman algorithm calculates the Smith Waterman matrix and defines the maximum value to be found within the matrix to be the distance between two sequences. No implementation a Smith Waterman algorithm using protein sequences has been implemented. The algorithm builds the matrix using the recurrence relation:

$$\begin{aligned} s_1 &= M(j-1, i-1) + F(A(i), B(j)), \\ s_2 &= M(j, i-1) - G, \\ s_3 &= M(j-1, i) - G, \\ M(i, j) &= \max(s_1, s_2, s_3). \end{aligned} \quad (4.12)$$

where  $A$  and  $B$  are two sequences of the dataset to be compared and  $A(i)$  is the  $i$ -th letter of sequence  $A$  and  $B(j)$  is the  $j$ -th letter of sequence  $B$ .

$$F(A(i), B(j)) = \begin{cases} A(i) = B(j) : 4 \\ A(i) \neq B(j) : -4, \end{cases} \quad (4.13)$$

is the comparison relation.  $G$  is the gap penalty defined to be 3 and  $M(i, j)$  the Smith Waterman matrix yielding the distance:

$$d(A, B) = \max(M(i, j)); \quad (4.14)$$

In the case of the GPU implementation only three subsequent rows of the matrix are stored in memory evaluating the maximum on the fly calculating row after row of the Smith Waterman Matrix. This shall allow to use more local memory that is closer to the compute unit, if not only local registers on the graphics card for the computation.

#### 4.4.4 Example

```
cluster_dbSCAN_PCA test.fasta test.pca 7 0.02 4 /tmp/outf /tmp/outv
```

Listing 4.8 – Example of the *cluster\_dbSCAN\_PCA* tool

This example performs DBSCAN based clustering using  $L_2$  distances in a 7 dimensional subspace with subspace projections stored in the test.pca file. The input values for the DBSCAN run are  $\epsilon = 0.02$  and minpoints  $n = 4$ . The resulting clusters will be stored in /tmp/ with in FASTA format with a file name of outf $N$ , and as projections onto the principal components outv $N$ , where  $N$  is an individual number for each clusters.

#### 4.4.5 Implementation

The DBSCAN algorithm for all variants is implemented in *dbscan.c*. The Interface for the Smith-Waterman variants is implemented in *cluster\_dbSCAN\_SW.c*. The Interface for k-mer distance based  $L_1$  and  $L_2$  variants is implemented in *cluster\_dbSCAN\_kmer.c*. The Interface for the PCA variant is implemented in *cluster\_dbSCAN\_pca.c*.

### 4.5 compareSW

#### 4.5.1 General

The *compareSW* tool calculates the mean distance and standard deviation between two datasets under the Smith Waterman distance.

#### 4.5.2 Usage

One can call the *compareSW* tool like:

```
compareSW [fasta1] [fasta2] [n-threads]
```

Listing 4.9 – Calling the *compareSW* tool

with the following arguments:

1. *fasta1*: A FASTA file containg a set of sequences.
2. *fasta2*: An other FASTA file containing a set of sequences.
3. *n-threads*: The number of threads that this computation might use.

### 4.5.3 Algorithm

calculates the mean and the standard deviation of the distances between the sequences in the two datasets. If the first dataset has  $n$  sequences and the second  $j$ , the number of total distances calculated is hence,  $N = nj$ . The Smith Waterman algorithm is the same as outlined in section 4.4.3.

### 4.5.4 Example

```
compareSW test1.fasta test2.fasta 8
```

Listing 4.10 - Example of the *compareSW* tool

This example computes the mean distance between the sequences in test1.fasta and test2.fasta using 8 cores.

### 4.5.5 Implementation

The functions to calculate the mean and standard deviation are implemented in *comparison.c*. The Interface is implemented in *compareSW.c*.

## 4.6 find\_satellite

### 4.6.1 General

The *find\_satellite* is a filtering tool narrowing down a FASTA dataset of similar sequences. The filter can filter sequences by length, and by Smith Waterman distance to a template sequence. The Smith Waterman algorithm is implemented as shown in 4.4.3. This tool only works with nucleotide sequences.

### 4.6.2 Usage

The tool is to be called like:

```
find_satellite [fasta-ds] [fasta-target] [size] [size+] [size-] \
[sw-dist] [n-threads] > [fasta-filtered]
```

Listing 4.11 - Calling the *find\_satellite* tool

where the arguments are:

1. *fasta-ds* A FASTA dataset of the sequences that should be searched for according to the pattern selected by the following arguments.
2. *fasta-target* A FASTA file holding the target sequence from which the Smith Waterman distance will be calculated.
3. *size* The size of the sequence to be searched for with the tolerance intervals given by the subsequent arguments.
4. *size+* How many nucleotides above the size sequences are accepted.
5. *size-* How many nucleotides below the size sequences are accepted.
6. *sw-dist* The maximum Smith Waterman distance from the target sequence to accept sequences.
7. *n-threads* The number of parallel threads this computation may use.
8. *fasta-filtered* A FASTA file where the sequences that met the filtering criteria above are written to.

### 4.6.3 Example

```
find_satellite test.fasta target.fasta 100 5 3 20 8 > filtered.fasta
```

Listing 4.12 – Example of the *find\_satellite* tool

Here the tool searches for sequences in *test.fasta* that are less than a Smith Waterman distance of 20 away from the sequence stored in *target.fasta*. Sequences that are between 97 and 105 nucleotides long are accepted. The resulting sequences are stored to *filtered.fasta*

### 4.6.4 Implementation

The tools interface is implemented in *find\_satellite.c*. The mechanism is implemented in *filter.c*.

## 4.7 silhouette

### 4.7.1 General

The *silhouette* tool calculates the silhouette index comparing two, clusters or datasets stored in two FASTA files.

### 4.7.2 Usage

One can call the *silhouette* tool like:

```
silhouette [fasta1] [fasta2] [n-threads]
```

Listing 4.13 – Calling the *silhouette* tool

with the following arguments:

1. *fasta1*: A FASTA file containing a set of sequences.
2. *fasta2*: An other FASTA file containing a set of sequences.
3. *n-threads*: The number of threads that this computation might use.

### 4.7.3 Algorithm

The tool calculates the silhouette index comparing distances of internal to each dataset, with distances external to each dataset.

The silhouette index is calculated using the following formula:

$$S = \frac{1}{n_1} \sum_{i=1}^{n_1} s_i, \quad (4.15)$$

with  $s_i$  defined to be:

$$s_i = \begin{cases} a_i < b_i : 1 - a_i/b_i \\ a_i = b_i : 0 \\ a_i > b_i : b_i/a_i - 1 \end{cases} \quad (4.16)$$

and  $a_i$  being the internal measure for distances to the  $i$ -th sequence  $A_i$  found in the first FASTA file *fasta1* and hence,

$$a_i = \frac{1}{n_a} \sum_{j=1}^{n_A} d_{SW}(A_i, A_j). \quad (4.17)$$

likewise  $b_i$  is the external measure for distances to the  $i$ -th sequence  $A_i$  in the first FASTA file and sequences  $B$  found in the second fasta file. We write as such:

$$b_i = \frac{1}{n_b} \sum_{j=1}^{n_B} d_{SW}(A_i, B_j). \quad (4.18)$$

with  $d_{SW}(X, Y)$  being the smith watherman distance between two sequences  $X$  and  $Y$  as defined in section 4.4.3.  $n_A$  and  $n_B$  are the number of sequences found in the first  $A$ , and second  $B$  FASTA file.

#### 4.7.4 Example

```
silhouette test1.fasta test2.fasta 8
```

Listing 4.14 - Example of the *silhouette* tool

This example calculates the silhouette index between the dataset of sequences found in *test1.fasta* and the dataset found in *test2.fasta*.

#### 4.7.5 Implementation

The functions to calculate the silhouette index are implemented in *comparison.c*. The interface is implemented in *silhouette.c*.

### 4.8 consens

#### 4.8.1 General

The *consens* tool calculates a consens sequence from sequences found in a fasta file. Further the *consens* tool gives a detailed information about the distribution of nucleotide sequences in a datasets. The sequences have to be aligned or quasi-aligned before calling the *consens* tool. The tool currently only works with nucleotide sequences.

#### 4.8.2 Usage

The *consens* tool can be called like:

```
consens [fasta]
```

Listing 4.15 - Calling the *consens* tool

where the *fasta* argument is a fasta file to perform the *consens* calculation on.

#### 4.8.3 Example

```
consens test.fasta
```

Listing 4.16 - Example of the *consens* tool

calculates the consens of the sequences found in test fasta and provides you with an output like:

Base	A	C	G	T	A	C	G	T
G	10	5	230	5	0.04000	0.02000	0.92000	0.02000
C	18	213	8	11	0.07200	0.85200	0.03200	0.04400
T	0	0	0	250	0.00000	0.00000	0.00000	1.00000
T	0	4	0	246	0.00000	0.01600	0.00000	0.98400
C	0	246	0	4	0.00000	0.98400	0.00000	0.01600
T	2	4	1	243	0.00800	0.01600	0.00400	0.97200
T	3	2	6	239	0.01200	0.00800	0.02400	0.95600
G	5	0	240	5	0.02000	0.00000	0.96000	0.02000
A	245	3	0	2	0.98000	0.01200	0.00000	0.00800
A	237	2	10	1	0.94800	0.00800	0.04000	0.00400
G	8	0	241	1	0.03200	0.00000	0.96400	0.00400
G	3	0	241	6	0.01200	0.00000	0.96400	0.02400
G	9	4	234	3	0.03600	0.01600	0.93600	0.01200
A	244	2	2	2	0.97600	0.00800	0.00800	0.00800
A	246	3	0	1	0.98400	0.01200	0.00000	0.00400
A	240	1	7	2	0.96000	0.00400	0.02800	0.00800
G	15	11	222	2	0.06000	0.04400	0.88800	0.00800

Listing 4.17 – Example output of the *consens* tool

where you are presented in the first column with the consensus sequence, from the second to the fifth column with the absolute number of the according bases found, and from the sixth to the ninth column with the percentage of each bases occurring at this position. The final line also shows some statistics:

MEAN MAX ACCURACY: 0.897734 SIGMA: 0.091085

Listing 4.18 – Example statistics output of the *consens* tool

where *MEAN MAX ACCURACY* is the mean of the highest percentage between the four nucleotide bases for each position and *SIGMA* the square of the standard deviation.

## 4.9 sequence\_multiplicity

### 4.9.1 General

The *sequence\_multiplicity* finds unique sequences in a FASTA file and allows one to know how often a unique sequence appears in the FASTA file.

### 4.9.2 Usage

One can call the *sequence\_multiplicity* tool like:

```
sequence_multiplicity [fasta] > [fasta-out]
```

Listing 4.19 – Calling the *sequence\_multiplicity* tool

with the arguments:

1. *fasta* a FASTA file to find unique sequences in.

2. *fasta-out* the unique sequences in FASTA format.

### 4.9.3 Example

```
sequence_multiplicity test.fasta > out.fasta
```

Listing 4.20 - Example of the *sequence\_multiplicity* tool

Finds unique sequences in *test.fasta* and writes them into *out.fasta*. Each sequence in the *out.fasta* file contains a suffix number telling you how many times this sequence occurred in the original dataset.

```
>sequence_2_2344
```

Listing 4.21 - Example a line in *out.fasta*

Tells you that the third FASTA file, we count from 0 onwards in the input file, appears 2344 times in the dataset.

### 4.9.4 Implementation

Finding unique sequences is implemented in *dataset.c* while the interface is implemented in *sequence\_multiplicity.c*.

## 4.10 lengths\_from\_fasta

The *lengths\_from\_fasta* tool prints the lengths of sequences stored in a FASTA file.

### 4.10.1 Usage

The *lengths\_from\_fasta* tool is to be called like:

```
lengths_from_fasta [fasta]
```

Listing 4.22 - Calling the *lengths\_from\_fasta* tool

where the *fasta* argument is a FASTA file to print out the length of sequences from.

### 4.10.2 Implementation

The code is implemented in *lengths\_from\_fasta.c*

## 4.11 compare\_norms

### 4.11.1 General

The compare norms utility allows you to calculate distances between sequences in a dataset using the Smith Waterman Distance or a distance in the a principal components subspace of k-mers.

### 4.11.2 Usage

The *compare\_norms* tool compares distances between all sequences in a fasta file in Smith Waterman distance and in a subspace yielded by PCA.

```
compare_norms [fasta] [pca] [dimensions] > [output]
```

*Listing 4.23 - Calling the compare\_norms tool*

where the arguments are:

1. *fasta* A fasta file containing the sequences to calculate the distances in between them.
2. *pca* A corresponding projection onto principal components from k-mers as obtained by the *kmer2pca* tool outlined in section 4.2
3. *dimensions* The number of projections per sequence to be found in the *pca* file.
4. *output* A file where to output the results.

### 4.11.3 Algorithm

The Smith Waterman algorithm is used in the same way as pointed out in section 4.4.3. The distance based on the projections is calculated in  $L_2$  and  $L_1$  norm.

### 4.11.4 Example

```
compare_norms test.fasta test.pca 7 > distances
```

*Listing 4.24 - Example of the compare\_norms tool*

The *compare\_norm* tool calculates the distances between the sequences within the *test.fasta* in Smith Waterman distance and in  $L_1$  and  $L_2$  norm in the 7 dimensional PCA subspace.

The distances file contains, per line as first and second integers the index of the sequences compared. The third position the Smith Waterman distance, the fourth distance position is the distance in the PCA subspace in  $L_2$  norm and the fifth in  $L_1$  norm. A sample output of looks like this:

### 4.11.5 Example

```
i j SW L1 L2
1 0 82.000000 2.546303 6.056748
2 0 122.000000 5.428422 10.829566
2 1 84.000000 4.875662 10.343906
3 0 96.000000 4.237413 8.629056
3 1 88.000000 4.427232 8.065469
3 2 118.000000 5.210451 12.005179
4 0 80.000000 3.201523 7.927563
4 1 72.000000 3.526202 7.919044
4 2 106.000000 6.843701 14.724481
```

Listing 4.25 – Example output of the *compare\_norms* tool

## 4.12 adaptive\_clustering\_X

### 4.12.1 General

The adaptive clustering tools are at the heart of the MNHN-Tree-Tools suite. The adaptive clustering tools perform a progressive scan using different densities in exploiting the DBSCAN [2] algorithm in order to build trees, and gain further insight into a dataset of sequences. Just as the DBSCAN clustering tools (c.f. section 4.4), the adaptive clustering algorithm exists in different variants, corresponding to sequence distance measure and underlying machine hardware. The tools are: *adaptive\_clustering\_PCA*, *adaptive\_clustering\_kmer\_L1*, *adaptive\_clustering\_kmer\_L2*, *adaptive\_clustering\_SW\_GPU*, *adaptive\_clustering\_SW\_MPI\_GPU*

### 4.12.2 Usage

One can call the utilities with the following commands:

```
adaptive_clustering_PCA [fasta] [initial-epsilon] [delta-epsilon] \
[minPoints] [split-sets] [n-threads] [dimensions] [pca-file] > [outfile]
adaptive_clustering_kmer_L1 [kmers] [initial-epsilon] [delta-epsilon] \
[minPoints] [split-sets] [n-threads] > [outfile]
```

```

adaptive_clustering_kmer_L2 [kmers] [initial-epsilon] [delta-epsilon] \
[minPoints] [split-sets] [n-threads] > [outfile]

adaptive_clustering_SW [fasta] [initial-epsilon] [delta-epsilon] \
[minPoints] [split-sets] [n-threads] > [outfile]

adaptive_clustering_SW_GPU [fasta] [initial-epsilon] [delta-epsilon] \
[minPoints] [split-sets] [n-threads] > [outfile]

adaptive_clustering_SW_MPI_GPU [fasta] [initial-epsilon] [delta-epsilon] \
[minPoints] [split-sets] [n-threads] > [outfile]

```

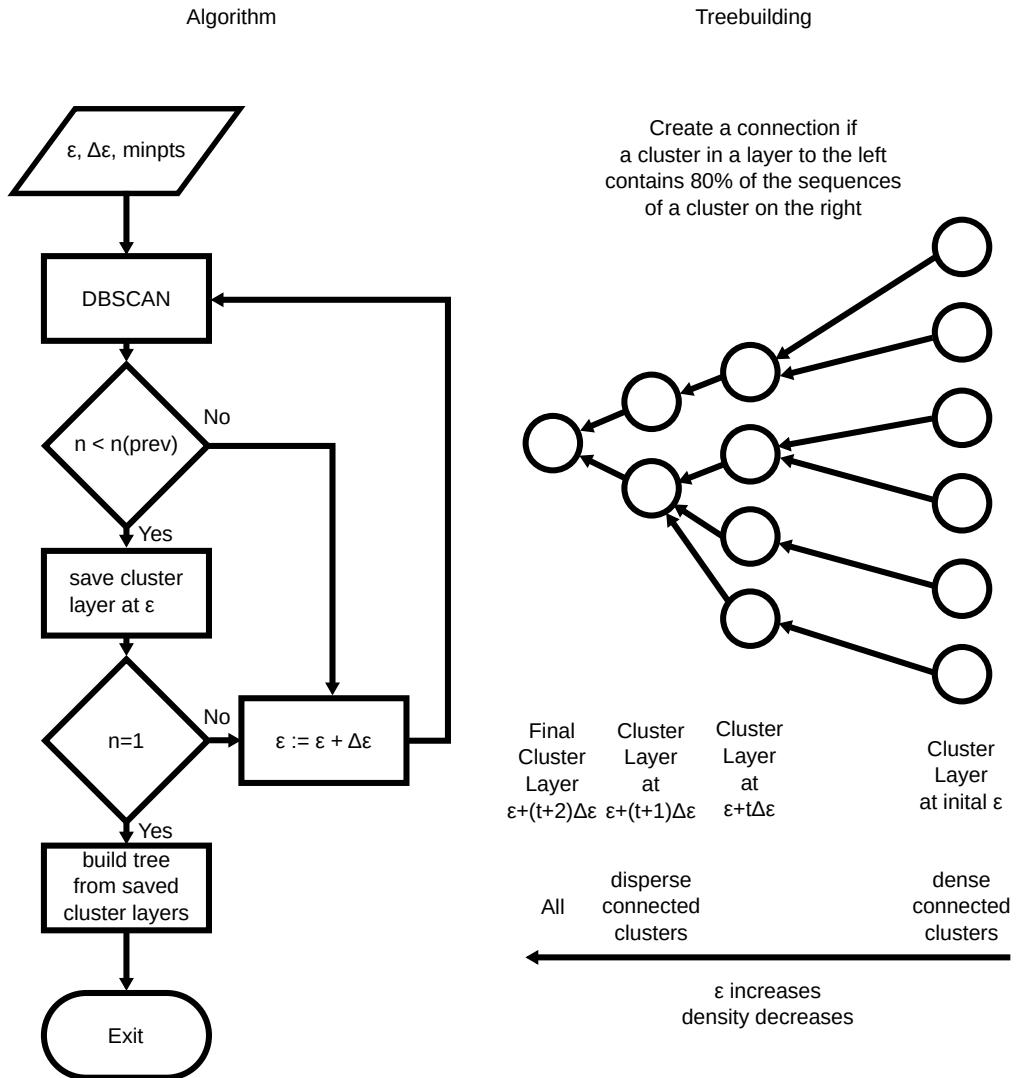
**Listing 4.26 - Calling the *adaptive\_clustering\_X* tools**

which have the following arguments:

1. *fasta* The FASTA file containing the dataset to be adaptively clustered and to have a tree built from.
2. *kmers* A k-mer representation of a dataset to be adaptively clustered and to have a tree built from. Such a set can be obtained by the *fasta2kmer* tool highlighted in section 4.1
3. *initial-epsilon* The initial epsilon to start an adaptive clustering run from.
4. *delta-epsilon* The increase in epsilon between successive calls of the DBSCAN algorithm.
5. *minPoints* The number of sequences need to be found within an epsilon neighbourhood to form or expand a cluster.
6. *split-sets* A split is a binary file to hold entire clustering results. As the algorithm creates multiple split sets ( one for each stage of the tree ) a path with a named prefix has to be given. The splitsets are named by the argument given with a running number at the end starting at 0 for the outermost leaves of the tree counting upwards until only a single cluster can be formed, the one with the lowest density.
7. *n-threads* The number of threads this run can be allowed to use. MPI\_GPU runs will fail if this value is not set to 1.
8. *outfile* The outfile of the adaptive clustering run which is valuable not only for informative purposes but also for further tools of this software suite.

### 4.12.3 Algorithm

The algorithm resides on the DBSCAN algorithm already outlined in section 4.4. The algorithm starts at a given sequence density defined by the parameters *initial-epsilon* and *minPoints*, as highlighted by equation 4.11. By successive increasing  $\epsilon$  by *delta-epsilon* and hence successive DBSCAN runs several clusterings with ever decreased densities are generated. Clusterings with contain a higher or the same number of clusters as the previous DBSCAN runs are discarded, only clustering results with less clusters in lesser density run are kept, until a system with a single cluster is reached. We further remark that a result can be obtained at any specified density à posteriori if the researcher requests to do so with the *cluster\_dbscan\_X* tools highlighted in section 4.4. We shall also outline that this forced monotony in an every decreasing number of clusters also makes the importance of a wisely chosen starting point: *initial-epsilon* and *minpoints* necessary. The starting point shall be chosen at a density where the highest number of clusters is suspected. From these multiple clustering results at different densities we can infer a hierarchy by comparing clusters of different layers and verifying whether they assumed smaller clusters at higher densities are a part of larger clusters in results from cluster runs at lesser densities. Connections between clusters are created and print to the output file if a cluster from a higher density is contained at least at 80% of its sequences in a cluster of a lesser density. The functioning of the algorithm is outlined in figure 4.2. The algorithm further highlights the same optimizations as pointed out in section 4.4.3. Further the algorithm is multithreaded precalculating layers at different  $\epsilon$ s in different threads in parallel. Finally a Message Passing Interface (MPI) version of the GPU based Smith Waterman DBSCAN region expand function exists. In this function the distance between a sequence in question with all other sequences of the dataset is to be evaluated to find out whether or not this distance is smaller than  $\epsilon$  and falls within an epsilon neighborhood or not. Using the MPI interfaces the algorithm has the possibility to profit from multiple available GPUs in multiple machines such as in a clustering environment. We have successfully tested to run this algorithm on six independent cluster nodes with four GPUs each using an infiniband network backbone.



**Figure 4.2 – Building trees from sequences:** Highlights our algorithm, and its three parameter input  $\epsilon$  the initial epsilon neighborhood radius for the DBSCAN algorithm,  $\Delta\epsilon$  the increase of  $\epsilon$  in every step and  $\text{minpts}$  the minimal number of points to be found in an epsilon neighbourhood to either extend or create a cluster.  $n$  in the diagram represents the number of clusters of the current DBSCAN run,  $n(\text{prev})$  the number of clusters of the previous run.  $t$  at the treebuilding side, the stepnumber and hence, how often DBSCAN has been run. The algorithm works itself from right to the left through the tree, detecting first using a small  $\epsilon$  very dense clusters of nucleic sequences and hence, clusters that are highly conserved in sequence. In increasing DBSCANS  $\epsilon$  from layer to layer, the clusters contain less and less conserved sequences, and basically fusion from layer to layer, until DBSCAN just detects a single cluster, the root of the tree.

#### 4.12.4 Example

```
adaptive_clustering_PCA test.fasta 0.02 0.01 4 /tmp/out-splits- 8 7 test.pca
```

Listing 4.27 – Calling the *adaptive\_clustering\_PCA* tools

Performs an adaptive clustering run on the sequences stored in *test.fasta* by applying DBSCAN on the 7 dimensional k-mer subspace defined by the projections onto the principal components stored in *test.pca*. The *initial-epsilon* is defined to be 0.02, the *delta-epsilon* is defined to be 0.01 and a minimum number of points for cluster expansion or creatin is defined to be four. The resulting clusters for each run are stored in binary form at */tmp/out-splits-X* where X counts from 0 to the number of layers calculated until only a single cluster was found due to the small density at a large  $\epsilon$ .

#### 4.12.5 Implementation

The algorithm and its interfaces are implemented in *adaptive\_clustering.c*. The code makes havly use of the DBSCAN algorithm implemented in *dbscan.c*.

### 4.13 split\_set\_to\_fasta

This tool generates FASTA files for clusters found in a split set stored in binary form like the ones generated by the *adaptive\_clustering\_X* tools, highlighted in section 4.12.

#### 4.13.1 Usage

The *split\_set\_to\_fasta* tool can be called in the following way

```
split_set_to_fasta [fasta] [split-set] [out-path-name]
```

Listing 4.28 – Calling the *split\_set\_to\_fasta* tool

with the arguments:

1. *fasta* A FASTA file containing the whole dataset from where the set of clusters was created.
2. *split-set* The binary split set containing the clusters.
3. *out-path-name* The path and name of the resulting FASTA files for that are suffixed with unique number for each cluster in stored in the *split-set*

### 4.13.2 example

```
split_set_to_fasta test.fasta /tmp/s0023 /tmp/fa0023-
```

Listing 4.29 – Example of the *split\_set\_to\_fasta* tool

Here the *split-set* /tmp/s0023 is written to individual FASTA files per clusters stored to /tmp/fa0023-X where X is a number for each individual cluster. *test.fasta* holds the original dataset which is needed as the *split-set* file only stores indices and not the sequences itself.

### 4.13.3 Implementation

The tool is implemented in *split\_set\_to\_fasta.c*.

## 4.14 split\_set\_to\_projections

The *split\_set\_to\_projections* tool allows you to get clusterwise principal component projections from the clustering information stored in binary *split-set* files obtained by an adaptive clustering run on a dataset, i.e. obtained from the tool shown in section 4.12 and a PCA projection of the same dataset that can be obtained using the *kmer2pca* tool presented in section 4.2. This tool is especially useful to plot individual clusters together with a PCA analysis of sequences and hence find sequences corresponding to clusters in such a diagram.

### 4.14.1 usage

The *split\_set\_to\_projections* tool is to be called like:

```
split_sets_to_projections [fasta] [pca] [dimensions] [split-set] [projections]
```

Listing 4.30 – Calling the *split\_set\_to\_projections* tool

with the following arguments:

1. *fasta* The FASTA file the clustering and hence the *split-set* was generated from.
2. *pca* A projection onto principal components of the affor-mentioned FASTA file, i.e. obtained with the *kmer2pca* tool mentioned in section 4.2.
3. *dimensions* The number of dimensions of the PCA subspace in the afformentioned *pca* file.

4. *split-set* The binary file that holds the clusters.
5. *projections* A *path/filename* combination where to put the resulting projections of the clusters in PCA space. Each cluster will generate a file with an individual number added to the name.

```
split_sets_to_projections test.fasta test.pca 7 /tmp/s0023 /tmp/s0023proj
```

Listing 4.31 – Calling the *split\_set\_to\_projections* tool

In this case clusters stored in */tmp/s0023* a binary *split-set* file are searched in the 7 dimensional subspace projection onto the principal components of a k-mer representation of the sequences found in the *test.fasta* FASTA file. The projections of each cluster are stored in a separate file */tmp/s0023projX* where X is a number corresponding to each cluster.

## 4.15 split\_sets\_to\_newick

This tool allows you to create a Newick [7] tree file from the *split-set* and *output* files that have been generated by an adaptive clustering run as outlined in section 4.12.

### 4.15.1 Usage

One can use the tool like:

```
split_sets_to_newick [filter] [lengths-from-output] [dimensions] \
[outfile] [minPoints] [file1 ... filen] > [newick]
```

Listing 4.32 – Calling the *split\_sets\_to\_newick* tool

with arguments as follows:

1. *filter* One may use 0, for no filter, an other number for the minimum cluster size to occur in the resulting tree. Filtering like this works but changes the labels on the tree. For correct labels we suggest the user to filter with the *filter\_split\_sets\_by\_min* tool, as shown in section 4.19, prior using this tool and apply no filter. This parameter is kept for convenience.
2. *lengths-from-output* Use distances between nodes of the Newick tree using parameters found in the output file of an adaptive clustering run with a tool shown in section 4.12.

If you do not choose to use lengths proportional to density differences choose 0 here, choose 1 if your run uses an  $L_1$  based distance calculation, and 2 if your run uses an  $L_2$  based distance calculation. 1 uses a hypercube with a side-length of  $2\epsilon$  to calculate the volume, 2 uses a hypersphere with radius  $\epsilon$ .

3. *dimensions* As the lengths between the nodes correspond to density differences we need the dimensions to calculate the volume from the different  $\epsilon$  radii found in the *outfile* file. As a basic guidance choose the number of projections onto principal components that you have used if this run was performed with PCA data. Otherwise choose 1 or a value to your convenience.
4. *outfile* The *output* file from an adaptive clustering run generated from one of the tools outlined in section 4.19.
5. *minPoints* The *minPoints* argument used during the adaptive clustering run, in order to correctly calculate the correct density differences.
6. *file1...filen* All the *split-set* files to use to build this tree from. In order to select them in general wildcards can be used on shells like the bash by i.e. using something like  
`/path/prefix*`  
`.`
7. *newick* The created Newick tree file.

The arguments *dimensions*, *outfile* and *minPoints* are to be omitted if 0 was chosen for the *lengths-from-output* argument.

## 4.15.2 Algorithm

If the lengths in the newick file are chosen to be density differences, the density is calculated from the  $\epsilon$  value in the output file of an adaptive clustering run. The formulas used for the volume are:

$$V = (2\epsilon)^d, \quad (4.19)$$

for the hypercube  $L_1$  distance case, and

$$V = \epsilon^d \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} \quad (4.20)$$

for the hyperspherical case. Here  $d$  is the number of dimensions and  $\Gamma$  is the  $\Gamma$  function.

The algorithm builds the Newick file using a recursive function traversing the the nodes. Connections between nodes in different layers are made if from a denser layer a cluster is contained at least at 80% in a cluster of a less dense layer.

### 4.15.3 Example

```
split_sets_to_newick 0 2 7 outfile 4 clusters/out-* > tree.dnd
```

Listing 4.33 - Example of the *split\_sets\_to\_newick* tool

Here we generate a Newick tree which we store in the *tree.dnd* whose lengths are density difference based, from an adaptive clustering run whose output has been stored in *outfile* and whose *split-sets* have been stored to *clusters/out-X* files. A *minPoints* argument of 4 was used during the adaptive clustering run, which was PCA based and operated on a seven dimensional subspace spun by the principal components.

### 4.15.4 Implementation

The interface for Newick tree generation is implemented in *split\_sets\_to\_newick.c*, algorithms for tree generation are found in *cluster\_io.c*.

## 4.16 split\_set\_to\_matrix\_X

This are two tools to generate a matrix view of a tree that can then be plotted with the *gnuplot* [5] program. The matrix view basically, highlights each sequents as a point on a line, coloring sequences belonging to the same cluster in the same color. This is useful for instance in the verification of simulated or real datasets where the sequences are already ordered by clusters and hence should generate lines in the same color. This also shows as one stacks these lines on top of each other how clusters merge as the  $\epsilon$  value changes.

### 4.16.1 Usage

The tools *split\_set\_to\_matrix\_annotation* and *split\_set\_to\_matrix\_line* can be used together to form a readable output file.

```
split_set_to_matrix_annotation [fasta] >> [matrix]
split_set_to_matrix_line [fasta] [split-set] >> [matrix]
```

Listing 4.34 – Calling the *split\_set\_to\_matrix* tools

where:

1. *fasta* Is a FASTA file that one has generated clusters from for instance in an adaptive clustering, as the ones outlined in section 4.12
2. *split-set* A binary file holding the clusters obtained from an adaptive clustering run.
3. *matrix* The matrix to store a line to.

The *split\_set\_to\_matrix\_annotation* call in itself is not needed, it basically annotates each sequence with the internal sequence number, or the number of sequence occurrence in the original FASTA file, starting with 0 for the first sequence.

### 4.16.2 Example

One has run an adaptive clustering run that yielded 86 different clustering layers (of a tree) and hence, 86 split set files in using one of the adaptive clustering tools shown in section 4.12

```
for((i=0;i<85;i++))
do split_set_to_matrix_line test.fasta /tmp/out`printf %04d $i` >> test.matrix
done
```

Listing 4.35 – Example usage of the *split\_set\_to\_matrix* tools

In this example the 86 *split-sets* for each layer of the adaptive clustering run are successively as lines written to the *test.matrix* file. The *test.matrix* file can then be visualized by *gnuplot*. On the *gnuplot* console one might issue the command:

```
plot 'test.matrix' matrix with image
```

Listing 4.36 – Using gnuplot to plot a matrix from *split\_set\_to\_matrix* tools

to obtain a map with different colors for each cluster.

## 4.17 split\_set\_from\_annotation

### 4.17.1 General

This tool allows you to generate binary clustering sets such as those created using by an adaptive clustering run by hand, annotating each cluster.

### 4.17.2 Usage

The tool can be used like:

```
split_set_from_annotation [fasta-ds] [annotation] [unique-annotations] \
[split-set]
```

Listing 4.37 – Calling the *split\_set\_from\_annotation* tool

with the following arguments:

1. *fasta-ds* A FASTA file that shall be clustered according to the annotations in the *annotation* file.
2. *annotation* The annotation file, containing a line for each sequence in the *fasta-ds* file. An annotation can be any chosen word. If a line contains only a dash '-', the sequence will not be part of any of the possible clusters.
3. *unique-annotations* The database of possible annotations. For each cluster the word used to annotate sequences in the *annotation* file an entry containing the word corresponding to this cluster is made. A sole entry. The entries shall be made in cluster order that you wish to have in the resulting binary *split-set* file.
4. *split-set* The resulting binary clustering or *split-set* file.

### 4.17.3 Example

We imagine a fasta file:

```
>sequence1
AACCTT
>sequence2
AGCCTT
>sequence3
TTTTTT
>sequence4
AAACTT
>sequence5
AAACGT
```

```
>sequence6
AAAGCT
```

**Listing 4.38** - Example of a dataset clustered with the *split\_set\_from\_annotation* tool

and we want to cluster this file into two clusters, one containing sequence 1, and sequence 2 while the second cluster is composed of sequence 4, 5 and 6. We would like to omit sequence 3 from the clustering. As such we create an annotations file like:

```
one
one
-
two
two
two
```

**Listing 4.39** - Example of an *annotation* file for the *split\_set\_from\_annotation* tool

As shown we annotate all sequences of the first cluster with one, and all sequences of the second cluster with two. We omit the third sequence by annotating it with a dash '-'. Further we need to write a file containing all the words that we used to annotate our clusters.

```
one
two
```

**Listing 4.40** - Example of an *unique-annotations* file for the *split\_set\_from\_annotation* tool

With the files at hand we can call the tool:

```
split_set_from_annotation test.fasta annotation unique-annotations /tmp/out-set
```

**Listing 4.41** - Example of the *split\_set\_from\_annotation* tool

and obtain the binary *split-set* clustering for the clustering at our wish in */tmp/out-set*.

#### 4.17.4 Implementation

The tool is implemented in *split\_set\_from\_annotation.c*.

## 4.18 split\_set\_from\_swarm

### 4.18.1 General

Using this tool one can generate a binary *split-set* file describing a clustering in the sense of *MNHN-Tree-Tools* from a clustering generated with Swarm version 3 [12].

### 4.18.2 Usage

The tool can be used in the following manner:

```
split_set_from_swarm [fasta] [swarm] [split-set]
```

**Listing 4.42** – Calling the *split\_set\_from\_swarm* tool

with the following arguments:

1. *fasta* The original FASTA file that has been clustered with Swarm version 3 [12]. This is not the downsampled file that you use to create the swarm clustering but the original dataset. (c.f. example).
2. *swarm* The clustering performed by swarm. Swarms output file
3. *split-set* The binary file containing the clustering to be created.

### 4.18.3 Example

We imagine we would like to generate a clustering with swarm of several sequences. Hence we first have to eliminate duplicates and name the duplicity in a way that swarm understands it. The *sequence\_multiplicity* tool outlined in section 4.9, is perfect for this task. Hence we prepare the swarm input file as follows:

```
sequence_multiplicity test.fasta > swarm-in.fasta
```

**Listing 4.43** – Generating an input for swarm using the *sequence\_multiplicity* tool

We than run the swarm tool to cluster our dataset.

```
swarm swarm-in.fasta > swarm-output
```

**Listing 4.44** – Calling Swarm before the *split\_set\_from\_swarm* tool

and can finally built a *split-set* cluster file that uses the interal structures of the tools that are presented in this manual.

```
split_set_from_swarm test.fasta swarm-output /tmp/split-set
```

Listing 4.45 – Example of the *split\_set\_from\_swarm* tool

Note that *test.fasta* is used in the final call and not *swarm-in.fasta*. The tool can by tracking the correct annotations find the correct sequences in the original dataset and built a binary *split-set* file accordingly.

#### 4.18.4 implementation

The tool is implemented in *split\_set\_from\_swarm.c*.

### 4.19 filter\_split\_sets\_by\_min

#### 4.19.1 General

The tools allows one to filter clusters obtained from an adaptive clustering run, such as with the tools outlined in section 4.12, that contain less sequences than a certain threshold, and retain sequences that contain more sequence than this threshold.

#### 4.19.2 Usage

The tool is to be called like:

```
filter_split_sets_by_min [minimum] [filtered-split-sets-path-name] \
[input-split-sets]
```

Listing 4.46 – Calling the *filter\_split\_sets\_by\_min* tool

where the arguments are:

1. *minimum* An interger representing the minimum number of sequences to be found in a cluster to be retained in the dataset
2. *filtered-split-sets-path-name* A path and name where the filtered *split-sets* should be stored.
3. *input-split-sets* The unfiltered *split-sets* to filter using this algorithm.

### 4.19.3 example

```
filter_split_sets_by_min 100 /tmp/f100s /tmp/s*
```

Listing 4.47 – Example of the *filter\_split\_sets\_by\_min* tool

Here the program takes the *split-sets*, clusters of each layer, and filters clusters that contain less than 100 sequences from the dataset. The new dataset without these clusters is written to */tmp/f100sX* where *X* is the number for each layer in the output splitset. The selection using the wildcard works as the original split sets are name */tmp/sX* and the numbers *X* are created in a way by the adaptive clustering tools such that the wildcard selection keeps them in correct order.

### 4.19.4 implementation

The code and the interface is implemented in *filter\_split\_set\_by\_min.c*. The filtering mechanism in *cluster\_io.c*

## 4.20 find\_sequence\_in\_split\_sets

### 4.20.1 General

The tool allows you to find a specific sequences among the multiple clusters obtained from an adaptive clustering run that was performed with the tools outlined in section 4.12.

### 4.20.2 Usage

The utility can be called like:

```
find_sequence_in_split_set [fasta-ds] [fasta-seq] [split-sets]
```

Listing 4.48 – Calling the *find\_sequence\_in\_split\_sets* tool

with the arguments:

1. *fasta-ds* A FASTA file containing the entire dataset of sequences that an adaptive clustering run was performed on.
2. *fasta-seq* A FASTA file containing a single sequence, that should be searched for in the obtained clusters
3. *split-sets* A list of binary files that represent the different layers of clusters obtained from an adaptive clustering run.

### 4.20.3 Example

```
find_sequence_in_split_set test.fasta seq.fasta /tmp/s*
```

Listing 4.49 - Example of the *find\_sequence\_in\_split\_sets* tool

This searches for the sequence in *seq.fasta* which has to reside in *test.fasta* in the different *split-sets* selected with */tmp/s\**. The wildcard here as with most tools that work with multiple *split-sets* works as they are numbered correctly for such a selection as they are generated by an adaptive clustering run. The tool then prints an output like this:

```
Sequence has internal ids:  
0  
Sequence is found in Clusters  
L22C0 1 times  
L23C0 1 times  
L24C0 1 times  
L25C0 1 times  
L26C0 1 times  
L27C0 1 times
```

Listing 4.50 - Output of the *find\_sequence\_in\_split\_sets* tool

Which means that in FASTA outputs of the clusters, such as those generate with the tool *split\_set\_to\_fasta* (c.f. section 4.13), the sequence is outlined in the fasta file with the name *>sequence\_0* as it is in this case the first sequence of the original dataset. The sequence is found in clusters named *LXY*. Where *X* represents the layer number and *C* the cluster number. As such the sequence is found the binary *split-set* with the number *X* at its end. Once this binary *split-set* is converted to clusters, i.e in FASTA form the sequence is found in the files with the suffix *Y*. One may also say that in the tree built from such an adaptive run the sequence is to be found in the tree layer *X* in cluster *Y*.

### 4.20.4 Implementation

The interface as well as the tool itself is implemented in *find\_sequence\_in\_sp*

## 4.21 print\_connections

The *print\_connections* tool allows you to highlight connections, just as printed in the output file of an adaptive clustering run. This means it shows which cluster of a more dense

stage of an adaptive clustering run is embedded into a cluster of less dense stage of an adaptive clustering run. A connection is formed if 80% of the sequences of a cluster of a higher density is found in a cluster with a lower density.

### 4.21.1 Usage

```
print_connections [lower-layer-split-set] [upper-layer-split-set]
```

Listing 4.51 - Calling the *print\_connections* tool

where *lower-layer-split-set* and *upper-layer-split-set* are binary clusterings obtained from an adaptive clustering run that was performed with the tools highlighted in section 4.12. The *upper-layer-split-set* should be a clustering with a higher density than the *lower-layer-split-set*.

## 4.22 tree\_map\_for\_sequence

This tool allows you to find a certain sequence in a Newick tree. This tool creates a *map* file that you can use in conjunction with the Newick utilities [8] published by the University of Geneva, in order to highlight clusters in different colors that contain a sequence in question.

### 4.22.1 Usage

The *tree\_map\_for\_sequence* tool can be used:

```
tree_map_for_sequence [fasta-ds] [fasta-seq] [color] [split-sets] > [map]
```

Listing 4.52 - Calling the *tree\_map\_for\_sequence* tool

with the following arguments:

1. *fasta-ds* A FASTA file containing the full dataset of that has been used to perform an adaptive clustering run with the tools shown in section 4.12.
2. *fasta-seq* A FASTA file containing the single sequence to be outlined in a Newick tree by using this tool in conjunction with the Newick utilities.
3. *color* A string telling which color to use to mark clusters that contains the sequence in question. The color has to be expressed in a Cascade Style Sheet (CSS) compatible

way. Hence simple color expressions such as red, green and blue should work. In general it is preferred to use colors in the format #RRGGBB where lead by a hash and three hexadecimal numbers for the color values of red, green and blue are used. Hence in order to get an orange color you might use, #FF8800 which corresponds to full red, about half green and no blue.

4. *split-sets* All the binary clusterings that built the tree. And that were used when running the *split\_sets\_to\_newick* tool, outlined in section 4.15, in order to generate the dendrogram that the tree visualization using the Newick tools will use finally.
5. *map* The resulting *map* file that can be used together with the Newick utilities to generate a tree where all clusters containing a specific sequence are colored.

### 4.22.2 Example

```
tree_map_for_sequence test.fasta seq.fasta '#FF0000' /tmp/s* > tree.map
```

Listing 4.53 – Example of the *tree\_map\_for\_sequence* tool

In this example we generate a map file *tree.map* to color all clusters that contain the sequence contained in *seq.fasta* red in a tree that we print using the Newick utilities. The tree was built from the complete dataset of sequences found in *test.fasta*. The binary *split-sets* that build the different layers of the tree are found at */tmp/sX* where *X* is the number of each layer. And as with all the previous tools the wild card selection should work perfectly. To better illustrate the usage of the tool we show in the following in concordance with the the *split\_sets\_to\_newick* found in section 4.15, and the Newick utilities [8] by the university of Geneva.

```
split_sets_to_newick 0 test.fasta 0 /tmp/s* > tree.dnd
tree_map_for_sequence test.fasta seq.fasta '#FF0000' /tmp/s* > tree.map
nw_display -w 600 -rs -i 'opacity:0' -b 'opacity:0' -l 'opacity:0' \
-c tree.map tree.dnd > tree.svg
```

Listing 4.54 – Further example of the *tree\_map\_for\_sequence* tool

where *tree.dnd* is the Newick file that corresponds to the map *tree.map* that we generate with the tool outlined in this section. The *nw\_display* from the Newick utilities [8] generates an svg

image without annotations of a with of 600 points in radial fashion. It uses the aforementioned map to color clusters in the tree red that contain the sequence in question. An example tree generated like this is shown in figure 4.3.

### 4.22.3 Implementation

The *tree\_map\_for\_sequence* tool is implemented in *tree\_map\_for\_sequence.c*.

## 4.23 tree\_map\_for\_split\_set

The *tree\_map\_for\_split\_set* tool allows you to highlight clusters in trees in different ways. First you can identify different clusters in a tree. Second you can visualize how many sequences of certain cluster are found in a different cluster. This tool can also be used in conjunction with different graphics software to build more complex trees by overlaying the graphics that can be generated with the help of this tool. This tool generates *map* files that you can use together with the Newick tools published by the university of Geneva [8].

### 4.23.1 Usage

The tool can be called in the following manner:

```
tree_map_for_split_set [split-set] [fasta-ds] [number-of-clusters] \
[gradient-type] [cluster-index-colors] [split-sets-full-tree] > [map]
```

Listing 4.55 – Calling the *tree\_map\_for\_split\_set* tool

where the arguments are the following:

1. *split-set* The split set to color, highlight on the Newick tree.
2. *fasta-ds* The complete dataset from which an adaptive clustering run has generated clusters and hence a tree structure.
3. *number-of-clusters* how many clusters to color in the same diagram. Currently more than 1 cluster might result in unpredictable behavior.
4. *gradient-type* Colors the clusters from white to the desired color ( by number of sequences of the clusters found in a cluster, 0 sequences for white and all, the color that you have chosen ). The possible values to select are:

#### 4.23. TREE\_MAP\_FOR\_SPLIT\_SET

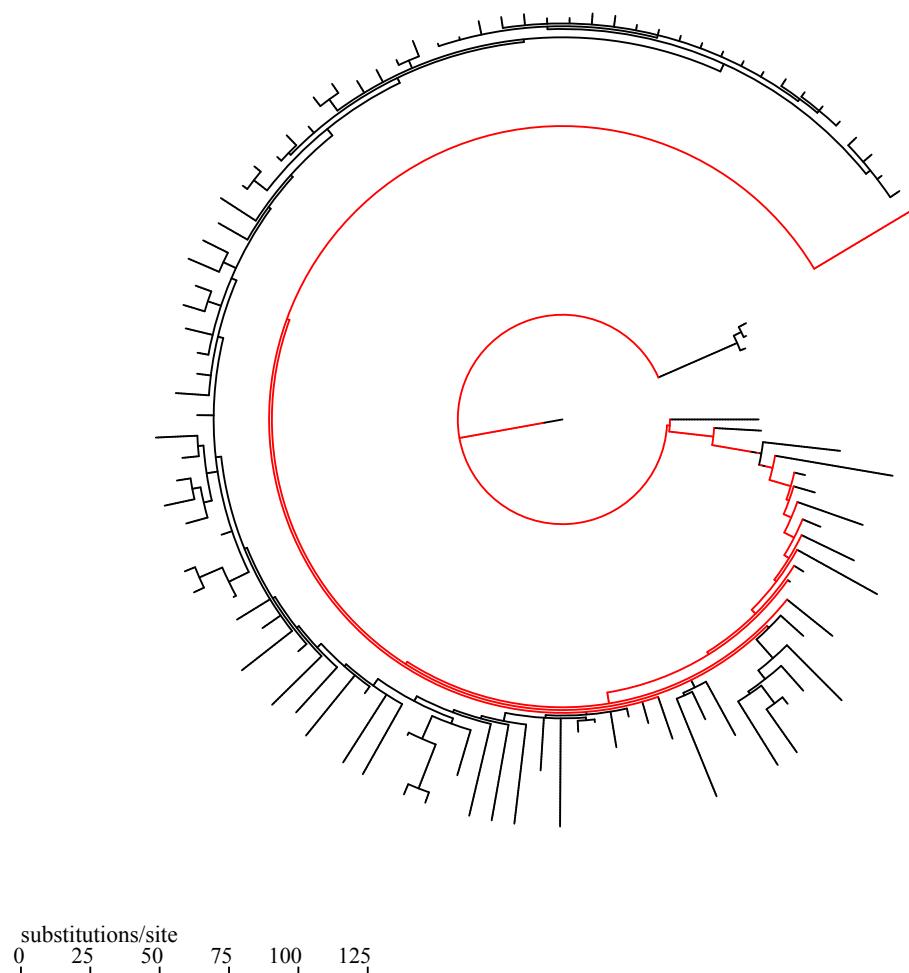


Figure 4.3 – An example of a tree that outlines with a red line the location of a certain sequence. A tree generated using the tools in order as outlined in listing 4.54. The substitutions/site scale automatically generated by the Newick utilities [8] can be ignored.

- 0: no gradient and hence all clusters that have at least 1 sequence are colored in the desired color.
  - 1: a linear gradient from white corresponding to: no sequences, to the desired color corresponding to all sequences.
  - 2: a logarithmic gradient. The logarithmic gradient nevertheless sets 0 sequences to all white.
5. *cluster-index-colors* allows you to select the different colors for the different clusters. As already mentioned this might result in unpredictable behavoir selecting more than 1 cluster and more than one color. This argument has to be formed as a set of strings like '2,#FF0000' '4,#00FF00' where the first number is the number of the cluster to select from the *split-set* which might hold  $n$  clusters. This number is followed by a comma "," and a color value, expressed by 3 hexadecimal number, with the first two being the red component, the second two to the green and the third pair the blue component.
  6. *split-sets* The clusters as obtained by an adaptive clustering run with the tools mentioned in section 4.12.
  7. *map* The resulting *map* file that can be used together with the Newick utilities to generate a tree coloring clusters containing sequences of the *split-set* according to the rules chosen above.

### 4.23.2 Example

```
tree_map_for_split_set cluster_set test.fasta 1 2 '5,#000000' /tmp/s* > tree.map
```

Listing 4.56 – Example of the *tree\_map\_for\_split\_set* tool

Here this tool takes the 5th cluster of from the binary cluster file *cluster\_set* that contains clusters of the sequence dataset found in *test.fasta*. The tool generates a *tree.map* file for a dendrogram that was built from all the binary cluster files stored at */tmp/sX*, where  $X$  corresponds to the layer number. In this case the map is built such that only clusters containing sequences from the 5th cluster of the *cluster\_set* are shown, in a logarithmic gradient from white ( no sequences ) to black, #000000, corresponding to clusters holding all sequences.

Also this tool was at one point designed to permit different clusters to be highlighted as different colors, subsequent modifications of this tool destroyed this functionality and selecting multiple clusters with different colors might result in unexpected behavoir, as such you may limit yourself to highlight only one cluster per dendrogram and use different imaging software to overlay results.

An example for generating a tree in conjunction with other tools might look like:

```
split_sets_to_newick 0 test.fasta 0 /tmp/s* > tree.dnd
tree_map_for_split_set cluster_set test.fasta 1 2 '5,#000000' /tmp/s* > tree.map
nw_display -w 600 -rs -i 'opacity:0' -b 'opacity:0' -l 'opacity:0' \
-c tree.map tree.dnd > tree.svg
```

Listing 4.57 – Extended example of the *tree\_map\_for\_split\_set* tool

where we use the *split\_sets\_to\_newick* tool (c.f. section 4.15) to generate a Newick file, and the *tree\_map\_for\_split\_set* tool shown in this section to generate a corresponding map, highlighting as explained above only sequences from the 5th cluster of the *cluster-set* binary file. We than use *nw\_display* from the Newick tools [8] to generate the final diagram.

A diagram generated with a logarithmic gradient might look like the one shown figure 4.4

### 4.23.3 Implementation

The *tree\_map\_for\_split\_set* tool is implemented in *tree\_map\_for\_split\_set.c*.

## 4.24 virtual\_evolution

The tools *virtual\_evolution* and *virtual\_evolution\_controlled* allow you to simulate simple evolutionary processes and generate nucleic sequences resembling to these evolutionary processes. Your simulated biosphere will then reside in a FASTA file.

### 4.24.1 Usage

The evolution simulating in MNHN-Tree-Tools comes in two versions in rougher *virtual\_evolution* tool and in a more controlled, later added *virtual\_evolution\_controlled* tool.

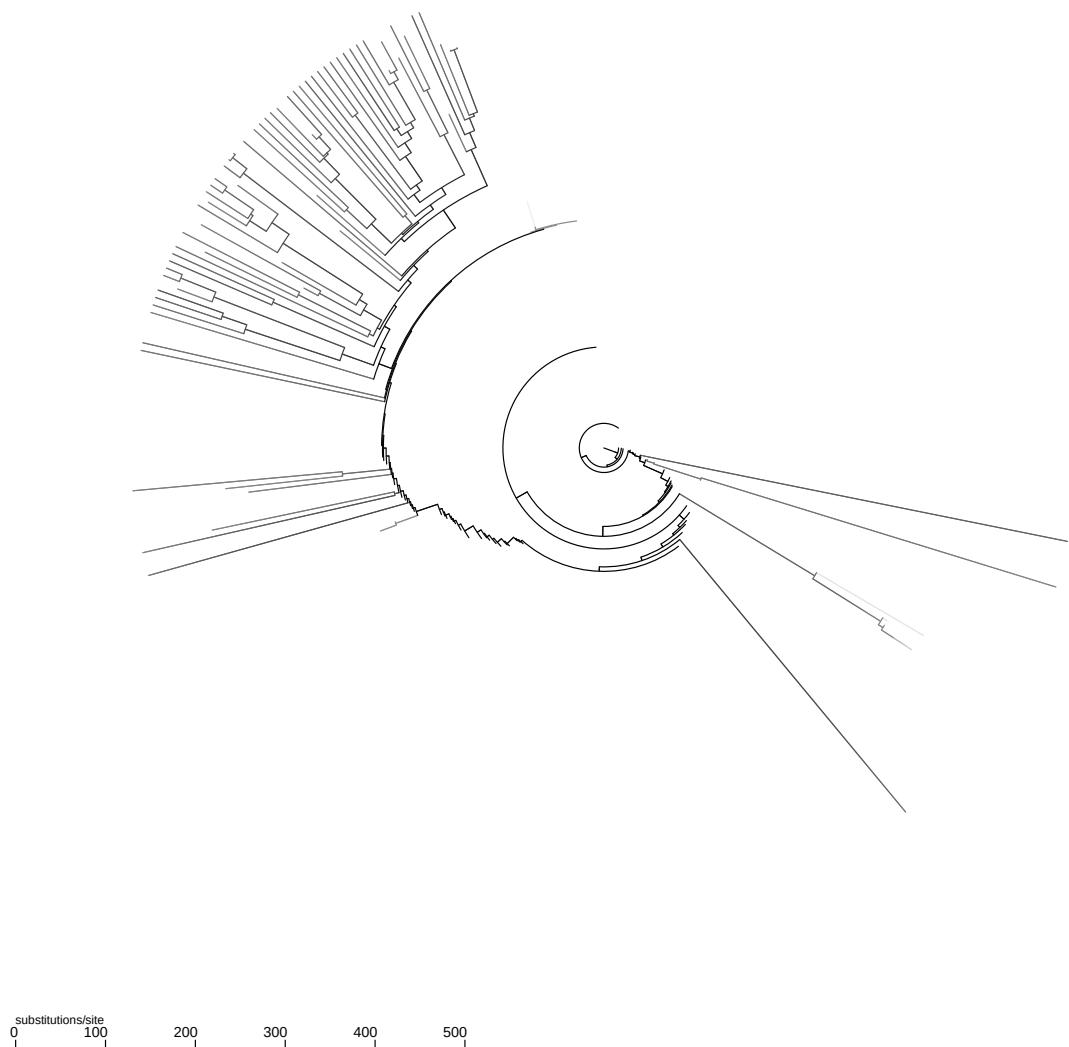


Figure 4.4 - An example of a tree that used the *tree\_map\_for\_split\_set* to only highlight a certain family in a tree. You see the logarithmic gradient as clusters with few sequences are shown in light gray, while clusters containing the entire set of sequences, that belong to the outlined family are dark and black. The substitutions/site are again automatically generated by the Newick tools [8] and are to be ignored.

The tools can be called like:

The *tree\_map\_for\_sequence* tool can be used:

```
virtual_evolution [seed] [seq-length] [n-sequences] [n-partitions] \
    [max-time] [rate] [n-mutations-to-amplification-template] [fasta]
virtual_evolution_controlled [seed] [seq-length] [n-sequences] [n-partitions] \
    [max-time] [rate] [n-mutations-to-previous-amplification-template] [fasta]
```

**Listing 4.58 – Calling the *virtual\_evolution* tools**

with the following parameters:

1. *seed* The seed, to pseudorandom number generators used, that permits you to generate the same set of sequences in subsequent calls of this tool.
2. *seq-length* The length of the sequences to be generated.
3. *n-sequences* How many sequences in total have to be generated.
4. *n-partitions* How many families/clusters shall be generated, i.e. how many amplification events should happen.
5. *max-time* If the evolution should occur at random times between amplifications and be limited to *max-time*. -1 here chooses a constant evolution cycle, before a new cluster expansion happens, otherwise a random number of evolution cycles may occur, between 0 and *max-time*.
6. *rate* substitution of bases in the nucleic sequence per evolution cycle.
7. *n-mutations-to-amplification-template* Adds the here chosen number of mutation to the amplification template before cluster expansion. In this case the nucleotide bases are only  $X$  times as chosen here replaced. This actual number of mutations might be lower as the replaced base might be the same as the nucleotide that was already in place.
8. *n-mutations-to-previous-amplification-template* Adds this number of real mutations to the previous amplification template, before generating a new family.
9. *fasta* A FASTA file where to output the sequences in question.

## 4.24.2 algorithm

The algorithms between the two simulation process are different as such we begin to present the,

- *virtual\_evolution* process. The evolutionary process herein is a basic expansion and diffusion process. The way this simulator works is that it generates a template sequence of a given size. It then copy this sequence until it fills a whole family as defined by the *n-sequences* and *n-partitions* arguments. A family or partition as such contains

$$N = n_{\text{seq}} / n_{\text{part}}, \quad (4.21)$$

sequences where  $n_{\text{seq}}$  are the number of total sequences and  $n_{\text{part}}$  are the number of partitions or families to be created. It than performs

$$M = ptr \quad (4.22)$$

mutations on the previous partitions generated. Where  $t$  is a random time, and hence number of cycles number, which is chosen between 0 and  $t_{\max}$  which is defined by the *max-time* parameter.  $r$  is the mutation rate as defined by the *rate* parameter.  $p$  is the number of partitions already generated. Once  $M$  mutation were perfomed on the current state of the dataset. A sequence is randomly chosen from the previous sequencies and will be the template for the next expension of  $N$  sequences. The algorithm is continued until all partitions are created, perfoming a last mutation run on the entire dataset as determined by equation 4.22. The whole process is further outlined in figure 4.5.

- *virtual\_evolution\_controlled* is a different, simpler but more controlable process. The main difference to the uncontrolled *virtual\_evolution\_process* is that the template for each family is chosen from the previous template to create a partition or family, and not randomly chosen from the existing sequences. Further the number of mutations is enforced not be just exchanges of nucleotides that might actually not lead to mutations but to actual mutations. Hence you perfectly know how far the templates between different families are away, and with a fixed rate of mutations can perfectly estimate the diffusion of each partition. Hence we have a perfectly controlled distance-diffusion mechanism. The process is outlined in figure 4.6.

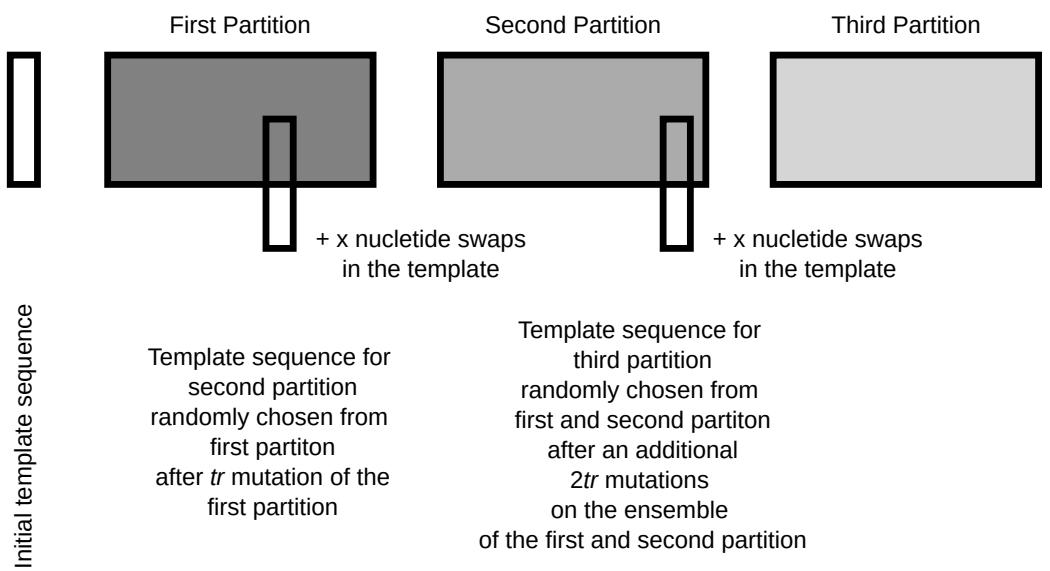


Figure 4.5 – The classic virtual evolution process. The different levels of gray show how diffused, (old) the sequences are at the end of the process. Please note that  $t$  is chosen at random after each partition expansion.

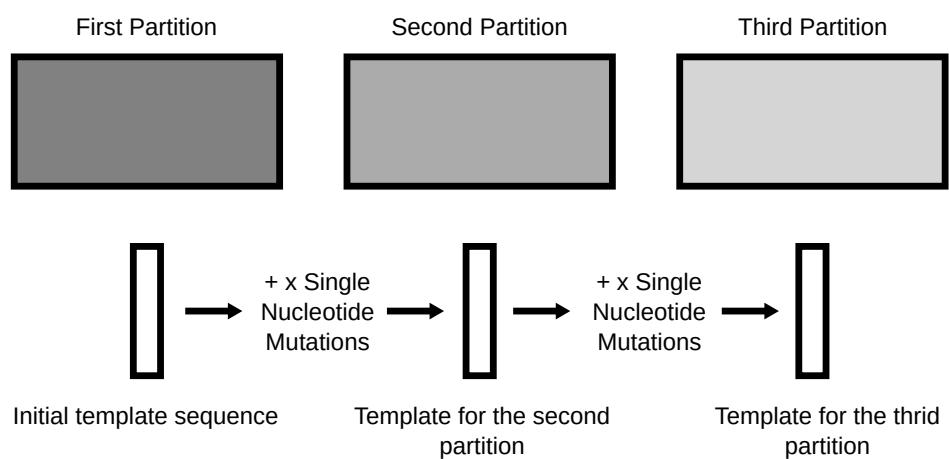


Figure 4.6 – The more restricted controlled virtual evolution process. The different level of gray show how diffused, (old) the sequences are at the end of the process. Again note there are  $M = ptr$  mutations happening across the available sequences after each expansion and that  $t$  might be variable if chosen so.

### 4.24.3 example

```
virtual_evolution_controlled 42 100 10000 10 -1 1000 4 biosphere.fasta
```

Listing 4.59 – Example of using the *virtual\_evolution* tools

Here the evolutionary process generates 10000 sequences of 100 nucleotides, in 10 partitions of 1000 sequences and a mutation rate  $r = 1000$ . The time  $t = 1$ , or number of cycles is fixed to one though that the diffusion stays constant during the process. The template undergoes 4 mutations before creating each new family. The resulting sequences are stored to *biosphere.fasta*.

### 4.24.4 implementation

*virtual\_evolution* and *virtual\_evolution\_controlled* are implemented in *virtual\_evolution.c* and *virtual\_evolution\_controlled.c* respectively.

## 4.25 simulation\_verification

### 4.25.1 General

This is a small tool that tells you how many families generated by a virtual evolution run created using the tools shown in section 4.24 can be found by an adaptive clustering run using the tools shown in section 4.12. This basically is an internal quality measurement tool.

### 4.25.2 Usage

The tool is to be called like:

```
simulation_verification [fasta] [partitions] [precision] [split-sets]
```

Listing 4.60 – Calling the *simulation\_verification* tool

with the arguments:

1. *fasta* A FASTA file containing a virtual evolution dataset as generated with the tools in section 4.24.
2. *partitions* The number of partitions in the dataset as chosen during the virtual evolution run.

3. *precision* How many sequences have to be attributed correctly to a family or cluster in order to be proven to be correct. If you chose for instance 0.8 here a family of 1000 sequences can only be found in clusters that have between 800 and 1200 sequences and hold at least 800 sequences of the original partition.
4. *split-sets* The binary clusters layer files that are the result from an adaptive clustering run with the tools presented in section 4.12.

### 4.25.3 Example

```
simulation_verification sim.fasta 10 0.8 /tmp/s*
```

Listing 4.61 – Example of the *simulation\_verification* tool

Here the simulation verification tool uses the simulated dataset *sim.fasta* which contains 10 partitions and sees if clusters during an adaptive clustering run yielded at least in one layer the correct attribution of a family under the precision criteria 0.8.

A possible output might be:

```
Found 8 out of 10 partitions
```

Listing 4.62 – Example Output of the *simulation\_verification* tool

### 4.25.4 Implementation

The tool is implemented in *simulation\_verification.c*.

## 4.26 pca2densitymap

### 4.26.1 General

The *pca2densitymap* tool allows you to visualize the sequence density in a pca file by generating a portable network graphics (PNG) image. This basically generates a histogram by laying a rectangular grid across a projection into two/three dimensional subspace, spun by principal components. The tool then counts how many sequences are found within a square/cube and attributes this value to a pixel/voxel in the final image/images. The 3D version creates a stack of images. To overcome the limitations of an image based tool such as this one, one may refer to the *pca2densityfile* tool outlined in section 4.27.

## 4.26.2 Usage

The tool can be called like:

```
pca2densitymap [fasta] [pca] [dimensions] [2D or 3D] \
[invertal-length] [png-file-or-files]
```

Listing 4.63 – Calling the *pca2densitymap* tool

and has the followgin arguments:

1. *fasta* The original FASTA file that has been transformed into a k-mer representation which was projected into a subspace spun by principal components. For instance by using the tools *fasta2kmer* and *kmer2pca* as described in sections 4.1 and 4.2.
2. *pca* A file containing the projections onto the principal components.
3. *dimensions* An integer discribing how many dimensions the PCA subspaces consists of.
4. *2D or 3D* Chose 2 for a 2-dimensional output or 3 for a 3-dimensional output. 3-dimensional outputs are not fully tested though.
5. *interval-length* The grid spacing for our histogram.
6. *png-file-or-files* A path for the resulting PNG file. In case of a 3D output a whole stack of images with individual numbers for each slice our created.

## 4.26.3 Example

```
pca2densitymap test.fasta test.pca 7 2 0.1 /temp/out.png
```

Listing 4.64 – Example of the *pca2densitymap* tool

Here an image, *out.png* will be generated that contains from white ,most sequences per grid tile, to black, no sequences per grid tile. As the image is 8 bits per pixel a range from 0 (black) to 255 (white) is available. Which might cause, due to the rescaling only a few white points if the dataset contains very large peaks and compleatly black background. *test.fasta* is the sequence dataset, *test.pca* its projection into a 7 dimensional PCA subspace. The tool always choses only to treat either the first two or three dimensions, for the 2D or 3D case respectivly. We

generate a two dimensional output with a grid spacing of 0.1 [k-mer frequencies]. The tool yields some information about the image generated:

```
shift x: 8.45373440
shift y: 5.56738997
n points x: 122
n points y: 89
factor 255/maximum(intensity): 0.10814250
```

**Listing 4.65 – Text output of the *pca2densitymap* tool**

where shift x and y are origin of the image, in general on the upper left. n Points x and y are the number of pixels according to the spacing chosen that are calculated and stored in the output PNG file. The last value is the scalefactor that has been used in order to scale the the densities from 0 to 255 or from black to dark. A sample of an image generated using this tool is highlighted in figure 4.7.

#### 4.26.4 Implementation

The interface to this tool is implemented in *pca2densitymap.c*. The engine of the tool resides in *density.c*.

### 4.27 *pca2densityfile*

#### 4.27.1 General

The *pca2densityfile* functions similar to the *pca2densitymap* tool outlined in section 4.26. It generates a two dimensional histogram accross a 2 dimensional PCA projection by selection of a grid spacing. The tool then outputs the densities for each tile to a text file that can be used for further processing, or external visualisation tools.

#### 4.27.2 Usage

The *pca2densityfile* tool can be called like:

```
pca2densityfile [fasta] [pca] [dimensions] [interval-length] > [densityfile]
```

**Listing 4.66 – Calling the *pca2densityfile* tool**

with the arguments:

1. *fasta* A FASTA file that has been transformed to a k-mer representation (i.e. using *fasta2kmer* as shown in section

4.27. PCA2DENSITYFILE



Figure 4.7 – The sequence density in a two dimensional PCA subspace of a k-mer representation. The image was upscaled and color inverted for printing using the Gnu Image Manipulation Program (GIMP) [13].

4.1) and where such a k-mer representation has been projected onto a PCA subspace (i.e. using *kmer2pca* as outlined in section 4.2.

2. *pca* The projections onto the principal components.
3. *dimensions* How many principal components have been retained, the dimensionaly of the PCA subspace stored in the *pca* file.
4. *interval-length* The grid distance, i.e. the tile size of the histogram.
5. *densityfile* A text file where for each tile a density value is stored.

### 4.27.3 Example

```
pca2densitymap test.fasta test.pca 7 0.1 > /tmp/out
```

Listing 4.67 – Example of the *pca2densityfile* tool

In this example a dataset of sequences *test.fasta* and a representation in a 7 dimensional PCA subspace obtained from a k-mer representation of the sequences *test.pca* is examined. A histogram is generated using a grid spacing (tile size) of 0.1 [k-mer frequencies]. How many frequencies are found in each tile is printed to the output file */tmp/out*.

```
shift x: 8.45373440
shift y: 5.56738997
n points x:    122
n points y:     89
0 0 0
1 0 0
2 0 0
3 0 0
4 0 0
5 0 0
6 0 0
7 0 0
8 0 0
9 0 0
10 0 0
11 0 0
12 0 0
13 0 0
14 0 0
15 0 0
```

Listing 4.68 – Example output of the *pca2densityfile* tool

In the beginning of the output file you can see a shift x and y value which tells us about the origin of the file, futher it tells you how many tiles it has generated accroding to the *interval-length* parameter. This header is followed by the data which is represented as three values per line. The x and y coordinate ( of the tile ) i.e. the number of tile that you are in, and the number of sequences in this square represents the third value. Such a file can then be used with 3rd party visualization software, such as Blender [14] in order to inspect the density of sequences in three dimensions. Such an example image is shown in figure 4.8. We also provide a Blender import script for blender 2.80 and 2.90 in the following:

```

import bpy

def add_map_from_density_file(filename, objname, delta, scalefactor):
    f = open(filename, "r")

    line = f.readline();
    line_array = line.split();
    shift_x = float(line_array[2]);

    line = f.readline();
    line_array = line.split();
    shift_y = float(line_array[2]);

    line = f.readline();
    line_array = line.split();
    n_points_x = int(line_array[3]);

    line = f.readline();
    line_array = line.split();
    n_points_y = int(line_array[3]);

    verticies = [];
    faces = [];

    n_points = n_points_x*n_points_y

    for i in range(n_points):
        line = f.readline()
        line_array = line.split();
        current_point = [(delta*float(line_array[0])-shift_x,
                          delta*float(line_array[1])-shift_y,
                          float(line_array[2])*scalefactor)]
        verticies.append(current_point)

    for j in range(n_points_y-1):
        for i in range(n_points_x-1):
            current_face = [(j*n_points_x+i,(j+1)*n_points_x+i,
                            (j+1)*n_points_x+i+1,j*n_points_x+i+1)]
            faces.append(current_face)

    themesh = bpy.data.meshes.new(objname)
    theobject = bpy.data.objects.new(objname, themesh)

```

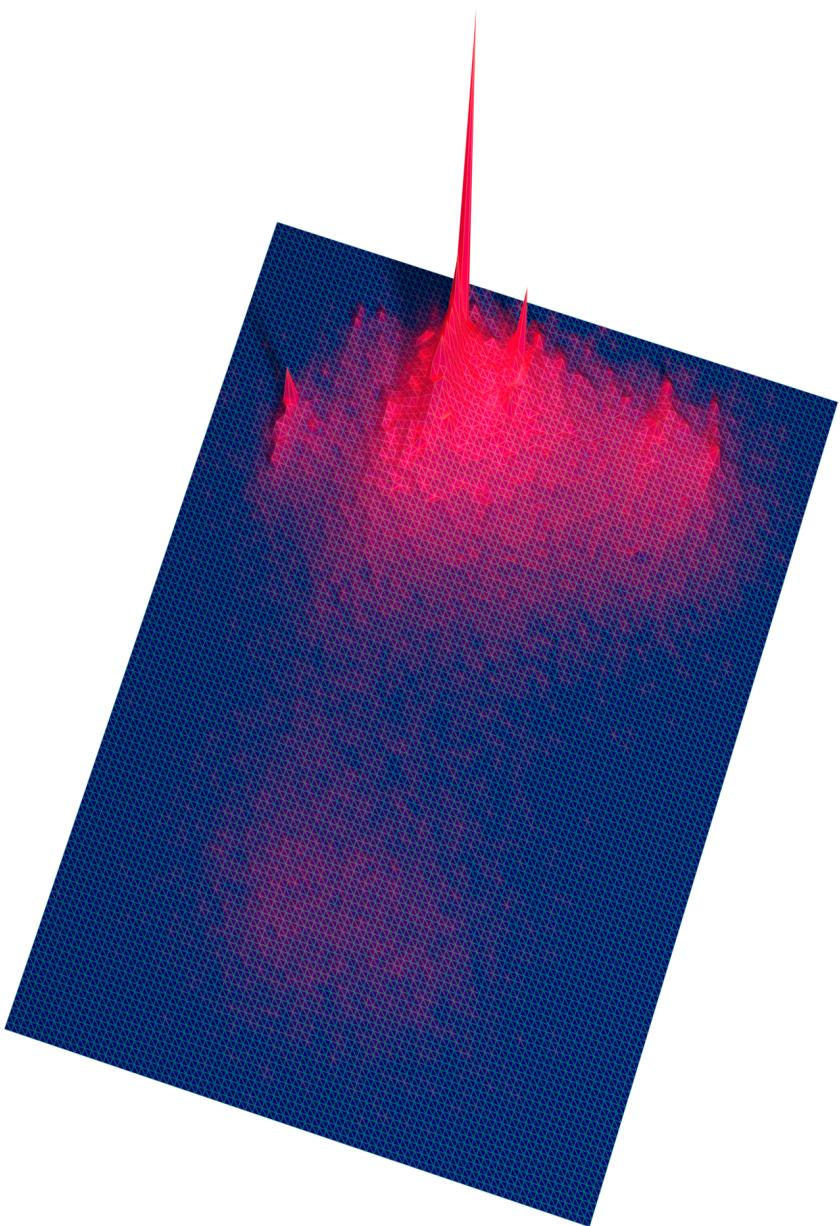


Figure 4.8 - The output of *pca2densityfile* visualized with Blender [14]

```

bpy.data.objects.new(objname, themesh)
bpy.context.scene.collection.objects.link(theobject)

themesh.from_pydata(verticies, [], faces)
themesh.update(calc_edges=True)

return theobject

rescale = 5.;

add_map_from_density_file("/tmp/densityfile", "surface-name", 0.1, 0.03/rescale)

```

Listing 4.69 - Python script to import the output of the *pca2densityfile* tool into blender

where you have to adapt */tmp/densityfile* to the location of your output, *surface-name* to the name that you want to give to your 3D surface and *0.1* to the grid spacing that you have chosen and *0.03* to your rescale factor.

#### 4.27.4 Implementation

The interface to this tool is implemented in *pca2denistyfile.c*. The inner workings are implemented in *density.c*.

## 4.28 reverse\_with\_mask

### 4.28.1 General

This tool allows you to reverse certain sequences in a FASTA file. A mask text file has to be provided to indicated which sequences shall be reversed. This tool only changes read direction of the selected sequences. A more elaborate tool is the *reverse\_complement\_with\_mask* tool found in section 4.29.

### 4.28.2 Usage

The *reverse\_with\_mask* tool can be called like:

```
reverse_with_mask [fasta] [mask] > [fasta-out]
```

Listing 4.70 - Calling the *reverse\_with\_mask\_tool* tool

with the following arguments:

1. *fasta* A FASTA file to reverse sequences in.
2. *mask* A mask file to tell the utility which sequences to reverse. The file has to contain a line with a number for each

sequence. If the number matches 1 this sequence is reversed in the output file.

3. *fasta-out* The FASTA dataset containing all sequences with those reversed where the mask matches.

### 4.28.3 Example

We imagine the following FASTA file with three sequences:

```
> sequence_1
ACTG
> sequence_2
ACCT
> sequence_3
GGAT
```

Listing 4.71 – Example of the *reverse\_with\_mask* tool - FASTA file

We would like to reverse the second sequence and hence write a mask file like the following:

```
0
1
0
```

Listing 4.72 – Example of the *reverse\_with\_mask* tool - Mask file

Calling the tool like:

```
reverse_with_mask test.fasta test.mask > out.fasta
```

Listing 4.73 – Example of the *reverse\_with\_mask* tool

will produce the output:

```
> sequence_1
ACTG
> sequence_2
TCCA
> sequence_3
GGAT
```

Listing 4.74 – Example of the *reverse\_with\_mask* tool - Output FASTA file

### 4.28.4 implementation

The tools interface is implemented in *reverse\_with\_mask.c*. The code that reverses the sequences resides in *dataset.c*.

## 4.29 reverse\_complement\_with\_mask

### 4.29.1 General

This tool allows you to reverse complement certain sequences in a FASTA file. A mask text file has to be provided to indicated which sequences shall be reversed. This tool only changes read direction of the of the selected but also correctly complements the nucleotides.

### 4.29.2 Usage

The *reverse\_complement\_with\_mask* tool can be called like:

```
reverse_with_mask [fasta] [mask] > [fasta-out]
```

**Listing 4.75** – Calling the *reverse\_complement\_with\_mask\_tool* tool

with the following arguments:

1. *fasta* A FASTA file to reverse sequences in.
2. *mask* A mask file to tell the utility which sequences to reverse complement. The file has to contain a line with a number for each sequence. If the number matches 1 this sequence is reverse complemented in the output file.
3. *fasta-out* The FASTA dataset containing all sequences with those reverse complemented where the mask matches.

### 4.29.3 Example

We imagine the following FASTA file with three sequences:

```
> sequence_1
ACTG
> sequence_2
ACCT
> sequence_3
GGAT
```

**Listing 4.76** – Example of the *reverse\_complement\_with\_mask* tool - FASTA file

We would like to reverse complement the second sequence and thus write a mask file like the following:

```
0
1
0
```

**Listing 4.77 – Example of the *reverse\_complement\_with\_mask* tool - Mask file**

Calling the tool like:

```
reverse_complement_with_mask test.fasta test.mask > out.fasta
```

**Listing 4.78 – Example of the *reverse\_complement\_with\_mask* tool**

will produce the output:

```
> sequence_1
ACTG
> sequence_2
AGGT
> sequence_3
GGAT
```

**Listing 4.79 – Example of the *reverse\_complement\_with\_mask* tool - Output FASTA file**

#### 4.29.4 Implementation

The tools interface is implemented in *reverse\_complement\_with\_mask.c*. The code that reverses the sequences resides in *dataset.c*.

### 4.30 digest\_X

The *digest\_X* tools allow you to virtually digest a sequence using restriction enzymes creating a FASTA file with the sequences cutted appart. Currently following tools corresponding to the following restriction enzymes are implemented: *digest\_XbaI*, *digest\_XmnI* and *digest\_HindIII*.

#### 4.30.1 Usage

The tools can be used in the following manner:

```
digest_XbaI [fasta] [n-threads] > [fasta-out]
digest_XmnI [fasta] [n-threads] > [fasta-out]
digest_HindIII [fasta] [n-threads] > [fasta-out]
```

**Listing 4.80 – Calling the *digest\_X* tools**

with the arguments:

1. *fasta* A FASTA file. Only the first sequence of the *fasta* file will be digested.
2. *n-threads* How many threads to use for this digestion process.
3. *fasta-out* A fasta file containing the digested sequences.

### 4.30.2 Example

```
digest_HindIII test.fasta 8 > digested.fasta
```

Listing 4.81 – Example of the *digest\_X* tools

Digests the first sequence stored in *test.fasta* and stores the resulting sequences in *digested.fasta*. The digestion process can use up to eight threads on the machine it runs on. The restriction enzyme used for this virtual digestion is HindIII.

### 4.30.3 Implementation

The interface to the *digest\_X* tools is implemented in *digester.c*. The digestion engine is implemented in *restriction\_digest.c*.

## 4.31 replace\_N\_sequence

### 4.31.1 General

*replace\_N\_sequence* Is a tool to replace sequences that are holding N letters in them with all possible combinations possible. This tool currently only works with nucleotide sequences.

### 4.31.2 Usage

The tool can be called like:

```
replace_N_sequence [fasta] [sequence-index] [fasta-out]
```

Listing 4.82 – Calling the *replace\_N\_sequence* tool

where the arguments are as follows:

1. *fasta* A FASTA file holding sequences
2. *sequence-index* An index to select a sequence from the *fasta* file. In this tool the first sequence is selected by 1.
3. *fasta-out* All possible permutations by replacing the N letters in the sequence.

### 4.31.3 Example

Let us imagine a fasta file like the following:

```
>sequence1
ACGT
>sequence2
ANNT
>sequence3
ANTG
```

**Listing 4.83** – Fasta example for the *replace\_N\_sequence* tool

We would like to get all sequences possible by replacing the N letters in the second sequence, hence we call the utility like:

```
replace_N_sequence test.fasta 2 /tmp/out.fasta
```

**Listing 4.84** – Fasta example for the *find\_satellite* tool

And get a */tmp/out.fasta* that looks like:

```
>sequence_0
AAAT
>sequence_1
ACAT
>sequence_2
AGAT
>sequence_3
ATAT
>sequence_4
AACT
>sequence_5
ACCT
>sequence_6
AGCT
>sequence_7
ATCT
>sequence_8
AAGT
>sequence_9
ACGT
>sequence_10
AGGT
>sequence_11
ATGT
>sequence_12
AATT
>sequence_13
ACTT
>sequence_14
AGTT
>sequence_15
ATTT
```

**Listing 4.85** – Fasta example for the *replace\_N\_sequence* tool

#### 4.31.4 implementation

The tools interface is implemented in *replace\_N\_sequence.c*. The mechanism is implemented in *kmers.c*.

### 4.32 tree\_pureness

#### 4.32.1 General

*tree\_pureness* is a tool that can compare a Tree and the layer wise cluster separations generated by an adaptive clustering run with the tools outlined in section 4.12 with a ground truth set. Ground truth sets can for instance be generated using the *split\_set\_from\_annotation* (c.f. section 4.17) or *split\_set\_from\_swarm* (c.f. 4.17) tool if a comparison against such tools is the reason for using this utility. *Tree\_pureness* calculates three indices:

1. Pureness Index,
2. Cluster correspondance, and
3. Impure over pure.

This tool as the major tree quality measurement tool if a ground truth for a tree is known.

#### 4.32.2 Usage

The tool can be called like

```
tree_pureness [fasta] [ground-truth] [split-sets] > [output]
```

Listing 4.86 – Calling the *tree\_pureness* tool

with the following arguments:

1. *fasta* A FASTA dataset of sequences that a tree has been built from, and that a ground truth clustering set is known from.
2. *ground-truth* A binary *split-set* file that contains the clustering of the ground truth.
3. *split-sets* The resulting split sets from an adaptive clustering run, that represent a full tree.
4. *output* Output is a text file containing each layer in a tree the three indices that allow for a comparison to ground truth.

### 4.32.3 Algorithm

The program defines the three indices:

1. *Pureness*: We define the Pureness Index to be:

$$P_l = \frac{1}{n_{l,\text{inp}}} \sum_{i=1}^{n_{l,\text{inp}}} \sum_{k=1}^{n_{\text{target}}} \frac{f_{l,i,k}^{<50\%}}{C_{l,i}}, \quad (4.23)$$

where  $l$  represents the layer of a tree,  $n_{l,\text{inp}}$  the number of inpure clusters, hence the clusters that are mixture of two partitions from the initial truth dataset to compare to,  $n_{\text{target}}$  the number of partitions in the original truth dataset,  $f_{l,i,k}^{<50\%}$  the number of elements belonging to partition  $k$  of the truth dataset in the inpure cluster indexed by  $i$  in layer  $l$  if this number represents less than 50% of the elements in the inpure cluster otherwise  $f_{l,i,k}^{<50\%}$  represents the total number of elements in the inpure cluster minus the elements in this cluster that belong to the partition of the truth dataset that is indexed by  $k$ . Finally  $C_{l,i}$  represents the total number of elements in the inpure cluster in layer  $l$  indexed by  $i$ .

2. *Cluster correspondance*: It is straight forward to reason that pureness as defined above is not enough, as a partitioning with enough small clusters would yield highly pure clusters but not a meaningful partitioning, as such we further define:

$$S_l = \frac{|n_l - n_{\text{target}}|}{n_{\text{target}}}, \quad (4.24)$$

where  $n_l$  is the number of clusters in layer  $l$  to be the measure of correspondance between the number of clusters in the truth partitioning and the number of clusters obtained in our partitioning.

3. *Impureness over pureness*: Finally we argue that we need a third value in order to describe the quality of our tree:

$$I_l = \frac{n_{l,\text{inp}}}{n_l}. \quad (4.25)$$

$I_l$  represents the number of inpure clusters, hence those who contain mixtures of partitions from the truth partitioning over the number of clusters that are pure.

An example of clustering and the calculation of the three indices is shown in figure 4.9.

#### 4.32. TREE\_PURENESS

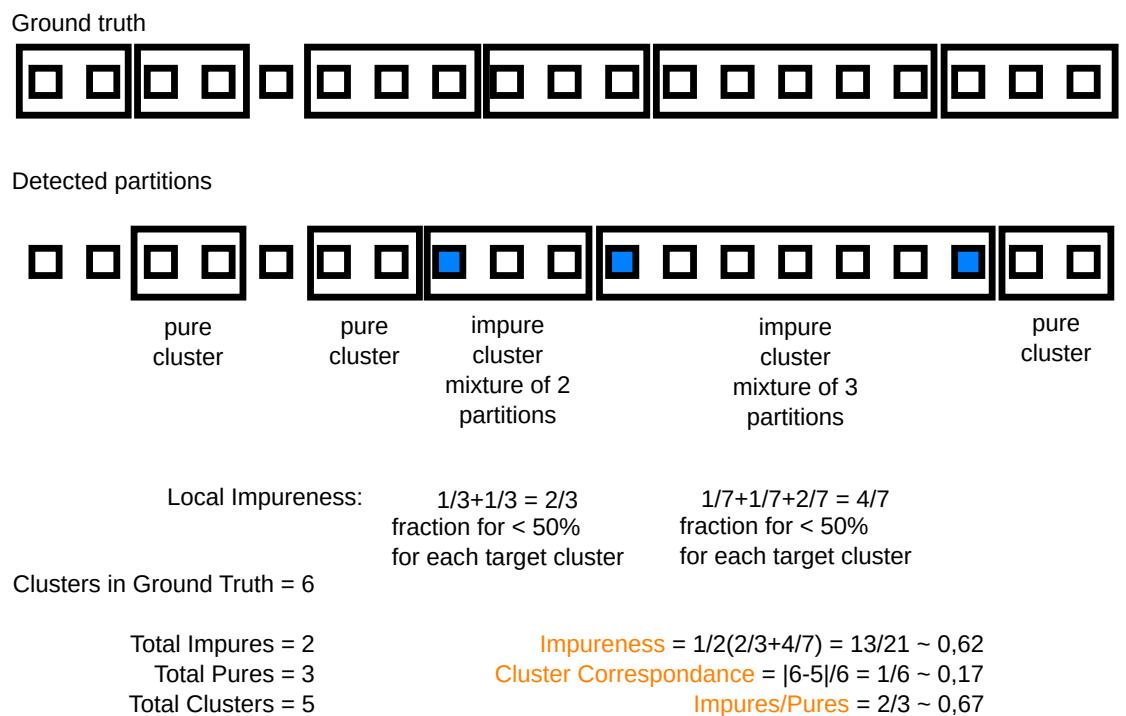


Figure 4.9 - A group truth and a computed clustering. The figure shows how the cluster indices: Purity, Cluster Correspondence and Impureness over Purity are calculated.

#### 4.32.4 Example

We call *tree\_purness* like

```
tree_purness test.fasta ground-truth /tmp/s* > /tmp/pureness
```

Listing 4.87 – Example of the *tree\_pureness* tool

where *test.fasta* is a dataset and *ground-truth* is a binary *split-set* describing the clustering we care comparing all the clusterings in the tree to. */tmp/s\** is the wildcard to select all the binary clusterings that have been produced for a complete tree using an adaptive clustering run (c.f. section 4.12). A resulting */tmp/pureness* then might look like this:

0.207187	1.015385	0.259542
0.207658	0.761538	0.288210
0.214303	0.407692	0.267760
0.267122	0.230769	0.281250
0.270319	0.007692	0.297710
0.278394	0.176923	0.280374
0.333774	0.330769	0.264368
0.301733	0.415385	0.315789
0.327885	0.569231	0.267857

Listing 4.88 – Example output of the *tree\_pureness* tool

with three tablator seperated values per line, where the first value is the Pureness, the scond the Cluster Correspondance and the third the Impureness over Pureness value. Using *gnuplot* for instance one can create a three value diagram in order to judge if a level of the tree yields the accuracy requested to the ground truth. Such an example is shown in figure 4.10.

#### 4.32.5 Implementation

The interface is implemented in *tree\_purness.c*. The functionality is to be found in *cluster\_io.c*.

#### 4.32. TREE\_PURENESS

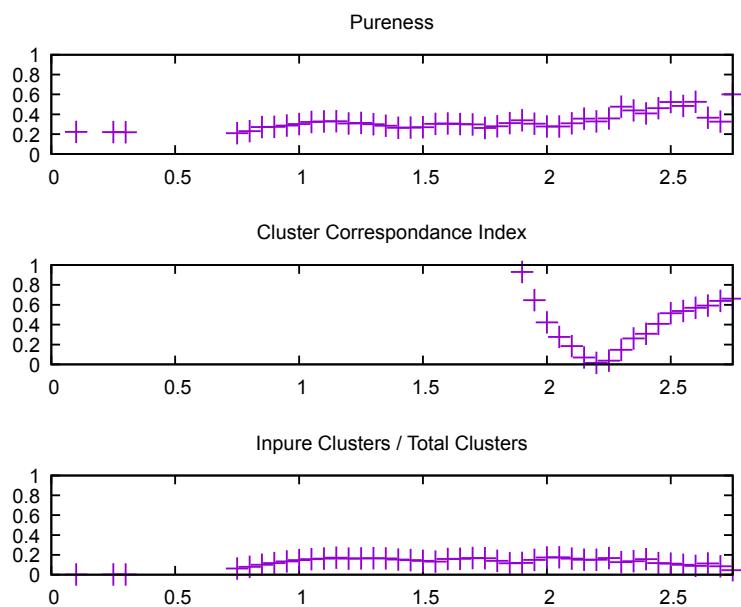


Figure 4.10 – An example of a visual representation of the data obtained from the *tree\_pureness* tool. On the abscissa the  $\epsilon$  value is shown, which resides in the output of the adaptive clustering run for each value. (c.f. section 4.12).

CHAPTER 4. TOOLS

## A | Bibliography

- [1] Lipman D, Pearson W. Rapid and sensitive protein similarity searches. *Science*. 1985;227(4693):1435–1441. Available from: <https://science.sciencemag.org/content/227/4693/1435>.
- [2] Ester M, Kriegel HP, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96. AAAI Press; 1996. p. 226–231.
- [3] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987;20:53 – 65. Available from: <http://www.sciencedirect.com/science/article/pii/0377042787901257>.
- [4] Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*. 2000 Jun;16(6):276–277. Available from: [https://doi.org/10.1016/S0168-9525\(00\)02024-2](https://doi.org/10.1016/S0168-9525(00)02024-2).
- [5] Williams T, Kelley C. Gnuplot 5.2: an interactive plotting program; 2017. <http://gnuplot.info/>. Available from: <http://gnuplot.info>.
- [6] Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: *Cercopithecus solatus*. *BMC Genomics*. 2016 Nov;17(1):916. Available from: <https://doi.org/10.1186/s12864-016-3246-5>.
- [7] Olsen G. Interpretation of the "Newick's 8:45" Tree Format Standard; 1990. [https://evolution.genetics.washington.edu/phylip/newick\\_doc.html](https://evolution.genetics.washington.edu/phylip/newick_doc.html).

## APPENDIX A. BIBLIOGRAPHY

- [8] Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*. 2010 May;26(13):1669–1670. Available from: <https://doi.org/10.1093/bioinformatics/btq243>.
- [9] GNU P. Free Software Foundation. Bash (3.2. 48)[Unix shell program]; 2007.
- [10] Kahan W. Pracniques: Further Remarks on Reducing Truncation Errors. *Commun ACM*. 1965 Jan;8(1):40. Available from: <https://doi.org/10.1145/363707.363723>.
- [11] Anderson E, Bai Z, Dongarra J, Greenbaum A, McKenney A, Du Croz J, et al. LAPACK: A Portable Linear Algebra Library for High-Performance Computers. In: Proceedings of the 1990 ACM/IEEE Conference on Supercomputing. Supercomputing '90. Washington, DC, USA: IEEE Computer Society Press; 1990. p. 2–11.
- [12] Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*. 2015;3:e1420.
- [13] The GIMP Development Team. GIMP;. Available from: <https://www.gimp.org>.
- [14] Community BO. Blender - a 3D modelling and rendering package. Stichting Blender Foundation, Amsterdam; 2020. Available from: <http://www.blender.org>.