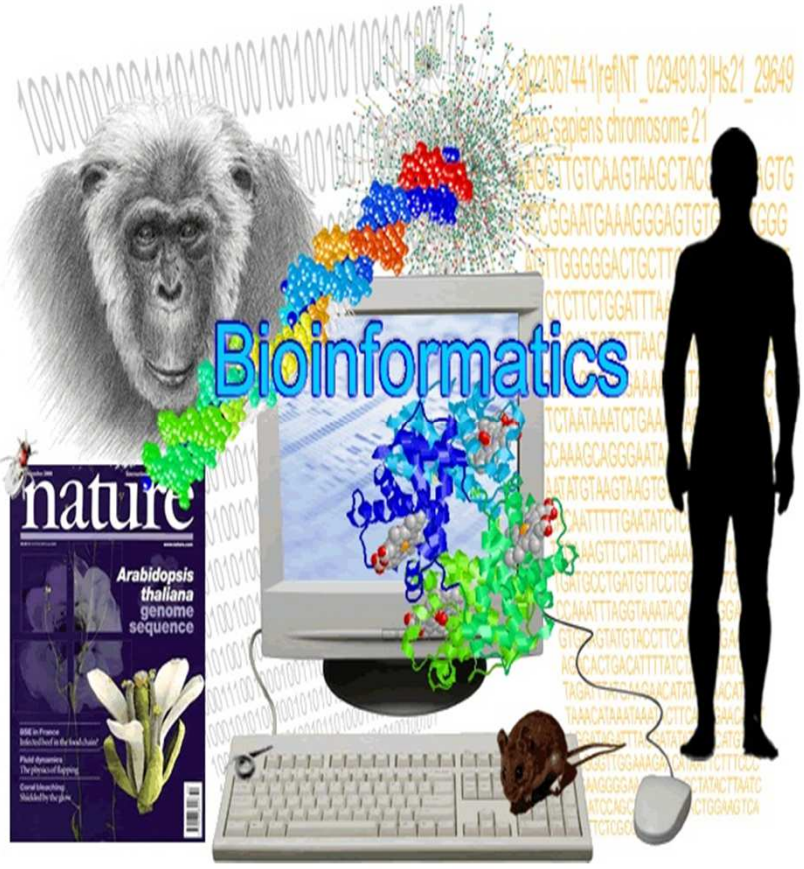
Bitmap Indexes and NoSQL for Identifying Species with DNA Signatures through Metagenomics Samples

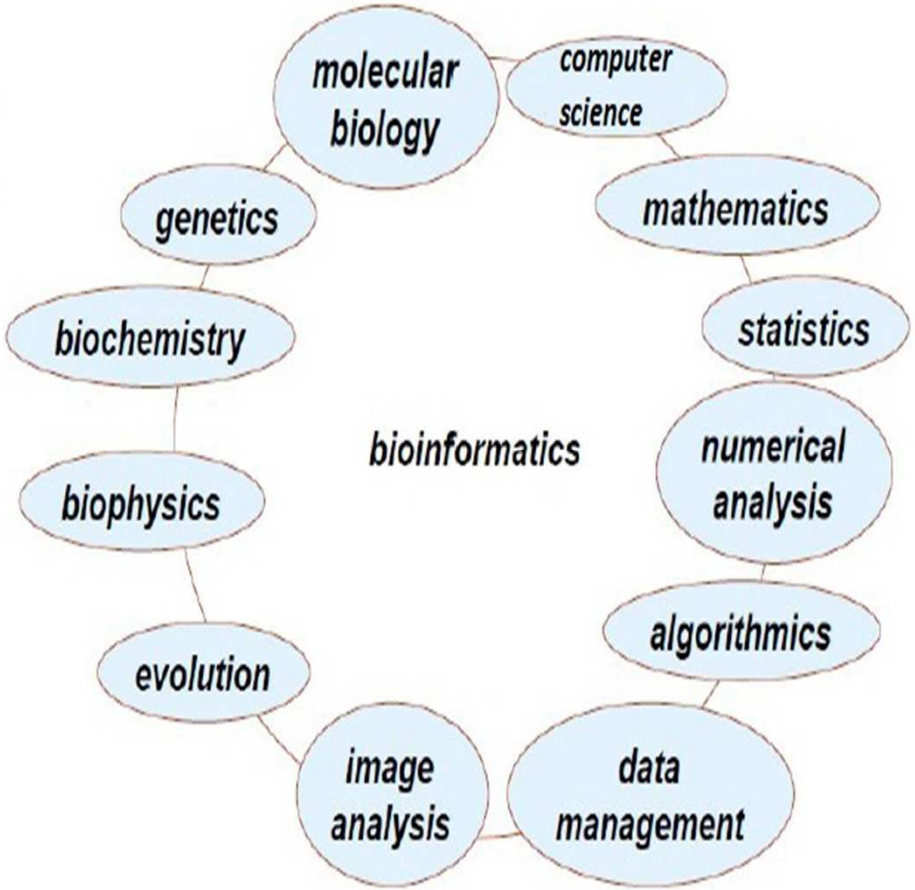
Présenté par : **Ramin Karimi**

Sous l'encadrement de : **Pr. Ladjel BELLATRECHE**
Pr. Patrick GIRARD

le 10 /04/2014

Bioinformatics is a combination of several fields of science



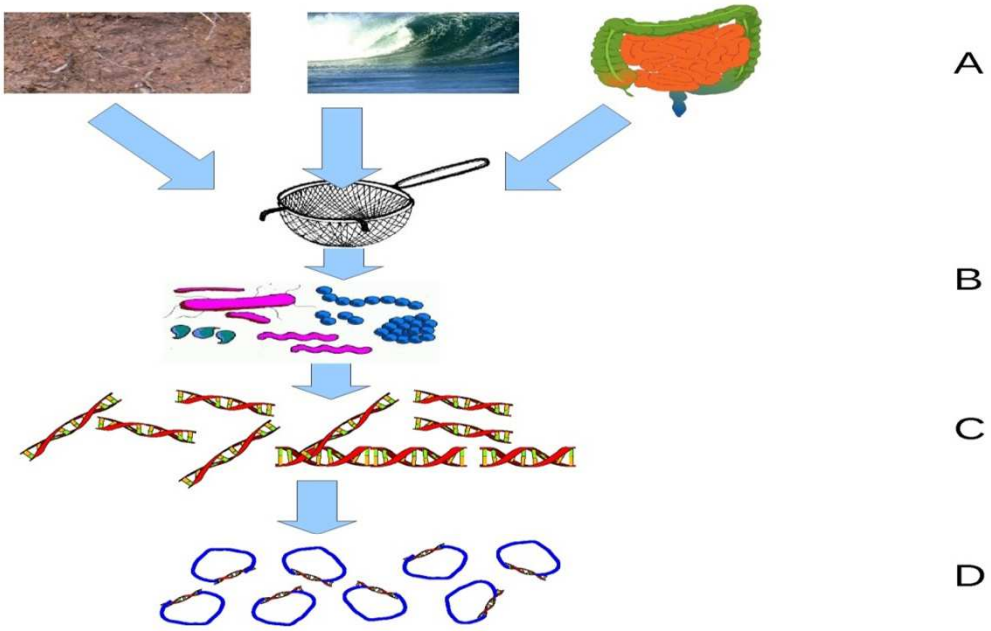


Metagenomics

Metagenomics is the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments.

The use of Metagenomics

- ✓ Medicine
- ✓ Biofuel
- ✓ Environmental remediation
- ✓ Biotechnology
- ✓ Agriculture
- ✓ Ecology



A T C G

TGCACCCTGAAAAATCAACGC
ACTTAAGAGCATCTTGGCAG
AAAGGGGCTGAGGTTGCTGT
TAATTAGCCGTT

Short Reads: The output of sequencing technology is short fragments of DNA sequence with 25 base pairs (bp) to 900 (bp) length

>vibrio011c021p1k
AGTGCATCTGCTTTAACTGCAATTTCCAGGTTAGTATAAGCTTGAGCTTCTTCTCTAACT
TCTAACCCTTCCAGTTTGCACACCAAACCTTTATGACGTTTCCAGCTTTCGACTGAAAAC
TCGACCCACTCA

>vibrio011b11.q1k
GCCGTCGTTCTGCAGCGTCGCGACTGGGTAAACCCATAAGCGCACACTTAAAATAGTCG
GATAGCAGAAGCGAGAATATTAACGAGGTTGTTGCCTATTAGAATTAAACCAATAAGTCG
GTCTGGGCGACTTAGTAATTTCTCAACACGCTGTGCCCCCTTATGACCTTGATTGGATAA
GTGTTTTAAGCGATATCGATTTAAAGACATCATGCCGGTTTCAGAGCTTGAAAATAACCA
GATGC

>vibrio011c02.p1k
GACAACGACGACGACTTAACCCCTTCATTATCACGTCCTTTCTTATCAATGTCCTCCTTA
TTTCTTTATCCACAACTAAGGTAGTGCAAGAACATCACCTTAAAAAAAAGCCCTAGAATT
GGAAACAATAAGCGACTGCGGTTATGAGTACTAACGTATTAGTACTTTGTATTTAAATAA
ATAAGTAAATAAATATTTAACAAAAGAATAAGAGCATAAGTACCTCCCATTCATTTTTAT
TAACCTTATCTCTCACACTAACTCCATCAATTAAAAATGAACGGGCCCCCAGCCAGTTCA
AACCATCGTTATAAATCAATTCATCGATATCACTGGCTGCGCTGAACATTTCTTTCGTCA
TCCGACAGGAAGATACAATGCACCCTGAAAAATCAACGCACTTAAGAGCATCTTGGCAG
AAAGGGGCTGAGGTTGCTGTTAATTAGCCGTTATTTCTTTCTCTATGGGCTTACAAACAA
TAACGTGGTCGTGAGTGACTCTGGTCTACTAACTTAACGCATCATCGAGTCATTACCCTT

>vibrio011c03.q1k
CGTCGTTCTGCAGCGTCGCGACTGGGAAAACCCTTCTCAATGTTCAATCTCTGAACCTG
ATTAAAGTAAGTCGAACTATATTTACCCAACCTTATCTTTTGATTTACGCTACTTCGGAT
TAAAACGATGCGCTTACATCTATCTGCGCGAACTATTTGTTCTCGAATCGCTTGCTTATT
ATTGGAAAGTTGTTTTTTGGCTATGTTTACTCTTTTTTCAAGATGAGATTGAAAACGGTA
ATCATCCGCAATATAACGGAATAGATAAAGTTGCTCACAGATCTCTTCAATTATCTCTAA
TCTATCCGCTTTAGCCGCGGTTCCAACCGCTTTTCTAATTAGTTTAATACGTTTAACTAG
CAATGCGTTACATTCAACGAGAGTATATGAGTGGGGAGC

>vibrio011c04p1k
CATTGGGTACGCCCGGCAGGGAAAGCTGGAATCAGCGTCTGGGCTTTTCCAGTCGCGAC
CTGCAGAACGACGGCTAAGCTCTCGAGGATCCGGGTACCGGGCCCATAAGGCCCTATAG
GAGTCGTATTAAGTCGACCATGGTTTTTTCCTCCTGTGTGAAATTGTTATCCGC

Genome:

>gi|158341503|ref|NC_009933.1| Acaryochloris marina MBIC11017 pREB8, complete genome

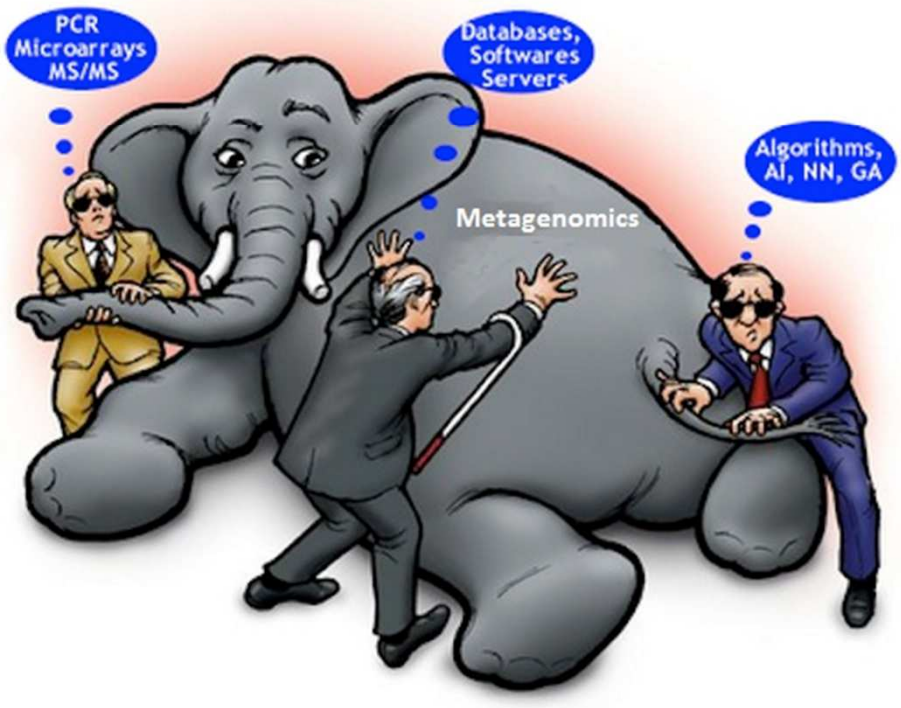
ATTGATTCCTATCATACTCGTGTTTATATGAACGGAAGAGACCTCGGCTTAAAACAAGCTGATGCAGC
TTACATTGCCGAAATATCTACAAGAACCGGACAACGAATTGAGGCCGGCACCCATCAACCTAATCGC
GGCCGCCTTCAGGACCAACGCACCGTTCCCGACCCCTTAGCTGATGTGTGGGAAGACGAGCTAGAA
CCAATGCTGCGTCGAGACCCACGCCTCAAACCCATGACCCTGTATGAGTACCTGCAGGATAAGTATC
CAGGCCAGTATCCCCAAGTCCTGCGGACCCTACAACGTCGGGTGAGAACGTGGAAAGCCTTACATG
GACCAAGCCCTGAAGTGATGTTTGAATTGCGTCATGAACCAGGGGTACAAGGGTTCTCCGATTTTAC
AGAACTCAAAGGCATCACGATTACCATTGCCGGCAAACCCTTTGAGCACCTGATTTACCATTACCGT
CTGGGATACAGCGGCTGGCGATATGCCCAGATCATCGAAGGAGGCGAAAGCTTTGTCGCCCTCTCA
GAAGGATTGCAAAATGCCTTTGCAGCCTGTGGAGGTGTTCCTACACAGCATCGTACTGATAGTTTGA
GTGCAGCCTATCGCAACATGGGCGGCCGCCGGTCCAAAAACCTCACTCGTCTGTACGACGAACTGT
GTGACCACTATCGGCTAGAACCCACTCGTAACAACAAAGGTGTAGCCCATGAGAATGGCTCCATTGA
ATCTCCCCATGGTCATCTGAAGAACCGAATTAAGCAGGCGATCTATCTGCGCGGCAGTGCAGATTTT
ACGAGCGTTGCTGAGTATCAAGCCTTAATTGATGCACAGGTTGCCAAGTTGAATCAGCAGTGCCAAA
CCAAGTATGAGCAAGAGAAAGACCATCTACAACCACTGCCCAAATATCGAACCCCTGACTATGAAGT
GCTCACGGCTAAAGTCAGCAAACGCAGCACCATCGATGTTCGCTGCATTCTATACACCGTCCCTTCT
CGACTGATTGGTCGTCAATTGGAACTGCATCTATACCATGACCGGATTGTCGGCTATCTGGAGCGAC
ACCCGGTGGTGGAATTGCCGAGGAAGCGCGTCAGTGGCAAAGGCAAACGTCGCGACCGTTGCATCA
ACTATCGCCATGTTATTGGTTCAATGCGATTGAAGCCTCGTGCTTTTATCTATTGCACCTGGCAATCAG
ACCTACTTCCCAATCCTGAATACCGCCAAATCTGGGAACAGCTCAAAGCCCAATTTGACCTGGAGCAG
GCTGCCAAGATCATCGTGGAAGCCCTGTATATTGCTGCGGTTCAAGATAAAGAACAGGCCGTAGCAG
TGTACTTACAGCAGCAGCTTCGCTCATCCAGCCTTACCCTCAATCGCCTGAAAAAACAGTTTGAGCCG
CCTCAGATGAAGCAGGTTCCTGAACTCAGCATTGAACAACATTCACTTGAACTTTATGACAAACTCCTC
CCCTCCTGCTCAGTCCCCGCTGAGCCCCTACCAGCACCTGAGCCTCTATTTAAAAAAGCTCAGGCTCT
CCCACATGTTGACCCATTGGGAATCTATCGAATCCCAAGCCATGCAGGAAAACTGGTCTTATGCGGAA
TTCTTACTAGCCTTGTGCGAAACGGAGGCCCAACGAAGAGAACAAGCTCGTCTAAAACGTGCCCTCAC
CGAAGCCAGGCTCCCAAACGCAAAAAGTTTTACCAACTTTGACTTTAGCCATTGTCCCCAGCTCAATCC
AGCTCCCTTGATGCAATTAGCCGCAGATCCGGGTTGGTTGGAGCGCGCCGAGAATTGCCTTATTCTGG
GGCCCTCGGGTGTTGGAAAAACACATCTGGCCACTGGGGTGTCCAAAAAGATGCTGGAATTCGGTAAG
CGGGTGAAGTTCTTTGCAGCCAACGCATTGGTCCAGCAACTGCAACAGGCGAAGCTCCAACTGCAGCT
GCATCCAATGCTCAAAAAACTGGACCGCTATGATCTGTTGATCTTGGATGACTTGGGCTATTGCAAAAA

Constructing the whole genome in order to identify species:

- Alignment
- Assembling

Problems :

- ✓ Huge datasets
- ✓ Expensive process
- ✓ Long time execution
- ✓ Most of the existing applications are just suitable for a single machine
- ✓ Very complicated Assembling and Alignment process
- ✓ Too many errors and missing parts
- ✓ Too many same copies of the same genomes in the sample



DNA Signature:

TAGGCCCACCTGTTTGTG
CAGGCCCTCAGATTAAAT
GTCGAGTAGACAGGGTAT
GATACTTAGAGCGAACCC
GAACCCTTTCTAACGCGC
ACCCTTTCTAACGCGCGA
TCTAACGCGCGAGTCAGA
TGGTATGGTCCAACGCTC
TAAAGAGTCCTCCCCCAG
CACAGCGGGGTACACATT
TGGTCTCGCTAATTCGTA
GGTCTAGCGGCATAACAT
CCCACACGCTGTATTTAC
CAGCCACACCCTACTATT
GCCAGTAAGTGTGAGGGT
GTCATTAAGCGAGATTCG
GATATATATGGGGGTCAC
TGTGTCTCTGAGCCCGCC
GTGGGCGGGAATAGCACT
CGGCTCCGTTGTATATAA
GGGTGTACCACCCCGCTA
TGTACCACCCCGCTAGGT
TACCACCCCGCTAGGTGA
CAGCTATAAGGTCCGGAC
TTCGGGGTTAGTGGAGAG
CGGAGAGGGGTAGCATAC
AGGTTTCTAGGTCACCCC
CTAGGTCACCCCATAGTA
AAGAAGGTCCTAAGGATC
GTCCTAAGGATCGTGTTC
AATAGACGGTAACCCGAG

DNA signature is a short nucleotide sequence which is used to distinguish one species apart from all other species.

Number of signatures for every species can differ from 1 to several millions.

rid	Short Reads	B1			B2								B3				B4								.....					B3000								
		S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	.	.	.	.	S
1	R																																					
2	R																																					
3	R																																					
4	R																																					
5	R																																					
6	R																																					
7	R																																					
8	R																																					
9	R																																					
10	R																																					
11	R																																					
12	.																																					
13	.																																					
.	.																																					
.	.																																					
.	.																																					
.	.																																					
4,000,000	R																																					

S600,000,000

4,000,000 **Short reads**

200,000 **The average number of signatures for every species**

3,000 **Number of species with identified signatures**

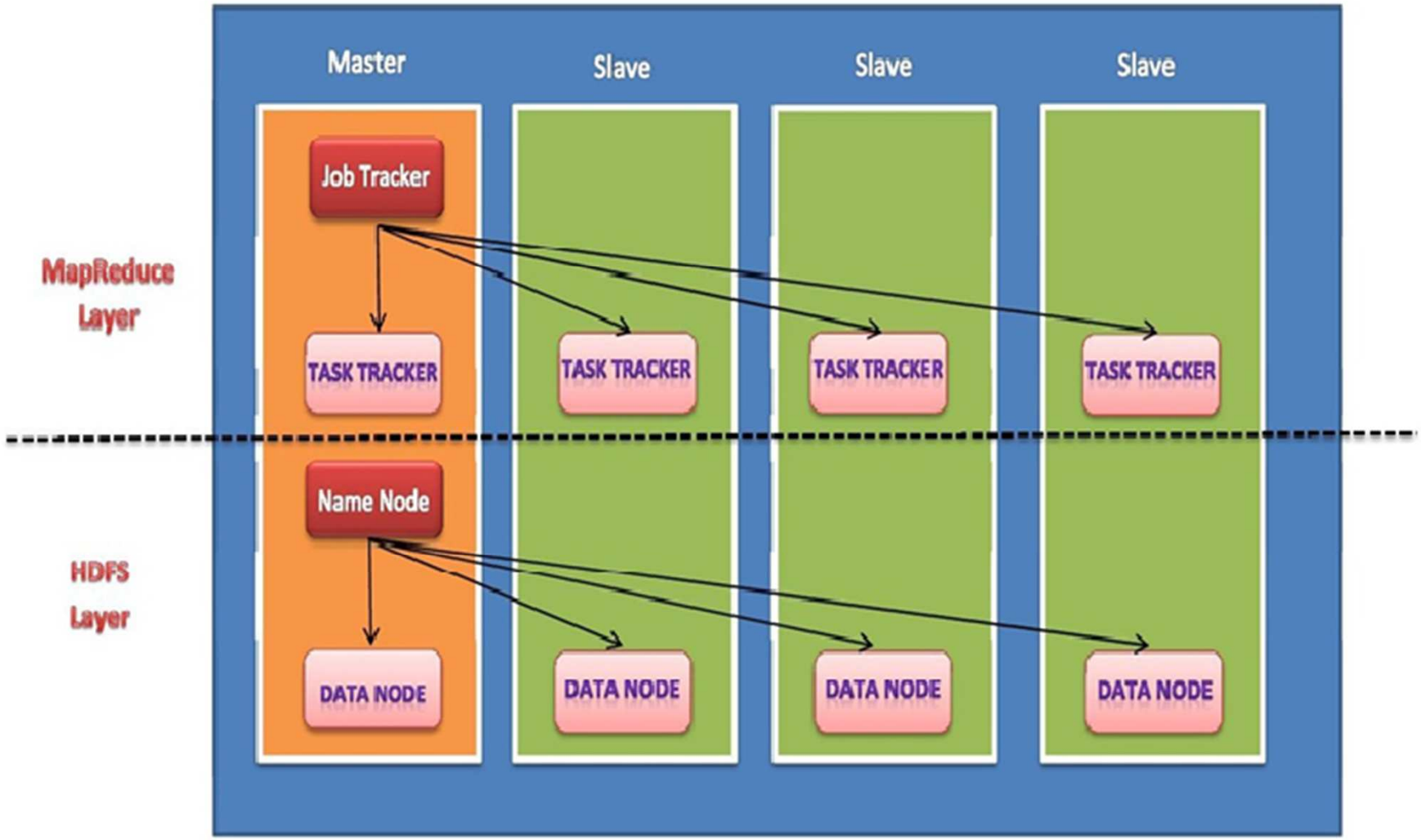
Problems of using signature to identify species in the Metagenome samples:

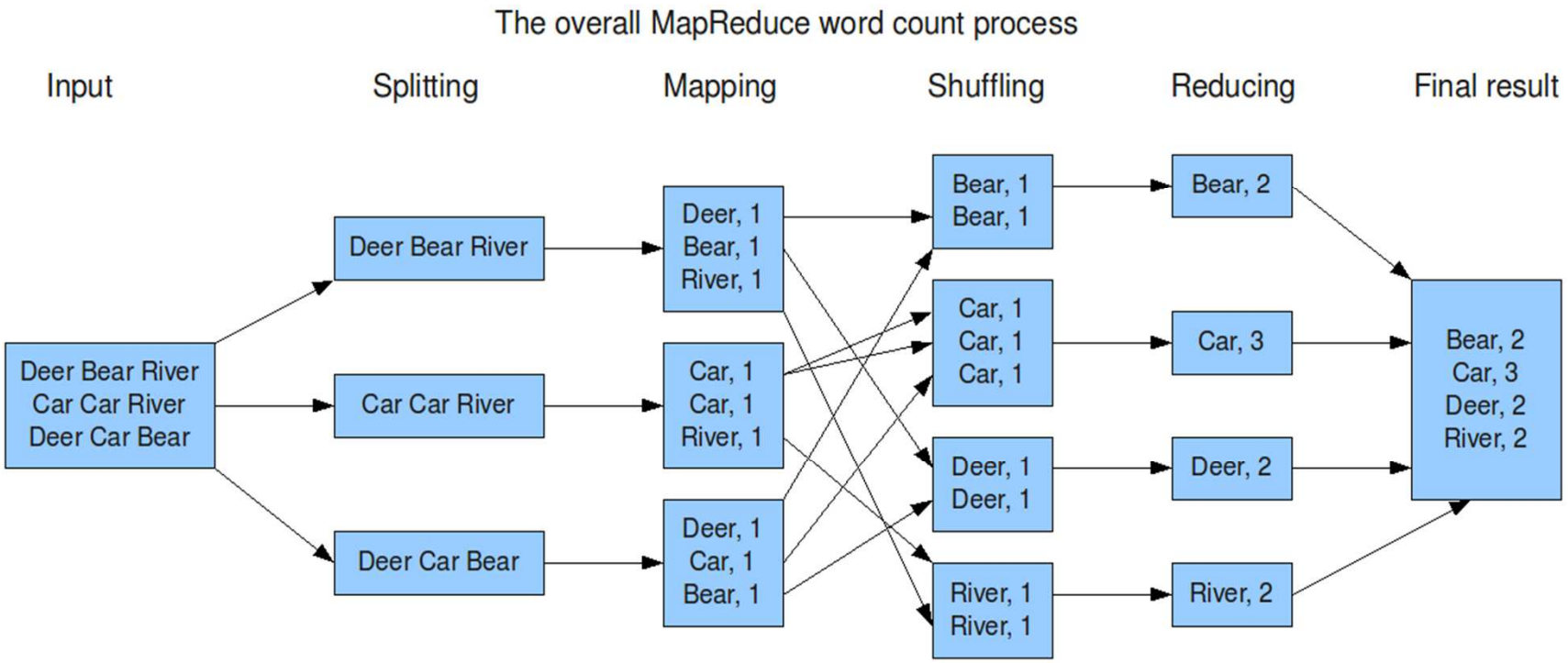
Large number of signatures
Large number of short reads
Large number of organisms

= Very big Datasets or Very big Tables

Using ordinary hardware and software is impossible or it takes a long time to implement from days to several months regardless of any failing during the process.

Hadoop as a parallel and distributed computing framework and Bitmap Indexing technique are suitable solutions to solve this problem.


Master
Slave
Slave
Slave
Job Tracker
MapReduce Layer
TASK TRACKER
TASK TRACKER
TASK TRACKER
TASK TRACKER
Name Node
HDFS Layer
DATA NODE
DATA NODE
DATA NODE
DATA NODE


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

index tables

RID	Reads	b1						
		s1	s2	s3	s4	s5	s6	s7
1	R1	0	0	0	0	0	0	0
2	R2	0	0	0	0	0	0	1
3	R3	0	0	0	0	0	0	0
4	R4	0	0	0	0	0	0	0
5	R5	0	0	0	0	0	0	0
6	R6	0	0	0	0	0	0	0
7	R7	0	0	0	1	0	0	0
8	R8	0	0	0	0	0	0	0
9	R9	0	0	0	0	0	0	0
10	R10	0	0	1	0	0	0	0

RID	Reads	b1	b2	b3	b4	b5
1	R1	0	0	0	1	0
2	R2	1	0	0	0	0
3	R3	0	0	0	1	0
4	R4	0	0	0	0	0
5	R5	0	0	1	0	0
6	R6	0	0	0	0	1
7	R7	1	0	0	0	0
8	R8	0	0	0	0	0
9	R9	0	0	0	0	0
10	R10	1	0	0	0	0

4,000,000*200000*1000 =
800,000,000,000,000 byte= 727.6 TB

4,000,000*1000=4,000,000,000 byte
= 3.7 GB

Results:

Real-time identification of species in the metagenome sample

Query optimization

Speed up the Queries

Using ordinary or normal hardware and software

Thank you for your attention