

# Map Reduce for Motif Search in Biological Networks

Rohan Kuber and Sathvik Laxminarayan

Computer and Information Sciences and Engineering (CISE)



## ABSTRACT

We present a map reduce approach for motif search in biological networks that would help identify gene regulatory expressions working on mutated genome data.

### Cancer Tumor Genomics – Personalized Therapy

"...10 years from now, each cancer patient is going to want to get a genomic analysis of their cancer and will expect a customized therapy based on that information."

- Director, The Cancer Genome Atlas (TCGA), Time Magazine, 6/13/11

## WAR ON CANCER

1000 genomes =  
5,994,000 processes =  
23,076,000 hours =  
2737 years

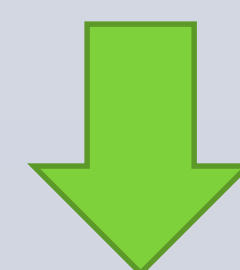
1000genomes.org/

- This approach would be one of the few experiments that would establish that large distributed data clusters can be employed for large scale gene network analysis
- Known pathway knowledge will be used in the identification of genes and sub-networks that are related to the disease which can help researchers in disease prognosis.

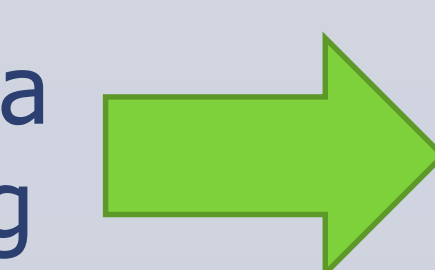
### Challenge Problem

we have sequences that are 600 nucleotide long with random and uniformly distributed bases. In each of these sequences, we plant a motif of length 15 with 4 random mutations (15, 4) in each of the sequence

Search for motif identification (150x)



Mutated genome data dynamic and evolving



## Data Used

- Motifs will be planted in randomly generated biological sequences to test whether algorithm is giving output with acceptable accuracy.
- Real life data sequences acquired from protein databases such as RCSB Protein Data Bank

### Methodology - SP STAR algorithm

Step 1: Consider every possible subsequence of length 'l' and find its hamming distance from every other l-mer of every sequence. If the distance between them is at most 2\*d, then store it in candidate array.

Find consensus of all such candidates.

Step 2: Now iteratively use this consensus string and find hamming distance of each l-mer.

against the consensus and find new consensus out of new candidates.

Continue doing this until you no longer get different consensus. The final consensus is our Motif.

Complexity:  $O((n-1+1)*m)^2$

- Obtain the results for an implementation on a single node
- Implement the scaled map-reduce approach over Spark

### Divide Data

```
JOPARMOWZDQYOXYTJBBHAWDYDCPRJBXPHOHPKQYUHRQZAMVAFYRARKSVKHTQDIHER
SIGBHZJZUJXMMYSNARAWEKGBSJOLLSQSGHMCPELSTFLBGSFNPUCZSRUPCHYNVZ
HCPQXNPNQJNBKCPDMOKALXAPEMVQLZSVXZKUTAPWPGZPDPYZKZCBBCJPDYJLIBITLLP
LXELDRKEXDTQVPTTEYHTLQLBBBVQVWZXSQXNJOMUYJNJUWRSYXWQYYXCSZPOKLWJD
RLTBSCIEKWLBMJAHTPUTIEBPPBMEGYRHLTHXWGJPRVSHZVAVQYEEQZNBHYCWMFYA
QQIANNYHQOUIZVEAAHACBBBBBVGZJWOBNAQFTONEQLNBTDDBMIVAZXZMCOJDHISW
KAXIAGTUTAFRYHJMJIIMDSPIZHWZIZCYUCEGLMBGYGFVTQVZQATXLGIBYBTUMILOCL
TGCOZGTZIUFUZIMXEKWNYSBICRFXDIOGCSEOBKAPBSZYSPDAAOXRAOJOTJUFSTURT
JHUDPETSGRZUIGUSAFSYVYRQPIZTEUOXGTHFXSMXJRWPSURSOWTHHILGDHZPZAEWKE
KFGRCYVZSAHPSZOBRIKXWKPICLQLJTKVNBIMPDKYECEPAYTFAVEOPSZRNVNEHRJSTZ
FLVYAHORKTKMGYCOUIERHLGDYWDFNZQLXVKHDJYALHVVFAANGODNLXLEXAMEYKVI
XDAQVSSDWWOEDKLOE
```

### Finding candidates for consensus

```
gactaTcTTCcTTCCaatattacctcatgctatt
tactgaagTTatTTTgCCCacgagaggctaaaaac
cgctcgactcttaacctacagTTTTaTTaCctCgtg
```

```
TcTTCcTTCCCa
gTTaCTTTgCCC
TTTTaTTaCctC
```

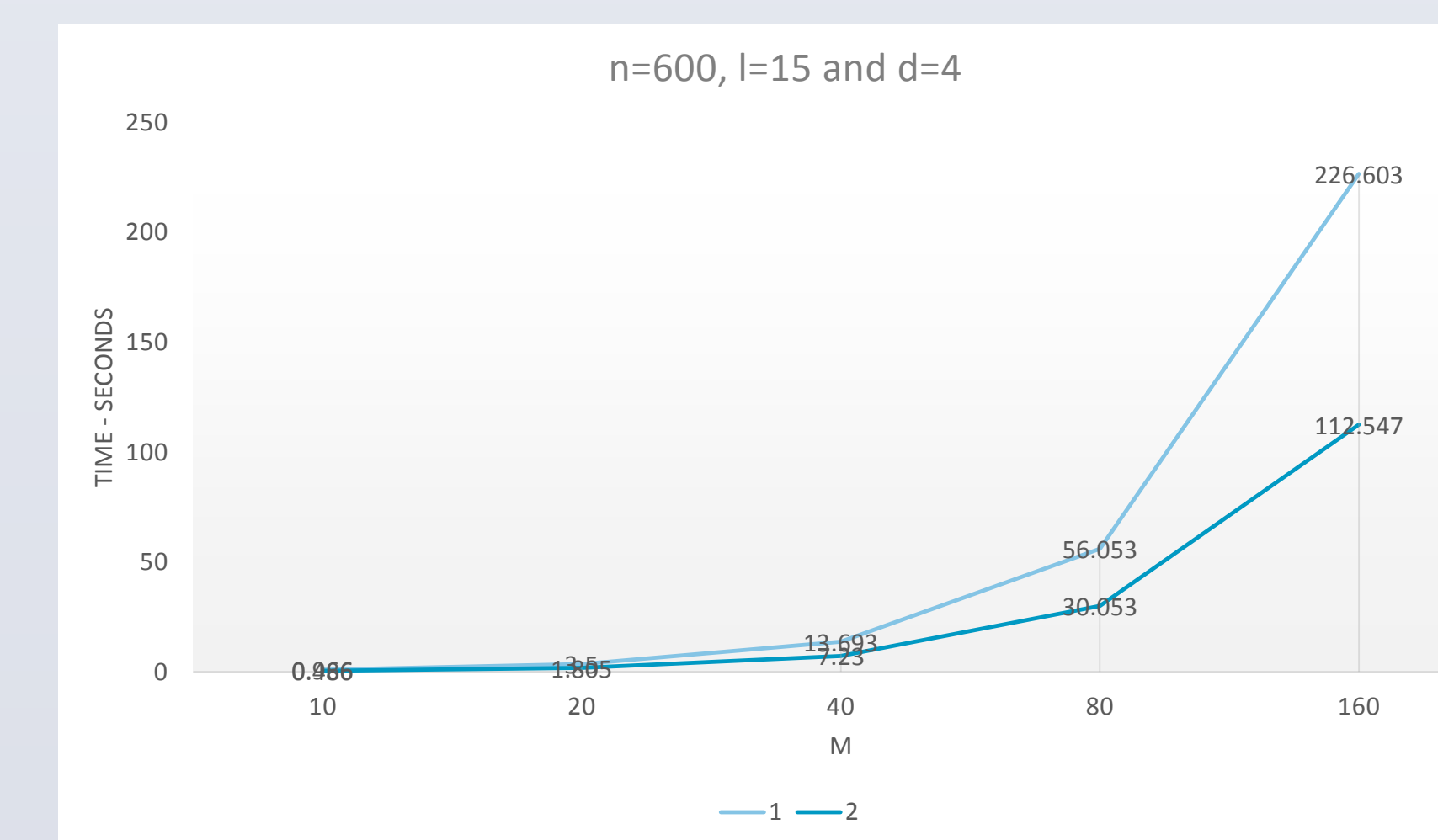
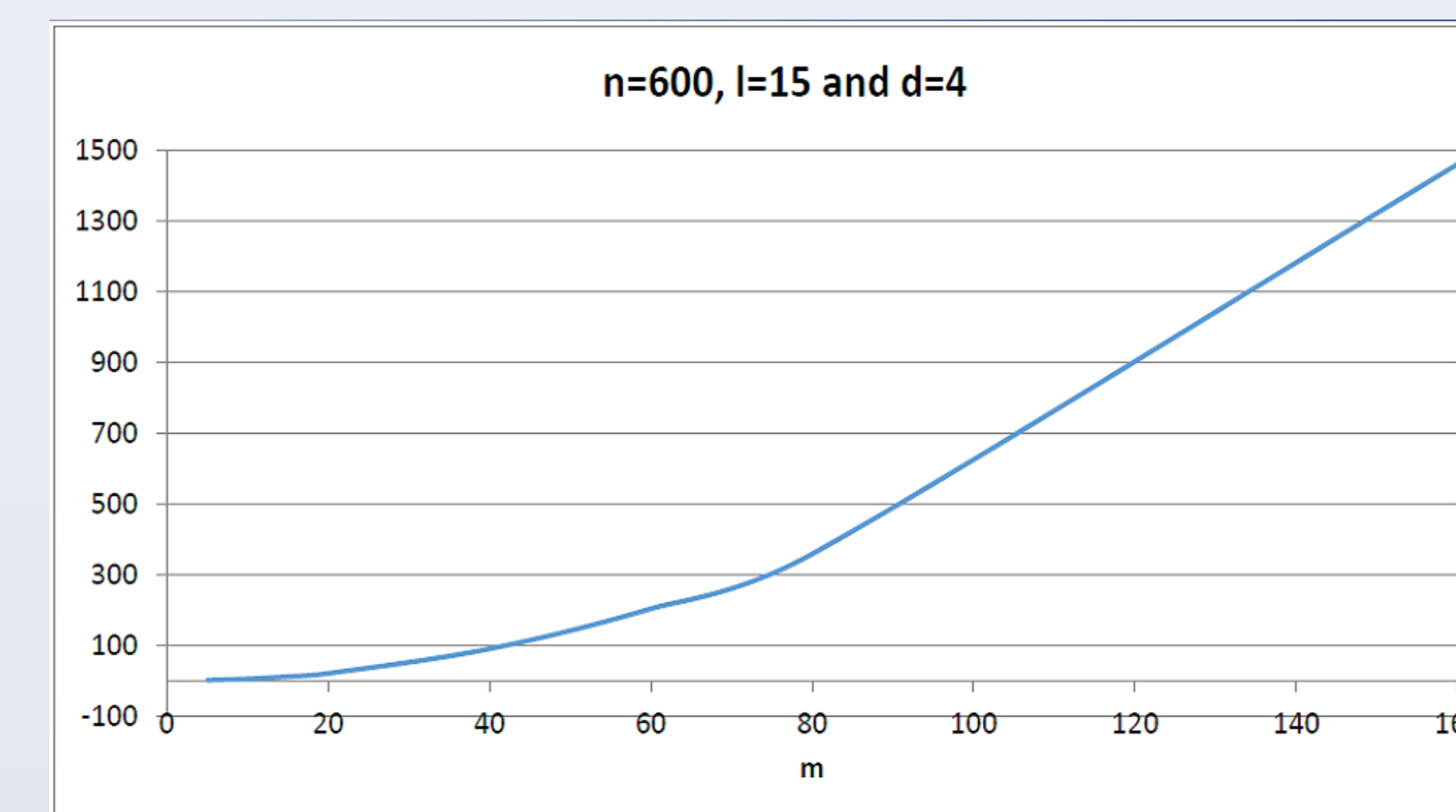
```
TTTTCTTTCCCC
```

← Consensus signal

## Why the heuristic of 2\*d ?

```
TTCCCTTAAACGAAAA
TTTTCCGGAAAAAAA
TGTTTTGAAAAcata
TTTGATAAAATAAAC
-----
TTTTTTAAAAAAA
```

### Results



- Fast, MapReduce like engine
- In-memory storage (RDD) for very fast iterative queries
- Up to 100x faster than Hadoop
- Let's users manipulate distributed collections ("Resilient Distributed Datasets" – RDDs ) with parallel operations
- Has language integrated APIs in Scala Python and Java shells



## Conclusions

We established that map-reduce approach can be successfully employed to perform the identification of motifs in gene regulatory networks which are dynamic in Why the heuristic of 2\*d nature subject constant mutations. It also provides the advantages of

- No data collision
- No dependency between parallel threads

## Future Work

- The modified algorithm works well on the synthetic sequence however on protein databank sequences it cannot match motif's as effectively due to the massive data scale and diverse mutations.
- A distributed platform built over shark and spark may provide even faster results.

## References

- P. Pevzner and S.-H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Eighth International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 269-278.
- A survey of DNA motif finding algorithms Modan K Das and Ho-Kwok Dai
- MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat
- <http://spark-project.org/>

## Acknowledgements

We would like to thank our professor Dr. Daisy Zhe Wang and our TA Abhiram Jagarlapudi for their support and guidance throughout this project

## CONTACT

Daisy Zhe Wang  
University of Florida  
[daisyw@cise.ufl.edu](mailto:daisyw@cise.ufl.edu)

Sathvik Laxminarayan  
University of Florida  
[sathvikl@ufl.edu](mailto:sathvikl@ufl.edu)

Rohan Kuber  
University of Florida  
[kuber@cise.ufl.edu](mailto:kuber@cise.ufl.edu)