Institiúid Teicneolaíochta Chorcaí
Cork Institute of Technology

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

Development of a Multiple Sequence Alignment Algorithm using Cloud Computing and Big Data Technologies

Jurate Daugelaite[1], Aisling O' Driscoll[2] and Roy D. Sleator[1]*

[1]*Department of Biological Sciences, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland.
[2]Department of Computing, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland.

Jurate.Daugelaite@cit.ie

Introduction

Multiple Sequence Alignment (MSA) of DNA, RNA and protein sequences is one of the most essential techniques in the fields of molecular biology, computational biology and bioinformatics [1]. Next-generation sequencing technologies are changing the biology landscape by flooding the databases with massive amounts of raw sequence data [2]. Combining MSA algorithms with distributed and parallelised computing solutions is therefore necessary in order to improve the speed, quality and capability for MSA algorithms. The **storage and analysis** of the growing genomic data represents the central challenge in computational biology today [3].

Multiple Sequence Alignment

MSA is a widely used computational procedure for biological sequence analysis. Sequences are compared in order to:

- Construct Phylogenetic trees
- Analyse secondary and tertiary protein structures
- Analyse protein functions

Finding mathematically perfect MSA can generally be defined as a complex optimization problem or **NP-complete problem,** therefore heuristic ("best guess") methods are used instead [4].

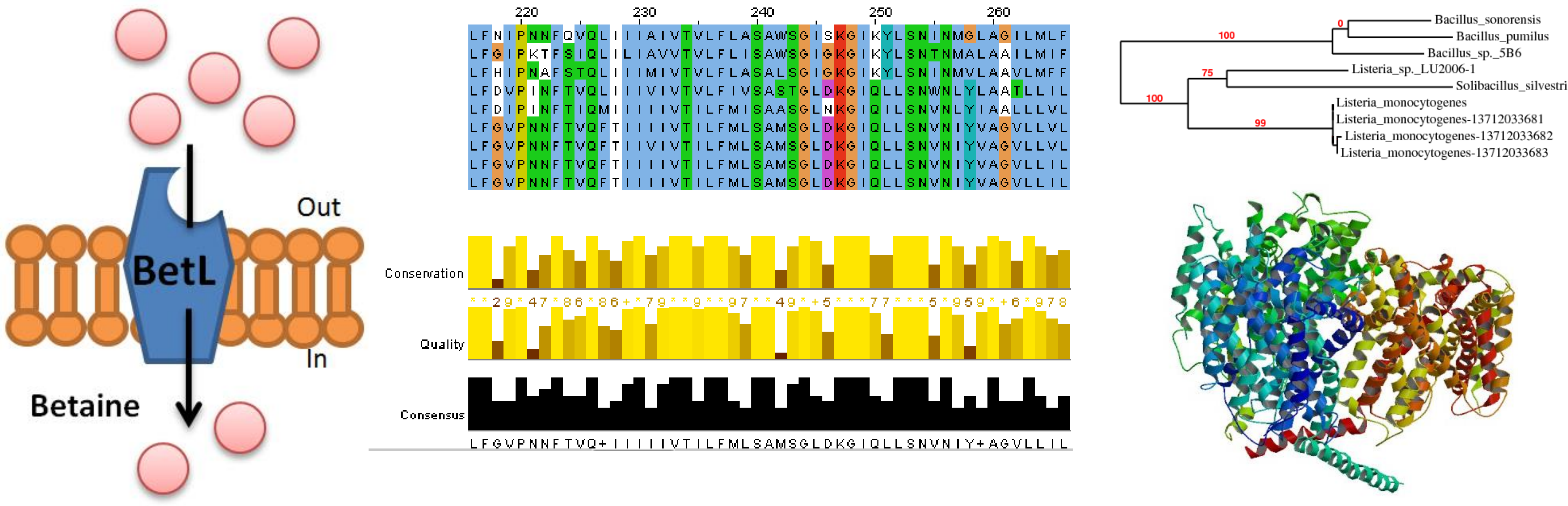

Fig 1: An example of Multiple Sequence Alignment of BetL protein (betaine transporter of *Listeria monocytogenes*) generated by Clustal Omega algorithm.

Big Data Technology

Hadoop is a software framework , consisting of MapReduce and HDFS. It is driven by big data, distributes the data over **commodity hardware** and provides **parallelised processing and analytics**. MapReduce is a software framework used for processing and analysing large amounts of data across distributed commodity servers.

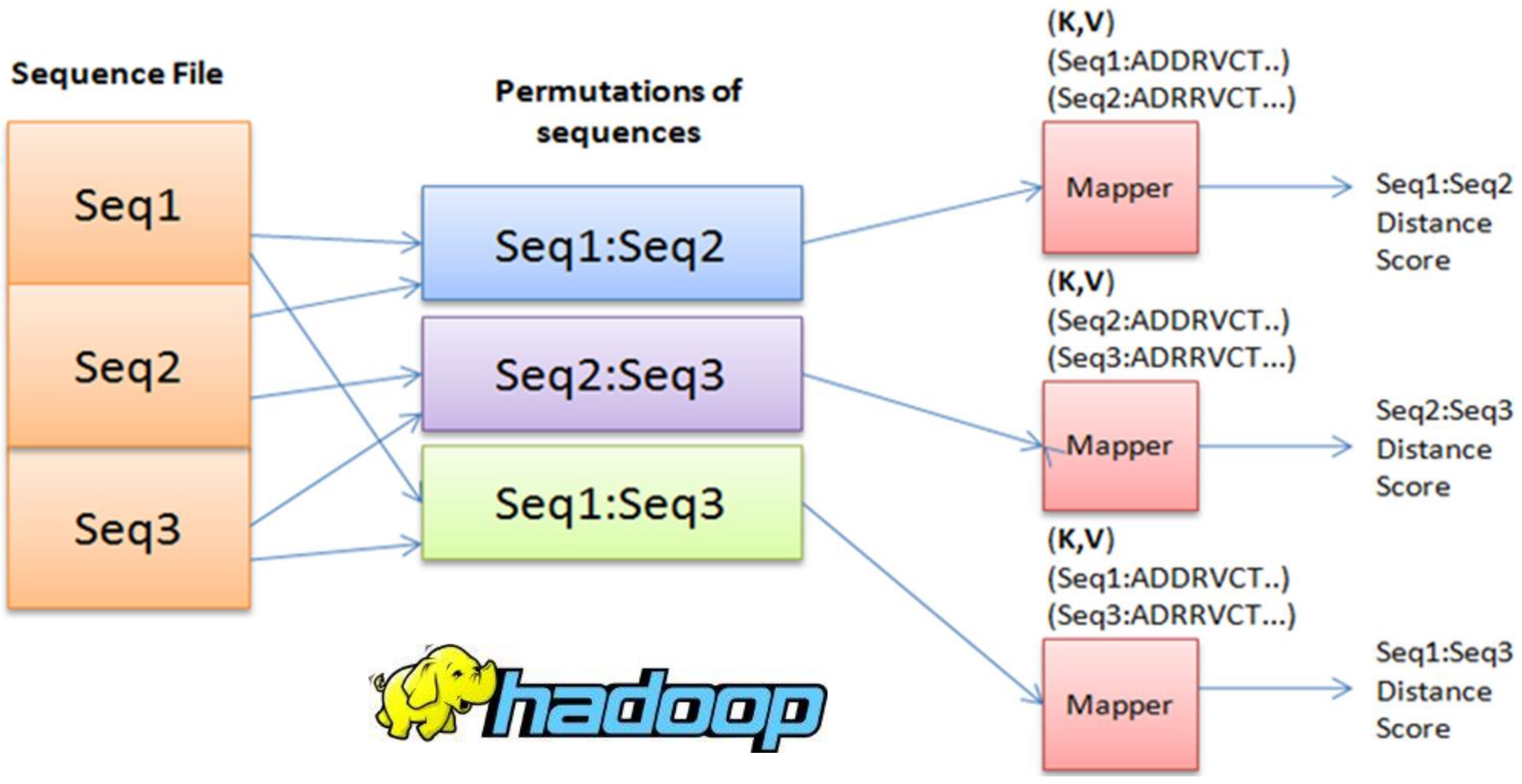

Fig 2:Diagram of k-tuple pairwise alignment method used by Clustal Omega as a MapReduce job, showing tree steps.

Cloud Computing

The National Institute of Standards and Technology (NIST) describes cloud computing as *"a **pay-per-use model** of enabling **available, convenient** and on-demand **network access** to a shared pool of **configurable computing resources** that can be rapidly provisioned and released with **minimal management effort** or service provider interaction"* [5].

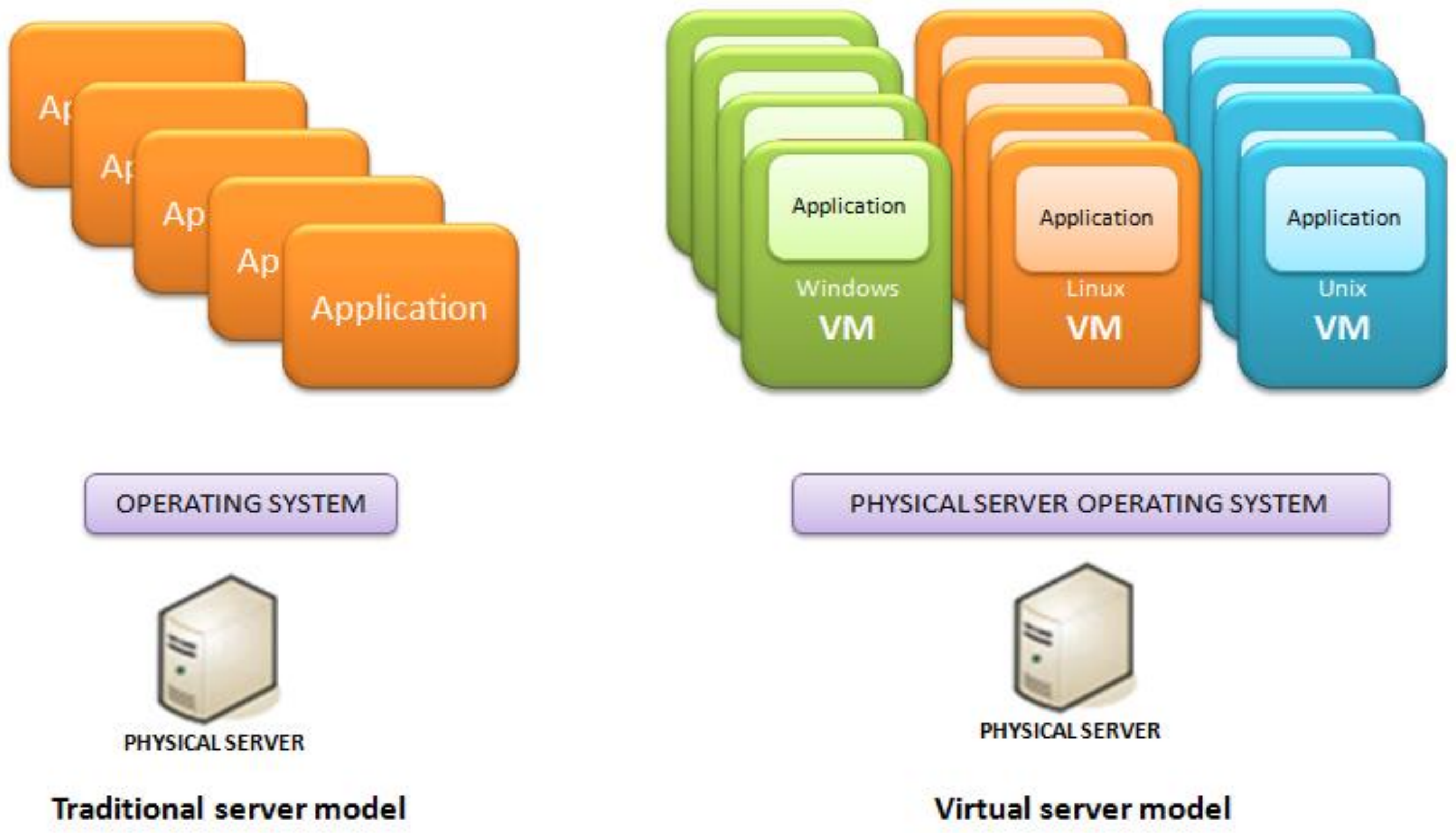

Fig 3:Transition from traditional computing to virtualised computing where multiple OS images share the hardware resources.

Conclusions

- The data from sequencing projects is increasing at exponential rates.
- There is no biologically perfect solution for MSA.
- Raw biological data storage and processing is at a bottleneck.
- Cloud computing and the big data technologies have the potential to aid in solving these problems, by offering distributed storage and faster processing times.

Acknowledgments

Jurate Daugelaite is funded under the Embark Initiative by the Irish Research Council (IRC). Aisling O'Driscoll and Dr. Roy D. Sleator are Principal Investigators on ClouDx-i an FP7 –PEOPLE-2012-IAPP project.

References

1. Kemena, C. and C. Notredame, *Upcoming challenges for multiple sequence alignment methods in the high-throughput era.* Bioinformatics, 2009. **25**(19): p. 2455-65.
2. Edgar, R.C. and S. Batzoglou, *Multiple sequence alignment.* Curr Opin Struct Biol, 2006. **16**(3): p. 368-73.
3. Dai, L., et al., *Bioinformatics clouds for big data manipulation.* Biology Direct, 2012. **7**(1): p. 43.
4. Katoh, K. and H. Toh, *Recent developments in the MAFFT multiple sequence alignment program.* Brief Bioinform, 2008. **9**(4): p. 286-98.
5. *A Definition of The Cloud at Last? - Web Performance Watch*. Available from: http://blogs.keynote.com/the_watch

EUROPEAN REGIONAL DEVELOPMENT FUND
SEVENTH FRAMEWORK PROGRAMME
MARIE CURIE ACTIONS
ClouDx-i