DNA SEQUENCE ANALYSIS USING DISTRIBUTED COMPUTING WITH SMITH-WATERMAN ALGORITHM

Shanu Verma¹, Balwant Ram², Bikramjit kaur³

¹, ³Student of M.tech Computer Science, Department of CSE, Lovely Professional University, Phagwara, Punjab, India
²Assistant Professor, Department of CSE, Lovely Professional University, Phagwara, Punjab, India.

ABSTRACT

Mostly everything is done on the internet in today’s world. There are number of jobs which have to be performed on the internet. Some of them can be easily executed but some of them are high computing jobs. They need high power computation to perform the task like jobs in image processing, bioinformatics, finance extra. Another way of computing high computing jobs are supercomputers but very expensive and out of reach from generality. Therefore performance of high computing jobs are very difficult problem in these days. To compute these jobs, there is another method which I am going to propose, that is, if we split a job into small scale fragments and perform these all fragments concurrently in a network on different nodes then this problem will be solved. So, by using web services with distributed computing, I am going to examine execution of high computing jobs and maximum job size needed for accurate results. SOAP is used by web services for interacting over the network. I am going to use Sequence alignment as an example of high computing job.

Keywords: Distributed Computing, Web Services, SOAP, WSDL, DNA Analysis, Sequence Alignment.

1. INTRODUCTION

For high computing jobs, parallel computing is needed for reducing the computational time. If we divide high computing job into small chunks and compute all that chunks separately and simultaneously then computational time will get reduced. Distributed computing is one of special type of parallel computing.
1.1 Distributed computing
Distributed computing is a computing in which large job split into small scale fragments and perform these all fragments concurrently in a network on different nodes concurrently to achieve a single goal [1].

1.2 Architectural model
There are layers in this model that are: services, middleware and platform. Services are one or more servers processes provide distributed service by interacting with one another where middleware products and standards are widely used. They include CORBA, Java RMI, Web services, DCOM. Platform is a combination of operating system and computer network layer.

1.3 System architecture
The division of responsibilities is done by the system architectures. There are two types of architectures, namely, Client-server and Peer-to-peer. In client-server, one server may be clients of another server. In peer-to-peer, all nodes play same role, can be client or server.

1.4 Challenges faced by distributed computing in early decades
Distributed computing faced the problem of how programming functionality encapsulated within “objects” is exchanged in early decades. Over the past few decades, many techniques are used to interchange programming functionality, including RPCs architecture such that CORBA and COM. Platform neutral technique is specified by XML. Through XML and web technologies for accessing computing functionality, that ability is known as web services.

1.5 Limitations of CORBA and DCOM
DCOM is a Microsoft’s-only architecture. CORBA is too complex and semantically ambiguous to provide cross-platform interoperability without a large number of manual integration works, in reality. RPC do specters of marshalling (serializing) the executable code and shipping it over the internet then unsuspected problems of the security concerns and compatibility start arising. Both CORBA and DCOM use binary wire protocols, which are not humanly reliable and have difficulties to move through firewalls.

1.6 Web Services
Web Services solves above problems by HTTP with SOAP [2]. Web services are set of standards which allow two computers to interact with each other and exchange data over the web. HTTP is supported by all browsers and servers.

1.7 Sequence alignment
Sequence alignment is using the technique of modifying two compared sequences so that patterns can match approximately with each other as closely as possible by inserting gaps [4].

2. RELATED WORK
Guan Qing and Guan Jianhe in 2012 focused on parallel computing on LAN which analyzes the problems of electromagnetic field exploration application with OO technique using parallel computer architecture and Message passing interface. To implement the message passing, Visual C++.Net+MPICH2 which forms a parallel programming platform.
Zhang Feng et.al, explained in 2009 service oriented architecture that is agent based for manufacturing enterprise collaborations. Services are platform-independent computational elements
that can be described, published, discovered using XML for the purpose of developing distributed interoperable applications.

CORBA allows interoperability between various components by providing their interfaces in meta-language [3]. But according to above challenges web services are much better than CORBA. Traditional distributed applications have importance as their own but the problem is to connect these services together, to work for high computing. This problem can be resolved by web services {[5], [6]}. To implement distributed computing, service-oriented architecture gives a framework. Web services are becoming the base of distributed computing. If an application calls service asynchronously then it doesn’t need to wait for result. It can do another operation. When it need result to do execution, then it can be stopped and fetch that result. This mechanism is important for number of applications in distributed execution environment. By solving the issue of interrelationship between the call and result, the module is capable of to execute asynchronous call of web services. Working as a proxy for calling application, module is also capable to execute the call dynamically {[7], [8]}. By using distributed computing with web services sequence alignment is going to be analyzed. There are two algorithms for sequence alignment, which are: Needleman-Wunsch algorithm and Smith-Waterman algorithm. Parallel strategy designed to explore the computational characteristics of the Needleman-Wunsch algorithm that are used for biological sequence comparisons. Needleman-Wunsch algorithm is used for global sequence alignment where Smith-Waterman algorithm is used for local sequence alignment [9].

3. OBJECTIVES

Processing of huge amount of data or computation of high computing jobs is a hot issue in these days. So I have following objectives: To analyze the factors affecting the performance of distributed computing using web services and to find out the optimal job size that is given by master node to the slave in distributed computing environment.

4. RESEARCH METHODOLOGY

Research methodology of the proposed work for the hypothesis is presented here. Problem of computation of high computing job can be resolved by diving this job into small parts and then process them in parallel. I am going to use distributed computing for analyze the performance of high computing job.

![Fig 4.1: Allocation of parts of job to the slaves by master in distributed computing](image)
Web services are included as middleware.

**4.1 Web services**

Web services provide a standard interoperability among various software applications which are running on different platforms [10]. Web services are using XML language. XML stands for eXtensible Markup Language. XML is a case sensitive language. XML must have root element.

**4.2 SOAP**

SOAP stands for Simple Object Access Protocol. SOAP is used for accessing web services and it is XML based protocol. It is a format used for sending messages over the network. It is language independent. By SOAP, applications can communicate over HTTP.

**Format of SOAP:** A SOAP message is an simple XML document containing the following components:

- An Envelope finds out the XML file.
- A Header element has header details.
- A Body element that contains detail about invocation and response.
- A Fault element has detail about errors.

A SOAP message must use XML.

**Syntax of SOAP Message:** Syntax of SOAP message is given below:

```xml
<?xml version="1.0"?><soap:Envelope
xmlns:soap="http://www.w3.org/2001/12/soap-envelope"
```

**4.3 DNA analysis**

DNA analysis is the process of analyzing the DNA sequence. Here sequence alignment is going to be done. Alignment can be done in two ways, Local alignment and Global alignment [11], [12]. In proposed work Local alignment is going to be used by using smith-waterman algorithm.

**4.4 Smith-Waterman Algorithm**

This algorithm performs local alignment on two sequences. This algorithm is applied when both the sequences are dissimilar and expected to have some similarity [13]. This algorithm is example of the dynamic programming and the aim of Smith-Waterman algorithm is to find the best alignment over all the other sequence alignments.
Three steps are there in this algorithm which is following:

- Initialization
- Scoring
- Trace back

In scoring scheme, Match Score = +1, Mismatch Score = -1, Gap penalty = -1 and Substitution Matrix are given:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Initialization: In this step do following instructions:

- Create a matrix with X +1 Rows and Y +1 Columns
- The 1st row and the 1st column of the score matrix are filled with zeros.

Scoring: The score of any cell C (i, j) is the maximum of:

- scorediag = C(i-1, j-1) + S(I, j)
- scoreup = C(i-1, j) + g
- scoreleft = C(i, j-1) + g

where S(I, j) is the substitution score for letters i and j, and g is the gap penalty [13].

Trace back: alignment is done in this step.

- Trace back starts from the last cell, i.e. position X, Y in the matrix
- Gives alignment in reverse order
- There can be three moves: diagonally, up, or left
- Trace back start from the cell having maximum value.
- The only possible predecessor is the diagonal match/mismatch neighbor. If more than one possible predecessor exists, any can be chosen [13]. This gives us a current alignment of both sequences.

5. CONCLUSION AND FUTURE WORK

The time consumption and memory consumption can be reduced by distributed computing. Therefore, purposed work is to examine execution of high computing jobs using web services and maximum job size needed for accurate results which will pass over the network. By applying web services, interaction will get easy over all the nodes in the network because of HTTP and SOAP. Hardware and expenditure will also decrease. There are number of fields where we have to compute high computing jobs, there we can use distributed computing with web services. In this work, sequence alignment is use as high computing job. So this is beneficial for future scope.
6. REFERENCES


