Designing and Building a Framework for DNA Sequence Alignment Using Grid Computing

EL-Sayed Orabi
IS Dept. MTI University
Cairo, Egypt.

Mohamed A. Assal
Dean of IT Center. MTI Univ.
Cairo, Egypt.

Mustafa Abdel Azim
Dean of CS Faculty. AASTMT
Cairo, Egypt.

Yasser Kamal
CS Dept. AASTMT
Cairo, Egypt.

***Abstract*—Deoxyribonucleic acid (DNA) is a molecule that encodes unique genetic instructions used in the development and functioning of all known living organisms and many viruses. This Genetic information is encoded as a sequence of nucleotides (adenine, cytosine, guanine, and thymine) recorded using the letters A, C, G, and T.DNA querying or alignment of these sequences required dynamic programming tools and very complex matrices and some heuristic methods like FASTA and BLAST that use massive force of processing and highly time consuming. We present a parallel solution to reduce the processing time. Smith waterman algorithm, Needleman-Wunsch, some weighting matrices and a grid of computers are used to find field of similarity between these sequences in large DNA datasets. This grid consists of master computer and unlimited number of agents. The master computer is the user interface for insert the queried sequence and coordinates the processing between the grid agents.**

***Keywords*—DNA fingerprint; Smith waterman algorithm; Needleman-Wunsch; Grid computing; Coordinator and Agent computers**

I. INTRODUCTION

DNA sequences are a string of characters (A, C, G and T) representing the genetic information of a living organisms (humans, animals, birds, bacteria, planets, etc.) and many known viruses. Every living culture has its own unique nucleotide code. Based on this fact the government agency all over the world use the DNA sequence to identify persons (criminals, army and police solider, terrorism, etc.) and can also be used to determine a child's paternity (genetic father) or a person's ancestry. The way of differentiates these sequences is called sequence alignment. Sequence alignment is also used to identify the breed (homologies) of unknown protein or nucleotide sequences. This can be solved by using dynamic programming in time proportional to the product of the length of the two sequences being compared, as in [1].

Sequence alignment is a tool used to compare the sequence to find a similarity between them based on complex algorithms and matrixes as in [2]. The Smith Waterman and Needleman Wunsch are used for local and global sequence alignment.

To solve the delay time of the comparison scientists all over the world proposed different models and techniques including hardware improving (sequencer machines)to reduce the length of the sequence using microarray as in [3].

II. DNA FINGERPRINT

DNA fingerprinting is a test to identify and evaluate the genetic information in a person's cells. It is called a "fingerprint" because it is very unlikely that any two people would have exactly the same DNA information, in the same way that it is very unlikely that any two people would have exactly the same physical fingerprint.

The test is used to determine whether a family relationship exists between two people, to identify organisms causing a disease, and to solve crimes. [4] With different DNA datasets contains millions of records, it may be impossible to find the target sequence (person) at the right time. So the need to get information fast from the large databases was raised rabidly. The parallel computing is a solution to reduce the time of querying and retrieving information from these databases by distributes the processing over numbers of devices, as in [2].

III. SMITH WATERMAN ALGORITHM

It performs a local alignment over two sequences. It is an example of dynamic programming. This algorithm is useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context or sequences with the same length. Initialization, Scoring and Trace back (Alignment) are three steps to find the best alignment over the conserved domain of two sequences.

The complexity of this algorithm is $O(N*M)$ where N is the length of queried sequence and M is the length of target sequence. Smith waterman is a pair wise sequence alignment on other word it is 1 to 1 alignment. Example of local alignment 2 sequence N and M with match= 4, mismatch = -1 and gap = -2 is illustrated in the following example. The First step is the initialization of matrix N*M as illustrates in table 1.

TABLE I. FILLING THE MATRIX WITH THE GIVEN SCORES

Initiates the matrix with gap		Sequence M						
			G	A	T	T	G	A
		0	0	0	0	0	0	0
Sequence N	A	0	0	4	2	0	0	4
	C	0	0	2	3	1	0	2
	G	0	4	2	1	2	5	3
	C	0	2	3	1	0	3	4

The second step is the trace back (local alignment) as shown in table 2. The trace back starts with the maximum value in the matrix then goes left or up or diagonal according to values next to the start point.

TABLE II. THE TRACING BACK (LOCAL ALIGNMENT)

Trace back from the Max. Value		Sequence M						
			G	A	T	T	G	A
		0	0	0	0	0	0	0
Sequence N	A	0	0	4	2	0	0	4
	C	0	0	2	3	1	0	2
	G	0	4	2	1	2	5	3
	C	0	2	3	1	0	3	4

The last step is the alignment of the two sequences as illustrates in figure 1.

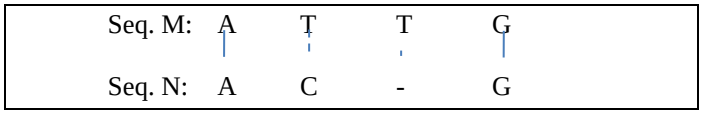

Fig. 1. The Final Local alignment of the given sequences using smith waterman algorithm

IV. NEEDLEMAN WUNSCH

The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was published in 1970 by Saul B. Needleman and Christian D. Wunsch. It uses dynamic programming, and was the first application of dynamic programming to biological sequence comparison. It is sometimes referred to as the optimal matching algorithm. Initialization, Scoring and Trace back (Alignment) are three steps to find the best alignment over the entire length two sequences. The following example shows a simple alignment between two sequences N and M with match score = 4, mismatch = -1 and gap = -2. The First step is the initialization of matrix N*M as illustrates in table 3.

TABLE III. THE SCORING MATRIX OF NEEDLEMAN WUNSCH ALGORITHM

Initiates the matrix with gap		Sequence M						
			G	A	T	T	G	A
		0	-1	-2	-3	-4	-5	-6
Sequence N	A	-1	-1	3	-3	-4	-5	-1
	C	-2	-2	2	2	1	0	-1
	G	-3	2	1	1	1	5	4
	C	-4	1	1	0	0	4	3

The second step is the trace back (Global alignment) as shown in table 4. The trace back starts with the last right point in the matrix then goes left or up or diagonal according to values next to the start point.

TABLE IV. THE TRACING BACK (GLOBAL ALIGNMENT)

Trace back from the last point		Sequence M						
			G	A	T	T	G	A
		0	-1	-2	-3	-4	-5	-6
Sequence N	A	-1	-1	3	-3	-4	-5	-1
	C	-2	-2	2	2	1	0	-1
	G	-3	2	1	1	1	5	4
	C	-4	1	1	0	0	4	3

The last step is the alignment of the two sequences as illustrates in figure 2.

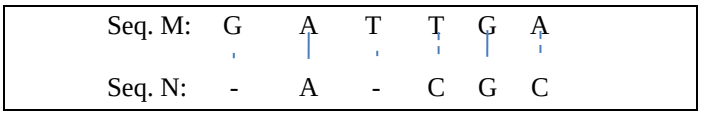

Fig. 2. The Final Global alignment of the given sequences using Needleman Wunsch algorithm

V. GRID COMPUTING

Grid computing is basically a paradigm that aims to enable access to high performance distributed resources in a simple and standard way. A grid is defined as a type of parallel and distributed autonomous resources dynamically at runtime depending on their availability, capability and performance as in [5]. The aim of grid computing is to enable coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organization as in [6][7][8].Grid computing is known as a distributed system that connects many computer systems having different hardware and platforms (operating system). It allows applications to run in parallel on multiple machines, clusters, or systems (Virtual Organization). The system is suitable for solving the problems that require a large amount of computation as well as storage capacity. In our research, the Grid system is used for solving the delay time of the global & local alignment. Figure 3 illustrated the proposed grid components.

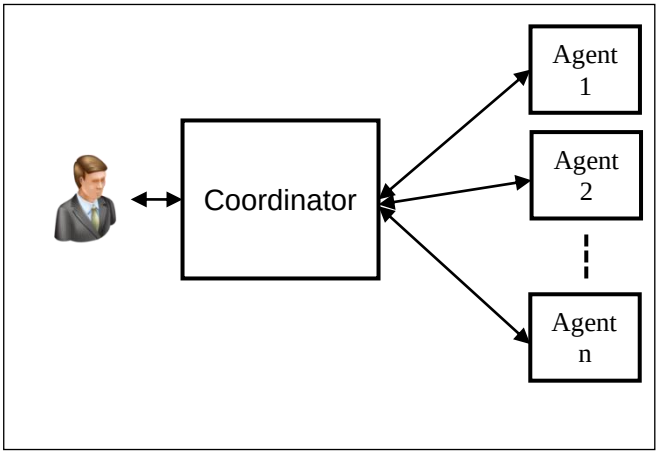

Fig. 3. The grid components

a) The coordinator

The coordinator sends the tasks to agents so agents are fully utilized as possible. The coordinator considers each physical core of each agent as a separate execution unit as in [8]. The user inters the queried sequence and XML file contains the dataset that will be searched through the coordinator application. Then the coordinator calculates the total number of sequences in the dataset counts the connected agents and divides the task equally for each agent. Each task consists of two sequences the first is the queried sequence and the second is a sequence from the dataset file. Example, dataset contains 1024 sequences and the grid consists of 8 agents so each agent receives 128 tasks. Figure 4 shows the components of the coordinator computer.

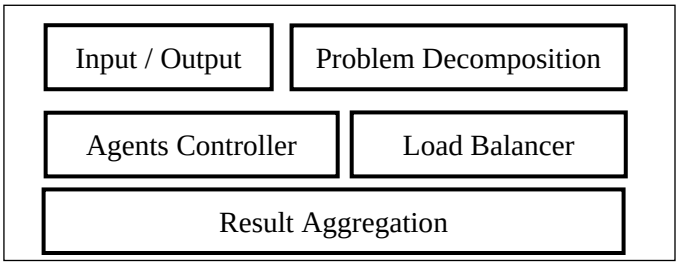

Fig. 4. Coordinator components

b) Agent

The agent registers itself with the coordinator and waits to receive grid tasks from coordinator as in [8]. The agent receives the sequence alignment tasks from the coordinator and executes them using Smith waterman algorithm. An agent is configured to be dedicated which mean that agent resources are centrally managed by the coordinator. Then each agent sends results of alignment back to the coordinator which selects the most similar sequence to the queried sequence. Figure 5 shows the main function of the agent computer.

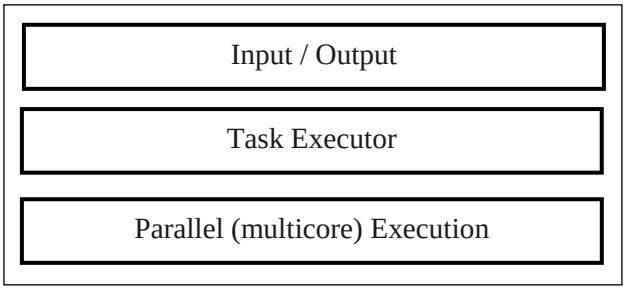

Fig. 5. The Agent functions

V. Related Work

In many published paper the researchers all over the world proposed a lot of models to make local alignment using parallel computing. The following section demonstrates some examples of those models.

- Propose parallel processing of optimal alignment between two sequences by exploiting parallel MPI/FORTRAN 90. The algorithm for optimal alignment is based on dynamic programming techniques. Two versions of algorithms have been developed: one versus one sequence alignment and one versus many sequence alignment. The second algorithm used "block" parallel dynamic programming algorithm and this technique will increase the amount of workloads done by each processor as in [9].

- DNA sequence alignment model under this hierarchical grid architecture. They used dynamic programming algorithm with linear space parallelism and is separated into two parts: parallelization of the similarity matrix and parallelization of the divide-and-conquer algorithm. Three clusters have been setup where each cluster has eight nodes. The clusters are connected by Ethernet switch where the bandwidth is about 8 MByte/s. Meanwhile the bandwidth between nodes in each cluster is about 190 Mbyte/s. The architecture of the software is based on two layers; upper layer uses MPICH-G2 and lower layer employs MPICH as a communication interface protocol as in [10].

- FASTA is a heuristic based technique in sequence similarity search. Parallelization of FASTA has been implemented in the Grid Application Development Software (GrADS) project as in [11]. The GrADS adapts the master-worker paradigm, scheduling and rescheduling the tasks on an appropriate set of resources, launching and monitoring the execution. The GrADSoft scheduler makes a static schedule for its application where the whole or a portion of sequence databases are replicated on some or all of the grid nodes. The master will inform each worker which portions of database should be loaded into memory. The master also sends the input query sequence to each worker and collects the results from the workers.

VI. The Implemented Framework

By using the previous grid topology and Smith Waterman, Needleman-Wunsch basic implementation the researcher proposes a DNA multiple sequences alignment framework. This framework is been based on pairwise comparison to query large databases by distribute one to one tasks to find the maximum complete or partial alignment over the grid computing. Those tasks are generated from user interface (coordinator). The coordinator counts the number of sequences of the database. Then calculates the number of connected agents and counts the number of available cores. Then the coordinator divides the tasks equally with load balance through the agent to execute the Smith waterman and Needleman-Wunsch algorithms. Then the agents start to execute the tasks one by one and send the results to the coordinator. If any failure is found in any agent, the coordinator reassigned the task for another available agent considering the load balance of each agent in the grid. At last the coordinator selects the sequence with maximum matching score. The second scenario is distributing the dataset as clustered sub datasets on each agent. Then the coordinator defines the cluster of the queried sequence with codon cluster technique. To increase the accuracy of the alignment to cover the analysis requirement, the researcher combined the original implementation of Smith waterman algorithm with some

weighting matrixes. The first will be blosum 62, the second is PAM 250 and the third is Gonnet160.Those weighting matrices are used in many DNA analysis applications. So the researcher thought it will be very helpful feature in the proposed model.

VII. Experiments and Result

The experiments were carried out in a computer laboratory contains 16 connected personal computer. One PC used as coordinator and the rest used as agents. The configuration of each PC is shown in table 5.

TABLE V. Demonstrates the Grid Nodes Configurations

#	The nodes configuration	
1	OS	Microsoft Windows 7 Enterprise Service Pack 1
2	Processor	Intel® Core™ i7-2600 CPU @ 3.40Ghz
3	Number	4 P / 4 V
4	Clock	3.701 Ghz
5	Memory	4 GB

The first iteration of the experiment was for test the performance of the grid through sending some tasks (sequences to be aligned) with different lengths. These tasks and the time to carry out them over number of nodes are illustrated in table 6 and figure 6.

TABLE VI. Illustrates First Iteration with Different Sequence Length

Seq. Length	No. Of Seq.	No. Of Cores (time per minute)				
		1	4	8	16	32
1000	128	0.3178	0.2347	0.1237	0.06727	0.0343
2000	128	1.2622	0.2409	0.1804	0.10385	0.0579
4000	128	5.0105	0.6445	0.6333	0.32808	0.1801
8000	128	9.6267	0.2641	0.2232	0.11123	0.0946

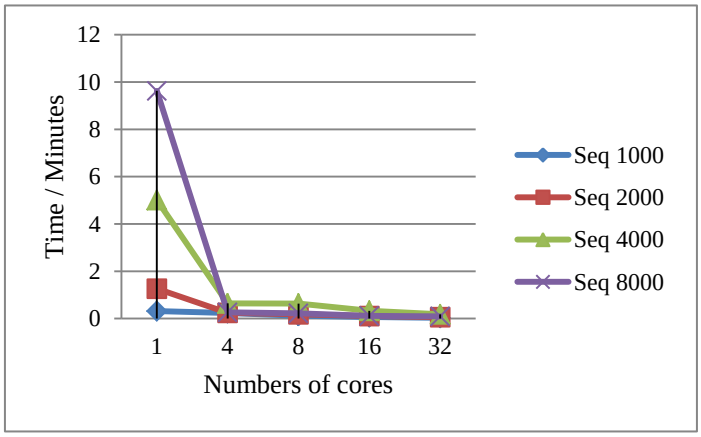

Fig. 6. Demonstrates the result of first iteration

After the first iteration it was found that the processing power of each agent using multi core processing is fully utilized when using sequence length larger than 2000 nucleotides. It was found that the sequence of length 4000 nucleotides takes more than double time of the sequence 2000 nucleotides length, so the second iteration is querying sequences with length 4000 nucleotides against datasets contain different number of sequences. The results of this iteration are demonstrated in table number 7 and figure number 7.

TABLE VII. Demonstrates the Result of Second Iteration (Same Sequence Length Against Different Datasets)

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	5.01052	0.6445	0.6333	0.3281	0.1801
4000	256	4.85454	1.1837	0.5745	0.3216	0.1718
4000	512	19.8005	3.2172	1.0123	0.5622	0.3454
4000	1024	39.6572	9.7471	2.0293	1.2582	0.6607
4000	2048	79.3692	19.494	9.7472	4.4924	2.4490

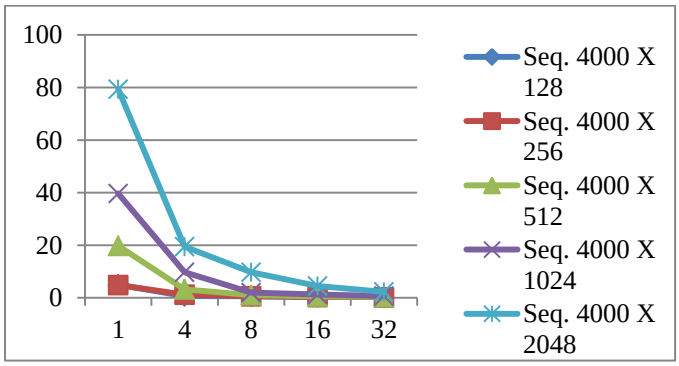

Fig. 7. Shows the relation between the sequences with the same length and different numbers of sequence to be aligned with

For more alignment details the researcher combined the original Smith Waterman & Needleman Wunsch algorithms with BLOSUM, PAM and Gonnet matrices. To calculate the effect of adding the blosum62 and PAM250 weight matrices to the original implementation the researcher run three more iterations. The results of those iterations are illustrated in tables and figures number 8, 9 and 10.

TABLE VIII. Demonstrates the Result of Blosum62 Weighting Matrix

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	7.3378	1.8514	0.9181	0.3148	0.24600
4000	256	14.676	3.6027	1.8361	0.6296	0.49200
4000	512	29.351	7.6054	3.6593	1.8726	0.94385
4000	1024	58.702	14.677	7.3186	3.5989	1.86914
4000	2048	117.97	29.255	14.627	7.0823	2.62731

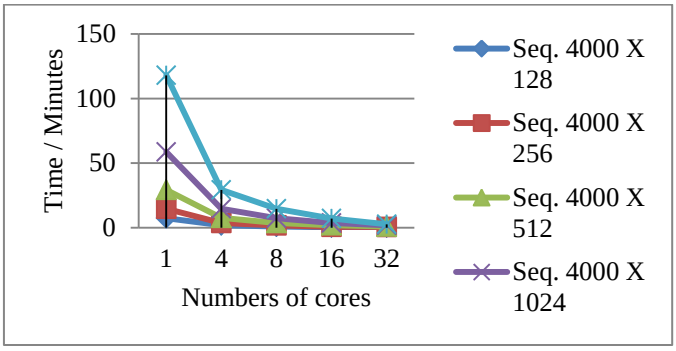

Fig. 8. Demonstrates the result of Blosum62 matrix

TABLE IX. THE RESULT OF PAM250 MATRIX ITERATION

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	7.51120	1.8779	0.9459	0.4748	0.2473
4000	256	15.0224	3.7557	1.8918	0.9497	0.4947
4000	512	30.0448	7.5112	3.6923	1.8517	0.9657
4000	1024	60.0896	15.022	7.5142	3.7202	1.8580
4000	2048	120.179	30.145	15.123	7.5231	3.7302

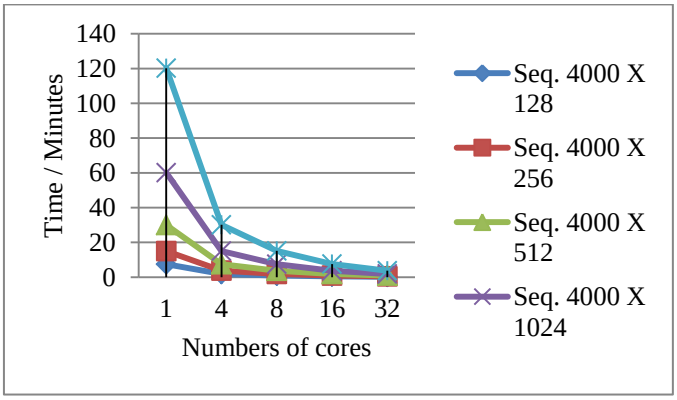

Fig. 9. Shows the result of PAM250 iteration

TABLE X. DEMONSTRATES THE RESULT OF GONNET160 WEIGHTING MATRIX

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	9.2305	3.1070	1.5535	0.7581	0.4259
4000	256	18.461	6.2142	3.1162	1.5161	0.8517
4000	512	36.922	12.428	6.2142	3.0321	1.7034
4000	1024	73.844	24.857	12.428	6.0642	3.4069
4000	2048	147.69	49.713	24.857	12.128	6.8138

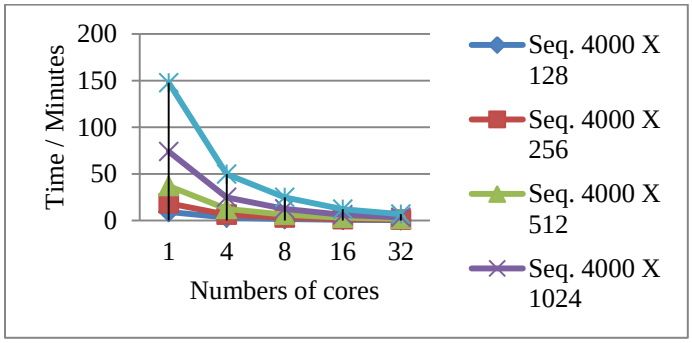

Fig. 10. Shows the result of Gonnet160 iteration

From previous experiments it is found that the time of processing decreased when the number of nodes (cores) increased. These tests make the complexity of Smith waterman and Needleman Wunsch algorithms for finding target sequence in multi sequence as O(N*M)S/C where S is number of sequences and C is number of cores connected to the grid.

VIII. CONCLUSION

The implementation of Smith waterman and Needleman Wunsch algorithms over a grid of computers decreases the time of processing. Grid computing makes querying large DNA datasets fast enough and with affordable cost. The use of grid and smith waterman algorithm to find a match between a sample and multi sequences is a negative relation as more cores in the grid the less computational time and vice versa. The complexity of the proposed model for Multi sequence alignment is O(N*M)S/C where N and M is the pairwise sequences, S is the number of sequences in the dataset file and C is number of cores in the grid agents.

Fig. 11. References

[1] "Algorithms in Bioinformatics",.I,WS06/07, C.Dieterich.

[2] "Multiple Sequence Alignment on the Grid Computing using Cache Technique, " International Journal of Computer Science and Telecommunications Volume 3, Issue 7, July 2012.

[3] "A Novel Algorithm for Fast Synthesis of DNA Probes on Microarrays" ACM Journal on Emerging Technologies in Computing Systems, Vol. 9, No. 1, February 2013.

[4] "DNA Testing in Criminal Justice: Background, Current Law, Grants, and Issues,"Congressional Research Service 7-5700.

[5] "Evolution of Cloud Computing and Enable Technologies," International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol.1, no.4, pp.182-198, October 2012.

[6] PerfCloud: Grid Services For Performance-Oriented Development of Cloud Computing Application," 18th IEEE International Workshop on Enabling Technologies: Infrastructures for Collaborative Enterprise (WETICE 09), pp. 201-206,2009.

[7] Ahmed Said Abo El-Ala, Mohamed Anwar Assal, and Mohamed Bakr," On the Design of a framework for Grid computing Developing System(GDS) ", In Managerial Research Journal, Consultancy Research & Development Centre, Sadat Academy for Management Sciences, July 2012.

[8] Ahmed Said Abo El-Ala, Mohamed Anwar Assal, and Mohamed Bakr," An Enhanced framework for Grid computing Developing System(EGDS) ", In Managerial Research Journal, Consultancy Research & Development Centre, Sadat Academy for Management Sciences, July 2012.

[9] Nguyen, E. N. D., Nguyen, D. N., Nguyen, D. T. and Tungkahotara,"Comparing DNA Sequences by Dynamic Programming in Sequential and Parallel Computer Environments", Proc. of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, 2006, 146 – 153.

[10] Chen, C. and Schmidt, B. "An Adaptive Grid Implementation of DNA Sequence Alignment", Future Generation Computer Systems,2005, 988 – 1003.

[11] YarKhan, A. and Dongarra, J. J. "Biological Sequence Alignment on the Computational Grid Using the GrADS Framework", Future Generation Computer Systems,2005, 980 – 986.