A Comprehensive Study by Using Different Alignment Algorithms to Demonstrate the Genetic Evolution of Heat Shock Factor 1 (HSF1) in Different Eukaryotic Organisms

Mohammed M. Saleh[1], Ahmed M. Alzohairy[2], Osama Abdo Mohamed[3,1], Gaber H. Alsayed[4]
[3] Computer and Information Technology Department, Khulais Faculty of Computer Science and Information Technology, King Abd El-Aziz University, Khulais, Saudi Arabia,
[1]Mathematics department, Faculty of science, Zagazig university, Egypt.
[2] Genetics department, Faculty of agriculture, Zagazig university, Egypt
[1] Mathematics department (computer science), Faculty of science, Zagazig university, Egypt
[4] Basic sciences department, Faculty of engineering sciences, Sinai university, Egypt

***Abstract*— The process of gene tracing of specific gene or of all genome in species is very important for biologist. Sometimes the aim of gene tracing process is to know the genetic evolution that occurs in a certain protein sequence in specific organisms and also to know the genetic evolution that occurs in a conserved domain in this protein sequence. This process is very difficult one for biologists, but using computer algorithms makes this problem easier.**
In this paper we introduce a comprehensive study by using different alignment algorithms through which we demonstrate the genetic evolution that occurs in heat shock factor protein 1 (HSF1) in some eukaryotic organisms such as human, Danio rerio, Taurus, mouse, plant (Arabidopsis) and yeast. In addition, this study will illustrate the molecular evolution in the conserved domains of HSF1 (HSF_DNA-bind) throughout the different eukaryotic organisms.

***Keywords*— Pairwise global alignment, multiple sequence alignment, pairwise local alignment, gene tracer algorithm, heat shock factor protein 1 (HSF1), phylogenetic tree.**

I. INTRODUCTION

The focus of much recent biological evolution research has been on the detection of similarities and dissimilarities among living species. As the study of nature continues, human knowledge of variations among species has grown gradually both in all directions. This knowledge is used to give names to species we know. Identifying, naming, and organizing species into groups is a science called Taxonomy. Up till now, millions of species have been identified [1], demonstrating the current vast knowledge about species.

Scientists go further to think about the cause and evolution of the observed similarities and dissimilarities between different organisms. Many biologists including Charles Darwin's and many others have suggested that all living organisms share a common ancestor .Meanwhile, the differences among species are partly caused by mutations accumulated over generations during the course of evolution . Thus, Phylogenetic is a science that study of the evolutionary relatedness among species. Researchers have established the links among seemingly different life forms, from bacteria, to animals and plants using [2].

Pairwise global alignment algorithms are intended for comparing two sequences that are entirely similar. A dynamic programming (DP) algorithm called Needleman and Wunsch [3] was proposed for pairwise global alignment. Those methods are very useful in analysis of DNA and protein sequences. There are other dynamic programming algorithms for pairwise global alignment such as Huang and Chao [4] and NGILA [5]. There are another algorithms for making pairwise local alignment such as the algorithm introduced by Smith and Michael Waterman [6]. Another algorithm is used to make multiple sequence alignment MSA, it introduced by Thompson et al [7]. The algorithm which introduced by Thompson et al to make multiple sequence alignment depend on the progressive alignment. This works by constructing a succession of pairwise alignment . Initially, two sequences are chosen and aligned by standard pairwise alignment; this alignment is fixed. Then, a third sequence is chosen and aligned to the first alignment ,and this process is iterated until all sequences have been aligned.

Progressive alignment is heuristic: it does not separate the process of scoring an alignment from the optimization algorithm. It does not directly optimize any global scoring function of alignment correctness. The advantage of

progressive alignment is that it is fast and efficient, and in many cases the resulting alignments are reasonable. General results are illustrated in conclusion.

Sudden increases in temperature promote Heat shock (HS) response in diverse organisms (e.g yeast, plants and animals). HS is characterized by elevated synthesis of a set of proteins called heat shock proteins (hsps) which provide resistance against a subsequent usual lethal dose of heat stress. The enhanced expression of hsps is regulated by heat shock transcription factors (HSFs).

This study outlines our current knowledge of the functions of regulation of Heat shock factors HSF1, and offers a comparative view of it structure in yeast, plants and animals using different alignments algorithms. HSF1 structure observations indicate that HSF1 in different organisms overlaps.

II. PAIRWISE GLOBAL ALIGNMENT (NEEDLEMAN AND CHRISTIAN WUNSCH ALGORITHM)

In our work we use Needleman and Christian Wunsch algorithm to make pairwise global alignment. This algorithm can be summarized as follows:

1. First we consider any two strings such as

$$A = a_1 a_2 a_3 \ldots a_n,$$
$$B = b_1 b_2 b_3 \ldots b_m$$

2. Scoring matrix of size (n+1)*(m+1) is constructed and initialized using a substitution matrix, e.g., PAM (Percent Accepted Mutations) [8], or BLOSUM (Blocks Substitution Matrix) [9].

3. S (a_i , b_j) = score of aligning a_i with b_j
(=+1if $a_i = b_j$, $\mu \leq 0$ if $a_i \neq b_j$, for example)
S (a_i ,)= S (, b_j) = $\delta \leqslant 0$ (for indels). (1)
μ and – δ refer to the score of mismatch and indel respectively.
Indel means insertions or deletions and it can be represented by using the symbol" ".

4. The best alignment is that produces the largest score for $S_{i,j}$:

$$S_{i,j} = \max \begin{Bmatrix} S_{i-1,j-1} + s(a_i, b_j) \\ S_{i-1,j} - \delta \\ S_{i,j-1} - \delta \end{Bmatrix} \quad (2)$$

5. The score for elements in the first row and column of the alignment matrix are given by

$$S_{i,0} = -i\delta, \qquad S_{0,j} = -j\delta \quad (3)$$

6. The score for the best global alignment of A with B is S (A, B) = $S_{n,m}$, and it corresponds to the highest-scoring path through the matrix and ending at element (n, m). It is determined by tracing back element by element along the path that yielded the maximum score into each matrix element. This algorithm has time complexity *O(nm)*.

This algorithm can be briefly summarized as in the following pseudocode:

```
Global alignment
    Input sequences A, B
        Set S_{i,0} ← −δi  for all i
        Set S_{0,j} ← −δj  for all j
        For i=1 to n
            J=1 to m
                S_{i,j} ← max{S_{i−1,j} − δ, S_{i−1,j−1} + s(a_i, b_j), S_{i,j−1} − δ}
            end
        end
```

III. PAIRWISE LOCAL ALIGNMENT (SMITH AND WATERMAN ALGORITHM)

Sometimes we need to look for local regions of similarity that suggest shared structural of functional subunits. In this case we need to make pairwise local alignment.

A local alignment of strings *s* and *t* is an alignment of a substring of *s* with a substring of *t*. A dynamic programing algorithm (DP) called Smith and Waterman [6] was proposed for pairwise local alignment. The equation used for calculation of the scoring matrix is:

$$sim(i,j) = \max \begin{Bmatrix} sim(i-1, j-i) \pm 1 \\ sim(i-1, j) - p \\ sim(i, j-1) - p \\ 0 \end{Bmatrix} \quad (4)$$

Where *sim (i, j)* is an *(m+1)* x *(n+1)* matrix the sign for the first case is chosen as follow:

We choose +1 in the last term if *s(i)=t(i);* otherwise we choose -1. *p* is the gap penalty. The time complexity of this algorithm is *O (nm)*.

IV. RESULTS AND DISCUSSIONS

A. First, we have made a code by using MATLAB program to make pairwise global alignment among the common conserved domains (HSF_DNA-bind) that is a part of the heat shock factor protein1 sequence (HSF1) in all sequences of eukaryotic organisms and also we made pairwise global alignment among the entire sequences of HSF1 protein .

In our work we use the following scoring matrices:-

1- Blosum50
2- Blosum30

The results were as follows:-

1) In case of the conserved domains (HSF_DNA-bind) in HSF1 protein sequences

Table 1: This table shows the score of the pairwise global alignment and also the percentage of similarity in the case of the pairwise global alignment among the conserved domains in HSF1 protein sequence of all previous organisms.

Conserved domains(HSF_DNA-bind)	Score by using Blosum50	Identities	Score by using Blosum30	Identities
Human & Danio rerio	210.667	89%	135.2	89%
Human &Yeast	114.333	46%	74	46%
Human & Plant	91.6667	51%	60.8	51%
Human & Mouse	244.667	100%	156.8	100%
Human & Taurus	240.333	98%	153.6	98%
Danio rerio & Mouse	210.667	89%	135.2	89%
Danio rerio & Taurus	206.333	87%	132	87%
Danio rerio & Plant	92	51%	61.2	51%
Danio rerio & Yeast	103.667	46%	66.6	46%
Mouse & Taurus	240.333	98%	153.6	98%
Mouse & Yeast	114.333	46%	74	46%
Mouse & Plant	91.667	51%	60.8	51%
Taurus & Yeast	116	47%	75.2	47%
Taurus & Plant	88.6667	50%	58.6	50%
Yeast & Plant	71.6667	43%	51	43%

The results in table (1) show that the highest similarity is between human and mouse with percentage 100%, but the least similarity is between yeast and plant (Arabidopsis) with percentage 43%.

2) In case of the entire sequence of HSF1 protein

Table 2: This table shows the score of the pairwise global alignment and also the percentage of similarity in case of the pairwise global alignment among the entire of HSF1 protein sequences of all previous species.

HSF1 protein sequence	Score by using Blosum50	Identities	Score by using Blosum30	Identities
Human & Danio rerio	544.333	56%	345.8	55%
Human &Yeast	440.333	22%	239.4	21%
Human & Plant	-8	27%	23.8	24%
Human & Mouse	649.667	70%	406.6	70%
Human & Taurus	1017.33	89%	624.6	89%
Danio rerio & Mouse	415.333	52%	269.2	51%
Danio rerio & Taurus	565.333	56%	355.8	56%
Danio rerio & Plant	-7.6666	26%	23.8	25%
Danio rerio & Yeast	-397.667	22%	-212.4	20%
Mouse & Taurus	670	70%	418.6	70%
Mouse & Yeast	-547	22%	-306.6	23%
Mouse & Plant	34.3333	27%	52.8	25%
Taurus & Yeast	-427	23%	-233.2	23%
Taurus & Plant	-5	27%	24.6	24%
Yeast & Plant	-450.333	24%	-237	23%

The results in table (2) show that the highest similarity is between human and Taurus with percentage 89%, but the least similarity is between human and yeast, Danio rerio and yeast, and mouse and yeast with percentage 22%.

B. Second we have made a code by using MATLAB program to create multiple sequence alignment (MSA) among all the conserved domains (HSF_DNA-bind) in HSF1 protein sequences of all previous organisms. We also have made MSA among all the entire sequences of HSF1 protein sequence of those organisms. The results are illustrated by using phylogenetic trees.

In our work we use the following scoring matrices:-

1- Blosum60
2- Blosum80
3- PAM10

1) Comparative phylogenetic study of the conserved domains (HSF_DNA-bind) in HSF1 protein sequences:-

An Eukaryotic organisms contain highly complex multigene families encoding Heat Shock Factors (HSFs) which binds specifically to heat shock promoter elements (HSE), which are palindromic sequences rich with repetitive purine and pyrimidine motifs [10], [11] and [12]. Promoters of eukaryotic heat stress (hs)-inducible genes share common HSF recognition elements (HSEs) with the palindromic consensus sequence (AGAAn) (nTTCT) located within a few hundred base pairs of the 5' flanking regions of heat shock genes [13], [14], [15], [16] and [17]. Deletion analyses of plant sHSP promoters initially pointed out that the cis-elements necessary for the heat stress response (HSEs) were also required for developmental regulation [17] and [18].

By blasting the DNA binding domain of some eukaryotic HSFs1 with different cellular organism, similar conserved domain was found among many of them discovered to date (Fig. 1). These results suggest that similar function of HSFs could be found in different organisms. That is explaining that swapping HSFs between different families can do the same effect as shown in [19].

Fig.1: Graphical summary of the conserved domains of HSF1 in the different organisms human, Danio rerio, Taurus, mouse, yeast, and plant.

Comparative study of HSF1 sequences using different matrices.

a) Matrix blosum60

A danderogram is constructed by using an HSF1 domain (HSF_DNA-bind) of human, mouse, Taurus, Danio rerio, plant (Arabidopsis), and yeast using the scoring matrix blosum60. Based on this specific matrix the Human and Mouse is the closet in the structure sequences of HSF1 domain (HSF_DNA-bind). It is also shown that the plant (Arabidopsis) and the yeast HSF1 domain (HSF_DNA-bind) are closer to each other than to the other organisms.

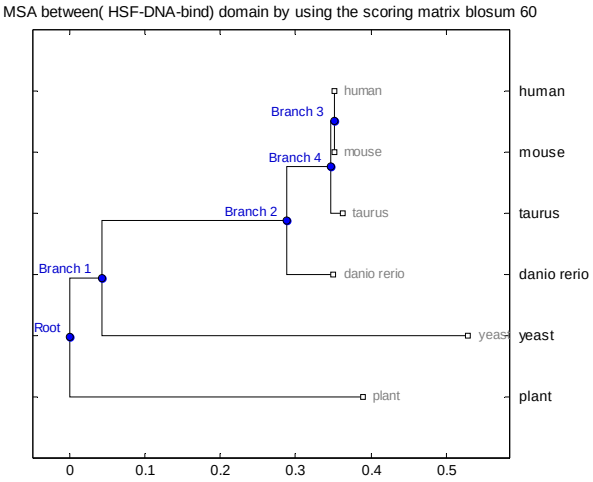

Fig. 2:.This figure shows the similarity between the conserved domains in HSF1 protein sequences in all the last species in case of MSA by using the scoring matrix blosum60.

b) Matrix blosum80

When we use the scoring matrix blosum80 the same results are obtained as when using blosum60.

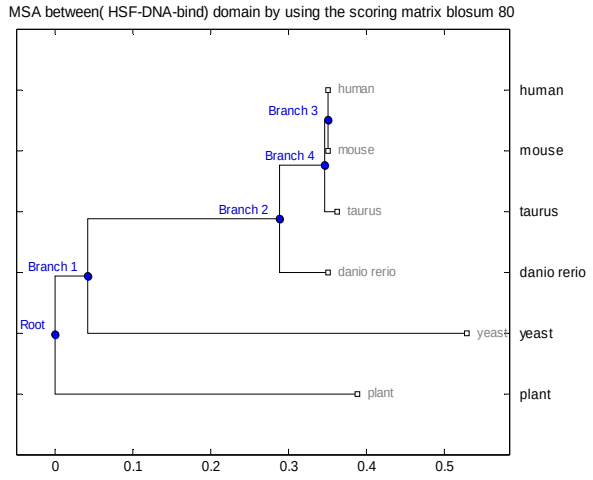

Fig. 3: This figure shows the similarity between the conserved domains in HSF1 protein sequences in all the last species in case of MSA by using the scoring matrix blosum80.

c) Matrix PAM10

When we use the scoring matrix PAM10 the same results obtained as when using blosum60 and blosum80.

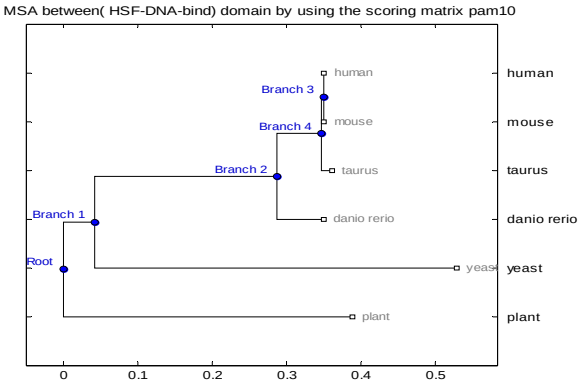

Fig. 4: This figure shows the similarity between the conserved domains in HSF1 protein sequences in all the last species in case of MSA by using the scoring matrix PAM10.

2) Comparative phylogenetic study in case of the entire length of HSF1 protein sequences are as in the following phylogenetic trees:-

a) Matrix blosum60

A danderogram is constructed using the entire length of HSF1 protein sequences of human, mouse, Taurus, Danio rerio, plant (Arabidopsis), and yeast using the scoring matrix blosum60. Based on this specific matrix the human and Taurus are the closest in the structure of the entire length of HSF1 protein sequences. It is also shown that the plant (Arabidopsis) and the yeast are much closer to each other than to the other organisms.

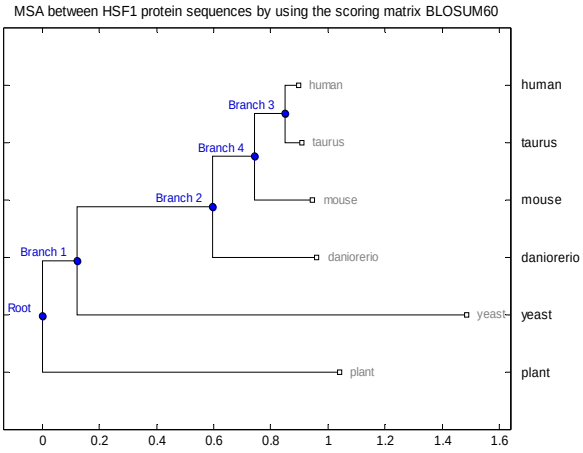

Fig 5: This figure shows the similarity between all sequences of the last species in case of HSF1 protein sequence by using the scoring matrix blosum60.

b) Matrix blosum80

When we use the scoring matrix blosum80 the same results are obtained as when using blosum60.

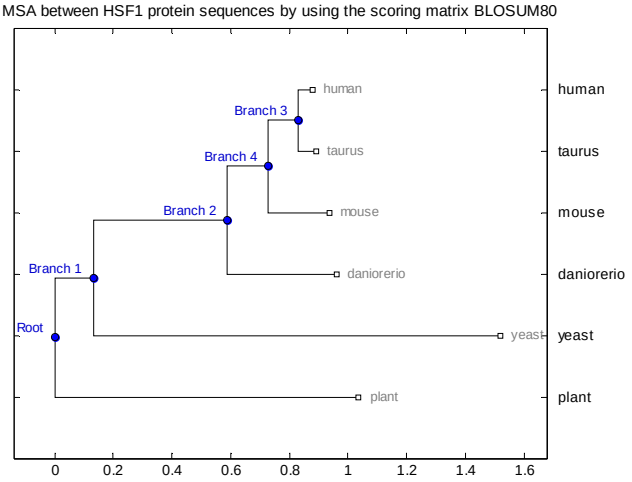

Fig. 6: This figure shows the similarity between all sequences of the last species in case of HSF1 protein sequence by using the scoring matrix blosum80.

c) Matrix PAM10

When we use the scoring matrix PAM10 the same results are obtained as when using blosum60 and blosum80.

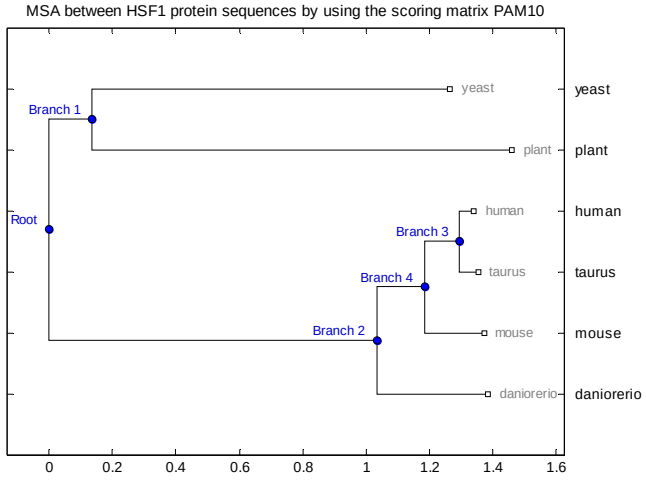

Fig. 7: This figure shows the similarity between all sequences of the last species in case of HSF1 protein sequence by using the scoring matrix PAM10.

C. Third, we use an algorithm, called Gene Tracer algorithm [20] which based on the pairwise local sequence alignment. Gene Tracer algorithm gives two ancestor sequences and their offspring one, tracks down genes modification in the ancestor sequences, and finds related parts of each ancestor in the offspring one. Gene Tracer algorithm is of complexity *O(max(M,N)*P)* in computing time and memory space, where *M, N* and *P* are respectively the lengths of the first ancestor sequence, the second ancestor one and the offspring one. As shown in table (1) and table (2) human and Taurus are more closer than other organisms on the entire length of the HSF1 sequence, but the HSF1 conserved domain (HSF_DNA-bind) sequence is more similar between human and mouse comparing to others. So that we use Gene Tracer algorithm to specify the related parts of each ancestor sequence in the offspring one. Moreover, Gene Tracer is used to find precisely the location of the ancestor sequences contribution inside the offspring one and gives statistical results that express the relationship between the two ancestor sequences and their offspring one. We consider Taurus as ancestor1 and mouse as ancestor2, and human as the offspring. We have coded Gene Tracer algorithm in Perl and have applied it on both the conserved domain sequences (HSF_DNA-bind) and the entire sequence of HSF1 protein. The results were as follow:

Table 3: This table shows the match percentage of the pairwise local alignment in the case of the entire sequence of HSF1 protein and also the conserved domain (HSF_DNA-bind) in HSF1 protein sequence.

Conserved domain (HSF_DNA-bind)	Match percentage	The entire sequence of HSF1 protein	Match percentage
Human & Mouse	100%	Human & Mouse	55.26%
Human & Taurus	96.19%	Human & Taurus	100%

The results in table (3) show that the highest similarity in case of the conserved domain sequence (HSF_DNA-bind) is between human and mouse with percentage 100% but in case of the entire sequence of HSF1 the highest similarity is between human and taurus .

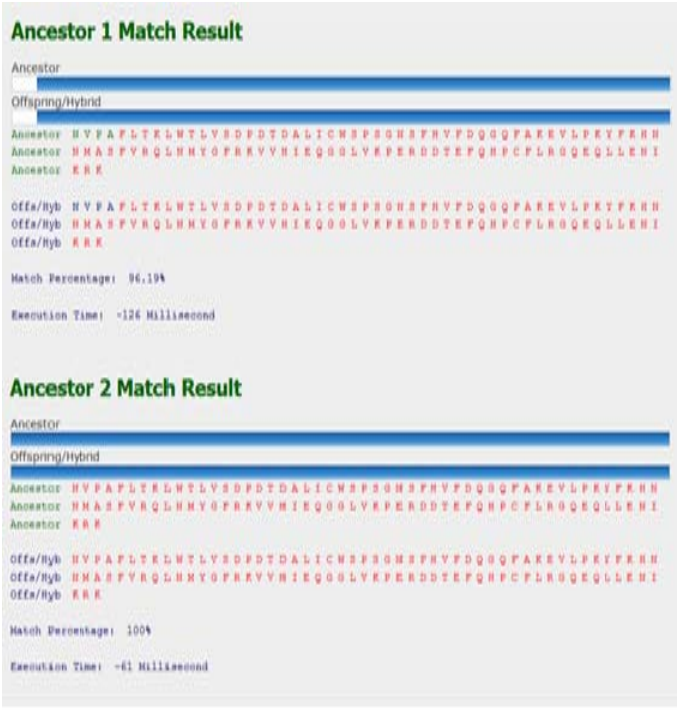

Fig. 8: Results of gene tracer program in the case of the conserved domain sequence (HSF_DNA-bind).

Fig. 9: Results of gene tracer program in the case of the entire sequence of HSF1 protein sequence

V. CONCLUSION

Heat shock factor (HSF) is a transcriptional activator of heat shock genes. Heat shock transcription factors (Hsfs) bind to conserved regulatory elements located in the promoters of HSP genes, known as heat shock elements (HSEs) [21]. Upon activation, the HSFs bind to HSEs and interact with proteins of the basal transcription machinery [22] and [15]. The presence of common HSF recognition elements explains the possibility for the same HSFs to be activated between different related organis [19]. The pairwise alignment comparison of HSF1 among different studied eukaryotic organism (eg. Human, Taurus, Dania rerio, Mouse, Plant (Arabidopsis), Yeast) shows that the human and Taurus are closer on the entire length of the HSF1 using the scoring matrix BLOSUM30 and BLOSUM50 as shown in table (2). However, the HSF1 conserved domain (HSF_DNA-bind) sequence was more similar between human and mouse comparing to others by using the same blosum matrices as shown in table (1). Similar results are obtained using multiple sequence alignment. As shown in table (3), results obtained by using gene tracer algorithm ensured that the conserved domain (HSF_DNA-bind) in mouse is the same in human and the entire sequence of HSF1 protein in human is the same as in Taurus. In our future work we intend to implement the GPU (graphical processing unit) capacities in studying HSF1 in larger database of sequences and higher number of organisms.

REFERENCES

[1] Report "IUCN Red List of Threatened Species" 2010.

[2] Maddison. "The Tree of Life Web Project".2007.

[3] C. S. B Needleman, C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of molecular biology, vol. 48, no. 1, pp. 443-453. 1970.

[4] X. Huang, K. M. Chao, "A generalized global alignment algorithm, Bioinformatics", Vol. 19, N°2: (2003), p228–233.

[5] R. A. Cartwright: Ngila, "global pairwise alignments with logarithmic and affine gap costs", Bioinformatics, Vol. 23, N°11:, p1427–1428. (2007).

[6] T. F. Smith, M. S. Waterman, "Identification of common molecular subsequences", J. Molecular Biology, no. 147, pp. 195-197, 1981.

[7] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CABIOS 10, 19-29.

[8] M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, "A model of evolutionary change in proteins, in Atlas of Protein Sequence and Structure", chapter 22, National Biomedical Research Foundation, Washington, DC: (1978), p345–358.

[9] S. Henikoff, J. G. Henikoff, "Amino acid substitution matrices from protein blocks", Proc. Natl. Acad. Sci. USA, Vol. 89, N°22: (1992), p10915-10919.

[10] Clos J, Westwood J T, Becker P B, Wilson S, Lambert K, Wu C "Molecular cloning and expression of a hexameric Drosophila heat shock factor subject to negative regulation". Cell. 63, 1085-1097. (1990).

[11] Wu C "Heat stress transcription factors. Annu" . Rev. Cell Biol. 11:441–469. (1995).

[12] Nover L, Bharti K, Dring P, Mishra S K, Ganguli A, and Scharf K D "Arabidopsis and the heat stress transcription factor world": How many heat stress transcription factors do we need? Cell Stress Chaperones 6, 177–189. (2001).

[13] Pelham H R B "A regulatory upstream promoter element in the Drosophila HSP70 heat-shock gene". Cell. 30, 517–528. (1982).

[14] Pelham H R B, Bienz M "A synthetic heat-shock promoter element confers heat-inducibility on the Herpes simplex virus thymidine kinase gene". EMBO J. 1, 1473–1477. (1982).

[15] Nover L, "Expression of heat shock genes in homologous and heterologous systems. Enzyme Microb". Technol. 9, 130–144. (1987).

[16] Schöffl F, Prändl R, Reindl A "Regulation of the heat-shock response". Plant Physiol 117,1135-1141. (1998).

[17] Prandl R, Schoffl F "Heat shock elements are involved in heat shock promoter activation during tobacco seed maturation". Plant Mol Biol 31,157–162. (1996).

[18] Coca M A, Almoguera C, Thomas T L, Jordano J "Differential regulation of small heat-shock genes in plants: analysis of a water stress inducible and developmentally activated sunflower promoter". Plant Mol Biol 31, 863–876. (1996).

[19] Yokotani N, Ichikawa T, Kondou Y, Matsui M, Hirochika H, Iwabuchi M, Oda K, "Expression of rice heat stress transcription factor OsHsfA2e enhances tolerance to environmental stresses in transgenic Arabidopsis". Planta. 1432-2048. (2007).

[20] M.Eissa, A.M.Alzohairy, H.Abobakr, I.Zidan, "Gene-Tracer: Algorithm Tracing Genes Modification from Ancestors through Offsprings". International Journal of Computer Applications (0975 – 8887) Volume 52– No.19, August 2012

[21] Sorger et al. "Stress-induced oligomerization and chromosomal relocalization of heat-shock factor". Nature 353, 822 - 827 (31 October 1991); doi:10.1038/353822a0. (1991).

[22] Morimoto, R. I. " Regulation of the heat shock transcriptional response:cross talk between a family of heat shock factors, molecular chaperones, andnegative regulators". Genes Dev. 12, 3788-3896. (1998).

Authors' Profile

Dr. Osama Abdo Mohamed is assistant professor in King abd El-Aziz University. He has got Msc. Degree in computer science in. 1998 & PH.D. degree in computer science in 2007. He has more than 17 years teaching experience and more than 17 years research experience in the field of signal processing and logic programming. He has published more than 10 national and international research papers in various refereed journals

Dr. Ahmed Mansour Mohamed Mansour Alzohairy, Ph.D. (2000–2005) in (Plant Signal transduction), Thesis entitled (Manipulation of Genetic information in studying plant performance) from Faculty Of agriculture, Zagazig University. Currently he is working as Assistant professor of Genetics andAgricultural Genetic Engineering in Faculty of Agriculture, Zagazig University.He has published 13 papers in bioinformatics

Dr. Mohamed Mohamed Saleh, Ph.D. (1984-1985) in Integral Equations from Mathematics Departement, Faculty of Science, Zagazig University. Currently he is working as Assistant Professor Mathematics Departement, Faculty of Science, Zagazig University, Egypt

Mr. Gaber Hassan Alsayed Ahmed,BSC of Math and Computer Science in 2007 from Faculty of Science, Zagazig University, Egypt.I worked as Aresearch Assistant in Basic Sciences Departement, Faculty of Engineering Science, Sinai University, Egypt.