International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 3, May- June 2017 ISSN 2278-6856

Computational Modeling for Evolution of HSP90A Homologues

Leonid Datta[1], Abhishek Mukherjee[2], Srijita Banerjee[3] , Shampa Sen[*]

[1]School of Computer Science and Engineering (SCOPE),VIT University, Vellore 632014, India

[2]School of Computer Science and Engineering (SCOPE),VIT University, Vellore 632014, India

[3]School of Bio Science and Technology (SBST),VIT University, Vellore 632014, India

Abstract

Changes in the amino acids resulting from nucleotide substitutions lead to change in the protein structure and/or function. This phenomenon is known as molecular evolution. Starting from simple unicellular systems to higher organisms up to humans, there have been changes in the DNA or amino acid sequences as the organisms have evolved. But when we consider all the proteins present in these organisms, they have not evolved with the same consistency in all the organisms. The changes in the proteins are the results of the substitutions of one nucleotide in place of other. These changes, also known as mutations, gave way to differences in the sequences at certain places in different organisms. The changes in HSP90A for different species (Human, common chimpanzee, rhesus monkey, mouse rat and Neurospora) were studied using three nucleotide substitution models (GTR, HKY85 and K80). This is the first report comprising of comparison of nucleotide substitutions for six organisms at a place using computational algorithm for implementing the substitution models. Results obtained suggested high similarity between common chimpanzee and rhesus monkey. HSP90A protein in human was found to be the most similar to rat and least to Neurospora as compared to the rest of the organisms.

Keywords: Nucleotide substitution model, Evolution, HSP90A, Bioinformatics, Computational biology.

1. INTRODUCTION

Molecular evolution deals with the changes in the nucleotides in case of DNA/RNA or amino acids in case of protein, respectively. These changes lead to alterations in protein structure, shape and composition which in turn give way to changes in the function of the proteins. This is known as protein evolution. In depth study of molecular evolution, which is driven by forces such as mutation, recombination, genetic drift and selection, is very important for understanding the pattern of changes in evolutionary biology.

However, the rate of evolution wasn't same for different proteins; for example, it was found that the rates of evolution in cytochrome c and hemoglobin were different [1]. Later it was concluded that the functionally important areas in the proteins were found to mutate differently at nucleotide level as compared to the other areas [1]. These changes that affected the functionality of the protein were reported to occur due to changes in the nucleotide sequence in the DNA or RNA.

The specific parameters for the changes in the sequences of different proteins at nucleotide level have been measured by various models. These models include: GTR, JC69, K80, F81, HKY85, T92, and T93. In JC69 the substitution rates (μ) are usually considered to be the same and the only parameter [2]. In case of F81 model the base frequencies (where base frequency of a specific nucleotide is calculated as the ratio of the number of mutations to that specific nucleotide and the total number of mutations) were considered to vary with a probability of 0.25 because this model does not distinguish between transition and transversion. Hence, when a particular nucleotide changes to any other nucleotide, it will be treated similarly by the model irrespective of the type of the change. [3]. Assuming that the base frequencies of the four nucleotides are same, the K80 model calculates the transition and transversion rates in the sequences, where the transition specifies the number of times base changes from purine to purine or pyrimidine to pyrimidine and transversion specifies number of times base changes from purine to pyrimidine and pyrimidine to purine [4]. T92 is another method which uses the ratio of total number of transitions to the total number of changes i.e. whenever a nucleotide changes from one base to another and also uses the ratio of total number of tranversions to the total number of changes for the calculations of the substitution model parameters [5]. HKY85 model considers the base frequencies differently and transition and transversion rates are taken into account [6]. In GTR model the most important factor is the condition that the number of forward and reverse mutations for a particular nucleotide must be same in all possible cases and the base frequencies of different nucleotides are differently calculated [7]. In the present study, more than two organisms were taken into consideration whereas in previous reports usually sequences (mostly, mitochondrial genome owing to its conserved gene sequences) of related species have been compared [8]. Here in this work, nucleotide sequences of HSP90A (α subunit of HSP90) from different animals and a fungus were compared using different substitution models namely, GTR, HKY 85 and K80 were used. HSP90 being the heat shock protein, gets activated under thermal stress and protects the cells from thermal shock [9]

2. METHOD

For the calculation of the parameters, the models were implemented in java and eclipse IDE was used for compiling and generating the output of the code. Sequences were taken from gene bank. Detailed methodology is given below.

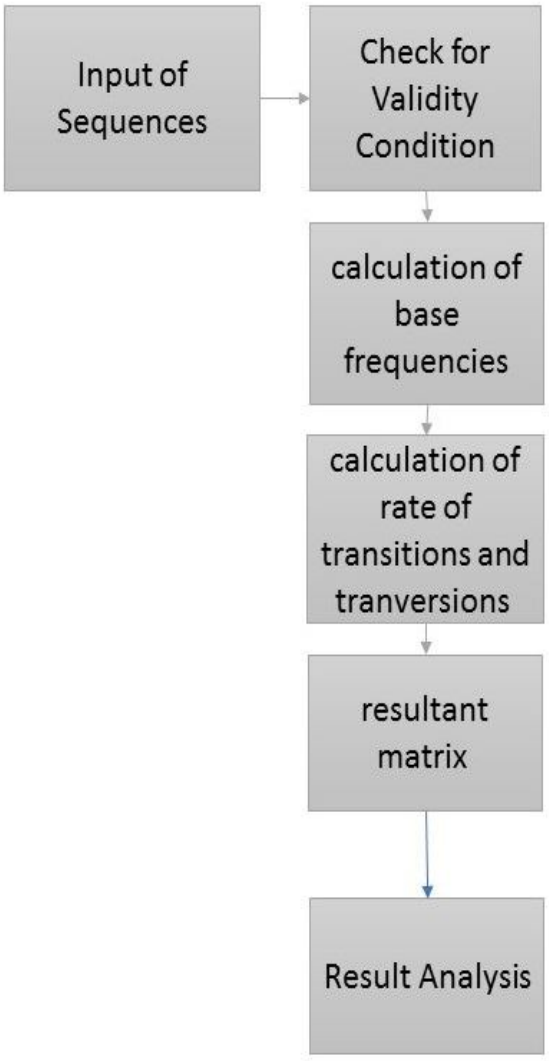

Fig. 1: Generalized Algorithmic Diagram

Substitution models help to mathematically analyze the mutation in gene sequences or in other words substitution of one nucleotide in place of other resulting from evolution. The tedious job of checking and calculating the type and rate of changes in the gene sequences containing large strings of nucleotides could be simplified by using java. While implementing the models two main factors were considered for code optimization purposes i.e. space complexity and time complexity. The program as an input takes, two gene sequences were used as input sequences, which were then compared to count the number of nucleotide substitutions using the program. Fig.1 explains method followed in general for calculation of the resultant matrix obtained from each model. Each Sequences of a protein expressed under stress were taken from different species i.e., gene sequence of α sub-unit of heat shock protein 90 (HSP90A) from six different species namely, human (Homo sapiens), chimpanzee (Pan troglodytes), rhesus monkey (Macaca mulatta), mouse (Mus musculus), rat (Rattus norvegicus), and a fungus (Neurospora crassa) were compared for the nucleotide substitutions using GTR,

HKY85 and K80 models.

There are two types of change or substitution of nucleotides possible in nature, namely, transition (where a purine is replaced by a purine or a pyrimidine by a pyrimidine) and transversion (where a pyrimidine is replaced by a purine or vice versa). It was required to iterate through the sequences to find the number of nucleotide sites where either type of change had taken place i.e., the number of places where each distinct type of changes namely, adenine to guanine, guanine to adenine etc. The number of iterations here depended upon the length of the sequences which were taken into consideration, as the length of the sequences could differ. However, here the number of iterations would be equal to the length of the smaller sequence.

GTR

In case of GTR model, the main criteria for validation was to check the six conditions where the number of changes in nucleotide sites between nucleotide of any two varieties and vice versa were the same i.e., changes from adenine (A) to guanine (G) were equal to changes from G to A and similarly changes from cytosine (C) to Thymine (T) were equal to changes from T to C, etc. The number of times any distinct type of change was taking place was stored to calculate base frequencies of all possible types and to check the conditions for the model to iterate through the two sequences.

The base frequencies were taken as π_A for base frequency of A, π_T for base frequency of T, π_G for base frequency of G, π_C for base frequency of C.

For the calculation of each type of changes taking place by comparing the two nucleotides of the sequences at a specific place, twelve variables were introduced and they were storing the values of changes as mentioned in table 1:

Table 1: List of parameters and corresponding types of mutation

Sl no	Parameter	Type of Mutation
1	α1	T->C
2	α2	C->T
3	β1	T->A
4	β2	A->T
5	γ1	T->G
6	γ2	G->T
7	δ1	C->A
8	δ2	A->C

9	ε1	C->G
10	ε2	G->C
11	η1	A->G
12	η2	G->A

Then, the specific cases were considered and checked if number of changes in terms of mutations and corresponding reverse mutations (namely C>T and T>C or A>G and G>A) were same in any of the six cases. If found that in any of the cases, it was not same, then it was evident that GTR model could not be applied for that set of sequences and also it was notified for which of the specific mutations the model failed.

In those cases, where GTR model could be applied, that the conditions, $\alpha1=\alpha2=\alpha$, $\beta1=\beta2=\beta$, $\gamma1=\gamma2=\gamma$, $\delta1=\delta2=\delta$, $\varepsilon1=\varepsilon2=\varepsilon$, $\eta1=\eta2=\eta$ would hold good. When the specific number of changes was being recorded, the total number of changes (i.e. A->A, T->T, G->G, C->C cases were not taken into account) are calculated and also the number of changes from any other nucleotide to a specific nucleotide (i.e., for nucleotide C, T->C, G->C, A->C were taken into account) were verified.

The base frequencies were calculated as
π_A = the number of changes from any other nucleotide to nucleotide A / total number of changes
π_T = the number of changes from any other nucleotide to nucleotide T / total number of changes
π_G = the number of changes from any other nucleotide to nucleotide G / total number of changes
π_C = the number of changes from any other nucleotide to nucleotide C / total number of changes

$$Q = \begin{pmatrix} -(\alpha\pi_C+\beta\pi_A+\gamma\pi_G) & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & -(\alpha\pi_T+\delta\pi_A+\varepsilon\pi_G) & \delta\pi_A & \varepsilon\pi_G \\ \beta\pi_T & \delta\pi_C & -(\beta\pi_T+\delta\pi_C+\eta\pi_G) & \eta\pi_G \\ \gamma\pi_T & \varepsilon\pi_C & \eta\pi_A & -(\gamma\pi_T+\varepsilon\pi_C+\eta\pi_A) \end{pmatrix}$$

The resultant matrix was calculated as given below [7]:

HKY 85

For the case of HKY model, in a similar way as stated above, two sequences were taken and nucleotides in specific spaces were compared with the condition that number of iterations was same as the length of smaller sequence. Base frequencies were calculated in similar manner as it was estimated in case of GTR model. But GTR model differs from HKY85 model in validity condition. The number of mutations and reverse mutations in a specific case (i.e. A->C and C->A) needs to be same in GTR model, but no such condition is applicable for HKY model. In HKY model, the total number of transitions and transversions were recorded.

The k factor was defined to be the ratio of number of transitions and transversions. The resultant rate matrix was calculated as given by Hasegawa et al [10]

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

K80

In case of K80 model, same procedure was followed as was done for GTR and HKY models. Two sequences were taken and comparing the specific position of different nucleotides in both the sequences, total number of transition and transversions were recorded. As in case of GTR and HKY85 models, in this model also, the smaller sequence was taken as reference sequence. K80 model assumes that all of the bases were equally frequent ($\pi_T=\pi_C=\pi_A=\pi_G=0.25$). p and q values were calculated as given below,

p = number of transitions/length of reference sequence
q = number of transversions/length of reference sequence
k value was calculated from the value of p and q as per the following equation [4]:

$$K = -\tfrac{1}{2}\ln((1-2p-q)\sqrt{(1-2q)})$$

2.RESULTS AND DISCUSSIONS

GTR

For the cases taken into consideration, the GTR model was not valid because in none of the pairs of sequences the number of mutations and reverse mutations as defined above, for example A→T and T→A, were same.

K80

The K80 model was applied on the same pairs of sequences. The transition and transversion rates were calculated and from that the k value was also obtained. The table below shows the results of p, q and k values between Homo sapiens and the other species.

Table 2: Result of K80 model

Sl. No.	P	Q	K	Sequence Sources
1	0.2489	0.4693	2.40	H. sapiens Vs. P. troglodytes
2	0.25	0.4690	2.42	H. sapiens Vs. M. mulatta
3	0.247	0.4792	2.60	H. sapiens Vs. M.

				musculus
4	0.2255	0.4671	1.93	H. sapiens Vs. R. norvegicus
5	0.2454	0.4863	2.78	H. sapiens Vs. N. crassa
6	0.2594	0.4678	2.54	M. musculus Vs. R. norvegicus
7	0.06	0.0943	0.17	P. troglodytes Vs. M. mulatta

From the values obtained from this model, maximum rate of transversion was found in Neurospora crassa followed by Mus musculus, Pan troglodytes, Macaca mulatta and finally the lowest rate was found in Rattus norvegicus as compared to that of Homo sapiens. Whereas, maximum transition rate was observed in Macaca mulatta, followed by Pan troglodytes, Mus musculus, Neurospora crassa and finally the least transition rate was observed in Rattus norvegicus. Transversion rate (q) was found to be double of the transition rate (p) for all the cases. Moreover, p and q rates were not the same when they were compared with different pairs of species. Additionally, it could be inferred from the results that the usual rate of change was not observed [11]. Rather, in this case the least number of changes were found to have occurred in Rattus norvegicus when it was compared to Homo sapiens. Reports suggesting similarity between Rat and human jejunum was found which correlated with the results from the present study [12].

Even, the k values suggested the same thing. The least value was obtained in Rattus norvegicus with Pan troglodytes, Macaca mulatta, Mus musculus in the between with the highest value in Neurospora crassa. The morphological differences between Mus musculus and Homo sapiens were reported earlier similar to the findings of the present research work [13] [13] [13]. So, it is inferred from the results that all the proteins did not evolve in similar way as it was reported earlier [14]. The result also suggested that Rattus norvegicus and Homo sapiens would respond similarly under heat stressed condition. In case of Mus musculus and Rattus norvegicus a lot of difference was observed in the k value which implies that their response is not similar under heat stress, whereas a lot of similarity was found in case of Pan troglodytes and Macaca mulatta as the k value was found to be very low. This suggested that they respond in the same way under heat stress.

HKY85

The results of HKY85 model were found to be in agreement with the k80 model. In this case, instead of the transition and transversion ratio, the base frequencies were obtained along with the k value. The lower the value of k, lower will be the similarity between the sequences under consideration which means both the animals will respond in the same way when they are in heat stress condition.

Table 3: Result of HKY85 model

Sl. No.	K	π_A	π_G	π_C	π_T	Sequence Sources
1	0.33	0.30	0.26	0.20	0.25	H. sapiens Vs. P. troglodytes
2	0.33	0.29	0.26	0.19	0.25	Homo sapiens Vs. Macaca mulatta
3	0.32	0.30	0.26	0.20	0.25	Homo sapiens Vs. Mus musculus
4	0.29	0.30	0.25	0.20	0.25	Homo sapiens Vs. Rattus norvegicus
5	0.33	0.23	0.29	0.20	0.28	Homo sapiens Vs. Neurospora crassa
6	0.34	0.30	0.25	0.20	0.25	Mus musculus Vs. Rattus norvegic

						us
7	**0.06**	**0.28**	**0.22**	**0.20**	**0.29**	**Pan troglodytes Vs. Macaca mulatta**

The base frequency of A was observed to be the highest in case of Homo sapiensandRattusnorvegicus whereas the lowest was observed in Homo sapiens and Neurosporacrassa. In case of base frequency of G, the lowest change was observed between Rat and Human whereas highest change was observed when the model was tested between Homo sapiens and Neurosporacrassa. On checking the base frequency with that of C, the test between Homosapiens and Pan troglodytes gave the lowest base frequency and the highest base frequency was observed between Homosapiens and Neurosporacrassa. Test between Homosapiens and Neurosporacrassa, gave highest base frequency of T and Homosapiens and Macacamulata, gave lowest base frequency. In most of the results, the high base frequencies were observed between Homosapiens and Neurosporacrassa and thus the k value in this case was also observed to be very high compared to the other values. This suggested the difference in the sequences of this protein in these two species which also gives us an idea how differently they should react when these two-species come under thermal stress condition. On the other hand, a very low k value was obtained when in case of Homo sapiens and Rattusnorvegicus which went along with the previous k80 results that these two species react in that same way during thermal stressed condition which also suggested the similarity in the bases of the two species and was also found correlating with other reports as stated above [12]. Moreover, a high base frequency of A was observed between these two species. From this result, it can also be said that, the high base frequency of A does not lead to much of a difference in the k value. When it was looked into from another angle, a very low base frequency of A was obtained in case of Neurospora crassa and Homo sapiens still a very high k value was obtained. This gives us an idea that the change to A does not cause much of a change in the protein as compared to the other nucleotides.

When we looked into the k values of the closely related species, a high k value was obtained in case of Mus musculus and Rattus norvegicus. This result was also found to be similar to that of the k80 and it was inferred that there is not much similarity in the HSP90A protein in these two species. In case of HKY85 also, a high k value was obtained which suggests the large differences in the proteins of these two species. On applying the model between Pan troglodytes and Macaca mulata, a very low k value was obtained which suggests that these two-species reacted approximately the same way during heat stressed condition which again was in agreement with the results obtained from K80.

3. Conclusions

From the results obtained, the highest similarity was observed between Rat and Human in both K80 and HKY85. This also gives us an idea that both of these species will react in the similar way during heat stressed condition. Other than this, common chimpanzee and rhesus monkey are also seen to have high similarity which suggested that these species would react in the same way under thermal stress.

4.Acknowledgements

The authors are thankful VIT University for providing supports in form of space and facilities required for this work.

References

[1] C.-I. W. Justin C. Fay, “Sequence Divergence, Functional Constraint, And Selection in Protein Evolution,” Annual Review of Genomics and Human Genetics, pp. 213-235, 2003.

[2] J. T. H. &. C. C. R, “How Many Nudeotide Substitutions actually took place?,” Evolution of protein molecules. (Munro H N, eel.), pp. 21-132, 1990.

[3] J. Felsenstein, “Evolutionary trees from DNA sequences: A maximum likelihood approach,” Journal of Molecular Evolution, pp. 368-376, 1981.

[4] M. Kimura, “A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequence,” Journal of Molecular Evolution, pp. 111-120, 1980.

[5] M. N. Koichiro Tamura, “Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees,” Molecular Biology Evolution, pp. 512-526, 1993.

[6] H. K. T.-a. Y. Masami Hasegawa, “Dating of the Human-Ape splitting by a Molecular Clock of Mitochondrial DNA,” Journal of Molecular Evolution , pp. 160-174, 1985.

[7] S. Travare, “Some probabilistic and statistical problems in the analysis of DNA sequences,” in Lectures on mathematics in the life sciences , American Mathematical Society , 1986, pp. 57-86.

[8] G. B. a. P. Lemey, “Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency,” BIOINFORMATICS, pp. 1970-1979, 2013.

[9] D. H. a. T. M. Michael Schroda, “The Chlamydomonas heat stress response,” The Plant Journal, pp. 466-480, 2015.

[10] H. K. a. T.-a. Y. Masami Hasegawa, “Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA,” Journal of Molecular Evolution , p. 1985, 160-174.

[11] R. J. Britten, “Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels,” PNAS, p. 13633–13635, 2002.

[12] U. F. J. Lennernäs, "Comparison Between Permeability Coefficients in Rat and Human Jejunum," Phamaceutical research, pp. 1336-1342, 1996.
[13] M. S. R. A. S. Jurij Dolensek, "Structural similarities and differences between the human and the mouse pancreas," Islets, pp. e1024405-1 - e1024405-16, 2015.
[14] W. R. Pearson, "An Introduction to Sequence Similarity ("Homology") Searching," Curr Protoc Bioinformatics, 2013.

AUTHORS

Leonid Datta is currently a final year under graduate student of School of Computer Science and Engineering (SCOPE), VIT University, Vellore, India.

Abhishek Mukherjee is currently in his final year of B.Tech. in computer science from VIT University, Vellore, Tamil Nadu, India.
He has presented various papers in various IEEE conferences. He has won the best paper award for his paper titled 'The amalgamated algorithm' in the 2ndInternational Conference on Human Computer Interactions (ICHCI). He has also presented his paper titled 'Big Data as service' in the 3rd International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS). His research interest includes parallel processing, Big data processing and data science. He is currently a software development intern at Citrya Innovations. He is a technical blogger as well and writes about various big data frameworks available on selected technical forums. He also takes technical sessions on big data related technologies.

Srijita Banerjee is currently perusing her 2nd year in B.Tech. program in Biotechnology from VIT University, Vellore, Tamil Nadu, India.
She has presented a poster in 17th All India Congress of Cytology and Genetics & Symposium on "Exploring Genomes: The New Frontier" organized by CSIR – Indian Institute of Chemical Biology & Archana Sharma Foundation of Calcutta on December 22-24, 2015. She was awarded with Merit Scholarship for best academic performance in the academic year 2016-17.

Dr. Shampa Sen has completed her graduation in Chemical Engineering (1998) and masters in biotechnology (2003) from Utkal University, Bhubaneswar, India and Jadavpur University, Kolkata, India respectively. After an year of academic job, she joined Ph.D. at Indian institute of technology, Guwahati and completed it by 2010. She had also served as Reader at Shree Hari Atmiya Centre for PG Studies, Rajkot, Gujarat, Lecturer at Meerut Institute of Engineering and Technology, Meerut (U.P.) and Jagannath Institute for Technology and Management, Paralakhemundi (Orissa). Her area of interest includes Nanobiotechnology and bioremediation of persistent pollutants. She has published her research in number of international journals and conferences.