Pseudo-Reverse Approach in Genetic Evolution: An Empirical Study with Enzymes

Sukanya Manna and Cheng-Yuan Liou[+]

Abstract— A pseudo-reverse approach is presented in this paper to analyze the evolutionary behaviour of enzymes. It employs the standard model of Nei and Gojobori [1] in a generalized form for determining the nucleotide substitutions and Jukes and Cantor's [2] model for finding the out their rates. Comparative genomics is also embedded in this model to calculate the lineages among the species like human, mouse and rat for these enzyme proteins. It is predicted from this study that the mutation for the enzymes are comparatively slower than ordinary proteins and the time of divergence for these enzymes with human and mouse or rat is almost five times more, around 400 Million years. Hence, this paper describes the methodology and the findings in details.

***Index Terms*— genetic evolution, pseudo-reverse approach, nucleotide substitutions, enzymes, comparative genomic, evolutionary time.**

I. Introduction

The rates of molecular evolution generally vary among lineages. Different studies have predicted that the source of this variation have differential effects on the synonymous and nonsynonymous substitution rates [3]. Changes in generation length or mutation rates are likely to have an impact on both the synonymous and nonsynonymous substitution rates. Hence, the number of substitutions per site between nucleotide sequences has become to be one of the most fundamental quantities for molecular evolution studies. It provides a valuable means for characterizing the evolutionary divergence of homologues. Thus accurate quantification of genetic evolutionary distances in terms of number of nucleotide substitutions between two homologous DNA sequences is an essential goal in evolutionary genetics. When two coding regions are under analysis, it is important to distinguish between the numbers of synonymous and nonsynonymous nucleotide substitutions per site. Estimation of calculation of these rates are not very simple, several methods have been developed to obtain these estimates from a comparison of two sequences [4], [5]. The early methods have been improved or simplified by many authors [1], [6], [7] and [8]. Those methods follow almost the same strategy. The numbers of synonymous (S) and nonsynonymous (N) sites in the sequence and the numbers of synonymous (S_d) and nonsynonymous (N_d) differences between the two sequences are counted. Corrections for multiple substitutions are then applied to calculate the numbers of synonymous (ds) and nonsynonymous substitutions per site (dn) between two sequences. These methods assume the equal base and codon frequencies.

Enzymes being protein in nature, belong the subset of existing proteins. Hence we believe that like other proteins they too play an important role in the evolutionary process. So, we have used them here for this case study. The approach used here is pseudo-reverse in the sense that we converted the amino acid sequences of the respective genes for the enzymes back to the nucleotide sequences based on the cumulative probability of the codons of the genomes of the species taken into account here. We then applied comparative genomics and the nucleotide substitution process to analyze and test this experiment. Comparative genomics is applied to align the sequences of each species' pairs: human, mouse and rat.

II. Method

A. Assumptions and Implementation

This work is proceeded on the basis of three major assumptions. Firstly, mammalian species, such as human and mouse share a vast majority of their genes [9], [10]. Secondly, most genes are subject to much stronger selective constraints on nonsynonymous changes than on synonymous ones [11], [12]. Finally, the genes found for an enzyme for a species are closely related to one another. The first two are common assumptions with [13] about comparative genomics.

Nei and Gojobori's model is the simplest model for nucleotide substitution schemes. Hence, we have used this along with Jukes and Cantor's model to find out the

[+] Correspondent: Cheng-Yuan Liou, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R. O. C.
cyliou@csie.ntu.edu.tw

nucleotide substitution rates. We implemented much generalized model of the above mentioned algorithm. Here, instead of using the transition matrix or codon substitution matrix, we directly calculated the aligned codon positions to compute this. Besides this, previously used models used phylogenetic approach of codon comparison [14], but we used here simple codon by codon comparison of the two sequences using a sliding buffer of length of three characters. We estimated the divergence time for the species pairs using the formula $T=K/2E$, where E is rate of evolution, T is species' divergence time and K is base pair substitutions per site [15].

B. Approach

We basically collected a set of enzymes from the enzyme database BRENDA [16]. Then we used Swiss-Prot knowledgebase [17] to collect the related genes' amino acid sequences for each enzyme for three species individually. For this case study, we considered only those enzymes for which we found valid genes in all three species. We then filtered out the data by separating the amino acid sequences having the terms like fragments, precursors, clones, and dominants and kept the mostly related sequences with respect to the enzymes considered. We assumed that the amino acid sequences obtained for each enzymes share a great similarity as what the genes belonging to the same group do. Instead of finding out the conserved regions between two species, we found out the least mismatch in their amino acid sequences for respective enzymes for each species' pair. We then collected those amino acid sequences, which satisfy this condition for each of the species pairs. In fact we believe that, more the similarity between the sequences, less is the mismatch between their amino acid sequences. We used here the amino acid sequences that have multiple numbers of least mismatches. So we explain here in brief about the concept of all pairs with least mismatch.

Let $H= [h_1, h_2, h_3,\ldots, h_n]$ be a set of genes for human, $M=[m_1, m_2, m_3,\ldots, m_m]$ be a set of genes for mouse and $R=[r_1, r_2,r_3,\ldots,r_k]$ be a set of genes for rat and n, m and k are the number of genes found in each species respectively for a particular enzyme. Suppose, h_1m_1, h_1m_3, h_2m_2, h_2m_3, h_1r_2, h_2r_5, h_1r_1, m_1r_1, m_1r_2, and m_2r_6 have the least mismatch in their sequences when compared among species pairs. We use all these pairs for the species wise sequence comparison for estimating the nucleotide substitution rates. So, for accomplishing this, we generated the random nucleotide sequences for amino acids h_1, h_2, m_1, m_2, m_3, r_1, r_2, r_5, and r_6 respectively for that particular enzyme.

The role of pseudo-reverse mechanism comes into picture when we convert the amino acid sequences back to the nucleotide sequences. But the conversion of all possible sequences was an absurd idea to be accomplished because of very high time as well as the space complexity. So we retrieved the total frequency of all the codons from the genomes of each species separately. Later we calculated the cumulative probability of the codons from the frequency obtained, and generated the random nucleotide sequences for all the amino acid sequences having the least mismatch for a particular enzyme. The Fig. 9 shows the frequencies of codons obtained. We generated 100 sequences for each of these amino acid sequences, because we were aware of the false positive and false negative outcomes. Next we compared these random sequences species' pair-wise (like human-mouse, mouse-rat, and mouse-rat respectively) to calculate the *dn/ds* ratio as mentioned earlier. There were 10,000 possible comparisons for each pairs per enzyme proteins (or genes). Out of these some returned valid results and others could not due to very low count of synonymous substitutions per site. We then plotted the graphs based on the valid results of the enzymes obtained.

III. Experimental Results

In this section, the results obtained from this work are illustrated in details. Figs 1 to 6 illustrate the variation of the *dn/ds* ratio with different enzymes along with the species pair comparison. Here the numbers in brackets along x-axis denotes the number of codons compared for each case. The last two figures, i.e., Fig. 7 and Fig. 8 depict the estimated time that we have obtained from this case study. The abbreviations like HM, MR, and HR signify Human-Mouse, Mouse-Rat, and Human-Rat species' pairwise sequence comparisons respectively. Now, in Fig. 1, the diagram clearly depicts the behaviour of the enzymes in different species pair comparisons. The ratio for the HM and HR is almost consistent here for these enzymes, but varies in case of MR. All these show purifying selection.

Figures 2 and 3, show similar kinds of results for two different enzymes, *Transaldolase* and *Carboxylesterase* respectively. But the former shows purifying selection and the latter shows diversifying selection for the corresponding species pair comparisons. In both the cases, we found more than one least mismatches in their corresponding amino acid sequences.

Now, Fig. 4 shows the enzymes found only in HM comparison for which we got a valid result, but not in the other pairs.The enzymes Trypsin and Alkaline phosphatase belong to this category. The Fig. 5 shows the comparison between the *dn/ds* ratio for the enzymes found only in the MR and HR species pairs. Both of them show purifying selection as *dn/ds* ratio is less than one. On the other hand, we see a drastic change in Fig. 6. It shows diversifying selection and the value for the enzyme *Ribonuclease* which shows a very high value. This means that the genes taken into account for this enzyme vary a lot in their behaviour. Here, we plotted the individual cases for MR and HR respectively like disjoint sets.

In the Figures 7 and 8, that the estimated divergence time for human and rat/mouse in cases of enzyme proteins seems to be 5 times higher as ordinary proteins which is around 80Myr [15]. The estimated range is ~400Myr. Fig. 8 shows the variation for the enzymes found in all the three species' pair comparisons. It is clearly seen from Fig. 7, the amino

acid replacements take longer times for all the species considered here in cases of these enzymes.

In table. 1, we have showed our results with the same set of enzymes as in [15]. We have calculated the *dn/ds* ratio from the original source, and used this here in the table. Here, we also see that many enzymes do not give us valid result using our reverse approach, inspite of having data in the already established work. We represent these in the form of NVR. We illustrated the results separately for HM as well HR.

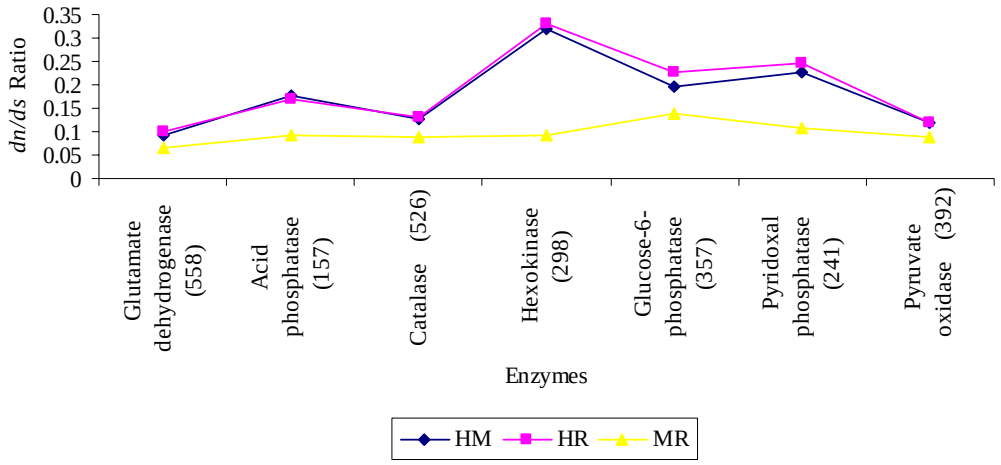

Figure 1: Comparison between dn/ds ratio of the enzymes common in all

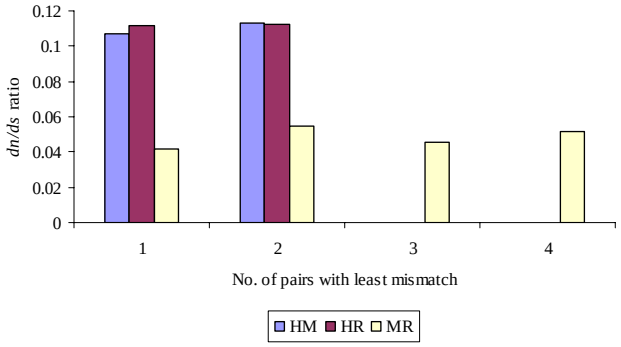

Figure 2: Comparison between dn/ds ratio of the enzyme *Transaldolase* for all pairs with least mismatches

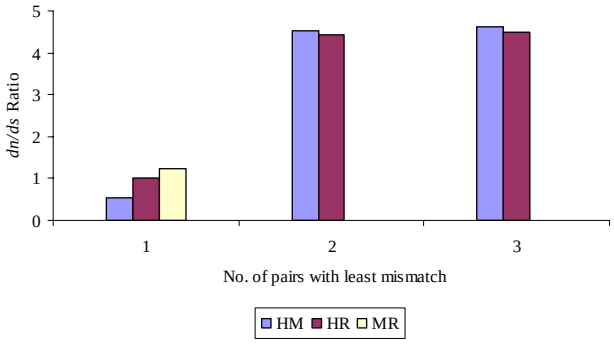

Figure 3: Comparison between dn/ds ratio of the enzyme *Carboxylesterase* for all pairs with least mismatches

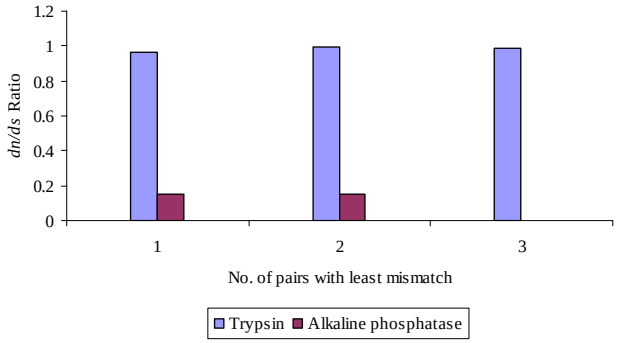

Figure 4: dn/ds ratio for the enzymes in HM having more than one least mismatch

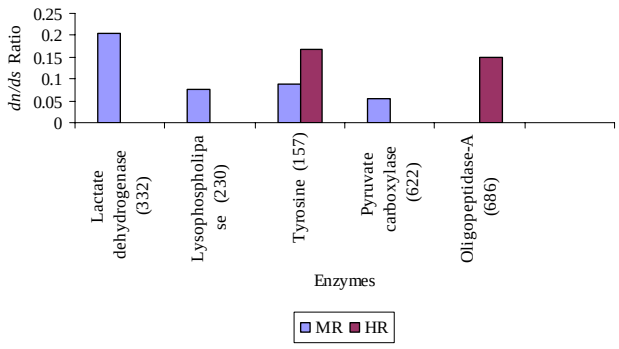

Figure 5: dn/ds ratio of species pairs with purifying result

Figure 6: dn/ds ratio of species pairs with diversifying result

IV. Conclusion

This work has emphasized some important facts in regard to the evolutionary trends of enzymes. Normally, the rates of nucleotide substitution vary considerably from gene to gene. But the ones closer to each other show almost similar type of behaviour. Here, we have noticed, that many enzymes, inspite of being proteins in nature, do not provide us any valid result as shown by NVR in table 1. In these cases, the rate of synonymous change was so small, that a proper valid ratio could not be computed. For such case the nonsynonymous sites were comparative higher. Thus in this approach we found the accuracy rate to be around 50 to 55 %. The possible reason behind this result may be the random generation of the nucleotide sequences from the amino acid sequences which might have deviated much from the original one or the divergence between the two species may be very high for certain genes in those enzymes. We estimated here the divergence time between the species. We found that it is almost 5 times higher (~400Myr) than the ordinary proteins. So, we can say that these enzymes are five times stronger than the ordinary proteins. Since enzymes are considered to be biocatalysts, it remains unchanged even after the reaction is over. Thus, these take much longer time to get mutated since during evolution, accumulation of mutation is very slow.

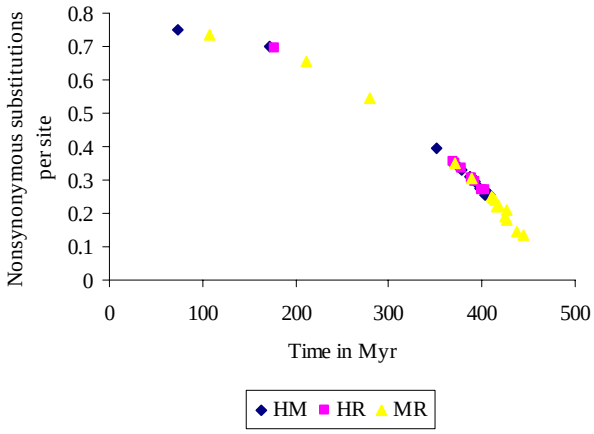

Figure 7: Estimated time for amino acid substitutions per site for enzymes

The table 1 shows a comparative study between the already established work, and our approach with the same set of enzymes. As far as the results are concerned, we can only classify them according to neutral or purifying or diversifying selection. We feel that this idea can establish some new concepts in the molecular evolution to trace back the relation between the genes and evolutionary times for the different protein.

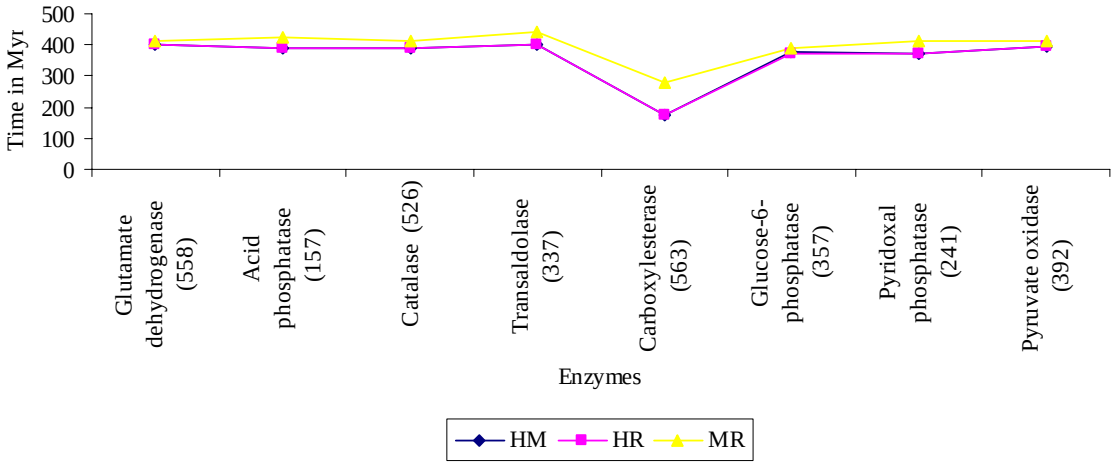

Figure 8: Estimated time for amino acid substitution per site for the enzymes common in all the three species

TABLE I
Comparison between already established result and our approach
(NVR – No Valid Results, H-Human, M-Mouse, R-Rat)

Enzymes	Li's Approach		Our Approach			
	Codons compared (H-M/R)	dn/ds ratio	Codons compared (H-M)	dn/ds ratio	Codons compared (H-R)	dn/ds ratio
Aldolase A	363	0.03	363	0.10	NVR	NVR
Creatine kinase M	380	0.06	381	0.10	381	0.10
Lactate dehydrogenase A	331	0.02	332	0.50	332	0.53
Glyceraldehyde-3-phosphate dehydrogense	332	0.09	332	NVR	332	NVR
Glutamine synthetase	371	0.08	372	0.10	372	0.11
Adenine phosphoribosyltransferase	179	0.19	179	NVR	179	NVR
Carbonyc anhydrase I	260	0.26	259	NVR	259	0.26

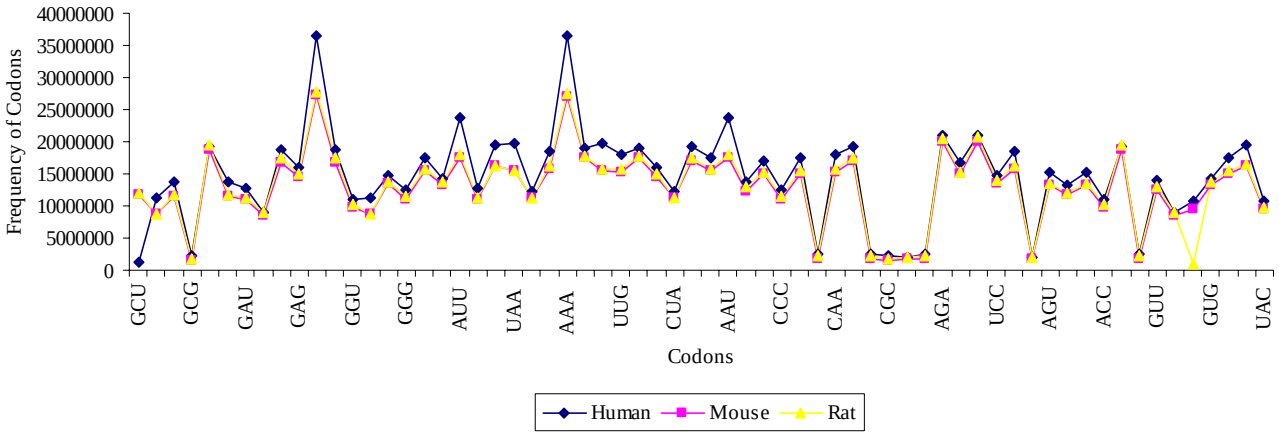

Figure 9: Frequency of the codons obtained from the genome of each species

ACKNOWLEDGMENT

This work is supported by National Science Council under the project no: NSC 94-2213-E-002-034.

REFERENCES

[1] M. Nei and T. Gojobori, "Simple Methods for Estimating Numbers of Synonymous and Nonsynonymous Substitutions," Mol. Biol. Evol., vol. 3, pp. 418-426, 1986.

[2] T. H. Jukes and C. R. Cantor, "Evolution of Protein Molecules. Mammalian Protein Metabolism", Academic Press, New York, 1969.

[3] T. K. Seo, H. Kishino, and J. L. Thorne, "Estimating Absolute Rates of Synonymous and Nonsynonymous Nucleotide Substitution in order to Characterize Natural Selection and Date Species Divergences", Mol. Biol. Evol., vol. 21, pp. 1201-1213, 2004.

[4] T. Miyata and T. Yasunaga, "Molecular Evolution of mRNA: A Method for Estimating Evolutionary Rates of Synonymous and Amino Acid Substitutions from Homologous Nucleotide Sequences and its Applications", J. Mol. Evol., vol. 16, pp. 23-36, 1980.

[5] W. –H. Li, C. –I. Wu, and C. –C. Luo, "A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitutions considering the Relative " J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982.

[7] P. Pamilo, and N. O. Bianchi, "Evolution of the Zfx and Zfy Genes: Rates and Interdependence between the Genes", Mol. Biol. Evol., vol. 10, pp. 271-281, 1993.

[8] J. M. Comeron, "A Method for Estimating the Numbers of Synonymous and Nonsynonymous Substitutions per Site", J. Mol. Evol., vol. 41, pp. 1152-1159.

[9] R. H. Waterston and *et al*., "Mouse Genome Sequencing Consortium: Initial Sequencing and Comparative Analysis of the Mouse Genome", Nature, vol. 420, pp. 520-562, 2000.

[10] E. S. Lamder and *et al*., "Initial Sequencing and Analysis of Human Genome", Nature, vol. 409, pp. 860-921, 2001.

[11] W. –H. Li, "Molecular Evolution", Sinauer, 1997.

[12] W. Makalowski and M. S. Boguski, "Evolutionary Parameters of the Transcribed Mammalian Genome: An Analysis of 2,820 Orthologous Rodent and Human Sequences", Proc. Natl. Acad. Sci. U.S.A., vol. 95, pp. 9407-9412, 1998.

[13] A. Nekrutenko, W. –Y. Wu, and W. –H. Li, "An Evolutionary Approach Reveals a High Protein-Coding Capacity of the Human Genome", TRENDS in Genetics, vol. 19, pp. 306-310, 2003.

[14] Z. Yang, "PAML: A Program Package for Phylogenetic Analysis by Maximum Liklihood", Comput. Appl. Biosci., vol. 13, pp. 555-556, 1997.

[15] D. Graur and W. –H. Li, "Fundamentals of Molecular Evolution", 2nd edn., Sinauer Associates Inc., 2000.

[16] BRENDA: Enzyme database; electronically available at http://www.brenda.uni-koeln.de/

[17] Swiss-Prot: Protein Knowledgebase; electronically available at http://ca.expasy.org/sprot/