Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 3 Number 1 (2010) pp. 71–77

http://www.ripublication.com/jcib.htm

The Distribution of Protein Sequences Based On Length

Chinnaiah Swaminathan Vinobha[1*], Ekambaram Rajasekaran[2] and Maruthamuthu Rajadurai[3]

*[1]Department of Bioinformatics, School of Bioengineering,
SRM University, Kattankulathur - 603 203, Kancheepuram Dt., Tamil Nadu, India.
[2]Department of Bioinformatics, School Biotechnology and Medical Sciences,
Karunya University, Karunya Nagar, Coimbatore – 641 114, Tamil Nadu, India.
[3]Department of Biotechnology and Bioinformatics,
Bishop Heber College, Tiruchirappalli - 620 017, Tamil Nadu, India.
Corresponding Author E-mail: csvinobha@gmail.com

Abstract

Despite a strong evolutionary pressure to reduce genome size, proteins vary in length over a surprisingly wide range also in very compact genomes. Here we investigated the hydrophobicity forces that act on protein size in 20 different living systems based on length. Overall the alteration in sequences length is subject to food habits and sexual behavior. Another observation is that only few sequences found with length less than 90. It is observed that during the evolution the fraction of large hydrophobic residues (FILMV) reduces considerably in animals compared to that of fungi and plant. To maintain the hydrophobicity, the length of the proteins increases in heterosexual species.

Keywords: Evolutionary; genome; hydrophobicity; fungi; plant.

Introduction

In general, the length of a protein sequence is determined by its function and the wide variance in the lengths of an organism's proteins reflects the diversity of specific functional roles for these proteins. However, additional evolutionary forces that affects the length of a protein, the length distributions of proteins evolving under weaker functional constraints. (Knight RD *et al.,* 2001)Proteins are the working force in all living systems. These proteins evolved to have a defined structure and specified function. These proteins are translated from mRNA. All proteins are not active in a given tissue. Only a cluster of them is active. The question is what the relationship between these tissue specific proteins is and why other proteins are not active? There are attempts to understand these complicated biological networks at gene level (DeRisi,J.L et al.,1997, Wen,X et al., 1998, Tamayo,P et al., 1999). In an attempt for example scientists try to identify clusters of genes having tissue specific profiles (Yeung,K.Y et al., 2001). Two of their algorithms able to cluster in line with experimental evidence while another algorithm separated them into different cluster. In another attempt the computational scientists apply different methods to identify the genes in a given genome. In this paper an attempt is made to understand the common features among diverse of species. In particular, the distribution profile of protein sequences based on length is studied in detail.

Methodology

The complete set of protein sequences of all 20 species studied here are downloaded from public NCBI database. The sequence lengths are measured as number of amino acid residues present in each sequence. The number of sequences found in each length is counted and tabulated using our own C programs. All possible lengths are taken into account for statistics. However, there are only few sequences having length greater than 5000. So the statistic is done up to 5000. And for better comparison, the plots are drawn up to the length of 1000. Total number sequences found with length less than 2000 are calculated and reported in terms of fraction. The number of sequences with length greater than 5000 is counted separately.

Results and Discussion

As can be seen in the Fig.1., the distribution of protein sequences based on length is uniform and skew type. In the figure only length 1000 are shown as not many sequences counted after that and no significant variation among the species at those lengths. The results show that there is no much variation in distribution among the fungi. At the length 250-300, there is a dip relative to the adjacent points in all three fungi and A.thaliana. The fraction at length 300-350 varies from species to species i.e. A.thaliana>S.pombe>K.lactis and S.cerevisiae. A.thaliana shows an increased number of sequences at length 300-350 compared to that of S.cerevisiae, K.lactis and A.pombe. The distribution profile of human, chimpanzee, cow and dog are shown figure 2. It shows clearly that the dog profile has a shift in right side and having highest number at length 300-350. That is length of protein sequences increased compared to that of human, chimpanzee and cow. The fractions at length 100-150 reduces from species to species. That is the chimpanzee has a higher fraction followed by human, cow and dog. This indicates that the alteration in sequence length is significant in dog and lesser in chimpanzee. Compared to human the chimpanzee has a higher degree of order in protein profile and dog has a lower degree of order. The decrease in fractions at length 100-150 and increase at length 300-350 is clear indication of increase in protein length. The food habits among the species result in change protein length and distribution profile.

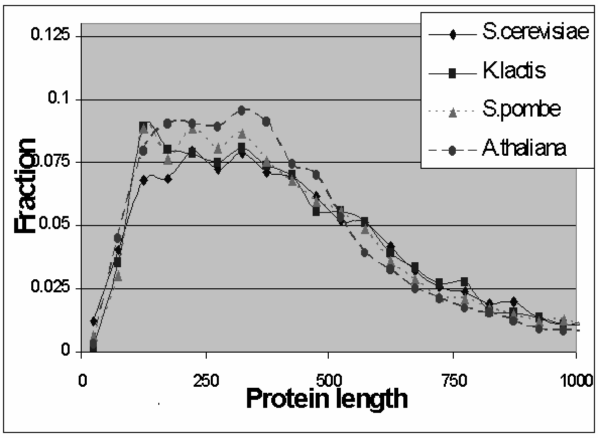

Figure 1: Distribution of protein sequences based on length in fungi and plant.

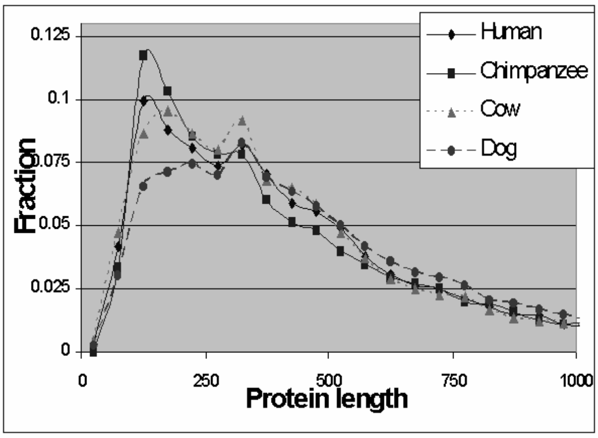

Figure 2: Variation of protein distribution profile in different heterosexuals with different food habit.

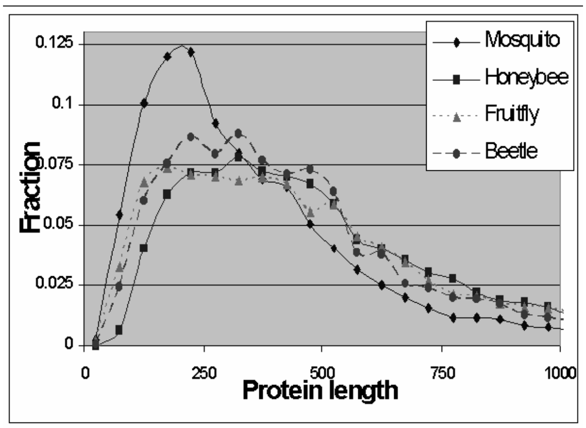

Figure 3: Comparison of protein distribution profile of mosquito, honeybee, fruitfly and beetle.

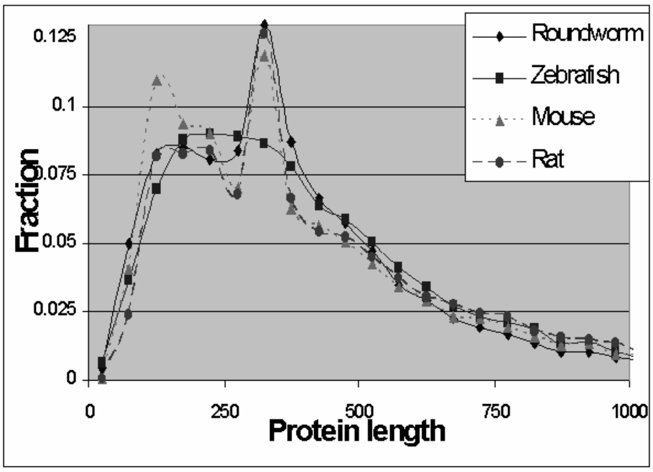

Figure 4: Variation of protein distribution profile of different heterosexuals lives in different environmental condition.

The distribution profile of mosquito, honeybee, fruitfly and beetle shown in figure 3. Almost all species show a relatively less fraction at length 250-300. Mosquito has a different distribution profile. The length variation is not very significant. The distribution profiles of some of the heterosexuals' lives in different living conditions are shown in figure 4. The increase in fractions of protein sequences at length 300-350 is significant in round worm, rat and mouse. Round worm, rat and mouse. The mouse has higher fractions at both lengths 100-150 and 300-350. The rat has increased amount of sequence at 300-350 and correspondingly decreased at length 100-150 compared to mouse. But at length 250-300 both rat and mouse has same amount of sequences. Rat and roundworm show higher amount of sequences at length 300-350.

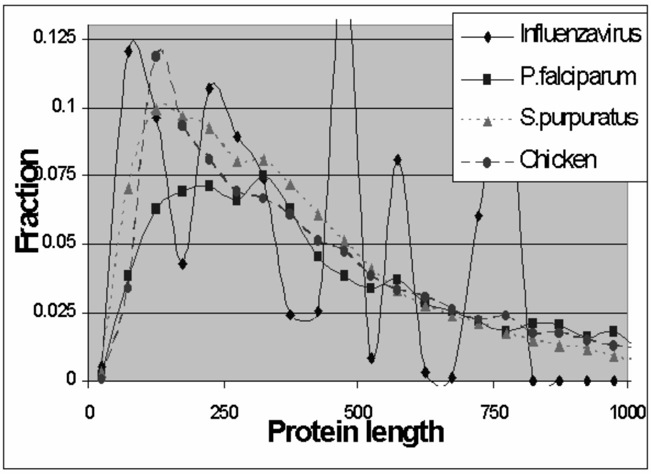

Figure 5: Difference in distribution profile of viruses are shown here.

The protein length based distribution profile of Influenza virus, P.falciparum, S.purpuratus and chicken are shown in figure 5. Both Influenza virus and P.falciparum has different profiles compared to other species. In particular the Influenza virus follow totally a random distribution. This virus has protein sequences of length not more than 800.

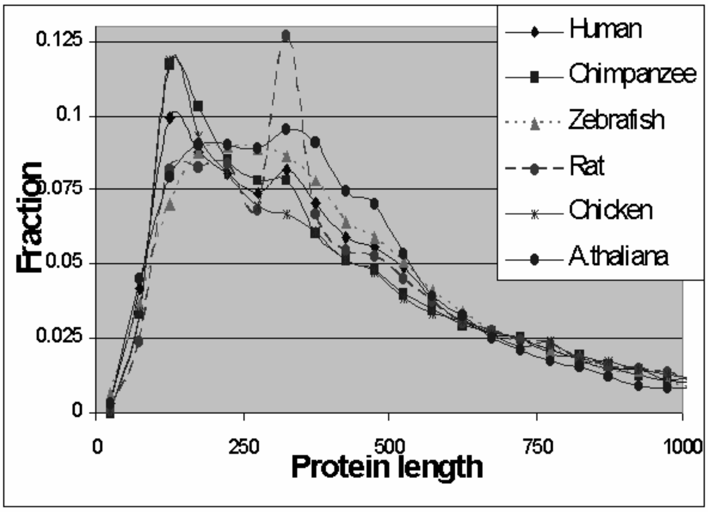

Figure 6: Comparison of distribution profile of human with animals, fish, bird and plant.

A comparison of distribution profile of human with animals, fish, bird and plant are made at figure 6. That the distribution profile of human, chimpanzee, zebrafish, rat, chicken and *A.thaliana* are plotted and compared. It is clear that the chimpanzee distribution profile is closer to human than the other species. Looking further into these profiles show that the fraction of sequences at length 100-150 is less and more at length 300-350 for human. It reveals that the lengthening of protein sequences higher in human compared to that of chimpanzee. That is lengthier proteins found more in human than chimpanzee. One reason for this alteration in protein length may be due to food habit. It seems that vegetarians are better lived than the counter part. The profiles show that the lengthening is less in chicken and more in rat. The distribution is uniform in plant compared to that of other heterosexuals. At length 250-300 almost all species show a dip in number of sequences. This length may not be a preferable one for protein to have a stable structure and function.

Table 1: The fraction of protein sequences having length less than 2000 in different species.

S.cerevisiae	K.lactis	S.pombe	A.thaliana
0.9937	0.9947	0.9936	0.9976
Human	Chimpanzee	Cow	Dog
0.9818	0.9788	0.9872	0.9780
Mosquito	Honeybee	Fruitfly	Beetle
0.9893	0.9701	0.9805	0.9829
Roundworm	Zebrafish	Mouse	Rat
0.9905	0.9892	0.9848	0.9780
Influenzavirus	P.falciparum	S.purpuratus	Chicken
1.0000	0.9250	0.9874	0.9706

The fractions of protein sequences with length less than 2000 are given in Table 1. *S.cerevisiae, K.lactis, S.pombe and A.thaliana* are having the fraction of 0.9937, 0.9947, 0.9936 and 0.9976 respectively. These species posses a higher degree of order ness in the protein sequences. At the same time the animals such as human, chimpanzee, cow and dog has the fraction of 0.9818, 0.9788, 0.9872 and 0.978 respectively. The dog has lower degree of order compared to human, chimpanzee and cow. The other species such as mosquito, honeybee, fruitfly, beetle, round worm, zebrafish, chicken and *S.purpuratus* has the fractions of 0.9893, 0.9702, 0.9805, 0.9829, 0.9905, 0.9892, 0.9706 and 9874 respectively. Again these species posses a lesser degree of order compared fungi and plant. Interestingly the rat and mouse has the fractions of 0.9780 and 0.9848. This clearly shows that the alteration at protein length is due to food habit and environmental factors. Overall, during the evolution the length of the protein increases in all species. It is phenomenal in heterosexuals because of mixing of DNA take place during reproduction. Apart from the heterosexual reproduction, the food habit also contribute towards this alteration ie., increase in length in protein sequences. The viruses, *P.falciparum* and influenza virus has the fractions of 0.925 and 1.0. This clearly indicates that the values are not comparable. The fractions either lower or higher compared to any of the other species studied here. Protein sequences with length greater than 5000 are counted for each species. That is the length of the proteins increases considerably in heterosexuals. There are no sequences having length greater than 5000 in *S.cerevisiae*, *K.lactis* and *S.pombe* and only two sequences *in A.thaliana* while dog have 33 sequences. The rat and mouse are having 29 and 18 sequences respectively. It confirms the presence of lengthier proteins in heterosexuals compared to plant and fungi.

Conclusion

The conclusion is that during evolution the length of the proteins increases considerably. The distribution is not uniform in all species but differ among them. Lengthier proteins found more in heterosexuals compared to that of plant and fungi. The heterosexual species are having increased number of sequences at the length of 300-325, which alters the distribution curve. This is significant in Mouse, Rat and *C.elegan*. There are hardly any sequences found up to the length of 89 in Human, Chicken, Chimpanzee, Rat and Mouse. But there are significant numbers of sequences counted at the length of 90, which alters the uniform skew distribution. Influenza virus has no uniform distribution and it is varying at random. It is observed that the fraction of large hydrophobic residues (FILMV) reduces considerably in animals compared to that of fungi and plant. To maintain the hydrophobicity, the length of the proteins increases in heterosexual species.

References

[1] Knight RD, Freeland SJ, Landweber LF 2001, “A simple model based on mutation and selection explains trends in codon and aminoacid usage and GC composition within and across genomes”. Genome Biol., 2: RESEARCH0010

[2] DeRisi,J.L., Iyer,V.R. and Brown,P.O. 1997, “Exploring the metabolic and genetic control of gene expression on a genomic scale”. Science., 278, pp 680-686.

[3] Wen,X., Fuhrman,S., Michaels,G.S., Carr,D.B., Smith,S., Barker,J.L. and Somofyi,R. 1998, “Large-scale temporal gene expression mapping of central nervous system development”. Proc. Natl. Acad. USA., 95.pp 334-339.

[4] Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. 1999, “Interpretting patterns of gene expression with self-organising maps: methods and application to hematopoietic differentiation”. Proc. Natl. Acad. USA., 96, pp2907-2912.

[5] Yeung,K.Y., Haynor,D.R. and Ruzzo,W.L. 2001, “Validating clustering for gene expression data”. Bioinformatics., 17(4), pp309-318.