Compositional analysis of protein sequences of different species

Vinobha C.S.[1]*, Rajasekaran E.[2] and Rajadurai M.[3]
[1]Department of Bioinformatics, School of Bioengineering, Faculty of Engg. & Technology, SRM University, Kattankulathur, 603203, Kancheepuram, Tamil Nadu, India
[2]Department of Bioinformatics, School Biotechnology and Medical Sciences, Karunya University, Karunya Nagar, Coimbatore, 641114, Tamil Nadu, India
[3]Department of Biotechnology and Bioinformatics, Bishop Heber College, Tiruchirappalli, Tamil Nadu, India, csvinobha@gmail.com

Abstract: We describe several protein sequence statistics designed to evaluate distinctive attributes of residue content and arrangement in primary structure. As per the global consideration, the compositional biases of clustering different residue types (charged residues, hydrophobic residues) of long runs of charged or uncharged residues, periodic patterns, counts and distribution of homooligopeptides, and unusual spacing between particular residue types. The computer program SEQUANA (statistical analysis of protein sequences) calculates all the statistics for any individual protein sequence input and is available for the WINDOWS environment through electronic mail on request to csvinobha@gmail.com
Keywords: bioinformatics, hydrophobic, compositional analysis, protein, software development, statistical analysis

Introduction

Advances in sequencing technology have taken the number of available sequences in database to unprecedented levels [1]. But unfortunately, the determination ability of the sequence of a particular gene does not accompany by an equally impressive gain in our ability to achieve insights into the biological function (including molecular and cellular) of these sequences [2, 3]. Living organisms are the product of a complex interplay of cellular processes. These processes are controlled through the interactions of molecules within and between cells [4, 5]. Much progress has been made in understanding the details of these phenomena through basic research in many areas of experimental biology [6]. Several protein sequence analysis algorithms are based on properties of amino acid composition, A number of groups have approached this problem by comparing amino acid composition and/or the distributions of pairs (or triplets *etc.*) of amino acids or nucleotides [7, 8]. Recently, protein search extended this approach to include global characteristics such as sequence length and calculated isoelectric point in addition to amino acid and pair composition. Using an optimized set of weigh for the various measurements, rather inconclusive sequence database searches were performed using multiple sequences as queries. Non-linear mapping of sequence composition data have also been used to cluster large set of sequences [9, 10].

METHODOLOGY

Sequence Collection:

The protein sequences can be downloaded from NCBI (National Center for Biotechnology Information) database. After concern time the protein sequences for the corresponding species are saved in the location as we mentioned.

SEQUANA:

The SEQUANA software was created so as to analyze the protein sequence at the residue level. The software computes compositional analysis for the given input sequence file. It calculates the total number of each amino acid present in every sequences, total number of polar amino acids, total number of non polar amino acids, and the net charge of each protein sequence for every species. The average values for each amino acid were determined using MS Excel worksheet. Using the average value, the graph was drawn which was available in the Excel worksheet.

Data input:

The downloaded Sequence file and the execution of the SEQUANA software kept in the same directory makes the software work conveniently. We can save the protein sequence either in the notepad format or word document file; it's according to our convenience.

RESULTS AND DISCUSSION

The protein sequences are analyzed for the following species, Human, *E.Coli,* Yeast and Virus. The sequences are directly downloaded from the ftp website of NCBI. The protein sequences having the extensions .faa are downloaded and it was analyzed using the SEQUANA software. All the proteins are computed by SEQUANA software and the results contains amino acid composition, total number of residues, total number of positive charged amino acids, total number of negative charged amino acids and fraction of each amino acid residues. Using the fraction value of each amino acid, graph was drawn in MS-Excel worksheet. The graph briefly explains the probability of each amino acid residue. Greater the probability will be the greater standard deviation. This shows that the positively charged (non polar) amino acids have either more or less probability whereas the negatively

charged (polar) amino acids have either more or moderate probability.

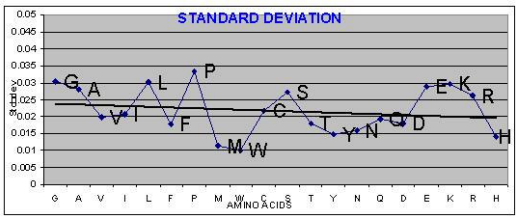

Fig 1: Standard deviation of amino acids in Human proteins.

The figure shows the Standard deviation of the amino acids in Human protein. The result obtained from the SEQUANA software was opened in MS Excel worksheet and the graph was drawn, so as to study the probable of each amino acid. The graph shows clearly that the amino acids on X-axis and the fraction values of each amino acid on Y-axis. The graph shows that all the amino acids present in Human protein lies between Standard deviation of 0.01 – 0.035 and the moderate amino acids lies at Standard deviation value of 0.02 – 0.025. In graph, for each amino acid linear line was drawn and also linear equation was found and the graph has some standard format like the Standard deviation values (0 – 0.5). This kind of analysis deeply explains / describes / shows each amino acid role in every protein of Human. The probable of each amino acid are analyzed using the chart. The entire chart had the equal scale so that we can easily analyze probable of each amino acid. The trend line and the equation were showed for each graph. The graph had the amino acid sequences at X-axis and the Fraction at Y-axis.

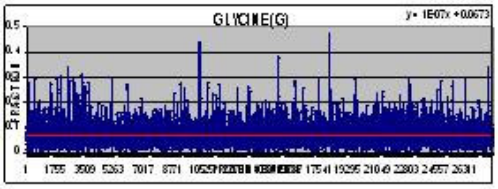

Fig 2: Distribution of Glycine in Human proteins

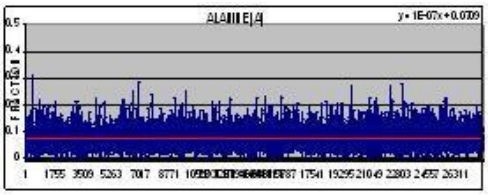

Fig 3: Distribution of Alanine in Human proteins

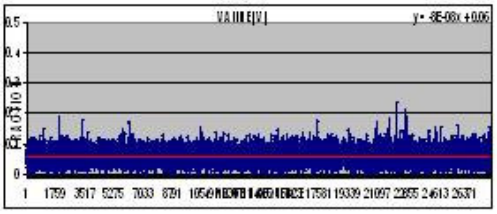

Fig 4: Distribution of Valine in Human proteins

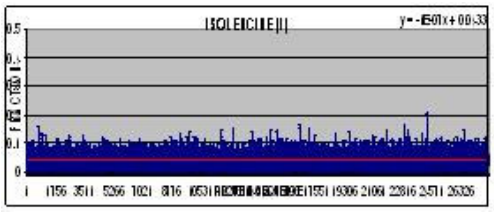

Fig 5: Distribution of Isoleucine in Human proteins

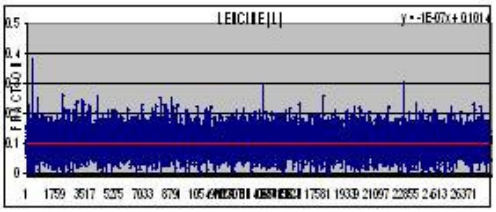

Fig 6: Distribution of Leuine in Human proteins

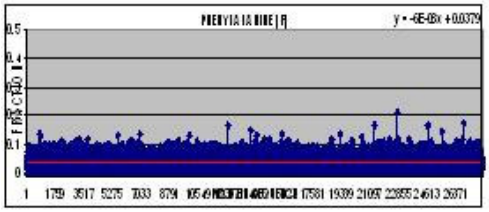

Fig 7: Distribution of Phenylalanine in Human proteins

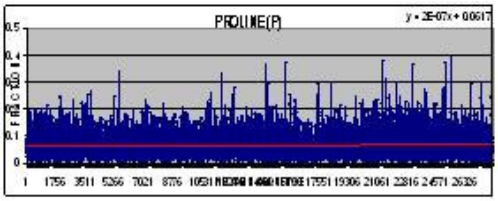

Fig 8: Distribution of Proline in Human proteins

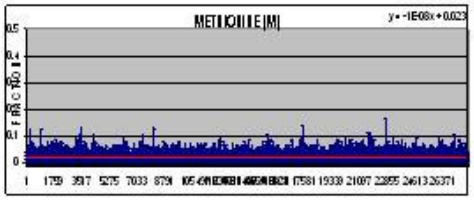

Fig 9: Distribution of Methionine in Human proteins

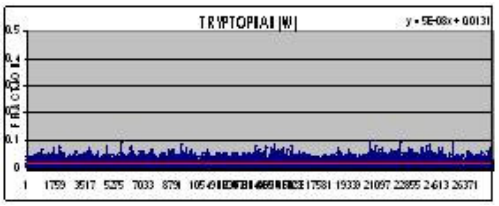

Fig 10: Distribution of Tryptophan in Human proteins

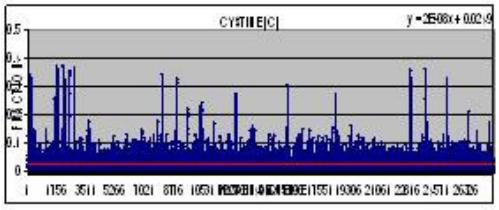

Fig 11: Distribution of Cystine in Human proteins

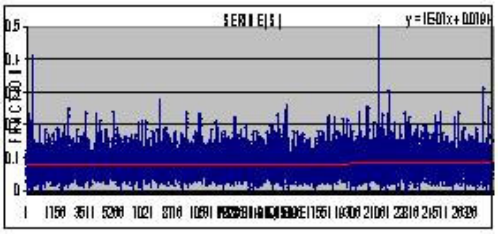

Fig 12: Distribution of Serine in Human proteins

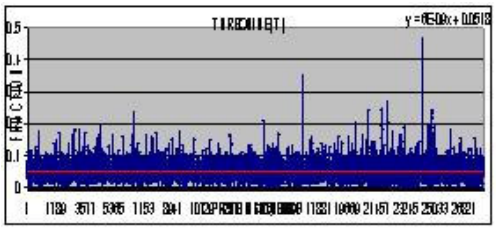

Fig 13: Distribution of Thrionine in Human proteins

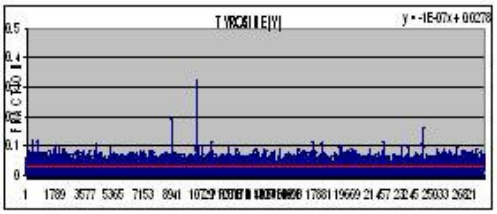

Fig 14: Distribution of Tyrosine in Human proteins

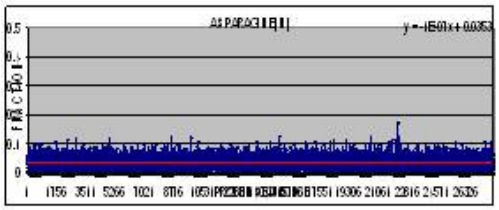

Fig 15: Distribution of Asparigine in Human proteins

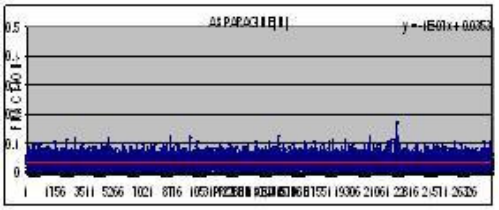

Fig 16: Distribution of Glutamine in Human proteins

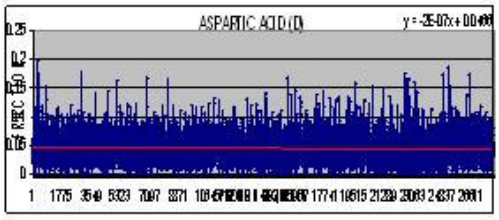

Fig 17: Distribution of Aspartic acid in Human proteins

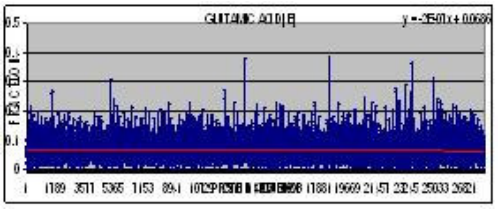

Fig 18: Distribution of Glutamic acid in Human proteins

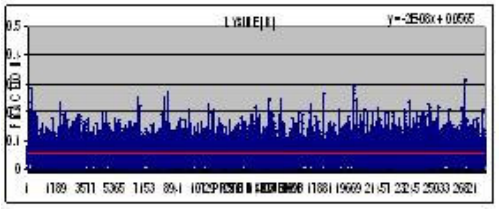

Fig 19:Distribution of Lysine in Human proteins

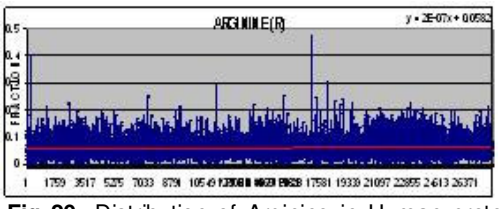

Fig 20: Distribution of Arginine in Human proteins

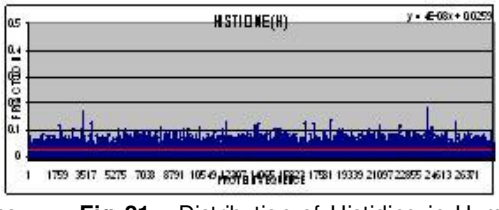

Fig 21: Distribution of Histidine in Human proteins

The fig 2 shows the probable of Glycine in human protein; the graph shows that it has more probable up to 0.5 and the trend line felt below o.1 in y-axis. The fig 3 shows the probable of Alanine in human protein, the trend line felt in between 0 and 0.1, this shows that it has more probable. The fig 4 shows the probable of Valine in human protein, the trend line felt somewhat closer to 0.1 and the probable of the amino acid was less. The fig 5 shows the probable of Isoleucine, the trend line felt in between o and 0.1 and it has less probable. The fig 6 shows the probable of Leucine, the trend line felt in between 0 and 0.1 and it has more probable. The fig 7 shows the probable of Phenylalanine, the trend line felt in between o and 0.1 and it has moderate probable. The fig 8 shows the probable of Proline, the trend line felt in between 0 and 0.1 and it has more probable. The fig 9 shows the probable of Methionine, the trend line felt in between 0 and 0.1 and it has moderate probable. The fig 10 shows the probable of Tryptophan, the trend line felt in between o and 0.1 and it has moderate probable. The fig 11 shows the probable of Cystine, the trend line felt in between 0 and 0.1 and it has less probable. The fig 12 shows the probable of Serine, the trend line felt in between o and 0.1 and it has more probable. The fig 13 shows the probable of Threonine, the trend line felt in between o and 0.1 and it has moderate probable. The fig 14 shows the probable of Tyrosine, the trend line felt in between o and 0.1 and it has moderate probable. The fig 15 shows the probable of Asparigine, the trend line felt in between o and 0.1 and it has moderate probable. The fig 16 shows the probable of Glutamine, the trend line felt in between o and 0.1 and it has less probable. The fig 17 shows the probable of Aspartic acid, the trend line felt in between o and 0.1 and it has less probable. The fig 18 shows the probable of Glutamic acid, the trend line felt in between o and 0.1 and it has more probable. The fig 19 shows the probable of Lysine, the trend line felt in between o and 0.1 and it has more probable. The fig 20 shows the probable of Arginine, the trend line felt in between o and 0.1 and it has more probable. The fig 21 shows the probable of Histidine, the trend line felt in between o and 0.1 and it has moderate probable.

Comparison of Human proteins with that of *E.Coli*, Yeast and Virus

Comparison of protein sequences is the major part, so the compositional analysis is done by means of the SEQUANA software for Human along with *E.Coli*, Yeast and Virus. The average values for each amino acid within the protein sequences of Human, *E.Coli*, Yeast and Virus were determined using MS Excel worksheet. The graph corresponding to the average values were also represented.

.

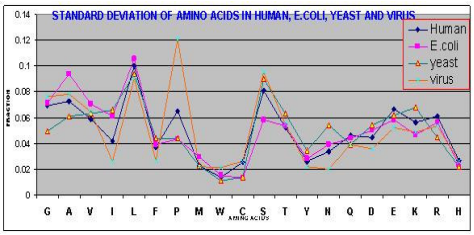

Fig 22: Standard deviation of amino acids in Human, *E.Coli*, Yeast and Virus.

The graph shows, the proteins of different species either they are closer to each other or deviated from one another. The proteins of species that were taken for analysis were follows almost same trend, only slight variation from one another. Especially the amino acid present in Virus, shows that either it has more probable or lesser probable, but it almost starts and ends at equal probable to other amino acids present in Human, *E.Coli* and Yeast.

Conclusion

We conclude that the amino acids distribution shows that the Virus sequences are having (i) larger (size) hydrophobic groups are less and (ii) smaller (size) hydrophobic groups are greater. The distribution of amino acids in *E.Coli* and Human are almost same with very slight variation whereas the Yeast having different distribution compared to Human.

References

[1] Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. (2008) GenBank. *Nucleic Acids Res.* 36:D25–D30.

[2] Brendel V., Bucher P., Nourbakhsh I.R., Blaisdell B.E., Karlin S. (1992) *Proc. Natl Acad. Sci. USA* 89.

[3] Friedberg I., Harder T., Godzik A. (2006) Nucleic Acids Res. 34:W379–W381.

[4] Boutet E., Lieberherr D., Tognolli M., Schneider M., Bairoch A. (2007) *Methods Mol. Biol.* 406:89–112.

[5] Ponting C.P. (2001) *Brief Bioinform.* 2:19–29.

[6] Krogh A., Larsson B., Von Heijne G., Sonnhammer E.L. (2001) *J. Mol. Biol.* 305:567–580.

[7] Van-Heel M. (1991). *J. Mol. Biol.* 220:877-887.

[8] Wu C., Whitson G., McLarty J., Ermongkoncha A., Chang T. C. (1992) *Prot. Sci.* 1:667-677.

[9] Ferran E.A., Pflugfelder B., Ferrara P., (1994) *Prot. Sci.* 3:507-521.

Hennig M., Darimont B., Sterner R., Kirschner K., Jansonius J.N. (1995) *Structure* 3:1295-1306.

Table 1: *Tabulation of Standard deviation values of amino acids Human, E.Coli, Yeast and Virus*

AMINO ACIDS	*E.COLI*	YEAST	HUMAN	VIRUS
GLYCINE (G)	0.071183	0.049574	0.069303	0.076344
ALANINE (A)	0.09371	0.061021	0.072901	0.078008
VALINE (V)	0.07091	0.063256	0.058928	0.064733
ISOLEUCINE (I)	0.060859	0.066086	0.041896	0.026079
LEUCINE (L)	0.105586	0.094064	0.099931	0.090442
PHENYLALANINE (F)	0.039206	0.043875	0.037049	0.026884
PROLINE (P)	0.043061	0.043982	0.065105	0.120799
METHIONINE (M)	0.0298	0.022719	0.022877	0.022054
TRYPTOPHAN (W)	0.015134	0.011164	0.013836	0.021473
CYSTINE (C)	0.012884	0.01363	0.025187	0.026166
SERINE (S)	0.058158	0.089921	0.081002	0.095414
THREONINE (T)	0.05377	0.06314	0.051925	0.054702
TYROSINE (Y)	0.027855	0.034002	0.025996	0.022191
ASPARIGINE (N)	0.038784	0.054004	0.033761	0.020018
GLUTAMI NE (Q)	0.044165	0.038846	0.046139	0.03896
ASPARTIC ACID (D)	0.050353	0.054429	0.044271	0.035626
GLUTAMIC ACID (E)	0.058293	0.061687	0.066362	0.051862
LYSINE (K)	0.004679	0.067619	0.056299	0.048818
ARGININE (R)	0.056067	0.044809	0.060767	0.053576
HISTIDINE (H)	0.023426	0.021878	0.026466	0.025852