International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:12, 2017

Computing the Similarity and the Diversity in the Species Based on Cronobacter Genome

E. Al Daoud

Abstract**—**The purpose of computing the similarity and the diversity in the species is to trace the process of evolution and to find the relationship between the species and discover the unique, the special, the common and the universal proteins. The proteins of the whole genome of 40 species are compared with the cronobacter genome which is used as reference genome. More than 3 billion pairwise alignments are performed using blastp. Several findings are introduced in this study, for example, we found 172 proteins in cronobacter genome which have insignificant hits in other species, 116 significant proteins in the all tested species with very high score value and 129 common proteins in the plants but have insignificant hits in mammals, birds, fishes, and insects.

Keywords**—**Genome, species, blastp, conserved genes, cronobacter.

I. Introduction

THE previous two decades have seen a blast of the hereditary information. Countless DNA sequences and genotypes have been produced, and they have prompted noteworthy biomedical advances and provided new insights into biology [1]. In addition, this information has significantly expanded our comprehension of patterns of hereditary variety among individuals and populations [2]. Interpreting of a given genomic sequence is one of the focal difficulties of science today. Maybe the most encouraging way to deal with this problem is based on the pairwise alignment and multiple sequences alignment methods. For example, protein-coding subsequences tend to be conserved between species. Subsequently, a straightforward strategy for recognizing a functional exon is to look for its homologue from related species using the whole genome alignment. Hence, enthusiasm for quicker, estimated, or heuristic (instead of ideal) alignment algorithms has increased [3]-[5]. Two of the most well known heuristic alignment procedures are implemented in the FASTA and BLAST packages. Comparisons of full genome sequences empower scientists to make inquiries that were unthinkable with small subsequences. Large-scale comparisons can uncover the genetic basis of speciation and variation, increase our understanding of the biological processes in living cells, recognize shared biochemical functions, expand our knowledge in human diseases and offer important information about evolutionary histories of extinct and living kinds [6], [7]. The whole genome is used in several studies such as utilizing data from one genome to understand another, identifying potential orthologs, comparison of genome content [8], genome alignment and genome signature analysis based on di-nucleotide abundance [9]-[11] among others.

Alignment of genomes implies identify differences that generated from mutational changes. In considering genome modifications, one differentiates between three important evolutionary operations: DNA mutations, genome rearrangements, and content alterations. DNA mutations impact on one or few nucleotides, while genome rearrangements work on bigger genomic subsequences and lead to change the orientation and the order of genes. Lastly, content alterations are an outcome of gene losses and duplications. Genome duplication has clearly permitted the development of more complex life forms; it equips an organism with a cornucopia of extra gene copies, which are allowed to change to fill unique needs. While one copy evolved for use in the brain, say, another evolved for use in the liver or adjusted for a novel reason. Therefore, the duplicated genes allow for increased sophistication and complexity [12]. In this study, we used 40 full genomes from 11 organisms to find the relationship between the species and discover the unique, the special, the common and the universal proteins. To trace the genes using bottom up approach, the cronobacter genome is used as reference genome.

II. Dataset

TABLE I
THE PROTEINS DETAILS OF BACTERIA GENOMES

ID	Species	# proteins	#AA	Avg. length
1	Cronobacter	3842	1244298	323.9
2	Salmonella	4770	1385186	290.4
3	Shigella	6409	1293263	201.8
4	Enterobacter	4289	1375730	320.8
5	Chlamydia	1013	356049	351.5
6	Cronobacter_sak	4442	1342730	302.3
7	Ecoli	4843	1508759	311.5

To find the distinguished genes and quantify sequence similarities, the full genome of 40 species from 11 organisms are downloaded from Kyoto Encyclopedia of Genes and Genomes site (KEGG) [13]. The species are selected to represent various branches of the phylogenetic tree of life and provide adequate coverage of main kinds within the evolutionary tree, including, seven bacteria, three protists, three fungi, three archaea, seven mammals, three birds, three fishes, five insects, a tick, a mollusk and four plants. Tables I-IV summarize the name of the selected species, the number of proteins and the average length (number of the amino acid) of each one.

Essam Al-Daoud is with the Faculty of Information Technology, Computer Science Department, Zarqa University, Jordan (phone: +96279668005, e-mail: essamdz@zu.edu.jo).

TABLE II
THE PROTEINS DETAILS OF PROTISTS, FUNGI AND ARCHAEA GENOMES

ID	Species	# proteins	#AA	Avg. length
8	Entamoeba_dispar	8811	3563877	404.5
9	Babesia_bovis	3706	1856394	500.9
10	Plasmodium_yoelii	7353	3382406	460.0
11	Laccaria_bicolor	18215	6700944	367.9
12	Aspergillus_nidulans	9541	5067689	531.1
13	Neurospora_crassa	10813	5632539	520.9
14	Pyrococcus_abyssi	1784	539209	302.2
15	Archaeoglobus_prof.	1823	478828	262.7
16	Methanotorris_igneus	1772	506747	286.0

TABLE III
THE PROTEINS DETAILS OF MAMMALS GENOMES

ID	Species	# proteins	#AA	Avg. length
17	Human	109052	73449745	673.5
18	Chimpanzee	79947	55635610	695.9
19	Mouse	76217	52262429	685.7
20	Cow	28901	18146954	627.9
21	Arabian_camel	26729	15276008	571.5
22	Elephant	29784	17488002	587.2
23	Minke_whale	34821	21600601	620.3

TABLE IV
THE PROTEINS DETAILS OF BIRDS AND FISHES GENOMES

ID	Species	# proteins	#AA	Avg. length
24	Chicken	46346	32575322	702.9
25	Saker_falcon	21235	12955188	610.1
26	Rock_pigeon	18582	11198213	602.6
27	Zebrafish	52829	38449214	727.8
28	Platyfish	23478	13384899	570.1
29	Coelacanth	34251	20280708	592.1

TABLE V
THE PROTEINS DETAILS OF INSECTS GENOMES

ID	Species	# proteins	#AA	Avg. length
30	House_fly	21304	13686004	642.4
31	Mosquito	14099	7371687	522.9
32	Honey_bee	22451	15287002	680.9
33	Leaf_cutting_ant	10657	6082041	570.7
34	Monarch_butterfly	15232	6424480	421.8

TABLE VI
THE PROTEINS DETAILS OF A TICK, A MOLLUSK AND PLANTS GENOMES

ID	Species	# proteins	#AA	Avg. length
35	Octopus_bimaculoide	23994	13806582	575.4
36	Black_legged_tick	20467	5810072	283.9
37	Thale_cress	48350	20856276	431.4
38	Rice1	28555	10301721	360.8
29	Wheat	33849	13570085	400.9
40	Chlamydomonas_rein	14489	6573428	453.7

III. GNOMES COMPARISONS AND MINING

To align two proteins, *blastp* is downloaded and called using MATLAB as:

system(['blastp -query crono.fa -db sp1 -out results.out -evalue .01 –num_alignments 5']);

where *crono.fa* is a query that is formatted as fasta file which will be compared with the genome *sp1*. The results are saved as NCBI file for each pair has expectation value < 0.01, and then, the results are interpreted and saved as a matrix:

M=ParseNCBI('results.out');

Four important values are extracted for each pair of the compared sequences, the values are the score, the expectation, the percentage of identities and the matches:

M.Hits(0).HSPs(1).Score;
M.Hits(0).HSPs(1).Expect;
M.Hits(0).HSPs(1).Identities.Percent;
M.Hits(0).HSPs(1).Identities.Match;

Algorithm 1 is used to find all universal genes with expectation value less than 10^{-33}:

Algorithm 1: Universal genes
```
For each protein in cronobacter j
  For each species i
    If exect(i, j)<1e-33
        count=count+1
  If count = num
    Print j
```

where num is equal to 40 for universal genes, more than 38 for near-universal genes and less than 3 for special and unique genes. Algorithm 2 is used to find the common proteins in one organism but not in the other organisms

Algorithm 2: Common genes
```
For each protein in cronobacter j
Flag=1
  For each species i in the target organism
    If exect(i, j)>Expet_value
        Flag=0
  For each species k not in the target organism
    If exect(i, j)< Expet_value
        Flag=0
  If flag = 1
    Print j
```

Algorithm 3 is used to find the maximum identical protein in a given species

Algorithm 3: Maximum identical protein
```
For each protein in cronobacter j
  Max=0
  For each species i
    If Ident(i, j)> max
      Max=iden(i,j)
```

IV. RESULTS AND DISCUSSIONS

Five algorithms are implemented using MATLAB and the package Blastp, where the cronobacter genome is used as reference genome, the implemented algorithms are to compare the proteins, interpret the results, find the common, the universal and maximum identical proteins. Cronobacter genome contains 3842 proteins, while the human genome

contains 109052 proteins. Hence, to compare the both genomes, we have to implement 3842*109052 pairwise alignments, which took 5.3 hours using 2.3 GHz dual-core CPU. To mine all the selected genomes, more than 3 billion pairwise alignments are implemented and took about 10 days. Fig. 1 shows the score of first 500 proteins of cronobacter after aligning it to E-coli and human genome, which illustrates the relationship between the both species. Fig. 2 shows the frequency of cronobacter proteins which have scoring value more than 250 when aligned with each species excluding bacteria genomes. The histogram suggests that the protists genomes (ID: 8-10) and archaea genomes (ID: 14-16) have the lowest homology and the plants genomes (ID: 37-40) have the highest homology with cronobacter genome and contains the most highly conserved proteins.

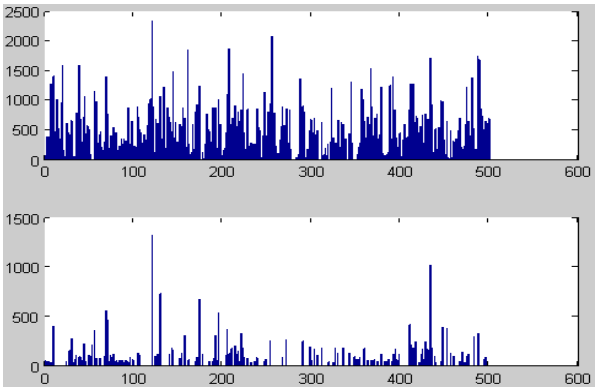

Fig. 1 The score of first 500 proteins with E-coli (top) and human genome

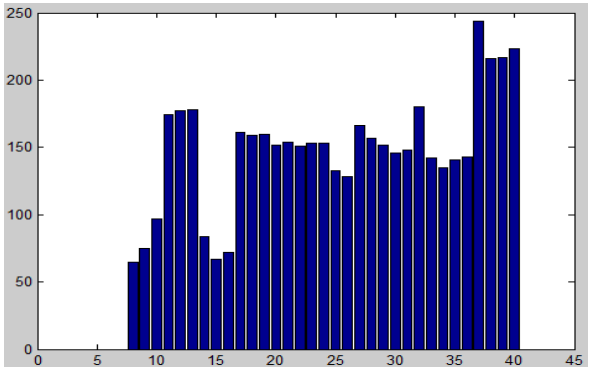

Fig. 2 Frequency of proteins which have scoring value more than 250 excluding Bacteria species

The following are some important findings:

- 172 unique proteins are found in Cronobacter and have insignificant hits in all tested genomes.
- Number of significant proteins with p-value $<10^{-10}$ and conserved in all tested species is 116
- Number of significant proteins with p-value $<10^{-50}$ and conserved in all tested species is 3, namely protein ID: 514, 2839 and 3047. The corresponding proteins name according to NCBI site are enolase, isoleucine tRNA ligase, and ATP-dependent metalloprotease. These proteins seem to be the core biological functions in the all living cells. The following is the amino acid sequence of the protein ID 514 in FASTA format:

> ALB69585.1 enolase
MSKIVKVIGREIIDSRGNPTVEAEVHLEGGFVGMAAA
PSGASTGSREALELRDGDKSRFLGKGVTKAVGAVNG
PIAQAIVGKDAKDQAGIDKIMIDLDGTENKSNFGANA
ILAVSLAAAKAAAASKGMPLYEHIAELNGTPGKFSMP
VPMMNIINGGEHADNNVDIQEFMIQPVGASSVKEAIR
MGSEVFHHLAKVLKGKGMNTAVGDEGGYAPNLGSN
AEALAVIAEAVKAAGYELGKDITLAMDCAASEFYKD
GKYVLAGEGNKAFTSEEFTHFLEDLTKQYPIVSIEDGL
DESDWDGFAYQTKVLGDKIQLVGDDLFVTNTKILKE
GIEKGIANSILIKFNQIGSLTETLAAIKMAKDAGYTAVI
SHRSGETEDATIADLAVGTAAGQIKTGSMSRSDRVAK
YNQLIRIEEALGEKAPYNGRKEIKGQA

- Protein ID 3666 has a significant hit (p-value $< 10^{-33}$) in human but insignificant in the Chimpanzee:

>ALB72737.1 gluconate kinase
MSTTNHDHHIYILMGVSGSGKSVVASEVAHRLKAA
FLDGDFLHPRRIMKMASGDPLNDDDRTPWLQALND
AAFAMQRTNKVSLIVCALKKRYRDILRSGNPNLSFI
WLKGDFEVIESRLRARKGHFFKPQMLVTQFEALEAP
QEDEKDVLFVDINQSLDDVIDSTIALINKGQ

The conserved proteins in the mammals are compared with other organisms, the following results are obtained with expectation value $<10^{-10}$:

- 738 conserved proteins are common in mammals and birds
- 510 conserved proteins are common in mammals, birds, fishes, insects and plants
- 19 conserved proteins are common in mammals, birds, fishes, insects but not in plants such as protein ID 2365 and 2890.
- 52 conserved proteins are common in mammals but not in insect such as protein ID 620 and 669.
- 300 conserved proteins are common in mammals but not in archaea such as protein ID 814 and 2329.
- 129 conserved proteins in plants but not in mammals, birds, fishes and insects such as protein ID 3671 and 14. Fig. 3 shows the distribution of protein ID 14. On the contrary, there are 47 conserved proteins are common in mammals but not in plant, such as protein ID 2365 and 2366.
- 11 conserved proteins are common in mammals but not in birds such as protein ID 1176.
- 87 conserved proteins are common in mammals but not in Fungi such as protein ID 121 and 3039.

Fig. 4 shows the scoring value of protein ID 1115, which is conserved (among other 42 proteins) in bacteria species but insignificant in the other tested species. On the contrary, protein ID 2839 is conserved in all tested species.

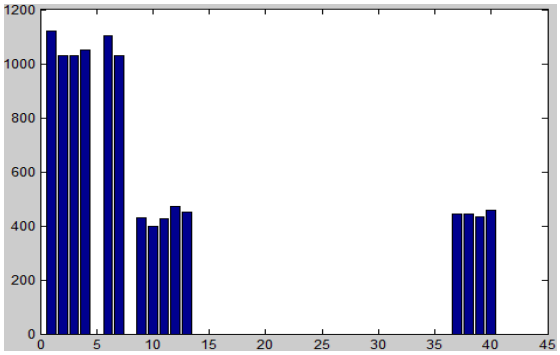

Fig. 3 The scoring of the protein ID 14 for each species

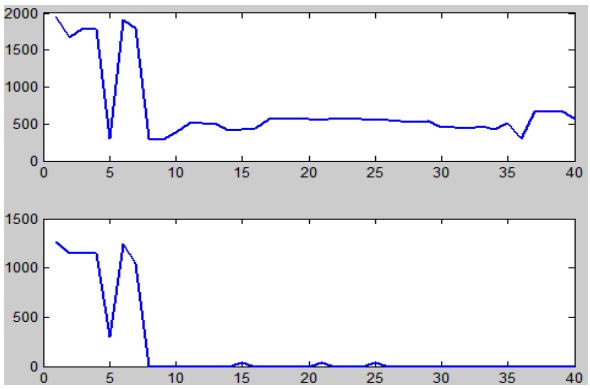

Fig. 4 The scoring value of protein 2839 (top) and protein 1115 (bottom)

Tables VII-IX highlight the number of significant, insignificant proteins and the maximum identical protein in the mammals, insects and plants. Protein ID 658 (ALB69729.1 scaffolding protein according to NCBI site) seems to be another important protein for mammals and insects. The Honey bee proteins appear to be odd when compared to other insects. The plants show more diversity than other organisms.

TABLE VII
NUMBER OF SIGNIFICANCE, IN SIGNIFICANCE AND MAXIMUM IDENTITIES OF MAMMALS GENOMES

ID	Species	<e-100	# insigf.	Max Iden. Protein ID
17	Human	103	2642	658
18	Chimpanzee	104	2633	658
19	Mouse	101	2656	658
20	Cow	102	2607	658
21	Arabian_camel	102	2608	658
22	Elephant	105	2608	658
23	Minke_whale	101	2627	658

TABLE VIII
NUMBER OF SIGNIFICANCE, IN SIGNIFICANCE AND MAXIMUM IDENTITIES OF INSECTS GENOMES

ID	Species	<e-100	# insigf.	Max Iden. Protein ID
30	House_fly	91	2737	658
31	Mosquito	89	2693	658
32	Honey_bee	110	2480	2958
33	Leaf_cutting_ant	83	2697	658
34	Monarch_butterfly	82	2710	658

TABLE IX
SIGNIFICANCE AND MAXIMUM IDENTITIES OF A TICK, A MOLLUSK AND PLANTS GENOMES

ID	Species	<e-100	# insigf.	Max Iden. Protein ID
37	Thale_cress	165	2348	67
38	Rice1	148	2370	1741
29	Wheat	140	2467	1262
40	Chlamydomonas_rein	140	2328	3600

V. CONCLUSION

The aim of whole genomes alignment is to utilize an ensemble of related genomes to better see every individual genome in the set and to discover the core biological functions. Albeit similar genomic investigations of many genomes are still generally uncommon contrasted of genomic investigations of specific groups of organisms, they are quickly expanding in number. Closing the gap between our capacity to create tremendous amounts of information utilizing computational techniques and our capacity to guarantee the resulting annotation will be a main objective of the following decade.

ACKNOWLEDGMENT

This research is funded by the Deanship of Scientific Research in Zarqa University /Jordan.

REFERENCES

[1] L. Sian, "Brain evolution: Genetic layering," *Nature reviews Neuroscience.* Vol.18, No. 6, pp 324-324, 2017.
[2] P. Blanco-Arias, C. A. Sargent, and N. A. Affara, "A comparative analysis of the pig, mouse, and human" *PCDHX* genes. *Mamm. Genome.* Vol 15, pp 296–306, 2004.
[3] A. Varki1 and K. T. Altheide "Comparing the human and chimpanzee genomes: Searching for needles in a haystack," *Genome Res.* Vol. 15, pp 1746-1758, 2005.
[4] Z. He, D. Han, O. Efimova, P. Guijarro, Q. Yu, A. Oleksiak, S. Jiang, K. Anokhin, B. Velichkovsky, S. Grünewald, et al. "Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques," *Nat. Neurosci* Vol. 20, pp 886–895, 2017.
[5] J. R. Dixon, D. U. Gorkin, B. Ren, "Chromatin Domains: The Unit of Chromosome Organization," *Mol Cell* Vol. 62, pp 668–680, 2016.
[6] B. I. Bae, D. Jayaraman, C. A. Walsh "Genetic changes shaping the human brain" *Dev Cell* Vol 32, pp 423–434, 2015.
[7] B. B. Lake, "Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain," *Science* Vol. 352, pp 1586–1590, 2016.
[8] A. Peltzer, G. Jäger, A. Herbig, A. Seitz, C. Kniep, J. Krause and K. Nieselt, "EAGER: efficient ancient genome reconstruction," *Genome Biol.* Vol. 17, No. 1, pp 60-74, 2016.
[9] M. A. McMahon, M. Rahdar, M. Porteus "Gene editing: not just for translation anymore," *Nat Methods* Vol. 9, pp 28–31, 2012.
[10] L. Pozzi, C. M. Bergey and A. S. Burrell, "The use (and misuse) of phylogenetic trees in comparative behavioral analyses," *International Journal of Primatology*, Vol 35, pp 32–54, 2014.
[11] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nat. Methods*, Vol. 12, pp 59–60, 2015.
[12] D. He, O. Fiz-Palacios, C. J. Fu, J. Fehling, C. C. Tsai and S. L. Baldauf, "An alternative root for the eukaryote tree of life," *Curr Biol,* Vol 24 pp 465-470, 2014.
[13] KEGG, http://www.genome.jp/kegg/catalog. Last access on October, 2017.