ANALYSIS OF THE GENOMIC ALIGNMENT OF MUTANT SEQUENCES OF P53 GENE

Antonio Alves dos Santos Neto[1]

Thomas Vaz Fluminhan[2]

Antonio Fluminhan[3]

Abstract: Bioinformatics tools available on the web were employed to analyze nucleotide sequences in p53 gene mutations from patients with skin carcinoma deposited in genomic banks in the United States and Europe. The feasibility of the proposed approach and the identification of common characteristics among the evaluated mutants were demonstrated. The gene regions most sensitive to changes in nucleotides, which may be related to the origin of the pathology, were identified. Mutations occurring in the p53 gene, which are derived from genetic sequences deposited at the NCBI Institute, are predominantly located in very restricted gene regions, which were identified between positions 23562 and 24735. Mutant sequences deposited at the EBI Institute have shown that the most frequent mutations occurr between positions 19844 and 21786 of the p53 gene. The occurrence of transitions (purine-purine and pyrimidine-pyrimidine) was predominant over the occurrence of transversions in mutants for the p53 gene. The present research demonstrated the multidisciplinary nature of these activities, increasing the relevance and productivity of bioinformatics analyses.

Keywords: Bioinformatics; Sequence similarity; Skin cancer; Mutations; p53 gene

1 São Paulo State Technological College - FATEC, Presidente Prudente, SP, Brazil
2 Pontifical Catholic University of Parana - PUCPR, Curitiba, PR, Brazil
3 São Paulo State University – FCAV UNESP, Jaboticabal, SP, Brazil

Introduction

The majority of noninherited malignant cutaneous tumors result from mutations caused by carcinogens that, in some way, promote damage to DNA and confer advantages that promote disordered cell growth and invasion of other tissues. Thompson et al. (1994) emphasized that the process of tumor formation begins only with a somatic cell somewhere in the organism.

The risk factors that contribute to the development of skin cancer are well known and include race, age, sex, chronic exposure to chemical and physical mutagens, and genetic factors (Parkin 2004; Martinez et al. 2006). According to the National Cancer Institute (INCA 2020), nonmelanoma skin cancer is the most common type in Brazil and accounts for approximately 25% of all malignant tumors registered in the country. If it is detected early, there is a high cure rate.

Among skin tumors, the nonmelanoma type has the highest incidence and lowest mortality. According to Corapcioglu et al. (2006), the p53 gene is one of the genes most studied by the global scientific community. According to Frebourg et at. (1992), which is located on the short arm of chromosome 17 (17p13.1) and encodes a tumor suppressor protein containing transcriptional activation, DNA binding and oligomerization domains. The encoded protein responds to several cellular mechanisms to regulate the expression of target genes, thus inducing, according to Varley et al. (2001), cell cycle arrest, apoptosis, senescence, DNA repair or changes in metabolism.

DNA damage was the first type of stress discovered in p53 activation; therefore, this gene is widely considered to be the “guardian of the human genome” (Efeyan and Serrano 2007). Mutations in this gene are associated with a variety of human cancers, including hereditary cancers (Hainaut and Hollstein 1999), and represent the most common genetic alteration in malignant cells (Morita et al. 2008).

As highlighted by Sansom and Smith (2000), biological data (DNA, RNA and protein sequences) stored in publicly accessible databases have facilitated the work of many scientists, especially in studies of comparisons of nucleotide sequences for different purposes, such as the

identification of genetic mutations or polymorphisms, and evolutionary and phylogenetic studies. The availability of this information and the ease of access to the internet have revolutionized the way research in this area is currently being carried out, significantly reducing the time allocated to work (Baxevanis and Ouellette 2004).

In the present research, bioinformatics tools were used for comparative analysis of mutant sequences of the p53 gene from patients affected by skin cancer and deposited in publicly accessible genomic banks. They were compared with each other and with the original gene sequence via multiple sequence similarity analysis tools to characterize the variations most commonly found in these mutants, as well as their possible relationships with pathology.

Materials and methods

Equipment: Computers with an Intel Core i5 processor, 8 GB of memory, and a 1 Tb HD with a 1 GB video card were used and configured with administrator permission and unrestricted internet access. The equipment was installed with the purpose of managing, processing and storing information cataloged from searches in genomic databases available on the internet and thus creating its own databases.

Databases: Mutant and original nucleotide and amino acid sequences corresponding to the p53 gene and its corresponding P53 protein, which are known to be related to the occurrence of skin cancer in humans, were searched and located in two publicly accessible databases on the internet (Benson et al. 2011). Existing databases were chosen from the United States - National Center for Biotechnology Information - NCBI (available at: http://www.ncbi.nlm.nih.gov/) and Europe - European Bioinformatics Institute - EBI (available at: http://www.ebi.ac.uk/).

Bioinformatics tools: A survey was carried out to identify which multiple alignment analysis tools for nucleotide and amino acid sequences were available, and a significant variety was found, as detailed in Table 1. Only public and free access software stored or made available by bioinformatics

research centers was considered.

Table 1. List of software for analyzing multiple alignments of genomic sequences that are freely available on the internet.

Software	Features
ClustalW (RRID:SCR_017277)	Series of widely used computer programs for multiple sequence clustering. Web service provided by DNA data bank of Japan: CLUSTAL, CLUSTALV, CLUSTALW, CLUSTALX and CLUSTAL OMEGA.
DIALIGN(RRID:SCR_003041)	Series of computer programs created for multiple sequence alignment. Variations: CHAOS-DIALIGN, DIALIGN-TX and DIALIGN-PFAM.
Kalign (RRID:SCR_011810)	A fast and accurate multiple alignment algorithm. It is widely used to align a large number of sequences.
MAFFT (RRID:SCR_011811)	Multiple sequence alignment program for Unix-like operating systems. It offers a range of several alignment methods, L-INS-i (accurate; for alignment of approximately 200 sequences), FFTNS-2 (fast, for alignment of approximately 30,000 sequences).
MUSCLE (RRID:SCR_011812)	Program that enables multiple Log-Expectation sequence comparison. Depending on the options chosen you can achieve better accuracy and greater speed than CLUSTAL W and T-COFFEE.
PCMA (RRID:SCR_024156)	It is a consistency-based multiple aligner that makes use of the progressive approach.
ProbCons (RRID:SCR_011813)	Tool to generate multiple alignments of protein sequences. Using a combination of probabilistic modeling and consistency-based alignment techniques.
T-Coffee (RRID:SCR_011818)	It can align protein sequences as well as DNA and RNA sequences.

Source: The authors and compiled from Edwards, Stajich, Hasen (2009).

Among the software programs listed in Table 1, ClustalW (RRID:SCR_017277), MAFFT (RRID:SCR_011811) and T-Coffee (RRID:SCR_011818) were chosen for the analyses. These tools are commonly cited and recognized for their use of fast algorithms with significant levels of precision. As not all software is available on the same platform, that is, some work on the Windows operating

system and others only on the Linux operating system, we opted for a web environment that does not require a specific operating system, just a browser, thus avoiding installing the software on the workstation, as previously described in Santos Neto and Fluminhan (2014).

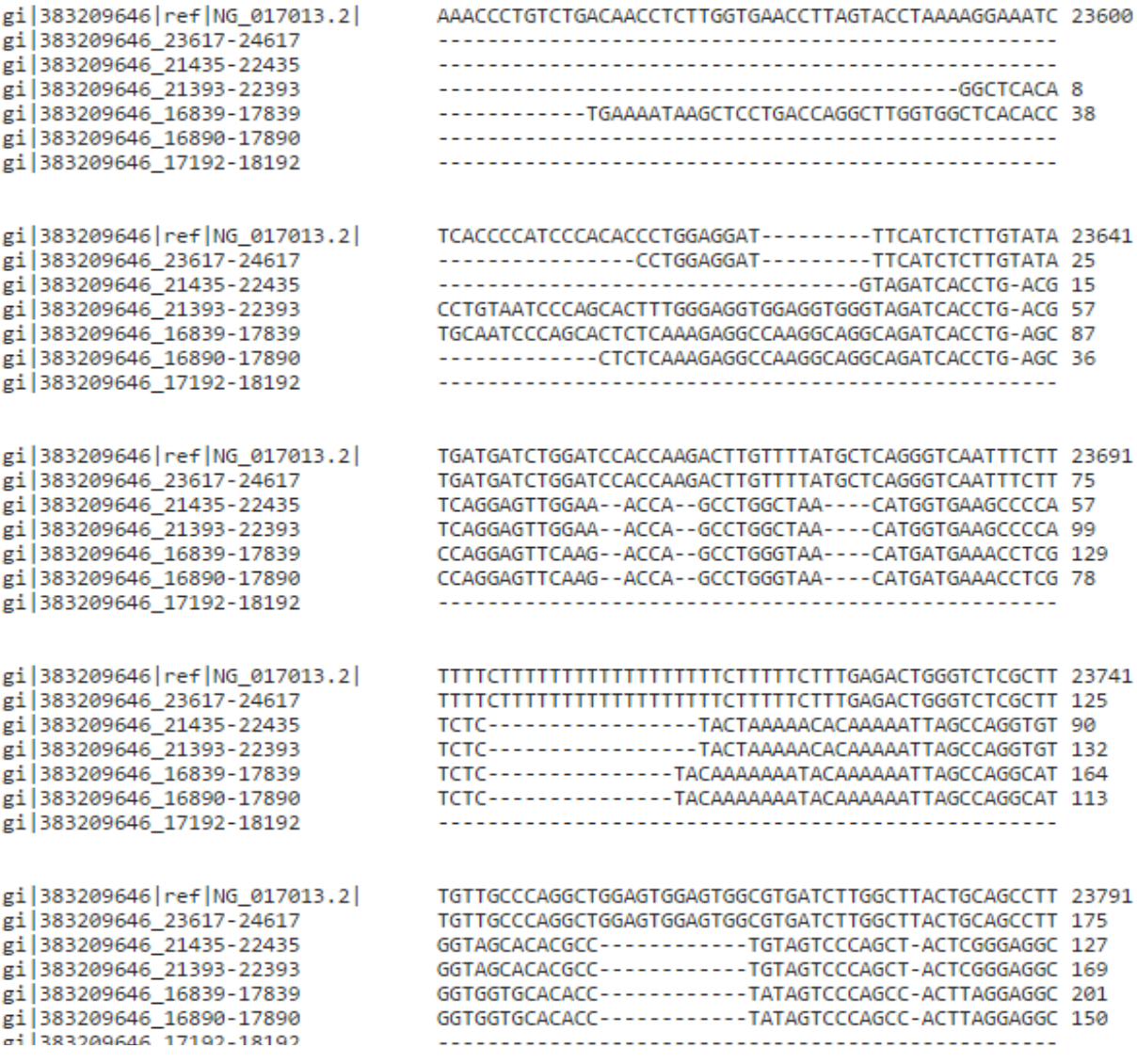
Results and Discussion

Nucleotide and amino acid sequences were cataloged and compared with the sequences of genetic variants (mutants) described in patients with proven diagnoses of skin cancer via the multiple sequence analysis tools ClustalW (RRID:SCR_017277), MAFFT (RRID:SCR_011811) and T-Coffee (RRID:SCR_011818). The original and mutant nucleotide and amino acid sequences of the p53 gene found in the NCBI and EBI genomic banks were subjected to analysis via the three tools mentioned above to evaluate their applicability, ease of study and feasibility in the analysis of each gene. Because these sequences are stored in genomic banks located in different countries, it is assumed that this implies different situations of environmental impact and contrasting differences in relation to ethnic groups, geographic location, latitude and the human development index (HDI), among other factors.

Sixty-two mutant amino acid sequences for the p53 gene were found in the EBI database, identified as coming from patients diagnosed with skin cancer, making it possible to consult these mutations by their respective phenotypes. To store the mutant sequences found in EBI, a new text file containing the original sequence of the p53 gene and the variant sequences inserted just below the original sequence was created.

Analysis of the results produced by the ClustalW (RRID:SCR_017277) tool made it possible to identify the regions of the p53 gene where the main mutations related to skin cancer occur. Figure 1 illustrates these results, where it is possible to visualize the regions where the mutations found in the EBI database for skin cancers are located.

Figure 1. Details of the analysis between the original sequence of the p53 gene and its corresponding mutants stored in the European Bioinformatics Institute (EBI) database via the ClustalW (RRID:SCR_017277) tool.



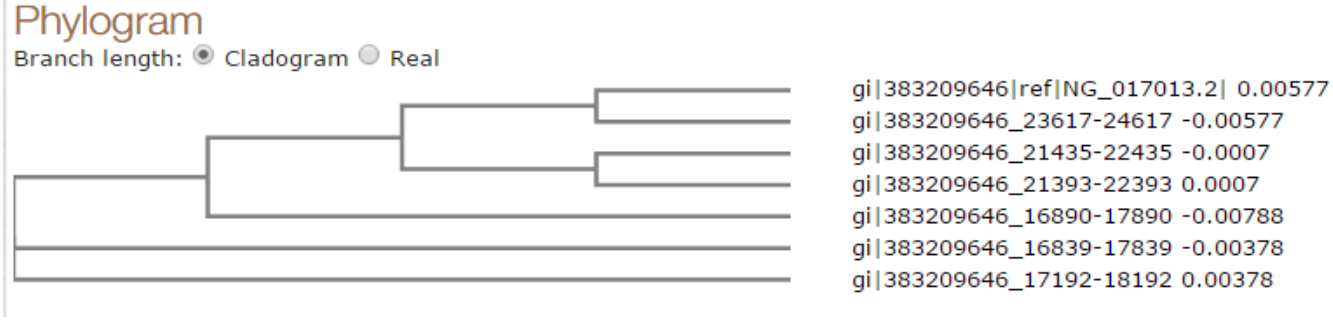
Available at: http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=CLUSTALW2-I20150209-114459-0326-41191369-pg (Accessed in August/2024)

In Figure 2, it is possible to visualize the sequence cladogram, also known as the phylogenetic tree, which shows the relationships between the original sequences and the analyzed mutations. This

cladogram allows inference of the degree of similarity between sequences, as the more distant they are from each other, the greater the divergences between the analyzed sequences.

In turn, the MAFFT (RRID:SCR_011811) tool presented results with a different layout than ClustalW (RRID:SCR_017277) tool. Initially, he presented the original sequence without mutation, and below, he presented the mutated sequences one after the other.

Figure 2. Display of the phylogenetic tree (phylogram) of the mutant sequences of the p53 gene, available on the web server of the European Bioinformatics Institute (EBI), assembled via the ClustalW (RRID:SCR_017277) tool.



Available at: http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=CLUSTALW2-I20150209-114459-0326-41191369-pg&analysis=tree (Accessed in August/2024).

The region of sequences that are outside the comparison area, that is, where one sequence is larger than the other, is represented by a sequence of characters identified by the term “gaps” or, in the free translation “spaces”. In addition to the cladogram display option, the MAFFT (RRID:SCR_011811) software also allows the display of a percentage identity matrix, as shown in Figure 3, where it is possible to verify the differences between the sequences inserted for analysis.

Figure 3. Identity percentage matrix generated by the MAFFT (RRID:SCR_011811) tool available on the European Bioinformatics Institute (EBI) web server.

1:	gi|383209646|ref|NG_017013.2|	100.00	100.00	100.00	100.00	97.50	94.01	93.11
2:	gi|383209646_16839-17839	100.00	100.00	100.00	100.00	57.63	75.31	75.79
3:	gi|383209646_16890-17890	100.00	100.00	100.00	100.00	57.63	75.31	74.05
4:	gi|383209646_17192-18192	100.00	100.00	100.00	100.00	57.63	-nan	-nan
5:	gi|383209646_23617-24617	97.50	57.63	57.63	57.63	100.00	-nan	-nan
6:	gi|383209646_21435-22435	94.01	75.31	75.31	-nan	-nan	100.00	100.00
7:	gi|383209646_21393-22393	93.11	75.79	74.05	-nan	-nan	100.00	100.00

Available at: http://www.ebi.ac.uk/Tools/services/rest/mafft/result/mafft-I20150209-120714-0961-63142518-oy/pim (Accessed in August/2024).

The numbers in the matrix indicate the degree of similarity between the sequences. Sequences identified with the number 100 indicate that the gene regions are identical to each other. The columns with '-nan' indicate the regions referring to the "gaps".

With respect to the data stored at NCBI, one hundred and ninety-four mutations were found in the p53 gene. All the mutant nucleotide and amino acid sequences available in this genomic bank, as well as the original nucleotide and amino acid sequences of the p53 gene, were copied. All these data were stored in a text file to be processed later with greater ease.

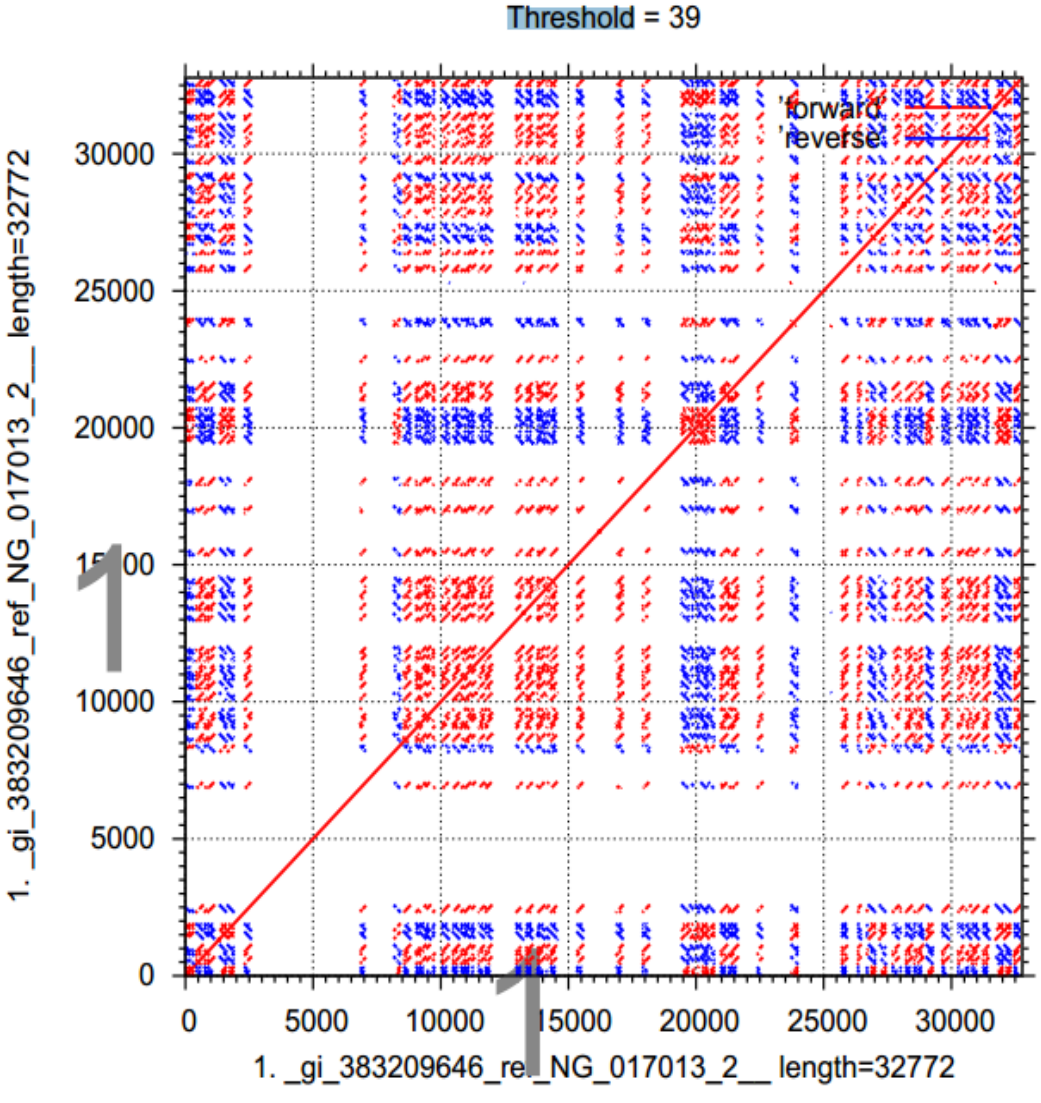
In the NCBI database, each mutant sequence provides access to dozens of pieces of information about the mutant gene, including references to publications on the mutation, the nucleotide sequence and its corresponding amino acid sequence, the location of the laboratory or institute where the genetic material was sequenced, the description of the mutation indicating the codon where the mutation occurs and the type of change, including change, insertion or deletion, among other information.

To visualize the mutations in a graphical format, sequence analyses were carried out via the MAFFT (RRID:SCR_011811) tool, which is stored on the web server of the Computational Biology Research Center (CBRC) located in Japan. This software has a very simple interface layout. Similar to that found on the EBI server but with a graphical resource, as shown in Figure 4. In this illustration,

the regions of the p53 gene containing the most frequently deposited mutations in the NCBI database are identified.

Figure 4 displays a rectangle filled with colored dots. The bordering positions of the rectangle represent the largest nucleotide sequence, which in this case is the wild-type sequence. A diagonal line crosses the image cutting through the region with the greatest number of points, which indicates the location where mutations occur, indicating a possibly fragile site. This form of data presentation is called the “threshold” (or edge of tolerance) of the mutant sequences in relation to the wild-type sequence.

Figure 4. ‘Threshold’ generated by the MAFFT (RRID:SCR_011811) tool available on the web server of the Computational Biology Research Center (CBRC) with data on mutant sequences for the p53 gene existing at the National Center for Biotechnology Information (NCBI).



Available at: http://mafft.cbrc.jp/alignment/server/spoolhar15020921309213h4NTV65b8KXIFlQeD2ZtU.pdf (Accessed in August/2024).

These results suggest that the mutations occurring in the p53 gene, which are derived from genetic material sequenced from patients from the United States and deposited at the NCBI Institute, are predominantly located in a very restricted gene region. Considering that the original nucleotide sequence of the p53 gene is composed of 32 thousand bases, the mutants were identified between positions 23562 and 24735. Analysis of the results obtained from the mutant sequences deposited at the EBI Institute allowed us to clearly identify the most frequent mutations occurring between positions 19844 and 21786 of the p53 gene.

The same sequences found in EBI were also analyzed via the T-Coffee (RRID:SCR_011818) tool (Notredame 2020), which is stored on another server with the domain www.tcoffee.org. This tool allows, among other functions, the evaluation of the type of mutation identified most frequently among the mutants evaluated, as summarized in Table 2.

Table 2. Identification of the most frequent types of mutations in the p53 gene found among the sequences cataloged in the European Bioinformatics Institute (EBI) database and related to skin cancer, analyzed via the T-Coffee (RRID:SCR_011818) tool.

rs12203592 - C/T	rs258322 - A/G	rs45430 - C/T	rs137854597 - C/T	rs149617956 - G/A
rs1805007 - C/G/T	rs4785763 - A/C	rs7023329 - A/G	rs121913388 - G/A	rs16891982 - C/G
rs12202284 - C/A	rs910873 - G/A	rs401681 - C/T	rs104894099 - A/C	rs121909233 - G/A
rs8015138 - A/C	rs1393350 - G/A	rs3219090 - T/C	rs104894109 - C/A	rs121909232 - C/A
rs113488022 - A/G/T/C	rs16953002 - G/A	rs228437 - C/T	rs104894095 - C/G/T	rs121909234 - G/A
rs137853080 - T/G	rs17119461 - T/C	rs35390 - C/A	rs104894097 - C/G	rs11547328 - G/A
rs137853081 - G/C	rs7412746 - C/T	rs1722784 - A/G	rs202187871 - C/T	rs104894098 - A/T
rs121913323 - C/T	rs13016963 - A/G	rs6001027 - A/G	rs104894340 - C/T	rs113798404 - C/A/G
rs121913315 - G/A/T	rs2284063 - A/G	rs36204594 - G/A	rs1805006 - C/A	rs104894094 - C/A
rs202187871 - C/T	rs1801516 - G/A	rs11552822 - C/A	rs861539 - G/A	rs137854599 - C/T

Source: The authors

The results made it possible to visualize the specific data of the mutant sequences, that is, the existing base substitutions identified between these sequences and the original sequence of the p53 gene. Table 2 shows the identification of the mutations in the p53 gene that occurred most frequently among the nucleotide sequences analyzed. There was no nucleotide deletion, but almost exclusively, base substitutions were detected.

The most frequently observed mutations were substitutions of cytosine for thymine and substitutions of guanine for adenine, both with the same occurrence rates, in accordance with what was recommended by Watson et al. (2006) of the predominance of the occurrence of transitions (purine-purine and pyrimidine-pyrimidine) compared with the occurrence of transversion in mutants for the p53 gene. To analyze the frequency of mutations, only the replacement of one nucleotide with another was considered, with other types of mutations not being considered.

In the EBI database, it was possible to analyze a greater number of mutant sequence variations in the p53 gene. Similar to the previous case, a total of 44 of the 50 mutant sequences were characterized by substitutions of nitrogenous bases, and among these, 30 were of the transition type and 14 were of the transversion type, that is, 68% transition and 31% transversion, similar to the NCBI database and in agreement with reports in the literature (Watson et al. 2006).

According to Santos Neto and Fluminhan (2014), observation of the data analyzed by the ClustalW (RRID:SCR_017277), MAFFT (RRID:SCR_011811) and T-Coffee (RRID:SCR_011818) tools can be very useful for identifying possible fragile sites in any gene. These tools also allow the analysis of the similarity of gene sequences (Malaman and Fluminhan 2014) and the establishment of phylogenetic relationships between the analyzed genomic sequences (Malaman and Fluminhan 2015).

Conclusion

The results allowed us to affirm that it is feasible to establish a genomic analysis strategy using the proposed tools and that other works can be developed and compared with data from various research institutes located in different parts of the planet, as recommended in the work of Prosdocimi (2007).

To transform the data from this research in a more dynamic way, future work may aim to integrate the mutant sequences into protein visualization software on the basis of their amino acid sequences. The construction of the P53 protein in a three-dimensional format will make it possible to analyze its active sites and correlate possible mutations with changes in its activity pattern.

The use of IT in the fields of life is essential for the interaction between sciences, thus providing better human and technological development. In recent years, Biology has eagerly appropriated the tools provided by IT. Thus, computing has become a key part of research being carried out in specialized areas such as Molecular Biology, and the storage of biological data in public databases has become increasingly common (Gibas and Jambeck 2001).

The present research demonstrated the multidisciplinary nature of these activities, increasing the relevance and productivity of bioinformatics analyses. In this context, the aim is to encourage preparations for significant insertion in the so-called postgenomic era to transform genomic information into scientific and technological knowledge.

In conclusion, the internet makes it possible for public databases of genome sequences to offer services through a uniform interface to a worldwide community of researchers. With a web browser, a molecular biologist can then compare an unknown DNA sequence to the complete collection of public DNA sequences. By making your results publicly available on networks, you can also contribute to other researchers, increasing the volume of information.

Acknowledgments

This work was supported by the National Council for Scientific and Technological Development / Brazilian Ministry of Science, Technology and Innovation (CNPq/MCTI) – Grant number 574.324/2008-9, Presidente Prudente – SP, Brazil.

References

Baxevanis AD, Ouellette BFF (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins – 3rd. Edition. John Wiley & Sons, Inc., New Jersey.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Research, v. 39, p. D32-D37 https://doi.org/10.1093/nar/gkq1079

Corapcioglu D, Sak SD, Delibasi T, Tonyukuk V, Kamel N, Uysal AR, Kocak S, Aydintug S, Erdogan G (2006) Papillary microcarcinomas of the thyroid gland and immunohistochemical analysis of expression of p53 protein in papillary microcarcinomas. Journal of Translational Medicine, v. 4, n. 28 https://doi.org/10.1186/1479-5876-4-28

Edwards D, Stajich J, Hasen D (2009) Bioinformatics: Tools and Applications. Springer, New York.

Efeyan A, Serrano M (2007) p53: guardian of the genome and policeman of the oncogenes. Cell Cycle, v. 6, p. 1006-1010 https://doi.org/10.4161/cc.6.9.4211

Frebourg T, Kassel J, Lam KT, Gryka MA, Barbier N, Andersen TI, Børresen AL, Friend SH (1992) Germ-line mutations of the p53 tumor suppressor gene in patients with high risk for cancer inactivate the p53 protein. Proceedings of the National Academy of Science of the United States of America, v. 89, p. 6413-6417 https://doi.org/10.1073/pnas.89.14.6413

Gibas C, Jambeck P (2001) Developing bioinformatics computer skills. O`Reilly Media Inc., New York.

Hainaut P, Hollstein M (1999) p53 and human cancer: the first ten thousand mutations. Advances in

Cancer Research, v. 77, p. 81–86, 86a, 87-137 https://doi.org/10.1016/S0065-230X(08)60785-X

INCA - Instituto Nacional de Câncer (BRASIL). Coordenação de prevenção e vigilância. Estimativa 2020: Incidência de câncer no Brasil. Rio de Janeiro: INCA, 2020. Available at: https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/estimativa-2020-incidenciade-cancer-no-brasil.pdf (Accessed in August/2024).

Malaman ACP, Fluminhan A (2015) Análise filogenética sistematizada de sequências gênicas de compostos antioxidantes encontrados em espécies vegetais. Colloquium Vitae, v. 7, p. 1-14 https://doi.org/10.5747/cv.2015.v07.n3.v139

Malaman ACP, Sevilha TL, Fluminhan A (2014) Análise da similaridade de sequências gênicas do ascorbato e do licopeno em espécies vegetais cultivadas. Colloquium Exactarum, v. 6, p. 9-20 https://doi.org/10.5747/ce.2014.v06.n4.e095

Martinez MAR, Francisco G, Cabral LS, Ruiz IRG, Neto CF (2006) Genética molecular aplicada ao câncer cutâneo não melanoma, Anais Brasileiros de Dermatologia, v. 81, p. 405-19. Available at: http://www.scielo.br/pdf/abd/v81n5/v81n05a03.pdf (Accessed in November/2024).

Morita N, Ikeda Y, Takami H (2008) Clinical significance of P53 protein expression in papillary thyroid carcinoma. World Journal of Surgery, v. 32, p. 2617-2622 https://doi.org/10.1007/s00268-008-9756-9

Notredame C (2020) T-Coffee: Multiple Sequence Alignment Tools. Center for Genomic Regulation (CRG), Barcelona. Available at: https://tcoffee.crg.eu/apps/tcoffee/do:regular (Accessed in November/2024).

Parkin DM (2004) International variation. Oncogene, v. 23, p. 6329-6340 https://doi.org/10.1038/sj.onc.1207726

Prosdocimi F (2007) Curso On Line: Introdução à Bioinformática: Biotecnologia, Ciência & Desenvolvimento. Brasília: Portal Biotecnologia. Available at: http://www.iq.usp.br/setubal/bmc/2015/FProsdocimi07_CursoBioinfo.pdf (Accessed in August/2024).

Sansom CE, Smith CA (2000) Computer applications in biomolecular sciences. Part 2: bioinformatics

and genome projects. Biochemical Education, v. 28, p. 127-131 https://doi.org/10.1111/j.1539-3429.2000.tb00043.x

Santos Neto AA, Fluminhan A (2014) Implantação de metodologias para análise comparativa de sequências genômicas relacionadas ao câncer de pele. Colloquium Exactarum, v. 6, p. 1–9 https://doi.org/10.5747/ce.2014.v06.n3.e084

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, v. 22, p. 4673-4680 https://doi.org/10.1093/nar/22.22.4673

Varley JM, Attwooll C, White G, McGown G, Thorncroft M, Kelsey AM, Greaves M, Boyle J, Birch JM (2001) Characterization of germline TP53 splicing mutations and their genetic and functional analysis. Oncogene, v. 20, p. 2647-2654 https://doi.org/10.1038/sj.onc.1204369

Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R (2006) Biologia molecular do gene. Artmed, Porto Alegre,