UNFORMATED PREPRINT
Published in: Revista Tecnologia da Informação, v. 3, n. 2, p. 87-89, dez. 2003

Extracting knowledge from the p53 mutation database: the influence of tabagism in the somatic mutations profile of lung cancer subjects

Heitor S. Lopes [1], Adilson Alves Dias [2]

[1] Lab. de Bioinformática / CPGEI / Centro Fed. de Educ. Tecnol. do Paraná - CEFET-PR,
Av. 7 de setembro, 3165 80230-901 Curitiba (PR) – Brazil
[2] Depto. de Ciências Fisiológicas / Faculdade de Ciências Médicas da Santa Casa de SP
Av. Doutor Cesário Mota Junior, 61 – Vila Buarque 01221-906 São Paulo (SP) – Brazil
hslopes@cefetpr.br adilson@lpcc.org.br

Abstract. This work analyses a p53 mutation database aiming to unveil whether of not tabagism can influence the mutation profile of lung cancer patients. We analyzed both the frequency of single nucleotide mutations and the frequency of site (exon) mutations. For the database used, we conclude that tabagism produces more dramatic and unexpected mutations in the DNA sequence of p53.

1 Introduction

The strong correlation between tabagism and the occurrence of lung cancer is well known in the medical literature. According to the World Health Organization - WHO[1], tabagism is the leading cause of preventable death in the world today. Nowadays there are 4.9 million tobacco-related deaths per year, and no other consumer product is as dangerous, or kills as many people, as tobacco. Tabagism is present in more than 90% of cases of lung cancer. Tobacco smoke contains more than 4,000 chemical components, including over 50 known human carcinogens. The carcinogenic effect of the chemical components present in the industrialized tobacco (Nicotina tabacum) in the pulmonary tissues as well as in other parts of the human body is another issue also widely studied (Mendelsohn et al 1995).

Lung cancer presents one of the largest mortality indexes among the human cancers. In spite of huge research efforts throughout the world, there is no therapy available yet to change this picture. Therefore, it is important the comprehension of the mechanisms of the lung cancer etiology so as to better understand the early diagnosis and treatment of such disease.

The human p53 gene (GenBank accession number: U94788) is located in chromosome 17, locus 17p13.1 and its DNA chain has 20303 base pairs. The p53 gene is composed by 10 introns and 11 exons, and its product is a protein of the same name. This gene is known as tumor suppressor gene due to its role in the negative control of

1 http://www.who.int/features/2003/08/en/

the cellular cycle. The protein p53 inhibits the ciclin/ciclin-kinasis-dependent complex and stops the cellular cycle at the end of G1 phase (this point is denominated checkpoint G1-S), where the DNA is repaired in order to avoid the transmission of mutations to the next generation. When the p53 is inactivated due to mutations or formation of complexes with viral proteins (oncoproteins E6 and E7 of HPV), its regulation function is lost and the cell accumulates genetic errors, which are transmitted to the next generations. This is the start of the patophysiologic process of most solid tumors in adults.

The International Agency for Research on Cancer - IARC (Lyon, France) is part of the WHO and has the mission is to coordinate and conduct research on the causes of human cancer, the mechanisms of carcinogenesis, and to develop scientific strategies for cancer control. Among the several sources of information maintained by IARC, there is a database of mutations of p53 gene related to cancer (Olivier et al., 2002). This database contains, in separated files, information about somatic mutations and germinal mutations of the p53 for patients with cancer, as well as for other patients without this disease.

There are some contradictory reports in recent literature regarding the mutation profile of the p53 gene in smoker and non-smoker patients with lung cancer, (Campling and El-Deiry, 2003; Hainault and Pfeifer, 2001; Paschke, 2000; Zalcman and Soussi, 2000; Hernandez-Boussard and Hainaut, 1998), and other types of cancer (Pfeifer et al 2002). Therefore, the objective of this work is to evaluate the hypothesis that tabagism can influence the mutation profile of the somatic mutations on the exons of the p53 gene.

2 Material and methods

In this work we used the IARC p53 somatic mutation database, version R7, which contains 17232 entries. Actually, this is not a real database, but a data collection extracted from publications. Therefore, this collection lacks consistence and, despite the efforts to cure it, many errors still remain. Hence, the first step in using this database is to purge (clearly) wrong registers and fix (evident) errors. This hard work was done manually using a MS Excel spreadsheet. Additionally, a C++ program was developed to search all possible values for each column attribute, as a helper to the manual work. After this preliminary step, we extracted out all records with lung cancer, following the WHO's International Classification of Diseases for Oncology - ICD[2] code C34, thus yielding 2003 cases. From this set, 57 cases of intron mutations were excluded, as well as 1046 cases where the information about smoking status of the patient was missing. The resulting data were divided into two classes: smokers (henceforth, S) with 673 cases, and non-smokers (henceforth, NS) with 227 cases.

Groups S and NS were compared each other to evaluate the following: frequency of the different possible mutation types that may occur in the DNA chain; frequency of location of mutation (exon); frequency of the mutated amino acids after translation.

[2] http://www.cog.ufl.edu/publ/apps/icdo/icdo_top.txt

Besides, we compared the group of 1946 lung cancer patients (excluding intron mutation cases) with the group of all other types of cancer (again excluding intron mutation cases). The objective of this comparison was to analyze whether or not tabagism can influence the expected profile of lung cancer mutations.

The statistical analysis aimed to compare the significance of sample proportions of the groups under study. Therefore, Fisher test was used for this purpose using the software Graphpad Instat version 3. For all tests, the rejection level of the null hypothesis (no significance) was fixed in 0.05 (5%).

3 Results

As mentioned before, intron mutations were not considered in this work, since most of them are splice or unspecified mutations, and they represent less than 3% of the lung cancer cases in the database.

In the first study we compared the number of events and the corresponding frequency within each group, of the several possible single mutations of the DNA chain. Comparing lung cancer subjects with all other types of cancer in the database, two types of mutations were significantly different ($p < 0.05$): G_to_A (respectively, 13.6% for lung cancer and 26.5% for other cancers) and G_to_T (respectively, 26.6% for lung cancer and 10.0% for other cancers). Comparing smoker and non-smoker groups of lung cancer subjects, we found four types of mutations significantly different: A_to_T (3.3% for S and 2.2% for NS), C_to_T (13.2% for S and 18.1% for NS), G_to_A (13.2% for S and 17.6% for NS), del (9.1% for S and 4.0% for NS).

In the second study we compared the frequency of the mutation site (exon) for the same groups as above. For lung cancer subjects versus all other type of cancers, there were no statistically significant differences ($p < 0.05$). For smokers and non-smokers groups there were two statistically significant differences: in exon 5 (33.6% for S and 23.8% for NS) and in exon 8 (23.6% for S and 34.8% for NS).

4 Discussion and conclusions

Medical and biological databases, especially those available in the Internet, many times are more an unstructured collection of data than a database by itself. This is really the case of the IARC p53 somatic mutation database used in this work. Although it is the largest available collection of information about mutation cases of the p53, it still has many missing and wrong data. This is explained by the way data is collected and, mainly, by the fact that it was not designed, since the beginning, to be a real database. Therefore, extracting knowledge from this database is a real challenge and request much manual work, before being processed by a data-mining tool.

Deamination of nucleotides (the removal of an amino group from the molecule) is a common event in pathological processes and in the evolution of the human genome, and C-T and G-A transitions (exchanges a purine for a purine or a pyrimidine for a

pyrimidine) are the most common (Krawczak and Cooper, 1996). Non-smokers presented a larger number of substitutions than smokers ($p < 0.05$), what is compatible with the natural tendency for deamination.

In lung cancer, G-T transversions (which exchanges a purine for a pyrimidine or a pyrimidine for a purine) predominate. This kind of mutation is infrequent in other types of cancers, except hepatobiliar carcinoma (Pfeifer et al 2002; Hainault and Pfeifer, 2001), and it is considered the "mutation signature" of lung cancer compared with other cancers. This statistically significant difference was confirmed here ($p < 0.05$). Tabagism did not affect this proportion ($p > 0.05$), suggesting that such transversion is a consequence of some intrinsic factor of the pulmonary microenvironment.

Smokers presented more A-T transversions and more deletions than non-smokers ($p < 0.05$). These events are rare (considering the pathology-related mutations as a whole), they request more energy to occur and they consequences are more dramatic.

Regarding the distribution of mutations by location (exons), there were no significant differences between lung cancer patients and other types of cancer. Smokers presented more mutations in exon 5 ($p < 0.05$) and non-smokers presented more mutations in exon 8 ($p < 0.05$). It was not possible to establish the biological importance of this fact up to now.

As final conclusion, tabagism clearly produces more dramatic and unexpected mutations in the DNA sequence of p53 gene.

References

1. Campling, B.G., El-Deiry, W.S.: Clinical implication of p53 mutation in lung cancer. Mol. Biotech. 4:2 (2003) 141-156
2. Hainault, P., Pfeifer, G.P.: Patterns of p53 G-T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis 22:3 (2001) 367-374
3. Hernandez-Boussard, T.M., Hainaut, P.: A specific spectrum of p53 mutations in lung cancer from smokers: review of mutations compiled in the IARC p53 database. Environ. Health Persp. 106:7 (1998) 385-391
4. Krawczak, M., Cooper, D.N.: Single base-pair substitutions in pathology and evolution: two sides of the same coin. Hum. Mutat. 8:1 (1996) 23-31
5. Mendelsohn, J., Hawly, P.M., Israel, M.A. and Liotta, L.A.: Molecular Basis of Cancer. WB Saunders Companhy, Philadelphia (1995)
6. Olivier, M., Eeles, R., Hollstein, M., Khan, M.A., Harris, C.C., Hainaut, P.: The IARC Tp53 database: new online mutation analysis and recommendations to users. Hum. Mutat. 19:2 (2002) 607-614
7. Paschke, T.: Analysis of different versions of the IARC p53 database with respect to G->T transversion mutation frequencies and mutation hotspots in lung cancer of smokers and non-smokers. Mutagenesis 15: 6 (2000) 457-458
8. Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., Hainaut, P.: Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 21:48 (2002) 7435-7451
9. Zalcman, G., Soussi, T.: Alterations in p53 tumor suppressor gene in lung cancer. Rev. Mal. Respir. 17:1 (suppl.) (2000) 329-349