FEBS 17681

FEBS Letters 396 (1996) 119–122

# Sequence specificity in CpG mutation hotspots

Juha Ollila[1], Ilkka Lappalainen, Mauno Vihinen*

*Department of Biosciences, Division of Biochemistry, P.O. Box 56, FIN-00014 University of Helsinki, Helsinki, Finland*

Received 29 August 1996

**Abstract CpG dinucleotides are efficiently methylated in vertebrate genomes except in the CpG islands having a high C+G content. Methylated CpGs are the single most mutated dinucleotide. Sequences surrounding disease causing CpG mutation sites were analyzed from locus-specific mutation databases. Both tetra- and heptanucleotide analyses indicated clear overall sequence preference for having pyrimidines 5′ and purines 3′ to the mutated 5-methylcytosine. The most mutated tetranucleotides are TCGA and TCGG, the former being also a frequent restriction and modification site. The results will help in elucidating the still controversial mutation mechanism of CpG doublets.**

*Key words:* CpG dinucleotide; CpG suppression; DNA methylation; Human mutation database

## 1. Introduction

Vertebrate genomes have a marked overall deficiency of the dinucleotide CpG. Most, but not all, cytosines in CpG dinucleotides of the mammalian genome carry a methyl group added by methyltransferase. Methylated CpG dinucleotides are known mutational hotspots (for review see [1]). Methylation of DNA affects transcription of several genes by preventing transcription factors and other proteins from binding. CpG sites are also frequently mutated in many cancer types [2]. Genome methylation patterns change at different development stages. Although the number of CpG dinucleotides is remarkably depleted, about one third of human point mutations appear in these sites [1]. Spontaneous deamination of 5-methylcytosine (m5C) to thymine leading to TpG and CpA mutations has been suggested to be the reaction mechanism in CpG alterations [3].

Although CpG mutations constitute the single most frequent mutational effect, the details of the mechanism are still unclear. To study the requirements on the DNA we have performed an extensive analysis of human CpG doublets and flanking nucleotides in hereditary disorders. We found clear sequence specificity indicating preference for pyrimidine bases 5′ and purines 3′ to the mutated cytosine. The results will prove important in elucidating structural mechanisms of human diseases and CpG suppression.

## 2. Methods

The mutations were taken from locus-specific mutation databases: BTKbase (version 2) for X-linked agammaglobulinemia [4,5], hemophilia A database HAMSTeRS [6], somatic point mutations in p53 [7], and PAHdb for phenylketonuria [8]. The sequences were taken from Genbank. The accession numbers of the sequences are X58957, X01179, X02469, and L47800, respectively. GCG program package [9] was used for sequence analysis.

The mutations were analyzed by comparing frequencies of the obtained mutations to those suggested based on base composition by using the $\chi^2$ test to determine the significance of the results. The formula for the test is:

$$\chi^2 = (O-E)^2/E$$

where O denotes the observed frequency and E is the expected frequency. The expectations were calculated by multiplying the frequencies of mononucleotides.

## 3. Results and discussion

CpG doublet is present only at about 20% of its expected frequency in vertebrate genomes. The mutated CpG sites were studied in four hereditary disorders. We analyzed registries for Bruton's tyrosine kinase (BTK) [4,5], factor VIII [6], tumor suppressor p53 [7], and phenylalanine hydroxylase (PAH) [8], which were chosen based on database size, format and availability. General data for the analyzed genes and defects are in Table 1. The CpG dinucleotides are remarkably suppressed when compared to expected frequencies (Table 1). The highest number of CpG sites is in factor VIII gene, where there are 45% of the expected number. In the three other genes less than one third of the calculated sites is present.

The mutations occur in cytosine C in either the coding or noncoding strand giving rise to CpG→TpG and CpA transitions. The analyzed sequence stretches were taken from the mutated strand. Altogether 1919 mutations in 94 CpG doublets were analyzed (Table 1). Only one patient was chosen to represent each affected family to avoid bias in calculations. The C+G contents (43–57%) and CpG frequencies (1.0–3.6%) (Table 1) are typical for human genes. Despite the rarity of the CpG doublets, CpG mutations constitute at least one third of missense and nonsense mutations in these defects. The ratios of CpG mutations in the coding vs. noncoding strand have a normal distribution.

### *3.1. NCGN sequences*

Analysis of the bases flanking CpG sites indicated that mutations appear in all the tetranucleotides, although at highly variable frequencies (Tables 2 and 3). The most abundant site in all four deficiencies is YCGR (Y denotes pyrimidine and R purine). The significance of the distributions was estimated with the $\chi^2$ test by comparing the observed mutations to those expected based on nucleotide composition. Of the tetranucleotides, only YCGR is significantly mutated in all four databases (Table 2). The other NCGN sequences are significantly mutated in some disorders, but they are also significantly underrepresented in some other disorders. Some tetranucleotides are even clearly underrepresented in mutations, such as

*Corresponding author. Fax: (358) (9) 708 59068.
E-mail: mauno.vihinen@helsinki.fi

[1]The first two authors contributed equally to the work.


*PII* S0014-5793(96)01075-7

Table 1
CpG mutations in human disorders

| | BTK | FVIII | p53 | PAH |
|---|---|---|---|---|
| C+G (%) | 46.8 | 42.6 | 56.9 | 47.3 |
| CpG (%) | 1.7 | 1.0 | 3.6 | 1.7 |
| CpG (*n*) | 33 | 71 | 42 | 23 |
| Mutations (total) | 269 | 541 | 4499 | 1162 |
| Missense+nonsense (*n*) | 124 | 396 | 3770 | 624 |
| CpG mutations (*n*) | 43 | 201 | 1437 | 238 |
| CpG sites mutated (%) | 42.4 | 43.6 | 81.0 | 65.2 |
| CpG mutations of missense and nonsense mutations (%) | 34.7 | 50.8 | 38.1 | 38.1 |
| Ratio (coding vs. noncoding strand) | 1.15 | 1.83 | 0.61 | 0.89 |

*n* denotes number.

ACGG. The results propose that the preceding base and maybe even the preceding triplet might be somehow correlated to CpG dinucleotides in mutational hotspots.

The YCGR sequences are mutated in at least 60% of their occurrence in the four studied genes, whereas the other tetranucleotides have on average markedly fewer mutations (Table 4). It is the most mutated tetranucleotide in all the studied disorders. The overall mutability of the sites is highest in the p53 database, presumably because of the large number of total mutations, altogether 4499 patients with 1437 CpG transitions. The most abundant sequences in the mutations in all four diseases are TCGA and TCGG, whereas the other two YCGR sequences, CCGA and CCGG, display a more random distribution of mutations (Table 3). None of the other tetranucleotides is significantly mutated in all the studied databases. It is interesting that the flanking bases in the inverted order, RCGY, contain mutations only in every fourth site (Table 4). Despite p53, mutations in RCGY tetranucleotides are rare or nonexistent except for some single hotspots (Table 3).

### 3.2. NNCGNN sequences

Because of the clear sequence preference in the CpG containing tetrapeptides it was tempting to analyze longer sequence stretches as well, to see whether sequence effects were more extensive. This was interesting also because of the suggested sequence periodicity around CpG sites [10]. Analysis of the mutated CpG containing heptapeptides indicated pronounced sequence specificity for YYCGRR and especially YYCGRY, which both contain the central YCGR pattern (Table 5). Both the tetra- and heptanucleotide studies indicate obvious sequence preference for neighbors in CpG mutations. The mixed occurrence of purines and pyrimidines on one or both sides is very rare, and often actually underrepresented. The overrepresentation at YRCGYR in p53 is due to a single hotspot. Of the symmetric sequences YYCGYY is significantly overrepresented in two diseases, but has not been identified at all in one, while in p53 it is significantly underrepresented. The largest registry, p53 database, contains a significant number of mutations also in the YRCGYR motif (Table 5). In Btk, only the sequences containing the central YCGR motif are significantly mutated, while in factor VIII most of the mutations appear in the central YCGR, with some other frequently diagnosed sites. Also heptanucleotides tend to have purines on one side and pyrimidines on the other side of the most mutated CpG dinucleotides. This could mean that these stretches are recognized or ignored by some mutation mechanisms or repair systems, respectively.

It would have been of interest to study even longer CpG containing sequences but the size of databases does not warrant statistical estimations.

### 3.3. CNG sequences

CpNpG sites have been shown to be stably methylated in mammalian cells at low frequency [11]. To see if mutation patterns are affected by the central base in this triplet the distribution of mutations was analyzed. No sequence preferences were noticed when mutations in CpG dinucleotides were omitted. The incidence of CNG methylation outside CpG is low and it does not affect mutation rates, at least not to a degree detectable in the statistical test.

Table 2
CpG mutations in NCGN sequences

| NCGN | BTK | | FVIII | | p53 | | PAH | |
|---|---|---|---|---|---|---|---|---|
| | *n* | $\chi^2$ | *n* | $\chi^2$ | *n* | $\chi^2$ | *n* | $\chi^2$ |
| RCGR | 8 | 0.007 | 54 | **57.5**$^{***}$ | 86 | *143.6*$^{***}$ | 28 | 1.11 |
| RCGY | 9 | 0.88 | 8 | *5.50*$^{*}$ | 666 | **448.0**$^{***}$ | 3 | *29.0*$^{***}$ |
| YCGR | 26 | **57.0**$^{***}$ | 95 | **331.1**$^{***}$ | 544 | **302.1**$^{***}$ | 164 | **480.2**$^{***}$ |
| YCGY | 0 | *5.60*$^{*}$ | 46 | **55.9**$^{***}$ | 140 | *92.8*$^{***}$ | 43 | 1.63 |

Bold letters indicate overrepresentation and italics underrepresentation. *n* gives the number of sites.
The results of the $\chi^2$ test are shown with significance level: $^{*}P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$.

Table 3
CpG mutations in tetranucleotides

| NCGN | BTK | FVIII | p53 | PAH |
|---|---|---|---|---|
| ACGA | 0.19 | **13.3**$^{***}$ | *63.2*$^{***}$ | 0.47 |
| ACGC | 0.86 | 1.97 | **430.**$^{***}$ | |
| ACGG | | 3.64 | *39.3*$^{***}$ | 0.72 |
| ACGT | | | *51.3*$^{***}$ | 4.59 |
| CCGA | **5.92**$^{*}$ | 3.44 | *14.8*$^{***}$ | **29.8**$^{***}$ |
| CCGC | | **48.3**$^{***}$ | *66.3*$^{***}$ | 1.47 |
| CCGG | **4.48**$^{*}$ | | **1042.**$^{***}$ | |
| CCGT | | 0.0001 | *60.2*$^{***}$ | **50.6**$^{***}$ |
| GCGA | 2.25 | **47.9**$^{***}$ | *55.5*$^{***}$ | 0.26 |
| GCGC | **8.69**$^{**}$ | | **87.4**$^{***}$ | |
| GCGG | | **7.13**$^{**}$ | *7.31*$^{**}$ | 0.0002 |
| GCGT | 0.18 | | **50.6**$^{***}$ | |
| TCGA | **35.8**$^{***}$ | **839.9**$^{***}$ | **7.79**$^{**}$ | *3.32* |
| TCGC | | **81.6**$^{***}$ | | 3.86 |
| TCGG | **19.7**$^{***}$ | **9.48**$^{**}$ | *53.5*$^{***}$ | **1976.**$^{***}$ |
| TCGT | 0.15 | 0.12 | *17.7*$^{***}$ | 1.89 |

The numbers are from the $\chi^2$ test.
The significance is as indicated in Table 2.

### 3.4. Mutations in CpG island

CpG islands are C+G rich regions having a high CpG frequency. These stretches are clearly undermethylated. CpG islands usually appear outside protein coding regions in promoter regions. The islands are more common among housekeeping genes. The gene involved in protein C deficiency (entry M11228) contains an island of about 320 nucleotides in the middle of the coding region. In the analysis of the database [12] containing 163 CpG mutations, apparently different results were obtained compared to those in Tables 2–4. The C+G content is also higher, 62%. Most of the CpG mutations are in sites other than YCGR, which in fact does not contain a significant number of mutations. Of the nonsense and missense mutations altogether 67% are in CpG dinucleotides. However, only three CpG sites were found outside the CpG island, and none of them is mutated. Thus, the appearance of a CpG island has a profound effect on the mutation pattern which is expected due to biased base composition. It remains to be seen if also the CpG island mutations are due to methylation, which seems likely.

### 3.5. Suppression of CpG containing sequences

CpG dinucleotides have been shown to have a periodicity of eight nucleotides in purines in both CpG-poor and -rich regions of the human genome [10]. CCGG sequences were noticed to be highly represented on CpG-poor coding regions [10].

High mutation rates at YCGR sequences, especially TCGA (Table 3), would during evolution lead to their suppression from coding genes. To test this idea, we analyzed the distribution of tetranucleotides in 99 randomly selected human mRNA sequences containing 172 507 bases and found that CCGG is indeed clearly overrepresented ($\chi^2$=82.1***), consistent with the previous results [10]. TCGG also appears more abundantly than expected (12.4***), CCGA has the predicted frequency, whereas the highly mutated TCGA is significantly suppressed (11.5***). The distribution data show at least indirectly the importance of nucleotides flanking mutated CpG dinucleotides as well as a relation to the observed mutabilities.

### 3.6. Comparison to other studies

Related observations have been made when studying the rate of base substitution in 40 mammalian processed pseudogenes [13]. Relative mutation rates were the highest for YCGR stretches in both strands in CpG containing tetranucleotides.

Cooper and Krawczak [1] have presented that there is no preference for 5′ nucleotide in human CpG mutations. Their assumption was based on analysis of mutations in the first position of codons. The preceding base was suggested to have random distribution of the third position base. However, when their data were analyzed at the sequence level we found a most significant bias in the mutated tetranucleotides. The YCGR sequence is clearly overrepresented ($\chi^2$=11.6***) in mutations, while RCGY was underrepresented (3.74). The most mutated tetranucleotide sequence was TCGA (29.0***), which alone constitutes 18% of the mutations in the 16 CpG tetranucleotides. Thus, also their data support our findings.

Table 4
Percentage of mutated NCGN sites

| NCGN | Btk | FVIII | p53 | PAH |
|---|---|---|---|---|
| RCGR | 42.9 | 42.1 | 77.8 | 66.7 |
| RCGY | 25 | 12.5 | 91.7 | 25 |
| YCGR | 80 | 61.1 | 90 | 100 |
| YCGY | 0 | 46.2 | 63.6 | 57.1 |

Table 5
CpG mutations in NNCGNN sequences

| NNCGNN | BTK | VIII | p53 | PAH |
|---|---|---|---|---|
| RRCGRR | | **25.5***** | *56.3**** | 0.926 |
| RRCGRY | 0.44 | 0.034 | *102.**** | |
| RRCGYR | | | *13.5**** | |
| RRCGYY | **17.6***** | 0.25 | **140.***** | *7.73*** |
| **RYCGRR** | 3.38 | **80.1***** | **10.5**** | 0.79 |
| **RYCGRY** | | | 1.57 | |
| RYCGYR | | | *111.**** | 3.80 |
| RYCGYY | | 4.17* | 5.17* | |
| YRCGRR | 0.005 | **9.66**** | *37.3**** | 0.26 |
| YRCGRY | 0.86 | 0.013 | *138.**** | *11.1**** |
| YRCGYR | | | **298.***** | |
| YRCGYY | | | *133.**** | |
| **YYCGRR** | **23.3***** | **34.9***** | *45.9**** | **466.***** |
| **YYCGRY** | **30.2***** | **204.2***** | **18.5***** | **144.1***** |
| YYCGYR | | **96.0***** | *114.**** | |
| YYCGYY | | **17.3***** | *144.**** | **40.8***** |

The numbers are from the $\chi^2$ test.
The significance is as indicated in Table 2.
The central YCGR sequence is indicated in bold letters.

### 3.7. Conclusions

The results seem to indicate that CpG dinucleotides, although appearing in all combinations of flanking bases, are much more prone to mutations when there are one or more pyrimidines 5′ and one or more purines downstream of the mutated cytosine. It is interesting that the most frequent mutation site, TCGA, has a palindromic sequence, which might have functional relevance. TCGA is also a common DNA restriction and methylation site. Spontaneous deamination is believed to be the mutation mechanism in 5mC sites [3]. Other possible ways are thought to be deficient mutation repair mechanisms, and recently cytosine methyltransferase has been shown to be able to convert 5mC to thymine [14]. Also the structure of the DNA might have some local alterations (e.g. in grooves) in the presence of one or two 5mCs in complementary strands and thereby affect e.g. protein binding. It remains to be seen what is the general mechanism, if any, but the inverted repeat of the most frequent mutation sequence could imply importance of protein-DNA interactions.

*Acknowledgements:* Financial support from Biocentrum Helsinki and the European BIOMED concerted action 'PL1321' is gratefully acknowledged.

## References

[1] Cooper, D.N. and Krawczak, M. (1993) Human Gene Mutation, Bios Science, Eynsham.
[2] Jones, P.A., Rideout, W.M., Shen, J.-C., Spruck, C.H. and Tsai, Y.C. (1992) BioEssays 14, 33–36.
[3] Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Nature 274, 775–780.
[4] Vihinen, M., Cooper, M.D., de Saint Basile, G., Fischer, A., Good, R.A., Hendriks, R.W., Kinnon, C., Kwan, S.-P., Litman, G.W., Notarangelo, L.D., Ochs, H.D., Rosen, F.S., Vetrie, D., Webster, A.D.B., Zegers, B.J.M. and Smith, C.I.E. (1995) Immunol. Today 16, 460–465.

[5] Vihinen, M., Iwata, T., Kinnon, C., Kwan, S.-P., Ochs, H.D., Vořechovský, I. and Smith, C.I.E. (1996) Nucleic Acids Res. 24, 160–165.
[6] Wacey, A.I., Kemball-Cook, G., Kazazian, H.H., Antonarakis, S.E., Schwaab, R., Lindley, P. and Tuddenham, E.G.D. (1996) Nucleic Acids Res. 24, 100–102.
[7] Hollstein, M., Shomer, B., Greenblatt, M., Soussi, T., Hovig, E., Montesano, R. and Harris, C.C. (1996) Nucleic Acids Res. 24, 141–146.
[8] Hoang, L., Byck, S., Prevost, L. and Scriver, C.R. (1996) Nucleic Acids Res. 24, 127–131.
[9] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[10] Clay, O., Schaffner, W. and Matsuo, K. (1995) Somat. Cell Mol. Genet. 21, 91–98.
[11] Clark, S.J., Harrison, J. and Frommer, M. (1996) Nature Genet. 10, 20–27.
[12] Reitsma, P.H. (1996) Nucleic Acids Res. 24, 157–159.
[13] Bains, W. and Bains, J. (1987) Mutat. Res. 179, 65–74.
[14] Yebra, M.J. and Bhagwat, A.S. (1995) Biochem. 34, 14752–14757.