Biophysical Chemistry, 37 (1990) 25–29
Elsevier

BIOCHE 01462

Frequency of abnormal human haemoglobins caused by C → T transitions in CpG dinucleotides *

M.F. Perutz

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

Received 29 January 1990
Accepted 1 February 1990

CpG; Mutation rate; Globin gene; Methylation

A large part of human genetic disease apparently arises from deamination of cytosines in methylated CpG dinucleotides. Their mutation rate is known to be high when C is present as 5-methylcytosine, but is believed to be normal when it is unmethylated. The β-globin gene contains five, the γ-globin gene two, and each of the α-globin genes contain 35 CpGs. The CpGs in the β- and γ-globin genes are methylated, while those in the α-globin genes are undermethylated. One would therefore have expected the CpGs to be a frequent source of mutations in the β- and γ-globin genes, but not in the α-globin genes. In fact, the evidence points to CpGs being a frequent source of mutations in both the α- and β-globin genes. This suggests either that the mutation rates of both methylated and unmethylated CpGs are abnormally high, which conflicts with published evidence, or that there is a finite chance of some CpGs in the α-globin genes of certain individuals being methylated and therefore subject to mutation.

Sinsheimer [1] appears to have been the first to draw attention to the under-representation of the dinucleotide CpG in mammalian DNA after he found that an enzymatic digest of calf thymus DNA contained only a low proportion of CpGs and no 5-methyl CpGs. His work was followed up by Schwartz et al. [2] who showed that CpG dinucleotides occurred in the DNA of animal tissues at no more than one third of the frequency expected from the base composition if the distribution of dinucleotides were random, while GpC dinucleotides occurred at the expected frequency. After Roy and Weissbach [3] had isolated a methylase from HeLa cells that specifically methylated CpGs, Coulondre et al. [4] found a methylase in *E. coli* that methylates the second cytosine in the sequence CCAGG. In the *lac*I gene of *E. coli* these 5-methylcytosines acted as 'hotspots', because deamination converted the 5-methylcytosines to thymines which then paired with adenines. In a methylase-deficient mutant of *E. coli* these hotspots were absent, apparently because uracil derived from deamination of unmethylated cytosine was excised by uracil-DNA-glycosidase and replaced by cytosine, while thymine derived from deamination of 5-methylcytosine was not excised. Recently, Jones et al. [5] discovered that *E. coli* does in fact have a system that specifically repairs T/G mismatches derived from deamination of CpGs and replaces them by C/G pairs. They speculated that Couloudre et al. did not observe that repair, because they used F-lac episomes that were present in large numbers in the *E. coli* cell, and that the number of repair enzymes in each cell was too small to cope with them. In mammalian cells mismatched T/G pairs are subject to excision repair that is biased in favour of C/G: when circular SV40 DNA containing one mismatched T/G pair was cloned in monkey kidney cells, most of the mispairs were corrected to C/G rather than T/A pairs [6–8]. In monkey kidney cells deamination of 5-methylcytosine in a (5-mC·

Correspondence address: M.F. Perutz, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

* Published originally in the Journal of Molecular Biology, vol. 212 (1990).

G/G · 5m-C) pair to (T · G/G · 5m-C) is followed by mismatch repair that preserves the methylated strand, thus ensuring restoration of the original (C · G/G · 5m-C) pair [9]. All the same, the CpG dinucleotide is present in mammalian genomes at only about 20% of its expected frequency, and the TpG/CpA dinucleotide occurs in a corresponding excess, thus testifying to a net loss of CpG during evolution [10].

That loss is not entirely random. Salser [11] found 22 CpGs in the mRNA for rabbit α-globin, but only three CpGs in the mRNA for β-globin. He suggested that most CpGs might have been eliminated from the β-globin gene because they were methylated and therefore formed hotspots for mutations. This prediction was borne out by the discovery of methylated CpGs on the human β- and γ-globin genes. On the other hand, the human α_1- and α_2-globin genes and many other regions of mammalian chromosomes contained islands of unmethylated CpGs where the frequency of this dinucleotide was higher than expected from the base composition [12–15]. The human α_1- and α_2-globin genes each contain 35 CpGs, compared to 26 expected on a random basis, while the β-globin gene contains only five and the γ-globin gene only two, consistent with the generally accepted notion that methylated CpGs have been lost during evolution, while non-methylated CpGs have been preserved. If only methylated CpGs gave rise to hotspots, then the amino acid substitutions to be expected from C → T transitions should have shown up frequently among the human haemoglobins with abnormal β- or γ-chains, but only rarely among those with abnormal α-chains. I set out to discover if this is really true.

Table 1 shows that deamination of 5-methylcytosine in the coding strand of DNA replaces CpG in the mRNA by CpA; deamination of 5-methylcytosine in the non-coding strand of DNA replaces CpG in the mRNA by UpG. In the β-globin chain, C → T transitions in the coding strand would cause three valines to be replaced by methionines, one valine by an isoleucine, and one glycine by a serine. C → T transitions in the non-coding strand would cause no amino acid substitutions, because they all occur in the third letters of the codons. Table 2 shows the number of apparently independent occurrences of the five predicted replacements in the β-globin chains that have either been published or reported to the International Hemoglobin Information Center in Augusta, Georgia. Hemoglobin Köln, which causes the severest symptom, i.e., inclusion body anaemia, has been reported 32 times. Of the other four replacements, two cause high oxygen affinity which is a mild symptom, and the other two are asymptomatic. They have been reported between two and five times. On the other hand, neither of the two electrophoretically silent substitutions predicted in the γ-chain have been found, probably because such mutants rarely manifest themselves symptomatically: only six electrophoretically silent mutants have been detected among the 44 abnormal foetal haemoglobins reported so far.

The α_1 and α_2 genes differ at only two codons, neither of which contains a CpG [16]; therefore, we have to deal with only a single set of α-globin codons. Table 3 lists the number of apparently independent occurrences of the amino acid substitutions predicted in the α-globin chain that have been either published or reported to the International Hemoglobin Information Center. Ten of the eleven predicted and electrophoretically distinct amino acid substitutions have been found, one eight times, one seven times, one five times, and three have been found four times. All occurred in unrelated families. Two of the electrophoretically silent ones have also been found. They involve replacements of prolines at the $\alpha_1\beta_1$ contact by leucines, which disturbs the allosteric equilibrium, raises the oxygen affinity and manifests itself clinically by polycythemia (too many red cells). 33 of the remaining 58 possible C → T transitions would produce no amino acid substitu-

Table 1

Nucleotide base changes caused by deamination of 5-methylcytosine in CpG dinucleotides

	DNA		mRNA
	Coding	Non-coding	
Wild type	CpG	CpG	CpG
Mutant	TpG	(CpA)	CpA
	(CpA)	TpG	UpG

Table 2

Abnormal human hemoglobins due to C → T transitions in CpG dinucleotides in the β- and γ-globin genes

HOA, high oxygen affinity; n.o., not observed.

	Wild type		Mutant			Number of apparently independent occurrences	Clinical manifestations
	mRNA	Residue	mRNA	Residue	Name		
β-chain							
Coding strand							
11	GCC-GUU	Val	AUU	Ile	Hamilton	5	–
20	AAC-GUG	Val	AUG	Met	Olympia	5	HOA
69	CUC-GGU	Gly	AGU	Ser	City of Hope	4	–
98	CAC-GUG	Val	AUG	Met	Köln	32	inclusion body anaemia, HOA
109	AAC-GUG	Val	AUG	Met	San Diego	2	HOA
Non-coding strand							
10	GCC	Ala	GCU	Ala			
19	AAC	Asn	AAU	Asn			
68	CUC	Leu	CUU	Leu			
97	CAC	His	CAU	His			
108	AAC	Asn	AAU	Asn			
γ-chain							
Coding strand							
113	ACC-GUU	Val	AUU	Ile			
118	UUC-GGC	Gly	AGC	Ser			
Non-coding strand, all silent							

Table 3

Abnormal human hemoglobins due to C → T transitions in CpG dinucleotides in the α-globin gene

The codons for the globin genes in tables 2 and 3 have been taken from ref. 18, and the abnormal haemoglobins from the list of the International Hemoglobin Information Center published in *Hemoglobin* 12 (1988) no. 3. HOA, high oxygen affinity; n.o., not observed.

	Wild type		Mutant			Number of apparently independent occurrences	Clinical manifestations
	mRNA	Residue	mRNA	Residue	Name		
α_2-chain							
Coding strand							
6	GCC-GAC	Asp	AAC	Asn	Dunn	4	HOA
23	GGC-GAC	Glu	AAG	Lys	Chad	5	–
47	UUC-GAC	Asp	AAC	Asn	Arya	2	slightly unstable
64	GCC-GAC	Asp	AAC	Asn	Aida	8	–
75	GAC-GAC	Asp	AAC	Asn	Matsui-Oki	7	–
85	AGC-GAC	Asp	AAC	Asn	G-Norfolk	4	–
92	CUU-CGG	Arg	CAG	Gln	Cape Town	4	HOA
116	GCC-GAG	Glu	AAG	Lys	O-Indonesia	7	–
141	UAC-CGU	Arg	CAU	His	Surèsnes	2	HOA
Plus 20 electrophoretically silent mutations unobserved							
α_2-chain							
Non-coding strand							
44	CCG	Pro	CUG	Leu	Milledgeville	1	HOA
92	CGG	Arg	UGG	Trp		n.o.	
95	CCG	Pro	CUG	Leu	G-Georgia	6	HOA
141	CGU	Arg	UGU	Cys	Nunobiki	1	HOA
Plus 18 changes in 3rd codon and 7 Ala → Val n.o.							

tions, which leaves 25 predicted amino acid substitutions that have not been reported. Only one of these would be visible electrophoretically: Arg 92α → Trp; the rest includes eleven Val → Ala; three Val → Met; one Val → Ile and two Gly → Ser substitutions. All these substitutions are conservative. The three valines that are replaced by methionines all lie in surface crevices, so that the bulkier methionine side chain can be accommodated without disturbing the structure. It is improbable, therefore, that any of these substitutions would manifest themselves clinically and they are therefore likely to escape detection.

The reported frequency of an abnormal haemoglobin is a function of its actual frequency in the population, divided by the probability of its detection. For haemoglobin Köln, which causes severe haemolytic anaemia, that factor would be unity. Haemoglobins with abnormal electrophoretic mobility show up either in routine hospital tests or in large-scale screening programmes, but the fraction of the population caught by such programmes must be small, say, no larger than one tenth. Haemoglobins with high oxygen affinity rarely produce disease, but may manifest themselves by polycythaemia in routine blood counts. Again the fraction detected may not be more than one tenth. Abnormal haemoglobins that are electrophoretically silent and asymptomatic are detected even more rarely. α-chain mutants are intrinsically harder to detect than β-chain mutants because they affect only about a quarter of the haemoglobin in the red cell. For all these mutants the probability of detection is low and the real frequency is likely to be much larger than given in tables 2 and 3.

The frequent occurrence of the amino acid substitutions predicted among the β-globins is consistent with the notion that methylated CpGs form hotspots, but the discovery of ten out of the eleven predicted, electrophoretically distinct substitutions among the α-globins, and their repeated presence in unrelated individuals, are unexpected if the CpGs in the α-globin genes are supposedly unmethylated. It has been suggested to me that the occurrence of the α-globin mutants may not necessarily imply abnormally high mutation rates at these CpGs, because the world-wide search for abnormal haemoglobins has already led to the discovery of all possible electrophoretically distinct abnormal haemoglobins. However, this is not true. At the eight loci in the α-globin gene shown at the top of table 3, transitions and transversions could give rise to 51 other electrophoretically visible abnormalities, but only 20 of these have been found.

The evidence for undermethylation of the α-globin genes comes from restriction mapping rather than sequencing of the germ-line DNA [15], whence some methylated CpGs might have escaped detection, but this can hardly be true of all ten CpGs involved in the electrophoretically observed amino acid replacements.

On the evidence presented here, one would be led to conclude that unmethylated CpGs also form hotspots, for example, because proof-reading by the polymerase-associated exonuclease is error-prone, but this would contradict previous work and raise the question of why unmethylated CpGs have not been largely eliminated from the α-globin genes in evolution, just as methylated ones have been eliminated from the β-globin genes. Alternatively, single CpGs in the α-globin genes might occasionally become methylated at random and these single methylated CpGs might then act as hotspots. If that were true, then sequence analysis of germ-line DNA of different individuals should reveal single methylated CpGs at different loci. Whatever the explanation, it is difficult to see what selective advantage, if any, can now be assigned to the unmethylated islands in the α-globin genes.

Cooper and Youssoufian [17] estimate that nearly one third of intragenic single base-pair mutations causing human inherited disease occurs in CpG dinucleotides, because CpG is up to 42-times more unstable than predicted from random mutations. My findings show that they are also responsible for a substantial fraction of haemoglobinopathies.

Acknowledgements

I thank Drs A.P. Bird and M. Radman for helpful criticism and Drs T.H.J. Huisman and

B.B. Webber at the International Hemoglobin Information Center in Augusta, GA, for kindly searching their records in order to supply me with the frequencies of the independently reported abnormal haemoglobins listed in tables 2 and 3. This work was supported by NIH grant no. HL 31461 and NSF grant no. DMB 8609842.

References

1 R.L. Sinsheimer, J. Biol. Chem. 215 (1955) 579.
2 M.N. Schwartz, T.A. Trautner and A. Kornberg, J. Biol. Chem. 237 (1962) 1961.
3 P.H. Roy and A. Weissbach, Nucleic Acids Res. 2 (1975) 1669.
4 C. Coulondre, J.H. Miller, P.J. Farabough and W. Gilbert, Nature 274 (1978) 775.
5 M. Jones, R. Wagner and M. Radman, J. Mol. Biol. 194 (1987) 155.
6 T.C. Brown and J. Jiricny, Cell 50 (1987) 945.
7 T.C. Brown and J. Jiricny, Cell 54 (1988) 705.
8 K. Wiebauer and J. Jiricny, Nature 339 (1989) 234.
9 J.T. Hare and J.H. Taylor, Proc. Natl. Acad. Sci. U.S.A. 82 (1985) 7350.
10 S. Ohno, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 9630.
11 W. Salser, Cold Spring Harbor Symp. Quant. Biol. 42 (1977) 985.
12 A. Bird, M. Taggart, M. Frommer, O.J. Miller and D. Macleod, Cell 40 (1985) 9199.
13 A.P. Bird, Nature 321 (1986) 209.
14 W.R.A. Brown and A.P. Bird, Nature 322 (1986) 477.
15 A.P. Bird, M.H. Taggart, R.D. Nichols and D.R. Higgs, EMBO J. 6 (1987) 999.
16 S.A. Liebhaber, M. Goossens and Y.W. Kan, Nature 290 (1981) 26.
17 D.N. Cooper and H. Youssoufian, Hum. Genet. 78 (1988) 151.
18 E.H. Bunn and B.G. Forget, Hemoglobin: Molecular, genetic and clinical aspects (W.B. Saunders, Philadelphia, 1986) p. 182.