Over-representation of the disease associated (CAG) and (CGG) repeats in the human genome

Jian Han*, Chuancheih Hsu[1], Zhou Zhu[2], John W.Longshore and Wayne H.Finley
Laboratory of Medical Genetics, [1]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294 and [2]Department of Internal Medicine, Yale University, New Haven, CT, USA

Received September 8, 1993; Revised and Accepted April 7, 1994

ABSTRACT

Expansion of trimer repeats has recently been described as a new type of human mutation. Of the 64 possible trimer compositions, only the CGG and CAG repeats have been implicated in genetic diseases. This study intends to address two questions: (1)What makes the CGG and CAG repeats unique? (2) Could other trimer repeats be involved in this type of mutation? By computer analysis of trimer and hexamer frequency distributions in approximately 10 Mb of human DNA, twenty trimer motifs (ten complementary pairs) have been identified that are the most likely to be expanded. The frequency distribution study also indicated that the expanded trimer motif in Fragile-X syndrome is GGC instead of CGG. DNA linguistics studies revealed that the GGC/GCC and CAG/CTG repeats were over-represented in the human genome. Further analysis of base composition suggested that the CCA/TGG repeats may be involved in the trimer expansion mutation since they possessed many similar characteristics to GGC/GCC and CAG/CTG. The computer aided sequence analysis studies reported here may help to understand the molecular mechanisms of trimer repeat expansion.

INTRODUCTION

Trimer repeat expansion is a novel type of human mutation that has only recently been described. Within the past three years, this type of mutation has been found to be associated with seven human genetic disorders. Of the 64 (4^3) possible trimer compositions, 4 of them (AAA/TTT and CCC/GGG) could only form a uniform string but not trimer repeats. The other 60 trimers are divided into ten groups, each group having 6 trimers related by frame shift and complementation (1, 2). For example, in the CGG group, depending on where the counting starts on a CGG repeat string, the expanded motifs could be CGG, GCG, GGC, or their complementary motifs CCG, CGC, and GCC. Interestingly enough, only two of the ten groups have been implicated in genetic diseases. Expansion of the CAG repeat has been shown to be responsible for Huntington's disease (3), myotonic dystrophy (4–7), spinobulbar muscular atrophy (8), spinalcerebellar ataxia type I (9), and dentatorubral-pallidoluysian atrophy (10). Fragile-X (11–14) and FRAXE (15) syndromes are caused by CGG repeat expansion.

Some observations have been made about this new type of mutation. All the unstable trimer repeats found so far are transcribed, but are not necessarily translated (1–3, 7–11, 15). Segregation analysis of flanking markers excludes meiotic crossing-over events (16). The length alteration is largely due to mitotic events, predominantly during early development. The disease associated trimers are CG-rich, and the CGG repeat has methylation involved. These repeats are polymorphic, and the copy numbers are unstable within and among families. Several models, including the DNA replication slippage (17) and the recombination gap repair (18), have been proposed to explain the mechanisms of trimer expansion.

In this study, we were interested in two questions: What makes the CGG and CAG repeats unique? Are other trimer repeats besides CGG and CAG unstable and involved in expansion? Two sequence analysis strategies were used to address these questions. First, the genomic distribution of trimers was studied with a previously described computer program, OLIGOMER (19). By a direct counting approach, the OLIGOMER program reported the appearance frequencies of all the possible di-, tri-, tetra-, penta-, and hexamer compositions in more than 9.7 Mb of human DNA sequence in GenBank. The result indicated that the distribution of oligonucleotides in the human genome is non-random which suggested that some high genomic frequency trimer motifs would have a higher probability of expansion. The second sequence analysis strategy was a DNA linguistic approach. A second order Markov chain model was proposed for the generation of a genetic text (DNA strings), then the expected number of occurrences of each word (hexamer) in the text was calculated. If some word occurred considerably more (or less) frequently than expected, it was termed 'meaningful' or 'functional' (20). When hexamers representing the trimer repeats were studied in this manner, it was found that the diseases associated trimer repeats $(CGG)_2$ and $(CAG)_2$ were extremely over-represented in the human genome. A third trimer repeat, $(CCA)_2$, come to attention because of its over-representation in the human genome. Using an 'odds' ratio approach, Burge *et al.* (21) also reached the same conclusion that the CCA/TGG triplet was over-represented in human sequences.

*To whom correspondence should be addressed at: 420G Sparks Center, UAB Station, Birmingham, AL 35294, USA

METHOD AND ALGORITHM

Oligonucleotide frequency analysis with the OLIGOMER program

A human genome specific database of 9.74Mb was constructed by downloading sequences from GenBank (Version 7.2). The STRINGSEARCH program in the GCG (22) sequence analysis software package allowed the identification of sequences that met the searching criteria. The FETCH program downloaded these identified sequences into a VAX mainframe under a specific directory. Repeated entries were identified and removed from the database. After the descriptive text was deleted, sequences were appended into a single file on VAX to serve as a genome specific database. The human specific database was built using the command STRINGSEARCH GENEMBL:HUM* COMPLETE CDS finds every entry in the GenBank and/or EMBL sequence libraries whose definitions contain the text pattern 'human' and 'complete cds.' Use of 'complete cds' (complete coding sequences) as a search pattern helped to eliminate most of the incomplete and duplicated sequence entries. The STRINGSEARCH program reported the findings in a file named GENEMBL.STRINGS. The FETCH @GENEMBL. STRINGS command was used to download selected human sequences into the VAX. Then, APPEND hum*.*;* command assembled these downloaded human sequences into a single file for further study. The human specific database constructed contained 9,739,600 bp of human DNA sequence from 3576 entries. The OLIGOMER program reported each of the 16 dimer, 64 trimer, and 4,096 hexamer frequencies in the assembled human genome specific database.

The Markov chain method and DNA linguistic studies

The Markov chain models have been used to predict oligonucleotide frequencies in prokaryotic and eukaryotic genomes (23–26). The second order Markov chain was used in this study for calculating hexamer frequencies. Using the observed dimer and trimer frequency data reported by the OLIGOMER program, the model can give expected frequency values for related hexamers. For example, the genomic frequency of GCCGCC can be calculated as the following:

$$f(GCCGCC) = \frac{f(GCC)f(CCG)f(CGC)f(GCC)}{f(CC)f(CG)f(GC)}$$

This model was used to predict the occurrence frequencies of the 4096 hexamers in human genome. The results were used in a DA linguistic study in which the expected and the observed hexamer frequencies were compared so that over- or under-represented hexamers can be identified. A standard deviation (std) value was calculated for each comparison according to the following:

$$std(f) = \frac{Observed\ (f) - Expected\ (f)}{\{Expected\ (f)\}^{1/2}}$$

The standard deviation values for the 4096 hexamers were ranked in descending order. The larger the standard deviation value, the more the hexamer been over-represented in the human genome.

RESULTS AND DISCUSSION

Related trimer motifs have unequal opportunity for expansion

Figure 1 shows the observed trimer frequencies in the ten groups. Significant frequency differences were observed both between and within trimer groups. Differences within the trimer groups, in particular, caught our attention. This difference may indicate an unequal opportunity for expansion, even for the trimers within the same group. Trimer pairs with a significantly higher genomic frequency than the rest of the group are denoted with an asterisk in Figure 1. In the CCG group for example, the frequencies for GGC and its complementary trimer GCC were about 2.4 times higher than that of the other trimers in the group (Figure 1). This data indicated that, by chance along, a GGC motif would be more

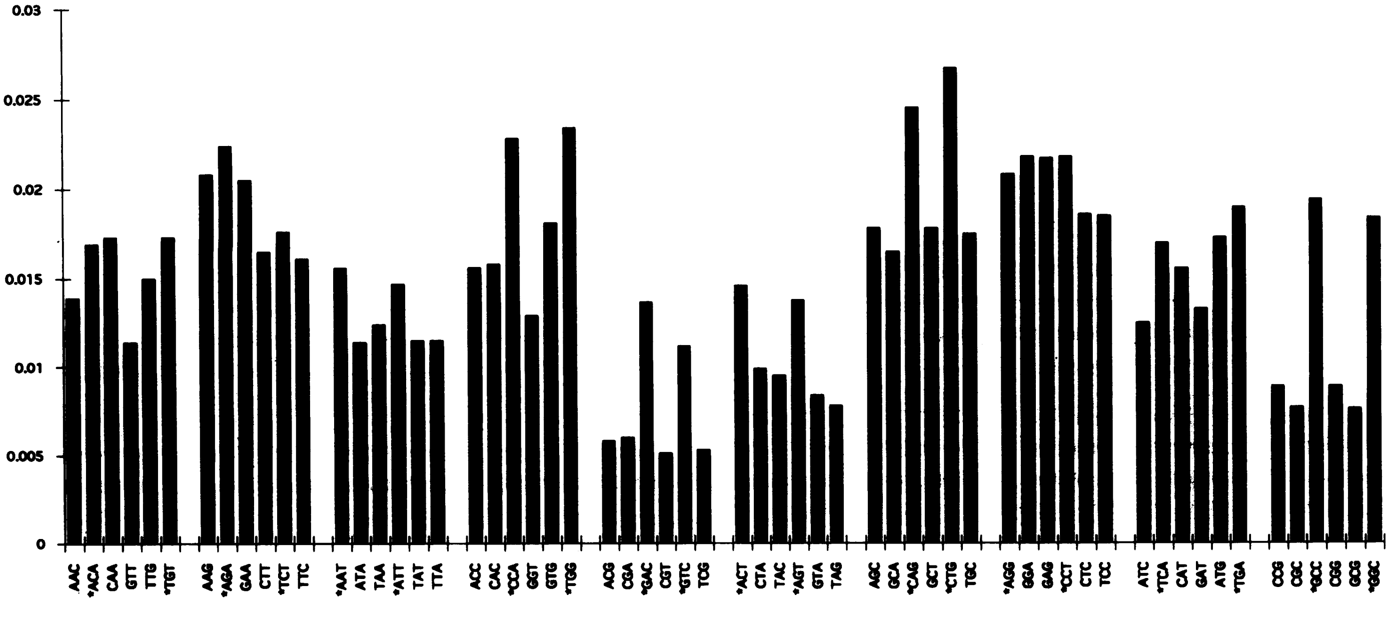

Figure 1. Genomic frequency of trimers in the ten related groups.

likely to expand than that of a CGG. In other words, a trimer repeat string was more likely to begin and end with GGC, or it was more likely to expand by a GGC motif than by a CGG motif.

To further study this issue, genes with (CGG)n repeats (n ≥ 4) were pulled out from GenBank, the trimer at the 5′ end of the repeats was determined for each gene (Table 1). Of the 51 genes, 22 of them (43%) initiate the repeated string with GGC, including the Fragile-X gene (3), 9 (17.6%) with GCC, including the FRAXE gene, 9 (17.6%) with GCG, 5 (9.8%) with CGC and 3 (5.9%) with CCG or CGG. As the most popular motifs, GGC and GCC initiate trimer repeat strings more than 60% of the time (31/51). As suggested by this observation, at the DNA level the expanded motif in Fragile-X and FRAXE genes should be GGC/GCC instead of CGG/CCG. Similar observations can be made for the other disease related trimer, the CAG/CTG pair, which is more frequent than others in the group. Since the trimer expansion is more likely related to a DNA replication or a damage repair event rather than a transcription or a translation related event, it is more appropriate to identify the expanded motifs at the DNA level. Our results indicated that for each of the ten trimer groups, at least one complementary pair was more likely to expand (those labeled with * in Figure 1).

The disease associated trimer repeats are over-represented in the human genome

The oligonucleotide frequency study presented above helped to identify trimer motifs that were more likely to be involved in

Table 1. Human genes with (GGC)n repeat

Locus	Definition	Sequence	5′ trimer
Humrxrb	Retinoid X receptor	5′ CCGCA(GCG)$_{7}$TGGCA 3′	GCG
Hhumglpex	Glutathione peroxidase	5′ GGCTA(GCG)$_{6}$GCCCAG 3′	GCG
Humrash	C-Ha-ras oncogene	5′ CGTAA(GCG)$_{6}$GGTGG 3′	GCG
Humrsc830	Random sequence	5′ CGGTA(GCC)$_{5}$GCGCC 3′	GCC
Humferh	Ferritin	5′ AGCCA(CCG)$_{5}$CCTCTC 3′	CCG
Humgdf1	GDF-1	5′ GTTTC(GCG)$_{6}$GCAGCC 3′	GCG
Humrgit	ITS1	5′ GGCCG(CGC)$_{6}$GGCGG 3′	CGC
Humgnas1	G-protein	5′ TAAGA(GCG)$_{4}$GCAGC 3′	GCG
Humia1x	z-finger DNABP	5′ GAGGG(GGC)$_{5}$GC 3′	GGC
Humthyp	Parathymosin	5′ CGTGT(CGC)$_{5}$CACCG3′	CGC
Humgprklg	G-protein receptor	5′ GCGCA(GGC)$_{5}$GCCCT 3′	GGC
Humtrlall1	ALL-1	5′ (GCG)$_{7}$GGAAGCAGC 3′	GCG
Humadra2r	Adrenergic receptor	5′ AGAAG(GGC)$_{5}$CCGCA 3′	GGC
Humb94	B94 protein	5′ GAGCT(GGC)$_{6}$GGGCG 3′	GGC
Humcoupii	COUP-TFII	5′ GGGCA(GGC)$_{6}$CCAGC 3′	GGC
Humodc1a	Ornithine decarboxylase	5′ CTGTA(GCC)$_{6}$CGCCG 3′	GCC
Humhdad	HDAD	5′ GCCCG(CGC)$_{4}$CTCAG 3′	CGC
Humen1aa	Engrailed protein	5′ GCAGT(GGC)$_{6}$CGCAG 3′	GGC
Humggf2a	Glial growth factor 2	5′ AGGAAGGC)$_{5}$GGGCG 3′	GGC
Humbcraa	BCR	5′ GAGGA(GGC)$_{7}$AGCGG 3′	GGC
Humpdgfa1	PDGFA	5′ GGTGT(GGC)$_{7}$CCCAG 3′	GGC
Humap2	AP-2	5′ ATGCT(GCC)$_{6}$GCTGC 3′	GCC
Humcanpo2	Neutral protease	5′ CCGGG(GGC)$_{8}$GGTGG 3′	GGC
Humk10a	Keratin 10	5′ AGTTCCGG)$_{6}$CTACG 3′	CGG
Humhhr6b	HHR6B	5′ AGTCT(CGG)$_{8}$CGATC3′	CGG
Humgab3r	GABA-A receptor	5′ CCAGC(GCG)$_{7}$GCAGC 3′	GCG
Humrgmb	28S rRNA	5′ GGAG(GGC)$_{9}$ 3′	GGC
Hum21seqh	Chromosome 21 sequence	5′ CACCG(CGC)$_{8}$GGGGC 3′	CGC
Hummrpx	MRP	5′ CCCTG(CGC)$_{7}$CCGCCGC 3′	CGC
Humclg4	Type-4 collagenase	5′ GACCT(GCG)$_{8}$GGGGC 3′	GCG
Humerk2a	ERK2A	5′ AACAT(GGC)$_{6}$GGGCGC 3′	GGC
Humar	Androgen receptor	5′ GTGGT(GGC)$_{10}$ 3′	GGC
Humror1a	Transmembrane receptor	5′ TGGGA(GCC)$_{6}$TCAGC 3′	GCC
Humfmr1	**Fragile X**	5′ GCGCG(GGC)$_{10-50}$ 3′	GGC
Hum*	**FRAXE**	5′ CCGCT(GCC)$_{6-25}$GCTGCCG3′	GCC
Humsef21b	SEF2−1B	5′ GTAGT(GGC)$_{7}$GGGGA 3′	GGC
Humegr2a	EGR2	5′ CAGCA(GCC)$_{6}$TATAA 3′	GCC
HumtefsII	Elongation factor	5′ (GCC)$_{9}$GCGGG 3′	GCC
Hummar	NR1−1	5′ GCGGA(GCC)$_{6}$GGGCC 3′	GCC
Humpura	Pur-alpha	5′ GCAGT(GGC)$_{5}$GGGGC 3′	GGC
Humadrb1	ADRB1	5′ CGCC(CCG)$_{5}$CCCACC 3′	CCG
Huma8seq	A8SEQ	5′ TCGTC(CGG)$_{4}$CAGCG 3′	CGG
Humafpebp	a-fetoprotein EBP	5′ AGCTC(CCG)$_{5}$TCGCC 3′	CCG
Humhox	homeobox protein	5′ GTCCT(GGC)$_{5}$AGCAGC 3′	GGC
Humdhpra	hDHPR	5′ GGAT(GGC)$_{5}$TGCAG 3′	GGC
Hummevkin	Mevalonate Kinase	5′ GGGGA(GGC)$_{5}$AGGAT 3′	GGC
Humarf6a	hARF6	5′ GTTTC(GCG)$_{5}$TTGTT 3′	GCG
Humcol4a2a	a-2 collagen	5′ GGGAA(GGC)$_{5}$TCCGT 3′	GGC
Humen2aa	Engrailed protein	5′ GGCCG(GGC)$_{5}$CGGAG 3′	GGC
Humgrpra	Glucocorticoid receptor	5′ CTTCT(GCC) $_{5}$TCGCA 3′	GCC
Humpcd	Potassium channel	5′ GCTGT(GGC)$_{4}$TGCGA 3′	GGC

Table 2. Standard deviation ranking of the trimer repeats related hexamers.

Hexamer	Observed(f)	Expected(f)	Std	Percentile
AACAAC	0.000271	0.000180	0.006783	0.057129
ACAACA	0.000330	0.000237	0.006041	0.069580
CAACAA	0.000396	0.000239	0.010155	0.018555
GTTGTT	0.000198	0.000136	0.005316	0.087646
TTGTTG	0.000275	0.000186	0.006526	0.061035
TGTTGT	0.000249	0.000208	0.002843	0.195313
				Mean 0.081543
AAGAAG	0.000816	0.000501	0.014073	0.005371
AGAAGA	0.000830	0.000540	0.012480	0.008057
GAAGAA	0.000829	0.000491	0.015254	0.003418
CTTCTT	0.000446	0.000324	0.006778	0.057373
TCTTCT	0.000485	0.000345	0.007537	0.043213
TTCTTC	0.000486	0.000315	0.009635	0.021973
				Mean 0.023234
AATAAT	0.000291	0.000168	0.009490	0.022705
ATAATA	0.000220	0.000118	0.009390	0.023926
TAATAA	0.000257	0.000125	0.011806	0.010742
ATTATT	0.000348	0.000199	0.010562	0.016113
TATTAT	0.000260	0.000152	0.008760	0.028564
TTATTA	0.000266	0.000152	0.009247	0.025146
				Mean 0.021200
ACCACC	0.000412	0.000303	0.006262	0.065186
CACCAC	0.000471	0.000315	0.008790	0.028076
CCACCA	0.000601	0.000443	0.007507	0.043701
GGTGGT	0.000365	0.000248	0.007430	0.045410
GTGGTG	0.000543	0.000354	0.010045	0.019287
TGGTGG	0.000695	0.000465	0.010666	0.015869
				Mean 0.036255
ACGACG	0.000067	0.000024	0.008777	0.028320
CGACGA	0.000070	0.000025	0.009000	0.026367
GACGAC	0.000115	0.000061	0.006914	0.054199
CGTCGT	0.000036	0.000017	0.004608	0.108643
GTCGTC	0.000063	0.000040	0.003637	0.150146
TCGTCG	0.000037	0.000018	0.004478	0.112793
				Mean 0.080078
ACTACT	0.000183	0.000122	0.005523	0.082520
CTACTA	0.000173	0.000086	0.009381	0.024170
TACTAC	0.000157	0.000083	0.008123	0.034180
AGTAGT	0.000116	0.000079	0.004163	0.126221
GTAGTA	0.000079	0.000046	0.004866	0.098877
TAGTAG	0.000121	0.000044	0.011608	0.011230
				Mean 0.062866
AGCAGC	0.000697	0.000420	0.013516	0.006104
GCAGCA	0.000602	0.000390	0.010735	0.015137
CAGCAG	0.000827	0.000577	0.010408	0.016846
GCTGCT	0.000736	0.000459	0.012929	0.007568
CTGCTG	0.001007	0.000683	0.012398	0.009033
TGCTGC	0.000690	0.000452	0.011195	0.013184
				Mean 0.011312
AGGAGG	0.000867	0.000525	0.014926	0.003662
GGAGGA	0.000864	0.000546	0.013609	0.005859
GAGGAG	0.000918	0.000543	0.016093	0.002441
CCTCCT	0.000713	0.000537	0.007595	0.041992
CTCCTC	0.000577	0.000451	0.005933	0.072021
TCCTCC	0.000598	0.000442	0.007420	0.045654
				Mean 0.028605
ATCATC	0.000335	0.000172	0.012429	0.008789
TCATCA	0.000421	0.000238	0.011862	0.010254
CATCAT	0.000413	0.000215	0.013503	0.006348
GATGAT	0.000358	0.000196	0.011571	0.011475
ATGATG	0.000442	0.000259	0.011371	0.012207
TGATGA	0.000510	0.000284	0.013411	0.006592
				Mean 0.009277
CCGCCG	0.000244	0.000093	0.015658	0.002930
CGCCGC	0.000274	0.000080	0.021690	0.000732
GCCGCC	0.000377	0.000210	0.011524	0.011719
CGGCGG	0.000272	0.000082	0.020982	0.001221
GCGGCG	0.000285	0.000071	0.025397	0.000488
GGCGGC	0.000388	0.000181	0.015386	0.003174
				Mean 0.003377

expansion within a particular group. The DNA linguistic study described below, however, may help to identify which of the ten groups are more likely to be involved.

DNA linguistics was used to investigate if the disease associated trimer repeats, $(GGC)_2$ and $(CAG)_2$, possess any unique feature that allow them to be differentiated from the 4096 possible hexamers. The Markov chain model was used to generate expected hexamer frequency values. The overall ratio of the observed-to-expected was 1.079 ± 0.3167 for the 4096 hexamers. The χ^2 value was 0.066 ($P < 0.001$) indicating that the model provided accurate estimates for most of the observed hexamer frequencies. A standard deviation value was calculated for each of the hexamers. After standard deviation values were obtained for all 4096 hexamers, they were ranked in descending order. The standard deviation value served as a measurement of the model's predictive power. A large standard deviation value indicated that the model has under-estimated the real frequency of a hexamer. Table 2 shows the position of trimer related hexamers in this ranking in percentile format. A low percentile number indicates that the hexamer was under-estimated by the model, or was over-represented in the genome. An over-represented hexamer does not necessarily mean that the hexamer is highly abundant in the human genome. For example, the Fragile-X associated trimer repeat $(GGC)_2$ has an observed genomic frequency of 3.77×10^{-4}, ranking at 1,517 among the 4,096 hexamers (37th percentile). However, the value for its standard deviation (0.021) was ranked 6th (or 0.15 percentile) of the most over-represented hexamers. The standard deviation ranking for $(CAG)_2$ was at the 1.68 percentile. As a related group, the average standard deviation ranking for the GGC and the CAG groups were 0.3 and 1.1 percentiles respectively. This ranking data indicates that the observed genomic frequency for $(GGC)_2$ and $(CAG)_2$ repeats could not be well explained by the mathematical model. In other words, their genomic frequencies were largely determined by biological factors. It is possible that the factors responsible for the over-representation also play a role in the trimer expansion mutation process.

The CCA/TGG trimer pair as a candidate for expansion

Several common characteristics were noticed for the diseases associated CAG and GGC trimer repeats: (i) Both trimers are over-represented in the human genome, (ii) They are CG-rich, (iii) and they contain pyrimidine as well as purine bases. If we hypothesize that these common features were connected with the mutagenesis mechanisms, we would expect that other diseases associated trimer repeats also sharing the same characteristics.

Seven of the ten hexamer groups (Table 2) are over-represented in the human genome, each has a mean percentile ranking of less than the 4th. These highly over-represented groups are denoted by GGC, ATC, CAG, AAG, AGG, AAT, and CCA. Four of the seven groups are CG-rich, including GGC, CAG, AGG, and CCA groups. Of the four CG-rich groups, CCA is similar to GGC and CAG, in that it contains both pyrimidine and purine bases. The AGG group has only purine bases. We also noticed that in the CCA group, as in the GGC and CAG groups, there is no significant frequency difference between the complementary motifs (see Figure 1). However, the frequencies for the AGG, GAG and GGA motifs are significantly different from that of their complements (CCT, CTC and TCC). This may be caused by the asymmetric strand distributions of the GA/TC pair, as we noticed in an earlier study (19).

In summary, a computer program was developed to study the genomic distributions of oligonucleotides. The direct frequency counting results indicated that the distribution of expandable trimer motifs in human genome was non-random. Of the 64 possible trimer compositions, 20 were more likely to become the expanded motifs (two for each of the 10 groups, see Figure 1). At the DNA level, GGC was identified as the expanded motif in the Fragile-X gene. When a DNA linguistics approach was used, it was learned that the disease associated trimers were over-represented in human genome. By base composition analysis, the CCA/TGG motif was recognized to share the most common characteristics with the CAG and GGC motifs. The possibility of CCA repeat being involved in disease is currently under investigation.

The OLIGOMER program is available upon request.

ACKNOWLEDGEMENTS

We thank Mr Eric Crawford and Ms Judy Knops for critical reading of this manuscript. We thank Dr Elliot J.Lefkowitz, Mr Tim Buckner and Ms Lily Yang for assisting in DNA sequence database development. We also thank the Center for Aids Research at the University of Alabama at Birmingham (P30 AI27767) for computer facility support.

REFERENCES

1. Richards, R.I., and Sutherland, G.R. (1992). Nature Genetics, **1**,7–9.
2. Mandel, J.L. Questions of expansion. (1993). Nature Genetics, **4**, 8–9.
3. Huntington's Disease Collaborative Research Group. (1993). Cell, **72**, 371–383.
4. Harley, H.G., Brook, J.D., Rundle, S.A., Crow, S., Reardon,W., Buckler, A.J., Harper, P.S., Housman D.E., and Shaw D.J. (1992). Nature, **355**,545–546.
5. Buxton, J., Shelbourne P., Davies, J., Jones, C., Van Tongeren T., Aslanidis, C., de Jong, P.,Jansen, G., Anvret, M., Riley, B., Williamson, R., and Johnson, K. (1992). Nature, **355**,547–548.
6. Aslanidis, C., Jansen G., Amemiya, C., Shutler G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman J., Wormskamp N.G.M., Vooijs M., Buxton J., Johnson K., Smeets H.J.M., Lennon, G.G., Carrano, A.V., Korneluk, R.G., Wieringa, B., and de Jong P.J. (1992). Nature, **355**,548–551.
7. Fu, Y.H., Pizzuti, A., Fenwick, R.G., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T., DeJong, P., Wieringa, B., Korneluk, R., Perryman, M.B., Epstein, H.F., and Caskey, C.T. (1992). Science, **255**, 1256–1258.
8. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., and Fischbeck K.H. (1991). Nature, **352**,77–79.
9. Orr, H.T., Chung, M., Banfi, S., Kwiatkowski, T.J., Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P.W., and Zoghbi, H.Y. (1993). Nature Genetics, **4**, 221–226.
10. Koide,R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi,S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F., and Tsuji, S. (1994). Nature Genetics, **6**, 9–13.
11. Sutherland, G.R., Haan, E.A., Kremer, E., *et al.* (1991). Lancet, **338**, 289–292 .
12. Yu, S., Pritchard, M., Kremer, E., Lynch,M., Nancarrow, J., Baker,E., Holman,K., Mulley J.C., Warren, S.T., Schlessinger, D., Sutherland, G.R., and Richards, R.I. (1991). Science, **252**,1179–1181.
13. Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, G.R., and Richards, R.I. (1991). Science, **252**,1711–1714.
14. Verkerk, A.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P.A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F., Eussen, B.E., van Ommen G.B., Blonden, L.A.J., Riggins, G.J., Chastain, J.L., Kunst, C.B., Galjaard, H., Caskey, C.T., Nelson, D.L., Oostra, B.A., and Warren S.T. (1991). Cell, **65**,905–914

15. Knight, SJL., Flannery, AN., Hirst, MC.Campbell, L., Christodoulou, Z., Phelps, S.R., Pointon, J., Middleton-Price, H.R., Barnicoat, A., Pembrey, M.E., Holland, J., Oostra, B.A., Bobrow, M., and Davies, K.E. (1993). Cell, **74**, 127–134.
16. Imbert, G., Kretz, C., Johnson, K., Mandel, J. L., (1993). Nature Genet. **4**, 72–76.
17. Wells, R.D., and Sinden, R.R.(1993). Defined Ordered Sequence DNA, DNA Structure, and DNA-directed Mutation. pp.107–138. in Genome Analysis Volume 7: Genome Rearrangement and Stability. Davies, K.E., and Warren, S.T. *editor). Cold Spring Harbor Laboratory Press, Plainview, New York. USA.
18. Jansen, G., Willems, P., Coerwinkel, M., Nillesen, W., Smeets, H., Vits, L., Höweler, C., Brunner, H, and Wieringa, B.(1994). Am. J. Hum. Genet. **54**, 575–585.
19. Han, J., Zhu, Z., Hsu, C., and Finley, W.H. (1994). Antisense Research and Development, **4**, 53–65.
20. Brendel V., Beckman J.S., and Trifonov E.N. (1986). J. Biomol. Struct. Dyn. **4**, 11
21. Burge, C., Campbell, A.M., and Karlin, S. (1992). Proc. Natl. Acad. Sci. USA. 89, 1358–1362.
22. Genetics Computer Group, (1991). Program Manual for the GCG package, Version 7. 575 Science Drive, Madison, Wisconsin, USA 53711.
23. Phillips, G.J., Arnold, J., and Ivarie, R. (1987). Nucleic Acids Res. **15**, 2611–2638.
24. Sharp, P.M. (1986). Mol. Biol. Evol. **3**, 75–83.
25. Blaisdell, B.E. (1985). J. Mol. Evol. **21**, 278–288.
26. McClelland, M. (1986). J. Mol. Evol. **21**, 317–322.