An efficient method for finding repeats in molecular sequences

Hugo M. Martinez

Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA

Received 4 April 1983; Revised 23 May 1983; Accepted 31 May 1983

ABSTRACT

The problem of finding repeats in molecular sequences is approached as a sorting problem. It leads to a method which is linear in space complexity and N log N in expected time complexity. The implementation is straightforward and can therefore be used to handle large sequences with relative ease. Of particular interest is that several sequences can be treated as a single sequence. This leads to an efficient method for finding dyads and for finding common features of many sequences, such as favorable alignments.

INTRODUCTION

There are a number of significant problems in the analysis of molecular sequences (here regarded as strings over finite alphabets) which can be reduced to one of finding repeats. A necessary condition, for instance, that two or more sequences be homologous is that they contain a significantly large subsequence (substring) in common. If the sequences are placed end to end, then such a common subsequence is an instance of a special kind of repeat. Another example arises in the stem-oriented approach to determining the secondary structure of RNA. This approach first requires the finding of all the potential double helices (stems). Such helices are instances of dyads which in turn can be regarded as special repeats when the RNA sequence and its reverse complement are placed end to end. Still another important example occurs relative to searching for control signals. Sequences which are known to contain the signal can be placed end to end and the signal can be regarded as a special repeat.

Computer science has provided a number of solutions to the repeats problem, the most efficient of which appears to be the
one based on the concept of position trees (1). It can find repeats with an expected time complexity which is linear in the length of the sequence but whose worst case gives a time complexity which depends quadratically on the length unless special precautions are taken. Such worst cases also involve a space complexity which depends quadratically on the sequence length, though linearly in the expected sense. Additionally, there is the complication that delineating specific repeats involves cumbersome though straightforward tracing of paths in the tree. The method we propose is linear in space complexity for both the expected and worst case, and it is of time complexity $N \log N$ ($N =$ sequence length) in the expected sense. Worst cases can give a time complexity which is quadratic, but special methods can be used to spot the unusual examples in which they arise and reduce them to essentially the expected case. Further, the method involves no path tracing. The repeats are immediate and are reported during the process of sorting to be explained below. An entire tree does not have to first be constructed.

DESCRIPTION OF THE METHOD

Our approach is exceedingly simple and requires no more than the repeated application of a sorting algorithm. The overall speed is essentially determined by the speed of the sorting algorithm employed.

There is first constructed a sequence $P$ of pointers such that pointer value $P[i]$ is the location of the $i$th element in the sequence $S$. We then sort $P$ so that it constitutes an ordering of $S$. That is, $P[i] < P[j]$ or $P[i] > P[j]$ or $P[i] = P[j]$ according to whether $S[P[i]] < S[P[j]]$ or $S[P[i]] > S[P[j]]$ or $S[P[i]] = S[P[j]]$ respectively. With such a sorting of $P$ all the pointer values which point to the same kind of element in $S$ are grouped together. If there are $m$ letters in the alphabet over which $S$ is defined, then there will be at most $m$ groups of pointer values in this first sorting.

We next sort each of these groups of $P$ so that in the resulting subgroups two pointer values belong to the same one if and only if the elements immediately following the ones they point to are equal. Each of these subgroups is then sorted.
according to the elements twice removed from the elements pointed to, etc.

When no subgroups contain more than one pointer value the process is complete and there results an ordering of P with the following characteristic. Starting at each element of S there is a unique substring (sequence of contiguous elements) which distinguishes it from any other element of S; that is, no other element of S is the start of such a substring. These substrings can be lexicographically ordered and it is precisely this ordering which the final ordering of P represents.

The repeats are generated during the sorting procedure in the following manner. Suppose that we have just produced a group of pointer values and that it is the result of k sorts. The elements pointed to then have the property that the substrings of length k of which they are the start are instances of a repeat of length k provided that it cannot be extended in length. To be extended in length means that an additional sorting of the group of pointers does not break it up into subgroups. Thus, every time a group of pointer values breaks up it signals the finding of a repeat, and where it occurs are the pointer values in that group.

A distinct merit of this approach is that no significant storage space is required beyond that necessary for the sequence S and its pointer sequence P. We also note that the sorting scheme is very similar to how one would go about the lexicographic ordering of a finite number of sequences defined over a finite alphabet. If there are m elements in this alphabet we would first group the sequences (pointers to them) into at most m groups G1,G2,...,Gm. The pointers in group G1 are those pointing to the sequences beginning with the first letter of the alphabet, those in group G2 are the ones pointing to the sequences starting with the second letter of the alphabet, etc. Each of these groups is now independently divided into at most m groups G11,G12,...,G1m such that G11 are those pointers in group G1 which point to sequences whose second element is the first letter of the alphabet, those in G12 are the ones of G1 pointing to sequences whose second element is the second letter of the alphabet, etc. If the average length of the sequences is k and if
there are n sequences, then the expected time to accomplish their full lexicographic ordering is just k(cN), in which cN is the time required to sort the N elements into m groups.

The basic difference between this sorting scheme and the one used for obtaining the repeats is that we do not have separately defined sequences. There is just one. But if its length is N, then it is as though we had N sequences of average length \( \log_m M \). To see this equivalence, we note that the first sorting of the N elements results, on the average, in m groups of size N/m. If it takes time cN to sort N elements into m groups, then it will take time cN to sort all of the m groups of size N/m. This second sorting results, on the average, in \( m^2 \) groups of size \( M/m^2 \), for which the combined resorting time will again be cN, etc. The highest power k for which \( N/m^k \) > 1 sets an upper limit to the number of such full sorts. We can therefore take k as \( \log N \) to the base m.

IMPLEMENTATION CONSIDERATIONS AND APPLICATIONS

A considerable improvement in speed can be achieved with a little pruning. For example, a group of element positions which can be extended backwards must necessarily be included in a set of repeats obtained by the forward extension rule. We therefore test a group of potential repeats for backward extendability prior to its resorting. If the test is positive it is disregarded thereafter.

In the case of finding repeats in multiple sequences we take advantage of the constraint that a potential repeat group must necessarily contain an instance of the repeat from all of the sequences. Sizeable groups which would otherwise require repeated resorting can thus be eliminated.

We have constructed four implementations of the algorithm, called qrepeats, qdyads, qstems and qalign. The first finds all the repeats in one or more sequences, the second is qrepeats specialized to accept but one sequence, construct its reverse complement, and then find the repeats common to these two sequences. Attention in qdyads is also given to filtering out duplicates which necessarily arise because of the mirror image effect produced by working with the reverse complement of a sequence.
The third implementation, qstems, is qdyads specialized to allow for internal U-G base pairing as required for secondary structure in RNA. The programs qdyads and qstems are necessarily geared to the DNA and RNA alphabets, but qrepeats is quite general. No alphabet need be specified. This is achieved by using a sorting algorithm which does not depend upon sequence structure imposed by an alphabet. The slight loss in speed is, for the most part, amply compensated for by the increased flexibility.

Another feature of qrepeats is to allow for intersymbol spacing. For instance, we normally regard repeats as referring to a sequence of contiguous elements which occurs at more than one place. But it is sometimes important to consider not strict contiguity but also a sequence of 'every other element' and hence a single spacing between elements. An appropriate parameter is used to select any periodic spacing desired. This flexibility gives a direct implementation of searching for control signals consisting, for instance, of sequences whose elements are a helical turn apart. But in addition it provides a means of investigating complex repeat structures viewed as the combination of simple, periodic ones.

The fourth implementation, qalign, offers a somewhat new approach to the general problem of determining to what extent two or more sequences are homologous. This means that simultaneous alignments must allow for insertions and/or deletions. Our approach to this problem is a generalization of the approach to the RNA secondary structure problem which first finds the potential stems and then pieces them together to find a combination which minimizes the total free energy. Thus, we first find the potential common substrings and then find compatible combinations of these whose total length is as large as possible. This latter optimization is equivalent to a "shortest path" kind of a problem and therefore has a solution of time complexity which is the number of common substrings squared (1). In applying the qalign program to a specific problem the option is given of selecting the minimum length which is to be allowed for a common substring. This gives control over what are to be regarded as statistically significant common substrings and hence on the number of common substrings from which to select compatible combinations. The
optimization algorithm employed allows for the weighting of gaps, if desired, and mention should also be made of the option for finding near optimal alignments. The specific implementation of this latter feature finds all the alignments which lie within a specified distance of the optimal alignment as measured in units of the optimizing function (such as total number of matches in an alignment).

We have run a number of tests as to actual speed. Typical figures obtained on a VAX 11/750 computer are:

<table>
<thead>
<tr>
<th>Seq. length</th>
<th>qrepeats</th>
<th>qdyads</th>
<th>cNlog₄N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>4.0 secs</td>
<td>16.0 secs</td>
<td>4.0 secs</td>
</tr>
<tr>
<td>2048</td>
<td>9.4</td>
<td>27.9</td>
<td>8.8</td>
</tr>
<tr>
<td>4096</td>
<td>22.1</td>
<td>103.4</td>
<td>19.2</td>
</tr>
<tr>
<td>8192</td>
<td>51.9</td>
<td>136.2</td>
<td>41.6</td>
</tr>
</tbody>
</table>

The qrepeat figures refer to repeats of length 5 or greater and those for qdyads to dyads for which each half is of length 5 or greater. The column headed cNlog₄N is obtained from the qrepeats column by assuming the 4.0 second figure to correspond to cNlog₄N. The constant c is thus evaluated and used to calculate what the computation speeds would be at the remaining sequence lengths if the expected time were cNlog₄N. That the actual computation times are more than this is attributed to the use of a sorting algorithm which has an expected time complexity of NlogN rather than cN.

The implementations are written in the C language. They are available separately or as part of the UCSF Biomathematics Computation Laboratory sequence analysis package, which provides a comprehensive set of programs geared to a UNIX operating system environment.

ACKNOWLEDGEMENT

I am indebted to Professor Frederick Blattner of the University of Wisconsin for his very helpful suggestions during the initial implementations of the algorithm and for his making a 27 kilobase DNA sequence available as an impetus to develop efficient methods for handling large sequences.

This work was in part supported by NSF grant # PCM 8022006.

REFERENCES

Gamma thalassemia resulting from the deletion of a γ-globin gene

P.K. Sukumaran
Bai Jerbai Wadia Hospital for Children, Parel, Bombay, India, and
Department of Cell and Molecular Biology*, Medical College of Georgia, Augusta, GA 30912, USA

Received 15 March 1983; Revised and Accepted 28 May 1983

ABSTRACT

The first example of a deletion of one of the two γ globin genes has been characterized through an analysis of the DNA of the heterozygous parent of a homozygous newborn, using restriction endonuclease mapping techniques. A deletion of ∼5 kb was observed which was probably caused by an unequal crossing-over between the -Gγ- and -Aγ- genes resulting in the formation of a -GγAγ- hybrid gene. Data on proportions of Gγ and Aγ chains in newborn babies assumed to be heterozygous for the hybrid and normal genes suggest that this hybrid gene may be producing its Aγ chain at levels normally seen only for the Gγ chain.

INTRODUCTION

In newborn babies the Gγ and Aγ types of γ chain of fetal hemoglobin (Hb F) are present in a ratio of 7 to 3. This ratio changes after birth and becomes about 4 to 6 at the age of 3 to 4 months (1,2). The two γ globin genes which produce these two types of chain are part of the ε-Gγ-Aγ-ψ81-δ-β gene cluster on chromosome 11 (3). The 5' ends of the two γ genes are about 5.0 kb apart (4).

Recently we observed a reduced Gγ to Aγ ratio of 4 to 6 in the Hb F of over 30 healthy newborn babies, mainly from Japan, Mainland China, India, and Yugoslavia. These may be heterozygous for a normal β-globin gene cluster and one which produces no Gγ chains, and indeed, one Chinese and one Indian baby each had Hb F with only α and Aγ chains (5). These two infants were clinically normal but had low levels of Hb F at birth (about 55%), which was almost completely replaced within 2 to 3 months by Hb A and Hb A2. The parents of both babies were normal.

Nine infants with the low Gγ to Aγ ratio had an additional heterozygosity for the AγT allele. The AγT variant is characterized by an Ile-Thr substitution at position γ75, and is a relatively frequent variation observed in numerous populations (6,7,8). The Indian baby with only α and Aγ chains in its Hb F had exclusively the AγT variant.

Two alternative hypotheses were proposed (5) to account for $G_Y$ chains being present in a low proportion or absent altogether: The presence on one or both chromosomes of a $g$-globin gene cluster with two $A_Y$ genes (the $A_Y^*A_Y$ genic arrangement), or a $G_Y^{O}$-thalassemia (the $A_Y^*A_Y$ genic arrangement).

The present study reports a test of these hypotheses by means of restriction endonuclease analyses of the DNA of one parent of the Indian newborn whose fetal hemoglobin had $A_Y^T$ and no $G_Y$ chains. Our data suggest that the lack of $G_Y$ chain production results from a form of $G_Y^{O}$-thalassemia, in which the normal $G_Y^*A_Y$ genic arrangement is replaced by a gene which produces $A_Y$ chains but which retains restriction mapping similarities with the $G_Y$ gene. We propose that such a hybrid $-G_YA_Y^*$ gene has resulted from homologous but unequal crossing-over between $G_Y$ and $A_Y$ genes, such as is thought to have occurred for the $G_Y^*$ hybrid chains of the Lepore hemoglobins (9,10) and for the $A_Y^*$ hybrid chain of Hb Kenya (11,12).

**MATERIALS AND METHODS**

**Patient.** The subject (FM) was the father of India baby #147 (5); at birth this baby had Hb F with only $\alpha$ and $A_Y^T$ chains. This 35-year-old male and his 24-year-old wife, who is also his first cousin, had a normal hematology with normal levels of Hb A2 and Hb F. This family comes from Karnataka State in India and belongs to the Hindu Schedule Caste (Madhiga). Attempts to isolate a sufficient quantity of Hb F from blood samples of both parents were not successful, preventing the characterization of the Hb F in these adults.

**Restriction Endonuclease Analyses.** About 40 ml of blood was collected from the father and from a normal control with EDTA as anticoagulant. DNA was isolated from the white cells using a modification of the procedure described by Poncz et al (13). Samples of 6 to 10 $\mu$g of DNA were digested with various restriction endonucleases as listed in Table I. The fractions were separated by electrophoresis on vertical agarose gels (0.8%) overnight at 1.5 V.cm$^{-1}$, and at 4°C. Transfer to nitrocellulose paper followed the technique of Southern (14). The filters were hybridized to various probes which were labelled with $^{32}$P by nick translation using a commercial kit. Four different probes were used; plasmids containing the fragments were all generously provided by Dr. Oliver Smithies and collaborators, Madison, Wisconsin. They included: (a) the 460 bp Bam HI/Pvu II $G_Y^*$IVS-II probe which is specific for the second intervening sequence (IVS) of both the $A_Y$ and $G_Y$ genes; (b) the 1000 bp Bam HI/Eco RI $\beta$IVS-II probe which is specific for the second IVS of the $\beta$ gene; (c) the $\delta$IVS-II probe which is specific for the second IVS of the
δ gene; (d) the γδ 1.6 BX probe, a 1.6 kb fragment containing the ψ81 pseudo-
gene obtained through digestion with Bgl II and Xba I. All probes except
δIVS-II were purified fragments obtained by digestion of the plasmid DNA with
the appropriate enzymes and polyacrylamide gel electrophoresis (15). Hybridi-
zation was carried out for 17 to 20 hours at 68°C in 3 x SSC, 0.1% SDS. The
nitrocellulose filters were washed for several hours with 3 X SSC and 0.5% SDS,
pH 7.0, at 68°C.

A few filters were rehybridized after denaturation for 4 minutes followed
by neutralization for 4 minutes. Radioactive bands were detected by auto-
radiography using XAR-5 Kodak X-ray film and DuPont Cronex Lightning Plus
intensifying screens. Differences in intensity of some bands were determined
by measuring the optical densities of the band patterns, using an Ortec Model
4310 densitometer.

RESULTS.

When the DNA from subject FM was digested with several enzymes and then
hybridized to the four different probes, normal band patterns were obtained
in all instances except for two digests (Table I). The abnormal bands were
present in the Bgl II and Bcl I digests and were obtained after hybridization
to the GγIVS-II probe.

Figure 1 illustrates that the abnormal fragment was seen for the Bgl II
digest after hybridization to the GγIVS-II probe, while only normal bands were
observed for the Bam HI and Xba I digests. The abnormal band was 7.7 kb, which

TABLE I The sizes of gene fragments identified in restriction en-
donuclease digests of DNA from the father of the propositus (FM)
and from a normal control.

<table>
<thead>
<tr>
<th>Type of Probe</th>
<th>Enzyme</th>
<th>FM</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>GγIVS-II</td>
<td>Bgl II</td>
<td>12.8; 7.7</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>Bcl I</td>
<td>20; 15</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Eco RI</td>
<td>7.2; 2.7</td>
<td>7.2; 2.7</td>
</tr>
<tr>
<td></td>
<td>Bam HI</td>
<td>15.5; 5.0</td>
<td>15.5; 5.0</td>
</tr>
<tr>
<td></td>
<td>Xba I</td>
<td>7.5; 5.0; 3.7</td>
<td>7.5; 5.0; 3.7</td>
</tr>
<tr>
<td></td>
<td>Bgl I</td>
<td>5.0; 3.1</td>
<td>5.0; 3.1</td>
</tr>
<tr>
<td></td>
<td>Hind III</td>
<td>7.5; 6.9; 3.4</td>
<td>7.5; 3.4</td>
</tr>
<tr>
<td>Gδ1.6 BX</td>
<td>Bam HI</td>
<td>15.5</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>Xba I</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>Bgl II</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>δIVS-II</td>
<td>Bam HI</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Xba I</td>
<td>10.8</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>Bgl II</td>
<td>7.7</td>
<td>7.7</td>
</tr>
<tr>
<td>δIVS-II</td>
<td>Eco RI</td>
<td>5.4</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>Bam HI</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>Xba I</td>
<td>10.8</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>Bgl II</td>
<td>5.2</td>
<td>5.2</td>
</tr>
</tbody>
</table>
Figure 1. DNA fragments obtained with the restriction endonucleases Bam HI (B), Xba I (X) and Bgl II (Bgl) and the GγIVS-II and γδ 1.6 BX probes. FM is the heterozygous father of a homozygous newborn; N is a normal control.

is about 5 kb less than the normally occurring band of about 13 kb. Rehybridization of the same filter to the γδ 1.6 BX probe showed normal banding patterns only (Figure 1).

The fact that the heterozygote FM showed only bands of normal size, except when the fragment should have included both Gγ and Aγ genes, suggested a deletion which went from a point in the Gγ gene to the precisely corresponding point in the Aγ gene. This possibility is reinforced by the size of the deletion (~5 kb), since 5 kb would be deleted if this assumption were correct (16).

To test this idea, and to eliminate the alternative possibility that the abnormal Bgl II band was due to a previously unreported polymorphism, a digest with Bcl I, expected to produce a fragment containing both Gγ and Aγ genes in normal DNA, was analyzed. Figure 2 shows that the normal DNA, digested with Bcl I, yielded only one fragment of about 20 kb after hybridization with the GγIVS-II probe, but the DNA of FM showed two bands, namely the normal one and one about 5 kb smaller. This is consistent with the 5 kb deletion suggested by the Bgl II digest. Figure 3 illustrates this proposed γ-gene deletion, compared to the normal γ-gene region, with respect to restriction sites relevant to the data for enzymes Bam HI, Bcl I, Bgl II, Eco RI and Xba I.

Additional significant observations were:

(i) The Hind III digest, followed by hybridization with the GγIVS-II probe,
Figure 2. DNA fragments obtained with three restriction endonucleases as listed, and the \( G_y \) IVS-II probe. FM is the heterozygous father of a homozygous newborn; N is a normal control.

showed three bands of about 7.5, 6.9, and 3.4 kb (Figure 2). The two larger fragments correspond to the two \( G_y \) fragments (of 8 and 7.2 kb, respectively) that occur because of a Hind III restriction site polymorphism (17,18). This was shown by a separate experiment, in which a Hind III digest of DNA of FM was compared with that of a control known to be heterozygous for the Hind III site in the \( G_y \) gene: both samples gave the same size \( \gamma \)-gene bands, of 7.9 and 7.2 kb, due to \( G_y \), and of 3.4 kb, due to \( A_y \) (data not shown). Note that DNA

Figure 3. Comparison of the normal \( \gamma \)-globin gene region with the \( \gamma \)-gene deletion. A partial restriction map of this region is shown for enzymes Bam HI (B), Bcl I (Bc), Bgl II (Bg), Eco RI (E), and Xba I (X); additional sites for these enzymes, which are not detected by the genomic mapping procedures of this paper, are not shown. Data for the normal gene map are from reference 20.
sequencing studies show the precise size of the \(A_\gamma\) fragment to be 3.3 kb (16). (ii) The Eco RI digest contained two fragments of normal size (7.2 and 2.7 kb) which hybridized to the \(G_\gamma\) probe (Figure 2). However, the relative intensities of the two bands differed between the DNA of FM and that of the control. Figure 4 shows densitometric tracings from the same autoradiograph; maximally-exposed film gave about 1.5 times the intensity as the maximum of these tracings. For the normal control, the peak areas of the 2.7 kb and 7.2 kb bands were nearly identical, while for FM the area of the 2.7 kb band was only about 60% that of the 7.2 kb band. The band of reduced intensity is that for the \(A_\gamma\) gene, which implies that FM may have the normal diploid complement of \(G_\gamma\) genes. Thus, restriction enzyme data with Hind III and Eco RI suggest that FM has two \(G_\gamma\) genes. This contrasts with the protein quantitation data on his homozygous child (5), which indicate that FM may have only one \(G_\gamma\) gene.

DISCUSSION

The data obtained for the Bgl II and Bcl I digests after hybridization to the \(G_\gamma\)Ivs-II probe indicate a deletion of a segment of DNA of some 5 kb which apparently contains one of the \(\gamma\) genes. Hybridizations to the other three probes revealed normal band patterns only. The size of the deletion corresponds remarkably well with the size of DNA which separates any particular point of the two \(\gamma\) genes. The deletion of one of the two \(\gamma\) genes is also supported by the Eco RI data which show quantitative differences between the 7.2 kb and 2.7 kb fragments for FM versus the control.

Most data indicate that the remaining \(\gamma\) gene is the \(G_\gamma\) gene. For instance, the presence of three Hind III fragments, two of which are specific for the \(G_\gamma\) gene, suggests the existence of at least one \(G_\gamma\) gene on each chromosome. The finding of only one \(A_\gamma\)-specific Hind III fragment is consistent with the possibility that only one \(A_\gamma\) gene is present in FM, although it does not exclude the presence of two such genes. Moreover, the decreased density of the 2.7 kb Eco RI fragment (Figure 4) suggests the deletion of one of the \(A_\gamma\) genes.

Protein data, however, suggest that \(G_\gamma\) is deleted. The homozygous child of FM produced Hb F with only \(\alpha\) and \(A_\gamma^T\) chains, but no \(G_\gamma\) chains. The DNA restriction enzyme analyses and the protein data may be reconciled by assuming that the \(\gamma\)-thalassemia-like condition arose through an unequal but homologous interchromosomal (19) or intrachromosomal (4) crossing-over between \(G_\gamma\) and \(A_\gamma\) genes. Such a crossover would have deleted 5 kb of DNA and resulted in a single \(\gamma\) gene.

The region within which the crossover may have occurred is suggested by

4640
Figure 4. Densitometry tracings of the Eco RI bands shown in Figure 2, indicating a decreased intensity in the 2.7 kb band of FM, who is the heterozygous father of a homozygous newborn; N is a normal control.

The data of Figure 4, which show that the γ-thalassemic DNA yields only the 7.2 kb γ-gene fragment upon Eco RI digestion. The crossover must therefore have occurred between the Eco RI site 1.2 kb 5' to the transcribed portion of the Aγ gene (16) (which site is missing in γ-thalassemic DNA), and the codon for position 136 of the amino acid sequence (which is responsible for the Gly→Ala difference between Gγ and Aγ chains). The high degree of homology between Gγ and Aγ genes extends only about 500 bp 5' to the cap site (16), so that the likeliest location for the crossover between Gγ and Aγ is within the transcribed portion of the γ gene 5' to the codon for γ136. We therefore propose that the unequal but homologous crossover produced a -GγAγ- hybrid gene, whose 5' section derived from the Gγ gene and 3' end from the Aγ gene.

The product of the -GγAγ-hybrid gene would be an Aγ chain, if the 3' end included the region of the Aγ gene with the codon for alanine at position γ136. In this particular case, the unusual gene produces the AγT chain with threonine at position γ75. The Ile→Thr mutation could have occurred after the formation of the -GγAγ-hybrid gene or, perhaps more likely, the Aγ gene that participated in the crossover was AγT. In the latter event, the crossover must have occurred 5' to the codon for position γ75.

If one assumes that the -GγAγ-hybrid gene produces Aγ at a level normally seen for Gγ, as the result of its partial Gγ nature in some undefined way, then heterozygotes for the hybrid gene and the normal γ-globin genic arrangement would produce a 4:6 ratio of Gγ:Aγ as newborns. This low Gγ:Aγ ratio has been observed in a number of Asian babies (5), and we hypothesize that it results from the heterozygous condition for this γ-thalassemia.

It appears probable that the -GγAγ-δ-β-/ haplotype is complementary to the Gγ-Gγ-Aγ-δ-β/ haplotype described by Trent et al (20). It might well be
that the second \( \gamma \) gene in this condition is in fact an \(-\Delta y\gamma\)-hybrid gene, and data provided by these authors do not exclude this possibility.

ACKNOWLEDGEMENTS

One of us (PKS) is grateful to Dr. S.M. Merchant, Dean, Bai Jerbai Wadia Hospital for Children for granting study leave which enabled him to work in the U.S.A. He is also thankful to Dr. S.N. Naik, Dr. J. Akolkar, Miss H.R. Master, Mr. S.S. Desai, and Mr. D.S. Chavan, Cancer Research Institute, Tata Memorial Centre, Bombay, for their help in obtaining and the initial processing of the blood sample. This research was in part supported by USPHS Research Grant HLB-05168.

*Contribution #0762

REFERENCES