Am. J. Hum. Genet. 55:777–787, 1994

Toward Fully Automated Genotyping: Allele Assignment, Pedigree Construction, Phase Determination, and Recombination Detection in Duchenne Muscular Dystrophy

Mark W. Perlin,[1] M. Blaine Burks,[1] Rita C. Hoop,[2] and Eric P. Hoffman[2]

[1]School of Computer Science, Carnegie Mellon University, and [2]Departments of Molecular Genetics and Biochemistry, Human Genetics, and Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh

Summary

Human genetic maps have made quantum leaps in the past few years, because of the characterization of >2,000 CA dinucleotide repeat loci: these PCR-based markers offer extraordinarily high PIC, and within the next year their density is expected to reach intervals of a few centimorgans per marker. These new genetic maps open new avenues for disease gene research, including large-scale genotyping for both simple and complex disease loci. However, the allele patterns of many dinucleotide repeat loci can be complex and difficult to interpret, with genotyping errors a recognized problem. Furthermore, the possibility of genotyping individuals at hundreds or thousands of polymorphic loci requires improvements in data handling and analysis. The automation of genotyping and analysis of computer-derived haplotypes would remove many of the barriers preventing optimal use of dense and informative dinucleotide genetic maps. Toward this end, we have automated the allele identification, genotyping, phase determinations, and inheritance consistency checks generated by four CA repeats within the 2.5-Mbp, 10-cM X-linked dystrophin gene, using fluorescein-labeled multiplexed PCR products analyzed on automated sequencers. The described algorithms can deconvolute and resolve closely spaced alleles, despite interfering stutter noise; set phase in females; propagate the phase through the family; and identify recombination events. We show the implementation of these algorithms for the completely automated interpretation of allele data and risk assessment for five Duchenne/Becker muscular dystrophy families. The described approach can be scaled up to perform genome-based analyses with hundreds or thousands of CA-repeat loci, using multiple fluorophors on automated sequencers.

Received April 20, 1994; accepted for publication June 2, 1994.

Address for correspondence and reprints: Dr. Mark Perlin, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.1994 94/5504-0000

0002-9297/94/5504-0021$02.00

Introduction

A primary goal of the NIH/DOE Human Genome Project during its initial 5-year phase of operation was to develop a genetic map of humans, with markers spaced 2–5 cM apart (Hoffman 1994). This task has already been largely accomplished in half the time anticipated, with markers that are far more informative than originally hoped for (Weissenbach et al. 1992). In these new genetic maps, RFLP loci have been entirely replaced by CA-repeat loci (dinucleotide repeats) (Weber and May 1989). One of the advantages of CA-repeat loci is their high density in the genome, with approximately 1 informative CA repeat every 50,000 bp: this permits a theoretical density of ~20/cM. Another advantage of CA-repeat polymorphisms is their informativeness, with most loci in common use having PIC values of >.70 (Weissenbach et al. 1992). Finally, these markers are PCR based, permitting rapid genotyping using minute quantities of input genomic DNA. Taken together, these advantages have facilitated linkage studies by orders of magnitude: a single full-time scientist can cover the entire genome at a 10-cM resolution and can map a disease gene in an autosomal dominant disease family in ~1 year (Stephan et al., in press).

The CA-repeat–based genetic maps are not without disadvantages. First, alleles are detected by size differences in PCR products, which often differ by as little as 2 bp in a 300-bp PCR product. Thus, these alleles must be distinguished using high-resolution sequencing gels, which are more labor intensive and technically demanding to use than most other electrophoresis systems. Second, CA-repeat loci often show secondary "stutter" or "shadow" bands in addition to the band corresponding to the primary allele, thereby complicating allele interpretation (fig. 1). These stutter bands may be due either to errors in *Taq* polymerase replication during PCR or to secondary structure in PCR products. Allele interpretation is further complicated by the differential mobility of the two complementary DNA strands of the PCR products when both are labeled. Finally, sequencing gels often show inconsistencies in mobility of DNA fragments, making it difficult to compare alleles of individuals, between gels and often within a single gel. The most common experimental ap-

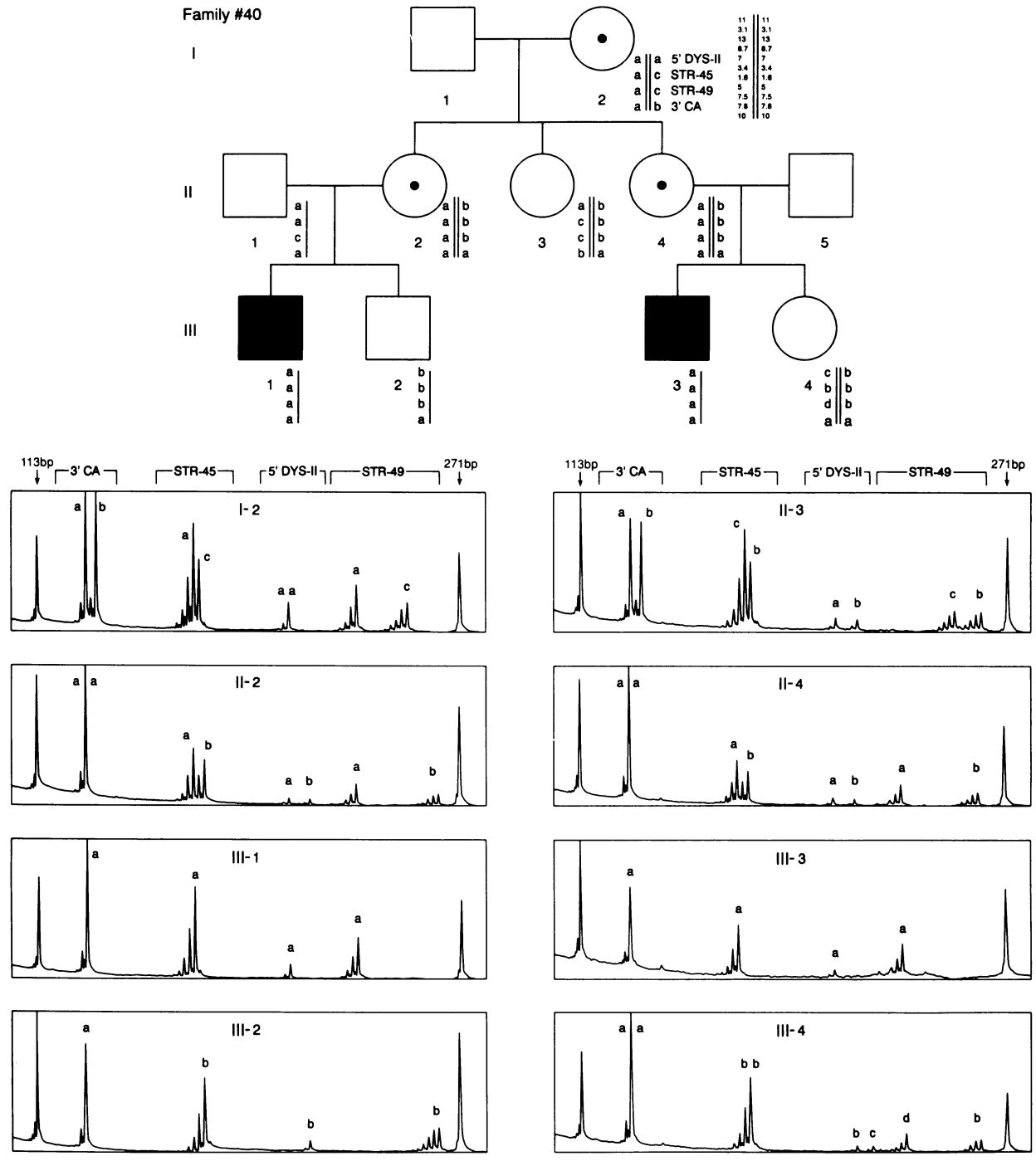

Figure 1 PMT voltage-vs.-time data used for input into automated genotyping. Shown is a BMD family (*top*), with representative lane data from the automated sequencer (*bottom*). Multiplex fluorescent CA-repeat analysis was done as described elsewhere (Schwartz et al. 1992). The time windows corresponding to each of four dinucleotide repeat loci are shown above the data traces. The four dystrophin gene CA-repeat loci show the full range of different patterns observed with most CA repeats: 3'CA shows very clean, distinct alleles but is not very informative, whereas STR-49 and STR-45 show complex patterns of six or seven peaks for each allele. (Reprinted from Schwartz et al. 1992, with permission)

proach used for typing CA-repeat alleles involves incorporation of radioactive nucleotide precursors into both strands of the PCR product. The combined consequence of stutter peaks and visualization of both strands of alleles differing by 2 bp often leads to considerable noise on the resulting autoradiograph signals, which then requires careful subjective interpretation by an experienced scientist, in order to determine the true underlying two alleles. Tri- and tetranucleotide repeat loci show substantially cleaner signals; however, they do not occur as frequently in the

genome, and maps based on these markers have not reached the density of dinucleotide repeats.

A corollary of highly dense and informative genetic maps is the need to accurately acquire, analyze, and store large volumes of data on each individual or family studied. For example, a genomewide linkage analysis on a 30-member pedigree at 10-cM resolution would generate data for ~30,000 alleles, with many markers showing ≥5 alleles. Currently, alleles are visually interpreted and then manually entered into spreadsheets for analysis and storage. This approach requires a large amount of time and effort and introduces the high likelihood of human error. Moreover, future studies of complex multifactorial disease loci will probably require large-scale genotyping on hundreds or thousands of individuals. Each of these features suggests that automation of genotype data generation, acquisition, interpretation, and storage is required to fully utilize the developing genetic maps. Some effort has been made to assist in allele identification and data storage (ABI Genotyper software); however, this software still requires substantial user interaction to place manually assigned alleles into a spreadsheet, and it is unable either to deconvolve (and hence cannot accurately genotype) closely spaced alleles or to perform other needed analyses.

The Duchenne/Becker muscular dystrophy (DMD/BMD) gene locus (dystrophin gene) (Monaco et al. 1986; Koenig et al. 1987) provides an experimental system in which to test the feasibility of automation of genetic analysis. The dystrophin gene can be considered a minigenome: it is by far the largest gene (2.5 Mbp) known to date; it has a high intragenic recombination rate (10 cM, i.e., 10% recombination between the 5′ and 3′ ends of the gene); and it has a considerable spontaneous mutation rate (10^{-4} meioses). Mutation of the dystrophin gene results in one of the most common human lethal genetic diseases, and the lack of therapies for DMD demands that molecular diagnostics be optimized. The gene is very well characterized, with both precise genetic maps (Oudet et al. 1990) and physical maps (Burmeister et al. 1988). Finally, approximately one dozen CA-repeat loci distributed throughout the dystrophin gene have been isolated and characterized (Beggs and Kunkel 1990; Oudet et al. 1990; Clemens et al. 1991; Feener et al. 1991).

We have recently shown that many of the problems with interpretation of dystrophin gene CA-repeat allele data can be overcome by multiplex fluorescent PCR and data acquisition on automated sequencers (Schwartz et al. 1992). This approach uses fluorescently labeled PCR primers to simultaneously amplify four CA-repeat loci in a single reaction. By visualizing only a single strand of the PCR product, and by reducing the cycle number, we were able to eliminate much of the noise associated with these CA-repeat loci. Moreover, the production of fluorescent multiplex reaction kits provides a standard source of reagents, which, in our hands, have not deteriorated 3 years after the fluorescent labeling reactions were performed. In our previous report, alleles were manually interpreted from the automated sequencer traces (fig. 1) (Schwartz et al. 1992).

Here we report successful efforts to automate data acquisition and interpretation. We have designed and implemented computer software that successfully identified each of the dystrophin gene alleles in pedigree members; deconvolved complex "stuttered" alleles that differed by only 2 bp, where signal/noise is a particular problem; reconstructed the pedigrees from lane assignment information; set phase in females; propagated haplotypes through the pedigree; and identified female carriers and affected males in the pedigree on the basis of computer derivation of an at-risk haplotype. We also artificially introduced recombination events into this data and then designed and implemented software that was able to detect each recombination event and localize it to the correct female meiosis in the pedigree.

Methods

Multiplex CA Repeats

Four CA-repeat markers (3′-CA [Oudet et al. 1990], 5′DYSII (Feener et al. 1991), and simple tandem repeats (STRs) 45 and 49 [Clemens et al. 1991]) distributed throughout the 2.5-Mb dystrophin gene were used. The forward primer of each pair of PCR amplimers was covalently linked to fluorescein, and all four loci were amplified in a single 25-cycle multiplex PCR reaction as described elsewhere (Schwartz et al. 1992). The mixed fluorescent primers have been stored for >3 years, with no loss of label intensity, obviating the need for relabeling prior to each experiment. Two fluorescent molecular-weight standards (dystrophin gene exons 50 [271 bp] and 52 [113 bp] [Beggs and Kunkel 1990; Schwartz et al. 1992]) were added to samples prior to electrophoresis.

Allele Data Acquisition

The PCR products of the four CA-repeat loci lie in four nonoverlapping size windows, and the alleles for all four loci and the molecular-weight markers can be read out as a size-multiplexed signal in one lane of a DNA sequencer. We have used the DuPont Genesis DNA sequencer, which can generate fluorescent intensity data for 10–12 lanes, with one lane assigned to each individual. Thus, 10 family members can be haplotyped for the dystrophin gene by a single sequencer run. Each lane's signal intensity is observed as photomultiplier tube (PMT) voltage units (12-bit resolution) and is sampled by the sequencer every 3 s, providing ~20 data points/base of DNA. Gels were run for a total of 4 h, generating ~5,000 data points/lane (individual). Machine-readable data files from the sequencer runs, recorded as a linear fluorescence signal (PMT voltage) trace for each lane (individual), were automatically generated by the Genesis 2000 software. These time-ver-

sus-voltage files were input into our software, as described below.

Signal Processing

Each individual's preprocessed DuPont data file contains a time-versus-intensity trace of the multiplexed PCR sequencer run generated from the corresponding gel lane. For quantitative processing, these data must be converted to DNA size–versus–DNA concentration units. Our software first searches predetermined time regions to find the molecular-weight markers (dystrophin gene exons 50 [271 bp] and 52 [113 bp]). A linear interpolation is then performed to construct a time-versus-size mapping grid. Each predefined CA-repeat locus is then processed independently within its predefined size window. Every peak within the CA-repeat marker region is identified and is assigned a time and an area. The apex of a peak is defined as the point of change between a monotonically increasing series and a monotonically decreasing series, left to right. The monotonicity predicate holds when the difference between an average of right values and an average of left values exceeds a predetermined threshold. With the linear time-to-size interpolation from the grid, the time of each peak apex's occurrence is converted to a DNA size estimate. The areas are computed as the full width at half-max peak and are considered to be proportional to the approximate DNA concentration for any specific locus.

Allele Separation by Deconvolution

We mathematically *deconvolved* overlapping stutter peaks of proximate alleles at a locus, thereby computing a single peak per allele. For any given marker, the allele stutter pattern is relatively fixed. The DNA concentrations for one allele at each discrete DNA size can be written as the pattern vector $<p_n, \ldots, p_2, p_1, p_0>$, or, equivalently, as the polynomial $p(x)$, $p(x) = p_n x^n + \ldots + p_2 x^2 + p_1 x + p_0$. Each coefficient p_k is the observed peak area in the allele's pattern for stutter peak n.

The superimposed stutter patterns observed in the sequencer data of heterozygotic markers can be similarly described by a polynomial, $q(x)$. The coefficients of $q(x)$ are the superimposed peak areas produced by PCR stuttering of the two alleles. The PCR stutter of each allele has a fixed pattern described by the polynomial $p(x)$. When the allele contains precisely r repeated dinucleotides, the pattern is shifted $2r$ bases on the sequencer gel lane. (With a tri- or tetranucleotide repeat, the pattern would be shifted $3r$ or $4r$ bases, respectively, and similar analyses would apply.) A $2r$-base shift in the stutter pattern mathematically corresponds to multiplication of the polynomial $p(x)$ by x^{2r}. Therefore, if the two allele sizes are s and t, then the two stuttered alleles produce the shifted polynomials $x^s p(x)$ and $x^t p(x)$, respectively. Superimposing these two allele stutter patterns produces the observed sum $q(x) = x^s p(x) + x^t p(x) = (x^s + x^t) p(x)$. Direct deconvolution to obtain the allele sizes s and t (hence, the genotype) by polynomial division via $q(x)/p(x) = x^s + x^t$ is not sufficiently robust with real data containing noise. Therefore, we devised a method based on statistical moment computations. Further, since the computing time for these moments is just proportional to the size of the data (i.e., linear time), the algorithm is fast, and it is asymptotically faster than direct (and noise intolerant) polynomial division, which requires quadratic time. This speed advantage may prove important in on-line real-time automated genotyping.

The kth moment of a polynomial $u(x)$ is $u_k = u^{(k)}(1)$, where $u^{(k)}$ is the kth algebraic derivative of $u(x)$. u_k can be rapidly computed by weighted summation of the coefficients of $u(x)$'s kth derivative. As derived in the appendix, $s + t = (q_1 - 2p_1)/p_0$, $s^2 + t^2 = [(q_2 - 2p_2) + (s+t)(p_0 - 2p_1)]/p_0$, and $(s-t)^2 = 2(s^2+t^2) - (s+t)^2$. The first expression for $s + t$ implies that, given the hemizygous distribution $p(x)$, the stuttered sequencer data $q(x)$, and the size t of the larger allele (corresponding to the peak of the largest PCR product), the size s of the smaller allele can be computed. The last two expressions are used for simultaneously genotyping both alleles. Applying all three expressions, we can directly calculate the allele sizes as $s = [(s+t)+(s-t)]/2$ and $t = [(s+t)-(s-t)]/2$. This computation has the effect of deconvolving the superimposed PCR stutter patterns of the heterozygotic alleles into the two discrete peaks, having sizes s and t, needed for straightforward genotyping. The real numbers s and t are rounded (up or down) to the nearest integer occurring in the observed peak data.

Setting Phase by Graph Propagation

Once the genotypes have been determined for a DMD pedigree, phase can be readily set on the X chromosome. This is done by treating the pedigree as a graph, where the nodes are the individuals and the links are the inheritance paths between them. Starting from a male descendant (e.g., the proband), the neighboring nodes that are one inheritance link away (whether child or parent) are explored. Individual haplotypes are locally determined from haplotyped neighbors, as follows:

- Male individuals are given the haplotype of their hemizygotic genotype.
- Female individuals are set from a male neighbor by assigning one haplotype to the male's haplotype and assigning the second haplotype as the difference, at each marker, of the individual's genotype and the male haplotype.
- Female individuals are set from a haplotyped female neighbor by first determining which (if either) of the neighbor's haplotypes is contained within the individual's genotype. This haplotype becomes the first haplotype of the individual, and the second haplotype is obtained as the difference, at each marker, of the individual's genotype and the first haplotype.

Other local computations, such as assessing consistency, can be done when visiting each node. Since the graph tra-

versal only propagates to unhaplotyped neighbors, the algorithm terminates when all individuals have been consistently haplotyped.

Independent graph propagations from each male descendant are done. The propagation locally terminates at an individual when a parent-child haplotype inconsistency is detected. This early termination can suggest both where recombinations (or other events) occur in the pedigree and how to correct for their occurrence.

Determining Carrier and Affected Individuals

For the purposes of program development, we have assumed DMD/BMD to be a fully penetrant X-linked recessive disorder with a *low* mutation rate. Once the entire pedigree has been haplotyped, the carrier and affected individuals can then be inferred. If no recombination events are found, then the dystrophin gene haplotype of the proband serves as a signature that indicates an affected disease gene. Any male with this disease gene haplotype signature on his dystrophin gene who is reachable in the X-linked recessive pedigree graph is interpreted as affected. Similarly, reachable females shown to have the disease gene signature (on one chromosome) are considered carriers. Since our propagation algorithm only communicates between individuals who can directly transmit or receive an X chromosome (i.e., immediate parents or children), independent (i.e., unreachable) paternal lineages that coincidentally share the disease gene signature are not incorrectly phenotyped.

Software System and User Interface

Individual software modules were written for signal processing on the DuPont Genesis data, allele separation by deconvolution, haplotyping via graph propagation, and carrier/affected status determination. All are modules of a single computer program developed in Macintosh Common LISP. A color graphic interface was also developed for presenting the pedigree and for displaying the processing and results of genetics computations.

Results

Identification of Dinucleotide Repeat Alleles

The first step in the automated linkage analysis is to determine the genotypes of family members at each of the four intragenic dystrophin gene CA-repeat loci. The output from the DuPont automated sequencer is a data file containing PMT voltage as a function of time for each individual (lane) of the gel (fig. 1). The two molecular-weight markers and complex allele patterns for each of the four dinucleotide repeat alleles are seen as peaks in these data.

To automatically define genotypes at each dinucleotide repeat locus, the software first searches the time windows corresponding to the molecular-weight markers. Once these are identified, the software conducts a linear interpolation between these markers, to derive a time-versus-size (in nucleotides of DNA) mapping. Predefined DNA size windows for each of the four dinucleotide repeat loci are then superimposed on the quantitative PMT signal data, and all peaks within those windows are identified. Peak areas are automatically calculated. The program scans the window for each marker and assesses the *pattern* of peaks, to classify the peaks into one of three classes: hemizygote/homozygote alleles, distinct heterozygote alleles, or superimposed heterozygote alleles.

These three classes of peak patterns are defined as follows. A hemizygote/homozygote allele comprises a single decay pattern of decreasing peak amplitudes, with DNA size decreasing from right to left (fig. 1); the rightmost and largest peak is considered to be the primary peak. For example, individual III-1 of family 40 is a male hemizygote. At locus STR-45, for the values shown in table 1, the peak occurs at length 171 nucleotides, with a concentration of 101,299. Thus, the genotype of individual III-1 at locus STR-45 is assigned the value 171.

The peak pattern is classified as distinct heterozygote when (*a*) two such decay patterns are found within the marker window and (*b*) the two primary peaks are of similar amplitude. For example, individual II-2 of family 40 is heterozygotic at locus STR-49. As seen in table 1, there is one peak at length 233 and a second peak at length 264. The stutter peaks are widely separated, so the genotype was readily determined to be (233,264).

The third class, superimposed heterozygote alleles, is invoked when no simple pattern of alleles satisfying the hemizygotic/homozygotic or distinct-heterozygotic criteria is detected. In this class, present only in female heterozygotes, the alleles are closely spaced and produce a complex pattern of overlapping peaks. Deconvolution of the peak pattern is then invoked to identify the two alleles. Since the peak decay patterns are similar for any given locus, the deconvolution of a complex heterozygous pattern at a locus can be done with respect to the hemizygous decay pattern (of a different individual) at the same locus.

Consider, for example, the STR-45 locus of individual I-2 of family 40. The DNA concentrations at the PCR product sizes 161–173 are given in table 1. The sizes and concentrations can be represented by the polynomial

$$q(x) = 61326x^{173} + 94852x^{171} + 47391x^{169} + 18115x^{167} + 5896x^{165} + 1928x^{163} + 930x^{161}.$$

This pattern does not conform to a simple uniform decay. In family 40, individual III-1's hemizygotic locus STR-45 does (as expected) have a simple decay pattern from the peak at size 171 down through size 161, as seen in table 1. These data can similarly be represented by the polynomial

$$p(x) = 101299x^{171} + 55373x^{169} + 20799x^{167} + 7242x^{165} + 2171x^{163} + 821x^{161},$$

Table 1

Computed Base Size versus Peak Area, for Representative Individuals and Loci

Size	Individual III-1	Individual I-2
Marker STR-45:		
161	821	930
163	2171	1928
165	7242	5896
167	20799	18115
169	55373	47391
171	101299	94852
173	0	61326
175	0	0
		Individual II-2
Marker STR-49:		
221		843
223		1217
225		2360
227		6123
229		11469
231		26811
233		48135
234		0
236		0
238		0
240		0
242		0
244		0
246		0
248		0
250		0
252		1695
254		2877
256		5410
258		11553
260		17482
262		25866
264		28672

NOTE.—The DNA concentrations shown were detected and quantitated at every DNA length (rows) for each genotyped individual (columns). The peak area values were computed by the system, from the raw-data files used to generate the graphs in fig. 1, are in arbitrary units, and have been rounded to the nearest integer. Zero values denote minimal signal. The numbers illustrate the three classes of CA-repeat genotype data: hemizygote/homozygote alleles, distinct heterozygote alleles, and superimposed heterozygote alleles.

and can be used to help recover the two alleles at individual I-2's STR-45 locus.

As described in Methods, I-2's peak pattern at locus STR-45 can be viewed as the superposition of two shifted copies of III-1's peak pattern at STR-45. Mathematically, the observed $q(x)$ pattern is the sum of two shifted copies of $p(x)$: $q(x) = x^s p(x) + x^t p(x)$ or $(x^s+x^t)p(x)$. Deconvolution of $q(x)$ with respect to $p(x)$ can determine (x^s+x^t), where s and t are the peaks of the shifted patterns. That is, s and t provide the genotype. The polynomial coefficients are first renormalized to account for the expectation that $p(x)$ measures a single chromosome dosage, whereas $q(x)$ measures two doses. Then, using the polynomial moment technique detailed in Methods and shifting the sizes to their correct origin, we compute $s = 173.061$ and $t = 170.832$. Rounding these numbers to the closest integers in the peak pattern, we obtain the genotype (173,171).

This example result illustrates how PCR stutter peaks can be effectively exploited using our deconvolution approach to automatically resolve CA-repeat alleles that are close in size. The computed genotypes for all tested members of family 40 were obtained by the program at every locus by invoking and applying the appropriate method (i.e., hemizgote/homozygote, distinct heterozygote, or superimposed heterozygote) to the data (fig. 2).

Establishing Haplotype Inheritance and Phenotype

Once the genotypes have been determined, the risk can be directly assessed in this X-linked disease gene system. This is done by setting phase to determine haplotypes and then inferring the phenotype of each all individuals, from their haplotype(s). The phenotype is inferred by comparing the proband's signature haplotype with the haplotypes of other related individuals in the pedigree. The use of multiple informative markers assures that, with high probability in our system, identity by state of the multiple markers implies identity by descent. Thus, an identical signature at a related individual in the DMD pedigree implies a shared chromosomal segment, including the diseased dystrophin gene region(s). Males sharing an affected proband's signature are presumed to be affected, whereas females sharing this signature are presumed to be carriers.

An example of setting phase from the allele data is illus-

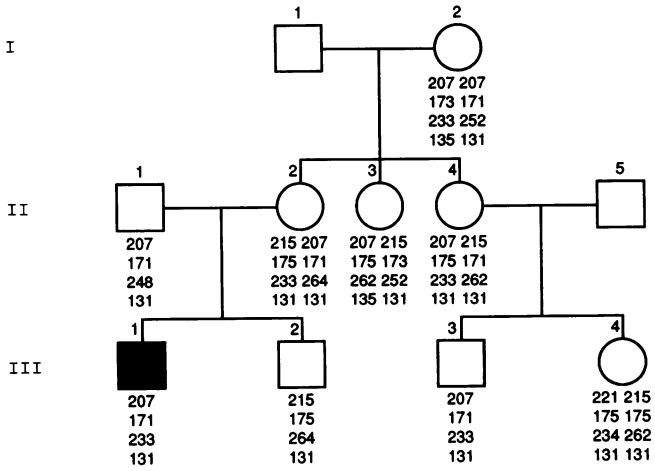

Figure 2 Output from the pedigree-construction and genotyping modules. Shown are the genotypes that the software automatically computed for each tested member of family 40 (fig. 1). The software automatically applied one of three methods (maximum of single peak, maxima of double peaks, or allele deconvolution) most appropriate to the locus data. This diagram was drawn by the graphic-display component of the system.

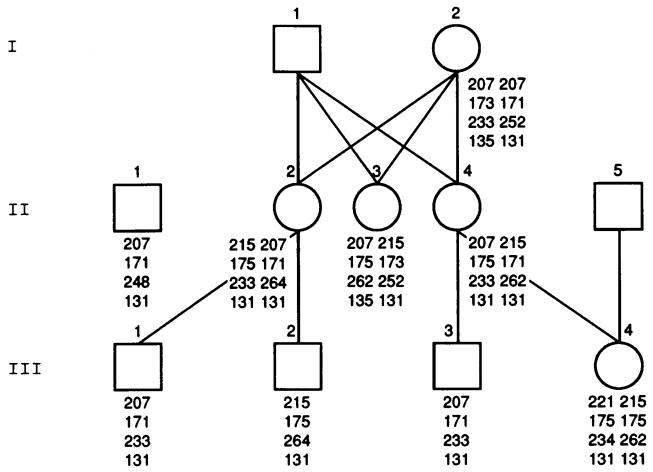

Figure 3 Inheritance graph construction and setting of phase. The links between the individuals in family 40 show the X-chromosome inheritance paths between parents and children. These links are traversed to generate the vertical, in-phase haplotypes shown. This is done by applying the haplotyping rules when graph nodes (i.e., individuals) are reached in the graph traversal. This diagram was drawn by the graphic-display component of the system.

trated with female individual II-2 and male proband III-1 from family 40. The genotype of II-2 across the four dystrophin markers 5 DYS-II, STR-45, STR-49, and 3-CA is the allele sequence (207,215), (171,175), (233,264), and (131,131). III-1's haplotype is 207,171,233,131. Extracting this haplotype from II-2's genotype leaves 215,175,264, 131; these two sequences describe II-2's two haplotypes.

The program applies the graph-propagation and consistency-analysis rules given in Methods to set phase for the entire pedigree. With the family 40 pedigree as an example, the graph traversal of the pedigree follows the X-chromosome inheritance links shown in figure 3 and generates the phase-known haplotypes shown for each individual. No inconsistencies are detected in family 40.

The program then determines phenotypes. In family 40, for example, proband III-1's allele signature at the four markers 5 DYS-II, STR-45, STR-49, and 3-CA is the allele sequence 207,171,233,131. All individuals in family 40 sharing this sequence on one of their haplotyped chromosomes are presumed to also share the affected proband's disease gene. Thus, individual III-3 is inferred to be another affected male, and individuals I-2, II-2, and II-4 are inferred to be carrier females. The phenotyped pedigree is shown in figure 4.

Identifying Recombinant Chromosomes

When a recombination event occurs, inconsistencies arise in the X-linked pedigree graph haplotype relationships between parents and children. These inconsistencies can be detected by the described methods and then can be used to localize the event within the pedigree.

When a recombination occurs, our straightforward rules for setting haplotype phase by taking set differences between sequences of alleles are no longer operable. For example, suppose that a recombination occurred between the STR-45 and STR-49 markers in a meiosis of individual I-2 in family 40. Then related individual II-2 would still inherit the paternal haplotype 215,175,264,131, but the maternal disease haplotype would be changed from 207,171,233,131 to 207,171,252,135. Further, the proband III-1 would then inherit from II-2 the haplotype 207,171,252,135. As described next, these changes propagate inconsistencies at certain points in the pedigree.

The program detects inconsistencies via early termination: it propagates the parent-child set-difference operation through the inheritance graph as completely as possible but terminates as soon as an inconsistency is detected. Thus, with the hypothetical recombination between loci STR-45 and STR-49 in grandparent I-2 of family 40, a propagation from the male grandchild III-1 or III-2 would correctly haplotype grandsons III-1 and III-2 and mother II-2 but would incorrectly haplotype grandmother I-2. At that point, I-2's haplotype would be inconsistent with the rest of the pedigree (i.e., individuals II-3, II-4, III-3, and III-4). This inconsistency would be detected by the local set-difference operations emanating from I-2, and the haplotyping computation would halt. The result of this partial computation is shown in the left-hand panel of figure 5.

These inconsistencies can be exploited to locate recombination events. If the same recombination example is continued with family 40, a propagation of the set-difference operation could also be initiated from grandson III-3, on the other half of the pedigree graph. This propagation correctly haplotypes individuals I-2, II-3, II-4, III-3, and III-4

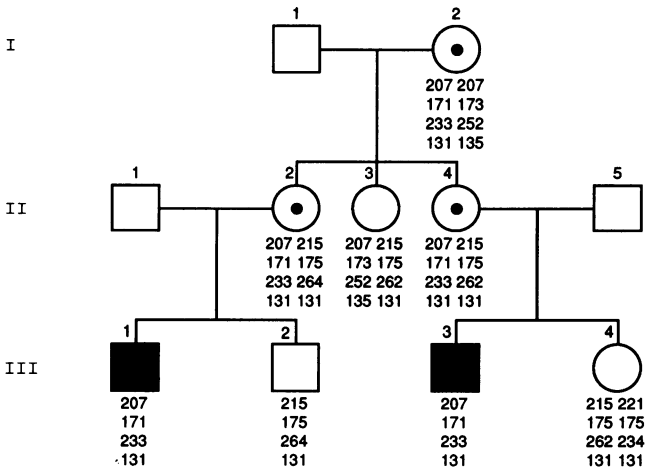

Figure 4 Identification of individuals having the at-risk haplotype. All individuals who share a chromosomal haplotype with proband III-1 are inferred to carry the disease gene. III-1's haplotype is the allele sequence 207,171,233,131. Male III-3 has this haplotype and is presumed to be affected. Females I-2, II-2, and II-4 have this haplotype on one of their X chromosomes and are inferred to be carriers. This diagram was drawn by the graphic-display component of the system.

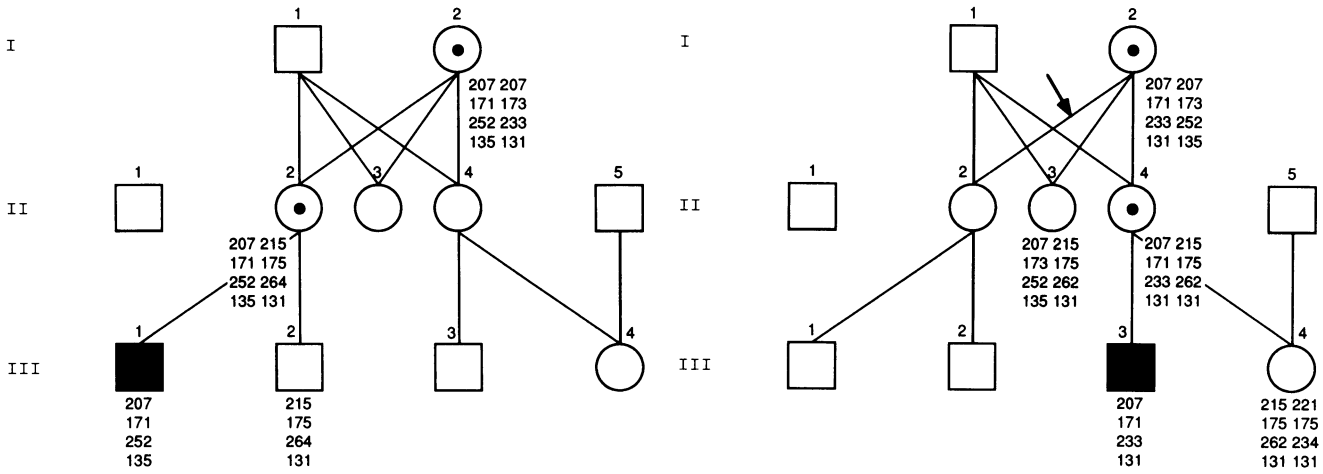

Figure 5 Detection of recombination events. Data on family 40 have been modified to simulate a recombination event between loci STR-45 and STR-49 in grandparent I-2, affecting individuals II-2 and III-1. Graph propagations are performed from grandsons until an inconsistency is detected; at that point, the graph traversal halts. Haplotypes are shown only for those individuals who are in the consistent region of the graph. These diagrams were drawn by the graphic-display component of the system. *Left,* Results of propagating from grandson III-1. *Right,* Results of propagating from grandson III-3. The arrow, indicating the inheritance link between individuals I-2 and II-2, locates the recombination event.

but would stop when the inconsistency is found between the grandmother I-2 and her daughter II-2. The resulting partial haplotyping is shown in the left-hand panel of figure 5. By combining the results of these two propagations, one immediately identifies the inconsistency as arising from the gametes of individual I-2. One infers that the single recombination event occurred between individuals I-2 and II-2, since this scenario entirely accounts for the data. (Otherwise, one would have to assume two identical recombination events between I-2 and her daughters II-3 and II-4, which is far less likely.) By inspecting the haplotypes of I-2, II-3, and II-4 and comparing them with those of II-2, we find that the recombination event is immediately localized to the region between the STR-45 and STR-49 markers.

Application to Family Data

Five families (40, 152, 154, 230, and 232) were analyzed using the described system. The number of genotyped individuals in each family was 9, 6, 6, 3, and 6, respectively. There were no recombination events in these families. In the initial signal processing of the DuPont sequencer data, some pedigrees had individual lane data with ambiguous reference marker peaks that were not detected by the computer. We added an interactive graphic software module for initial signal processing of lane data, which allowed a user to optionally interrupt the signal processing of lane data and to indicate the locations of (1) the left reference marker peak, (2) the right reference marker peak, and (3) any spurious data peaks. In 42 haplotypings, only one spurious data peak needed to be suppressed (an individual's lane data in family 232).

For some families, there was ambiguity in the quantitative sizing of alleles, because of local variations of 1 or 2 bp in gel migration that were not corrected for by linear interpolation between the two reference markers used. In the 168 observed alleles in the five families, 89% (149/168) showed no size variation, 9% (15/168) showed 1-bp variation, and 1% (3/168) showed 2-bp variation; there was one new mutation at a locus, as well. This ambiguity caused early termination of the program, during the graph propagation for setting phase. The size variation was handled by incorporating a tolerance (in base pairs) into the graph propagation, so that allele sizes within the specified tolerance were considered equivalent. Analysis of a family then proceeded by starting with a 0 tolerance and then incrementing by 1 until the program ran to completion. Family 40 ran at 0 tolerance, families 152 and 232 ran at a 1-bp tolerance, and families 154 and 230 required a tolerance of 2 bp.

The automated genotyping revealed two grandpaternal alleles in family 40, at STR-49 (allele 264, which is inherited by individuals II-2 and III-2, and allele 262, which is inherited by individuals II-3, II-4, and III-4). The inheritance pattern and single dinucleotide allele difference suggest mutation as the most likely explanation. Interestingly, this mutation was mistyped in (Schwartz et al. 1992); however, quantitative reexamination of the sequencer traces (fig. 1) confirms the computer interpretation. Since I-1's grandpaternal haplotype is neither used nor inferred by the program to set phase, no inconsistencies were detected.

After allele determination and setting of the size tolerance, all subsequent processing was fully automatic. The constraint propagation process used the allele information to unambiguously set phase; the phenotypes were determined; and the results were visually presented to the user, as a pedigree, on the computer display, annotated with carrier and affected status. Once the initial signal analysis was done, all this subsequent processing for a family was completed in several seconds on a Macintosh Quadra-class computer.

Discussion

Automation of genotype data acquisition, interpretation, and storage would benefit human molecular medicine at a number of levels. (1) The identification of disease genes through genetic linkage analysis would be dramatically accelerated: the currently labor-intensive and tedious process of linkage studies would be largely replaced by highly multiplexed, computerized allele identification and haplotype interpretation. Localization of disease genes would then be limited only by the availability of adequate disease families for analysis. (2) Automated data acquisition of dense genetic information should also open avenues for novel approaches to elucidation of disease molecular genetics, for both simple and complex genetic disorders, as well as in cancer research. (2a) For example, concordance mapping, where each meiotic recombination breakpoint is localized on each chromosome in each progeny, becomes possible. In this approach, analogous to physical mapping (Perlin and Chakravarti 1993), specific regions of the genome are identified that are concordant between all affected individuals but discordant with all unaffected individuals in a given pedigree. This approach could eliminate the need for much of the complex statistical analyses intrinsic to more traditional linkage studies. (2b) Complex genetic loci may be dissected by the identification of shared, localized haplotypes in unrelated affected patients (linkage disequilibrium mapping). (2c) Small regions demonstrating loss of heterozygosity could be mapped by comparing tumor DNA with peripheral blood DNA in cancer patients, thereby identifying genome regions important in cancer. (3) In the clinical setting, automated genotyping would increase the speed and accuracy of diagnostic studies. In addition, automation could dramatically decrease costs associated with molecular diagnostics. Molecular diagnostic tests are currently quite expensive, primarily because of the substantial personnel requirement for data generation and interpretation.

The goal of genotype automation is to generate large amounts of allele data in as few experimental analyses as possible and to use computers to acquire and interpret the data. In this report, we have successfully implemented many of the steps required for complete automation of genotyping. We used segregation of multiplexed CA repeats distributed throughout the 10-cM dystrophin gene, as a model system to develop software for computer acquisition and interpretation of genotype data. The multiplex PCR reaction contained four fluorescein-labeled CA-repeat loci and used internal molecular-weight markers, as described elsewhere (Schwartz et al. 1992). We used the DuPont Genesis 2000 automated sequencer system, which has automatic lane tracking after initial assignment of lanes and which, for each sequencer lane, provides raw digital data as PMT voltage as a function of time. We developed computer protocols that used the raw sequencer data files as input and that automatically interpreted the data and analyzed the genotype information.

Our deconvolution approach for genotyping STRs from complex band patterns is not limited to the X chromosome. Since the stutter pattern is associated with the STR locus and does not depend on a given family or allele, a database can be constructed of single-allele decay (e.g., derived from homozygotes or distinct heterozygotes) for each STR locus of interest. These patterns can then be used in our deconvolution algorithm to genotype STRs located anywhere in the genome, including on autosomes.

There are a number of aspects of our computer approach that must be modified before the system is robust enough for practical use. First, trial runs with some pedigrees encountered sufficient noise to lead to incorrect interpretations: the program needs to be made more tolerant of the variability of experimental data. We plan to introduce comparison algorithms that use pattern superposition of first-degree relatives, to help assign alleles in noisy data: an analogous process is used when data is interpreted by the human eye. Second, our program currently uses DuPont Genesis 2000 raw lane data as input. This automated sequencer system is extraordinarily sensitive and provides raw individual data as a single digital stream of voltage-versus-time lane data. Unfortunately, the Genesis 2000 sequencer is capable of detecting only a single color (fluorescein), as the proprietary dyes developed by DuPont are not available for custom incorporation into PCR primers. Furthermore, production of this automated sequencer has been halted, and it is no longer commercially available or serviceable. Finally, determination and propagation of phase is more complex for autosomal markers, where individuals with phase-known haplotypes (i.e., males in X-linked pedigrees) are not available. The high degree of informativeness should facilitate future autosomal studies with statistical (Ott 1991) or deductive (Wijsman 1987) analyses.

We are currently adapting our automated genotype-data-acquisition system to the more common ABI 373A automated sequencers. These sequencers have the advantage of simultaneous detection of multiple fluorophors. For example, we have recently described the analysis of 22 CA-repeat loci in two lanes of this sequencer (450 alleles/gel) (H. Kobayashi, personal communication). However, the ABI sequencer has the disadvantage of providing output as a "gel file": a two-dimensional, 20-megabyte array of data points that requires considerable processing before it can be used as input to our programs. The development of image-processing software as a front end to our programs is currently underway.

In addition to adaptation for the more common and flexible ABI automated sequencer, future goals are to automate risk assessment in DMD/BMD dystrophy families, via automated incorporation of serum creatine kinase data and Bayesian risk estimates into the computer haplotype analysis, and to detect nonpaternity. We are also extending our system to automated interpretation of genetic mapping data (Matise et al. 1994), including automated con-

cordance mapping of meiotic recombination data across the entire X chromosome. Eventually, we hope to incorporate LOD score calculation and automated experiment design, in order to optimize the statistical power of experiments using genetic map data.

Acknowledgments

Dr. Clark Tibbetts provided the formatting descriptions of the DuPont Genesis DNA sequencer system data files, required for fully mechanized data entry. This work was supported by from the NIH National Institute of Neurological Disorders and Stroke grant R01 NS32084 to M.W.P. and E.P.H.

Appendix

Key to our full automation is the ability to compute the allele sizes s and t at a locus, in the presence of PCR stutter peaks. As above, the data pattern $q(x)$ represents one or two alleles, whose separation may give rise to complex stutter peaks, and the data pattern $p(x)$ represents the simpler stutter peaks of just one allele. $p(x)$ and $q(x)$ measure the same locus for different individuals and are renormalized to reflect the dosage from one or two chromosomes. Our model is that $q(x)$ is constructed by the convolution (repeated shifting and adding) of $p(x)$: $q(x) = (x^s+x^t)p(x)$, where s and t denote the unknown true allele sizes.

We give here a detailed derivation of our deconvolution procedure for recovering the alleles s and t in the presence of PCR stutter peaks from $q(x)$, using $p(x)$. $p(x)$ is immediately known in X-chromosome family data and can be derived via similar deconvolution procedures for autosomal loci. We proceed in four steps.

Computing an Expression for the Allele Sum s + t

Taking the derivatives of both sides of $q(x) = p(x)(x^s+x^t)$, we obtain

$$\begin{aligned}\frac{d}{dx}[q(x)] &= \frac{d}{dx}[p(x)\times(x^s+x^t)] \\ &= \frac{d}{dx}[p(x)]\times(x^s+x^t) + p(x)\times\frac{d}{dx}(x^s+x^t) \\ &= p^{(1)}(x)\times(x^s+x^t) + p(x)\times(s\times x^{s-1} + t\times x^{t-1})\,.\end{aligned}$$

Evaluating at $x = 1$, we obtain

$$\begin{aligned}q^{(1)}(1) &= p^{(1)}(1)\times(1^s+1^t) + p(1)\times(s-\mathrm{ts}1^{s-1}+t\times 1^{t-1}) \\ &= p^{(1)}(1)\times(2) + p^{(0)}(1)\times(s+t)\,.\end{aligned}$$

The nth moment of a polynomial $u(x)$ is $u_n = u^{(n)}(1)$. This may be very efficiently computed, in linear time, as the sum of the coefficients of the polynomial's nth derivative. The moments are related to more intuitive function statistics, such as the mean and variance: $E(u) = u_1/u_0$ and $E(u^2) = u_2/u_0 + u_1/u_0 - (u_1/u_0)^2$, respectively. We can rewrite the above derivation as (easily computable) moment statistics: $q_1 = 2p_1 + (s+t)p_0$, or $q_1/p_0 = 2p_1/p_0 + s + t$, so

$$\begin{aligned}s + t &= q_1/p_0 - 2p_1/p_0 \\ &= (q_1-2p_1)/p_0\,.\end{aligned} \qquad \text{(A1)}$$

Thus, given the hemizygous (or homozygous) distribution $p(x)$ and the sequencer data $q(x)$, if either s or t is known, then so is the other. When the position t of the larger allele is determined by identifying the peak of the largest PCR product in the locus region, this algorithm will unambiguously determine the location s of the smaller allele.

Computing an Expression for the Allele Sum $s^2 + t^2$

To extract second moments, we compute the second derivative of the relation $q(x) = p(x)\times(x^s+x^t)$. After simplification, this produces

$$\begin{aligned}q^{(2)}(x) = p^{(2)}(x)&\times(x^s+x^t) \\ &+ 2[p^{(1)}(x)\times(sx^{s-1}+tx^{t-1})] \\ &+ p(x)[s(s-1)x^{s-2}+t(t-1)x^{t-2}]\,.\end{aligned}$$

Setting $x = 1$ to calculate moments and rearranging to group the constant, linear, and quadratic terms in s and t, we obtain the equality $0 = (2p_2-q_2) + (s+t)(2p_1-p_0) + (s^2+t^2)p_0$. Rearranging this equality gives the equivalence

$$s^2 + t^2 = [(q_2-2p_2) + (s+t)(p_0-2p_1)]/p_0. \qquad \text{(A2)}$$

Each right-hand-side term is directly or indirectly computable from moment properties of the data. For example, $s + t$ is known via equation (A1).

Computing an Expression for the Allele Difference s − t

From $s + t$, given in equation (A1), and $s^2 + t^2$, given in equation (A2), we can now obtain $s - t$, as follows:

$$\begin{aligned}(s-t)^2 &= s^2 - 2st + t^2 \\ &= s^2 + t^2 - 2st \\ &= 2s^2 + 2t^2 - (s^2+t^2+2st) \\ &= 2(s^2+t^2) - (s+t)^2\,.\end{aligned}$$

This provides a closed-form expression for $s - t$, as the square root of $2(s^2+t^2) - (s+t)^2$.

Computing Alleles s and t

Combining $s + t$ and $s - t$, we have $s = [(s+t)+(s-t)]/2$, and $t = [(s+t)-(s-t)]/2$. Thus, by taking the zero, first, and second moments of the multiallelic sequence data $q(x)$, together with the known haplotype $p(x)$, we can robustly and rapidly (in linear time) compute the absolute positions of nucleotide repeat alleles s and t.

References

Beggs A, Kunkel L (1990) A polymorphic CACA repeat in the 3′ untranslated region of dystrophin. Nucleic Acids Res 18:1931

Burmeister M, Monaco A, Gillard E, van Ommen G, Affara N, Ferguson-Smith M, Kunkel L, et al (1988) A 10-megabase physical map of human Xp21, including the Duchenne muscular dystrophy gene. Genomics 2:189–202

Clemens PR, Fenwick RG, Chamberlain JS, Gibbs RA, de Andrade M, Chakraborty R, Caskey CT (1991) Carrier detection and prenatal diagnosis in Duchenne and Becker muscular dystrophy families, using dinucleotide repeat polymorphisms. Am J Hum Genet 49:951–960

Feener CA, Boyce FM, Kunkel LM (1991) Rapid detection of CA polymorphisms in cloned DNA: application to the 5′ region of the dystrophin gene. Am J Hum Genet 48:621–627

Hoffman EP (1994) The evolving human genome project: current and future impact. Am J Hum Genet 54:129–136

Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM (1987) Complete cloning of the Duchenne muscular dystrophy cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 50: 509–517

Matise TC, Perlin MW, Chakravarti A (1994) Automated construction of genetic linkage maps using an expert system (MultiMap): application to 1268 human microsatellite markers. Nature Genet 6:384–390

Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, Kunkel LM (1986) Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 323: 646–650

Ott J (1991) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore

Oudet C, Heilig R, Mandel J (1990) An informative polymorphism detectable by polymerase chain reaction at the 3′ end of the dystrophin gene. Hum Genet 84:283–285

Perlin MW, Chakravarti A (1993) Efficient construction of high-resolution physical maps from yeast artificial chromosomes using radiation hybrids: inner product mapping. Genomics 18: 283–289

Schwartz LS, Tarleton J, Popovich B, Seltzer WK, Hoffman EP (1992) Fluorescent multiplex linkage analysis and carrier detection for Duchenne/Becker muscular dystrophy. Am J Hum Genet 51:721–729

Stephan DA, Buist NRM, Chittenden AB, Ricker K, Zhou J, Hoffman EP. A rippling muscle disease gene is localized to 1q41: evidence for multiple genes. Neurology (in press)

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al (1992) A second generation linkage map of the human genome. Nature 359:794–801

Wijsman EM (1987) A deductive method of haplotype analysis in pedigrees. Am J Hum Genet 41:356–373