IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861. Volume 13, Issue 1 Ver. III (Jan. 2014), PP 34-39
www.iosrjournals.org

Detection of Mutation-Hotspots and Exon-Deletions using Digital Signal Processing in Duchenne Muscular Dystrophy (DMD) Gene

J. K. Meher[1], A. K. Rath[2]*
[1]*Vikash College of Engineering for Women, Bargarh-768028, Odisha, India*
[2]*A.E.S. College, Tarbha-767016, Subarnapur, Odisha, India*

Abstract: *Duchenne muscular dystrophy (DMD) is a life threatening disease which mostly occurs due to deletions in the dystrophin gene. The deletions are reported to be clustered in two main regions in the gene involving nearly one fourth of exons. These regions also represent major meiotic recombination hotspots. The detection of deletions/mutation-hotspots in the coding sequences (exons) is not much possible till date due to the gene specific-potential variations or the complex organization of introns and exons in the gene. A Digital signal processing (DSP) method based on antinotch filter that exploits the period-3 property of a DNA sequence is applied for the identification of exons in DNA sequences by providing concerned peaks in magnitude-plot. A two digit numerical representation has been proposed for generating indicator sequence. In our approach to detect deletion in DMD gene, all-exon-sequence was taken and analyzed through digital filter to obtain normal peaks in the plot, which was than compared with the plots obtained after deletion of hotspot exons one by one. It was found that most of the hotspot exons share the peaks and deletions are marked by significant modifications in peaks. Thus, it was known that the peaks in filter-plot of DMD gene are associated with the location of the hotspots and their modifications/absence can be utilized to detect the deletions of exons.*
Keywords: *Exon, exon-deletion, hotspot, DMD, antinotch filter*

I. Introduction

Dystrophinopathies refers to a spectrum of muscle disease caused by the absence or malformation of a protein, dystrophin. The commonest type of muscular dystrophy is dystrophinopathy that includes the clinically more severe form of Duchenne muscular dystrophy (DMD) and the less severe form, Becker muscular dystrophy (BMD) [1]. Both the forms are characterized by progressive wasting and weakness affecting the voluntary muscles. Duchenne's disease (or pseudohypertrophic) is the most common type and affects boys between two and six. It begins in the large muscles of the lower trunk and upper legs. Progression is usually rapid; a wheelchair is usually needed by adolescence, and life expectancy is short [2].

Duchenne muscular dystrophy in Man is an X-linked recessive trait that affects 1 in 3600–6000 live male births [3-5] caused by mutations in the gene, DMD (chromosomal locus Xp21.3-p21.2), that encodes the protein, dystrophin. The dystrophin gene has been identified by positional cloning in 1986 on chromosome X [6, 7]. With 79 exons and 8 promoters and consisting of 2.4 million base pairs, the DMD gene is one of the largest genes known to date [8].

Approximately 60%-70% of mutations in individuals with DMD and BMD occur due to the deletions of one or more exons [9-12]. It was documented by Tan et al. in 2010 [13] that many of the large gene deletions within the dystrophin gene can be detected in specific "hotspot areas" of the gene. These "hotspots" are clustered in two main regions - at the 5' proximal portion of the gene (exons 1, 3, 4, 5, 8, 13, 19) and within the mid-distal region (exons 42 - 45, 47, 48, 50 - 53, 60). Further, the documentation of Sironi et al. in 2006 [14] and Walmsley et al. in 2010 [15] revealed the deletion hotspots domains are in the exons 40–55 and exons 45–53 respectively.

Identification of gene or the protein-coding regions in DNA sequences through computation means is of great importance nowadays. It is now established that a DNA sequence can be divided into gene and intergenic regions. Further, split gene concept revealed that a gene in eukaryote is divided into two sub-regions called coding regions (exons) and non-coding regions (introns). Some exons within the protein-coding regions of DNA sequences of eukaryotes tend to exhibit a period-three pattern [16-19]. The period-three pattern of the exons can be utilized either to predict gene locations or to predict specific exons within the genes of eukaryotic cells [16-19].

Common digital signal processing (DSP) methods for the identification of exons in DNA sequences include the Discrete Fourier Transform (DFT) on overlapping windows[17,18, 20], the anti-notch IIR filter and multi-stage bandpass filter centered at $2\pi/3$ [19].

Although computational-identification of the coding sequences (exons) is achieved in some cases, the detection of mutation-hotspots in the coding sequences (exons) is not much possible till date due to the gene specific-potential variations in these hotspot regions. Due to the complex organisation of the introns and exons

and considerable variations involving deletions/duplications/point mutations in the large DMD gene, the detection of hot spots and/or deletions in the hotspot arena is a challenging job for analysts. Application of computational methodologies in this field is also in infancy. In our approach we have successfully used the digital signal processing method such as digital antinotch filter to specify the hotspot clusters in the normal gene and detected deletion in a mutated DMD gene.

Section II describes the mode of dataset preparation, the proposed numerical representation of the DMD sequence and the digital signal processing method. Section III explains the simulation results and discussion of the method applied for the analysis of DMD dataset. In Section IV a conclusion is drawn from the analysis.

II. Materials and Methods

2.1 Dataset Preparation

The DMD cDNA Reference Sequence was retrieved from the Leiden Muscular Dystrophy pages[21], which has all the 79 exons with 11058 bases. The exons representing deletion hotspots in mutated DMD were found out from various references through literature search in Tan et al. 2010 [13]; Sironi et al. 2006 [14]; Walmsley et al. 2010 [15]. Of these 22 hotspot exons (1, 3-5, 8, 13, 19, 40-45, 47, 48, 50-55, 60) were studied using antinotch filter.

The exons were then lined up continuously without gaps to be used as the exon-data-set for filter analysis.

2.2 Proposed Numerical Representation

Since digital signal processing deals with analysis of numerical sequences, various approaches for numerical representation of genomic data and subsequent analysis have been made. As DNA sequence consists of four alphabets i.e. 'A', 'T', 'C' and 'G, it is much easier to represent it numerically by substituting the numerical values in a number of ways.' The sensitivity of the digital signal processing methods also depends on the numerical representations. There are some physico-chemical properties of the elements of DNA sequence which play major role in the analysis of the sequence. The existing numerical encoding method of Voss is the numerical representation [23] that maps the alphabets *A*, *G*, *C* and *T* into four-indicator sequences $x_A(n)$, $x_T(n)$, $x_C(n)$ and $x_G(n)$ substituting 1 for presence or 0 for absence of respective nucleotide. Another four-indicator sequence called relative frequency indicator sequence, based on various coding statistics like single-nucleotide, dinucleotide and trinucleotide biases was incorporated into the algorithm to improve the selectivity and sensitivity of filter methods [24,25]. The paired numeric [26] method deals with complementary property of nucleotides and assign +1 and -1 to show the presence of *A*-*T* and *C*-*G* nucleotide pairs respectively.

In this paper a two digit binary representation of the DNA bases is used. Since the alphabets are four such as A, T, C and G and binary numbers are two such as 0 and 1, hence four combinations are possible and suitable for substitution in the DNA sequence. The two digits 00, 01, 10 and 11 are assigned to A, T, C and G respectively. Like the real number representations, nucleotide bases A, T, C and G can be substituted with two-digit binary numbers as shown in Table 1. It follows the symmetric properties of nucleotides, i.e, A-G and C-T. The pairs are thus represented as 00-11, 01-10 respectively.

Table 1: Two Digit Representation

Nucleotide	Two Digit
A	00
T	01
C	10
G	11

2.3 Digital Filter Method

It has been established that DNA sequence exhibit period-3 property which determines the coding regions specifically the exons in the eukaryotic DNA sequence. Digital signal processing plays a major role in predicting these coding regions as DSP tools exhibit perod-3 property by suitable modifications. Thus these methods can be applied in predicting the coding regions of the DMD gene sequence. The digital filtering techniques such as the antinotch filter have been used to identify period-3 property in DNA sequences [18-19]. In digital filtering method for indicator sequence $x_B(n)$, corresponding filter output $Y_B(n)$ is computed where *B*=*A*, *T*, *C* and *G*. The sum of the square of filter outputs is expressed as :

$$Y(n) = \sum_{n=0}^{N-1} | Y_B(n) |^2 \quad (1)$$

An IIR anti-notch filter H(z) can be used for gene prediction [19]. The IIR filters can be obtained from a second order all pass filter with poles at $R\text{-}e^{\pm J\theta}$ as follows :

$$A(z) = \frac{R^2 - 2R\cos\theta z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \quad (2)$$

Where $R^2 < 1$, e.g. R =0.992, for stability. The IIR anti-notch filter is then can be calculated as follows :

$$H(z) = \frac{1 - A(z)}{2} \quad (3)$$

A plot of $Y(n)$ is used to extract the period-3 region of the DNA sequence effectively. The period-3 property of a DNA sequence implies that the peak at the gene locations is large. While this is generally true, the strength of the peak depends on the gene, and is thus sometimes is very pronounced and sometimes quite weak. The window length should be sufficiently large, e.g. 351, so that the periodicity effect dominates that background spectrum which makes its strong presence in DNA sequences.

The process was initiated with the retrieval of complete exon dataset of the DMD sequence from the public domain. It was arranged in a suitable form to be analysed by the DSP tool. The complete exon-data-set of DMD sequence was subjected to the digital filter to exploit its period three property and the Relative base location vs. Magnitude plot for the complete exon-data-set was obtained. Mutated gene sequences were generated by deleting the exons one by one or in association from the exon-data-set. The mutated exon-data-sets were subjected to antinotch filter to generate transformed domain signals and then plotted (Relative Base Locations vs. Magnitude) to get the concerned magnitude-peaks. Plots of mutated sequences were compared with the plot of normal exon-data-set to find the absence or modification of peaks.

The overall process has been described in a block diagram as shown in Fig.1.

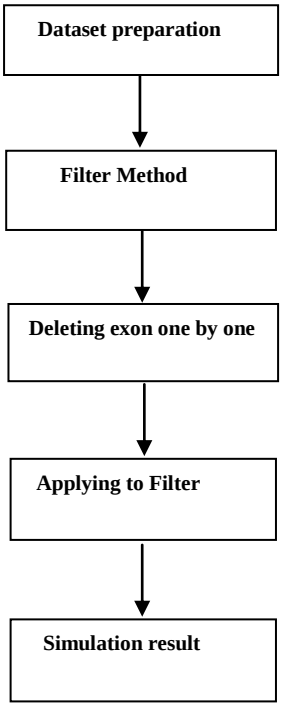

Fig 1. Block diagram to show the overall process of DMD gene detection.

III. Simulation Results and Discussion

The DMD exon sequence was subjected to the filter for the detection of the coding regions, especially, the hotspots. The result was obtained as magnitude versus the relative nucleotide base locations plots. The filter-response-plots of normal exon-data-set (comprising of all the 79 exons) showed more than 25 magnitude-peaks. About 17 (3-5, 8, 13, 19, 40-43, 47, 48, 51-53, 55, 60) of the 22 designated hotspot exons were located in

the peak regions, either at the apex or in ascending-descending arms (Fig. 2a). Again, filter-analyses of exon-data-sets after individual-deletion of hotspot exons revealed that modifications of the peaks in the plot though occurred for most of the exons, deletion of at least 14 hotspot exons could induce major changes in the peaks of filter-plots, which was marked by either absence or significant modifications of the corresponding peaks (Fig. 2b-f).

It is documented that large intragenic deletions and duplications together account for more than two-thirds of the mutations leading to DMD and BMD and, despite heterogeneity in deletion size and location, two hot spots have been identified. Of which, the major one involves exons 40-55 [27, 28]. These gene regions also represent major meiotic recombination hotspots [29]. Sequencing of deletion junctions involving both hotspots revealed no clear clustering of chromosome breaks and failed to identify any element or feature that could account for deletion formation [30, 31, 32, 33]. Though DNA double-strand breaks (DSBs) formation believed to be the primary cause [14], the process of exon deletions within the DMD locus is extremely unpredictable.

This period-three behaviour has been observed in many genes and is useful for locating the coding regions, which is exploited by filter method to obtain peaks in Relative Base Locations vs. Magnitude plot, where, the strength of the peak depends on the gene. The DMD exon-data-set, which, behaved as a prokaryotic (non-intronic) data set produced peaks in succession in the filter-plot. Systematic analysis of the plot revealed that most of the hotspots in DMD gene share the peak regions. Thus, revealing the correlation between the period-three pattern of the exons and the location of the hotspots (regions in DNA vulnerable to DSBs) in this gene. The fact was further supported by the absence or considerable modifications of the peaks in the filter-plots of mutated exon-data-set with deletions relating to at least 14 out of 22 hotspot exons. Though appearance of some peaks in non-hotspot area and location of a few hotspots in non-peak zones make the process more complicated in DMD gene, still, antinotch filter can be utilized both for the detection of hotspots and exon-deletions in this gene.

The performance analysis of various methods can be made by prediction measures such as sensitivity (S_N), specificity (S_P)

$$S_N = \frac{T_P}{T_P + F_N} \tag{4}$$

$$S_P = \frac{T_P}{T_P + F_P} \tag{5}$$

Whereas T_P=true positive, F_P=false positive and F_N=false negative [26]. T_P corresponds to those genes that are correctly predicted by the algorithm. F_P corresponds to the coding regions identified by a given algorithm which are not present in the standard annotation. F_N is coding region that is present but not predicted to be coding by the algorithm being used. True negatives are those regions which are not present in the dataset and also predicted as non-coding regions by not showing any peak at the respective locations.

The sensitivity is the ratio of true positive to the sum of true positive and false negative. The specificity is the ratio of true positive to sum of true positive and false positive. i.e. sensitivity depends on number of false negative and specificity depends on number of false positive. The sensitivity and specificity have been found out as 91% and 90.4% respectively.

IV. Conclusion

In this paper the analysis of DMD gene was made using digital signal processing i.e. antinotch filter having two digit numerical representations. The preliminary experiment provided encouraging result, as the detection of deletions/mutation-hotspots in the coding sequences was successfully done with a prediction accuracy of 91%. It happens to be a pioneer work using signal processing method.

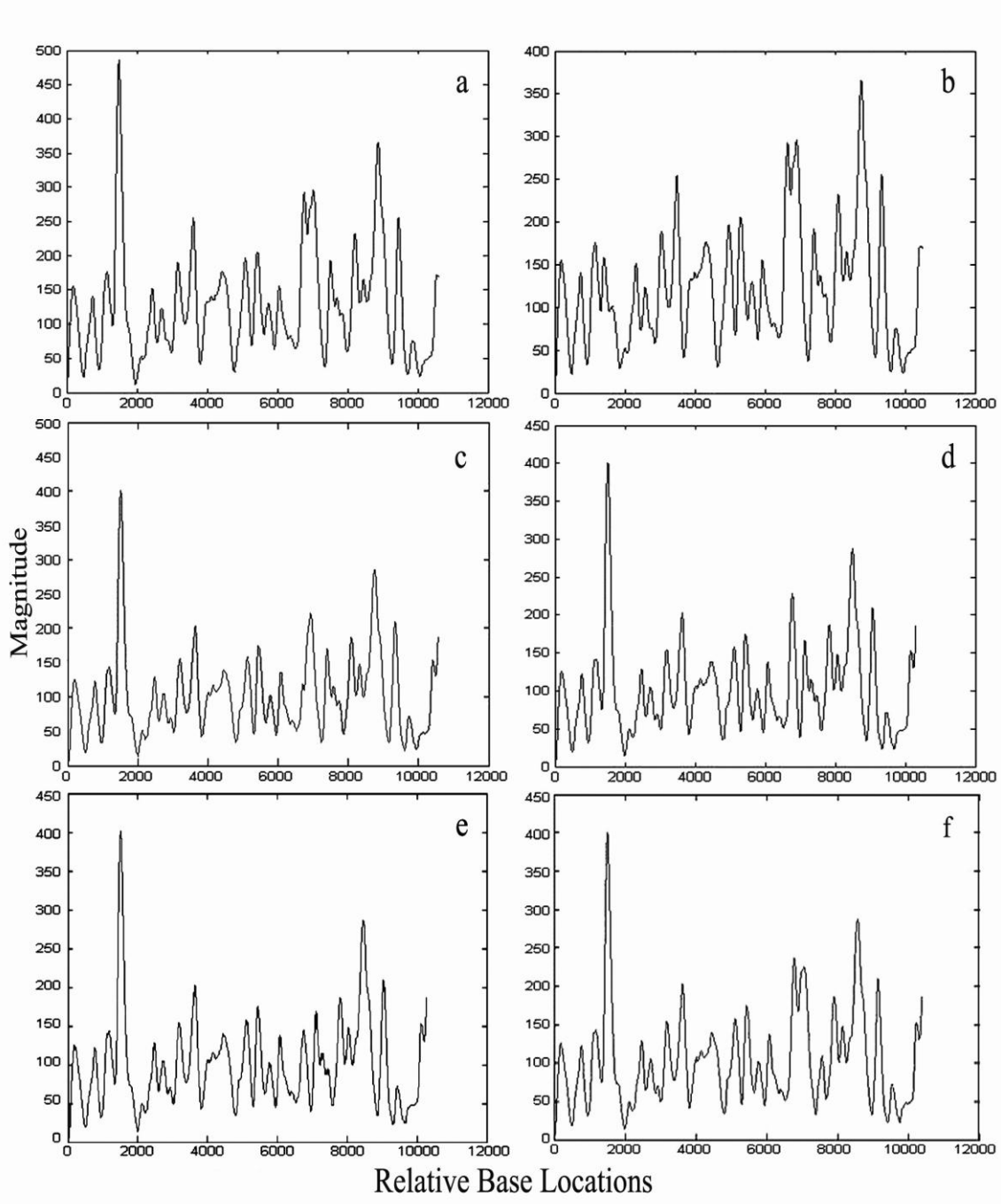

Figure 2. Plot of IIR antinotch filter technique, a) all exon-data-set, b) exon-data-set with deletion of exon 13 (bp 1483-1602), c) exon-data-set with deletion of exon 47 (bp (6763-6912), d) exon-data-set with deletion of exon 48 (bp (6913-7200), e) exon-data-set with deletion of exon 47-48 (bp (6763-7200), f) exon-data-set with deletion of exon 52-53 (bp (7543-7872).

Note : Authors have equal contributions; *Corresponding author.

References

[1] A.G. Engel and E. Ozawa, Dystrophinopathies, in A.G. Engel and L. C. Franzini-Armstrong, (Ed.) Myology. Basic and clinical, 3rd ed. Vol 2. (New York: McGraw Hill, 2004) 961-1025.

[2] R. Howard and M.E. Lewis, The Home Medical Manual (G-O) Time, Vol. III (New York: Doubleday & company, Inc., 1986). PP. 180.

[3] A. Drousiotou, P. Ioannou, T. Georgiou, et al., Neonatal screening for Duchenne muscular dystrophy: a novel semiquantitative application of the bioluminescence test for creatine kinase in a pilot national program in Cyprus, Genet Test., 2, 1998, 55–60.

[4] D. Bradley and E. Parsons, Newborn screening for Duchenne muscular dystrophy. Semin Neonatol., 3, 1998, 27–34.

[5] A.E. Emery, Population frequencies of inherited neuromuscular diseases—a world survey, Neuromuscul Disord, 1991, 1: 19–29.

[6] A.P. Monaco, R. L. Neve, C. Colletti-Feener, et al., Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene, Nature, Oct 16-22. 323(6089), 1986, 646-650.
[7] M. Koenig, E. P. Hoffman, C. J. Bertelson, et al., Complete cloning of the duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals, Cell, 50, Issue 3, 1987, 509-517.
[8] A. H. Ahn, L. M. Kunkel, The structural functional diversity of dystrophin, Nature Genetics, 3(4), 1993, 283-91.
[9] J. Yan, J. Feng, C. H., Buzin, et al., Three-tiered noninvasive diagnosis in 96% of patients with Duchenne muscular dystrophy (DMD), Hum Mutat., 23, 2004, 203–204.
[10] K. M. Dent, D.M., Dunn, A. C. von Niederhausern, et al., Improved molecular diagnosis of dystrophinopathies in an unselected clinical cohort, Am J Med Genet A, 134, 2005, 295–298.
[11] T. W. Prior, S. J. Bridgeman, Experience and strategy for the molecular testing of Duchenne muscular dystrophy, J. Mol Diagn, 7, 2005, 317–326.
[12] Y. Takeshima, M. Yagi, Y. Okizuka, et al., Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center, J Hum Genet., 55, 2010, 379–388.
[13] J.M.A Tan, J. H. Chan, K. Tan, et al., Dystrophin gene analysis in Duchenne/Becker dystrophy in a Malaysian population using multiplex polymerase chain reaction, Neurology Asia, 15(1), 2010, 19 – 25.
[14] M. Sironi, U. Pozzoli, G. Comi, et al., A region in the dystrophin gene major hot spot harbors a cluster of deletion breakpoints and generates double-strand breaks in yeast, doi: 10.1096/fj.05-5635fje September 2006. The FASEB Journal. 20, no. 11, 2006, 1910-1912.
[15] G. L. Walmsley, V. Arechavala-Gomeza, M. Fernandez-Fuente, et al. A Duchenne Muscular Dystrophy Gene Hot Spot Mutation in Dystrophin-Deficient Cavalier King Charles Spaniels Is Amenable to Exon 51 Skipping, PLoS ONE 5(1): e8647. doi:10.1371/journal.pone.0008647, 2010..
[16] J.W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Res., 10(17), 1982, 5303–5318.
[17] S. Tiwari, S. Ramachandran, A. Bhattacharya, et al., Prediction of probable genes by Fourier analysis of genomic sequences, Comput. Appl. Biosci., 13(3), 1997, 263–270.
[18] D. Anastassiou, DSP in genomics, Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Salt Lake City, Utah, USA, May 2001, 1053–1056.
[19] P. P. Vaidyanathan and B. J. Yoon, Digital filters for gene prediction applications, Proc. Asilomar Conference on Signals, Systems, and Computers. Pacific Grove, Calif, USA, November 2002, 306-310.
[20] D. Anastassiou, Genomic Signal Processing, IEEE Signal Processing Magazine, 18, 2001b, 8-20.
[21] http://www.dmd.nl/seqs/murefDMD.html#51
[22] S. Datta, A. Asif, H. Wang, Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics, IEEE 6th International Symposium on Multimedia Software Engineering, Miami, Florida, USA, 2004, 160-163.
[23] R. F. Voss, Evolution of long-range fractal correlations and 1/f noise inDNA base sequences, Physical Review Letters, 68, 1992, 3805.
[24] A. S. Nair, and S. P. Sreenathan, An improved digital filtering technique using frequency indicators for locating exons. Journal of CSI, vol.36, no.1, 2006.
[25] G. Aggarwal and R. Ramaswamy, Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER, J. Biosci., 27(Suppl. 1), 2002, 7-14.
[26] M. Akhtar, J. Epps, and E. Ambikairajah, Proc. IEEE GENSIPS on DNA numerical representations for period-3 based exon prediction, Tuusula, Finland, 2007, 1-4.
[27] J. T. Den Dunnen, P. M. Grootscholten, E. Bakker, et al., Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications, Am. J. Hum. Genet., 45, 1989, 835-847.
[28] M. Koenig, A. H. Beggs, M. Moyer, et al., The molecular basis for Duchenne versus Backer muscular dystrophy: correlation of severity with type of deletion. Am. J. Hum. Genet., 45, 1989, 498-506.
[29] C. Oudet, A. Hanauer, P. Clemens, et al., Two hot spots of recombination in the DMD gene correlate with the deletion prone regions, Hum. Mol. Genet., 1, 1992, 599-603.
[30] J. C. McNaughton, D. J. Cockburn, G. Hughes, et al., Is gene deletion in eukaryotes sequence-dependent? A study of nine deletion junctions and nineteen other deletion breakpoints in intron 7 of the human dystrophin gene, Gene, 222, 1998, 41-51.
[31] L. Toffolatti, B. Cardazzo, C. Nobile, et al., Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene, Genomics, 80, 2002. 523-530.
[32] C. Nobile, L. Toffolatti, F. Rizzi, et al., Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum. Genet., 110, 2002, 418-421.
[33] M. Sironi, U. Pozzoli, R. Cagliani, et al., Relevance of sequence and structure elements for deletion events in the dystrophin gene major hot-spot. Hum. Genet., 112, 2003, 272-288.