INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

COMPARATIVE SEQUENCING OF HUMAN AND MOUSE SYNTENIC REGIONS FROM DGCR AND MND2

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

Axin Hua

Norman, Oklahoma

1999

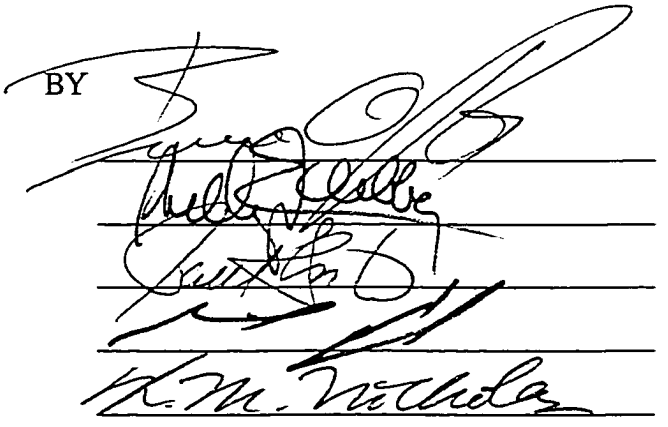
COMPARATIVE SEQUENCING OF HUMAN AND MOUSE

SYNTENIC REGIONS FROM DGCR AND MND2

A Dissertation APPROVED FOR THE

DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY



Acknowledgments

I wish to thank everyone in Dr. Roe's lab for their kind support and helpful discussion. Thanks especially to Dennis Burian, Linda Cantu, Lingzhi Chu, Sandy Clifton, Stephen Confer, Judy Crabtree, Stephen Deschamps, Trang Do, Angela Dorman, Jennifer Gray, Jennifer Hausner, Akbar Khan, Doris Kupfer, Lisa Lane, Victoria Lao, Michele Lasley, Sharon Lewis, Shaoping Lin, Eda Malaj, George Martin, Wendy Martin, Fares Najar, Huaqin Pan, Qun Ren, Lin Song, Yonathan Tilahun, YingPing Wang, Zhili Wang, Heather Wright, Hui Wu, Ziyun Yao, Xiling Yuan, Min Zhan, Guozhong Zhang, and Hua Zhu for their advice and help, and Steve Kenton, Hongshing Lai, Jing Tian, James White for expanding my computer knowledge.

I want to acknowledge my committee members past and present, Paul Cook, Michael Engel, Phil Klebba, Kenneth Nicholas and Greg Reinhardt for their time, support and advice. A special thanks goes to Michael Engel for his advice and continuous encouragement even after I left his group to pursue a career in chemistry and biochemistry. Support from the Department of Chemistry and Biochemistry, the Department of Geology and Geophysics, the Graduate college, the International Office also is appreciated.

A very special thanks goes to my host family, Mary Jonette McKee and Michael McKee, for their continuous help and encouragement during my more than five years stay in Norman.

Thanks to my major professor, Bruce Roe, for his guidance, support, patience, challenge and encouragement. I appreciate all he has done for me and I will do my best in the future to make him proud.

Finally, my deepest gratitude goes to my wife, Jing Tian, and daughters, Qianwen Hua and June Hua, for their love, patience and encouragement . Without their support, the dissertation would never have been written.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2xTY - medium containing tryptone, yeast extract and salt

ABI - Applied Biosystems Incorporated

AGCT - the Advanced Center for Genome Technology, Department of Chemistry and Biochemistry, the University of Oklahoma

ALS - amyotrophic lateral sclerosis

ALU - a human DNA repeat element with homology to 7SL RNA

ATF - activating transcription factor

B1 - a family of SINEs found in rodents

B2 - a family of SINEs found in rodents

BAC - bacterial artificial chromosome

BCR - breakpoint cluster region

BLAST - basic local alignment search tool

CAF - chromatin assembly factor

CATCH22 - Cardiac defect, Abnormal faces, Thymic hypoplasia, Cleft palate, Hypocalcaemia and 22q11 deletions

CCD - charge coupled device

cDNA - complementary deoxyribonucleic acid

CDP - CCAAT displacement protein

c/EBP - CCAAT/enhancer binding protein

CLTCL - clathrin heavy chain-like gene, also called CLTD or CLH22

CML - chronic myelogenous leukemia

CPE - Clostridium perfringens enterotoxin

CRE - cAMP-responsive elements

CREB - cyclic AMP regulatory element binding protein

CTF - CCAAT transcription factor

dbEST - a Genbank database of ESTs

ddNTP - dideoxynucleotide

DGCR - DiGeorge syndrome critical region

DGS - DiGeorge syndrome

dITP - 2'-deoxy-inosine triphosphate

DKH - Dok homology

DMSO - dimethyl sulfoxide

DNA - deoxyribonucleic acid

DNase - deoxyribonuclease

dNTP - deoxynucleotide

Dok - downstream of tyrosine kinase, genes coding for a p60 protein.

dsRBD - double stranded RNA-binding domains

E2F - a family of transcription factors.

EDTA - disodium ethylenediamine tetraacetate

EST - expressed sequence tag

ETS - a family of transcription factors

FGFR - fibroblast growth factor receptor

FISH - fluorescence *in situ* hibridization

GABP - GA-binding proteins

GAP - GTPase-activating protein

GCG - a software package from the Genetic Computer Group in Madison, WI.

Gene7 - putative genes found in human chromosome 2 and mouse chromosome 6

GET - buffer containing glucose, Tris-HCl and EDTA

GRE - glucocorticoid responsive elements

GSCL - a goosecoid-like homeobox gene

HARA - homolog of yeast repressors 1 and 2 of core histone gene transcription

HIV - human immunodeficiency virus

HTF - HpaII tiny fragment

Htf9c - one of two genes found in association with a HTF fragment in chromosome 22

IAP - intracisternal A particle

ID - rodent identifier repeat

IPTG - isopropyl-b-D-thiogalactopyranoside

LB - medium containing tryptone, yeast extract and salt

LINE - long interspersed repeat element, such as L1

Lor2 - lysyl-oxidase related gene 2

LTR - long terminal repeat

MER - Medium Reiteration frequency Repeat

MIR - mammalian-wide interspersed repeats

MND - motor neuron disease

mRNA - messenger ribonucleic acid

NF - nuclear factor

NNPP - promoter prediction by neural network

ORF - open reading frame

PAC - P1-derived artificial chromosome

PCR - polymerase chain reaction

PDGF - platelet-derived growth factor

PH - pleckstrin homology

Phred/Phrap - Phil's read editor/Phil's read assembly program

Porc-P1 - a human gene in human chromosome 22 which shows significant homology to the yeast protein CDC45

Powblast - a computer program to perform database homology search

PutA - putative gene A

PutB - putative gene B

RACE - rapid amplification of cDNA ends

RAN - a protein homologous to Ras family members

RanBP1 - RAN binding protein 1

RanGAP- a Ran GTPase-activating protein

RCC1 - the regulator of chromosome condensation

RFLP - restriction fragment length polymorphism

RNA - ribonucleic acid

RNase - ribonuclease

rRNA - ribosomal ribonucleic acid

RT-PCR - reverse transcriptase PCR

RUMT - tRNA (uracil-5-)methyltransferase

SDS - sodium dodecyl sulfate

SH - src homology

SINE - Short INterspersed Element

snRNA - small nuclear RNA

SP1 - a transcription factor

SRCR - scavenger receptor cysteine-rich

SRF - serum response factor

STS - sequence tagged site

T10 - a gene coding a serine/threonine rich protein

TB - terrific broth

TBE - buffer containing Tris-HCL, boric acid and EDTA

TBP - TATA binding protein

TE - buffer containing Tris-HCl and EDTA

TM - melting temperature or buffer containing Tris-HCl and magnesium

TMpred - a program for predicting transmembrane domain of protein

TMVCF - transmembrane protein deleted in Velo-Cardio-Facial syndrome

TOH - a male patient with DGS

tRNA - transfer ribonucleic acid

TSSW - a promoter prediction program

UTR - untranslated region

UV - ultraviolet

VCFS - velo-cardio-facial syndrome

VDU - a female patient with DGS, mother of ADU

WD - a number of highly conserved repeating units ending with trp-asp (W-D)

WMM - weight matrix model

WW - a protein domain probably involved in protein-protein interaction

X-Gal - 5-bromo-4-chloro-3-indolyl-b-D-galactoside

XGAP - X windows based genome assembly program

Xgrail - X windows based gene recognition and analysis internet link. Used for sequence feature prediction.

YAC - yeast artificial chromosome

YENB - medium containing yeast extract and nutrient broth

Abstract

With successful implementation and improvement of the shotgun sequencing strategy, two syntenic regions between human and mouse, totaling about 0.8 Mb, have been sequenced. One is from the DiGeorge syndrome critical region (DGCR) from human chromosome 22q11 and its syntenic region from mouse chromosome 16. The other is from the mouse motor neuron disease critical region (mnd2) from mouse chromosome 6 and its syntenic region in human chromosome 2. Based on results from database similarity search and computational gene prediction, a total of 13 genes have been predicted in the sequenced regions. Specifically, four known genes, i.e. CLTCL, HIRA, Porc-P1, and TMVCF and one putative gene, PutA, have been found in the human DGCR region. Another seven genes (four known: T10, Htf9c, RanBP1 and Dok; three putative: PutA, Gene7 and Lor2) have been found in both human and mouse. An additional gene, semaphorin M, was predicted and confirmed in the mouse mnd2 region. Genomic structure, putative regulatory elements and possible functional implications of all the genes have been examined in detail.

Sequence comparison between these human and mouse syntenic regions revealed a high conservation in gene order, gene structure as well as coding sequences. Regulatory elements including promoters and polyadenylation signals had a moderate to high degree of sequence conservation. A low level and occasionally an absence of sequence conservation was observed in the syntenic intronic and intergenic regions. No significant difference in the evolution rate of coding sequence was found between the two syntenic regions, which are located in different chromosomes in both species. Varying amounts of repeat sequences, ranging from 22% - 60% for human and 26-47% for mouse occurred in the sequenced regions, with similar amounts found in the human and mouse syntenic regions.

A new gene prediction method based on human and mouse coding sequence conservation was designed and implemented. The parameters evaluated in the new method include nucleic acid and amino acid sequence level conservation, exon frame and length

conservation, splicing site quality, presence of conserved CpG islands and polyadenylation signals. A program package, named Geneligner, was developed to facilitate annotation of genomic sequences. In addition to the compartative gene prediction, major features of the program package also include graphic alignment of the results of database similarity searches and computational gene predictions, and gene model building from those results. Other features of the program package include examination of GC content, CpG ilands, open reading frames and repeat statistics of genomic sequences.

Chapter I
Introduction

1.1 Human and mouse genome

Genetic material: Since the discovery of the double-helical structure of deoxyribonucleic acid (DNA) by Watson and Crick (1953), it has been established that DNA is the major carrier of genetic information in most life forms. According to Watson and Crick, DNA is composed of two associated polynucleotide strands that wind together in an antiparallel, helical fashion (Figure 1.1). Four types of nucleotides form the polynucleotide chains in the double helix. Each nucleotide contains a distinctive heterocyclic nitrogenous base, either adenine (A), guanine (G), cytosine (C) or thymine (T), joined to a deoxyribose phosphate. The 5' end of a polynucleotide chain either has a free hydroxyl or phosphate group at the 5' carbon of a sugar while the 3' end typically has a free hydroxyl group at the 3' carbon of a sugar. The two antiparallel strands of a double helix are held together by hydrogen bonds formed by specific base pairings, A with T and G with C, between the two strands, and favorable stacking of the bases. Normally, the two polynucleotide strands form a right-handed helix with about 10 base pairs per turn, which is referred to as the B form of DNA. The total genetic material, i.e. DNA for most life forms, is referred to as the genome of an organism.

At first glance, it is remarkable that only four types of bases code for almost all the information needed for complex life. However, considering the fact that a computer can store vast amounts of information using only two letters (0 and 1), it is obvious that if the string of letters is long enough, even greater amounts of information can be stored. Indeed, the amount of DNA (haploid genome size or C value) contained within organisms generally is very large, ranging from 35 Mb for *Navicola pelliculosa* (diatom) to 670,000 Mb for *Amoeba dubia* (amoeba) among eukaryotes (Li, 1997). The sizes of both the human and mouse genomes were estimated to be 3.4 billion bp (John and Miklos, 1988). The total 3.4

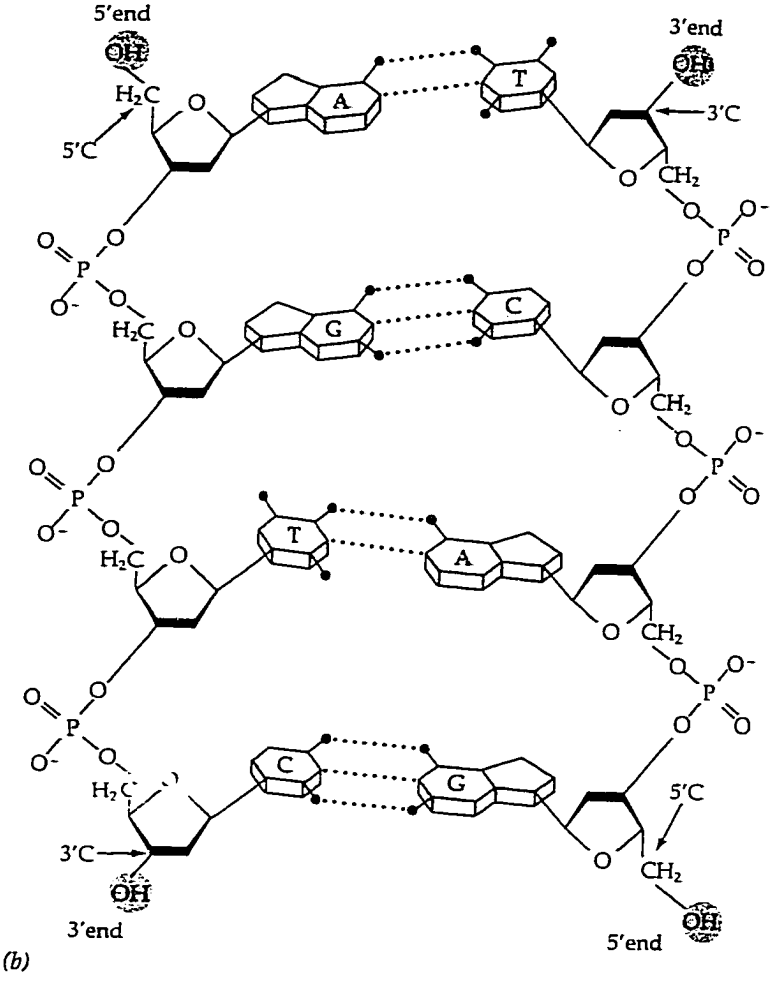


Figure 1.1 Diagram showing base pairing and antiparallel nature of DNA double helix

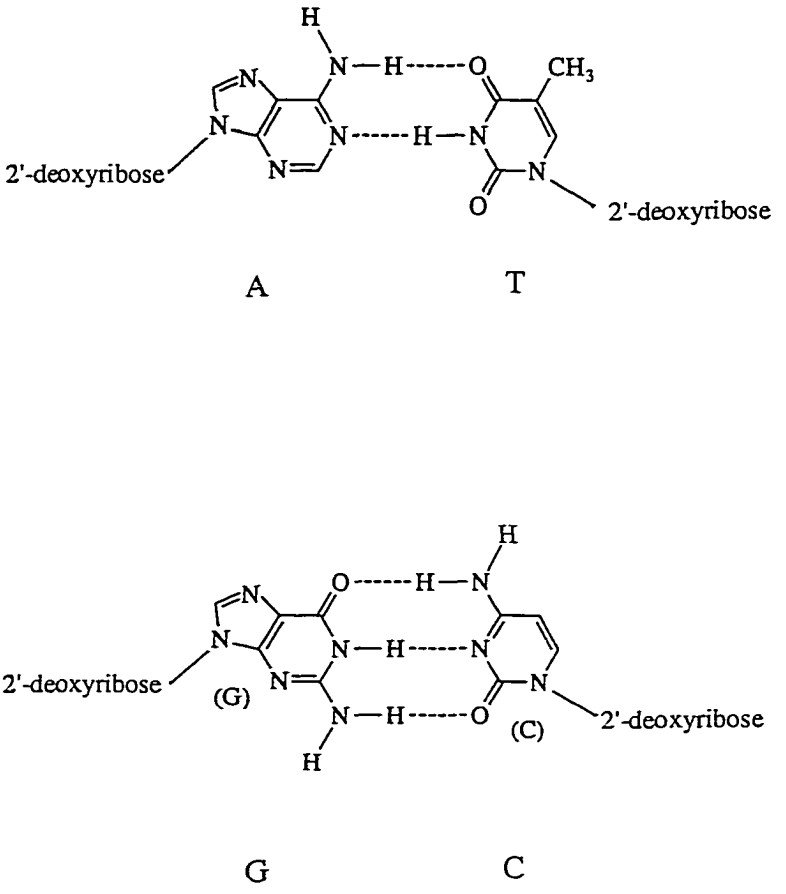


Figure 1.1b Base pairing between A and T and between G and C
Dash lines represent hydrogen bonds between the base pairs

billion bp of DNA in the human genome is separated into 23 chromosomes and 20 chromosomes in mouse.

Genes as the primary storage of genetic information: More than forty years of research has established the basic route of information flow in most organisms. Specifically, DNA serves as the template for its own replication and for transcription into ribonucleic acids (RNAs). Then, messenger RNA (mRNA, protein-coding RNA) serves as the template for translation into a protein. The relationship between DNA, RNA and protein often is termed the central dogma of molecular biology (Crick, 1958). Those DNA sequences which are necessary for synthesis of a functional peptide or RNA molecule are called genes. Eukaryotic genes have been grouped into three classes based on the use of three discrete transcription systems (Singer and Berg, 1991). Class I genes encode most ribosomal RNAs (rRNAs) and are transcribed by RNA polymerase I. Class II genes encode all messenger RNAs (mRNAs, those with protein as the final product) as well as small nuclear RNAs (snRNAs) and are transcribed by RNA polymerase II. Class III genes encode transfer RNAs (tRNAs), 5S rRNAs and other small RNAs and are transcribed by RNA polymerase III. Since class II genes encode all the proteins necessary for proper function of the cell and are the most well characterized gene class, genes in the remaining part of this dissertation refer to only class II genes unless otherwise specified.

The structure of a typical eukaryotic gene is shown in Figure 1.2. In eukaryotic organisms, including human and mouse, genes generally are composed of exons (coding for mature mRNA) that are separated by introns (non-coding regions transcribed into heterogeneous nuclear RNA or hnRNA but not present in mature mRNA). The length of exons varies between several to thousands of bp, averaging about 130 bp based upon examination of 4731 human exons (Zhang, 1998). The number of exons per gene also varies, from single exon genes to genes containing more than 70 exons, such as the human dystrophin gene which contains 79 exons spanning more than 2.5 Mb of the genome (Matsuo, 1995). While a typical exon averages 130 bp in length, introns generally are

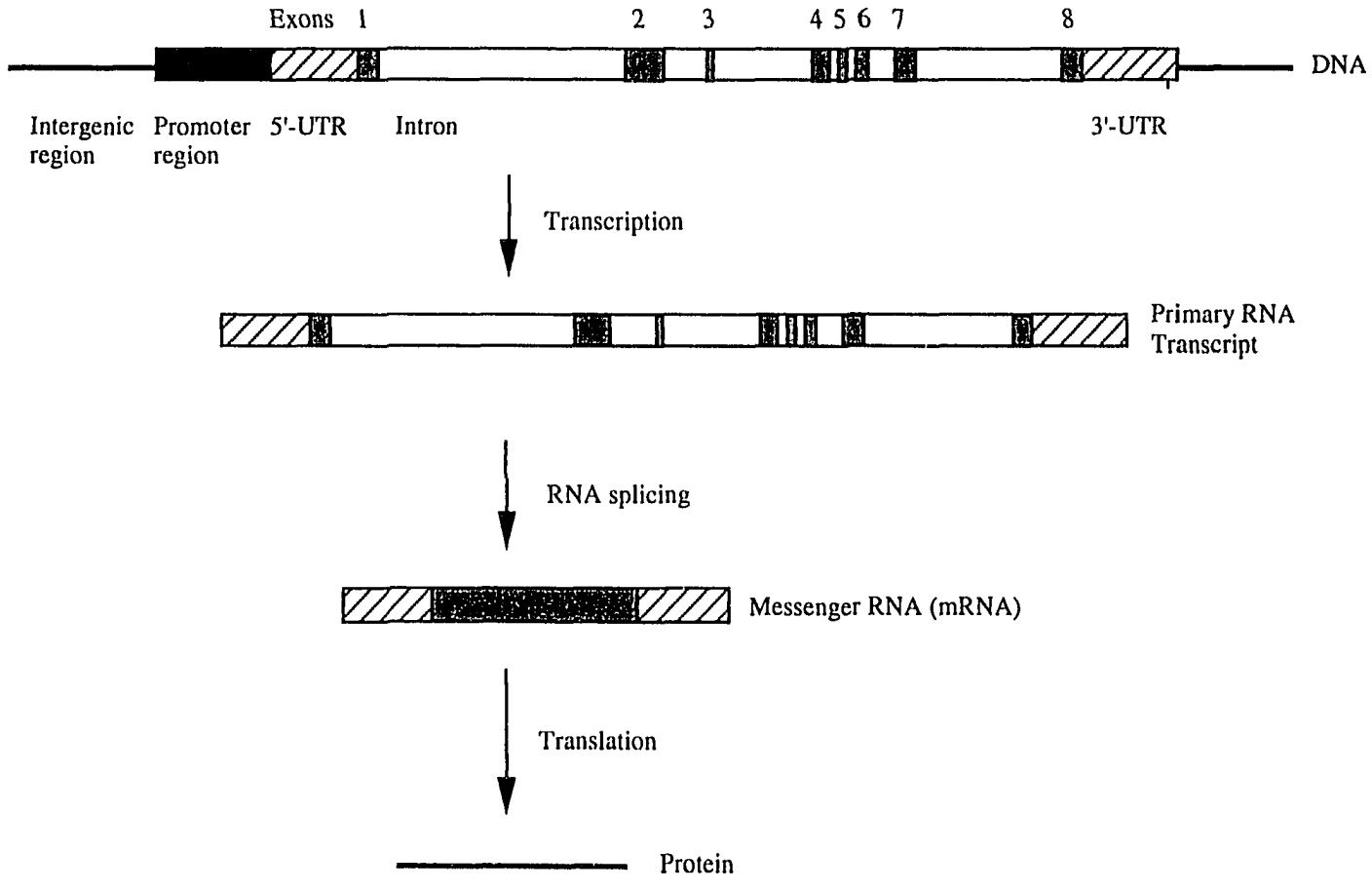


Figure 1.2 Diagram showing typical structure of an eukaryotic protein coding gene and its expression

much larger in length, frequently several thousand base pairs or even larger. Most introns are bounded by the consensus splice site sequence AG/GT at their 3' and 5' ends respectively. The expression of a gene often is regulated at both the transcription and translation levels. The major functional elements in a DNA sequence which control the transcription process are the promoters which regulate the initiation and the polyadenylation signals which likely regulate transcription termination. A promoter generally spans a region of more than 200 bp (Lewin, 1997) upstream from the transcription start site. Within a promoter, varying numbers of short sequence elements form transcription factor binding sites. One of the well studied transcription factor binding sites found in many genes is the TATA box which is located at about 30 bp upstream from the transcription start site and has a consensus sequence of TATAAAA (Grosschedl and Birnstiel, 1980; Lewin, 1997). The TATA box, which is the binding site for TATA-binding protein (TBP), is important in determining the start point of transcription. Other well studied sequence elements include the CCAAT box (consensus GGCCAATCT) and the GC box (GGGCGG). The CCAAT box, often located at about 60 to 100 bp upstream of the transcription start site, is the binding site of a number of transcription factors, including c/EBP (CCAAT/enhancer binding protein) (Osada et al., 1996b), CTF/NF-I (CCAAT transcription factor) (Osada et al., 1996a), CDP (CCAAT displacement protein) (Barberis et al., 1987) and NF-Y (nuclear factor Y) (Dorn et al., 1987). The GC box is the binding site for transcription factor SP1 and multiple GC boxes may present in a promoter (Toyoda et al., 1995). Both CCAAT box and GC box are functional in either orientation within a promoter and transcription activation is achieved through the binding of transcription factors. Although early studies show that most promoters contain a functional TATA box, it has been found in recent years that many genes lack a TATA box. Instead, sequence elements termed initiators (Inr) may be important in those TATA-less promoters and it has been suggested that the two types of promoters (TATA-containing and TATA-less) may be functionally and mechanistically different (Smale, 1997).

Different methods have been used to predict the total number of genes in the human genome and the estimated numbers range from 20,000 to 300,000 (Fields et al., 1994). The lower end of 20,000 genes was based on the assumption that about 12% of the human genome may be transcribed and genes may have an average size of 18 kb (Wagner et al., 1993). Since more than 50% of the genome sequence may be transcribed at least for the regions sequenced in this dissertation research (Table 6.1), this method seems to have underestimated the total number of genes. The higher end of 300,000 genes was obtained by assuming an average genomic size of 10 kb per gene and this number likely is an overestimate since at least in the DiGeorge syndrome critical region (about 1.25 Mb), only about 30 genes are found, which corresponds to about 42 kb per gene. Recently, an estimation based on EST data predicted that the human genome may contain 60,000-70,000 genes, which is similar to those based on the estimated numbers of CpG islands (67,000 genes) and calculated from the average gene density in the genomic regions already sequenced (71,000 genes) (Fields, 1994).

Human and mouse genomes rich in repetitive elements. In addition to genes, it is known that the human and mouse genomes contain large amounts of repetitive sequences. In humans, it is estimated that about 35% of the genome consists of interspersed repetitive sequences which primarily are degenerate copies of transposable elements (Smit, 1996). Two major classes of interspersed repeats in mammalian genomes are SINEs (Short INterspersed Element) and LINEs (Long INterspersed Element). The human genome contains two predominant families of interspersed repeats, i.e. the Alu family and the L1 family. Alu repeats, a family of SINEs, typically are about 300 bp in length and are composed of two head-to-tail repeats of a sequence which shares significant homology to 5' and 3' portions of 7SL RNA (Quentin, 1992). The other predominant repeat family in the human genome, as well as other mammalian genomes, is the L1 repeat family (a family of LINEs) and members from this family vary in length, from 1.5-6 kb (Adams et al., 1980). Other less frequently encountered interspersed repeats include the MIR

(Mammalian-wide Interspersed Repeat), the MER (Medium Reiteration frequency Repeat) and the LTR (Long Terminal Repeat). Compared to human, a different set of SINEs are present in mice. The two major family of SINEs in the mouse genome are the B1 and B2 families. B1 is thought to be the Alu-equivalent in rodents (Maraia, 1991). It is composed of ~140 bp of consensus sequence followed by a short A-rich region, sharing ~78% sequence homology with both 7SL RNA and Alu repeats (Maraia, 1991). In contrast, the B2 family of repeats, which is specific to the rodent species, is abundant in the mouse genome (Kass et al., 1997). It is estimated that about 800,000 copies of B2 repeats are present in the mouse genome (Kass et al., 1997). The consensus sequence of mouse B2 repeats is about 180 bp in length and shares significant homology with serine tRNA genes (Daniels and Deininger, 1985). Another less abundant SINE is the rodent identifier (ID) which also is thought to be derived from tRNA genes (Daniels and Deinger, 1985).

Striking similarities exist between human and mouse genomes: It is estimated that the human and the mouse diverged about 80 million years ago (Li, 1997), which is relatively short compared to a life history of billions of years. Given this short time span, human and mouse genomes are expected to be conserved at least in some aspects. However, the exact extent of this conservation was not known until recent years. Numerous studies now have shown that the coding regions of genes generally are well conserved between human and mouse (e.g. Whitehead and Sackstein, 1985). Comparison of over one thousand human and mouse cDNA sequences revealed an average of 85% identity of coding sequences (Makalowski et al., 1996). In addition to the conservation of individual genes, organization of genes often is conserved between human and mouse. For example, Oakey et al. (1992) constructed physical maps for a syntenic region of human and mouse chromosomes 1 which spans 11 and 13 Mb, respectively, and showed that the 14 genes in the mapped regions are at least 90% identical in human and mouse, based on gene content, order and spacing. To date, about 200 linkage groups have been found to be conserved in human and mouse (DeBry and Seldin, 1996). A limited number of large scale

sequencing experiments (see Section 4.1) have confirmed the high level of conservation in gene organization. Recent comparative mapping experiments also have demonstrated extensive similarity between human and mouse genomes (Figure 1.3; reviewed in Carver et al., 1997). Nevertheless, the mapping experiments also revealed significant differences between human and mouse genomes. Many syntenic regions are interrupted by insertions, deletions, transpositions, inversions and other types of rearrangement (Carver et al., 1997). Given the similarities and differences between the human and mouse genome, sequencing these two genomes are expected to provide evolutionary and functional information for both human and mouse sequences.

Although much has been learned about the human and other genomes in recent years, many unanswered questions still remain. First, a few thousand human genes have been characterized to date. What are the other genes and their functions? Second, how are the genes and other sequence segments organized into a functional genome? Third, as many human diseases are associated with alterations in genetic material, which specific genes cause which diseases? A full understanding of all aspects of the human genome is needed to understand the practical as well as theoretical aspects of genome structure and function. Therefore, the human genome project (with this dissertation being part of this project) should help to answer these questions.

1.2 DiGeorge Syndrome

DiGeorge syndrome (DGS), first described by DiGeorge in 1965 (Cooper et al., 1965), is a congenital disorder characterized by two or more of the following symptoms (Conley et al., 1979): (1) complete or partial absence of the thymus or cellular immune deficiency or both, (2) symptomatic hypocalcaemia or parathyroid hypoplasia or both, or (3) congenital heart disease. It is estimated that the syndrome occurs in one in every 4000 new births. Patients with DGS often die within 3 month of life, primarily due to congenital heart disease or chronic infection. The symptoms for DGS overlap with those of velo-

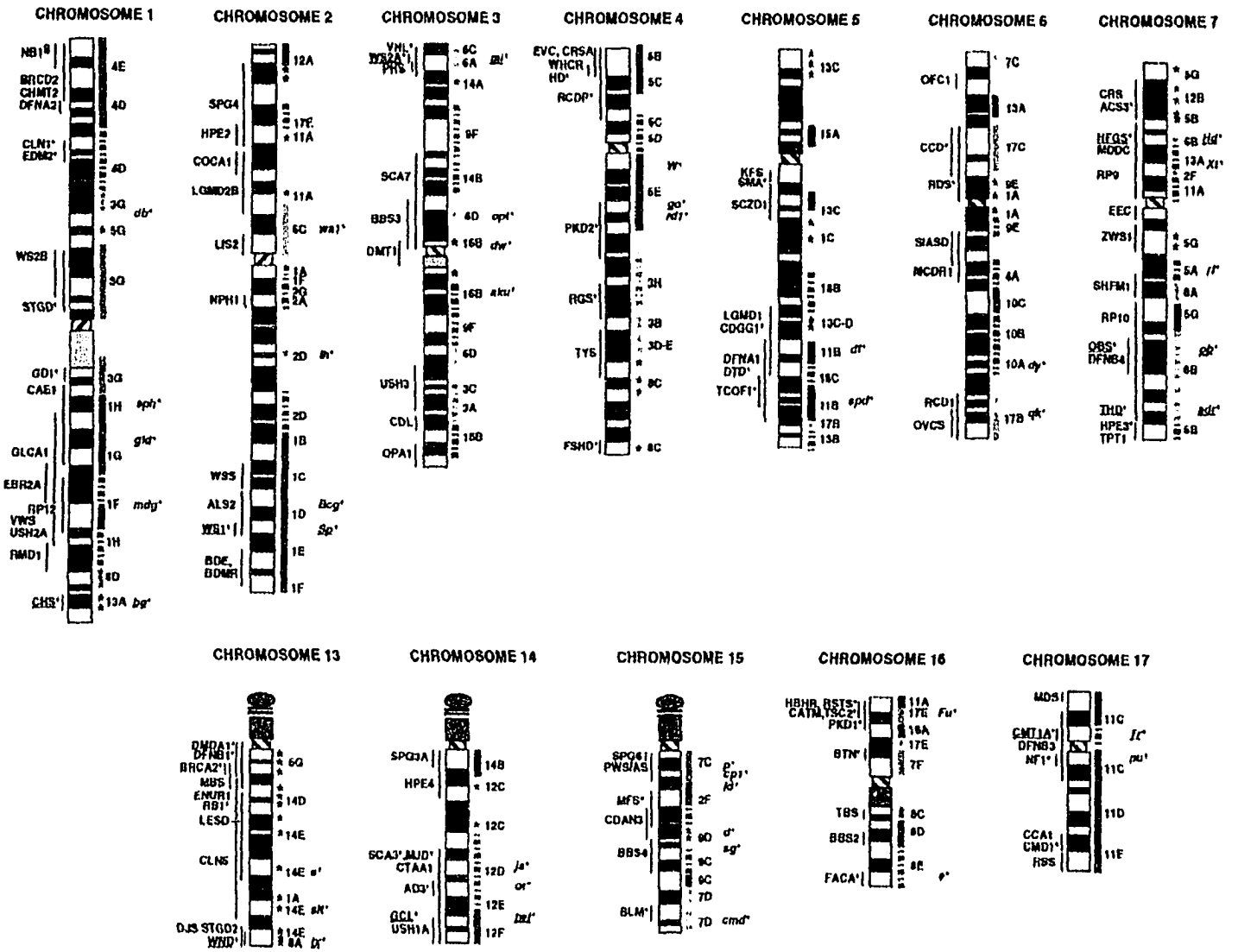


Figure 1.3 A comparative map of the human genome and the homologous mouse regions

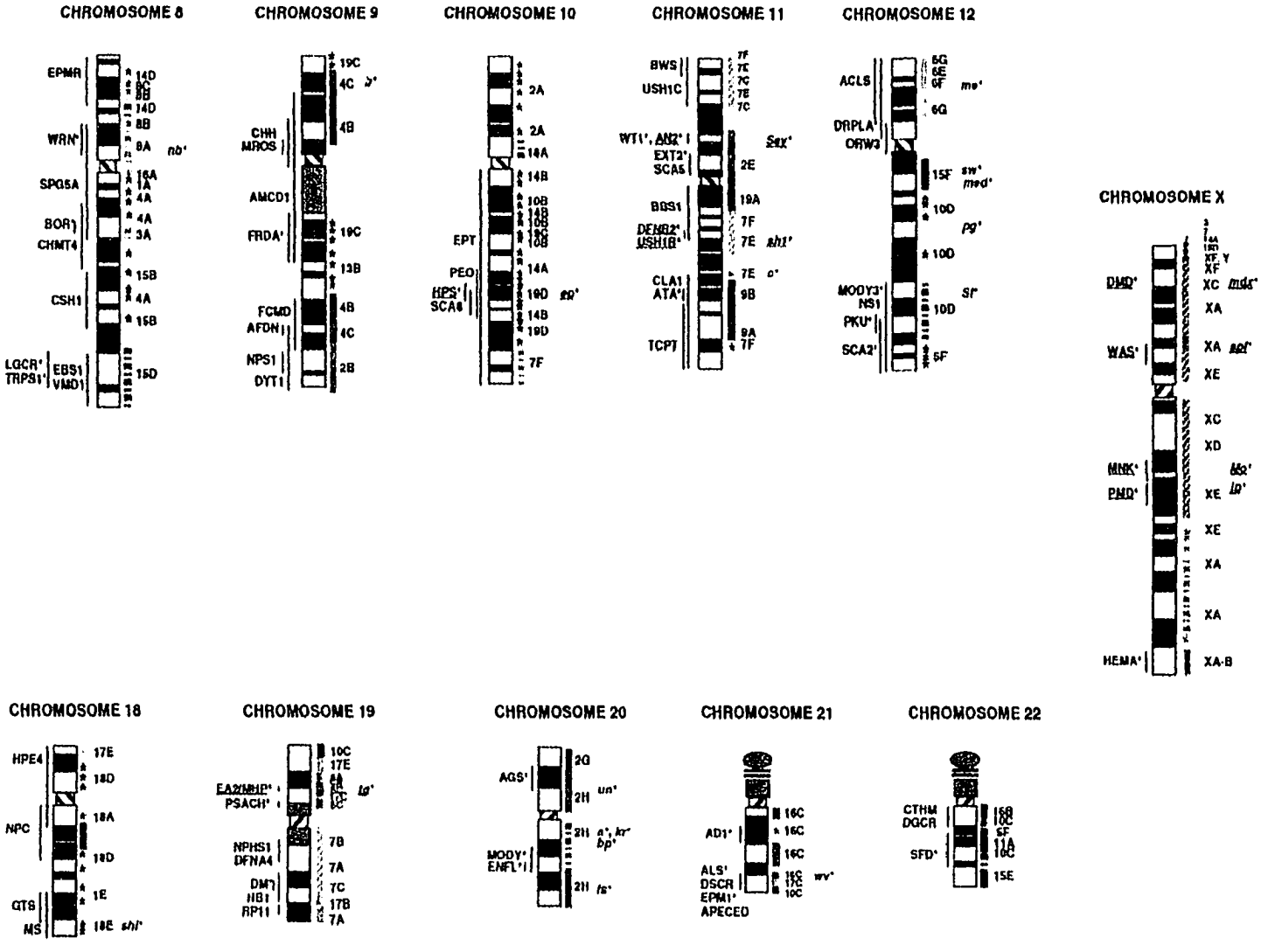


Figure 1.3 continued.

Figure 1.3 (continued) An ideogram of the human karyotype is shown, with murine homology region indicated to the right of each map. Solid homology regions are indicated by solid bars. Hatched bars denote regions to which fewer makers have been mapped and for which human and/or mouse map assignments have been established with less accuracy. Asterisks (*) indicate the approximate positions of single markers that have been mapped to both species. Each murine homology region is color-coded to correspond to specific mouse chromosomes. The cytogenetic positions determined for several human diseases, represented by locus symbols, are indicated to the left of each human map. A selection of locus symbols corresponding to mouse mutations that are known to map within a particular homology region are included at far right of the map in italics. An asterisk denotes the human gene or mouse mutant has been cloned; underlining indicates that human diseases and murine model, aligned in parallel positions on the map, are known to be associated with orthologous genes. (Adapted from Carver and Stubbs, 1997).

cardiofacial syndrome (VCFS) whose major features include cleft palate, cardiac anomalies, typical facies and learning disabilities (Shprintzen et al., 1981). The VCFS also is associated with microdeletions of chromosome 22q11 (Scambler et al., 1992; Kelly et al., 1993), as with the DGS which will be discussed later in this introduction. Due to the association of these disorders with human chromosome 22 deletions, these disorders have been given the acronym CATCH 22 (Cardiac defect, Abnormal faces, Thymic hypoplasia, Cleft palate, Hypocalcaemia and 22q11 deletions) (Glover, 1995). During the past thirty years, significant progress has been made toward a full understanding of the etiology, cytogenetics and molecular genetics of this disorder.

DiGeorge syndrome is associated with an abnormality in human chromosome 22: The association of DGS with human chromosome 22 started in 1981 when Chapelle et al. (1981) reported that four infants in a Finnish family displaying features of DGS had an unbalanced autosomal translocation involving chromosomes 20 and 22. Since this translocation resulted in the loss of proximal 22q and a derivative of chromosome 20, the disruption of an unknown gene in chromosome 22 was suggested to be the cause of DGS. Kelley et al. (1982) then reported three additional cases of unbalanced translocations, involving chromosome 22 and one of three other chromosomes 3, 10 or 20. Subsequent studies revealed additional DGS cases and all were associated with chromosome 22 abnormalities (e.g. Mascarelllo et al., 1989; Greenberg et al., 1984, 1988). However, there also were several cases of DGS which did not have a translocation on chromosome 22 but rather had a monosomy of 10p13 (Greenberg et al., 1988), or deletion in chromosome 17 (Takahashi et al., 1992). Although there are DGS patients without any apparent visually detectable chromosome abnormalities (Greenberg., 1993), many of these patients have an interstitial deletion in chromosome 22q11.

Most DGS patients have a deletion in chromosome 22q11. In their study of 27 DGS cases, Greenberg et al.(1984) noticed that based on cytogenetics, one case showed an interstitial deletion of 22q11. This study then was followed by several reports (e.g.

Scambler et al., 1991; Driscoll et al., 1992) showing that the vast majority of DGS cases contained deletions within chromosome 22q11. Using dosage analysis and FISH, Carey et al. (1992) reported the detection of deletions in 22q11 in 33 of 35 DGS cases. Using RFLP and/or dosage analysis, Driscoll et al. (1992) reported that a deletion in 22q11 was detected in all 14 probands studied. This study also delineated a common deleted region of approximately 1.5 Mb between markers D22S36 and BCRL2 as the DiGeorge Syndrome Critical Region (DGCR). Later, Driscoll et al. (1993) reported studies on another 76 DGS/VCFS patients and found that 83% had deletions in the DGCR region. These and other reports strongly suggested that one or more genes in the DGCR region was responsible for most, if not all, cases of DiGeorge syndrome. Thus, one objective of this dissertation research was to locate known genes and predict putative genes in the DGCR region through large scale genomic sequencing.

The gene(s) responsible for DGS remains to be determined. Determination of the gene(s) responsible for DGS is hindered because there is no obvious correlation between the extent of 22q11 deletion and the severity of the disease (Glover, 1995), suggesting that more than one gene may be involved in the syndrome. Although more than a dozen genes already have been mapped to the region as a result of transcript mapping (e.g. Gong et al., 1996) and sequencing studies by members at this lab, intensive efforts now are underway to evaluate the association of these genes with DGS. At this stage, several genes have been proposed to be candidates for DGS, including the HIRA gene (Halford S. et al., 1993), the CLTCL gene (Holmes et al., 1997) and the GSCL gene (Gottlieb et al., 1997). To evaluate these candidates, a suitable model system will be needed to determine the exact cause of DGS, as well as for future therapeutic research. Since the mouse often has been used as a model organism in investigating many human diseases, another objective of this dissertation research was to compare a portion of the DGCR region with its syntenic region(s) in mouse through large scale genomic sequencing. Such a comparison should provide new information about similarities and differences in these syntenic genomic

regions and knowledge of the mouse sequence will aid in the production of transgenic mice.

1.3 Motor neuron disease

Motor neuron disease (MND) is a fatal neurological disorder characterized by degeneration of spinal and bulbar motor neurons (Jones et al., 1993). It is estimated that one in 20,000 people is affected by the disease and about 10% of all cases have a family history (Brady, 1995). Although the most common form of the disorder, amyotrophic lateral sclerosis (ALS), has been known for over 150 years (Bradley, 1995), its cause is unknown and no cure is available. In the past several years, significant progress has been made in locating the genes responsible for inherited human MND. Since more than 70 types of the disease have been observed, considerable genetic heterogeneity may exist (Brady, 1995). For example, genes responsible for ALS have been mapped to regions in chromosome 21q22 (Siddique et al., 1991) and 2q2-q35 (Hentati et al., 1992). Genomic regions in chromosome 5q13 also may contain genes responsible for another type of MND, spinal motor atrophy (SMA) (Brzustowicz et al., 1990). Collaborators at University of Michigan have developed a mouse model (mnd2) which shows rapidly progressive paralysis, severe muscle wasting and other symptoms of MND (Jones et al., 1993). Recently, this group mapped the mnd2 mutation to the central region of mouse chromosome 6 and the clones covering this region were supplied to this laboratory. As part of the efforts to identify the genes responsible for the mnd2 mutation, one objective of this dissertation research was to perform large scale genomic sequencing and gene prediction in two BAC clones which harbor inserts from the mouse mnd2 mutation region (b245c12) and its syntenic region in human chromosome 2 (h173).

1.4 Large scale genomic sequencing

Basic methods: Two basic methods, the Maxam and Gilbert method (Maxam and Gilbert, 1977) and the Sanger method (Sanger et al., 1977), have been developed to determine the sequence of DNA. In the Maxam and Gilbert method, the DNA sample is divided into four aliquots and then four different chemical reactions are performed to cleave the DNA only at a particular base (A, G, C or T) or base type (A and G, or C and T). With careful control of the reactions, all possible fragment lengths are generated by the four reactions. Since each fragment has a known base or combination of bases at its end and the 5' end of the original DNA sample is radiolabeled, the sequence of the original DNA sample can be deduced after electrophoresis and subsequent autoradiography.

The Sanger DNA sequencing method (also called dideoxy sequencing), takes advantage of the enzymatic activity of DNA polymerases. Instead of generating DNA fragments by chemical methods as in the Maxam and Gilbert method, DNA fragments are synthesized by DNA polymerase. Again, four separate reactions are performed, each containing all four deoxynucleotides, the building blocks of a DNA molecule, and one of four dideoxynucleotide analogs, 2', 3'-dideoxynucleotides (ddNTPs, terminators). Since DNA polymerase can not add nucleotides to a polynucleotide DNA if the 3' hydroxyl group is missing, chain elongation is terminated when the added terminator (ddNTP) is esterified to the 3' end. Thus, when a small amount of terminator, say ddATP, is added to a reaction, the product will be a mixture of DNA molecules of different length all terminating at ddATP. When four reactions, one with each of the four different dideoxynucleotide terminators, are performed and the products are electrophoresed in four parallel lanes, the nucleotide sequence of the original DNA molecule can be deduced. The Sanger method has been widely used in large scale genomic sequencing and is used in this dissertation research.

Recent improvements to the dideoxy method: Several modifications have been made to the original Sanger method to improve both the efficiency and accuracy of DNA sequencing. Modifications in three areas are especially important. First, the original

radiolabelling of primers or the extension products was replaced by fluorescent labeled primers (Smith, 1986) and then by fluorescent labeled terminators (Prober, 1987). These improvements not only eliminated the possible dangers associated with radioactivity but also made it possible to perform four reactions in one tube. Second, linear constant temperature sequencing was replaced by cycle sequencing (Carothers et al, 1989; Smith et al., 1990). In cycle sequencing (Figure 1.4), sequencing reactions are thermally cycled, with each cycle involving a different temperature, one for each denaturation, primer annealing and chain elongation step. Cycle sequencing has the advantage over constant temperature sequencing in that the amount of DNA template required is significantly less and that more product per mole of template can be produced and thus more signal can be obtained. Third, new enzymes have been developed to replace the Klenow fragment of *E. coli* DNA polymerase I used in the original Sanger method. For example, Tabor and Richardson (1989) introduced the modified bacteriophage T7 DNA polymerase, which has an increased polymerase activity and higher efficiency in incorporating ddNTPs than other DNA polymerases. Innis et al. (1988) then introduced the Taq (*Thermus aquaticus*) DNA polymerase, which is active over a broad range of temperatures with an optimum reaction temperature of 72°C and is thus suitable for being used in cycle sequencing. However, the wild-type Taq DNA polymerase has a high discrimination against incorporating ddNTPs and thus requires a higher concentration of dideoxy terminators. Tabor and Richardon (1995) then replaced the phenylalanine residue at position 667 of wide-type Taq polymerase with a tyrosine (Taq F667Y), causing the polymerase to incorporate ddNTPs much more efficiently. New enzymes carrying the F667Y and other mutations, such as KlenTaquenase TR and Taq-FS, now have been developed and are widely used in large scale sequencing. Although several automated DNA sequencing instruments have been developed, the ABI automated sequencers (Perkin Elmer) were used in this dissertation research. During the past few years, these instruments have been upgraded from the original ABI Model 370A to the Model 373 when this research begun, and recently to the

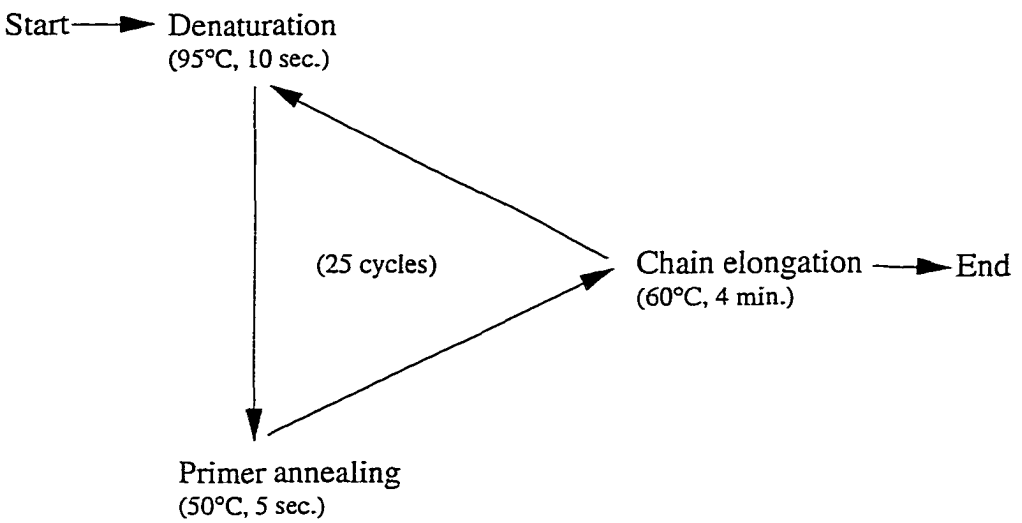


Figure 1.4 Diagram showing the cycle sequencing strategy
Typical cycling conditions are shown in parenthesis

Model 377 (Perkin Elmer, 1995). The ABI sequencers consist of three major components, an electrophoresis apparatus, a fluorescent signal detection unit and a computer. The major differences between the Models 377 and 373 include: (1) A photomultiplier tube in Model 373 is replaced by a charge coupled device (CCD) camera in Model 377 to detect light emitted from laser-excited fluorescent dyes, leading to higher sensitivity of the detection system; (2) In addition to the built-in heat-transfer plate located in the back of a mounted gel plate in Model 373, another heat-transfer plate is added to the front of the gel plate in the Model 377. With this more efficient temperature control, the electrophoresis can be performed at higher voltage in Model 377 as compared to that in Model 373 and (3) Longer and thinner gels are used in Model 377 (48 cm long and 0.2 mm thick) as compared to those in Model 373 (36 cm long and 0.4 mm thick). Also, the transition from the Model 373 to Model 377 was accompanied by a 2-fold increased speed of the detector's left to right movement and a 2-fold more focused lens on the detector which facilitate increased pixel resolution and allow an increase in the number of lanes per gel from 36 to 48 and now 64. Finally, a 25% increase in base resolution was obtained by lowering the collection read position and using longer gel plates. These changes increased the well-to-read length from 36 cm on the Model 373 to 48 cm on the model 377 and resulted in increasing of the read length per sample from approximately 400 bases to over 550 bases per run.

Sequencing large DNA molecules: Using the method described above, about 500-600 bp sequence per sequencing reaction can be obtained. How can we sequence a large insert of, say, 100 kb? One immediate answer might be "repeat the reaction". Indeed, this is one strategy to sequencing large DNA molecules, that has been implemented by direct sequencing with custom synthesized primers. This approach is not widely used because the cost of primers is prohibitive. However, this approach, which requires synthesis of a new primer from the end of the previous read and used in the next round of dideoxy sequencing, is useful for primer walking of small, 1-2 kb regions. But, to sequence a large insert, we would need many rounds of reaction-primer synthesis-reaction. To avoid this

problem, a second strategy, namely the random shotgun sequencing strategy, is widely used to collect the initial sequencing data for large genomic inserts which then can be completed by primer walking. With this strategy, a large DNA molecule firstly is broken into short fragments of about 2 kb by physical shearing. These fragments then are cloned into plasmid-based sequencing vectors. Since the sequence of the plasmid used in the cloning is known, universal primers can be made from the vector sequence near the cloning site and these primers can be used to sequence the ends of the cloned DNA fragments. After the end sequences for the fragments are obtained, computer programs assemble the individual sequences into a large sequence representing the sequence of the original insert. One drawback of the random shotgun strategy is that much more sequence data must be collected to insure an accurate assembly of the relatively short sequences into the original sequence based on sequence overlap. However, since the redundancy ensures the higher overall accuracy of the obtained final sequence, the higher overall efficiency of the random shotgun strategy made it the major method by which the vast majority of data in this dissertation research was collected prior to sequencing via custom synthesized primers for direct closure of any sequence gaps.

Chapter II
Material and Methods

2.1 Overall sequencing strategy

The shotgun sequencing approach used in this dissertation research entailed seven major steps. (1) A large scale DNA preparation (typically 0.5-1.0 mg, depending on size of insert) was performed for each clone harboring the DNA insert of interest; (2) A portion of the large scale DNA preparation was physically sheared into fragments of 1.5-3 kb through nebulization. The fragments then were size-selected with low-melting agarose gel electrophoresis and end-repaired; (3) The DNA fragments were ligated into pUC18 vectors and transformation then was performed with *E. coli* XL1-Blue as the host strain; (4) Mini-prep isolations were performed to obtain subcloned DNA; (5) Cycle sequencing reactions were performed and sequences of the DNA subclones were determined with automated sequencer ABI 373 or 377; (6) These sequences were assembled into large contigs with assembly program XGAP or PhredPhrap. (7) When sequence gaps existed, directed sequencing with custom synthesized primers was used the majority of the times to close those gaps. The overall strategy is diagrammed in Figure 2.1 and detailed in the following sections.

2.2 Large scale DNA isolation

A modified procedure (Roe, 1995) which was based on the alkaline lysis method (Birnboim and Doly, 1979) was used as the basic method for the large scale isolation.

In the first step, a colony of *E. coli* which harbored the target DNA of interest was transferred into a 12X75 mm Falcon tube containing 3 ml of LB medium supplemented with appropriate antibiotics. After incubation at 37°C for 8-10 hours with shaking at 250 rpm, the culture was transferred to a 250 ml Erlenmeyer flask containing 50 ml of same medium and incubated for another 8-10 hours as before. The culture then was divided into

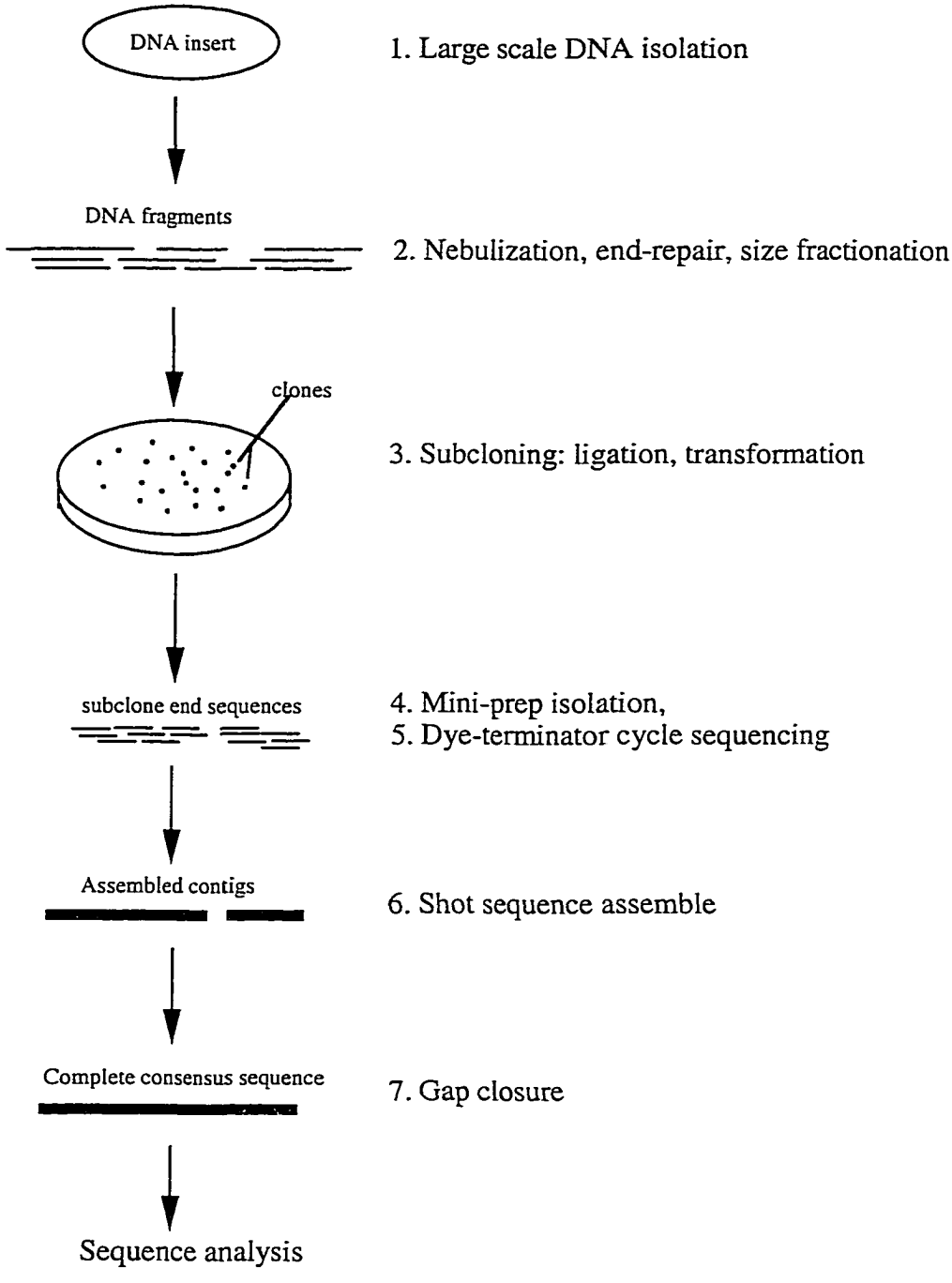


Figure 2.1 Diagram showing the overall large scale sequencing strategy

two 2 liter flasks, each of which contained one liter of the same media, and further incubated for another 8-10 hours. Cells were harvested by centrifugation at 7000 rpm for 20 minutes into 500 ml centrifuge bottles in a RC5-B centrifuge. The cells then were stored at -70°C at this point.

The insert DNA in the cultured cells was isolated by alkaline lysis followed by salting out high molecular weight genomic DNA and proteins followed by ethanol precipitation. Specifically, the cells were gently resuspended in 70 ml of GET/Lysozyme solution. After the cells were incubated at room temperature for 10 minutes, 140 ml of alkaline lysis solution was added. The solution then was gently mixed and incubated for 5 minutes in an ice-water bath. *E.coli* genomic DNA and proteins were precipitated by adding 105 ml of 3M of NaOAc or KOAc, followed by gentle mixing and incubation in an ice-water bath for 30-60 minutes. The lysate was cleared of insoluble SDS, proteins, membranes and *E.coli* genomic DNA by pouring the lysate through a double-layer of cheesecloth. Further clearing of the supernatant was performed by centrifugation in a 250 ml centrifuge bottle at 10,000 rpm for 30 minutes at 4°C using the RC5-B centrifuge. This last step was repeated until the supernatant did not form a significant amount of residue after centrifugation.

The supernatant, which contained the desired DNA insert, was further purified by several additional steps. First, the clear supernatant obtained in the previous step was transferred into a 500 ml bottle and DNase-free RNase A and RNase T1 were added, with the final concentration of RNase A and RNase T1 being 40 ug/ml and 40 U/ml, respectively. The solution then was incubated in a 37°C water bath for 30 minutes. After incubation, an equal volume of isopropanol was added. After mixing, the solution was incubated at room temperature for 5 minutes followed by centrifugation at 9000 rpm for 30 minutes in the RC5-B centrifuge. The supernatant was decanted and the DNA pellet drained. To perform purification using diatomaceous earth, the pellet was resuspended in 20 ml 10:1 TE buffer followed by addition of 40 ml of de-fined diatomaceous earth in

guanidine-HCl (100 mg/ml). The mixture then was incubated at room temperature for 5 minutes with occasional mixing. The diatomaceous earth with bound DNA was obtained by centrifugation at 9,000 rpm for 10 minutes in the RC5-B centrifuge and removal of the supernatant. The pellet then was resuspended in 40 ml of diatomaceous earth-wash buffer and centrifugation was performed as above. After dried in a vacuum oven, the pellet was resuspended in 20 ml of 10:1 TE buffer and bound DNA was eluted by incubation at 65°C for 10 minutes with intermittent mixing. The diatomaceous earth was removed by centrifugation at 9,000 rpm for 10 minutes in the RC5-B centrifuge. The DNA in the supernatant then was precipitated in 35 ml Corex glass tubes by addition of 2.5 volumes of cold 95% ethanol/acetate. After dried in a vacuum oven, the pellet was resuspended in 2 ml of 10:0.1 TE buffer and DNA concentration determined by absorbance readings at 260 nm or by agarose gel electrophoresis.

One important parameter in evaluating the quality of a large scale preparation for shotgun cloning is the extent of contamination by *E.coli* genomic DNA. Since different researchers or even the same researcher often observes significantly different amounts of contamination in different preparations, an attempt to determine the most critical step(s) in this isolation procedure that results in *E.coli* DNA contamination was performed. Since DNA purification using diatomaceous earth was thought to lower the extent of *E.coli* contamination, different concentrations of diatomaceous earth (50 mg/ml, 100 mg/ml, 200 mg/ml) was used in the large scale isolation of cosmid 83e8, with all the other steps being the same as described above. After shotgun sequencing of each of the three preparations, it was found that all three preparations had *E. coli* contamination ranging from 5 to 10%, suggesting that the diatomaceous earth at the three different concentrations was equally effective or, alternatively, had no effect in lowering the *E. coli* contamination. During large scale isolation of the next project, p201m18, the extent of *E. coli* contamination for preparations with and without the diatomaceous earth wash step was compared and it was found that the *E. coli* contamination for both preparations was about the same (12% vs.

11%). These results indicate that the diatomaceous earth wash step had little or no effect in the further purification of large scale DNA preparations in terms of the extent of *E. coli* contamination and thus, this step was eliminated from subsequent large scale isolations performed during this dissertation research. This modification significantly reduced the time and cost needed for the preparation of large scale DNA. During the large scale DNA isolation of p201m18, incubation at the alkaline lysis step also was performed for different time lengths (5 min, 10 min, 15 min) and the results indicate that prolonged incubation may result in a higher extent of *E. coli* contamination (11% for 5 min, 12% for 10 min and 23% for 15 min). Other large scale DNA isolations performed in this dissertation research used an incubation time of 2-5 min in the alkaline lysis step and the extent of *E. coli* contamination was generally found to be about 5%.

2.3 Physical shearing, end repair and size fractionation

Physical shearing of large DNA molecules into fragments of 1.5-3 kb was performed by using a device called a nebulizer. With a solution containing approximately 50 ug of DNA, 200 µl of 10X TM buffer and 0.5-1 ml of sterile glycerol added to the bottom of a nebulizer, sterile ddH_2O was added to a final volume of 2 ml. Nubulization was performed in an ice-water bath for 2.5 minutes with nitrogen pressure at 8-12 psi, depending on the size of the original DNA molecules and the desired size range of DNA fragments. The nebulized sample then was collected by briefly centrifuging the nebulizer at 1500 rpm in a table top centrifuge. The DNA sample then was ethanol precipitated, vacuum dried, and resuspended in a total of 27 µl of 1X TM buffer. Care was taken not to overdry the sample otherwise difficulty may be encountered in later subcloning steps.

Since the nebulization step generates single-stranded ends, these ends were repaired by DNA polymerase. Specifically, a solution containing 27 µl of the nebulized DNA solution, 5 µl of 10X kinase buffer, 5 µl of 10 mM rATP, 7 µl of 0.25 mM dNTPs, 3 units of T4 polynucleotide kinase and 10 units of Klenow DNA polymerase was prepared

and incubated at 37°C for 30 minutes. This reaction also ensured that the 5' ends of DNA fragments were phosphorylated.

The reaction mixture from the last step, with 6 μl of agarose loading buffer added, was loaded into a 0.7% low melting temperature agarose gel. Size markers of Hind III-digested λDNA and Hae III-digested ϕX174 DNA were loaded in parallel with the sample lanes. Electrophoresis was performed for 30-60 minutes at 100-120 mA. After electrophoresis, DNA fragments with sizes in the range 1.5 - 3 kb were excised from the gel into a microcentrifuge tube with the aid of a long wave length UV light in a dark room. After being frozen at -70°C for at least 15 minutes, the gel slice was melted by incubation in an Eppendorf tube at 65°C for about 10 minutes. An equal volume of TE-saturated phenol then was added to the tube and the tube was vortexed for 5 minutes. After the tube was centrifuged at 12,000 rpm for 5 minutes at room temperature in a microcentrifuge to separate the aqueous and organic phases, the upper aqueous phase was transferred into a clean tube. The sample then was further extracted with half volume of the phenol and half volume of chloroform. A final round of extraction was performed using an equal volume of chloroform or ether. The purified sample then was ethanol precipitated, and rinsed with 75% of ethanol. After dried, the pellet was resuspended in 20 μl of 10:0.1 TE buffer and the sample was ready for ligation.

2.4 Shotgun library construction

Library construction involves two steps: ligation and transformation. The ligation step was performed in a solution containing 100-1000 ng of DNA fragments, 20 ng of SmaI-linearized and calf intestinal alkaline phosphatase-dephosphorylated pUC18 vector, 1 μl of 10X ligation buffer and 1 μl of T4 DNA ligase, with sterile ddH_2O added to a total volume of 10 μl. The reaction mixture then was incubated at 4°C overnight. To obtain optimal results, several different ligations generally were performed, each with a different amount of DNA fragments. A parallel reaction with AluI-digested cosmid DNA also was

performed in parallel as the positive control while a negative control ligation was performed in the absence of insert DNA fragments to test for any possible vector self ligation.

Before the transformation experiment was performed, competent cells, cells which have the increased ability to take up foreign DNA, were prepared. Two methods of preparing competent cells were used during this dissertation research. The $CaCl_2$-based method was used to prepare competent cells for transformation by heat shock. The other method was used to prepare competent cells for electroporation.

With the first method, a frozen glycerol stock of *E. coli* strain XL1 Blue was thawed and transferred to a flask containing 50 ml of pre-warmed 2xTY media. After incubation for 1 hour without shaking in a 37°C water bath, the culture was further incubated for 2-3 hours at 37°C with shaking at 250 rpm. The culture then was transferred into a 50 ml polypropylene centrifuge tube, and the cells recovered by centrifugation at 3000 rpm for 8 minutes at 4°C in a GPR centrifuge. The cells then were resuspended in 20 ml of cold, sterile 50 mM calcium chloride and incubated in an ice-water bath for 20 minutes. The cells then were collected again by centrifugation as above and resuspended in 4 ml of the same calcium chloride solution. The final solution contained competent cells ready for transformation.

The competent cells used for electroporation were prepared by spreading glycerol stock of XL1 Blue MRF' cells on a LB plate (supplemented with 20 ug/ml of tetracycline) in the first step. After the plate was incubated at 37°C overnight, colonies from the plate were used to inoculate 3 ml of YENB broth supplemented with 12.5 ug/ml of tetracycline. After the culture was incubated overnight, 5-10 μl of the culture was inoculated into 1 liter of fresh YENB medium. The culture then was incubated with shaking at 250 rpm until the A_{600} for the culture was about 0.5. Cells were harvested in four 500 ml centrifuge bottles by centrifugation at 5000 rpm in the RC5-B in GS3 rotor for 10 minutes at 4°C. The cells then were washed twice with 100 ml of cold sterile water and once with 40 ml of 10% cold

sterile glycerol solution. Finally, the cells were resuspended in 2-3 ml of cold sterile glycerol solution and ready for use fresh or frozen at -70°C for future use.

For transformation using the competent cells prepared with calcium chloride, the 10 µl ligation reaction was added to a 12X75 Falcon tube containing 0.1-0.2 ml of competent cells. After the cells were gently mixed, the tube was incubated in an ice-water bath for 40-60 minutes. The cells then were heat shocked by incubation at 42°C for 2 minutes. With 1 ml of fresh 2xTY added, the transformation mixture was incubated at 37°C with shaking for 30 minutes. The cells were collected by centrifugation and gently resuspended in 0.2 ml of fresh 2xTY. The sample, with 25 µl IPTG and 25 µl of X-Gal added, was mixed and plated onto the surface of a pre-warmed LB plate supplemented with ampicillin (100 ug/ml). The plate then was incubated at 37°C overnight. For electroporation, two microliters of ligation mix was added to 40 µl of competent cells. After the cells briefly were mixed and incubated on ice, they were electroporated with a *E. coli* Pulser. One milliliter of YENB medium then was added and the sample was incubated at 37°C with shaking for 40 minutes. The remaining steps were the same as for the calcium chloride-based method described above.

2.5 Mini-prep isolation of subcloned DNA

A strategy similar to that used for large scale DNA isolation as described in Section 2.2 was used for mini-prep isolation of the subcloned DNA.

White colonies from the plates prepared as described in Section 2.4 were picked with sterile toothpicks and used to inoculate each well of a 96 well block containing 1.5 ml of TB media supplemented with 100 ug/ml of ampicillin. The block then was incubated at 37°C with shaking at 300 rpm for 24 hours. The cells were harvested by centrifugation and stored at -20°C for later use. Typically, two blocks of cells were prepared in the initial stage of a project and more blocks were prepared after checking the extent of *E. coli* genomic

DNA contamination and the quality of previous ligation reactions. The following isolation steps were performed either manually or with a Biomek 2000 workstation.

After thawing, the cells in each well of the sample block were resuspended in 200 µl of TE-RNase-A solution. The cell suspension was incubated at room temperature for 10 minutes followed by the addition of 200 µl of alkaline lysis solution to each well. After the sample was mixed and incubated at room temperature for another 30 minutes, 200 µl of 3M potassium or sodium acetate was added into each well. The samples then were mixed and incubated at -20°C for 30 minutes. After centrifugation at 3000 rpm in a GPR centrifuge, 400 ml of the supernatant from each well was transferred to another 96 well block. The DNA sample in each well then was ethanol precipitated with one milliliter of ethanol and washed with 500 µl of 75% ethanol. After being drained and dried, the DNA sample was resuspended in 50 µl of ddH_2O and stored at -20°C for later use.

2.6 Cycle sequencing reaction and data collection

During this dissertation research, two types of DNA polymerases, KlenTaq-TR and AmpliTaq-FS, and three types of dye-labeled terminators including Rhodamine-, dRhodamine- and BigDye-labeled terminators, were used in cycle sequencing reactions.

To perform the cycle sequencing reaction, about 50 ug of plasmid DNA and 1 µl of forward or reverse primers (6.5 uM) were pooled with 8 µl of reaction mix (provided by supplier) containing dye-labeled ddNTPs, dNTPs, DNA polymerase and $MgCl_2$ and the total volume was brought to 20 µl with ddH_2O. Typically, 96 reactions were prepared in a 96-well cycle plate simultaneously with the aid of the automatic pipetting machine Hydra96. The cycle reactions then were performed in a Perkin Elmer GeneAmp PCR system 9600 thermal cycler which was programmed typically for 25 cycles of 96°C for ten seconds, 50°C for five seconds and 60°C for four minutes. After incubation, unincorporated dye terminators were removed by chromatography on Sephadex G-50 spin

columns. The purified reaction products then were dried in a vacuum oven and stored at -20°C.

Before the samples were loaded, each sample was resuspended in 1 μl of loading buffer (10 mg/ml blue dextran and 5 mM EDTA in deionized formamide) and heated at 95°C for 2 minutes. The samples then were loaded onto a polyacrylamide gel (5-5.5%). Electrophoresis and data collection then were performed for 3.5 or 7 hours (depending on the mode chosen) by the automatic sequencers ABI 373 or 377. After preliminary processing, the raw data were transferred to a SUN workstation for project assembly and analysis.

2.7 Shotgun sequence assembly

Since the resultant sequences generated from the subcloned DNA were typically 500 bp and their positions relative to the original large insert DNA were unknown, computer programs were used to assemble the shotgun sequences into larger contig(s). During the course of this dissertation research, two programs were used. The first program, XGAP (X-window Genome Assembly Program, Gleeson and Staden, 1991), was used during the early stage of this dissertation research. More recently, a second program, PhredPhrap (an unpublished program package developed by Phil Green at the University of Washington), replaced XGAP as the primary assembly tool. Although both programs assemble short sequences into larger contigs based on sequence overlap, PhredPhrap has a higher accuracy than XGAP, since it estimates the quality of each base in each gel reading and uses this quality information in the assembly process. Nevertheless, assembly errors occasionally occur, especially in complex repeat regions. Such regions then could be corrected by manually breaking the false joins and then performing more sequencing to resolve the problem regions. Project assembly not only resulted in the final sequence of the original insert DNA but as a project was progressing it also could be used to estimate the progress. For example, the average coverage of each base by existing gel

readings and the number of sequence gaps remaining could be estimated from an earlier assembly prior to project completion. This information then could be used to determine whether more shotgun sequencing was needed or whether the gap closure stage should begin. Typically, gap closure for a project was performed when the average coverage of each base in the assembled contigs was between 5- and 7-fold.

2.8 Gap closure

Sequence gaps generally come from two sources. First, because the shotgun approach is based on cloning randomly generated fragments, some regions of the original DNA insert may not be covered by sequenced subclones. Second, some regions, typically GC-rich or polynucleotide repeat regions, are difficult to be sequenced. Typically, the gap closure stage is one of the most challenging steps in large scale sequencing and often requires more effort than the initial shotgun data collection phase.

The gap closure strategy developed in this dissertation research was as follows. First, the gel readings bounding a gap were examined to determine whether the gap resulted from regions difficult for DNA polymerase to read through. Then, primer walking with custom synthesized primers was performed using typically two subclones covering the gap as the templates. If the gap was not covered by any subclone, polymerase chain reaction (PCR) was performed to amplify the region using custom synthesized primers. The PCR product then was purified (see below) and used as the template for primer walking. The use of PCR products directly as templates for primer walking first was exploited and found practical and time-saving in this lab as a result of this dissertation research. Previously, PCR products were purified and subcloned into pUC18 vectors, a procedure that required a series of time-consuming steps involving subcloning and DNA isolation. During this dissertation research, all of the gaps which resulted simply from lack of coverage but not from difficult regions were closed by no more than two rounds of primer walking.

For those sequencing gaps difficult to be closed, a series of further experiments often is needed. Inclusion of 5-10% of DMSO (Winship, 1989), formamide or glycerol allows DNA polymerases to read through a difficult region, presumably by effectively increasing the reaction temperature and denaturation of secondary structures formed in the template (Innis, 1995). Different DNA polymerase/reaction pre-mixes often can have different effects on different templates. For example, a newly introduced reaction pre-mix in which dGTP replaced dITP in the regular reaction pre-mix was found effective when used to sequence G-rich regions. If all the methods failed to close a sequence gap, subclones covering the gap or purified PCR products can be used as the target DNA to perform another round of shotgun sequencing. With the strategy described above, all remaining gaps in the seven projects at the end of the shotgun phase were closed during this dissertation research.

Several methods, which were involved in the gap closure stage or during other parts of this dissertation research, were as follows. The oligonucleotide primers required for primer walking or PCR experiments were synthesized with the Beckman 1000M eight column oligo synthesizer. The oligonucleotides were cleaved from the column and deprotected by gas-phase AMA (Beckman), eluted with ddH_2O and quantified by measuring the absorption at 260 nm.

A PCR reaction mixture typically contained 10-20 ng of target DNA, 100 pmol of each primer, 5 units of AmpliTaq DNA polymerase, 20 nmole of each deoxynucleotide and 10 µl of 10X PCR buffer with total volume of 100 µl. The PCR buffer contained 500 mM KCl, 100 mM Tris-HCl and 15 mM $MgCl_2$ with pH at 8.5. The reaction was performed in a thermal cycler typically programmed for 25 cycles of 95°C for 1 minute, 55°C for 1 minute and 72°C for 2 minutes. After incubation, the reaction was held at 72°C for 10 minutes and then at 4°C until further processing. Modifications were made to the above conditions when amplification of a sequence region was difficult. For example, KCl was

removed from the PCR buffer to successfully amplify a GC-rich simple repeat region of clone p201m18.

Two methods were used to purify the PCR products. The first method used low-melting temperature agarose gel electrophoresis and was performed as described in Section 2.3. The other method used exonuclease I and shrimp alkaline phosphatase to remove unreacted oligonucleotide primers and dNTPs (http://genome.wustl.edu/gsc/Protocols/EXOSAP.shtml). With this method, 10U of exonuclease I and 2U of shrimp alkaline phosphatase were added to a 20 μl aliquot of the PCR product. The reaction was performed in a thermal cycler programmed as follows: 37°C for 30 minutes; 80°C for 15 minutes and finally held at 4°C. The samples purified with this second method were used only for sequencing purpose but not for subcloning.

2.9 Sequence analysis methods

At the final stage of a sequencing project, the consensus sequence is analyzed to extract biological information. During this dissertation research, emphasis was placed on identifying known genes and predicting potential genes. As a general strategy, a database similarity search was performed for the first step. The search would reveal matches to known genes or conserved homologous genes from other species, as well as EST database matches. When known genes existed, genomic structures of those genes could be deduced by sequence alignment of genomic sequence with sequences of genes and/or cDNAs already in the database. Second, gene prediction could be performed by either computational methods or by comparison of the contig sequence with sequences of homologous genes and ESTs. Third, potential features and/or functions of known or predicted genes could be explored by methods that include prediction of promoters, transcription factor binding sites, polyadenylation signals, protein structures (transmembrane domains), potentially functional protein motifs and protein targeting sites.

After the known or putative genes were located, other features such as repetitive elements and pseudogenes then could be located.

Database similarity searches were performed using Powblast (Zhang and Madden, 1997), which masks repetitive elements and low complexity regions of a query sequence , and then splits a large sequence into pieces. After performing a BLAST (basic local alignment search tool) search for each piece, Powblast merges and presents the results as if only one single large search was done.

Sequence comparison was performed with Dotter (Sonnhammer and Durbin, 1996) and alignment was performed using several programs including Bestfit or Gap of the GCG (Genetics Computer Group) computer package, or Crossmatch of the PhredPhrap program package.

Computational gene prediction was performed by Genscan (Burge and Karlin, 1997) and Xgrail (X-window based Gene Recognition and Analysis Internet Link, Xu et al., 1994). Although there were many gene prediction programs currently available, Genscan was chosen due to its superior performance to other programs (Claverie, 1997) while Xgrail historically is one of the most popular and robust programs.

Although promoters are one of the most important elements in gene function and there exist at least ten programs available for their prediction, the performance in terms of sensitivity and specificity of those programs generally was less than 50% (Fickett and Hatzigeorgiou, 1997). Two programs with relatively high performance, i.e. NNPP (Promoter prediction by Neural Network, Reese and Eeckman, 1995) and TSSW (Solovyev and Salamov, 1997) were tested for promoter prediction during this dissertation research. While NNPP predicts promoters using a neural network, TSSW predicts based on the presence of a TATA box, distribution of transcription factor binding sites and hexamer preference. Using a test set of sequences containing 24 known promoters (Fickett and Hatzigeorgiou, 1997), NNPP correctly predicted 54% of real promoters while giving 72 false positives. TSSW correctly predicted 42% of real promoters while giving 42 false

positives. Due to the relatively low performance, promoter prediction during this dissertation research generally was performed only on relatively narrow sequence regions where a promoter was expected, such as the 5' region of a predicted gene. Xgrail also was used to locate TATA boxes and other promoter elements.

Compared to the prediction of promoters, the prediction of transcription factor binding sites by different programs generally exhibit high sensitivity but vary in specificity (Frech et al., 1997). For example, the program Signal Scan (Prestridge, 1996) correctly predicted 28 sites but with more than 900 positive matches in its output when applied to a eukaryotic sequence containing 29 known sites (Frech et al., 1997). Using another program, Consinspector (Frech et al. 1993; 1997), 28 sites also were correctly predicted but with only 4 false positives when applied to the same sequences. The higher performance of Consinspector was achieved partly by analyzing not only a relatively short potential binding site as Signal Scan does but also regions surrounding the site (Frech et al., 1997). One drawback of Consinspector is that currently it only provides libraries for 16 families of vertebrate transcription binding sites. Given the better performance of Consinspector over other programs, it was chosen for the prediction of transcription factor binding sites during this dissertation research.

The program TMpred predicts transmembrane domains by comparing a query protein sequence with statistics from a database of membrane spanning protein segments (Hofman and Stoffel, 1993). The performance of TMpred was unknown but had been used previously in transmembrane domain prediction by Suhan and Hovde (1998).

Prediction of functional protein motifs was performed with ProfileScan which is based on comparison of a query sequence with a protein profile database (Gribskov et al., 1987; Bucher et al., 1996). Protein targeting sites were predicted by Psort II which evaluated a number of parameters that include signal sequence and protein secondary structure (Horton and Nakai, 1997). The prediction accuracy of Psort was estimated to be

57% for the yeast and 86% for *E. coli* sequences (Horton and Nakai, 1997) but was not determined for vertebrate sequences.

Finally, searches for repetitive sequences were performed with Repeatmasker (unpublished program of Smit AFA). Although the database search program Powblast also identifies repetitive elements, it could not be performed for certain sequences unless repetitive elements are masked. For consistency, identification of repetitive elements and their statistics were based on results of Repeatmasker.

Chapter III
Sequence Analysis and Discussion

3.1 Introduction

Using the strategy detailed in Chapter 2, a total of seven clones have been sequenced to completion during this dissertation research. Two Cosmid clones (59f8 and 83e8), one BAC clone (72f8) and one PAC clone (p201m18) from the DGCR region in human chromosome 22q11 have been sequenced. One BAC clone (72k21) which is syntenic to the DGCR remutageisisso was sequenced. One BAC clone (b245c12) in the mouse motor neuron disease critical region (mnd2) from mouse chromosome 6 has been sequenced and one BAC clone (h173) from human chromosome 2 which is syntenic to the mouse clone b245c12 also was sequenced. A total of about 720 kb unique nucleic acid sequences have been determined from these seven clones. A summary of clones sequenced is shown in Table 3.1 and their chromosomal locations are shown in Figure 3.1.

Table 3.1 Summary of clones sequenced

Contig name	Clone name	Vector type	Source	Chr. region	Size of inserts
Contig 1	72f8	BAC	Human	22q11(DGCR)	100,067 bps
Contig 2	59f	Cosmid	Human	22q11(DGCR)	42,017 bps
	83e8	Cosmid	Human	22q11(DGCR)	40,676 bps
Contig 3	p201m18	PAC	Human	22q11(DGCR)	162,268 bps
Contig 4	72k21	BAC	Mouse	16	118,243 bps
Contig 5	h173	BAC	Human	2	90,832 bps
Contig 6	b245c12	BAC	mouse	6	162,750 bps

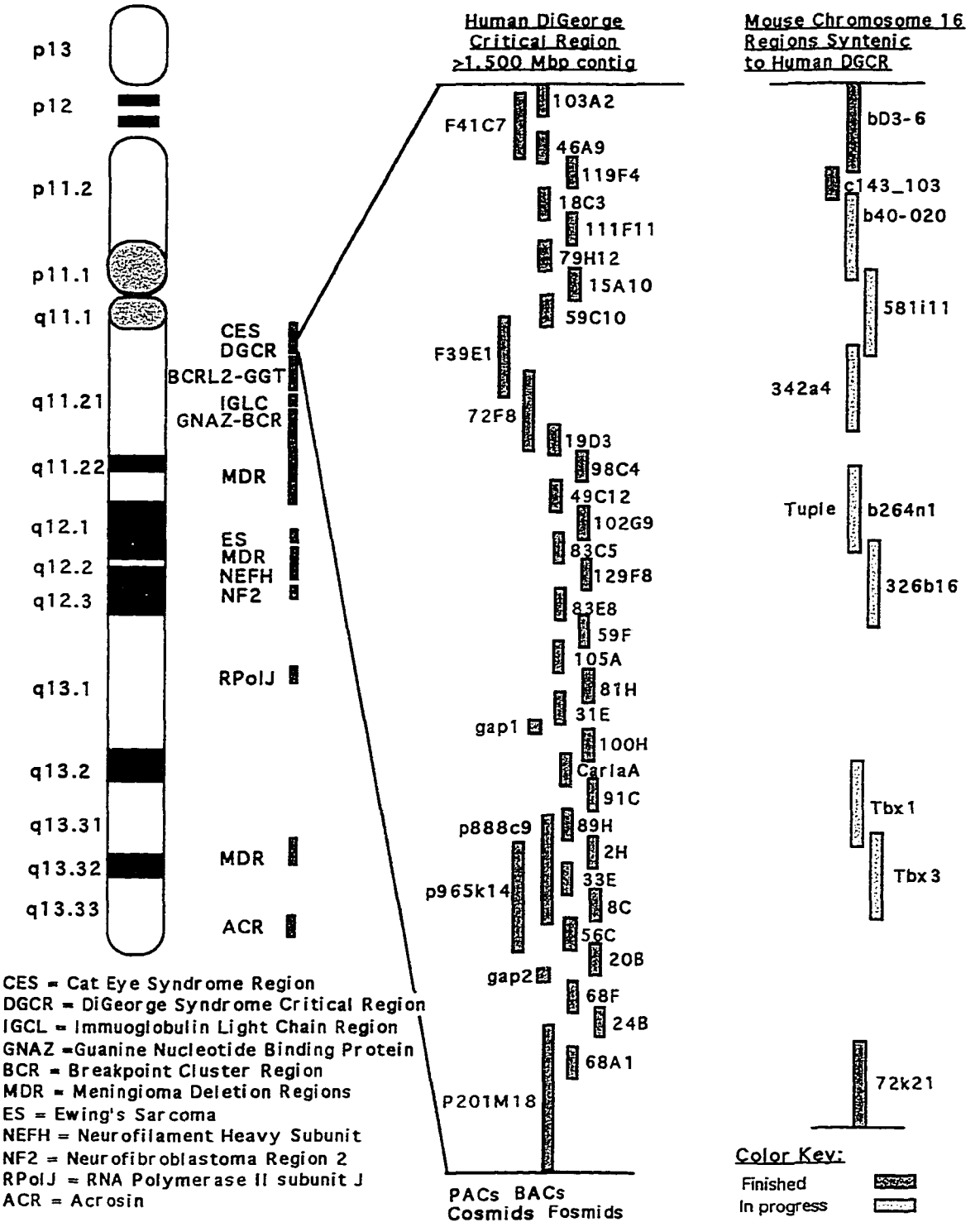


Figure 3.1a Cosmid, BAC, and PAC clones in the DiGeorge Critical Region of Human Chromosome 22 and Syntenic Regions of Mouse Chromosome 16

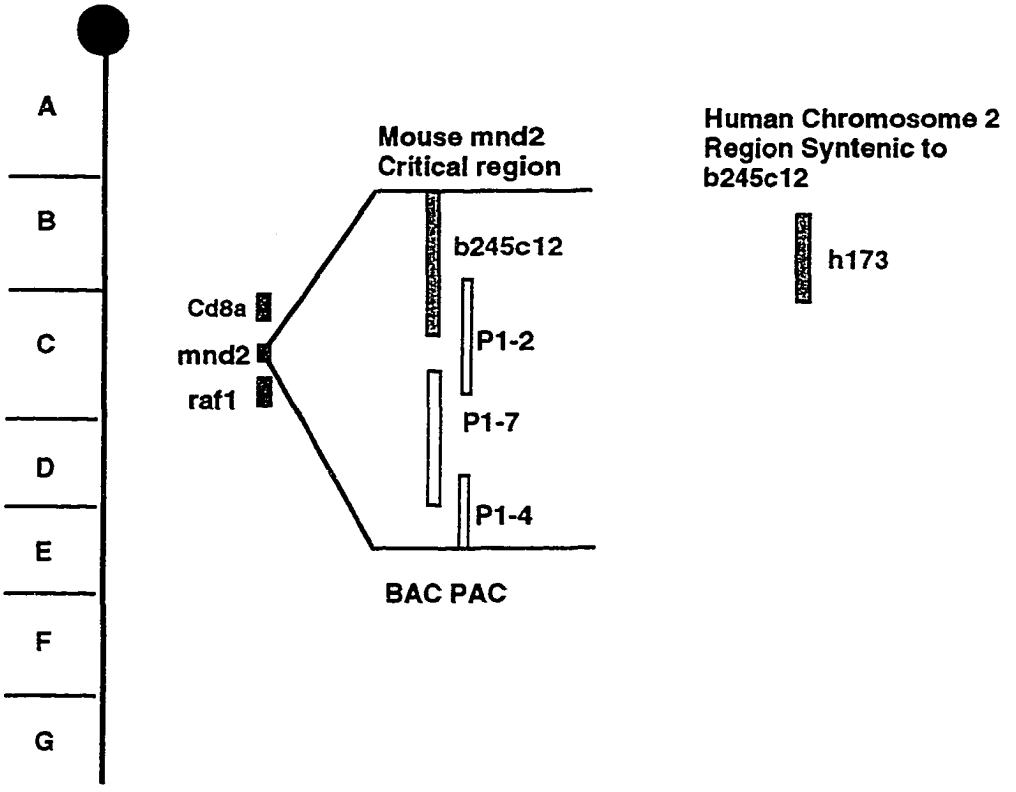


Figure 3-1b Chromosome locations of BAC b245c12 in the mouse mnd2 critical region of mouse chromosome 6 and syntenic region in human chromosome 2

3.2 Analysis of contig 1

This contig is composed of a single BAC clone, 72f8. It overlaps Fosmid f39e1 at its centromeric end and cosmid 19d3 at its telomeric end. Both flanking clones have been sequenced by others from this laboratory. Nine exons from two known genes, specifically the 5' end two exons of the clathrin heavy chain-like gene (CLTCL) and the 3' end seven exons of HIRA gene (Figure 3.2) are contained in this contig.

3.2.1 Human Clathrin heavy chain-like gene (CLTCL)

As shown in Figure 3.2, a Powblast search of the Genbank non-redundant nucleic acid database revealed that two segments (16697-16904 and 32672-32782) of this contig were homologous to several Genbank entries (X95486. L48481, U60803, U41763). These two segments have been confirmed to be the 5' end two exons of a recently characterized human gene, clathrin heavy chain-like gene from human chromosome 22 (CLTCL) (Kedra et al., 1996; Holmes et al., 1997). The CLTCL gene (also named CLTD by Sirotkin et al., 1996; CLH22 by Kedra et al., 1996) contains 33 exons spanning 116 kb of genomic DNA (Holmes et al., 1997) and is oriented from telomere to centromere in chromosome 22. The CLTCL cDNA is 5506 bp with a putative translation product of 1640 amino acid residues. The two exons in this contig are 111 bp and 208 bp in length respectively, separated by a 15,768 bp intron. The remaining 3' exons are located in clones centromeric to this contig. A putative translation start site (AUG) is located at contig position 32713 and is contained within a consensus Kozak signal (GCCATGG) (Kozak, 1984). These two exons encode 82 of the 1640 amino acid residues predicted for the CLTCL gene.

The function of CLTCL gene is unknown but it shares high similarity with the human clathrin heavy chain gene (CLTC), which has been mapped to human chromosome 17, and encodes for 1765 amino acids. Both the CLTCL and the chromosome 17 encoded CLTC have homologies of 74.3% at the nucleic acid level and of 85% at amino acid level

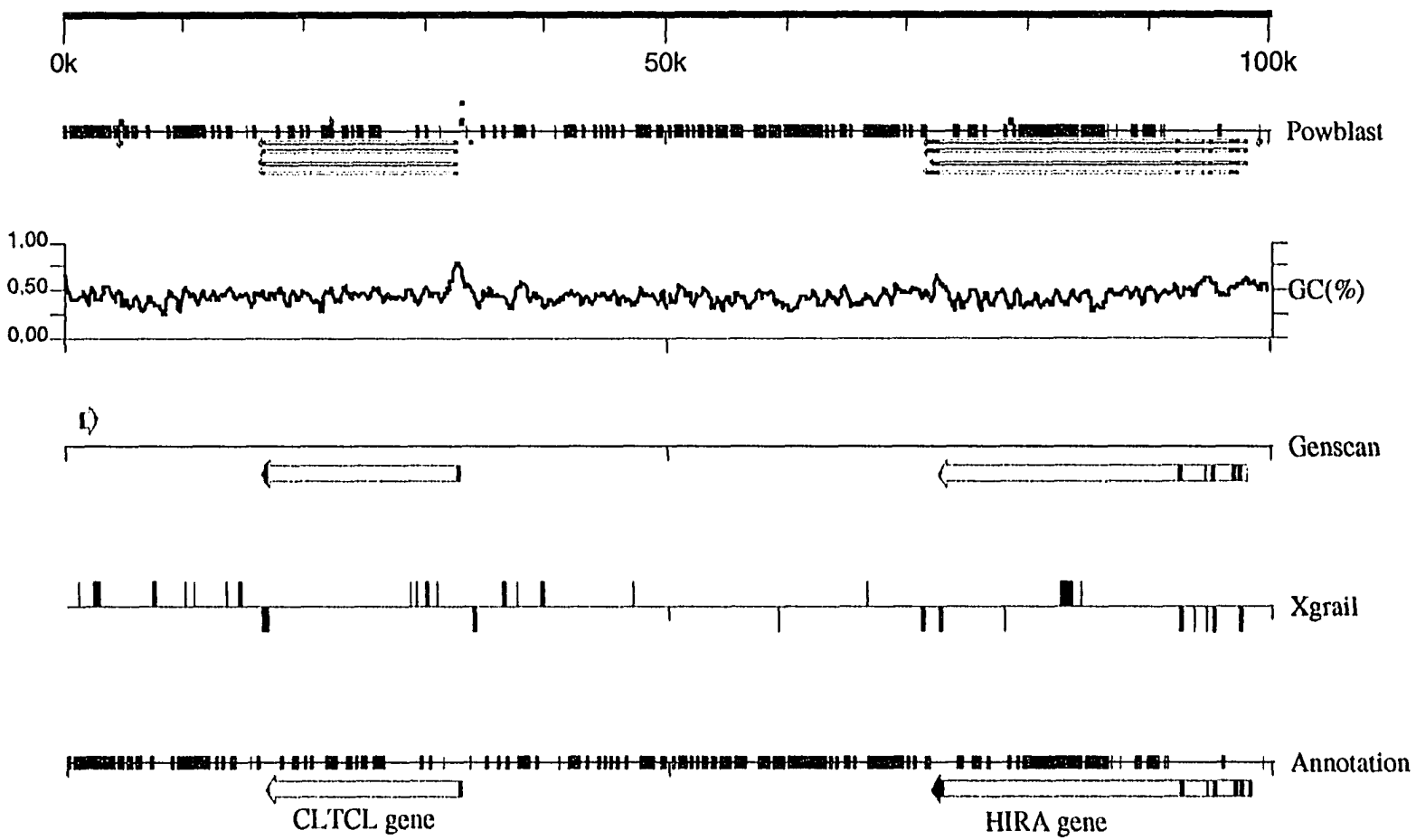


Figure 3.2a Features and annotation of contig 1 (see Fig.3.2b for legend)

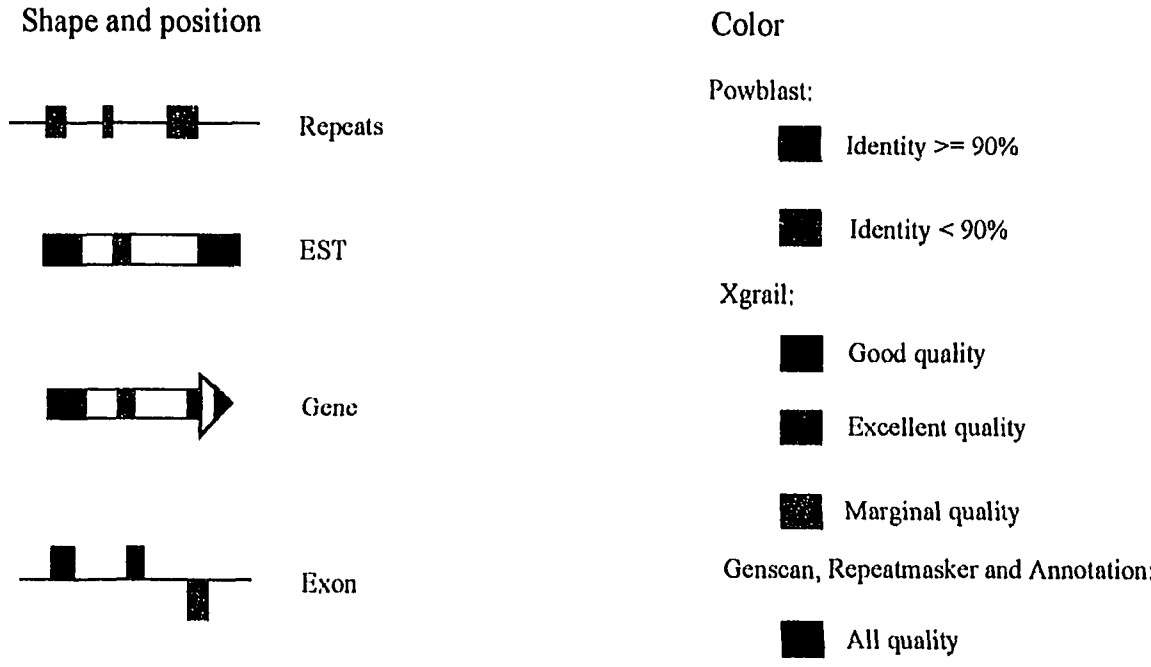


Figure 3.2b Legend for feature and annotation figures.

(Kedra et al., 1996). Clathrin is known to be involved in cellular endocytosis as it promotes the formation of and serves as a major protein component of clathrin-coated pits and vesicles (Robinson, 1994). The C-terminus of bovine CLTC forms a coiled-coil structure, which is required for its binding to the clathrin light chain (Pley and Parham, 1993). Protein secondary structure predication suggests that CLTCL also has a similar structure in the same position of the CLTC (Sirotkin et al., 1996). These results suggest that CLTCL and the clathrin heavy chain may have similar functions. Nevertheless, differences exist between the two genes. The expression of clathrin heavy chain is ubiquitous and occurs at a high level in all tissues tested while the CLTCL gene is highly expressed in skeletal muscle but expressed at a much lower level in the other tissues tested (Holmes et al., 1996). In addition, Northern blot analysis has revealed that there exist alternatively spliced forms of CLTCL but no alternative splicing has been reported in clathrin heavy chain genes (Kedra et al., 1996).

Even though the function of CLTCL is unknown, it has been implicated in the etiology of DiGeorge Syndrome (DGS) as outlined in chapter I. The CLTCL gene is interrupted by a balanced translocation in the patient TOH (Budarf et al., 1995). This break point is within intron 28 in a region proximal to contig 1. This translocation results in a truncated protein with 1496 amino acids, which is 142 amino acids shorter than the normal form of CLTCL protein. Since CLTCL is the only known gene affected by the translocation event in TOH, it has been suggested that CLTCL is one of the several candidate genes that produce DGS (Budarf et al., 1995).

Regulatory elements: The first exon of the CLTCL gene is located in a CpG island (32213-33155 bp) as revealed both by Xgrail and manual inspection of the sequence. The central 500 bp region surrounding the first exon has a GC content of 75% and this region lacks any noticeable TATA-like promoter. Within the CpG island, the promoter prediction program TSSW (see Section 2.9) predicted a promoter at 32846, which was 65 bp upstream from the first exon. Within the same region, the NNPP predicted a promoter

(33160-33210) which was about 400 bps upstream from the first exon and close to the upstream boundary of the CpG island. Since the 5' end of CLTCL cDNA has been determined by a 5' RACE experiment specifically designed to obtain the 5' end sequence of the gene (Holmes et al., 1997), the TSSW predicted promoter at contig position 32846 is more likely to be the favored promoter region for the CLTCL gene. Within the 300 bp region around this predicted promoter (32846-33146), ConsInspector predicted multiple transcription factor binding sites, including seven SP1 binding sites, one ETS1 binding site, one NF-Y binding site and one SRF binding site. Six of the seven SP1 binding sites were clustered in a 70 bp region (32876-32945).

3.2.2 Human HIRA gene

The human HIRA gene, originally named TUPLE1, was one of the initial genes detected in the DiGeorge Syndrome Critical Region of human chromosome 22 (Halford et al., 1993; Lamour et al., 1995). This gene contains 25 exons and spans a genomic region of approximately 100 kb (Lorain et al., 1996; Chen, 1997). The HIRA cDNA (X89887) is 4018 bp long and codes for a putative protein with 1017 amino acids (Lamour et al., 1995). Alignment of the HIRA cDNA sequence with the contig 1 sequence shows that seven exons from the 3' end of the HIRA gene reside in this contig as shown in Figure 3.2. The sizes of the seven exons and six introns of the human HIRA gene, listed in Table 3.2, encode the C-terminal 273 residues of the HIRA protein.

Table 3.2 Exon and intron sizes of HIRA gene in contig 1

no.	Exon size (bp)	Intron size (bp)
19	162	601
20	59	364
21	106	1641
22	123	476
23	164	1909
24	89	19802
25	856	

It has been shown that the HIRA protein sequence shares significant similarity with two yeast proteins, HIR1 and HIR2, both of which are repressors of core histone gene transcription (Lamour et al., 1995). Both the N-terminal portion of the HIRA protein and that of yeast HIR1 contain seven homologous WD repeat motifs, which are present in a wide range of different proteins and are believed to be involved in protein-protein interaction (Sondek et al., 1996). Due to these similarities, the HIRA protein has been implicated in the transcriptional regulation of histone synthesis (Lamour et al., 1995). Recently, it was reported that the HIRA protein also might interact with histone H2B (Roberts et al., 1997). Further database searches show that the human HIRA protein also has significant similarity to CAF-I (chromatin assembly factor I: Q13112) (identity 27% and similarity 45%), a protein known to assemble histone H3 and acetylated H4 onto newly synthesized DNA (Kaufman et al., 1995). This observed similarity thus suggests that the HIRA protein also could be involved in effecting chromatin structure. These database searches also revealed that homologues of human HIRA gene recently had been identified from several other species, including mouse (Q61666), chicken (P79987), fruit fly (O17468) and *Fugu rubripes* (O42611). The wide conservation of this gene suggests that it plays important roles in cellular processes. Although the exact function of HIRA gene currently is unknown, it also has been suggested to be a candidate gene responsible for the DiGeorge syndrome as described in chapter I.

Regulatory element: The promoter region for the HIRA gene was not located in this contig and thus not examined. The 3' end exons of HIRA gene which were in this contig encodes a 740 bp UTR that is AT-rich (locally as high as 75% calculated with a window size of 50 bp) and has a single putative polyadenylation signal (AATAAA) 21 bp upstream from the 3' end of the HIRA cDNA which is located at position 71812.

3.2.3 Repetitive elements in contig 1

In the 100,067 bp nucleic acid sequence of this contig, exons account for 1,878 bp, or only 1.8% of the contig. Although it has been estimated that about 35% of the human

genome consists of interspersed repetitive sequences (Smit, 1996), analysis of the repeat sequences in contig 1 shows that they make up 61% of this contig (Table 3.3), a figure which is almost twice the estimated average for the human genome. As expected, Alu and L1 sequences comprise most of the repeat sequences, accounting for 28.94% and 22.44% of the contig, respectively. Inspection of the distribution of repeats within contig 1 (Figure 3.3) shows that they generally have a lower abundance in regions surrounding exons or in short intron regions and are abundant in large intron regions as well as intergenic regions. Alu repeats are distributed randomly through contig 1, with no observable preference for their orientation. However, L1 repeats tend to be distributed as clusters. The orientations of L1 repeats also seem to show a preferred orientation in the direction opposite to that of gene transcription when located within introns of a gene. For example, within the intronic region of the CLTCL gene, there are 11 segments of L1 repeats, all of which are oriented opposite to the direction of the transcription of the CLTCL gene. Within the intron region of HIRA gene, there are 15 segments of L1 repeats, 11 of which are in the opposite direction to that of HIRA gene transcription. In comparison, 29 segments of L1 repeats are present in the intergenic region between the CLTCL and HIRA gene, with an almost equal number of segments (15 and 14) in opposite orientation.

Table 3.3 Content of repetitive sequences in contig 1

		number of elements	length occupied	percentage of sequence
SINEs	ALUs	108	28961 bp	28.94 %
	MIRs	3	533 bp	0.53 %
LINEs	L1	23	22455 bp	22.44 %
	L2	5	967 bp	0.97 %
MER		6	3738 bp	3.74 %
LTR elements		7	2702 bp	2.70 %
Unclassified		2	1463 bp	1.46 %
Simple repeats		3	273 bp	0.3 %
Total repeats		157	61092 bp	61.08 %

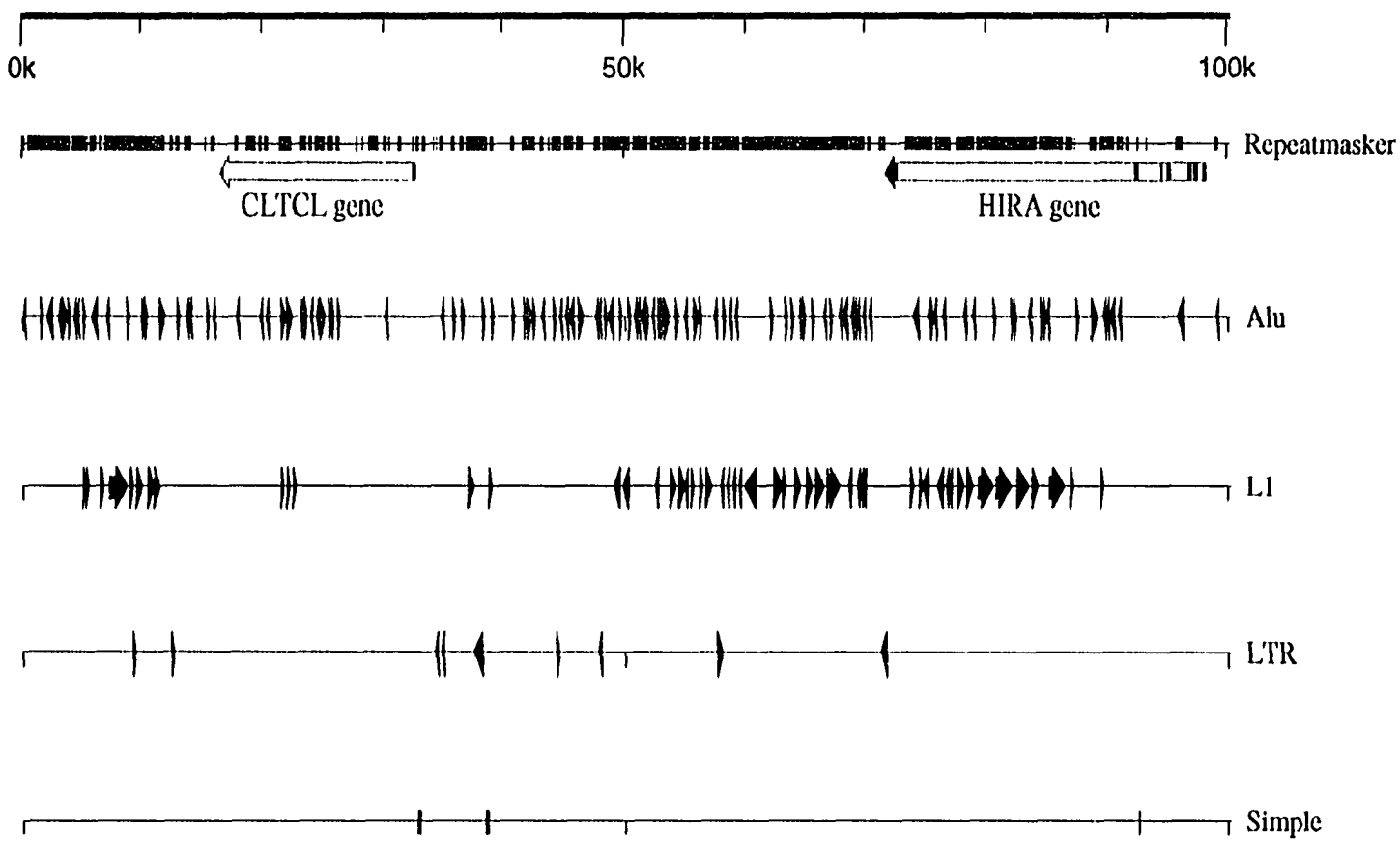


Figure 3.3 Repetitive elements in contig 1

3.3 Analysis of contig 2

Contig 2 consists of cosmids 83e8 and 59f, both of which are located in the DiGeorge Syndrome Critical Region. These two cosmids overlap by 17,274 bp and have a total combined unique sequence length of 65,420 bp. A Powblast database search revealed that two genes, Porc-P1 and TMVCF, were located in this region (Figure 3.4). Since seven of the total 19 Porc-P1 exons lie in a region of 20 kb telomeric to the contig in cosmid 129f8, which has been sequenced by Feng Chen at the Advanced Center for Genome Technology (ACGT), Department of Chemistry and Biochemistry, the University of Oklahoma, analysis of the entire gene region is included. Thus, the total analyzed sequence for contig 2 is 85420 bp.

3.3.1 Human Porc-P1 gene:

The Porc-P1 cDNA recently was cloned and sequenced by a group in England (AJ223728). The cDNA is 1871 bp long and has a putative translation product of 566 amino acids. A dotter comparison of the cDNA sequence with the sequence of contig 2 shows that there are a total of 19 exons spanning a region of about 40 kb. The exon and intron sizes of the Porc-P1 gene are listed in Table 3.4. Alignment of the Porc-P1 cDNA with the genomic sequence revealed one discrepancy at contig position 32111 where a G was substituted for an A in the cDNA sequence. This mismatch does not change the coded amino acid for the putative Porc-Pl protein.

Table 3.4 Exon and intron sizes of Porc-P1 gene

no.	Exon size(bp)	Intron size (bp)	no.	Exon size(bp)	Intron size (bp)
1	81	138	11	132	248
2	60	735	12	99	664
3	93	1644	13	162	6057
4	138	1034	14	139	77
5	144	10318	15	84	1478
6	56	1598	16	119	171
7	49	1356	17	77	1950
8	62	1653	18	66	1571
9	51	6240	19	120	
10	120	1904			

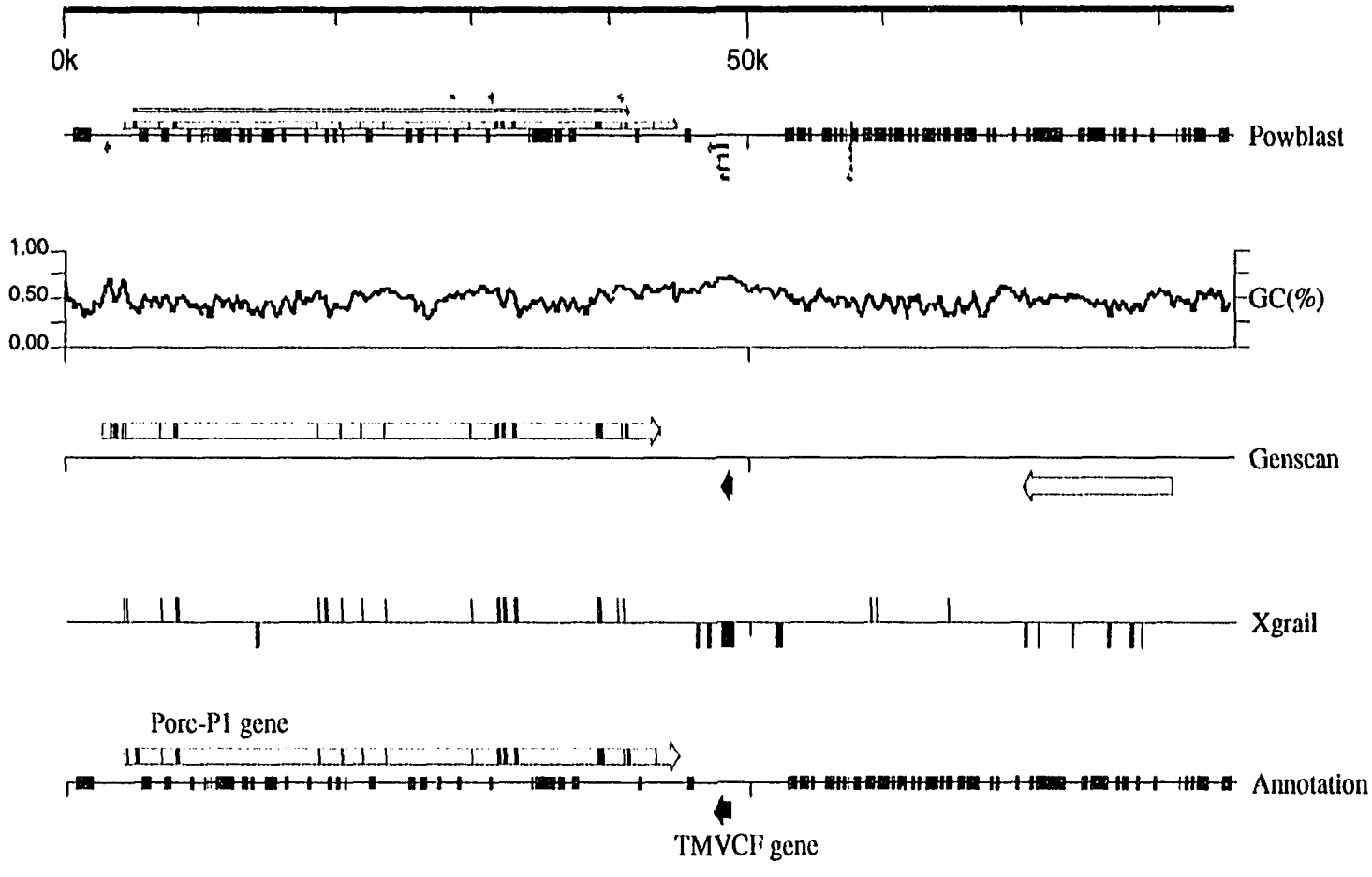


Figure 3.4 Features and annotation of contig 2

The function of Porc-P1 gene currently is unknown. A protein database similarity search shows that Porc-P1 is similar to the *Saccharomyces cerevisiae* protein CDC45(Q08032), with an identity of 32% and a similarity of 45%. An alignment of the Porc-Pl gene and CDC45 gene protein sequence is shown in Figure 3.5. The similarity between the two proteins also is seen from their Kyte & Doolittle hydropath plots (Figure 3.6). Protein CDC45 (651 amino acids in length) is a nuclear protein involved in the initiation of DNA replication (Zou L. and Stillman B., 1998). The Porc-Pl protein also shows some similarity to segments of a Ustilago mayids gene TSD2 (U50276), with an identity of 24-30% and a similarity of 41-49%. The TSD2 gene product also has been implicated in DNA synthesis (U50276).

A search of the profile database indicated that the predicted protein would have a bipartite nuclear localization signal at residues 156-173. Psort predicted that the protein would be localized in the nucleus and a prosite database search revealed one potential aminoacyl-tRNA synthetase class-II signature I at residues 7-27. All class-II aminoacyl-tRNA synthetases share protein sequence similarities and contain at least three conserved regions (Cusack et al., 1991). The signature I region was thought to contain an active site of those RNA synthetases but the absence of the other signatures and the lack of overall protein sequence similarity between Porc-P1 and class-II aminoacyl-tRNA synthetases suggest that Porc-pl is not a class-II aminoacyl-tRNA synthetase. However, the presence of the potential signature is consistent with the possible implication of the protein in nucleic acid synthesis. Overall, these results suggest that Porc-Pl might be a nuclear protein, possibly involved in DNA synthesis.

Inspection of the exon organization of Porc-Pl gene shows that the 19 exons are clustered into four groups. Exons 1-5 reside in a region of about four kilobases, Exons 6-9 about 5 kb, exons 10-13 about 3.5 kb, and finally exons 14-19 about 6 kb. These exons are relatively small, varying between 49 - 162 bp. The three introns between the four groups are about 10 kb, 6 kb and 6 kb, respectively.

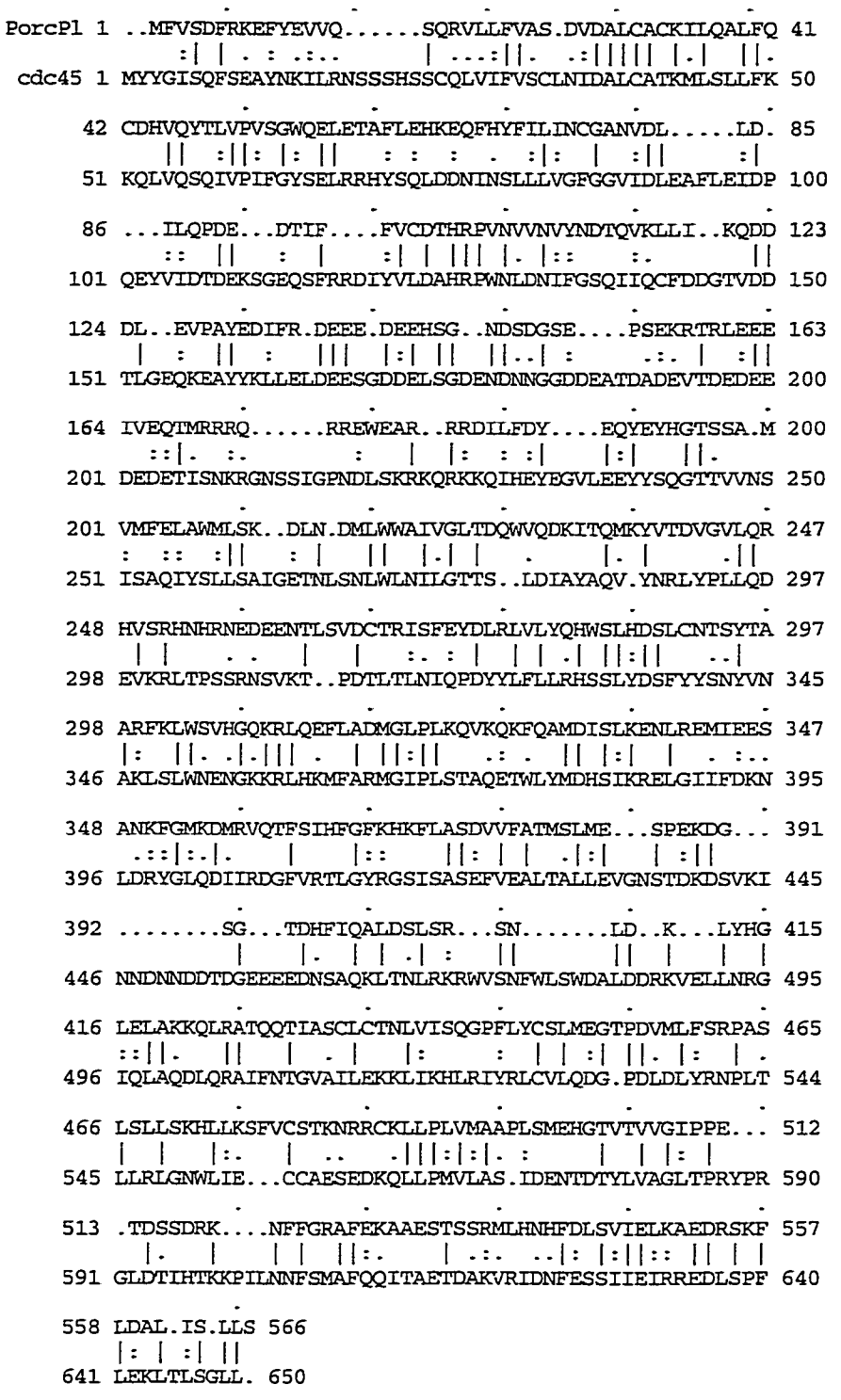


Figure 3.5 Bestfit alignment of Porc-p1 and CDC45 protein
Note: | indicates identity :. indicate similarity

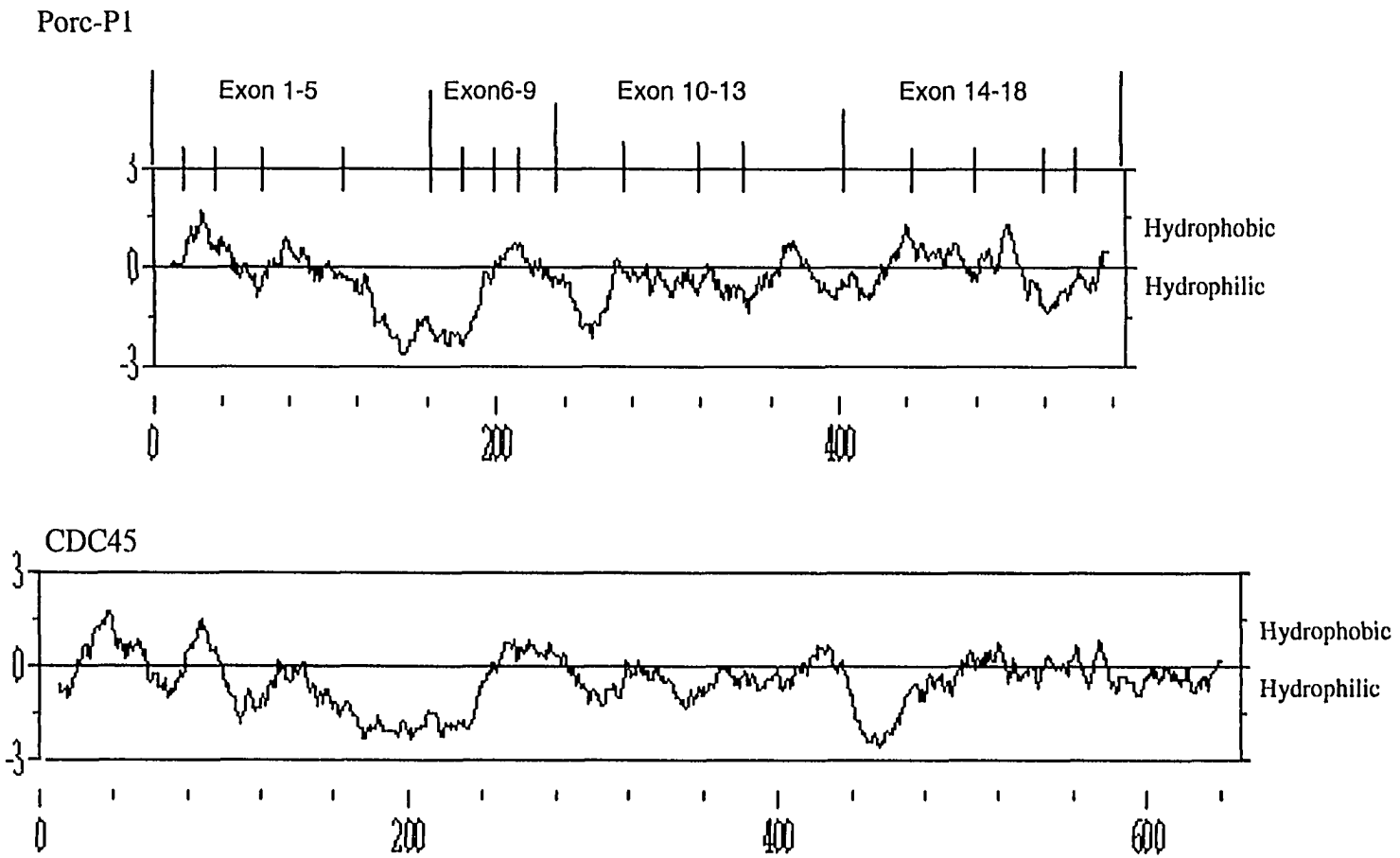


Figure 3.6 Comparison of Kyte-Doolittle hydrophobicity plot between Porc-p1 protein and CDC45 protein

Upper panel: Porc-Pl protein, approximate exon boundaries also are shown.
Down Panel: CDC45 protein

Regulatory Elements: The first two exons of the Porc-P1 gene reside in a CpG island (4132-4572) which is recognized by Xgrail. However, no TATA box is found within one kb upstream of the first exon and Xgrail does not predict a promoter in the region. Promoter prediction by TSSW revealed a promoter at contig position 4124, which was 92 bp upstream from the 5' end of the first exon of the gene. Another program, NNPP, predicted two promoters preceding the first exon. The first predicted promoter was at position 3580-3630, 586 bp upstream from the first exon. The second predicted promoter was at position 3947-3997, 219 bp upstream from the first exon. The 300 bp region (3824-4124) thus most likely contained the promoter as predicted by both programs and therefore was used in a subsequent transcription factor binding site search. Within this putative promoter region, ConsInspector predicted a number of transcription factor binding sites as shown in Figure 3.7. Among those factors, Octamer, SP1 (GC box binding) and NF-Y (CCAAT box binding) are ubiquitous and found in promoters of many genes (Lewin, 1997). Both ETS1 and GABP belong to the ETS family of transcription factors which have been identified in the transcriptional regulatory regions of multiple T cell genes (Wang et al., 1992). The serum response factor (SRF) is thought to be essential for serum and growth factor activation (Johansen and Prywes, 1993). How these factors are related to the putative involvement of the Porc-P1 gene in DNA synthesis is unclear.

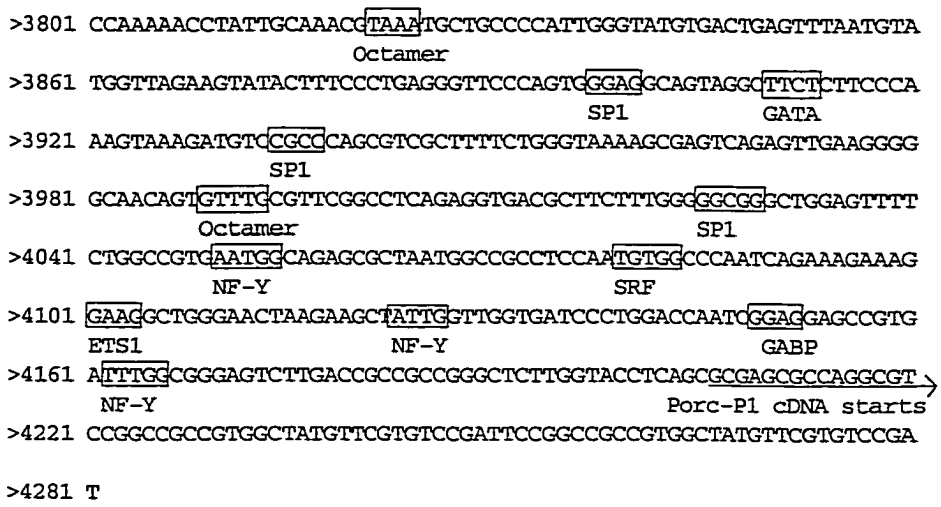


Figure 3.7 Distribution of predicted transcription factor binding sites around the putative promoter region of Porc-P1 gene

3.3.2 Human TMVCF: a putative transmembrane protein

This contig also contains a gene with unknown function which has been called TMVCF (TransMembrane protein deleted in Velo-Cardio-Facial syndrome) (Sirotkin et al., 1997). Comparison of genomic sequence with that of the TMVCF cDNA (AF000959, 1348 bp) indicates that the gene contains only one exon that spans contig positions 47292-48636 and that the direction of transcription is from telomere to centromere. One mismatch between the gene and the cDNA sequence was found in the coding region which does not change the amino acid composition of the putative TMVCF protein. A Kozak signal sequence is located at contig position 48519 and an in-frame stop codon is present at contig position 47865. Therefore, the predicted TMVCF protein has 218 amino acid residues (Figure 3.8) and the 5' and 3' untranslated regions are 118 and 570 bps long, respectively.

```
>TMVCF protein; Length 218 aa
MGSAALEILGLVLCLVGWGGLILACGLPMWQVTAFLDHNIVTAQTTWKGL
WMSCVVQSTGHMQCKVYDSVLALSTEVQAARALTVSAVLLAFVALFVTLA
GAQCTTCVAPGPAKARVALTGGVLYLFCGLLALVPLCWFANIVVREFYDP
SVPVSQKYELGAALYIGWAATALLMVGGCLLCCGAWVCTGRPDLSFPVKY
SAPRRPTATGDYDKKNYV
```

Figure 3.8 Putative protein sequence of human TMVCF gene

Nucleic acid and protein database searches using either the gene sequence or the putative translated protein sequence revealed multiple significant matches, with the most significant including a mouse brain endothelial cell 1 (MBEC1) gene (AF035814) (identity 85%), and both the human and mouse Clostridium perfringens enterotoxin receptors (hCPE-receptor: AB000713 and mCPE-receptor: AB000712, respectively). Although the function of MBEC1 is unknown, it has been reported to be the mouse homologue of TMVCF based on sequence homology (Chen et al., 1998). As receptors for clostridium perfringens enterotoxin (Katahira et al., 1997a), the hCPE and mCPE receptors have been predicted to contain four putative transmembrane segments and are thought to be required for both target cell recognition and pore formation in cell membrane (Katahira et al., 1997b).

Based on hydropathic analysis and transmembrane domain prediction, Sirotkin et al. (1997) proposed that TMVCF had two transmembrane helices (residues 69-100 and 156-176), with both the N-terminal and C-terminal located in the cytoplasm. In their prediction, the Kyte-Doolittle algorithm predicted four hydrophobic regions but the PHDhtm program (Rost et al., 1995) predicted only two transmembrane helices. To investigate these conflicting results, the TMVCF protein sequence was analyzed by the most recent version of the PHDhtm program for protein topology prediction. The new results indicated four transmembrane helices. The program TMpred also predicted four transmembrane helices, with a total score of 9704 and individual scores of more than 2000, where scores above 500 are considered to be significant (Figure 3.9, lower panel). Both programs also predicted that the N- and C-termini of the TMVCF protein would be on the cytoplasmic side of the membrane. Because the positions of the predicated transmembrane helices from both programs are very similar, as shown in Table 3.5, these present results suggest that a model with four transmembrane helices might be more favorable than a model with two transmembrane helices as proposed by Sirotkin et al. (1997).

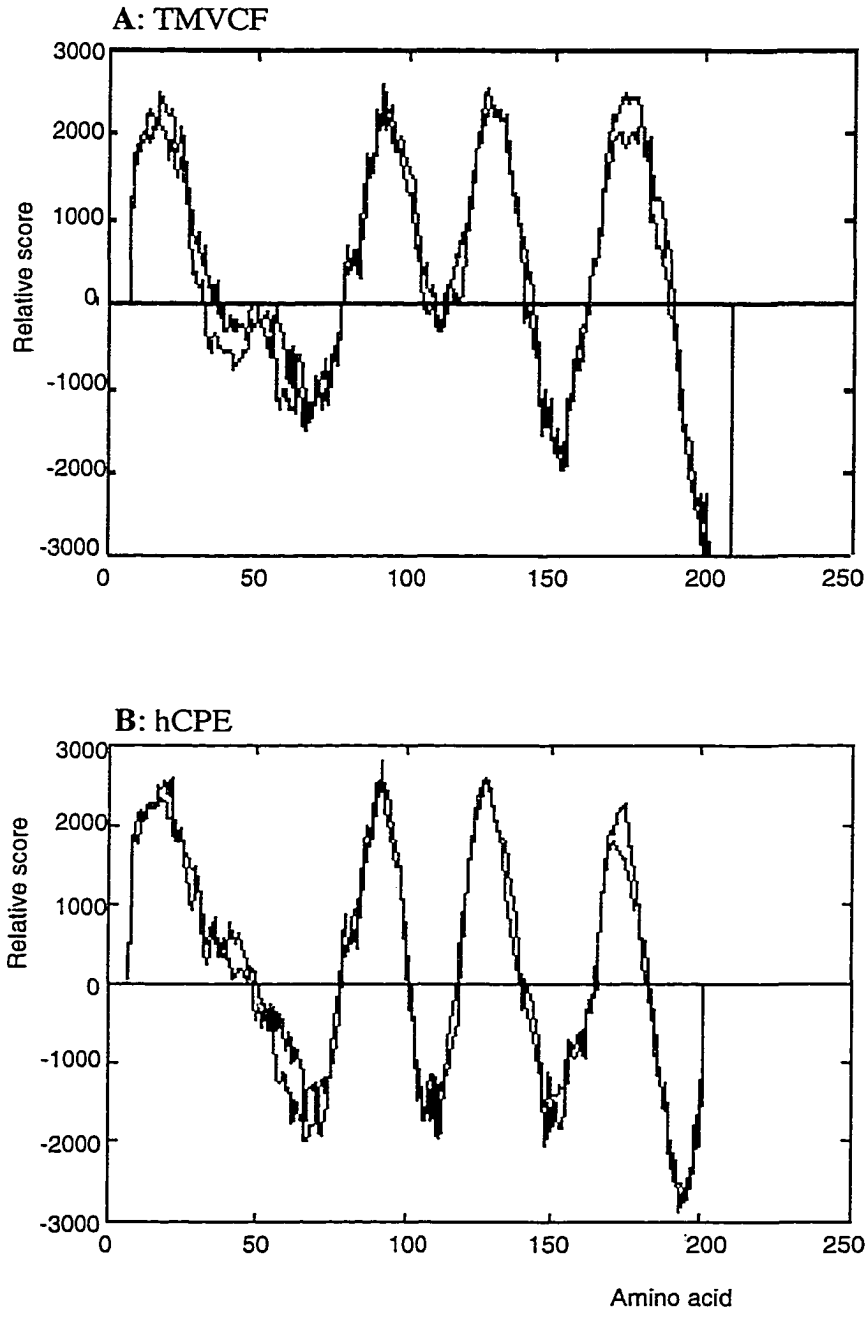


Figure 3.9 Transmembrane helices prediction of TMVCF protein by TMPred

For comparison, the TMPred output for hCPE protein is shown in panel B

The TMpred output for TMVCF protein resembles that for human CPE receptor (see Figure 3.9) which, as discussed above, most likely has four membrane-spanning segments.

Table 3.5 Predicted transmembrane domains of TMVCF protein

no. of tm helices	TMpred prediction (residue position)	PHDhtm prediction (residue positions)
1	8-27	12-44
2	82-102	79-102
3	117-139	114-145
4	166-187	162-188

A profile search with the TMVCF protein sequence did not reveal any homologous protein domains in the profile database. However, a prosite search revealed several potential phosphorylation sites, two potential membrane lipoprotein lipid attachment sites at residues 21-31 and 175-185 respectively, and six other prenyl group binding sites. These latter sites are close to or within the predicted transmembrane domains and therefore might be involved in attaching the protein to the membrane. This result is consistent with the results from Psort, which predicted that the protein was localized in the plasma membrane, and had a potential N-terminal signal sequence.

Regulatory elements: The genomic region of TMVCF gene has a high GC content (~70%). Interestingly, Xgrail indicated that the entire coding region is located in a CpG island (47570-49026). Although no TATA box was found upstream of the putative coding region, a promoter was predicted by TSSW 140 bp upstream of the 5' end of the cDNA sequence at contig position 48776. NNPP also predicted one promoter 200 bp upstream of the 5' end of the cDNA spanning contig positions 48833-48883. Both GENSCAN and Xgrail predicted the presence of a gene in this region, and both predicted that translation of the gene starts 138 bp upstream from the 5' end of the cDNA sequence at contig position 48774. These positions for the predicted promoter and coding region raise the possibility that the cDNA clone of TMVCF (Sirotkin et al., 1997) may not contain the complete

sequence of the TMVCF transcript. However, a search of the dbEST database did not reveal any ESTs which would align to any region more upstream from this cDNA sequence. ConsInspector predicted one NF-Y, two GA binding protein, one SP1 and one ETS-1 binding site in the predicted promoter region (48776-49076) and a putative polyA signal (AGTAAA) (Sirotkin et al., 1997) located at contig position 47305, 15 bp upstream from the 3' end of the TMVCF cDNA sequence.

Thus, TMVCF is a putative transmembrane protein, possibly having four transmembrane helices and localized in the plasma membrane. It also might be involved in target cell recognition and pore formation as suggested by its high sequence similarity with the CPE receptors.

3.3.3 Repetitive elements in contig 2

Contig 2 contains a total of 48.92% repeat elements (Table 3.6). Alu repeats and L1 repeats are the two most prevalent repeats with 25.76% and 16.69%, respectively. The Porc-P1 gene region contains two clusters of L1 repeats each within a relatively large intron region (Figure 3.10). There are 16 fragments of L1 repeats in the two clusters, all of which are oriented in the opposite direction of Porc-P1 transcription. There also are 9 LTR repeats in the Porc-P1 gene region, 8 of which are oriented in the opposite direction of the gene transcription.

Table 3.6 Content of repetitive sequences in contig 2

	number of elements	length occupied	percentage of sequence
SINEs:			
ALUs	82	22005 bp	25.76 %
MIRs	7	699 bp	0.82 %
LINEs:			
L1	26	14253 bp	16.69 %
L2	6	883 bp	1.03 %
LTR elements:	8	3057 bp	3.58 %
MER	3	465 bp	0.54 %
Simple repeats	6	395 bp	0.5 %
Total interspersed repeats:	138	41757 bp	48.92 %

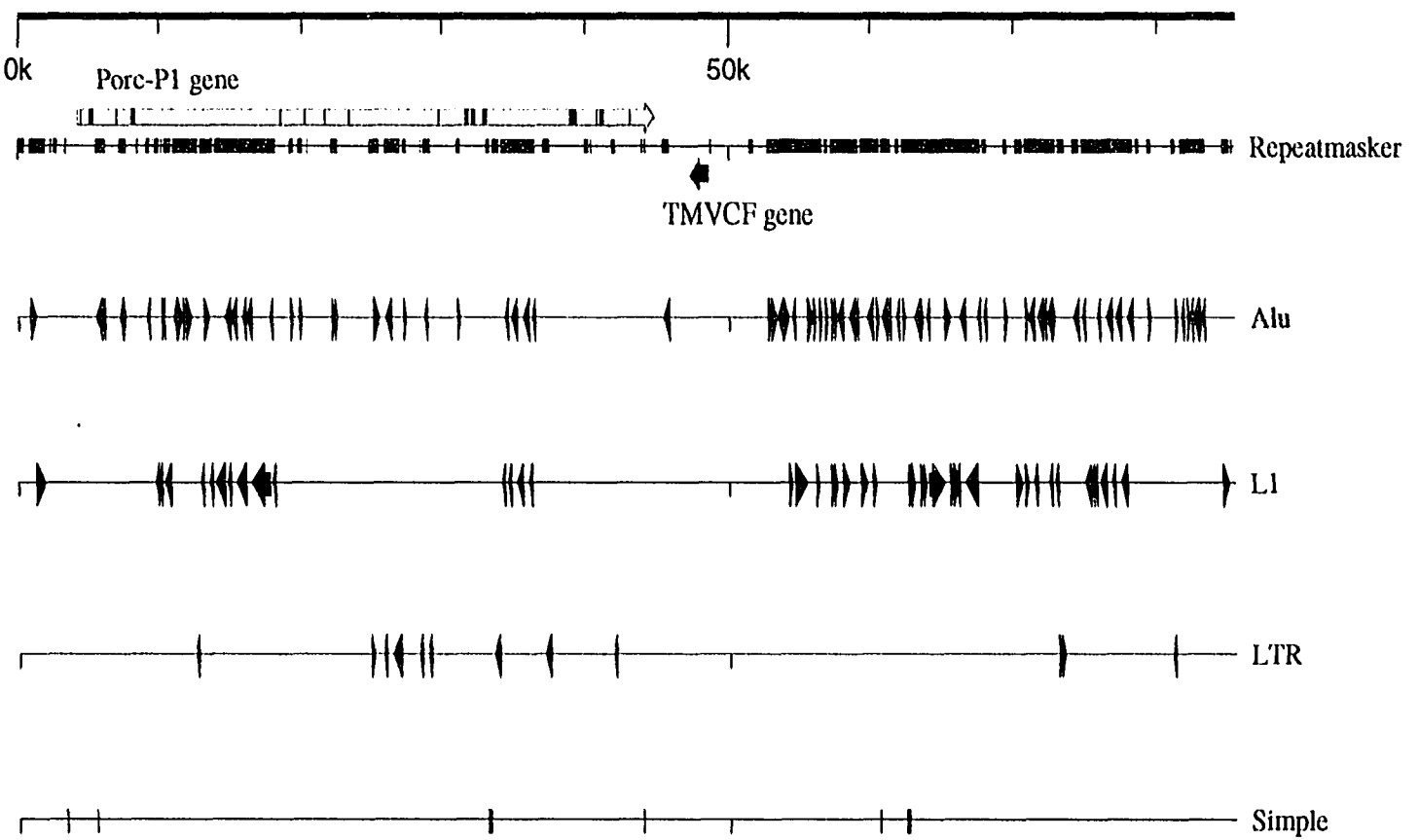


Figure 3.10 Distribution of repetitive elements in contig 2

3.4 Analysis of contig 3

Contig 3 is composed of a single PAC clone, p201m18. This contig, located in the telomeric end of the DiGeorge Syndrome Critical Region (DGCR), has a total length of 162,268 bp. A Genbank database search reveals that this contig contains three genes, the human RanBP1 gene, and genes homologous to the mouse T10 gene and the mouse Htf9c gene (Figure 3.11). Two additional putative genes also are predicted by their EST similarities and results from computational gene prediction programs.

3.4.1 Human T10 gene

A powblast database search of the Genbank non-redundant database revealed that a 15 kb region on the centromeric end of this contig has significant homology to mouse T10 cDNA (X74504). Dotter analysis between the contig sequence with mouse T10 cDNA shows that five exons (exons 4-8) are well conserved between human and mouse, but the 3' terminal exon (mouse exon 9) is not (Figure 3.12). However, two conserved segments in exon 9 occur in both the mouse exon 9 and human genomic sequence. The first conserved segment includes the 5' ~160 bp, which contains the coding region of the mouse exon 9. The other conserved segment is about 50 bp from the 3' end of exon 9, where a putative polyadenylation signal is located. When the sequence of mouse and human exon 9 are compared, an insertion of about 600 bp is seen in the human gene between contig positions 13720-14450. This inserted sequence contains a ~200 bp (13750-13950) CA-rich simple repeat and a ~280 bp Alu repeat (14160-14440). Interestingly, the rat T10 gene has the region corresponding to the mouse exon 9 split to two exons (9 and 10) and one intron (intron 9) (see Section 3.5). Although this also may be possible for the human T10 gene, it is likely that, as in the mouse, the human T10 gene has a single 3' terminal exon 9 because (a) Although there is no single human EST hit spanning the putative human exon 9, a number of overlapping ESTs cover the entire region as shown in Figure 3.11; (b) The splicing site consensus sequence at both ends of rat intron 9 (as revealed by comparing

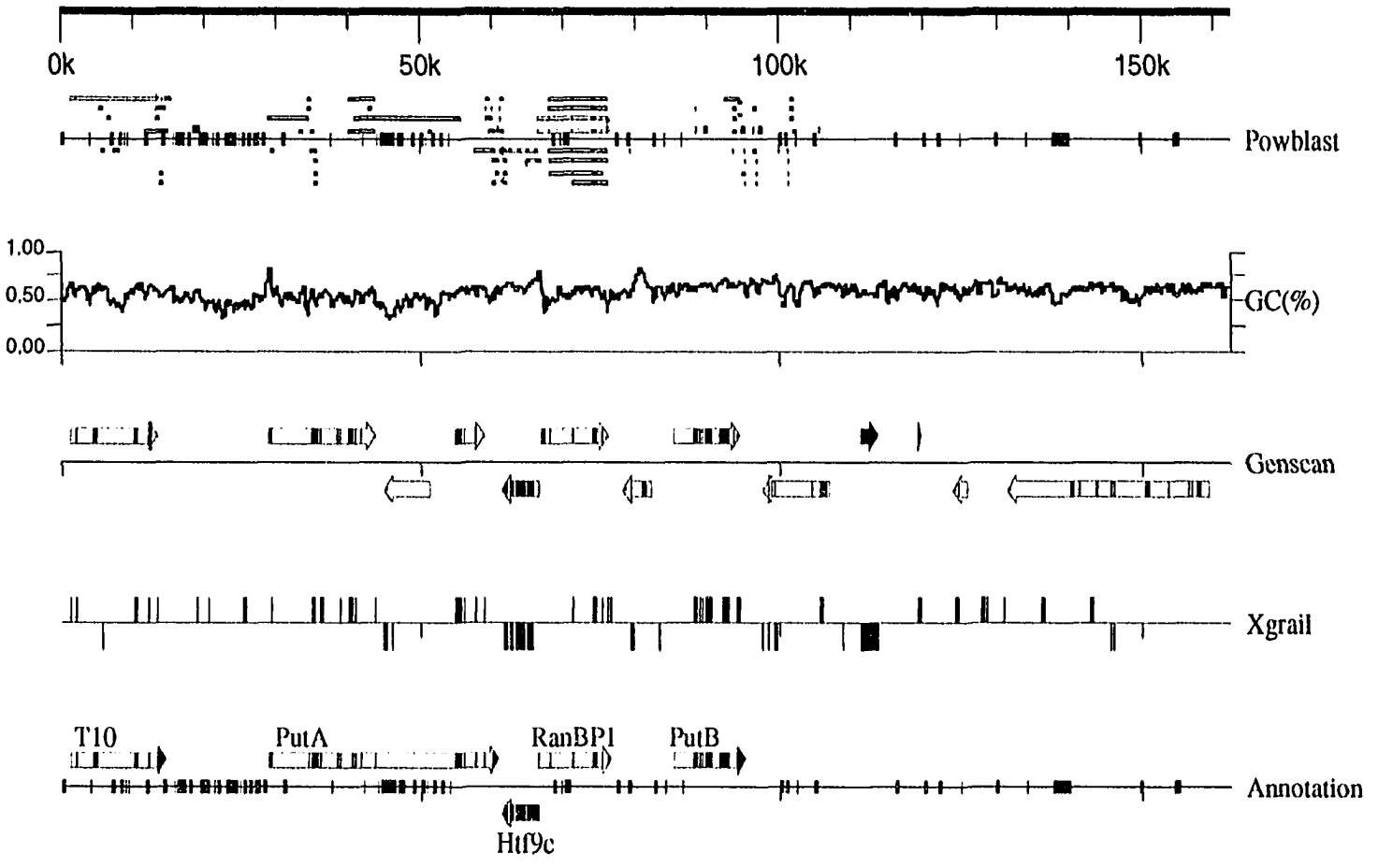


Figure 3.11 Features and annotation of contig 3

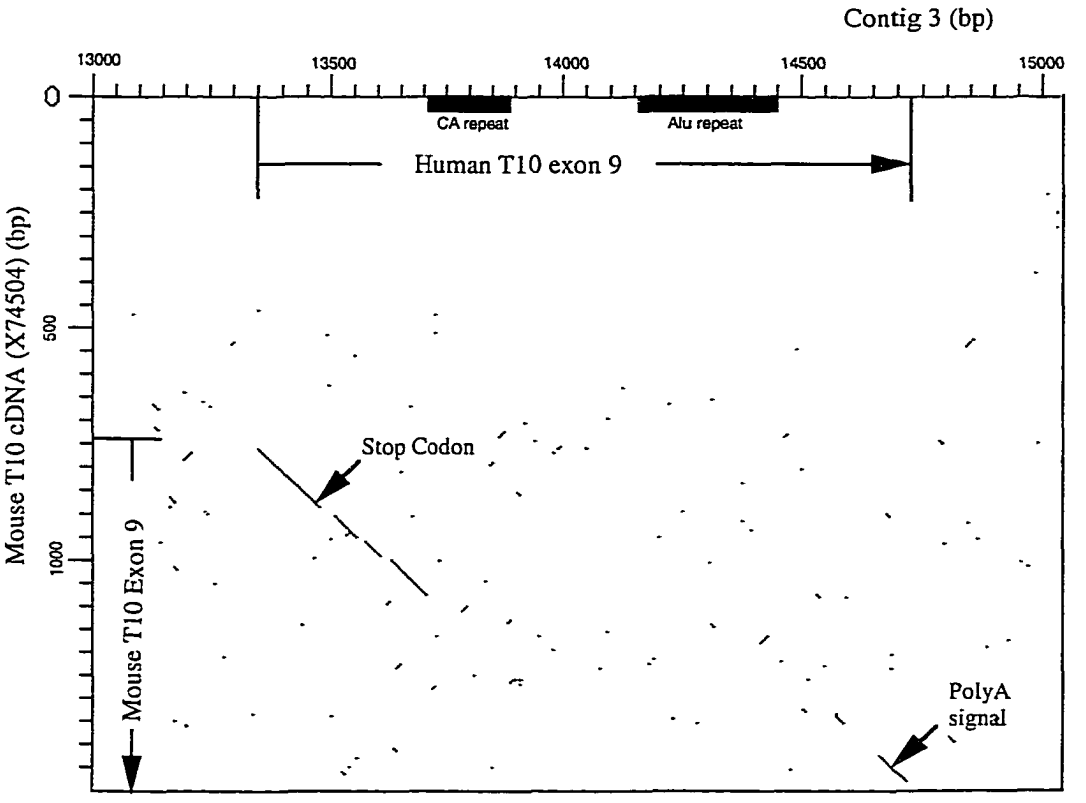


Figure 3.12 Dotter plot of contig 3 sequence (13000-15000) with mouse T10 cDNA

rat EST AI010662 with mouse genomic sequence) is not conserved in human; and (c) Previous Northern blot analysis with human fetal tissues indicated that the human T10 transcript was about 2 kb in length, 500 bp longer than the length of the mouse T10 transcript (Halford, 1993). It was proposed earlier that the human T10 gene might have a longer 5' UTR or the gene was alternately spiced at later stages of development (Halford, 1993). A more recent study (Chen, 1997) indicated that the human T10 gene indeed had a longer 5' UTR because of the presence of an additional 120 bp 5' end exon. However, the present study showed that the mouse T10 gene had an additional 120 bp 5' end exon as well (see section 4.1) which was not present in the sequenced mouse T10 cDNA. Both the human and mouse 5' end exons are non-coding, reside in major CpG islands, and thus most likely represent the 5' end of these T10 genes. Thus, it now is clear that the ~500 bp difference between human and mouse T10 transcripts is not due to differences in the 5' UTRs of the two genes although alternative splicing can not be excluded as the source of the length difference between the human and mouse T10 transcripts. However, examination of the human genomic sequence strongly suggests that the longer human T10 transcript results from the insertion of repetitive elements in the 3' untranslated region of the T10 gene. The size of the six exons and five introns of the putative human T10 gene in this contig are listed in Table 3.7.

Table 3.7 Exon and intron sizes of human T10 gene

No.	Exon Range	Exon size(bp)	Intron Size (bp)
(exons 1-3 are not located in this contig)			
4	1272-1391	120	852
5	2244-2358	115	2390
6	4749-4819	71	5516
7	10336-10489	154	1654
8	12144-12248	105	1099
9	13348-14722	1375	

Both the putative human and mouse T10 proteins are 276 amino acid residues long and have 84% identity and 87% similarity (Chen, 1997). The six exons in this contig

encode the carboxy-terminal 227 amino acid residues of the human T10 protein. Even though the mouse T10 gene was isolated as a serine/threonine-rich protein of unknown function(Halford et al., 1993), additional protein and nucleotide database searches failed to reveal any homologies to any known genes and a profile search also did not reveal any similar protein domains in the profile database. Thus, the function of T10 gene remains to be discovered.

3.4.2 Human Htf9 locus: human Htf9c and RanBP1 gene

In mouse, the Htf9 locus contains two genes, Htf9c and RanBP1. Both genes were discovered because of their proximity to a CpG-rich sequence (Lavia, 1987). The two genes are transcribed in opposite orientations and have coincident transcription start sites (Lavia., 1987).

The presence of the human Htf9c gene in this contig is revealed by the extensive sequence similarity between the contig sequence with that of the mouse Htf9c gene. Comparing the human genomic sequence with the mouse Htf9c cDNA sequence (see Section 3.4) indicates that both the human and mouse Htf9c genes likely contain 12 exons (Figure 3.13). Two human EST entries (R72799 and T88997) aligned perfectly to the genomic sequence and coincide with three of the 3' end exons of the predicted human Htf9c gene. It is observed that while the 5' half of the putative 5' end exon of human Htf9c gene is well conserved in mouse, the 3' half has much lower homology to that of mouse. However, alignment of the human and mouse genomic sequences in this region shows that the sequence at the 3' end of the first exon is moderately conserved, allowing assignment of the putative 3' end of this exon. In addition, the 3' end of exon 12 also is not well conserved and the predicted 3' end of this exon is based on the human ESTs mentioned above. The predicted cDNA sequence for the human Htf9c gene is 2766 bp, and the size of the exons and introns are listed in Table 3.8. An ORF encoding 624 amino acid residues occurs in the predicted human Htf9c cDNA sequence (Figure 3. 14). The sequence

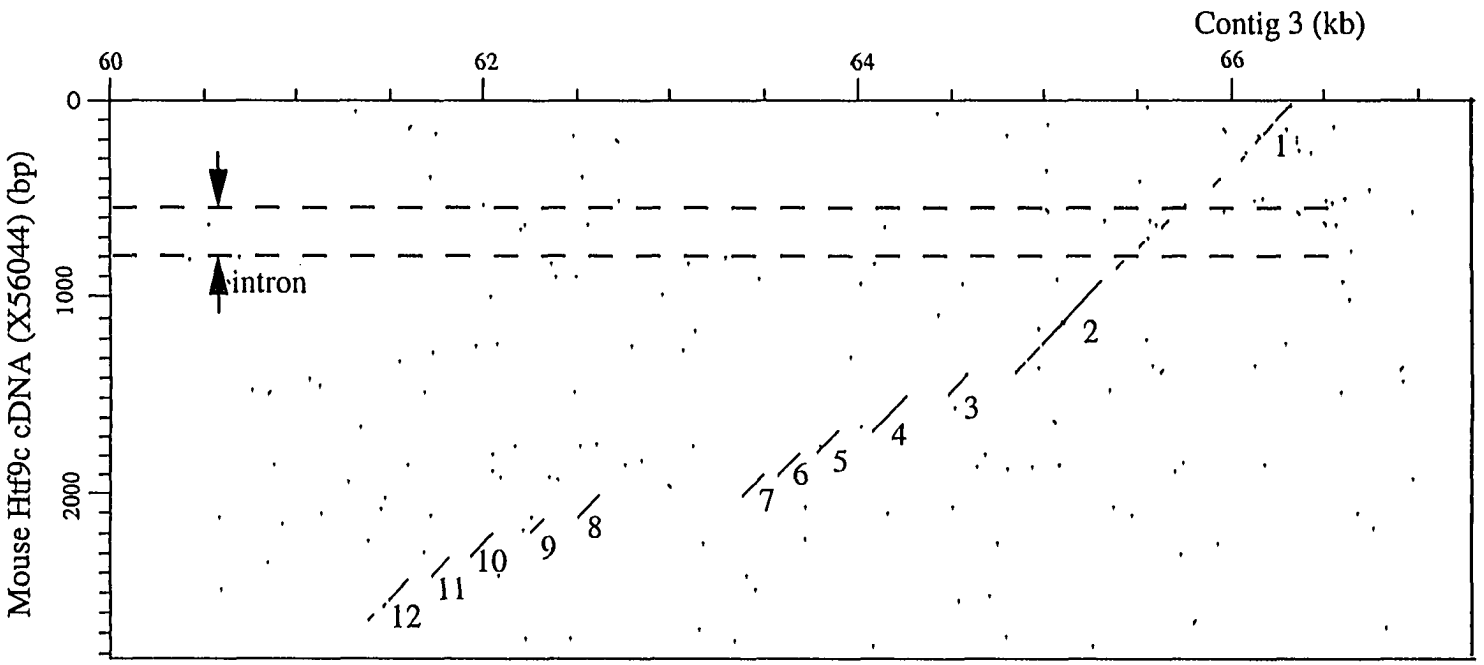


Figure 3.13 Dottor plot of contig 3 sequence with mouse Htf9c cDNA sequence

Predicted exons are numbered. The mouse Htf9c cDNA may contain an intron (575-809bp) as indicated by compariison with mouse ESTs AI037312 and AI037089.

>human Htf9c cDNA 2766bp

CTTCCCTCCGCCCCACGACCCGAACCTCGACCCCTGACCCCAGCGGCGGGAAACGCGCCG
CGATTCAAAACCAGTACAGCTTTATTGGCTCAGGAGGCGGACACTCTCATTGTCCGGCCC
GCCCCGCGCTTCCCGCCAAAACGCCCGCGGCGCATGCCCGGCCCGCCCGTCAGCCAGAGG
AGCGCGCACTCCCGGCGACACCCGCGCGCACTTCCGTCCCGGCCACGCGCGCCTCTAGAG
CTGGCTAAAGTGGTTGGGGGTTGGATCCGGCGCGTGACGATGACGTTGCCAGAAGCCAGG
ATCATAGAGACGCTCGGCCTCCGCCGCACGCGCTGGCGGACTAAGAGTGGCTGGCGAAGC
GAGCGGCCGGCGCGGGCCCCTGGCGGGCGGGCGGTACAGCCCCAAGCCTGAGACCCGGAC
CTGAGCATCGCAGGTTCGAGTCCCGCCCCGCCTGGGGCGAAGCCGGGGGTGGCGGCGACC
TCGCGGCGTTGCACCGGCTCTGTGAGCACCTCCCCTCTGAGCACTTCCCTTGTGACAGGC
CACTTCCCTTGTGACAGGCCCAGGACGAGGTGGCCAGGCGGCCCCCATGGCGTCCCTGGT
CTAGGCGGAGAACCGCCTGGGCGATGAGTGAGAACCTCGACAACGAGGGGCCCGAAGCCC
ATGGAGAGCTGTGGCCAGGAGAGCAGCAGTGCCCTGAGCTGCCCTACCGTCTCGGTGCCC
CCTGCAGCCCCGGCAGCCCTGGAGGAGGTGGAGAAAGAGGGCGCTGGGGCGGCTACAGGG
CCGGGGCCTCAGCCCGGGCTCTACAGCTACATCAGGGATGACTTGTTTACCTCTGAGATC
TTTAAACTGGAGCTGCAGAACGTGCCTCGCCACGCCAGCTTCAGCGACGTCCGGCGCTTC
CTGGGCCGCTTTGGTCTGCAGCCCCACAAAACCAAACTCTTTGGGCAACCACCCTGCGCC
TTTGTGACATTCCGCAGCGCTGCAGAGAGGGACAAGGCCCTGCGCGTTTTGCATGGTGCC
CTCTGGAAAGGCCGCCCACTCAGTGTGCGCCTGGCCCGGCCCAAGGCCGACCCCATGGCC
AGGAGGAGGCGACAGGAGGGTGAGAGTGAGCCACCAGTAACACGAGTGGCCGACGTGGTG
ACCCCTCTATGGACAGTGCCCTATGCTGAGCAGCTTGAGCGGAAGCAGCTGGAGTGCGAG
CAGGTGCTGCAGAAACTTGCCAAGGAAATCGGGAGCACCAACCGTGCCTTGCTGCCCTGG
CTGCTCGAGCAGAGGCACAAGCACAACAAGGCCTGCTGCCCGCTGGAGGGGGTCAGGCCA
TCACCCCAGCAGACTGAGTATCGTAATAAGTGTGAGTTTCTGGTTGGCGTCGGGGTGGAT
GGGGAGGATAACACCGTGGGCTGTCGGCTCGGCAAGTACAAGGGCGGGACGTGTGCTGTG
GCAGCCCCGTTTGACACCGTGCACATCCCCGAAGCCACCAAGCAGGTGGTGAAGGCCTTC
CAGGAGTTCATCCGGTCCACTCCATACTCGGCATACGACCCAGAGACGTACACAGGCCAC
TGGAAGCAGCTGACTGTGCGCACCAGCCGCCGCCACCAGGCCATGGCCATTGCCTACTTC
CACCCCCAGAAGCTGAGCCCTGAGGAGCTGGCAGAGCTGAAGACCTCCCTAGCGCAGCAC
TTCACAGCAGGGCCAGGCAGGGCCAGTGGAGTGACCTGCCTCTACTTCGTGGAGGAGGGA
CAGCGAAAGACTCCTAGCCAGGAGGGCCTGCCCCTGGAGCATGTGGCTGGGGACCGGTGC
ATCCACGAGGACCTGCTAGGGCTGACCTTCCGGATCTCTCCACACGCCTTCTTCCAGGTG
AACACACCCGCAGCCGAGGTGCTCTACACAGTCATCCAGGACTGGGCCCAATTGGATGCG
GGGAGCATGGTGCTGGACGTGTGCTGTGGCACCGGCACCATTGGCCTGGCCCTGGCCCGG
AAGGTAAAGAGGGTCATTGGGGTCGAGCTATGCCCAGAGGCTGTGGAGGACGCCCGGGTG
AACGCCCAGGACAATGAGTTGAGTAATGTGGAGTTCCACTGCGGGAGGGCCGAGGACCTG
GTGCCCACCCTGGTGAGCAGACTGGCCTCCCAGCACCTCGTGGCCATCCTGGACCCACCC
CGTGCTGGCTTGCATTCCAAGGTGATCCTGGCCATCCGGAGAGCTAAGAACCTCAGGCGG
CTGCTGTACGTCTCATGCAACCCCCGGGCAGCCATGGGCAACTTTGTGGACCTCTGCAGA
GCCCCATCTAACCGGGTGAAGGGCATTCCCTTCCGGCCGGTCAAGGCTGTGGCAGTGGAC
CTGTTCCCGCAGACCCCGCACTGTGAGATGCTCATCCTGTTTGAGAGGGTGGAGCACCCC
AATGGCACAGGGGTCTTGGGGCCCCACAGGCCTCCAGCTCAACCCACACCAGGACCCCCA
GATAACACCCTACAAGAAACTGGGACCTTCCCCTCATCCTAGGCCCAGGACCCATTATGG
GGACAGGAGCTGAAGGGCTCCCAGGGCCCCAGGGGCGGGGAAGGCATGGCTTGCTGGCCA
GGATATTCAGAGGTAGGCAGCCAAAGGGCACAGAACCCTGTGCCCCACCAGGATTGGCCT
GCTGCTAGGGGCCAGGGGCCCTGACCAGGAAGACAGGTGGGTTCCTGTAGCCTGTTGCCC
ATCCCCAGGGCCTTTGTGGACTAGCCAAGGCTGTGAGGGCCAGAATAAACAACTGCTCAA
CAAAGG

Figure 3.14 Predicted human Htf9c cDNA sequence
Start, stop codons and polyadenylation signal is underlined

surrounding the first methionine codon(CCCATGG) is similar to the Kozak signal (GCCATGG). Assuming translation starts at this methionine residue, the putative human Htf9c protein is 613 amino acid residues in length.

Table 3.8 Exon and intron sizes of human Htf9c gene

No.	Exon Range	Exon size (bp)	Intron Size (bp)
1	66332-65686	647	269
2	65416-64841	576	257
3	64583-64475	109	224
4	64250-64069	182	186
5	63882-63768	115	88
6	63679-63564	116	75
7	63488-63377	112	765
8	62611-62489	123	178
9	62310-62235	76	197
10	62037-61921	117	110
11	61810-61714	97	116
12	61597-61102	496	

An alignment of the putative human and mouse Htf9c protein sequences is shown in Figure 3.15. The two proteins have an overall identity of 85% and similarity of 87%, with the N- and C- terminal ~40 amino acid residues having lower similarities.

The biological function of Htf9c gene currently is unknown. Since mouse Htf9c protein shares sequence similarity with yeast and bacterial nucleic acid-modifying enzymes, it has been suggested that the mouse Htf9c gene also may function in nucleic acid metabolism and/or modification (Guarguaglini et al., 1997). A database search with the human Htf9c protein sequence reveals that the protein also shares partial similarity with *E. coli* and *Haemophilus influenzae* tRNA (uracil-5-)methyltransferase (RUMT) (P23003 and P44643 respectively), *Aquifex aeolicus* RNA methyltransferase (AE000681), yeast endo-exonuclease (P33753), the *Chlamidia trachomatis* protein HOM1 (P55137) and several other hypothetical proteins. These similarities (identity ~30% and similarities ~40%) span the C-terminal two thirds of the protein. A profile database search predicts the presence of a RNA binding motif (residues 61-134, encoded by exon 2) and a methyltransferase motif (residues 417-472, encoded by exon 8-10), again suggesting that the protein may be

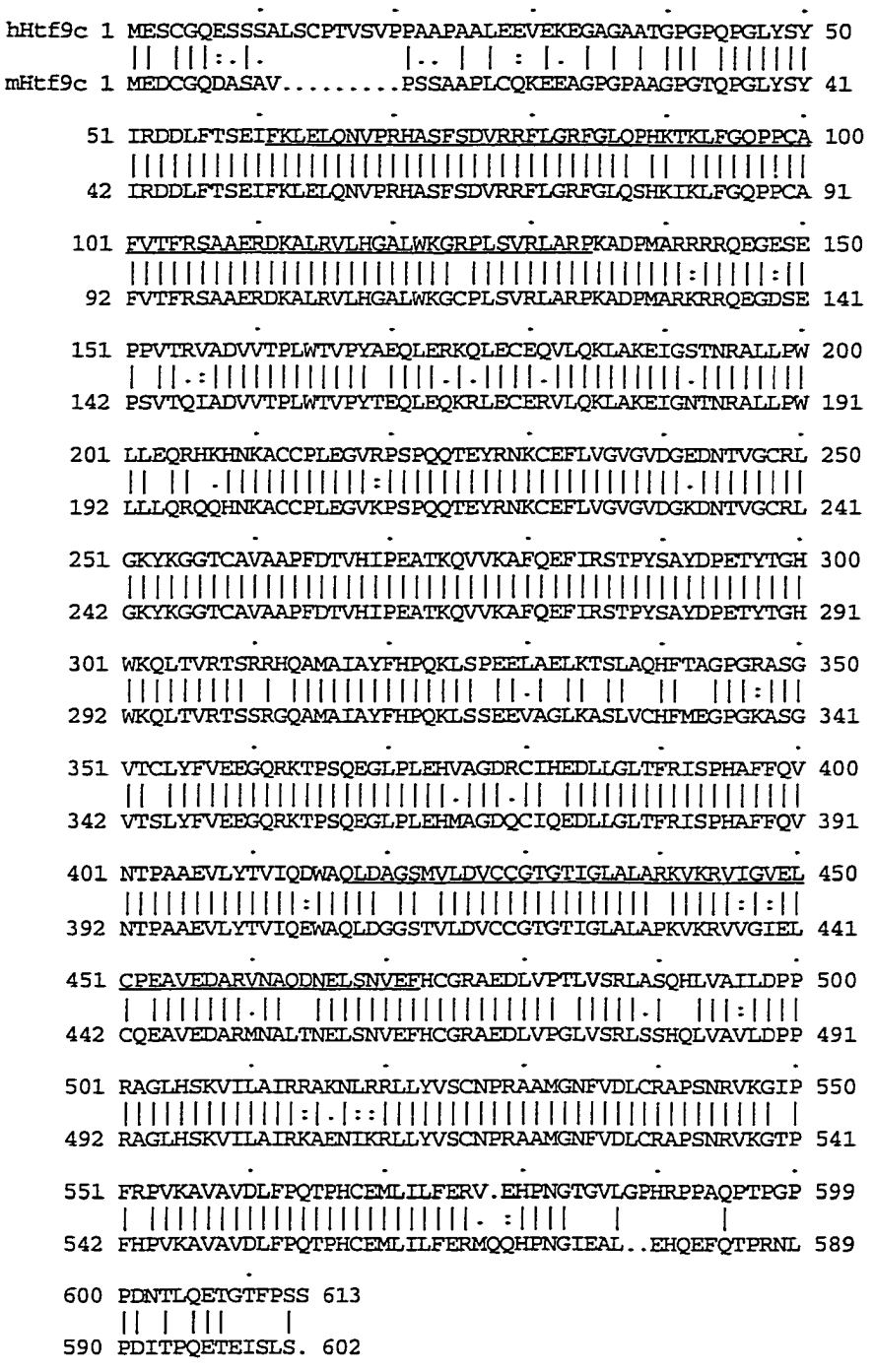


Figure 3.15 GAP alignment of human and mouse Htf9c protein
The putative RNA-binding domain (61-134) and the putative methyltransferase motif (417-472) are underlined.

involved in nucleic acid modification. Although Guarguaglini et al. (1997) indicated a putative nuclear targeting signal (KKRK) in their predicted mouse Htf9c protein, this signal sequence is not present in the newly interpreted protein sequence from this dissertation research and the protein localization program Psort predicted that this protein would be cytoplasmic.

In contrast to human Htf9c gene which has not been isolated, the human RanBP1 cDNA was isolated several years ago in a search for members of the RCC1 (the regulator of chromosome condensation)-Ran signal pathway (Bischoff et al., 1995). Alignment of genomic sequence with that of human RanBP1 cDNA sequence (D38076, X38067, differ in the length of 5' UTR) reveals that the human RanBP1 gene is composed of six exons (Table 3.9). As in the case of mouse, the human RanBP1 gene also is located in the region upstream of the Htf9c gene and in the opposite orientation.

Table 3.9 Exon and intron sizes of human RanBP1 gene

No.	Exon Range	Exon size (bp)	Intron Size (bp)
1	66304-66467	164	1349
2	67817-67953	137	3114
3	71068-71225	158	2886
4	74112-74240	129	866
5	75107-75172	66	586
6	75759-76159	401	

The six exons of the human RanBP1 gene encode a protein of 200 amino acid residues in length that is very rich in hydrophilic amino acid residues (21.6% Glu + Asp, 17.6% Lys+Arg). The human RanBP1 protein has been shown to be a cytoplasmic protein serving as one of the molecular partners of RAN (Bischoff et al., 1995). RAN was first identified due to its homology to the Ras family members in a GTP-binding domain (Drivas et al., 1990) and subsequently was shown to be involved in a number of cellular processes including the initiation of DNA replication, cell cycle progression, monitoring completion of DNA replication before entry into mitosis, nuclear protein import and RNA export (Rush et al, 1996). In addition to RanBP1 and RAN, other proteins of the RAN network have

been identified, such as RCC1, a regulator of chromosome condensation, and RanGAP, a Ran GTPase-activating protein. RCC1, a nuclear protein, specifically stimulates the exchange of guanine nucleotides on RAN (Bischoff and Ponstingl, 1991) while RanGAP1, a cytoplasmic protein, stimulates the GTPase activity of RAN (Bischoff et al., 1994). *In vitro*, it was shown (Bischoff et al., 1995) that RanBP1 protein, does not influence the intrinsic RanGTPase but co-activates RanGTPase if RanGAP1 is present. When complexed with Ran-GTP and RCC1, RanBP1 strongly inhibits the RCC1-induced exchange of Ran-bound GTP. RanBP1 also serves as a dissociation factor for the stable RanGTP-transport factor complex, thereby releasing the transport factor for a new round of transport and exposing RanGTP to nucleotide hydrolysis (Pu and Dasso, 1997).

Regulatory elements: Comparison of the start positions of the first exons of both human Htf9c and RanBP1 genes (see Tables 3.8 and 3.9) indicates that the 5' end exons of both genes overlap by 29 bp. Thus, it is expected that the promoter regions for both human genes are in close proximity or a common promoter region serves as a bidirectional promoter for both genes, as is the case for the syntenic mouse Htf9 locus (Lavia et al., 1987; Somma et al., 1991). Promoter-mapping experiments showed that the mouse bidirectional promoter was within a 273 bp region required for expression of both mouse genes (Somma et al., 1991). Since the promoter region of mouse Htf9 locus is well conserved in the human Htf9 locus (see below) and the expression pattern of human Htf9 locus is similar to that of mouse (Bressan et al., 1991), the expression of the human Htf9 locus also likely is controlled by a bidirectional promoter. Examination of the genomic sequence shows that there is a CpG island but no TATA box located near the 5' end of both the human Htf9c and RanBP1 genes. Alignment of this region with the promoter region identified for the mouse Htf9 locus (Somma et al., 1991; Di Matteo et al., 1998) shows that eight transcription factor binding sites, including three E2F binding sites, three SP1 binding sites and one HBF binding site and a CCAAT box, are conserved in the human Htf9 locus (Figure 3.16).

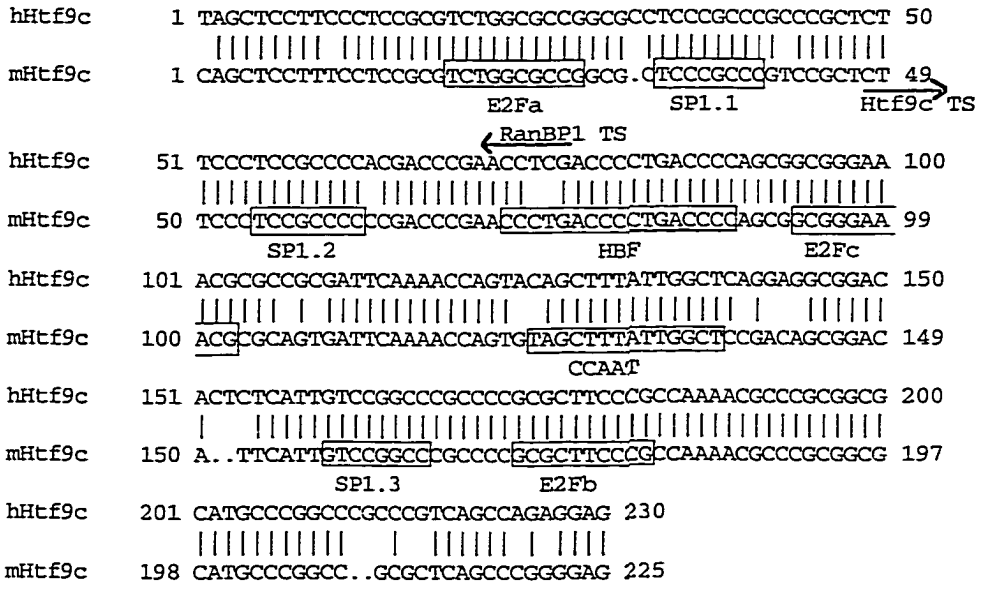


Figure 3.16 Alignment of promoter region of human and mouse Htf9 locus
Human sequence is the complemented sequence of contig positions 66270-66500. Locations of transcription factor binding sites within the mouse promoter are according to Di Matteo et al. (1998). Putative human Htf9c gene transcription start (TS) site was based on the mouse Htf9c cDNA sequence (X56044) and Matteo et al.(1998). Putative human RanBP1 transcription start site was located based on the human RanBP1 cDNA sequence (D38076).

The expression of both mouse Htf9c and RanBP1 genes are cell-cycle dependent, a process where regulation is controlled by E2F activators and repressors of the retinoblastoma gene family (Guarguaglini et al., 1997). Promoter expression assays have demonstrated that within the 273 bp mouse Htf9 promoter region, a 74 bp sequence including the SP1.2 binding site, the major transcription start sites, and the E2Fc site binding site is sufficient for basal transcription activity in both orientations. The E2Fa region is important for cell cycle control of Htf9c transcription. The region containing the SP1.3 and E2Fb sites contributes to transcriptional activation in both promoter orientations and influences both basal and cell cycle-dependent transcription activation. The high degree of sequence conservation between the human and mouse Htf9 promoters may indicate that expression of both the human and mouse Htf9 genes are controlled by a similar transcription regulation mechanism. Indeed, the expression patterns of the human Htf9c

and RanBP1 genes are very similar to that of mouse (Bressan et al., 1991), indicating that the expression of both human genes also is cell cycle-dependent.

As revealed by a blastn search of dbEST, there are about 50 EST entries which are well-aligned to the 3' end of the human RanBP1 gene in the genomic sequence. More than half of these ESTs terminate near contig position 75985 and approximately one third terminate at or near contig position 76160 to produce a transcript which would be 170 bp longer. Examination of the sequences near the 3' ends of these two categories of transcripts reveals that there is one putative polyadenylation signal (TATAAA) at contig position 75970 (II) and another (AATAAA) at contig position 76135 (III). Thus, transcripts with different 3' ends most likely would be obtained because of alternative polyadenylation positions. Another possible polyadenylation signal (I) (AATAAA) also occurs just downstream from the coding region of the sixth exon at contig position 75852. Since very few ESTs terminate around this position, the signal most likely functions, but less efficiently than the other two. The position of the three putative polyadenylation signals and related data are summarized in Figure 3.17. Detailed examination of the tissue source of these ESTs indicates that the alternative polyadenylation may be somewhat tissue specific. Such tissue specific regulation of alternative polyadenylation has been shown for a few other m-RNAs (Gautheret et al., 1998). The sequences between the three potential polyA signals are very T rich and slightly A rich. For example, the 112 bp sequence between signals I and II is 70% A+T and 56% T. The 189 bp sequence between signals II and III is 68% A+T and 39% T. It has been noted by Shaw and Kamen (1986) that AT-rich sequences near a polyadenylation signal may be involved in RNA degradation.

In contrast to the multiple potential poly-A signal sequences downstream of the RanBP1 gene, the human Htf9c gene only has a single putative polyadenylation signal (AATAAA) located at contig position 61119, 17 bp upstream of the 3' end of the terminal exon of the gene.

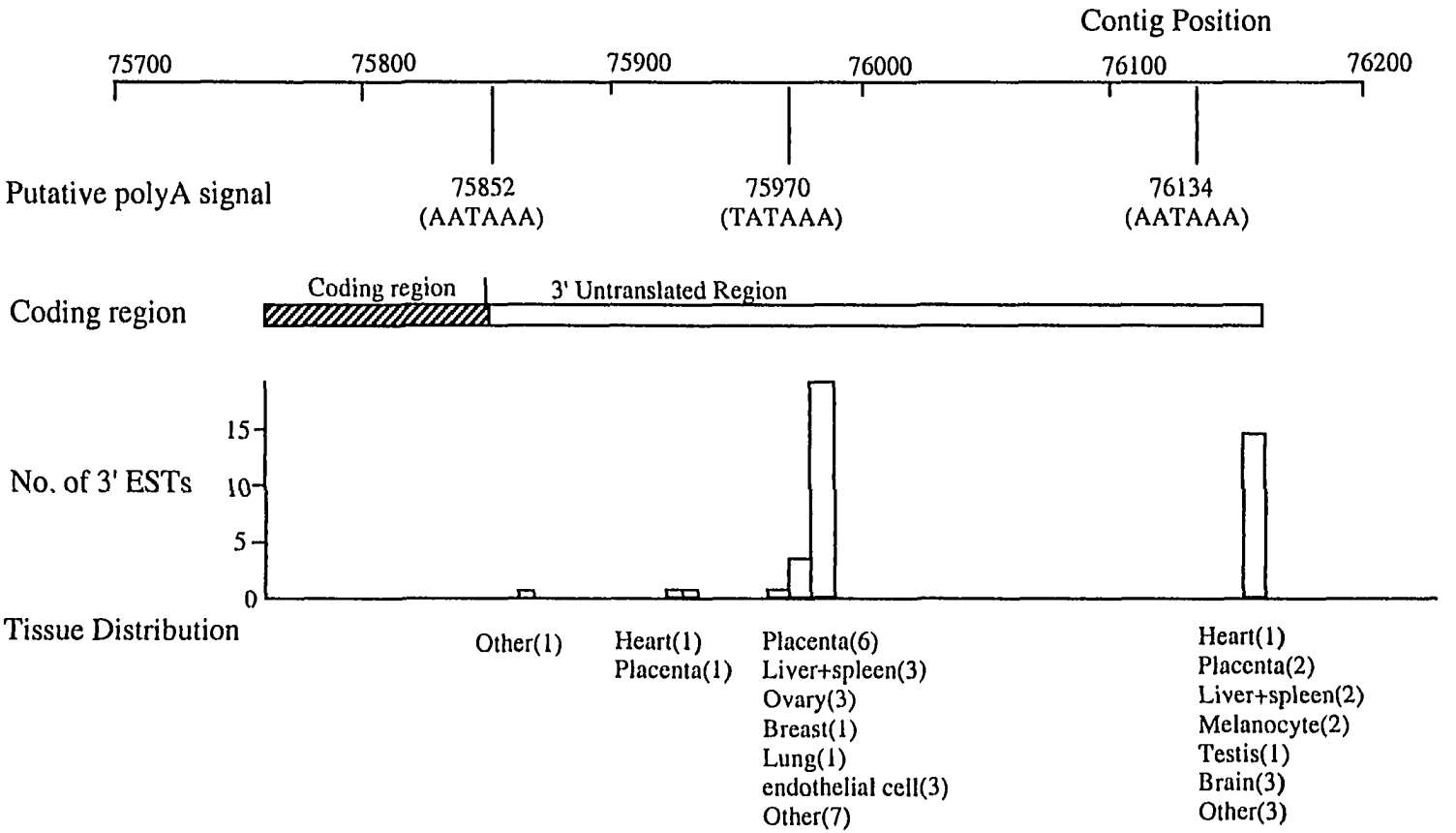


Figure 3.17 Distribution of 3' ends of RanBp1 transcripts as revealed by 3' ESTs and their tissue distribution.
Data extracted from Powblast output. Nos of ESTs from same tissue areincluded in the parenthesis.

3.4.3 Putative human gene PutA

The region located at contig position 25000 to 60000 has significant homology to several ESTs. By aligning the EST sequences to the contig sequence, 12 exons (or 11 exons, see below) coding for a putative gene, termed PutA, are revealed (Table 3.10). These 12 exons form two clusters. The first cluster (exons 1-2) is represented by one EST (T06489) and contains stop codons in all three frames, suggesting that they may represent the 5' or 3' end of a gene. The second cluster (exons 7-16) is represented by multiple ESTs and has an open reading frame which terminates at exon 16. Since there are no obvious promoter elements in the genomic DNA sequence upstream of the second exon cluster, and the first exon (exon 1) from the first EST cluster is located in a major CpG island, it is likely that these 12 exons may be expressed from a promoter in this CpG island. Genscan also predicts three genes in this same region with two on the forward strand and one on the reverse strand. Comparison of the results from Genscan with the EST data shows that one of the predicted genes likely is incorrect since it is located within an intron whose flanking exons are spanned by an EST (AA672501). Both Genscan and Xgrail predicted four more exons between the two clusters of exons, three of which are at identical or similar contig positions. Human and mouse sequence comparison by Geneligner also predicted four exons, all of which are at identical or similar contig positions to the four exons predicted by Genscan. These four exons (3-6) form an ORF which is compatible with the exon clusters located at 5' and 3' sides of these exons. Thus, the putative PutA gene most likely has 16 exons. However, an alternative form of PutA with only 15 exons also is possible since the 3' end of the putative gene has homology to more than 20 human ESTs and some of those ESTs, such as AA203219, overlap the 3' end exon but extend in the 5' direction into the last intron while still others, such as AA312704, overlap exon 15 but extend in the 3' direction into the last intron. These observations strongly suggest that the PutA gene may have an alternative splicing pattern where intron 15 is part of a large 3' end exon. The

predicted exon and intron sizes of the putative gene PutA are listed in table 3.10. The predicted protein sequence is shown in Figure 3.18.

Table 3.10 Exon and intron sizes of putative PutA gene

No.	Exon Range	Exon	Intron(bp)	Evidence
1	29088-29188	101	5302	T06489
2	34491-34625	135	132	T06489
3	34899-35487	589	478	Geneligner, Genscan
4	35966-36125	160	2347	Geneligner, Genscan
5	38473-38615	143	164	Geneligner, Genscan
6	38780-39062	283	1176	Geneligner, Genscan
7	40239-40436	198	236	AA000671, AA431537
8	40673-40774	102	838	AA672501, AA431537
9	41613-41711	99	1805	AA672501, AA431537
10	43517-43599	83	11380	AA672501, AA431537
11	54980-55080	101	314	AA672501
12	55395-55501	107	567	AA672501
13	56074-56201	128	1491	AA217940,AA431213
14	57693-57806	114	1021	AA217940,AA431213
15	58828-58882	55	1306	AA217940,AA431213
16	60189-60675	487		AA431213, AI028241
(15'	58828-60675	1848		see text)

```
>hputa.prot
MKRAIGAEVMDVGSGGDGQSELPAEDPFNFYGASLLSKGSFSKGRLLIDP
NCSGHSPRTARHAPAVRKFSPDLKLLKDVKISVSFTESCRSKDRKVLYTG
AERDVRAECGLLLSPVSGDVHACPFGGSVGDGVGIGGESADKKDEENELD
QEKRVEYAVLDELEDFTDNLELDEEGAGGFTAKAIVQRDRVDEEALNFPY
EDDFDNDVDALLEEGLCAPKKRRTEEKYGGDSDHPSDGETSVQPMMTKIK
TVLKSRGRPPTEPLPDGWIMTFHNSGVPVYLHRESRVVTWSRPYFLGTGS
          WW domain
IRKHDPPLSSIPCLHYKKMKDNEEREQSSDLTPSGDVSPVKPLSRSAELE
FPLDEPDSMGADPGPPDEKDPLGAEAAPGALGQVKAKVEVCKDESVDLEE
FRSYLEKRFDFEQVTVKKFRTWAERRQFNREMKRKQAESERPILPANQKL
ITLSVQDAPTKKEFVINPNGKSEVCILHEYMQRVLKVRPVYNFFECENPS
                      dsRBP domain
EPFGASVTIDGVTYGSGTASSKKLAKNKAARATLEILIPDFVKQTSEEKP
KDSEELEYFNHISIEDSRVYELTSKAGLLSPYQILHECLKRNHGMGDTSI
                                dsRBP domain
KFEVVPGKNQKSEYVMACGKHTVRGWCKNKRVGKQLASQKILQLLHPHVK
NWGSLLRMYGRESSKMVKQETSDKSVIELQQYAKKNKPNLHILSKLQEEM
KRLAEEREETRKKPKMSIVASAQPGG(EPLCTVDV*)KQCWPVGRPQHIL
PSLPV*
```

Figure 3.18 Predicted protein sequence of PutA gene
Sequence in parenthesis indicates the region of an alternative splicing pattern resulting in only 15 exons

The predicted PutA cDNA encodes for putative proteins with either 744 or 734 amino acids beginning at the AUG methionine codon at contig position 34612 in exon 2. A protein database search failed to reveal any proteins with high homology to PutA although partial regions of homologies were observed with a mouse leucine-rich acidic nuclear protein (P49911) and a human double-stranded RNA-specific adenosine deaminase (P55265). A profile search revealed three putative protein domains. The first is a WW domain, which may be involved in protein-protein interaction (Chen and Sudol, 1995) and is predicted at residues 262-295. The other two are both double stranded RNA-binding domains (dsRBD) at residues 472-539 and 580-647. TMpred does not predict a transmembrane domain and protein targeting site analysis with Psort predicts that this protein would be targeted to the nucleus. Thus, the putative PutA gene most likely encodes for a nuclear protein involved in nucleic acid modification.

3.4.4 Human putative gene PutB

The region from contig positions 84000 to 96000 also contains homology to several ESTs and from the EST data alone, Geneligner predicts the presence of a gene encoded on the forward strand with 10 exons. Genscan also predicts one gene on the forward strand with 12 exons, 10 of which are identical or similar to the 10 exons revealed by EST data. By incorporating both EST data and Genscan predicted exons, a single gene with 11 exons is predicted. Examination of these predicted exons shows that the 3' end exon has a number of ESTs aligned to it and contains an in frame stop codon, and thus, should be the 3' end coding exon. There also is a sub-optimal polyadenylation signal (ATTAAA) located at contig position 95237, 18 bp upstream from the end of a predicted exon, exon 11. Further, at least six more ESTs, four of which are 3' ESTs, are located in the region 96165-96811, about 900 bp downstream from exon 11. The four 3' ESTs terminate at the same contig position, 96811, and have a polyadelynation signal (AATAAA) corresponding to contig position 96789, 22 bp upstream of the end of these ESTs. These features suggest

that the region between 96165-96811 is a 3' end exon (exon 12) of an upstream gene. Since there are no ESTs spanning this exon and upstream exons, the 5' boundary of this exon can not be determined. Also, whether this exon represents the result of alternative polyadenylation or alternative splicing can not be determined. While the 3' end of this putative gene can be located by EST data, no ESTs were found that would signal the 5' end. Examination of the genomic sequence also revealed no TATA box in the 10 kb contig region upstream from the first EST-revealed exon (exon 2). However, a prominent CpG island is found within the region at contig positions 79600-81300, about 4 kb upstream of the exon that Genscan predicted as the initial coding exon of this putative gene. Thus, it is likely that the predicted 5' end exon may represent the 5' end coding exon of the putative gene. The resulting gene model for the putative gene (referred to as PutB) contains 12 exons, corresponding to a translation product of 729 amino acid residues (Table 3.11 and Figure 3.19).

Table 3.11 Exon and intron sizes of putative PutB gene

No.	Exon Range	Exon (bp)	Intron (bp)
1	84976-85064	89	2933
2	87998-88119	122	162
3	88282-88439	158	84
4	88524-88696	173	222
5	88919-88985	67	434
6	89421-89512	92	162
7	89675-89816	142	204
8	90021-90146	126	82
9	90229-90276	48	1286
10	91560-92560	1001	1472
11	94033-95255	1223	
12*	96165-96811	>646	

* The 5' boundary of exon 12 not determined. see text for detail.

Figure 3.19 Predicted cDNA and protein sequence of PutB gene

Note: DNA sequence of exon 12 is not included since it's 5 end can not be determined at this stage.

A protein database search with the sequence of the putative PutB protein revealed no significant homology with any known proteins. However, the N-terminal ~200 residues of the PutB protein did show significant homology to 4 yeast hypothetical proteins (P34679, Q09701, P42836 and O14345), two of which were suggested to be transmembrane proteins in their Genbank entries, and a putative protein of *C. elegans* (AL032657). A profile search revealed a putative Zn-finger motif at residues 99-149 and Psort predicted the protein to be targeted to plasma membrane. Since two of the homologous hypothetical proteins also were suggested to be transmembrane proteins, the putative protein sequence was analyzed by transmembrane domain prediction with TMpred. The result (Figure 3.20 and Table 3.12) suggests that the protein contains 4 transmembrane domains and the N-terminal is predicted to be within the cytoplasmic membrane.

Table 3.12 TMpred prediction result of PutB gene

No. of tm helices	Amino acid residues
1	14 - 38
2	44 - 63
3	144 - 167
4	180 - 204

3.4.5 Other putative genes in contig 3

In the remaining regions of this contig (95000-162000), Genscan predicted five genes and Xgrail also predicted multiple exons. However, a blastn search of the dbEST database unexpectedly revealed no significant EST homologies. To evaluate the possibility that there could be genes within this region, the predicted protein sequences for the five putative genes were searched against the protein database. The N-terminal of ~160 residues of one of the Genscan predicted proteins had a modest degree of homology to segments of several known proteins, most of which were collagens, such as the sea urchin collagen alpha 2 chain homologue (S23809). Examination of the protein sequences of the

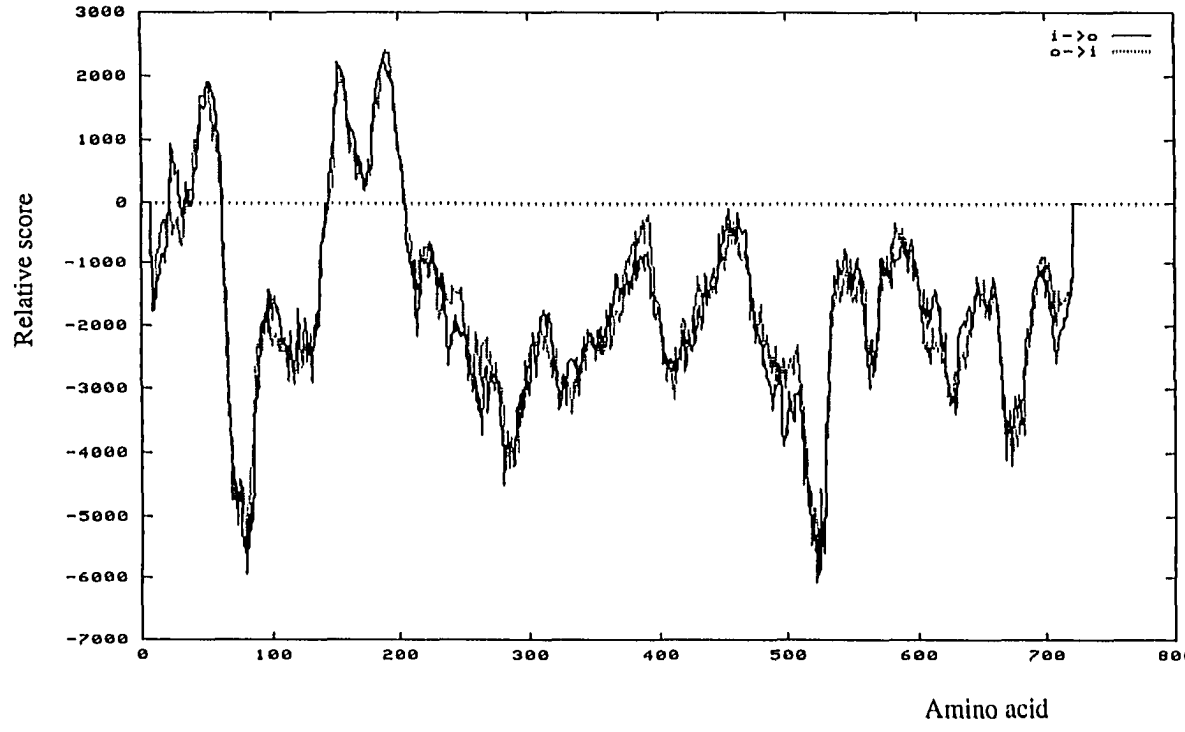


Figure 3.20 Transmembrane domain prediction result of PutB gene by TMpred

homologous regions shows that the collagen sequences are composed of recursively repeating of several amino acid residues while the predicted protein sequence is not. Thus, there is no evidence for the presence of any genes in the remaining region of this contig.

3.4.6 Repetitive elements in contig 3

In contrast to contigs 1 and 2, both of which contains over 40% of repetitive sequences, contig 3 contains only 22.3% repeated sequences (Figure 3.21 and Table 3.13). Furthermore, these repeated sequences are not evenly distributed along the length of the contig but are at a higher density in the first 100 kb, where five known or putative genes also are located. This trend holds for each individual repeat family, including Alu, L1 and LTR repeats. Again, there does not appear to be a preference for orientation or location of Alu repeats in this region but the L1 repeats have a preference for relatively large introns and intergenic regions. Interestingly, the three segments of L1 repeats in the intron region of human T10 gene and the ten segments of L1 repeats in the intron region of PutA gene all are oriented in the opposite direction of transcription of both genes.

Table 3.13 Repetitive elements in contig 3

	number of elements	length occupied	percentage of sequence
SINEs:			
ALUs	68	18515 bp	11.41 %
MIRs	9	837 bp	0.52 %
LINEs:			
L1	10	5011 bp	3.09 %
L2	3	654 bp	0.40 %
LTR elements	10	5938 bp	3.66 %
MER	5	1613 bp	0.99 %
Simple Repeat	33	3188 bp	2.00 %
Total interspersed repeats:		32933 bp	22.30 %

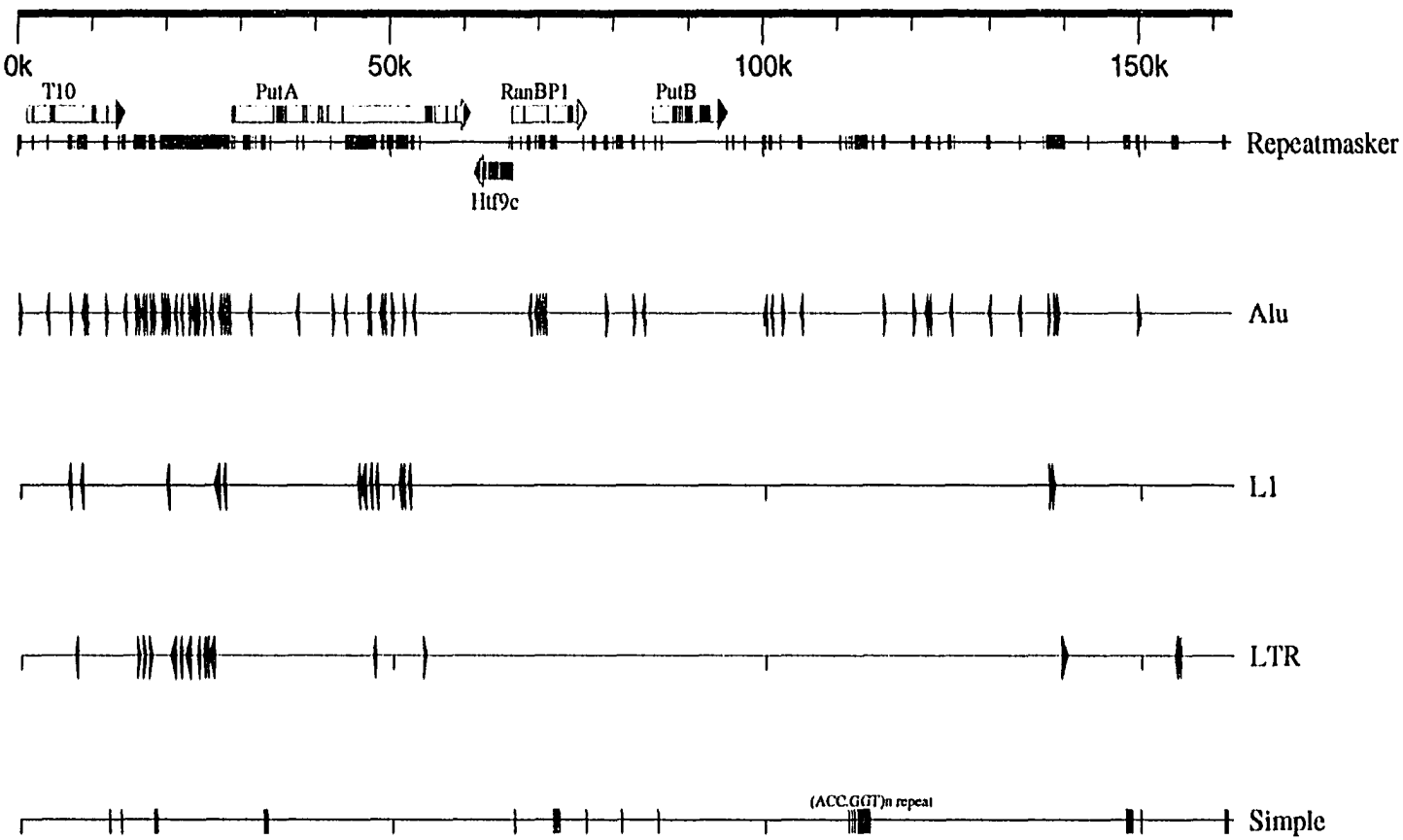


Figure 3.21 Repetitive elements in contig 3

A repeatmasker analysis shows that the region from contig positions 110849-113658 is rich in simple repeats composed of repeating units of $(ACC)_{n=1-9}$ (or GGT in the complemented strand) frequently separated by ATCACT (Figure 3.22). Surprisingly, both Genscan and Xgrail predicted a majority of the region to be coding. Genscan predicted a forward strand single exon gene while Xgrail predicted two exons in the opposite strand (see Figure 3.11). The Genscan predicted protein consists primarily of threonine in the N-terminal two-thirds and proline in the C-terminal one-third. A protein database search with the predicted amino acid sequences failed to reveal any similarities to known proteins. Currently, there are two known eukaryote proteins which are encoded largely by simple tandem repeats (Djian et al., 1998). One is involucrin, a keratinocyte protein , which contains 39 repeats of three glutamines and two glutamic acids (Eckert and Green, 1986) while the other is a nuclear protein of unknown function composed largely of glutamine and histidine (Cox et al., 1996). The repeating units which code both of the known proteins are CAG. Examination of the protein sequences of both of the known proteins shows that at least a portion of them consists of unique amino acid sequences while the predicted protein sequence in this contig consists almost exclusively of repeats. In addition, there are no potential TATA boxes, CpG islands or polyA signals found on either side of the repeat sequence. Given these observations, it is unlikely that the repeat in this contig actually codes for a gene although this possibility can not be excluded.

Tandem repeat sequences are abundant in eukaryote genomes. A survey of the Genbank nucleic acid database shows that on average, one simple tandem repeat is present in every 6 kb of human genomic sequence (Beckman and Weber, 1992). Two classes of tandem repeat sequence, i.e. minisatellites and microsatellites, have been intensively studied in recent years, primarily because of their association with some human disorders (Caskey et al., 1992) and their usefulness in genome mapping (Jeffreys et al., 1985). Minisatellites, which are predominantly located at the terminal regions of human chromosomes, are tandem repetitive sequences of up to 10 kilobases with repeating units

>Contig sequence from 110849 - 113658
CATCATCACCACTGTCACCATGAGCATGGCCATAACCACCATCATGGTCAGCTCCACCAG
CATCACCACCACCACCACCACCATCACCATCACTGCCACCACCACCGCCACCATGACTAC
CACCACCATCACCACCACCACCACCATCACTACCACCACCACCATCACTACCACCACCAC
CACCACGATCACTACCACCACCACCACCATCACCATCACTACCACCACCACCACCATCAC
CATCACCACCACCATCACTACCACCACCACCACCACCACCACCACCACCATCACTACCACCAT
CACCACCACCATCACCACCACCACCACCACCACCATCACTACAACCACTACCACCACCAC
CACCATCACTACCACCACCACCATAACCATCACTACCAGCATCACCACCACCACCATCAC
CATCACTACCACCACCATCACCATCACTACCACCACCACCACCACCACCACCATCACTACCAC
CACCACCATCACTACCACCATCACCATCACTACCACCACCACCACCACCACCACCATCACTAC
CACCACCACCACCACCATCATCACCACCACCACCACCACCACCACCATCACTACCACCAC
CACCACCATCACTACCACTACCACCACCATCACTGCCACCACCACCACCACCACCATCAC
TACCACCACCACCACCATCACTGCCACCACCACCACCACCACCACCATCACTACCACCACCAC
CATCACCACCACCACCACCACCATCACTACCACCACCACCACCACCACCACCACCATCACTAC
CACCACCATCACTACCACCACCACCACCACCACCATCACTACCACCACCACCACCACCATCAC
CACCACCACCATCACTACCACCACCACCACCACCATCACTACCACCACCACCACCACCATCAC
TACCACCACCACCACCACCACCATCACTACCACCATCACCACCACCACCACCACCATCACCAC
CATCACCACCACCACCACCACCACCATCACCACCACCACCATCACCATCACTACCACCACCAC
CACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCAT
CACCACCACCACCACCACCATCACTACCACCACCACCACCACCACCACCACCATCACTACCACCAC
CACCACCACCACCACCACCATCACTACCACCACCACCACCACCACCACCACCACCACCACCAC
TACCACCACCACCACCATCACCATCACTACCACCATCACCACCACCACCACCACCATCACCAC
CACCACCATCACCATCACTACCACCACCACCACCACCACCACCACCACCACCACCACCACCAT
CACCACCACCACCACCATCACTACCACCACCACCACCACCACCACCACCACCACCACCACCAC
CACCACCACCATCACTACCACCACCACCACCACCACCACCACCACCACCACCACCACCACCAC
TACCACCACCACCACCATCACCATCACTACCACCATCACCACCACCACCACCACCATCACCAC
CACCACCATCACCATCACTACCACCACCACCACCACCACCACCACCACCACCACCACCACCAT
CACCACCACCATCACTACAACCGCCATCACTACCACCATCACCACTACCACCATCACTAC
CACCACCACCACCACCATCACCATCACTTCCACCACCACCACCACCACCACCATCACTACCAC
CACCACCACCACCATCACCATCACTTCCACCACCACCACCACCACCACCATCACTACCACCAC
CACCACCACCATCACCATCACTTCCACCACCACCACCACCACCACCATCACTACCACCACTAC
CACCAGGACCACCATCACTACCACCACCACCACCACCACCATCACTACCACCATCACCACCAC
CACCACCACCATCACTACCACCACCACCACCACCACCATCACTACCACCACCACCACCATCACCAT
CACCACCACCACCACCACCACCACCATCACTACCACCACCACCACCACCACCACCACCACCAC
CACCACCATCACTACCACCACCACCACCACCACCACCATCACTACCACCACCACCACCACCATCAC
TTCCACCACCACCACTGCCATCACTACCACCACCACCACCACCATCACCATCACTACCACCAC
CCCACCACTACCACCATCACCACCACCACCACCACCATCACTACCACCACCACCACCACCACCATC
ACTACCACCACCACCATCACCATCACTTCCACCACCACCACTGCCATCACTACCACCACCC
ACCACCATCACCATCACTACCACCACCACCACCACCACCATCACCATCACTACCACCACCACCC
ACCACCACCATCACTACCACCACCACCACCACCACCATCACTACCACCACCACCACCATCACTACCACCC
ATCACCACCATCACTATCCCCACCATCACTACCACTGCCACCACCACCACCACCATCACTACCC
ACCACCACCACCATCACTACCACCACTACCACCACCACCACCACCACCATCACCACCACCACCACCC
ATCACTACCACCACCACCACCAGCACCACCACCATCACTACCACCATCACCACCACAACC
ACCATCACCACCACCATCATCAGAACCACCACCATCACCATCACTACCACCACCACCACCC
ATCACCACCACCACCATCACCATCACTATCACCACCACCATCACCATCGTTACCATCACC
ATCACCATCACCACCACCACCACCATCACCATCACTATCACCACCACCATCACCATCGTT
ACCATCACCATCACCATCACCATCACCATCACCATCACCACACCTCCATCACTATGGTCG
CTGTGACTGAGCCCATCCTCACTACTGTCATCAGCACTGTCACAGCACCATGCTGACAAG
CCCACAGTGGGTTTGAGGGCTGCCATTGCTTGGGACTAGTTCCTCAGGCACAGGAGGCTC
CACGCTGGTCTGATTGCATTGGAGGGGATGCCTGACAAAGGTCTTTATGGGGTGTGCATT
GTGGAGTGGGGAGGGGGAGGCTCAGGGAGGTCTCGAGGACTTCATGCCGTGGGCATTCTC
CCCGGCCCCCTTCTGTCTCAGCAGCTCCCCACGCTTCTTGACTGAGCTGGAGCAGAGGAA
GCAGGCAGGGGTGGGAGGAGGGGCAGGGGCAGGAGAAGGGGCAGGGGCAG

Figure 3.22 Nucleic acid sequence of the minisatellite in contig 3

of 9-100 bp (Sertedaki and Lindsay, 1996). In contrast, microsatellites have shorter repeating units (1-6 bp), span regions of about 50-100 bp and are dispersed throughout the human genome. The large repeat sequence in this contig might be classified as a minisatellite due to its overall size and that it has a repeating unit of $(ACC)_{5-7}(ATC)(ACT)$.

Although minisatellites are common in human genomes, the repeat from this contig, which is composed of primarily trinucleotide repeats, seems unusual. In hybridization experiments for trinucleotide tandem repeats in the human genome, Armour et al. (1994) identified multiple alleles of trinucleotide repeats with lengths generally less than 300 bps. Longer triplet repeats often are reported in association with human diseases such as mytonic dystrophy, a disease that results from expansion of a CTG repeat in the 3' untranslated region of the myotonin kinase gene (reviewed by Riggins et al., 1992). Specifically, a short triplet repeat with 5-30 copies of CTG in normal individuals was expanded into a triplet repeat with up to 2000 copies in affected patients (Harley et al., 1992). In a search for other trinucleotide repeats, Armour et al. (1995) also isolated two repeat loci from normal individuals which contain thousands of irregular repeating GGA-GGT triplets which they claimed were "greatly exceeding repeat copy number at most trinucleotide and other simple repeat loci". These repeat loci have been mapped to human chromosomes 15 and 22 and the repeat locus which maps to chromosome 22 ranges in size from 1.6 - 3.5 kb between different individuals based on PCR experiments. In view of these observations, the tandem ACC/GGT repeat that is described in this present work most likely is among the largest trinucleotide repeats in the human genome as it spans about 2.5 kb.

Recent studies of repeat sequences, especially trinucleotide repeats, also suggest that the presence of large trinucleotide repeats may have important biological significance. Firstly, tandem repeat sequences, including both minisatellites and microsatellites frequently are unstable and undergo either deletion or expansion (Buard and Jeffreys, 1997; Schlotterer, 1998). Secondly, trinucleotide repeats tend to form alternative secondary

structures as suggested from *in vitro* experiments (Pearson and Sinden, 1996; Maueler et al., 1998). Although there were no such experiments directly involving the (ACC/GGT)n triplet repeat as observed here, it is likely that an alternative secondary structure readily could be formed. It is known that potassium ions stabilize four strand DNA structures formed by guanine-rich DNA (Sen and Gilbert, 1990) and that extremely stable intramolecular tetrads are formed in oligonucleotides composed of $(GGGT)_4$ in the presence of potassium ion (Jing et al., 1997). Therefore it is not surprising that only in the absence of potassium ions could the repeat region of this contig be specifically amplified by a PCR reaction as discussed above and shown in Figure 3.23. Other studies supporting the biological significance of trinucleotide repeat sequence include those indicating that several single- strand DNA binding proteins bind to a specific subset of trinucleotide repeat sequence including (GGT)n (Yano-Yanogisawa et al., 1995; Maueler et al., 1998) and that chromatin structure may have special property when assembled at trinucleotide repeats (Rossetti et al., 1998).

Therefore, the presence of the minisatellite sequence in this contig may have some as yet unknown significance that could range from being involved in chromosomal rearrangements or deletions, such as observed in DGS, or one of the above discussed functions.

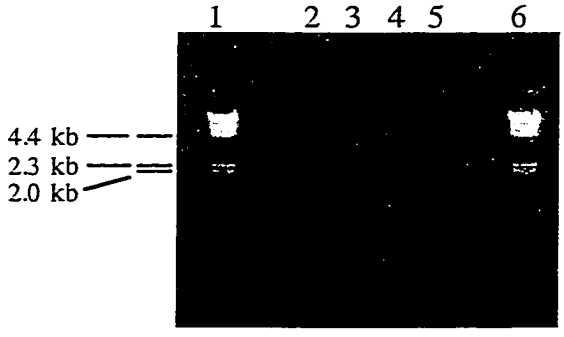


Figure 3.23 PCR amplification of the ACC/GGT repeat region in Contig 3
Primer 1: GTCAGGCTCAGTCCACGCTCTGCTCTTC
Primer 2: GGCATCCCCTCCAATGCAATCAGACCAG
Cycle condition: 95°C 1 min, 65°C 45 sec, 72°C 2 min, 25 cycle
Lanes 1 and 6: size marker
Lanes 2 and 3: PCR buffer without KCl
Lanes 3 and 4: PCR buffer with 50 mM KCl

3.5 Analysis of contig 4

Contig 4 is composed of a single mouse BAC clone 72k21 which has a 118,243 bp insert. Detailed analysis of this contig shows that the contig contains 3 known genes, i.e. mouse T10, mouse Htf9c and mouse RanBP1, and a putative gene PutA (Figure 3.24).

3.5.1 Mouse T10 gene

Eight exons are revealed by aligning the mouse T10 cDNA (X74504) with the first 50 kb sequence of this contig (Figure 3.24). Although the 5' 18 bp of this mouse T10 cDNA does not match the genomic sequence, two other mouse ESTs (AA274508, AA200446) do match the genomic sequence exactly but do not match the 5' 18 bp of the mouse T10 cDNA. In addition, an EST entry (AA200446) which overlaps four existing exons reveals one more exon 5' to the previously known 8 exons. This exon is located in a CpG island and probably represents the 5' end exon of the mouse T10 gene. The nine exons are 1575 bp long and contain a single open reading frame, starting from a methionine codon within the second exon and terminating within the 9th exon to produce a putative protein with 276 amino acids. Within the coding region, six discrepancies between the cDNA sequence and the genomic sequence were observed that would result in 3 amino acid differences in the putative translation products. The exon and intron sizes of the mouse T10 gene are listed in Table 3.14 and the translated protein sequence is shown in Figure 3.25.

Table 3. 14 Exon and intron sizes of mouse T10 gene

No.	Exon Range	Exon size (bp)	Intron Size (bp)
1	8072-8170	99	19348
2	27519-27612	94	7622
3	35235-35323	89	3889
4	39213-39332	120	1655
5	0988-41102	115	2828
6	43931-44001	71	4133
7	48135-48288	154	892
8	49181-49285	105	1133
9	50419-51146	728	

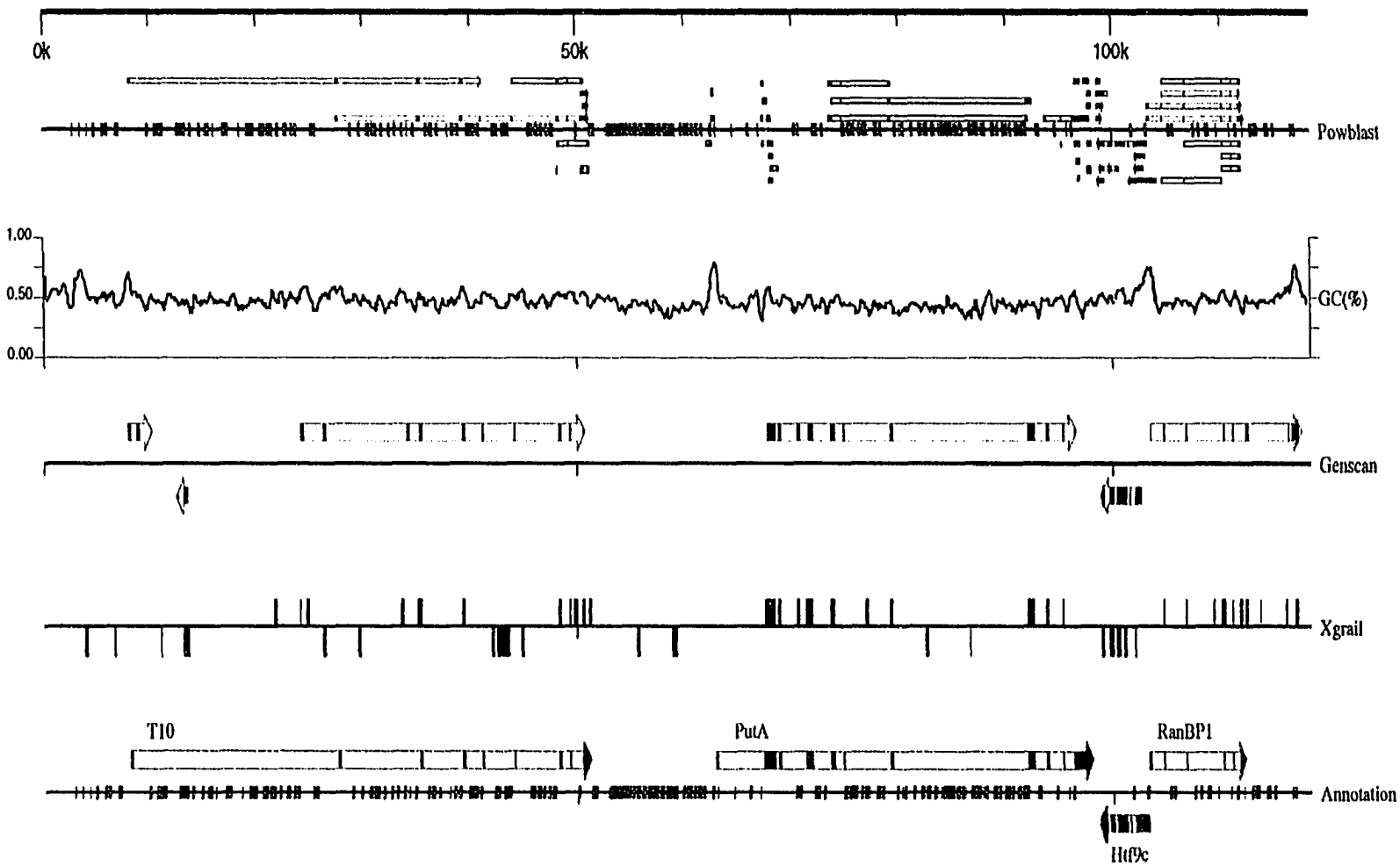


Figure 3.24 Features and annotation of contig 4 (see Figure 3.2b for legend)

>mouse T10 protein
MCIIFFKFDPRPVSKNAYRLILAANRDEFYNRPSKLADFWGNNSEILSGLDMEEGKAGGT
WLGISTRGKLGALTNYLQPRQEPDARGRGELVSHFLTSDMDSLSYLKKVSTEGHLYNGFN
IIAADLSTSKGDVVCYYGNRGEPEPIVLTPGTYGLSNALLETPWKKLCFGKQLFMEAVEQ
SEALPKDVLVTQLLDVLNNEEAQLPDPAIEDQGQEYVQPILNKYAAVCVRCATYGTRTNT
IILVDANGHVTFTERSMLDKDTSRWETNTYEFTLQS

Figure 3.25 Protein sequence of mouse T10 gene
The three amino acids which differ from published mouse T10 protein sequence are underlined.

Although the function of the T10 gene currently is unknown, a protein database similarity search indicates that the protein is similar to a putative *C. elegans* protein (e257662) of unknown function, with an identity of 31% and similarity of 49% but a profile database search did not reveal any known protein motif. Since this gene is conserved in several different species, including human, mouse, rat and *C. elegans*, it likely plays an important biological role.

Analysis of the genomic sequence also reveals that two rat ESTs align to the 3' terminal exons of the mouse T10 gene. One of the ESTs (AA799663) aligned to exons 7, 8 and the 3' terminal 253 bp of exon 9 of the mouse T10 gene and therefore the 5' 473 bp of mouse exon 9 was not present in the rat EST. The other EST entry (AI010662) aligned to the 5' end 115 bp and the 3' end 253 bp of the terminal exon (exon 9) of the mouse T10 gene and therefore an internal 357 bp of mouse exon 9 was not present in the rat EST sequence. When aligning both of the EST entries to the mouse genomic sequence, typical splicing site consensus sequences were found at all gap boundaries. These results suggest that the mouse exon 9 might be split into two exons in rat. This possible lack of conservation between mouse and rat correlates with the relatively low conservation of exon 9 between the human and mouse T10 genes, suggesting that part of the terminal exon of T10 genes might be subjected to lower selection pressure than other portions of the genes.

Regulatory elements: There is no TATA box predicted by Xgrail in the 5' region of the mouse T10 gene. However, as revealed by Xgrail, the first exon is located in a CpG island (7877-8238). Promoter prediction with NNPP in the region (7000-8200) predicted four promoters at contig positions 7702-7752, 7710-7760, 7798-7848 and 8019-8069,

respectively, while promoter prediction with TSSW predicted a promoter at contig position 7931. Considering that the fourth promoter region predicted by NNPP (8019-8069) and the promoter predicted by TSSW (7931) are relatively close to the first exon which starts at contig position 8072, and that the region (7870- 8070) contains several segments which are conserved in the human T10 gene promoter region as revealed by Dotter analysis, this 200 bp region likely contains the promoter of the mouse T10 gene. When the putative promoter region subsequently was subjected to transcription factor-binding site prediction with Consinspector, a total of seven transcription factor binding sites, including 5 SP1 binding sites, 1 glucocorticoid response element and 1 GA binding site were found. A polyadenylation signal (AATAAA) is observed at contig position 51125, which is 22 bp upstream of the 3' end of the mouse T10 cDNA sequence.

3.5.2 Mouse Htf9 locus: mouse Htf9c gene and mouse RanBP1 gene

The mouse Htf9c and RanBP1 genes were first isolated because of their association with a CpG island (Htf9) (Lavia, 1987). These two genes are transcribed in opposite directions and share coincident transcription start sites (Larvia, 1987). A Powblast search of the Genbank databases reveals that these two genes are located in the contig within a 15 kb genomic region (contig position 98k-113k).

Alignment of the cDNA sequence of mouse Htf9c gene (X56044, 2815 bp) with the contig sequence revealed 28 mismatches, 8 insertions and 4 deletions in the cDNA sequence relative to the genomic sequence. The alignment also revealed 11 exons for the mouse Htf9c gene. However, three mouse ESTs (AI037089, AI037312, AA049825) aligned perfectly with the genomic sequence but had a gap in the region where the first exon is located (the sequence of EST AA049825 is extended by interpretation of its trace file obtained from the Washington University ftp site). The gap (102484-102717) also is bound by consensus splicing site sequences, indicating the presence of an intron in this region. Thus, these data suggest that the mouse Htf9c gene transcript might be alternatively

spliced, with the major form possibly containing 12 exons. The total length of the 12 exons is 2570 bps and the exon and intron sizes are listed in Table 3.15

Table 3.15 Exon and intron sizes of mouse Htf9c gene

No.	Exon Range	Exon size (bp)	Intron Size (bp)
1	103290-102718	573	234
2	102483-101936	547	426
3	101515-101407	109	257
4	101149-100968	182	154
5	100813-100699	115	88
6	100610-100495	116	74
7	100420-100309	112	465
8	99843-99721	123	153
9	99567-99492	76	241
10	99250-99134	117	88
11	99045-98949	97	79
12	98869-98467	403	

The 12 exons of the Htf9c gene contain an ORF which codes for a putative protein of 602 amino acid residues starting from the second codon which codes for methionine. The second codon is tentatively chosen as the translation start site because the nucleotide sequence around the codon (CCCATGG) is more similar to the Kozak signal (GCCATGG) than the sequence around the first codon (GAGATGA) and the conceptually translated amino acid sequence between the first and the second methionine codons is not conserved in humans. The putative amino acid sequence of the mouse Htf9c gene is shown in Figure 3.26.

>Htf9c protein

```
MEDCGQDASAVPSSAAPLCQKEEAGPGPAAGPGTQPGLYSYIRDDLFTSEIFKLELQNVP
RHASFSDVRRFLGRFGLQSHKIKLFGQPPCAFVTFRSAAERDKALRVLHGALWKGCPLSV
RLARPKADPMARKRRQEGDSEPSVTQIADVVTPLWTVPYTEQLEQKRLECERVLQKLAKE
IGNTNRALLPWLLLQRQQHNKACCPLEGVKPSPQQTEYRNKCEFLVGVGVDGKDNTVGCR
LGKYKGGTCAVAAPFDTVHIPEATKQVVKAFQEFIRSTPYSAYDPETYTGHWKQLTVRTS
SRGQAMAIAYFHPQKLSSEEVAGLKASLVCHFMEGPGKASGVTSLYFVEEGQRKTPSQEG
LPLEHMAGDQCIQEDLLGLTFRISPHAFFQVNTPAAEVLYTVIQEWAQLDGGSTVLDVCC
GTGTIGLALAPKVKRVVGIELCQEAVEDARMNALTNELSNVEFHCGRAEDLVPGLVSRLS
SHQLVAVLDPPRAGLHSKVILAIRKAENIKRLLYVSCNPRAAMGNFVDLCRAPSNRVKGT
PFHPVKAVAVDLFPQTPHCEMLILFERMQQHPNGIEALEHQEFQTPRNLPDITPQETEIS
LS
```

Figure 3.26 Putative protein sequence of mouse Htf9c gene

The function of the Htf9c gene currently is unknown. A protein database search reveals that this protein is similar to a number of RNA methyltransferases, a result also seen with the human Htf9c protein sequence (see Section 3.4). A profile search reveals a RNA-binding domain (residues 52-125) and a SAM binding domain (residues 408-463). Psort predicts the mouse Htf9c protein to be cytoplasmic as it did for the human homologue of the Htf9c gene.

The RanBP1 gene is located on the opposite strand distal to the Htf9c gene. There are three Genbank entries for the mouse RanBP1 cDNA sequence(X56045, 1044 bp; X56046, 872 bp and L25255, 641 bp) and several ESTs. Alignment of these sequences with the contig sequence revealed that the mouse RanBP1 gene contains 6 exons. The exon and intron sizes of the mouse RanBP1 gene is listed in Table 3.16

Table 3.16 Exon and intron sizes of mouse RanBP1 gene

No.	Exon Range	Exon size (bp)	Intron Size (bp)
1	103267-103421	155	1122
2	104544-104680	137	1930
3	106611-106768	158	3327
4	110096-110224	129	785
5	111010-111075	66	700
6	111776-112185	410	

The six exons of mouse RanBP1 gene contains an open reading frame which encodes the RanBP1 protein of 203 amino acid residues in length (Figure 3.27). In the coding region, one discrepancy occurs between the genomic sequence and two of the cDNA sequences (X56045 and X56046), which results in an amino acid change at position 126 from threonine to alanine. One mismatch also is found between the genomic sequence with the other cDNA (L25255) which results in an amino acid change at position 33 from valine to leucine. It is not known whether the differences in the nucleic acid sequences are true microheterogeneities or due to errors in the earlier reported cDNA sequences.

>mRanBP1 protein sequence
MAAAKDSHEDHDTSTENADESNHDPQFEPIVSLPEQEIKTLEEDEEELFKMRAKLFRFAS
ENDLPEWKERGTGDVKLLKHKEKGTIRLLMRRDKTLKICANHYITPMMELKPNAGSDRAW
VWNTHADFADECPKPELLAIRFLNAENAQKFKTKFEECRKEIEEREKKGPGKNDNAEKVA
EKLEALSVREAREEAEEKSEEKQ

Figure 3.27 Protein sequence of mouse RanBP1 gene

Regulatory elements: It has been shown that both the mouse Htf9c and RanBP1 genes share a common promoter region (Lavia et al., 1987; Bressan et al., 1991). The 5' ends overlap 24 bp based on aligning the cDNA sequences of both genes. A number of transcription factor binding sites, including three SP1 binding sites and three E2F binding sites, important for transcription of both genes, are conserved in the promoter region of the human Htf9c gene and RanBP1 gene, as shown previously in Figure 3.16. A polyadenylation signal sequence (AATAAA) is present at contig position 98488 3' to the coding region of the mouse Htf9c gene and two polyadenylation signals (TATAAA and AATAAA) are found at contig positions 111983 and 112169, respectively, 3' to the coding region of the mouse RanBP1 gene.

3.5.3 A putative gene: mouse PutA

The region between the mouse T10 and Htf9c genes (70-98k) contains several EST homologies which align well to the genomic sequence. By aligning these ESTs with the genomic sequence and by incorporating results of gene prediction methods, including Genscan, Xgrail and sequence comparison between the mouse sequence with its syntenic human sequence, a putative gene with a total of 15 exons, of which 14 are coding exons, is predicted, as shown in Figure 3.24.

Six exons (exons 7-12) of the 15 exons are revealed by aligning two overlapping EST sequences (AA000671 and AA672501) to the genomic sequence. Another group of three exons (exons 13-15) are revealed by aligning a mouse EST (AA217940) to the genomic sequence. The 3' end 14 bp of EST AA672501 matches the 5' end of exon 13, suggesting that the two groups of exons belong to a single unknown gene. In the 5' region

of these nine exons, there are a number of mouse EST matches but none of them revealed exons with appropriate exon-intron boundaries. Within its human syntenic region, two additional exons are revealed by an EST match (Section 3.4). Although a database homology search with the mouse genomic sequence under default stringency conditions failed to find these two additional exons, a dotter alignment of the human EST sequence with the mouse genomic sequence does reveal significant sequence conservation for these syntenic regions, suggesting the presence of these two mouse exons. As with the human counterpart, the first exon of this mouse gene also is located in a major CpG island, and thus most likely represents the 5' end exon of the mouse gene. In addition, four and five exons were predicted by Genscan and Xgrail between these two 5' end exons and the downstream 9 exons. Sequence comparison (Geneligner) between human and mouse genomic sequence also predicted four potential exons. Although most of these exons predicted by the three methods overlap, only one exon (exon 4) was predicted with exactly the same 5' and 3' boundaries by all of the three methods. Without additional EST evidence, it is impossible to evaluate the correctness of these predictions. By comparing the four exons predicted for the human PutA gene (Section 3.4), four mouse exons (exons 3-6, see below) were tentatively identified. Among these four exons, exons 3 and 4 were predicted by all three methods but Xgrail split exon 3 into two exons. Exon 5 was predicted by Geneligner and Xgrail but not predicted by Genscan. Exon 6 was predicted by Geneligner and Genscan and by Xgrail as part of a larger exon. Thus, the putative mouse gene (referred to as PutA) with a total of 15 exons is predicted. As also observed with the human PutA gene, a number of overlapping mouse ESTs also aligned to the 3' end region of the mouse gene. Two mouse ESTs (AA032352 and W53243) contained more than 20 A's at their 3' end and since no other overlapping ESTs extended beyond their 3' end, the 3' end exon of the mouse PutA gene could be located to contig position 98019. However, unlike its human homologue for which two alternative splicing pathways were indicated by human EST data (Section 3.4), there were no EST matches which would indicate that part

of the 3' end exon might be alternatively spliced. However, alignment of the 3' end genomic sequence of the human and mouse genes shows that the splicing site sequences for the intron between human exon 15 and 16 were conserved in mouse, suggesting that an alternative splicing pathway also might be possible.

The 15 exons of PutA gene are 4006 bp long and contain an open reading frame which would encode a predicted 734 amino acid protein. The sizes of exons and introns are listed in Table 3.17 and the putative protein sequence is shown in Figure 3.28.

Table 3.17 Exon and intron sizes of mouse PutA gene

No.	Exon Range	Exon size (bp)	Intron Size (bp)
1	62763 - 62870	108	4403
2	67274 - 67408	135	276
3	67685 - 68273	589	413
4	68687 - 68846	160	2369
5	71214 - 71356	143	111
6	71468 - 71750	283	1703
7	73454-73651	198	178
8	73830-73931	102	774
9	74706-74804	99	4363
10	79168-79250	83	12741
11	91992-92092	101	223
12	92316-92422	107	1225
13	93648-93775	128	1408
14	95184-95297	114	1066
15	96364-98019	1656	

>mPutA.seq
MKRAIDAEVMDVGSGGDGQSEPPADDPFNFYGASLLSKGSFSKGRLLIDP
NCSGHSPRTARHAPAVRKFSPDLKLLKDVKISVSFTESCRSKDRKVLYTG
VERSTRPECGQLLSPVSGDVHACPFGGSVGNGVGLGGESADKKDEENELD
QEKRVEYAVLDELEDFTDNLELDEEGTGGFTAKAIVQRDRVDEEALNFSY
EDDFDNDVDALLEEGLCAPKKRRMEEKYGGDSDHPSDGETSVQPMMTKIK
TVLKSRGRPPTEPLPDGWIMTFHNSGVPVYLHRESRVVTWSRPYFLGTGS
IRKHDPPLSSIPCLHYKKMKDNEEREQNCDLAPSGEVSPVKPLGRSAELD
FPLEEPDSMGGDSGSMDEKDPLGAEAAAGALGQVKAKVEVCKDESVDLEE
FRNYLEKRFDFEQVTVKKFRTWAERRQFNREMKRKQAESERPILPANQKL
ITLSVQDAPTKKEFVINPNGKSEVCILHEYMQRVLKVRPVYNFFECENPS
EPFGASVTIDGVTYGSGTASSKKLAKNKAARATLEILIPDFVKQTSEEKP
KDSEELEYFNHISIEDSRVYELTSKAGLLSPYQILHECLKRNHGMGDTSI
KFEVVPGKNQKSEYVMACGKHTVRGWCKNKRVGKQLASQKILQLLHPHVK
NWGSLLRMYGRESSKMVKQETSDKSVIELQQYAKKNRPNLHILSKLQEEM
KRLAAEREETRKKPKMSIVASAQPGGEPLCTVDV

Figure 3.28 Predicted protein sequence of mouse PutA gene

The predicted protein sequence of the mouse PutA gene is very similar to its human homologue, with a GAP alignment identity of 96.2%. A protein database and profile search gave similar results for both the human and mouse PutA genes (Section 3.4), suggesting that both the human and mouse genes may encode nuclear proteins possibly involved in nucleic acid modification.

3.5.4 Repetitive elements in contig 4

Repeatmasker analysis reveals that contig 3 contains a total of 26.77% of repeat sequences (Figure 3.29 and Table 3.18). These repeats are dispersed throughout the contig except for those short intron regions of the Htf9c gene. Short interspersed elements (SINEs) are the predominant repeat elements in this contig, accounting for about 70% of the total repeat sequence.

Table 3.18 Repetitive elements in contig 4

	number of elements	length occupied	percentage of sequence
SINEs			
B1	73	8521 bp	7.21 %
B2	61	10851 bp	9.18 %
ID	3	181 bp	0.15 %
Other	16	2086 bp	1.76 %
LINEs			
L1	4	591 bp	0.50 %
LTR elements	28	5931 bp	5.02 %
MER	4	498 bp	0.42 %
Simple	37	2174 bp	1.80 %
Total repeats:	226	31694 bp	26.77 %

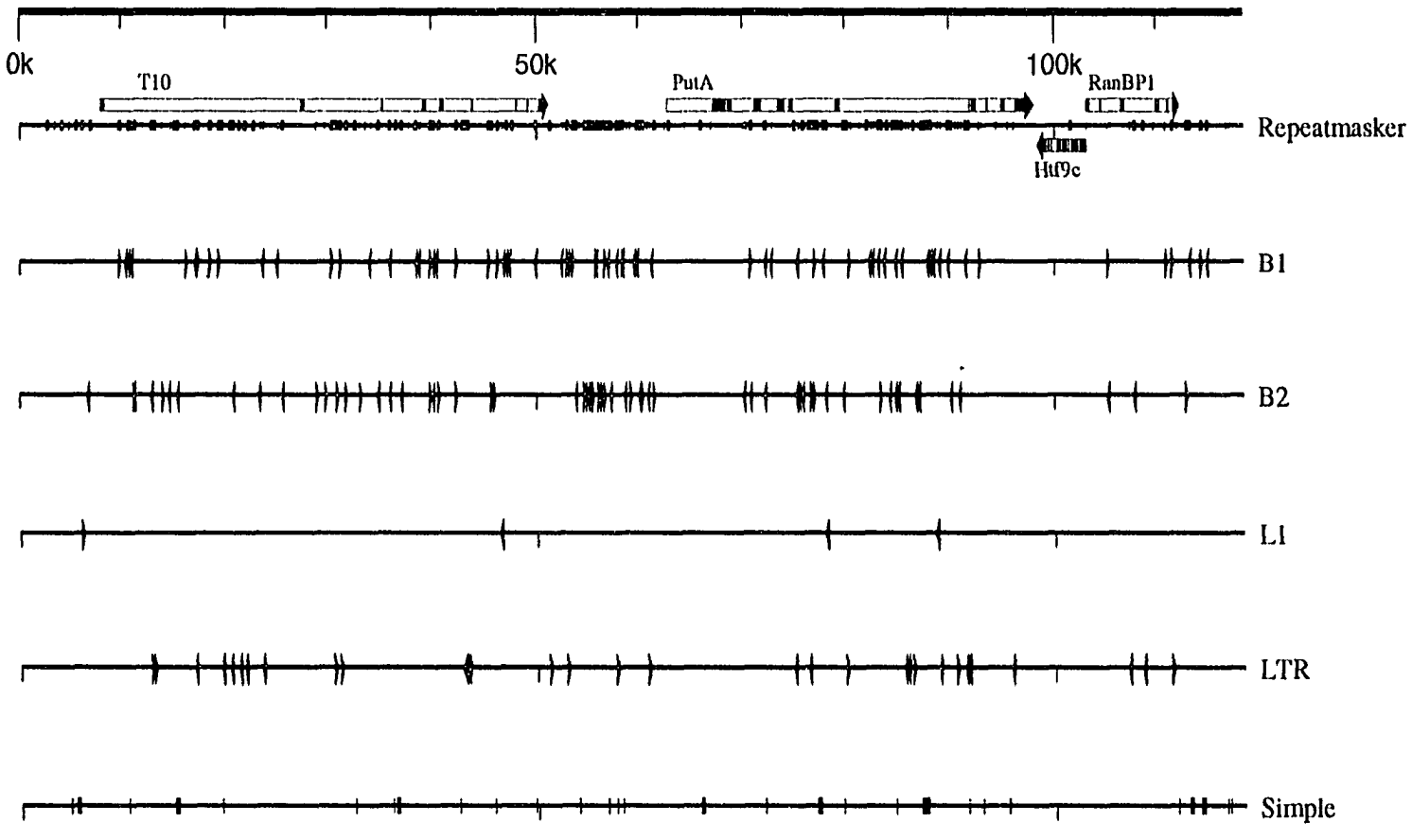


Figure 3.29 Repetitive elements in contig 4

3.6 Analysis of contig 5

Contig 5 is composed of a single BAC clone h173, with a total length of 90,832 bp. One known gene, the human Dok gene, and one putative gene, human Gene7, are found in this contig (Figure 3.30).

3.6.1 Human Dok gene

Human chronic myelogenous leukemia (CML) is a myeloproliferative disease that usually results from a chromosomal translocation between human chromosomes 9 and 22 which is known as the Philadelphia chromosome (Clarkson and Strife, 1993). This chromosomal translocation results in the c-abl protooncogene on chromosome 9 being fused to the 5' exons of the bcr (breakpoint cluster region) gene on chromosome 22 (DeKlein et al., 1982). The fusion usually generates the chimeric protein $p210^{bcr\text{-}abl}$ which causes CML (Groffen et al., 1987). It was found that relative to the non-receptor tyrosine kinase c-Abl, Bcr-Abl has increased tyrosine kinase activity which is directly associated with its transforming abilities (McWhirter and Wang, 1991). Abnormal protein-tyrosine kinase activity also has been found associated with cell transformation by oncogenes such as v-fps and v-Src (Ellis et al., 1990). This increased kinase activity has been proposed to be responsible for transformation-related phosphorylation of several proteins including p62, p190 and a GTPase-activating protein (GAP) (Ellis et al., 1990). In addition, protein p62 also was reported to be rapidly tyrosine-phosphorylated upon stimulation with growth factors such as insulin (Hosomi et al., 1994) and PDGF (Heidaran et al., 1992). Although these studies indicated that the p62 protein plays a central role in the signal transduction pathway downstream of the tyrosine kinases, the identity of the p62 was unknown until recent purification of the protein and molecular cloning of its corresponding cDNA from human (Carpino et al., 1997) and mouse (Yamanashi and Baltimore, 1997) and the genes were named Dok (downstream of tyrosine kinases). A powblast search of the Genbank databases revealed that the human and mouse Dok genes now have been sequenced as part

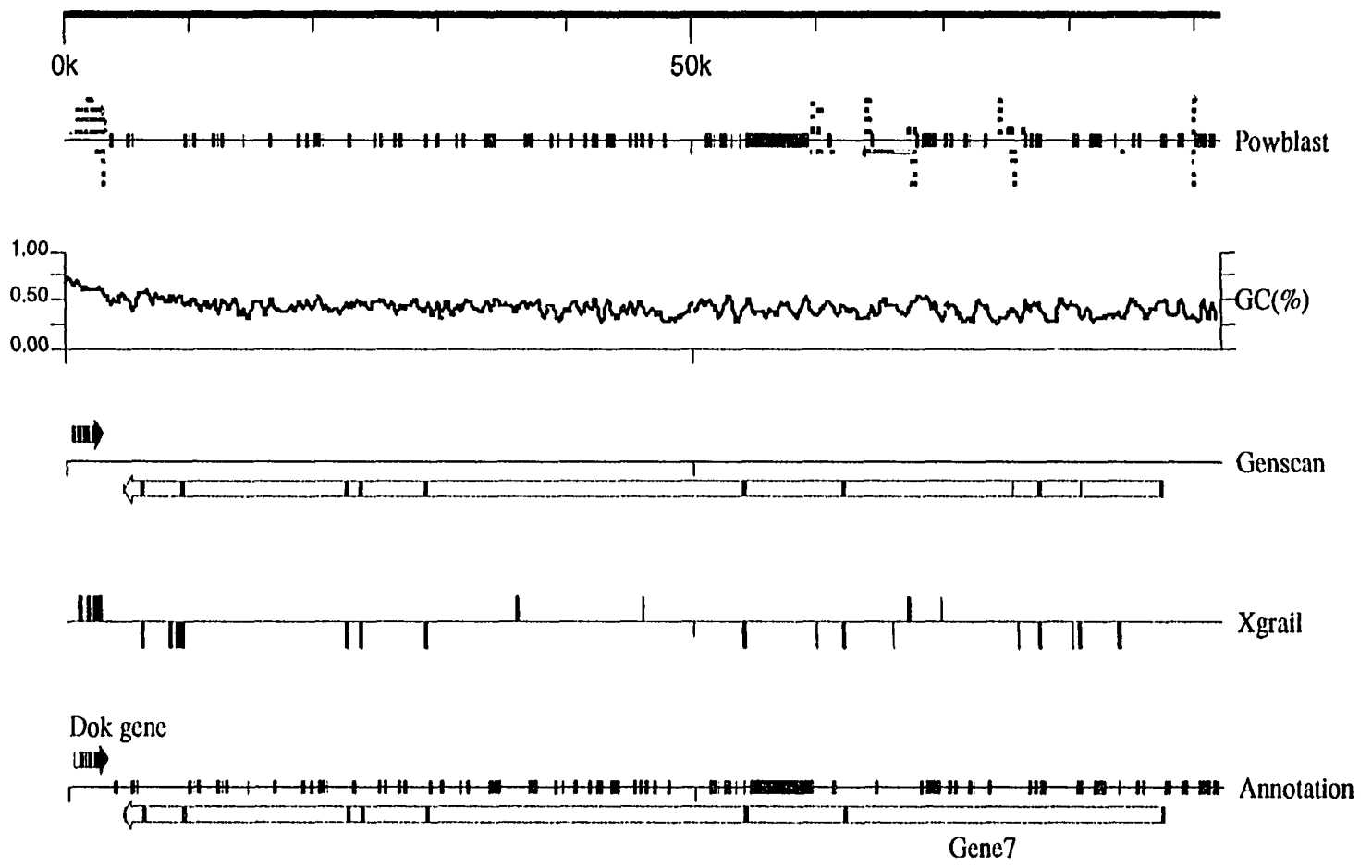


Figure 3.30 Features and annotation of contig 5 (see Figure 3.2b for legend)

of this dissertation research as they are located in this contig and contig 6, respectively.

Alignment of the human Dok cDNA (U70987) with the genomic sequence shows that the Dok gene contains 5 exons spanning about 3 kb. No discrepancies were observed between the genomic and the 1873 bp cDNA sequence. The sizes of exons and introns of human Dok gene are listed in Table 3.19. The 3' end of exon 5, which is 29 bp longer than the human Dok cDNA, is based on the sequences of two human ESTs (AA142943 and D20997).

Table 3.19 Exon and intron size of human Dok gene

no.	contig position	Exon size (bp)	Intron size (bp)
1	564-645	82	339
2	985-1284	300	121
3	1406-1499	94	225
4	1725-1909	185	229
5	2139-3379	1241	

The human Dok cDNA has one open reading frame that would code for a 481 amino acid long protein (Figure 3.31; Carpino et al., 1997). The predicted translation start site likely is authentic since this start site sequence fits the Kozak consensus (GCCATGG) and there is a high level of amino acid sequence conservation between human and mouse downstream of the AUG start codon.

```
>1      MDGAVMEGPLFLQSQRFGTKRWRKTWAVLYPASPHGVARLEFFDHKGSSSGGGRGSSRRL
                      pleckstrin homology
>61     DCKVIRLAECVSVAPVTVETPPEPGATAFRLDTAQRSHLLAADAPSSAAWVQTLCRNAFP

>121    KGSWTLAPTDNPPKLSALEMLENSLYSPTWEGSQFWVTVQRTEAAERCGLHGSYVLRVEA

>181    ERLTLLTVGAQSQILEPLLSWPYTLLRRYGRDKVMFSFEAGRRCPSGPGTFTFQTAQGND
                                    Dok Homology
>241    IFQAVETAIHRQKAQGKAGQGHDVLRADSHEGEVAEGKLPSPPGPQELLDSPPALYAEPL

>301    DSLRIAPCPSQDSLYSDPLDSTSAQAGEGVQRKKPLYWDLYEHAQQQLLKAKLTDPKEDP

>361    IYDEPEGLAPVPPQGLYDLPREPKDAWWCQARVKEEGYELPYNPATDDYAVPPPRSTKPL

>421    LAPKPQGPAFPEPGTATGSGIKSHNSALYSQVQKSGASGSWDCGLSRVGTDKTGVKSEGS

>481    T
```

Figure 3.31 Protein sequence of human Dok gene
The 15 tyrosines conserved in mouse Dok protein are highlighted.

The exact function of both the human and mouse $p62^{Dok}$ protein currently is unknown, although previous studies, as described above, and the results presented below, indicate that they are involved in a signal transduction pathway. A protein database search shows that $p62^{Dok}$ shares partial similarity with signaling adapter SNT-1 (AF036718) and SNT-2 (AF036717) of human FGFR (fibroblast growth factor receptor). Both SNT-1 and SNT-2 were identified as specific targets of differentiation factor-induced tyrosine kinase activity in neuronal cells (Rabin et al., 1993; Xu et al., 1998) and they were believed to be direct protein links between FGFR-1 and multiple downstream pathways (Xu et al., 1998).

Carpino et al. (1997) noted that there is a putative pleckstrin homology (PH) domain in the N-terminus of the human $p62^{Dok}$ protein and a profile search confirmed the potential pleckstrin homology (PH) domain between amino acid positions 3-119 at the N-terminus. The PH domain that occurs in a variety of proteins involved in signaling and/or cytoskeletal organization (Musacchio et al., 1993), may mediate protein-protein interactions or interact with cellular membranes (Lemmon et al., 1996). A proline-rich region (279-292), also indicated by the profile search, may mediate interactions with other motifs in potential partner proteins (Yamanashi and Baltimore, 1997). Since the human $p62^{Dok}$ contains 15 tyrosines, all of which are conserved in its mouse homologue (Figure 3.31), it is likely some of these tyrosine residues may interact with SH2 domain-containing signaling proteins (Carpino et al., 1997).

Recently, Di Cristofano et al. (1998) characterized a protein from $p210^{bcr\text{-}abl}$ expressing cells that is 35% identical to the human $p62^{Dok}$ and named it $p56^{Dok\text{-}2}$. Their searches of the EST database found at least four more ESTs encoding proteins which had homologies with the $p62^{dok}$ and $p56^{Dok\text{-}2}$. All six proteins share a common homologous region of about 50 amino acids which they named Dok homology (DKH) sequence motif. Based on the Dok homology, they defined a new family of Dok proteins and speculated that the Dok association regulates GAP activity toward Ras and serves as a mediator of Bcr-Abl signaling.

In addition to the Dok cDNA sequence, a Powblast search of the Genbank non-redundant database also revealed a cDNA sequence (AF035299) which was homologous to the genomic sequence in the region of the Dok gene. An alignment of the cDNA and genomic sequences showed that the cDNA sequence perfectly matched exons 2-5 of the Dok gene. However, the 5' 146 bp of the cDNA sequence did not match any segment of the genomic sequence. A further Blast search of the Genbank Htgs database (high throughput genomic sequences) showed that the 146 bp sequence matched a segment of the genomic sequence of a human clone (AC005033) which was at phase one in four unordered pieces. Dotter analysis of the two genomic sequences, i.e. the sequence of the human clone AC005033 and that determined in this dissertation research (this contig), showed that the former overlapped the entire genomic region of this contig and that the segment which matched the 146 bp cDNA sequence was located in a region about 5 kb upstream of the Dok gene (Figure 3.32) but not in the genomic sequence determined in this present research. There are a deletion and a mismatch in the cDNA sequence relative to the sequence of the Genbank clone and the ORF of exons 2-5 of the Dok gene extends in the 5' direction through the 146 bp sequence for both the cDNA sequence and the genomic sequence. In addition, there are no in-frame methionine (AUG) or stop codons in this 146 bp sequence. Although it is not known whether the putative exon represents the 5' end coding exon of the alternative form of the Dok gene, a major CpG island spans the putative exon, which suggests that the putative exon may be close to or at the 5' end of the coding region of an alternative form of the Dok gene. The presence of this cDNA sequence suggests that the human Dok gene may have two alternative forms and possibly have two promoters, where one promoter is located in this contig and the other is located about 5 kb upstream. Transcription from these two promoters would result in two distinct transcripts, which most likely would correspond to two distinct protein products.

As shown in Figure 3.32 and described in Section 3.7, the human homologue of the mouse Lor2 gene is located upstream of the human Dok gene and transcription of the

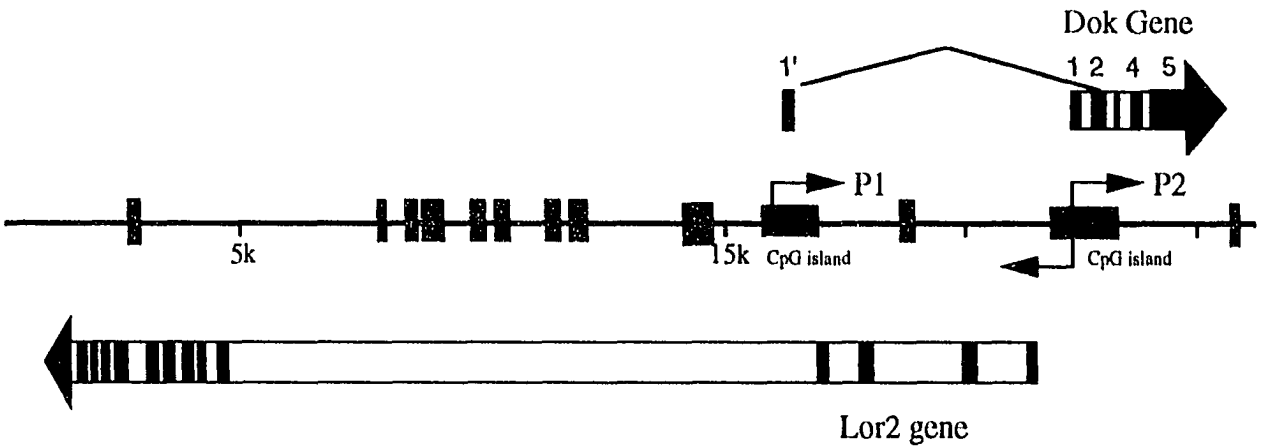


Figure 3.32 Diagram showing the structure of human Dok and Lor2 genes

alternative form of the Dok gene would overlap the human Lor2 gene. About 800 bp separate the 5' ends of the regular form of Dok gene and the human Lor2 gene. Promoter prediction by NNPP and TSSW predicted multiple promoters in this 800 bp region and all the predicted promoters are located within 500 bp from the 5' end of Dok gene. Since a similar result was obtained for the mouse Dok and Lor2 genes, it is most likely that the human Lor2 gene and the regular form of human Dok gene may be under the control of a bidirectional promoter located close to the 5' end of the Dok gene.

Bidirectional promoters have been observed for a number of mammalian genes and in most cases, they are shared by genes encoding either related products or proteins required in a common metabolic or regulatory pathway (Guarguaglini et al., 1997). The possible involvement of the Dok protein in a signal transduction pathway and the Lor2 gene in cell adhesion and other extracellular processes seems to suggest that the Lor2 gene serves as an extracellular or cell surface component of a signal transduction pathway.

Overall, the $p62^{Dok}$ sequence has features of a signaling molecule. Its association with GTPase-activating protein (GAP) (Carpino et al., 1997) and phosphorylation after cell transformation strongly suggests that it is a component of a signal transduction pathway and may play an important role in the oncogenesis of CML.

Regulatory elements: Even though this contig contains 564 bps 5' to the first exon of human Dok gene, no TATA box was found in the region. However, a major CpG island is located at contig positions 51 through 1416 that encompasses the first two exons of the Dok gene. The NNPP promoter prediction program indicated two putative promoters at positions 133-183 and 177-227. Promoter prediction by TSSW also predicted two promoters but at positions 178 and 549. Thus, both methods predicted the presence of promoters in the 400 bp region upstream of the first exon of the Dok gene. Within this region, Consinspector predicted 3 ATF, 3 SP1 and several other transcription factor binding sites as discussed in Section 4.3.

Although no polyadelynation consensus (AATAAA) could be found near the 3' end, a potential polyA signal of ATTAAA is observed 22 bp upstream of the 3' end of the exon 5 of the Dok gene at contig position 3358.

3.6.2 A putative gene: Human Gene7

A Powblast search of Genbank databases in the remaining region (4-92k) of this contig, did not reveal any sequences with extensive homology to any known genes. However, Genscan did predict a gene in this region that was in the complementary strand and contained 12 exons. Xgrail also predicted several exons, 15 of which also were in the complementary strand (see Figure 3.30). Since a gene, mouse Gene7, was described in the mouse syntenic region (Section 3.7), a dotter analysis was performed to compare the cDNA sequence of the mouse Gene7 (Jiang, 1998) and the human genomic sequence. The results show that all nine exons of the mouse Gene7 are conserved in human and thus, the region likely contains the human Gene7 equivalent which also has 9 exons. As in mouse, the splicing site sequence of 5' end of intron 7 of this gene is GC rather than the usual GT. A Geneligner comparison of the human and mouse genomic sequences predicted a gene with 7 exons on the complementary strand as well. These 7 exons match 7 exons of human Gene7 exactly except for the 3' end of exon 7 which contains the non-typical splicing site sequence (GC). The 9 exons of the human Gene7 would result in a 1876 bp mRNA and code for a putative protein of 526 amino acids. The exon and intron sizes of the human Gene7 are listed in Table 3.20. Alignment of the putative human Gene7 and mouse Gene7 protein sequences shows an identity of 73% and similarity of 78% (Figure 3.33).

Table 3.20 Exon and intron size of human Gene7

no.	contig position	Exon size(bp)	Intron size(bp)
1	87202-86911	292	24886
2	62024-61839	186	7742
3	54096-53928	169	25207
4	28720-28547	174	5055
5	23491-23329	163	876
6	22452-22311	142	13012
7	9298-9092	207	2969
8	6122-5982	141	1276
9	4705-4281	425	

```
hGene7   1 MHPGRTTGKGPSTHTQIDQQPPRLLIVHIALPSWADICTNLCEALQNFFS  50
mGene7   1 MNRRKTTSRGTSAAMKISHQPPRLLIVNIAVPSWVDICPNLCEALQNFFS  50

        51 LACSLMGPSRMSLFSLYMVQDQHECILPFVQVKGNFARLQTCISELRMLQ 100
        51 IACSLMGPSRMSLFSLYTVQNQHECVLPFVQVRGNFIRLQACISELRMLQ 100

       101 REGCFRSQGASLRLAVEDGLQQFKQYSRHVTTRAALTYTSLEITILTSQP 150
       101 VEGCHRPPHALLPLAIEDGLQQFKQYSSHMASSAAQPWTSLEITVLTSRP 150

       151 GKEVVKQLEEGLKDTDLARVRRFQVVEVTKGILEHVDSASPVEDTSNDES 200
       151 GKEVVKELEEGLKDINLLSVRRLQVAEVTKGIQERSDSPSPTEEPSNDES 200

       201 SILGTDIDLQTIDNDIVSMEIFFKAWLHNSGTDQEQIHLLLSSQCFSNIS 250
       201 SILEADIVLETLDNDVVSMEVFFKAWLHNSETDQENIHLLLTPQSLPPPS 250

       251 RPRDNPMCLKCDLQERLLCPSLLAGTADGSLRMDDPKGDFITLYQMASQS 300
       251 RAKDHPICLKCDLQERFLSPSLLPGTADGVSRIDDPKGDISTLYQMASLA 300

       301 SASHYKLQVIKALKSSGLCESLTYGLPFILRPTSCWQLDWDELETNQQHF 350
       301 SASPDKLQVVKALKSSGICESLTYGLPFILRPTSCWQLDWDELETNQQHF 350

       351 HALCHSLLKREWLLLAKGEPPGPGHSQRIPASTFYVIMPSHSLTLLVKAV 400
       351 HALCHCLLKRDWLLLARGEPLIPKHNQSLPACSFYVITPSHSLTLLVKLV 400

       401 ATRELMLPSTFPLLPEDPHDDSLKNVESMLDSLELEPTYNPLHVQSHLYS 450
       401 ATRELMLPGFFPLLSEDPPEDSLKIIESTLDSLDLGLTYNPLHVGSHLYS 450

       451 HLSSIYAKPQGRLHPHWESRAPRKTGQLQTNRARATVAPLPMTPVPGRAS 500
       451 HLSSAHAKPQGRLYTSCASRGLRKGGQLQTNRVRAAVVPLPVAPAPRRAL 500

       501 KMPAASKSSSDAFFLPSEWEK.DPSRP... 526
       501 KMTAASKASS.AAFLPSDSEEGEEERPSHT 529
```

Figure 3.33 GAP alignment of human and mouse Gene7 protein

Although the function of Gene7 is unknown, a protein database homology search with the putative protein sequence shows that amino acid residues 100-230 are moderately homologous to a putative *C. elegans* protein (AL021498) of unknown function, the chain length determinant proteins of *Salmonella typhimurium* (Q04866) and *E. coli* (AF011916) and a *Schizosaccharomyces pombe* hypothetical protein (AL023860). A profile search did not reveal any homologous protein domain in the profile database, but Psort predicted the protein to be localized in the nucleus.

A Powblast search of the Genbank databases also revealed that the region between contig positions 25736-29439 within human Gene7 has a high degree of similarity with a cDNA named BH30 (Y12839). This cDNA is 3786 bp long and includes 11 A's at its 3' end. Alignment of this transcript with genomic sequence reveals that 3697 bp (78-3775) of the cDNA match the genomic contig sequence perfectly except for two deletions in the cDNA sequence. The 5' end 77 bp of the cDNA is not found in the contig sequence and a nucleic acid database search shows that the 77 bp perfectly matches a portion of the 5' long terminal repeat of type I HIV virus (HIVU89860). Since the BH30 cDNA was isolated from HIV-infected cells (Y12839), it is likely that the BH30 cDNA resulted from the insertion of a HIV genome to the genomic position in those HIV-infected cells. The insertion site sequence is a short simple repeat (ACATAT-insert-ATATAT). Examination of the nucleic acid sequence of the BH30 cDNA shows that it contains two Alu repeats and a MIR repeat, with several ORFs of up to 120 amino acids. The consensus AATAAA polyadenylation signal sequence is located toward the 3' end of the cDNA. The transcribed region of BH30 spans part of intron 3 and intron 4, and the entire exon 4 of Gene7.

Regulatory elements: There are no TATA box elements found in the vicinity of the 5' end exon of the putative human gene Gene7. Promoter prediction in the mouse Gene7 region predicted the presence of a promoter which is about 9.5 kb upstream from the first coding exon of the mouse Gene7 and located in a relatively GC-rich region (Section 3.7). Since the first coding exon of human Gene7 is only about 3.5 kb from the end of the contig

and a GC-rich region is not present in the 3.5 kb region, it is likely that the promoter for the human Gene7 would be located out of this contig. However, the ATTAAA sequence located near the 3' end of the gene at contig position 4286 which also is conserved in the mouse, may represent the Gene7 polyadenylation signal in both human and mouse.

3.6.3 Human TorsinB pseudogene and other pseudogenes

A Powblast search revealed several homologous human ESTs and the TorsinB gene within the first intron of the human Gene7 between contig positions 63800-67800. A dotter plot of the TorsinB cDNA sequence and the contig sequence is shown in Fig. 3.34.

The TorsinB gene was one of the several genes isolated from the early-onset dystonia critical region in human chromosome 9 (Ozelius et al., 1997). Although the genomic structure of the TorsinB gene is unknown, a detailed comparison of the genomic sequence in this contig and the cDNA sequence of the TorsinB gene reveals that the cDNA contains an Alu repeat (960-1080) which is not present in the genomic sequence while the genomic sequence contains a L1 repeat at positions 64290-64560 and another sequence segment at positions 66300-67500 which are not present in the cDNA sequence. Examination of the genomic region homologous to the 803 bp open reading frame of the TorsinB cDNA (AF007872) also reveals multiple stop codons in all reading frames. These observations are consistent with the idea that this region of human chromosome 22 contains a pseudogene of the human TorsinB gene whose true location is human chromosome 9.

The Powblast search also revealed that the contig region 74414-75820, also within the first intron of the Gene7 gene, is similar to a number of ESTs and segments of three human BAC clones. One of the human BAC clones is from human chromosome 16p13 (U91322). Detailed examination showed that the 1.4 kb contig sequence is co-linear with a segment of similar size from that BAC clone (bacA), with an identity of 88%. A similar result also was obtained when the contig sequence was compared with that of a second

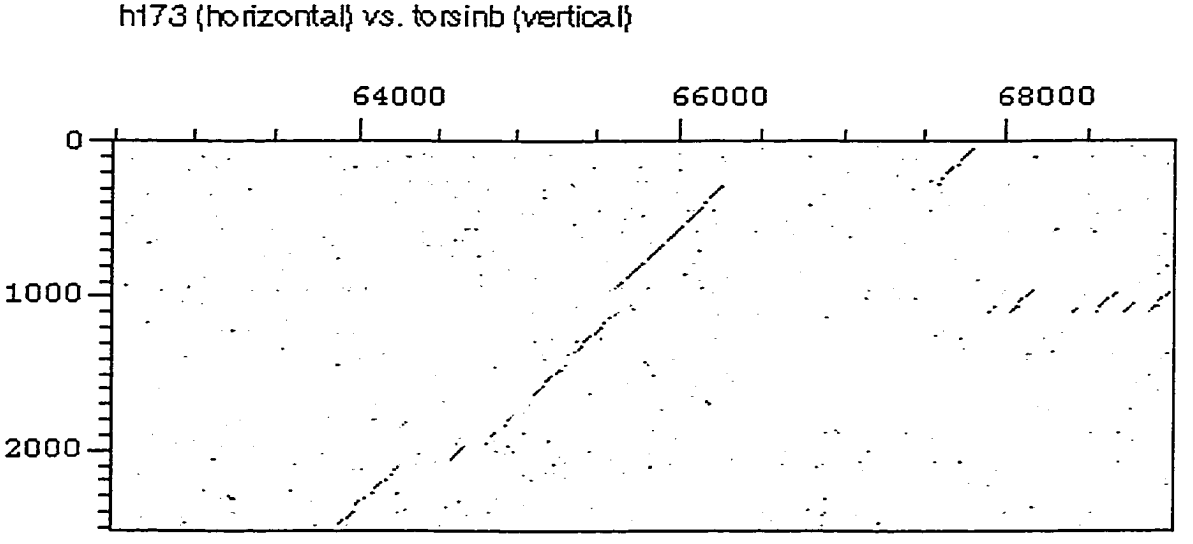


Figure 3.34 Dotter plot between contig 5 sequence with TorsinB cDNA sequence

human BAC clone which is from human chromosome 7q21 (bacB, AC000059). However, when the contig sequence was compared to that of the third human BAC clone, which was from human chromosome 17 (bacC, AC005838), the 1.4 kb contig sequence was similar to six segments of the BAC sequence from chromosome 17. Comparison of sequences of ESTs homologous to the 1.4 kb contig sequence showed multiple substitutions, insertions and deletions, suggesting that these ESTs did not result from transcription from the genomic region cloned in this contig. Examination of the 1.4 kb contig sequence also revealed that it was bounded by two 10 bp direct repeats of AAAGTATTCT and that multiple stop codons occur in all six frames. A protein database search with all six possible conceptual translation products of this region (with multiple stop codons) revealed modest similarity with a yeast hypothetical 23.1 kd protein (P38962) and the putative translation product of the *C.elegans* C34D4.4 gene (U58755). These observations suggested that the 1.4 kb contig region may contain a pseudogene. To further study this possibility, the sequence of the human BAC from chromosome 17 (bacC) was retrieved from Genbank and further analysis of the sequence was performed. A Powblast database search against the Genbank non-redundant and dbEST databases revealed multiple ESTs in the sequence regions of bacC which matched the 1.4 kb contig sequence. Based on these EST data, Geneligner built a gene model with seven exons and six of these exons overlapped the six segments of bacC which were homologous to the 1.4 kb contig sequence (Figure 3.35). Dotter analysis (Figure 3.36) showed that the sequence segments from chromosomes 16, 7 and this contig have high sequence similarity to the predicted cDNA sequence of the putative gene in chromosome 17. Both the sequence segments from chromosomes 16 and 7 have poly-A or A-rich segments at their 3' ends but only the sequence segment from chromosome 16 contains two 9 bp direct repeats at its 5' and 3' ends and a blastn search showed that the sequence segment in chromosome 7 contains a Alu repeat (see Figure 3.36). A protein database search showed that the protein sequence for the gene predicted in human chromosome 17 was 58% similar to the putative protein of C34D4.4 in *C. elegans*

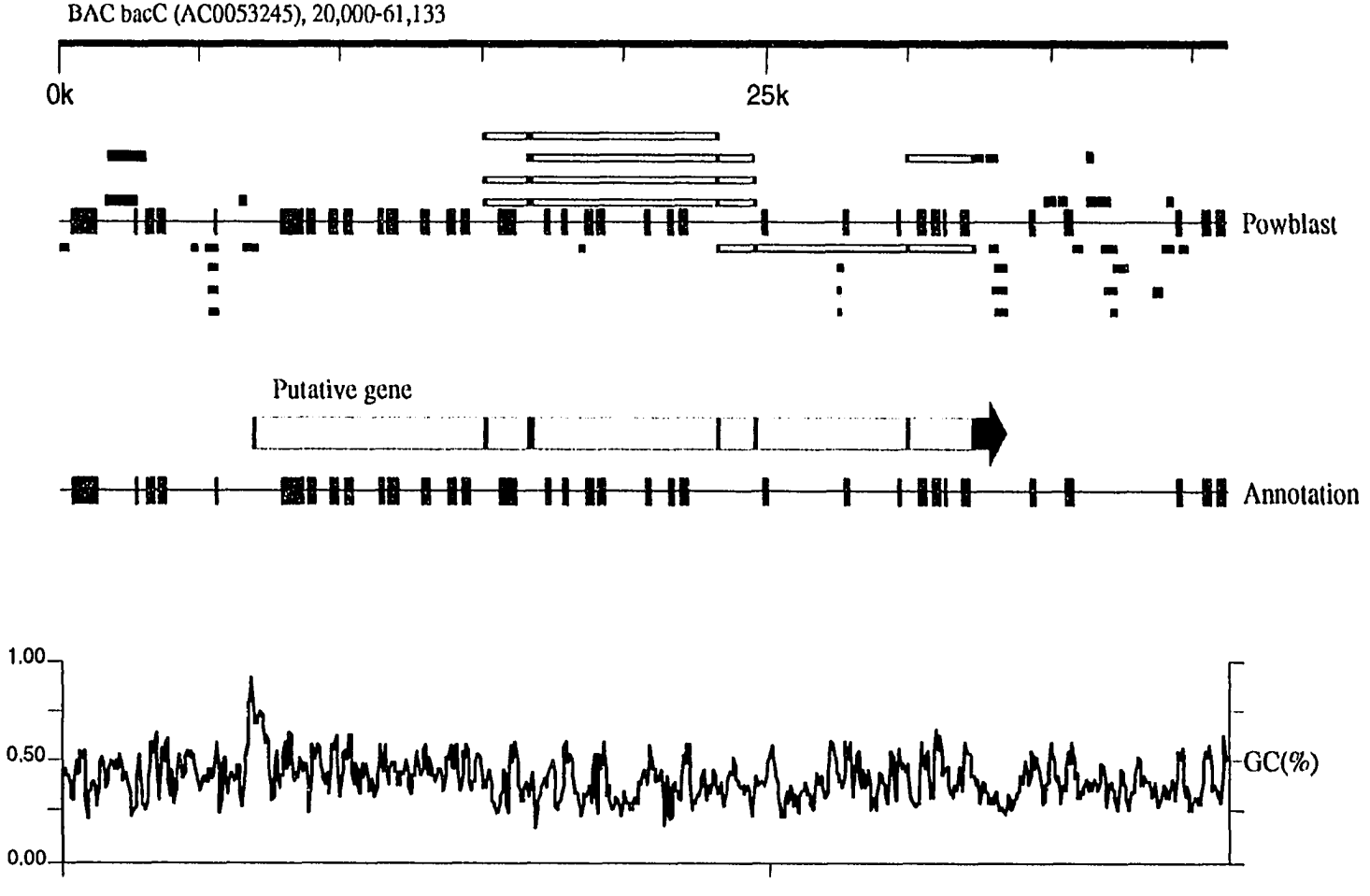


Figure 3.35 A gene model built from EST data by Geneligner

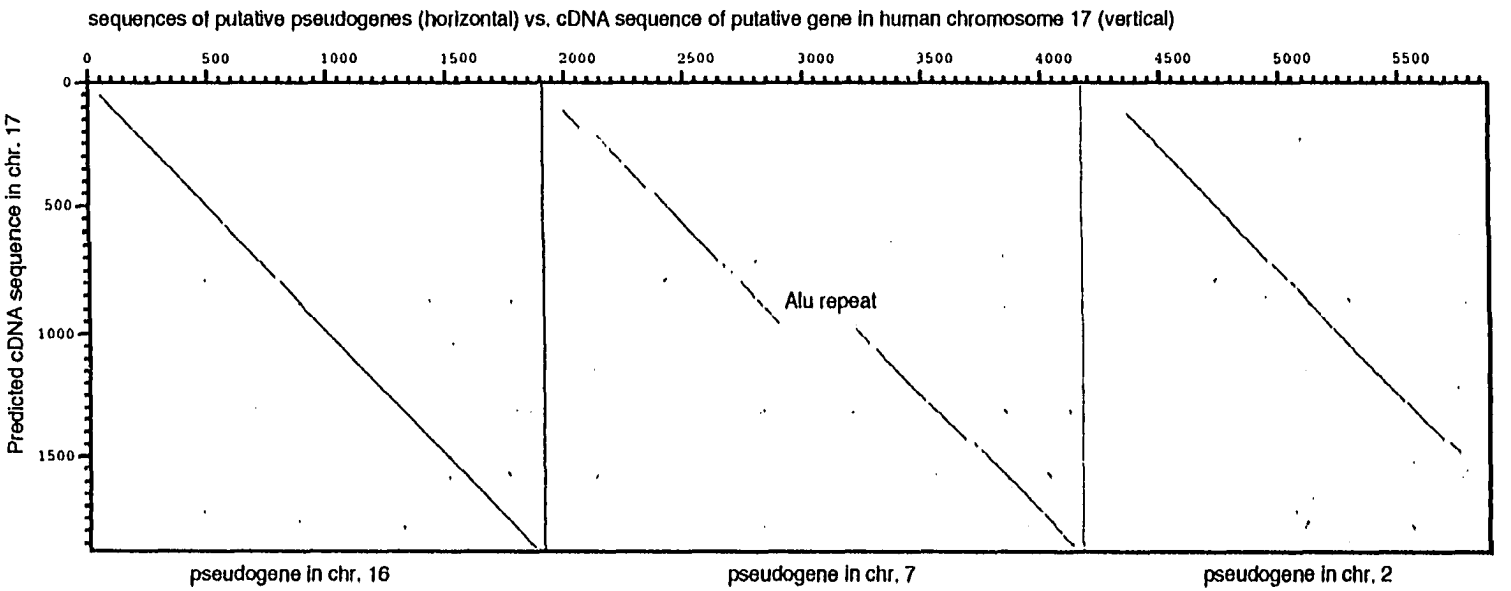


Figure 3.36 Dotter plot of sequences of three putative pseudogenes with the cDNA sequence of the putative gene in chromosome 17

and 54% similar to the yeast hypothetical 23.1 kd protein. These results strongly suggest that this contig region of chromosome 22 and the two other regions of chromosomes 16 and 7 contain processed pseudogenes which most likely were derived from transcripts of a gene from human chromosome 17.

3.6.4 Repetitive elements in contig 5

A Repeatmasker analysis revealed that 39429 bp or 42.87% of contig 5 is repeat sequences (Table 3.21 and Figure 3.37) where Alu repeats and L1 repeats are in about equal abundance and account for 16.00% and 17.89% of the sequence, respectively. The Alu repeats are more randomly dispersed throughout this contig than are the L1 repeats. Also, while the Dok gene region has virtually no repeats detected by Repeatmasker, the region of the putative Gene7 gene is rich in both Alu and L1 repeats. However, the observation from previous analysis of contigs 1 through 4 indicating that the L1 repeats are preferentially oriented in the opposite direction of gene transcription is not maintained. Instead, the L1 repeats which are located in the 5' 40 kb region of Gene7 mostly are oriented in the same direction of Gene7 transcription and those located in the 3' 40 kb region largely are oriented in the opposite direction of Gene7 transcription (see Figure 3.37). To investigate whether the preference of orientation of L1 repeats within genes is a general trend or merely an artifact from small sampling size, L1 repeats from the entire DiGeorge region (~1.5 Mb) were analyzed. A total of 11 genes which contain L1 repeats in their intron regions were investigated and their orientations relative to gene transcription are summarized in Table 3.22. From the table, the preference of L1 repeat orientation likely exists at least for the DGCR region although exceptions do occur.

Table 3.21 Content of repetitive elements in contig 5

	number of elements	length occupied	percentage of sequence
SINEs			
ALUs	56	14697 bp	16.00 %
MIRs	8	1277 bp	1.39 %
LINEs			
L1	28	16436 bp	17.89 %
L2	8	1727 bp	1.88 %
LTR elements			
MER elements	11	2628 bp	2.86 %
Unclassified	1	36 bp	0.04 %
Simple repeats	12	685 bp	0.70 %
Total interspersed repeats:		39429 bp	42.87 %

Table 3.22 Orientation of L1 repeats within genes in the DGCR region

Gene name	forward no.*	reverse no.	percentage of reverse
DGCR5	3	2	40 %
LAN	7	15	68 %
CLTCL	8	14	63 %
HIRA	5	17	77 %
UFD1L	3	2	40 %
Porc-P1	0	6	100 %
Putative gene	0	5	100 %
COMT	0	3	100 %
ARVCF	0	4	100 %
T10	0	3	100 %
PutA	0	3	100 %
Total	26	74	74 %

*forward when the repeat is oriented in the same direction as gene transcription; reverse when in opposite direction.

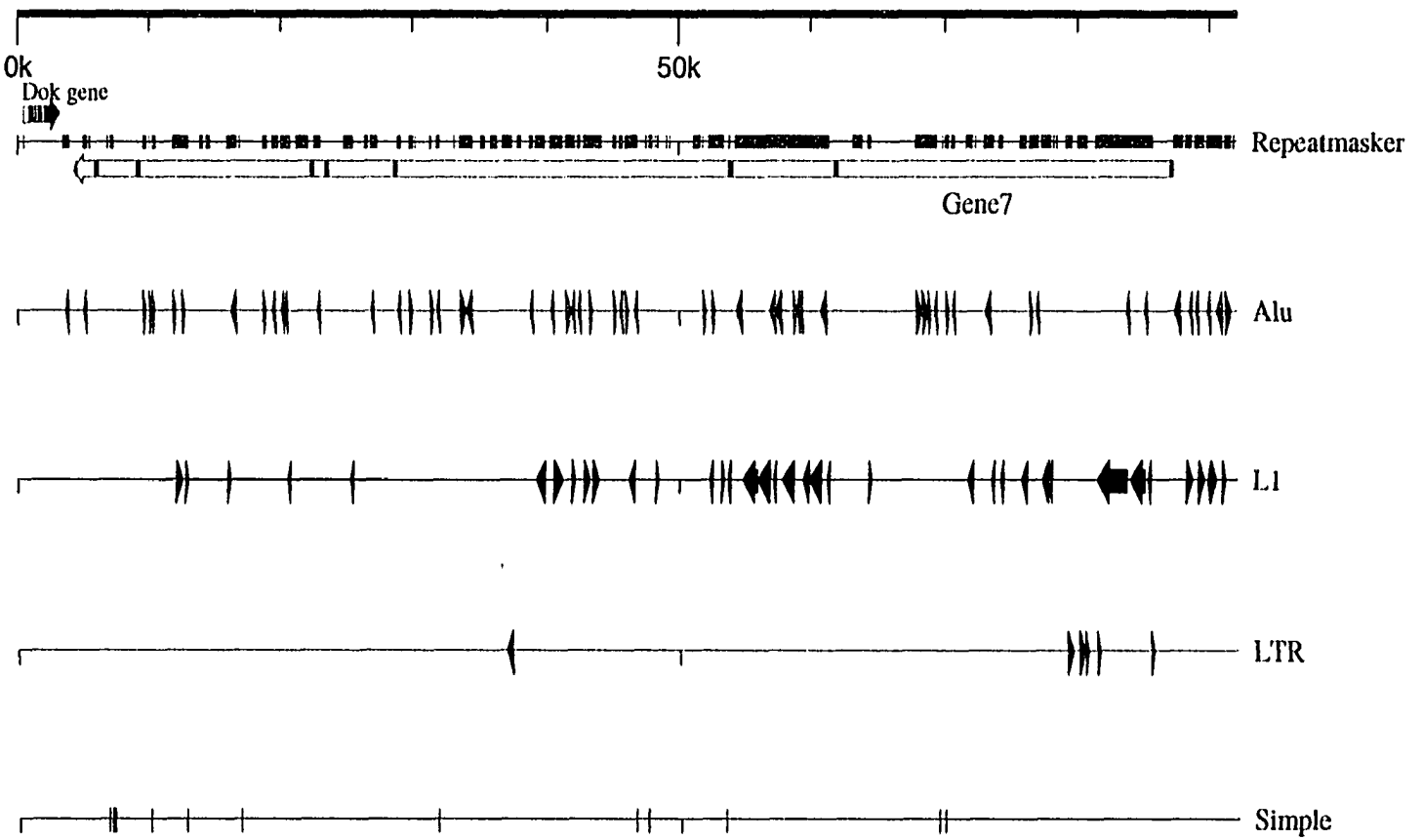


Figure 3.37 Repetitive elements in contig 5

3.7 Analysis of contig 6

This contig is composed of one mouse BAC clone b245c12 which contains an insert of 162, 750 bp. The mouse genomic region of this contig is syntenic to a region of human chromosome 2, part of which (human BAC h173) has been sequenced and was described above in section 3.6. As expected, the mouse Dok and Gene7 genes are found in this contig. Two additional genes, the mouse semaphorin M and Lor2 genes also are found in this contig (Figure 3.38).

3.7.1 Mouse Dok gene

As described in section 3.6, the human Dok gene cDNA was isolated during the search for the target of the $p210^{bcr\text{-}abl}$ kinase fusion protein that resulted from the chromosomal translocation between human chromosomes 9 and 22. The mouse homologue of the human Dok gene cDNA also recently was isolated from a v-abl-transformed murine precursor B cell line (Yamanashi and Baltimore, 1997). An alignment of the mouse Dok cDNA (U78818) with the genomic sequence of this contig reveals that like the human Dok gene, the mouse Dok gene also contains five exons spanning a genomic region of about 3 kb (contig positions 6-9k). This alignment also shows a total of 7 mismatches between the two sequences, including six substitutions and one insertion in the mouse cDNA. Five substitutions are located in the coding region, resulting in three amino acid differences from the published sequence. Several mouse ESTs (AA19595, AA832762 and AI035509) also were homologous to the contig sequence near the 3' end of the mouse Dok gene but extended up to 20 bp beyond the end point of the mouse Dok cDNA (U78818) in the genomic sequence. Considering that a putative polyA signal (ATTAAA) is located within the extended 20 bp, and one EST (AI035509) contains a polyA tail, the 3' terminal exon is expected to be 20 bp longer than that revealed by the published mouse cDNA. The exon and intron sizes of mouse Dok gene are listed in Table 3.23.

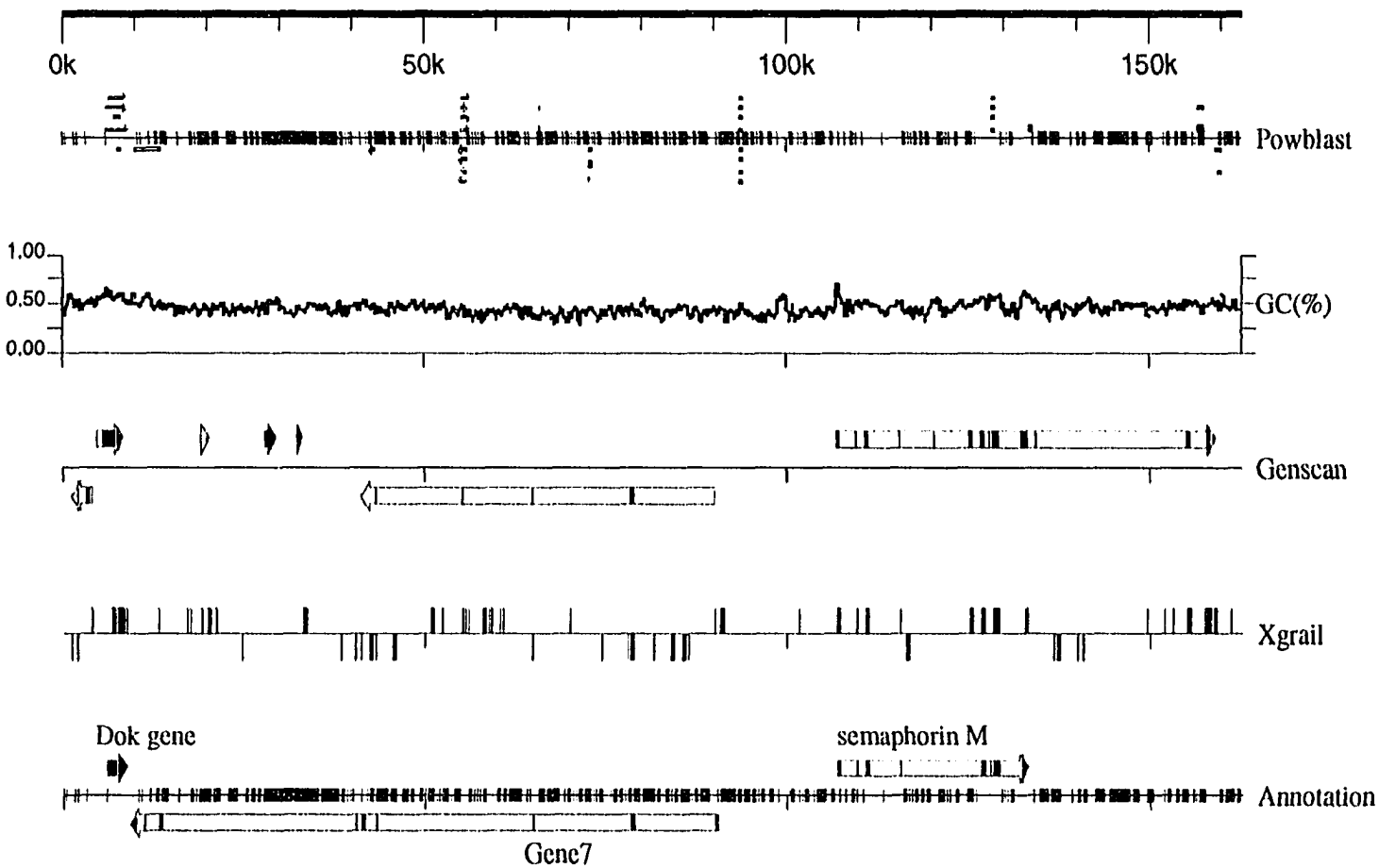


Figure 3.38 Features and annotation of contig 6

Table 3.23 Exon and intron size of mouse Dok gene

no.	contig position	Exon size(bp)	Intron size(bp)
1	6381-6446	66	304
2	6751-7050	300	138
3	7189-7282	94	119
4	7402-7586	185	170
5	7757-8879	1123	

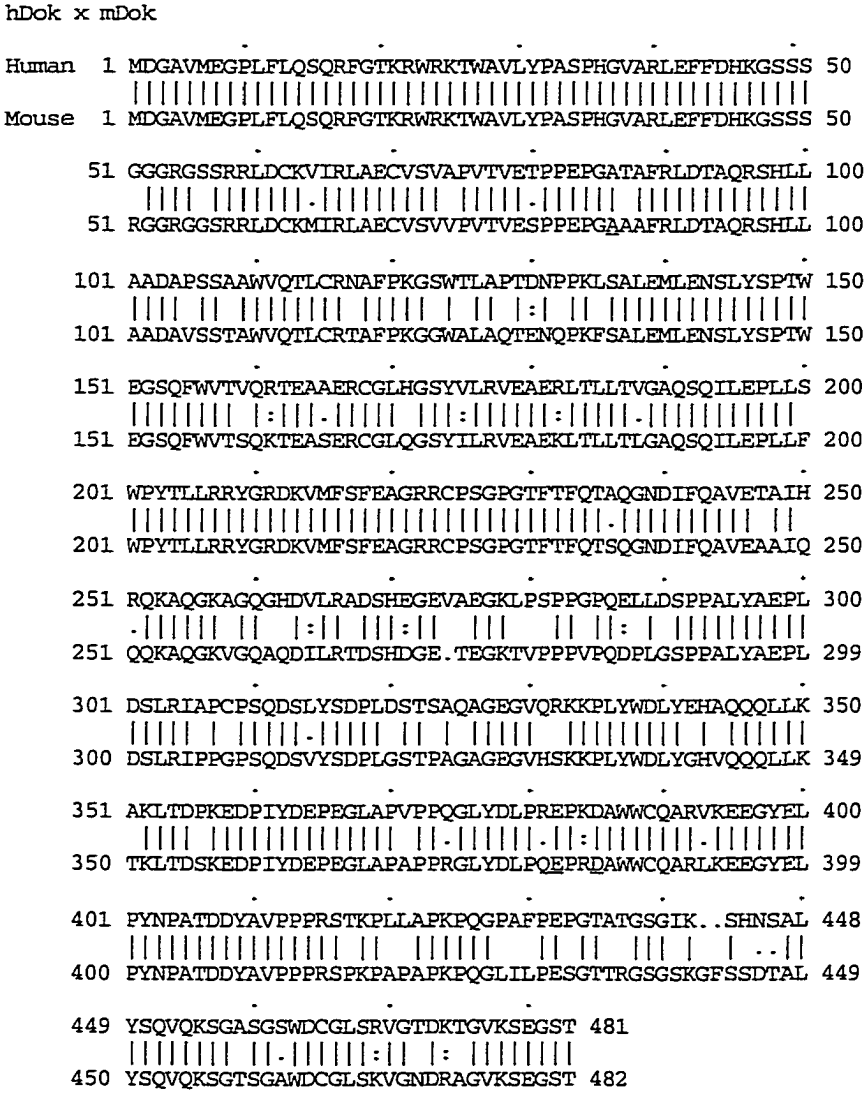


Figure 3.39 GAP alignment of human and mouse Dok protein

The three amino acid residues which are different from published mouse Dok protein (Yamanashi and Baltimore, 1997) are underlined.

The mouse Dok gene encodes a mouse $p62^{Dok}$ protein with 482 amino acids, one amino acid residue longer than the human homologue. The mouse and human Dok protein sequences have an identity of 84% and similarity of 86% (Figure 3.39, previous page). As with the human Dok, a PH domain (3-119) and a proline-rich region (411-424) also are predicted by profile search. Because of their similar primary sequences, the mouse and human Dok proteins, as discussed in section 3.6, most likely have similar functions.

Regulatory elements: There is no TATA box in the region 5' to the coding region of the gene. However, a CpG island occurs at positions 6238-7050. This CpG island also spans the first two exons of the gene as it did with the human Dok gene. A single promoter was predicted by NNPP (6290-6340) and by TSSW (6381). While the NNPP predicted promoter is about 40 bp upstream to the start position of the mouse Dok cDNA (U78818), the TSSW predicted site coincides with the start site of the cDNA. Thus, the 300 bp region (6081-6381) likely contains the promoter of the mouse Dok gene. Consinspector predicted the presence of 3 ATF, 2 GABP, 3 SP1 and several other transcription factor binding sites (see Figure 4.9 and Section 4.3) in this putative promoter region.

As was the case for human Dok gene, although the typical polyadenylation signal consensus AATAAA was not present 3' to the coding region, the less-frequently used ATTAAA signal did occur near the end of the gene at contig position 8860.

3.7.2 Mouse Gene7

A Powblast search of the Genbank non-redundant and dbEST databases revealed that the 150 kb region centromeric to the mouse Dok gene contains regions with homology to several ESTs. The presence of ESTs which significantly matched the contig sequence suggests the possibility of genes in this region. One EST entry (W88205) aligned to the contig with 99% identity in three regions of the complementary strand, indicating the possible presence of three exons in the contig. This possible gene, which contains nine

exons (see below) and spans positions 9-100k of this contig, was named Gene7 by our collaborators at University of Michigan (Jiang, 1998).

During this dissertation research, we and our collaborators at the University of Michigan identified several genes in this contig region (Jiang et al, 1999). By exon trapping, a cDNA clone was isolated and its sequence revealed four exons, three of which overlapped the EST entry W88205. Since it was not possible to isolate a full length cDNA clone for the putative gene using as many as eight commercial and custom-made cDNA libraries, our collaborators then performed a series of RACE (rapid amplification of cDNA ends) and RT-PCR experiments. Here, the DNA sequences collected by this dissertation research were used to predict exons by Genscan and Xgrail and the predicted exons were amplified by RT-PCR and further cDNA cloning. These experiments identified nine putative exons for the gene and six other presumably false positive exons (Jiang et al., 1999). Based on human and mouse sequence comparison, Geneligner predicted 7 exons in the region, six of which matched the putative exons identified by Jiang et al. (1999). The exon and intron sizes of the mouse Gene7 gene are shown in Table 3.24. Beginning with a methionine codon in the first exon of Gene7, the putative protein is encoded by nine exons that would be translated into a 529 amino acid peptide.

Table 3.24 Exon and intron size of mouse Gene7

no.	contig position	Exon size(bp)	Intron size(bp)
1	90398-90101	298	11264
2	78836-78651	186	13632
3	65018-64850	169	21477
4	43372-43199	174	1607
5	41591-41429	163	669
6	40759-40618	142	26989
7	13628-13422	207	1998
8	11423-11283	141	1231
9	10051-9652	400	

Since a database search with the putative protein sequence of Gene7 did not reveal any homology to known proteins and a profile database search also did not reveal any homology to known protein domains, little is known about the possible function of Gene7.

Regulatory elements: Since no TATA boxes or CpG islands were found in the vicinity of the 5' end exon of the mouse Gene7 gene, the sequence between the 5' end exon of Gene7 and the first exon of upstream semaphorin M gene (next section) (about 17 kb) was subjected to promoter prediction by TSSW. Five promoters were predicted by TSSW and the promoter with the highest score was predicted at contig position 99835 in a region with relatively high GC content (60%) relative to the surrounding region (40%) although this region was not predicted by Xgrail to be a CpG island, probably due to its relatively low GC and CpG contents. In the ~9 kb between this potential promoter and the first Gene7 exon, one promoter with a lower score was predicted by TSSW at contig position 93370 in a region overlapping a repetitive B1 element. Thus, the region around contig position 99835 likely contains the promoter of the Gene7. The promoter prediction results by NNPP were not considered relevant since more than 30 weak promoters were predicted between the 5' end exons of Gene7 and upstream semaphorin M gene. Consinspector predicted two ETS-1, two ETS-2, one NFkB and one GABP transcription factor binding sites within the potential promoter at contig positions 99830-100130.

A consensus polyA signal is found at contig position 9532, about 130 bp downstream of the 3' end of the terminal exon of mouse Gene7. The less frequently used polyA signal (ATTAAA) also was observed immediately at the end of the terminal exon of Gene7.

3.7.3 Mouse semaphorin M gene

Several ESTs were aligned to the contig sequence in the region from 128-133k. These include several mouse ESTs (e.g. H24181, Z45329) as well as human ESTs (e.g. AA459837, T09073) among others. Although these ESTs aligned to the contig sequence

with multiple mismatches, consensus splicing sites and an open reading frame were maintained within six potential exons. Genscan predicted one gene containing 20 exons and Xgrail also predicted multiple exons in this region, giving a good indication that this region contains a gene. As was the case with the mouse Gene7, as discussed above, our collaborators at the University of Michigan performed a series of experiments including RT-PCR and cDNA cloning based on information from the database homology search and gene prediction programs using the genomic sequence collected during this dissertation research (Jiang, 1998). Eight additional exons subsequently were identified upstream from the six exons predicted by EST homology. Since the most 5' exon resides in a CpG island, and the most 3' exon has a stop codon, a 2271 bp open reading frame encompassing 14 exons likely contains all of the coding region for this putative gene. The sizes of the exons and introns are listed in Table 3.25.

Table 3.25 Sizes of exons and introns of mouse semaphorin M

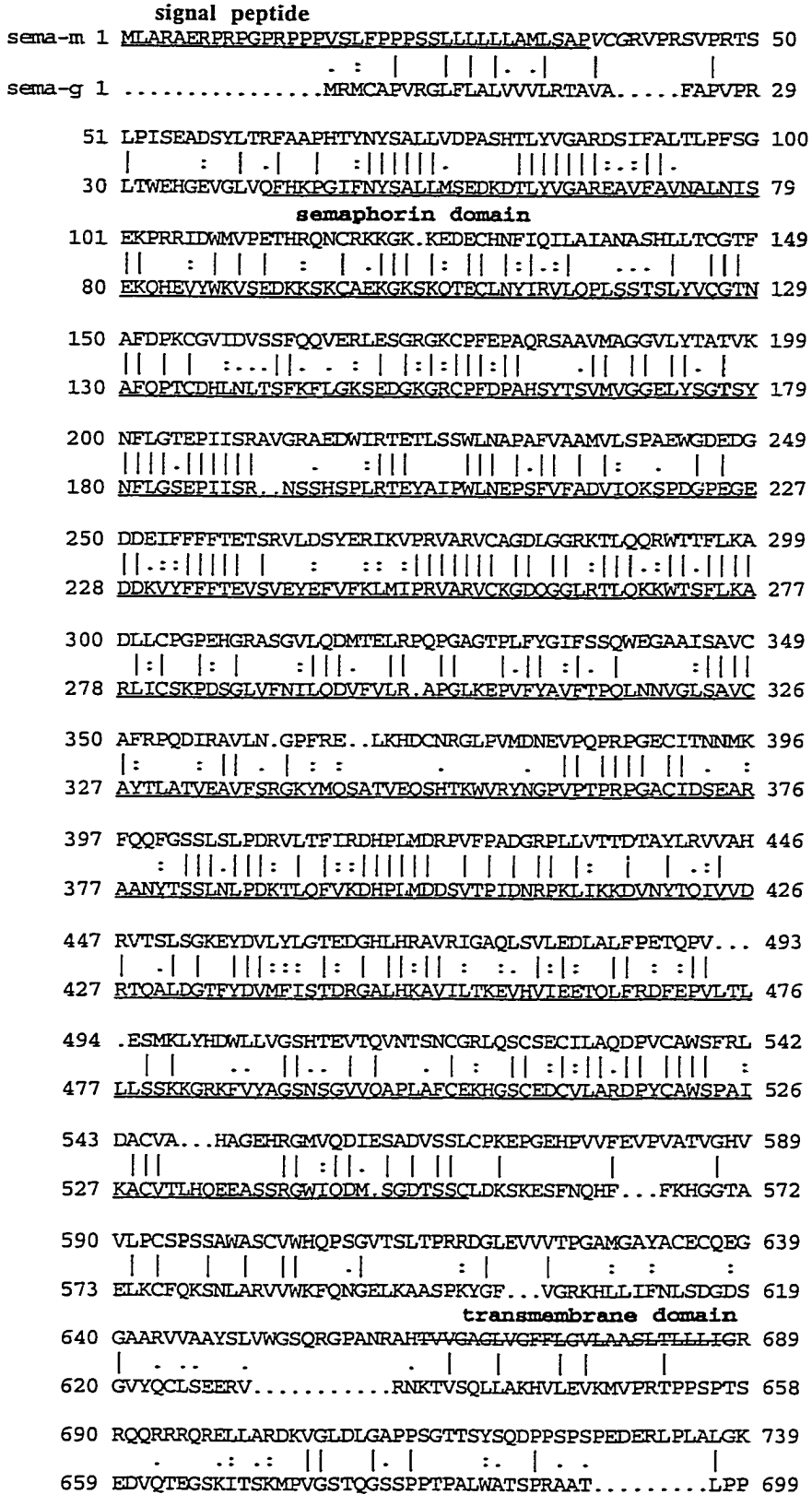
no.	contig position	Exon size(bp)	Intron size(bp)
1	107166-107328	163	2252
2	109581-109732	152	1107
3	110840-110899	60	209
4	111109-111207	99	4595
5	115803-115896	94	11219
6	127116-127235	120	94
7	127330-127481	152	606
8	128088-128266	179	200
9	128467-128614	148	138
10	128753-128975	223	124
11	129100-129209	110	106
12	129316-129476	161	2896
13	132373-132431	59	262
14	132694-133491	798	

These 14 exons encode a putative protein with 777 amino acids and a protein database homology search revealed a high degree of sequence similarities with several semaphorin family members. The region with the highest homologies extends from residues 50 to 600 and contains ~500 amino acids that are conserved in all semaphorin

family members (Kolodkin, 1996). Since these results suggest that this region encodes a gene belonging to the semaphorin family, it has been named semaphorin M (*Semam*) (Jiang, 1998).

Semaphorins, also known as collapsins, are a family of proteins that typically are about 750 amino acid residues long and are believed to be involved in neuronal axon guidance during neural network formation (Kolodkin, 1996). Since the first semaphorin was identified in grasshopper (Kolodkin, 1992), many members of the family have been identified in other species including both human and mouse. It is estimated that at least 10 unique semaphorin family members may exist in any given higher vertebrate (Koppel and Raper, 1998). In addition to the common semaphorin domain, these proteins generally also contain an N-terminal signal peptide and an immunoglobulin-like (Ig) domain. A stretch of highly basic, C-terminal amino acid residues also exist in some secreted Semaphorins and a transmembrane domain in transmembrane semaphorins (Kolodkin, 1996). In mouse, at least 8 family members (M-semaA, B, C, D, E, F, G, H) have been characterized (Furuyama et al., 1996; Christensen et al., 1998). Since the nucleic acid sequence of semaphorin M differs from those semaphorins already characterized, we thus have identified a new member for the mouse semaphorin family.

A protein database homology search revealed that semaphorin M had the highest similarity with two transmembrane semaphorins, the mouse M-semaG (U69535) (Furuyama et al., 1996) and the human semaphorin CD100 (Hall et al., 1996). An alignment of semaphorin M with M-semaG is shown in Figure 3.40. As can be seen from Figure 3.40, the similarity between semaphorin M and M-semaG mainly is in the semaphorin domain. A blastp search with the protein sequence segments outside the semaphorin domain revealed no significant matches to any semaphorin family members or any other known proteins. However, M-semaG was predicted to contain two stretches of hydrophobic sequence, one located at the N terminus which might function as a signal peptide and the other located close to the C terminus which could serve as a potential


signal peptide
sema-m 1 MLARAERPRPGPRPPPVSLFPPPSSLLLLLLAMLSAPVCGRVPRSVPRTS 50
sema-g 1 .....MRMCAPVRGLFLALVVVLRTAVA.....FAPVPR 29
51 LPISEADSYLTRFAAPHTYNYSALLVDPASHTLYVGARDSIFALTLPFSG 100
30 LTWEHGEVGLVQFHKPGIFNYSALLMSEDKDTLYVGAREAVFAVNALNIS 79
semaphorin domain
101 EKPRRIDWMVPETHRQNCRKKGK.KEDECHNFIQILAIANASHLLTCGTF 149
80 EKQHEVYWKVSEDKKSKCAEKGKSKQTECLNYIRVLQPLSSTSLYVCGTN 129
150 AFDPKCGVIDVSSFQQVERLESGRGKCPFEPAQRSAAVMAGGVLYTATVK 199
130 AFQPTCDHLNLTSFKFLGKSEDGKGRCPFDPAHSYTSVMVGGELYSGTSY 179
200 NFLGTEPIISRAVGRAEDWIRTETLSSWLNAPAFVAAMVLSPAEWGDEDG 249
180 NFLGSEPIISR..NSSHSPLRTEYAIPWLNEPSFVFADVIQKSPDGPEGE 227
250 DDEIFFFFTETSRVLDSYERIKVPRVARVCAGDLGGRKTLQQRWTTFLKA 299
228 DDKVYFFFTEVSVEYEFVFKLMIPRVARVCKGDQGGLRTLQKKWTSFLKA 277
300 DLLCPGPEHGRASGVLQDMTELRPQPGAGTPLFYGIFSSQWEGAAISAVC 349
278 RLICSKPDSGLVFNILQDVFVLR.APGLKEPVFYAVFTPQLNNVGLSAVC 326
350 AFRPQDIRAVLN.GPFRE..LKHDCNRGLPVMDNEVPQPRPGECITNNMK 396
327 AYTLATVEAVFSRGKYMQSATVEQSHTKWVRYNGPVPTPRPGACIDSEAR 376
397 FQQFGSSLSLPDRVLTFIRDHPLMDRPVFPADGRPLLVTTDTAYLRVVAH 446
377 AANYTSSLNLPDKTLQFVKDHPLMDDSVTPIDNRPKLIKKDVNYTQIVVD 426
447 RVTSLSGKEYDVLYLGTEDGHLHRAVRIGAQLSVLEDLALFPETQPV... 493
427 RTQALDGTFYDVMFISTDRGALHKAVILTKEVHVIEETQLFRDFEPVLTL 476
494 .ESMKLYHDWLLVGSHTEVTQVNTSNCGRLQSCSECILAQDPVCAWSFRL 542
477 LLSSKKGRKFVYAGSNSGVVQAPLAFCEKHGSCEDCVLARDPYCAWSPAI 526
543 DACVA...HAGEHRGMVQDIESADVSSLCPKEPGEHPVVFEVPVATVGHV 589
527 KACVTLHQEEASSRGWIQDM.SGDTSSCLDKSKESFNQHF...FKHGGTA 572
590 VLPCSPSSAWASCVWHQPSGVTSLTPRRDGLEVVVTPGAMGAYACECQEG 639
573 ELKCFQKSNLARVVWKFQNGELKAASPKYGF...VGRKHLLIFNLSDGDS 619
transmembrane domain
640 GAARVVAAYSLVWGSQRGPANRAHTVVGAGLVGFFLGVLAASLTLLLIGR 689
620 GVYQCLSEERV.....RNKTVSQLLAKHVLEVKMVPRTPPSPTS 658
690 RQQRRRQRELLARDKVGLDLGAPPSGTTSYSQDPPSPSPEDERLPLALGK 739
659 EDVQTEGSKITSKMPVGSTQGSSPPTPALWATSPRAAT.....LPP 699

```
740 RGSGFGGFPPPFLLDSCPSPAHIRLTGAPLATCDETSI..... 777
    : |       |  .:.. |   |   |     .. .   .
700 KSSSGTSCEPKMVINTVPQ.LHSEKTVYLKSSDNRLLMSLLLFIFVLFLC 748

778 ..................................................... 778

749 LFSYNCYKGYLPGQCLKFRSALLLGKKTPKSDFSDLEQSVKETLVEPGSF 798

778 ..................................................... 778

799 SQQNGDHPKPALDTGYETEQDTITSKVPTDREDSQRIDELSARDKPFDVK 848

778 ............. 778

849 CELKFADSDADGD 861
```

Figure 3.40 Alignment of mouse semaphorin M and M-semaG (U69535)
Potential signal peptide, transmembrane domain and semaphorin domain are as indicated.

transmembrane-spanning domain (Furuyama et al., 1996). To investigate the possibility that semaphorin M also is a transmembrane semaphorin, transmembrane domain prediction was performed. Here, the TMpred program predicted 2 transmembrane domains (residues 19-40 and 664-688) with scores over 1000. This result is similar to that for M-semaG although there was no protein sequence similarity detected within the predicted transmembrane regions between the two semaphorin members. Signal peptide prediction by SignalP (Nielsen et al., 1997) indicated the presence of a signal peptide at residues 1-40, which spans the first transmembrane domains predicted by TMpred. Thus, the mouse semaphorin M likely also is a transmembrane protein with one potential transmembrane spanning domain. Northern blotting experiments (Jiang, 1998) show that the semaphorin M transcript is enriched in mouse brain and spinal cord. This result is consistent with the association of the semaphorin family proteins with neuronal network formation/function although the exact role of semaphorin M is unknown.

Regulatory elements: Although no TATA box was observed in the vicinity of the first exon of the semaphorin M gene, the presence of a 300 bp CpG island (107017-107350) spanning the first exon suggests that the promoter of the gene may be located in this CpG island. Within a 1000 bp sequence spanning the CpG island, NNPP predicted

three promoters at contig positions 107040-107090, 107096-107146 and 107193-107243, respectively, while TSSW predicted one promoter at position 107203, 38 bp downstream from the start of the 5' end exon. When the region between the upstream Gene7 and the first exon of semaphorin M (~17 kb) was searched for possible promoters by the TSSW program, five promoters were predicted in the forward strand, with the promoter at position 107203 having the highest score. Based on these results, the region 107000-107200 most likely contains the promoter for the mouse semaphorin gene. Within this region, Consinspector predicted one NF-Y, three SP1 and one AP1 transcription factor binding sites. Since there was no consensus polyadenylation sequence observed near the end of the 3' end coding exon, Jiang (1998) noted that their preliminary data suggested that this gene has an additional 1.4 kb 3' UTR. Consistent with their results, examination of genomic sequence revealed a polyadenylation consensus sequence (AATAAA) at contig position 134900, 1508 bp downstream from the end of the 3' end coding exon of the mouse semaphorin M gene.

3.7.4 Mouse lysyl-oxidase related gene 2 (Lor2)

One mouse EST entry (AA522066) matched the contig within the first 6 kb of this sequence with an identity of 99% over a region with three potential exons for an unknown gene. The cDNA clone corresponding to the mouse EST subsequently was sequenced by collaborators at the University of Michigan (Jiang, 1998) and was found to contain an insert of 2871 bp (AF053368). Alignment of the contig sequence with the complete sequence of the cDNA clone reveals that the contig contains the 5' end four exons of the unknown gene (Jiang, 1998). The size of the four exons and the three introns is shown in Table 3.26.

Table 3.26 Size of exons and introns of mouse Lor2 gene

no.	contig position	Exon size(bp)	Intron size(bp)
1	5612-5528	85	1164
2	4363-4021	343	1634
3	2386-2223	164	805
4	1417-1203	215	
other exons are out of this contig			

The cDNA sequence of this unknown gene contains an ORF of 2262 bp, with in-frame stop codons at both the 5' and 3' ends of the ORF. From a methionine codon near the 5' end of the ORF, a putative protein of 754 amino acids is predicted (Figure 3.41). A protein database search shows that this predicted protein has a high sequence similarity to the human lysyl oxidase-related protein WS9-14 (U89942) and a *Perca favescens* lysyl oxidase homologue (AF103901), with similarities of 64% and 52%, respectively, and thus this gene has been given the name lysyl-oxidase related gene 2 (Lor2). The function of these homologous genes are unknown but the human WS9-14 protein was suggested to be an extracellular protein possibly involved in cell adhesion and senescence (Saito et al., 1997). Other protein similarities were observed in two domains. The N-terminal 550 amino acids of Lor2 protein have a high sequence similarity to the human DMBT1 protein (AJ000342), the mouse CRP-ductin-alpha protein (U37438), and several human and beef antigens (I38004-6, S56744) because all of these proteins contain several scavenger receptor cysteine-rich (SRCR) domains (2-9) in regions homologous to the Lor2 protein (Law et al., 1993; Nunes et al., 1995; Cheng et al., 1996; Mollenhauer et al., 1997). A profile database search also revealed four SRCR domains for the Lor2 protein at amino acids 45-146, 170-283, 308-408, and 418-526, respectively. The SRCR domain, possibly involved in binding to other cell-surface and extracellular molecules, is about 100 amino acids long and has been found in at least eight cell surface or secreted proteins (reviewed in Resnick et al., 1994). The C-terminal 200 amino acids of Lor2 protein shows a high sequence similarity to the human (M94054), mouse (M65142), rat (J02903) and chicken

(M97881) lysyl oxidases which are copper-dependent enzymes catalyzing collagen and elastin cross-linking (Hamalainen et al., 1993) whose C-terminal regions were thought to contain the active site for these proteins (Krebs and Krawetz, 1993). These results thus suggest that the Lor2 protein also is a cell surface or secreted protein possibly involved in cell adhesion and other extracellular processes as suggested for WS9-14 protein to which Lor2 protein has the highest similarity. Consistent with this hypothesis, Psort also predicted that the Lor2 protein would be localized to extracellular space including cell wall, with a potential signal sequence predicted for amino acids 1-26.

```
lor2  Length: 754

   1  MRAVSVWYCCPWGLLLLHCLCSFSVGSPSPSISPEKKVGSQGLRFRLAGF
            Signal peptide
  51  PRKPYEGRVEIQRAGEWGTICDDDFTLQAAHVLCRELGFTEATGWTHSAK
      SRCR 1
 101  YGPGTGRIWLDNLSCRGTEGSVTECASRGWGNSDCTHDEDAGVICKDQRL

 151  PGFSDSNVIEVEHQLQVEEVRLRPAVEWGRRPLPVTEGLVEVRLPEGWSQ
                                SRCR 2
 201  VCDKGWSAHNSHVVCGMLGFPGEKRVNMAFYRMLAQKKQHSFGLHSVACV

 251  GTEAHLSLCSLEFYRANDTTRCSGGNPAVVSCVLGPLYATFTGQKKQQHS

 301  KPQGEARVRLKGGAHQGEGRVEVLKAGTWGTVCDRKWDLQAASVVCPELG
                    SRCR 3
 351  FGTAREALSGARMGQGMGAIHLSEVRCSGQEPSLWRCPSKNITAEDCSHS

 401  QDAGVRCNLPYTGVETKIRLSGGRSRYEGRVEVQIGIPGHLRWGLICGDD
                           SRCR 4
 451  WGTLEAMVACRQLGLGYANHGLQETWYWDSGNVTEVVMSGVRCTGSELSL

 501  NQCAHHSSHITCKKTGTRFTAGVICSETASDLLLHSALVQETAYIEDRPL

 551  HMLYCAAEENCLASSARSANWPYGHRRLLRFSSQIHNLGRADFRPKAGRH
      Conserved region in lysyl oxidases
 601  SWVWHECHGHYHSMDIFTHYDILTPNGTKVAEGHKASFCLEDTECQEDVS

 651  KRYECANFGEQGITVGCWDLYRHDIDCQWIDITDVKPGNYILQVVINPNF

 701  EVAESDFTNNAMKCNCKYDGHRIWVHNCHIGDAFSEEANRRFERYPGQTS

 751  NQIV
```

Figure 3.41 Predicted protein sequence of mouse Lor2 gene
The four conserved histidines (highlighted) are the proposed lysyl oxidase Cu-binding site (Krebs and Krawetz, 1993)

The 5' end exons of the mouse Lor2 and Dok genes are 768 bp apart. Since the human syntenic region sequenced in this dissertation research (Contig 5) only contains 564 bp upstream from the human Dok gene, the human homologue of the mouse Lor2 gene would not be contained within contig 5, a prediction which was confirmed by a dotter analysis of the mouse Lor2 cDNA with the sequence of contig 5. However, as discussed in Section 3.6, part of the human genomic region upstream the human Dok gene has been sequenced and deposited in Genbank by another group (AC005033). A dotter analysis of the Genbank sequence with the mouse Lor2 cDNA sequence revealed that the human Lor2 gene is located about 800 bp upstream the human Dok gene in the opposite orientation and contains 14 exons (Table 3.27). Its coding region is well conserved but its 5' and 3' UTRs are only moderately conserved in humans. If translation begins from the methionine codon located in the second exon, the human lor2 gene would be predicted to encode a 754 amino acid protein that is one amino acid more than predicted mouse Lor2 protein.

Table 3.27 Exon and Intron sizes of human Lor2 gene

exon no.	Exon (bp)	Intron (bp)	Note
1	>58	~1.2 kb	boundaries can not be determined
2	>313	1973	5' end boundary can not be determined
3	164	601	
4	215	12440	
5	220	237	
6	181	140	
7	155	240	
8	168	133	
9	163	517	
10	244	99	
11	116	79	
12	137	106	
13	112	202	
14	545		

The alignment of the predicted human and mouse Lor2 proteins (Figure 3.42) revealed a similarity of 94.2% and an identity of 92.7%. This high degree of sequence similarity suggests that they most likely have similar functions in both human and mouse.

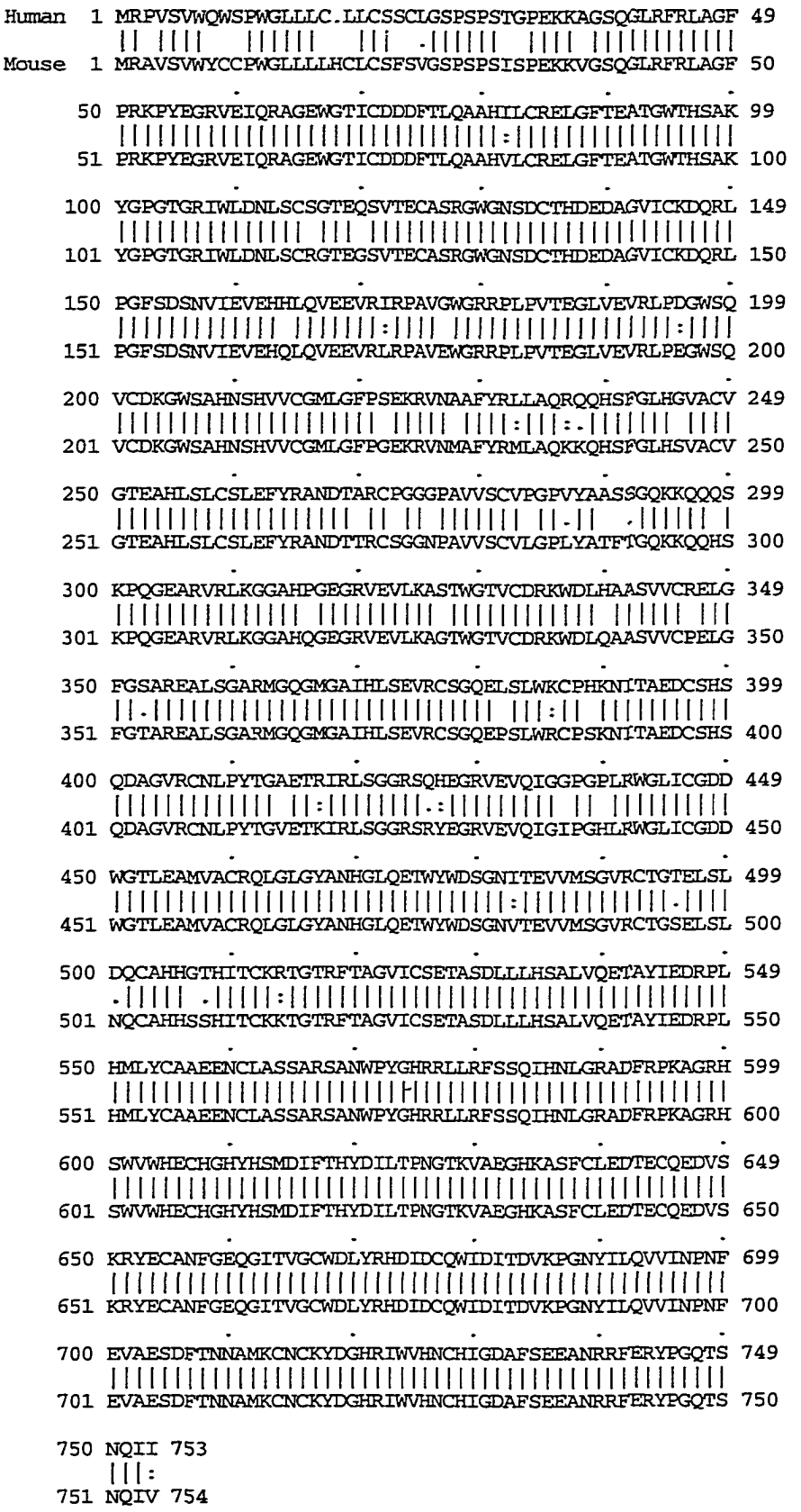

Human 1 MRPVSVWQWSPWGLLLC.LLCSSCLGSPSPSTGPEKKAGSQGLRFRLAGF 49
Mouse 1 MRAVSVWYCCPWGLLLLHCLCSFSVGSPSPSISPEKKVGSQGLRFRLAGF 50
50 PRKPYEGRVEIQRAGEWGTICDDDFTLQAAHILCRELGFTEATGWTHSAK 99
51 PRKPYEGRVEIQRAGEWGTICDDDFTLQAAHVLCRELGFTEATGWTHSAK 100
100 YGPGTGRIWLDNLSCSGTEQSVTECASRGWGNSDCTHDEDAGVICKDQRL 149
101 YGPGTGRIWLDNLSCRGTEGSVTECASRGWGNSDCTHDEDAGVICKDQRL 150
150 PGFSDSNVIEVEHHLQVEEVRIRPAVGWGRRPLPVTEGLVEVRLPDGWSQ 199
151 PGFSDSNVIEVEHQLQVEEVRLRPAVEWGRRPLPVTEGLVEVRLPEGWSQ 200
200 VCDKGWSAHNSHVVCGMLGFPSEKRVNAAFYRLLAQRQQHSFGLHGVACV 249
201 VCDKGWSAHNSHVVCGMLGFPGEKRVNMAFYRMLAQKKQHSFGLHSVACV 250
250 GTEAHLSLCSLEFYRANDTARCPGGGPAVVSCVPGPVYAASSGQKKQQQS 299
251 GTEAHLSLCSLEFYRANDTTRCSGGNPAVVSCVLGPLYATFTGQKKQQHS 300
300 KPQGEARVRLKGGAHPGEGRVEVLKASTWGTVCDRKWDLHAASVVCRELG 349
301 KPQGEARVRLKGGAHQGEGRVEVLKAGTWGTVCDRKWDLQAASVVCPELG 350
350 FGSAREALSGARMGQGMGAIHLSEVRCSGQELSLWKCPHKNITAEDCSHS 399
351 FGTAREALSGARMGQGMGAIHLSEVRCSGQEPSLWRCPSKNITAEDCSHS 400
400 QDAGVRCNLPYTGAETRIRLSGGRSQHEGRVEVQIGGPGPLRWGLICGDD 449
401 QDAGVRCNLPYTGVETKIRLSGGRSRYEGRVEVQIGIPGHLRWGLICGDD 450
450 WGTLEAMVACRQLGLGYANHGLQETWYWDSGNITEVVMSGVRCTGTELSL 499
451 WGTLEAMVACRQLGLGYANHGLQETWYWDSGNVTEVVMSGVRCTGSELSL 500
500 DQCAHHGTHITCKRTGTRFTAGVICSETASDLLLHSALVQETAYIEDRPL 549
501 NQCAHHSSHITCKKTGTRFTAGVICSETASDLLLHSALVQETAYIEDRPL 550
550 HMLYCAAEENCLASSARSANWPYGHRRLLRFSSQIHNLGRADFRPKAGRH 599
551 HMLYCAAEENCLASSARSANWPYGHRRLLRFSSQIHNLGRADFRPKAGRH 600
600 SWVWHECHGHYHSMDIFTHYDILTPNGTKVAEGHKASFCLEDTECQEDVS 649
601 SWVWHECHGHYHSMDIFTHYDILTPNGTKVAEGHKASFCLEDTECQEDVS 650
650 KRYECANFGEQGITVGCWDLYRHDIDCQWIDITDVKPGNYILQVVINPNF 699
651 KRYECANFGEQGITVGCWDLYRHDIDCQWIDITDVKPGNYILQVVINPNF 700
700 EVAESDFTNNAMKCNCKYDGHRIWVHNCHIGDAFSEEANRRFERYPGQTS 749
701 EVAESDFTNNAMKCNCKYDGHRIWVHNCHIGDAFSEEANRRFERYPGQTS 750
750 NQII 753
751 NQIV 754

Figure 3.42 Alignment of predicted human and mouse Lor2 proteins (previous page)

Regulatory elements: A sequence segment of 768 bp is located between the 5' end of the first exon of the Lor2 gene and 5' end of the first exon of the downstream mouse Dok gene (5613 - 6380). Within this region, TSSW predicted a promoter at contig position 6235 which was 622 bp upstream from the first exon of the Lor2 gene, and falls within the predicted promoter region of the upstream Dok gene. Similarly, NNPP predicted a promoter at contig position 6410 which was 797 bp upstream from the first exon of the Lor2 gene and falls within the 5' end exon of the upstream Dok gene. Both of the promoters are within the major CpG island (6238-7050) described above in Section 3.7.1. These results suggest that the mouse Lor2 gene might share a common promoter with the Dok gene as seems to be the case with the human Lor2 and Dok genes. In addition, since this contig only covers exons 1-4 of the mouse Lor2 gene, it is not known whether the intron between exons 4-5 of the Lor2 gene also contains an alternative promoter for the mouse Dok gene as in human.

3.7.5 A retrotransposon: intracisternal A particle (IAP)

Intracisternal A-particles are endogenous defective retrovirus-like elements present at about a thousand copies in the mouse and other rodent genomes (Lueders and Mietz, 1986; Heidmann and Heidmann, 1991). These IAP elements in mouse are highly polymorphic, ranging in size from 3.5-7.3 kb (Ymer et al., 1986; Heidmann and Heidmann, 1991). They possess long terminal repeats, with U3-R-U5 organization that is homologous to the retroviral LTRs (Heidmann and Heidmann, 1991) and they also encode open reading frames for the viral gag-pol-like proteins (Kuff and Lueders, 1988). IAPs have been grouped into two classes (class I and II) and several subclasses. Class I IAPs are collinear with the genomic RNA of the particles which codes for the IAP main structural protein p73 but usually have variably sized deletions (Paterson et al., 1978). Class II IAPs

comprise of about 10% of the IAPs in the mouse genome and have a characteristic 0.3 kb insertion as well as internal deletions of variable size (Shen-Ong and Cole, 1982; Lueders and Mietz, 1986).

A Powblast database search revealed that the sequence in the contig region between position 25k-35k was highly homologous to several IAP elements in the Genbank database. A dotter plot comparative analysis of this contig and a full length IAP (M17551, clone MIA14, 7096 bp) indicates homology to a full length IAP element of 7089 bp at contig positions 27673 through 34761 (referred to as IAPg7 due to its position within an intron of mouse Gene7). A dotter analysis between IAPg7 and a type II IAP (X03520) shows that the IAPg7 does not contain the 0.3 kb insertion characteristic of type II IAPs, and thus, is likely a class I IAP element. Examination of the IAPg7 sequence shows that the LTRs on both ends of IAPg7 are identical both in length (353 bp) and in nucleic acid sequence. There are no other significant repeat regions within the sequence of IAPg7 but a short 6 bp direct repeat (GTAAGC) is located on both sides of the IAP element.

Examination of the IAPg7 sequence also shows that there are several open reading frames with more than 100 codons between the LTRs. These open reading frames encode putative gag, pol and protenase proteins based on protein database homology search results. The ORF for proteinase overlaps with that of putative gag and pol proteins. The pol gene likely is non-functional since an insertion of T at contig position 32498 introduced an in- frame stop codon within the otherwise long ORF which codes for a full length pol protein. Likewise, there is no discernible open reading frame for the retroviral env gene, which is a common feature of IAP elements and thus renders these elements incapable of forming infectious extracellular viruses (Kuff and Fewell, 1985; Ono et al., 1985; Ymer et al., 1986).

Although most known IAP elements, including IAPg7, contain one or more non-functional genes, an observation that implies they most likely are not competent for autonomous transposition, it was suggested that transposition of IAP elements might be a

powerful source of insertional mutagenesis and might play a role in both transformation and tumor progression (Heidmann and Heidmann, 1991). This proposal was based on several detailed studies on IAP elements and their associated genes, since transcription of a truncated cellular c-mos oncogene in a mouse plasmacytoma was correlated with the insertion of an IAP sequence (Rechavi et al., 1982). An IAP element (IAP-IL3) also was found inserted upstream of the promoter of the interleukin-3 gene of the leukaemia cell line WEHI-3B (Ymer et al., 1986). However, a detailed mechanism for the transposition of these seemingly non-competent IAP elements is unknown.

Another aspect of IAP research concerns its role in transcription and aging. IAP elements generally are repressed in normal cells and transcribed at a low level in some adult tissues (Piko et al., 1984; Dupressoir and Heidmann, 1996). Dupressoir et al. (1995) reported that two IAP-related transcripts were present in organs of aging mice but not detected in young mice. Although the exact mechanism is unknown, the association of the expression of the IAP element with aging has lead these authors to propose that the IAP retrotransposons might serve as reporters of aging.

As described above, the IAPg7 is located in the intron region of the putative gene Gene7. Therefore it might be expected that if IAPg7 transcription is activated, transcription into the flanking cellular sequence (as in the case reported by Puech et al., 1997) followed by unique splicing events would result in an aberrant Gene7 product, which could in turn have an adverse effect on the cells.

3.7.6 A pseudogene of ribosomal protein S6

A blast search reveals that the region encompassing positions 55 -57 kb of this contig has significant sequence similarity to the mouse ribosomal protein S6 cDNA (Rps6: Y00348) and a number of ESTs. The mouse Rps6 gene contains 5 exons (Pata and Metspalu, 1996). However, a dotter comparison (Figure 3.43) shows that the contig sequence and the mouse Rps6 cDNA sequence are collinear, and thus this genomic region

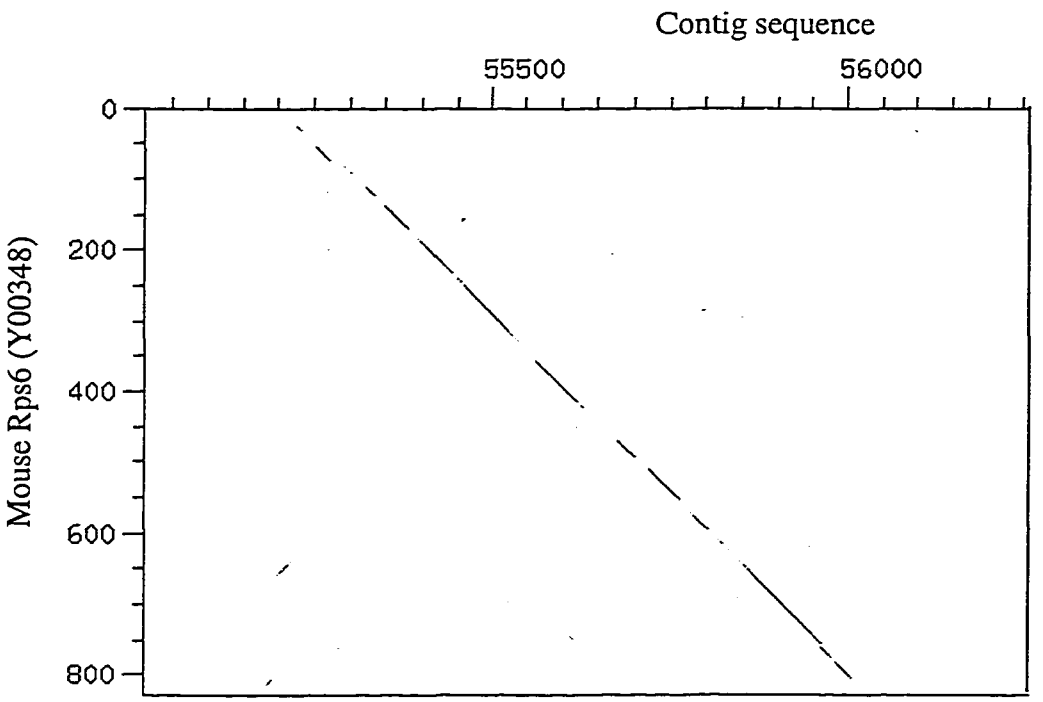


Figure 3.43 Dotter plot of contig sequence with mouse Rps6 cDNA sequence

only contains an exon sequence but lacks any introns. Further examination of this region indicates that multiple stop codons are present in the homologous coding region of Rps6 cDNA. Taken together, the lack of introns and the lack of a suitable open reading frame suggest that this is a mouse Rps6 pseudogene.

Many mouse ribosomal protein genes have non-functional copies integrated into the genome as pseudogenes. Pseudogenes of the mouse ribosomal protein L7 (Klein and Meyuhas, 1984), ribosomal protein L32 (Dudov and Perry, 1984) and now ribosomal protein S6, are but three examples. It recently has been estimated that there are up to 15 processed pseudogenes of the mouse Rps6 gene dispersed throughout the mouse genome (Pata and Metspalu, 1996). The presence of a Rps6 pseudogene in this contig indicates that a member of the mouse Rps6 family of pseudogenes is on human chromosome 22.

3.7.7 A pseudogene of ribosomal protein L34

The region between contig positions 93580-94070 also contains a copy of ribosomal protein L34 (RPL34) cDNA, as revealed by a Powblast search of the Genbank databases. A sequence alignment (Figure 3.44) indicates that this region of the contig is 84% identical to the human RPL34 cDNA sequence (L38941). However, comparison of this genomic sequence and that of the RPL34 coding region revealed two in-frame stop codons in the genomic sequence and no introns. The genomic region also was bounded by two 13 bp direct repeat and a 11 bp poly-A sequence was present near its 3' end. Although the genomic structure for human RPL34 gene is unknown, it is known that human RPL34 contains at least one 80 bp intron (Kenmochi et al., 1998). The lack of introns in this genomic sequence and the other features described above indicate that contig region 93580-94070 most likely contains a pseudogene of RPL34.

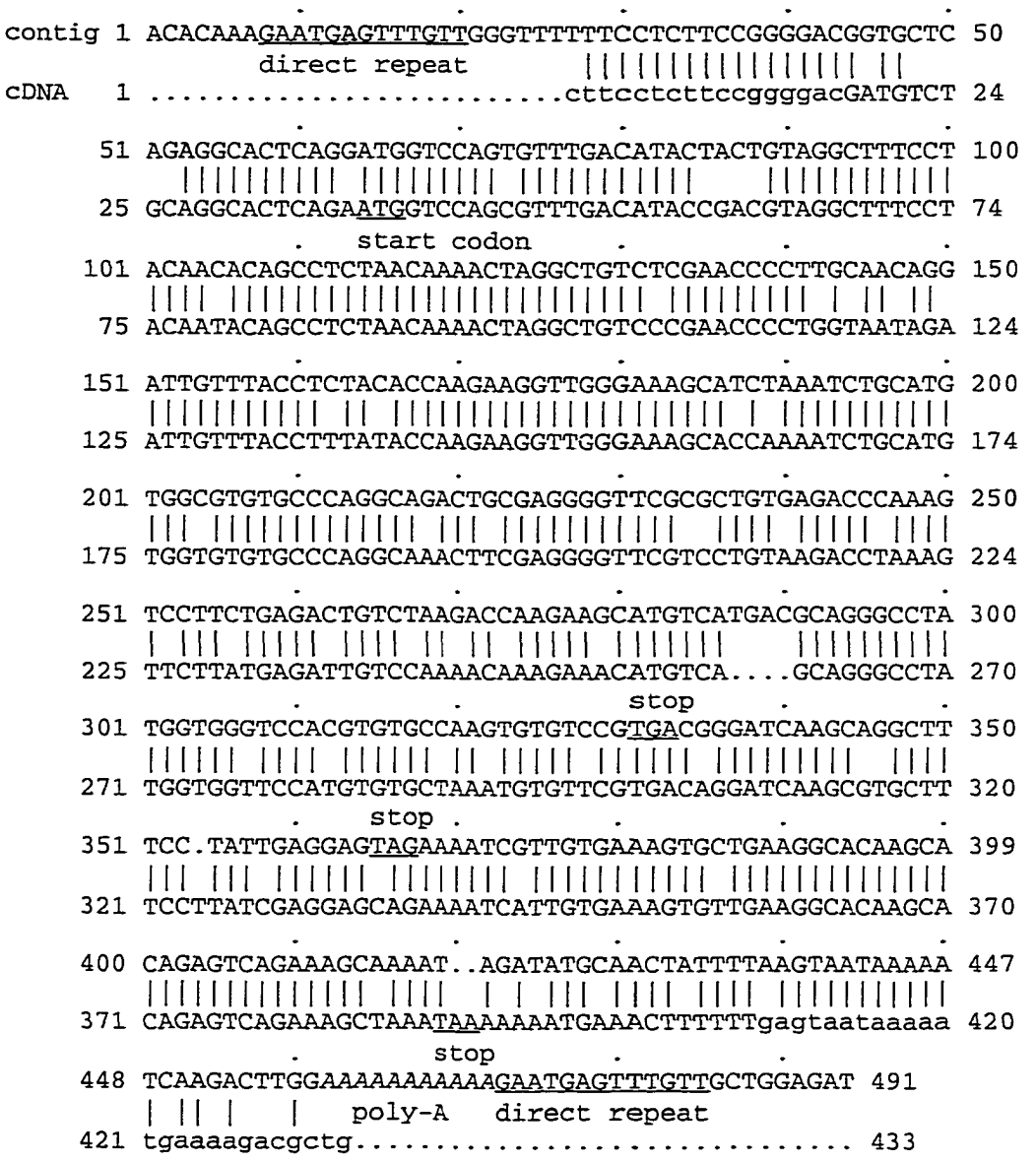


Figure 3.44 Alignment of contig sequence with human RPL34 cDNA

Contig sequence was from contig positions 93580-94070 and was complemented. Both 5' and 3' end of RPL34 cDNA (L38941) were extended according to EST AA825612 and the extended parts were shown in small letters.

3.7.8 Other repetitive elements in contig 6

Repeatmasker analysis of contig 6 revealed that 37.98% of this contig was repeat sequences. The major repeat families present include B1, B2, L1 and LTR as shown in Figure 3.45 and Table 3.28. Relative to the contigs analyzed in previous sections above, contig 6 is relatively rich in simple repeat sequences as they account for 4.3% of this contig. Interestingly, the short intron regions of the Dok gene lack any noticeable repeat sequences while the intron regions of semaphorin M gene which contain B1 and B2 repeats do not contain any L1 and LTR repeats and the Gene7 introns contain B1 and B2 as well as L1 and LTR repeats. The orientation of L1 repeat in the Gene7 region does not show the overall preference in orientation as described above in section 3.6, an observation similar to that for the human Gene7 region.

Table 3.28 Content of repetitive sequence in contig 6

	number of elements	length occupied	percentage of sequence
SINEs			
B1	71	8698 bp	5.34 %
B2	60	10445 bp	6.42 %
B4	34	5112 bp	3.14 %
ID	7	511 bp	0.31 %
LINEs			
L1	23	9862 bp	6.06 %
LTR elements	47	18220 bp	11.20 %
MER elements	3	313 bp	0.19 %
Unclassified	7	1303 bp	0.80 %
Simple Repeat	91	6946 bp	4.30 %
Total repeats:	343	61765 bp	37.98 %

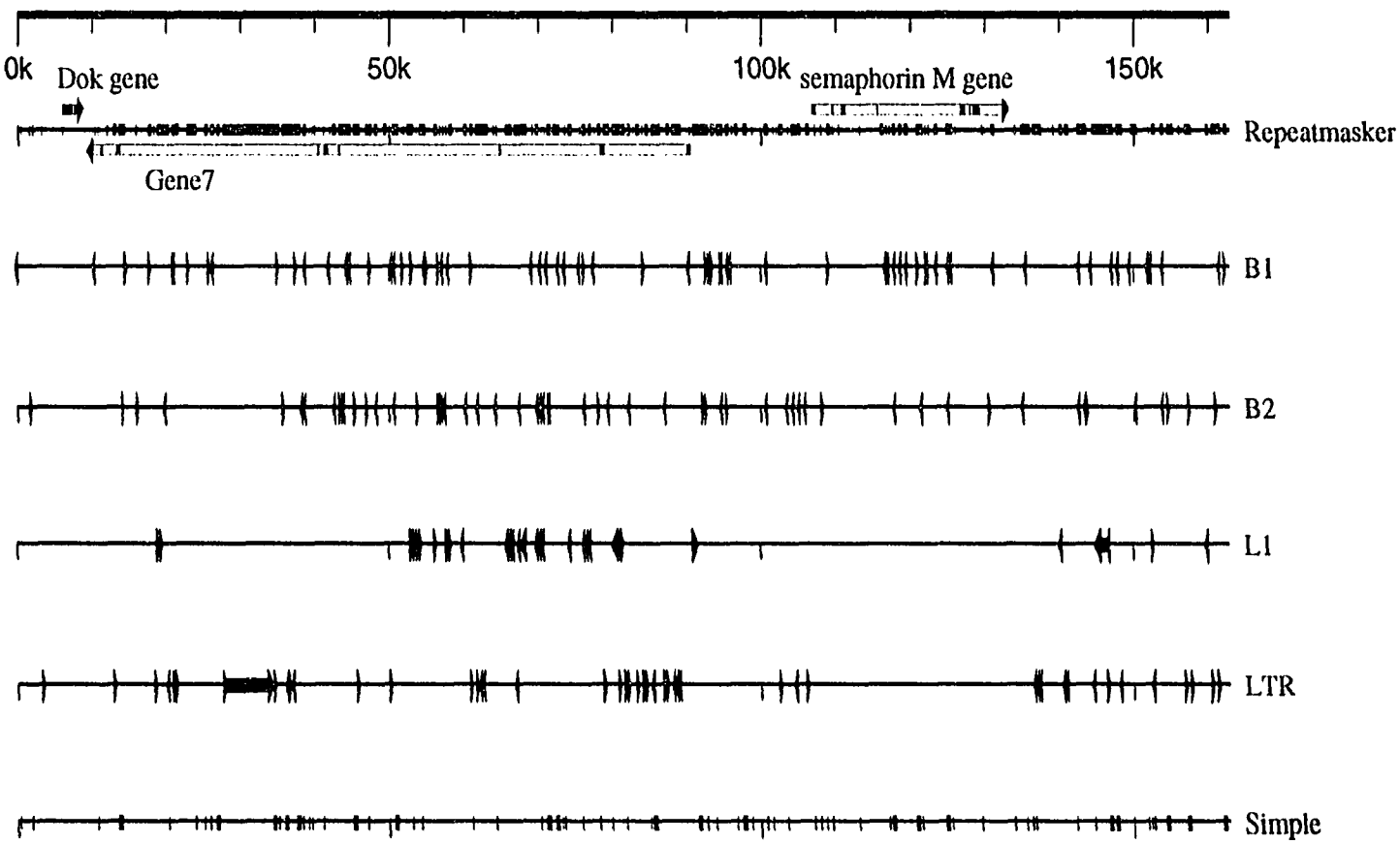


Figure 3.45 Distribution of repetitive elements in contig 6

Chapter IV
Comparison of human and mouse sequences

4.1 Introduction

Human and mouse genomic DNA sequence comparison offers several advantages when determining the functional significance of a DNA segment because functional elements tend to be conserved between closely related species. For example, Makalowski et al. (1996) compared 1196 full-length cDNA sequences from human and mouse and found that most of the coding regions were highly conserved, with a nucleic acid level identity averaging 85%. Regulatory elements of gene transcription also frequently were found conserved (e.g. Baumgartner, 1994). These observations, which were obtained from comparison of known genes, can be applied to uncharacterized genomic sequences and therefore facilitate identifying new genes and their associated regulatory elements. As described in Chapter 3, the identification of the human and mouse genes Gene7 was based in part on comparing human and mouse sequences determined in this dissertation research. In addition, comparative genomics also can provide information that improves our understanding of genome organization and evolution (Li, 1997), and the results from such comparisons can be useful in producing transgenic mice as a model system for specific genetic based diseases (Searle et al., 1994). Thus, the mouse genome also is being actively sequenced along with the human genome, with an expected completion date of 2005 (Collins et al., 1998).

A limited number of large scale (>20 kb) human and mouse sequence comparisons have been published in the past several years (e.g. Koop and Hood, 1994; Oeltjen et al., 1997; Chen, 1997; Ansari-Lari et al., 1998; Jiang et al., 1999). These studies confirmed the notion that coding regions are well conserved between human and mouse. In contrast, non-coding regions are conserved to varying degrees, leading to the proposal that different regions of the human and mouse genomes have evolved at different rates (Koop, 1995). It

also is known that both the human and mouse genomes contain a large number of randomly interspersed repeats. However, when the present studies were begun, only a small fraction of the human and mouse genomes had been sequenced and additional work is needed to provide a clearer and more accurate comparative picture of the human and mouse genomes.

As part of this dissertation research, two pairs of syntenic regions of the human and mouse genomes were sequenced. The first pair was from human chromosome 22 (PAC p201m18, contig 3) and its syntenic region from mouse chromosome 16 (BAC 72k21, contig 4). A dotter comparison of these two sequences is shown in Figure 4.1. Crossmatch analysis also was performed and the results revealed several large, highly conserved regions as shown in Figure 4.2. Based on the distribution of conserved segments, it can be estimated that the centromeric 82.3 kb of the human sequence is syntenic to the telomeric 81.2 kb of the mouse sequence. Within these syntenic regions, four genes, i.e. the T10, PutA, Htf9c and RanBP1 genes occur in both the human and mouse sequence (Chapter 3). These genes are organized in the same order, direction and occupy similar lengths of genomic sequences. It also can be seen that the highly conserved segments generally are gene-rich. The other pair of syntenic regions sequenced in this dissertation research was from human chromosome 2 (BAC h173, contig 5) and its syntenic region in mouse chromosome 6 (BAC b245c12, contig 6). A comparison of these two sequences (Figures 4.3 and 4.4) showed that approximately 89.3 kb of the mouse sequence matched the 91.8 kb human sequence. Within these syntenic regions, two genes, i.e. Dok and Gene7, occur both in human and mouse (Chapter 3) and they are organized in the same relative orientation. It is interesting to note that even though the sizes of individual introns differ by up to 13 kb (first intron of Gene7) between these human and mouse genomic regions, the sizes of the entire syntenic regions differ only by 2.5 kb. This also is true for the first pair of syntenic regions, which differ in size by only 1.1 kb while the entire regions encompass more than 80 kb of human and mouse sequences. A detailed comparison of these two pairs of syntenic human and mouse regions are presented in the following sections.

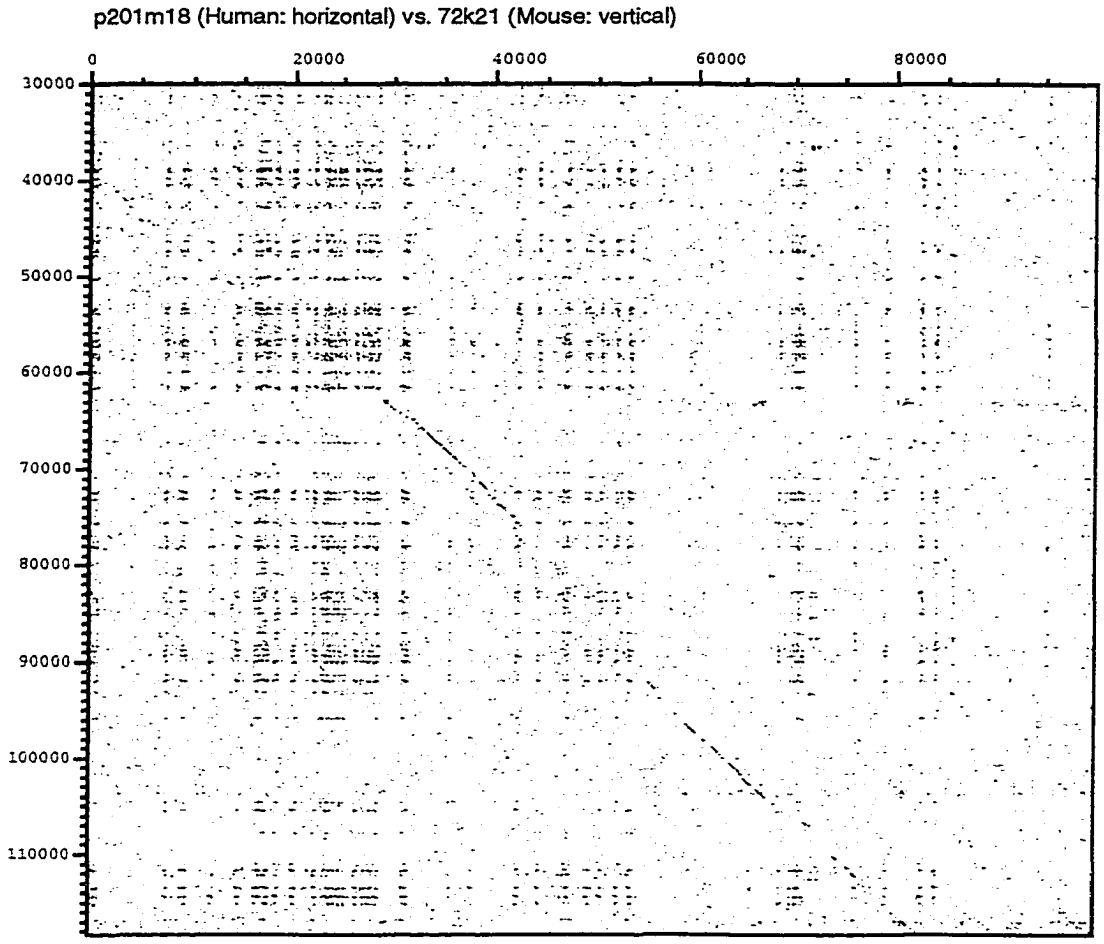


Figure 4.1 Dotter plot of the first human and mouse syntenic regions

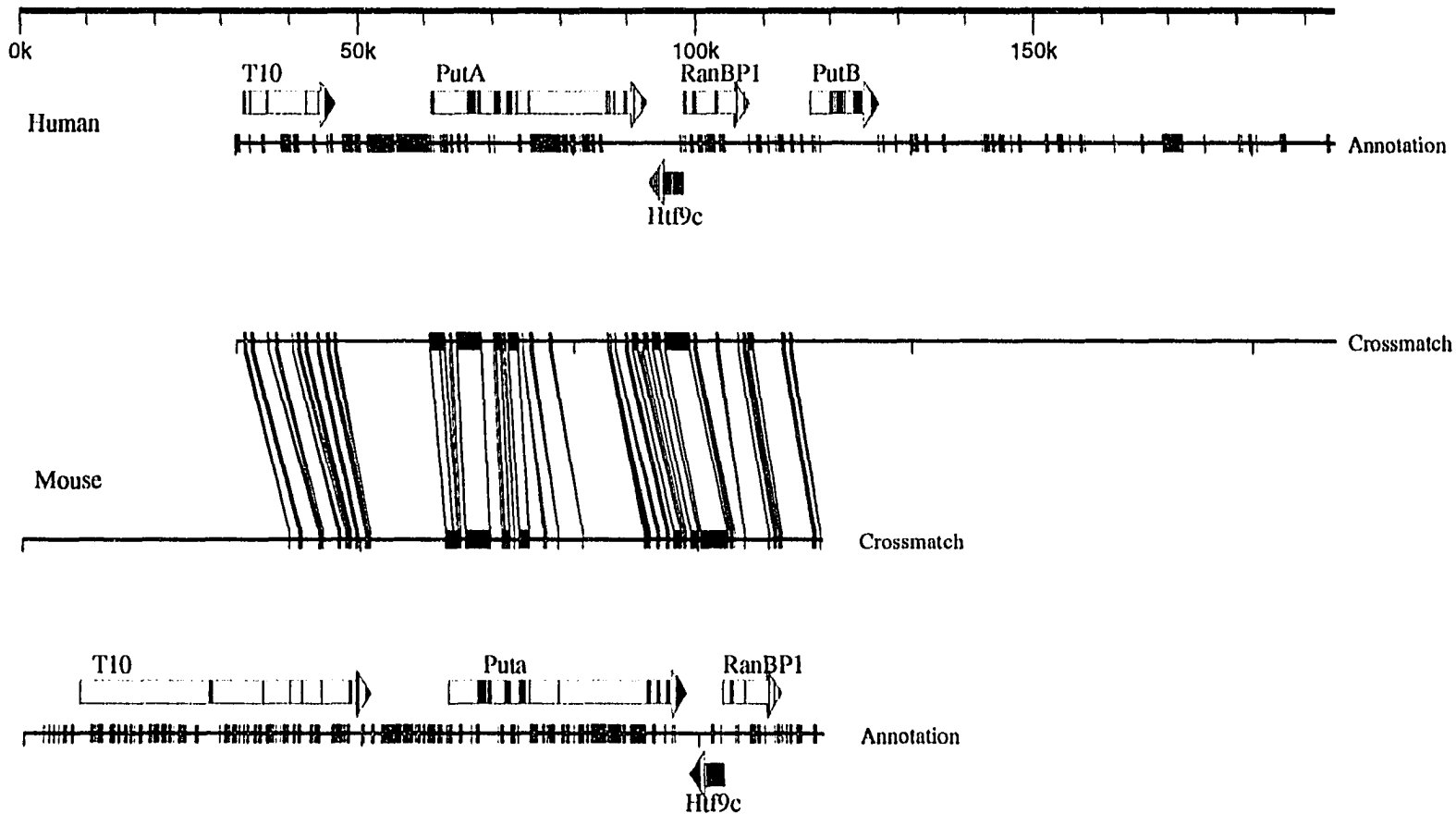


Figure 4.2 Crossmatch result comparing the human (P201m18) and the mouse (72k21) sequences
Crossmatch: Blue=excellent match, Green=good match, Red=poor match
Annotation of both contigs is shown for comparison

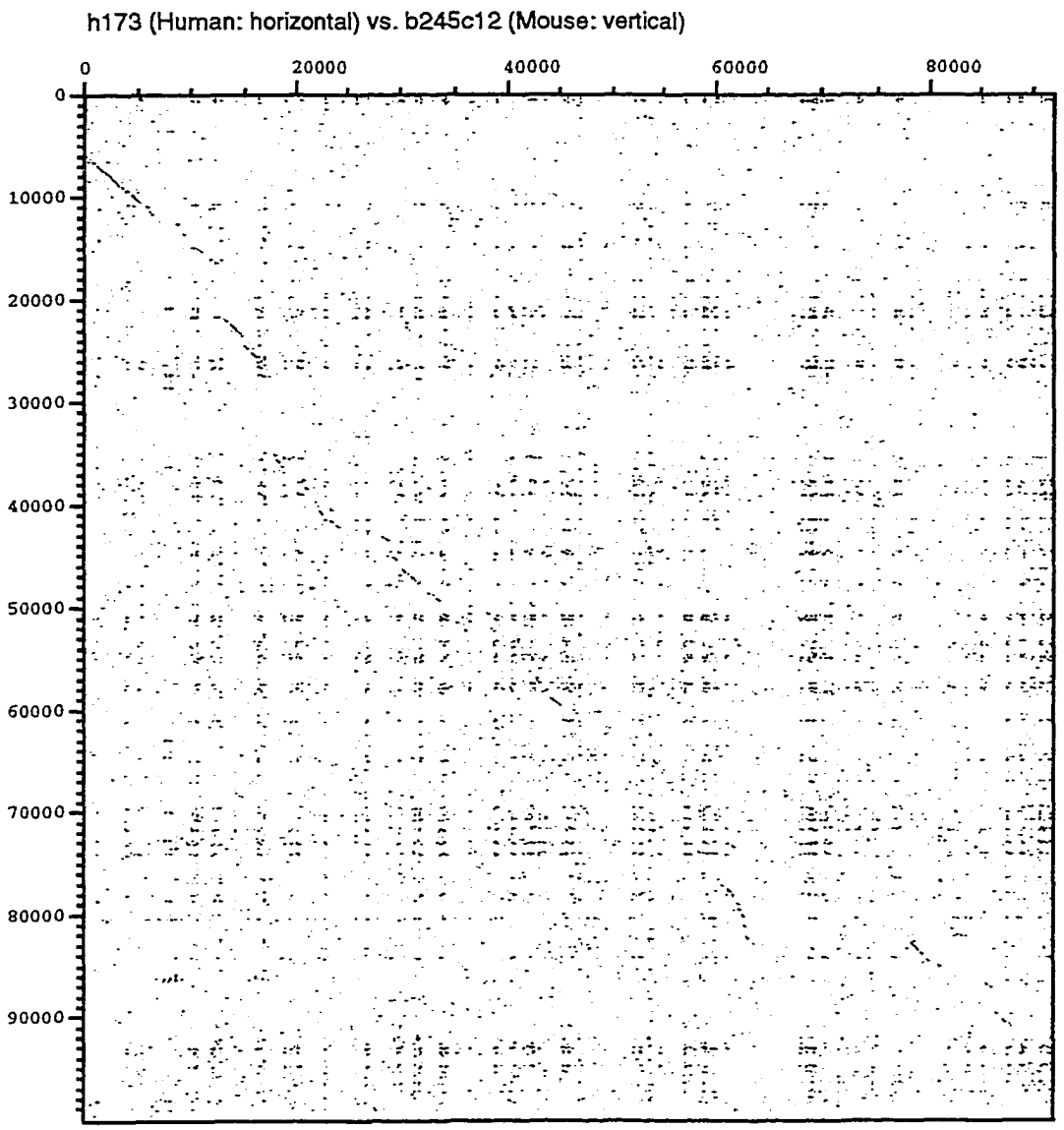


Figure 4.3 Dotter plot of the second human and mouse syntenic regions

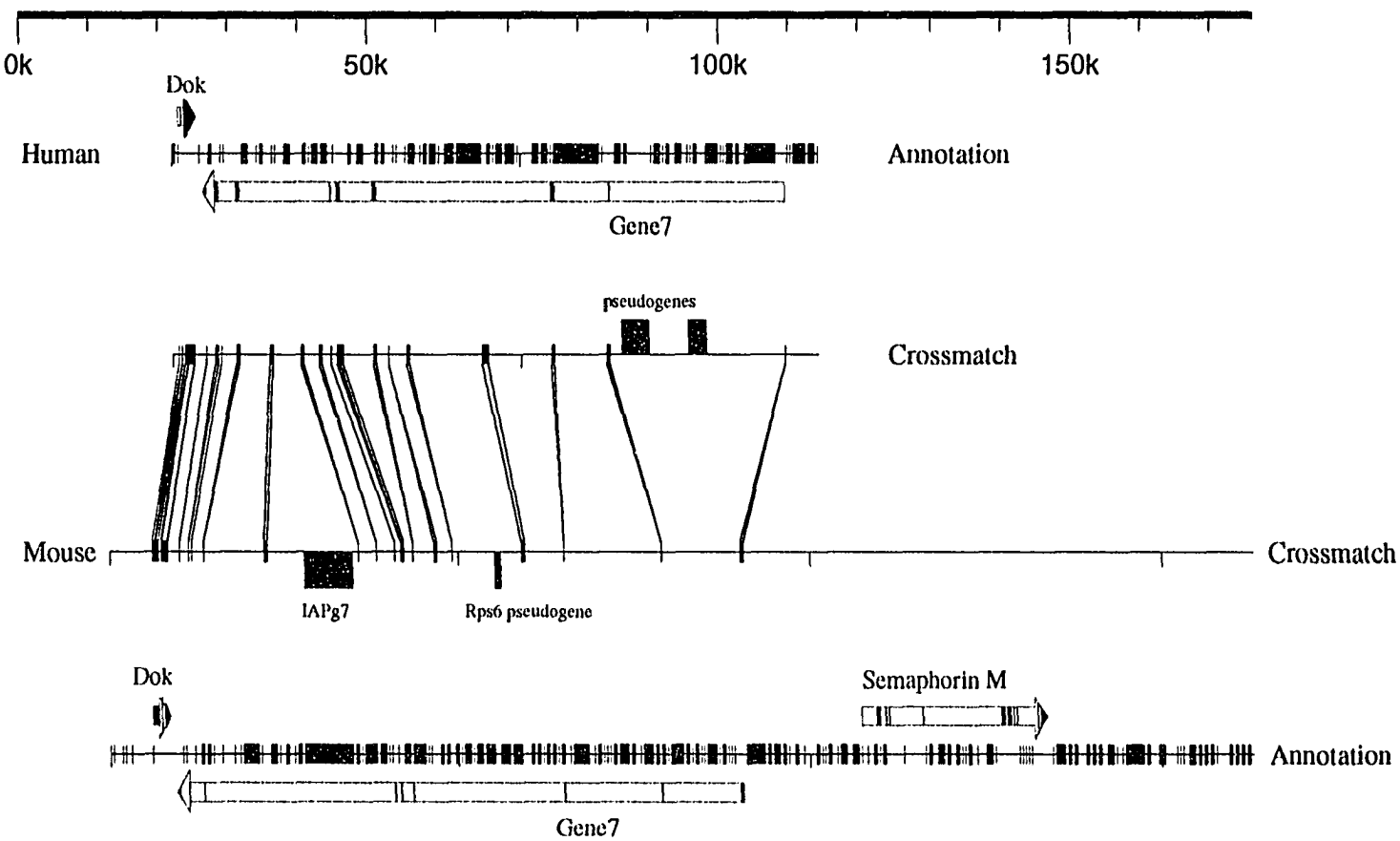


Figure 4.4 Crossmatch result comparing the human (h173) and the mouse (b245c12) sequences

Crossmatch: Blue=excellent match, Green=good match, Red=poor match
Annotation of both contigs is shown for comparison

4.2 Comparison of gene organization

Comparison of the gene organization of the six homologous gene pairs within the two pairs of syntenic regions (Table 4.1) shows that the overall gene structures are well conserved between human and mouse. The number of coding and non-coding exons all are conserved for these six gene pairs except that the human PutA gene may have an alternative splicing pattern for its 16 exons. The coding regions are conserved in both their length and nucleic acid sequence, with the coding region identities varying between 79.4-90.6%. In contrast to the overall conservation in the length and sequences of coding region between human and mouse, the 5' and 3' UTR of the six pairs of homologous genes are less conserved, with the identities varying between 30.0-85.6%. However, the level of sequence conservation is not evenly distributed along the length of the UTRs. Instead, there are both conserved and non-conserved segments within the UTRs (Figure 4.5), and it is suspected that the conserved regions may contain regulatory elements of gene expression (see Section 4.4).

Table 4.1 Comparison of gene organization and sequence identity

gene pair	total exons	coding exons	5' UTR length	5' UTR ident.	coding region length	coding region ident.	3' UTR length	3' UTR ident.	protein similarity
hDok	5	5	22		1443		438		
mDok	5	5	6	22.7%	1446	83.2%	316	49.9%	86.0%
hGene7	9	9	52		1578		269bp		
mGene7	9	9	58	67.3%	1587	79.4%	205	59.1%	77.2%
hT10	9	8	163		828		1257		
mT10	9	8	137	55.3%	828	84.3%	610bp	31.2%	90.9%
hHtf9c	12	11	660		1839		267		
mHtf9c	12	11	585	60.7%	1806	80.4%	180	30.0%	86.0%
hRanBP1	6	6	149		600		306		
mRanBP1	6	6	140	85.6%	609	87.5%	306	72.9%	96.0%
hPutA*	15	14	222		2202		1767		
mPutA	15	14	229	82.1%	2202	90.6%	1575	54.8%	97.0%

*The alternative splicing pattern of 15 exons is used for comparison

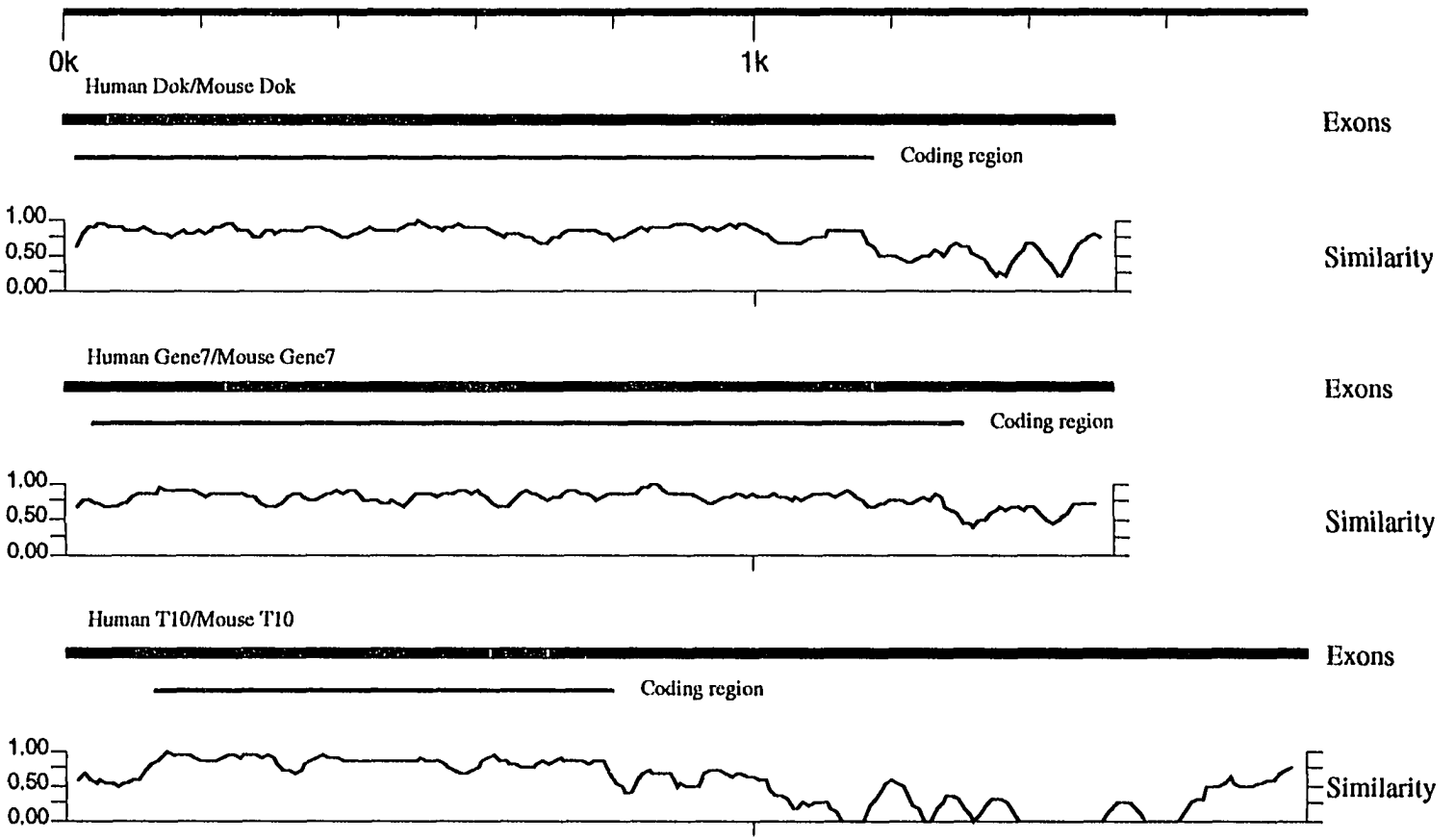


Figure 4.5a DNA sequence similarity comparison of homologous human and mouse genes

(to be continued) Similarity is calculated with a window size of 50 bp

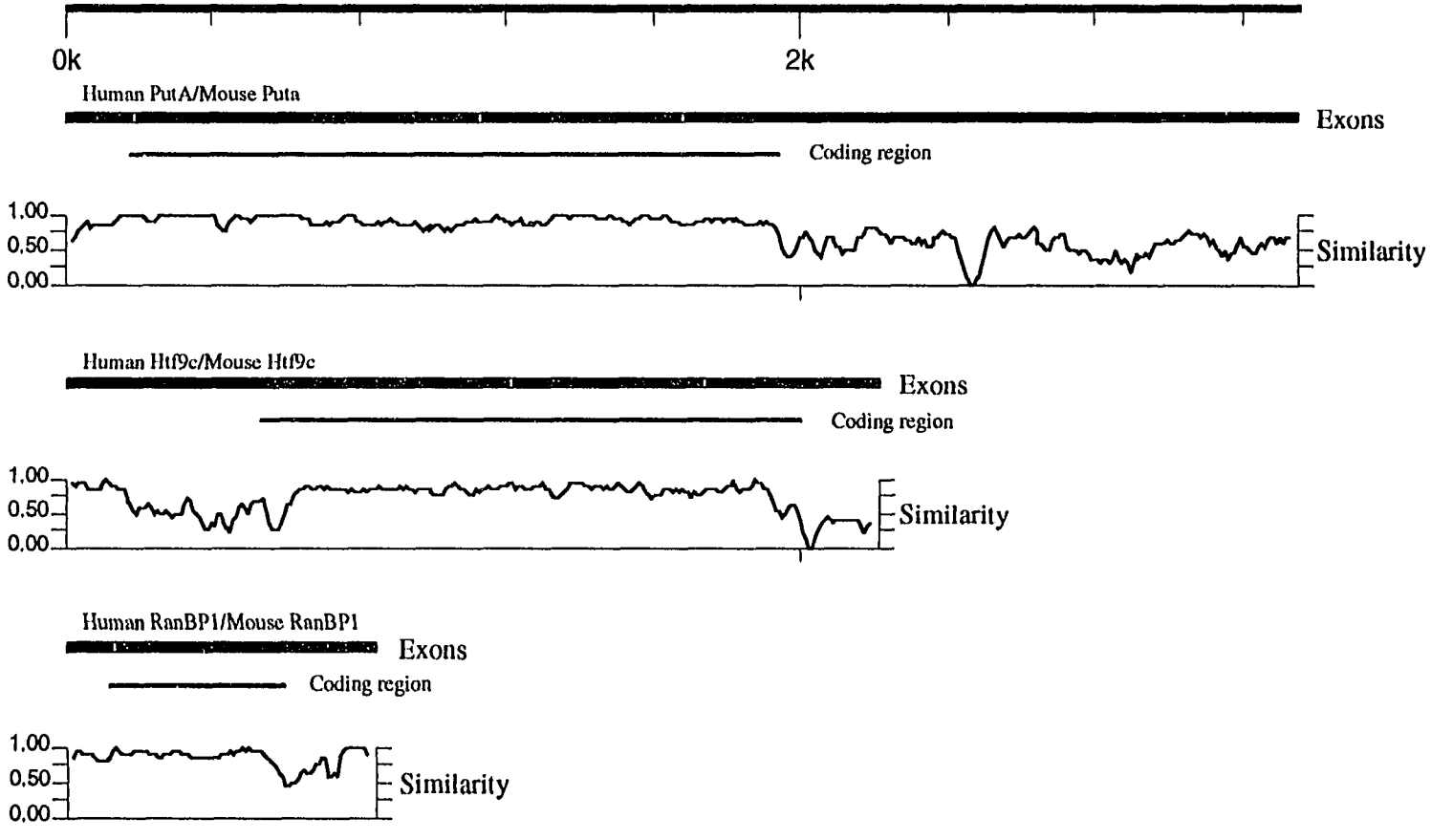


Figure 4.5b (continued)

4.3 Comparison of coding region

As discussed in the previous section, the coding regions of homologous human and mouse genes are highly conserved, even though four of the six gene pairs show slight differences of up to 33 bp in their coding region length. Specifically, the Dok gene pair shows that the human sequence has an insertion of 3 bp and a deletion of 6 bp when compared to the mouse Dok gene in their 3' most exons. For the Gene7 gene pair, a 9 bp difference due to an insertion of 3 bp and a deletion of 12 bp occurs in the human sequence relative to the mouse gene exon 9. When comparing the Htf9c gene pair, a 27 bp deletion in the first coding exon and 6 bp deletion at the last coding exon of the mouse gene result in a total difference of 33 bp. For the RanBP1 gene pair, a 9 bp difference also occurs in the 3' end coding exon. Thus, for all the four gene pairs which show different coding region lengths, the differences occur either in the first or in the last coding exon. Furthermore, the differences between coding exons of homologous genes always are multiples of three and all of the internal coding exons for all the six gene pairs are of equal length.

The high level sequence conservation of homologous coding regions contrasted with the lower level sequence conservation of 5' and 3' UTRs as shown above in Table 4.1. For all the six gene pairs, the transition from high to low level of sequence conservation occurs immediately at the coding region - 3' UTR boundaries (Figure 4.6). In contrast, sequences of 5' UTRs near the 5' UTR - coding region boundaries are relatively conserved although insertions or deletions occur, suggesting that these regions play an important biological function such as initiation of translation.

With a high level of sequence conservation in coding regions and highly accurate comparable DNA sequences available, a detailed analysis of sequence differences is possible. The differences can be deletions, insertions or substitutions, but substitutions are most prevalent in the six pairs of human and mouse genes compared here. Substitutions can be classified as either transitions or transversions. Transitions are intraexchanges between purines (A and G) or pyrimidines (T and C). Transversions are interexchanges

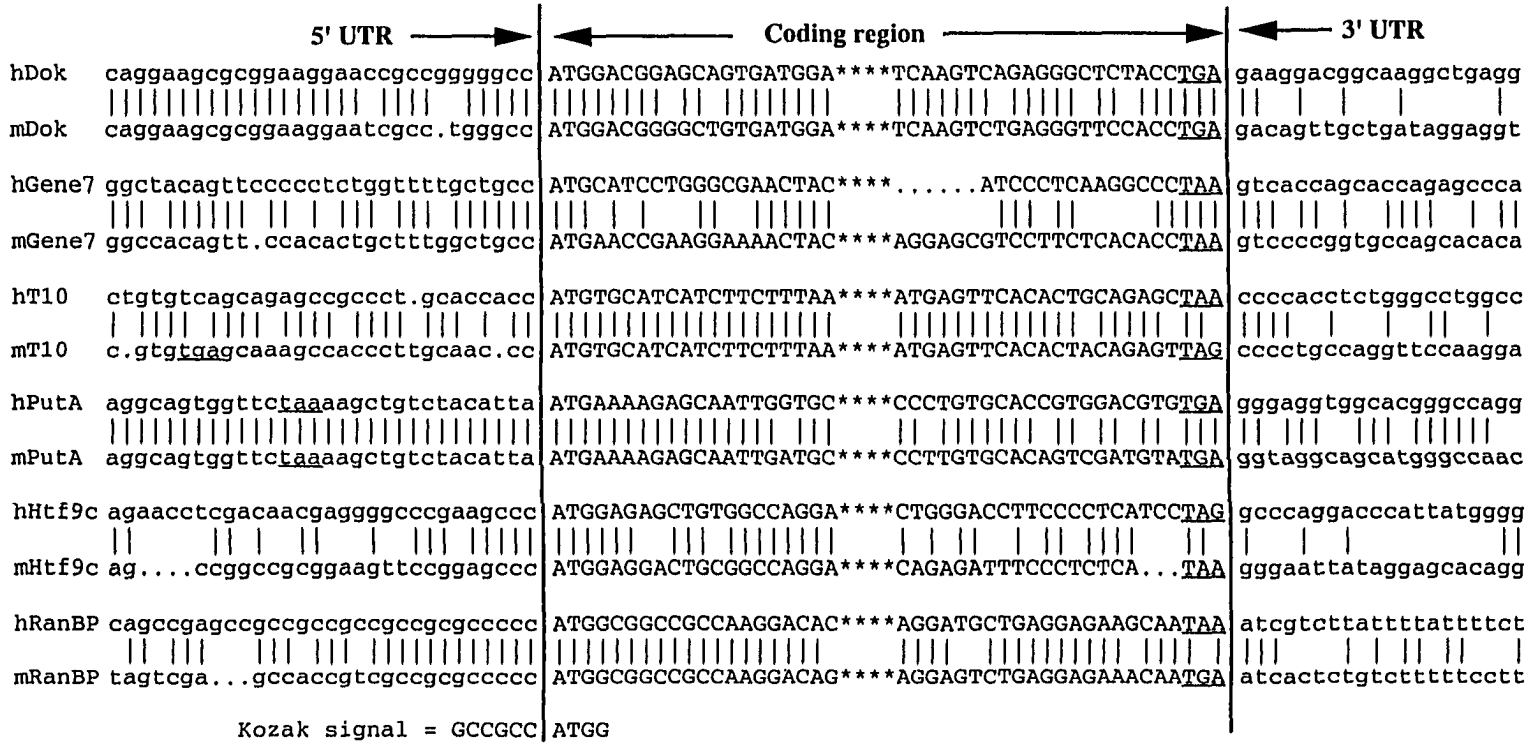


Figure 4.6 DNA sequence alignment at boundaries of coding regions between homologous genes

**** represents omitted internal coding sequence; ... represents sequence gap; | represents identity; Stop codons are underlined. 5' end of mouse Dok gene is extended according to genomic sequence.

between purines and pyrimidines. In coding regions, substitutions also are classified as synonymous substitutions or nonsynonymous substitutions. Synonymous substitutions are those that do not result in amino acid change while nonsynonymous substitutions result in amino acid change (replacement). A summary of coding region substitutions between the six pairs of human and mouse genes are listed in Table 4.2. For all the six pairs of genes, transitions occur more often than transversions, with the apparent ratio between them varying between 1.4 and 3.2. Previous studies on pseudogenes (Li et al., 1984), which were assumed to have no cellular functions, showed that transition had a relative frequency of 59.3%, with a transition/transversion ratio of 1.46. Thus, generally higher transition and transversion ratios likely reflect that there are functional constraints imposed on the genes. Detailed examination of the distribution of these substitutions (Table 4.3) shows that five of the six gene pairs have more than 60% of the substitutions in the third codon positions. Ten out of the twenty amino acids have codons which only allow transitions at their third positions while other amino acids have codons which allow both transition and transversion for synonymous substitution. Thus, both the higher transition/transversion ratios and the predominance of substitutions at the third codon positions indicates that cellular constraints on the substitution process tend to preserve the amino acid sequence. However, this preservation of amino acid sequence can not explain the overall predominance on transition over transversion in the other two codon positions (see Table 4.3), especially since substitutions at second coding positions invariably result in amino acid replacement. The transitional excess in substitutions resulting in amino acid replacement previously has been observed in analyzing 337 human and mouse genes but it was thought likely due to mutational biases rather than selective forces (Collins and Jukes, 1994).

Table 4.2 Summary of substitutions between human and mouse gene pairs

gene pair	ORF length	Ts	Tv	Ts/Tv	(Ts+Tv)/ORF length
Dok	1443	158	76	2.1	16.2%
Gene7	1578	187	133	1.4	20.2%
T10	828	84	46	1.8	15.7%
Htf9c	1839	219	104	2.1	17.5%
RanBP1	600	50	25	2.0	12.5%
PutA	2202	158	50	3.2	9.4%

Note: coding region length are that of human gene of a human and moue gene pair.
Ts: transition Tv: transversion

Table 4.3 Distribution of substitutions between different codon positions

Gene pair	1st position			2nd Position			3rd position		
	Ts	Tv	%Total	Ts	Tv	%Total	Ts	Tv	%Total
Dok	36	17	22.6%	27	12	16.7%	95	47	60.7%
Gene7	50	51	31.6%	42	26	21.3%	95	56	47.2%
T10	8	13	16.2%	12	6	13.8%	64	27	70.0%
Htf9c	33	23	17.3%	36	17	16.4%	150	64	66.3%
RanBP1	6	4	13.3%	1	6	9.3%	43	15	77.3%
PutA	17	7	11.5%	10	3	6.3%	131	49	82.2%

As mentioned above, substitutions also are classified as synonymous or nonsynonymous. Synonymous substitution rates were thought to reflect the rates of evolutionary divergence of nucleic acid sequences (Oeltjen et al., 1997) while nonsynonymous substitutions were believed to reflect the rate of protein evolution (Li, 1997, page 179). Assuming the human and rodent divergence occurred 80 million years ago (Li and Graur, 1991), the synonymous and nonsynonymous substitution rates for the six gene pairs are calculated and shown in Table 4.4. These results overall are similar to

those obtained from comparison of 47 human and rodent genes (Li, 1993; Pamilo and Bianchi, 1993; Li, 1997; average value for synonymous substitution rate: 3.51±1.01; average nonsynonymous substitution rate: 0.74 ± 0.67). As noted above, synonymous substitution rates for the six gene pairs vary in a relatively narrow range (1.99-3.17) compared to that of nonsynonymous substitutions (0.13-1.12). The higher rate of synonymous substitutions than nonsynonymous substitutions reflects functional constraints on nonsynonymous substitutions. From these substitution rates, no significant differences in gene and protein evolution rates were observed with different chromosomal regions, i.e. between the region in human chromosome 2 (Dok and Gene7 gene pair) and the region in human chromosome 22 (the remaining four gene pairs).

Table 4.4 Synonymous and nonsynonymous substitution rates*

Gene pair	ORF length	Synonymous substitution rate	Nonsynonymous substitution rate
Dok	1443	2.64 ± 0.24	0.61 ± 0.06
Gene7	1578	2.54 ± 0.23	1.12 ± 0.09
T10	828	2.85 ± 0.34	0.48 ± 0.08
Htf9c	1839	3.17 ± 0.24	0.59 ± 0.06
RanBP1	600	2.64 ± 0.38	0.16 ± 0.05
PutA	2202	1.99 ± 0.16	0.13 ± 0.02

*the rates, with units of per site per 10^9 years, are calculated by the method of Ina (1995).

Examination of the base composition of human and mouse coding sequences (Table 4.5) show several distinct features. First, the third codon position of all six human genes have a higher GC content than the other codon positions. Second, the GC content of the third codon position of all six human genes is consistently higher than that of the mouse gene, with differences up to 20%. Third, the second codon position of both the human and mouse genes all have a lower GC content than do the other codon positions. These observations are consistent with the earlier research indicating that the third codon position

Table 4.5 Comparison of GC content at each codon position

gene pair	Total	1st position	2nd position	3rd position	intron	contig
hDok	62.0%	64.4%	51.9%	69.8%	61.5%	41.4%
mDok	59.4%	62.3%	51.7%	64.2%	53.9%	45.0%
hGene7	52.4%	55.3%	43.1%	58.7%	40.3%	41.4%
mGene7	52.5%	57.0%	43.7%	57.0%	43.4%	45.0%
hT10	57.7%	56.5%	40.5%	76.0%	56.8%	57.6%
mT10	52.7%	55.0%	40.2%	63.0%	47.4%	46.6%
hHtf9c	63.0%	65.2%	47.1%	76.8%	62.8%	57.6%
mHtf9c	54.1%	62.4%	43.2%	56.7%	47.5%	46.6%
hRanBP1	51.0%	59.0%	33.0%	61.0%	57.1%	57.6%
mRanBP1	50.0%	58.0%	34.5%	57.5%	47.3%	46.6%
hPutA	53.3%	58.5%	39.7%	61.7%	51.9%	57.6%
mputa	49.3%	57.6%	39.6%	50.8%	43.2%	46.6%

of human genes frequently but does not always have higher GC content than the other codon positions. Aota and Ikemura (1986) studied the GC content at the third codon position of various human genes and found that the GC content varies between 33-94%. Interestingly, they also found that the variation of GC content at the third codon position was correlated to (and generally higher than) the GC content of introns and bulk genomic regions surrounding a gene. They suggested that the variation of GC content likely is associated with "isochore" organization of the genome. The correlation between the third codon GC content and that of introns also was found for 117 genes in Drosophila (Kliman and Eyre-Walker, 1998). Thus, the observation that the GC content at the third codon positions of all the six human genes is higher than that at other codon positions likely is a local phenomenon and could be associated with specific isochores in which those genes reside. Compositional differences for the human and mouse isochores which harbor the genes, thus, could have contributed the observed difference in GC content between homologous human and mouse genes. To examine such possibilities, the GC contents of

introns for each gene and those of the bulk contig sequence were calculated and are shown in Table 4.5. Indeed, an overall correlation between GC content at the third codon position to that of intron and contig sequence can be discerned for the six pairs of human and mouse genes, which indicates that coding exons and introns may be subjected to similar mutational pressure (Li, 1997).

In addition to substitutions, insertions or deletions of nucleotides of syntenic human and mouse coding sequences also exist. However, insertions and deletions occur at a much lower frequency and typically are at terminal exons and in groups of three nucleotides to maintain the reading frame. The insertions or deletions in the six pairs of homologous genes investigated during this dissertation research all were located at 5' or 3' end coding exons, as for example, was observed at the 3' end coding exon of human and mouse RanBP1 gene. An alignment of the sequences at the 3' end of human and mouse RanBP1 coding region is shown in Figure 4.7.

```
mouse   AAGCCCTTTCAGTGAGGGAGGCCAGAGAGGAGGCTGAAGAGAAGTCTGAGGAGAAACAATGA
        |||| || || |||| ||||          ||| |  | ||| |  |||||||||| |||| |
human   AAGCTCTCTCGGTGAAGGAG.........GAGACCAAGGAGGATGCTGAGGAGAAGCAATAA
```

Figure 4.7 Sequences at the 3' end of RanBP1 coding regions.
stop codons are underlined.

The above alignment suggests that an internal insertion/deletion of 9 bp might have occurred after human and rodent divergence. However, whether this sequence difference is because of an insertion or deletion can not be directly determined from this alignment. However, a Genbank database search for potential homologues of RanBP1 gene from other species reveals a *Xenopus laevis* (African clawed frog) RanBP1 cDNA and several rat ESTs (including AI010261) with high sequence similarity to the RanBP1 cDNA sequences. Assuming the rat ESTs represent the homologue of the RanBP1 gene in rat, an alignment of the coding sequences from all four species, shown in Figure 4.8, reveals that the 9 bp sequence which exists in mouse but not human, is completely conserved in rat. The presence (with less conservation) of the 9 bp sequence in frog at the expected position

seems to indicate that a deletion in the human RanBP1 gene rather than an insertion in the mouse, rat and frog RanBP1 genes most likely had occurred.

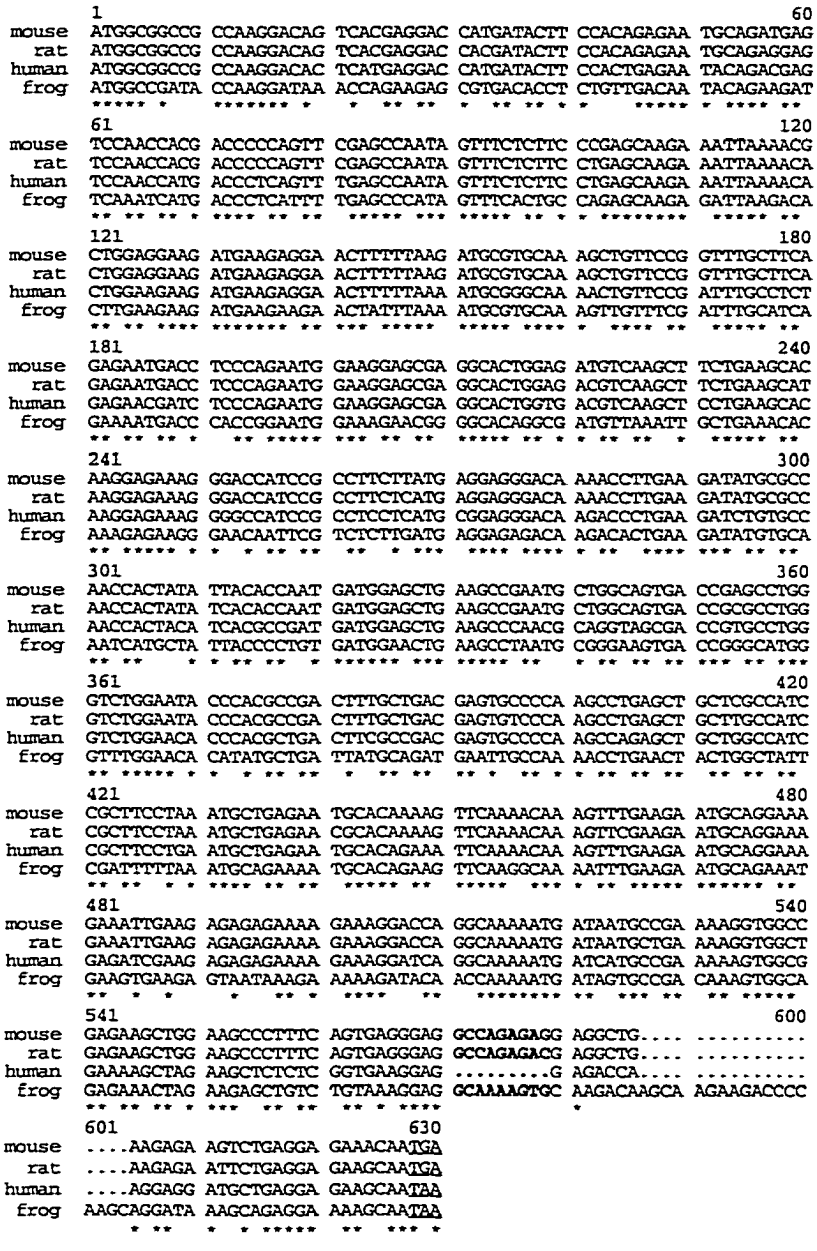


Figure 4.8 A Pileup comparison of the RanBP1 homologues from mouse, rat, human and frog. Stop codons are underlined, the 9 bases deleted in human are highlighted.

4.4 Comparison of regulatory elements

For a protein coding gene to perform properly, its product must be present in the cell at an optimal concentration and therefore, its transcription must be carefully regulated. It might be expected that mutations which destroy regulatory elements in a promoter could have detrimental effects on the cell. Thus, regulatory elements are expected to be conserved between closely related species such as human and mouse. However, unlike protein coding sequences which code amino acid residues in a specific order over a significant length, regulatory elements such as those found in promoter region of a gene likely are composed of relatively short cis elements organized in a modular fashion (Lewin, 1997). These properties suggest that conservation of promoter function may not necessitate conservation of the whole promoter region between human and mouse. Even though only a few sequence comparisons of experimentally determined promoters between human and mouse have been performed, the available data indicate that many promoter elements are conserved between human and mouse. In Section 3.4, it has been shown that the promoter region of mouse RanBP1 and Htf9c genes, which was experimentally determined, is well conserved in humans. Previous studies, including those on polo-like kinase gene promoters (Brauninger et al., 1995), type I DNA topoisomerase gene promoters (Baumgartner, 1994) and lipoprotein lipase gene promoter (Bey et al., 1998), also showed significant conservation of promoter elements between human and mouse although species specific regulatory elements also were reported (Vuillaumier et al., 1997). Thus, further research on the comparison of human and mouse promoter sequences may improve our understanding of the patterns of promoter conservation, and perhaps more importantly, identify putative regulatory elements.

Except for the promoter regions of the mouse Htf9c and RanBP1 genes, which have been examined in Section 3.4, there are no experimental data on any of the promoters of the other five homologous genes pairs. Therefore the initial step in promoter analysis is

to compare predicted promoters based on the assumption that evolutionarily conserved regions have a similar biological function.

The alignment of the predicted promoter regions of the human and mouse Dok genes shown in Figure 4.9 reveals several well conserved segments that most likely represent common transcription factor binding sites. Among these are the binding sites for 3 ATF, 2 SP1 and one ETS2 transcription factors. The 3 ATF (activating transcription factor) binding sites are tandemly arranged, with the consensus sequences spaced by 12 and 11 bps. Since one ATF site occupies about 20 bp nucleic acid sequence (Lewin, 1997), this arrangement would allow three ATFs to bind to the promoter sequence without space conflicts. The ATF transcription factor family members typically form homodimers or ATF-Jun heterodimers which bind to cAMP-responsive elements (CRE). Such elements are present in the promoter regions of many genes that are involved in skeletal and central nervous system development (Reimold et al., 1996). Even though the exact function of the human and mouse Dok genes is unknown, it is known that the Dok proteins are rapidly phosphorylated after cell transformation and have many of the features that usually are associated with signaling molecules (Section 3.6). The presence and conservation of the three ATF binding sites (cAMP-responsive elements) suggests that not only the phosphorylation of Dok protein but also the transcription of the Dok gene likely are under the control of a signal transduction pathway. In addition to the conserved putative transcription factor binding sites, there also are several other predicted sites which are not located within conserved sequence segments. Some of them, such as a ETS1 and a SP1 binding sites, are predicted in both human and mouse although at different positions relative to other elements while still others are predicted only in human or mouse. Further experiments will be needed to determine if these other elements are functional and whether they constitute differences in the regulation of the Dok gene transcription between human and mouse.

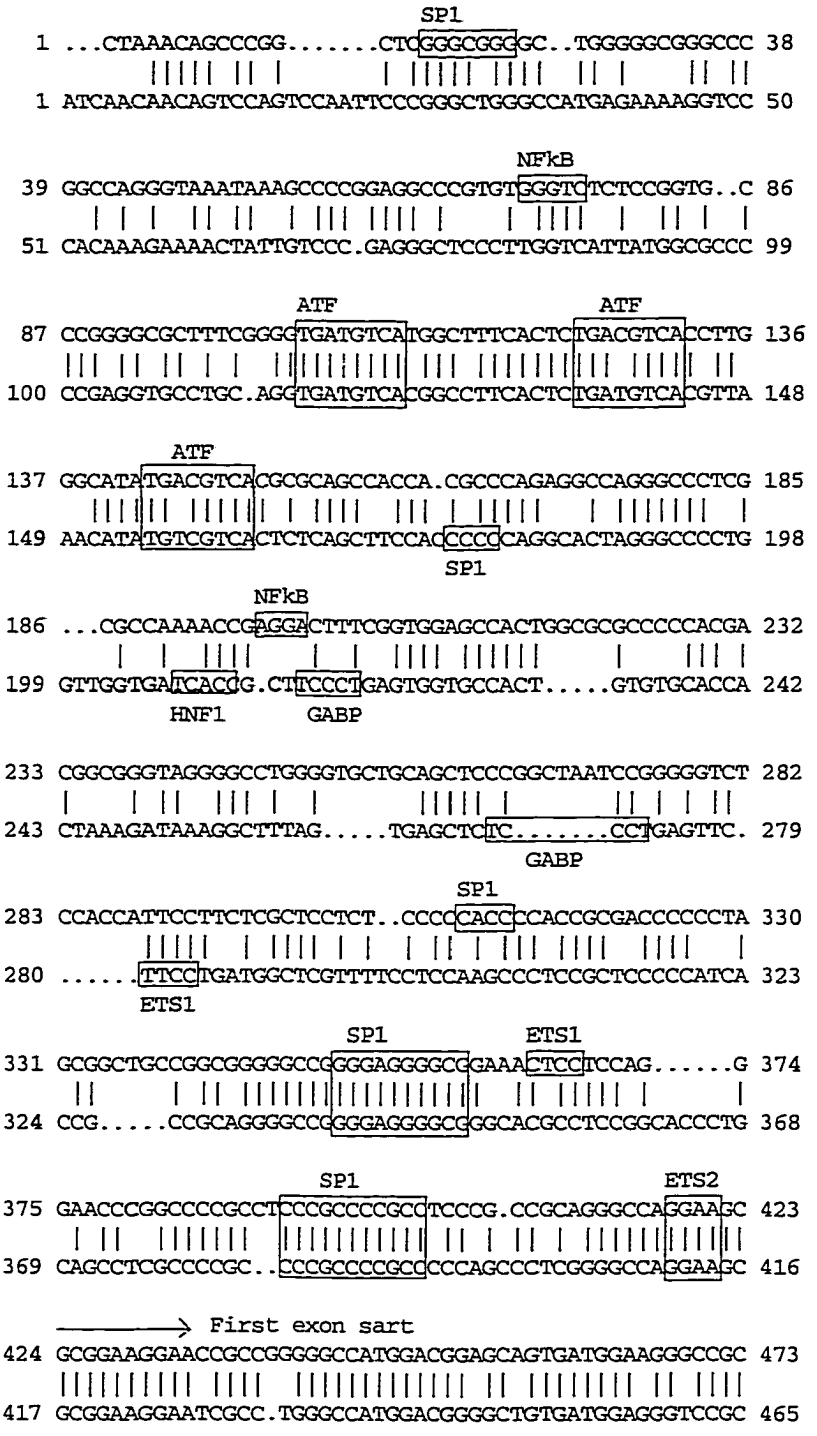


Figure 4.9 Alignment of predicted promoter region of human/mouse Dok genes. Transcription factor binding sites are predicted by Consinspector.

A somewhat different pattern of conservation is observed in the predicted human and mouse T10 gene promoter region (Figure 4.10). The human T10 promoter region is located in cosmid 68a1 which previously has been sequenced (Chen, 1997). For further

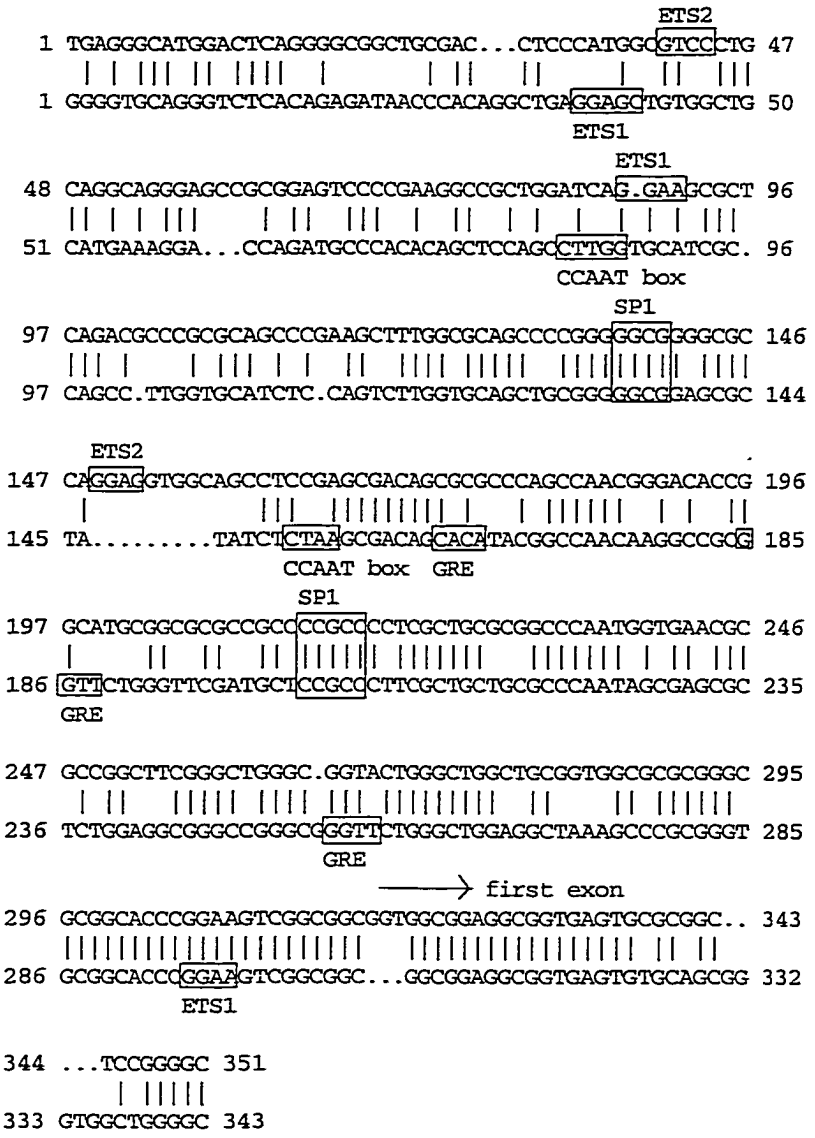


Figure 4.10 Alignment of predicted promoter region of human/mouse T10 gene

analysis, the region surrounding the 5' end exon of the human T10 gene was extracted from the sequence and studied. As in the mouse T10 gene promoter region (Section 3.4), the human T10 gene promoter also is located in a CpG island and no TATA box is found within this predicted promoter region (also see Chen, 1997). As can be seen from Figure 4.10, the promoter region also contains several well conserved sequence segments.

Specifically, two SP1 binding sites are predicted within these segments both in the human and mouse sequences but the three GRE binding sites and two CCAAT boxes that are predicted in mouse are not predicted in human. In contrast, two ETS2 binding sites are predicted in human but not in mouse. Since these different elements constitute the majority of the predicted sites both in human and mouse, it might be possible that the human and mouse T10 gene transcription are controlled by different sets of regulatory elements.

A similar pattern for the distribution of regulatory elements is present in the predicted promoter regions of the human and mouse PutA genes (Figure 4.11). Both of the promoter regions are located in CpG islands without any recognized TATA box nearby. Although the sequences show about 70% identity, only three transcription factor binding sites, namely an SP1, a CCAAT box and a GABP binding site occur in conserved regions of the human and mouse genomic DNA. In addition, although five NFkb binding sites are predicted in the human sequence, none are predicted in mouse and instead, four additional GABP binding sites are predicted in mouse than are predicted in human. Thus, it seems that species specific regulatory elements dominate the promoter region of both the human and mouse PutA genes.

It should be noted that the current version of Consinspector includes libraries for 16 families of transcription factor binding sites. It is likely that other regulatory elements also are present in the analyzed sequences but not predicted. Although some of the above predicted sites may not be functional, comparing the promoter regions of the five pairs of genes (the promoter region of human Gene7 likely is not located in the sequenced region and thus unavailable for comparison) shows that the human and mouse promoter regions of different genes are conserved at different levels of nucleic acid identity, ranging from about 90% in the promoter regions of the Htf9c and RanBP1 genes, to about 65% in the T10 gene promoters. Because many predicted transcription factor binding sites are highly conserved, such as the three tandem cAMP response elements in the promoter region of Dok genes, these elements most likely have similar functions.

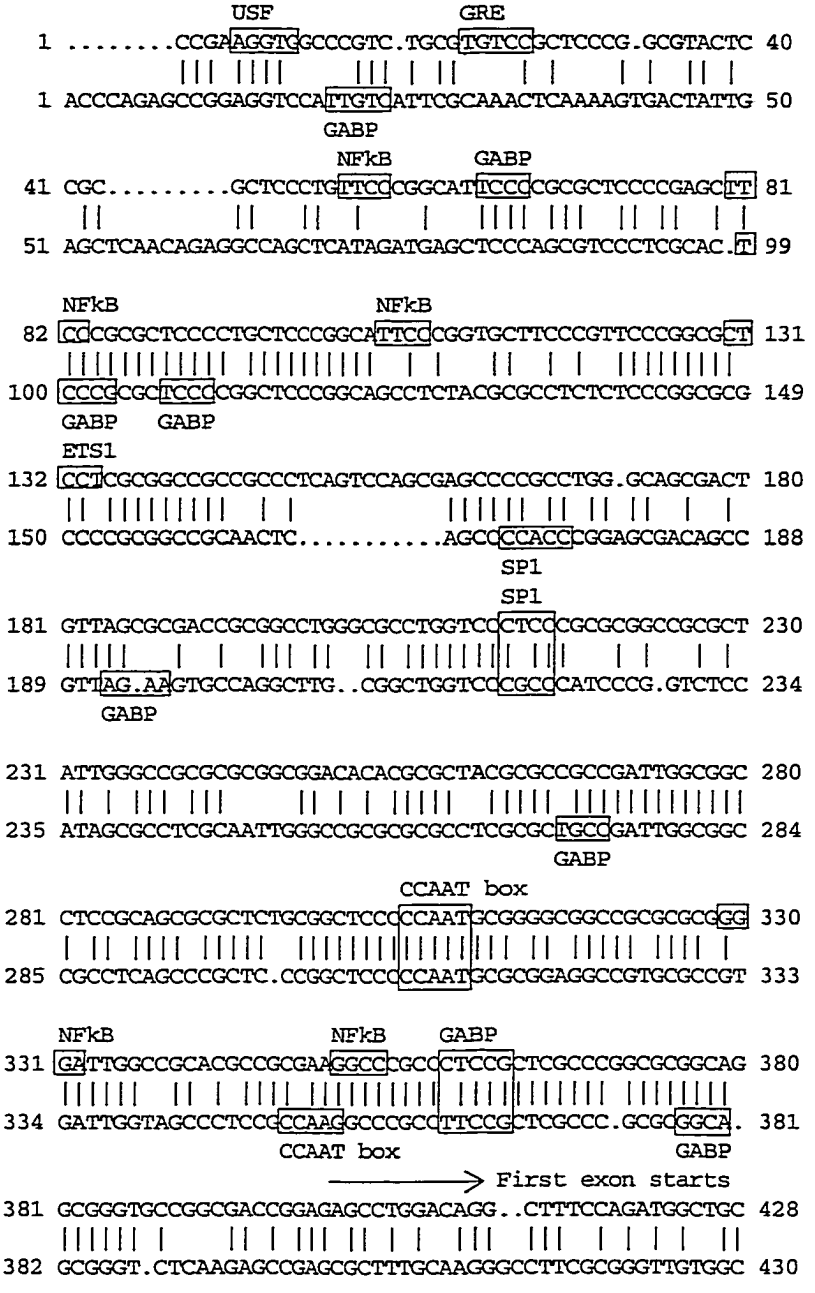


Figure 4.11 Alignment of predicted promoter region of human/mouse PutA genes

In addition to promoter elements, polyadenylation signals of the compared genes also are well conserved between human and mouse. Table 4.6 shows the sequences at the 3' ends of each of the six gene pairs. The conservation is interesting in two aspects. First,

when the gene in one species has an atypical polyadenylation signal such as ATTAAA, the homologous partner also has the same sequence. Second, although significant sequence similarity seems not to exist outside the canonical sequence AATAAA, or its variant between different homologous gene pairs, the sequence similarity within homologous gene pairs often extends beyond these six base pairs. A detailed alignment of the 3' UTRs of the six gene pairs show that these sequences show varying degrees of conservation in terms of both conserved length and sequence similarity. On one extreme, the human and mouse RanBP1 gene pair contains the most conserved 3' UTR (Figure 4.12). About 100 bp of sequence surrounding the terminal putative polyA signal is exactly conserved between human and mouse. Sequences of about 100 bp around the other polyA signal also are well conserved, although to a lesser extent. The high conservation between the RanBP1 genes might be related to the alternative polyadenylation, as detailed in Section 3.3. It might be that additional regulatory elements, such as protein binding sites may be required to control the process of alternative polyadenylation. This would result in a higher conservation of sequences surrounding the alternative polyadenylation signal than might be seen with other gene pairs which contain only a single, more typical polyadenylation signal. The other extreme occurs at the 3' end of human and mouse Htf9c genes. Although an 11 bp sequence spanning the polyadenylation signal is conserved between human and mouse (see Table 4.6), this segment is the most conserved segment within the 3' end 200 bp of both genes. The level of sequence conservation outside the AATAAA signal reinforces our current understanding that sequences upstream of ATTAAA may enhance the efficiency of polyadenylation but they are not essential for the process (Wahle and Keller, 1996).

Table 4.6 Comparison of sequences around polyadenylation signals

gene pair	putative polyadenylation signal sequence
hDok	TTTAA.AAATTTTAAAATTCTCATTAAAGTCAGCTTGGGTTTAAG
mDok	TTTAACTTATAAGACAAATCTCATTAAAGCCAGCTTGGGTTG...
hGene7	GCTACTGTTTACACAGGTGAAGATTAAACACCCAGTAAGCTTCT
mGene7	GATAGTGTTTACACA...CATAATTAAATGCCCAGTAAGCCTCC
hT10	TACCTGAATGTGAGCTCCTGATAATAAAACTCTGAGGCTTT..
mT10	AACTTGAATGTGCACAGTTGATAATAAATCTTTGAAGTTGTAA
hHtf9c	TAGCCAAGGCTGTGAGGGCCAGAATAAACAACTGCTCAACAAAGG....
mHtf9c	..................GAGAATAAACAATTGCTGATTTTACTAACA
hRanBP1	TTTATAGTCATCAAAAATAGTGAATAAAAAACACATTTGGAACCTGGG.
mRanBP1	TTTATAGTCATCAAAAATAGTGAATAAAAAATGCCTTTGAAAACCTGGA
hPutA	GCGAGCCTTGATTGTCTGAACACATAAAGCAAACTGTCCAGAA...
mputa	GTGTACCCTGATGGTCTGAACACATAAA....ACTGTCCAAAAGGG

```
                .         .         .         .         .
750 TAAATCGTCTTATTTTATTTTCTTTTCCTCTCTTTCCTTTCCTTTTTTTAA 800
    | |                 ||  |   ||| ||||||||||||||| ||
734 TGA.................ATCACTCTGTCTTTTTCCTTTCCTTTTCTTTT 784
                .         .         .         .         .
801 AAAATTTTACCCTGCCCCTCTTTTTCGGTTTGTTTTTATTCTTTCATTTT 850
     |||  ||    |||||  |  |||  | ||  ||||  | |||  ||||
785 TAAAAATT....TGCCCTACCCTTTAAGGTTTGTTTTTCTGTTTTGTTTT 830
                .         .         .         .         .
851 TACAAGGGACGTTATATAAAGAACTGAACT.CAACATTCAGGTTGTTTTT 899
    ||||||||||  | ||||||||||||||| | |||  ||||||||||||||
831 TACAAGGGAC..TTTATAAAGAACTGAATTCCAATTTTCAGGTTGTTTTT 878
                .         .         .         .         .
900 TTTTTTTGTT.....................TCTAAGTTTTTGCCCTATT 928
    || |||||||                     || |||||||  | | |||
879 TTGTTTTGTTTTTTTTTTAACTTTTTTTTCCTCCAAGTTTTGACACCATT 928
                .         .         .         .         .
929 GAAGATGACTTCAGAAAATCCATTCCCCAGTCATGAAAATGTACTGTGCT 978
    ||| |||||||||| ||||||||||||||||||||||||||||||||||||
929 GAACATGACTTCAG.AAATCCATTCCCCAGTCATGAAAATGTACTGTGCT 977
                .         .         .         .         .
979 AACTTTCTTTTCCATAGTGGAAACACTTATTTATAGTCATCAAAAATAGT 1028
    ||||||||||||||||||||||||||||||||||||||||||||||||||
978 AACTTTCTTTTCCATAGTGGAAACACTTATTTATAGTCATCAAAAATAGT 1027
                .         .
1029 GAATAAAAAACACATTTGGAACCTGGG. 1055
     ||||||||||  | ||||  || |  ||
1028 GAATAAAAAATGCCTTTGAAAACCTGGA 1055
```

Figure 4.12 Alignment of the 3' UTR of human and mouse RanBP1 gene.
Putative polyadenylation signals are underlined.

4.5 Comparison of introns

Previous studies (e.g. Lamerdin et al., 1995; Chen, 1997) have reported that relative intron sizes of homologous genes are generally conserved between human and mouse. This observation also is supported by comparison of intron sizes of the six pairs of homologous genes sequenced in this dissertation research (Figure 4.13). However, significant size differences between homologous introns do exist. Two striking examples occur in the first and sixth introns of the almost identically sized human and mouse Gene7 genes. The first intron of the human and mouse genes are 24886 and 11264 bp, respectively, while the sixth intron of the human and mouse genes are 13012 and 26989 bp, respectively, with the mouse intron more than twice of that in human. Examination of the sequences in those intron regions reveals that the first intron of the human gene contains two pseudogenes (total about 6.5 kb, see Section 3.6), which are not found in mouse, while an IAP element is found in the sixth intron of the mouse gene (about 7 kb, see Section 3.7), which is absent in human (see Figure 4.4). Although insertions of these pseudogenes/transposons into the respective intron regions do not explain all the differences between these introns, they clearly make a significant contribution to the observed differences in these introns while maintaining the approximately equally sized total (exon and intron) gene sequences.

Examination of intron sequences through Dotter analysis shows that for the six gene pairs, they have varying degrees of conservation at the nucleic acid level. The Dok gene pair introns have a relatively small size and moderately well conserved sequences (Figure 4.14), as revealed by the multiple short segments (around 30 bps) distributed along the diagonal line in Figure 4.14. Alignment of these intron sequences (Figure 4.15) showed that they had an identity of 58.2% and that a similar level of conservation also occurs in both the 5' and 3' UTRs (see Table 4.1).

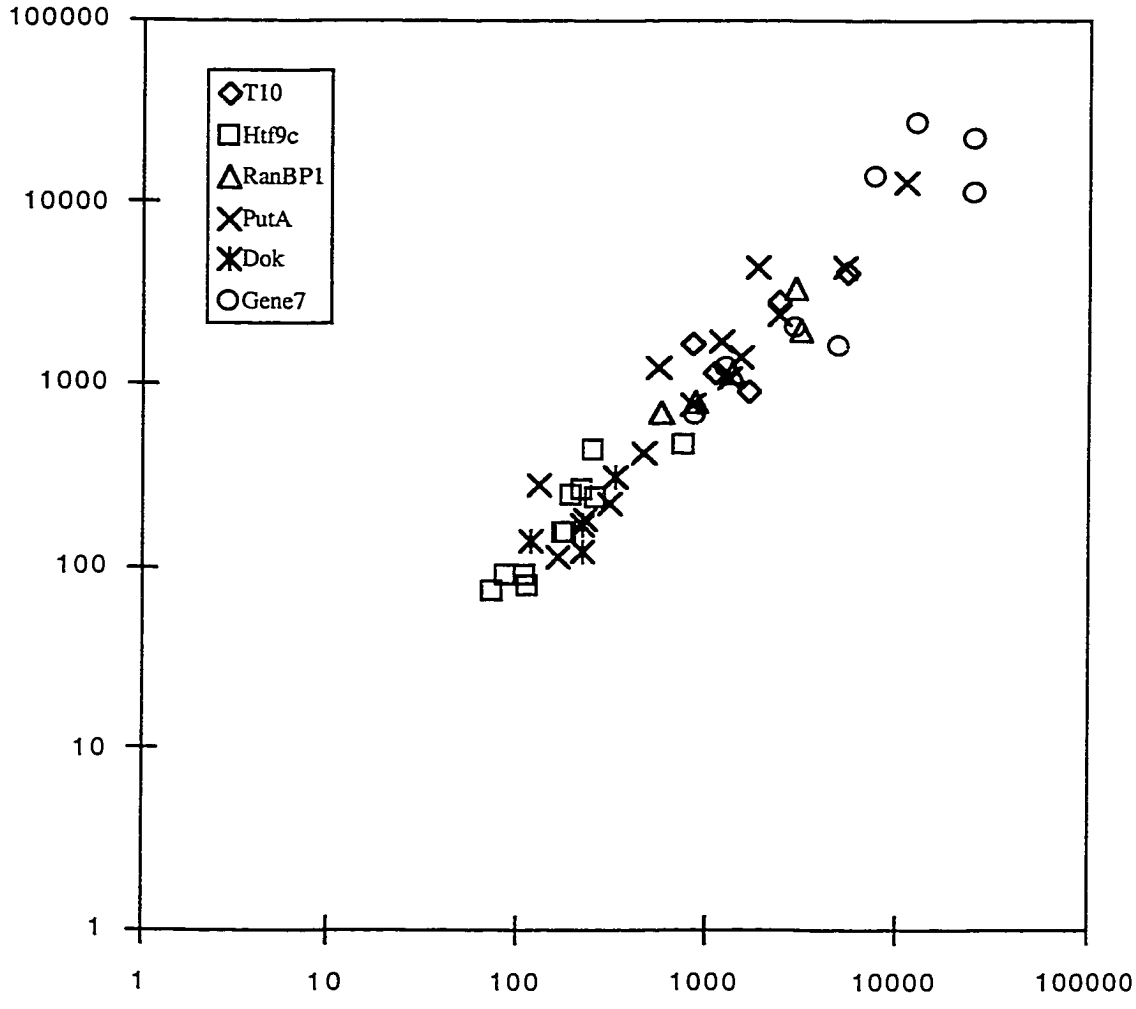


Figure 4.13 Comparison of intron sizes between human (horizontal) and mouse (vertical) genes

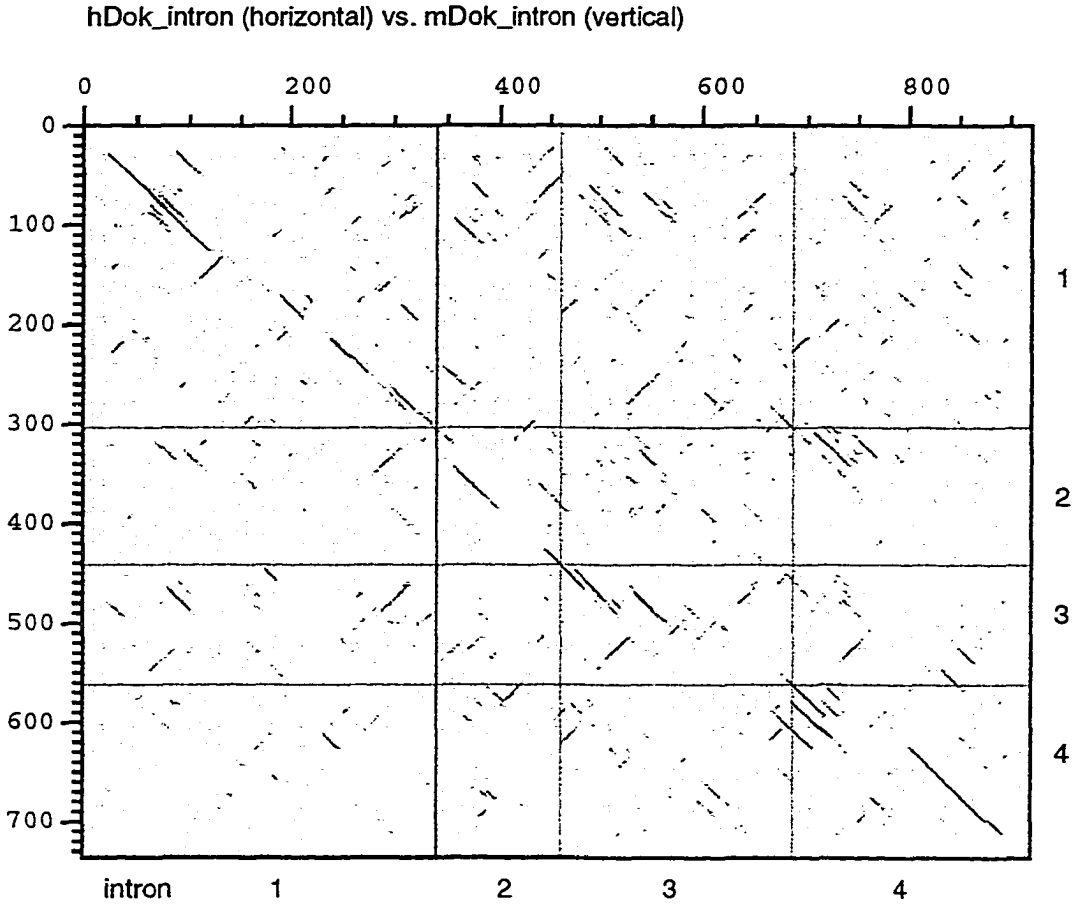


Figure 4.14 Dotter plot between intron sequences of human and mouse Dok genes

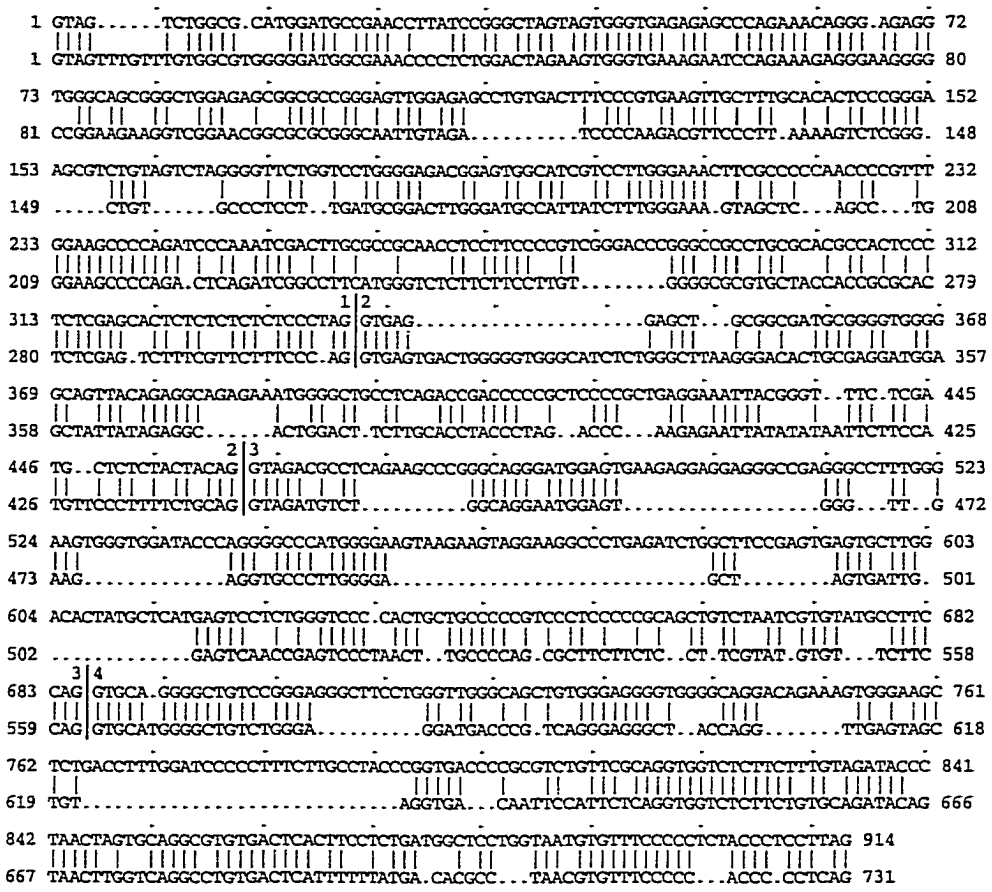


Figure 4.15 Alignment of intron sequences of human and mouse Dok gene
The boundaries for each of the four introns are as indicated.

The intron sequences of the other gene pairs subsequently were examined in a similar fashion. In contrast to the results for the Dok gene pairs, a Dotter analysis revealed a relatively low level of sequence conservation, as only relatively short segments of homology were observed distributed sparsely along the diagonal. However, these short conserved segments are present in every gene pair although not necessarily in every intron pair as short conserved segments are found in all introns of the Gene7 and T10 gene pairs, in most introns of the RanBP1 gene pairs, and about half the introns of the Htf9c and PutA gene pairs. A typical example of introns with only slight but detectable homology is shown in the Dotter plot of the eight intron regions of the human and mouse Gene7 genes shown

in Figure 4.16. Although little is known as to why there is a relatively high degree of identity of some comparable introns and a low identity of other comparable introns, it has been hypothesized that these variations may be due to differences in the evolutionary rate for the intron regions of different gene pairs (see below). It also may be that shorter introns evolve slower than longer introns and the introns in the Dok genes are the shortest (average about 200 bp) among all the six homologous gene pairs. Alternatively, introns with relatively high sequence similarities may contain regulatory elements which are required to be maintained. Although neither idea can be proven, neither can be excluded. Finally, it is interesting to note that no repeated sequences of any type are detected within the Dok gene introns. The effect of repeated sequences on the degree of apparent conservation is expected to be significant since both of the human and mouse sequences contains abundant short interspersed repeats, many of which most likely were inserted into the human and mouse genome after their divergence. In addition, insertion of pseudogenes and transposons that occurred after human and mouse divergence obviously would result in different intronic sequences.

Since repeated sequences inserted into the human and mouse introns after their divergence would result in apparent non-conservation, it would be interesting to examine the degree of conservation when repeated sequences are excluded from the comparison because the resulting sequences would be derived, with a larger proportion, from the intron sequences before the human and mouse divergence. For this purpose, the repeated sequences that could be identified by Repeatmasker were masked from all the intron regions, and then a dotter analysis was performed. The apparent conservation level between intron regions of homologous genes are significantly improved for all the remaining gene pairs except for the Htf9 gene pairs, which do not contain detectable repeat sequences. An example of this analysis for the human and mouse Gene7 gene pairs is shown in Figure 4.17. The shifts of diagonal traces within introns 1 and 6 in Figure 4.17 partly are attributed to the two pseudogenes inserted in intron 1 of human and one

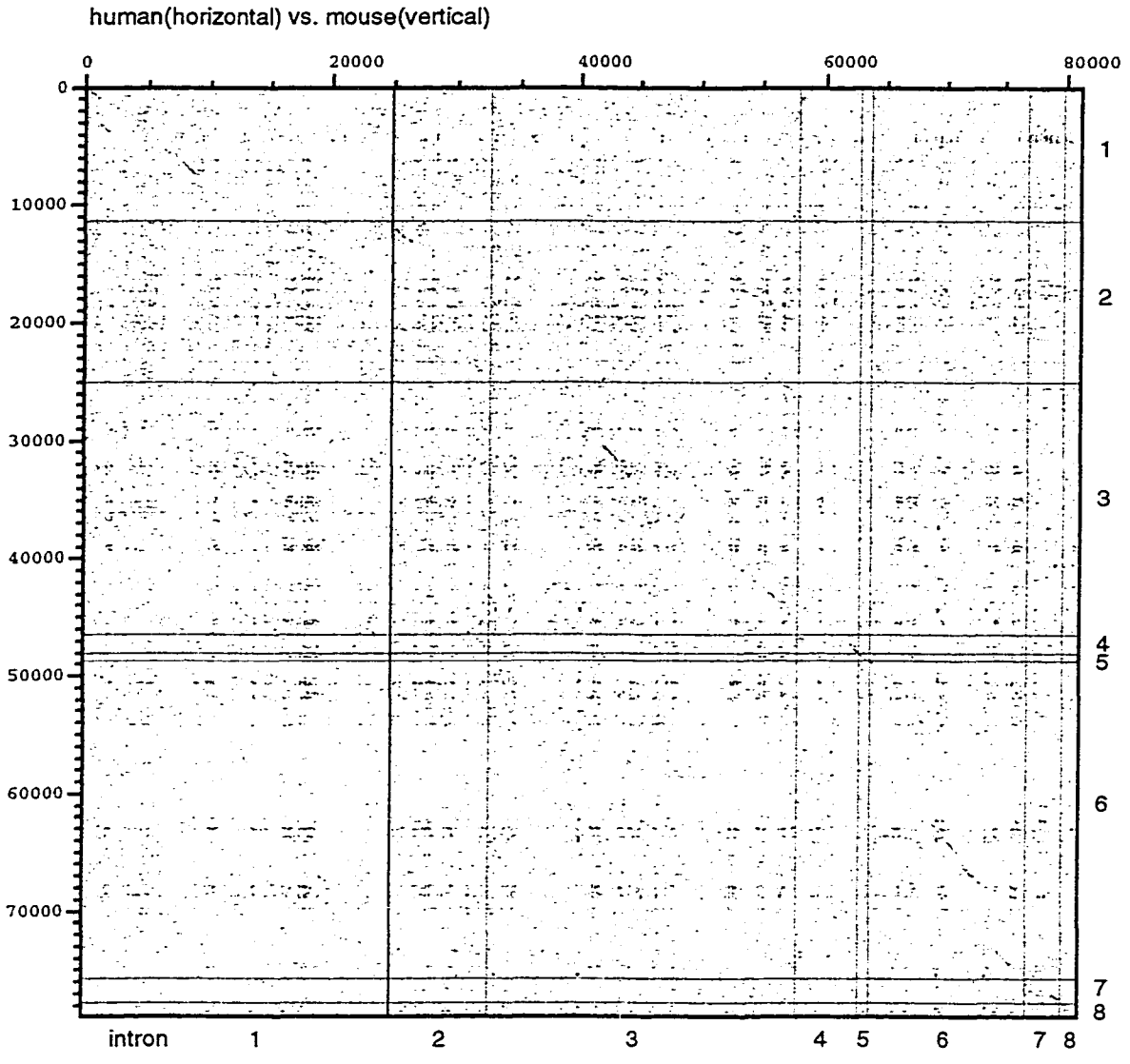


Figure 4.16 Dotter plot comparing the intron sequences of the human and mouse Gene7

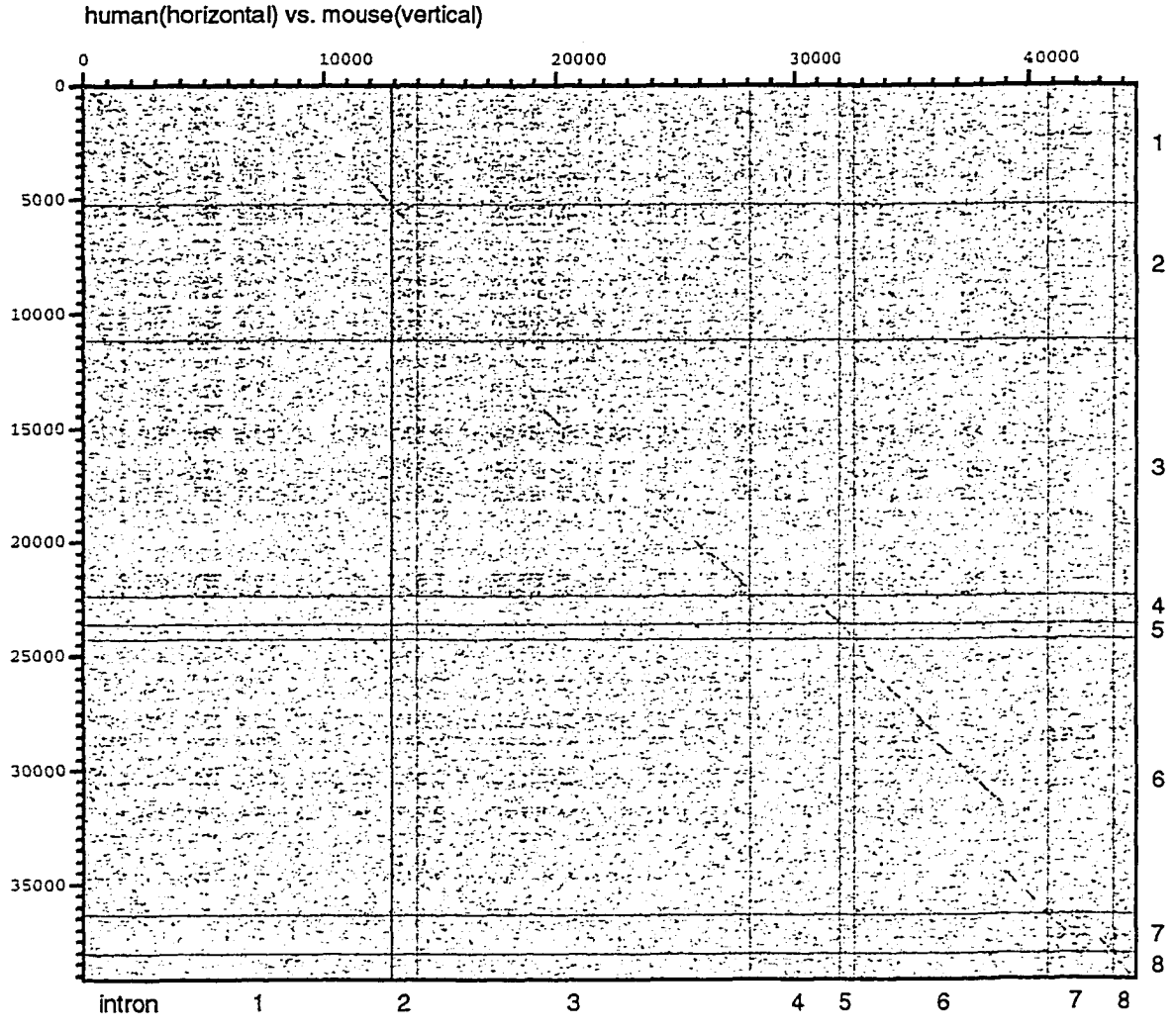


Figure 4.17 Dotter plot of the intron sequences (with repeats removed) of human and mouse Gene7

transposon inserted in intron 6 of mouse Gene7. It thus can be concluded that in at least four pairs of homologous genes from this dissertation research, the presence of repeat sequences is one of the major factors responsible for the apparent absence of conservation of intron sequences. After assessing the effect of repeats, it should be noted that only a few introns (with known repeats removed) do not show observable conservation between human and mouse, as is found, for example, in several introns (introns 9-14) of the human and mouse PutA genes (Figure 4.18). The reason for the lack of observable conservation is not clear. However, it could partially be due to the possible presence of unidentified repeat sequences and/or pseudogenes/transposons in those introns. Examination of synonymous substitution rate of internal exons does not reveal a systematic increase for those exons adjacent to the nonconserved introns, and thus, there is no evidence for a higher evolution rate of those nonconserved introns than the moderately conserved introns.

Overall, a comparison of intron sequences of the six pairs of homologous genes shows that most introns (37/47) are weakly to moderately conserved between human and mouse, if repeat sequences and pseudogenes/transposons are excluded from the comparison. This observation may explain the general conservation of the lengths of introns if we assume that the human and mouse received a similar amount of insertions of interspersed repeats after their divergence.

Previous studies have shown some strikingly different results regarding conservation of intron sequences. Koop and Hood (1994) compared about 100 kb of human and mouse sequences in a T-cell receptor region. They showed that although the coding sequences only account for <6% of the studied region, the 100 kb sequence from human and mouse has a similarity of about 71%. Similarly, a high level conservation of intron sequences also was found in the human and mouse Bruton's tyrosine kinase loci (Oeltjen et al., 1997) as well as recently in a gene-rich region in human chromosome 12 syntenic to mouse chromosome 6 (Ansari-Lari et al., 1998). In contrast to these studies and partly consistent with this present work, little or no conservation of intron sequences was

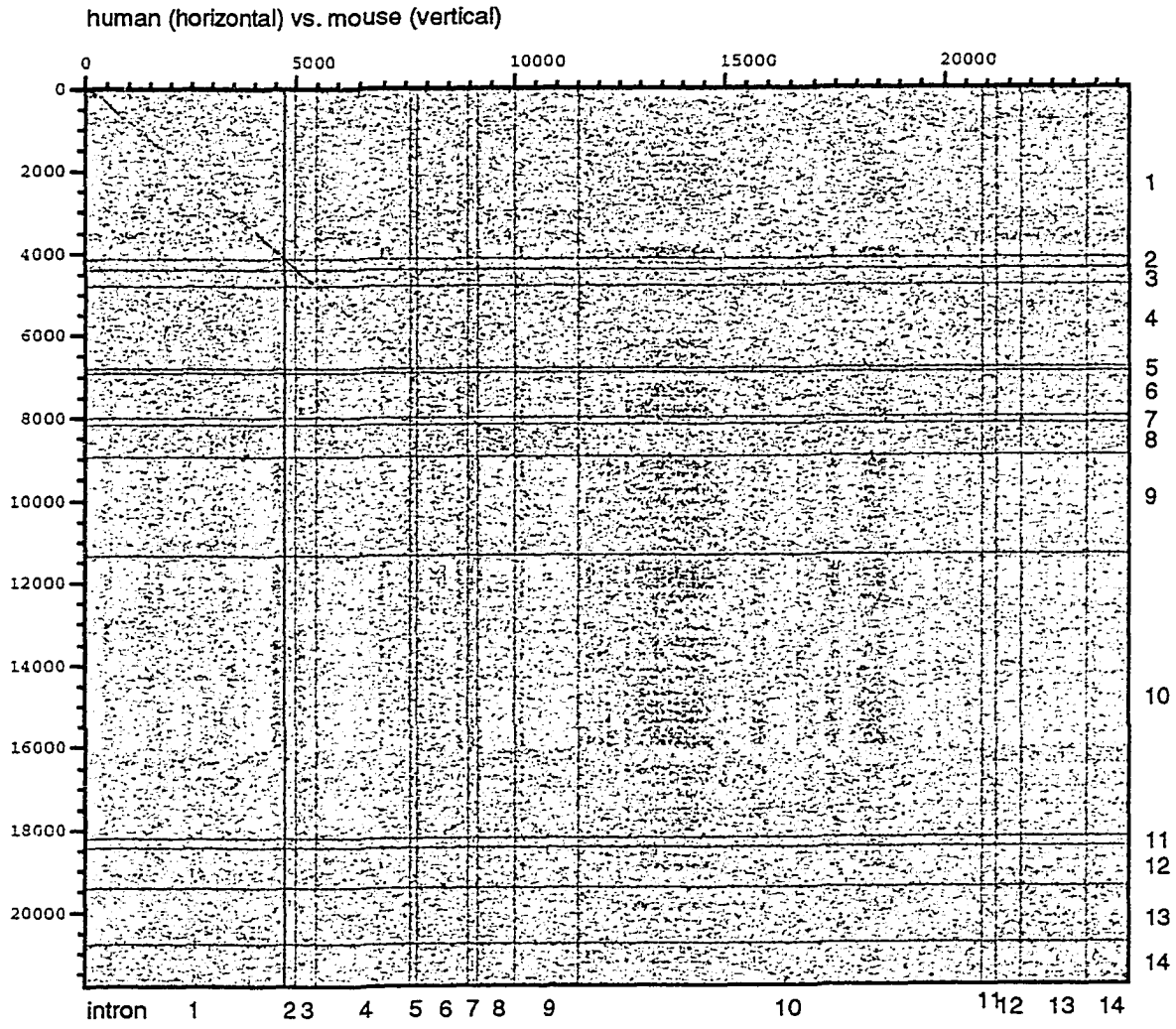


Figure 4.18 Dotter plot comparing the intron sequences of human and mouse PutA genes with repeat sequences identified by Repeatmasker removed

found in the human and mouse beta-globin gene cluster region (Shehee et al., 1989), XRRC1 DNA repair gene regions (Lamerdin et al., 1995) and ERCC2 region (Lamerdin et al., 1996). Based on the observed differences, Koop et al. (1995) proposed a mosaic model of genome evolution in which different regions of human and mouse genomes most likely have "evolved" at different rates. Based on the limited number of sequence comparisons performed to date, this prediction seems unlikely because the extent of intronic sequence differences in different genomic regions may vary depending on the need for maintaining, or not maintaining a region which may or may not have biological function.

The human and mouse genomic regions spanning the six homologous gene pairs contain four intergenic regions (Table 4.7) whose sizes are quite well conserved. However, at the sequence level, one intergenic region (between Dok and Gene7 genes) has only a low to moderate sequence similarity while two other intergenic regions have very little sequence similarity. Therefore, it is likely that the conserved regions might represent sequences needed for appropriate gene expression beyond the typical promoter domains while the other non-conserved intergenic regions do not contain sequences with significant functional relevance.

Table 4.7 Sizes of intergenic regions between adjacent genes

	Dok-Gene7	T10-PutA	PutA-Htf9c	Htf9c-RanBP1
Human	902 bp	14366 bp	426 bp	0 bp
Mouse	772 bp	11615 bp	444 bp	0 bp

4.6 Comparison of repetitive elements

Table 4.8 summarizes the level of repetitive elements within the two syntenic regions in human and mouse and reveals several notable features. First, different genomic regions contain different amounts of repetitive elements and the tendency for a region to have more repeats seems to be conserved between human and mouse. Specifically, the first

syntenic region has a lower amount of total repeats (32.2 % in human and 25.4 % in mouse) than the second syntenic region (42.9% in human and 46.8 % in mouse). This observation may be related to the gene density of the two syntenic regions as the first region contains four genes while the second contains only two. It may also be due to the presence of common repetitive elements inserted before human and mouse divergence (such as L1 repeats discussed below). Second, in terms of total repeats, no tendency is observed for either mammalian species to have more repetitive elements than the other species, although the mouse seems to contain more simple repeats than does the human. Third, any significant variation in the level of repeats between the two syntenic regions mainly is because of differences in the amount of L1 and LTR repeats. In Chapter 3, it was observed that L1 repeats are frequently oriented in the opposite direction of gene transcription, with the notable exception of the human and mouse Gene7 regions even though comparison of the distribution of L1 repeats in the human and mouse Gene7 region (Figure 3.37 and Figure 3.45) shows that the exceptions occur in the same intron regions. Therefore, it might be that these L1 repeats were inserted before the human and mouse diverged.

Table 4.8 Summary of repetitive elements within syntenic regions

	Region 1 (p201m18-72k21)		Region 2 (h173-b245c12)	
	Human	Mouse	Human	Mouse
Alu (human) B1+B2 (mouse)	16.3 % (51)	17.0 % (97)	16.0% (56)	13.6 % (80)
L1	5.8 % (9)	0.3 % (2)	17.8 % (28)	8.8 % (19)
LTR	5.5 % (8)	3.9 % (20)	3.1 % (5)	15.7 % (29)
MERs	1.8 % (4)	0.5 % (3)	2.9 % (11)	0
Simple repeats	1.1 % (13)	2.0 % (26)	0.7 % (12)	4.2 % (52)
Total repeats	32.2 %	25.4 %	42.9 %	46.8 %

Note: numbers in parenthesis represent numbers of respective repeats found.

Chapter V
Gene Prediction

5.1 Overview

One major goal of genomic sequencing is to locate the structural features that can be correlated directly with genomic functions. Current methods of predicting genes and other genomic features include computation-based methods such as Genscan and Xgrail, and database similarity searches.

Computational methods have been quite successful in recent years to decipher the complexity of vertebrate genomes. Claverie (1997) evaluated the performance of various programs and found that the most accurate program available only can perfectly locate approximately 80% of all the internal coding exons while being less successful at predicting the 5' and 3' end exons that contain untranslated regions. Most of the available programs evaluate several parameters in their prediction process. These parameters include (1) codon usage; (2) start, stop and splice sites; (3) open reading frames and (4) other signals such as promoter and polyadenylation signals. Unlike gene prediction in bacterial sequences, open reading frames are not very useful in predicting vertebrate genes since the average internal exon size typically is only approximately 150 bp while every kb sequence is expected to contain open reading frames larger than 150 bp just by chance (Claverie, 1997). Therefore, the most often used parameter to determine whether a region is coding or non-coding is the coding measure which, in most cases, is based on the observation that exons and introns exhibit different codon usage pairs or hexamer frequency (Claverie and Bougueleret, 1986; reviewed in Fickett and Tung, 1992). Thus, for an exon to be correctly predicted, it should have a hexamer or "word" composition similar to that of known genes compiled into a training set for each organism. Therefore, if a non-coding region has a word composition similar to known genes, it often will be falsely predicted as an exon. An

example of such false prediction is the falsely predicted exons within the (GGT)n repeat region in PAC p201m18 (contig 3, see Chapter 3) by both Genscan and Xgrail. To further improve the accuracy of prediction, it is desirable to evaluate predicted coding regions by a fundamentally different criteria. During this dissertation research, the hypothesis that the conservation of coding region between human and mouse can serve as a basis for coding region prediction by comparative gene analysis was tested.

Coding region sequence conservation between different species previously has been used to predict genes whose transcripts could be isolated experimentally. For example, Page et al. (1987) found that a nucleic acid probe from human Y chromosome cross-hybridized to a region of Y chromosome of other mammals including mouse. Subsequently, the sequence of the probe led the authors to describe a new gene termed zfy. However, for gene prediction based on coding region conservation to be useful, it should be assumed that most if not all genes are conserved between the human and mouse, and available data do suggest that this is reasonable. Even though there only is a limited number of comparative genomic sequences available, a comprehensive study (Makalowski et al., 1996) of 1196 orthologous human and mouse full-length cDNA sequences showed that the coding sequences were well conserved, with an average identity of about 85% both at the nucleic acid and amino acid levels. Although sequence conservation between closely related species such as human and mouse has been suggested as a source of information for gene prediction (Ansari-lari et al., 1998), there has been no widely available program which takes advantage of comparative sequence information.

5.2 Prediction of exons and regulatory elements

As discussed above, the strategy for predicting coding regions is to find segments of a nucleic acid sequence which maintain an open reading frame, code for an amino acid sequence that is conserved between human and mouse and is bounded by start, stop or consensus splicing site sequences (Figure 5.1). Since interspersed repetitive elements

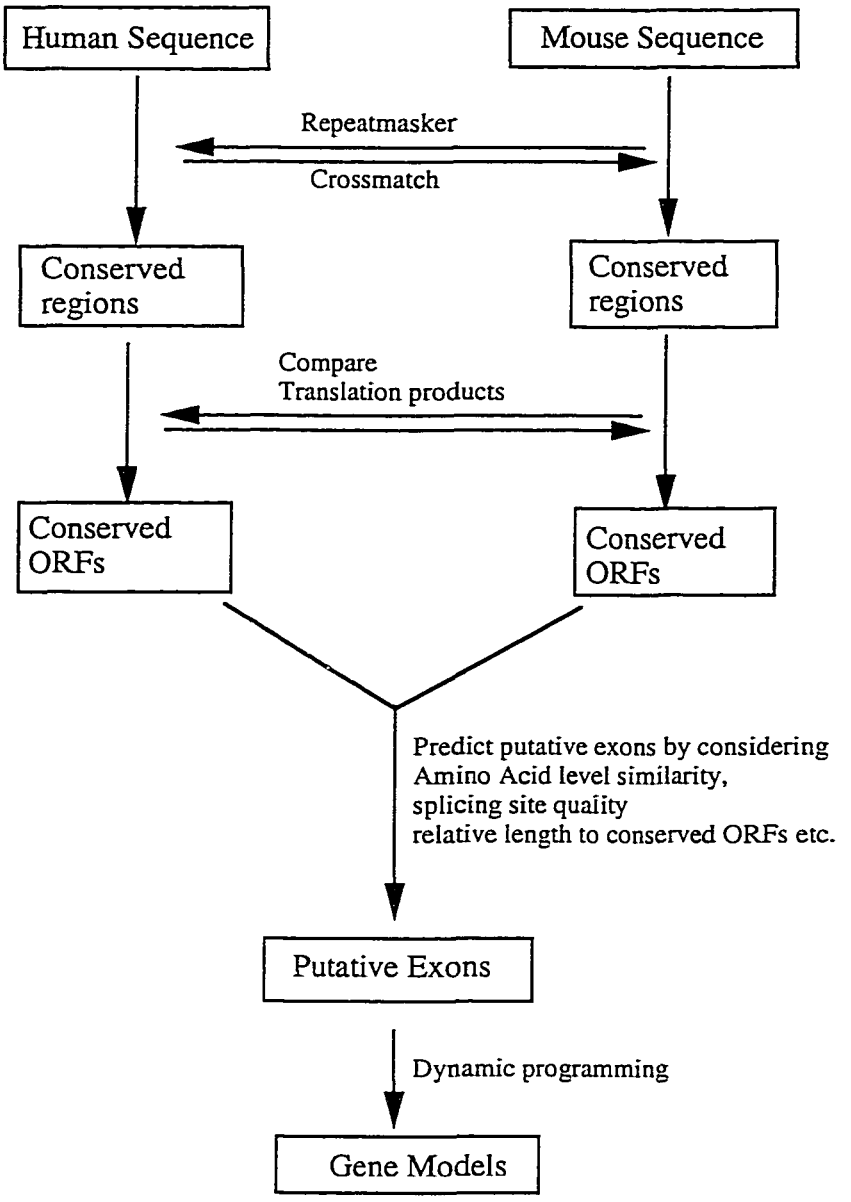


Figure 5.1 Diagram showing the procedures used in comparative gene prediction by Geneligner

generally do not occur in gene coding regions (Adams et al., 1995), the first step in the gene prediction approach developed during this dissertation research was to mask the repetitive sequences with Repeatmasker. The conserved nucleic acid segments then were identified by Crossmatch with the minimum match length changed from the default 14 bp to 10 bp. Subsequently, the sequences in and around the Crossmatch-identified conserved segments were evaluated as potential exons by determining the following: (1) sequence similarity between human and mouse at the amino acid level; (2) splicing site quality; (3) change of sequence similarity surrounding a splicing site; (4) relative length of a potential exon when compared to the conserved ORF length; (5) percentage of substitutions at different codon positions; and (6) conservation of exon length and reading frames.

Several assumptions typically were made when investigating potential exons. First, although the length of coding exons vary and can be very small, a minimum exon length of 30 bp was assumed since the vast majority (>98%, based on results of Zhang, 1998) of human exons are larger than 30 bp. Second, exons from different strands do not overlap, i.e. only one potential exons in either of the two strands would be predicted; Lastly, the splicing sites conform to the AG/GT rule and the splicing site quality was calculated using a Weight Matrix Model (WMM, Staden, 1984) where the nucleotide distribution around a potential splicing site is compared to the distribution of nucleotides at known splicing sites. The profile data used for known splicing sites was obtained from 1254 splicing sites of 238 multi-exon genes (Burge, 1997). The qualities for translation start sites (ATG) and stop sites also were calculated using the WMM model with profile data from Zhang (1998). In calculating an overall score, all parameters were weighted to obtain the overall probability for a potential exon. Here, the weight for each parameter was empirically determined such that an overall high accuracy was obtained when applied to a training set of sequences (see below). When several overlapping but not identical exons were predicted with similar probability of being candidate exons, all were chosen as potential exons and further analyzed by building gene models.

Figure 5.2 illustrates the typically chosen parameters for an example of known genes, i.e. the human and mouse Dok genes, which are located in contigs 5 and 6 (Sections 3.6 and 3.7). As expected, the coding exons have a relatively high level of amino acid level similarity when compared to non-coding regions and most known splicing sites also have scores higher than any pseudosites. By using just the two parameters of amino acid level similarity and splicing site score, more than half of the exons in the training set could be identified correctly. Examination of the false predictions revealed that errors often occur when high score pseudo-splicing sites occur. For example, if a pseudosite was located in a less conserved region outside of a large exon, the overall amino acid level similarity would be only slightly lower than that in the true exon causing this pseudosite to be incorrectly chosen in the final prediction. On the other hand, if the pseudosite is located within an exon, it may be falsely chosen as the final prediction since a potential exon with this false site may have an equal or even higher amino acid similarity than the true exon. To avoid such false predictions, the third and fourth parameters, i.e. the level of sequence similarity surrounding a splicing site and the relative length of a potential exon as compared to the length of conserved ORF were considered. The other parameters, such as the percentage of third codon substitutions also were designed to eliminate false predictions since, as detailed in Chapter 4, substitution at the third codon position generally contribute more than those of the other codon positions to the total substitutions in a known coding region.

In addition to exons, potential regulatory elements such as promoters and polyadenylation signals also can be predicted. Promoters obviously are important elements in gene function and they determine the 5' end of a gene. However, the existing promoter predicting methods often are inefficient. As discussed in Chapter 2, even the best program available predicted a large number of false positives when analyzing genomic sequences. However, when the syntenic human and mouse sequences were compared, as detailed in Chapter 4, all the known genes in the genomic regions sequenced in this dissertation research were associated with conserved CpG islands in both human and mouse. These

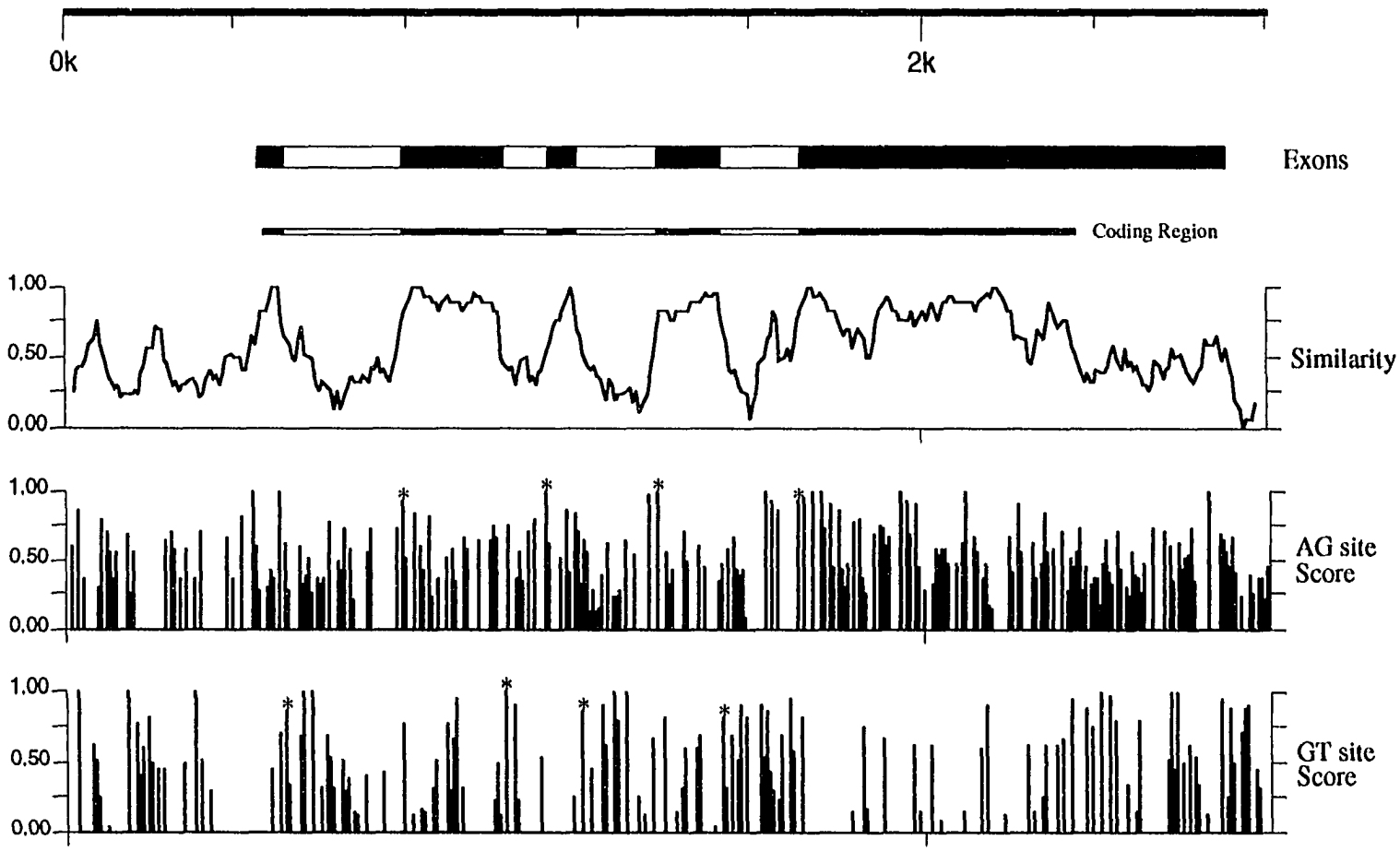


Figure 5.2 Amino acid level similarity and splicing site quality around the human and mouse Dok gene region

Note: Amino acid level similarity is calculated as the maximum for all three frames with a window size of 15 aa.
Splicing site quality is calculated as described in the text, * indicate known splicing sites

observations are consistent with the prediction that 75% of human genes are associated with CpG islands (Bird, 1986; Garden and Frommer, 1987; Fields et al., 1994) and conversely, Lavia et al. (1987) showed that CpG islands typically are associated with genes. Given these observations, conserved CpG islands can be good predictors of either promoters or the 5' ends of genes. The prediction of CpG islands was implemented based on the definition by Garden and Frommer (1987). Specifically, they defined a CpG island as a segment of genomic DNA at least 200 bp long which has a GC content of greater than 50% and a ratio of the number of observed CpG to the number of expected CpG greater than 0.6. Another feature, the polyadenylation signals, which also has a relatively well defined sequence, most frequently AATAAA, occurs in many sites randomly throughout the genome. In this current study, only those potential polyadenylation sites (AATAAA or the less frequently used ATTAAA) which were located close to or within conserved regions were examined.

5.3 Building gene models

After the gene elements described above are located, a prediction based on gene models then is attempted. Here a dynamic programming technique was used to solve two of the major problems with building gene models. First, the predicted exons, including falsely predicted ones must be optimally organized into an open reading frame assuming that only one gene occurs in the region and second, when more than one gene are present in a region, the individual start and stop positions must be determined.

Dynamic programming, a method that builds a global optimal answer on local optimum to a potentially complex problem and is often used to solve optimization problems (Cormen et al., 1990), has been used in gene prediction programs (e.g. Snyder and Stormo, 1993). To illustrate how this approach was employed in this dissertation research to solve the gene structure prediction problem, a simplified example with three non-overlapping potential exons (A, B and C), each of which has a score and associated ORF

compatibility, is as follows. First, it is determined whether exons A and B can form an ORF and if so, a new score is calculated, otherwise the higher scoring exon, either A or B, is selected to represent the optimal path from exons A to B, or (A, B). Similarly, the combination of exons B and C is examined and subsequently, the combinations of ((A,B), C) and (A, (B, C)) are examined. The higher score of the two combinations is the final score for the three exons and the path from which the higher score is obtained, represents the optimal exon organization of the three exons. Generally, the score and optimal path from predicted exons i to j is calculated by the recursive formula:

$$m(i, j) = \max_{\substack{0 \le s \le 4 \\ 0 \le t \le 4}} \{ \max_{\substack{i \le k < j \\ 0 \le u \le 4}} \{ m_{su}(i, k) + m_{ut}(k+1, j) \} \}$$

where $m_{su}(i, k)$ is the score for optimal path from exons i to k, with the start frame of s and next compatible frame of u. A start frame of 0 represents an initial exon and a end frame of 4 represents a terminal exon. Thus, when applying dynamic programming to solve the exon structure problem, the objective is to search for an optimal path which has a maximum total score among all possible paths through the exons. However, when the predicted exons are joined to form an open reading frame in the manner as described above, the optimal path will result in only a single ORF representing a single gene. To allow for the possibility that more than one gene could be detected, the algorithm also allows for a potential 3' end exon to form a valid path with a neighboring 5' end exon or even an internal exon to accommodate cases where a 5' end exon is missed in the exon prediction process. Similarly, the algorithm also allows for searching alternative paths around the 5' end exons. Promoters and polyadenylation signals are assigned frames of 0 and 4, respectively.

5.4 Testing of comparative gene prediction

A training set of human and mouse sequences, including the two pairs of human and mouse sequences collected in this dissertation research, and another sequence pair that was retrieved from Genbank (Table 5.1) was used to adjust the various parameters to obtain the optimized prediction results. Then, the program (part of Geneligner, see Section 5.6) was run on a test sequence set (Table 5.1). The predicted genes then were compared to gene annotation either directly from the Genbank or from aligning cDNA sequences with genomic sequences. The results are shown in Table 5.2.

Table 5.1 Human and mouse sequences used in comparative gene prediction

	Source	Clone name (Accession)	Length(bp)	Genes*	Reference
Training set:					
1	Human	p201m18(AC000097)	162,268	3	This dissertation
	mouse	72k21(AC003060)	118,243		This dissertation
2	human	H173(AC003065)	90,832	2	This dissertation
	mouse	b245c12(AC003061)	162,750		This dissertation
3	human	(U47924)	222,930	16	Ansari-Lari et al.(1998)
	mouse	(AC002397)	227,538		Ansari-Lari et al.(1998)
Test set					
1	human	(U78027)	99,014	5	Oeltjen et al.(1997)
	mouse	(U58105)	88,871		Oeltjen et al.(1997)
2	human	(L47234)	54,336	3	Lamerdin et al. (1996)
	mouse	(L47235)	32,595		Lamerdin et al. (1996)
3	human	(L34079)	37,785	1	Lamerdin et al. (1995)
	mouse	(L34078)	37,349		Lamerdin et al. (1995)
4	human	(L77570)	108,400	2	ACGT
	mouse	(AC003063)	128,148		ACGT
5	human	83c5(AC00087)+102g9(AC000068)+ 49c12(AC000079)+ 98c4(AC000029)			
			~100,000	1	ACGT
	mouse	b264n1(AC003062)	89,743		ACGT

* Only known genes are counted for the testing purpose.

As listed in Table 5.2, the performance of the current version of Geneligner is close to that of Genscan and significantly better than that of Xgrail in terms of sensitivity and specificity. A close examination of the predictions reveals that errors may occur in several situations: (1) low conservation of an exon between the species compared; (2) non-typical splicing sites; (3) potential sequence errors; and (4) coding exons shorter than 30 bp. It is

Table 5.2 Performance of gene prediction programs on testing set

	Exact exon	Overlap exon	Missed exon	False exon	Exact gene ends
Genscan	102	10	11	8	12
Xgrail	68	36	18	27	N/A
Geneligner	95	17	11	0	11

	Sensitivity(%)	Specificity(%)
Genscan	83	85
Xgrail	55	52
Geneligner	77	85

Sensitivity: % of the actual coding exons been predicted exactly right
Specificity: % of the predicted exons perfectly matching an actual exon

expected that later versions of Geneligner will deal with these situations and as a result, improve its overall accuracy. For example, if two internal coding exons are predicted with high probability but with an incompatible open reading frame, a less conserved exon between them may be predicted. Currently, all exons are predicted independently without considering the frames of adjacent exons. An example of a comparative prediction result vs. the prediction of Genscan and Xgrail is shown in Figure 5.3. Here a comparison of the exons in the two gene model predicted by Geneligner (panel 3) with those of the two known multi-exon genes (Idd gene: X95480 and ES2 gene: AF037256, both contain 10 coding exons) shows that Geneligner correctly predicted 18 of the 20 exons, with no false exons, one exon predicted with an incorrect exon range and one exon not predicted. Although at the position of the missed exon (position labeled with * in panel 3 of Figure 5.3) an exon was originally predicted in the exon prediction phase (see panel 1) but the final gene model failed to include this exon since the exon range was falsely predicted due to a deletion in its corresponding human exon. In contrast, the first coding exon of the ES2 gene was predicted to be shorter than the actual exon since part of the region was not recognized as conserved by Crossmatch. Compared to Geneligner, Genscan correctly predicted 16 exons, with three more overlapping the actual exons, the first coding exon of the Idd gene being missed entirely and three false positive exons being predicted. Xgrail predicted 16 exons which overlap the known exons while it predicted more than 10 false

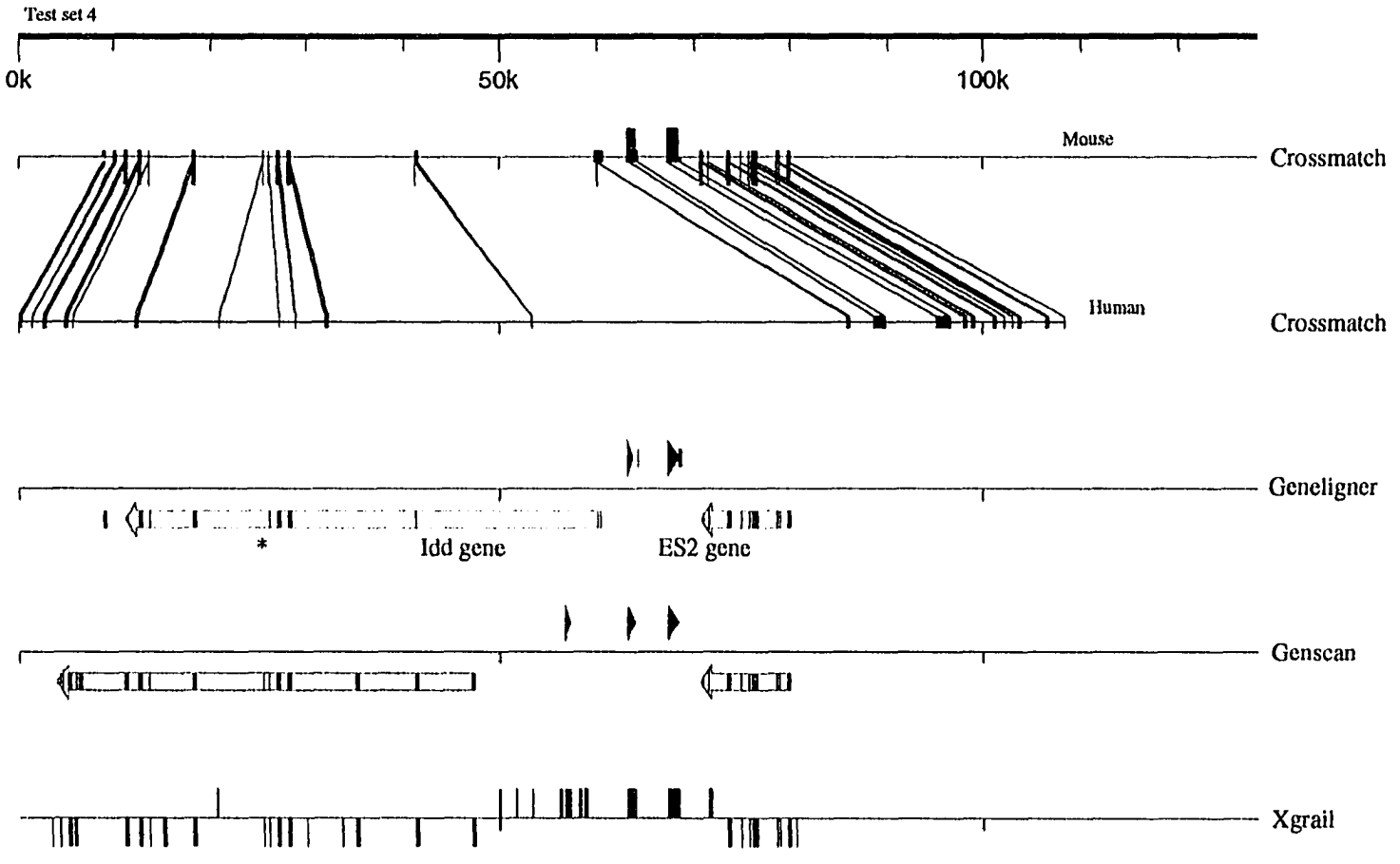


Figure 5.3 An example of gene prediction by different programs

Crossmatch result with predicted gene elements are shown in the top two panals (Boxes: Blue=good exon; green=fair exon; black=CpG island; yellow=polyA signal)
Third and fourth panels:blue=exon; green=promoter location; red=polyA signal

positives. Comparison of the prediction results on the test set by all three programs showed that Geneligner made the fewest false positive predictions because the regions falsely predicted by the other programs were not conserved to a level that could be detected by Crossmatch. Interestingly, Genscan and Xgrail often made similar false predictions most likely because of the exon-like nucleotide composition of those regions.

It also can be seen from Table 5.2 that as with all known gene prediction programs (Claverie, 1997), gene boundaries are much less efficiently predicted. The major difficulty in predicting gene boundaries is that terminal coding exons often are not as conserved as internal exons (also see Chapter 4). However, since regions spanning terminal exons generally are revealed by Crossmatch, it is expected that future improvement in the algorithm should improve the performance of Geneligner in predicting terminal exons.

5.5 Building gene models by integrating data from database similarity search and computational gene prediction

As discussed above, another important aspect of gene prediction is to include the results from a database similarity search. The growing Genbank EST database provides important information that can be used to predict gene presence and location. Since EST sequences generally represent the 5' or 3' end of a gene and gene prediction programs do not perform well in predicting these gene boundaries, this information has the potential of being extremely useful. Thus, integrating the results from gene prediction programs and results from database similarity searches most likely will improve the overall accuracy of gene prediction. During this dissertation research, a program which builds gene models based on information from database searches and from gene prediction programs has been written into the Geneligner programs in an attempt to improve the automatic annotation of genomic sequences.

The strategy is similar to that used for comparative gene prediction as detailed in the above sections. Specifically, individual elements, including exons, promoters and polyadenylation signals, first are identified from individual data sources and then gene

models are built. When available, the results from Genscan, Xgrail, comparative gene prediction and other gene prediction programs are directly inputted into the gene model building routines while special routines process database similarity search results.

Database searches are performed with the program Powblast. Since the Powblast output generally does not contain the full length sequence and annotation of the homologous items in Genbank, the first step is to retrieve these data from Genebank. Then, the homologous sequences are evaluated to determine whether all segments of the full length sequence (such as an EST sequence) are present in the Powblast results. This step is necessary since it was observed that Powblast often fails to identify short exons, or low homology exons. If such segments exist, alignment with the genomic sequence is performed using Crossmatch. In addition to the above, Powblast often does not correctly identify boundaries of homologous regions as proper exon boundaries, since the homologies frequently extend into the introns and therefore are longer than the actual exons. In such cases, further local alignment will give an optimal alignment of the homologous sequence with the genomic sequence. In addition, the direction of an EST, i.e. whether the EST corresponds to the 5' end of a potential gene or from the 3' end is determined. Although this information can be obtained from Genbank annotation of the EST, often the annotation is incorrect. However, when a Genbank entry matches several segments (potential exons) of a genomic sequence, the boundaries of those matches in the genomic sequence are examined for potential splicing site consensus sequence (AG|GT) in both strands. If the obtained results contradict the Genbank annotation, the direction of the EST is reversed. Finally, if the Genbank annotation of an EST indicates the presence of repetitive elements, the EST is removed from further consideration since ESTs containing repetitive elements such as Alu repeats may arise by contamination from various sources (Adams et al., 1995). After these initial processing steps, the remaining homologous entries are analyzed for the exon characteristics of potential genes as described above. The obtained result then serves as input for gene model building routines as described in

Section 5.3. The overall strategy for analyzing database similarity search results is shown in Figure 5.4.

During the course of this work, it became clear that further restrictions were needed when building gene models from input data containing information from database similarity searches. First, if an EST or cDNA revealed two or more exons with appropriate splicing sites, any other exons which were predicted within the region must be removed from further consideration because the experimental EST or cDNA evidence should take precedence over a computationally-based prediction. Second, when two potential exons are revealed by the same EST, they must be organized into the same gene model, whether or not the two exons have compatible open reading frames. Therefore, the overall strategy of building gene models from data that includes extensive database similarity search results is shown in Figure 5.5. In the example shown in Figure 5.6, which was the sequence from contig 2 described in Chapter 3, Powblast was performed against only the Genbank EST database to exclude the known Porc-P1 cDNA sequence in this contig for illustrational purposes. Figure 5.6 shows that the 5' and 3' ends of the gene are delimited by a series of overlapping EST sequences. From the EST data alone, Geneligner predicts a putative gene with 15 exons where the exon content in the middle of the gene is uncertain as indicated as red boxes in panel 2 of Figure 5.6. Genscan also predicted a gene in the region, but both the 5' end and 3' end of the Genscan prediction were incorrect. Specifically, the 5' end coding exon was predicted as an internal exon and the correct 3' end coding exon was not predicted by Genscan. Nevertheless, Genscan predicted most of the internal exons correctly except for an internal coding exon (exon 3). When combining the results of Powblast and Genscan, Geneligner predicted all the exons and gene boundaries correctly. It should be noted that this feature of Geneligner has not been systematically evaluated in terms of its accuracy although it has analyzed the entire DiGeorge Syndrome Critical Region and gave satisfactory results (Chen et al., 1997).

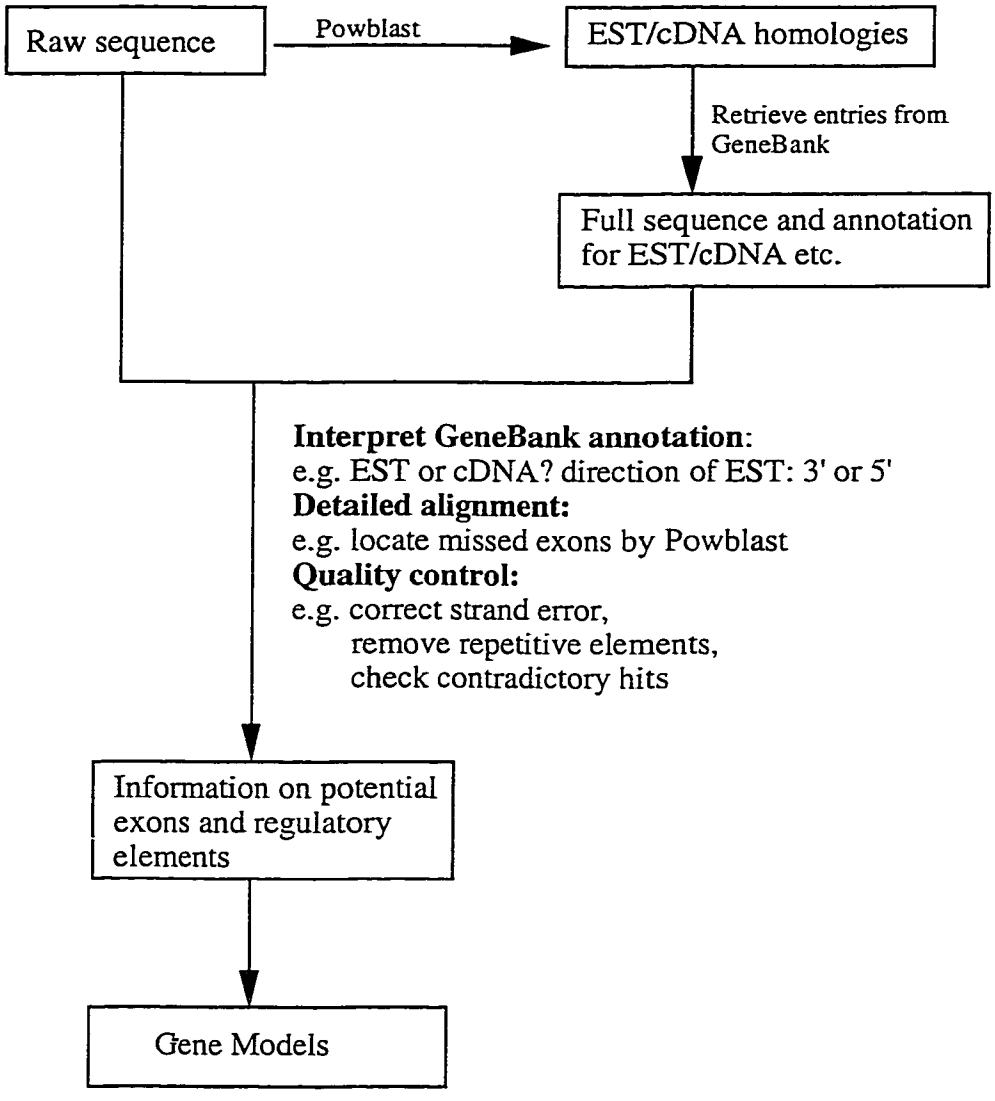


Figure 5.4 Diagram showing the procedure for building gene models from GeneBank similarity search results by Geneligner

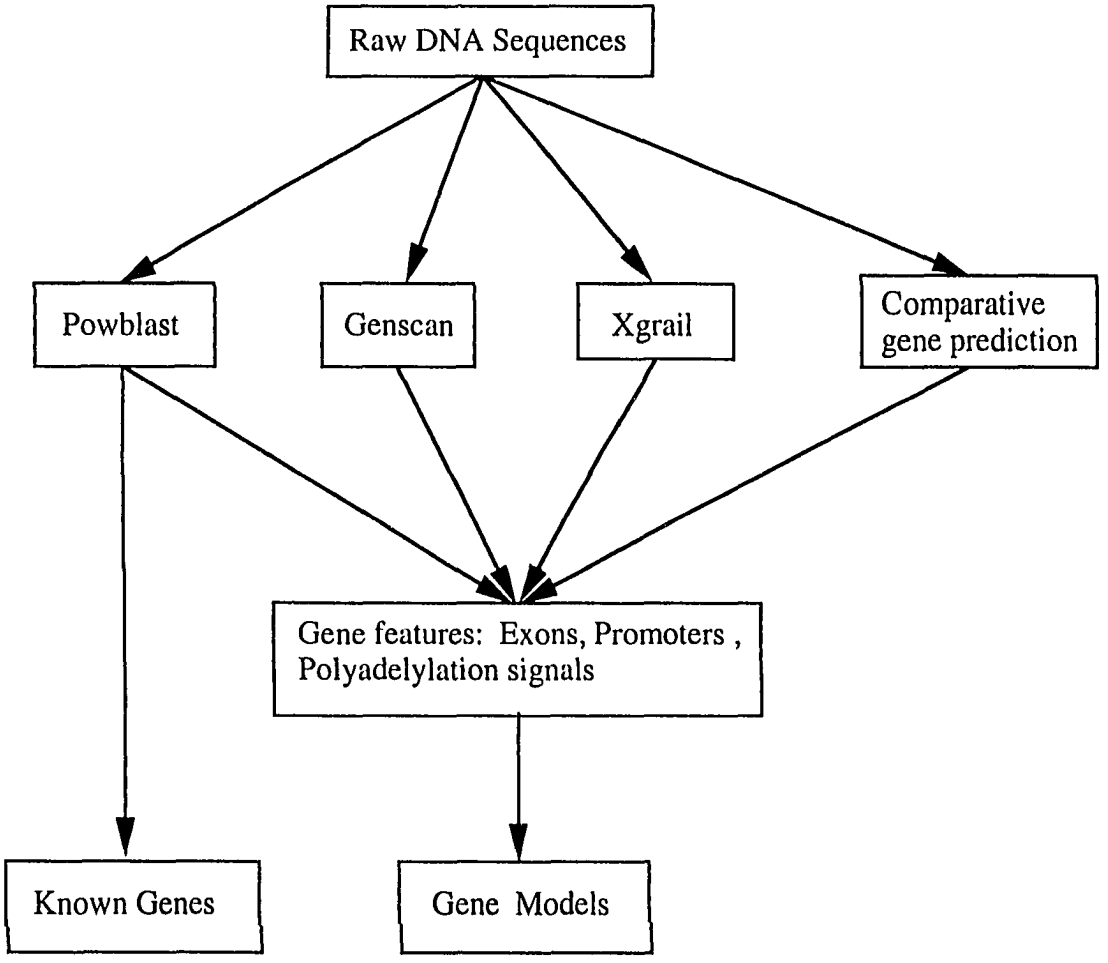


Figure 5.5 Diagram showing dataflow for the gene model building routines used in Geneligner

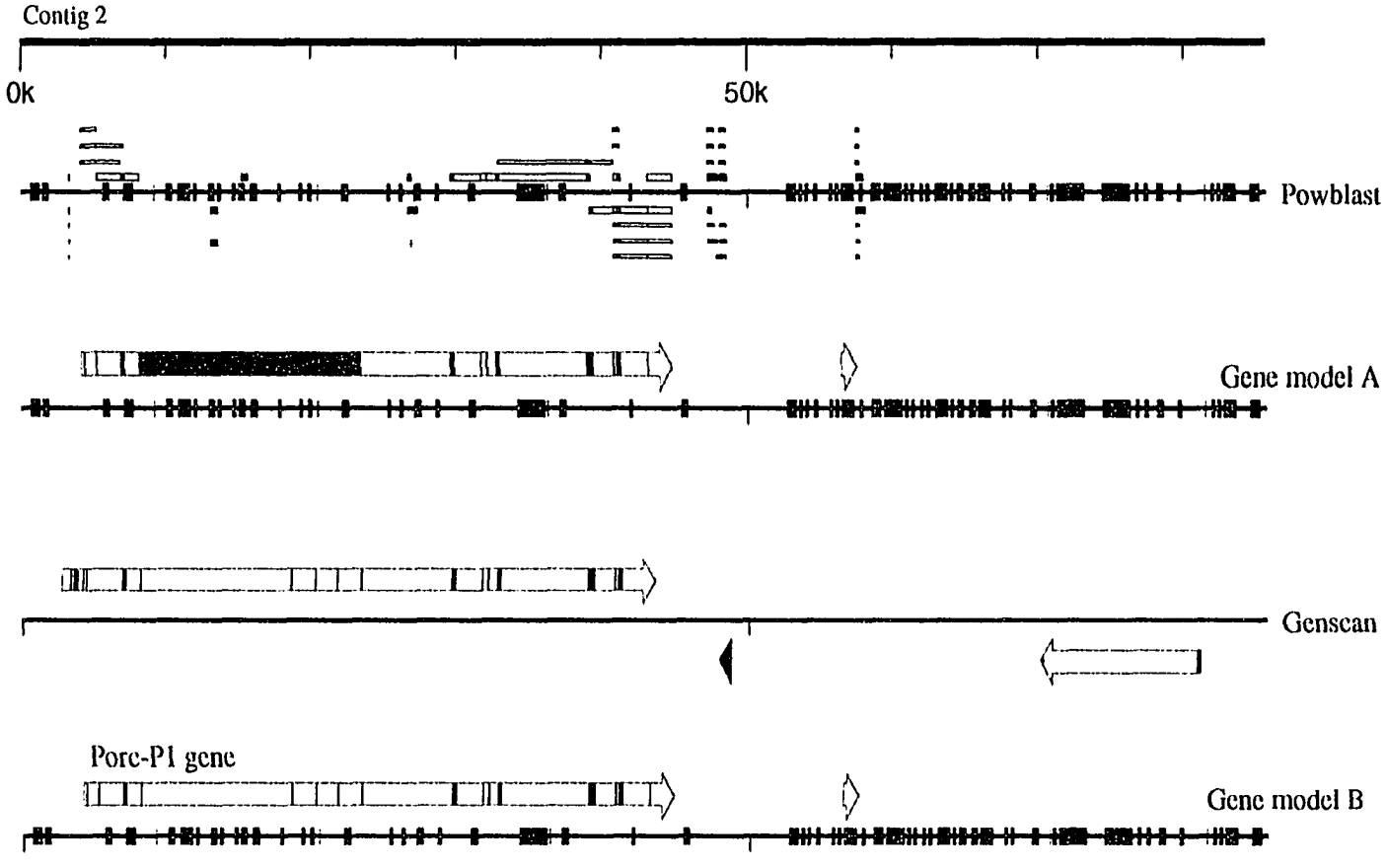


Figure 5.6 An example of building gene models from database similarity search and gene prediction results

Panel1: For illustrational purpose, Powblast performed against only Genbank EST database
Second and Fourth panels: gene models built from Powblast data with and without Genscan prediction
Red zone in gene models indicates uncertainty

5.6 The program package: Geneligner

The programs for comparative gene prediction and gene model building as described in the above sections were written as part of a program package named Geneligner. One of the major goals of this program package was to provide a platform for either automatic or manual annotation of human and mouse genomic sequences. In the current version of Geneligner, computational gene predictions and database homology search results can be graphically examined in a main window area and to facilitate examination of various features at nucleotide sequence level, sequence alignments are shown in a separate window. The main window area, which can be zoomed in or out, is used to obtain a detailed or overall view of the sequence features from one or more projects (such as syntenic regions of human and mouse). In addition, the sequence alignments can be adjusted to facilitate feature comparison. Annotations can be added to or removed from existing ones, and the graphs can be outputted as a postscript file for printing.

Geneligner also includes the ability to search for specific annotations and sequences and the results can give either graphic or tabular output. For example, a sequence can be queried for potential polyadenylation signals, splicing sites (with quality), CpG islands and open reading frames. In addition, any portion of a sequence can be chosen either directly or as its complemented sequence and outputted into a separate file in FASTA format with or without translation in all six frames.

This UNIX operating system specific program package was written in the C programming language and the graphic interface was written in MOTIF.

Chapter VI
Conclusion

Since its advent seventeen years ago (Anderson, 1981), the random shotgun sequencing strategy has been evolving in almost every step of its implementation. For example, the original fragmentation of DNA with DNase I digestion was replaced by sonication and now by nebulization, which is more economic, convenient, and consistently gives a more random distribution of DNA fragments of desired size than any other methods (Hengen, 1997). The radioactively labeled sequencing terminator was replaced by dye-labeled primer and now by more and more sensitive dye-labeled terminator. Many technologic improvements, together with advances in bioinformatics, have made it possible to efficiently sequence large genomic inserts of more than 150 kb in length, as evidenced by the results presented in this dissertation research. As detailed in the previous chapters, special efforts also were made to improve the efficiency of both the large scale and mini-prep DNA isolation as well as sequence gap closure during this dissertation research.

Through successful implementation of the random shotgun sequencing strategy, seven cosmid, BAC and PAC clones, five from the DiGeorge Syndrome Critical Region (DGCR) from human chromosome 22q11 and its syntenic region from mouse chromosome 16, and two from the Motor Neuron Disease Critical Region (mnd2) from mouse chromosome 6 and its syntenic region from human chromosome 2, have been sequenced to completion. Analysis of the resulting approximately 800 kb sequence revealed 13 known or putative genes (Table 6.1) along with several other genome features such as pseudogenes, repeats and regulatory elements. These features, both structural genes and other elements, have been examined in detail.

In humans, five known genes, including CLTCL, HIRA, Porc-P1, TMVCF and RanBP1, were found located in the DGCR region sequenced during this dissertation research. Although the function of CLTCL gene is unknown, it shares significant amino

Table 6.1 Summary of Genes predicted in the sequenced contigs*

Contig no.	Genes	Genomic size (kb)	Number of Exons	Σgene size / contig size	Gene density (kb)
1 (human)	CLTCL	(73.7)	(2)		
	HIRA	(28.2)	(7)	61%	(50)
2 (human)	Porc-P1	41.7	19		
	TMVCF	1.4	1	53%	42.7
3 (human)	T10	(14.7)	(6)		
	Htf9c	5.2	12		
	RanBP1	9.8	6	45%	32.5
	PutA	31.6	15		
	PutB	11.8	11		
4 (mouse)	T10	43.1	9		
	Htf9c	4.8	12	78%	29.6
	RanBP1	8.9	6		
	PutA	35.3	15		
5 (human)	Dok	2.8	5		
	Gene7	82.9	9	94%	45.4
6 (mouse)	Dok	2.5	5		
	Gene7	80.3	9		
	SemaM	26.3	14	70%	40.7
	Lor2	(5.6)	(4)		
Average	human	23.4	9.8	60%	37.2
	mouse	28.7	10	73%	35.1

* Numbers in parenthesis indicate incomplete genes and average values are calculated on complete genes.

acid level sequence similarity with the clathrin heavy chain gene. The HIRA gene may be involved in transcriptional regulation of histone synthesis and/or modification of chromatin structure. Both the CLTCL gene and HIRA genes also have been proposed to be the candidate genes responsible for DGS. The Porc-P1 gene codes a putative nuclear protein that possibly is involved in DNA synthesis. The TMVCF gene codes a putative transmembrane protein that might be involved in target cell recognition and pore formation. The RanBP1 gene, a member of the Ran network, most likely is involved in a number of cellular processes including DNA replication, cell cycle progression, protein and RNA transportation. In addition to the above known genes, two other genes, i.e. the T10 and Htf9c genes, were predicted in the DGCR region based on sequence similarity with two known mouse genes as well as the human EST data. While the function of the T10 gene currently is unknown, the Htf9c gene most likely is involved in nucleic acid metabolism and/or modification. Two other genes, PutA and PutB, were predicted in the DGCR region based on the combination of database similarity search and computational gene prediction. The putative PutA gene probably is a nuclear protein involved in nucleic acid modification and the PutB gene contains transmembrane domains and a Zn-finger motif. Thus, a total of nine genes were discovered in the approximately 500 kb human DGCR region sequence. In the human syntenic region of mouse mnd2 region, one known gene, the Dok gene, was located and one putative gene, Gene7, were predicted. The Dok gene most likely is a component of a signal transduction pathway while no function has yet been assigned to the putative Gene7 gene.

In mouse, two known genes, the T10 and Htf9c genes, were localized and one predicted gene, PutA, was discovered in the mouse region syntenic to the human DGCR region. In the mouse mnd2 region, one known gene, Dok, was located and three genes, Gene7, semaphorin M and Lor2, were predicted and partially confirmed. The semaphorin M gene is a member of the semaphorin family which is believed to be involved in neuronal

axon guidance during neural network formation. The Dok, Gene7 and Lor2 genes found in the mouse presumably have functions similar to their human homologues.

A detailed comparison of the two syntenic human and mouse genomic regions determined in this dissertation research reveals that both gene order and genomic structure are well conserved between human and mouse. At the DNA sequence level, the coding regions are well conserved, with a sequence similarity varying between 79.4% and 90.6%. Relatively less conservation is seen at the 5' and 3' ends of coding regions. The 5' and 3' untranslated regions (UTR) are even less conserved in human and mouse but often well conserved sequence stretches that presumably represent conserved regulatory elements important for transcription and/or translation are scattered within these UTRs. Comparisons of the promoter regions revealed varying degrees of conservation. Some promoter regions, such as those of the human and mouse RanBP1 genes, are well conserved, especially at known or predicted transcription factor binding sites. Transcription factor binding sites for other genes, such as the human and mouse T10 genes, are less conserved since more unconserved transcription factor binding sites were predicted than conserved ones. If those predictions are correct, it might be that there is a different method of gene regulation for the human and mouse T10 genes. Comparison of intronic and intergenic regions generally revealed high conservation in their length with low but generally discernible conservation at sequence level. An examination of the distribution of repetitive sequences showed that both the human and mouse genomes are rich in repetitive elements, with the amounts varying between 32.2 - 42.9% in human and 25.4 -46.8% in mouse for the two syntenic regions. While different genomic regions show different repeat percentages, the variation seems to be locally conserved between human and mouse. Specifically, the first syntenic region contains 32.2% and 25.4% repeats in human and mouse, respectively, while the second syntenic region contains 42.9% and 46.8% repeats in human and mouse, respectively. The variation in the repeat percentage may be related to gene density of the regions since the

first syntenic region contains four known or predicted genes while the second syntenic region only contains two genes even though both regions span about 100 kb.

Coding and protein sequence substitution rates were examined for the two syntenic regions sequenced by this dissertation research. Specifically, for the synonymous substitutions which are thought to be an indication of coding sequence evolution, a rate of 1.99-3.17 per site per 10^9 year was observed for the human DGCR region and its syntenic region in mouse while a rate of 2.54-2.64 per site per 10^9 year was observed for the mouse mnd2 and its syntenic region in human. The nonsynonymous substitution rate, which is thought to be an indication of protein evolution, is 0.13-0.59 per site per 10^9 year and 0.61-1.12 per site per 10^9 year for the two syntenic regions, respectively. Since these two syntenic regions are located in different chromosomes for both species, the numbers indicate a similar evolution rate for both coding DNA sequence and the resulting protein sequence from these two different regions of the human and mouse genomes. Further studies involving more syntenic regions will be required to determine whether different portions of the human and mouse genomes evolved at a similar rate.

Finally, a new gene prediction approach which includes analysis of the coding sequence conservation between human and mouse was developed. Results from the detailed sequence comparison between human and mouse guided the selection of parameters used in this new prediction method. A preliminary test of the program on a set of sequences showed that the performance of this new approach was slightly superior to the gene prediction programs currently available. This gene prediction program was designed as part of a program package named Geneligner developed by this dissertation research. The major purpose of the Geneligner program package was to provide a platform for annotation of large genomic sequences. Many of the figures and tables in this dissertation research were generated by this program package whose features include a graphic alignment of the results from database similarity searches, computational gene prediction programs, and gene model building.

Glossary

Amyotrophic lateral sclerosis. A neurologic disease that rapidly progresses from mild motor symptoms to severe motor paralysis and premature death.

Atrophy. Diminution in the size of a tissue or organ after full normal development has been attained.

Centromere. The region of an eukaryotic chromosome that is attached to the spindle during nuclear division.

Complementary DNA (cDNA). A single-stranded DNA molecule that has a complementary base sequence to a molecule of a messenger RNA.

Competent cells. Cells which have an increased ability to take up foreign DNA.

Conservation. Macromolecules showing a high degree of invariance of their primary structure.

Cytogenetics. The branch of genetics concerned with correlating inherited traits with cytological features, especially the appearance, structure, and behavior of chromosomes.

Divergence. Biological evolution that tends to produce an increasing difference in some characteristics between initially similar groups of organisms or gene products.

Dosage analysis. A method to determine the copy number of a locus. Gene dosage often is quantified by Southern blotting experiments using a DNA probe with genomic DNA as the substrate.

Expressed sequence tag (EST). A particle sequence of a clone picked at random from a cDNA library.

Etiology. The cause or causes of disease and their study

Exon trapping. A method to functionally clone expressed sequences from genomic DNA. In this method, a genomic DNA fragment is inserted into a plasmid

vector. After the plasmid is introduced into a suitable host cell and gene expression is stimulated, exons in the target genomic DNA segment are identified by their ability to remain as a mature spliced mRNA.

Fluorescence in situ hybridization (FISH). A procedure that involves the use of DNA probes to locate specific regions of DNA in the chromosomes.

Homologous. DNA or proteins that have the same or similar residues at corresponding positions.

Hypocalcaemia. The presence of a lower than normal concentration of calcium in the blood.

Hypoplasia. Underdevelopment or defective formation of an organ or tissue.

Isochore. A genomic region which has a relatively homogenous base composition.

Linkage group. A group of genes whose members show a marked tendency to segregate as a unit.

Microheterogeneity. A slight difference between sequences of different individual.

Monosomy. A condition in which one member of a chromosome pair is missing.

Mytonic dystrophy. An autosomal dominant disorder characterized by myotonia (muscle malformations), muscular dystrophy (abnormal structure and function of a muscle), and other changes.

Pleckstrin homology. A protein domain which was first noted in the platelet protein pleckstrin and has unknown function.

Proband. The first individual to present as a patient, in a series of cases related to a heritable disorder, and from whom the line of descent is traced.

Profile analysis. A method for detecting related proteins by comparing a query sequence with a profile. A profile is a position-specific scoring table created from a group of sequences previously aligned by structural or sequence similarity.

Pseudogene. A sequence in DNA that is related to a functional gene but can not be transcribed.

Rapid amplification of cDNA ends (RACE). A variation of the polymerase chain reaction designed for amplification of cDNA corresponding to low-abundance mRNA.

Restriction fragment length polymorphism (RFLP). Variation in the length of DNA fragments generated by a restriction endonuclease.

Shotgun cloning. A method of cloning in which the DNA is randomly broken into fragments of a desired size and the fragments are then cloned into suitable vectors.

Splicing. The enzymatic process in eukaryotic cells by which introns are excised from heterogeneous nuclear RNA (hnRNA) following its transcription from DNA, and the cut ends are rejoined to form messenger RNA.

Synteny. In genetics, synteny refers to the condition that gene loci or genes are located on the same chromosome. In recent years, synteny has been used to indicate that similar genomic sequences are located in homologous chromosomal regions of different species.

Telomere. The structure at the end of a chromosome.

Transcription factor. Any protein required to initiate or regulate transcription; They often bind to DNA sequence elements in the promoter region and stimulate transcription.

Transformation. 1. Transfer of genetic information by means of extracellular DNA in bacteria. 2. Characteristic and inherited changes produced in cells in culture when they have been treated with certain viruses or carcinogens.

Transgenic. The process by which a new gene is added to the germ line of an organism using recombinant DNA technology.

Translocation. A chromosomal mutation associated with the transfer of a chromosomal segment from one chromosome to another. When there is no net loss of the genetic material, the translocation is said to be balanced, otherwise it is unbalanced.

Zinc finger. A DNA-binding domain of a protein that has a characteristic pattern of cysteine and histidine residues that complex with zinc ions.

Literature Cited

Adams J.W., Kaufman R.E., Kretschmer P.J., Harrison M. and Nienhuis A.W. (1980) A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. *Nucleic Acids Res.* 8(24): 6113-28.

Adams M.D., Kerlavage A.R., Fleischmann R.D., Fuldner R.A., Bult C.J., Lee N.H., Kirkness E.F., Weinstock K.G., Gocayne J.D. and White O. et al. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377(6547 Suppl): 3-174.

Anderson S. (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* 9(13): 3015-27.

ano-Yanagisawa H., Li Y., Wang H. and Kohwi Y. (1995) Single-stranded DNA binding proteins isolated from mouse brain recognize specific trinucleotide repeat sequences in vitro. *Nucleic Acids Res.* 23(14): 2654-60.

Ansari-Lari M.A., Oeltjen J.C., Schwartz S., Zhang Z., Muzny D.M., Lu J., Gorrell J.H., Chinault A.C., Belmont J.W., Miller W. and Gibbs R.A. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* 8(1): 29-40.

Aota S. and Ikemura T. (1986) Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* 14(16): 6345-55.

Armour J.A., Crosier M., Malcolm S., Chan J.C. and Jeffreys A.J. (1995) Human minisatellite loci composed of interspersed GGA-GGT triplet repeats. *Proc. R. Soc. Lond. B. Biol. Sci.* 261(1362): 345-9.

Armour J.A., Neumann R., Gobert S. and Jeffreys A.J. (1994) Isolation of human simple repeat loci by hybridization selection. *Hum. Mol. Genet.* 3(4): 599-65.

Barberis A., Superti-Furga G. and Busslinger M. (1987) Mutually exclusive interaction of the CCAAT-binding factor and of a displacement protein with overlapping sequences of a histone gene promoter. *Cell* 50(3): 347-59

Baumgartner B., Heiland S., Kunze N., Richter A. and Knippers R. (1994) Conserved regulatory elements in the type I DNA topoisomerase gene promoters of mouse and man. *Biochim. Biophys. Acta* 1218(1): 123-7.

Beckman J.S. and Weber J.L. (1992) Survey of human and rat microsatellites. *Genomics* 12(4): 627-31.

Bey L., Etienne J., Tse C., Brault D., Noe L., Raisonnier A., Arnault F., Hamilton M.T. and Galibert F. (1998) Cloning, sequencing and structural analysis of 976 base pairs of the promoter sequence for the rat lipoprotein lipase gene. Comparison with the mouse and human sequences. *Gene* 209(1-2): 31-8.

Bird A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature* 321(6067): 209-13.

Birnboim H.C. and Doly J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* 7(6): 1513-23.

Bischoff F.R. and Ponstingl H. (1991) Catalysis of guanine nucleotide exchange on Ran by the mitotic regulator RCC1. *Nature* 354(6348): 80-2.

Bischoff F.R., Klebe C., Kretschmer J., Wittinghofer A. and Ponstingl H. (1994) RanGAP1 induces GTPase activity of nuclear Ras-related Ran. *Proc. Natl. Acad. Sci. USA* 91(7): 2587-91.

Bischoff F.R., Krebber H., Smirnova E., Dong W. and Ponstingl H. (1995) Co-activation of RanGTPase and inhibition of GTP dissociation by Ran-GTP binding protein RanBP1. *EMBO J.* 14(4): 705-15.

Bradley W.G. (1995) Overview of motor neuron disease: classification and nomenclature. *Clin. Neurosci.* 3(6): 323-6.

Brady S. (1995) Motor-neuron diseases. Interfering with the runners. *Nature* 375(6526): 12-3.

Brauninger A., Strebhardt K. and Rubsamen-Waigmann H. (1995) Identification and functional characterization of the human and murine polo-like kinase (Plk) promoter. *Oncogene* 11(9): 1793-800.

Bressan A., Somma M.P., Lewis J., Santolamazza C., Copeland N.G., Gilbert D.J., Jenkins N.A. and Lavia P. (1991) Characterization of the opposite-strand genes from the mouse bidirectionally transcribed HTF9 locus. *Gene* 103(2): 201-9.

Brzustowicz L.M., Lehner T., Castilla L.H., Penchaszadeh G.K., Wilhelmsen K.C., Daniels R., Davies K.E., Leppert M., Ziter F., Wood D., Dubowitz V., Zerres K., Hausmanowa-Petrusewicz I., Ott J., Munsat T.L. and Gilliam T.C. (1990) Genetic mapping of chronic childhood-onset spinal muscular atrophy to chromosome 5q11.2-13.3. *Nature* 344(6266):540-1

Buard J. and Jeffreys A.J. (1997) Big, bad minisatellites. *Nat. Genet.* 15(4): 327-8.

Bucher P. Karplus K. Moeri N. and Hofmann K. (1996) A flexible search technique based on generalized profiles. *Computers and Chemistry* 20: 3-24.

Budarf M.L., Collins J., Gong W., Roe B., Wang Z., Bailey L.C., Sellinger B., Michaud D., Driscoll D.A. and Emanuel B.S. (1995) Cloning a balanced translocation associated with DiGeorge syndrome and identification of a disrupted candidate gene. *Nat. Genet.* 10(3): 269-78.

Burge C. (1997) Identification of genes in human genomic DNA. PhD thesis, Stanford University, Stanford, CA.

Burge C. and Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268(1): 78-94.

Carey A.H., Kelly D., Halford S., Wadey R., Wilson D., Goodship J., Burn J., Paul T., Sharkey A. and Dumanski J. et al. (1992) Molecular genetic study of the frequency of monosomy 22q11 in DiGeorge syndrome. *Am. J. Hum. Genet.* 51(5): 964-70.

Carothers A.M., Urlaub G., Mucha J., Grunberger D. and Chasin L.A. (1989) Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. *Biotechniques* 7(5): 494-9.

Carpino N., Wisniewski D., Strife A., Marshak D., Kobayashi R., Stillman B. and Clarkson B. (1997) p62(dok): a constitutively tyrosine-phosphorylated, GAP-associated protein in chronic myelogenous leukemia progenitor cells. *Cell* 88(2): 197-204.

Carver E.A. and Stubbs L. (1997) Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale. *Genome Res.* 7(12): 1123-37.

Caskey C.T., Pizzuti A., Fu Y.H., Fenwick R.G. Jr and Nelson D.L. (1992) Triplet repeat mutations in human disease. *Science* 256(5058): 784-9.

Chapelle de la A., Herva R., Koivisto M. and Aula P. (1981) A deletion in chromosome 22 can cause DiGeorge syndrome. *Hum Genet.* 57(3): 253-6.

Chen F. (1997) Sequence and analysis of approximately 0.5 MB of the DiGeorge syndrome critical region in human chromosome 22q11 and a syntenic mouse BAC. Ph.D. Thesis, University of Oklahoma.

Chen F., Hua A., Yao Z., Chu L. Roe B. et al. (1998) Sequence and analysis of over 2 Mb from the human DGCR and distal region on chromosome 22 and over 1 Mb from the syntenic region of mouse chromosome 16. *Microbial & Comparative genomics* 3 (3): C-60.

Chen H.I. and Sudol M. (1995) The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules. *Proc. Natl. Acad. Sci. USA* 92(17): 7819-23.

Chen Z., Zandonatti M., Jakubowski D. and Fox H.S. (1998) Brain capillary endothelial cells express MBEC1, a protein that is related to the Clostridium perfringens enterotoxin receptors. *Lab Invest.* 78(3): 353-63.

Cheng H., Bjerknes M. and Chen H. (1996) CRP-ductin: a gene expressed in intestinal crypts and in pancreatic and hepatic ducts. *Anat. Rec.* 244(3): 327-43.

Christensen C.R., Klingelhofer J., Tarabykina S., Hulgaard E.F., Kramerov D. and Lukanidin E. (1998) Transcription of a novel mouse semaphorin gene, M-semaH, correlates with the metastatic ability of mouse tumor cell lines. *Cancer Res.* 58(6): 1238-44.

Clarkson B. and Strife A. (1993) Linkage of proliferative and maturational abnormalities in chronic myelogenous leukemia and relevance to treatment. *Leukemia* 7(11): 1683-721.

Claverie J.M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 6(10): 1735-44.

Claverie J.M. and Bougueleret L. (1986) Heuristic informational analysis of sequences. *Nucleic Acids Res.* 14(1): 179-96.

Collins D.W. and Jukes T.H. (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* 20(3): 386-96.

Collins F.S., Patrinos A., Jordan E., Chakravarti A., Gesteland R. and Walters L. (1998) New goals for the U.S. Human Genome Project: 1998-2003. *Science* 282: 682-9.

Conley M.E., Beckwith J.B., Mancer J.F. and Tenckhoff L. (1979) The spectrum of the DiGeorge syndrome. *J. Pediatrr.* 94(6): 883-90.

Cooper M.D., Peterson R.D.A. and Good R.A. (1965) A new concept of the cellular basis of immunology. *J. Pediat.* 67: 907-8.

Cormen T., Leiserson C. and Rivest R. (1990) Introduction to Algorithms. The MIT Press, Cambridege, MA.

Cox G.W., Taylor L.S., Willis J.D., Melillo G., White R.L. 3rd, Anderson S.K. and Lin J.J. (1996) Molecular cloning and characterization of a novel mouse macrophage gene that encodes a nuclear protein comprising polyglutamine repeats and interspersing histidines. *J. Biol. Chem.* 271(41): 25515-23.

Crick F.H.C. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.* 12: 548-555.

Cusack S., Hartlein M. and Leberman R. (1991) Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. *Nucleic Acids Res.* 19(13): 3489-98.

Daniels G.R. and Deininger P.L. (1985) Repeat sequence families derived from mammalian tRNA genes. *Nature* 317(6040): 819-22.

DeBry R.W. and Seldin M.F. (1996) Human/mouse homology relationships. *Genomics* 33(3): 337-51.

de Klein A., van Kessel A.G., Grosveld G., Bartram C.R., Hagemeijer A., Bootsma D., Spurr N.K., Heisterkamp N., Groffen J. and Stephenson J.R. (1982) A cellular *Oncogene* is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* 300(5894): 765-7.

Di Cristofano A., Carpino N., Dunant N., Friedland G., Kobayashi R., Strife A., Wisniewski D., Clarkson B., Pandolfi P.P. and Resh M.D. (1998) Molecular cloning and characterization of p56dok-2 defines a new family of RasGAP-binding proteins. *J. Biol. Chem.* 273(9): 4827-30.

Di Matteo G., Salerno M., Guarguaglini G., Di Fiore B., Palitti F., Lavia P. (1998)Interactions with single-stranded and double-stranded DNA-binding factors and alternative promoter conformation upon transcriptional activation of the Htf9-a/RanBP1 and Htf9-c genes. *J. Biol. Chem.* 273(1): 495-505.

Djian P. (1998) Evolution of simple repeats in DNA and their relation to human disease. *Cell* 94(2): 155-60.

Dorn A., Bollekens J., Staub A., Benoist C. and Mathis D. (1987) A multiplicity of CCAAT box-binding proteins. *Cell* 50(6): 863-72.

Driscoll D.A., Budarf M.L. and Emanuel B.S. (1992) A genetic etiology for DiGeorge syndrome: consistent deletions and microdeletions of 22q11. *Am. J. Hum. Genet.* 50(5): 924-33.

Drivas G.T., Shih A., Coutavas E., Rush M.G. and D'Eustachio P. (1990) Characterization of four novel ras-like genes expressed in a human teratocarcinoma cell line. *Mol. Cell Biol.* 10(4): 1793-8.

Dudov K.P. and Perry R.P. (1984) The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene. *Cell* 37(2): 457-68.

Dupressoir A. and Heidmann T. (1996) Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. *Mol. Cell Biol.* 16(8): 4495-503.

Dupressoir A., Puech A. and Heidmann T. (1995) IAP retrotransposons in the mouse liver as reporters of ageing. *Biochim. Biophys. Acta* 1264(3): 397-402.

Eckert R.L. and Green H. (1986) Structure and evolution of the human involucrin gene. *Cell* 46(4): 583-9.

Ellis C., Moran M., McCormick F. and Pawson T. (1990) Phosphorylation of GAP and GAP-associated proteins by transforming and mitogenic tyrosine kinases. *Nature*343(6256): 377-81.

Fickett J.W. and Hatzigeorgiou A.G. (1997) Eukaryotic promoter recognition. *Genome Res.* 7(9): 861-78.

Fickett J.W. and Tung C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.* 20(24): 6441-50.

Fields C., Adams M.D., White O. and Venter J.C. (1994) How many genes in the human genome? *Nat Genet.* 7(3): 345-6.

Frech K., Dietze P., Werner T. (1997b) ConsInspector 3.0: new library and enhanced functionality. *Comput. Appl. Biosci.* 13(1): 109-10.

Frech K., Quandt K. and Werner T. (1997a) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.* 22(3): 103-4.

Furuyama T., Inagaki S., Kosugi A., Noda S., Saitoh S., Ogata M., Iwahashi Y., Miyazaki N., Hamaoka T. and Tohyama M. (1996) Identification of a novel transmembrane semaphorin expressed on lymphocytes. *J. Biol. Chem.* 271(52): 33376-81.

Gardiner-Garden M. and Frommer M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.* 196(2): 261-82.

Gautheret D., Poirot O., Lopez F., Audic S. and Claverie J.M. (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* 8(5): 524-30.

Gottlieb S., Emanuel B.S., Driscoll D.A., Sellinger B., Wang Z., Roe B. and Budarf M.L. (1997) The DiGeorge syndrome minimal critical region contains a ggosecoid-

like (GSCL) homeobox gene that is expressed early in human development. Am. J. Hum. Genet. 60(5): 1194-201.

Gleeson T.J. and Staden R. (1991) An X windows and UNIX implementation of our sequence analysis package. *Comput. Appl. Biosci.* 7(3): 398.

Glover T.W. (1995) CATCHing a break on 22. *Nat Genet.* 10(3): 257-8

Gong W., Emanuel B.S., Collins J., Kim D.H., Wang Z., Chen F., Zhang G., Roe B. and Budarf M.L. (1996) A transcription map of the DiGeorge and velo-cardio-facial syndrome minimal critical region on 22q11. *Hum. Mol .Genet.* 5(6): 789-800.

Greenberg F. (1993) DiGeorge syndrome: an historical review of clinical and cytogenetic features. *J. Med. Genet.* 30(10): 803-6.

Greenberg F., Courtney K.B., Wessels R.A., Huhta J., Carpenter R.J., Rich D.C. and Ledbetter D.H. (1988b) Prenatal diagnosis of deletion 17p13 associated with DiGeorge anomaly. *Am. J .Med. Genet.* 31(1): 1-4.

Greenberg F., Crowder W.E., Paschall V., Colon-Linares J., Lubianski B. and Ledbetter D.H. (1984) Familial DiGeorge syndrome and associated partial monosomy of chromosome 22. *Hum. Genet.* 65(4): 317-9.

Greenberg F., Elder F.F., Haffner P., Northrup H., Ledbetter D.H. (1988a) Cytogenetic findings in a prospective series of patients with DiGeorge anomaly. *Am. J .Hum. Genet.* 43(5): 605-11

Gribskov M., McLachlan A.D. and Eisenberg D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84: 4355-4358.

Groffen J. and Heisterkamp N. (1987) The BCR/ABL hybrid gene. *Baillieres Clin. Haematol.* 1(4): 983-99.

Grosschedl R. and Birnstiel M.L. (1980) Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc Natl Acad Sci USA.* 77(3): 1432-6.

Guarguaglini G., Battistoni A., Pittoggi C., Di Matteo G., Di Fiore B. and Lavia P. (1997) Expression of the murine RanBP1 and Htf9-c genes is regulated from a shared bidirectional promoter during cell cycle progression. *Biochem. J.* 325: 277-86

Halford S., Wadey R., Roberts C., Daw S.C., Whiting J.A., O'Donnell H., Dunham I., Bentley D., Lindsay E. and Baldini A et al. (1993) Isolation of a putative transcriptional regulator from the region of 22q11 deleted in DiGeorge syndrome, Shprintzen syndrome and familial congenital heart disease. *Hum. Mol. Genet.* 2(12): 2099-107.

Halford S., Wilson D.I., Daw S.C., Roberts C., Wadey R., Kamath S., Wickremasinghe A., Burn J., Goodship J. and Mattei M.G. et al. (1993b) Isolation of a gene expressed during early embryogenesis from the region of 22q11 commonly deleted in DiGeorge syndrome. *Hum. Mol. Genet.* 2(10): 1577-82.

Hall K.T., Boumsell L., Schultze J.L., Boussiotis V.A., Dorfman D.M., Cardoso A.A., Bensussan A., Nadler L.M. and Freeman G.J. (1996) Human CD100, a novel

leukocyte semaphorin that promotes B-cell aggregation and differentiation. *Proc. Natl. Acad. Sci. USA* 93(21): 11780-5.

Hamalainen E.R., Kemppainen R., Pihlajaniemi T. and Kivirikko K.I. (1993) Structure of the human lysyl oxidase gene. *Genomics* 17(3): 544-8.

Harley H.G., Brook J.D., Rundle S.A., Crow S., Reardon W., Buckler A.J., Harper P.S., Housman D.E. and Shaw D.J. (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature* 355(6360): 545-6.

Heidaran M.A., Molloy C.J., Pangelinan M., Choudhury G.G., Wang L.M., Fleming T.P., Sakaguchi A.Y. and Pierce J.H. (1992) Activation of the colony-stimulating factor 1 receptor leads to the rapid tyrosine phosphorylation of GTPase-activating protein and activation of cellular p21ras. *Oncogene* 7(1): 147-52.

Heidmann O. and Heidmann T. (1991) Retrotransposition of a mouse IAP sequence tagged with an indicator gene. *Cell* 64(1): 159-70.

Hengen P.N. (1997) Shearing DNA for genomic library construction. *Trends Biochem. Sci.* 22(7): 273-4.

Hentati A., Bejaoui K., Pericak-Vance M.A., Hentati F., Hung W-Y., Figlewicz D.A., Ben Hamida C., Ben Hamida M., Brown R.H. and Siddique T. (1992) The gene locus for one form of juvenile amyotrophic lateral sclerosis maps to chromosome 2. *Neurology* 42(Suppl): 201.

Hofmann K. and Stoffel W. (1993) TMBASE - A database of membrane spanning protein segments. *Biol. Chem.* Hoppe-Seyler 374, 166.

Holmes S.E., Riazi M.A., Gong W., McDermid H.E., Sellinger B.T., Hua A., Chen F., Wang Z., Zhang G., Roe B., Gonzalez I., McDonald-McGinn D.M., Zackai E., Emanuel B.S. and Budarf M.L. (1997) Disruption of the clathrin heavy chain-like gene (CLTCL) associated with features of DGS/VCFS: a balanced (21;22)(p12;q11) translocation. *Hum Mol Genet.* 6(3): 357-67.

Horton P. and Nakai K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Ismb* 5: 147-52.

Hosomi Y., Shii K., Ogawa W., Matsuba H., Yoshida M., Okada Y., Yokono K., Kasuga M., Baba S. and Roth R.A. (1994) Characterization of a 60-kilodalton substrate of the insulin receptor kinase. *J. Biol. Chem.* 269(15): 11498-502.

Ina Y. (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40(2): 190-226.

Innis M.A., Gelfand D.H. and Sninksy (1995) PCR Strategies. Academic Press.

Innis M.A., Myambo K.B., Gelfand D.H. and Brow M.A. (1988) DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. Natl. Acad. Sci. USA* 85(24): 9436-40.

Jeffreys A.J., Wilson V. and Thein S.L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* 314(6006): 67-73.

Jiang W. (1998) Transcript characterization in the region of the mouse neuromuscular mutation mnd2, Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Jiang W., Hua A., Spilson S.V., Miller W., Roe, B. and Meisler M.H. (1999) Gene identification in the mnd2 target region by comparative sequencing of human and mouse BAC clones: implications for analysis of rare transcripts. *Genome Res.* Accepted.

Jing N., Gao X., Rando R.F. and Hogan M.E. (1997) Potassium-induced loop conformational transition of a potent anti-HIV oligonucleotide. *J. Biomol. Struct. Dyn.* 15(3): 573-85.

Johansen F.E. and Prywes R. (1993) Identification of transcriptional activation and inhibitory domains in serum response factor (SRF) by using GAL4-SRF constructs. *Mol. Cell. Biol.* 13(8): 4640-7.

John B and Miklos G.L.G. (1988) The Eukaryote Genome in Development and Evolution. Allen and Unwin, London.

Jones J.M., Albin R.L., Feldman E.L., Simin K., Schuster T.G., Dunnick W.A., Collins J.T., Chrisp C.E., Taylor B.A. and Meisler M.H. (1993) mnd2: a new mouse model of inherited motor neuron disease. *Genomics* 16(3): 669-77.

Kass D.H.,Kim J., Rao A. and Deininger P.L. (1997) Evolution of B2 repeats: the muroid explosion. *Genetica.* 99(1): 1-13.

Katahira J., Inoue N., Horiguchi Y., Matsuda M. and Sugimoto N. (1997b) Molecular cloning and functional characterization of the receptor for Clostridium perfringens enterotoxin. *J. Cell Biol.* 136(6): 1239-47.

Katahira J., Sugiyama H., Inoue N., Horiguchi Y., Matsuda M. and Sugimoto N. (1997a) Clostridium perfringens enterotoxin utilizes two structurally related membrane proteins as functional receptors in vivo. *J. Biol. Chem.* 272(42): 26652-8.

Kaufman P.D., Kobayashi R. and Stillman B. (1997) Ultraviolet radiation sensitivity and reduction of telomeric silencing in Saccharomyces cerevisiae cells lacking chromatin assembly factor-I. *Genes Dev.* 11(3):345-57.

Kedra D., Peyrard M., Fransson I., Collins J.E., Dunham I., Roe B.A. and Dumanski J.P. (1996) Characterization of a second human clathrin heavy chain polypeptide gene (CLH-22) from chromosome 22q11. *Hum. Mol. Genet.* 5(5): 625-31.

Kelley R.I., Zackai E.H., Emanuel B.S., Kistenmacher M., Greenberg F. and Punnett H.H. (1982) The association of the DiGeorge anomalad with partial monosomy of chromosome 22. *J. Pediatr.* 101(2): 197-200.

Kelly D., Goldberg R., Wilson D., Lindsay E., Carey A., Goodship J., Burn J., Cross I., Shprintzen R. and Scambler P. (1993) Confirmation that the velo-cardio-facial syndrome is associated with haplo-insufficiency of genes at chromosome 22q11. Am. J. Med. Genet. 45(3): 308-12.

Kenmochi N., Kawaguchi T., Rozen S., Davis E., Goodman N., Hudson T.J., Tanaka T. and Page D.C. (1998) A map of 75 human ribosomal protein genes. *Genome Res.* 8(5): 509-23.

Klein A. and Meyuhas O. (1984) A multigene family of intron lacking and containing genes, encoding for mouse ribosomal protein L7. *Nucleic Acids Res.* 12(9): 3763-76.

Kliman R.M. and Eyre-Walker A. (1998) Patterns of base composition within the genes of Drosophila melanogaster. *J. Mol. Evol.* 46(5): 534-41.

Kolodkin A.L. (1996) Growth cones and the cues that repel them. *Trends Neurosci* 19(11): 507-13.

Kolodkin A.L., Matthes D.J., O'Connor T.P., Patel N.H., Admon A., Bentley D. and Goodman C.S. (1992) Fasciclin IV: sequence, expression, and function during growth cone guidance in the grasshopper embryo. *Neuron* 9(5): 831-45.

Koop B.F. (1995) Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet.* 11(9): 367-71.

Koop B.F. and Hood L. (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7(1): 48-53.

Koppel A.M. and Raper J.A. (1998) Collapsin-1 covalently dimerizes, and dimerization is necessary for collapsing activity. *J. Biol. Chem.* 273(25): 15708-13.

Kozak M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12(2): 857-72.

Krebs C.J. and Krawetz S.A. (1993) Lysyl oxidase copper-talon complex: a model. *Biochim. Biophys. Acta* 1202(1): 7-12.

Kuff E.L. and Fewell J.W. (1985) Intracisternal A-particle gene expression in normal mouse thymus tissue: gene products and strain-related variability. *Mol. Cell Biol.* 5(3): 474-83.

Kuff E.L. and Lueders K.K. (1988) The intracisternal A-particle gene family: structure and functional aspects. *Adv. Cancer Res.* 51:183-276

Lamerdin J.E., Montgomery M.A., Stilwagen S.A., Scheidecker L.K., Tebbs R.S., Brookman K.W., Thompson L.H. and Carrano A.V. (1995) Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics* 25(2): 547-54.

Lamerdin J.E., Stilwagen S.A., Ramirez M.H., Stubbs L. and Carrano A.V. (1996) Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* 34(3): 399-409.

Lamour V., Lecluse Y., Desmaze C., Spector M., Bodescot M., Aurias A., Osley M.A. and Lipinski M. (1995) A human homolog of the S. cerevisiae HIR1 and HIR2 transcriptional repressors cloned from the DiGeorge syndrome critical region. *Hum. Mol. Genet.* 4(5): 791-9.

Lavia P., Macleod D. and Bird A. (1987) Coincident start sites for divergent transcripts at a randomly selected CpG-rich island of mouse. *EMBO J.* 6(9): 2773-9.

Law S.K., Micklem K.J., Shaw J.M., Zhang X.P., Dong Y., Willis A.C. and Mason D.Y. (1993) A new macrophage differentiation antigen which is a member of the scavenger receptor superfamily. *Eur. J. Immunol.* 23(9): 2320-5.

Lemmon M.A., Ferguson K.M. and Schlessinger J. (1996) PH domains: diverse sequences with a common fold recruit signaling molecules to the cell surface. *Cell* 85(5): 621-4.

Lewin B. (1997) Genes VI. Oxford University Press, New York.

Li W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36(1): 96-9.

Li W.H. (1997) Molecular Evolution. Sinauer Associates, Inc., Massachusetts.

Li W.H. and Graur D. (1991) Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, MA.

Li W.H., Wu C.I. and Luo C.C. (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* 21(1): 58-71.

Lorain S., Demczuk S., Lamour V., Toth S., Aurias A., Roe B.A. and Lipinski M. (1996) Structural Organization of the WD repeat protein-encoding gene HIRA in the DiGeorge syndrome critical region of human chromosome 22. *Genome Res.* 6(1): 43-50.

Lueders K.K. and Mietz J.A. (1986) Structural analysis of type II variants within the mouse intracisternal A-particle sequence family. *Nucleic Acids Res.* 14(3): 1495-510.

Lueders K.K. and Mietz J.A. (1986) Structural analysis of type II variants within the mouse intracisternal A-particle sequence family. *Nucleic Acids Res.* 14(3): 1495-510.

Makalowski W., Zhang J. and Boguski M.S. (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* 6(9): 846-57.

Maraia R.J. (1991) The subset of mouse B1 (Alu-equivalent) sequences expressed as small processed cytoplasmic transcripts. *Nucleic Acids Res.* 19(20):5695-702.

Mascarello J.T., Bastian J.F. andJones M.C. (1989) Interstitial deletion of chromosome 22 in a patient with the DiGeorge malformation sequence. *Am. J .Med. Genet.* 32(1): 112-4.

Matsuo M. (1995) Duchenne muscular dystrophy. *Southeast Asian J Trop Med Public Health.* 26 Suppl 1: 166-71.

Maueler W., Kyas A., Keyl H.G., Epplen J.T. (1998) A genome-derived (gaa.ttc)24 trinucleotide block binds nuclear protein(s) specifically and forms triple helices. *Gene* 215(2): 389-403.

Maxam A.M. and Gilbert W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* USA 74(2): 560-4.

McWhirter J.R. and Wang J.Y. (1991) Activation of tyrosinase kinase and microfilament-binding functions of c-abl by bcr sequences in bcr/abl fusion proteins. *Mol. Cell Biol.* 11(3): 1553-65.

Mollenhauer J., Wiemann S., Scheurlen W., Korn B., Hayashi Y., Wilgenbus K.K., von Deimling A. and Poustka A. (1997) DMBT1, a new member of the SRCR superfamily, on chromosome 10q25.3-26.1 is deleted in malignant brain tumours. *Nat. Genet.* 17(1): 32-9.

Musacchio A., Gibson T., Rice P., Thompson J. and Saraste M. (1993) The PH domain: a common piece in the structural patchwork of signalling proteins. *Trends Biochem. Sci.* 18(9): 343-8.

Nielsen H., Engelbrecht J., Brunak S. and von Heijne G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10(1): 1-6.

Nunes D.P., Keates A.C., Afdhal N.H. and Offner G.D. (1995) Bovine gall-bladder mucin contains two distinct tandem repeating sequences: evidence for scavenger receptor cysteine-rich repeats. *Biochem. J.* 310 (Pt 1): 41-8.

Oakey R.J., Watson M.L. and Seldin M.F. (1992) Construction of a physical map on mouse and human chromosome 1: comparison of 13 Mb of mouse and 11 Mb of human DNA. *Hum. Mol. Genet.* 1(8): 613-20.

Oeltjen J.C., Malley T.M., Muzny D.M., Miller W., Gibbs R.A. and Belmont J.W. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* 7(4): 315-29.

Ono M., Toh H., Miyata T. and Awaya T. (1985) Nucleotide sequence of the Syrian hamster intracisternal A-particle gene: close evolutionary relationship of type A particle gene to types B and D oncovirus genes. *J. Virol.* 55(2): 387-94.

Osada S., Daimon S., Nishihara T. and Imagawa M. (1996a) Identification of DNA binding-site preferences for nuclear factor I-A. *FEBS Lett* 390(1): 44-6.

Osada S., Yamamoto H., Nishihara T. and Imagawa M. (1996b) DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. *Biol. Chem.* 271(7): 3891-6.

Ozelius L.J., Hewett J.W., Page C.E., Bressman S.B., Kramer P.L., Shalish C., de Leon D., Brin M.F., Raymond D., Corey D.P., Fahn S., Risch N.J., Buckler A.J., Gusella J.F. and Breakefield X.O. (1997) The early-onset torsion dystonia gene (DYT1) encodes an ATP-binding protein. *Nat. Genet.* 17(1): 40-8.

Page D.C., Mosher R., Simpson E.M., Fisher E.M., Mardon G., Pollack J., McGillivray B., de la Chapelle A. and Brown L.G. (1987) The sex-determining region of the human Y chromosome encodes a finger protein. *Cell* 51(6): 1091-104.

Pamilo P. and Bianchi N.O. (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.* 10(2): 271-81.

Pata I. and Metspalu A. (1996) Structural characterization of the mouse ribosomal protein S6-encoding gene. *Gene* 175(1-2): 241-5.

Paterson B.M., Segal S., Lueders K.K. and Kuff E.L. (1978) RNA associated with murine intracisternal type A particles codes for the main particle protein. *J. Virol.* 27(1): 118-26.

Pergin Elmer (1995) ABI Prism 377 DNA sequencer, User's Manual. The Perkin-Elmer Corporation.

Piko L., Hammons M.D. and Taylor K.D. (1984) Amounts, synthesis, and some properties of intracisternal A particle-related RNA in early mouse embryos. *Proc. Natl. Acad. Sci. USA* 81(2): 488-92.

Pley U. and Parham P. (1993) Clathrin: its role in receptor-mediated vesicular transport and specialized functions in neurons. *Crit. Rev. Biochem. Mol. Biol.* 28(5): 431-64.

Prestridge D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Appl. Biosci.* 12(2): 157-60.

Prober J.M., Trainor G.L., Dam R.J., Hobbs F.W., Robertson C.W., Zagursky R.J., Cocuzza A.J., Jensen M.A. and Baumeister K. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238(4825): 336-41.

Pu R.T. and Dasso M. (1997) The balance of RanBP1 and RCC1 is critical for nuclear assembly and nuclear transport. *Mol. Biol. Cell* 8(10): 1955-70.

Puech A., Dupressoir A., Loireau M.P., Mattei M.G. and Heidmann T. (1997) Characterization of two age-induced intracisternal A-particle-related transcripts in the mouse liver. Transcriptional read-through into an open reading frame with similarities to the yeast ccr4 transcription factor. *J. Biol. Chem.* 272(9): 5995-6003.

Quentin Y. (1992) Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Res.* 20(13): 3397-401.

Rabin S.J., Cleghon V. and Kaplan D.R. (1993) SNT, a differentiation-specific target of neurotrophic factor-induced tyrosine kinase activity in neurons and PC12 cells. *Mol. Cell Biol.* 13(4): 2203-13.

Rechavi G., Givol D. and Canaani E. (1982) Activation of a cellular Oncogene by DNA rearrangement: possible involvement of an IS-like element. *Nature* 300(5893): 607-11.

Reese M.G. and Eeckman F.H. (1995) in The Seventh International Genome Sequencing and Analysis Conference, Hilton Head Island, Sourth Carolina.

Reimold A.M., Grusby M.J., Kosaras B., Fries J.W., Mori R., Maniwa S., Clauss I.M., Collins T., Sidman R.L., Glimcher M.J. and Glimcher L.H. (1996) Chondrodysplasia and neurological abnormalities in ATF-2-deficient mice. *Nature* 379(6562): 262-5.

Resnick D., Pearson A. and Krieger M. (1994) The SRCR superfamily: a family reminiscent of the Ig superfamily. *Trends Biochem. Sci.* 19(1): 5-8.

Riggins G.J., Lokey L.K., Chastain J.L., Leiner H.A., Sherman S.L., Wilkinson K.D. and Warren S.T. (1992) Human genes containing polymorphic trinucleotide repeats. *Nat. Genet.* 2(3): 186-91.

Roberts C., Daw S.C., Halford S. and Scambler P.J. (1997) Cloning and developmental expression analysis of chick Hira (Chira), a candidate gene for DiGeorge syndrome. *Hum. Mol. Genet.* 6(2): 237-45.

Robertson H.D., Barrell B.G., Weith H.L. and Donelson J.E. (1973) Isolation and sequence analysis of a ribosome-protected fragment from bacteriophage phiX 174 DNA. *Nat. New Biol.* 241(106): 38-40.

Robinson M.S. (1994) The role of clathrin, adaptors and dynamin in endocytosis. *Curr. Opin. Cell Biol.* 6(4): 538-44.

Roe B. (1995) Protocols for Recombinant DNA Isolation, Cloning and Sequencing. University of Oklahoma.

Rossetti L., Cacchione S., Fua M. and Savino M. (1998) Nucleosome assembly on telomeric sequences. *Biochemistry* 37(19): 6727-37.

Rost B., Casadio R., Fariselli P. and Sander C. (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci.* 4(3): 521-33.

Rush M.G., Drivas G. and D'Eustachio P. (1996) The small nuclear GTPase Ran: how much does it run? *Bioessays* 18(2): 103-12.

Saito H., Papaconstantinou J., Sato H. and Goldstein S. (1997) Regulation of a novel gene encoding a lysyl oxidase-related protein in cellular adhesion and senescence. *J. Biol. Chem.* 272(13): 8157-60.

Sanger F., Nicklen S. and Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74(12): 5463-7.

Scambler P., Carey A.H., Wyse R.K., Roach S., Dumanski J.P., Nordenskjold M. and Williamson R. (1991) Microdeletions within 22q11 associated with sporadic and familial DiGeorge syndrome. *Genomics* 10(1): 201-6.

Scambler P., Kelly D., Lindsay E., Williamson R., Goldberg R., Shprintzen R., Wilson D., Goodship J., Cross I. and Burn J. (1992) Velo-cardio-facial syndrome associated with chromosome 22 deletions encompassing the DiGeorge locus. *Lancet* 339(8802): 1138-9

Schlotterer C. (1998) Genome evolution: are microsatellites really simple sequences? *Curr. Biol.* 8(4): R132-4.

Searle A.G., Edwards J.H. and Hall J.G. (1994) Mouse homologues of human hereditary disease. *J. Med. Genet.* 31(1): 1-19.

Sen D. and Gilbert W. (1990) A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* 344(6265): 410-4.

Sertedaki A. and Lindsay S. (1996) CAC--the neglected repeat. *Bioessays* 18(3): 237-42.

Shaw G. and Kamen R. (1986) A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell* 46(5): 659-67.

Shehee W.R., Loeb D.D., Adey N.B., Burton F.H., Casavant N.C., Cole P., Davies C.J., McGraw R.A., Schichman S.A. and Severynse D.M. et al. (1989) Nucleotide sequence of the BALB/c mouse beta-globin complex. *J. Mol. Biol.* 205(1): 41-62.

Shen-Ong G.L. and Cole M.D. (1982) Differing populations of intracisternal A-particle genes in myeloma tumors and mouse subspecies. *J. Virol.* 42(2): 411-21.

Shprintzen R., Goldberg R., Young D. and Wolford L. (1981) The velo-cardio-facial syndrome: a clinical and genetic analysis. *Pediatrics* 67: 167-172.

Siddique T., Figlewicz D.A., Pericak-Vance M.A., Haines J.L., Rouleau G., Jeffers A.J., Sapp P., Hung W.Y., Bebout J., McKenna-Yasek D., et al. (1991) Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *N. Engl .J .Med.* 324(20): 1381-4.

Singer M. and Berg P. (1991) Genes and Genomes. University *Science* Books, California.

Sirotkin H., Morrow B., DasGupta R., Goldberg R., Patanjali S.R., Shi G., Cannizzaro L., Shprintzen R., Weissman S.M. and Kucherlapati R. (1996) Isolation of a new clathrin heavy chain gene with muscle-specific expression from the region commonly deleted in velo-cardio-facial syndrome. *Hum. Mol. Genet.* 5(5): 617-24.

Sirotkin H., Morrow B., Saint-Jore B., Puech A., Das Gupta R., Patanjali S.R., Skoultchi A., Weissman S.M. and Kucherlapati R. (1997) Identification, characterization, and precise mapping of a human gene encoding a novel membrane-spanning protein from the 22q11 region deleted in velo-cardio-facial syndrome. *Genomics* 42(2): 245-51.

Smale S.T. (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim Biophys Acta* 1351(1-2): 73-88.

Smit A.F. (1996) The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev.* 6(6): 743-8.

Smith D.P., Johnstone E.M., Little S.P. and Hsiung H.M. (1990) Direct DNA sequencing of cDNA inserts from plaques using the linear polymerase chain reaction. *Biotechniques* 9(1): 48-52.

Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B. and Hood L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321(6071): 674-9.

Snyder E.E. and Stormo G.D. (1993) Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* 21(3): 607-13.

Solovyev V. and Salamov A. (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Ismb* 5: 294-302.

Somma M.P., Pisano C. and Lavia P. (1991) The housekeeping promoter from the mouse CpG island HTF9 contains multiple protein-binding elements that are functionally redundant. *Nucleic Acids Res.* 19(11): 2817-24.

Sondek J., Bohm A., Lambright D.G., Hamm H.E. and Sigler P.B. (1996) Crystal structure of a G-protein beta gamma dimer at 2.1A resolution. *Nature* 379(6563): 369-74.

Sonnhammer E.L. and Durbin R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167(1-2): GC1-10.

Staden R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12(1 Pt 2): 505-19.

Suhan M.L. and Hovde C.J. (1998) Disruption of an internal membrane-spanning region in Shiga toxin 1 reduces cytotoxicity. *Infect. Immun.* 66(11): 5252-9.

Tabor S. and Richardson C.C. (1989) Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J. Biol. Chem.* 264(11): 6447-58.

Tabor S. and Richardson C.C. (1995) A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. USA* 92(14): 6339-43.

Toyoda A., Fukumaki Y., Hattori M. and Sakaki Y. (1995) Mode of activation of the GC box/Sp1-dependent promoter of the human NADH-cytochrome b5 reductase-encoding gene.

Whitehead A.S. and Sackstein R. (1985) Molecular biology of the human and mouse MHC class III genes: phylogenetic conservation, genetics and regulation of expression. *Immunol. Rev.* 87: 185-208.

Vuillaumier S., Dixmeras I., Messai H., Lapoumeroulie C., Lallemand D., Gekas J., Chehab F.F., Perret C., Elion J. and Denamur E. (1997) Cross-species characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene reveals multiple levels of regulation. *Biochem. J.* 327 (Pt 3): 651-62.

Wagner R., Maguire M. and Stallings R. (1993) Chromosomes: A Synthesis. Wiley, New York.

Wahle E. and Keller W. (1996) The biochemistry of polyadenylation. *Trends Biochem. Sci.* 21(7): 247-50.

Wang C.Y., Petryniak B., Ho I.C., Thompson C.B. and Leiden J.M. (1992) Evolutionarily conserved Ets family members display distinct DNA binding specificities. *J. Exp. Med.* 175(5): 1391-9.

Watson J.D. and Crick F.H.C. (1953) Molecular structure of nucleic acid. A structure of the deoxyribose nucleic acid. *Nature* 171: 737-738.

Winship P.R. (1989) An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. *Nucleic Acids Res.* 17(3): 1266.

Xu H., Lee K.W. and Goldfarb M. (1998) Novel recognition motif on fibroblast growth factor receptor mediates direct association and activation of SNT adapter proteins. *J. Biol. Chem.* 273(29): 17987-90.

Xu Y., Mural R. J., Shah M., Uberbacher E.C. (1994) Recognizing Exons in Genomic Sequence Using GRAIL II. *Genetic Engineering* 16: 241-53.

Yamanashi Y. and Baltimore D. (1997) Identification of the Abl- and rasGAP-associated 62 kDa protein as a docking protein, Dok. *Cell* 88(2): 205-11.

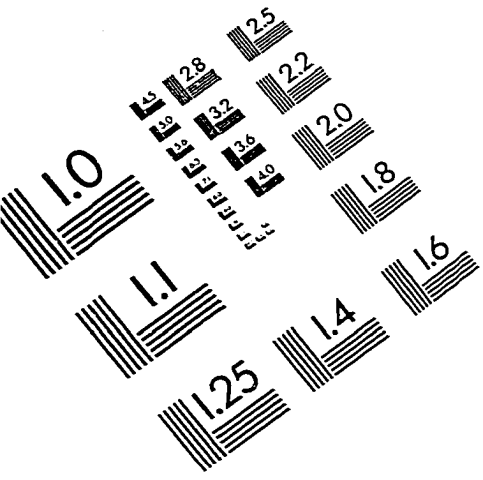
Ymer S., Tucker W.Q., Campbell H.D. and Young I.G. (1986) Nucleotide sequence of the intracisternal A-particle genome inserted 5' to the interleukin-3 gene of the leukemia cell line WEHI-3B. *Nucleic Acids Res.* 14(14): 5901-18.

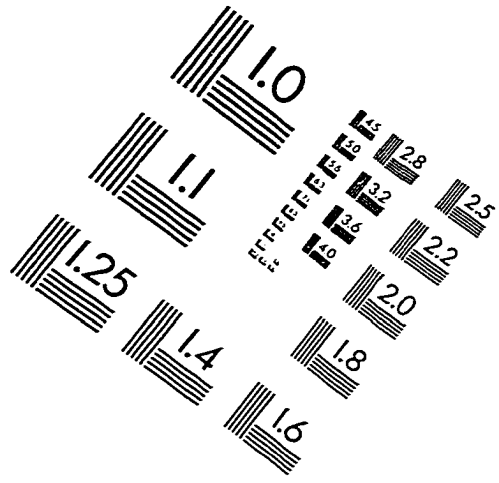
Zhang J. and Madden T.L. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* 7(6): 649-56

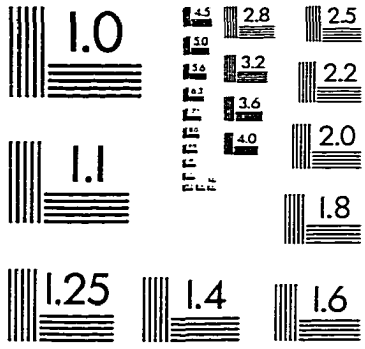
Zhang M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum Mol Genet.*; 7(5): 919-32.

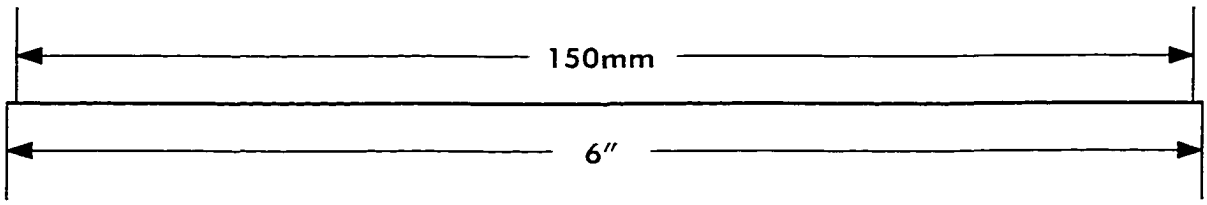
Zou L. and Stillman B. (1998) Formation of a preinitiation complex by S-phase cyclin CDK-dependent loading of Cdc45p onto chromatin. *Science* 280(5363): 593-6.

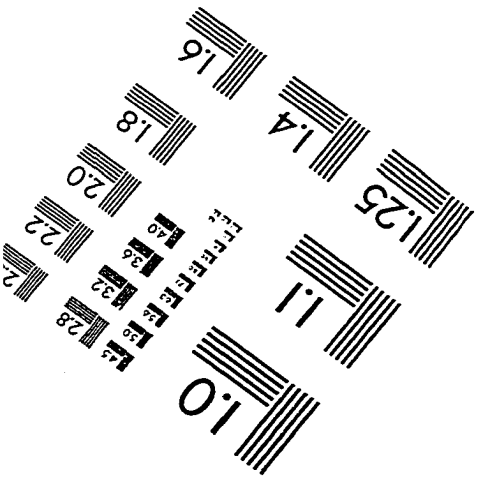
IMAGE EVALUATION
TEST TARGET (QA-3)

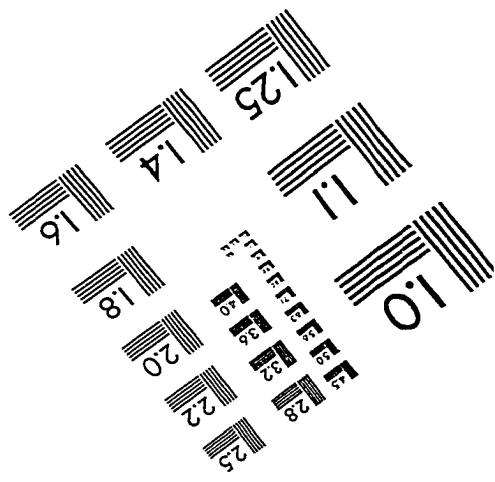












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989