CHARACTERIZATION OF THE 5' REGION OF THE HUMAN METHYLENETETRAHYDROFOLATE REDUCTASE (MTHFR) GENE

by

MANUEL CHAN

Department of Biology

McGill University

Montreal, Quebec, Canada

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirement of the degree of

Master's of Science

May 1999

0-612-50734-3

“Gloria in excelsis Deo”

This thesis is dedicated to my parents and my brothers Antonio and Carlos.

Table of contents Page

Abstract

Methylenetetrahydrofolate reductase (MTHFR) catalyses the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a methyl donor for the re-methylation of homocysteine to methionine. A thermolabile variant of this enzyme, present in approximately 35% of alleles in the North American population, has been associated with cardiovascular disease, neural tube defects, and colon cancer. A cDNA of 2.2kb for human MTHFR has been expressed and results in an active enzyme, but the cDNA and genomic sequences 5' to the ATG start site have not been adequately investigated. The characterization of the 5' region of the human MTHFR gene is reported here. Four additional 5' exonic sequences were localized to a 4kb genomic fragment. The original exon 1 extends directly upstream into a 5' UTR. Three other 5' exons (two with open reading frames) are alternatively spliced into a common splice acceptor site, generating cDNAs with 4 possible 5' ends. The N-terminal peptide sequence of the porcine MTHFR has not been identified in the human sequence suggesting that the missing human coding sequence might be localized further upstream or not conserved across species. A putative chloride ion channel gene (ClC-6) was located in the opposite orientation, at 3.5kb upstream of the original ATG codon, suggesting an overlap with the MTHFR gene and potential co-localization of regulatory elements. A CpG island was identified in the region of a 5' exon (43S) suggesting that a transcription start site and a promoter might be nearby. This work is relevant in understanding the regulation of this important enzyme in folate metabolism.

Résumé

L'enzyme méthylènetétrahydrofolate réductase (MTHFR) catalyse la réduction du 5, 10-méthylènetétrahydrofolate en 5-méthyltétrahydrofolate, un donneur de groupe méthyl dans la reméthylation de l'homocysteïne en méthionine. Une variante thermolabile de cette enzyme, présente dans environ 35% des allèles de la population en Amerique du nord, a été associée aux maladies cardio-vasculaires, aux défauts du tube neural, au cancer du colon. Un ADNc de 2.2 kb provenant de MTHFR humain a été exprimé et a produit un enzyme actif, mais cet ADNc et les séquences en 5' du codon de départ ATG n'ont pas été adéquatement étudiés. La caractérisation de la région en 5' du gène MTHFR humain est rapportée ici. Quatre séquences additionnelles exoniques en 5' ont été localisées dans un fragment génomique de 4 kb. L'exon 1 original s'étend directement en avant dans le 5' UTR. Trois autres exons en 5' (2 avec cadre de lecture ouvert) sont alternativement excisés dans un site accepteur commun d'excision, générant des ADNc avec 4 extrêmités 5' possibles. La séquence peptide en N-terminal du MTHFR porcine n'a pas été retrouvée dans la séquence humaine ce qui suggère que la séquence codante qui manque chez l'homme pouvrait être localisée plus en amont ou ne pas être conservée en passant d'une espèce à l'autre. Ce qui à tout l'air d'être un gène pour un canal d'ion chlorure a été localisé dans l'orientation contraire, à 3.5 kb en amont du codon original ATG, ce qui suggère un chevauchement du gène de MTHFR et des éléments régulateurs potentiels. Un ilôt CpG a été identifié dans la région de l'exon en 5' (43S), ce qui suggère qu'un site de départ de la transcription et un promoteur puissent être presents dans le voisinage. Ce travail est pertinent pour comprendre la régulation de cette enzyme important du métabolisme du folate.

Acknowledgements

I would like to thank my supervisor Dr. Rima Rozen for giving me this great opportunity to work on this project and for her advice and encouragement that helped me overcome the challenges of the project and the thesis. I would also like to thank Dr. Roy Gravel and Dr. Greg Brown for their invaluable comments and suggestions. Thanks to Daniel Leclerc for his technical advice. Thanks to François for translating the abstract into French and helping me inside and outside the lab. Thanks to Philippe for helping me when I first started the project and bringing delicious cheesecakes to the lab. Thanks to Renate, Maria and Carmen for showing me the basic techniques when I first came. Thanks to Nelly for sharing her thoughts and experience as well as putting through numerous rush orders for me. Thanks to Irfan for his encouraging suggestions when I got stuck. Thanks to Pam for listening to my never-ending story about the 5' end and being a cheerful friend. Thanks to Peppy for splitting the cells for me and introducing me to the "Tragically Hip". Thanks to Zhoutao for listening to my "big" theory about the experiments as well as sharing many different aspects of life outside science. Thanks to Ilan for taking care of my gels sometimes and buying me coffees. Thanks to Sahar for offering her PCR slots sometimes and sharing her hilarious jokes and good candies. Thanks to Benedicte, Josée, Wei, Patricia, Qing, Xin, Tak, Mei, Liyuan, Erin, Thida and Dan for their company.

I thank my parents for bringing me to Canada and taking care of me in all these years. Thanks to my brothers for sharing many home responsibilities. Although unable to mention all the names, I would also like to thank all my friends for their understanding and acceptance during the pass few years. Finally, I would like to thank Connie for her love, care, and patience and for being there always to listen, to accept, and to encourage.

ABBREVIATIONS

ATCC	American Type Culture Collection
BHMT	Betaine homocysteine methyl transferase
BSA	Bovine serum albumin
CAD	Coronary artery disease
CBS	Cystathionine β synthase
DEPC	Diethyl pyrocarbonate
DHF	Dihydrofolate
DTT	Dithiothreitol
EDTA	Disodium ethylenediamine tetraacetate
EST	Expressed sequence tag
FAD	Flavin-adenine dinucleotide
Hcy	Homocysteine
IPTG	Isopropyl-1-thio-β-D-galactoside
MS	Methionine synthase
MTHFR	Methylenetetrahydrofolate reductase
NCBI	National Center for Biotechnology Information
NTD	Neural tube defects
ORF	Open reading frame
PAC	P1 artificial chromosome
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
PI	Internal peptide of porcine MTHFR

PLP	Pyridoxal 5'-phosphate
PN	N-terminal peptide of porcine MTHFR
RACE-PCR	Rapid Amplification of cDNA ends PCR
RT	Reverse transcriptase (transcription)
RT-PCR	Reverse transcription PCR
SAM	S-adenosyl methionine (Ado-Met)
SDS	Sodium dodecyl sulfate
SSC	Sodium chloride/ sodium citrate
THF	Tetrahydrofolate
UTR	Untranslated region
XGAL	5-bromo-4-chloro-3-indolyl-β-D-galactoside

List of Figures Page

List of Tables Page

1. Introduction

1.1 Folate

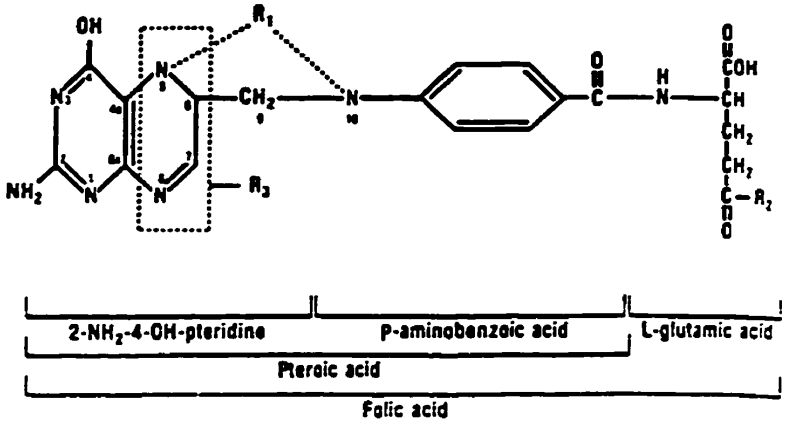
Folic acid (pteroylmonoglutamic acid), a synthetic group B vitamin, is the parent compound of a large group of naturally-occurring derivatives called folates (pteroylglutamates). Folates are coenzymes that are involved in several critical single-carbon transfer reactions such as the biosynthesis of purines, pyrimidines, serine and methionine and degradation of histidine and purines (Rosenblatt 1995). Since mammals are unable to synthesize folates de novo, they have to obtain folates in their diet. Folates are mainly found in liver, dry beans, and green vegetables such as spinach and broccoli. As early as the 1930's, Wills and her colleagues showed that folate deficiency resulted in macrocytic megaloblastic anemia (Wills 1931, Wills et al. 1935), at about the same time that pteridine pigments were isolated from butterfly wings. Folate has now been recognized as an important vitamin that reduces the risk of vascular disease (e.g. Boushey 1995), neural tube defects (e.g. MRC 1991), and cancer (e.g. Glynn 1994).

1.1.1 Folate derivatives

The term "folate" is often used as a generic term to represent the folic acid derivatives that vary at the one-carbon substitution at N5 and N10 (R1), the number of glutamate residues (R2) and the levels of reduction of the pteridine ring (R3) (Fig. 1). The biologically active forms of folate are the derivatives of 5,6,7,8-tetrahydrofolate (THF). Among all the others, 5-methyltetrahydrofolate is the predominant form of folate in the plasma and in many tissues (Rosenblatt 1995). Dietary folates, mostly in the form of

Figure 1. The structure of folic acid and its derivatives.

Reproduced from: Rosenblatt DS (1995) Inherited disorders of folate transport and metabolism. *The metabolic basis of Inherited Disease, 7th ed. Vol. 2* (eds Scriver CR, Beaudet AL, Sly WS and Valle D).



R_1

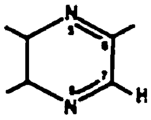
$—CH_3$ at N^5, — H at N^{10}:	5-methyl
$—CHO$ at N^5, —H at N^{10}:	5-formyl (citrovorum factor, folinic acid, leucovorin)
$—CHNH$ at N^5, —H at N^{10}:	5-formimino
$\overset{+}{=}CH=$ at N^5, N^{10}:	5,10-methenyl
$—CH_2—$ at N^5, N^{10}:	5,10-methylene
—H at N^5, CHO at N^{10}:	10-formyl

R_2

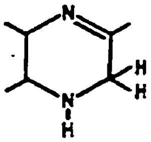
Poly(γ-glutamyl)$_n$ glutamic acids

—OH

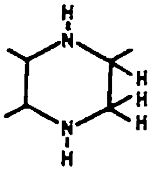
R_3



Unreduced pteridine group



7,8 Dihydro (H_2Folate)



5,6,7,8 Tetrahydro (H_4Folate)

polyglutamate, have to be hydrolyzed into their monoglutamate form before they are taken up into the cell (Fan 1992, Rosenblatt 1995).

1.1.2 Folate transport

The concentration of folates in red blood cells is 30 times higher than in the plasma; and in the cerebrospinal fluid, it is 3 times higher than in the plasma (Fan 1992). To transport against the gradient, an active transport system across the cell membrane and the blood brain barrier is necessary. Two folate transport systems across the mammalian cell membrane have been reported. One works at high folate concentration (μM scale) while the other works at low folate concentration (nM scale). Patients with inherited folate malabsorption have been described (Rosenblatt 1995). The concentration of intracellular folate is critical for the activity of folate-dependent enzymes and for the optimal growth rate of the cell. For human cells, the amount of folate ranges from 50nM in fibroblasts to 1μM in lymphocytes and tumor cells (Watkins 1983). In fact, all cells with a high turnover rate, such as lymphocytes, have a high demand for folate because of the involvement of folate in DNA synthesis.

1.1.3 Folate Metabolism

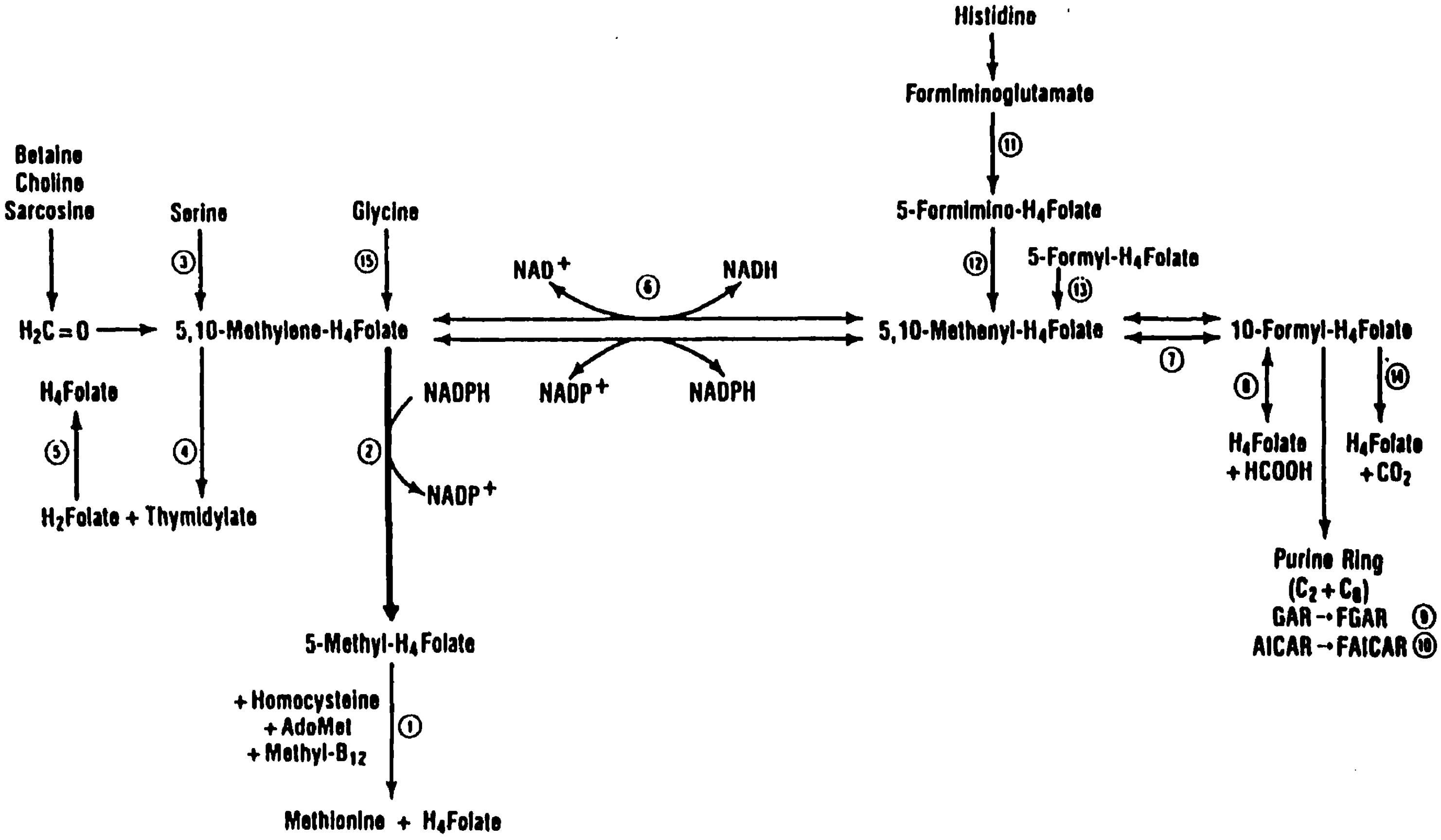
As mentioned earlier, folic acid is a synthetic vitamin that does not exist in nature. Dietary folates are mainly in polyglutamate form. Polyglutamate folates are hydrolyzed into their monoglutamate forms, by pteroylpolyglutamate hydrolase in the intestine, which are then absorbed into the cell by active transport. Monoglutamate folates are released into the circulation, with 5-methyltetrahydrofolate (5-methylTHF) being the most predominant form (Rosenblatt 1995). After entering the cell, 5-methylTHF donates the methyl group in the biosynthesis of methionine from homocysteine, regenerating THF. THF then becomes the substrate of polyglutamate synthetase, which converts monoglutamate folates into polyglutamate folates, the storage form of folates (Zittoun 1995). Folic acid and 5-methylTHF, however, are poor substrates for polyglutamate synthetase. Polyglutamate folates have much lower Km values than monoglutamate folates for some folate-dependent reactions, hence allowing the reaction to occur at lower intracellular folate concentration (Rosenblatt 1995).

Folates, as mentioned earlier, are involved in many critical single-carbon transfer reactions such as the biosynthesis of purines, pyrimidines, serine and methionine and degradation of histidine and purines (Rosenblatt 1995). Single carbon units from serine and glycine enter this single-carbon pool in the form of 5,10-methyleneTHF (Fig. 2, Reaction 3 and 15 respectively). 5,10-methyleneTHF can be used directly in thymidylate synthesis by thymidylate synthase (Fig. 2, Reaction 4). Formation of thymidylate also oxidizes THF into dihydrofolate (DHF). Synthetic folic acid and DHF enter the folate metabolic pathway by the action of dihydrofolate reductase which reduces them into THF (Fig. 2, Reaction 5). Methylenetetrahydrofolate reductase (MTHFR) reduces 5,10-methyleneTHF into 5-methylTHF, a methyl donor in the biosynthesis of methionine (Fig. 2, Reaction 2 and 4 respectively). 5,10-methyleneTHF can also be oxidized to 10-formylTHF for purine synthesis (Fig. 2, Reaction 6 and 7).

Figure 2. Pathway of folate metabolism

1) Methionine synthase (methyltetrahydrofolate:homocysteine methyltransferase). 2) Methylenetetrahydrofolate reductase. 3) Serine hydroxymethyltransferase. 4) Thymidylate synthase. 5) Dihydrofolate reductase. 6) Methylenetetrahydrofolate dehydrogenase. 7) Methenyl-tetrahydrofolate cyclohydrolase. 8) 10-formyltetrahydrofolate reductase. 9) GAR (5-phosphoribosylglycineamide) transformylase. 10) AICAR (5-phosphoribosyl-5-aminoimidazole-4-carboxamide) transformase. 11) Glutamate formiminotransferase. 12) formiminotetrahydrofolate cyclodeaminase. 13) 5,10-methylenyl-tetrahydrofolate synthase. 14) 10-formyl-tetrahydrofolate dehydrogenase. 15) Glycine cleavage pathway.

[Reproduced from: Rosenblatt DS (1995) Inherited disorders of folate transport and metabolism. *The metabolic basis of Inherited Disease, 7th ed. Vol. 2* (eds Scriver CR, Beaudet AL, Sly WS and Valle D).]


Histidine
Formiminoglutamate
11
5-Formimino-H_4Folate
12
5-Formyl-H_4Folate
13
Betaine
Choline
Sarcosine
Serine
3
Glycine
15
NAD^+
6
NADH
$H_2C=O$
5,10-Methylene-H_4Folate
5,10-Methenyl-H_4Folate
10-Formyl-H_4Folate
7
$NADP^+$
NADPH
H_4Folate
5
4
2
NADPH
$NADP^+$
H_2Folate + Thymidylate
8
14
H_4Folate
+ HCOOH
H_4Folate
$+ CO_2$
Purine Ring
$(C_2 + C_8)$
GAR → FGAR 9
AICAR → FAICAR 10
5-Methyl-H_4 Folate
+ Homocysteine
+ AdoMet
+ Methyl-B_{12}
1
Methionine + H_4Folate

1.1.4 Folate and multifactorial diseases

i) Folate and vascular disease

Increased folate intake has been shown to reduce plasma homocysteine, a risk factor in vascular disease. In a meta-analysis of 27 studies on the effects of folate on homocysteine and vascular disease, it was calculated that the fortification of 350μg folic acid per 100g food could prevent 9% of male and 54% of female coronary artery deaths in the United States (Boushey et al. 1995). Folate reduces homocysteine by the re-methylation pathway; the metabolism of homocysteine, which will be discussed in section 1.2.2. Although the mechanisms of how homocysteine causes vascular disease is still not completely clear, mild hyperhomocysteinemia is recognized as a risk factor for many forms of vascular disease (see section 1.2.5).

ii) Folate and Neural Tube Defects

Periconceptional use of folic acid by pregnant woman has been shown to reduce the occurrence and recurrence of neural tube defects (NTDs) (Rieder 1994, Czeizel et al. 1992, MRC 1991). Maternal folate and vitamin B12 have been suggested as independent risk factors of NTDs (Kirke et al. 1993). Hyperhomocysteinemia has been associated with NTDs (Steegers-Theunissen et al. 1994, Mills et al. 1995). To prevent the occurrence of NTDs, a supplementation of 0.4 mg folic acid is recommended to women who might become pregnant (Centers for Disease Control 1992). Although the underlying etiology of NTDs is still unclear, it has long been known that there exists a genetics-environmental interaction. Both genetic defects and teratogen exposure could lead to increased risk for NTDs (Minns 1996). Several studies have reported an association between homozygosity for the 677C→T mutation in MTHFR and an increased risk of NTDs (Whitehead et al. 1995, van der Put et al. 1995 and 1996, Ou et al. 1996, Kirke et al. 1996). This mutation is discussed in greater details in sections 1.3.3 and 1.3.4.

iii) Folate and Cancers

Cancer is known to be caused by accumulation of DNA changes. Activation of proto-oncogenes and inactivation of tumor suppressor genes may result from structural changes in DNA such as translocations, rearrangements and mutations as well as changes in DNA methylation. Although the mechanism of cancer induction by folate deficiency is not established, folate deficiency has been shown to correlate with increased risk for cancer (Butterworth 1993, Giovannucci et al. 1993, 1995), and folate supplementation has been shown to reduce the risk of colon cancer (Baron et al. 1998). Two hypotheses have been proposed for carcinogenesis due to folate deficiency. Folate deficiency could cause increased cancer risk by enhancing chromosome fragility because of uracil misincorporation into DNA, or by abnormal DNA methylation (Herbert 1986). Folate deficiency reduces thymidylate synthesis from dUMP and hence increases the chance of uracil misincorporation. SAM level is also reduced due to folate deficiency, since folates are necessary for methionine and SAM synthesis. SAM is involved in many important methylation reactions such as DNA methylation. Interestingly, the homozygous 677C→T mutation in MTHFR has been correlated with reduced risk in colon cancer (Ma et al. 1997), despite its role in plasma homocysteine elevation and as a genetic risk factor for cardiovascular disease.

1.2 Homocysteine

1.2.1 Homocysteine

Homocysteine is a sulphur amino acid discovered in 1932 by deVigneaud as an intermediate in the metabolic pathway of methionine (DeVigneaud 1952). The term "homocysteine" often refers to total homocysteine (tHcy) which is the sum of the amino acid homocysteine, the homocysteine-homocysteine disulfide (homocystine) and the cysteine-homocysteine mixed disulfide. Homocysteine is unstable and is readily oxidized to homocystine and cysteine-homocysteine disulfide. Since there is a cellular mechanism to excrete homocysteine into the plasma, the plasma of a normal individual contains a small amount of homocysteine of about 5-15 μM (Ueland et al. 1993). About 80% of plasma homocysteine is present in the protein-bound form in a normal individual (Kang et al. 1979). Hyperhomocysteinemia refers to the above normal homocysteine level in blood, plasma or serum, which can be moderate (about 16-30 μM), intermediate (about 31-100μM) or severe (above 100μM) (Kang et al. 1992). An elevated level of homocysteine is now recognized as a risk factor for coronary artery disease (CAD), cerebrovascular disease and peripheral arterial vascular disease (Boushey et al. 1995, Malinow 1994).

1.2.1 Homocysteine metabolism

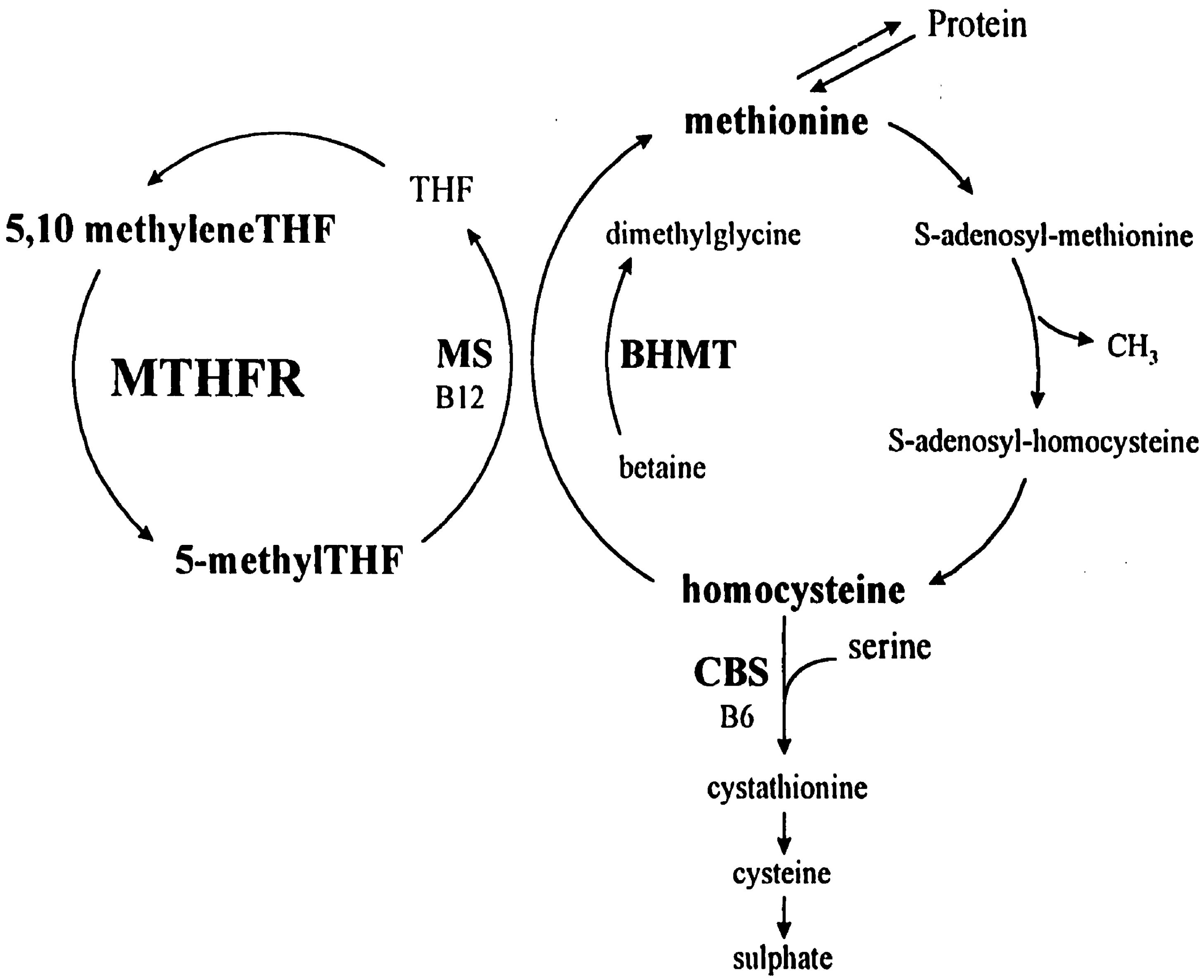
Homocysteine levels are regulated by two major pathways, the re-methylation and the trans-sulfuration pathways (Fig. 3). In the trans-sulfuration pathway, cystathionine β synthase (CBS) is the first enzyme involved in the degradation of homocysteine. In the re-methylation pathway, methionine synthase (MS) re-methylates homocysteine into

Figure 3. Pathways of homocysteine metabolism

The re-methylation and the trans-sulfuration pathways of homocysteine are shown.

THF = Tetrahydrofolate
MTHFR = Methylenetetrahydrofolate reductase
MS = Methionine synthase
BHMT = Betaine homocysteine methyl transferase
CBS = Cystathionine β synthase
B6 = vitamin B6
B12 = vitamin B12

[adapted from Gallagher et al. (1996) *Circulation* 94: 2154-2158]


Protein
methionine
THF
5,10 methyleneTHF
dimethylglycine
S-adenosyl-methionine
MTHFR
MS
B12
BHMT
CH_3
betaine
S-adenosyl-homocysteine
5-methylTHF
homocysteine
serine
CBS
B6
cystathionine
cysteine
sulphate

methionine by using 5-methyltetrahydrofolate (5methylTHF) as methyl donor. 5-methylTHF is provided by the enzyme methylenetetrahydrofolate reductase (MTHFR). In mammals, homocysteine is re-methylated into methionine by using 5-methylTHF as methyl donor in almost all tissues. However, in the liver, the remethylation of homocysteine can also be carried out by the enzyme betaine-homocysteine methytransferase (BHMT), which uses betaine as the methyl donor. Methionine can be used directly in protein synthesis or converted into S-adenosylmethionine (SAM or AdoMet). Increased levels of SAM due to high methionine favour the trans-sulfuration pathway by activating CBS as well as by inhibiting MS and MTHFR and hence inhibiting the re-methylation of homocysteine. SAM can also be converted back into homocysteine through subsequent reactions in the re-methylation pathway that are part of the methionine cycle. In addition to being an allosteric activator of the trans-sulfuration pathway and inhibitor in the re-methylation pathway, SAM is also an important methyl donor in many methylation reactions such as methylation of DNA, protein, phospholipids and neurotransmitters (Rosenblatt 1995 and Fowler 1997).

i) Cystathionine Synthase (CBS)

Cystathionine β Synthase (CBS), a pyridoxal 5'-phosphate (PLP; an active vitamin B6)-requiring enzyme, catalyses the first step of the trans-sulfuration pathway of homocysteine metabolism (Mudd et al. 1995). Homocysteine irreversibly condenses with serine forming cystathionine which is then broken down by cystathionase into cysteine and α-oxobutyrate. Cysteine is eventually broken down into inorganic sulfate by subsequent enzymatic reactions. SAM binding to CBS activates the enzyme and favors the trans-sulfuration pathway. Although remethylation and trans-sulfuration are both major pathways in the liver, not all tissues have the trans-sulfuration pathway (Finkelstein 1990).

ii) Methionine Synthase (MS)

Methionine Synthase (5-methyltetrahydrofolate:homocysteine methyltransferase) catalyses the remethylation of homocysteine to methionine by transferring a methyl group from 5-methylTHF to homocysteine (Fenton and Rosenberg 1995). MS activity is dependent upon its cofactor cobalamin (B12). In bacteria, the methyl group from 5-methylTHF converts the enzyme-bound cobalamin into methylcobalamin. The methyl group is then transferred to homocysteine. The enzyme-bound cobalamin is designated as cob(I)alamin because it has a cobalt atom with the valence of 1+. However, the enzyme-bound cob(I)alamin oxidizes spontaneously to cob(II)alamin or cob(III)alamin. For the enzyme to be active, the oxidized cobalamin has to be reduced for the generation of methylcobalamin.

1.2.3 Homocystinuria

Classical homocystinuria is an autosomal recessive disease characterized by dramatic elevation of homocysteine in the plasma (hyperhomocysteinemia) and urine (homocystinuria). Note that the term "homocystinuria" can be used to signify the disease and the biochemical condition. Patients usually have neurological and vascular complications. While homocystinuria due to disruption of the trans-sulfuration pathway is mainly a result of mutations in cystathionine β synthase (CBS), homocystinuria can also be caused by MTHFR and methionine synthase deficiencies which disrupt the remethylation pathway.

Clinical and biochemical features of CBS and methionine synthase deficiency are described below, while severe and mild MTHFR deficiencies will be discussed in sections 1.3.2 and 1.3.3 respectively.

i) Cystathionine Synthase (CBS) Deficiency

The common clinical features of CBS deficiency are dislocation of the optic lens, osteoporosis, thinning and lengthening of the long bones, mental retardation, and thromboembolisms of the arteries and veins (Mudd et al. 1995). Biochemically, CBS deficiency can be distinguished from MTHFR deficiency by means of methionine loading. The procedure involves oral intake of a standard dose of methionine, and homocysteine is measured after a fixed time interval. Patients with disrupted homocysteine remethylation usually have hyperhomocysteinemia during fasting, with a normal increase in homocysteine after methionine loading. Patients with disrupted homocysteine trans-sulfuration, however, usually have normal fasting homocysteine but have an abnormally high increase in homocysteine after a methionine load (Refsum et al. 1997). In CBS deficiency, both homocysteine and methionine levels are high, whereas in remethylation defects, the homocysteine level is high but the methionine level is low. The reason is that in CBS deficiency, the remethylation of homocysteine to methionine is normal; a high level of homocysteine leads to an increase of methionine. In remethylation defects, methionine synthesis is reduced.

ii) Methionine Synthase (MS) Deficiency

Methionine Synthase deficiency is classified as a cobalamin disorder due to the close association with deficiencies in methylcobalamin metabolism. Patients with methionine synthesis deficiency may have neurologic deterioration, but they also have hematologic abnormalities such as megaloblastic anemia. Patients with MTHFR deficiency, in contrast to those with deficiency in methionine synthesis due to abnormalities in methylcobalamin formation (cblC, cblD, cblE, cblF and cblG complementary groups), do not have megaloblastic anemia (Rosenblatt 1995)

The only known function of 5-methyl THF is to serve as a methyl donor for the remethylation of homocysteine to methionine catalyzed by methionine synthase. If there is a blockage in methionine synthesis due to methionine synthase deficiency, the 5-methylTHF is unable to donate its methyl group and for regeneration of THF. This is called the "folate trap hypothesis", where the THF is trapped in the form of 5-methylTHF making it unavailable for other biochemical reactions such as DNA synthesis (Rosenblatt 1995, Fenton and Rosenberg 1995). This problem results in megaloblastic anemia due to the inability of lymphoid cells to proliferate normally.

1.2.5 Elevated homocysteine as a risk factor of vascular disease

Thrombotic and arteriosclerotic complications were often found in patients with homocystinuria (Mudd et al 1995). It was therefore proposed that carriers (heterozygotes) of homocystinuria with slightly elevated homocysteine would be at an increased risk for vascular disease (McCully 1969, Wilcken and Wilcken 1976). Since the 1970's, there have been an increasing number of reports supporting the concept that elevated homocysteine is associated with vascular disease. Elevated homocysteine levels have been found in patients with cardiovascular, cerebrovascular and peripheral vascular disease (Clark et al 1991, Boers et al 1985). Homocysteine is one of the 200 identified risk factors for cardiovascular disease. Increased dietary intake of folate, B6 and B12 reduce homocysteine level, while age smoking and alcohol are some of the factors that increase homocysteine. In 1995, a meta-analysis of 27 studies on homocysteine and vascular disease showed that 10% of the risk for cardiovascular disease (CAD) could be accounted for by elevated homocysteine. The same analysis showed that a level of 15μM homocysteine increased the risk for CAD by 60% in men and 80% in women, an effect comparable to a cholesterol increase of as much as 0.5 mM. Folic acid supplementation is a promising and inexpensive way of reducing the risk of vascular disease. Clinical trials of vitamin supplementation are in progress.

Many mechanisms of how hyperhomocysteinemia causes vascular disease have been proposed. Homocysteine may induce vascular lesions in the following ways: a) direct toxicity to the endothelium, b) increase DNA synthesis in vascular smooth muscle cells, c) cause oxidation of low-density lipoprotein and d) decrease thrombomodulin cell surface expression and inhibition of protein C activation thus increasing thrombosis.

Homocysteine levels, although influenced by numerous non-genetic factors, are also determined by genetic factors. Any genetic disruptions of homocysteine metabolism will result in elevated homocysteine. Severe cystathionine β synthase (CBS) and methylenetetrahydrofolate reductase (MTHFR) deficiencies are the most common genetic defects in homocystinuria. However, heterozygotes for homocystinuria with slightly elevated homocysteine are extremely rare and can hardly be a significant cause of vascular disease, which is so common in the general population. The cloning and molecular analysis of MTHFR have brought new insights to the story of genetic causes of hyperhomocysteinemia and vascular disease.

1.3 Methylenetetrahydrofolate Reductase (MTHFR)

1.3.1 Biochemistry of MTHFR

Methylenetetrahydrofolate reductase (EC 1.5.1.20), a flavoprotein that uses NADPH as electron donor, catalyses the reduction of 5, 10-methylenetetrahydrofoate to 5-methyltetrahydrofolate (5-methyl THF) (Jencks and Matthews 1987). The only known function of 5-methyl THF is to serve as a methyl donor for the remethylation of homocysteine to methionine catalyzed by methionine synthase (Fig. 3). Homocysteine can be converted into S-adenosylmethionine (SAM) which is an allosteric inhibitor of MTHFR. Porcine liver MTHFR is the only mammalian MTHFR purified to homogeneity (Daubner and Metthews 1982). The *E. coli* MTHFR gene (metF) has been isolated and sequenced (Saint-Girons et al. 1983). The *E. coli* MTHFR protein has been crystallized and studied by x-ray crystallography (Guenther et al, in press). Porcine MTHFR is a homodimer of two 77 kDa subunits. The dimeric enzyme binds noncovalently to two FAD molecules per dimer (Daubner and Matthews 1982) and remains dimeric in the presence of SAM (Jencks and Matthews 1987). Recent in vitro studies have shown that FAD dissociation is accompanied by loss of enzyme activity, and increased folate reduces the rate of FAD dissociation and enzyme activity loss (Guenther et al, in press). Partial proteolytic digestion of porcine MTHFR with trypsin revealed that each subunit has two spatially distinct domains, a 40 kDa N-terminal domain and a 37 kDa C-terminal domain (Matthews et al. 1984). While the substrates (FAD, NADPH and 5-methylTHF) bind to the 40kDa N-terminal domain, the allosteric inhibitor (SAM) binds to a 3kDa region near one end of the 37 kDa C-terminal domain (Sumner et al. 1986). The SAM binding site was found in the N-terminal region of the 37 kDa C-terminal domain, after the alignment of porcine peptide sequences with the human sequence derived from the human cDNA (Goyette et al 1994). Several porcine internal peptides have been sequenced by Edman degradation. A total of 12 internal peptides and the most N-terminal peptide were sequenced (some sequences are unpublished data from Dr. R. Matthews, University of Michigan). The internal peptides have various lengths ranging from 8 to 30 amino acids, and all showed a high percentage of homology to the human sequence. The eukaryotic and prokaryotic MTHFR sequences have been aligned (Goyette et al. 1994, Yang et al. 1984, Guenther et al., in press) (Fig. 4). The bacterial MTHFR lacks the regulatory C-terminal domain although it has the same catalytic activity (Matthews et al. 1984). The bacterial enzyme has ~30% amino acid sequence identity to the N-terminal domain of the derived human sequence (from the cloned cDNA). The pig and mouse sequences are ~90% identical to the human sequence. The yeast (*Saccharomyces cerevisiae*) MTHFR has also been identified (Genbank #1709159) showing about 30% sequence identity to the human sequence.

1.3.2 Severe MTHFR deficiency

Severe MTHFR deficiency (0-20% residual activity), the most common inborn error of folate metabolism, results in hyperhomocysteinemia, homocystinuria, low levels of plasma methionine and decreased neurotransmitter level in the central nervous system

Figure 4. Conservation of MTHFR amino acid sequence across eukaryotes and prokaryotes

The MTHFR amino acid sequence for the human (hMTHFR), mouse (mMTHFR), pig (pMTHFR), *E. coli* (eMTHFR) and *Saccharomyces cerevisiae* (yMTHFR) are aligned. Sequences that are similar or identical to the human sequence are shown in bold. Identical sequence with the human sequence is shown in upper case. Identical sequences across all species are underlined. Porcine sequences are a cDNA reported in this thesis (P600) and the internal peptides that are the only other available sequences for this species. The conserved alanine residue for the common polymorphism 677C→T (Ala to Val) is indicated by a vertical arrow.

```
hMTHFR        MVNEARGNSSLNPCLEGSASSGSESSKDSSRCSTPGLDPERHERLREKMR
mMTHFR        MVNEARGsgSpsPrsEGS-SSGSESSKDSSRCSTPsLDPERHERLREKMR
eMTHFR        -------------------------------------msffHasqRdaln
yMTHFR        -----------------------------------------------msi
pMTHFR        MVNEARGNggpgPrcEGS-SSGSESSKESSRCSTtGLDPERHERLREKMk


hMTHFR        RRLES--GDKWFSLEFFPPRTAEGAVNLISRFDRMAAGGPLYIDVTWHPA
mMTHFR        RRmdS--GDKWFSLEFFPPRTAEGAVNLISRFDRMAAGGPLfvDVTWHPA
eMTHFR        qsLaevqGqinvSfEFFPPRTsEmeqtLwnsiDRlsslkPkfvsVTy--g
yMTHFR        RdLyharaspfiSLEFFPPkTelGtrNLmeRmhRMtAldPLfItVTW--g
pMTHFR        RRMES--GDKWFSLEFFPPRTAQGAVNLIS :(P600)

pMTHFR          (Pi-3): KXFSLEFFPPRTAEXAVNLISsFDRMgAGGP


hMTHFR        GDP-GSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEITGHLHKAKQ
mMTHFR        GDP-GSDKETSSMMIASTAVNYCGLETILdMTCCqQRpEEITGHLHrAKQ
eMTHFR        ansgerDrthSiikgikdrt---GLEaapHlTCidatpdElrtiardywn
yMTHFR        a--gGttaEktlt-lASlAqqtlnipvcmHlTCtntekaiIddaLdrcyn


hMTHFR        LGLKNIMALRGD-PIGDQWEEEEGG---FNYAVDLVKHIRSEFGDYFDIC
mMTHFR        LGLKNIMALRGD-PvGDhWEaEEGG---FsYAtDLVKHIRtEFaDYFDIC
eMTHFR        nGirhIvALRGDlPpgsgkp--------emYAsDLVtllkeva--dFDIs
yMTHFR        aGirNIlALRGDpPIGedWldsqsnespFkYAVDLVryIkqsyGDkFcvg


                                                  ↓
hMTHFR        VAGYPKGHPEAGSFEADLKH------LKEKVSAGADFIITQLFFEADTFF
mMTHFR        VAGYPrGHPdAeSFEdDLKH------LKEKVSAGADFIITQLFFEAsTFF
eMTHFR        VAaYPevHPEAkSaqADLln------LKrKVdAGAnraITQLFFdvesyl
yMTHFR        VAAYPeGHcEgeaegheqdplkdlvyLKEKVeAGADFvITQLFydvekFl

pMTHFR                             (Pi-2):   KVaAGADFITgQLFFEADTF


hMTHFR        RFVKACTDMGITC-PIVPGIFPIQGYHSLRQLVKLSKLEVPQEIKDVIEP
mMTHFR        sFVKACTeiGIsC-PIlPGIFPIQGYtSLRQLVKLSKLEVPQkIKDVIEP
eMTHFR        rFrdrcvsaGIdv-eItPGIlPvsnfkqakkfadmtnvriPawmaqmfdg
yMTHFR        tFemlfrerisqdlPlfPGlmPInsYllfhraaKLShasiPpaIlsrfpP


hMTHFR        IKDNDAAIRNYGIELAVSLCQELLASGLVPG---LHFYTLNREMATTEVL
mMTHFR        IKDNDAAIRNYGIELAVxLCrELLdSGLVPG---LHFYTLNREvATmEVL
eMTHFR        ldDDaetrklvGaniAmdmvkiLsreG-Vkd---fHFYTLNRaemsyaic
yMTHFR        eiqsDdnavksigvdilieliqeiyqrtsgrikgfHFYTLNlEkAiaqiv
```

```
hMTHFR    KRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYRTQEWDE
mMTHFR    KqLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYRTQdWDE
eMTHFR    htLGvrpgl-----------------------------------------
yMTHFR    sqspvlshivnesseeegedetsgeigsienvpiedadgdivlddsneet

pMTHFR    KaLGlWiEDPRRPLPWA :(Pi-1)
pMTHFR                (Pi-9):   KgREiDVprIFWASRPK
pMTHFR                                (Pi-8):   KXYIYXTQXXXX


hMTHFR    FPNGRWGNSSSPAFGELKDYYLFYLKSKSPKEELLKMWGEELTSEASVFE
mMTHFR    FPNGRWGNSSSPAFGELKDYYLFYLKSKSPrEELLKMWGEELTSEeSVFE
eMTHFR    --------------------------------------------------
yMTHFR    vaNrkrrrhSSldsakLifnraivtekglrynnengsmpskkalisiskg

pMTHFR    FPNGRXGaS :(Pi-8)          (Pi-4): KMWGqELTSEeSVFE


hMTHFR    VFVLYLSGEPNRNGHKVTCLPWNDEPLAAETSLLKEELLRVNRQGILTIN
mMTHFR    VFehYLSGEPNRhGyrVTCLPWNDEPLAAETSLmKEELLRVNRlGILTIN
eMTHFR    --------------------------------------------------
yMTHFR    hgtlgrdatwdefpngrfgdsrspaygeidgygpsikvskskalelwgIp

pMTHFR    VFahYXSGEPNq :(Pi-4)      (Pi-5): KEELLXVNRrGILTIN


hMTHFR    SQPNINGKPSSDPIVGWGPSGGYVFQKAYLEFFTSRETAEALLQVLKKYE
mMTHFR    SQPNINaKPSSDPvVGWGPSGGYVFQKAYLEFFTSRETvEALLQVLKtYE
eMTHFR    --------------------------------------------------
yMTHFR    ktigdlkdifikyleGstdaipwsdlglsaEtaliqEeliqLnyrgyltl

pMTHFR    SQPNInGK   :(Pi-5)                      (Pi-7): KXYE


hMTHFR    LRVNYHLVNVKGENITNAPELQPNAVTWGIFPGREIIQPTVVDPVSFMFW
mMTHFR    LRVNYHiVdVKGENITNAPELQPNAVTWGIFPGREIIQPTVVDPiSFMFW
eMTHFR    --------------------------------------------------
yMTHFR    asqpatnatlssdkIfgwgpakgrlyqkafvemfihrQqwettlkpkldh

pMTHFR    LRVNYaiVdk


hMTHFR    KDEAFALWIERWGKLYEEESPSRTIIQYIHDNYFLVNLVDNDFPLDNCLW
mMTHFR    KDEAFALWIEqWGKLYEEESPSmTIIQYIHDNYFLVNLVDNeFPLDsCLW
eMTHFR    --------------------------------------------------
yMTHFR    ygrrkfsyyagdssgsfEtnldphsssvvtwgvFpnspVkqttiieeesf

pMTHFR         (Pi-6): KLYEEESPSRmlIQYIHDNYFLVNLVDNenP
```

```
hMTHFR    QVVEDTLELLNRPTQNARETEAP---------------------------
mMTHFR    QVVEDTfELLNRh-pteRETqAP---------------------------
eMTHFR    --------------------------------------------------
yMTHFR    kawrDeafsiwsewaklfprntPanillrlvhkdyclvsivhhdfketde

hMTHFR    ---------
mMTHFR    ---------
eMTHFR    ---------
yMTHFR    lwemlldqa
```

(Rosenblatt 1995). Clinical severity correlates with the degree of enzyme deficiency, and the age of onset ranges from neonatal to adolescence. The most common clinical symptoms include developmental delay, motor and gait abnormalities, seizures, psychiatric manifestations, and neurological and vascular complications (Rosenblatt 1995). The unique neurological symptoms observed in severe MTHFR deficiency, such as demyelination of neurons, are absent in CBS deficiency. These unique symptoms in severe MTHFR deficiency are believed to be caused by the reduced production of SAM and hence the decrease in methylation reactions. Severe MTHFR deficiency is rare, with approximately 50 cases worldwide. Since the cloning of the MTHFR cDNA, 18 severe mutations have been reported in patients with homocystinuria (Goyette et al. 1994, 1995, 1996 and Kluijtmans et al. 1998).

1.3.3 Mild MTHFR Deficiency

A milder form of MTHFR deficiency (35%-50% residual activity), characterized by a thermolabile enzyme (Kang et al. 1988), has been recognized as a genetic risk factor for mild hyperhomocysteinemia. The thermolabile enzyme has <35% residual activity after heat inactivation at 46°C for 5 min (Frosst et al. 1995). This variant was first identified in 17% of North American patients with coronary artery disease and in 5% of controls (Kang et al. 1991). In the same study, the thermolabile variant was found to be inherited as an autosomal recessive trait (Kang et al. 1991). The cloning of MTHFR cDNA allowed the study of MTHFR at the molecular level. The thermolability of the enzyme was found to be caused by a point mutation (677C→T) in MTHFR changing an alanine to a valine codon (Frosst et al. 1995). This mutation was found to be a common polymorphism with an allele frequency of about 35 % in the North American population (Frosst et al. 1995, Jacques et al. 1996). The MTHFR variant has been suggested as an inherited risk factor for vascular disease independent of other risk factors such as age, smoking, hypercholesterolemia and hypertension (Kang et al. 1993).

Homozygosity and heterozygosity for the 677C→T mutation is associated with reduced specific activity and increased thermolability at 46°C in lymphocyte extracts (Frosst et al. 1995, van der Putt et al. 1995, Kluijtmans et al. 1996). Individuals who are homozygous for the mutation have significantly elevated plasma homocysteine when their plasma folate is below the median value (15nM) (Jacques et al. 1996). The mutation was not associated with elevated homocysteine when plasma folate levels were above the median value.

Since the alanine residue at nucleotide position 677 is conserved across species, the *E. coli* MTHFR was also mutagenized to create the alanine to valine codon and used as a model (Guenther et al., in press). As mentioned above, in vitro studies with *E. coli* MTHFR showed that FAD dissociation is accompanied with enzyme activity loss, and increased folate reduces the rate of FAD dissociation and enzyme activity loss. The enzyme activity loss due to FAD dissociation was also observed in the mutagenized *E. coli* MTHFR. Increased folate concentration had a more profound protective effect on FAD dissociation and enzyme activity in the mutant enzyme compared to the wild type enzyme (Guenther et al., in press), suggesting that the alanine to valine change affects FAD binding. However, x-ray crystallography of the *E. coli* MTHFR showed that the alanine residue does not reside in the FAD binding site. It has been suggested that the change of the alanine residue changes the conformation of the FAD binding site and hence affects

FAD binding and reduces enzyme activity (Guenther et al., in press). Studies with the human enzyme have also shown that folate and FAD protects the normal and mutant MTHFRs from destabilization. Therefore, increased folate intake has been suggested as a compensation for individuals with mild MTHFR deficiency to maintain the homocysteine level in the lower range.

Recently, a second common polymorphism (1298A→C), changing a glutamate to an alanine residue, was reported (van der Put et al. 1998 and Weisberg et al. 1998). Homozygosity for the mutation was found in approximately 10% of Canadian individuals and is associated with decreased specific activity (60% of control). The significance of this polymorphism in other populations and in folate-dependent multifactorial diseases remains to be determined.

1.3.4 Thermolabile MTHFR as a genetic risk factor for vascular disease

Hyperhomocysteinemia has been recognized as an independent risk factor for vascular disease (Boushey et al 1995). The common 677C→T polymorphism in MTHFR, encoding the thermolabile MTHFR variant, has been shown to associate with elevated homocysteine (e.g. Jacques et al. 1996, Kluijtmans et al 1996). The most interesting and important question is whether this common polymorphism does indeed correlate with vascular disease. Since the cloning of MTHFR and the identification of the polymorphism, more than a hundred studies have been done world wide in trying to verify the correlation between this single base pair change in MTHFR with elevated homocysteine and vascular disease. However, the results may appear contradicting when interpreted without care.

Almost all studies that try to correlate the MTHFR polymorphism with homocysteine levels were able to find a significant correlation. The 677C→T polymorphism, in the homozygous state, does correlate with elevated plasma homocysteine level (e.g. Morita et al 1997, Jacques et al 1996, Kluijtmans et al 1996, and Harmon et al 1996). The situation for the studies on the correlation between the polymorphism and vascular disease is, on the contrary, more complicated. Although some studies report a significant correlation between the MTHFR polymorphism with vascular disease and/or myocardial infarction (e.g. Morita et al 1997, Kluijtmans et al 1996, Gallagher et al 1996, Arruda et al 1997 and Izumi et al 1995), there are also studies that report the absence of such correlation (Girelli et al 1998, Markus et al 1998, Deloughery et al 1996, Ma et al 1996, Schmitz et al 1996) or a weak correlation (Christensen et al 1997). To interpret these results correctly, one has to keep in mind how MTHFR influences the metabolism of homocysteine.

First of all, thermolabile MTHFR can only predispose to vascular disease when there is mild hyperhomocysteinemia. Since it is mild hyperhomocysteinemia that is the risk factor for vascular disease, studies on the correlation between MTHFR polymorphism and vascular disease without considering homocysteine levels could be misleading (e.g. Arruda et al. 1997, Adams et al. 1996, Wilcken et al. 1996). Homocysteine levels are known to be influenced by folate intake. Thermolabile MTHFR is correlated with elevated homocysteine levels particularly when plasma folate is below the median level (e.g. Jacques et al. 1996, Kluijtmans et al. 1996, Harmon et al. 1996). Studies that do not include folate status may weaken the effect of thermolabile MTHFR on elevated homocysteine, if there is high folate level due to proper diet or vitamin use. Another concern of these association studies is the elimination of known risk factors. Since there are numerous environmental risk factors that contribute to the development of vascular disease, patients with significant correlations with other known risk factors may influence the results.

There is no doubt that thermolabile MTHFR is correlated with elevated plasma homocysteine, particularly when folate is low. Whether the thermolabile MTHFR does cause vascular disease is still under investigation. Perhaps the study of a knockout mouse model will allow us to determine if the absence of MTHFR leads to vascular complications in these animals and if increased folate intake would compensate for the genetic defect. Furthermore, there could be other common polymorphisms in MTHFR or other enzymes involved in homocysteine metabolism such as methionine synthase (MS) that may predispose to elevated homocysteine and vascular disease.

1.3.5 Thermolabile MTHFR and other multifactorial diseases

Several studies have reported an association between homozygosity for the 677C→T mutation in MTHFR and an increased risk of NTDs (Whitehead et al. 1995, van der Put et al. 1995, 1996, Ou et al. 1996, Kirke et al. 1996). Although the underlined mechanism is still unclear, hyperhomcysteinemia has been correlated with NTDs. The homozygous 677C→T mutation of MTHFR has also been correlated with reduced risk in colon cancer (Ma et al. 1997). A possible mechanism is that reduced MTHFR activity ensures that more methylenetetrahydrofolate is available for thymidylate synthesis and hence reduce uracil misincorporation. MTHFR deficiency may also reduce the level of SAM and DNA methylation and hence ensure the activation of tumor suppresser genes.

1.4 Molecular biology of MTHFR

1.4.1 Cloning of human MTHFR cDNA

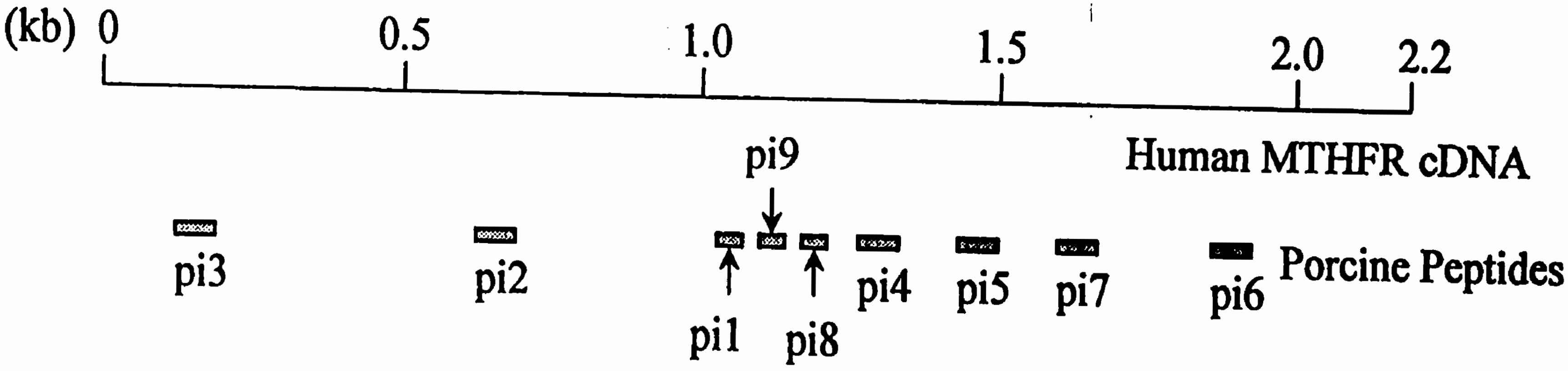
By using the 30 amino acid porcine internal peptide sequence, degenerate oligonucleotides were designed to amplify a 90 bp porcine cDNA by RT-PCR from pig liver RNA (Goyette et al. 1994). A non-degenerate oligonucleotide was then designed from the 90 bp porcine cDNA and was used to screen a human cDNA library by PCR. The first human MTHFR cDNA clone was 1.3 kb. A 2.2 kb cDNA was isolated by using the 1.3 cDNA as a probe (Goyette et al. 1994). Conserved amino acid sequences have been identified in the deduced amino acid sequence of the human cDNA when aligned with the amino acid sequences of several porcine internal peptides (Fig. 4 and 5), the metF (MTHFR) gene in *E. coli* (Goyette et al. 1994) and the yeast MTHFR in *S. Cerevisiae* (Fig. 5). An ATG (met) codon at base pair position +13 of the 2.2 kb cDNA (numbered according to the published cDNA) was used as a translation start site in an *E. coli* expression system, giving rise to a 70 kDa protein with MTHFR enzyme activity (Frosst et al. 1995).

1.4.2 Human and mouse MTHFR genes

The human gene has been localized to chromosome 1p36.3-36.2 (Goyette et al. 1994) and the mouse gene to distal Chromosome 4 (Frosst et al. 1996). By using the human 2.2 kb cDNA, both the human and mouse MTHFR genes were cloned and characterized (Goyette et al. 1998). The characterized region of the human gene contains 11 exons and is approximately 17 kb in length. The human and mouse genes are very

Figure 5. **Alignment of all the available MTHFR porcine peptide sequences on the human 2.2kb cDNA.**

Nine porcine peptide sequences have been aligned with the human cDNA. The amino acid sequence homology between the human and porcine MTHFR is about 90%. Filled boxes represent porcine peptide sequences. The base pair position is given according to the published human 2.2kb cDNA. The equivalent human sequence of three internal peptides (Pi10, Pi11, Pi12) as well as the N-terminal peptide (PN) of the porcine MTHFR have not been identified. (PN = porcine N-terminal peptide. Pi = porcine internal peptide).



Location of porcine peptides on human cDNA

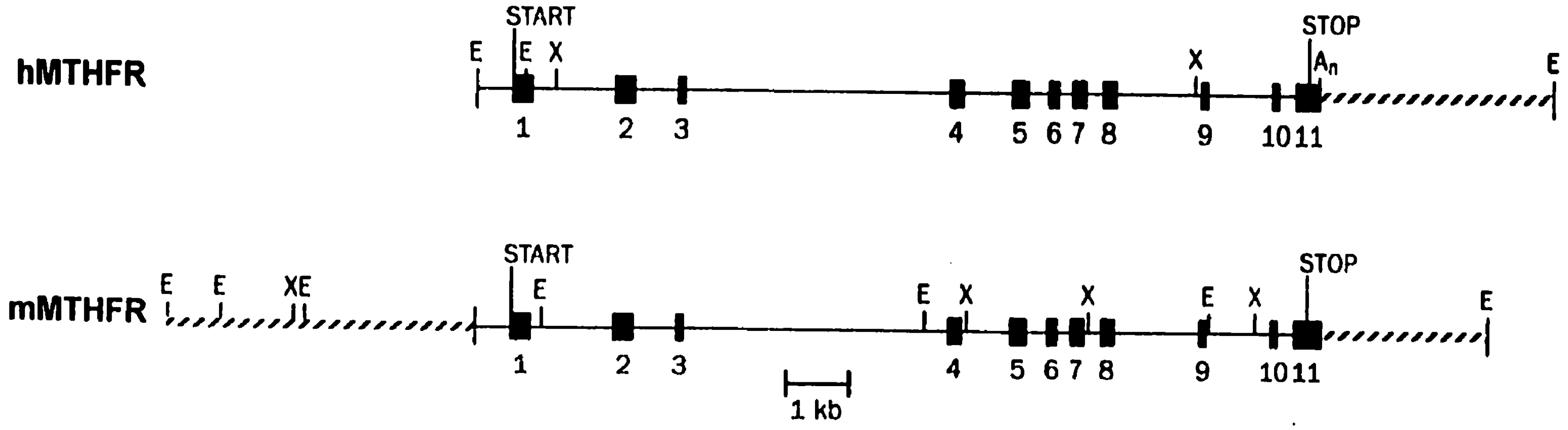
pi1(17 a.a) : bp 1012-1061
pi2(20 a.a) : bp 661-720
pi3(31 a.a) : bp 190-276
pi4(27 a.a) : bp 1267-1374
pi5(24 a.a) : bp 1411-1485
pi6(31 a.a) : bp 1801-1890
pi7(14 a.a) : bp 1599- 1641

pi8(21 a.a) : bp 1125-1188
pi9(17 a.a) : bp 1077- 1128
pi10(8 a.a) : location unknown
pi11(19 a.a): location unknown
pi12(8 a.a): location unknown
pn(10 a.a) : at n-terminal

Figure 6. The gene structure of human and mouse MTHFR

The characterized region of the human and mouse MTHFR gene reported in Goyette et al (1998) have identical exon/intron organization. There are 11 exons in both the human and mouse gene, spanning approximately 17kb. The cloned 2.2kb human cDNA was expressed using the ATG (Start) codon in exon 1 (as indicated), giving a catalytically active protein of 70kDa. The hatched boxes and the 5' end of human MTHFR were not characterized in Goyette et al (1998). MTHFR=human MTHFR. mMTHFR=mouse MTHFR. STOP=Stop codon, An=polyadenylation signal. E=EcoRI. X=XbaI.

[Reproduced from: Goyette et al (1998) Gene structure of human and mouse methylenetetrahydrofolate reductase (MTHFR). *Mammalian Genome* **9**: 652-656]


hMTHFR
mMTHFR
START
STOP
E
X
XE
A_n
1
2
3
4
5
6
7
8
9
10
11
1 kb

similar in size and exon/intron boundaries (Fig. 6). The human and mouse amino acid sequences are about 90% identical (Goyette et al. 1998). Both the human and the mouse cDNA have multiple 5' ends. Northern analysis revealed that both the human and the mouse MTHFR transcripts are large, approximately 7-9 kb. Since the cloned coding sequence is only 2.2 kb, it suggests that the transcript has large UTR(s). The transcription start sites and the promoters have not been identified in either the human or the mouse gene. The 5' regions of human and mouse MTHFR are therefore under investigation, as part of this thesis.

1.4.3 Missing coding exon(s) in human MTHFR

Based on three observations, we believe that the cloned 2.2 kb human cDNA is incomplete and it is still missing a coding exon(s) at the 5' end of the cDNA.

i) Presence of Isozymes

The purified porcine liver enzyme has been shown to be 77 kDa (Daubner and Matthews 1982). Western analysis using porcine MTHFR antibody also revealed that most human tissues have a 77 kDa protein (Frosst et al. 1995). When the 2.2 kb human cDNA was expressed in *E. coli*, the protein obtained was only 70 kDa, suggesting that the cDNA is incomplete. The existence of both 77kDa and 70 kDa proteins in human fetal liver and other tissues on the Western blots clearly suggests the presence of isozymes. Since MTHFR is a cytosolic protein, post-translation modification such as glycosylation is unlikely and has not been reported.

ii) Missing Porcine N-terminal sequence

The human and porcine MTHFR amino acid sequences are highly homologous (~90% identity). Various porcine peptides have been aligned to the amino acid sequence predicted from the human cDNA (Fig. 4 and 5). However, the N-terminal amino acid sequence of the porcine peptide has not been identified in the predicted amino acid sequence of our human cDNA. It therefore suggests that the missing coding cDNA sequence is at the 5' end.

iii) Multiple cDNA 5' ends

Various cDNA species with different 5' exons upstream of the ATG (met) start codon have been isolated, suggesting the presence of 5' alternative splicing. The transcription start site of the human MTHFR cDNA has not been defined. Although most of these 5' exons are believed to be 5' UTR, it certainly raises the possibility that a coding exon may be alternatively spliced into the 5' end of the 2.2 kb cDNA and translated into the 77kDa isoform.

1.4.4 The significance of characterizing the 5' region of a gene

In the next couple of years, the entire human genome will be completely sequenced according to the Human Genome Project. It is becoming more popular to clone genes by analyzing the genomic sequences with the help of computer technologies and the Genbank databases. However, whether these "genes" are indeed coding for a functional protein remains to be determined experimentally. Successful cloning of a gene begins with the isolation a cDNA, since a cDNA is made from a messenger RNA that is expressed in the cell and is potentially coding for a functional protein. Even with the advancement of biotechnology, obtaining a cDNA is still dependent upon the enzyme reverse transcriptase. Reverse transcriptase has some limitations regardless of the modification made by biotechnology companies. The most challenging task for the generation of a full-length cDNA has been the presence of strong secondary structure of RNA and CG rich regions near the 5' end of the mRNA. These sequence-related obstacles often lead to pre-mature termination of reverse transcription before the 5' end of the mRNA is reached. The longer the mRNA and the stronger the secondary structure, the harder it is to obtain a full-length cDNA. It is therefore common to have most part of the cDNA characterized, but the 5' ends of the cDNA and the promoter of the gene are left uncharacterized.

The presence of multiple splice variants at the 5' untranslated region (UTR) is a common phenomenon in the human genome. Gene expression can be regulated both at the transcriptional level and the translational level. Alternate use of transcription start sites and multiple 5' exons are also common in the human genome as part of transcriptional regulation. The alternate use of 5' UTRs provides options for the mRNA to form various secondary structures hence provide fine regulation at the translational level.

RNA polymerase binds to a promoter and starts transcribing at the transcription start site. A promoter with a TATA box is usually located within 100bp of the transcription start site. A promoter in a mammalian gene can be a well-defined TATA box or simply a CG rich region without any specific consensus sequences. Housekeeping genes often have CpG islands near the promoter regions that may be involved in DNA methylation-related gene regulation. To define a promoter, transfection studies using a reporter gene construct to test for promoter activity are required.

RNA polymerase alone is not enough to drive transcription. The transcription of mammalian genes requires the binding of proper transcription factors to the transcription factor binding site. The 5' region of a mammalian gene, therefore, has a number of cis-elements that are involved in the binding of transcription factors. Studying the cis-elements of a gene reveals valuable information on how a gene is regulated and what factors are involved.

Many complex biological phenomena could occur at the 5' region of every gene in the human genome. To fully understand the regulation at both the translational and transcriptional level of a gene requires the characterization of the 5' region. In this thesis, I report the characterization of a 12kb region in the 5' end of the human MTHFR gene.

2. **Research Proposal**

A 2.2 kb cDNA and gene of human methylenetetrahydrofolate reductase (MTHFR) have been cloned. Although many analyses have been performed on the cloned region of MTHFR, the promoter and the transcription start site of this gene remain unidentified. Current experimental data indicate that the MTHFR cDNA has multiple cDNA 5' ends and missing 5' exon(s). The complete characterization of the 5' ends of the cDNAs and of the gene is essential in understanding the putative splicing events in the 5' region as well as in identifing the missing exon(s), the transcription start site and the promoter. The isolation of these sequences is critical for studying the regulation of the MTHFR gene.

The characterization of the 5' region of the MTHFR gene will be carried out by isolating more cDNA clones and genomic fragments. cDNA clones will be obtained from cDNA libraries by library screening techniques and by various PCR methods (such as RACE-PCR and RT-PCR). Genomic fragments will be obtained from genomic libraries and PAC (P1 artificial chromosome) clones using library screening and Southern analysis techniques. Once new cDNAs and genomic fragments are obtained, they will be sequenced and analyzed for the presence of open reading frames, sequences homologous to the porcine N-terminal peptide and mouse genomic sequences, consensus sequence for splice sites, transcription and translation start sites as well as promoter elements. DNA sequences will also be compared to Genbank sequences such as the EST (Expressed Sequence Taq) database in an attempt to search for expressed sequences. These data will be used to generate a physical map of the genomic region containing all the exonic sequences identified 5' of the existing 2.2 kb cDNA.

3. Methods and materials

3.1. Libraries and PAC clone

Four human genomic λ phage libraries were obtained from ATCC: 1) Chromosome 1 specific-HindIII complete digest (ATCC# 57754), 2) Genomic-EcoRI partial digest (ATCC# 37385), 3) Genomic-AluI or HaeIII partial digest (ATCC# 57760) and 4) Genomic-MboI partial digest (ATCC# 37458). The human PAC (P1 Artificial Chromosome) clone was obtained from J. Rommens, University of Toronto.

3.2. Frozen tissues and cell cultures

Frozen human fetal lung and fetal kidney tissues were obtained from Dr. C. Goodyer (Montreal Children's Hospital). Frozen pig liver tissue was obtained from Dr. T. Perrault (Montreal Children's Hospital). Colon carcinoma cell lines SW620 and SW1222 were obtained from ATCC (#CCL 227) and Dr. Beauchemin (Montreal General Hospital) respectively.

3.3. Western Blotting

Cell pellets from frozen tissues were obtained by crushing the tissues in liquid nitrogen. Cell pellets for the colon carcinoma cell line were obtained from tissue culture dishes by washing the cells with 1X PBS followed by treating the cells with trypsin. The cells were then removed from the culture dish and centrifuged to obtain a pellet. The cell pellets were then homogenized in 0.25M sucrose with 2μg/ml aprotinin and leupeptin (0.1g tissue/ml of homogenization buffer) by sonication on ice (3 X 15s). Homogenized samples were centrifuged at 14,000rpm for 30 minutes at 4°C. 100μg of crude protein extract was denatured by adding DTT containing loading buffer and boiling and was then loaded on 10% SDS-polyacrylamide gel. Protein was then transferred onto a nitrocellulose membrane (Hybond ECL, Amersham). The membrane was blocked in blocking solution (1X PBS, 0.5% Tween-20 and 5% powdered milk) overnight. The membrane was then incubated with a rabbit polyclonal anti-porcine MTHFR antibody containing blocking solution for 1 hour. The membrane was then washed with 1X PBS with 0.5% Tween-20 followed by incubation of a second antibody (anti-rabbit Ig horseradish peroxidase linked antibody) (Amersham). The membrane was then washed and a light reaction was carried out (ECL kit, Amersham). The membrane was then exposed to film (X-OMAT, Kodak).

3.4. Library screening

Genomic libraries were screened using standard DNA hybridization method (Sambrook et al 1989). The λ phage library was serially diluted in SM buffer. Diluted samples of the λ phage library were mixed with melted soft agarose and bacterial host and then spread on an agar plate. 80% confluence (~30,000 pfu) was required to avoid overlapping plaques (which might allow DNA recombination) but to allow efficient use of the plate. Duplicate nitrocellulose membranes were lifted from the plate. Membranes were treated with DNA denaturing solution (1.5M NaCl, 0.5M NaOH for 1.5 minutes) followed by neutralization solution (1.5M NaCl, 0.5 Tris-HCl, pH 8.0 for 5 minutes). The membranes were rinsed in 3X SSC. The membrane was baked at 80°C for 1 hour. The membrane was pre-hybridized in the pre-hybridization solution (5X SSC, 5X Denhardt's solution, 0.1% SDS and 100μg/ml denatured salmon sperm DNA) for 2-4 hours followed by an overnight hybridization in a fresh pre-hybridization solution containing a radioactive probe. The membrane was washed to remove non-specific binding (2X SSC and 0.05% SDS for 1 hour at room temp., then 1X SSC and 0.1% SDS at 68°C). If the background was high, the membranes were washed with 0.2X SSC and 0.1%SDS at 68°C. The membranes were wrapped in Saran Wrap and exposed to film (X-OMAT, Kodak).

3.5. Southern Blotting

DNA samples were run on 1% agarose gels. After electrophoresis, the agarose gel was soaked in 0.25M HCl for about 30 minutes. The gel was then treated with denaturation buffer (1.5M NaCl and 0.5M NaOH) for 30 minutes. The gel was then placed in a neutralization buffer (1.5M NaCl, 0.5M Tris-HCl pH7.2 and 0.001M EDTA) for 15 minutes. The DNA was then transferred to a nitrocellulose membrane (Hybond-N, Amersham). The DNA was fixed to the membrane by baking at 80°C for 1-2 hours. The membrane was pre-hybridized and hybridized in the hybridization solution (5X SSC, 5X Denhardt's solution, 0.5% SDS and 200μg/ml denatured salmon sperm DNA). The membrane was pre-hybridized at 65°C for 1hour and hybridized with a radioactive labeled probe overnight at 65°C. After incubation, the membrane was washed with first 2X SSC, 0.1% SDS at room temperature for 10 minutes twice, then with 1X SSC and 0.1% SDS at 65°C for 15 minutes, and finally with 0.1X SSC and 0.1% SDS at 65°C for

10 minutes. The membrane was then wrapped in SaranWrap and exposed to film (X-OMAT, Kodak).

3.6. DNA Probes

Probes were obtained from cDNA or genomic clones by restriction digesteion or PCR. The 800 bp MRE-5' cDNA extension was released from the Bluescript plasmid cDNA clone MRE-5' by EcoRI. The "300 bp" 43S cDNA extension was released from the Bluescript plasmid cDNA clone 43S by BamHI and NcoI. The 80 bp M3 cDNA extension was generated by PCR using one primer on the plasmid arm and another within the M3 cDNA extension. The 300bp genomic DNA extension of the genomic fragment MRM-5' was released from the Bluescript plasmid by BamHI and EcoRI. A 750bp or 1kb MTHFR cDNA probe was obtained by digesting the Bluescript cDNA clone—MTHFR_all with EcoRI and MseI.

All DNA probes were labeled with $\alpha[^{32}P]$dCTP by random priming according to the manufacturer's protocol (Multiprimed Kit, Amersham). After labeling, the DNA probe was purified in a Sephadex G-50 column (Pharmacia Biotech). 1-2µl of the probe was sampled for scintillation counting. Generally, a minimum of 10 million cpm was used in each hybridization.

3.7 PCR

PCR was carried out using Taq DNA polymerase (GibcoBRL) in a Perkin Elmer DNA Thermal Cycler. In a 50µl PCR reaction, DNA or cDNA samples were added to a mixture of 5µl 10X PCR buffer [200mM Tris-HCl (pH8.4), 500mM KCl], 1.5µl of 50mM $MgCl_2$, 2.5µl of 4mM dNTPs, 250-500ng of each primer and 1.25U of Taq DNA polymerase. A standard PCR reaction had 35 cycles and each cycle included a denaturation step of 1 minute at 94°C, an annealing step of various temperature (usually between 55°C-68°C), and an extension step of 72°C.

3.8. RNA isolation

Pellets of frozen tissues and carcinoma cell lines were obtained as described above in section 3.3. Total RNA was isolated according to the protocol recommended by the manufacturer of the TRIzol Reagent (GibcoBRL). Cell pellet (100mg) was homogenized in RNA isolation reagent (1 ml) using a power homogenizer (Polytron). Homogenate was incubated at room temperature for 5 minutes followed by chloroform extraction (0.2ml per 1ml of reagent). The sample was then centrifuged at 11,000 rpm for 15minutes at 4°C. The upper aqueous phase was transferred to a fresh tube to which 0.5 ml of isopropanol per 1 ml of reagent was added. After an incubation of 10 minutes at room temperature, the RNA was pelleted by centrifugation at 11,000 rpm at 4°C for 10 minutes. The RNA pellet was washed with 70% ethanol (Rnase free). The RNA pellet was then re-dissolved in Rnase free water. Poly A^+ mRNA was purified from total RNA using the Oligotex mRNA Kit (QIAGEN).

3.9. Northern blotting

Northern blotting was carried out according to standard protocols (Sambrook 1989). RNA was run on 1.5% agarose gel (containing 50mM boric acid, 5mM sodium borate, 10mM sodium sulphate, 1mM EDTA and 37% formaldehyde). The RNA was then transferred to a nitrocellulose membrane (Zetabind). The membrane was pre-hybridized in hybridization buffer [0.5M phosphate buffer (pH7.0) , 1mM EDTA, 7% SDS, 1%BSA and 100μg/ml salmon sperm DNA] at 65°C for 1-2 hours and then hybridized overnight with a radioactive probe at 65°C. A 750bp EcoRI-MseI fragment of the MTHFR cDNA was used as the probe. After hybridization, the membrane was washed twice with a buffer containing 1% SDS, 0.5% BSA, 40mM phosphate buffer (pH 7.0) and 1mM EDTA at 65°C for 20 minutes. The membrane was then washed twice with another buffer containing 1% SDS, 40mM phosphate buffer (pH 7.0) and 1mM EDTA at 65C°C for 20 minutes. Finally, the membrane was washed with 40mM phosphate buffer (pH 7.0). The membrane was then wrapped in SaranWrap and exposed to film (X-OMAT, Kodak).

3.10. RT-PCR

RT-PCR was carried out according to the protocol recommended by the manufacturer of SuperScript II Reverse Transcriptase (GibcoBRL). Total RNA (5-8μg) and random hexamer (0.1μg, Boehringer-Mannheim) were heated at 70°C for 10 minutes and then chilled on ice immediately. The denatured RNA was then added to a pre-mixed and pre-warmed (42°C) mixture of 5X First strand buffer [250mM Tris-HCl (pH8.3), 375 mM KCl, 15 mM $MgCl_2$] (4μl), 0.1M DTT (2μl), 4mM dNTPs (2.5μl), RNase inhibitor

(20u, Promega) and reverse transcriptase (200u) in a total volume of 20 μl. The reverse transcription reaction mix was incubated at 42°C for 1 hour. The reaction was inactivated by heating at 70°C for 15 minutes. 3-5μl of the first strand cDNA mix was used in 50μl of PCR.

3.11. 5' RACE-PCR

5' Rapid Amplification of cDNA Ends (RACE)-PCR was carried out according to the protocol of the Marathon cDNA Amplification Kit (Clontech). 1μg of poly A^{+} mRNA in a minimum of 0.25μg/μl was required for the protocol. The mRNA was first reverse transcribed by MMLV reverse transcriptase generating the first strand cDNA. The second strand cDNA was synthesized using a mixture of *E. coli* DNA polymerase I, *E. coli* DNA ligase and *E.coli* Rnase H. A linker (or cDNA adaptor) was then ligated to the ends of the double stranded cDNA using T4 DNA ligase. The adaptor ligated cDNA was then subjected to PCR using a gene specific primer and a primer that binds to the cDNA adaptor. The PCR amplification was carried out according to the protocol recommended by the manufacturer. If a distinct band was not observed, a nested PCR was carried out using a nested adaptor primer provided by the kit and another gene specific primer closer to the 5' end in the case of 5' RACE.

3.12. Inverse-PCR

Inverse-PCR was carried out on cDNA to amplify the cDNA 5' end. This protocol contained various steps in which the first strand and the complementary strand were synthesized, followed by template digestion and ligation which generated a circular cDNA molecule. Total RNA (25μg) in 25μl of DEPC water was denatured at 65°C for 5 minutes and then chilled on ice. The denatured RNA was added to a mixture of Rnase inhibitor (80u, Promega), gene specific primer (0.5 μg), 2X RT buffer [200mM Tris-HCl (pH8.0), 280 mM KCl, 20mM $MgCl_2$, 2mM dNTPs, 10mM DTT] (25μl) and AMV reverse transcriptase (3-5u, GibcoBRL) and then incubated at 42°C for 1 hour. After incubation, water (50μl) and tRNA (25μg) were added to the reaction mix which was then extracted by chloroform. The cDNA was precipitated by adding 5M NaCl (5μl) and cold 99% ethanol (250μl), incubating on dry ice, and centrifugation. The cDNA pellet was then resuspended in water (31μl) and split into two samples. A complementary strand was synthesized using the Multiprime DNA labeling kit (Amersham) with modification. To the cDNA solution, a mixture containing reaction buffer mix (containing dATP, dGTP and dTTP) (10μl), 62.5nM dCTP (2μl), random hexamer (5 μl) and Klenow DNA polymerase I (2u) was added, bringing the final volume to 50 μl. The reaction was incubated at 37°C for 30 minutes. The double stranded cDNA was extracted, precipitated as in the previous step, and resuspended in water (40μl). The cDNA was then digested overnight with restriction enzymes that cut only 3' to two other gene specific primers that were designed for the PCR later in the experiment. (The two gene specific primers were designed such that the sense primer is located 3' to the antisense primer and the two primers were within about 100 bp downstream to the known 5' region of the cDNA). The digested cDNA was then precipitated as described above. The digested cDNA was ligated overnight with T4DNA ligase (GibcoBRL) at 15°C. Some cDNA molecules would contain new 5' sequence ligated to the 3' end of the digested cDNA forming a circular molecule. 2µl of this re-ligated cDNA was used in PCR using the two gene specific primers that initiated DNA synthesis toward the new 5' region.

3.13. Subcloning

DNA fragments digested with restriction enzyme were subcloned into Bluescript plasmid (PBS-KS II) (Stratagene). The vector was cut with the same restriction enzyme(s) that generated the DNA insert. The vector was treated with alkaline phosphatase (Boehringer-Mannheim) to prevent re-ligation of the vector if the vector was linearized with one single enzyme. To achieve an efficient ligation of DNA insert into the vector, a 3:1 insert to vector ratio was used in the ligation reaction. About 5µg of linearized vector and 15-30µg DNA insert were run on a 0.8% low melting agarose. The bands were then purified by phenol-chloroform extraction and precipitated in 99% ethanol. In a 20µl ligation reaction, the purified DNA insert and vector were mixed with the reaction buffer containing 25mM Tris-HCl (pH 7.6), 5mM $MgCl_2$, 1mM ATP, 0.5mM DTT, 2.5%(w/v) polyethylene glycol-8000 and 1 unit of T4DNA ligase (GibcoBRL). The ligation reaction of the insert and vector was incubated overnight at 15°C. 10µl of the ligation product was transformed into 100µl of DH5α competent cells (GibcoBRL). The transformed cells were then plated on agar plates containing ampicillin, X-gal and IPTG. White colonies that contained the plasmid with the DNA insert were grown in LB broth. Plasmid mini-preps and restriction digestion were carried out to determined the authenticity of the subclones.

Subcloning of PCR fragments was carried out according to the protocol provided by the manufacturer of the TA-cloning Kit (Invitrogen). PCR inserts could be released from the pCRII vector by EcoRI digestion.

3.14. Sequencing

Sequencing of plasmid and PCR products was performed using the sequencing kit provided by Amersham. The plasmid sequencing procedure involved the denaturation of 3-5μg plasmid DNA by 0.1 volume of 2mM EDTA and 0.1 volume of 1N NaOH. The denatured plasmid DNA was then annealed to 20ng of gene specific primer in reaction buffer. The annealing mixture was then cooled to below 35°C in a heating block slowly (~30 minutes). The following steps were followed according to the manufacturer's protocol. PCR product sequencing was different from plasmid sequencing in that the PCR product was treated with exonuclease and shrimp alkaline phosphatase to remove residual single stranded PCR products and primers and to remove free dNTPs, respectively.

3.15. Genomic sequence analysis

The genomic DNA sequence was analyzed by several methods.

i) Homology to porcine N-terminal peptide

The porcine N-terminal peptide (KQVTQSYEXL), PN for short, was searched for in the genomic sequence. Since a perfect match was not expected in a comparison across species, the following method was used to identify a match. First, low stringency comparisons were made between PN and the genomic sequence. This would identified a number of poor as well as reasonable homologies. A homology was considered significant if: 1) there was a perfect match of at least 2 amino acids consecutively or intervened by the same number of amino acids, 2) there were flanking amino acids that matched in terms of their properties such as polarity and charges, 3) there was an ATG codon with sequence similar to the Kozak consensus sequence for the translation start site in the ORF and 4) there was a splice donor site at the 3' end of the ORF.

ii) Mouse genomic sequence homology

There is a 4kb genomic fragment at the 5' end of the both human and the mouse gene that is equivalent in terms of the location. Genomic sequences were entered and analyzed in the same database. DNA analysis revealed homology between the human and the mouse genomic sequence. Sequences that had more than 20 base pair identity consecutively were considered highly significant.

iii) Genbank database search on the Internet

DNA or amino acid sequences could be compared to Genbank sequences by using the NCBI-BLAST search engine (Altschul et al 1990). Sequences were either sent to an automated email server (blast@ncbi.nlm.nih.gov) or were entered online (www.ncbi.nlm.nih.gov) for comparison to Genbank sequences. The genomic sequences (0.5–3kb) were compared to the EST (Expressed Sequence Taq) database and the NR database (all non-redundant Genbank+EMBL+DDBJ+PDB but without EST, STS, GSS or HTGS). A match was considered significant if there is at least 90% of homology and a long stretch of unbroken matches of at least 100bp. When the sequence was compared with that from the same species, homologies of less than 70% and matches with very short pieces (around 30bp) were considered as insignificant.

An example of a sequence search that was sent to the automated server is shown below.

```
program blastn
datalib dbest
expect 1000
descriptions 50
alignments 50
begin
>mthfr-cDNA-est
AATTCCGGAGCCATGGTGAACGAAGCCAGAGGAAACAGCAGCCTCAACCCCTGCTTG
GAGGGCAGTGCCAGCAGTGGCAGTGAGAGCTCC
```

The first line "program" of the above program indicates to the BLAST search engine that nucleotide sequence (blast"n") is being compared. The second line "datalib" instructs the search engine to compare the given DNA sequence to the EST database (dbest). The third line "expect" indicates that 1000 matches are expected to be found by chance. The matches will be reported according to the statistical significance of the match. In this case, only 1000 matches will be reported. The line "description" restricts the number of short sequences to be reported to 50. The line "alignment" restricts the number of sequences that have high-scoring significance to be reported.

Amino acid sequence was compared by setting the "program" to "blastp" where the "p" indicates peptide sequence is being compared. The details of the automated server can be obtained at the NCBI Webpage.

4. Results

4.1 Cloning of human MTHFR cDNA 5' ends

4.1.1 Tissues chosen for cDNA cloning

Human fetal lung, fetal kidney and lymphocytes were shown to have the higher molecular weight 77kDa isoform of MTHFR (Frosst 1995). Human fetal lung and fetal kidney tissues, and 2 colon carcinoma cell lines were tested for the presence of the 77 kDa isoform using an antibody generated against purified porcine liver MTHFR. Western analysis showed that the 77kDa MTHFR is expressed in human fetal lung and kidney but not in the colon carcinoma cell lines (Fig. 7a). RNA isolated from human fetal lung and fetal kidney were therefore chosen for cDNA cloning. Human placental RNA, which was obtained in a RACE-PCR kit as a control sample, was also used for cDNA cloning.

4.1.2 Northern analysis

Northern analysis (using a 750bp EcoRI-MseI fragment of the human MTHFR cDNA) on human lymphocytes and colon carcinoma cell lines (Goyette P. and Perreira P. respectively, personnel communication, data not shown) and mouse tissues (Fig. 7b) showed that MTHFR transcript (7-9 kb) is much larger than the cloned 2.2kb cDNA, suggesting large UTRs.

4.1.3 5' RACE-PCR and inverse-PCR

5' RACE-PCR using human placental RNA resulted in the cloning of previously isolated cDNA 5' ends–MRE-5' and M3 (data not shown, see Fig. 8 for clone

Figure 7. a) Western analysis of human MTHFR

Western analysis (using a rabbit polyclonal antibody to purified pig MTHFR) showed that human fetal lung and fetal kidney express the 77kDa isoform of MTHFR while colon carcinoma cell lines express the 70kDa isoform. Lane 1: Expressed fusion protein of the 2.2kb cDNA (72kDa). Lane 2: Colon carcinoma cell line SW 620 (70kDa). Lane 3: Colon carcinoma cell line SW 1222 (70kDa). Lane 4: Human fetal lung (77kDa). Lane 5: Human fetal kidney (77kDa). A non-specific band is often observed on Western blotting of approximately 65 kDa and is seen in all the above lanes.

b) Northern analysis of mouse MTHFR

Northern analysis (probed with a 750bp EcoRI-MseI fragment of the MTHFR cDNA) showed that the mouse kidney has two major MTHFR transcripts (7kb and 9kb) and the mouse brain has a major (9kb) and a minor (7kb) transcript. Lane 1: Mouse brain. Lane 2: Mouse kidney.

c) The presence of MRC-5’ cDNA and MRE-5’ extensions in human tissues

RT-PCR using a sense primer in the MRC-5’ cDNA extension and an antisense primer in exon 1 showed that the MRC-5’ cDNA extension is present in the two human tissues and that the MRE-5’ cDNA extension is also present in these tissues. The 180bp band is the amplification of MRC-5’ and the 450bp band is the amplification of MRE-5’. Lane 1: Human lymphocytes. Lane 2: Human fetal lung. Lane 3: Negative control (without template in PCR).

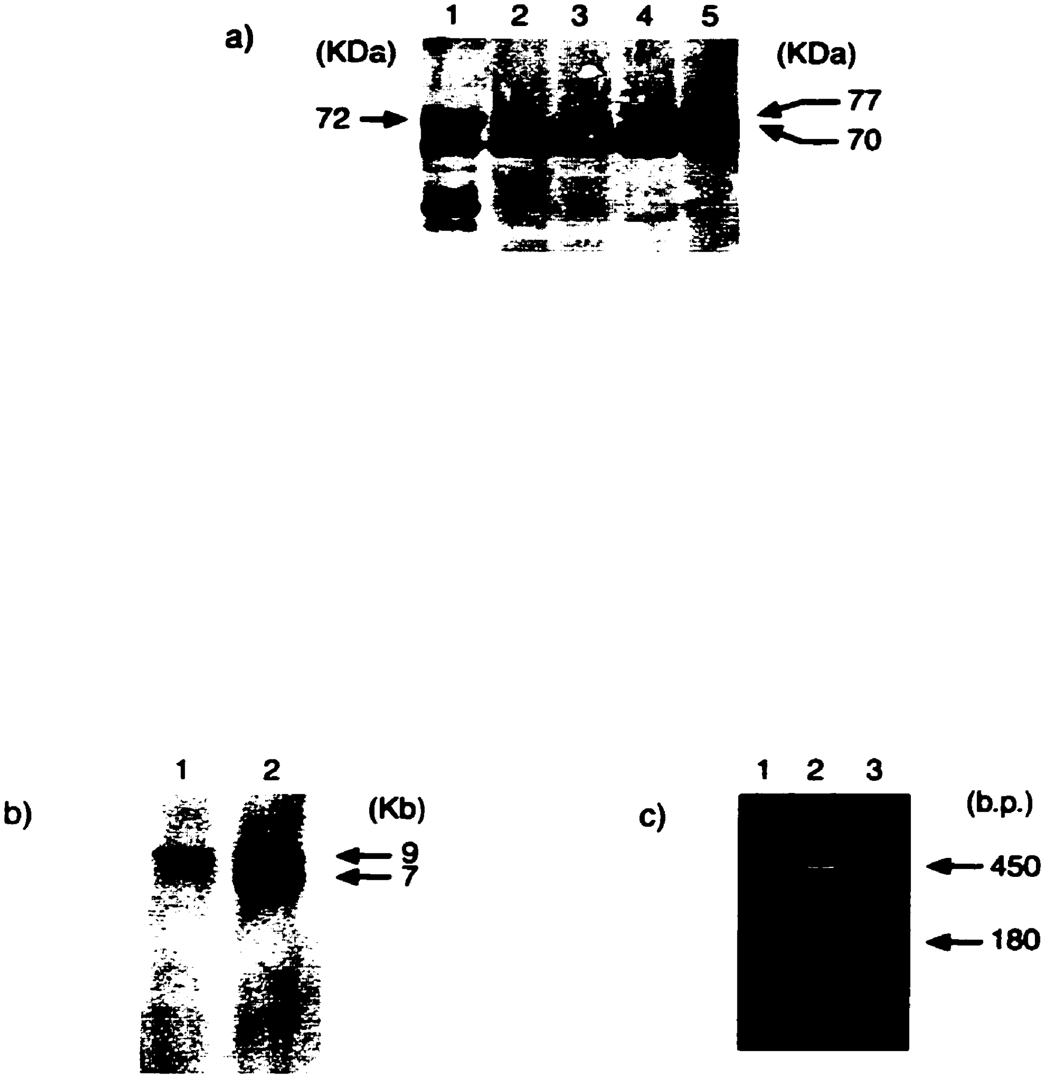

a)
1 2 3 4 5
(KDa)
72
(KDa)
77
70
b)
1 2
(Kb)
9
7
c)
1 2 3
(b.p.)
450
180

Figure 8. Four species of human cDNAs and a pig cDNA of MTHFR.

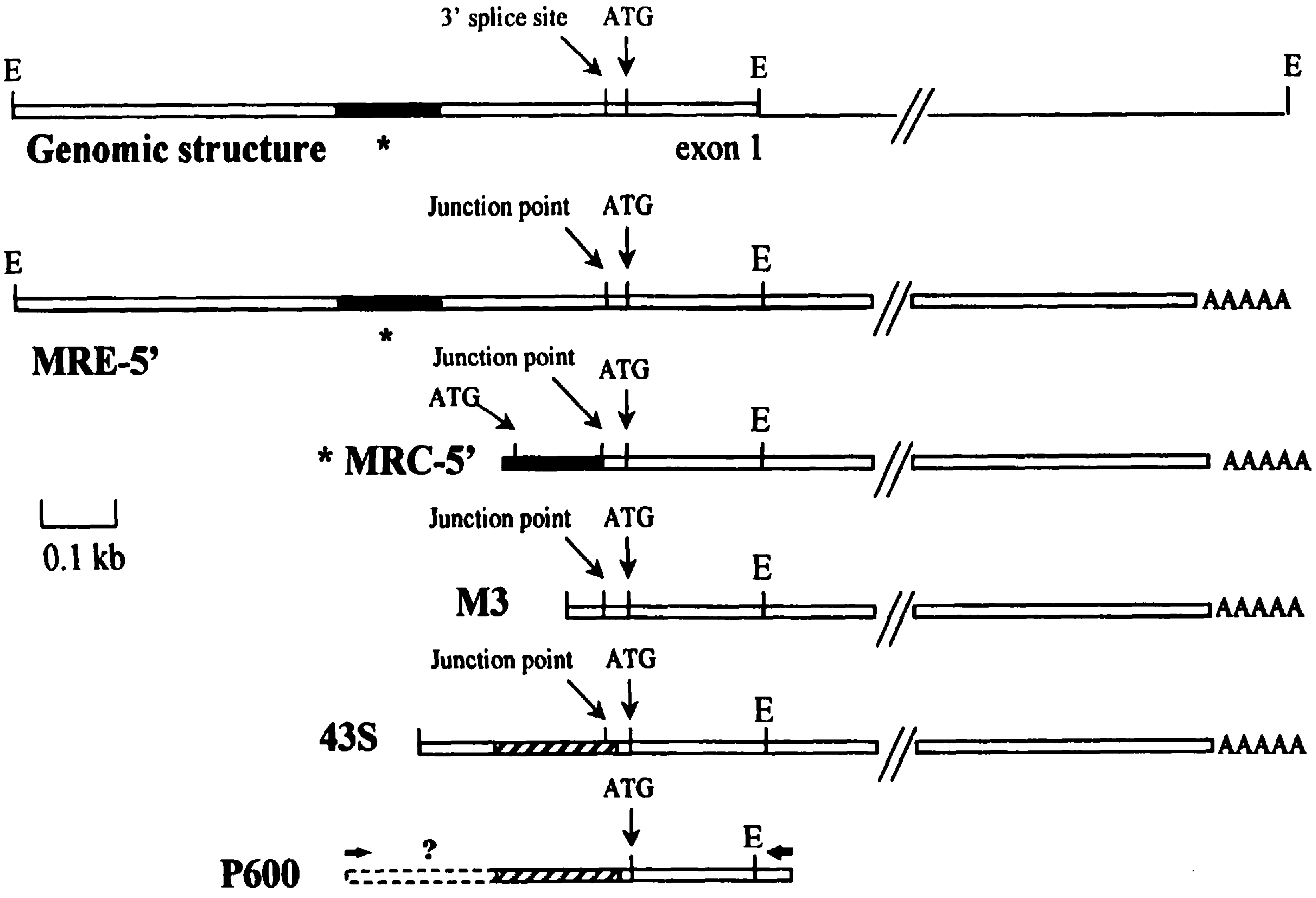
MRE-5' is a 750bp cDNA extension that is contiguous with the original exon 1. It does not contain an ORF or another ATG in frame with the original ATG.

MRC-5' (shown as filled box) is a 130bp exon located within MRE-5' that is spliced into the common splice site in exon 1. It has an ORF and an additional ATG that is in frame with the original ATG in exon 1.

M3 is an 80bp exon that is spliced into the common splice site in exon 1. It has an ORF that is in frame with the ATG of exon 1, but no additional in frame ATG was found.

43S is a 250bp exon that is spliced into the common splice site in exon 1. It does not contain an ORF that is contiguous with the ATG in exon 1. The hatched box represents a 60% DNA sequence identity with the pig cDNA.

P600 is a 600bp porcine cDNA. The amino acid sequence homology between the two species is 90% at exon 1. An ATG codon was found in the same relative location in the human exon 1. Porcine exon 1 contains an ORF that extends into the 5' cDNA extension. The hatched box represents 60% DNA sequence homology between the 5' sequence and the 43S exon. No sequence identity was found at the 5' most sequence.
(see section Fig. 19 for more details)


3' splice site
ATG
E
E
E
Genomic structure
*
exon 1
Junction point
ATG
E
E
AAAAA
*
MRE-5'
Junction point
ATG
ATG
E
* MRC-5'
AAAAA
0.1 kb
Junction point
ATG
E
M3
AAAAA
Junction point
ATG
E
43S
AAAAA
ATG
?
E
P600

information). 5'-RACE-PCR and inverse-PCR using human fetal lung resulted in the cloning of these same two cDNAs as well as a new cDNA 5' extension–MRC-5'. MRC-5' was found to localize within the MRE-5' exon by comparing the sequence of both exons. The authenticity of MRC-5' as a cDNA was confirmed by RT-PCR using a primer within the MRC-5' cDNA extension and a second primer in exon 1 (Fig. 7c). RT-PCR suggests the presence of both MRE-5' and MRC-5' cDNA extensions in the human fetal kidney (data not shown), fetal lung, and lymphocytes (Fig. 7c).

4.1.4 Four species of MTHFR cDNAs

All MTHFR cDNAs vary at the 5' ends with 4 possible exons–MRE-5', MRC-5', 43S and M3 (Fig. 8). MRE-5' and 43S appear to be 5' UTRs because of the lack of an open reading frame (ORF) that continues with the rest of the MTHFR cDNA. Both MRC-5' and M3, on the other hand, possess an ORF that does continue with the rest of the MTHFR cDNA (see section 4.3.1 for details).

4.2 Cloning of 5' genomic fragments of MTHFR

A total of 5 genomic fragments were isolated from the human genomic library and the human PAC clone (Fig. 9) by library screening and Southern analysis respectively.

Figure 9. **Five overlapping genomic fragments at the 5' end of human MTHFR.**

Clone 1 (MRM-5') is a 3 kb fragment isolated from human genomic λ phage library (Mbo1 partial digest) using the 1kb-MRE-5' exon as the probe.

Clone 2 (H2) is an 8 kb fragment isolated from a PAC clone using the first 300 bp of clone 1 as the probe.

Clone 3 (X5) is a 4.5 kb fragment isolated from a PAC clone using the first 300 bp of clone 1 and the M3(80 bp) exon as the probes in different hybridization experiments.

Clone 4 (H4) is a 7.5 kb fragment isolated from a PAC clone using the M3 (80 bp) exon as the probe.

Clone 5 (PAC-3K) is a 3.8 kb fragment isolated by PCR from PAC clone (the same anti-sense primer bound at both the 5' and 3' end).

(B = BamH I, H = Hind III, X = Xba I)

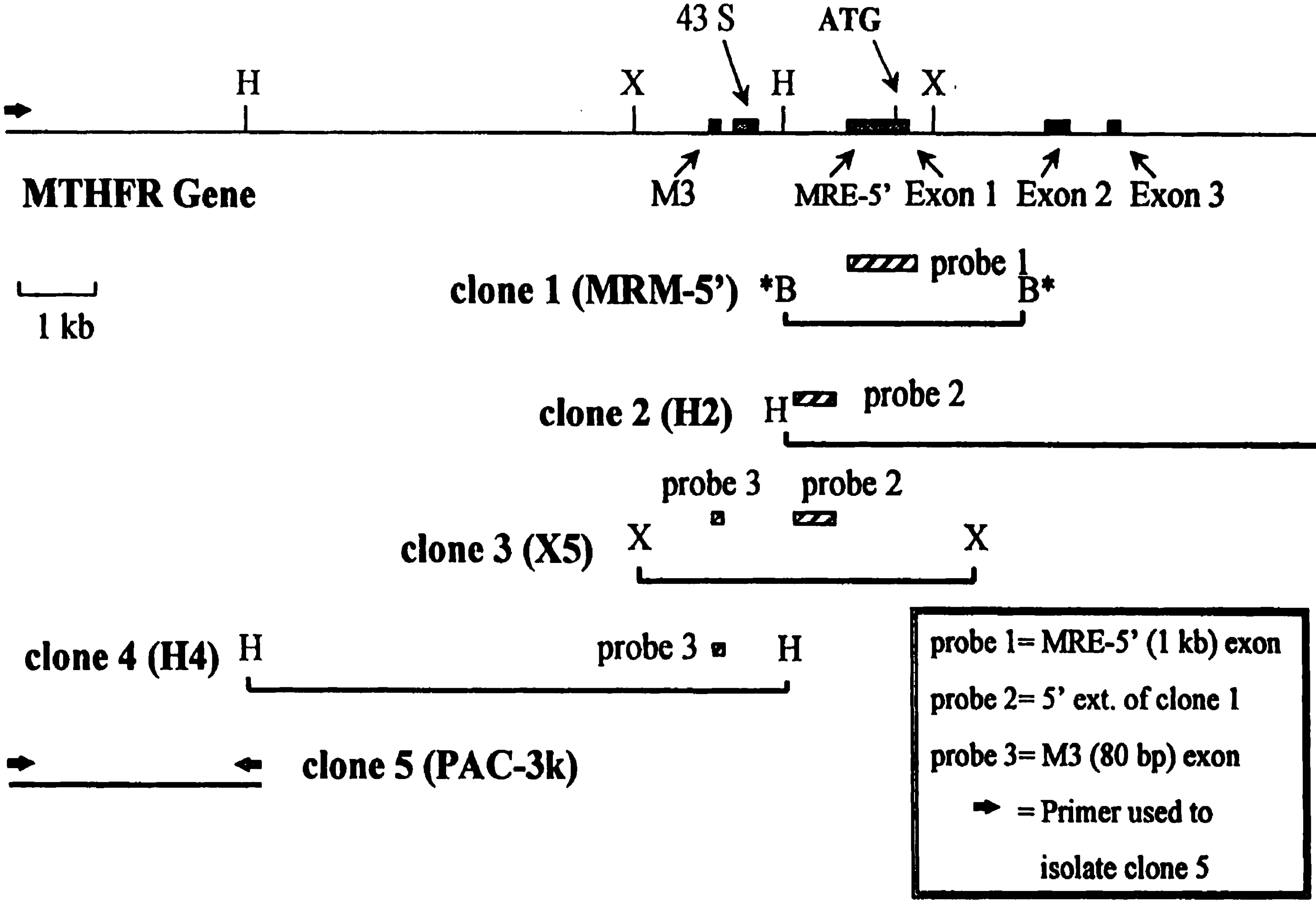

43 S
ATG
H
X
H
X
H
M3
MRE-5'
Exon 1
Exon 2
Exon 3
MTHFR Gene
1 kb
probe 1
clone 1 (MRM-5') *B
B*
clone 2 (H2) H
probe 2
H
probe 3
probe 2
clone 3 (X5) X
X
clone 4 (H4) H
probe 3
H
clone 5 (PAC-3k)
probe 1= MRE-5' (1 kb) exon
probe 2= 5' ext. of clone 1
probe 3= M3 (80 bp) exon
➨ = Primer used to isolate clone 5

Figure 10. **a) Southern analysis of the human genomic fragment (MRM-5')**

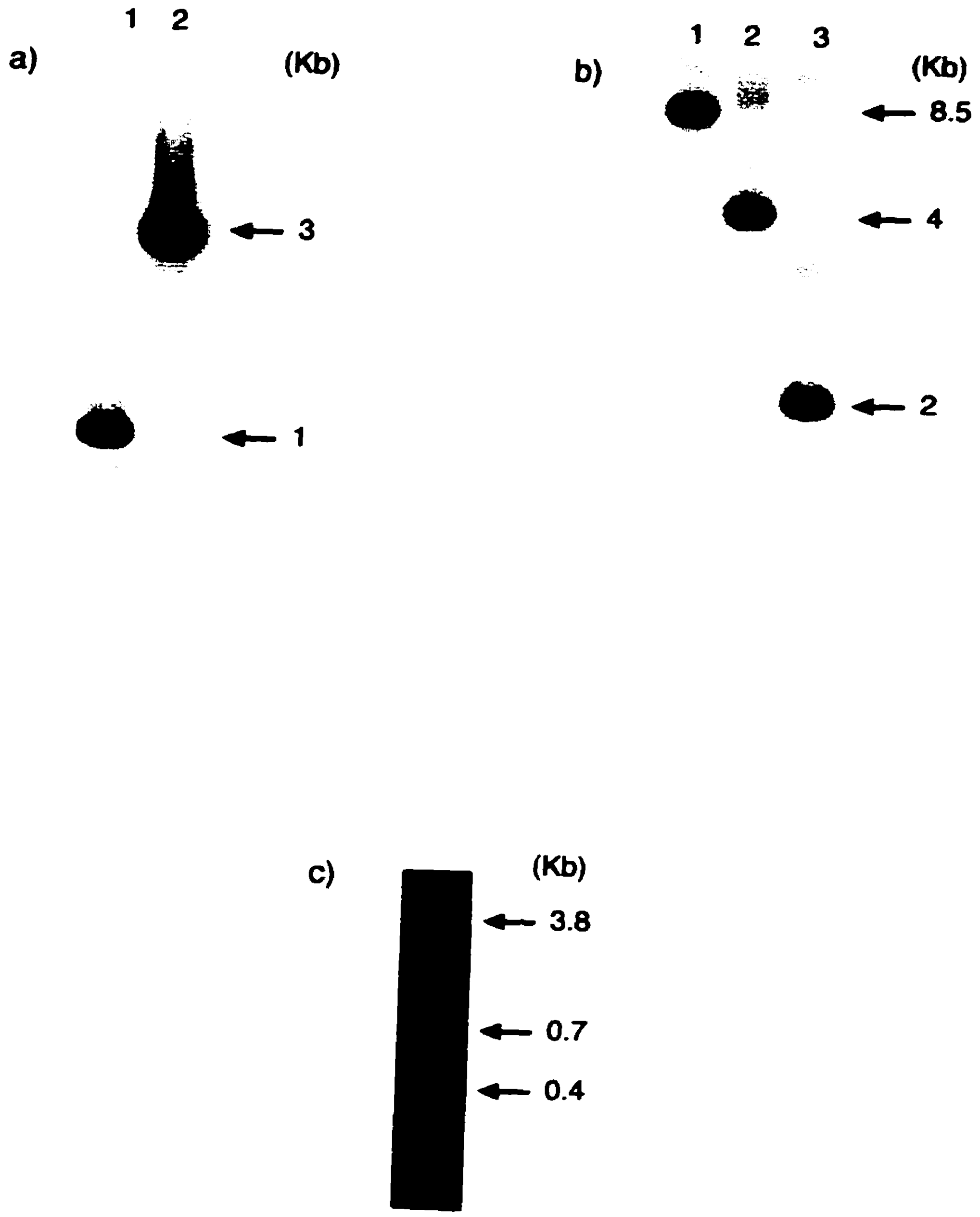
A positive clone from the human lambda phage genomic library was digested with BamH I and hybridized with the MRE-5' cDNA extension probe. A 3kb fragment was identified and was named MRM-5'. Digestion of the phage clone with BamH I and EcoR I release the 1 kb fragment to which the MRE-5' cDNA extension probe hybridized.

b) Southern analysis of a human PAC clone

A human PAC clone (129_L13) was digested with two different restriction enzmyes and hybridized with a 165bp probe generated from the 5' end of the MRM-5' genomic fragment. Lane 1: Digestion with Hind III revealed an 8.5kb genomic fragment (H2). Lane 2: Digestion with Xba I revealed a 4kb genomic fragment (X5). Lane 3: Digestion with Hind III and Xba I revealed a 2kb fragment that was the overlapping region between the genomic clones H2 and X5.

c) PCR product of PAC clone generating a 3.8kb genomic fragment (PAC_3K)

PCR was carried out using the human PAC clone as DNA template, an antisense primer designed from the most 5' human genomic sequences (located in the genomic fragment H4_06) and a degenerate sense primer designed from the porcine N-terminal peptide sequence. From top to bottom, the first band is a 3.8kb genomic fragment generated by specific binding of the antisense primer at the 3' end and non-specific binding at the 5' end. Two other bands (0.7kb and 0.4kb) were found to be non-specific at both ends.


a)
1
2
(Kb)
3
1
b)
1
2
3
(Kb)
8.5
4
2
c)
(Kb)
3.8
0.7
0.4

4.2.1 **Clone 1 (MRM-5')** is a 3kb MboI fragment obtained from library screening using the cDNA extension MRE-5' as the probe (Fig. 10a). This fragment contains exon 1 and the cDNA extension MRE-5', as well as about 300bp of sequence 5' to MRE-5'.

4.2.2 **Clone 2 (H2)** is a 8kb HindIII fragment isolated from the PAC clone using the first 300 bp of clone 1 as the probe (Fig. 10b). This fragment contains the cDNA extension- MRE-5', exons 1, 2 and 3 as well as introns 1, 2 and 3.

4.2.3 **Clone 3 (X5)** is a 4kb XbaI fragment isolated from the PAC clone using cDNA extension-M3 and the first 300bp of clone 1 as the probes (Fig. 10b). This fragment contains the cDNA extensions M3, 43S, MRE-5' as well as exons 1 and 2.

4.2.4 **Clone 4 (H4)** is a 7kb fragment isolated from the PAC clone using the cDNA extension-M3 as the probe (data not shown). This fragment contains the cDNA extensions M3 and 43S. To facilitate the sequencing of this fragment, this fragment was digested with XbaI and the smaller fragments were further subcloned. These smaller subcloned fragments are, in the orientation of 5' to 3', a 0.6kb HindIII-XbaI fragment (H4_06), a 3kb XbaI fragment (H4_3X) and a 1.5kb XbaI-HindIII fragment (H4_15).

4.2.5 **Clone 5 (PAC_3K)** is a 3.8kb PCR generated fragment using the PAC clone as DNA template. The original purpose of the experiment was to use a degenerate oligo designed from the pig N-terminal peptide as the sense primer and the most

5' genomic human sequence as the anti-sense primer, in an attempt to isolate the sequence(s) that are homologous to the pig N-terminal peptide. Since the human MTHFR sequence might be missing the N-terminal sequences, this approach might result in isolation of the human region that is homologous to the N-terminal peptide. Among the various nonspecific bands obtained from this experiment is a 3.8kb band that has the anti-sense primer bound specifically at the 3' end and also non-specifically at the 5' end (Fig.10c). Although the original aim was not achieved, an extra genomic fragment was obtained.

4.3 Genomic sequence of the 5' end of MTHFR

The significant findings from sequencing a total of 11 kb of genomic sequence are best summarized in Fig. 11. Details of the sequences are described in the following subsections. In summary, sequencing of genomic fragments localized 4 human MTHFR 5' exons to a 4kb XbaI fragment. A possible CpG island was identified in the region of the 43S exon. An overlapping gene ClC-6, a putative chloride ion channel transcribed in the opposite strand and in the opposite direction was also identified. The first exon of the ClC-6 gene is located approximately 3.5 kb upstream of the start codon of MTHFR. Genbank EST database search localized 2 EST clones to two genomic fragments.

Figure 11. The 5' region of the human MTHFR gene.

Black and open boxes represent human MTHFR exons (MRE-5', MRC-5', 43S and M3). Dotted boxes represent exons of the human putative chloride ion channel (CLC-6) gene, which is transcribed on the opposite strand and in the opposite direction. Hatched boxes represent sequences that were identified in EST clones. The EST sequences are transcribed in the opposite direction to MTHFR. Horizontal striped box represents sequences that are thought to represent a CpG island. The names of the subclones are provided at the bottom of the diagram. (Note that the PAC-3k fragment has not been completely sequenced. Exon 3 of the ClC-6 gene has not been localized but is thought to be located within this fragment.)

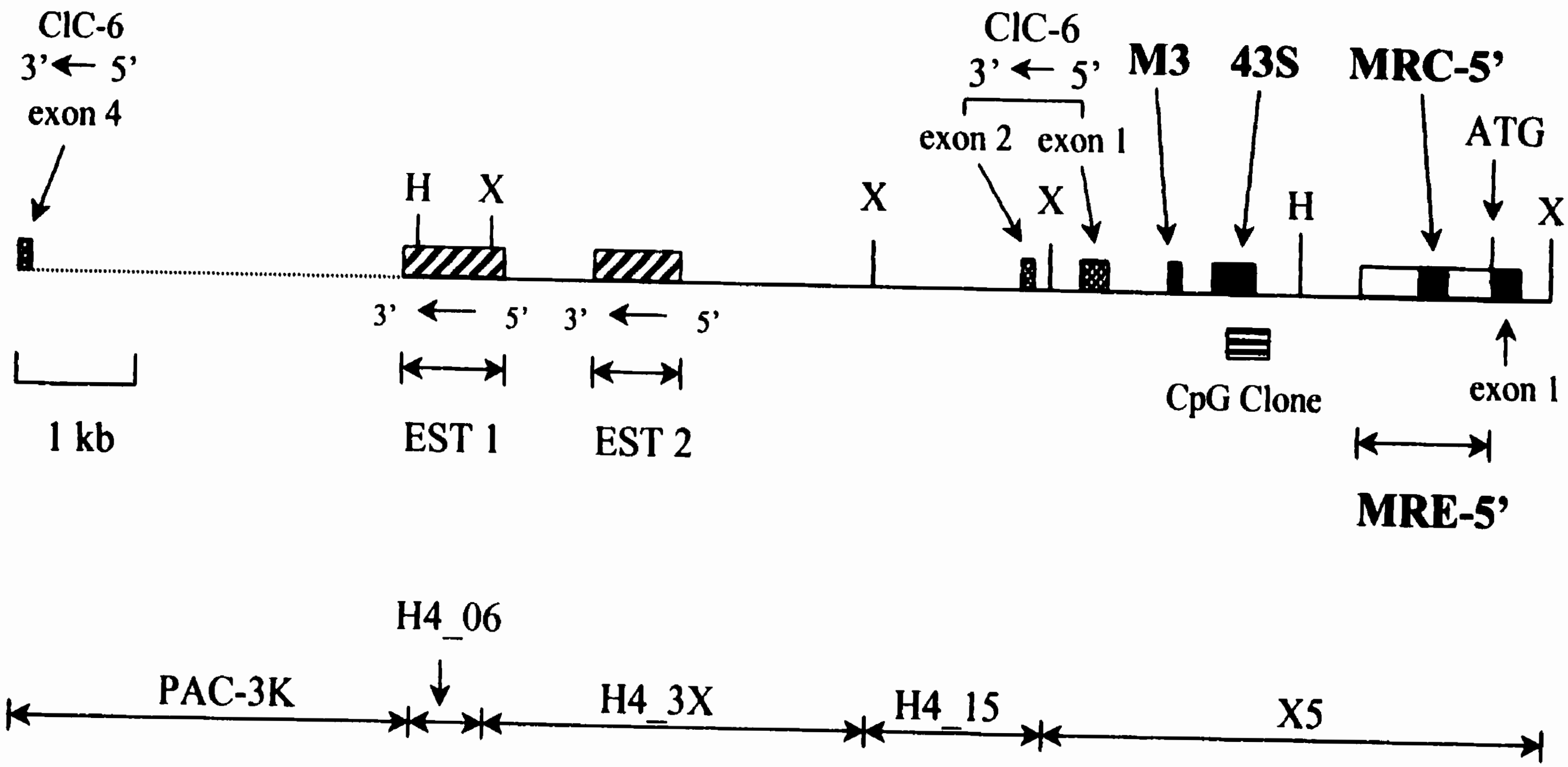
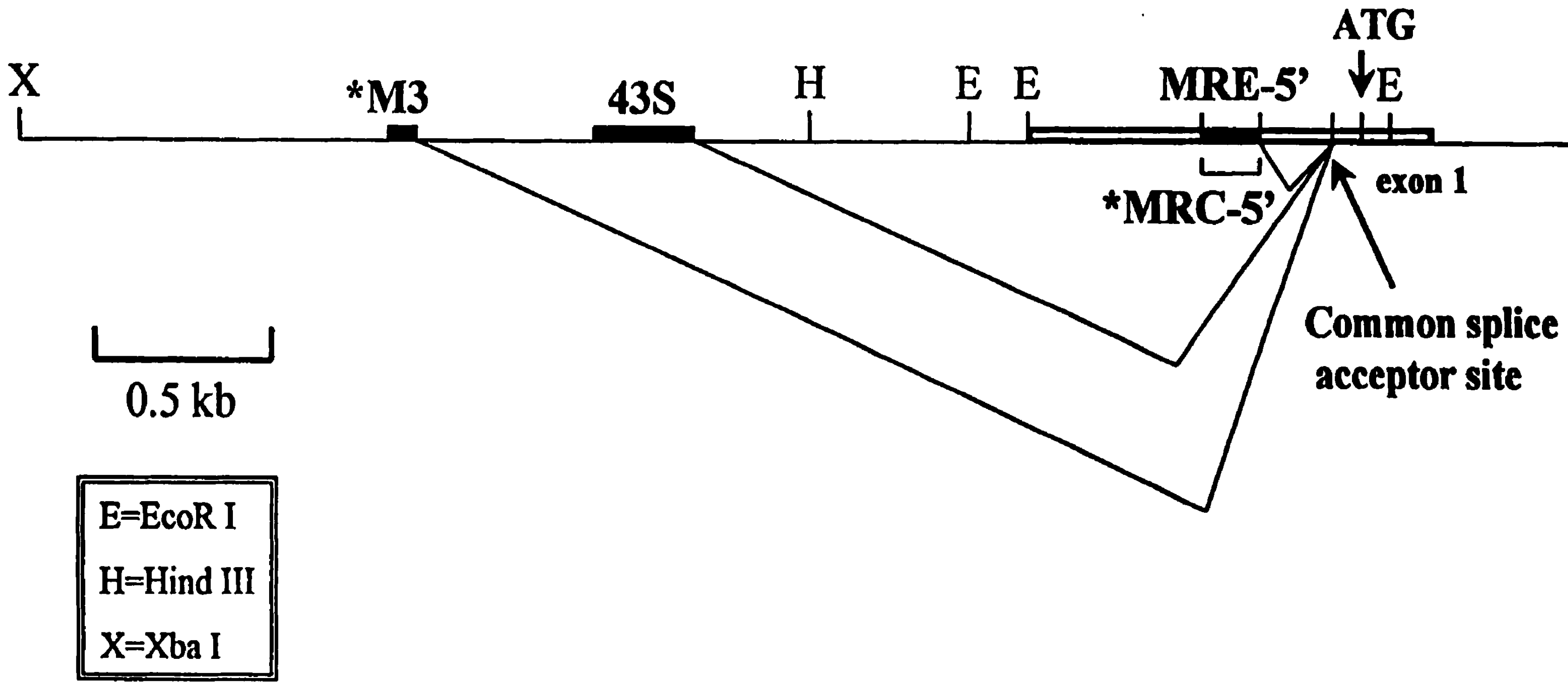

ClC-6
3'← 5'
exon 4
H
X
X
ClC-6
3' ← 5'
exon 2
exon 1
X
M3
43S
H
MRC-5'
ATG
X
3' ← 5'
3' ← 5'
CpG Clone
exon 1
1 kb
EST 1
EST 2
MRE-5'
H4_06
PAC-3K
H4_3X
H4_15
X5

Figure 12. Alternative splicing of 4 exons at the 5' end of human MTHFR gene.

Black and open boxes represent exons. Both MRC-5' and M3 have an ORF that is contiguous with the ATG in exon 1 of MTHFR. MRE-5' and 43S are believed to be part of the 5' UTR. The 5' end of each exon has not been confirmed. The 5' boundaries of MRC-5' and M3 were defined by the primers which allowed amplification of the cDNA in RT-PCR. The 5' boundaries of MRE-5' and 43S were defined by the end of the cDNA clone.


X
*M3
43S
H
E
E
MRE-5'
ATG
E
X
*MRC-5'
exon 1
Common splice acceptor site
0.5 kb
E=EcoR I
H=Hind III
X=Xba I

4.3.1 Sequence of the X5-4kb genomic fragment

Sequencing of this genomic fragment revealed the following:

i) Alternative splicing of 5' exons

Alternative splicing had been suggested based on the presence of various cDNA 5' ends. Sequencing of this fragment provided molecular evidence on the mechanism of the alternative splicing event. The 5' exons MRC-5', M3 and 43S are alternatively spliced into a common splice acceptor site (Fig. 12). A splice acceptor site was identified at base pair 3969 in the clone and is 13bp upstream of the ATG start codon (Fig. 13). Splice donor sites at the 3' end of each of these exons were also identified.

ii) MRE-5' (base pair 3201-3969)

The MRE-5' is a cDNA extension of approximately 750bp that is directly linked to the ATG in exon 1 without an intron in between (Fig. 13). The 5' boundary of this exon is defined by the end of the cDNA clone that contains this exon. The MRE-5' cDNA extension has an ORF of 80bp at the 3' end that continues with the ATG in exon 1 without any splicing. But this ORF is relatively short and it does not contain a ATG site with a proper start site consensus sequence nor does it have a match with the pig N-terminal peptide sequence. Furthermore, no splice site is identified at the 5' end of this ORF. Therefore, this exon is likely to be a 5' UTR. The counterpart of this exon is also present in the mouse gene (data obtained from P. Tran)

Figure 13. The DNA sequence of the X5-4kb genomic fragment in the 5' region of human MTHFR gene.

Sequence of a 4kb Xba I genomic fragment subcloned from the PAC clone. All human exons are shown in boxes and in bold. Dotted lines on top of the sequence indicate the location and size of the human or mouse exons, primers and a Genbank CpG DNA clone (Z58297). Short underlined sequences are consensus sequences of splice sites (GT for splice donor site or AG for splice acceptor site), the Kozak ATG translation start site (A/G-C-C-A-T-G-G/A) or sequences of restriction sites. Open reading frames (ORFs) of the sequences where an exon is found are shown in bold.

Four human MTHFR exons are found to splice into a common splice acceptor site at bp3969. The 5' boundary of the MTHFR exons M3, MRC-5' are defined by the end of a primer (that allows amplification in RT-PCR) while the 5' boundary of 43S and MRE-5' are defined by the end of the cDNA clone. The amino acid sequence of the mouse counterpart for the human exon MRC-5' has been aligned and shown in the corresponding location. Underlined sequence between bp 340-420 and bp 780-950 represents identical sequence between the human and mouse. A potential match with the porcine internal peptide (Pi-12) is identified at bp 2380-2405. The human ClC-6 gene (Genbank AF 009247) has an ORF that is transcribed on the opposite strand and in the opposite direction.

```
              Xba 1
         10     ↓    20         30         40         50         60
     *    *     *    *     *    *     *    *     *    *     *    *
GGTGCGGCCG CTCTAGAAAA ACAGATCCGC AGAGGGAGCC TTTTTTAGTG GTTAGGAGTG

         70         80         90        100        110        120
     *    *     *    *     *    *     *    *     *    *     *    *
GAAGGAGGCA GAAGTATTTA CTGAAATCAG TGGCTAGCTT CTAAATCATC TGGAGCCATT

        130        140        150        160        170        180
     *    *     *    *     *    *     *    *     *    *     *    *
ATCAAATAGG AACCAGCCCT CAAAAAAAAC CTTTCGGGGG TGAGGGACCA TGTGGGTGAG

        190        200        210        220        230        240
     *    *     *    *     *    *     *    *     *    *     *    *
GATCTACAGC CATCAGCTGA GCTCTTCATT TCCACTGATA GTCTCCAAAT AACCACCCTC

        250        260        270        280          290
     *    *     *    *     *    *     *    *     *     *     *
CTCTTCCAGG ACACCTCAAA GATGTCCAAC NNCAGCTGA AAA GGG GGT AAA ATG CAG
                                          Lys Gly Gly Lys Met Gln>

300              310         320         330         340
 *        *       *     *     *     *     *     *     *     *
GTT CCA TTT GAC AGT GTG ACG TAT CTG AAA TCA GAA AGG ACT TGT CAA CTC
Val Pro Phe Asp Ser Val Thr Tyr Leu Lys Ser Glu Arg Thr Cys Gln Leu>
                                         Asn Gln Lys Gly Leu Val Asn Ser>
                   Gln Cys Asp Val Ser Glu Ile Arg Lys Asp Leu Ser Thr>

    |------- primer X5S352 -------→|  |--- mouse MTHFR 5' exon ----

350         360         370         380         390
 *     *     *     *     *     *     *     *     *     *
TGG GAA CAC AAC TCA AGT TTT CCC AGG ATG CTT TGC AGG GGG AAG CTG GAC
Trp Glu His Asn Ser Ser Phe Pro Arg Met Leu Cys Arg Gly Lys Leu Asp>
 Gly Asn Thr Thr Gln Val Phe Pro Gly Cys Phe Ala Gly Gly Ser Trp Thr>
Leu Gly Thr Gln Leu Lys Phe Ser Gln Asp Ala Leu Gln Gly Glu Ala Gly>

---------------------- mouse MTHFR 5' exon ------------------

400         410         420         430         440         450
 *     *     *     *     *     *     *     *     *     *     *
TGT CAG TG ACC CAG AAA GGT GAG GGA TAGAG TGAGAAAGGA CTGGAGAAGC
Cys Gln>
 Val Ser  Asp Pro Glu Arg>
Leu Ser Val Thr Gln Lys Gly Glu Gly>
```

```
----------------------- mouse MTHFR 5' exon --------------------

           460          470          480           490         500
      *      *     *      *      *     *      *      *     *      *
TAA AAC AGC AGC ATG ATA AGC ACA AAG TCC TGT GAG GAA GCT CAT TCT GAA
    Asn Ser Ser Met Ile Ser Thr Lys Ser Cys Glu Glu Ala His Ser Glu>

                                                                  |-

----------------------- mouse MTHFR 5' exon ---------------------

         510           520          530          540           550
    *     *      *      *     *      *     *      *     *      *
AAC GCT TGT TTC ATT CCA AAC TCT TTT CAG ATG GAA ATA AAA GGA AAC ATG
Asn Ala Cys Phe Ile Pro Asn Ser Phe Gln Met Glu Ile Lys Gly Asn Met>


----- primer X5S552 ---->|

----------------------- mouse MTHFR 5' exon ---------------------

        560          570           580          590          600
  *      *      *      *      *      *     *      *      *     *
GGT GGG ATT TAC TGG AGC TGG CCT GGA TCT CCC TCA GAT TCC AGG AGG GGT
Gly Gly Ile Tyr Trp Ser Trp Pro Gly Ser Pro Ser Asp Ser Arg Arg Gly>


--------->|

       610          620         630           640          650
  *      *      *      *      *     *      *      *      *     *      *
TAT GAG AAA AGA CCC CAG ACT TAGGCA CGTGA AGC AGG GTA GAC GCT TCG AGA
Tyr Glu Lys Arg Pro Gln Thr                            Thr Leu Arg Glu>
                                         Ser Arg Val Asp Ala Ser Arg>


   660          670          680           690          700
    *      *      *      *     *      *      *      *      *     *
GCC CTG GCT GCG GTC CCC AGG CCC CAC CCN CTG CCA CCT GCG GCC CAG ATT
 Pro Trp Leu Arg Ser Pro Gly Pro Thr Xxx Cys His Leu Arg Pro Arg Leu>
Ala Leu Ala Ala Val Pro Arg Pro His Pro Leu Pro Pro Ala Ala Gln Ile>


710           720           730          740           750
  *      *      *      *      *      *      *      *      *      *
GGG CCC CCA CCC CCG GCA ACG CTC TCT CAG TCC CTT AGC AAC CGC CCC CTC
 Gly Pro His Pro Arg Gln Arg Ser Leu Ser Pro Leu Ala Thr Ala Pro Ser>
                                                   Gln Pro Pro Pro>
Gly Pro Pro Pro Pro Ala Thr Leu Ser Gln Ser Leu Ser Asn Arg Pro Leu>
```

```
                                        | 3' <--- ClC-6 gene (exon 1) --- 5'

                                    Splice donor site (for ClC-6 gene)
  760          770          780    ↓     790          800
    *      *      *      •      *      *      *      *      •      •
 CCC ACC GAG CTC GCC GGC TTC TTA CCA GCT CCT CGG GGG TGC GGG TCT CAC
                             AAT GGT CGA GGA GCC CCC ACG CCC AGA GTG
                                <Leu Glu Glu Pro Thr Arg Thr Glu Arg

  Pro Pro Ser Ser Pro Ala Ser Tyr Gln Leu Leu Gly Gly Ala Gly Leu Thr>
 Pro His Arg Ala Arg Arg Leu Leu Thr Ser Ser Ser Gly Val Arg Val Ser>
  Pro Thr Glu Leu Ala Gly Phe Leu Pro Ala Pro Arg Gly Cys Gly Ser His>



                                                             |---------

 3' <------------------- ClC-6 gene (exon 1) ------------------ 5'

 810          820          830          840          850          860
   *      *      *      *      *      •      •      *      •      •      •
 GCT CAC CGC AGC AGC AGC ACC ACC TGC AGC AGC AGC ACA GAG ACC CCC TGC
 CGA GTG GCG TCG TCG TCG TGG TGG ACG TCG TCG TCG TGT CTC TGG GGG ACG
 <Glu Gly Cys Cys Cys Cys Trp Arg Cys Cys Cys Cys Leu Ser Gly Arg Cys

  Leu Thr Ala Ala Ala Ala Pro Pro Ala Ala Ala Ala Gln Arg Pro Pro Ala>
 Arg Ser Pro Gln Gln Gln His His Leu Gln Gln Gln His Arg Asp Pro Leu>
  Ala His Arg Ser Ser Ser Thr Thr Cys Ser Ser Ser Thr Glu Thr Pro Cys>



 -- primer X5S854--->|

 3'<---------|<--------- ClC-6 gene (partial promoter) --------- 5'

              870          880          890          900          910
     *      *      *      *      *      *      •      *      •      *
 ACC CCG CCA TCT TCC TCC TTT ACT GCC ACT CTG GAC CCC TCT ACC AAC CCC
 TGG GGC GGT AGA AGG AGG AA
  Gly Ala Met

  Pro Arg His Leu Pro Pro Leu Leu Pro Leu Trp Thr Pro Leu Pro Thr Pro>
 His Pro Ala Ile Phe Leu Leu Tyr Cys His Ser Gly Pro Leu Tyr Gln Pro>
  Thr Pro Pro Ser Ser Ser Phe Thr Ala Thr Leu Asp Pro Ser Thr Asn Pro>



 3'<------------------- ClC-6 gene (partial promoter) --------- 5'

            920          930          940          950          960
    *      *      *      *      *      *      *      *      *      *
 CTC CCA GCC AGG ATC TGC GCC TCA CGT GAC TGG CCC GGC CGC CAC GGG TCA
  Ser Gln Pro Gly Ser Ala Pro His Val Thr Gly Pro Ala Ala Thr Gly His>
 Pro Pro Ser Gln Asp Leu Arg Leu Thr>
  Leu Pro Ala Arg Ile Cys Ala Ser Arg Asp Trp Pro Gly Arg His Gly Ser>
```

```
3'←------------------- ClC-6 gene (partial promoter)-------------- 5'

      970       980        990       1000      1010
 *     *     *     *     *     *     *     *     *     *
CGT GGC CCT CTC GAG CTC TGG GAC GTG GAC CGA ACA GAC GGG TGG GGC GAG
 Val Ala Leu Ser Ser Ser Gly Thr Trp Thr Glu Gln Thr Gly Gly Ala Arg>
Arg Gly Pro Leu Glu Leu Trp Asp Val Asp Arg Thr Asp Gly Trp Gly Glu>


3'←------------------- ClC-6 gene (partial promoter) ------------ 5'

    1020      1030      1040       1050      1060
 *     *     *     *     *     *     *     *     *     *
GAC TCG CGT CAC ATG ACG ATA AAG GCA CGG CCT CCA ACG AGA CCT GTG GGC
 Thr Arg Val Thr>
Asp Ser Arg His Met Thr Ile Lys Ala Arg Pro Pro Thr Arg Pro Val Gly>


3'←------------------- ClC-6 gene (partial promoter) ------------- 5'

  1070       1080      1090      1100       1110
*     *     *     *     *     *     *     *     *     *     *
ACG GCC ATG TTG GGG GCG GGG CTT CCG GTC ACC CGC GCC GGT GGT TTC CGC
Thr Ala Met Leu Gly Ala Gly Leu Pro Val Thr Arg Ala Gly Gly Phe Arg>


           |------- primer X5 S1124 ---------→|

←more 5'?|------------------ MTHFR exon (M3-120bp)------------------→

3'←--- ClC-6 ----- 5'|

  1120       1130      1140      1150      1160
     *     *     *     *     *     *     *     *     *     *
CCT GTA G|GC CCG CCT CTC CAG CAA CCT GAC ACC TGC GCC GCG CCC CTT CAC|
          Ala Arg Leu Ser Ser Asn Leu Thr Pro Ala Pro Arg Pro Phe Thr>
Pro Val  Gly Pro Pro Leu Gln Gln Pro Asp Thr Cys Ala Ala Pro Leu His>


---------------------- MTHFR exon (M3-120bp)----------------------→

 1170       1180      1190       1200      1210
     *     *     *     *     *     *     *     *     *     *
|TGC GTT CCC CGC CCC TGC AGC GGC CAC AGT GGT GCG GCC GGC GGC CGA GCG|
 Ala Phe Pro Ala Pro Ala Ala Ala Thr Val Val Arg Pro Ala Ala Glu Arg>
Cys Val Pro Arg Pro Cys Ser Gly His Ser Gly Ala Ala Gly Gly Arg Ala>
```

```
-----MTHFR exon (M3-120bp)------→| Splice donor site
                                   ↓
1220        1230        1240       1250      1260      1270
   *      *     *      *      *      *    *      *    *     *   *
 TTC T GAG TCA CCC GGG ACT GGA GG G TGAGTGAC GGCGAGGCGG GGTCGNNGGG
  Ser  Glu Ser Pro Gly Thr Gly Gly>
 Phe>


       1280      1290      1300      1310      1320      1330
          *    *     *    *     *    *     *    *     *    *     *    *
 AGGGAGATCC TGGAGCCGGC AAACAACCTC CCGGGGGCAA GGACGTGCTT GTGGGCGGGG


       1340      1350      1360      1370      1380      1390
          *    *     *    *     *    *     *    *     *    *     *    *
 AGCGCTGGAG GCCGGCCTGC CTCTCTTCTT GGGGGGGGCC TGCCGCTCCT TGCGCACCCT


       1400      1410      1420      1430      1440      1450
          *    *     *    *     *    *     *    *     *    *     *    *
 TCGCGGATTA GTGTAACTCC AATGGCTACC ACTTCCAGCG ACCGCCAACC CTCAAGCGAA


       1460      1470      1480      1490      1500      1510
          *    *     *    *     *    *     *    *     *    *     *    *
 GACTGACTTT NGCTCCCTGC CTGGACGGAG GGNCCCCTGA GCAGGTGACG ATCCCGCCCC


             ←more 5' ??  |---------- MTHFR exon (43S-240bp) -----→

       1520      1530       1540      1550      1560      1570
          *    *     *    *      *    *     *    *     *    *     *    *
 TCTGACCGGC NCANNCCGTG TN CTCGCCCC CATCGGTGAC TCAGTGACCT GGTGACTGGA

                                                  Mse I
                                                   ↓---CpG DNA clone--→

--------------------- MTHFR exon (43S-240bp)-------------------→

       1580      1590      1600      1610      1620      1630
          *    *     *    *     *    *     *    *     *    *     *    *
 TTCTCGGCCA CCTGGGGCCG ACGCGTTCCG GCTCCTGCTT TTAAACCTGC CTCCCCGGCG


----------------- Genbank CpG DNA clone (Z58297) ----------------→

--------------------- MTHFR exon (43S-240bp)-------------------→

       1640      1650      1660      1670      1680      1690
          *    *     *    *     *    *     *    *     *    *     *    *
 ATCACCTGGA GAAGAGCGCT GGGCCCGGGG CACTGCTCCC TGGCGCCCAC TGCGTCCCCG
```

```
----------------- Genbank CpG DNA clone (Z58297) -------------->

---------------------- MTHFR exon (43S-240bp) ----------------->

         1700      1710      1720      1730      1740      1750
     *    *     *    *     *    *     *    *     *    *     *    *
 TGCGCACGGG GGTCCGCCGG GCCATTTCTG GGAGTCGTAG GCTTAGTATC CCAGTGCTTG


----------------- Genbank CpG DNA clone (Z58297) --------------->

--- (exon 43S) --->| Splice donor site
                   ↓
         1760      1770      1780      1790      1800      1810
     ●    ●     *    *     ●    ●     *    *     ●    *     *    ●
 GCGCAGACTA GTTGTTCA GT AGTGGCAGAG GCTTATTTGA GAGAGTGGCA GCACCTGGCC


----------------- Genbank CpG DNA clone (Z58297) --------------->

         1820      1830      1840      1850      1860      1870
     *    ●     *    *     *    *     ●    *     *    *     ●    *
 CTTTNNCTCA GTGAATGTTG GCTATCACCG TGTGCCAAAC TCTGGGGATA CCCCAGNCAG

 ------------->|

         1880      1890      1900      1910      1920      1930
     *    ●     ●    ●     ●    *     ●    ●     ●    *     *    *
 GACACCGGTC TTCTCAGGGA ACTGGGGAAA GAGAAAGGAG ACAGGCCTTT TCACCCACAG

         1940      1950      1960      1970      1980      1990
     *    ●     *    *     *    *     *    *     *    ●     *    ●
 TTACAACCCA GGGTGCTATG GGAGTCCAGC AGATAACGGA TAAATCGTGG GAGTTGGCTT

         2000      2010      2020      2030      2040      2050
     *    *     ●    *     ●    ●     ●    *     ●    ●     ●    *
 ACAAASATGG CACATGCGTG GATATACTAG GAATGCAATA AGTCTTTGAA AATCAGAGGG

         2060      2070      2080      2090      2100      2110
     *    *     *    *     *    *     *    *     *    *     *    *
 TTTACAGGTG GTTCAGCTTC CTCCTACTCT AGGTTCTGTT CCAGCAAGCA ATTAACGAGG

         2120      2130      2140      2150      2160      2170
     ●    *     *    *     ●    *     *    *     *    ●     ●    *
 TGCGCCCTTA AACGCTGGAG GAAACGCAAC TGGCTGCTCT TGCTGTTACT CCTCCCCCCC

         2180      2190      2200      2210      2220      2230
     *    ●     *    ●     *    *     ●    ●     *    ●     ●    *
 GCCCCGTTCC TCACTCCCCA CAGCCATCCC CACTGAGAAT CTGGAGTTTG AGGTCAGAAT
```

```
                        Hind III
      2240       2250     ↓2260        2270       2280      2290
    *     •     •    •    *    •     *    *     •    •     *    •
GAAAGAGAGC AGCCTAGAGG GAGAAAGCTT TGGCCCAGGG TTCTTAGTCT GGAATCAACT

      2300       2310       2320       2330       2340      2350
    *     •     •    *    •    *     *    •     •    *     *    *
CCTTGTCTTT GGATGTATCC CCGTGTAGTC TGTGCACCTG TGTGTGTATT TCAGGGGAAG


                                                 |---primer x52382---

      2360         2370         2380         2390         2400
    *    *     *      •     *      •      •     •      *      •     *
GGAGCAGTGC ATTTAA TCA GAT TGT CAA AAG AGT CTA AGA CCC CAA ATG GTT AGG
                  Ser Asp Cys Gln Lys Ser Leu Arg Pro Gln Met Val Arg>
```

Homology to **porcine peptide** (**Pi-12**): Arg Ala Gly **Pro Gln** Val **Val Arg**

```
--------->|

  2410         2420         2430          2440        2450
     *      •     *      *      •     •      •      *     •      *
TAC ACA GGG TTA GTG GTG GAC AGT CTG AAA GAA ATG AAC CTC ACC TGG GCT
Tyr Thr Gly Leu Val Val Asp Ser Leu Lys Glu Met Asn Leu Thr Trp Ala>

 2460         2470        2480          2490         2500
    *     *      *      •     •      *      *     *      *      *
TTC CTC TGT TGT GCC ATG TCA CCA CAC ACC CAT TCA CTA CTG TGT GTT TGC
Phe Leu Cys Cys Ala Met Ser Pro His Thr His Ser Leu Leu Cys Val Cys>

2510           2520         2530         2540      2550       2560
    •      •      •     •      *      *     *     *   *      •    *
 CCA TTG CTG TGC AAG TGT TTT GTT TGT TTT TAA GTGTTTGTCT TATTTTCTTA
 Pro Leu Leu Cys Lys Cys Phe Val Cys Phe>


      2570       2580      2590       2600       2610       2620
    *    *      •    •    *    *     •    *     *    *     *    *
ACCAGACTGC CAGATGACCC TATGCCCTCT GTTGGCCTGT CTGTGCCCTG GTGGCTCTGA

      2630       2640      2650       2660       2670       2680
    *    •     *    *     •    •     •    •     •    •     *   *
TTACTTGTTT CTAGTTTTTT GTTTTTGTTT TTTTTTTTTG AGATGGAGTT TTGCTTTTGT
```

```
      2690       2700       2710       2720       2730       2740
    ●    *     *    *     *    ●     *    ●     ●    *     *    ●
GCCCAGGCTG GATGCTATGG CACAATCTTG GTTCACTGCA ACCTCTGCTT CCGGGTTCAA

      2750       2760       2770       2780       2790       2800
    *    ●     ●    *     *    *     ●    *     ●    *     ●    ●
GCGATTCTCC TGCCTCAGCC TCCAGGTAGC TGGGTTACAN GCCCGCCACC AAACCTGGCT

      2810       2820       2830       2840       2850       2860
    ●    *     ●    ●     ●    *     ●    *     *    *     ●    *
AATTTTTATA TTTTTAGTAG AGTCGGGATT TCACCATGTT GGCCAGGCTG GTCTCAAACT

      2870       2880       2890       2900       2910       2920
    *    *     *    *     ●    *     ●    ●     *    ●     *    *
CCTGGCCTCA GGTGATCCAC CCACCTCGGC CTCCCAAAGT GCTGGGATTA CAGGTGTGTT

      2930       2940       2950       2960       2970       2980
    *    ●     ●    *     ●    *     ●    ●     *    ●     ●    *
TTTTGTTTTT TTAAGAGATG GAGTCTCGCT ATGTTCGCCA GGCTGGCCTT GAACTCTGGG

      2990       3000       3010       3020       3030       3040
    *    *     ●    *     *    *     ●    *     ●    ●     *    *
CTCAGGCAAT CCCTTCTGCC TCAGCTTCCC AGTAGCTGGT CACTGTGATG ATTTGAATTG

      3050       3060       3070       3080       3090       3100
    *    *     *    *     *    *     ●    ●     ●    ●     ●    *
AATTCTGTGA TGTGTAAGAA GAGCAGCCTG CAAGGCAAGC AGCAGATGGG GCAGCTTTTG

      3110       3120       3130       3140       3150       3160
    *    *     *    ●     ●    ●     ●    *     *    *     ●    ●
TTCTGAGAAA TTCGTGCCCT TACTGAACCT TGGGTCTGGC TATTTTTGGA ACATGGCCAG

                                              |---- (MRE-5') --->
                                             EcoR I
                                              ↓
      3170       3180       3190       3200        3210       3220
    ●    *     ●    *     *    *     *    *     ●    ●     *    *
NATCAAGTTC TAACCCACAA CACGGTCAAA AAGGATAGCA T GAATTCAGG AGAAATCTGG

---------------- MTHFR exon (MRE-5'- 800bp) --------------------→

      3230       3240       3250       3260       3270       3280
    *    *     *    *     *    *     *    *     *    *     *    *
CTGCATAGTC AAGCCCTCAC CCCTTCCATC CTGTGCACGA ACTGTTTCAA GTAACAGATG
```

```
---------------------- MTHFR exon (MRE-5'-800bp) --------------->

       3290       3300       3310       3320       3330       3340
          *    *     *    *     *    *     *    *     *    *     *    *
 TTCCAGCAAG AGCCAGCCAG AGTGAGCTGT TCCTTCTCTG GAGGGTGATC TGGTATCCCT


---------------------- MTHFR exon (MRE-5'-800bp) --------------->

       3350       3360       3370       3380       3390       3400
          *    *     *    *     *    *     *    *     *    *     *    *
 GAACGCCTGT TGGCCTCATC TCCACCAACC CCTGCAGTCT CTGCCCCTGA GTCCCCCTCC


---------------------- MTHFR exon (MRE-5'-800bp) ------------------>

       3410       3420          3430         3440         3450
          *    *     *    *     *      *      *      *      *      *
 TTCCATCCGC CTCCCCTTAC TAG AGC CTC AGC CCT CCC TCC TCG CCT GGA AGC CTT
                           Ser Leu Ser Pro Pro Ser Ser Pro Gly Ser Leu>

     (Mouse sequence):     Phe Glu Leu Arg Leu Ser Leu Ser Pro Ala Gly->


---------------------- MTHFR exon (MRE-5'-800bp) ---------------->

 3460         3470         3480          3490         3500
      *    *       *    *       *    *       *    *       *    *
 GCC CCC GCC CCC TTG TGC TGG CTG GAG CTC AAG CCT CTT CCT TTG TCG CAG
 Ala Pro Ala Pro Leu Cys Trp Leu Glu Leu Lys Pro Leu Pro Leu Ser Gln>

 Lys Pro Tyr Ser Arg Pro Phe Ala Trp Pro Gln Leu Gln Leu Gly Ala Leu->


                                    |----- primer MRC-5'-S5 -->|

                                    |--- MTHFR exon (MRC-5'-160bp)->

---------------------- MTHFR exon (MRE-5'-800bp) ---------------->

 3510          3520        3530        3540          3550
    *      *     *     *     *     *     *     *      *     *
 CTC CGC CCA GTT GAA CAC ACC CGC TGG GGA AGG TGC CTC TGT TCC CTC CCC
 Leu Arg Pro Val Glu His Thr Arg Trp Gly Arg Cys Leu Cys Ser Leu Pro>

 Cys His Ser Ser Ala Gln Leu Gly Thr Pro Ser Arg Lys Gly Ser Phe Pro->
```

```
------------- Human/Mouse MTHFR exon (MRC-5'-160bp) -------------→

---------------------- MTHFR exon (MRE-5'-800bp) ---------------→

3560      3570       3580      3590      3600
  *    *    *    *    *    *    *    *    *    *
ACG CAC TCT GGG CCT GAG CTG ACA GAG ATG GAC CAT CGA AAA GCC AGG GTC
Thr His Ser Gly Pro Glu Leu Thr Glu Met Asp His Arg Lys Ala Arg Val>

Thr His Ser Gly Ser Glu Gln Ile Gly Met Asn His Gln Lys Ala Lys Phe→


--------------- Human/Mouse MTHFR exon (MRC-5'-160bp) -----------→

---------------------- MTHFR exon (MRE-5'-800bp) ---------------→

3610      3620       3630       3640      3650      3660
 *    *    *    *    *    *    *    *    *    *    *
CTC CCA GCT GGG CAC TAC TGC CCC TCG CTA GGA ATA TGG GCC TCG CAG GTC
Leu Pro Ala Gly His Tyr Cys Pro Ser Leu Gly Ile Trp Ala Ser Gln Val>

Phe Pro Ala Gly His Cys Tyr Pro Ser Leu Gly Met Trp Ala Ser Glu Ala→


                                              Splice donor site
                                                   ↓
--------- MTHFR exon (MRC-5'-160bp) --------→|

---------------------- MTHFR exon (MRE-5'-800bp) -------------→

          3670      3680      3690      3700      3710
   *    *    *    *    *    *    *    *    *    *
GGC AGC GTG AGG TCC TCT GTG CCA CCT TCC ATC AGG TAGC TGTCACCGAG
Gly Ser Val Arg Ser Ser Val Pro Pro Ser Ile>

Gly Cys Val Arg Leu Ser Val Pro Pro Ser Ile (mouse sequence)


---------------------- MTHFR exon (MRE-5'-800bp) -------------→

       3720      3730      3740      3750      3760      3770
   *    *    *    *    *    *    *    *    *    *    *    *
GAGCATGTTG CAGTGCCGGG TGGGGGCTGC TTGCATGCAA GGAGCCTGGC AGCAGCGGAG


---------------------- MTHFR exon (MRE-5'-800bp) -------------→

       3780      3790      3800      3810      3820      3830
   *    *    *    *    *    *    *    *    *    *    *    *
GGCAAGGCTT TGAGTGAGGC GGCCCGGACA GCCATAGCTG AGGAGCATGG AGCCACTGGG
```

```
---------------------- MTHFR exon (MRE-5'-800bp) --------------->

        3840       3850       3860       3870       3880       3890
       *    *     *    *     *    *     *    *     *    *     *    *
 AGGGGGCAGT GTCACCTTTT TTGGCCTTCT TCCTGTGTGG AAATACAGCG CCTCCGGCTT
```

```
---------------------- MTHFR exon (MRE-5'-800bp) --------------->

           3900          3910         3920         3930         3940
       *        *        *      *        *        *      *        *       *      *
 GA ACC TGC CAC TCA GGT GTC TTG ATG TGT CGG GGG TGT GGC TGC CTG CCC
    Thr Cys His Ser Gly Val Leu Met Cys Arg Gly Cys Gly Cys Leu Pro>
```

```
                                             Common splice
                                             acceptor site
 -----MTHFR exon (MRE-5'-800bp)----->|↓                  |---------->

           3950          3960          3970          3980          3990
       *       *       *       *       *       *       *       *       *       *
 CCT GAT GCT CCC TGC CCC ACC CTG TGC AGT AGG AAC CCA GCC ATG GTG AAC
 Pro Asp Ala Pro Cys Pro Thr Leu Cys Ser Arg Asn Pro Ala Met Val Asn>
```

```
--------------------------- MTHFR exon 1 ------------------------->

          4000          4010          4020          4030          4040
      *       *       *       *       *       *       *       *       *       *
 GAA GCC AGA GGA AAC AGC AGC CTC AAC CCC TGC TTG GAG GGC AGT GCC AGC
 Glu Ala Arg Gly Asn Ser Ser Leu Asn Pro Cys Leu Glu Gly Ser Ala Ser>
```

```
--------------------------- MTHFR exon 1 ------------------------->

         4050          4060          4070          4080          4090
    *       *       *       *       *       *       *       *       *       *
 AGT GGC AGT GAG AGC TCC AAA GAT AGT TCG AGA TGT TCC ACC CCG GGC CTG
 Ser Gly Ser Glu Ser Ser Lys Asp Ser Ser Arg Cys Ser Thr Pro Gly Leu>
```

```
--------------------------- MTHFR exon 1 ------------------------->

        4100          4110          4120          4130          4140
    *       *       *       *       *       *       *       *       *       *
 GAC CCT GAG CGG CAT GAG AGA CTC CGG GAG AAG ATG AGG CGG CGA TTG GAA
 Asp Pro Glu Arg His Glu Arg Leu Arg Glu Lys Met Arg Arg Arg Leu Glu>
```

```
--------------------------- MTHFR exon 1 -------------------------→

     4150          4160          4170          4180          4190
 *     *      *      *     *      *      *     *      *      *     *
TCT GGT GAC AAG TGG TTC TCC CTG GAA TTC TTC CCT CCT CGA ACT GCT GAG
Ser Gly Asp Lys Trp Phe Ser Leu Glu Phe Phe Pro Pro Arg Thr Ala Glu>


------- MTHFR exon 1 -------→|

   4200         4210           4220       4230        4240       4250
      ●      *      *      *      ●     *     *     *     *     *    ●
GGA GCT GTC AAT CTC ATC TCA AG G T AAACTCATGC AAGGTTAAGG TGGGAGGNNN
Gly Ala Val Asn Leu Ile Ser Arg>


        4260       4270       4280       4290       4300       4310
         *    *     *    *     *    *     ●    *     *    ●     *    *
GAGTGGTGGT GCCTGGGGAG NAAACTGTAN CCTGAGGCTG GGCCTGCCCT TTAACCCGGN

        4320       4330       4340       4350       4360       4370
         ●    *     *    ●     *    *     *    *     *    *     *    *
AGAAAAAAGC AGCCGGAGGG GGCAGGAAGG AGAAGGCAGA GTGGGTCATC TGATCTGCAA

        4380       4390       4400       4410       4420       4430
         ●    *     *    *     *    ●     ●    *     *    ●     *    ●
CTTCTCTGTT GCCNGTAAAA GTCCCTCCAC AGTGTCAAGG TATTTTAGAC ACACTCTTCA

        4440       4450       4460       4470       4480       4490
         *    *     ●    *     *    *     ●    *     *    *     ●    ●
ATGGGTTAAA AATGTTGATT CTCACACTGT ATTCTTTGAG GGAACTTGGG GGCTGGATCA

        4500       4510       4520       4530       4540       4550
         *    *     *    *     *    *     *    *     *    *     *    *
CGAGTGCAGT TTAGGAAGTC ACAAAGCAGT GAGATTGTTT TCAGCAAGTT AGAAATTTGG

        4560       4570 Xba I
         *    *     *    * ↓
ATTATGCTTC TTGAACTCTA GA
```

iii) **MRC-5' (base pair 3535-3695)**

The MRC-5' is a 160bp exon located 280bp upstream of the ATG start codon of the cDNA (Fig. 13). The 5' boundary of this exon is defined by the 5'end of a primer (MRC-5'-S5) which allows the amplification of the cDNA in RT-PCR. Although located within an UTR (MRE-5'), MRC-5' has a at least 160bp ORF that continues with the MTHFR cDNA when spliced into the common splice acceptor site 13 base pair 5' of ATG start codon of the 2.2kb human cDNA. The ORF extends more than 160bp upstream according to the genomic sequence, but whether the cDNA does contain this upstream sequence has not been confirmed (Fig. 13). The mouse counterpart for the MRC-5' exon has also been identified (data obtained from P. Tran). A 66% sequence identity between the human and the mouse predicted amino acid sequence was observed for this exon. An ATG codon (Met) containing sequence that resembles the Kozak start site consensus sequence was identified at base pair 3586 of the clone. This ATG codon is also present in the mouse gene.

iv) **43S (bp 1533-1768)**

The 43S is a 240bp exon located 2.2 kb upstream of the ATG start codon for the MTHFR cDNA (Fig. 13). The 5' boundary of the exon is only defined by the end of the cDNA clone. This exon is alternatively spliced into the common splice acceptor site 13bp upstream of the start codon. This exon does not contain an ORF that continues with the MTHFR cDNA and is likely to be an UTR. Interestingly, a Genebank database search reveal a CpG island clone (Z58297) that has sequences from this region of the gene. The CpG island was originally cloned by a method which was designed to purify CpG islands in the genome (Cross et al 1994). The CpG clone is about 150bp covering most of the 3' end of the 43S exon and the beginning of the intron downstream (Fig. 13). This suggests that the 43S exon may be close to the end of a transcript, since it is common to have CpG islands in the promoter and the beginning of a transcript (Cross et al 1994).

v) M3 (bp 1123-1241)

The M3 was an exon thought to be 80bp in length when first isolated from a cDNA library (Fig. 13). RT-PCR using primers from genomic cloned showed that this exon is in fact at least 120bp in length. The exon is 2.7 kb upstream of the ATG start site of the MTHFR cDNA. The 5' boundary of this exon is not well defined. The current 5' boundary is defined by the 5' end of a sense primer (X5-S1124) that allows the amplification of the MTHFR cDNA (with an antisense primer in exon 1) in RT-PCR. The M3 exon contains an ORF when spliced into the common splice acceptor site 13bp upstream of the start codon. No ATG codon was identified in this ORF.

vi) Sequence homology between the human and mouse genes

Genomic sequences between the human and the mouse have been compared and homologies have been identified. As mentioned earlier, the counterpart for the human MRE-5' exon and the MRC-5' exon are present in the mouse. MRC-5' share 66% identity of the predicated amino acid between the two species suggesting that this exon may be a coding exon (Fig. 13). Furthermore, an ATG codon (Met), with a strong homology to the Kozak translation start site consensus sequence (A/G-C-C-A-T-G-G/A) has been identified at the same location in both species.

Another region with a strong DNA homology between the two species is between base pair 340 and 420 of the X5 clone (Fig. 13). Part of this sequence was found in a mouse 5' exon. However, RT-PCR failed to prove that this genomic region is transcribed as an exon in the human.

vii) **Overlapping gene (ClC-6-exon 1)**

Another area of strong DNA sequence homology between the human and the mouse was identified between bp 780-950 (Fig. 13). Despite possible ORFs in this region, RT-PCR failed to prove the sequence to be a MTHFR exon. A Genbank search later discovered that this region also belongs to another human gene, a putative chloride channel gene ClC-6 (Brandt and Jentsh 1995, Genbank X83378). However, this gene is transcribed on the opposite strand and in the opposite direction to the MTHFR gene suggesting that the two genes may overlap. Exon 1 of the ClC-6 gene falls into the region of strong DNA sequence homology between the mouse and the human, explaining the reason of the homology. The mouse ClC-6 gene was also identified in the mouse gene (P. Tran, personal communication). The ORF of the ClC-6 gene reading from the opposite direction to the MTHFR gene, as well as the splice sites are shown in Fig. 13.

viii) **Homology between the human and the pig sequences**

A search for porcine peptide sequences in the ORFs of this 4kb genomic sequence was performed with various stringencies. No homology was found between the porcine N-terminal peptide sequence and the predicted ORFs in the human sequence. A 50% homology to the porcine internal peptide (Pi-12) was found between base pair 2380-2405 (Fig. 13). However, RT-PCR failed to prove the presence of such an exon.

Figure 14. The DNA sequence of the H4-15 genomic fragment in the 5' region of human MTHFR gene.

Sequence of a 1.6kb Xba I genomic fragment subcloned from the PAC clone. Homology to the porcine N-terminal peptide sequence was identified at bp 801-833. Identical amino acids are underlined and shown in bold. Similar amino acids are underlined. The ATG codon and the homologous Kozak translation start site consensus sequence are underlined. Primer designed to test the authenticity of the ORF is shown as dotted line. Exon 2 of the human ClC-6 gene is located at bp 1670-1610, being transcribed in the opposite strand and opposite direction. Consensus sequences of the splice donor site and acceptor site for the ClC-6 gene exon are underlined.

```
                                                              Xba I
        10         20         30         40         50        ↓60
    *    •     *    •     •    *     •    •     •    •     •    *
CGACGGTATC GATAAGCTTG ATATCGAATT CCTGCAGCCC GGGGGATCCA CTAGTTCTAG

        70         80         90        100        110        120
    *    *     *    •     •    •     •    *     *    •     •    •
ACAGTAGAAA ATCCCAGGTC CTAATTATGT TGTCAAATAA CTTACAAAAA AAAAAAAAGA

       130        140        150        160        170        180
    •    •     •    *     •    *     •    •     •    *     •    •
TGAAGGAAGA AGCTACAGAT TAAAAGAGAT TTAACAGATT TATCAAACAA ATAAAAAAGG

       190        200        210        220        230        240
    *    *     •    •     •    *     •    •     •    •     •    *
ACTTTTGGAA CCTAATTCAA ATAAAGTTAA AATACATATG ACATTTACAA GAAAACTGGA

       250        260        270        280        290        300
    •    •     *    •     •    *     •    *     •    *     *    *
AAATCTGGAC AGTAGATATT TGATATCAAG AAATTCCTAA TTTTTTAGGA TGTGGTAATA

       310        320        330        340        350        360
    •    *     •    •     •    *     *    •     •    *     •    *
GTATTGGGTT ATGCTTTTTA AAAGTCCTTA TCCATTAGAG ATGCATATAG AATTATGGGT

       370        380        390        400        410        420
    •    *     *    *     •    *     *    *     *    *     •    *
GAAATAATAT GATGTACAGA ATTTCTTTCA AAATAACATG GGAAGGGCGG CCAGGTATGG

       430        440        450        460        470        480
    •    *     *    *     •    *     *    •     •    *     •    *
TAGATCATGA CTGTAATCCC AGAACTTTGG GAGGCTGACA TGGGAGGATC GCTGGTGTCA

       490        500        510        520        530        540
    *    *     •    •     •    •     •    •     *    *     •    •
AGGAGTTCGA GAACAGCCTG GGCAACATAG GGAGACTGTC TTTACTTTAA AACACACACA

       550        560        570        580        590        600
    *    •     •    *     *    *     *    *     •    *     •    •
CACACACACA CACACACACA CACACACACN AACACACACA CACATACAAT AAAAAATAAT

       610        620        630        640        650        660
    *    *     *    •     *    *     *    *     •    *     *    *
AATATGAGAA AGGAGAAGTG GATAACGAAA GCACAGACAA AAGAAGACTG GGCTTGAGCT

       670        680        690        700        710        720
    •    *     *    •     *    •     •    •     *    •     •    •
AAATCACTGT GGAAGCTGTG TGATGAGGAC AAGGGGTTTC AACCAGATTA TTCTGTTTTT
```

```
        730          740           750          760          770
    •     *      •     •      •      •     •      •      •     •
GCATATAG TTG AAA TCC TTC ATA ATA AAA GTT AAA CAA AAA AGT GCT ACA AAT
         Leu Lys Ser Phe Ile Ile Lys Val Lys Gln Lys Ser Ala Thr Asn>


                                     |------- primer H4S803 -------

      780          790          800           810          820
 *      *     •      •      *     •      •      •     •      •
CCT AAC GCT GAC ACT AAA TTC TAC AGC ATG AAG AAT TTA TGG AAG AGT GAG
GGA TTG CGA CTG TGA TTT AAG ATG TCG TAC TTC TTA AAT ACC TTC TCA CTC
Pro Asn Ala Asp Thr Lys Phe Tyr Ser Met Lys Asn Leu Trp Lys Ser Glu>

               Homology to porcine
               N-terminal peptide (PN): Lys Gln Val Thr Gln Ser Try

---→|

    830           840          850          860           870
•     •      •      •     •      •      •     *      •      •     •
ACT CTG TCA AAA AAA AGA AAA GGT TGT GGA AAA TTA CCT AAT GAT ACA AAA
Thr Leu Ser Lys Lys Arg Lys Gly Cys Gly Lys Leu Pro Asn Asp Thr Lys>

Glu xxx Leu   ( porcine N-term. peptide )


   880          890           900         910        920        930
     •      •     •      •      *     •     *     •     •     •    •
AGT ATA ATA AAG GGG GTG AGT GGG ACA GGT TAAAA TAGATGATTT AGTGGAATCT
Ser Ile Ile Lys Gly Val Ser Gly Thr Gly>


       940        950        960        970        980        990
    *    •     •    *     •    •     •    •     •    *     *    •
CATTTTTGTT ACTTAAAAAA AGCATGTATC ATGTATATAA AAGCATGGGT AAGGATGTAG

      1000       1010       1020       1030       1040       1050
    *    *     •    •     •    •     •    •     •    •     •    •
AAAGATACTT AGTAAGGTGT TAACAGTCAT CTTTTAATGA ATGGAGTTAT AGGCTTTTCT

      1060       1070       1080       1090       1100       1110
    *    •     •    •     •    •     *    •     •    *     •    *
CTACCTTCTT TATTACGTTT AAGAATTTAA TGATAAAAAG TCACGCATGT TGTTTCACCC

      1120       1130       1140       1150       1160       1170
    *    •     *    •     •    •     •    •     •    *     •    •
CAGCAGTGTT GGCTTTACTA AAAACACATA CTCTAAGAAT CCCATACATG AAAGGTGGAA
```

```
          1180           1190           1200           1210           1220           1230
      *     •      •     •      •     •      •     *      •     •      •     •
CTTAATCAAG AACATGTCAG AGAGAAGCTG CTTGAGCAAG AAAATATCTG AAAAGCAAGC

          1240           1250           1260           1270           1280           1290
      •     •      •     •      •     *      •     •      •     •      •     •
TCTCCTAGAA GAGAAGCAAG GCCTCCTTAA CAGTTTATAA GCACACACCT GTTAGGGGCA

          1300           1310           1320           1330           1340           1350
      •     *      *     *      •     •      •     •      *     •      *     •
CATACTGCCT GCCAGTGCCT TTCACTCTAA GTCTCTCTGT AAGCTGTAAA AGCAAACAGT

          1360           1370           1380           1390           1400           1410
      *     *      *     •      *     •      *     •      •     *      *     •
TTAATGGCAA ATACATACTC TATCCAGGTT AAGGTGCCAC AGATTTATAA ACGTGTAAAA

          1420           1430           1440           1450           1460           1470
      •     •      *     •      •     *      *     •      •     *      •     •
ACACCAGACC TTGTTTTGGG GTTAAGGTGA AGAAAAGGGC TTGCTAAACA GTCTCTTTCC

          1480           1490           1500           1510              1520
      •     *      *     •      *     •      •     •      •         •      •
TAGAAGGAAG TTCTCTTACC CACTCACTCC CTGGGACAGA AGGTCA AAG AGC ATC TCT

 1530            1540            1550            1560            1570
   •      *       *       •       *       •       *       •       •       *
GGT TGG AAT CAT AAT GCT TCG GCC TGA AGT GAC CAG GCC ACT CAC TAC TTT


                                           |3'←---ClC-6 (Exon 2)---- 5'
1580          1590           1600          1610           1620           1630
  *      •      •      *      •      *      •      •      •      •      •
AGT TGC CCA AGC AGT ATC AAA GGA GCT CAC CTC ATA GTC TTT CCT TGG AAG
TCA ACG GGT TCG TCA TAG TTT CCT CGA GTG GAG TAT CAG AAA GGA ACC TTC
                                  <Val Glu Tyr Asp Lys Arg Pro Leu
                                     ↑
                      Splice donor site for ClC-6 gene


  3'←-------------- CLC-6 gene (exon 2) ------------ 5'|
                                                      Xba I
             1640          1650          1660          1670 ↓  1680
     *      •      •      •      *      •      •      *      •    *
AAT CTC ATC CTC CTC CTC CTG TGT TTC TCC AAG GAT GGT C TAGAGCGGCC
TTA GAG TAG GAG GAG GAG GAC ACA AAG AGG TTC CTA CCA G ATCTCGCCGG
<Ile Glu Asp Glu Glu Glu Gln Thr Glu Gly Leu Ile Thr
                                                     ↑
                                           Splice acceptor site
                                              for ClC-6 gene
```

4.3.2 Sequence of H4_15 genomic fragment

The H4_15 is the next genomic fragment 5' to theX5-4kb fragment. Sequencing of this fragment revealed the following:

i) **Overlapping gene (ClC-6-exon2)**

Exon 2 of the human ClC-6 gene was identified at the end of this 1.6 kb genomic fragment, between bp 1610-1670. The ORF and the splice sites for the ClC-6 gene are shown in Fig. 14.

ii) **Homology to the porcine N-terminal peptide**

A 20% identity and a 50% similarity was found between the porcine N-terminal peptide and the ORF predicted from the region at bp800-834 of this clone (Fig 14). The entire ORF is 177bp in length. An ATG codon with 86% homology to the Kozak translation start site consensus sequence was identified at bp 801 (Fig 14). Despite the homologies, RT-PCR failed to prove this sequence to be exonic.

4.3.3 Sequence of the H4_3X genomic fragment

The H4_3X is the next genomic fragment 5' to the genomic fragment H4_15. The sequencing of this fragment revealed the following:

i) **ORFs with homology to the porcine N-terminal peptide**

An ORF of 252bp was found at bp1220-1470 of the clone (Fig. 15). A 40% identity to the porcine N-terminal peptide sequence was found between bp 1249-1279. Another ORF of 81bp was found at bp 1470-1554. A 30% identity and 50% similarity to the porcine N-

Figure 15. The DNA sequence of the H4_3X genomic fragment in the 5' region of the human MTHFR gene.

Sequence of a 3.1kb Xba I genomic fragment subcloned from the PAC clone. The dotted line represents open reading frames (ORF), primers or the sequence that was found in Genbank EST clones. EST clone AA058363 covers the first 165 bp of this fragment and EST clone H71759 covers bp 1086-1704. The sequence homologous to the Kozak sequence are underlined and the ORFs that contain such sequence and a homology to the porcine N-terminal peptide sequence (PN) are shown in bold. The identical amino acids to the porcine peptide sequence are underlined and shown in bold.

```
←--- Genbank EST clone (AA058363) continues 650bp upstream -----

         10  Xba I  20         30         40         50         60
    •    •    ↓ *    *    *    •    •    •    •    *    *    *
ATCCACTAGT TCTAGAATAG GCAAATAGAC AGAAAGGAGA TCACTGGTTG CCTGGAGTCA


----------------- Genbank EST clone (AA058363) ----------------

         70         80         90        100        110        120
    *    *    •    *    *    •    •    •    *    •    *    *
GGGGCAAGGA GAACAGGCAG TGGCTGCTAA TGAGTGCAGG GTTTCCTCTC AGGTGATGAA


-------- Genbank EST clone (AA058363) ---------|

        130        140        150        160        170        180
    •    *    *    •    *    •    *    *    *    *    *    *
AATATTCTAA AATTGACTGT AGGAATTAGC TGGGCGCAGT GGCTCACGTC TGTAATCTCA

        190        200        210        220        230        240
    *    *    •    *    *    *    •    *    •    *    •    •
GCACTTTGGG AAAGCCAANG CGGGCGGATC ACCCGAGGCC AGGAGTTCAA NACCATCCTG

        250        260        270        280        290        300
    *    *    •    *    •    •    *    *    *    *    •    *
ACCAACATGG AAGAACCTCC TCTCTACTAA AAATACAAAA TTAGCCGGGT GTGGTGGCAG

        310        320        330        340        350        360
    *    •    *    *    *    *    *    *    *    *    *    *
ACGCCTGTAA TCCCCAGCTA CTTGGGAGGC TGAGGCAGGA GAATCACTTG AACCCGGGAG

        370        380        390        400        410        420
    *    *    •    •    *    *    *    *    •    *    *    •
GCAGAGGTTG CAGTGAGCTG AGATCACAAC ATTGCACTCT TGCCTGGGCA ACAAGAGCGA

        430        440        450        460        470        480
    *    *    *    •    *    *    *    *    *    *    *    *
AACTCCGTCT CAAAAAAAAA AAAAAATTGA CTGTAGGAAT GGTTGTACAA CTCTGTGAAT

        490        500        510        520        530        540
    *    *    *    •    *    *    •    •    •    •    *    *
ATACTAATTT TTAAAACCCC ATTGAATTGT ACACTCTACA TGGGTGAACT GAATAGTATG
```

```
       550       560       570       580       590       600
     *   *     *   *     *   *     *   *     *   *     *   *
CCAATTATAG TGCAATAAAA CTGCTAAGAA GGGCATTTAG AAACAACCGC AACGCTAGAA

       610       620       630       640       650       660
     *   *     *   *     *   *     *   *     *   *     *   *
ACACCAGAAA TCCCTTTTTT TTTTGTCAGG GTCTCGCTCT GTCACCCAGG CTGGAGTGCA

       670       680       690       700       710       720
     *   *     *   *     *   *     *   *     *   *     *   *
GTGGCGTGAT CACAGCTCTC TGCAACCTCC GTCTCCTGGG CTCAAGCAAT TCACCTGCTT

       730       740       750       760       770       780
     *   *     *   *     *   *     *   *     *   *     *   *
CGATCTCCTG AGTAGGGGGG ACCACAGGTT AGCCACCACA CCTCGTTAAT TTTTTCGTAT

       790       800       810       820       830       840
     *   *     *   *     *   *     *   *     *   *     *   *
TTGTTGTTGA GATGGAGTCT CGCCATGTTG CCCAGGCTGG TATGAAGTCC TGGGCTCAAG

       850       860       870       880       890       900
     *   *     *   *     *   *     *   *     *   *     *   *
TGATCCACCT GCCTCAGCCT CCTAAAGTGT GGGGGTAACA GGTGTGGGCC ACTGCACCCG

       910       920       930       940       950       960
     *   *     *   *     *   *     *   *     *   *     *   *
GCCCCCTGCT AATTTTGAAC AAATTTTTTT GTAGGAACAA AGTCTCACTT TATTGCCCAG

       970       980       990      1000      1010      1020
     *   *     *   *     *   *     *   *     *   *     *   *
GATGGTCTTG AAACTCCTGG CTTTAAGCGA TCCTCCTGCC TTGGCCTCCC AAAGTGCTGA

      1030      1040      1050      1060      1070
     *   *     *   *     *   *     *   *     *   *     *
GCATATAGGC GTGAGCCACT GTGCCACGCT TCTTTTTTTT TTAAATAAAT TAG TTG AGT
                                                           Leu Ser>


           | 3'<------------ Genbank EST clone (H71759) ----------- 5'

1080          1090        1100          1110        1120        1130
  *    *    *    *    *    *    *    *    *    *    *    *    *
TTG GCC CTT TGT AAA TAT GTA CCA ATG GGA ATA TAT GAG CTA TTA TTA AAA
                                                Ala Ile Ile Lys>
Leu Ala Leu Cys Lys Tyr Val Pro Met Gly Ile Tyr Glu Leu Leu Leu Lys>
```

```
 3'<---------------- Genbank EST clone (H71759) --------------- 5'

          1140        1150        1160        1170        1180
       *     *     •     •     •     *     •     *     •     *
 CCA GAT ACT GTT ACT TCT CCT TTG CTA TG ACT AAC TTT TTA TTT CTG CAG
Thr Arg Tyr Cys Tyr Phe Ser Phe Ala Met Thr Asn Phe Leu Phe Leu Gln>
 Pro Asp Thr Val Thr Ser Pro Leu Leu>


                                                     |-- 252 bp ORF--

 3'<---------------- Genbank EST clone (H71759) ---------------- 5'

          1190        1200        1210        1220        1230
       *     •     *     *     *     *     *     *     •     •
 TAT ATG TTG CTC TTT GGT TTA GGG GAC TAA CAA CTT TTT AAA AAA ATA AAA
                                                        Lys Asn Lys>
 Tyr Met Leu Leu Phe Gly Leu Gly Asp    Gln Leu Phe Lys Lys Ile Lys>
                                                                Asn>


       |----------- primer H4-3XS1240 ------------→|

 -------------------------- 252bp ORF ---------------------------

 3'<----------------- Genbank EST clone (H71759) ---------------- 5'

        1240        1250        1260        1270        1280
      *     •     *     *     •     *     *     *     •     *
 TGT TAT CTG TGT CTT GCA GGG CTG TTA CAG GGG GTT ACG AGC AAA ATG GCA
Met Leu Ser Val Ser Cys Arg Ala Val Thr Gly Gly Tyr Glu Gln Asn Gly>
 Cys Tyr Leu Cys Leu Ala Gly Leu Leu Gln Gly Val Thr Ser Lys Met Ala>
  Val Ile Cys Val Leu Gln Gly Cys Tyr Arg Gly Leu Arg Ala Lys Trp His>


  Homology to porcine
  N-term. peptide (PN): Lys Gln Val Thr Gln Ser Tyr Glu xxx Leu



 -------------------------- 252bp ORF --------------------------

 3'<----------------- Genbank EST clone (H71759) --------------- 5'

       1290        1300        1310        1320        1330
     *     *     *     *     *     *     •     •     *     *
 CCT GGC CGG GAG GGA GGT GGG GGC TG AAC CAG CCC CAA TTC AAC CAG GAA
Thr Trp Pro Gly Gly Arg Trp Gly Leu Asn Gln Pro Gln Phe Asn Gln Glu>
 Pro Gly Arg Glu Gly Gly Gly Gly>
  Leu Ala Gly Arg Glu Val Gly Ala  Glu Pro Ala Pro Ile Gln Pro Gly>
 ----------------------------- 252bp ORF -----------------------------

 3'<------------------- Genbank EST clone (H71759) -------------- 5'

       1340         1350          1360          1370         1380
   *       *      *      *      ●       *     ●      *       *     *
 AGG GGC ATC TTA GTG TTC ACA CCA AAA CCA CCT TCA ACT GGT TGT GAG CTA
 Arg Gly Ile Leu Val Phe Thr Pro Lys Pro Pro Ser Thr Gly Cys Glu Leu>
Lys Gly His Leu Ser Val His Thr Lys Thr Thr Phe Asn Trp Leu>


 ----------------------------- 252bp ORF ------------------------------

 3'<------------------ Genbank EST clone (H71759) --------------- 5'

      1390         1400          1410          1420         1430
  ●       *      *      *       *      *      *       ●      ●      ●
 GGG GTG CTG CTA TGG CCA GTG GGG GAG GCC TTT AAT TCT AAA TCT TAC GTT
 Gly Val Leu Leu Trp Pro Val Gly Glu Ala Phe Asn Ser Lys Ser Tyr Val>


 ------------------ 252bp ORF ----------------------|----- 81bp ORF ---

 3'<------------------ Genbank EST clone (H71759) ----------------- 5'

     1440           1450          1460          1470          1480
 *       *       ●      ●      *       *      *       ●      *     *     *
 CTG AGT GTG CCT CTT TCT AAT TCA AAG GTA CAA TAT AAA T AGA CTG TGT GAC
 Leu Ser Val Pro Leu Ser Asn Ser Lys Val Gln Tyr Lys>
      Val Cys Leu Phe Leu Ile Gln Arg Tyr Asn Ile Asn  Arg Leu Cys Asp>
                                                      Ile Asp Cys Val Thr>


                              |-------- primer H4-3xS1512 -------→|

 ---------------------------- 81bp ORF ----------------------------

 3'<------------------- Genbank EST clone (H71759) -------------- 5'

   1490         1500          1510          1520          1530
     *      *       *       *      *       *       *      *      *      *
 CTG CAT GAA AAC ATT CCT GTT TGC AAT GGC ACA TCT CAC AGA AAG TTA CCA
            Lys His Ser Cys Leu Gln Trp His Ile Ser Gln Lys Val Thr>
 Leu His Glu Asn Ile Pro Val Cys Asn Gly Thr Ser His Arg Lys Leu Pro>
  Cys Met Lys Thr Phe Leu Phe Ala Met Ala His Leu Thr Glu Ser Tyr His>


            Homology to porcine
            N-terminal peptide (PN) : Lys Gln Val Thr Gln Ser Tyr Glu
 ----- 81bp ORF -----|

 3'<------------------- Genbank EST clone (H71759) --------------- 5'

1540        1550        1560        1570        1580
  *     *     *     *     *     *     *     *     *     *
 TTT TTC AAA CCA CCT ATA ATG AGA ATG GAT GTC AAA ATG GGA TTC TA ACA
Ile Phe Gln Thr Thr Tyr Asn Glu Asn Gly Cys Gln Asn Gly Ile Leu Thr>
 Phe Phe Lys Pro Pro Ile Met Arg Met Asp Val Lys Met Gly Phe>
  Phe Ser Asn His Leu>


 xxx Leu  (porcine N-term. peptide)


 3'<------------------- Genbank EST clone (H71759) --------------- 5'

1590        1600        1610        1620        1630
  *     *     *     *     *     *     *     *     *     *
 TTT TGT AAG ACC CTA AAT TTA TCA AAT AAA GAG TCA AAT ATG TGC AAT GAA
 Phe Cys Lys Thr Leu Asn Leu Ser Asn Lys Glu Ser Asn Met Cys Asn Glu>


 3'<--------------------- Genbank EST clone (H71759) -------------- 5'

1640        1650        1660        1670        1680
  *     *     *     *     *     *     *     *     *     *
 CTC TTC CTG AGA CAG GAA GAA GGG GCA CTT TTT AGA AAG ATA TTC AAC AGA
 Leu Phe Leu Arg Gln Glu Glu Gly Ala Leu Phe Arg Lys Ile Phe Asn Arg>


 3'<------------ 5'|

1690        1700        1710        1720        1730        1740        1750
  *     *     *     *     *     *     *     *     *     *     *     *     *
 AGG GTT TTC TA GTTGACTGAC TATCAAGGTA GAGAAAACCC TATAACTCCA TTCATTAAAA
 Arg Val Phe>


      1760        1770        1780        1790        1800        1810
    *     *     *     *     *     *     *     *     *     *     *     *
 TATGACTATT AATACCTTAC TAAATATCCT TCTACATCCT TACCCACTTT TTTATACACA


      1820        1830        1840        1850        1860        1870
    *     *     *     *     *     *     *     *     *     *     *     *
 TGCATACACA TGCTTTTAAC ACGTGTTTTA AATAACAAAA ATGNAGATTC CACTAAATCA


      1880        1890        1900        1910        1920        1930
    *     *     *     *     *     *     *     *     *     *     *     *
 ACTATTTTAA CCTGTCCCAC TCACTCCCCT CTTTATTATA CTTTTTGTGT CAAGTTGATT
```

```
      1940       1950       1960       1970       1980       1990
    *    ●     ●    ●     ●    ●     ●    ●     ●    ●     ●    *
TATTTATTGA AATAAATAAA ATCCAGACTT TTCTGTTTAA GTGCCTGAAT TTGTTCTTAC

      2000       2010       2020       2030       2040       2050
    *    ●     ●    *     *    *     ●    *     ●    ●     *    *
ACAAAGCAAG TGCTCTCAAG AAGCTCATTC TTGAAGATGG TTGAGAAATA TCAATGACCA

      2060       2070       2080       2090       2100       2110
    *    *     *    *     *    ●     ●    ●     *    *     ●    ●
ATATCAAAAG AACATTAACC CATGAAATGT ACATGAATAT GGCCACGCGT GGTGGCTCAC

      2120       2130       2140       2150       2160       2170
    *    *     *    *     *    ●     ●    ●     *    *     *    *
GCTTGTAATC CCAGTACTTT GGGAGGCTGA AGTGGGTGGA TCACCTGAGG TCAGGAGTTC

      2180       2190       2200       2210       2220       2230
    ●    *     *    *     ●    ●     *    *     *    *     *    *
AAGACCAGCC TGACCAACAT GGTGAAAACC CCATCTCTAC TAAAAATACA AAACTTAGCC

      2240       2250       2260       2270       2280       2290
    *    *     *    *     *    *     *    *     *    *     *    *
AGGCATGGTG GTGTATGCAT GTAGTCCCAG CTACTTGGGA GGCTGAGGTA GGAGAACAGC

      2300       2310       2320       2330       2340       2350
    *    ●     *    *     *    *     ●    ●     ●    *     *    *
TTGAACTCAG AGGCGGGGGT GGAGGTCGCA GTGATCCGAG ACCGTACCAT TGCACTCCAA

      2360       2370       2380       2390       2400       2410
    *    *     ●    ●     ●    *     ●    *     ●    ●     ●    ●
CCTGGGCGAC AGAGTGAGAC TCTGTCTCAA ACAATGACAA CAACAACAAA AATTTACATG

      2420       2430       2440       2450       2460       2470
    ●    *     ●    ●     ●    *     ●    *     ●    ●     *    ●
AATATGCATA TCAGCTGAAG GCAAGCACAT TACAAATGTA AGTACAATTT AGTAGAAGGG

      2480       2490       2500       2510       2520       2530
    ●    *     ●    *     *    ●     *    ●     ●    ●     ●    ●
GTTGAGGGAG CCTGGGTGTC AATGGGGAAA GCCTTGGTGA GCTCCTAGAA GTAGGATGAA

      2540       2550       2560       2570       2580       2590
    *    ●     ●    *     *    ●     *    *     ●    ●     ●    *
GGCTGTGAGA CAGTCTCAGT CCTGGTGATA CAGTCAGGAT GATGTAACAG GGTTCAGTGA
```

```
      2600       2610       2620       2630       2640       2650
     *    ●     ●    *     ●    *     ●    ●     ●    ●     *    *
GGGAGTGCAA GGTGACAGAA CAATATGGAG TATCTTATAC TAATTCCTGC AAACTTCTCA

      2660       2670       2680       2690       2700       2710
     *    *     ●    *     ●    ●     *    *     *    *     ●    *
ACTATGTACA AGAAAACAAA TACTTACAAA ACAGAAGAAA ATGGCCAGGT ATAGTGGCTC

      2720       2730       2740       2750       2760       2770
     ●    *     ●    *     *    *     ●    ●     ●    *     *    ●
ATACCTGTAA TCCCAGCACT CTGGGAGACC GAAGCAGTAT CACCTGAGGT CAGGAGTTCG

      2780       2790       2800       2810       2820       2830
     *    *     *    *     *    *     *    *     *    *     *    *
AGACCAGCCT GGCCAACATG GTGAAACCCT ATCTCTACTA AAAATACAAA AATTAGCCGG

      2840       2850       2860       2870       2880       2890
     ●    *     ●    *     ●    ●     ●    ●     *    *     *    *
GCATGGTGGT GGGCGCCTGT AATCCCATCT ACTCGGGAGG CTGGGGCAGA AGAACCGCTT

      2900       2910       2920       2930       2940       2950
     *    ●     ●    ●     *    ●     ●    ●     *    *     *    *
GAACCTAGGG TGGGGTGGAG GTTGCAGTGA GCTGAGATCG AGCCATTGCA CTCCAACCTG

      2960       2970       2980       2990       3000       3010
     ●    *     *    *     *    ●     *    *     *    *     *    *
GGCACTAAGA GTGAGACTCC ATCTCAAATT TAAAAAAAAG AAAAAAAAAG AAATAGAACA

      3020       3030       3040       3050       3060       3070
     *    ●     ●    *     *    *     ●    ●     ●    *     *    *
AAACACAGCA GTGGATACAG TGCTTTTTGG TTCATTCAAA GCAGTCCCAT CAGAATTAGC

      3080       3090       3100       3110       3120
     *    ●     *    ●     *    *     *    *     *    ●     *
ATATTAAACT ATACAAACGG CTCCAAATGG CAAAATCTAG AACCCGCCGC CACCG
                                      Xba I
```

terminal peptide sequence was found between bp 1514-1543. Primers were made from these two regions but RT-PCR failed to prove the sequence to be exonic.

ii) **Sequence contained in Genbank EST clones**

A Genbank EST database search revealed that a EST clone (H71759), 440bp in length, contained the sequences between bp1087-1704 (Fig. 15). Another EST clone (AA058363), 790bp in length, contained sequences at the beginning of the clone and sequences from the next genomic fragment upstream. Regardless of the presence of such sequences in the EST clones and the homology to the porcine sequence in certain areas in this fragment, RT-PCR failed to prove the sequence to be a MTHFR exon.

4.3.4 Sequence of the H4_06 genomic fragment

The H4_06 is the next genomic fragment 5' to the genomic fragment H4_3X. The sequencing of this fragment revealed the following:

i) **ORF with ATG codon (Met)**

An ORF of 121bp was found between bp 57-179 of this clone (Fig. 16). An ATG codon with 72% homology to the Kozak ATG start site consensus sequence was found at bp99. RT-PCR failed to prove the sequence to be exonic.

ii) **ORF with homology to porcine sequence**

An ORF of 141bp was found between bp 293-432 of this clone (Fig. 16). A 40% identity to the porcine N-terminal peptide sequence was found between bp 296-325. Two ATG codons with strong homology to the Kozak start site consensus sequence were

Figure 16. The DNA sequence of the H4_06 genomic fragment in the 5' region of human MTHFR gene.

Sequence of a 600 bp Hind III-Xba I genomic fragment subcloned from The PAC clone. The dotted line indicates that a 780 bp Genbank human EST clone (AA058363), transcribed in the opposite direction to MTHFR, is found to contain the sequence of this entire genomic fragment with the ends extending into the flanking genomic fragments. All the open reading frames (ORFs) that encode for >25 amino acids are shown. ATG codon (Met) with a homology to the Kozak translation start site consensus sequence (G/A-C-C-A-T-G-G) is underlined and the ORFs that contain the ATG codon are shown in bold. A homology to the porcine N-terminal peptide (PN) is identified at bp 295-325 and the identical amino acids are underlined and shown in bold. Primers designed to test the authenticity of the possible ORF's are shown as dotted lines.

```
 3'<----- Genbank EST clone (AA058363) continues 200 upstream ------ 5'

  Hind III 10         20         30         40          50
     ↓ *    *      *      *     *      *     *      *      *     *
 ATAA GCT TCC ATA AAA CTC CAA AAT GAC AAG GTT CAG AGA GCT TCC AGA CAG
      Ala Ser Ile Lys Leu Gln Asn Asp Lys Val Gln Arg Ala Ser Arg Gln>


 3'<-------------------- EST clone (AA058363) -------------------- 5'

          60         70         80          90         100
   *      *      *      *      *      *      *      *      *      *
 CTG AAC CGT TTG GAG GTT CCT GGA GGG T AGC GTG CCT GGA GAG GGC ATG AAA
 Leu Asn Arg Leu Glu Val Pro Gly Gly      Arg Ala Trp Arg Gly His Glu>
     Thr Val Trp Arg Phe Leu Glu Gly  Ser Val Pro Gly Glu Gly Met Lys>


 3'<-------------------- EST clone (AA058363) ------------------- 5'

                               |------ primer H4_06S194 ----→|  |--------

      110        120        130        140         150
  *     *      *      *      *      *      *      *      *      *      *
 GCT CCG TGC CCC CTC CCA CGT TCC TTG AAT CTC TCC ATC TGG CTG TTC TTC
                                         Ile Ser Pro Ser Gly Cys Ser Ser>
Ser Ser Val Pro Pro Pro Thr Phe Leu Glu Ser Leu His Leu Ala Val Leu>
 Ala Pro Cys Pro Leu Pro Arg Ser Leu Asn Leu Ser Ile Trp Leu Phe Phe>


 3'<-------------------- EST clone (AA058363) ------------------- 5'

 --- primer H4_06S221 --→|

     160        170        180         190        200
     *      *      *      *      *      *      *      *      *      *
 TGT ATC CTT TAT GAT ATC CTT TAT T AAT AAA CTA GCA AAT GTT AAT TAC ATT
  Val Ser Phe Met Ile Ser Phe Ile  Asn Lys Leu Ala Asn Val Asn Tyr Ile>
Leu Tyr Pro Leu>
 Cys Ile Leu Tyr Asp Ile Leu Tyr>
```

```
3'<--------------------- EST clone (AA058363) ---------------- 5'

210           220        230        240           250           260
  *      *      *     *    *     *    *      ●      *      *     ●
TCT GTG TTT TCC T GACCCGCTCT AGCAAGTTAA CCA AAC TCA AGG AGG GAG TCC
Ser Val Phe Ser                         Pro Asn Ser Arg Arg Glu Ser>


                                        |----- primer H4_06S365-----

3'<--------------------- EST clone (AA058363) ----------------- 5'

                                        |------- 140 bp ORF --------

        270           280          290          300           310
    *     ●      ●      ●     *      ●      *     *      *      *
TGT GAA CCT TCA ATT CAC TGC CAG CTG CTG AGA AAT GCA GGT GAC AAC CTA
Cys Glu Pro Ser Ile His Cys Gln Leu Leu Arg Asn Ala Gly Asp Asn Leu>
                                         Glu Met Gln Val Thr Thr Tyr>

Homology to porcine N-terminal peptide(PN):  Lys Gln Val Thr Gln Ser


--->|                                                   |-- primer--

3'<--------------------- EST clone (AA058363) ----------------- 5'

-------------------------- 140 bp ORF --------------------------

      320           330           340          350          360
  *      *      *     *      *      *     *      *       *     *
CTA CTT TTT TTT TTT GAG ATG GGG TCT TTC TCT GTC ACC CAG CAT GGA GTG
Leu Leu Phe Phe Phe Glu Met Gly Ser Phe Ser Val Thr Gln His Gly Val>
 Tyr Phe Phe Phe Leu Arg Trp Gly Leu Ser Leu Ser Pro Ser Met Glu Cys>

 Tyr Glu xxx Leu (porcine peptide, PN)


--- H4_06S427 --->|

3'<--------------------- EST clone (AA058363) ------------------ 5'

-------------------------- 140 bp ORF --------------------------

     370           380          390           400          410
 ●      *     *      *      *     *      *      *     ●      ●
CAG TGG CAT AAT CTC GGT TCA CTG CAA CCT CCA ACT CCT GGG CTT AAG CCA
Gln Trp His Asn Leu Gly Ser Leu Gln Pro Pro Thr Pro Gly Leu Lys Pro>
 Ser Gly Ile Ile Ser Val His Cys Asn Leu Gln Leu Leu Gly Leu Ser His>
 3'<---------------------- EST clone (AA058363) ------------------ 5'

 ----- 140 bp ORF ------|

     420           430           440          450           460
 *     •      *      *     *      •      •      *     •      *     *
 TCC TCC CAC CTC AGC CTC CTG AGT AGT GGG GAA ACA ACC TCC ACA TAT CGG
 Ser Ser His Leu Ser Leu Leu Ser Ser Gly Glu Thr Thr Ser Thr Tyr Arg>
  Pro Pro Thr Ser Ala Ser     Val Val Gly Lys Gln Pro Pro His Ile Gly>
                                    Trp Gly Asn Asn Leu His Ile Ser>



 3'<---------------------- EST clone (AA058363) ------------------- 5'

     470          480           490            500          510
       *     •      *      •     •      *      *     *      *      *
 TGT T AGC AGT GTT CCG CGA GTG GTG CCT GAG AGC AGA GGA GAA ACA ATA CCC
 Cys                                        Glu Gln Arg Arg Asn Asn Thr>
  Val  Ser Ser Val Pro Arg Val Val Pro Glu Ser Arg Gly Glu Thr Ile Pro>
Val Leu Ala Val Phe Arg Glu Trp Cys Leu Arg Ala Glu Glu Lys Gln Tyr Pro>



 3'<----------------------- EST clone (AA058363) ------------------ 5'

 520           530           540           550           560
   •      *      •     *      *      *      *      •      *     *
 TTG CCT TCA TCC AAA ATA TGT ATG GAA ATT ACT GGT CAA GTA CTT CAT T GAT
Leu Ala Phe Ile Gln Asn Met Tyr Gly Asn Tyr Trp Ser Ser Thr Ser Leu Met>
 Leu Pro Ser Ser Lys Ile Cys Met Glu Ile Thr Gly Gln Val Leu His>
  Cys Leu His Pro Lys Tyr Val Trp Lys Leu Leu Val Lys Tyr Phe Ile  Asp>


 3'<---- EST clone (AA058363) continues 150 bp downstream ------ 5'

570           580           590           600           610        620
 *     •       •     *      *      *      •      *      *     *     •
 GGT TCC ATC T AAT AAT GAT ATG GTT TCC GTA AAC ATT TAAG AAGGTTTTTA
  Val Pro Ser  Asn Asn Asp Met Val Ser Val Asn Ile>
 Gly Ser Ile>


3'<----- 5'
    Xba I
      ↓ 630
    •    •
 GATATCTAGA
```

Figure 17. Partial DNA sequence of the PAC-3K genomic fragment in the most 5' characterized region of the human MTHFR gene.

Sequence of a 3.5kb genomic fragment generated by PCR using the PAC clone as DNA template. The antisense primer bound specifically at the 3' end and non-specifically at the 5' end, as shown on the sequence. The 5' end of this genomic fragment contains exon 4 of the ClC-6 gene, which is again transcribed on the opposite strand and opposite direction. The consensus sequence of the possible splice sites for this exon in the 3' end of this fragment contains some sequence of the genomic subclone H4_06 and some extra sequences of the Genbank EST clone (AA058363). Final sequencing of this genomic subclone is in progress.

```
|-- primer H4_06A180 --→|              |3'←ClC-6 gene (exon 4)--5'

         10         20         30         40         50         60
    *    ●    *    *    ●    ●    *    *    ●    ●    *    ●
GAGCTTTCAT GCCCTCTCCA GGCCCTGCTC CTCACCAGGC CAGTGCAGAC TCCAATGGCA
CTCGAAAGTA CGGGAGAGGT CCGGGACGAG GAGTGGTCCG GTCACGTCTG AGGTTACCGT

          possible splice donor site ↑


3'←---------- ClC-6 gene (exon 4) ------ 5'|

         70         80         90        100        110        120
    ●    ●    *    *    *    *    ●    ●    *    *    *    *
AACACCACCA TCCACTTCAC CGCCTCATAT CTTCGACCTT CTGTTCACAC AGGAAAAAGA
TTGTGGTGGT AGGTGAAGTG GCGGAGTATA GAAGCTGGAA GACAAGTGTG TCCTTTTTCT

               possible splice acceptor site ↑


        130        140        150
    *    ●    *    *    *    ●    ●
ATGATTCAGT TTACAGGGTT ATGGTGGCAG AGGGG ------- /~3.5 kb/--------
TACTAAGTCA AATGTCCCAA TACCACCGTC TCCCC ------------------------


←---- Genbank EST clone (AA058363) continues 140bp upstream-----

    *    ●    *    *    *    *    ●    *    *    *    *    *
GCTGAAGGTT AAGCTGATCA CCAAGGGCTG GTAGTGTAAT CAATCATGAC TGCATAAAAA
CGACTTCCAA TTCGACTAGT GGTTCCCGAC CATCACATTA GTTAGTACTG ACGTATTTTT


 ------- Genbank EST clone (AA058363) continues in H4_06 ------→

 |--------------- H4_06 starts here --------------------------→|
Hind III
↓   *    ●    *    *    *    ●    ●    ●    *    ●    *    *
AGCTTCCATA AAACTCCAAA ATGACAAGGT TCAGAGAGCT TCCAGACAGC TGAACCGTTT
TCGAAGGTAT TTTGAGGTTT TACTGTTCCA AGTCTCTCGA AGGTCTGTCG ACTTGGCAAA


----------- part of H4_06 subclone -----------→

    ●    ●    *    *    *    ●    ●    *    *
GGAGGTTCCT GGGAAGGTAG CTGCCTGGAG AGGGCATGAA AGCTC
CCTCCAAGGA CCCTTCCATC GACGGACCTC TCCCGTACTT TCGAG

                      |←-primer H4_06A180-----|
```

identified, one at the beginning of the ORF at bp296, another downstream in the ORF at bp356. RT-PCR was unable to prove the sequence to be exonic.

iii) **Sequence contained in Genebank EST clone**

An EST clone (Genebank AA058363) was found to contain the entire H4_06 fragment with the ends extending in both directions to the flanking genomic fragment. However, it has not been proven that the EST belongs to MTHFR.

4.3.5 The sequence of the PAC_3K genomic fragment

This is the next genomic fragment 5' to the H4_06 genomic fragment (Fig. 17) (Note that the first and the last 23 base pairs are identical because this fragment was generated by PCR in which the anti-sense primer bound non-specifically at the 5' end.) Final sequencing of this clone is in progress, but preliminary result revealed the following:

i) **Sequence contained in Genebank EST clone**

The 3' end of this genomic fragment contains some EST sequence which has most of the sequence in the 3' genomic fragment H4_06.

ii) **Overlapping gene (ClC-6- exon 4)**

The 5' end of this clone contains exon 4 of the human ClC-6 gene with the ORF in the opposite strand and in the opposite direction consistent with the previous findings in other genomic fragments.

4.4 Cloning porcine cDNA

Porcine liver, like many human tissues, was shown to have the 77kDa isoform of MTHFR (Frosst 1995). The porcine internal peptide sequences show strong (90%) identity to the human sequence suggesting that the N-terminal sequence between the two species may also be homologous. Therefore, a great deal of homology searches for the porcine N-terminal peptide sequence were performed on the human genomic sequence. However, the assumption that the porcine sequence is highly homologous between the two species at the N-terminus was never tested experimentally. An attempt to compare the cDNA sequence between the two species close to the 5' end led to the cloning of the porcine cDNA, depicted in Fig 18 and 19.

A degenerate oligo designed from the porcine N-terminal peptide sequence and a non-degenerate primer from a 90 bp porcine cDNA sequence (identified in Goyette et al. 1994) were used to amplify porcine cDNA by RT-PCR using porcine liver RNA. A 600bp PCR fragment was generated. Sequencing of this fragment showed that the degenerate oligo did not bind while the anti-sense primer bound at both the 5' and 3' ends. Regardless, this cDNA does contain authentic porcine MTHFR cDNA based on the close amino acid sequence homology between the human and pig (Fig. 18).

The sequence of this porcine cDNA showed 84% amino acid sequence identity with the human sequence at exon 1, starting with the original ATG (Fig. 18). An ATG codon (met) with a perfect Kozak's start site consensus sequence was found in the same location as the original in the human. The 5' extension of this porcine cDNA contains sequences that resemble the human exon 43S, with 60% DNA sequence homology (Fig. 19). Although the human exon 43S does not contain an ORF that continues with the rest

Figure 18. The DNA sequence of the porcine cDNA clone P600.

Sequence of a 600 bp cDNA clone generated by RT-PCR using porcine liver RNA. The anti-sense primer bound specifically at the 3' end and non-specifically at the 5' end. The porcine cDNA contains the counterpart of human exon 1 starting at the original ATG, sharing 84% amino acid identity and 90% similarity. The human exon 1 amino acid sequence is also shown. There is a 60% DNA homology between human exon 43S and this porcine cDNA at bp 266 – 367 (identical sequences are underlined), but the ORF at this region of the porcine cDNA is not found in human exon 43S.

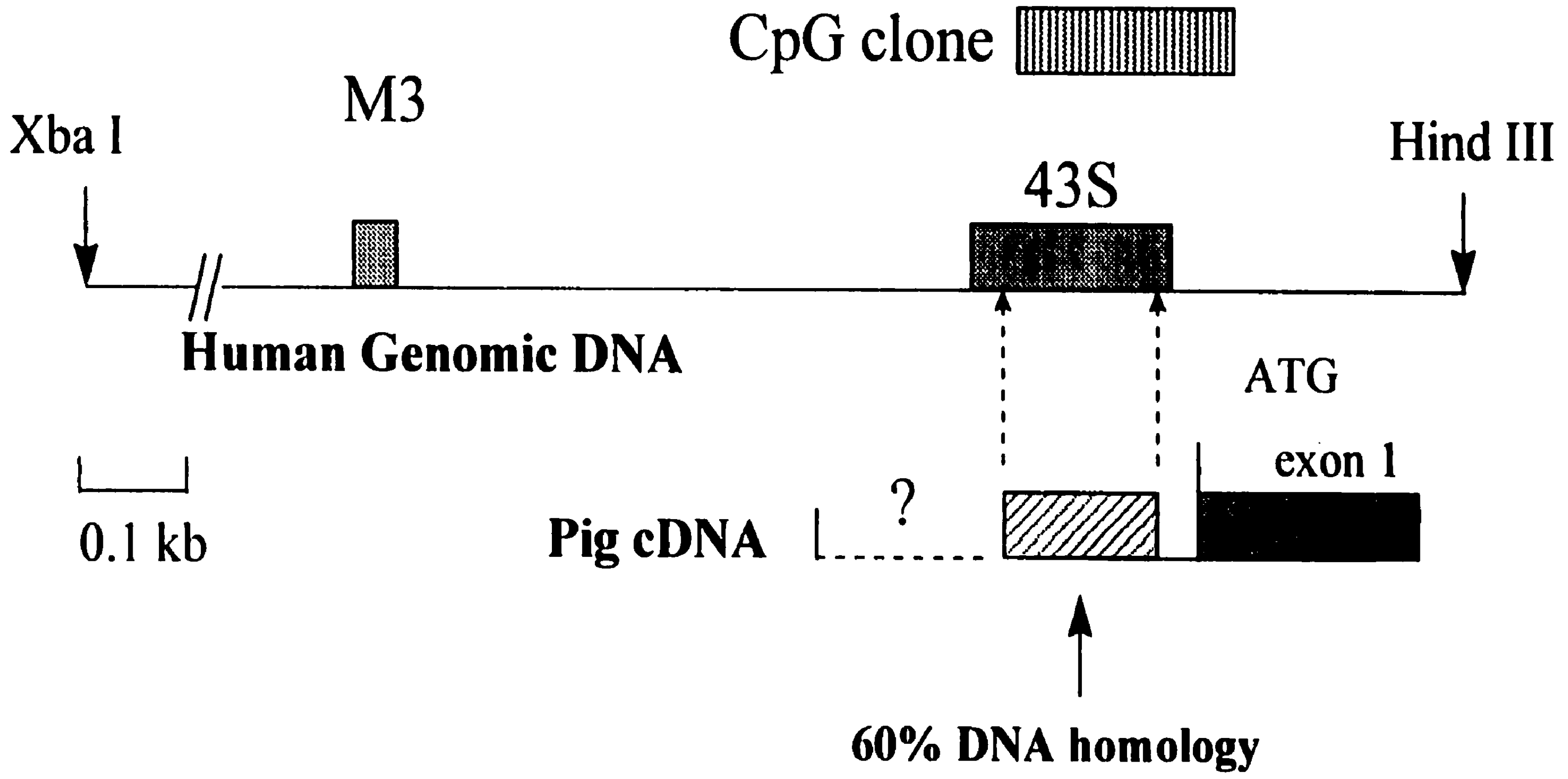
```
        10         20         30         40         50         60
    *    *     •    *     •    •     •    •     •    •     •    •
GGCTTAAACC TAGAGATGAG ATTGACGGCC CCTCCTCGCT CGTCCTCCGA GAGAGTAATG

        70         80         90        100        110        120
    *    *     *    *     •    •     *    •     •    •     •    •
TAATCCTTAA TGGCTGCCTC CTGCACGACC CCCAGCAGCG AAGAGACTCA CTTTCCTCCC

       130        140          150           160          170
    •    •     *    •      •     •      *      *     •      •
CGGCCTGGTC GGAGGCGGCC CTGA GTC GGG GTG TCC GTG CCG GAC TGG GCC TGG
                           Val Gly Val Ser Val Pro Asp Trp Ala Trp>

    180           190          200          210           220
•     •      *      *     •      •      *     *      *      *     •
GCT TCA TTT CTT CAC CCC GAC CAT GAG CCG GTG ACC CGT CGA ATC GAT CCA
Ala Ser Phe Leu His Pro Asp His Glu Pro Val Thr Arg Arg Ile Asp Pro>

   230          240           250          260          270
     *      *     *      *      *     *      *      *     *      *
CTG CCC CCG AGC CAC CCG GGC GCC AAA GAC TGT TCG CCG GTT CCC CGC CGA
Leu Pro Pro Ser His Pro Gly Ala Lys Asp Cys Ser Pro Val Pro Arg Arg>

  280          290          300           310          320
    *     *      *      *     *      *      *     *      *      *
GCG CCT GGA GAA GAG CGG TGG TCG CTG GCC CTG TTT CCG GCG TCC ACT GCA
Ala Pro Gly Glu Glu Arg Trp Ser Leu Ala Leu Phe Pro Ala Ser Thr Ala>

330           340          350          360           370
  •      *      *     *      *      •     *      •      *     •
TCG GGC TGC GCA CGG GTG CCC ACC CGA CCT TTC TGG GAG TGG TGG CCT CCG
Ser Gly Cys Ala Arg Val Pro Thr Arg Pro Phe Trp Glu Trp Trp Pro Pro>

                                                        Start codon
                                                                 ↓
380         390           400          410          420
 *      •     •      *      •     •      *      •     *      •
TAT TTC CCC ACT GCC CCG TGC TCG GCA CTG ATT AAC AGG AGC TCA GCC ATG
Tyr Phe Pro Thr Ala Pro Cys Ser Ala Leu Ile Asn Arg Ser Ser Ala Met>
```

Amino Acid sequence of Human MTHFR exon 1: **Asn Arg** Asn Pro **Ala Met**→

```
430         440          450           460          470          480
*      •      •      *     •      •      •     •      •      •    •
GTG AAC GAA GCC AGA GGG AAC GGC GGC CCC GGC CCC CGC TGT GAG GGC AGC
Val Asn Glu Ala Arg Gly Asn Gly Gly Pro Gly Pro Arg Cys Glu Gly Ser>

Val Asn Glu Ala Arg Gly Asn Ser Ser Leu Asn Pro Cys Leu Glu Gly Ser→


           490          500          510           520          530
      *      •     •      •      •     *      •      •    •      •
AGC AGT GGC AGC GAG AGC TCC AAG GAG AGC TCA AGG TGC TCT ACC ACG GGC
Ser Ser Gly Ser Glu Ser Ser Lys Glu Ser Ser Arg Cys Ser Thr Thr Gly>

Ala Ser Gly Ser Glu Ser Ser Lys Glu Ser Ser Arg Cys Ser Thr Pro Gly→


         540          550           560          570          580
    *      •     •      *     *       •      •    •      *      •
CTG GAC CCC GAG CGT CAC GAG AGG CTC AGG GAG AAG ATG AAG CGG AGG ATG
Leu Asp Pro Glu Arg His Glu Arg Leu Arg Glu Lys Met Lys Arg Arg Met>

Leu Asp Pro Glu Arg His Glu Arg Leu Arg Glu Lys Met Arg Arg Arg Leu→


        590          600          610          620          630
  •       *      •    •       •      •     •     •      •     *
GAG TCA GGT GAC AAG TGG TTC TCC CTA GAA TTC TTC CCT CCT CGA ACT GCT
Glu Ser Gly Asp Lys Trp Phe Ser Leu Glu Phe Phe Pro Pro Arg Thr Ala>

Glu Ser Gly Asp Lys Trp Phe Ser Leu Glu Phe Phe Pro Pro Arg Thr Ala→


       640          650          660
 *       •     *     •      •      •     •
CAG GGC GCC GTC AAT CTC ATC TCT AGG TTT AAG CC
Gln Gly Ala Val Asn Leu Ile Ser Arg Phe Lys>

Glu Gly Ala Val Asn Leu Ile Ser Arg  (Human MTHFR exon 1)
```

Figure 19. The sequence homology between the porcine cDNA, the human CpG DNA clone and the human genomic fragment.

The first 2kb of the 4kb Xba I genomic fragment, the relative location of the CpG DNA clone and the porcine cDNA are shown. The human exons 43S and M3 are shown as filled boxes. The porcine cDNA contains the counterpart of human exon 1 and shares an amino acid identity of 84%. The hatched box represents the presence of a 60% DNA sequence homology between the porcine cDNA and human genomic sequence. The ORF found in the porcine cDNA is not found in human exon 43S and the human exon 43S does not have an ORF that is contiguous with the original ATG in exon 1. The human CpG DNA clone falls into the region of DNA homology between the two species. The dotted line indicates that the identity of the sequence is unknown.


CpG clone
M3
Xba I
43S
Hind III
Human Genomic DNA
ATG
exon 1
0.1 kb
?
Pig cDNA
60% DNA homology

of the human cDNA, it is interesting that the porcine cDNA does contain an ORF that continues 280bp upstream of the ATG codon. But the porcine N-terminal and the other internal peptide sequences were not identified in this porcine cDNA. The porcine ORF extension does not contain another ATG codon upstream. The location of human exon 43S falls into a possible CpG island suggesting that perhaps the 5' extension of the porcine cDNA is also part of a CpG island, which is a common location for the beginning of transcription in mammalian genomes (Cross et al 1994).

4.5 SUMMARY OF RESULTS

The cloning of various cDNAs has revealed that human MTHFR has four possible cDNA 5' ends. Sequencing of the cDNA clones and genomic fragments showed that the original exon 1 extends into a 5' UTR (called MRE-5') of about 750bp without intervention by an intron. Three other 5' exons are alternatively spliced into a common splice acceptor site 13bp 5' to the original ATG codon of the human cDNA. Within this MRE-5' cDNA extension is another 5' exon that is alternatively spliced into the common splice acceptor site generating another 5' cDNA extension. This exon, MRC-5', located 280bp upstream of the original ATG codon, has an ORF of at least 160bp that continues with the rest of the human cDNA and another upstream ATG codon. M3 is another 5' exon that is alternatively spliced into the common splice site. Located 2.7kb upstream of the original ATG, M3 also has an ORF of at least 120bp that continues with the rest of the cDNA, but an ATG codon was not found in this ORF. The 43S exon, another alternatively spliced 5' exon, has no ORF that continues with the rest of the cDNA. A porcine cDNA was cloned with 84% amino acid sequence identity to the human sequence starting at the original ATG codon of the human. The same ATG codon was found in the porcine cDNA. The 5' end of the porcine cDNA has a 60% DNA sequence identity with human exon 43S. Although the porcine cDNA has 280bp ORF upstream of the ATG codon, the same ORF was not found in human exon 43S. A CpG island was identified in the region of the 43S exon, suggesting that the transcription start site and a promoter may be near.

An overlapping gene, a putative chloride ion channel gene (ClC-6) was found to be transcribed on the opposite strand and the opposite direction to the MTHFR gene. The significance of the overlapping gene is not yet known but co-regulation of gene expression is possible. Two EST clones were found to contain sequences from two human genomic fragments. Since these two EST sequences are transcribed in the opposite direction to MTHFR, they are likely to be unrelated to MTHFR.

5. Discussion

The cloning of the human MTHFR cDNA and the analysis of the gene structure enabled the study of MTHFR deficiency in homocystinuria, mild hyperhomocysteinemia and multifactorial diseases at the molecular level. The characterization of the 5' region of the human MTHFR gene is essential to understand the alternative splicing event at the 5' end as well as to identify the missing coding sequence, the transcription start site and the promoter. The identification of the regulatory elements would facilitate the study of gene regulation which may be crucial in the MTHFR related biochemical pathways.

Gene expression can be regulated at the transcriptional or translational level. Regulation of translation allows the cell to respond more rapidly than regulation of transcription. A 5'UTR of mRNA that forms stem-and-loop secondary structure between the 5' cap and the AUG codon, depending on the strength and position can impair translation by preventing the ribosome from reaching the translation initiation codon (Kozak 1991). Alternative splicing at the 5' end is a common phenomenon in the human genome. Transcripts with various 5'UTRs can be generated by alternative splicing or alternate promoters. For instance, the human Ataxia-telangiectasia gene (*ATM*), with 66 exons and spanning 150kb, is known to have 12 different 5'UTRs with various lengths and sequences generated by alternative splicing (Savitsky et al 1997). These 5' UTRs were determined to form different secondary structures and hence subject to complex post-transcriptional regulation.

The cloning of various human MTHFR cDNAs and the analysis of the 5' genomic fragments reveal a complex alternative splicing event at the 5' end of the cDNA, suggesting that post-transcriptional regulation may be involved. Since the transcription start site is still not defined, it is not known if more than one promoter is used to generate these 5' splice variants. A total of four 5' exons were localized to a 4kb genomic fragment. The original exon1 was found to extend upstream into a 800bp 5' UTR (MRE-5'), and three other 5' exons were found to be alternatively spliced into a common splice acceptor site generating cDNAs with 4 possible 5' ends. The 5' boundaries of these exons have not yet been defined by classical methods, such as primer extension and S1 nuclease mapping, and hence should not be considered as the exon/intron boundary.

The identification of these 5' exons suggests that the exon numbers for the human MTHFR gene may have to be modified. However, since the exon number for the original cDNA is already well established, it would be more convenient to rename the 5' exons. Since MRE-5' is in fact an extension of exon 1, the entire exon including the original exon 1 and the extension could be called exon 1a. The MRC-5' exon is located within the MRE-5' exon and could therefore be named as exon 1b. The 43S exon is located more upstream and could be called as exon 1c. The M3 exon is the most 5' exon of the human MTHFR identified so far and can be called as exon 1d.

The missing coding sequence has not been isolated although two 5' exons remain as strong candidates. Since the two isoforms of MTHFR have a difference of 7kDa, the missing coding sequence is calculated to be approximately 200bp. The MRC-5' and M3 are two 5' exons that can generate an extended open reading frame (ORF) upstream to the MTHFR cDNA when alternatively spliced into the common splice acceptor site. The extended ORFs for M3 and MRC-5' are at least 160bp and 120bp, respectively. While M3 has an ORF that ends with another possible splice acceptor site at the 5' end, MRC-5' has an ORF that contains a potential translation start site. In addition, the MRC-5' exon and the potential translation start site exist in both human and mouse, further supporting that the hypothesis that this sequence may be coding. Both exons could be part of a larger coding sequence if one or more exons are spliced into the 5' end of these two exons. In the case of MRC-5', the translation start site might be used in generating the larger 77kDa isoform of the MTHFR protein. On the other hand, the conserved ATG codon in the MRC-5' exon could simply encode for an internal methionine (Met) (assuming that the exon is coding) rather than being a translation start site because the sequence homology between the human and mouse continues upstream to this ATG codon. However, it is not uncommon to have sequence homology across species in the 5' UTR. For instance, in the human and mouse HEXA gene, DNA sequence homology was found between the cap site and the initiation codon (Wakamatsu et al 1994).

The missing coding sequences have been hypothesized to be located at the 5' end of the cDNA because the porcine N-terminal peptide sequence has not been identified in the human cDNA. One approach to isolate this coding exon(s) was to search for the porcine N-terminal peptide sequence in the human genomic sequence. However, both the MRC-5' and the M3 exons do not contain any sequences homologous to the porcine N-terminal peptide. If the N-terminal peptide sequence between the human and pig are indeed conserved, then the MRC-5' and M3 are probably only part of a larger coding sequence and more 5' sequence would have to be spliced into one of these two candidate exons.

Sequence analysis of about 10kb of human genomic sequence resulted in several candidate sequences that might encode the sequence equivalent to the porcine N-terminal peptide (Table 1). Some sequences were found to contain up to 50% homology to the porcine sequence and an ATG codon with a strong homology to the Kozak consensus sequence for the translation start site (Kozak 1996) in the open reading frame. One of these sequences was even present in a Genbank EST clone, suggesting the sequence is expressed. However, none of these sequences was proven to be exonic sequence by RT-PCR. One possible explanation is that since the available porcine N-terminal peptide is only 10 amino acid in length, it is possible to find homologous sequences by chance. Another possible explanation is that the N-terminal sequence of the human and pig is not conserved.

An RT-PCR experiment using pig liver RNA (a tissue in which the 77kDa protein is expressed) and a sense degenerate primer designed from the porcine N-terminal peptide and an antisense primer in exon 1 was carried out. The purpose of this experiment was to obtain cDNA sequences for the porcine N-terminal peptide region and the sequence immediately downstream, to increase the length of porcine sequences as well as to understand if sequences are conserved when approaching the N-terminal. A 600 bp cDNA was obtained but the degenerate primer did not bind. Rather, the anti-sense primer (located in porcine exon 1) bound at both ends with non-specific binding at the 5' end. Regardless, this 600 bp cDNA contains exon 1 (230 bp) and 400 bp of 5' sequence. The porcine exon 1 shows 84 % amino acid sequence identity with the human sequence. A translation start site is also identified in the porcine cDNA with the same location as that in the human. The porcine cDNA sequence that encodes the N-terminal peptide and the sequence immediately downstream remain unknown. Interestingly, the 5' sequence of this porcine cDNA shares 60% DNA sequence homology with human exon 43S.

Table 1. A summary of the ORFs in the 5' region of human MTHFR.

All ORFs listed here contain sequences that have a homology to the porcine or mouse sequence or sequences that were found in EST clones. The location of these ORFs is designated by the name of the genomic fragment (clone) and the base pair position of the ORF in the corresponding clone. The position of any possible ATG start site consensus sequence (A/G-C-C-A-T-G-G/A) in the ORF, the location of the primers that were designed to test the authenticity of the ORF by RT-PCR and the confirmed or speculated identity of the ORF are shown. It has been confirmed that the Clc-6 gene exon 1, the MTHFR exons- M3 and MRC-5' are real exons. Other ORFs provide negative results in RT-PCR despite the strong homology to the mouse or porcine sequence and the presence of the sequence in EST clones.

Clone	Primer	Homology (%)	ORF (base pair in clone)	ATG (base pair in clone)	Identity
X5	S352	mouse DNA (90)	1) 280-405	292	possible exon
			2) 309-425	-	
			3) 326-418	-	
	S552	mouse DNA (90)	454-624	531, 550	possible exon
	S854	mouse DNA (90)	1) 636-1220	1026	ClC-6 gene-exon 1
			2) 646-1026	-	
			3) 746-937	-	
	S1124	-	1123-1241	-	MTHFR exon (M3) (coding?)
	S2382	Pi-12 a.a (50)	2367-2537	2439	possible exon
	MRC5'-S5	mouse a.a (84)	3424-3695	3586	MTHFR exon MRC-5' (coding?)
H4_15	H4S803	PN (20-50)	729-905	801	possible exon
H4_3X	S1240	PN (40-50)	1222-1473	1231	possible exon
	S1512	PN (30-50)+EST (99)	1) 1474-1553	1491 and 1512	possible exon
			2) 1439-1582	1556, 1563, 1574	
H4_06	S221	EST (99)	1) 57-179	99	possible exon
			2) 133-219	167	
	S365	PN (40-50)	296-433	296 and 356	possible exon

Although the porcine cDNA has an ORF extension of 260 bp upstream to the ATG start site, no ORF is found in human exon 43S. Genbank database search identified a CpG DNA clone covering almost the entire 43S exon, suggesting that the 43S exon might be a CpG island. The homology found between the 5' end of the porcine cDNA and human exon 43S suggests that perhaps the porcine cDNA may also contain CpG sequence. The identification of these additional porcine sequence also demonstrates multiple cDNAs for the enzyme.

The presence of a CpG island in the human and pig sequence may indicate that the transcription start site is near. CpG islands are known to be sites for DNA methylation in the mammalian genome, which are involved in gene inactivation or, in rare occasions, gene activation (Cross et al 1994). There can be multiple CpG islands spread throughout the entire gene, but it is also common to have CpG islands at the promoter and the beginning of a transcript (Cross et al 1994). A CpG island can be as large as one kilobase. A 150bp CpG DNA clone containing part of human exon 43S was identified through a Genbank database search. These CpG islands had been originally purified by a method that made use of the rat chromosomal protein, MeCP$_2$, that binds to methylated DNA. Genomic DNA was first digested with the restriction enzyme Mse I (TTAA) which rarely cuts at CpG islands because of the high GC content. Digested DNA was then purified by the methylated-DNA binding enzyme. Since there is a Mse I site in the human-43S exon, it is possible that the CpG island may in fact continue even more 5' but was cut by the restriction enzyme before being extracted by the methylated DNA binding enzyme. Therefore, assuming that the presence of the CpG island is indeed an indication of the presence of a transcription start site, the distance between the transcription start site and the 43S exon could be as large as one kilobase. DNA sequence analysis of the X5-4kb genomic fragment was unable to provide concrete information on whether a promoter or transcription start site was present upstream of the 43S exon. Primer extension analysis is in progress, trying to identify the end of at least one transcript which might be in the 4kb genomic fragment.

Genbank database search also revealed the presence of an overlapping gene. A human putative chloride channel gene (ClC-6) (Brandt and Jentsch 1995) was identified approximately 3.5kb upstream of the ATG start codon for the expressed MTHFR cDNA. This ClC-6 gene is transcribed in the opposite direction and in the opposite strand to the MTHFR gene. The mouse ClC-6 gene has also been identified in the same location in the mouse gene supporting that the fact that the ClC-6 gene in this location is not a cloning artifact. Furthermore, both the ClC-6 gene and the MTHFR gene have been shown to be located at chromosome 1p36. More and more mammalian genes have been found to overlap suggesting that this is a common phenomenon. Examples of mammalian genes that are found to overlap with other genes include the tenescin-X gene (Speek et al. 1996), the neuropeptide Y receptor (Herzog et al 1997) in man and the androgen-binding protein (ABP/SHBG) gene in the rat (Joseph 1998). The identification of this gene explains why a sequence homology between the human and mouse previously found in the X5-4kb genomic fragment failed show that it was a MTHFR exon. The homologous sequence was later found to be exon 1 of the ClC-6 gene. Exon 1 of the ClC-6 gene is located only about 250 bp upstream of MTHFR exon-M3, implying that the two genes might overlap. In addition, a mouse exon has been identified 5' to the ClC-6 exon 1, clearly showing that the MTHFR gene and ClC-6 gene are indeed overlapping, at least in the mouse. In the human Ataxia-telangiectasia (*ATM*) gene, an overlapping gene (*E4* or *NPAT*) was observed at the 5' end (Savitsky et al 1997). The two genes are localized in the opposite orientation, similar to the case for MTHFR and ClC-6. The transcription start sites of these two genes were found to be separated by ~500bp and a bi-directional promoter was hypothesized (Savitsky 1997).

The significance of the ClC-6 gene as an overlapping gene is still undetermined, but a speculated co-regulation mechanism could be involved. For instance, assuming that the two genes are indeed overlapping, it means that the promoter of the MTHFR is located in the ClC-6 gene and vice versa. When a RNA polymerase binds to the promoter of one gene and is transcribing in one direction, the binding of another RNA polymerase to the promoter of the other gene and transcription in the opposite direction is probably impossible because the two RNA polymerases might prevent each other from moving forward and would eventually fall off the gene. If this is indeed the case, when one gene is being transcribed, the other gene has to be silent. A co-regulation mechanism may imply that the ClC-6 gene be directly or indirectly involved in the MTHFR related pathways such as methionine or SAM synthesis and proper DNA synthesis.

Genbank EST database search identified 2 EST clones in 2 genomic fragments in the 5' region of the MTHFR gene. Since these sequences are transcribed in the opposite direction to MTHFR, they are therefore likely to be unrelated to MTHFR. The function of these ESTs sequences is still unknown.

In summary, the alternative splicing at the 5' end of the human MTHFR has been confirmed and is probably involved in post-transcriptional regulation. The missing coding sequence is still not identified although the MRC-5' and the M3 exons remain as candidates for at least part of the coding sequence. The porcine N-terminal peptide and other internal peptide sequences were not identified in the human genomic sequence. The inability to identify these porcine sequences in the human genomic sequence is probably due to two reasons. First, the available porcine sequences are relatively short, with a maximum length of 10 amino acids which makes it difficult to achieve significant results when comparing sequences across species. Second, it is possible that the N-terminal sequence between the human and the pig is not conserved. The fact that the two most 5' human exons were not identified in the mouse and at least one mouse exon was not identified in the human suggests that the 5' sequences between different species may not be conserved. On the other, the MRC-5' appears to be strong candidate as a coding exon since it is the only 5' exon that shares a 66% amino acid identity with the mouse. The transcription start site and the promoter are still undetermined although the identification of a CpG island at the 43S exon suggests that the transcription start site might be near. The presence of an overlapping gene, ClC-6, suggests that co-regulation might be involved. If there is indeed a promoter near the 43S exon and the missing coding sequence is localized further upstream in the uncharacterized region of the gene, an alternate promoter and transcription start site might be involved in generating the 77kDa isoform of the MTHFR.

Several directions could be taken in the future to complete the characterization of the 5' region of the human MTHFR gene. First of all, human genomic sequences obtained from sequencing human PAC clones in the human genome project are available in Genbank. Analysis of the 5' region could be performed on genomic DNA sequences if they are available in Genbank. Hundreds if not thousands of EST (Expressed Sequence

Taq) sequences are deposited in Genbank everyday. An extended MTHFR cDNA with the missing coding sequence could be identified in these EST clones. The transcript of MTHFR is believed to contain large UTRs and the 3' end of the MTHFR gene has not been thoroughly investigated. It is possible that the human MTHFR has large 3' UTRs and 3' alternative splicing events.

The human exons MRC-5' and M3 are the two main candidates for being part of the missing coding sequences. More cDNA extensions to these two exons have to be isolated and analyzed for the presence of an upstream ATG start site. If a cDNA extension that contains an ORF and an upstream ATG start site in frame with the original ATG of the MTHFR cDNA is identified, this cDNA extension could potentially be expressed with the entire MTHFR cDNA. The size of this expressed protein could be compared to the 77kDa protein expressed in many human tissues.

The assumption that the N-terminal peptide is conserved across species needs further support. The isolation of the porcine cDNA 5' end that contains the N-terminal peptide of the porcine MTHFR would indicate that the N-terminal porcine sequences are indeed authentic. To identify the 5' boundaries of each of the human 5' exons and the transcription start site, primer extension analysis, S1 nuclease mapping or RNase protection assay could be performed. After the transcription start site is identified, promoter elements could be identified by genomic sequence analysis. To confirm promoter activity, transfection study with a reporter gene construct has to be performed.

6. References

Adams M, Smith PD, Martin D et al. (1996) Genetic analysis of thermolabile methylenetetrahydrofolate reductase as a risk factor for myocardial infarction. *Q J Med* **89**: 437-444

Altschul SF, Gish W, Miller W et al. (1990) Basic local alignment search tool. *J. Mol. Biol* **215**: 403-410

Arruda VR, Von Zuben PM, Chiaparini LC et al. (1997) The mutation Ala 677→ Val in the methylenetetrahydrofolate reducatse gene: arisk factor for arterial disease and venous thrombosis. *Thromb Haemost* **77**: 818-821

Baron JA, Sandler RS, Haile RW, Mandel JS et al. (1998) Folate intake, alcohol consumption, cigarette smoking, and risk of colorectal adenomas. *J National Cancer Institute* **90**: 57-62

Boers GHI, Smals AGH, Trijbels FJM et al (1985) Heterozygosity for homocystinuria in premature peripheral and cerebral occlusive arterial disease. *N Eng J Med* **313**: 709-715

Boushey CJ, Beresford SAA, Omen GS et al (1995) A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. *JAMA* **274**: 1049-1057

Brandt S and Jentsch TJ (1995) ClC-6 and ClC-7 are two broadly expressed members of the ClC chloride channel family. *FEBS Letter* **337**:15-20

Butterworth CE (1993) Folate deficiency and cancer. In *Micronutrients in Health and Disease*. Bendich A and Butterworth CE, eds., (New York: Marcel Dekker) pp165-185

Centers for Disease Control (1992) Recommendation for the use of folic acid to reduce the number of cases of spina bifida and other neural tube defects. *MMWR Morb Mortal Wkly Rep* **41**:1-7

Christensen B, Frosst P, Lussier-Cacan S et al. (1997) Correlation of a common mutation in the methylentetrahydrofolate reductase gene with plasma homcysteine in patients with premature coronary artery disease. *Arterioscler Thromb Vasc Biol.* **17**: 569-573

Clarke R, Daly L, Robinson K et al (1991) Hyperhomocysteinemia. *N Eng J Med* **324**: 1149-1155

Cross SH, Charlton JA, Nan X and Bird A (1994) Purification of CpG islands using a methylated DNA binding column. *Nature Genet* **6**: 236-244

Czeizel AE, Dudas I (1992) Prevention of the first occurrence of neural tube defects by periconceptional vitamins supplementation. *N Engl J Med* **327**: 1832-1835

Daubner SC, Matthews RG (1982) Purification and properties of methylenetetrahydrofolate reductase from pig liver. *J Biol Chem* **257**: 140-145

Deloughery TG, Evans A, Sadeghi A et al. (1996) Common mutation in methylenetetrahydrofolate reductase: a correlation with homocysteine metabolism an late onset vascular disease. *Circulation* **94**: 3074-3078

DeVigneaud VE (1952) *Trial of Research in Sulfur Chemistry and Metabolism, and Related Fields*. Ithaca, NY: Cornell University Press.

Fan J, Vitols KS, Huennekens FM (1992) Multiple folate transport systems in L1210 cells. *Adv Enzyme Regul* **32**: 3

Fenton WA and Rosenberg LE (1995) Inherited disorders of cobalamin transport and metabolism. In *The Metabolic and Molecular Bases of Inherited Disease*, 7th *ed.* Scriver CR, Beaudet AL, Sly S, Valle D, eds., (New York: McGraw-Hill), pp.3129-3149

Finkelstein JD (1990) Methionine metabolism in mammals. *J Nutr Biochem* **1**: 228-237

Fowler B (1997) Disorders of homocysteine metabolism. *J Inher Metab Dis* **20**: 270-285

Frosst P, Zhang ZX, Pai A and Rozen R (1996) The methylentetrahydrofoalte reductase (Mthfr) gene maps to distal mouse chromosome 4. *Mammalian Genome* **7**: 864-869

Frosst P, Blom HJ, Milos, R et al (1995) A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature Genetics* **10**: 111-113

Gallagher PM, Meleady R, Shield DC et al. (1996) Homocysteine and risk of premature coronary heart disease: evidence for a common gene mutation. *Circulation* **94**: 2154-2158

Giovannucci E, Stampfer MJ, Colditz GA, Rimm EB et al. (1993) Folate, methionine and alcohol intake and risk of colorectal adenoma. *J National Cancer Institute* Vol **85**, no 11: 875-884

Giovannucci E, Rimm EB, Ascherio A, Stampfer MJ et al. (1995) Alcohol, low-methioine—low-folate diets, and risk of colon cancer in men. *J National Cancer Institute* Vol **87** no 4: 265-273

Girelli D, Friso S, Trabetti E et al. (1998) Methylenetetrahydrofolate reductase C677T mutation, plasma homocysteine, and folate in subjects from northern Italy with or without angiographyically documented severe coronary atherosclerotic disease: evidence for an important genetic-environmental interaction. *Blood* **91**: 4158-4163

Glynn SA, Albanes D (1994) Folate and cancer. A review of the literature. *Nutrition and Cancer* **22**:101-119

Goyette P, Pai A, Milos R et al. (1998) Gene structure of human and mouse methylenetetrahydrofolate reductase (MTHFR). *Mammalian Genome* **9**: 652-656

Goyette P, Christensen B, Rosenblatt DA and Rozen R (1996) Severe and mild mutations in *cis* for the methylenetetrahydrofolate reductase (MTHFR) gene, and description of five novel mutations in MTHFR. *Am J Hum Gent* **59**: 1268-1275

Goyette P, Frosst P, Rosenblatt DS and Rozen R (1995) Seven novel mutations in the methylenetetrahydrofolate reductase gene and genotype/phenotype correlations in severe MTHFR deficiency. *Am J Hum Genet* **10**: 111-113

Goyette P, Sumner JS, Milos R et al. (1994) Human methylenetetrahydrofolate reductase: isolation of cDNA, mapping and mutation identification. *Nature Genetics* **7**: 195-200

Guenther BD, Sheppard CA, Tran P et al. (1999) The structure and properties of methylenetetrahydrofolate reductase from E. coli: a model for the role of folate in ameliorating hyperhomocysteinemia in humans. *In press.*

Harmon DL, Woodside JV, Yarnell JWG et al. (1996) The common "thermolabile" variant of methylenetetrahydrofolate reductase is a major determinant of mild hyperhomocysteinemia. *Q J Med* **89**: 571-577

Herbert V (1986) The role of vitamin B12 and folate in carcinogenesis. *Advances in Experimental Medicine and Biology* **206**: 293-311

Herzog H, Darby K, Ball H et al. (1997) Overlapping gene structure of the human neuropeptide Y receptor subtypes Y1 and Y5 suggests coordinate transcriptional regulation. *Genomics* **41**: 315-319

Izumi M, Iwai N, Ohmichi N et al. (1995) Molecular varaint of 5,10-methylenetetrahydrofolate reducatase is a risk factor of ishemic heart disease in the Japanese population. *Artherosclerosis* **121**: 293-294

Jacques PF, Bostom AG, Williams RR et al. (1996) Relation between folate status, a common mutation in methylenetetrahydrofolate reductase, and plasma homocysteine concentrations. *Circulation* **93**: 7-9

Jencks D and Matthews RG (1987) Allosteric inhibition of methylentetrahydrofolate reductase by adenosylmethionine. *J. Biol Chem* **262**: 2485-2493.

Joseph DR (1998) The rat androgen-binding protein (ABP/SHBG) gene contains triplet repeats similar to unstable triples: eveidence that the ABPS/SHBG and the fragile X-related genes overlap. *Steriods* **63**: 2-4

Kang SS, Passen EL, Ruggie N etn al. (1993) Thermolabile defect of methylenetetrahydrofolate reductase in coronary artery disease. *Circulation* **88**: 1463-1469

Kang SS, Wong PWK, Malinow MR (1992) Hyperhomcyst(e)inemia as a risk factor for occlusive vascular disease. *Annu Rev Nutr* **12**: 279-298

Kang SS, Wong PWK, Susmano A et al. (1991) Thermolabile methylenetetrahydrofolate reductase. AN inherited risk factor for coronary artery disease. *Am J Hum Genet* **48**: 536-545

Kang SS, Zhou J, Wong PWK, Kowalisyn J and Strokosch G (1988) Intermediate homocysteinemia: A thermolabile variant of methylenetetrahydrofolate reductase. *Am J Hum Genet* **43**: 414-421

Kang SS, Wong PWK, Becker N (1979) Protein-bound homocyst(e)ine in normal subjects
and in patients with homocystinuria. *Pediatr Res* **13**: 1141-1143

Kirke PN, Molloy AM, Daly LE, Burke H, Weir DG, Scott JM (1993) Maternal plasma folate and vitamin B12 are independent risk factors for neural tube defects. *Q J Med* **86**: 703-708

Kirke PN, Mills JL, Whitehead As et al. (1996) Methylenetetrahydrofolate reductase mutation and neural tube defects (Letter) *Lancet* **348**: 1037-1038

Kluijtmans LAJ, Wendel U, Stevens EMB et al. (1998) identification of four novel mutations in severe methylenetetrahydrofolate reductase deficiency. *Euro J Hum Gent* **6**: 257-265

Kluijitmans LAJ, Van den Heuvel LPWJ, Boers GHJ et al. (1996) Molecular genetic analysis in mild hyperhomocysteinemia: a common mutation in the methylentetrahydrofolate reductase gene is a risk factor for cardiovascular disease. *Am J Hum Gent* **58**: 35-41

Kozak M (1996) Interpreting cDNA sequences: some insights from studies on translation. *Mammalian Genome* 7: 563-571

Kozak M (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* **266**: 19867-19870

Ma J, Stampfer MJ, Giovannucci E et al. (1997) Methylenetetrahydrofolate reductase polymorphism, dietry interactions and risk of colorectal cancer. *Cancer Research* **57** (6): 1098-1102

Ma J, Stamper MJ, Hennekens CH et al. (1996) Methylenetetrahydrofolate reductase polymorphism, plasma folate, homocyateine, and risk of myocardial infarction in US physicians. *Circulation* **94**: 2410-2416.

Malinow (1994) Homocyst(e)ine and arterial occlusive diseases. *J Int Med* **236**: 603-617

Markus HS, Ali N, Swaminatham R et al (1997) A common polymorphism in the methylterahydrofolate reductase gene, homocyteine, and ischemic cerebrovascular disease. *Stroke* **28**: 1739-1743

Matthews EG, Vanoni MA, Hainfeld JF and Wall J (1984) Methylentetrahydrofolate reductase. Evidence for spatially distinct subunit domians obtained by scanning transmission electron microscopy and limited proteolysis. *J Biol Chem* **259**: 11647-11640

Medical Research Council Vitamin Study Research group (1991) Prevention of the first occurrence of neural tube defects: results of the Medical Research Council vitamin study. *Lancet* **338**: 131-137

Mills JL. McPartin JM, Kirke PN et al. (1995) Homocysteine metabolism in pregnancies complicated by neural tube defects. *Lancet* **345**: 149-151

Minns RA (1996) Folic acid and neural tube defects. *Spinal cord* **34**: 460- 465

Morita H, Taguchi JI, Kurihara H et al. (1997) Genetic polymorphism of 5,10-methylenetetrahydrofolate reductase (MTHFR) as a risk factor for coronary artery disease. *Circulation* **95**: 2032-2036

Mudd SH, Levy HL, Skovby B (1995) Disorders of transsulfuration. In *The Metabolic and Molecular Bases of Inherited Disease*, 7th *ed.* Scriver CR, Beaudet AL, Sly S, Valle D, eds., (New York: McGraw-Hill), pp.1279-1327

Ou CY, Stevenson RE, Brown VK et al. (1996) 5, 10-methylenetetrahydrofolate reductase genetic polymorphism as risk factor for neural tube defects. *Am J Med Gent* **63**: 610-614

Refsum H, Ueland PM, Nygard O and Vollset SE (1998) Homocysteine and cardiovascular disease. *Annu Rev Medicine* **49**: 31-62

Refsum H, Fiskerstrand T, Guttormsen AB et al. (1997) Assessment of homocysteine status. J *Inher Metab Dis* **20**: 286-294

Rieder MJ (1994) Prevention of neural tube defects with periconceptional folic acid. *Clin Perinatol* **21**: 483-503

Rosenblatt, D.S. (1995) Inherited disorders of folate transport and metabolism. In *The Metabolic and Molecular Bases of Inherited Disease*, 7th *ed.* Scriver CR, Beaudet AL, Sly S, Valle D, eds., (New York: McGraw-Hill), pp.3111-3128

Rozen R (1996) Molecular genetic aspects of hyperhomocysteinemia and its relation to folic acid. *Clin Invest Med* **19**: 171-178

Saint-Giron I et al (1983) Nucleotide sequence of metF, the *E. coli* structural gene for 5-10 methylenetetrahydrofolate reductase and of its homology to control region. *Nucl. Acid Res.* **11**: 6723-6732.

Sambrook J, Fritsch EF and Maniatia T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed. (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Savitsky K, platzer M, Uzeil T et al. (1997) Ataxia-telengiectasia: structural diversity of untranslated sequences suggests complex post-transcriptional regulation of *ATM* gene expression. *Nucl Acids Res* **25**: 1678-1684

Schmitz C, Lindpaintner K, Verhoef P et al. (1996) Genetic polymorphism of methylenetetrahydrofolate reductase and myocardial infarction: a case-control study. *Circulation* **94**: 1812-1814

Speek M, Barry F and Miller WL (1996) Alternative promoters and alternate splicing of human tenescin-X, a gene with 5' and 3' ends buried in other genes. *Hum Mol Gent* **5**: 1749-1758

Steegers-Theunissen RPM, Boers GHJ, Trijbels FJM et al (1994) Maternal hyperhomocysteinemia: A risk factor for neural tube defects? *Metabolism* **43**: 1475-1480

Sumner J, Jencks DA, Khani S and Matthews RG (1986) Photoaffinity labeling of methylenetetrahydrofolate reductase with 8-azido-S-adenosylmethionine. *J Biol Chem* **261**: 7697-7700

Ueland PM, Refsum H, Stabker SP et al. (1993) Total homocysteine in plasma or serum: methods and applications. *Clin Chem* **39**: 1764-1779

Van der Pt NM, Gabreels F, Stevens EMB et al. (1998) A second common mutation in the methylenetetrahydrofolate reductase gene: an additional risk factor for neural-tube defects?

Van der Put NM, van den Heuviel LP, Steegers-Theunissen RP et al. (1996) Decreased methylenetetrahydrofolate reductase activity due to the 677C→T mutation in families with spina bifida offspring. *J Mol Med* **74**: 691-694

Van der Put NMJ, Steegers-Theunissen RPM, Frosst P et al. (1995) Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. *Lancet* **346**: 1070-1071

Wakamatsu N, Benoit G, Lamhonwah AM et al. (1994) Structural organization, sequence, and expression of the mouse *HEXA* gene encoding the α subunit of Hexosaminidase A. *Genomics* **24**: 110-119

Watkins D, Cooper BA (1983) A critical intracellular concentration of fully reduced non methylated folate polyglutamate prevents macrocytosis and diminished growth rate in human cell K562 in culture. *Biochem J* **214**: 456

Weisberg I, Tran P, Christensen B et al. (1998) A second genetic polymorphism in methylenetetrahydrofolate reductase (MTHFR) associated with decreased enzyme activity. *Mol Genet Metab* **64**: 169-172

Whitehead AS, Gallagher P, Mills JL et al. (1995) A genetic defect in 5, 10 methylenetetrahydrofolate reductase in neural tube defects. *Q J Med* **88**: 763-766

Wilcken DEL, Wnag XL, Sim AS and McCredie RM (1996) Distribution in healthy and coronary populations of the methylenetetrahydrofolate reductase (MTHFR) C-677T mutation. *Arterioscler Thromob Vasc Biol* **16**: 878-882

Wills L (1931) Treatment of "pernicious anaemia of pregnancy" and "tropical anaemia." *Br Med J* **1**: 1059

Wills L, Stewart A (1935) Experimental anemia in monkeys with special reference to macrocytic nutritional anemia. *Br J exp Pathol* **16**: 444

Zittoun J (1995) Congenital errors of folate metabolism. In *Baillière 's Clincal Haematology* Vol **8** no 3: 603-616