CHARACTERIZATION OF THE HUMAN CERULOPLASMIN cDNA AND GENE

By

Marlys Laverne Koschinsky

B.Sc. (Hons), The University of Winnipeg, 1982

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Department of Biochemistry)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

March 1988

© Marlys Laverne Koschinsky, 1988
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of **BIOCHEMISTRY**

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date  **April 19 88**
A cDNA for human ceruloplasmin was identified in a human liver cDNA library by screening with mixtures of synthetic oligonucleotides complementary to two regions of the ceruloplasmin mRNA. The resulting clone (phCP-1) contained DNA coding for amino acid residues 202 - 1046 of the protein, followed by a 3' untranslated region of 123 bp and a poly(A) tail. To isolate additional clones extending in a 5' direction, two randomly-primed human liver cDNA libraries were constructed in the bacteriophage vectors λgt10 and λgt11. From the former library, a clone was isolated (λhCP-1) that contained DNA coding for a putative signal peptide consisting of 19 amino acid residues, followed by DNA encoding residues 1 - 380 of plasma ceruloplasmin. From the λgt11 library, six ceruloplasmin cDNA clones were purified, two of which were shown to contain 10 and 38 bp of non-coding sequence extending 5' to λhCP-1. Blot hybridization analysis using cDNA probes showed that ceruloplasmin mRNA from the human hepatoma cell line HepG2 is 3700 nucleotides in size, while human liver RNA contained an additional hybridizing species of 4500 nucleotides in size.

Ceruloplasmin genomic DNA clones (spanning a region of approximately 45 Kbp) were obtained by the screening of several human genomic phage libraries using cDNA probes. These clones were initially characterized by restriction endonuclease mapping. Using DNA sequence analysis, the positions of intron/exon boundaries were determined. To date, 14 exons (average size of 183 bp) have been identified in the ceruloplasmin gene, corresponding to nucleotide residues 1 - 2565 of the coding sequence. The
majority of the 14 introns localized within this region were located in analogous positions in the factor VIII gene, thereby suggesting that these two proteins have evolved from a common ancestral gene.

At least 4 exons have been localized within the 5' untranslated region of the human ceruloplasmin gene, although typical eukaryotic promoter elements have not yet been identified. The significance of this novel organization remains unclear at present.

In addition to the wild-type gene, a processed pseudogene for human ceruloplasmin was identified and contained DNA corresponding to the functional gene sequence encoding the carboxy-terminal 563 amino acid residues and the 3' untranslated region. The pseudogene appears to have arisen from a processed RNA species, since intervening sequences coincident with those of the functional gene have been removed with the exception of a short segment of intronic sequence which denotes the 5' boundary of the pseudogene. Based on genomic Southern blot analysis performed under high stringency conditions, the pseudogene seems to comprise the only sequence in the human genome that is closely related to the wild-type gene. Using somatic cell hybridization, the pseudogene was localized to human chromosome 8; this differs from the location of the wild-type ceruloplasmin gene, which has been mapped to chromosome 3.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Abstract</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>xiii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xiv</td>
</tr>
<tr>
<td>Dedication</td>
<td>xv</td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>I.A PERSPECTIVES</td>
<td>1</td>
</tr>
<tr>
<td>I.B PROPERTIES OF CERULOPLASMIN</td>
<td></td>
</tr>
<tr>
<td>I.B.1 Structure of the Protein</td>
<td>1</td>
</tr>
<tr>
<td>I.B.2 Sites of Ceruloplasmin Biosynthesis</td>
<td>9</td>
</tr>
<tr>
<td>I.B.3 Functions of Ceruloplasmin</td>
<td>10</td>
</tr>
<tr>
<td>I.B.3.1 Ferroxidase activity</td>
<td>10</td>
</tr>
<tr>
<td>I.B.3.2 Serum antioxidant activity</td>
<td>10</td>
</tr>
<tr>
<td>I.B.3.3 Amine oxidase activity</td>
<td>11</td>
</tr>
<tr>
<td>I.B.3.4 Role of ceruloplasmin in copper transport</td>
<td>11</td>
</tr>
<tr>
<td>I.B.4 Regulation of Ceruloplasmin Expression</td>
<td>12</td>
</tr>
<tr>
<td>I.B.4.1 Hormonal regulation of ceruloplasmin synthesis</td>
<td>12</td>
</tr>
<tr>
<td>I.B.4.2 Copper induction of ceruloplasmin expression</td>
<td>13</td>
</tr>
<tr>
<td>I.B.4.3 Regulation of ceruloplasmin synthesis during inflammation</td>
<td>14</td>
</tr>
<tr>
<td>I.C ABNORMALITIES IN COPPER HOMEOSTASIS - WILSON'S DISEASE</td>
<td>15</td>
</tr>
<tr>
<td>I.D CHARACTERIZATION OF RAT CERULOPLASMIN</td>
<td>17</td>
</tr>
<tr>
<td>I.E THE RELATIONSHIP OF CERULOPLASMIN TO OTHER COPPER-CONTAINING PROTEINS</td>
<td>19</td>
</tr>
<tr>
<td>I.F THE RELATIONSHIP BETWEEN CERULOPLASMIN AND PROTEINS INVOLVED IN BLOOD COAGULATION</td>
<td>26</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>I.G.</td>
<td>CHARACTERIZATION OF THE HUMAN FACTOR VIII GENE</td>
</tr>
<tr>
<td>I.G.1</td>
<td>Historical Perspectives</td>
</tr>
<tr>
<td>I.G.2</td>
<td>Organization of the Factor VIII Gene</td>
</tr>
<tr>
<td>I.G.3</td>
<td>Evolutionary Aspects of Intron Positions Within the Factor VIII Gene</td>
</tr>
<tr>
<td>I.H</td>
<td>DYNAMICS OF PROTEIN AND GENE EVOLUTION</td>
</tr>
<tr>
<td>I.H.1</td>
<td>The Molecular Clock</td>
</tr>
<tr>
<td>I.I</td>
<td>MECHANISMS OF GENE EVOLUTION</td>
</tr>
<tr>
<td>I.I.1</td>
<td>Gene Duplication</td>
</tr>
<tr>
<td>I.I.2</td>
<td>Gene Fusion</td>
</tr>
<tr>
<td>I.I.3</td>
<td>Exon Shuffling</td>
</tr>
<tr>
<td>I.I.4</td>
<td>Intron Insertion and Intron Sliding</td>
</tr>
<tr>
<td>I.J</td>
<td>PSEUDOGENES</td>
</tr>
<tr>
<td>I.J.1</td>
<td>Non-Processed Pseudogenes</td>
</tr>
<tr>
<td>I.J.2</td>
<td>Processed Pseudogenes</td>
</tr>
<tr>
<td>I.K.</td>
<td>The Present Study</td>
</tr>
<tr>
<td>II.</td>
<td>MATERIALS AND METHODS</td>
</tr>
<tr>
<td>II.A</td>
<td>BACTERIAL HOSTS AND MEDIA</td>
</tr>
<tr>
<td>II.B</td>
<td>HYBRIDIZATION PROBES</td>
</tr>
<tr>
<td>II.B.1</td>
<td>Purification and Labelling of Oligodeoxyribonucleotides</td>
</tr>
<tr>
<td>II.B.2</td>
<td>Nick Translation</td>
</tr>
<tr>
<td>II.B.3</td>
<td>Klenow Labelling</td>
</tr>
<tr>
<td>II.B.4</td>
<td>Preparation of M13 Probes</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>II.C</td>
<td>IDENTIFICATION OF cDNAS FOR HUMAN CERULOPLASMIN</td>
</tr>
<tr>
<td>II.C.1</td>
<td>Screening of a Human Liver cDNA Library</td>
</tr>
<tr>
<td>II.C.2</td>
<td>Preparation and Screening of Randomly-Primed Human Liver cDNA Libraries</td>
</tr>
<tr>
<td>II.D</td>
<td>SCREENING OF HUMAN GENOMIC LIBRARIES</td>
</tr>
<tr>
<td>II.E</td>
<td>ISOLATION OF NUCLEIC ACIDS</td>
</tr>
<tr>
<td>II.E.1</td>
<td>Purification of Plasmid DNA</td>
</tr>
<tr>
<td>II.E.2</td>
<td>Isolation of Bacteriophage DNA</td>
</tr>
<tr>
<td>II.E.3</td>
<td>Preparation of Human Genomic DNA</td>
</tr>
<tr>
<td>II.E.4</td>
<td>Isolation of RNA</td>
</tr>
<tr>
<td>II.E.4.1</td>
<td>Preparation of human liver poly(A)$^+$ RNA</td>
</tr>
<tr>
<td>II.E.4.2</td>
<td>Preparation of total RNA from HepG2 cells</td>
</tr>
<tr>
<td>II.F</td>
<td>BASIC DNA TECHNIQUES</td>
</tr>
<tr>
<td>II.F.1</td>
<td>Restriction Enzyme Digestion</td>
</tr>
<tr>
<td>II.F.2</td>
<td>End-Labelling of DNA Fragments</td>
</tr>
<tr>
<td>II.F.3</td>
<td>Electrophoresis of DNA</td>
</tr>
<tr>
<td>II.F.3.1</td>
<td>Agarose gel electrophoresis</td>
</tr>
<tr>
<td>II.F.3.2</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>II.F.4</td>
<td>Southern Transfers</td>
</tr>
<tr>
<td>II.G</td>
<td>DNA CLONING</td>
</tr>
<tr>
<td>II.G.1</td>
<td>Fragment Production</td>
</tr>
<tr>
<td>II.G.2</td>
<td>Ligation of DNA into pUC or M13 Vectors</td>
</tr>
<tr>
<td>II.G.3</td>
<td>Transformations</td>
</tr>
<tr>
<td>II.H</td>
<td>DNA SEQUENCE ANALYSIS</td>
</tr>
<tr>
<td>II.H.1</td>
<td>Screening of M13 Clones</td>
</tr>
<tr>
<td>II.H.2</td>
<td>Isolation of M13 Template DNA</td>
</tr>
<tr>
<td>II.H.3</td>
<td>DNA Sequence Analysis</td>
</tr>
</tbody>
</table>
## II.I RNA ANALYSIS

- **II.I.1 Northern Blot Analysis**
- **II.I.2 RNA Dot Blots**
- **II.I.3 Nuclease S1 Mapping**

## II.J CHROMOSOME MAPPING

## II.K SUMMARY OF HYBRIDIZATION/WASHING CONDITIONS

- **II.K.1 Genomic Southern Blot Analysis**
- **II.K.2 Hybridization Conditions Other Than For Genomic Southern Blots**

## III. RESULTS

### III.A CHARACTERIZATION OF THE HUMAN PRECERULOPLASMIN cDNA

- **III.A.1 Initial Screening of a Human Liver cDNA Library**
- **III.A.2 Isolation of cDNA Clones Encoding the 5' End of Ceruloplasmin**
- **III.A.3 DNA Sequence Analysis of Human Preceruloplasmin**
- **III.A.4 Ceruloplasmin Transcript Analysis**

### III.B CHARACTERIZATION OF THE WILD-TYPE HUMAN CERULOPLASMIN GENE

- **III.B.1 Isolation and Restriction Endonuclease Mapping of Genomic Clones**
- **III.B.2 Localization of Intron/Exon Junctions Corresponding to the Ceruloplasmin Coding Region**
- **III.B.3 Partial Nucleotide Sequence Analysis of the Human Ceruloplasmin Gene**
- **III.B.4 Organization of the 5' End of the Human Ceruloplasmin Gene**
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>III.B.4.1</td>
<td>Comparison of genomic and cDNA sequence data</td>
<td>98</td>
</tr>
<tr>
<td>III.B.4.2</td>
<td>SI mapping analysis of exon 1</td>
<td>98</td>
</tr>
<tr>
<td>III.B.4.3</td>
<td>Southern blot analysis of exon L4</td>
<td>98</td>
</tr>
<tr>
<td>III.B.4.4</td>
<td>Northern blot analysis of the 5' end of the human ceruloplasmin gene</td>
<td>105</td>
</tr>
<tr>
<td>III.B.4.5</td>
<td>RNA dot blot analysis</td>
<td>105</td>
</tr>
<tr>
<td>III.C</td>
<td>ISOLATION AND COMPLETE CHARACTERIZATION OF A PSEUODOGENE FOR HUMAN CERULOPLASMIN</td>
<td>111</td>
</tr>
<tr>
<td>III.C.1</td>
<td>Isolation of Genomic Clones Containing the Human Ceruloplasmin Pseudogene</td>
<td>111</td>
</tr>
<tr>
<td>III.C.2</td>
<td>DNA Sequence of the Human Ceruloplasmin Pseudogene</td>
<td>114</td>
</tr>
<tr>
<td>III.C.3</td>
<td>Nuclease SI Analysis of the Human Ceruloplasmin Pseudogene</td>
<td>118</td>
</tr>
<tr>
<td>III.C.4</td>
<td>Chromosome Localization of the Human Ceruloplasmin Pseudogene</td>
<td>124</td>
</tr>
<tr>
<td>III.D</td>
<td>GENOMIC SOUTHERN BLOT ANALYSIS OF THE HUMAN CERULOPLASMIN GENE</td>
<td>124</td>
</tr>
<tr>
<td>IV.</td>
<td>DISCUSSION</td>
<td>132</td>
</tr>
<tr>
<td>IV.A</td>
<td>CHARACTERIZATION OF THE HUMAN PRECERULOPLASMIN cDNA</td>
<td>132</td>
</tr>
<tr>
<td>IV.A.1</td>
<td>DNA Sequence Analysis of the Human Preceruloplasmin cDNA</td>
<td>132</td>
</tr>
<tr>
<td>IV.A.2</td>
<td>Internal Homology Within the Ceruloplasmin cDNA Sequence</td>
<td>136</td>
</tr>
<tr>
<td>IV.A.3</td>
<td>Analysis of the Human Ceruloplasmin Transcript</td>
<td>137</td>
</tr>
<tr>
<td>IV.B</td>
<td>CHARACTERIZATION OF THE WILD-TYPE HUMAN CERULOPLASMIN GENE</td>
<td>138</td>
</tr>
<tr>
<td>IV.B.1</td>
<td>Ceruloplasmin Gene Organization Corresponding to the Coding Sequence</td>
<td>138</td>
</tr>
<tr>
<td>IV.B.2</td>
<td>DNA Sequence Analysis of the Wild-Type Ceruloplasmin Gene</td>
<td>139</td>
</tr>
<tr>
<td>IV.B.3</td>
<td>Intron Positions Within the Triplicated A Domain of Human Ceruloplasmin</td>
<td>139</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>IV.B.4</td>
<td>Comparison of the Gene Organization of Ceruloplasmin and Factor VIII</td>
<td>145</td>
</tr>
<tr>
<td>IV.B.5</td>
<td>Characterization of the 5' End of the Human Ceruloplasmin Gene</td>
<td>150</td>
</tr>
<tr>
<td>IV.C</td>
<td>CHARACTERIZATION OF A PSEUDOGENE FOR HUMAN CERULOPLASMIN</td>
<td>153</td>
</tr>
<tr>
<td>IV.C.1</td>
<td>DNA Sequence Analysis of the Human Ceruloplasmin Pseudogene</td>
<td>153</td>
</tr>
<tr>
<td>IV.C.2</td>
<td>Chromosome Localization of the Human Ceruloplasmin Pseudogene</td>
<td>156</td>
</tr>
<tr>
<td>IV.C.3</td>
<td>Speculations on the Evolutionary Origin of the Human Ceruloplasmin Pseudogene</td>
<td>157</td>
</tr>
<tr>
<td>IV.D</td>
<td>A MODEL FOR THE EVOLUTION OF CERULOPLASMIN, FACTOR V AND FACTOR VIII</td>
<td>159</td>
</tr>
<tr>
<td>IV.E</td>
<td>CONCLUDING REMARKS</td>
<td>163</td>
</tr>
<tr>
<td>V.</td>
<td>REFERENCES</td>
<td>165</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>I</th>
<th>Comparison of A Domains in Factor V, Factor VIII and Ceruloplasmin</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>Summary of the Genotypes of Bacterial Hosts Used in the Present Study</td>
<td>46</td>
</tr>
<tr>
<td>III</td>
<td>Subcloning Strategy for the Wild-Type Human Ceruloplasmin Gene</td>
<td>67</td>
</tr>
<tr>
<td>IV</td>
<td>Compositions of M13 DNA Sequencing Mixes</td>
<td>70</td>
</tr>
<tr>
<td>V</td>
<td>Nucleotide Sequence of Intron/Exon Junctions in the Human Ceruloplasmin Gene</td>
<td>94</td>
</tr>
<tr>
<td>VI</td>
<td>Frequency of Nucleotides at Intron/Exon Junctions</td>
<td>95</td>
</tr>
<tr>
<td>VII</td>
<td>Sizes and Positions of Introns and Exons Within the Ceruloplasmin Gene</td>
<td>97</td>
</tr>
<tr>
<td>VIII</td>
<td>Segregation of the Human Ceruloplasmin Pseudogene with Human Chromosomes in Human-Hamster Somatic Hybrids</td>
<td>123</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1. Structural Model of the Human Ceruloplasmin Molecule ........................................ 5
2. Relationship between Ceruloplasmin and Other Copper-Containing Proteins .................. 21
3. Comparison of the Structural Organization of Ceruloplasmin, Factor V and Factor VIII .......... 28
4. Intron Positions within the A and C Domains of Human Factor VIII .......................... 36
5. Schematic Summary of the Cloning of the Human Preceruloplasmin cDNA ......................... 77
6. Sequencing Strategy for the Human Preceruloplasmin cDNA .................................. 80
7. Nucleotide Sequence of Human Preceruloplasmin cDNA .......................................... 83
8. Comparative Analysis of cDNA Clones \(\lambda\)hCP-2 to \(\lambda\)hCP-6 ............................... 85
9. Blot Hybridization Analysis of Human Ceruloplasmin mRNA .................................. 88
10. Partial Restriction Map and Intron/Exon Organization of the Human Ceruloplasmin Gene .......... 91
11. Comparison of the Sequence of \(\lambda\)hCP-6 with Overlapping Genomic Sequence Derived from \(\lambda\)WT2 ................................................................. 99
12. Nuclease S1 Mapping of Exon 1 ........................................................................ 101
13. Southern Blot Analysis of Exon L4 .................................................................... 103
14. Northern Blot Analysis of the 5' End of the Ceruloplasmin Gene .............................. 106
15. RNA Dot Blot Analysis of the 5' Untranslated Region of the Ceruloplasmin Gene ........ 109
16. Partial Restriction Map and Sequencing Strategy for the Human Ceruloplasmin Pseudogene ...... 112
17. Nucleotide Sequence of the Human Ceruloplasmin Pseudogene and Comparison with the Corresponding Region of the Ceruloplasmin cDNA Sequence ........................................... 115
18. Nuclease S1 Mapping Analysis of the Ceruloplasmin Pseudogene .......................... 119
## Page 121

19. Chromosome Mapping of the Human Ceruloplasmin Pseudogene Using Somatic Cell Hybrid Analysis

## Page 125

20. Genomic Southern Blot Analysis of the Human Ceruloplasmin Gene

## Page 128

21. Genomic Southern Analysis of the Human Ceruloplasmin Pseudogene and Related Sequences

## Page 130

22. Genomic Southern Blot Analysis of the 3' End of the Human Ceruloplasmin Gene

## Page 140

23. Intron Positions Within the Three Repeated Units of Human Ceruloplasmin

## Page 143

24. Positions of Introns in the A Domains of Ceruloplasmin and Factor VIII

## Page 147

25. Comparative Positions of Introns in Ceruloplasmin with Corresponding Introns in Factor VIII

## Page 160

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>adenosine</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine-5'-triphosphate</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>C</td>
<td>cytosine</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>cpm</td>
<td>counts per minute</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxyribonucleotidetriphosphate</td>
</tr>
<tr>
<td>ddNTP</td>
<td>dideoxyribonucleotidetriphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediamine tetraacetic acid</td>
</tr>
<tr>
<td>G</td>
<td>guanosine</td>
</tr>
<tr>
<td>HEPES</td>
<td>N-2-hydroxyethylpiperazine-N-2-ethanesulfonic acid</td>
</tr>
<tr>
<td>Kb, Kbp</td>
<td>Kilobases, Kilobase pairs</td>
</tr>
<tr>
<td>Kda</td>
<td>Kilodaltons</td>
</tr>
<tr>
<td>Krpm</td>
<td>thousand revolutions per minute</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>Mr</td>
<td>relative molecular weight</td>
</tr>
<tr>
<td>N</td>
<td>adenine, cytosine, guanine or thymine</td>
</tr>
<tr>
<td>OD</td>
<td>optical density</td>
</tr>
<tr>
<td>PIPES</td>
<td>piperazine-N,N'-bis (2-ethanesulfonic acid)</td>
</tr>
<tr>
<td>poly(A)+</td>
<td>polyadenylated</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulfate</td>
</tr>
<tr>
<td>T</td>
<td>thymidine</td>
</tr>
<tr>
<td>TEMED</td>
<td>N,N,N',N'-Tetramethylethylene diamine</td>
</tr>
<tr>
<td>Tris</td>
<td>Tri(hydroxymethyl)aminoethane</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>U</td>
<td>uridine</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>W</td>
<td>watts</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to thank my supervisor, Ross MacGillivray, for his enthusiastic support of my career ("great"), and his willingness to allow each member of the laboratory to develop an independent approach to scientific research. Ross has generated an atmosphere of warmth and friendship within his research group that has filled my past five years with many happy memories. I would also like to thank Walter (Waltman) Funk for the special camaraderie that we have shared over the years. In addition, I would like to acknowledge Heather Kirk, for technical advice during several aspects of this project, and Susan Heming, for her patience in the preparation of the manuscript.

A very special thanks to Roland Russnak for his relentless optimism -- it might be contagious!
DEDICATION

To Mother and Father
for their tireless love and
belief in me that has given
me the strength to keep trying.
I. INTRODUCTION

A. PERSPECTIVES

Copper is required for the function of a number of metalloenzymes and metalloproteins present in both prokaryotic and eukaryotic cells. Therefore, strict maintenance of copper homeostasis is essential for many vital processes. On this basis, it is not surprising that since its initial discovery by Holmberg in 1944, ceruloplasmin (the principal copper transport protein in vertebrate plasma) has been the focus of intensive biochemical study. This, in turn, has resulted in the generation of several thousand papers dealing with the properties of this large multicopper oxidase. In addition to its multifunctional nature which renders it an interesting protein for study, ceruloplasmin has also received attention in analyses tracing the evolution of metal-containing enzymes and proteins in aerobic cells (e.g. Frieden, 1974).

B. PROPERTIES OF HUMAN CERULOPLASMIN

B.1 Structure of the Protein

Ceruloplasmin is a blue, crystallizable (Morell, 1969; Nakagawa, 1972) α-2 glycoprotein that binds 90 – 95% of plasma copper in vertebrates. The remaining copper exists in complexes of amino acids (Harris and Sass-Kortsak, 1967), serum albumin (Sarkar and Wigfield, 1968) and a tripeptide composed of glycine-histidine-lysine (Pickart et al., 1980), all of which may provide auxiliary transport mechanisms (Frieden, 1981).

Multiple enzymatic functions have been ascribed to ceruloplasmin (see Section I.B.3; for reviews dealing with the functions of ceruloplasmin,
see Frieden, 1981; Owens, 1982; Cousins, 1985), all of which likely involve the presence of intrinsically-bound cupric ions. Each ceruloplasmin molecule (Mr = 132 Kda) contains at least 6 or 7 copper-binding sites (Rydén and Björk, 1976), which can be categorized based upon the physical properties of light absorbance and electromagnetic behavior. Although the stoichiometry of the different copper types of ceruloplasmin is not well-established, the following composition has been proposed: two type I and one type II [based on quantitative electron paramagnetic resonance (EPR) measurements of Deinum and Vännegard, 1973], and three (or four) type III (Rydén and Björk, 1976). Type I copper is present in small blue electron-transfer proteins and blue copper oxidases of which ceruloplasmin is the only mammalian representative. Type I copper centres are characterized by strong absorbance around 600 nm (red light), and typically result in narrow EPR hyperfine splitting. Based on circular dichroism (CD) and magnetic circular dichroism (MCD) studies (Dawson et al., 1979), type I sites in ceruloplasmin are proposed to be coordinated by a cysteine, methionine, and two histidine ligands, in a distorted tetrahedral geometry. Type I copper has a redox potential appreciably higher than that of the Cu (II)/Cu (I) couple in aqueous solution (Fee, 1975). This may be due to the distortion of the copper geometry which results in decreased activation energy required for electron transfer (Williams, 1971).

Type II copper absorbs weakly at 600 nm, and features an EPR hyperfine structure similar to that found in tetragonal copper complexes. Type II copper is present in large blue multicopper oxidases, such as laccase, cytochrome oxidase, ascorbate oxidase and ceruloplasmin, as well
as in non-blue oxidases, such as monoamine oxidase and galactose oxidase. Studies of the coordination environment of type II copper in ceruloplasmin (Dawson et al., 1978) and galactose oxidase (Bereman and Kosman, 1977) provide evidence for histidine coordination of the copper in these type II centres.

The type III copper centre is binuclear, composed of two copper (II) ions antiferromagnetically coupled and therefore EPR nondetectable. These copper atoms are associated with an intense absorption band at 330 nm. Type III copper is present in copper oxidases that catalyze the reduction of dioxygen to two water molecules, such as ceruloplasmin, ascorbate oxidase and laccase. Very little is known about the coordination geometry or specific ligand groups for type III centres (Urbach, 1981). An EPR signal has been detected in the latter two proteins which differs from the signals resulting from type I or type II Cu (II) ions (Reinhammar et al., 1980). This new signal has been attributed to one of the pair of Cu (II) ions existing in the binuclear centre, and is comparable to signals observed in half-met hemocyanin (Himmelwright et al., 1978) and superoxide dismutase (Fielden et al., 1974). Both of the latter proteins have been shown to contain copper-binding sites in binuclear centres.

Despite persistent study, the assignment of copper-binding sites to specific regions of the ceruloplasmin polypeptide chain remains largely undetermined. However, it has been recently demonstrated by Raju (1983) that 50% of the non-blue copper-binding occurs in an 11 Kda fragment derived from the carboxyl terminus of ceruloplasmin (corresponding to amino acid residues 885 - 1046). Additionally, based on sequence
similarity with known blue and non-blue copper binding sites, several putative copper-binding centres have been identified in ceruloplasmin (see Section I.E).

Early studies suggested that ceruloplasmin had a subunit structure consisting of 2 – 8 polypeptide chains (Freeman and Daniel, 1973; Poulik and Weiss, 1975; McCombs and Bowman, 1976). Rydén (1972) demonstrated, however, that the observed subunits corresponded to proteolytic fragments that could be eliminated when fresh plasma, supplemented with protease inhibitors, was used for the isolation of ceruloplasmin. This led to a proposed single chain structure for the molecule, which was later confirmed by amino acid sequence analysis (Takahashi et al., 1984). In vitro and possibly in vivo, spontaneous proteolytic cleavage occurs rapidly at two highly sensitive sites following basic amino acid residues, producing fragments with molecular weights of 67 Kda, 50 Kda, and 19 Kda, corresponding to the amino-terminal, middle, and carboxy-terminal portions of the protein, respectively (Kingston et al., 1980; Dwulet and Putnam, 1981a; see Figure 1). Limited tryptic digestion in vitro results in cleavage at the two labile sites described above, and additionally degrades the 50 Kda fragment, producing 25- and 26 Kda fragments, and slowly cleaves the 67 Kda fragment, yielding 49- and 18 Kda fragments (Takahashi et al., 1983; see Figure 1). The physiological significance of the limited proteolytic cleavage is unclear, but may function in protein regulation as is proposed for the deactivation of anaphylatoxins and bradykinin (see below).

The complete amino acid sequence of human ceruloplasmin, consisting of 1046 residues, was determined by analysis of the 19 Kda (Kingston
Figure 1. **Structural model of the human ceruloplasmin molecule based on proteolytic cleavage sites and internal amino acid sequence identity** (modified from Ortel et al., 1984; see text for details).

The polypeptide chain is cleaved autolytically into the 67, 50 and 19 Kda fragments as shown. In the intact polypeptide chain, these fragments are connected by single amino acid residues arginine (R) and lysine (K). Tryptic cleavage sites are indicated by vertical arrows, with broad arrows identifying major sites of cleavage. The sizes of proteolytic fragments are given in Kilodaltons (Kda). The positions of glucosamine oligosaccharide attachment sites are indicated by solid diamonds; the carbohydrate moiety that is missing in the type II ceruloplasmin variant is starred. The extent of the triplicated units A1, A2 and A3 is also shown.
Proteolytic Fragments (Kda)

Triplicated Units (amino acids)
et al., 1979) and 50 Kda (Dwulet and Putnam, 1981a) proteolytic fragments; the sequence of the 67 Kda fragment and overlapping peptides was reported by Takahashi et al. (1984). The latter strategy revealed the presence of an additional amino acid residue (Arg 481) between the 67- and 50 Kda fragments, and a single residue (Lys 667) connecting the 50- and 19 Kda fragments. It is assumed that during preparation, these two basic residues are excised by an enzyme with carboxypeptidase-like specificity. A similar mechanism is involved in the carboxypeptidase-B catalyzed removal of carboxy-terminal arginine or lysine residues that results in inactivation of kinins, and the C3a and C5a analphylatoxins (Putnam, 1984).

Ceruloplasmin possesses attachment sites for four glucosamine-linked (GlcN) oligosaccharides (Tetaert et al., 1982; Takahaski et al., 1984). The location of these carbohydrate attachment points was determined by separation of GlcN-containing peptides using reverse-phase high pressure liquid chromatography, followed by amino acid sequence analysis. Three sites are located in the 67 Kda fragment, while one resides in the 50 Kda proteolytic fragment (see Figure 1). All four GlcN oligosaccharides are linked in the obligatory tripeptide acceptor sequence Asn-X-Ser/Thr (Clamp, 1975). Ceruloplasmin also has three GlcN acceptor sequences that are not glycosylated. This is probably the result of their location in buried, hydrophobic regions of the protein (Ortel et al., 1984).

Two variants of ceruloplasmin have been identified based on carbohydrate composition. In type I ceruloplasmin (predominant form), all four oligosaccharides are present, while the second oligosaccharide (i.e. Asn-339) is missing in the less abundant type II form (Takahashi et al.,
1984) (see Figure 1). The physiological significance of these two variant forms is unknown at present.

Based on computer analysis of the amino acid sequence, the entire human ceruloplasmin molecule has been shown to exhibit an internal 3-fold homology, with each homology unit consisting of approximately 350 amino acids (Takahashi et al., 1984; see Figure 1). These units (arbitrarily designated A1, A2 and A3 from amino- to carboxy-terminus) share nearly 40% sequence identity when compared pairwise (see Table 1), including a high degree of conservation of the four least frequent amino acids in proteins: methionine, histidine, tryptophan, and cysteine (Ortel et al., 1984). The amino acid sequence conservation has been interpreted to suggest a high level of structural conservation between these related segments (Ortel et al., 1984). The proteolytic cleavage pattern of non-denatured ceruloplasmin suggests that each of these three regions is subdivided into 2 or 3 domains (see Figure 1). The boundaries between the individual homology units do not correspond to the sites of proteolytic cleavage described previously, which generate the 67-, 50- and 19 Kda proteolytic fragments. However, trypsin cleaves at a site between the A1 and A3 domains. Due to additional sites of cleavage, these 2 units are further divided into subdomains, as is detailed in Figure 1. Differential sensitivity to proteolytic cleavage at homologous sites between proposed domains has been attributed to observed differences in both primary and secondary structures at interdomain boundaries (Ortel et al., 1984).

Local secondary structure within the ceruloplasmin polypeptide chain was determined based on measurements of CD (Noyer and Putnam, 1981) as well as the calculation of parameters predictive of secondary structure
Ortel et al., 1984). On this basis, it has been proposed that ceruloplasmin consists of 33% β sheet organization, 33% β turns and 20% α-helices. Coupled with a calculated hydropathy profile for the protein (Ortel et al., 1984), these data are in accordance with the domain model presented in Figure 1. It is attractive to speculate that different domains within the ceruloplasmin molecule may correspond to its various biological activities (see Section I.B.3).

B.2 Sites of Ceruloplasmin Biosynthesis

Originally, ceruloplasmin biosynthesis was thought to occur exclusively within the parenchymal cells of the liver (Neifakh et al., 1969). However, it has since been demonstrated that the choroid plexus, yolk sac, placenta and testis represent extrahepatic sites of ceruloplasmin synthesis in the rat (Aldred et al., 1987). It has been proposed that ceruloplasmin expression by these tissues may be important in the transport of copper across natural barriers existing between these compartments (i.e. blood/cerebrospinal fluid, maternal/fetal circulation, and blood/testis barriers) (Aldred et al., 1987). Synthesis of the transport proteins transferrin and transthyretin in these tissues (Dickson et al. 1985; Schreiber, 1987) further implicates the importance of transport proteins at the interface between extracellular compartments. Furthermore, the synthesis of transferrin in these tissues, coupled with the known ferroxidase activity of ceruloplasmin (see Section I.B.3) may be important in the transport of iron across compartment barriers. Similarly, the ability of ceruloplasmin to oxidize serotonin and various catecholamines (see Section I.B.3) may be of physiological significance in
the regulation of cerebrospinal fluid composition. Using the technique of \textit{in situ} histohybridization, Yang \textit{et al.} (1986) have also observed ceruloplasmin mRNA within circulating macrophages and lymphocytes. Interestingly, gene expression of transferrin has also been shown in lymphocytes (Lum \textit{et al.}, 1987) again suggesting coordinated functional roles for these two genetically-linked (Weitkamp, 1983; Yang \textit{et al.}, 1984) plasma proteins.

\textbf{B.3 Functions of Ceruloplasmin}

Ceruloplasmin possesses a number of enzymatic activities which cannot be attributed to a subunit organization, since there is conclusive evidence that ceruloplasmin is synthesized as a single polypeptide chain (see Section I.B.1). Rather, the multifunctional nature of the protein has been ascribed to the catalytic activities of the bound cupric ions (Frieden, 1981; see Section I.B.1). The various functions of ceruloplasmin are detailed below:

\textbf{B.3.1 Ferrooxidase activity} (Curzon and O'Reilly, 1960; Scheinberg and Morell, 1973; Frieden and Hseih, 1976). Although capable of oxidizing a variety of substrates \textit{in vitro}, ferrous [Fe (II)] iron has been proposed as the principal physiological substrate for ceruloplasmin. In this capacity, ceruloplasmin oxidizes Fe (II) released from ferritin to the Fe (III) form, for subsequent binding to apotransferrin [i.e. Fe \textit{ferrooxidase} → Fe (III)-transferrin]. Thus, ceruloplasmin is directly involved in the regulation of hepatic iron mobilization.

\textbf{B.3.2 Serum antioxidant activity} (Al-Timimi and Dormandy, 1977; Goldstein \textit{et al.}, 1979). It is well established that ceruloplasmin can
serve as a scavenger of free radicals and superoxide ions. Frieden (1981) has estimated that the collective radical-scavenging potential of ceruloplasmin in serum is less than that of superoxide dismutase. However, since ceruloplasmin is extracellular in location while superoxide dismutase is primarily intracellular, ceruloplasmin acts as the major scavenger in plasma, particularly during the acute phase response (see below) when ceruloplasmin levels are characteristically increased.

B.3.3 Amine oxidase activity (Peisach and Levine, 1963; Frieden, 1981). Ceruloplasmin possesses significant oxidase activity toward numerous aromatic amines and phenols in vitro. Physiologically, this activity may be important in the regulation of levels of biogenic amines such as serotonin and various catecholamines.

B.3.4 The role of ceruloplasmin in copper transport. The proposal that ceruloplasmin is a copper transport protein is strengthened by experiments demonstrating that a ceruloplasmin molecule will reversibly bind up to ten cupric ions, in addition to the intrinsically-bound, catalytically active copper atoms (McKee and Frieden, 1971). This is consistent with proposals that ceruloplasmin is the primary source of copper for intracellular metalloenzymes present in various extrahepatic tissues (Owen, 1965; Linder and Moor, 1977; Campbell et al., 1981). In this regard, there is evidence that the copper ions of ceruloplasmin are a pre-requisite for copper utilization in the biosynthesis of cytochrome c oxidase (Marceau and Aspin, 1973a,b; Hsieh and Frieden, 1975). It has been proposed that ceruloplasmin Cu (II) is reduced at cell membrane receptors and that Cu (I) is subsequently transferred to an unidentified intracellular acceptor(s) (Frieden, 1981). Alternatively, ceruloplasmin
may be taken up by endocytosis, and Cu (I) may then be released by proteolysis, accompanied by recycling of the protein to the plasma membrane for release (Cousins, 1985). This latter model precludes the necessity of an acceptor for intracellular transport. The net result of either route of copper entry into the cells is that the labile Cu (I) form would be oxidatively transferred to intracellular apoenzymes, where it could then be fixed with the aid of oxygen into holoenzyme-Cu (II) form. Recently, evidence has been presented for the presence of a specific ceruloplasmin receptor in membranes from chicken aorta and cardiac tissues (Stevens et al., 1984). In this study, membrane fragments derived from the latter tissues bound \[ \text{[}^{125}\text{I} \text{]} \] labelled chicken ceruloplasmin with a dissociation constant \( K_d \) of approximately \( 10^{-8} \) M. This is consistent with studies showing that the activation of aortic lysyl oxidase is correlated with elevated plasma ceruloplasmin levels (Harris and DiSilvestro, 1981).

B.4 Regulation of Ceruloplasmin Expression

B.4.1 Hormonal regulation of ceruloplasmin synthesis. Hormonal factors have been shown to influence ceruloplasmin production by the liver (see Cousins, 1985 for a recent review). Meyer et al. (1958) have shown that epinephrine and estradiol increase serum ceruloplasmin levels in the rat, while both adrenocorticotrophic hormone (ACTH) and hydrocortisone have been shown to increase ceruloplasmin levels in chickens (Starcher and Hill, 1965). Based on these observations, it has been proposed that any stress-related change in ceruloplasmin involves adrenal steroids. Although adrenal hormones have historically received particular attention
regarding their effect on ceruloplasmin levels, it has been demonstrated more recently that leukocyte endogenous mediator (interleukin I) can also elevate serum ceruloplasmin levels (Wannemacher et al., 1975). The stimulatory effects of leukocyte endogenous mediator on ceruloplasmin and several other plasma proteins have been proposed in the regulation of the acute phase inflammatory response (see below).

B.4.2 Copper induction of ceruloplasmin expression. Linder et al. (1979) have reported that copper directly controls the plasma concentration of ceruloplasmin in diet-induced copper deficient rats by regulation of its level of synthesis. In these studies, the effects of oral administration of copper to copper-deficient rats was assessed by monitoring the incorporation of a two hour pulse dose of $[^{3}\text{H}]$-leucine into plasma proteins. No effect of copper administration was observed on synthesis of plasma proteins in general. However, a marked effect on ceruloplasmin synthesis was observed, resulting in a significant enhancement (nearly three-fold) after 6 - 8 hours, thereby resembling the effect of iron on ferritin synthesis (Drysdale et al., 1966). It has been proposed that the sudden influx of copper associated with large administered doses may be sufficient to bypass normal control mechanisms, and may either activate ceruloplasmin gene transcription specifically, or may alter some aspect of translational regulation. These data are consistent with the observations of Weiner and Cousins (1980) using rat parenchymal cells. In the latter study, incubation with 50 μM copper for 12 hours or more significantly increased $[^{3}\text{H}]$-ceruloplasmin secretion by the cells suggesting that when extracellular copper content
was sufficiently high, ceruloplasmin gene expression may have been enhanced.

B.4.3 **Regulation of ceruloplasmin synthesis during inflammation - the acute phase response.** The acute phase reactants comprise a group of mainly glycoproteins which show characteristically altered rates of synthesis in the liver (Schreiber et al., 1982) resulting in changes in their plasma concentrations in response to a wide variety of inflammatory stimuli [see Koj (1974) for a review]. Ceruloplasmin is an acute phase reactant (Larson, 1974) and as such, its serum levels can become increased by 2 to 3-fold (from the normal level of 15 – 60 mg/dl serum; Owen, 1982) in response to inflammation. The property of ceruloplasmin as a serum antioxidant is attractive in this respect, since increased ceruloplasmin levels would be useful in the subsequent neutralization of lipid peroxidation products released into the serum upon tissue damage (Bonta, 1978).

The observed stimulatory effect of leukocyte endogenous mediator (LEM) on hepatocytes has been proposed to result in increased ceruloplasmin synthesis during the inflammatory response (Frieden, 1981) (see above). This protein, which is released by leukocytes, stimulates the uptake of iron, zinc, and amino acids by liver cells and also enhances the synthesis and release of acute phase reactants including ceruloplasmin. The mechanism underlying the stimulatory effects of LEM is not understood.

Ceruloplasmin expression during the acute phase response has recently been studied by analysis of rat liver samples following induction of inflammation by injection of the animals with turpentine (Aldred et al.,
Ceruloplasmin mRNA levels increased to a peak corresponding to 350% of the normal value by 36 hours. By 60 hours post-inflammation, ceruloplasmin has decreased to normal levels. As has been the case for other acute phase reactants studied (transferrin, $\alpha_2$-macroglobulin, $\beta$ chain of fibrinogen, $\alpha_1$-acid glycoprotein, and metallothionein-I) (Schreiber et al., 1986), the regulation of ceruloplasmin synthesis during the inflammatory response appears to occur at the mRNA level, affecting the rate of transcription and/or mRNA stability as opposed to the rate of protein translation.

C. ABNORMALITIES IN COPPER HOMEOSTASIS — WILSON'S DISEASE

In this autosomal recessive disorder (also referred to as hepatolenticular degeneration), pathogenesis is related to abnormal copper deposition in body tissues, especially the brain and liver (Cumings, 1948). Ceruloplasmin levels are characteristically decreased in Wilson's disease. This was initially demonstrated by Scheinberg and Gitlin (1952) who isolated ceruloplasmin from normals and Wilson's disease patients and quantified its levels both immunochemically, and by monitoring the decrease in absorbance (610 nm) of ceruloplasmin following reduction. The observed correlation between decreased ceruloplasmin levels and Wilson's disease has since been interpreted to suggest a defect in the rate of ceruloplasmin biosynthesis (Poulik and Weiss, 1975). Additionally, there have been reports that the disease is associated with structural anomalies of the ceruloplasmin protein. In this regard, Verbina and Puchkova (1985) have reported the isolation of ceruloplasmin from a Wilson's disease patient that differs from normal ceruloplasmin in physicochemical,
immunological, and catalytic properties. They have postulated that this anomalous ceruloplasmin may be the result of incorrect post-translational modification of the protein.

Gaitskhoki et al. (1975) reported comparative immunochemical analysis (using $^{125}\text{I}$-ceruloplasmin antibodies) of ceruloplasmin-synthesizing polysomes in liver biopsies obtained from control subjects and Wilson's disease patients. This study clearly demonstrated that the amount of ceruloplasmin-forming polysomes in patients affected with Wilson's disease was 10 - 20 times lower than that determined for normal controls. On this basis, it was proposed that a decreased level of translatable mRNA is the likely cause of the genetic block in ceruloplasmin synthesis that is characteristic of Wilson's disease. This hypothesis is consistent with the recent findings of Czaja et al. (1987). The latter study showed that ceruloplasmin mRNA levels in five Wilson's disease patients were decreased to 33% that of control mRNA levels. In contrast, levels of albumin mRNA (also synthesized in the liver) were elevated to 161% in Wilson's disease patients compared with normal levels, suggesting that liver function in the former group was not compromised due to the disease state. Using nuclear run-on assays to analyze transcriptional rates, it was found that the amount of ceruloplasmin gene transcription in the Wilson's disease patients was decreased to 44% that of control levels. This has been interpreted to indicate that in at least some cases of Wilson's disease, observed reduction of plasma ceruloplasmin in levels may be due to decreased ceruloplasmin gene transcription, as opposed to a defect in the rate of protein synthesis.
The causal relationship between ceruloplasmin and Wilson's disease has become complicated by the recent studies of Frydman et al. (1985), showing linkage between the gene for Wilson's disease and the esterase D locus on chromosome 13. These data have been further substantiated by the more recent assignment of the Wilson's disease locus to chromosome 13q (Yuzbasiyan-Gurkan et al., 1987). Since the wild-type ceruloplasmin gene has been unequivocally mapped to chromosome 3 using both somatic cell and in situ hybridization techniques (Yang et al., 1986; Royle et al., 1987), it seems likely that reduced ceruloplasmin-specific transcription observed in Wilson's disease patients studied by Czaja et al. (1987) is not the result of defect(s) within the structural gene, but may involve a trans-acting factor(s) mapping to chromosome 13.

D. CHARACTERIZATION OF RAT CERULOPLASMIN

Gaitskhoki et al. (1980) have previously reported the isolation of highly-purified rat ceruloplasmin mRNA by indirect immunoprecipitation of ceruloplasmin-synthesizing rat liver polysomes. Ceruloplasmin mRNA was then translated in a heterologous wheat germ system; products of cell-free translation were immunoprecipitated and subjected to SDS-polyacrylamide gel electrophoresis. The results indicated that the primary translation product was 84 Kda. Addition of rat liver membranes to the translation system resulted in the visualization of two polypeptides (80 Kda and 65 Kda) following immunoprecipitation. The 65 Kda product was identical to the secreted form of ceruloplasmin isolated from the Golgi complex following in vivo pulse-labelling and subcellular fractionation (Neifakh et al., 1979). On this basis it was proposed that a maturation step,
resulting in the appearance of the 130 Kda plasma form of ceruloplasmin, involved ligation of two 65 Kda polypeptides (Puchkova et al., 1981). In support of this theory, Prozorovski et al. (1982) have emphasized the existence of sequence homology within each half of human holoceruloplasmin based on structural studies using [125I]-peptide mapping. However Takahashi et al. (1984) have conclusively demonstrated a triplicated structure for the protein (see Section I.B.1). More recently, it has been shown that a single mRNA species of approximately 3.8 Kb corresponds to the ceruloplasmin transcript from rat liver, which is more than sufficient in size to encode the entire protein (Aldred et al., 1987).

Aldred et al. (1987) have also reported the isolation of a rat ceruloplasmin cDNA clone. Using a partial human ceruloplasmin cDNA clone to screen a rat liver cDNA library, a clone was isolated that contained DNA coding for the equivalent of residues 194 - 879 of the human ceruloplasmin amino acid sequence (Takahashi et al., 1984). The predicted amino acid sequence derived from the rat cDNA shows close sequence identity (nearly 75%) with the determined amino acid sequence of human ceruloplasmin from amino acid residues 194 - 276 (i.e. within the A1 domain). There is striking (approximately 98%) amino acid sequence identity in the carboxy-terminal region of this latter sequence when compared with the human ceruloplasmin protein sequence, showing complete conservation from residues 227 - 276 with the exception of one conservative amino acid change occurring at residue 243. The predicted rat ceruloplasmin amino acid sequence also closely matches that determined for human ceruloplasmin between amino acids 810 - 879 (corresponding to the A3 domain), showing 83% sequence identity in this region.
E. THE RELATIONSHIP OF CERULOPLASMIN TO OTHER COPPER-CONTAINING PROTEINS

The amino acid sequence of human ceruloplasmin (Takahashi et al., 1984) was analyzed for segments similar to known type I copper-binding sites in small blue electron transfer proteins (e.g. azurin and plastocyanin), and in some large multicopper oxidases whose crystallographic structures and/or amino acid sequences have been determined.

The three-dimensional structure of azurin (a bacterial electron transport protein with Mr = 14 Kda) has been determined from a 3.0 Å resolution electron density map using X-ray diffraction methods (Adman et al., 1978). Based on these data, the single type I copper ion present in azurin is thought to be coordinated to cys-112, met-121, his-46 and his-117 residues. This is consistent with a distorted tetrahedral geometry for that type I site that has been proposed based on near-infrared absorption, CD, and MCD studies (Solomon et al., 1976).

Like azurin, plastocyanin (Mr = 10,500) also belongs to the group of small blue electron transfer proteins and possesses a single type I blue Cu (II) ion. Plastocyanin has been identified as a component in the photosynthetic chain of a number of green plants and algae. Its X-ray crystal structure at 2.7 Å resolution has been reported by Colman et al. (1978). As is the case for azurin, the coordination geometry of the type I Cu (II) is consistent with a distorted tetrahedron, liganded by a cysteine thiol group (cys-84), a methionine thioether group (met-92) and histidine imidazole groups (his-37 and his-87). When compared to the type I centres proposed for azurin and plastocyanin, the type I copper present
in stellacyanin (a small blue protein with proposed electron transfer function) differs with respect to at least one coordination position. This is based on the reported amino acid sequence of stellacyanin (Bergman et al., 1977) which is devoid of methionine residues. This may explain observed differences in the type I centres in stellacyanin and plastocyanin as determined by EPR analysis (Peisach et al., 1967) and a comparison of redox potentials.

On the basis of homology with the identified ligands to the single type I copper in azurin and plastocyanin (see above), it has been suggested that a type I copper centre is present in the carboxy terminal 19 Kda proteolytic fragment of ceruloplasmin (Kingston et al., 1979; see Figures 2A and 2B). Three of the proposed ligands are the clustered cys-1021, his-1026, and met-1031 residues; the fourth type I ligand suggested by Kingston et al. (1979) is his-956, which is 65 residues amino terminal to the cys-1021 residue. The predicted cysteine, histidine and methionine ligands are in agreement with the three dimensional model of the copper active site in human ceruloplasmin presented by Rydén (1982). However, the latter study indicates that the fourth ligand is his-975 which is more likely, since this position corresponds to the location of type I histidine ligands in both plastocyanin and azurin (see Figure 2A). It has been postulated further that a second type I copper binding site may be located in the 50 Kda proteolytic fragment (Dwulet and Putnam, 1981b), where the ligands cys-680, his-685 and met-690 are located in homologous positions to the proposed type I site in the 19 Kda fragment. The fourth residue comprising the type I site in this fragment is proposed to be his-637 (Dwulet and Putnam, 1981b), which also corresponds to the
Figure 2. The relationship between human ceruloplasmin (Cp), *Neurospora crassa* laccase (Lac), *Pseudomonas aeruginosa* azurin (Azn), poplar plastocyanin (Pl), and bovine superoxide dismutase (B.SOD) in two regions (Figures 2A and 2B) containing proposed copper ligands (taken from Dwulet and Putnam, 1981b; Germann and Lerch, 1986). Identical amino acid residues are boxed, and potential ligands to the three types of copper centers are indicated by *1, *2 and *3 respectively. Arrows identify known type I ligands in azurin and plastocyanin (see text for details). Histidine ligands implicated in copper-binding in bovine superoxide dismutase are circled. The fourth potential type I ligand in laccase (i.e. Met-169) is enclosed in a diamond. Numbers on the left of each sequence identify positions within the proteins of the first residues given. For human ceruloplasmin, the corresponding homology unit is given in brackets.
position of type I histidine ligands identified in azurin and plastocyanin (see Figure 2A).

Although the 67 Kda fragment, representing the amino terminal portion of ceruloplasmin has cysteine and histidine residues in identical positions to those present in the 19 and 50 Kda fragments, there is no corresponding methionine residue (see Figure 2B), which may preclude the binding of a type I copper ion in this region (Takahashi et al., 1983). However, the type I copper centers in ceruloplasmin have been differentiated into two subtypes, based on the kinetics of reoxidation (Hervé et al., 1978). In this regard, it has been shown more recently that the type I "fast" copper center does not require methionine as a coordination ligand, while the type I "slow" center does (Hervé et al., 1981). Thus, the 67 Kda proteolytic fragment also possesses a putative type I copper binding site which may resemble the blue copper center in stellacyanin which also lacks methionine as a ligand (see above).

A second region within the 19 Kda proteolytic fragment of ceruloplasmin is homologous to a known non-blue copper binding site in bovine and human superoxide dismutase (Richardson et al., 1975; Jabusch et al., 1980) and human cytochrome c oxidase (Barrell et al., 1979), all of which contain copper in binuclear centres. The X-ray crystal structure of bovine superoxide dismutase has been determined at 3 Å resolution. The two copper ions on opposite subunits within the dimer are 34 Å apart. The copper and zinc in each subunit are approximately 6 Å apart, and they both form ligands to the imidazole ring of his-61. The protein ligands to the copper are proposed to be his-44, his-46, his-61 and his-118 (i.e. his-X-his motif), arranged in a slightly distorted square plane and
thereby resembling a type II centre. The 19 Kda fragment of ceruloplasmin has a histidine-rich sequence element (see Figure 2A) that is homologous to the non-blue copper centre described in bovine and human superoxide dismutase (Ortel et al., 1984). Based on a three-dimensional model of the copper active site of the human ceruloplasmin 19 Kda proteolytic fragment (Rydén, 1982), it has been proposed that these histidine residues (i.e. his-980 and his-982, in addition to his-978 and his-1020) (see Figures 2A and 2B) may function as type III ligands in ceruloplasmin. The latter model identifies human ceruloplasmin residues his-1022 and his-1028 as potential type II ligands (see Figure 2B). A corresponding histidine-rich cluster resembling that identified in the 19 Kda fragment of ceruloplasmin is absent from the 50 Kda proteolytic peptide (Ortel et al., 1984) (see Figure 2A). As a general observation, there is an unusually large number of his-X-his sequences in ceruloplasmin, some of which may be implicated in non-blue copper binding.

Fungal laccase (Mr = 62 Kda) is a blue oxidase containing four copper ions and a unique cysteine residue. Initially, peptides containing the single sulfhydryl group were isolated and characterized (Briving et al., 1980). Additional amino acid sequence information has been provided from partial nucleotide sequence analysis of the laccase gene (Germann and Lerch, 1986). Comparison of the available amino acid sequence for laccase with that of human ceruloplasmin has revealed the presence of several highly-conserved sequence elements in these two multicopper oxidases (see Figures 2A and 2B). Due to its conserved position with proposed type I cysteine ligands in ceruloplasmin, azurin, and plastocyanin (see Figure 2B), the unique cysteine residue in laccase
is thought to coordinate type I copper. Additionally, laccase contains two histidine residues in similar positions to the proposed type I histidine ligands in the above copper-containing proteins (see Figures 2A and 2B). However, the methionine ligand which is conserved in the latter proteins is absent in laccase, as is the case for stellacyanin (see above) (Germann and Lerch, 1986). It has been proposed (Germann and Lerch, 1986) that met-169 in laccase (see Figure 2B) may be involved in type I copper coordination. In addition to similarity observed with respect to type I ligands, histidine residues have been identified in laccase that occur in identical positions to proposed non-blue copper ligands in ceruloplasmin (see Figures 2A and 2B).

Based on earlier studies (Richardson et al., 1975; Rydén, 1982), it has been suggested that the conserved sequence elements in ceruloplasmin and laccase may coordinate the binding of type II and/or type III copper, in addition to their proposed involvement in type I copper centres (Briving et al., 1980; Germann and Lerch, 1986). Thus, these sequences may form a link between type I, II and III copper centres present in multicopper oxidases.

Taken together, the above data suggest that the type I copper binding site is similar in both the small blue electron transfer proteins as well as large multicopper oxidases (e.g. laccase and ceruloplasmin). It also seems that in addition to the presence of a putative type I copper centre, the 19 Kda fragment of ceruloplasmin is structurally related to non-blue copper binding sites identified in other copper oxidases. It is likely, therefore, that these proteins share a common evolutionary origin, perhaps
deriving from a primordial gene encoding a small blue protein possessing either electron transfer or oxidase function.

There are several sites within the ceruloplasmin molecule that are characterized by the positioning of histidines adjacent to basic amino acid residues. These sites might be involved in copper binding as is the case for serum albumin, or the plasma tripeptide gly-his-lys (Pickart et al., 1980). The first two residues of the latter motif are involved in the binding of copper, while the side chain of the lysyl residue is proposed to be necessary for the recognition by cell surface receptors. This is analogous to plasma albumin or α-fetoprotein, in which the histidine residue that binds copper is immediately adjacent to either lysine or arginine residues (Aoyagi et al., 1980). Thus, the positioning of a histidine residue next to a basic residue may be a biologically active structure for copper uptake. The presence of such sequences in ceruloplasmin may correspond to observed sites of reversible copper binding (McKee and Frieden, 1971) which may in turn play an essential role in copper transport (see Section I.B.3.4).

F. THE RELATIONSHIP BETWEEN CERULOPLASMIN AND PROTEINS INVOLVED IN BLOOD COAGULATION

An interesting structural relationship has been shown to exist between ceruloplasmin and the blood coagulation factors V and VIII. These latter two proteins (both with Mr > 300 Kda) share a high degree of structural and functional similarity (Suzuki et al., 1982; Nesheim et al., 1984). Both proteins function in the intrinsic blood clotting cascade (Jackson and Nemerson, 1980) in conjunction with an activated, vitamin-K
dependent clotting factor (factors IXa and Xa for factors VIII and V, respectively). Both complexes require a phospholipid surface and calcium ions, and subsequently result in the specific activation of a second vitamin K-dependent coagulation protein (factor X and prothrombin for factors VIII and V, respectively). Analysis of the complete amino acid sequences of human factor V (Jenny et al., 1987) and human factor VIII (Wood et al., 1984; Toole et al., 1984) predicted from the corresponding cDNAs revealed the existence of 3 types of domains within the two proteins: a triplicated "A" domain (approximately 320 - 380 amino acid residues), a unique "B" domain (925 and 886 amino acid residues in factors VIII and V, respectively), and a duplicated "C" domain consisting of approximately 100 - 150 amino acid residues (see Figure 3). Organization of these units from amino-to carboxyl terminal within the proteins are as follows: A1 - A2 - B - A3 - C1 - C2. The "A" domains of factors V and VIII show a high level of similarity with the triplicated units in the human ceruloplasmin molecule, sharing approximately 30 - 40% sequence identity when compared pairwise (see Table I). Of particular note is the clustering of cysteine residues at similar positions within the triplicated "A" domains of factor VIII (Vehar et al., 1984) and ceruloplasmin (Takahashi et al., 1983), indicating a high degree of structural conservation between these repeated units. The duplicated "C" domain present in factors V and VIII is unrelated to ceruloplasmin, but shows approximately 20% sequence identity when compared with discoidins, which are phospholipid-binding lectins from Dictyostelium discoideum (Poole et al., 1981). The "B" domains in factors V and VIII are each
Figure 3. Comparison of the structural organization of ceruloplasmin, factor V and factor VIII.

The triplicated A domain (designated A1, A2 and A3 in an amino-to-carboxyl direction) is identified in the 3 molecules by cross-hatched bars. The B domain (present in factors V and VIII only) is represented by an open bar. The duplicated C domain (also present only in factors V and VIII) is shown by stippled bars. Sizes of the domains correspond to amino acid residues.
Comparison of the structures of factor VIII, factor V and ceruloplasmin.

Factor VIII:
- 350 350 1000 350

Ceruloplasmin:
- 350 350 350

Factor V:
- 350 350 1000 350

Amino acids:
- 150 150
Table I. Comparison of "A" domains in factor V, factor VIII and ceruloplasmin (from Jenny et al., 1987).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V-A1</td>
<td>31</td>
<td>32</td>
<td>36</td>
<td>30</td>
<td>27</td>
<td>31</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>V-A2</td>
<td>-</td>
<td>29</td>
<td>29</td>
<td>44</td>
<td>31</td>
<td>35</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>V-A3</td>
<td>-</td>
<td>-</td>
<td>33</td>
<td>31</td>
<td>39</td>
<td>32</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>VIII-A1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30</td>
<td>32</td>
<td>34</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>VIII-A2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>33</td>
<td>38</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>VIII-A3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>34</td>
<td>36</td>
</tr>
<tr>
<td>Cer-A1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>Cer-A2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40</td>
</tr>
</tbody>
</table>

Values (expressed as percentages) represent total identical amino acid matches divided by overlapping lengths (including gaps).
structurally unique when compared with the rest of the molecule, and in both cases are removed upon activation of the proteins.

The functional relationship between ceruloplasmin and these two seemingly unrelated blood clotting factors is unclear. However, it has been proposed previously that factor V is a metalloprotein (Greenquist and Colman, 1975) requiring calcium and a variety of metal ions for expression of activity (Esmon, 1979; Hibbard and Mann, 1980). Recently, using both atomic emission and atomic absorption spectroscopy, factor V has been shown to contain copper in the ratio of 1 copper ion per mol of factor V (Mann et al., 1984). Since factor V exhibits an absorption spectrum with no maximum at either 310 nm (type III copper) or 610 nm (type I copper), the copper binding is assumed to be the equivalent of the type II copper center present in ceruloplasmin. Although factor VIII has not yet been assessed with respect to its metal-binding content, the peptide chains are non-covalently associated in a process that is EDTA sensitive (Fass et al., 1982). Thus, the association of the peptide chains in the factor VIII molecule may be dependent upon bound metal and/or calcium ions. The four amino acids proposed as ligands for type I copper binding in the ceruloplasmin molecule (see Section I-E) are located in analogous positions in the A1 and A3 units of factor VIII (Vehar et al., 1984). Conservation of these copper ligand residues may imply similar metal binding characteristics for both factor VIII and ceruloplasmin.

In terms of physiological significance, the demonstrated structural relationship between factors V and VIII and ceruloplasmin suggests the possible involvement of copper and/or other metal ions, in addition to calcium, in the function of factors V and VIII. The possible role of
metal binding in this process is implied by studies showing the binding capacity of $\gamma$-carboxyglutamic acid residues for lanthanide ions (Sperling et al., 1978). These modified glutamic acid residues (present in blood-clotting factors II, VII, IX, X, protein Z, protein S and protein C) are thought to be involved in calcium binding, which is in turn proposed to mediate the interaction of these proteins with platelet surfaces in vivo. It is also conceivable that potential metal ligands present in factor V and possibly factor VIII can interact with a different metal ion jointly bound by the $\gamma$-carboxyglutamic acid regions of factors IX and X, thereby promoting complex formation.

Ceruloplasmin has a number of enzymatic functions ascribed to it in addition to its role in copper transport and homeostasis (see Section I.B.3). It is unknown at this time whether any of these catalytic activities are also associated with factors V and VIII.

G. CHARACTERIZATION OF THE HUMAN FACTOR VIII GENE

G.1 Historical Perspective

In the majority of cases, the bleeding disorder hemophilia A (or classic hemophilia) results from mutations (either single base changes or gross rearrangements) within the structural gene coding for factor VIII. Since the factor VIII gene is located on the X chromosome, there is a high frequency of hemophilia A relative to autosomal clotting disorders (Haldane, 1935). Historically, studies addressing the nature of hemophilia A have been seriously hampered by difficulties encountered in the purification of factor VIII since it is an unusually large (Mr \(\approx 330\) Kda), unstable protein, present in low concentrations in plasma (100 -
200 ng/ml) (Wood et al., 1984). Studies involving the characterization of the factor VIII gene were initiated to facilitate an understanding of the molecular basis of hemophilia A. In addition, expression of the cloned gene in vitro would provide a virus-free preparation of recombinant factor VIII for treatment of hemophiliacs.

G.2 Organization of the Factor VIII Gene

The complete organization of the 186 Kbp human factor VIII gene was reported by Gitschier et al. (1984). Initially, factor VIII clones were isolated from a genomic library constructed using a lymphoblast cell line which was derived from an individual with 4X chromosomes (Karyotype 49, XXXY). This library was screened using a unique 36-base synthetic oligonucleotide probe constructed on the basis of a previously characterized factor VIII tryptic peptide (Vehar et al., 1984). Using chromosome walking to extend the initially identified clones, 200 Kbp (nearly 0.1%) of the X chromosome was characterized that encompassed the complete factor VIII gene.

DNA sequence analysis of intron/exon boundaries revealed the presence of 26 exons in the factor VIII gene. Although most exon sizes were consistent with reported distributions (Naora and Deacon, 1982), two of the exons are unusually large. The largest exon is 3106 bp in length, and corresponds to the 100 Kda connecting peptide (i.e. the B domain; see Figure 3) joining amino- and carboxy-terminal fragments of 90- and 80 Kda, respectively. This exon corresponds to a physiological unit, since the 100 Kda peptide is proteolytically excised upon thrombin activation (Fulcher et al., 1983). The other large exon is 1958 bp long, 1805 bp of which correspond to the 3' untranslated region of the gene. Intron sizes
in the factor VIII gene were found to be highly variable (Naora and Deacon, 1982) with the largest intron spanning 32.4 Kbp. Overall, the factor VIII gene consists of 9 Kbp of exon sequence interrupted by 177 Kbp of intervening sequence, suggesting a lack of selective pressure to decrease intron size.

RNAse protection experiments using mRNA derived from either the AL-7 T cell hybridoma line or human liver indicate that the transcription initiation site in the factor VIII gene is positioned at -170 or -172 respectively (+1 denotes the position of the initiator methionine residue). At 30 bp 5' to the predicted mRNA start site is located the sequence "GATAAA", which closely resembles the Goldberg-Hogness consensus sequence (i.e. the "TATA" box), proposed to be required for precise initiation of transcription by eukaryotic RNA polymerase II (Goldberg, 1979; Breathnach and Chambon, 1981). No "CAT" sequence element (Breathnach and Chambon, 1981) was observed upstream to the "ATA" sequence in the factor VIII gene.

Following the 5' untranslated region is a typical 19 amino acid signal peptide (von Heijne, 1982) containing two charged residues flanking a core of hydrophobic amino acids. This secretory signal precedes the mature protein sequence of 2332 amino acid residues, which is followed by a "TGA" stop codon and a subsequent 3' untranslated region of 1802 bp. The conserved polyadenylation signal "AATAAA" (Proudfoot and Brownlee, 1976) is contained in the latter sequence, occurring 19 bp prior to the position of poly (A) addition.
G.3 Evolutionary Aspects of Intron Positions within the Factor VIII Gene

As discussed previously (see Section I.F), the factor VIII protein is composed of three different domains; the order of the domains in the protein is A1 – A1 – B – A3 – C1 – C2 (see Figures 3 and 4). If tandem gene duplication events (see Section I.I.1) have occurred in the evolution of the factor VIII gene, as is strongly suggested by the repeated domain organization, conservation of intron boundaries within the A and C repeats would be predicted (Doolittle, 1985). For the C duplication, intron/exon boundaries occur precisely at the borders of the C1/C2 repeat units, as would be expected if a gene duplication event has occurred (see Figure 4). Again, there is an intron at the boundary of the A3 and C1 units (see Figure 4) thus also supporting a mechanism involving intron joining. However, the A1/A2 and A2/A3 junctions are each contained on one exon (see Figure 4).

Within the A and C repeated units, only some of the intron boundaries are conserved, suggesting that these introns were present in the ancestral gene prior to duplication (see Figure 4). The differing number of exons within each of the repeats is reflective of either intron loss or intron insertion following the initial duplication events.

The origin of the unique B domain is highly speculative. This region is contained almost entirely within a 3106 bp exon as described previously, where the end of the A2 repeat and the beginning of the A3 repeat are also found. Due to its anomalous size, it has been postulated that the B domain may have arisen by insertion of a processed gene (mRNA-derived; see Section I.J) into a short exon containing the A2/A3
Figure 4. **Location of introns (vertical lines) within the triplicated A and duplicated C domains of human factor VIII** (from Gitschier et al., 1984).

For the A and C repeated units, numbers identify the position of the first amino acid residue in each line. The location and extent of the B domain is indicated; numbers represent amino acid residues. The exons are numbered consecutively.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1649</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2173</td>
<td></td>
<td></td>
<td>24</td>
<td>25</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

100 Amino Acids
boundary. Although the exon containing the B domain corresponds to a functional unit, which is excised upon thrombin activation of factor VIII, correspondence of other exons to functional units of the protein is unclear at this time.

H. THE DYNAMICS OF PROTEIN AND GENE EVOLUTION

H.1 The Molecular Clock

Initial studies involving the comparison of protein sequences from species whose times of evolutionary divergence are established has allowed an estimate of the rate at which mutations have been accumulating in genomes. On this basis, it was suggested that random mutation of DNA occurs at a nearly constant rate, which is in turn manifested as a constant rate of amino acid substitution (Zuckerkandl and Pauling, 1965; Wilson et al., 1977). Thus, protein sequences can serve as approximate molecular clocks. This empirical finding has been useful in the construction of phylogenetic trees (Wilson et al., 1977; Li et al., 1985), allowing major contributions to our knowledge of evolutionary relationships among organisms. However, if the function of a protein is modified (as can occur with independent products of gene duplication events), its evolutionary rate will likely change due to the different set of selective pressures. Even if a protein maintains a constant function, the rate of evolution may still be subject to change in response to variation within the organism's cellular environment.

Observed differences in the evolutionary rates of different proteins likely depend upon differences in probability that mutations will retain protein function (Afinsen, 1959; Ohno, 1970). This concept of "functional
constraint" also illustrates why the rate of evolution is variable at different sites within a particular protein, depending upon which areas can tolerate variation without resulting in loss of function (Wilson et al., 1977). Thus, the identification of highly conserved regions within a protein allows the detection of potential functional domains within a molecule.

I. MECHANISMS OF GENE EVOLUTION

1.1 Gene Duplication

The process of gene duplication has been used to explain the occurrence of multigene families either having similar functions, thereby allowing them to act synergistically (e.g. the globin gene family; Edgell et al., 1983) or differing in function (e.g. the lysozyme-lactalbumin family; Hall et al. 1982).

Increased protein size compared with ancestral forms can be the result of internal gene duplication (Doolittle, 1985; Li, 1983). In such events, the entire ancestral molecule can be duplicated or triplicated as is the case for transferrin (MacGillivray et al., 1982) or ceruloplasmin (Takahashi et al., 1983), respectively. Alternatively, portions of proteins can be duplicated to generate larger, more complex forms (Doolittle, 1985). Internal gene duplications are reflected by homologous amino acid sequences and/or similar three-dimensional structures between duplicated regions (McLachlan, 1979).

1.2 Gene Fusion

By the process of gene fusion, new proteins are created by the joining of protein domains derived from different sources (Doolittle,
1985). Coupled with gene duplications, gene fusion events have been proposed in the formation of complex molecules, such as the coagulation factors (Doolittle, 1985).

### I.3 Exon Shuffling

It has been proposed that exons are structural motifs which can be assorted by recombination within intervening sequences to yield novel proteins with different functions (Gilbert, 1978; Blake, 1983a,b; Rogers, 1985). This phenomenon of modular unit shuffling is thought to have been a dominant force in the evolution of the serine protease gene superfamily (Rogers, 1985; Neurath, 1985). For example, the gene for tissue plasminogen activator provides the first example of exon transfer between otherwise unrelated genes (Ny et al., 1984).

### I.4 Intron Insertion and Intron Sliding

Intron insertion has been commonly observed between related genes, such as within the protease domain of the trypsin gene family (Rogers, 1985). Some regularity with respect to the insertion of introns has been shown. For example, it seems that new insertions tend to occur near the middle of pre-existing exons, thus allowing evolution toward a consistency with respect to exon size (Naora and Deacon, 1982; Lonberg and Gilbert, 1985). Secondly, it has been suggested that small observed length variations between members of protein families can be the result of substitution of alternative intron splice sites, thereby permitting extension or contraction of exons at intron junctions (Craik et al., 1982a,b; Craik et al., 1983). Resulting small insertions (usually 2 to 17 amino acid residues) often map to the surface of molecules, where
resultant structural modifications are the least disruptive to the overall tertiary structure (Craik et al., 1982a,b; Craik et al., 1983). In this process of intron/exon sliding, it is expected that only those variations which result in maintenance of the translational reading frame can be tolerated. The mechanism of intron sliding has been postulated to explain small variations in the length of related gene products in both the dihydrofolate reductase (DHFR) and serine protease gene families (Craik et al., 1983).

J. **PSEUDOGENES**

Higher eukaryotic genes (including human ceruloplasmin) commonly exist in multigene families that contain both functional genes, as well as closely related sequences that have lost the ability to produce a functional product due to mutational changes. These latter sequences have been termed "pseudogenes" (Jacq et al., 1977; Proudfoot, 1980). Pseudogenes fall into two categories: 1) non-processed pseudogenes and, 2) processed pseudogenes; the characteristics of each category are summarized below.

1. **Non-Processed Pseudogenes**

This group (mainly composed of the globin pseudogenes from a variety of species; Vanin, 1983) includes those pseudogenes that have retained the intervening sequences found in their functional counterparts. In the majority of cases, the chromosomal location of these pseudogenes is adjacent to the respective wild-type gene, suggesting that the pseudogene sequences have arisen from gene duplication events (Vanin, 1983).
2. **Processed Pseudogenes**

This second, more abundant category of pseudogenes is represented in many different gene families from a number of mammalian species (see Vanin, 1985 for a recent review). Most processed pseudogenes contain genetic lesions that preclude the production of a functional gene product. These lesions include: 1) the presence of in-frame termination codons as a result of single base substitutions and 2) the insertion or deletion of non-integral nucleotide triplets that cause frameshift mutations, thereby resulting in premature termination of translation. There are four examples, however, of processed pseudogenes that contain no deleterious mutations: the human metallothionein II pseudogene (Varshney and Gedamu, 1984), the rat RC9 cytochrome c pseudogene (Scarpulla, 1984), the mouse L32 ribosomal protein pseudogene rpL32-4A (Dudov and Perry, 1984), and the DHFRψ1 pseudogene (Chen et al., 1982).

In addition to various genetic mutations, processed pseudogenes possess a number of characteristic features, the most distinguishing of which is the loss of introns relative to their functional counterparts. The absence of introns is usually precise - i.e. sequences 5' and 3' to the introns are contiguous. Processed pseudogenes are also characterized by the observed divergence of sequence homology with the wild-type counterparts at points corresponding to the beginning and end of the functional genes. Notable exceptions to this include the human immunoglobulin c (Ueda et al., 1982) and λψ1 pseudogenes (Hollis et al., 1982) and the mouse corticotropin β-lipotropin precursor pseudogene (Notake et al., 1983), which appear to be DNA copies of only a portion of the wild-type mRNA transcripts, as well as the mouse ψα3
pseudogenes which contain additional sequences compared to the functional transcripts (Vanin et al., 1980). Many processed pseudogenes have a poly(A) tract located immediately 3' to the point at which homology between the pseudogene and wild-type gene ceases, and are often characterized by short (7 - 17 bp) direct repeats flanking the pseudogene sequence. Lastly, almost without exception, processed pseudogenes do not share the same chromosomal location as the corresponding functional genes (Battey et al., 1982; Czosnek et al., 1984).

A number of mechanisms were originally proposed for the origin of processed pseudogenes (Vanin, 1985). However, based on the characteristics summarized above for a number of compiled processed pseudogene sequences, it is now generally accepted that processed pseudogenes have arisen from reverse transcription of mature mRNA species. In this model, cDNA copies of corresponding mRNAs are then randomly integrated into the genome, as has been postulated for the dispersion of the human snRNA pseudogenes (Denison et al., 1982; Van Arsdell et al., 1981) and the human Alu family of repetitive sequences (Jagadeeswaran et al., 1981).

K. THE PRESENT STUDY

The analysis of the structure of a number of representatives of different gene families has led to an enhanced understanding of the nature of protein and gene evolution [e.g. the globin gene family (Edgell et al., 1983), the serine protease supergene family (Rogers, 1985), and the immunoglobulin supergene family (Hood et al., 1985)]. Based on their shared structural similarities as demonstrated by amino acid sequence
determination (see Section I.F), it has been proposed that ceruloplasmin
and blood clotting factors V and VIII constitute a gene family (Doolittle,
1985). In 1984, the present investigation was initiated in order to
characterize the human ceruloplasmin cDNA and gene, thereby facilitating a
comparison of the gene organization of this multicopper oxidase to that
reported for human factor VIII (Gitschier et al., 1984; see Section I.G).
Since 1984, several partial human ceruloplasmin cDNA clones have been
reported by other groups (Mercer and Grimes, 1986; Yang et al., 1986).
Details concerning the relevance of these latter studies will be addressed
in the context of subsequent sections.
II. MATERIALS AND METHODS

A. BACTERIAL HOSTS AND MEDIA

The media used for both growth of appropriate bacterial hosts and screening of λ phage clones was NZYC (Maniatis et al., 1982) (10 g NZ amine type A, 2 g MgCl₂, 5 g NaCl, 5 g Yeast Extract, 1 g Casamino Acids per litre, adjusted to pH 7.5 by NaOH addition). Phage libraries were plated on NZYC-agar (1.5% w/v) plates, with an overlay of NZYC-agarose (0.7% w/v). The media used for the growth of bacteria transformed with pUC plasmids was Luria broth (LB) (Maniatis et al., 1982) (5 g Yeast Extract, 10 g Bactotryptone and 10 g NaCl per litre). For the selection of bacteria containing pUC plasmids, clones were plated on LB agar (1.5% w/v) plates supplemented with 50 - 100 µg/ml ampicillin. LB medium was also used for screening the human liver cDNA library constructed in the pKT218 vector, except that tetracycline (12.5 µg/ml) replaced ampicillin as the antibiotic. Bacteria containing phage M13 clones were grown in YT medium (Maniatis et al., 1982) (5 g Yeast Extract, 8 g Bactotryptone, 5 g NaCl per litre); M13 transformants were plated on YT agar (1.5% w/v) plates overlayed with YT containing 0.75% (w/v) agar.

E. coli strains JM101 and JM103 were maintained on minimal media plates, prepared as follows: 3 g of agar in a total of 160 ml dH₂O was autoclaved, cooled to 55°C, and mixed with 40 ml of 5X salts [2.1 g K₂HPO₄, 0.9 g KH₂PO₄, 0.2 g (NH₄)₂SO₄, 0.1 g Na Citrate·7H₂O per 40 ml], 2 ml 20% glucose, 0.2 ml 20% MgSO₄·7H₂O, and 0.1 ml 10 mg/ml thiamine. Bacteria for large-scale plasmid preparations were grown either in LB medium, supplemented with the appropriate antibiotic, or in M9 minimal media (Maniatis et al., 1982).
Table II. Summary of the genotypes of bacterial hosts used in the present study. Vectors utilized in conjunction with each host are also given.

<table>
<thead>
<tr>
<th>Bacterial Strain</th>
<th>Genotype</th>
<th>Reference</th>
<th>Compatible Vector System</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) <em>E. coli</em> MC1061</td>
<td>araD139, Δ(ara, leu)7697, Δlac74, galU−, galK−, hsr−, hsm+, strA</td>
<td>Casadaban and Cohen, 1980</td>
<td>human liver cDNA library in pKT218 (Prochownik et al., 1983)</td>
</tr>
<tr>
<td>b) <em>E. coli</em> K802</td>
<td>hsdR+, hsdM+, gal−, met−, SupE</td>
<td>Maniatis et al., 1982</td>
<td>human genomic library in Charon 4A (Lawn et al., 1978)</td>
</tr>
<tr>
<td>c) <em>E. coli</em> LE392</td>
<td>CF−, hsdR514(k-m) supE4, supF58, lacY1, Δ(lacIZY)6, galK2, galT22, metB1, trpR55, lambda−</td>
<td>Maniatis et al., 1982</td>
<td>human genomic library in EMBL 3 (Frischauf et al., 1983)</td>
</tr>
<tr>
<td>d) <em>E. coli</em> P2 392</td>
<td>P2 lysogen of LE392</td>
<td>Maniatis et al., 1982</td>
<td>human genomic library in EMBL 3 (Frischauf et al., 1983)</td>
</tr>
<tr>
<td>Bacterial Strain</td>
<td>Genotype</td>
<td>Reference</td>
<td>Compatible Vector System</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>-----------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>e) E. coli</strong> RY1088</td>
<td>Δlac U169, supE, supF, hsdR&lt;sup&gt;−&lt;/sup&gt;, hsdM&lt;sup&gt;+&lt;/sup&gt;, met B, trpR, tonA21, proC::Tn5 (pmc9)</td>
<td>Young and Davis, 1983a,b</td>
<td>human liver cDNA library in λgt11. (Young and Davis, 1983a,b)</td>
</tr>
<tr>
<td><strong>f) E. coli</strong> C600 Hfl&lt;sup&gt;+&lt;/sup&gt;</td>
<td>F&lt;sup&gt;−&lt;/sup&gt;, thi&lt;sup&gt;−&lt;/sup&gt;-1, thr&lt;sup&gt;−&lt;/sup&gt;-1, leuB6, lacY1, tonA21, supE44, λ&lt;sup&gt;−&lt;/sup&gt;, Hfl&lt;sup&gt;+&lt;/sup&gt;</td>
<td>Appleyard, 1954</td>
<td>human liver cDNA library in λgt10</td>
</tr>
<tr>
<td><strong>g) E. coli</strong> JM101</td>
<td>Δlacpro, supE, thi&lt;sup&gt;−&lt;/sup&gt;, F&lt;sup&gt;+&lt;/sup&gt;, traD36, proAB, lacIQ, lacZΔM15</td>
<td>Messing, 1983</td>
<td>pUC vectors (Vieira and Messing, 1982) M13 vectors (Messing, 1983)</td>
</tr>
<tr>
<td><strong>h) E. coli</strong> JM103</td>
<td>Δlacpro, supE, thi&lt;sup&gt;−&lt;/sup&gt;, strA, sbcB15, endA, hsdR&lt;sup&gt;−&lt;/sup&gt;, F&lt;sup&gt;+&lt;/sup&gt;, traD36, proAB, lacIQ, lacZΔM15</td>
<td>Messing, 1983</td>
<td>pUC vectors (Vieira and Messing, 1982) M13 vectors (Messing, 1983)</td>
</tr>
</tbody>
</table>
containing 840 ml dH$_2$O, 100 ml 10X salts (7 g Na$_2$HPO$_4$, 3 g KH$_2$PO$_4$, 0.5 g NaCl, 1 g NH$_4$Cl per 100 ml), 10 ml MgSO$_4$·7H$_2$O, 20 ml 20% glucose, 10 ml 0.01 M CaCl$_2$, 20 ml 20% Casamino Acids, 0.2 ml 10 mg/ml thiamine and 0.2 g uridine.

A summary of the genotypes of various bacterial strains used in this study is given in Table II.

B. HYBRIDIZATION PROBES

B.1 Purification and Labelling of Oligodeoxyribonucleotides

Oligodeoxyribonucleotide mixtures were synthesized by Tom Atkinson in the laboratory of Dr. M. Smith, University of British Columbia, using an Applied Biosystems 380 A DNA synthesizer. Crude oligonucleotide preparations were purified through 20% denaturing gels (containing 8.3 M urea) and isolated by reverse-phase chromatography using a C$_{18}$ SEP-PAK (Millipore) column as described by Atkinson and Smith (1984). The oligonucleotides were labelled using [$\gamma$-$^{32}$P]-ATP and T4 polynucleotide kinase (Chaconas and van de Sande, 1980), and unincorporated ATP was subsequently removed by chromatography on G25 Sephadex.

Three pools of heptadecadeoxyribonucleotides were used as hybridization probes for the screening of human liver cDNA libraries:

POOL I: 5'd(TARTARTGTYTTYTCYT)3'

POOL II: 5'd(ATNGCRTGCATYTTRTT)3'

POOL III: 5'd(CCCATNARRTACCARTT)3'

where "R" represents both C and A, "Y" represents T and C, and "N" represents G, A, T and C. The three nucleotide pools are complementary to
the mRNA encoding amino acid residues 1 - 6, 937-942, and 962-967 of ceruloplasmin, respectively, as predicted from the amino acid sequence (Takahashi et al., 1984).

B.2 Nick Translation

Purified DNA fragments or entire plasmids were labeled with $^32$P by nick-translation as described by Maniatis et al. (1975). Approximately 200 - 500 ng of DNA was labelled in 50 µl of reaction mixture, containing 50 mM Tris-HCl pH 7.5, 5 mM MgCl$_2$, 0.05 mg/ml BSA, 10 mM β-mercaptoethanol, 20 µM dGTP, 20 µM dTTP, 1.4 mM dATP, 1.4 mM dCTP, 1.4 µCi/µl [α$^32$P]-dATP (3000 Ci/mMole), 1.4 µCi/µl [α$^32$P]-dCTP (3000 Ci/mMole), 0.2 mM CaCl$_2$, 1 pg/µl DNase I (Sigma) and 0.4 u/µl E. coli DNA polymerase I (Kornberg fragment) (Boehringer-Mannheim). The above reaction was incubated for 60 - 120 min at 15°C, and was subsequently terminated by the addition of 3 volumes of 1% SDS/10 mM EDTA containing 25 µg tRNA, followed by heating for 10 min at 68°C. Free triphosphates were separated from the labelled strands by chromatography on 1.0 ml spun columns of Sephadex G-50 (Maniatis et al., 1982). Specific activities of resultant probes ranged from 0.5 - 1.0 × 10$^8$ cpm/µg.

B.3 Klenow Labelling

DNA was also labelled using the method described by Feinberg and Vogelstein (1983). Reaction mixtures typically contained 50 - 200 ng of DNA (either purified restriction fragments or entire plasmids) in a total volume of 50 µl. The DNA sample (in 30 µl of dH$_2$O) was denatured by boiling for 3 minutes followed by rapid cooling on ice for 5 minutes. The labelling reaction was subsequently carried out in 50 µl total volume,
containing 50 mM Tris-HCl pH 8.0, 10 mM MgCl₂, 10 mM β-mercaptoethanol, 20 μM dCTP, 20 μM dGTP, 20 μM dTTP, 1 μCi/μl [α³²P]-dATP (3000 Ci/mMole), 200 mM Hepes pH 6.6, 60 OD 260 nm/ml random hexadeoxyribonucleotides [p(dN6)] (P-L Biochemicals), 0.4 mg/ml BSA and 0.1 U/μl E. coli DNA polymerase I (Klenow fragment) [Bethesda Research Laboratories (BRL) or P-L Biochemicals]. Extension was allowed to continue at 37°C for either 3 - 4 hours or overnight. The reaction was terminated by heating the sample at 68°C for 10 minutes, and unincorporated nucleotides were separated as described for nick-translation (see above). Typically, specific activities of resultant probes ranged from 2 - 5 x 10⁸ cpm/μg.

B.4 Preparation of M13 Probes

M13 templates containing ceruloplasmin DNA fragments were also used to generate probes with high specific activities (Russnak and Candido, 1985). An annealing mixture containing 3.0 μl template (0.5 - 1 μg), 2.0 μl universal primer (P-L Biochemicals; 0.03 A₂₆₀ units/ml) and 2.0 μl of 10x annealing buffer (100 mM Tris-HCl pH 7.5, 600 mM NaCl, 70 mM MgCl₂), was incubated at 65°C for 15 minutes in a 1.5 ml microfuge tube. After cooling to room temperature, 1.0 μl of 20 mM DTT, 2.0 μl of 0.5 mM dGTP, 2.0 μl of 0.5 mM dTTP, 2.5 μl each of [α³²P]-dATP and [α³²P]-dCTP (25 μCi; 3000 Ci/mMol) and 0.5 U of E. coli DNA Polymerase I (Klenow fragment) were added. The reaction was allowed to proceed for 10 minutes at room temperature, and was then followed by a 5 minute chase initiated by the addition of 2.0 μl of 0.5 mM dGTP and 2.0 μl of 0.5 mM dCTP before termination of the reaction by
heating at 68°C for 10 minutes. Unincorporated dNTPs were separated from labelled strands as described for nick translation (see above).

C. IDENTIFICATION OF cDNAS FOR HUMAN CERULOPLASMIN

C.1 Screening of a Human Liver cDNA Library

An adult liver cDNA library (Prochownik et al., 1983) was kindly provided by Dr.S.H. Orkin (Children's Hospital Medical Center, Boston). This library contains cDNA inserts of > 500 bp inserted into the Pst1 site of pKT218 by homopolymeric dG•dC tailing. The cDNA library was screened using the colony hybridization method of Grunstein and Hogness (1975). Ampicillin-resistant colonies (approximately 5,000 per 100 x 15 mm petri plate) were transferred to nitrocellulose filters (82 mm; BA-85, Schleicher and Schuell). A second set of replicas was prepared from the original filter lifts, and bacteria on both sets of filters were allowed to grow at 37°C on LB plates containing tetracycline until colonies were 1 - 2 mm in diameter. Filters were subsequently transferred to LB plates containing 170 μg/ml chloramphenicol, and the plasmids were allowed to amplify overnight at 37°C.

Cell lysis was carried out by placing filters onto Whatman 3MM paper soaked with 0.5 M NaOH, followed by incubation at room temperature for 20 minutes. In a similar manner, filters were subsequently denatured again with NaOH for 20 minutes, neutralized with 1 M Tris-HCl pH 7.5 for 20 minutes, and finally treated with 0.5 M Tris-HCl pH 7.5/1.5 M NaCl for 20 minutes. The filters were then air-dried and baked at 68°C overnight. Prior to hybridization, filters were washed 3 times in 2X SSC buffer (1X SSC is 0.15 M NaCl, 0.015 M Na Citrate pH 7) in order to remove cellular
debris. The replica filters were then screened by using Pool II and Pool III oligonucleotide mixtures (see Section II.B.1) as hybridization probes. Hybridization and washing conditions were essentially those described by Fung et al. (1984) and are summarized in Section II.K. Putative positive clones were purified from the master plates, and the recombinant plasmids were analyzed (see Section II.E.1). In order to obtain additional ceruloplasmin clones, the library was rescreened using a nick-translated restriction fragment as a probe. Hybridization and washing conditions for the latter screen were essentially as described by Maniatis et al. (1982), and are summarized in Section II.K.

C.2 Preparation and Screening of Randomly-Primed Human Liver cDNA Libraries

To isolate cDNAs encoding the 5' end of the ceruloplasmin transcript, several randomly-primed human liver cDNA libraries were constructed in the vectors λgt10 or λgt11 (Huynh et al., 1984) by Walter Funk (Biochemistry Department, University of British Columbia). Briefly, human liver poly (A)+ RNA (for isolation procedure, see Section II.E.4) was used as a template for first strand cDNA synthesis by reverse transcriptase. DNase I-digested rat thymus DNA (average length 20 nucleotides) was used as a primer (Goelet and Karn, 1984). Second strand synthesis was performed as described by Gubler and Hoffman (1983), using ribonuclease H (BRL), DNA polymerase I (BRL) and E. coli DNA ligase (P-L Biochemicals). After S1 nuclease treatment to generate blunt-ends, the resultant double-stranded cDNA was methylated using EcoRI methylase and S-adenosylmethionine (BRL), EcoRI linkers (P-L Biochemicals) were ligated to the ends, and the linkers were subsequently digested with EcoRI. The
cDNA was then chromatographed on a column (30 x 0.2 cm) of Bio-Gel A-50 m (Bio-Rad), equilibrated with 0.01 M Tris-HCl pH 7.5/0.3 M NaCl/0.001 M EDTA. Fractions comprising the leading edge of the cDNA peak (corresponding to cDNA fragments > 1 Kb) were pooled and the DNA (~ 50 ng) was ligated with 1 µg of EcoRI-digested, dephosphorylated λgt10 or λgt11 DNA (Vector Cloning Systems).

Half of the resulting DNA was packaged into phage particles in vitro using a Gigapak (Vector Cloning Systems). The libraries constructed in λgt10 and λgt11 contained 400,000 and 200,000 independent recombinant clones respectively, and were subsequently plated at high density for screening (4 x 10^4 or 2 x 10^4 plaques per 150 mm petri plate for λgt10 and λgt11 libraries, respectively).

For plating, appropriate dilutions of phage prepared in SM buffer (5.8 g NaCl, 2 g MgSO_4, 50 ml 1 M Tris-HCl pH 7.5, 5 ml 2% gelatin per litre) were incubated with overnight cultures of appropriate host cells at 37°C for 10 minutes in order to allow phage attachment. Phage were then plated on NZYC agarose. Plates were subsequently incubated at 37°C until plaques were visible but not confluent. Replicas (2 sets) of the plaques were then transferred to nitrocellulose circles (132 mm; Schleicher and Schuell). DNA was denatured by treatment of the nitrocellulose filters with 0.5 N NaOH/1.5 m NaCl for 5 minutes. The filters were then neutralized by treatment with 1 M Tris-HCl pH 7.5 for 5 minutes, followed by treatment with 0.5 M Tris-HCl pH 7.5/1.5 M NaCl for 5 minutes. Ceruloplasmin cDNA clones were identified by plaque-hybridization (Benton and Davis, 1977) to appropriate restriction fragments labelled with ^32P by either nick-translation or Klenow
extension (see above). Recombinant clones of interest were detected by autoradiography and purified to homogeneity by successive rounds of replating and rescreening at decreased phage densities. For the λgt10 library, purified phage clones were further screened using the Pool I oligonucleotide mixture (see Section II.B.1) in order to identify cDNAs extending furthest in a 5' direction.

Hybridization and washing conditions for the above screens are summarized in Section II.K.

D. SCREENING OF HUMAN GENOMIC LIBRARIES

Two different human genomic DNA libraries were used in this study. One was a partial AluI/HaeIII digest of human genomic DNA, constructed in the λ Charon 4A vector (Lawn et al., 1978). This library was generously provided by Dr. T. Maniatis and amplified as described (Maniatis et al., 1982). The other library used was a partial Sau3A digest of human lymphocyte genomic DNA, cloned into the BamHI site of the λ derivative EMBL 3. This library was constructed by Val Geddes (1987) (Department of Biochemistry, University of British Columbia) and was screened prior to amplification.

For both libraries, initial screening was carried out on $5 \times 10^5$ - $1 \times 10^6$ plaques, representing 2.5 - 5 genome equivalents, based on the conservative estimate that each clone contained approximately 10 Kbp of genomic DNA. Phage were plated at a density of $3 - 5 \times 10^4$ plaques per 150 mm petri dish. The plating and screening procedures were essentially performed as previously described for the λgt10 and λgt11 cDNA libraries (Section II.C.2) with the exception that replicas of the plaques
were transferred to nitrocellulose filters and then incubated on fresh NZYC plates at 37°C to allow amplification of the phage (Woo, 1980). For screens other than the initial high density screen, this amplification step was omitted.

E. ISOLATION OF NUCLEIC ACIDS

E.1 Purification of Plasmid DNA

Small amounts of plasmid DNA were routinely purified using a modified alkaline lysis procedure of Birnboim and Doly (1979), as described by Maniatis et al. (1982). Briefly, a 1.5 ml aliquot of an overnight bacterial culture containing a recombinant plasmid of interest was placed in a microfuge tube, and bacteria were harvested by centrifugation in an Eppendorf centrifuge for 2.5 minutes. The pellet was resuspended in 100 μl of ice-cold solution containing 50 mM glucose, 10 mM EDTA, 25 mM Tris-Cl pH 8.0 and 4 mg/ml lysozyme (Sigma). This mixture was incubated for 5 minutes at room temperature, followed by the addition of 200 μl of a solution containing 0.2 N NaOH/1% SDS. This solution was incubated at 4°C for 5 minutes, and then 150 μl of potassium acetate solution (60 ml 5 M potassium acetate, 11.5 ml glacial acetic acid, 28.5 ml dH₂O; pH 4.8) was added. This suspension was incubated on ice for 5 minutes, following which the precipitate was removed by centrifugation at 4°C for 5 minutes. The resultant supernatant was removed and extracted with an equal volume of phenol:chloroform (1:1 v/v). Nucleic acids were precipitated by the addition of 2 volumes of ethanol (room temperature) and were recovered by centrifugation for 5
minutes. The pellet was washed with 70% ethanol, air-dried, and resuspended in 50 μl of TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA).

Two different procedures were used for large-scale plasmid isolations. The Triton lysis procedure (Katz et al., 1973, 1977) was used for the large scale preparation of plasmids in the pKT218 cloning vector. An aliquot (5 ml) of an overnight bacterial culture was used to inoculate 1 litre of M9 medium at 37°C. When the OD$_{600}$ of the culture reached 0.6 - 0.7, 250 mg of chloramphenicol was added and the culture was incubated a further 12 - 16 hours at 37°C. Cells were harvested by centrifugation at 6 K RPM in a GS-3 rotor for 20 minutes, followed by freezing of the pellets at -20°C for 2 hours. The cells were then resuspended in 6.25 ml of solution containing 25% w/v sucrose and 50 mM Tris-HCl pH 8.0. Lysozyme (1.5 ml of 10 mg/ml solution) was added, and the solution was mixed by swirling on ice for 5 minutes. EDTA (1.25 ml of a 0.5 M solution, pH 8.0) was added and mixed on ice by swirling for an additional 5 minutes. Triton solution [10 ml of a solution containing 10 ml of 10% (w/v) Triton X-100, 125 ml 0.5 M EDTA pH 8.0, 50 ml 1 M Tris-HCl pH 8.0, 800 ml dH$_2$O] was then added, and the solution was mixed again for 5 minutes on ice. Following lysis, cell debris was removed by centrifugation at 19 K RPM in an SS-34 rotor for 30 minutes (4°C). Plasmid DNA was subsequently isolated by isopycnic centrifugation (Ti 70.1 rotor/20 hours/20°C) using CsCl-ethidium bromide density gradients.

The large-scale alkaline lysis procedure of Birnboim and Doly (1979) as described by Maniatis et al. (1982) was used for the isolation of recombinant pUC plasmids and M13 replicative form (RF) DNA (Messing, 1983). In the former case, 500 ml aliquots of L broth were inoculated
with 10 ml of the appropriate bacterial culture (grown in the presence of a selective antibiotic) and incubated at 37°C for 12 - 16 hours without chloramphenicol amplification. For the large-scale isolation of M13 RF DNA, 5.0 ml of an exponentially-growing JM103 or JM101 culture and 300 µl of an M13 infectious phage supernatant (see Section II.H.2) were added to 500 ml of YT broth and incubated at 37°C for 8 - 10 hours. In both cases, the purification procedure was a large-scale version of that described for the isolation of small amounts of plasmid DNA (see above), with several modifications. Following addition of the potassium acetate solution, cell debris was removed by centrifugation at 19 K RPM in an SS-34 rotor for 30 minutes. The supernatant was mixed directly with 0.6 volumes of isopropanol, and incubated at room temperature for 15 minutes. Nucleic acids were pelleted by centrifugation for 20 minutes at 10 K RPM in an SS-34 rotor. The pellets were resuspended in 11.0 ml TE, to which 11 g CsCl and 600 µl ethidium bromide (10 mg/ml in H2O) were added. This mixture was incubated on ice for 60 minutes, followed by low-speed centrifugation (4 K RPM for 5 minutes in an HB-4 rotor) in order to remove the flocculant precipitate. The supernatant was loaded into 13 ml Quick-Seal Tubes (Beckman), and the DNA was banded by isopycnic density gradient centrifugation under the conditions described above.

E.2 Isolation of Bacteriophage DNA

Phage DNA was purified routinely from 20 ml lysates. For Charon 4A phage isolates, consistent lytic infections were obtained using the following conditions: 200 µl of SM buffer containing 3.0 - 4.0 x 10^6 phage was incubated at 37°C for 20 minutes with 100 µl of an overnight host bacterial culture. This mixture was added to 20 ml of appropriate
growth media in a 125 ml Erlenmeyer flask, and incubated at 37°C with vigorous shaking. Lysis was usually observed within 4.5 – 7 hours post-innoculum. For phage isolates from either the EMBL 3 genomic library or λgt11 and λgt10 cDNA libraries, a single plaque was added directly to a solution containing 100 μl of stationary-phase bacterial host cells and 100 μl of SM buffer, and incubated at 37°C for 20 minutes to allow attachment. The mixture (containing the phage plug) was then added to 20 ml of appropriate growth medium in a 125 ml Erlenmeyer flask, and incubated at 37°C with vigorous shaking. Under these conditions, lysis was usually observed within 3 – 4 hours.

At the time of cell lysis, chloroform (3 mL) was added to the culture, which was then left shaking slowly for a further 5 – 10 minutes. At that time, the contents were carefully transferred to a 30 ml Corex tube, such that most of the chloroform was left behind, and the sample was centrifuged at 10 K RPM for 10 minutes in an SS-34 rotor. To the resulting supernatant, 3 ml of 5 M NaCl and 3 g of polyethylene glycol (PEG) 8000 (average molecular weight 7000 – 9000) were added. The contents were mixed, and left at 4°C overnight in order to precipitate phage particles. Following centrifugation of the suspension for 10 minutes at 10 K RPM in an HB-4 rotor, the phage pellet was resuspended in 500 μl of DNase I buffer (50 mM HEPES pH 7.5, 5.0 mM MgCl₂, and 0.5 mM CaCl₂), and transferred to a microfuge tube. RNAse A (Sigma) (10 μl of a 5 mg/ml stock) and DNase I (Boehringer Mannheim) (5 μl of a 10 mg/ml stock) were added, and the solution was incubated at 37°C for 1 hour. Following digestion, the solution was centrifuged for 5 minutes in a microfuge in order to remove any remaining cellular debris. To the
supernatant, 50 µl of 10X SET buffer (0.1 M Tris-HCl pH 7.5, 0.2 M EDTA and 5% SDS) was added prior to digestion with 4 µl of proteinase K (40 mg/ml stock solution) (Boehringer Mannheim) for 30 minutes at 68°C. The solution was extracted twice with an equal volume of phenol:chloroform (3:1 v/v followed by 1:1 v/v) and once with an equal volume of chloroform. The DNA was precipitated from the aqueous phase by addition of 2 volumes of 95% ethanol at room temperature for 2 minutes, and collected by centrifugation. The DNA pellet was washed with 1.0 ml of 70% ethanol, air-dried, and resuspended in 50 µl of TE buffer.

E.3 Preparation of Human Genomic DNA

Genomic DNA from human liver was prepared essentially as described by Blin and Stafford (1976). Liver tissue was ground to a fine powder in liquid nitrogen in a Waring blender. The resulting powder was dissolved in a buffer (10 ml/g tissue) consisting of 0.5 M EDTA pH 8.0, 0.5% SDS, and 100 µg/ml proteinase K, and was incubated for 12 - 16 hours at 50°C. The solution was extracted 3 times with equal volumes of phenol and then dialyzed against buffer (50 mM Tris-HCl pH 8.0, 10 mM NaCl, 10 mM EDTA) until the OD<sub>270</sub> of the dialysate was < 0.05. RNAse A was added to a final concentration of 100 µg/ml and the solution was incubated at 37°C for 60 minutes. The DNA solution was extracted 3 times with equal volumes of phenol:chloroform (1:1 v/v) and then dialyzed against TE. Insoluble material was removed by centrifugation at 10 K RPM in an SS-34 rotor at 4°C for 15 minutes. Genomic DNA was precipitated by the addition of Gilbert Salts (5X Salts is 2.5 M NH<sub>4</sub>Ac, 100 mM MgCl<sub>2</sub> and 1 mM EDTA) to a final concentration of 1X, followed by the addition of 2 volumes of 95% ethanol. Following collection by
centrifugation, the DNA pellet was allowed to rehydrate for several days. At this time, any remaining insoluble material was removed by centrifugation as previously described. The final DNA pellet was resuspended in TE, at a concentration of 0.5 – 1.0 mg/ml.

Human genomic DNA was also prepared from white blood cells by Heather Kirk (Department of Biochemistry, University of British Columbia) according to a modified procedure of Kunkel et al. (1979), and was generously supplied by Ms. Kirk for various aspects of this study.

E.4 Isolation of RNA

E.4.1 Preparation of human liver poly(A)$^+$ RNA. All glassware, utensils, and solutions were autoclaved prior to use in order to inactivate contaminating ribonucleases. RNA was isolated from human liver by the guanidine hydrochloride method (Chirgwin et al., 1979). Powdered human liver (obtained from brain-dead donors and immediately frozen in liquid nitrogen) was added to buffer (10 ml/10 mg of tissue) consisting of 7.5 M guanidine hydrochloride (GuHCl) pH 7.5, 25 mM sodium citrate pH 7.0 and 0.1 M DTT. The suspension was disrupted using a Polytron homogenizer. N-lauryl sarcosine was added to 0.5% (w/v) and the insoluble material was removed by centrifugation (10 K RPM for 15 minutes at 4°C in an SS-34 rotor). Following addition of ethanol to a final concentration of 33%, RNA was precipitated overnight at -20°C. The precipitate was collected by centrifugation under the conditions described above, and dissolved in one half of the starting volume with the GuHCl buffer. Again, insoluble material was removed by centrifugation. RNA was precipitated as described above, resuspended in one quarter the starting volume with the GuHCl buffer, and reprecipitated. The final RNA pellet was resuspended in
sterile dH$_2$O, and any precipitate removed by centrifugation (10 K RPM for 15 minutes at 4°C in the SS-34 rotor).

Human poly(A)$^+$ RNA was isolated by chromatography on a column of oligo d(T) cellulose (Sigma) (Edmonds et al., 1971; Aviv and Leder, 1972). Total human or bovine RNA samples (in buffer containing 10 mM Tris-HCl pH 7.5, 4 mM EDTA, 0.5 M NaCl) were applied to the column. The unbound RNA fraction was reapplied to the column twice, and the column was washed with the buffer described above until the OD$_{260}$ of the eluate was less than 0.05. Poly(A)$^+$ RNA was subsequently eluted from the column with sterile dH$_2$O; fractions containing RNA were identified spectrophotometrically and pooled. RNA was precipitated by the addition of 0.1 V of 3 M NaOAc pH 4.8 and 2 volumes of ethanol. Poly(A)$^+$ RNA was resuspended in sterile dH$_2$O at a concentration of 0.6 – 1.0 mg/ml and stored in aliquots at -70°C.

E.4.2 Preparation of total RNA from HepG2 cells. Total cellular RNA from HepG2 cells (Knowles, 1980) was purified as previously described by van Oost et al. (1985) for RNA isolation from the hybrid endothelial cell line EA.hy926. Cells grown to confluency on a total area of 5225 cm$^2$ were dissolved in 135 ml of GuHCl buffer, and total RNA was purified from this solution by successive ethanol precipitations as described above.

F. BASIC DNA TECHNIQUES

F.1 Restriction Enzyme Digestion

With the exception of genomic digests, DNA (usually 0.5 – 2 μg) was routinely digested in a total volume of 20 μl, using the buffer system described by Maniatis et al. (1982). BSA (BRL) was added to
a final concentration of 100 µg/ml. In most cases, 1 - 5 units of restriction enzyme were used per reaction. For digests of small-scale plasmid and phage DNA preparations, 5 µg of RNase A was included. Genomic DNA (5 - 10 µg) was usually digested in a total volume of 30 µl using 30 - 50 units of appropriate restriction enzyme. Enzymes were purchased from BRL, New England Biolabs, Boehringer Mannheim, and P-L Biochemicals.

Restriction enzyme digestion mixtures were analyzed by electrophoresis in agarose or polyacrylamide gels (see below), following the addition of 0.1 volumes of loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol and 25% ficoll).

F.2 End-labelling of DNA Fragments

Where necessary, ceruloplasmin cDNA inserts, cloned into the EcoRI site in λgt10 or λgt11 vectors (see Section II.C.2), were visualized by labelling the 5' overhang generated by EcoRI digestion. These reactions were carried out according to Maniatis et al. (1982), using \([\alpha^{32}P]\)-dATP and the Klenow fragment of E. coli. DNA polymerase I. Reaction products were analyzed by polyacrylamide gel electrophoresis followed by autoradiography, as described below (see Section F.3.1.).

F.3 Electrophoresis of DNA

F.3.1 Agarose gel electrophoresis. DNA samples of restriction endonuclease digestion mixtures were analyzed on 0.7 - 1.5% agarose gels, which were poured and run in either 1XTBE buffer (89 mM Tris-borate pH 8.3, 89 mM borate and 2.0 mM EDTA) (Maniatis et al., 1982) or 1XTAE buffer (50XTAE is 2 M Tris base, 1 M glacial acetic acid) (Maniatis et al., 1982), containing 0.5 µg/ml ethidium bromide. DNA fragments
were visualized under UV light (260 nm) and photographs were taken with a Polaroid camera using Type 57 film.

F.3.2 Polycrylamide gel electrophoresis. Non-denaturing polyacrylamide gels (5 - 10% prepared from a stock of 29:1 acrylamide:bis-acrylamide) were poured and run in 1XTBE buffer. Polymerization was initiated by the addition of ammonium persulfate and TEMED to final concentrations of 0.066% (w/v) and 0.04% (w/v) respectively. DNA fragments were visualized either by staining of the gels with 10 µg/ml ethidium bromide in 0.5XTBE followed by UV irradiation or by autoradiography if the fragments were end-labelled. In the latter case, gel were dried under vaccuum at 80°C for 40 minutes using a Bio-Rad gel drier and exposed to X-ray film (Kodak XK-1). Where required, intensifying screens (Lightning Plus, Dupont) were used at -70°C.

For denaturing polyacrylamide gels (6 - 20%, prepared from a stock of 38:2 acrylamide:bis-acrylamide), urea (final concentration of 8.3 M) was added as a denaturant, and the gels were poured and run in 1XTBE buffer. Polymerization of the gels was initiated by the addition of ammonium persulfate and TEMED to final concentrations of 0.066% (w/v) and 0.024% (w/v) respectively. Gels were dried, and the DNA was visualized by autoradiography as described above.

F.4 Southern Transfers

DNA separated by electrophoresis in agarose gels was transferred to nitrocellulose (Schleicher and Schuell), Nytran (Schleicher and Schuell) or Zetaprobe (Bio-Rad) essentially as described by Southern (1975), except that the acid depurination step was routinely omitted. DNA fragments in the gels were denatured for 45 - 60 minutes in 0.5 N NaOH,
1.5 M NaCl, and then neutralized by treating twice for 20 minutes each with 0.5 M Tris-HCl pH 7.5/1.5 M NaCl. DNA transfer to various membranes was carried out overnight in either 10XSSC or in 1.0 M ammonium acetate (pH 7), the latter being more efficient. Filters were then air dried and baked at 68°C for 4 - 16 hours prior to hybridization.

G. DNA CLONING

G.1 Fragment Production

DNA fragments for ligation into either pUC or M13 vectors were produced by several methods, including sonication (Deininger, 1983) or by restriction enzyme digestions and subsequent fragment isolation. In the latter case, DNA fragments were recovered from agarose gels by electroelution into dialysis tubing containing either 0.5XTBE or 0.5XTAE buffer. The DNA was then purified by chromatography through NACS PREPAC cartridges (BRL). In this procedure, DNA was loaded and washed in TE containing 0.2 M NaCl. The DNA was eluted in 500 µl of TE containing 2.0 M NaCl. After addition of 0.1 volume of 3.0 M Na acetate (pH 4.8) the fragment was precipitated with 2 volumes of 95% ethanol and resuspended in a small volume of sterile TE buffer.

Random DNA fragments were produced by sonication (Deininger, 1983) using a Heat Systems Sonifier. Plasmid DNA (10 - 20 µg in 500 µl of 0.5 M NaCl, 0.1 M Tris-HCl pH 7.4, 10 mM EDTA) was sheared by 5 power bursts of 5 seconds each. Following ethanol precipitation, DNA fragments of 300 - 600 bp were separated by electrophoresis in a 5% non-denaturing polyacrylamide gel and isolated by electroelution. The DNA fragments were made blunt-ended by incubation at 37°C for 90 - 120 minutes with 33 mM
Tris-acetate pH 7.8, 66 mM potassium acetate, 10 mM Mg acetate, 100 mg/ml BSA, and 0.2 mM of each deoxyribonucleotide triphosphate in 50 µl total volume, containing 6 units of T4 DNA polymerase (BRL). Following phenol/chloroform (1:1 v/v) extraction, the DNA was precipitated with ethanol and resuspended at a concentration of 10 ng/µl of TE buffer.

G.2 Ligation of DNA into pUC or M13 Vectors

For restriction endonuclease mapping analysis or sonication, DNA fragments were subcloned into pUC12 or pUC13 vectors (Vieira and Messing, 1982) (pUC plasmids were kindly provided by Dr. Roland Russnak, Department of Biochemistry, University of British Columbia). For DNA sequence analysis, the M13 vectors mp8, mp9, mpl8 and mpl9 were used.

M13 ligations were usually carried out with 10 ng of vector DNA and 20 - 60 ng insert DNA in 20 µl of 50 mM Tris-HCl pH 7.4, 10 mM MgCl₂, 10 mM DTT, 1.0 mM spermidine, 1.0 mM ATP and 100 µg/ml BSA. For ligations using pUC vectors, 100 ng of vector DNA was ligated to a 3-fold molar excess of insert DNA, using the ligation buffer described above. Phage T4 DNA ligase (BRL or P-L Biochemicals) was added (0.5 or 1 unit for cohesive or blunt-ended ligations respectively) and the reaction was allowed to proceed at 15°C overnight for blunt-ended ligations or at 4°C overnight for cohesive-ended ligations.

G.3 Transformations

E. coli strains JM103 or JM101 were used as bacterial hosts for all pUC and M13 transformations (see Table II). Bacterial cells were made competent by treatment with 50 mM calcium chloride (Messing, 1983). Aliquots of competent cells (0.3 ml) were typically transformed with 3 - 5 µl of ligation mixture (see preceding section). Cells were incubated
with DNA at 4°C for 40 - 60 minutes and then heat-shocked at 42°C for 3 minutes prior to plating. Bacteria infected with recombinant M13 phage were assayed for their inability to cleave 5-bromo4-chloro3-indolyl-galactoside (X-Gal) as described by Messing (1983), resulting in the appearance of clear plaques. The same colour assay was used to detect bacterial colonies containing pUC plasmids.

For clarity, a summary of the different subclones used in analysis of the wild-type human ceruloplasmin gene is presented in Table III, and will be referred to in subsequent sections.

H. DNA SEQUENCE ANALYSIS

H.1 Screening of M13 Clones

In the case of sonication, mixtures of randomly-sheared DNA fragments were cloned into M13 vectors (see Section II.G.1). In order to identify recombinant clones carrying either cDNA or exon-encoding sequences, M13 plaques were screened by plaque hybridization as previously described (Benton and Davis, 1977) (see Section II.C.2). Hybridization and washing conditions varied, depending of the nature of the probe (see Section II.K for a summary of hybridization and washing conditions used in the present study).

H.2 Isolation of M13 Template DNA

Single-stranded M13 phage DNA from clones of interest was prepared as described by Messing (1983). Aliquots (2 ml) of YT containing 20 µl of host bacteria (JM101 or JM103) were each inoculated with a single plaque, and incubated at 37°C for 8 - 10 hours. Bacterial cells were pelleted by centrifugation in a 1.5 ml microfuge tube. Phage
Table III. Cloning Strategy for the Wild-Type Human Ceruloplasmin Gene.

Various restriction fragments containing exon sequences were subcloned from ceruloplasmin genomic clones (Figure 10) into appropriate pUC and/or M13 vectors. Asterisks follow those pUC subclones that were analyzed by sonication. Crosses identify subclones that were used in Northern blot analysis of the 5' end of the gene. L1 - L4 designate exons present in the 5' untranslated region of the gene.

<table>
<thead>
<tr>
<th>RESTRICTION FRAGMENT</th>
<th>CORRESPONDING EXON(S)</th>
<th>PHAGE CLONE DERIVED FROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6 Kbp EcoRI (pUC13)</td>
<td>L1, L2</td>
<td>XWT7</td>
</tr>
<tr>
<td>700 bp XbaI/EcoRI (pUC13)</td>
<td>→ 85 bp Sau3A/EcoRI (mpl8)</td>
<td>L1</td>
</tr>
<tr>
<td>1.5 Kbp EcoRI† (pUC13)</td>
<td>190 bp Sau3A (mpl8)</td>
<td>XWT2</td>
</tr>
<tr>
<td>800 bp XbaI/EcoRI (pUC13)</td>
<td>→ 70 bp HinfI/EcoRI (mpl8)</td>
<td>L2</td>
</tr>
<tr>
<td>85 bp Sau3A/EcoRI (mpl8)</td>
<td>→ 220 bp HinfI (mpl8)</td>
<td></td>
</tr>
<tr>
<td>420 bp EcoRI† (pUC13)</td>
<td>mp18,19</td>
<td>XWT7</td>
</tr>
<tr>
<td>2.7 Kbp HinfIII/EcoRI (pUC13)†</td>
<td>→ 160 bp BdeI (mpl8)</td>
<td>L3</td>
</tr>
<tr>
<td>1.3 Kbp HinfIII/EcoRI† (pUC13)</td>
<td>→ 450 bp BdeI (mpl8)</td>
<td></td>
</tr>
<tr>
<td>1.4 Kbp EcoRI (pUC13)†</td>
<td></td>
<td>XWT7</td>
</tr>
<tr>
<td>3.1 Kbp EcoRI (pUC12)†</td>
<td>L4, L1</td>
<td>XWT2</td>
</tr>
<tr>
<td>Restriction Fragment</td>
<td>Corresponding Exon(s)</td>
<td>Phage Clone Derived From</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>4.6 Kbp EcoRI (pUC12)*</td>
<td>2</td>
<td>λWT2</td>
</tr>
<tr>
<td>800 bp EcoRI/SallI (mpl18,19)</td>
<td>3</td>
<td>λWT6</td>
</tr>
<tr>
<td>1.8 Kbp EcoRI (pUC12)</td>
<td>4</td>
<td>λWT5</td>
</tr>
<tr>
<td>3.2 Kbp EcoRI/SallI (pUC12)</td>
<td>5</td>
<td>λWT5</td>
</tr>
<tr>
<td>2.6 Kbp HindIII (pUC12)*</td>
<td>7</td>
<td>λWT1</td>
</tr>
<tr>
<td>1.6 Kbp BamHI/HindIII (pUC12)*</td>
<td>8,9</td>
<td>λWT1</td>
</tr>
<tr>
<td>670 bp BamHI/SastI (mpl8)</td>
<td>9</td>
<td>λWT1</td>
</tr>
<tr>
<td>450 bp EcoRI (pUC12, mpl8, mpl19)</td>
<td>10</td>
<td>λWT1</td>
</tr>
<tr>
<td>1.2 Kbp EcoRI (pUC12)</td>
<td>10</td>
<td>λWT3</td>
</tr>
<tr>
<td>4.8 Kbp EcoRI (pUC13)*</td>
<td>10,11,12,13</td>
<td>λWT3</td>
</tr>
<tr>
<td>2.1 Kbp EcoRI/SallI (pUC13*, mpl19)</td>
<td>13,14</td>
<td>λWT3</td>
</tr>
</tbody>
</table>
particles in 1.3 ml of supernatant were precipitated by the addition of
0.3 ml of a solution containing 20% PEG, 2.5 M NaCl, followed by
incubation at room temperature for 15 minutes. M13 phage were collected
by centrifugation for 5 minutes. The phage pellet was then resuspended in
200 μl of low tris buffer (50 mM NaCl, 10 mM Tris-HCl, pH 7.5, 1 mM
EDTA). DNA was purified by extraction with phenol/chloroform (1:1 v/v),
followed by a 1:1 extraction with chloroform. The DNA was then
precipitated by the addition of 0.1 volumes of 3 M Na Acetate and 2
volumes of ethanol. Final M13 phage pellets were resuspended in 50 μl
of low tris buffer.

H.3 DNA Sequence Analysis

All DNA sequence was determined using the dideoxynucleotide
chain termination technique of Sanger et al. (1977) as modified for M13
phage templates (Messing et al., 1981). Sequencing reactions were
performed using the dideoxy/deoxyribonucleotide concentrations given in
Table III. DNA sequence analysis was carried out by hybridizing 4 μl of
M13 template DNA with 1 μl of universal sequencing primer (17mer; 0.03
OD 260 units/ml), 1 μl dH₂O, and 2 μl 10 x Hin buffer (600 mM NaCl,
100 mM Tris-HCl pH 7.5, 70 mM MgCl₂) at 68°C for 10 minutes.
Following slow cooling to room temperature, 1 μl of 15 μM dATP and 1.5
μl of [α³²P]-dATP (10 μCi/μl; 3000 Ci/mMole) was added; 2 μl
of this mixture was then added to 1.5 μl of the appropriate
dideoxy/deoxyribonucleotide mix (see Table IV). To initiate the extension
reaction, 0.4 units DNA polymerase 1 Klenow fragment (BRL, Pharmacia) was
added to each tube. Following incubation for 15 minutes at room
temperature, 1 μl of 0.5 mM dATP was added to each reaction, and the
### Composition of Sequencing Mixes

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>d/ddG</th>
<th>d/ddA</th>
<th>d/ddT</th>
<th>d/ddC</th>
</tr>
</thead>
<tbody>
<tr>
<td>dG</td>
<td>7.9</td>
<td>109.4</td>
<td>158.7</td>
<td>157.9</td>
</tr>
<tr>
<td>dT</td>
<td>157.6</td>
<td>109.4</td>
<td>7.9</td>
<td>157.9</td>
</tr>
<tr>
<td>dC</td>
<td>157.4</td>
<td>109.4</td>
<td>158.7</td>
<td>10.5</td>
</tr>
<tr>
<td>ddG</td>
<td>157.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ddA</td>
<td>-</td>
<td>116.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ddT</td>
<td>-</td>
<td>-</td>
<td>550.3</td>
<td>-</td>
</tr>
<tr>
<td>ddC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>191.6</td>
</tr>
</tbody>
</table>

The numbers refer to concentrations (µM) of dideoxy- and deoxyribonucleotide triphosphates used for the preparation of M13 DNA sequencing mixes. The values given were determined empirically by Dr. Joan McPherson, Department of Plant Sciences, University of British Columbia.
cold "chase" was allowed to proceed for an additional 15 minutes at room temperature. At this time, 5 µl of dye mix (98% formamide, 10 mM EDTA pH 8.0, 0.02% xylene cyanol, 0.02% bromophenol blue) was added. Reaction products were denatured by heating at 90°C for 3 minutes, and 2 µl of each reaction was analyzed on 6% denaturing polyacrylamide gels (see Section II.F.3.2). After electrophoresis (0.9 W/cm), gels were dried and exposed overnight to Kodak XK-1 film at room temperature.

In order to avoid redundancy, most M13 clones generated by sonication were first analyzed using only ddTTP reactions (Sanger et al., 1980). Where necessary, fragments cloned into the M13 in opposite orientation were identified according to their ability to form figure 8-like configurations, which migrate more slowly in agarose gels (Messing, 1983).

I. RNA ANALYSIS

1.1 Northern Blot Analysis

RNA samples (10 µg poly(A)+ human liver RNA or 20 µg HepG2 total cellular RNA) were denatured by addition of loading buffer containing formaldehyde, and subsequently separated by electrophoresis in formaldehyde-containing agarose gels (Lehrach et al., 1977; Maniatis et al., 1982). Prior to transfer, gels were treated with 50 mM NaOH, 10 mM NaCl for 45 minutes, neutralized by treatment for 45 minutes with 0.1 M Tris-HCl pH 7.5, and finally soaked in 20XSSC for 60 minutes. RNA was transferred to nitrocellulose in 20XSSC buffer, by the method of Southern (1975). Following transfer, blots were air-dried and baked for 4 - 16 hours at 68°C. Ceruloplasmin mRNA species were detected using
hybridization and washing conditions described for double-stranded DNA probes (see Section II.K).

I.2 RNA Dot Blots

Dot blots for rapid analyses were prepared by spotting 7.5 - 10 µg of human liver mRNA in sterile 10XSSC onto nitrocellulose filters. Following air drying, filters were baked at 68°C for 4 - 16 hours. The dot blots were probed subsequently with labelled M13 templates containing ceruloplasmin genomic DNA fragments (see Section II.K for hybridization and washing conditions).

I.3 Nuclease S1 Mapping

Single-stranded probes were prepared from recombinant M13 templates as described previously (see Section II.B.4). For nuclease S1 protection assays (as described by Kay et al., 1986), 100,000 - 150,000 cpm of single-stranded DNA probe (specific activity = 10⁸ cpm/µg) was mixed with 0.35 - 1.0 µg of human liver poly(A)⁺ RNA in a final volume of 30 µl hybridization buffer (50% formamide, 10 mM PIPES pH 6.9, 400 mM NaCl, 1 mM EDTA). The hybridization mixture was then heated to 80°C for 15 minutes and subsequently incubated at 42°C for 12 hours. Following annealing, a solution (200 µl) was added that contained 70 mM Na Acetate, 600 mM NaCl, 2.5 mM ZnSO₄, and 150 - 200 U of nuclease S1. After digestion (60 minutes at 37°C), intact probe DNA was precipitated by the addition of 30 µl of a solution containing 100 mM EDTA, 4 M ammonium acetate and 100 µg of tRNA per ml, followed by 230 µl isopropanol. Reaction products were separated by electrophoresis on a 6% denaturing polyacrylamide gel (see Section II.F.3.2) and were visualized
by exposure to Kodak XK-1 film overnight at \(-70^\circ C\) with an intensifying screen.

J. CHROMOSOME MAPPING

Chromosome localization studies for the human ceruloplasmin pseudogene were performed in collaboration with Dr. John Hamerton (Department of Human Genetics, University of Manitoba) using human-hamster somatic cell hybrids which had been characterized previously by cytogenetic and isozyme analysis (Donald et al., 1983; Riddell et al., 1985). DNA from cultured cell lines was isolated as described by Riddell et al. (1986). DNA (5 µg) from each hybrid line, as well as control human placental and hamster DNA was digested with \(\text{EcoRI}\), electrophoresed on 1% agarose gels and transferred to Zetaprobe (Bio-Rad) according to the manufacturer's specifications. Blots were probed with a \(^{32}\text{P}\)-labeled restriction fragment specific for the human ceruloplasmin pseudogene (see Section II.K for hybridization and washing conditions).

K. SUMMARY OF HYBRIDIZATION/WASHING CONDITIONS

K.1 Genomic Southern Blot Hybridization

DNA fragments were detected by hybridization to \(^{32}\text{P}\)-labelled DNA probes as described by Kan and Dozy (1978). Membranes were first wetted with 3XSSC and then prehybridized for 2 - 4 hours at 37°C in a solution containing 50% formamide, 6XSSC, 1 mM EDTA, 0.1% SDS, 10 mM Tris-HCl pH 7.5, 10X Denhardts (DH) solution (1XDH is 0.02% BSA, 0.02% ficoll, 0.02% polyvinylpyrrolidone), 0.05% sodium pyrophosphate, 100 µg/ml denatured herring sperm DNA, and 25 µg/ml poly(A)+ RNA.
Hybridizations were carried out in the above buffer, with the addition of denatured probe. After hybridization, which was allowed to proceed for 24 - 36 hours at 37°C, blots were washed in 2XSSC, 1XDH, and then washed twice for 45 minutes each at 50°C in 0.1XSSC, 0.1% SDS. Following a final room temperature rinse in 0.1XSSC, filters were air dried and exposed to film (see below). For the chromosome localization study, human-hamster hybrid panels were prehybridized overnight at 42°C in 50% formamide, 3XSSPE, (1XSSPE is 0.1 mM EDTA, 10 mM NaH₂PO₄ pH 7.0 and 0.18 M NaCl), 1% SDS, 0.5% non-fat powdered milk, 10% dextran sulfate and 200 μg/ml salmon sperm DNA. The hybridization reaction was carried out overnight in the same buffer with the addition of nick-translated probe at a concentration of 20 ng/ml. Following hybridization, blots were washed twice in 2XSSC at room temperature for a total of 10 minutes. Filters were then washed for 15 minutes in 0.2XSSC, 0.1% SDS at 55°C, and finally for 15 minutes in 0.2XSSC, 0.1% SDS at 55°C.

K.2 Hybridization Conditions Other Than for Genomic Southern Blots

For all other hybridizations using double-stranded DNA probes, filters were prehybridized in a solution containing 6XSSC, 2XDH solution at 68°C for 1 - 4 hours. Filters were then hybridized overnight at 68°C in 6XSSC, 2XDH, 1 mM EDTA, 0.5% SDS. and the denatured probe DNA (at least 1 x 10⁶ cpm/ml, with specific activity ≥ 0.5 x 10⁸ cpm/μg). Following hybridization, filters were rinsed once at room temperature in 2XSSC, and then washed three times at 68°C in 1XSSC, 0.5% SDS for 30 - 40 minutes each, and finally rinsed in 2XSSC at room temperature.
For 5'-labelled oligodeoxyribonucleotide probes, prehybridization was carried out at 37°C for 2 – 16 hours in 6XSSC, 2XDH, and 0.2% SDS. Filters were then hybridized at 68°C with the addition of 5'-end labeled oligonucleotide (at least 1 x 10^6 cpm/ml, with specific activity ≈ 10^6 cpm/p mole). All washes contained 6XSSC, and were usually carried out at room temperature for 15 minutes, and then twice for 15 minutes each at 37°C. This was followed by higher temperature washings at 40 – 55°C, depending on the base composition of the oligonucleotide.

After washing, nitrocellulose filters from all hybridizations described above were air-dried and exposed overnight to either Kodak XK-1 or Kodak X-Omat AR film (the latter being approximately 5-fold more sensitive) at -70°C with intensifying screens.
III. RESULTS

A. CHARACTERIZATION OF THE HUMAN PRECERULOPLASMIN cDNA

A.1 Initial Screening of a Human Liver cDNA Library

Two hundred thousand recombinant clones from a human liver cDNA library (provided by Dr. Stuart Orkin at Children's Hospital, Harvard University) were screened at high density by using pool II and pool III oligonucleotide mixtures (corresponding to amino acid residues 937 – 942 and 962 – 967 of the ceruloplasmin protein sequence respectively (see Figures 5 and 7) as hybridization probes. One recombinant clone (designated phCP-1) hybridized specifically to both oligonucleotide mixtures. Restriction endonuclease mapping of the purified plasmid showed that phCP-1 contained an insert of approximately 2.7 Kbp cloned into the PstI site of the pKT218 vector. Subsequent DNA sequence analysis (see Section III.A.3) showed that the phCP-1 insert contained DNA encoding amino acid residues 202 – 1046 of plasma ceruloplasmin (Takahashi et al., 1984), in addition to a 123 bp 3' untranslated region and a poly(A) tract (see Figures 5 and 6). In order to isolate a clone(s) coding for the 5' region of the ceruloplasmin mRNA, the cDNA library was rescreened using a 322 bp HaeIII – PstI fragment as hybridization probe (Probe A; see Figure 6). This fragment was derived from the 5' end of the phCP-1 insert. A singular clone was identified and found to contain DNA corresponding to amino acid residues 202 – 432 of ceruloplasmin. Therefore, the latter clone extended no further 5' than phCP-1.

A.2 Isolation of cDNA Clones Encoding the 5' End of Human Ceruloplasmin

For the isolation of a clone(s) containing cDNA sequence corresponding to the remainder of the ceruloplasmin mRNA, two
Figure 5. Schematic summary of the cloning of the human preceruloplasmin cDNA.

The phCP-1 clone (derived from the Stuart Orkin human liver cDNA library) encodes amino acid residues 202 - 1046 of ceruloplasmin, followed by a 123 bp 3' untranslated region (cross-hatched bar) and a poly(A) tail. The λhCP-1 clone (isolated from a randomly-primed human liver cDNA library) was found to contain sequence corresponding to amino acid residues 1 - 380 of plasma ceruloplasmin, preceded by a 19 amino acid long signal peptide peptide (solid bar). An arrow identifies the site of signal peptide cleavage. Stars indicate the positions of synthetic oligonucleotide probes used to screen the respective cDNA libraries.
RANDOM PRIMER USED IN FIRST STRAND CDNA SYNTHESIS

DOUBLE-STRANDED CDNA CLONED INTO λgt 10

OLIGO dT PRIMER USED IN FIRST STRAND CDNA SYNTHESIS

DOUBLE-STRANDED CDNA CLONED INTO pKT 218 PLASMID

EcoR1

Xhaf

Signal sequence

Signal peptide

-57 0 600 1200 1800 2400 3000 3321

3' NON-CODING SEQUENCE

EcoR1

EcoR1

polyA phCP-1

NUCLEOTIDES

AMINO ACIDS
randomly-primed cDNA libraries were constructed (see Section II.C.2). The first of these, cloned into the EcoRI site of the phage vector λgt10, was screened by plaque hybridization using a 1071 bp PstI-EcoRI fragment isolated from phCP-1 (Probe B, Figure 6) as a probe. Of the 16 positive clones that were identified, Southern blot analysis showed that 13 hybridized to the 322 bp HaeIII-PstI fragment derived from the 5' end of the phCP-1 cDNA insert (Probe A, Figure 6). Of these latter clones, only one was found to hybridize specifically to the pool I oligonucleotide mixture, which corresponds to the amino-terminal 6 amino acids of plasma ceruloplasmin (see Figures 4 and 6). This clone (designated λhCP-1; see Figures 5 and 6) was characterized by restriction endonuclease mapping and DNA sequence analysis (see below) and was found to contain an EcoRI insert of approximately 1.2 Kbp, corresponding to amino acid residues 1 - 380 of the mature ceruloplasmin protein. This sequence was preceded by a signal peptide of 19 amino acid residues beginning with a putative initiator methionine (see Figures 5 and 7).

A second randomly-primed human liver cDNA library was constructed in the EcoRI site of the phage vector λgt11 and screened in order to identify a clone(s) containing the nucleotide sequence of the 5' untranslated region of the ceruloplasmin mRNA. The probe used for library screening was a 150 bp EcoRI-HindIII fragment obtained from the 5' end of the λhCP-1 clone (designated Probe C, Figure 6). Five positive cDNA clones were identified (designated λhCP-2 to λhCP-6) and were subsequently purified to homogeneity. From restriction endonuclease analysis, it was determined that these clones contained EcoRI inserts of 780 bp, 920 bp, 1.1 Kbp, 730, and 1.0 kbp, for λhCP-2 to λhCP-6.
Figure 6. Restriction Map and Sequencing Strategy for Human Preceruloplasmin cDNA Clones.

The longer bars below the restriction map represent the clones phCP-1 and λhCP-1 that together include regions coding for the leader peptide (hatched bar), the plasma protein (open bar), and the 3' untranslated sequence (solid bar). Arrows indicate the extent and direction of nucleotide sequence obtained from various M13 clones. Restriction fragment probes A, B, and C, which were used in library screening (see text for details) are indicated directly below the restriction map (solid bars). The PstI and EcoRI sites in parentheses result from the cloning procedures used in the construction of the cDNA libraries. Kb, Kilobases.
respectively. Subsequent DNA sequence analysis (see below) showed that \( \lambda hCP-2 \) and \( \lambda hCP-6 \) extended an additional 19 and 38 bp respectively, 5' to the previously characterized \( \lambda hCP-1 \) clone (see Figure 8).

A.3 DNA Sequence Analysis of Human Preceruloplasmin

The complete nucleotide sequence of the phCP-1 and \( \lambda hCP-1 \) cDNA inserts was determined using the strategy shown in Figure 6. The majority of the sequence of phCP-1 was obtained by analysis of randomly-sheared fragments cloned into M13. The remainder of the phCP-1 sequence, as well as the entire \( \lambda hCP-1 \) nucleotide sequence was determined by analysis of specific restriction endonuclease fragments cloned into M13 vectors. The complete nucleotide sequence of these two cDNA clones that together encode human preceruloplasmin is shown in Figure 7. The position of each nucleotide was determined an average of 3.4 times, and 62% of the sequence was obtained on both strands. In the region where they overlap (nucleotide residues 648 to 1197), the sequences of phCP-1 and \( \lambda hCP-1 \) were found to be identical.

Nucleotide residues 1 - 57 code for an amino terminal leader sequence, which is removed prior to the appearance of ceruloplasmin in plasma (Takahaski et al., 1984). Nucleotides 58 - 3195 of the cDNA sequence encode the plasma form of ceruloplasmin (Takahaski et al., 1984). The open reading frame is followed by a 'TGA' stop codon (encoded by nucleotides 3196 - 3198), a 3' untranslated region of 123 bp (nucleotide residues 3199 - 3321) and a poly(A) tail. The 3' untranslated sequence contains a putative polyadenylation signal, 'ATTAAA', which is situated 14 nucleotides upstream of the poly(A) tract.
Figure 7. **Nucleotide Sequence of Human Preceruloplasmin cDNA.**

The sequence was determined by analysis of the overlapping clones shown in Figure 6 (see text for details). The predicted amino acid sequence of human preceruloplasmin is indicated above the DNA sequence. The putative signal peptidase cleavage site is shown by a solid arrow. Potential carbohydrate sites (Takahashi et al., 1984) are represented by solid diamonds. Boxed sequences are complementary to oligonucleotide probes used to screen cDNA libraries. The polyadenylation signal ATTAAA is underlined.
```
<table>
<thead>
<tr>
<th>2.865</th>
<th>2,880</th>
<th>2,895</th>
<th>2,910</th>
<th>2,925</th>
<th>2,940</th>
<th>2,955</th>
<th>2,970</th>
<th>2,985</th>
<th>3,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>2,565</td>
<td>2,580</td>
<td>2,595</td>
<td>2,610</td>
<td>2,625</td>
<td>2,640</td>
<td>2,655</td>
<td>2,670</td>
<td>2,685</td>
<td>2,700</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>2,115</td>
<td>2,130</td>
<td>2,145</td>
<td>2,160</td>
<td>2,175</td>
<td>2,190</td>
<td>2,205</td>
<td>2,220</td>
<td>2,235</td>
<td>2,250</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1,815</td>
<td>1,830</td>
<td>1,845</td>
<td>1,860</td>
<td>1,875</td>
<td>1,890</td>
<td>1,905</td>
<td>1,920</td>
<td>1,935</td>
<td>1,950</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1,665</td>
<td>1,680</td>
<td>1,695</td>
<td>1,710</td>
<td>1,725</td>
<td>1,740</td>
<td>1,755</td>
<td>1,770</td>
<td>1,785</td>
<td>1,800</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1,515</td>
<td>1,530</td>
<td>1,545</td>
<td>1,560</td>
<td>1,575</td>
<td>1,590</td>
<td>1,605</td>
<td>1,620</td>
<td>1,635</td>
<td>1,650</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1,065</td>
<td>1,080</td>
<td>1,095</td>
<td>1,110</td>
<td>1,125</td>
<td>1,140</td>
<td>1,155</td>
<td>1,170</td>
<td>1,185</td>
<td>1,200</td>
</tr>
</tbody>
</table>

```

```
```
```
EcoRI inserts isolated from the above clones were subcloned into M13 and characterized by T-tracking analysis. Their positions relative to the previously characterized λhCP-1 cDNA clone are shown. The location of Probe C (an EcoRI-HindIII restriction fragment used as a hybridization probe for library screening) is identified by a solid bar. Arrowheads indicate that the clones extend 3' to Probe C (see text for details). λhCP-2 and λhCP-6 were found to extend 19 and 38 bp respectively 5' to λhCP-1. The additional nucleotide sequence contained in these two clones is given.
5' -38
CACTTCATTTCTCTCAGGCTCCAAGAAGGGGAAAAA

(EcoR1) Probe C

139

HincII

3'

λhCP-1

λhCP-2

λhCP-3

λhCP-4

λhCP-5

λhCP-6
EcoRI inserts isolated from λhCP-2, λhCP-3, λhCP-4, λhCP-5, and λhCP-6 clones were subcloned into M13 and initially characterized by T-tracking analysis (Sanger et al., 1980) in order to determine their positions relative to the λhCP-1 cDNA clone (see Figure 8). Two of these clones (λhCP-2 and λhCP-6) were found to extend 19 and 38 nucleotides respectively further 5' than λhCP-1. The additional 5' sequence contained in these two clones is also presented in Figure 8.

A.4 Ceruloplasmin Transcript Analysis

A Northern blot prepared using samples of both poly(A)$^+$ RNA from human liver and total RNA purified from HepG2 cells was hybridized to the $^{32}$P-labelled cDNA insert isolated from phCP-1. In both liver and HepG2 RNA samples, the cDNA hybridized to an mRNA species that is 3700 ± 200 nucleotides in size (Figure 9). The cDNA hybridized to an additional species of 4500 ± 250 nucleotides in the human liver poly(A)$^+$ RNA sample (Figure 9, lane 1). Given that eukaryotic mRNAs usually contain poly(A) tails of 180 – 200 nucleotides (Perry, 1976), both of the ceruloplasmin transcripts are much larger in size than that required to encode the entire coding region of the ceruloplasmin polypeptide chain (Takahashi et al., 1984).

A Northern blot containing samples of bovine liver poly(A)$^+$ RNA was probed at high stringency using the $^{32}$P-labelled insert isolated from phCP-1 (results not shown). Interestingly, only a single hybridizing species was detected which was 3800 ± 200 nucleotides in size.
Figure 9. **Blot hybridization analysis of human ceruloplasmin mRNA.**

RNA was separated by electrophoresis in a denaturing agarose/formaldehyde gel and transferred to nitrocellulose. The filter was hybridized with $^{32}$P-labelled phCP-1. The filter was exposed to X-ray film for 18 hours at $-70^\circ$C with intensifying screens. Lane 1, 10 μg of human liver poly(A)$^+$ RNA; lane 2, 20 μg of total HepG2 cell RNA. The positions of HindIII fragments of λ phage DNA used as size markers are shown. Sizes are given in Kilobases (Kb).
B. CHARACTERIZATION OF THE WILD-TYPE HUMAN CERULOPLASMIN GENE

B.1 Isolation and Restriction Endonuclease Mapping of Genomic Clones

Initially, five genomic equivalents \(1 \times 10^6\) plaque-forming units (pfu) of the Maniatis human genomic phage library, constructed in the phage lambda vector Charon 4A, were screened by using the phCP-1 cDNA clone as a hybridization probe. As described previously, this cDNA encodes amino acid residues 202–1046 of ceruloplasmin, and also contains a 3' untranslated region of 123 bp, terminating with a poly(A) tail. The fourteen positive clones identified from this screen were purified to homogeneity, and analyzed by restriction endonuclease mapping. On this basis, the clones were representative of only two independently-derived genomic clones. One of these clones (designated 3\(\psi\)10) corresponded to a pseudogene for human ceruloplasmin (see Section III.C). The other clone (designated \(\lambda\)WT1; see Figure 10) corresponded to the wild-type ceruloplasmin gene, as was determined initially by restriction endonuclease mapping and Southern blot analysis, and subsequently confirmed by using DNA sequence analysis (see below). In order to obtain phage clones corresponding to the 5' end of the ceruloplasmin gene, the Maniatis library was subsequently rescreened using the 1.2 Kbp EcoRI insert from \(\lambda\)hCP-1 (encoding the signal peptide and amino acids 1–380 of the mature protein) as a hybridization probe. From this screen, one positive clone (designated \(\lambda\)WT2; see Figure 10) was identified.

To isolate additional ceruloplasmin genomic clones, \(1 \times 10^6\) pfu from a second genomic library (constructed in the lambda phage vector EMBL 3) were screened, using \(^{32}\)P-labelled cDNA inserts from both phCP-1 and \(\lambda\)hCP-1 as hybridization probes. From this screen, 10 different
Figure 10. Partial restriction map and intron/exon organization of the human ceruloplasmin gene.

A complete map of the 45 Kbp region using the enzymes BamHI (B), SalI (S), HindIII (H), and EcoRI (E) is shown; an incomplete map for AccI (A), XbaI (X), Kpn (K), SstI (T), and BalI (L) is also given. Exons located within the coding region are shown as black boxes in a bar above the restriction map and are numbered 1 - 14. Corresponding introns are also labelled (A - N). Approximate positions of introns (LA - LD) and exons (L1 - L4) present in the 5' untranslated region of the gene are indicated; lengths of these exons and introns are undetermined. Genomic phage clones λWT1 - λWT9 are shown below the restriction map. Open circles at the ends of EMBL 3 phage clones represent Sau3A sites, while solid diamonds at the ends of clones isolated from the Maniatis library represent EcoRI linkers. The relative positions of the 4.6 Kbp and 2.4 Kbp EcoRI fragments identified from genomic Southern analysis are also shown, corresponding to the 3' end of the gene (see text for details). The scale represents 1 Kilobase pair.
recombinant clones were obtained. Seven of these were found to correspond to the wild-type ceruloplasmin locus (designated λWT3 to λWT9; see Figure 10), while the remaining three clones (designated 3ψ21, 3ψ29, and 3ψ9) were identified as human ceruloplasmin pseudogene clones (see Section III.C).

DNA was prepared from small scale lysates of wild-type phage clones, and analyzed by multiple restriction enzyme digestions and Southern blot analyses using appropriate hybridization probes derived from cDNA clones. On this basis, a partial restriction enzyme map of cloned ceruloplasmin genomic DNA was constructed (Figure 10). The nine wild-type ceruloplasmin clones identified (Figure 10) span a region of approximately 45 Kbp of genomic DNA. Southern blot analysis using M13 probes derived from the 3' end of the ceruloplasmin cDNA (i.e., containing sequences derived from nucleotide residues 2565 – 3321, Figure 7) indicated that the 3' end of the ceruloplasmin gene was not represented in the above phage clones.

B.2 Localization of Intron/Exon Junctions Corresponding to the Ceruloplasmin Coding Region

In order to identify exon-containing DNA sequences corresponding to the cDNA sequence of human preceruloplasmin, either specific restriction endonuclease fragments hybridizing to cDNA-derived probes or fragments generated by sonication of appropriate genomic clones were ligated into M13 vectors for analysis (see Table 3, section II.G.2). In the latter case, coding sequences were identified by screening M13 subclones with \(^{32}\)P-labelled cDNA fragments as hybridization probes. The intron/exon boundary sequences for exons 1 – 14 are presented in Table V. All introns characterized to date are spliced according to the GT...AG rule
Table V. Nucleotide Sequence of Intron/Exon Junctions in the Human Ceruloplasmin Gene.

<table>
<thead>
<tr>
<th>EXON</th>
<th>5' SPLICE DONOR</th>
<th>INTRON</th>
<th>3' SPLICE ACCEPTOR</th>
<th>CODON PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4</td>
<td>not determined</td>
<td>c g c t t c t c c</td>
<td>a a a a a A G G G</td>
<td>II</td>
</tr>
<tr>
<td>1</td>
<td>A C A C g t a a g a</td>
<td>t g t c t t g</td>
<td>t t t t t t t t g c a g G G A</td>
<td>II</td>
</tr>
<tr>
<td>2</td>
<td>G A G G g t a a g t</td>
<td>c a t t g c a t g</td>
<td>t t g c t t c c t a g G G G C</td>
<td>I</td>
</tr>
<tr>
<td>3</td>
<td>A A A G g t a c a t</td>
<td>a a a g t a a c t t t a a c t c c a g A T T C</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>C T C C g t a a g a</td>
<td>a c t c t g c t t c t g a c t t a c a c a g C T G T</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A A A G g t a g g a</td>
<td>c t c t t t t t g t t t c a t t t c a g C C G G</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>G A A G g t a a t t</td>
<td>a a c a c t t t t t t t c c c c c c t g a g G T G A C</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>C T G G g t g a g t</td>
<td>a a t t c c t c t c t t t g t t c c c c a g G T C C</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>A G A A g t g a g t</td>
<td>t a a t g g a c c t t t t c t c c a c a g C G T G T</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>A C A C g t a a g t</td>
<td>t c a t t t t t g t t t t a t t a c a c a g G A T C</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>C A C T g t a a g t</td>
<td>t t a a t t t c c e a a c t t t t a c a c a g C C A T</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>not determined</td>
<td>k t t t c t a t t t c c a c t c c a g G G A C</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>A G A A g t a a t t</td>
<td>t t t c t a t t t c c a c t c c a g G G A C</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>C T A G g t a g t t</td>
<td>t t c c t c t c a c t t t t g c c a g G T C C</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>C C A C g t a c t c</td>
<td>n t t t t t t t t c c a c t c c a g G G A C</td>
<td>II</td>
<td></td>
</tr>
</tbody>
</table>

Exon sequence is shown in upper case, while intron sequence is given in lower case. The codon phase refers to the position of the intron relative to the codon triplet (Sharp, 1981), i.e., 0: intron occurs between codons, I: intron occurs after the first nucleotide of the codon, and II: intron occurs following the second nucleotide of the codon. "L" designates introns or exons occurring within the 5' untranslated region.
Table VI. Frequencies of Nucleotides at Intron/Exon Junctions.

**DONOR FREQUENCIES**

<table>
<thead>
<tr>
<th></th>
<th>+4</th>
<th>+3</th>
<th>+2</th>
<th>+1</th>
<th>-1</th>
<th>-2</th>
<th>-3</th>
<th>-4</th>
<th>-5</th>
<th>-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>8</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**ACCEPTOR FREQUENCIES**

<table>
<thead>
<tr>
<th></th>
<th>-20</th>
<th>-19</th>
<th>-18</th>
<th>-17</th>
<th>-16</th>
<th>-15</th>
<th>-14</th>
<th>-13</th>
<th>-12</th>
<th>-11</th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
<th>+4</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

CON   N   A   A   G   G   T   R   A   G   T

CON   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   Y   N   Y   A   G   G   N   N   N   N

The frequencies of nucleotides occurring at intron/extron boundaries in the human ceruloplasmin gene are compared to the consensus sequence (CON) of Mount (1982). "N" represents G, A, T or C, while "Y" denotes pyrimidine residues (i.e. C or T). Splice junctions are located between -1 and +1.
(Breathnach and Chambon, 1981; Cech, 1983). The consensus sequences surrounding splice junctions in eukaryotic RNA polymerase II-transcribed genes (Breathnach and Chambon, 1981; Mount, 1982) were also found to be in agreement with those characterized in the human ceruloplasmin gene (see Table VI). Although the splice donor and acceptor sequences were not determined for introns K and L respectively, intron positions in both cases could be assigned unequivocally from sequences obtained on one side. Introns present in the coding sequence corresponding to the 3' end of the gene (i.e. the region containing nucleotide residues 2656 - 3321 of the cDNA sequence) were not determined, since phage clones containing this area were not obtained from genomic screening (see Figure 10).

B.3 Partial Nucleotide Sequence Analysis of the Human Ceruloplasmin Gene

Partial DNA sequence of the ceruloplasmin gene corresponding to the coding region was determined by analysis of M13 subclones. Approximately 40% of the sequence data was determined on both strands, and DNA sequence obtained on one strand only was determined at least twice. Sizes and relative positions of introns and exons in the gene are summarized in Table VII and shown schematically in Figure 10. Intron sizes were determined in all cases by restriction endonuclease mapping, and therefore represent close approximations to actual sizes. Exons range in size from 129 - 255 bp, with a calculated average length of 183 bp. Introns typically were found to be variable in size, ranging from approximately 800 bp to 9.5 Kbp. The partial amino acid sequence determined from the characterization of exons 1 - 14 was identical to that
Table VII. Sizes and positions of introns and exons within the ceruloplasmin gene.

<table>
<thead>
<tr>
<th>Exon</th>
<th>Nucleotide Position</th>
<th>Length (bp)</th>
<th>Intron</th>
<th>Nucleotide Position</th>
<th>Length (bp)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-12 - 146</td>
<td>158</td>
<td>A</td>
<td>159 - 2459</td>
<td>2300</td>
</tr>
<tr>
<td>2</td>
<td>2460 - 2708</td>
<td>248</td>
<td>B</td>
<td>2709 - 4759</td>
<td>2050</td>
</tr>
<tr>
<td>3</td>
<td>4760 - 4973</td>
<td>213</td>
<td>C</td>
<td>4974 - 5774</td>
<td>800</td>
</tr>
<tr>
<td>4</td>
<td>5775 - 5949</td>
<td>174</td>
<td>D</td>
<td>5950 - 7400</td>
<td>1450</td>
</tr>
<tr>
<td>5</td>
<td>7401 - 7656</td>
<td>255</td>
<td>E</td>
<td>7657 - 8957</td>
<td>1300</td>
</tr>
<tr>
<td>6</td>
<td>8958 - 9130</td>
<td>172</td>
<td>F</td>
<td>9131 - 12231</td>
<td>3100</td>
</tr>
<tr>
<td>7</td>
<td>12232 - 12372</td>
<td>140</td>
<td>G</td>
<td>12373 - 14273</td>
<td>1900</td>
</tr>
<tr>
<td>8</td>
<td>14274 - 14427</td>
<td>153</td>
<td>H</td>
<td>14428 - 15578</td>
<td>1150</td>
</tr>
<tr>
<td>9</td>
<td>15579 - 15791</td>
<td>212</td>
<td>I</td>
<td>15792 - 25192</td>
<td>9400</td>
</tr>
<tr>
<td>10</td>
<td>25193 - 25344</td>
<td>151</td>
<td>J</td>
<td>25345 - 26645</td>
<td>1300</td>
</tr>
<tr>
<td>11</td>
<td>26646 - 26859</td>
<td>213</td>
<td>K</td>
<td>26860 - 28110</td>
<td>1250</td>
</tr>
<tr>
<td>12</td>
<td>28111 - 28319</td>
<td>208</td>
<td>L</td>
<td>28320 - 29920</td>
<td>1600</td>
</tr>
<tr>
<td>13</td>
<td>29921 - 30061</td>
<td>140</td>
<td>M</td>
<td>30062 - 31311</td>
<td>1250</td>
</tr>
<tr>
<td>14</td>
<td>31312 - 31441</td>
<td>129</td>
<td>N</td>
<td>31442 -&gt; 32242</td>
<td>800+</td>
</tr>
</tbody>
</table>

*Sizes of all introns were estimated from restriction enzyme analysis.
determined for the cDNA sequence corresponding to nucleotide residues 1 - 2555.

B.4 Organization of the 5' End of the Human Ceruloplasmin Gene

B.4.1 Comparison of genomic and cDNA sequence data. Initially, the nucleotide sequence of the λhCP-6 cDNA clone (Figure 8) was compared with overlapping genomic DNA sequence (corresponding to nucleotide residues -288 - +412) derived from sonication of a 3.1 Kbp EcoRI fragment (see Table III). Alignment of the nucleotide sequences is presented in Figure 11. The observed divergence of the genomic sequence with that of the cDNA sequence correlated with the presence of a 3' consensus splice acceptor site at nucleotide position -13 in the 5' untranslated region of the human ceruloplasmin gene. Location of this splice site was subsequently confirmed using nuclease S1 mapping analysis (see below).

B.4.2 Nuclease S1 mapping analysis of exon 1. The presence of a splice acceptor site at -13 was confirmed using the 700 bp genomic clone described above as a hybridization probe for nuclease S1 mapping analysis (see Figure 12). The observed protected fragment of 158 bp indicates the presence of an intron splice site at -13 bp, since a splice junction had been identified following nucleotide residue 146 in the coding region. This result is in agreement with DNA sequence data (Figure 11), and confirms that the λhCP-6 cDNA clone contains 26 bp corresponding to an exon in the 5' untranslated region.

B.4.3 Southern blot analysis. The location of an exon in the 5' untranslated region (designated L4; see Figures 8 and 10) was determined by Southern blot analysis of an AccI/EcoRI digest of the 3.1 Kbp EcoRI
Figure 11. Comparison of the sequence of the \( \lambda hCP-6 \) cDNA clone with overlapping genomic DNA sequence derived from the phage clone \( \lambda WT2 \).

\( +1 \) indicates the first nucleotide of the initiator methionine residue. The point of sequence divergence between genomic and cDNA sequences coincides with the location of a 3' splice acceptor site in the genomic sequence (vertical arrow). Sequences corresponding to exon L4 (from the cDNA) or intron LD (from the genomic clone) are indicated.
Intron LP

38

Genomic

GTCCGCCGCTTTCTCCCTCGGAAGAGGGGAAAAAAA
CACCTCATTTCTTCAGGCTCCAAGAGGGGAAAAAAA
cDNA

Exon L4

Intron LD
Figure 12. **Nuclease S1 mapping analysis of exon 1**.

A 700 bp fragment (containing nucleotide residues -288 - +412 of the genomic DNA sequence) was used as a probe for S1 nuclease protection analysis. Following hybridization to 0.5 μg of human liver poly(A)$^+$ RNA and nuclease S1 digestion, S1-resistant products were separated on a denaturing polyacrylamide gel and visualized by autoradiography. A protected band of 158 bp was detected (lane 2), corresponding to exon 1. A *Hinf*I digest of pBR322 was used as markers (lane 1). The sizes of resulting fragments are given in base pairs (bp).
Figure 13. Southern blot analysis localizing exon L4 in the 5' untranslated region of human ceruloplasmin.

AccI/EcoRI (panel A, lane 1) or AccI (panel A, lane 2) digests of the 3.1 Kbp genomic EcoRI fragment containing exon 1 were transferred to nitrocellulose and probed with the λhCP-6 cDNA clone. In addition to the pUC vector band (designated "P"), hybridizing species corresponding to 1.7 Kbp and 0.85 Kbp AccI fragments were detected. Positions of HindIII fragments of phage λ DNA used as size markers are shown. Fragment sizes are given in Kilobase pairs (Kbp). Restriction analysis of the 3.1 Kbp EcoRI (E) genomic fragment using AccI (A) is shown in panel B. The relative positions of exon L4 and exon 1 are indicated; the latter exon is designated by a solid bar. The precise location of exon L4 is uncertain, as indicated by the dashed lines. An arrow identifies the direction of transcription.
genomic fragment (Figure 13; see above), using \( \lambda hCp-6 \) cloned into M13 to generate a hybridization probe. The hybridization and washing conditions used were similar to those employed for oligonucleotide probes (see Section II.B.1). Hybridizing DNA fragments of 1.7 Kbp and 0.85 Kbp were detected, the latter of which contains Exon 1 (see Figure 10) based on previous Southern analysis (data not shown). This indicates that the 3.0 Kbp EcoRI genomic fragment contains at least two exons (see Figure 10).

B.4.4 Northern blot analysis of the 5' end of the human ceruloplasmin gene. For further analysis of the 5' end gene organization, EcoRI fragments derived from \( \lambda WT7 \) and \( \lambda WT2 \) phage clones were subcloned (see Table III) and used as hybridization probes for Northern blot analysis of poly(A)+ RNA. Hybridization to both 3.7 Kb and 4.5 Kb ceruloplasmin-specific transcripts (see Section II.A.4) was detected with all fragments tested (see Fig. 14) except the 1.4 Kbp EcoRI fragment. The hybridization signal obtained using the 420 bp EcoRI probe was observed to be very weak following a 5-day exposure, suggesting the presence of a short exon sequence within this fragment. For more precise assignment of mRNA-encoding sequences in the 5' untranslated segment, the 4.0 Kbp EcoRI fragment was subcloned further utilizing an internal HindIII site (see Table III and Figure 10), and the resulting two fragments were subsequently used as hybridization probes for Northern blot analysis. As shown in Figure 14A, hybridization was detected with only the 1.3 Kbp HindIII/EcoRI fragment.

B.4.5 RNA dot blot analysis. In order to localize further the exons within the 5' untranslated region of the human ceruloplasmin gene, EcoRI genomic fragments in pUC13 were subcloned into M13 vectors (see Table
Figure 14. **Northern blot analysis of the 5' end of the ceruloplasmin gene.**

Samples of human liver poly(A)$^+$ RNA (10 µg each) were electrophoresed in a denaturing agarose-formaldehyde gel and transferred to nitrocellulose. Filters were hybridized with $^{32}$P-labelled EcoRI (E) and HindIII/EcoRI (HE) restriction fragments derived from the 5' end of the gene (panel A). Fragment sizes of probes (Kilobases) are shown above each blot; bracketed numbers represent lengths of exposure times in days. Sizes of the two hybridizing RNA species detected are given in Kilobases, based on positions of $^{32}$P-labelled HindIII fragments of λ DNA. "C" designates a positive control, since this 3.1 Kbp genomic fragment contains exon 1 (see text for details). Blots are shown sequentially in a 5' to 3' direction from left to right. For clarity, a restriction map in EcoRI (E) and HindIII (H) corresponding to this region is shown in panel B; fragment sizes are given in Kilobase pairs (Kbp). The bracketed E identifies an EcoRI linker. An arrow indicates the direction of transcription.
Figure 15. RNA dot blot analysis of the 5' untranslated region of the ceruloplasmin gene.

Human liver poly(A)$^+$ RNA (7.5 μg) was spotted onto nitrocellulose and probed with M13 clones derived from appropriate restriction fragments (see Table III). A partial map of the region is shown; abbreviations used are E - EcoRI; S - Sau3A; F - HinfI; D - DdeI; H - HindIII. The bracketed E represents an EcoRI linker. The extent and direction of sequence analysis of the M13 clones used for dot blot hybridizations are shown by thin arrows below the restriction map. The thick arrow designates the direction of transcription. Corresponding exons L1, L2, and L3 identified by dot blot analysis are shown. The scale represents 0.1 Kilobase pairs (Kbp).
III). M13 clones were analyzed initially by DNA sequence analysis and were subsequently used to generate double-stranded probes for dot blot hybridization analysis (Figure 15). The results indicate the presence of at least 4 exons in the 5' untranslated sequence of the gene, with a minimum of 2 separate exons within the 1.5 Kbp EcoRI fragment (designated L1 and L2; see Figure 10), one exon within the 1.3 Kbp HindIII/EcoRI fragment (designated L3; see Figure 10) and one exon (L4) determined previously within the 1.7 Kbp AccI fragment (see above).

C. Isolation and Complete Characterization of a Pseudogene for Human Ceruloplasmin

C.1 Isolation of Genomic DNA Clones Containing the Human Ceruloplasmin Pseudogene

As described previously (see Section III.B.1), one ceruloplasmin pseudogene clone (designated 3ψ10) was obtained from the screening of a human genomic library constructed in the Charon 4A vector, using the phCP-1 cDNA clone as a hybridization probe. The identity of this initial clone as a pseudogene for human ceruloplasmin was based on restriction endonuclease mapping and Southern blot analysis (results not shown). This was subsequently confirmed using DNA sequence analysis and nuclease S1 analysis (see following sections). Three additional pseudogene clones (designated 3ψ9, 3ψ21, and 3ψ29) were isolated from a human genomic library in the phage vector λ EMBL 3, using both phCP-1 and λhCP-1 cDNA clones as hybridization probes. Restriction enzyme mapping analysis indicated that the latter three phage clones overlapped the 3ψ10 clone (Figure 16). A total of nearly 21 Kbp of contiguous genomic DNA is
Figure 16. Partial Restriction Map and Sequencing Strategy for the Human Ceruloplasmin Pseudogene.

The complete restriction map for EcoRI (E) is shown. The lines above the restriction map represent the four overlapping phage clones 3ψ9, 3ψ21, 3ψ29 (isolated from the Geddes library), and 3ψ10 (from the Maniatis library). The solid bar represents the region of the ceruloplasmin pseudogene that is homologous to the ceruloplasmin cDNA sequence. Solid circles at the ends of EMBL 3 phage clones represent Sau3A sites, while solid squares at the end of the 3ψ10 phage clone isolated from the Maniatis library represent EcoRI linkers. The locations of BamHI (B) and HindIII (H) sites used for subcloning and sequence analysis are shown within the 3ψ10 clone; the remainder of the map is incomplete for these two enzymes. The region containing the BamHI/HindIII fragment has been expanded below the restriction map. Arrows below this region indicate the direction and extent of nucleotide sequence obtained from various M13 clones. Probe A was used as a hybridization probe for chromosomal localization studies (see text for details). The scale represents 1 Kbp.
represented by these four recombinant clones, with the human ceruloplasmin pseudogene mapping to approximately 1.7 Kbp within this region (see Figure 16).

C.2 DNA Sequence Analysis of the Human Ceruloplasmin Pseudogene

The nucleotide sequence of the ceruloplasmin pseudogene (Figure 17) was determined by using the strategy outlined in Figure 16. Each nucleotide was determined an average of 3.4 times, and 54% of the sequence was obtained on both strands. Nucleotides 53 - 1644 of the pseudogene sequence are very similar to residues 1502 - 3318 of the ceruloplasmin cDNA (nearly 97% identical; see Figure 17), extending through the 123 bp segment corresponding to the 3' untranslated region of the phCP-1 cDNA clone (see Section III.A). The pseudogene, however, is not characterized by a poly(A) tract at the expected polyadenylation site. DNA sequence analysis in a 3' direction revealed the presence of an unusual 54 bp segment, composed primarily of repeated CT dinucleotides (nucleotide residues 1867 - 1920; see Figure 17).

Following the sequence corresponding to the 3' untranslated region of phCP-1, the next 43 bp of pseudogene sequence (nucleotides 1645 - 1687) correspond to the 3' untranslated sequence of a ceruloplasmin cDNA clone (designated Cp-1) characterized by Yang et al. (1986). This clone differs from phCP-1 in that the 3' untranslated region extends for an additional 120 bp prior to the point of poly(A) addition. Following this 43 bp segment, the remainder of the pseudogene sequence prior to the poly(CT) segment shares little similarity with the 3' untranslated region from the cDNA clone described by Yang et al. (1986).
Figure 17. Nucleotide sequence of the human ceruloplasmin pseudogene and comparison with the corresponding region of the ceruloplasmin cDNA sequence.

The pseudogene sequence was determined by analysis of the overlapping clones shown in Figure 16. I/E denotes positions of intron-exon boundaries in the wild-type gene (see Section III.B). The position of the AGCT insertion that results in a frameshift mutation is enclosed in a box. The sizes and positions of deletions (Δ) relative to the ceruloplasmin cDNA sequence are also shown; the region of the cDNA corresponding to the 213 bp deletion has been omitted. The places (5' and 3') where the pseudogene sequence diverges from the wild-type sequence are indicated by arrowheads (see text for details). The cDNA sequence represents nucleotides 1502 - 1864 and 2078 - 3318 of phCP-1. Asterisks indicate identical nucleotides in corresponding positions in the two sequences. Dashes were introduced at points of insertion or deletion in order to maximize homology.
The 5' end of the pseudogene sequence (nucleotide residues 10 - 52; see Figure 17) corresponds to the sequence of 3' end of intron present in the wild-type gene (data not shown). A consensus 3' splice acceptor sequence (Breathnach and Chambon, 1981; Cech, 1983) is located immediately prior to nucleotide residue 53. This splice junction is in the corresponding position to a splice site present in the wild-type gene (see Table V). On the basis of DNA sequence analysis, nucleotide residue 9 marks the point of divergence from the wild-type gene. Southern blot analysis confirms that the 2.6 Kbp EcoRI fragment located directly 5' to the 0.8 Kbp EcoRI fragment (see Figure 16) does not hybridize to the corresponding area of the wild-type gene (results not shown). In contrast to most processed pseudogenes characterized to date (Vanin, 1985), the points of divergence at the ceruloplasmin pseudogene are not flanked by short direct repeats.

A comparison of the nucleotide sequences of the pseudogene and ceruloplasmin cDNA is also shown in Figure 17. A deletion of 213 bp was observed in the pseudogene sequence, corresponding to nucleotides 1865 - 2077 of the human ceruloplasmin cDNA sequence. Prior to this deletion, the pseudogene sequence contains an open reading frame corresponding to amino acid residues 483 - 602 of the ceruloplasmin coding sequence. Occasional base changes within this open reading frame result in the occurrence of amino acid substitutions (Figure 17). The 213 bp deletion in the pseudogene sequence maintains this open reading frame, which then continues for an additional 9 amino acid residues. At this point, an insertion of 4 duplicated nucleotide residues (AGCT) (nucleotides 445 - 448; see Figure 17) causes a frameshift mutation, such that a TCA
termination codon occurs immediately following this insertion. The remainder of the pseudogene sequence is similar to the phCP-1 cDNA clone and the cDNA clones reported by Yang et al. (1986) but contains a number of base substitutions compared with the cDNA sequence. There is also a small deletion of 17 bp corresponding to nucleotide residues 2942 - 2958 of the cDNA (see Figure 17). However, this deletion does not result in the resumption of an open reading frame in the pseudogene sequence.

C.3 Nuclease S1 Analysis of the Human Ceruloplasmin Pseudogene

The presence of the 213 bp deletion in the pseudogene sequence as compared to the cDNA sequence was utilized to analyze the presence of pseudogene-specific transcripts. A single-stranded DNA fragment (286 nucleotides in length) derived from the human ceruloplasmin cDNA sequence was used as a hybridization probe for nuclease S1 analysis. This probe corresponds to a region spanning nucleotide residues 1927 - 2213 of the cDNA sequence, thereby containing 150 bp within the deleted region identified in the pseudogene sequence (see Figure 18B). Protection of human liver poly(A)\(^+\) RNA with this probe resulted in a single protected band of 286 bp, corresponding to the wild-type transcript (Figure 18A). Additionally, a band corresponding to the full-length probe was observed (Figure 18A) which contains M13 sequences in addition to ceruloplasmin sequences. If the pseudogene was transcribed in liver, and assuming that nuclease S1 cleaves all single-base mismatches between the probe and the putative transcript, protected DNA fragments of 47 bp and 55 bp would be expected (corresponding to nucleotides 453 - 499 and 501 - 555 respectively; see Figure 17). However, no protected fragment of a smaller size, which would correspond to an RNA species containing the 213 bp
Figure 18. **Nuclease S1 transcript analysis for the ceruloplasmin pseudogene.**

Panel b shows the location of the 286 bp probe used for S1 nuclease protection analysis relative to the 213 bp deletion observed in the pseudogene sequence. Part of this probe (150 bp) is within the deleted region (corresponding to nucleotides 1865 - 2077 of the cDNA), while 136 bp of the probe are 3' to the deletion. Following hybridization to 1 μg (lane A) or 0.35 μg (lane B) of poly(A)$^+$ mRNA and nuclease S1-digestion, nuclease-resistant products were separated on a denaturing polyacrylamide gel and visualized by autoradiography (panel a). A band corresponding to the size of the undigested full-length probe (FLP) was observed, as well as a protected band of 286 bp corresponding to the wild-type transcript. A *Hinf1* digest of pBR322 was used for markers. The sizes of resulting fragments are given in base pairs.
probe (286 bp)

wild-type transcript

putative pseudogene transcript

FLP

286 bp

506/517
396
344
298
220/221
154
75

5' 1865 2077 3'

△ 213 bp
Figure 19. Chromosome Mapping of the Human Ceruloplasmin Pseudogene Using Somatic Cell Hybrid Analysis.

A 1.1 Kbp pseudogene-specific probe (Probe A; see Figure 16) was hybridized to *EcoRI*-digested DNA from human-hamster hybrid cell lines (lanes 1-22). Numbering of the cell lines corresponds to that shown in Table 8. *EcoRI*-digested hamster DNA (lane H) and human placental DNA (lane HP) were included as controls. The positions of $^{32}$P-labelled *HindIII*-*EcoRI* fragments of λ DNA used as size markers are shown. Fragment sizes are given in Kilobase pairs (Kbp).
| Cell Line | Response to Probe A | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | X | Y |
| 1         | 41.06             | - | - | - | + | - | - | + | - | + | - | + | - | - | - | + | - | + | - | - | + | - | - |
| 2         | 45.01             | - | - | - | + | - | - | - | + | - | - | + | - | - | + | - | - | + | - | + | - | - | + |
| 3         | 45.43             | + | - | - | - | - | - | + | - | + | - | - | - | + | + | - | + | - | - | + | - | + | + |
| 4         | 76.14             | + | - | - | + | - | - | + | - | + | - | - | - | + | - | - | - | + | - | + | - | + | - |
| 5         | 76.31             | - | + | - | + | - | - | + | - | - | + | - | - | + | - | - | + | - | + | - | + | - |
| 6         | 76.33             | + | - | + | - | - | + | - | + | - | + | - | + | + | + | - | + | - | + | - | + | - |
| 7         | 79.05b            | - | - | - | - | - | - | - | - | - | + | - | - | + | + | - | + | - | + | - | + | - |
| 8         | 80.05d            | - | - | - | - | + | - | - | - | - | + | - | - | + | + | - | + | - | + | - | + | - |
| 9         | 80.14c            | - | - | - | - | - | - | - | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 10        | 80.17a            | - | - | - | - | - | - | - | + | - | - | + | - | + | + | - | + | - | + | - | + | - |
| 11        | 82.82a            | - | - | - | - | - | - | - | - | - | - | + | - | + | + | - | + | - | + | - | + | - |
| 12        | 85.16a            | - | - | + | + | + | + | - | + | - | - | - | + | - | + | - | + | - | + | - | + | - |
| 13        | 89.27a            | - | - | - | - | - | - | - | - | - | - | + | - | + | + | - | + | - | + | - | + | - |
| 14        | 100.02b           | + | - | - | - | - | - | - | - | + | - | - | - | + | + | - | + | - | + | - | + | - |
| 15        | 102.05b           | + | - | - | - | - | - | - | + | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 16        | 103.04            | - | - | - | - | + | - | + | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 17        | 111.02a           | - | - | - | - | + | - | + | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 18        | 112.10a           | + | - | - | - | - | - | + | - | - | - | + | - | + | + | - | + | - | + | - | + | - |
| 19        | 120.33            | - | - | - | - | - | + | - | + | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 20        | 120.35            | + | - | - | - | - | - | + | - | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 21        | 133.05            | + | - | - | - | - | - | - | + | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| 22        | 134.02a           | + | - | - | - | + | - | - | + | - | - | - | - | - | + | + | - | + | - | + | - | + | - |
| X Discordancy |                  | 27 18 32 23 45 45 32 0 41 27 27 41 27 45 36 64 50 50 32 55 55 45 41 41 |

aPresence (+) or absence (-) of human chromosomes as determined by cytogenetic analysis and confirmed by isozyme analysis (Donald et al., 1983; Riddell et al., 1985).

bPresence (+) or absence (-) of human EcoRI-digested sequences homologous to the human ceruloplasmin pseudogene probe (Probe A).
deletion was observed at either of the RNA concentrations used (Figure 18A).

C.4 Chromosomal Localization of the Human Ceruloplasmin Pseudogene

The chromosomal location of the ceruloplasmin pseudogene was determined by somatic cell hybrid analysis. For this purpose, an isolated 1.1 Kbp EcoRI restriction fragment (Probe A, Figure 16), located 3' to the pseudogene, was used as a hybridization probe. Southern blot analysis of EcoRI-digested human-hamster somatic cell hybrid DNA (Figure 19) showed that all 22 hybrid cell lines tested (lanes 1 – 22) were concordant for the presence or absence of the 1.1 Kbp band and chromosome 8 (Table VIII). Control lanes containing either EcoRI-digested human placental DNA (lane HP, Figure 19) or hamster DNA (lane H, Figure 19), were included in the Southern blot analysis. No cross-hybridization was detected in the lane containing hamster DNA, while the expected 1.1 Kbp EcoRI band corresponding to Probe A was observed in the human placental DNA.

D. Genomic Southern Blot Analysis of the Human Ceruloplasmin Gene

Initially, human genomic DNA isolated from liver tissue was digested with several restriction endonucleases, and the resulting fragments were separated by agarose gel electrophoresis. Following transfer to nitrocellulose, DNA fragments were analyzed with $^{32}$P-labelled hybridization probes derived from the previously characterized human ceruloplasmin cDNA clones phCP-1 and $\lambda$hCP-1 (see Section III.A). Using the $\lambda$hCP-1 hybridization probe (Figure 20B), a large single band (> 23 Kbp) was detected with a BamHI digest; this is predicted based on restriction endonuclease mapping of the wild-type gene
Figure 20. **Genomic Southern Blot Analysis of the Human Ceruloplasmin Gene.**

Human liver DNA (10 μg) was digested with various restriction enzymes (BamHI, EcoRI, PstI, and HindIII) and electrophoresed on 1.0% agarose gels. Following transfer to nitrocellulose, the blot was hybridized using $^{32}$P-labelled phCP-1 (panel A) or λhCP-1 (panel B) as probes. Positions of $^{32}$P-labelled HindIII fragments of phage λ DNA used as size markers are given in Kilobase pairs (Kbp).
(see Figure 10) and suggests that a single gene is being detected with this probe. Other hybridizing fragments present in EcoRI (15 Kbp, 4.6 Kbp, 3.1 Kbp, 1.8 Kbp), PstI (9.5 Kbp) and HindIII (5.7 Kbp, 4.5 Kbp) genomic digests all correspond to fragments identified in the partial restriction enzyme map of the wild-type gene (see Figure 10).

Using the phCP-1 cDNA clone as a hybridization probe for EcoRI, BamHI, HindIII and PstI human genomic digests (Figure 20A), the resulting pattern is more complex. Multiple bands were detected with each restriction endonuclease used. EcoRI bands which could not be identified in the restriction map of the wild-type gene (see Figure 10) were assigned subsequently to either a pseudogene for human ceruloplasmin (corresponding to nucleotide residues 1502 - 3318 of the cDNA sequence; see Section III.C) or to the 3' end of the wild-type gene (see below). In the former case, the 0.45 Kbp, 0.8 Kbp, and 1.5 Kbp EcoRI fragments (indicated by arrows in Figure 21) correspond to the pseudogene locus (see Section III.C and Figure 16). The remainder of the EcoRI fragments in Figure 20A, as well as all of those detected using the λhCP-1 hybridization probe (Figure 20B) have been assigned unambiguously to the wild-type ceruloplasmin gene (see Figures 10 and 22).

The 3' end of the ceruloplasmin gene was characterized by Southern blot analysis of EcoRI-digested human genomic DNA (Figure 22A) using two specific hybridization probes derived from the cDNA sequence (Figure 22B). Two bands were detected with each hybridization probe, one of which in each case corresponds to the pseudogene locus (see Figure 21). The remaining two fragments (4.6 Kbp and 2.4 Kbp) were thus unequivocally
Figure 21. Genomic Southern Analysis of the Human Ceruloplasmin Pseudogene and Related Sequences.

EcoRI-digested human liver genomic DNA was hybridized with either the phCP-1 (panel A) or λhCP-1 (panel B) ceruloplasmin cDNA clones, under conditions of high stringency. Arrows indicate bands that have been assigned to the human ceruloplasmin pseudogene (see Figure 16). The remaining bands correspond to the wild-type locus (see text for details). Positions of $^{32}$P-labelled HindIII fragments of phage λ DNA used as size markers are shown. Fragment sizes are given in Kilobase pairs (Kbp).
Figure 22. Genomic Southern blot Analysis of the 3' End of the Human Ceruloplasmin Gene.

EcoRI-digested human lymphocyte genomic DNA was hybridized using either Probe 1 or Probe 2 (panel A), derived from the 3' end of the human ceruloplasmin cDNA sequence (panel B). Arrows indicate bands that have been assigned to the ceruloplasmin pseudogene locus (see text for details); the remaining bands correspond to the 3' end of the wild-type ceruloplasmin gene. Positions of $^{32}$P-labelled HindIII fragments of phage λ DNA used as size markers are shown. Fragment sizes are given in Kilobase pairs (Kbp).
assigned to the wild-type gene (see Figure 10), with the 2.4 Kbp EcoRI hybridizing species representing the 3'-most fragment.

IV. DISCUSSION

A. Characterization of the Human Pre Ceruloplasmin cDNA

A.1 DNA Sequence Analysis of the Human Pre Ceruloplasmin cDNA

Initially, a human ceruloplasmin cDNA clone of 2.7 Kbp (phCP-1) was isolated from a cDNA library prepared from human liver mRNA, by using mixtures of synthetic oligonucleotides as hybridization probes. This cDNA clone was found to encode amino acid residues 202 - 1046 of the ceruloplasmin protein sequence, followed by 123 bp of 3' untranslated sequence (preceded by a TGA stop codon) and terminating with a poly (A) tract. Two randomly-primed human liver cDNA libraries were subsequently screened using appropriate 32P-labelled restriction fragments. Two clones of 1.2 Kbp and 1.0 Kbp (designated λhCP-1 and λhCP-6, respectively) were identified that together contained cDNA sequence encoding amino acid residues 1 - 380 of the mature protein preceded by a putative 19 amino acid long signal peptide and 38 bp of 5' untranslated sequence.

Nucleotides 58 - 3195 of the cDNA sequence code for the plasma form of ceruloplasmin; the predicted amino acid sequence agrees completely with that reported previously by Takahashi et al. (1984) who used protein chemistry techniques. Overall, the base composition of ceruloplasmin mRNA is somewhat A + U rich (33% A, 26% U, 22% G, 19% C), reflective of the coding region in which 60% of the codons end in either A or U. This observation is in contrast to the codon usage in other liver mRNAs, such
as those for prothrombin (MacGillivray and Davie, 1984), factor X (Fung et al., 1985), or factor XII (Cool et al., 1985), in which approximately 90% of the codons end in G or C. One codon is not used in the coding region of the ceruloplasmin mRNA (CGC for arginine) and others are used rarely (2 of 51 alanine residues are encoded by GCG).

The 3' untranslated region (spanning nucleotide residues 3199 - 3321) contains a putative polyadenylation signal ATTAAA (Proudfoot and Brownlee, 1976) that is located 14 nucleotides upstream of the poly(A) tail. This polyadenylation signal is observed in 12% of such 3' terminal sequences from vertebrates (Wickens and Stephenson, 1984), and is a variant of the more commonly observed signal AATAAA (Birnstiel et al., 1985).

Nucleotide residues 1 - 57 code for an amino terminal extension of 19 amino acids that is removed prior to the appearance of ceruloplasmin in plasma. A methionine residue occurs at residue -19 in this peptide sequence and may function as the initiator methionine. The leader peptide is rich in hydrophobic amino acids, and thus resembles a typical signal peptide (von Heijne, 1982; Watson, 1984). Such sequences function in the initiation of export of nascent polypeptide chains across the rough endoplasmic reticulum (Blobel et al., 1979). The ceruloplasmin cDNA sequence predicts that an Ala-Lys bond (encoded by nucleotides 55 - 60; see Figure 7) is cleaved during removal of this leader peptide. This is consistent with the specificity of signal peptidase cleavage (von Heijne, 1983) which typically occurs following basic residues that are preceded by small hydrophobic amino acids. This suggests that ceruloplasmin is synthesized in liver as a typical preprotein, containing a signal peptide sequence of at least 19 amino acids.
Mercer and Grimes (1986) have reported the characterization of a partial human ceruloplasmin cDNA clone that was isolated from a human liver cDNA library constructed in the phage vector λgt10. DNA sequence analysis showed that the clone extended from the amino terminal leader sequence to 114 amino acids short of the carboxy-terminus. The proposed signal peptide sequence obtained from DNA sequence analysis of λhCP-1 (see Figure 7) was found to agree completely with that reported by Mercer and Grimes (1986). However, the 38 bp of 5' untranslated sequence determined by analysis of both λhCP-2 and λhCP-6 cDNA clones (see Figure 8) differs completely from the 14 bp of 5' untranslated sequence determined by Mercer and Grimes (1986). The sequence divergence occurs immediately 5' to the proposed initiator methionine residue and does not appear to occur as the result of a frameshift mutation. The reason for this discrepancy is unclear at present, since the nucleotide sequences of λhCP-2 and λhCP-6 are in complete agreement with that determined from the analysis of the corresponding genomic region (see Figure 11). The possibility that the observed difference is the result of the utilization of alternative splicing patterns is unlikely, since the point of sequence divergence (i.e. immediately 5' to the initiator methionine) between the two cDNA sequences does not correspond to the location of an intron in the wild-type gene; an intron/exon junction was identified at nucleotide position -13 on the basis of comparative DNA sequence analysis and nuclease S1 mapping (see Section III.B.4.2). More probably, the sequence reported by Mercer and Grimes (1986) represents a cloning artifact.

The nucleotide sequence of two partial cDNA clones for human ceruloplasmin have been reported independently by Yang et al. (1986).
These clones (isolated from the Stuart Orkin human liver cDNA library) are nearly identical to the previously described 2.7 Kbp phCP-1 clone (see Section III.A). However, one of the clones analyzed by Yang et al. (1986) (designated CP-1) contains an insertion of four amino acid residues (occurring between residues 1041 and 1042 of the ceruloplasmin protein sequence) as well as an additional 121 bp (extending 3' to phCP-1) in the 3' untranslated region. The commonly used polyadenylation signal AATAAA is located 16 bp upstream from the poly(A) tract in the CP-1 clone. The origin of these different cDNA clones is uncertain at present, although these data clearly suggest some heterogeneity in ceruloplasmin transcripts with respect to the point of polyadenylation. Although the corresponding region of the wild-type ceruloplasmin gene has not yet been characterized (see Section III.B), several explanations for the observed amino acid insertion in the CP-1 clone described by Yang et al. (1986) can be postulated. Examination of the factor VIII genomic sequence (Gitschier et al., 1986) reveals the presence of an intron between exons 19 and 20 (i.e., corresponding to the ceruloplasmin cDNA sequence encoding amino acid residues 1041 – 1042, Vehar et al., 1984). Additional amino acids present in the CP-1 clone may have arisen from alternative splicing of an intron in some liver cells during RNA processing. Other plausible explanations could include the presence of two ceruloplasmin alleles in the donor's genome, or the existence of two ceruloplasmin loci. The latter assumption is somewhat doubtful, since a second wild-type gene was not detectable by either genomic Southern blot analysis (see Section III.D) or chromosome mapping (Yang et al., 1986; Royle et al., 1987).
A.2 Internal Homology Within the Ceruloplasmin cDNA Sequence

Extensive amino acid sequence identity has been reported previously within the three repeated A domains of ceruloplasmin (Takahashi et al., 1984). As expected, this homology extends to the nucleotide sequences of the repeated units when they are aligned. The three domains of ceruloplasmin exhibit approximately 46 - 51% identity when compared pairwise (Yang et al., 1986; Koschinsky et al., 1986). Within these three domains, specific regions show higher levels of sequence conservation (Yang et al., 1986). A comparison of the nucleotide sequence encoding amino acid residues 204 - 276 with that encoding residues 903 - 975 reveals 66.2% identity, while residues 204 - 276 compared with residues 565 - 637 shows 61.2% identity, and residues 565 - 637 aligned with residues 903 - 975 shows 59.4% nucleotide identity (Yang et al., 1986). Interestingly, a similar region from the derived amino acid sequence of a partial rat ceruloplasmin cDNA clone (Aldred et al., 1987) (see Section I.D) shows remarkable similarity with the published human amino acid sequence from residues 194 - 276 (75% identity), with almost complete conservation (98%) from residues 227 - 276, except for a conservative amino acid change at residue 243. This highly conserved region, present in each of the triplicated units in the human ceruloplasmin molecule may represent a region subjected to functional constraint. Interestingly, these regions are not coincident with proposed copper-binding centers in ceruloplasmin (see Section I.E). These highly conserved, structurally-related areas may thus be indicative of the evolutionary maintainence of an alternative ceruloplasmin function (Yang et al., 1986).
A.3 Analysis of the Human Ceruloplasmin Transcript

Northern blot analysis of the human ceruloplasmin transcript revealed the presence of two hybridizing species in human liver poly(A)$^+$ RNA, (3.7 ± 200 Kb and 4.5 ± 250 Kb), using the phCP-1 cDNA clone as a hybridization probe. In contrast, total RNA isolated from the human hepatoma cell line HepG2 (Knowles, 1980) (also probed with phCP-1) contained only the smaller mRNA species. Ceruloplasmin is synthesized and secreted by HepG2 cells (Knowles et al., 1980), suggesting that the 3.7 Kb RNA species encodes a functional ceruloplasmin transcript. The ceruloplasmin cDNA clones previously described in this study contain a total of 3359 bp of sequence in addition to a poly(A) tract which is usually 180 - 200 bp long in eukaryotic mRNAs (Perry, 1976). On this basis, both species appear to be larger than that required to encode the preceruloplasmin mRNA.

These transcripts may reflect the presence of long 5' or 3' untranslated segments which are differentially processed, resulting in the generation of two RNA species. The possibility of alternative splicing patterns occurring at the 3' end of the ceruloplasmin transcript has been suggested by Yang et al. (1986) to explain an insertion of 4 amino acids in the protein coding sequence (see Section IV.A.1). The extra sequence identified in the latter clone when compared to phCP-1 (including the additional 121 bp observed in the 3' untranslated segment) does not, however, account for the size difference (approximately 800 nucleotides) observed between the two hybridizing RNA species detected using Northern blot analysis. It is still possible that an as yet unidentified cDNA clone may contain additional sequence in the 3' untranslated segment,
thereby differing from previously characterized cDNA clones with respect to the point of polyadenylation. This proposal is feasible, since heterogeneity with respect to the position of polyadenylation in the phCP-1 clone compared to the cDNA clone described by Yang et al. (1986) has already been demonstrated. However, based on the results presented in this study, it is most likely that the two different ceruloplasmin mRNA species arise from differential processing and/or promoter function at the 5' end of the gene (see Section IV.B.5).

In contrast to the results described above using human liver poly(A)$^+$ RNA, it is interesting to note that a single ceruloplasmin transcript of 3.8 kb is detected using rat liver mRNA, and that comparable results have been obtained using RNA derived from rat testis, yolk sac, placenta, and choroid plexus (Aldred et al., 1987). Similarly, only one ceruloplasmin mRNA species is detected using bovine poly(A)$^+$ RNA. This reinforces the proposal that the 3.7 kb human ceruloplasmin RNA species is likely sufficient to encode a functional ceruloplasmin protein.

B. CHARACTERIZATION OF THE WILD-TYPE HUMAN CERULOPLASMIN GENE

B.1 Ceruloplasmin Gene Organization Corresponding to the Coding Sequence

Seven recombinant phage containing wild-type human ceruloplasmin genomic sequences were isolated from human genomic λ phage libraries. From restriction endonuclease mapping analysis coupled with Southern blot analysis, the ceruloplasmin gene was found to span at least 50 kbp in length. The location of 14 exons was determined (corresponding to nucleotides 1 - 2555 of the cDNA sequence); the 3' end of the gene (i.e.
the region corresponding to nucleotide residues 2556 - 3321 of the cDNA sequence) was not represented in the phage clones obtained from library screening. However, genomic Southern blot analysis allowed the identification of two genomic EcoRI fragments (4.6 kbp and 2.4 kbp) encompassing this region.

B.2 DNA Sequence Analysis of the Wild-Type Ceruloplasmin Gene

All splice donor/acceptor sequences conform to the GT..AG rule for nucleotides immediately flanking exon boundaries (Breathnach and Chambon, 1981; Cech, 1983). Further flanking sequences are in general accordance with compiled nucleotide frequencies (Breathnach and Chambon, 1981; Mount, 1982). The sequences of the exons were found to be identical to the previously determined cDNA sequence. The sizes of characterized exons range from 129 bp - 255 bp, with a calculated median size of 183 bp. This is consistent with published exon size distributions (Naora and Deacon, 1982). As found in other eukaryotic genes, intron sizes in the ceruloplasmin gene are highly variable (0.8 kbp - approximately 9.5 kbp). The total size of introns determined to date is approximately 30 kbp which is already in excess of that predicted by Naora and Deacon (1982), according to the observed variability of total intron size as a function of total exon length.

B.3 Intron Positions Within the Triplicated A Domains of Human Ceruloplasmin

As has been described previously (see Section I.A), the entire human ceruloplasmin molecule possesses an internal triplicated structure, which likely represents the product of successive gene duplication events (Dwulet and Putnam, 1981b; Doolittle, 1984; see Section IV.D). If
Figure 23. **Intron positions within the three repeated units of human ceruloplasmin.**

The top, middle and bottom lines correspond to the amino acid sequence of the A1, A2 and A3 triplicated regions respectively. Except for gaps introduced in order to maximize homology, the complete amino acid sequence is given continuously in single letter code. Residues boxed at a given position are identical; non-identical residues in the middle sequence are enclosed in dashed boxes. Triangles indicate determined intron positions. Carbohydrate attachment sites are designated by asterisks. Putative type I copper ligands are identified by solid circles. Homologous pairs of cysteine residues are shaded and vertical arrows identify four sites of trptic cleavage (modified from Ortel et al., 1984).
segmental DNA duplication has played a role in the evolution of the ceruloplasmin gene, one may expect the occurrence of intron boundaries between the A domains within the ceruloplasmin gene (Doolittle, 1985). The comparative alignment of intron boundaries within the repeated A units of ceruloplasmin (Figure 23) is facilitated by the relatively high degree of similarity between the related segments. As is clearly illustrated in Figure 21, the boundaries of the A1/A2 and A2/A3 repeats are each contained on one exon (exons 6 and 12, respectively), which is inconsistent with a DNA duplication mechanism where the boundaries of repeated units would predictably fall within introns. This is reminiscent of the factor VIII gene organization (see Figure 4, Section I.G.3) in which the A1/A2 and A2/A3 junctions are each contained on one exon. This observation is likely the result of intron loss within these two genes subsequent to their initial formation, such that the present-day gene organizations are no longer reflective of the original gene duplication events.

As is also the case for the factor VIII gene (see Figure 4, Section I.G), only some of the intron boundaries within the A repeats are conserved in the ceruloplasmin gene (see Figure 23). Furthermore, each of the three A domains within the factor VIII and ceruloplasmin genes contain a different number of exons (see Figure 24). The precise alignment of some intron boundaries within the repeated units of ceruloplasmin suggests that these introns may have arisen following the successive gene duplication events, either by loss of one member of a duplicated intron, or by insertion of new introns. These latter two events are essentially indistinguishable.
Figure 24. Positions of introns (↑) in the A domains of ceruloplasmin and factor VIII.

The numbers identifying the extent of the triplicated A units represent nucleotide residues. Codon phases (i.e., I, II, or 0 indicating the position of each intron relative to the codon triplet) are shown. Ceruloplasmin introns are labelled A - N, while introns in factor VIII are designated a - r. Dashed lines designate areas in which introns have not yet been localized in the ceruloplasmin gene. The relative position of the B domain in factor VIII is indicated.
In an analogous situation, α-fetoprotein (Eiferman et al., 1981) and albumin (Sargent et al., 1981) each have a triplicated gene structure (i.e. composed of 3 similar sets of 4 exons). On this basis, it has been proposed that these genes have arisen through the duplication of an ancestral gene approximately 300 - 500 million years ago (Eiferman et al., 1981). Unlike ceruloplasmin or factor VIII, all intron boundaries within the repeated units in α-fetoprotein and albumin are almost precisely conserved. Since it has been observed that in a number of genes, introns separate protein domains (Blake, 1983a, b; Gilbert, 1978; Go, 1981, 1983), the occurrence of introns in α-fetoprotein and albumin at the borders of the three genetic domains may reflect divisions between functional units. Indeed, although active binding sites in mammalian α-fetoproteins are poorly characterized, specific functions have been assigned to the 3 genetic units in albumin (Peters and Reid, 1977) (Domain 1: long-chain fatty acid binding, Domain 2: bilirubin binding, and Domain 3: indole binding). However, for both the ceruloplasmin and factor VIII genes, correlations between exons and functional domains are unclear at present. Thus, the conservation of some intron boundaries within the repeated units of these latter two genes cannot be correlated with the incidence of functional units.

B.4 Comparison of the Gene Organizations of Ceruloplasmin and Factor VIII

Comparative analysis of the gene organizations of ceruloplasmin and blood coagulation factor VIII (Gitschier et al., 1984) is useful in tracing the evolutionary history of these two proteins. As can be clearly seen, introns A - L in the ceruloplasmin gene correspond closely in
position to introns present in the factor VIII gene, and except for intron H, interrupt the reading frame in the same phase (Figure 24). When the protein sequences of ceruloplasmin and factor VIII are aligned to maximize identity, some of these introns interrupt either the same amino acid in the two proteins (e.g., ceruloplasmin introns B, I, J, and K) or amino acids in identical positions (e.g., ceruloplasmin introns C and E) (see Figure 25). This strongly suggests that these introns were present in the progenitor gene, prior to duplication. In two cases (introns A and L in the ceruloplasmin gene), intron positions vary in the factor VIII and ceruloplasmin genes by one amino acid (3 nucleotide residues). These introns also may have been present in the ancestral gene (i.e. prior to duplication), and differ slightly in position in the present-day genes due to insertion or deletion of single amino acids within the coding sequences of either factor VIII or ceruloplasmin. It is also possible that these small variations may be the result of intron sliding (see below), as has been postulated to explain similar differences in the organizations of the triphosphate isomerase genes from *Aspergillus nidulans* and maize (McKnight et al., 1986).

Introns D, F, and G in the ceruloplasmin gene differ in position by 10, 7, and 12 amino acids respectively in the factor VIII gene (see Figure 25), although the phases of these three introns in the two genes are the same (see Figure 24). For introns D and G, a model involving intron sliding seems unlikely, since no corresponding insertions or deletions are observed at these positions in the two genes (Figure 25). This suggests that independent intron insertion has occurred at these locations in the ceruloplasmin and factor VIII genes and these introns
Figure 25. **Comparative positions of introns in ceruloplasmin (A) with corresponding introns in factor VIII (V).**

The factor VIII amino acid sequence is aligned with that of ceruloplasmin in the region of the A domain. The consensus line designates residues which are identical in the two proteins. The numbering above the sequences corresponds to that of factor VIII. The numbers preceding the ceruloplasmin lines represents the number of the first amino acid in the line. Introns are labelled as described in Figure 24.
have coincidentally interrupted amino acids in the same codon phase.
Alternatively, introns may have originally been present in identical
positions in factor VIII and ceruloplasmin, and may have been subsequently
lost. This model, however, dictates the presence of very small exons (10
and 7 bp, respectively) in the ancestral gene, thereby making the former
theory more attractive.

The difference in the position of intron h in the factor VIII gene
corresponding to intron F in ceruloplasmin may be the result of intron
sliding. Amino acid sequence comparison in the region of these intron
junctions in factor VIII and ceruloplasmin shows a low degree of
similarity, with a deletion observed in the factor VIII protein sequence
relative to that of ceruloplasmin (Figure 25). Since insertions or
deletions are a consequence of the intron sliding mechanism, it is
possible that junctional sliding (due to the formation and utilization of
alternative splice donor/acceptor sites within exons or introns) may
account for this observed variation in intron position, while maintaining
the coding phase of the intron.

The remainder of the introns identified in the ceruloplasmin gene to
date (i.e., introns M and N, contained with the A3 repeated unit; see
Figure 24) do not seem to correlate with the positions of introns in the
factor VIII gene. It is thus likely that introns present in this region
of the two genes have resulted from independent intron insertion or
deletion events.

Figure 24 clearly shows that in several corresponding regions, the
factor VIII gene possesses more introns than does the ceruloplasmin gene.
As has been stated previously, however, it is not possible to distinguish
whether this reflects intron insertion in the factor VIII gene, or corresponding intron loss in the ceruloplasmin gene.

B.5 Characterization of the 5' End of the Human Ceruloplasmin Gene

The interruption of the 5' untranslated region by at least four introns renders the organization of the 5' end of the human ceruloplasmin gene somewhat unusual. Although not a common occurrence, the presence of one intron in the 5' untranslated sequence of a number of genes from diverse organisms has been reported [e.g., the human and chicken insulin gene (Perler et al., 1980; Bell et al., 1980), the heat shock protein 83 gene from Drosophila (Hackett and Lis, 1983), and the nuclear COX4 gene of yeast (Schneider and Guarente, 1987)]. However, it has been reported (Irminger et al., 1987) that the single-copy, human insulin-like growth factor II (IGFII) gene possesses at least 3 exons (designated 1, 2, and 3, in a 5' to 3' direction) in the 5' untranslated region. These exons encode alternative 5' untranslated regions in a tissue-specific manner resulting in IGF-II mRNA species of variable lengths. Brain, placenta, and adrenal gland contain a 6.0 kbp IGF-II transcript, utilizing exon 3 in the 5' untranslated region, while liver contains a 5.3 kbp IGF-II transcript, with the 5' untranslated segment derived from exons 1 and 2. The novel organization of these transcripts was first detected when it was found that the 5' untranslated regions of characterized cDNA clones differed from the 5' sequence reported in the genomic clone (Dull et al., 1984). It has not been established whether the heterogenous transcripts are due to the existence of two tissue-specific promoters (i.e. one upstream of exon 1 and one upstream of exon 3) as is the case for the mouse α-amylase gene (Young et al., 1981; Schibler et al., 1983) (see
below), or whether different transcripts arise from alternative splicing patterns in various tissues. The 5' extent of the IGF-II gene has not yet been determined, thereby rendering the possibility of additional exons in the 5' untranslated segment.

The α-amylase transcripts accumulate in two different tissues of the mouse - the salivary gland and the liver. The nucleotide sequences of the two mRNA species are identical, with the exception of the 5' non-translated regions. The 5' terminal 158 nucleotides of the major liver species are separated from the first exon in the coding sequence by a 4.5 Kbp intron. This leader sequence is unrelated to the 5' terminal 47 nucleotides present in the salivary gland counterpart and is separated from the first exon in the coding sequence by an additional intron of 2.6 Kbp. The two distinct mRNA species arise due to the presence of a specific promoter for each transcript (Schibler et al., 1983); the relative activity of these two promoters is tissue-specific. Interestingly, the use of alternative promoters has also been described for the generation of two transcripts encoding intracellular and secreted forms of yeast invertase (Carlson and Botstein, 1982).

It is interesting to speculate as to the biological significance of the heterogenous 5' non-translated leaders observed in the IGF-II and α-amylase genes. Perhaps variable transcripts result in different levels of translational efficiencies, thereby selectively regulating protein concentrations in different tissues. It has been reported that RNA species containing greater than two hundred residues in the 5' non-coding region are translated less efficiently than are shorter leader sequences by eukaryotic ribosomes in vitro (Young et al., 1981). Thus,
the length of the 5' non-translated segment may regulate the efficiency of translation or perhaps affect mRNA stability. The significance of multiple exons in the 5' untranslated region of the ceruloplasmin gene is unclear at present since both the 3.7 Kb and 4.5 Kb mRNA species are detectable with 5' end probes tested to date. This would not be expected if alternative splicing patterns give rise to these two transcripts. Utilization of two different promoters (as has been described for the mouse α-amylase gene and the yeast invertase gene) cannot be ruled out at present. It is possible, however, that this unusual organization may be implicated in the regulation of tissue-specific expression of the ceruloplasmin gene, as described above for the α-amylase and IGF-II genes. Unfortunately, data have not yet been obtained regarding the synthesis of ceruloplasmin in extrahepatic tissues. Alternatively, it is possible that the novel organization of the ceruloplasmin gene may be required for the regulation of ceruloplasmin levels by exogenous copper, or during the acute phase inflammatory response (see Section I.B.4). With reference to the latter proposal, the complete gene organizations of several acute phase reactants have been determined [e.g. transferrin (Lucero et al., 1986) and fibrinogen (Crabtree et al., 1985)]. On the basis of DNA sequence determinations, these latter genes do not possess unusual organizations in their respective 5' ends. Further analysis of the 5' end of the ceruloplasmin mRNA using primer extension, and nuclease S1 mapping techniques to estimate exon sizes should allow determination of the size of the 5' untranslated leader, and identification of the transcriptional start site. These results, coupled with analysis of extrahepatic human ceruloplasmin biosynthesis and complete
characterization of the 5' end of the gene may resolve the existence of the two ceruloplasmin mRNA species detectable with Northern blot analysis. Thus far, DNA sequence analysis of the 5' untranslated region has not revealed the presence of any typical eukaryotic promoter elements, strongly suggesting that the 5' most end of the ceruloplasmin transcript has not yet been identified.

C. CHARACTERIZATION OF A PSEUDOGENE FOR HUMAN CERULOPLASMIN

C.1 DNA Sequence Analysis of the Human Ceruloplasmin Pseudogene

Using previously characterized cDNA clones for human ceruloplasmin as hybridization probes for human genomic libraries, four overlapping recombinant phage clones encoding approximately 21 kbp of contiguous genomic DNA were obtained. Within this region, a pseudogene for human ceruloplasmin was identified (nearly 2 Kbp in length), corresponding to nucleotide residues 1502 - 3198 of the human ceruloplasmin cDNA sequence. Additionally, the pseudogene extends through the 123 bp of the 3' untranslated sequence that is present in the phCP-1 cDNA clone, and continues for a further 40 bp, the sequence of which corresponds to the 3' untranslated sequence of a ceruloplasmin cDNA clone described by Yang et al. (1986). On this basis, it appears that the pseudogene has been derived from an mRNA species corresponding to the cDNA clone described by Yang et al. (1986), as opposed to the shorter mRNA species represented by phCP-1. The derivation of the ceruloplasmin pseudogene from this particular mRNA species is interesting, since it has been shown that for different rat cytochrome c mRNA species occurring at the same intracellular concentration, multiple pseudogenes have arisen
preferentially from one mRNA (Scarpulla, 1984). It has been postulated that this may be due to local secondary structure in the 3' end of the mRNA, which facilitates binding of enzymes involved in reverse transcription or subsequent integration into the genome (Scarpulla, 1984).

As is characteristic of processed pseudogenes, complete DNA sequence analysis revealed that the human ceruloplasmin pseudogene lacks introns present in the wild-type gene. In contrast to numerous other examples of processed pseudogenes in which the intervening sequences are precisely removed (e.g., Vanin et al., 1980), the ceruloplasmin pseudogene contains a 213 bp deletion (corresponding to nucleotides 1865 - 2077 of the cDNA sequence) that occurs exactly at the boundaries of exon 11 in the wild-type gene. There is also a small 17 bp deletion in the pseudogene sequence, beginning at nucleotide 2943 of the cDNA sequence. It has not been determined whether this deletion also corresponds to the location of intron/exon boundaries in the wild-type gene. It is unclear if the 213 bp deletion observed in the pseudogene occurred at the time of intron processing, or whether the deletion is the result of a subsequent mutation event.

The 5' boundary of the pseudogene is characterized by a short sequence that is homologous to the 3' end of intron H in the wild-type gene. The presence of this intronic segment is expected, since the pseudogene diverges from the wild-type gene prior to the 5' end of this intron. Therefore, the appropriate 5' slicer donor site (Breathnach and Chambon, 1981; Cech, 1983), required for intron removal is absent.

Although the pseudogene appears to have been derived from a processed RNA species (i.e. lacking intervening sequences), there is no poly(A)
tract present in the sequence. While the majority of processed pseudogenes have a poly(A) tail, several exceptions have been reported (Vanin et al., 1980; Notake et al., 1983). It is possible that the absence of a poly(A) tract in the human ceruloplasmin pseudogene sequence may be the result of the mechanism of its formation, possibly involving base pairing between the poly(A) tail of the mRNA and U-rich region in the 3' untranslated region. In this case, the six T residues observed following the 3' untranslated segment corresponding to that described by Yang et al. (1986) may represent a site of mRNA self-priming. This explanation may also account for the divergence that occurs at this point between the pseudogene sequence and the remainder of the 3' untranslated region reported by Yang et al. (1986).

The presence of a repeated CT dinucleotide segment at the 3' end of the ceruloplasmin pseudogene sequence corresponding to the coding strand is interesting. A 116 bp segment, composed mainly of repeated GA dinucleotides corresponding to the coding strand has been reported at the 3' end of the mouse corticotropin B-lipotropin precursor pseudogene (Notake et al., 1983). In the latter case, this repeated segment occurs immediately following the point at which the pseudogene diverges from the wild-type gene. In the human ceruloplasmin pseudogene, the repeated TC region occurs 172 bp 3' to the point at which the pseudogene diverges from the wild-type gene. The rat metallothionein pseudogene 14b, which has been characterized by Andersen et al. (1986), contains a 42 bp poly(CA) tract, located approximately 300 bp 3' to the site of polyadenylation. While stretches of repeating CA residues have been found in eukaryotic DNA at the site of recombination events such as gene conversion (Shen et al.,
1981), and are thought to induce Z-DNA conformational changes (Nordheim and Rich, 1983), the function of significantly long purine or pyrimidine stretches remains unclear. Tracts of $d(GA)_n • d(TC)_n$ have been found at many sites in eukaryotic genomes [e.g. human U1 RNA genes (Htun et al., 1984) and the murine immunoglobulin μ-σ heavy chain gene (Richards et al., 1983)]. Interestingly, a $d(GA)_{27} • d(TC)_{27}$ tract has been demonstrated in a polyomavirus-transformed cell line near the end of a host DNA segment that is responsible for arrest of the viral replication process (Baran et al., 1987). In the latter case, it has been postulated that this repeated sequence, in conjunction with an inverted repeat, may serve as an arrest site for chromatin replication in vivo. In pseudogenes, it is attractive to speculate that repeated pyrimidine or purine stretches may be involved in the process of pseudogene integration into the genome, since very little is known about the mechanism(s) mediating this event (Vanin, 1985).

C.2 Chromosomal Location of the Human Ceruloplasmin Pseudogene

All processed pseudogenes studied to date are located on different chromosomes than their functional counterparts. This is in contrast to non-processed pseudogenes, which have likely arisen from gene duplication events and are therefore on the same chromosome as the respective wild-type gene (Vanin, 1985). Using previously characterized human-hamster hybrid cell lines (Donald et al., 1983), the human ceruloplasmin pseudogene has been assigned to chromosome 8. This result has been recently verified, using the technique of in situ hybridization (Wang et al., 1987). This differs from the unequivocal assignment of the functional ceruloplasmin locus to chromosome 3q25 (Yang et al., 1986;
Royle et al., 1987). It has been reported previously by Yang et al. (1986) that a 0.8 kbp EcoRI fragment that can be identified in genomic Southern blots probed with the ceruloplasmin cDNA segregates with human chromosome 11. However, the present study suggests that this 0.8 kbp EcoRI fragment is part of the human ceruloplasmin pseudogene, and maps to chromosome 8 (see Section III.C). The reason for this discrepancy is unclear at present, since genomic Southern blot analysis indicates that there is only one pseudogene for human ceruloplasmin. However, since the mapping analysis reported by Yang et al. (1986) was performed using the ceruloplasmin cDNA to probe human-mouse hybrid cell lines, the 0.8 kbp band detected may represent a cross-reacting species in the mouse genome. This is in agreement with previous difficulties encountered in chromosome mapping when using cDNA fragments as hybridization probes (J. Hamerton, personal communication).

C.3 Speculations on the Evolutionary Origin of the Human Ceruloplasmin Pseudogene

The human ceruloplasmin pseudogene shares approximately 97% nucleotide sequence identity compared to the wild-type ceruloplasmin coding sequence, suggesting that it has been formed relatively recently in evolutionary time. This is characteristic of processed pseudogenes analyzed to date [e.g. Li et al., 1981; Freytag et al., 1984], all of which have arisen following mammalian radiation (approximately 80 million years ago) (Vanin, 1985)]. Interestingly, the existence of a processed pseudogene for rat ceruloplasmin has been suggested recently by Shvartsman et al. (1985) based on preliminary restriction endonuclease mapping analysis of rat ceruloplasmin genomic clones. This suggests that the
formation of the ceruloplasmin pseudogene occurred prior to the divergence of rats and humans. This is consistent with the notion that processed pseudogenes have arisen in genomes following the widespread appearance of mammals on earth.

Presumably pseudogenes are not subject to functional constraints. This has been interpreted to suggest that the pattern of nucleotide substitutions in pseudogenes should reflect patterns of intrinsic mutation (Li, 1983). Thus, transversions (i.e. purine-pyrimidine or pyrimidine-purine base substitutions) should occur twice as frequently as transitions (i.e. purine-purine or pyrimidine-pyrimidine base changes), if mutations occur randomly (Li, 1983). Compared with the nucleotide substitution pattern at the first and second positions of codons in functional genes, it has been shown for pseudogenes that the relative frequency of transitions is much greater than that of transversions (Gojobori, 1982). This suggests that a sequence under no functional constraint will become A - T rich, due to the spontaneous deamination of cytosine and 5-methylcytosine. Of the 44 nucleotide substitutions observed in the human ceruloplasmin pseudogene sequence, 34 of these (77%) are transition mutations.

Genomic Southern blot analysis indicates that there are no human ceruloplasmin pseudogene sequences corresponding to the 5' end of the wild-type gene. This suggests that the pseudogene may have arisen from an aberrant transcript, as the result of initiation within the gene. Such a model has been proposed for the mouse corticotropin ß-lipotropin pseudogene (Notake et al., 1983). This pseudogene is similar to the human ceruloplasmin pseudogene, in that it is only a partial copy of the
functional gene, encoding the carboxy-terminal 143 amino acid residues and the 3' untranslated region (Notake et al., 1983). In the case of these two pseudogenes, RNA polymerase III initiation could have occurred within the respective genes. In vitro expression studies suggest that aberrant initiation of transcription of the human corticotropin B-lipotropin precursor can occur at a site within the gene corresponding to the 5' end of the homologous pseudogene sequence (Mishina et al., 1982). Additionally, it has been shown that RNA polymerase III activity can result in transcripts initiating upstream of the human B globin gene, which is normally transcribed by RNA polymerase II (Carlson and Ross, 1983). Furthermore, these latter aberrant transcripts have been shown to be both polyadenylated and spliced (Carlson and Ross, 1983). Therefore, the possible generation of the human ceruloplasmin pseudogene from an aberrant RNA polymerase III event does not account for incorrect splicing or the observed lack of polyadenylation.

D. A MODEL FOR THE EVOLUTION OF CERULOPLASMIN, FACTOR V AND FACTOR VIII

Based on its internal 3-fold repeated structure (Section I.B.1) and comparison to known copper-binding centers in several blue copper proteins (Section I.E), Dwulet and Putnam (1981b) have proposed a model for the evolution of the ceruloplasmin gene. This model suggests that ceruloplasmin has evolved from a small primordial copper-binding protein (approximately 350 amino acids in length). Tandem duplications of this ancestral gene could then have given rise to the present-day gene for ceruloplasmin (see Figure 26). Because the pairwise comparison of the three repeated units in the human ceruloplasmin molecule shows very
Figure 26. **A proposed model for the evolution of ceruloplasmin (CP), factor V (FV) and factor VIII (FVIII)** (see text for details).

The triplicated A domain is shown by cross-hatched bars while the B and C domains (present in factors V and VIII) are identified by open and stippled bars, respectively. The ceruloplasmin pseudogene is represented by a solid bar. **MY** designates million years.
ANCESTRAL GENE

TRIPLEXICATION (400 MY)

DUPLICATION

FUSION

B

C

DUPLICATION

PRESENT CP FV FVIII

FORMATION (<80 MY)
similar values (see Table 1), the order of the triplication cannot be determined; two consecutive elongations likely occurred close together on an evolutionary time scale (Doolittle, 1984). Without comparative amino acid sequence data and/or molecular weights for ceruloplasmin from more primitive species, the time of the first duplication event is difficult to estimate. Interestingly, the degree of similarity between the three repeated units in ceruloplasmin is approximately the same as that observed for the duplicated halves of transferrin (Doolittle, 1984). Thus, if the rate of change of ceruloplasmin was similar to that of transferrin, both sets of duplications may have occurred within the same time frame (Doolittle, 1984). Since lampreys have a full-sized transferrin, the latter event has been placed at > 400 million years ago (Doolittle, 1984). On this basis, it would seem reasonable that ceruloplasmin has grown by duplication at the same time. This is in agreement with the proposal of Dwulet and Putnam (1981b) who suggested that the triplication event resulting in the formation of ceruloplasmin was coincident with the appearance of vertebrate animals (i.e. 500 million years ago), possessing a closed vascular system and a urogenital system. The latter anatomical developments necessitated that plasma proteins have molecular weights of > 60 Kda to avoid renal excretion.

As has been previously discussed, factor VIII was shown to be structurally related to human ceruloplasmin on the basis of amino acid similarity (Wood et al., 1984; Vehar et al., 1984). Although initially indicated on the basis of limited amino acid sequence derived from bovine factor V (Church et al., 1984), the complete amino acid sequence of human factor V (Jenny et al., 1987) has shown clearly that this protein is
structurally related to both factor VIII and ceruloplasmin. It has therefore been proposed that ceruloplasmin, factor V and factor VIII form a gene family derived from a common ancestral gene (see Figure 26). In this scheme, the entire triplicated gene for human ceruloplasmin would have undergone duplication, thereby forming the progenitor species for factors V and VIII. As is depicted schematically in Figure 26, this event likely occurred relatively close in evolutionary time to the formation of the initial triplicated structure, since when compared pairwise, the A repeated units in ceruloplasmin, factor V and factor VIII share very similar levels of amino acid identity (Table 1), implying that they have been evolving for a similar length of time. Prior to the third gene duplication event leading to the formation of factors V and VIII, the B and C domains were likely inserted by independent gene fusion events (Figure 26), thereby creating the larger and more complex factor V and VIII molecules. The evolution of ceruloplasmin and factors V and VIII provides an example of divergent evolution in which factors V and VIII have adopted diverse functions compared to ceruloplasmin.

E. CONCLUDING REMARKS

Comparative analyses of the gene organizations of structurally related proteins contribute greatly to our knowledge of gene evolution, aiding in the unambiguous classification of genes into various multigene families. For example, although the sequence of angiotensinogen is only 20% identical to that of α-antitrypsin, the distribution of introns in these two genes is precisely conserved (Tanaka et al., 1984) thus allowing their unequivocal assignment to the same gene family (Doolittle, 1985).
Examination of the gene organization of ceruloplasmin with that previously reported for factor VIII (Gitschier et al., 1984) confirms that these genes are members of the same gene family, and as such have likely arisen from a common ancestral gene. Characterization of the factor V gene (which as also been proposed as a member of the latter multigene family, based on shared structural properties) will prove interesting since one might predict that it will exhibit a more similar organization to factor VIII than does ceruloplasmin, based on similar functional constraints shared by the two blood coagulation factors. As has been discussed previously, development of a more complete understanding of the evolutionary history of ceruloplasmin itself will require comparative phylogenetic data, involving the isolation and characterization of ceruloplasmin from primitive vertebrates such as hagfish or lampreys.

Just as ceruloplasmin has been rendered a fascinating protein for chemical study, the molecular biology of this multicopper oxidase is proving to be equally intriguing. As has been shown in the various aspects of this study, the ceruloplasmin gene is highly complex in nature, possessing an unusual organization at the 5' end with multiple introns interrupting the 5' untranslated leader segment, and heterogeneity with respect to the site of polyadenylation at the 3' end. Therefore, complete knowledge of the organization of this gene will enhance our understanding of gene regulation in eukaryotic systems. Finally, complete characterization of the human preceruloplasmin cDNA sequence as detailed in this study should facilitate studies involving site-directed mutagenesis of various regions of the ceruloplasmin molecule, in order to identify specific copper-binding centers within the protein.
V. REFERENCES


Blobel, G., Walter, P., Chang, C.N., Goldman, B.M., Erickson, A.H., and
   Lingappa, V.R. (1979) in Secretory Mechanisms, Vol. 33 (Hopkins,
   9-36.
Bonta, I.L. (1978) in The Handbook of Experimental Pharmacology (Born,
   G.V. et al., Eds.), Springer-Verlag, New York, pp. 523-567.
   Acta 678, 27-38.
Chen, M.-J., Shimada, T., Moulton, A.D., Harrison, M., and Nienhuis, A.W.
   Biochemistry 18, 5294-5299.
Church, W.R., Jernigan, R.L., Toole, J., Hewick, R.M., Knopf, J., Knutson,
Colman, P.M., Freeman, H.C., Guss, J.M., Murata, M., Norris, V.A.,
   319-324.
Cool, D.E., Edgell, C.-J.S., Louie, G.V., Zoller, M.J., Brayer, G.D., and


