Characterization of the gene encoding a 40 kDa major antigen of the sealworm (*Pseudoterranova decipiens*).

by

Brian Dixon

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
February, 1994

ISBN 0-315-93771-8

Canada

Name Brian Dixon

Dissertation Abstracts International is arranged by broad, general subject categories Please select the one subject which most nearly describes the content of your dissertation Enter the corresponding four-digit code in the spaces provided

Molecular Biology
SUBJECT TERM

0 3 0 7
SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS
Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION
General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS
Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY
Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES
American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women s Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES
Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES
Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES
Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Econom cs 0386

PHYSICAL SCIENCES
Pure Sciences
Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences
Applied Mechan cs 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY
General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

STATEMENT OF RETENTION OF COPYRIGHT

Some of the research described in this thesis has been published in the scientific literature. I include references for these publications below:

The research described in Chapter 2 is also described in:

Dixon, B., Walker, B., Kimmins, W. and Pohajdak, B. 1991. Isolation and sequencing of an unusual haemoglobin from the parasitic nematode *Pseudoterranova decipiens*. *Proc. Nat. Acad. Sci. USA* **88**, 5655-5659. I retain copyright to this material.

The research described in Chapter 3 is also described in:

Dixon, B., Kimmins, W. and Pohajdak, B. 1993. Variation of colour in *Pseudoterranova decipiens* (Nematoda; Anisakidae) larvae correlates with haemoglobin concentration in the pseudocoelomic fluid. *Can. J. Fisheries Aquat. Sci.* **50**, 767-771. I retain copyright to this material.

The research described in Chapter 4 is also described in:

Dixon, B., Walker, B., Kimmins, W. and Pohajdak, B. 1992. A nematode haemoglobin gene contains an intron previously thought to be unique to plants. *J. Mol. Evol.* **35**, 131-136. I retain copyright to this material.

Some of the ideas described in Chapter 6 are also discussed in:

Dixon, B. and Pohajdak, B. Did the ancestral globin gene of plants and animals contain only two introns? *Trends in Biochem. Sci.* **17**, 486-488 (1992). Copyright to this material is held by Elsevier Science Publishers Ltd., with exception of its use in this thesis.

ABSTRACT

Larvae of the parasitic nematode *Pseudoterranova decipiens* (commonly known as sealworm) infect several species of demersal fish in the North Atlantic, including Atlantic cod (*Gadus morhua*). The most important definitive hosts of the parasite appear to be grey seals (*Halichoerus grypus*). Immunological control of the *P. decipiens* population might be achieved by immunizing seals against the major antigens of the worm. Polyclonal sera raised in adult seals against *P. decipiens* indicated that there were two major antigens, of 40 kDa and 105 kDa. A cDNA clone of the 40 kDa antigen, encoding a unique 333 amino acid haemoglobin was isolated from the nematode. The predicted protein sequence contains an 18 amino acid hydrophobic signal sequence and has a calculated mass of 37.6 kDa in the mature form. The predicted protein sequence also reveals an internal duplication of a 154 amino acid domain (51% identity). Both domains have significant sequence homology to other haemoglobins, in agreement with a duplication event. The nematode globin gene has a unique six-intron, seven-exon structure that contains an intron in a position previously thought to be unique to plants. The nematode globin also contains a unique intron, absent in other extracellular invertebrate globin genes, between its secretory peptide leader sequence and its coding sequence. The variation in colour of the larval nematodes from reddish-brown to white is due to variations in the concentration of haemoglobin in the pseudocoelomic fluid. While reddish brown worms contain more haemoglobin, this variation is not due to greater quantities of globin mRNA. Thus the level of haemoglobin may be controlled at the post-transcriptional level. The full length cDNA clone and fragments encoding the first and second repeats of this protein were expressed in bacteria to produce sufficient quantities of protein for vaccines.

LIST OF ABBREVIATIONS

bp	basepairs
cDNA	complementary DNA
DNA	deoxyribonucleic acid
F	F statistic
FMN	flavin mononucleotide
kb	kilobases
kDa	kiloDalton
mg	milligrams
min	minutes
ml	millilitre
mRNA	messenger ribonucleic acid
nm	nanometres
P	significance level of statistical test
PAGE	polyacrylamide gel electrophoresis
poly(A)$^+$	polyadenylate
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
R1	repeat 1 of haemoglobin sequence
R2	repeat 2 of haemoglobin sequence
SDS	sodium dodecyl sulphate
V	volts
Z	significance value of similarity test
μg	micrograms

ACKNOWLEDGEMENTS

There are many people I would like to thank for aiding me in producing this thesis. First and foremost, Dr Bill Pohajdak, who has been a constant source of support and inspiration over the past three years. Thanks also to all of the people in the Sealworm Intervention Project who made this thesis possible. Greg Stuart's technical assistance was much appreciated, as was the aid of the numerous summer students and honours students who helped with this research. I'd like to thank Dr Dave Hoskin for his guidance, not only concerning this thesis, but also regarding life in the scientific community. I owe Dr Mike Gray thanks for being on my supervisory committee.

I'd also like to thank all my friends, who have been very supportive over the last few years. Special thanks are due to Anna Metaxas and Heather Hunt for taking me to the beach with them. I'd also like to thank Emily for her understanding and patience while I was putting the finishing touches on this work.

Finally, I thank my family, particularly my parents for their love and support during the many years I've spent in various universities. Cath and George: thanks for everything!

CHAPTER 1: GENERAL INTRODUCTION

Pseudoterranova decipiens is an economically important parasite of the Atlantic Cod (*Gadus morhua*) (Templeman et al., 1957, Chandra and Khan, 1988, Brattey et al., 1990, McClelland et al., 1990), but its definitive host is the grey seal (*Halichoerus grypus*) (Mansfield and Beck, 1977, Malouf, 1986). This nematode has cost the Atlantic fishing industry millions of dollars over the past several years in removal costs and unsalable product: about $30 million in 1984 alone (Malouf, 1986).

Immunity to several nematode parasites has been induced by vaccination with surface or excretory/secretory antigens, for example *Dirofilaria immitis* (Culpepper et al., 1992), *Ascaris suum* (Stromberg and Soulsby, 1977) and *Trichostrongylus colubriformis* (Connan, 1965, Rothwell and Love, 1974, Urban and Romanowski, 1985, O'Donnell et al., 1989 and Frenkel et al., 1992). The purpose of the project described here was original to characterize the nematode antigen which appears to be approximately 40 kDa, recognized by polyclonal seal sera on Western blots, and to produce enough of this protein to use as a vaccine. Some very interesting theoretical observations arose from this project, however. Thus while the development of the vaccine was still a goal of this project, it has taken a smaller role than originally anticipated. The introduction that follows provides the background against which

these interesting observations will be discussed, but each chapter contains a relevant introduction to its particular subject matter.

The Discovery of Introns

In the 1970's, with the development of restriction enzymes, blotting techniques and protein and nucleic acid sequencing methods, molecular biology radically altered the perceived view of the molecular mechanisms occurring within cells. Perhaps the best example of this is the startling discoveries that led to the elucidation of gene structure. The first genes for which nucleotide sequences were determined were bacterial in origin, due to the convenience of the small genome sizes involved. In agreement with all expectations, these genes corresponded exactly in size and structure to the mRNAs they produced. Genes produced mRNAs which produced protein in a neat, efficient manner, following the "one-gene, one-polypeptide" hypothesis.

In 1977, however, restriction enzymes made it possible to study the genes in the complex genomes of eukaryotes. The first experiments involving the hybridization of restriction fragments to sequences derived from mRNA were performed and to everyone's surprise eukaryotic genes were found to be much larger than the mRNAs they encoded! Even more startling was the discovery that this size difference was caused by stretches of DNA that did not encode protein but interrupted stretches of DNA that did (Breathnach et al., 1977, Doel et al., 1977, Brack and Tonegawa,

1977, Jeffreys and Flavell, 1977). The function and origin of these "intervening sequences" were the most intriguing questions of the day. Gilbert (1978) dubbed the intervening sequences "introns" and the coding regions they separated "exons."

Origins: Late Versus Early

The evolutionary origin of introns is also a puzzling question since they are absent from prokaryotes and lower eukaryotes, but present in the higher eukaryotes. Since any argument regarding intron evolution is closely tied to the question of their function, I will outline the main theories regarding their origin, then discuss the support for these theories found in the work on intron function.

Late Introns

Gilbert's (1978) original hypothesis was that introns were introduced into eukaryotes following their split from the prokaryotes, providing an evolutionary advantage by allowing the shuffling of exons. Several other scientists (particularly Cavalier-Smith, 1978 and Crick, 1979, but including Borst and Grivell, 1981, and Koch et al., 1981,) ascribed to this point of view, which became known as "Introns-Late". This theory suggests that the ancestral genes were continuous, like modern prokaryotic genes, but following the eukaryotic/prokaryotic split, eukaryotic genes acquired introns. Why they should do so is discussed in the next section, but how did eukaryotic genes gain introns?

Most proponents of this theory suggest that introns developed from transposons, stretches of DNA that encode polymerases and nucleases that enable them to copy themselves and insert themselves elsewhere in the genome. Cavalier-Smith (1985) suggests that this hypothesis explains the origin of introns and the RNA splicing mechanism. Introns are defective transposons and the RNA splicing enzymes that remove them are derived from the DNA splicing enzymes that were specific for transposon termini. Introns that encode nucleases have been reported for genes from fungal mitochondria (Cavalier-Smith, 1991), chloroplasts (Rochaix et al., 1985) and the phage T4 thymidylate synthetase gene (West et al., 1989). The alteration of the nuclease activity from DNA to RNA would allow the cell to remove any other transposons that infected its genome, thus assuring that the RNA splicing mechanism would be maintained.

The discovery of introns that were capable of splicing themselves from an RNA molecule with no enzymatic aid presented a problem for this theory (Cech et al., 1981, Kruger et al., 1982). However Cavalier-Smith (1991) constructed a phylogeny that accounts for these introns. Self-splicing introns arose very early in evolution (about 3.5 billion years ago), but only in tRNA genes, as evidenced by their presence in eubacterial tRNA genes (Kuhsel et al., 1990, Xu et al., 1990). A radical change in the splicing mechanism occurred about 1.7 billion years ago when proteins, encoded by the introns and capable of splicing out introns, arose in

archaebacterial rRNA and tRNA genes and nuclear tRNA genes. Introns excised by spliceosomes (groups of proteins and small nuclear RNAs) arose following the development of the nucleus 1.7 to 1 billion years ago and were inserted into previously uninterrupted genes following the development of mitochondria 1 billion years ago. Cavalier-Smith bases this phylogeny on bacterial phylogenies and the facts that spliceosomal introns are never found in bacteria, and that the introns in mitochondrial and chloroplast genes are not in corresponding positions, which would be expected if they were present in ancestral genes prior to the development of organelles. The similarity of the splicing consensus sequence for all types of introns suggests that they may all have a common evolutionary origin (reviewed in Cech, 1986). Cavalier-Smith (1985) considers it highly unlikely that introns were present in the ancestral genes and were subsequently deleted from every prokaryotic gene.

In support of the Introns-Late theory there is evidence that at least some introns were inserted into genes much later than the point in evolution at which eukaryotes appeared. Some examples are the intermediate filament multigene family (Lewis and Cowan, 1986, Dibb and Newman, 1989, and van Daal et al., 1990). All three of these cases involve comparisons of intron positions which suggest that novel introns have arisen at unique sites in certain genes and do not appear in genes of common evolutionary origin. In order to support the idea that such events are selected against

proponents of this theory usually cite the observation of Craik et al., (1982) that introns do not divide genes into regions that encode functional domains of proteins and that introns are generally inserted between regions encoding domains. They propose that introns are generally inserted between regions encoding domains, perhaps at pre-existing proto-splice sites (Dibb and Newman, 1989)

Introns in the Primordial Genes

The alternate view of intron evolution, that introns were present in the very earliest genes is called "Introns-Early." This possibility was first presented by Doolittle (1978), who responded to Gilbert's (1978) proposal that introns originated in eukaryotes by pointing out that perhaps prokaryotes eliminated introns and RNA splicing to reduce energy costs as an adaptation to their accelerated life-cycle. Doolittle also pointed out that the assertion of an evolutionary advantage as justification for the development and maintenance of introns was in conflict with Darwinian evolution. Both Doolittle (1978) and Darnell (1978) disagreed with the "Introns-Late" theory from the stand-point that the interruption of a continuous coding sequence would be extremely disruptive, even if a splicing mechanism existed to compensate.

Gilbert (1985) pointed out that the test of an early origin for introns would be to examine genes in diverse eukaryotic taxa that appeared very early in evolutionary history. If introns appeared very early their positions would be conserved in these modern

genes. Shortly thereafter Gilbert's group (Gilbert et al. 1986, Marchionni and Gilbert, 1986) presented evidence of the conservation of intron position over evolutionary time periods. The conservation of five intron positions in the triosephosphate isomerase (TIM) gene between maize and vertebrates suggested that these particular introns were present in the gene of the common ancestor of plants and animals over 1 billion years ago. The positions of three of these introns were also very close to those of the TIM gene of fungi in the genus of *Aspergillus*. Other examples of conservation of intron position were reported (e.g. Quigley et al., 1988, Obaru et al., 1988), including observations that intron positions predated the eukaryote/prokaryote split (Shih et al., 1988). Gilbert (1987) had used his observations to formulate the "Exon Theory of Genes", which proposed that modern multidomain genes were built up by recombinations within intervening sequences that separated "minigenes." These minigenes encoded short polypeptides capable of a single function, which when linked formed novel proteins. The nature of these minigenes is discussed further in the following section on exon shuffling. The Exon Theory of Genes also stated that while introns could be inserted or move position slightly by sliding (Craik et al., 1983) these events were rare and intron loss was the main process shaping modern genes. The loss of introns from genes was first noted by Perler et al. (1980), and was soon reported for many other genes (recent e.g.'s Nojima, 1987, Brown et al., 1987, Perret et

al., 1988, Downie et al., 1991). The discovery of pseudogenes produced from a retrotranscribed transcript (Vanin et al., 1980, Lewin, 1983) suggested the possibility that a cross-over between a gene and a retrotranscribed partially-spliced transcript could eliminate introns (Vanin, 1984). The loss of introns formed more complex exons encoding more complex protein domains. The existence of gene families where introns were present in some genes but not the others was attributed to the loss of introns from the ancestral gene, which had contained all possible introns (Marchionni and Gilbert, 1986, Gilbert, 1987). Gilbert's theory suggested that modern proteins were built from a small number of original "building blocks" and not from the association of individual amino acids. He presented evidence of this by examining a sequence database for the minimum number of exons required to produce every gene present (Dorit et al., 1990). This analysis suggested that there were a mere 7000 exons from which all modern genes were built.

The Big Picture of Intron Evolution

Stone et al. (1985) proposed that introns remaining from the original assembly of genes, and introns derived from recently inserted retrotransposons could co-exist in the same gene. Gilbert's proposal that the ancestral genes contained the maximum number of introns is too general a rule given the evidence for intron insertion, and the fact that, following this rule, either some

ancestral exons were only a few nucleotides in length or the junctions have undergone sliding, a mechanism for which there is little evidence (Rogers, 1986, 1989). This suggests a model of intron evolution in which some introns are the remnants of the assembly of the primordial genes but any general pattern remaining from this process has been obscured by the subsequent insertion and deletion of introns (Rogers, 1989, 1990, Hickey et al., 1989).

Do Introns Have a General Function?

Shuffling Introns

Any evolutionary scheme for introns must explain the reason for their appearance. Do they have a general role in gene function, or are they simply serving their own purposes? While it is now clear that introns can function in such processes as immunoglobulin gene rearrangement (Early et al., 1980, Cerg et al., 1980, Moore et al., 1981, Alt and Baltimore, 1982), regulation of gene expression (e.g. Banerji et al., 1983, Gillies et al., 1983, Behringer et al., 1987, Chung and Perry, 1989, Jonsson et al., 1992) and in alternative splicing of mRNAs to produce tissue-specific forms of various proteins (reviewed in Smith et al., 1990), these are obviously not the primary functions of introns.

Gilbert (1978), one of the first people to see the potential functions of introns, visualized their role as sites for recombination events that would combine exons encoding functional domains of

proteins not previously linked. This process was viewed as producing novel genes encoding proteins capable of performing new combinations of functions. Prior to this it was thought that new genes arose following gene duplications that allowed the original enzyme to be preserved by one copy, while the other mutated to produce new functions. Since mutations are often deleterious, exon shuffling would allow a more rapid and specific evolution of genes. Gilbert (1978) argued that the introduction of introns into the genes of eukaryotes had speeded their evolution with respect to prokaryotes. Blake (1978) carried the Exon Shuffling theory one step further by suggesting that each exon encoded a discrete functional domain of a protein, thus new combinations of exons would be stable since each domain would be a stable, self-contained unit. This idea was derived from the observations of Rossman et al. (1974), who noted that different enzymes contained domains of similar structure and function. Thus introns and the exons they separated were considered to be relics of the process by which multi-domain proteins were built. A similar correlation of exons with protein structural domains was proposed by Go (1981). Structural domains or modules were defined as groups of amino acid residues within a compact radius of 27Å. The intron/exon structure of genes did not always correlate with modules, but it seemed more reasonable than Blake's suggested correlation with functional domains, which often encompassed several modules.

The structural domain hypothesis was further supported by Craik et al. (1982) who noted that introns usually interrupted amino acid sequences in regions located on protein surfaces. This was due to the prevalence of hydrophilic amino acids in the exon on the 5' side of the intron/exon boundary. This was seen by Craik et al. as a consequence of the joining of exons and thus protein segments. New sequences of amino acids produced by shuffling of exons would be isolated on the protein surface and thus could not disrupt the pre-existing three-dimensional structure (and thus function) of the protein. In addition, if exons did correlate with protein domains, the functional region would be isolated in the centre of the exon, while the ends of the exons would encode joining segments used to link functional regions together, presumably out on the protein surface.

Definitive evidence for exon shuffling would be the discovery that the same exon was used to encode protein domains in two or more different genes when a similar structural or functional element was required. This was demonstrated for the gene encoding the low-density lipoprotein (LDL) receptor (Sudhoff et al., 1985a, 1985b). Cell surface receptors usually perform several functions, each carried out by a discrete protein domain. Sudhoff et al. (1985a) found that the domains of the LDL receptor were strongly correlated with the intron/exon structure of its gene. Following a 21 amino acid (aa) hydrophobic signal sequence, the first domain of the LDL receptor consists of approximately 300 aa's

made up of seven repeats of 40 aa's, each of which contained six cysteine residues. Sudhoff et al. (1985a) suggested that this 40 aa repeated unit was homologous to the cysteine-rich 40 aa sequence of the human complement C9 protein. The second domain of the LDL receptor is a sequence of 400 aa with about 33% sequence identity to the external domain of the epidermal growth factor (EGF) precursor. The next three domains of the LDL receptor are: a 48 aa sequence that attaches carbohydrate chains; a membrane-spanning domain; and a 50 aa cytoplasmic region. The organization of the LDL receptor gene reflects this domain structure, since all of the above-mentioned domains are separated from each other by an intron. In addition, introns precisely divide exons encoding the 40 aa basic unit in the first domain, with the exception of the third, fourth and fifth repeats, which are grouped within a single exon. The domain with sequence similarity to the EGF precursor is divided into eight exons, five of which are located in identical positions to those found in the EGF precursor gene. Sudhoff et al. (1985b) concluded that the LDL receptor gene is a mosaic derived from an exon encoding the 40 aa cysteine-rich region, duplicated seven times, a cassette of eight exons from the EGF precursor gene-like region, and exons encoding the other four functional domains. At least some of these exons had been shuffled to produce other genes.

Other examples of shuffled exons have been reported. Duester et al. (1986) suggested that the nucleotide-binding

domains of several alcohol dehydrogenases, chicken glyceraldehyde phosphate dehydrogenase, mouse lactate dehydrogenase, and chicken pyruvate kinase were all derived from one ancestral domain which was subsequently shuffled to produce many genes. Domain duplication and intron loss were used to explain the differences in structure of the descendant genes. Kimura et al. (1989) proposed that exon shuffling has added four domains onto the end of the human thyroid peroxidase (TPO) gene, including an EGF-LDF receptor domain and a transmembrane domain. These domains are not present in the human myeloperoxidase gene, which is similar in structure and sequence to the TPO gene to the end of the 12th exon.

An example of exon shuffling in action was reported by Hall et al. (1989). They showed that two spontaneous deletion mutants of T4 bacteriophage had joined exons from different genes by recombination between similar sequences in self-splicing introns. One of the resultant hybrid introns was capable of self-splicing.

Problems With Shuffling

Traut (1988) examined many genes for correspondence between exons and functional or structural domains of the encoded protein. He suggests that the average size of exons is too small for them to encode domains , and too large for them to correspond to the smallest structural features. Their size is comparable to that of functional regions of proteins. Traut points out, however, that

except for several good examples there is a general lack of correspondence between exons and functional domains of proteins. This is probably due to an obscuring of ancestral gene structures by intron gain, loss and sliding. Traut's greatest criticism of the exon shuffling hypothesis lies in the fact that the phases of introns and exons are not consistent. The phase of an intron/exon junction is defined by the position at which it divides the codon; those that divide the coding sequence between codons are in phase 0, while a junction dividing the first nucleotide of the codon from the other two is called phase 1, and one that divides the first two nucleotides from the third is defined as phase 2 (reviewed in Li and Graur, 1991). In a recombination within an intron the splicing junctions of the exons that are brought together must be in the same phase to produce a functional recombinant protein. Traut noted that in all the genes he surveyed, 54% of all intron/exon junctions were phase 0, 27% were phase 1, and 18% were phase 2. Phase 0 junctions would be the ideal condition for exon shuffling, since the recombined exons would not require extra nucleotides to complete codons, but Traut argues that the prevalence of phase 0 junctions is due to codon usage.

Patthy (1987) had also seen the phase problem in relation to the exon shuffling theory. He surmised that if exon shuffling was to be effective, all exons should start and end in the same phase (i.e. be symmetrical). Patthy noticed that some of the best examples of shuffled exons from the glyceraldehyde-3-phosphate

dehydrogenase and phosphoglycerate kinase genes were not symmetrical. He concluded that non-symmetrical exons were produced by the loss or insertion of introns. He also theorized that exon shuffling was limited early in evolutionary history, as all introns were self-splicing. The argument is that self-splicing introns required internal sequences to maintain their ability to splice, thus successful recombinations between them would be rare. Exon shuffling would be more prevalent later in evolutionary history when cellular splicing mechanisms that left intron sequences free to drift developed. Gilbert (1985) had suggested that exon shuffling was used to produce those genes that were the late products of evolution. On the other hand, it could be argued that the sequence similarity of self-splicing introns promoted recombination and thus enhanced exon shuffling early in evolutionary history.

In conclusion, proteins appear to have been built up from ancestral exon-sized minigenes, by shuffling, but the pattern of exon-protein structure correlation has been obscured in modern genes by intron loss and insertion. The insertion of introns has destroyed the original phase symmetry of the exon-like minigenes. This correlates well with the consensus model of intron evolution presented above (Rogers, 1989, 1990, Hickey et al., 1989).

Do Introns Have Any Function?

Orgel and Crick (1980) and Doolittle and Sapienza (1980) simultaneously published the theory that introns may not have a general function at all but may, like repetitive DNA, be an example of "selfish DNA". This is DNA that does not serve a function, but merely ensures its own propagation. Doolittle and Sapienza (1980) stated that such DNA would not necessarily be selected against at the phenotypic level, since it does not present a significant energy burden to its host. They suggest that the successful propagation of such sequences throughout the genome would be enough to ensure their survival. Both papers suggest that introns arose from transposable elements. This model suggests that the splicing mechanism in modern eukaryotes developed as a defense against these sequences.

CHAPTER 2: ISOLATION AND SEQUENCING OF A cDNA CLONE

Introduction

The transport of oxygen in multicellular organisms can be facilitated by various types of proteins. While most vertebrates transport oxygen by an intracellular heme-containing haemoglobin, oxygen transport in invertebrates is more diversified. The transport proteins in invertebrates can be subdivided into three major classes (Reviewed in Wood, 1980 and Mangum, 1985). First, hemocyanins are non-heme extracellular proteins containing Cu(I) atoms and are found in many molluscs and arthropods (van Holde and Miller, 1982, Volbeda and Hol, 1989). Second, intracellular non-heme hemerythrins utilize Fe(II) atoms and are found in annelids, brachiopods, priapulids and sipunculids (Klippenstein, 1980). The last group are the extracellular heme-iron containing haemoglobins that resemble vertebrate haemoglobins (Terwilliger, 1980). The invertebrate haemoglobins can also be grouped by their quaternary structure (Vinogradov, 1985). These structures include single-repeat polypeptides organized as single or multiple subunits, two repeat polypeptides arranged as multi-subunit and multi-repeat polypeptides, also organized as multiple subunits. An evolutionary tree for various invertebrate globins has been constructed (Goodman et al., 1988). Recently, cDNA clones have been sequenced that code for a two repeat haemoglobin from the

clam *Barbatia reeveana* (Riggs and Riggs, 1991) and a nine repeat globin from the brine shrimp *Artemia salina* (Manning et al., 1990).

One of the most unusual haemoglobins found was that isolated from the parasitic nematode *Ascaris*. This haemoglobin was shown to have the highest affinity for oxygen among haemoglobins due to a decreased dissociation rate (Davenport, 1949a). This high oxygen affinity was also found in haemoglobins of other parasitic nematodes that inhabit the alimentary system (Vinogradov, 1985). Furthermore, *Ascaris* haemoglobin has an unusually large (40 kDa) monomeric size (Wittenberg et al., 1985). This unusual size has placed nematode haemoglobins into their own group. Several groups have estimated that this monomer contains only one heme group (Terwilliger, 1980, Wittenberg et al., 1985). Darawshe et al. (1987) have shown that the full heme binding potential of *Ascaris* haemoglobin is not normally utilized and the exogenous addition of heme results in a single polypeptide carrying two hemes.

The codworm, *Pseudoterranova decipiens*, is a major problem to the Atlantic fishing industry. As part of a study that is characterizing the parasite proteins that are immunogenic in grey seals, the nematodes definitive host, I wanted to characterize the 40 kDamajor antigen. To this end a cDNA library was prepared from the worms. If a cDNA clone which encoded this antigen could be obtained, it could be used to further characterize this molecule, and perhaps to produce large amounts of this antigen in a bacterial expression system. The cDNA library was prepared in λgt11, in

order to facilitate screening with antibodies raised against the antigen in adult Grey seals.

In this report I present the complete derived sequence of an immunogenic haemoglobin from this parasitic nematode. This molecule shows a duplication of the normal monomeric haemoglobin-type protein, resulting in a single chain protein containing two potential heme and oxygen binding sites. This molecule also contains several potential copper binding sites. A portion of the molecule also may contain the remnants of an FMN binding site similar to that found in other electron transport molecules (e.g. cytochromes). These unusual characteristics for a haemoglobin molecule further support the theory that the haemoglobins may have a common ancestor with the copper and heme containing cytochromes.

Materials and Methods

RNA Preparation

Larval nematodes were removed from fillets of Atlantic Cod and stored at 4°C in Hanks buffered saline solution (HBSS). The nematodes were washed three times in HBSS prior to extraction. Total cellular RNA was obtained from homogenization of 50 larval nematodes by the method of Chirgwin et al. (1979). The integrity and quantity of the RNA was analysed by electrophoretic fractionation on a denaturing 0.8% agarose gel at 90 V for 2 hours.

Construction of cDNA Library and Screening

The complete cDNA clones were made available to me by Dr Pohajdak. The method used to obtain these clones is as follows. Approximately 120 μg of total RNA was primed with oligo dT and cDNA was synthesized using the method of Gubler and Hoffman (1983). Eco R1 linker-adaptors (Invitrogen) were added to the cDNA that was then size-selected for >500 bp by electrophoresis on a 1% agarose gel. The cDNA was ligated into λ-gt11 (Young and Davis, 1983). The cDNA library was screened using a polyclonal antisera that were obtained from immunizing adult grey seals with both larval and adult *P. decipiens*. Antibodies were detected using an alkaline phosphatase conjugated protein A (Sigma) followed by the substrates NBT and BCIP (BRL). Filters were blocked, washed and probed by following the standard Western blot procedures (Towbin et al., 1979). Six positive clones were isolated and subcloned into pUC18 and M13mp18 for further characterization.

DNA Sequencing and Analysis

Single stranded DNA templates were prepared in M13mp18 and sequenced by the dideoxy chain termination method of Sanger et al. (1977) using T7 polymerase (Sequenase, United States Biochemical). Sequence data and analysis is presented for clone 18 which was one of the longest cDNAs isolated. The sequence of clone 18 was also confirmed by sequencing two independent cDNA

clones, 10 (1311 bp) and 38 (1033 bp). Sequences were compared with the GenBank data base (Release 63) and the National Biomedical Research Foundation, Protein Identification Resource (Release 23) using the FASTA sequence-analysis software package (Lipman and Pearson, 1985, Pearson and Lipman, 1988) and the program BLAST (Altschul et al., 1990).

Northern Analysis

Approximately 15µg of total RNA and 5µg of poly A^+ RNA were denatured and loaded on a denaturing 0.8% agarose gel and electrophoresed at 90 V for 2 hours. RNA was transferred to a nylon membrane (Amersham). The blot was probed with ^{32}P dCTP labelled random primed (Feinberg and Vogelstein, 1983) clone 10 cDNA insert, isolated from pUC18 clones with Eco R1. The probe was hybridized in 50% formamide overnight and the blot was washed at 55°C in 0.2X SSPE, 0.1% SDS for 30 minutes. Blots were autoradiographed overnight at room temperature.

Results

A λ-gt11 cDNA expression library was constructed from mRNA of the larval form of the parasitic nematode. By screening this library using seal antibodies directed against the nematode, six individual cDNA clones were isolated. They were all shown to contain similar sequences by cross-hybridization and restriction

mapping (data not shown). Three independent cDNA clones have been completely sequenced and the 1371 bp DNA sequence reported here is clone 18 (Fig. 2.1). The two other cDNA clones which were sequenced are nearly identical (> 99% identity) and the corresponding protein sequences show only two conserved amino acid substitutions, that most likely reflect individual polymorphisms (data not shown). The mRNA is polyadenylated and contains the polyadenylation consensus sequence at position 1339 (Fig. 2.1, underlined). The mRNA sequence contains only one long open reading frame (ORF), that codes for a 333 amino acid protein (Fig. 2.1). This protein has a hydrophobic leader sequence with a perfect signal peptide consensus sequence (Pearlman and Halvorson, 1983) for cleavage after the 18th amino acid. The calculated molecular weight (Mr) of the mature protein is 37.6 kDa. There are two potential N-linked glycosylation sites at amino acid positions 68 and 215. The mature protein has an unusually high histidine content (10.5%) and five of the histidine-rich regions (amino acids 50, 114, 128, 160 and 199) potentially represent the copper binding histidine motif. In this motif the copper ion is bound by three histidine residues, two of which are In an a-helix and are three amino acid residues apart. The third histidine residue is located in an antiparallel α-helix approximately 30 to 40 amino acid residues towards the carboxyl terminal (Volbeda and Hol, 1989, Gaykema et al., 1984).

```
   1   GGAACCATT ATG CAC TCT TCA ATA GTT TTG GCC ATT GTG CTC TTC GTA GCG ATC GCT TCA
                 M  H  S  S  I  V  L  A  I  V  L  F  V  A  I  A  S              17
  61 GCA TCA AAA ACG CGA GAG CTA TGC ATG AAA TCG CTC GAG CAT GCC AAG GTT GGC ACC AGC
     A  S  K  T  R  E  L  C  M  K  S  L  E  H  A  K  V  G  T  S              37
 121 AAG GAG GCG AAG CAG GAC GGC ATC GAC CTC TAC AAA CAT ATG TTC GAG CAC TAT CCA GCA
      K  E  A  K  Q  D  G  I  D  L  Y  K  H  M  F  E  H  Y  P  A              57
 181 ATG AAG AAA TAC TTC AAG CAT CGT GAA AAT TAT ACA CCG GCC GAT GTC CAA AAG GAT CCC
      M  K  K  Y  F  K  H  R  E  N  Y  T  P  A  D  V  Q  K  D  P              77
 241 TTC TTT ATT AAA CAA GGT CAA AAT ATC TTG CTC GCC TGT CAC GTT TTG TGC GCC ACA TAC
      F  F  I  K  Q  G  Q  N  I  L  L  A  C  H  V  L  C  A  T  Y              97
 301 GAC GAT CGT GAG ACA TTC GAC GCG TAC GTT GGT GAG CTG ATG GCA CGA CAC GAG CGG GAC
      D  D  R  E  T  F  D  A  Y  V  G  E  L  M  A  R  H  E  R  D             117
 361 CAT GTT AAA ATA CCG AAT GAT GTT TGG AAT CAC TTC TGG GAA CAT TTC ATC GAG TTT CTG
      H  V  K  I  P  N  D  V  W  N  H  F  W  E  H  F  I  E  F  L             137
 421 GGA AGT AAG ACC ACG TTG GAC GAG CCA ACC AAG CAC GCA TGG CAA GAG ATC GGT AAA GAA
      G  S  K  T  T  L  D  E  P  T  K  H  A  W  Q  E  I  G  K  E             157
 481 TTC TCA CAT GAA ATC AGC CAC CAC GGT CGA CAT TCG GTT CGC GAC CAT TGC ATG AAC TCG
      F  S  H  E  I  S  H  H  G  R  H  S  V  R  D  H  C  M  N  S             177
 541 TTG GAG TAT ATC GCG ATC GGC GAT AAG GAA CAT CAA AAG CAG AAT GGC ATT GAC CTT TAC
      L  E  Y  I  A  I  G  D  K  E  H  Q  K  Q  N  G  I  D  L  Y             197
 601 AAG CAT ATG TTC GAG CAT TAT CCA CAT ATG AGA AAG GCA TTC AAG GGA CGC GAA AAC TTC
      K  H  M  F  E  H  Y  P  H  M  R  K  A  F  K  G  R  E  N  F             217
 661 ACG AAA GAA GAC GTT CAA AAG GAC GCA TTC TTC GTT AAA CAA GGA CAC AAG ATT CTG TTG
      T  K  E  D  V  Q  K  D  A  F  F  V  K  Q  G  H  K  I  L  L             237
 721 GCC CTT CGT ATG CTG TGC TCC TCA TAC GAT GAC GAG CCG ACA TTT GAC TAT TTT GTT GAT
      A  L  R  M  L  C  S  S  Y  D  D  E  P  T  F  D  Y  F  V  D             257
 781 GCC CTA ATG GAT CGT CAT ATC AAA GAT GAT ATT CAT CTA CCT CAG GAA CAA TGG CAT GAG
      A  L  M  D  R  H  I  K  D  D  I  H  L  P  Q  E  Q  W  H  E             277
 841 TTC TGG AAA TTG TTT GCC GAA TAT TTG AAC GAA AAG AGT CAC CAG CAT TTG ACA GAA GCC
      F  W  K  L  F  A  E  Y  L  N  E  K  S  H  Q  H  L  T  E  A             297
 901 GAG AAA CAT GCA TGG AGT ACA ATA GGT GAG GAC TTC GCG CAT GAG GCC GAT AAG CAT GCA
      E  K  H  A  W  S  T  I  G  E  D  F  A  H  E  A  D  K  H  A             317
 961 AAG GCC GAA AAA GAC CAT CAT GAA GGA GAG CAC AAA GAG GAA CAC CAC TGA ACCAACCCGTC
      K  A  E  K  D  H  H  E  G  E  H  K  E  E  H  H                         333

1023 GTCGTTCAACTTAAGCCTTCAGCTTAAGCTCGAGCTAAAGCCTCAGCTTGAGCTCAATCTTATGTCCTCAGGCCTAAAC
1102 TTGAATTTTAAAAGCATTTTGTTGAAGCAGTGCTAGCCAATCTCTTATCTTATCGGTGCTATTATCAATTTACTCTATG
1181 CCACCCCCCCCCCCCTCTCTCTGTTCTCTATTTGATATTCTGTTCTTTTAGTGCCAGATGTTAGTACCAGATGTTATTT
1260 TCTGCATAATTTTCTTCTCTTTACTTCGTTATTTTTTCGTTCTTCTATTTTTATGGCCAATTTTTGTGATGTCGAAGTC
1339 AATAAAACCATTTTTAAAAAAAAAAAAAAAAAA
```

Figure 2.1. DNA sequence of the 1371 bp cDNA clone of *P. decipiens* haemoglobin. The deduced amino acid sequence is shown below the DNA sequences. The polyadenylation signal sequence and the signal peptide sequences are underlined. Numbering on the left side of the diagram refers to nucleotide position and on the right side denotes amino acid position.

Probing a Northern blot with labelled cDNA revealed a single polyadenylated mRNA species of approximately 1.8 kb (Fig. 2.2a). The high expression of this mRNA in total RNA (Fig. 2.2b) suggests that this mRNA is present at a high steady-state level.

Analysis of the nucleotide sequence indicated that the gene is composed of two large duplicated regions. Nucleotide sequence from positions 73 to 486 shows 62.2% identity to nucleotides 521-945 (data not shown). This duplication results in a duplication of the protein sequence (Fig. 2.3). Using the program LFASTA the protein was shown to have two large repeated units of approximately 154 amino acids. The two repeats are 51% identical and 69% of the amino acid substitutions are conservative. Using the program RDF2.1 a statistical Z value of 33.4 was obtained. The RDF2.1 program compares a sequence with a permuted related sequence using three scoring methods (Lipman and Pearson, 1985, Pearson and Lipman, 1988). Scores of Z>3 are possibly significant, Z>6 are probably significant and Z>10 are significant. Both of the N-linked glycosylation sites (Fig. 2.3, second underlined region) are also conserved in this alignment. Two of the five potential copper binding motifs are also aligned in both repeats (Fig. 2.3, first underlined region). Several key amino acids which are important in maintaining secondary structure (e.g. proline, cysteine) are also aligned in both repeats.

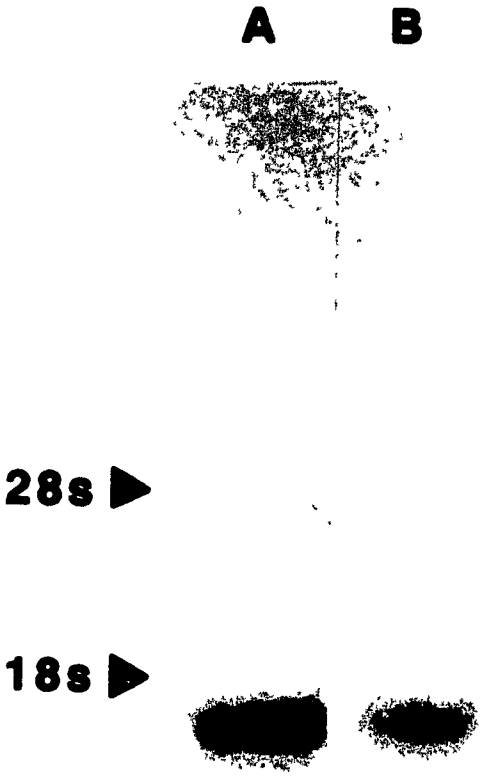


Figure 2.2. Northern blot of *P. decipiens* total (A) and poly A^+ (B) RNA probed with ^{32}P labelled nematode haemoglobin cDNA insert. Position of RNA size markers 28s and 18s was determined using ethidium bromide staining.

```
      10         20         30         40         50         60
R1 RELCMKSLEHAKVGTSKEAKQDGIDLYKHMFEHYPAMKKYFKHRENYTPADVQKDPFFIK
   :. ::.:::.  .:.....::.:::::::::::::: :.: :: :::.: .::::::.::.:
R2 RDHCMNSLEYIAIGDKEHQKQNGIDLYKHMFEHYPHMRKAFKGRENFTKEDVQKDAFFVK
        160        170        180        190        200        210

        70         80         90        100        110        120
R1 QGQNILLACHVLCATYDDRETFDAYVGELMARHERDHVKIPNDVWNHFWEHFIEFLGSKT
   ::..:::: ..::..::::  ::: .:..::.:: .:....:.. :..::. : :.:..:.
R2 QGHKILLALRMLCSSYDDEPTFDYFVDALMDRHIKDDIHLPQEQWHEFWKLFAEYLNEKS
        220        230        240        250        260        270

         130        140        150
R1 --TLDEPTKHAWQEIGKEFSHEISHHGRHSVRDH
      :.:..:::: .::..:.:: ..:.. .  .:
R2 HQHLTEAEKHAWSTIGEDFAHEADKHAKAEKDHH
        280        290        300
```

Figure 2.3. Sequence similarity between the first (R1) and second (R2) repeats of nematode haemoglobin. Numbering represents amino acid positions of the mature protein. There is 51% amino acid sequence identity between the two repeats (:) and 61% of the amino acid substitutions are conservative (.). Conserved sequences for probable N-linked glycosylation sites (N*T/S), the proximal portions of potential copper binding sites (H***H) and potential distal copper binding histidines are underlined (positions 74 and 215). Alignment was achieved using the program LFASTA from the FASTA sequence analysis package.

The secondary structure determination (Fig. 2.4) showed striking similarity between both of the repeats. The Kyte and Doolittle hydropathic index plot (Kyte and Doolittle, 1982) showed a strongly hydrophobic leader sequence. Both repeats have two identical hydrophobic pockets (Fig. 2.4, positions 85-90 and 235-240). By using the program of Garnier et al. (1978) for secondary structure analysis it was deduced that the protein contained 77.6% alpha helix. Several of these helices are also conserved in both repeats (Fig. 2.4).

Searching the invertebrate section of GenBank (Ver. 63) using the FASTA program revealed no overall similarities to sequences in this data base. However a search of the NBRF Protein Identification Resource (PIR) December 1989 version revealed significant sequence similarity to three invertebrate haemoglobins. Both repeats one and two show a high degree of sequence similarity to the myoglobins from the water snail, *Cerithidea rhizophorarum* (Takashi et al., 1983), the whelk, *Busychon canaliculatum* (Bonner and Laursen, 1977), and the extracellular haemoglobin of the polychaete worm, *Tylorrhyncus heterochaetus* (Suzuki and Gotoh, 1986). Both repeats also have significant sequence similarity to the human a globin chain (Dickerson and Geis, 1986) . A portion of the alignment of these molecules is shown in Fig. 2.5. The overall similarities and Z values for the comparisons are: repeat 1 versus water snail, 17.2% identity, Z=10.4; whelk, 20% identity, Z=9.6; polychaete worm, 22% identity, Z=3.8; human a chain, 13.6%

identity, Z=7.2; repeat 2 versus water snail, 15.1% identity, Z=10.9; whelk, 17.6% identity, Z=8.1; polychaete worm, 23.5%identity, Z=8.7; human a chain, 16.1% identity, Z=1.41. Other haemoglobins which have possible significant similarities (i.e. Z > 3), that are not shown above, include *Lumbricus terrestris* (Garlick and Riggs, 1982), *Petromyzon marinus* (Li and Riggs, 1970), *Glycera dibranchiata* (Imamura et al., 1972), *Aplysia limacina* myoglobin(Tentori et al., 1973) and *Heterodontus portjacksonii* α-globin(Nash et al., 1976) and myoglobin (Thompson, 1980). The non-heme containing hemocyanin from tarantula is close to significantly similar to repeat 1 (Z= 2.72)(Voit and Schneider, 1986).

Several key amino acid positions, which are invariant in all or most other haemoglobins are also conserved in both repeats 1 and 2 (Fig. 2.5). Most notable are the the proximal histidine at position F8 and the distal position E7, which can be either a histidine or a glutamine (Fig. 2.5). The similarity search also revealed a strong region of sequence similarity with NADPH-ferrihemoprotein reductase (E.C. 1.6.2.4) from four different organisms. This molecule is also referred to as NADPH-cytochrome P-450 reductase. The region of sequence similarity (25.5% identity, 66% conserved substitutions in a 47 amino acid overlap, Z=3.9) was between amino acids 1-47 in repeat 1 and amino acids 91-136 of the pig cytochrome reductase (Haniu et al., 1989). It is extremely interesting that this region corresponds exactly to the FMN binding region of the latter molecule (Yabusaki et al., 1988).

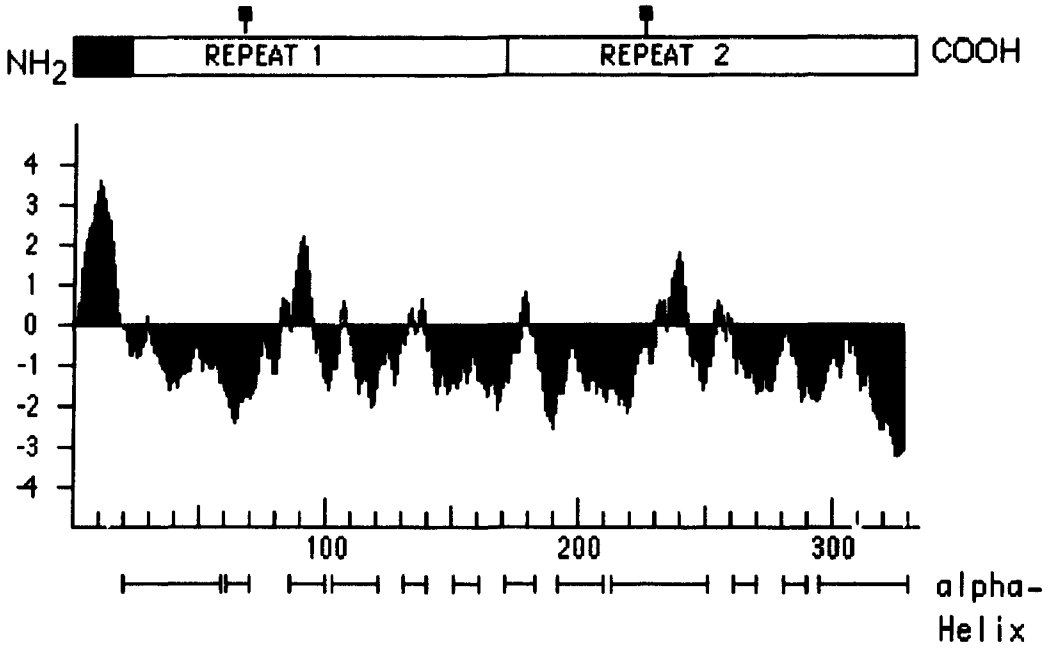


Figure 2.4. Secondary structure of nematode haemoglobin. The Kyte and Doolittle hydrophobicity plot in the centre shows the three major hydrophobic regions. The schematic diagram at the top demonstrates the leader sequence (black) and the proposed division between the duplicated repeats. The two possible N-linked glycosylation sites are marked with black squares. The bars at the bottom shows the location of the alpha helices that make up 77.6% of the secondary structure.

Figure 2.5. Two repeats of nematode haemoglobin (amino acids 43-115 and 192-264) aligned with human alpha globin (Human A) (Dickerson and Geis, 1986), globin IIA from the polychaete worm, *Tylorrhynchus heterochaetus* (worm IIA) (Suzuki and Gotoh, 1986), and haemoglobins from the water snail, *Cerithidea rhizophorarum* (Takashi et al., 1983), and the whelk, *Busychon canaliculatum* (Bonner and Laursen, 1977). Conserved amino acids are numbered as positions equivalent to those of human haemoglobin

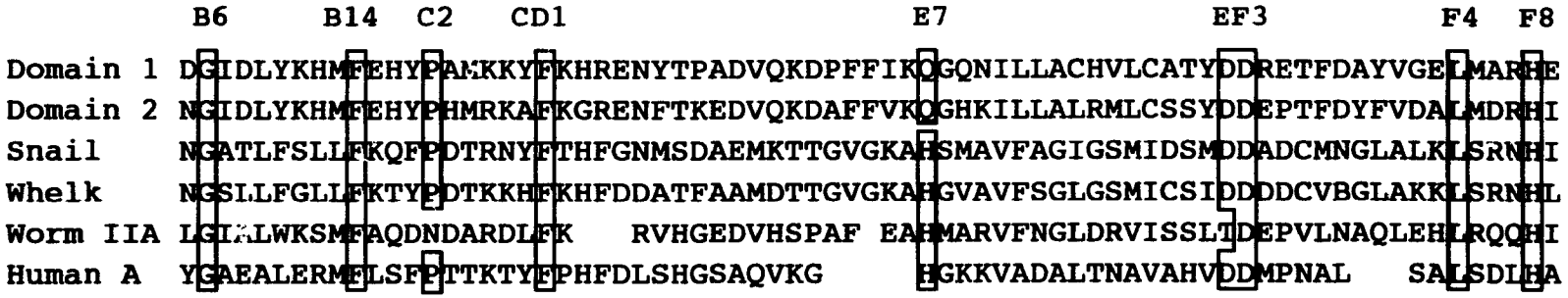

B6
B14
C2
CD1
E7
EF3
F4
F8
Domain 1
Domain 2
Snail
Whelk
Worm IIA
Human A

Figure 2.5

Discussion

Nematode haemoglobins have always been characterized as unusual due to their large size and their exceptionally high oxygen affinity (Davenport, 1949a, 1949b, Wittenberg et al., 1985, Darawshe et al.,1987). This report describes the complete sequence of a haemoglobin from the parasitic nematode *P. decipiens*. The data shows that the large size results from the protein being composed of two nearly identical repeats. These two repeats probably arose due to an ancient duplication event. *P. decipiens* haemoglobin is very similar to the haemoglobins isolated and characterized from the nematodes *Ascaris lumbricoides*, Mr 40.6 kDa (Wittenberg et al., 1985) and *A. suum*, Mr 41.6 kDa (Darawshe et al., 1987). Our calculated molecular weight of 37.6 kDa does not include the addition of two hemes (616 Daltons each) nor any increase due to glycosylation. I propose that nematode haemoglobin is glycosylated due to the exact alignment of the N-linked glycosylation sites in both of the repeats with that of the snail haemoglobin which has previously been shown to be glycosylated (Alfonso et al., 1980). SDS-PAGE analysis of nematode haemoglobin agrees with a size near 40 kDa (data not shown). Wittenberg et al. (1985) have estimated that *Ascaris* haemoglobin contains 332 amino acids by amino acid analysis. While this value differs from mature nematode haemoglobin (318 amino acids), the amino acid composition of *A. lumbricoides* (Wittenberg et al., 1985)

and *A. suum* is nearly identical to the values obtained from our deduced amino acid sequence (F=47.6, p=0.0001).

The haemoglobins from a variety of nematodes have all been shown to have an extremely high oxygen affinity (P 0.5 ~0.001-0.004 mmHg) due to a low dissociation rate (Davenport, 1949a). This may be because many nematodes live in anaerobic environments (e.g. digestive system). At the present time it is difficult to determine why nematode haemoglobins have this increased affinity from our sequence data. Both of our repeats utilize glutamine at the E7 position which is usually a histidine residue in other haemoglobins. Since this residue is intimately involved in hydrogen bonding to oxygen and decreases oxygen dissociation (Perutz, 1989), the substitution of a glutamine residue in this position may cause an alteration in oxygen affinity. Further molecular modelling, site directed mutagenesis of this residue and X-ray crystallography data may provide further mechanisms for this high oxygen affinity.

The haemoglobins from *Ascaris* were also unusual in their larger than normal size (Wittenberg et al., 1985, Darawshe et al., 1987). Each 40 kDa chain contained approximately one heme group. This was later shown to be an under utilization of the second heme-pocket as reconstitution with hemin revealed that the nematode haemoglobin was capable of binding two heme groups (Darawshe et al., 1987). The sequence data presented here suggests that both repeats are well conserved in sequence, secondary

structure and several of the invariant amino acids that play a key role in heme binding. Preliminary results suggest that various preparations of our nematode haemoglobin from perienteric fluid also have between one to two hemes per chain. The reason why this family of nematode haemoglobins is heterogeneous in their heme content is presently unknown. Future modelling and X-ray crystallography data may determine which repeat contains the heme group and what mutations or alterations are responsible for the decreased heme content. The rate of heme synthesis may not be proportional to globin protein production, therefore many of the globins may not be fully loaded with heme.

In addition to having a duplication event, this nematode haemoglobin sequence has several unique features not previously reported in the literature. Our molecule has two conserved potential copper binding sites. This motif is usually found in hemocyanin-type molecules (Volbeda and Hol, 1989) and in other select copper-containing proteins such as tyrosinase (Lerch, 1978). Whether copper can be bound by this nematode haemoglobin or plays a role in its function is presently unclear. This molecule has a region of sequence similarity with the electron transport protein NADPH cytochrome P450 reductase (E.C. 1.6.2.4). This region of similarity is the exact region that FMN binds to the cytochrome molecule (Yabusaki et al., 1988). Whether this FMN binding site is functional or a remnant of an ancient FMN binding domain in nematode haemoglobins is under investigation. Preliminary

phylogenetic analysis data indicate that this molecule may occupy it's own separate branch on a phylogenetic tree. The histidine copper binding motif and the FMN binding site may provide further evidence that the ancestral haemoglobins were heme containing cytochrome-like molecules.

CHAPTER 3: VARIATION IN COLOUR OF WORMS CORRELATES WITH HAEMOGLOBIN CONCENTRATION IN THE PSEUDOCOELOMIC FLUID.

Introduction

Larvae of the parasitic nematode *Pseudoterranova decipiens* (commonly known as sealworm) infect several species of demersal fish in the North Atlantic, including Atlantic cod (*Gadus morhua*) (Templeman et al., 1957, Chandra and Khan, 1988, Brattey et al., 1990, McClelland et al., 1990). The most important definitive hosts of the parasite appear to be grey seals (*Halichoerus grypus*) (Mansfield and Beck, 1977, Malouf, 1986). Adult worms found anchored in the gastric submucosa (McClelland, 1980) produce partly embryonated eggs which pass with the seals' feces. The eggs hatch into larvae that infect benthic macroinvertebrates. Secondary fish hosts ingest the macroinvertebrates and thus acquire the larvae. Larval codworms mature into adults after entering the definitive host's stomach (McClelland, 1990, Templeman, 1990).

While investigating immunological methods for controlling sealworm infestation, I characterized the gene for one of the major antigens of the worm, a 40 kDa haemoglobin from the pseudocoelomic fluid (Dixon et al., 1991). This report shows a direct correlation between the content of this haemoglobin and the colour of individual worms which have been described variously as white, yellow, reddish-brown or brown (Margolis, 1977, Bowen, 1990,

Sinderman, 1990, Appleton and Burt, 1991). The function of haemoglobin in parasitic nematodes is still unclear. Most investigations of the function of nematode haemoglobin have involved adult *Ascaris suum* or *Ascaris lumbricoides*. *Ascaris* and *P. decipiens* belong to the same order of nematodes and their pseudocoelomic haemoglobins are similar in size and protein sequence (64% identity, with 83% conservative substitution) (Dixon et al., 1991, 1992). Thus it is possible that haemoglobin in the pseudocoelomic fluid of *P. decipiens* is performing the same function as that found in *Ascaris*. Pseudocoelomic fluid haemoglobin, however, has such a high oxygen affinity that it is not likely to be deoxygenated *in vivo* and hence would not function in oxygen transport (Davenport, 1949a, Smith and Lee, 1965). Some researchers have suggested that *Ascaris* haemoglobin serves as a hematin sink, because these organisms are apparently unable to synthesize their own heme (Lee and Smith, 1963) while others speculate that it supplies oxygen for the eggs as they pass through the anaerobic lower gut of the seal (reviewed in Smith and Lee, 1965). Evidently the amount of pseudocoelomic fluid haemoglobin can vary a great deal with no apparent adverse effect on the organism (Smith and Lee, 1965). Recently Burr et al. (1989) report that nematode haemoglobins may also function in the photosensitivity of the ocelli of a free-living nematode. This suggests that the use of haemoglobin for functions other than oxygen transport by nematodes may not be unusual.

Materials and Methods

Source of Larval Codworm

Larval nematodes were removed from fillets of Atlantic cod supplied by National Sea Products, Lunenburg Division and stored at 4°C in Hanks buffered saline solution (HBSS). Worms of uniform size were separated according to colour into three groups, red, brown, and white. All worms were washed three times in HBSS.

Extraction of Pseudocoelomic Fluid

In the first experiment 30 worms were sliced into segments less than 1 mm in length to release the pseudocoelomic fluid into 1 mL of sterile water. In the second experiment the body wall of 20 worms was sliced carefully with a scalpel along their length with the aid of a dissecting microscope. Pooled pseudocoelomic fluid was centrifuged at 14,000 rpm (16,000 xg) for 10 minutes to remove tissue and insoluble material.

Protein Content Determination

The concentration of protein in the pseudocoelomic fluid was assayed using a bicinchoninic acid protein assay kit (Sigma). One ml of pooled pseudocoelomic fluid from each group was scanned using a Beckman model DU-64 UV/visible spectrophometer at wavelengths from 330 to 700 nm. The colour in haemoglobins is

derived from the heme group (Lee and Smith, 1965), which gives a characteristic spectrophotometric pattern.

Purification of Haemoglobin

The haemoglobin was purified from body fluid using the method of Darawshe et al. (1987). Pseudocoelomic fluid was bound on to DEAE-sephadex A50 in a column and eluted using 250 mM NaCl/ 25 mM Tris/HCl. The eluate was then passed through a Sephadex G-200 column. The coloured fractions were pooled and concentrated in an Amicon centriprep 10 concentrator.

Protein Composition and Western Blotting

The composition of the body fluid was determined by separating 10 mg of protein from each sample on a denaturing polyacrylamide gel (Laemmli, 1970), and staining with Coomassie brilliant blue (Sigma). The relative amount of haemoglobin in the pseudocoelomic fluid was determined by Western blotting. Each well of the denaturing polyacrylamide gel was loaded with 10 mg of protein from samples of each colour group. The protein was transferred from the gel onto a nitrocellulose filter by electroblotting. This filter was screened using a polyclonal antiserum obtained by immunizing juvenile grey seals with purified *P. decipiens* pseudocoelomic fluid haemoglobin. Antibodies were detected using an alkaline phosphatase conjugated protein A (Sigma) followed by the substrates NBT and BCIP (BRL).

Filters were blocked, washed and probed by following the standard Western blot procedures (Towbin et al., 1979). The relative intensity of the haemoglobin band from each sample was obtained by scanning the developed filter with an Abaton 300-gs scanner (Everex Systems, Inc, Fremont CA). The immunoreactive bands from the digitized image were analysed using Scan Analysis software (Biosoft, Ferguson, MO) on a Apple Macintosh computer (Apple Computer, Inc, Cupertino, CA).

RNA Preparation

Total cellular RNA was obtained from homogenization of 100 larval nematodes from each colour group by the lithium precipitation method of Palmiter (1974). The integrity and quantity of the RNA was analysed by electrophoretic fractionation on a denaturing 0.8% agarose gel at 90 V for 2 hours.

Northern Analysis

Approximately 20 μg of total RNA from worms of each colour group was denatured, loaded on a denaturing 0.8% agarose gel and electrophoresed at 90 V for 2 hours. RNA was transferred to a nylon membrane (Amersham). The blot was probed with ^{32}P dCTP labelled random primed (Feinberg and Vogelstein, 1983) clone 10 cDNA insert (Dixon et al., 1991), isolated from pUC18 clones with Eco RI. The probe was hybridized in 50% formamide overnight and the blot was washed at 55°C in 0.2X SSPE, 0.1% SDS for 30 minutes.

Blots were autoradiographed overnight at room temperature. Relative amounts of haemoglobin mRNA in each sample were determined by scanning the developed autoradiogram using the method outlined above for Western blots.

Results

Spectrophotometric analysis of pseudocoelomic fluid revealed that absorbance and, therefore, heme concentration, was highest in the fluid of the red worms and lowest in fluid from white worms (Fig. 3.1). Similar results were obtained from the pooled pseudocoelomic fluid of the three groups of twenty longitudinally sliced worms. Protein composition, as analysed by polyacrylamide gel electrophoresis and Coomassie brilliant blue staining (Fig. 3.2), did not differ in the pseudocoelomic fluid from the three sealworm colour groups and in all cases haemoglobin comprised more than 30% of the total protein. This is apparent in Fig. 3.2A; the predominant band is the 40 kDa haemoglobin molecule. Western analysis of each group of pooled pseudocoelomic fluid using polyclonal antiserum against the haemoglobin revealed that pseudocoelomic fluid from red worms contained the greatest concentration of haemoglobin, that from brown worms, intermediate amounts and that from white worms, the lowest concentrations (Fig 3.2B). These differences were quantified using

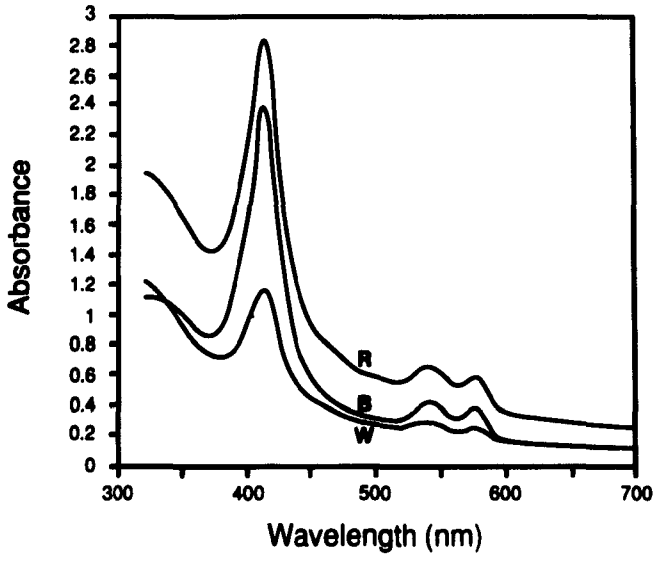


Figure 3.1. Absorbance of *P. decipiens* body fluid from different colour groups at wavelengths from 300 nm to 700nm. R: body fluid pooled from red worms. B: body fluid pooled from brown worms. W: body fluid pooled from white worms. The large peak at 400 nm and the secondary peaks at 540 and 580 nm are due to heme.

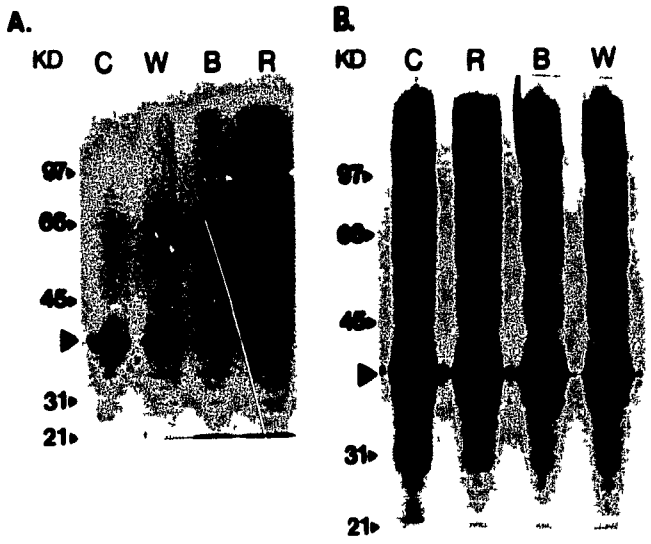


Figure 3.2. A) Protein constituents of *P. decipiens* body fluid analysed by polyacrylamide gel electrophoresis and B) Western blot of body fluid probed with grey seal anti-haemoglobin sera. Molecular weight in KiloDaltons is indicated on the left. C: enriched *P. decipiens* haemoglobin. R: body fluid pooled from red worms. B: body fluid pooled from brown worms. W: body fluid pooled from white worms. The arrow indicates the *P. decipiens* haemoglobin.

densitometry. The ratio of haemoglobin concentration in the pseudocoelomic fluids of red, brown and white worms was 1.0 : 0.77 : 0.56. This experiment was repeated three times. Since equal amounts of protein were loaded in each well of the gel this difference is in addition to differences in total protein concentration of the pseudocoelomic fluids from each sealworm colour group. The ratio of total protein concentrations in pseudocoelomic fluid was 1.0 : 0.84 : 0.66 (Red : Brown : White nematodes).

Probing of a Northern blot containing 20 mg of total RNA from each group revealed that there was no difference in haemoglobin mRNA content of the three colour groups of worms (Fig. 3.3). Quantification of the mRNA content by densitometry showed the ratio of band intensities to be 0.87 : 1.0 : 0.92 (Red : Brown : White nematodes).

Discussion

Variation of colour of larval sealworm from the flesh of cod has caused some investigators to speculate that there are more than one species or variant of *P. decipiens* (Appleton and Burt, 1991, Paggi et al., 1991). Multilocus electrophoresis has revealed the existence of three sibling species of *P. decipiens* in the North Atlantic, but of two siblings, *P. decipiens* B and C, found in Canadian waters only sibling B has been identified from important commercial fisheries (Paggi et al., 1991). We report here a variation

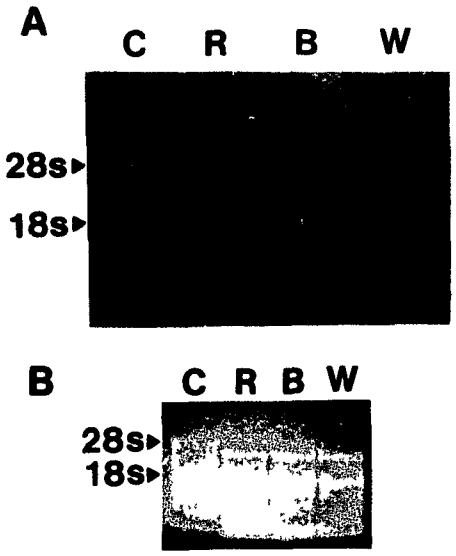


Figure 3.3. A) Northern Blot indicating the relative levels of haemoglobin mRNA. C: human Jurkatt cell line RNA. R: RNA pooled from Red worms. B: RNA pooled from Brown worms. W: RNA pooled from White worms. B) Agarose gel of 20 μg of total RNA from the three colour groups of worms. The arrows indicate the migration of the human 28s and 18s rRNA bands.

in body fluid protein content that does not indicate a difference in species, but may indicate the differential regulation of haemoglobin production. We tested the hypothesis that colour differences of *P. decipiens* larvae (red, brown and white) were due to the concentration of haemoglobin, a prominent protein in the pseudocoelomic fluid of nematodes. The colour in haemoglobins comes from the heme group (Smith and Lee, 1965) apparent in the spectrophotometric analysis of one mL of pooled pseudocoelomic fluid from thirty worms of each colour class (Fig. 3.1) as the large peaks at 400 nm, 540 nm and 580 nm (Hamada et al., 1963, Wittenberg et al., 1965). Heme concentration was not due to free heme since all the colour from the body fluid co-elutes from the G-200 column with the haemoglobin molecule during purification.

It was also noted that the pelleted body tissue contained very little colour (data not shown), and hence may differ from anothe ascaridoid nematode, *Ascaris lumbricoides,* which has a second form of haemoglobin in the body wall. Alternatively, the concentration of body wall haemoglobin in *P. decipiens* may be significantly lower than the pseudocoelomic form. Nevertheless, my findings indicate a marked difference in the heme content of pseudocoelomic fluids of different coloured worms.

Haemoglobin extracted from the nematodes could not include ingested host proteins as all vertebrate haemoglobins are approximately 17 kDa in size (Dickerson and Geis, 1983). *P.*

decipiens haemoglobin is 40 kDa in size as the gene encoding it has undergone an unusual head-to-tail duplication event (Dixon et al., 1992). The protein composition of pseudocoelomic fluid protein herein appeared to be similar in all three colour groups of nematodes (Fig. 3.2A). Haemoglobin constitutes more than 30% of the *P. decipiens* pseudocoelomic fluid protein. This contrasts with the body fluid of *Ascaris* where an 11 kDa protein which has proven allergenic in the definitive host predominates (McGibbon et al., 1990). *P. decipiens* also has an 11 kDa allergen. Antibodies against the *Ascaris* allergen (ABA-1) show a strong reactivity to the *P. decipiens* 11 kDa protein on a Western blot (McGibbon, Dixon, Lee and Pohajdak; unpublished data).

Coomassie staining of a gel containing equivalent amounts of protein from the pseudocoelomic fluid of each nematode colour group showed no apparent variation in the proportion of the haemoglobin band (Fig. 3.2A). Western blotting and screening with polyclonal sera to purified pseudocoelomic haemoglobin, however, revealed that the concentration of pseudocoelomic haemoglobin correlates with the variation in colour. The combination of the difference in proportion of haemoglobin in the pseudocoelomic fluid with the difference in total protein concentration between the three groups indicates that the brown worms have only 64.7% of the haemoglobin found in red worms, while white worms contain only 36.7% of the red worm level. Visually the difference in colour between the red and white worms appeared to be more than a 66%

decrease in red colour, but the haemoglobin in the white worms may not contain heme and thus would not contribute to the colour. The correlation of haemoglobin content and variation in colour has been reported for *A. lumbricoides*, an ascaridoid nematode with a similar body fluid haemoglobin (Smith and Lee, 1962).

Since haemoglobin is the major protein in the body fluid (Fig. 3.2) the increased protein content of red worms might reflect increased haemoglobin expression. To test this theory I isolated total RNA from worms of each group and examined the haemoglobin mRNA content by Northern analysis and densitometry (Fig. 3.3). The difference in mRNA content, as measured by densitometry, is not significant enough to account for the large difference in protein content in the three groups of worms. Thus difference in gene expression between the colour groups of worms is probably not due to differential regulation at the transcriptional level. Since haemoglobin mRNA from other organisms has been shown to have an extremely long half-life (Aviv et al., 1976, Lowenhaupt and Lingrel, 1978) perhaps the difference in protein levels is due to regulation at the translational or post-translational level.

Recently, I have reported the sequence of a genomic clone of *P. decipiens* haemoglobin (Dixon et al., 1992) and I am currently sequencing the upstream regulatory region of this gene to determine if its expression is oxygen or heme dependent. Regulation of this gene by heme or heme precursors is suggested

by the strong correlation of the decrease in colour derived from heme and the haemoglobin content. Heme and oxygen regulation of gene expression has been reported for the haemoglobin gene of *E. coli* (Khosla and Bailey, 1989) and for other genes (Zagorec et al., 1988, Minagawa et al., 1990, Belazzi et al., 1991). In *A. lumbricoides* haemoglobin content increases following the addition of heme precursors (Smith and Lee, 1965) as, evidently, this nematode cannot synthesize heme *de novo* due to an inability to synthesize porphyrin ring heme precursors (Lee and Smith, 1962).

CHAPTER 4: A NEMATODE HAEMOGLOBIN GENE CONTAINS AN INTRON PREVIOUSLY THOUGHT TO BE UNIQUE TO PLANTS

Introduction

Haemoglobin genes have previously been shown to possess one of three basic chromosomal organizations. Animal globin genes have a characteristic three-exon, two intron structure (Maniatis et al., 1980). While the size of the introns can vary considerably, as in the case of seal myoglobins (Blanchetot et al., 1983), the position of the introns is highly conserved, usually within six or seven codons of intron positions in globin genes of other species (Hardison, 1991). Plant haemoglobins have an extra, central intron which divides the heme-binding domain into two modules which bind the heme molecule from opposite sides (Jensen et al., 1981, Landsmann et al., 1986, Bogusz et al., 1988). The discovery of this central intron substantiated the earlier hypothesis of Go, who suggested that the ancestral haemoglobin gene would possess such an intron based on protein domain analysis (Go, 1981). This chromosomal organization was also found in other plant haemoglobin genes but not in any animal haemoglobin gene (Landsmann et al., 1986, Bogusz et al., 1988). It was hypothesized that this central intron was lost early in animal globin gene evolution (Go, 1981). The third haemoglobin gene organization is found only in the insect *Chironomus thummi*. This haemoglobin gene contains no introns

and is present in several copies throughout the genome of the animal (Antoine et al., 1987).

Extracellular proteins often contain hydrophobic signal leader peptides that are cleaved from the mature protein during secretion. Usually an intron separates the coding region for the leader sequence from the coding region of the mature protein (Breathnach and Chambon, 1981). An unusual observation is that several genes coding for extracellular haemoglobins from invertebrates do not have introns separating secretory leader sequences from the coding region for mature protein (Antoine and Nessing, 1984, Jhiang et al., 1988).

Previously, I reported the complete cDNA sequence of an internally duplicated haemoglobin from the parasitic nematode, *Pseudoterranova decipiens* (codworm, Dixon et al., 1991). Northern analysis indicated that this haemoglobin was encoded by an abundant 1.8 kb mRNA (Dixon et al., 1991). This extracellular haemoglobin contains 333 amino acids, has an 18-amino acid hydrophobic leader sequence and has the potential to bind two molecules of heme. In addition, a portion of this molecule had sequence similarity to cytochromes. This led me to conclude that this haemoglobin was similar to the ancestral haemoglobin molecule. I report here the isolation and complete sequence of the gene for this haemoglobin. The chromosomal organization of this duplicated haemoglobin consists of seven exons and six introns. There are four introns in the first repeat, one intron which

separates the two repeats and only one intron in the second repeat. The second and fourth introns of the first repeat correspond to the two introns found in animal globin genes, but the third intron is located in a position corresponding to the central intron previously found in only plant haemoglobin genes. This unusual chromosomal structure suggests that either animal haemoglobin genes lost the central intron much later than previously hypothesized or that the central intron was lost then re-acquired by either *P. decipiens* or plant haemoglobin genes. The similarity in gene organization to an ancestral haemoglobin gene and the protein sequence similarity to cytochromes suggests that *P. decipiens* haemoglobin was structurally similar to an ancestral haemoglobin prior to its duplication.

Materials and Methods

Library Construction

The genomic clones described in this chapter were provided to me by Dr. Pohajdak. They were obtained as follows. Genomic DNA from *P. decipiens* was partially digested with Sau 3AI and partially filled in using the Klenow fragment of *E. coli* DNA polymerase (Sambrook et al., 1989). The DNA was then ligated into Xho I half-site arms of the vector EMBL 3 (Promega) and packaged using the Promega packaging kit. The resultant library was then

screened with a cDNA clone of *P. decipiens* haemoglobin (Dixon et al., 1991). Five positive clones were obtained.

Template Preparation and DNA Sequencing

The sequence presented here was obtained by completely sequencing both strands of two clones, W1 and W5. These two clones were either digested with Eco RI and subcloned into M13mp18, or used as templates in polymerase chain reactions (PCR) with primers to cDNA sequences (Dixon et al., 1991). PCR products were ligated into either the vector pCR1000 (Invitrogen), or pUCBM20 (Boehringer-Mannheim) and prepared for sequencing using the alkaline lysis mini-preparation technique (Sambrook et al., 1989). Single stranded DNA templates were prepared for sequencing using the method of Sanger et al. (1977). All sequencing reactions were carried out by the dideoxy chain termination technique (Sanger et al., 1977) using T7 polymerase (Sequenase, United States Biochemical).

Southern Blotting

For the Southern analysis 15 μg of *P. decipiens* genomic DNA pooled from 50 individuals was digested to completion with Eco RI, and electrophoresed at 25 V for 13 hours in a 0.8% agarose gel run in 1X TBE (pH 8). The gel was stained with 0.5 mg/mL ethidium bromide and photographed. The gel was transferred to nylon (Amersham Corporation) (Sambrook *et al.*, 1989). The radiolabelled

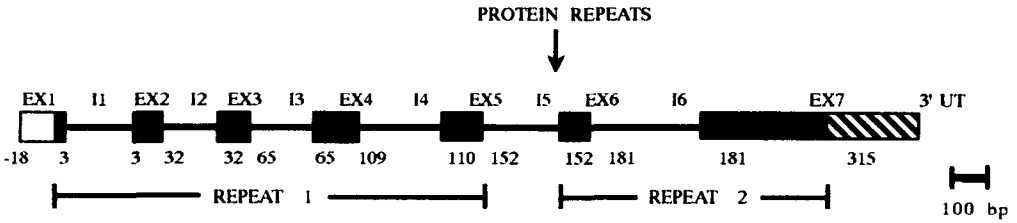
DNA probe was prepared by random priming of *P. decipiens* haemoglobin cDNA (Dixon et al., 1991). The blot was probed with 2 x 10^6 cpm per mL at 42 ^oC for 18 hours in 5X SSPE pH 7.4 (0.75 M NaCl; 50 mM $NaHPO_4$; 5 mM Na_2EDTA), 0.1% w/v sodium dodecylsulfate, 1% w/v bovine serum albumin, 150 mg/mL tRNA, 10% w/v sodium dextran sulphate, 50% w/v formamide, 1X Denhardt's solution (0.02% w/v polyvinylpyrrolidine; 0.02% w/v bovine serum albumin; 0.02% w/v Ficoll 400). The blot was washed six times with 0.2X SSPE pH 7.4, 0.1% w/v sodium dodecylsulfate at 55 ^oC, and exposed for 72 hours at -80 ^oC with an intensifying screen.

Results and Discussion

All invertebrate haemoglobin genes sequenced to date, such as those of the annelid *Lumbricus terrestris* (Jhiang et al., 1988), the mollusc *Aradara trapezia* (Titchen et al., 1991), and the clam *Barbatia reeveana* (Naito et al., 1991), have the common three-exon, two-intron pattern (i.e. they lack the central intron). However the genomic organization of a duplicated nematode haemoglobin gene isolated from *P. decipiens* contained seven exons and six introns (Fig. 4.1). The introns vary in size from from 164 (I2) to 335 (I6) base pairs and all exon/intron boundaries have the expected acceptor/donor splice sites. One of the introns starts at amino acid position 65 (Fig. 4.1). When the amino acid sequences of *P. decipiens* and plant haemoglobins are aligned the position of

Figure 4.1. (Next two pages) The structure and sequence of the *P. decipiens* haemoglobin gene. I) Schematic diagram showing the intron (line)/exon (black box) arrangement, the leader sequence is shown as a white box and the 3' untranslated region is shown as a hatched box. The numbers below denote the amino acid position of the intron/exon boundaries in the mature protein. The point of gene duplication is shown, as well as the boundaries of the duplicated repeats. II) The DNA sequence of the *P. decipiens* haemoglobin gene with the deduced amino acid sequence presented below. The numbers indicate the nucleotide number of the gene or amino acid number of the mature protein. Bases corresponding to the intron splice consensus sequences are underlined, as is the proposed peptide leader sequence.

I



II

```
1   GGAACCATTATGCACTCTTCAATAGTTTTGGCCACCGTGCTCTTTGTAGCGATTGCTTCAGCATCAAAAAGTAAGTTTCACCTCCCTCCCCCCCCCCTTTTATCAATACTTTCCCAAGCT
-18          M  H  S  S  I  V  L  A  T  V  L  F  V  A  I  A  S  A  S  K

121 TCCCGAAACGAATCCTTCAAGTTTCCCAACCTCTTCAACTAGTTCCCCAAATTTCACTACTTCCCGAGCTTTGGGAAATAAACTAGAATTCTCCCAAATCCCGTGAGCAAGACGTGAAAA

241 TTTTCGACATTTCAGCGCGAGAGCTATGCATGAAATCGCTCGAGCATGCCAAGGTTGGCACCAGCAAGGAGGCGAAGCAGGACGGCATCGACCTCTACAAACAGTACGTTTGGCCCTCCA
3               T  R  E  L  C  M  K  S  L  E  H  A  K  V  G  T  S  K  E  A  K  Q  D  G  I  D  L  Y  K  H

361 TTTTTCCCCGCTTTCCTCGAAATTTCGGCACCATTTTCGTGGAAAAATGAGAAAAAATTTTTTTGAAAAATTACGAAAAAAATATTTCAGTTAGACTAAAAAATACTAAATATTTATAGAT

481 ACTGAATACTAAATATTAAATATTTCAGTATGTTCGAGCACTATCCAGCAATGAAGAAATACTTCAAGCATCGTGAAAATTATACACCGGCCGATGTCCAAAAGGATCCCTTCTTTATTA
33                                  M  F  E  H  Y  P  A  M  K  K  Y  F  K  H  R  E  N  Y  T  P  A  D  V  Q  K  D  P  F  I

601 AACAAGGTAGGTTCCACAATCCCCTCCCCCGCCGCCCCCCCCCCCACGCCAATCTCCAATCTCTCGAGAAAACTCCCAAAGAGAAAACACCATCTAGAAAAATTTAGAAGGTCCCGATGT
63  K  Q

721 CGGACCAATTTCAGGTCTCATCTGAACCCAATTTCAGGTCCCATCCAACCTGAATTTCAGGTCTCATCCAACCTCCCTTTCAGGTCAAAATATCTTGCTCGCCTGTCACGTTTTGTGCGC
65                                                                                  G  Q  N  I  L  L  A  C  H  V  L  C  A

841 CACATACGACGATCGTGAGACATTCGACGCGTACGTTGGTGAGCTGATGGCACGACACGAGCGGGACCATGTTAAAATACCGAATGATGTTTGGAATGTAAGCAATCTAGAAGTCTAGTC
78    T  Y  D  D  R  E  T  F  D  A  Y  V  G  E  L  M  A  R  H  E  R  D  H  V  K  I  P  N  D  V  W  N

961 TATCCCCCCCTCTTCCCCTGTCTGCACACACCTCAGCTAATTTCAGGTCTCGTATCTAGAGATTTTAGGTTCGGTACCACAGATGAAGTATGAATTCATAAAATTGAGGGTTTTAAAATT
```

Figure 4.1

```
1081 GCAGCACAAAACCAAAATTTTAAATTTTAAAATTTAATTTCAAATTTTTAAAAATTTTTTAAAATATTTTTTAAAATCCGGTTCAATCTTTGAGCCTCGAGAAAATTTCCAGCACTTCTG
 110                                                                                                                  H  F  W

1201 GGAACATTTCATCGAGTTTCTGGGAAGTAAGACCACGTTGGACGAGCCAACCAAGCACGCATGGCAAGAGATCGGTAAAGAATTCTCACATGAAATCAGCCACCACGGTCGACATTCCGG
 113   E  H  F  I  E  F  L  G  S  K  T  T  L  D  E  P  T  K  H  A  W  Q  E  I  G  K  E  F  S  H  E  I  S  H  H  G  R  H  S

1321 TAAGTCACACGATTCATAAAGTTTTGGGCAAATATCCGGGCAAAACCGGACCACCCTGCGGCATTGCAAACGTTTTTTGGGTAGTTTCTTTTGGGTAAGTTAGGCGGACGGGCCGCTCAA

1441 ATAAAACTGCCCCACGTCTTCTTCGCAGCGTGATACTCCAGTCATTTTGAGTGGCTGTCATGCCTGCGGGTAAAAATTGCCCATTCGGGTATTTTTTATTCCCGAAATTCATCCACTTTA

1561 AGGTTGAACTCACATTGTGTTGAGCATTTCAGTTCGCGACCATTGCATGAACTCGTTGGAGTATATCGCGATCGGCGATAAGGAACATCAAAAGCAGAATGGCATTGACCTTTACAAGCA
 157                                V  R  D  H  C  M  N  S  L  E  Y  I  A  I  G  D  K  E  H  Q  K  Q  N  G  I  D  L  Y  K  H

1681 GTAAGTTTACAAGCTGAAAGGTATTAGGGTTCAAAATTTTGACAAATTCCCCGGGAAACAAATCATTTCCCGGGAACCATCTCCATATTCCAGAAATTACACCCCCTATTCCCATAAATA

1801 CACGGGGTTTGCCTATTATTGGGAATAGGGGTGGGGAAAGGGGGATCAATTTCGTTAAAATTTAGCGTTTCCAGTGGATTGTGGGTGATATTCTAGCCGCAAAAATGATGACTTAGTAGT

1921 ATCGACAGTGAAATATTACAGTAACATTTATTTTTGTACATTACAGTAATATTTATTACTGTACATTACAGTAATATTTGTTTGTGTACATTACAGTATGTTCGAGCATTATCCACATAT
 182                                                                                                  M  F  E  H  Y  P  H  M

2041 GAGAAAGGCATTCAAGGGACGCGAAAACTTCACGAAAGAAGACGTTCAAAAGGACGCATTCTTCGTTAAACAAGGACACAAGATTCTGTTGGCCCTTCGTATGCTGTGCTCCTCATACGA
 190   R  K  A  F  K  G  R  E  N  F  T  K  E  D  V  Q  K  D  A  F  F  V  K  Q  G  H  K  I  L  L  A  L  R  M  L  C  S  S  Y  D

2161 TGACGAGCCAACATTCGACTATTTTGTTGATGCCCTAATGGATCGTCATATCAAAGATGATATTCATCTACCTCAGGAACAATGGCATGAGTTCTGGAAATTGTTTGCCGAATATTTGAA
 230   D  E  P  T  F  D  Y  F  V  D  A  L  M  D  R  H  I  K  D  D  I  H  L  P  Q  E  Q  W  H  E  F  W  K  L  F  A  E  Y  L  N

2281 CGAAAAGAGTCACCAACATTTGACAGAAGCCGAGAAACATGCATGGGGTACAATAGGTGAGGACTTCGCGCATGAGGCCGATAAGCATGCAAAGGCCGAAAAAGACCATCATGAAGGAGA
 270   E  K  S  H  Q  H  L  T  E  A  E  K  H  A  W  G  T  I  G  E  D  F  A  H  E  A  D  K  H  A  K  A  E  K  D  H  H  E  G  E

2401 GCACAAAGAGGAACACCACTGAACCAACCCGTCGTCGTTCAACTTAAGCCTTCAGCTTAAGCTCGAGCTAAAGCCTCAGCTTGAGCTCAATCTTATGTCCTCAGGCCTAAACTTGAATTT
 310   H  K  E  E  H  H

2521 TAAAAGCATTTTGTTGAAGCAGTGCTAGCCAATCTCTTATCTTATCGGTGCTATTATCAATTTACTCTATGCCACCCCCCCCCCCCTCTCTCTGTTCTCTATTTGATATTCTGTTCTTTT

2641 AGTGCCAGATGTTAGTACCAGATGTTATTTTCTGCATAATTTTCTTCTCTTTACTTCGTTATTTTTTCGTTCTTCTATTTTTATGGCCAATTTTTGTGATGTCGAA
```

Figure 4.1 (Continued)

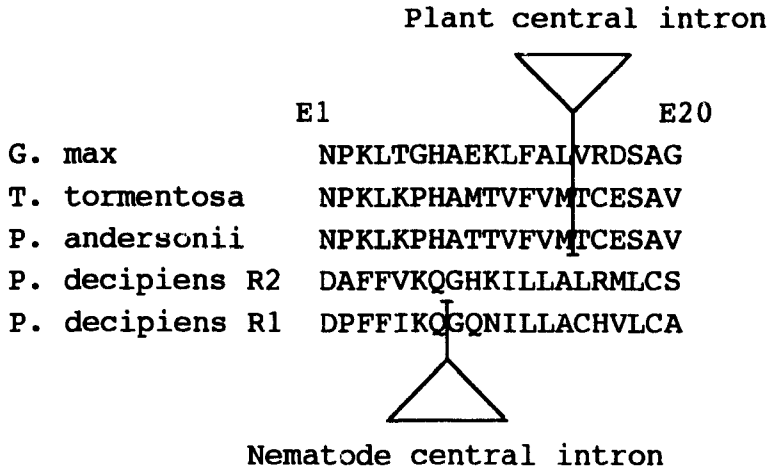


Figure 4.2. An alignment of the E domain of three plant (*Glycine max*, soybean, Dickerson and Geis, 1983, and two non-nodulating legumes *Parasponia andersonii* and *Trema tormentosa*, Bogusz et al., 1988) haemoglobins with the E domain sequence of both repeats (R1, R2) of *P. decipiens*. The location of the plant and nematode central introns with respect to the alignment are indicated.

the first and third plant introns are in identical positions to those of *P. decipiens* but the central intron position varies by seven amino acids (Fig 4.2, Jensen et al., 1981, Landsmann et al., 1986, Bogusz et al., 1988). The question of the equivalence of these two introns is difficult to ascertain as there is poor sequence similarity in this region and the position of the intron may have drifted in the 1.5 billion years since the divergence of plants and animals. There might easily have been an addition or deletion of seven codons in the appropriate gene to produce the intron locations present in both genes. Since the position of introns can vary as much as six codons within mammalian globin genes (Hardison, 1991), I suggest that the nematode haemoglobin gene contains an intron in a similar position to that found previously only in plants. As in plant haemoglobin genes, this central intron separates the heme-binding domain into two modules that bind the heme molecule from opposite sides, as Go hypothesized (1981). The position of the central intron predicted by Go (1981) is between amino acids E10 and E11, three amino acids away from the position of the *P. decipiens* central intror and four amino acids away from that of plants. The presence of this intron indicates that this nematode gene, as well as plant haemoglobin genes, may have retained the chromosomal structure of an ancestral haemoglobin gene. Another possibility is that the central intron was lost shortly after the divergence of plants and animals, but was re-acquired subsequently by either or both of these genes.

Genes coding for extracellular haemoglobins in invertebrates do not possess introns between the region coding for the hydrophobic leader sequence and the rest of the coding region (Jhiang et al., 1988). In contrast, the nematode gene contains an intron at amino acid position 21 of the unprocessed protein (Fig. 4.1), indicating that the nematode haemoglobin gene organization is similar to other eukaryotic extracellular protein coding genes in that it contains an intron following its secretory leader sequence (Breathnach and Chambon, 1981). This unique intron may have been recruited before or after the duplication event. The second and fourth introns are found in the equivalent positions to the introns found in all other animal haemoglobin genes.

The amino acid sequence for repeat one ends after the fifth exon. The protein sequence then repeats at the start of exon six. Interestingly, the coding sequence for this second repeat only contains one intron, at amino acid position 181, corresponding to the same position as intron two. There is no sequence similarity between introns two and six (data not shown). However, any similarity between these regions following the duplication event most probably was lost due to sequence divergence over time. This is also the probable reason why intron six differs in size from intron two. Using the hypothesis that a one percent protein sequence difference occurs every five million years in globins (Dickerson and Geis, 1983) and given that there is 51% sequence identity between the derived amino acid sequences of repeats one

and two (Dixon et al., 1991), the gene duplication probably occurred approximately 245 Myr ago. Both the central (or "plant" intron) and intron four have been deleted from the second repeat. The loss of both of these introns specifically from the 3' end probably reflects an unequal cross-over event with a reverse transcribed copy of a partially spliced mRNA. I am uncertain whether this cross-over event occurred at the time of, or subsequent to, the duplication event. Such a loss of introns has been observed in many genes, including globins (Nishioka et al., 1980, Vanin et al., 1980), and often results in processed pseudogenes that are usually incorporated into other regions of the chromosome (Lewin, 1983, Vanin, 1984).

Southern analysis of genomic DNA (Fig. 4.3) revealed that this gene is present in a single copy in the genome of *P. decipiens*. The 846 bp band is diagnostic for this gene. In addition, the bands observed when these genomic clones are Eco RI digested and probed are identical to those shown in Fig. 4.3. This single copy gene is expressed in the larval worms and produces a 1.8 kb mRNA (Dixon et al., 1991). Any putative mechanism of duplication must address the fact that this duplicated gene has remained single copy.

This haemoglobin gene duplication event may differ from those that produced multimeric haemoglobin genes in other organisms. A two-repeat haemoglobin from the clam, *Barbatia reeveana* (Naito et al., 1991) and a multiple-repeat haemoglobin

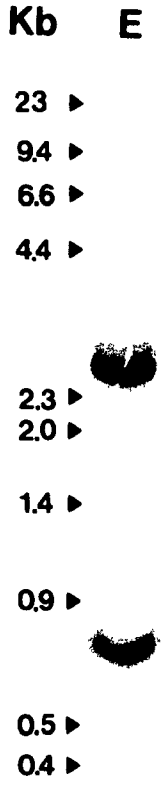


Figure 4.3. Southern blot of *P. decipiens* genomic DNA digested with Eco RI (E) and probed with labelled cDNA clones coding for *P. decipiens* haemoglobin (Dixon et al., 1991). Arrows mark the position of molecular weight standards.

from the brine shrimp *Artemia salina* (Manning et al., 1990) have recently been reported. The nematode globin is unusual because the postulated duplication appears to have resulted in a direct head-to-tail arrangement with the original genomic copy. An alignment of the protein sequences from both repeats shows that repeat two is 13 amino acids longer than repeat one. It is possible that during duplication an unequal cross-over occurred resulting in a truncation of 13 amino acids. This duplication event therefore maintained the duplicated copy at "full-length", but truncated the original copy. How can one explain this type of duplication? I suggest that an unequal cross-over event occurred involving the coding sequence in exon 5, the last exon of the first repeat. Examining the nucleotide sequence near the inferred cross-over point, I have found eleven of fourteen nucleotides in the beginning of repeat one which match with sequences in both the beginning and end of repeat two (Fig. 4.4). These similar coding regions may have been involved in a cross-over event, perhaps with a a partially processed cDNA transcript, resulting in the truncation. If this is the case how did the intron separating the two repeats occur? Presently the only possible explanation is that it was recruited later. The mechanism of duplication suggested for the clam haemoglobin did not involve an exon truncation (Naito et al., 1991). Instead, a cross-over between the intron preceding the coding sequence and the 3' non-coding region of the clam haemoglobin gene occurred, as indicated by sequence similarity in

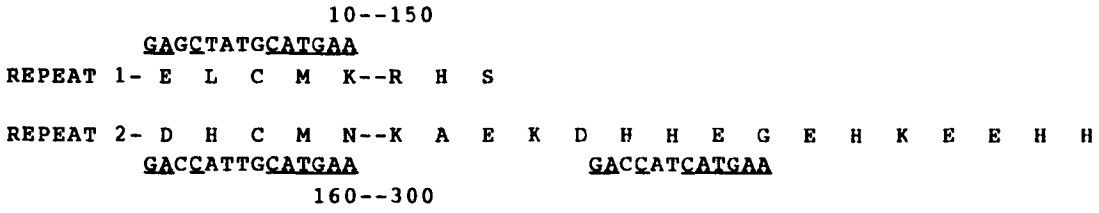


Figure 4.4. The homologous regions of the 5' end of repeat 1 and the 5' and 3' ends of repeat 2 proposed to have been involved in the duplication of the *P. decipiens* haemoglobin gene. Nucleotide sequences are indicated above and below the relevant amino acid sequences. Similar nucleotide sequences are underlined. Numbers indicate amino acid position numbers for the mature protein.

these regions (Naito et al., 1991). The two repeats of the clam haemoglobin are 78% identical, indicating a more recent duplication event than that which occurred in the nematode haemoglobin. It is possible that in the nematode haemoglobin gene a cross-over event occurred between a sequence in intron one and a sequence near the 3' coding end of the original gene (exon 5). Unfortunately, time has eroded any sequence similarity between the present introns one and five to support this theory. This theory may be further supported by the similar size of exons two and six. However, this mechanism of intron/exon cross-over must have also created the correct splicing consensus sequence and also maintained the subsequent coding region in frame. There is also a possibility that the duplication of this gene occurred by recombinational events involving genomic DNA. Efstratiadis et al. (1980) suggested a model for slipped mispairing of small (2-8 bp) regions of sequence similarity during DNA replication. If the gene was duplicated and the duplicate copy was inserted downstream from the original copy the regions of sequence similarity shown in Fig. 4.4 could facilitate the head to tail joining of the two copies. This mechanism would explain the truncation of repeat 1, but would not explain the presence of intron 5, which may have arisen later. A fourth possibility is that the gene duplicated during a mispaired gene conversion event such as that outlined by Slightom et al. (1980), that is capable of producing genes that are a fusion of two original genes. This mechanism also leaves the origin of intron 5

unexplained. The region of sequence similarity noted in Fig. 4.4 may have mediated such a gene conversion event between two chromosomes containing single or duplicated genes resulting in a heteroduplex gene which then became the predominant gene in the population. I am currently trying to discriminate between all the proposed mechanisms of duplication. An interesting observation is that the cDNA for the body fluid haemoglobin of the nematode *Trichostrongylus colubriformis* is only 605 bp long and the resultant mature protein is only 18 kDa in size (Frenkel et al., 1992). This indicates that the *T. colubriformis* body fluid haemoglobin gene is not duplicated although it has 32.2% sequence identity (75% conservative substitutions) with repeat 1 of *P. decipiens* haemoglobin and 29.9% sequence identity (69% conservative substitutions) with repeat two at the protein level (data not shown). Therefore, unlike the duplicate haemoglobin proteins found in the nematodes *P. decipiens* (Dixon et al., 1991) and *Ascaris* (L. Moens, personal communication) other nematodes may not contain a duplicated body fluid haemoglobin.

The finding of this central intron in nematodes suggests that this intron may have existed in animals after the divergence of plants and animals 1,500 Myr ago. This is consistent with the nematode globin gene retaining the structure of the ancestral globin gene. It has been suggested that plant leghaemoglobin has been acquired in evolution by lateral or horizontal gene transfer through a viral vector (Jeffreys, 1981). However, the fact that this

haemoglobin is widespread in plants would seem to weaken this hypothesis, and it is presently thought that the gene evolved by vertical descent from a common ancestor with animal globin genes (Go, 1981). Most likely this putative ancestral haemoglobin gene contained the additional central intron that was retained or reacquired by either or both plants and nematodes (this manuscript). However, many plants are infected with nematodes (Zuckerman et al., 1971) and the similar chromosomal organization of this gene in plants and this nematode could be used to support the theory of lateral transfer of this gene. I consider this hypothesis unlikely, however, due to the low sequence similarity between the nematode globin and plant globins (results not shown). Phylogenetic analysis has indicated that the nematode haemoglobin is unusual and occupies its own branch of an evolutionary globin tree. The *P. decipiens* haemoglobin does not group with any of the other non-nematode invertebrate haemoglobins (data not shown). The lack of the central intron in other primitive invertebrates may suggest that the loss of this intron occurred just after the evolutionary divergence of nematode and other animal haemoglobins.

CHAPTER 5: PRODUCTION OF RECOMBINANT HAEMOGLOBIN IN A PROKARYOTIC EXPRESSION SYSTEM

Introduction

One of the main goals of the project described in this thesis was to produce vaccines that could be used to eliminate the sealworm from the Atlantic Grey seal population, and thus eliminate the worm from the Atlantic cod population. Vaccination of the definitive host using proteins derived from infectious worms has proved successful in protecting animals for several related species (Connan, 1965, Rothwell and Love, 1974, Stromberg and Soulsby, 1977, Urban and Romanowski, 1985, O'Donnell et al., 1989, and Culpepper et al., 1992). In one case the protein used was the equivalent pseudocoelomic haemoglobin of the related nematode *T. colubriformis* (Frenkel et al., 1992). Preliminary studies using haemoglobin from *P. decipiens* indicate that vaccination with this protein may provide protection to the seal hosts (Pohajdak, personal communication). In this case purifying large quantities of haemoglobin protein from worm pseudocoelomic fluid would be inefficient. A more efficient method would be to produce the protein in an expression system that allowed easy, large-scale purification.

Another reason for producing this protein in a prokaryotic expression system is that the purified protein could then be used to answer several outstanding questions regarding this molecule.

Since this gene has undergone a tandem duplication event (Dixon et al., 1991), some of these questions could best be answered by expressing each repeat of the "standard" haemoglobin polypeptide independently. First, since this duplicated molecule contains two heme-binding domains does it in fact bind two heme groups? Wittenberg et al. (1965) and Terwilliger (1980) reported that *Ascaris* haemoglobin monomers bind only one heme group, while the results of Darawshe et al. (1987) showed that they bind two. Also, it has been reported that the *Ascaris* haemoglobin molecule binds squalene and NADPH (Sherman et al., 1992). Perhaps one of the repeats binds heme and the other binds these components. Separation and independent testing of both repeats for heme, squalene and NADPH binding would clarify these findings. Second, the monomeric form of the very similar *Ascaris* haemoglobin appears to form large multimers *in vivo* when viewed with electron micrography (Darawshe and Daniel, 1991). The exact form of this multimer has been postulated (DeBaere et al., 1992) but cannot be confirmed using x-ray crystallography as the crystal form of this protein is extremely difficult to produce. This may be due to associated glycosyl groups interfering with the crystallization process. Production of the polypeptide monomer in a prokaryotic system would preclude glycosylation and may aid in producing the crystals necessary for this experiment.

Materials and Methods

Vector Construction

The pRSET prokaryotic expression vectors as described by Schoepfer (1993) were obtained from Invitrogen (San Diego, CA). Each member of this vector set consists of the same basic plasmid with a varied multiple cloning site that allows insertion of a fragment in one of the three possible reading frames. The recombinant protein is under the control of the T7 promoter and is terminated by the T7 terminator. The vector encodes a 31-amino acid leader for the recombinant protein that contains a poly-histidine tract, used in isolating the recombinant protein on nickel columns (Jankneckt et al., 1991, see below), and an enterokinse cleavage site that can be used to remove the vector-encoded amino acids.

The schematic diagram in Figure 5.1 shows the relevant restriction enzyme sites of the *P. decipiens* haemoglobin cDNA clone number 18 (Dixon et al., 1991). In order to express the full-length protein the 957 bp *Xho* I fragment from bp 94 to 1051 was ligated into the *Xho* I restriction site of pRSET A (pFL). Two strategies were pursued to produce recombinant proteins that separated the first and second repeats. The 384 bp *Xho* I-*Eco* RI fragment from the first repeat was ligated into *Xho* I and *Eco* RI cut pRSET A (p1X/E) and the 412 bp *Xho* I-*Sal* I fragment was ligated into *Xho* I cut pRSET A (p1X/S) to produce vectors containing only the sequence

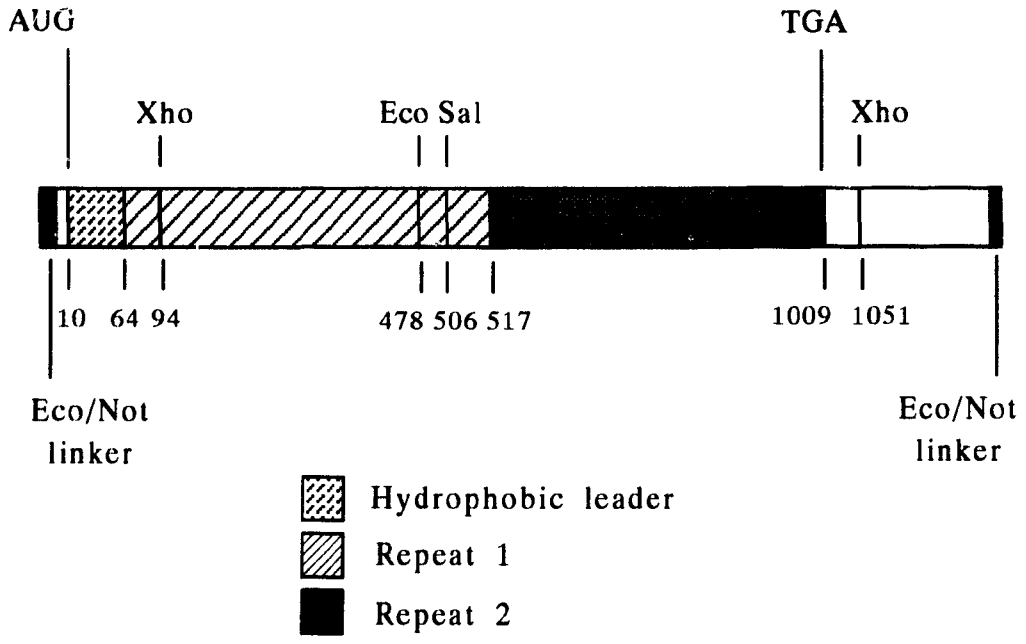


Figure 5.1. A schematic diagram of the *P. decipiens* haemoglobin cDNA clone used to produce the expression plasmids. AUG; start codon. TGA; stop codon. Xho; *Xho* I restriction enzyme site. Eco; *Eco* RI restriction enzyme site. Sal; *Sal* I restriction enzyme site. Eco/Not linker; the 13 bp linker containing *Eco* RI and *Not* I restriction enzyme sites used to insert the cDNA into the phage vector during library construction (see Chapter 2, materials and methods). The numbers indicate distance in base pairs from the 5' end of the clone.

encoding the first repeat. Vectors consisting of only the sequence encoding the second repeat were produced in the following way: the 906 bp fragment, from the *Eco* RI site at bp 478 to the *Eco* RI site in the Eco/Not linker, was ligated into the *Eco* RI site of pRSET B (p2E) and the 545 bp fragment, from the *Sal* I site at bp 506 to the *Xho* I site at bp 1051, was ligated into *Xho* I cut pRSET C (p2X).

The expression constructs described above were transformed into the *Escherichia coli* strain BL21(DE3) plysS (Novagen, Madison, WI) using the technique described on pages 1.84 to 1.84 of Sambrook et al. (1989). Cultures of these transformants were induced with 0.8 mM IPTG for 18 hours at 27°C.

Purification of Recombinant Protein

Recombinant proteins were isolated using the Xpress System (Invitrogen, San Diego, CA) that employs immobilized metal ion affinity chromatography (IMAC, Hemdan et al., 1989). The methods described in the Xpress System manual were followed for both denaturing and native lysis procedures.

Purification of Control Haemoglobin

The haemoglobin was purified from body fluid using the method of Darawshe et al. (1987). Pseudocoelomic fluid was bound to DEAE-Sephadex A50 in a column and eluted using 250 mM NaCl/ 25 mM Tris.HCl. The eluate was then passed through a Sephadex

G-200 column. The coloured fractions were pooled and concentrated in an Amicon centriprep 10 concentrator.

Recombinant Protein Analysis

The molecular weight and relative amount of recombinant protein produced by each clone was determined by separating protein from 10 ml of culture on a denaturing polyacrylamide gel (Laemmli, 1970). The samples, along with 0.5 µg of the purified control haemoglobin, were loaded in duplicate on each half of the gel. One half was stained with Coomassie brilliant blue (Sigma), while the other half was used to produce a Western Blot (see below). The relative amount of recombinant haemoglobin produced by each culture was determined by measuring the relative intensity of the recombinant protein. This value was obtained by scanning a photograph of the gel with an Hewlitt-Packard ScanJet IIc scanner (Hewlitt-Packard, Inc, Boisie, Idaho). The protein bands from the digitized image were analyzed using Scan Analysis software (Biosoft, Ferguson, MO) on a Apple Macintosh IIvx computer (Apple Computer, Inc, Cupertino, CA).

Western Blotting

The protein was transferred from the gel onto a nitrocellulose filter by electroblotting. This filter was screened using a polyclonal antiserum obtained by immunizing juvenile grey seals with purified *P. decipiens* pseudocoelomic fluid haemoglobin.

Antibodies were detected using alkaline phosphatase-conjugated protein A (Sigma) followed by the substrates NBT and BCIP (BRL). Filters were blocked, washed and probed by following the standard Western blot procedures (Towbin et al., 1979). The relative intensity of the immunoreactive bands was measured by scanning as described above.

Results

Recombinant proteins were isolated from cultures stimulated with 0.8 mM IPTG for 18 h at 27°C as this gave a much better yield of recombinant protein than cultures stimulated with an equivalent concentration of IPTG for 4 hours at 37°C (Figure 5.2). The recombinant proteins produced by each construct are shown in Figure 5.3. These samples were obtained using the denaturing lysis procedure. The plasmid, pFL, which contains the 957 bp *Xho* I fragment having all but 30 bp of the sequence encoding the mature protein, produced a recombinant protein of approximately 37 kDa, as shown in Figure 5.3 lane B. Similarly the size of the recombinant proteins produced by the other plasmid constructs are; p2X, 22 kDa, p2E, 25 kDa, p1X/S, 22 kDa and p1X/E, 19 kDa (Fig. 5.3 lanes C to F). Analysis of the intensities of the bands relative to the 0.5 µg purified haemoglobin standard is presented in Table 5.1.

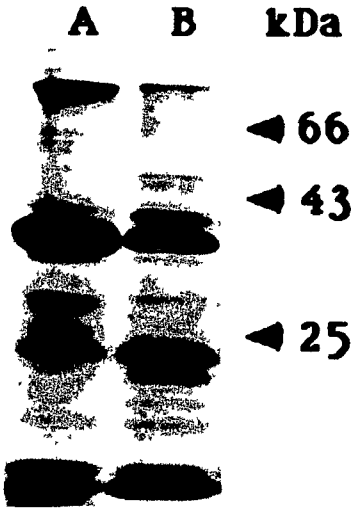


Figure 5.2. SDS-PAGE gel showing recombinant *P. decipiens* haemoglobin proteins produced by the plasmid pFL at different growth temperatures. A: recombinant protein isolated from a culture grown 18 h at 27°C. B: recombinant protein isolated from a culture grown 4 h at 37°C.

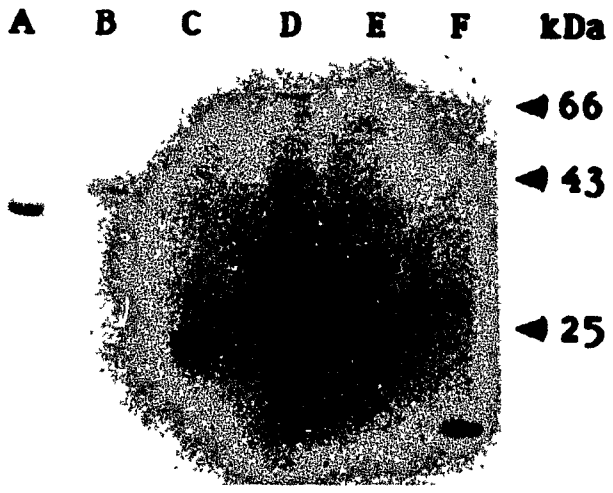


Figure 5.3. SDS-PAGE gel showing recombinant *P. decipiens* haemoglobin proteins. A: control lane (0.5 µg of purified haemoglobin), B to F: Protein purified by IMAC from 10 ml of culture from bacteria transformed with the following plasmids: B; pFL, C; p2X, D; p2E, E; p1X/S, F; p1X/E. See text for details. Numbers on the right indicate protein molecular weight in KiloDaltons.

The Western blot shown in Figure 5.4 is a result of probing a nitrocellulose membrane blotted with the second half of the duplicate loaded gel with polyclonal sera obtained from Grey seals immunized with purified *P. decipiens* haemoglobin. All of the major bands seen in the Coomassie-stained gel showed very strong immunoreactivity with the sera, indicating that they are indeed *P. decipiens* haemoglobin recombinant proteins. In addition, there was a high molecular weight band in the p2E lane (Fig. 5.4, lane D) that also showed immunoreactivity. The relative immunoreactivity of each of these bands, as compared to the 0.5 μg control band is shown in Table 5.1.

The recombinant proteins obtained when the samples were extracted using the native lysis method are shown in Figure 5.5. The samples loaded on this gel are not from an equivalent amount of culture as those loaded on the gel in Figure 5.3. Bands of equivalent size and amount of recombinant protein to those obtained using denaturing lysis were observed for the products of the plasmids p1X/S and p2E (Figure 5.5, lanes A and C, respectively). The product of pFL showed some degradation, as did the product of p2X, although some intact product was observed in each case (Figure 5.5, lanes E and D, respectively). Using native lysis p1X/E yielded less recombinant protein than it did using the denaturing lysis, but no degradation of the recombinant protein was observed (Figure 5.5, lane B).

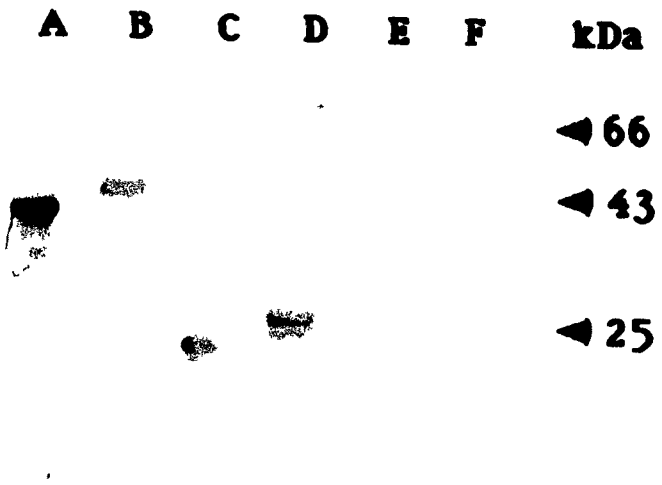


Figure 5.4. Western blot showing recombinant haemoglobin proteins probed with Grey seal polyclonal sera to *P. decipiens* haemoglobin. A: control lane (0.5 μg of purified haemoglobin), B to F: Protein purified by IMAC from 10 ml of culture from bacteria transformed with the following plasmids: B; pFL, C; p2X, D; p2E, E; p1X/S, F; p1X/E. See text for details. Numbers on the right indicate protein molecular weight in KiloDaltons.

Table 5.1. Intensities of recombinant protein bands from Coomassie-stained gel and Western blot relative to 0.5 µg haemoglobin control.

Plasmid	Intensity (% control)	
	Coomassie Gel	Western blot
pFL	90	55
p2X	175	44
p2E	182	63.5
p1X/S	179	30
p1X/E	134	37

Discussion

The production of recombinant protein including portions of the pseudocoelomic fluid haemoglobin from the nematode *P. decipiens* would facilitate both vaccine production and further research into the nature of this molecule. I have produced recombinant haemoglobin using a prokaryotic expression vector and restriction fragments from a previously isolated cDNA clone (Dixon et al., 1991).

The optimal temperature for production of recombinant protein was 27°C (Fig 5.2). It has been previously noted that larger yields of recombinant protein can be obtained by growing the

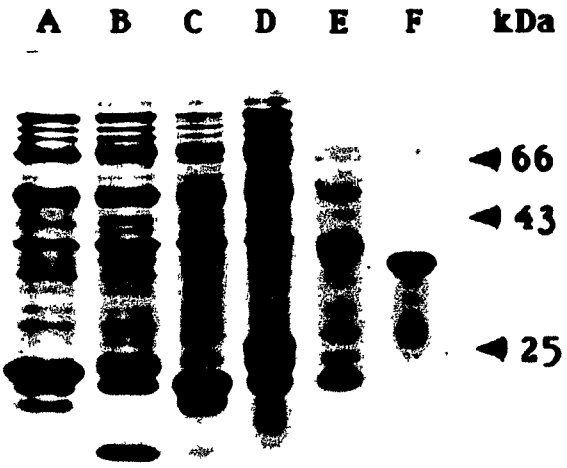


Figure 5.5. SDS-PAGE gel showing recombinant *P. decipiens* haemoglobin proteins isolated using native lysis procedure. A to E: Protein purified by IMAC from 10 ml of culture from bacteria transformed with the following plasmids: A; p1X/S, B; p1X/E, C; p2X, D; p2E, E; pFL. F: control lane (0.5 μg of purified haemoglobin). See text for details. Numbers on the right indicate protein molecular weight in KiloDaltons.

bacteria at 25°C-30°C, as opposed to the optimal growth temperature of 37°C (Schein, 1989).

As seen in Fig. 5.3 lane B, the plasmid that should express the mature protein, pFL, produces a major band at approximately 37 kDa, slightly larger than the purified haemoglobin control, since the leader sequence encoded by the vector has not been cleaved off. Similarly the expected sizes of the other plasmid constructs based on predicted amino acid sequence are; p2X, 22 kDa, p2E, 23.5 kDa, p1X/S, 21.2 kDa and p1X/E, 19 kDa. Fig. 5.3 lanes C to F show that the major bands produced these plasmids; p2X, 22 kDa, p2E, 25 kDa, p1X/S, 22 kDa and p1X/E, 19 kDa (Fig. 5.3 lanes C to F) are all very close to these expected sizes.

In order to test whether the major bands that appear at the expected sizes are genuine recombinant haemoglobin, a Western blot of the gel was probed with polyclonal sera to purified haemoglobin raised in Grey seals. As seen in Fig. 5.4 all of the major bands were strongly recognized by the sera. It is interesting to note that the recombinant proteins containing the first repeat showed less reactivity (Fig 5.4 lanes E and F) than those containing the second repeat or full-length protein. In order to quantify these apparent differences in reactivity, the Western and a photograph of the gel were scanned and these scanned images were analyzed using the densitometry program Scan Analysis (Biosoft, Ferguson, MO). As seen in Table 5.1, the two recombinant proteins containing the first repeat show lower reactivity relative to the control lane

than those containing the second repeat. Comparing the recombinant protein produced by p1X/S with that produced by p2X, an equivalent amount of protein shows only 70% of the reactivity relative to control (Table 5.1). This may be due to the fact that the recombinant protein does not show conformational epitopes on a Western blot, or it may reflect the fact that the major epitope recognized by the polyclonal sera is within the second repeat of the natural haemoglobin.

Some larger, minor bands were produced by the plasmids encoding the second repeat of the haemoglobin (Fig 5.3, lanes C and D). These bands are in the size range of dimers and trimers of the recombinant proteins. This is interesting since a histidine-rich region at the terminus of the protein is thought to be responsible for assembling the large multimers (DeBaere et al., 1992) seen *in vivo* (Darawshe and Daniel, 1991). These bands are not seen in the lanes containing the recombinant first repeats of the haemoglobin. The possibility that these larger minor bands represent multimers is suggested by their cross reactivity with the seal polyclonal antisera used to confirm the major bands (Fig 5.4, lane D). These observations needs confirmation by further experimentation, however, before firm conclusions can be drawn.

The recombinant proteins obtained from each culture using the native lysis procedure were identical to those obtained using the denaturing lysis procedure, although in some cases (pFL and p2X) there was some degradation of the recombinant protein (Fig

5.5, lanes D and E) and there was a reduced yield of the product of p1X/E (Fig 5.5, lane B). The extraction of proteins using the native lysis procedure is more desirable for the heme- and squalene-binding studies described in the Introduction, as well as for vaccine production since antibodies may react strongly to the secondary structure of proteins.

The production of these recombinant proteins will allow future experiments into the nature of this unique haemoglobin. The full-length recombinant protein can be used to produce crystals for X-ray crystallography, which was not possible with purified protein. The recombinant proteins composed of the first and second repeats can be used to investigate heme-binding by each repeat, as well as the possibility that one of the repeats will bind NADPH or squalene (Sherman et al., 1992).

CHAPTER 6: GENERAL DISCUSSION AND CONCLUSIONS

The discussion section in each chapter has dealt with the conclusions from each separate body of work. The goal of characterizing this antigen and producing large amounts of this protein has been reached. The following section will tie together some conclusions from the gene evolution aspect of this project into a novel theory.

Did the Ancestral Globin Gene of Plants and Animals Contain Only Two Introns?

Prior to 1981, the only known globin gene structures were obtained from vertebrates. The structure is very regular, containing two introns of varying size in highly conserved positions in the regions encoding the B and G domains (Blanchetot et al., 1983, Hardison, 1991). In 1981 Go proposed that a third intron should be present in the middle exon of globin genes in the region encoding the E domain (Go, 1981). Go based this proposal on the results of protein structure analysis and Gilbert's "genes-in-pieces" hypothesis (Gilbert, 1981, 1985, 1987) since this intron would divide the exon encoding the heme-binding domain into two smaller domains that would each bind the heme group from opposite sides. This proposal was confirmed almost immediately when Jensen et al. (1981) reported a central intron in the gene encoding the globin of the root nodules of the legume soybean

(*Glycine max*). This observation was further confirmed by the sequencing of two more globin genes from another nodulating plant(*Parasponia andersonii*), as well as a non-nodulating plant (*Trema tormentosa,* Landsmann et al., 1986, Bogusz et al., 1988). Blake (1981) postulated that the four-exon, three-intron globin gene structure was the ancestral form of the gene. Since then several investigators have concluded that the common ancestor of all animals lost the central intron following the divergence of plants and animals one to one and a half billion years ago (Lewin, 1984, Appleby et al., 1990, Hardison, 1991).

Central Introns of Nematode Genes: New Evidence

The sequence and structure of the genes encoding an extracellular haemoglobin in three nematodes, *Caenorhabditis elegans* (Sulston et al., 1992, Kloek et al., 1993), *Ascaris suum* (Sherman et al., 1992) and *Pseudoterranova decipiens* (Dixon et al., 1991, 1992) cast doubt on the theory that the ancestral globin gene contained three introns. The *C. elegans* globin gene contains only two exons and one introns. However the first exon starts with a perfect 3' splice sequence, so a section encoding hydrophobic leader may be trans-spliced onto the mRNA (reviewed in Sharp, 1987, Bonen, 1993), or there may be an upstream exon equivalent to those of *P. decipiens* and *A. suum* . The sole intron of the *C. elegans* globin gene is in the region of the central intron (Fig. 6.1).

The *A. suum* and *P. decipiens* globin genes are much more complex in structure. Both of these nematodes are members of the Order Ascaridida, and both have undergone a unique head-to-tail gene duplication event and thus produce a polypeptide twice the length of other globin proteins, containing two heme-binding pockets (Dixon et al., 1991, 1992, Sherman et al., 1992). The *A. suum* gene contains nine exons and eight introns, while the *P. decipiens* gene contains seven exons and six introns. These introns are in identical positions and phases, the only difference in gene structure is that the *P. decipiens* globin gene has lost the introns equivalent to the central intron and the G domain intron from the second repeat of the gene, presumably in a cross-over with partially processed cDNA of the gene following the divergence of *A. suum* and *P. decipiens*. For both genes the first intron separates the exon containing the hydrophobic leader sequence from the rest of the gene, while the second and fourth introns, aa well as the sixth and eighth introns in the *A. suum* globin gene, are in the regions encoding the B and G domains (i.e. equivalent to the "vertebrate" introns). The third intron (and seventh of *A. suum*) is near the position of the central intron of plants.

While both the *C. elegans* and the Ascarid globin genes contain a central intron in the region near the plant central intron and the position predicted by Go (1981), these introns are not in conserved positions when the proteins are aligned (Fig. 6.1). The alignment presented in Fig. 6.1 presents only the relevant E domain

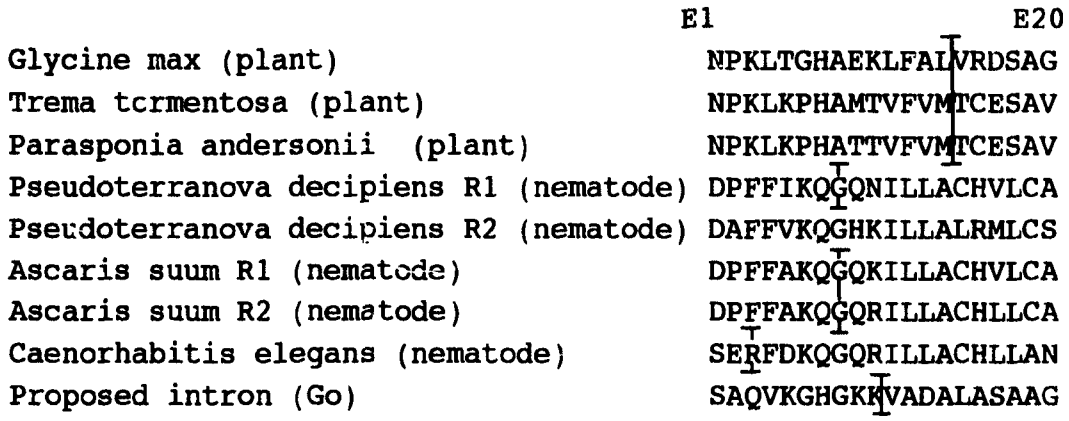


Figure 6.1. An alignment of the E domain of several haemoglobins, showing the position of the central intron in globin genes from three plants and three nematodes. The position of Go's proposed central intron was based on the sequence of Balb/c murine a globin. Sequences were obtained from references stated in the text. A bar through an amino acid indicates that the intron divides the codon for that amino acid residue.

of the haemoglobins, but it is derived from several previously published alignments, notably that of Dickerson and Geis (1983). In addition to this difference in the introns, the central intron of all plant globin genes falls precisely between two codons (phase 0 intron), while the central intron of the Ascarid globin genes separate the first base of a codon from the remaining two (phase 1 intron) and the central intron of the *C. elegans* globin gene divides the first two bases of a codon from the remaining base (phase 2 intron, Table 6.1). In contrast, when these haemoglobins are aligned together with typical vertebrate haemoglobins and both position and phase of the B and G domain introns are compared, these introns are highly conserved (Table 6.1). A protein alignment also shows that these "vertebrate" introns are in identical positions in all haemoglobins (not shown). This conservation over such a great evolutionary distance argues that these two introns were derived from introns present in a common ancestral globin gene. The large variation in both position and phase of the central introns of plant and of both nematode globin genes, however, argues that these introns were gained independently, following the divergence of these groups. This is consistent with previous examples of introns which have been independently gained in a similar region of a gene but not in precisely the same position (reviewed by Rogers, 1990). In the case of the Ascarid and *C. elegans* globin genes protein sequence identity suggests that this divergence occurred approximately 500 million years ago (483±56 million

years; 1% sequence difference equals 5 million years; Dickerson and Geis, 1983; corrected for multiple substitutions using the method of Dickerson, 1971). Thus the central intron in these nematodes is a relatively recent development.

Table 6.1. A comparison of the phases of the introns of globin genes. Phase 0 means that the intron separates two codons precisely, while phase 1 means that the codon is split between the first and second bases, while phase 2 indicates that the codon is split between the second and third bases. R1, repeat 1; R2, repeat 2. Mouse α and β gene sequences used are from Nishioka and Leder (1979) and Konkel et al. (1978) respectively. The rest of the sequences were obtained from references quoted in the text.

Organism	Intron Position			
	Leader	B Domain	Central	G Domain
Glycine max (plant)	-	2	0	0
Trema tormentosa (plant)	-	2	0	0
Parasponia andersonii (plant)	-	2	0	0
Mus musculus α (vertebrate)	-	2	-	0
Mus musculus β (vertebrate)	-	2	-	0
Pseudoterranova decipiens R1 (nematode)	1	2	1	0
Pseudoterranova decipiens R2 (nematode)	-	2	-	-
Ascaris suum R1 (nematode)	1	2	1	0
Ascaris suum R2 (nematode)	-	2	1	0
Caenorhabitis elegans (nematode)	-	-	2	-

A Proposed New Phylogeny for Globin Genes

Using the late gain of the central introns of nematodes and plants I have constructed a new phylogeny for globin genes (Fig. 6.2). The ancestral eukaryotic globin gene contained three exons and two introns, which have been maintained in both position and phase in all modern globin genes, except cases where they have been lost (e.g. *Chironomus, C. elegans*). All vertebrate globin genes (Blanchetot et al., 1983, Hardison, 1991) and several invertebrate globin genes (e.g. the worm, *Lumbricus terrestris,* Jhiang et al., 1988; the mollusc, *Anadara trapezia*, Titchen et al., 1991; and the clam *Barbatia reevena*, Naito et al., 1991) have maintained the ancestral three-exon, two-intron structure with little variation. A common ancestor of the haemoglobin-containing plants gained a central intron following the plant-animal divergence 1 to 1.5 billion years ago which has been conserved in position and phase in all modern plant globin genes. (Jensen et al., 1981, Landsmann et al., 1986, Bogusz et al., 1988). Another possibility is that the plant central intron was retained from that of the ancestral globin gene, while the animal line lost the ancestral central intron. Subsequently, several members of the animal lineage regained an intron in a similar, but not identical, position. However, this possibility presumes that the plant central intron is more likely to be derived from an ancestral one than the animal central introns. There is no evidence for this presumption. A four-exon, three-intron globin

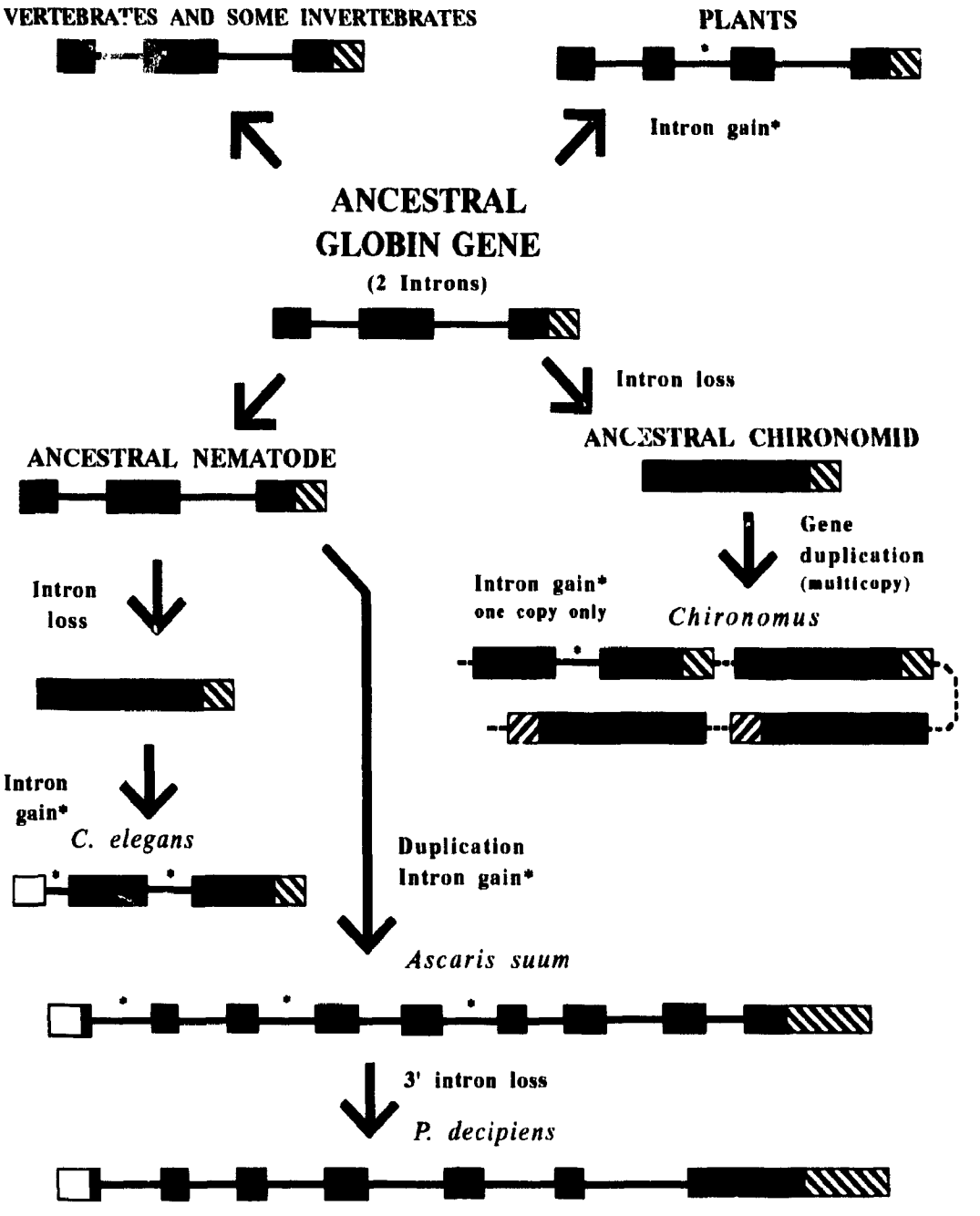


Figure 6.2. The proposed phylogeny of globin genes. Black boxes represent exons, while solid lines represent introns. Hatched boxes indicate 3' non-coding sequence and white boxes indicate regions encoding hydrophobic leader sequences. Dashed lines indicate DNA separating multiple copies of genes.

gene from *Chlamydomonas eugametos* has been recently reported (Guertin et al., 1992, Genbank accession number X65870). The central intron from this globin gene does not align with the location of the other plant central introns. This supports the theory that all central introns were gained following the plant-animal divergence.

The ancestor of the insect *Chironomus thummi thummi* lost both the B and G domain introns following its divergence from its last common ancestor with the nematodes approximately 855±71 million years ago (based on protein sequence identity). Following this intron loss the *Chironomus* globin gene underwent several duplication events to produce a multicopy gene family (Lewin, 1984, Guertin et al., 1992). Finally, the *Chironomus* IIB gene alone gained a single intron, but not in the position of the conserved vertebrate introns but one codon 3' of Go's (1981) predicted central intron (G. Bergtrom, personal communication, Kao and Bergtrom, 1992).

The Phylogeny of Nematodes and Their Globin Genes

The nematode globin genes appear to have diverged a great deal as a group from the ancestral form. The precursor to the *C. elegans* globin gene lost both vertebrate introns in a fashion similar to the *Chironomus* globin genes, but later re-gained an intron in the region of the other central introns. The sequence of haemoglobins of two other nematodes, *Trichostrongylus colubriformis* (Frenkel et al., 1992) and *Nippostrongylus brasiliensis* are known, but their

gene structures remain undetermined. *N. brasiliensis* contains two forms of haemoglobin, a bodywall form (Genbank accession number LO20895) and cuticular form (Genbank accession number LO20896). The lack of knowledge regarding their gene structures prevents us from speculating about their intron/exon structure, but their protein sequences can be used to help date events in the evolution of the nematode globin gene. Similar to the C. elegans globin gene, all three of these globin genes are unduplicated, based on cDNA sequence. The Strongylids diverged from *C. elegans* 385±16 million years ago, and thus share a more recent common ancestor with *C. elegans* than the Ascarids. The structures of the Strongylid globin genes would reveal whether the loss of all introns, then the regain of the central one occurred before or after this divergence 385 million years ago. This loss and regain is definitely subsequent to the divergence of the Ascarids from *C. elegans* and the Strongylids approximately 509±41 million years ago.

In a previous paper (Chapter 2) I concluded that the unduplicated precursor of the *P. decipiens* globin gene was similar in structure to an ancestral globin gene (Dixon et al., 1991). However, I have recently realized that the ascarid globin gene is probably one of the most divergent from the ancestral form of the globin gene. The gene originally had a typical three-exon, two-intron structure, but gained a central intron before it underwent a unique head-to-tail duplication event approximately 281±55

million years ago, based on protein sequence identity between the two repeats (Dixon et al., 1992, Sherman et al., 1992). An intron was gained to segregate the exon encoding the hydrophobic leader sequence in the common ancestor of the Ascarids and an intron that separated the two repeats was gained. The introns equivalent to the central intron and the G domain intron in the second repeat were lost in the *P. decipiens* gene following its divergence from *A. suum* roughly 250 million years ago.

It may be argued that the ancestral globin gene may have contained three introns in which the central intron was lost and subsequently re-gained in a similar but not identical position. We find the loss and re-gain of the central intron in several independent cases to be less likely than the phylogeny presented in Fig. 6.2. The possibility that the central intron was recruited once then moved by sliding is unlikely (Craik et al., 1982, 1983). It is difficult to imagine a process of nucleotide loss and gain on either side of a splice junction maintaining the degree of sequence similarity observed between the *C. elegans* and *P. decipiens* genes in this region (Fig. 6.1) while also altering the phase of the junction (Table 6.1).

In all the other genes containing central introns this intron separates an exon encoding the proximal heme binding histidine residue (position F8) from that encoding the distal heme binding histidine residue (E7, glutamine in *P. decipiens*). However the *C. elegans* central intron leaves them on the same exon. Go and

Gilbert's hypotheses suggest that the function of the central intron is to separate the two functional modules encoded by the exons. What then is the purpose of this intron? It cannot be a remnant of the assembly of large modern genes from ancestral minigenes (Gilbert, 1978, 1985). The failure of the *C. elegans* central intron to divide the two functional modules encoded by the exon may support the theories of Blake (1981) and Traut (1988) that introns do not always divide exons encoding functional modules of proteins but rather have some other function.

REFERENCES

Afonso, A. M. M., Santoro, M. M., and Neves, A. G. A. (1980) *Comp. Biochem. Physiol.* **67B**, 143-146.

Alt, F. W. and Baltimore, D. (1982) *Proc. Nat. Acad. Sci. USA* **79**, 4118-4122.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, P. J. (1990) *J. mol. Biol.* **215**, 403-410.

Antoine, M. and Niessing, J. (1984) *Nature* **310**, 795-798.

Antoine, M., Erbil, C., Munch, E., Schnell, S., and Niessing, J. (1987) *Gene* **56**, 41-51

Appleby, C.A., Dennis, E. S. and Peacock, W. J. (1990) *Aust. Syst. Bot.* **3**, 81-89.

Appleton, T. E. and Burt, M. D. B. (1991) *Can. J. Fish. Aquat. Sci.* **48**, 1800-1803.

Aviv, H., Valloch, V., Bastos, R. and Levy, S. (1976) *Cell* **8**, 495-503.

Banerji, J., Olson, L. and Schaffner, W. (1983) *Cell* **33**, 729-740.

Behringer, R. R., Hammer, R. E., Brinster, R. L., Palmiter, R. D. and Townes, T. M. (1987) *Proc. Nat. Acad. Sci. USA* **84**, 7056-7060.

Bellazi, T., Wagner, A.,Wieser, R., Schanz, M., Adam, G., Hartig, A. and Ruis, H. (1991) *EMBO J.* **10**, 585-592.

Blake, C. C. F. (1978) *Nature* **273**, 267.

Blake, C. C. F. (1979) *Nature* **277**, 598.

Blake, C. C. F. (1981) *Nature* **291**, 616.

Blanchetot, A., Wilson, V., Wood, D. and Jeffreys, A. J. (1983) *Nature* **301**, 732-734.

Bogusz, D., Appleby, C.A., Landsmann, J., Dennis, E. S., Trinick, M. J. and Peacock, W. J. (1988) *Nature* **331**, 178-180.

Bonner, A. G. and Laursen, R. A. (1977) *FEBS Lett.* **73**, 201-203.

Borst, P. and Grivell, L. A. (1981) *Nature* **289**, 439-440.

Bowen, W. D. (1990) In: Population biology of sealworm (*Pseudoterranova decipiens*) in relation to its intermediate and seal hosts. W.D. Bowen (ed.) *Can. Bull. Fish. Aquat. Sci.* **222**, vii-viii.

Bonen, L. (1993) *FASEB J.* **7**, 40-46

Brack, C. and Tonegawa, S. (1977) *Proc. Nat. Acad. Sci. USA* **74**, 5652-5656.

Brattey, J., Bishop, C. A. and Myers, R. A. (1990) In: Population biology of sealworm (*Pseudoterranova decipiens*) in relation to its intermediate and seal hosts. W.D. Bowen (ed.) *Can. Bull. Fish. Aquat. Sci.* **222**, 67-82.

Breathnach, R. and Chambon, P. A. (1981) *Ann. Rev. Biochem.* **50**, 349-383.

Breathnach, R., Mandel, J. L. and Chambon, P. (1977) *Nature* **270**, 314-319.

Brown, E. R., Coker, G. T. III and O'Malley, K. L. (1987) *Biochemistry* **26**, 5208-5212.

Burr, A. H., Eggleton, D. K., Patterson, R. K. and Leutscher-Hazelhoff, J. T. (1989) *Phytochem. Phytobiol.* **49**, 89-95.

Cavalier-Smith, T. (1978) *J. Cell. Sci.* **34**, 247-278.

Cavalier-Smith, T. (1985) *Nature* **315**, 283-284.

Cavalier-Smith, T. (1991) *Trends Genet.* **7**, 145-148.

Cech, T. R. (1986) *Cell* **44**, 207-210.

Cech, T. R., Zaug, A.J. and Grabowski, P. J. (1981) *Cell* **27**, 487-496.

Chandra, C. V. and Kahn, R. A. (1988) *J. Parasitol.* **74**, 1038-1040.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979)*Biochemistry* **18**, 5294-5299.

Chung, S. and Perry, R. P. (1989) *Mol. Cell. Biol.* **9**, 2075-2082.

Connan, R. (1965) *Parasitology* **55**, 10.

Cory, S., Jackson, J. and Adams, J. M. (1980) *Nature* **285**, 450-456.

Craik, C. S., Sprang, S., Fletterick, R. and Rutter, W. J. (1982) *Nature* **299**, 180-182.

Craik, C. S., Rutter, W. J. and Fletterick, R. (1983) *Science* **220**, 1125-1129.

Crick, F. H. C. (1979) *Science* **204**, 264-271.

Culpepper, J., Grieve, R. B., Friedman, L., Mika-Grieve, M., Frank, G. R. and Dale, B. (1992) *Mol. Biochem. Parasitol.* **54**, 51-62.

Darawshe, S. and Daniel, E. (1991) *Eur. J. Biochem.* **201**, 169-173.

Darawshe, S., Tsafadyah, Y. and Daniel, E. (1987) *Biochem. J.* **242**, 689-694.

Darnell, J.E. (1978) *Science* **202**, 1257-1260.

Davenport, H. E. (1949a) *Proc. R. Soc. B.* **136**, 255-270.

Davenport, H. E. (1949b) *Proc. R. Soc. B.* **163**, 206-214.

De Baere, I., Liu, L., Moens, L., Van Beeumen, J., Gielens, C., Richelle, J., Trotman, C., Finch, J., Gerstein, M. and Perutz, M. (1992) *Proc. Nat. Acad. Sci. USA* **89**, 4638-4642.

Dibb, N. J. and Newman, A. J. (1989) *EMBO J.* **8**, 2015-2021.

Dickerson, R.E. (1971) *J. Mol. Evol.* **1**, 26-45.

Dickerson, R.E. and Geis, I. (1983) Haemoglobin: Structure, Function, Evolution, and Pathology. Benjamin Cummings, Menlo Park, p97.

Dixon, B., Walker, B., Kimmins, W. and Pohajdak, B. (1991) *Proc. Nat. Acad. Sci. USA* **88**, 5655-5659.

Dixon, B., Walker, B., Kimmins, W. and Pohajdak, B. (1992) *J. Mol. Evol.* **35**, 131-136.

Doel, M. T., Houghton, M., Cook, E. A. and Carey, N. H. (1977) *Nucl. Acids Res.* **4**, 3701-3713.

Doolittle, W. F. (1978) *Nature* **272**, 581.

Doolittle, W. F. and Sapienza, C. (1980) *Nature* **284**, 601-603.

Dorit, R. L., Shoenbach, L. and Gilbert, W. (1990) *Science* **250**, 1377-1342.

Downie, S. R., Olmstead, R. G., Zurawski, G., Soltis, D. E., Soltis, P. S., Watson, J. C. and Palmer, J. D. (1991) *Evolution* **45**, 1245-1259.

Duester, G., Jornvall, H. and Hatfield, G. W. (1986) *Nucl. Acids Res.* **14**, 1931-1941.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. and Hood, L. (1980) *Cell* **20**, 313-319.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980) *Cell* **21**, 653-668.

Feinberg, A. P. and Vogelstein, B. (1983) *Anal Biochem.* **132**, 6-13.

Frenkel, M. J., Dopheide, T. A. A., Wagland, B. . and Ward, C. W. (1992) *Mol. Biochem. Parasitol.* **50**, 27-36.

Garlick, R. L. and Riggs, A. (1982) *J. Biol. Chem.* **257**, 9005-9015.

Garnier, J., Osguthorpe, D. J. and Robson, B. (1978) *J. Mol. Biol.* **120**, 97-120.

Gaykema, W. P. J., Hol, W. G. J., Vereijken, J. M., Soeter, N. M., Bak, H. J. and Beintema, J. J. (1984) *Nature* **309**, 23-29.

Gilbert, W. (1978) *Nature* **271**, 501.

Gilbert, W. (1985) *Science* **228**, 823-824.

Gilbert, W. (1987) *Cold Spring Harbour Symp. Quant. Biol.* **52**, 901-905.

Gilbert, W., Marchionni, M. and McKnight, G. (1986) *Cell* **46**, 151-154.

Gillies, S. D., Morrison, S. L., Oi, V. T. and Tonegawa, S. (1983) *Cell* **33**, 717-728.

Go, M. (1981) *Nature* **291**, 90-92.

Goodman, M., Pedwaydon, J., Czelusniak, J., Tomohiko, S., Gotoh, T., Moens, L., Shishikura, F., Walz, D. and Vinogradov, S. (1988) *J. Mol. Evol.* **27**, 236-249.

Gubler, U. and Hoffman, B. (1983) *Gene* **25**, 263-269.

Guertin, M., Couture, M., Hamel, D., St.-Pierre, B. and Savard, F. 1992) Fifth Int. Conf. on the Cell and Molecular Biology of *Chlamydomonas*. Pacific Grove, CA, USA, p.50.

Hall, D. H., Liu, Y. and Shub, D. A. (1989) *Nature* **340**, 574-576.

Hamada, K., Okazaki, T. , Shukuya, R. and Kaziro, K. (1963) *J. Biochem.* **53**, 484-488.

Haniu, M., McManus, M. E., Birkett, D. J., Lee, T.D. and Shively, J. E. (1989) *Biochemistry* **28**, 8639-8645.

Hardison, R. C. (1991) In: Evolution at the molecular level. Selander RK, Clark AG, Whittam TS (eds), Sinauer Associates, Sunderland, p272-290.

Hemdan, E. S., Zhao, Y.-J., Sulkowski, E. and Porath, J. (1989) *Proc. Nat. Acad. Sci. USA* **86**, 1811-1815.

Hickey, D. A., Benkel, B. F. and Abukashawa, S. M. (1989) *J. Theor. Biol.* **137**, 41-53.

Imamura, T., Baldwin, T. O. and Riggs, A. (1972) *J. Biol. Chem.* **247**, 2785-2797.

Jankneckt, R., De Matrynoff, G., Lou, J., Hipskind, R. A., Nordheim, A. and Stunnenberg, H. G. (1991) *Proc. Nat. Acad. Sci. USA* **88**, 8972-8976.

Jefferys, A. J. (1981) In: Genetic Engineering. R. Williamson (Ed.) Academic Press, London New York, Vol. **2**, pp1-48.

Jeffreys, A. J. and Flavell, R. A. (1977) *Cell* **12**, 1097-1108.

Jensen, E. O., Paludan, K., Hyldig-Neilsen, J. J., Jorgensen, P. and Marcker, K. A. (1981) *Nature* **291**, 677-679.

Jhiang, S. M., Garey, J.R. and Riggs, A. F. (1988) *Science* **240**, 334-336.

Jonsson, J. J., Foresman, M. D., Wilson, N. and McIvor, R. S. (1992) *Nucl. Acids Res.* **20**, 3191-3198.

Kao W.-K. and Bergtrom, G. (1992) Int. Cong. on Invert. Dioxygen Carriers. Abstract II-L4, Lunteren, The Netherlands.

Khosla, C. and Bailey, J. E. (1989) *J. Bacteriol.* **171**, 5995-6004.

Kimura, S., Hong, Y.-S., Kotani, T., Ohtaki, S. and Kikkawa, F. (1989) *Biochemistry* **28**, 4481-4489.

Klippenstein, G. L. (1980) *Amer. Zool.* **20**, 39-51.

Kloek, A. P., Sherman, D. R. and Goldberg, D. E. (1993) *Gene* **129**, 215-221.

Koch, W., Edwards, K. and Kossel, H. (1981) *Cell* **25**, 203-213.

Konkel, D. A., Tilghman, S. M. and Leder, P. (1978) *Cell* **15**, 1125-1132.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. and Cech, T. R. (1982) *Cell* **31**, 147-157.

Kuhsel, M. G., Stickland, R. and Palmer, J. D. (1990) *Science* **250**, 1570-1573.

Kyte, J. and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105-132.

Laemelli, U. K. (1970) *Nature* **227**, 680-685.

Landsmann, J., Dennis, E.S., Higgins, T. J. V., Appleby, C. A. , Krott, A. A. and Peacock, W. J. (1986) *Nature* **324**, 166-168.

Lee, M. H. and Smith, D. L. (1965) *Exp. Parasitol.* **16**, 392-424.

Lerch, K. (1978) *Proc. Nat. Acad. Sci. USA* **75**, 3635-3639.

Lewin, R. (1983) *Science* **219**,1052-1054.

Lewin, R. (1984) *Science* **226**, 328.

Lewis, S. A. and Cowan, N. J. (1986) *Mol. Cel. Biol.* **6**, 1529-1534.

Li, S. L. and Riggs, A. (1970) *J. Biol. Chem.* **245**, 6149-6169.

Li, W.-H. and Graur, D. (1991) Fundamentals of Molecular Evolution Sinauer Associates, Inc., Sunderland, Massachusetts. pp. 136-171.

Lipman, D. J. and Pearson, W. R. (1985) *Science* **227**, 1435-1441.

Lowenhaupt, K. and Lingrel, J. B. (1978) *Cell* **14**, 337-344.

Malouf, A. H. [Chair](1986) Seals and sealing in Canada. Report of the Royal Commision. 3 vol. Department of Supply and Services Canada, Hull Que.

Mangum, C. P. (1985) *Am. J. Physiol.* **248**, R505-R514.

Maniatis, T., Fritsch, E. F., Aner, J. and Lawn, R. M. (1980) *Ann. Rev. Genet.* **14**,145-178

Manning, A. M., Trotman, C. N. A. and Tate, W. P. (1990) *Nature* **348**, 653-656.

Mansfield, A. W. and Beck, B. (1977) *Environ. Can. Fish. Mar Serv. Tech. Rep.* **704**, 81p.

Marchionni, M. and Gilbert, W. (1986) *Cell* **46**, 133-141.

Margolis, L. (1977) *J. Fish. Res. Board Can.* **34**, 887-898.

McClelland, G. (1980) *Exp. Parasitol.* **49**, 405-419.

McClelland, G. (1990) In: Population biology of sealworm (*Pseudoterranova decipiens*) in relation to its intermediate and seal hosts. W.D. Bowen (ed.) *Can. Bull. Fish. Aquat. Sci.* **222**, 47-65.

McClelland, G., Misra, R. K. and Martell, D. J. (1990) In: Population biology of sealworm (*Pseudoterranova decipiens*) in relation to its intermediate and seal hosts. W.D. Bowen (ed.) *Can. Bull. Fish. Aquat. Sci.* **222**,83-118.

McGibbon, A., Christie, J. F. , Kennedy, M. W. and Lee, T. D. G. (1990) *Mol. Biochem. Parasitol.* **39**, 163-172.

Minagawa, J., Nakamura, H., Yamato, I., Mogi, T. and Anraku, Y. (1990) *J. Biol. Chem.* **265**, 11198-11203.

Moore, K. W., Rogers, J., Hunkapiller, T., Early, P., Nottenburg, C., Weissman, I., Bazin, H., Wall, R. and Hood, L. E. (1981) *Proc. Nat. Acad. Sci. USA* **78**, 1800-1804.

Naito, Y., Riggs, C. K., Vandergon, T. L., Riggs, A. F. (1991) *Proc. Nat. Acad. Sci. USA* **88**, 6672-6676.

Nash, A. R., Fisher, W. K. and Thompson, E. O. P. (1976) *Aust. J. Biol. Sci.* **29**, 73-97.

Nishioka, Y. and Leder, P. (1979) *Cell* **18**, 875-882.

Nishioka, Y., Leder, A. and Leder, P. (1980) *Proc. Nat. Acad. Sci. USA* **77**, 2806-2809.

Nojima, H. (1987) *FEBS Lett.* **217**, 187-190.

Obaru, K., Tsuzuki, T., Setoyama, C. and Shimada, K. (1988) *J. Mol. Biol.* **200**, 13-22.

O'Donnell, I. J., Dineen, J. K., Wagland, B. M., Letho, S., Dopheide, T. A. A., Grant, W. N. and Ward, C. W. (1989) *Int. J. Parasitol.* **19**, 793-802.

Orgel, L. E. and Crick, F. H. C. (1980) *Nature* **284**, 604-607.

Paggi, L., Nascetti, G., Cianchi, R., Orecchia, P., Mattiuchi, S., D'Amelio, S., Berland, B., Brattey, J., Smith, J. W. and Bullini, L. (1991) *Int. J. Parasitol.* **21**, 195-212.

Palmiter, R. D. (1974) *Biochemistry* **13**, 3606-3615.

Patthy, L. (1987) *FEBS Lett.* **214**, 1-7.

Pearson, W. R. and Lipman, D. J. (1988) *Proc. Nat. Acad. Sci. USA* **85**, 2444-2448.

Perler, F., Efstratiadis, A., Lomedico, P. Gilbert, W. Kolodner, R. and Dodgson, J. (1980) *Cell* **20**, 555-566.

Perlman, D. and Halvorson, H. O. (1983) *J. Mol. Biol.* **167**, 391-409.

Perret, C., Lomri, N. and Thomasset, M. (1988) *J. Mol. Evol.* **27**, 351-364.

Perutz, M. F. (1989) *TIBS* **14**, 42-44.

Qiugley, F., Martin, W. F. and Cerff, R. (1988) *Proc. Nat. Acad. Sci. USA* **85**, 2672-2676.

Riggs, C. K. and Riggs, A. F. (1991) In: Invertebrate Dioxygen Carriers. Preaux, G. (Ed.), Leuven Univ. Press, Leuven, Belgium.

Rochaix, J. D., Rahire, M. and Michel, F. (1985) *Nucl. Acids Res.* **13**, 975-984.

Rogers, J. H. (1986) *Trends Genet.* **2**, 223.

Rogers, J. H. (1989) *Trends Genet.* **5**, 213-216.

Rogers, J. H. (1990) *FEBS Lett.* **268**, 339-343.

Rossman, M. G., Moras, D. and Olsen, K. W. (1974) *Nature* **250**, 194-199.

Rothwell, T. L. W. and Love, R. L. (1974) *Int. J. Parasitol.* **4**, 293-299.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning. Cold Spring Harbour Laboratory Press, 2nd Edition.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Nat. Acad. Sci. USA* **74**, 5463-5467.

Schein, C. H. (1989) *Bio/Technology* **7**, 1141-1148.

Schoepfer, R. (1993) *Gene* **124**, 83-85.

Sharp, P. (1987) *Cell* **50**, 147-148.

Sherman, D. R., Guinn, B., Perdok, M. M. and Goldberg, D. E. (1992) *Science* **258**, 1930-1932.

Sherman, D. R., Kloek, A. P., Krishnan, R., Guinn, B. and Goldberg, D. E. (1992)*Proc. Nat. Acad. Sci. USA* **89**, 11696-11700.

Shih, M.-C., Heinrich, P. and Goodman, H. M. (1988) *Science* **242**, 1164-1166.

Sinderman, C. J. (1990) Principal diseases of marine fish and shellfish. Toronto, Academic Press. pp. 134-135.

Slightom, J. L., Blechl, A. E. and Smithies, O. (1980) *Cell* **21**, 627-638.

Smith, M. H. and Lee, D. L. (1962) *Proc. Royal Soc. (London) Series B.* **157**, 234-257.

Smith, C. W. J., Knaack, D. and Nadal-Ginard, B. (1990) in Intervening Sequences in Evolution and Development Stone, E. M. and Schwartz, R. J. (Eds) Oxford University Press, Oxford and New York. pp. 162-195.

Stone, E. M., Rothblum, K. N. and Schwartz, R. J. (1985) *Nature* **313**, 498-500.

Stromberg, B. E. and Soulsby, E. J. L. (1977) *Int. J. Parasitol.* **7**, 287-291.

Sudhoff, T. C., Goldstein, J. L., Brown, M. S. and Russell, D. W.(1985a) *Science* **228**, 815-822.

Sudhoff, T. C., Russell, D. W., Goldstein, J. L. and Brown, M. S. (1985b) *Science* **228**, 893-895.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Meig, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hakins, T., Ainscough, R. and Waterson, R. (1992) *Nature* **356**, 37-40.

Suzuki, T. and Gotoh, T. (1986) *J. Biol. Chem.* **261**, 9257-9267.

Takashi, T., Tobita, M. and Shikama, K. (1983) *Biochim. Biophys. Acta* **745**, 32-36.

Templeman, W., Squires, H. J. and Fleming, A. M. (1957) *J. Fish. Res. Board Can.* **14**, 831-897.

Templeman, W. (1990) In: Population biology of sealworm (*Pseudoterranova decipiens*) in relation to its intermediate and seal hosts. W.D. Bowen (ed.) *Can. Bull. Fish. Aquat. Sci.* **222**,1-16.

Tentori, L., Vivaldi, G., Carta, S., Marinucci, M., Massa, A., Antonini, E. and Brunori, M. (1973) *Int. J. Pept. Protein Res.* **5**, 182-200.

Terwilliger, R. C. (1980) *Amer. Zool.* **20**, 53-67.

Thompson, E. O. P. (1980) In *The Evolution of Protein Structure and Function,* Sigman, D. S. and Brazier A. B. (Eds)., Academic Press, New York, pp 267-298.

Titchen, D. A., Glenn, W. K., Nassif, N., Thompson, A. R. and Thompson E. O. P. (1991) *Biochim. Biophys. Acta* **1089**, 61-67.

Towbin, H., Staehelin, T. and Gordon, J. (1979) *Proc. Nat. Acad. Sci. USA* **76**, 4350-4354.

Traut, T. W. (1988) *Proc Nat Acad Sci USA* **85**, 2944-2948.

Urban, J. F. Jr. and Romanowski, R. D. (1985) *Exp. Parasitol.* **60**, 245-254.

van Daal, A., White, E. M., Elgin, S. C. R. and Gorovsky, M. A. (1990) *J. Mol. Evol.* **30**, 449-455.

van Holde, K. E. and Miller, K. I. (1982) *Q. Rev. Biophys.* **15**, 1-129.

Vanin, E. F. (1984) *Biochem Biophys Acta* **782**, 231-241.

Vanin, E. F., Goldberg, G. I., Tucker, P. W. and Smithies, O. (1980) *Nature* **286**, 222-226.

Vinogradov. S. N. (1985) *Comp. Biochem. Physiol.* **82B**, 1-15.

Voit, R. and Schneider, H.-J. (1986) *Eur. J. Biochem.* **159**, 23-29.

Volbeda, A., and Hol, W. J. G. (1989) *J. Mol. Biol.* **206**, 531-546.

West, D. K., Changchien, L.-M., Maley, G. F. and Maley, F. (1989) *J. Biol. Chem.* **264**, 10343-10346.

Wieringa, Hofer, E. and Weissmann, C. (1984) *Cell* **37**, 915-925.

Wittenberg, B. A., Okazaki, T., and Wittenberg, J. B. (1965) *Biochim. Biophys. Acta* **111**, 485-495.

Wodak, S. J. and Janin, J. (1981) *Biochemistry* **20**, 6544-6552.

Wood, E. J. (1980) *Essays Biochem.* **16**, 1-47.

Xu, M.-G., Kathe, S. D., Goodrich-Blair, H., Nierzwicki-Bauer, S. A. and Shub, D. A. (1990) *Science* **250**, 1566-1570.

Yabusaki, Y., Murakami, H., and Ohkawa, H. (1988) *J. Biochem.* **103**,1004-1010.

Young, R., and Davis, R. (1983) *Proc. Nat. Acad. Sci. USA* **80**, 1194-1198.

Zagorec, M., Buhler, J.-M., Treich, I., Keng, T., Guarente, L. and Labbe-Bois, R. (1988) *J. Biol. Chem.* **263**, 9718-9724.

Zuckerman, B. M., Mai, W. F. and Rohde, R. A. (Eds) (1971) Plant Parasitic Nematodes. Academic Press, New York.