Grasshopper lectin : cDNA sequence, amino acid sequence and computer-based homology model
by Jay Richard Radke

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biochemistry
Montana State University

Abstract:
This work extends research directed at the structural and functional characteristics of the lectin(s) from grasshopper Melanoplus differentialis and their role in immunity and molecular defense. Insect lectins mediate the progression of cellular defense and are considered a form of primordial recognition molecule. These proteins represent a potential immuno-recognitory surveillance molecule of broad carbohydrate specificity and their biochemical and genetic characterization are prerequisite to a complete understanding of invertebrate immune defense. Seventeen (17) clones have been isolated from a cDNA library constructed from fat body mRNA. Three clones were sequenced (Clones 2, 3 and 4). The sequences for clones 2 and 3 represent the same mRNA. Clone 3 was 879 bp long, but missing ≈ 200 bp at the 3' end including the stop codon and polyA tail. The clone 4 cDNA is 1079 bp long and includes sequence representing a stop codon and the 3' non-translated region through the polyA tail, but is incomplete at the 5' end missing sequence representing the initiating Met and signal sequence. Comparison of these sequences indicates the two are similar but represent distinct mRNAs, and presumably distinct genes. PCR evaluation of the remaining 14 clones, using gene-specific primers, has determined that a clone containing either the missing 3' end of clone 3 or the 5' end of clone 4 is not available in the cDNA library.

3' RACE (Rapid Amplification of cDNA Ends) was used to amplify and isolate the missing 3' end of clone 3. Reverse transcriptase-polymerase chain reaction (RT-PCR) using a gene-specific primer resulted in a single band of 800 bp representing the 3' end of clone 3. This fragment was cloned and sequenced. In combination with the original clone 3, this sequence completes a cDNA representative of a full-length message for a lectin. The cDNA is 1220 bp long including sequence for the signal peptide and the 3' non-translated region through the polyA tail. The 972 bp open reading frame (ORF) encodes a 324 amino acid polypeptide including a 20 residue signal sequence. The calculated molecular weight of the 304 amino acid protein is 34056 Da. Two potential glycosylation sites are present. Grasshopper lectin is unique among invertebrate lectins having two homologous carbohydrate recognition domains (CRD). This duplication may contribute to the dual carbohydrate specificity (galactose/glucose) shown by this protein. The sequence defining each CRD demonstrates ≈ 30% identity with other vertebrate and invertebrate C-type lectins and 100% identity at 7 invariant amino acid positions within the CRD.

5' RACE was used to amplify and isolate the missing 5' end of clone 4. PCR using a clone 4-specific primer resulted in a single band of ≈ 1000 bp which contained the missing 5' end of this clone. This fragment was cloned and partially sequenced. In combination with sequence known for the original clone 4, the new sequence completes a cDNA representative of a full-length message for a second lectin. The complete sequence is 1213 bp long including sequence representing the signal peptide and the 3' non-translated region through the polyA tail. A 978 bp ORF encodes a 326 amino acid polypeptide including a 21 residue signal sequence. The calculated molecular weight of this 305 amino acid protein is 34401 Da. Clone 3 and Clone 4 cDNAs are 81% identical.

A 3D homology model of CRD2 from clone 3 was constructed using the crystal structures for rat

mannose binding protein (MBP) and E-Selectin (ESEL) as references. Assessment of the model using φψ angles in a Ramachandran plot, dihedral angles and 3D Profiles suggests the model CRD structure is a reasonable 3D representation of the CRD sequence.

GRASSHOPPER LECTIN: cDNA SEQUENCE, AMINO ACID SEQUENCE AND COMPUTER-BASED HOMOLOGY MODEL

by

Jay Richard Radke

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY-BOZEMAN
Bozeman, Montana

November 1996

APPROVAL

of a thesis submitted by

Jay Richard Radke

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Nov-27, 1996
Date

Kenneth D Hapner
Chairperson, Graduate Committee

Approved for the Major Department

Nov. 27, 1996
Date

David M Dooley
Head, Major Department

Approved for the College of Graduate Studies

12/3/96
Date

R L Brown
Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a doctoral degree at Montana State University-Bozeman, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature

Nov 26, 1996

Date

ACKNOWLEDGMENTS

I wish to thank my mentor and advisor Dr. Kenneth D. Hapner for his efforts and consideration during my work in his laboratory. I have found his passion for teaching, science and his delight in solving problems professionally motivating and his compassion for colleagues and students personally inspiring. I also acknowledge the support and expertise of my graduate committee; Dr. Larry Jackson, Dr. Martin Teintze, Dr. John Amend, Dr. Bill Dyer and Dr. Jack Olson. Finally, I wish to thank my family for the sacrifices they have made for me to be here; my wife Michelle for never telling me "I shouldn't have done this", my son Josh for loaning us money when we needed groceries, my son Daniel for being perpetually happy no matter the situation and my daughters Miranda and Arica for helping me remember that there are infinitely more important things to worry about than the failure of my last experiment. I certainly would not have finished this work without them.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

ABSTRACT

This work extends research directed at the structural and functional characteristics of the lectin(s) from grasshopper *Melanoplus differentialis* and their role in immunity and molecular defense. Insect lectins mediate the progression of cellular defense and are considered a form of primordial recognition molecule. These proteins represent a potential immuno-recognitory surveillance molecule of broad carbohydrate specificity and their biochemical and genetic characterization are prerequisite to a complete understanding of invertebrate immune defense. Seventeen (17) clones have been isolated from a cDNA library constructed from fat body mRNA. Three clones were sequenced (Clones 2, 3 and 4). The sequences for clones 2 and 3 represent the same mRNA. Clone 3 was 879 bp long, but missing ≈ 200 bp at the 3' end including the stop codon and polyA tail. The clone 4 cDNA is 1079 bp long and includes sequence representing a stop codon and the 3' non-translated region through the polyA tail, but is incomplete at the 5' end missing sequence representing the initiating Met and signal sequence. Comparison of these sequences indicates the two are similar but represent distinct mRNAs, and presumably distinct genes. PCR evaluation of the remaining 14 clones, using gene-specific primers, has determined that a clone containing either the missing 3' end of clone 3 or the 5' end of clone 4 is not available in the cDNA library.

3' RACE (Rapid Amplification of cDNA Ends) was used to amplify and isolate the missing 3' end of clone 3. Reverse transcriptase-polymerase chain reaction (RT-PCR) using a gene-specific primer resulted in a single band of 800 bp representing the 3' end of clone 3. This fragment was cloned and sequenced. In combination with the original clone 3, this sequence completes a cDNA representative of a full-length message for a lectin. The cDNA is 1220 bp long including sequence for the signal peptide and the 3' non-translated region through the polyA tail. The 972 bp open reading frame (ORF) encodes a 324 amino acid polypeptide including a 20 residue signal sequence. The calculated molecular weight of the 304 amino acid protein is 34056 Da. Two potential glycosylation sites are present. Grasshopper lectin is unique among invertebrate lectins having two homologous carbohydrate recognition domains (CRD). This duplication may contribute to the dual carbohydrate specificity (galactose/glucose) shown by this protein. The sequence defining each CRD demonstrates ≈ 30% identity with other vertebrate and invertebrate C-type lectins and 100% identity at 7 invariant amino acid positions within the CRD.

5' RACE was used to amplify and isolate the missing 5' end of clone 4. PCR using a clone 4-specific primer resulted in a single band of ≈ 1000 bp which contained the missing 5' end of this clone. This fragment was cloned and partially sequenced. In combination with sequence known for the original clone 4, the new sequence completes a cDNA representative of a full-length message for a second lectin. The complete sequence is 1213 bp long including sequence representing the signal peptide and the 3' non-translated region through the polyA tail. A 978 bp ORF encodes a 326 amino acid polypeptide including a 21 residue signal sequence. The calculated molecular weight of this 305 amino acid protein is 34401 Da. Clone 3 and Clone 4 cDNAs are 81% identical.

A 3D homology model of CRD2 from clone 3 was constructed using the crystal structures for rat mannose binding protein (MBP) and E-Selectin (ESEL) as references. Assessment of the model using ϕ-ψ angles in a Ramachandran plot, dihedral angles and 3D Profiles suggests the model CRD structure is a reasonable 3D representation of the CRD sequence.

INTRODUCTION

Statement of Purpose

The purpose of this research was to extend the characterization of a lectin(s) from the grasshopper, *Melanoplus differentialis*. This work was addressed using two strategies; 1) the completion of cDNA sequences representing full-length mRNAs for two distinct lectins from this grasshopper, and 2) the construction of a computer-based homology model for a carbohydrate recognition domain from one lectin, using the amino acid sequence and the crystal structure coordinates from two homologous reference proteins.

Lectins are proteins or glycoproteins defined by their ability to bind carbohydrate without covalent modification (Rini, 1995). These proteins are found in the hemolymph of a wide variety of invertebrates (Kennedy, 1995) and are considered to be ubiquitous among living organisms (Sharon, 1993). Carbohydrate structures are often the primary determinants of cellular identity (Gaveriaux, 1987). Distinct carbohydrate specificity suggests lectins may distinguish between cell types and thereby mediate a wide range of cellular interactions (Marschal, 1992). Thus lectins that bind carbohydrate on a foreign surface may serve as a discriminatory link between nonself material and the hemocytes involved in immune defense (Tsuboi, 1993). A lectin from grasshopper has been shown to opsonize fungal blastospores (Wheeler, 1993) suggesting this protein represents a defense molecule with selective specificity (Bradley, 1989).

Insect Immune-Defense

Insects proliferate in habitats where microbial predators use them as a source of nutrition (Hoffman, 1995). The evolution of effective molecular and cellular defense mechanisms has thus been prerequisite to their survival (Ratcliffe, 1985). Invertebrates lack the immunoglobulin-based defense common to vertebrate organisms (Gupta, 1986), but employ well defined physical, cellular and molecular defense strategies. The primary defense against pathogenic infection is prevention via passive structural barriers to penetration of the hemocoel (Dunn, 1986). These barriers include a sclerotized outer integument (Gupta, 1992) and the peritrophic membrane (Spence, 1993), which isolates the midgut epithelium and hemolymph from ingested food. Insects have also evolved effective cellular (Hajek, 1994) and humoral (Vasta, 1987) defense strategies to act on pathogens that do breach the hemocoel.

Cellular Defense: Phagocytosis and Encapsulation

The first defense response to foreign entities in the hemocoel is mediated by circulating granulocytes and plasmatocytes (Marmaras, 1993). This cellular response occurs via random encounters with the pathogen (Dunn, 1990) and leads to its phagocytosis (Ratcliffe, 1977) or encapsulation (Schmit, 1977). Granulocytes are proposed to be involved with early recognition of non-self entities (Hagen, 1994). Following recognition of nonself substrate, these cells degranulate to release antibacterial factors such as lysozyme (Zachary, 1974) and components of the prophenoloxidase cascade (Leonard, 1985). Phenoloxidase causes the hydroxylation of mono-phenols such as tyrosine to o-diphenols and o-quinones which mediate adhesion of the pathogen to hemocytes (Marmaras, 1996) and promote the subsequent encapsulation or phagocytosis by plasmatocytes (Ratcliffe, 1984).

Humoral Defense: Antibacterial Proteins

The induced synthesis of specific antibacterial proteins in response to infection defines the humoral defense (Gupta, 1986). Lysozyme (Morishima, 1994), pore-forming peptides termed cecropins (Steiner, 1981) and bacteriocidal proteins termed attacins (Hultmark, 1983) all represent antibacterial proteins. These proteins bind distinct components of the bacterial cell wall and effect destruction of the invading pathogen (Vasta, 1985). These proteins were first demonstrated in the wax moth (Stephens, 1962) and the silk moth (Hink, 1968). Sarcotoxin I (Okada, 1985), attacin-like Sarcrotoxin II (Ando, 1983) and defensin (Matsuyama, 1988) are antibacterial proteins in the flesh fly *Sarcophaga peregrina* . These proteins are synthesized in the fat body and secreted into the hemolymph where they exhibit antibacterial activity against invading bacteria and pathogenic fungi.

The immunoglobulin-like protein, hemolin or P4 (Sun, 1990), is one of the first hemolymph components to bind the surface of bacteria in the tobacco hornworm. A similar protein initiates complex formation that mediates a cellular immune response in the giant silk moth (Ladendorff,1991). This protein increases hemocyte adhesion to nonself substrate during the putative recognition phase of the immune response (Kanost, 1993).

Invertebrate Immuno-recognition

Cellular and humoral defense strategies require discrimination between 'self' and 'nonself' to be effective. A definitive model for immuno-recognition (Mullett, 1993) and the molecular strategy employed to recognize potential pathogenic substrate remains problematic (Hoffman, 1996). Invertebrate lectins are considered a form of primordial recognition molecule (Kennedy, 1995). The ability of these proteins to bind specific carbohydrate moieties suggests they serve an immuno-surveillance function by providing a discriminatory link between nonself substrate and circulating hemocytes (Vasta, 1987).

Insect Lectins and Immuno-surveillance

Lectins that demonstrate putative immuno-recognitory properties have been isolated and purified from several insects. The beet armyworm (*Spodoptera exigua*) contains high levels of a naturally occurring endogenous galactose-binding lectin (Pendland, 1986). This protein (Pendland, 1988) has demonstrated opsonic properties toward blastospores of the entomopathogenic fungi *Paecilomyces farinosus* and *Beauveria bassiana* (Boucias, 1993). Recent studies have shown that injected cells of *P. farinosus*, that present surface galactosidic residues, are rapidly cleared from the hemolymph while blastospores of *Nomuraea rileyi* lacking such residues are not. A lectin receptor has been localized to an outer coat associated with these blastospores (Boucias, 1991). The beet armyworm lectin is a large molecular weight aggregate comprised of two subunits which are 33 and 34 kDa (Boucias, 1993). It is believed the 34 kDa subunit is a glycosylated version of the 33 kDa subunit. Each is present in an equimolar amount. This protein requires Ca^{2+} for binding carbohydrate.

A Ca^{2+}-dependent lectin from the tobacco hornworm *Manduca sexta* is induced via bacterial challenge and its presence associated with hemocyte-mediated coagulation of the invading pathogen (Minnick, 1986). It has been suggested this protein may regulate the related process of encapsulation (Kanost, 1994).

The flesh fly *Sarcophaga peregrina* (Natori, 1987) has a lectin which is induced by injury to the cuticle. Synthesized in the fat body, this protein mediates lysis and clearance of sheep erythrocytes from the hemocoel (Komano, 1985). The protein is also present during pupation, suggesting a role in removing histolytic debris during morphogenesis. This Ca^{2+} dependent protein is a 190 kDa aggregate consisting of α and β subunits of 32 kDa and 35 kDa (Komano, 1980). It is proposed the β subunit is a posttranslationally modified version of the α subunit (Takahashi, 1985). A cDNA representing this lectin

codes for a 279 amino acid polypeptide including a 19 residue signal sequence (Takahashi, 1985). The gene coding for this lectin has been cloned into puc118 and expressed in a nuclear extract of NIH-Sape 4 insect cells suggesting this extract contains transcription factors for this gene (Kobayashi, 1989).

Periplaneta lectin, first purified from cockroach hemolymph (Kubo, 1987) recognizes 2-keto-3-deoxy octanate (KDO) in the carbohydrate portion of *Escherichia coli* and *Salmonella minnesota* lipopolysaccharides (LPS). This lectin requires Ca^{2+} to bind KDO and has also been shown to mediate the neutralization and clearance of bacterial cells from the hemocoel (Kawasaki, 1993). A second cockroach lectin, LPS binding protein (Jomori, 1990) was shown to specifically recognize only *E. coli* LPS. A transient third lectin (Kubo, 1993) mediates the regeneration of cockroach legs implying a role in morphogenic development, but not defense.

A Ca^{2+}-dependent lectin from the grasshopper *Melanoplus differentialis* specifically binds D-galactose and D-glucose (Stebbins, 1985). This protein is a $\approx$ 700 kDa aggregate of individual 70 kDa units, each consisting of 40 and 30 kDa subunits (Stebbins, 1985). Recent evidence from the purified protein indicate the 70 kDa unit consists of two similar or identical $\approx$ 36 kDa subunits (Wenzlick, 1996) and the 30 kDa subunit previously identified is a contaminant. The grasshopper lectin is synthesized in the fat body, testis and ovaries (Stiles, 1988) and secreted into the hemolymph where it has demonstrated opsonic properties towards blastospores of the fungus *Beavaria bassiana* (Wheeler, 1993). The protein has also been associated with the outer membrane of hemocytes (Bradley, 1989). Synthesis of the grasshopper lectin is not induced by injury to the body wall or bacterial infection. The protein is considered present in the hemolymph at a level sufficient to allow function as an effective defense molecule (Wheeler, 1993). cDNAs coding for portions of two lectins from grasshopper have been isolated and sequenced (Rognlie, 1991). The coded amino acid sequences are similar, but do not match sequence obtained from the

lectins isolated in this laboratory. Completion of these cDNAs represents one focus of this work.

Amino Acid Sequences for Invertebrate Lectins

The amino acid sequences for several invertebrate lectins have been determined and 25-35% homology between these sequences in regions defining their carbohydrate recognition domains (CRD) has been demonstrated. This alignment is shown in Figure 1 and includes two vertebrate lectins, rat mannose binding protein (MBP) and human E-selectin (ESEL). The amino acid sequences for the Periplaneta lectin from cockroach (Jomori, 1991), the lectin from flesh fly (Takahashi, 1985) and the partial sequence from grasshopper (Rognlie, 1991) each harbor invariant residues at key positions suggesting the presence of a C-type lectin 'fold' (Drickamer, 1992). This 'fold' is defined by the crystal structure (Weis, 1992) for the expressed rat mannose binding protein (MBP). The C-type architecture consists of four strictly conserved Cysteine residues connected via two disulfide bridges such that a 3 dimensional "loop-within-a-loop" motif is formed. The loops are stabilized by two Ca^{2+} ions which are also required for carbohydrate binding, thus the name "C-type". The Ca^{2+} ions are unique within this family of proteins because they are bound to both protein and carbohydrate substituents (Weis, 1996).

The lectin(s) from grasshopper is a C-type lectin (Stebbins, 1985). The strict conservation of amino acids in key positions implies the sequence defining the CRD from grasshopper may occupy a similar C-type fold as defined by MBP and ESEL. Characterization of the structures for known C-type CRDs like MBP, and comparison with homologous sequences provide the basis for predicting the 3D structure for a CRD from a grasshopper lectin. Construction of a homology model (Sutcliffe, 1987) for a CRD from the grasshopper lectin represents a second focus of this project.

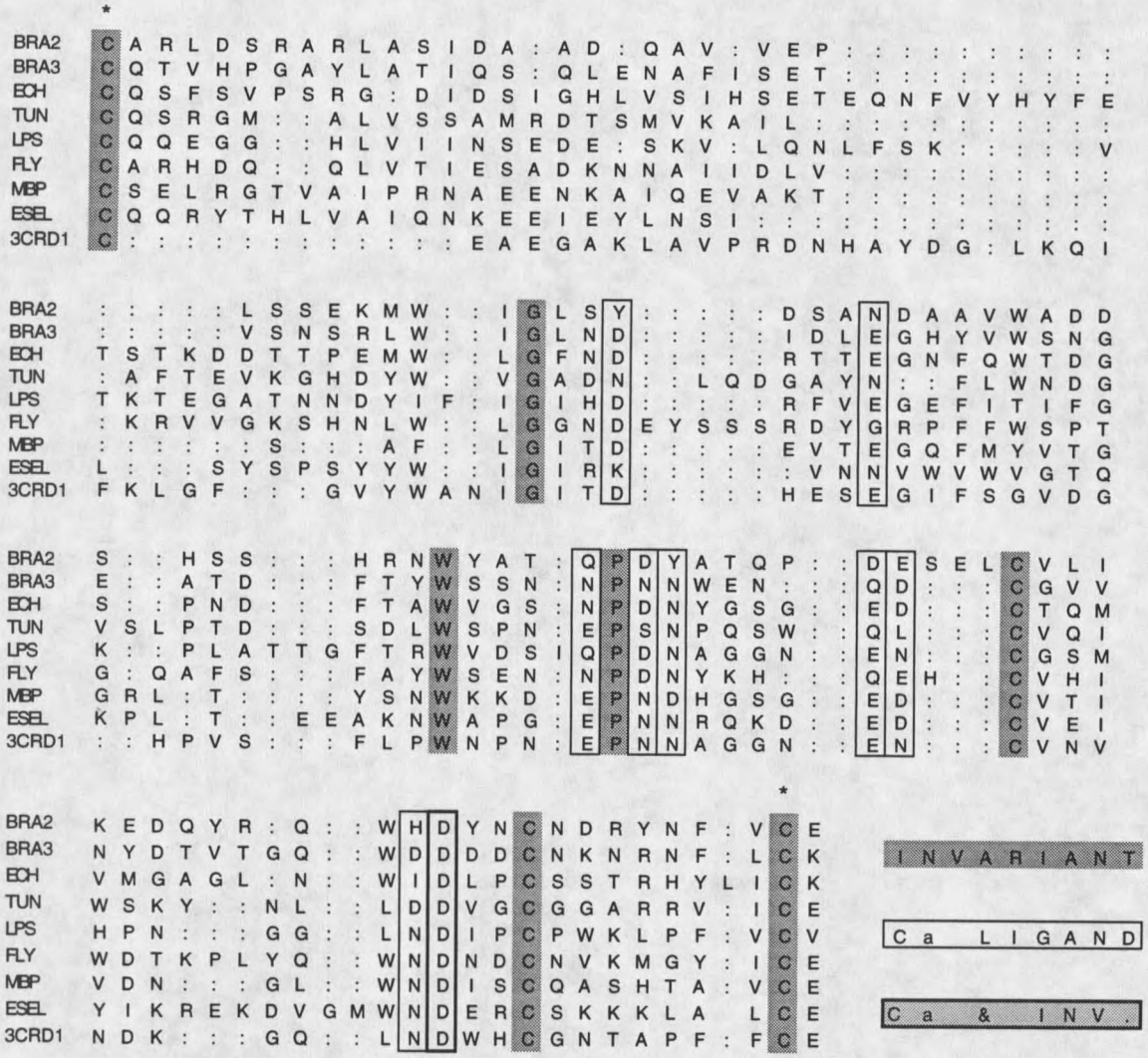


Figure 1. Amino acid sequences for nine invertebrate and vertebrate C-type lectin CRDs. A CRD is defined by the amino acid sequence between the terminal Cys residues disulfide bridged to form the large loop (Weis, 1992 & Graves, 1994). These positions are marked with an asterisk. Seven invariant positions are shaded gray and nine positions conserved or conservatively substituted are boxed. 3CRD1 is the N-terminal CRD encoded by the clone 3 sequence (Rognlie, 1991). ESEL=human E-selectin, MBP=rat mannose binding protein, FLY=fleshfly, LPS=cockroach, TUN=tunicate, ECH=echinoidin and BRA2 & 3=barnacle.

Structure Prediction by Homology

Sequences for over 100,000 proteins have been reported yet structure determination from X-ray crystallography and NMR spectroscopy has provided less than 2000 solved structures (Sali, 1993). The growth of this disparity has stimulated the development of predictive tools to generate reasonable model structures given only the amino acid sequence. Modeling by homology to known structures is one example (Bajorath, 1993). Numerous protein structures e. g. insulins, globins and serine proteases confirm that related proteins from different species adopt similar tertiary structures characteristic of the family (Blundell, 1993).

There are no crystal structures available for a C-type domain from an invertebrate lectin. However, the amino acid sequences of several invertebrate C-type CRDs have been determined. When aligned with each other and representative members of the vertebrate C-type lectin family including MBP and ESEL, these sequences demonstrate ≈ 30% overall identity and 100% identity at 8 invariant amino acids positions (Figure 1). It is logical that these invertebrate sequences may occupy a 3D structure similar to that defined by MBP and ESEL (Drickamer, 1994). The crystal structures for MBP and ESEL represent structural templates on which the sequence for the C-type CRD from the grasshopper lectin may be modeled by homology (Blundell, 1993). A 3D model for a grasshopper CRD represents the second focus of this study.

Classification of Lectins

Lectins comprise a structurally diverse class of proteins defined by their ability to bind carbohydrate (Rini, 1995). It is possible to group lectins into distinct families of homologous proteins that share common structural and functional properties (Sharon, 1993). Animal lectins can be divided into two distinct families; 1) C-type lectins

(Drickamer, 1988) and 2) S-type lectins or galectins (Barondes, 1994). The lectin from grasshopper is considered a C-type lectin, thus discussion will focus on C-type lectins and the two representative crystal structures from this family of proteins.

The C-type lectins are a family of carbohydrate binding proteins characterized by the presence of a ≈ 15 kDa , Ca^{2+}-dependent CRD. This family includes endocytic receptors (Lodish, 1991) such as hepatic asialoglycoprotein receptor, 'selectins' (Geng, 1992 and Whelan, 1996) which mediate targeting and adhesion of leukocytes to endothelial cells and 'collectins', innate defense proteins found in the extracellular matrix and in serum (Holmskov, 1994). These proteins contain one or more CRDs that range from ≈ 115 to 134 amino acid residues in length (Drickamer, 1994). The C-type lectin CRD contains 8 invariant amino acid positions and 8 positions which are conserved in character (Drickamer, 1993). These residues play critical roles in defining the architecture of the authentic C-type lectin CRD, including the characteristic pattern of disulfide bonds between conserved Cys residues and the ligation of bound Ca^{2+} by aspartate, asparagine and glutamate (Weis, 1992).

The most studied of the vertebrate C-type lectins are the Type III lectins. Also called collectins (Hoppe, 1994), these proteins have an amino-terminal collagen-like domain with CRDs located at the carboxy-most end of the polypeptide (Holmskov, 1994). Mannose binding protein (MBP) is a member of this family. These proteins are found in the serum (Hoppe, 1994) or secreted as part of the lung surfactant (Kuroki, 1994) where they provide an innate immune response in which "nonself" is distinguished from "self" based on the presence of cell-specific carbohydrate moieties on the surface of potentially pathogenic microorganisms (Holmskov, 1994). Serum mannose binding protein fixes complement and acts as an opsonin (Lu, 1990) to invoke the antibody-dependent arm of the immune system. The x-ray crystal structure for rat MBP has been determined (Weis, 1992).

E-selectin is a Type IV C-type lectin (Drickamer, 1994) containing a CRD similar to that seen in rat MBP (Graves, 1994). These adhesion molecules are expressed on endothelial cells at a site of inflammation and serve to bind carbohydrate on the surface of circulating neutrophils and monocytes responding to the perturbation (Whelan, 1996).

Structure of Rat MBP

Two X-ray structures for a single CRD from rat mannose-binding protein A have been solved (Weis, 1991 and Weis, 1992). The first structure was solved at 2.3 Å resolution (Weis, 1991; PDB filename=pdb1msb.ent). This protein is a collectin serving an antibody-independent host defense function as discussed above. A second structure, in complex with an oligo-mannose asparaginyl oligosaccharide, has been reported at 1.7 Å resolution (Weis, 1992; PDB filename =pdb2msb.ent). A ribbon diagram and space-filling model for the MBP structure are illustrated side-by-side in Figure 2 (top). Four conserved Cys residues are shown as space-filled atoms within the ribbon diagram. These residues are colored by atom e. g. yellow=S, red=O, blue=N, green=C and white=H. Cys128 and Cys217 are bridged by a disulfide bond forming the outer loop of the 'loop-within-a-loop' motif. The ribbon defining this loop is purple. Cys195 and Cys209 are bridged to form the small 15 amino acid loop which is colored white. This arrangement defines the archetypal C-type motif discussed on page 7. Conserved residues and those serving as Ca^{2+} ligands are colored orange and labeled using the single letter abbreviation for the amino acid residue and its position number in the amino acid sequence. The space-filling model of the CRD helps illustrate the globular nature of this domain and the relative orientation of the small loops, conserved residues and Ca^{2+} ligands. Atoms of the big loop are colored purple as defined in the ribbon diagram while atoms representing the small loop are white. Conserved residues and Ca^{2+} ligands are colored orange. Roughly 50% of

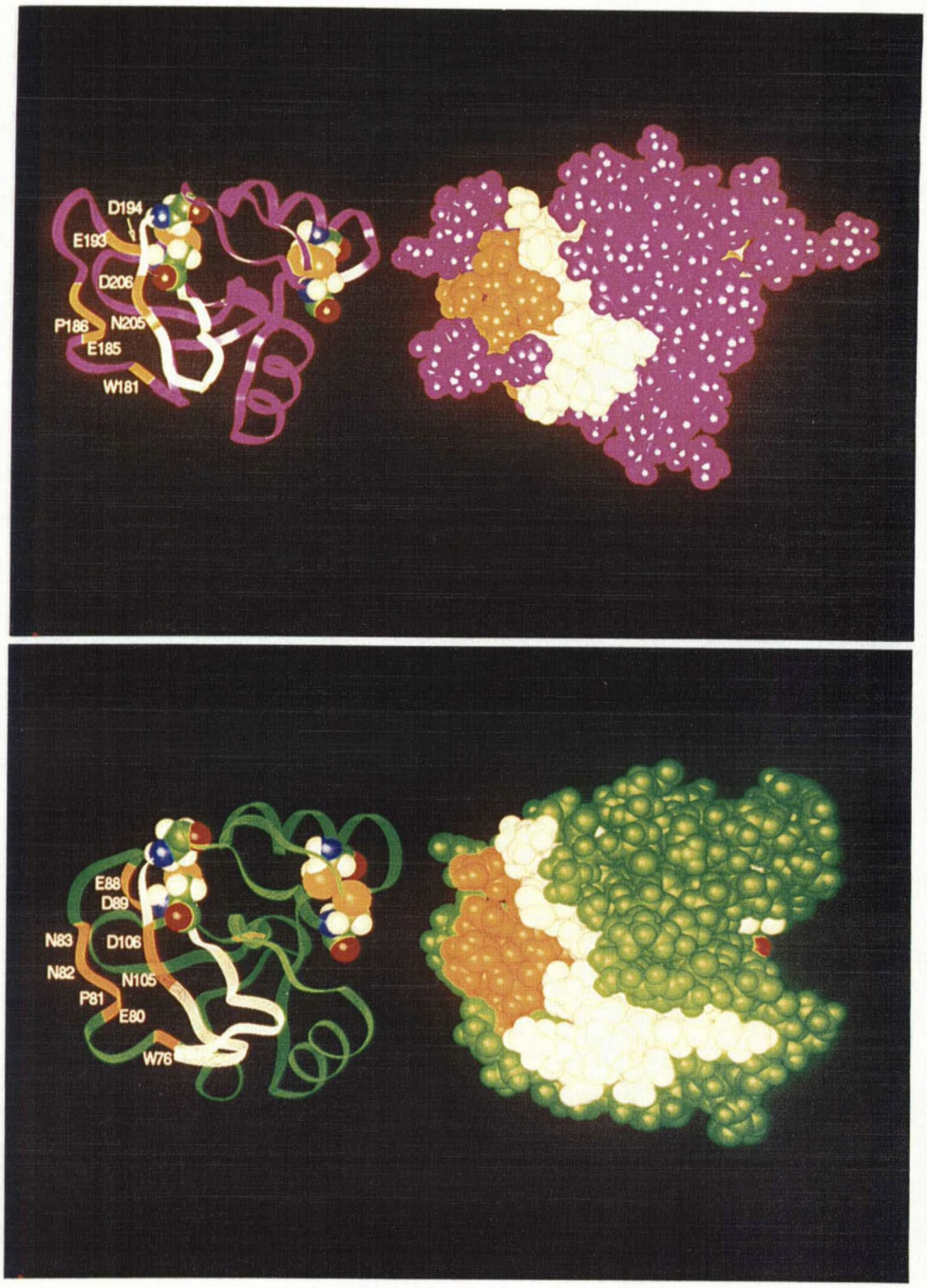


Figure 2. Ribbon structures and space-filling models for the crystallized CRDs from rat MBP (top) and human E-selectin (bottom). Space-filled atoms in each ribbon represent conserved Cys residues bridged to form the C-type motif. Ca^{2+} ligands are orange and labeled by residue and sequence position in each ribbon structure. The small loop in the motif is colored white in both ribbon diagrams. The large loop in MBP is purple and green in ESEL. Comparison of MBP and ESEL illustrates qualitative similarities including the spatial orientation of conserved Cys residues forming the large and small loops and the location of Ca^{2+} ligands.

this structure is represented by loops and extended conformations and 50% is comprised of two α helices and five β sheets (Rini, 1995).

The C3 and C4 hydroxyl groups of the terminal mannose interact directly with one Ca^{2+} ion. In addition, Glu185, Asn187, Glu193 and Asn205 sidechains form coordinate covalent bonds with Ca^{2+} (Iobst, 1994). The β carbon of His189 is also within a van der Waals distance of C_1 on the terminal mannose ring (Iobst, 1994). The His189 C_β forms one edge of the carbohydrate binding site.

Human E-selectin: Structure and Comparison to Rat MBP

The X-ray structure for the human C-type lectin domain of E-selectin (ESEL) has been solved at 2.0 Å resolution (Graves, 1994; PDB filename=pdb1esl.ent). A ribbon diagram and space-filling model for this structure are also illustrated in Figure 2 (bottom). The four conserved Cys residues are shown as space-filling models within the ribbon diagram and colored as described for MBP. For ESEL, Cys19 and Cys117 are bridged by a disulfide bond forming the outer loop of the motif which is colored green: Cys90 and Cys109 are bridged to form the small loop which is shown in white. As in the structure for MBP, this arrangement defines the archetypal C-type motif. Conserved residues and those serving as Ca^{2+} ligands are colored orange and labeled by amino acid residue and sequence position. In the space-filling model, atoms representing the big loop are colored green. Atoms representing the small loop are colored white and conserved residues and/or Ca^{2+} ligands are orange. As described for MBP above, roughly 50% of the CRD from ESEL is represented by loops and extended conformations while 50% occupies two α helices and five β sheets (Rini, 1995).

Quantitatively, the structures for CRDs from MBP and ESEL are similar. The root-mean-square difference (RMSD) calculated over equivalent C_α-backbone atoms is 1.94 Å. Comparison in Figure 2 also illustrates four qualitative similarities: 1) the spatial orientation of disulfide bridges between conserved Cys residues defining the 102 residue and 111 residue large loops in the MBP and ESEL, 2) the 15 residue and 20 residue small loops for MBP and ESEL, 3) the space occupied and the orientation of these loops with regard to one another and 4) the space and orientation of conserved residues and Ca^{2+} ligands. The five residue insert in the small loop of ESEL occupies a 3D space which forces the second a helix in this domain to occupy a slightly different conformation when compared to MBP.

Using Homology Models

Computer-based modeling of an amino acid sequence based on the known crystal structures for one or more reference proteins is now commonly used to generate a 3D structure for a polypeptide (Bajorath, 1993). Model building of the HIV protease (Weber, 1989) based on the structure of Rous Sarcoma virus protease (Wlodawer, 1989) has allowed assessment of the homology generated model based on the later determination of the HIV protease structure (Weber, 1990). The model differed from the crystal structure by an RMSD of 1.4 Å over all equivalent C_α-backbone atoms.

Four homology models have been constructed based on the crystal structure for rat MBP. The lectin-domain of the human and murine low-affinity Fcε receptor (Padlan, 1993) were the first to be constructed using MBP. The neural cell adhesion protein, fasciclin III (Castonguay, 1995), the putative T-cell activation antigen CD69 (Bajorath, 1994) as well the type II antifreeze protein (Sonnichsen, 1995) from smelt and herring have all since been modeled based on the known structure for MBP. A homology model for ESEL (Erbe, 1992), prior to solution of its crystal structure, was also built based on the

structure for MBP. Comparison of the final model with the crystal structure (Graves, 1994) suggests that use of only MBP as a template limited the accuracy of the model in regions of unusual secondary structure (Bajorath, 1996). The crystal structures for both rat MBP and human ESEL were recently used to model the lectin binding domain of the human cell surface receptor CD23 (Bajorath, 1996). Two reference structures allow for better definition of the conserved core regions in this family of proteins (Sutcliffe, 1987), and also help define structurally conserved regions where typical secondary structures are not evident. The increased confidence in defining legitimately conserved regions of the CRD provides an improved basis for comparative modeling of other members of the C-type lectin family.

Project Rationale

The rationale for this work is based on the assumption that insect lectins represent potential immuno-recognitory molecules. Current models for invertebrate defense fail to define the molecular details mediating immuno-recognition (Hoffman, 1995). Insect pests compete with man for agricultural products and may act as vectors for disease (Richman, 1995) in both man and animals. Interference or restriction of immune-defense systems may serve to increase the organisms vulnerability to natural predation or managed control agents and to enhance the efficacy of biological pest control strategies. This may contribute to a reduced dependence on chemical pesticides. Thus, the molecular details related to immune factors such as possible recognition molecules are important. Basic structural characterization of the lectin(s) from grasshopper may help refine models for invertebrate immune-defense and provide a platform from which further studies will continue to investigate the structural characteristics of these proteins.

Previous Work

A cDNA library was previously constructed using a λgtll expression vector and mRNA isolated from the fat body of the grasshopper *Melanoplus differentialis* (Hapner, unpublished). Immuno-screening with rabbit anti-agglutinin antibody identified a lectin-related cDNA coding for an antigenic β-galactosidase fusion protein. The cDNA insert from one purified plaque was subcloned into pGEM7fz(+) plasmid vector and sequenced (Rognlie, 1991). This 300 bp cDNA clone was ^{32}P-labeled via nick translation and used as a hybridization probe to screen the library for additional positive plaques. An additional 17 plaques were identified and three cDNAs (clones 2, 3 & 4) were selected, based on their larger size, and subcloned into pGEM7fz(+) plasmid vector. These clones and their relative size are illustrated in Figure 3. The cDNA inserts in each of clones 2 and 4 were not completely sequenced. The clone 3 insert was sequenced in both directions (Rognlie, 1991). This 879 bp cDNA contained a single open-reading frame (ORF) encoding a presumed initiating Met, a 20 residue signal sequence and 268 amino acids. The cDNA did not contain a stop codon or any 3' nontranslated sequence through the polyA tail. Limited cDNA sequence from the 3' region of clone 4 contained an ORF, a stop codon (TGA) and polyA tail in addition to a homologous sequence that overlapped the 3' terminus of clone 3. The clone 4 sequence was combined with that of clone 3 to complete a hybrid cDNA representative of a full-length mRNA as defined by the presence of a start codon (ATG), stop codon and polyA tail. The composite cDNA contained a 972 bp ORF and encoded a 324 amino acid polypeptide with a calculated molecular weight of 36 kDa. This value is similar to that measured for a purified lectin isolated from grasshopper in the laboratory (Stebbins, 1985). This work represents the starting point for the research discussed in this thesis.

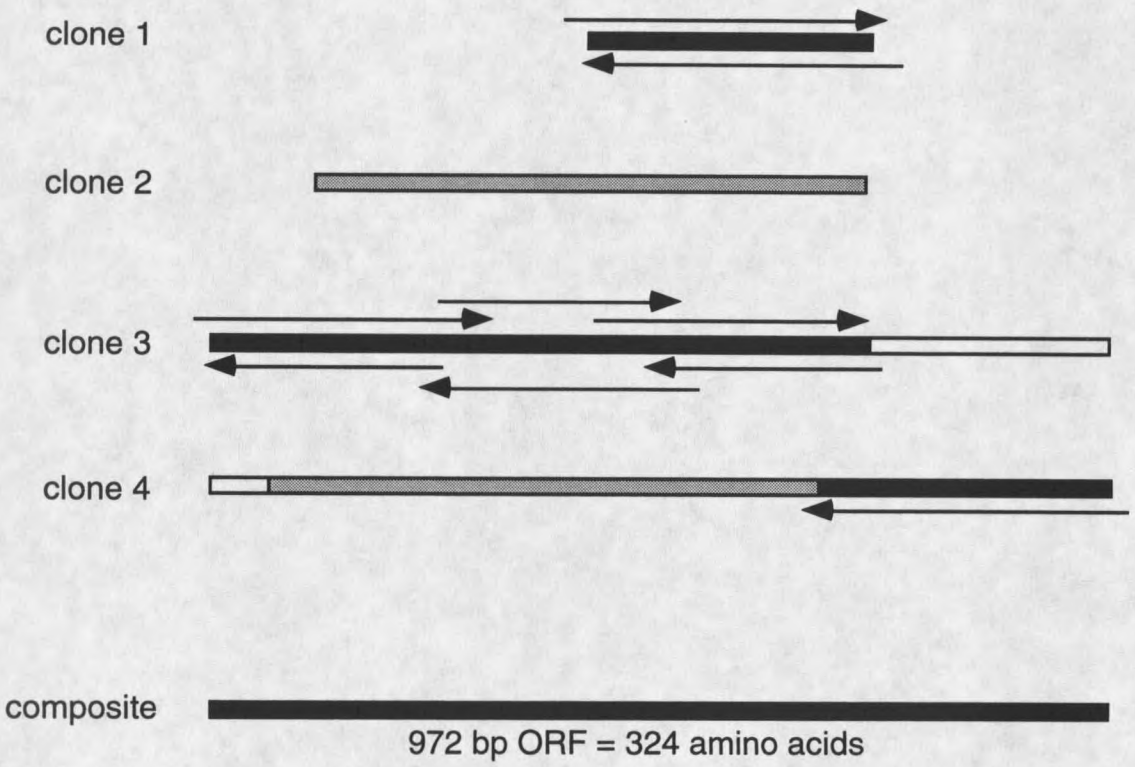


Figure 3. cDNA from clones 1, 2, 3 and 4 are each represented by a horizontal bar. Black portions of each bar represent cDNA that was previously sequenced (Rognlie, 1992). Gray portions were not sequenced in previous work. White portions represent the incomplete 3' and 5' ends of the original clones 3 and 4. Clone 1 is the 300 bp cDNA used as a probe to identify clones 2, 3 & 4. A composite cDNA using sequence from the 879 bp clone 3 and limited sequence from clone 4 represents a 972 bp ORF coding for 324 amino acids. The composite ORF represents the status of this project at the start of research discussed in this thesis. Arrows indicate the start of sequencing primers and the amount of sequence data obtained from each.

Project Goals and Approach

The work discussed in this thesis to extend the characterization of the lectin from grasshopper focuses on three goals: 1) to complete a clone 3-specific cDNA sequence containing a full-length ORF coding for a lectin from the grasshopper, *Melanoplus differentialis*, 2) to complete a clone 4-specific cDNA containing a full-length ORF for a second lectin and 3) construct a computer-based homology model for a CRD from grasshopper lectin using the deduced amino acid sequence and the crystal structure coordinates from rat MBP and human ESEL.

The cDNA sequences from clones 3 and 4 represent incomplete portions of two homologous lectins. Alignment of these sequences suggested clone 3 was missing ≈ 200 bp at the 3' end including a stop codon and polyA tail. Two strategies for obtaining the ≈ 200 bp to complete the clone 3 sequence were considered. One, the 14 remaining positive plaques would be screened using the polymerase chain reaction (PCR) and a clone 3-specific primer to determine whether a clone 3 sequence representing the missing 3' end was present. If successful, the combination of this newly obtained sequence with that known for clone 3 would provide two cDNAs that together represent the entire coding region for this lectin. Two, the alternative strategy was to use the Rapid Amplification of cDNA Ends (3' RACE) to isolate and clone a cDNA containing the complete 3' sequence using RT•PCR and polyA$^+$ mRNA. Clone 3-specific sequence obtained using this strategy, combined with the original 879 bp clone 3 sequence would similarly complete a cDNA representing a full-length ORF for this lectin.

The original clone 4 cDNA sequence was only partially complete at the start of this research (Figure 3). This clone 4 sequence would be completed first. Alignment of sequences for clones 3 and 4 suggested clone 4 was missing ≥ 69 bp at the 5' end

including sequence encoding the initiating Met, signal peptide and 2-4 N-terminal amino acids. As discussed for clone 3 above, two strategies for obtaining the clone 4-specific sequence representing the complete 5' end were outlined. Positive plaques would be screened using PCR and a clone 4-specific primer to determine whether the 5' sequence for this clone is present. If successful, the combination of this sequence with that known for clone 4 would provide two cDNAs that together represent the entire coding region for a second lectin from grasshopper. The alternative strategy was to use 5' RACE to isolate and clone a cDNA containing the complete 5' sequence. Similarly, this cDNA would be used in combination with that known for the original clone 4 to complete a sequence containing a full-length ORF for a second lectin.

The deduced amino acid sequences for clones 3 or 4 would provide complete sequence each containing two carbohydrate recognition domains (CRDs). The 3D structure for the primary sequence representing the C-terminal CRD from clone 3 would be predicted based on sequence homology with two reference structures, MBP and ESEL. Structurally conserved regions between crystal structures for MBP and ESEL were defined and these coordinates used to build the 3D structure for the grasshopper lectin (GHA) sequence. The reasonableness of the final model was assessed using Ramachandran plots and 3D Profiles (Luthy, 1992).

METHODS

Standard Molecular Biology Techniques

Competent Cells and Cell Transformation

Recombinant plasmid DNA was amplified in *Escherichia coli.*(*E. coli*; JM101, JM109) rendered competent by $CaCl_2$ (Sambrook, 1989). A single colony was picked from minimal media (1X M9 salts, 0.1 M glucose, 0.1 M vitamin B_1, 0.2 M $MgSO_4$ & 0.01 M $CaCl_2$) and grown to OD=.25-.40 in 50 ml of LB medium at 37°C. Cells were isolated by centrifugation (4000 g) for 10 min at 4°C. The pellet was resuspended in 10 ml of ice-cold 0.1 M $CaCl_2$ and allowed to stand on ice for several minutes. The cell suspension was centrifuged as above and the pellet resuspended in 2 ml of ice-cold 0.1 M $CaCl_2$. A 200 µl aliquot was transformed with ≈ 25 ng (volume < 10 µl) of a recombinant plasmid. Following a 30 min incubation period at 0°C, cells were transferred to a 42°C water bath for 90 sec and immediately transferred back to ice for several minutes. 800 µl SOC medium (2% bacto-tryptone, 0.5% bacto-yeast, 8 mM NaCl, 2.5 mM KCl & 20 mM glucose) was added and the cells incubated 45 min at 37°C. 25-100 µl aliquots were plated on LB-AXI plates (100 µg/ml Ampicillin, 80 µg/ml 5-bromo-4-chloro-3-indolyl-β-D-galactoside, X-Gal & 0.5 mM isopropyl-β-D-thio-galactoside, IPTG) and grown inverted overnight at 37°C.

Promega Magic™ Mini-Prep

The Magic™ Mini-Prep DNA purification system (Promega) is a resin/adhesion-based strategy which requires approximately 15 minutes to prepare sequenceable plasmid DNA from 1.5-3 ml of an overnight culture of transformed bacteria. The resin is a proprietary formulation that binds DNA, leaving associated impurities unbound and readily washed away. The DNA was eluted from the resin using TE buffer (pH=8) or water. Concentration was measured using OD_{260} (1 OD = 50 μg/ml). The purity of each prep was evaluated using the ratio of $OD_{260/280}$ where ≥ 1.8 was considered sufficient for sequence analysis.

Maxi-Prep of Plasmid DNA

A single bacterial colony containing the recombinant plasmid of interest was grown to OD_{600}=0.6 at 37°C in 30 ml LB medium containing 100 μg/ml Ampicillin. This culture was used to inoculate 200-500 ml of LB medium containing 100 μg/ml Ampicillin. Incubation was continued overnight at 37°C. Cells were harvested by centrifugation at 4000 g for 15 min and the pellet resuspended in 15 ml of buffer (50 mM Tris-HCl, 10 mM EDTA & 100 μg/ml RNAase A, pH=7.5). Cells were lysed using 15 ml 0.2 M NaOH (1% SDS). The lysate was neutralized using 15 ml potassium acetate (2.55 M $K^+CH3COO^-$, pH=3.5). This solution was centrifuged at 4000 g for 15 min and the supernatant filtered through cheesecloth. Plasmid DNA in the filtrate was recovered by precipitation with 0.6 volumes of isopropanol. The DNA pellet was washed with 70% ethanol, dried and resuspended in 1 ml of TE buffer.

Cycle Sequencing

A Sequitherm™ (Epicentre Technologies, Madison WI) cycle sequencing kit was used according to manufacturers protocol. A minimum of 500 fmols of template DNA was

used in each reaction. A "premix" of 15 pmol of primer, 10 μCi of α[^{35}S] dATP, target DNA, 10X sequencing buffer and 1 U of Sequitherm™ was combined in a volume of 16 μl. Four (4) μl of this "premix" was added to each of 4 extension/termination mixtures (e. g. dT/ddT, dC/ddC etc.). Tubes were overlayed with mineral oil, denatured for 5 min at 95°C and subjected to 35 cycles of amplification. One cycle was 30 sec at 95°C and 1 min at 70°C. Stop solution, 3 μl was added immediately following the final cycle (95% formamide, 20 mM EDTA, 0.05% bromophenol blue and 0.05% xylene cyanol FF).

Sanger Nucleotide Sequencing

Determination of cDNA nucleotide sequence was carried out using the Sanger dideoxy nucleotide-mediated chain termination method (Sanger, 1977). Sequenase ver 2.0 DNA polymerase (United States Biochemical, Cleveland OH) was used in all reactions according to manufacturers protocol. DNA (3-5 μg) was denatured in 0.2 M NaOH and 0.1 mM EDTA using a 30 min incubation at 37°C. The solution was neutralized with 0.1 volume 3 M sodium acetate and the DNA precipitated with 3 volumes of absolute ethanol at -70°C for 15 min. The dried pellet was brought up in 7 μl of H_2O to make ready for sequencing. The radio-label used in all sequence reactions was α[^{35}S] dATP (10 mCi/ml, Spec Activity ≈ 1150 Ci/mmol; Du Pont Company, Wilmington DE). All templates were sequenced completely in both directions.

Polyacrylamide Gel Electrophoresis

Polyacrylamide electrophoresis gels (0.4 mm x 21cm x 50cm) and a nucleic acid sequencing cell (Bio-Rad Laboratories, Richmond CA) were used to separate DNA fragments. Gels were 8% acrylamide and 8 M urea in 1X TBE (0.09 M Tris-Base, 0.09 M Boric Acid & .004 M EDTA, pH=8). Two different 2.5 μl aliquots of each reaction were loaded ≈ 3 hrs apart for a total of 6 hrs. This prevented short fragments representing

sequence near the primer from running off the gel and resolved longer fragments representing sequence further from the primer. The gel was maintained at 50-55°C using a constant 46 W of power over the duration of the electrophoresis. Gels were fixed in a solution of 10% methanol and 10% acetic acid for 15 min and transferred to Whatmann #3 chromatography paper (Whatmann International LTD, Maidstone UK). A BioRad® Model 583 Vacuum Gel Dryer was used to dry gels for ≈ 1.5 hrs at 80°C. Dried gels were exposed to X-ray film (Kodak X-OMAT AR, 35 cm x 43 cm) for 24 to 48 hours. Sequence data were analyzed manually using a light box and magnifying viewer where necessary.

Storage and Analysis of Sequence Data

GenePro™ (Riverside Scientific, Seattle WA) sequence analysis software was used for storage, manipulation, alignment and comparison of all sequence data. This package was used on a Zenith Data Systems desktop PC. In addition to storage and alignment facilities, the software features translation, open reading frame (ORF) and restriction site analysis.

Selection and Design of Sequence Primers

Design of new sequencing primers and the selection of optimal priming sites followed three general guidelines. All primers for this project were designed with attention to the following criteria (Berger, 1987); 1) ≈ 50% total GC content 2) 3' terminus is a G or a C and 3) ≤ 50% identity specifically within the last 10 (3') bases of the primer. Oligo™ (National Biosciences Inc., Plymouth MN) was used to calculate melting temperatures (T_m & T_d) and analyze primer sequences for significant hairpin loops or the potential to form 'primer dimers'. Sequencing primers used to "walk" the length of a target

DNA were also selected to insure ≈ 40 bp of overlap between known and newly obtained sequence.

Preparation of Custom Ordered Primers

The sequence and commercial source of each primer used to sequence clones 2, 4 and the 3' and 5' RACE products are reported in Table 1. The relative location of the these primers on clone 4 and the 3' and 5' RACE products is illustrated in Appendix A. Primers 5'γ, 5'δ, 3'β, 3'γ and 3'δ were used to complete the sequence of the original 1093 bp clone 4. These oligos were synthesized by the department of Veterinary Molecular Biology (MSU-Bozeman) and lyophilized. Primers were resuspended in 500 µl of TE buffer, vortexed and spun at 13000 g for 5 min. The supernatant was removed and sterile filtered using a 0.22 micron disposable filter (Gelman Sciences, Ann Arbor MI). The DNA was then precipitated in 3 volumes of absolute ethanol. Following centrifugation at 13000 g (15 min), the pellet was washed with 500 µl of 70% ethanol and centrifuged at 13000 g for 5 min. The final pellet was dried and diluted to 200-400 µl of TE buffer. $OD_{260/280}$ measurements were used to establish purity and concentration. An aliquot of each primer was diluted to 1 pmol/µl for use in sequencing reactions.

Primers labeled 3132, 3152 and 3153 were used to sequence the RACE product containing the complete 3' end of the original clone 3. These oligos were obtained from National Biosciences Inc. (Plymouth MN) in lyophilized form. Each was diluted in 300 µl of H_2O and concentration evaluated via optical density as described above. Aliquots were diluted to a working concentration of 1 pmol/µl.

Table 1. Nucleotide sequences for primers used to sequence cDNA of clones 2, 4 and the 3' and 5' RACE products.

Primer Name	Nucleotide Sequence	Source
λF	5'-GGT GGC GAC GAC TCC TGG AGC CCG-3'	[1]Promega
λR	5'-TTT GAC ACC AGA CCA ACT GGT AAT-3'	[1]Promega
SP6	5'-GAT TTA GGT GAC ACT ATA G-3'	[1]Promega
T7	5'-TAA TAC GAC TCA CTA TAG GG-3'	[1]Promega
5B	5'-ACT GTG TTT ACG TCG ACG AC-3'	[2]VMB
5γ	5'-ACT GTG TTT ACG TCG ACG AC-3'	[2]VMB
5δ	5'-CTA CCT GAC AGG ATT CAC AG-3'	[2]VMB
3β	5'-GTT GTT ATT TGG CTC ACC AG-3'	[2]VMB
3γ	5'-GGC GTG CAC CTT GTA GAA GC-3'	[2]VMB
3δ	5'-GAT CTG TGA TTC CGA TGT TG-3'	[2]VMB
3132	5'-CAG CTC CAC TTC GCA GCT CC-3'	[3]NBI
3152	5'-TCT ACA AGG TGC ACG CCG AG-3'	[3]NBI
3153	5'-TTC ATC TGT GAG ATA GCA CC-3'	[3]NBI
879PCR	5'-CTT CAA GTT AGG GTT TGG GG-3'	[2]VMB
1150PROB	5'-CCC CTT CCT CTT TAT GCT CT-3'	[2]VMB
4036	5'-GCA CAG TTT TTC GGG TTC AC-3'	[3]NBI

[1]Promega Corporation, Madison WI, [2]Dept of Veterinary Molecular Biology, MSU-Bozeman, Bozeman MT, [3]National Biosciences Incorporated, Plymouth MN

Screening Remaining Positive Clones

PCR Amplification of Phage DNA

Phage suspensions (Rognlie, 1991) representing each of the 14 remaining positive clones were stored in SM buffer (0.01 % gelatin, 50 mM Tris-HCl, 100 mM NaCl, 8 mM $MgSO_4$) and 0.03% $CHCl_3$. A single phenol extraction followed by ethanol precipitation served to concentrate the DNA and eliminate residual buffer prior to amplification. The phage suspension was extracted with an equal volume of phenol followed by 500 µl of choroform/isoamylalcohol (24:1). After centrifugation at 13000 rpm for 2 min, the aqueous layer was transferred to a new tube and the DNA precipitated in 0.1 vol 3 M

$NaCH_3COO$ (pH=5.5) and 3 vol of ethanol. The final pellet was diluted to 10 µl in TE buffer. One to five (1-5) µl was used directly in PCR.

PCR was performed using AmpliTaq™ Gold DNA polymerase (Perkin-Elmer Applied Biosystems Division, Foster City CA). All reactions were completed according to manufacturers protocol. A single 50 µl reaction contained 0.02 uM of each primer, 0.2 mM each of 4 dNTPs, 1X amplification buffer and 1-5 µl target DNA as discussed above. Reactions were denatured at 95°C for 10 min and then subjected to 30 cycles of amplification using an automated thermal-cycler (Perkin-Elmer Applied Biosystems). One cycle consists of 45 sec at 95°C, 45 sec at 55°C and 1 min at 72°C. PCR products are analyzed by size following electrophoresis on 1.2% agarose gel containing ethidium bromide.

Phage Growth and Amplification

A single bacterial colony of Y1090 E. coli (Promega Corporation, Madison WI) was grown overnight at 37°C in 5 ml LB medium containing 1% Maltose and 10 mM $MgSO_4$. In a 15 ml screw-cap tube 100 µl of the overnight culture was combined with ≈ 100 µl of phage suspension and 4 ml molten LB/agarose. This mixture was poured over an LB plate containing hardened bottom agar. The top agarose was allowed to harden and the inverted plate incubated at 37°C overnight or until plaques appeared. To replace original phage suspensions depleted during screening, single plaques were picked and stored at 4°C in 1 ml SM buffer containing 0.3% $CHCl_3$. For phage preps, plaques were allowed to grow to confluence. 3 ml SM buffer was poured over the plate and the top agarose was gently scraped into a 50 ml centrifuge tube. This mixture of top agar, SM buffer and phage DNA was incubated for 30 min at room temperature and centrifuged (10000 g) for 10 min at 4°C. The solid top agar is removed from the mixture by transferring the supernatant

containing the phage to a new microfuge tube. The phage was treated with a nuclease mixture (RNAase A and DNAase I at 0.25 mg/ml, 150 mM NaCl and 5% glycerol) for 15 min at 37°C and precipitated in polyethylene glycol (33% PEG 8000, 3.3 mM NaCl) for 30 min on ice. Following centrifigation for 10 min at 10000 g, the pellet was resuspended in 500 µl buffer (150 µM NaCl, 40 mM Tris-HCl, 10 mM $MgSO_4$, pH=7.4). This mixture was purified using a purification matrix (Promega) and eluted in 100 µl H_2O.

Rapid Amplification of cDNA Ends: 3' RACE

Isolation of mRNA

Total RNA was isolated from freshly dissected grasshopper fat body using Tri-Reagent™ (Molecular Research Center Inc., Cincinnati OH). Aqueous reagents, glassware and microfuge tubes were made RNAase-free by treatment with diethylpyrocarbonate (DEPC). Dissecting instruments were baked in a conventional oven at 350°F for ≥ 5 hrs. Adult grasshoppers were washed by immersion in a mild detergent followed by 5% bleach and finally in distilled water. Subjects were pinned and dissected through a ventral incision from the anus to the head. The entire gut was removed with forceps by clamping at the posterior and carefully lifting toward the head. Fat body, which lines the internal walls of the cuticle, was gathered and removed with a forceps. Tissue was immediately "snap-frozen" on a tared container in contact with a solid cube of dry ice. Roughly 100 mg of fat body can be obtained from 2 grasshoppers. Fat body was homogenized in Tri-Reagent™ (5-100 mg/1 ml) using a 10 ml manual tissue homogenizer and allowed to stand for 5 min at room temperature to allow for separation of nucleoprotein complexes. Addition of chloroform (0.2 ml/1ml Tri-Reagent™), vigorous shaking (15 sec) and subsequent centrifugation at 12000 g induces the separation of the mixture into three phases. The lower phenol-chloroform phase contains proteins while the interphase contains DNA and

the upper aqueous phase contains only RNA. This phase was transferred directly to an oligo-(dT)-cellulose affinity chromatography column (MRC, Cincinnati OH). The column was washed twice with 1 ml binding buffer (0.5 M LiCl, 50 mM sodium citrate, 3 Na^+ $C_5H_7O_6^{-3}$ & 0.1 % SDS) and eluted with 0.5 ml elution buffer (1 mM $Na^+_3C_5H_7O_6^{-3}$, 0.1 % SDS). PolyA^+ mRNA was recovered by precipitation in two volumes of isopropanol at room temperature. Following centrifigation at 10000 g for 15 min, the pellet was air dried and brought up in 20 µl H_2O. Final concentration was measured using optical density at 260 nm (1 OD_{260}=40 µg/ml).

cDNA Synthesis

Synthesis of cDNA was done according to the protocol for 3' RACE (Life Technologies™, Gaithersburg MD). This protocol is based on a strategy first outlined by Frohman (1989). The complete procedure is outlined in Figure 4. The 20 µl synthesis mixture contained 1 µg of polyA^+ mRNA, 20 mM Tris-HCl, 50 mM KCl, 2.5 mM $MgCl_2$, 100 µg/ml BSA (Bovine Serum Albumin), 10 mM DTT, 500 nM adapter primer (AP) and 500 µM each of 4 dNTPs. Five (5) units of SuperscriptII™ reverse transcriptase were added and the reaction incubated at 42°C for 30 min. The mRNA template strand was digested with 10 U of RNAase H for 30 minutes at 37°C. Two (2) µl of this reaction were used in RT-PCR.

RT-PCR of the 3' End of Clone 3

The 3' end of clone 3 was amplified using a gene-specific primer (GSP=879PCR) and the universal amplification primer (UAP). The 50 µl reaction mixture contains 2 µl of newly synthesized first-strand cDNA and 10 mM Tris-HCl, 50 mM KCl, 2.5 mM $MgCl_2$, 100 µg/ml BSA , 200 nM gene-specific primer (879PCR), 200 nM universal amplification

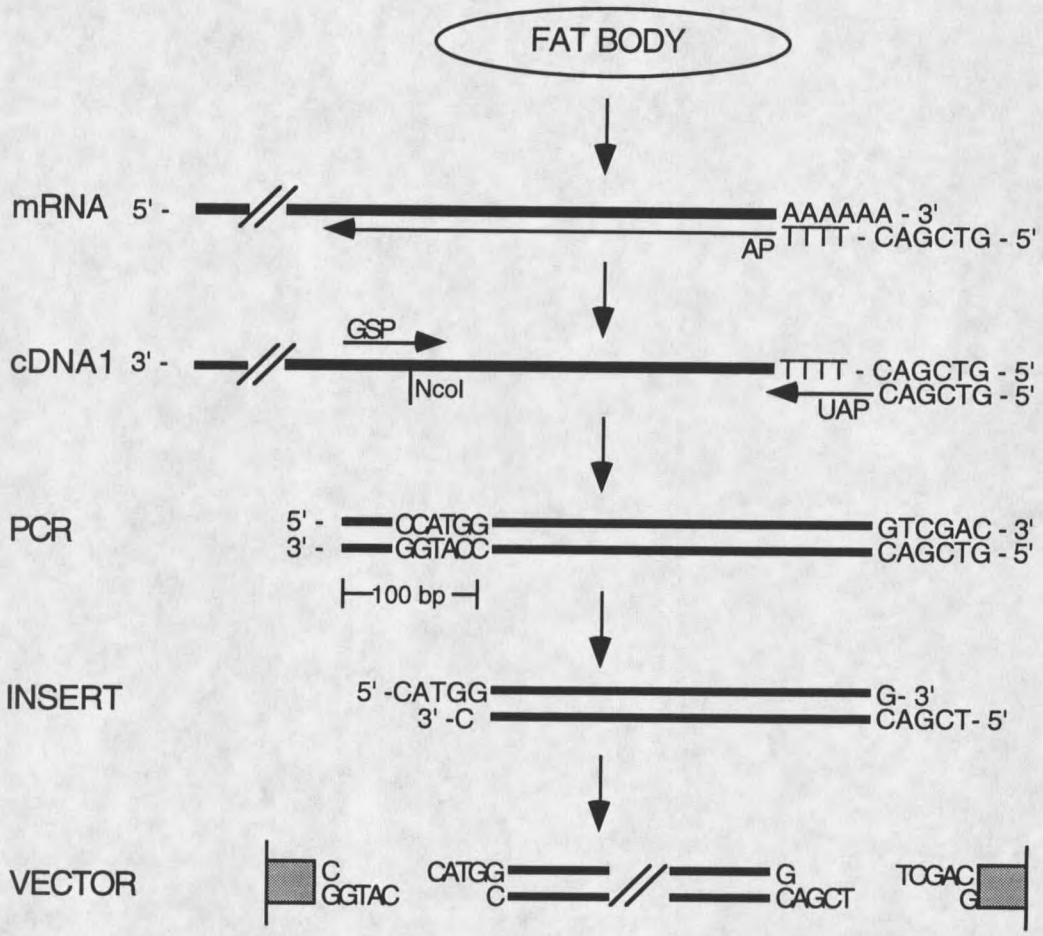


Figure 4. The 3' RACE strategy used to isolate and clone the 3' end of clone 3. mRNA is isolated from fat body and first strand cDNA is synthesized using a polyT primer and SuperscriptII™ reverse transcriptase. The primer contains a SalI restriction site. mRNA is digested with RNAase H and PCR is completed using the gene-specific primer (GSP) '879PCR' and a universal amplification primer (UAP). The 800 bp PCR product is digested with NcoI and SalI restriction enzymes creating 'sticky ends' for ligation into pGEM5fz(+) plasmid vector. Restricting the original 800 bp product with NcoI cleaves 100 bp from the fragment making the insert ≈ 700 bp.

primer (UAP), 200 μM dNTPs and 0.5 U Taq polymerase. This mixture was subject to 35 cycles of the following amplification sequence: 45 sec at 94°C, 45 sec at 60°C and 2.0 min at 72°C. Amplification products were visualized in 1.2% agarose gels containing ethidium bromide.

Cloning the 3' End of Clone 3

A search of the known sequence for clone 3 revealed an NcoI restriction site 388 bases from the 3' end of this sequence. The UAP primer contains a SalI restriction site. Restriction of this RT-PCR product with these enzymes produces "sticky ends" for ligation into a complementary plasmid vector. pGEM-5zf(+) (Promega Corporation, Madison WI) contains both NcoI and SalI sites and the amp^r -lacZ genes for blue/white color selection of transformed cells. 1 U of T4 DNA ligase and a 3:1 molar ratio of insert to vector (50 ng/35 ng) in a total volume of 10 μl was incubated overnight at 15°C to complete the ligation reaction. Competent *E. Coli* (JM101) were transformed with 2 μl of this mixture and plated on AXI Plates.

Rapid Amplification of cDNA Ends: 5' RACE

Isolation of mRNA.

PolyA$^+$ mRNA was isolated from freshly dissected grasshopper fat body using Tri-Reagent™. This protocol is discussed in detail on page 26.

cDNA Synthesis

cDNA synthesis was done according to the protocol for 5' RACE (CLONTECH™ Laboratories Inc., Palo Alto CA). The complete 5' RACE procedure is outlined in Figure 5. The 30 μl synthesis reaction contained 2 μg of polyA$^+$ mRNA, 40 units RNAase

inhibitor, 25 units MMLV RT (Moloney Murine Leukemia Virus Reverse Transcriptase) and was 0.33 mM each of 4 dNTPs, 1X RT buffer and 0.03 μM polyT primer. This mixture was incubated at 52°C for 30 min. Second strand synthesis was completed using a 10 μl aliquot of the 1st strand synthesis reaction and an "enzyme cocktail" containing RNAase H, *E. coli* DNA polymerase I and *E. coli* DNA ligase. The mixture was incubated at 16°C for 2 hr. Double stranded cDNA was blunt-ended by adding 5 U of T4 DNA polymerase and incubating for an additional 30 min at 16°C. cDNA was isolated using Prep-a-Gene™ DNA purification matrix (BioRad Corp., Hercules CA).

Ligation of DS-Anchor Sequences to cDNA

The double-stranded anchor sequence (AS) was ligated to each end of the cDNA. The 10 μl ligation reaction contained 2 μl of double-stranded cDNA discussed above, 4 pmol of anchor sequence, buffer and 10 U of T4 DNA Ligase. This mixture was incubated at room temperature for 24 hrs. The anchor ligated-cDNA was diluted 1:25 prior to use in PCR.

PCR Amplification of the 5' End of Clone 4

The 5' end of clone 4 was amplified using a gene-specific primer (4036) and the anchor primer (AP). A 50 μl reaction containing 2 μl of the anchor-ligated cDNA (1:25) and 1 U AmpliTaq Gold™ (Perkin Elmer Applied Biosystems Division, Foster City CA) contained 10 mM dNTPs and 10 μM of each primer. This mixture was heated 10 min at 95°C and subjected to 35 cycles of amplification (45 sec at 94°C, 45 sec at 55°C, 2 min at 72°C). Amplification products were visualized using a 1.2% agarose gel containing ethidium bromide.

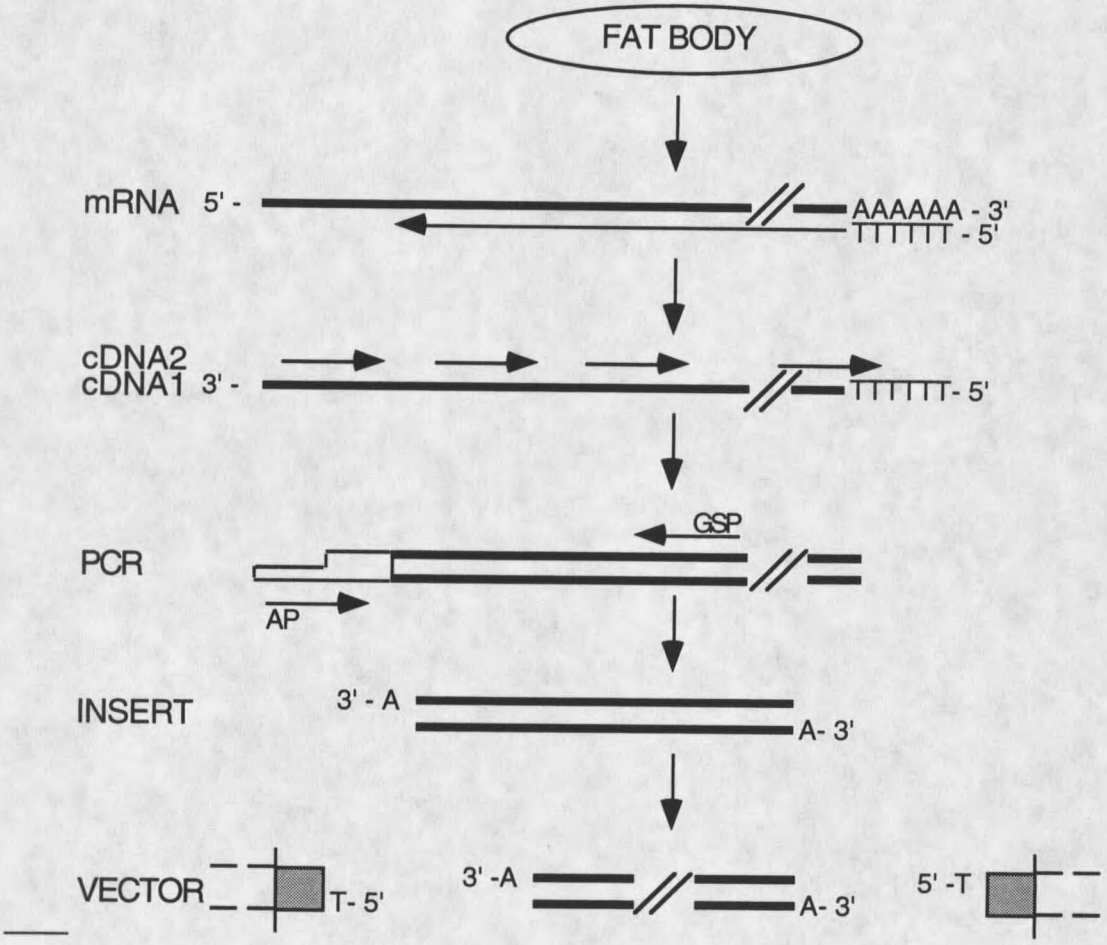


Figure 5. The 5' RACE strategy used to isolate and clone the 5' end of clone 4. mRNA was isolated from fat body and first strand cDNA synthesized using a polyT primer and MMLV reverse transcriptase. Following second-strand cDNA synthesis, the double-strand cDNA was blunt ended using T4 DNA Polymerase and isolated using Prep-a-Gene™. Double-stranded 'anchor sequences' were blunt-end ligated to the DNA and PCR completed using the gene-specific primer (GSP) '4036' and an anchor primer, 'AP'. The PCR product was ≈ 1100 bp and contains sequence representing the 5' end of the original clone 4. This product was ligated directly into pGEM-T vector.

Cloning 5' End of Clone 4

The pGEM-T plasmid vector (Promega, Madison WI) contains amp^r and lacZ genes for blue/white color selection of transformed cells. This vector was commercially modified to have overhanging thymine (T) residues for the direct sticky end ligation of PCR products with complimentary adenine residues (A). This "A/T" cloning approach takes advantage of the non-template directed addition of 'A' by Taq polymerase to the 3' end of the DNA. 1 U of T4 DNA ligase and a 3:1 molar ratio of PCR product (insert) to vector (50 ng/35 ng) in a total volume of 10 µl was incubated overnight at 15°C to complete the ligation.

Homology Model Development

A molecular model for a CRD from a grasshopper lectin was generated using InsightII™ v2.3.5 and Homology v2.3 (Biosym Technologies Inc., San Diego CA). This software package was employed using a Silicon Graphics 4D/35 workstation running IRIX v5.0. The deduced amino acid sequence from clone 3 encoding the C-terminal CRD was used to represent the model sequence. This CRD has the highest sequence homology with each of MBP (31%) and ESEL (30%). The CRD begins at Ala191 and ends at Cys300. The reference structure coordinates for rat MBP and human ESEL were obtained from the Brookhaven Protein Data Bank (Brookhaven NY). A structure for MBP is available in complex with oligomannose at 1.7 Å resolution (Wies, 1992;file=pdb2msb.ent). The C-type CRD is represented by 102 residues beginning at His115 and extending through Cys 217. The coordinates for ESEL have been determined at 2.0 Å resolution (Graves, 1994;file=pdbesel.ent). The C-type CRD in this structure is represented by 110 residues from Thr7 to Cys117. ESEL was used as the second reference structure during construction of the model. No other crystal structures for C-type lectins are available.

Conserved Regions Within Reference Proteins

Structurally conserved regions (SCRs) within MBP and ESEL were defined in two steps. Sequences were first aligned automatically using GenePro™ to maximize alignment of identical residues. The result was manually modified by introducing gaps to insure alignment of the twelve (12) strictly conserved residues in this region of the sequence (Drickamer, 1993). These residues include Gly158, Trp181, Pro186 and known Ca^{2+} ligands Glu185, Asn187, Asp188, Glu193 and Asp194 in addition to four Cys residues 128, 195, 209 and 217 defining the C-type architecture. Gaps were introduced between regions of defined secondary structure, based on the known structures for each reference protein. SCRs were then defined by superimposing equivalent C_α backbone atoms of the two reference structures over regions of the sequence representing known secondary structure elements. Structural conservation was defined using root-mean-square (RMS) differences (Bajorath, 1993). Any region of the sequences where the RMS deviation between the two references was ≤ 0.75 Å was defined as an SCR and used in construction of the final CRD model.

Model Sequence-Reference Structure Alignment

The grasshopper lectin sequence defining the model CRD extends from Ala 191 to Cys300. Corresponding residues in rat MBP and ESEL are His115 to Cys217 and Thr 7 to Cys117, respectively. Alignment of the model CRD from grasshopper lectin and the C-type lectin domains of MBP and ESEL were generated using GenePro™. Automatic alignment with each reference protein separately provided alignments based on maximum identity. These alignments guided the manual adjustments needed to introduce gaps in the model sequence to insure that the twelve (12) strictly conserved residues were aligned and that gaps were not introduced in the GHA sequence where regions of formal secondary structures may exist.

Construction of the GHA Homology Model

The GHA model CRD was constructed using the atomic coordinates from rat MBP and ESEL. In SCRs, model coordinates were assigned from the reference structure with the higher sequence homology to the GHA sequence. The loops between SCRs were modeled using two different strategies. Loops having the same number of residues as the corresponding loop in either (or both) reference structure(s), were modeled using coordinates directly from MBP or ESEL (Villoutreix, 1994). As with SCRs, reference coordinates (either MBP or ESEL) were selected based on homology with the model sequence. Coordinates for loops containing insertions or deletions were extracted from a structurally nonredundant database of high-resolution atomic structures using a loop search algorithm (Protein Data Bank, Brookhaven NY). Potential loops were evaluated using RMS differences following superposition of C_α atoms at the junction of the loop with corresponding residues defining the flanking termini of SCRs on each side of the loop. The best loop is one in which the SCR on either side will superimpose onto adjacent regions of the test loop with a low RMS deviation. A suitable loop conformation is defined as one that does not create unfavorable steric clashes with the rest of the structure or introduce unreasonable peptide bond dihedral angles at the splice-junction.

Side chains in the SCRs were mutated to reflect the sequence of GHA from grasshopper. The stereochemistry of new side-chain residues in the model were adjusted manually if necessary to avoid the introduction of non-allowed torsional angles and severe van der Waals overlap.

GHA Model Refinement

To relax backbone structures and refine stereochemical contacts, the initial model was subjected to constrained energy minimization calculations using Discover™ (Biosym Technologies, Inc., San Diego, CA). All calculations were carried out using the CVFF

forcefield (Consistent-Valence Forcefield) and an 8 Å cutoff distance for non-bonded interactions (Castonguay, 1995). All calculations were done in vacuo, without water.

Initially, harmonic constraints of 100 kcal/mol/Å^2 were applied to all backbone atoms except the loop regions, and constrained energy minimization was carried out for 1000 iterations using a steepest descents algorithm. This was followed by 1000 iterations using a conjugate gradient algorithm until the RMS derivative of the energy function was less than 5 kcal/mol/Å^2. This was done to refine the loop-SCR splice junctions. To refine loop backbone atoms, harmonic constraints of 30 kcal/mol/Å^2 were placed on all loop side chains and minimization was carried out for 1000 iterations using steepest-descents, or until the RMS derivative of the energy function was less than 5 kcal/mol/Å^2. To complete the model refinement, unconstrained minimization was done using a conjugate gradient algorithm until the RMS derivative of the energy function was $\leq .2$ kcal/mol/Å^2 (Castonguay, 1995).

Assessment of the GHA Model CRD

Assessment of the model seeks to answer the following question: Is the 3 dimensional structure for the model a reasonable representation of the primary sequence? Evaluation is based on the comparison of assessment measures made on the model itself as well as the two reference structures. The results of assessment for the model should compare favorably to those of the reference structures to conclude the final CRD model is a reasonable 3 dimensional representation of the sequence.

Initial assessment of the CRD models was carried out using a Ramachandran plot of the Psi (ψ) and Phi (ϕ) angles for each residue in the structure. This assessment is based on two related properties, 1) the spatial exclusions placed on conformations of any polypeptide based on a finite range of ϕ and ψ, and 2) the comparison of the ϕ and ψ

angles in the model structure which fall within 'allowed' regions of the plot vs those which fall outside these regions for both the model and the reference structures. Further comparison of these angles to other homology models built based on the same crystal structure also provide an assessment tool to define the reasonableness of the final structure. Plots were generated using Excel 5.0 (Microsoft Corporation, Bellevue WA) on a 33MHz/486-based PC running Windows 3.1. A ".txt" file of ϕ-ψ angles was first generated using the Biopolymer™ module (Biosym Technologies, San Diego CA) of InsightII™. A check of dihedral angles in the model to insure adherence to the $\pm 180^{\circ}$ required for a trans peptide bond was also done using the Biopolymer™ module within InsightII™. Dihedral angles can be used to evaluate loop/SCR splice-junctions and determine the quality of the loop coordinates selected in the early steps of the model building process. Poor dihedral angles may indicate the need for selection of different loop coordinates.

Assessment of Sequence-Structure Compatibility

The sequence-structure compatibility of the GHA model was assessed using the a three-dimensional profile analysis as implemented by 3D-Profiles™ (Eisenberg, 1991). 3D profile assessment is a quantitative evaluation of the model CRD and seeks to assess the compatibility of the model sequence with the 3D structure it has been folded into (Luthy, 1992). This represents the inverse protein folding problem (Godzik, 1992). The strategy is based on conversion of three-dimensional residue environments found in the final model into a one-dimensional sequence of environments. Eighteen categories of environments have been defined based on the area of the side chain that is buried, the fraction covered by polar atoms and the type of secondary structure it occupies (Bowie, 1991). The 3D profile is constructed as follows, 1) each sidechain in the model is assigned a category and 2) a 3D probability score representing the probability of finding that sidechain in its assigned

category, 3) the sequence of probability scores are summed to provide the S_{score} and/or plotted to provide the physical profile. A comparison of S_{scores} for the model CRD and each of the reference structures represents the assessment. The folded model sequence should provide scores compatible to each reference if the overall structure is a reasonable 3D representation of the sequence. Likewise, the physical profile plot of the sequence of 3D-1D probability scores should also be similar. 3D profile assessment of the model CRD in this project suggests the folded structure of the model is a reasonable 3D representation of the sequence.

RESULTS

A cDNA library was previously constructed from fat body mRNA using a λgtll expression vector (Hapner, unpublished). A 300 bp cDNA was isolated via an antigenic fusion protein, ^{32}P-labeled and used as a hybridization probe to screen the library. Seventeen plaques were identified and three (clones 2, 3 & 4) were isolated and subcloned These clones were diagrammed in Figure 3 (p 16). The 879 bp clone 3 cDNA contained an ORF encoding an initiating Met, a 20 residue signal sequence and 268 amino acids, but not a stop codon or 3' nontranslated sequence through the polyA tail (Rognlie, 1991). Limited cDNA sequence from clone 4 was combined with that of clone 3 to complete a hybrid cDNA representative of a full-length mRNA as defined by the presence of a start codon, stop codon and polyA tail. The overlapping sequence between each of clones 3 and 4 was not identical, suggesting each represented distinct mRNAs. Alignment of these sequences indicated clone 3 was missing ≈ 200 bp of 3' sequence representing a stop codon and polyA tail. Clone 4 was missing ≈ 69 bp at the 5' end representing 2-4 N-terminal amino acid residues, a signal peptide and start codon. Completion of these sequences and the use of the deduced amino acid sequence to construct a homology model for one CRD is reported below.

cDNA Sequence for Clone 2

Clone 2 was sequenced in one direction and determined to be 737 bp in length. Alignment of this cDNA sequence with that for clone 3 proved the two to be identical and thus representative of the same mRNA. This clone was not studied further.

cDNA Sequence for Clone 3

Search for the 3' end of Clone 3

The 3' end of the original 879 bp clone 3 was incomplete. Seventeen positive plaques were originally identified, but only three were subcloned and sequenced (clones 2, 3 & 4). The 14 remaining plaques were screened using PCR and a clone 3-specific primer to selectively amplify an insert(s) which contains the ≈ 200 bp representing the 3' end of clone 3. The size of a clone 3-specific insert containing the missing 3' sequence would be ≈ 800 bp, based on alignment with clone 4. Twelve of the 14 remaining positive plaques were amplified with the clone 3-specific 879PCR primer, but no products proved to be of sufficient size to contain the ≈ 200 bp representing the 3' end of this clone. Figure 6 illustrates the PCR products obtained for clones 1, 2, 5, 6a, 6b, 7c, 8, 9, 10, 11, 12, 13, 14 and 15 (Rognlie, 1991). Plaques 7c and 13 did not amplify using the clone 3-specific primer. All amplified products are ≤ 600 bp. These data suggest the clones do not contain the ≈ 200 bp that would complete the cDNA sequence for clone 3. It was concluded this sequence was not accessible in the current cDNA library using this method. The alternative 3' RACE strategy was employed to selectively isolate and clone this sequence from polyA$^+$ mRNA. This work is discussed below.

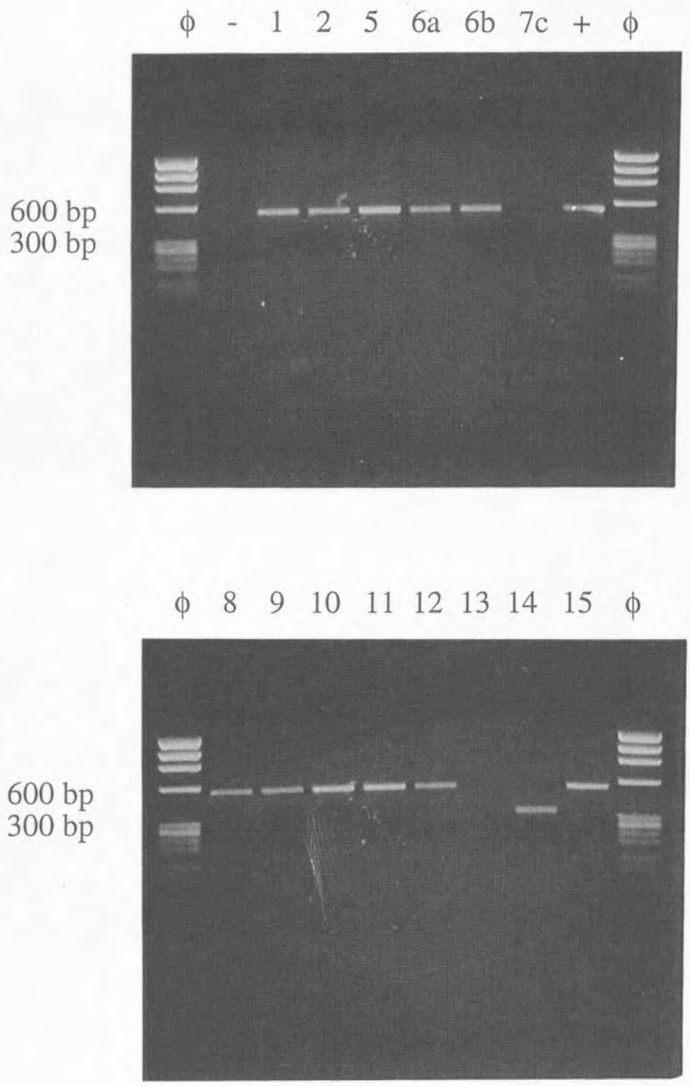


Figure 6. Screening the remaining 14 positive plaques using PCR with the gene-specific primer, 879PCR. ϕ = ϕx174/Hae III DNA ladder. "-" = negative control using 879PCR with a clone 4-specific target. The absence of an amplified product in this lane illustrates the clone 3-specificity of the 879PCR primer. "+" = positive control using the 879PCR primer to amplify the original 879 bp clone 3 insert from a recombinant plasmid. The ≈ 600 bp band in this positive control illustrates the expected size of a clone 3-specific PCR product without the 3' end. Plaques 1, 2, 5, 6a, 6b, 7c, 8, 9, 10, 11, 12, 13, 14, 15 are the fourteen positive plaques originally isolated by hybridization with the 300 bp cDNA.

Rapid Amplification of cDNA Ends: 3' RACE

mRNA Isolation. mRNA was isolated using Tri-Reagent™. A total of 1.5 g of fat body tissue was dissected and stored at -85°C. Approximately 350 µg of total RNA and 16 µg of polyA$^+$ mRNA was isolated from ≈ 100 mg fat body tissue ($OD_{260/280}$ = 1.7). Two (2) µg of polyA$^+$ mRNA was used per RT•PCR reaction.

RT•PCR Amplification. First strand cDNA synthesis and subsequent PCR amplification (RT•PCR) using the gene-specific primer (879PCR) and the universal amplification primer (UAP) produced a single band ≈ 800 bp as visualized by agarose electrophoresis and ethidium bromide staining. This band is pictured in each of lanes 1 and 2 of Figure 7. Estimated yield is ≈ 30 ng. The 800 bp product is large enough include the complete 3' sequence for the original clone 3.

Restriction Analysis. Verification of the clone 3-authenticity of the 800 bp RT•PCR product before cloning was done using restriction analyses with NcoI, XhoI and EcoRI restriction enzymes. These enzymes were selected based on the location of their restriction sites in the sequence known for the original clone 3. The ≈ 800 bp band was excised from the gel using Prep-a-Gene™ and subjected to separate restriction analyses. The results of these restriction analyses are pictured in Figure 8. For a clone 3-specific fragment, restriction with NcoI should cleave off 100 bp leaving a 700 bp product. Restriction with XhoI should cut the fragment approximately in half leaving two fragments of ≈ 440 bp and ≈ 390 bp and restriction with EcoRI should cut this fragment one time producing fragments of ≈ 500 bp and ≈ 350 bp. In each case, the predicted size products were obtained. These results verify the ≈ 800 bp 3' RACE product is clone 3-specific and contains sequence that represents the complete 3' end of the original clone 3.

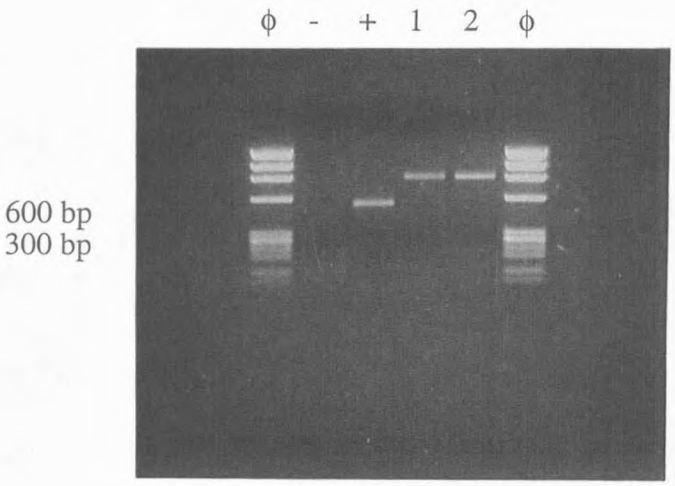
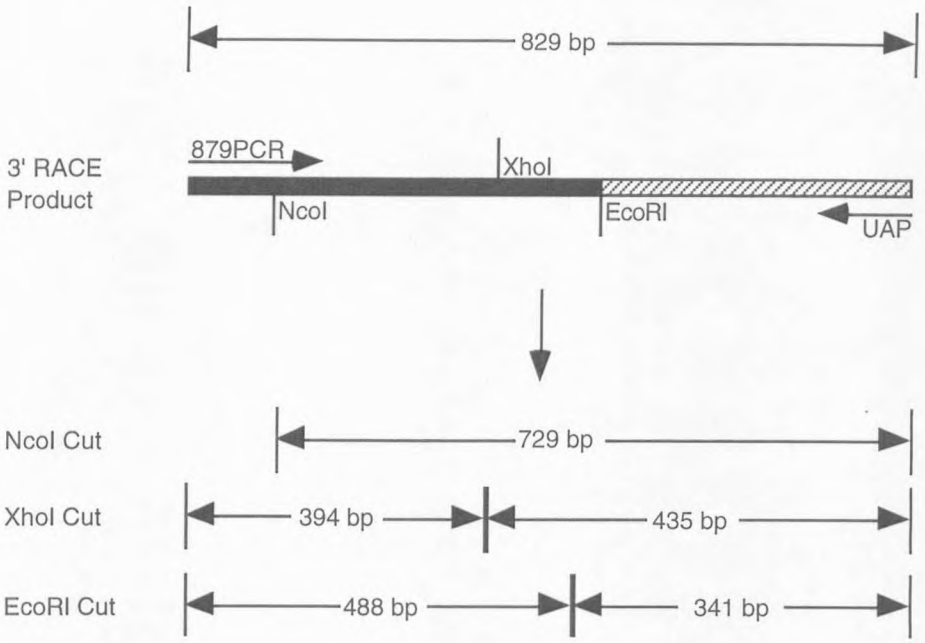


Figure 7. 3' RACE products obtained using 879PCR and UAP primers with 1st strand cDNA made directly from grasshopper fat body. ϕ = ϕx174/Hae III DNA ladder. "-" = negative control with 879PCR and UAP primers, but no DNA template. "+" = positive control using 879PCR and λF with clone 3 plasmid DNA as template. 1 and 2 are 3' RACE products. These products appear ≈ 800 bp in size, large enough to contain the complete 3' end of the original 879 bp clone 3.

Ligation into pGEM5fz(+) and Transformation

of Competent Cells

The ≈ 700 bp NcoI restricted product was excised from the gel and restricted with SalI, creating a second sticky end for cloning. This product was ligated into a pGEM5fz plasmid vector opened with NcoI and SalI. Transformation of competent cells and overnight growth on AXI plates revealed 176 blue and 25 white colonies. White colonies should harbor a recombinant plasmid based on α-complementation and blue/white color selection. Transformation efficiency was 5.34 x 10^6 colonies/μg DNA. Five white



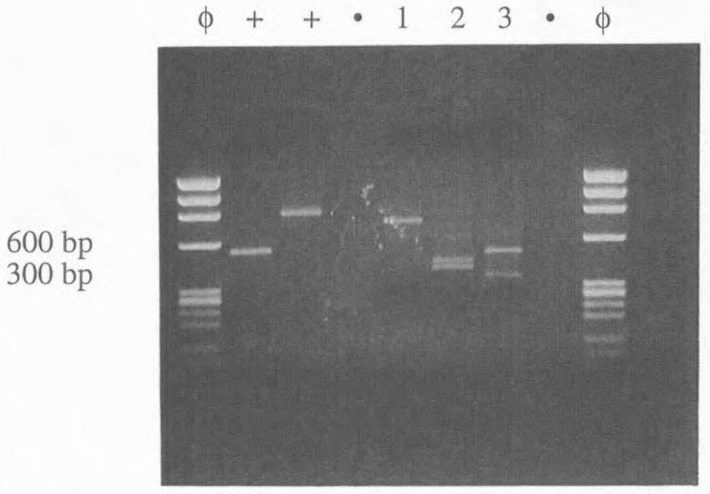


Figure 8. Verification of the clone 3-authenticity of the ≈ 800 bp 3' RACE product. This 829 product is represented at the top with a thick horizontal line. Black portions of the line represents sequence known from the original clone 3. Hatched portions represent the missing portion of clone 3 isolated in the RACE procedure. The location of NcoI, XhoI and EcoRI restriction sites are labeled. Expected product sizes for each restriction are illustrated with thin lines and arrows. Below, the actual results of the restriction analyses. ϕ=ϕx174 DNA ladder, '+'= PCR product using the 879PCR primer on clone 3 plasmid, '+'=the original 3' RACE product, •=empty lane, the 829 bp RACE product was restricted with 1=NcoI, 2=XhoI and 3=EcoRI.

colonies were picked and grown separately overnight in LB (100 μg/ml ampicillin). Roughly 10 μg of plasmid DNA for each colony was isolated from 1.5 ml of overnight LB culture. To obtain sufficient DNA for multiple sequence reactions, a larger prep of colony 5 was completed and yielded 178 ug of recombinant plasmid DNA ($OD_{260/280}$ = 1.86). This recombinant plasmid was labeled pGEM31.

Complete cDNA Sequence for Clone 3

Determination of cDNA sequence of this newly cloned fragment was done using Sequenase v2.0. Approximately 388 bp of sequence on the 5' end of this fragment overlaps known sequence from the original clone 3. The two sequences in the region of this overlap are identical and confirm the identity of the pGEM31 insert as that representing the 3' end of the original clone 3 (Figure 9). The sequence from this insert, in combination with that already known for the original clone 3, provides a cDNA sequence representative of a complete ORF. The complete clone 3 sequence is reported in Figure 10. This cDNA is 1220 bp and includes sequence representing the initiating Met and signal peptide through the 3' non-translated region, polyadenylation consensus sequence and the 20 residue polyA tail (Figure 10). The ORF is 972 bp in length and codes for 324 amino acids including the signal peptide.

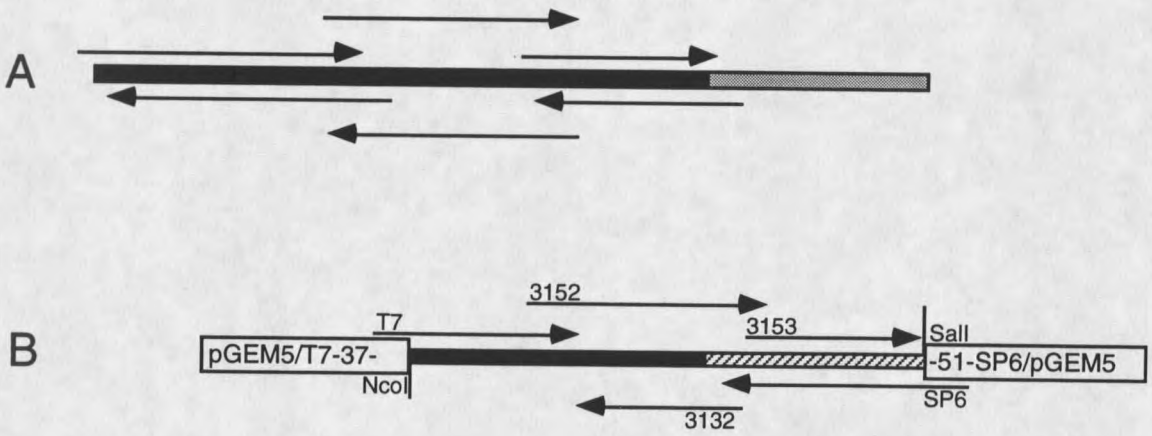


Figure 9. The combination of sequence known from the original 879 bp clone 3 (A) with the complete 3' end, pGEM31 (B) provide a complete cDNA sequence representative of a full-length ORF for a lectin from grasshopper. Sequencing primers and location with respect to the cDNA insert representing the 3' end of clone 3 are shown. The black portion of the pGEM31 insert represents the 388 bp (491-879) of sequence that overlaps with that known for the original clone 3. The cross-hatched portion of the insert is new sequence data representing the previously missing 3' end of clone 3.

```
GGGAGAGCAC GGCTGCAGCT CGCGCCAGGA GGAGGAGTTG ATAAGCAGGA TGCAGCTGGT GACGGTGTGC    70
                                                          (M)
GCGGCGCTGG TGGCAGCGAC AGTACCCTGC ACCCTGGCCG CCGTGGACCT GTTCTGCAGC TGCCAGGTGC   140

GCCACCACAG GGACTCGACG ACGGCCGTGC ACTGCTCAGG GGAACAGAGT GGGAACAAAA CGATTTCTTG   210

CCAAAAAGCT CAAGTGCCGG ACATTCCACG TGACTACCAC TACGTGCCAG GCTACGCCCT CGTCAAGCTG   280

TACCGCATAA TGATGACATG GGAGGAAGCC AAAAAGGCCT GCGAAGCCGA GGGAGCAAAA TTAGCAGTCC   350

CAAGAGACAA CCACGCCTAC GATGGCCTGA AGCAGATCTT CAAGTTAGGG TTTGGGGTGT ACTGGGCCAA   420
                                                879PCR -->
CATCGGAATC ACAGATCATG AGAGCGAGGG AATATTCAGC GGAGTGGATG GTCATCCAGT GTCGTTCCTG   490

CCATGGAATC CTAATGAACC CAACAACGCC GGAGGCAACG AGAACTGTGT TAACGTCAAC GACAAAGGAC   560
 NcoI
AGCTGAACGA CTGGCATTGC GGGAATACAG CGCCATTCTT CTGCGAGCGC CGGCCCTCGG TGGGCATACC   630

ACCCTCCTAT GTGTGGCTGA AGGACGCGAG CCGCTTCTAC AAGGTGCACG CCGAGAAGCA CGTGTACGCG   700
                                        3152 -->
GAGGCAGCCA GGGTGTGCCG ATCCGAGAAC GCGACGCTCG CTGTGCCCGA CACCTGGGAC CGTGTCGAGA   770

CCCTGCTGCG ACTCCTCGAG CCGAAAGAAG AGTTCTACCT GACAGGATTC ACAGATGAGG CTGTGGAAGG   840
               XhoI
TGACTTCGTT ACCGAAACAG GAAGACACCT AAAAGGCATG GAATTCCAGG TGTGGAGCCC TGGTGAGCCA   910
                                           EcoRI
AATAACGACG TCGATGGGAA GCCCGAGAAT TGCCTAGCCT TTTCGGGCCG TGGCTACTAC GGCGACAGGA   980

GCTGCGAAGT GGAGCTGCCC TTCATCTGCG AGATAGCGCC CTGACGTGGC GCTCTCGGGA CACAATTCTG  1050
          <-- 3132     3153 -->                *
AGGACGCAAC AACTGCGAAT GGAAACGACG TACGCGGAGA GCATGGACTG GTGCGAAGCC GAGTGACATT  1120

CAGAAGACAT TGTATAATTT ATATGTGAAT AAATATTCGT TAGCAACCCC TAAAAAAAAA AAAAAAAAAA  1190

AGTACTAGTC GACCATATGG GAGAGCTCCC                                             1220
```

Figure 10. Nucleotide sequence for the complete 1220 bp clone 3. The location of 3152, 3153 and 3132 sequencing primers are underlined. The clone 3-specific primer (879PCR) used in the 3' RACE procedure is also underlined. The NcoI, XhoI and EcoRI restriction sites used to verify the clone 3 authenticity of the original RACE product are underlined as well as the start codon (ATG), stop codon (TGA), consensus polyadenylation signal (AATAAA) and polyA tail. The complete ORF is 972 bp and codes for a polypeptide of 324 amino acids.

cDNA Sequence for Clone 4

Partial sequence from the original clone 4 cDNA was first used in combination with partial sequence from clone 3 to form a hybrid cDNA suggestive of a full-length ORF for a lectin from grasshopper (Figure 3 p 16; Rognlie, 1991). This partial sequence was homologous with that of clone 3 in the region of overlap and included 3' sequence coding for a stop codon and polyA tail. Sequence analysis of the entire clone 4 cDNA determined the insert to be 1079 bp in length. This cDNA contained a single ORF encoding a stop codon and polyA tail at the 3' end. The 5' end was lacked sequence coding for several N-terminal residues, the signal peptide and initiating Met. Alignment of this partial clone 4 sequence with the complete clone 3 sequence suggested that at least 69 bp were missing from the 5' end of this sequence. It became evident later that the truncated versions of both clone 3 and clone 4 were the result of unintended cleavage at the internal EcoRI sites during library construction.

Search for the 5' end of Clone 4

Seventeen positive plaques were originally identified using the ^{32}P-labeled 300 bp cDNA as discussed previously (Rognlie, 1991). Three were subcloned and sequenced (clones 2, 3 & 4) and 12 of the remaining plaques were determined to contain clone 3-specific inserts (Figure 6 p 40). Two plaques that did not amplify with the clone 3-specific primer, 57c and 13, were screened using PCR and a clone 4-specific primer to amplify an insert that would contain the $\geq$ 69 bp of sequence representing the 5' end of clone 4. The size of an insert containing this sequence would be $\geq$ 300 bp, based on sequence alignment with clone 3. Figure 11 illustrates the PCR products obtained for clones 57c, 13 and 43a. Plaque 43a was from a library of cDNA fragments known to be $\leq$ 600 bp (Hapner, unpublished). These products are $\approx$ 300 bp and not of sufficient length to contain the

additional sequence representing the missing 5' end of clone 4. It was concluded the complete 5' end of clone 4 is not accessible in the current cDNA library using this method. The alternative 5' RACE strategy was employed to selectively isolate and clone this sequence from polyA$^+$ mRNA. This work is discussed below.

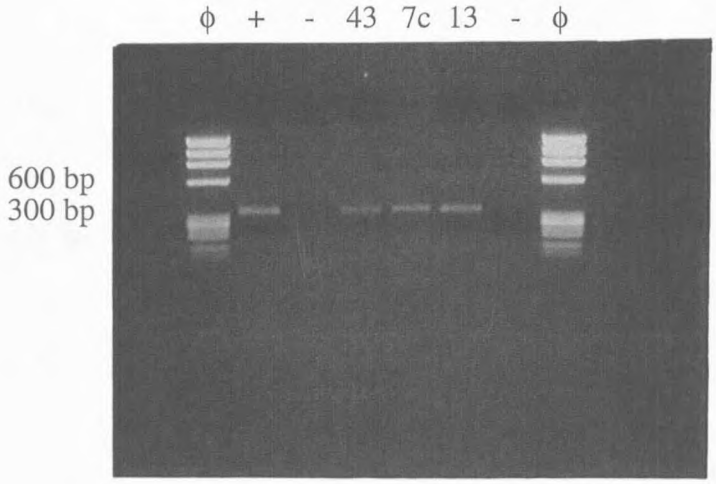


Figure 11. Screening clones 43, 7c and 13 using PCR with the clone 4-specific primer 1150PROB. ϕ = ϕx174/Hae III DNA ladder. "-" = negative control using 1150PROB with a clone 3-specific template and serves to illustrate the gene-selectivity of the primer 1150PROB. "+" = positive control using 1150PROB with a clone 4-specific template. The ≈ 300 bp product in this positive control illustrates the expected size ORF a clone 4-specific PCR product without the 5' end. Plaques 43a, 57c and 13 are positive clones isolated using the original 300 bp cDNA probe. The similar sized PCR products indicate that clones 43, 7c and 13 do not contain sequence representing the missing 5' end of the original 1079 bp clone 4.

Rapid Amplification of cDNA Ends: 5' RACE

mRNA Isolation. mRNA was isolated using Tri-Reagent™. Roughly 350 μg of total RNA and 12 μg of polyA$^+$ mRNA were obtained from 100 mg of fat body tissue ($OD_{260/280}$ = 1.6). One (1) μg of polyA$^+$ mRNA was used for first strand cDNA synthesis.

Anchor Ligation and PCR Using a Gene-Specific Primer. Following synthesis of the second-strand cDNA, the double-stranded DNA was blunt-ended using T4 DNA polymerase and purified with Prep-a-gene™. Double-stranded anchor sequences were ligated to each end of the cDNA and an aliquot of the completed reaction diluted 1:25 in water. Two (2) µl of this solution was used with a clone 4-specific primer '4036' to PCR amplify a clone 4 product containing the complete 5' sequence. The size of this RACE product is ≈ 1100 bp. This product is pictured in Lanes 1 and 2 of Figure 12. The 1100 bp product is sufficiently long to contain the complete 5' sequence for the original clone 4. The '4036' primer anneals at nucleotide 78 of the original clone 4 cDNA thus the 1100 bp fragment would represent 78 bp of the original clone 4 sequence and ≈ 1022 bp of new 5' sequence including that encoding the N-terminal amino acids, the signal peptide and initiating Met. Restriction analysis of this fragment to verify its clone 4-authenticity before cloning and sequence analysis was not possible. Only 78 bp of the fragment represent known sequence. This limits the region of known sequence in which to find a clone 4-specific restriction enzyme. Further, the fragment cleaved would not substantially change the size of the product thus making it difficult to determine if restriction took place. Therefore, this fragment was cloned directly into pGEM-T plasmid vector for subsequent sequence analysis.

Ligation into pGEM-T and Transformation of Competent Cells. The ≈ 1100 bp 5' RACE product was isolated from the agarose gel and ligated directly into a pGEM-T plasmid vector. Following transformation of competent cells (JM109) and overnight growth, 2 white colonies and 76 blue colonies were observed. Transformation efficiency was 2.3×10^4 colonies/µg DNA. Each white colony was selected, grown overnight in LB and a mini-prep completed on 3 ml of each culture. This resulted in ≈ 30 µg of recombinant plasmid for each colony. This plasmid was labeled pGEM41.

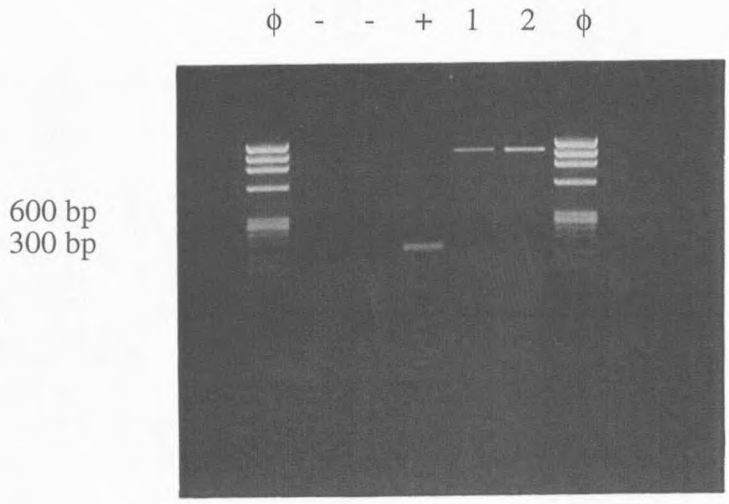


Figure 12. 5' RACE product obtained using 4036 and AP primers with ds-strand cDNA made directly from fat body mRNA. ϕ = ϕx174/Hae III DNA ladder. "-" = negative control with 4036 and AP primers, but no DNA template. "-" = negative control with 4036 and AP primers using clone 3 plasmid DNA template. "+" = positive control using 4036 and λR with clone 4 plasmid DNA template. Lanes 1 and 2 are 5' RACE products. These products appear ≈ 1100 bp in size, large enough to contain the complete 5' end of the original clone 4 sequence.

Complete Clone 4 cDNA Sequence

Sequence analysis of the pGEM41 insert was completed using Sequenase v2.0. Seventy-eight (78) bp of sequence at the 5' end of the original clone 4 overlaps with new sequence in pGEM41. Overlapping sequences are identical and confirms the clone 4 authenticity of the pGEM41 insert. The combination of sequence from the original 1079 bp clone 4 with that from pGEM41 provides a complete clone 4-specific cDNA representing a full-length ORF for a second lectin from this grasshopper (Figure 13). The complete

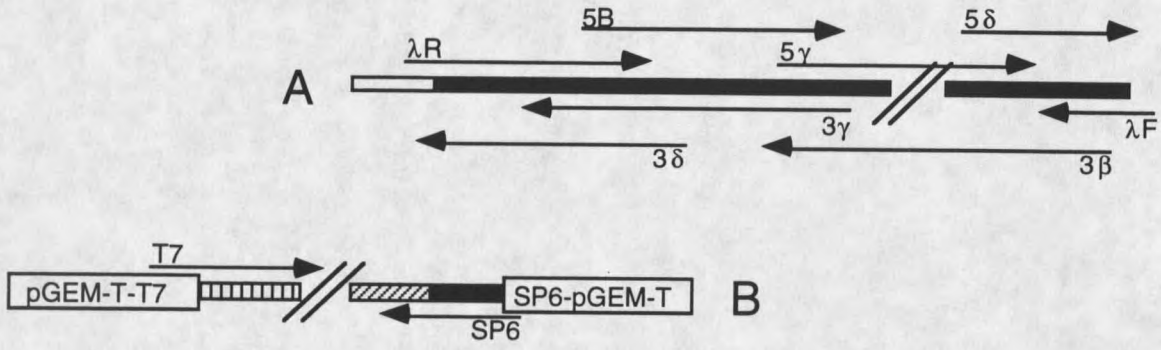


Figure 13. The combination of sequence known from the original 1079 bp clone 4 (A) with the complete 5' end in pGEM41 (B) provide a complete cDNA sequence for a full-length ORF representative of a second lectin from this grasshopper. cDNA sequences are represented by horizontal bars. Primers 5B, 5γ, 5δ, 3β, 3γ and 3δ were used to sequence the original 1079 bp clone 4 cDNA. SP6 and T7 primers were used to sequence the two end regions. Primers are indicated with arrows. The black portion of the pGEM41 insert represents the 78 bp of sequence that overlaps with that known for the original clone 4. The cross-hatched portion of the insert is new sequence data representing the previously missing 5' end of clone 4. Vertical bars represent the unsequenced portion of the pGEM41 insert.

```
GCCACAGCG GTGGTTCCGG CAGTACAAGG ATTTCCTTAC GCGAAATACG GGCAGACATG     59
                                                                *
GCCTGCCCCC TTATTATTAT TTTTGAGACC AGACCAACTG TGTTAATGGT AGGGACCGGG    119

GCTCAGCAGA ATTCCGGCGG CTGCAAGCTG CGCCACCACA AGGAATCGAC GATGGCCCTG    179
        EcoRI
CACTGCTCAA GGGATCAGAG TGGGAACAAA ACGTGTCAAA AAGCCCAAGT GCCGGACATT    239
                                                  <-- 4036
CCACGTGGCT ACCACTACGT GCCAGGCTAC GCTCTCGTCA AGATGTACCG CATAATGATG    299
                                          5B -->
ACATGGGAGG AAGCCAGGAA GGCCTGCGAA GCCGAGGGAG CCATACTGGC ACTCCCAAAG    359

GACAGCCACG CCTATGATGG ACTGAAGCAG GTCATCATAG CAGAGCATAA AGAGGAAGGG    419

GTTTATTGGG CCAACATCGG AATCACAGAT CAGTACAGCG AGGGAATATT CGTGGGAGTG    479
                                  <-- 3δ
GATGGTCTTC CAGTGTCGTA CCTGCCATGG AGACCTAATG AGCCCAACAA CTTCGGAGGC    539

AACGAGAACT GTGTTTACGT CGACGACAAA GGACAGCTGA ACGACTGGGG GTGCGCGAAT    599
        5γ -->
GCAGAGCCAT TCTTCTGCGA GCGCCGGCTC TCGGTGGGCA TACCACCCTC CTATGTGTGG    659

CTGAAGGACG CGAGCCGCTT CTACAAGGTG CACGCCGAGA AGCACGTGCA CGCGGAGGCA    719
                                      <-- 3γ
GCCAGGGTGT GCCGATCCGA GAACGCGACG CTCGCTGTGC CCGACACCTG GGACCGTGTC    779

GAGGCCCTGC TGCGACTCCT CGAGCCGAAA GAAGAGTTCT ACCTGACAGG ATTCACAGAT    839
                                          5δ -->
GAGGCTGCGG AAGGTGACTT CGTTGCCGAA ACAGGAAGAC ACCTAAGAGA CATGGAATTC    899
                                                              EcoRI
CAAGTGTGGA AACCTGGTGA GCCAAATAAC AACTTTCTTG GGAAGCCCGA GAACTGCTTA    959
                                  <-- 3β
GGCTTTGGGG GCAAAGGCTA CTACGACGAC AAGAGCTGCG ATTTGGAGCT GCCCTTCATC   1019

TGCGAGATAG CGCCCTGACC TGGCGCTCTC GTGACACGTC TGAGGACGCA ACAACTGCGA   1079
                  *
ATGGAAACGA CGTACTCGGA AAGTATGGAC TGTTGCGAGG CAGAGTTACG TTCAGAAGAC   1139

GTTGTATAAT TGATATGCGA ATAAATATTC GTTAGTAACC TAAAAAAAAA AAAAAAAAAA   1199

AACGGAATCC GCGGAATTCC                                               1213
```

Figure 14. Nucleotide sequence for the complete 1213 bp cDNA of clone 4. 5B, 5γ, 5δ, 3β, 3γ and 3δ sequencing primers are underlined. EcoRI sites at #128 and #894, a start codon (ATG) at nucleotide #57, a stop codon (TGA) at nucleotide #1035 and the polyadenylation consensus sequence (AATAAA) and polyA tail are underlined. The complete 975 bp ORF encodes a polypeptide of 326 amino acid residues.

sequence is 1213 bp long and includes sequence encoding a start codon, signal peptide and stop codon through the polyA tail. This complete clone 4 sequence is reported in Figure 14. The ORF is 975 bp in length and codes of a 326 amino acid polypeptide.

Comparison of cDNA Sequences for Clones 3 and 4

The complete cDNA sequences for clone 3 and clone 4 show 81% identity including the 3' and 5' non translated region (Figure 15). Identity within the ORFs is 89%. These sequences represent distinctly different, but homologous genes. Both clone 3 and clone 4 are complete as defined by sequence representing an initiating Met, signal peptide, stop codon and polyA tail. In the Figure 15 alignment, the location of the start codons for clones 3 and 4 differ by one codon. This difference can be attributed to the gaps inserted in each sequence to optimize maximum identity in the alignment. Sequences encoding the signal peptide in each clone differ by 3 nucleotides making the signal in clone 3 one amino acid residue shorter (20) when compared to that of clone 4 (21). Stop codons in each of clones 3 and 4 are identical to one another and their alignment also evident in Figure 15. The size of the ORF for clone 3 is 972 bp, three nucleotides shorter than the 975 bp ORF reported for clone 4. The sequences are very similar at the 3' ends including the ≈ 200 bp of non-translated sequence following the stop codon, but distinctly different in the 5' non-translated region. The major differences are at the 5' end where clone 3 has a 6 nucleotide insert at position #203, and clone 4 has a 9 nucleotide insert at position #397. The net difference between the two sequences is one amino acid in the mature protein. Each of these regions is underlined in Figure 15. These differences provided the basis for the design of gene-specific primers used in the 3' and 5' RACE strategies. The clone 3-specific primer (879PCR) was used to amplify and clone the complete 3' end of clone 3 and was designed around the 9 nucleotide gap at position #394.

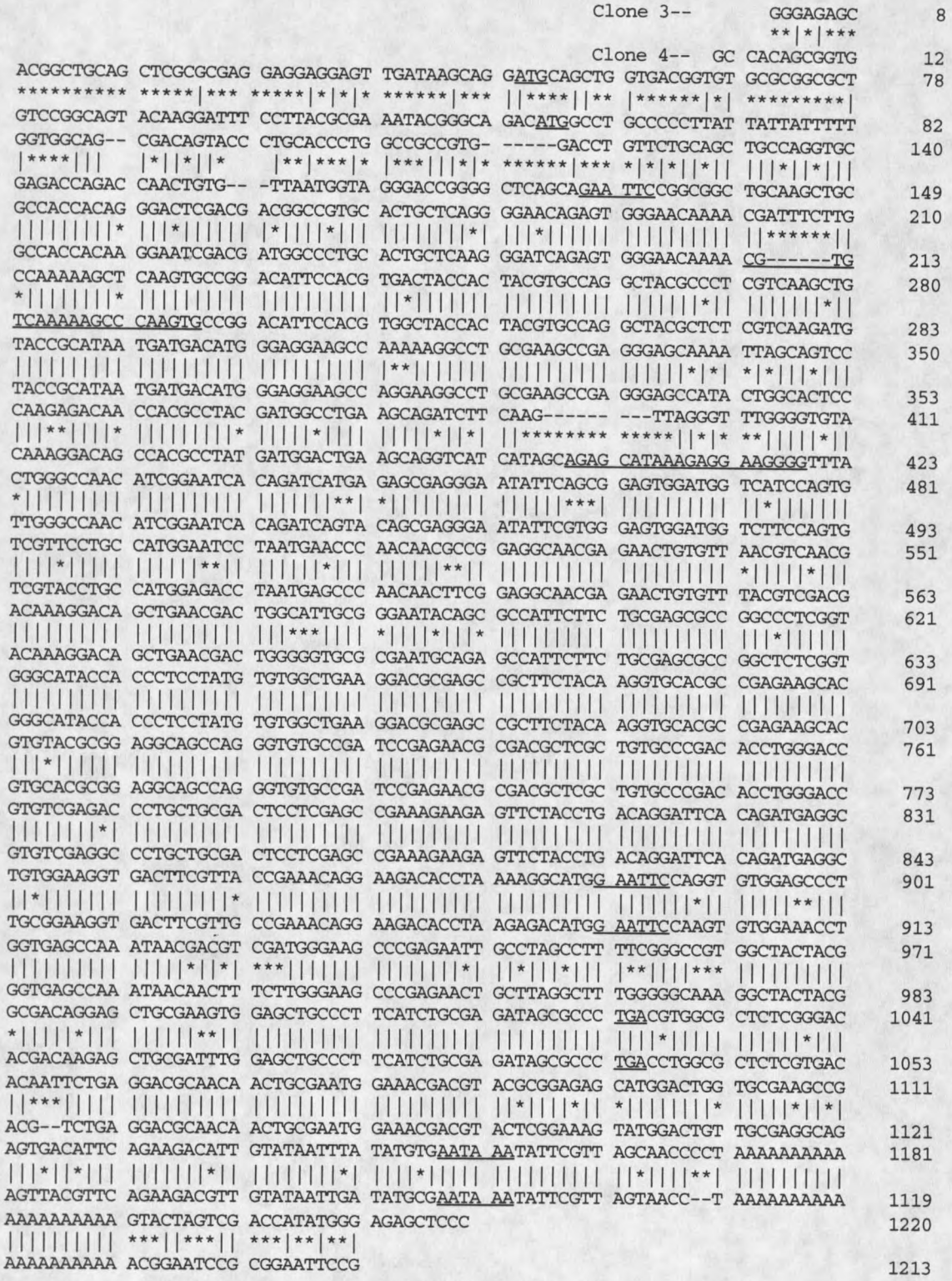

Clone 3-- GGGAGAGC 8
Clone 4-- GC CACAGCGGTG 12
ACGGCTGCAG CTCGCGCGAG GAGGAGGAGT TGATAAGCAG GATGCAGCTG GTGACGGTGT GCGCGGCGCT 78
GTCCGGCAGT ACAAGGATTT CCTTACGCGA AATACGGGCA GACATGGCCT GCCCCCTTAT TATTATTTTT 82
GGTGGCAG-- CGACAGTACC CTGCACCCTG GCCGCCGTG- -----GACCT GTTCTGCAGC TGCCAGGTGC 140
GAGACCAGAC CAACTGTG-- -TTAATGGTA GGGACCGGGG CTCAGCAGAA TTCCGGCGGC TGCAAGCTGC 149
GCCACCACAG GGACTCGACG ACGGCCGTGC ACTGCTCAGG GGAACAGAGT GGGAACAAAA CGATTTCTTG 210
GCCACCACAA GGAATCGACG ATGGCCCTGC ACTGCTCAAG GGATCAGAGT GGGAACAAAA CG------TG 213
CCAAAAAGCT CAAGTGCCGG ACATTCCACG TGACTACCAC TACGTGCCAG GCTACGCCCT CGTCAAGCTG 280
TCAAAAAGCC CAAGTGCCGG ACATTCCACG TGGCTACCAC TACGTGCCAG GCTACGCTCT CGTCAAGATG 283
TACCGCATAA TGATGACATG GGAGGAAGCC AAAAAGGCCT GCGAAGCCGA GGGAGCAAAA TTAGCAGTCC 350
TACCGCATAA TGATGACATG GGAGGAAGCC AGGAAGGCCT GCGAAGCCGA GGGAGCCATA CTGGCACTCC 353
CAAGAGACAA CCACGCCTAC GATGGCCTGA AGCAGATCTT CAAG------ ---TTAGGGT TTGGGGTGTA 411
CAAAGGACAG CCACGCCTAT GATGGACTGA AGCAGGTCAT CATAGCAGAG CATAAAGAGG AAGGGGTTTA 423
CTGGGCCAAC ATCGGAATCA CAGATCATGA GAGCGAGGGA ATATTCAGCG GAGTGGATGG TCATCCAGTG 481
TTGGGCCAAC ATCGGAATCA CAGATCAGTA CAGCGAGGGA ATATTCGTGG GAGTGGATGG TCTTCCAGTG 493
TCGTTCCTGC CATGGAATCC TAATGAACCC AACAACGCCG GAGGCAACGA GAACTGTGTT AACGTCAACG 551
TCGTACCTGC CATGGAGACC TAATGAGCCC AACAACTTCG GAGGCAACGA GAACTGTGTT TACGTCGACG 563
ACAAAGGACA GCTGAACGAC TGGCATTGCG GGAATACAGC GCCATTCTTC TGCGAGCGCC GGCCCTCGGT 621
ACAAAGGACA GCTGAACGAC TGGGGGTGCG CGAATGCAGA GCCATTCTTC TGCGAGCGCC GGCTCTCGGT 633
GGGCATACCA CCCTCCTATG TGTGGCTGAA GGACGCGAGC CGCTTCTACA AGGTGCACGC CGAGAAGCAC 691
GGGCATACCA CCCTCCTATG TGTGGCTGAA GGACGCGAGC CGCTTCTACA AGGTGCACGC CGAGAAGCAC 703
GTGTACGCGG AGGCAGCCAG GGTGTGCCGA TCCGAGAACG CGACGCTCGC TGTGCCCGAC ACCTGGGACC 761
GTGCACGCGG AGGCAGCCAG GGTGTGCCGA TCCGAGAACG CGACGCTCGC TGTGCCCGAC ACCTGGGACC 773
GTGTCGAGAC CCTGCTGCGA CTCCTCGAGC CGAAAGAAGA GTTCTACCTG ACAGGATTCA CAGATGAGGC 831
GTGTCGAGGC CCTGCTGCGA CTCCTCGAGC CGAAAGAAGA GTTCTACCTG ACAGGATTCA CAGATGAGGC 843
TGTGGAAGGT GACTTCGTTA CCGAAACAGG AAGACACCTA AAAGGCATGG AATTCCAGGT GTGGAGCCCT 901
TGCGGAAGGT GACTTCGTTG CCGAAACAGG AAGACACCTA AGAGACATGG AATTCCAAGT GTGGAAACCT 913
GGTGAGCCAA ATAACGACGT CGATGGGAAG CCCGAGAATT GCCTAGCCTT TTCGGGCCGT GGCTACTACG 971
GGTGAGCCAA ATAACAACTT TCTTGGGAAG CCCGAGAACT GCTTAGGCTT TGGGGGCAAA GGCTACTACG 983
GCGACAGGAG CTGCGAAGTG GAGCTGCCCT TCATCTGCGA GATAGCGCCC TGACGTGGCG CTCTCGGGAC 1041
ACGACAAGAG CTGCGATTTG GAGCTGCCCT TCATCTGCGA GATAGCGCCC TGACCTGGCG CTCTCGTGAC 1053
ACAATTCTGA GGACGCAACA ACTGCGAATG GAAACGACGT ACGCGGAGAG CATGGACTGG TGCGAAGCCG 1111
ACG--TCTGA GGACGCAACA ACTGCGAATG GAAACGACGT ACTCGGAAAG TATGGACTGT TGCGAGGCAG 1121
AGTGACATTC AGAAGACATT GTATAATTTA TATGTGAATA AATATTCGTT AGCAACCCCT AAAAAAAAAA 1181
AGTTACGTTC AGAAGACGTT GTATAATTGA TATGCGAATA AATATTCGTT AGTAACC--T AAAAAAAAAA 1119
AAAAAAAAAA GTACTAGTCG ACCATATGGG AGAGCTCCC 1220
AAAAAAAAAA ACGGAATCCG CGGAATTCCG 1213

Figure 15. (previous page) Alignment of the cDNA sequences for clones 3 and 4. These sequences are 81% identical when including the 3' and 5' non-translated regions, and 89% identical within sequence defining the ORFs. Two regions of significant difference occur at nucleotide position #203 in clone 3 and #397 in clone 4 where 6 and 9 nucleotide insertions are present. These sequence differences are underlined. Each was the basis for the design of gene-specific primers used in the 3' and 5' RACE procedures. EcoRI sites at #127 and #893 in clone 4 and #881 in clone 3 are underlined as well as start codons (ATG), stop codons (TGA) and the polyadenylation signal sequences in each clone. Sequences were aligned using GenePro™. Identities are represented by vertical bars, non-identities by asterisks.

The clone 4-specific primer (4036) used to amplify and clone the complete 5' end of clone 4 and was based on the 6 nucleotide insertion at position #203. No other regions between the two sequences are sufficiently different to allow for the design of primers that would effectively distinguish clone 3 from clone 4 sequence. Clone 4 has an EcoRI site at position #126 that is not present in the clone 3 sequence.

Amino Acid Sequence for Clone 3

The deduced amino acid sequence for clone 3 is 324 residues including a 20 residue signal peptide. This sequence is shown in Figure 16. The initiating Met and 20 residue signal peptide is underlined in this sequence. The N-terminal Alanine residue was established by Edman sequence from a grasshopper lectin isolated in this laboratory. The calculated weight of the mature 304 residue protein is 34056 Da. This value is in agreement with published data for a grasshopper lectin (Stebbins, 1985) and further is consistent with the mass of determined via Matrix Assisted Laser Desorption Ionization/Time of Flight (MALDI/TOF) mass spectrometry (Wenzlick, 1996). Two potential N-glycosylation sites (NXT, S or C) are present at amino acid positions #29 and #207. Cys residues at positions #71, #143, #157 and #165 are conserved within this protein family and are typical for C-type lectin CRDs. The motif defined by these residues

is repeated at positions #203, #278, #292 and #300 such that the single polypeptide contains two separate C-type CRDs. This duplication of CRDs is novel among invertebrate C-type lectins.

Amino Acid Sequence for Clone 4

The deduced amino acid sequence for clone 4 is 326 residues including a 21 residue signal peptide. The complete sequence is shown in Figure 17. The initiating Met and 21 residue signal peptide is underlined, and as discussed for the clone 3 sequence, the N-terminal Alanine established by N-terminal Edman sequence data from a lectin protein isolated in this laboratory. The calculated molecular weight of the mature 305 residue protein is 34401 Da. This weight is also consistent with that observed via SDS PAGE (Stebbins, 1985) and MALDI/TOF mass spectrometry (Wenzlick, 1996) for a previously isolated grasshopper lectin. Two potential N-glycosylation sites are also present in this sequence at positions #29 (NKT) and #208 (NAT). The presence of conserved Cys residues at positions #69, #144, #158 and #166 define the type C-type lectin motif for members of this protein family. This motif is repeated at positions #204, #279, #293 and #301 are typical for other lectins with C-type carbohydrate recognition domains (Drickamer, 1994) and suggests this protein sequence also represents an authentic C-type lectin. As discussed above, the duplication of CRDs is novel among invertebrate lectins.

Comparison of Amino Acid Sequence for Clones 3 and 4

Alignment of the deduced amino acid sequences for clones 3 and 4 is shown in Figure 18. These sequences are 80% identical and 85% homologous when conservative substitutions are considered. The clone 4 amino acid sequence is 326 amino acids, one

```
  G GGA GAG CAC GGC TGC AGC TCG CGC CAG GAG GAG GAG TTG ATA AGC AGG ATG CAG     55
    •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   M   Q
CTG GTG ACG GTG TGC GCG GCG CTG GTG GCA GCG ACA GTA CCC TGC ACC CTG GCC GCC   112
L   V   T   V   C   A   A   L   V   A   A   T   V   P   C   T   L   A   A     1
GTG GAC CTG TTC TGC AGC TGC CAG GTG CGC CAC CAC AGG GAC TCG ACG ACG GCC GTG   169
V   D   L   F   C   S   C   Q   V   R   H   H   R   D   S   T   T   A   V    20
CAC TGC TCA GGG GAA CAG AGT GGG AAC AAA ACG ATT TCT TGC CAA AAA GCT CAA GTG   226
H   C   S   G   E   Q   S   G   N   K   T   I   S   C   Q   K   A   Q   V    39
CCG GAC ATT CCA CGT GAC TAC CAC TAC GTG CCA GGC TAC GCC CTC GTC AAG CTG TAC   283
P   D   I   P   R   D   Y   H   Y   V   P   G   Y   A   L   V   K   L   Y    58
CGC ATA ATG ATG ACA TGG GAG GAA GCC AAA AAG GCC TGC GAA GCC GAG GGA GCA AAA   340
R   I   M   M   T   W   E   E   A   K   K   A   C   E   A   E   G   A   K    77
TTA GCA GTC CCA AGA GAC AAC CAC GCC TAC GAT GGC CTG AAG CAG ATC TTC AAG TTA   397
L   A   V   P   R   D   N   H   A   Y   D   G   L   K   Q   I   F   K   L    96
GGG TTT GGG GTG TAC TGG GCC AAC ATC GGA ATC ACA GAT CAT GAG AGC GAG GGA ATA   454
G   F   G   V   Y   W   A   N   I   G   I   T   D   H   E   S   E   G   I   115
TTC AGC GGA GTG GAT GGT CAT CCA GTG TCG TTC CTG CCA TGG AAT CCT AAT GAA CCC   511
F   S   G   V   D   G   H   P   V   S   F   L   P   W   N   P   N   E   P   134
AAC AAC GCC GGA GGC AAC GAG AAC TGT GTT AAC GTC AAC GAC AAA GGA CAG CTG AAC   568
N   N   N   G   G   N   E   N   C   V   N   V   N   D   K   G   Q   L   N   153
GAC TGG CAT TGC GGG AAT ACA GCG CCA TTC TTC TGC GAG CGC CGG CCC TCG GTG GGC   625
D   W   H   C   G   N   T   A   P   F   F   C   E   R   R   P   S   V   G   172
ATA CCA CCC TCC TAT GTG TGG CTG AAG GAC GCG AGC CGC TTC TAC AAG GTG CAC GCC   682
I   P   P   S   Y   V   W   L   K   D   A   S   R   F   Y   K   V   H   A   191
GAG AAG CAC GTG TAC GCG GAG GCA GCC AGG GTG TGC CGA TCC GAG AAC GCG ACG CTC   739
E   K   H   V   Y   A   E   A   A   R   V   C   R   S   E   N   A   T   L   210
GCT GTG CCC GAC ACC TGG GAC CGT GTC GAG ACC CTG CTG CGA CTC CTC GAG CCG AAA   796
A   V   P   D   T   W   D   R   V   E   T   L   L   R   L   L   E   P   K   229
GAA GAG TTC TAC CTG ACA GGA TTC ACA GAT GAG GCT GTG GAA GGT GAC TTC GTT ACC   853
E   E   F   Y   L   T   G   F   T   D   E   A   V   E   G   D   F   V   T   248
GAA ACA GGA AGA CAC CTA AAA GGC ATG GAA TTC CAG GTG TGG AGC CCT GGT GAG CCA   910
E   T   G   R   H   L   K   G   M   E   F   Q   V   W   S   P   G   E   P   267
AAT AAC GAC GTC GAT GGG AAG CCC GAG AAT TGC CTA GCC TTT TCG GGC CGT GGC TAC   967
N   N   D   V   D   G   K   P   E   N   C   L   A   F   S   G   R   G   Y   286
TAC GGC GAC AGG AGC TGC GAA GTG GAG CTG CCC TTC ATC TGC GAG ATA GCG CCC TGA  1024
Y   G   D   R   S   C   E   V   E   L   P   F   I   C   E   I   A   P   *   304
CGT GGC GCT CTC GGG ACA CAA TTC TGA GGA CGC AAC AAC TGC GAA TGG AAA CGA CGT  1081
•   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •     *
ACG CGG AGA GCA TGG ACT GGT GCG AAG CCG AGT GAC ATT CAG AAG ACA TTG TAT AAT  1138
•   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •     *
TTA TAT GTG AAT AAA TAT TCG TTA GCA ACC CCT AAA AAA AAA AAA AAA AAA AAG TAC  1195
•   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •   •     *
TAG TCG ACC ATA TGG GAG AGC TCC C                                           1220
•   •   •   •   •   •   •   •
```

Figure 16. cDNA and coded amino acid sequence for the 1220 bp clone 3. The 972 bp ORF encodes a 324 residue polypeptide including the signal sequence. The 304 amino acid protein has a calculated molecular weight of 34056 Da. Potential N-glycosylation sites at residues #29 and #207 are boxed. The start codon (ATG), initiating Met and N-terminal Ala are underlined. The stop codon at nucleotide 1022 is marked with an asterisk (*). 5' and 3' nontranslated sequences are indicated with an "•" upstream of the initiating Met and downstream of the stop codon. The signal peptide is underlined.

```
                                                                GCC ACA GCG GTG GTT CCG GCA      21
                                                                 .   .   .   .   .   .   .       .
GTA CAA GGA TTT CCT TAA CGC GAA ATA CGG GCA GAC ATG GCC TGC CCC CTT ATT      75
 .   .   .   .   .   .   .   .   .   .   .   .   M   A   C   P   L   I       .
ATT ATT TTT GAG ACC AGA CCA ACT GTG TTA ATG GTA GGG ACC GGG GCT CAG CAG     129
 I   I   F   E   T   R   P   T   V   L   M   V   G   T   G   A   Q   Q       3
AAT TCC GGC GGC TGC AAG CTG CGC CAC CAC AAG GAA TCG ACG ATG GCC CTG CAC     183
 N   S   G   G   C   K   L   R   H   H   K   E   S   T   M   A   L   H      21
TGC TCA AGG GAT CAG AGT GGG AAC AAA ACG TGT CAA AAA GCC CAA GTG CCG GAC     237
 C   S   R   D   Q   S   G   N   K   T   C   Q   K   A   Q   V   P   D      39
ATT CCA CGT GGC TAC CAC TAC GTG CCA GGC TAC GCT CTC GTC AAG ATG TAC CGC     291
 I   P   R   G   Y   H   Y   V   P   G   Y   A   L   V   K   M   Y   R      57
ATA ATG ATG ACA TGG GAG GAA GCC AGG AAG GCC TGC GAA GCC GAG GGA GCC ATA     345
 I   M   M   T   W   E   E   A   R   K   A   C   E   A   E   G   A   I      75
CTG GCA CTC CCA AAG GAC AGC CAC GCC TAT GAT GGA CTG AAG CAG GTC ATC ATA     399
 L   A   L   P   K   D   S   H   A   Y   D   G   L   K   Q   V   I   I      93
GCA GAG CAT AAA GAG GAA GGG GTT TAT TGG GCC AAC ATC GGA ATC ACA GAT CAG     453
 A   E   H   K   E   E   G   V   Y   W   A   N   I   G   I   T   D   Q     111
TAC AGC GAG GGA ATA TTC GTG GGA GTG GAT GGT CTT CCA GTG TCG TAC CTG CCA     507
 Y   S   E   G   I   F   V   G   V   D   G   L   P   V   S   Y   L   P     129
TGG AGA CCT AAT GAG CCC AAC AAC TTC GGA GGC AAC GAG AAC TGT GTT TAC GTC     561
 W   R   P   N   E   P   N   N   F   G   G   N   E   N   C   V   Y   V     147
GAC GAC AAA GGA CAG CTG AAC GAC TGG GGG TGC GCG AAT GCA GAG CCA TTC TTC     615
 D   D   K   G   Q   L   N   D   W   G   C   A   N   A   E   P   F   F     165
TGC GAG CGC CGG CTC TCG GTG GGC ATA CCA CCC TCC TAT GTG TGG CTG AAG GAC     669
 C   E   R   R   L   S   V   G   I   P   P   S   Y   V   W   L   K   D     183
GCG AGC CGC TTC TAC AAG GTG CAC GCC GAG AAG CAC GTG CAC GCG GAG GCA GCC     723
 A   S   R   F   Y   K   V   H   A   E   K   H   V   H   A   E   A   A     201
AGG GTG TGC CGA TCC GAG AAC GCG ACG CTC GCT GTG CCC GAC ACC TGG GAC CGT     777
 R   V   C   R   S   E   N   A   T   L   A   V   P   D   T   W   D   R     219
GTC GAG GCC CTG CTG CGA CTC CTC GAG CCG AAA GAA GAG TTC TAC CTG ACA GGA     831
 V   E   A   L   L   R   L   L   E   P   K   E   E   F   Y   L   T   G     237
TTC ACA GAT GAG GCT GCG GAA GGT GAC TTC GTT GCC GAA ACA GGA AGA CAC CTA     885
 F   T   D   E   A   A   E   G   D   F   V   A   E   T   G   R   H   L     255
AGA GAC ATG GAA TTC CAA GTG TGG AAA CCT GGT GAG CCA AAT AAC AAC TTT CTT     939
 R   D   M   E   F   Q   V   W   K   P   G   E   P   N   N   N   F   L     273
GGG AAG CCC GAG AAC TGC TTA GGC TTT GGG GGC AAA GGC TAC TAC GAC GAC AAG     993
 G   K   P   E   N   C   L   G   F   G   G   K   G   Y   Y   D   D   K     291
AGC TGC GAT TTG GAG CTG CCC TTC ATC TGC GAG ATA GCG CCC TGA CCT GGC GCT    1047
 S   C   D   L   E   L   P   F   I   C   E   I   A   P   *   .   .   .     305
CTC GTG ACA CGT CTG AGG ACG CAA CAA CTG CGA ATG GAA ACG ACG TAC TCG GAA    1101
 .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .       *
AGT ATG GAC TGT TGC GAG GCA GAG TTA CGT TCA GAA GAC GTT GTA TAA TTG ATA    1155
 .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .       *
TGC GAA TAA ATA TTC GTT AGT AAC CTA AAA AAA AAA AAA AAA AAA AAC GGA ATC    1209
 .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .       *
CGC GGA ATT CCG                                                           1213
 .   .   .   .
```

Figure 17. cDNA and coded amino acid sequence for the 1213 bp clone 4. The 978 bp ORF encodes a 326 residue polypeptide including the signal sequence. The 305 amino acid protein has a calculated molecular weight of 34401 Da. Potential N-glycosylation sites at residues #29 and #208 are boxed. The start codon (ATG), initiating Met and N-terminal Ala are underlined. The stop codon at nucleotide 1022 is marked with an asterisk (*). 5' and 3' nontranslated sequences are indicated as in Figure 16. The signal peptide is underlined.

residue longer than that of clone 3. The 21 residue signal peptide for the clone 4 sequence is also one residue longer than the 20 residue signal for clone 3. Each sequence has two CRDs as defined by the four conserved Cys residues. Within each CRD, the 16 invariant residues which help define these proteins as authentic C-type lectins are present. These residues are later discussed in detail. The N-terminal CRD in each sequence is 95 residues in length. Sequences from clones 3 and 4 in the region of this CRD are 78% homologous when considering conservative substitutions. The C-terminal CRDs in each sequence are 98 residues in length and 90% homologous.

```
      Clone 3  MQLVTVCAALVAATVPCTLAAVDLFCSCQVRHHRDSTTAVHCSG
                                        |******|*•|||••||*|•|||*
      Clone 4 MACPLIIIFETRPTVLMVGTGAQQNSGGCKLRHHKESTMALHCSR

EQSGNKTISCQKAQVPDIPRDYHYVPGYALVKLYRIMMTWEEAKKACEAEGAKLAVPRDN
•|||||||**|||||||||||*|||||||||||*|||||||||•||||||||*||•|•|*
DQSGNKT--CQKAQVPDIPRGYHYVPGYALVKMYRIMMTWEEARKACEAEGAILALPKDS

HAYDGLKQIFKLGF---GVYWANIGITDHESEGIFSGVDGHPVSFLPWNPNEPNNAGGNE
||||||||•********||||||||||**||||||*||||*|||•|||*|||||||*|||||
HAYDGLKQVIIAEHKEEGVYWANIGITDQYSEGIFVGVDGLPVSYLPWRPNEPNNFGGNE

NCVNVNDKGQLNDWHCGNTAPFFCERRPSVGIPPSYVWLKDASRFYKVHAEKHVYAEAAR
|||*|*|||||||||*|•|**|||||||*|||||||||||||||||||||||||||*||||||
NCVYVDDKGQLNDWGCANAEPFFCERRLSVGIPPSYVWLKDASRFYKVHAEKHVHAEAAR

VCRSENATLAVPDTWDRVETLLRLLEPKEEFYLTGFTDEAVEGDFVTETGRHLKGMEFQV
||||||||||||||||||*||||||||||||||||||*|||||*||||||•*||||||
VCRSENATLAVPDTWDRVEALLRLLEPKEEFYLTGFTDEAAEGDFVAETGRHLRDMEFQV

WSPGEPNNDVDGKPENCLAFSGRGYYGDRSCEVELPFICEIAP-COOH-CLONE 3
|*|||||***||||||||•|*|•|||*|•||••||||||||||
WKPGEPNNNFLGKPENCLGFGGKGYYDDKSCDLELPFICEIAP-COOH-CLONE 4
```

Figure 18. Alignment of the deduced amino acid sequences for clones 3 and 4. The complete sequences are 80% identical and 85% homologous when considering conservative substitutions. N-terminal CRDs are 78% homologous when comparing clones 3 and 4, C-terminal CRDs are 90% homologous. The signal peptide for each is underlined as well as conserved N-glycosylation sites. 'l'=match, '*'=mismatch, '•'=conservative substitution.

3D Model Carbohydrate Recognition Domain

Conserved Regions Within Reference Proteins

The sequence defining the C-terminal CRD from clone 3 was modeled. Relative to the four known CRDs from the clone 3 and clone 4 proteins, this sequence has the highest homology to MBP and ESEL (31%). To define conserved regions in both MBP and ESEL, the two sequences were first aligned. This alignment is shown in Figure 19. The alignment meets two criteria: 1) conserved residues or residues with similarly defined functions (e. g. Ca^{2+} ligands) occupy identical positions, and 2) insertions or deletions as a result of sequence differences or manual manipulation do not occur in regions of known secondary structure. Using this approach, eight (8) structurally conserved regions were defined by superimposing the two reference structures over specified regions of the sequence (Table 2). Known secondary structures were used as a starting point. Regions

Table 2. Structurally conserved regions in rat MBP and human E-selectin used to construct the 3D model of the C-terminal CRD from clone 3. RMSD refers to the root-mean-square distances between analogous backbone atoms in the respective reference structures.

SCR	MBP aa#	ESEL aa#	Length*	RMSD (Å)
1 - α 1	116-132	7-23	17	0.42
2 - α 2	139-151	30-42	13	0.58
3 - β 2	154-161	48-55	8	0.64
4 - β 3	167-171	59-63	5	0.37
5 - LP	174-177	67-70	4	0.37
6 - LP	179-188	74-83	10	0.68
7 - β 3	193-199	88-94	7	0.82
8 - β 4/5	202-217	102-117	16	0.50

*73% of the 3D structure for the grasshopper CRD is defined by structurally conserved regions within the reference proteins.

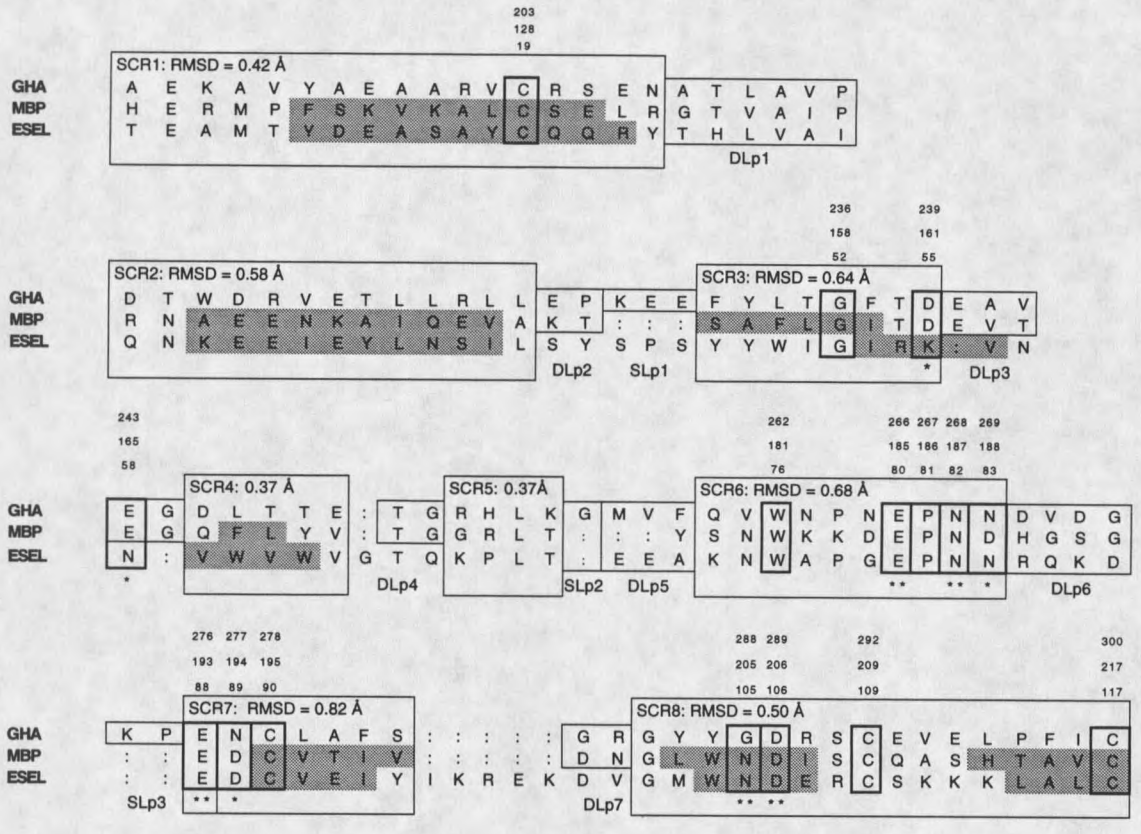


Figure 19. Alignment of sequences representing CRDs for GHA, MBP and ESEL. This format optimizes the alignment of structurally conserved regions between MBP and ESEL as well as conserved residues in MBP, ESEL and GHA. Secondary structure elements in MBP and ESEL are boxed gray. SCRs defined using the reference structures are outlined with a large box and labeled SCRs 1-8. The RMSD values for each SCR are included in the label e.g. SCR1: 0.42 Å. Consensus residues important for the C-type fold or known Ca^{2+} ligands are boxed and the amino acid position in the sequence is provided immediately above. The Ca^{2+} ligands involved in mannose binding in MBP are marked with two asterisks. Single asterisks indicate ligands which ligate a second Ca^{2+} .

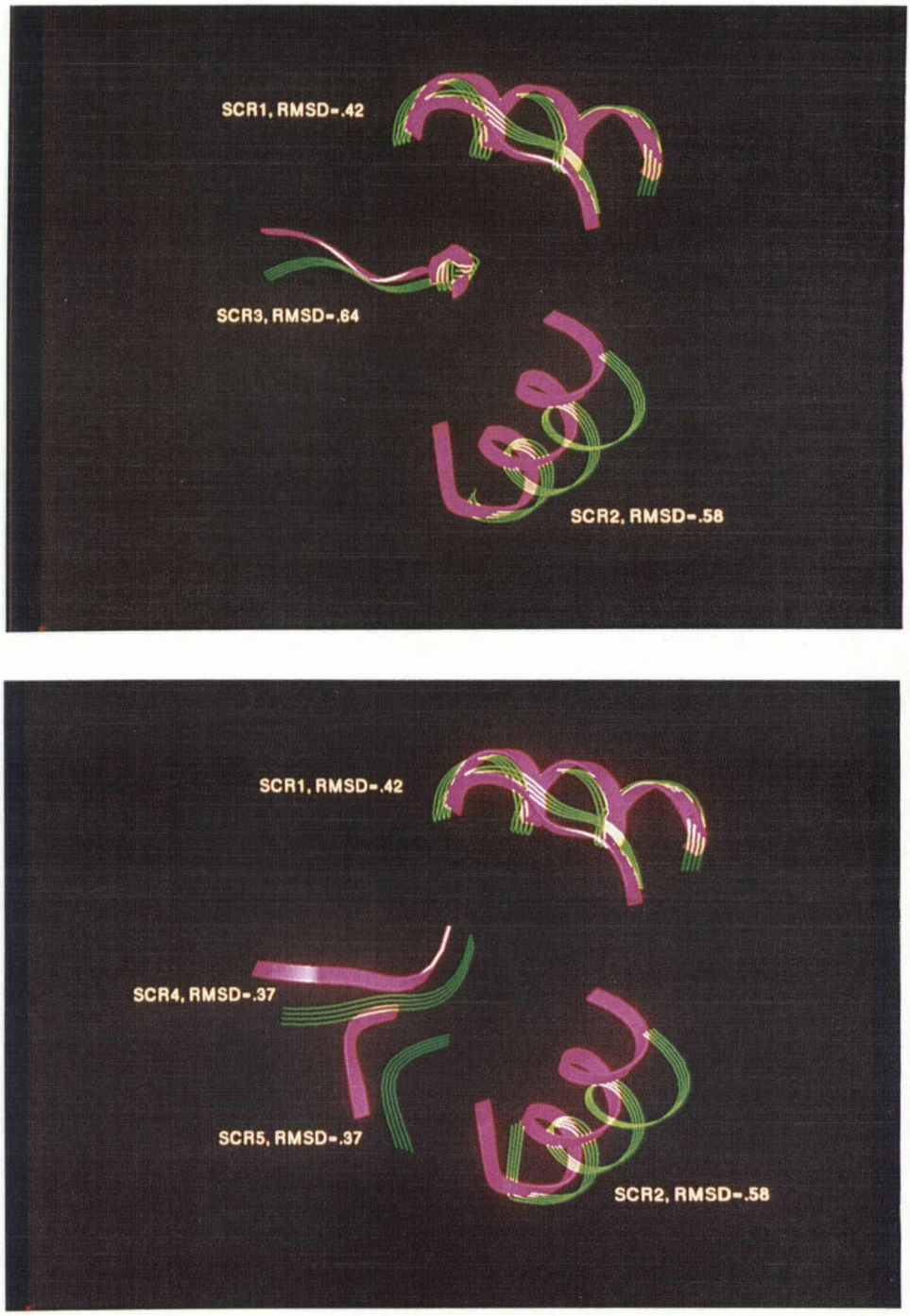


Figure 20. Structurally conserved regions 1, 2, 3, 4 and 5 between MBP (purple) and ESEL (green). The complete structures were first superimposed using the four conserved Cys residues defining the C-type architecture. Portions of the ribbons were removed to selectively illustrate each SCR. SCRs 1 and 2 are pictured throughout to help orient the viewer.

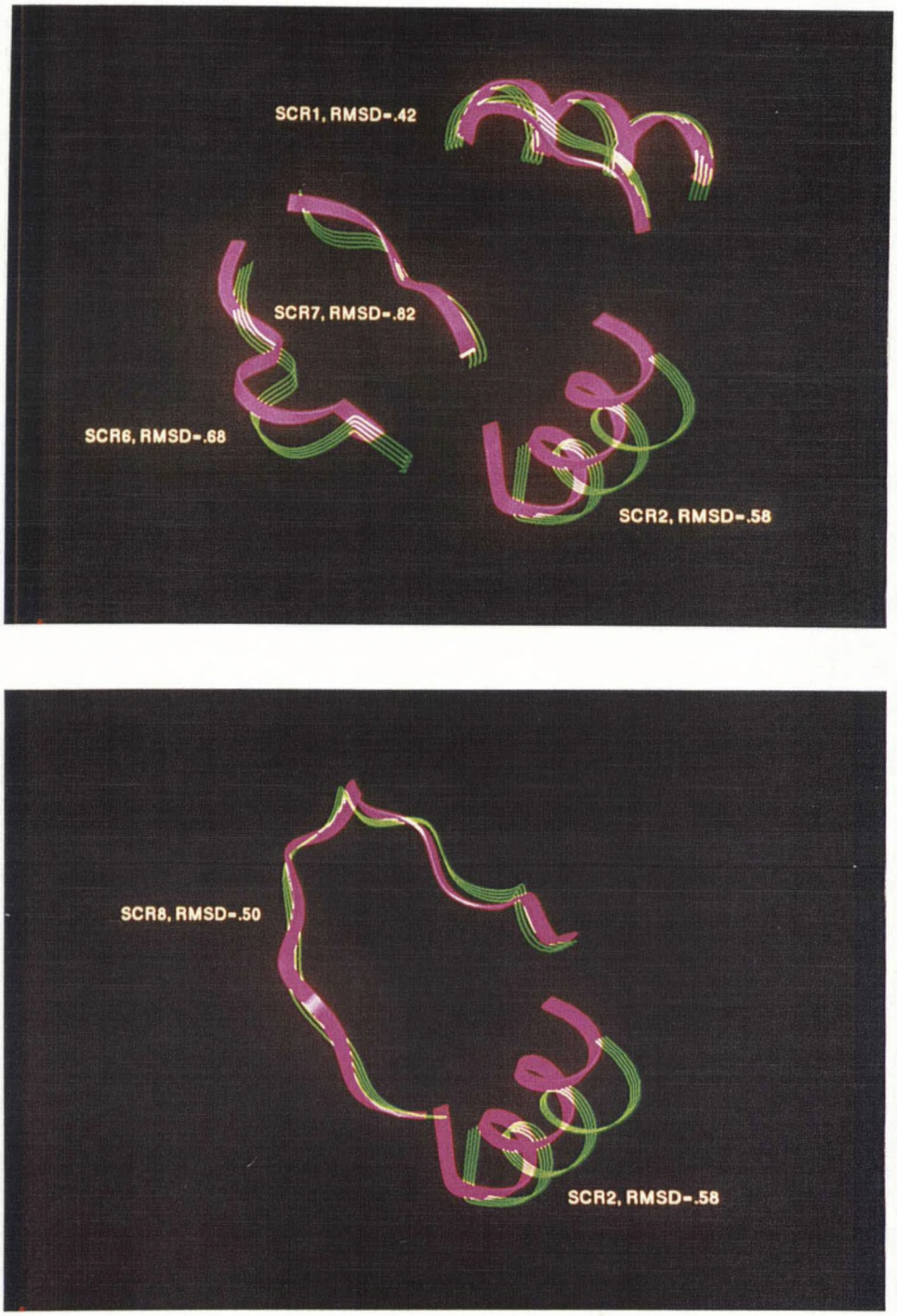


Figure 21. Structurally conserved regions 1, 2, 6, 7 and 8 between rat MBP (purple) and human ESEL (green). The two ribbon structures were superimposed using the four conserved Cys residues defining the C-type CRD.

of sequence between MBP and ESEL with RMSD values of ≤ .82 Å were defined as structurally conserved. These SCRs are listed in Table 2 and illustrated in Figure 19. RMSD values range from 0.37 Å for SCR4 & 5 to 0.82 Å for SCR7. Figures 20 and 21 illustrate the 3D similarities when comparing the references proteins within SCRs 1-8. Between regions defined as structurally conserved, there are seven "loops" (LP). Loops arise in regions where the reference proteins are significantly different. Analogous structural regions in MBP and ESEL with RMSD values ≥ .82 Å were defined as loops.

GHA CRD Sequence/Reference Alignment

The second stage of model construction requires alignment of the GHA sequence with SCRs of the reference proteins. Optimal alignment requires insertion of a one (1) residue gap between Glu59 and Thr60, and a five (5) residue gap between Ser92 and Gly 93. These gaps are illustrated in the Figure 19 alignment. All gaps are in regions that do not disrupt the 'C-type fold' by limiting changes to loop regions between SCRs and/or known 2° structures. The locations of gaps in this alignment are similar to those defined by Graves (1994) when comparing ESEL and MBP crystal structures.

Based on sequence homology, coordinates were assigned to the GHA model from either MBP or ESEL. For each SCR, a score for the homology between each reference sequence and GHA is computed. The coordinates from the reference sequence with the higher homology to GHA were used to construct that region of the model. The homology scores from each SCR and the source of coordinates used to build the model are listed in Table 3. Coordinates for loop regions were obtained using two different strategies; 1) if the size of the loop in MBP or ESEL was equal to that of GHA, the corresponding coordinates were used to construct the model loop, otherwise 2) the coordinates were found searching a structurally non-redundant database (loop search), as discussed in Methods (p 34). Loops 1-4, 6, 7 & 9 were constructed using coordinates from MBP or

ESEL. Assignment of coordinates from either MBP or ESEL in loops 1 and 6 was based on sequence homology as discussed for SCRs. For loops 2, 3, 4, 7 & 9, only one of the two reference proteins had loops of equal size to that of GHA. In each case, these coordinates were used to construct the model. Coordinates for loops 5 and 8 were found using the loop search procedure. The source of SCR and loop coordinates used to build the model are summarized in Table 3.

Table 3. Assignment of reference coordinates to the CRD model from MBP or ESEL.

SCR	[1]Homology	[2]Source	LP	Homology	Source
1	8.20/10.5	MBP	1	25.00/8.33	ESEL
2	10.00/11.5	MBP	2	na	ESEL
3	18.75/08.75	ESEL	3	na	MBP
4	8.00/00.00	ESEL	4	na	MBP
5	2.50/22.5	MBP	5	na	Loop Search
6	41.20/46.25	MBP	6	12.50/7.5	MBP
7	34.30/35.7	MBP	7	na	ESEL
8	11.80/14.37	MBP	8	na	Loop Search
			9	na	Loop Search
			10	na	MBP

[1]Homology score computed between GHA model sequence and each of the two reference proteins, MBP and ESEL. A score is computed separately for the sequence defining each SCR or LP.
[2]Source is the reference protein from which the coordinates were taken to construct portions of the model represented by each SCR or Loop.
[3]na=not applicable
[4]Loop Search indicates that the coordinates used to construct this part of the model CRD were obtained by searching a structurally non-redundant database of structures as discussed in the Methods (p 34).

GHA Model Structure

The final model structure for the C-terminal CRD from clone 3 is shown as a ribbon structure along side the ribbon structure for MBP in Figure 22 (top). The overall fold of

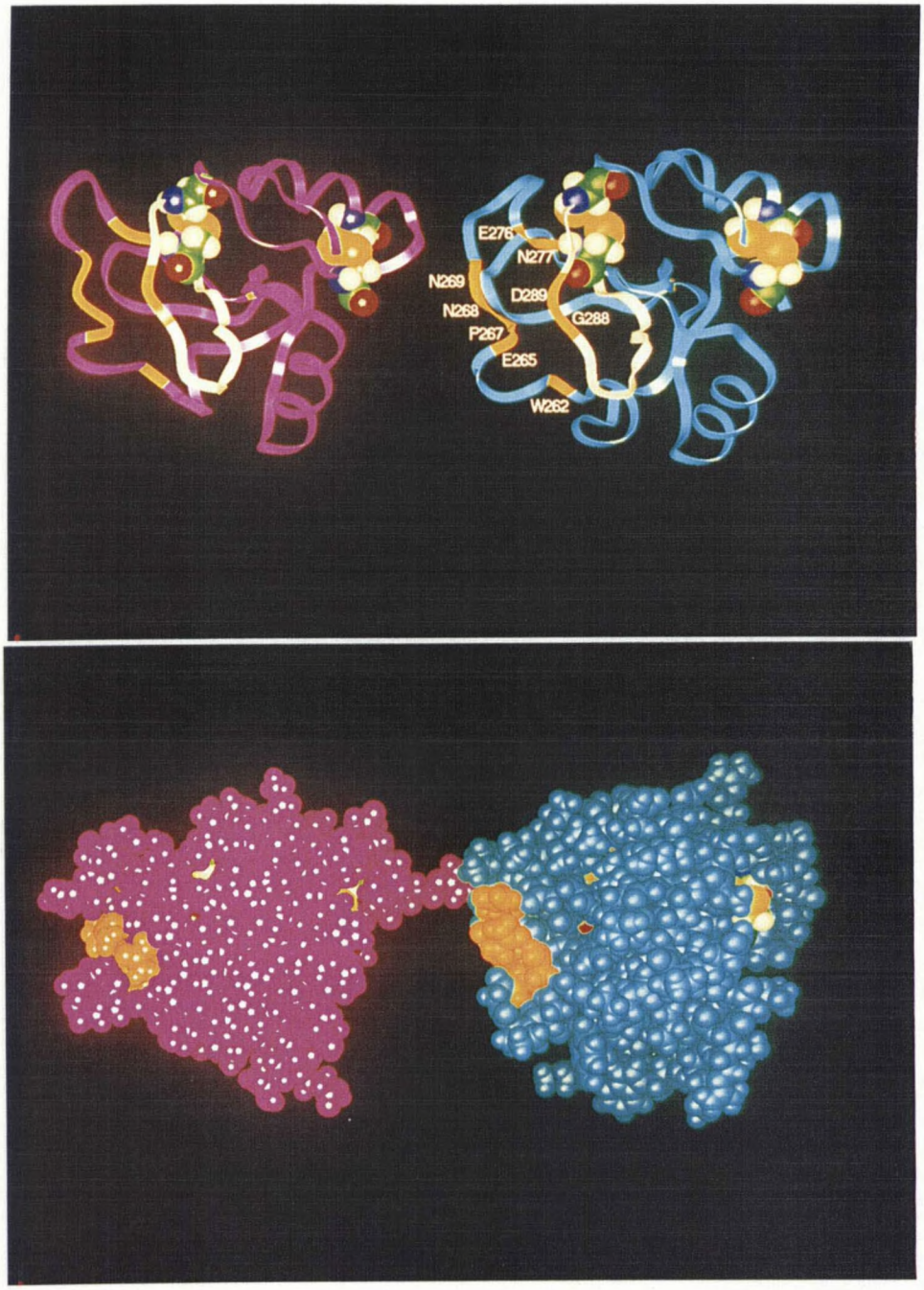


Figure 22. Ribbon structures (top) for the homology model of the C-terminal CRD from the grasshopper lectin, side-by-side with the MBP crystal structure. Space filling models of both MBP and GHA ribbon structures are shown at the bottom. MBP is purple. GHA is blue. The large loop formed by the characteristic bridging of conserved Cys residues is purple in MBP and blue in GHA. The small loop is white in each. Conserved residues representing Ca^{2+} ligands in MBP are orange in both MBP and GHA. Cys residues are represented with space-filling atoms and colored by atom e. g. green=C, blue=N, red=O, white=H and yellow=S. Conserved residues for GHA are labeled and numbered by sequence position.

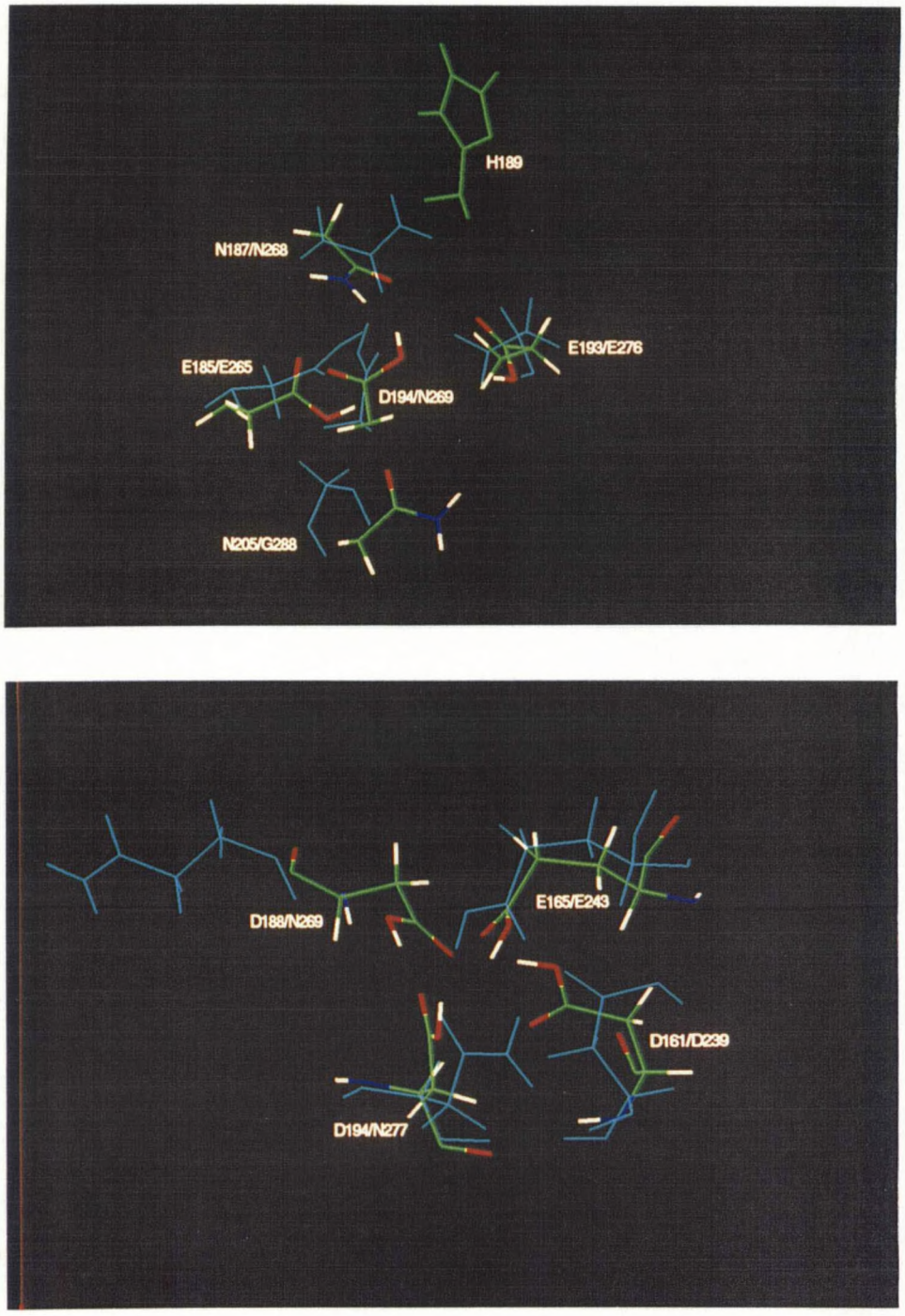


Figure 23. Ca^{2+} ligands for sites 1 (bottom) and 2 (top) from MBP on which equivalent residues in GHA have been superimposed. Residues from MBP are colored by atom e. g. green=carbon, blue=nitrogen, red=oxygen and white=hydrogen. Residues from GHA are colored blue. The similar orientation of equivalent residues in GHA suggest this CRD may bind Ca^{2+} in the same way.

the final model is similar to that seen for both MBP and ESEL (Figure 2 p 11). The four conserved Cys residues defining the C-type motif in MBP are shown as space-filled models and colored by atom e. g. yellow=S, blue=N, red=O, green=C and white=H. The relative location and orientation of these side chains is similar for both MBP and GHA. Sulfur atoms in the model are oriented such that the distances between sulfurs allows for the potential formation of the disulfide bridges as seen in MBP and ESEL. The large loop in the C-type motif is colored purple for MBP and blue for GHA. The small loop in each is colored white. The insect CRD has ≈ 50% regular secondary structure defined as α helices, β sheets and turns. The remaining structure is irregular and cannot be classified as α helix or β sheets. These general features are also characteristic of MBP and ESEL. Strictly conserved residues in this family of CRDs are found in these extended regions. These residues are colored orange in both MBP and GHA. With the exception of G236 which is found in a tight turn at the core of the model structure W262, E265, P267, N268, N269, E276, N277, G288 and D289 are all located at or near the surface in the same general region as seen for both MBP and ESEL. This similarity is demonstrated in Figure 22 (bottom) using space-filling models for GHA and MBP CRDs. These residues are strictly conserved and their counterparts in MBP and ESEL are demonstrated Ca^{2+} ligands. This suggests the CRD from grasshopper may ligate Ca^{2+} using a similar strategy. Ca^{2+} ligands for sites 1 and 2 in the MBP CRD are pictured in Figure 23. In each, equivalent residues from GHA are superimposed to demonstrate the similar 3D orientation of these potential ligands. The CRD sequence folded as seen in this model also positions one of two potential N-glycosylation sites at the surface of the CRD with N207 positioned for potential covalent attachment to carbohydrate (Figure 24). The glycosylation site at N29 lies outside of either CRD in both clones 3 and 4.

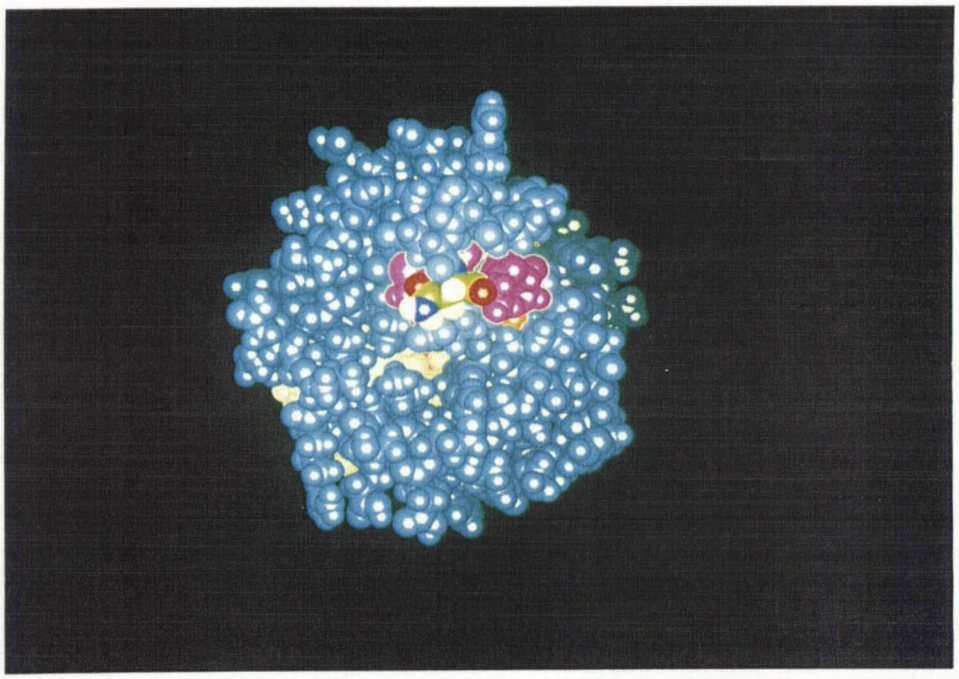


Figure 24. Model CRD in Figure 22 rotated 90° left on the Y-axis to show the location of Asn207 at the surface of the structure. This residue represents one of two glycosylation consensus sequences (NXT, s or C) available for covalent interaction with carbohydrate. The N atom from Asn207 is colored dark blue. This potential glycosylation site is ≈ 180° opposite of the protein's carbohydrate binding site and corresponding Ca^{2+} ion/ligands.

GHA Model CRD Structure Assessment

Assessment of RMS Differences

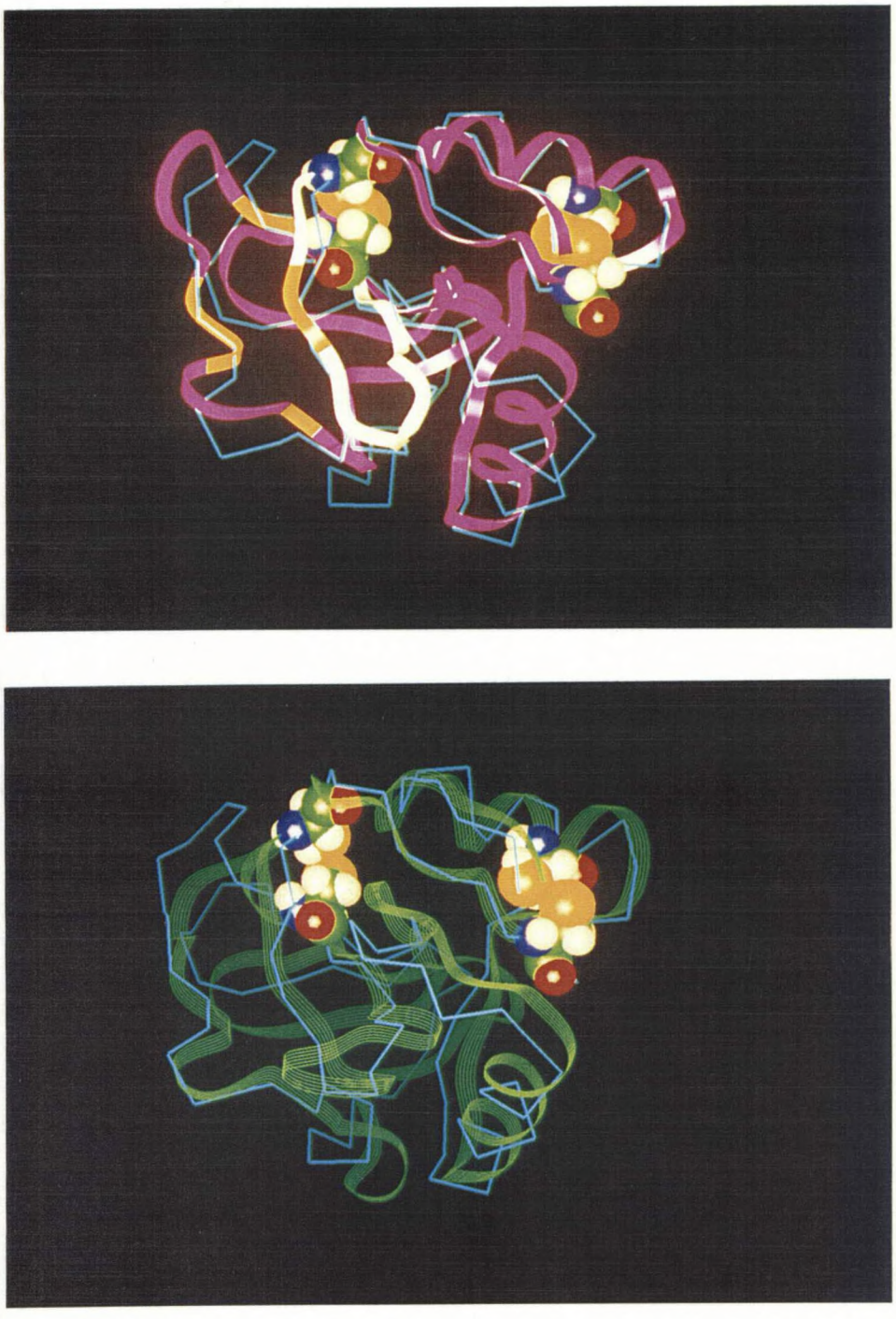
The RMS difference between the GHA model, MBP and ESEL were measured to assess the similarity of the new model to each reference structure. The model CRD was superimposed separately on each of the reference proteins using the Cα from each of the four conserved Cys residues known to form the C-type lectin fold in MBP and ESEL. The difference between the model CRD and MBP, and the model CRD and ESEL was 0.96 Å and 0.82 Å respectively. Superimposition of these structures is shown in Figure 25. This

compares favorably with the RMS difference of 0.74 Å when comparing the two reference structures as shown in Figure 26. These measurements demonstrate quantitatively the new model structure is as similar to each reference as each reference is to one another.

Assessment of $\phi-\psi$ Geometry

Comparison of the $\phi-\psi$ angles in the new model with those from ESEL and MBP help to assess the final structure. Ramachandran plots of MBP, ESEL and the CRD model are similar (Figures 27, 28 and 29). The ϕ-ψ geometry in the model CRD suggests the 3 dimensional structure occupied by this sequence is reasonable. Among non-glycine residues, 99% (97/98) maintain ϕ-ψ angles that fall within allowed regions of the Ramachandran plot. Of these 97 residues, 15 maintain ϕ-ψ geometries in the 'allowed' region of plot, while the remaining 82 residues maintain angles within the 'core' region. 'Allowed', 'disallowed', 'generously allowed' and 'core' regions have been defined previously by Morris (1992). The ϕ-ψ angles for Phe232 lie in the 'disallowed' region of the plot. This residue is labeled in Figure 27. These values are 83.7 and 145.8 for ϕ and ψ respectively. Corresponding values in MBP are -61.6 and 122.2. This alone does not detract from the 'reasonableness' of the model. However rationalization of these geometries may lead to useful information about the molecule or the ability to improve the model. This residue represents the N-terminal residue in Loop1 suggesting the spurious ϕ-ψ geometries may be the result of a poor loop/SCR splice-junction.

The MBP CRD contains 101 residues. Of the 94 non-glycine amino acids, 99% (93/94) fall within allowed regions of the plot and 88% of these residues maintain geometries that lie within the "core" region. Eleven residues fall within the 'allowed' region. As in the model CRD, one residue has angles which are "disallowed" (Lys37). ESEL has 109 total residues. 100% (106/106) of the non-glycine residues have ϕ-ψ



Figures 25. GHA model CRD (blue) superimposed on MBP (purple). Bottom, GHA model CRD superimposed on ESEL (green). RMS differences equal 0.96 Å and 0.87 Å respectively.

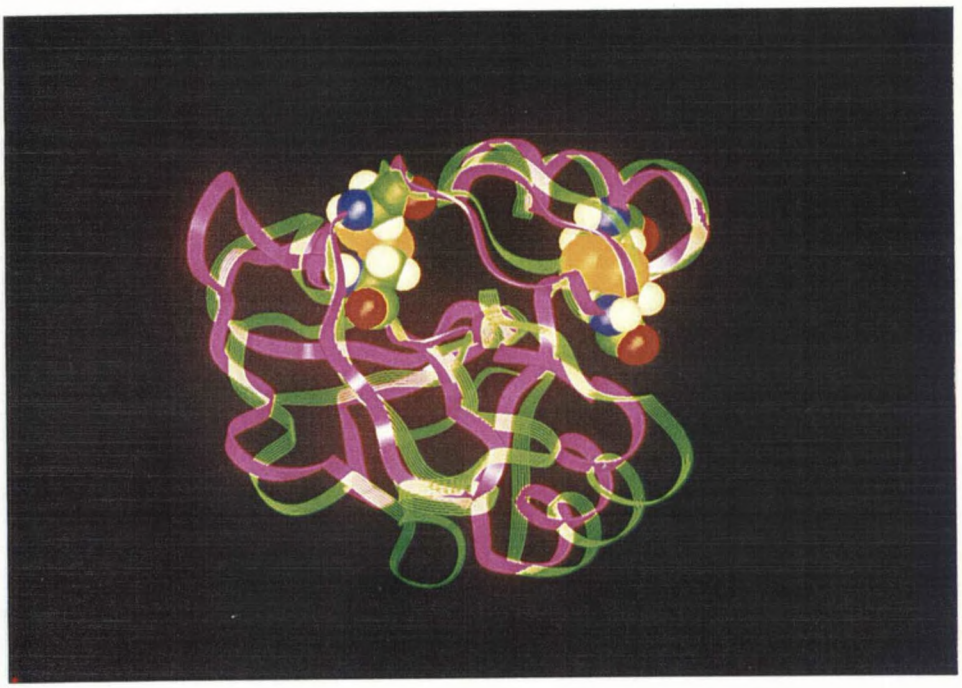


Figure 26. Reference structures MBP (purple) and ESEL (green) when superimposed using the four conserved Cys residues have an RMSD of 0.74 Å.

geometries falling within acceptable regions of the plot. 86% of these residues fall within "core" regions while 10 residues (11%) maintain geometries which falling within the "allowed" regions and two residues Tyr49 and Thr7 fall in the "generously allowed" regions of the plot. No residues have ϕ-ψ geometries which are disallowed. Together, these comparative data suggest the GHA model CRD is a reasonable 3D structure for the sequence representing the C-terminal CRD from clone 3.

Assessment of Dihedral Angles

The range of peptide bond dihedral angles for the GHA model were compared to that observed in both MBP and ESEL. Deviations in the range of angles, relative to those

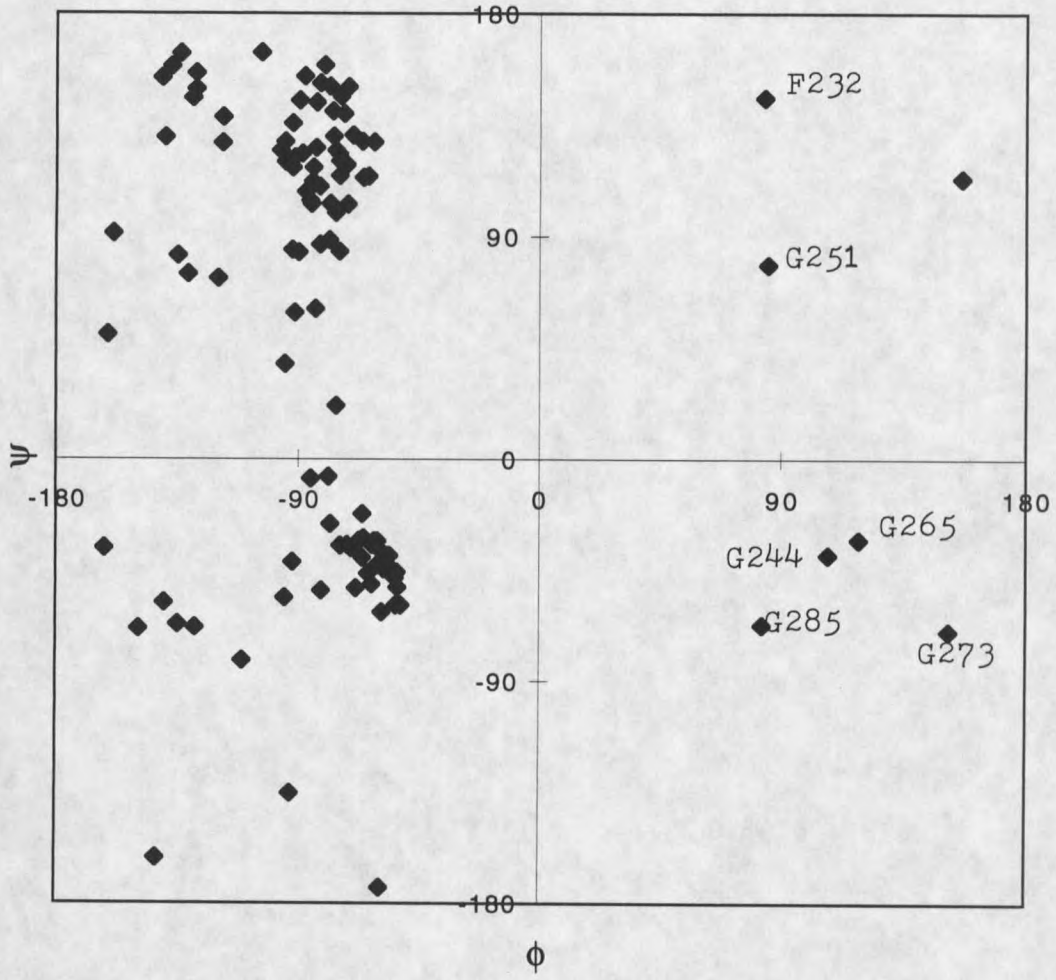


Figure 27. Ramachandran plot of ϕ-ψ angles in the GHA model CRD. 99% (97/98) of non-glycine residues maintain allowed ϕ-ψ angles. Of these residues, 15 maintain ϕ-ψ geometries in the 'allowed' region of the plot while the remaining 82 residues maintain angles within the 'core' region. The ϕ-ψ angles for Phe232 lie in a 'disallowed' region. 'Core', 'allowed', 'generously allowed' and 'disallowed' regions of the plot have been previously defined by Morris (1992).

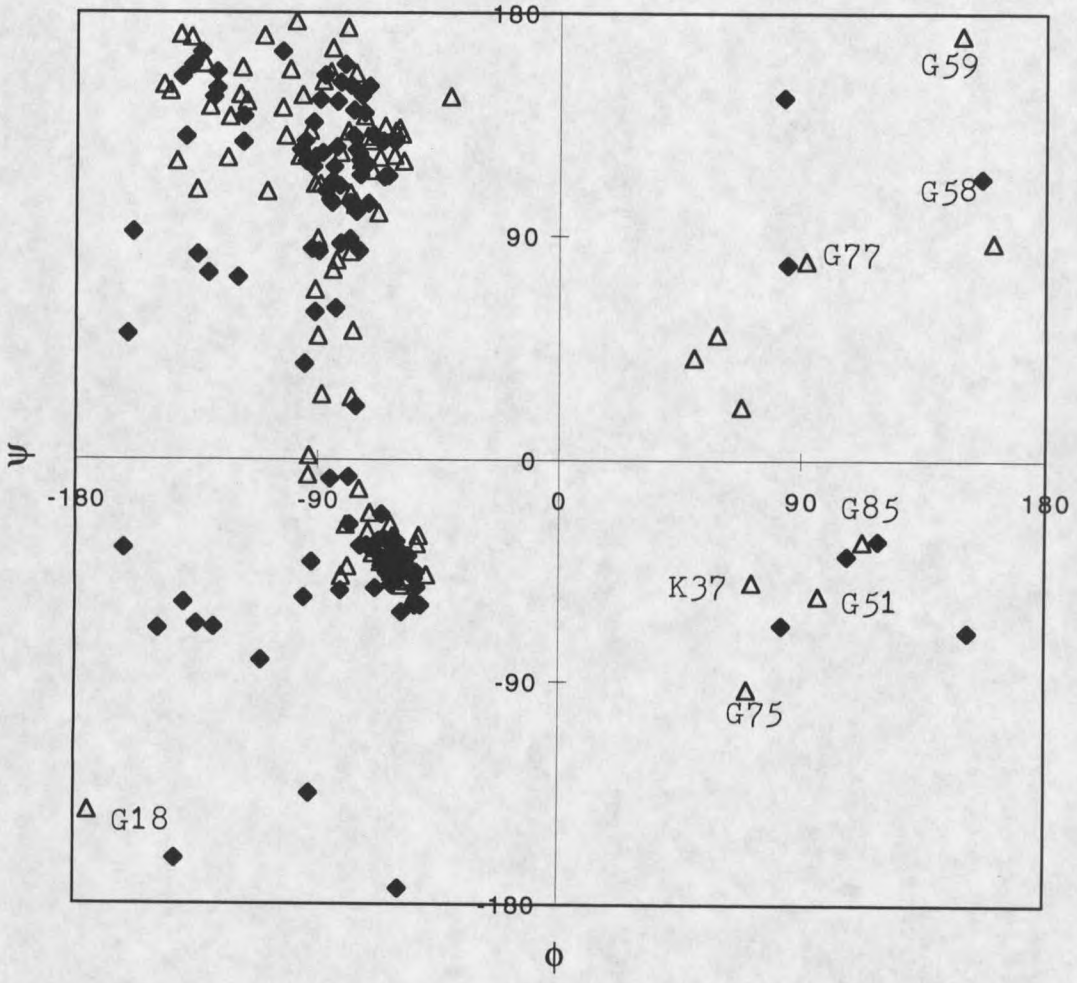


Figure 28. Comparison of Ramachandran plots for the GHA model CRD and rat MBP. The plot of MBP is similar to that of GHA. Of the 94 non-glycine amino acids, 99% (93/94) fall within allowed regions of the plot. 88% (83/94) lie within the "core" region and 11 maintain ϕ-ψ geometries that fall within the "allowed" region. Lys37 has angles which are "disallowed". This residue is labeled on the plot. Open triangles represent the ϕ-ψ angles for each residue in MBP. Closed triangles represent GHA.

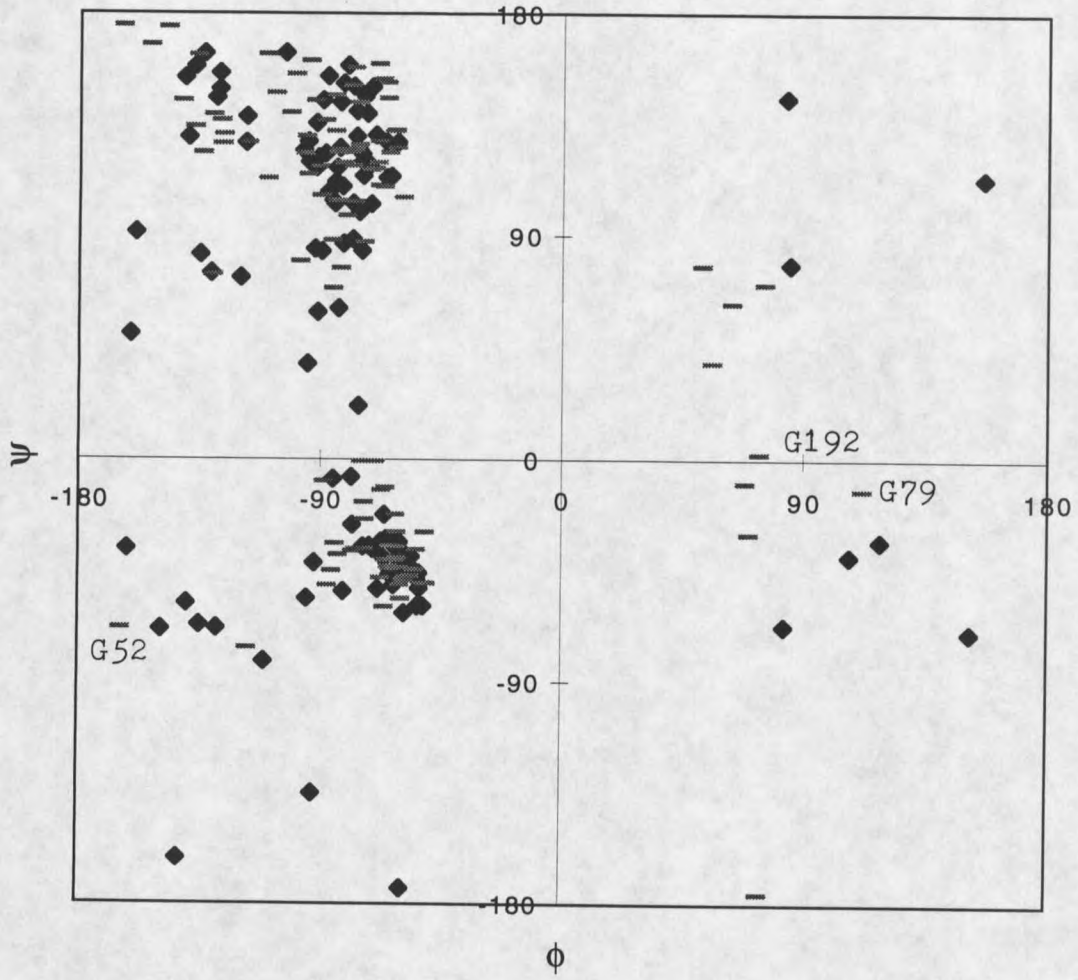


Figure 29. Comparison of Ramachandran plots for the GHA model CRD and human ESEL. The plot for ESEL is similar to that of GHA. 100% (106/106) of the non-glycine residues have ϕ-ψ geometries falling within acceptable regions of the plot. 86% of these residues fall within "core" regions while 10 residues (11%) maintain geometries falling within the "allowed" regions. No residues have ϕ-ψ geometries which are disallowed. Gray dashes represent the ϕ-ψ angles for each residue in ESEL. Closed triangles represent GHA.

of ESEL and MBP, may imply the presence of a structural flaw or provide information that may be used to improve the model. A list of peptide dihedral angles for each reference structure and the model CRD can be found in Appendix E. In the GHA model, 105 of 107 peptide bond angles (98%) are within 25° of ±180°. In MBP, 100 of 101 dihedrals (99%) occupy this range and 99% of the peptide bonds in ESEL fall within 31° of ±180°. The major exceptions in each structure are equivalent cis Prolines at positions 267, 186 and 81 in GHA, MBP and ESEL respectively (Figure 19). The peptide bond at Pro267 in the model CRD occupies a dihedral geometry of 7.1°. Similarly, for MBP and ESEL, these geometries are 5.1° and 3.8°.

In the GHA model, Thr209 occupies a dihedral angle of -142.3°. This angle is a significant deviation from the favored ±180°. However, Asp87 in ESEL maintains a similar 149.1° dihedral angle suggesting the Thr209 dihedral may not significantly detract from the model structure. The large deviation at position Thr209 likely arises from the loop coordinates used to construct this region. Overall, the range of peptide bond dihedral angles in the GHA model indicate the 3D structure is a reasonable representation of the GHA CRD sequence.

Comparison of the disulfide dihedral angles between bridged Cys residues in this family of CRDs was also used to assess the reasonableness of the final model CRD. Dihedrals from the grasshopper model CRD are 101° for the disulfide dihedral between Cys278 and Cys292 bridged to form the small loop, and -86° for the dihedral between Cys203 and Cys300 bridged to form the large, outer loop. The dihedral between conserved Cys residues 195 and 209 forming the small inner loop in MBP is 96.7°. This angle between Cys residues 128-217 forming the large outer loop is -79.70°. Similarly for the CRD from ESEL, equivalent dihedrals between Cys residues disulfide bridged to form

the small and large loops are -84° and -89° respectively. These similarities suggest the orientation of Cys residues to form the C-type motif in the GHA CRD is reasonable.

Assessment of Sequence/Structure Compatibility

Three dimensional (3D) Profiles were used to assess the compatibility of the sequence defining the C-terminal CRD from clone 3 with the 3D dimensional model structure (Eisenberg, 1992). Figure 30 illustrates profiles for MBP and ESEL reference proteins. These plots are somewhat different, but the 3D probability scores for each remain relatively high and thus each represents a profile which may be typical for any homologous sequences folded in this manner. Favorable comparison of the 3D profile for the model CRD from grasshopper lectin with each of the reference profiles (Figure 31 and 32) suggests the overall model CRD is folded correctly and that there are no significant errors in the core regions. Such errors would show up in the profile as regions radically different when compared to the reference profiles or very low average 3D probability scores in certain regions of the plot. Total probability scores (S) between the three structures compare favorably. $S_{MBP} = 33.56$ as calculated in this study and is consistent the $S_{MBP}=38.43$ reported by Bajorath (1995). The difference can be attributed to the number of amino acids used to define the MBP CRD. For ESEL, the calculated value of $S_{ESEL}=$ 48.13. This value is substantially higher than that seen for MBP because the portion of the graph for residues 1-50 for ESEL contains probability scores substantially higher than those seen for either MBP (Figure 30) or the model CRD in this region of the sequence (Figure 32). Also, the sequence defining the CRD for this reference protein is longer than either of those for MBP or the model CRD and this too will make the value of S_{ESEL} higher since the score is a sum of all scores for each residue in the sequence. Finally, the reported value of $S_{GHA} = 33.11$ for the GHA model CRD, compares favorably to both MBP and ESEL. These 3D profile data suggest the model structure for the CRD from this

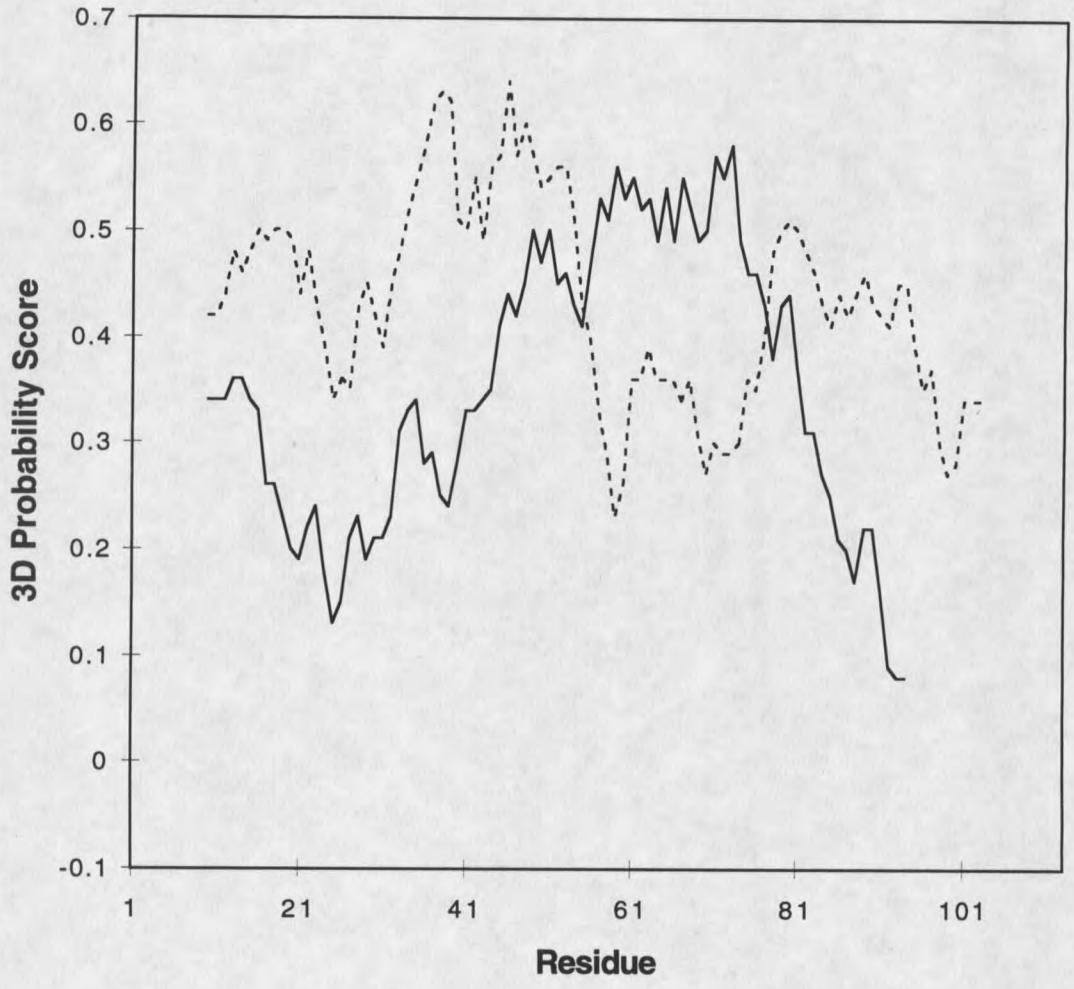


Figure 30. 3D profiles for MBP (black) and ESEL (dashed). $S_{MBP} = 33.56$, $S_{ESEL} = 48.13$. These values differ because of the difference in 3D probability scores seen in the first part of the two sequences.

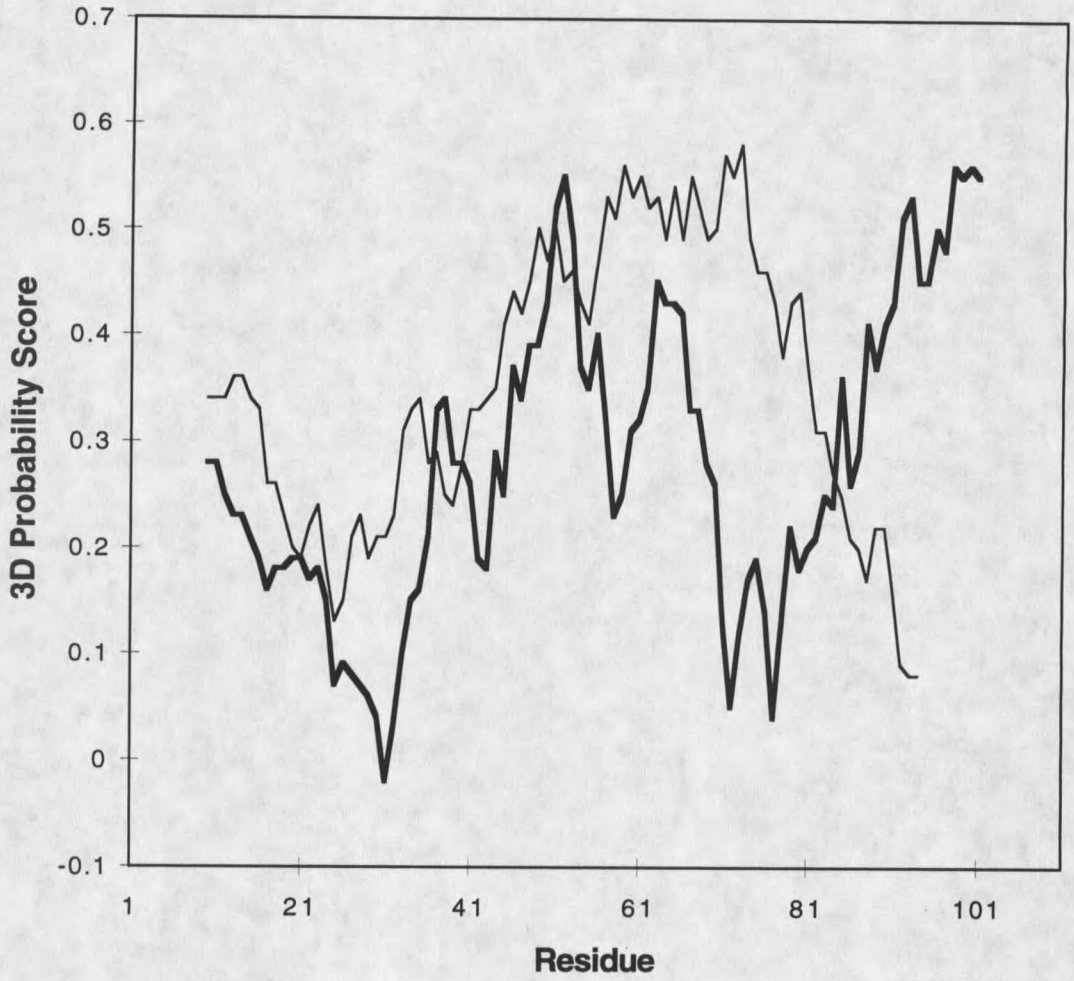


Figure 31. 3D profiles for MBP (black) and the model CRD (heavy black). $S_{MBP} = 33.56$ and $S_{GHA} = 33.11$. These values compare favorably and the 3D profiles for each are similar.

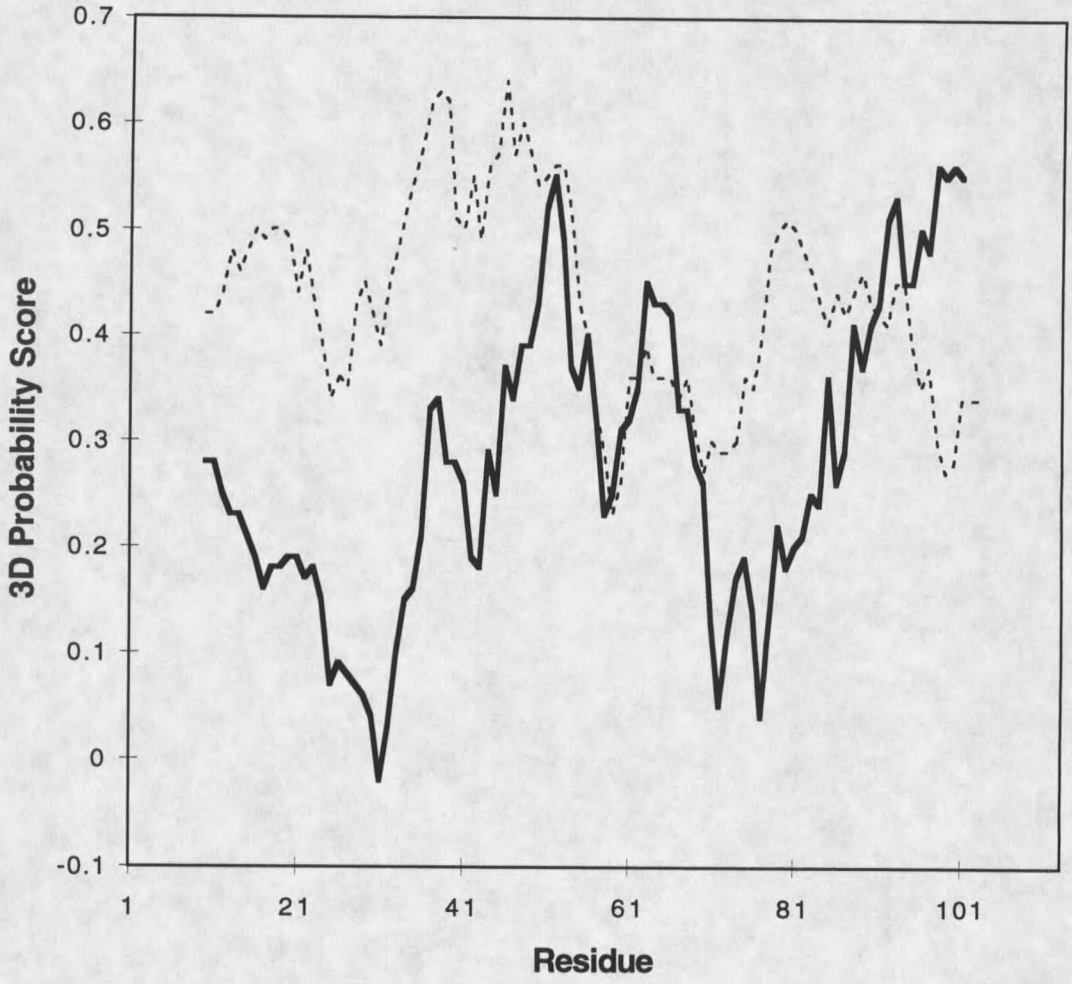


Figure 32. 3D profiles for ESEL (dotted) and the model CRD (heavy black). S_{ESEL} = 48.13 and S_{GHA} = 33.11.

grasshopper lectin is a "reasonable" 3 dimensional representation of this primary sequence (Bajorath, 1994 & Bajorath, 1996).

DISCUSSION

The purpose of this work was to extend the characterization of two lectins from the grasshopper, *Melanoplus differentialis.* The focus was defined by three goals: 1) to complete a clone 3-specific cDNA sequence representing a full-length ORF for a lectin from the grasshopper, *Melanoplus differentialis*, 2) to complete a clone 4-specific cDNA representing a full-length ORF for a second lectin from this grasshopper and 3) to construct a computer-based homology model for a CRD from grasshopper lectin using the deduced amino acid sequence and the crystal structure coordinates from rat MBP and human ESEL. Results presented in this thesis indicate these goals to characterize a lectin(s) from grasshopper have been achieved.

Previous Work and Origin of Clones 2, 3 and 4

Seventeen positive plaques were isolated from the original cDNA library. Three inserts were selected based on their larger size, subcloned in pGEM plasmid vector and labeled clones 2, 3 and 4 (Rognlie, 1991). Clone 3 is the original 879 bp cDNA sequenced and reported by Rognlie (1991). This cDNA contained an ORF including sequence coding for an initiating Met and signal peptide, but lacked sequence for the stop codon and polyA tail. The ends of clones 2 and 4 were sequenced in search of a clone 3-specific fragment containing a stop codon and polyA tail. Clone 2 was partially sequenced, but did not contain sequence representing the 3' end of the original clone 3. Approximately 200 bp of

clone 4 sequence was homologous to that of clone 3 and contained an ORF, a stop codon and polyA tail. This sequence was used to complete a composite cDNA representing a full-length mRNA for a grasshopper lectin. This 3' sequence from clone 4 was not exactly identical to that of clone 3 in the region of overlap, suggesting the clone 4 cDNA may represent an mRNA coding for a second lectin. This presumption was the basis for work to complete clone 3 and clone 4-specific cDNAs representing full-length ORFs for two lectins from the grasshopper.

Strategy to Complete the Clone 3 cDNA

The complete cDNA sequence for clone 3 is 1220 bp long and contains a 324 residue ORF coding for an initiating Met, signal peptide and stop codon through the polyA tail (Figure 10 p 46). This sequence is a product of the original 879 bp clone 3 and a newly cloned cDNA containing the complete 3' end. This recombinant plasmid was designated pGEM31. The original 879 bp clone 3 was incomplete. Sequence from clones 2 and 4 did not represent the authentic 3' end of the clone 3 sequence. The most efficient strategy for completing the 3' sequence was to search the remaining 14 clones for a either a single cDNA representing a full-length mRNA for clone 3, or a fragment representing the 3' end. This strategy takes advantage of the increased probability of finding the 3' end of clone 3 since during cDNA synthesis, the 3' end is transcribed first. RACE, or the construction of a new cDNA library were viable alternatives to this strategy, but were considered premature before analysis of the remaining 14 clones.

A PCR based search of the remaining 14 positive plaques using a clone 3-specific primer indicated that 12 of the remaining 14 were indeed clone 3-specific, but no products

long enough to include the 3' sequence were found (Figure 6 p 40). Eleven of the 12 clone 3-specific PCR products were the same size, ≈ 550 bp. The expected size of a clone 3-specific product containing the complete 3' end was ≈ 730 bp. These products are identical in size to that of the positive control (clone 3 plasmid) suggesting all 11 are truncated as observed in the original clone 3. Based on alignment with clone 4 sequence, it was evident an EcoRI restriction site was present at the point of truncation (Figure 10 p 46). Because the point of truncation is identical in all 11 clone 3-specific cDNAs, it is likely the result of inefficient protective methylation of the cDNA prior to ligation into the original λgt11 library vector. This would allow the cDNA to be restricted when creating sticky ends for ligation into the vector. It was concluded the missing 3' end of the original clone 3 cDNA was not accessible via these methods and may not be available in this cDNA library.

RACE was used to isolate and clone the 3' end of the original 879 bp cDNA. This strategy is based on RT• PCR with a gene-specific primer. The clone 3-specificity of the '879PCR' primer was established, thus making 3' RACE a rational means to selectively isolate the 3' end of clone 3 from polyA$^+$ mRNA. Figure 10 (p 46) illustrates the complete clone 3 cDNA sequence. The position of the clone 3-specific oligo (879PCR) is underlined and labeled in this figure. The 5' end of this primer anneals at nucleotide #392. At nucleotide #491, #785 and #881 are NcoI, XhoI and EcoRI restriction sites. These sites were used to verify the clone 3-authenticity of the ≈ 800 bp RACE product (Figure 7 p 42 & Figure 8 p 43). The NcoI site is 100 bp 3' of the primer annealing site. The ability to cleave the ≈ 800 bp product at this site served two purposes: 1) successful restriction would help verify the clone 3 authenticity of this fragment before cloning and sequence analysis, and 2) restriction at this site would also create a sticky end for ligation into

pGEM5fz(+) plasmid vector. The location of this NcoI site also allowed for the overlap of 388 bp of the original 879 bp clone 3 with the new fragment containing sequence representing the complete 3' end. Identical sequence in this region proves the clone 3 authenticity of the newly obtained RACE product (Figure 9 p 45). The combination of sequence for the original 879 bp clone 3 with that sequence isolated via RACE provides a complete clone 3-specific cDNA.

Strategy to Complete the Clone 4 cDNA

The complete clone 4 cDNA sequence is 1213 bp long and codes for a 326 residue ORF including an initiating Met, signal peptide and stop codon through the polyA tail. The sequence is the product of the original 1079 bp clone 4 and its newly cloned 5' end. Sequence analysis of the complete clone 4 was done first and determined this cDNA to be 1079 bp long. Alignment with the clone 3 cDNA indicated this sequence was homologous, but not identical and thus clone 4 represented an mRNA for a second lectin from this grasshopper. The alignment also showed the 5' end of this cDNA to be incomplete, missing sequence coding for an initiating Met, signal peptide and several N-terminal amino acids. As discussed above for clone 3, one strategy for obtaining a cDNA representing this 5' sequence was to search the remaining positive clones for this fragment. The PCR-based search of the remaining 14 positive plaques using the clone 3-specific primer discussed above clearly indicated that 12 of the 14 were clone 3-specific. However, two clones did not amplify (Figure 6 p 40). These clones were amplified using the clone 4 specific primer '1150PROB'. These products were ≈ 350 bp and nominally of sufficient length to contain any new 5' sequence for the original clone 4 (Figure 12 p 49). Based on alignment with

clone 3, an additional 69 bp would provide 5' sequence encoding the initiating Met and signal peptide, therefore two clones were cycle sequenced. Each was determined to be clone 4-specific but truncated at the same position as the original 1079 bp cDNA. These sequence results are not reported in this thesis. It was concluded the 5' end of clone 4 was not accessible using this method.

RACE was used to isolate and clone the 5' end of the original 1079 bp clone 4 cDNA. Figure 14 (p 52) illustrates the complete clone 4 cDNA sequence. The position of the clone 4-specific oligo (4036) is underlined. The 5' end of this primer anneals at nucleotide #225 and its 3' end is oriented towards the 5' end of the sequence. The RACE product is ≈1100 bp (Figure 12 p 50). A clone 4-specific restriction site to verify the authenticity of this fragment prior to cloning and sequence analysis was not available. This fragment was cloned directly into pGEM-T vector and sequenced. This eliminated the use of restriction enzymes for ligation and allowed the entire fragment to be cloned without risk of cutting it into smaller pieces. The position of the 4036 primer allowed for the overlap of 96 bp of the original 1079 bp clone 4 sequence with the new cDNA fragment containing sequence representing the 5' end. Identical sequence in this overlapping region proves the clone 4 authenticity of the new sequence (Figure 13 p 51). The combination of sequence for the original 1079 bp clone 4 with that sequence isolated via 5' RACE provides a complete clone 4-specific cDNA.

The 5' RACE procedure was attempted unsuccessfully with two primers prior to the use of the 4036 primer discussed above. The 1150PROB primer was used successfully to screen the remaining positive plaques, but was not capable of isolating a clone 4-specific 5' RACE product. Two fragments generated using this primer were cloned and partially sequenced, but determined not to be clone 4-specific. A second primer, pGEM41NP, was

also employed and two resulting DNA fragments were cloned and partially sequenced. Here again, neither was determined to be clone 4-specific. The 4036 primer which was ultimately successful in the 5' RACE procedure arose from work to determine the presence or absence of introns within the coding regions of clones 3 and 4 (Gedik, 1996).

cDNA Sequences for Clones 3 and 4

The cDNA sequences for clone 3 (Figure 10 p 46) and clone 4 (Figure 14 p 52) are 81% identical (Figure 15 p 54) including the 3' nontranslated regions and polyA tail. These sequences represent distinct genes coding for two separate lectins from the grasshopper, *Melanoplus differentialis*. Each is complete as defined by sequence coding for an ORF, initiating Met, signal peptide and stop codon through the polyA tail. The most important differences between these two sequences are at nucleotide #203 of the clone 3 sequence and #398 of the clone 4 sequence where there exist 6 and 9 nucleotide inserts, respectively. The net difference is 3 additional nucleotides in the clone 4 sequence coding for 1 additional amino acid in the mature protein. These regions represent the two greatest differences between the sequences and were the basis for the design of clone 3 and clone 4-specific primers used in the RACE procedures discussed above. No other regions between the two sequences are sufficiently different to allow for the design of primers which would distinguish clone 3 from 4. Clone 3 and 4 have identical EcoRI sites at positions #881 and #891, respectively. It is interesting that all of the clone 3 positive inserts are truncated at this site, but the two clone 4 positive inserts are not. Clone 4 has an EcoRI site at position #126 that is not present in the clone 3 sequence. The clone 4 positive cDNAs are truncated at this position. This explains why there were no complete clone 3 or clone 4-specific

cDNAs in the original library. As discussed for the clone 3 sequence, this truncation must be the result of incomplete protective methylation of cDNA prior to restriction and ligation during construction of the original library.

The cDNAs for each of clones 3 and 4 encode two duplicated, tandem CRDs. cDNAs representing these CRDs are similar, but encode distinctly different CRDs when compared to one another, and when comparing CRDs between clones 3 and 4. This duplication is unique among invertebrate lectins thus far reported in the literature and likely arose as a result of early gene duplication. The cDNA encoding the cockroach LPS binding protein reported by Jomori (1991) is 2271 bp long, but encodes a single CRD. Roughly 1500 bp of this sequence is at the 3' end following the stop codon. Kawasaki (1996) reports five cDNAs encoding C-type lectin domains similar to those seen in the grasshopper. Partially complete cDNAs range from 901-2881 bp in length, but again sequence representing only a single C-type CRD is present. Drosophila melanogaster also encodes a single CRD within a 774 bp cDNA (Haq, 1996). Similarly, the C-type lectin cDNA (Takahashi, 1985) from the fleshfly is 976 bp, but encodes only one CRD. Sequence identity between cDNA representing each grasshopper lectin and that from cockroach, fleshfly and Drosophila is $\leq 25\%$. However, in each sequence, codons for the eight strictly conserved residues defining the C-type CRD are present.

Amino Acid Sequences for Clones 3 and 4

Clone 3 Amino Acid Sequence

The amino acid sequences from clone 3 (Figure 16 p 56) and clone 4 (Figure 17 p 57) are 80% identical between sequence defining each ORF. Alignment of these sequences

is shown in Figure 18 (p 58). The 304 amino acid sequence for the mature clone 3 protein has a calculated molecular weight of 34056 Da. This weight is consistent with that reported by Stebbins (1985) via SDS-PAGE and Wenzlick (1996) for a laboratory purified lectin from this grasshopper. Amino acid sequences representing two distinct C-type carbohydrate recognition domains are evident in this protein. Cys residues at positions #71, #143, #157 and #165 are conserved in CRD1. Based on sequence alignment with MBP and ESEL, these residues are likely disulfide bridged to form the C-type architecture common to this protein family. All strictly conserved positions are colored gray in Figure 33. It is proposed that Cys71 and Cys165 are bridged to form the large outer loop of the motif, while Cys143 and Cys157 form the small loop. Each loop is evident in the crystal structures for MBP, ESEL and the GHA model CRD (Figure 22 p 65). This conserved sequence pattern is repeated in the second CRD where it is proposed Cys203 and Cys300 are bridged to form the large loop while Cys278 and Cys 292 form the small loop in the motif. The large loop in CRD1 of this sequence is 95 residues while that in CRD2 is 98. The significance of this size difference is unknown. The small loops are each 15 residues. In each CRD, four additional invariant residues are present. Gly102, Trp125, Pro130 and Asp150 are conserved in CRD1 and equivalent residues Gly232, Trp258, Pro260 and Asp285 in CRD2. Nine potential Ca^{2+} ligands in each GHA CRD are also conserved or conservatively substituted relative to the known ligands in MBP and ESEL (Figure 23 p 56). These positions are boxed in Figure 33. The exception is position 288 which is Asp and Asn in MBP and ESEL, but Gly in CRD2 of clone 3. This position is a proposed Ca^{2+} ligand and therefore requires an available -C=O or $-NH_2$ for donation of electron pairs. It may be that Gly serves as a ligand by donating a backbone -C=O. This position is occupied by His and Ile in the barnacle (BRA2) and the echinoidin (ECH) shown in Figure

33. Based on alignment with the other sequences in this family, these residues are unusual but do not detract from the C-type classification of this CRD.

Clone 4 Amino Acid Sequence

The deduced amino acid sequence for clone 4 is similar in size and content to that of clone 3. The 305 residue protein has a calculated molecular weight of 34401 Da. This mass is also consistent with that reported by Stebbins (1985) and Wenzlick (1996) for a grasshopper lectin isolated and purified in this laboratory. Conserved amino acid sequences representing two distinct C-type carbohydrate recognition domains are present in the deduced clone 4 sequence. Cys residues at positions #69, #144, #158 and #165 are conserved in CRD1 and are likely disulfide bridged to form the archetypal C-type motif (Figure 33). It is proposed that Cys69 and Cys166 are bridged to form the large outer loop, while Cys144 and Cys158 form the small loop. This pattern of conserved Cys residues is repeated again in the second CRD (Figure 33) where it is proposed Cys204 and Cys301 are bridged to form the large loop while Cys279 and Cys 293 form the small loop in the motif. The large loop in each of the clone 4 CRDs is 98 residues while the small loops are 15. Gly107, Trp130, Pro135 and Asp155 in CRD1 equivalent residues Gly237, Trp263, Pro265 and Asp290 in CRD2 are also strictly conserved in this family of C-type CRDs. Nine potential Ca^{2+} ligands in each CRD are also conserved or conservatively substituted relative to the known ligands in MBP and ESEL. These positions are boxed in Figure 33.

Comparison of C-type CRDs

Figure 33 illustrates the alignment of sequences representing 12 known vertebrate and invertebrate C-type lectin CRDs, including the four reported here from grasshopper. Sequences extend between the two conserved Cys residues forming the large loop in the C-type motif, based on the known structures for MBP and ESEL. The size of the large and small loops vary slightly between species, but sizes throughout are generally consistent. The size of the large loops range from 80 residues in the echinoderm (Giga, 1987) to 105 residues in the cockroach (Jomori, 1991) and fleshfly (Takahashi, 1985). Small loops extend from 15 residues in the cockroach, fleshfly and grasshopper proteins to 20 residues in ESEL (Graves, 1994). Each sequence maintains eight strictly conserved positions which help conserve the typical C-type architecture as seen in MBP and ESEL. The four Cys residues forming the large and small loops are conserved throughout. These residues are shown colored gray in Figure 33. Four additional positions are strictly conserved. Gly, Trp, Pro and Asp are colored gray and/or boxed in Figure 33. The positions of Ca^{2+} ligands in MBP and ESEL are boxed. These positions are less well conserved throughout this family suggesting some members may bind Ca^{2+} differently using different sidechain ligands. This is not unreasonable since MBP and ESEL employ different ligands to bind Ca^{2+} differently at one position. It is concluded that each of the four CRDs encoded by cDNA clones 3 and 4 represent authentic C-type lectins.

Signal Peptides for Clones 3 and 4

Proteins destined for export are generally synthesized with a 20-40 residue N-terminal extension (Briggs, 1985). This is the signal peptide and as such mediates export

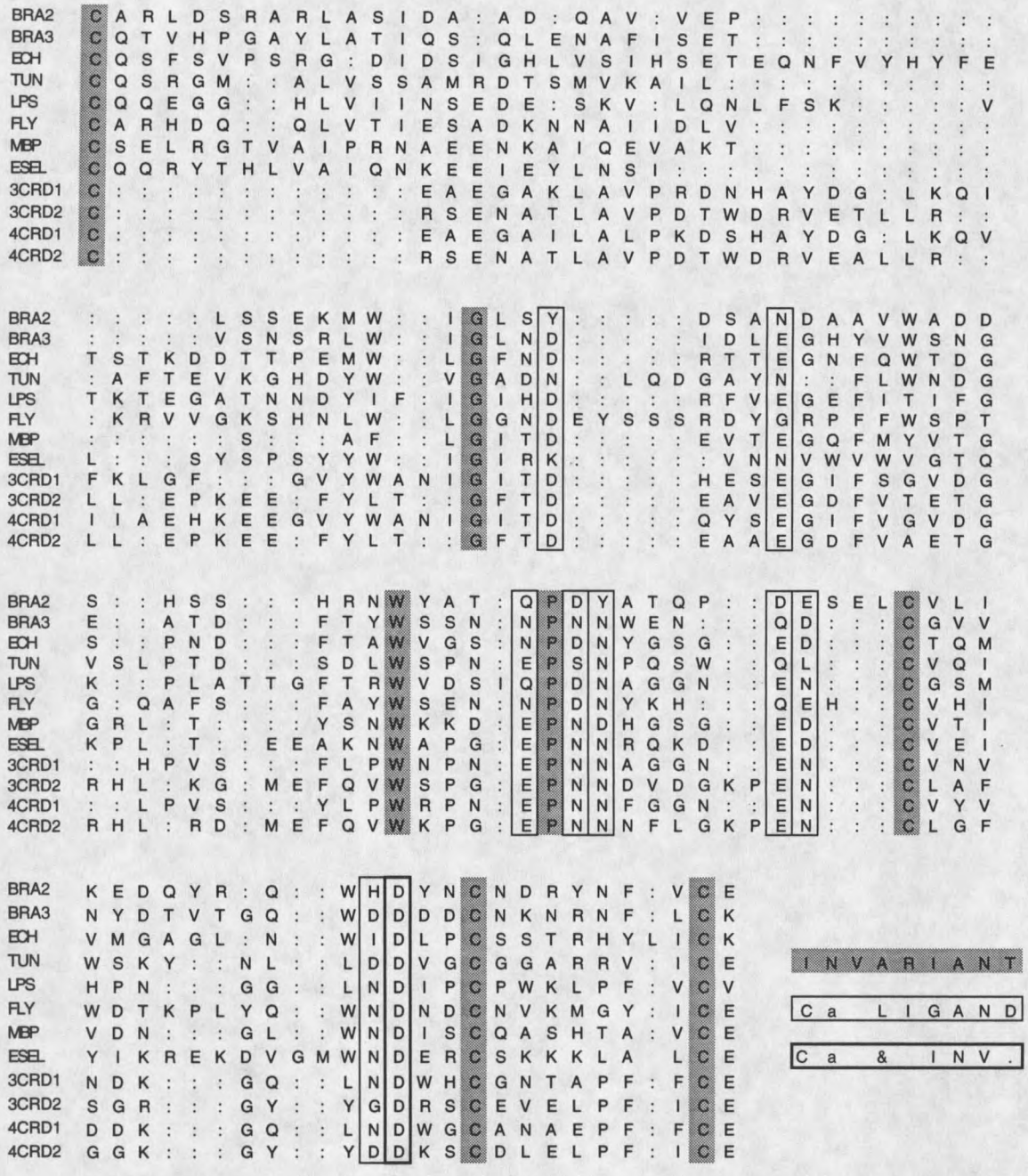


Figure 33. Alignment of amino acid sequences representing 12 invertebrate and vertebrate C-type lectin domains. GRAY = conserved, BOXED=Ca^{2+} ligands in MBP, HEAVY BOXED=conserved and a Ca^{2+} ligand in MBP. BRA2 & 3=Barnacle, ECH=echinoidin, TUN=tunicate, LPS=cockroach, MBP=rat mannose binding protein, ESEL=human E - selectin, 3CRD1 & 2=CRD1 and CRD2 from the clone 3 sequence, 4CRD1 & 2 =carbohydrate recognition domain from the clone 4 sequence.

of the protein across membranes (von Heijne, 1984). The format of a signal sequence has been defined by Watson (1984) based on the analysis of 277 known signals. These sequences; 1) are 20-40 residues in length and usually contain a charged residue among the first five amino acids, but not exclusively, 2) contain a core of hydrophobic residues (up to 9) followed by 3) a 'helix-breaking' residue commonly Gly, Pro or a large polar amino acid like Gln, which frequently occurs 4-8 residues prior to the cleavage site. The pattern of amino acids near the cleavage site has been defined (von Heijne, 1984) for 78 eukaryotic proteins. This pattern has been summarized as the '-3, -1' rule (cleavage occurs between -1 and +1, where +1 is the N-terminal residue in the protein). The '-1' position must be occupied by Ala, Ser, Gly, Cys, Thr or Gln. Phe, His, Tyr or Trp cannot occupy this position. Asp, Glu, Lys, Arg, Asn or Gln usually reside in position '-3'. Pro is not allowed in positions '-3' to '+1', but is common at position '-5'.

The 20 residue signal from the clone 3 sequence adheres to the format discussed above. The clone 3 signal sequence is underlined in Figure 16 (p 56). The first residue in the protein sequence is Ala (MQLVTVCAALVAATVPCTLAA). This residue is '+1' position and underlined. At the '-1' position is Ala which is allowed by the '-3, -1' rule. The '-2' position is occupied by Leu. The only stipulation put on this position is that Pro is not allowed there. The '-3' position is occupied by Thr, which is unusual for this position. However, in a study of 90 signal peptides, von Heijne (1983) reports eight sequences in which Thr occupies the '-3' position, indicating this residue is not exclusively prohibited. The '-5' position in the putative clone 3 signal is occupied by Pro, a common residue for this location. The clone 3 signal peptide contains a core sequence of eight hydrophobic residues (M--VCAALVAA--), but is missing a charged sidechain within the first 5 residues of the sequence. Here again, of 90 known signal sequences reported by von Heijne (1983)

only 26 contained a charged residue prior to the hydrophobic core. This suggests that while a charged residue is common in this region, its absence does not preclude the sequence from functioning as the signal. There are no alternative start sites in this sequence. Additionally, N-terminal amino acid sequence suggests the designated N-terminal Ala is accurate and the molecular weight reported for the deduced amino acid sequence beginning with this Ala is consistent with that measured for grasshopper lectin isolated and purified in this laboratory (Wenzlick, 1996). This evidence combined, suggests the designation of the signal sequence in clone 3 is rational.

The clone 4 signal is 21 residues and is shown underlined in Figure 17 (p 57). The '+1' position representing the N-terminal amino acid is Ala and is also underlined (MACP LIIIFETRPTVLMVGTGA). The '-1' position in this sequence is occupied by Gly and so adheres to the '-3,-1' rule as defined above. Thr occupies the '-2' position. This also meets the criteria outlined by von Heijne (1983) since only Pro is prohibited from occupying '-2'. The '-3' position is occupied by Gly. This residue is very unusual at this position, but not prohibited (von Heijne, 1983). Pro occupies '-9' and most likely represents the 'helix breaking' residue common to many signal sequences (Watson, 1984). A nine residue hydrophobic core (MACPLIIIF--) is present in the sequence, but as discussed for the clone 3 signal peptide, no charged residue is present. Glu at position '-12' immediately following the hydrophobic core is peculiar. Polar residues are not common in the C-terminal side of the hydrophobic core region unless at position '-3', but exceptions exist. The signal sequence for the LPS binding protein from cockroach (Jomori, 1991) also has a single Glu at position '-13' immediately following the hydrophobic core. As seen with the clone 3 signal sequence, there are no alternative start sites present in the designated clone 4 signal. The calculated molecular weight of the

mature protein beginning at Ala is also consistent with that measured for a purified lectin from this grasshopper. Considering these data, it is concluded that the proposed signal peptides for both clones 3 and 4 are rational.

Use of Unique Protocols

PCR Amplification Using a "Hotstart"

Hotstart PCR was used throughout work described in this thesis. This technique prevents the premature extension of randomly annealed primers at room temperature. Manual hotstarts were first employed, but two strategies 1) hotwax™ beads (Clontech™, Palo Alto CA) and 2) Amplitaq Gold™ (Perkin Elmer Corp., Applied Biosystems Division) were used with greater ease and success. Manual hotstarts are done by setting up the reactions on ice, thus preventing premature extension by Taq polymerase. The hotwax™ bead allows the combination of all reactants at room temperature, but without the Mg^{2+} cofactor required by Taq. This cofactor is supplied in the wax bead added to the tube immediately prior to cycling. Denaturing temperatures melt the wax and liberate Mg^{2+} to activate the enzyme. The molten wax serves as a vapor barrier during cycling. The use of Amplitaq Gold™ has proven to be the most efficient and cost effective. This Taq is a proprietary formulation which is inactive before prolonged heating at 95° C. Reaction components can be combined at room temperature and overlayed with mineral oil. The reaction is heated for 10 minutes at 95° C to activate the enzyme before cycling. This is the recommended hotstart strategy for future work in this laboratory.

Isolating the RT•PCR Product from Agarose

PCR products were visualized using agarose gel electrophoresis with ethidium bromide staining. To remove DNA from agarose gel for restriction analysis or cloning, Prep-a-gene™ (BioRAD, Hercules CA) was employed. This strategy is convenient and requires ≈ 20 minutes to recover DNA. Prep-a-gene™ is highly recommended for future work in this laboratory.

Use of pGEM-T Vector

The commercial pGEM-T plasmid vector was used to clone the 5' RACE products discussed in this thesis. This vector uses a modified 'sticky-end' approach and allows for the cloning of PCR products directly. The plasmid is a pGEM5fz(+) vector opened within the polylinker and blunt ended. Overhanging thymines (T) are enzymatically added to each side. This strategy takes advantage of the non-template directed addition of a single adenine (A) by Taq polymerase during PCR. Overhanging 'As' on the product base-pair with overhanging Ts in the vector providing for a sticky end ligation of PCR products directly into the vector. Restriction sites in newly amplified DNA are often unknown prior to sequence analysis. Use of a T-vector allows PCR products to be directly cloned without further enzymatic modification or purification. This makes ligation into this vector easy and fast. This cloning procedure is recommended for future work in this laboratory.

Model of GHA CRD

A 3D structure for the C-terminal carbohydrate recognition domain from the clone 3 amino acid sequence was predicted using sequence homology with two reference proteins, MBP and ESEL (Figure 22 p 65). The GHA model is qualitatively similar to the structure

seen for both MBP and ESEL. Approximately 50% of the structure is represented by α helices and β sheets. The remaining structure is very non regular, and best described as extended loops. The orientation of the large and small loops forming the C-type motif is also similar to that seen in each reference (Figure 22 p 65). Potential Ca^{2+} ligands are oriented at the surface of the domain (Figure 24 p 67) and occupy space suggesting the GHA domain could ligate Ca^{2+} as seen in MBP. Figure 25 (p 69) illustrates the overall similarity of GHA to the MBP and ESEL references. Quantitatively, the overall GHA structure is as similar to MBP and ESEL as these two references are to one another. This suggests the GHA model is a rational 3D structure for the sequence defining the C-terminal CRD from clone 3.

Reference Structures MBP and ESEL

MBP and ESEL have been used as references to predict 3D structures for other C-type lectin domains. A structure for Fcε/CD23 was first predicted by Padlan (1993) using MBP. A second structure for this domain was later generated using both MBP and ESEL (Bajorath, 1996), concluding that use of both references helped to more accurately predict the model structure in regions of atypical secondary structures. The type II antifreeze protein in fish was also modeled using the MBP crystal structure (Sonnichsen, 1995), as well as the T-cell activation antigen CD69 (Bajorath, 1994). Homology based modeling to predict each of these structures were used to predict and assess the structure for the GHA model.

Conservation of Disulfide Geometry

The C-type CRD in MBP and ESEL contains two disulfide bridges. The GHA model is predicted to have equivalent disulfide bridges between residues Cys278 and Cys292 proposed to form the small loop in the motif, and Cys203 and Cys300 forming the large loop. The side chain conformation of these four Cys residues were modeled specifically to provide good proximity for disulfide bridging (Bajorath, 1994). This was accomplished by constraining the movement of these residues during refinement (Castonguay, 1995). This restraint serves to maintain the integrity of the angle for each dihedral (Sonnichsen, 1995). The disulfide dihedral angles in GHA are 101° between Cys278 and Cys292 and -86° for that between Cys203 and Cys300. These angles are consistent with those seen in both MBP (96.7° & -79.70°) and ESEL(-84° & -89°) suggesting the constrained refinement used to construct the model and preserve the two disulfide linkages forming the C-type motif was effective.

ϕ–ψ and Peptide Bond Geometries in the GHA Model

Ramachandran plots were used to assess ϕ–ψ geometry in the final model. The main chain torsion angles ϕ and ψ are key variables for describing protein conformations (Karplus, 1996) and the initial assessment of predicted models by analysis of ϕ and ψ is essential to judging the overall quality of the final model structure. The model CRD exhibits 99% of all non-Gly residues within allowed regions of the plot as shown in Figure 27 p 71). Of these residues, 84% fall within the 'core' regions as defined by Morris (1992), while 15% are within 'generously allowed' regions. This distribution is consistent with the predicted structure of the type II antifreeze protein. Here, 97% of non-Gly residues fall within 'allowed' regions of the plot, 84% within core regions (Sonnichsen,

1995). In the model structure for CD69, 100% of the non-Gly residues lie within allowed regions of the plot (Bajorath, 1994). These distributions suggest the ϕ–ψ geometries in the predicted structure for a GHA CRD are reasonable and the quality of the GHA model is comparable with those in the literature. Based on these comparisons, it may be concluded the GHA model is a reasonable 3D structure for this sequence.

Peptide bond dihedral angles for the GHA model and the two reference structures MBP and ESEL are summarized in Appendix E. Within the GHA model, 98% of all peptide bond dihedral angles fall within 25° of ±180°. This distribution is reasonable when compared with the those for MBP (99%) and ESEL (99%) and confirms that no unreasonable dihedrals were introduced into GHA during the model building procedure. The major exceptions in each structure are equivalent cis Pro at positions 267, 186 and 81 in GHA, MBP and ESEL respectively. These cis peptide bond dihedrals are similar for GHA (7.1°), MBP (5.1°) and ESEL (3.8°), here again suggesting the GHA CRD was modeled appropriately in the region of this Pro residue. Models for CD23 (Padlan, 1993 & Bajorath, 1996) and CD69 (Bajorath, 1994) do not contain equivalent Pro residues at this position. The range of peptide bond dihedral angles in the GHA model, when compared to those of MBP and ESEL, again suggests the 3D structure is a reasonable representation of the GHA CRD sequence.

Comparison of 3D Profiles

The 3D profile for the GHA model was compared to that for MBP (Figure 31 p 77) and ESEL (Figure 32 p 78) to assess the compatibility of the GHA sequence with the structure it was folded into. The comparison is based on the assumption that profiles for MBP and ESEL crystal structures represent properly folded sequences. Profiles for GHA,

MBP and ESEL are generally similar and represent comparable 3D probability scores. This general similarity suggests the GHA profile would be typical for any homologous sequences folded in this manner and suggests the overall model CRD is folded correctly. Further, the lack of low scoring regions in the plot relative to plots for MBP and ESEL suggests there are no significant errors in the overall structure. The GHA profile is consistent with those for the models of CD69 (Bajorath, 1994) constructed using MBP as a reference, and CD23 (Bajorath, 1996) constructed using both MBP and ESEL as reference proteins. This suggests the GHA model is an equally good 3D representation of the sequence defining the grasshopper CRD.

Recommended Future Work

In the immediate future of this project, the following priorities should take precedence.

1) The complete sequence for the ≈ 1100 bp 5' RACE product reported in Figure 12 (p 50) should be completed. Approximately 400 bp of this cDNA was sequenced to obtain the complete 5' end of the original clone 4, leaving ≈ 600 bp of unknown sequence. Similarly, the 5' RACE procedure was completed using a clone 3-specific primer generating a ≈ 1500 bp fragment of which ≈ 1100 bp represents new 5' non-translated sequence.

2) RT•PCR should be used to confirm that fragments used to complete clones 3 and 4 belong to one another. For clone 3, this can be accomplished by using a clone 3-specific primer in the 5' non-translated region of the sequence to amplify the entire coding region. Sequence analysis of this fragment would serve to verify that the

two cDNAs, discussed in this thesis, and used to obtain the complete clone 3 ORF belong together. This strategy should also be applied to the clone 4 sequence using a gene-specific primer in the 5' non-translated region.

3) Amino acid sequence for CNBr fragments from the protein(s) isolated in the laboratory should be repeated. If enough sequence data of sufficient quality can be obtained, this may provide the basis for the design of gene-specific primers of potential use in a RT•PCR reaction to isolate cDNA representing the protein isolated in the lab.

4) The 450 bp PCR product obtained using the clone 3-specific primer for screening the remaining 14 positive clones (Figure 6 p 40, lane 14) should be isolated, cloned and sequenced. This fragment may represent a portion of the protein isolated in the laboratory. Cloning and sequencing this fragment is straightforward and cost-effective.

5) Finally, laboratory personnel should collaborate to express all or a portion of clones 3 or 4. Expression of a CRD may allow for the analysis of its CHO specificity. Expression of each domain separately may help determine if each has similar or different specificities e. g. galactose or glucose.

Summary of Results

The cDNA sequences from the original 879 bp clone 3 and the 1079 bp clone 4 represented incomplete portions of two homologous sequences. 3' RACE was employed to isolate and clone a cDNA containing the complete 3' sequence for clone 3. This new

sequence was combined with the original 879 bp clone 3 sequence to complete a cDNA representing a full-length ORF for a grasshopper lectin. 5' RACE was used to isolate and amplify a cDNA containing the complete 5' sequence for clone 4. Similarly, this new sequence was combined with that known for the original 1079 bp clone 4 to complete a cDNA sequence representative of a full-length ORF for second lectin.

A computer-based 3D structure for the deduced amino acid sequence representing a the C-terminal CRD from the clone 3 lectin was predicted by homology using the crystal coordinates for rat MBP and human ESEL. Assessment of this model indicates the final structure is a reasonable 3D representation of this sequence suggesting the C-type CRD from grasshopper may occupy a fold similar to that seen in rat MBP and human ESEL.

Conclusions

The work discussed in this thesis represents the successful completion of three goals to extend the molecular characterization of lectins from the grasshopper. Specific accomplishments are:

1) The sequence for a complete clone 3-specific cDNA representing a full-length ORF for a lectin from the grasshopper *Melanoplus differentialis* was completed. This cDNA is 1220 bp long and contains sequence representing the initiating Met, signal peptide, stop codon and the 3' non-translated region through the polyA tail. The deduced amino acid sequence contains 324 amino acids. The calculated molecular weight of the protein is 34056 Da.

2) The sequence for a complete clone 4-specific cDNA representing a full-length ORF for a second lectin from the grasshopper was completed. This cDNA is 1213 bp long and contains sequence representing the initiating Met, signal peptide, stop codon and the 3' non-translated region through the polyA tail. The deduced amino acid sequence is 325 amino acids. The calculated molecular weight of the protein is 34441 Da. Together, results confirm the existence of multiple lectin proteins in the grasshopper.

3) A computer-based 3D model for a grasshopper CRD was completed by homology modeling using the crystal structure coordinates for MBP and ESEL. Based on the assessment of the model using Ramachandran plots and 3D Profiles, this model is a reasonable 3D representation of the primary sequence.

REFERENCES

Ando, K., Okada, M. & Natori, S. (1983). Purification of sarcotoxin II, antibacterial proteins of *Sarcophaga peregrina* (flesh fly) larvae. Biochemistry, 26, 226-230.

Bajorath, J., Stenkamp, R. & Arufo, A., (1993). Knowledge-based model building of proteins: Concepts and examples. Protein Science, 2, 1798-1810.

Bajorath, J. & Aruffo, A. (1994). Molecular model of the extracellular lectin-like domain in CD69. Journal of Biological Chemistry, 269, 32457-32463.

Bajorath, J. & Aruffo, A. (1996). Structure-based modeling of the ligand binding domain of the human cell surface receptor CD23 and comparison of two independently derived molecular models. Protein Science, 5, 240-247.

Barondes, S. H., Cooper, D. N. W., Gitt, M. A. & Leffler, H. (1994). Galectins. Journal of Biological Chemistry, 269, 20807-208010.

BioRad Corporation, 1000 Alfred Nobel Drive, Hercules CA.

Blundell, T. L. & Johnson, M. S. (1993). Catching a common fold. Protein Science, 2, 877-883.

Boehringer Mannheim Inc., 700 Massechusetts Avenue, Cambridge MA.

Boman, H. G., Faye, I., van Hofsten, P., Kockum, K., Lee, J. Y., Xanthopoulos, K. G., Bennich, H., Engstrom, A., Merrifield, B. R. & Andreu, D. (1986). Antibacterial Immune Proteins in Insects: A Review of Some Current Perspectives. M. Grehelin (Ed). Immunity in Invertebrates. New York: Springer-Verlag.

Boman, H. G. & Hultmark, D. (1987). Cell-free immunity in insects. Annual Review of Microbiology, 41, 103-126.

Boucias, D. G. & Pendland, J. C. (1991). The Fungal Cell Wall and its Involvement in the Pathogenic Process in Insect Hosts. J. P. Latge & D. G. Boucias (Eds). Fungal Cell Wall and Immune Response. Heidelberg: Springer-Verlag, 303-316.

Boucias, D. G. & Pendland, J. C. (1993). The galactose binding lectin from the Beet Army Worm, *Spodoptera exigua*: Distribution and site of synthesis. Insect Biochemistry and Molecular Biology, 23, 233-242.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253, 164-170.

Bradley, R. S., Stuart, G. S., Stiles, B., & Hapner, K. D. (1989). Grasshopper haemagglutinin: Immunochemical localization in haemocytes and investigation of opsonic properties. Journal of Insect Physiology, 35, 353-361.

Briggs, M. S., Gierasch, L. M. Zlotnick, A., Lear, J. D. & DeGrado, W. F. (1985). In vivo function and membrane binding properties are correlated for Escherichia coli LamB signal peptides. Science, 228, 1096-1099.

Castonguay, L. A., Bryant, S. H., Snow, P. M. & Fetrow, J. S. (1995). A proposed structural model of domain 1 fasciclin III neural cell adhesion protein based on an inverse folding algorithm. Protein Science, 4, 472-483.

Chen, C., Ratcliffe, N. A. & Rowley, A. F. (1993). Detection, isolation and characterization of multiple lectins from the haemolymph of the cockroach Blaberus discoidalis. Biochemistry Journal, 294, 181-190.

CLONTECH Laboratories Inc., 4030 Fabian Way, Palo Alto CA 94303-4607.

Dunn, P. E. (1986). Biochemical aspects of insect immunology. Annual Review of Entomology, 31, 321-339.

Dunn, P. E. (1990). Humoral immunity in insects. BioScience, 40, 738-743.

Dupont Company Biotechnology Systems, Barley Moll Plaza P-24, Wilmington DE 19898.

Drickamer, K. (1988). Two distinct classes of carbohydrate-recognition domains in animals lectins. Journal of Biological Chemistry, 263, 9557-9560.

Drickamer, K. (1992). Engineering galactose-binding activity into a C-type mannose-binding protein. Nature, 360, 183-186.

Drickamer, K. (1993). Biology of animal lectins. Annual Review of Cell Biology, 9, 273-264.

Drickamer, K. (1994). Evolution of Ca^{2+} dependent animal lectins. Progress in Nucleic Acid Research and Molecular Biology, 45, 207-233.

Eastman Kodak Company, 343 State Street, Rochester NY 14650-0207.

Epicentre Technologies Corporation, 1202 Ann Street, Madison WI 53713.

Erbe, D. V., Wolitzky, B. A., Presta, L. G., Norton, C. R., Ramos, R. J., Burns, D. K., Rumberger, J. M., Rao, N. B. N., Foxall, C., Brandley, B. K. & lasky, L. A. (1992). Identification of an E-selectin region critical for carbohydrate recognition and cell adhesion. Journal of Cell Biology, 119, 215-227.

Frohman, M. A., Dush, M. K. & Martin, G. R. (1989). Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. Procedings of the National Academy of Sciences, 85, 8998-9002.

Gaveriaux, C., & Loor, F. (1987). An enzyme-linked lectin-binding assay on cells (CELLBA) for the comparison of lectin receptor expression on cell surfaces. Journal of Immunological Methods, 104, 173.

Gelman Sciences, 600 Sth Wagner Rd, Ann Arbor MI 48601.

Geng, J. G., Heavner, G. A. & McEver, R. P. (1992). Lectin domain peptides from selectins interact with both cell surface ligands and Ca^{2+} ions. Journal of Biological Chemistry, 267, 19846-19853.

Gedik, T. (1996). Grasshopper Lectin Genes: Southern Analysis and Polymerase Chain Reaction. Masters Thesis. Montana State University, Bozeman MT.

Giga, Y., Ikai, A. & Takahashi, K. (1987). The complete amino acid sequence of Echinoidin, a lectin from the coelomic fluid of the sea urchin Anthocidaris crassispina. Journal of Biological Chemistry, 262, 6197-6203.

Godzik, A., Kolinski, A. & Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. Journal of Molecular Biology, 227, 227-238.

Graves, B. J., Growther, R. L., Chandran, C., Rumberger, J. M., Li, S., Huang, K. S., Presky, D. H., Familletti, P. C., Wolitzky, B. A. & Burns, D. K. (1994). Insight into E-selectin/ligand interaction from the crystal structure and mutagenesis of the lec/EFG domains. Nature, 367, 532-538.

Gupta, A. P. (1986). Hemocytic and Humoral Immuniy in Arthropods (1st ed). New York: John Wiley & Sons.

Gupta, A. P. (1992). Immunology of Insects and Other Arthropods. New York: John Wiley & Sons.

Hagen, H. E., Grunewald, J. & Ham, P. J. (1994). Induction of the prophenoloxidase-activating system of *Simulium* (Diptera: Simuliidae) following *Onchorcerca* (Nematoda: filarioidae) infections. Parasitology, 109, 649-655.

Haq, S., Kubo, T., Kurata, S., Kobayashi, A. & Natori, S. (1996). Purification, characterization, and cDNA cloning of a galactose-specific C-type lectin from Drosophila melanogaster. Journal of Biological Chemistry, 271, 20213-20218.

Hajek, A. E. & St Leger, R. J. (1994). Interactions between fungal pathogens and insect hosts. Annual Review of Entomology, 39, 293-322.

Hink, W. F. & Briggs, J. D. (1968). Bactericidal factors in hemolymph from normal and immune wax moth larvae, *Galleria mellonella*. Journal of Insect Physiology, 14, 1025-1034.

Hoffman, J. A. (1995). Innate immunity of insects. Current Opinion in Immunology, 7, 4-10.

Hoffman, J. A., Reichart, J. M. & Hetru, C. (1996). Innate immunity in higher insects. Current Opinion in Immunology, 8, 7-13.

Holmskov, U., Malhotra, R., Sim, R. & Jensenius, J. (1994). The innate immune system: Collectins; collagenous C-type lectins for the innate immune defense system. Immunology Today, 15, 67-74.

Hoppe, H. J. & Reid, K. B. M. (1994). Collectins-soluble proteins containing collagenous regions and lectin domains-and their roles in innate immunity. Protein Science, 3, 1143-1156.

Iobst, S. T., Wormald, M. R., Weis, W. I., Dwek, R. A. & Drickamer, K. (1994). Binding of sugar ligands to Ca^{2+}-dependent animal lectins; Analysis of mannose binding by site-directed mutagenesis and NMR. Journal of Biological Chemistry, 269, 1505-15511.

Iobst, S. T. & Drickamer, K. (1994). Binding of sugar ligands to Ca^{2+}-dependent animal lectins; Generation of high-affinity galactose binding by site-directed mutagenesis. Journal of Biological Chemistry, 269, 15512-15519.

Jomori, T., Kubo, T. & Natori, S. (1990). Purification and characterization of lipopolysaccharide-binding protein from hemolymph of the American Cockroach *Periplaneta americana*. European Journal of Biochemistry, 190, 201-206

Jomori, T. & Natori, S. (1991). Molecular cloning of cDNA for lipopolysaccharide-binding protein from the hemolymph of the American Cockroach, *Periplaneta americana*. The Journal of Biological Chemistry, 266, 13318-13323.

Kanost, M. R. (1993). Hemolin: an insect hemolymph protein from the immunogolbulin superfamily. Conference Proceedings: Society of Invertebrate Pathology, Asheville NC, 26-27.

Kanost, M. R., Zepp, M. K., Ladendorff, N. E. & Anderson, L. A. (1994). Isolation and characterization of a hemocyte aggregation inhibitor from hemolymph of *Manduca sexta* larvae. Archives of Insect Biochemistry and Physiology, 27, 123-136.

Karplus, P. A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Science, 5, 1406-1420.

Kawasaki, K., Kubo, T. & Natori, S. (1993). A novel role of Periplaneta lectin as an opsonin to recognize 2-keto-3-deoxy octanate residues of bacterial lipopoly-saccharides. Comparative Biochemistry and Physiology, 106B, 675-680.

Kawasaki, K., Kubo, T. & Natori, S. (1996). Presence of the Periplaneta lectin-related protein family in the American Cockroach *Periplaneta americana*. Insect Biochemistry and Molecular Biology, 26, 335-364.

Kennedy, J. F., Palva, P. M. G., Corella, M. T. S., Cavalcanti, M. S. M. & Coelho, L. C. B. B. (1995). Lectins, versatile proteins of recognition: A review. Carbohydrate Polymers, 26, 219-230.

Kobayashi, A., Hirai, H., Kubo, T., Ueno, K., Nakanishi, Y. & Natori, S. (1989). Cloning and in vitro transcription of the Sarcophaga lectin gene. Biochimica et Biophysica Acta, 1009, 244-250.

Komano, H., Mizuno, D., & Natori, S. (1980). Purification of lectin induced in the hemolymph of *Sarcophaga peregrina* larvae on injury. Journal of Biological Chemistry, 255, 2919-2924.

Komano, H., Nozawa, R., Mizuno, D., & Natori, S. (1983). Measurement of *Sarchphaga peregrina* lectin under various physiological conditions by radioimmunoassay. The Journal of Biological Chemistry, 258, 2143-2147.

Komano, H., & Natori, S. (1985). Participation of *Sarcophaga peregrina* humoral lectin in the lysis of sheep red blood cells injected into the abdominal cavity of larvae. Developmental Comparative Immunology, 9, 31-40.

Kubo, T., & S., N. (1987). Purification and some properties of a lectin from the hemolymph of *Periplaneta amercana* (American Cockroach). European Journal of Biochemistry, 168, 75-82.

Kubo, T. Kawasaki, K. & Natori, S. (1993). Transient appearance and localization of a 26 kDa lectin, a novel member of the Periplaneta lectin family, in regenerating cockroach leg. Developmental Biology, 156, 381-390.

Kuroki, Y. & Voelkers, D. R. (1994). Pulmonary surfactant proteins. Journal of Biological Chemistry, 269, 25943-25946.

Ladenjorff, N. E. & Kanost, M. R. (1991). Bacteria-induced protein P4 (Hemolin) from Manduca sexta: A member of the immunoglobulin superfamily which can inhibit hemocyte aggregation. Archives of Insect Biochemistry and Physiology, 18, 285-300.

Leonard, C., Ratcliffe, N. A. & Rowley, A. F. (1985). The role of prophenoloxidase activation in non-self recognition and phagocytosis by insect blood cells. Journal of Insect Physiology, 31, 789-799.

Lodish, H. F. (1991). recognition of complex oligosaccharides by the multi-subunit asialoglycoprotein receptor. Trends in Biological Science, 16, 374-377.

Lu, J., Thiel, S., Wiedemann, H., Timple, R. & Reid, K. B. M. (1990). Binding of the pentamer/hexamer forms of mannan binding protein to zymosan activates the proenzyme clr2clr3 complex, of the classical pathway of compliment, without involvement of c1q. Journal of Immunology, 144, 22287-22295.

Luthy, R., Bowie, J. U. & Eisenber, D. (1992). Assessment of protein models with three-dimensional profiles. Nature, 356, 83-85.

Marmaras, V. J., Bournazos, S. N., Katsoris, P. G. & Lambropoulou, M. (1993). Defense mechanisms in insects: Certain integumental proteins and tyrosinase are responsible for nonself-recognition and immobilization of *Eschericia coli* in the cuticle of developing *Ceratitis capitata*. Archives of Insect Biochemistry and Physiology, 23, 169-180.

Marmaras, V. J., Charalambidis, N. D. & Zervas, C. G. (1996). Immune response in insects: The role of phenoloxidase in defense reactions in relation to melanization sclerotization. Archives of Insect Biochemistry and Physiology, 31, 119-133.

Matsuyama, K. & Natori, S. (1988). Purification of three antibacterial proteins from culture medium of NIH-Sape 4, an embryonic cell line of *Sarcophaga peregrina*. Journal of Biological Chemistry, 260, 17112-17116.

Marschal, P., Herrmann, J., Leffler, H., Barondes, S. H., & Cooper, D. N. W. (1992). Sequence and specificity of a soluble lactose-binding lectin from *Xenopus laevis* skin. The Journal of Biological Chemistry, 267, 12942-12949.

Minnick, M. E., Rupp, R. A., & Spence, K. D. (1986). A bacterial-induced lectin which triggers hemocyte coagulation in Manduca sexta. Biochimica et Biophysica Resource Communication, 137, 729-736.

Molecular Research Center Inc., 5645 Montgomery Rd, Cinncinnati OH 45212.

Morishima, I., Horiba, T. & Yamano, Y. (1994). Lysozyme activity in immunized and non-immunized hemolymph during the development of the silworm, *Bombyx mori*. Comparative Biochemistry and Physiology, 108A, 311-314.

Morris, A. L., MacArthur, M. W.; Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of portein structure coordinates. Proteins; Structure, Function and Genetics, 12, 345-364.

Mullett, H., Ratcliffe, N. A. & Rowley, A. F. (1993). Analysis of immune defense of the Wax Moth, *Galleria mellonella*, with anti-haemocytic monoclonal antibodies. Journal of Insect Physiology, 39, 897-902.

Muramoto, K. & Kamiya, H. (1986). the amino-acid sequence of a lectin of the Acorn Barnacle *Megabalanus rosa*. Biochimica et Biophysica Acta, 874, 285-295.

Muramoto, K. & Kamiya, H. (1992). the amino-acid sequence of a lectin from conger eel, *Conger myriaster*, skin mucus. Biochimica et Biophysica Acta, 1116, 129-136.

Natori, S. (1987). Hemolymph proteins participating in the defense system of *Sarcophaga peregrina*. Molecular Entomology, J. H. Law (Ed). New York: UCLA Symposia on Molecular Biology, 369-378.

OLIGO™, National Biosciences, 725 Tower Drive, Hamel MN 55340.

Okada, M. & Natori, S. (1985). Primary structure of sarcotoxin I, an antibacterial protein induced in the hemolymph of *Sarcophaga peregrina* (flesh fly) larvae. Journal of Biological Chemistry, 260, 7174-7177.

Padlan, E. A. & Helm, B. A. (1993). Modeling of the lectin-homology domains of the human and murine low-affinity FcE receptor (FcERII/CD23). Receptor, 3, 325-341.

Pendland, J. C. & Boucias, D. G. (1986). Lectin binding characteristics of several entomogenous hyphomycetes: Possible relationship to insect hemagglutinins. Mycologia, 78, 818-824.

Pendland, J. C., Heath, M. A. & Boucias, D. G. (1988). Function of a galactose-binding lectin from *Spodoptera exigua* larval haemolymph: Opsonization of blastospores from entomogenous hyphomycetes. Journal of Insect Physiology, 34, 533-540.

Perkin-Elmer Applied Biosystems Division, 850 Lincoln Centre Drive, Foster City CA 94404.

Promega Corporation, 2800 Woods Hollow Road, Madison WI 53711-5399.

Ratcliffe, N. A. & Gagen, S. H. (1977). Studies on the *in vivo* cellular reactions of insects: An ultrastructural analysis of nodule formation in *Galleria mellonella.* Tissue Cell, 9, 73-85.

Ratcliffe, N. A., Leonard, C. & Rowley, A. F. (1984). Prophenoloxidase activation: non-self recognition and cell cooperation in insect immunity. Science, 226, 557-559.

Ratcliffe, N. A. (1985). Invertebrate immunity - A primer for the non-specialist. Immunology Letters, 10, 253-270.

Richman, A. & Kafatos, F. C. (1995). Immunity to eukaryotic parasites in vector insects. Current Opinion in Immunology, 8, 14-19.

Rini, J. M. (1995). Lectin structure. Annual Review of Biophysics and Biomolecular Structure, 24, 551-557.

Rognlie, M. C. (1991). Molecular cloning and sequencing of a cDNA from the grasshopper *Melanoplus differentialis*. Masters Thesis. Montana State University, Bozeman MT.

Sali, A. & Blundell, T. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology, 234, 779-815.

Sali, A. & Overington, J. P. (1994). Derivation of rules for comparative modeling from a database of protein structure alignments. Journal Protein Science, 3, 1582-1586.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Chris Nolan, Ed). Cold Spring Harbor Laboratory Press: Cold Spring Harbor.

Sanger, G., Niklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terting inhibitors. Procedures of the National Academy of Science USA, 74, 5463-5467.

Schmit, A. R. & Ratcliffe, N. A. (1977). The encapsulation of foreign tissue implants in *Galleria mellonella* larvae. Journal of Insect Physiology, 23, 175-184.

Sharon, N. (1993). Lectin-carbohydrate complexes of plants and animals: An atomic view. Trends in Biological Science, 18, 221-226.

Sonnichsen, F. D., Sykes, B. D. & Davies, P. L. (1995). Comparative modeling of the three-dimensional structure of Type II antifreeze protein. Protein Science, 4, 460-471.

Spence, K. D. & Kawata, M. Y. (1993). Permeability characteristics of the peritrophic membranes of Manduca sexta larvae. Journal of Insect Physiology, 39, 785-790.

Stebbins, M., & Hapner, K. (1985). Preparation and Properties of the Haemagglutinin from Haemolymph of Acrididae (Grasshoppers). Insect Biochemistry, 15, 451-462.

Steiner, H., Hultmark, D., Engstrom, A., Bennich, H. & Boman, H. G. (1981). Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature, 292, 246-248.

Stephens, J. M. (1962). Bactericidal activity of the blood of actively immunized wax moth larvae. Canadian Journal of Microbiology, 8, 491-499.

Stiles, B., Bradley, R., Stuart, G., & Hapner, K. (1988). Site of synthesis of the haemolymph agglutinin of *Melanoplus differentialis* (Acrididae: Orthoptera). Journal of Insect Physiology, 34, 1077-1085.

Sun, S. C., Lindstrom, I., Boman, H. G. & Schmidt, O. (1990). Hemolin: an insect immune protein belonging to the immunoglobulin superfamily. Science, 250, 1729-1732.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). Knowledge-based modeling of homologous proteins: I. Three-dimensional frameworks. Protein Engineering, 1, 385-392.

Takahashi, H., Komano, H., Kawaguchi, N., Kitamura, N., Nakanishi, S., & Natori, S. (1985). Cloning and sequencing of cDNA of *Sarcophaga peregrina* humoral lectin induced n injury of the body wall. Journal of Biological Chemistry, 260, 12228-12233.

Tsuboi, I., Matsukawa, M., Sato, N. & Kimura, S. (1993). Isolation and characterization of a sialic acid-specific binding lectin from the hemolymph of Asian horseshoe crab, *Tachypleus tridentatus*. Biochimica et Biophysica Acta, 1156, 255-262.

United States Biochemical, PO Box 22400, Cleveland OH 44122.

Vasta, G. & Marchalonis, J. J. (1985). Humoral and cell membrane-associated lectins from invertebrates and lower chordates: Specificity, molecular characterization and their structural relationships with putative recognition molecules from vertebrates. Developmental and Comparative Immunology, 9, 531-539.

Vasta, G. R. & Marchalonis, J. J. (1987). Invertebrate Agglutinins and Evolution of Humoral Cellular Recognition Factors. A. H. Greenberg, Ed. Invertebrate Models: Cell Receptors and Cell Communication. Basel: A. G. Karger.

Villoutreix, B. O., Getzoff, E. D. & Grifin, J. H. (1994). A structural model for the prostate disease marker, human prostate-specific antigen. Protein Science, 3, 2033-2044.

von Heijne, G. (1984). How signal sequences maintain cleavage specificity. European Journal of Biochemistry, 173, 243-251.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. Nucleic Acids Research, 14, 4683-4690.

Weber, I. T., Miller, M., Jaskolski, M., Skalka, A. M. & Wlodawer, A. (1989). Molecular modeling of the HIV-1 protease and its substrate binding site. Science, 243, 928-931.

Weber, I. T. (1990). Evaluation of homology modeling of HIV protease. Proteins Structure & Function of Genetics, 7, 172-184.

Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. & Hendrickson, W. A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. Science, 254, 1608-1615.

Weis, W. I., Drickamer, D. & Hendrickson, W. A. (1992). Structure of a C-type mannose-binding protein complexed with an oligosaccharide. Nature, 360, 127-134.

Weis, W. I. & Drickamer, K. (1996). Structural basis of lectin-carbohydrate recognition. Annual Review of Biochemistry, 65, 441-473.

Wenzlick, D. (1996). Grasshopper Agglutinin: Preparation and Characterization by MALDI/TOF-MS. Masters Thesis. Montana State University, Bozeman MT.

Whatmann International Limited, Maidstone England, UK.

Wheeler, M. B., Stuart, G. S. & Hapner, K. D. (1993). Agglutinin mediated opsonization of fungal blastospores in *Melanoplus differentialis* (Insecta). Journal of Insect Physiology, 39, 477-483.

Whelan, J. (1996). Selectin synthesis and inflammation. Trends in Biological Science, 21, 65-69.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L. Schneider, J. & Kent, S. (1989). Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. Science, 245, 616-621.

Zachary, D. & Hoffman, J. A. (1974). Lysozyme is stored in the granules of certain haemocyte types of Locusta. Journal of Insect Physiology, 30, 405-413.

APPENDICES

NOTE: The following list defines abbreviations at the top of each column in Appendices B, C, D and E.

#	= position of amino acid residue in the polypeptide sequence.
R	= amino acid residue.
Ab	= area of the amino acid sidechain that is buried.
Fp	= fraction of the amino acid sidechain that is covered by polar atoms.
Score	= 3D probability for individual residue.
Acc	= accumulated 3D probability score.
Avg	= average 3D probability score for a 21 residue window (10 residues on each side)
ϕ	= ϕ angle used to plot ϕ vs ψ in a Ramachandran plot.
ψ	= ψ angle used to plot ψ vs ϕ in a Ramachandran plot.
Amide	= peptide bond dihedral angle.
$\chi 1$	= side chain angle

Data presented inside a box is discussed in the Results and or Discussion section of this thesis.

APPENDIX A

Sequencing Primer Summary

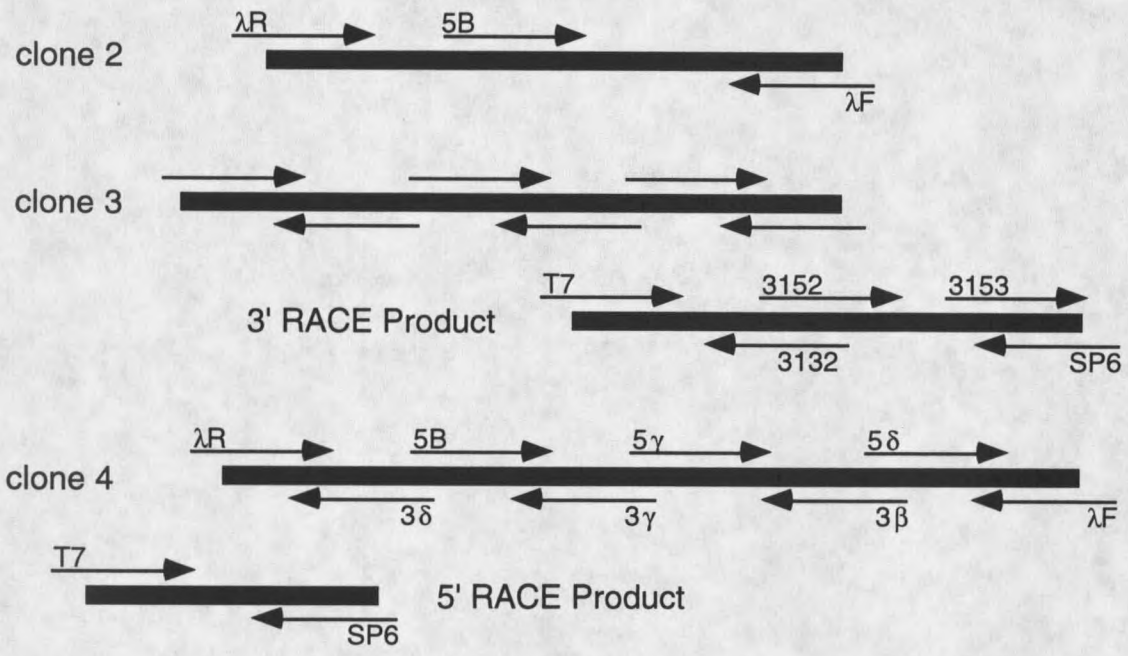


Illustration of clones 2, 3 and 4 including the cloned products from 3' and 5' RACE showing the primers used for sequence analysis and their relative location to one another. λF and λR anneal to the λgt11 cloning vector flanking the insert. SP6 and T7 primers anneal to vector sequence flanking the insert.

APPENDIX B

GHA Model Assessment Data

#	R	Ab	Fp	Env	Acc	Scor	Avg21	ϕ	ψ	Amide	$\chi 1$
191	A	3.6	0.94	E	0.14	0.14	0.28	-59	-173.9		
192	E	69.2	0.65	P2	0.42	0.28	0.28	-74.5	114.3	-167.8	-174.2
193	K	100.5	0.74	P2	0.89	0.47	0.28	-92.4	84.6	178.1	-166.6
194	A	53.5	0.55	P1	1.38	0.49	0.28	-142.4	-161.7	-170.5	
195	V	89.3	0.73	P2	0.64	-0.74	0.28	-87.9	154.6	-178.6	64
196	Y	163.3	0.45	B2	1.78	1.14	0.28	-60.2	-33.5	166.1	-161.2
197	A	21.9	0.91	E	1.92	0.14	0.28	-61.3	-33.6	161.4	
198	E	56.7	0.64	P2	2.2	0.28	0.28	-70.1	-36.2	-176.1	-66.1
199	A	71	0.11	P1	2.69	0.49	0.28	-53.4	-59.3	172.9	
200	A	53.5	0.68	P2	2.44	-0.25	0.28	-57.8	-44.9	169.7	
201	R	53.5	0.94	P2	2.68	0.24	0.28	-53.6	-48.4	173.5	-170.6
202	V	88.6	0.6	P2	1.94	-0.74	0.25	-71.6	-35	170.1	-157.9
203	C	56	0.45	P1	3.23	1.29	0.23	-65.9	-40.2	161.9	-69.6
204	R	69.5	0.81	P2	3.47	0.24	0.23	-52.8	-46.1	167.2	-70.1
205	S	15.9	0.87	E	3.81	0.34	0.21	-75.8	21.3	178	-174.2
206	E	83.2	0.62	P2	4.09	0.28	0.19	-139.5	-58.5	-171.3	-72.6
207	N	81.7	0.7	P2	4.6	0.51	0.16	-128.3	-68.5	-179.3	-66.8
208	A	24	0.72	E	4.74	0.14	0.18	-149	-68.8	-168.2	
209	T	76.1	0.55	P1	5.29	0.55	0.18	-161.7	-36.3	-142.3	-162.2
210	L	154	0.44	B2	6.06	0.77	0.19	156.6	113.2	-173.6	-46.8
211	A	59.1	0.67	P2	5.81	-0.25	0.19	-82.2	109.9	-175.3	
212	V	126.5	0.55	B3	5.41	-0.4	0.17	-96.9	124.7	165.1	-61.5
213	P	122.3	0.36	B2	5.34	-0.07	0.18	-78.2	103	-175.5	33
214	D	100.6	0.48	P1	5.68	0.34	0.15	-92.1	-42	-175.2	-63.8
215	T	65.4	0.64	P2	5.76	0.08	0.07	-140.6	154.2	176.2	67.2
216	W	96.6	0.71	P2	4.67	-1.09	0.09	-56.4	-39.3	170	-71.8
217	D	14.2	0.92	E	5.11	0.44	0.08	-67.1	-33.3	163.8	-71.4
218	R	102	0.8	P2	5.67	0.56	0.07	-63.8	-46.5	168.8	-66.9
219	V	116	0.38	B2	6.08	0.41	0.06	-55.9	-45.6	168.3	-56.4
220	E	75.7	0.87	P2	6.7	0.62	0.04	-51.3	-59.5	174.7	-178.1
221	T	21.8	0.84	E	6.5	-0.2	-0.02	-65.3	-32.1	164.9	-165.4
222	L	93.6	0.47	P1	6.2	-0.3	0.03	-62.3	-51	170.6	-70
223	L	105.6	0.6	P2	5.74	-0.46	0.1	-56.4	-43	169	-170.3
224	R	46.7	0.9	P2	6.3	0.56	0.15	-62.3	-34.9	169.9	-66.3
225	L	17.9	0.91	E	4.93	-1.37	0.16	-74	-35.3	-176	-172.9
226	L	121	0.35	B2	5.7	0.77	0.21	-78.3	150.5	163.4	-69.8
227	E	3.4	0.9	E	5.74	0.04	0.33	-73	139.5	-176.5	-68.9
228	P	36.7	0.82	E	5.99	0.25	0.34	-74.6	83.7	-170.9	29.8
229	K	139.5	0.59	B3	6.07	0.08	0.28	-92.2	117.7	-168	-161.1
230	E	17.4	0.9	E	6.11	0.04	0.28	-61.5	-35	170.2	62.8
231	E	116.7	0.49	B3	5.65	-0.46	0.26	-78.5	-7.3	167.2	72
232	F	160.2	0.69	B3	6.36	0.71	0.19	83.7	145.8	-171.2	-49.7
233	Y	189.4	0.42	B2	7.5	1.14	0.18	-95.2	128.1	-173.9	-61.3

APPENDIX B

GHA Model Assessment Data

234	L	153.3	0.28	B1	8.56	1.06	0.29	-66	128.1	-176.6	-72.7
235	T	111.6	0.32	P1	9.11	0.55	0.25	-103.9	164.2	164.6	67.1
236	G	40	0.53	E	10.21	1.1	0.37	-83.5	60.6	-168	
237	F	189	0.37	B2	11.53	1.32	0.34	-158.6	91.2	-157.2	-164
238	T	85.6	0.6	P2	12.32	0.79	0.39	-139.3	130	-179.2	-66.1
239	D	115.3	0.53	B3	11.56	-0.76	0.39	-76	124.7	-168.8	177.2
240	E	53.4	0.73	P2	11.97	0.41	0.43	-110.8	-81.8	-177.4	-67.1
241	A	1.5	0.92	E	12.11	0.14	0.52	-95	-56.4	-179.7	
242	V	25.4	0.9	E	10.45	-1.66	0.55	-85.8	104.4	-172.3	-56.9
243	E	114.5	0.74	B3	10.03	-0.42	0.49	-61.5	127.8	-163.6	-165.1
244	G	17.5	0.85	E	11.78	1.75	0.37	107.4	-39.1	-179.9	
245	D	26.7	0.76	E	11.48	-0.3	0.35	-84.8	111.9	-178.9	-71.3
246	F	186.9	0.45	B2	12.8	1.32	0.4	-88.8	123.3	-171.1	-71.6
247	V	68.9	0.85	P2	12.88	0.08	0.32	-91.5	121.7	178.1	-57.5
248	T	110.9	0.26	P1	13.83	0.95	0.23	-72	118.7	-178.7	-57
249	E	92.9	0.63	P2	14.24	0.41	0.25	-59.8	-36.6	170.2	-168.1
250	T	77.2	0.51	P1	15.19	0.95	0.31	-65.8	-22.5	-177.5	53.5
251	G	7	0.97	E	16.94	1.75	0.32	85.1	78.3	-166.9	
252	R	113.7	0.66	P2	17.13	0.19	0.35	-133.6	163.9	159.9	-65.4
253	H	52.2	0.8	P2	16.64	-0.49	0.45	-65.5	113.6	179.6	67.6
254	L	95.8	0.67	P2	15.31	-1.33	0.43	-81.8	86.5	167.1	179.8
255	K	102	0.78	P2	15.86	0.55	0.43	-78.3	88.5	-157	55.9
256	G	21.1	0.47	E	17.61	1.75	0.42	-52.6	-52.3	-177.9	
257	M	17.9	0.89	E	16.99	-0.62	0.33	-60.3	-43.2	171.5	-68.7
258	E	105.7	0.51	P1	16.31	-0.68	0.33	-78	-26.6	163.7	71.9
259	F	187.6	0.33	B2	17.63	1.32	0.28	-68.3	-52.6	-161	-93.4
260	Q	120.1	0.47	B3	18.15	0.52	0.26	-119.8	73	-171	-169.3
261	V	104	0.45	P1	18.59	0.44	0.13	-89.9	83.6	-176.2	-169.4
262	W	197.6	0.4	B2	19.42	0.83	0.05	-71.2	150.2	160.1	-60.6
263	S	72.1	0.47	P1	19.91	0.49	0.12	-74	146.5	-178.2	65.7
264	P	19.8	0.84	E	19.03	-0.88	0.17	-63.7	114	-159.4	25.8
265	G	7	0.98	E	20.78	1.75	0.19	118.7	-33.1	176.2	
266	E	128.3	0.66	B3	20.36	-0.42	0.14	-89.8	144.9	2.6	-71.3
267	P	99.1	0.5	P1	19.71	-0.65	0.04	-84.4	117.7	-164.9	32.4
268	N	67.5	0.96	P2	19.71	0	0.13	-81.2	-53.6	175.3	79
269	N	52.1	0.79	P2	19.71	0	0.22	-94.8	119.9	-171.5	-74.7
270	D	84.8	0.71	P2	19.68	-0.03	0.18	-130.8	74.7	178.9	-62.8
271	V	35.3	0.9	E	18.02	-1.66	0.2	-118.2	127.6	-171.9	-58.4
272	D	60	0.81	P2	17.99	-0.03	0.21	-71.4	102.8	-176.3	-164.9
273	G	0	1	E	19.74	1.75	0.25	151.9	-70	-178.2	
274	K	85.8	0.67	P2	20.29	0.55	0.24	-134.7	-67.2	-175.8	174.5
275	P	60.5	0.59	P2	19.28	-1.01	0.36	-85.1	102.9	169	33.5
276	E	137.8	0.66	B3	18.86	-0.42	0.26	-134.7	82.3	-170.2	-77.1
277	N	118.2	0.7	B3	18.53	-0.33	0.29	-85.3	-8.2	166	-84.7

APPENDIX B

GHA Model Assessment Data

278	C	56	0.24	P1	19.81	1.28	0.41	-81.9	151.8	162.3	-63.2
279	L	154	0.18	B1	20.88	1.07	0.37	-92.2	135.7	173.2	-172.8
280	A	71	0.34	P1	21.47	0.59	0.41	-118.1	138.1	176	
281	F	183.4	0.22	B1	22.43	0.96	0.43	-74.6	120.1	-172.9	-164.1
282	S	86	0.4	P1	22.92	0.49	0.51	-77.1	140.4	162.6	-66.6
283	G	37.2	0.62	E	24.67	1.75	0.53	-55.7	-39.5	163.8	
284	R	76.6	0.74	P2	24.86	0.19	0.45	-95	38.3	167	52.1
285	G	33.7	0.3	E	26.61	1.75	0.45	83.1	-67.3	-174.7	
286	Y	66.1	0.75	P2	26.31	-0.3	0.5	-77	130.1	-159.1	-73.7
287	Y	197	0.25	B1	26.48	0.17	0.48	-80.3	159	169.6	-65.3
288	G	40	0.38	E	28.23	1.75	0.56	-136.9	158.6	176.8	
289	D	116	0.59	B3	27.47	-0.76	0.55	-83.3	144	-162.8	82.2
290	R	151.7	0.47	B3	28.31	0.84	0.56	-129.1	146	-179.7	-160
291	S	72.1	0.32	P1	28.8	0.49	0.55	-74.8	123	-167.1	-170.7
292	C	37.8	0.45	E	28.76	-0.04	0.55	-58.4	-62.2	175.2	-57
293	E	89.9	0.62	P2	29.17	0.41	0.55	-91.3	59.1	173.1	-67.6
294	V	64.7	0.74	P2	29.25	0.08	0.55	-127.9	156	-176.8	-62.6
295	E	43.2	0.79	P2	29.66	0.41	0.55	-75.9	99.4	-177.8	-61.7
296	L	149.8	0.46	B3	29.8	0.14	0.55	-128	149.5	-173.7	-76.3
297	P	123	0.33	B1	29.03	-0.77	0.55	-69.7	130.8	-174.4	27
298	F	187.6	0.42	B2	30.35	1.32	0.55	-87.4	107.9	165.9	-60
299	I	143	0.34	B2	31.42	1.07	0.55	-83.5	125.5	170.5	-173.1
300	C	45.6	0.52	P1	32.7	1.28	0.55	-92.7	-135.3	-172.3	-53
301	E	75.9	0.68	P2	33.11	0.41	0.55	-160.6	50.3		142

APPENDIX C

Rat MBP Assessment Data

#	R	Ab	Fp	Env	Score	Acc	Avg	ϕ	ψ	Amide	$\chi 1$
116	H	32.7	0.92	E	-0.41	-0.41	0.34		-39.8	-174.4	-65.2
117	E	-1	-1		0	-0.41	0.34	-65.1	134	-172.6	-69
118	R	112.7	0.79	P2	0.24	-0.17	0.34	-91.5	110.6	-173.6	-65.5
119	M	160.8	0.43	B2	1	0.83	0.34	-147.5	150.8	167.1	68.3
120	P	98.4	0.69	P2	0.44	1.27	0.34	-40.6	146.2	-176.2	-36
121	F	186.9	0.26	B1	1.4	2.67	0.34	-57.8	-39.7	166.6	-167.6
122	S	20.1	0.69	E	0.34	3.01	0.34	-61.8	-31.5	162.4	-168.7
123	K	-1	-1		0	3.01	0.34	-69.3	-38.3	168.4	-66.1
124	V	130	0.34	B2	0.66	3.67	0.34	-60.5	-45.6	168	-171.1
125	K	96.6	0.59	P2	0.47	4.14	0.34	-56.1	-45.5	166.1	-67.7
126	A	36.6	0.85	E	0.14	4.28	0.34	-58.6	-51.4	173.6	
127	L	94.3	0.5	P1	-0.33	3.95	0.34	-56.5	-49.6	172.5	-171.9
128	C	56	0.32	P1	1.29	5.24	0.36	-64.3	-41.1	166.6	-64
129	S	23.9	0.87	E	0.34	5.58	0.36	-63	-40.7	163.5	61.4
130	E	23.7	0.87	E	0.04	5.62	0.34	-57	-40.7	167	-70.4
131	L	59.3	0.68	P2	-0.68	4.94	0.33	-77.4	25.1	177.6	-61.8
132	R	0	0.98	E	-0.41	4.53	0.26	59.1	50.3	178.4	-63.7
133	G	19.6	0.85	E	1.1	5.63	0.26	-174.9	-142.5	-175.5	
134	T	57.7	0.58	P2	0.08	5.71	0.23	-137.3	169.9	176.9	50.6
135	V	127.2	0.3	B1	1	6.71	0.2	-73	138.1	-164.1	-167.6
136	A	68.9	0.29	P1	0.49	7.2	0.19	-58	120	-177.4	
137	I	155.6	0.52	B3	-0.54	6.66	0.22	-109.2	107.6	158.3	-69.6
138	P	123	0.29	B1	0.59	7.25	0.24	-76.3	102.2	161.9	30.5
139	R	99.1	0.74	P2	0.24	7.49	0.18	-70.5	-22	178.6	-57.5
140	N	58.5	0.72	P2	0.51	8	0.13	-141.4	170.9	177.5	70
141	A	23.3	0.75	E	0.14	8.14	0.15	-64	-33.9	170.7	
142	E	17.9	0.89	E	0.04	8.18	0.21	-65.8	-40.1	158.9	-167.2
143	E	99.1	0.64	P2	0.28	8.46	0.23	-55.8	-48.1	178.1	-73.1
144	N	119.6	0.37	B2	-0.56	7.9	0.19	-56.1	-49.4	170.8	-157.6
145	K	-1	-1		0	7.9	0.21	-59.9	-44.5	170.2	-65.5
146	A	19.8	0.71	E	0.14	8.04	0.21	-62.4	-45.9	173	
147	I	141.6	0.32	B1	0.93	8.97	0.23	-65.1	-41.4	161.3	-63.3
148	Q	117.1	0.57	B3	0.1	9.07	0.31	-49.1	-47.1	169.9	-176.5
149	E	-1	-1		0	9.07	0.33	-63.1	-39.2	167.6	-68.6
150	V	40.9	0.74	P2	-0.74	8.33	0.34	-64	-42.7	176.1	179
151	A	70.3	0.32	P1	0.49	8.82	0.28	-54.7	-58.4	178.6	
152	K	57.6	0.96	P2	0.47	9.29	0.29	72	-50	154.1	-58.6
153	T	65.1	0.85	P2	0.08	9.37	0.25	-110.6	170.6	-168.8	-173.9
154	S	74.9	0.66	P2	0.17	9.54	0.24	-61.6	122.2	-163.3	70.6
155	A	71	0.23	P1	0.49	10.03	0.28	-119.4	146.8	170.5	
156	F	189	0.38	B2	1.04	11.07	0.33	-70.6	128.2	169.4	-82.5
157	L	154	0.21	B1	1.06	12.13	0.33	-100.9	156.9	175.9	-60
158	G	40	0.21	E	1.1	13.23	0.34	-77	51.7	-171.5	

APPENDIX C

Rat MBP Assessment Data

159	I	157	0.2	B1	0.93	14.16	0.35	-135.1	108.3	-168.6	-56.7
160	T	100	0.42	P1	0.55	14.71	0.41	-145.1	148.2	163	68.2
161	D	114.6	0.59	B3	-0.83	13.88	0.44	-90	49.7	172.5	65.8
162	E	89.5	0.59	P2	0.28	14.16	0.42	-78.6	-43.7	178.4	-70
163	V	28.2	0.8	E	-0.8	13.36	0.45	-66.2	-42	168.1	178.2
164	T	46.8	0.78	P2	0.08	13.44	0.5	-142.6	120	-176.1	-55.9
165	E	86	0.86	P2	0.28	13.72	0.47	-61.2	132.1	-162.7	-169.7
166	G	23.9	0.73	E	1.1	14.82	0.5	96.7	-55.2	178.9	
167	Q	48.8	0.77	P2	0.25	15.07	0.45	-76.4	103.5	173.9	-69.4
168	F	181.3	0.36	B2	1.04	16.11	0.46	-71.8	129.9	-169.9	-87
169	M	115.5	0.69	B3	0.23	16.34	0.43	-130.5	142.1	170.1	-67.6
170	Y	190.7	0.51	B3	1.25	17.59	0.41	-70	116.1	-174.8	-54.6
171	V	106.8	0.38	P1	-0.09	17.5	0.47	-62.5	-26.6	166.1	-57.4
172	T	68.8	0.59	P2	0.08	17.58	0.53	-71.1	-29.5	175.8	55.6
173	G	5.6	0.99	E	1.1	18.68	0.51	160.8	87.3	168.2	
174	G	0	0.94	E	1.1	19.78	0.56	149.3	171.1	165.7	
175	R	13.9	0.88	E	-0.41	19.37	0.53	-85.8	155.8	-172.1	72.3
176	L	142.1	0.31	B1	1.06	20.43	0.55	-67.6	98.8	170.8	-178.5
177	T	38.2	0.92	E	0.08	20.51	0.52	-81.1	-47.7	174.1	55.9
178	Y	157.7	0.58	B3	1.25	21.76	0.53	-83.1	79.9	-175	179.1
179	S	64.8	0.46	P1	0.59	22.35	0.49	-84.7	76	-169.1	72.2
180	N	83.1	0.7	P2	0.51	22.86	0.54	-91.3	68.2	-165.3	-168.6
181	W	200.6	0.35	B2	1.62	24.48	0.49	-96.4	146.6	171.7	-70.8
182	K	107.4	0.63	P2	0.47	24.95	0.55	-79.5	173.9	-177.8	-156.5
183	K	0	0.99	E	-0.1	24.85	0.52	-58.9	130.9	-172.1	-169.6
184	D	12	0.92	E	0.23	25.08	0.49	68	21.3	-176	-60.4
185	E	131.2	0.62	B3	-0.46	24.62	0.5	-123.2	137.9	5.1	-44.9
186	P	85.1	0.49	P1	0.64	25.26	0.57	-89.7	110.7	171.7	33.4
187	N	81.4	0.76	P2	0.51	25.77	0.55	-79.3	-26.9	-172.6	60.7
188	D	74.5	0.76	P2	0.51	26.28	0.58	50.5	41.1	-158.4	-135.1
189	H	57.9	0.93	P2	0.2	26.48	0.49	-98.7	176.4	173.8	-155.5
190	G	11.9	0.84	E	1.1	27.58	0.46	70.4	-92.9	167.7	
191	S	0	1	E	0.34	27.92	0.46	-93.4	1.3	-161.3	73.5
192	G	29.5	0.94	E	1.1	29.02	0.43	91.9	80	179.5	
193	E	132.4	0.65	B3	-0.46	28.56	0.38	-77.9	83.2	-168	-71.8
194	D	77.4	0.74	P2	0.51	29.07	0.43	-93.7	-6.5	-176.5	-66.7
195	C	56	0.23	P1	1.29	30.36	0.44	-117.1	144.1	175.9	-67.4
196	V	130	0.22	B1	1	31.36	0.37	-79.1	132.1	168.9	167.6
197	T	110.3	0.34	P1	0.55	31.91	0.31	-124	121.3	172.5	44.7
198	I	156.3	0.44	B2	0.81	32.72	0.31	-82	123.2	170	-169
199	V	104	0.66	P2	-0.74	31.98	0.27	-85	165.7	179.3	-66.3
200	D	-1	-1		0	31.98	0.25	-51.9	-31.3	167.4	-72.1
201	N	63.9	0.69	P2	0.51	32.49	0.21	-74.4	-12.1	178.2	62.9
202	G	37.9	0.75	E	1.1	33.59	0.2	112.9	-33.2	-178	

APPENDIX C

Rat MBP Assessment Data

203	L	109.8	0.73	P2	-0.68	32.91	0.17	-76.3	155.4	-167.4	-64.2
204	W	234	0.32	B1	0.96	33.87	0.22	-102.5	130.1	166.7	-63.6
205	N	90.6	0.57	P2	0.51	34.38	0.22	-93.6	132.4	160.4	177.4
206	D	115.3	0.37	B2	-1.97	32.41	0.16	-74.3	120.5	177.6	65.3
207	I	114.9	0.64	B3	-0.54	31.87	0.09	-118.7	157.7	164	-160.6
208	S	38.7	0.73	E	0.34	32.21	0.08	-67.1	121.5	-164.3	-172.1
209	C	35	0.37	E	-0.35	31.86	0.08	-52.7	-34.4	176.4	-64.5
210	Q	17.1	0.82	E	-0.03	31.83	0.08	-88.6	26.1	164.1	-63.6
211	A	35.9	0.76	E	0.14	31.97	0.08	-88.7	152.2	-176.2	
212	S	43.6	0.62	P2	0.17	32.14	0.08	-78.3	106.9	170.6	-64.2
213	H	153.4	0.45	B2	0.54	32.68	0.08	-132.5	159.3	-171.3	-75
214	T	108.1	0.37	P1	0.55	33.23	0.08	-59.7	133.4	-173.3	-81.3
215	A	66.1	0.23	P1	0.49	33.72	0.08	-90.2	88.7	179.6	
216	V	113.9	0.37	P1	-0.09	33.63	0.08	-97.3	121.9	178	-176.2
217	C	39.9	0.41	E	-0.35	33.28	0.08	-103.8	141.4	169.1	-57.7
218	E	70.5	0.84	P2	0.28	33.56	0.08	-93.3			60

APPENDIX D

Human ESEL Assessment Data

#	R	Ab	Fp	Env	Score	Acc	Avg	ϕ	ψ	AMIDE	$\chi 1$
7	T	0	0.92	E	0.08	0.08	0.42	-72.2	-170.7	63.3	
8	E	36.7	0.92	E	0.04	0.12	0.42	-127.4	137	-171.7	-169.7
9	A	52.8	0.54	P1	0.49	0.61	0.42	-77.5	103.8	-175.9	
10	M	150.3	0.53	B3	0.23	0.84	0.42	-153.4	167.4	177.5	66.4
11	T	86.4	0.67	P2	0.08	0.92	0.42	-65.4	152.3	-174	68.5
12	Y	185.8	0.38	B2	0.86	1.78	0.42	-54.5	-44.5	172.1	-162.6
13	D	17	0.8	E	0.44	2.22	0.42	-64.2	-34.9	159.6	-63.7
14	E	65	0.77	P2	0.62	2.84	0.42	-63.9	-44.3	178.2	-63.5
15	A	71	0.22	P1	0.76	3.6	0.42	-60	-43.8	166	
16	S	70.1	0.47	P1	0.47	4.07	0.42	-59.1	-48.7	169.4	-167.3
17	A	17.7	0.86	E	0.44	4.51	0.42	-56.1	-47	170	
18	Y	141.9	0.54	B3	0.5	5.01	0.44	-60.2	-50.7	173.5	-176.9
19	C	56	0.39	P1	0.95	5.96	0.48	-61.1	-39.6	165.8	-68.2
20	Q	66.8	0.78	P2	0.62	6.58	0.46	-65.5	-40.7	176.9	-67.2
21	Q	26.7	0.9	E	0.29	6.87	0.48	-67.3	-47.9	175.7	-175.4
22	R	96.2	0.73	P2	0.56	7.43	0.5	-85.8	-45.1	-154.8	-77.3
23	Y	69.3	0.67	P2	-0.43	7	0.49	-117.2	-76.3	175	-52.2
24	T	2.3	0.93	E	0.08	7.08	0.5	-164.2	-68.4	175	-63.9
25	H	151.3	0.59	B3	1.04	8.12	0.5	-163.9	175.1	167.6	-142.6
26	L	154	0.41	B2	0.77	8.89	0.49	-65.2	145.9	-160.1	-71.7
27	V	102.6	0.43	P1	-0.09	8.8	0.44	-64.5	123.9	175.1	73.3
28	A	57	0.41	P1	0.49	9.29	0.48	-95.3	115.1	-179.4	
29	I	156.3	0.33	B2	0.81	10.1	0.43	-90.4	106.4	174.3	-63.5
30	Q	41.2	0.92	P2	0.25	10.35	0.39	-77.9	-37.2	171	-66.6
31	N	72.5	0.67	P2	0.51	10.86	0.34	-147.2	174.7	174.2	71.3
32	K	90.3	0.8	P2	0.66	11.52	0.36	-50.7	-29.4	172.4	-61.4
33	E	37.1	0.93	E	0.6	12.12	0.35	-61.5	-29.3	162.2	-62.9
34	E	91.7	0.73	P2	0.62	12.74	0.43	-64.5	-43.3	167.7	-76.2
35	I	143.7	0.38	B2	0.55	13.29	0.45	-58.4	-49	168.6	-62.4
36	E	42.3	0.92	P2	0.62	13.91	0.42	-50.2	-50.2	169.1	-169.5
37	Y	44.8	0.78	P2	-0.55	13.36	0.39	-59.7	-56.3	179.6	-180
38	L	131.5	0.2	B1	1.3	14.66	0.45	-58.3	-47.4	167.7	-62.6
39	N	90.7	0.44	P1	-0.58	14.08	0.48	-67.1	-12.4	168.5	-159.7
40	S	18.1	0.87	E	0.16	14.24	0.52	-87.5	-51	-179.5	-58.6
41	I	63	0.69	P2	-0.59	13.65	0.55	-76.7	-36.2	-177.3	-160.7
42	L	118.9	0.34	B2	0.77	14.42	0.58	-85.1	132.5	176.9	-58.1
43	S	8.3	1	E	0.34	14.76	0.62	-77.8	158.2	174.9	-65.9
44	Y	115.4	0.63	B3	1.25	16.01	0.63	-61.7	125.8	161.4	177.7
45	S	74.9	0.49	P1	0.59	16.6	0.62	-126.4	131.3	170	-176.8
46	P	38.1	0.8	E	0.25	16.85	0.51	-65.6	-11.5	154.5	30.1
47	S	57.2	0.68	P2	0.17	17.02	0.5	-61.8	-40.6	-176.3	73.9
48	Y	165.6	0.63	B3	1.25	18.27	0.55	74.3	-176.6	-162.9	-53.6
49	Y	179.6	0.37	B2	1.14	19.41	0.49	-127.1	127.3	179.3	-70.3

APPENDIX D

Human ESEL Assessment Data

50	W	234	0.37	B2	1.62	21.03	0.56	-77.6	125.1	175.2	-69.2
51	I	157	0.2	B1	0.93	21.96	0.57	-99.9	155.6	-175.3	67.3
52	G	40	0.4	E	1.1	23.06	0.64	-85.6	68.8	179.8	
53	I	157	0.17	B1	1.5	24.56	0.57	-137.1	134.2	179.6	-154.6
54	R	156.8	0.62	B3	0.84	25.4	0.6	-130.2	139.3	179.2	-67.8
55	K	140.1	0.73	B3	0.35	25.75	0.57	-90.9	118.1	167.9	179.2
56	V	35.3	0.77	E	-1.66	24.09	0.54	-97.8	79.6	-176.3	-179.4
57	N	0	0.94	E	0.41	24.5	0.55	52.5	77.8	-165.5	-165.5
58	N	36	0.94	E	0.41	24.91	0.56	69.9	-30.2	-174.9	-157.2
59	V	50	0.77	P2	0.08	24.99	0.56	-80.7	98.2	178.6	-177.8
60	W	205.4	0.34	B2	0.83	25.82	0.51	-78.5	118.5	-177	-72.7
61	V	101.2	0.53	P1	0.44	26.26	0.43	-117.4	127.9	170.8	-179.2
62	W	222.8	0.36	B2	0.83	27.09	0.4	-71.8	116.8	-177.5	-66.1
63	V	101.2	0.63	P2	-0.74	26.35	0.33	-73.7	-36	175.8	-64.4
64	G	36.5	0.24	E	1.1	27.45	0.29	-66	-59.7	169.8	
65	T	57.6	0.52	P1	0.55	28	0.23	-73.7	-17.3	-173	58.3
66	Q	23.9	0.93	E	-0.03	27.97	0.26	63.5	62.8	-169.8	-62.9
67	K	51.9	0.91	P2	0.47	28.44	0.36	-141.7	144.9	167.5	-71.4
68	P	56.3	0.63	P2	0.44	28.88	0.36	-68.7	159.8	-178.1	31.9
69	L	151.2	0.32	B1	1.06	29.94	0.39	-68.2	110.5	174.5	179.4
70	T	55.5	0.86	P2	0.08	30.02	0.36	-76.2	144	173.2	63
71	Q	37.7	0.88	E	0.04	30.06	0.36	-61.4	-35.2	163.4	-60.7
72	Q	45.8	0.91	P2	0.28	30.34	0.36	-61.9	-22.1	164.1	69.9
73	A	71	0.6	P2	-0.25	30.09	0.34	-84.7	-34.1	-169.5	
74	K	101.4	0.64	P2	0.47	30.56	0.36	-59.4	105.6	-179.9	-173.8
75	N	101	0.41	P1	-0.26	30.3	0.31	-130.2	74.4	179.2	-163.4
76	W	214.4	0.22	B1	0.96	31.26	0.27	-85	147	167.1	-67.4
77	A	68.2	0.33	P1	0.49	31.75	0.3	-67	153.6	-177	
78	P	34.6	0.75	E	0.25	32	0.29	-66.4	110	-161.3	26.2
79	G	0	0.98	E	1.1	33.1	0.29	111.9	-12.5	-175.9	
80	E	133.2	0.7	B3	-0.46	32.64	0.3	-107.3	147.8	3.8	-71
81	P	99.8	0.45	P1	0.64	33.28	0.36	-84.1	123.8	176.1	31.7
82	N	52.3	0.96	P2	0.51	33.79	0.35	-83.5	-38.5	-173.2	79.8
83	N	100.3	0.74	P2	0.51	34.3	0.4	-84.7	104.4	177.6	-80.2
84	R	30.4	0.99	E	-0.41	33.89	0.48	-73.4	-24.2	-171.3	-70.4
85	Q	28	0.92	E	-0.03	33.86	0.5	-79	151.1	172	-174.1
86	K	19.1	1	E	-0.1	33.76	0.51	-62.1	132.7	-166.8	-172.7
87	D	50.3	0.89	P2	0.51	34.27	0.5	68.5	-9.6	149.1	-69.3
88	E	110.4	0.65	P2	0.28	34.55	0.48	-88.9	-8.9	173.5	-58.6
89	D	99.2	0.62	P2	0.51	35.06	0.46	-74.8	-0.8	-168.6	70.7
90	C	56	0.3	P1	1.28	36.34	0.43	-79.1	124.1	-159.6	-69.8
91	V	130	0.26	B1	1.18	37.52	0.41	-96.4	123.5	164.1	-177
92	E	144.1	0.44	B2	-0.2	37.32	0.44	-88.9	136.8	167.5	71.2
93	I	157	0.2	B1	1.5	38.82	0.42	-96	130.4	-177.2	-170.5

APPENDIX D

Human ESEL Assessment Data

94	Y	174.9	0.58	B3	1.25	40.07	0.44	-82.8	76.8	-174.9	-68.4
95	I	157	0.3	B1	0.93	41	0.46	-69.1	119.8	178.4	-62.6
96	K	118.4	0.71	B3	0.08	41.08	0.43	56.1	38.8	-172.1	-64.5
97	R	144.8	0.56	B3	0.71	41.79	0.42	-92.6	144.9	167.6	-168.9
98	E	29.4	0.83	E	0.04	41.83	0.41	-52	-50	175.5	76.5
99	K	40	0.92	E	-0.1	41.73	0.45	-110.2	113.4	178.5	-72
100	D	73.4	0.45	P1	0.34	42.07	0.45	75.6	70.5	-177.5	-54.7
101	V	67.5	0.85	P2	-0.74	41.33	0.39	-76.1	126.8	-173.2	67.6
102	G	37.2	0.25	E	1.1	42.43	0.35	73.7	2.3	-167.9	
103	M	116	0.68	B3	0.23	42.66	0.37	-94.6	161	-169.5	-68.6
104	W	234	0.21	B1	0.92	43.58	0.3	-101.8	140	173.3	-64.1
105	N	106.5	0.68	P2	0	43.58	0.27	-134.2	123.8	176.6	-178.8
106	D	113.3	0.48	P1	-0.61	42.97	0.28	-83.4	125.2	-161.3	82.8
107	E	118	0.68	B3	-0.42	42.55	0.34	-136	163.2	171.3	-176.3
108	R	84.3	0.87	P2	0.24	42.79	0.34	-62.8	132.6	178.3	-165.6
109	C	42	0.47	P1	1.29	44.08	0.34	-54.1	-36.5	179.3	-75.8

APPENDIX E

Dihedral Angle Assessment Data

Human ESEL			Rat MBP			GHA Model		
#	R	Amide	#	R	Amide	#	R	Amide
7	T		116	H		191	A	
8	E	-171.7	117	E	-172.6	192	E	-167.8
9	A	-175.9	118	R	-173.6	193	K	178.1
10	M	177.5	119	M	167.1	194	A	-170.5
11	T	-174	120	P	-176.2	195	V	-178.6
12	Y	172.1	121	F	166.6	196	Y	166.1
13	D	159.6	122	S	162.4	197	A	161.4
14	E	178.2	123	K	168.4	198	E	-176.1
15	A	166	124	V	168	199	A	172.9
16	S	169.4	125	K	166.1	200	A	169.7
17	A	170	126	A	173.6	201	R	173.5
18	Y	173.5	127	L	172.5	202	V	170.1
19	C	165.8	128	C	166.6	203	C	161.9
20	Q	176.9	129	S	163.5	204	R	167.2
21	Q	175.7	130	E	167	205	S	178
22	R	-154.8	131	L	177.6	206	E	-171.3
23	Y	175	132	R	178.4	207	N	-179.3
24	T	175	133	G	-175.5	208	A	-168.2
25	H	167.6	134	T	176.9	209	T	-142.3
26	L	-160.1	135	V	-164.1	210	L	-173.6
27	V	175.1	136	A	-177.4	211	A	-175.3
28	A	-179.4	137	I	158.3	212	V	165.1
29	I	174.3	138	P	161.9	213	P	-175.5
30	Q	171	139	R	178.6	214	D	-175.2
31	N	174.2	140	N	177.5	215	T	176.2
32	K	172.4	141	A	170.7	216	W	170
33	E	162.2	142	E	158.9	217	D	163.8
34	E	167.7	143	E	178.1	218	R	168.8
35	I	168.6	144	N	170.8	219	V	168.3
36	E	169.1	145	K	170.2	220	E	174.7
37	Y	179.6	146	A	173	221	T	164.9
38	L	167.7	147	I	161.3	222	L	170.6
39	N	168.5	148	Q	169.9	223	L	169
40	S	-179.5	149	E	167.6	224	R	169.9
41	I	-177.3	150	V	176.1	225	L	-176
42	L	176.9	151	A	178.6	226	L	163.4
43	S	174.9	152	K	154.1	227	E	-176.5
44	Y	161.4	153	T	-168.8	228	P	-170.9
45	S	170	154	S	-163.3	229	K	-168
46	P	154.5	155	A	170.5	230	E	170.2
47	S	-176.3	156	F	169.4	231	E	167.2
48	Y	-162.9	157	L	175.9	232	F	-171.2

APPENDIX E

Dihedral Angle Assessment Data

49	Y	179.3	158	G	-171.5	233	Y	-173.9
50	W	175.2	159	I	-168.6	234	L	-176.6
51	I	-175.3	160	T	163	235	T	164.6
52	G	179.8	161	D	172.5	236	G	-168
53	I	179.6	162	E	178.4	237	F	-157.2
54	R	179.2	163	V	168.1	238	T	-179.2
55	K	167.9	164	T	-176.1	239	D	-168.8
56	V	-176.3	165	E	-162.7	240	E	-177.4
57	N	-165.5	166	G	178.9	241	A	-179.7
58	N	-174.9	167	Q	173.9	242	V	-172.3
59	V	178.6	168	F	-169.9	243	E	-163.6
60	W	-177	169	M	170.1	244	G	-179.9
61	V	170.8	170	Y	-174.8	245	D	-178.9
62	W	-177.5	171	V	166.1	246	F	-171.1
63	V	175.8	172	T	175.8	247	V	178.1
64	G	169.8	173	G	168.2	248	T	-178.7
65	T	-173	174	G	165.7	249	E	170.2
66	Q	-169.8	175	R	-172.1	250	T	-177.5
67	K	167.5	176	L	170.8	251	G	-166.9
68	P	-178.1	177	T	174.1	252	R	159.9
69	L	174.5	178	Y	-175	253	H	179.6
70	T	173.2	179	S	-169.1	254	L	167.1
71	Q	163.4	180	N	-165.3	255	K	-157
72	Q	164.1	181	W	171.7	256	G	-177.9
73	A	-169.5	182	K	-177.8	257	M	171.5
74	K	-179.9	183	K	-172.1	258	E	163.7
75	N	179.2	184	D	-176	259	F	-161
76	W	167.1	185	E	5.1	260	Q	-171
77	A	-177	186	P	171.7	261	V	-176.2
78	P	-161.3	187	N	-172.6	262	W	160.1
79	G	-175.9	188	D	-158.4	263	S	-178.2
80	E	3.8	189	H	173.8	264	P	-159.4
81	P	176.1	190	G	167.7	265	G	176.2
82	N	-173.2	191	S	-161.3	266	E	7.1
83	N	177.6	192	G	179.5	267	P	-164.9
84	R	-171.3	193	E	-168	268	N	175.3
85	Q	172	194	D	-176.5	269	N	-171.5
86	K	-166.8	195	C	175.9	270	D	178.9
87	D	149.1	196	V	168.9	271	V	-171.9
88	E	173.5	197	T	172.5	272	D	-176.3
89	D	-168.6	198	I	170	273	G	-178.2
90	C	-159.6	199	V	179.3	274	K	-175.8
91	V	164.1	200	D	167.4	275	P	-169.6
92	E	167.5	201	N	178.2	276	E	-170.2

APPENDIX E

Dihedral Angle Assessment Data

93	I	-177.2
94	Y	-174.9
95	I	178.4
96	K	-172.1
97	R	167.6
98	E	175.5
99	K	178.5
100	D	-177.5
101	V	-173.2
102	G	-167.9
103	M	-169.5
104	W	173.3
105	N	176.6
106	D	-161.3
107	E	171.3
108	R	178.3
109	C	179.3
202	G	-178
203	L	-167.4
204	W	166.7
205	N	160.4
206	D	177.6
207	I	164
208	S	-164.3
209	C	176.4
210	Q	164.1
211	A	-176.2
212	S	170.6
213	H	-171.3
214	T	-173.3
215	A	179.6
216	V	178
217	C	169.1
218	E	
277	N	166
278	C	162.3
279	L	173.2
280	A	176
281	F	-172.9
282	S	162.6
283	G	163.8
284	R	167
285	G	-174.7
286	Y	-159.1
287	Y	169.6
288	G	176.8
289	D	-162.8
290	R	-179.7
291	S	-167.1
292	C	175.2
293	E	173.1
294	V	-176.8
295	E	-177.8
296	L	-173.7
297	P	-174.4
298	F	165.9
299	I	170.5
300	C	-172.3
301	E	-77.2