School of Biomedical Sciences

Molecular Modelling of Platelet Endothelial Cell Adhesion Molecule 1 and its Interactions with Glycosaminoglycans.

Neha Sureshchandra Gandhi

This thesis is presented for the Degree of
Master of Science
of
Curtin University of Technology

October 2007

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: ………………………………………….

Date: ………………………...

I dedicate this thesis to my parents who support and encourage me on all of my endeavours. I would also like to dedicate it to my other family members and friends who touch my heart and gave me strength to move forward to something better in life. I also dedicate my thesis to disabled children everywhere, with the hope that their families realise their potential and small accomplishments and help them to cherish their dreams.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my supervisor Dr. Ricardo Mancera who has supported and encouraged me throughout this project. He was always there to meet, listen and advice about my ideas, to proofread and mark up my papers and chapters, and to ask me good questions to help me throughout the course of this research work. His mentorship has been invaluable throughout the research. No words could express my gratitude.

I would like to thank to Assoc. Prof. Deirdre Coombe for her much appreciated advice. I thank her for serving on my supervisory committee and for her careful perusal of this thesis. Special thanks go to Prof. Erik Helmerhorst, Head of the School of Biomedical Sciences, for his support and allowing me to carry out the project in the school.

I would also like to thank the staff and students of Biomedical Sciences and Pharmacy for their support, technical assistance and friendship afforded to me during this study. Darran Carey at IVEC supercomputing centre in Bentley helped me getting started with the Cognac computing facility and offered extensive help, especially during the initial stage of the project.

I appreciate the financial support of Bank of India, Navsari branch, Gujarat during the year of my Master's study.

On a more personal note, I would like to thank my parents, my brother, sister-in-law and friends Manisha Parmar, Shabana Kapadia, Pragnya Anglay, Preeti Raval, Pradeep Shukla, Pankaj Bhatt, Peeush Tomar, Vaishali Munot and many more. Without their love, support and constant encouragement this work would not have been possible. I also appreciate help from my colleagues Jestin Mandumpal and Miguel Renteria who were a constant source of entertainment and support. I also thank many of my flatmates at Rotary International House specially Michael Connaughton and friends at Kurrajong Village, Curtin University for making my stay in Bentley memorable.

Table of Contents

Page

ABBREVIATIONS

Uronosyl-2-*O*-Sulphotransferase	2OST
Glucosaminyl 3-*O*-Sulfotransferases	3OST
Glucosaminyl 6-*O*-Sulfotransferases	6OST
Beta-Amyloid Peptide	Aβ
Adapted Basis-Set Newton-Raphson	ABNR
Assisted Model Building and Energy Refinement	AMBER
Activated Protein C	APC
Antithrombin III	AT-III
Basic Local Alignment Search Tool	BLAST
Biomolecular Ligand Energy Evaluation Protocol	BLEEP
BLOck SUbstitution Matrix	BLOSUM
Cell Adhesion Molecule	CAM
Cluster of Differentiation	CD
Conserved Domain Database	CDD
Chemokine Domain of Fractalkine	CDF
Complementarity-Determining Regions	CDR
Consistent Force Field	CFF
Chemistry at HARvard Macromolecular Mechanics	CHARMM
Conjugate Gradient	CONJ
Chondroitin Sulfate	CS
Carbohydrate Solution Force Field	CSFF
Discrete Optimised Protein Energy	DOPE
Dermatan Sulfate	DS
Endothelial Cell	EC
Extra-Cellular Matrix	ECM
Exostoses	EXT
Fibroblast Growth Factor	FGF
Fibroblast Growth Factor Receptor	FGFR
Factor Va	FVa
Genetic Algorithm	GA

Generalized AMBER Force Field	GAFF
Glycosaminoglycans	GAGs
Galactose	Gal
2-deoxy-2-acetamido-α-D-glucopyranosyl	GalNAc
Generalised Born	GB
β-L-glucuronic acid	GlcA
2-deoxy-2-sulfamido-α-D-glucopyranosyl	GlcNS
2-deoxy-2-sulfamido-α-D-glucopyranosyl-6-*O*-sulfate	GlcNS6S
GROningen MOlecular Simulation package	GROMOS
Heparan Sulfate	HS
Heparan Sulphate Proteoglycan	HSPG
Intercellular Cell Adhesion Molecule	ICAM
α-L-iduronic acid	IdoA
2-*O*-sulfo-α-L-iduronic acid	IdoA2S
1,4-dideoxy-*O*-2-sulfo-glucuronic acid	IDU
Interferon-Gamma	IFNγ
Immunoglobulin domains	Ig-domains
Immunoglobulin Super Family	IgSF
Immunoreceptor Tyrosine-based Inhibitory Motifs	ITIMs
Junction Adhesion Molecules	JAM
Dissociation Constant	K_d
Lamarckian Genetic Algorithm	LGA
Mucosal Addressin Cell Adhesion Molecule	MADCAM
Molecular Dynamics	MD
Monocyte Chemoattractant Protein-1	MCP-1
Mast Cell Protease 6	MCP-6
Major Histocompatibility Complex	MHC
Metal Ion-Dependent Adhesion Site	MIDAS
Macrophage Inflammatory Protein 1 alpha	MIP-1a
Neural Cell Adhesion Molecule	NCAM
N-deacetylase/N-sulfotransferase enzymes	NDSTs
Optimised Potential for Liquid Simulations	OPLS
Point Accepted Mutation	PAM

3’-PhosphoAdenosyl-5’-PhosphoSulfate	PAPS
Poisson Boltzmann	PB
Protein Data Bank	PDB
Platelet Endothelial Cell Adhesion Molecule	PECAM
Potential Energy Surface	PES
Platelet Factor 4	PF4
Proteoglycan	PG
Protein Homology/analogY Recognition Engine	Phyre
Potential of Mean Force score	PMFscore
Position Specific Iterated BLAST	PSI-BLAST
Position Specific Scoring Matrix	PSSM
Restrained ElectroStatic Potential fit	RESP
Root Mean Square Deviation	RMSD
Simulated Annealing	SA
Structurally Conserved Regions	SCRs
Steepest Descent	SD
Stromal Cell-Derived Factor-1	SDF-1
N,*O*-6-disulfo-glucosamine	SGN
Src Homology 2-containing tyrosine Phosphatases	SHP-2
Simple Point Charge/Extended	SPC/E
Support Vector Machines	SVM
Transforming Growth Factor β-1	TGFβ-1
1,4-dideoxy-5-dehydro-*O*2-sulfo-glucuronic acid	UAP
Vascular Cell Adhesion Molecule	VCAM
Van der Waals	VDW
Vascular Endothelial Growth Factor	VEGF
4-deoxy-L-threo-2-sulfohex-4-enopyranosyluronic acid	ΔUA2S

POSTERS AND PUBLICATIONS

Platelet Endothelial Cell Adhesion Molecule (PECAM-1) and its Interactions with Glycosaminoglycans I: Molecular Modeling Studies. Neha S. Gandhi, Deirdre R. Coombe and Ricardo L. Mancera, 2008 (Biochemistry, accepted for publication).

Platelet Endothelial Cell Adhesion Molecule (PECAM-1) and its Interactions with Glycosaminoglycans II : Biochemical Analyses. Deirdre R. Coombe, Sandra S. Stevenson, Beverley F. Kinnear, Neha S. Gandhi, Ricardo L. Mancera, Ronald I.W. Osmond and Warren C. Kett, 2008 (Biochemistry, accepted for publication).

Free energy of binding of heparin fragments to the PECAM-1 receptor. Neha Gandhi and Ricardo Mancera, 2007 (manuscript in preparation).

Molecular modelling of the structure of human PECAM-1 and its interactions with glycosaminoglycans. Neha Gandhi, Deirdre Coombe and Ricardo Mancera, Poster presentation at MM2007, a meeting of the Association of Molecular Modellers of Australasia.

Molecular Modelling of the PECAM-1 Receptor and its Interactions with Glycosaminoglycans. Neha Gandhi, Deirdre Coombe and Ricardo Mancera, Bioinfosummer, 2007.

LIST OF FIGURES

Page

LIST OF TABLES

Page

SUMMARY

The Platelet Endothelial Cell Adhesion Molecule 1 (PECAM-1) has many functions including its roles in leukocyte extravasation as part of the inflammatory response, and in the maintenance of vascular integrity through its contribution to endothelial cell-cell adhesion. Various heterophilic ligands of PECAM-1 have been proposed. The possible interaction of PECAM-1 with glycosaminoglycans (GAGs) is the focus of this thesis. The three dimensional structure of the extracellular immunoglobulin (Ig)-domains of PECAM-1 was constructed using homology modelling and threading methods. Potential heparin/heparan sulfate binding sites were predicted on the basis of their amino acid consensus sequences and a comparison with known structures of sulfate binding proteins. Heparin and other GAG fragments have been docked to investigate the structural determinants of their protein binding specificity and selectivity. It is predicted that two regions in PECAM-1 appear to bind heparin oligosaccharides. A high affinity binding region was located in Ig-domains 2 and 3 and a low affinity region was located in Ig-domains 5 and 6. These GAG binding regions are distinct from regions involved in PECAM-1 homophilic interactions. Docking of heparin fragments of different size revealed that fragments as small as a pentasaccharide appear to be able to bind to domains 2 and 3 with high affinity. Binding of longer heparin fragments suggests that key interactions can occur between six sulfates in a hexasaccharide with no further increase in binding affinity for longer fragments. Molecular dynamics simulations were also used to characterise and quantify the interactions of heparin fragments with PECAM-1. These simulations confirmed the existence of regions of high and low affinity for GAG binding and revealed that both electrostatic and van der Waals interactions determine the specificity and binding affinity of GAG fragments to PECAM-1. The simulations also suggested the existence of ‘open’ and ‘closed’ conformations of PECAM-1 around domains 2 and 3.

This is a brief synopsis of the thesis:

CHAPTER 1. This chapter provides a historical and general introduction about GAGs and GAG-binding proteins. The biological characterisation of the interactions of

GAGs with various proteins such as cell adhesion molecules is reviewed, with particular attention to the biology of PECAM-1 and its implication in heterophilic interactions with other molecules including GAGs. The structural and molecular modelling studies of GAG-protein interactions are reviewed in detail.

CHAPTER 2. This chapter discusses the theoretical background of various molecular modelling techniques such as homology modelling, threading, ligand-protein docking and MM/PBSA.

CHAPTER 3. This chapter describes how homology modelling and threading methods were used to construct a three-dimensional model of PECAM-1. The model of PECAM-1 is analysed in detail, including a prediction of its likely heparin/HS binding regions.

CHAPTER 4. This chapter describes how molecular docking was used to model the interactions of GAGs with PECAM-1.

CHAPTER 5. This chapter describes how MM/PBSA simulations were used to characterise the interactions of heparin fragments with PECAM-1.

CHAPTER 6. In this chapter the main conclusions derived from this study are outlined and the likely direction of future work is discussed.

Chapter 1

LITERATURE REVIEW

INTRODUCTION

The ubiquitous presence of glycosaminoglycans (GAGs) and their interactions with numerous immunologically-relevant proteins is now attracting considerable interest as a source of new therapeutics for the treatment of infectious diseases, inflammation, allergies and cancers. One of the key functions of GAGs is to regulate the activity of proteins that they bind. Site-directed mutagenesis, protein sequence mapping using synthetic peptides, NMR, X-ray crystallography and molecular modelling have assisted in understanding the molecular basis of such interactions. The complexity of the interactions occurring between GAGs and proteins is in part due to the conformational flexibility and underlying sulfation patterns of GAGs, the binding of metal ions to both protein and GAGs and the effects of pH on GAG-protein binding. New approaches to carbohydrate synthetic chemistry have allowed the synthesis of GAG oligosaccharides and GAG mimetics, some of which may be novel glycotherapeutics. This literature review provides an overview of the understanding of the structural attributes involved in GAG-protein interactions. The focus of this thesis is the interaction of GAGs with PECAM-1, a member of the Ig-super family of proteins.

1.1 THE IMMUNOGLOBULIN SUPERFAMILY

The immunoglobulin superfamily (IgSF) is a group of proteins recognised as one of the largest protein families in vertebrate genomes. All members of the IgSF have at least one immunoglobulin domain (Ig-domain). Ig-domains are named after immunoglobulin molecules, which have two identical heavy chains and two identical light chains connected by disulfide bonds. The immunoglobulin fold was first discovered nearly 35 years ago (Poljak *et al.* 1973). The characteristic structure of an Ig-domain which identifies this gene family is a two-layer sandwich of varying number of antiparallel β-strands stabilised by disulfide bridges, taking the form of a conserved disulfide bridge packed against a tryptophan or tyrosine amino acid that stabilises the Ig-fold. The backbone switches repeatedly between the two β-sheets as they form a ‘pin’ structure or "X" arrangement in such a way that the N- and C-terminal hairpins are facing each other.

Two basic types of Ig-domains have been defined from crystallographic analysis, namely the variable (IgV; Interpro database: IPR013106) and constant (IgV; Interpro database: IPR013106) regions (Williams & Barclay 1988). IgV-domains are generally longer (with 9 beta-strands) than IgC-domains (with 7 beta-strands), as shown in Figure 1.1. The domain structure of a V region is dominated by a series of nine anti-parallel β strands, connected by variable-length loop sequences and a conserved disulfide bridge between strands B and F. The IgV domains encompass five beta strands in one sheet and four beta strands in another, whilst the “a” strand can occupy a variable position (Bork *et al.* 1994). The IgC domains have four beta strands in one sheet and three on the other. The IgC domain lacks the pair of internal β strands (the c’ and c” strands are missing in these domains), but they otherwise assume the same general structure with a distinct but overlapping series of conserved residues. The lack of this extra pair of c’ and c” strands decreases the distance between the two cysteine residues in the strands relative to that of V regions. The Ig domains of some IgSF members resemble IgV-domains in their amino acid composition, yet they are similar in size to IgC-domains (Harpaz & Chothia 1994). These are the C2-set (constant-2; Interpro database: IPR008424) and the I-set (intermediate; Interpro database: IPR013098) Ig-domains.

Proteins of the IgSF carry out numerous functions in the immune system, in cell-cell recognition and in structural organisation of muscle (Barclay 2003). IgSF domains are known to be present in cell surface receptors such as NCAM (Neural Cell Adhesion Molecule) and mediate homophilic (antiparallel inter-digitation of opposing receptor or protein molecules on adjacent cells) and heterophilic interactions (to other ligands). Members of IgSF share very low sequence similarity. The IgV domains are found in myelin membrane adhesion molecules, T-cell surface glycoproteins, junction adhesion molecules (JAM), Coxsackie virus, adenovirus Car receptors and viral haemagglutinin. IgC domains are found in immunoglobulin light and heavy chains and in the major histocompatibility complex (MHC) class I and II molecules. The C2-set topology is found primarily in mammalian T-cell surface antigens (Cluster of Differentiation) CD2, CD4 and CD80, in Vascular Cell Adhesion Molecule (VCAM) and Intercellular Cell Adhesion Molecule (ICAM). I-set topology is the key feature of cell adhesion molecules of the Ig-superfamily such as ICAM, VCAM, NCAM (Neural Cell Adhesion Molecule), Mucosal Addressin Cell Adhesion Molecule (MADCAM) as well as JAM. IgSF members with an I-set topology set have been shown to be involved in a variety of cell-cell interactions (Chothia & Jones 1997). PECAM-1, the biological target considered in this project is a member the Ig superfamily.

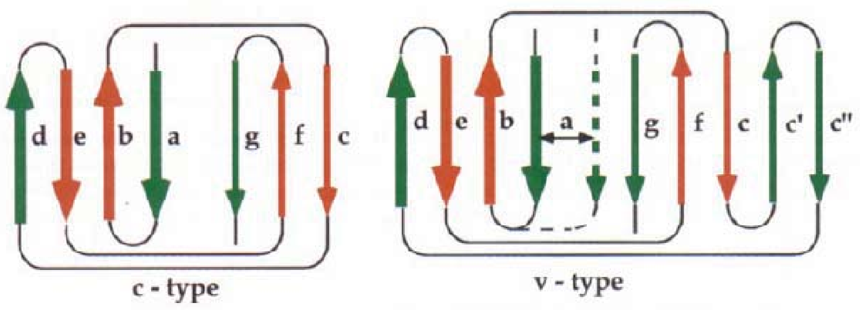


Figure 1.1. *Different topologies of the immunoglobulin fold based on the composition of beta strands and conserved disulfide bridges. The c-type domain has seven strands and v-type domain has nine strands forming a sandwich of 2 sheets. The structural core beta strands b, c, e and f (coloured in red) common to most Ig-domains are surrounded by structurally more variable strands (coloured in green). The thin arrows represent the back sheet (a-b-e-d) whereas the thick arrows represent the front sheet (g-f-c-c'-c''). Strand 'a' in V-type can assume slightly different positions. The figure was taken from Bork et al. (1994).*

1.2 BIOLOGY OF PECAM-1

PECAM-1/CD31 is a member of the cell adhesion molecule (CAM) subgroup of IgSF, and is expressed on the surface of circulating platelets, monocytes, neutrophils and a subpopulation of circulating T-lymphocytes. It is also found in CD43+ haemopoietic progenitor cells in the bone marrow (Newman 1997). Its mRNA is highly expressed in the kidney, lung and trachea and, at lower levels, in the brain, heart and liver. It is not expressed by fibroblasts, epithelial cells or red blood cells (Y. Wang *et al.* 2003).

1.2.1 Structural features of PECAM-1

The 130-kD translated sequence of PECAM-1/CD31 contains six extracellular C2-type Ig-like domains (574 amino acids), one transmembrane domain and a 118-amino acid cytoplasmic domain (see Figure 1.2) (Newman 1997, 1999; Newman *et al.* 1990). The extracellular domains of PECAM-1 are characterised by the Ig fold. Five out of the six Ig domains in PECAM-1 comprise a beta-sandwich made up of seven anti-parallel β strands joined with a Greek key topology forming a C2-type fold. Domain 5 is incomplete, having β strands that form only one side of the sandwich (Newman *et al.* 1990). When PECAM-1 was cloned, it was assigned a C-set topology; however, with the introduction of I-set topologies, new classifications for PECAM-1 extracellular domains were inevitable. Six disulfide bridges are present in the Ig-domains of human PECAM-1. The major structural features of the Ig-domains of human PECAM-1, as described in Swiss-Prot (P16284), are given in Table 1.1.

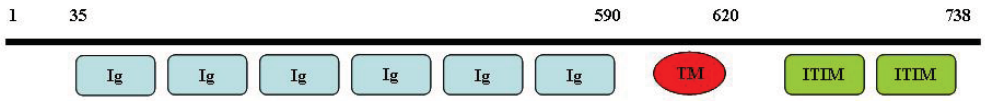


Figure 1.2. *Structural organization of the domains of PECAM-1. The line indicates the relative positions of the amino acids with various structural features of PECAM-1 based on Swiss-Prot (P16284). Ig=Ig-like domains; TM= Transmembrane domains; ITIM= Immunoreceptor Tyrosine-based Inhibitory Motifs*

The cytoplasmic tail of PECAM-1 contains amino acid motifs recognised as having a role in cell signaling. These motifs are called immunoreceptor tyrosine-based inhibitory motifs (ITIMs). This intracellular region contains two tyrosines (Y663 and Y686) conforming to the ITIM (I/VxYxxL/V/I>20aa.I/VxxYxxL/V/I) sequence, where V is valine, L is leucine, I is isoleucine, Y is tyrosine and X can be any other amino acid. These two motifs are separated by more than 20 amino acids. When PECAM-1 interacts with a ligand, these ITIM motifs become phosphorylated by enzymes of the Src family of kinases, allowing them to recruit other enzymes such as the phosphotyrosine phosphatases. These phosphatases decrease the activation of molecules involved in cell signalling.

PECAM-1 is differentially glycosylated in both N-linked and O-linked glycosylation sites (Newton *et al.* 1999), which are important features in the regulation of PECAM-1 cell surface interactions and signal transduction. PECAM-1 is glycosylated to the extent that approximately 39% of its molecular weight is attributable to carbohydrates, and the mature protein has nine putative consensus asparagine-linked (N-linked) glycosylation sites. Other post-translational modifications of PECAM-1 include palmitoylation of cysteine 595 (D. E. Jackson 2006) and phosphorylation of the cytoplasmic tail (D. E. Jackson 2003).

Table 1.1. Structural features of human PECAM-1 as described in Swiss-Prot (P16284).

Domain	Sequence range in PECAM-1
Signal	1-27
Transmembrane	602-620
Cytoplasmic	621-738
Ig-like C2-type 1 Ig-like C2-type 2 Ig-like C2-type 3 Ig-like C2-type 4 Ig-like C2-type 5 Ig-like C2-type 6	35-121 145-233 236-315 328-401 424-493 499-591
Disulfide bridges	57-109 152-206 256-304 347-386 431-476 523-572
N-linked glycosylation sites	52, 84, 151, 301, 320, 344, 356, 453, 551

1.2.2 Genomic organisation of PECAM-1

PECAM-1 has been mapped to human chromosome 17 in the region 17q23 and to mouse chromosome 6. The open reading frame of PECAM-1 is composed of 16 exons. Exons 1 and 2 encode the 5'-untranslated region and the signal peptide, exons 3–8 encode six Ig-like homology domains, exon 9 encodes the transmembrane portion of the protein and exons 10–16 encode the cytoplasmic domain. There are several alternatively spliced variants of PECAM-1 that are expressed in a cell-type, tissue and species-specific pattern in human, rat and mouse. Figure 1.3 shows that there is 79% sequence identity between human PECAM-1 and PECAM-1 from mouse, pig, rat and bovine. Different isoforms of PECAM-1 are known to arise due to the alternative splicing of either the transmembrane or cytoplasmic domain exons (Y. Wang & Sheibani 2002; Y. Wang *et al.* 2003). The human PECAM-1 gene encodes full-length human PECAM-1, which is predominant in human tissue and endothelial cells, and five other isoforms, which lack exon 12, 13, 14, or 15, or exons 14 and 15 (Figure 1.4). The PECAM-1 isoform lacking exons 14 and 15 is the predominant isoform in murine endothelium. The PECAM-1 isoform lacking exon 13 has been detected in human hematopoietic cells and endothelial cells. This isoform is absent in murine endothelium.

1.2.3 PECAM-1 as a therapeutic target

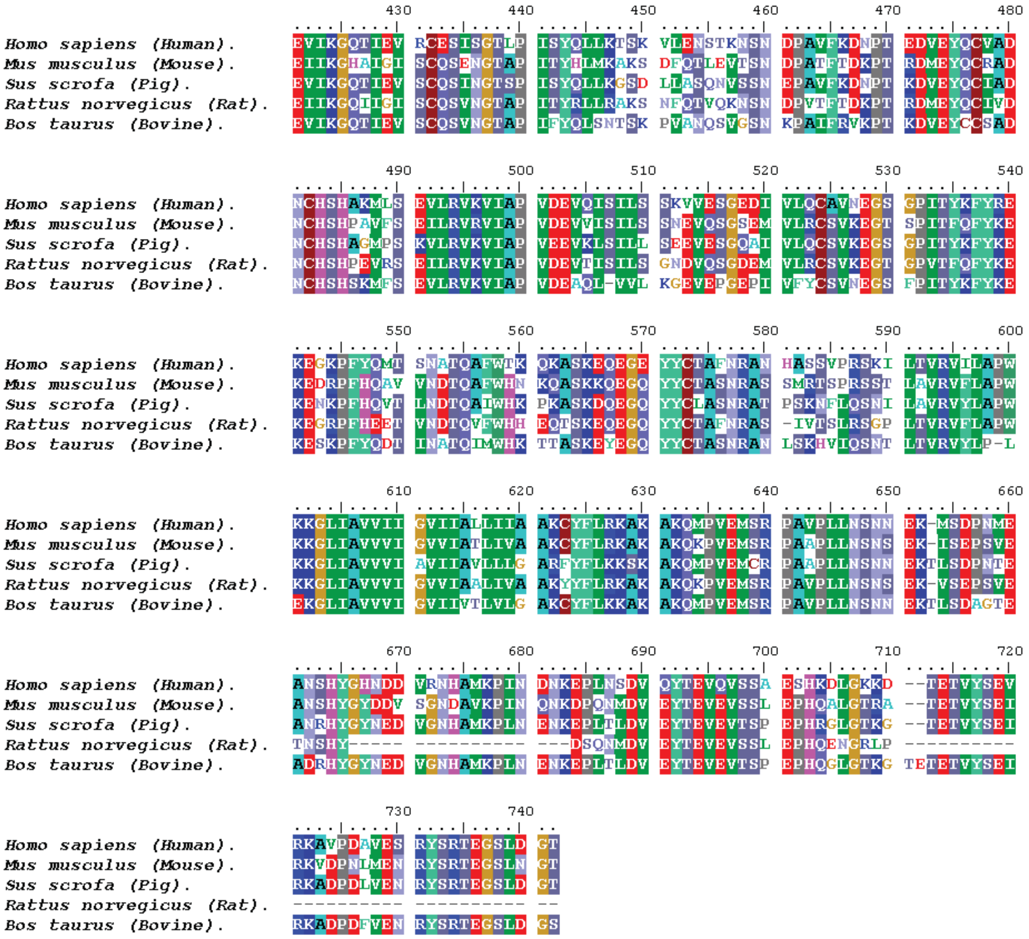
PECAM-1 has clinical importance in pathological disorders such as thrombosis, neuroinflammatory (Kalinowska & Losy 2006) and infectious diseases (Moseley & Jackson 2004; Newman 1994, 1999). It is also implicated in numerous functions, including angiogenesis (Cao *et al.* 2002) and neutrophil transmigration and T-cell activation (Zehnder *et al.* 1995). PECAM-1 governs endothelial cell (EC) migration during angiogenesis by binding to integrins (DeLisser *et al.* 1997). Initial studies using blocking anti-PECAM-1 antibodies inhibited cytokine and tumour induced angiogenesis and suggested that interactions of endothelial PECAM-1 are important in the formation of new vessels. This conclusion has been supported by the observation of reduced angiogenesis in mice deficient in PECAM-1 expression (O'Brien* *et al.* 2004).

PECAM-1 is a key participant in the adhesion cascade leading to extravasation of leukocytes during the inflammatory process. Leukocyte extravasation refers to the movement of leukocytes in post-capillary venules from the circulatory system into the interstitial fluid, towards the site of tissue damage or infection. The process of leukocyte extravasation (Figure 1.5) can be dissected into three distinct phases: rolling, adhesion and transmigration. These phases are mediated by the actions of

```
                                  10         20         30         40         50         60
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     MQPRWAQGAT MWLGVLLTLL LCSSLEGQEN SFTINSVDMK SLPDWTVQNG KNLTLQCFAD
Mus musculus (Mouse).     ---------- MLLALGLTLV LYASLQAEEN SFTINSIHME SLPSWEVMNG QQLTLECLVD
Sus scrofa (Pig).         MRLRWTQGGN MWLGVLLTLQ LCSSLEGQEN SFTINSIHME MLPGQEVHNG ENLTLQCIVD
Rattus norvegicus (Rat).  ---------- MLLALLLTML LYASLQAQEN SFTINSIHME SRPSWEVSNG QKLTLQCLVD
Bos taurus (Bovine).      MQLRWTQRGM MWLGALLTLL LCSSLKGQEN SFTINSIHMQ ILPHSTVQNG ENLTLQCLVD

                                  70         80         90        100        110        120
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     VSTTSHVKPQ HQMLFYKDDV LFYNISSMKS TESYFIPEVR IYDSGTYKCT VIVNNKEKTT
Mus musculus (Mouse).     ISTTSKSRSQ HRVLFYKDDA MVYNVTSREH TESYVIPQAR VFHSGKYKCT VMLNNKEKTT
Sus scrofa (Pig).         VSTTSSVKPQ HQVLFYKDDV LFHNVSSTKN TESYFISEAR VYNSGRYKCT VILNNKEKTT
Rattus norvegicus (Rat).  ISTTSKSRPQ HQVLFYKDDA LVYNVSSSEH TESFVIPQSR VFHAGKYKCT VILNSKEKTT
Bos taurus (Bovine).      VSTTSRVKPL HQVLFYKDDV LLHNVSSRRN TESYLIPHVR VCDSGRYKCN VILNNKEKTT

                                 130        140        150        160        170        180
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     AEYQLLVEGV PSPRVTLDKK EAIQGGIVRV NCSVPEEKAP IHFTIEKLEL NEKMVKLKRE
Mus musculus (Mouse).     IEYEVKVHGV SKPKVTLDKK EVTEGGVVTV NCSLQEEKPP IFFKIEKLEV GTKFVKRRID
Sus scrofa (Pig).         AEYKVVVEGV SNPRVTLDKK EVIEGGVVKV TCSVPEEKPP VHFIIEKFEL NVRDVKQRRE
Rattus norvegicus (Rat).  IEYQLTVNGV PMPEVTVDKK EVTEGGIVTV NCSMQEEKPP IYFKIEKVEL GTKNVKLSRE
Bos taurus (Bovine).      PEYEVWVKGV SDPRVTLDKK EVIEGGVVVV NCSVPEEKAP VHFTIEKFEL NIRGAKKKRE

                                 190        200        210        220        230        240
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     KNSRDQNFVI LEFPVEEQDR VLSFRCQARI ISGIHMQTSE STKSELVTVT ESFSTPKFHI
Mus musculus (Mouse).     KTS-NENFVL MEFPIEAQDH VLVFRCQAGI LSGFKLQESE PIRSEYVTVQ ESFSTPKFEI
Sus scrofa (Pig).         KTANNQNSVT LEFTVEEQDR VILFSCQANV IFGTRVEISD SVRSDLVTVR ESFSNPKFHI
Rattus norvegicus (Rat).  KTS-NMNFVL IEFPIEEQDH LLVFRCQAGV LSGIKMQTSE FIRSEYVTVQ EFFSTPKFQI
Bos taurus (Bovine).      KTSQNQNFVT LEFTVEEQDR TIRFQCQAKI FSGSNVESSR PIQSDLVTVR ESFSNPKFHI

                                 250        260        270        280        290        300
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     SPTGMIMEGA QLHIKCTIQV THLAQEFPEI IIQKDKAIVA HNRHGNKAVY SVMAMVEHSG
Mus musculus (Mouse).     KPPGMIIEGD QLHIRCIVQV THLVQEFTEI IIQKDKAIVA TSKQSSEAVY SVMAMVEYSG
Sus scrofa (Pig).         SPKGVIIEGD QLLIKCTIQV THQAQSFPEI IIQKDKEIVA HSRNGSEAVY SVMATVEHNS
Rattus norvegicus (Rat).  QPPEMIIEGN QLHIKCSVQV AHLAQEFPEI IIQKDKAIVA TSKQSKEAVY SVMALVEHSG
Bos taurus (Bovine).      IPEGKVMEGD DLQVKCTVQV THQAQSFPEI IIQKDREIVA HNSLSSEAVY SVMATTEHNG

                                 310        320        330        340        350        360
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     NYTCKVESSR ISKVSSIVVN ITELFSKPEL ESSFTHLDQG ERLNLSCSIP GAPP-ANFTI
Mus musculus (Mouse).     HYTCKVESNR ISKASSIMVN ITELFPKPKL EFSSSRLDQG ELLDLSCSVS GTPV-ANFTI
Sus scrofa (Pig).         NYTCKVEASR ISKVSSIMVN ITELFSRPKL KSSATRLDQG ESLRLWCSIP GAPPEANFTI
Rattus norvegicus (Rat).  HYTCKVESNR ISKASSILVN ITELFPRPKL ELSSSRLDQG EMLDLSCSVS GAPV-ANFTI
Bos taurus (Bovine).      NYTCKVEASR ISKVSSVVVN VTELFSKPKL ESSATHLDQG EDLNLLCSIP GAPP-ANFTI

                                 370        380        390        400        410        420
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     QKEDTIVSQT QDFTKIASKS DSGTYICTAG IDKVVKKSNT VQIVVCEMLS QPRISYDAQF
Mus musculus (Mouse).     QKEETVLSQY QNFSKIAEES DSGEYSCTAG IGKVVKRSGL VPIQVCEMLS KPSIFHDAKS
Sus scrofa (Pig).         QKGGMMMLQD QNLTKVASER DSGTYTCVAG IGKVVKRSNE VQIAVCEMLS KPSIFHDSGS
Rattus norvegicus (Rat).  QKEETVLSQY QNFSKIAEER DSGLYSCTAG IGKVVKRSNL VPVQVCEMLS KPRIFHDAKF
Bos taurus (Bovine).      QKGGMTVSQT QNFTKRVSEW DSGLYTCVAG VGRVFKRSNT VQITVCEMLS KPSIFHDSRS
```

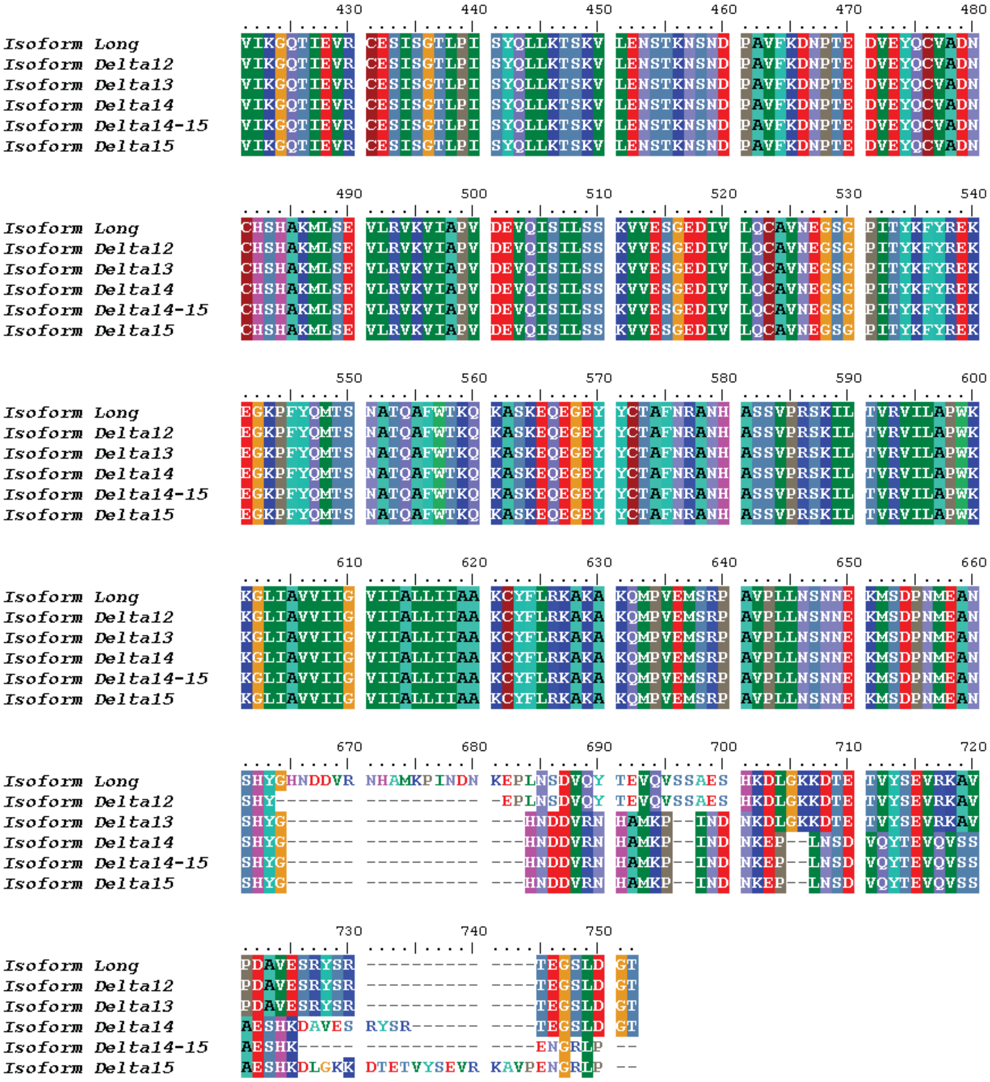
Continue...

Continued…



```
                                 430        440        450        460        470        480
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     EVIKGQTIEV RCESISGTLP ISYQLLKTSK VLENSTKNSN DPAVFKDNPT EDVEYQCVAD
Mus musculus (Mouse).     EIIKGHAIGI SCQSENGTAP ITYHLMKAKS DFQTLEVTSN DPATFTDKPT RDMEYQCRAD
Sus scrofa (Pig).         EVIKGQTIEV SCQSINGTSP ISYQLLKGSD LLASQNVSSN EPAVFKDNPT KDVEYQCIAD
Rattus norvegicus (Rat).  EIIKGQIIGI SCQSVNGTAP ITYRLLRAKS NFQTVQKNSN DPVTFTDKPT RDMEYQCIVD
Bos taurus (Bovine).      EVIKGQTIEV SCQSVNGTAP IFYQLSNTSK PVANQSVGSN KPAIFRVKPT KDVEYCCSAD

                                 490        500        510        520        530        540
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     NCHSHAKMLS EVLRVKVIAP VDEVQISILS SKVVESGEDI VLQCAVNEGS GPITYKFYRE
Mus musculus (Mouse).     NCHSHPAVFS EILRVRVIAP VDEVVISILS SNEVQSGSEM VLRCSVKEGT SPITFQFYKE
Sus scrofa (Pig).         NCHSHAGMPS KVLRVKVIAP VEEVKLSILL SEEVESGQAI VLQCSVKEGS GPITYKFYKE
Rattus norvegicus (Rat).  NCHSHPEVRS EILRVKVIAP VDEVTISILS GNDVQSGDEM VLRCSVKEGT GPVTFQFYKE
Bos taurus (Bovine).      NCHSHSKMFS EVLRVKVIAP VDEAQL-VVL KGEVEPGEPI VFYCSVNEGS FPITYKFYKE

                                 550        560        570        580        590        600
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     KEGKPFYQMT SNATQAFWTK QKASKEQEGE YYCTAFNRAN HASSVPRSKI LTVRVILAPW
Mus musculus (Mouse).     KEDRPFHQAV VNDTQAFWHN KQASKKQEGQ YYCTASNRAS SMRTSPRSST LAVRVFLAPW
Sus scrofa (Pig).         KENKPFHQVT LNDTQAIWHK PKASKDQEGQ YYCLASNRAT PSKNFLQSNI LAVRVYLAPW
Rattus norvegicus (Rat).  KEGRPFHEET VNDTQVFWHH EQTSKEQEGQ YYCTAFNRAS -IVTSLRSGP LTVRVFLAPW
Bos taurus (Bovine).      KESKPFYQDT INATQIMWHK TTASKEYEGQ YYCTASNRAN LSKHVIQSNT LTVRVYLP-L

                                 610        620        630        640        650        660
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     KKGLIAVVII GVIIALLIIA AKCYFLRKAK AKQMPVEMSR PAVPLLNSNN EK-MSDPNME
Mus musculus (Mouse).     KKGLIAVVVI GVVIATLIVA AKCYFLRKAK AKQKPVEMSR PAAPLLNSNS EK-ISEPSVE
Sus scrofa (Pig).         KKGLIAVVVI AVIIAVLLLG ARFYFLKKSK AKQMPVEMCR PAAPLLNSNN EKTLSDPNTE
Rattus norvegicus (Rat).  KKGLIAVVVI GVVIAALIVA AKYYFLRKAK AKQKPVEMSR PAVPLLNSNS EK-VSEPSVE
Bos taurus (Bovine).      EKGLIAVVVI GVIIVTLVLG AKCYFLKKAK AKQMPVEMSR PAVPLLNSNN EKTLSDAGTE

                                 670        680        690        700        710        720
                          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Homo sapiens (Human).     ANSHYGHNDD VRNHAMKPIN DNKEPLNSDV QYTEVQVSSA ESHKDLGKKD --TETVYSEV
Mus musculus (Mouse).     ANSHYGYDDV SGNDAVKPIN QNKDPQNMDV EYTEVEVSSL EPHQALGTRA --TETVYSEI
Sus scrofa (Pig).         ANRHYGYNED VGNHAMKPLN ENKEPLTLDV EYTEVEVTSP EPHRGLGTKG --TETVYSEI
Rattus norvegicus (Rat).  TNSHY----- ---------- ---DSQNMDV EYTEVEVSSL EPHQENGRLP ----------
Bos taurus (Bovine).      ADRHYGYNED VGNHAMKPLN ENKEPLTLDV EYTEVEVTSP EPHQGLGTKG TETETVYSEI

                                 730        740
                          ....|....| ....|....| ..
Homo sapiens (Human).     RKAVPDAVES RYSRTEGSLD GT
Mus musculus (Mouse).     RKVDPNLMEN RYSRTEGSLN GT
Sus scrofa (Pig).         RKADPDLVEN RYSRTEGSLD GT
Rattus norvegicus (Rat).  ---------- ---------- --
Bos taurus (Bovine).      RKADPDFVEN RYSRTEGSLD GS
```

Figure 1.3. *Multiple sequence alignment of PECAM-1 sequences from human, mouse, pig, rat and bovine using ClustalX. Residues are coloured according to their physicochemical properties.*



```
                                  430        440        450        460        470        480
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Isoform Long             VIKGQTIEVR CESISGTLPI SYQLLKTSKV LENSTKNSND PAVFKDNPTE DVEYQCVADN
Isoform Delta12          VIKGQTIEVR CESISGTLPI SYQLLKTSKV LENSTKNSND PAVFKDNPTE DVEYQCVADN
Isoform Delta13          VIKGQTIEVR CESISGTLPI SYQLLKTSKV LENSTKNSND PAVFKDNPTE DVEYQCVADN
Isoform Delta14          VIKGQTIEVR CESISGTLPI SYQLLKTSKV LENSTKNSND PAVFKDNPTE DVEYQCVADN
Isoform Delta14-15       VIKGQTIEVR CESISGTLPI SYQLLKTSKV LENSTKNSND PAVFKDNPTE DVEYQCVADN
Isoform Delta15          VIKGQTIEVR CESISGTLPI SYQLLKTSKV LENSTKNSND PAVFKDNPTE DVEYQCVADN

                                  490        500        510        520        530        540
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Isoform Long             CHSHAKMLSE VLRVKVIAPV DEVQISILSS KVVESGEDIV LQCAVNEGSG PITYKFYREK
Isoform Delta12          CHSHAKMLSE VLRVKVIAPV DEVQISILSS KVVESGEDIV LQCAVNEGSG PITYKFYREK
Isoform Delta13          CHSHAKMLSE VLRVKVIAPV DEVQISILSS KVVESGEDIV LQCAVNEGSG PITYKFYREK
Isoform Delta14          CHSHAKMLSE VLRVKVIAPV DEVQISILSS KVVESGEDIV LQCAVNEGSG PITYKFYREK
Isoform Delta14-15       CHSHAKMLSE VLRVKVIAPV DEVQISILSS KVVESGEDIV LQCAVNEGSG PITYKFYREK
Isoform Delta15          CHSHAKMLSE VLRVKVIAPV DEVQISILSS KVVESGEDIV LQCAVNEGSG PITYKFYREK

                                  550        560        570        580        590        600
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Isoform Long             EGKPFYQMTS NATQAFWTKQ KASKEQEGEY YCTAFNRANH ASSVPRSKIL TVRVILAPWK
Isoform Delta12          EGKPFYQMTS NATQAFWTKQ KASKEQEGEY YCTAFNRANH ASSVPRSKIL TVRVILAPWK
Isoform Delta13          EGKPFYQMTS NATQAFWTKQ KASKEQEGEY YCTAFNRANH ASSVPRSKIL TVRVILAPWK
Isoform Delta14          EGKPFYQMTS NATQAFWTKQ KASKEQEGEY YCTAFNRANH ASSVPRSKIL TVRVILAPWK
Isoform Delta14-15       EGKPFYQMTS NATQAFWTKQ KASKEQEGEY YCTAFNRANH ASSVPRSKIL TVRVILAPWK
Isoform Delta15          EGKPFYQMTS NATQAFWTKQ KASKEQEGEY YCTAFNRANH ASSVPRSKIL TVRVILAPWK

                                  610        620        630        640        650        660
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Isoform Long             KGLIAVVIIG VIIALLIIAA KCYFLRKAKA KQMPVEMSRP AVPLLNSNNE KMSDPNMEAN
Isoform Delta12          KGLIAVVIIG VIIALLIIAA KCYFLRKAKA KQMPVEMSRP AVPLLNSNNE KMSDPNMEAN
Isoform Delta13          KGLIAVVIIG VIIALLIIAA KCYFLRKAKA KQMPVEMSRP AVPLLNSNNE KMSDPNMEAN
Isoform Delta14          KGLIAVVIIG VIIALLIIAA KCYFLRKAKA KQMPVEMSRP AVPLLNSNNE KMSDPNMEAN
Isoform Delta14-15       KGLIAVVIIG VIIALLIIAA KCYFLRKAKA KQMPVEMSRP AVPLLNSNNE KMSDPNMEAN
Isoform Delta15          KGLIAVVIIG VIIALLIIAA KCYFLRKAKA KQMPVEMSRP AVPLLNSNNE KMSDPNMEAN

                                  670        680        690        700        710        720
                         ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
Isoform Long             SHYGHNDDVR NHAMKPINDN KEPLNSDVQY TEVQVSSAES HKDLGKKDTE TVYSEVRKAV
Isoform Delta12          SHY------- ---------- -EPLNSDVQY TEVQVSSAES HKDLGKKDTE TVYSEVRKAV
Isoform Delta13          SHYG------ ---------- ---HNDDVRN HAMKP--IND NKDLGKKDTE TVYSEVRKAV
Isoform Delta14          SHYG------ ---------- ---HNDDVRN HAMKP--IND NKEP--LNSD VQYTEVQVSS
Isoform Delta14-15       SHYG------ ---------- ---HNDDVRN HAMKP--IND NKEP--LNSD VQYTEVQVSS
Isoform Delta15          SHYG------ ---------- ---HNDDVRN HAMKP--IND NKEP--LNSD VQYTEVQVSS

                                  730        740        750
                         ....|....| ....|....| ....|....| ..
Isoform Long             PDAVESRYSR ---------- ----TEGSLD GT
Isoform Delta12          PDAVESRYSR ---------- ----TEGSLD GT
Isoform Delta13          PDAVESRYSR ---------- ----TEGSLD GT
Isoform Delta14          AESHKDAVES RYSR------ ----TEGSLD GT
Isoform Delta14-15       AESHK----- ---------- ----ENGRLP --
Isoform Delta15          AESHKDLGKK DTETVYSEVR KAVPENGRLP --
```

Figure 1.4. *Multiple sequence alignment of alternatively spliced isoforms of human PECAM-1 using ClustalX. Isoforms of PECAM-1 have difference in their C-terminal regions due to the presence or absence of cytoplasmic domain exons. Residues in the multiple sequence alignment are colored according to their physicochemical properties.*

three distinct classes of adhesion molecules: selectins, integrins and IgSF members (Johnson-Leger 2000). Upon encountering an inflammatory stimulus, leukocytes move rapidly into the blood, come into contact with the endothelial surface and commence rolling along the endothelium. Leukocyte rolling is mediated predominantly by selectins and α_4-integrins. Rolling leukocytes are able to detect chemoattractant molecules such as IL-1, TNFα and chemokines on the endothelial surface, inducing cell and integrin activation. Activated β_2-integrins such as LFA-1 and Mac-1 interact with ICAM-1 to mediate the firm adhesion of cells to the endothelium. Once adhered, leukocytes emigrate into the extravascular tissue through interendothelial junctions.

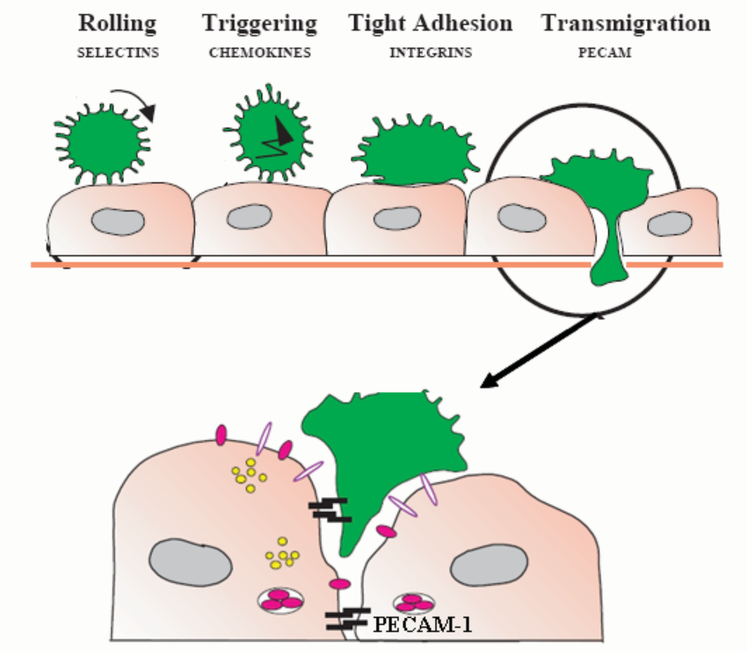


Figure 1.5. *Multi-step cascade of leukocyte extravasation. The interendothelial junction is closed due to the homophilic binding of adhesion molecules such as PECAM-1 and JAM. During the multi-step adhesion cascade, a leukocyte (green in color) emigrates into the extravascular tissue resulting in loosening of the junctions. PECAM-1 (black bars) is localized on the intercellular junctions of endothelial cells directing cells to migrate through the vessel wall via homophilic interactions on the leukocyte surface. The figure was extracted from Johnson-Leger et al. (2000).*

PECAM-1 is localised on the intercellular junctions of endothelial cells and directs cells to migrate through the vessel wall through homophilic interactions on the leukocyte surface (Figure 1.5). Once in the interstitial fluid, leukocytes migrate along a chemotactic gradient towards the site of injury or infection. Blocking antibodies which recognise PECAM-1 can block diapedesis (migration of leukocytes across the endothelium upon inflammatory stimulus), with leukocytes being arrested on the apical surface of the endothelial cell border.

PECAM-1 is a target glycoprotein in drug-induced immune thrombocytopenia (Kroll *et al.* 2000). The mechanism behind drug-induced thrombocytopenia is a decrease in platelet production caused by drug-dependent antibodies. Targets for drug-dependent antibodies are glycoproteins on the cell membrane of the platelets, such as glycoprotein (GP) Ib/IX and GPIIb/IIIa, which contain sites for platelet antigens. The drug classes that are most often associated with drug-induced immune thrombocytopenia are quinine, quinidine, sulfonamides, NSAIDs, anticonvulsants, disease modifying antirheumatic drugs and diuretics. Antibodies from patients with acute mild thrombocytopenia following treatment with anti-thyroid drug carbimazole were found to be specific for PECAM-1 (Kroll *et al.* 2000). In the same study, antibodies from patients with quinidine-induced thrombocytopenia reacted very weakly with PECAM-1.

1.2.4 Role of PECAM-1 in signalling

PECAM-1 has been shown to serve as a scaffolding molecule in a number of signalling pathways. Its cytoplasmic tail becomes phosphorylated on serine and tyrosine residues following cellular activation, creating binding sites for SHP-2 (Src homology 2-containing tyrosine phosphatases) and perhaps other cytosolic signaling molecules (D. E. Jackson 2003) as shown in Figure 1.6. As a scaffold for SHP-2, PECAM-1 modulates its recruitment and activation in a number of cellular pathways involving catenins, STATs and PI3 kinase. The intracellular interaction of PECAM-1 with molecules of MAP kinase cascades occur independently of the ITIM motifs, whereas interaction of PECAM-1 with calmodulin is dependent on the sequence “599-RKAKAK-604” in the PECAM-1 intracellular region.

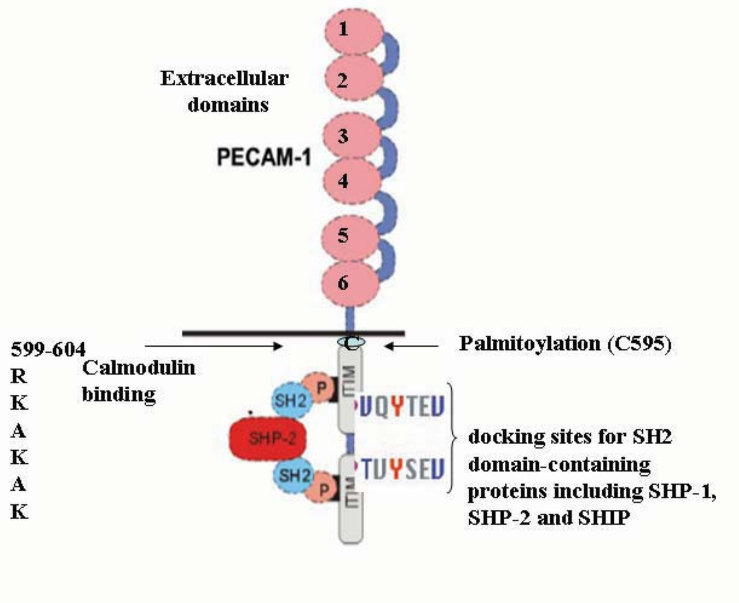


Figure 1.6. *The intracellular processes activated by PECAM-1 are illustrated. Phosphorylation of tyrosines Y663/Y686 in the ITIM regions encoded in the cytoplasmic region occurs following ligation of PECAM-1 or activation of other receptors which may be dependent or independent of the ITIM motifs. The cysteine palmitoylation site at amino acid position 595 is also shown.*

1.2.5 Role of PECAM-1 in homophilc and heterophilic interactions

The interactions of PECAM-1 with its ligands are complex. PECAM-1 exists in equilibrium between monomeric and dimeric forms (Sun 2000). It engages in both homophilic (antiparallel inter-digitation of opposing PECAM-1 molecules on adjacent cells) (Holness & Simmons 1994) and heterophilic binding (to other ligands), depending on the conditions of the interaction and the ligands available (Newman 1997). PECAM-1-PECAM-1 interactions can be both *cis* (interactions between adjacent PECAM-1 molecules in the same membrane) and *trans* associations (interactions between two PECAM-1 molecules in distinct membranes) (Newton *et al.* 1997; Zhao & Newman 2001). PECAM-1 trans-homophilic interactions require Ig-domain 1 (Nakada *et al.* 2000), with residues Asp 11, Asp 33, Lys 50, Asp 51 and Lys 89 having been implicated in this binding (Newton *et al.* 1997; Sun *et al.* 2004; Sun *et al.* 1996). Human-mouse chimeric studies have suggested that these homophilic interactions are species specific (Sun *et al.* 2004). Moreover, antibodies that recognise the epitope “CAVNEG” in Ig-domain 6 have been shown to enhance homophilic adhesion (Yan *et al.* 1995).

A number of heterophilic ligands of PECAM-1 have been proposed to interact with regions encompassing Ig-domains 1-3. These include integrin $\alpha_v\beta_3$, CD38 and heparan sulfate proteoglycans (HSPGs) (Deaglio *et al.* 1998; DeLisser *et al.* 1993; Piali 1995; Prager 1996). The heterophilic interactions between PECAM-1 and $\alpha_v\beta_3$ may be the result of direct binding of PECAM-1 to another molecule or the result of secondary interactions mediated by non-PECAM-1 molecules whose activation is PECAM-1-dependent. The likelihood that PECAM-1 binds directly to $\alpha_v\beta_3$ has been questioned (Brugge *et al.* 2000). PECAM-1 signaling is reported to activate integrin $\alpha_v\beta_3$ on endothelial cells, and this interaction in turn induces β_1 integrin-mediated firm adhesion of eosinophils to endothelial cells (Chiba *et al.* 1999), thus suggesting that PECAM-1 does not bind directly to $\alpha_v\beta_3$. However, other data provide evidence for a direct interaction (Piali 1995).

PECAM-1 has been reported to have high affinity binding sites for Mn^{2+} cations, involving acidic residues in region 436-448 of Ig-domain 5 and a cluster of acidic residues in regions 485-495 and 534-549 in Ig-domain 6 (D. E. Jackson *et al.* 1997). The heterophilic interaction with CD177 involves the cation binding site of Ig-domain 6 (Sachs *et al.* 2007). Monoclonal antibodies against CD177 and Ig-domain 6 of PECAM-1 inhibited adhesion of cells expressing CD177 to immobilised PECAM-1 in the presence of cations. These monoclonal antibodies also inhibited the transendothelial migration of human neutrophils, indicating a role for this divalent cation-dependent heterophilic association in the neutrophil transmigration pathway.

A number of studies have suggested that cell surface GAGs act as ligands for PECAM-1 (DeLisser *et al.* 1993) by binding to a GAG consensus binding sequence (L-K-R-E-K-N) in Ig-domain 2 (DeLisser *et al.* 1993). Interestingly, a similar GAG recognition sequence (residues 131-148) is located in loop 2 of Ig-domain 2 of NCAM. NCAM is known as a heparin and heparan sulphate (HS) binding protein (Albelda 1991; DeLisser *et al.* 1993). The crystal structures of NCAM Ig-domains forming zipper adhesion complexes (a dimerisation pattern where at least two monomers are intertwined) reveal that heparin and HS binding sites are solvent exposed, suggesting that the association of NCAM with heterophilic and homophilic ligands could occur simultaneously (Kasper *et al.* 2000; Soroka *et al.* 2003). However, whether this also occurs with PECAM-1 is not known. Experimental data has suggested that PECAM-1 could mediate L-cell aggregation by binding in a heterophilic fashion to HS on cell surfaces. Cell aggregation was blocked by iduronic acid-containing GAGs, including heparin, HS and dermatan sulfate (DS) (weakly effective), but not by hyaluronic acid (HA) or the chondroitin sulfates (CS), thus suggesting that PECAM-1 may bind iduronic acid containing GAGs (DeLisser *et al.* 1993; Watt *et al.* 1993).

Some research groups argued that cell surface GAGs are not ligands for PECAM-1 (Sun *et al.* 1998) and that heparin affects PECAM-1 adhesion by indirect mechanisms that are downstream of the interactions of PECAM-1 with its ligands. The Molecular Immunology Group at Curtin University, headed by Assoc. Prof. Deirdre Coombe, have reinvestigated the binding of heparin/HS to PECAM-1 and report that the extracellular domains of PECAM-1 can bind heparin and HS in biochemical experiments, and that HS can also bind PECAM-1 on cell surfaces. Binding is pH sensitive, as stable binding occurs at a slightly acidic pH, which would allow the protonation of histidines. Moreover, domain deletion experiments revealed that the heparin/HS binding regions of PECAM-1 are distinct from those involved in homophilic binding and that these domains contain high and low affinity binding sites for heparin. The molecular modelling studies reported in this thesis were undertaken to provide a rationalisation at the molecular level of these experimental studies by seeking to determine whether PECAM-1 is indeed a GAG binding molecule.

1.3. BASIC FEATURES AND FUNCTIONS OF GAGs

GAGs are polyanionic molecules that bind to a wide range of proteins involved in physiological and pathological processes (R. L. Jackson *et al.* 1991). GAGs are sometimes known as mucopolysaccharides because of their viscous, lubricating properties, as found in mucous secretions. These molecules are present on all animal cell surfaces in the extracellular matrix (ECM), and some are known to bind and regulate a number of distinct proteins, including chemokines, cytokines, growth factors, morphogens, enzymes and adhesion molecules (Lindahl & Kjellen 1991). The key properties of GAGs are summarised in Table 1.2.

Table 1.2. Key properties of glycosaminoglycans.

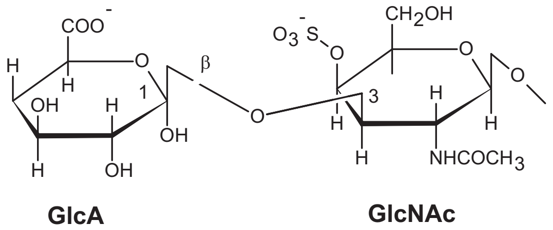
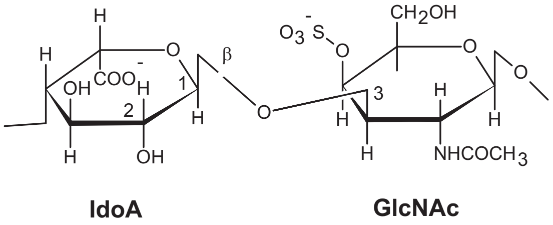
***Physico-chemical properties of GAGs*:** Negatively charged, viscous, lubricating, Unbranched polysaccharides, Repeating disaccharide units, bind large amounts of water, low compressibility.
Classification of GAGs: Chondrotin sulfates, Keratan sulfate, Dermatan sulfate, Hyaluron, Heparin and Heparan sulfate.
Function of GAGs: Cell adhesion, cell growth and differentiation, cell signalling.

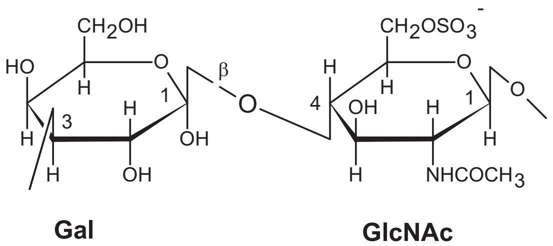
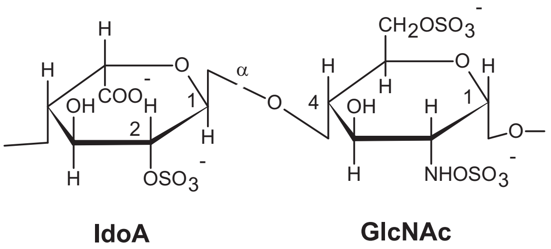
1.3.1 Classification of GAGs

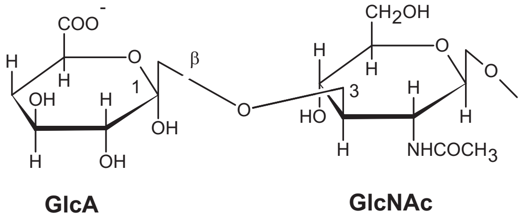
There are two main types of GAGs: non-sulfated (hyaluronic acid) and sulfated (chondroitin sulfate (CS), dermatan sulfate (DS), heparin and HS). The highly sulfated analogues heparin and HS have been studied extensively (Capila & Linhardt 2002) due to their well understood functions in anti-coagulation.

These linear, sulfated polysaccharides have molecular weights of roughly 10 ~ 100 kDa. GAG chains are composed of disaccharide repeating units called disaccharide repeating regions (see Table 1.3). The repeating units are composed of uronic acid (D-glucoronic acid or L-iduronic acid) and amino sugar (D-galactosamine or D-glucosamine). The amino sugar may be sulfated on carbon 4 or 6 or on non-acetylated nitrogen. At physiological pH, the carboxylate groups in the acidic sugars and the sulfate groups are deprotonated and hence negatively charged. CS and DS that contain galactosamine are called galactosaminoglycans, whereas heparin and HS that contain glucosamine are called glucosaminoglycans. The sugar backbone can be sulfated at various positions. As a result, a simple octasaccharide can have over 1,000,000 different sulfation sequences (Sasisekharan & Venkataraman 2000). GAGs also vary in the geometry of the glycosidic linkage (α or β).

Table 1.3. Repeating disaccharide units of various glycosaminoglycans.

Glycosaminoglycan	Disaccharide units	Features
Chondroitin sulfates	 COO- O β 1 OH H O3- S O CH2OH 3 NHCOCH3 GlcA GlcNAc * The figure contains GalNAc 4-sulfate.	• Disaccharide unit: N-acetylgalatosamine (GalNAc) with sulfate (SO_3^-) on either C-4 or C-6 and glucoronic acid (GlcA). • Glycosidic linkage: beta (1, 3). • Molecular weight 5-50 kDa. • Most abundant GAG in the body. • Found in cartilage, bone, heart valves.
Dermatan sulfate	 H COO- OH O β 1 2 OH O3- S O CH2OH 3 NHCOCH3 IdoA GlcNAc * IdoA may be sulfated on C-2 position. In the figure, no sulfation is shown on IdoA.	• Disaccharide unit: N-acetylgalatosamine (GalNAc) and L-iduronic acid (IdoA) with variable amounts of glucuronic acid. • Glycosidic linkage: beta (1, 3). • Molecular weight 15-40 kDa. • Localised in skin, blood vessels, heart

		valves
Keratan sulfates I and II	 CH2OH CH2OSO3- HO O β 1 3 4 OH H NHCOCH3 Gal GlcNAc Gal may or maynot be sulfated at position 6. In the figure, no sulfation is shown on C-6 of Gal.	• Disaccharide unit: N-acetylglucosamine (GlcNAc) and galactose (Gal). Sulfate (SO_3^-) content is variable and may be present on C-6 of either sugar. • No uronic acid. • Glycosidic linkage: beta (1, 4). • Molecular weight 4-19 kDa. • Most heterogeneous GAG. • KS I is localized in the cornea. • KS II is found in cartilage aggregated with chondroitin sulfates.
Heparin/Heparan sulfate	 H O COO- OH α 1 2 4 OSO3- CH2OSO3- NHOSO3- IdoA GlcNAc In the figure, IdoA is sulfated at C-2 whereas GlcNAc have sulfation on C-2 and C-6.	• Disaccharide unit: N-acetylgucosamine (GlcNAc) and L-iduronic acid (IdoA) or glucuronic acid (GlcA). Most

		glucosamine residues are bound in sulfamide linkages. Sulfate (SO_3^-) is found on C-3 or C-6 of glucosamine and C-2 of uronic acid. • Glycosidic linkage: alpha (1, 4).
Hyaluronic acid	 COO H H O OH H 1 β H OH OH O CH2OH H O H 3 H O HO H H NHCOCH3 GlcA GlcNAc	• Disaccharide unit: N-acetylglucosamine (GlcNAc) and glucoronic acid (GlcA). • Glycosidic linkage: beta (1, 3). • Molecular weight 4-8000 kDa. • Non-sulfated, not covalently attached to the protein in the ECM, also found in bacteria. • Usually localised in synovial fluid, vitreous humor, ECM of loose connective tissue. • Excellent lubricators and shock absorbers.

1.3.2 Clinical significance of GAGs

GAGs play a major role in cell signalling and development, angiogenesis (Iozzo & San Antonio 2001), axonal growth (Holt & Dickson 2005), tumour progression, metastasis and anti-coagulation (Casu *et al.* 2004; Fareed *et al.* 2000). Anti-coagulation was the first described function for a GAG. Heparin was first discovered in 1917 because of its capacity to prolong the process of blood clotting, an effect due to its potentiating interaction with the natural inhibitor of thrombin, antithrombin III (AT-III), with only about one third of all heparin chains processing the structures required for AT-III binding. Heparin is mainly used in pharmaceutical products as an anti-coagulant for the treatment of thrombosis, phlebitis and embolism. Pharmaceutical heparin is usually derived from bovine lung or porcine intestinal mucosa. It was originally isolated from canine liver cells, hence its name (from the Greek *hepar* for liver). It has different molecular weights due to variations in chain length and is structurally heterogeneous.

More recently, it has become clear that, in addition to anti-coagulation, other roles can be attributed to various GAGs. In cancer, progenitor cell proliferation is no longer restricted, leading to malignant transformation (Sasisekharan *et al.* 2002; Yip *et al.* 2006). GAGs and proteoglycans (PGs) are believed to play a very important role in cell proliferation because they serve as co-receptors for growth factors of the FGF (Fibroblast Growth Factor) family. Indeed, members of the FGF family need to interact with both a GAG chain and their high affinity receptor to realise their full signalling potential. Overexpression of these molecules could contribute to tumour progression.

Sulfated GAGs are a common constituent in many different types of amyloid. They play an important role in the pathology of amyloid diseases such as Amyloid A-amyloidosis, Alzheimer’s disease, type-2 diabetes, Parkinson’s disease and prion diseases (Kisilevsky *et al.* 2007). These diseases are characterised by deposition in tissues of fibrillar aggregates of polypeptides. HS is known to bind amyloidogenic peptides *in vitro* and *in vivo,* and this binding promotes fibril formation and enhances the disease condition. Sometimes HS is present within the amyloid β-containing amyloid deposits in Alzheimer’s diseased brains (Snow *et al.* 1987).

Diseases such as rheumatoid arthritis, inflammatory bowel disease and microbial infections are associated with inflammatory responses. Many proteins play a role in the inflammation cascade that leads to the activation of leukocytes and endothelial cells, and ultimately to the extravasation of leukocytes and leukocyte migration into the inflamed or diseased tissue. GAGs such as heparin have important roles in these processes, as adhesion ligands in leukocyte extravasation and carriers/presenters of chemokines and growth factors.

GAGs are also known to promote microbial pathogenesis and invasion (Fry *et al.* 1999; Rostand & Esko 1997) by interacting with several microbial pathogens on cell surfaces and in the ECM. Many pathogenic micro-organisms such as bacteria (e.g., *Helicobacter pylori*, *Bordetella pertussis*, *Mycobacterium tuberculosis* and *Chlamydia trachomatis*), viruses (e.g., herpes simplex), and protozoa (e.g., *Plasmodium* and *Leishmania*) express proteins capable of binding to HS, DS and CS on cell surfaces, and these interactions appear to mediate infection (Rostand & Esko 1997). Dengue and foot-and-mouth viruses interact with cell surface HSPGs (heparan sulfate proteoglycans) and promote the concentration of virus particles at the cell surface after subsequent binding to integrin receptors. Heparin is known to exert its anti-HIV-1 activity by binding to the viral surface glycoprotein, gp120 (Rider 1997), thus blocking HIV-1 entry into cells.

1.4. PROTEOGLYCANS

In nature, all GAG chains with the exception of HA are covalently linked to a core protein (Figure 1.7) to give a PG. The linkage of GAGs to the protein core involves a specific trisaccharide composed of two galactose (Gal) residues and a xylose (Xyl) residue (GAG-GalGalXyl-O-CH_2-protein). These saccharide residues are coupled to the protein core through an *O*-glycosidic bond to a serine residue in the protein. Some forms of keratan sulfates are linked to the protein core through an *N*-asparaginyl bond.

Virtually all mammalian cells produce PGs and either secrete them into the ECM, insert them into the plasma membrane, or store them in secretory granules. A number of PGs have been characterised and named according to their structure and functional location. Examples of large PGs are aggrecan, the major PG in cartilage, and versican, which is present in many adult tissues including blood vessels and skin. A variety of core proteins have been shown to carry HS chains in the ECM and at the cell surface. Some membrane PGs such as sydecan-1 (Figure 1.8) are hybrid structures known to contain both HS and CS (Kokenyesi & Bernfield 1994).

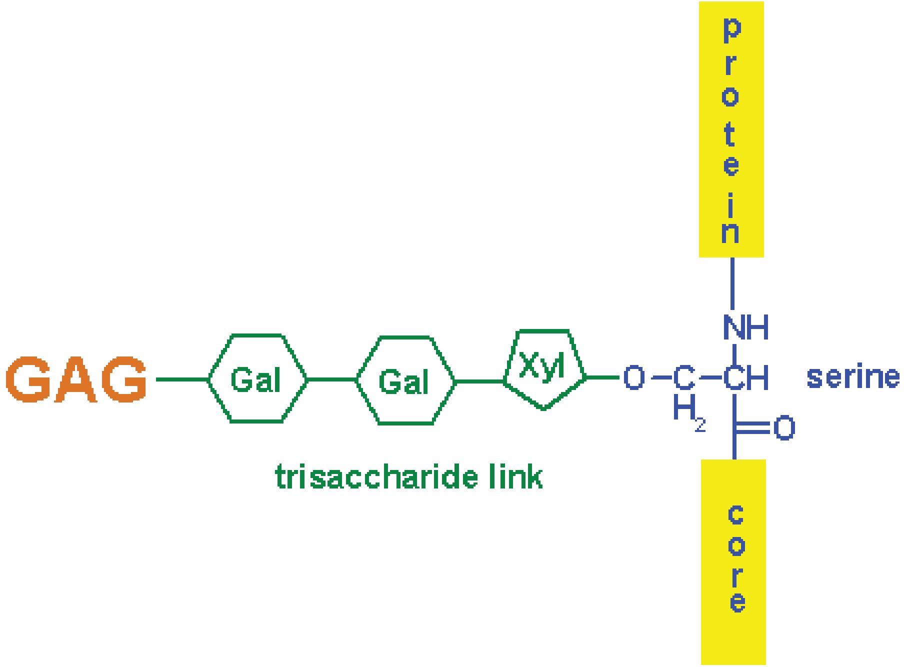


Figure 1.7. *Structure of the GAG linkage to proteins in proteoglycans.*

PGs exhibit tremendous structural variation due to a number of factors. Different numbers of GAG chains having different saccharide sequences can be attached to the various serine residues present in the core protein. There are two major types of HSPGs (Bernfield *et al.* 1999): the syndecans (Figure 1.8) and the glypicans. The core protein of each family differs: the syndecans are composed of an integral membrane protein whereas the glypicans have a GPI-anchored protein as their core protein. An HSPG with a core protein of a completely different structure is formed in ECM (Iozzo *et al.* 1994). Thus, HSPGs are formed both on the cell surface and in extracellular matrices.

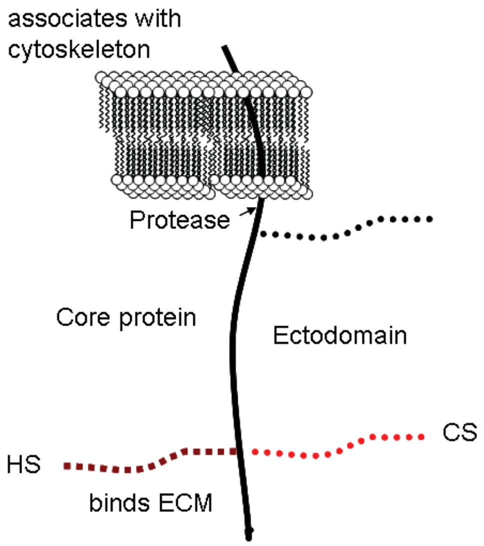


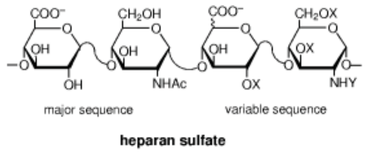
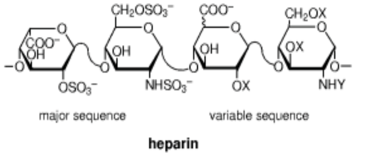
Figure 1.8. *Schematic representation of cell surface proteoglycans. Syndecans are transmembrane proteins that bear HS and CS chains distal from the plasma membrane.*

1.5 SULFATED GAGs: HEPARIN/HEPARAN SULFATE

Heparin and HS are highly evolutionarily conserved in a broad range of organisms belonging to many different kingdoms (Nader *et al.* 1999). The difference between HS and heparin is quantitative and not qualitative, as can be seen from Table 1.4. HS contains higher acetylated glucosamine and is less sulfated than heparin (Lindahl & Kjellen 1991). Heparin is synthesised by and stored exclusively in mast cells, whereas HS is expressed, as part of a PG, on cell surfaces and in ECM (Varki 1999). Heparin has the highest negative charge density of any known biological macromolecule because of its high content of negatively charged sulfate and carboxylate groups.

Heparin consists of repeating units of 1→4 linked pyranosyluronic acid and 2-amino-2-deoxyglucopyranose (glucosamine) residues. The uronic acid residues typically consist of 90 % L-idopyranosyluronic acid (L-iduronic acid) and 10 % D-glucopyranosyluronic acid (D-glucuronic acid). The amino group of the glucosamine residue may be substituted with an acetyl or sulfate group, or remain unsubstituted. The 3- and 6-positions of the glucosamine residues can either be substituted with an *O*-sulfate group or remain unsubstituted. The uronic acid, which can either be L-iduronic or D-glucuronic acid, may also contain a 2-*O*-sulfate group. HS is structurally related to heparin but is much less substituted with sulfate groups than heparin. Like heparin, HS is a repeating linear copolymer of an uronic acid 1⟶4 linked to glucosamine. D-glucuronic acid predominates in HS, but HS can also contain substantial amounts of L-iduronic acid. HS generally contains about one sulfate group per disaccharide, but it may have higher sulphate contents (Varki 1999). On the cell surface, the ester and amide sulfate groups are deprotonated in HS and attract positively charged counter ions to form a salt under physiological conditions.

Table 1.4. Key differences between heparan sulfate and heparin.

Property	Heparan sulfate	Heparin
Sulfate *versus* hexosamine content	0.8-1.8	1.8-2.4
2-deoxy-2-sulfamido-α-D-glucopyranosyl content	40-60%	>85%
α-L-iduronic acid content	30-50%	>70%
Site of synthesis	Extracellular component found in basement membrane and as a ubiquitous component of cell surfaces.	Intracellular component of mast cells especially in liver, lungs and skin.
Mass	10-70 kDa	10-12 kDa
Major and minor disaccharide repeating units (X=H or SO_3^-, Y=Ac, SO_3^-, or H).	 COO⁻ CH₂OH COO⁻ CH₂OX OH OH OH OX OH NHAc OX NHY major sequence variable sequence heparan sulfate	 CH₂OSO₃⁻ COO⁻ CH₂OX COO⁻ OH OH OH OX OSO₃⁻ NHSO₃⁻ OX NHY major sequence variable sequence heparin

HS chains also often contain domains of extended sequences having low sulfation compared to heparin, as illustrated in Figure 1.9. The non-sulfated regions that have a GlcA-GlcNAc (Acetylated glucosamine) sequence are the most common in the HS chain, with IdoA-containing sulfated regions (called S-domains) usually of about 5–10 disaccharides (Lyon & Gallagher 1998). There are also relatively minor proportions of mixed sequences, which contain both $GlcNSO_3$ and GlcNAc (called NA-domain). A substantial proportion of the HS chain may consist of alternating GlcA-GlcNAc residues with no sulfate substitution.

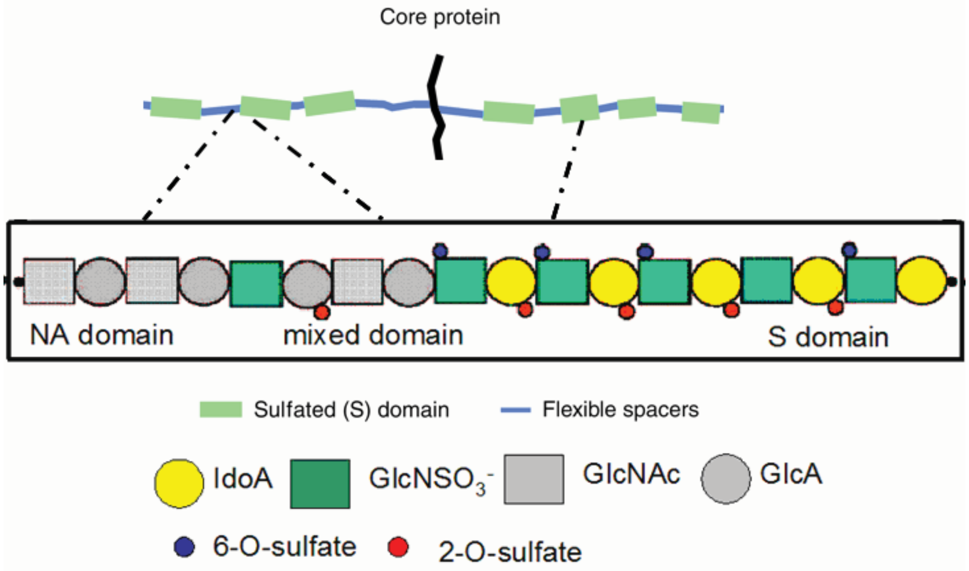


Figure 1.9. *Multidomain structure of HS. The distributed sulfated domains of HS are separated by flexible spacers of low sulfation. The mixed sequences define transition zones between the S-domains and the unmodified N-acetyl–rich regions. Several monomeric or oligomeric proteins can bind to GAGs, often by recognising different structural features of the domain.*

The sulfated–acetylated–sulfated domain has been subsequently found to be recognised by a number of chemokines such as interleukin-8 (IL-8) (Spillmann *et al.* 1998), platelet factor 4 (PF4) (Stringer & Gallagher 1997) and macrophage inflammatory protein 1 alpha (MIP-1α) (Stringer *et al.* 2002). The IL-8 dimer consists of two α-helical monomers lying on top of two β sheets forming basic clusters on one face of the dimer. The two S-domains, each consisting of 5-6 saccharides in HS, accelerate the rate of dimer formation in IL-8. The flexibility in the N-acetyl–rich "spacer" or NA domain (6-7 saccharides) in HS allows more conformational freedom for simultaneous interactions of two S-domains and brings the monomers of IL-8 in close proximity to form dimers in an anti-parallel arrangement (Figure 1.10). On the other hand, interferon-gamma (IFNγ) does not significantly interact with isolated S-domains (Lortat-Jacob *et al.* 1995), in contrast to many other heparin binding proteins. Similarly, basic residues are clustered on both faces of the tetramer of PF4, requiring 21 saccharides in the HS to form a more extended binding site on the charged surface of PF4. Heparin is assumed to be an analogue of the S-domains of HS, consisting mainly of sequences of sulfated disaccharides with IdoA2S (Iduronic acid sulfated at C-2) and GlcNS6S (2,6-disulfoglucosamine).

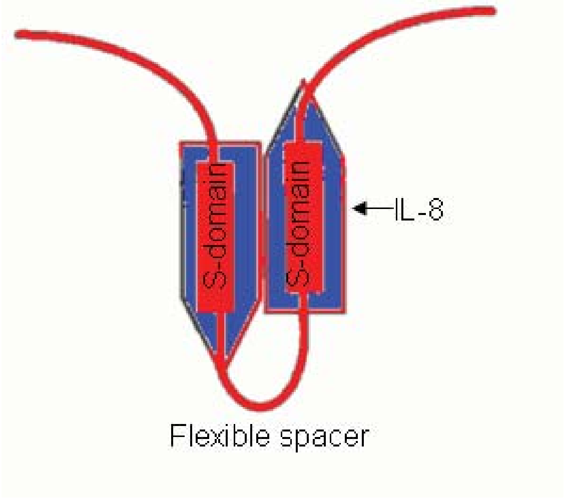


Figure 1.10. *Schematic representation of the dimerisation of IL-8 by HS. HS is colored red and the IL-8 monomer in blue. The rate of dimer of formation of IL-8 is accelerated by the S-domains whereas the flexible spacer sequence (NA-domain) allows appropriate folding of the monomer in an antiparallel arrangement.*

Heparin and HS GAGs can often be structurally distinguished through their sensitivity towards microbial GAG degrading enzymes, the heparin lyases. Three major polysaccharide lyases heparin lyases I, II, and III, isolated from *Flavobacterium heparinum*, are capable of cleaving linkages present in heparin and HS chains (Lohse & Linhardt 1992). These three enzymes share very little homology at the DNA, protein or even structural level, which imparts specificity towards the substrates. Heparin lyase I is involved in heparin binding whereas heparin lyase III exhibits a strong specificity for heparan sulfate. Heparin lyase II is believed to act on heparin and as well as on HS through two distinct active sites.

1.6 BIOSYNTHESIS OF HEPARAN SULFATE

The structural heterogeneity of HS with respect to the size of the polysaccharide chain, the ratio of IdoA to GlcA units, and the amount and distribution of sulfate groups along the carbohydrate backbone is the result of variations in the biosynthesis of HSPGs. The fine structure of the chains depends on the regulated expression and action of multiple biosynthetic enzymes, such as glycosyltransferases, sulfotransferases and an epimerase, which are arrayed in the lumen of the Golgi apparatus i.e. the enzymic reactions do not go to completion, yielding individual chains whose sequence is likely to be distinct from all other chains (Lindahl *et al.* 1998). The biosynthesis of heparin/HS (Figure 1.11) can be conveniently separated into three steps:

a. Formation of the linkage region to the core protein. Polysaccharide formation is initiated by the transfer of a xylose (Xyl) unit from UDP-Xyl by xylotransferase (XT) to a serine residue in the core protein. Two galactose units are then transferred by galactosyltransferases I and II (GalTI and GalTII) from corresponding UDP nucleotides to the xylosylated core protein. Amino acid sequences flanking the linking serine residue and/or the overall structure of the core protein seem to act as signals for directing the assembly of heparin/HS chains or CS or DS. Xylose attachment to the core protein is thought to occur in the endoplasmic reticulum (ER), with further assembly of the linkage region and the remainder of the chain occurring in the Golgi apparatus (Champe & Harvey 2005).

b. Chain elongation or polymerisation. The non-reducing end of the neutral trisaccharide xylosyl-galactosyl-galactose becomes the primer for the elongation of the polysaccharide. In the case of HS and heparin, polymer formation occurs through an alternating transfer of GlcA and GlcNAc units to the growing chain. This is carried out by one or more related enzymes whose genes are members of the exostoses (EXT) gene family of tumour suppressor genes. The mechanisms that control the length of the fully grown polysaccharides have not been fully elucidated. In general, the length of the final chain increases with the availability of the UDP-sugar precursor and decreases with the availability of the core protein.

c. Chain modifications. Subsequent to polymer formation, the repeating GlcA-GlcNAc disaccharide chain undergoes a number of enzymatic modifications that occur in a specific order, carried out by four classes of sulfotransferases and an epimerase. The sulfate group, which is crucial to the activity of the sulfotransferases, is made available by PAPS (3'-phosphoadenosyl-5'-

phosphosulfate), an AMP molecule with a sulfate group attached to its 5'-phosphate). The various enzymatic chain modifications are:

1) N-deacetylation of GlcNAc units originates GlcN residues and the N-sulfation of newly formed GlcN residues. N-deacetylation/N-sulfation is carried out by one or more members of a family of four GlcNAc N-deacetylase/N-sulfotransferase enzymes (NDSTs).

2) C5 epimerization of GlcA residues, which leads to the formation of IdoA units. Epimerisation is catalysed by either GlcA C5 epimerase or heparosan-N-sulfate-glucuronate 5-epimerase.

3) 2-*O*-sulfation of newly originated IdoA residues is catalysed by uronosyl-2-*O*-sulphotransferase (2OST).

4) *6-O*-sulfation of GlcN residues. Three glucosaminyl 6-*O*-transferases (6OSTs) enzymes have been identified that act on the formation of GlcNS6S adjacent to sulfated or non-sulfated IdoA.

5) Sulfation can also occur at C3 of GlcN units in the presence of at least five glucosaminyl 3-*O*-sulfotransferases (3OSTs) and to a limited extent, at C2 or C3 of GlcA units.

The regulation of the chain modification process leads to cell- or organ-specific HS structures that may allow fine modulation of their biological functions and specific binding with macromolecules such as growth factors, enzymes, ECM proteins and the cell surface proteins of pathogens. The enzymes described above also synthesise heparin.

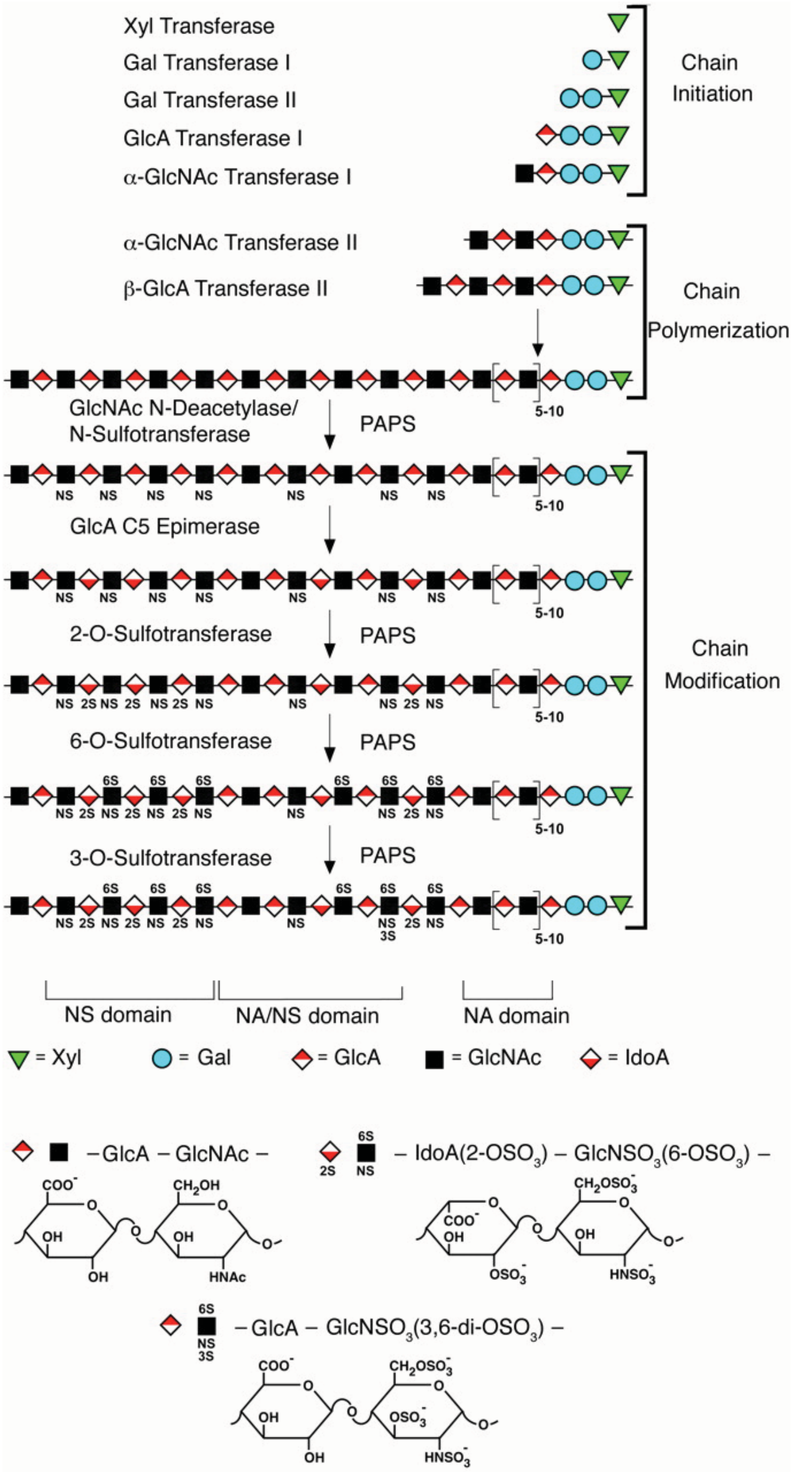


Figure 1.11. *Heparan sulfate chain biosynthesis. This figure was adapted from the reference by (Esko & Lindahl 2001). The symbols used are defined by the structures shown below the scheme. Structural domains (NA, NA/NS, NS) are defined with regard to the distribution of GlcN N-substituents, as indicated.*

1.7 CONFORMATION OF HEPARIN

Heparin is a linear, unbranched, highly sulfated polysaccharide. GAGs such as HS/heparin tend to have extended conformations in solution due to their strong hydrophilic nature resulting from their extensive degree of sulfation. These molecules are surrounded by a shell of water molecules and occupy an enormous hydrodynamic volume in solution. They tend to repel each other when brought together due to their same net electrostatic charge. When a solution of GAGs is compressed, the water is squeezed out and the GAGs are forced to occupy a smaller volume. When the compression is removed, GAGs regain their original hydrated volume because of the repulsion arising from their negative charges.

Analysis of the conformations of individual sugars within heparin (Figure 1.12) indicates that unsubstituted IdoA residues exist predominantly in the $^1C_4$ chair form, whereas IdoA residues, when bearing a sulfate group at position 2 (IdoA2S), exist in equilibrium between a number of different conformations, the most important being the chair ($^1C_4$) and skew-boat ($^2S_0$) forms (D. R. Ferro *et al.* 1990). Solution NMR studies suggest that IdoA2S prefers a $^2S_0$ conformation, whereas glucosamine sulfated at the N and O positions (GlcNS6S) prefers a $^4C_1$ conformation (Mikhailov *et al.* 1996). It seems that glucosamine and its derivatives are stable in the $^4C_1$ chair conformation irrespective of substitution (Desai *et al.* 1993; D. R. Ferro *et al.* 1986; Yates *et al.* 1996).

Heparin oligosaccharides sometimes contain a non-reducing terminal 4-deoxy-L-threo-2-sulfohex-4-enopyranosyluronic acid (unsaturated Δ4-uronic acid, ΔUA2S) residue arising from heparin lyase cleavage of an HS chain. Based on the conformation of the 4,5-double bond, ΔUA2S can exist in either the $^2H_1$ or $^1H_2$ conformations (Figure 1.12 I) and the equilibrium between these two conformations is controlled by their substitution pattern. The solution structures of heparin-derived oligosaccharides determined by NMR spectroscopy suggest that the terminal ΔUA2S residue exists predominantly in the $^1H_2$ form, with a minor contribution from the $^2H_1$ form (Mikhailov *et al.* 1997).

The solution structure of a heparin dodecasacchride composed of six GlcNS6S-IdoA2S repeat units has been determined using a combination of NMR spectroscopy and molecular modelling techniques (Mulloy *et al.* 1993). These two structures (Figure 1.12 II) have been deposited in the protein data bank (PDB) under code 1HPN. One structure has all IdoA2S residues in the $^2S_0$ conformation (A) and the other one has all IdoA2S residues in the $^1C_4$ conformation (B). The three dimensional structure of heparin is complicated by the fact that iduronic acid may be present in either of two low energy conformations when internally positioned within an oligosaccharide. This conformational equilibrium can be influenced by the sulfation state of adjacent glucosamine sugars (van Boeckel *et al.* 1987). The $^2S_0$ form appears to be slightly favoured in terms of conformational stability, as it tends to minimise the unfavourable 1,3 diaxial non-bonded interactions that are expected in the $^1C_4$ form, where four of the substituents are axially oriented and only the carboxylate group is equatorial (Mikhailov *et al.* 1996). Whilst the spatial orientation of the 2-*O*-sulfate group in the IdoA2S residues is altered during $^1C_4$-$^2S_0$ intercoversion, no significant conformational change can be seen in the backbone of the polysaccharide chain in the NMR structures. In these NMR structures heparin adopts a helical conformation, the rotation of which places clusters of sulfate groups at regular intervals of about 17 Å on either side of the helical axis.

It is possible for the iduronate ring to adopt either the $^2S_0$ and $^1C_4$ forms in the protein-bound state, which enables it to make specific electrostatic interactions with the electropositive surface regions of a protein. The helical parameters for heparin oligosaccharides are conserved in spite of the conformational flexibility of the L-iduronate residues. NMR studies on a series of modified heparins with systematically altered substitution patterns indicate that all derivatives in the unbound form, regardless of the sulfation pattern, exhibit similar glycosidic bond ψ and φ conformations (Mulloy *et al.* 1994). These conserved glycosidic linkages are also consistent with the X-ray structures of heparin in complex with proteins such as acidic fibroblast growth factors (DiGabriele *et al.* 1998) (PDB codes 1AXM and 2AXM) and many other heparin structures bound to proteins (PDB codes 1AZX, 1BFC, 1E03, 1E0O, 1FQ9, 1G5N, 1GMN, 1QQP and 1TB6) .

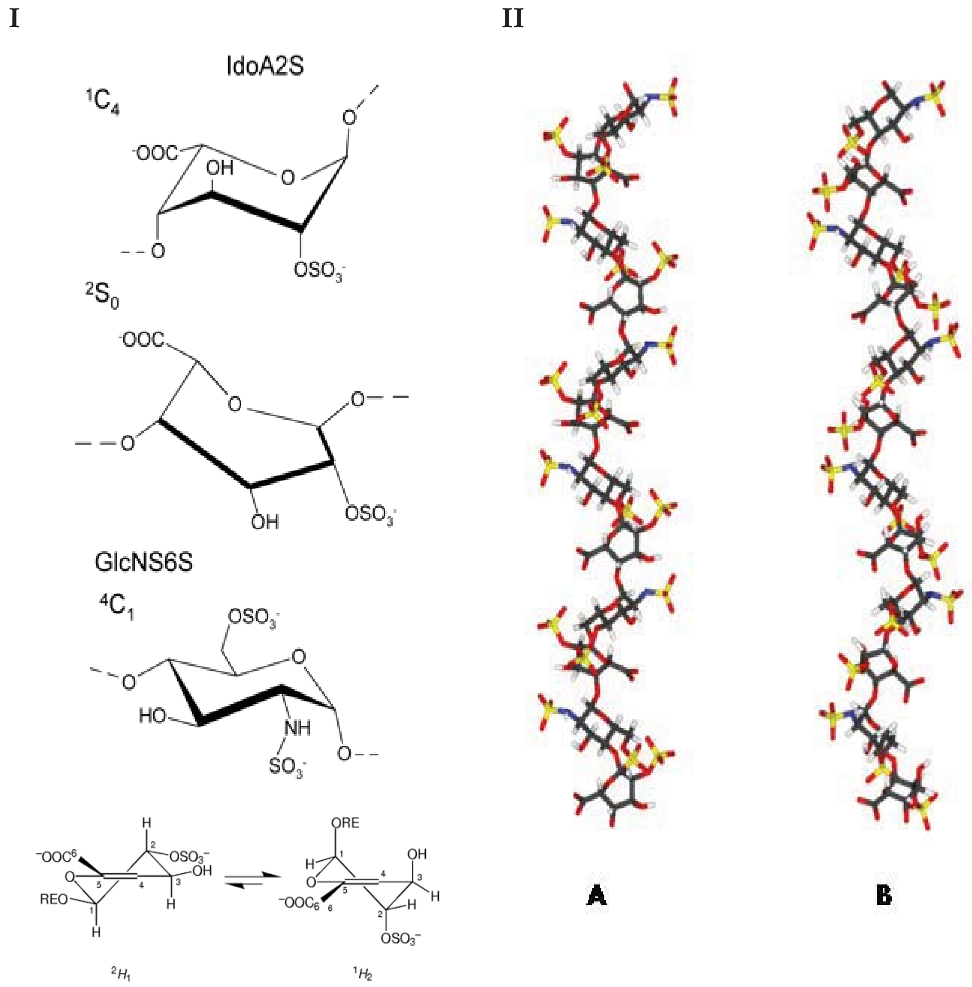


Figure 1.12. I. Conformations of sulphated iduronate, glucosamine, glucuronic acid and Δ4-uronic acid derivatives. **II.** The solution structure of a heparin dodecasacchride (PDB code 1HPN), in which all IdoA(2S) are in the $^{2}S_0$ conformation (A) and in which they are in the $^{1}C_4$ conformation (B).

It is also possible that the degree of flexibility in a disaccharide, the surrounding solvent water and cations could considerably affect the conformation of HS/heparin oligosaccharides depending on their local sequence. A theoretical study was recently undertaken (Remko *et al.* 2007) to determine the stable conformations of 1-OMe . IdoA2SNa$_2$ ($^{2}H_1$ and $^{1}H_2$ forms), 1-OMe GlcNS6SNa$_2$, 1,4-DiOMe GlcNa, 1,4-DiOMe GlcNS3S6SNa$_3$, 1,4DiOMe IdoA2SNa$_2$ ($^{4}C_1$, $^{1}C_4$, and $^{2}S_o$ conformations) of 1,4-DiOMe GlcNS6SNa$_2$ monomers and their ionised forms in the presence of solvent, cations as well as in isolation. In the gas-phase, the $^{2}H_1$ conformation of the uronate residue is more stable than the $^1H_2$ form observed in the presence of water. The most stable structure of the sodium salt of heparin confirmed the presence of a skew-boat ^2S_o conformation in water. With anions, the $^1C_4$ conformation is the most stable form. In general, the results indicate that in water, the relative stability of cation-heparin ionic bonds is considerably diminished.

Various studies of heparin conformations have revealed similar, well-defined molecular structures in terms of overall chain conformation, with versatility in the pyranose ring of iduronic acid (Remko & Hricovíni 2006). However, chemical parameters such as the primary sequence of GAGs and its degree of sulfation can result in different binding modes with proteins that can affect their activity.

1.8 CONFORMATION OF HEPARIN FRAGMENTS BOUND TO PROTEINS

Iduronate may exist in skew-boat, chair and intermediate ring forms in heparin-protein complex crystal structures. The central iduronate in the crystal structure of the foot and mouth virus complexed with a pentasaccharide (PDB code 1QQP) fits to the mixture of $^1C_4$ and $^2S_0$ forms (intermediate conformation), whereas the outer iduronates are in the $^1C_4$ and $^{2,5}B$ conformations (Fry *et al.* 1999). One of the iduronate rings in the hexasaccharide-bFGF complex crystal structure (PDB code 1BFC) adopts a $^1C_4$ chair conformation and the other a $^2S_0$ skew boat conformation (Faham *et al.* 1996). However, in the annexin V–heparin tetrasaccharide complex, the IdoA2S residue in the $^2S_0$ skew conformation was found to interact with the protein, whereas the non-interacting IdoA2S residue is in the $^1C_4$ conformation (Capila *et al.* 2001). These data suggest that when heparin binds to a protein, a change in the conformation of the IdoA2S residue may be induced, resulting in a better fit and enhanced binding, whilst the conformation of the less flexible GlcNS6S residue remains unaltered.

Various studies of the conformation and dynamics of heparin pentasaccharides have investigated their high affinity interactions with AT-III, both in the solid (Jin *et al.* 1997) and solution states (Ragazzi *et al.* 1990). The protein-bound pentasaccharide has a conformation roughly similar to one of the conformations predicted for the pentasaccharide wherein the iduronate residue adopts a conformation between the $^2S_0$ skew-boat and $^{2,5}B$ forms (Ragazzi *et al.* 1986). In contrast, NMR studies of the dynamics of the conformation of heparin oligosaccharides bound to AT-III and FGFs suggest that their proposed structure when complexed with proteins may be different to that in solution (Hricovini *et al.* 1999; Hricovini *et al.* 2002). NMR studies of heparin tetrasaccharides in the presence of fibroblast growth factors aFGF and bFGF indicate that FGF binding stabilizes the $^1C_4$ conformation of 2-*O*-sulfated iduronic acid (IdoA2S) residue directly involved in binding. On the other hand, the IdoA2S residue which is not directly involved in binding adopts an exclusively skew-boat $^2S_0$ conformation in the AT-III complex. In addition, complexation in both cases induces a change in the geometry around the glycosidic linkage between the non-reducing end glucosamine and the adjacent sugar residue as compared with the free aqueous solution state.

1.9 INTERACTIONS OF HEPARIN/HEPARAN SULFATE WITH PROTEINS

Extensive studies have identified common structural features in the heparin/HS binding sites of proteins. Different structural (NMR spectroscopy and X-ray crystallography) and molecular modelling approaches have been used to elucidate the three-dimensional features and structure-activity relationships of GAG–protein interactions (Sasisekharan *et al.* 2006). A list of the different proteins that have been crystallised in complex with heparin oligosaccharides and their characteristics such as the optimal length required for binding and their binding affinities can be found in Table 1.5. Crystal structures of some of the proteins such as IL-8, PF-4 and NCAM are not available in complex with GAGs.

1.9.1 Sequence considerations of GAG binding proteins

The use of structure and sequence-based statistical methods (Malik & Ahmad 2007; Shionyu-Mitsuyama *et al.* 2003; Taroni *et al.* 2000) indicate that residues Asn, Asp, Glu, Gln, Arg, His and Trp are more likely to form binding sites for non-sulfated carbohydrates than other amino acids. The aromatic residue Trp has a significantly higher mean solvent accessibility in carbohydrate binding locations, whereas aliphatic residues Ala, Gly, Ile and Leu, hydrophobic residues which are usually buried inside proteins, do not apparently participate in sugar binding. The aromatic ring in Trp can pack against the hydrophobic face of a sugar molecule. In the case of polyanionic carbohydrates such as GAGs, the hydroxyls on the sugars, charged groups such as sulfates and the sugar backbone can mediate electrostatic, van der Waals (VDW) and hydrogen bonding interactions with the proteins.

Table 1.5. Characteristics of a few of the known complexes between proteins and heparan sulfate or heparin fragments.

PDB code	**Name of protein**	**Type of protein**	**Size of oligosaccharide**	***K*d**	**Reference**
1G5N	Annexin V	Extracellular protein	8-mer	20 nM	(Capila *et al.* 2001)
2HYU, 2HYV	Annexin A2	Extracellular protein	4- to 5-mer	30 nM	(Shao *et al.* 2006)
1XT3	Cardiotoxin A3, A5, M4 and M1	Toxin	5- to 7-mer	μM	(S. C. Lee *et al.* 2005)
-	IL-8	Chemokine	18- to 20-mer	6 μM	(Spillmann *et al.* 1998)
-	PF-4	Chemokine	12-mer	nM	(Stringer & Gallagher 1997)
1U4L, 1U4M	RANTES	Chemokine	16- to 18-mer	32 nM	(Shaw *et al.* 2004)
1BFB, 1BFC	Basic fibroblast growth factor (bFGF)	Growth factor	4- to 6-mer	nM	(Faham *et al.* 1996)
1AXM, 2AXM	Acidic fibroblast growth factor (aFGF)	Growth factor	4- to 6-mer	nM	(DiGabriele *et al.* 1998)
1E0O	aFGF/ecto-domain of FGF receptor 2 (FGFR2)	Growth factor/receptor	12-mer	nM	(Pellegrini *et al.* 2000)

1FQ9	bFGF/ecto-domain of FGF receptor 1 (FGFR1)	Growth factor/receptor	12-mer	nM	(Schlessinger *et al.* 2000)
1AZX, 1E03, 1NQ9	AT-III	Serpin	5-mer	20 nM	(Jin *et al.* 1997)
1XMN	Thrombin	Protease	8-mer	7 μM	(Carter *et al.* 2005)
2GD4	AT-III/factor Xa	Serpin/protease	5-mer	100-200 nM	(D. J. D. Johnson *et al.* 2006)
-	NCAM	Adhesion protein	5-mer	52 nM	(Kasper *et al.* 2000; Soroka *et al.* 2003)
1FNH	Fibronectin	Adhesion protein	8- to 14-mer	μM	(Calaycay *et al.* 1985)

GAGs interact with residues that are prominently exposed on the surface of proteins. Ionic interactions are the most dominant type of interaction between heparin and a protein. Clusters of positively charged basic amino acids on proteins form ion pairs with spatially defined negatively charged sulfate or carboxylate groups on the heparin chain. The main contribution to binding affinity comes from an ionic interaction between the highly acidic sulphate groups and the basic side chains of arginine, lysine and, to a lesser extent, histidine (Fromm *et al.* 1997). The relative strength of heparin binding by basic amino acid residues has been compared and arginine was shown to bind 2.5 times more tightly than lysine. The guanidinio group in arginine forms more stable hydrogen bonds as well as stronger electrostatic interactions with sulfate groups. The ratio of these two residues is said to define, in part, the affinity of a protein site for GAGs (Hileman *et al.* 1998a).

Protein–GAG binding also involves a variety of different types of interactions including VDW forces, hydrogen bonds and hydrophobic interactions with the carbohydrate backbone. Amino acids such as asparagine and glutamine present in heparin-binding domains are capable of hydrogen bonding. Polar residues with smaller side chains like serine and glycine have been found to be the most frequent non-basic residues in heparin-binding proteins, providing minimal steric constraints and good flexibility for protein interaction with GAGs (Caldwell *et al.* 1996). The affinity of heparin to bFGF is partly due to the ionic and non-ionic interactions (Faham *et al.* 1996; L. D. Thompson *et al.* 1994b). Studies of the interaction of brain natriuretic peptide (BNP) with heparin revealed that only a small portion of the free energy of binding arises from ionic interactions, whereas the major contribution arises from hydrogen bonding between the polar amino acids on BNP and heparin (Hileman *et al.* 1998b). Hydrophobic forces may also play an important role in heparin-protein interactions. Based on NMR data, a tyrosine residue in a synthetic AT-III peptide has been reported to make specific, hydrophobic interactions with the N-acetyl group of a GAG pentasaccharide in porcine mucosal heparin (Bae *et al.* 1994).

Structural studies of the heparin–AT-III complex showed that basic amino acids participate in 5 to 6 ionic interactions, contributing 40% of the binding energy, whereas non-ionic interactions are responsible for the remaining 60% of the binding energy. The two aromatic residues, Phe 121 and Phe 122, residing near basic amino acids of the heparin-binding domain make direct contact with the pentasaccharide (Jairajpuri *et al.* 2003). The Phe 121 was mutated to Ala and Phe 122 to Leu resulting in decreased affinity of heparin for AT-III. These residues thus appear to play a critical role in heparin binding and AT-III activation.

1.9.2 Consensus sequences in GAG binding proteins

The X-ray crystal structures of many GAG-binding proteins helped to explore the existence of a consensus sequence for GAG binding with common features such as the arrangement of basic amino acids. Cardin and Weintraub (Cardin & Weintraub 1989) analysed the structures of 21 heparin-binding proteins and proposed that typical heparin-binding sites have the sequence XBBXBX or XBBBXXBX, where B is a lysine or arginine (with a very rare occurrence of His) and X is a hydropathic residue. The “X” in the consensus sequences was defined as hydropathic residue using matrices based on the frequency of occurrence of residues at specific position from known heparin binding proteins. The residues Asn, Ser, Ala, Gly, Ile, Leu, and Tyr were preferred at position “X”. Residues such as Cys, Glu, Asp, Met, Phe, and Trp exhibited a very low occurrence at position “X” in either the α helical or β sheet topology of heparin binding proteins.

Depending on the secondary structure of the protein, very few residues in these consensus sequences may actually participate in GAG binding. GAG-binding sites are often found along one exposed face of a protein and sometimes wrapping around multiple faces in case of beta sheets. Spacing of clusters of basic residues can also provide information on the structural features within heparin-binding sites that are important for GAG interaction and can facilitate the design of peptides that bind heparin efficiently (Hileman *et al.* 1998a).

The basic amino acids of the sequence XBBBXXBX, when modelled into an α-helix, are displayed on one side forming an amphiphatic helix arrangement (Figure 1.13A). Therefore, in order to interact with a linear GAG chain, it would be predicted that the positively charged amino acid residues in the alpha helical proteins would have to line up along the same side of the protein segment. Comparative analysis of heparin binding sequences have shown that basic amino acids are generally located about 20 Å apart (Figure 1.13B) in an amphipathic helix structure, and the same spatial arrangement is preserved in a beta-strand structure (Margalit *et al.* 1993). For example, the sulfates that mimic HS in the Artemin crystal structure (Silvian *et al.* 2006) were found to be separated by approximately 8-9 Å and arranged at the vertices of an approximate equilateral triangle in the pre-helix (having a positively charged heparin consensus sequence XBBXBX) and amino-terminal regions.

In β strands, the positively charged residues in a GAG-binding protein should be located very differently compared to what is seen in α-helical structures. The basic amino acids in the sequence XBBXBX line up on one face of the ß-strand, whereas the hydropathic residues points back into the protein core. Examples of β sheet heparin binding proteins are the CTXs (cobra cardiotoxins), which contain 9 discontinuous basic residues (-B-X_{2n-1}-B-, where X is any residue and B is basic residue) separated by an odd number of any residue (Vyas *et al.* 2005).

A third consensus sequence was similarly proposed in the heparin binding protein von Willebrand factor: XBBBXXBBBXXBBX, where ‘‘B” represents a cationic residue (Sobel *et al.* 1992). The consensus sequence TXXBXXTBXXXTBB as shown in

Figure 1.13C was also observed in aFGF, bFGF and transforming growth factor β-1 (TGFβ-1), where T defines a turn, B a basic amino acid (arginine or lysine) and X a hydropathic residue. The spatial distance between each of the three turns present in the consensus of these crystal structures was 12 to 18 Å (Hileman *et al.* 1998a).

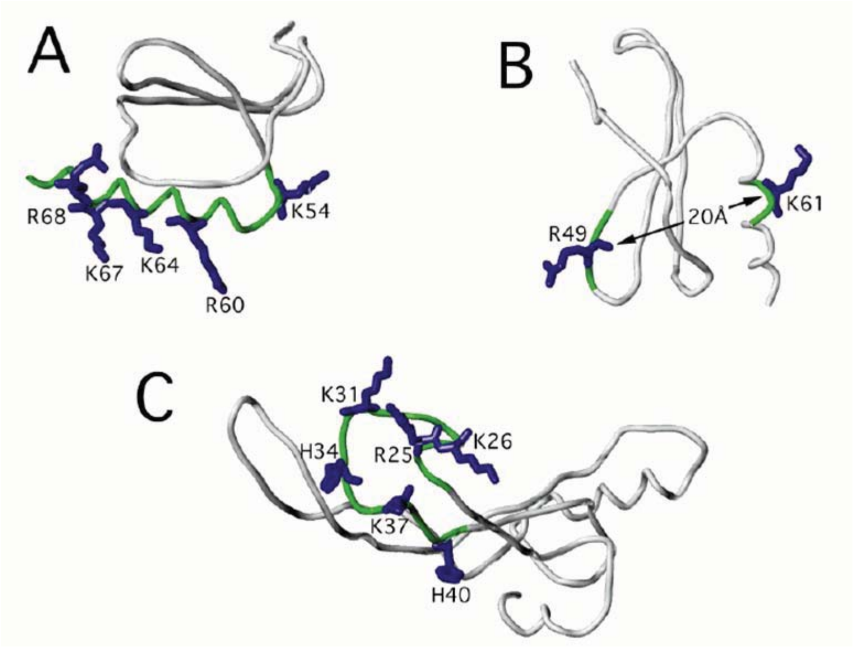


Figure 1.13. *Types of GAG consensus sequences present in proteins.* ***A:*** *Example of a linear XBBBXXBX motif with basic arginine and lysine residues (blue) oriented on one surface of a helix (green, residues 53–72) based on the structure of interleukin-8 X-ray coordinates (PDB code 3IL8)* ***B:*** *A linear motif, having basic arginine and lysine residues (blue) spaced at a 20 Å linear distance, located on opposite surfaces (green, residues 48–50 and 60–62) as observed in the X-ray monomeric structure of platelet factor-4 X-ray coordinates (PDB code HPF4).* ***C:*** *Linearly contiguous GAG-binding domains with the consensus TXXBXXTBXXXTBB based on the structure of TGFβ-1 (PDB code 1KLC) shown in white with the consensus sequence shown in green (residues 23–41). The figures were adapted from Hileman et al. (1998).*

GAG binding sites are often not conserved between protein structures, as observed in the case of chemokines, which have high structural similarity (Z. Johnson *et al.* 2004).

PF4 and IL-8 are members of the α-chemokine family that have very similar monomeric three dimensional structures, with anti-parallel β-strands and the α-helix in the C-terminus. PF4 has a heparin/HS binding consensus sequence "KKIIKK", where K is lysine and I is isoleucine protruding from the α-helix. The GAG consensus sequence in the equivalent α-helical domain of IL-8 is "KENWVQRVVEKFLKR", which is responsible for heparin/HS binding. The heparin/HS binding proteins of the β-chemokine subfamily (e.g. MIP-1α, RANTES) use a different structural motif, "KRNR". Members of both chemokine α and β families have additional residues and hence lack conservation in the GAG binding regions, allowing specificity and selectivity of HS binding across the chemokines.

1.9.3 Structural considerations of GAG binding proteins

The X-ray crystal structures of heparin-protein complexes have provided information on the structural features required for heparin binding, including the folding of the protein and the periodicity of clusters of basic residues, as well as the periodicity of sulfate clusters on the GAG chains and the sulfation level required for interactions with the binding site. Heparin binding sites can be formed by basic amino acids that are distant in sequence but are brought spatially close together through the folding of the protein. The end-to-end lengths of these extended clusters are comparable to the minimum GAG chain lengths that are required for binding (typically 6-12 monosaccharide units, approximately 25-50 Å long). The binding of GAG fragments to chemokines has strong length dependence but it is clearly not the only determinant of selectivity (Kuschert *et al.* 1999).

The periodicity of sulfate group clusters along an oligosaccharide chain can play a key role in determining the structure of a GAG binding site on the surface of either helical or β-sheet proteins. The regular periodicity of sulfate group clusters along one side of an oligosaccharide chain was consistent with the ability of heparin to induce an α-helical structure in polylysine peptides, allowing electrostatic interactions every three peptide turns between a HS cluster and a zeta-amino group of the polylysine peptide (Mulloy *et al.* 1996). The heparin octasaccharide was the minimal fragment size required for such interactions to occur with the polylysine peptides. A similar phenomenon has been detected for several lysine-rich regions in the Tau protein (Sibille *et al.* 2006), wherein the heparin oligosaccharide wraps tightly around the outer surface of the (double) pleated sheets, inducing secondary structural changes and thereby neutralising the inhibitory charge repulsions that would occur in a parallel stacking of the repeat regions formed by a polylysine stretch.

The varying extent of *N*- and *O*-linked sulfate groups and *N*-linked acetyl groups in a GAG oligosaccharide can effect the interaction of proteins with heparin/HS. In the case of RANTES, *O*-sulfation was more important than *N*-sulfation (Kuschert *et al.* 1999). MIP-1α, MCP-1 (monocyte chemoattractant protein-1) and IL-8 showed preference for both *N*- and *O*-sulfation. The binding of chemokines to GAG fragments requires both *N*- and *O*-sulfation (Kuschert *et al.* 1999). In addition, binding studies involving chemically modified heparins or HS preparations have shown that 2-*O*- and *N*-sulfate groups are important for interactions with bFGF (Figure 1.14) and doesnot require 6-O-sulfate group for binding. The HIV-Tat protein requires 2-*O*-, 6-*O*- and *N*-sulfate groups for optimal interaction with heparin (Marco Rusnati *et al.* 1997).

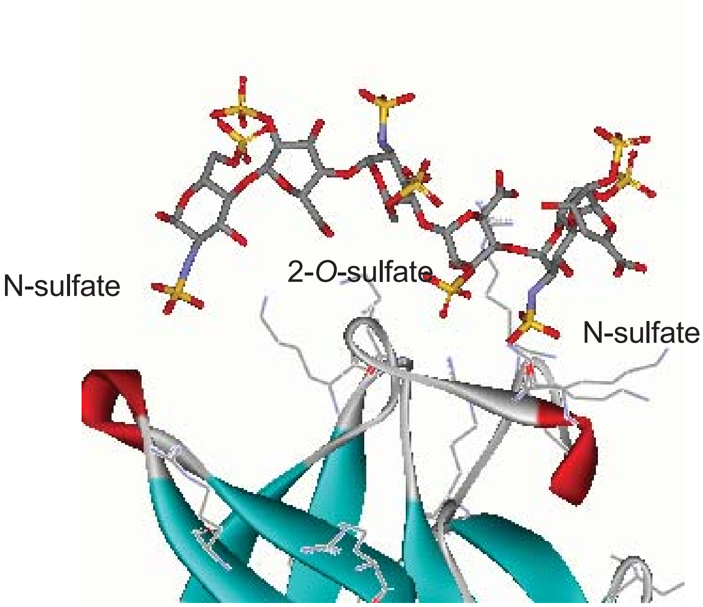


Figure 1.14. *Distinct role of sulfate groups of heparin(represented in sticks) in interactions with basic residues (represented as lines) of bFGF. bFGF is represented by secondary structure.*

1.9.4 Heparin-AT-III interactions: a case study of GAG-protein binding

The anti-coagulant activity of heparin arises primarily through activation of AT-III-mediated inhibition of blood coagulation factors such as thrombin and Factor Xa, as depicted in Figure 1.15. The interaction of AT-III and its coagulation factors with heparin moves through a number of affinity states to terminate in a high affinity interaction. First, the interaction between GAG and AT-III is mediated by a well-defined unique pentasaccharide sequence within heparin. This binding generates a conformational change in the structure of AT-III, which enables additional interactions between AT-III and heparin, resulting in stronger binding. The conformational change also expels a protease reactive loop in AT-III. A ternary complex is formed, after which the AT-III interaction reverts to low-affinity binding, resulting in the release of heparin from the covalent AT-III–protease complex. Several theories have involved in relation to the length dependence of the interaction of heparin with AT-III and serine proteases. Heparin chains at least 16 saccharides in length are required to accelerate the reaction of AT-III with thrombin, even though only the pentasaccharide sequence is necessary to bind AT-III (Petitou *et al.* 1998). In contrast, heparin chains as small as the AT-III-binding pentasaccharide are able to accelerate the inactivation of the other target coagulation enzymes, such as Factor Xa.

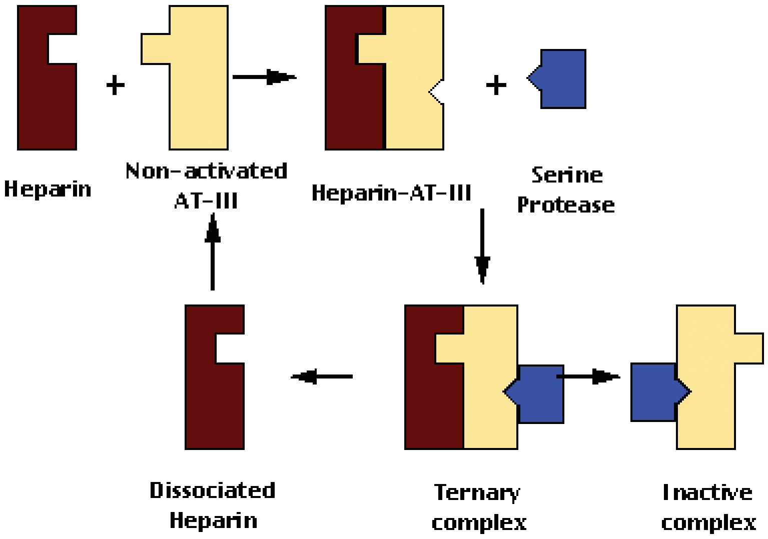


Figure 1.15. *Heparin-binding domain of AT-III. Heparin enhances the actin of the plasma protease inhibitor AT-III, followed by inhibition of clotting factor proteases, (e.g. FIIa, Xa, Ixa and Xia), by forming stable complexes with them. Heparin speeds up the formation of these complexes by binding to AT-III and causing a conformational change, thereby activating AT-III. This figure was adapted from*

Ontario Veterinary College's online course modules, taken from the University of Guelph's website (Guelph).

1.9.5 GAG-fibroblast growth factor interactions

Extracellular domains of fibroblast growth factors aFGF (FGF-1) and bFGF (FGF-2) have been extensively studied to determine the thermodynamics and kinetics of their interactions with heparin. These growth factors exert their biological effects by binding to different, specific cell surface FGFRs. High-resolution X-ray crystal structures of complexes of FGF, FGFR, and a heparin oligosaccharide provided insight into the stoichiometry and structural features of this physiologically relevant interaction. In the crystal structure of a 2:2:2 dimeric ternary complex of bFGF, FGFR-1, and a heparin decasaccharide, heparin makes numerous contacts with both bFGF and FGFR-1, stabilising the FGF-FGFR interaction (Schlessinger *et al.* 2000). Heparin also makes contacts with the FGFR-1 of the adjacent FGF-FGFR complex, thus seeming to promote FGFR dimerisation (Figure 1.16). The 6-*O*-sulfate group of heparin plays a major role in promoting these interactions (M. Rusnati *et al.* 1994).

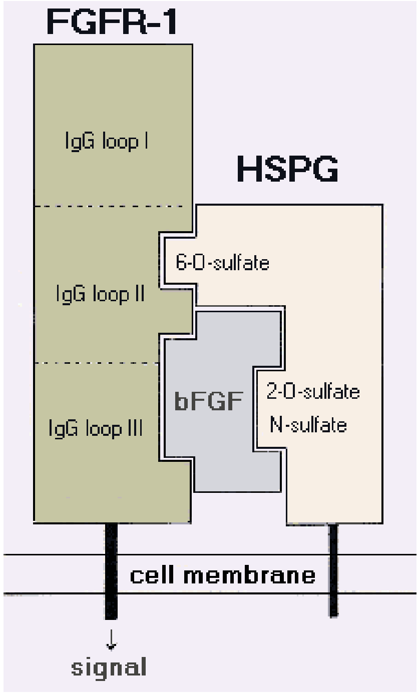


Figure 1.16. *Schematic representation of the bFGF-FGFR1 complex. Heparin requires both 2-O, 6-O-sulfate and N-sulfate groups, to promote the binding of*

bFGF to soluble FGFR-1. The binding of heparin/HS to bFGF, without 6-O-sulfate groups is not sufficient to induce bFGF interaction with FGFR. The figure was taken from the angiogenesis portal from the Department of Biomedical Sciences and Biotechnology, University of Brescia – Italy (Presta 2005).

The crystal structure of a 2:2:1 complex of aFGF, FGFR-2, and a heparin decasaccharide has also been determined to a 2.8 Å resolution (Pellegrini *et al.* 2000). The complex is assembled around a central asymmetric heparin molecule linking two aFGF ligands into a dimer that bridges between two receptor chains (Figure 1.17). The heparin fragment makes contact with both aFGF molecules but only with one receptor chain. It is clear that different member of the FGF family and their respective receptors (FGFRs) may interact differently with heparin/HS due to the heterogeneity in the structure of HSPGs and FGF receptors on cell surfaces in different tissues. It has been reported that aFGF may recognise several conformations of the iduronic residues of a GAG hexasaccharide. It is believed that the hexasaccharide undergoes local $^1C_4$-$^2S_0$ equilibrium conformational changes as a result of ionic interactions with flexible Arg and Lys side chains present in the protein (Canales *et al.* 2005).

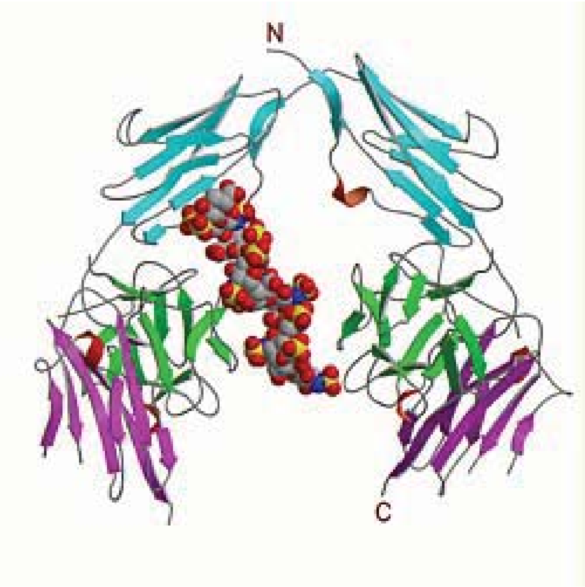


Figure 1.17. *Ribbon diagram of the aFGF-FGFR2-heparin complex (PDB code 1E0O). The heparin fragment (shown in CPK) makes contact with both aFGF molecules (beta strands shown in green) but only with one FGFR2 receptor chain*

(immunoglobulin domains shown in cyan and magenta). The figure was adapted from Pellegrini et al. (2000).

1.10. ROLE OF pH IN GAG BINDING

Certain HS-protein interactions are regulated by pH. Alteration of the pH can have profound effects on the ability of some proteins to bind heparin or HS. This is the case of the synthetic beta-amyloid peptide (Aβ) (Fraser *et al.* 1991), selenoprotein P (Arteel *et al.* 2000), granulocyte macrophage colony stimulating factor (GM-CSF) (Wettreich *et al.* 1999), the mouse mast cell protease 7 (Matsumoto 1995) and stromal cell-derived factor-1 (SDF-1) (Veldkamp *et al.* 2005). This occurs particularly when the GAG binding site contains histidines, since these amino acids have a pKa of approximately 6. Hence, if the pH falls closer to 6 an increasingly larger proportion of histidines will become protonated and hence positively charged, thus favouring electrostatic interactions with the negatively charged sulfate groups of GAGs.

A further example is that of mouse mast cell protease 6 (MCP-6). Molecular modelling of MCP-6 identified four conserved, pH dependent and surface exposed histidine residues, His 35, His 106, His 108, and His 238 (Figure 1.18), that mediate the interaction of mast cell tryptase 6 with heparin in a pH dependent fashion (Hallgren *et al.* 2004). The electropositive nature of the surface of the protease, as shown in Figure 1.18, is due to presence of pronated histidines that can make favourable interactions with GAGs, as compared to the surface accessible in the presence of deprotonated histidines. Histidine proline-rich glycoprotein (HPRG) is another example wherein binding to heparin is minimal at neutral pH but increases rapidly to a maximum at pH 6.5 (Borza & Morgan 1998). At an intermediate pH, both protonation of histidines and the binding of zinc promote the interaction of HPRG with heparin. It is probable that there is a pH range where all histidines will be protonated, whereas most, if not all, of the glutamic and aspartic acid residues will still be negatively charged. This is likely to be most favourable situation for heparin binding.

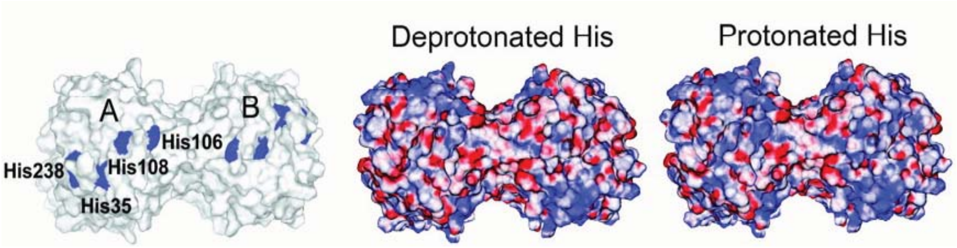


Figure 1.18. *Model of mouse Mast Cell Protease (mMCP-6) model reported by Hallgren et al 2004. His residues (shown in blue) are located at the edge of the A-B surface. The electrostatic potential surface is shown for both neutral (deprotonated His residues, positive charge contributed by Lys/Arg residues) and acidic pH (protonated His residues).*

1.11. EFFECT OF METAL IONS ON GAG BINDING

Sulfated GAG chains also bind strongly to divalent metal ions present in proteins or in solution. The binding of heparin/HS to proteins is enhanced in the presence of divalent cation such as zinc. The binding of endostatin to heparin and HS requires the presence of divalent cations (Ricard-Blum *et al.* 2004). The presence of Zn^{2+} metal ions enhances the binding of endostatin to heparin/HS.

Crystallographic studies of human annexin A2 in complex with heparin-derived oligosaccharides suggest that annexin A2 exhibits significant Ca^{2+}-dependent heparin-binding properties (Figure 1.19) at pH 7.4, either as a monomeric protein or as a component of the A2t heterotetramer (Shao *et al.* 2006). In the complex of annexin V with a heparin oligosaccharide the calcium cation does not interact directly with the GAG fragment but it induces the conformation of protein loops necessary for binding (Capila *et al.* 2001). Prion proteins (PrP) also bind GAGs at pH values above the pKa of histidine and in a metal ion-dependent fashion (Gonza'lez-Iglesias *et al.* 2002). Prion protein-GAG complexes are stabilised by Cu^{2+} or Zn^{2+} and prion protein-GAG interactions are mediated largely by protonated and Cu(II)-bound His side-chains present at the N-terminal domain of PrP. Divalent cations were not a prerequisite for the interaction of GAGs with lipoproteins but were found to stabilise the lipoprotein complexes of heparin. It was observed that Mn^{2+} is better than Mg^{+2} or Ca^{+2} at promoting stronger binding between the acidic groups of heparin and the phospholipid portion of LDL (Srinivasan *et al.* 1975).

It is known that Zn^{2+} binds selectively to heparin rather than to other GAGs (Parrish & Fair 1981), which suggests that binding of divalent cations to GAG chains is not always a simple electrostatic interaction between the negatively charged groups on the carbohydrate and the positively charged metal ion. NMR evidence indicates that iduronic acid is the main binding site in heparin for divalent cations. It is also known that Zn^{2+} metal ion binding controls the ring conformation of iduronate in heparin and HS, as suggested by spectral data, showing that the $^{1}C_{4}$ ring conformation of iduronic acid is stabilised over the $^{2}S_{o}$ conformation (Whitfield *et al.* 1992; Whitfield & Sarkar 1992). Consequently, divalent cation binding may be expected to influence the specificity and affinity of protein interactions with GAGs.

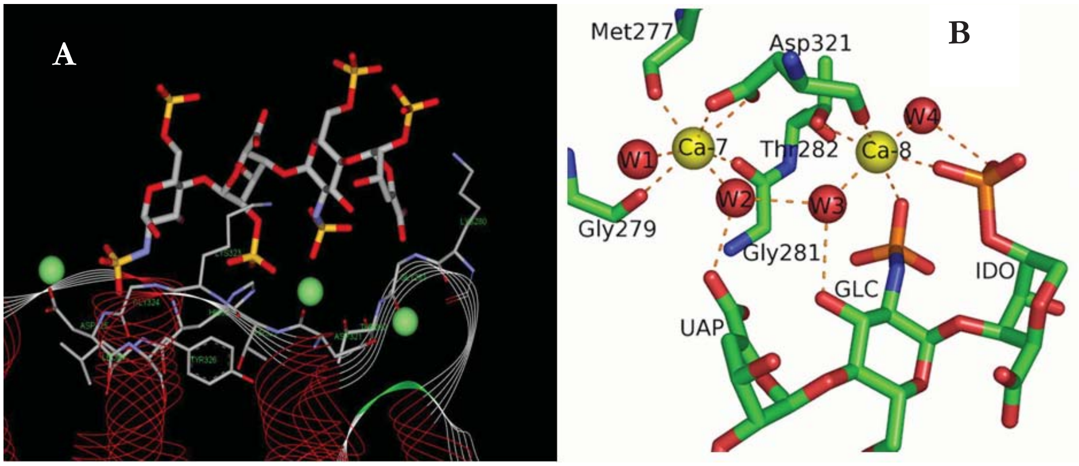


Figure 1.19. *Calcium coordination at the heparin-binding site in the crystal structure of Annexin A2.* ***A:*** *The heparin tetrasaccharide binding site. The Ca^{2+} ions are shown in green.* ***B:*** *The Ca^{+2} ions are shown as yellow spheres and water molecules as red spheres. Orange dashed lines denote the Ca^{+2} coordination bonds and interactions between water molecules and the oligosaccharide. The figures were extracted from (Shao et al. 2006).*

1.12. MOLECULAR MODELLING STUDIES OF GAGS AND GAG-PROTEIN INTERACTIONS

In view of the limited structural knowledge available on GAG-protein interactions and the phenomenal structural diversity of heparin and HS, molecular modelling approaches have assisted the understanding of GAG binding affinity and specificity. GAGs are challenging from a molecular modelling perspective because of their high negative charge density and their conformational flexibility. Protein side chains also have a high degree of conformational flexibility. This means that, if all possible conformations of the sulfate and hydroxyl groups on the oligosaccharide and all rotamers of charged side chains in the protein are taken into account, an accurate prediction of GAG-protein binding becomes an extremely challenging task.

Several molecular modelling techniques have been described in the literature for the successful prediction of sulfated GAG binding sites on the surface of proteins and for the prediction of their relative affinities. These methods include energy mapping of ligand probes on the surface of proteins, molecular docking and scoring, and molecular dynamics simulations.

1.12.1 Prediction of GAG binding sites on protein surfaces using GRID

The prediction of the location of GAG binding sites on the surface of proteins has been attempted by searching for the most positively charged patches of amino acids. The GRID algorithm (Goodford 1985) has been useful for mapping the most energetically favourable positions where sulfate groups may bind to the surface of proteins (Figure 1.20). Such studies have been performed with a number of proteins such as aFGF, bFGF, antithrombin and IL-8 (Bitomsky & Wade 1999). The GRID program uses atom probes to represent polar or charged groups on saccharide molecules. Mapping sulfate interaction energies can be first computed using GRID and then followed by docking. In a different study, different HS binding modes were proposed for its interactions with chemokines RANTES, MIP-1α, and CDF (Chemokine Domain of Fractalkine), illustrating that the types of interactions that may exist on the surface of proteins are determined by the three-dimernsional structure of the proteins (Lortat-Jacob *et al.* 2002). This study first used the GRID program, followed by docking to predict the most favourable anchoring position for a charged sulfate group on the surface of the chemokines. However, this study did not allow for a fine analysis of the GAG sequence optimal for binding (i.e., effect of *N* and *O*-sulfation).

1.12.2 Molecular docking

Several ligand-protein docking studies have been reported for the prediction of heparin-binding sites on AT III, aFGFand bFGF. In these studies the docking predictions have been compared to crystallographic data available for complexes of these proteins with oligosaccharide fragments (Bitomsky & Wade 1999). After correctly predicting the binding sites for AT III, aFGF and bFGF, these authors used docking to predict the heparin-binding site on IL-8. Other molecular modelling studies have been carried out to predict the binding of a hexasaccharide to the multi-component complex between bFGF and FGFR1. The results were consistent with experimental data of the binding mechanism of bFGF to its receptor, the receptor dimerisation, and site-specific mutagenesis and biochemical cross-linking data (Lam *et al.* 1998). Molecular docking studies have also been used to predict that a long heparin fragment such as a dodecasaccharide or tetradecasaccharide is required for

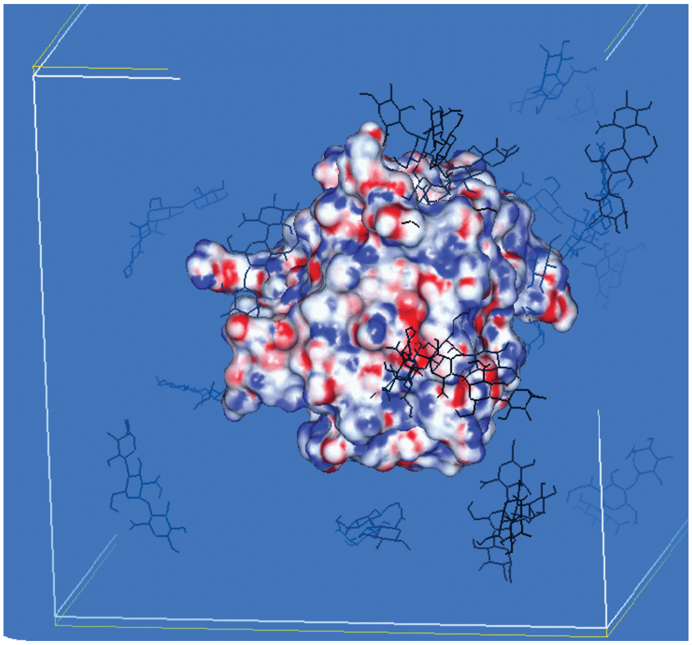


Figure 1.20. *Example of the docking of GAG saccharides onto bFGF using the GRID algorithm. The molecular surface of bFGF is shown with its electrostatic potential. The figure was adapted from the online course on computing methods in biochemistry (Pagel 1999).*

binding to the chemokine SDF-1α dimers (Sadir *et al.* 2001). In another study, since the crystal structure of the chemokine was not available, different protein models for a MIP-1α dimer were built based on the crystal structures of PF4 and IL-8 (Stringer *et al.* 2002). Docking simulations using heparin penta- and endecasaccharides predicted the interaction of the S-domains (usually 12-14 saccharides long) and the electropositive surface on opposite faces of the MIP-1α dimer.

A study of the interaction between a heparin pentasaccharide and AT-III has been carried out, despite the difficulty associated with the known conformational change that occurs in the protein upon ligand binding. Homology modelling of the protein structure and manual docking of the pentasaccharide were used to determine the basic amino acids involved in the recognition of the sulfate and carboxylate groups of the oligosaccharide. These predictions were confirmed by automated docking simulations (Grootenhuis & Van Boeckel 1991). The crystal structure of the complex between anti-thrombin and the pentasaccharide revealed the existence of contacts between heparin and arginine and lysine residues on three different helices of the protein (Jin *et al.* 1997). The crystal structures of ternary complexes of anti-thrombin, thrombin and heparin and anti-thrombin, Factor Xa and heparin provided further information about the large conformational changes that occur in anti-thrombin upon activation.

Docking simulations have also been used to predict the binding mode of a heparin oligosaccharide on the surface of endostatin (Ricard-Blum *et al.* 2004), as well as to determine the binding mode of a hexasaccharide to aFGF (Canales *et al.* 2006). In the aFGF study, most of the low energy docked conformers of a hexasaccharide orientated towards Lys127 and Lys142 on the surface of the growth factor. Other studies have suggested the likely amino acid residues that comprise heparin binding sites in proteins such as the aFGF, bFGF and AT-III (Mulloy & Forster 2000).

Docking methods have also been used for the screening of a combinatorial virtual library of hexasaccharides, identifying high specificity heparin/HS sequences using the AT-III-heparin crystal complex (Raghuraman *et al.* 2006). The combinatorial library consisted of 6859 unique heparin hexasaccharides which were generated on the basis of an 'average backbone'. The intra- and inter-glycosidic conformations were constrained, irrespective of sequence and intra-ring conformational variability to mimic the crystallised hexasaccharide, and AT-III was considered rigid. The linear correlation between the GOLD docking score and the predicted free energy of binding suggested that GOLD scores correlated with protein binding affinity. Twenty-eight hexasaccharide sequences were predicted to bind with higher affinity to AT-III on the basis of higher GOLD score. These 28 sequences were again subjected to triplicate docking runs wherein a larger conformational space was searched for multiple geometries of ligands in the binding site. Out of the 28 high-affinity sequences, 10 sequences were predicted to have high specificity of interaction. However, these binding affinity and specificity predictions lacked experimental validation.

Different methods were used to dock heparin and activated protein C (APC) (Fernandez-Recio *et al.* 2002). A structure-based virtual screening approach has also been used to dock short heparin oligosaccharides onto APC. The modelling study supported by experimental data indicated that short heparin oligosaccharides bind to loop structures 37, 60 and 70 in APC and this binding impairs the interaction of APC with FVa (factor Va) during APC-catalysed cleavage. Recent developments in docking techniques of short oligosaccharide chains onto APC and hexasaccharides onto AT-III further demonstrate the effectiveness of virtual screening in glycobiology.

1.12.3 Scoring methods to rank docked ligand conformations

Specific scoring functions have been developed for ranking the binding modes of non-GAG carbohydrates to proteins (Kerzmann *et al.* 2006; Laederach & Reilly 2003). These functions can also be used for GAGs. A structure-activity relationship study using docking calculations with various scoring functions has been done for calculated and observed binding affinities for the complexation of oligosaccharides to aFGF and bFGF. The predicted binding modes in both FGFs were similar and good correlations were obtained between the predicted and experimental binding affinities.

The BLEEP (Biomolecular Ligand Energy Evaluation Protocol) method has been used successfully to identify low-energy binding modes of heparin fragments (Mitchell *et al.* 1999). This study was carried out in presence of a shell of water molecules. Various conformations for heparin were generated and the structure of human bFGF was kept rigid. The method correctly assigned the lowest energy to the binding modes observed in the crystal structure, indicating that its PMFscore

(Potential of Mean Force score) scoring function is able to rank well the interaction energies of molecules such as GAGs.

1.12.4 Molecular dynamics simulations

Molecular dynamics (MD) simulations have been reported for oligosaccharide complexes with proteins such as galectin-1 (Goodford 1985) and endo-1,4-b-xylanase II (XynII) (Laitinen *et al.* 2003), but very few MD simulations have been performed for sulfated GAGs such as heparin and HS.

Some of the MD simulations of heparin fragments have been performed in aqueous solution. Simulations of a heparin decasaccharide-water-sodium system using the GROMACS forcefield and the SPC (Simple Point Charge) and SPC/E (Simple Point Charge/Extended) water models were in agreement with NMR data of the conformation of heparin in solution under physiological conditions (Verli & Guimarães 2004). In these simulations the conformational change in iduronic acid and the conformational flexibility of the glycosidic linkage were investigated. These simulations reported great variability in the conformation of heparin compared with the previously determined NMR structures of heparin, due possibly to the use of different partial atomic charges (Löwdin atomic charges). MD simulations have also been performed for a complex of a heparin pentasaccharide with AT-III in order to characterise the energetic contribution of important amino acids required for interactions with GAG fragments and the ability of GAG fragments to induce the observed conformational change in AT-III (Verli & Guimarães 2005). These simulations revealed that there is no specific conformational requirement for IdoA, as either of the skew-boat or chair conformations are appropriate for binding with a similar enthalpy to AT-III.

The set of parameters representing force constants, equilibrium bond lengths and angles, partial charges and VDW interactions can significantly affect the accuracy of simulations of ligand-protein interactions. There is a variety of molecular mechanics force fields that have been designed for the modelling of carbohydrates. This include OPLS (Optimized Potential for Liquid Simulations) (Kony *et al.* 2002), GROMOS (GROningen MOlecular Simulation package) (Lins & Hünenberger 2005), CSFF (Carbohydrate Solution Force Field) (Kuttel *et al.* 2002), CHARMM (Brooks *et al.* 1983), CHARMM CHEAT95 (Grootenhuis & Haasnoot 1993), Glycam/AMBER

(Woods *et al.* 1995), MM2 (Allinger 1977) and MM3 (Allinger *et al.* 1989), PEF95SAC (Fabricius *et al.* 1997) and PIM (set of carbohydrate parameters) (Imberty *et al.* 1999). These force-fields do not always contain parameters for sulfated carbohydrates such as GAGs, but various approaches can be followed to develop specific parameters for GAGs using the MM2 (D. R. Ferro *et al.* 1995; D. R. Ferro *et al.* 1997), AMBER and CHARMm force fields. Some non-bonded parameters not available from the work of Huige and Altona (Huige & Altona 1995) can be approximated from those for phosphates available from AMBER or CHARMm.

1.13. THERAPEUTIC POTENTIAL OF GAG MOLECULES AND GAG MIMETICS

In living cells, carbohydrates such as GAGs derive their activity through binding to their protein receptors. These carbohydrate-protein interactions could be mimicked to enhance the binding and affinity of the interaction for drug discovery purposes. X-ray crystallography, NMR spectroscopy and structure-based design have been used to investigate carbohydrate-protein interactions. The affinities of such interactions start at millimolar levels, whereas small molecule chemical entities used in drug discovery often have submicromolar or nanomolar binding affinities. Specifically designed synthetic compounds that can mimic the structure and interactions of carbohydrate ligands, such as GAG mimetics, may bind their receptors with higher affinity than the natural GAG oligosaccharide.

The molecular diversity of heparin/HS interactions has led to the clinical progression of GAG mimetics (Fugedi 2003). Discrete GAG sequences can bind specifically and make unique interactions with a large number of proteins including chemokines (Z. Johnson *et al.* 2005), growth factors (Spillmann & Lindahl 1994), proteases such as the AT-III (Jin *et al.* 1997), and adhesion molecules (Lyon & Gallagher 1998). Nevertheless, the design of GAG mimetics requires an understanding of the pathophysiological role of a given GAG-protein interaction and its specificity. Potential strategies based on heparin/HS-protein interactions have recently been described to assist GAG-based drug discovery (Lindahl 2007). As shown in Figure 1.21, GAG-based drugs can act in several ways:

1) Endogenous heparin, once released from mast cell granules, tends to exist as free GAG chains. The negatively charged sulfated regions provide interaction sequences for a variety of proteins, including growth factors, chemokines, enzymes/enzyme inhibitors, and various extracellular-matrix proteins (Figure 1.21A).

2) Activate (agonists) or inactivate (antagonists) protein-based receptors. An example of a GAG acting as an agonist can be found in the interaction of a specific heparin pentasaccharide with AT-III, which potentiates AT-III to inhibit the serine proteases involved in blood coagulation (Petitou & van Boeckel 2004) (Figure 1.21B).

3) Compete with endogenous GAGs. Receptor signalling can be inhibited by a GAG-based drug that displaces a ligand from its receptor. The interactions of a GAG with the protein depend mainly on the charge density of the GAG. This charge density is due to the content of either N- or O-sulfated regions. These regions can be fine tuned to develop an effective drug. Another drug discovery strategy has been used in the case of endostatin (Figure 1.21D). An oligosaccharide comprising two *N*-sulfated regions separated by at least one *N*-acetylated glucosamine unit was reported to compete with endogenous HS for binding to endostatin (Ricard-Blum *et al.* 2004).

4) Inhibit GAG biosynthesis. For example, some O-xyloside inhibitors specifically target xylosyltransferases that initiate HS biosynthesis (Figure 1.21E) and are known to have a role in cancer therapy (Belting *et al.* 2002).

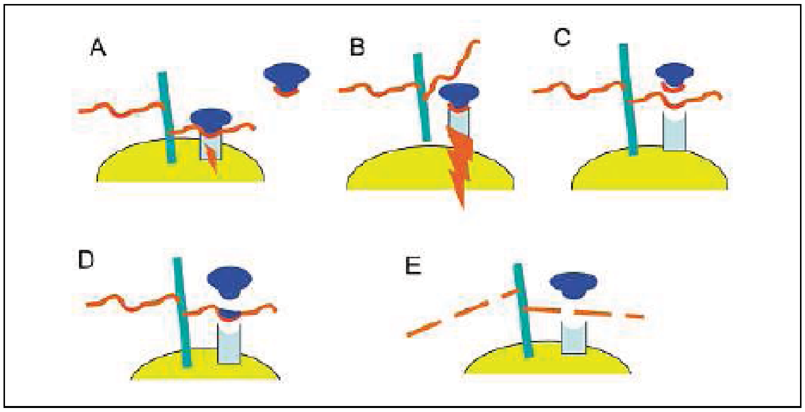
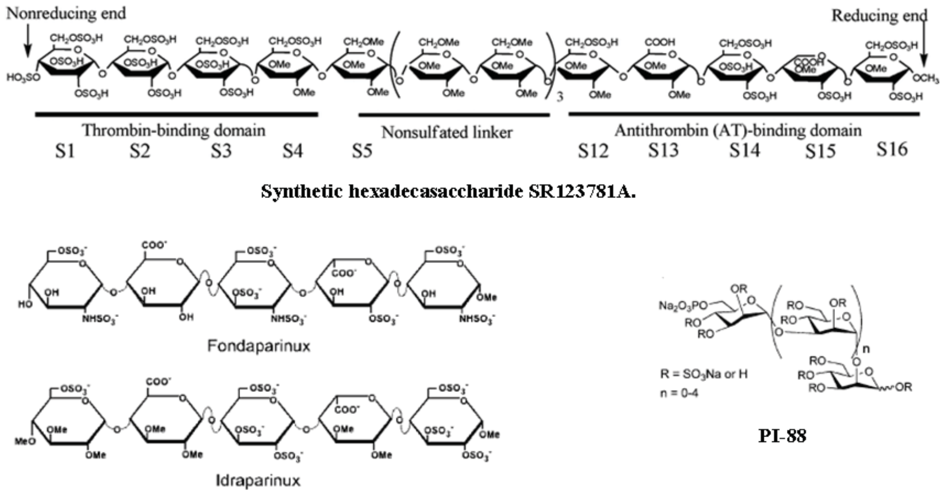


Figure 1.21. *Potential strategies for drug development based on HS-protein interactions, as illustrated by U. Lindahl (2007a). Most of the examples shown relate to HS (shown in red)-dependent binding of a protein ligand (shown in dark*

blue, e.g. a growth factor) to its cell-surface receptor (shown in light blue); however, similar principles would apply to a variety of interaction systems. ***A)*** *Binding of protein ligand to the receptor, assisted by endogenous HS.* ***B)*** *Activation of the receptor by a GAG mimetic that forms a ternary complex with the ligand and receptor, and displaces endogenous HS. Direct binding of a GAG mimetic to target a protein may promote or inhibit bioactivity.* ***C)*** *Inhibition of receptor signalling by a drug that displaces a ligand from its receptor.* ***D)*** *Inhibition of receptor signalling by a GAG mimetic that blocks the protein binding site of HS.* ***E)*** *GAG mimetics (not indicated) interfering with HS biosynthesis.*

Very few GAG fragments have been developed for therapeutic use (Figure 1.22), mostly because the synthesis of such fragments is chemically challenging. The synthetic challenges posed by the complex structure of these oligosaccharides are the availability of $_{\text{L}}$-idose and $_{\text{L}}$-iduronic acid from commercial or natural sources and the lack of efficient synthetic routes to access sufficient amounts of these monosaccharides. Other challenges are the development of a suitable protecting-group strategy to allow the implementation of a high degree of functionalisation of heparin/HS fragments and the stereo selective and efficient formation of interglycosidic bonds in the carbohydrate backbone (Codée *et al.* 2004).

The most recognised pharmaceutical application of GAGs is in anti-coagulation. Many pharmaceutical companies like Organon and Sanofi-Aventis are working on the development of commercial GAG-based drugs that can bind AT-III and thereby cause anti-coagulation. Their goal is to produce a GAG-based drug that is efficacious but is less frequently administered than full length heparin. An example of such drug is the synthetic pentasaccharide Arixtra® (fondaparinux or SR90107/Org31540) (Choay *et al.* 1983). Fondaparinux is known to bind AT-III and to have better efficacy at low doses (half-life of 17 hours). The crystal structure of Arixtra complexed with AT-III confirms the importance of basic residues Arg 46, Arg 47, Lys 114, Lys 125, Arg 129 and Lys 114 for this interaction (Jin *et al.* 1997).



Synthetic hexadecasaccharide SR123781A.

Fondaparinux

Idraparinux

PI-88

Figure 1.22. *Heparin sequences of therapeutic significance.*

Fondaparinux has been followed by the development of several other clinical candidates, such as idraparinux (SANORG 34006), which also selectively inhibits coagulation Factor Xa as well as binds AT-III (Herbert *et al.* 1998). Idraparinux is currently in phase III trials for the treatment of venous thromboembolism. The synthesis of idraparinux is much easier than that of fondaparinux or heparin. It also has higher affinity (K_d of 1 nM) and better efficacy than fondaparinux (K_d of 25 nM). Idraparinux and fondaparinux differ from each other in the type of sulfation and the methylation of all hydroxyl groups. The hydroxyl groups in idraparinux are methylated and the *N*-sulfate groups in fondaparinux are replaced by *O*-sulfates in idraparinux. Idraparinux has an increased half-life (120 hours) in the bloodstream. The higher activity was observed in an idraparinux pentasaccharide due to the presence of methyl ethers, which interact with complementary lipophilic groups at the protein surface.

AT-III in the absence of coagulation factors has been crystallised complexed with fondaparinux. The structural requirements for heparin binding to AT-III were determined on the basis of the crystal structure and structure-activity relationships for a series of pentasaccharides, as shown in Figure 1.23 below, with various combinations of sulfate and carboxylate groups (Petitou & van Boeckel 2004). The 3-*O* sulfate group at position H of fondaparinux exhibits stronger binding to AT-III by interacting with positively charged amino acids, whereas the lack of the 3-*O*-sulfate group at position F results in a decreased binding affinity to AT-III of nearly 20,000-fold.

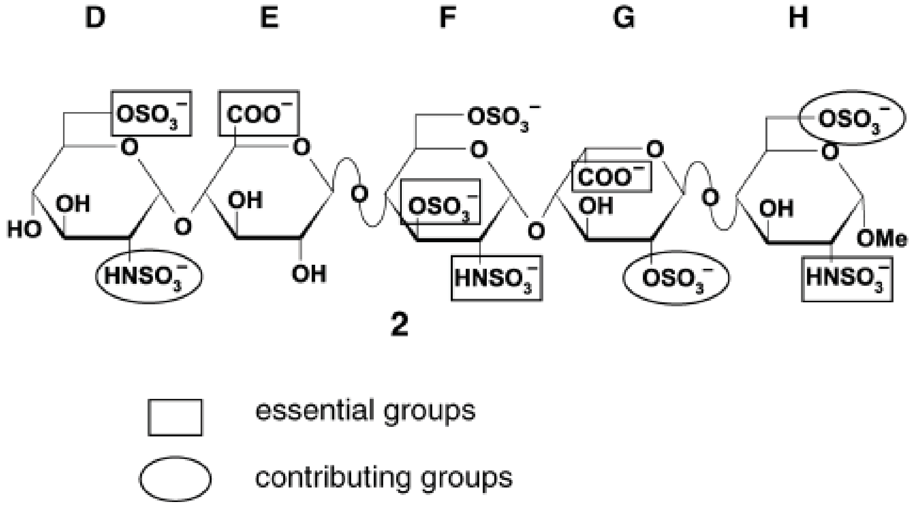


Figure 1.23. *The structural requirement for heparin binding to AT-III based on the structure activity relationships of fondaparinux (Maurice Petitou & Boeckel 2004). The groups highlighted in the boxes are absolutely essential for the activation of AT-III, whereas the groups in the circles only help to increase the biological activity. Sulfate group at the 3-O position H of the fondaparinux pentasaccharide can exhibit stronger binding to AT-III and the N-sulfated groups can be replaced by methyl groups to form more potent pentasaccharide idraparinux. The 3-O-sulfate group on the GlcN unit F of pentasaccharide is very specific for to the AT-III binding sequence, and is absent in the heparin molecules. The figure was adapted from Petitou et al. (2004).*

Heparin binds both AT-III and thrombin simultaneously to form a ternary complex, as well as bind and inhibit Factor Xa. The required size of an oligosaccharide that can inhibit thrombin activity is much larger than the specific pentasaccharide that is required to bind AT-III and inhibit Factor Xa. The synthetic hexadecasaccharide SR123781 has tailor-made Factor Xa and Thrombin inhibitory activities combined with less specific binding. The molecular interactions of this hexadecasaccharide have been determined from X-ray crystal structures of ternary complexes of AT-III/thrombin (Li *et al.* 2004). This oligosaccharide consists of an AT-III binding domain (S12–S16) at the reducing end of the non-sulfated linker, a non-sulfated linker region (S6-S11), and a thrombin-binding domain (S1–S5) at the non-reducing end of the linker (Figure 1.22). This synthetic oligosaccharide contains methylated hydroxyls and 2-*O*-sulfo substituted glucose in the AT-III binding domain instead of the *N*-sulfo substituted glucosamine (S12–S16) that occurs in the natural pentasaccharide. The highly sulfated glucose units allow non-specific binding to thrombin. This binding to thrombin is thus dependent primarily on the overall charge density of the GAG fragment rather than on a precise sequence of variously substituted sugar residues. All of the monosaccharide units in the AT-III binding domain are in the chair conformation, with the exception of the iduronic acid (S15) at the reducing end, which is in the "skew-boat" conformation. The eight non-sulfated linker region (S6-S11) does not interact with any protein residues, but rather it enhances the formation of the ternary complex giving rise to increased AT-III activity but with minimal interaction with PF4.

PI-88 (Progen) (Figure 1.22) has progressed to clinical trials to treat inflammatory diseases, thrombosis, virus infections and cancer (V. Ferro & Don 2003). PI-88 acts as a substrate analogue to inhibit heparanase activity and so prevents HS degradation. It is targeted for conditions such as tumour cell invasion, metastasis, and angiogenesis. PI-88 is a phosphomannopentose sulphate (6-*O*-PO_3H_2-α-ᴅ-Man-(1→3)-α-ᴅ-Man-(1→3)-α-ᴅ-Man-(1→3)-α-ᴅ-Man-(1→2)-ᴅ-Man), wherein the chain length, sugar composition and glycosidic linkages α1->3 and α1->2 play important roles in its anti-coagulation activity compared to the anti-coagulant activity of sulfated glucose-containing oligosaccharides with β1->4, β1->3 linkages (Wall *et al.* 2001).

A variety of different approaches such as solution-phase and solid-phase chemistry to the polymer-supported synthesis of GAG and non-GAG derivatives has been reported over the years for the development of a large variety of GAGs (Codée *et al.* 2004). Nonetheless, the introduction of non-anionic structural motifs into heparin/HS should provide a route for the development of novel, potent drug-like GAG mimetic molecules to treat various diseases.

SIGNIFICANCE AND AIMS OF THIS STUDY

PECAM-1 is important in the extravasation of leukocytes during inflammation. In PECAM-1 knock-outs, cells (leukocytes) get caught between the endothelium and the basement membrane. It is possible that the interaction of HS with PECAM-1 *in vivo* is critical for both the initial interaction of leukocytes with endothelial cells and the final passaging across the basement membrane. There is a large amount of HS on the endothelial cell surface which could be important for interactions with the extracellular domains of PECAM-1. In contrast to other heparin binding cell adhesion molecules such as NCAM, GAG interactions with PECAM-1 are still controversial. An understanding of the nature of the interactions of GAGs with PECAM-1 will play an important role in the discovery of small molecule selective inhibitors of these interactions. The crystal structure of PECAM-1 has not yet been determined, either on its own or in complex with heparin fragments. Consequently, the use of molecular modelling techniques in this research project provides an alternative route to the investigation of the structure of PECAM-1 and its interactions with GAGs.

The aims of this study are:

- To predict the structure of the extracellular domains of the PECAM-1 molecule using protein modelling techniques

- To identify and characterise the homophilic, heterophilic, metal binding and sulfate binding sites of the PECAM-1 molecule.

- To predict the binding of various GAG fragments to the putative GAG binding sites of PECAM-1 and their associated binding affinities.

- To predict the free energies of binding in aqueous solution of various GAG fragments using molecular dynamics simulation methods.

- To rationalise the structural determinants of binding specificity and selectivity in the interaction of PECAM-1 with various GAG fragments.

Chapter 2

MOLECULAR MODELLING METHODS

A variety of computational approaches have been used in this research to construct a model of the three dimensional structure of PECAM-1 and investigate its intermolecular interactions with GAGs. This chapter provides an introduction to the concepts and methods of homology modelling, fold recognition, ligand-protein docking and molecular dynamics simulations.

2.1 HOMOLOGY MODELLING

One of the main challenges in biochemistry is the ‘protein folding problem’, an understanding of how the overall fold of a protein is determined by its amino acid sequence (Anfinsen 1972). The function of a protein is a consequence of its three-dimensional (3D) structure (i.e. its fold) and hence the determination of its structure is essential. The 3D structure of a protein can be determined through X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, both of these methods are expensive, cumbersome and time-consuming techniques.

A 3D model of the structure of a protein is necessary when an X-ray or NMR structure is not available. Homology modelling or comparative modelling methods are able to predict the 3D structure of a protein sequence by using information derived from homologous proteins of common evolutionary origin whose structures are known. Homology modelling involves combining the sequence of a macromolecule of unknown structure with the structure (template) of another structurally similar macromolecule in order to obtain an approximate model of the structure of the protein of interest. Homology models of proteins are used to understand protein stability and function, perform structure-based drug design and optimisation, or design experiments such as site-directed mutagenesis.

Figure 2.1 outlines certain features that apply to all protein structure prediction methods (Marti-Renom *et al.* 2000), in particular:

- Homology modelling requires the availability of similar structures.

- 40% amino acid identity or higher is best for performing comparative modelling as shown by the conservation in structural folds in protein 3D structure databases.

- 20% to 40% or lower amino acid identity may be of limited value, although successful examples have been reported. With such low amino acid identity it is better to use methods like threading (modelling with folds) (D. T. Jones *et al.* 1992). 20% - 35% amino acid sequence identity is often referred to as the "twilight" zone.

- 0% - 20% amino acid sequence identity is often referred to as the "midnight zone" (previously referred to as "twilight zone" as shown in Figure 2.1). A*b initio* prediction methods can be used. These methods predict structure on the basis of identifying low-energy conformations of the target protein. This field is of great theoretical interest but, so far, of little practical application. However, threading and *ab initio* methods have been applied to the modelling of membrane-bound proteins such as GPCRs (Becker *et al.* 2004; Fleishman & Ben-Tal 2006).

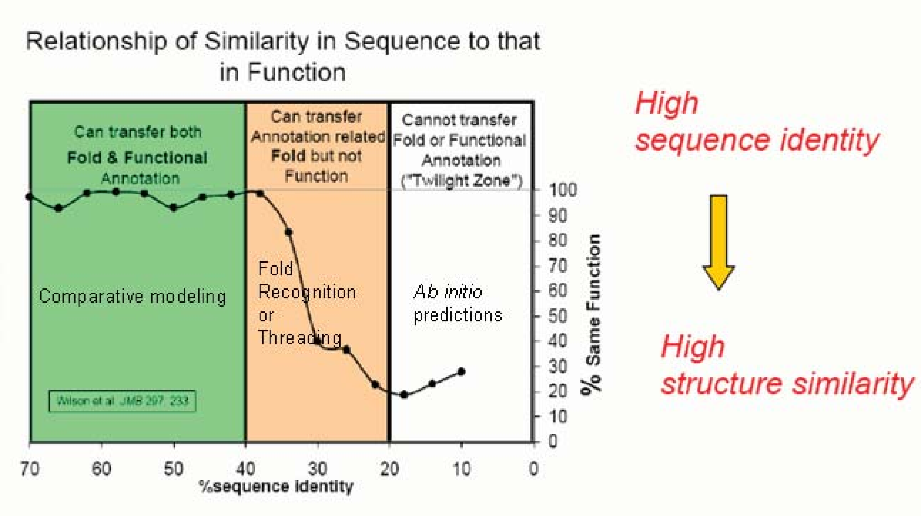


Figure 2.1. *Structure prediction methods. Comparative or knowledge-based modelling is used when sequence identity is greater than 40%, whilst threading or*

fold recognition are preferable when sequence identity is between 20% and 40%. These methods can model 3D structures reliably because structural identity increases with the increase in similarity between the query sequence and the homologous templates. If sequence identity drops below 20% (i.e. the 'midnight' or 'twilight zone') it becomes difficult to predict the structure of the sequence in consideration.

In order to build the 3D homology model of a protein its amino acid sequence is required, along with the high-resolution structure(s) and the sequences of related proteins. Figure 2.2 provides a flow diagram of the standard process of homology modelling. A number of key steps are involved in the construction of the 3D model of a protein using homology modelling:

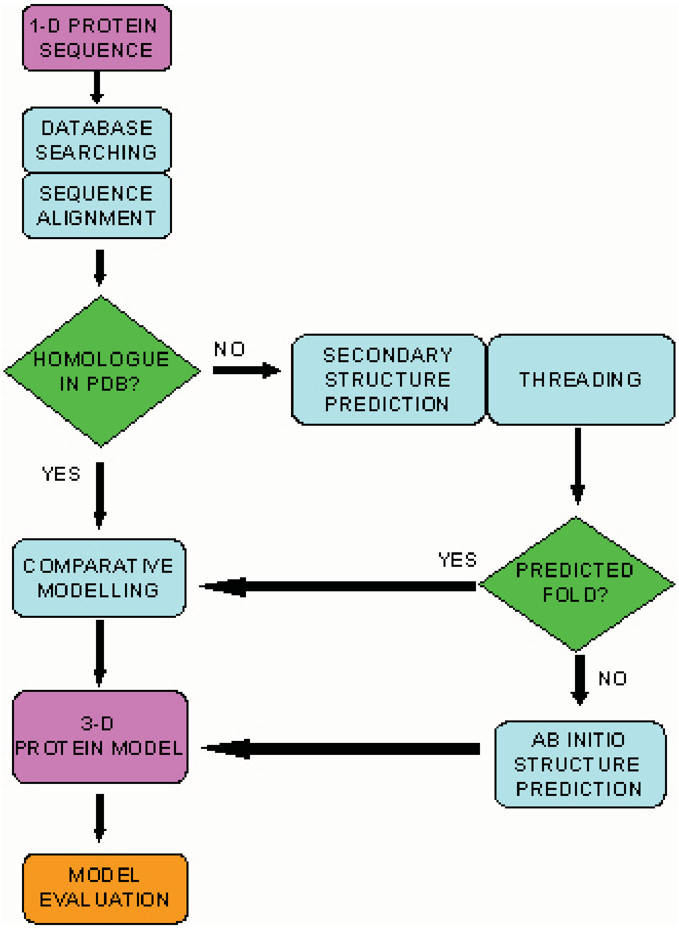


Figure 2.2. *Homology modelling flow chart. A similarity search is performed with the query sequence against a 3D structure database such as PDB. Sequence and*

structural alignments are carried out between the homologous template and the query sequence. The SCRs (structurally conserved regions) and VRs are modeled using comparative modelling followed by optimisation algorithms. Finally, the model is verified using Ramachandran plots and other structure prediction methods. In the absence of a known template from the database, modelling is performed using fold recognition or ab initio methods.

2.1.1 Template detection

Comparative model building of a new (target) protein sequence involves the extrapolation of a known 3D structure of one or more related family members (templates). Comparative protein modelling requires at least one sequence of known 3D structure with significant similarity to the target sequence. BLAST (Altschul *et al.* 1997) and FASTA (Pearson & Lipman 1988) searches against structural databases like the PDB enable detection of homologous templates. Statistical methods implemented in BLAST and FASTA searches can determine the likelihood of a particular alignment between sequences arising by chance given the size and composition of the database being searched. BLAST can also perform genome specific similarity searches and conserved domain database searches. FASTA uses the Smith-Waterman dynamic programming algorithm for protein and nucleotide searches, which are slower but more sensitive when full-length protein sequences are used as queries. As a result FASTA is more specific compared to BLAST algorithm when identifying long regions of low similarity, especially for highly divergent sequences. BLAST is designed to find local regions of similarity whereas FASTA is preferred for global pair wise alignments. Position-specific iterated (PSI)-BLAST (Altschul *et al.* 1997) is the most sensitive BLAST program, which is used for finding very distantly related proteins. PSI-BLAST iteratively expands the set of homologues of the target sequence based on a position-specific scoring matrix (PSSM or profile). This matrix is created from an alignment of the sequences returned with higher score (E-values). This PSSM created in the first iteration becomes the query in the next iteration search. Any new database hits below the inclusion threshold are included in a new PSSM. The search converges when no more new database sequences are added in subsequent iterations.

These similarity searches allow the selection of several suitable templates for a given target sequence in the modelling process. The function of the target protein can also be predicted on the basis of the homology between the target and template sequence. The best template(s) structure(s) is the one with the highest sequence similarity to the target and with better crystallographic parameters (such as resolution) and the completeness of the structure. Such template serves as the reference structure. In the case of distantly related proteins, the sequence alignment may not indicate the correct fold assignment of the target sequence. In this case, the templates are superimposed onto the query sequence depending on their structural folds.

The next step is to predict secondary structure. The aim of secondary structure prediction is to look for patterns of residue conservation that are indicative of known secondary structures, such as alpha helices and beta strands, within a protein or protein family. PSIPRED (Bryson *et al.* 2005) and PredictProtein (Rost *et al.* 2004) are some of the tools available used to predict secondary structure with greater confidence (although the prediction of β-strands is still imperfect).

PSIPRED is a secondary structure prediction method based on two feed-forward neural networks. Sequence similarity searches are made using PSI-BLAST and a final position-specific scoring matrix (PSSM) after three iterations of PSI-BLAST. This PSSM is used as input to a single hidden layer neural network. This method achieved an average Q_3 score of between 76.5% and 78.3% in CASP3 (Critical Assessment of techniques for protein Structure Prediction) because of its ability to predict secondary structure precisely amongst 187 unique folds, achieving the highest published score compared to any other method in that competition (D. T. Jones 1999). Recent versions of PSIPRED average the output from up to four separate neural networks in the prediction process, and achieved an average Q_3 score of 80% in CASP4.

PredictProtein is a secondary structure prediction method that considers various aspects of protein sequence and structure analysis, such as multiple sequence alignments and database search, ProSite sequence motifs, low-complexity regions, ProDom domain assignments, nuclear localisation signals, disulfide bridges, secondary structure, solvent accessibility, globular regions, transmembrane helices and coiled-coil regions. This method uses algorithms based on PHD methods, namely PHDsec (Rost 2001; Rost & Sander 1994) or PROFsec for secondary structure prediction. PHDsec prediction is based on the generation of a pair wise profile-based multiple sequence alignment created by the program MaxHom. This alignment is then fed into a neural network with three layers (input, hidden, and output). The output of the first level (sequence-to-structure network) based on the conversation profile from the neural network is fed into a second level consisting of structure-to-structure network. The three output units in PHDsec code for α-helix, strand, and unconserved regions. PHDsec focuses on predicting hydrogen bonds and, as a consequence, helices may be renamed as strands after highly reliable secondary structure predictions.

2.1.2 Sequence alignment and optimisation

This is the most important step in the construction of the 3D model of a protein. The target sequence needs to be aligned with the template sequence or, if several templates have been selected, with the structurally corrected multiple sequence alignment. This can be achieved by using multiple sequence alignment methods such as ClustalW (Chenna *et al.* 2003). ClustalW uses pair wise alignments between all the input sequences on the basis of similarity using scoring matrices like PAM (Point Accepted Mutation) and BLOSUM (BLOck SUbstitution Matrix), and assigning gap penalties for insertions or deletions in the sequence alignment. The distances are calculated by looking at the non-gapped positions and counting the number of mismatches between the two sequences, and then dividing this value by the number of non-gapped pairs from the alignments. A matrix is formed once all distances corresponding to all pairs have been calculated. ClustalW constructs a similarity tree using this matrix and a neighbour joining algorithm is used to construct the phylogenetic tree. In the end all alignments are clustered on the basis of this progressive guiding tree starting from the closest related groups.

The factors that need to be considered when performing sequence alignments are (1) algorithm used for the alignment, (2) scoring methods applied and (3) assignment of gap penalties. Residues such as those located in non-conserved loops should be modeled after modelling of the conserved regions. For proteins with low homology sequence with the query protein (~<40% percentage sequence identity), the model can be improved by using secondary structure prediction (i.e. align-model-realign-remodel).

The next step is to transfer the coordinates of the atoms in SCRs from the template structure(s) to the target. Methods based on the satisfaction of spatial restraints like MODELLER (Sali & Blundell 1993) are based on generating as many constraints (or restraints) as possible from the structural alignments of the parents and building the target structure, in a similar fashion to NMR structure determination methods (using additional energy restraints according to the correct stereochemistry of the protein chain). MODELLER starts to build a model using distance and dihedral angle restraints on the target sequence derived from its alignment with template 3D structures. Spatial restraints and CHARMM force field terms, which enforce proper stereochemistry, are then combined into an objective function. Restraints can include distances between alpha carbons, other distances within the main chain, and main chain and side chain dihedral angles. Finally, the model is generated by optimising the objective function in Cartesian space. One of the strengths of carrying out modelling in this way is that constraints or restraints derived from a number of different sources can easily be added to homology-derived restraints. It is clear that regions where the structure of the homologous templates cannot be structurally aligned, or where an alignment between the target and the multiple alignments of the templates is not given, need to be built with an additional function. Most of the structural changes are produced in loop regions, but occasional secondary structures may also be involved in variable regions. In the case of multiple superimposed template structures, the coordinates are separated into conserved secondary structural elements and conserved loops.

2.1.3 Modelling of variable regions

Almost every protein model contains non-conserved loops (variable regions), which are expected to be the least reliable portions of a protein model. In most cases, these loops also correspond to the most flexible parts of the structure, as evidenced by high crystallographic temperature factors in template structure(s). After the backbone of the target protein is generated, loops for which no structural information is available in the template structures (non-conserved regions in the alignment which are not defined) need to be constructed. This can be done by finding peptide segments in other proteins that fit into the spatial constraints of the model after a search for high resolution fragments in databases such as the PDB. There are two methods for predicting loop conformations: *ab initio* methods (Moult & James 1986) and database searching techniques or knowledge-based approaches (van Vlijmen & Karplus 1997).

In the case of *ab initio* loop prediction methods, a conformational search or enumeration of conformations in a given environment is carried out guided by a scoring or energy function. There are many such methods which use different protein representations, energy function terms, and optimisation or enumeration algorithms. Search algorithms include sampling of main chain dihedral angles biased by their distributions in known protein structures, the minimum perturbation random tweak method, systematic conformational searches, Monte Carlo simulated annealing, Monte Carlo and molecular dynamics simulations, search of discrete conformations by dynamic programming, random sampling of conformations that rely on dimers from known protein structures, enumeration based on graph theory, etc. (Olson *et al.* 2007; Samudrala & Moult 1998)

Database approaches to loop prediction aim to find a segment of main chain that fits between two stem regions of a loop (Greer 1981). A residue range is chosen to include the undefined loop as well as a few residues (usually three) on either side of the loop for which coordinates have been defined. Segments are examined for their ability to fit in the undefined region without making bad contacts with other atoms but overlapping well with the residues on either side of the loop. The loop may then be subjected to conformational searches to identify low energy conformers. Coordinates for side chain atoms in these loop regions may be copied if the residues are similar, although often side chain rotamer libraries are used to define coordinates in these regions. Loops are modelled from database searches consisting of 1) homologous structures, 2) a cluster database of loops, and 3) a non-redundant database of proteins with less than 25% homology and resolution higher than 2.5 Å (van Vlijmen & Karplus 1997).

The database search is valid only for short and medium sized loops or for special cases where homologous proteins share structural commonalities in the loops although still being considered variable regions. Hybrid methods have been proposed (Martin *et al.* 1989) which use both database search and *ab initio* methods to predict loops in antibodies. CODA (Deane & Blundell 2001) is a combination of two algorithms: FREAD, a knowledge-based method, and PETRA, an *ab initio* method.

Two types of energy functions are supported by MODELER's Loop Refinement: DOPE_Loop (Shen & Sali 2006) and (Fiser) Loop (Fiser *et al.* 2000). Both include bonded terms (bond length, bond angle, main chain dihedral angles, side chain dihedral angles, etc.) but differ mainly in their non-bonded terms. DOPE_Loop uses non-bonded terms such as Lennard-Jones, DOPE (Discrete Optimised Protein Energy) statistical potential (an all-atom potential that computes a residue-by-residue energy profile of a homology model), charge-charge interactions and an electrostatic contribution to the solvation free energy, whereas (Fiser) Loop uses the Melo statistical potential, which is a residue based distance-dependent statistical potential of mean force.

2.1.4 Replacement of template side chains with model side chains

The coordinates of the side chains are transferred to the model if the residue type in the target structure is identical or very similar to that in the known homologues. The number of side chains that need to be built is dictated by the degree of sequence identity between target and template sequences. In the case of disulphide bridges, these are modelled using secondary structural information from proteins using program like PredictProtein and from conserved disulfide bridges in related structures. For other side chains a rotamer library can be used in conjunction with a systematic search to explore possible side chain conformations depending on the associated backbone conformation (Dunbrack Jr & Karplus 1993). The rotamer library generally provides lists of χ_1 and χ_2 angle pairs for residues for given ψ and φ angle values, and explores these pairs to try to minimise side chain-backbone clashes and side chain-side chain clashes. Consequently, a library of rotamers taken from a database of protein structures can be used as an alternative to model the conformations of side chains.

Force field terms can be incorporated for the prediction of side chain conformations to include solvation corrections. Side Chain Refinement modules in Accelrys tools can optimise side chain conformations based on systematic searches of side chain conformations and CHARMm energy minimisation using the ChiRotor algorithm (V. Z. Spassov *et al.* 2007). In this algorithm, the side chain atoms from the residues to be optimised are minimised keeping the backbone fixed. This step is followed by conformational sampling of the side-chains, followed by minimisation using

CHARMm. The lowest two energy conformations are then saved while the atoms of side-chains with higher energy conformations are deleted, and this process is repeated iteratively through all the residues which are selected for optimisation. The lowest energy conformation of all side chains are assembled into the structure framework and energy minimised using CHARMm. The first residue starting from the N-terminus is replaced by the second lowest energy side-chain conformer in the framework and energy minimised. If the second side-chain conformer has lower energy after minimisation, it replaces the first one. For some residues such as Trp, His, Asn, and Gln, in the second cycle an additional rotation is performed corresponding to a change of 180^o in the terminal χ angle, due to the presence of asymmetric groups in the side-chains of these residues. Residues Ala, Gly, Cys in disulfide bridges and Pro are not subject to side-chain refinement but are kept in their original confirmation.

2.1.5 Optimisation of the model

The homology modelling procedure continues with a molecular mechanics minimisation in order to reduce irregularities in the structure and find an optimum molecular geometry. Various energy minimisation algorithms can be utilised, such as the Newton Raphson method, steepest descents and conjugate gradients methods, using force fields such as CFF (Consistent Force Field), CHARMM or AMBER. The refinement of a primary model is initially performed by approximate 100 steps of steepest descents, followed by 200-300 steps of conjugate gradient energy minimisation. This process can be repeated until some convergence criteria are satisfied. Sometimes models optimised by energy minimisation (or molecular dynamics) methods usually change their conformation away from their initial structure. Constraining the positions of selected atoms (such as Cα atoms or the backbone atoms of transmembrane regions in GPCRs) in each residue generally helps to avoid excessive structural drift during minimisations and molecular dynamics simulations (Patny *et al.* 2006).

Optimisation of the model also involves refining the flexible regions formed by a loop. Loop Refinement is a CHARMm based protocol integrated in Accelrys Discovery Studio. The initial stage in the algorithm “LOOPER” (V.Z. Spassov *et al.* 2008) includes a systematic conformational search of loop structures by sampling the backbone phi and psi dihedral angles keeping the rest of the protein fixed. A minimum set of dihedral angles for each residue is chosen based on the lowest energy conformation. The loop is divided into two halves: the N-terminal half and the C-terminal half. Each half has N/2 residues for an even number residue loop, whereas in case of an odd number residue loop, the N-terminal half is one residue longer than the C-terminal. Two conformational states are sampled for each residue [(φ=-90, ψ=120) and (φ=-60, ψ=-40)] in each half with the exception of glycine. Energy minimisation is carried out for half of the loop in absence of the other half, wherein 50 conformations for each half are retained. This step is followed by ranking of the N and C- terminal half loop conformations on the basis of CHARMm energy evaluations. Full loop conformations are constructed from all combinations of the retained half loops by constructing the peptide bond between the two half loops and energy minimising using CHARMm. The side-chain atoms in these loops are positioned using the approach of ChiRotor (V. Z. Spassov *et al.* 2007). Finally, a full energy minimisation of the resulting loop is carried out and the top ranked conformation is retained on the basis of CHARMm energies.

2.1.6 Detection of errors (model verification)

All models built by homology have errors. Side chains can be placed incorrectly, whole loops can be misplaced or novel folds may not be predicted correctly. In the latter case, the model will be more similar to the template than the real structure. It has thus been necessary to develop criteria with sufficient discriminatory power to distinguish a good model from a bad one. The quality of protein models can be assessed by measuring the root mean square deviation from the crystal structure, the proportion of main chain conformations in acceptable regions of the Ramachandran plot, the presence of planar peptide bonds, the existence of side chain conformations that correspond to those in the rotamer library, the presence of hydrogen bonds between polar atoms if they are buried, the presence of proper environments for hydrophobic and hydrophilic residues, and the lack of bad atom-atom contacts. These parameters can be evaluated using the program WHAT IF (Vriend 1990).

Analyses of Ramachandran plots (PROCHECK) (Laskowski *et al.* 1993) as well as Profile 3D and Verify 3D methods can be used to evaluate the quality of a protein model. PROCHECK is based on an analysis of φ/ψ angles, peptide bond planarity, bond lengths, bond angles, hydrogen-bond geometry, and side chain conformations through a comparison with the expected values of these parameters obtained from known protein structures, as a function of atomic resolution of the structures from which the model was developed.

The Profile 3D method is based on the statistical preferences of each of the 20 amino acids for particular environments within the protein. Preferred environments for amino acids are derived from known three-dimensional structures and are defined by three parameters: (1) the area of each buried residue, (2) the fraction of side chain area that is covered by polar atoms (*i.e.*, O and N), and (3) the local secondary structure. Based on these environment variables, a 3D structure is converted into a 1-D profile that describes each residue in the folded protein structure. Examination of these profiles reveals the regions of a sequence that appear to be folded correctly.

Verify 3D is used to evaluate sequence-structure compatibility in a crystal structure or homology model with at least 100 residues (Eisenberg *et al.* 1997). Verify 3D assesses the environment of the 3D structure or model based on the solvent exposed side chains. The compatibility score of the sequence to the structure segments (1D to 3D profile) is plotted for all residues according to their sequence number. Scores are averaged over a 21-residue window. The Verify 3D scores below or near 0.0 reflect structures that are almost certainly incorrect, whilst scores near 1.0 reflect scores similar to those expected for a valid protein of the same size.

2.1.7 Iteration over all steps to remove errors

After verification of the structure using the above methods, a final step in the construction of a protein model may required an iterative process through the steps outlined in the flowchart in Figure 2.2, particularly errors in template selection or a correction of the sequence alignments.

2.2 THREADING OR FOLD RECOGNITION

Homology modelling becomes increasingly unreliable when the sequence identity between two proteins falls to 40% or less. In this case, threading or fold recognition methods can be more useful for assessing protein sequence-structure compatibility. The terms threading and fold recognition are frequently used synonymously. In these methods, a protein sequence of interest is firstly used to search a database of known protein structures with the aim of finding the overall protein fold that the sequence is likely to adopt. Apart from these methods, a database search using fragments can be carried out using the query sequence against the known protein(s) to identify a number of structurally conserved regions/motifs or regions of well defined secondary structure motifs or signatures (i.e. helices or strands). This method is sometimes referred to as fragment based homology modelling (Kolodny *et al.* 2002).

Fold recognition methods can be broadly divided into two types:

1. Methods that uses the information like secondary structure or accessible surface area of the query protein - 2D Threading or Prediction Based Methods

   This method uses databases such as DSSP (Database of Secondary Structures for Proteins), containing sequences, secondary structures and solvent accessible surface area (Rost *et al.* 1997). The method tries to align the query sequence against the database using dynamic programming and ranks the alignment in accordance with the fold. The performance of these algorithms increases if the database size is considerably large. 2D fold recognition methods are much faster than the 3D counterparts but the limitation of this method is that it cannot produce a 3D model at the end of the process.

2. Methods that consider the full 3D structure of the protein template - 3D Threading or Distance Based Methods (DBM).

   In the 3D representation, the structure is modelled as sets of interatomic distances i.e. the distances are calculated between some or all of the atom pairs in the structure (Bryant & Lawrence 1993). This is a much better description of the structure, but these methods generate poor sequence alignments. This method can be based on profile or PSSM. The method based on profile is the 1D-3D profile (Bowie *et al.* 1991; Luthy *et al.* 1992). This method derives a 1D profile for each structure in the fold library and aligns the target sequence to these profiles. A simple example of a profile representation takes each amino acid in the structure and labels it according to 18 structural environments on the basis of secondary structure, solvent accessibility and burial by polar atoms. Scoring functions are then used to verify the statistical significance of the profiles with the preliminary sequence in the homology model or the known structure. Those sequences that produce high compatibility scores are likely to be structurally related to the probe (3D profile), and regions having low scores are likely to be placed where the backbone has been incorrectly modelled. The disadvantage of this method is that the structural environments or residue classes do not define the structural context, e.g. amphiphatic helices or strands with the distant sequence homologues.

Web-servers such as 3D-PSSM and Phyre (Protein Homology/analogY Recognition Engine) perform a profile-profile matching algorithm (PSSM, Position Specific Scoring Matrices) together with predicted secondary structure matching (Kelley *et al.* 2000). These methods generate structural alignments of homologous proteins using three passes of a global dynamic programming algorithm to search the structural classification of proteins (SCOP) database. The resulting multiple alignment based on superfamily classification is converted into a PSSM. The score (residue equivalence) for a match between a residue in the query sequence and a residue in the library sequence is calculated as the sum of the secondary structure, solvation potential and PSSM scores. Each iteration of search differs in the PSSM used for the scoring, with secondary structure and solvation held constant. Finally, the program uses a statistical parameter-E value, which is a measure of confidence in the prediction. Combined with secondary structure matching and solvation potentials, 3D-PSSM and Phyre can confidently model proteins undetectable by PSI-BLAST (Bennett-Lovsey *et al.* 2007). 3D-PSSM or Phyre is also used for large scale annotation of genomes. Recently, Phyre has been developed as an ensemble (meta or cluster) fold recognition system (Bennett-Lovsey *et al.* 2007; Kelley *et al.* 2000). A protein query sequence is processed by a pool of fold recognition algorithms such as profile-profile or sequence-profile to detect homologies from the SCOP database as described in 3D-PSSM. This results in formation of a pool of candidate protein structural models. These models are clustered according to one of the SVM (Support Vector Machines) and Greedy protocols.

2.3 FORCE FIELD METHODS

A force field is a term used to describe the functional form and parameter sets of the molecular mechanics potential energy of a molecular system. ‘All atom’ force fields contain parameters for every atom in a molecule, including hydrogens. The set of parameters describe the force constants and equilibrium lengths and angles of chemical bonds, as well as partial charges and VDW interaction parameters. These parameters are usually derived from experimental and *ab initio* quantum mechanical calculations (Burkert & Allinger 1982). The energy, E, is a function of the atomic positions, R, of all the atoms in the system. The energy (equation 1) is calculated as a sum of 'bonded' terms (E_{bonded}), which describe the bonds, angles and conformations in a molecule (equation 2), and of non-bonded terms ($E_{non\text{-}bonded}$) shown in equation 3, which describe electrostatic and VDW interactions (Burkert & Allinger 1982).

$$V(R) = E_{bonded} + E_{non\text{-}bonded} \tag{1}$$

$$\text{Where } E_{bonded} = E_{bond} + E_{angle} + E_{torsion};\ E_{non\text{-}bonded} = E_{electrostatic} + E_{vdW} \tag{2}$$

Equation 2 can be written as the sum of all the bonded (equations 3, 4 and 5) and non-bonded (equation 6) interactions (refer to Figure 2.3).

$$E_{bond} = \Sigma \frac{K_b}{2}(r - r_0)^2 \tag{3}$$

Equation 3 describes the interaction between atom pairs separated by a covalent bond through a harmonic potential (following Hooke's law). This equation is an approximation to the energy of a bond as a function of deviations from its ideal bond length, r_0. r is the length of the bond (*i.e.,* the distance between the two nuclei of the atoms). The force constant, K_b, determines the steepness of the potential, controlling how difficult it is to stretch a bond.

$$E_{angle} = \Sigma \frac{K_\theta}{2}(\theta - \theta_0)^2 \tag{4}$$

Equation 4 describes the harmonic potential associated with the alteration of a bond angle theta θ (the angle between two bonds) from its ideal value θ_0. The values of θ_0 and K_θ depend on the chemical type of atoms constituting the angle.

$$E_{torsion} = \Sigma \frac{K_\phi}{2}\left(1 + \cos\left(n\phi - \gamma\right)\right) \tag{5}$$

Equation 5 describes the torsional potential of a dihedral angle. Such angles are assumed to be periodic and are often expressed as a cosine function, which models the periodic presence of steric barriers between atoms separated by three covalent bonds. K_{Φ} represents the energy barrier to rotation, n is the multiplicity (the number of maxima or minima in one full rotation), Φ is the torsion angle and γ determines the angular offset.

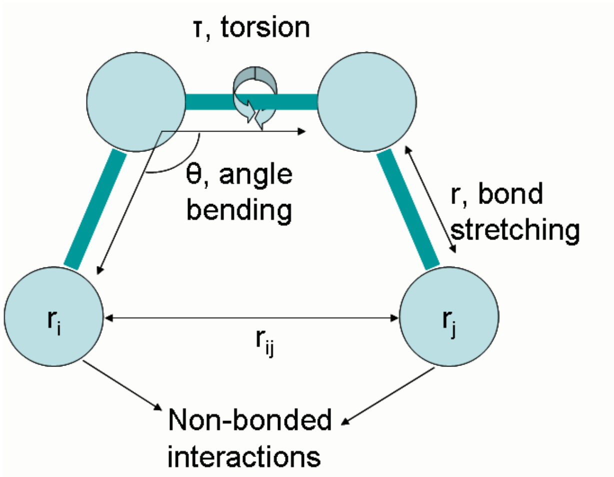


Figure 2.3. *Energy terms in a molecular mechanics force field. The total energy of the system is given by the sum of bonded interactions (bond stretching, bond angle bending and torsional changes) and non-bonded terms such as VDW and electrostatic interactions.*

VDW interactions are most often modelled using the Lennard-Jones potential, which expresses the interaction energy as a sum of a repulsive and an attractive term, using atom-type dependent constants A and B. The electrostatic interaction between a pair of atoms r_i and r_j is represented by Coulomb's equation. ε_0 is the permittivity of free space, ε_r, is the relative dielectric constant of the medium in which the charges are placed, and r_{ij} is the separation between two atoms having charges q_i and q_j.

$$E_{VDW} + E_{electrostatic} = \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \frac{(q_i \, q_j)}{4\Pi\varepsilon_0 \varepsilon_r r_{ij}} \qquad (6)$$

A force field can thus describe the energy of a molecule as a function of the coordinates of its atoms. Some of the most popular force fields are:

AMBER (Assisted Model Building and Energy Refinement): This force field uses a united atom approach (wherein non-polar hydrogen atoms are not represented explicitly), and was developed by Peter Kollman and his group at the University of California, San Francisco (Cornell *et al.* 1995). The parameter sets for proteins and DNA in AMBER is referred to as “ff94” or “ff99”. GAFF (Generalized AMBER force field) provides parameters for small organic molecules (J. Wang *et al.* 2004). Parameters for carbohydrates have been developed in the form of the GLYCAM force field by Robert Woods (Woods *et al.* 1995). The AMBER force field functional form is:

$$V = \sum_{bond} \frac{K_r}{2}(r - r_0)^2 + \sum_{angles} \frac{K_\theta}{2}\left(\theta - \theta_0\right)^2 + \frac{1}{2}\sum_{torsion} K_\phi\left(1 + \cos\left(n\phi - \gamma\right)\right) + \sum_{non-bonded}\left[\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \frac{\left(q_i\ q_j\right)}{4\Pi\varepsilon r_{ij}}\right] + \sum_{H-bonds}\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) \tag{7}$$

where K_r, K_θ *and* K_ϕ are the respective force constants, r is the bond length, r_0 is the equilibrium bond length, θ is the angle value, θ_0 is the equilibrium bond angle value, φ is the dihedral angle value, n is the periodicity, q_i and q_j are the charges of the atoms, ε is the effective dielectric constant and r_{ij} is the distance between two atoms.

GLYCAM: The GLYCAM force fields (Woods *et al.* 1995) has been developed for oligosaccharides but lacks parameters for charged carbohydrates such as GAGs. Dihedral angle parameters and ensemble-averaged partial atomic charges are derived by selecting 100-200 conformations from a solvated MD run (Basma *et al.* 2001). The initial charges are derived after a quantum mechanical energy optimisation of the geometries at the HF/6-31G* level, keeping exocyclic torsion angles fixed in their MD conformation. RESP (Restrained ElectroStatic Potential fit) charges are used to reproduce the electrostatic potential and dipole moment of the molecule using basis set theory (Bayly *et al.* 1993). The RESP method involves generation of multiple conformations of molecule to perfom the fiting of the charges on the atom. GLYCAM versions 93, 94 and 2000 were augmented with AMBER parameter sets. The recent version GLYCAM06 (Kirschner *et al.* 2007) has parameter sets applicable to any stereoisomer and to all monosaccharide ring sizes and conformations, and is independent of AMBER force field parameters.

CHARMM (Chemistry at HARvard Macromolecular Mechanics): This molecular force field and dynamics program was developed by Martin Karplus and his group at Harvard (Brooks *et al.* 1983). The CHARMM22 and CHARMM27 sets of parameters are used for the simulation of protein, DNA and lipids. The commercial version of this force field implemented in Accelrys programs is referred to as CHARMm. CHARMM uses the following energy function:

$$E_{potential} = E_{bond} + E_{angle} + E_{torsion} + E_{oop} + E_{UB} + E_{elec} + E_{vdw}$$

$$\begin{aligned}
&= \sum_{bond} K_b\left(b - b_0\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_0\right)^2 + \sum_{torsion} K_\phi\left(1 + \cos\left(n\phi - \gamma\right)\right) + \\
&\sum_{oop} K_\omega\left(\omega - \omega_0\right)^2 + \sum_{UB} K_\upsilon\left(\upsilon - \upsilon_0\right)^2 + \\
&\sum_{non-bonded} \varepsilon\left[\left(\frac{R_{\min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{\min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{\left(q_i\ q_j\right)}{\varepsilon_1\, r_{ij}}
\end{aligned} \tag{8}$$

where E_{oop} is the out-of-plane potential or improper torsion used to select the correct geometry or chirality of atoms, E_{UB} is the Urey-Bradley component (cross-term accounting for angle bending using 1,3 nonbonded interactions). $K_b, K_\theta, K_\phi, K_\omega$ and K_υ are the respective force constants, b is the bond length, b_0 is the equilibrium bond length, θ is the angle value, θ_0 is the equilibrium bond angle value, φ is the dihedral angle value, n is the periodicity, ω is the improper angle value, ω_0 is the ideal improper angle value, υ is UB 1,3 distance, υ_0 is the ideal UB 1,3 distance, ε is the Lennard-Jones well depth, R_{minij} is the distance at the Lennard-Jones minimum, q_i and q_j are the charges of the atoms, ε is the effective dielectric constant and r_{ij} is the distance between two atoms.

CFF (Consistent Force Field): This force field was developed by Halgren and the Biosym Consortium based upon *ab initio* quantum mechanical calculations on small molecules (Hagler & Ewig 1994). The CFF force field uses quartic polynomials for bond stretching and angle bending, and uses three-term Fourier expansion for torsions. VDW interactions are represented by an inverse 9th power term for repulsive behaviour instead of the inverse Lennard-Jones 12th power term.

2.4 ENERGY MINIMISATION METHODS

The potential energy surface (PES) is a mathematical representation of the total potential energy of a molecule or molecular system as a function of its coordinates and resulting from all interactions between the atoms (Leach 2001), as described in the previous section. A complex PES can be compared with a mountain range having energy barriers (peaks), energy minima (valleys), and saddle points (passes). Energy minimisation is performed on macromolecular structures and small molecules to relax their conformation and remove any steric overlap that produces bad contacts (high energy states) between atoms. Different minimisation algorithms can be applied on a molecule to find an energetically preferred conformation of a molecule (which will rarely be the absolute energy minimum). These algorithms are classified according to the order of the Taylor series expansion of the energy as a function of all coordinates (x):

$$U(x) = u(x_k) + (x - x_k)\frac{\partial U(x_k)}{\partial x_k} + \frac{(x - x_k)^2}{2}\frac{\left(\partial^2 U\right)(x_k)}{\partial x_k \, \partial x_j} + \dots \qquad (9)$$

where the second term in Equation 9 is known as the gradient (force) and the third term is known as the Hessian (force constant). The most important first and second order energy minimisation methods (Leach 2001) are the steepest descent, conjugate gradient and adapted basis set Newton-Raphson methods.

Steepest Descent (SD): This is a very simple and first derivative minimisation method. It considers only the current location of the coordinates from iteration to iteration, resulting in the gradient and the direction of successive steps parallel to each other. SD is superior to other methods when the starting structure is far from the minimum, as it can rapidly move away from high energy conformations. However, this method converges very slowly to a local minimum in a complex PES and, consequently, it is used mostly in the early stages of a minimisation algorithm to remove unfavorable steric contacts.

Conjugate Gradient (CONJ): This is also a first order derivative method. It is an iterative method that makes use of the previous history of minimisation steps and the current gradient to determine the next step. The gradients and the directions of successive steps are orthogonal. This method exhibits better convergence than the steepest descent method. A variation of the conjugate gradient method with improved efficiency is called the Powell algorithm.

Adapted Basis-set Newton-Raphson (ABNR): This is a second order derivative method. It performs energy minimisation using a Newton-Raphson algorithm applied to a subspace of the coordinate vector spanned by the displacement coordinates of the last positions. The second derivative matrix is constructed numerically from the change in the gradient vectors, and is inverted by an eigenvector analysis that allows the algorithm to recognise and avoid saddle points in the energy surface. At each step, the residual gradient vector is calculated and used to add a steepest descent step, incorporating the new direction into the basis set.

The performance of these algorithms is monitored on the basis of convergence criteria (a sufficiently small change in the energy gradient and RMS gradient), the number of minimisation steps carried out within a specific time, and the memory storage requirement. SD requires little memory whereas Newton-Raphson methods require lot of memory and computation. SD or CONJ is usually recommended for the initial minimisation of a system, followed by a few steps of ABNR (if there is enough memory).

2.5 MOLECULAR DOCKING

The non-covalent binding of a small molecule to its target protein receptor is one of the most common biomolecular interactions. Molecular docking is a computational method that seeks to model interactions between the ligand and protein and make predictions about the geometry of the ligand-protein complex (the ‘binding mode’) and the associated free energy of binding. Most proteins contain pockets, cavities, surface depressions or grooves where small molecules can easily bind. The small molecule or substrate complements the shape and physio-chemical properties of the binding/active site of a protein/enzyme. In the case of enzymes, binding of a substrate to the active site leads to the catalysis a particular chemical reaction and this term in biochemistry is called “docking”. This is based on the “lock-and-key” principle postulated in 1894 by Emil Fischer, wherein receptors (in analogy with lock) and ligands or substrates (analogy with the keys) fit together tightly on the basis of structural (shape) and interaction complementarities. The “lock and key” model is rigid model but most often the substrate plays a role in determining the final shape of the enzyme inducing partial flexibility within the enzyme (Koshland 1995). This flexible model is called induced fit theory. Both these theories form the basis of docking algorithms.

The most common technique used in many docking programs is the shape complementarity method, which aims to match the receptor and the ligand by finding an optimal binding pose (DesJarlais *et al.* 1988). Structural complementarity may require matching of the solvent accessible surface area and of overall shape and geometric constraints between atoms in the protein and ligand. Interaction complementarity takes into account hydrogen bonding interactions, hydrophobic contacts and VDW interactions, in order to describe how well a particular ligand binds to a protein.

The challenge in molecular docking is to search for and accurately predict the binding mode of a ligand and its associated binding affinity. The correct binding mode of a ligand molecule may be found after extensive sampling of the conformational space of the molecule in the protein binding site. The typical output of a docking program is a set of ligand binding modes ranked by a docking score. The docking score is assigned by a scoring function, which should be able to distinguish the correct binding mode from other putative modes. The protein-ligand complex with the highest ranking score should then resemble the actual (observed) binding mode.

There are a variety of methods that have been developed to perform conformational searches, such as incremental step construction, as implemented in PatchDock, and simulated annealing and genetic algorithms, as integrated in AutoDock.

2.5.1 PatchDock

This is a rigid molecular docking algorithm for small molecule-protein, protein-protein and antibody-antigen interactions based on shape complementarity principles (Duhovny *et al.* 2002; Schneidman-Duhovny *et al.* 2005; Schneidman-Duhovny *et al.* 2003). It is used for protein-protein docking based on the identification of ’hot spot’ residues. It is also used for docking of antigen-antibody molecules on the basis of the complementarity-determining regions (CDRs) or hyper variable (HV) regions. The

CDRs are detected in PatchDock by aligning the sequence of a given antibody to a consensus sequence of a library of antibodies.

PatchDock makes use of a geometry-based molecular docking algorithm that aims to find docking transformations that yield good molecular shape complementarity. The algorithm divides the Connolly dot surface representation of the molecules into concave, convex and flat patches. These complementary patches (convex, concave and flat) are calculated using a hybrid of the geometric hashing and pose-clustering matching techniques in order to generate rigid molecule candidate transformations. Each candidate transformation is further evaluated by a scoring function that considers both geometric fit (the complementarity molecular shape score) and atomic desolvation energy (C. Zhang *et al.* 1997). Finally, a clustering method can be applied to the candidate solutions based on the complementarity molecular shape score in order to reduce the number of potential solutions.

PatchDock uses a high density representation of the molecular surface using Connolly's MS Surface algorithm (Connolly 1983) and a low density representation, using a sparse surface for the unbound docking of rigid molecules. The denser surface is used to detect steric clashes and for fine geometric scoring. The algorithm divides the receptor into shells (according to the distance from the molecular surface) for primary scoring of the transformations. The number of surface points in each shell is counted at each stage of a candidate transformation. The geometric score is a weighted average of all the shells, with a preference for candidate complexes with a large number of points in the outer shell, and a lower preference for possible points in the 'penetrating' inner shells.

A PatchDock run reports the surface area, atomic contact energy, distance transformations and the geometric fit score of the ligand-protein complex. The atomic contact energy (ACE) is a desolvation free energy score (C. Zhang *et al.* 1997) based on the method of Miyazawa and Jernigan (Miyazawa & Jernigan 1985) with improvements. It is defined as the free energy of replacing a protein-atom/water contact, by a protein-atom/protein-atom contact. The ACE scores were obtained for all pairs of 18 atom types as observed in case of 91 representatives protein monomeric structures. The total ACE score of a protein (equation 10), ΔE_C, is calculated in dimensionless RT units and can be given as the difference between two terms: the total number of atom-water contacts in the fully solvated conformation and the total number of atom-water contacts in the native conformation:

$$
\begin{aligned}
\Delta E_C = {} & E_C(\text{native structure}) - E_C(\text{solvated conformation}) \\
& = \sum_{i=1}^{18} e_{i\,,p}\, n_{ir\,,p} \\
& = \sum_{i=1}^{18} e_{i\,,p} \frac{q_{e\,,i}^{\;0} \left(n_{i\,,p} \right)}{2} - \sum_{i=1}^{18} q_{i\,,p}\, n_{i0\,,p} \\
& = e_{v\,,p} \sum_{i=1}^{18} e_{i\,,p} \frac{\left(q_{e\,,i}^{\;0} \right) \left(n_{i\,,p} \right)}{2} - e_{s\,,p}\, n_{r0\,,p} \qquad (10)
\end{aligned}
$$

where different terms are defined for an individual protein *p*. e_i is the average contact energy for the *i* th atom type of the protein *p*, n_r is the total number of solute atoms, n_i is the number of atoms of type *i*, n_{ir} is the total number of solute contacts made by atom type *i*, n_{r0} are the total numbers of solute-solvent contacts, q_i is the coordination number for residues of type *i*, e_v and e_s are the contact energies averaged, respectively over all the atom-water contacts in the solvated state and over all the atom-water contacts in the native state.

2.5.2 AutoDock

AutoDock is a set of docking algorithms developed at the Scripps Research Institute and the University of California at San Diego. AutoDock uses three docking protocols: simulated annealing (SA), genetic algorithm (GA) and Lamarckian genetic algorithms (LGA).

SA was the first method used for optimisation in AutoDock (Morris *et al.* 1996). This algorithm translates the ligand from an arbitrary point in space into the protein binding site through a series of translation and rotation steps. The ligand is treated as a flexible entity, whilst the protein target remains rigid. SA docking is a global optimisation technique based on the Metropolis Monte Carlo method. During each constant temperature cycle of simulated annealing, random changes are made to the ligand's current position, orientation, and conformation (if flexible). The new configuration is then compared to its predecessor. If its new energy is lower than the previous, this new configuration is immediately accepted. However, if the energy of the new configuration is higher than the previous one, it is accepted with a probability given by its Boltzmann factor (equation 11):

$$P(\Delta E) = e^{(\Delta E / KT)} \tag{11}$$

where *ΔE* is the difference in energy between the new and previous configurations, and *K* is the Boltzmann constant. This probability depends upon the energy and cycle temperature. Each cycle contains a large number of individual steps which are accepted or rejected upon the current temperature. After a specified number of acceptances or rejections, the next cycle begins with a lower temperature as specified by equation 12:

$$T_i = g\, T_{i-1} \tag{12}$$

where T_i is the temperature at cycle *i* , and *g* is a constant between 0 and 1. At high temperatures, many high energy configurations will be accepted, whilst at low temperatures, the majority of these configurations will be rejected. In general, this method performs a global energy search when the high temperature allows the exploration of the PES of the interaction to predict the energy minima, and performs a local search at low temperature. This method performs better than steepest descent where SA is able to accept all the low-lying energy states of the clusters in the narrow valley with probability P as per the equation 11 that are rejected by SD.

GA constitutes a general purpose optimisation method that works by mimicking the process of Darwinian evolution. GA is used for multidimensional global search problems where the search space potentially contains multiple local minima. An advantage of GAs over many search or optimisation algorithms is that derivatives of the scoring functions are not required.

All living organisms consist of cells. In each cell there is the same set of chromosomes. A chromosome (large portions of DNA) consists of genes which encode proteins. The genome encodes all of the physical characteristics of the organism, known as the "phenome". A particular set of genetic information is a "genotype", and likewise a particular set of physical characteristics, or "traits", is a "phenotype".

During reproduction, genes are transferred from parents to their offsprings through recombination or crossover. The new created offspring can have mutations, which are. mainly caused by errors in copying genes from parents. The suitability of a given organism to its environment is usually measured by its "fitness" in analogy with the idea of the "Survival of the fittest", as introduced by Darwin. Computationally, it is usual to evaluate the "fitness" of an organism directly, without considering any kind of phenome.

In an optimisation algorithm, a chromosome contains information about the system that it represents. In AutoDock, the phenotype is described by the set of Cartesian co-ordinates of the protein-ligand complex, whilst the genotype encodes information describing how to put together the ligand and protein into a bound complex. The particular arrangement of a ligand and a protein can be defined by a set of variables describing the translation, orientation and conformation of the ligand with respect to the protein. They are composed of a 3D translational vector, Eulerian rotation angles and a collection of torsion angles that describe bond rotations in the ligand and protein. These values correspond to a ligand's state variables and each state variable is referred to as one gene. The configuration of the ligand can now be referred to as genotype, whilst the conversion of the configuration into atomic co-ordinates defines its phenotype. A docking simulation generates a series of generations. Each generation is composed of a population of individuals, i.e. protein-ligand complexes. A population of different genes is generated at random, and each is scored using a fitness function such as the AutoDock energy function. Genes are selected to form the next population based on the scoring function, with better scoring poses or conformations more likely to be selected. Pairs of the selected genes or poses are allowed to cross over with each other in order to gradually find a better solution.

A basic GA applied for molecular docking (G. Jones *et al.* 1995b) in general, is outlined below:

1. A set of reproduction operators such as crossover and mutations are chosen and each operator is assigned weight.

2. An initial population is randomly created and the fitness of its members is determined.

3. An operator is chosen using a selection algorithm based on the weights of the operators.

4. The parents required by the operator are selected using selection algorithms based on scaled fitness.

5. The operator is applied and the child chromosomes are produced. The fitness of the offspring is evaluated.

6. If not already present in the population, the children replace the least fit members of the population.

7. If an acceptable solution is found, stop, or else go to step 3.

The characteristics of the GA that have a major impact on the outcome are the implementation of crossover and the fitness factor (see further below). GA implemented in AutoDock also follows the basic genetic algorithm but differs in terms of few parameters as described below:

Generations/Initial Populations: The initial population for a GA optimisation is usually chosen at random in AutoDock. For each random individual in the initial population, a translation is assigned on the basis of the randomly distributed values between the minimum and maximum x, y, and z extents of the grid maps, an orientation is assigned on the basis of a quaternion having a random vector and a random rotation angle between -180° and 180°, with torsion angles assigned random values between -180° and 180°. The generation of populations continues until the maximum number of generations or the maximum number of energy evaluations is reached, whichever is encountered first.

Fitness factors: Setting the values of the fitness factors in GA involves assigning a value, its probability, to one or more strings (strings in AutoDock is referred to as the set of genes in the chromosomes consisting of three cartesian coordinates for the ligand translation; four variables defining a quaternion specifying the ligand orientation and one real-value for each ligand torsion, in that order) as a measure of improvement of the solution compared to other strings that result from reproduction, crossover and mutation. The fitness function in AutoDock is the docking energy function, which is the sum of the intermolecular interaction energy between the ligand and the protein, and the intramolecular energy of the ligand.

Selection: Chromosomes are selected from the population to become parents to reproduce, and these are selected according following Darwinian evolution ('survival of the fittest'). Hence, the best chromosomes survive and create new offspring. In AutoDock, the selection is made in accordance with:

$$n_o \, \frac{f_w f_i}{f_w <f>} \; f_w <f> \tag{13}$$

where n_o is the integer number of offspring to be allocated to the individual; f_i is the fitness of the individual (i.e. the energy of the ligand); f_w is the fitness of the worst individual (i.e the ligand with the highest energy) in the last N generations (number of docking runs) and $<f>$ is the mean fitness of the population. If the numerator in equation 13, $f_w f_i$, is greater than the denominator $f_w <f>$ then such individuals will be allocated at least one child and will be able to reproduce.

Cross-over: The cross-over in GA refers to swapping of a single bit of a chromosome, which is more like a single-point mutation, or of several bits, where a distinction is made between the two parents (bit strings, chromosomes) as being identical, different and single parent. Cross-over can occur at multiple sites. Two-point crossover is used in AutoDock with breaks occurring only between genes in parents chromosomes resulting into three pieces. For example, ABC and abc are the set of genes in each parent chromosomes. The chromosomes of the resulting offspring after twopoint crossover would be AbC and aBc.

Mutation: After a crossover is performed, a mutation takes place. This is to prevent all chromosomes in a population from falling into a local optimum. Mutations change randomly the new offspring. The mutation depends on the encoding as well as the crossover. Mutation in AutoDock is performed by adding a random real number that has a Cauchy distribution to the real variables (i.e the translational, orientational, and torsional genes are represented by real variables in AutoDock).

Other Parameters: The user-defined integer parameter *elitism* determines how many of the top individuals automatically survive into the next generation. If the *elitism* parameter is non-zero, the new population that has resulted from the proportional selection, crossover, and mutation is sorted according to its fitness.

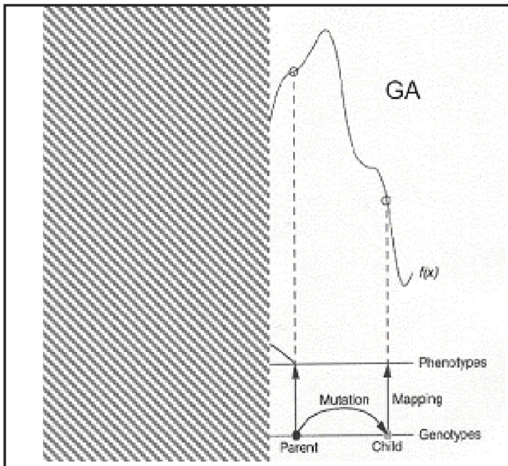
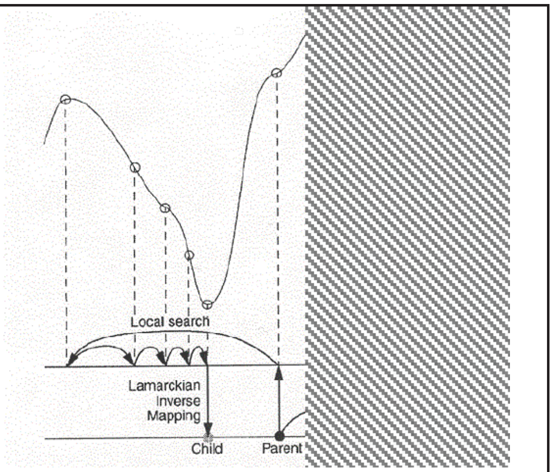
There are two primary parameters concerning the behaviour of genetic algorithms: the crossover rate (Cr) and the mutation rate (Mr). The crossover rate controls the frequency with which the crossover operator is applied. If there are N individuals (population size=N) in each generation then, in each generation, N*Cr individuals will undergo crossover. The higher the crossover rate, the more quickly new individuals are added to the population. If the crossover is too high, high-performance individuals are discarded faster than selection can produce improvements. A high crossover rate of about 80%-95% for GAs is often recommended. On the other hand, a low crossover rate may stagnate the search due to loss of exploration power. Mutation is the operator that maintains diversity in the population. A genetic algorithm with a mutation rate too high will become a random search. After the selection phase, each bit position of each individual in the intermediate population undergoes a random change with a probability equal to the mutation rate Mr. Consequently, approximately Mr*N*L mutations occur per generation, where L is the length of the chromosome. A low mutation rate is recommended for chromosomes with binary encoding. Best rates reported are about 0.5%-1%.

The rate of genetic crossover is set to zero in AutoDock. The rate of genetic mutation is increased compared to the rate of genetic crossover. Another important parameter is the population size. This defines how many chromosomes exist in a population (in one generation). If there are too few chromosomes, the GA has a small chance to perform crossover and only a small part of search space will be explored. If there are too many chromosomes, the GA will slow down. Population sizes of 50-100 are recommended.

AutoDock uses both Darwinian and Lamarckian inheritance (Morris *et al.* 1998). A LGA is similar to a standard (Darwinian) GA except that each conformation or chromosome is subjected to energy minimisation before scoring. The LGA decodes the chromosomes and optimises obtained parameters, i.e. the phenome. The optimised version is scored and the genetic data is re-evaluated. That is, during the 'life' of a conformer or chromosome, it may experience local changes so that some characteristics can be transmitted to its offspring. Since the local search alters the phenotype of the strings, and is then recorded into its genotype, it is described as a

Lamarckian process. The differences between traditional GA and LGA are outlined in Table 2.1 (Blansché *et al.* 2005).

Table 2.1. Differences between genetic algorithms and Lamarckian genetic algorithms. Figures adapted from Morris *et al*. (1998).

 GA f(x) Phenotypes Mutation Mapping Genotypes Parent Child	 Local search Lamarckian Inverse Mapping Child Parent
1. In standard GA, the genotype (x, y, z coordinates and rotational/torsional angles) are mapped onto the fitness function f(x). 2. New generation based on parent's genes. 3. Genotypes of parents with high f(x) values are mutated, forming child genotypes with lower f(x) values.	1. LGA finds lowest fitness function (energy) values first and then maps these values onto their respective genotypes. 2. Each new child is allowed to create a new generation. 3. It also includes Soils and Wets local search. 4. Better performance than GA or simulated annealing algorithm.

The traditional and Lamarckian GAs in more recent versions can handle ligands with more degrees of freedom than the SA method used in earlier versions of AutoDock.

The LGA algorithm implements an adaptive global optimiser with local search. The local search method is based on the optimisation algorithm of *Solis and Wets* (*SW*), which is independent of gradient information (Solis & Wets. 1981). The local search modifies the phenotype, which is then allowed to update the genotype. The SW local search uses fixed variances for probabilistically determining the change to a particular state variable, like a translation or rotation. These variances are either doubled or halved during the search, depending on the number of consecutive successful or failed moves resulting in a drop in energy. AutoDock also has a modified version of SW called pseudo-Solis and Wets (*pSW*) to take into account in the variances the relative magnitudes of translations and rotations.

An energy evaluation is performed every time the GA or the local search computes the fitness of a candidate docking. AutoDock stops a docking simulation if either the maximum number of evaluations or the maximum number of generations is reached, whichever occurs first. The number of energy evaluations needed for a docking simulation will depend on the number of torsion angles in the ligand (and receptor, if it is flexible). For rigid ligands and rigid receptors some general guidelines are outlined in Table 2.2. In the case of docking simulations of highly flexible molecules, such as carbohydrates, 'ga_num_evals' is set to a very large number as observed in the case of blind docking (Hetenyi & van der Spoel 2002). In the case of blind docking, it is important to increase 'ga_pop_size' from 50 to 300 in steps of 50 whilst keeping other parameters constant. The most robust docking results are obtained with a population size of 300. The AutoDock authors recommend to run at least 50 docking runs, specified by the 'ga_run' parameter.

Table 2.2. Recommended values for AutoDock parameters for docking rigid ligands to rigid receptors.

Number of Torsions	**ga_num_evals**	**ga_num_generations**
0	25 000 to 250 000	27 000
1-10	250 000 to 25 000 000	27 000
>10	>25 000 000	27 000

AutoDock uses a force field scoring function to provide a fast calculation of the potential energy term in the free energy of binding (Morris *et al.* 1998). The rigid receptor is represented as a potential energy grid and an atom is treated as a probe. For each atom type, charge, and placement within the grid, an energy value is computed, according to the scoring function (as described below). In version 3.0 of AutoGrid and AutoDock, the scoring function is based on the principles of quantitative structure-activity relationships (QSAR) and it was parameterised using a large number of protein-ligand complexes for which both their structures and inhibition constants K_i were known. AutoDock employs a molecular mechanics term (equation 19) in the scoring function with solvation and entropy terms:

Molecular Mechanics Terms:

- VDW (Lennard-Jones 12-6 attraction/repulsion)

$$\Delta G_{vdw} = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \qquad (14)$$

  where A_{ij} / r_{ij}^{12} is the repulsive term and - B_{ij} / r_{ij}^{6} is the dispersion interaction, both shown to vary according to the inverse powers of the distance between the two atoms $r_{ij.}$

- Hydrogen Bonding

$$\Delta G_{Hbond} = W_{hbond} \sum_{i,j} \mathrm{E}\,(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{rij^{10}} + E_{hbond} \right) \qquad (15)$$

  where *E(t)* is a directional weight based on the angle, *t*, between the probe and the target atom and r_{ij} = distance between atoms *i* and *j*.

- Electrostatics – according to Coulomb's Law

$$\Delta G_{elec} = W_{elec} \sum_{i,j} \frac{q_i\, q_j}{\varepsilon\left(r_{ij}\right) r_{ij}} \qquad (16)$$

  where q_i and q_j are the magnitude of the charges and r_{ij} is their separation.

- Desolvation (AutoDock 3)

$$\Delta G_{desolv} = \sum_{i_c, j} \left(S_i V_j \, e^{-\frac{r_{ij}^2}{2\sigma^2}} \right) \qquad (17)$$

  where *i* and *j* are the index of atoms, S_i = solvation term for atom *I* , V_j = atomic fragmental volume of atom *I*, r_{ij} = distance between atom *i* and atom *j* (in Å) and σ = gaussian distance constant = 3.5 Å

The estimated change in torsional free energy when a ligand goes from an unbound to a bound state is calculated as:

- Torsional

  $\Delta G_{tor} = N_{tor}$ (non-H rotors) (18)

Finally, the AutoDock scoring function can be expressed (equation 16) as

$$\begin{aligned}\Delta G = \Delta G_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j} \mathrm{E}\,(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{rij^{10}} + E_{hbond} \right) + \\ \Delta G_{elec} \sum_{i,j} \frac{q_i\, q_j}{\varepsilon \left(r_{ij} \right) r_{ij}} + \Delta G_{solv} \sum_{i_c, j} \left(S_i V_j \; e^{-\frac{r_{ij}^2}{2\sigma^2}} \right) + \Delta G_{tor} N_{tor}\end{aligned} \qquad (19)$$

The inhibition constant (Ki) is calculated from the computed free energy of binding, ΔG_{bind}:

$\Delta G_{bind} = RT \ln K_i$ (20)

where R is the gas constant, 1.987 cal K^{-1} mol^{-1}, and T is the absolute temperature (assumed to be room temperature, 298.15 K).

Various terms in the molecular mechanics energy function in Autodock 3.0 and 4.0 have been re-scaled by coefficients that were derived using linear regression analysis. New terms include the desolvation free energy of the ligand and an estimate of the loss of conformational degrees of freedom of the ligand upon binding. The parameters for the different terms are self-consistent and are based on the AMBER force field for the protein and any metallic ions, along with the GLYCAM force field for the carbohydrate. The values used in the AutoDock 3.0 scoring function are given as the VDW coefficients and well depth energies, e, scaled by 0.1485, the electrostatic energy scaled by 0.1146, and the hydrogen bonding terms scaled by 0.0656. The new terms for loss of torsional degrees of freedom upon binding and the ligand desolvation free energy are scaled by 0.3113 and 0.1711, respectively. The torsional term is actually the number of rotatable bonds in the ligand involving heavy atoms, multiplied by the coefficient, 0.3113. Hydroxyl groups are not considered.

2.6 MOLECULAR DYNAMICS SIMULATIONS

Molecular dynamics (MD) allows the study of the fluctuations and conformational changes of large macromolecules such as proteins, nucleic acids (DNA, RNA), membranes, as well as small molecules. MD simulations track the time evolution of a molecular system after the atoms are given an initial velocity and are allowed to move according to the forces exerted upon them. The time evolution of the system is computed by solving numerically the equations of motion according to the laws of Newtonian mechanics (Leach 2001).

In Newtonian mechanics, the force *F* exerted on a particle is equal to the produce of the mass *m* of the particle and the acceleration *a*: *F*=*ma*. It is thus possible to determine the acceleration of each atom in a system using the above equation if the forces can be calculated. Integration of the equations of motion yields a trajectory (a series of 'snapshots' of the configuration of all atoms) that describes the positions, velocities and accelerations of the particles at each time step.

$$F_i = m_i\, a_i = m_i \frac{dv_i}{dt} = m_i \frac{d^2 x_i}{dt^2} \tag{21}$$

Equation 21 denotes Newton's second law of motion for *i* atoms in a system where *m* is the mass and *a* is the acceleration, which is given by the rate of change of velocity, which is in turn the rate of change of displacement.

The force can also be expressed as the gradient of the potential energy:

$$F_i = -\Delta_i\, V \tag{22}$$

where *V* is the potential energy of the system. Combining equations 18 and 19 yields the derivative of the potential energy, which relates to the change in position as a function of time.

$$-\frac{dV}{dx_i} = m_i \frac{d^2 x_i}{dt^2} \tag{23}$$

Many force fields such as AMBER and CHARMM have been developed to approximate the forces exerted on atoms using classical mechanics. The position of all atoms in a system thus gives complete information about the system's forces and energy, whilst the velocities of the atoms are determined by the forces.

The initial positions of the atoms may be taken from experimental structures, such as the X-ray crystal structure of a protein or the crystal lattice of ice. The initial velocities, v_i, are often chosen randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature (equation 24).

$$p\left(v_{ix}\right)=\left(\frac{m_i}{2\Pi k_{bT}}\right)^{\frac{1}{2}} e^{\frac{-m_i v_{ix}^2}{2k_b T}} \tag{24}$$

where p is the probability that an atom i of mass m_i has a velocity v_x in the x direction at a temperature T and k_b is the Boltzmann constant.

In general, the positions (equation 22), velocities (equation 23) and accelerations (equation 24) can be approximated by a Taylor series expansion:

$$r\left(t+\delta t\right)=r\left(t\right)+v\left(t\right)\delta t+\frac{1}{2}a\left(t\right)\delta t^2+\ldots \tag{25}$$

$$v\left(t+\delta t\right)=v\left(t\right)+a\left(t\right)\delta t+\frac{1}{2}b\left(t\right)\delta t^2+\ldots \tag{26}$$

$$a\left(t+\delta t\right)=a\left(t\right)+b\left(t\right)\delta t+\ldots \tag{27}$$

where r is the position, v is the velocity (the first derivative with respect to time), a is the acceleration (the second derivative with respect to time), etc. Integrating the equations of motions can be computationally expensive for a large macromolecular system and various integration methods have been developed for this purpose.

2.6.1 Molecular Mechanics/Poisson Boltzmann Surface Area (MM/PBSA) method

Electrostatic interactions, solvation effects and hydrophobic interactions play a crucial role in ligand-protein and protein–protein binding. Contributions to the free energy of binding arising from these interactions can be calculated using continuum solvation models based on Poisson-Boltzmann (PB) calculations of the solvation energy and solvent accessible surface area (SA) approximations to hydrophobic hydration (Reddy & Erion 2001). These methods can be used in conjunction with MD simulations to accurately model the behaviour of a macromolecule in aqueous solution. The MM/PBSA (Molecular Mechanics-Poisson-Boltzmann Surface Area) and MM/GBSA (Molecular Mechanics-Generalized Born Surface Area) methods have been developed to calculate the free energy difference between two states, such as the bound and unbound states of two solvated molecules, or two different solvated conformations (Fogolari *et al.* 2003). The MM/PBSA/GBSA approach has been successfully applied to study protein-peptide/protein interactions (Massova & Kollman 1999), protein-ligand interactions (Chong *et al.* 1999), protein-carbohydrate interactions (Goodford 1985; Laitinen *et al.* 2003), protein-nucleic acid interactions (Reyes & Kollman 2000) and protein folding (M. R. Lee *et al.* 2000). The MM/PBSA method has also been used to compute changes in the free energy of binding upon alanine scanning mutagenesis between mutant and the wild-type protein complexes (Laitinen *et al.* 2003).

The MM/PBSA is referred to as continuum solvation method because calculations of the electrostatic contribution to the solvation free energy are carried out assuming that the molecule (i.e protein or small molecule) can be modelled as a dielectric continuum of low polarisability embedded in a dielectric continuum (the solvent) of high polarisability. The solvent is represented as a homogeneous continuum with a dielectric constant of 80, which is taken to be equal to the value for pure water. The solute (i.e. protein or ligand) is represented by a dielectric constant between 1 and 20. The protein dielectric constant can vary depending on the protein. The electrostatic contribution to the solvation free energy arises from the non-uniform charge distribution between the solute charge distribution and the dielectric continuum, which is calculated using the Poisson–Boltzmann equation given below:

$$-\Delta\left[\varepsilon\left(\overrightarrow{r}\right)\Delta\phi\left(\overrightarrow{r}\right)\right]=4\pi\rho_{solute}\left(\overrightarrow{r}\right)+\underset{i}{\mathrm{E}}z_i^0\;\overset{\frac{-z_i\phi\overrightarrow{\quad(r)}}{kT}}{e} \qquad (28)$$

where k is the Boltzmann constant, T is the temperature, ϕ is the electrostatics potential, ρ_{solute} is the fixed charge density and ε is the dielectric constant. ε, ρ and ϕ are all functions of position vector r.

In the MM/PBSA method the different contributions to the free energy of binding are calculated on the basis of a thermodynamic cycle (Massova & Kollman 1999). It is of utmost importance to calculate all terms that contribute to the free energy of binding, such as the solvation energies of the ligand, protein and complex, the interaction energy between the protein and the ligand, the vibrational and conformational entropy changes in ligand and protein upon complex formation.

Figure 2.4 shows the thermodynamic cycle for the binding of an enzyme, *E*, and an inhibitor, *I*, in both the solvated phase and *in vacuo*. The solvent molecules are indicated by filled circles. Solvent molecules tend to be ordered around the larger molecules, but when *E* and *I* bind, several solvent molecules are released and become disordered. This is an entropic effect and is the basis of the hydrophobic effect. The solvent ordering around *E* and *I*, when both bound and unbound, is strongly influenced by the hydrogen bonding between these molecules. These hydrogen bonds between solvent and *E,* and solvent and *I,* contribute to the enthalpy stabilisation of the complex.

Since the free energy is function of state, the change in free energy between two states is the same regardless of the path between the two states. Hence the free energy of binding in solution can be expressed as

$$\Delta G_{binding,\ sol} = \Delta G_{binding,\ vaccum} + \Delta G_{sol\ (EI)} - \Delta G_{sol\ (E+I)} \qquad (29)$$

The $\Delta G_{binding,\ vaccum}$ can be calculated using molecular mechanics force fields, whilst the free energy changes upon solvation for the separate molecules *E* and *I*, and for the complex, *EI*, $\Delta G_{sol(EI)}$ and $\Delta G_{sol(E+I)}$ respectively, can be calculated by the above continuum methods. This allows the calculation of the free energy change upon binding of the inhibitor to the enzyme in solution, $\Delta G_{binding,sol}$. The inhibition constant, *Ki*, for the inhibitor, *I*, can also be estimated.

The electrostatic contribution to the solvation free energy can be calculated by either solving the linearised Poisson Boltzmann (PB) or the Generalised Born (GB) equation for each of the three states. The GB equation calculates the solvation free energy by assigning effective radius (Born radii) to each atom as shown in Equation 27. It captures the physics of the Poisson-Boltzmann equation, whist improving the speed of calculations (Bashford & Case 2000).

$$\Delta G^{elec}_{\varepsilon_p - \varepsilon_w} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\sum_{ij} \frac{q_{iq_j}}{\sqrt{\left(r_{ij}^2 \alpha_i \alpha_j\right) e^{\frac{-r_{ij}^2}{F\alpha_i \alpha_j}}}} \qquad (30)$$

where ε_p and ε_w are the interior and exterior dielectric constants, respectively, r_{ij} is the distance between atoms *i* and *j*, and α_i is the so-called generalized Born radius of atom *i*. *F* is the empirical factor, which modulates the Gaussian factor scaling the Born radii. It may range from 2 to 10, with 4 being the most commonly used value. The atomic Born radius is the distance of a given charge location from the solvent boundary; for atoms at the center of a spherical cavity.

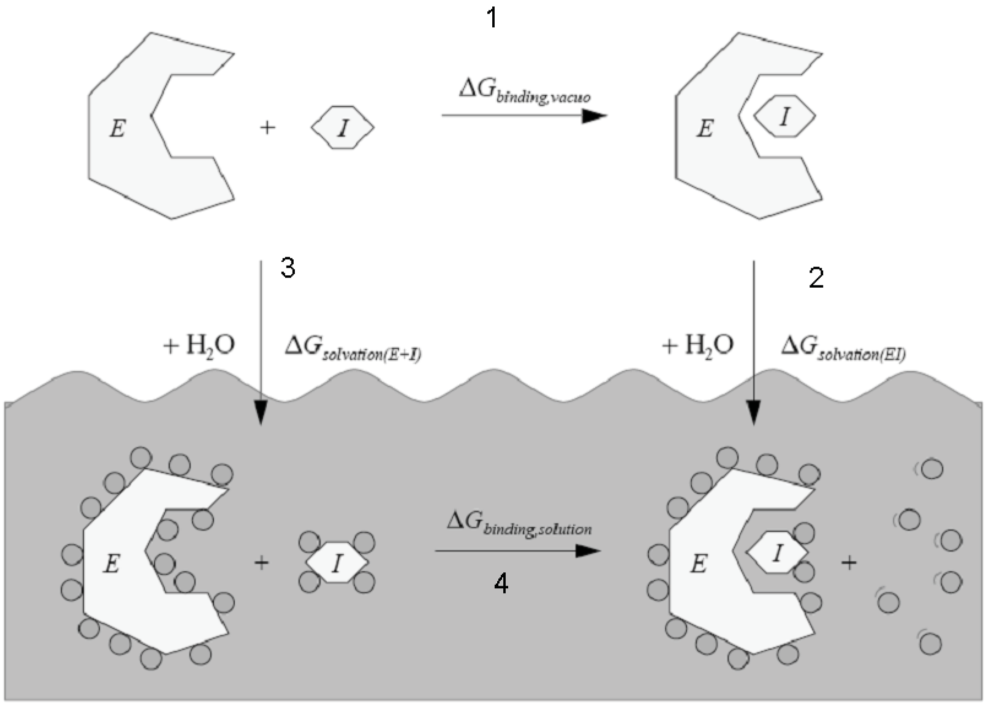

1
E
+
I
$\Delta G_{binding,vacuo}$
E
I
3
2
$+ H_2O$
$\Delta G_{solvation(E+I)}$
$+ H_2O$
$\Delta G_{solvation(EI)}$
E
+
I
$\Delta G_{binding,solution}$
4
E
I
+

Figure 2.4. *The free energy of binding for a ligand-receptor complex is determined by the thermodynamic equation:* $\Delta G_{binding,\ solution} = \Delta G_{binding,\ vaccum} + \Delta G_{sol\ (EI)} - \Delta G_{sol}$

$_{(E+I)}$ where E is the receptor, I is the ligand and EI is the ligand-receptor complex. In the figure, 4 = 1 + 2 – 3. The figure was adapted from AutoDock 3.0.5. User Guide.

An empirical term for the hydrophobic hydration component can then be added as shown in equation 28.

$$\Delta G_{sol} = G_{electrostatic,\, \varepsilon = 80} - G_{electrostatic,\, \varepsilon = 1} + \Delta G_{hydrophobic} \qquad (31)$$

where ε is the permittivity of the medium, also known as dielectric constant.

ΔG_{vacuum} is obtained by calculating the average molecular mechanics interaction energy between receptor and ligand and the vibrational entropy change upon binding.

$$\Delta G_{vaccum} = \Delta H_{molecular\ mechanics} - T\Delta S \qquad (32)$$

The vibrational entropy change ΔS is approximated from quasi-harmonic models which assume the protein to be a system of coupled harmonic oscillators. The vibrational entropy can be computed by performing normal modes analysis on the three species. In practice vibrational entropy contributions can be neglected if, for example, two ligands binding to the same protein are compared, as they will have similar vibrational entropies. Normal mode analysis is computationally expensive and provides an approximation to the vibrational entropy. The average interaction energies of receptor and ligand are usually obtained by performing calculations on an ensemble of uncorrelated snapshots collected from an equilibrated MD simulation.

The MM/PBSA and MM/GBSA methods are integrated in MD simulation packages like AMBER to evaluate free energies of binding in solution. The electrostatic contribution to the solvation free energy is calculated with the Poisson-Boltzmann method implemented in the DelPhi program (Gilson & Honig 1987) or with the program GB in AMBER 8.0 and 9.0, which uses the generalised Born equation to estimate the electrostatic contribution to the solvation free energy (Jayaram *et al.* 1998). The *mm_pbsa* script in AMBER can be used to analyse a simulation trajectory by extracting the energies of the species of interest and calculating the corresponding PBSA energies. Vibrational entropies can be computed using the *nmode* module in AMBER.

Chapter 3

HOMOLOGY MODELLING OF THE EXTRACELLULAR DOMAINS OF PECAM-1

The preceding chapter provided an overview of the theoretical background of the various molecular modelling techniques that have been used in this work. This chapter describes the use of homology modelling, threading methods and other techniques to construct a three-dimensional model of PECAM-1. This structural model is discussed in detail along with predictions of the likely heparin/HS binding regions of PECAM-1.

3.1 HOMOLOGY MODELLING OF PECAM-1

Sequences of the human, mouse, pig, rat and porcine PECAM-1, as well as various alternatively spliced isoforms, were retrieved from the Swiss-Prot protein sequence database (Boeckmann *et al.* 2003). The sequences of human, mouse, pig, rat and bovine PECAM-1 correspond to P16284, Q08481, Q95242, Q3SWT0 and P51866 Swiss-Prot accession numbers. Multiple sequence alignment was performed with ClustalW (J. D. Thompson *et al.* 1994a) using BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) matrices in order to quantify the sequence similarity between individual subunits of PECAM-1 in different species (Figure 1.3) and different isoforms (Figure 1.4). Initially, a PSI-BLAST (Altschul *et al.* 1997) search against the Protein Data Bank (PDB) was performed in order to find sequences that were homologous with human PECAM-1, so that proteins of known structure could be identified and used as a global template for standard homology modelling. Proteins with the required sequence similarity (>35%) were not found following a global search using the entire extracellular region of PECAM-1 or when individual PECAM-1 Ig-domains were considered. In local sequence similarity searches, the identities of the aligned sequences varied from 18% to 23%.

Different sequence analysis tools were used for analysis of the PECAM-1 secondary structural features. Secondary structure prediction algorithms such as PredictProtein

(Rost *et al.* 2004) and PSIPRED (Bryson *et al.* 2005) were used to predict the residues comprising the different subunits. The Ig-domain sequences, as classified by Swiss-Prot (see Table 1.1), were submitted to the fold recognition servers Phyre (Kelley *et al.* 2000) and CBS Meta Server (Douguet & Labesse 2001a). These servers predicted various I set topologies (Harpaz & Chothia 1994) for Ig folds with various templates, including VCAM-1 (E. Y. Jones *et al.* 1995a; J. H. Wang *et al.* 1995), NCAM (Kasper *et al.* 2000; Soroka *et al.* 2003), ICAM-1 (Casasnovas *et al.* 1998), CEA (Boehm *et al.* 2000) and different isoforms of FcγR (Maxwell *et al.* 1999; Powell *et al.* 1999; Sondermann *et al.* 1999a; Sondermann *et al.* 1999b; Sondermann *et al.* 2001; Y. Zhang *et al.* 2000). The crystal structures of VCAM-1, NCAM and ICAM-1 confirmed the existence of the I-type topology in these cell adhesion molecules (Kasper *et al.* 2000).

In the past, various sequence similarity searches were performed to determine relationships between PECAM-1 and various cell adhesion molecules of the Ig superfamily (Ig-CAM) (Newman *et al.* 1990; Simmons 1990; Stockinger 1990). Similarly, we have performed alignments of the Ig-domains of PECAM-1 with crystal structures of known Ig folds using LALIGN/PLALIGN (Pearson & Lipman 1988) and PAM 120 matrices. This allowed us to include new structural information available from time to time in the subsequent stages of our modelling studies. The statistical significance of an alignment was computed by aligning the two sequences and then shuffling the second sequence between 200 and 1000 times using the PRSS module (Pearson & Lipman 1988).

A preliminary sequence analysis with the collected sequences of PECAM-1 and its relatives was carried out in order to investigate its evolutionary relationships and, consequently, we chose to derive the models with the above templates. The Ig-domain 1 of PECAM-1 was modelled with NCAM (Ig-domain 2) and VCAM (Ig-domain 1) as templates. Ig-domain 2 of PECAM-1 was modelled using the structural features of NCAM (Ig-domain 3) and VCAM (Ig-domains 1 and 2). The Ig I set topology in domain 3 of PECAM-1 was modelled with multiple templates including CD8 and ICAM. Extracellular domains 4-6 of PECAM-1 showed preference for various folds of FcγR, as predicted by fold recognition programs like Phyre (Kelley *et al.* 2000). The alignment between the PECAM-1 sequence and the template obtained from Phyre

(Kelley *et al.* 2000) was used to build the global alignment. The signal, transmembrane and cytoplasmic regions were not modelled due to a lack of proper protein folds. The amino acid numbering from Swiss-Prot was retained label the PECAM-1 sequence.

Threading and comparative modelling techniques were used to model the extracellular domains of PECAM-1. Firstly, individual models of Ig-domains were obtained using the information obtained from Phyre (Kelley *et al.* 2000). A complete structure of the human PECAM-1 model was obtained by merging all the individual extracellular domains together by constructing the loops spanning the two domains. Assignment of six disulfide bridges, optimisation and visualisation were carried out using DS Modelling 1.7 (Accelrys, Inc). Loops were built using the loop modelling protocol implemented in MODELLER (Fiser *et al.* 2000; Sali & Blundell 1993). Essential hydrogens and charges were added to the structure. Different side chains rotamers of residues Asp 38, Asp 60, Lys 77, Asp 78 and Lys 116 were searched and replaced in order to make these residues exposed to the surface. The metal coordination site in Ig-domain 6 was modelled with a metal-oxygen distance of between 2.15 and 2.25 Å, which is a key characteristic of metals bound to proteins (Harding 2006). Energy minimisation of the modelled structure was carried out in order to remove any unfavourable interactions. The CHARMm force field (Brooks *et al.* 1983) was used with the smart minimiser method, which begins with the steepest descent method and is followed by the conjugate gradients method until the gradient reached a value below 0.001 kcal/mol. This was followed by molecular dynamics simulations (with a non-bonded cut-off of 10 Å, a dielectric constant of 4, at a temperature of 300K for 20 ps using a time step of 1 fs) with the backbone atoms of the Ig-domains kept fixed. The structural quality of the resultant protein structure was tested using PROCHECK (Laskowski *et al.* 1993), Eval23D (Douguet & Labesse 2001b) and Verify3D (Douguet & Labesse 2001b). Electrostatic potential calculations were done using the DELPHI program (Gilson & Honig 1987) implemented in DS Modelling 1.7 (Accelrys, inc.) using the atomic partial charges assigned by CHARMm with a protein interior dielectric constant of 4, a solvent dielectric constant of 80 and an ionic strength of 0.145 M.

3.2 SURVEY OF THE SULFATE BINDING REGIONS

The PDB was surveyed for sulfate binding motifs using BLAST searches for short overlapping segments for the six Ig-domains. It was assumed that a preliminary search for sulfates in the known crystal structures having similar residue composition to that of Ig-domains of PECAM-1 might give an indication of likely binding sites of GAGs having charged sulfate groups attached to their pyranose rings. For added precision, the Ig-domains were split into a set of overlapping fragments 15 amino acids long, each overlapping by 5 amino acids. For each fragment a sequence similarity search was performed from the PDB. Each hit from the similarity search for PECAM-1 fragments was checked for sulfate interactions in the protein of corresponding crystal structures in PDBsum (Laskowski *et al.* 2005).

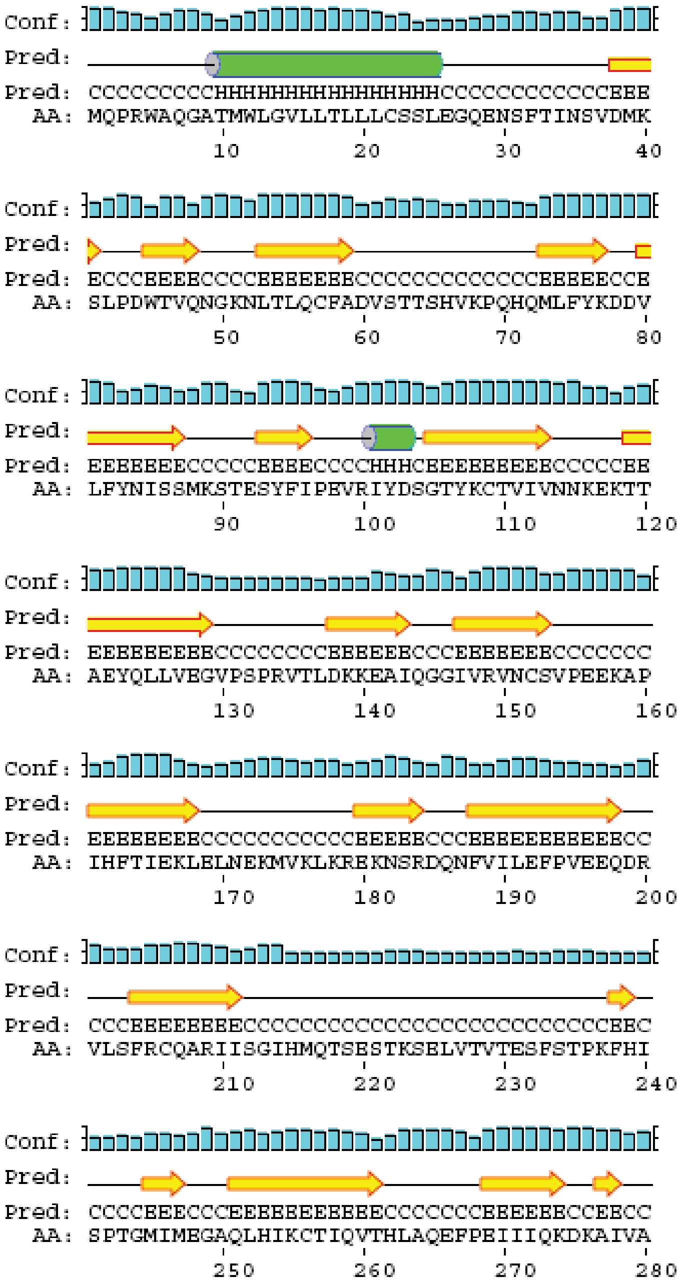
3.3 RESULTS

A standard homology building procedure was adopted to construct a three-dimensional model of PECAM-1, starting with similarity searches and followed by solvent accessibility composition (core/surface ratio) and secondary structure predictions, as shown in Table 3.1 and Figure 3.1, respectively. The predictions of secondary structure correlated well with the known annotation of human PECAM-1 sequence from Swiss-Prot. PECAM-1 Ig-domains are predominantly classified as all beta proteins.These predictions also suggest the presence of a helix in the N-terminal and transmembrane regions, and between two beta sheets in Ig-domains 1, 3 and 6. The presence of a smaller percentage of alpha helix is also observed in the crystal structures of NCAM (PDB codes 1QZ1 and 1EPF). BLAST searches detected conserved Ig regions in domains 1, 4 and 6. The percentage identity of various alignments was found to be very low, as shown in Table 3.2**.**

Table 3.1. Predicted secondary structure composition and solvent accessibility composition for human PECAM-1 using PredictProtein (Rost *et al.* 2004).

	Secondary structure type			**Solvent accessibility composition**	
	Helix	**β-sheet**	**Loop**	**B**[*]	**E**[*]
% in protein	3.93	40.79	55.28	45.53	54.47

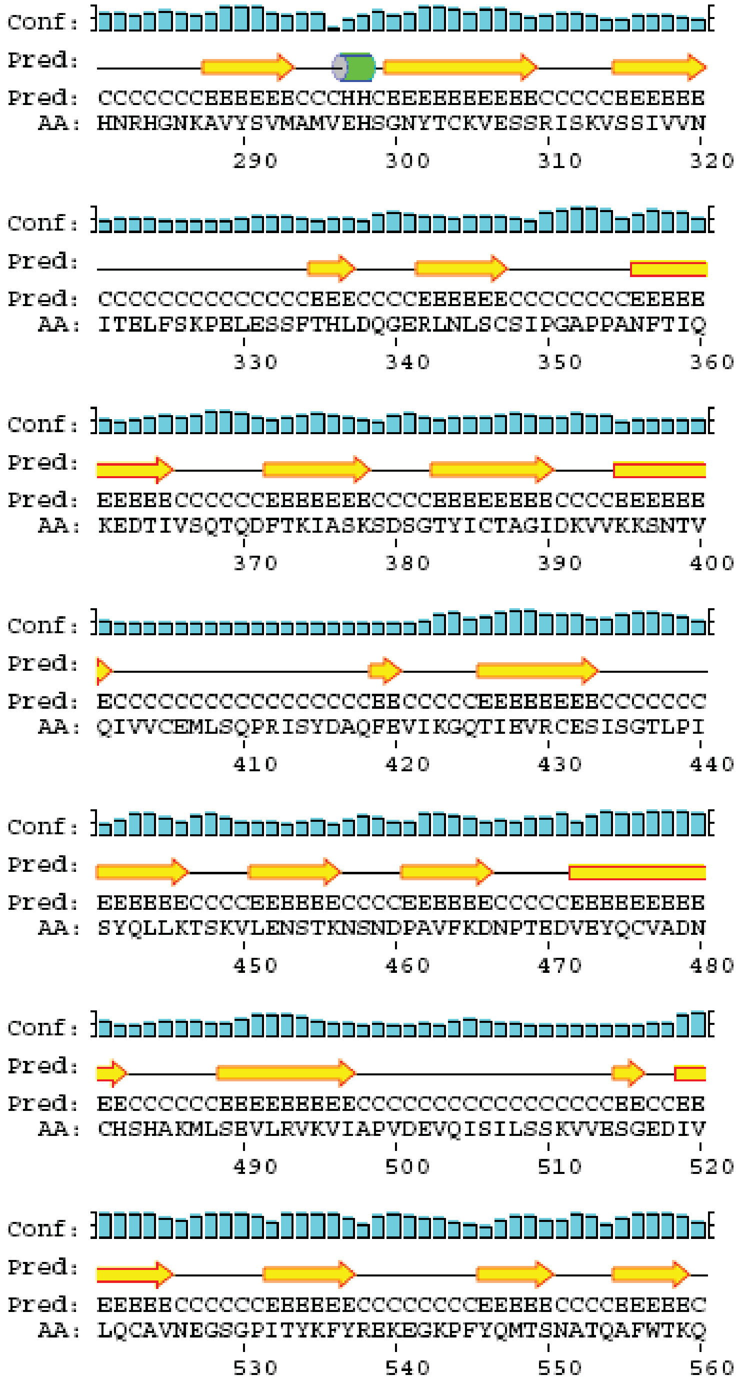
[*] Classes used: E: residues exposed with more than 16% of their surface; B: all other residues.



```
Conf:
Pred:
Pred: CCCCCCCCCHHHHHHHHHHHHHHHHCCCCCCCCCCCCEEE
  AA: MQPRWAQGATMWLGVLLTLLLCSSLEGQENSFTINSVDMK
                10        20        30        40

Conf:
Pred:
Pred: ECCCEEEECCCCEEEEEEEECCCCCCCCCCCEEEEEECCE
  AA: SLPDWTVQNGKNLTLQCFADVSTTSHVKPQHQMLFYKDDV
                50        60        70        80

Conf:
Pred:
Pred: EEEEEEECCCCCEEEECCCCHHHCEEEEEEEEECCCCCEE
  AA: LFYNISSMKSTESYFIPEVRIYDSGTYKCTVIVNNKEKTT
                90        100       110       120

Conf:
Pred:
Pred: EEEEEEEEECCCCCCCCEEEEEECCCEEEEEEECCCCCCC
  AA: AEYQLLVEGVPSPRVTLDKKEAIQGGIVRVNCSVPEEKAP
                130       140       150       160

Conf:
Pred:
Pred: EEEEEEEECCCCCCCCCCCEEEEECCCEEEEEEEEEEECC
  AA: IHFTIEKLELNEKMVKLKREKNSRDQNFVILEFPVEEQDR
                170       180       190       200

Conf:
Pred:
Pred: CCCEEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCEEC
  AA: VLSFRCQARIISGIHMQTSESTKSELVTVTESFSTPKFHI
                210       220       230       240

Conf:
Pred:
Pred: CCCCEEECCCEEEEEEEEEEECCCCCCCEEEEEECCEECC
  AA: SPTGMIMEGAQLHIKCTIQVTHLAQEFPEIIIQKDKAIVA
                250       260       270       280
```

Continue…

Continued…



```
Conf:
Pred:
Pred: CCCCCCCEEEEEECCCHHCEEEEEEEEEECCCCCEEEEEE
  AA: HNRHGNKAVYSVMAMVEHSGNYTCKVESSRISKVSSIVVN
                290       300       310       320

Conf:
Pred:
Pred: CCCCCCCCCCCCCCEEECCCCEEEEEECCCCCCCCEEEEE
  AA: ITELFSKPELESSFTHLDQGERLNLSCSIPGAPPANFTIQ
                330       340       350       360

Conf:
Pred:
Pred: EEEEECCCCCCEEEEEEECCCCEEEEEEEECCCCEEEEEE
  AA: KEDTIVSQTQDFTKIASKSDSGTYICTAGIDKVVKKSNTV
                370       380       390       400

Conf:
Pred:
Pred: ECCCCCCCCCCCCCCCCCEECCCCCEEEEEEEECCCCCCC
  AA: QIVVCEMLSQPRISYDAQFEVIKGQTIEVRCESISGTLPI
                410       420       430       440

Conf:
Pred:
Pred: EEEEEECCCCEEEEEECCCCEEEEEECCCCCEEEEEEEEE
  AA: SYQLLKTSKVLENSTKNSNDPAVFKDNPTEDVEYQCVADN
                450       460       470       480

Conf:
Pred:
Pred: EECCCCCCEEEEEEEEECCCCCCCCCCCCCCCCCEECCEE
  AA: CHSHAKMLSEVLRVKVIAPVDEVQISILSSKVVESGEDIV
                490       500       510       520

Conf:
Pred:
Pred: EEEEECCCCCCEEEEEECCCCCCCCEEEEECCCCEEEEEC
  AA: LQCAVNEGSGPITYKFYREKEGKPFYQMTSNATQAFWTKQ
                530       540       550       560
```

Continue…

Continued…

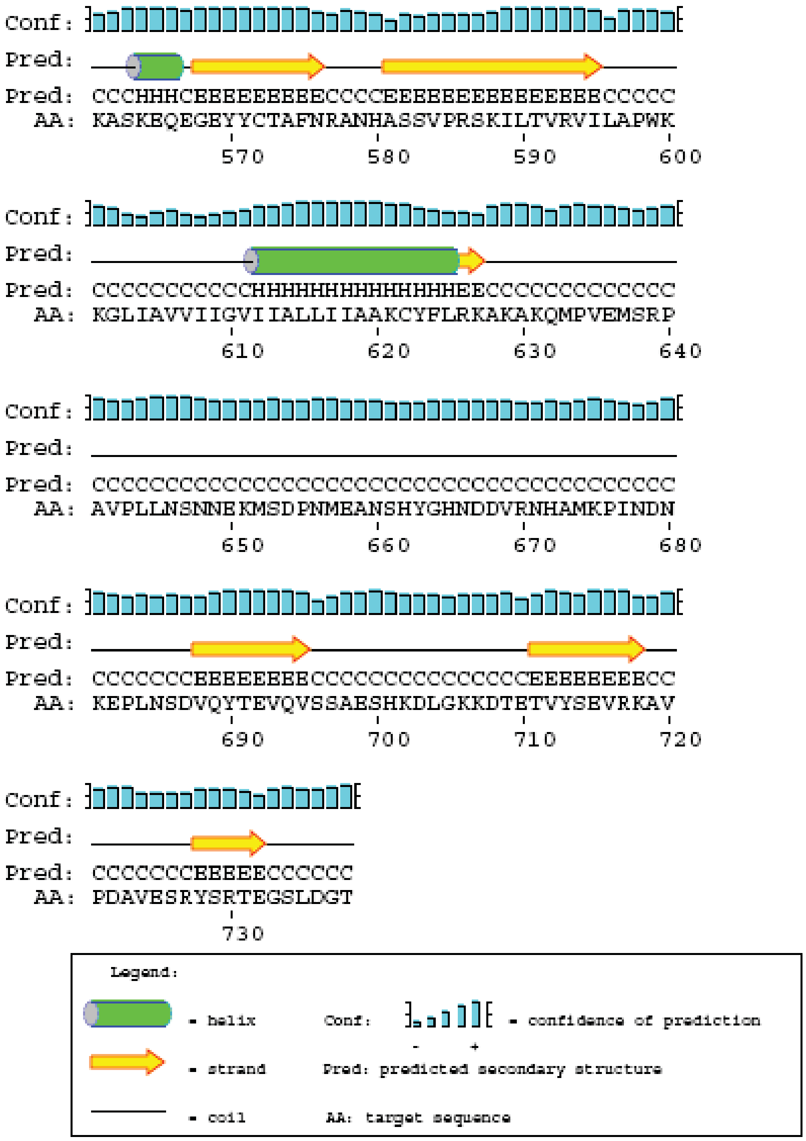

Conf:
Pred:
Pred: CCCHHHCEEEEEEEEECCCCEEEEEEEEEEEEEEECCCCC
AA: KASKEQEGEYYCTAFNRANHASSVPRSKILTVRVILAPWK
570
580
590
600
Conf:
Pred:
Pred: CCCCCCCCCCCHHHHHHHHHHHHHHEECCCCCCCCCCCCC
AA: KGLIAVVIIGVIIALLIIAAKCYFLRKAKAKQMPVEMSRP
610
620
630
640
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
AA: AVPLLNSNNEKMSDPNMEANSHYGHNDDVRNHAMKPINDN
650
660
670
680
Conf:
Pred:
Pred: CCCCCCCEEEEEEEECCCCCCCCCCCCCCCEEEEEEEECC
AA: KEPLNSDVQYTEVQVSSAESHKDLGKKDTETVYSEVRKAV
690
700
710
720
Conf:
Pred:
Pred: CCCCCCCEEEEECCCCCC
AA: PDAVESRYSRTEGSLDGT
730
Legend:
- helix
- strand
- coil
Conf:
- confidence of prediction
-
+
Pred: predicted secondary structure
AA: target sequence

Figure 3.1. *Secondary structure predictions for the aligned protein sequences of the subunits of human PECAM-1 using PSIPRED (Bryson et al. 2005). Alpha helices and coils were predicted with higher confidence as indicated by the blue bars by PSIPRED as compared to the coil prediction. The secondary structure prediction is in agreement with Swiss-Prot annotation.*

Table 3.2. Templates used for building the initial models of the Ig-domains of PECAM-1 identified by LALIGN/PLALIGN (Pearson & Lipman 1988). The data in the table report the percentage identity and the length of the amino acid (AA) aligned range that shares identity with the query sequence (extracellular domains of PECAM-1) obtained by local sequence alignments.

Template crystal structures	PDB code	Amino acid range in PECAM-1											
		Ig-domain 1 (34-128)		Ig-domain 2 (139-234)		Ig-domain 3 (232-342)		Ig-domain 4 (328-414)		Ig-domain 5 (415-510)		Ig-domain 6 (499-604)	
		% identity	AA	% identity	AA	% identity	AA	% identity	AA	% identity	AA	% identity	AA
VCAM (I) - C2 set	1VCA	19.8	81	18.7	75	41.2	17	25.9	81	19	76	21.6	74
VCAM(II) - I set	1VCA	43.8	16	22.2	72	30.4	23	21	81	25	45	33.3	18
CD8 (I) – V set	1HNF	25	40	27.8	18	21.7	83	27	37	-	-	27.3	22
CD8 (II) - C2 set	1HNF	-	-	50	6	31.6	19	27	63	26.9	52	23.1	78
CEA (I) – V set	1L6Z (mouse)	22.5	40	36.6	11	15.2	33	25.6	39	-	-	24.6	61
CEA (II) – I set	1L6Z	20.9	67	24.1	83	25	40	23.5	85	24.6	61	-	-
CD16/FcGR (I) - I set	1FCG	23.7	76	50	6	33.3	36	27.7	65	20.7	58	50	12
CD16/FcGR (II) - I set	1FCG	26.5	68	22.7	44	30.2	63	26.2	84	24.6	65	24.7	81
NCAM (I) - I set	1QZ1 (Rat)	28.1	32	33.3	12	20.4	49	20.4	49	20	40	20	30
NCAM (II) - I set	1QZ1	20.7	82	40	10	35	40	21.2	66	38.1	21	31.2	64
NCAM (III) - I set	1QZ1	38.1	21	27.3	55	22.7	22	23.8	63	22.4	67	21.2	99
ICAM (I) - C2 set	1P53	37.5	24	31.2	32	25.3	79	25	88	31.8	22	21.4	28
ICAM (II) - I set	1P53	38.5	13	35.3	34	29.2	24	19.3	83	42.9	14	29.4	17
ICAM (III) - I set	1P53	29.2	24	36.4	11	25.8	31	23.8	42	33.3	33	38.5	13

The alignment of the PECAM-1 Ig-like domains with other members of the Ig-CAM family performed with LALIGN/PLALIGN suggests that PECAM-1 evolved through a process of gene duplication (Newman *et al.* 1990). The evolutionary analysis using ClustalW suggested that NCAM, CEA and FcγR are more closely related to human PECAM-1 (Figure 3.2) Ig-domains 1 and 2 were modelled with multiple templates including NCAM and VCAM, as reported by others (Nakada *et al.* 2000; Newton *et al.* 1997).

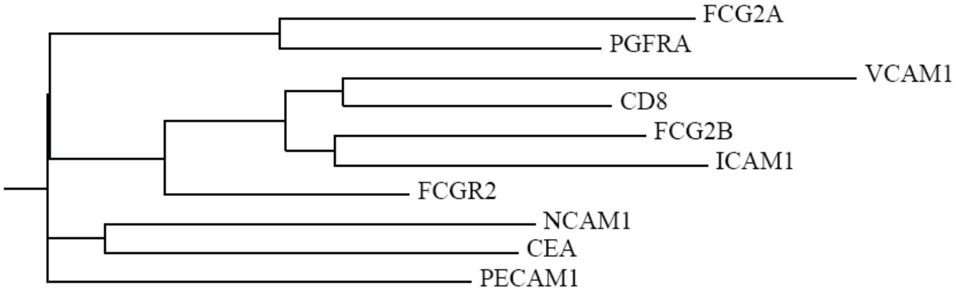


Figure 3.2. *Phylogenetic analysis of human PECAM-1 with sequences from known crystal structures of human ICAM, CD8 and isoforms of FcγR, NCAM-1 (Rat), and NMR structure of CEA.*

The quality of the final model of human PECAM-1 (having all the extracellular domains together) was evaluated by means of the PROCHECK program (Laskowski *et al.* 1993) at a 2.5 Å resolution, which gives Ramachandran plots and a quantitative distribution of the geometric parameters within the allowed conformational space. The percentage of residues in the most favoured, allowed and disallowed conformations, were 82.7, 13.7 and 1.0, respectively. As shown in Figure 3.3, the distribution of the Psi/Phi angles of the model is within the allowed regions and only five residues are in disallowed regions. These residues were not further optimised due to the extensive presence of loops and residues like glycine and proline in the structure. Eval23D (Douguet & Labesse 2001b) and Verify 3D (Douguet & Labesse 2001b) also predicted a good model, with 3D profile scores of 0.035 and 0.117, respectively. The models of various Ig-domains 4, 1, 2 and 3, and 5 and 6 of the human PECAM-1 model that were derived are illustrated in Figure 3.4, Figure 3.5, Figure 3.6 and Figure

3.7, respectively. The electrostatic potential surface representations show that Ig-domains 2, 3, 4, 5 and 6 have positive electrostatic surfaces (coloured blue in Figure 3.4, Figure 3.6 and Figure 3.7) and may constitute binding surfaces for predominantly anionic ligands such as GAGs.

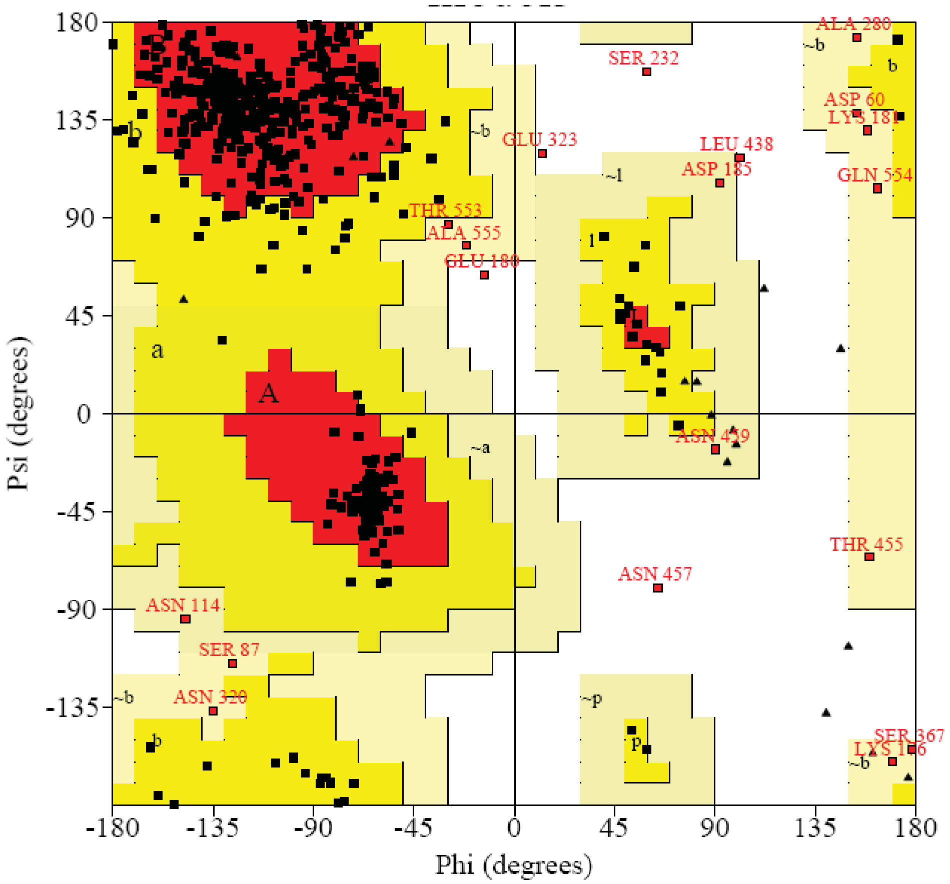


Figure 3.3. *Ramachandran plot for the final model (with all the six extracellular domains) of PECAM-1.*

The identification of sulfate binding sites is important for predicting the binding of heparin and other GAGs. Various approaches have been used to identify heparin/GAG binding sites on the surface of proteins on the basis of amino acid composition (Caldwell *et al.* 1996; Fromm *et al.* 1997), secondary structure (Hileman *et al.* 1998a), spatial distribution of the basic amino acids (Margalit *et al.* 1993) and the surface properties of proteins (Forster & Mulloy 2006). While consensus sequences such as XBBXBX and XBBBXXBX (where B is a basic residue and X can be any residue) have been suggested for heparin binding (Cardin & Weintraub 1989), they are neither necessary nor sufficient to define a GAG binding site. GAG binding sites generally consist of a cluster of basic residues on the protein surface, but not necessarily in a continuous sequence. Consequently, it was decided to perform a database search for sulfate binding structural motifs. It has been observed that arginine and lysine have the highest propensity to bind to GAGs in general. There are fewer tryptophans than any other amino acids in the PECAM-1 sequence. This is indicated by the percentage composition of amino acids predicted using the PHD module of PredictProtein (Rost *et al.* 2004) (see Table 3.3). Nonetheless, basic residues like arginine and lysine (approximately 13%) present in various domains of PECAM-1 may contribute towards GAG binding.

Table 3.3. Percentage residue composition for human PECAM-1 sequence.

Ala: 5.7	Cys: 2.2	Asp: 4.2	Glu: 7.7	Phe: 3.1
Gly: 4.1	His: 2.4	Ile: 6.9	Lys: 8.0	Leu: 6.5
Met: 2.6	Asn: 5.0	Pro: 4.3	Gln: 4.9	Arg: 3.4
Ser: 9.8	Thr: 6.5	Val: 9.1	Trp: 0.7	Tyr: 3.0

A survey of sulfate binding regions revealed the existence of several positively charged regions in Ig-domains 2 as reported earlier (DeLisser *et al.* 1993), 3, 4 and 6 as described in Table 3.4. Residues 177-182 in Ig-domain 2 have a high sequence identity with the sulfate binding site in bacterial protein disulfide oxidoreductase (PDB code 2AYT). Other sulfate binding motifs were also found: (1) residues 207-223 in Ig-domain 2, which are homologous to sulfate binding motifs in bacterial SecA translocation ATPase (PDB code 1M6N and PDB code 1M74); (2) residues 254-258 and 278-286 in Ig-domain 3, which are homologous to sulfate binding motifs in snake phospholipase A2 (MIPLA3) (PDB code 1OZY); (3) residues 330-342 in Ig-domain 4, which are homologous to sulfate binding motifs in the multiple sugar binding transport ATP-binding protein (PDB code 2D62) and ribosomal protein S6 kinase alpha 5 (PDB code 1VZO). Interestingly, the region 563-571 in Ig-domain 6 of PECAM-1 showed homology to the sulfate binding motifs in the crystal structure of HIV-2 reverse transcriptase (PDB code 1MU2). In the final model of human PECAM-1, the relative distance between each sulfate was found to be approximately 8-9 Å in Ig-domains 2 and 3, as observed in the crystal structure of Artemin (PDB code 2ASK)

(Silvian *et al.* 2006). Two sulfate binding motifs were identified in Ig-domain 4 (Figure 3.4) placed at a distance of 24 Å from each other. No sulfate binding motif was identified in Ig-domain 5. A single sulfate binding motif was predicted by the survey in Ig-domain 6. The clusters of basic amino acids in Ig-domains 2 and 3 are located approximately 20 Å apart. A comparison of the spatial distribution of basic amino acids in other heparin binding proteins that have a β-sheet topology (such as apolipoprotein E, AT-III and NCAM) suggests that this 20 Å-long region can accommodate a GAG pentasaccharide (Margalit *et al.* 1993). All the predicted sulfate binding sites (and the accompanying sulfates) were incorporated into the complete model of human PECAM-1.

Table 3.4. Prediction of the sulfate binding motifs in Ig-domains of PECAM-1 from known crystal structures.

Sulphate binding motif predicted in Ig-domains of PECAM-1	**Corresponding alignment of sulphate binding motifs from known PDB structure**	**PDB code**
176-KLKREK-181	221-KLKREK-226	2AYT: A, B
207-QAR---IISGIHMQTSESTK-223	216-EARTPLIISG---QAAKSTK-232	1M6N: A 1M74: A
254-IKCTI-258	9-IKCTI-13	1OZY: A, B
278-IVAHNRHGN-286	133-IVAHQR-GN-140	1AQZ: A, B
331-ESSFTHLD-338	229-DASFTHLD-306	2D62: A
329-ELESSFTHLDQGER-342	145-EL---FTHLDQGER-155	1VZO: A
563-SKEQEGEYY-571	306-SQEQEGHYY-314	1MU2: A,B

The colour encoding in blue indicates that residues in the structural fold of protein are known to bind sulfates directly according to the interactions defined in PDBsum.

Ig-domain 5 of PECAM-1 is connected to Ig-domain 4 through a long flexible loop. Ig-domain 5 forms only half of the Ig-fold since it lacks the characteristic β strands a, b, d and e, as reported earlier (Newman *et al.* 1990). The b strand was obtained from the structure of the KK50.4 T-cell receptor beta chain (PBD code 2ESV: E). The region 449-507, including the d and e strands in Ig-domain 5, showed 49% homology with monomeric isocitrate dehydrogenase in complex with isocitrate and Mn^{2+} (PDB code 1ITW) when a PSI-BLAST search was performed. However, after surveying 3D fragments of domain 5 in the PBD no sulfate binding motifs were detected.

Construction of the Ig-like domains 1 and 2 of PECAM-1 was carried out by modelling the side chains of specific surface-exposed amino acids in order to facilitate the interactions involved in homophilic and heterophilic binding of ligands. The side chains of the key residues Asp 38, Asp 60, Lys 77, Asp 78 and Lys 116 (Swiss-Prot numbering) that mediate PECAM-1 homophilic binding were repositioned in Ig-domain 1 (see Figure 3.5) on the basis of PECAM-1 models described earlier based on the VCAM-1 structure (Nakada *et al.* 2000; Newton *et al.* 1997). A ClustalW (J. D. Thompson *et al.* 1994a) sequence analysis of PECAM-1 shows that residues Asp 60, Lys 77, Asp 78 and Lys 116 are highly conserved in species such as human, bovine, pig, rat and mouse, whereas Asp 38 in human PECAM-1 is replaced by His 38 in other species. This correlates with the fact that homophilic interactions of PECAM-1 mediated by Ig-domain 1 are species specific, as reported earlier (Nakada *et al.* 2000).

Ig-domains 1-3 of PECAM-1 have also been proposed to participate in αvβ3-mediated heterotypic binding. This binding was inferred to be cation and temperature dependent (Buckley 1996); however, the validity of some of these presumed ligand interactions has been challenged (Sun *et al.* 1996). Ig-domains 1-3 lack the integrin binding motifs found in members of the integrin-binding Ig superfamily (IgSF) described to date, including the Ig-domains of VCAM, ICAM-1 and ICAM-2. The integrin and metal binding motifs were not found upon survey of structural motifs from the PDB in Ig-domains 1 and 2, anticipating the fact that PECAM-1 may modulate heterophilic adhesion by an indirect mechanism, as reported previously (Sun *et al.* 1996).

PECAM-1 has a large presence of charged residues on the surface of Ig-domains 1, 2 and 3. Modelling of the protonated and unprotonated states of these residues may provide useful information. Basic amino acids are clustered in the model of PECAM-1, with histidine side chains from the β-sheets positioned within 3.0 to 5.0 Å of the sulfate binding motifs. Ig-domains 2 and 3 are brought into close contact by a loop of only 3 residues long. However, the flexibility of this loop results in either an open or a closed conformation of that portion of the PECAM-1 molecule. In the homology model, the best fit loop modeled connects Ig-domains 2 and 3 in a closed conformation. In a closed conformation, it is expected that a GAG fragment may be able to bind to PECAM-1 in a way that involves domains 2 and 3. Modelling of these domains (see Figure 3.6) showed the sulfates bound to two clusters of residues: Q-A-R (207-209) and L-K-R-E-K-N (177-182), with His 162 in close vicinity in Ig-domain 2. The side chains of His 239 and His 253 are oriented towards the region I-K-C-T-I (254-258), and His 281 and His 298 are in close vicinity of the region 278-286 in Ig-domain 3 (see Figure 3.6). Histidines were modelled in a positively charged form with both N_δ and N_ε atoms protonated. Modelling of surface-exposed histidines in Ig-domains 2 and 3 in close vicinity to the sulfate binding motifs suggests that these histidines may assist in the binding of PECAM-1 to heparin or HS in a pH-dependent manner, as was observed for the chemokine CXCL 12 (Veldkamp *et al.* 2005), the cytokine Vascular Endothelial Growth Factor (VEGF) (Coombe & Kett 2005), and the mouse mast cell protease 7 (Matsumoto 1995). A fully open conformation was seen to result in the clusters of basic residues in Ig-domains 2 and 3 being too far apart to constitute a GAG binding site due to a twist in the relative orientation of these domains.

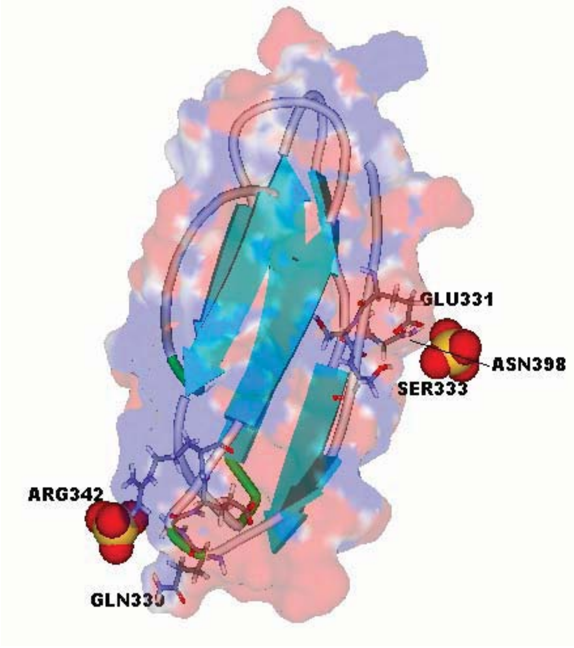


Figure 3.4. *Ribbon diagram of Ig-domain 4 is shown in a schematic representation according to secondary structure with the sulfate binding motifs indicated by highlighted residues (in stick representation). The surface of Ig-domain 4 is represented by electrostatic potential (negative potential in red and positive potential in blue). The sulfates are shown in CPK.*

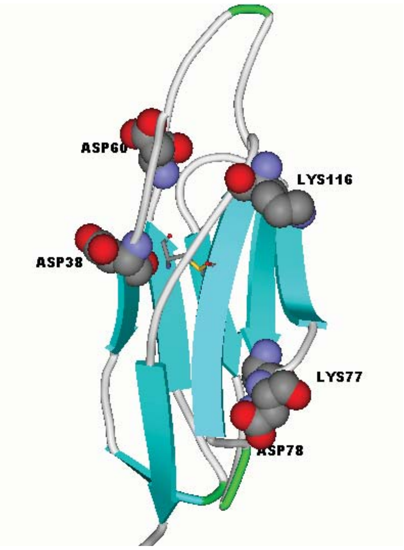


Figure 3.5. *Ribbon diagram of Ig-domain 1 is shown in a schematic representation according to secondary structure. Homophilic binding sites reported by Newton et al. (1997) in Ig-domain 1 are shown in CPK representation. The disulfide bond between Cys 57 and Cys 109 is shown in sticks.*

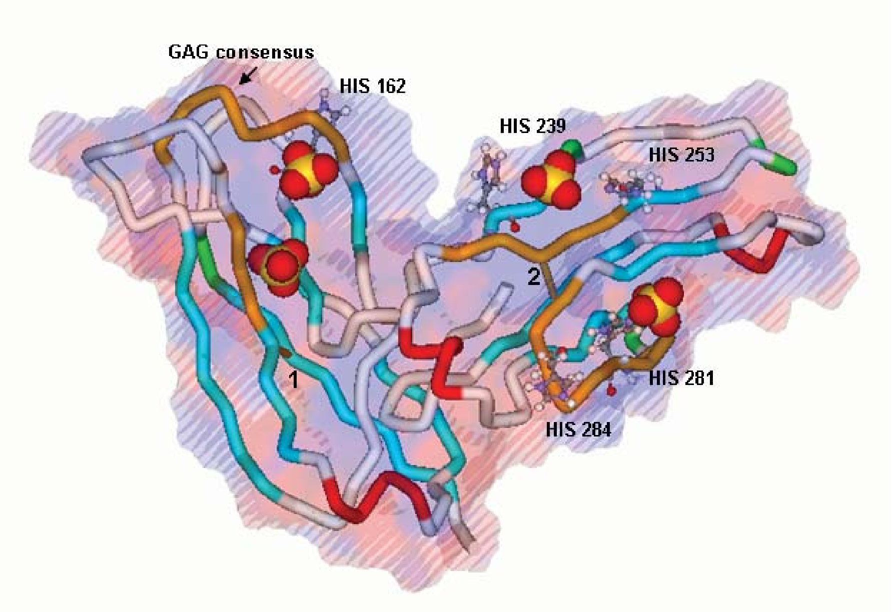


Figure 3.6. *Toothpaste representation of Ig-domains 2 and 3 according to secondary structure. Numbers 1 and 2 indicate the disulfide bridges. The protonated histidines are shown in ball and stick representation and the positive electrostatic potential (shown in blue) surface represents the regions consisting of basic residues*

found by a survey of sulfate binding motifs (in orange), which may constitute high affinity binding sites for GAGs.

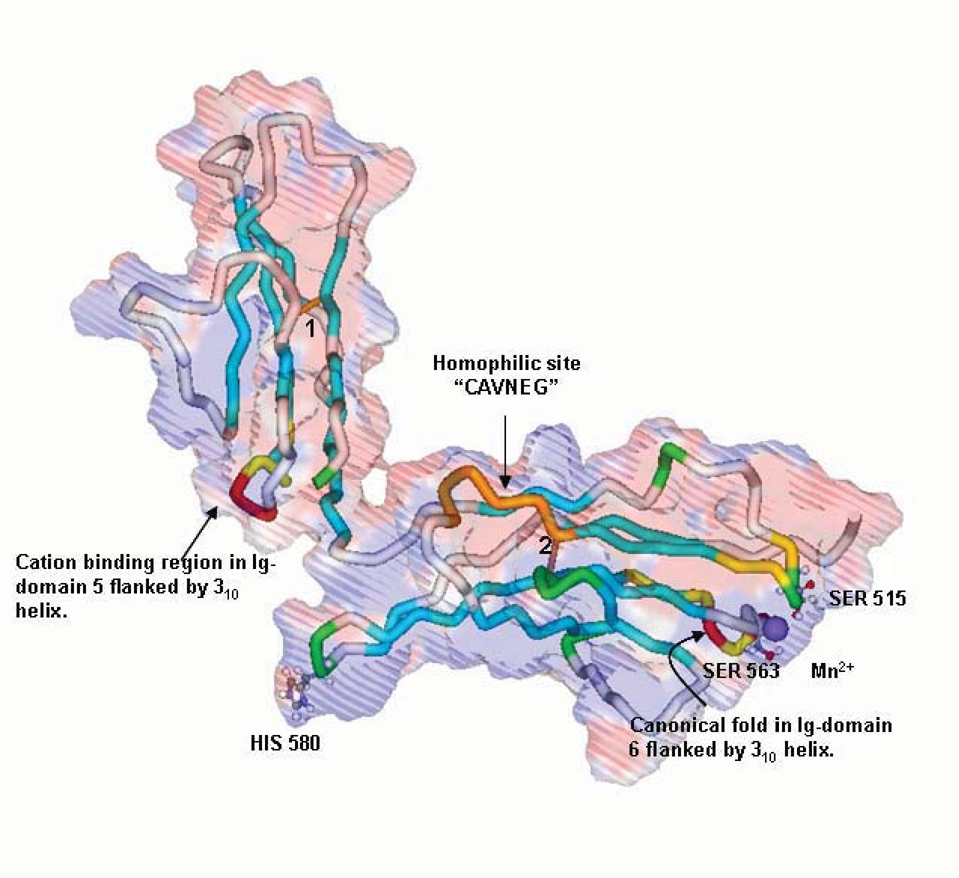

1
Homophilic site
"CAVNEG"
2
Cation binding region in Ig-domain 5 flanked by 3_{10} helix.
SER 515
SER 563
Mn^{2+}
HIS 580
Canonical fold in Ig-domain 6 flanked by 3_{10} helix.

Figure 3.7. *The cation coordination site in Ig-domains 5 and 6, as described by Jackson et al. (D. E. Jackson et al. 1997), is represented by a yellow tube. Numbers 1 and 2 indicate the disulfide bridges***.** *The cation binding regions in Ig-domains 5 and 6 are flanked by 3_{10}-helix. The homophilic binding site identified in Ig-domain 6 (Yan et al. 1995) is shown in orange. The regions of positive electrostatic potential (shown in blue) may contribute to low affinity binding site for GAGs.*

Cation coordination sites were also modelled in Ig-domains 5 and 6 of PECAM-1. A survey of the PDB for metal binding sites in this region could not detect any metal binding motif. It was noticed that the acidic residues in the region 463-475 are placed at a distance of 8-9 Å, making them too far apart to constitute a metal coordination site (see Figure 3.7). However, the presence of a cluster of basic residues in the vicinity (second coordination shell) of the cation binding site may nonetheless act as an electrostatic anchor for the metal ion. A new canonical fold was found in region 565-572 in Ig-domain 6, as described for the anti-HIV-1 V3 Fab 2219 structure (PDB codes 2B0S: L and 2B1A: L) and the CD8 alpha ectodomain fragment (PDB code 1BQH: G, H, I, K), which consists of a beta-sheet followed by a 3_{10}-helix. This fold is similar to the predicted secondary structure in the region 565-572 of Ig-domain 6 (Figure 3.7).

The PECAM-1 cation binding site in Ig-domain 6 has been demonstrated to have higher affinity for Mn^{2+} than for Ca^{2+} or Zn^{2+} (D. E. Jackson *et al.* 1997). Consequently, it was decided to model the conformation of the cation coordination site MIDAS (metal ion-dependent adhesion site) for Ig-domain 6 in the presence of Mn^{2+}, as described in the past (Legge *et al.* 2002; Leitinger & Hogg 2000; Yang *et al.* 2005). In those reports, Mac-1(CD11b/CD18/αMβ2), LFA-1(αLβ2) and CD2 had the metal coordinated to water molecules as well as serine, aspartate and glutamate residues in both open and closed conformation of the receptor as shown in figure below (Figure 3.8).

In the final model, Glu 514, Asp 518 and Glu 569 are located at a distance of 6 Å from each other and so it is possible that these residues may coordinate to the Mn^{2+} through water molecules in the MIDAS site as observed for Mac-1 and LFA-1 cell adhesion molecules (Leitinger & Hogg 2000). It is predicted that serine residues 515 and 563 present in the metal binding motif in domain 6 are available to complete the metal ion coordination, but there are no experimental data available to support this prediction. The mutagenesis data is only available for the acidic residues in the cation binding regions. The model of Ig-domain 6 appears to be different in terms of the position of the metal coordination site flanked by 3_{10}-helix to that of the hypothetical model described by Jackson and coworkers (D. E. Jackson *et al.* 1997). The modelling of the topology of the cation binding site was further validated by performing searches in the Conserved Domain Database (CDD) (Marchler-Bauer *et al.* 2005; Marchler-Bauer *et al.* 2007) for Ig-domains 5 and 6. These searches suggested structures with similar folds to those we derived in the final model. The aromatic residues Phe 545 and Tyr 546 located near the metal-binding sites as observed in case of metal binding protein CD2 (Yang *et al.* 2005). The rationale behind the existence of divalent cation binding sites proximal to the protein surface is that Mn^{2+} might interact efficiently with histidine-containing ligands (Babor *et al.* 2005; Bock *et al.* 1999). Such interactions would stabilise the adhesive and downstream signaling properties of PECAM-1, providing a structural basis for PECAM-1 mediated cellular interactions.

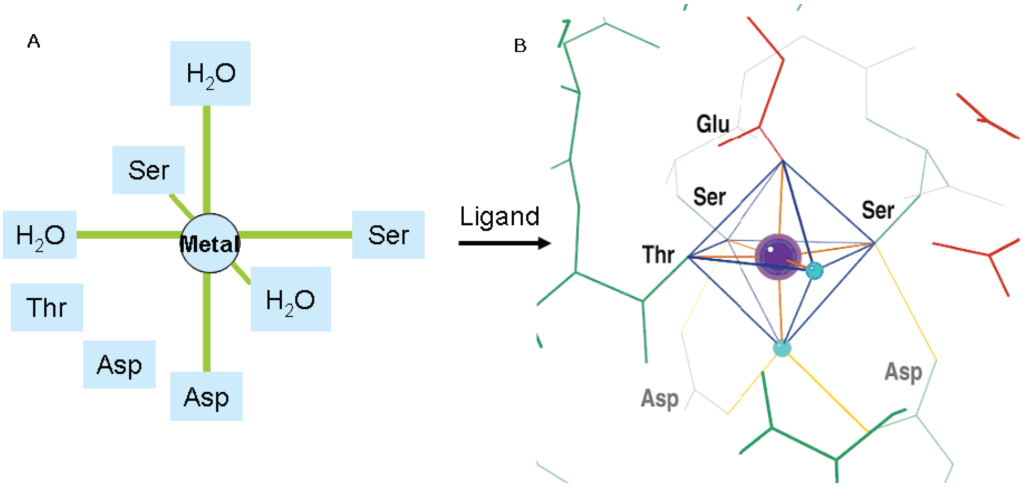


Figure 3.8. *Schematic view the metal ion coordination of the "I domain" (MIDAS motif) upon ligand binding as described for Mac-1 and LFA-1 (Bella & Berman 2000; Leitinger & Hogg 2000).* ***A****: In the absence of ligands, the metal ion is directly coordinated by three side chains (2 Ser and 1 Asp residues) and three water molecules. Another Asp and Thr are indirectly involved in metal coordination via hydrogen bonds.* ***B****: Octahedral coordination of ions such as Mg^{2+} or Mn^{2+} (purple sphere) in the open or liganded conformation of an "I domain". An aspartate residue (labelled in grey) that coordinates directly to the metal ion in the closed form does so indirectly through a water molecule in the open form (upon ligand binding such the "I domain" of integrin receptor αLβ2 interacts with its ligand ICAM-1). Yellow lines represent hydrogen bonds. Side chains in red correspond to the ligand molecule. Residues labelled in black are directly coordinated to the metal ion, including the threonine residue that was indirectly coordinated in the closed form.*

Chapter 4

MOLECULAR DOCKING OF GAGS TO THE EXTRACELLULAR DOMAINS OF PECAM-1

The previous chapter described the construction of a homology model of PECAM-1 and the prediction of likely sulfate binding sites. This chapter describes the use of this homology model in molecular docking simulations in order to predict and model the interactions of various GAG fragments with PECAM-1.

4.1 DOCKING OF GAG FRAGMENTS

Two programs were used for docking GAG fragments to domains 2 and 3 as mentioned in Figure 3.6 and domains 5 and 6 as mentioned in Figure 3.7. The sulfates were removed from the homology models of the Ig-domains 2 and 3, and 5 and 6. PatchDock (Schneidman-Duhovny *et al.* 2005; Schneidman-Duhovny *et al.* 2003) was used to dock heparin and other GAG fragments to the homology model of PECAM-1. PatchDock is a fast geometry-based molecular docking algorithm that works by optimizing shape complementarity (hence, it is not an energy grid-based method). No constraints were used to define the binding site in order to allow the program to explore the entire surface of PECAM-1 and find appropriate interaction regions using an RMSD clustering in order to reduce the number of potential binding modes (Schneidman-Duhovny *et al.* 2005).

Most three-dimensional X-ray structures of GAG-protein complexes determined so far involve relatively small oligosaccharides (di- to hexasaccharides) of varying affinity for their protein targets. Hence, in order to determine the *minimal* length of the heparin fragments required for binding to the Ig-domains of PECAM-1, docking simulations with di- and pentasaccharides were performed. The structure of the heparin pentasaccharide was obtained from the crystal structure of annexin A2 complexed with a Δhexasaccharide (Δ indicates the presence of 1,4–dideoxy–5-dehydro glucoronic acid at the non-reducing end) in PDB code 2HYV. Since no electron density was observed for the sixth saccharide (Shao *et al.* 2006), the pentasaccharide was extracted directly from the structure. The residue at the non-reducing end was modified from the unsaturated UA2S by *in-silico* addition of hydrogen to the double bond between C-4 and C-5 to create a 4-deoxy IdoA2S residue (4dIdoA2S). The sequence of the modelled pentasaccharide (Figure 4.1) consisted of IdoA2S(1→4)Glc*N*S6S(1→4)IdoA2S(1→4)Glc*N*S6S(1→4)IdoA2S. The pyranose rings of the glucosamine residues adopt a $^4C_1$ chair conformation whereas iduronic acids can adopt either a $^1C_4$ chair or a ^2S_o skew-boat conformation. The pentasaccharide was modelled with the iduronic acid at the non-reducing end in $^1H_{2,}$ the central and terminal iduronic acids in the $^1C_4$ chair conformation. The structure of the disaccharide (IdoA2S(1→4)Glc*N*S6S) was extracted from the reported NMR structures of a heparin dodecasaccharide fragment (PDB structure 1HPN) (Mulloy *et al.* 1993). The iduronic acid was chosen in the $^1C_4$ chair conformation and the glucosamine in the $^4C_1$ chair conformation to compare the docking results with those obtained for a similar conformation of the pentasaccharide.

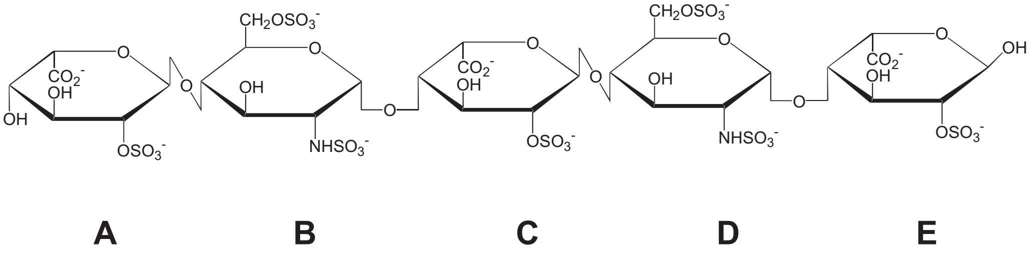


Figure 4.1. *Structure of pentasaccharide (UAP-SGN-IDU-SGN-IDU; SGN - N,o6-disulfo-glucosamine, IDU - 1,4-dideoxy-o2-sulfo-glucuronic acid, UAP - 1,4-dideoxy-5-dehydro-o2-sulfo-glucuronic acid) extracted from its crystal structure (Shao et al. 2006). The UAP residue was modified to IDU for the docking simulations. A, B, C, D, and E refer to the labels of each of the residues of the pentamer UAP-SGN-IDU-SGN-IDU, respectively.*

Docking of the DS tetrasaccharide (PDB code 1HM2) and a pentasaccharide extracted from CS (PDB code 1C4S) was also performed. There is no crystal structure available for the DS pentasaccharide. The modelled DS tetrasaccharide consisted of

IdoA(1→3)Gal*N*Ac4S(1→4)IdoA(1→3)Gal*N*Ac4S and the modeled CS consisted of GlcA(1→3)Gal*N*Ac4S(1→4)GlcA(1→3)Gal*N*Ac4S(1→4)GlcA. Hydrogen atoms were added to these oligosaccharides and the resultant structures were energy minimised to optimise the orientation of rotatable groups. The surface area, atomic contact energy and the binding score computed by PatchDock for the heparin pentasaccharide were extracted.

Further docking simulations were carried out using the program AutoDock 3.0 (Morris *et al.* 1998). This program allows for flexibility in the ligand structure but uses a rigid body approximation for the protein receptor in order to speed up the calculation. This assumes that no conformational changes affect the structure of the receptor, which is a necessary approximation given the many degrees of conformational freedom in the GAG molecules. Chapter 5 discusses the modelling of the conformation of the receptor through molecular dynamics simulations. AutoDock Tools (ADT) (Sanner 1999) were used to prepare the PECAM-1 molecule by adding appropriate hydrogens, partial atomic charges and solvation parameters. The atom type of sulfur and oxygen atoms in sulfate groups of all oligosaccharide ligands were modified to S.o2 and O.co2, respectively, and bond type between these atoms was modified to aromatic bond in SYBYL (Tripos, Inc.). Ligand rotatable bonds for all docked ligands were defined using the AutoTors module of AutoDock. The ligands were atom-typed manually to ensure that they complied with the carbohydrate force field in AMBER (Weiner *et al.* 1984). The ligands were energy minimised in order to optimise the orientation of its hydrogen atoms. A grid spacing of 0.37 Å and a distance-dependent dielectric constant of 4.0 (Mehler & Solmajer 1991) were used for the binding energy calculations, covering the putative binding site surface. Using AutoDock's Lamarckian genetic algorithm, heparin fragments were subjected to 200 search runs using a population of 200 individuals. A grid box was defined with a constant grid spacing of 0.37 Å around each heparin fragment using the binding poses obtained from PatchDock with respect to Ig-domains 2 and 3, Ig-domain 4, and Ig-domains 5 and 6 of PECAM-1.

Due to the flexibility and size of the di- and pentasaccharides of heparin, the number of energy evaluations and the size of the genetic population were optimised in order to ensure convergence of the calculated energies, starting with a minimum of $5x10^6$ and a maximum of 50x10^6 energy evaluations, as reported earlier for blind docking (Hetenyi & van der Spoel 2002). Cluster analysis was performed on the resulting binding poses using an RMSD tolerance of 1.0 Å. Since AutoDock cannot handle more than 32 rotatable bonds, the docking of the heparin fragments was performed keeping the hydroxyl groups fixed. The lowest docking energy binding scores of the disaccharides with full rotational freedom of their hydroxyl groups were similar to those obtained when the hydroxyl groups were fixed, confirming that the initial orientation of the hydroxyl groups was appropriate for interactions with PECAM-1 (these interactions were more important for non-heparin fragments, as discussed below).

Docking of heparin fragments of various sizes (refer to Table 4.1 further below) was also performed with AutoDock version 3.0, using the same docking protocol to identify the optimal length of heparin fragment required for binding to Ig-domains 2 and 3 of PECAM-1. The heparin fragments that were modelled varied in size from two to six saccharide subunits. The di- and trisaccharides were defined to have fixed glycosidic torsion angles as taken from a reported NMR structure of heparin (PDB code 1HPN). The tetrasaccharide (obtained from PDB code 2HYU) consisted of residue A UA2S ($^1H_2$ form), residues B and D GlcNS6S (in the predominant $^4C_1$ conformation) and residue C IdoA2S (in the $^1C_4$ chair form). Different conformations of the pentasaccharide and hexasaccharide were considered. The first pentasaccharide (obtained from PDB code 2HYV) consisted of 3 iduronic residues separated by glucosamine residues, as mentioned above. Another pentasaccharide was also considered (obtained from PDB code 1QQP (Fry *et al.* 1999)), wherein the GlcNS6S sugar rings and one of the terminal IdoAp2S rings adopt the energetically favourable $^4C_1$ or $^1C_4$ chair conformations, the other terminal iduronic ring appears to adopt a $^{2,5}B$ conformation and the central iduronic ring adopts a skew boat ^2S_o conformation. The different conformations of the hexasaccharide were extracted from two different structures. In the bFGF-heparin complex (PDB code 1BFC) the iduronic rings exists in two conformations: iduronic acid in the $^1C_4$ chair conformation and in the ^2S_o skew boat conformation (Faham *et al.* 1996). In the cobra cardiotoxin A3-heparan sulfate complex (PDB code 1XT3) (S. C. Lee *et al.* 2005) the ring conformations of glucosamine residues are all in $^4C_1$, the first and second iduronic acids are in the $^2S_0$ and $^1C_4$ conformation, respectively, and the terminal uronate adopts a $^1H_2$

conformation. A third hexasaccharide (referred to hereafter as Construct 1) was also built from a pentasaccharide (PDB code 2HYV) by adding an extra glucosamine residue "A" in $^4C_1$ conformation to the non-reducing end of the sequence. The $^1H_2$ conformation of the iduronic acid "B" was substituted with the $^1C_4$ conformation similar to the iduronic acid "D". The terminal saccharide in 1BFC structure is glucuronic acid where as in all the other structures from pdb do not have any glucuronic acids.

4.2 RESULTS

A heparin pentasaccharide (Figure 4.1) was docked to all Ig domains of PECAM-1 on the basis of shape complementarity using PatchDock in order to attempt to obtain initial binding modes of the saccharides with each domain. The best binding mode of the pentasaccharide to Ig-domains 2 and 3 was determined to have an approximate interaction surface area of 1200 Å^2, an atomic contact energy (ACE) (C. Zhang *et al.* 1997) of 290 and a geometric shape complementarity score of 9830. The second best binding mode of the pentasaccharide was obtained with Ig-domains 5 and 6, having an approximate interaction surface area of 865 Å^2, an ACE of 152 and a geometric shape complementarity score of 7598. While there is no evidence of the accuracy of these measures for carbohydrate-protein interactions, these scores suggest the presence of high and low affinity GAG binding sites in Ig-domains 2 and 3 and Ig-domains 5 and 6, respectively.

In order to obtain more accurate free energies of binding, the AutoDock program was used to dock heparin fragments to PECAM-1. The top ranking binding mode of heparin obtained resulted in an improved fit between the negatively charged pentasaccharide and the positively charged regions in both Ig-domains 2 and 3 (see Figure 4.2). Docking of disaccharides to Ig-domains 5 and 6 of PECAM-1 using AutoDock suggested a better fit and lower free energies of binding on the electropositive surface of these domains (see below) compared with the pentasaccharide as predicted by PatchDock.

The results of the docking studies indicating a heparin pentasaccharide binds a region on Ig-domains 2 and 3 are consistent with the predicted location of the sulfate binding motifs. The key interactions of the heparin pentasaccharide and Ig-domains 2 and 3 identified by the docking simulations involve Lys 176, Leu 177, Arg 179, His 239, Lys 255, Gln 259 and Ile 258 (Figure 4.3). The protonated $N_{\varepsilon2}$ in His 239 makes an electrostatic interaction with the 2-*O*-sulfate of the central iduronic acid of the pentasaccharide (as defined in Figure 4.1). Residues Ile 258 (main chain) and Lys 255 (side chain) make hydrogen bonding and electrostatic contacts with the 2-*O*-sulfate in residue A and the *O*-sulfate in residue B, respectively. The side chain of Gln 259 makes a hydrogen bond with the *N*-sulfate in residue B in the top ranked binding mode obtained, but makes a hydrogen bond with the hydroxyl group of residue B in the second ranked binding mode. The *O*-sulfates in residue E and D establish ionic interactions with the charged side chains of Lys 176 and Arg 179, respectively. Furthermore, Arg 209 in Ig-domain 2 is in close proximity to of the GAG consensus region ("L-K-R-E-K-N") and hence it is possible that its guanidine group may interact with the charged residues of the pentasaccharide. However, this was not observed in the docking simulations, as it would require a change in the conformation of the main chain of residues 207-209 to bring Arg 209 closer to this cluster of basic residues.

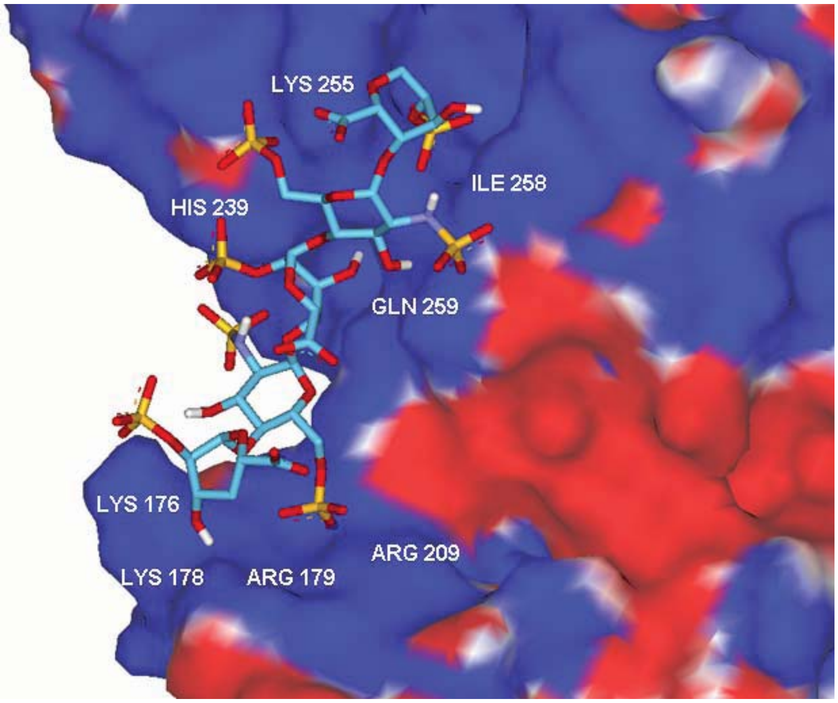


Figure 4.2. *Predicted binding modes for sulfated pentasaccharide with Ig-domains 2 and 3 of human PECAM-1, which is represented with an electrostatic potential surface (negative potential in red and positive potential in blue). The electrostatic*

potential surfaces were calculated and displayed using the DELPHI module in Discovery Studio (Accelrys, Inc.). The pentasaccharide fragment is shown in sticks.

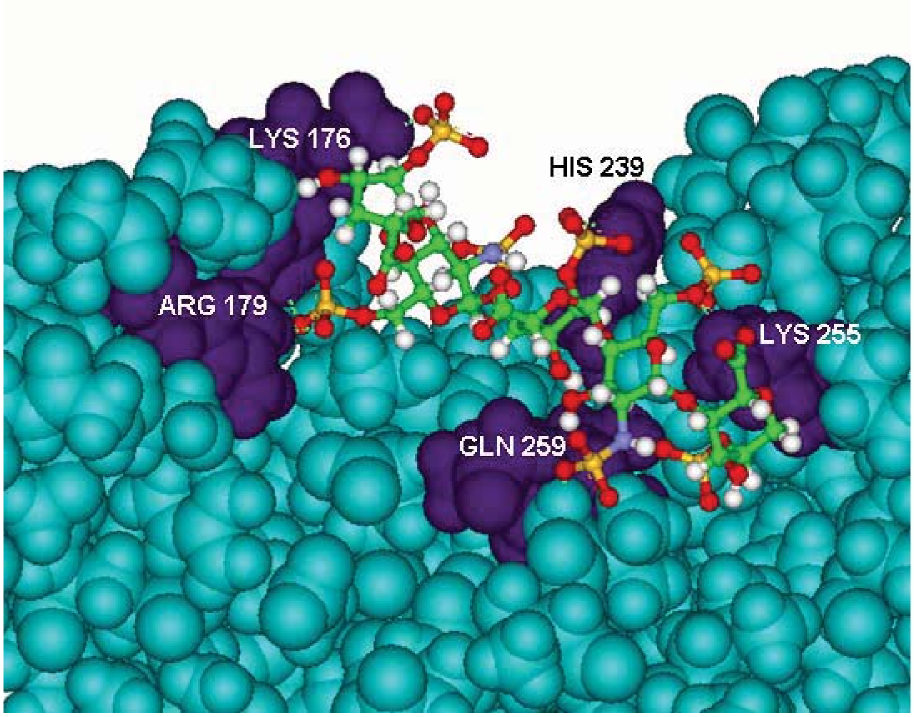


Figure 4.3. *Predicted binding mode for a sulfated pentasaccharide in Ig-domains 2 and 3 of PECAM-1, showing those amino acids (in purple) that interact with the fragment. The pentasaccharide fragment is shown in sticks.*

The docking calculations predicted the free energy of binding and the dissociation constant (K_d) of the ligands with the extracellular domains of PECAM-1 at slightly acidic pH. The predicted free energy of binding of the best binding mode of the heparin pentasaccharide with Ig-domains 2 and 3 was computed to be -17.22 kcal/mol, which results in a predicted dissociation constant of 4.93 nM. The second ranked binding mode was predicted to have a free energy of binding of -16.91 kcal/mol, resulting in a predicted dissociation constant of 9.04 nM. These calculations were repeated at a neutral pH (leaving the histidine residues in an unprotonated, neutral state). This resulted in the free energy of binding increasing to approximately -3 kcal/mol, due to the loss of ionic interactions with His 239 (Figure 4.4).

These calculations were made assuming a closed configuration of Ig-domains 2 and 3 and the $^{1}C_{4}$ chair conformation of the IdoA2S. It is likely that in nature Ig-domains 2 and 3 will not always be in such close proximity, hence allowing longer fragments to interact with both domains. Moreover, although IdoA residues exist in two conformations of nearly equal energy ($^{1}C_{4}$ chair and $^{2}S_{o}$ skew-boat), internal IdoA residues will favor the $^{2}S_{o}$ skew-boat conformation because in the $^{1}C_{4}$ form the bulky carboxylate group is equatorial and all other substituents are in axial positions (Capila & Linhardt 2002). Docking of the pentasaccharide with these alternative IdoA conformations would likely result in lower affinity of binding, and hence experimental determinations would reflect an average lower binding affinity, as we discuss further below.

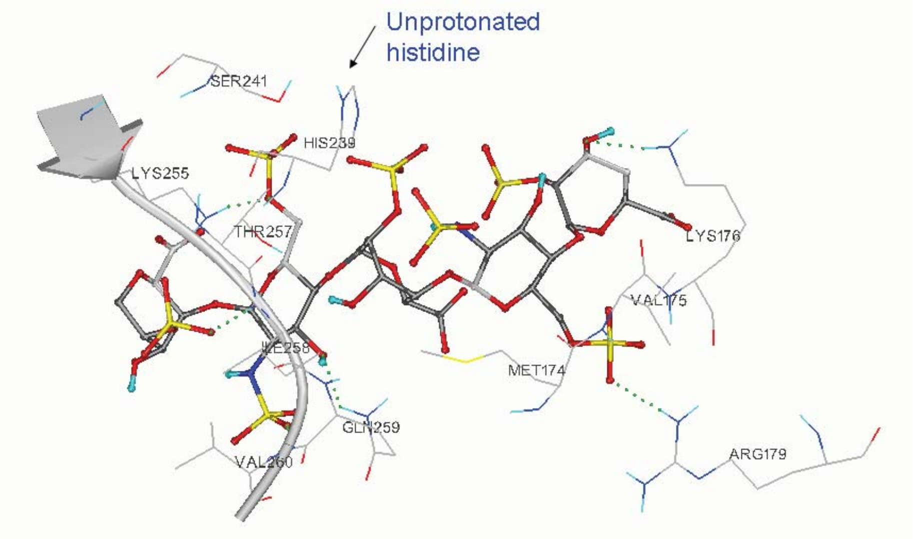


Figure 4.4. *Predicted binding mode for a sulfated pentasaccharide in Ig-domains 2 and 3 of PECAM-1, showing histidines in an unprotonated state. The free energy of binding of the heparin pentasaccharide is comparatively lower to the docked solution obtained with histidine in protonated form.*

The docking simulations also predicted the binding of heparin fragments to Ig-domains 5 and 6. Docking of heparin fragments such as di, tri, tetra and penta saccharides were carried out for prediction of binding affinities for Ig-domains 5 and 6. In this case, docking of sulfated disaccharides resulted in two clusters with significant numbers of related binding modes (see Figure 4.5). The disaccharides with the lowest energies of binding in each cluster were seen to interact with the positively charged accessible surfaces of Ig-domains 5 and 6. However, the interactions are predicted to be weak: the computed free energies of binding in clusters 1 (Ig-domain 5) and 2 (Ig-domain 6) were -6.54 kcal/mol and -6.23 kcal/mol, respectively, resulting in dissociation constants of 15.4 μM and 26.9 μM, respectively. The third lowest energy cluster was predicted to have a free energy of binding and dissociation constant of -6.13 kcal/mol and 32.2 μM, respectively. A number of amino acids (Lys 423, Lys 446, Lys 449, Asn 467, Arg 577 and His 580) in Ig-domains 5 and 6 interact with the disaccharides, mostly through ionic interactions with negatively charged sulfates. In addition, the main chain of Phe 464 interacts with the ionised carboxylate, and Thr 533 (main chain) interacts with the 2-*O*-sulfate of the IdoA residue of the disaccharide. The side chain of Glu 470 makes a hydrogen bond with the amide group of glucosamine, whereas the main chains of Gly 528 and Ser 529 make hydrogen bonds with the hydroxyl groups of the disaccharides. No significant free energies were detected upon docking of tri, tetra and pentasaccharides for these Ig-domains.

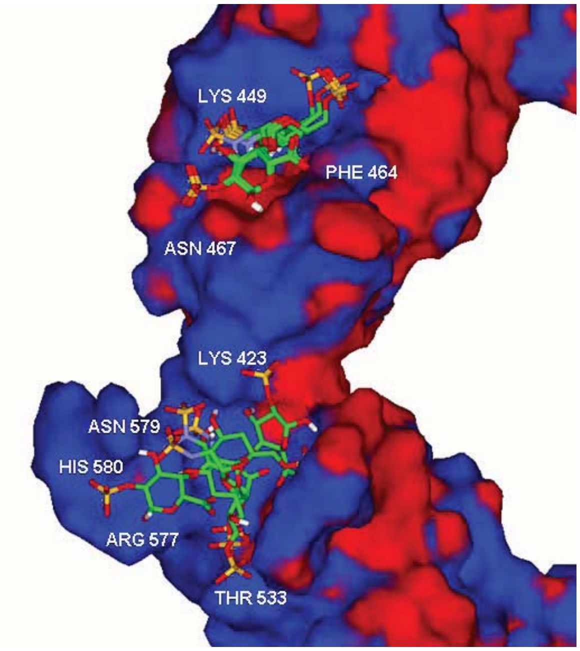


Figure 4.5. *Predicted binding modes for sulfated disaccharides with Ig-domains 5 and 6. The protein surface is colored according to the sign of the electrostatic potential (blue for positive and red for negative areas). Low binding energy clusters are depicted for the disaccharides, showing that amino acids form basic (positively charged) clusters on the surface of the protein. The electrostatic potential surfaces*

were calculated and displayed using the DELPHI module in Discovery Studio (Accelrys, Inc.). The disaccharide fragments are shown in sticks.

The binding of the disaccharides to Ig-domains 5 and 6 did not involve the cation binding region. The predicted sulfate binding motif (region 563-571) in Ig-domain 6 partially overlaps with the cation binding region (formed by residues 512-522 and 560-572). The surface of this region in Ig-domain 6 is electropositive in nature. Consequently docking simulations of disaccharides to the cation binding region in Ig-domain 6 were carried out. However, these simulations resulted in predicted positive free energies of binding, so no binding would be expected. Despite the fact that the sulfate binding region 563-571 in Ig-domain 6 showed a high level of identity and a similar structural topology with the template structure 1MU2 (HIV-2 reverse transcriptase), the presence of Glu 569 in Ig-domain 6 of PECAM-1, instead of histidine, eliminates the possibility of forming favorable ionic interactions with a sulfate group, as it occurs in the template 1MU2 structure.

In the case of Ig-domain 4, docking of the sulfated disaccharide fragments resulted in extremely low binding affinity (dissociation constants in the molar range). A cluster of binding poses for the disaccharides were found to interact with Ser 333, Arg 342 and Asn 344, as predicted by the survey of sulfate binding motifs, but no significant affinity was measured. It is likely that the lack of conservation of the protein fold required to coordinate the sulfate on the surface of Ig-domain 4, as compared with the templates 2D62 (multiple sugar binding transport ATP-binding protein) and 1VZO (ribosomal protein S6 kinase alpha 5), result in a lack of binding affinity of GAGs for Ig-domain 4 compared to Ig-domains 2, 3, 5 and 6. The sulfates in the template structures are also coordinated by additional residues from neighboring structural folds (residues Ser 346 and Lys 347 in the case of template 2D62, and residues Thr 151 and Glu 151 in the 1VZO structure). The fold in Ig-domain 4 of PECAM-1 lacks this coordination from the neighboring β-sheet despite the high sequence similarity with the template structures in the sulfate binding regions.

The distribution of amino acids in the predicted sulfate binding regions of the Ig-domains of PECAM-1 in other species was also examined. Multiple sequence alignment of the human sequence with *Bos Taurus* (bovine), *Mus musculus* (mouse), *Sus scrofa* (pig) and *Rattus norvegicus* (rat) species reflect the fact that the Ig-domains in other species may bind GAGs with different affinity due to differences in the sequence conservation of sulfate binding motifs in regions 177-182 and 207-209 in Ig-domain 2, and 239, 254-258 and 278-286 in Ig-domain 3 (Figure 4.6). The charged residues Arg 179, involved in the interaction of Ig-domain 2 of human PECAM-1 with the GAG pentasaccharide, are replaced by Ile 179 in mouse. Arg 209 present in the sulfate binding motif 207-209 is mutated to Gly 209 in the mouse and Asn 209 in the pig. His 239 in human PECAM-1, predicted to interact strongly with the GAG pentasaccharide at slightly acidic pH, is replaced by the more acidic residues Glu 239 in mouse and Gln 239 in rat. His 281, found in the predicted sulfate binding motif of Ig-domain 3 of human PECAM-1, is replaced by Thr 281 in mouse and rat.

A DS tetrasaccharide was also docked to Ig-domains 2 and 3 in order to determine the effect of the number of electrostatic interactions on the binding affinity of GAGs. The predicted binding mode showed that the hydroxyl groups present in the sugars of DS tetrasaccharide contribute to hydrogen bonding with Arg 179 and Gln 259, whereas these residues in Ig-domains 2 and 3 of PECAM-1 make ionic interactions with the sulfate group of heparin/HS. The sulfate group of DS makes an ionic interaction with Lys 255, resulting in dissociation constant and free energy of binding of 18 μM and -6.46 kcal/mol, respectively. The binding modes of dermatan sulfate and heparin tetrasaccharides are shown in Figure 4.7. It is important to note that residue A at the non-reducing end of the heparin tetrasaccharide is in the $^{1}H_{2}$ form, residues B and D are in the $^{4}C_{1}$ conformation, and residue C of IdoA2S adopts the $^{1}C_{4}$ chair form, whereas the uronic acids of dermatan sulfate are in the $^{1}C_{4}$ conformation, resulting in different protein-oligosaccharide interactions. This suggests that the biological activities of heparin and DS are modulated not only by electrostatic interactions but also by the flexibility of the iduronic acid subunits and the presence of optimal VDW interactions between the oligosaccharides and the protein.

Docking of a CS pentasaccharide to Ig-domains 2 and 3 resulted in a predicted positive free energy of binding due to the presence of a lower degree of sulfation and lower flexibility of saccharide subunits compared to heparin/HS and DS tetrasaccharides, giving rise to fewer interactions with the electropositive regions found on the surface of Ig-domains 2 and 3 of PECAM-1 (see Figure 4.8). The charged sulfate and carboxylate groups make ionic interactions with Lys 255 and Lys 176, respectively, and hydrogen bonding is observed between the hydroxyl of CS and the backbone of Ile 258.

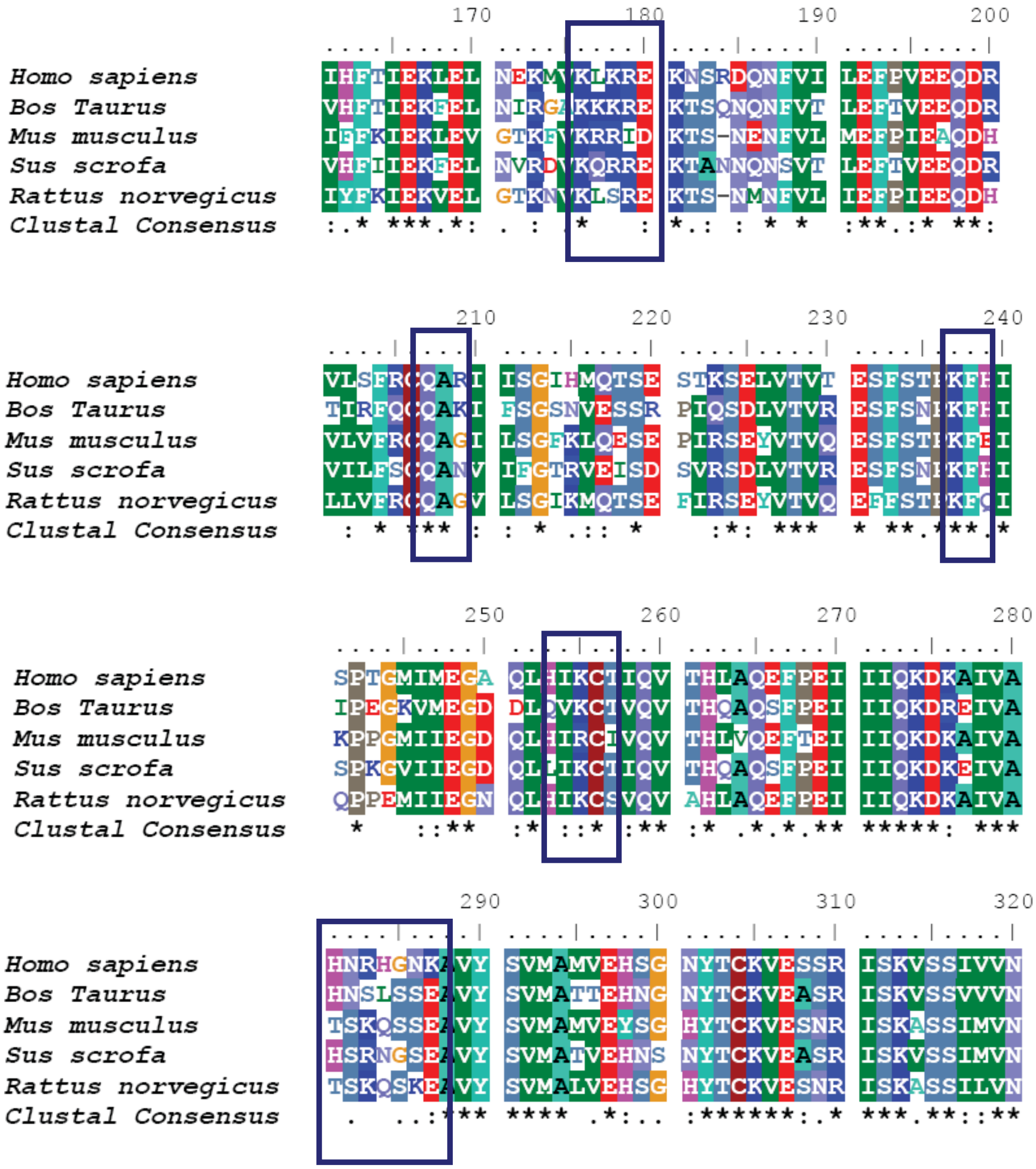

170
180
190
200
Homo sapiens IHFTIEKLEL NEKMYKLKRE KNSRDQNFVI LEFPVEEQDR
Bos Taurus VHFTIEKFEL NIRGAKKKRE KTSQNQNFVT LEFTVEEQDR
Mus musculus IPFKIEKLEV GTKFVKRRID KTS-NENFVL MEFPIEAQDH
Sus scrofa VHFIIEKFEL NVRDVKQRRE KTANNQNSVT LEFTVEEQDR
Rattus norvegicus IYFKIEKVEL GTKNVKLSRE KTS-NMNFVL IEFPIEEQDH
Clustal Consensus :.* ***.*: . : .* : *.: : * * :**.:* **:
210
220
230
240
Homo sapiens VLSFRCQARI ISGIHMQTSE STKSELVTVT ESFSTPKFHI
Bos Taurus TIRFQCQAKI FSGSNVESSR PIQSDLVTVR ESFSNPKFHI
Mus musculus VLVFRCQAGI LSGFKLQESE PIRSEYVTVQ ESFSTPKFEI
Sus scrofa VILFSCQANV IFGTRVEISD SVRSDLVTVR ESFSNPKFHI
Rattus norvegicus LLVFRCQAGV LSGIKMQTSE FIRSEYVTVQ EFFSTPKFQI
Clustal Consensus : * *** : : * .:: * :*: *** * **.***.*
250
260
270
280
Homo sapiens SPTGMIMEGA QLHIKCTIQV THLAQEFPEI IIQKDKAIVA
Bos Taurus IPEGKVMEGD DLQVKCTVQV THQAQSFPEI IIQKDREIVA
Mus musculus KPPGMIIEGD QLHIRCIVQV THLVQEFTEI IIQKDKAIVA
Sus scrofa SPKGVIIEGD QLLIKCTIQV THQAQSFPEI IIQKDKEIVA
Rattus norvegicus QPPEMIIEGN QLHIKCSVQV AHLAQEFPEI IIQKDKAIVA
Clustal Consensus * ::** :* ::* :** :* .*.*.** *****: ***
290
300
310
320
Homo sapiens HNRHGNKAVY SVMAMVEHSG NYTCKVESSR ISKVSSIVVN
Bos Taurus HNSLSSEAVY SVMATTEHNG NYTCKVEASR ISKVSSVVVN
Mus musculus TSKQSSEAVY SVMAMVEYSG HYTCKVESNR ISKASSIMVN
Sus scrofa HSRNGSEAVY SVMATVEHNS NYTCKVEASR ISKVSSIMVN
Rattus norvegicus TSKQSKEAVY SVMALVEHSG HYTCKVESNR ISKASSILVN
Clustal Consensus . ..:*** **** .*:.. :******:.* ***.**::**

Figure 4.6. *Multiple sequence alignment of PECAM-1 sequences from various species including Homo sapiens, Bos Taurus, Mus musculus, Sus scrofa and Rattus norvegicus, showing sequence conservation in the sulfate binding motifs in Ig-domains 2 and 3. Residues are colored according to their physicochemical properties. Amino acids in the sulfate binding motifs are indicated with blue boxes.*

These results support the experimental evidence that PECAM-1 cannot bind HA and CS, but may bind DS with low affinity (Watt *et al.* 1993). The simulations suggest that this occurs because heparin can easily establish electrostatic interactions with the Ig-domains of PECAM-1 due to the conformational flexibility of its iduronic acid-containing sugars, compared to rigid glucuronic acid-rich GAGs. These results also indicate that Ig-domains 2 and 3 of PECAM-1 are critical for heparin recognition and binding. These observations explain the main effects of the sulfation pattern of iduronic acid-containing GAGs, including heparin and HS, on the binding to the extracellular domains of PECAM-1 at slightly acidic pH. The sulfation patterns in GAG oligosaccharides are the result of the chemical identity of the saccharide subunits, the linkage between glucosamine and uronic acid, the distribution of sulfate groups along the chain and the conformational flexibility of the saccharide subunits. This model of the interactions of various GAG molecules with PECAM-1 is in agreement with a study demonstrating that GAG consensus “LKREKN” (177-182) in the Ig-domain 2 of PECAM-1 is involved in GAG binding (DeLisser *et al.* 1993).

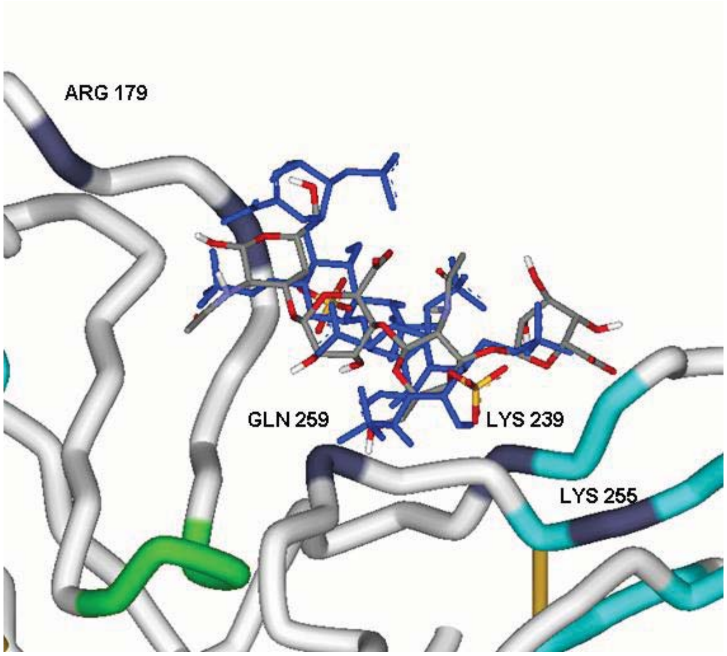


Figure 4.7. *Predicted binding mode for a dermatan sulfate tetrasaccharide (shown in sticks and colored by atom type), superimposed on a heparin tetrasaccharide (shown in blue) in Ig-domains 2 and 3 of PECAM-1. The amino acids that interact with the fragment are shown in purple.*

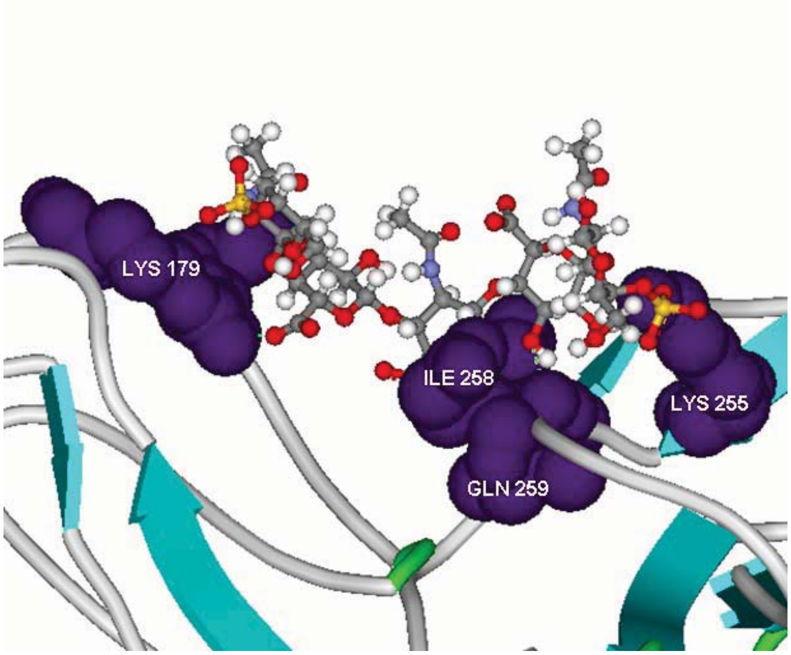


Figure 4.8. *Predicted binding modes for chondroitin sulfate pentasaccharides with Ig-domains 2 and 3 of PECAM-1. The amino acids that interact with the fragment are shown in purple.*

Heparin is predicted to have direct electrostatic interactions with positively charged residues located in loops in Ig-domains 2 and 3, providing the basis for the existence of a high affinity GAG binding region in PECAM-1. The docking simulations indicate that this GAG binding region involves major interactions from Ig-domain 3 (residues His 239, Lys 255 and Gln 259), with further contributions from Ig-domain 2 (residue Arg 179). An additional low affinity heparin binding region appears to be located in Ig-domain 5, with contributions from Ig-domain 6. Importantly, these two putative GAG binding regions are distinct from those involved in homophilic and heterophilic interactions in Ig-domains 1 and 6, as well as the cation binding sites in Ig-domains 5 and 6 of PECAM-1.

Earlier studies have reported that binding of GAG fragments to chemokines has strong GAG-length dependence (Kuschert *et al.* 1999). It would thus be desirable to provide a quantitative assessment at the molecular level of the effect of varying the size and conformation of a GAG fragment on its binding affinity to PECAM-1. An attempt was made to determine the size range of heparin fragments that can bind to Ig-domains 2 and 3 of PECAM-1 by docking heparin fragments of various lengths (from the disaccharide to the hexasaccharide) and comparing their binding affinities as computed by AutoDock. In AutoDock, the free energy of binding is the sum of the ligand-receptor intermolecular energy and the internal energy of the ligand (Morris *et al.* 1998), which can be separated to distinguish the variations in energy due to changes in the nature and number of intermolecular interactions and those that arise from changes in the geometry of the interacting molecules.

Docking of the smaller fragments showed that the binding affinity of heparin to Ig-domains 2 and 3 increased with increasing length of the heparin fragment (see Table 4.1). The disaccharide showed the highest free energy of binding (approximately -5 kcal/mol, with a dissociation constant of 42 μM), which decreased steadily as the heparin fragment increased in size until it reached a minimum for the pentasaccharide in the conformation found in PDB structure 1HYV (with a free energy of binding of approximately -11.3 kcal/mol and a dissociation constant of 4.93 nM).

Docking of one of the heparin hexasaccharide conformations (from PDB structure 1BFC) resulted in a substantial increase in the predicted free energy of binding to +15.0 kcal/mol, as can be seen in Table 4.1. This is accompanied by an increase in the intermolecular energy (4.49 kcal/mol) and a large increase in the internal energy (+21.68 kcal/mol) of the hexasaccharide. It was observed that none of the docked binding poses showed a good fit between the negative charged sulfates in this hexasaccharide conformation and the positively charged basic amino acid clusters of Ig-domains 2 and 3. This was due to the distinct conformations of iduronic acid and glucosamine in residues D, E and F. The hydroxyl of the GlcNS6S, residue B, made a hydrogen bond with the side chain of Lys 255, and the 6-O-sulfate makes electrostatic interactions with His 239, but most of the 2-*O*-sulfate groups on the IdoA2S residues protrude away from the protein surface.

This apparent loss of interactions between the hexasaccharide and Ig-domains 2 and 3 (and the resulting increase in intermolecular energy) as a result of unfavourable changes in molecular conformation required additional validation. For this purpose alternative conformations of the hexasaccharide were docked, as found in another reported crystal structure (PDB structure 1XT3) and in Construct 1. Additionally,

Table 4.1. Predicted energies of interaction of heparin fragments with Ig-domains 2 and 3.

Oligosaccharide[1]	**Template structure (PDB code)**	**Intermolecular Energy (kcal/mol)**	**Internal energy (kcal/mol)**	**Free energy (kcal/mol)**
Disaccharide[2] **IdoA2S – GlcNS6S** $^{1}C_{4}$ $^{4}C_{1}$	1HPN	-9.73	-1.87	-5.0
Trisaccharide[3] **IdoA2S – GlcNS6S – IdoA2S** $^{1}C_{4}$ $^{4}C_{1}$ $^{1}C_{4}$	1HPN	-10.84	-3.82	-5.50
Tetrasaccharide[4] **IdoA2S – GlcNS6S – IdoA2S – GlcNS6S** $^{1}H_{2}$ $^{4}C_{1}$ $^{1}C_{4}$ $^{4}C_{1}$	2HYU	-13.69	+0.94	-6.78
*Pentasaccharide[5] **IdoA2S – GlcNS6S – IdoA2S – GlcNS6S – IdoA2S** $^{1}H_{2}$ $^{4}C_{1}$ $^{1}C_{4}$ $^{4}C_{1}$ $^{1}C_{4}$	2HYV	-19.74	+2.52	-11.33
Pentasaccharide[5] **IdoA2S – GlcNS6S – IdoA2S – GlcNS6S – IdoA2S** $^{1}C_{4}$ $^{4}C_{1}$ $^{2}S_{0}$ $^{4}C_{1}$ $^{2,5}B$	1QQP	-9.31	-5.33	-0.91
Hexasaccharide[6] **ΔUA2S – GlcNS6S – IdoA2S – GlcNS6S – IdoA2S – GlcNS6S** $^{1}H_{2}$ $\mathbf{^{4}C_{1}}$ $\mathbf{^{1}C_{4}}$ $\mathbf{^{4}C_{1}}$ $\mathbf{^{2}S_{0}}$ $\mathbf{^{4}C_{1}}$	1BFC	+4.49	+21.68	+15.0
Hexasaccharide[6] **ΔUA2S – GlcNS6S – IdoA2S - GlcNS6S – IdoA2S – GlcNS6S** $^{1}H_{2}$ $^{4}C_{1}$ $^{1}C_{4}$ $^{4}C_{1}$ $^{2}S_{0}$ $^{4}C_{1}$	1XT3	-14.67	-5.42	-4.09
Construct 1[7] **GlcNS6S – IdoA2S – GlcNS6S – IdoA2S – GlcNS6S – IdoA2S** $^{4}C_{1}$ $^{1}C_{4}$ $^{4}C_{1}$ $^{1}C_{4}$ $^{4}C_{1}$ $^{1}C_{4}$	2HYV	-18.25	-4.20	-7.66

[1] Oligosaccharides structures are shown with the ring conformations of the monosaccharide units given below.

[2] Monosaccharide units are identified as A – B

[3] Monosaccharide units are identified as A – C

[4] Monosaccharide units are identified as A – D

[5] Monosaccharide units are identified as A – E

[6] Monosaccharide units are identified as A – F

[7] Monosaccharide units are identified as A', A – E. The structure A – E is identical to the 2HYV pentasaccharide.

* In the crystal structure this residue is ΔUA2S, for modelling purposes the double bond was removed but the $^{1}H_{2}$ conformation was retained.

an alternative conformation of the pentasaccharide extracted from a different crystal structure (PDB structure 1QQP) was docked.

Docking of the pentasaccharide extracted from PDB structure 1QQP showed a reasonable low binding affinity as compared to the pentasaccharide extracted from PDB structure 2HYV due to the difference in the conformation of their iduronic rings and their glycosidic angles. The 2-*O*-sulfate groups in subunits D and E of the pentasaccharide make electrostatic interactions with Lys 176 and Arg 179, whereas the carboxylate groups of residue A and E make electrostatic interactions with the side chain of His 239 and the backbone of Leu 177. Additional hydrogen bonding interactions were detected between the hydroxyl groups of subunits A, C and D and residues Lys 255, Thr 257 and Gln 259, respectively. The best docking pose obtained with the pentasaccharide extracted from PDB 1QQP resulted in a free energy of binding of -0.91 kcal/mol (see Table 4.1), with a significant loss in interactions between the GlcNS6S of residue B and Ig-domains 2 and 3 of PECAM-1.

Docking of heparin hexasaccharide fragments to Ig-domains 2 and 3 was also performed using several conformations. Docking results using the hexasaccharide conformation extracted from PDB structure 1XT3 resulted in a worsening of the binding affinity. The top ranking binding pose resulted in a free energy of binding of -4.09 kcal/mol (see Table 4.1). Analysis of the binding modes of the hexasaccharide fragment with respect to Ig-domains 2 and 3 in the conformations extracted from PDB structures 1XT3 and 1BFC reveals that it has significantly fewer electrostatic interactions and hydrogen bonds than the pentasaccharide fragment in the conformation extracted from PDB structure 2HYV. In the case of Construct 1, docking was carried out with restrained inter-glycosidic torsions but flexible substituents (except hydroxyl groups). A significantly more favourable predicted free energy of binding of -7.66 kcal/mol was computed as shown in Table 4.1. This free energy of binding was corrected to take into account the missing conformational entropy arising from the additional glycosidic bond constraints with respect to the docking of smaller fragments. The predicted binding mode of Construct 1 was similar to that of the pentasaccharide fragment from which it was constructed (PDB structure 2HYV), with an RMSD of 0.759 Å.

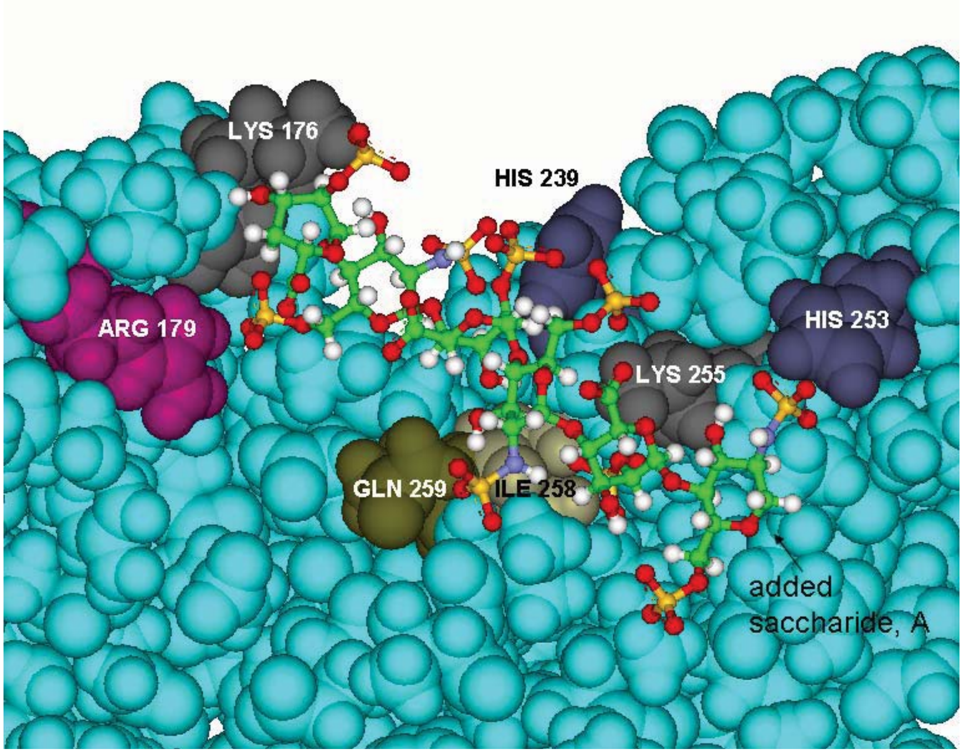


Figure 4.9. *Predicted binding mode of the heparin hexasaccharide (Construct 1) to Ig-domains 2 and 3 of PECAM-1. The interactions between the hexasaccharide fragment and the protein are similar to those predicted to be important for the binding of the pentasaccharide fragment, with the additional sixth residue interacting with His 253. The protonated histidines (purple in color) and arginine(magenta in color) are shown in CPK representation and the Ig-domains are also represented as CPK.*

Despite its greater length, the hexasaccharide fragment Construct 1 makes the same interactions with Ig-domains 2 and 3 through its five saccharides, with the *N*-sulfate of residue 6 establishing an additional electrostatic interaction with the protonated $N_{\varepsilon 2}$ in His 253, as shown in Figure 4.9. In this longer fragment, the other 6-*O*-sulfate group and sugar of subunit 6 point away from the protein surface and does not contribute to an increase in binding affinity with PECAM-1. The weaker electrostatic interactions of Construct 1 with the surface of the molecule have contributed to the relatively large RMSD value as compared to pentasaccharide. The different binding modes of the various hexasaccharides seem to arise as a consequence of differences in the conformation of their iduronic acid subunits. In the case of the hexasaccharide in the conformations extracted from PDB structures 1XT3 and 1QQP the iduronic acid subunits are in the $^{2}S_{o}$ conformation, whereas all the iduronic acid subunits in Construct 1 have a $^{4}C_{1}$ conformation.

The intermolecular and free energies for each heparin fragment predicted by the docking simulations have also been plotted (Figure 4.10). These energies correspond to oligosaccharides that have similar conformations as it became obvious (see previous discussion) that alternative subunit conformations resulted in significantly higher (less favourable) energies of interaction. Hence the energies plotted for the penta and hexasaccharides correspond to the conformations extracted from PDB structure 2HYV and Construct 1 (Figure 4.10), respectively. The shape of the energy plots suggests that the optimum size required for a GAG fragment to have maximal affinity of binding to Ig-domains 2 and 3 is the pentasaccharide, with a clear preference for iduronic acid subunits in either $^{4}C_{1}$ or $^{1}C_{4}$ conformations. Comparison of the docking of various heparin fragments suggests that five saccharides are critical for recognition and binding of heparin to Ig-domains 2 and 3 of PECAM-1 but no experimental evidence of this has been obtained. The comparison between the pentasaccharide, 2HYV, and the hexasaccharide, Construct 1, which includes the 2HYV structure, suggests that key interactions occur between six sulfates in the 2HYV structure and, although saccharide A’ of Construct 1 established an additional electrostatic interaction with His 253, no increase in binding affinity was recorded.

These observations are consistent with previous crystallographic studies of annexin A2-heparin complexes, wherein the electron density beyond the pentasaccharide was not observed (Shao *et al.* 2006), suggesting a high level of disorder due to either a large degree of molecular flexibility and/or weak binding to the protein. This is also consistent with previous suggestions that proteins with patches of basic residues on their surface at a distance of about 20 Å between each other can best accommodate a heparin pentasaccharide (Margalit *et al.* 1993). Overall, these results indicate that the conformational flexibility of sugar residues, the chemical identity of the saccharide subunits, the linkage between glucosamine and uronic acid, the distribution of sulfate groups along the chain and the substitution pattern of sugar residues attached to the non-reducing end of the oligosaccharide may play a key role in the recognition and binding properties of heparin fragments to a protein molecule.

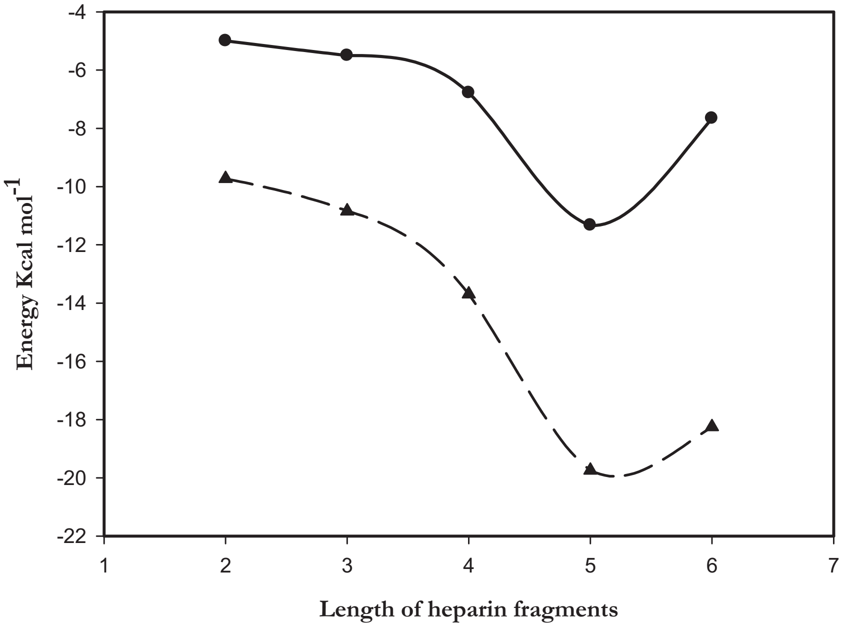


Figure 4.10. *Plot of the free energy (filled circles) and intermolecular energy (filled triangles) predicted by AutoDock for the binding of heparin fragments to Ig-domains 2 and 3 in PECAM-1. The trend lines are indicative only. The energies for penta and hexasaccharide refer to structures 2HYV and Construct 1, respectively.*

Chapter 5

MM/PBSA SIMULATIONS

This chapter details the use of the MM/PBSA method within MD simulations in explicit solvent to predict the free energy of interaction between heparin fragments and the high and low affinity GAG-binding regions of the PECAM-1 receptor. The simulations also provide information about the conformational changes that affect the receptor and heparin fragments upon their interaction.

5.1 MM/PBSA ANALYSIS

The coordinates of Ig-domains 2-3 and 5-6 were extracted from the homology model of the extracellular domains of PECAM-1 discussed in Chapter 3. Complexes of a heparin pentasaccharide consisting of a sequence of IdoAp2S(1→4)GlcNpS6S(1→4)IdoAp2S(1→4)GlcNpS6S(1→4)IdoAp2S with Ig-domains 2-3, and of a disaccharide consisting of a sequence of IdoAp2S(1→4)GlcNpS6S with Ig-domains 5-6 were taken from our previous docking simulations, as discussed in Chapter 4. All histidine sidechains in the binding regions were protonated as these residues have been determined to be essential for the interaction of GAGs at slightly acidic pH (see Chapter 4).

The docking simulations discussed in Chapter 4 identified a number of residues in Ig-domains 2, 3, 5 and 6 that are involved in the binding of heparin. The docking simulations identified a high affinity region in Ig-domains 2 and 3 involving Lys 176, Leu 177, Arg 179, His 239, Lys 255, Gln 259 and Ile 258 (Figure 5.1), and a low affinity region in domains 5 and 6 involving residues Lys 423, Lys 446, Lys 449, Asn 467, Arg 577 and His 580. The MD simulations of the disaccharide fragment described here considered the third cluster found in the docking simulations, which involve interactions with both Ig-domains 5 and 6 and were predicted to have a free energy of binding and dissociation constant of -6.13 kcal/mol and 32.2 μM, respectively (Figure 5.2).

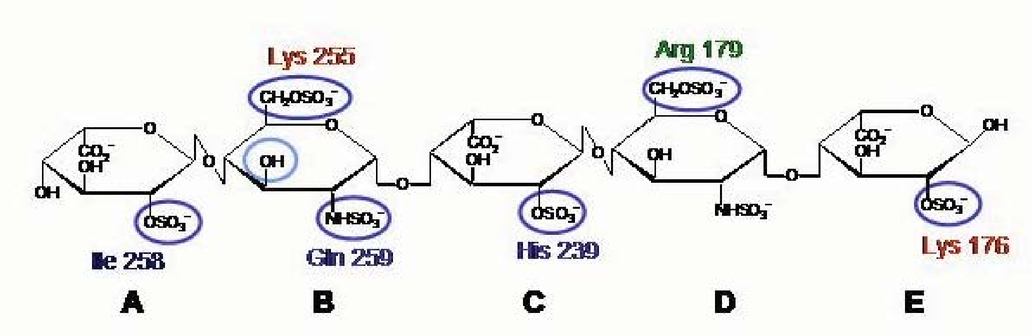


Figure 5.1. *Schematic representation of a heparin pentasacchride ABCDE which was docked to Ig-domains 2 and 3 of PECAM-1. Circled anionic groups are critical for high-affinity interactions with Ig-domains 2 and 3. The amino acids with which they interact are indicated.*

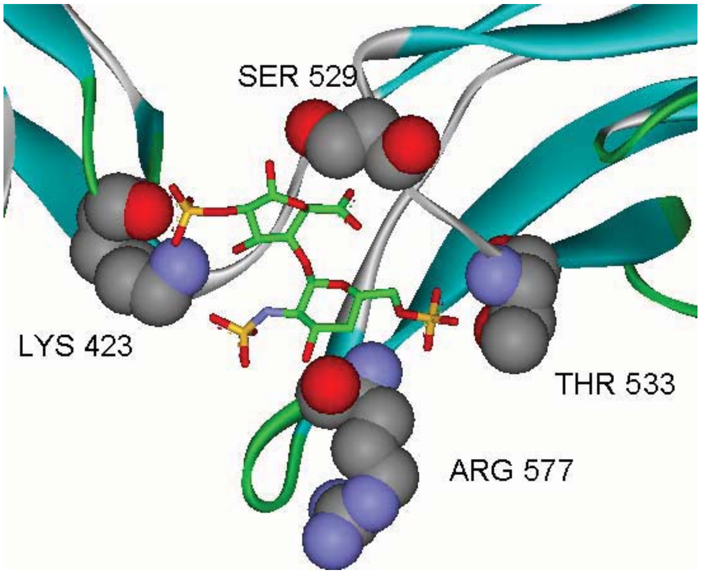


Figure 5.2. *Predicted binding mode of a heparin disaccharide docked to Ig-domains 5 and 6. The disaccharide fragment is shown in sticks.*

5.1.1 Parameterisation of the AMBER/GLYCAM force field for heparin fragments.

The Parm94 (Cornell *et al.* 1995) force field in AMBER 9.0 (Case *et al.* 2005) was used with the GLYCAM04 extension for carbohydrates (Woods *et al.* 1995). Existing non-bonded parameters for sulfates and sulfamates were used (Huige & Altona 1995). Some parameters such as bond, angles and torsion parameters that were not available for sulfates were approximated by taking those for phosphates available in the GLYCAM04 force field.

Partial atomic charges for the disaccharide (Figure 5.3) and pentasaccharide (Figure 5.4) were obtained using the restricted electrostatic potential (RESP) method (Bayly *et al.* 1993; Cornell *et al.* 1993) using the *leap* and *sander* modules from AMBER 9.0. For this purpose both molecules were initially subjected to a full geometry optimisation with a 6-31G* basis set using Gaussian 98 (Frisch *et al.* 1998). A SCF convergence criterion of 10^{-8} kcal/mol and a 'tight' optimisation threshold were used. The resulting minimum energy conformation of each saccharide was then subjected to a single point energy calculation with a 6-31G* basis set and the POP=CHelpG charge option. The resulting RESP partial charges of these oligosaccharides are shown in Figure 5.3 for the disaccharide and Figure 5.4 for the pentasaccharide.

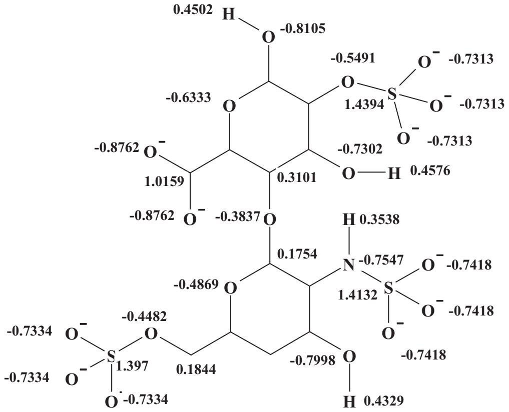


Figure 5.3. *Partial atomic RESP charges of a heparin disaccharide and pentasaccharide. The charges on the hydrogens and carbons of the pyranose ring are not shown for clarity.*

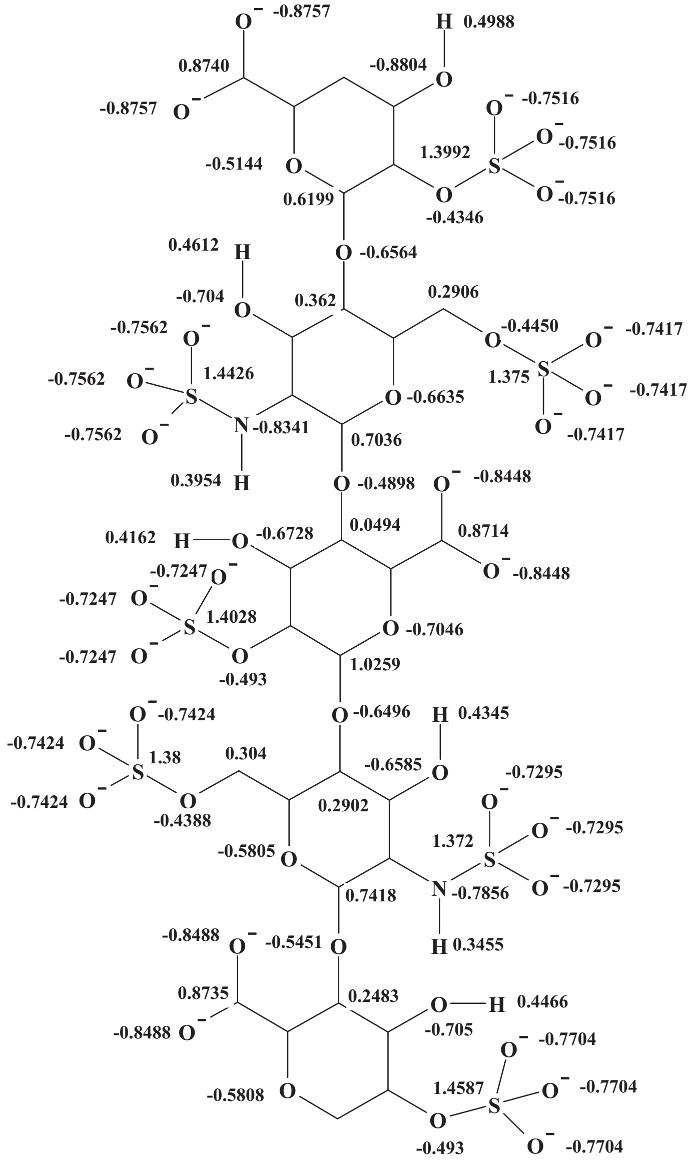


Figure 5.4. *Partial atomic RESP charges of a heparin pentasaccharide. The charges on the hydrogens and carbons of the pyranose ring are not shown for clarity.*

5.1.2 MD simulations.

Following the MM/PBSA protocol (see below), separate MD simulations were carried out for the relevant Ig-domains of PECAM-1, the heparin fragment of interest, and a complex between the two. During heating and equilibration, weak restraints (with a force constant of 25 kcal/(molxÅ^2) were applied to all heavy atoms in the protein domains, except those in the binding sites. The homology modelling studies indicated that the receptor may exist in an open or closed conformation due to the presence of loops connecting the two domains (Chapter 3). As a consequence, full flexibility of the receptor and ligand were allowed during the production stage of the simulations. The binding site regions included residues 176-182, 207-209, 250-260 and 278-288 of Ig-domains 2 and 3. The selected complex of a dissacharide with Ig-domains 5 and 6 has the following initial interactions: the 2-*O*-sulfate of IdoA2S makes an electrostatic interaction with Lys 423, the 6-*O*-sulfate of GlcNS6S makes hydrogen bonds with the backbone of Thr 533 and Arg 577, and the *N*-sulfate makes an electrostatic interaction with Lys 423.

All energy minimisations and MD simulations were performed using the AMBER 9.0 program (Case *et al.* 2005). A box of TIP3P water molecules (Jorgensen *et al.* 1983) was added to solvate the complex, keeping a minimum distance of 12.0 Å between each face of the box and the solute. The number of water molecules added to the complex of the pentasaccharide with Ig-domains 2 and 3 was 7181, whereas 5651 water molecules were added to the complex of the disaccharide with Ig-domains 5 and 6. Net charges in the protein or heparin fragment were neutralised by adding counter ions (Na^+ or Cl^-) as required. The Particle Mesh Ewald (PME) method was used to compute long range electrostatic interactions (Tom *et al.* 1993), using a 1.0 Å grid spacing and a fourth-order spline for interpolation. The non-bonded cutoff was set to 8.0 Å and the SHAKE algorithm (Ryckaert *et al.* 1977) was used to constrain all bonds involving hydrogen atoms. The MD simulation was carried out using isobaric-isothermal ensemble (NPT). Isotropic scaling for pressure regulation was used. The external pressure was set to 1 atm. The temperature was kept at 300 K using Langevin dynamics (Pastor *et al.* 1988) with a collision frequency of 2 ps^{-1}. A timestep of 1.0 fs was used in all simulations. Periodic boundary conditions were applied throughout.

In each simulation initial unfavourable contacts with the solvent were removed by energy minimisation after performing 10 steps of steepest descents followed by 990 steps of conjugate gradients. A 150-ps period of simulated annealing was then carried out, during which the temperature was raised from 5 to 300 K over 50 ps, with a further 50 ps at 300 K, before cooling back to 5 K over 50 ps. The system was energy minimised again as before, followed by heating from 5 K to 300 K over 50 ps, upon which the systems were deemed to have equilibrated. The production phases of the simulations without constraints were then run at 300 K for 8.0 ns for the complex and protein, and for 4.0 ns for the heparin fragments. Various parameters (density, temperature, pressure, kinetic energy and potential energy) were monitored during the simulations to ensure that proper equilibration had been achieved.

For the 8.0 ns simulations 800 snapshots were taken at regular intervals for the binding energy analyses and post-processed after removing all solvent molecules and counter ions. The free energies of binding reported are averages over these 800 snapshots or portions thereof.

5.1.3 MM/PBSA calculations.

The MM/PBSA module of AMBER 9.0 was used to compute the components of the free energy as dicssued in Chapter 2. For the 8.0 ns simulations, 800 snapshots of the coordinates of the system were taken at 10 ps intervals. All solvent molecules and counterions were removed in order to analyse the snapshots. These snapshots were analysed with the modified GB solvation model (Tsui & Case 2001), modified for use with the PARM94 parameters to obtain energies of solvation. Poisson-Boltzmann calculations were also used to obtain solvation energies, with an ionic strength of 0.14 M, a dielectric constant (ε) of 1 for the solute and 80 for the solvent. A probe solvent radius of 1.4 Å and the PARSE atomic radii parameter set (Sitkoff *et al.* 1994) were used to determine the molecular surface. Different surface parameters were used: in the case of GB calculations, $\gamma = 0.0072$ kcal/Å^2 and $b = 0.0$ kcal/mol, and in the case of PB calculations, $\gamma = 0.00542$ kcal/Å^2 and $b = 0.92$ kcal/mol.

The vibrational entropy of the systems was computed by performing normal modes calculations (Kollman *et al.* 2000) using the Nmode module of AMBER (Kollman *et al.* 2000) on 40 snapshots, corresponding to 200 ps intervals. In the case of the simulations of Ig-domains 2/3, 20 snapshots were collected for each 2 ns portion of the trajectory. Prior to these normal modes calculations, the selected snapshots of the complex, protein and ligand were subjected to a full conjugate gradient energy minimisation using a ε = 4r and a convergence criterion of 0.0001 kcal/mol. The reported vibrational entropies are the averages over all selected snapshots.

MM/PBSA calculations were carried out using only the 8.0 ns trajectories of the heparin fragment complexes with either Ig-domains 2/3 or 5/6. This was done as the protein exhibited significant conformational changes, which made it difficult to compare the various free energy contributions from independent simulations of the protein and saccharides.

5.2 RESULTS

Upon equilibration, the temperature and potential energy were monitored during the course of the simulations. Figure 5.5 (A and B) and Figure 5.6 (A and B) show corresponding plots for the simulations of the pentasaccharide complexed to Ig-domains 2 and 3 and the disaccharide complexed with Ig-domains 5 and 6, respectively. It can be seen that the potential energy and temperature fluctuate around converged average values. Recently, the role of potential energy as a function of conformation has been reviewed using MM/PBSA methods (Gilson & Zhou 2007). In the case of Ig-domain 2 and 3 of PECAM-1, its fluctuations are related to changes in the structure of the protein, as discussed below (see Figure 5.7).

The root mean square deviation (RMSD) of the coordinates of the heparin fragment complexes with their corresponding Ig-domains in each snapshot of the simulation with respect to the coordinates in the initial snapshot were also monitored. A significant amount of backbone motion (up to ~ 9 Å) in Ig-domains 2 and 3 can be observed in Figure 5.5 (C). The high RMSD values indicate that there is a significant conformational change in Ig-domains 2 and 3. Similarly, the high RMSD values (up to ~ 9 Å) observed with Ig-domains 5 and 6 indicate a conformational change in the backbone structure of domains Figure 5.6 (C).

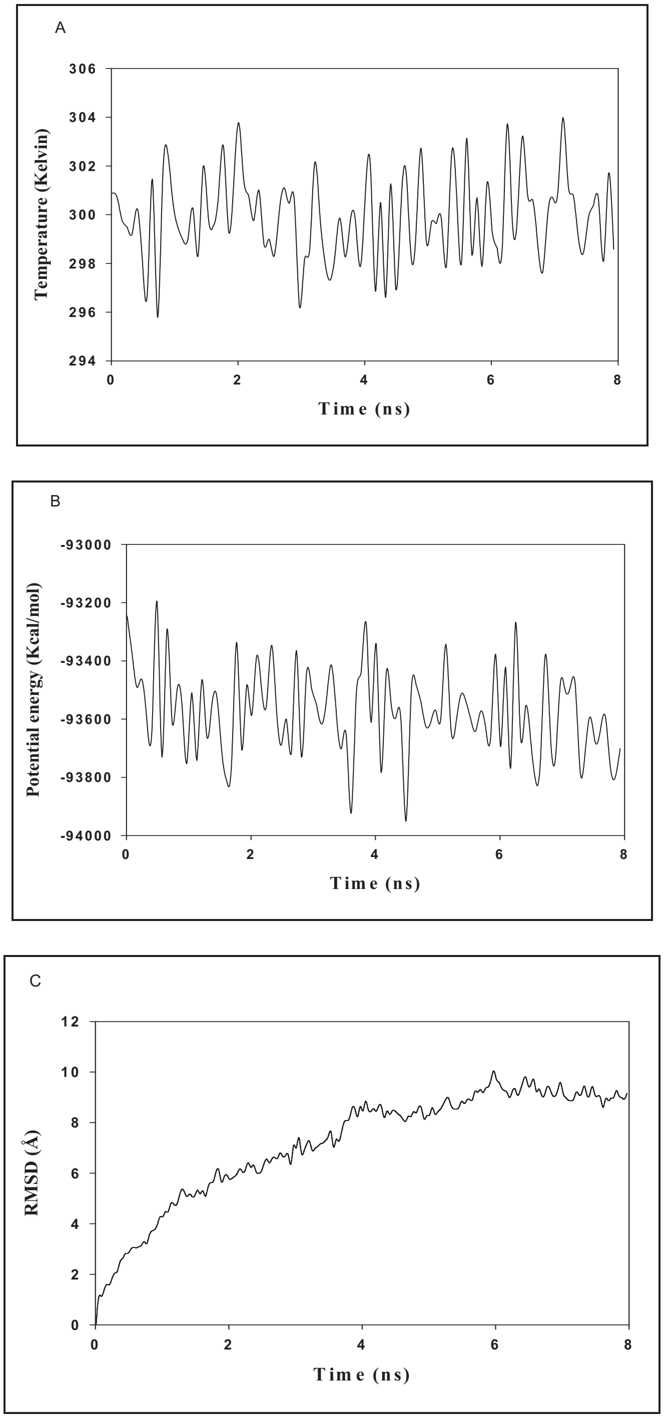


Figure 5.5. *Time evolution of MD simulations of the complex of a heparin pentasaccharide complex with Ig-domains 2 and 3. (**A**) Temperature, (**B**) Potential energy, (**C**) RMSD of the coordinates of protein main chain atoms (C_{α}, C and N) in each snapshot with respect to the coordinates in the first snapshot.*

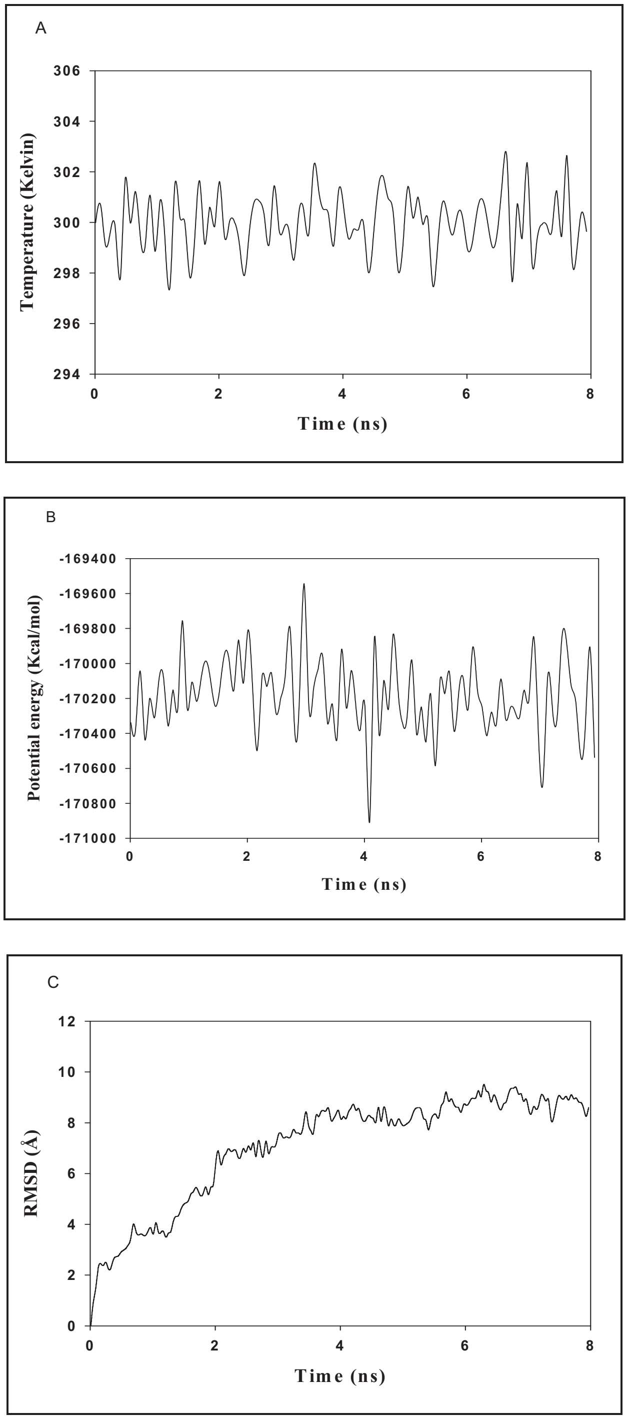


Figure 5.6. *Time evolution of MD simulations of the complex of a heparin disaccharide complex with Ig-domains 5 and 6. (**A**) Temperature, (**B**) Potential energy, (**C**) RMSD of the coordinates of protein main chain atoms (C_{α}, C and N) in each snapshot with respect to the coordinates in the first snapshot.*

5.2.1 Structural analysis of the Ig-domains of PECAM-1 and their interactions with heparin fragments.

The docking studies discussed in Chapter 4 suggested that electrostatic interactions between heparin fragments and positively charged residues located on the surface of Ig-domains 2 and 3 were responsible for the existence of a high affinity GAG binding region in PECAM-1. This GAG binding region involves major contributions from Ig-domain 3 (residues His 239, Lys 255 and Gln 259), with further contributions from Ig-domain 2 (residue Arg 179).

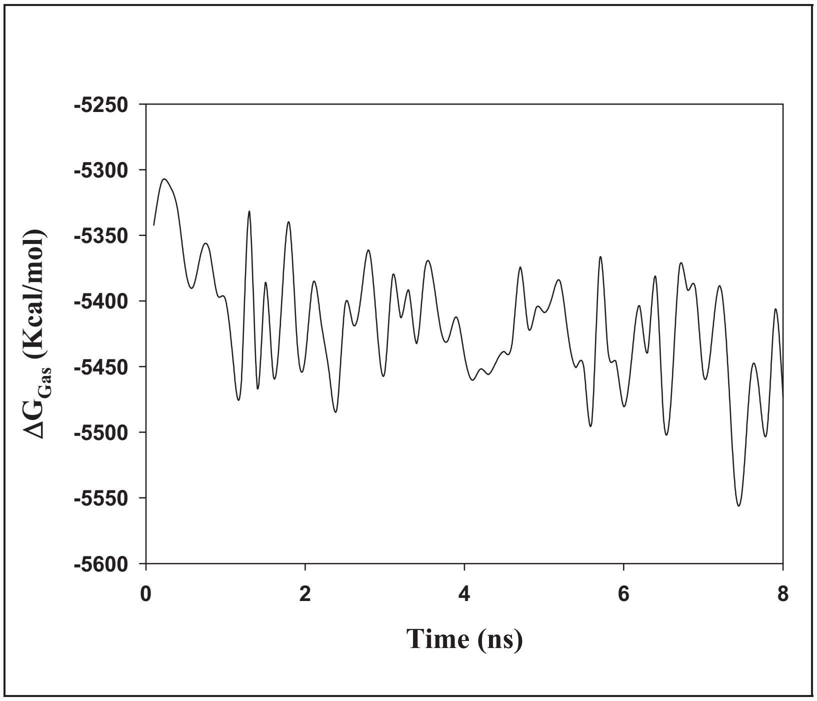


Figure 5.7. *Time evolution of internal potential energy of Ig-domains 2 and 3. The gradual decrease in the potential energy is associated with a favourable conformational change. The plot has been scaled by a factor of 10 for clarity.*

Analysis of the simulation trajectory of the complex of the heparin pentasaccharide fragment with Ig-domains 2 and 3 revealed that the 2-*O*-sulfate and carboxylate of the first IdoA2S (residue A) form ionic interactions with the positively charged side chains of Lys 287 and His 239 (protonated $N_{\varepsilon 2}$), respectively. The first GlcNS6S (residue B) is not involved either in electrostatic interactions or hydrogen bonding. The 2-*O*-sulfate group of the second IdoA2S (residue C) forms a strong hydrogen bond with the side chain of Gln 259. The *N*-sulfate of the second GlcNS6S (residue D) forms an ionic interaction with the positively charged side chain of Lys 237. The 6-*O*-sulfate group in this residue makes a strong hydrogen bond with the backbone of Val 175. The third IdoA2S (residue E) is not involved in any interaction. The interactions with amino acids Lys 176 and Arg 179 observed in earlier docking studies were not seen in the MD simulations due to the conformational change in the hinge region connecting the Ig-domains 2 and 3.

Analysis of the first 2 ns of the simulation trajectory showed that the carboxylate and 2-*O*-sulfate of IdoA2S at position E make electrostatic interactions with the protonated $N_{\varepsilon 2}$ in His 239 and the positively charged Lys 287, respectively. The *N*-sulfate of GlcNS6S at position B and the 2-*O*-sulfate of IdoA2S at position C also make electrostatic interactions with Lys 237. The 6-*O*-sulfate of GlcNS6S makes a strong hydrogen bond with the backbone of Val 175. None of the sulfates in residues A and D make any electrostatic or hydrogen bonding interactions with Ig-domains 2 and 3, in contrast to what was observed in the docking simulations discussed in Chapter 4. Analysis of the simulation trajectory between 2-4 ns revealed that the pentasaccharide retains the interactions of residues B, C and E with the protein observed during the first 2 ns. In addition, the 2-*O*-sulfate of the first IdoA (residue A) makes a further ionic interaction with Lys 176.

The last 4 ns of the simulation trajectory showed the formation of favourable electrostatic interactions of the ligand with the receptor. The 2-*O*-sulfate of the first IdoA2S (residue A) makes an electrostatic interaction with Lys 255. The 6-*O*-sulfate of the second GlcNS6S (residue D) makes an electrostatic interaction with His 239. The *N*-sulfate in the first GlcNS6S (residue B) makes an electrostatic interaction with Lys 237 and the 2-*O*-sulfate of the second IdoA2S (residue C) makes a strong hydrogen bond with Gln 259. The 2-*O*-sulfate in the first IdoA2S (residue A) makes further electrostatic interactions with His 162 in Ig-domain 2.

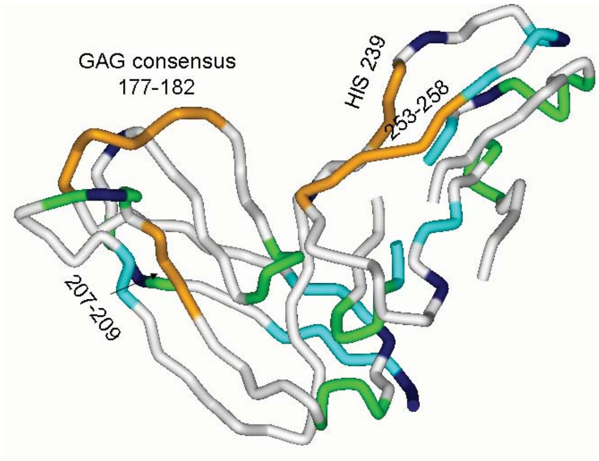


Figure 5.8. *Tube representation of the average structure of Ig-domains 2 and 3 obtained during the 0-2 ns portion of the simulation. Coils are coloured in green and the beta sheets in blue. Purple colour marks the presence of glycines and prolines. Sulfate binding regions are shown in orange. The grey colour shading indicates the loss of beta propensity.*

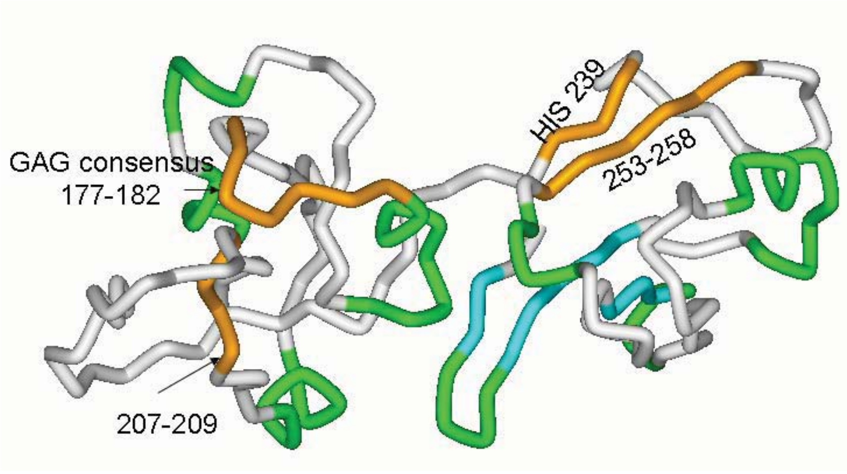


Figure 5.9. *Tube representation of the average structure of Ig-domains 2 and 3 obtained during 2-4 ns portion of the simulation. Coils are coloured in green and the beta sheets in blue. Sulfate binding regions are shown in orange. The grey colour shading indicates the loss of beta propensity.*

Inspection of the whole 8.0 ns trajectory of the complex of the pentasaccharide with Ig-domains 2 and 3 revealed the occurrence of a significant conformational change in Ig-domains 2 and 3 of PECAM-1. There is a transition from the predominant beta-sheet structure of Ig-domains 2 and 3 to a disordered random coil structure (Figure 5.8 and Figure 5.9). This transition is similar to the one observed in the globular structure of Fibronectin-III (FN-III), where its beta sheeted Ig-domains adopt a random coil structure (Penkett *et al.* 1997) at acidic pH in solution. This conformational transition in Ig-domains can be rationalised by the presence of large numbers of glycine residues (which impart conformational flexibility) and proline (which have structure breaking properties).

A hinge region is present in the binding site shared by Ig-domains 2 and 3 of PECAM-1, which can open up to expose more basic residues that may interact with a longer heparin oligosaccharide. As a consequence, during the simulation a hinge movement opened up and increased the size of the binding site (see Figure 5.10). This conformational change may allow a longer oligosaccharide (such as an octasaccharide) to interact with basic residues such as Lys 176 and Arg 179. Since the change in conformation of the binding site is likely to affect the computed free energy of binding of the pentasaccharide, the MM/PBSA analysis described further below was carried out separately for four 2 ns portions of the whole trajectory with a normal mode analysis of 20 snapshots.

Docking studies suggested the existence of a low affinity GAG-binding region in Ig-domain 5, with contributions from Ig-domain 6 (Chapter 4). The MD simulations have confirmed that the interactions between the heparin disaccharide and the protein are stable. Analysis of the 8 ns trajectory revealed that the carboxylate of Glu 527 makes a strong hydrogen bond with the amine of GlcNS6S, whereas the main chain NH of Gly 528 makes a hydrogen bond with the 2-*O*-sulfate of the iduronic acid, resulting in the loss of the electrostatic interaction between the 2-*O*-sulfate of IdoA2S and Lys 423. The 2-*O*-sulfate of IdoA2S makes a hydrogen bond with the sidechain of Ser 529, the 6-*O*-sulfate of GlcNS6S makes an electrostatic interaction with Lys 423 and the amine of GlcNS6S makes a strong hydrogen bond with Glu 527. No interactions between the disaccharide and Thr 533 and Arg 577 were detected, as opposed to what was seen in docking simulations discussed in Chapter 4. This loss of interactions of the disaccharide with the Ig-domains 5/6 is due a conformational change from a predominantly β-sheet structure in Ig-domains 5/6 to a disordered random coil structure (Figure 5.6 C).

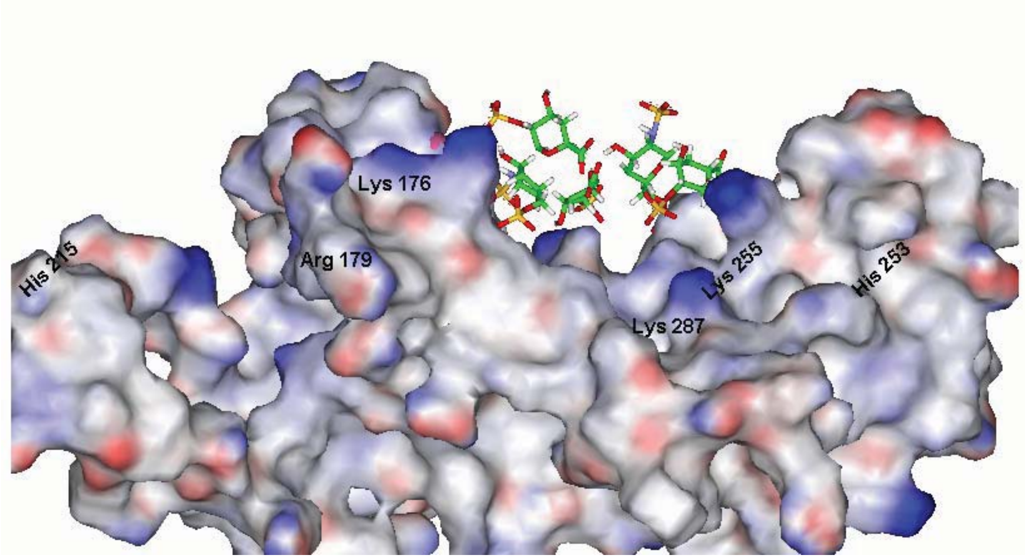


Figure 5.10. *Final binding mode of a heparin pentasaccharide complexed with Ig domains 2 and 3 of PECAM-1 after 8 ns. The open conformation of the Ig-domains can interact with a longer heparin fragment through its basic residues exposed on the surface. The Ig-domains 2 and 3 of human PECAM-1 are represented with an solvent accessible surface (negative potential in red and positive potential in blue). The potential surfaces were calculated and displayed using the DELPHI module in Discovery Studio (Accelrys, Inc.). The pentasaccharide fragment is shown as sticks. The glycosidic bonds of the pentasaccharide are not shown for clarity.*

5.2.2 Structural analysis of the conformation of free and protein-bound heparin fragments.

An analysis and comparison of the structure of the heparin pentasaccharide bound to Annexin-A2 (PDB code 2HYV), bound to Ig-domains 2 and 3 of PECAM-1 and in aqueous solution were carried out to investigate any differences in the conformation of the glycosidic linkages between each saccharide monomer. The conformations of the heparin pentasaccharide bound to Ig-domains 2 and 3 and in aqueous solution were analysed from the 4 ns trajectories. Table 5.1 lists the average values of the four glycosidic torsion angles in the pentasaccharide for each case.

Table 5.1. Average values of the glycosidic torsion angles for the heparin pentasaccharide extracted from Annexin crystal structure 2HYV, the pentasaccharide complexed with Ig-domains 2 and 3 of PECAM-1 after 4 ns and the pentasaccharide in solution after 4 ns.

Glycosidic Linkage	Torsion angles	PBD structure 2HYV	Protein-bound	In solution
α (1,4)$_1$	Ψ	–0.16	140.12 (95.31)	156.58 (68.72)
	φ	56.18	11.19 (9.35)	14.10 (8.40)
α(1,4)$_2$	Ψ	–23.12	–27.78 (7.89)	–38.08 (8.88)
	φ	–36.56	–23.39 (10.29)	–41.09 (8.39)
α(1,4)$_3$	Ψ	–39.05	–1.00 (17.08)	8.93 (11.89)
	φ	0.25	55.61 (8.81)	48.91 (11.73)
α(1,4)$_4$	Ψ	–21.77	10.64 (8.01)	–24.66 (13.50)
	φ	–45.26	28.02 (9.35)	–34.99 (10.51)

The Ψ and φ angles in α (1, 4) linkages are defined as C1-Ox-Cx-Hx and H1-C1-Ox-Cx, respectively. The standard deviations are shown in brackets.

The glycosidic linkage $\alpha(1,4)_1$, $\alpha(1,4)_2$, $\alpha(1,4)_3$ and $\alpha(1,4)_4$ of the pentasaccharide exhibited greater fluctuations when bound to PECAM-1 than in aqueous solution. The torsion angles Ψ of the $\alpha(1,4)_1$ linkage of the pentasaccharide in aqueous solution as well as in the protein bound structure exhibited large fluctuations due the change in the conformation of the first iduronic acid residue (present in the $^1H_2$ conformation). The conformation of the glycosidic linkage $\alpha(1,4)_2$ in solution and in bound form to the Ig-domains 2 and 3 is similar to that found in the crystal structure of Annexin-A2. There is no clear pattern in the conformations of linkage $\alpha(1,4)_3$ when comparing the structures in solution, protein bound or Annexin-A2. A similar conformation is observed in the $\alpha(1,4)_4$ glycosidic linkage between the crystal structure of Annexin-A2 and the aqueous solution form. These observations suggest that the heparin pentasaccharide undergoes a conformational change upon binding to a protein. In the case of Annexin-2, this conformational change may be also due to interactions between the heparin fragment and calcium ions present on the protein surface. The larger fluctuation in the case of the PECAM-1 bound pentasaccharide fragment with respect to the conformation in aqueous solution is likely to be due to the change in the receptor conformation, as described above. Comparison of these average glycosidic linkages with values obtained by NMR measurements, MD simulations and crystal structures of heparin bound proteins like aFGF (Mikhailov *et al.* 1997) shows that the $\alpha(1,4)_{2,}$ $\alpha(1,4)_3$ and $\alpha(1,4)_4$ linkages remain quite stable, whereas rather large changes in the $\alpha(1,4)_1$ linkage are observed. The fluctuations in the $\alpha(1,4)_1$ linkage may occur because of the modification at the non-reducing end of the unsaturated UA2S to create a 4-deoxy IdoA2S residue (4dIdoA2S), since the residue adopts a different conformation to its original ${}^1H_2$ conformation (see Chapter 4).

5.2.3 Calculations of the free energy of binding.

Docking studies predicted that the free energy of binding of the heparin pentasaccharide fragment to Ig-domains 2 and 3 is -17.22 kcal/mol, which results in a predicted dissociation constant of 4.93 nM. Table 5.2 summarises the results of the calculations of the free energies of binding using the MM/PBSA and MM/GBSA methods. Calculation of energy terms was carried out for 800 snapshots whilst vibrational entropy calculations were done on 40 snapshots. The predicted free energies of binding were -18 kcal/mol with MM/PBSA and -15 kcal/mol with MM/GBSA, which translate into dissociation constants of 0.43 and 3.85 pM, respectively.

These affinity values for the heparin pentasaccharide fragment are influenced by changes in the conformation of the Ig-domain. The existence of an open conformation suggests that a longer oligosaccharide with appropriate sulfation may interact with Ig-domains 2/3 with varying affinity. The calculated $\Delta G_{binding\text{-}PBSA}$ for 200 snapshots for different portions of the trajectory were -6.7 kcal/mol for 0-2 ns (Table 5.3), -10.76 kcal/mol for 2-4 ns (Table 5.4), -22.81 kcal/mol for 4-6 ns (Table 5.5) and -19.79 kcal/mol for 6-8 ns (Table 5.6). The vibrational entropy change contributions to the free energy of binding are nearly of the same magnitude as the other terms, suggesting that both enthalpy and entropy play a key role in determining the free energy of binding.

There is a gradual decrease in free energy of binding as the simulation progresses. During the first 4 ns of the simulation there is an absence of interactions between residues A and D of the pentasaccharide and the protein. The binding affinity then increases as Ig-domains 2/3 of the receptor adopt an open conformation and the pentasaccharide makes more electrostatic and hydrogen bonding interactions with residues Lys 255, His 239, Lys 237, Gln 259 and His 162 during the last 4 ns of the simulation. There is a gradual decrease in the potential energy of Ig domains 2/3 along the simulation (Figure 5.7). This is also mirrored by a gradual decrease in the free energy of solvation of the pentasaccharide (Figure 5.11) and the protein. This suggests that the conformational change is thermodynamically favourable and is independent of the interactions with the heparin fragment.

Docking studies predicted that the free energy of binding of the heparin disaccharide fragment with Ig-domains 5 and 6 is -6.5 kcal/mol, resulting in a dissociation constant of 15.4 μM, suggesting weak binding. Table 5.7 summarises the results of the calculations of the free energies of binding using the MM/PBSA and MM/GBSA methods. The predicted free energies of binding (+4.21 kcal/mol with MM/PBSA and +9.21 kcal/mol with MM/GBSA) translate into dissociation constants in the mM range, indicating very weak binding. In this case, the favourable sum of the interaction and solvation energy terms (PBTOTAL and GBTOTAL) was not large enough to overcome the unfavorable contribution of the vibrational entropy change.

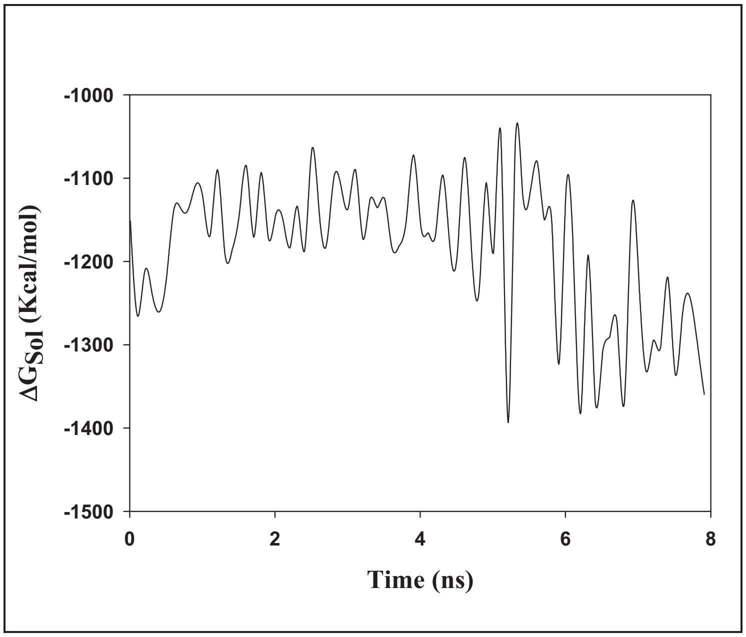


Figure 5.11. *Time evolution of the (polar + non-polar) solvation energy of the heparin pentasaccharide. The graph has been scaled by a factor of 10 for clarity.*

Both MM/PBSA and MM/GBSA calculations suggest that electrostatic interactions contribute significantly to the interactions between heparin fragments and PECAM-1. Interestingly, these calculations also reveal that VDW interactions play an equally important role in driving the interaction with the protein. The results of calculations using MM/GBSA differ only by a few kcal/mol with respect to MM/PBSA energy values.

The PBTOTAL and GBTOTAL energies (the sum of all interaction and solvation terms) for the complex of the heparin disaccharide with Ig-domains 5 and 6 are much lower than those for the complex of the heparin pentasaccharide with Ig-domains 2 and 3. This confirms previous docking and experimental studies indicating the presence of a high affinity GAG binding site in Ig-domains 2 and 3 and of a low affinity GAG binding site in Ig-domains 5 and 6 of PECAM-1.

Table 5.2. MM/PBSA energy component analysis of the interactions of the heparin pentasaccharide with Ig-domains 2-3 averaged over 8 ns[a].

	Complex		**Receptor**		**Ligand**		**Δ**	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
ELE	-4696.88	118.36	-3374.69	142.57	1277.96	32.82	-2600.15	113.69
VDW	-476.13	27.96	-445.64	24.08	3.95	5.28	-34.44	8.48
INT	2154.66	45.29	2092.51	43.86	62.14	9.88	0	0
GAS	-3018.36	118.67	-1727.81	143.34	1344.05	31.43	-2634.6	111.89
PBSUR	73.9	2.67	71.8	2.5	8.11	0.14	-6.01	0.51
PBCAL	-3338.24	127.35	-3635.36	150.61	-2287.91	28.23	2585.03	108.51
PBSOL	-3264.34	125.31	-3563.56	148.67	-2279.8	28.34	2579.02	108.49
PBELE	-8035.12	43.56	-7010.05	36.25	-1009.95	8.92	-15.12	14.83
PBTOT	-6282.69	51.84	-5291.37	49.14	-935.75	8.3	-55.57	8.68
GBSUR	96.95	3.55	94.16	3.32	9.55	0.18	-6.76	0.68
GB	-3321.74	127.89	-3687.6	154.28	-2222.68	28.94	2588.55	108.7
GBSOL	-3224.79	125.25	-3593.44	151.73	-2213.13	29.08	2581.79	108.71
GBELE	-8018.62	38.91	-7062.29	34.05	-944.73	8.7	-11.6	11.9
GBTOT	-6243.15	49.77	-5321.26	49.2	-869.08	8.96	-52.81	7.89
TΔS[b]	-2090.38	10.03	-1987.91	10.47	-139.80	0.37	-37.34	7.88
$\Delta G_{binding\text{-}PBSA}$							-18.23	
$\Delta G_{binding\text{-}GBSA}$							-15.47	

a Average over 800 snapshots from trajectory.

b Entropy calculations were based on normal mode analysis using only 40 snapshots.

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; PBSUR, hydrophobic contribution to solvation free energy for PB calculations; PBCAL, reaction field energy calculated by PB; PBSOL=PBSUR+PBCAL; PBELE=PBCAL+ELE; PBTOTAL=PBSOL+GAS; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GBCAL+ELE; GBTOTAL=GBSOL+GAS; TΔS, T(temperature)* ΔS(sum of rotational, translational and vibrational entropies); $\Delta G_{binding}$, total binding energy of the system.

Table 5.3. MM/PBSA energy component analysis of the interactions of the heparin pentasaccharide with Ig-domains 2-3 averaged during 0-2 ns[a].

	Complex		**Receptor**		**Ligand**		**Δ**	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
ELE	-4751.91	103.21	-3457.97	69.47	1267.03	22.39	-2560.96	107.24
VDW	-499.8	22.7	-462.07	24.48	2.43	4.63	-40.16	7.69
INT	2904.75	46.41	2842.88	46.13	61.87	9.76	0	0
GAS	-2346.95	101.21	-1077.16	75.5	1331.33	20.92	-2601.12	103.57
PBSUR	71.05	1.74	69.14	1.92	8.14	0.11	-6.23	0.36
PBCAL	-3239.32	110.57	-3518.58	63.8	-2275.54	17.98	2554.8	100.39
PBSOL	-3168.26	109.3	-3449.44	63.23	-2267.4	18.05	2548.57	100.47
PBELE	-7991.22	33.79	-6976.55	35.84	-1008.51	8.59	-6.16	12.54
PBTOT	-5515.22	47.64	-4526.6	47.99	-936.07	8.14	-52.54	7.98
GBSUR	93.17	2.31	90.62	2.55	9.59	0.14	-7.05	0.48
GB	-3225.66	109.32	-3568.82	65.23	-2211.85	18.6	2555.01	101.2
GBSOL	-3132.49	107.72	-3478.19	64.66	-2202.26	18.69	2547.96	101.31
GBELE	-7977.56	27.14	-7026.79	28.98	-944.83	8.44	-5.95	10.58
GBTOT	-5479.44	46.07	-4555.35	45.87	-870.93	8.57	-53.16	6.23
TΔS[b]	-2078.93	9.55	-1982.92	5.75	-141.84	0.15	-45.84	9.43
$\Delta G_{binding-PBSA}$							-6.7	
$\Delta G_{binding-GBSA}$							-7.32	

a Average over 200 snapshots from trajectory.

b Entropy calculations were based on normal mode analysis using only 20 snapshots.

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; PBSUR, hydrophobic contribution to solvation free energy for PB calculations; PBCAL, reaction field energy calculated by PB; PBSOL=PBSUR+PBCAL; PBELE=PBCAL+ELE; PBTOTAL=PBSOL+GAS; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GBCAL+ELE; GBTOTAL=GBSOL+GAS; TΔS, T(temperature)* ΔS(sum of rotational, translational and vibrational entropies); $\Delta G_{binding}$, total binding energy of the system.

Table 5.4. Energy MM/PBSA energy component analysis of the interactions of the heparin pentasaccharide with Ig-domains 2-3 averaged during 2-4 ns[a].

	Complex		**Receptor**		**Ligand**		**Δ**	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
ELE	-4731.55	83.16	-3456.73	81.11	1260.58	20.12	-2535.4	74.86
VDW	-487.6	20.98	-452.69	18.87	4.67	4.85	-39.58	6.9
INT	2112.77	38.13	2097.54	36.06	15.23	10.01	0	0
GAS	-3106.39	88.36	-1811.88	86.81	1280.48	21.13	-2574.99	75.55
PBSUR	72.2	1.01	70.49	0.91	8.17	0.07	-6.46	0.3
PBCAL	-3292.78	73.21	-3549.73	67.5	-2273.59	17.82	2530.54	74.52
PBSOL	-3220.58	72.99	-3479.24	67.48	-2265.41	17.87	2524.08	74.41
PBELE	-8024.33	33.58	-7006.46	27.99	-1013.01	7.83	-4.86	12.04
PBTOT	-6326.97	40.86	-5291.13	38.09	-984.93	8.25	-50.91	7.27
GBSUR	94.69	1.34	92.42	1.21	9.64	0.1	-7.36	0.4
GB	-3273.23	72.16	-3595.96	69.86	-2208.2	17.8	2530.93	73.58
GBSOL	-3178.54	71.98	-3503.54	69.89	-2198.57	17.86	2523.57	73.46
GBELE	-8004.79	26.8	-7052.69	22.93	-947.62	8.1	-4.48	8.52
GBTOT	-6284.93	38.49	-5315.42	36.35	-918.08	8.82	-51.42	5.81
TΔS[b]	-2093.52	4.49	-1992.42	8.06	-141.25	0.04	-40.15	10.66
$\Delta G_{binding\text{-}PBSA}$							-10.76	
$\Delta G_{binding\text{-}GBSA}$							-11.27	

a Average over 200 snapshots from trajectory.

b Entropy calculations were based on normal mode analysis using only 20 snapshots.

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; PBSUR, hydrophobic contribution to solvation free energy for PB calculations; PBCAL, reaction field energy calculated by PB; PBSOL=PBSUR+PBCAL; PBELE=PBCAL+ELE; PBTOTAL=PBSOL+GAS; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GBCAL+ELE; GBTOTAL=GBSOL+GAS; TΔS, T(temperature)* ΔS(sum of rotational, translational and vibrational entropies); $\Delta G_{binding}$, total binding energy of the system.

Table 5.5. MM/PBSA energy component analysis of the interactions of the heparin pentasaccharide with Ig-domains 2-3 averaged during 4-6 ns[a].

	Complex		**Receptor**		**Ligand**		**Δ**	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
ELE	-4713.93	128.46	-3397.42	110.24	1267.91	23.43	-2584.42	91.61
VDW	-456.35	20.26	-437.48	18.97	7.11	4.83	-25.98	4.91
INT	2141.04	39.59	2078.84	37.75	62.19	9.73	0	0
GAS	-3029.24	129.92	-1756.06	110.9	1337.21	24.21	-2610.39	92.92
PBSUR	75.56	1.12	72.88	0.98	8.17	0.11	-5.49	0.44
PBCAL	-3356.68	119.87	-3629.63	104.31	-2281.69	20.58	2554.64	88.15
PBSOL	-3281.12	119.12	-3556.75	103.71	-2273.52	20.65	2549.15	87.85
PBELE	-8070.6	27.34	-7027.05	25.18	-1013.78	7.84	-29.78	9.32
PBTOT	-6310.36	38.99	-5312.8	36.29	-936.31	8.27	-61.25	7.92
GBSUR	99.16	1.49	95.6	1.3	9.63	0.15	-6.07	0.59
GB	-3338.42	122.08	-3686.56	108.9	-2213.76	21	2561.89	86.11
GBSOL	-3239.27	121.09	-3590.96	108.12	-2204.13	21.09	2555.82	85.72
GBELE	-8052.35	23.6	-7083.98	20.63	-945.85	8.19	-22.53	9.48
GBTOT	-6268.51	40.48	-5347.02	36.96	-866.92	8.69	-54.58	9.51
$T\Delta S$[b]	-2090.56	5.36	-1989.04	12.41	-139.96	0.21	-38.44	15.56
$\Delta G_{binding-PBSA}$							-22.81	
$\Delta G_{binding-GBSA}$							-16.14	

a Average over 200 snapshots from trajectory.

b Entropy calculations were based on normal mode analysis using only 20 snapshots.

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; PBSUR, hydrophobic contribution to solvation free energy for PB calculations; PBCAL, reaction field energy calculated by PB; PBSOL=PBSUR+PBCAL; PBELE=PBCAL+ELE; PBTOTAL=PBSOL+GAS; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GBCAL+ELE; GBTOTAL=GBSOL+GAS; TΔS, T(temperature)* ΔS(sum of rotational, translational and vibrational entropies); $\Delta G_{binding}$, total binding energy of the system.

Table 5.6. Energy MM/PBSA energy component analysis of the interactions of the heparin pentasaccharide with Ig-domains 2-3 averaged during 6-8 ns[a].

	Complex		**Receptor**		**Ligand**		**Δ**	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
ELE	-4590.15	78.65	-3186.62	90.85	1316.31	29.44	-2719.84	77.44
VDW	-460.77	21.12	-430.33	19.73	1.6	5.01	-32.05	4.54
INT	2128.75	39.04	2073.07	36.53	55.68	10.27	0	0
GAS	-2922.17	85.1	-1543.88	98.71	1373.59	27.8	-2751.88	78
PBSUR	76.78	1.08	74.69	1.09	7.96	0.11	-5.86	0.32
PBCAL	-3464.16	70.5	-3843.49	83.17	-2320.83	25.43	2700.16	75.62
PBSOL	-3387.38	69.91	-3768.81	82.55	-2312.88	25.52	2694.3	75.43
PBELE	-8054.31	30.23	-7030.11	27.16	-1004.52	8.17	-19.68	8.43
PBTOT	-6309.56	42.58	-5312.69	40.48	-939.28	8.22	-57.59	7.47
GBSUR	100.77	1.43	97.99	1.44	9.35	0.15	-6.57	0.42
GB	-3449.63	72.46	-3899.08	83.66	-2256.93	25.89	2706.37	72.39
GBSOL	-3348.86	71.71	-3801.08	82.87	-2247.58	26.01	2699.81	72.14
GBELE	-8039.78	24.25	-7085.7	21.93	-940.62	8.56	-13.46	9.37
GBTOT	-6271.03	39.89	-5344.96	39.25	-873.99	9.15	-52.08	9.02
TΔS[b]	-2096.65	9.33	-1993.95	7.64	-140.5	0.61	-37.8	11.8
$\Delta G_{\text{binding-PBSA}}$							-19.79	
$\Delta G_{\text{binding-GBSA}}$							-14.28	

a Average over 200 snapshots from trajectory.

b Entropy calculations were based on normal mode analysis using only 20 snapshots.

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; PBSUR, hydrophobic contribution to solvation free energy for PB calculations; PBCAL, reaction field energy calculated by PB; PBSOL=PBSUR+PBCAL; PBELE=PBCAL+ELE; PBTOTAL=PBSOL+GAS; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GBCAL+ELE; GBTOTAL=GBSOL+GAS; TΔS, T (temperature)* ΔS (sum of rotational, translational and vibrational entropies); $\Delta G_{\text{binding}}$, total binding energy of the system.

Table 5.7. Energy MM/PBSA energy component analysis of the interactions of the heparin disaccharide with Ig-domains 5-6 averaged during 0-8 ns[a].

	Complex		**Receptor**		**Ligand**		**Δ**	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
ELE	-4827.81	111.08	-4711.05	112.35	64.67	8.76	-181.43	31.25
VDW	-600.46	20.82	-587.52	20.34	8.08	3.54	-21.03	3.9
INT	2323.65	44.32	2299.91	43.92	23.74	6.07	0	0
GAS	-3104.62	118.29	-2998.65	121.27	96.49	9.11	-202.46	31.56
PBSUR	79.15	1.07	78.5	1.09	4.58	0.04	-3.93	0.26
PBCAL	-3726.98	93.86	-3373.83	98.8	-548.34	6.7	195.19	30.28
PBSOL	-3647.83	93.41	-3295.33	98.25	-543.76	6.72	191.25	30.14
PBELE	-8554.79	35.68	-8084.88	30.73	-483.67	4.79	13.76	7.67
PBTOT	-6752.45	53.64	-6293.98	51.99	-447.27	5.36	-11.2	5.96
GBSUR	103.92	1.43	103.06	1.45	4.86	0.06	-4	0.35
GB	-3741.22	99.36	-3417.39	101.68	-524.7	6.56	200.86	29.6
GBSOL	-3637.3	98.82	-3314.32	100.97	-519.83	6.59	196.86	29.45
GBELE	-8569.03	26.42	-8128.44	24.99	-460.03	4.49	19.43	4.45
GBTOT	-6741.91	48.33	-6312.98	49.01	-423.34	5.36	-5.6	4.47
TΔS[b]	-2302.42	11.52	-2250.36	10.2	-67.46	0.53	-15.41	17.74
$\Delta G_{binding\text{-}PBSA}$							+4.21	
$\Delta G_{binding\text{-}GBSA}$							+9.81	

a Average over 800 snapshots from trajectory.

b Entropy calculations were based on normal mode analysis using only 40 snapshots.

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; PBSUR, hydrophobic contribution to solvation free energy for PB calculations; PBCAL, reaction field energy calculated by PB; PBSOL=PBSUR+PBCAL; PBELE=PBCAL+ELE; PBTOTAL=PBSOL+GAS; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GBCAL+ELE; GBTOTAL=GBSOL+GAS; TΔS, T(temperature)* ΔS (sum of rotational, translational and vibrational entropies); $\Delta G_{binding}$, total binding energy of the system.

Chapter 6

CONCLUSIONS AND SCOPE FOR FUTURE WORK

SUMMARY OF MAIN FINDINGS

A homology model of human PECAM-1 was successfully constructed using homology modelling and threading techniques. Similarity searches against structural databases were used to locate putative sulfate binding motifs in the Ig-domains of PECAM-1. This homology model was used in combination with docking simulations of representative heparin fragments to predict their interactions with PECAM-1.

The docking simulations have predicted that the heparin fragments may have direct electrostatic interactions with positively charged residues located in loops in Ig-domains 2 and 3, providing the basis for the existence of a high affinity GAG binding region in PECAM-1 at slightly acidic pH. The docking simulations indicate that the GAG binding region involves Ig-domain 3, residues His 239, Lys 255 and Gln 259, with further contributions from Ig-domain 2 residue Arg 179. An additional low affinity heparin binding region appears to be located in Ig-domain 5, with contributions from Ig-domain 6. Importantly, these two putative GAG binding regions are distinct from regions involved in homophilic and heterophilic interactions in Ig-domains 1 and 6, as well as the cation binding sites in Ig-domains 5 and 6. These findings suggest that PECAM-1 may be capable of mediating heterophilic aggregation through interactions with specific GAGs on adjacent cells, as in the case of NCAM. Furthermore, the binding of homophilic and heterophilic ligands like heparin/HS to PECAM-1 may occur simultaneously.

The docking simulations also suggest that PECAM-1 cannot bind HA and CS, but may bind DS with low affinity due to differences in sulfation and the glycosidic linkages present in these oligosaccharides. This is consistent with experimental data. Docking of smaller heparin fragments showed that the binding affinity of heparin to Ig-domains 2 and 3 increased with increasing length of the heparin fragment. For a *closed* conformation of Ig-domains 2 and 3, the simulations suggest that a heparin pentasaccharide is the optimal fragment for binding to Ig-domain 2 and 3. Docking of various oligosaccharides with diverse conformations suggest that the preferred conformation of iduronic acids is the $^1C_4$ conformation for binding to Ig-domains 2 and 3.

MD simulations using the MM/PBSA and MM/GBSA methods have been used here for the first time to investigate the binding of heparin fragments to the receptor PECAM-1. There are differences in the affinity prediction from both methods, i.e. docking and MM/PBSA. The binding affinties predicted using MM/PBSA are considerable higher due to the lack of proper forcefield parameterisation for GAGs and the combination of different forcefields for geometry generation and energy calculations. The high binding affinities using MM/PBSA were also observed for galectin-1-oligosaccharide complexes (Ford *et al.* 2003). The magnitude of such affinities using MM/PBSA is the result of changes in the electrostatic interactions due to the flexibility induced in the Ig-domains of PECAM-1 and the heparin fragments, which was not taken into account whilst performing docking studies. The MM/PBSA calculations have been shown to be in good agreement with the docking simulations, confirming the prediction of the existence of high and low affinity GAG-binding regions in the receptor, as has been recently found experimentally (Coombe *et al.* 2008).

Binding of heparin fragments to the Ig-domains of PECAM-1 appears to be dominated by favourable VDW and electrostatic interactions, as expected from the polyanionic nature of heparin and the cationic nature of the binding site of the receptor. The vibrational entropy contribution has similar magnitude to the VDW and electrostatic interactions and hence it also plays a critical role in determining the free energy of binding of the oligosaccharides to the Ig-domains. Calculations of the solvation free energies using the generalised Born model as well as the Poisson-Boltzmann approach resulted in similar predictions of the free energy of binding of heparin fragments to PECAM-1.

The MD simulations also revealed the existence of a hinge-type conformational change affecting Ig-domains 2 and 3. This conformational change in the receptor exposes more basic residues on the surface, which may facilitate the binding of longer-sized heparin fragments, such as an octasaccharide. This conformational change is responsible for fluctuations in the free energy of binding of heparin fragments. The MD simulations revealed the presence of conformational changes from a predominantly beta-sheet structure in Ig-domains 2/3 and 5/6 to a disordered random coil due to the presence of large numbers of glycine residues (which impart conformational flexibility) and proline (which have structure breaking properties).

In conclusion, the analysis of the predicted interactions of GAGs with PECAM-1 provide further understanding of the interaction forces, the specific sulfation patterns and the conformational preferences of GAGs involved in determining the specificity and selectivity of GAG binding.

SCOPE FOR FUTURE WORK

The purpose of docking GAG fragments to the surface of a protein such as a cell adhesion molecule is to identify the likely position of the heparin-binding site(s), predict the binding mode of GAG fragments and obtain a rough estimate of the free energy of binding (and dissociation constant). Most docking methods perform coarse docking. As a consequence, two model oligosaccharide fragments were needed for improved accuracy, one in which all IdoA2S residues are in the $^1C_4$ ring form and another one in which the $^2S_0$ ring form is adopted. Nowadays, there are plenty of docking programs which would allow flexibility of both, the glycosidic torsions as well as the exocyclic torsion angles.

Current methods aimed at predicting high affinity GAG sequences fail to take into account any conformational changes that may occur in the protein receptor. In addition, many docking programs impose limits on the number of rotatable bonds that can be modelled (AutoDock, for example, can handle up to 32 rotatable bonds in the ligand), resulting in oligosaccharides longer than a pentasaccharide being treated as semi-rigid molecules. The accurate computational prediction of the affinity of binding for GAG-protein complexes is still in its infancy, particularly because of the poorly defined contribution of water (solvation/desolvation) to the binding interaction and limitations in the force field and scoring functions used to represent GAG structure, dynamics and interactions. These limitations can be overcomed by using advanced methods for docking and scoring such as the one implemented in Glide and AutoDock

4.0. These programs can take into account partial flexibility of the protein and allow full treatment of ligand flexibility. Statistical analysis of protein ensembles obtained from molecular dynamics simulations can also help in studying the fluctutaions corresponding to different conformations and the correlations of the amino acids in the transition state.

There is much further work to be done to build on the predictions reported in this thesis. Of particular interest is the use of structure-based drug design methods to provide a rationale for the development of new selective, potent drug-like GAG-mimetic molecules for the treatment of inflammatory diseases. The combination of these approaches with experimental information (structural and binding activity data) may assist the identification of chemical modifications that should be performed on known GAG fragments to optimise their binding. Similarly, these approaches may be used to predict the structure of new small molecules that can mimic the sulfation patterns required for binding specificity and selectivity to PECAM-1.

Recent progress in the understanding of GAG biosynthesis, structure and function create the opportunity to capitalise on the large structural diversity of GAGs in drug discovery programs. Heparin/HS GAGs are an important subset of complex polysaccharides that can be exploited to treat inflammatory diseases, thrombosis, virus infections and cancer. Approaches like combinatorial virtual screening (Raghuraman *et al.* 2006) and focused libraries (introduction of non-anionic structural motifs into heparin/HS) (Fernandez *et al.* 2006; Huang & Kerns 2006) can further provide a rationale for the development of novel charge-reduced, potent drug-like GAG mimetic molecules.

Molecular dynamics methods such as computational alanine mutagenesis can aid in elucidating the nature of the molecular interactions between GAGs and PECAM-1. A systematic scanning mutagenesis of the GAG binding sites in PECAM-1 could be carried out to measure the contributions of specific amino acid residues of Ig-domains 2 and 3 to the specificity and affinity of GAG interactions using MM/PBSA methods or free energy perturbation methods. These methods are capable of providing predictions that can be tested experimentally and hence have an important role to play in early drug discovery programs.

REFERENCES

Albelda, SM 1991, 'Molecular and cellular properties of PECAM-1 (endoCAM/CD31): a novel vascular cell-cell adhesion molecule', *The Journal of Cell Biology,* vol. 114, no. 5, pp. 1059-68.

Allinger, NL 1977, 'Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms', *Journal of the American Chemical Society,* vol. 99, no. 25, pp. 8127-34.

Allinger, NL, Yuh, YH & Lii, JH 1989, 'Molecular mechanics. The MM3 force field for hydrocarbons. 1', *Journal of the American Chemical Society,* vol. 111, no. 23, pp. 8551-66.

Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W & Lipman, DJ 1997, 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Research,* vol. 25, no. 17, pp. 3389-402.

Anfinsen, CB 1972, 'The formation and stabilization of protein structure', *The Biochemical Journal,* vol. 128, no. 4, pp. 737-49.

Arteel, GE, Franken, S, Kappler, J & Sies, H 2000, 'Binding of Selenoprotein P to heparin: characterization with surface plasmon resonance', *Biological Chemistry,* vol. 381, no. 3, pp. 265-8.

Babor, M, Greenblatt, HM, Edelman, M & Sobolev, V 2005, 'Flexibility of metal binding sites in proteins on a database scale', *Proteins: Structure, Function, and Bioinformatics,* vol. 59, no. 2, pp. 221-30.

Bae, J, Desai, UR, Pervin, A, Caldwell, EE, Weiler, JM & Linhardt, RJ 1994, 'Interaction of heparin with synthetic antithrombin III peptide analogues', *The Biochemical Journal,* vol. 301, no. Pt 1, pp. 121-9.

Barclay, AN 2003, 'Membrane proteins with immunoglobulin-like domains--a master superfamily of interaction molecules', *Seminars in Immunology,* vol. 15, no. 4, pp. 215-23.

Bashford, D & Case, DA 2000, 'Generalized Born models of macromolecular solvation effects', *Annual Review of Physical Chemistry,* vol. 51, no. 1, pp. 129-52.

Basma, M, Sundara, S, Calgan, D, Vernali, T & Woods, RJ 2001, 'Solvated ensemble averaging in the calculation of partial atomic charges', *Journal of Computational Chemistry,* vol. 22, no. 11, pp. 1125-37.

Bayly, CI, Cieplak, P, Cornell, W & Kollman, PA 1993, 'A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model', *The Journal of Physical Chemistry,* vol. 97, no. 40, pp. 10269-80.

Becker, OM, Marantz, Y, Shacham, S, Inbal, B, Heifetz, A, Kalid, O, Bar-Haim, S, Warshaviak, D, Fichman, M & Noiman, S 2004, 'G protein-coupled receptors: In silico drug discovery in 3D', *Proceedings of the National Academy of Sciences,* vol. 101, no. 31, pp. 11304-9.

Bella, J & Berman, HM 2000, 'Integrin-collagen complex: a metal-glutamate handshake', *Structure,* vol. 8, no. 6, pp. R121-6.

Belting, M, Borsig, L, Fuster, MM, Brown, JR, Persson, L, Fransson, LA & Esko, JD 2002, 'Tumor attenuation by combined heparan sulfate and polyamine depletion', *Proceedings of the National Academy of Sciences,* vol. 99, no. 1, p. 371.

Bennett-Lovsey, RM, Herbert, AD, Sternberg, MJE & Kelley, LA 2007, 'Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre', *Proteins: Structure, Function, and Bioinformatics,* vol. 70, no. 3, pp. 611 - 25.

Bernfield, M, Gotte, M, Park, PW, Reizes, O, Fitzgerald, ML, Lincecum, J & Zako, M 1999, 'Functions of cell surface heparan sulfate proteoglycans', *Annual Review of Biochemistry,* vol. 68, no. 1, pp. 729-77.

Bitomsky, W & Wade, RC 1999, 'Docking of glycosaminoglycans to heparin-binding proteins: validation for aFGF, bFGF, and antithrombin and application to IL-8', *Journal of American Chemical Society,* vol. 121, no. 13, pp. 3004-13.

Blansché, A, Gançarski, P & Korczak, J 2005, 'Genetic Algorithms for Feature Weighting: Evolution vs. Coevolution and Darwin vs. Lamarck', in *MICAI 2005: Advances in Artificial Intelligence*, pp. 682-91.

Bock, CW, Katz, AK, Markham, GD & Glusker, JP 1999, 'Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions', *Journal of American Chemical Society,* vol. 121, no. 32, pp. 7360-72.

Boeckmann, B, Bairoch, A, Apweiler, R, Blatter, MC, Estreicher, A, Gasteiger, E, Martin, MJ, Michoud, K, O'Donovan, C, Phan, I, Pilbout, S & Schneider, M 2003, 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003', *Nucleic Acids Research,* vol. 31, no. 1, pp. 365-70.

Boehm, MK, Corper, AL, Wan, T, Sohi, MK, Sutton, BJ, Thornton, JD, Keep, PA, Chester, KA, Begent, RH & Perkins, SJ 2000, 'Crystal structure of the anti-(carcinoembryonic antigen) single-chain Fv antibody MFE-23 and a model for antigen binding based on intermolecular contacts', *The Biochemical Journal,* vol. 346, no. Pt 2, pp. 519-28.

Bork, P, Holm, L & Sander, C 1994, 'The immunoglobulin fold. Structural classification, sequence patterns and common core', *Journal of Molecular Biology,* vol. 242, no. 4, pp. 309-20.

Borza, DB & Morgan, WT 1998, 'Histidine-Proline-rich Glycoprotein as a Plasma pH Sensor. Modulation of its interaction with glycosaminoglycans by pH and metals', *Journal of Biological Chemistry,* vol. 273, no. 10, pp. 5493-9.

Bowie, JU, Luthy, R & Eisenberg, D 1991, 'A method to identify protein sequences that fold into a known three-dimensional structure', *Science,* vol. 253, no. 5016, pp. 164-70.

Brooks, BR, Bruccoleri, RE, Olafson, BD, States, DJ, Swaminathan, S & Karplus, M 1983, 'CHARMM: a program for macromolecular energy, minimization, and dynamics calculations', *Journal of computational chemistry,* vol. 4, no. 2, pp. 187-217.

Brugge, J, Wong, CWY, Wiedle, G, Ballestrem, C, Wehrle-Haller, B, Etteldorf, S, Bruckner, M, Engelhardt, B, Gisler, RH & Imhof, BA 2000, 'PECAM-1/CD31 trans-homophilic binding at the intercellular junctions is independent of its cytoplasmic domain; evidence for heterophilic interaction with Integrin avß3 in cis', *Molecular Biology of the Cell,* vol. 11, no. 9, pp. 3109-21.

Bryant, SH & Lawrence, CE 1993, 'An empirical energy function for threading protein sequence through the folding motif', *Proteins,* vol. 16, no. 1, pp. 92-112.

Bryson, K, McGuffin, LJ, Marsden, RL, Ward, JJ, Sodhi, JS & Jones, DT 2005, 'Protein structure prediction servers at University College London', *Nucleic Acids Research,* vol. 33, no. Web server, pp. W36-8.

Buckley, CD 1996, 'Identification of alpha v beta 3 as a heterotypic ligand for CD31/PECAM-1', *Journal of Cell Science,* vol. 109, no. 2, pp. 437-45.

Burkert, U & Allinger, NL 1982, *Molecular mechanics*, American Chemical Society, Washington, D.C.

Calaycay, J, Pande, H, Lee, T, Borsi, L, Siri, A, Shively, JE & Zardi, L 1985, 'Primary structure of a DNA-and heparin-binding domain (Domain III) in human plasma fibronectin', *Journal of Biological Chemistry,* vol. 260, no. 22, pp. 12136-41.

Caldwell, EE, Nadkarni, VD, Fromm, JR, Linhardt, RJ & Weiler, JM 1996, 'Importance of specific amino acids in protein binding sites for heparin and heparan sulfate', *The International Journal of Biochemistry & Cell Biology,* vol. 28, no. 2, pp. 203-16.

Canales, A, Angulo, J, Ojeda, R, Bruix, M, Fayos, R, Lozano, R, Giménez-Gallego, G, Martín-Lomas, M, Nieto, PM & Jiménez-Barbero, J 2005, 'Conformational flexibility of a synthetic glycosylaminoglycan bound to a fibroblast growth factor. FGF-1 recognizes both the (1) C (4) and (2) S (O) conformations of a bioactive heparin-like hexasaccharide', *Journal of American Chemical Society,* vol. 127, no. 16, pp. 5778-9.

Canales, A, Lozano, R, Lopez-Mendez, B, Angulo, J, Ojeda, R, Nieto, PM, Martin-Lomas, M, Gimenez-Gallego, G & Jimenez-Barbero, J 2006, 'Solution NMR structure of a human FGF-1 monomer, activated by a hexasaccharide heparin-analogue', *FEBS Journal,* vol. 273, no. 20, p. 4716.

Cao, G, O'Brien, CD, Zhou, Z, Sanders, SM, Greenbaum, JN, Makrigiannakis, A & DeLisser, HM 2002, 'Involvement of human PECAM-1 in angiogenesis and in vitro endothelial cell migration', *American Journal of Physiology- Cell Physiology,* vol. 282, no. 5, pp. 1181-90.

Capila, I, Hernáiz, MJ, Mo, YD, Mealy, TR, Campos, B, Dedman, JR, Linhardt, RJ & Seaton, BA 2001, 'Research article Annexin V–heparin oligosaccharide complex suggests heparan sulfate–mediated assembly on cell surfaces', *Structure,* vol. 9, pp. 57-64.

Capila, I & Linhardt, RJ 2002, 'Heparin–protein interactions', *Angewandte Chemie International Edition,* vol. 41, no. 3, pp. 390-412.

Cardin, AD & Weintraub, HJ 1989, 'Molecular modeling of protein-glycosaminoglycan interactions', *Arteriosclerosis, Thrombosis, and Vascular Biology,* vol. 9, no. 1, pp. 21-32.

Carter, WJ, Cama, E & Huntington, JA 2005, 'Crystal Structure of Thrombin Bound to Heparin', *Journal of Biological Chemistry,* vol. 280, no. 4, pp. 2745-9.

Casasnovas, JM, Stehle, T, Liu, JH, Wang, JH & Springer, TA 1998, 'A dimeric crystal structure for the N-terminal two domains of intercellular adhesion molecule-1', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 95, no. 8, pp. 4134-9.

Case, DA, Cheatham Iii, TE, Darden, T, Gohlke, H, Luo, R, Merz Jr, KM, Onufriev, A, Simmerling, C, Wang, B & Woods, RJ 2005, 'The Amber biomolecular simulation programs', *Journal of Computational Chemistry,* vol. 26, no. 16, pp. 1668-88.

Casu, B, Guerrini, M & Torri, G 2004, 'Structural and conformational aspects of the anticoagulant and anti-thrombotic activity of heparin and dermatan sulfate', *Current Pharmaceutical Design,* vol. 10, no. 9, pp. 939-49.

Champe, PC & Harvey, RA 2005, *Biochemistry*, Lippincott Williams & Wilkins.

Chenna, R, Sugawara, H, Koike, T, Lopez, R, Gibson, TJ, Higgins, DG & Thompson, JD 2003, 'Multiple sequence alignment with the Clustal series of programs', *Nucleic Acids Research,* vol. 31, no. 13, pp. 3497-500.

Chiba, R, Nakagawa, N, Kurasawa, K, Tanaka, Y, Saito, Y & Iwamoto, I 1999, 'Ligation of CD31 (PECAM-1) on Endothelial Cells Increases Adhesive Function of alpha vbeta 3 Integrin and Enhances beta 1 Integrin-Mediated Adhesion of Eosinophils to Endothelial Cells', *Blood,* vol. 94, no. 4, p. 1319.

Choay, J, Petitou, M, Lormeau, JC, Sinay, P, Casu, B & Gatti, G 1983, 'Structure-activity relationship in heparin: a synthetic pentasaccharide with high affinity for antithrombin III and eliciting high anti-factor Xa activity', *Biochemical and Biophysical Research Communications,* vol. 116, no. 2, pp. 492-9.

Chong, LT, Duan, Y, Wang, L, Massova, I & Kollman, PA 1999, 'Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 96, no. 25, pp. 14330-5.

Chothia, C & Jones, EY 1997, 'The molecular structure of cell adhesion molecules', *Annual Review of Biochemistry,* vol. 66, no. 1, pp. 823-62.

Codée, JDC, Overkleeft, HS, van der Marel, GA & van Boeckel, CAA 2004, 'The synthesis of well-defined heparin and heparan sulfate fragments', *Drug Discovery Today: Technologies,* vol. 1, pp. 317–26.

Connolly, M 1983, 'Analytical molecular surface calculation', *Journal of Applied Crystallography,* vol. 16, no. 5, pp. 548-58.

Coombe, DR & Kett, WC 2005, 'Heparan sulfate-protein interactions: therapeutic potential through structure-function insights', *Cellular and Molecular Life Sciences (CMLS),* vol. 62, no. 4, pp. 410-24.

Coombe, DR, Stevenson, SS, Kinnear, BF, Gandhi, NS, Mancera, RL, Osmond, RIW & Kett, WC 2008, 'Platelet Endothelial Cell Adhesion Molecule (PECAM-1) and its interactions with glycosaminoglycans II : Biochemical analyses', *Biochemistry,* vol. in press.

Cornell, WD, Cieplak, P, Bayly, CI, Gould, IR, Merz, KM, Ferguson, DM, Spellmeyer, DC, Fox, T, Caldwell, JW & Kollman, PA 1995, 'A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules', *Journal of the American Chemical Society,* vol. 117, no. 19, pp. 5179-97.

Cornell, WD, Cieplak, P, Bayly, CI & Kollmann, PA 1993, 'Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation', *Journal of the American Chemical Society,* vol. 115, no. 21, pp. 9620-31.

Deaglio, S, Morra, M, Mallone, R, Ausiello, CM, Prager, E, Garbarino, G, Dianzani, U, Stockinger, H & Malavasi, F 1998, 'Human CD38 (ADP-Ribosyl Cyclase) is a counter-receptor of CD31, an Ig Superfamily Member 1', *The Journal of Immunology,* vol. 160, no. 1, pp. 395-402.

Deane, CM & Blundell, TL 2001, 'CODA: A combined algorithm for predicting the structurally variable regions of protein models', *Protein Science,* vol. 10, no. 3, pp. 599-612.

DeLisser, HM, Christofidou-Solomidou, M, Strieter, RM, Burdick, MD, Robinson, CS, Wexler, RS, Kerr, JS, Garlanda, C, Merwin, JR & Madri, JA 1997, 'Involvement of endothelial PECAM-1/CD31 in angiogenesis', *American Journal of Pathology,* vol. 151, no. 3, pp. 671-7.

DeLisser, HM, Yan, HC, Newman, PJ, Muller, WA, Buck, CA & Albelda, SM 1993, 'Platelet/endothelial cell adhesion molecule-1 (CD31)-mediated cellular aggregation involves cell surface glycosaminoglycans', *Journal of Biological Chemistry,* vol. 268, no. 21, pp. 16037-46.

Desai, UR, Wang, HM, Kelly, TR & Linhardt, RJ 1993, 'Structure elucidation of a novel acidic tetrasaccharide and hexasaccharide derived from a chemically modified heparin', *Carbohydrate Research,* vol. 241, pp. 249–59.

DesJarlais, RL, Sheridan, RP, Seibel, GL, Dixon, JS, Kuntz, ID & Venkataraghavan, R 1988, 'Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure', *Journal of Medicinal Chemistry,* vol. 31, no. 4, pp. 722-9.

DiGabriele, AD, Lax, I, Chen, DI, Svahn, CM, Jaye, M, Schlessinger, J & Hendrickson, WA 1998, 'Structure of a heparin-linked biologically active dimer of fibroblast growth factor', *Nature,* vol. 393, no. 6687, pp. 812-7.

Douguet, D & Labesse, G 2001a, 'Easier threading through web-based comparisons and cross-validations', *Bioinformatics,* vol. 17, no. 8, pp. 752-3.

Douguet, D & Labesse, G 2001b, 'Easier threading through web-based comparisons and cross-validations', *Bioinformatics,* vol. 17, no. 8, pp. 752-3.

Duhovny, D, Nussinov, R & Wolfson, HJ 2002, 'Efficient unbound docking of rigid molecules', *Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pp. 185-200.

Dunbrack Jr, RL & Karplus, M 1993, 'Backbone-dependent rotamer library for proteins. Application to side-chain prediction', *Journal of Molecular Biology,* vol. 230, no. 2, pp. 543-74.

Eisenberg, D, Luthy, R & Bowie, JU 1997, 'VERIFY3D: assessment of protein models with three-dimensional profiles', *Methods in Enzymology,* vol. 277, pp. 396-404.

Esko, JD & Lindahl, U 2001, 'Molecular diversity of heparan sulfate', *The Journal of Clinical Investigation,* vol. 108, no. 2, pp. 169-73.

Fabricius, J, Engelsen, SB & Rasmussen, K 1997, 'The Consistent Force Field. 5. PEF95SAC: Optimized potential energy function for alcohols and carbohydrates', *Journal of Carbohydrate Chemistry,* vol. 16, no. 6, pp. 751-72.

Faham, S, Hileman, RE, Fromm, JR, Linhardt, RJ & Rees, DC 1996, 'Heparin structure and interactions with Basic Fibroblast Growth Factor', *Science,* vol. 271, no. 5252, p. 1116.

Fareed, J, Hoppensteadt, DA & Bick, RL 2000, 'An update on heparins at the beginning of the new millennium', *Seminars in Thrombosis and Hemostasis,* vol. 26, no. number s 1, pp. 5-18.

Fernandez-Recio, J, Totrov, M & Abagyan, R 2002, 'Soft protein-protein docking in internal coordinates', *Protein Science,* vol. 11, no. 2, pp. 280-91.

Fernandez, C, Hattan, CM & Kerns, RJ 2006, 'Semi-synthetic heparin derivatives: chemical modifications of heparin beyond chain length, sulfate substitution pattern and N-sulfo/N-acetyl groups', *Carbohydrate Research,* vol. 341, no. 10, pp. 1253-65.

Ferro, DR, Provasoli, A, Ragazzi, M, Casu, B, Torri, G, Bossennec, V, Perly, B, Sinay, P, Petitou, M & Choay, J 1990, 'Conformer populations of L-iduronic acid residues in glycosaminoglycan sequences', *Carbohydrate Research,* vol. 195, no. 2, pp. 157-67.

Ferro, DR, Provasoli, A, Ragazzi, M, Torri, G, Casu, B, Gatti, G, Jacquinet, JC, Sinay, P, Petitou, M & Choay, J 1986, 'Evidence for conformational equilibrium of the sulfated L-iduronate residue in heparin and in synthetic heparin mono-and oligo-saccharides: NMR and force-field studies', *Journal of the American Chemical Society,* vol. 108, no. 21, pp. 6773-8.

Ferro, DR, Pumilia, P, Cassinari, A & Ragazzi, M 1995, 'Treatment of ionic species in force-field calculations: sulfate and carboxylate groups in carbohydrates', *International Journal of Biological Macromolecules,* vol. 17, no. 3-4, pp. 131-36.

Ferro, DR, Pumilia, P & Ragazzi, M 1997, 'An improved force field for conformational analysis of sulfated polysaccharides', *Journal of Computational Chemistry,* vol. 18, no. 3, pp. 351-67.

Ferro, V & Don, R 2003, 'The development of the novel angiogenesis inhibitor PI-88 as an anticancer drug', *Australasian Biotechnology,* vol. 13, pp. 38-9.

Fiser, A, Do, RKG & ŠAli, A 2000, 'Modeling of loops in protein structures', *Protein Science,* vol. 9, no. 09, pp. 1753-73.

Fleishman, SJ & Ben-Tal, N 2006, 'Progress in structure prediction of alpha-helical membrane proteins', *Current Opinion in Structural Biology,* vol. 16, pp. 496-504.

Fogolari, F, Brigo, A & Molinari, H 2003, 'Protocol for MM/PBSA Molecular Dynamics Simulations of Proteins', *Biophysical Journal,* vol. 85, no. 1, pp. 159-66.

Ford, MG, Weimar, T, Kohli, T & Woods, RJ 2003, 'Molecular dynamics simulations of galectin-1-oligosaccharide complexes reveal the molecular basis for ligand diversity', *Proteins: Structure, Function and Bioinformatics,* vol. 53, no. 2, pp. 229-40.

Forster, M & Mulloy, B 2006, 'Computational approaches to the identification of heparin-binding sites on the surfaces of proteins', *Biochemical Society Transactions,* vol. 34, no. Pt 3, pp. 431-4.

Fraser, PE, Nguyen, JT, Surewicz, WK & Kirschner, DA 1991, 'pH-dependent structural transitions of Alzheimer amyloid peptides', *Biophysical Journal,* vol. 60, no. 5, pp. 1190-201.

Frisch, MJ, Trucks, GW, Schlegel, HB, Scuseria, GE, Robb, MA, Cheeseman, JR, Zakrzewski, VG, Montgomery Jr, JA, Stratmann, RE & Burant, JC 1998, 'Gaussian 98, Revision A. 7', *Gaussian Inc., Pittsburgh, PA*.

Fromm, JR, Hileman, RE, Caldwell, EEO, Weiler, JM & Linhardt, RJ 1997, 'Pattern and spacing of basic amino acids in heparin binding sites', *Archives of Biochemistry and Biophysics,* vol. 343, no. 1, pp. 92-100.

Fry, EE, Lea, SM, Jackson, T, Newman, JWI, Ellard, FM, Blakemore, WE, Abu-Ghazaleh, R, Samuel, A, King, AMQ & Stuart, DI 1999, 'The structure and function of a foot-and-mouth disease virus- oligosaccharide receptor complex', *The EMBO Journal,* vol. 18, pp. 543-54.

Fugedi, P 2003, 'The potential of the molecular diversity of heparin and heparan sulfate for drug development', *Mini Reviews in Medicinal Chemistry,* vol. 3, no. 7, pp. 659–67.

Gilson, MK & Honig, BH 1987, 'Calculation of electrostatic potentials in an enzyme active site', *Nature,* vol. 330, no. 6143, pp. 84-6.

Gilson, MK & Zhou, HX 2007, 'Calculation of Protein-Ligand Binding Affinities', *Annual Reviews of Biophysics and Biomolecular Structure,* vol. 36, pp. 21-42.

Gonza'lez-Iglesias, R, Pajares, MA, Ocal, C, Oesch, B & Gasset, M 2002, 'Prion protein interaction with glycosaminoglycan occurs with the formation of oligomeric complexes stabilized by Cu (II) bridges', *Journal of Molecular Biology,* vol. 319, no. 2, pp. 527-40.

Goodford, PJ 1985, 'A computational procedure for determining energetically favorable binding sites on biologically important macromolecules', *Journal of Medicinal Chemistry,* vol. 28, no. 7, pp. 849-57.

Greer, J 1981, 'Comparative model-building of the mammalian serine proteases', *Journal of Molecular Biology,* vol. 153, no. 4, pp. 1027-42.

Grootenhuis, PDJ & Haasnoot, CAG 1993, 'A CHARMm based force field for carbohydrates using the CHEAT approach: carbohydrate hydroxyl groups represented by extended atoms', *Molecular Simulation,* vol. 10, no. 2, pp. 75-95.

Grootenhuis, PDJ & Van Boeckel, CAA 1991, 'Constructing a molecular model of the interaction between antithrombin III and a potent heparin analog', *Journal of the American Chemical Society,* vol. 113, no. 7, pp. 2743-7.

Guelph, Uo, *Clinical Pharmacology Review*, Ontario veterinary college's online course modules from http://www.ovc.uoguelph.ca/BioMed/Courses/Public/Pharmacology/pharmsite/98-409/Blood/hem_thromb.html

Hagler, AT & Ewig, CS 1994, 'On the use of quantum energy surfaces in the derivation of molecular force fields', *Computer Physics Communications,* vol. 84, no. 1-3, pp. 131-55.

Hallgren, J, Backstrom, S, Estrada, S, Thuveson, M & Pejler, G 2004, 'Histidines are critical for heparin-dependent activation of Mast Cell Tryptase 1', *The Journal of Immunology,* vol. 173, no. 3, pp. 1868-75.

Harding, MM 2006, 'Small revisions to predicted distances around metal sites in proteins', *Acta Crystallographica Section D Biological Crystallography,* vol. 62, no. Part 6, pp. 678-82.

Harpaz, Y & Chothia, C 1994, 'Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains', *Journal of Molecular Biology,* vol. 238, no. 4, pp. 528-39.

Herbert, JM, Herault, JP, Bernat, A, van Amsterdam, RGM, Lormeau, JC, Petitou, M, van Boeckel, C, Hoffmann, P & Meuleman, DG 1998, 'Biochemical and pharmacological properties of SANORG 34006, a potent and long-acting synthetic pentasaccharide', *Blood,* vol. 91, no. 11, p. 4197.

Hetenyi, C & van der Spoel, D 2002, 'Efficient docking of peptides to proteins without prior knowledge of the binding site', *Protein Science,* vol. 11, no. 7, pp. 1729-37.

Hileman, RE, Fromm, JR, Weiler, JM & Linhardt, RJ 1998a, 'Glycosaminoglycan-protein interactions: definition of consensus sites in glycosaminoglycan binding proteins', *BioEssays,* vol. 20, no. 2, pp. 156-67.

Hileman, RE, Jennings, RN & Linhardt, RJ 1998b, 'Thermodynamic analysis of the heparin interaction with a basic cyclic peptide using isothermal titration calorimetry', *Biochemistry,* vol. 37, no. 43, pp. 15231-7.

Holness, CL & Simmons, DL 1994, 'Structural motifs for recognition and adhesion in members of the immunoglobulin superfamily', *Journal of Cell Science,* vol. 107, no. Pt 8, pp. 2065-70.

Holt, CE & Dickson, BJ 2005, 'Sugar Codes for Axons?' *Neuron,* vol. 46, no. 2, pp. 169-72.

Hricovini, M, Guerrini, M & Bisio, A 1999, 'Structure of heparin-derived tetrasaccharide complexed to the plasma protein antithrombin derived from NOEs, J-couplings and chemical shifts', *European Journal of Biochemistry,* vol. 261, no. 3, pp. 789-801.

Hricovini, M, Guerrini, M, Bisio, A, Torri, G, Naggi, A & Casu, B 2002, 'Active conformations of glycosaminoglycans. NMR determination of the conformation of heparin sequences complexed with Antithrombin and Fibroblast Growth Factors in solution', *Seminars in Thrombosis and Hemostasis,* vol. 28, no. 4, pp. 325-34.

Huang, L & Kerns, RJ 2006, 'Diversity-oriented chemical modification of heparin: Identification of charge-reduced N-acyl heparin derivatives having increased selectivity for heparin-binding proteins', *Bioorganic Medicinal Chemistry,* vol. 14, no. 7, pp. 2300-13.

Huige, CJM & Altona, C 1995, 'Force field parameters for sulfates and sulfamates based on Ab Initio calculations: Extensions of AMBER and CHARMm fields', *Journal of Computational Chemistry,* vol. 16, no. 1, pp. 56-79.

Imberty, A, Bettler, E, Karababa, M, Mazeau, K, Petrova, P & Pérez, S 1999, 'Building sugars: The sweet part of structural biology', *Perspectives in Structural Biology. Indian Academy of Sciences and Universities Press, Hyderabad, India*, pp. 392–409.

Iozzo, RV, Cohen, IR, Grassel, S & Murdoch, AD 1994, 'The biology of perlecan: the multifaceted heparan sulphate proteoglycan of basement membranes and pericellular matrices', *The Biochemical Journal,* vol. 302, no. Pt 3, pp. 625-39.

Iozzo, RV & San Antonio, JD 2001, 'Heparan sulfate proteoglycans: heavy hitters in the angiogenesis arena', *The Journal of Clinical Investigation,* vol. 108, no. 3, pp. 349-55.

Jackson, DE 2003, 'The unfolding tale of PECAM-1', *FEBS Letters,* vol. 540, no. 1-3, pp. 7-14.

Jackson, DE 2006, 'Palmitoylation at Cys 595 is essential for PECAM-1 localisation into membrane microdomains and for efficient PECAM-1-mediated cytoprotection', *Thrombosis and Haemostasis,* vol. 96, pp. 756-66.

Jackson, DE, Loo, RO, Holyst, MT & Newman, PJ 1997, 'Identification and characterization of functional cation coordination sites in platelet endothelial cell adhesion molecule-1', *Biochemistry,* vol. 36, no. 31, pp. 9395-404.

Jackson, RL, Busch, SJ & Cardin, AD 1991, 'Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes', *Physiological Reviews,* vol. 71, no. 2, pp. 481-539.

Jairajpuri, MA, Lu, A, Desai, U, Olson, ST, Bjork, I & Bock, SC 2003, 'Antithrombin III phenylalanines 122 and 121 contribute to its high affinity for heparin and its conformational activation', *Journal of Biological Chemistry,* vol. 278, no. 18, pp. 15941-50.

Jayaram, B, Sprous, D & Beveridge, DL 1998, 'Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the AMBER Force Field', *Journal of Physical Chemistry B,* vol. 102, no. 47, pp. 9571-6.

Jin, L, Abrahams, JP, Skinner, R, Petitou, M, Pike, RN & Carrell, RW 1997, 'The anticoagulant activation of antithrombin by heparin', *Proceedings of the National Academy of Sciences of the United States of America,* vol. 94, no. 26, pp. 14683-8.

Johnson-Leger, C 2000, 'The parting of the endothelium: miracle, or simply a junctional affair?' *Journal of Cell Science,* vol. 113, no. 6, pp. 921-33.

Johnson, DJD, Li, W, Adams, TE & Huntington, JA 2006, 'Antithrombin–S195A factor Xa-heparin structure reveals the allosteric mechanism of antithrombin activation', *The EMBO Journal,* vol. 25, no. 9, pp. 2029-37.

Johnson, Z, Power, CA, Weiss, C, Rintelen, F, Ji, H, Ruckle, T, Camps, M, Wells, TN, Schwarz, MK & Proudfoot, AE 2004, 'Chemokine inhibition–why, when, where, which and how', *Biochemical Society Transactions,* vol. 32, no. Pt 2, pp. 366-77.

Johnson, Z, Proudfoot, AE & Handel, TM 2005, 'Interaction of chemokines and glycosaminoglycans: A new twist in the regulation of chemokine function with opportunities for therapeutic intervention', *Cytokine and Growth Factor Reviews,* vol. 16, no. 6, pp. 625-36.

Jones, DT 1999, 'Protein secondary structure prediction based on position-specific scoring matrices', *Journal of Molecular Biology,* vol. 292, no. 2, pp. 195-202.

Jones, DT, Taylort, WR & Thornton, JM 1992, 'A new approach to protein fold recognition', *Nature,* vol. 358, no. 6381, pp. 86-9.

Jones, EY, Harlos, K, Bottomley, MJ, Robinson, RC, Driscoll, PC, Edwards, RM, Clements, JM, Dudgeon, TJ & Stuart, DI 1995a, 'Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1. 8 A resolution', *Nature,* vol. 373, no. 6514, pp. 539-44.

Jones, G, Willett, P & Glen, RC 1995b, 'Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation', *Journal of Molecular Biology,* vol. 245, no. 1, pp. 43-53.

Jorgensen, WL, Chandrasekhar, J, Madura, JD, Impey, RW & Klein, ML 1983, 'Comparison of simple potential functions for simulating liquid water', *The Journal of Chemical Physics,* vol. 79, p. 926.

Kalinowska, A & Losy, J 2006, 'PECAM-1, a key player in neuroinflammation', *European Journal of Neurology,* vol. 13, no. 12, pp. 1284-90.

Kasper, C, Rasmussen, H, Kastrup, JS, Ikemizu, S, Jones, EY, Berezin, V, Bock, E & Larsen, IK 2000, 'Structural basis of cell- cell adhesion by NCAM', *Nature Structural Biology,* vol. 7, no. 5, pp. 389-93.

Kelley, LA, MacCallum, RM & Sternberg, MJ 2000, 'Enhanced genome annotation using structural profiles in the program 3D-PSSM', *Journal of Molecular Biology,* vol. 299, no. 2, pp. 501-22.

Kerzmann, A, Neumann, D & Kohlbacher, O 2006, 'SLICK-Scoring and energy functions for protein-carbohydrate interactions', *Journal of chemical information and modeling,* vol. 46, no. 4, pp. 1635-42.

Kirschner, KN, Yongye, AB, Tschampel, SM, GonzÃ¡lez-OuteiriÃ±o, J, Daniels, CR, Foley, BL & Woods, RJ 2007, 'GLYCAM06: A generalizable biomolecular force field for carbohydrates', *Journal of Computational Chemistry,* vol. 29, no. 4, pp. 622-55.

Kisilevsky, R, Ancsin, JB, Szarek, WA & Petanceska, S 2007, 'Heparan sulfate as a therapeutic target in amyloidogenesis: prospects and possible complications', *Amyloid,* vol. 14, no. 1, pp. 21-32.

Kokenyesi, R & Bernfield, M 1994, 'Core protein structure and sequence determine the site and presence of heparan sulfate and chondroitin sulfate on syndecan-1', *Journal of Biological Chemistry,* vol. 269, no. 16, pp. 12304-9.

Kollman, PA, Massova, I, Reyes, C, Kuhn, B, Huo, S, Chong, L, Lee, M, Lee, T, Duan, Y & Wang, W 2000, 'Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models', *Accounts of Chemical Research,* vol. 33, no. 12, pp. 889-97.

Kolodny, R, Koehl, P, Guibas, L & Levitt, M 2002, 'Small libraries of protein fragments model native protein structures accurately', *Journal of Molecular Biology,* vol. 323, no. 2, pp. 297-307.

Kony, D, Damm, W, Stoll, S & Gunsteren, WFV 2002, 'An improved OPLS-AA force field for carbohydrates', *Journal of Computational Chemistry,* vol. 23, no. 15, pp. 1416-29.

Koshland, DE 1995, 'The Key–Lock Theory and the Induced Fit Theory', *Angewandte Chemie International Edition in English,* vol. 33, no. 2324, pp. 2375-8.

Kroll, H, Sun, QH & Santoso, S 2000, 'Platelet endothelial cell adhesion molecule-1 (PECAM-1) is a target glycoprotein in drug-induced thrombocytopenia', *Blood,* vol. 96, no. 4, p. 1409.

Kuschert, GS, Coulin, F, Power, CA, Proudfoot, AE, Hubbard, RE, Hoogewerf, AJ & Wells, TN 1999, 'Glycosaminoglycans interact selectively with chemokines and modulate receptor binding and cellular responses', *Biochemistry,* vol. 38, no. 39, pp. 12959-68.

Kuttel, M, Brady, JW & Naidoo, KJ 2002, 'Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations', *Journal of Computational Chemistry,* vol. 23, no. 13, pp. 1236-43.

Laederach, A & Reilly, PJ 2003, 'Specific empirical free energy function for automated docking of carbohydrates to proteins', *Journal of Computational Chemistry,* vol. 24, no. 14, pp. 1748-57.

Laitinen, T, Rouvinen, J & Peräkylä, M 2003, 'MM-PBSA free energy analysis of endo-1, 4-xylanase II (XynII)–substrate complexes: binding of the reactive sugar in a skew boat and chair conformation', *Organic and Biomolecular Chemistry,* vol. 1, pp. 3535-40.

Lam, K, Rao, VS & Qasba, PK 1998, 'Molecular modeling studies on binding of bFGF to heparin and its receptor FGFR1', *Journal of Biomolecular Structure and Dynamics,* vol. 15, no. 6, pp. 1009-27.

Laskowski, RA, Chistyakov, VV & Thornton, JM 2005, 'PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids', *Nucleic Acids Research,* vol. 33, no. Database issue, pp. D266-8.

Laskowski, RA, MacArthur, MW, Moss, DS & Thornton, JM 1993, 'PROCHECK: a program to check the stereochemical quality of protein structures', *Journal of Applied Crystallography,* vol. 26, no. 2, pp. 283-91.

Leach, AR 2001, *Molecular Modelling: Principles and Applications*, Prentice Hall.

Lee, MR, Duan, Y & Kollman, PA 2000, 'Use of MM-PB/SA in estimating the free energies of proteins: Application to native, intermediates, and unfolded villin headpiece', *Proteins: Structure, Function and Genetics,* vol. 39, no. 4, pp. 309-16.

Lee, SC, Guan, HH, Wang, CH, Huang, WN, Tjong, SC, Chen, CJ & Wu, W 2005, 'Structural basis of citrate-dependent and heparan sulfate-mediated cell surface retention of cobra cardiotoxin A3', *Journal of Biological Chemistry,* vol. 280, no. 10, pp. 9567-77.

Legge, GB, Morris, GM, Sanner, MF, Takada, Y, Olson, AJ & Grynszpan, F 2002, 'Model of the alphaLbeta2 integrin I-domain/ICAM-1 DI interface suggests that subtle changes in loop orientation determine ligand specificity', *Proteins,* vol. 48, no. 2, pp. 151-60.

Leitinger, B & Hogg, N 2000, 'From crystal clear ligand binding to designer I domains', *Nature Structural Biology,* vol. 7, pp. 614-6.

Li, W, Johnson, DJ, Esmon, CT & Huntington, JA 2004, 'Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin', *Nature Structural Molecular Biology,* vol. 11, no. 9, pp. 857-62.

Lindahl, U 2007, 'Heparan sulfate-protein interactions–A concept for drug design?' *Thrombosis and Haemostasis,* vol. 98, no. 1, pp. 109-15.

Lindahl, U & Kjellen, L 1991, 'Heparin or heparan sulfate--what is the difference?' *Thrombosis and Haemostasis,* vol. 66, no. 1, pp. 44-8.

Lindahl, U, Kusche-Gullberg, M & Kjellen, L 1998, 'Regulated Diversity of Heparan Sulfate', *Journal of Biological Chemistry,* vol. 273, no. 39, pp. 24979-82.

Lins, RD & Hünenberger, PH 2005, 'A new GROMOS force field for hexopyranose-based carbohydrates', *Journal of Computational Chemistry,* vol. 26, no. 13, pp. 1400-12.

Lohse, DL & Linhardt, RJ 1992, 'Purification and characterization of heparin lyases from Flavobacterium heparinum', *Journal of Biological Chemistry,* vol. 267, no. 34, pp. 24347-55.

Lortat-Jacob, H, Grosdidier, A & Imberty, A 2002, 'Structural diversity of heparan sulfate binding domains in chemokines', *Proceedings of the National Academy of Sciences,* vol. 99, no. 3, pp. 1229-34.

Lortat-Jacob, H, Turnbull, JE & Grimaud, JA 1995, 'Molecular organization of the interferon gamma-binding domain in heparan sulphate', *The Biochemical Journal,* vol. 310, no. Pt 2, pp. 497-505.

Luthy, R, Bowie, JU & Eisenberg, D 1992, 'Assessment of protein models with three-dimensional profiles', *Nature,* vol. 356, no. 6364, pp. 83-5.

Lyon, M & Gallagher, JT 1998, 'Bio-specific sequences and domains in heparan sulphate and the regulation of cell growth and adhesion', *Matrix Biology,* vol. 17, no. 7, pp. 485-93.

Malik, A & Ahmad, S 2007, 'Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network', *BMC Structural Biology,* vol. 7, no. 1, p. 1.

Marchler-Bauer, A, Anderson, JB, Cherukuri, PF, DeWeese-Scott, C, Geer, LY, Gwadz, M, He, S, Hurwitz, DI, Jackson, JD & Ke, Z 2005, 'CDD: a Conserved Domain Database for protein classification', *Nucleic Acids Research,* vol. 33, pp. D192-D6.

Marchler-Bauer, A, Anderson, JB, Derbyshire, MK, DeWeese-Scott, C, Gonzales, NR, Gwadz, M, Hao, L, He, S, Hurwitz, DI & Jackson, JD 2007, 'CDD: a conserved domain database for interactive domain family analysis', *Nucleic Acids Research,* vol. 35, no. Database issue, p. D237.

Margalit, H, Fischer, N & Ben-Sasson, SA 1993, 'Comparative analysis of structurally defined heparin binding sequences reveals a distinct spatial distribution of basic residues', *Journal of Biological Chemistry,* vol. 268, no. 26, pp. 19228-31.

Marti-Renom, MA, Stuart, AC & Fiser, A 2000, 'Comparative protein structure modeling of genes and genomes', *Annual Review of Biophysics and Biomolecular Structure,* vol. 29, pp. 291-325.

Martin, ACR, Cheetham, JC & Rees, AR 1989, 'Modeling antibody hypervariable loops: A combined algorithm', *Proceedings of the National Academy of Sciences,* vol. 86, no. 23, pp. 9268-72.

Massova, I & Kollman, PA 1999, 'Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies', *Journal of American Chemical Society,* vol. 121, no. 36, pp. 8133-43.

Matsumoto, R 1995, 'Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells', *Journal of Biological Chemistry,* vol. 270, no. 33, pp. 19524-31.

Maurice Petitou & Boeckel, CAAv 2004, 'A Synthetic Antithrombin III Binding Pentasaccharide Is Now a Drug! What Comes Next?' *Angewandte Chemie International Edition,* vol. 43, no. 24, pp. 3118-33.

Maxwell, KF, Powell, MS, Hulett, MD, Barton, PA, McKenzie, IF, Garrett, TP & Hogarth, PM 1999, 'Crystal structure of the human leukocyte Fc receptor, Fc gammaRIIa', *Nature Structural Biology,* vol. 6, no. 5, pp. 437-42.

Mehler, EL & Solmajer, T 1991, 'Electrostatic effects in proteins: comparison of dielectric and charge models', *Protein Engineering* vol. 4, no. 8, pp. 903-10.

Mikhailov, D, Linhardt, RJ & Mayo, KH 1997, 'NMR solution conformation of heparin-derived hexasaccharide', *The Biochemical Journal,* vol. 328, no. Pt 1, pp. 51-61.

Mikhailov, D, Mayo, KH, Vlahov, IR, Toida, T, Pervin, A & Linhardt, RJ 1996, 'NMR solution conformation of heparin-derived tetrasaccharide', *The Biochemical Journal,* vol. 318, no. 1, pp. 93-102.

Mitchell, JBO, Laskowski, RA, Alex, A & Thornton, JM 1999, 'BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential', *Journal of Computational Chemistry,* vol. 20, no. 11, pp. 1165-76.

Miyazawa, S & Jernigan, RL 1985, 'Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation', *Macromolecules,* vol. 18, no. 3, pp. 534-52.

Morris, GM, Goodsell, DS, Halliday, RS, Huey, R, Hart, WE, Belew, RK & Olson, AJ 1998, 'Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function', *Journal of Computational Chemistry,* vol. 19, no. 14, pp. 1639-62.

Morris, GM, Goodsell, DS, Huey, R & Olson, AJ 1996, 'Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4', *Journal of Computer-Aided Molecular Design,* vol. 10, no. 4, pp. 293-304.

Moseley, GW & Jackson, DE 2004, 'The multiple functions of PECAM-1', *Australian Biochemist,* vol. 35, no. 2, pp. 9-12.

Moult, J & James, MNG 1986, 'An algorithm for determining the conformation of polypeptide segments in proteins by systematic search', *Proteins: Structure, Function and Genetics,* vol. 1, no. 2, pp. 146-63.

Mulloy, B, Crane, DT, Drake, AF & Davies, DB 1996, 'The interaction between heparin and polylysine: a circular dichroism and molecular modelling study', *Brazilian Journal of Medical and Biological Research,* vol. 29, no. 6, pp. 721-9.

Mulloy, B & Forster, MJ 2000, 'Conformation and dynamics of heparin and heparan sulfate', *Glycobiology,* vol. 10, no. 11, pp. 1147-56.

Mulloy, B, Forster, MJ, Jones, C & Davies, DB 1993, 'Nmr and molecular-modelling studies of the solution conformation of heparin', *The Biochemical Journal,* vol. 293, no. Pt 3, pp. 849-58.

Mulloy, B, Forster, MJ, Jones, C, Drake, AF, Johnson, EA & Davies, DB 1994, 'The effect of variation of substitution on the solution conformation of heparin: a spectroscopic and molecular modelling study', *Carbohydrate Research,* vol. 255, pp. 1-26.

Nader, HB, Chavante, SF, dos-Santos, EA, Oliveira, FW, de-Paiva, JF, Jerônimo, SMB, Medeiros, GF, de-Abreu, LRD, Leite, EL & de-Sousa-Filho, JF 1999, 'Heparan sulfates and heparins: similar compounds performing the same functions in vertebrates and invertebrates?' *Brazilian Journal of Medical and Biological Research,* vol. 32, pp. 529-38.

Nakada, MT, Amin, K, Christofidou-Solomidou, M, O'Brien, CD, Sun, J, Gurubhagavatula, I, Heavner, GA, Taylor, AH, Paddock, C & Sun, QH 2000, 'Antibodies against the first Ig-like domain of human Platelet Endothelial Cell

Adhesion Molecule-1 (PECAM-1) that inhibit PECAM-1-dependent homophilic adhesion block in vivo neutrophil recruitment 1', *The Journal of Immunology,* vol. 164, no. 1, pp. 452-62.

Newman, PJ 1994, 'The role of PECAM-1 in vascular cell biology', *Annals of the New York Academy of Sciences,* vol. 714, no. 1, pp. 165-74.

Newman, PJ 1997, 'The Biology of PECAM-1', *The Journal of Clinical Investigation,* vol. 99, no. 1, pp. 3-8.

Newman, PJ 1999, 'Switched at birth: a new family for PECAM-1', *The Journal of Clinical Investigation,* vol. 103, no. 1, pp. 5-9.

Newman, PJ, Berndt, MC, Gorski, J, White 2nd, GC, Lyman, S, Paddock, C & Muller, WA 1990, 'PECAM-1 (CD31) cloning and relation to adhesion molecules of the immunoglobulin gene superfamily', *Science,* vol. 247, no. 4947, pp. 1219-22.

Newton, JP, Buckley, CD, Jones, EY & Simmons, DL 1997, 'Residues on both faces of the first immunoglobulin fold contribute to homophilic binding sites of PECAM-1/CD31', *Journal of Biological Chemistry,* vol. 272, no. 33, pp. 20555-63.

Newton, JP, Hunter, AP, Simmons, DL, Buckley, CD & Harvey, DJ 1999, 'CD31 (PECAM-1) exists as a dimer and is heavily N-glycosylated', *Biochemical and Biophysical Research Communications,* vol. 261, no. 2, pp. 283-91.

O'Brien*, CD, Cao, G, Makrigiannakis, A & DeLisser, HM 2004, 'Role of immunoreceptor tyrosine-based inhibitory motifs of PECAM-1 in PECAM-1-dependent cell migration', *American Journal of Physiology- Cell Physiology,* vol. 287, no. 4, pp. 1103-13.

Olson, MA, Feig, M & Brooks 3rd, CL 2007, 'Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions', *Journal of Computational Chemistry*.

Pagel, M 1999, *FGF Binding and FGF Receptor Activation by Synthetic Heparan-Derived Di- and Trisaccharides*,

Parrish, RF & Fair, WR 1981, 'Selective binding of zinc ions to heparin rather than to other glycosaminoglycans', *The Biochemical Journal,* vol. 193, no. 2, pp. 407-10.

Pastor, R, Brooks, B & Szabo, A 1988, 'An analysis of the accuracy of Langevin and molecular dynamics algorithms', *Molecular Physics,* vol. 65, pp. 1409-19.

Patny, A, Desai, PV & Avery, MA 2006, 'Homology Modeling of G-Protein-Coupled Receptors and Implications in Drug Design', *Current Medicinal Chemistry,* vol. 13, pp. 1667-91.

Pearson, WR & Lipman, DJ 1988, 'Improved Tools for Biological Sequence Comparison', *Proceedings of the National Academy of Sciences,* vol. 85, no. 8, pp. 2444-8.

Pellegrini, L, Burke, DF, von Delft, F, Mulloy, B & Blundell, TL 2000, 'Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin', *Nature,* vol. 407, no. 6807, pp. 1029-34.

Penkett, CJ, Redfield, C, Dodd, I, Hubbard, J, McBay, DL, Mossakowska, DE, Smith, RAG, Dobson, CM & Smith, LJ 1997, 'NMR Analysis of Main-chain Conformational Preferences in an Unfolded Fibronectin-binding Protein', *Journal of Molecular Biology,* vol. 274, no. 2, pp. 152-9.

Petitou, M, Duchaussoy, P, Driguez, PA, Jaurand, G, Herault, JP, Lormeau, JC, Van Boeckel, CAA & Herbert, JM 1998, 'First synthetic carbohydrates with the full antocoagulant properties of heparin', *Angewandte Chemie. International edition in English,* vol. 37, no. 21, pp. 3009-14.

Petitou, M & van Boeckel, CAA 2004, 'A synthetic antithrombin III binding pentasaccharide is now a drug! What comes next', *Angewandte Chemie International Edition,* vol. 43, no. 24, pp. 3118-33.

Piali, L 1995, 'CD31/PECAM-1 is a ligand for alpha v beta 3 integrin involved in adhesion of leukocytes to endothelium', *The Journal of Cell Biology,* vol. 130, no. 2, pp. 451-60.

Poljak, RJ, Amzel, LM, Avey, HP, Chen, BL, Phizackerley, RP & Saul, F 1973, 'Three-Dimensional Structure of the Fab'Fragment of a Human Immunoglobulin at 2.8-angstrom Resolution', *Proceedings of the National Academy of Sciences,* vol. 70, no. 12, pp. 3305-10.

Powell, MS, Barton, PA, Emmanouilidis, D, Wines, BD, Neumann, GM, Peitersz, GA, Maxwell, KF, Garrett, TP & Hogarth, PM 1999, 'Biochemical analysis and crystallisation of Fc gamma RIIa, the low affinity receptor for IgG', *Immunol Lett,* vol. 68, no. 1, pp. 17-23.

Prager, E 1996, 'Interaction of CD31 with a heterophilic counterreceptor involved in downregulation of human T cell responses', *Journal of Experimental Medicine,* vol. 184, no. 1, pp. 41-50.

Presta, M 2005, from http://www.med.unibs.it/~airc/gfs.html

Ragazzi, M, Ferro, DR, Perly, B, Sinay, P, Petitou, M & Choay, J 1990, 'Conformation of the pentasaccharide corresponding to the binding site of heparin for antithrombin III', *Carbohydrate Research,* vol. 195, no. 2, pp. 169-85.

Ragazzi, M, Ferro, DR & Provasoli, A 1986, 'A Force-Field Study of the conformational characteristics of the iduraonate ring', *Journal of Computational Chemistry,* vol. 7, no. 2, pp. 105-12.

Raghuraman, A, Mosier, PD & Desai, UR 2006, 'Finding a needle in a haystack: development of a combinatorial virtual screening approach for identifying high specificity heparin/heparan sulfate sequence (s)', *Journal of Medicinal Chemistry,* vol. 49, no. 12, pp. 3553–62.

Reddy, MR & Erion, MD 2001, *Free Energy Calculations in Rational Drug Design*, Springer.

Remko, M & Hricovíni, M 2006, 'Theoretical study of structure and properties of hexuronic acid and d-glucosamine structural units of glycosaminoglycans', *Structural Chemistry*, pp. 1-11.

Remko, M, Swart, M & Bickelhaupt, FM 2007, 'Conformational behavior of basic monomeric building units of glycosaminoglycans: Isolated systems and solvent effect', *Journal of Physical Chemistry B,* vol. 111, no. 9, pp. 2313-21.

Reyes, CM & Kollman, PA 2000, 'Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change', *Journal of Molecular Biology,* vol. 297, pp. 1145-58.

Ricard-Blum, S, Feraud, O, Lortat-Jacob, H, Rencurosi, A, Fukai, N, Dkhissi, F, Vittet, D, Imberty, A, Olsen, BR & van der Rest, M 2004, 'Characterization of endostatin binding to heparin and heparan sulfate by surface plasmon resonance and molecular modeling: Role of divalent cations', *Journal of Biological Chemistry,* vol. 279, no. 4, p. 2927.

Rider, CC 1997, 'The potential for heparin and its derivatives in the therapy and prevention of HIV-1 infection', *Glycoconjugate Journal,* vol. 14, no. 5, pp. 639-42.

Rost, B 2001, 'Protein secondary structure prediction continues to rise', *Journal of Structural Biology,* vol. 134, no. 2-3, pp. 204-18.

Rost, B & Sander, C 1994, 'Combining evolutionary information and neural networks to predict protein secondary structure', *Proteins: Structure, Function and Genetics,* vol. 19, no. 1, pp. 55-72.

Rost, B, Schneider, R & Sander, C 1997, 'Protein fold recognition by prediction-based threading', *Journal of Molecular Biology,* vol. 270, no. 1-10, p. 26.

Rost, B, Yachdav, G & Liu, J 2004, 'The PredictProtein server', *Nucleic Acids Research,* vol. 32, no. web server issue, pp. W321-6.

Rostand, KS & Esko, JD 1997, 'Microbial adherence to and invasion through proteoglycans', *Infection and Immunity,* vol. 65, no. 1, pp. 1-8.

Rusnati, M, Coltrini, D, Caccia, P, Dell'Era, P, Zoppetti, G, Oreste, P, Valsasina, B & Presta, M 1994, 'Distinct role of 2-O-, N-, and 6-O-sulfate groups of heparin in the formation of the ternary complex with basic fibroblast growth factor and soluble FGF receptor-1', *Biochemical and Biophysical Research Communications,* vol. 203, no. 1, pp. 450-8.

Rusnati, M, Coltrini, D, Oreste, P, Zoppetti, G, Albini, A, Noonan, D, di Fagagna, FdA, Giacca, M & Presta, M 1997, 'Interaction of HIV-1 Tat Protein with Heparin. Role of the backbone structure, sulfation and size', *Journal of Biological Chemistry,* vol. 272, no. 17, pp. 11313-20.

Ryckaert, J-P, Ciccotti, G & Berendsen, HJC 1977, 'Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes', *Journal of Computational Physics,* vol. 23, no. 3, pp. 327-41.

Sachs, UJH, Andrei-Selmer, CL, Maniar, A, Weiss, T, Paddock, C, Orlova, VV, Young Choi, E, Newman, PJ, Preissner, KT & Chavakis, T 2007, 'The neutrophil specific antigen CD177 is a counter-receptor for endothelial PECAM-1 (CD31)', *Journal of Biological Chemistry,* vol. 282, no. 32, pp. 23603-12.

Sadir, R, Baleux, F, Grosdidier, A, Imberty, A & Lortat-Jacob, H 2001, 'Characterization of the Stromal Cell-derived Factor-1a-heparin complex', *Journal of Biological Chemistry,* vol. 276, no. 11, pp. 8288-96.

Sali, A & Blundell, TL 1993, 'Comparative protein modelling by satisfaction of spatial restraints', *Journal of Molecular Biology,* vol. 234, no. 3, pp. 779-815.

Samudrala, R & Moult, J 1998, 'A graph-theoretic algorithm for comparative modeling of protein structure', *Journal of Molecular Biology,* vol. 279, no. 1, pp. 287-302.

Sanner, MF 1999, 'Python: a programming language for software integration and development', *Journal of Molecular Graphics and Modeling,* vol. 17, no. 1, pp. 57-61.

Sasisekharan, R, Raman, R & Prabhakar, V 2006, 'Glycomics approach to structure-function relationships of glycosaminoglycans', *The Annual Review of Biomedical Engineering,* vol. 8, pp. 181-231.

Sasisekharan, R, Shriver, Z, Venkataraman, G & Narayanasami, U 2002, 'Roles of heparan-sulphate glycosaminoglycans in cancer', *Nature Reviews Cancer,* vol. 2, no. 7, pp. 521-8.

Sasisekharan, R & Venkataraman, G 2000, 'Heparin and heparan sulfate: biosynthesis, structure and function', *Current Opinion in Chemical Biology,* vol. 4, no. 6, pp. 626-31.

Schlessinger, J, Plotnikov, AN, Ibrahimi, OA, Eliseenkova, AV, Yeh, BK, Yayon, A, Linhardt, RJ & Mohammadi, M 2000, 'Crystal structure of a ternary FGF-

FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization', *Molecular Cell,* vol. 6, no. 3, pp. 743-50.

Schneidman-Duhovny, D, Inbar, Y, Nussinov, R & Wolfson, HJ 2005, 'Geometry-Based Flexible and Symmetric Protein Docking', *Proteins: Structure, Function and Bioinformatics,* vol. 60, no. 2, pp. 224-31.

Schneidman-Duhovny, D, Inbar, Y, Polak, V, Shatsky, M, Halperin, I, Benyamini, H, Barzilai, A, Dror, O, Haspel, N & Nussinov, R 2003, 'Taking geometry to its edge: Fast unbound rigid(and hinge-bent) docking', *Proteins: Structure, Function and Genetics,* vol. 52, no. 1, pp. 107-12.

Shao, C, Zhang, F, Kemp, MM, Linhardt, RJ, Waisman, DM, Head, JF & Seaton, BA 2006, 'Crystallographic analysis of calcium-dependent heparin binding to Annexin A2', *Journal of Biological Chemistry,* vol. 281, no. 42, pp. 31689-95.

Shaw, JP, Johnson, Z, Borlat, F, Zwahlen, C, Kungl, A, Roulin, K, Harrenga, A, Wells, TN & Proudfoot, AE 2004, 'The X-ray structure of RANTES: Heparin-derived disaccharides allows the rational design of chemokine inhibitors', *Structure,* vol. 12, no. 11, pp. 2081-93.

Shen, MY & Sali, A 2006, 'Statistical potential for assessment and prediction of protein structures', *Protein Science,* vol. 15, no. 11, pp. 2507-24.

Shionyu-Mitsuyama, C, Shirai, T, Ishida, H & Yamane, T 2003, 'An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins', *Protein Engineering Design and Selection,* vol. 16, no. 7, pp. 467-78.

Sibille, N, Sillen, A, Leroy, A, Wieruszeski, JM, Mulloy, B, Landrieu, I & Lippens, G 2006, 'Structural impact of heparin binding to full-length Tau as studied by NMR spectroscopy', *Biochemistry,* vol. 45, no. 41, pp. 12560-72.

Silvian, L, Jin, P, Carmillo, P, Boriack-Sjodin, PA, Pelletier, C, Rushe, M, Gong, B, Sah, D, Pepinsky, B & Rossomando, A 2006, 'Artemin crystal structure reveals insights into heparan sulfate binding', *Biochemistry,* vol. 45, no. 22, pp. 6801-12.

Simmons, DL 1990, 'Molecular cloning of CD31, a putative intercellular adhesion molecule closely related to carcinoembryonic antigen', *Journal of Experimental Medicine,* vol. 171, no. 6, pp. 2147-52.

Sitkoff, D, Sharp, KA & Honig, B 1994, 'Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models', *Journal of Physical Chemistry,* vol. 98, no. 7, pp. 1978-88.

Snow, AD, Willmer, J & Kisilevsky, R 1987, 'Sulfated glycosaminoglycans: a common constituent of all amyloids?' *Laboratory investigation; a journal of technical methods and pathology,* vol. 56, no. 1, pp. 120-3.

Sobel, M, Soler, DF, Kermode, JC & Harris, RB 1992, 'Localization and characterization of a heparin binding domain peptide of human von Willebrand factor', *Journal of Biological Chemistry,* vol. 267, no. 13, pp. 8857-62.

Solis, FJ & Wets., RJ-B 1981, 'Minimization by random search techniques', *Mathematical Operations Research,* vol. 6, pp. 19-30.

Sondermann, P, Huber, R & Jacob, U 1999a, 'Crystal structure of the soluble form of the human fcgamma-receptor IIb: a new member of the immunoglobulin superfamily at 1.7 A resolution', *Embo J,* vol. 18, no. 5, pp. 1095-103.

Sondermann, P, Jacob, U, Kutscher, C & Frey, J 1999b, 'Characterization and crystallization of soluble human Fc gamma receptor II (CD32) isoforms produced in insect cells', *Biochemistry,* vol. 38, no. 26, pp. 8469-77.

Sondermann, P, Kaiser, J & Jacob, U 2001, 'Molecular basis for immune complex recognition: a comparison of Fc-receptor structures', *Journal of Molecular Biology,* vol. 309, no. 3, pp. 737-49.

Soroka, V, Kolkova, K, Kastrup, JS, Diederichs, K, Breed, J, Kiselyov, VV, Poulsen, FM, Larsen, IK, Welte, W & Berezin, V 2003, 'Structure and interactions of NCAM Ig1-2-3 suggest a novel zipper mechanism for homophilic adhesion', *Structure,* vol. 11, no. 10, pp. 1291-301.

Spassov, VZ, Yan, L & Flook, PK 2007, 'The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: A side-chain prediction algorithm based on side-chain backbone interactions', *Protein Science,* vol. 16, no. 3, pp. 494-506.

Spassov, VZ, Yan, L & Flook, PK 2008, 'LOOPER: A CHARMm Based Algorithm for Loop Prediction', *Protein engineering, design & selection : PEDS.*

Spillmann, D & Lindahl, U 1994, 'Glycosaminoglycan-protein interactions: a question of specificity', *Current Opinion in Structural Biology,* vol. 4, pp. 677-82.

Spillmann, D, Witt, D & Lindahl, U 1998, 'Defining the Interleukin-8-binding domain of heparan sulfate', *Journal of Biological Chemistry,* vol. 273, no. 25, pp. 15487-93.

Srinivasan, SR, Radhakrishnamurthy, B & Berenson, GS 1975, 'Studies on the interaction of heparin with serum lipoproteins in the presence of Ca2+, Mg2+, and Mn2+', *Archives of Biochemistry and Biophysics,* vol. 170, no. 1, pp. 334-40.

Stockinger, H 1990, 'Molecular characterization and functional analysis of the leukocyte surface protein CD31', *The Journal of Immunology,* vol. 145, no. 11, pp. 3889-97.

Stringer, SE, Forster, MJ, Mulloy, B, Bishop, CR, Graham, GJ & Gallagher, JT 2002, 'Characterization of the binding site on heparan sulfate for macrophage inflammatory protein 1alpha', *Blood,* vol. 100, no. 5, p. 1543.

Stringer, SE & Gallagher, JT 1997, 'Specific Binding of the Chemokine Platelet Factor 4 to Heparan Sulfate', *Journal of Biological Chemistry,* vol. 272, no. 33, pp. 20508-14.

Sun, J 2000, 'Contributions of the extracellular and cytoplasmic domains of platelet-endothelial cell adhesion molecule-1 (PECAM-1/CD31) in regulating cell-cell localization', *Journal of Cell Science,* vol. 113, no. 8, pp. 1459-69.

Sun, J, Williams, J, Yan, HC, Amin, KM, Albelda, SM & DeLisser, HM 2004, 'Platelet Endothelial Cell Adhesion Molecule-1 (PECAM-1) homophilic adhesion is mediated by Immunoglobulin-like domains 1 and 2 and depends on the cytoplasmic domain and the level of surface expression', *Journal of Biological Chemistry,* vol. 81, no. 2, pp. 408-18.

Sun, QH, DeLisser, HM, Zukowski, MM, Paddock, C, Albelda, SM & Newman, PJ 1996, 'Individually distinct Ig homology domains in PECAM-1 regulate homophilic binding and modulate receptor affinity', *Journal of Biological Chemistry,* vol. 271, no. 19, pp. 11090-8.

Sun, QH, Paddock, C, Visentin, GP, Zukowski, MM, Muller, WA & Newman, PJ 1998, 'Cell surface glycosaminoglycans do not serve as ligands for PECAM-1. PECAM-1 is not a heparin-binding protein', *Journal of Biological Chemistry,* vol. 273, no. 19, pp. 11483-90.

Taroni, C, Jones, S & Thornton, JM 2000, 'Analysis and prediction of carbohydrate binding sites', *Protein Engineering Design and Selection,* vol. 13, no. 2, pp. 89-98.

Thompson, JD, Higgins, DG & Gibson, TJ 1994a, 'CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', *Nucleic Acids Research,* vol. 22, no. 22, pp. 4673-80.

Thompson, LD, Pantoliano, MW & Springer, BA 1994b, 'Energetic characterization of the Basic Fibroblast Growth Factor-heparin interaction: Identification of the heparin binding domain', *Biochemistry,* vol. 33, no. 13, pp. 3831-40.

Tom, D, Darrin, Y & Lee, P 1993, 'Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems', *The Journal of Chemical Physics,* vol. 98, no. 12, pp. 10089-92.

Tsui, V & Case, DA 2001, 'Theory and applications of the generalized Born solvation model in macromolecular simulations', *Biopolymers (Nucleic Acid Sciences),* vol. 56, pp. 257-91.

van Boeckel, CA, van Aelst, SF, Wagenaars, GN, Mellema, JR, Paulsen, H, Peters, T, Pollex, A & Sinnwell, V 1987, 'Conformational analysis of synthetic heparin-like oligosaccharides containing {alpha}-L-idopyranosyluronic acid', *Recueil des travaux chimiques des Pays-Bas,* vol. 106, pp. 19–29.

van Vlijmen, HWT & Karplus, M 1997, 'PDB-based Protein Loop Prediction: Parameters for Selection and Methods for Optimization', *Journal of Molecular Biology,* vol. 267, no. 4, pp. 975-1001.

Varki, A 1999, *The Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press.

Veldkamp, CT, Peterson, FC, Pelzek, AJ & Volkman, BF 2005, 'The monomer-dimer equilibrium of stromal cell-derived factor-1 (CXCL 12) is altered by pH, phosphate, sulfate, and heparin', *Protein Science,* vol. 14, no. 4, pp. 1071-81.

Verli, H & Guimarães, JA 2004, 'Molecular dynamics simulation of a decasaccharide fragment of heparin in aqueous solution', *Carbohydrate Research,* vol. 339, no. 2, pp. 281-90.

Verli, H & Guimarães, JA 2005, 'Insights into the induced fit mechanism in antithrombin–heparin interaction using molecular dynamics simulations', *Journal of Molecular Graphics and Modeling,* vol. 24, no. 3, pp. 203–12.

Vriend, G 1990, 'WHAT IF: a molecular modeling and drug design program', *Journal of Molecular Graphics,* vol. 8, no. 1, pp. 52-6.

Vyas, AA, Pan, JJ, Patel, HV, Vyas, KA, Chiang, CM, Sheu, YC, Hwang, JK & Wu, W 2005, 'Analysis of binding of cobra cardiotoxins to heparin reveals a new beta-sheet heparin-binding structural motif', *Journal of Biological Chemistry,* vol. 280, no. 10, pp. 9567-77.

Wall, D, Douglas, S, Ferro, V, Cowden, W & Parish, C 2001, 'Characterisation of the anticoagulant properties of a range of structurally diverse sulfated oligosaccharides', *Thrombosis Research,* vol. 103, no. 4, pp. 325-35.

Wang, J, Wolf, RM, Caldwell, JW, Kollman, PA & Case, DA 2004, 'Development and testing of a general amber force field', *Journal of Computational Chemistry,* vol. 25, no. 9, pp. 1157-74.

Wang, JH, Pepinsky, RB, Stehle, T, Liu, JH, Karpusas, M, Browning, B & Osborn, L 1995, 'The crystal structure of an N-terminal two-domain fragment of vascular cell adhesion molecule 1 (VCAM-1): a cyclic peptide based on the domain 1 CD loop can inhibit VCAM-1-a4 integrin interaction', *Proceedings of the National Academy of Sciences,* vol. 92, no. 12, pp. 5714-8.

Wang, Y & Sheibani, N 2002, 'Expression pattern of alternatively spliced PECAM-1 isoforms in hematopoietic cells and platelets', *Journal of Cellular Biochemistry,* vol. 87, no. 4, pp. 424-38.

Wang, Y, Su, X, Sorenson, CM & Sheibani, N 2003, 'Tissue-specific distributions of alternatively spliced human PECAM-1 isoforms', *American Journal of Physiology- Heart and Circulatory Physiology,* vol. 284, no. 3, pp. 1008-17.

Watt, SM, Williamson, J, Genevier, H, Fawcett, J, Simmons, DL, Hatzfeld, A, Nesbitt, SA & Coombe, DR 1993, 'The heparin binding PECAM-1 adhesion molecule is expressed by CD34+ hematopoietic precursor cells with early myeloid and B-lymphoid cell phenotypes', *Blood,* vol. 82, no. 9, pp. 2649-63.

Weiner, SJ, Kollman, PA, Case, DA, Singh, UC, Ghio, C, Alagona, G, Profeta, S & Weiner, P 1984, 'A new force field for molecular mechanical simulation of nucleic acids and proteins', *Journal of the American Chemical Society,* vol. 106, no. 3, pp. 765-84.

Wettreich, A, Sebollela, A, Carvalho, MA, Azevedo, SP, Borojevic, R, Ferreira, ST & Coelho-Sampaio, T 1999, 'Acidic pH modulates the interaction between human Granulocyte-Macrophage Colony-stimulating Factor and glycosaminoglycans', *Journal of Biological Chemistry,* vol. 274, no. 44, pp. 31468-75.

Whitfield, DM, Choay, J & Sarkar, B 1992, 'Heavy metal binding to heparin disaccharides. I. Iduronic acid is the main binding site', *Biopolymers,* vol. 32, no. 6, pp. 585-96.

Whitfield, DM & Sarkar, B 1992, 'Heavy metal binding to heparin disaccharides. II. First evidence for zinc chelation', *Biopolymers,* vol. 32, no. 6, pp. 597-619.

Williams, AF & Barclay, AN 1988, 'The Immunoglobulin superfamily-domains for cell surface recognition', *Annual Review of Immunology,* vol. 6, no. 1, pp. 381-405.

Woods, RJ, Dwek, RA, Edge, CJ & Fraser-Reid, B 1995, 'Molecular mechanical and molecular dynamic simulations of glycoproteins and oligosaccharides. 1. GLYCAM_93 parameter development', *The Journal of Physical Chemistry,* vol. 99, no. 11, pp. 3832-46.

Yan, HC, Pilewski, JM, Zhang, Q, DeLisser, HM, Romer, L & Albelda, SM 1995, 'Localization of multiple functional domains on human PECAM-1 (CD31) by monoclonal antibody epitope mapping', *Cell Adhesion and Communication,* vol. 3, no. 1, pp. 45-66.

Yang, W, Wilkins, AL, Ye, Y, Liu, ZR, Li, SY, Urbauer, JL, Hellinga, HW, Kearney, A, van der Merwe, PA & Yang, JJ 2005, 'Design of a calcium-binding protein with desired structure in a cell adhesion molecule', *Journal of American Chemical Society,* vol. 127, no. 7, pp. 2085-93.

Yates, EA, Santini, F, Guerrini, M, Naggi, A, Torri, G & Casu, B 1996, '1H and 13C NMR spectral assignments of the major sequences of twelve systematically modified heparin derivatives', *Carbohydrate Research,* vol. 294, no. 1, pp. 15-27.

Yip, GW, Smollich, M & Gotte, M 2006, 'Therapeutic value of glycosaminoglycans in cancer', *Molecular Cancer Therapeutics,* vol. 5, no. 9, pp. 2139-48.

Zehnder, JL, Shatsky, M, Leung, LL, Butcher, EC, McGregor, JL & Levitt, LJ 1995, 'Involvement of CD31 in lymphocyte-mediated immune responses: Importance of the membrane-proximal immunoglobulin domain and identification of an inhibiting CD31 peptide', *Blood,* vol. 85, no. 5, pp. 1282-8.

Zhang, C, Vasmatzis, G, Cornette, JL & DeLisi, C 1997, 'Determination of atomic desolvation energies from the structures of crystallized proteins', *Journal of Molecular Biology,* vol. 267, no. 3, pp. 707-26.

Zhang, Y, Boesen, CC, Radaev, S, Brooks, AG, Fridman, WH, Sautes-Fridman, C & Sun, PD 2000, 'Crystal structure of the extracellular domain of a human Fc gamma RIII', *Immunity,* vol. 13, no. 3, pp. 387-95.

Zhao, T & Newman, PJ 2001, 'Integrin activation by regulated dimerization and oligomerization of Platelet Endothelial Cell Adhesion Molecule (PECAM)-1 from within the cell', *The Journal of Cell Biology,* vol. 152, no. 1, pp. 65-74.

APPENDIX A

The coordinates and charges derived using RESP method for the GAG pentasaccharide are listed below:

Atom	Resi -due Unit	x	y	z	GLY-CAM atom type	GAG residue	Charge
1	A	-4.5094	6.1614	-11.5252	CG	IdoA2S	0.6199
2	A	-4.8853	5.4091	-12.6686	OS	IdoA2S	-0.5144
3	A	-4.6315	7.6707	-11.7596	CG	IdoA2S	-0.0031
4	A	-4.7103	5.9857	-13.9517	CG	IdoA2S	0.3865
5	A	-3.5573	6.9996	-13.9594	CG	IdoA2S	-0.3701
6	A	-4.4728	4.8862	-14.9455	C	IdoA2S	0.874
7	A	-5.0142	4.7823	-16.0785	O2	IdoA2S	-0.8757
8	A	-3.6507	4.0136	-14.5578	O2	IdoA2S	-0.8757
9	A	-3.8332	8.1548	-12.9887	CG	IdoA2S	0.6564
10	A	-6.0094	7.935	-11.9955	OS	IdoA2S	-0.4346
11	A	-6.9372	8.6047	-10.8929	S	IdoA2S	1.3992
12	A	-6.6647	10.0181	-10.8552	O2	IdoA2S	-0.7516
13	A	-8.227	8.4257	-11.5078	O2	IdoA2S	-0.7516
14	A	-6.9024	7.778	-9.7144	O2	IdoA2S	-0.7516
15	A	-2.5856	8.7458	-12.6535	OH	IdoA2S	-0.8804
16	A	-5.364	5.9519	-10.8874	H2	IdoA2S	0.015
17	A	-4.3342	8.2898	-10.9032	H1	IdoA2S	0.1016

18	A	-5.6412	6.5115	-14.219	H1	IdoA2S	-0.1117
19	A	-2.648	6.4648	-13.6392	HC	IdoA2S	0.0447
20	A	-3.3792	7.3852	-14.9773	HC	IdoA2S	0.0447
21	A	-4.4338	8.9183	-13.5174	H1	IdoA2S	-0.0437
22	A	-2.6655	9.528	-12.1162	HO	IdoA2S	0.4988
23	B	-3.9379	5.1591	-7.0007	CG	GlcNS6S	0.7036
24	B	-2.9935	6.2999	-7.3977	CG	GlcNS6S	0.0849
25	B	-3.8993	4.1107	-7.948	OS	GlcNS6S	-0.6635
26	B	-3.15	6.6881	-8.8726	CG	GlcNS6S	0.228
27	B	-3.0565	5.4366	-9.7438	CG	GlcNS6S	0.362
28	B	-2.1353	7.5956	-9.259	OH	GlcNS6S	-0.704
29	B	-4.1549	4.4565	-9.2951	CG	GlcNS6S	0.2757
30	B	-4.1923	3.1579	-10.1081	CG	GlcNS6S	0.2906
31	B	-2.898	2.5705	-10.1708	OS	GlcNS6S	-0.4445
32	B	-2.0919	2.243	-11.5004	S	GlcNS6S	1.375
33	B	-3.0463	1.9713	-12.5439	O2	GlcNS6S	-0.7417
34	B	-1.4745	3.514	-11.7776	O2	GlcNS6S	-0.7417
35	B	-1.0332	1.3262	-11.1654	O2	GlcNS6S	-0.7417
36	B	-3.2013	5.8136	-11.1039	OS	GlcNS6S	-0.6564
37	B	-3.2003	7.454	-6.5318	N	GlcNS6S	-0.8341
38	B	-2.092	7.7591	-5.3635	S	GlcNS6S	1.4426
39	B	-0.9335	6.9375	-5.6012	O2	GlcNS6S	-0.7562
40	B	-1.6937	9.1368	-5.4936	O2	GlcNS6S	-0.7562
41	B	-2.6828	7.3591	-4.1127	O2	GlcNS6S	-0.7562

42	B	-3.5293	4.7393	-6.0669	H2	GlcNS6S	-0.0107
43	B	-1.9703	5.8968	-7.3565	H1	GlcNS6S	0.0796
44	B	-4.1444	7.1544	-8.9454	H1	GlcNS6S	0.034
45	B	-2.0581	4.9861	-9.6303	H1	GlcNS6S	-0.0114
46	B	-2.1488	7.8226	-10.1819	HO	GlcNS6S	0.4612
47	B	-5.1881	4.8114	-9.1735	H1	GlcNS6S	0.0048
48	B	-4.5797	3.3733	-11.1143	H1	GlcNS6S	0.004
49	B	-4.8684	2.4282	-9.635	H1	GlcNS6S	0.004
50	B	-4.2076	7.513	-6.1642	H	GlcNS6S	0.3954
51	C	-7.7931	6.9653	-4.6179	CG	IdoA2S	1.0259
52	C	-6.2928	7.0027	-4.3175	CG	IdoA2S	-0.009
53	C	-5.7119	5.5855	-4.4837	CG	IdoA2S	0.2153
54	C	-6.0275	4.9943	-5.8636	CG	IdoA2S	0.0494
55	C	-7.5331	5.1203	-6.1437	CG	IdoA2S	0.4008
56	C	-7.8283	4.726	-7.5616	C	IdoA2S	0.8714
57	C	-7.6078	3.5466	-7.9464	O2	IdoA2S	-0.8448
58	C	-8.2358	7.9249	-4.8854	H2	IdoA2S	-0.1116
59	C	-8.2912	5.5759	-8.3684	O2	IdoA2S	-0.8448
60	C	-7.9966	6.4388	-5.9161	OS	IdoA2S	-0.7046
61	C	-6.2399	4.6704	-3.5391	OH	IdoA2S	-0.6728
62	C	-6.1422	7.3393	-3.2784	H1	IdoA2S	0.1381
63	C	-5.582	7.9113	-5.1579	OS	IdoA2S	-0.493
64	C	-5.8721	9.4672	-5.2987	S	IdoA2S	1.4028
65	C	-4.6803	10.1226	-5.7715	O2	IdoA2S	-0.7247

66	C	-6.7945	9.4391	-6.4041	O2	IdoA2S	-0.7247
67	C	-6.619	9.9134	-4.1512	O2	IdoA2S	-0.7247
68	C	-4.6307	5.6552	-4.2882	H1	IdoA2S	0.0456
69	C	-5.8051	3.9121	-5.9114	H1	IdoA2S	0.0887
70	C	-8.1013	4.4373	-5.49	H1	IdoA2S	-0.119
71	C	-5.8365	3.8073	-3.5456	HO	IdoA2S	0.4162
72	C	-5.26	5.6489	-6.8545	OS	IdoA2S	-0.4898
73	D	-8.3575	6.139	-3.6217	OS	GlcNS6S	-0.6496
74	D	-9.737	6.2156	-3.3666	CG	GlcNS6S	0.2902
75	D	-10.1813	4.9394	-2.6537	CG	GlcNS6S	0.1918
76	D	-9.9607	7.3926	-2.4151	CG	GlcNS6S	0.1109
77	D	-11.6235	5.0695	-2.1507	CG	GlcNS6S	0.1047
78	D	-10.0693	3.8406	-3.5387	OH	GlcNS6S	-0.6585
79	D	-11.9211	6.4214	-1.4684	CG	GlcNS6S	0.7418
80	D	-11.9288	3.9494	-1.2665	N	GlcNS6S	-0.7856
81	D	-12.8917	2.747	-1.8264	S	GlcNS6S	1.372
82	D	-12.1419	2.021	-2.8185	O2	GlcNS6S	-0.7295
83	D	-14.1291	3.3555	-2.2412	O2	GlcNS6S	-0.7295
84	D	-13.1724	1.8603	-0.7271	O2	GlcNS6S	-0.7295
85	D	-11.351	7.4955	-2.195	OS	GlcNS6S	-0.5805
86	D	-9.4262	8.7301	-2.9402	CG	GlcNS6S	0.304
87	D	-10.4623	9.4634	-3.5837	OS	GlcNS6S	-0.4388
88	D	-10.8842	10.9018	-3.0565	S	GlcNS6S	1.38
89	D	-11.9529	10.7489	-2.1035	O2	GlcNS6S	-0.7424

90	D	-11.4493	11.4398	-4.2668	O2	GlcNS6S	-0.7424
91	D	-9.6902	11.6581	-2.7811	O2	GlcNS6S	-0.7424
92	D	-10.3133	6.3351	-4.2997	H1	GlcNS6S	0.0313
93	D	-9.5213	4.7668	-1.7844	H1	GlcNS6S	0.0445
94	D	-9.4281	7.1269	-1.4793	H1	GlcNS6S	0.0453
95	D	-12.2518	5.0261	-3.0585	H1	GlcNS6S	0.076
96	D	-10.3024	3.0104	-3.1357	HO	GlcNS6S	0.4345
97	D	-12.9954	6.6408	-1.5841	H2	GlcNS6S	-0.0711
98	D	-12.5492	4.3295	-0.4999	H	GlcNS6S	0.3455
99	D	-8.3692	8.9222	-3.0663	H1	GlcNS6S	0.0166
100	D	-9.4804	9.2239	-1.9522	H1	GlcNS6S	0.0166
101	E	-11.481	6.4363	-0.1209	OS	IdoA2S	-0.5451
102	E	-12.4747	6.371	0.8918	CG	IdoA2S	0.2483
103	E	-12.2069	5.2317	1.8939	CG	IdoA2S	0.2489
104	E	-12.5398	7.707	1.6591	CG	IdoA2S	0.2938
105	E	-13.377	5.1087	2.6903	OH	IdoA2S	-0.705
106	E	-10.9569	5.546	2.7408	CG	IdoA2S	0.1847
107	E	-11.1128	6.9266	3.3977	CG	IdoA2S	0.2955
108	E	-11.3599	7.9056	2.4105	OS	IdoA2S	-0.5808
109	E	-12.7212	8.8758	0.735	C	IdoA2S	0.8735
110	E	-13.321	8.8576	-0.3729	O2	IdoA2S	-0.8488
111	E	-12.2221	9.9504	1.1637	O2	IdoA2S	-0.8488
112	E	-9.807	5.4781	1.8946	OS	IdoA2S	-0.493
113	E	-8.6211	6.532	1.806	S	IdoA2S	1.4587

114	E	-7.7233	5.7913	0.958	O2	IdoA2S	-0.7704
115	E	-7.9896	6.636	3.0959	O2	IdoA2S	-0.7704
116	E	-9.0545	7.6741	1.0436	O2	IdoA2S	-0.7704
117	E	-13.4963	6.1913	0.5117	H1	IdoA2S	0.0164
118	E	-11.9855	4.2288	1.5029	H1	IdoA2S	0.0415
119	E	-13.4084	7.7213	2.3396	H1	IdoA2S	-0.1036
120	E	-13.3095	4.4546	3.3792	HO	IdoA2S	0.4466
121	E	-10.8086	4.7943	3.5347	H1	IdoA2S	0.0759
122	E	-11.948	6.9055	4.1171	H1	IdoA2S	-0.0241
123	E	-10.2175	7.2146	3.9649	H1	IdoA2S	-0.0241

APPENDIX B

The coordinates and charges derived using RESP method for the GAG disaccharide are listed below:

Atom	Residue Unit	x	y	z	GLY-CAM atom type	GAG residue	Charge
1	A	27.999	40.517	56.351	CG	IdoA2S	0.3101
2	A	26.473	40.594	56.527	CG	IdoA2S	0.1793
3	A	28.429	41.255	55.069	CG	IdoA2S	0.0639
4	A	25.786	40.873	55.188	CG	IdoA2S	0.2559
5	A	26.325	42.173	54.567	CG	IdoA2S	0.784
6	A	27.652	42.426	54.929	OS	IdoA2S	-0.6333
7	A	26.146	42.091	53.183	OH	IdoA2S	-0.8105
8	A	29.916	41.587	55.014	C	IdoA2S	1.0159
9	A	30.392	42.516	55.72	O2	IdoA2S	-0.8762
10	A	30.71	40.941	54.28	O2	IdoA2S	-0.8762
11	A	24.4	41.029	55.422	OS	IdoA2S	-0.5491
12	A	23.291	39.929	55.438	S	IdoA2S	1.4394
13	A	22.265	40.43	54.466	O2	IdoA2S	-0.7313
14	A	23.919	38.641	54.998	O2	IdoA2S	-0.7313
15	A	22.782	39.89	56.849	O2	IdoA2S	-0.7313
16	A	26.01	39.37	57.054	OH	IdoA2S	-0.7302
17	A	28.3077	39.474	56.2807	H1	IdoA2S	0.0311

18	A	26.2344	41.4095	57.2098	H1	IdoA2S	0.0037
19	A	28.2542	40.5755	54.2349	H1	IdoA2S	-0.049
20	A	25.9796	40.0449	54.5062	H1	IdoA2S	0.049
21	A	25.7692	43.0271	54.9539	H2	IdoA2S	-0.0642
22	A	26.4315	42.9984	52.7383	HO	IdoA2S	0.4502
23	A	24.9715	39.4219	57.2004	HO	IdoA2S	0.4576
24	B	29.338	40.365	58.342	CG	GlcNS6S	0.1754
25	B	28.744	40.43	59.758	CG	GlcNS6S	0.1293
26	B	28.928	41.839	60.339	CG	GlcNS6S	0.514
27	B	30.41	42.236	60.265	CG	GlcNS6S	-0.2763
28	B	30.887	42.09	58.81	CG	GlcNS6S	0.327
29	B	30.678	40.757	58.363	OS	GlcNS6S	-0.4869
30	B	28.598	41.164	57.465	OS	GlcNS6S	-0.3837
31	B	27.333	40.049	59.717	N	GlcNS6S	-0.7547
32	B	26.81	38.849	60.697	S	GlcNS6S	1.4132
33	B	27.803	37.727	60.634	O2	GlcNS6S	-0.7418
34	B	26.684	39.454	62.065	O2	GlcNS6S	-0.7418
35	B	25.482	38.446	60.126	O2	GlcNS6S	-0.7418
36	B	28.535	41.839	61.695	OH	GlcNS6S	-0.7998
37	B	32.374	42.396	58.584	CG	GlcNS6S	0.1844
38	B	33.136	41.637	59.497	OS	GlcNS6S	-0.4482
39	B	34.691	41.631	59.656	S	GlcNS6S	1.397
40	B	35.14	43.059	59.732	O2	GlcNS6S	-0.7334
41	B	34.92	40.895	60.944	O2	GlcNS6S	-0.7334

42	B	35.233	40.896	58.467	O2	GlcNS6S	-0.7334
43	B	29.2874	39.3359	57.9863	H2	GlcNS6S	0.1214
44	B	29.2668	39.7291	60.4087	H1	GlcNS6S	0.1029
45	B	28.3226	42.5435	59.7686	H1	GlcNS6S	-0.0282
46	B	30.9973	41.5831	60.9107	H1	GlcNS6S	0.0505
47	B	30.5292	43.2702	60.588	H1	GlcNS6S	0.0505
48	B	30.3025	42.8252	58.2568	H1	GlcNS6S	0.0788
49	B	27.1175	39.7503	58.7124	H	GlcNS6S	0.3538
50	B	27.5259	41.558	61.7674	HO	GlcNS6S	0.4329
51	B	32.6522	42.13	57.5642	H1	GlcNS6S	0.0073
52	B	32.5586	43.4579	58.7463	H1	GlcNS6S	0.0073