# Prediction and identification of B-cell epitopes using protein structure information

**Andersen, Pernille; Lund, Ole**

*Publication date:*
2007

*Document Version*
Early version, also known as pre-print

Ph.D. Thesis

# Prediction and identification of B-cell epitopes using protein structure information

Pernille Andersen

September 14th, 2007

Center for Biological Sequence Analysis
BioCentrum
Technical University of Denmark

# Abstract

Recognition of antigens by antibodies is a fundamental mechanism in the immune system. A major challenge in immunological research is to identify the molecular entities recognized by antibodies (B-cell epitopes). The identification of B-cell epitopes can lead to a better understanding of the mechanisms involved in host-pathogen interactions, and can additionally facilitate vaccine development. The use of protein structure information in B-cell epitope identification is the general topic of this thesis.

Bioinformatic methods for prediction of B-cell epitopes have been used for more than 20 years with the major aim of reducing experimental work in the identification process. These methods have mostly been based on protein sequence information. Recently, the field of bioinformatics has developed considerably, and the amount of structural information on antigen-antibody complexes has increased. This allows for the development of new prediction methods for B-cell epitopes. This thesis describes the development of a new method entitled DiscoTope; the method is based on protein three-dimensional information and predicts residues part of B-cell epitopes. The development of the method was based on a dataset of 76 epitopes from experimentally determined structures of antigen-antibody complexes. It is shown that the new method has a better performance than a classical sequence-based method for predicting residues in contact with antibodies. Additionally, it is demonstrated that the method successfully predicts residues in epitopes which have been identified by using different experimental and sequence analysis methods. The DiscoTope method has a potential for guiding B-cell epitope identification experiments and is available in public web-servers at the Centre for Biological Sequence Analysis (CBS) and via the Immune Epitope Database (IEDB).

Pregnancy-associated malaria (PAM) causes thousands of cases of low birth weight and severe maternal anemia every year. The protein VAR2CSA is thought to play a major role in pregnancy-associated disease, and is expressed on the surface of erythrocytes infected with malaria. Here, the identification of B-cell epitopes in the VAR2CSA protein is described. Human serum from women that are immune to the disease was used for pepscan analysis, which identified several epitopes in the different domains of the protein. The structures of these VAR2CSA domains were predicted and this allowed for the analysis of identified

B-cell epitopes in a three-dimensional context. It is shown, that in general, the identified surface-exposed B-cell epitopes are found in particular sub-domains, and on the same side of the structures. These findings have implications for further development of a VAR2CSA-based vaccine and lead to a better understanding of the overall structure of the VAR2CSA protein on the surface of erythrocytes.

The main conclusion of this thesis is that protein structure information in general has several useful applications in the field of B-cell epitope identification and vaccine design.

# Dansk Resume

Antistoffers genkendelse af antigener er en af immunsystemets fundamentale mekanismer. En af de store udfordringer i immunologisk forskning er at identificere de molekylære enheder, som kan genkendes af antistoffer (B-celle epitoper). Identifikation af B-celle epitoper kan føre til bedre forståelse af de mekanismer som er involveret i vært-patogen interaktioner, og kan yderligere bidrage til udviklingen af vacciner. Brugen af information om proteiners strukturer til identifikation af B-celle epitoper er det overordnede emne i denne afhandling.

Bioinformatiske metoder til forudsigelse af B-celle epitoper har været brugt i mere end 20 år med det primære formål at reducere det eksperimentelle arbejde i identifikationsprocessen. Metoderne har hovedsageligt været baseret på information om proteinsekvenser. Igennem de sidste år har bioinformatikfeltet udviklet sig betydeligt og mængden af strukturel information om antigen-antistofkomplekser er øget. Dette har muliggjort udviklingen af nye metoder til forudsigelse af B-celle epitoper. I denne afhandling beskrives udviklingen af DiscoTope, en ny metode som forudsiger aminosyre-rester i B-celle epitoper, og som er baseret på proteiners tredimensionelle struktur. Udviklingen af metoden er baseret på et datasæt bestående af 76 epitoper fra eksperimentelt bestemte strukturer af antigen-antistofkomplekser. Det vises, at metoden forudsiger aminosyre-rester, som er i kontakt med antistoffer, mere nøjagtigt end en klassisk, sekvensbaseret metode. Desuden vises det, at metoden kan forudsige rester i epitoper, som er identificeret ved hjælp af forskellige eksperimentelle og sekvensanalytiske metoder. Metoden har potentiale til at støtte eksperimenter til identifikation af B-celle epitoper og er tilgængelig igennem offentlige web-servere på Center for Biologisk Sekvens Analyse (CBS) og via the Immune Epitope Database (IEDB).

Graviditets-associeret malaria forårsager årligt tusinder af tilfælde af lav fødselsvægt hos spædbørn og alvorlig anæmi hos mødre. Proteinet VAR2CSA spiller sandsynligvis en stor rolle i denne graviditets-associerede sygdom, og er udtrykt på overfladen af malaria-inficerede erythrocytter. I denne afhandling beskrives identifikationen af B-celle epitoper i VAR2CSA. Humane sera fra kvinder med immunitet overfor sygdommen blev brugt i pepscan-analyse til at identificere adskillige epitoper i proteinets forskellige domæner. Strukturerne af flere af VAR2CSA proteinets domæner blev forudsagt. Dette muliggjorde analyse af

de identificerede B-celle epitoper i en tredimensionel kontekst. Vores resultater viser, at de identificerede overflade-eksponerede B-celle epitoper generelt findes i bestemte subdomæner af strukturerne og på den samme side af strukturerne. Disse opdagelser har betydning for den videre udvikling af en VAR2CSA-baseret vaccine og giver bedre forståelse af VAR2CSA-proteinets overordnede struktur på overfladen af erythrocytter.

Den generelle konklusion af afhandlingen er, at information om proteiners strukturer har mange nyttige anvendelser inden for identifikation af B-celle epitoper og i vaccinedesign.

# Preface

Protein structure is one of my major interests. Coming from the field of protein structural biology, my view of the world is quite biased. The work in my years as a Ph.D. student has allowed me to combine my interest in protein structure with research in the promising field of B-cell epitopes and vaccine design. It has been highly motivating for me to work on several projects in immunological bioinformatics with the overall goal of improving health in society.

This Ph.D. thesis was written at BioCentrum at the Technical University of Denmark under the supervision of Associate Professor Ole Lund at the Center for Biological Sequence Analysis (CBS). The work was funded by NIH contracts HHSN26620040083C and HHSN266200400025C. My work presented in this thesis falls within two major areas: Prediction of B-cell epitopes in general, and identification of B-cell epitopes in a malaria vaccine project. The work in both projects was done in collaboration with other researchers from CBS, and the latter project included additional collaboration with researchers at the Centre for Medical Parasitology, University of Copenhagen and Copenhagen University Hospital. I have chosen to present my results from the two projects in the form of published and submitted articles. This way of presenting scientific results was chosen because this is the most common way of publishing of scientific results today, as opposed to the more traditional approach where all results are assembled into one report.

During my years in the immunological bioinformatics group, I have worked on several projects. Initially, I started training artificial neural networks for protein secondary structure prediction. During this project, I learned many basic principles of bioinformatics, which helped shaping my mind for the work presented in the papers within this thesis. The project led to an in-house program for secondary structure prediction but did not lead to a publication, since the performance of the method was comparable to other methods available in the field, but not significantly better. Therefore, the work on this project has not been included in this Ph.D. thesis.

As part of my Ph.D. program, I had the wonderful opportunity of visiting the group of Professor David Baker, Department of Biochemistry, University of Washington, Seattle, US. I was working 5 months in the start-up phase of an

HIV gp120 redesign project, which was supervised by Ph.D. Bill Schief. The visit was an absolutely outstanding opportunity to interact with and learn from top-researchers in the field of protein structure prediction, protein design and HIV vaccine design. The project goals however, turned out to be harder to reach than expected, and were not pursued further. Despite this experience, I still see a bright future for the combination of protein structure prediction and protein design in vaccine development.

In order to facilitate the reading and understanding of the different presented projects, this thesis starts with an introduction to antibodies, B-cell epitopes, and prediction methods of the latter. This general introduction is followed by two main chapters presenting the different areas and projects I have been working on. Each of these chapters are initiated by introductions on the subjects and some methods mentioned in the papers. The final chapter summarizes the results, conclusions and perspectives of my work.

Pernille Andersen
September 2007

# Acknowledgements

Working on this Ph.d. thesis has been one of the major challenges of my life. I am deeply thankful to all the people I have met on my way through these past years, who have helped me and supported me in my work.

Special thanks to my supervisor Ole Lund. Thank you for all the hours of valuable discussions we have had about the projects, and all the things you have taught me about bioinformatics. Also, thank you for your friendly attitude, your open door and your trustful approach letting me follow my own ways. You have really been "a bridge over troubled water". Thank you for believing in me from the very beginning and giving me the opportunity to work in your group.

Thanks to my other co-authors at CBS: Morten Nielsen, who has always been extremely effective, helpful, and open to give me advise in general. Thank you for all your C-programs, which are useful for almost everything. Thanks to Thomas S. Rask for being my fellow at CBS in malaria research and for insight-bringing evolutionary studies.

Additionally, I am grateful to all my collaborators at the Centre for Medical Parasitology, University of Copenhagen and Copenhagen University Hospital, who have excellent expertise in PAM research and contributed with profound insight and interesting experimental studies to the projects. I have enjoyed the openness and sharing attitude of many skilled members of the group, especially I would like to thank Ali Salanti who has been truly inspiring, motivating and contributing to fruitful discussions.

I would like to thank Anne Mølgaard and Thomas Blicher, my two fellow protein structure nerds. It has been a great pleasure to work with you, and I look back with gratitude on all the hours we have discussed the wonders of protein structures and tools to analyze them. I have learned so much from both of you, and hope to continue sharing our interests, including those not related to protein structure.

Thanks to everybody at CBS, particularly people in the vaccine group, the lunch club, and my former and present office mates for making my daily life run smoothly and joyfully. So many of you have become close friends, and I

truly value these friendships, both in the way they contribute to my work and in personal aspects. I greatly appreciate Søren Brunak for making CBS an interesting and fruitful environment for research. Also, thank you for being formal supervisor in the beginning of my study.

I would like to acknowledge Mette Voldby Larsen, Mikala Grubb, Thomas Blicher and Rodrigo Gouveia-Oliveira, who have helped me with valuable comments to the manuscript of this thesis. I value your support and interest in my work highly.

Thanks to the administrative and technical personnel at CBS, who have been a great help during these years. In particular, I would like to thank Lone Boesen, Johanne Keiding and Dorthe Kjærsgaard for excellent help and friendly attitudes. Thanks to Anne Christiansen, I have had skilled assistance and many joyful train-rides. I further acknowledge the expert assistance from Kristoffer Rapacki in the construction of the DiscoTope web-server and software package and maintaining the high-quality lifestyle of my office plant.

Finally, I am grateful for all the support and inspiration I have had from my parents Lilian and Jørgen, my sisters Malene and Mette, my grandparents Esther and Henning, all my other friends and from Tue. Thank you for being excellent role-models, for believing in me and helping me to pursue my dreams.

# Papers included in the thesis

In this thesis, I present my work in on the following three papers:

- **Prediction of residues in discontinuous B-cell epitopes using protein 3D structures**

  **Haste Andersen P**, Nielsen M, Lund O.

  Protein Sci. 2006 Nov;15(11):2558-67. Epub 2006 Sep 25.

- **Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in *P. falciparum* placental sequestration**

  Dahlbäck M, Rask TS, **Andersen PH**, Nielsen MA, Ndam NT, Resende M, Turner L, Deloron P, Hviid L, Lund O, Pedersen AG, Theander TG, Salanti A.

  PLoS Pathog. 2006 Nov;2(11):e124.

- **Structural insight into epitopes in the pregnancy-associated malaria protein VAR2CSA**

  **Andersen P**, Nielsen MA, Rask TS, Dahlbäck M, Theander TG, Lund O, Salanti A.

  Submitted to PloS Pathogens

# Papers not included in the thesis

During the work on my Ph. D. thesis, I have been a co-author on the additional publications listed below. In order to simplify the thesis, and because I am not the main author on these publications, I have chosen not to include them here.

- **Human pregnancy-associated malaria-specific B cells target polymorphic, conformational epitopes in VAR2CSA**

  Barfod L, Bernasconi NL, Dahlbäck M, Jarrossay D, **Andersen PH**, Salanti A, Ofori MF, Turner L, Resende M, Nielsen MA, Theander TG, Sallusto F, Lanzavecchia A, Hviid L.

  Mol. Microbiol. 2007 Jan;63(2):335-47.

- **Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools**

  Greenbaum JA, **Andersen PH**, Blythe M, Bui HH, Cachau RE, Crowe J, Davies M, Kolaskar AS, Lund O, Morrison S, Mumey B, Ofran Y, Pellequer JL, Pinilla C, Ponomarenko JV, Raghava GP, van Regenmortel MH, Roggen EL, Sette A, Schlessinger A, Sollner J, Zand M, Peters B.

  J Mol Recognit. 2007 Mar-Apr;20(2):75-82.

# Abbreviations

| | |
|---|---|
| 3D | Three dimensional |
| AMA1 | Apical membrane antigen 1 |
| ANN | Artificial neural network |
| AUC | Area under the (ROC) curve |
| BCR | B-cell receptor |
| bp | Base pair |
| C | Constant (domain of an antibody) |
| CDR | Complementarity determining region |
| CEP | Conformational epitope prediction |
| CIDR | Cysteine-rich domain |
| CSA | Chondroitin sulphate A |
| DARC | Duffy antigen receptor for chemokines |
| DBL | Duffy-binding like |
| $D_{KL}$ | Kullback-Leibler distance |
| dN | Non-synonymous |
| dS | Synonymous |
| DV | Depletion values |
| EBA | Erythrocyte-binding antigen |
| ELISA | Enzyme-linked immunosorbent assay |
| EM | Electron microscopy |
| Fab | Fragment antibody binding |
| Fc | Fragment crystallizable |
| GAG | Glycosaminglycan |
| ID2 | Inter-domain 2 |
| IEDB | Immune epitope database and analysis resource |
| Ig | Immunoglobulin |
| IgG | Immunoglobulin gamma |
| H | Heavy (chain of an antibody) |
| HBV | Hepatitis B virus |
| HBsAg | HBV surface antigen |
| HIV | Human immunodeficiency virus |
| L | Light (chain of an antibody) |
| LBW | Low birth weight |
| MHC | Major histocompatibility complex |

| | |
|---|---|
| MSP1 | Merozoite surface protein 1 |
| HMM | Hidden Markow model |
| OD | Optical density |
| OspA | Outer surface protein A |
| PAM | Pregnancy-associated malaria |
| PDB | Protein Data Bank |
| PfEMP1 | *Plasmodium falciparum* erythrocyte membrane protein 1 |
| Pk$\alpha$-DBL | *Plasmodium knowlesi* $\alpha$-DBL |
| RIA | Radioimmuno assay |
| ROC | Receiver operator characteristics |
| RSA | Relative surface area |
| RT | Room temperature |
| S1-3 | Sub-domains 1-3 |
| SNPs | Single nucleotide polymorphisms |
| V | Variable (domain of an antibody) |
| VLP | Virus-like particle |
| VNTR | Variable number of tandem repeats |

# Contents

Chapter 1

# Background

Recognition of antigens by antibodies is an essential mechanism in the immune system. A general challenge in immunological research is to identify the molecular entities recognized by antibodies (B-cell epitopes). This identification can lead to a better understanding of the mechanisms involved in host-pathogen interactions, and can additionally facilitate vaccine development. This chapter is devoted to a general introduction to antibodies and B-cell epitopes. First in this introduction, is a short description of the humoral immune response, antibodies, their structures and interaction with B-cell epitopes. Secondly, a number of sections describe B-cell epitope classification, identification and B-cell epitopes in vaccine design. The last sections of the chapter describe prediction methods for B-cell epitopes. A major part of the following sections, which describe the immune system in general, is based on informations from the book "'Immunobiology"' by Janeway, Travers, Walport and Shlomchik (Janeway et al. 2005).

## 1.1 The vertebrate immune system

The immune system is the body's defense against foreign agents, such as infectious organisms, toxins and other molecules that are not part of the body. The defense can be divided into the innate and adaptive responses, which constitute the first and second lines of defense, respectively. The innate immune response is rapid and non-specific; it is mediated by the recognition of conserved structural patterns found primarily in microorganisms. The adaptive immune response is directed specifically to target individual infectious agents and is developed over time. After the infection is resolved, memory cells persist in the body and can induce a rapid and effective response if the harmful agent or infectious organism is encountered again.
The adaptive immune system has two major branches: the cellular immune response mediated by the T lymphocytes (T cells) and the humoral immune response mediated by antibody-secreting B lymphocytes (B cells). The responses of both systems are based on receptors which specifically bind minor parts of the foreign agent called epitopes. The molecules containing the epitopes are called

antigens. Antibodies recognize epitopes exposed on the surface of infectious organisms, where the antigens are in the native conformations. In contrast, T-cell receptors recognize short linear peptides, which are products of intra-cellular enzymatic digestion. T-cell epitopes are presented by the major histocompatibility complex (MHC) proteins to bind the T-cell receptors, and the binding may result in a $CD4^+$ T-helper cell response or a $CD8^+$ cytotoxic T-cell response.

## 1.2 Mechanisms of antibody-mediated immunity

Antibodies contribute to the immunity against pathogens in three main ways. The first is by neutralization, a process where antibodies bind functional sites of antigens, thereby hindering the binding of antigens to receptors on target cells. Neutralization is important for immunity against microorganisms, which infect host cells by adhesion and subsequently enter the target cell to multiply. Neutralization can additionally prevent bacterial toxins from entering cells. Opsonization is the second way of antibody-mediated immunity; in this process, antibodies cover the surface of pathogens and mediate the destruction of the pathogen by phagocytosis. Finally, binding of antibodies can activate the complement system, which leads to binding of complement proteins to the pathogen. Complement proteins facilitate opsonization of the pathogen, and can form pores in the membrane of pathogens, thus leading to lysis of the cells.

## 1.3 B-cell development and antibody production

B lymphocytes originate in the bone marrow. Here, in the early lymphoid development, rearrangement of the V, D, J genes ensures expression of a variety of B-cell receptors (BCRs) on the surface of the immature B cells. These BCRs are membrane-anchored versions of the antibody type immunoglobulin (Ig) M, which in general bind epitopes with low affinity. Subsequently, B cells develop further outside the bone marrow into mature, naïve B cells expressing antibodies of the IgM and IgD types on the surface (Radbruch et al. 2006).
Binding of protein antigens to BCRs on the surface of B cells leads to the internalization and intra-cellular degradation of antigens. Single IgM binding sites generally bind epitopes with low affinity, but if the density of epitopes on the surface of an antigen or organism is high, the simultaneous binding of epitopes by multiple IgM binding sites can lead to high-avidity interactions. After internalization and degradation of the antigen, peptide fragments of antigens are presented on the surface of the B cell by MHC class II molecules. If $CD4^+$ T-helper cells recognize peptides presented by the MHC class II molecules as foreign, they can activate the B cells, which then undergo further development of antibodies in the germinal centers of secondary lymphoid tissues. This development leads to affinity maturation of the antibodies by somatic hypermutation, and to isotype switching.
Somatic hypermutation is an iterative process of directed mutations in the V gene segments, which encode the antibodies, and of subsequent antigen-mediated selection. It leads to the development of antibodies that bind epitopes with high affinity. Following affinity maturation, the process of isotype switching involves changes in the expression of C domain genes, which lead to

the expression of high-affinity antibodies of the types IgG, IgA, and IgE.

After the switching has occurred, the B cell may differentiate into a plasma cell or a memory B cell. Plasma B cells are specialized for high expression of antibodies. Antibodies produced by plasma B cells are secreted and do not contain a trans-membrane region. Long-lived plasma cells ensure the secretion of high-affinity antibodies, which provide immediate protection if the antigen is encountered again during a second infection.

Memory B cells are additionally long-lived; they do not secrete antibodies, but express antibodies on the cell surface. When antigens are bound to these antibodies during a second infection, memory cells can differentiate into plasma cells. These plasma cells can mediate rapid responses of greater magnitude with antibodies of higher affinity than those secreted during the primary infections. Therefore, both long-lived plasma cells and memory cells are crucial to immunity (Tarlinton and Lew 2007).

Activation of B cells through the BCR can also occur without T-helper cells. Especially the binding of bacterial polysaccharides to IgM on the surface of B cells can lead to the secretion of IgM and generation of memory B cells, which have not gone through somatic hypermutations or isotype switching. The immunity provided by these cells is based on high-avidity binding of the antigens. In addition, the binding of several repeated epitopes can lead to cross-linking of BCRs. This cross-linking leads the B cell to further differentiation and mediates immunological memory (Weintraub 2003).

## 1.4 Antibody structure

A single antibody is formed by four polypeptide chains, two heavy (H) chains and two light (L) chains. The H and L chains together form heterodimers which interact to form a Y or T-shaped homodimer (Figure 1.1). The basic tertiary structure of both H and L chains is the immunoglobulin domain, a sandwich of $\beta$-sheets stabilized by disulfide bridges, which is also found in the T-cell receptor. The L chains contain two immunoglobulin domains, and the H chains contain four or more. The conformation between the H and L chains in the antibody is stabilized by disulfide bridges, hydrophobic interactions and hydrogen bonds (Paul 2003).

### 1.4.1 The variable antibody regions

Antibody domains are divided into variable (V) domains and constant (C) domains. The V domains contain the complementarity determining regions (CDRs), which are varied by the process of affinity maturation to bind epitopes. These CDRs are loop regions in the V immunoglobulin domains, which define the overall shape of the epitope binding regions, called paratopes. A monomeric antibody has two paratopes, each at the end of the two antibody arms. Each paratope can be formed by residues of three CDR loops from the H chain (H1-3) and three CDR loops from the L chain (L1-3) (see Figure 1.1).

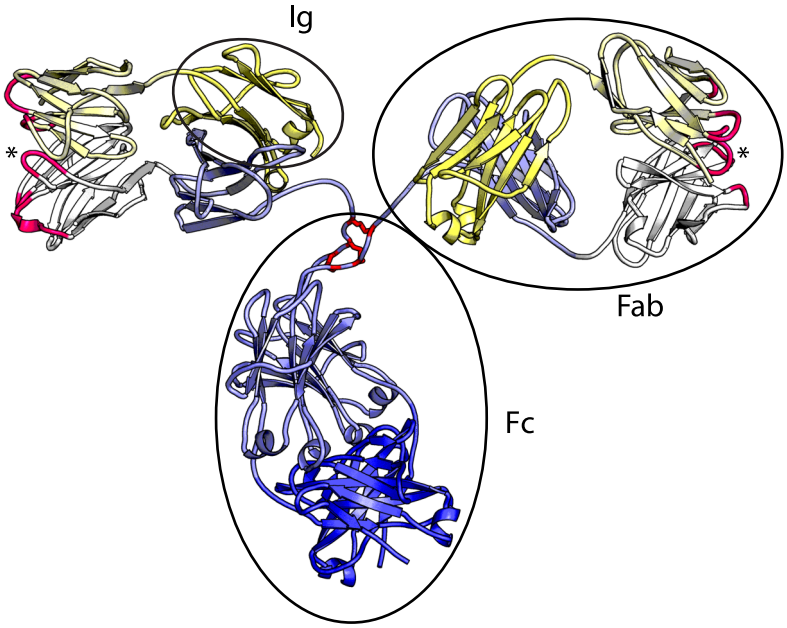


Figure 1.1: **An example of an IgG antibody structure.** A secreted mouse antibody (PDB code 1GTT)(Harris et al. 1997) is shown here. Light chains are highlighted in yellow colors and heavy chains are highlighted in white and blue colors. Ellipses indicate a single Ig domain (Ig), a Fab fragment (Fab) and the Fc fragment (Fc). Intra chain disulfide bridges linking the heavy chains in the Fc fragment are shown in red. Complementarity determining regions (CDRs) are colored in pink and indicated with asterisks.

### 1.4.2 The constant antibody regions

As mentioned previously, isotype switching is caused by changes in expressions of C domain genes. The isotype of secreted antibodies influence their properties of multimerization: IgG, IgD and IgE are always in monomeric forms, while IgM forms pentamers and more rarely hexamers, and IgA can form dimers in addition to monomers. The isotype of the antibodies additionally mediates a different localization of antibodies. IgG is the most prevalent type of Ig in plasma. IgM is also found in plasma, while IgA is the major isotype of the mucosa and IgE is more frequent in the skin (Paul 2003).

### 1.4.3 Fragments of antibodies

Antibodies can be cleaved enzymatically in a flexible hinge region of the heavy chains, and this leads to different types of fragments. The Fab (fragment antibody binding) consists of a single arm of the antibody, including both C and V domains and binds antigens monovalently. The remaining fragment, the Fc (fragment crystallizable) contains the rest of the C domains. Despite the naming of the Fc fragments, it is the Fab fragments that have played the primary

role in crystallographic studies of antibodies, because they are used to study the antibody-antigen interactions.

## 1.5 Antibody-antigen interactions

B-cell epitopes are defined by their binding of antibodies. Therefore, detailed analysis of antibody paratopes can provide more insight in the characteristics of epitope-antibody binding, and this knowledge is essential for a general understanding of B-cell epitopes, which can potentially be used for B-cell epitope prediction or identification.

### 1.5.1 Paratope topography

A diversity in length of the CDRs contributes to different overall topographies of the paratopes. It has been shown that paratope-binding haptens (which are small organic molecules) often are concave or forming deep pockets, whereas peptide-binding paratopes more frequently are either moderately concave or forming grooves. Protein-binding paratopes are more likely either moderately concave or planar (MacCallum et al. 1996; Collis et al. 2003). However, paratope shapes against protein antigens are diverse. For instance, Saphire et al. have shown that the human antibody IgG b12 paratope has a unusually long, protruding CDR H3 loop (Saphire et al. 2001).

### 1.5.2 Amino acid composition

Sequence composition in paratopes from a number of X-ray crystallography structures has been investigated by Collis et al. In general, it was shown that amino acids tyrosine, tryptophan, isoleucine and serine are more abundant in paratopes binding protein antigens than in a set of generic protein loops. Especially tyrosine and tryptophan are both capable of forming a variety of different interactions with other amino acids because of their aromatic, hydrophobic, and hydrophilic character. The amino acids glutamate, aspartate, lysine, cysteine and proline had a lower abundance in protein binding paratopes, compared to generic protein loops (Collis et al. 2003).

### 1.5.3 Binding forces

The binding of antibodies to antigens is mediated by a number of different forces. Electrostatic forces play a role in salt bridges and hydrogen bonds between polar and charged groups in the binding interface. These types of electrostatic interactions are relatively long-ranged, whereas electrostatic van der Waals forces are short-ranged. Van der Waals forces are also important for the packing of non-polar, hydrophobic groups in the interface. In addition, hydrophobic forces can stabilize the binding by exclusion of water molecules from the local environment of hydrophobic groups (Lo Conte et al. 1999). However, water molecules in the interface can also stabilize the binding by mediating hydrogen bonds between the paratope and epitope (Cohen et al. 2005). Crystal structures of antibody-antigen complexes have additionally shown that interfaces have a high degree of shape-complementarity. This ensures a high degree of packing, which is similar to protein interiors, and facilitates binding interactions (Lo Conte et al. 1999).

### 1.5.4 Models of binding mechanisms

Conformational changes in antibody-antigen binding have been studied by a number of groups; X-ray crystallography has been used to determine structures of antibodies and antigens in bound and unbound conformations. These have shown that small conformational changes are not uncommon for antibodies binding protein antigens. The mechanism for this type of binding has been described as the "induced-fit" model (Wilson and Stanfield 1993; Stanfield et al. 2007). However, a more rigid binding has been described for antibodies binding haptens, leading to the conclusion that the binding mechanism is closer to a "lock-and-key" fit (Wedemayer et al. 1997).

## 1.6 B-cell epitopes

As described previously, antibodies can bind a variety of different types of molecules including proteins, peptides, haptens, and polysaccharides. Often, a B-cell epitope is characterized by the type of antigen it is derived from, and this can be misleading. An antibody recognizes an entity composed of atoms with specific chemical features and spatial arrangement. For protein epitopes, this often means that only part of the residues in the epitope are included in the antibody binding entity (Van Regenmortel 1996). The entity may also be formed by antigens of different nature. For instance, a peptide may be able to cross-react with an antibody raised against a carbohydrate epitope (Lo Passo et al. 2007). Such substances, which can mimic the natural epitopes, are called mimotopes. However, because the main focus in this report is the identification of epitopes in protein antigens, the term *epitope* in the following text will refer to residues in proteinaceous B-cell epitopes.

### 1.6.1 Classification of B-cell epitopes

B-cell epitopes are classified into two different groups. The first group consists of linear or continuous epitopes. A continuous epitope comprises a single, consecutive stretch of amino acids in the protein sequence, which is specifically recognized by an antibody raised against the intact protein.
The second group is formed by conformational or discontinuous epitopes. These are epitopes which are composed of residues separated in the protein sequence, but in spatial proximity because of the protein fold. An example of a discontinuous epitope is shown in Figure 1.2. The mouse antibody 184.1 binds on the surface of the *Borrelia burgdorferi* outer surface protein A (OspA), and the epitope is formed by several loops which are proximal in the folded protein (see Figure 1.2A and B), but separated in the sequence. Together, the residues of the discontinuous epitope form a more continuous surface which can interact with the antibody (see Figure 1.2C)(Li et al. 1997).
However, the limit separating the two groups of B-cell epitopes is not completely clear. A discontinuous epitope may consist of several linear epitopes, which together form the antibody interaction site. In addition, continuous epitopes may contain residues which are not interacting with the antibody (Van Regenmortel 1996). Since the majority of antibodies raised against complete proteins does not cross-react with peptide fragments, which are derived from the same protein, it is thought that the most epitopes are discontinuous. It has been estimated

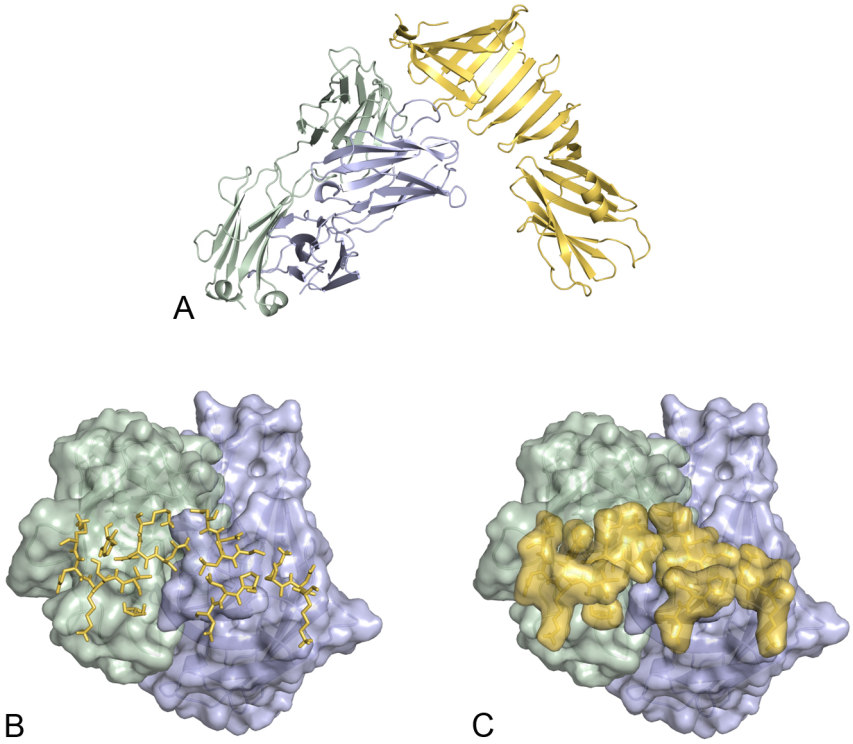


Figure 1.2: **An example of a discontinuous epitope.** The structure of the *Borrelia burgdorferi* outer surface protein A (OspA) in a complex with the mouse antibody 184.1 (PDB code 1OSP) (Li et al. 1997). The OspA protein is shown in yellow, the Fab heavy chain is shown in blue and the Fab light chain in green. (A) The overall structure of OspA and the Fab fragment. Note the different loops interacting with the Fab fragment. (B) The Fab fragment surface and the OspA epitope residues shown as sticks. (C) Surface representations of the Fab and the epitope.

that approximately 90% of B-cell epitopes in globular proteins are discontinuous in nature (Thornton et al. 1986; Barlow et al. 1986).

### 1.6.2 B-cell epitope identification

Several different experimental methods can be used for B-cell epitope mapping or identification, and most of them are useful for studies of protein-protein interactions in general. Here is described a number of different experimental methods, which are referred to or applied in the articles of this thesis.

#### Phage display

In phage display, a set of peptides (called a library) is expressed on the surface of bacteriophages by using recombinant gene technology techniques. In B-cell epitope identification experiments, phages are incubated with antibodies to select phages which express specifically binding peptides. These selected phages are then amplified by infection of *E.coli* and subsequently sequenced to identify

the binding peptides. Libraries expressed in phages can be a random set of peptides or a genome fragment library; the latter can represent peptides from a specific organism or protein of interest (Wang and Yu 2004). Phage display is most commonly used for identifying linear epitopes, but discontinuous epitopes have also been identified using this technique. The latter is a more complex task, because an antibody directed against a discontinuous epitope may only cross-bind peptides with sequences that are not found in the native protein. Lately, a number of bioinformatic tools have been developed which can be used to map identified cross-binding peptides to the three-dimensional (3D) antigen structures (Enshell-Seijffers et al. 2003; Mumey et al. 2003; Schreiber et al. 2005; Moreau et al. 2006; Bublil et al. 2007; Castrignano et al. 2007). In addition to epitope mapping, phage display technology can also be used for discovering mimotopes and carbohydrate epitopes binding to antibodies of special interest (Wang and Yu 2004; Untersmayr et al. 2006; Lo Passo et al. 2007; Mayrose et al. 2007; Saphire et al. 2007).

### Pepscan analysis

In pepscan analysis, a library of synthetic peptides is tested for antibody binding (Geysen et al. 1984). The peptides are often overlapping fragments of the antigen of interest, but can also be libraries of random peptides (Slootstra et al. 1997). Synthetic peptides are immobilized on surfaces of pins or cards, for example by inserting a cysteine residue in the middle of the peptide and using it for covalent linking. Subsequently, enzyme linked immunosorbent assays (ELISAs) are used to test the binding of antibodies to the individual peptides (Uthaipibull et al. 2001; Hijnen et al. 2004). Similarly to phage display, the method is mostly used for identifying linear epitopes, but can also be used for mapping discontinuous epitopes and discovery of mimotopes (Timmerman et al. 2004). In chapter 3 research results are described from a malaria vaccine project, where pepscan analysis of 31-mer peptides was used. The peptides were overlapping by 6 residues, and attached to a card by a covalent link to a cysteine inserted in the middle of the peptide.

### Site directed mutagenesis

Mutational studies of antigens have been shown to be useful for detailed epitope mapping. A simple approach is alanine-scanning, where single amino acids are replaced by alanine residues, and the effects on the antibody binding affinity are measured (Tarr et al. 2006). Mutagenesis of a single amino acid in an epitope can help to determine if the residue contributes to binding. More elegant substitutions can also be used: for instance, the role of the positive charge in a lysine residue can be investigated by replacement with isoleucine (Uthaipibull et al. 2001). Isoleucine is similar in size and hydrophobicity, but lacks the positive charge of lysine. Multiple amino acid substitutions can in this way lead to profound insights into the type of binding interactions between epitope and antibody.

### X-ray crystallography

X-ray crystallography is a powerful tool for analysis of protein interactions in general. It has been used for detailed epitope mapping in studies of a large

variety of antigens (Fischmann et al. 1991; Chitarra et al. 1993; Fleury et al. 2000; Mirza et al. 2000; Igonet et al. 2007). Before the determination of protein structure, different techniques are used for proteins to form crystals. This step can be a limiting factor because some proteins are less crystallizable than others, and the optimal conditions for crystallization can be difficult to reach. When suitable crystals are obtained, these are irradiated with X-rays and the resulting diffraction pattern can be transformed into a 3D electron density map, which can be used to build a model of the protein structure (Branden and Tooze 1998). For analysis of antibody-antigen interactions, X-ray crystallography is typically used to determine the structure of the antigen in complex with a Fab fragment. Crystal structures of antibody-antigen complexes result in high quality epitope mapping, because information about antibody-antigen interactions is obtained on the atomic level, in contrast to many other methods which only assign interactions to amino acid residues. The level of information allows for in-depth analysis of the interactions, thus obtaining more knowledge of shape complementarity, types of interactions, and water molecules in the interface. Information about the antigen structure alone is also useful. The mapping of experimental results on the structure can allow for further analysis and bring one dimensional (1D) sequence-based results into a 3D context. For instance, the bioinformatic tools for mapping mimotopes identified by phage display (mentioned in above) use the 3D structure of the antigen. Also, the mapping of linear epitopes identified by pepscan analysis can help to identify discontinuous epitopes in the native protein (Hijnen et al. 2004).

## 1.7 B-cell epitope vaccines and diagnostic tools

The main goal of research in B-cell epitopes is to develop vaccines or diagnostic tools. Historically, vaccines have been based on responses to the entire pathogen. Killed or attenuated organisms have been used for the vaccines to raise the immunity while avoiding hazardous effects. These strategies have been effective in diminishing the occurrence of major diseases, such as smallpox and polio. However, there are several drawbacks of these types of vaccines. In general, the practical use of killed or attenuated pathogens can be affected by problems caused by producing of pathogen in sufficient amounts, safety, and genetic evolution of pathogens (Arnon and Ben-Yedidia 2003).
Today the field is moving more toward “rational vaccine design”. The basic idea of this approach is to use knowledge about the pathogen, the immune responses against the pathogen, and general host-pathogen interactions in order to design more efficient vaccines. However, even the rational approach to vaccine design is still heavily dependent on experimental trials, since it is hard to predict the response of a new vaccine in complex systems such as the human body (Van Regenmortel 2007). Some of the problems in rational vaccine design are mentioned in sections 1.7.1 and 1.7.2. One general approach of modern vaccine design is the use of more simple vaccine formulations containing non-infecting subunits of the pathogen (Ellis 1999). Subunit vaccines have shown to be useful for vaccination. For instance, virus-like particles (VLPs) are assembled of the human papilloma virus (HPV) proteins L1 and L2 (see Figure 1.3), and the VLPs are used for HPV vaccines preventing viral infection and lowering the risk of genital cancer (Ljubojevic 2006). One of the advantages of this approach is

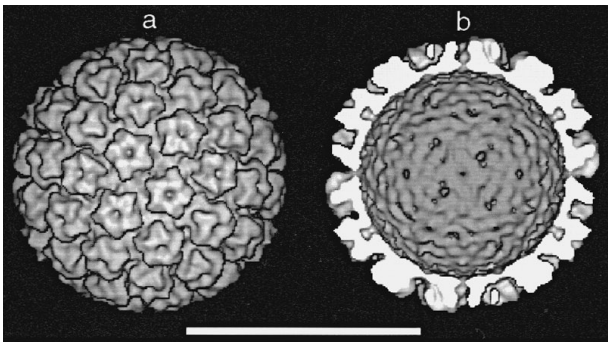


Figure 1.3: **A model of a virus-like particle (VLP).** The structure of a VLP formed by the human papilloma virus proteins L1 and L2. (A) Outside view of the VLP. (B) Inside view of the back half of a VLP. The bar represents 50 nm. The VLP structure was investigated by using cryo-electron microscopy (cryo-EM) (Hagensee et al. 1994) and 3D modeled by Belnap et al. (Belnap et al. 1996).

that these new vaccines should be more safe and not lead to infections.

Another example of a recombinant subunit vaccine, which has been approved for human use, is the hepatitis B virus (HBV) DNA vaccine (Keating and Noble 2003). This vaccine is based on the HBV surface antigen (HBsAg) and produced in genetically modified yeast.

Recently, one of the major research-topics in rational vaccine design has been human immunodeficiency virus (HIV) vaccines. A vaccine, which elicits neutralizing antibodies and effectively protects against HIV infection, has shown to be extremely difficult to develop, and multiple approaches using modern vaccine design technologies are still used in the pursuit of an HIV vaccine (Douek et al. 2006). Other examples of major research projects are B cell-based vaccines against infection of malaria parasites. The RTS,S subunit vaccine has been an outcome of these malaria research projects. RTS,S vaccine is based on a fusion-protein of a polypeptide from *Plasmodium falciparum* circumsporozoite protein and HBsAg (Heppner et al. 2005) and clinical trials of the vaccine have been promising (reviewed by (Matuschewski 2006)). However, even though the number of technologies for vaccine design is growing, developing a vaccine is a complex task that is still mostly based on labor-intensive experimental studies.

B-cell epitope-based diagnostic tools also constitute a major research topic. In the diagnosis of infectious diseases, B-cell epitope binding assays can be used to detect humoral responses against pathogen-specific epitopes (Hsueh et al. 2004; Hamby et al. 2005). In addition, B-cell epitopes have a potential for diagnosis of autoimmune diseases (Selak et al. 2003; Mahler et al. 2003), allergy (Eigenmann 2004) and cancer (Valmori et al. 2005). The development of effective B-cell epitope-based diagnostic tools is not trivial, however, because non-specific antibody cross-reactivity and reactivity resulting from exposure, but not infection, can affect the rate of false positives.

### 1.7.1 Vaccines based on linear epitopes or peptides

Peptides containing linear epitopes are considered to have a high potential for vaccines. In addition to the advantages of subunit vaccines mentioned above, peptides are easily synthesized, purified, stored and handled. However, it has become clear that efficient peptide-based vaccines in general are complex to develop.

Peptides in vaccines must be immunogenic, i.e. have the ability to elicit antibodies which cross-react with the native protein and protect against infection or pathogenesis. Studies testing peptide epitopes are based on cross-reactive antigenicity: the ability for a peptide to be recognized by an antibody raised against a native protein. However, antigenic peptides are usually not very immunogenic and there can be several reasons for this lack of immunogenicity: Most peptide-based vaccines would rely on $CD4^+$ T-helper cell initiated immune responses, and the B-cell epitope itself may not contain an MHC class II epitope. This can be solved by fusing the peptide with residues containing an efficient MHC class II epitope (Sabhnani et al. 2003).

Another major problem is that the binding of antibodies to continuous epitopes is conformation-dependent. A peptide in solution may have a broad variety of structural conformations, and the conformations to which antibodies are developed may not be similar to the conformation in the native protein. To solve this problem, conformationally restricted peptides have been tested; for instance disulphide bridges or other covalent links have been introduced in peptides to stabilize loop conformations (Cabezas et al. 2000; Sabo et al. 2007). Conformational restriction of peptides can also help to circumvent another problem of short peptides, which is fast degradation by peptidases in the human body (Hans et al. 2006).

Another problem in peptide-based vaccines is that the humoral immune response is more efficiently initiated when the epitopes are repeatedly presented on larger antigens. To circumvent this problem, a number of different adjuvant and carrier systems have been studied. For instance, VLPs are used in development of a foot-and-mouth disease virus vaccine to present continuous epitopes in a manner that facilitates an immune response (Zhang et al. 2007).

### 1.7.2 Vaccines based on discontinuous epitopes

Discontinuous epitopes are more difficult than linear to use in general vaccine design. Because the epitope is composed of different parts of the protein sequence, it is more complex to conserve the native conformation of the epitope in a recombinant protein or a peptide. The majority of natural epitopes are thought to be discontinuous as mentioned in section 1.6. Therefore, much effort is put into the development of vaccines based on discontinuous epitopes. Mimotopes are considered to have a high potential in vaccines by mimicking discontinuous epitopes (Untersmayr et al. 2006; Saphire et al. 2007). Recombinant proteins are also considered useful for presentation of discontinuous epitopes in vaccines. Koide et al. successfully applied structure-based design for a lyme disease vaccine candidate; the authors removed approximately 45% of the residues from the OspA and stabilized the engineered protein by promoting hydrophobic interactions. The engineered protein presented discontinuous epitopes and had an affinity to monoclonal antibodies similar to the full-length OspA protein (Koide

| Feature | Year | Reference |
|---|---|---|
| Hydrophilicity | 1981 | (Hopp and Woods 1981) |
| Hydrophilicity | 1986 | (Parker et al. 1986) |
| Hydrophobicity | 1982 | (Kyte and Doolittle 1982) |
| Antigenicity | 1985 | (Welling et al. 1985) |
| Accessibility | 1976 | (Chothia 1976) |
| Surface probability | 1978 | (Janin and Wodak 1978) |
| Surface accessibility | 1985 | (Emini et al. 1985) |
| Backbone flexibility | 1985 | (Karplus and Schulz 1985) |
| Secondary structure | 1978 | (Chou and Fasman 1978) |
| Secondary structure | 1978 | (Garnier et al. 1978) |
| Turn prediction | 1993 | (Pellequer et al. 1993) |

Table 1.1: **A variety of propensity scales used for B-cell epitope prediction**

et al. 2005). Recently, Zhou et al. investigated the binding of the neutralizing antibody b12 to stabilized protein constructs derived from the HIV protein gp120 (Zhou et al. 2007), and such constructs have been suggested to have high potential for HIV vaccine design (Douek et al. 2006).
However, the use of engineered proteins also has disadvantages. Recombinant proteins for vaccines must be stable and easily produced in sufficient amounts; for some proteins this is a limiting factor. In addition, the engineered proteins may contain new epitopes on the surface which could be immunodominant and lead to unwanted immune responses.

## 1.8 B-cell epitope prediction

Antibodies have been studied for many decades, and much effort has been put into the delineation of interactions between antibodies and epitopes. Even though "immunological bioinformatics" is a rather new term, computational analysis and prediction of B-cell epitopes have been major areas of research for more than 20 years (Hopp and Woods 1983; Barlow et al. 1986; Thornton et al. 1986). As the entire field of biotechnology has expanded tremendously within the last two decades, development of new methods have led to more insight into the antibody-antigen interactions as well as larger amounts of data. This, in turn, allows for development of new bioinformatic tools and databases (Korber et al. 2006).
Prediction of B-cell epitopes has mostly been based on analysis of protein sequences. However, these predictions are mostly limited to linear epitopes, because the prediction of discontinuous epitopes is thought to require knowledge of the antigen structure.

### 1.8.1 Amino acid propensity scales

Sequence-based epitope prediction methods typically use propensity scales for calculation of a prediction score. Propensity scales are composed by values which describe intrinsic features for each of the 20 amino acids. No single physico-chemical feature has been definitively used for epitope prediction, but

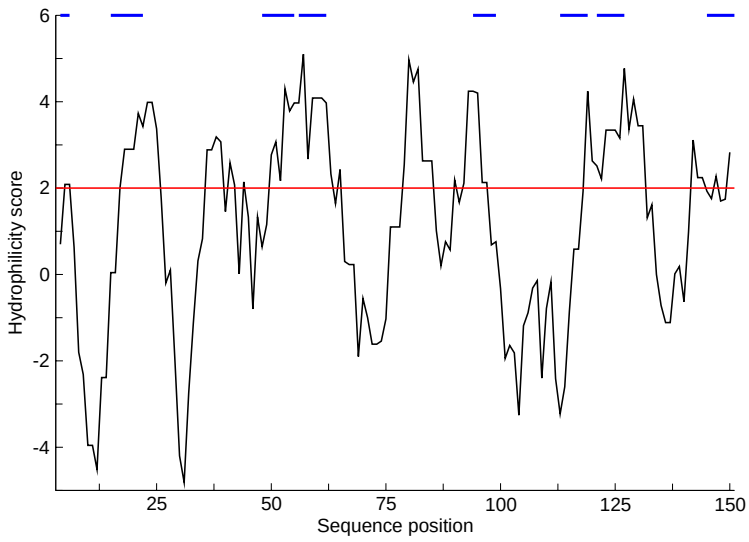


Figure 1.4: **A propensity score plot of the sperm whale myoglobulin sequence as provided by the Immune Epitope Database and Analysis source (IEDB) (Peters et al. 2005).** The Parker hydrophilicity scale and a smoothing window size of 7 residues was used for calculation of scores shown in black. The red line corresponds to prediction threshold value as used in the IEDB. Blue lines indicate experimentally identified continuous myoglobulin epitopes, which were compiled in a data set by Pellequer et al. (Pellequer et al. 1993).

atoms which interact with the paratope have to be surface exposed. In B-cell epitope prediction, the most successful features have been hydrophilicity, accessibility, flexibility or loop/turn structures. In general, the predictions of the propensity scales correlate with features of surface exposure. In Table 1.1 is listed a number of the different propensity scales which have been used for epitope prediction.

Propensity scales are often used in combination with smoothing procedures. The simplest type of smoothing is based on the sliding of a window through the protein sequence and averaging the propensities of the residues within the window. The mean value of the window is then assigned to the residue in the middle of the window. This simple type of smoothing has been used frequently for B-cell epitope prediction, but more sophisticated methods using a weighted average, or a Gaussian smoothing curve have also been applied (Pellequer et al. 1991). The smoothing results in a scoring profile, where the high-scoring regions are predicted to be antigenic. An example of a propensity scoring profile is shown in Figure 1.4. Several tools for antigenicity prediction using combinations of propensity scales have been developed (Jameson and Wolf 1988; Kolaskar and Tongaonkar 1990; Maksyutov and Zagrebelnaya 1993; Pellequer and Westhof 1993; Alix 1999). However, the most extensive sequence-based study, involving 484 different propensity scale methods on a new data set of 50 proteins, concluded that most propensity scales perform close to random and the use of more sophisticated machine learning methods such as artificial neural networks

(ANNs) was proposed (Blythe and Flower 2005).

### 1.8.2 New methods for prediction of continuous epitopes

In 2006, three different approaches using advanced machine learning methods were published: Saha et al. proposed a prediction method based on recurrent ANNs trained on data set of 700 continuous epitopes from the Bcipep database (Saha and Raghava 2006; Saha et al. 2005). Larsen et al. published the Bepipred method (Larsen et al. 2006) based on predictions of a hidden Markow model (HMM) in combination with the Parker hydrophilicity scale (Parker et al. 1986). The method was trained on continuous epitopes of 127 proteins from the AntiJen database (Toseland et al. 2005). Finally, Sollner et al. published a classification algorithm based on the combinations of propensities with sequential residue neighborhood parameters (Sollner and Mayer 2006). The classification algorithm was based on decision trees and nearest neighbor approaches, and was trained on publicly available data sets from Bcipep and FIMM (Saha and Raghava 2006; Schonbach et al. 2002) and a large amount of proprietary data. In total 1211 epitopes and 1211 nonepitope sequences were used for training, and the performance was shown to be greatly increased compared to single propensity scale methods (Sollner and Mayer 2006). Although the performance of these new methods is improved compared to the more simple propensity scale methods, accurate prediction of continuous epitopes remains a challenge in the field of immunological bioinformatics.

### 1.8.3 Prediction of B-cell epitopes based on protein structure

Protein 3D structures have also been used in the prediction of B-cell epitopes. Particularly, the prediction of discontinuous epitopes is thought to require this information. The first prediction methods based on protein structure were published in 1986 by Thornton et al. and Novotny et al. (Thornton et al. 1986; Novotny et al. 1986).

#### Early structure-based methods

The method proposed by Novotny et al. was based on the contacts of a large sphere (called a probe) on the Van der Waals surface of the protein. A similar method is used for calculating solvent accessible areas; here a 1.4 Å probe size, corresponding to the van der Waals radius of a water molecule, is used in a method first reported by Lee et al. (Lee and Richards 1971). Novotny et al. found a probe size of 10 Å radius to correlate well with antigenic epitopes in hen-egg-white lysozyme, sperm whale myoglobulin, cytochrome *c* and myohemerythrin. It was additionally observed that the contacting residues of the large probe also had high solvent accessibility scores, as determined by using a probe size of 1.4 Å.
Thornton et al. used ellipsoids with the same inertia moment as the protein structure in order to model the overall shape of the protein. The sizes of the ellipsoids were varied and chosen so that for protrusion index 9 (PI 9), 90% of the atoms in the structure were inside the ellipsoid. The rest of the atoms (10%) were protruding from (sticking out of) the structure. In general, it was found

that antigenic peptides tend to protrude from the structures of the proteins lysozyme, myoglobulin and myohemerythrin. In addition, it was found that protruding residues had a tendency to be more flexible and accessible.

### Recent structure-based methods

After the publication of studies by Novotny et al. and Thornton et al., only little was reported of epitope prediction on the basis of protein structure. However, the number of available structures of both antigens and antibody-antigens increased, and new approaches have recently been used for this type of epitope prediction. In chapter 2, the DiscoTope method is presented. It was published in 2006, and developed based on discontinuous epitopes in 76 structures of antibody-antigen complexes. Other methods using protein structures have been published recently as well. Some of these are mention below.
The server for Conformational Epitope Prediction (CEP) (Kulkarni-Kale et al. 2005) is based on the calculation of the relative surface accessibility (RSA). Sequence fragments of surface-exposed residues are identified and condensed with other proximal exposed fragments in the structure, this defines regions on the 3D surface which are exposed and possibly act as antigenic regions.
The Epitope Mapping Tool (EMT) (Batori et al. 2006) is based on a sequence library of epitopes in different proteins identified by phage display. In the prediction method, the antigen structure is searched to find surface exposed regions containing motifs of the library.
Rapberger et al. recently published a study of antigen-antibody interaction-site prediction (Rapberger et al. 2007). Their method uses surface accessibility measured with a probe radius of 3Å, inspired by the results published by Novotny et al. (Novotny et al. 1986). Additionally, the shape complementarity to paratopes and interaction energies are included to identify residues with high probability of being part of discontinuous epitopes. Similarly to the DiscoTope method, the work was based on discontinuous epitopes derived from PDB structures of antibody-antigen complexes. The method was tested using structures of free antigens in unbound conformations. It was shown to have a moderate performance and captured 3 of 8 residues in the 1F9 epitope of the apical membrane antigen 1 (AMA1) of *P. falciparum*, which were identified by using phage display (Coley et al. 2006).
As mentioned in section 1.6.2, a number of tools have been developed to facilitate the mapping of epitopes on protein structures by using mimotopes. The methods analyze the protein 3D structure to find regions which can be mimicked by peptides.
In general, the increased number of publicly available methods for B-cell epitope prediction shows that even though the performances are still quite low, when compared to MHC class I epitope prediction, there is much optimism among researchers in the field. Many research groups are continuously working on improving methods, building databases, and evaluating the various methods (Greenbaum et al. 2007). Thus, the field of B-cell epitope prediction is expected to lead to further improved methods, which can aid experimental epitope mapping and vaccine design.

Chapter 2

# DiscoTope, a prediction method for B-cell epitopes

This chapter describes a new method for prediction of B-cell epitopes based on protein 3D structures. The chapter is initiated by introductions to the methods used in the presented paper. Following the introduction, my work is presented in the paper "Prediction of residues in discontinuous B-cell epitopes based on protein 3D structures".

## 2.1 Performance measures

The concept of receiver operator characteristic (ROC) curves (Swets 1988) is widely used to measure the performance of bioinformatic tools. A strength of this method for evaluations of performances, is that the success rate for finding the feature that is predicted for is taken into account, at the same time accounting for the rates of non-successful predictions.
For general evaluation, and in calculation of ROC curves of epitope prediction methods, the following six defined measures are used (Lund et al. 2005):

- True Positives $TP$ is the number of residues that are predicted to be part of epitopes and which are annotated as epitope residues in the data set.
- Actual positives $AP$, is the number of residues annotated epitopes in the data set.
- True negatives $TN$, is the number of residues predicted to be non-epitopes and which are annotated as non-epitopes in the data set.
- Actual negatives $AN$, is the number of residues in the data set which are annotated as non-epitopes.
- *Sensitivity* is the fraction of actual positives which are correctly predicted. It is defined as follows:

$$Sensitivity = \frac{TP}{AP}$$

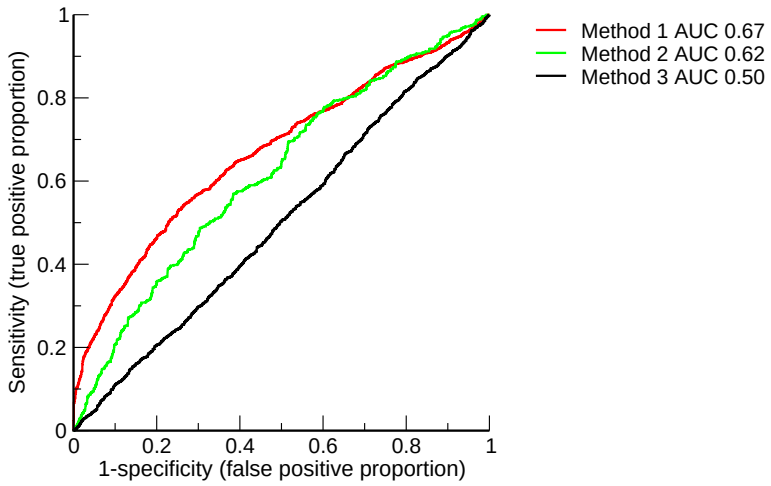


Figure 2.1: **An example of ROC curves for three different methods.** The areas under the curves (AUC) are noted in the legends. Based on the AUC values, the method 1 has the highest performance, method 2 has the second highest performance, and method 3 performs randomly. Note that the curves for method 1 and 2 are overlapping for sensitivity values >0.8. If only the thresholds corresponding to sensitivity values >0.8 were used for comparing method 1 and 2, only a very minor difference in performance would be observed.

- $Specificity$ is the fraction of actual negatives which are correctly predicted:

$$Specificity = \frac{TN}{AN}$$

For many prediction methods, the prediction outcome is a prediction score per residue. A threshold is then used to distinguish predicted positives from predicted negatives. However, depending on the value of the threshold, the values of $sensitivity$ and $specificity$ change. ROC curves are constructed by plotting the $Sensitivity$ as a function of $1 - specificity$ calculated with varied thresholds (see Figure 2.1). The area under the curve (AUC) is used to evaluate the overall performance of the methods. This has the advantage that methods are compared on a more general basis than if only a single threshold is used for each method. For a method predicting randomly, the AUC is 0.5, while the perfect prediction method has an AUC of 1.

## 2.2 Measuring surface exposure

### 2.2.1 Contact numbers

Several methods can be used for measuring the surface exposure in protein structures. A simple approach is based on the contact number of each residue (Nishikawa and Ooi 1980). The contact number for a residue can be calculated by summing the number of C$\alpha$ atoms in other residues within a specific distance from the C$\alpha$ atom of the residue. The number of contacts correlates with surface

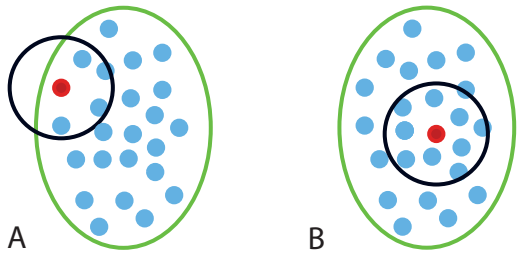


Figure 2.2: **Contacts numbers correlate with surface exposure and structural protrusion.** The surface of a protein is marked here with a green line. C$\alpha$ atoms in the protein are shown in dots. Blue dots denote other residues and red dots denote the residue for which contacts are counted. The area defined by a specific distance threshold is shown as a black circle, and only residues with C$\alpha$ atoms within the circle are counted as contacts. (A) Contacts for a residue located close to the surface in a protein structure, (B) Contacts for a residue located in the interior of a protein structure.

exposure and structural protrusion, because residues with C$\alpha$ atoms close to the surface and in protruding areas have a low number of contacts (see Figure 2.2.)

### 2.2.2 NACCESS surface exposure

The surface exposure of residues in a protein is frequently measured by using a method proposed by Lee and Richards (Lee and Richards 1971). In this approach, the area which is accessible for interaction with the solvent is calculated for each residue. The program NACCESS (Hubbard and Thornton 1993) allows for application of the method on protein structures. The accessible area is calculated by using van der Waals radii of the atoms on the protein surface and a probe of 1.4 Å, which is equal to the van der Waals radius of a water molecule. Values of relative solvent accessibility ($RSA$) for residues are often used to reflect the solvent accessibility. The $RSA$ is calculated as follows:

$$RSA = 100\% \times \frac{SA_{measured}}{SA_{ref,x}}$$

where the $SA_{measured}$ is the measured solvent accessible area of a particular residue. $SA_{ref,x}$ is the reference solvent accessible area for the specific amino acid type X. The reference area is the solvent accessible area in a tripeptide G-X-G, where the amino acid X is surrounded by the small amino acid glycine (G) and therefore assumed to have maximum solvent accessibility.

## 2.3 Amino acid log-odds ratios

In prediction methods for protein sequences, log-odds ratios are often used to reflect the probability that a given amino acid has the predicted property. For example, log-odds ratios can be calculated for each amino acid to reflect how likely for it to be part of a B-cell epitope. The probability of finding a given type of amino acid in an epitope of a data set can be described as:

$$p_{aa} = \frac{n_{aa}}{n}$$

where $p_{aa}$ is the probability for the amino acid, $n_{aa}$ is the number of times the amino acid is observed in epitopes of the data set, and $n$ is the total number amino acids in the data set. Because amino acids have different background frequencies, $q_{aa}$, the probabilities are often divided by the background frequencies observed in a large database such as SwissProt (Bairoch and Apweiler 2000). The resulting ratio is called the odds ratio.
The logarithm to the ratio is often used for practical reasons, so the final log-odds ratio is calculated as:

$$L = \log_k \left( \frac{p_{aa}}{q_{aa}} \right)$$

where $\log_k$ is the logarithm with base $k$. The epitope log-odds ratios presented in the following paper are based on half-bit units using $log_2$. In this example, a high score indicates that the amino acid is more frequently observed in B-cell epitopes than in the SwissProt database.
If a data set contains many similar examples for training, the log-odds ratios can easily be biased toward the redundant examples. To avoid this problem, several refinement techniques can be applied (Lund et al. 2005). For instance, sequence weighting can be used; first the similar sequences are clustered, then weights are assigned for each sequence to down-regulate the influence of highly similar sequences. For data sets of limited size, pseudo-count correction is helpful for types of amino acids which are observed rarely, or not at all. The strength of this technique is that the frequency of rarely observed amino acids are adjusted by using the frequencies of similar amino acids in the data set. The BLOSUM62 substitution matrix is widely used as a similarity measure (Nielsen et al. 2004).
In the paper presented in this chapter, log-odds ratios are calculated for both epitope residues and non-epitope residues and then subtracted to get the final epitope log-odds ratios:

$$L_{e-ne} = L_e - L_{ne}$$

where $L_e$ is the log-odds ratio of a given amino acid type derived from epitope residues, $L_{e-ne}$ is the log-odds ratio of a given amino acid type derived from non-epitope residues and $L_{e-ne}$ is the final epitope log-odds ratio for the given amino acid type.

## 2.4 Contributions to the paper

The following scientific paper was written in collaboration with Morten Nielsen and Ole Lund. My part of the work included compilation and analysis of the data set, development and evaluation of the method, set-up of the web-server, producing all figures and writing the manuscript.

## 2.5 Paper I

*Protein Science,2006 Nov;15(11):2558-67. Epub 2006 Sep 25*

# Prediction of residues in discontinuous B-cell epitopes using protein 3D structures

**Pernille Haste Andersen, Morten Nielsen and Ole Lund***

Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark.
(Received June 16, 2006; Final Revision August 9, 2006; Accepted August 12,2006)

* *Corresponding author:*
Tel: (+45) 45 25 24 25
Fax: (+45) 45 93 15 85
Email: lund@cbs.dtu.dk

## Abstract

Discovery of discontinuous B-cell epitopes is a major challenge in vaccine design. Previous epitope prediction methods have mostly been based on protein sequences and are not very effective. Here we present DiscoTope, a novel method for discontinuous epitope prediction which uses protein three dimensional structural data. The method is based on amino acid statistics, spatial information and surface accessibility in a compiled data set of discontinuous epitopes determined by X-ray crystallography of antibody/antigen protein complexes. DiscoTope is the first method to focus explicitly on discontinuous epitopes. We show that the new structure based method has a better performance for predicting residues of discontinuous epitopes than methods based solely on sequence information and that it can successfully predict epitope residues which have been identified by different techniques. DiscoTope detects 15.5% of residues located in discontinuous epitopes with a specificity of 95%. At this level of specificity, the conventional Parker hydrophilicity scale for predicting linear B-cell epitopes identifies only 11.0% of residues located in discontinuous epitopes. Predictions by the DiscoTope method can guide experimental epitope mapping in both rational vaccine design and development of diagnostic tools and may lead to more efficient epitope identification.

**Keywords:** Discontinuous epitopes; B-cell epitope; antibody; vaccine design; protein structure; antigen; accessibility; hydrophilicity

## Introduction

A major task in vaccine design is to select and design proteins containing antibody binding epitopes (B-cell epitopes) able to induce an efficient immune response. The selection can be aided by epitope prediction in relevant proteins or regions of proteins. In addition, prediction of B-cell epitopes may help to identify epitopes in proteins that have been analysed using experimental techniques

based on antibody affinity binding e.g. western blotting, immunohistochemistry, radioimmunoassay (RIA) and enzyme-linked immunosorbent assay (ELISA).
Most existing methods for prediction of B-cell epitopes exclusively use protein sequences as input and are best suited to predict epitopes composed of a continuous stretch of amino acids (linear epitopes) (Hopp and Woods 1981; Parker et al. 1986; Jameson and Wolf 1988; Debelle et al. 1992; Maksyutov and Zagrebelnaya 1993; Alix 1999; Odorico and Pellequer 2003). In general, these methods are based on prediction of hydrophilicity, flexibility, beta-turns and surface accessibility using a number of amino acid propensity scales. A large amount of data exists on linear epitopes (Leitner et al. 2003; Saha et al. 2005; Toseland et al. 2005), since the annotation can be done by measuring the binding of antigen peptide fragments to antibodies. However, this method of annotation may lead to annotation errors, because a peptide can specifically bind an antibody even if some residues of the peptide are not interacting with the antibody. Predicting linear epitopes is still a non-trivial task and the obtainable prediction accuracy is quite poor (Van Regenmortel and Pellequer 1994; Van Regenmortel 1996; Blythe and Flower 2005). However, combination of a hidden Markow model and a hydrophilicity scale constructed by Parker et al. (Parker et al. 1986) has recently lead to some improvement in linear B-cell epitope prediction (Larsen et al. 2006).
It has been estimated that more than 90% of B-cell epitopes are discontinuous, i.e., consist of segments which are distantly separated in the pathogen protein sequence and brought into proximity by the folding of the protein (Barlow et al. 1986; Van Regenmortel 1996). Identification of discontinuous epitopes is difficult, since the complete analysis must be done in context of the native antigen structure. The most informative and accurate method for identification of discontinuous epitopes is determination of structures of antigen-antibody complexes by X-ray crystallography (Fleury et al. 2000; Mirza et al. 2000). The use of discontinuous epitopes derived from presently available X-ray structures is complicated by two major problems: First, the available data on discontinuous epitopes in different antigens is much reduced compared to linear epitopes. Second, very few antigens have been studied to completely identify various discontinuous epitopes in the same antigen. The existence of undetected epitopes which are not identified in the data set can make it harder to develop good prediction algorithms because they influence on the measured performance. However, detailed structural knowledge on antibody-antigen complexes is growing and allows for broader analysis of discontinuous epitopes in various antigens and development of better prediction methods.
Correlation between surface exposure and B-cell epitopes has been known for many years (Novotny et al. 1986; Thornton et al. 1986). Recently, two new methods using protein structure and surface exposure for prediction of B-cell epitopes have been published (Batori et al. 2006; Kulkarni-Kale et al. 2005). However, none of these new methods using protein structure as input have the primary focus on discontinuous epitopes.
Here we present a prediction method for residues located in discontinuous B-cell epitopes. DiscoTope uses a combination of amino acid statistics, spatial information and surface exposure. It is trained on a compiled data set of discontinuous epitopes from 76 X-ray structures of antibody/antigen protein complexes. We present the performance of DiscoTope compared to the Parker hydrophilicity scale (Parker et al. 1986) for a comparison to a classical, sequence based method

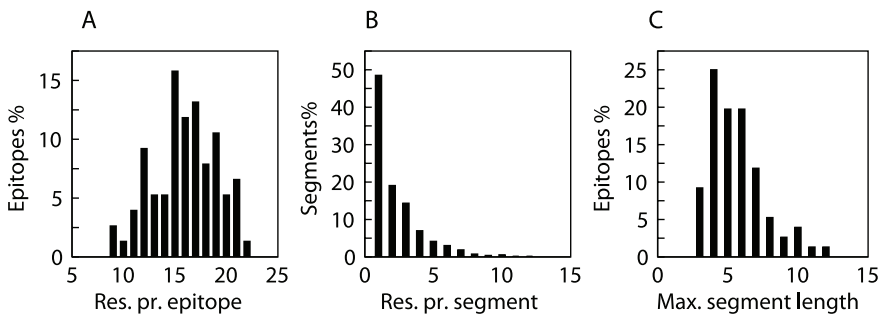


Figure 2.3: **Analysis of the complete dataset of discontinuous B-cell epitopes.** Analysis of the complete dataset of discontinuous B-cell epitopes. (A) Distribution of the number of residues pr. epitope. (B) Distribution of the number of residues pr. sequential stretch of epitopes. (C) Distribution of the maximum length of a sequential stretch pr. epitope.

which has been shown recently to perform well for prediction of linear epitopes (Larsen et al. 2006). In addition, we compare the performance with predictions based on surface accessibility measured on antigen structures using the program NACCESS (Hubbard and Thornton 1993). We demonstrate that DiscoTope is generally the best performing of all methods described here. Finally we present the delineation of epitopes in the malaria protein apical membrane antigen 1 (AMA1) where DiscoTope successfully predicts epitope residues that have been identified using either various experimental or sequence analysis techniques.

# Results

### Properties of discontinuous B-cell epitopes

In order to get a well-established basis for development and evaluation of the prediction method, we compiled a discontinuous epitope data set from 76 X-ray structures of complexes between antibodies and protein antigens. We analyzed the data set to find distributions for the number of residues per epitope, the number of residues per sequential stretch in epitopes, and the longest sequential stretch per epitope. These distributions are shown in Figure 2.3. The total number of residues per epitope was ranging from 9-22 and more than 60% of the epitopes consisted of 14-19 residues (Figure 2.3a). Segments with a single epitope residue represented more than 45% of the 528 segments in the data set (Figure 2.3b). The longest sequential stretch of identified residues per epitope was ranging between 3-12 residues and more than 75% of epitopes comprised a sequential stretch of a maximum length of 4-7 residues (Figure 2.3c). These findings confirm that most epitopes in the data set are indeed discontinuous and composed by small parts of the antigen sequence forming a binding region for the antibody.

The data set was analyzed with respect to surface exposure by determining the number of intra-molecular C$\alpha$ atom contacts for each residue (Figure 2.4). A low contact number correlates with localization close to the surface or in protruding regions of antigen structures. A t-test showed that residues identified as part of

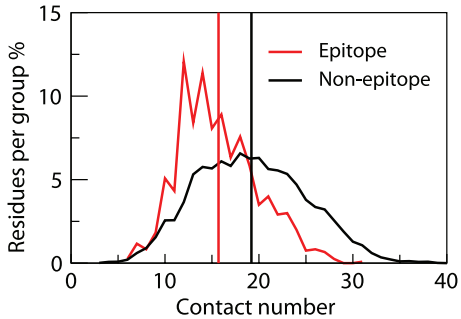


Figure 2.4: **Contact numbers of epitope residues in the dataset compared to non-epitope residues.** The curves show the distribution of contact numbers for epitope residues (red curve) compared to non-epitope residues (black curve). The vertical lines represent the mean value of contact numbers for the epitope residues (red line) and for the non-epitope residues (black line).

epitopes in the data set had significantly lower numbers of contacts compared to the non-epitope residues (P $<$ 10-5). The average number of contacts and standard error of mean for epitope residues was 15.7±0.12 and for non-epitope residues the average contact number and standard error of mean was 19.2±0.05 (see Figure 2.4, vertical lines). The finding that epitopes are in exposed or protruding regions is in agreement with previous analysis of B-cell epitopes (Novotny et al. 1986; Thornton et al. 1986). As shown in Figure 2.4, the two distributions are overlapping. This is most probably caused by the incomplete annotation of the data set or because other factors than contact numbers are important in defining an epitope.
For the development and evaluation of prediction methods, the 76 antigens in the data set were grouped into 25 non-homologous groups (see Materials and Methods for more details). From these 25 groups, 5 sets (of 5 groups each) were constructed and used for 5 fold cross-validated training and evaluation to avoid optimizing and evaluating on similar antigens.

### Log-odds ratios calculated from the epitope data set

We analyzed the statistics of amino acids in epitopes and non-epitopes of the data set by calculation of log-odds ratios from peptides of the data set. A peptide-based approach of similarity reduction was chosen to avoid skewing log-odds ratios toward highly redundant epitopes in the data set. Peptides with high similarity in the data set were weighted lower than peptides with low similarity and therefore the length of the peptides played an important role in the derivation of log-odds ratios. We used raw log-odds ratios as epitope propensities for prediction of epitopes in the training sets and found a peptide length of 9 residues to be optimal.
Table 2.1 shows epitope log-odds ratios calculated from homology-reduced peptides of the total data set of 76 proteins. Of the 20 amino acids, asparagines (N), arginine (R), proline (P) and lysine (K) had the highest log-odds ratios, meaning that they are over-represented in epitopes compared to non-epitopes of the data set. Cysteine (C), alanine (A), leucine (L), valine (V) and phenylalanine

| Amino acid* | Parker | Log-odds ratios |
|---|---|---|
| D | 2.460 | 0.691 |
| E | 1.860 | 0.346 |
| N | 1.640 | 1.242 |
| S | 1.500 | -0.145 |
| Q | 1.370 | 1.082 |
| G | 1.280 | 0.189 |
| K | 1.260 | 1.136 |
| T | 1.150 | -0.233 |
| R | 0.870 | 1.180 |
| P | 0.300 | 1.164 |
| H | 0.300 | 1.098 |
| C | 0.110 | -3.519 |
| A | 0.030 | -1.522 |
| Y | -0.780 | 0.030 |
| V | -1.270 | -1.474 |
| M | -1.410 | 0.273 |
| I | -2.450 | -0.713 |
| F | -2.780 | -1.147 |
| L | -2.870 | -1.836 |
| W | -3.000 | -0.064 |

*Amino acids are listed with descending hydrophilicity using the values of the Parker scale.

Table 2.1: **The Parker hydrophilicity scale and epitope log-odds ratios**

(F) had very low log-odds ratios and are correspondingly under-represented in epitopes. Interestingly, we found several discrepancies between the Parker hydrophilicity scale and the log-odds ratios (Table 2.1). For example, the most hydrophobic residue, tryptophan (W) did not have a particularly low log-odds ratio. The most hydrophilic residues, aspartate (D) and glutamate (E) had relatively moderate log-odds ratios. Arginine (R) and proline (P) had some of the highest log-odds ratios but are ranked close to the middle of the Parker hydrophilicity scale. Cysteine (C) and alanine (A) are ranked close to the middle of the Parker scale but had some of the lowest log-odds ratios.

### Evaluation of uncombined methods for B-cell epitope prediction

To test the predictive strength of contact numbers and the epitope propensity scale of log-odds ratios on discontinuous epitopes, we used the area under receiver operator curves (AUC) averages over different evaluation sets (see details in Materials and Methods). We additionally tested a sequential average of log-odds ratios as prediction score similar to the approach recommended for the hydrophilicity scale by Parker et al. (Parker et al. 1986). The optimal window size for sequential averaging of log-odds ratios was found to be 9 residues based on the predictive performance on the training sets (data not shown). We found that the epitope log-odds ratios used with sequential averaging performed better than the sequentially averaged Parker hydrophilicity scale on the discontinuous epitopes (Figure 2.5a). The raw epitope log-odds propensity scale gave an average performance of 0.604 on the evaluation sets. Smoothing of the log-odds ratios using a sequential average of 9 residues improved the performance

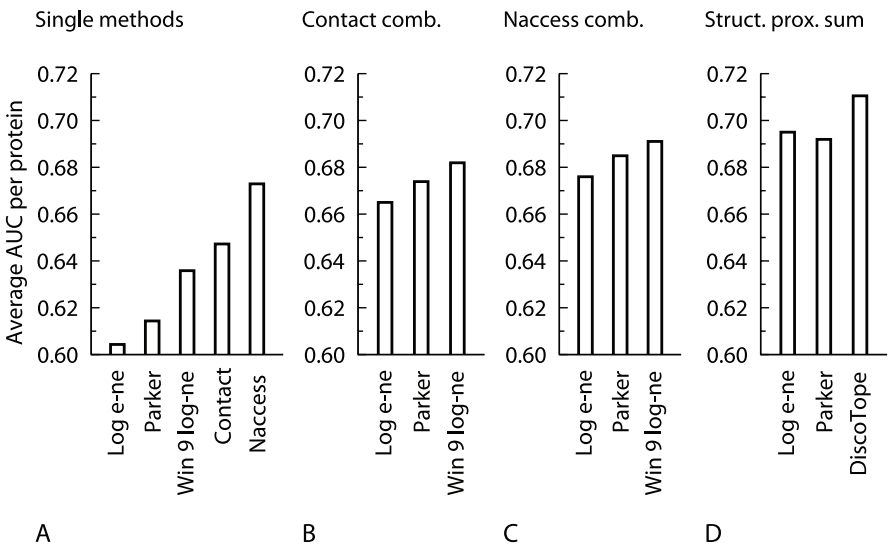


Figure 2.5: **Evaluation of B-cell epitope prediction methods.** The average AUC of various methods on the 5 evaluation sets. Log e-ne denote raw log-odds ratios, Parker denotes the Parker hydrophilicity scale, Win9 log e-ne denotes the log-odds ratios used with a smoothing window of 9 residues, Contact denotes contact numbers and Naccess denotes NACCESS RSA values. (A) The performance of single methods. (B) The performance of simple combination methods using contact numbers. (C) The performance of simple combination methods using NACCESS RSA values. (D) The performance of the structural proximity sum methods.

to 0.636. The Parker scale was used with a smoothing window of 7 residues and had a performance of 0.614. Compared to the methods based on propensity scales, the methods based on contact numbers and NACCESS relative surface area (RSA) values had considerably higher performances of 0.647 and 0.673 respectively (Figure 2.5a).

### Combination methods for epitope prediction

We additionally tested the prediction of epitope residues using surface localization values based on contact numbers or NACCESS RSAs in combination with

| | | Sequential average | | Structural proximity sum | | |
|---|---|---|---|---|---|---|
| | Log-odds | Win 9 log-odds | Parker | Log-odds | Win 9 log-odds | Parker |
| Contacts | 1.5 | 0.25 | 1.75 | -1.5 | -0.5 | -1.5 |
| NACCESS | 2.0 | 0.25 | 1.75 | | | |

Win 9 denotes epitope log-odds ratios smoothed by a window of 9 sequential residues

Table 2.2: **Optimal weights on surface localization scores for combination methods**

epitope log-odds ratios or the Parker hydrophilicity scale. One combination approach was to use a sum of weighted prediction scores from surface localization measures and methods based on sequential information (log-odds ratios or hydrophilicity scores). A second approach was tested by summing log-odds ratios, sequentially averaged log-odds ratios or Parker scale scores of residues in spatial proximity and adding the contact numbers to give a prediction score. For each combination, we estimated the relative weight on the surface localization score by optimizing the predictive performance on the training sets measured in average AUC. The optimized weights are listed in Table 2.2.
The predictive performances of the combination methods were tested by calculating the average AUC from predictions on the evaluation data sets (Figure 2.5b, c, d). Simple linear combinations of the Parker scale, raw log-odds ratios and smoothed log-odds ratios with structure based methods (contact numbers and NACCESS RSA values) in general improved the performance (Figure 2.5b, c). Combination methods using raw log-odds ratios had a performance of 0.665 for the combinations with contact numbers and 0.676 for the combination with NACCESS RSA values. The linear combinations with the Parker method had performances of 0.674 for the contact number combination and 0.685 for the NACCESS RSA combination. Using a combination of smoothed log-odds ratios combined with contact numbers yielded a performance of 0.682. The best performing method of the simple linear combinations was the combination of smoothed log-odds ratios with NACCESS RSAs. This method had a performance of 0.691 on the evaluation sets.
Methods based on a combination of structural proximity sums of propensity scales with contact numbers gave the best performances on the evaluation sets (Figure 2.5d). The performance of the structural proximity sum method based on Parker predictive values combined with contact numbers had a performance of 0.692. The corresponding structural proximity sum method using raw log-odds ratios had a performance of 0.695. The best performing method on the evaluation data sets was the structural proximity sum of sequentially smoothed epitope log-odds ratios combined with contact numbers. This method was shown to have a performance of 0.711, which is significantly better than the method based on structural averaging using raw log-odds ratios ($P = 0.040$). The method is also significantly better than the Parker method ($P = 0.007$) and marginally better than the NACCESS RSA method ($P = 0.105$). We call this method DiscoTope.

### Analysis of the DiscoTope method for discontinuous B-cell epitope prediction

We decided to further analyze the Parker hydrophilicity, NACCESS RSA and DiscoTope predictions to get a more detailed comparison of the performances of the methods. A comparison of the sensitivity of the three methods was done based on a number of selected specificities (Table 2.3). In Table 2.3, we have additionally listed prediction threshold values to facilitate general use of all three methods for B-cell epitope prediction. For all five specificity levels, DiscoTope had the highest sensitivity of the three methods. At a level of 95% specificity (which means only 5% false positive predictions) DiscoTope detected 15% of the epitopes. The Parker method had higher sensitivity than the NACCESS RSA method for the 95% and 90% specificity levels. This is in contrast to the

| | DiscoTope | | NACCESS RSA | | Parker | |
|---|---|---|---|---|---|---|
| Specificity | Sensitivity | T* | Sensitivity | T* | Sensitivity | T* |
| 95% | 15.5% | -3.1 | 8.7% | 88% | 11.0% | 1.07 |
| 90% | 24,2% | -4.7 | 18.9% | 74% | 19.6% | 0.88 |
| 85% | 32,3% | -6.0 | 27,2% | 67% | 27.1% | 0.74 |
| 80% | 40,2% | -6.9 | 37.3% | 60% | 36.0% | 0.60 |
| 75% | 47,3% | -7.7 | 44.1% | 55% | 39.8% | 0.50 |

* Residues with a prediction value above the thresholds (T) were predicted as parts of epitopes.

Table 2.3: **Sensitivity of methods corresponding to a number of selected specificity levels**

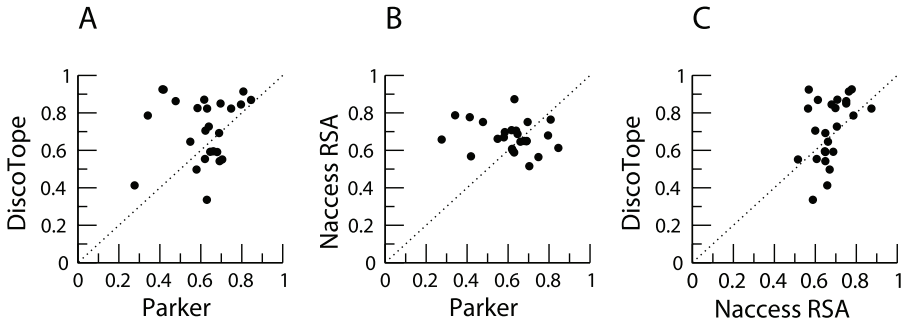


Figure 2.6: **Dot plots showing comparisons of performances of the Parker method, the NACCESS RSA method and the DiscoTope method.** Circles indicate average AUC per group showed for the 25 groups of different antigens. The dotted lines indicate points where the methods perform equally.

averaged AUC value on the 5 evaluation sets, which was found to be higher for the NACCESS method than for the Parker method (Figure 2.5a).

In order to analyze the performances of the three methods on different groups of antigens, we compared prediction AUC values for each of the 25 non-homologous antigen groups (Figure 2.6). For the majority of the groups of antigens in the data set, the DiscoTope method had a better performance than the Parker method (Figure 2.6a). However, in 8 groups of antigens the epitope residues were more accurately predicted using the Parker method. The same tendency was observed for the NACCESS RSA method, where the Parker method performed best for 12 groups (Figure 2.6b). Comparison of the DiscoTope and NACCESS RSA methods showed that, even though the average AUC value for the 25 groups was highest for the DiscoTope method, the NACCESS RSA method performed best for 10 of the antigen groups (Figure 2.6c). We found that the DiscoTope and the NACCESS RSA methods had 6 groups in common for which the Parker method performed best. These groups were represented by the PDB antigen entries 1JPS, 2JEL, 1TQB, 1AR1, 1OAZ, 1EO8. The fact that both surface accessibility based methods had lower performance than the Parker scale method suggests that the measured surface accessibility for single antigen chains is not sufficient for epitope prediction in all types of antigens. Three of the six groups (represented by antigens 1JPS, 1AR1 and 1EO8) contained antigens which have

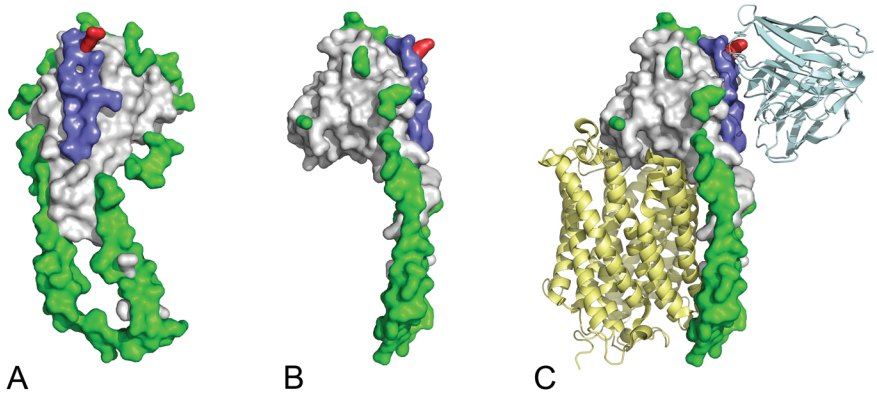


Figure 2.7: **Structure of the 1AR1 antigen.** The antigen is a subunit of the cytochrome c oxidase (Ostermeier et al. 1997). (A) The 30% of residues with lowest contact numbers are shown in green. In red is shown a residue which is part of the 30% with lowest contact numbers and of the epitope from the data set. The rest of the epitope is shown in blue. (B) The structure mentioned above and rotated 90 degrees. (C) The complex of the cytochrome c oxidase with antibody fragments. The 1AR1 antigen is color coded as in (A) and (B). Antibody fragments are shown in light blue. The other subunit of the cytochrome c oxidase is shown in yellow. Membrane spanning helices of the 1AR1 antigen are part of the lower half of the structure.

elongated structures. Furthermore, antigens of 1JPS, 1AR1 and 1EO8 are all known as subunits of larger biological complexes associated with membranes (Ostermeier et al. 1997; Fleury et al. 2000; Faelber et al. 2001). Perhaps not surprising, the single antigen chain approach taken by the DiscoTope method clearly could not correctly measure the surface accessibility of all residues in such proteins. In Figure 2.7 is shown the structure of the antigen of 1AR1 as an example. On the plot, most of the residues in the antigen that had the lowest contact numbers are not in proximity of the epitope (Figure 2.7a and b). In fact, only one residue of the epitope was among the 30% residues in the antigen with lowest contact numbers. The antigen of 1AR1 is a subunit of a membrane spanning cytocrome *c* oxidase (Figure 2.7c) and the largest continuous region of residues with low contact numbers corresponds to a region of the protein which is described as membrane-spanning (Ostermeier et al. 1997).

### Prediction of B-cell epitope residues in Apical Membrane antigen 1

To evaluate our method on B-cell epitopes, which are mapped using other types of methods than X-ray crystallography, we tested the predictions of DiscoTope on the structure of the ectodomain from AMA1 (Bai et al. 2005; Pizarro et al. 2005). No AMA1 epitopes are included in the data set of discontinuous epitopes derived from the PDB. However, two separate epitopes recognized by monoclonal antibodies Mab1F9 and Mab4G2 have been experimentally mapped on the AMA1 ectodomain: The Mab1F9 epitope was mapped using phage-display of peptides and point mutations of E197 (Coley et al. 2006). The discontinuous Mab4G2 epitope was mapped in detail by point mutation of 9 residues (Pizarro et al. 2005). In addition, Bai and coauthors have classified 5 residues (including

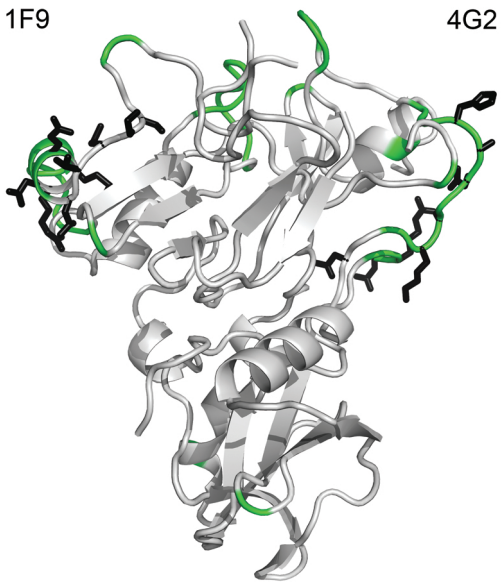


Figure 2.8: **Predicted epitope residues of the AMA1 ectodomain.** Backbone atoms of residues predicted by DiscoTope as parts of epitopes are highlighted in green. Side chains of the residues mapped to the monoclonal antibodies 1F9 and 4G2 are shown in black.

E197 and other residues in the same region of the structure) as highly polymorphic in *Plasmodium falciparum* AMA1 sequences. It has been suggested that the polymorphism is caused by selection pressure on the antigen to avoid the host immune system (Bai et al. 2005). We used a DiscoTope prediction threshold of -4.7 which corresponds to a specificity of 90% and 24% sensitivity (Table 2.3). In AMA1, 43 of 311 residues were predicted as epitope residues. Most of the predicted epitope residues cluster in three separate regions of the AMA1 structure (Figure 2.8). DiscoTope successfully identified 2 of the 8 residues in the 1F9 epitope which were mapped using phage-display (D196 and E197). In the discontinuous 4G2 epitope, all 9 residues except D348 were predicted to be part of epitopes. All of the 5 highly polymorphic residues described by Bai et al. were predicted to be located in epitopes. Thus, DiscoTope successfully predicted epitope residues of AMA1 which have been mapped by using diverse methods.

## Discussion

In this paper we have presented DiscoTope, a novel method for prediction of residues located in discontinuous B-cell epitopes. DiscoTope combines surface localization and spatial properties of a protein structure with a novel epitope propensity scale. The combination is defined in terms of a simple weighted sum of the contact number and a sum of sequentially averaged epitope log-odds ratios of spatially proximate residues. We propose to use DiscoTope for prediction of discontinuous epitope residues for several reasons: First, we have shown on

a data set of discontinuous epitopes that the average predictive performance of the DiscoTope is significantly higher than the Parker propensity scale and marginally higher than the surface localization score defined by the NACCESS RSA score. Second, we have shown that DiscoTope correctly predicts residues in epitopes which have been identified using different techniques such as phage-display, point mutation and sequence analysis. Third, the DiscoTope prediction method is publicly available on www.cbs.dtu.dk/services/DiscoTope and the output of the method is easily interpreted.

The Parker hydrophilicity scale is often used for prediction of linear B-cell epitopes by smoothing values in a 7 residue window (Parker et al. 1986). Compared to the epitope log-odds ratios smoothed over a window of 9 residues developed here, the Parker scale was not as accurate for prediction of discontinuous B-cell epitopes in the data set. The difference in ranking between the two scales suggests that our log-odds ratios represent more characteristics of the epitopes than only hydrophilicity. Possibly, this difference contributes to a better predictive performance on the data set since combinations of various propensity scales including hydrophilicity, flexibility, accessibility and beta-turn prediction are better than single propensity scales for epitope prediction (Pellequer et al. 1991). Our finding, that surface accessibility values improved the prediction of residues in B-cell epitopes, is in agreement with recently reported results by Batori et al. (Batori et al. 2006). In addition, the combination of propensity scale methods with structural information improved the performance considerably. This suggests that both accessibility and chemical characteristics are important in descriptors of discontinuous B-cell epitopes. Combination methods using a number of propensity scales have been used for B-cell epitopes for more than 15 years (Pellequer et al. 1991). However, DiscoTope is the first reported method combining a propensity scale with 3D structural information, such as spatial proximity.

Van Regenmortel (Van Regenmortel 1996) has addressed the problem of using protein sequences for prediction of B-cell epitopes which are in reality multi-dimensional. He concluded that more input data such as the antigen three dimensional structure is needed for accurate prediction. The requirement of structural input for B-cell epitope prediction is a limiting factor for the general use of the method. However, structural genomics projects help to increase the number of X-ray crystallography structures determined of proteins in general, and to cover larger areas of the structure space. Therefore, the requirement of protein structures as input for prediction methods will become a decreasing problem, because more structures will be determined and better homology models can be obtained.

In general, methods based on structural information were shown to predict residues in discontinuous B-cell epitopes with a higher performance measured in average AUC, than propensity scale methods which only used sequential information. In all methods of evaluation the DiscoTope method was shown to have the highest performance. However, we found that the Parker hydrophilicity scale had a higher sensitivity than the NACCESS RSA method on the 95% and 90% specificity levels. These results illustrate the importance of using other measures of performance for evaluation in addition to the AUC.

We found that for antigen groups that contain antigens part of larger biological complexes, the performances of both the NACCESS RSA method and the DiscoTope method were relatively low. The low performances were due to an

incorrect measure of surface accessibility of regions which are part of protein-protein interaction sites or embedded in a membrane. Therefore, we believe that the outcome of prediction methods for B-cell epitopes should be combined with additional information about properties such as biological complex formation, membrane interaction and glycosylation.

The accuracy of the described methods for B-cell epitope prediction was still relatively moderate. This may partly be caused by the incomplete identification of epitopes in the antigens of the data set. If the methods correctly predicted an epitope that was not bound by the antibody in the corresponding complex PDB file it counted as a false positive. However, since the same data set was used in the evaluation of all methods described here, we assumed that incomplete identification had the same influence on the predictive performance of all methods, and hence negligible influence on their relative ranking. The predictive performance of the method developed by Batori et al. (Batori et al. 2006) was evaluated on 6 epitopes of one single antigen. This evaluation approach, using an antigen where all epitopes are more completely identified, possibly had the effect that the false positive proportion was lower and the measured performance was higher. In our approach of evaluation we chose to include as much variation as possible and thereby avoid biasing the method towards a certain type of antigens or epitopes. However, a future evaluation of our DiscoTope method using a data set of antigens with more completely identified epitopes would be of interest.

Recently, Schlessinger et al. have developed a sophisticated method for identification of epitopes in antibody/antigen complex structures (Schlessinger et al. 2006). The method is based on an analysis and identification of complementarity determining regions (CDRs) of the antibody and a subsequent identification of epitopes by mapping residues in the antigen in proximity to CDRs. The identification described in this paper was simply based on antigen residues in proximity to antibody residues in general and it is plausible that a future application of the identification method developed by Schlessinger et al. could improve the DiscoTope method.

Because of their non-linearity, discontinuous epitopes pose other problems than linear epitopes in vaccine design. Not only must the new vaccine contain the amino acids or atoms that are necessary for binding and eliciting specific antibodies, but a conservation of the correct spatial conformation is also needed. DiscoTope can predict residues that are likely to be part of discontinuous epitopes. Subsequently, antibody binding studies and site-directed mutagenesis may help to group predicted epitope residues into epitopes and validate binding. Analysis of the local conformations of epitope residues in the antigen structure may also aid the design of vaccines, because a vaccine based on a discontinuous epitope must have these conformations preserved. The preservation may be obtained using native proteins, sub-domains of a protein, redesigned proteins carrying the epitope or mimotope peptides in vaccines. Therefore we consider discontinuous epitopes useful for rational vaccine design.

# Materials and methods

### Preparation of the data set

A list of experimentally determined protein antigen-antibody structures was obtained from the SACS database of antibody crystal structure information (Allcorn and Martin 2002). The list was filtered to only contain structures determined to a resolution $< 3$ Å with protein antigens of more than 25 amino acids. Coordinate files corresponding to the filtered list were downloaded from the Protein Data Bank (PDB, http://www.rcsb.org/pdb). The final data set contained 76 complexes of antibody-antigen pairs. Epitope residues in the data set were defined as antigen amino acids having atoms within a 4 Å distance from antibody atoms. Comparisons based on a subset of five identified epitopes with residues reported as antibody interacting (Padlan et al. 1989; Muller et al. 1998; Fleury et al. 2000; Mirza et al. 2000; Romijn et al. 2003) showed that a distance threshold of 4 Å gave an annotation corresponding well to that made by human experts. (92% of the epitope residues were correctly identified and only 1% of the non-epitope residues were identified as epitope residues). Only a single epitope was represented in each PDB file. All other epitopes that might exist in a given antigen were treated as non-epitopes in our analysis. Certain antigens were represented multiple times in the data set (29 antigens are variants of Lysozyme). Therefore we grouped the data set according to antigen homology. Homology in the data set of 76 proteins was determined using a BLAST search (Altschul et al. 1997) with the BLOSUM80 matrix against all other antigens in the data set combined with a homology threshold as described by Lund et al. (Lund et al. 1997). Antigens were then split into 25 groups with low homology between the groups (BLAST E-values $>0.30$ between groups). The data set annotations and the groups of antigens are publicly available at http://www.cbs.dtu.dk/suppl/immunology/DiscoTope. Finally, the 25 non-homologous groups of antigens were divided into 5 data sets used for cross-validated training and evaluation.

### Use of the Parker hydrophilicity scale

The average Parker scale value over a window of 7 residues was used for the per-residue epitope prediction value as proposed by Parker et al. (Parker et al. 1986).

### Definition of surface residues

A combined measure of amino acid surface localization and structural protrusion was obtained by using residue contact numbers. The residue contact number is the number of C$\alpha$ atoms in the antigen within a distance of 10 Å of the residue C$\alpha$ atom (Nishikawa and Ooi 1980). For a more direct measure of residue solvent accessibility, the relative solvent accessible surface area per residue was calculated for antigen chains extracted from each PDB file using the NACCESS program (Hubbard and Thornton 1993). NACCESS default options were used with a probe radius of 1.4 Å.

### Performance measures

The area under a receiver operator characteristics curve (AUC) (Swets 1988) was used as performance measure. A receiver operator characteristics curve is constructed by varying the prediction threshold and plot the false-positive proportion, or 1-specificity, on the x-axis against the true positive proportion, or sensitivity, on the y-axis (Swets 1988; Lund et al. 2005). We calculate the AUC on a per protein basis. This ensures that a prediction where all residues in a protein are predicted as only epitopes or only non-epitopes has an AUC of 0.5 corresponding to a random prediction. The performance of each method was measured as the average AUC, average specificity and average sensitivity for the 25 antigen groups.

### Statistical analysis

Mean values of contact numbers for epitope residues and non-epitope residues were analyzed using a double sided t-test (standard deviation = 0.121, n = 1202 for epitope residues and standard deviation = 0.050, n = 13242 for non-epitope residues.) A bootstrapping approach was used for pair-wise comparisons of the average AUC values to determine the significance of the performances (Efron and Tibshirani 1993). For each method, the 25 values of average AUC value pr. antigen group were re-sampled 100,000 times in order to obtain a robust estimate of the P-values.

### Derivation of epitope log-odds ratios

Four of the five data sets (the training sets) were used for derivation of epitope log-odds ratios. A series of peptides were produced by sliding an odd-sized window through the sequences of antigens in the training sets. The peptides were then sorted into an epitope group and a non-epitope group depending on the identification of the residue in middle position as epitope residue, or as non-epitope residue. Weight matrices were calculated from the peptides in each group using the method described by Nielsen et al. (Nielsen et al. 2004), including sequence clustering, sequence weighting and pseudo counts with a weight of 200. Finally the log-odds ratios at the central matrix position for each of the 20 amino acids in epitope group relative to non-epitope group were calculated in half bits and used as an epitope propensity scale.

### Using log-odds ratios for epitope prediction

For prediction of epitope residues, the raw log-odds ratios were used alone or in combination with a smoothing window calculating the sequential average of the epitope propensity scale values. The optimal length of peptides used for the derivation of log-odds ratios and the optimal size of the smoothing window were determined with respect to the predictive performance on the training sets used for calculating the log-odds ratios. The performance reported is the five-fold cross-validated performance on the data set. This reduces the risk of overestimating the performance, since the calculation of the log-odds ratios and optimization of other parameters, such as the peptide length and the smoothing window size, are estimated on the training set and hence not biased by the evaluation set data.

### Simple combinations of propensity scales with structure based methods

Contact numbers, NACCESS RSAs and Parker hydrophilicity values were normalized by subtracting the mean and dividing with the standard deviation. The normalized contact numbers were multiplied by -1 in order to correlate high values with surface localization. Subsequently, the different propensity scales were combined with contact numbers or NACCESS RSAs using a linear combination with a weight on the surface measure ranging from 0.001-100. Optimal weights were determined using the training sets. Finally, the performance was evaluated on evaluation sets.

### Structural proximity sum of epitope log-odds ratios

Alternatively, the epitope log-odds ratios or the Parker hydrophilicity scale were used by summing values for all residues with C$\alpha$ atoms within a 10Å distance of each residue. We tested a number of weighting schemes for the proximity sums, for instance based on the distance to the central residue, the contact number for the residue, and a combination of the two. However, the simple approach where all residues carry equal weight gave the highest performance on the training sets (data not shown).

### Prediction of epitopes in AMA1

Chain A of the AMA1 ectodomain from *P. falciparum* (PDB code 1Z40) was used for DiscoTope epitope prediction. We chose to use 1Z40 instead of a full-length AMA1 ectodomain structure (1W8K) because the main part of the residues in the 4G2 epitope was not observed in the latter. Residues 348, 351, 352, 354-356, 385 and 388-389 were counted as residues in the 4G2 epitope (Pizarro et al. 2005), residues 191-199 were counted as part of the 1F9 epitope (Coley et al. 2006) and residues 187, 197, 200, 230 and 243 were counted as highly polymorphic residues (Bai et al. 2005).

## Acknowledgements

This work was in part funded by NIH Contract No. HHSN266200400083C. The authors acknowledge Claus Lundegaard for helpful comments.

Chapter 3

# Identifications of B-cell epitopes in the *Plasmodium falciparum* protein VAR2CSA

This chapter is focused on research in B-cell epitopes with the final goal of developing a pregnancy-associated malaria (PAM) vaccine. First I will give a short introduction to malaria, PAM and to the vaccine-candidate; the *Plasmodium falciparum* protein VAR2CSA. Subsequently, is given an introduction to the methods used in this project: prediction of protein structure, affinity purification, and depletion of antibodies combined with pepscan analysis. Following the introduction, two papers present my work on the project, and finally the chapter is ended by a short discussion of the results obtained in papers II and III.

## 3.1 Pregnancy-associated malaria

Malaria is a major threat to public health in tropical and subtropical regions of the world. The World Health Organization has estimated the occurrence of 300-500 million cases of malaria per year, and more than 1 million deaths from malaria annually (Korenromp et al. 2005). The disease is caused by infection with parasites of the genus *Plasmodium*, which are transmitted by the *Anopheles* mosquitoes. The species *P. falciparum* is the most lethal of the four species which infect humans (Gardner et al. 2002). Several antimalarial drugs have been developed, but the eradication of the disease has failed. One reason for the failure is the development of drug resistance in *Plasmodium* species. Additionally, because malaria is a disease of poverty, the availability of affordable drugs play a major role (Kooij et al. 2006; Menendez et al. 2007). Several approaches have been taken to develop malaria vaccines, but no effective malaria vaccine has been developed yet (Epstein et al. 2007). However, the fact that immunity to the disease can be acquired after natural infection suggests that a malaria vaccine is attainable.
The *P. falciparum* parasite has a complex life cycle, and goes through many

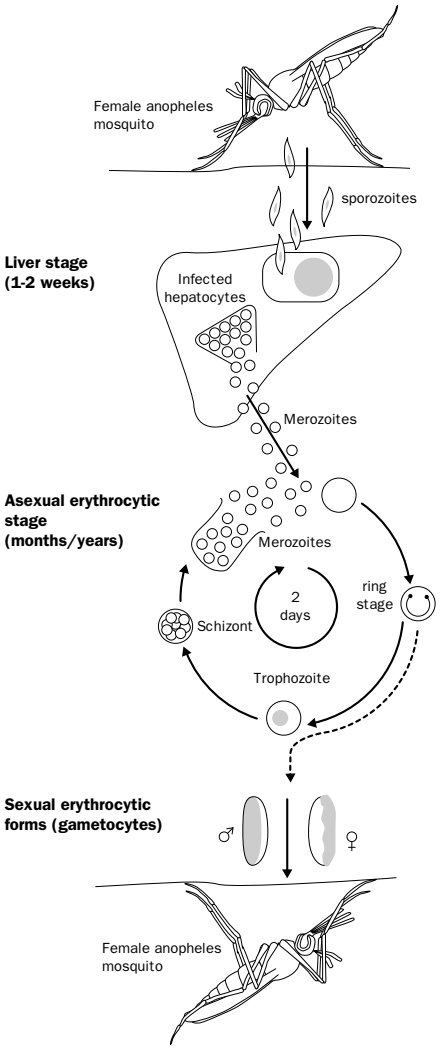


Figure 3.1: **Stages of *Plasmodium* development in humans.** The figure is adapted from Kwiakowski et. al. (Kwiatkowski and Marsh 1997)

stages of development (Kooij et al. 2006). It has been suggested that a malaria vaccine should target parasites in several different stages of development to be effective (Heppner et al. 2005). Here is summary of the different stages (see Figure 3.1). In the first stage, sporozoites enter the bloodstream of a human with the saliva of a feeding mosquito. Then the sporozoites infect liver cells (hepatocytes), grow and replicate to produce many parasites of the next developmental stage called merozoites. The merozoites invade red blood cells (erythocytes), go through the ring, trophozoite and schizont stages and reproduce new merozoites. These blood-stages are asexual and the major cause of the pathology of malaria. When infected erythrocytes burst to release a number of newly developed merozoites, a number of toxins are released with the parasites and cause fever, which is one of the most characteristic symptoms of malaria. Additionally, the destruction of erythocytes can cause anaemia, another symptom of the disease. A few merozoites develop into female or male gametocytes, and their development is arrested until reactivation in the midgut of a female *Anopheles* mosquito, which has been feeding on the infected human. Upon reactivation of development in the mosquito, the gametocytes first develop into gametes that are fertilized. These fertilized gametes develop into ookinetes, which traverses the midgut wall and forms oocysts in the basolateral lamina. In these oocysts, new sporozoites are developed and these migrate into the salivary glands of the mosquitoes, ready for infection of a new victim (Kwiatkowski and Marsh 1997; Kooij et al. 2006).

The spleen is a key site for removal of parasitized erythrocytes (Engwerda et al. 2005). However, the species *P. falciparum* has developed mechanisms for adhesion of infected erythrocytes to human endothelium (sequestration), which prevent the spleen-dependent clearance (Miller et al. 2002). This sequestration contributes greatly to the severity of the disease. For instance, sequestration of infected erythrocytes in capillaries of the brain results in cerebral malaria, which is a life-threatening condition. Additionally, another form of malaria, PAM, is critical in pregnant women. PAM is caused by the sequestration of infected erythrocytes in the placenta. This results in a high risk of still birth, low infant birth weight, and severe maternal anaemia. In sub-Saharan Africa, where *P. falciparum* is widely spread, and the transmission of infection is stable, approximately one in four pregnant women have evidence of placental infection at the time of delivery. Malaria-associated low birth weight is estimated to cause the deaths of 100,000 infants in Africa every year (Desai et al. 2007). Thus, a PAM vaccine obstructing the placental sequestration could have a high impact on the effects on *P. falciparum* infections, even if it would not directly prevent infection. Since women in areas of stable transmission develop immunity to the disease after a few pregnancies, a PAM vaccine would be most effective if given to young women who have never been pregnant.

The adhesion properties are determined by the *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) family of proteins. These proteins are expressed on the surface of the infected erythrocytes and are encoded by approximately 50-60 different *var* genes (Kraemer and Smith 2006). The proteins of the PfEMP1 family are capable of binding a variety of different receptors on the surface of host endothelial cells. The extracellular part of PfEMP1 proteins have similar compositions: all contain Duffy-binding like (DBL) domains and most contain inter-domains called cysteine-rich domains (CIDR). Despite this similarity, the sequence variability between different PfEMP1s and DBL domains within the

same protein is extremely high (Smith et al. 2000). The sequestration of infected erythrocytes to placenta during pregnancy is caused by adhesion to chondroitin sulphate A (CSA) (Fried and Duffy 1996; Pouvelle et al. 1998; Scherf et al. 2001). Women in high transmission areas develop increasing resistance to PAM over successive pregnancies (McGregor et al. 1983), and it has been shown that maternal antibodies, which inhibit the binding to CSA, are associated with a reduction in the symptoms of PAM (Duffy and Fried 2003; Fried et al. 1998). VAR2CSA is the major protein of the PfEMP1 family, which is thought to mediate this adhesion. Accordingly, VAR2CSA adhesion mechanisms and interactions with the human immune system are major research topics in PAM vaccine development projects (Oleinikov et al. 2007; Rogerson et al. 2007).
Recently, the first two structures of DBL domains from different *Plasmodium* proteins were published (Tolia et al. 2005; Singh et al. 2006). The structure of the *P. falciparum* erythrocyte binding antigen (EBA)-175 DBL domains F1 and F2 revealed that the two DBL domains were very similar in structure, despite sequence variations. Additionally the structure of the *Plasmodium knowlesi* $\alpha$-DBL domain showed that this DBL domain has a similar structure to the EBA-175 DBL domains, even though the sequence variation between the domains is high and functional residues were found to be located on the opposite sides of the domain structures (Singh et al. 2006). The finding that DBL domains from different proteins and species of *Plasmodium* share structural homology strengthens the hypothesis that DBL domains from other proteins can be predicted by comparative modeling using the determined DBL domains as templates. This modeling, and the interpretations of experimental data it allows for, can potentially lead to new insight into the mechanisms of *Plasmodium* infection and malaria disease (Tolia et al. 2005; Howell et al. 2006). Because of that, we found it suitable to use comparative modeling in our research of VAR2CSA and PAM.

## 3.2 Protein structure prediction

Knowledge of three-dimensional protein structure is essential for understanding many biological questions. Much effort has been put into solving the 3D structures of proteins, and the number of determined structures in the Protein Data Bank (PDB) has been remarkably increased during the last 20 years. However, the number of sequenced genes is increasing even more rapidly, and protein structure determination is still a time-consuming process. For some proteins, the structure determination is still not within reach despite the progress in improving technical methods. Protein structure prediction is a relatively fast alternative to structure determination. A general observation in structure research, is that evolutionary linked sequences often have similar 3D structures (Chothia and Lesk 1986); and this is the basic assumption of protein structure prediction by comparative modeling. In comparative modeling, a model is built from target protein sequence by using related proteins of known structure (templates).
A comparative modeling process normally consists of the following steps (Dunbrack 2006; Fiser and Sali 2003):

- Finding suitable template proteins which are related to the target protein sequence

- Aligning the target protein sequence to the sequences of templates
- Building coordinates of the 3D model
- Assessing the accuracy of the model

The correct identification of templates, and the alignment of target sequences to template structures are the primary determinants of the quality of the final model (Venclovas et al. 2003). In general, alignments are often close to optimal for sequence identities higher than 30%, but for sequence identities below 30%, the general alignment quality is decreasing greatly and as many as 50% of the residues may be misaligned when the sequence identity is below 20% (Kryshtafovych et al. 2005). Therefore, it is important to select methods for comparative modeling carefully, when modeling difficult targets with low sequence identity between target and template. For the comparative modeling of protein structures presented in this chapter, the HHpred server has been used (Soding 2005; Soding et al. 2005). The server has been chosen over other methods because it is fast, user-friendly and optimized for remote homology modeling. This makes it very suitable for the modeling of VAR2CSA domains. Here is described an introduction to the methods used by the HHpred server.

### 3.2.1 Template finding and alignment

HHpred is based on three methods: sequence profiles, HMM-comparisons and secondary structure prediction. These methods have been shown to greatly enhance the performance of comparative modeling for proteins which are remotely related to available target sequences (Moult 2005). Firstly, these methods have improved the detection of homology to identify a higher number of suitable templates. Secondly, the alignments between target protein sequence and template structures have been improved.
Sequence profiles are based on iterative sequence searches and can be built using the PSI-BLAST program (Altschul et al. 1997). Their main strength is that they capture both sequence variation and conservation in positions of related protein sequences. This can be used to identify farther related proteins with similar structures, but different sequences. In the HHpred prediction method, sequence profiles are build for both target protein sequences and template sequences, and these are compared to detect remote homologous proteins and improve alignments (Soding 2005).
HMMs have additionally improved the comparative modeling performance greatly. HMMs integrate probabilities of gaps and insertions at positions in the target protein sequence compared to templates (Dunbrack 2006). This helps to identify correct template structures, because insertions and deletions tend to locate at specific positions in profiles build from proteins of the same fold (Chothia and Lesk 1986).
Secondary structure prediction is also used in the HHpred method to improve the comparative modeling. The known and predicted template secondary structure is compared to secondary structure predictions of the target sequence. Because the performance of secondary structure prediction methods is relatively high, this information of local structure improves the identification of templates and alignments between target and template, for instance by shifting insertions in the target sequence to loop areas of the template structure.

The HHpred server first identifies template structures and generates alignments of target sequences to template structures. These alignments are then used as input for the MODELLER program (Sali and Blundell 1993), which is building 3D models based on the structures of the templates.

### 3.2.2 Building 3D coordinates

The MODELLER program is widely used for this purpose in protein prediction protocols from many groups (Dunbrack 2006). Comparative modeling using MODELLER is based on the extraction of a number of C$\alpha$-C$\alpha$ distance restraints and dihedral angle restraints (both sidechain and mainchain) from the template structure. In addition, the modeling method uses stereochemical restraints (i.e. chemical bonds, bond angles etc.) from the CHARMM molecular mechanics force field (MacKerell et al. 1998), and a potential based on general statistical preferences of amino acids from a data set of protein structures. The different restraints are combined to result in an objective energy-function that is optimized using molecular dynamics; in this way, 3D structure coordinates are found which minimizes the energy of the model (Sali and Blundell 1993).
After the modeling of residues aligned to template structures, the insertions are modeled using a loop modeling procedure. In general, loops are much more difficult to model correctly than core protein structure, and the difficulty increases with the loop length (Fiser et al. 2000). In loop regions, the number of restraints obtained from template structures is low, and the prediction is mainly based on energy functions derived from statistical energy potentials. Current energy potentials do not describe protein structures accurately enough; in this way loop modeling is closely related to the problem of *ab initio* protein structure prediction (Moult 2005). In the MODELLER loop modeling protocol the rest of the model is kept unchanged, while the loops (insertions) are optimized. Like in the protocol for modeling the general structure, the combination energy function is optimized using molecular dynamics to get the loop conformations with the lowest energy (Fiser et al. 2000).

### 3.2.3 Assessing model accuracy

The HHpred server has integrated model assessment and two of the integrated methods are Verify3D and ANOLEA. These two methods for evaluation of models are based on general statistical preferences for amino acids in protein structures.
Verify3D is based on the preferences for each amino acid to be in specific environments (Luthy et al. 1992). In the Verify3D method, 18 different environments are defined on the basis of solvent exposure, the degree of polarity of the amino acid sidechain and the secondary structure. The statistical probability for each type of amino acid to be found in each class is pre-calculated from determined protein structures, and these probabilities are used to score the amino acids in the models (Bowie et al. 1991).
The ANOLEA method is based on non-local interactions of residues which are separated by more than 11 residues in the protein sequence (Melo and Feytmans 1998). Native, globular proteins have a high number of non-local interactions, for instance between residues packing in the hydrophobic core of the protein. Models of protein structures may not have the same degree of packing, and

ANOLEA can be used to detect this lack of packing. In the ANOLEA method, any given atom in a residue is scored by encountering all heavy atoms within a distance of 7 Å and farther than 11 residues in the sequence (Melo and Feytmans 1998). The heavy atoms are divided into 40 classes dependent on the connectivity, chemical nature and location in sidechain or mainchain. Interaction probabilities for the different classes derived from structures of native proteins are used to score observed interactions in the models (Melo and Feytmans 1997). A surface exposure term based on the preferred total numbers of heavy atoms within 7 Å from the atom is further included in the scoring (Melo and Feytmans 1998).

## 3.3 Antibody affinity purification and depletion studies

As mentioned in section 1.6.2, pepscan analysis is a useful technique for identification of peptide fragments which cross-react with antiprotein antibodies. In the following two papers presented here, we have used pepscan analysis for identifying epitopes of antibodies binding the *P. falciparum* VAR2CSA protein. In our approach, we have combined pepscan analysis with antibody depletion and affinity purification techniques. These are useful for identification of surface exposed regions, and for detection of cross-reactivity between variants of a protein from different lines.

In the first paper presented in this chapter (see section 3.5), we have used affinity purification studies based on the heterologously expressed VAR2CSA DBL3X domain. Here, antibodies binding the heterologously expressed DBL3X domain are purified and analyzed by pepscan; antibodies which remain after the purification are assumed to bind on the surface of the folded DBL3X domain as opposed to non-surface exposed linear epitopes. However, because only the heterologously expressed DBL3X domain is used, it is possible that the reactive regions identified using this approach are not surface exposed in the native VAR2CSA protein.

Depletion studies were carried out by incubating human serum pools with infected erythrocytes expressing VAR2CSA. In infected erythrocytes, the whole VAR2CSA protein is expressed in an environment that is considered to be close to the native environment during a natural infection. Thus, the peaks of reactive antibodies, which bind surface-exposed epitopes of native-like VAR2CSA, are diminished after incubation. In the depletion studies presented in the second paper (see section 3.6), we have used erythrocytes infected with two different lines of *P. falciparum*, 3d7 and FCR3, to investigate if naturally acquired antibodies bind VAR2CSA regions of both lines.

Finally, we have analyzed the binding of rabbit antibodies raised against the VAR2CSA protein. Epitopes of rabbit antibodies may not be directly relevant for vaccine design, but are used here to map which regions are surface exposed in the VAR2CSA protein or the DBL3X domain alone.

## 3.4 Contributions to the papers

The following papers present the results of my work on the VAR2CSA PAM-vaccine project. In the first paper of this chapter, "Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in *P. falciparum* placental sequestration", my work consisted of prediction of the DBL3X structure, analysis of the structure model, mapping of identified B-cell epitopes and sequence variance analysis results on the structure model, preparation of figures including the structure model and writing part of the manuscript. In the second paper of this chapter, "Structural insight into epitopes in the pregnancy-associated malaria protein", my work included prediction of structure models, analysis of these models, analysis of the pepscan data, mapping of identified B-cell epitopes on the structure models, preparation of all figures and writing the main part of the manuscript.

## 3.5 Paper II

*PLOS Pathogens, November 2006, 2,11:1069-1080,e124*

# Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in *P. falciparum* placental sequestration

**Madeleine Dahlbäck*[1], Thomas S Rask*[2], Pernille H Andersen[2], Morten A Nielsen[1], Nicaise T Ndam[1,3], Mafalda Resende[1], Louise Turner[1], Philippe Deloron[3], Lars Hviid[1], Ole Lund[2], Anders Gorm Pedersen[2], Thor G Theander[1] and Ali Salanti[1]**

[1]Centre for Medical Parasitology at University of Copenhagen and Copenhagen University Hospital (Rigshospitalet), Denmark. [2]Center for Biological Sequence Analysis, BioCentrum-DTU, Denmark. [3]Institut de Recherche pour le Développement (IRD), Faculté de Pharmacie, France.

*These authors contributed equally to this work.

* *Corresponding author:*

Email: salanti@cmp.dk

**Pregnancy-associated malaria (PAM) is a major health problem, which mainly affects primigravidae living in malaria endemic areas. The syndrome is precipitated by accumulation of infected erythrocytes in placental tissue through an interaction between chondroitin sulphate A (CSA) on syncytiotrophoblasts and a parasite encoded protein on the surface of infected erythrocytes, believed to be VAR2CSA. VAR2CSA is a polymorphic protein of approximately 3000 amino acids forming six Duffy-binding-like (DBL) domains. For vaccine development it is important to define the antigenic targets for protective antibodies and to characterize the consequences of sequence variation. In this study, we used a combination of in silico tools, peptide arrays and structural modeling to show that sequence variation mainly occurs in regions under strong diversifying selection, predicted to form flexible loops. These regions are the main targets of naturally acquired IgG and accessible for antibodies reacting with native VAR2CSA on infected erythrocytes. Interestingly, surface reactive anti-VAR2CSA antibodies also target a conserved DBL3X region predicted to form an $\alpha$-helix. Finally, we could identify DBL3X sequence motifs that were more likely to occur in parasites isolated from primi- and multigravidae, respectively. These findings strengthen the vaccine candidacy of VAR2CSA and will be important for choosing epitopes and variants of DBL3X to be included in a vaccine protecting women against PAM.**

**Keywords:** VAR2CSA, Malaria, Pregnancy, Vaccine, PfEMP1, DBL, Positive selection, Recombination, PepScan, Epitope.

**Synopsis**

Pregnancy-associated malaria caused by *Plasmodium falciparum* is characterized by the accumulation of parasite-infected red blood cells in the placenta and is a major health problem in Africa. VAR2CSA is a parasite protein expressed on the surface of malaria-infected red blood cells and mediates the binding to the placental receptor, chondroitin sulphate A. It is believed that a vaccine based on VAR2CSA will protect pregnant women against the adverse effects of pregnancy-associated malaria. However, due to the size and polymorphism of VAR2CSA it is required to define smaller regions that can be included in a vaccine and to analyze the degree and consequences of sequence variation to ensure a broadly protective immune response. The authors have characterized the chondroitin sulphate A-binding DBL3X domain of VAR2CSA with regard to epitopes targeted by naturally acquired antibodies and the influence of sequence variation by bioinformatics and experimental data based on a VAR2CSA peptide array. They identify both variable and conserved surface-exposed epitopes that are targets of naturally acquired immunoglobulin gamma in pregnant women with placental malaria. These findings will be imperative for choosing epitopes and variants of DBL3X to be included in a vaccine protecting pregnant women against malaria.

## Introduction

Individuals living in areas with high *P. falciparum* transmission acquire immunity to malaria over time and adults have markedly reduced risk of getting severe disease (Christophers 1924). Pregnant women constitute an important exception to this rule and this has severe consequences for both mother and child (Brabin 1983). Pregnancy-associated malaria (PAM) is characterized by selective accumulation of *P. falciparum* in the intervillous blood spaces of the placenta (Walter et al. 1982; Bray and Anderson 1979). The main pathophysiological consequences of PAM are delivery of low birth weight (LBW) babies and maternal anaemia (McGregor et al. 1983). In areas of high parasite transmission PAM affects mainly primigravidae as immunity is acquired as a function of parity (Brabin 1983). Parasite sequestration in the placenta is mediated by an interaction between chondroitin sulphate proteoglycans (CSA) on the syncytiotrophoblasts and proteins expressed on the surface of infected erythrocytes (Fried and Duffy 1996). VAR2CSA, a single and uniquely structured molecule belonging to the *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) family, is currently believed to be the main parasite ligand for placental binding (Salanti et al. 2003). *Var2csa* is markedly up-regulated in *P. falciparum* selected in vitro to bind to CSA (Salanti et al. 2003) and in parasites isolated from the placenta (Tuikue Ndam et al. 2005). Antibodies to the surface-expressed VAR2CSA are acquired by women exposed to malaria during pregnancy (Salanti et al. 2004; Tuikue Ndam et al. 2006) and high levels of anti-VAR2CSA antibodies at delivery are associated with protection from LBW (Salanti et al. 2004). Furthermore, it has been demonstrated that targeted disruption of *var2csa* results in the loss of (Viebig et al. 2005), or a marked reduction (Duffy et al. 2006) in the ability of parasites to adhere to CSA. Based on these findings VAR2CSA is recognized as the leading PAM vaccine

candidate, however *var2csa* is a polymorphic gene and the sequence variation between genes from different parasites ranges from 10 to 30%, at the nucleic acid level (Salanti et al. 2003; Trimnell et al. 2006). It is thus a major challenge for vaccine development to characterize the importance of the sequence variation and to define smaller epitopes that can be used in a vaccine to protect women against PAM. This study had two objectives. Firstly, to characterize the epitopes of the CSA binding Duffy-binding-like (DBL) 3X domain of VAR2CSA, which are recognised by naturally acquired antibodies. Secondly, to analyze the degree and consequences of sequence variation and selection pressure within the *var2csa* subfamily, using the *var2csa* cDNA sequences of a large number of fresh placental parasite isolates. These studies would also test the validity of B-cell epitope predictions and structural modelling of the DBL3X domain.

## Results and Discussion

### Recombinant VAR2CSA DBL3X binds CSA and the affinity depends on the primary amino acid sequence

It has previously been shown that DBL3X expressed on the surface of CHO cells binds CSA in vitro (Gamain et al. 2005). However it is important to test the CSA binding properties of secreted VAR2CSA proteins produced in expression systems that could allow for larger scale production of a vaccine. For this study recombinant HIS-tagged proteins were produced in *Baculovirus*-transfected insect cells and binding to CSA was determined in an ELISA system. It has previously been suggested that the FCR3 DBL3X domain binds CSA, whereas the 3D7 DBL3X domain does not (Gamain et al. 2005). We found that both variants had affinity to CSA and that the 3D7 variant exhibited the strongest binding. Binding of both FCR3 and 3D7 DBL3X was concentration dependent (Figure 3.2A) and could be inhibited by soluble CSA in a dose dependent manner (Figure 3.2B).

### The structure of the VAR2CSA DBL3X domain can be modeled on the basis of the structure of the DBL domains of Erythrocyte Binding protein (EBA)-175

Recently, the structures of the two DBL domains (F1 and F2) of EBA-175 and a DBL domain in *Plasmodium knowlesi* (Pk)$\alpha$-DBL were solved (Tolia et al. 2005; Singh et al. 2006). Using the crystal structure of EBA-175 F1 (PDB code 1ZRO) as template, a three-dimensional model of DBL3X VAR2CSA was constructed by comparative modeling on the basis of the 3D7 sequence (Figure 3.2C). The sequence identity between DBL3X and EBA-175 F1 was 28%. From sequence alignments of DBL3X, Pk$\alpha$-DBL and EBA-175 (Supplementary figure 3.9) it was apparent that a number of cysteines were conserved and 10 of these from DBL3X aligned with cysteines which form disulfide bridges in the determined structures of EBA-175 (Tolia et al. 2005) and Pk$\alpha$-DBL (Singh et al. 2006). In addition, we identified 34 buried hydrophobic residues in EBA-175 F1, EBA-175 F2 and Pk$\alpha$-DBL, which corresponded to hydrophobic residues buried in the DBL3X model (Supplementary figure 3.9). Compared to EBA-175 F1, a number of insertions were found in DBL3X and the majority of these were predicted to have coil secondary structure (Figure 3.2C). One of the insertions in DBL3X (L1:

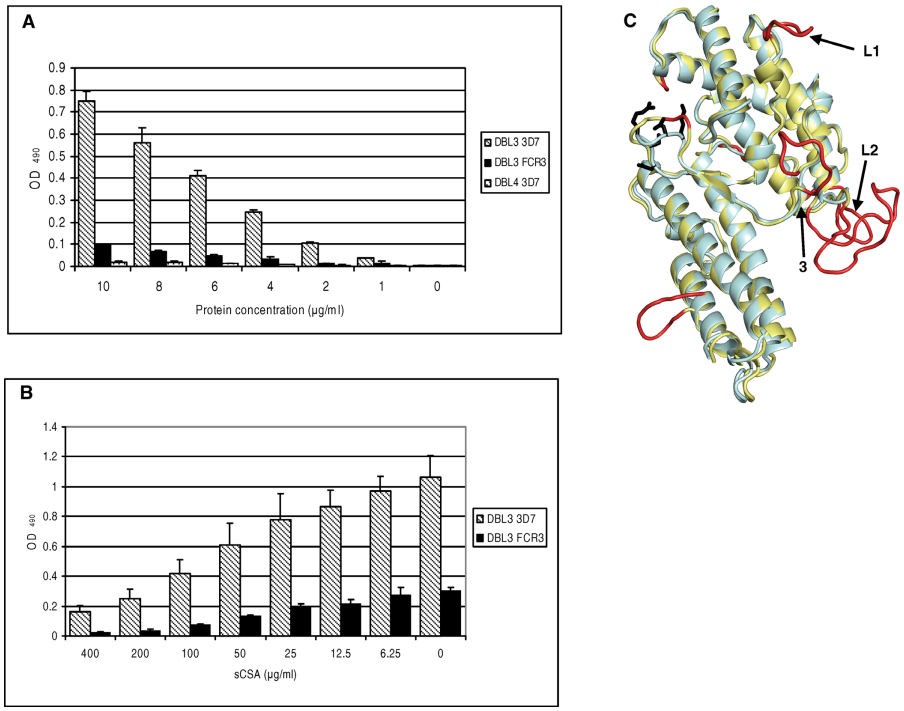


Figure 3.2: **Recombinant DBL3X VAR2CSA binds CSA and the structure of the domain can be modeled on the basis of the structure of the DBL domains of EBA-175.** (A) Binding assay. DBL3X binds to CSA in a protein concentration specific manner. ELISA plates were coated with soluble CSA and binding of recombinant proteins at different concentrations were determined. DBL3X from 3D7 shows better binding compared to DBL3X from FCR3. The non-CSA binding VAR2CSA DBL4$\epsilon$ was included as a control. Results are the mean of three binding assays and the error bars indicate the standard deviation. (B) Inhibition assay. Recombinant DBL3X proteins (7$\mu$g/ml) were pre-incubated with increasing amounts of soluble CSA and binding to CSA coated plates were determined. Binding of DBL3X is dose dependently inhibited by soluble CSA. Results are the mean of four inhibition binding assays and error bars indicate the standard deviation. (C) Model of DBL3X using EBA-175 as template. The figure shows a superposition of the EBA-175 F1 domain in light blue and the DBL3X (PFL0030c amino acids 1217 - 1559) model in yellow with insertions highlighted in red. Arrows indicate insertions, which align with flexible loop regions in Pk$\alpha$-DBL (L1) and EBA-175 F1 (L2). Side chains of glycan binding residues in EBA-175 F1 are shown in black. Arrow 3 indicates the corresponding Pk$\alpha$-DBL DARC binding site.

R56-I63) was found to align structurally next to a region in Pk$\alpha$-DBL, which is described as a flexible linker with an experimentally determined proteolytic cleavage site (Singh et al. 2006). A second insertion (L2: N1417-E1430) was aligned in a loop region where two residues were missing structural information in the structure of EBA-175 F1 (Tolia et al. 2005), which also indicates high flexibility. Thus, the alignment of DBL3X to the solved DBL domain structures seems to match characteristics such as disulfide-bridges, stabilizing hydrophobic

interactions and flexible loop regions. These findings suggests that VAR2CSA DBL3X has the same basic structure as the solved DBL structures (Tolia et al. 2005; Singh et al. 2006; Howell et al. 2006) in spite of the extensive sequence variation. The ability of proteins to form similar structures despite marked sequence variation has also been described for the VSG molecules covering the surface of Trypanosomes (Blum et al. 1993; Carrington and Boothroyd 1996).

### Sequence variation within the *var2csa* family is present as small hypervariable blocks and parasite diversity is similar on a local and global scale

*Var2csa* is a relatively conserved *var* gene carried by all *P. falciparum* genomes, however sequence polymorphisms are present in the gene. In a previous study, it was established that *var2csa* is transcribed at high levels by placental parasites isolated at delivery from Senegalese women (Tuikue Ndam et al. 2005; Tuikue Ndam et al. 2004). Using cDNA from 24 placentas from that study, the region encoding DBL3X of *var2csa* was cloned and sequenced. A multiple alignment of 43 sequences showed that the average nucleotide diversity was low ($\pi = 8.48\% \pm 0.37\%$, see materials and methods for details) reflecting a limited inter-parasite diversity in these isolates collected from a geographically small and well-defined area of West Africa. To test whether the Senegalese placental DBL3X sequences represented a monophyletic group compared to non-Senegalese DBL3X sequences, a phylogenetic tree, which included VAR2CSA DBL3X sequences available in GenBank, was constructed (Figure 3.3). These non-Senegalese sequences included four lab-strains parasites with different geographic origin (DD2 from Laos, MC from Thailand, IT4 from Brazil and 3D7 with unknown origin) and 21 sequences from a sequencing study from Malawi (Duffy et al. 2006). The four database sequences and the Malawi sequences were scattered evenly in the tree indicating that the Senegalese sequences were representative for the *var2csa* repertoire in general and that parasite nucleotide diversity is similar on a local and global scale. It is also apparent that there is no clear subgrouping within the DBL3X sequences. The protein alignment of the DBL3X sequences (Figure 3.4A and Appendix II) suggested that DBL3X could be divided into four relatively conserved regions (C1-4) separated by three shorter variable regions (V1-3). When the variable regions were mapped to the three-dimensional model (Figure 3.4B), V1 and V3 were part of flexible loops, whereas V2 included another flexible loop but also extended into a helical region. The length of V1 and V3 varied between sequences and the 3D7 sequence was relatively short in both regions.

### VAR2CSA sequence motifs can be linked to the parity of the infected woman

We have previously shown that the expression of certain *var* genes is associated with severe malaria in young children (Jensen et al. 2004) and suggested that *var* gene expression is hierarchically structured. This could occur because the progeny of parasites expressing *var* gene products that mediate the most effective sequestration outgrow the progeny of parasites expressing a molecule mediating less effective binding (Jensen et al. 2004; Hviid and Staalsoe 2004; Lavstsen et al. 2005). A similar process could shape the expression of mole-

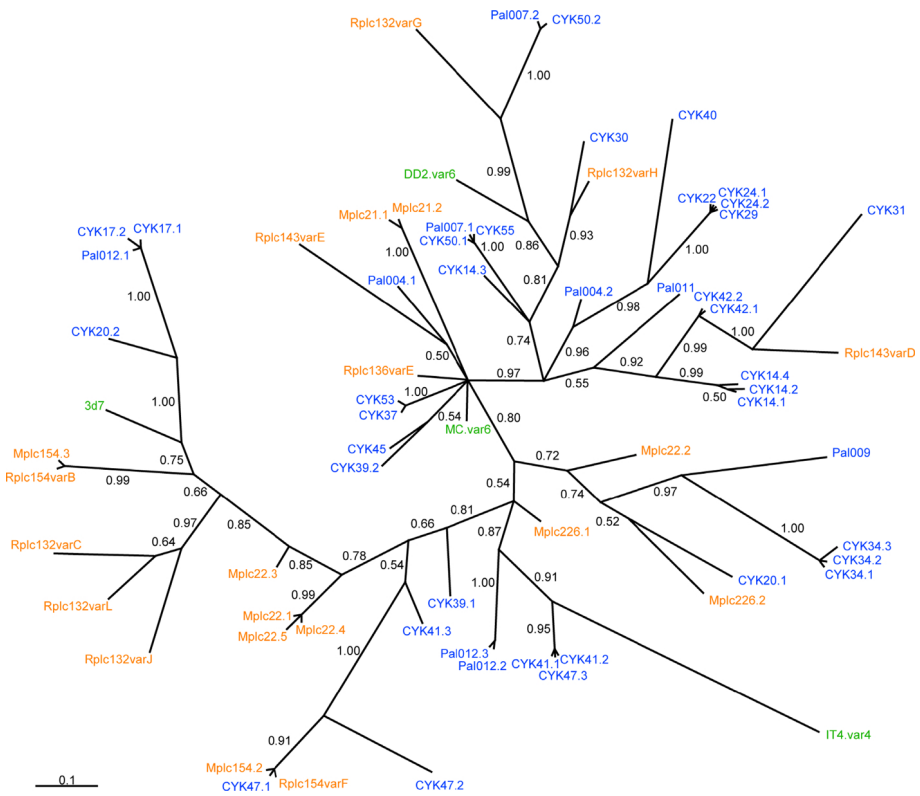


Figure 3.3: **Phylogentic relationship between different VAR2CSA DBL3X sequences.** Bayesian inference tree of 43 placental cDNA sequences from Senegal (blue), 21 Malawian sequences (orange) and database sequences from four isolates (green) of different geographic origin. The posterior probability of the clades is shown at the branches.

cules mediating binding in the placenta. If that was the case, it would be predicted that a systematic difference between molecules mediating binding in primigravidae and multigravidae women could be detected in areas of high *P. falciparum* transmission where women are exposed to several parasite clones during pregnancy (Jafari-Guemouri et al. 2006). This was addressed by calculating the Kullback-Leibler distance between 21 VAR2CSA sequences originating from primigravidae and 21 sequences from multigravidae. The overall cumulated Kullback-Leibler distance ($D_{KL}$) between sequences from the primigravidae and the multigravidae was higher than for randomly chosen groupings of the sequences ($p$=0.0075). The $D_{KL}$ was also calculated for each position in the alignment to determine which polymorphisms contributed to the difference between the sequence of parasites from primi- and multigravidae women and visualized in a Kullback-Leibler sequence logo (Figure 3.5A). This implicated two stretches of amino acids in positions 135-175 and 233-245 located in the V2 and V3 regions, respectively. The region from position 158 to 162 was of special interest since the motif "EIEKD" was mainly found in primigravidae and the motif "GIEGE" mainly in the multigravidae (Table 3.1). Intermediate motifs

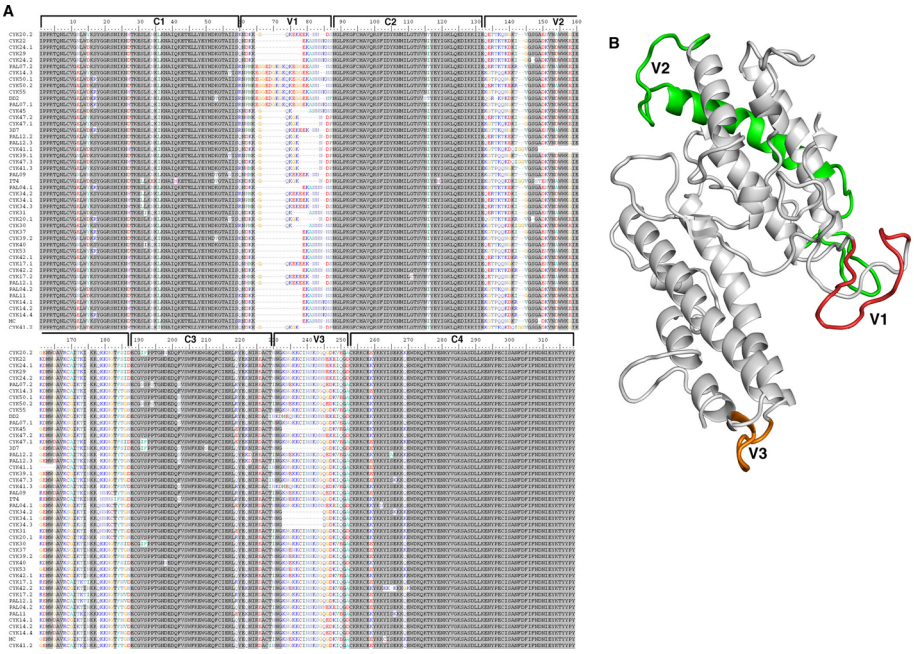


Figure 3.4: **Multiple alignment of VAR2CSA DBL3X sequences.** (A) cDNA from 43 placental parasite samples were amplified with conserved DBL3X primers and sequenced. Sequences were curated for primer sequence, translated and aligned. The text to the left indicate the identity of the samples. The DBL3X domain can be divided into four regions, which are highly conserved, C1-C4, and three regions, which are polymorphic and/or harbour deletions V1-V3. A larger version of the alignment can be found in the Appendix II of this thesis. (B) Model of DBL3X showing the position of V1 (red), V2 (green) and V3 (orange).

like "(G/E)IERE" where the fourth position had changed from lysine to arginine were also found, implying that the following evolutionary pathway may have been operating: lysine (AAA or AAG) ↔ arginine (AGG) ↔ glycine (GGG). The change from glutamate and lysine/arginine, which have large charged side chains into glycine without a side chain could result in marked functional and antigenic changes. Thus, it was interesting that positions 1, 4, and 5 in the motif (position 158, 161 and 162) were predicted to be surface exposed in the structural model, which was also the case for all of the other amino acid positions in V2 that differed significantly in the Kullback-Leibler sequence logo (Figure 3.5B and 3.5C). The finding that parasites expressing the EIEKD motif were more prevalent in primigravidae than in multigravidae women (Table 1, Chi-square test: $p<0.001$) could indicate that these parasites have a biological advantage in women experiencing their first pregnancy and that parasites expressing the other motifs (G/EIERE or GIEGE) have an advantage in women who have been pregnant before. This could arise because parasites carrying the EIEKD motif are the most effective mediators of binding and therefore dominate in women with limited immunity against PAM. As immunity develops against these parasite forms, parasites expressing other motifs that are less effective in binding but not serologically cross-reactive take over. Interestingly a monoclonal Ghanaian

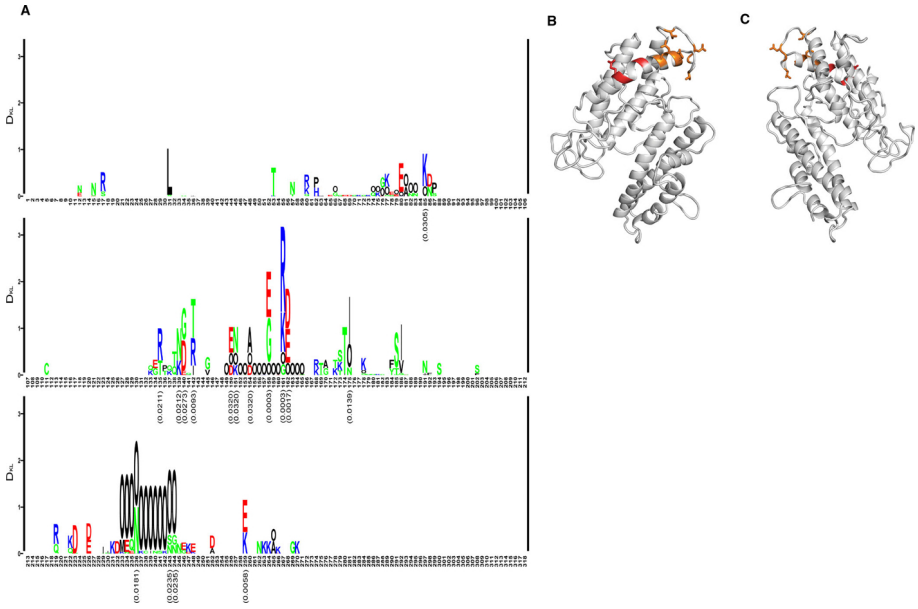


Figure 3.5: **Sequence differences in infections with different parity.** (A) Kullback-Leibler sequence logo based on a multiple alignment of the DBL3X region (Figure 3.4). The sequences in the alignment were split into two groups of equal size: 21 sequences obtained from primigravidae and 21 sequences from multigravidae women. The x-axis shows amino acid positions in the alignment. The height of a position indicates the amount of difference between the two groups according to the symmetric Kullback-Leibler distance, and the letters indicate the amino acids contributing to this difference. The letter "O" has been chosen to signify gaps in the alignment. Polar amino acids are green, basic are blue, acidic are red and hydrophobic amino acids are black. P-values indicating the significance of DKL based on the parity grouping compared to random groupings are shown below positions where $p<0.05$. (B) DBL3X model showing the position of amino acids that were differentially found in primigravidae and multigravidae women. Residues in the region 135-175 of moderately significant difference are shown in orange and highly significant residues (positions 158, 161, and 162) are shown in red. (C) The model (B) rotated 180 degrees so it is positioned similar to Figure 3.2C.

field isolate undergoing full genome sequencing at the Sanger institute has two copies of *var2csa*, one with the EIEKD motif and one with the GIEGE. Although the functional background for the observed phenomenon is presently not clear, the systematic sequence variation at positions predicted to be surface exposed between parasites from primi- and multigravidae women strengthen the concept that VAR2CSA is the main parasite ligand for sequestration of malaria parasites in the placenta.

### VAR2CSA is under both positive and purifying selection

A recent study has suggested that sequence polymorphisms in a region of VAR2CSA upstream to DBL3X largely are due to positive natural selection pressure (Trimnell et al. 2006). To further investigate the nature of sequence diversity in

| Sequence motif in DBL3X | Primigravidae ($n$=21) | Multigravidae ($n$=21) |
|---|---|---|
| EIEKD | 16[a] | 4 |
| EIERD/EIEGE/GIERE | 1 | 6 |
| GEIGE | 4 | 10 |
| Gap | 0 | 1 |

[a]Significantly higher represented in primigravidae, $p<0.001$, Chi-square.

Table 3.1: **Sequence signature in DBL3X positions 158-162 of VAR2CSA expressed by 42 parasites infecting either primigravidae or multigravidae.** The motif "EIEKD" is overrepresented iin primigravidae and underrepresented in multigravidae. The intermediate motifs and "GEIGE" are mainly present in multigravidae. One of the sequences from multigravidae contains a gap in the region

*var2csa*, the dN/dS ratios (dN: rate of non-synonymous mutations per non-synonymous site and dS: the rate of synonymous mutations per synonymous site) for DBL3X were calculated on the basis of the sequenced DBL3X domains. A dN/dS less than 1 indicates that the position is under negative or purifying selection pressure leading to conservation of the residue, while dN/dS>1 suggests positive or diversifying selection pressure, and suggest that amino acid changes are evolutionarily advantageous at the position. Purifying selection was mainly found in the conserved regions C1-4 and it was especially pronounced in the C2 and C4 regions, although single sites were observed to be under diversifying selection (Figure 3.6A). Several blocks appeared to be under strong diversifying selection and these appeared most prominently in regions V1, V2, and V3. It is interesting to note that residues under diversifying selection mainly were situated in regions predicted to be surface exposed and concentrated on one side of the molecule (Figure 3.6D and 3.6E). The two DBL domains of EBA-175 are predicted to form a reverse handshake dimer with the F1 domain of each molecule interacting with F2 of the other (Tolia et al. 2005). In this four-domain DBL structure, we replaced one of the EBA-175 F1 domains with our VAR2CSA DBL3X model (Figure 3.6F). It was noticeable that the largely conserved C2 and C4 regions of DBL3X were predicted to take part in the lining of the central cavity of the four-domain structure and form the region next to the cavity facing the membrane in native configuration. The model presented here is unlikely to fatefully reflect the structure of the native molecule, but the positioning of conserved DBL3X regions in the model makes biological sense, and the finding underscores the need to obtain knowledge about possible interactions between VAR2CSA DBL domains. The lack of amino acid positions under diversifying selection in the regions adjacent to the cavity might be due to that these sites are involved in ligand binding and thus functional constraint. Another explanation could be that the regions forming the predicted cavity and the area predicted to face the membrane are not accessible for antibodies in natively folded molecules.

### Recombination is a factor in the generation of *var2csa* sequence variation

Previous studies have reported that frequent recombination events generate sequence diversity in the PfEMP1 family (Taylor et al. 2000; Freitas-Junior et al. 2000; Mu et al. 2005). To determine the role of recombination for DBL3X se-

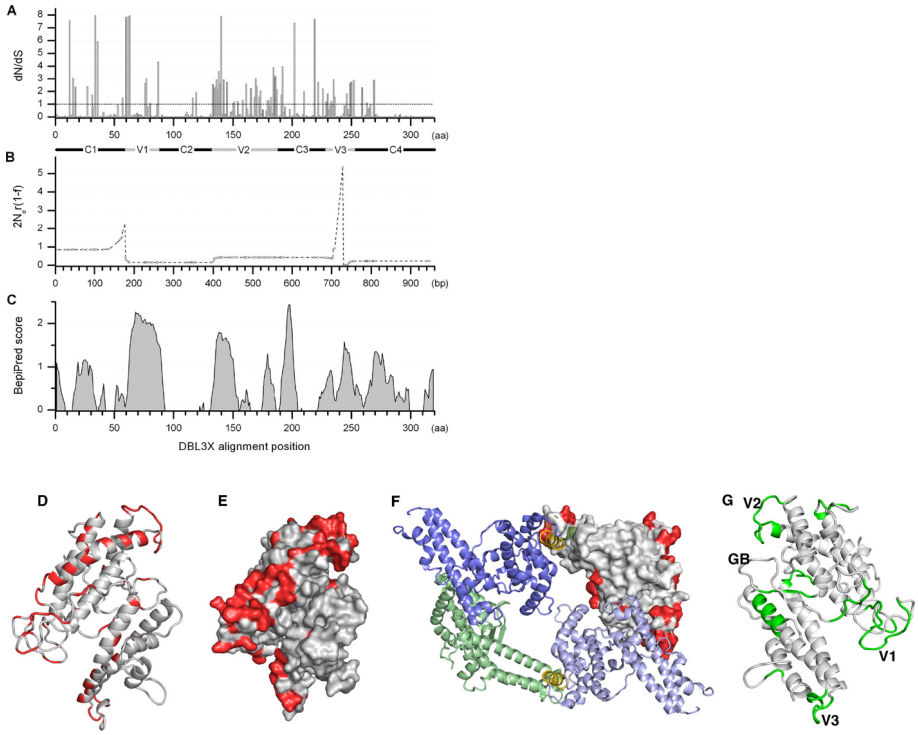


Figure 3.6: **Sequence variation and B-cell epitopes in DBL3X.** (A) dN/dS ratio per amino acid based on an alignment of 47 DBL3X sequences. dN/dS ratio higher than the dashed line indicate that the position is under positive selection. V1-V3 and C1-C4 are indicated with solid bars. It should be noted that the inference of selection pressure is less reliable in positions with gaps, which are mainly found in the variable regions (see Figure 3.4). (B) The population recombination rate $\rho = 2N_e r(1-f)$ estimated per base pair across DBL3X. Circles indicate the single nucleotide polymorphisms (SNPs) for which $\rho$ is estimated, linear interpolation is shown between SNPs. Two hotspots corresponding to a local raise in $r$ are observed in association with the variable regions V1 and V3. (C) B-cell epitope predictions of the 3D7 sequence performed by the BepiPred server. DD2 and IT-4 were used to fill 3D7 sequence gaps. (D) Cartoon representation of the DBL3X model showing amino acid positions under diversifying selection in red (dN/dS ratios >1). (E) DBL3X model with surface topography prediction showing amino acid positions under diversifying selection in red (dN/dS ratios >1). (F) The homo-dimer of the two domains of RII of EBA-175 has been shown to form a handshake structure with F1 of one molecule dimerizing with F2 of another molecule and vice-versa. Here VAR2CSA DBL3X has been superimposed on the F1 domain of one of the EBA-175 molecules in the dimer. The F1 domain is shown in green; F2 domains are blue and the linker regions are gold. Amino acids under diversifying selection in DBL3X are shown in red (dN/dS ratios >1). It is clear that most residues inwards of the dimer are not under positive selection. (G) DBL3X model showing residues with BepiPred scores higher than 0.9 in green. Locations of V1-V3 are indicated on the model. GB indicates the glycan binding site of EBA-175.

quence variation, we estimated the population recombination rate $\rho$ defined for partially inbreeding haploid species by the compound parameter $2N_e r(1-f)$,

where $N_e$ represents the effective population size, $r$ is the per-generation cross-over recombination rate per base pair (bp) and f is the inbreeding coefficient. Variations in $\rho$ across DBL3X correspond to variations in the recombination rate $r$, as $N_e$ and $f$ are constant for the data set. Two recombinational hotspots were observed at bp positions 138-178 and 704-730 (Figure 3.6B). The two best defined recombination breakpoints were present in the C1/V1 borderline at bp 177-179 and within V3 at bp 728-730. Both the V1 and V3 region harboured major deletions/insertions in several of the sequences, which may have arisen as a consequence of unequal cross-over during recombination at the hotspots. Unequal cross-over results in either deletions or insertions of variable number of tandem repeats (VNTR), and the DBL3X V1 region did indeed contain a high amount of VNTR. The loop region of V2 contained a small VNTR insert in some sequences, while VNTR were less obvious in the V3 region. In a recent study of the *P. falciparum* chromosome 3, high recombination rates were found in sub-telomeric regions, and African *P. falciparum* strains showed a much higher population recombination rate than strains from other regions of the world (Yang et al. 2003). The overall population recombination rate for the DBL3X region was estimated to $\rho = 0.54$ per bp (95% CI: 0.41-0.90). This is in accordance with a recent report for a VAR2CSA region upstream to DBL3X in interdomain 1, where the rate was estimated to 0.71 per bp (Trimnell et al. 2006), and in agreement with recombination rates in chromosome 3 of African parasite lines summarized as $\rho > 0.1$ per bp (Mu et al. 2005). The high population recombination rate in DBL3X combined with the observed sequence variation adjacent to the detected recombination hotspots, suggests that recombination is an important factor in generating sequence variation. The V1 region which seems to be most strongly affected by recombination is predicted to form a structurally unrestricted flexible loop allowing for sequence variation, and it is possible that the whole V1 region is under diversifying selection pressure exerted by the immune system, even though this could not be predicted by the dN/dS method due to gaps. This notion is supported by the findings discussed below, showing that this region is part of a major B-cell epitope.

### VAR2CSA DBL3X B-cell epitope prediction

VAR2CSA epitopes exposed on the surface of the protein and accessible for IgG binding could be under diversifying selection resulting in escape mutations and high dN/dS ratios, whereas residues involved in protein folding, stability and anchoring could be under purifying selection with low dN/dS ratios. To predict linear B-cell epitopes the 3D7 sequence was submitted to the BepiPred server (Larsen et al. 2006) and seven epitopes were predicted within the DBL3X sequence (Figure 3.6C). Some of these epitopes were located in areas with high dN/dS ratios and there was a weak but statistically significant association between the BepiPred score and the dN/dS ratio (Pearson's $r$=0.18 and $p = 2.5 \times 10^{-9}$). The reason for this weak association could be that antibody epitopes are situated in regions that are functionally constrained or that positive selection pressure is also driven by other forces like MHC-2 binding and T-helper cell activation as found for HIV-1 (Yang et al. 2003). Furthermore, the BepiPred algorithm predicts linear epitopes and some of these could be located in parts of the molecule that are not accessible to antibodies in the native folded molecule. Nevertheless, most of the predicted epitopes (Figure 3.6G)

were situated in surface exposed loop regions of the model, and one of the highest scoring epitopes was in the V1 region. Residues that align directly to the glycan-binding residues of EBA-175 F1 and to the Duffy antigen receptor for chemokines (DARC) binding site in Pk$\alpha$-DBL were not predicted to be part of epitopes. However, a part of the V2 region in proximity to the putative glycan-binding loop was predicted to be targeted by antibodies and had high dN/dS values.

### Fine epitope mapping of VAR2CSA antibodies acquired during pregnancy

To verify the above mentioned bioinformatical predictions we evaluated the fine specificity of naturally acquired human antibodies to VAR2CSA DBL3X in a peptide array consisting of 442 overlapping 31mer peptides covering exon I of VAR2CSA of the 3D7 sequence. Antibody reactivity to individual amino acids was assigned on the basis of an algorithm based on the observation that a major part of antibody binding motifs in a set of conformational epitopes are from 2 to 7 amino acids long, containing either 2 or 3 defined residues spaced by undefined amino acids (Haste Andersen et al. 2006). The VAR2CSA of 3D7 contains 31149 such motifs and each was assigned an average pepscan value by adding the measured reactivity from the 31mers in which the motif was present and dividing with the number of times the motif occurred. The method, described in the material and methods section, was validated by testing serum from rabbits immunized with a VAR2CSA construct, which showed that both the measurements based on individual peptide readings and the algorithm described above, mapped antibody responses to regions present in the antigen used for immunization (Supplementary figure 3.10). Furthermore, there was a high concurrence between antibody peaks defined by the reactivity of the individual peptides and the peaks defined by the algorithm defining single amino acid scores (Supplementary figure 3.10).
Plasma from individuals not exposed to malaria did not react with any of the peptides in the array (data not shown). The IgG reactivity in the plasma of eight Ghanaian women with a known history of a placental malaria infection was analyzed and the peptide array data was visualized on the DBL3X model (Figure 3.7). The regions with the highest reactivity were on the side of the domain where glycan-binding is found in EBA-175 F1 (Figure 3.7A and 3.7B) and therefore IgG reactivity was visualized on a model positioned with this side in the front (Figure 3.7A and 3.7C - I). It was clear that the majority of the individuals had specific IgG against the variable regions, V1 or V2. V3 is partly deleted in 3D7 and IgG reactivity could thus not be measured in the peptide array based on the 3D7 sequence. Remarkably, a short $\alpha$-helix (Figure 3.7A-arrow 1) in proximity to the loop for glycan-binding in EBA-175 F1 showed the highest antibody reactivity in all serum samples, despite the fact that this region had very low dN/dS ratios (Figure 3.6A, positions 120-132). Another $\alpha$-helix was also often recognized by antibodies (Figure 3.7A-arrow 2). This helix was predicted to contain a B-cell epitope by the BepiPred algorithm and had polymorphic residues (Figure 3.6A and 3.6C, positions 251-281). The region corresponding to the loops containing the EBA-175 F1 glycan-binding sites was not recognized by any of the serum samples. Conserved regions (Figure 3.7A-arrow 1) targeted by naturally acquired IgG are of considerable interest in

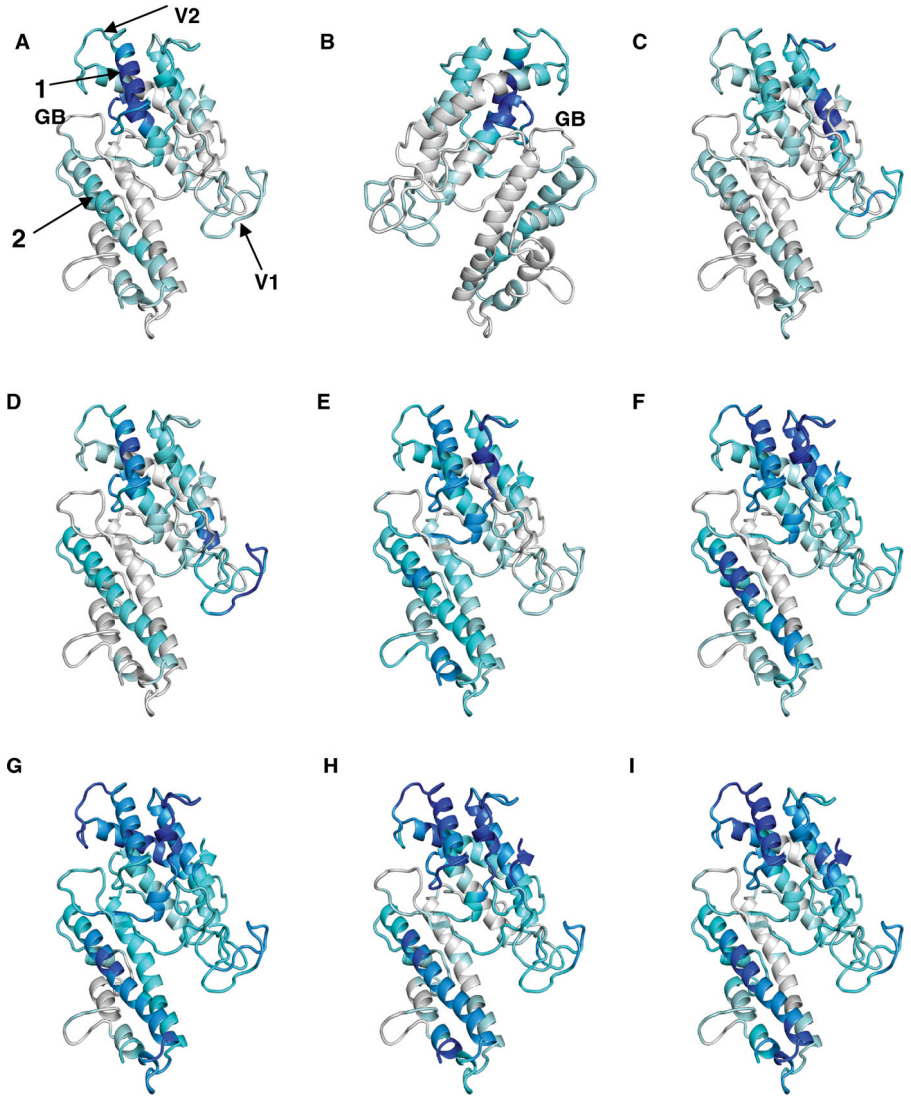


Figure 3.7: **Defining targets of antibodies on DBL3X.** Models of DBL3X in which the intensity of the grey and blue indicates reactivity of plasma tested on a peptide array (pepscan scores). Grey indicates lowest scores and dark blue indicates highest scores. (A and C-I) represent reactivity in eight individual plasma samples. (B) shows the reactivity in the same plasma as tested in (A), but the model has been rotated 180 degrees. Arrow 1 indicates a highly recognized $\alpha$-helix region, which was not predicted to be an epitope by BepiPred and was characterized by low dN/dS ratios and high sequence variation. Arrow 2 indicates an $\alpha$-helix with a predicted B-cell epitope containing residues with high sequence variation. Variable regions V1 and V2 are fairly well recognized by most of the serum samples. GB indicates the glycan binding site of EBA-175.

the search for vaccine constructs which could elicit a broad protective immune response.

### Pepscan analysis of affinity purified antibodies

During infection VAR2CSA will be degraded and antibodies will be acquired against epitopes that are not accessible for antibodies when the protein is in its natural conformation. It is therefore possible that some of the antibody reactivities measured in the peptide array were directed against such epitopes. To address this question, analysis was performed on plasma, which had been affinity purified on recombinant DBL3X or antibody-depleted by incubation with erythrocytes infected with VAR2CSA expressing parasites.

Plasma from a rabbit immunized with recombinant DBL3X and plasma from women who had suffered a placental infection were affinity purified on the recombinant CSA binding DBL3X protein and analyzed by the peptide array. Before affinity purification, the plasma pool from Ghanaian women showed reactivity corresponding to eight peaks distributed throughout the domain (Figure 3.8A). By contrast, the reactivity of the affinity-purified IgG was concentrated to three peaks (SE1-SE3) in the C1, V1, and C2/V2 regions. The immunized rabbit had strong IgG reactivity against the epitope in the C2/V2, which was also affinity purified from the female plasma pool (Figure 3.8B), but the rabbit had not raised an IgG response against the epitopes in C1 and V1. A plasma pool from Tanzanian women was also analyzed and in this case surface reactive antibodies were depleted from the pool by incubation with VAR2CSA expressing infected erythrocytes. In this pool the main reactivity was against the surface exposed epitopes (SE1-SE3) defined in the Ghanaian plasma, and the depletion experiment indicated that absorption with infected erythrocytes caused a marked reduction in the reactivity against these epitopes (Figure 3.8C). These findings indicate that the three identified regions were accessible to antibodies on the native protein and that the folding of the *Baculovirus* produced recombinant protein was close to the natural configuration. SE2 and SE3 corresponded to loop regions V1 and V2, which are both protruding from the structure of the DBL3X model (Figure 3.8D). Interestingly SE1 and SE3, which are located in separate parts of the primary structure of the domain, form a continuous region on the surface of the predicted DBL3X protein structure (Figure 3.8E). The model also predicts that all surface exposed sites are located on one part of the domain indicating that the other parts are buried or engaged in the intact PfEMP1 molecule expressed on the surface of the infected erythrocyte (Figure 3.8E). Unexpectedly, the highly conserved part of C2, which was well recognized in the peptide array by all women, corresponded to a part of the surface exposed epitope SE3.

When exchanging the F1 domain of one of the EBA-175 molecules in the EBA-175 dimer with VAR2CSA DBL3X and mapping SE1-SE3 to the model, it is apparent that the surface exposed regions are on the opposite side of the central cavity of the dimer (Figure 3.8F). However, the part of DBL3X that may be directly involved in the dimerization extends into the SE3 region (shown in green), indicating that in this model the potential dimerization motifs are accessible for antibodies. EBA-175 is suggested to dimerize upon ligand interaction and it could be that SE3 was "unengaged" due to the lack of CSA during antibody depletion. If several domains need to interact to form a buried CSA binding site, antibodies targeting conserved residues like the region in C2 that forms parts of SE3 could function by inhibiting the dimerization of domains.

Our results demonstrate that both conserved and polymorphic surface exposed

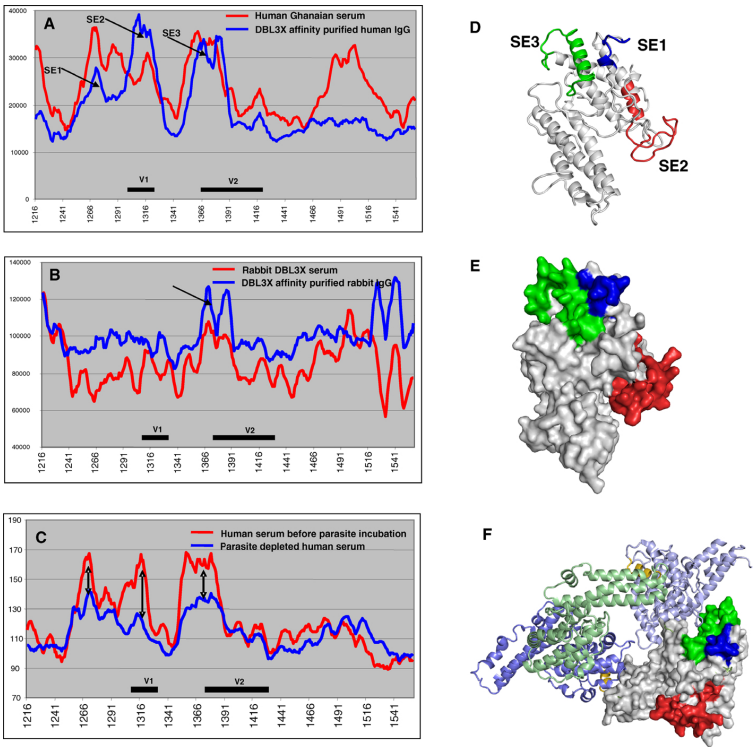


Figure 3.8: **Reactivity of plasma affinity purified on recombinant DBL3X and human IgG depleted for surface reactivity by incubation with infected erythrocytes expressing VAR2CSA.** (A) pepscan IgG reactivity in a Ghanaian pool of pregnancy plasma before (red) and after (blue) affinity purification on recombinant DBL3X protein. Variable regions V1 and V2 are indicated with a black line. Arrows indicates 3 surface exposed regions (SE). The X-axis is amino acid position in 3D7 VAR2CSA. (B) Pepscan IgG reactivity in plasma from a rabbit immunized with DBL3X before (red) and after (blue) affinity purification on recombinant DBL3X protein. Arrow indicates the surface exposed region. N and C-terminal parts of the recombinant proteins did also appear to be surface exposed; this could be due to improper folding of the N and C terminal parts of the protein. (C) Pepscan IgG reactivity in a Tanzanian pool of pregnancy plasma before (red) and after (blue) depletion of antibodies directed against surfaced exposed VAR2CSA regions by incubation with infected erythrocytes expressing VAR2CSA on the surface. Arrows indicates three regions where the reactivity was markedly reduced after depletion largely corresponding to SE1, SE2 and SE3 on panel A. (D) DBL3X model showing the location of surface exposed epitopes, SE1 (blue), SE2 (red) and SE3 (green). (E) DBL3X surface topography showing SE1-SE3. From the model it is apparent that the SE1 and SE3 form a discontinuous epitope. (F) VAR2CSA DBL3X superimposed on the F1 domain of one of the molecules in the EBA-175 dimer. The F1 domain is shown in green; F2 domains are blue and the linker regions are gold. On the DBL3X domain the three experimentally verified surface exposed regions are shown in similar colouring as (D and E). The three regions are mainly predicted to be on the opposite side of the central cavity of the dimer, and parts of SE3 correspond to amino acid positions directly involved in the dimerization of EBA-175.

regions are targets for VAR2CSA DBL3X antibodies acquired during pregnancy by malaria-infected women. This opens for vaccine strategies similar to those being pursued for the polymorphic Merozoite Surface Protein (MSP) 1 (Holder et al. 1999). The proteolytic processing of MSP1 is a prerequisite for successful parasite invasion of erythrocytes and one vaccine strategy is based on the induction of antibodies against the conserved C-terminal part of the molecule that inhibit processing (Blackman et al. 1994). Another MSP1 vaccine strategy employs chimeric vaccine constructs designed to induce antibodies targeting the polymorphic types present in the N-terminal part of the molecule (Tetteh et al. 2005). In a similar fashion VAR2CSA vaccine constructs could target conserved epitopes like those identified in SE3 or alternatively constructs should induce antibodies targeting different serological variants like those predicted to be generated by the sequence polymorphisms present in regions, SE1 and SE2. The human antibody pools used in this study to identify surface exposed antigenic targets inhibit parasite binding to CSA in vitro (data not shown). However, the molecular targets of the inhibitory antibodies have not been identified and knowledge is required about the antigenic targets for antibodies on the other VAR2CSA DBL domains. The high similarity between the DBL structures of EBA-175 (Tolia et al. 2005), Pk$\alpha$-DBL (Singh et al. 2006) and the VAR2CSA DBL3X model suggests that the DBL structures in *Plasmodium* are relatively conserved and that the antigenic characteristics of the DBL3X might be comparable to those of the remaining VAR2CSA DBL domains. However, a more comprehensive analysis of sequence variation, antibody epitopes and structure of the VAR2CSA DBL domains not belonging to DBL3X is needed to establish the extent of the structural conservation between the domains.
Development of PAM vaccines requires a much better understanding about the molecular interaction between placental parasites and the ligand on the syncytiotrophoblasts, as well as knowledge about the fine specificity of the targets for antibodies inhibiting binding. Native PfEMP1 molecules are difficult to isolate and with current technologies it is difficult to produce correctly folded recombinant material in quantities allowing structure elucidation by crystallography. In this paper, we have combined in silico methods such as model building and sequence analysis and the analysis of antibody reactivity to obtain new information and generate hypotheses about the structure and functional relationship of VAR2CSA.

## Materials and methods

### Cloning and expression of VAR2CSA domains

DBL3X and DBL4$\epsilon$ of VAR2CSA was amplified from FCR3 and 3D7 genomic DNA with the following primers: FCR3 DBL3X - 5' CG GAA TTC ACC AAT ATT AAT AAA AGT GAA and 3' ATT TGC GGC CGC CAG CAT TAT TAT ATT TGT A, 3D7 DBL3X - 5'CG GAA TTC AAG ATG AAG TCC TCC GAG and 3'ATT TGC GGC CGC CAA AAC AGC CAA GCT GGA, 3D7 DBL4$\epsilon$ - 5'CG GAA TTC CAG GTG AAG TAC TAC GAA and 3'CTG TTC CTC CAC GTG CTC CAG. PCR products were digested with *Eco*RI and *Not*I for cloning into the *Baculovirus* vector, pAcGP67-A (BD Biosciences), which was modified to contain the V5 epitope upstream of a histidine tag in the C-terminal end of the constructs. Linearised Bakpak6 *Baculovirus* DNA (BD

Biosciences Clontech) was co-transfected with pAcGP67-A into Sf9 insect cells for generation of recombinant virus particles. Recombinant protein was purified on $Co^{2+}$ metal-chelate agarose columns as secreted histidine-tagged proteins from the supernatant of infected High-Five insect cells.

### CSA binding assay

Binding to CSA (C9819 Sigma-Aldrich, Brøndby, Denmark) was determined in an ELISA system. ELISA plates (Falcon 351172) were coated overnight with CSA (50 $\mu$g/ml) in PBS at 4°C. Coating with 1% BSA in PBS (blocking buffer) was used as negative control. Plates were incubated with blocking buffer for 1 hour at room temperature (RT) to inhibit non-specific adsorption to the plate. The VAR2CSA proteins were diluted in blocking buffer (1-10 $\mu$g/ml protein), added to the wells and incubated for 1 hour at RT. For the inhibition assays, proteins (7$\mu$g/ml) were pre-incubated with different concentrations of soluble CSA for 30 min. Plates were washed 4× in PBS in between the different steps. Specific binding was visualized by adding an HRP-conjugated antibody (R960-25, Invitrogen, Taastrup, Denmark) targeting the V5 epitope of the constructs. Plates were incubated for 1 hour with the anti-V5 antibody diluted 1:3000 in blocking buffer. The color reactions were developed for 15 min by the addition of o-phenylenediamine substrate and stopped by adding 2.5 M $H_2SO_4$. The optical density (OD) was measured at 490 nm.

### Cloning and sequencing of placental *var2csa* genes

All DBL3X sequences were obtained from cDNA, whereas DBL2X and the overlapping region of DBL4$\epsilon$ and DBL5$\epsilon$ of *var2csa* were cloned from genomic DNA of placental parasites. In brief, parasites were dissolved in Trizol LS (Invitrogen) and RNA was prepared according to the manufacturer's instruction. RNA pellets were dissolved in 10 $\mu$l of RNase-free water and treated with *DNase*I (Sigma-Aldrich, Brøndby, Denmark) for 25 min at RT, followed by 10 min heat inactivation at 65°C. All RNA samples were subsequently tested in real-time PCR for contamination with genomic DNA using a primer set for the housekeeping gene, *seryl-tRNA synthetase*. DNA-free RNA samples were used for synthesis of cDNA by reverse transcriptase (Superscript II, Invitrogen) and random hexamer primers as described by the manufacturer. Following primer sets were used for cloning DBL3X from cDNA into either the *Baculovirus* vector, pAcGP67-A (BD Biosciences) or the pCR2.1-TOPO® vector (Invitrogen): p509 5' CG GAA TTC GAT ACA AAT GGT GCC TGT and p510 3' ATT TGC GGC CGC ATA TAC TGC TAT AAT CTC C, p508 5' CG GAA TTC ACA CAA AAT TTA TGT GTT and p510, p503 5' GAG ATA CAA ATG GTG CCT GT and p505 3' AAA TTT GCT GAT ATA CAT TCA G. PCR products aimed for the *Baculovirus* vector were digested with *Eco*RI and *Not*I before ligation. Three to six colonies of each cloning and corresponding plasmids were sequenced by Macrogen (Seoul, Korea). Genomic DNA was extracted from filter paper using a chelex based method (Pearce et al. 2003). Briefly, filter spots were dissolved in 0.5% saponin in PBS using a microtiter plate and incubated overnight on a shaker at RT. After washing the filter spots twice in PBS, a solution of 50 $\mu$l of H2O and 100 $\mu$l of 10% chelex mixture was added to each well. The plate was boiled for 8 min and subsequently cooled down for 10 min at RT.

A PCR reaction was run with primers for the *seryl-tRNA synthetase* gene to control for the DNA content. Around 1-3 $\mu$l of DNA was used for the PCR reactions amplifying the different *var2csa* regions. All PCR products were cloned into the pCR2.1-TOPO® vector and the inserts sequenced on a 3100-*Avant* Genetic Analyzer (Applied Biosystems). The origin of parasites are described in (Tuikue Ndam et al. 2004). All sequence data are available at GenBank accession numbers DQ995590-DQ995632.

### Phylogenetic reconstruction

The alignment of 43 placental and 4 database VAR2CSA DBL3X sequences, covering the 3D7 amino acid positions 1256 to 1549, was constructed using the software RevTrans (Wernersson and Pedersen 2003), and subsequently manually corrected for errors. To cover more of the DBL3X domain, an alignment of 17 database sequences was constructed in the same way, covering the 3D7 positions 1217 to 1255. For the phylogenetic tree, 21 Malawian sequences (Duffy et al. 2006) were aligned with the sequences mentioned above, again using RevTrans and manual correction, resulting in a 609 bp alignment. The program MrModeltest version 2.2 (Nylander 2006) was used to find the most appropriate nucleotide substitution model based on the Akaike information criterion (Posada and Buckley 2004). Phylogenetic trees based on the above alignments were constructed by Bayesian inference using the program MrBayes version 3.1.1 (Ronquist and Huelsenbeck 2003). In all cases Markov chain Monte Carlo (MCMC) sampling was performed for 10,000,000 generations with 8 chains. Convergence was confirmed by comparing the results of two independent runs. Burn-in was determined using Tracer (Rambaut and Drummond 2006) and 50 % majority rule consensus trees were constructed.

### Model fitting and Akaike weighted dN/dS average

The program codeml from the PAML package version 3.14 (Yang 1997) was used to fit a range of codon-based evolutionary models to the VAR2CSA DBL3X region using the alignments and Bayesian trees mentioned above. Eleven codon-based models were tested using codeml: M0, M1a, M2a, M3 (with either 3, 4, 5, 6, or 7 site categories), M5, M7 and M8 (Goldman and Yang 1994; Nielsen and Yang 1998; Yang et al. 1998). All 11 models were fitted using the F3x4 (different nucleotide frequencies for each codon position) approach for estimating codon frequencies. Convergence was confirmed by comparing the results of several independent runs started with different parameter vectors. The Akaike information criterion (AIC) was used to assess model fits (Posada and Buckley 2004; Akaike 1973; Burnham and Anderson 2002). Briefly, $AIC$ estimates the expected relative Kullback-Leibler distance (i.e. $AIC$ is an estimate of the amount of information that is lost when a given model is used to approximate the full truth). $AIC$ is a function of the maximized log-likelihood ($lnL$) and the number of estimated parameters ($K$) for a model. Specifically, $AIC = -2lnL + 2K$ where lower $AIC$ values indicate better models. From $AIC$ it was possible to compute Akaike weights, which can be interpreted as the conditional probability of the model given the data and the set of initial models (Posada and Buckley 2004; Burnham and Anderson 2002). On this basis, dN/dS ratios for codon positions were calculated as an average of the dN/dS ratios estimated from

each of the eleven models, weighted by the Akaike weights for the corresponding model.

### Recombination and mutation rates, diversity and sequence logo creation

The population recombination rate $\rho$ was estimated for the VAR2CSA DBL3X domain in LDhat version 2.0 (McVean et al. 2004), which based on population genetics uses coalescent methods specially adapted to account for the possibility of recurrent or back mutation and for an AT-rich genome such as that of *P. falciparum* (McVean et al. 2002). As argued by Mu et al. (Mu et al. 2005), the coalescent recombination estimate can for partially inbreeding haploid species such as *P. falciparum*, be interpreted as the compound parameter $\rho = 2N_e r(1-f)$, where $N_e$ represents the effective population size of the DBL3X population, $r$ denotes the rate of recombination cross-over events per generation per bp and $f$ is the inbreeding coefficient. The effective population size should be thought of as the size of an ideal population (McVean et al. 2002; Hudson 2001) with the same magnitude of random genetic drift as the actual population with size $N$ (Hartl and Clark 1997). To test if the placental DBL3X sequence data showed evidence for deviation from the neutral model of evolution assumed by the coalescent method, we calculated Tajima's D statistic (Tajima 1989) and Fu and Li's D* and F* statistic (Fu and Li 1993). All three statistics were insignificantly different from zero ($p>0.1$ in all cases), indicating that the coalescent approach could be applied. The hypothesis of no recombination was rejected using the likelihood permutation test (McVean et al. 2002) with 1000 permutations of segregating sites, of which none produced a higher maximum composite likelihood than for the DBL3X data. The hypothesis of a constant recombination rate across the analyzed region was also rejected ($p = 0.048$ with 10,000 simulations) using the method described in (McVean et al. 2004), indicating significance in the recombination rate variations over DBL3X. For calculation of the population recombination rate $\rho$, we used the Bayesian reversible-jump Markov chain Monte Carlo (RJMCMC) method with a block penalty of 10, running for 10,000,000 iterations with 2,000 iterations per sample and a burn in of 50 samples. The overall region estimate was converted to base pair units using the average length of the analyzed sequences. Recombination hotspots were defined as intervals containing SNPs where the population recombination rate mean was above the upper limit of the 95% confidence interval for the overall region estimate of $2N_e r(1-f)$. Fu and Li's D were calculated using DnaSP version 4.10.6 (Rozas et al. 2003). The average nucleotide diversity and its variance were calculated according to Nei (Nei 1997) equation 10.5 and 10.7, respectively (gapped columns were included). The Kullback-Leibler sequence logo was created by calculating the distance between the two groups of sequences for each amino acid position in the alignment using the symmetric Kullback-Leibler distance:

$$D_{KL} = \sum_{AA} (p - p') \times \log\left(\frac{p}{p'}\right)$$

where $p$ and $p'$ are the frequencies of an amino acid type in each of the two groups, and AA indicates that the sum is over all the amino acid types. The

cumulated Kullback-Leibler distance was calculated as the sum of $D_{KL}$ for all positions in the alignment. Gaps in the alignment were in this analysis assigned the letter "O" and treated as an amino acid class. To test if the mentioned grouping according to parity gave two significantly different sequence groups, we created 10,000 random groupings and for each of these summed the DKL over all amino acid positions. Similarly for the individual positions in the logo, the distribution of $D_{KL}$ for 10,000 random shuffles of sequence grouping was noted specifically for each position, and the $p$-values were based on these distributions.

### Pepscan motif analysis

442 overlapping 31mer peptides covering the exon1 of 3D7 VAR2CSA were synthesized as Solid Phase Peptide Synthesis (SPPS) with a stepwise addition of the different amino acids attached to a solid resin. The long peptides were synthesized with a cysteine at aa position 15 allowing some secondary structure. This approach allows identification of antigenic sites that cannot be mapped using short, linear peptides (Pepscan systems, the Netherlands). The raw data from a pepscan experiment consists of figures measuring the amount of IgG bound to each of the overlapping 31 mer peptides. We used the overlap in primary sequence to determine more specifically what the antibodies have affinity for. The motif analysis is based on the concept that polyclonal IgG response consists of subpopulations of monoclonal antibodies each binding a certain set of amino acid sequence motifs. We then made a list of all possible binding motifs and transferred the information from the peptide array to these, giving each motif a score indicating the IgG affinity for the motif. The pepscan assay is designed primarily to determine linear epitopes, and thus we are mainly interested in short binding motifs and gaps. On this basis we performed the pepscan motif analysis, using motifs containing either 2 or 3 defined residues spaced by undefined amino acids up to a certain maximal length of 5, 7, 10, 15, 20 or 25 residues. We found that the different maximal motif lengths gave similar results, but with different detail resolutions (long motifs had a smoothing effect), and a binding motif with a maximal length of 7 residues was selected as being most informative. Thus, the presented results are based on the assumption that motifs are 2 - 7 amino acids long, and contain either 2 or 3 defined residues spaced by undefined amino acids, e.g. a possible motif could be "WXXXDXE" or simply "KN". The method was validated by comparison to the raw data, where rabbit serum from a rabbit immunized with a DBL5 construct was used in the pepscan assay (Supplementary figure 3.10). The figure shows that the motif analysis produces peaks approximately at the same positions as in the raw data, and that the analysis does not introduce bias in the other regions of the protein. As control for the human IgG used in Figure 6 we used non-immune Dutch serum as well as a malaria exposed nulliparous woman.

### Affinity purification of antibodies and depletion of serum on parasites

Affinity purification of antibodies was done according to manufacturers instructions. In brief, 1 mg of recombinant protein was dialyzed against 0.2 M $NaHCO_3$, 0.5 M NaCl, pH 8.3 and applied to a NHS-activated HiTrap 1 ml column (GE Healthcare) that had been equilibrated with 3×2 ml 4°C 1 mM HCl. After coupling the columns were washed with 0.5 M ethanolamine 0.5

M NaCl, pH 8 and 0.1 M acetate, 0.5 M NaCl, pH 4 and a final wash with PBS pH 7.4. One ml of a plasma pool (28 women from Ghana) was then applied to the column. After washing in 10 ml PBS, affinity bound antibodies were eluted by $CH_3COONH_4$, pH 3 and neutralized in 1 M Tris pH 7.5. The specificity of the purified antibodies was tested in ELISA against 1) the domain used for affinity purification, 2) other VAR2CSA domains, and 3) glutamine rich protein (GLURP) (Dodoo et al. 2000). Affinity purified antibodies used for pepscan analysis were all negative in ELISA against control proteins and positive against the homologous domain (data not shown). Surface reactive antibodies in a pool of pregnancy plasma (from 15 pregnant women from Korogwe, Tanzania) were depleted using a parasite line selected for VAR2CSA expression using VAR2CSA specific antibodies (Salanti et al. 2004). In brief, 40 $\mu$l of the plasma pool were incubated with $2.0\times10^8$ MACS purified intact late stage trophozoite and schizont infected erythrocytes for 20 min at 4 °C. Hereafter, the cells were centrifuged at 800 *g* for 8 min, and the supernatant used to suspend a new pellet of $2.0\times10^8$ infected erythrocytes. This procedure was repeated four times. The depletion of surface reactive antibodies from the plasma pool was confirmed using a flow cytometry assay (Salanti et al. 2004) after the final depletion.

### 3D modelling of the 3D7 DBL3X domain

The three-dimensional structure of the 3D7 sequence (PFL0030c aa 1217 to 1559) was modeled using the HHpred server with default settings (Soding et al. 2005). Briefly, the HHpred method is specialized in remote homology detection using hidden Markow models (HMMs) build from PSI-BLAST profiles and secondary structure. The crystal structure of EBA-175 F1 (PDB code 1ZRO chain A, (Tolia et al. 2005)) was used as template and had the highest sequence and secondary structure alignment scores. The HHpred alignment was corrected in a short template loop sequence (Supplementary figure 3.9, positions 215-219) positioned next to a gap. The correction shifted the position of the gap and allowed for the modeling of a disulfide bridge in DBL3X, which was conserved in the EBA-175 F1, F2 and Pk$\alpha$-DBL domains. HHpred HMMs for DBL3X and the template continued to match. Finally the corrected alignment was used to generate a three-dimensional model using MODELLER (Sali et al. 1995) with the protocol setup in the HHpred server toolkit. A superimposition of the EBA-175 F1 structure and the DBL3X model was obtained by the HHpred toolkit. NACCESS (Hubbard and Thornton 1993) was used to calculate relative surface exposed areas (RSAs) in single chains of EBA-175 F1, F2 and the Pk$\alpha$-DBL domain (Singh et al. 2006). The MAMMOTH-mult alignment server (Lupyan et al. 2005) was used to make a multiple structure superimposition of DBL3X model on the EBA-175 F1 and F2 DBL domains (Tolia et al. 2005) and the Pk$\alpha$-DBL domain (Singh et al. 2006). The resulting alignment was inspected to identify conserved positions of cysteines and buried hydrophobic residues (RSA $< 30\%$). Structural visualizations were made using PyMol (DeLano 2002).

## Acknowledgments

We acknowledge the women who donated serum and parasites for this study. We thank Gilean McVean for advice regarding recombination analysis, Kristoffer

Rapacki for help with logo generation and Anne Corfitz for excellent technical assistance. This study was primarily funded by the EMVI grant #0012-2004. This study also received financial support from the Danish Medical Research Council (SSVF) (grants 22-02-0220 and 22-03-0333), and the Danish Research Council for Development Research (RUF) (grant 104.Dan.8.L.306 and 8.L.306) and the Danish National Research Foundation (Danish Platform for Integrative Biology). Salanti is supported by a postdoctoral grant from SSVF. Dahlbäck is supported by a Ph.D. studentship from RUF. Nielsen is supported by a postdoctoral grant from Hovedstadens Sygehusfaellesskab. Tuikue Ndam is supported by a postdoctoral grant from the Fondation Recherche Médicale. Pepscan systems (The Netherlands) are thanked for technical assistance on the data work on the serum pepscan data.

**Competing interests** The authors have declared that no competing interests exist.

# Supplementary figures to Paper II

```
                        10        20        30        40        50        60        70        80
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
DBL3X           YIRGCQPKIYDGKIFPGKGGEKQWICKDTIIHGDTNGACIPPRTQNLCVGELWDKRYGGRSNIKNDTKESLKQKIKNAIQ
EBA-175 F1      VLSNCREK..........RKGMKWDCKKK..NDRSNYVCIPDRRIQLCIVNLAII........KTYTKETMKDHFIEASK
EBA-175 F2      SKFGCDKNS.......VDTNTKVWECKKPYKLS.TKDVCVPPRRQELCLGNIDRI........YDKNLLMIKEHILAIAI
Pkalpha DBL     ...KCNDKR........KRGERDWDCPAE......KDICISDRRYQLCMKELTNL........ITFLKLNLKRKLMYDAA

                         90       100       110       120       130       140       150       160
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
DBL3X           KETELLYEYHDKGTAIISRNPMKGQKEKEEKNNDSNGLPKGFCHAVQRSFIDYKNMILGTSVNIYEYIGKLQEDIKKIIE
EBA-175 F1      KESQLLLKK.........N...............DNKYNSKFCNDLKNSFLDYGHLAMGNDMDFGGYSTKAENKIQEVFK
EBA-175 F2      YESRILKRKY........K................NKDDKEVCKIINKTFADIRDIIGGTDYWNDLSNRKLVGKINTNS.
Pkalpha DBL     VEGDLLLKK.........N...............NYQYNKEFCKDIRWGLGDFGDIIMGTNMEG...VENNLRSIFGTDE

                        170       180       190       200       210       220       230       240
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
DBL3X           KGTTKQNGKTVGSGAENVNAWWKGIEGEMWDAVRCAITKINKKQKKNGTFSIDECGIFPPTGNDEDQSVSWFKEWSEQFC
EBA-175 F1      GAHGEISEH...KIKNFRKKWWNEFREKLWEAMLSEHKNNI.............CKNIP...QEELQITQWIKEWHGEFL
EBA-175 F2      .NYVHRNKQ...NDKLFRDEWWKVIKKDVWNVISWVFK............DKTVCKEDDI..ENIPQFFRWFSEWGDDYC
Pkalpha DBL     .....KA.....KQ..DRKQWWNESKEHIWRAMMFSLRS......RLKEKFVWICKKDV.....PQIY.RWIREWGRDYM

                        250       260       270       280       290       300       310       320
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
DBL3X           IERLQYEKNIRDACTN....NGQGDKIQGDCKRKCEEYKKYISEKKQEWDKQKTKYENKY...........VGKSASDLL
EBA-175 F1      LERDNRAKLPKSKCKN....NALYEACEKECIDPCMKYRDWIIRSKFEWHTLSKEYETQK..........VPKENAENYL
EBA-175 F2      QDKTKMIETLKVEC...........CEDDNCKRKCNSYKEWISKKKEEYNKQAKQYQEYQKGNNYKMYSEFKSIKPEVYL
Pkalpha DBL     SELPKEQGKLNEKCASKLYYNNMAICMLPLCHDACKSYDQWITRKKKQWDVLSTKFSSVKK.........TNIATAYDIL

                        330       340       350       360
                ....|....|....|....|....|....|....|....|.
DBL3X           KENYPECISANFDFIFNDNIEYKTYYPYGDYSSICSC....
EBA-175 F1      IKISENKNDAKVSLLLNNC........DAEYSKYCDCK...
EBA-175 F2      KKYSEKCSNLNFEDEFKE.......ELHSDYKNKCTMCPEV
Pkalpha DBL     KQELN...GFKEATFENEINK.....RDNLYNHLCPCV...
```

Figure 3.9: **Multiple Structural Alignment of VAR2CSA DBL3X, EBA-175 F1, EBA-175 F2, and Pk$\alpha$-DBL.** Cysteines in conserved disulfide bonds are highlighted in blue, conserved tryptophans buried in the structures are shown in red, and other conserved, buried, and hydrophobic positions are shown in green.

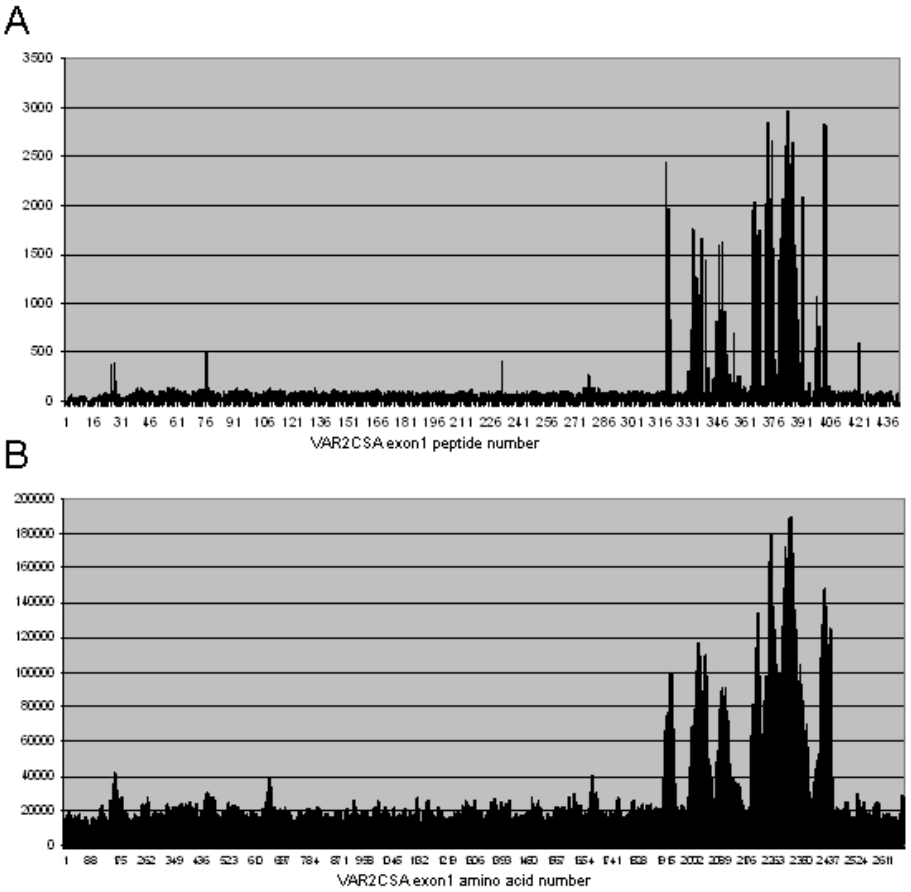


Figure 3.10: **Pepscan Analysis of Rabbit Serum Immunized with a DBL5 Recombinant Construct.** (A) Plasma from a rabbit immunized with a DBL5 construct was tested in 1:1000 on the VAR2CSA pepscan array. The upper diagram is the raw pepscan values with the pepscan score (y-axis) for each of the 420 peptides (x-axis). (B) The same PepScan scores as (A) but calculated using a motif algorithm assigning a value to each amino acid.

## 3.6 Paper III

*Submitted to PLoS Pathogens*

# Structural insight into epitopes in the pregnancy-associated malaria protein VAR2CSA

**Pernille Andersen[1], Morten A Nielsen[2], Thomas S Rask[1], Madeleine Dahlbäck[2], Thor Theander[2], Ole Lund[1], Ali Salanti[2*]**

[1]Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark.[2]Centre for Medical Parasitology at University of Copenhagen and Copenhagen University Hospital (Rigshospitalet), Denmark.

[*] *Corresponding author:*
Tel: (+45) 35 32 76 76
Fax: (+45) 35 32 78 51
Email: salanti@cmp.dk

## Abstract

**Pregnancy associated malaria is caused by *P. falciparum* malaria parasites binding specifically to chondroitin sulfate A (CSA) in the placenta. This sequestration of parasites is a major cause of low birth weight in infants and anemia in the mothers. VAR2CSA is the main parasite ligand for CSA-binding, and identifying epitopes of protective antibodies is essential for VAR2CSA vaccine development. VAR2CSA is a multi-domain protein and vaccine research has mainly focused on individual domains. In this study we propose 3D models for each of the VAR2CSA DBL domains. Additionally, we show that regions of ID2 and a PfEMP1 CIDR domain could be homologous to the EBA-175 and Pk$\alpha$-DBL fold, indicating that ID2 could be a functional domain. We identified regions of VAR2CSA present on the surface of the native VAR2CSA by comparing reactivity of plasma containing anti-VAR2CSA antibodies in peptide array experiments before and after incubation with native VAR2CSA. When these data were mapped onto the DBL models it was clear that the S1+S2 DBL sub-domains were surface-exposed in most domains, whereas sub-domains S3 seemed to be less exposed in native VAR2CSA. These results comprise an important step towards understanding the structure of VAR2CSA on the surface of CSA binding infected erythrocytes.**

**Keywords:** VAR2CSA, DBL, PfEMP1 Malaria, Pepscan, Pregnancy, Structure modeling

# Introduction

The invasion of erythrocytes and the subsequent adhesion of parasite-infected erythrocytes (IE) to vascular endothelium or placenta are key events in the asexual lifecycle of *Plasmodium falciparum* and thus of major importance for virulence of this parasite. Erythrocyte invasion is mediated by proteins belonging to the Erythrocyte Binding Ligand family (EBL) and in *P. falciparum* the Erythrocyte Binding Antigen (EBA)-175 is the best described EBL protein. EBA-175 contains two Duffy binding like (DBL) domains (called F1 and F2) which binds to the heavly sialated glycophorin A on the erythrocyte surface (Orlandi et al. 1992). The monomeric structure of EBA-175 was determined by X-ray crystallography and the primary feature of the two DBL domains were $\alpha$-helices and an anti-parallel $\beta$-hairpin (Tolia et al. 2005). EBA-175 also crystallized as a dimer, and the structure of this complex showed that the DBL domains of EBA-175 interacted in a reverse handshake orientation (Tolia et al. 2005). The simian malaria parasite, *Plasmodium knowlesi* invade erythrocytes through the host receptor "Duffy antigen receptor for chemokines" (DARC) (Miller et al. 1975). This interaction is also mediated by a parasite-encoded DBL containing protein, Pk$\alpha$-DBL, and the crystal structure of Pk$\alpha$-DBL has been shown to be very similar to PfEBA-175 despite of extensive sequence variation (Singh et al. 2006). Based on the structure of Pk$\alpha$-DBL the DBL domain could be divided into three sub-domains named S1-S3 which are connected by short linkers. Both glycan binding and DARC binding are predominantly located in S1 and S2 (Tolia et al. 2005; Singh et al. 2006).

Adhesion of IE to the vascular bed is thought to be a mechanism developed by the parasite to avoid being filtered through the spleen, where erythrocytes infected with late stage asexual parasites are removed from the circulation (Weiss 1990). The adhesion is mediated by *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), which interacts specifically with receptors on the vascular endothelium or placenta (Baruch et al. 1995; Salanti et al. 2004). Antibodies targeting PfEMP1 and abrogating binding are believed to be important mediators of acquired malaria immunity (reviewed in (Kraemer and Smith 2006).

Pregnancy-associated malaria (PAM) is caused by *P. falciparum* sequestering in the placenta by binding to chondroitin sulfate A (CSA) on syncytiotrophoblasts (Fried and Duffy 1996). Women suffering from PAM develop antibodies which protect them and their offspring during subsequent pregnancies (Duffy and Fried 2003). These protective antibodies are thought to recognize a relatively conserved antigen as plasma and parasites from pregnant women from different malaria areas cross-react (Fried et al. 1998). The PfEMP1 variant mediating placental binding was recently discovered and named VAR2CSA (Salanti et al. 2003; Salanti et al. 2004). The extra-cellular part of VAR2CSA consists of six Duffy Binding Like (DBL) domains, a large inter-domain (ID2) and a C-terminal region predicted to be cytoplasmic. Most PfEMP1 molecules, but not VAR2CSA, contain two cysteine-rich interdomain regions (CIDR domains) (Gardner et al. 2002; Smith et al. 2000). Some CIDR domains bind to CD36 (Smith et al. 1998) and they have been described as degenerated DBL domains (Su et al. 1995).

With the aim of making a vaccine that can reverse or inhibit parasite binding in the placenta, much effort is put into defining the specific part/parts of VAR2CSA that bind to CSA (reviewed in (Gamain et al. 2007)). The best way

of determining this interaction would be to produce the extra-cellular part of VAR2CSA and co-crystallize this multidomain protein with CSA. However, it is very difficult to express such a large protein and previous attempts to crystallize even single VAR2CSA DBL domains have failed. Thus, novel methods are required to generate models and hypotheses on the 3D structure of PfEMP1 molecules. The DBL domains of PfEMP1 are often illustrated as pearls on a string, and vaccine development strategies are focusing on the VAR2CSA DBL domains as single entities in a larger protein. We have recently published data showing a structural model of VAR2CSA DBL3X and mapped areas of DBL3X that are surface exposed on the native protein (Dahlback et al. 2006).
In this study, we modelled the remaining five VAR2CSA DBL domains, the VAR2CSA inter-domain 2 (ID2) and a number of CIDR domains from different PfEMP1 molecules. The models indicate that DBL domains contain features that are structurally conserved. Furthermore it appears that there is homology between the ID2, CIDR and part of the resolved structures of the DBL fold. By absorbing antibodies on native VAR2CSA on the surface of infected erythrocytes and comparing antibody reactivity on a VAR2CSA peptide array before and after absorption, we define which areas of the VAR2CSA molecule are accessible to antibodies in the native protein. Interestingly, the surface exposed epitopes on the six VAR2CSA domains are largely found within S1 and S2, whereas S3 appears to be hidden in the complete VAR2CSA structure. Based on these data we discuss the domain architecture of VAR2CSA and suggest a model where the protein is surface-exposed as a globular or multimerized structured protein stabilized by long $\alpha$-helices in the S3 region.

# Results and Discussion

### Modeling of the VAR2CSA DBL domains

The 3D structures of the 3D7 VAR2CSA DBL domains (Figure 3.11) were modeled using the HHpred server (Soding et al. 2005) developed for low-homology modeling. HHpred template searches confirmed that the determined structures of the *P. falciparum* EBA-175 DBL domains F1 and F2 and the *P. knowlesi* DBL domain Pk$\alpha$-DBL (Singh et al. 2006; Tolia et al. 2005) could be used

| VAR2CSA DBL domain | Modeling template | Template sequence identity | Insertions relative to template | Missing info inserts* | Gaps relative to template |
|---|---|---|---|---|---|
| DBL1 | Pka-DBL | 16.4% | 10 | 4 | 2 |
| DBL2 | EBA-175 F1 | 17.2% | 10 | 1 | 0 |
| DBL3 | EBA-175 F1 | 20.7% | 6 | 1 | 1 |
| DBL4 | Pka-DBL | 17.8% | 9 | 4 | 1 |
| DBL5 | EBA-175 F1 | 18.8% | 8 | 1 | 2 |
| DBL6 | Pka-DBL | 18.7% | 8 | 4 | 1 |

*Inserts in regions with missing structural information in the template structure.

Table 3.2: **Modeling details of the VAR2CSA DBL domains**

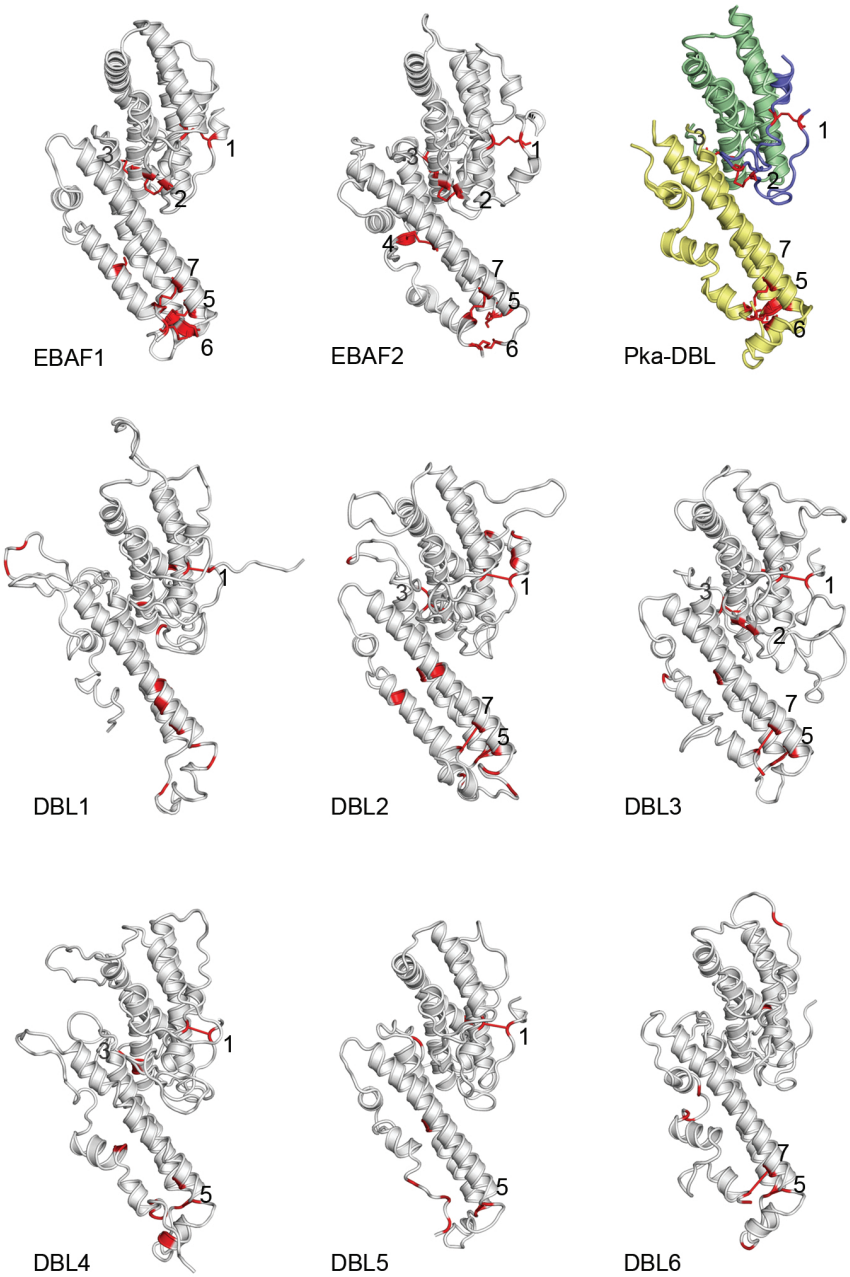


Figure 3.11: **Structure models of the VAR2CSA DBL domains** The experimental structures of the template EBA-175 and Pk$\alpha$-DBL domains are shown in the first row, and DBL models are shown in the middle and last rows. Sub-domains 1-3 of Pk$\alpha$-DBL (Pka-DBL) are colored in blue, green and yellow, respectively. Cysteines in all domains are highlighted in red. Disulfide bonds in templates are numbered according to occurrence in the sequences. Cysteine pairs in the models which are aligned with disulfide bonding cysteines in templates are numbered with respect to the templates.

as templates for modeling (HHpred probability scores were all 100%), although the sequence identity between the VAR2CSA DBL domains and templates was between 16-20% (see modeling details in Table 3.2). The determined structures of the two EBA-175 DBL domains each contain a region where structural information is missing and the structure of Pk$\alpha$-DBL has four such regions (Table 3.2 and Figure 3.12). In crystallographic experiments, the local occurrence of missing structural information indicates regions of flexibility and loosely defined secondary structure. In Pk$\alpha$-DBL, regions of missing structural information were used to divide the DBL fold into the subdomains S1-S3 (Tolia et al. 2005) and we adapted a similar classification for the VAR2CSA DBL domains (Figure 3.11 and 3.12). VAR2CSA residues corresponding to regions of missing structural information in the templates were modeled as insertions relative to the template structure. The structure of such inserted regions is difficult to predict correctly (Ginalski 2006). Secondary structure predictions using the PSIPRED method (Jones 1999) is part of the HHpred modeling protocol. The prediction results are divided into helix, strand and coil, where the coil class consists of secondary structure types mostly found in loops. Interestingly, the template regions of missing structural information aligned with VAR2CSA DBL sequences predicted to have coil secondary structure. In general, the VAR2CSA sequence variation is high within these regions (Figure 3.12). Taken together, this suggests that these VAR2CSA DBL regions form a variety of flexible loops of different structures.

### Evaluation of the VAR2CSA DBL structure models

The quality of a structure model obviously has a pronounced effect on the information that can be deduced from the model. We used the automated structure analysis tool ANOLEA (Melo and Feytmans 1998) for evaluation, and Z-scores ranging from 5.00 to 9.61 with 52-68% high-energy residues were obtained. The results indicated that the quality was lowest in the loop regions. Analysis using Verify3d (Eisenberg et al. 1997) resulted in a similar conclusion (data not shown). Since these results did not convince us that the models were correct, we decided to further investigate the quality of the models by inspecting them for conserved residues stabilizing the determined structures of EBA-175 and Pk$\alpha$-DBL domains.

Structural alignment of the EBA-175 and Pk$\alpha$-DBL domains identified a number of positions, which can be assumed to be important for the stabilization of the DBL fold. Thirty-four buried conserved positions of hydrophobic residues, 10 buried conserved positions of polar/charged residues, four helix-capping residues and 12 conserved cysteines which form six disulfide bridges were identified. The positions of these residues were distributed throughout the domains in blocks. We then made a multiple structural alignment including the six VAR2CSA DBL models and the EBA-175 and Pk$\alpha$-DBL structure to identify VAR2CSA residues at the positions corresponding to the positions identified as conserved and stabilizing EBA-175 and Pk$\alpha$-DBL (Figure 3.12). We found that the models have preserved a high number of buried hydrophobic positions, which help to form a hydrophobic core of the DBL domain structure (Table 3.3 and Figure 3.12). The buried polar and charged amino acids, which stabilize template structures by forming hydrogen bonding networks or salt bridges, were also conserved. The models also had conserved helix-capping residues, which stabilize the ends

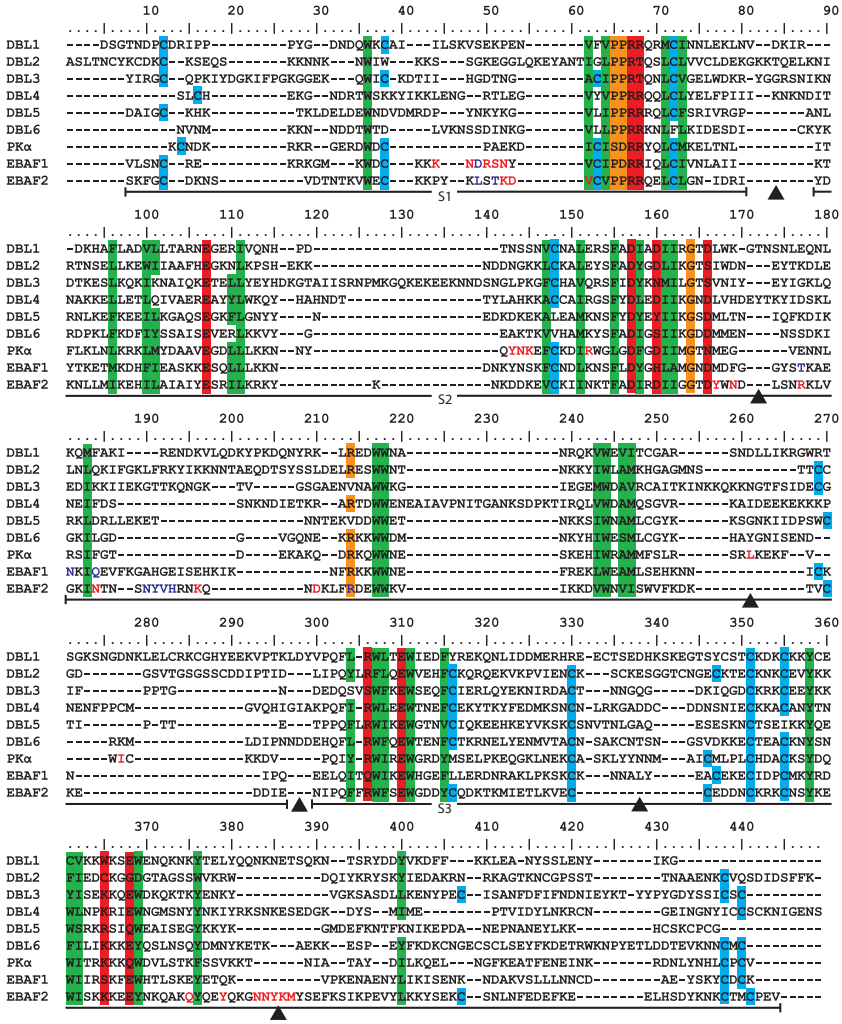


Figure 3.12: **A structural alignment of the template structures EBA-175 F1 and F2 and Pkα-DBL and the six VAR2CSA DBL models.** Vertical color bars denote positions of stabilizing residues in template structures. Blue bars denote cysteines forming disulfide bonds, green bars denote buried hydrophobic residues, orange bars denote helix capping residues and red bars denote buried polar/charged residues. The location of sub-domains S1-S3 are marked by horizontal black bars, and regions of missing structural information in the template structures are marked by black arrow heads. Residues involved in glycosaminglycan (GAG) binding of EBA-175 DBL domains or DARC receptor binding of Pkα-DBL are marked with red letters. EBA-175 residues involved in dimerization are marked in blue letters.

of helices in the template structures. This conservation of stabilizing positions indicates that the alignments of the model sequences to the template sequences are correct in regions surrounding these positions.

The structural alignment and sequence alignments used by HHpred for modeling were analyzed for conservation of cysteines forming disulfide bonds in the

| VAR2CSA DBL domain | Buried hydrophobic | Buried polar/ charged | Helix capping | Cysteines in positions of disulfide bridges* |
|---|---|---|---|---|
| Templates | 34 | 11 | 4 | 12(14) |
| DBL1 | 33 | 10 | 4 | 6 |
| DBL2 | 32 | 8 | 3 | 10 |
| DBL3 | 31 | 9 | 4 | 13 |
| DBL4 | 33 | 11 | 4 | 8 |
| DBL5 | 33 | 9 | 3 | 6 |
| DBL6 | 32 | 11 | 4 | 6 |

*Counted in HHpred alignments used for modeling.
The parenthesis denotes EBA175 F2.

Table 3.3: **Summary of conserved stabilizing positions in models of VAR2CSA DBL domains**

template DBL structures. The models of the VAR2CSA DBL domains all contain conserved cysteine positions likely to form disulfide bonds (Figure 3.11 and 3.12). The disulfide bonds were numbered according to the occurrence of the cysteines in the sequence. Cysteines of disulfide bond 1 are conserved in all models except DBL6. Likewise, the cysteines of disulfide bond 5 are conserved in all models except DBL1. DBL3 is the model with the highest number of conserved cysteines forming disulfide bonds. DBL1 has a high number of conserved cysteines, but many of them lack the partner cysteine to make the disulfide bond. An example is disulfide bond 2, where only one cysteine is conserved (Figure 3.12, positions 38 and 63). A number of cysteines in the models are in close proximity to a disulfide bond-forming template cysteine. This suggests that the local alignments used for the modeling are sub-optimal or that alternative disulfide bonds are formed in the VAR2CSA DBL domains. An example of an alternative bond is the disulfide bond 4 (Figure 3.11, number 4) which is only found in the EBA-175 F2 DBL (Figure 3.12, positions 316 and 407). All VAR2CSA DBL domains except DBL1 have cysteines aligned in position 316. Only VAR2CSA DBL3 has a cysteine aligned to 407, but the models of DBL2 and DBL6 have cysteines in the proximity and it is likely that they form disulfide bonds with the conserved cysteine in the native structure. Therefore we suggest that this disulfide bond is common among VAR2CSA DBL domains, although not present in EBA-175 F1 and Pk$\alpha$-DBL. Another interesting example is the disulfide bond 2 (Figure 3.11, number 2 and Figure 3.12, positions 38 and 64). The disulfide bond is proximal to a region containing glycan binding amino acids in the EBA-175 F1 and F2 domains (Figure 3.12, positions 44, 48, and 50-53) and it may play a role for the function of these domains. Among the VAR2CSA DBL domains, only DBL3 has both cysteines conserved. None of the other VAR2CSA DBL sequences have two cysteines in the proximity, and it is thus unlikely that the apparent variation stems from incorrect alignments. The lack of the disulfide bonds in some regions of the VAR2CSA DBL domains may suggest higher flexibility and a more dynamic structure than in DBL domains stabilized by a higher number of disulfide bonds.

The analysis of different types of stabilizing characteristics shows that these to

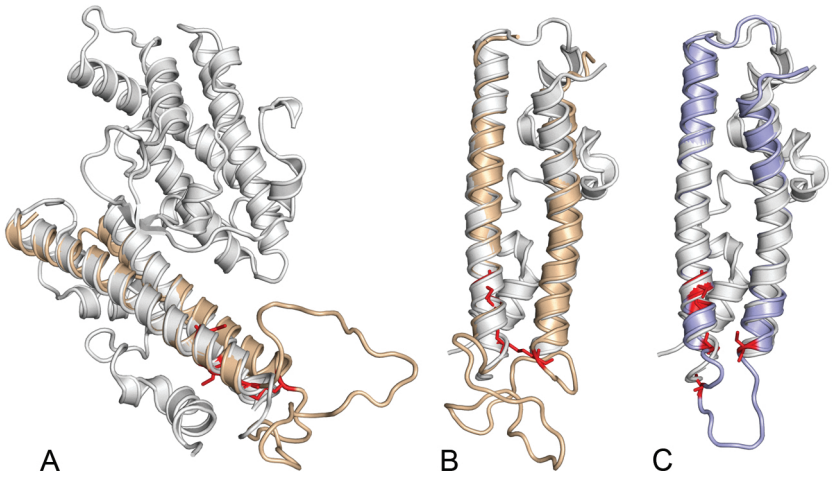


Figure 3.13: **Modeled regions of ID2 and CIDR domains.** The S3 domain of the template EBA-175 F2 is shown in white. Cysteines in the models are highlighted in red. (A) The model of the VAR2CSA ID2 domain superimposed in yellow on the EBA-175 F2. (B) Similar to (A) except showing only the S3 domain. (C) Model of a CIDR domain.

a large extent are conserved between the template and the VAR2CSA models. Since our aim was to map experimental data onto the DBL models, rather than analyzing the structural conformations in detail, we concluded that the models were of sufficient quality for mapping of these data.

### Modeling of VAR2CSA inter-domain 2

The extra-cellular part of VAR2CSA consists of six DBL domains and a sequence stretch consisting of 337 amino acids named inter-domain 2 (ID2) (Gardner et al. 2002). This part of the molecule has attracted little attention and has been viewed as an inter-domain spacer sequence. We analyzed the ID2 sequence (PFL0030c positions 879-1216) for homology to other proteins using the HHpred search and alignment tool. Interestingly, the EBA-175 and Pk$\alpha$-DBL domains were all identified as homologous to the ID2 domain with very high HHpred probability scores (Probabilities = 98.3-99.9%). The similarity was pronounced in a region of 100 residues (PFL0030c positions 1017-1116), which aligned to the first two $\alpha$-helices in S3 with a sequence identity of 19.8%. Using the EBA-175 F2 DBL domain as template, we modeled the structure of ID2 positions 1017-1116 (Figure 3.13A and B). The secondary structure of the whole ID2 domain was then predicted using PSIPRED (data not shown). The ID2 region subsequent to the DBL-homologous region was predicted to consist of a short $\beta$-sheet and four $\alpha$-helices interspersed with loop regions. In the EBA-175 and Pk$\alpha$-DBL structures, these sequences consist of $\alpha$-helices interspersed with a loop region. The ID2 region preceding the DBL homologous region was predicted to contain six $\beta$-strands and two $\alpha$-helices. These data suggest that the C-terminal part of ID2 form a structure similar to DBL S3, but does not support the notion that the N-terminal part of ID2 has a fold similar fold to DBL S1 and S2.

In most PfEMP1 molecules the first DBL domains are separated by a cysteine-rich inter-domain region (CIDR1) on which there is no structural data available. Since VAR2CSA ID2 and CIDR1 domains are placed at the same position in the sequence, we examined CIDR1 sequences representative for the three subgroups CIDR-$\alpha$, $\beta$ and $\gamma$. Using the HHpred server, homology between all CIDRs and EBA-175 was identified. Similarly to the ID2 domain, the homology was detected in the first two helices of the DBL S3 and a structure model was made using the EBA-175 F2 DBL as template structure (Figure 3.13B). The homologous region was part of the CIDR M2 region defined by Smith et. al (Smith et al. 2000). The secondary structure of the CIDR region preceding the DBL homologous region was predicted to contain three $\beta$-strands and four $\alpha$-helices; the subsequent region was predicted to contain six $\alpha$-helices (data not shown). These predictions suggest that like ID2, the C-terminal part of CIDR1 forms a structure similar to that of the DBL S3.

### Mapping surface-expressed continuous epitopes on VAR2CSA

In rational PAM vaccine design, it is important to establish which parts of native VAR2CSA are accessible to antibodies acquired by women who have developed immunity to pregnancy-associated malaria. To gain such information we absorbed anti-VAR2CSA IgG from pools of plasma using VAR2CSA-expressing infected erythrocytes and compared the antibody reactivity of the pools in a VAR2CSA peptide array before and after absorption. Based on prior knowledge of high levels of anti-VARCSA IgG two pools were made using plasma from 32 and 10 Tanzanian pregnant women, respectively. One of the plasma pools (Human serum pool 1) was depleted using erythrocytes infected with 3D7 parasites. From the other pool (Human serum pool 2), two depleted plasma pools were generated by depleting one part with erythrocytes infected with 3D7, and depleting the other part on erythrocytes infected with the FCR3 parasite line. The antibody assays were performed on an array containing 442 overlapping 31mer peptides corresponding to the extra-cellular part of VAR2CSA based on the sequence of 3D7.

Before absorption, the two pools contained antibodies targeting all the peptides (data not shown) but some regions distributed in different parts of the domains showed high reactivity, which can be seen as peaks in Supplementary figure 3.17 -3.22. The antibody reactivity toward most of the peptides was not affected by the depletion, but depletion on native VAR2CSA consistently removed antibody reactivity against some peptides (Supplementary figure 3.17 -3.22).

In addition to the human plasma pool, a pool of plasma from six rabbits each immunized by a different VAR2CSA DBL domain was absorbed and tested. The pattern of reactivity in the peptide array with this pool before absorption was slightly different from the reactivity obtained with the human plasma pools and this difference was also reflected in the absorption experiments.

These results indicated that all domains including the N-terminal segment contained continuous peptides sequences accessible to antibodies, when the VAR2CSA protein was expressed on the surface of CSA-binding infected erythrocytes. It was difficult however, to detect a pattern for this reactivity between the domains when depicting the reactivity on a string of residues. We therefore went on to visualize the reactivity on the DBL models.

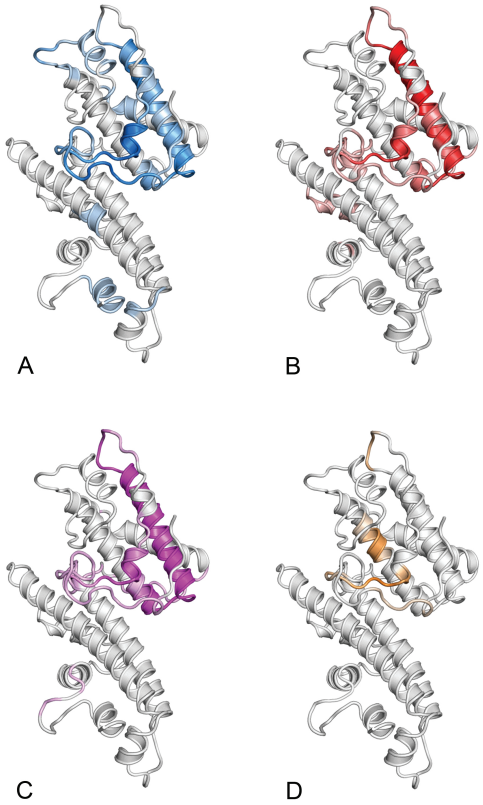


Figure 3.14: **Mapping of surface exposed antibody reactive regions onto the model of the DBL6 domain.** The color intensity denotes the level of antibody depletion when VAR2CSA containing plasma is incubated with parasites expressing native VAR2CSA. The following plasma and parasite combinations were used for the depletion experiments: (A) Human serum pool 1 depleted with 3D7 parasites. (B) Human serum pool 2 depleted with 3D7 parasites. (C) Human serum pool 2 depleted with FCR3 parasites. (D) Rabbit serum pool depleted with 3D7 parasites.

### 3D aspects of surface exposed epitopes in VAR2CSA DBL domains

To assign a value reflecting the accessibility to antibodies of a residue in the native protein we subtracted the depleted reactivity from the non-depleted reactivity. These values, Depletion values (DV) were calculated for each of the four depletion experiments (human pool 1 vs. 3D7 or FCR3, human pool 2 vs. 3D7, rabbit pool vs. 3D7) and mapped onto the six structural DBL models (Supplementary figure 3.23 -3.27). Figure 3.14 shows the results for DBL6 and it is apparent that the depletion experiments using the two human plasma pools identified essentially the same regions as target for surface reactive antibodies (Figure 3.14A versus 3.14B). The results obtained in the absorption experiment using FCR3-infected erythrocytes gave essentially the same results as the experiment using 3D7-infected cells (Figure 3.14A versus 3.14C). Since the peptide array was based on the 3D7 sequence, this indicates that the surface-exposed epitopes on the 3D7 and FCR3 versions of VAR2CSA are cross-reactive or target-

conserved epitopes. The absorption experiment using the rabbit plasma pool only showed depletion of antibodies targeting peptides residing in S1 and S2 (Figure 3.14D). There was no indication of depletion of antibodies targeting S3 in any of the experiments.

To facilitate a comparison between DBL domains, we created consensus DVs for each domain by calculating a sum of normalized DVs from each of the four absorption experiments and scoring the residue as positive if the value was above a fixed threshold (Figure 3.15). Overall there was a reasonable agreement between the models based on the individual absorption experiments and the consensus DVs (Figure 3.14A and B versus Fig. 3.15, DBL6). The results for DBL3 were also in agreement with the surface exposed epitopes identified previously (Dahlback et al. 2006). When evaluating the consensus models it should be kept in mind that surface-exposed VAR2CSA regions can only be detected if the plasma pool contains antibodies against the region and if the antibodies can be detected in the peptide array assay. Thus, regions not targeted by antibodies or targeted by antibodies, which cannot be detected in the peptide array experiment because they either target non-linear sequences or polymorphic sequences not represented in 3D7 VAR2CSA, will not be scored as surface reactive. Likewise, residues buried in the native molecule but residing close to a region representing a surface exposed epitope on the peptides will score as positive.

When comparing the consensus models for the six DBL domains, it was evident that the pattern of reactivity was comparable in DBL domains 1, 2, 3, 5 and 6, whereas the pattern in DBL4 was unique (Figure 3.15). For the former domains the targets of surface reactive antibodies were mainly located in S1 and S2, whereas little reactivity was found in S3, which is located on the lower left side of the models in Figure 3.15. In S1 and S2 both loops and $\alpha$-helices were targets of surface reactive antibodies. The loop between S1 and S2 (Figure 3.12, positions 81-87 and appearing most prominently in the upper left corner of the DBL2 model on Figure 3.15) is flexible in the Pk$\alpha$-DBL structure and we observe a high sequence variation between the DBL domains in the region, suggesting that the corresponding loops in the VAR2CSA DBL domains are correspondingly flexible, reinforcing the possibility that VAR2CSA domains can be divided into sub-domains. In DBL domains 2, 3, 5 and 6, a loop in S2 (appearing most prominently in the lower right corner of DBL3 on Figure 3.15) was also targeted by surface reactive antibodies. A loop region in S1 (Figure 3.12, positions 44-52, appearing most prominently in the center of DBL6 on Figure 3.15) was recognized in DBL domains 1, 2, 3, and 6. Interestingly, the corresponding regions in EBA-175 contain glycan-binding residues. The S2 $\alpha$-helix appearing in the upper right on all models was recognized in all domains but DBL4, whereas some of the $\alpha$-helices in S1 and S2 were recognized to a varying degree.

The regions of DBL4 targeted by surface reactive antibodies differed markedly from the other domains. Relatively more reactivity was detected against S3 and the reactivity against S2 was mainly against regions on opposite site of the domain compared to the other domains (Figure 3.15 and Figure 3.16). The possibility that DBL4 is positioned different from the other domains in the quaternary VAR2CSA structure is in agreement with the finding that the DBL4 specific rabbit antibodies are not reacting with the native VAR2CSA on IE (Nielsen et al. 2007, and unpublished data).

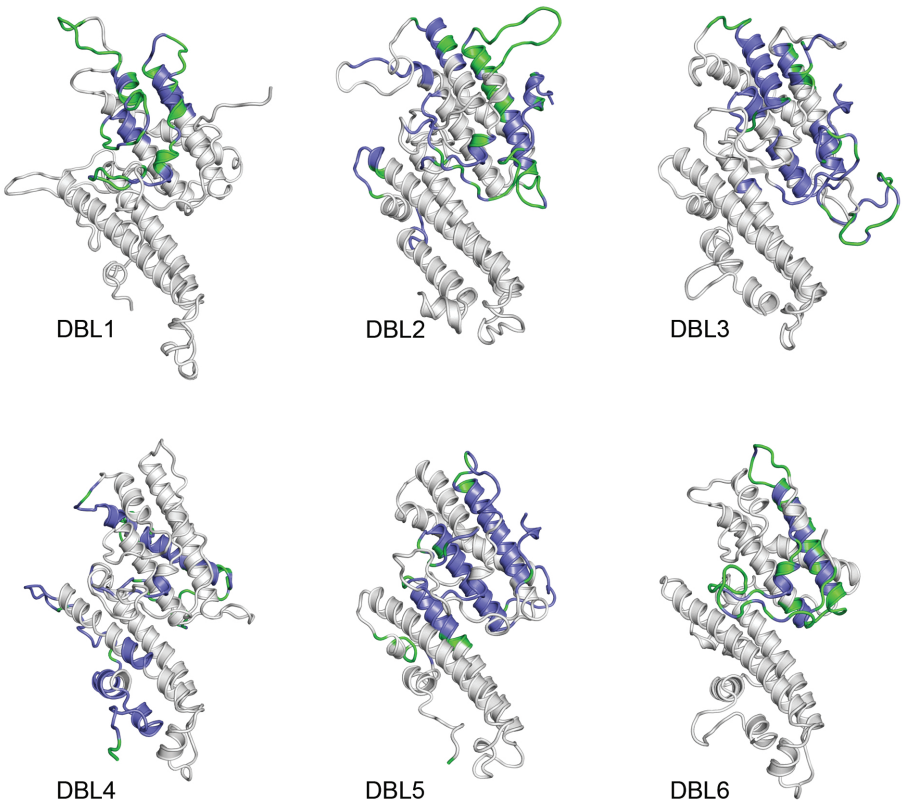


Figure 3.15: **Areas predicted to be targeted by surface reactive antibodies on the six VAR2CSA DBL domains.** The highlighted regions were defined on the basis of the results of several experiments where anti-VAR2CSA antibody containing plasma was incubated with parasites expressing native VAR2CSA. Residues with consensus DVs>1 which are conserved in sequences of VAR2CSA are highlighted in blue and variable residues with consensus DVs>1 are highlighted in green. See text for details.

Using a multiple sequence alignment of seven full-length VAR2CSA sequences, residues were classified as conserved if they were all identical and as polymorphic if any of the sequences showed variation in the particular position. It is apparent from Figure 3.14 and 3.15 that all domains contained both conserved and polymorphic regions targeted by surface reactive antibodies, but the conserved regions were most prominent in DBL3 and DBL5. This is in agreement with data showing that antibodies raised against recombinant proteins representing DBL3 and DBL5 are more likely to cross react with heterologous parasites, than antibodies raised against the other domains (Nielsen et al. 2007). The data is also in agreement with the reactivity of human monoclonal antibodies produced by immortalized B cells from malaria-exposed pregnant women which are directed predominantly against these two domains (Barfod et al. 2007).

Conserved surface-exposed epitopes appear to be attractive vaccine targets. However, protective immunity is acquired through successive pregnancies (Duffy and Fried 2003; Staalsoe et al. 2004), and is a function of transmission intensity

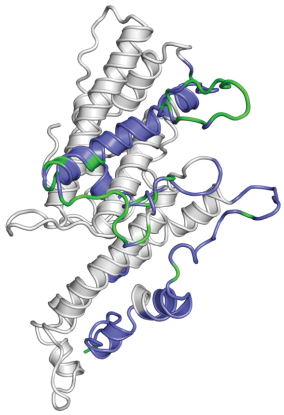

Figure 3.16: **VAR2CSA DBL4 regions predicted to be targeted by surface reactive antibodies.** The plot is similar to Figure 3.14, except viewed from the reverse side.

(McGregor et al. 1983). Natural acquired protection could therefore depend on the ability to recognize several polymorphic VAR2CSA variants. This is consistent with the finding that some targets of naturally acquired antibodies are under diversifying selection (Dahlback et al. 2006). The levels of antibodies against pregnancy-associated parasite-encoded antigens on the surface of the erythrocyte increase with the number of pregnancies, and are correlated to the adhesion inhibitory capacity (Fried et al. 1998; Ricke et al. 2000). Therefore, it is also possible that protective adhesion-blocking antibodies against conserved epitopes only appear after a number of infections, and that the antibodies appearing after the first infection have less protective value.

### VAR2CSA, a globular protein?

So far, little is known about the overall structure of VAR2CSA or any other PfEMP1. It has been suggested that the PfEMP1 protein architecture is comprised by a compact conserved head structure which defines the binding affinity and a number of variable C-terminal domains (Su et al. 1995). Previous data have indicated that the general DBL fold is relatively conserved (Dahlback et al. 2006; Howell et al. 2006; Singh et al. 2006; Tolia et al. 2005) and that DBL domains can interact with each other as building block to form binding sites (Tolia et al. 2005). The data presented here indicates that DBL S3 is less surface-exposed than S1 and S2. Sub-domain 3 contains two long $\alpha$-helices which are conserved in the template structures. A number of multimeric protein complexes have been reported to be stabilized by interactions between long $\alpha$-helices; a well-studied example is the trimer of influenza virus hemagglutinin (Gamblin et al. 2004). The data presented here does not lend support to the idea of a compact conserved head structure in VAR2CSA. Rather it is tempting to propose models of VAR2CSA where $\alpha$-helices of different DBL domains interact to bury a considerable part of S3 in the interface between the domains. The interaction could be formed between DBL domains of a single VAR2CSA molecule to form a more globular shape of VAR2CSA. Another possibility is

that several VAR2CSA molecules form dimers or multimers with S3 buried in the middle. The possibility that VAR2CSA is present as a globular protein opens up for the possibility that the CSA binding site is comprised of regions from different DBL domains, like the glycan binding sites of EBA-175 (Tolia et al. 2005). Further exploration of the quaternary structure of VAR2CSA is therefore of the utmost importance.

# Experimental procedures:

### Modeling the structure of the VAR2CSA domains

Structures of the CIDR domains and the 3D7 VAR2CSA DBL and ID2 domains were modeled using the HHpred server with default settings (Soding et al. 2005). The HHpred method is based on comparisons and alignments of hidden Markow models (HMMs), which include gaps and insertion probabilities. The modeling of the DBL3 domain was done as described previously (Dahlback et al. 2006). VAR2CSA domains were modeled separately by splitting the PFL0030c sequence into separate domains (DBL1 aa 57-400, DBL2 aa 531-879, DBL4 aa 1575-1911, DBL5 aa 1999-2283, DBL6 aa 2340-2633, ID2 aa 879-1216). All HMM databases available in web-server were used for template structure search, including the Protein Data Bank (PDB). For all the VAR2CSA DBL domains described, the structures of the EBA-175 DBL domains (Tolia et al. 2005) and the Pk$\alpha$-DBL domain (Singh et al. 2006) had HHpred probability scores significantly higher than other structures detected. The DBL structure with the highest sequence and secondary structure alignment scores were chosen as template for each domain. Template alignments proposed by the HHpred method were used to generate 3D models by using a HHpred server toolkit protocol for MODELLER (Sali and Blundell 1993). Models were evaluated using Verify3d (Eisenberg et al. 1997) and ANOLEA (Melo and Feytmans 1998), available in the HHpred server toolkit.
Superimpositions of EBA-175 F1, F2, Pk$\alpha$-DBL and VAR2CSA domains were made using the MAMMOTH-mult alignment server (Lupyan et al. 2005). NACCESS version 2.1.1 (Hubbard and Thornton 1993) was used for analysis of relative accessible surface areas (RSAs) in the template structures. Residues with RSA$<$30% were considerer buried. Separate secondary structure predictions of CIDR and ID2 were made using the PSIPRED (Jones 1999). All structural visualizations were produced using PyMol (DeLano 2002).

### Selection and depletion of plasma samples on parasites

Plasma levels of DBL1-DBL6 VAR2CSA specific IgG were measured in standard ELISA assays (Dodoo et al. 2000). Plasma samples positive against more than two domains were used to make two pools of plasma samples. No single plasma sample was used in both pools. Rabbits were immunized with DBL1-DBL6 and ID2 recombinant protein as described (Barfod et al. 2006) and the seven serum samples were pooled to create the VAR2CSA rabbit pool.
3D7 and FCR3 parasites were panned on BeWo cells to create CSA adhering parasite lines. Using real time PCR and flow cytometry with VAR2CSA specific reagents as well as plasma from Tanzania, we verified that the parasites were

gender-specifically recognized and expressing high levels of VAR2CSA on the surface of the infected erythrocyte (Haase et al. 2006).
For the parasite IgG depletion 40 $\mu$l of the plasma pool were incubated with 2.0 $\times$ 108 MACS purified intact late stage trophozoite- and schizont-infected erythrocytes for 20 min at 4 °C. Hereafter, the cells were centrifuged at 800 g for 8 min, and the supernatant used to suspend a new pellet of 2.0 $\times$ 108 infected erythrocytes. This procedure was repeated four times. The depletion of surface reactive antibodies from the plasma pool was confirmed using a flow cytometry assay after the final depletion.

### Analysis of Pepcan data

All raw pepscan data were analyzed for short motifs as described previously (Dahlback et al. 2006). For each of the four depletion studies, DVs were calculated by subtracting depleted pepscan data points from non depleted points. Positive DVs of each individual experiment was split into five equally sized intervals ranging from the highest to the lowest data points and with increasing color intensity. The five intervals were then plotted on the models. For consensus DVs, the DVs for each depletion study on individual DBL domain were first normalized by subtracting the mean and dividing with the standard deviation. Subsequently, the consensus DVs was calculated by summing the normalized values for each position in the DBL domains. Consensus DVs $>1$ were plotted on the DBL models.

## Acknowledgements

The authors would like to thank Thomas Lavstsen, Anne Mølgaard and Thomas Blicher for valuable suggestions and discussions of the work presented here.

# Supplementary figures to Paper III

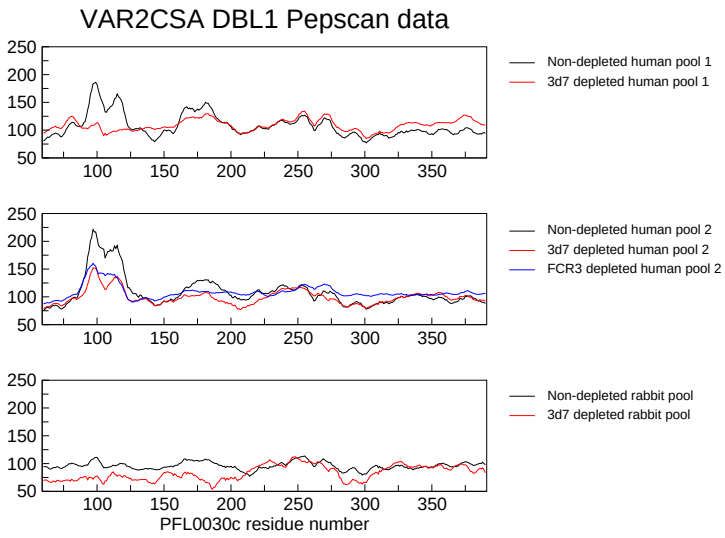


Figure 3.17: **Plots of pepscan analysis results for the DBL1 domain.** The top figure represent results from anti-VAR2CSA antibody human plasma pool 1 before and after depletion with infected parasites. The middle figure represents similar results obtained by using anti-VAR2CSA antibody human plasma pool 2, and the bottom figure represents similar results obtained by using a rabbit anti-VAR2CSA antibody plasma pool. All data presented here was analyzed for short motifs.

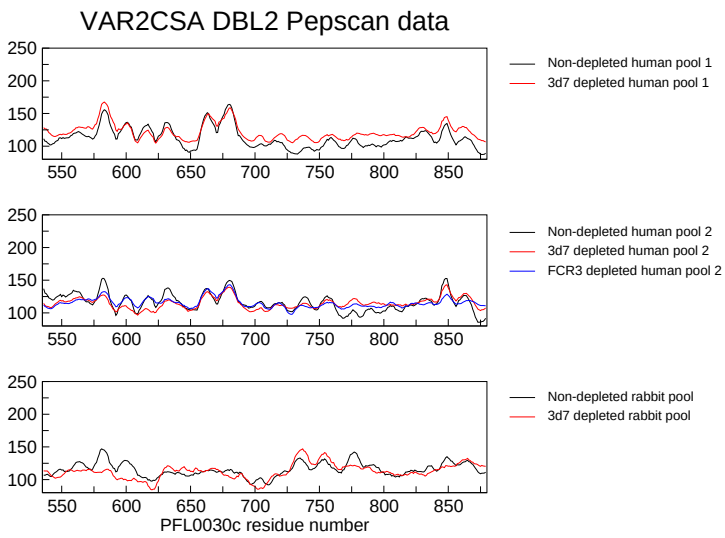


Figure 3.18: **Plots of pepscan analysis results for the DBL2 domain.** The top figure represent results from anti-VAR2CSA antibody human plasma pool 1 before and after depletion with infected parasites. The middle figure represents similar results obtained by using anti-VAR2CSA antibody human plasma pool 2, and the bottom figure represents similar results obtained by using a rabbit anti-VAR2CSA antibody plasma pool. All data shown in this figure was analyzed for short motifs.

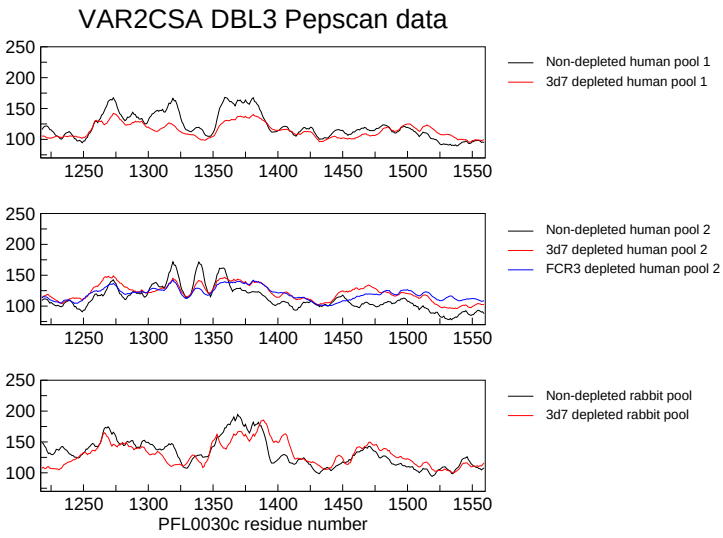


Figure 3.19: **Plots of pepscan analysis results for the DBL3 domain.** The top figure represent results from anti-VAR2CSA antibody human plasma pool 1 before and after depletion with infected parasites. The middle figure represents similar results obtained by using anti-VAR2CSA antibody human plasma pool 2, and the bottom figure represents similar results obtained by using a rabbit anti-VAR2CSA antibody plasma pool. All data shown in this figure was analyzed for short motifs.

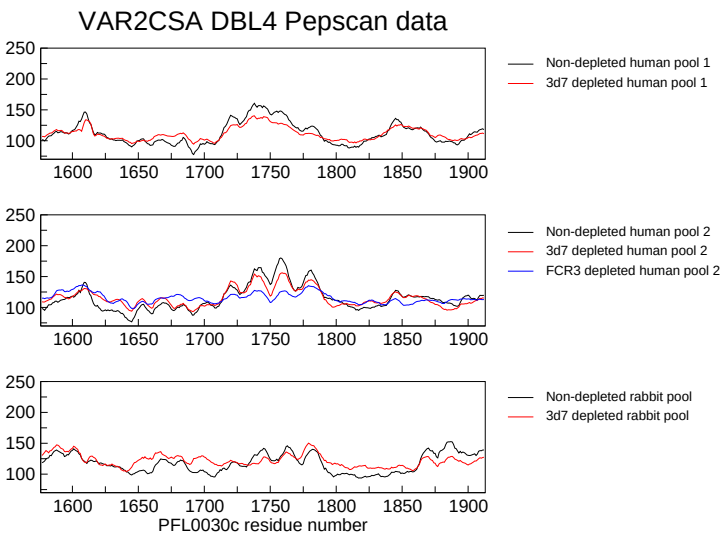


Figure 3.20: **Plots of pepscan analysis results for the DBL4 domain.** The top figure represent results from anti-VAR2CSA antibody human plasma pool 1 before and after depletion with infected parasites. The middle figure represents similar results obtained by using anti-VAR2CSA antibody human plasma pool 2, and the bottom figure represents similar results obtained by using a rabbit anti-VAR2CSA antibody plasma pool. All data shown in this figure was analyzed for short motifs.

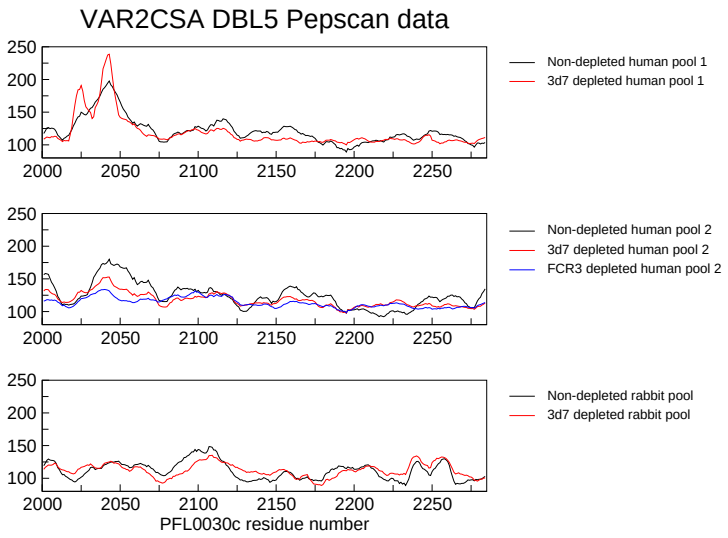


Figure 3.21: **Plots of pepscan analysis results for the DBL5 domain.** The top figure represent results from anti-VAR2CSA antibody human plasma pool 1 before and after depletion with infected parasites. The middle figure represents similar results obtained by using anti-VAR2CSA antibody human plasma pool 2, and the bottom figure represents similar results obtained by using a rabbit anti-VAR2CSA antibody plasma pool. All data shown in this figure was analyzed for short motifs.

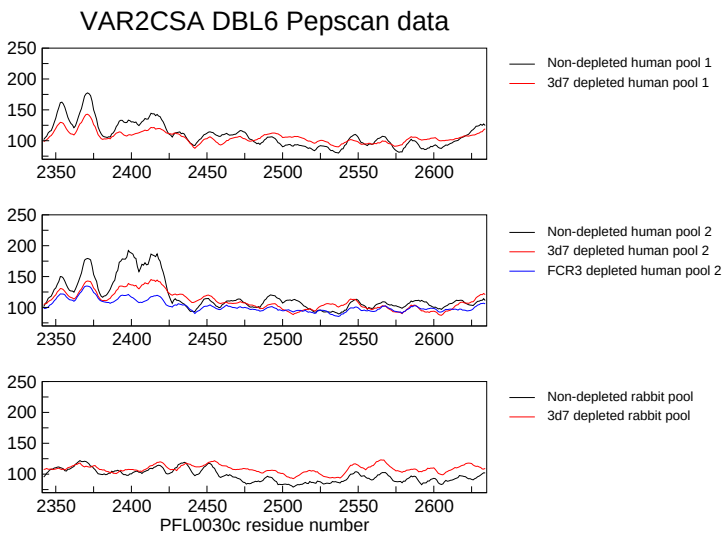


Figure 3.22: **Plots of pepscan analysis results for the DBL6 domain.** The top figure represent results from anti-VAR2CSA antibody human plasma pool 1 before and after depletion with infected parasites. The middle figure represents similar results obtained by using anti-VAR2CSA antibody human plasma pool 2, and the bottom figure represents similar results obtained by using a rabbit anti-VAR2CSA antibody plasma pool. All data shown in this figure was analyzed for short motifs.

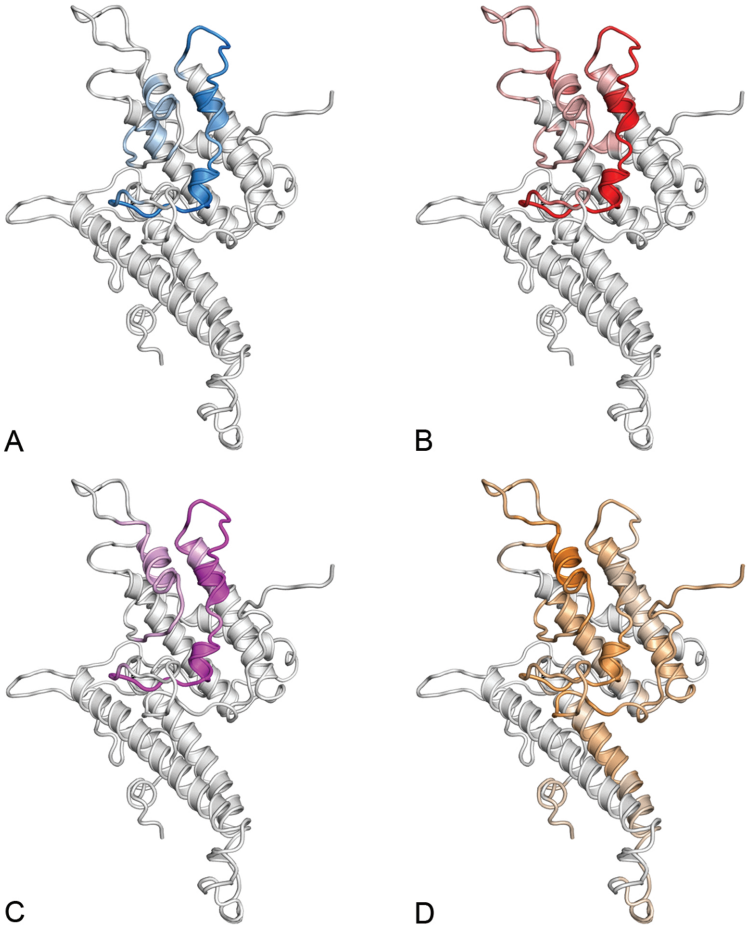


Figure 3.23: **Mapping surface exposed antibody reactive regions on the model of the DBL1 domain.** The color intensity denotes the values of positive depletion values. (A) Human serum pool 1 depleted with 3D7 parasites. (B) Human serum pool 2 depleted with 3D7 parasites. (C) Human serum pool 2 depleted with FCR3 parasites. (D) Rabbit serum pool depleted with 3D7 parasites.

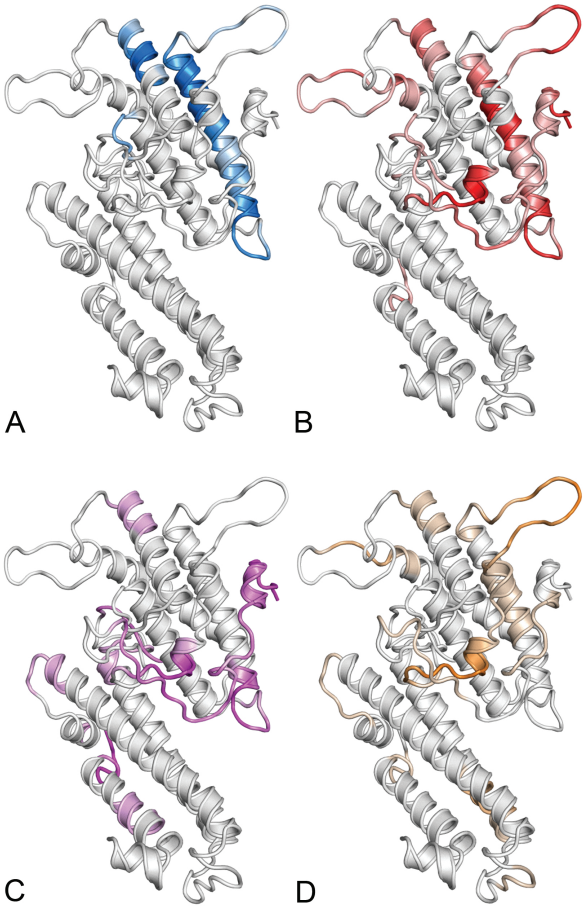


Figure 3.24: **Mapping surface exposed antibody reactive regions on the model of the DBL2 domain.** The color intensity denotes the values of positive depletion values. (A) Human serum pool 1 depleted with 3D7 parasites. (B) Human serum pool 2 depleted with 3D7 parasites. (C) Human serum pool 2 depleted with FCR3 parasites. (D) Rabbit serum pool depleted with 3D7 parasites.

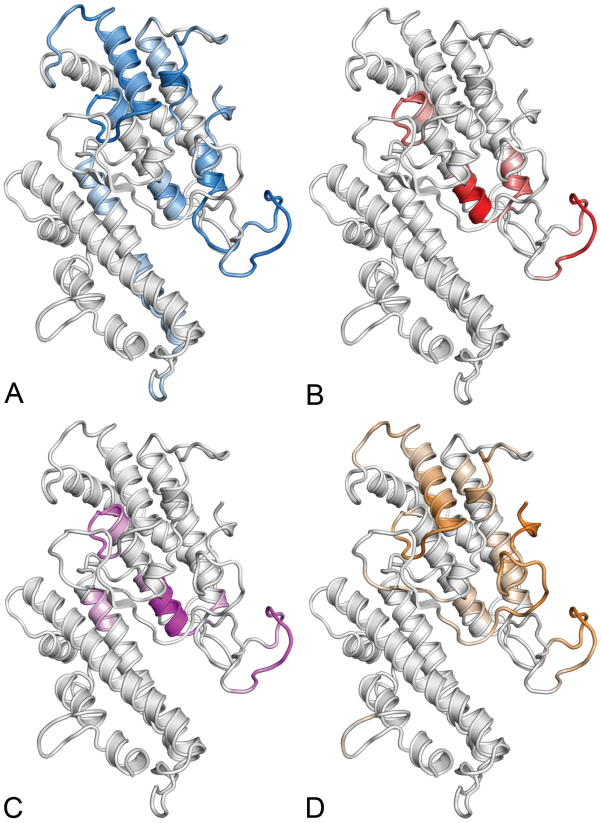


Figure 3.25: **Mapping surface exposed antibody reactive regions on the model of the DBL3 domain.** The color intensity denotes the values of positive depletion values. (A) Human serum pool 1 depleted with 3D7 parasites. (B) Human serum pool 2 depleted with 3D7 parasites. (C) Human serum pool 2 depleted with FCR3 parasites. (D) Rabbit serum pool depleted with 3D7 parasites.

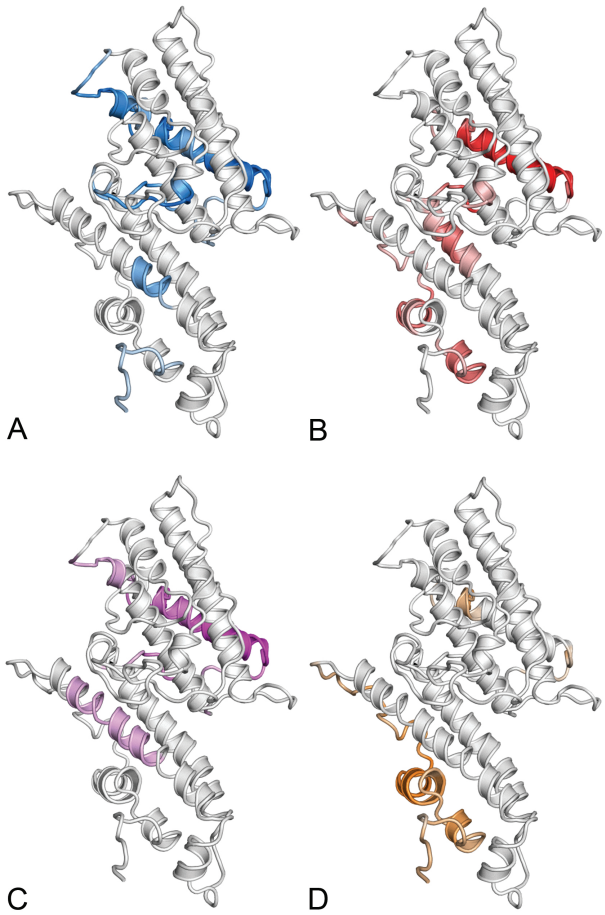


Figure 3.26: **Mapping surface exposed antibody reactive regions on the model of the DBL4 domain.** The color intensity denotes the values of positive depletion values. (A) Human serum pool 1 depleted with 3D7 parasites. (B) Human serum pool 2 depleted with 3D7 parasites. (C) Human serum pool 2 depleted with FCR3 parasites. (D) Rabbit serum pool depleted with 3D7 parasites.

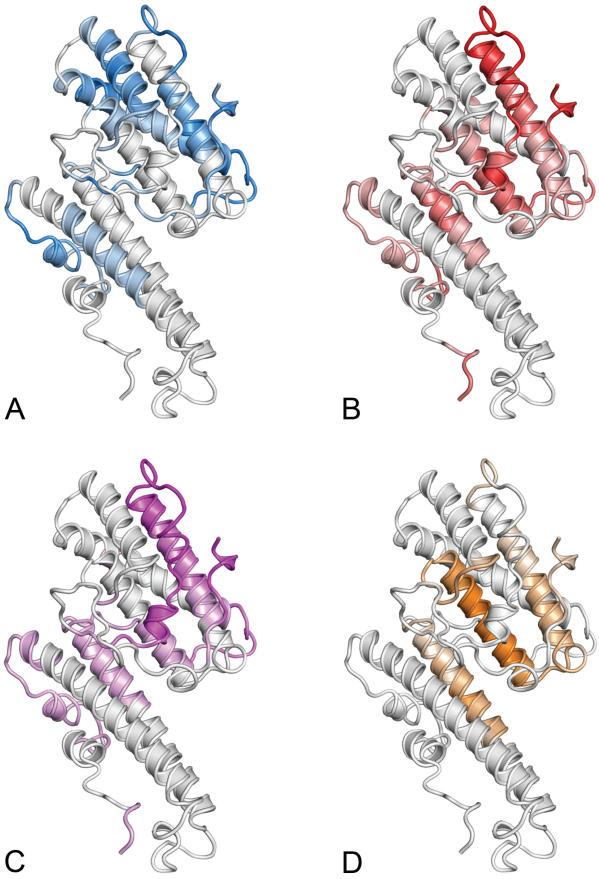


Figure 3.27: **Mapping surface exposed antibody reactive regions on the model of the DBL5 domain.** The color intensity denotes the values of positive depletion values. (A) Human serum pool 1 depleted with 3D7 parasites. (B) Human serum pool 2 depleted with 3D7 parasites. (C) Human serum pool 2 depleted with FCR3 parasites. (D) Rabbit serum pool depleted with 3D7 parasites.

## 3.7 Discussion of results in papers II and III

In the two papers presented in this chapter, we identified peptides from VAR2CSA DBL domains, which cross-bind to antibodies binding the native VAR2CSA protein in conformations found on the surface of infected erythrocytes. The results based on the peptide arrays were then mapped on the 3D models of DBL domain structures, and this suggested that S1 and S2 are generally more targeted by antibodies from women immune to PAM, than S3. As mentioned in the papers, it is important to realize that these peptide array experiments do not identify all epitopes of these domains. There is two reasons for this: First, the peptide arrays only identify linear epitopes. In general, many natural epitopes are discontinuous and conformation-dependent; therefore all residues of these epitopes cannot be expected to be identified by using peptide arrays. Second, the identified peptides are restricted to only cross-react to antibodies in the serum pools used. Because the sequence variability of VAR2CSA is high, it is possible that the serum pools contain antibodies which cannot bind the 3d7 line, but bind surface exposed epitopes in VAR2CSA DBL domains from other lines with high affinity.

It is important to note that the DBL3X residues which were found to vary significantly with the parity of women, are not found to be highly reactive in the pepscan experiments. This can be seen in figure 3.7, where only the data shown in figure 3.7G-I have notable pepscan scores in that region. In the results of paper III, which are based on other serum pools, the region is not seen to be reactive either (see figure 3.25). Also, we observed a difference in reactivity of depletion studies between different serum samples in papers II and III. One of the depletion studies (figure 3.25A) showed similar reactive regions to the SE1-SE3 regions identified in paper II (figure 3.8D and E), but the other depletion studies of DBL3X (figure 3.25B and C) showed different reactive regions, and only the SE2 region was found to be have reactivity. Taken together, these results suggest that more epitopes remain to be identified. However, all of the DBL domains except DBL4 have little reactivity in the S3 part of the DBL structure and this strengthens the hypothesis that the S3 part is less prone to targeting by neutralizing antibodies.

The final aim of the PAM research projects is to develop vaccines which can prevent the effects of PAM. However, little is known about the mechanisms of VAR2CSA function and host interaction. For the development of a vaccine, it is useful to gain more knowledge of these mechanisms because epitopes used in a vaccine should elicit neutralizing antibodies which abrupt binding of CSA. The results presented in the papers II and III provide more insight in antibody reactive regions, and future experiments should investigate whether the epitopes of these regions can elicit neutralizing antibodies.

Chapter 4

# Concluding remarks and perspectives

The chapters of this thesis have described my work on different research projects in the fields of B-cell epitope prediction and identification. In this chapter is a final summary of some findings of the projects and a discussion of the perspectives of the results.

## 4.1 Concluding remarks

### Prediction of B-cell epitopes

In the development of a B-cell epitope prediction method, 3D structures of antibody-antigen complexes provided detailed information of binding interactions which were used to derive a data set of discontinuous B-cell epitopes. The data set was used to develop a method for prediction of discontinuous B-cell epitopes, which uses protein 3D structures to measure surface exposure and determine spatial proximity. We have shown that the method performs better than a classical sequence-based method, and that it can predict residues in epitopes which have been identified on the basis of different experimental techniques. Thus, in the B-cell epitope prediction project, 3D protein structures were useful both for definition of epitopes in the data set, and for providing information about proteins which can be used for prediction of epitope residues.

### Identification of B-cell epitopes in the PAM-related protein VAR2CSA

The experimentally determined structures of the EBA-175 and Pk$\alpha$ DBL domains were essential for this project, as structure predictions showed that the VAR2CSA DBL domains have similar folds to the EBA-175 and Pk$\alpha$ DBL domains, despite a high sequence variation between the domains. Furthermore, 3D structure prediction of the VAR2CSA ID2 and various CIDR domains indicated that parts of these domains have similar folds to the S3 region of the DBL domains. Secondary structure prediction showed that other parts of the ID2

and CIDR domains are less likely to fold as DBL sub-domains 1 and 2. The homology modeling of the VAR2CSA gave some indications about the structures of several domains of the VAR2CSA protein, which can be used for mapping of experimental data. B-cell epitopes in VAR2CSA were identified using pepscan peptide-arrays. The study showed that all DBL domains in the VAR2CSA protein contain peptides that can cross-react with naturally acquired antibodies. The identified VAR2CSA B-cell epitopes were analyzed in the context of the predicted structures, and this showed that these epitopes are mostly located on one side of the DBL domain structures and primarily in sub-domains 1 and 2. Furthermore, the mapping of identified linear epitopes on the 3D structures showed that many of the epitopes are proximal in space and may together form discontinuous epitopes. In the VAR2CSA DBL3X domain, it was shown that residues with high sequence variability were mostly located on one side of the domain, in a region which is not aligned to the dimer interfaces of EBA-175 DBL domains. In addition, it was found that both positions of high recombination rates, and surface-exposed epitopes were located in predicted loop regions which were suggested to be flexible.

## 4.2 Perspectives

The developed method and findings presented in this thesis may contribute in many ways to vaccine design. It has been shown that the DiscoTope method has a reasonable performance, and hopefully the server will be applied broadly among researchers for prediction studies previous to experimental identification studies.

However, the method could be improved further. For instance, more sophisticated surface accessibility measures would probably improve the method. Additionally, protein structure information could be useful for allocating predicted B-cell epitope residues into groups representing whole B-cell epitope entities. The server could be improved by integrating the DiscoTope predictions with information of glycosylation or trans-membrane regions; and plots of sequence variability could also be used to help identify predicted epitopes located in conserved areas of the structure. Finally, the prediction of protein 3D structure could be part of the server; this would be useful for a number of pathogen proteins with no determined structure.

The presented results of the VAR2CSA PAM-vaccine project gave more insight in the locations of B-cell epitopes of naturally acquired human antibodies. Because B-cell epitopes of naturally acquired antibodies were identified in all DBL domains, it is probable that a PAM vaccine would have to present epitopes from several of these DBL domains to be effective.

In further investigation and modeling of the native VAR2CSA quaternary structure, the modeled DBL structures may be useful in combination with cryo-EM experiments. These studies could give more insight in the native conformation of VAR2CSA and indicate whether the overall shape of the molecule is like a string of pearls (where the DBL domains represent pearls), or if the VAR2CSA protein folds into specific quaternary structures by interactions between DBL domains. Cryo-EM is also useful for improving comparative models, because shapes observed in the structures can be used as restraints in comparative modeling (Topf and Sali 2005). In this way, it is likely that the DBL domain models could be

further improved by combination with cryo-EM structures of the VAR2CSA protein.
The mapping of functional, CSA-binding residues in VAR2CSA could lead to identify which of the identified B-cell epitopes are overlapping, or proximal, to the functional site; when used in a vaccine, such epitopes could more probably elicit neutralizing antibodies protecting against sequestration. ELISA experiments using peptides shown to bind naturally acquired antibodies could be used to investigate for CSA binding. Similarly, mutant DBL domains could be used to test antibody reactive regions for CSA binding properties. Another approach could be to immunize mice or rabbits with reactive peptides identified in the pepscan experiments, and test the serum for antibodies which inhibit the binding of CSA. Additionally, the serum could be affinity purified on infected erythrocytes, to identify peptides which elicit antibodies that bind native-like VAR2CSA.
The X-ray structures of the EBA-175 DBL domains suggested that the dimerization of this protein is important for GAG binding. If dimerization (or multimerization) of VAR2CSA DBL domains is similarly important for CSA binding, it would be interesting to determine if the binding of antibodies to epitopes that are overlapping the dimerization site can prevent CSA binding. Thus, investigation of dimerization properties of VAR2CSA domains also have a potential for providing valuable knowledge which can be useful in vaccine design.
The sequencing of more *var2csa* genes is essential in the investigation of sequence variability in the rest of the VAR2CSA DBL domains; this would help to determine which epitopes are more conserved than others. Conserved epitopes may be more effective if used in a vaccine, because the vaccine has a potential to protect from infections of several *P. falciparum* lines.
We found that naturally acquired antibodies are directed against all the DBL domains of VAR2CSA. This suggests that a PAM vaccine must present epitopes from several DBL domains to be effective. In contrast, the finding, that most epitopes are located in sub-domains 1 and 2, suggests that a VAR2CSA-based vaccine may only contain these sub-domains expressed as recombinant, stable proteins. The exclusion of sub-domain 3 could potentially prevent unwanted immune responses, which are directed against irrelevant epitopes and do not lead to protection against PAM. Energy potentials and protein structure prediction could probably contribute to the design of a vaccine including only sub-domains 1 and 2 by prediction of stable constructs that can be expressed as recombinant proteins. However, these constructs would need to be tested intensively in experimental studies to identify which constructs present relevant epitopes that elicit protecting antibodies, and at the same time are not leading to unwanted side-effects.
At present, both B-cell epitope prediction, and PAM vaccines are subjects of research for many different groups which are working on interesting projects. The outcome of this research has potential to impact the health of many individuals on a global scale. In this thesis I have presented results contributing to research in both subjects, and I am very interested in seeing how these contributions may facilitate future research in vaccine development.

# Bibliography


Akaike, H. (1973). Information theory and an extension of the extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second international symposium on information theory*, Budapest, pp. 267–281.

Alix, A. J. (1999, Sep). Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine 18*(3-4), 311–314.

Allcorn, L. C. and A. C. Martin (2002). Sacs–self-maintaining database of antibody crystal structure information. *Bioinformatics 18*(1), 175–81.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997, Sep). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25*(17), 3389–3402.

Arnon, R. and T. Ben-Yedidia (2003, Aug). Old and new vaccine approaches. *Int Immunopharmacol 3*(8), 1195–1204.

Bai, T., M. Becker, A. Gupta, P. Strike, V. J. Murphy, R. F. Anders, and A. H. Batchelor (2005). Structure of ama1 from plasmodium falciparum reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc Natl Acad Sci U S A 102*(36), 12736–41.

Bairoch, A. and R. Apweiler (2000, Jan). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res 28*(1), 45–48.

Barfod, L., N. L. Bernasconi, M. Dahlback, D. Jarrossay, P. H. Andersen, A. Salanti, M. F. Ofori, L. Turner, M. Resende, M. A. Nielsen, T. G. Theander, F. Sallusto, A. Lanzavecchia, and L. Hviid (2007, Jan). Human pregnancy-associated malaria-specific B cells target polymorphic, conformational epitopes in VAR2CSA. *Mol Microbiol 63*(2), 335–347.

Barfod, L., M. A. Nielsen, L. Turner, M. Dahlback, A. T. Jensen, L. Hviid, T. G. Theander, and A. Salanti (2006). Baculovirus-expressed constructs induce immunoglobulin G that recognizes VAR2CSA on Plasmodium falciparum-infected erythrocytes. *Infect Immun 74*(7), 4357–60.

Barlow, D. J., M. S. Edwards, and J. M. Thornton (1986, Aug). Continuous and discontinuous protein antigenic determinants. *Nature 322*(6081), 747–748.

Baruch, D. I., B. L. Pasloske, H. B. Singh, X. Bi, X. C. Ma, M. Feldman, T. F. Taraschi, and R. J. Howard (1995). Cloning the p. falciparum gene encoding pfemp1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell 82*(1), 77–87.

Batori, V., E. P. Friis, H. Nielsen, and E. L. Roggen (2006, Jan). An in silico method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context. *J Mol Recognit 19*(1), 21–29.

Belnap, D. M., N. H. Olson, N. M. Cladel, W. W. Newcomb, J. C. Brown, J. W. Kreider, N. D. Christensen, and T. S. Baker (1996, Jun). Conserved features in papillomavirus and polyomavirus capsids. *J Mol Biol 259*(2), 249–263.

Blackman, M. J., T. J. Scott-Finnigan, S. Shai, and A. A. Holder (1994, Jul). Antibodies inhibit the protease-mediated processing of a malaria merozoite surface protein. *J Exp Med 180*(1), 389–393.

Blum, M. L., J. A. Down, A. M. Gurnett, M. Carrington, M. J. Turner, and D. C. Wiley (1993, Apr). A structural motif in the variant surface glycoproteins of Trypanosoma brucei. *Nature 362*(6421), 603–609.

Blythe, M. J. and D. R. Flower (2005, Jan). Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci 14*(1), 246–248.

Bowie, J. U., R. Luthy, and D. Eisenberg (1991, Jul). A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*(5016), 164–170.

Brabin, B. J. (1983). An analysis of malaria in pregnancy in Africa. *Bull World Health Organ 61*, 1005–1016.

Branden, C. and J. Tooze (1998). *Introduction to protein structure* (2 ed.). Garland.

Bray, R. S. and M. J. Anderson (1979). Falciparum malaria and pregnancy. *Trans R Soc Trop Med Hyg 73*(4), 427–431.

Bublil, E. M., N. T. Freund, I. Mayrose, O. Penn, A. Roitburd-Berman, N. D. Rubinstein, T. Pupko, and J. M. Gershoni (2007, Jul). Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. *Proteins 68*(1), 294–304.

Burnham, K. P. and D. R. Anderson (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Cabezas, E., M. Wang, P. W. Parren, R. L. Stanfield, and A. C. Satterthwait (2000, Nov). A structure-based approach to a synthetic vaccine for HIV-1. *Biochemistry 39*(47), 14377–14391.

Carrington, M. and J. Boothroyd (1996, Oct). Implications of conserved structural motifs in disparate trypanosome surface proteins. *Mol Biochem Parasitol 81*(2), 119–126.

Castrignano, T., P. D. De Meo, D. Carrabino, M. Orsini, M. Floris, and A. Tramontano (2007). The MEPS server for identifying protein conformational epitopes. *BMC Bioinformatics 8 Suppl 1*, S6.

Chitarra, V., P. M. Alzari, G. A. Bentley, T. N. Bhat, J. L. Eisele, A. Houdusse, J. Lescar, H. Souchon, and R. J. Poljak (1993, Aug). Three-dimensional structure of a heteroclitic antigen-antibody cross-reaction complex. *Proc Natl Acad Sci U S A 90*(16), 7711–7715.

Chothia, C. (1976, Jul). The nature of the accessible and buried surfaces in proteins. *J Mol Biol 105*(1), 1–12.

Chothia, C. and A. M. Lesk (1986, Apr). The relation between the divergence of sequence and structure in proteins. *EMBO J 5*(4), 823–826.

Chou, P. Y. and G. D. Fasman (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol 47*, 45–148.

Christophers, S. R. (1924). The mechanism of immunity against malaria in communities living under hyper-endemic conditions. *Ind J Med Res 12*, 273–294.

Cohen, G. H., E. W. Silverton, E. A. Padlan, F. Dyda, J. A. Wibbenmeyer, R. C. Willson, and D. R. Davies (2005, May). Water molecules in the antibody–antigen interface of the structure of the Fab HyHEL-5–lysozyme complex at 1.7 Å resolution: comparison with results from isothermal titration calorimetry. *Acta Crystallographica Section D 61*(5), 628–633.

Coley, A. M., K. Parisi, R. Masciantonio, J. Hoeck, J. L. Casey, V. J. Murphy, K. S. Harris, A. H. Batchelor, R. F. Anders, and M. Foley (2006). The most polymorphic residue on plasmodium falciparum apical membrane antigen 1 determines binding of an invasion-inhibitory antibody. *Infect Immun 74*(5), 2628–36.

Collis, A. V. J., A. P. Brouwer, and A. C. R. Martin (2003, Jan). Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol 325*(2), 337–354.

Dahlback, M., T. S. Rask, P. H. Andersen, M. A. Nielsen, N. T. Ndam, M. Resende, L. Turner, P. Deloron, L. Hviid, O. Lund, A. G. Pedersen, T. G. Theander, and A. Salanti (2006, Nov). Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in P. falciparum placental sequestration. *PLoS Pathog 2*(11), e124.

Debelle, L., S. M. Wei, M. P. Jacob, W. Hornebeck, and A. J. Alix (1992). Predictions of the secondary structure and antigenicity of human and bovine tropoelastins. *Eur Biophys J 21*(5), 321–9.

DeLano, W. (2002). *The PyMol Molecular Graphics System.*

Desai, M., F. O. ter Kuile, F. Nosten, R. McGready, K. Asamoa, B. Brabin, and R. D. Newman (2007, Feb). Epidemiology and burden of malaria in pregnancy. *Lancet Infect Dis 7*(2), 93–104.

Dodoo, D., M. Theisen, J. A. Kurtzhals, B. D. Akanmori, K. A. Koram, S. Jepsen, F. K. Nkrumah, T. G. Theander, and L. Hviid (2000). Naturally acquired antibodies to the glutamate-rich protein are associated with protection against plasmodium falciparum malaria. *J Infect Dis 181*(3), 1202–5.

Douek, D. C., P. D. Kwong, and G. J. Nabel (2006, Feb). The rational design of an AIDS vaccine. *Cell 124*(4), 677–681.

Duffy, M. F., A. Caragounis, R. Noviyanti, H. M. Kyriacou, E. K. Choong, K. Boysen, J. Healer, J. A. Rowe, M. E. Molyneux, G. V. Brown, and S. J. Rogerson (2006, Aug). Transcribed var genes associated with placental malaria in Malawian women. *Infect Immun 74*(8), 4875–4883.

Duffy, M. F., A. G. Maier, T. J. Byrne, A. J. Marty, S. R. Elliott, M. T. O'Neill, P. D. Payne, S. J. Rogerson, A. F. Cowman, B. S. Crabb, and G. V. Brown (2006, Aug). VAR2CSA is the principal ligand for chondroitin sulfate A in two allogeneic isolates of Plasmodium falciparum. *Mol Biochem Parasitol 148*(2), 117–124.

Duffy, P. E. and M. Fried (2003, Nov). Antibodies that inhibit Plasmodium falciparum adhesion to chondroitin sulfate A are associated with increased birth weight and the gestational age of newborns. *Infect Immun 71*(11), 6620–6623.

Dunbrack, R. L. J. (2006, Jun). Sequence comparison and protein structure prediction. *Curr Opin Struct Biol 16*(3), 374–384.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the Bootstrap* (First edition ed.). London: Chapman and Hall.

Eigenmann, P. A. (2004, Jun). Do we have suitable in-vitro diagnostic tests for the diagnosis of food allergy? *Curr Opin Allergy Clin Immunol 4*(3), 211–213.

Eisenberg, D., R. Luthy, and J. U. Bowie (1997). Verify3d: assessment of protein models with three-dimensional profiles. *Methods Enzymol 277*, 396–404.

Ellis, R. W. (1999, Mar). New technologies for making vaccines. *Vaccine 17*(13-14), 1596–1604.

Emini, E. A., J. V. Hughes, D. S. Perlow, and J. Boger (1985, Sep). Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol 55*(3), 836–839.

Engwerda, C. R., L. Beattie, and F. H. Amante (2005, Feb). The importance of the spleen in malaria. *Trends Parasitol 21*(2), 75–80.

Enshell-Seijffers, D., D. Denisov, B. Groisman, L. Smelyanski, R. Meyuhas, G. Gross, G. Denisova, and J. M. Gershoni (2003, Nov). The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1. *J Mol Biol 334*(1), 87–101.

Epstein, J. E., B. Giersing, G. Mullen, V. Moorthy, and T. L. Richie (2007, Feb). Malaria vaccines: are we getting closer? *Curr Opin Mol Ther 9*(1), 12–24.

Faelber, K., D. Kirchhofer, L. Presta, R. F. Kelley, and Y. A. Muller (2001). The 1.85 a resolution crystal structures of tissue factor in complex with humanized fab d3h44 and of free humanized fab d3h44: revisiting the solvation of antigen combining sites. *J Mol Biol 313*(1), 83–97.

Fischmann, T. O., G. A. Bentley, T. N. Bhat, G. Boulot, R. A. Mariuzza, S. E. Phillips, D. Tello, and R. J. Poljak (1991, Jul). Crystallographic

refinement of the three-dimensional structure of the FabD1.3-lysozyme complex at 2.5-A resolution. *J Biol Chem 266*(20), 12915–12920.

Fiser, A., R. K. Do, and A. Sali (2000, Sep). Modeling of loops in protein structures. *Protein Sci 9*(9), 1753–1773.

Fiser, A. and A. Sali (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol 374*, 461–491.

Fleury, D., R. S. Daniels, J. J. Skehel, M. Knossow, and T. Bizebard (2000, Sep). Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins 40*(4), 572–578.

Freitas-Junior, L. H., E. Bottius, L. A. Pirrit, K. W. Deitsch, C. Scheidig, F. Guinet, U. Nehrbass, T. E. Wellems, and A. Scherf (2000, Oct). Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature 407*(6807), 1018–1022.

Fried, M. and P. E. Duffy (1996, Jun). Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. *Science 272*(5267), 1502–1504.

Fried, M., F. Nosten, A. Brockman, B. J. Brabin, and P. E. Duffy (1998, Oct). Maternal antibodies block malaria. *Nature 395*(6705), 851–852.

Fu, Y. X. and W. H. Li (1993, Mar). Statistical tests of neutrality of mutations. *Genetics 133*(3), 693–709.

Gamain, B., J. D. Smith, N. K. Viebig, J. Gysin, and A. Scherf (2007, Mar). Pregnancy-associated malaria: parasite binding, natural immunity and vaccine development. *Int J Parasitol 37*(3-4), 273–283.

Gamain, B., A. R. Trimnell, C. Scheidig, A. Scherf, L. H. Miller, and J. D. Smith (2005, Mar). Identification of multiple chondroitin sulfate A (CSA)-binding domains in the var2CSA gene transcribed in CSA-binding parasites. *J Infect Dis 191*(6), 1010–1013.

Gamblin, S. J., L. F. Haire, R. J. Russell, D. J. Stevens, B. Xiao, Y. Ha, N. Vasisht, D. A. Steinhauer, R. S. Daniels, A. Elliot, D. C. Wiley, and J. J. Skehel (2004). The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science 303*(5665), 1838–42.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M.-S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. A. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell (2002, Oct). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature 419*(6906), 498–511.

Garnier, J., D. J. Osguthorpe, and B. Robson (1978, Mar). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol 120*(1), 97–120.

Geysen, H. M., R. H. Meloen, and S. J. Barteling (1984, Jul). Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc Natl Acad Sci U S A 81*(13), 3998–4002.

Ginalski, K. (2006, Apr). Comparative modeling for protein structure prediction. *Curr Opin Struct Biol 16*(2), 172–177.

Goldman, N. and Z. Yang (1994, Sep). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol 11*(5), 725–736.

Greenbaum, J. A., P. H. Andersen, M. Blythe, H.-H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. S. Kolaskar, O. Lund, S. Morrison, B. Mumey, Y. Ofran, J.-L. Pellequer, C. Pinilla, J. V. Ponomarenko, G. P. S. Raghava, M. H. V. van Regenmortel, E. L. Roggen, A. Sette, A. Schlessinger, J. Sollner, M. Zand, and B. Peters (2007, Mar). Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit 20*(2), 75–82.

Haase, R. N., R. Megnekou, M. Lundquist, M. F. Ofori, L. Hviid, and T. Staalsoe (2006). Plasmodium falciparum parasites expressing pregnancy-specific variant surface antigens adhere strongly to the choriocarcinoma cell line bewo. *Infect Immun 74*(5), 3035–8.

Hagensee, M. E., N. H. Olson, T. S. Baker, and D. A. Galloway (1994, Jul). Three-dimensional structure of vaccinia virus-produced human papillomavirus type 1 capsids. *J Virol 68*(7), 4503–4505.

Hamby, C. V., M. Llibre, S. Utpat, and G. P. Wormser (2005, Jul). Use of Peptide library screening to detect a previously unknown linear diagnostic epitope: proof of principle by use of lyme disease sera. *Clin Diagn Lab Immunol 12*(7), 801–807.

Hans, D., P. R. Young, and D. P. Fairlie (2006, Nov). Current status of short synthetic peptides as vaccines. *Med Chem 2*(6), 627–646.

Harris, L. J., S. B. Larson, K. W. Hasel, and A. McPherson (1997, Feb). Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry 36*(7), 1581–1597.

Hartl, D. L. and A. G. Clark (1997). *Principles of population genetics* (3 ed.). Sunderland, Massachusetts: Sinauer.

Haste Andersen, P., M. Nielsen, and O. Lund (2006, Nov). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci 15*(11), 2558–2567.

Heppner, D. G. J., K. E. Kester, C. F. Ockenhouse, N. Tornieporth, O. Ofori, J. A. Lyon, V. A. Stewart, P. Dubois, D. E. Lanar, U. Krzych, P. Moris, E. Angov, J. F. Cummings, A. Leach, B. T. Hall, S. Dutta, R. Schwenk, C. Hillier, A. Barbosa, L. A. Ware, L. Nair, C. A. Darko, M. R. Withers, B. Ogutu, M. E. Polhemus, M. Fukuda, S. Pichyangkul, M. Gettyacamin, C. Diggs, L. Soisson, J. Milman, M.-C. Dubois, N. Garcon, K. Tucker, J. Wittes, C. V. Plowe, M. A. Thera, O. K. Duombo, M. G. Pau, J. Goudsmit, W. R. Ballou, and J. Cohen (2005, Mar). Towards an RTS,S-based, multi-stage, multi-antigen vaccine against falciparum malaria: progress at the Walter Reed Army Institute of Research. *Vaccine 23*(17-18), 2243–2250.

Hijnen, M., F. R. Mooi, P. G. M. van Gageldonk, P. Hoogerhout, A. J. King, and G. A. M. Berbers (2004, Jul). Epitope structure of the Bordetella pertussis protein P.69 pertactin, a major vaccine component and protective antigen. *Infect Immun 72*(7), 3716–3723.

Holder, A. A., J. A. Guevara Patino, C. Uthaipibull, S. E. Syed, I. T. Ling, T. Scott-Finnigan, and M. J. Blackman (1999, Sep). Merozoite surface protein 1, immune evasion, and vaccines against asexual blood stage malaria. *Parassitologia 41*(1-3), 409–414.

Hopp, T. P. and K. R. Woods (1981, Jun). Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A 78*(6), 3824–3828.

Hopp, T. P. and K. R. Woods (1983, Apr). A computer program for predicting protein antigenic determinants. *Mol Immunol 20*(4), 483–489.

Howell, D. P.-G., R. Samudrala, and J. D. Smith (2006, Jul). Disguising itself–insights into Plasmodium falciparum binding and immune evasion from the DBL crystal structure. *Mol Biochem Parasitol 148*(1), 1–9.

Hsueh, P.-R., C.-L. Kao, C.-N. Lee, L.-K. Chen, M.-S. Ho, C. Sia, X. D. Fang, S. Lynn, T. Y. Chang, S. K. Liu, A. M. Walfield, and C. Y. Wang (2004, Sep). SARS antibody test for serosurveillance. *Emerg Infect Dis 10*(9), 1558–1562.

Hubbard, S. J. and J. M. Thornton (1993). NACCESS, Computer Program. Technical report, Department of Biochemistry and Molecular Biology, University College London, UK.

Hudson, R. R. (2001, Dec). Two-locus sampling distributions and their application. *Genetics 159*(4), 1805–1817.

Hviid, L. and T. Staalsoe (2004, Feb). Malaria immunity in infants: a special case of a general phenomenon? *Trends Parasitol 20*(2), 66–72.

Igonet, S., B. Vulliez-Le Normand, G. Faure, M.-M. Riottot, C. H. M. Kocken, A. W. Thomas, and G. A. Bentley (2007, Mar). Cross-reactivity studies of an anti-Plasmodium vivax apical membrane antigen 1 monoclonal antibody: binding and structural characterisation. *J Mol Biol 366*(5), 1523–1537.

Jafari-Guemouri, S., C. Boudin, N. Fievet, P. Ndiaye, and P. Deloron (2006, Jun). Plasmodium falciparum genotype population dynamics in asymptomatic children from Senegal. *Microbes Infect 8*(7), 1663–1670.

Jameson, B. A. and H. Wolf (1988, Mar). The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput Appl Biosci 4*(1), 181–186.

Janeway, C. A. J., P. Travers, M. Walport, and M. J. Shlomchik (2005). *Immunobiology* (6rd ed.). New York, US: Garland Science Publishing.

Janin, J. and S. Wodak (1978, Nov). Conformation of amino acid side-chains in proteins. *J Mol Biol 125*(3), 357–386.

Jensen, A. T. R., P. Magistrado, S. Sharp, L. Joergensen, T. Lavstsen, A. Chiucchiuini, A. Salanti, L. S. Vestergaard, J. P. Lusingu, R. Hermsen, R. Sauerwein, J. Christensen, M. A. Nielsen, L. Hviid, C. Sutherland,

T. Staalsoe, and T. G. Theander (2004, May). Plasmodium falciparum associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A var genes. *J Exp Med 199*(9), 1179–1190.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol 292*(2), 195–202.

Karplus, P. A. and G. E. Schulz (1985). Prediction of Chain Flexibility in Proteins - A tool for the Selection of Peptide Antigens. *Naturwissenschaften 72*, 212–213.

Keating, G. M. and S. Noble (2003). Recombinant hepatitis B vaccine (Engerix-B): a review of its immunogenicity and protective efficacy against hepatitis B. *Drugs 63*(10), 1021–1051.

Koide, S., X. Yang, X. Huang, J. J. Dunn, and B. J. Luft (2005, Jul). Structure-based design of a second-generation Lyme disease vaccine based on a C-terminal fragment of Borrelia burgdorferi OspA. *J Mol Biol 350*(2), 290–299.

Kolaskar, A. S. and P. C. Tongaonkar (1990, Dec). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett 276*(1-2), 172–174.

Kooij, T. W. A., C. J. Janse, and A. P. Waters (2006, May). Plasmodium post-genomics: better the bug you know? *Nat Rev Microbiol 4*(5), 344–357.

Korber, B., M. LaBute, and K. Yusim (2006, Jun). Immunoinformatics comes of age. *PLoS Comput Biol 2*(6), e71.

Korenromp, E., J. Miller, B. Nahlen, T. Wardlaw, and M. Young (2005). Africa malaria report 2005. Technical Report WHO/CDS/MAL/2003.1093, World Health Organization.

Kraemer, S. M. and J. D. Smith (2006, Aug). A family affair: var genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol 9*(4), 374–380.

Kryshtafovych, A., C. Venclovas, K. Fidelis, and J. Moult (2005). Progress over the first decade of CASP experiments. *Proteins 61 Suppl 7*, 225–236.

Kulkarni-Kale, U., S. Bhosle, and A. S. Kolaskar (2005, Jul). CEP: a conformational epitope prediction server. *Nucleic Acids Res 33*(Web Server issue), 168–171.

Kwiatkowski, D. and K. Marsh (1997, Dec). Development of a malaria vaccine. *Lancet 350*(9092), 1696–1701.

Kyte, J. and R. F. Doolittle (1982, May). A simple method for displaying the hydropathic character of a protein. *J Mol Biol 157*(1), 105–132.

Larsen, J. E. P., O. Lund, and M. Nielsen (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res 2*, 2.

Lavstsen, T., P. Magistrado, C. C. Hermsen, A. Salanti, A. T. R. Jensen, R. Sauerwein, L. Hviid, T. G. Theander, and T. Staalsoe (2005). Expression of Plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans. *Malar J 4*(1), 21.

Lee, B. K. and F. M. Richards (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol 55*, 379–4000.

Leitner, T., B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (2003). Hiv sequence compendium 2003. *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR04-7420.*.

Li, H., J. J. Dunn, B. J. Luft, and C. L. Lawson (1997, Apr). Crystal structure of Lyme disease antigen outer surface protein A complexed with an Fab. *Proc Natl Acad Sci U S A 94*(8), 3584–3589.

Ljubojevic, S. (2006). The human papillomavirus vaccines. *Acta Dermatovenerol Croat 14*(3), 208.

Lo Conte, L., C. Chothia, and J. Janin (1999, Feb). The atomic structure of protein-protein recognition sites. *J Mol Biol 285*(5), 2177–2198.

Lo Passo, C., A. Romeo, I. Pernice, P. Donato, A. Midiri, G. Mancuso, M. Arigo, C. Biondo, R. Galbo, S. Papasergi, F. Felici, G. Teti, and C. Beninati (2007, Apr). Peptide mimics of the group B meningococcal capsule induce bactericidal and protective antibodies after immunization. *J Immunol 178*(7), 4417–4423.

Lund, O., K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng 10*(11), 1241–8.

Lund, O., M. Nielsen, C. Lundegaard, C. Kesmir, and S. Brunak (2005, September). *Immunological Bioinformatics (Computational Molecular Biology)* (1 ed.). MIT press.

Lupyan, D., A. Leo-Macias, and A. R. Ortiz (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics 21*(15), 3255–63.

Luthy, R., J. U. Bowie, and D. Eisenberg (1992, Mar). Assessment of protein models with three-dimensional profiles. *Nature 356*(6364), 83–85.

MacCallum, R. M., A. C. Martin, and J. M. Thornton (1996, Oct). Antibody-antigen interactions: contact analysis and binding site topography. *J Mol Biol 262*(5), 732–745.

MacKerell, A., D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B 102*(18), 3586–3616.

Mahler, M., M. Bluthner, and K. M. Pollard (2003, May). Advances in B-cell epitope analysis of autoantigens in connective tissue diseases. *Clin Immunol 107*(2), 65–79.

Maksyutov, A. Z. and E. S. Zagrebelnaya (1993, Jun). ADEPT: a computer program for prediction of protein antigenic determinants. *Comput Appl Biosci 9*(3), 291–297.

Matuschewski, K. (2006, Aug). Vaccine development against malaria. *Curr Opin Immunol 18*(4), 449–457.

Mayrose, I., T. Shlomi, N. D. Rubinstein, J. M. Gershoni, E. Ruppin, R. Sharan, and T. Pupko (2007). Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res 35*(1), 69–78.

McGregor, I. A., M. E. Wilson, and W. Z. Billewicz (1983). Malaria infection of the placenta in The Gambia, West Africa; its incidence and relationship to stillbirth, birthweight and placental weight. *Trans R Soc Trop Med Hyg 77*(2), 232–244.

McVean, G., P. Awadalla, and P. Fearnhead (2002, Mar). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics 160*(3), 1231–1241.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly (2004, Apr). The fine-scale structure of recombination rate variation in the human genome. *Science 304*(5670), 581–584.

Melo, F. and E. Feytmans (1997, Mar). Novel knowledge-based mean force potential at atomic level. *J Mol Biol 267*(1), 207–222.

Melo, F. and E. Feytmans (1998, Apr). Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol 277*(5), 1141–1152.

Menendez, C., U. D'Alessandro, and F. O. ter Kuile (2007, Feb). Reducing the burden of malaria in pregnancy by preventive strategies. *Lancet Infect Dis 7*(2), 126–135.

Miller, L. H., D. I. Baruch, K. Marsh, and O. K. Doumbo (2002, Feb). The pathogenic basis of malaria. *Nature 415*(6872), 673–679.

Miller, L. H., S. J. Mason, J. A. Dvorak, M. H. McGinniss, and I. K. Rothman (1975). Erythrocyte receptors for (plasmodium knowlesi) malaria: Duffy blood group determinants. *Science 189*(4202), 561–3.

Mirza, O., A. Henriksen, H. Ipsen, J. N. Larsen, M. Wissenbach, M. D. Spangfort, and M. Gajhede (2000, Jul). Dominant epitopes and allergic cross-reactivity: complex formation between a Fab fragment of a monoclonal murine IgG antibody and the major allergen from birch pollen Bet v 1. *J Immunol 165*(1), 331–338.

Moreau, V., C. Granier, S. Villard, D. Laune, and F. Molina (2006, May). Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics 22*(9), 1088–1095.

Moult, J. (2005, Jun). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol 15*(3), 285–289.

Mu, J., P. Awadalla, J. Duan, K. M. McGee, D. A. Joy, G. A. T. McVean, and X.-z. Su (2005, Oct). Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol 3*(10), e335.

Muller, Y. A., Y. Chen, H. W. Christinger, B. Li, B. C. Cunningham, H. B. Lowman, and A. M. de Vos (1998). Vegf and the fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 a resolution and mutational analysis of the interface. *Structure 6*(9), 1153–67.

Mumey, B. M., B. W. Bailey, B. Kirkpatrick, A. J. Jesaitis, T. Angel, and E. A. Dratz (2003). A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins. *J Comput Biol 10*(3-4), 555–567.

Nei, M. (1997). *Molecular evolutionary genetics* (1 ed.). New York: Columbia University Press.

Nielsen, M., C. Lundegaard, P. Worning, C. S. Hvid, K. Lamberth, S. Buus, S. Brunak, and O. Lund (2004, Jun). Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics 20*(9), 1388–1397.

Nielsen, M. A., M. Resende, M. Alifrangis, L. Turner, L. Hviid, T. G. Theander, and A. Salanti (2007, Sep). Plasmodium falciparum: VAR2CSA expressed during pregnancy-associated malaria is partially resistant to proteolytic cleavage by trypsin. *Exp Parasitol 117*(1), 1–8.

Nielsen, R. and Z. Yang (1998, Mar). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics 148*(3), 929–936.

Nishikawa, K. and T. Ooi (1980). Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int J Pept Protein Res 16*(1), 19–32.

Novotny, J., M. Handschumacher, E. Haber, R. E. Bruccoleri, W. B. Carlson, D. W. Fanning, J. A. Smith, and G. D. Rose (1986, Jan). Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc Natl Acad Sci U S A 83*(2), 226–230.

Nylander, J. A. (2006). MrModeltest, version 2.2. Technical report, Department of systematic zoology, Uppsala Univerisity. Available: http//:www.ebc.uu.se/systzoo/staff/nylander.html. Accessed April 2006.

Odorico, M. and J. L. Pellequer (2003). Bepitope: predicting the location of continuous epitopes and patterns in proteins. *J Mol Recognit 16*(1), 20–2.

Oleinikov, A. V., E. Rossnagle, S. Francis, T. K. Mutabingwa, M. Fried, and P. E. Duffy (2007, Jul). Effects of sex, parity, and sequence variation on seroreactivity to candidate pregnancy malaria vaccine antigens. *J Infect Dis 196*(1), 155–164.

Orlandi, P. A., F. W. Klotz, and J. D. Haynes (1992). A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of plasmodium falciparum recognizes the terminal neu5ac(alpha 2-3)gal- sequences of glycophorin a. *J Cell Biol 116*(4), 901–9.

Ostermeier, C., A. Harrenga, U. Ermler, and H. Michel (1997). Structure at 2.7 Å resolution of the paracoccus denitrificans two-subunit cytochrome c oxidase complexed with an antibody fv fragment. *Proc Natl Acad Sci U S A 94*(20), 10547–53.

Padlan, E. A., E. W. Silverton, S. Sheriff, G. H. Cohen, S. J. Smith-Gill, and D. R. Davies (1989). Structure of an antibody-antigen complex: crystal structure of the hyhel-10 fab-lysozyme complex. *Proc Natl Acad Sci U S A 86*(15), 5938–42.

Parker, J. M., D. Guo, and R. S. Hodges (1986, Sep). New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry 25*(19), 5425–5432.

Paul, W. E. (2003). *Fundamental Immunology* (5th ed.). Philadelphia, US: Lippincott Williams and Wilkins.

Pearce, R. J., C. Drakeley, D. Chandramohan, F. Mosha, and C. Roper (2003, Apr). Molecular determination of point mutation haplotypes in the dihydrofolate reductase and dihydropteroate synthase of Plasmodium falciparum in three districts of northern Tanzania. *Antimicrob Agents Chemother 47*(4), 1347–1354.

Pellequer, J. L. and E. Westhof (1993, Sep). PREDITOP: a program for antigenicity prediction. *J Mol Graph 11*(3), 204–210.

Pellequer, J. L., E. Westhof, and M. H. Van Regenmortel (1991). Predicting location of continuous epitopes in proteins from their primary structures. *Methods Enzymol 203*, 176–201.

Pellequer, J. L., E. Westhof, and M. H. Van Regenmortel (1993, Apr). Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol Lett 36*(1), 83–99.

Peters, B., J. Sidney, P. Bourne, H.-H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J. V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, and A. Sette (2005, Mar). The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol 3*(3), e91. http://tools.immuneepitope.org/tools/bcell/example.jsp.

Pizarro, J. C., B. Vulliez-Le Normand, M. L. Chesne-Seck, C. R. Collins, C. Withers-Martinez, F. Hackett, M. J. Blackman, B. W. Faber, E. J. Remarque, C. H. Kocken, A. W. Thomas, and G. A. Bentley (2005). Crystal structure of the malaria vaccine candidate apical membrane antigen 1. *Science 308*(5720), 408–11.

Posada, D. and T. R. Buckley (2004, Oct). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol 53*(5), 793–808.

Pouvelle, B., T. Fusai, C. Lepolard, and J. Gysin (1998, Oct). Biological and biochemical characteristics of cytoadhesion of Plasmodium falciparum-infected erythrocytes to chondroitin-4-sulfate. *Infect Immun 66*(10), 4950–4956.

Radbruch, A., G. Muehlinghaus, E. O. luger, A. Inamine, K. G. C. Smith, T. Dörner, and F. Hiepe (2006). Competence and competition: the challenge of becominig a long-lived plasma cell. *Nature Reviews Immunology 6*, 741–750.

Rambaut, A. and A. Drummond (2006). Tracer. Technical report, Evolutionary biology group, University of Oxford. Available: http//:www.evolve.zoo.ox.ac.uk/software/html?id=tracer. Accessed April 2006.

Rapberger, R., A. Lukas, and B. Mayer (2007, Mar). Identification of discontinuous antigenic determinants on proteins based on shape complementarities. *J Mol Recognit 20*(2), 113–121.

Ricke, C. H., T. Staalsoe, K. Koram, B. D. Akanmori, E. M. Riley, T. G. Theander, and L. Hviid (2000). Plasma antibodies from malaria-exposed pregnant women recognize variant surface antigens on plasmodium falciparum-infected erythrocytes in a parity-dependent manner and block parasite adhesion to chondroitin sulfate a. *J Immunol 165*(6), 3309–16.

Rogerson, S. J., L. Hviid, P. E. Duffy, R. F. G. Leke, and D. W. Taylor (2007, Feb). Malaria in pregnancy: pathogenesis and immunity. *Lancet Infect Dis 7*(2), 105–117.

Romijn, R. A., E. Westein, B. Bouma, M. E. Schiphorst, J. J. Sixma, P. J. Lenting, and E. G. Huizinga (2003). Mapping the collagen-binding site in the von willebrand factor-a3 domain. *J Biol Chem 278*(17), 15035–9.

Ronquist, F. and J. P. Huelsenbeck (2003, Aug). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*(12), 1572–1574.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas (2003, Dec). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics 19*(18), 2496–2497.

Sabhnani, L., M. Manocha, K. Sridevi, D. Shashikiran, R. Rayanade, and D. N. Rao (2003, Oct). Developing subunit immunogens using B and T cell epitopes and their constructs derived from the F1 antigen of Yersinia pestis using novel delivery vehicles. *FEMS Immunol Med Microbiol 38*(3), 215–229.

Sabo, J. K., D. W. Keizer, Z.-P. Feng, J. L. Casey, K. Parisi, A. M. Coley, M. Foley, and R. S. Norton (2007, Jan). Mimotopes of apical membrane antigen 1: Structures of phage-derived peptides recognized by the inhibitory monoclonal antibody 4G2dc1 and design of a more active analogue. *Infect Immun 75*(1), 61–73.

Saha, S., M. Bhasin, and G. P. S. Raghava (2005). Bcipep: a database of B-cell epitopes. *BMC Genomics 6*(1), 79.

Saha, S. and G. P. S. Raghava (2006, Oct). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins 65*(1), 40–48.

Salanti, A., M. Dahlback, L. Turner, M. A. Nielsen, L. Barfod, P. Magistrado, A. T. Jensen, T. Lavstsen, M. F. Ofori, K. Marsh, L. Hviid, and T. G. Theander (2004). Evidence for the involvement of var2csa in pregnancy-associated malaria. *J Exp Med 200*(9), 1197–203.

Salanti, A., T. Staalsoe, T. Lavstsen, A. T. Jensen, M. P. Sowa, D. E. Arnot, L. Hviid, and T. G. Theander (2003). Selective upregulation of a single distinctly structured var gene in chondroitin sulphate a-adhering plasmodium falciparum involved in pregnancy-associated malaria. *Mol Microbiol 49*(1), 179–91.

Sali, A. and T. L. Blundell (1993, Dec). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol 234*(3), 779–815.

Sali, A., L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus (1995, Nov). Evaluation of comparative protein modeling by MODELLER. *Proteins 23*(3), 318–326.

Saphire, E. O., M. Montero, A. Menendez, N. E. van Houten, M. B. Irving, R. Pantophlet, M. B. Zwick, P. W. H. I. Parren, D. R. Burton, J. K. Scott, and I. A. Wilson (2007, Jun). Structure of a High-affinity "Mimotope" Peptide Bound to HIV-1-neutralizing Antibody b12 Explains its Inability to Elicit gp120 Cross-reactive Antibodies. *J Mol Biol 369*(3), 696–709.

Saphire, E. O., P. W. Parren, R. Pantophlet, M. B. Zwick, G. M. Morris, P. M. Rudd, R. A. Dwek, R. L. Stanfield, D. R. Burton, and I. A. Wilson (2001, Aug). Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. *Science 293*(5532), 1155–1159.

Scherf, A., B. Pouvelle, P. A. Buffet, and J. Gysin (2001, Mar). Molecular mechanisms of Plasmodium falciparum placental adhesion. *Cell Microbiol 3*(3), 125–131.

Schlessinger, A., Y. Ofran, G. Yachdav, and B. Rost (2006). Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res 34*(Database issue), D777–80.

Schonbach, C., J. L. Y. Koh, D. R. Flower, L. Wong, and V. Brusic (2002, Jan). FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res 30*(1), 226–229.

Schreiber, A., M. Humbert, A. Benz, and U. Dietrich (2005, Jul). 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J Comput Chem 26*(9), 879–887.

Selak, S., M. Mahler, K. Miyachi, M. L. Fritzler, and M. J. Fritzler (2003, Nov). Identification of the B-cell epitopes of the early endosome antigen 1 (EEA1). *Clin Immunol 109*(2), 154–164.

Singh, S. K., R. Hora, H. Belrhali, C. E. Chitnis, and A. Sharma (2006, Feb). Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain. *Nature 439*(7077), 741–744.

Slootstra, J. W., W. C. Puijk, G. J. Ligtvoet, D. Kuperus, W. M. Schaaper, and R. H. Meloen (1997, Sep). Screening of a small set of random peptides: a new strategy to identify synthetic peptides that mimic epitopes. *J Mol Recognit 10*(5), 217–224.

Smith, J. D., S. Kyes, A. G. Craig, T. Fagan, D. Hudson-Taylor, L. H. Miller, D. I. Baruch, and C. I. Newbold (1998). Analysis of adhesive domains from the a4var plasmodium falciparum erythrocyte membrane protein-1 identifies a cd36 binding domain. *Mol Biochem Parasitol 97*(1-2), 133–48.

Smith, J. D., G. Subramanian, B. Gamain, D. I. Baruch, and L. H. Miller (2000, Oct). Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family. *Mol Biochem Parasitol 110*(2), 293–310.

Soding, J. (2005, Apr). Protein homology detection by HMM-HMM comparison. *Bioinformatics 21*(7), 951–960.

Soding, J., A. Biegert, and A. N. Lupas (2005, Jul). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res 33*(Web Server issue), 244–248.

Sollner, J. and B. Mayer (2006, May). Machine learning approaches for prediction of linear B-cell epitopes on proteins. *J Mol Recognit 19*(3), 200–208.

Staalsoe, T., C. E. Shulman, J. N. Bulmer, K. Kawuondo, K. Marsh, and L. Hviid (2004). Variant surface antigen-specific igg and protection against clinical consequences of pregnancy-associated plasmodium falciparum malaria. *Lancet 363*(9405), 283–9.

Stanfield, R. L., H. Dooley, P. Verdino, M. F. Flajnik, and I. A. Wilson (2007, Mar). Maturation of shark single-domain (IgNAR) antibodies: evidence for induced-fit binding. *J Mol Biol 367*(2), 358–372.

Su, X. Z., V. M. Heatwole, S. P. Wertheimer, F. Guinet, J. A. Herrfeldt, D. S. Peterson, J. A. Ravetch, and T. E. Wellems (1995). The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. *Cell 82*(1), 89–100.

Swets, J. A. (1988, Jun). Measuring the accuracy of diagnostic systems. *Science 240*(4857), 1285–1293.

Tajima, F. (1989, Nov). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics 123*(3), 585–595.

Tarlinton, D. and A. Lew (2007, Apr). Antigen to the node: B cells go native. *Immunity 26*(4), 388–390.

Tarr, A. W., A. M. Owsianka, J. M. Timms, C. P. McClure, R. J. P. Brown, T. P. Hickling, T. Pietschmann, R. Bartenschlager, A. H. Patel, and J. K. Ball (2006, Mar). Characterization of the hepatitis C virus E2 epitope defined by the broadly neutralizing monoclonal antibody AP33. *Hepatology 43*(3), 592–601.

Taylor, H. M., S. A. Kyes, and C. I. Newbold (2000, Oct). Var gene diversity in Plasmodium falciparum is generated by frequent recombination events. *Mol Biochem Parasitol 110*(2), 391–397.

Tetteh, K. K. A., D. R. Cavanagh, P. Corran, R. Musonda, J. S. McBride, and D. J. Conway (2005, Sep). Extensive antigenic polymorphism within the repeat sequence of the Plasmodium falciparum merozoite surface protein 1 block 2 is incorporated in a minimal polyvalent immunogen. *Infect Immun 73*(9), 5928–5935.

Thornton, J. M., M. S. Edwards, W. R. Taylor, and D. J. Barlow (1986, Feb). Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO J 5*(2), 409–413.

Timmerman, P., E. Van Dijk, W. Puijk, W. Schaaper, J. Slootstra, S. J. Carlisle, J. Coley, S. Eida, M. Gani, T. Hunt, P. Perry, G. Piron, and R. H. Meloen (2004). Mapping of a discontinuous and highly conformational binding site on follicle stimulating hormone subunit-beta (FSH-beta) using domain Scan and Matrix Scan technology. *Mol Divers 8*(2), 61–77.

Tolia, N. H., E. J. Enemark, B. K. L. Sim, and L. Joshua-Tor (2005, Jul). Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite Plasmodium falciparum. *Cell 122*(2), 183–193.

Topf, M. and A. Sali (2005, Oct). Combining electron microscopy and comparative protein structure modeling. *Curr Opin Struct Biol 15*(5), 578–585.

Toseland, C. P., D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I. A. Doytchinova, P. Guan, C. K. Hattotuwagama, and D. R. Flower (2005, Oct). AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res 1*(1), 4.

Trimnell, A. R., S. M. Kraemer, S. Mukherjee, D. J. Phippard, J. H. Janes, E. Flamoe, X.-z. Su, P. Awadalla, and J. D. Smith (2006, Aug). Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Mol Biochem Parasitol 148*(2), 169–180.

Tuikue Ndam, N. G., N. Fievet, G. Bertin, G. Cottrell, A. Gaye, and P. Deloron (2004, Dec). Variable adhesion abilities and overlapping antigenic properties in placental Plasmodium falciparum isolates. *J Infect Dis 190*(11), 2001–2009.

Tuikue Ndam, N. G., A. Salanti, G. Bertin, M. Dahlback, N. Fievet, L. Turner, A. Gaye, T. Theander, and P. Deloron (2005, Jul). High level of var2csa transcription by Plasmodium falciparum isolated from the placenta. *J Infect Dis 192*(2), 331–335.

Tuikue Ndam, N. G., A. Salanti, J.-Y. Le-Hesran, G. Cottrell, N. Fievet, L. Turner, S. Sow, J.-M. Dangou, T. Theander, and P. Deloron (2006, Mar). Dynamics of anti-VAR2CSA immunoglobulin G response in a cohort of senegalese pregnant women. *J Infect Dis 193*(5), 713–720.

Untersmayr, E., K. Szalai, A. B. Riemer, W. Hemmer, I. Swoboda, B. Hantusch, I. Scholl, S. Spitzauer, O. Scheiner, R. Jarisch, G. Boltz-Nitulescu, and E. Jensen-Jarolim (2006, Mar). Mimotopes identify conformational epitopes on parvalbumin, the major fish allergen. *Mol Immunol 43*(9), 1454–1461.

Uthaipibull, C., B. Aufiero, S. E. Syed, B. Hansen, J. A. Guevara Patino, E. Angov, I. T. Ling, K. Fegeding, W. D. Morgan, C. Ockenhouse, B. Birdsall, J. Feeney, J. A. Lyon, and A. A. Holder (2001, Apr). Inhibitory and blocking monoclonal antibody epitopes on merozoite surface protein 1 of the malaria parasite Plasmodium falciparum. *J Mol Biol 307*(5), 1381–1394.

Valmori, D., N. E. Souleimanian, C. S. Hesdorffer, G. Ritter, L. J. Old, and M. Ayyoub (2005, Oct). Identification of B cell epitopes recognized by antibodies specific for the tumor antigen NY-ESO-1 in cancer patients with spontaneous immune responses. *Clin Immunol 117*(1), 24–30.

Van Regenmortel, M. H. and J. L. Pellequer (1994). Predicting antigenic determinants in proteins: looking for unidimensional solutions to a three-dimensional problem? *Pept Res 7*(4), 224–8.

Van Regenmortel, M. H. V. (1996, Jun). Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. *Methods 9*(3), 465–472.

Van Regenmortel, M. H. V. (2007, Mar). The rational design of biological complexity: a deceptive metaphor. *Proteomics 7*(6), 965–975.

Venclovas, C., A. Zemla, K. Fidelis, and J. Moult (2003). Assessment of progress over the CASP experiments. *Proteins 53 Suppl 6*, 585–595.

Viebig, N. K., B. Gamain, C. Scheidig, C. Lepolard, J. Przyborski, M. Lanzer, J. Gysin, and A. Scherf (2005, Aug). A single member of the Plasmodium falciparum var multigene family determines cytoadhesion to the placental receptor chondroitin sulphate A. *EMBO Rep 6*(8), 775–781.

Walter, P. R., Y. Garin, and P. Blot (1982, Dec). Placental pathologic changes in malaria. A histologic and ultrastructural study. *Am J Pathol 109*(3), 330–342.

Wang, L.-F. and M. Yu (2004, Jan). Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Curr Drug Targets 5*(1), 1–15.

Wedemayer, G. J., P. A. Patten, L. H. Wang, P. G. Schultz, and R. C. Stevens (1997, Jun). Structural insights into the evolution of an antibody combining site. *Science 276*(5319), 1665–1669.

Weintraub, A. (2003, Nov). Immunology of bacterial polysaccharide antigens. *Carbohydr Res 338*(23), 2539–2547.

Weiss, L. (1990). The spleen in malaria: the role of barrier cells. *Immunol Lett 25*(1-3), 165–72.

Welling, G. W., W. J. Weijer, R. van der Zee, and S. Welling-Wester (1985, Sep). Prediction of sequential antigenic regions in proteins. *FEBS Lett 188*(2), 215–218.

Wernersson, R. and A. G. Pedersen (2003, Jul). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res 31*(13), 3537–3539.

Wilson, I. A. and R. L. Stanfield (1993, Feb). Antibody-antigen interactions: new structures and new conformational changes. *Curr Opin Struct Biol 3*(1), 113–118.

Yang, W., J. P. Bielawski, and Z. Yang (2003, Aug). Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol 57*(2), 212–221.

Yang, Z. (1997, Oct). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci 13*(5), 555–556.

Yang, Z., R. Nielsen, and M. Hasegawa (1998, Dec). Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol 15*(12), 1600–1611.

Zhang, Y.-L., Y.-J. Guo, K.-Y. Wang, K. Lu, K. Li, Y. Zhu, and S.-H. Sun (2007, Apr). Enhanced immunogenicity of modified hepatitis B virus core particle fused with multiepitopes of foot-and-mouth disease virus. *Scand J Immunol 65*(4), 320–328.

Zhou, T., L. Xu, B. Dey, A. J. Hessell, D. Van Ryk, S.-H. Xiang, X. Yang, M.-Y. Zhang, M. B. Zwick, J. Arthos, D. R. Burton, D. S. Dimitrov,

J. Sodroski, R. Wyatt, G. J. Nabel, and P. D. Kwong (2007, Feb). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature 445*(7129), 732–737.

# List of Figures

# Appendix I, Data sets used in Paper I

## DiscoTope data set 1

Each entry is marked with the PDB code. For each entry the antigen sequence is shown first. The next lines mark the identification status of the residues: "." marks non-epitope residue and epitopes are marked with capital letters according to the identifier of the contacting antibody chain.

```
200 1JPS.T
NTVAAYNLTWKSTNFKTILEWEPKPVNQVYTVQISTKSGDWKSKCFYTTDTECDLTDEIVKDVKQTYLARVFSYPAGNVE
PLYENSPEFTPYLETNLGQPTIQSFEQVGTKVNVTVEDERTLVRRNNTFLSLRDVFGKDLIYTLYYWKSSSSGKKTAKTN
TNEFLIDVDKGENYCFSVQAVIPSRTVNRKSTDSPVECMG
.............................................................................
.............................................................H.H.H.......HHLLLLLL
..................H.H.H..HHHHH.HH.......

95 1JRH.I
SVPTPTNVTIESYNMNPIVYWEYQIMPQVPVFTVEVKNYGVKNSEWIDACINISHHYCNISDHVGDPSNSLWVRVKARVG
QKESAYAKSEEFAVS
....................................L.LHHHHLLL...................H.HH..L.L......
.......L.......

91 1XIW.A
QTPYKVSISGTTVILTCPQYPGSEILWQHNDKNIGGDEDDKNIGSDEDHLSLKEFSELEQSGYYVCYPRGSKPEDANFYL
YLRARVCENCM
.......................D........DDDDDD......DD....................D.DCDCDCC.....
...........

251 1FJ1.F
SLDEKNSVSVDLPGEMKVLVSKEKNKDGKYDLIATVDKLELKGTSDKNNGSGVLEGVKADKCKVKLTISDDLGQTTLEVF
KEDGKTLVSKKVTSKDKSSTEEKFNEKGEVSEKIITRADGTRLEYTGIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVVKE
GTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKENTITVQQYDSNGTKLEGSAVEI
TKLDEIKNALK
.............................................................................
.............................................................................
.......................AAAB.A.AAA..........A.BAB..................BBBAB.........
...........

251 1OSP.O
SLDEKNSVSVDLPGEMKVLVSKEKNKDGKYDLIATVDKLELKGTSDKNNGSGVLEGVKADKCKVKLTISDDLGQTTLEVF
KEDGKTLVSKKVTSKDKSSTEEKFNEKGEVSEKIITRADGTRLEYTGIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVVKE
```

```
GTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKENTITVQQYDSNGTKLEGSAVEI
TKLDEIKNALK
.......L..HHH......LLLLL.....L................LLH....................HHHH.......
..............HHH..............................................................
.......................................................................................
...........

196 1FNS.A
MYCSRLLDLVFLLDGSSRLSEAEFEVLKAFVVDMMERLRVSQKWVRVAVVEYHDGSHAYIGLKDRKRPSELRRIASQVKY
AGSQVASTSEVLKYTLFQIFSKIDRPEASRIALLLMASQEPQRMSRNFVRYVQGLKKKKVIVIPVGIGPHANLKQIRLIE
KQAPENKAFVLSSVDELEQQRDEIVSYLCDLAPEAP
.......................................................................................
.......................................HHH.HHHHHL.......................HH..H..
...................................

196 1OAK.A
MYCSRLLDLVFLLDGSSRLSEAEFEVLKAFVVDMMERLRISQKWVRVAVVEYHDGSHAYIGLKDRKRPSELRRIASQVKY
AGSQVASTSEVLKYTLFQIFSKIDRPEASRIALLLMASQEPQRMSRNFVRYVQGLKKKKVIVIPVGIGPHANLKQIRLIE
KQAPENKAFVLSSVDELEQQRDEIVSYLCDLAPEAP
.......................................................................................
.......................................HHH.HLHHHL.....................H.HH..H..
...................................

186 1FE8.A
PDCSQPLDVILLLDGSSSFPASYFDEKSFAKAFISKANIGPRLTQVSVLQYGSITTIDVPWNVVPEKAHLLSLVDVQREG
GPSQIGDALGFAVRYLTSEHGARPGASKAVVILVTDVSVDSVDAAADAARSNRVTVFPIGIGDRYDAAQLRILAGPAGDS
NVVKLQRIEDLPTVTLGNSFLHKLCS
.......................................LL..L..............HHHHHLL.HH..HH.HH....
...................LL.L.............................................................
..........................

184 1MHP.A
TQLDIVIVLDGSNSIYPWESVIAFLNDLLKRMDIGPKQTQVGIVQYGENVTHEFNLNKYSSTEEVLVAANKIVQRGGRQT
MTALGIDTARKEAFTEARGARRGVKKVMVIVTDGESHDNYRLKQVIQDCEDENIQRFSIAILGTEKFVEEIKSIASEPTE
KHFFNVSDELALVTIVKALGERIF
...........HHH.H...............................L........................HHHHHHH
..................................L.H.LL.....................................
........................

94 1TZH.V
VVKFMDVYQRSYCHPIETLVDIFQEYPDEIEYIFKPSCVPLMRCGGCCNDEGLECVPTEESNITMQIMRIKPHQGQHIGE
MSFLQHNKCECRPK
...............................H................................H..HLHLLLHHHHH.
..............

94 1CZ8.W
VVKFMDVYQRSYCHPIETLVDIFQEYPDEIEYIFKPSCVPLMRCGGCCNDEGLECVPTEESNITMQIMRIKPHQGQHIGE
MSFLQHNKCECRPK
...............................H..H.............................HHHHH..HHHHHHHH
H.............

94 1BJ1.W
VVKFMDVYQRSYCHPIETLVDIFQEYPDEIEYIFKPSCVPLMRCGGCCNDEGLECVPTEESNITMQIMRIKPHQGQHIGE
MSFLQHNKCECRPK
...............................H..H.............................HHHHH..HHHHHHHH
H.............
```

# DiscoTope data set 2

Each entry is marked with the PDB code. For each entry the antigen sequence is shown first. The next lines mark the identification status of the residues: "." marks non-epitope residue and epitopes are marked with capital letters according to the identifier of the contacting antibody chain.

```
129 1A2Y.C
KVFGRCELAAAMKRHGLANYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..................A..BBB..B...................................................
.....................BB...........BBBBBB..AA....

129 1BVK.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B...................................................
.....................BB...........BBBBBB..AA....

129 3HFL.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.......................................H.HLLLLLL...H.............HHHH..........
...H.........................................

129 1IC4.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............LL..LLL.......................................HH.........H.H.....
........L...L..LHH.HHHH.......................

129 1FDL.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
................LL..HHH..H...................................................
..................................HHHHHL..LL....

129 1IC5.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............LL..LLL.......................................H.........H.H.....
........L...L..LHH.HHH.......................

129 1C08.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............AA..AAA.......................................B.........B.B.....
........A...A..ABB.BBBB.......................

129 1NDM.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............AA.AAAA.......................................B.........B.B.....
........A...A..ABB.ABBB.......................

129 1IC7.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............LL..LLL.......................................H.........H.H.....
........L...L..LHH.HHHH.......................

129 1MLC.E
```

```
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.......................................B.B.BAAAAAA.B.............BA.A........B.
B..B.............................................

129 1KIP.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B...................................................
.....................B............BBBBBB..AA....

129 1G7L.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B...................................................
.....................B............BBBBBA..AA....

129 1DQJ.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.............AAA.AAAA.......................................BB.........B.B.B...
........A...A..ABB.ABBB.........................

129 1KIR.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..................A..BBB..B....................................................
.....................BB...........BBBBBB...A....

129 1G7M.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B...................................................
.....................BB...........BBBBBA..A.....

129 1KIQ.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B...................................................
.....................B............BBBBBA..AA....

129 1G7H.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B...................................................
.....................BB...........BBBBBA..AA....

129 1J1X.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............LL..LLL.......................................HH.........H.H.H...
........L...L..LHH.LHHH.........................

129 1J1P.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.............LLL..LLL.......................................HH.........H.H.H...
........L...L..LHH.HHHH.........................

129 1J1O.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
```

```
..............LL..LLL.......................................HH.........H.H.H...
........L...L..HHH.HHHH.........................

129 1G7I.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B.................................................
.....................B............BBBBBA..AA....

129 1G7J.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.................AA..BBB..B.................................................
.....................B............BBBBBA..AA....

129 1NBZ.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKAIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
..............AA.AAAA.......................................B.........B.B.B...
........A...A..ABB.BBBB.........................

129 1NBY.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAAKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
............AAAA..AAA.......................................B.........B.B.....
........A...A..ABB.ABBB.........................

129 1NDG.C
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSAWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.............AAA..AAB.......................................BB........BBBB.B...
........A...A..ABB.BBB..........................

127 1MEL.L
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC
SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGC
...................................A..........AAA...A....AAA.AAA.........A.A.....
.................A..A.A..AAAA..A...A...........

129 1BQL.Y
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNTDGSTDYGVLQINSRWWCNDGKTPGSRNLCNIPC
SALLSSDITATVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
.......................................H.HLLL.L....H.............HHHH..........
H..H.............................................

129 1DZB.X
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCNIPC
SALLSSDITASVNCAKKIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL
....................A.A.....................................AAA.........A.A.....
...............AAA.AAAAA.A.....A...A............

129 1JHL.A
KVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESNFNTGATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCHIPC
SALLSSDITASVNCAKKIVSDGDGMNAWVAWRKHCKGTDVNVWIRGCRL
.....................H.H.........................................................
......................H..H.....HH..HLLL.L........

132 1ORS.C
DVMEHPLVELGVSYAALLSVIVVVVEYTMQLSGEYLVRLYLVDLILVIILWADYAYRAYKSGDPAGYVKKTLYEIPALVP
AGLLALIEGHLAGLGLFRLVRLLRFLRILLIISRGSKFLSAIADAADKLVPR
.............................................................................
.......B...BABBBBB.B...B........................
```

```
103 1K4D.C
SALHWRAAGAATVLLVIVLLAGSYLAVLAERGAPGAQLITYPRALWWSVETATTVGYGDLYPVTLWGRCVAVVVMVAGIT
SFGLVTAALATWFVGREQERRGH
.......................A...A..BBBBBBB.BAA.B....................................
.......................

103 1K4C.C
SALHWRAAGAATVLLVIVLLAGSYLAVLAERGAPGAQLITYPRALWWSVETATTVGYGDLYPVTLWGRCVAVVVMVAGIT
SFGLVTAALATWFVGREQERRGH
.......................A...AA.BBBBBBB.BAA.B....................................
.......................

136 1LK3.A
NMLRDLRDAFSRVKTFFQMKDQLDNLLLKESLLEDFKGYLGCQALSEMIQFYLEEVMPQAENQDPDIKAHVNSLGENLKT
LRLRLRRCHRFLPCEKSKAVEQVKNAFNKLQEKGIYKAMSEFDIFINYIEAYMTMK
..................L...............................H..HH..HH..HH.................
.......................H..HHLLLLL.HL....................

85 2JEL.P
MFQQEVTITAPNGLHTRPAAQFVKEAKGFTSEITVTSNGKSASAKSLFKLQTLGLTQGTVVTISAEGEDEQKAVEHLVKL
MAELE
LHLH............................L......L......................L.LHH.HHH..HH....
.....

581 1N8Z.C
TQVCTGTDMKLRLPASPETHLDMLRHLYQGCQVVQGNLELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIV
RGTQLFEDNYALAVLDNGDPLSPGGLRELQLRSLTEILKGGVLIQRNPQLCYQDTILWKDIFHKNNQLALTLIDTNRSRA
CHPCSPMCKGSRCWGESSEDCQSLTRTVCAGGCARCKGPLPTDCCHEQCAAGCTGPKHSDCLACLHFNHSGICELHCPAL
VTYNTDTFESMPNPEGRYTFGASCVTACPYNYLSTDVGSCTLVCPLHNQEVTATQRCEKCSKPCARVCYGLGMEHLREVR
AVTSANIQEFAGCKKIFGSLAFLPESFDSNTAPLQPEQLQVFETLEEITGYLYISAWPDSLPDLSVFQNLQVIRGRILHN
GAYSLTLQGLGISWLGLRSLRELGSGLALIHHNTHLCFVHTVPWDQLFRNPHQALLHTANRPEDECVGEGLACHQLCARG
HCWGPGPTQCVNCSQFLRGQECVEECRVLQGLPREYVNARHCLPCHPECQPQNGSVTCFGPEADQCVACAHYKDPPFCVA
RCPSIWKFPDEEGACQPCPIN
.......................................................................
.......................................................................
.......................................................................
.......................................................................
.......................................................................
.......................................................................
.......................................................BB.AB.......ABAAB...
....B.B..A.A.A.AAA...
```

# DiscoTope data set 3

Each entry is marked with the PDB code. For each entry the antigen sequence is shown first. The next lines mark the identification status of the residues: "." marks non-epitope residue and epitopes are marked with capital letters according to the identifier of the contacting antibody chain.

```
122 1HOD.C
DNSRYTHFLTQHYDAKPQGRDDRYCESIMRRRGLTSPCKDINTFIHGNKRSIKAICENKNGNPHRENLRISKSSFQVTTC
KLHGGSPWPPCQYRATAGFRNVVVACENGLPVHLDQSIFRRP
...............................AAAAABBB.....................................
.B.BBBBABB.B............................

156 1IQD.C
CSMPLGMESKAISDAQITASSYFTNMFATWSPSKARLHLQGRSNAWRPQVNNPKEWLQVDFQKTMKVTGVTTQGVKSLLT
SMYVKEFLISSSQDGHQWTLFFQNGKVKVFQGNQDSFTPVVNCLDPPLLTRYLRIHPQSWVHQIALRMEVLGCEAQ
......................AAAAB..............B....B.AA.........................BBBB
.B..........................................................AA.............

185 1KYO.E
KSTYRTPNFDDVLKENNDADKGRSYAYFMVGAMGLLSSAGAKSTVETFISSMTATADVLAMAKVEVNLAAIPLGKNVVVK
WQGKPVFIRHRTPHEIQEANSVDMSALKDPQTDADRVKDPQWLIMLGICTHLGCVPIGEAGDFGGWFCPCHGSHYDISGR
IRKGPAPLNLEIPAYEFDGDKVIVG
...............................................................................
............K..KJ..JJ.JJ.......JJJ..JJKJK....................................
.........................

185 1EZV.E
KSTYRTPNFDDVLKENNDADKGRSYAYFMVGAMGLLSSAGAKSTVETFISSMTATADVLAMAKVEVNLAAIPLGKNVVVK
WQGKPVFIRHRTPHEIQEANSVDMSALKDPQTDADRVKDPQWLIMLGICTHLGCVPIGEAGDFGGWFCPCHGSHYDISGR
IRKGPAPLNLEIPAYEFDGDKVIVG
...............................................................................
............YY.YY..XXXXX.......XXX..XXYXY....................................
.........................

389 1NCA.N
IRDFNNLTKGLCTINSWHIYGKDNAVRIGEDSDVLVTREPYVSCDPDECRFYALSQGTTIRGKHSNGTIHDRSQYRALIS
WPLSSPPTVYNSRVECIGWSSTSCHDGKTRMSICISGPNNNASAVIWYNRRPVTEINTWARNILRTQESECVCHNGVCPV
VFTDGSATGPAETRIYYFKEGKILKWEPLAGTAKHIEECSCYGERAEITCTCRDNWQGSNRPVIRIDPVAMTHTSQYICS
PVLTDNPRPNDPTVGKCNDPYPGNNNNGVKGFSYLDGVNTWLGRTISIASRSGYEMLKVPNALTDDKSKPTQGQTIVLNT
DWSGYSGSFMDYWAEGECYRACFYVELIRGRPKEDKVWWTSNSIVSMCSSTEFLGQWDWPDGAKIEYFL
...............................................................................
...............................................................................
...............................................................................
......LLLL............LLL.L..................HHHLH.H.........................HHH
HH.............................LL...................................

388 1NMC.N
RDFNNLTKGLCTINSWHIYGKDNAVRIGEDSDVLVTREPYVSCDPDECRFYALSQGTTIRGKHSNGTIHDRSQYRALISW
PLSSPPTVYNSRVECIGWSSTSCHDGKTRMSICISGPNNNASAVIWYNRRPVTEINTWARNILRTQESECVCHNGVCPVV
FTDGSATGPAETRIYYFKEGKILKWEPLAGTAKHIEECSCYGERAEITCTCRDNWQGSNRPVIRIDPVAMTHTSQYICSP
VLTDNPRPNDPTVGKCNDPYPGNNNNGVKGFSYLDGVNTWLGRTISIASRSGYEMLKVPNALTDDKSKPTQGQTIVLNTD
WSGYSGSFMDYWAEGECYRACFYVELIRGRPKEDKVWWTSNSIVSMCSSTEFLGQWDWPDGAKIEYFL
...............................................................................
...............................................................................
...............................................................................
.......LLLL..........LL.....................HHLHH...........................HH.
H..............................H...................................

388 1A14.N
RDFNNLTKGLCTINSWHIYGKDNAVRIGEDSDVLVTREPYVSCDPDECRFYALSQGTTIRGKHSNGTIHDRSQYRALISW
PLSSPPTVYNSRVECIGWSSTSCHDGKTRMSICISGPNNNASAVIWYNRRPVTEINTWARNILRTQESECVCHNGVCPVV
```

```
FTDGSATGPAETRIYYFKEGKILKWEPLAGTAKHIEECSCYGERAEITCTCRDNWQGSNRPVIRIDPVAMTHTSQYICSP
VLTDNPRPNDPTVGKCNDPYPGNNNNGVKGFSYLDGVNTWLGRTISIASRSGYEMLKVPNALTDDKSKPTQGQTIVLNTD
WSGYSGSFMDYWAEGECYRACFYVELIRGRPKEDKVWWTSNSIVSMCSSTEFLGQWDWPDGAKIEYFL
.............................................................................
.............................................................................
.............................................................................
.......LLLL........LLLL.....................H.LHH.H.........................HH.
H.............................H...................................

389 1NCB.N
IRDFNNLTKGLCTINSWHIYGKDNAVRIGEDSDVLVTREPYVSCDPDECRFYALSQGTTIRGKHSNGTIHDRSQYRALIS
WPLSSPPTVYNSRVECIGWSSTSCHDGKTRMSICISGPNNNASAVIWYNRRPVTEINTWARNILRTQESECVCHNGVCPV
VFTDGSATGPAETRIYYFKEGKILKWEPLAGTAKHIEECSCYGERAEITCTCRDNWQGSNRPVIRIDPVAMTHTSQYICS
PVLTDNPRPDDPTVGKCNDPYPGNNNNGVKGFSYLDGVNTWLGRTISIASRSGYEMLKVPNALTDDKSKPTQGQTIVLNT
DWSGYSGSFMDYWAEGECYRACFYVELIRGRPKEDKVWWTSNSIVSMCSSTEFLGQWDWPDGAKIEYFL
.............................................................................
.............................................................................
.............................................................................
......LLLL............LL..L..................HHHHH.H........................HHH
HH............................LL...................................

389 1NCC.N
IRDFNNLTKGLCTINSWHIYGKDNAVRIGEDSDVLVTREPYVSCDPDECRFYALSQGTTIRGKHSNGTIHDRSQYRALIS
WPLSSPPTVYNSRVECIGWSSTSCHDGKTRMSICISGPNNNASAVIWYNRRPVTEINTWARNILRTQESECVCHNGVCPV
VFTDGSATGPAETRIYYFKEGKILKWEPLAGTAKHIEECSCYGERAEITCTCRDNWQGSNRPVIRIDPVAMTHTSQYICS
PVLTDNPRPNDPTVGKCNDPYPGNNNNGVKGFSYLDGVNTWLGRTISRASRSGYEMLKVPNALTDDKSKPTQGQTIVLNT
DWSGYSGSFMDYWAEGECYRACFYVELIRGRPKEDKVWWTSNSIVSMCSSTEFLGQWDWPDGAKIEYFL
.............................................................................
.............................................................................
.............................................................................
......LLLL............LLL.L.................HHHHH.H........................HHH
HH............................LL...................................

389 1NCD.N
IREFNNLTKGLCTINSWHIYGKDNAVRIGEDSDVLVTREPYVSCDPDECRFYALSQGTTIRGKHSNGTIHDRSQYRDLIS
WPLSSPPTVYNSRVECIGWSSTSCHDGRARMSICISGPNNNASAVIWYNRRPVTEINTWARNILRTQESECVCQNGVCPV
VFTDGSATGPAETRIYYFKEGKILKWEPLTGTAKHIEECSCYGEQAGVTCTCRDNWQGSNRPVIQIDPVAMTHTSQYICS
PVLTDNPRPNDPTVGKCNDPYPGNNNNGVKGFSYLDGGNTWLGRTISIASRSGYEMLKVPNALTDDRSKPTQGQTIVLNT
DWSGYSGSFMDYWAEGECYRACFYVELIRGRPKEDKVWWTSNSIVSMCSSTEFLGQWNWPDGAKIEYFL
.............................................................................
.............................................................................
.............................................................................
......LLLL............LL..L.................HHHLH.H........................HHH
.H............................LL...................................

444 1OTS.A
RRRQLIRQLLERDKTPLAILFMAAVVGTLVGLAAVAFDKGVAWLQNQRMGALVHTADNYPLLLTVAFLCSAVLAMFGYFL
VRKYAPEAGGSGIPEIEGALEDQRPVRWWRVLPVKFFGGLGTLGGGMVLGREGPTVQIGGNIGRMVLDIFRLKGDEARHT
LLATGAAAGLAAAFNAPLAGILFIIEEMRPQFRYTLISIKAVFIGVIMSTIMYRIFNHEVALIDVGKLSDAPLNTLWLYL
ILGIIFGIFGPIFNKWVLGMQDLLHRVHGGNITKWVLMGGAIGGLCGLLGFVAPATSGGGFNLIPIATAGNFSMGMLVFI
FVARVITTLLCFSSGAPGGIFAPMLALGTVLGTAFGMVAVELFPQYHLEAGTFAIAGMGALLAASIRAPLTGIILVLEMT
DNYQLILPMIITGLGATLLAQFTGGKPLYSAILARTLAKQEAEQ
.............................................................................
.............................................................................
...............................................................C..C.CCC......
.............................................................................
.........................................DDCD.................................
.........................................
```

# DiscoTope data set 4

Each entry is marked with the PDB code. For each entry the antigen sequence is shown first. The next lines mark the identification status of the residues: "." marks non-epitope residue and epitopes are marked with capital letters according to the identifier of the contacting antibody chain.

```
305 1RZJ.G
EVVLVNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVGAGSCNTSVITQACPKVSFEPIPIHYCAPAGFAIL
KCNNKTFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSVNFTDNAKTIIVQLNTSVEINCTGAGHCNISR
AKWNNTLKQIASKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFNSTWFNGSDTITLPCRIKQIINMW
QKVGKAMYAPPISGQIRCSSNITGLLLTRDGGNSNNESEIFRPGGGDMRDNWRSELYKYKVVKIE
.................................LH.H.............LL.....................
.........................................................................
.................................................................H.HHH....
......H..H.................................................

305 1G9M.G
EVVLVNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVGAGSCNTSVITQACPKVSFEPIPIHYCAPAGFAIL
KCNNKTFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSVNFTDNAKTIIVQLNTSVEINCTGAGHCNISR
AKWNNTLKQIASKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFNSTWFNGSDTITLPCRIKQIINMW
QKVGKAMYAPPISGQIRCSSNITGLLLTRDGGNSNNESEIFRPGGGDMRDNWRSELYKYKVVKIE
.................................LH.H............H.LL.....................
.........................................................................
.................................................................H.HHH....
......L..H.................................................

297 1GC1.G
TENFNMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVGAGSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNNKTF
NGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIRSVNFTDNAKTIIVQLNTSVEINCTGAGHCNISRAKWNNTL
KQIASKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFNSTWFGSDTITLPCRIKQIINMWQKVGKAMY
APPISGQIRCSSNITGLLLTRDGGNSNNESEIFRPGGGDMRDNWRSELYKYKVVKIE
...........................L.H.............H.LL...........................
.........................................................................
.............................................................H.HHH.........L.
.H.................................................

306 1RZK.G
LENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVKLTPLCVGAGSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCN
DKKFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEIVIRSENFTNNAKTIIVQLNESVVINCTGAGHCNLSKTQW
ENTLEQIAIKLKEQFGNNKTIIFNPSSGGDPEIVTHSFNCGGEFFYCNSTQLFTWNDTRKLNNTGRNITLPCRIKQIINM
WQEVGKAMYAPPIRGQIRCSSNITGLLLTRDGGKDTNGTEIFRPGGGDMRDNWRSELYKYKVVKIE
...............................LH.H.............H.LL.......................
.........................................................................
.................................................................H.HHH...
.......H..H.................................................

306 1G9N.G
LENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVKLTPLCVGAGSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCN
DKKFNGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEIVIRSENFTNNAKTIIVQLNESVVINCTGAGHCNLSKTQW
ENTLEQIAIKLKEQFGNNKTIIFNPSSGGDPEIVTHSFNCGGEFFYCNSTQLFTWNDTRKLNNTGRNITLPCRIKQIINM
WQEVGKAMYAPPIRGQIRCSSNITGLLLTRDGGKDTNGTEIFRPGGGDMRDNWRSELYKYKVVKIE
...............................LH.H.............H.LL.......................
.........................................................................
.................................................................H.HHH...
.......H..H.................................................

102 1TQB.A
GLGGYMLGSVMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDQYSNQNNFVHDCVNITVKQHTVTTTTKGENFTETDI
KIMERVVEQMCITQYQRESQAY
BB..........................C.CC........................B..BB.BBBCBCBCCC......
.....................
```

```
102 1TPX.A
GLGGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDQYSNQNNFVHDCVNITVKQHTVTTTTKGENFTETDI
KIMERVVEQMCITQYQRESQAY
BB..........................C.CC........................B..BB.BBBCBCBCCC......
.....................

102 1TQC.A
GLGGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDRYSNQNNFVHDCVNITVKQHTVTTTTKGENFTETDI
KIMERVVEQMCITQYQRESQAY
BB......................C...C.CC........................B..BB.BBBCBBCCCC......
.....................

252 1AR1.B
QDVLGDLPVIGKPVNGGMNFQPASSPLAHDQQWLDHFVLYIITAVTIFVCLLLLICIVRFNRRANPVPARFTHNTPIEVI
WTLVPVLILVAIGAFSLPILFRSQEMPNDPDLVIKAIGHQWYWSYEYPNDGVAFDALMLEKEALADAGYSEDEYLLATDN
PVVVPVGKKVLVQVTATDVIHAWTIPAFAVKQDAVPGRIAQLWFSVDQEGVYFGQCSELCGINHAYMPIVVKAVSQEKYE
AWLAGAKEEFAA
.......................DD..DD.................................................
.....................................................................................
.....CCC....................................CCDD.........................CCC..C
............

115 1OAZ.A
SDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPIEESDDRRYDLVGPCKMIAPILDEILTVAKLNIDQNPGTAPKYGIR
GIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLA
...............................HHH.H...H.L..L.L..............................H
HH................HHLL..L..........

239 1NFD.B
DSGVVQSPRHIIKEKGGRSVLTCIPISGHSNVVWYQQTLGKELKFLIQHYEKVERDKGFLPSRFSVQQFDDYHSEMNMSA
LELEDSAMYFCASSLRWGDEQYFGPGTRLTVLEDLRNVTPPKVSLFEPSKAEIANKQKATLVCLARGFFPDHVELSWWVN
GKEVHSGVSTDPQAYKESNYSYSLSSRLRVSATFWHNPRNHFRCQVQFHGLSEEDKWPEGSPKPVTQNISAEAWGRADC
.............E.......................................................................
................................E.F...........................................
...................................................E.EEEEFFFF................
```

# DiscoTope data set 5

Each entry is marked with the PDB code. For each entry the antigen sequence is shown first. The next lines mark the identification status of the residues: "." marks non-epitope residue and epitopes are marked with capital letters according to the identifier of the contacting antibody chain.

```
430 2HMI.B
PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFREL
NKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIF
QSSMTKILEPFKKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWT
VQPIVLPEKDSWTVNDIQKLVGKLNWASQIYPGIKVRQLSKLLRGTKALTEVIPLTEEAELELAENREILKEPVHGVYYD
PSKDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMRGAHTNDVKQLTEAVQKITTESIVIWGKTPKFKLPIQKETWET
WWTEYWQATWIPEWEFVNTPPLVKLWYQLE
.....................................D.......................DDCDD.DD..........
.....................................C.........................................
..............................

101 1EGJ.A
IQMAPPSLNVTKDGDSYSLRWETMKMRYEHIDHTFEIQYRKDTATWKDSKTETLQNAHSMALPALEPSTRYWARVRVRTS
RTGYNGIWSEWSEARSWDTES
.......................LHHHHL...........................H....................LL
LL.L................

319 1EO8.A
STATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNNPHRILDGIDCTLIDALLGDPHCDVFQNETWDLFV
ERSKAFSNCYPYDVPDYASLRSLVASSGTLEFITEGFTWTGVTQNGGSNACKRGPGSGFFSRLNWLTKSGSTYPVLNVTM
PNNDNFDKLYIWGIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPGDVLVINSN
GNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKITYGACPKYVKQNTLKLATGMRNVPEKQ
.......................................H........HHHHH..........LH.HLH..H......
.H.H.H.....................................................................
.....................................................................
......................H.H.....................................................

317 1QFU.A
STATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNNPHRILDGIDCTLIDALLGDPHCDVFQNETWDLFV
ERSKAFSNCYPYDVPDYASLRSLVASSGTLEFITEGFTWTGVTQNGGSNACKRGPGSGFFSRLNWLTKSGSTYPVLNVTM
PNNDNFDKLYIWGIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPGDVLVINSN
GNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKITYGACPKYVKQNTLKLATGMRNVPE
.......................................HHH........HHHHH..........LL..LH........
.H.H.H.....................................L.........................
.....................................................................
......................HHHH...................................................

159 1FSK.A
GVFNYETETTSVIPAARLFKAFILDGDNLFPKVAPQAISSVENIEGNGGPGTIKKISFPEGLPFKYVKDRVDEVDHTNFK
YNYSVIEGGPIGDTLEKISNEIKIVATPDGGSILKISNKYHTKGDHEVKAEQVKASKEMGETLLRAVESYLLAHSDAYN
.......................................CCCCCBBBBCCC...............C.C...B....
......C.........C..........................................................

452 1TY6.A
LNLDPVQLTFYAGPNGSQFGFSLDFHKDSHGRVAIVVGAPRTLGPSQEETGGVFLCPWRAEGGQCPSLLFDLRDETRNVG
SQTLQTFKARQGLGASVVSWSDVIVACAPWQHWNVLEKTEEAEKTPVGSCFLAQPESGRRAEYSPCRGNTLSRIYVENDF
SWDKRYCEAGFSSVVTQAGELVLGAPGGYYFLGLLAQAPVADIFSSYRPGILLWHVSSQSLSFDSSNPEYFDGYWGYSVA
VGEFDGDLNTTEYVVGAPTWSWTLGAVEILDSYYQRLHRLRAEQMASYFGHSVAVTDVNGDGRHDLLVGAPLYMESRADR
KLAEVGRVYLFLQPRGPHALGAPSLLLTGTQLYGRFGSAIAPLGDLDRDGYNDIAVAAPYGGPSGRGQVLVFLGQSEGLR
SRPSQVLDSPFPTGSAFGFSLRGAVDIDDNGYPDLIVGAYGANQVAVYRAQP
.....................................................................LLLL
LH.H.......................................................H....H...H..
```

```
.........................................HH.LLLL.HHH........................
.......................................................................
.......................................................................
.................................................
```

# Appendix II

**A larger version of figure 3.4**

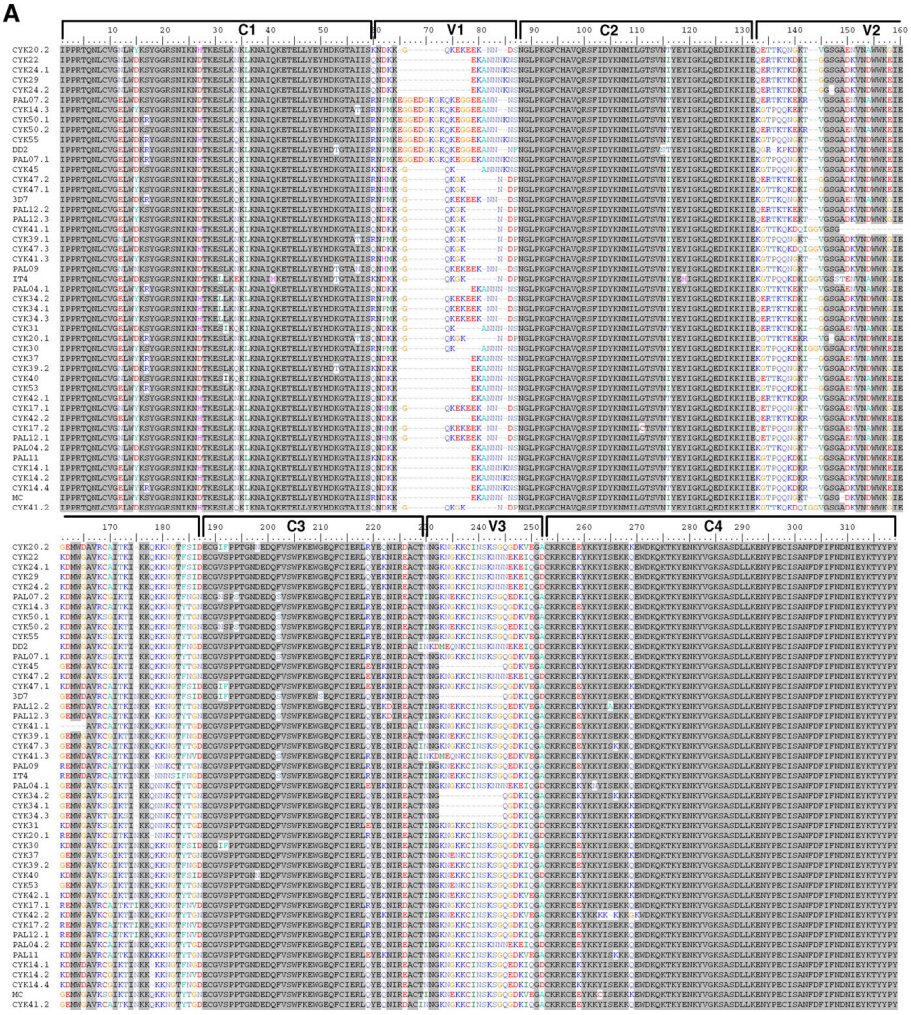


**Figure 3.4 Multiple alignment of VAR2CSA DBL3X sequences.** (A) cDNA from 43 placental parasite samples were amplified with conserved DBL3X primers and sequenced. Sequences were curated for primer sequence, translated and aligned. The text to the left indicate the identity of the samples. The DBL3X domain can be divided into four regions, which are highly conserved, C1-C4, and three regions, which are polymorphic and/or harbour deletions V1-V3.