UPTEC X 03 018 ISSN 1401-2138
JUN 2003

ÅSA BJÖRKLUND

Peptide filtration - a computational method for identification of unique protein motifs in allergens

Master's degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 03 018

Date of issue 2003-06-13

Author

Åsa Björklund

Title (English)

Peptide filtration – a computational method for identification of unique protein motifs in allergens

Title (Swedish)

Abstract

A bioinformatics method for the assessment of allergenicity when introducing new proteins with genetically modified organisms was developed. Prediction was based on similarity to allergen-specific protein motifs that were discovered with a developed algorithm called Peptide filtration. Classification performance was compared with current assessment methods recommended by the FAO/WHO and it was found to be notably more accurate. Attempts were made to identify some protein motifs with implications in allergenic responses, but the results were not conclusive.

Keywords

Allergy, Atopy, Allergen, GMO, Bioinformatics, Protein motifs

Supervisors

Ulf Hammerling and Daniel Soeria-Atmadja
Swedish National Food Administration

Scientific reviewer

Tomas Olofsson
Signals and System Group, Uppsala University

Project name	Sponsors
Language **English**	Security **Until 2004-03-01**
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **58**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Peptide filtration – a computational method for identification of unique protein motifs in allergens

Åsa Björklund

Sammanfattning

Allergi har blivit ett allt vanligare problem i vårt västerländska samhälle och intensiv forskning sker för att försöka klargöra varför. Allergier finns mot många olika substanser, så som pälsdjur, mat, mögel, kvalster och pollen. Det är vissa proteiner i dessa substanser som framkallar allergiska reaktioner, dessa proteiner kallas allergener. Många av dem har nu blivit kartlagda, och sekvenser för dess byggstenar, aminosyrorna, har blivit bestämda, men man vet ännu inte tillräckligt för att veta varför just de proteinerna ger allergier.

När nya födoämnen förs in på vår marknad, som t e x genmanipulerade grödor, genomförs grundliga tester för att säkerställa att inga nya allergener introduceras. Förutom en mängd laborativa tester så undersöks även aminosyrasekvensen med hjälp av bioinformatiska hjälpmedel för att se om de nya proteinerna är lika några kända allergener. Syftet med detta projekt var att försöka förbättra de rådande metoderna, och på ett bättre sätt kunna förutsäga om ett protein är en allergen. Korta bitar (proteinmotiv) i allergenernas sekvenser, som är särskilt viktiga för att framkalla allergier, har sökts genom att identifiera de proteinmotiv som finns hos allergener men inte hos icke-allergener. Om dessa proteinmotiv sedan finns i de proteiner man vill testa så klassas de som allergener.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet juni 2003

CONTENTS

1. INTRODUCTION

The prevalence of atopic allergy is increasing in today's Western Society and it is becoming a growing health-care concern. The reasons for this augment of allergic diseases is not clear, but factors such as allergen exposure, urban living, changed breast-feeding habits, smaller families, smoking, less childhood infections and higher standards of hygiene have been suggested (1,2). The proteins that causes allergy, the allergens, have intrigued the scientist for decades. Most allergens are glycoproteins and they come from various sources such as animal hair and dander, foods, pollen, insects, dust-mites and moulds. So far no one has been able to determine why these particular proteins induce allergic responses, whereas other similar proteins do not (3-5). It is clear however, that some families of proteins give allergic responses more often than others and proteins bearing high similarity with known allergens are more likely to be allergenic themselves (6,7).

When introducing new foods to our market there are several safety aspects to consider. One of them is whether the new food product will give rise to allergic reactions. Many new crops are being developed using genetic manipulation, and before these genetically modified organisms (GMOs) are allowed to our market they undergo safety assessment including tests for allergenic potential. In addition to immunological and chemical tests the sequences of introduced genes are compared with known allergens (8). These bioinformatics methods, based on alignments, are rather crude and give rise to many false positives. The aim of this project is to develop new and better methods for detecting allergens by looking at motifs in the proteins that are essential for their allergenic properties.

In the first part of this report the immune system and the mechanisms involved in allergy are introduced and a review of the current knowledge about allergens is given for a better perception of the problems involved in this project. The present safety assessment methods when introducing new GMOs are presented and some commonly used bioinformatics methods with relevance to this project are explained. After the theoretical background the central problem of this project and the ideas on how to solve them are discussed in Chapter 3. In the methods section in Chapter 4 the peptide filtration and classification of allergens used in this project are explained. Chapter 5 presents the results and finally these results are discussed in Chapter 6. A list of abbreviations frequently used in this report is provided in Appendix 1.

2. THEORETICAL BACKGROUND

2. 1. Immunology

The environment contains a great variety of infectious microbes that can cause disease and even kill its host. Evolution has provided us with the very sophisticated defence mechanisms, enabling us to deal with diverse types of microorganisms, that make our immune system.

2.1.1. Innate immunity

The innate immune system reacts rapidly and in a rather simple way to defend us from invaders. The first defences that a pathogen will encounter are the natural barriers such as skin, mucous membranes, stomach acid and tears. But if they manage to pass these obstacles there are some other non-specific mechanisms that deal with them, such as phagocytic white blood cells (neutophils, monocytes, macrophages, eosinophils and natural killer cells), antimicrobial proteins (complement proteins and interferons) and the inflammatory response with vasodilation, release of histamine, prostaglandin and other factors. These are all part of the so-called non-specific or innate immune system (9).

2.1.2. Adaptive immunity

The adaptive, or acquired immune system provides a highly specific defence response to foreign structures. Still it provides protection from a great variety of invaders. Characteristic for the adaptive response is the immunological memory and the ability to distinguish between self/non-self. The mechanisms of adaptive immunity involve several steps of recognition and complex reaction pathways where several different cell types are involved.

It is common to distinguish between humoral immunity and cell-mediated immunity. The prior refers to immune responses leading to the production and secretion of antibodies, and the latter to the direct action of lymphocytes (9). The main players in the immune response are the white blood cells, the leukocytes, circulating in the blood. They consist of lymphocytes (T and B cells), monocytes, granulocytes and others. Reviewed in Table 1 (10).

2.1.2.1. Lymphocytes

The two most abundant types of lymphocytes in the human body, both originating from stem cells in the bone marrow, are called T-cells and B-cells. Early T-cells migrate to the thymus where they mature while the B-cells remain in the bone marrow for maturation. Each B-cell and T-cell is specific for a particular antigen (an antigen is any substance which causes an immune response). They both have membrane-bound receptors for antigen recognition, the B-cell receptor (BCR) and the T-cell receptor (TCR), each present in thousands of identical copies on the cell surface. Both these receptors are encoded by genes assembled by the recombination of segments of DNA, and therefore have variable regions with great diversity in antigen binding. It has been estimated that we can synthesise over $2.5x10^7$ different TCR's and about the same number of BCR's (10).

The Cells of the Immune System

Lymphocytes:	**T-cells:** Involved in cell-mediated immune response. Differentiate into cytotoxic T-cells and destroy infected cells or helper T-cells and regulate the immune response by secreting cytokines.
	B-cells: Responsible for the humoral (anti-body secreting) immune response. Can differentiate into plasma cells or memory B-cells
	Plasma cells: Anti-body secreting cells.
	Natural killer cells: Destroy the body's own infected cells, especially those harbouring viruses. Attacks membrane causing lysis.
Monocytes: 5% of white blood cells. Circulate, but migrate to tissues and become macrophages.	
Macrophages: Large amoeboid cells that use pseudopodia to phagocytize bacteria, viruses and cell debris.	
Granulocytes:	**Neutrophils:** Becomes phagocytic in infected tissue. 60-70% of white blood cells. Attracted by chemical signals.
	Eosinophils: 1.5% of white cells, Have a limited phagocytic activity. Contain destructive enzymes in cytoplasmic granules. Contributes in defence against larger invaders such as parasitic worms.
	Basophils: Granulocyte with basophilic granules that contain histamine bound to a protein and heparin like mucopolysaccharide matrix. Similar to mast cells.
Megakaryocytes: Very large bone marrow cells which release mature blood platelets.	
Mast cells: Resident in connective tissue. Contains many granules rich in histamine and heparan sulphate.	
Dendritic cells: Found in tissues where they capture and process antigens, and presents them to T-cells.	
Antigen Presenting Cells: Cells that capture and process antigen and then presents them in complex with MHC II to T-cells. Include Langerhans cells, Dendritic cells, Interdigitating cells, B-cells and macrophages.	

Table 1. The cells of the immune system

2.1.2.2. Activation of T-cells

The lymphocytes are activated upon binding of an antigen to their receptor. The nascent T-cell, often called the T_0-cell can differentiate into either cytotoxic (T_C) cells, helper (T_H) cells or regulatory T-cells (also called suppressor T-cells). The T_C -cell can destroy infected cells by releasing substances that will lyse the plasma membrane. The T_H -cells work by secreting cytokines, thereby influencing the actions of many other cells of the immune system. T_C and T_H cells both act against pathogens that have entered into cells and have been fragmented and presented on the cell surface in complex with the major histocompatibility complex (MHC). They are both activated upon binding of a peptide/MHC complex that the TCR is specific for, as illustrated for T_H cells in figure 1.

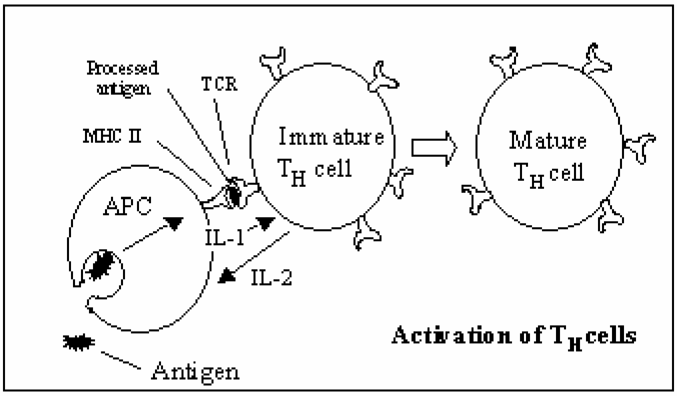


Figure 1. The activation of T-helper cells. The immature T_H cell encounters an antigen presenting cell that has endocytosed and processed an antigen and is presenting an antigen fragment in complex with MHC Class II. When this complex is recognised by the antigen specific TCR on the T-cell surface it triggers the release of stimulating cytokines IL-1 and IL-2. This promotes the maturation of the T_H cell to from T_H1 or T_H2 cells.

2.1.2.3. The major histocompatibility complex

The major histocompatibility complex (MHC) is a group of glycoproteins embedded in the plasma membrane of all nucleated cells in the body. They are important self-markers and are coded for by a set of gene loci with at least 20 genes and more than 100 alleles for each gene. MHC is a member of the immunoglobulin supergene family and is the most polymophic protein so far identified. The probability that two individuals, that are not identical twins, will have matching MHC sets is virtually zero (11).

The MHC system is called H2 in mice and HLA (human lymphocyte antigen) in humans. There are two main classes of MHC in the body: Class I MHC molecules are located on all nucleated cells of the body, it is encoded by three genes called HLA-A, B, and C. Class II MHC molecules are found only on specialised cells such as macrophages, dendritic cells, B-cells and activated T-cells. The class II genes are called HLA-DR, DQ, and DP.

MHC Class I plays an important role in the recognition of self/non-self. During maturation in thymus and bone marrow, T- and B-cells with receptors that bind and react to self-proteins in complex with MHC will be eliminated, resulting in a immune system that will not react to endogenous proteins. But when introducing foreign tissue into the body, as in grafts and transplants, the MHC molecule of another individual will function as a foreign antigen, and the new tissue will be destroyed by the immune system.

The MHC molecules bind to short (8-20 amino acids) in specialised compartments in the cells and are transported to the cell surface where the peptide-MHC complex can be recognised by the TCR on the surface of T-cells. Class I molecules bind to peptides from proteins that have been synthesised and degraded inside the cell (endogenous antigens) and present them to cytotoxic T-cells, thereby signalling that the T_C cell can destroy the cell. This is useful, for example when a cell has been infected with a virus or when mutated cancerous cells produce proteins that normally are not present. Class II molecules present foreign peptides, from proteins or microbes that have been endocytosed from the cell surface. They are recognised by T helper lymphocytes leading to their activation. The activated T_H cells will stimulate B-cells to produce antibody against the foreign substance and will recruit other actors of the immune system to the site (9, 10, 12).

2.1.2.4. Helper T-cells

When nacent T-cells acquire the helper cell marker, CD4, they are called pre-T-helper cells, or T_H0 cells. T_H0 cells can differentiate into two types of helper T-cells designated T_H1 and T_H2. The fate of a T_H0 cell is determined by many factors, such as the cytokine environment, dose of antigen and the affinity of the TCR for the antigen. Typical cytokines secreted by T_H1 cells are interleukin 12 (IL-12), tumour-necrosis factor-beta (TNF-β) and interferon-gamma (IFN-γ). These molecules stimulate macrophages to kill bacteria, and recruit other leukocytes to the site producing inflammation. IFN-γ and IL-12 suppresses the T_H2-pathway and promotes the T_H1-pathway. IFN-γ also inhibits the IgE production by B-cells (12).

The cytokine profile of T_H2 cells include IL-4, IL-5, IL-10 and IL-13. IL-4 and IL-13 stimulate B-cell class-switching promoting the synthesis of IgE antibodies. IL-5 attracts eosinophils and IL-10 inhibits the IL-12 production by dendritic cells, thereby inhibiting the formation of T_H1. IL-4 acts as positive feedback loop promoting more T_H0 cells to enter the T_H2 pathway and in the same time blocking the expression of IL-12 receptor. The many factors influencing the polarity of the T_H cell is illustrated in Figure 2 (12).

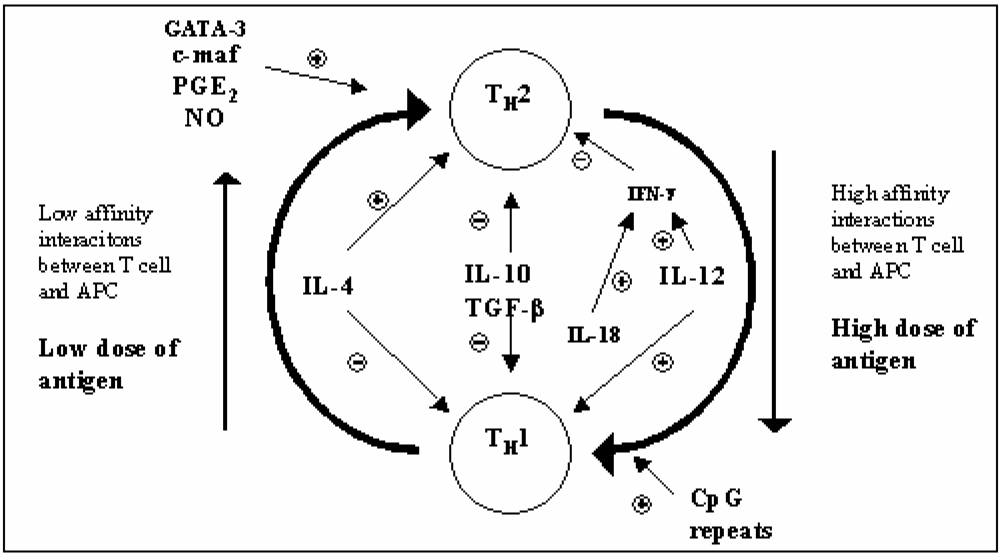


Figure 2. Factors regulating the T_H cell phenotype. Polarisation to T_H1 or T_H2 depends on IL-4 and IL-12, respectively. Other factors include interactions with APC and dose of antigen. CpG nucleotide repeats derived from bacteria favours T_H1 while factors such as GATA-3, c-maf and PGE_2 induces T_H2. NO is less inhibitory for T_H2 than T_H1 thereby promoting T_H2. IL-10 and TGF-β dampens both kinds of responses. IL-12 and IL-18 promotes the release of IFN-γ from T-cells and IFN-γ inhibits T_H2 (1).

2.1.2.5. Activation of B-cells

Upon activation B-cells will differentiate into antibody-secreting plasma cells, playing a major part in the humoral immune response. Activated B-cells also differentiate into memory B-cells that will help the immune system to react faster when exposed to the same antigen a second time. The activation of a B-cell is first triggered by the binding of an antigen to the BCR. The antigen is then internalised into proteolytic vesicles, cleaved and presented at the cell surface in complex with MHC class II. For full activation the B-cell requires the right cytokine environment provided by a T_H-cell bound to the MHC-II/antigen-peptide complex. See Figure 3. Although bystander activation of B-cell, where there is no contact between the T-cell and the B-cell, occurs it is not very common in the immune response (13).

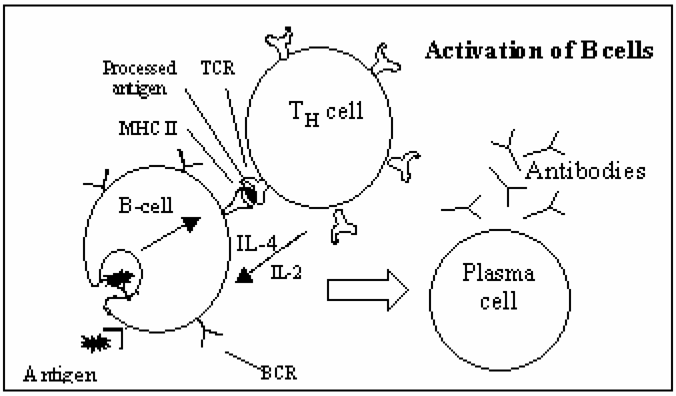


Figure 3 Activation of B-cells. The antigen specific BCR recognise, bind and internalise an antigen. The antigen is processed and presented in complex with MHC Class II. When the antigen is recognised by a mature T_H cell cytokines released by the T-cell will promote proliferation of the B-cell into a antibody secreting Plasma cell. Memory cells are also formed from the B-cell.

2.1.2.6. The immunoglobulin antibodies.

Antibodies are of a specific class of glycoproteins called immunoglobulins (Igs). Some are carried on the surfaces of B-cells and act as B-cell receptors or attached to other cell types. Others circulate freely in the blood or lymph. They are synthesised by B-cells and plasma cells. Antibodies are Y-shaped molecules with four polypeptide chains. All four chains consist of a constant region and a variable region. The variable regions at the two tips of the Y form the antigen-binding sites.

Antibodies do not destroy pathogens directly, but by binding to the antigen they tag the invader for destruction by one of several mechanisms: *Neutralization* is when the antibody blocks viral attachment sites or coats bacterial toxins making them ineffective. Each antibody has two or more antigen-binding sites and can cross-link adjacent antigen resulting in clumps of bacteria being held together. This is called *agglutination*. Similar to agglutination is *precipitation*, where cross-linked soluble antigen molecules are immobilised. Antibodies can also activate complement proteins so that they lyse foreign cell membranes. *Opsonisation* of microorganisms by antibodies make them more attractive to the phagocytotic white blood cells.

There are five types of constant regions, each characterising one of the five major classes of mammalian immunoglobulins. The type of Ig produced is determined by the cytokine environment surrounding the B-cell. The five types of Igs are:

- IgM. Consists of five monomers arranged in a pentamer structure. They are circulating antibodies that occur in the first response to an antigen. IgM act together with complement proteins to lyse cells.
- IgG acts as a monomer. It is the most abundant circulating antibody accounting for 70-75%. IgG can act on pathogens by agglutinating them, by opsonising them, by activating the complement system and by neutralising toxins.
- IgA is a dimer and is produced primarily by cells abundant in mucous membranes. They prevent the attachment of bacteria and viruses to the epithelial surfaces.
- IgD is a monomer found primarily on outer membranes of B-cell where they may play a role in antigen recognition.
- IgE is a monomer. Its stem regions attach to receptors on mast cells and basophils and can thereby stimulate the release of histamine and other chemicals associated with allergy.

Allergy is often referred to as “the IgE-mediated disease” since IgE plays such a central role in the disease, so the rest of this work is mainly focused on IgE antibodies (9).

2.2. Atopy and allergy

The term ‘allergy’ was introduced in 1906 by Von Pirquet (14) when he observed a ‘changed reactivity’ to an antigen. The term is now often used synonymously with IgE-mediated allergic disease, but this is not the meaning Von Pirquet initially intended. Another commonly used term for describing IgE-mediated disease is ‘atopy’ from the Greek *atopos*, meaning ‘out of place’(1). This work is focused on IgE-mediated atopic disease, also called immediate (Type I) hypersensitivity reactions, not to be confused with other sensitivity reactions such as gluten or lactose intolerance.

All individuals have the ability to produce IgE as a defence against large quantities of allergens, as in the case of helminth (parasitic worm) infections. But not everybody produces IgE against common allergens such as house dust mite. Individuals with an immune system inclined towards IgE-production are said to be ‘atopic’. It is evident that genetic predisposition is implicated in atopy, but the exact genes have not yet been identified. It has been proposed that several genes are involved (2). Since there has been a raise in atopy over the last decades it is clear that some environmental factors also influence the development of atopic diseases. Some suggested factors are: allergen exposure, maternal smoking, Western life-style, pollution, smaller families, changes in breast-feeding habits, possible lack of infections and higher standards of hygiene (1, 2).

2.2.1. The allergic diseases

Diseases associated with increased levels of IgE are allergic rhinitis, asthma, anaphylaxis, atopic eczema, urticaria and angioedema (reviewed in 1). **Allergic rhinitis** is a recurrent or persistent inflammation of the nostrils with symptoms such as nasal congestion, rhinorrhoes, sneezing and itching. The most common type of allergic rhinitis is often called hay fever. **Asthma** is a chronic inflammatory disease in the airways of the lung characterized by airway obstruction and airway hyper-responsiveness accompanied by wheeze, breathlessness or cough. Most asthmatics are atopic, but this is not always the case.

Anaphylaxis is a severe systematic allergic reaction induced by massive release of histamine leading to shortness of breath, rash, wheezing and a quick drop in blood pressure. The symptoms can sometimes be life-threatening or lead to permanent brain damage. The common causes of anaphylaxis are hypersensitivity to foods, bee and wasp stings, certain drugs and latex. **Atopic eczema** or **dermatitis** is most prominent in early childhood, and it affects 10-20% of children in Western populations. It is characterised by a red itchy rash, normally due to IgE antibodies against aero-allergens or food allergens. **Urticaria** (widespread itchy weals or hives) and **angioedema** (deep mucocutaneous swelling) normally occur together. They are often associated with sensitivity to foods, drugs or latex (1).

2.2.2. The allergic reaction

Mast cells and basophils both can attach IgE antibodies to the FcεRI receptor on their surface. In immediate allergic responses two or more such IgE/ FcεRI complexes bind to the same antigen, thereby causing the cross-linking of the FcεRI receptor. The cross-linking initiates a cascade of reactions eventually leading to degranulation and release of granule associated mediators. The granules of mast cells and basophils are particularly rich in histamine, but also contain serotonin, lipid mediators, proteases, chemokines and cytokines. The release of these substances produce a rapid increase in blood flow, enhanced vascular permeability, increased loss of intravascular fluid, itching, wheezing and sneezing. In severe cases it can lead to anaphylactic reactions (15).

The released cytokines also stimulates the production of more IgE and the recruitment of eosinophils to the tissue. Late phase allergic reactions (LPR) are associated with the primary accumulation of eosinophils and neutrophils, and later recruitment of T_H cells and basophils. LPR are developed approximately 8-24 hours after the immediate reaction (16). They cause further wheezing, oedemas and congestion of the nose. Antigen presenting cells, especially dendritic cells play an important role in the induction of LPR. They present antigen fragments together with MHC class II to the T_H cells and the activated T_H cells release cytokines that will attract eosinophils and neutrophils. The LPR can occur IgE-independently; the action of T-cells alone seems to be sufficient (1,11).

For these allergic reactions to occur the immune system must have been exposed to the allergen previously. At first encounter with the allergen the so-called sensitation reaction takes place. When a T_H2 cell is activated by the allergen it stimulates the proliferation of more allergen-specific T-cell clones and the production of allergen-specific IgE antibodies.

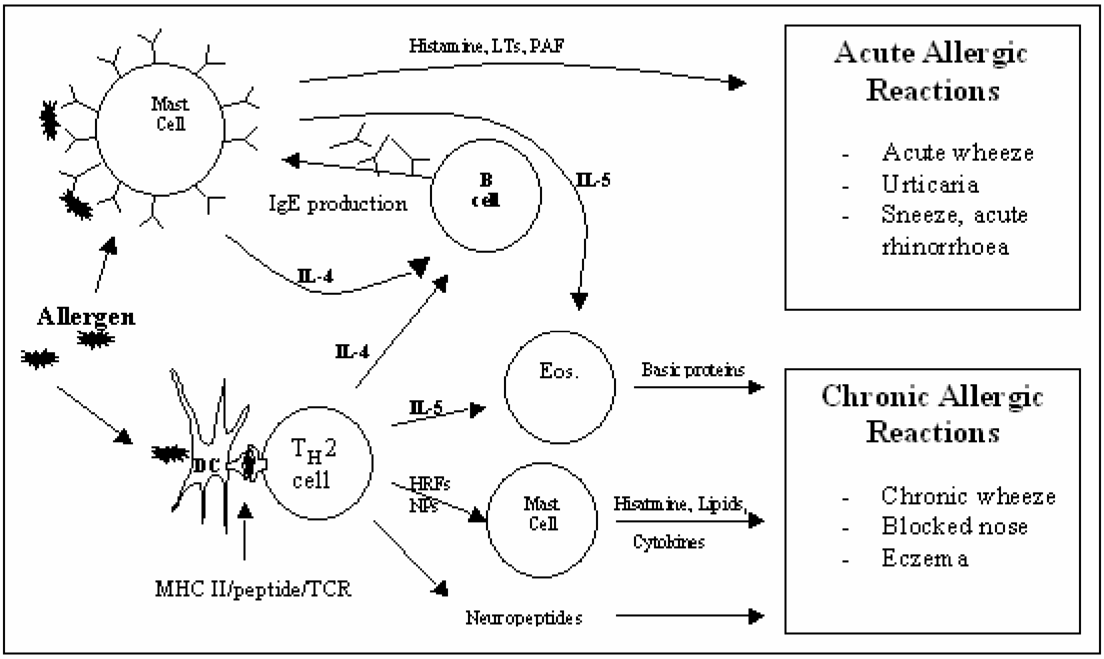


Figure 4. Pathways leading to acute and chronic allergic reactions. Acute reactions are due to histamine and lipid mediators released by mast cells. Chronic reactions may depend on a combination of factors including eosinophil recruitment, release of mast cell products and neorogenic inflammation (1).

However, acquisition of sensitisation to and subsequent allergic disease is known to be influenced by a variety of environmental factors and the timing, duration and extent of exposure. Moreover, the nature of the allergen itself may have an important impact on the allergic response (17).

2.2.3. T_H2 polarity in allergy

All atopic individuals have their T_H cell response shifted to a T_H2 profile in affected tissues. The cytokine environment, the TCR-MHC II-peptide interaction, genetic predisposition and many other factors seem to play a role in the induction of a T_H2 response. In newborns the T_H2 type is dominating and during the first months of life it reverses to T_H1 in non-atopic children, probably as a consequence of stimulation by infectious agents. It has been suggested that decreased postnatal exposure to microbes leads to a T_H2-skewed immune system, and also that increased postnatal allergen exposure can promote a T_H2 response (18). Several studies have implicated that microbial gut flora has influence on the development of atopy (19). But also prenatal exposure to allergens through the cord blood and amniotic fluid has been suggested to affect the development of a T_H2-response (18).

2.2.4. Treatment of allergy

Since the prevalence of allergic disease is increasing in Western society, the large health-care cost is becoming a burden. Much work is put into finding methods to treat and prevent the allergic reactions. The most obvious treatment of allergy is avoidance of the allergen. But since this is not always possible or convenient, other methods are developed. It is common to use anti-allergic treatment, to supress the allergic symthoms. Some drugs employed for this purpose are antihistamines, anticholinergic agents and corticosteroids (1).

Specific immunotherapy (SIT) has been used to treat allergies for nearly 100 years. It involves the administration of increasing concentrations of allergenic extract to the patient over long periods of time, thereby desensitising the patient to the allergens. The results are often good, and they last for years after terminating the treatment. But there are some risks associated with this treatment; the risk of developing severe, and potentially fatal, anaphylaxis. The mechanism by which SIT works is still unclear, but there is evidence that it induces a shift from T_H2 to T_H1 cytokine profile, and there is an increase of anti-allergen IgG antibodies (20).

T-cell peptide epitope immunotherapy has been shown to give good results in clinical trials. It involves the administration of short allergen-derived peptides, which can bind to MHC class II and induce a T-cell response, but still are unable to cross-link IgE and induce anaphylaxis. This method resembles specific immunotherapy, and the results can sometimes be as good, but without the risks associated with SIT. More research to identify allergenic T-cell epitopes will be of great use for this type of treatment (21).

Other methods under development are: DNA vaccines include the use of immunostimulatory CpG nucleotide motifs, which induce a strong T_H1 response (22). Virus-like particles can induce IFN-γ producing T_C lymphocytes rather than having a T_H2 response (23). Other treatments focus on blocking IgE or IgE synthesis. Humanized anti-IgE monoclonal antibodies have been shown to virtually eliminate all circulating IgE in allergic patients (24). IL-4 is an important inducer of IgE production. Several ways of inhibiting IL-4 are being investigated (1).

2.3. Allergens

An allergen is a protein capable of triggering immediate (Type 1) hypersensitivity reaction, i.e. what we commonly call allergy, in susceptible individuals. It is clear that some proteins are intrinsically more allergenic than others and many of them have been characterised. But what is it then that distinguishes them from other proteins? It is unlikely that the overall structure of the allergen is responsible for allergenicity (3). The list of allergens include a structurally and functionally heterogeneous group (4). But one thing that is clear is that for a protein to be allergenic it must have T-cell epitopes capable of inducing a type 2 T-cell response and it must have at least two IgE binding epitopes to cross-link the FCεR on mast cells and basophils, and most allergens have more than two. But it is not clear if the presence of appropriate epitopes alone is sufficient (5). Many features influencing allergenicity have been suggested, such as resistance to proteolysis, glycosylation status, size, heat stability, solubility, enzymatic activity and dose of allergen (3,5).

2.3.1. Protein stability

One suggested common feature to allergens is resistance to digestion and heat stability. For a food allergen to be able to sensitise the immune system it must resist degradation in the stomach. It has been found that most food allergens are stable in Simulated Gastric Fluid (SGF) (25). This is not true for many inhaled allergens, such as pollen and mite allergens. Many food allergens associated with oral allergy syndrome are not stable (26). There are also many non-allergenic proteins that are just as stable, but still do not induce an allergic response, so using just protein stability as a marker for allergenicity would not be enough. Since it seems like resistance to proteolytic cleavage is more common among allergens than other proteins, it is possible that the stability does not only reflect stability in the stomach, but also resistance to processing in the vesicles of antigen presenting cells (5).

Food allergens are normally resistant to heating and other food processing effects (27). Heating of a protein may induce conformational changes leading to the disappearance of some epitopes, but at the same time new epitopes may be created. Heating is also associated with various other reactions such as attachment of reducing sugars, oxidation, scrambling of disulphide bonds and deamination (28). Many fruit and vegetable allergens can be eliminated by heating, giving hypoallergenic products such as jams and juices (29). Cooking of egg eliminates the allergen response to egg white in many patients (30), but it is not possible to reduce the offensive properties of all allergens with heat treatment (31). There are even some cases where heating creates new allergens; an example of this are cooked pecan nuts (32). And some patients allergic to cooked cod and shrimp are not allergic to the raw meat (33, 28). In the case of peanuts it has been shown that roasting increases the allergenicity of the allergens Ara h 1 and Ara h 2 remarkably (34). Similar reactions to those that take place during heat treatment can also occur at a slow rate during storage of food, neoantigens appeared in wheat flour after storage for 7 month at ambient temperature (28,35).

2.3.2. IgE binding epitopes and cross-reactivity

An IgE epitope is the protein structure that the IgE antibody can recognise and bind to. They can be either linear or conformational. A conformational epitope is created when the three-dimensional structure of the protein brings together amino acids, not adjacent in the protein sequence, on the surface to form a site where the IgE antibody can bind. These epitopes can either be formed or broken due to denaturing of the protein. A linear epitope has sequential amino acids on the surface of the protein. These epitopes are easier to predict, and less vulnerable to changes in the three-dimensional protein structure. Many IgE binding epitopes of allergens, both linear and conformational, have been documented. But so far no one has been able to find any common feature among them that would distinguish them from non-allergenic epitopes (4). The shortest reported epitope required to bind IgE has 5 amino acid residues (36,37). However, some small linear epitopes may in fact be fragments of larger conformational epitopes (4).

When two proteins have the same or similar IgE epitopes it is possible that they are cross-reactive, meaning that they both give the same allergic response due to binding to the same IgE antibodies on mast cells or B-cells (38). Normally IgE cross-reactivity occurs between homologous proteins, since high homology often reflects high similarity in 3D structure. For example, serum albumins from vertebrates are often cross-reactive (39), and many related grasses are cross-reactive (38). However there are many examples of cross-reactivity between more distantly related organisms, such as ragweed/banana, birch/apple, latex/banana/avocado and mugworth/celery (reviewed in 38). In all reported cases the cross-reactive allergens have high sequence identity. So far there are no well-characterised example of cross-reactivity between proteins with different folds but with identical shorter amino acid stretches (38).

2.3.3. T-cell epitopes

For an allergen to be a true allergen it does not only require the property to elicit an IgE-mediated allergenic reaction it must also be able to *de novo* sensitise susceptible individuals (4). This requires T-cell epitopes (TCEs) capable of inducing type 2 T-cell responses. With epitope-mapping all allergens studied to date have been found to contain multiple TCE that are present throughout the molecule. But there is no difference between the epitopes that non-allergic patients recognise and the epitopes recognised by allergic patients, and immunotherapy does not induce an epitope shift (16). This may indicate that the epitope specificity does not have a direct influence on the T_H2 type response.

If a TCE is located in a conserved region the allergen specific T-cells may cross-react with homologous allergens from different species (16). This is seen for grass pollens where allergen specific T-cells are very diverse, they recognise multiple proteins in allergenic extracts, react with a vide variety of TCE and cross-react between many grass species (41).

T cell epitopes presented by MHC class II are of variable length ranging from 9 to 24 amino acids (aa). The actual binding groove of MHC class II is capable of accommodating 15 aa, but allows for additional peptide overhang outside of the groove. There is a large hydrophobic pocket at one end of the groove suggesting that an anchor residue, preferable an aromatic amino acid, binds there. There are differences in binding patterns for different HLA alleles. Restriction specificity appears to be at position 1, the anchor residue and at position 4, 6 and 9 (11).

Many studies have been done to determine which specific peptides different HLA alleles bind to using epitope elution or mapping with synthetic peptides (reviewed in 42). These data have been used to build computer algorithms for predicting T cell epitopes. Most of them are matrix-based prediction algorithms such as ProPred (43), DRGen (44), SYFPEITHI (45) and PAP (46), where matrices with probabilities for each amino acid are employed to search for the peptides. Other more complex algorithms such as neural networks in combination with an evolutionary algorithm has been applied to this problem by Honeyman and Brusic (47). Mallios has developed an iterative system that uses binding matrices in combination with suggested motifs (48, 49).

The data on MHC class II -binding motifs do not cover all the hundreds of different HLA alleles, even though the most common ones are mapped. This limits the TCE prediction methods to possible TCE's that are bound to the studied alleles. The fact that a peptide binds efficiently to MHC class II does not directly implicate that it is a T-cell epitope and the information on peptides that are recognised by the TCR is even more limited (42).

Since the size of the peptides bound to MHC class II and the site where they have been cut depend on features of the antigen processing machinery, the binding properties of naturally cleaved peptides may differ from those of synthetic peptides. Several programs for prediction of cleavage by some proteases are currently available on the Internet, such as FRAGPREDICT (50), NETCHOP (51) and PAPROC (52) But since the milieu in the antigen processing vesicles is quite complex and differing from individual to individual accurate prediction for the processing of antigens is difficult (42).

The T-cell epitopes of some allergens have been mapped. But so far no one has been able to identify any special feature that would distinguish them from other TCE. But there is still too little data to rule out the possibility that there might be some common feature among allergen TCE.

2.3.4. Glycosylation

Most allergens are glycoproteins (53), but a functional connection between protein glycosylation and the induction of allergenic response has not yet been demonstrated. It is known that glycosylation influences stability, hydrophobicity, solubility, electric charge and sometimes uptake of a protein into cells and organelles (5). Glycosylation can alter the structure of IgE epitopes (54), and carbohydrate epitopes of plants have been found responsible for cross-reactivity (55,56). But it has not been shown whether glycosylation affect the ability of proteins to sensitise the immune system (5). There might be a bias towards Th2 response for glycosylated antigens, since the type 2 specific interleukin-10 increases the expression and activity of mannose receptor on dendritic cells leading to increased uptake of glycans (57).

2.3.5. Enzymatic activity

Many studies support the idea that enzymatic activity contributes to the allergenicity of some allergens. One example of this is the Der p 1 allergen of house dust mite that can cleave CD23 (the low affinity IgE-receptor) on B-cells and CD25 (the α-subunit of the IL-2 receptor) on T-cells. Der p 1 significantly enhances IgE responses in mice, as compared to a enzymatically inactive mutant allergen (58,59). Mite proteolytical allergens have been shown to increase the permeability in the bronchial epithelium leading to enhanced uptake of the allergens (60).

Many allergens are not enzymes, especially most mammalian allergens, so as a rule, enzymatic activity is not a good determinant for allergenicity (58). But enzymes have some features that make them more probable to be allergens. Enzymes are often stable in hostile environments. They bind substrates in hydrophobic pockets that might have high antigenic potential. Enzymes often have flexible parts, which might facilitate binding of IgE and the B-cell receptor (5).

2.3.6. Allergen families

Even though there are no obvious common features among allergens, there are some discrete protein families where allergens are more frequent (61). Among mammalian allergens some common families are lysozymes, lipocalins and serum albumins (39). Napins, non-specific lipid transfer proteins, lipocalins, profilins, chitinases, cupins and Bet v 1-related proteins are some common allergens in plants (6, 7). Nevertheless, not all proteins belonging to these families are allergenic despite high homology. Serum Albumins, for example, are the most common source of allergic cross-reactivity between mammals, still there are no reports on cross-reactivity with avian serum albumins despite a homology of 43% (6).

2.4. Bioinformatics and computer analysis

With the help of engineering technology there have been great and fast advances in many fields of medical and biological research over the last decades. This has lead to the production of massive quantities of data, for example nucleic acid sequences and gene expression patterns. Therefore it has become necessary to integrate computer science with biological knowledge to develop tools to organise and analyse these data. The term bioinformatics was first introduced in the late 1980s and refers to the development of computational methods and the application of those methods to solve biological problems (62, 63).

Bioinformatics has many applications in diverse fields of biological research. Some of them are genomic sequencing, genome annotation, comparison of multiple genomes, analysis of gene expression data, analysis of protein sequences, protein structure, protein abundance and protein interactions, simulation of molecular pathways and gene regulation and studies of evolution and phylogeny (63).

2.4.1. Sequence alignment

Very important in bioinformatics are the tools for analysis of protein and DNA sequences. Sequence alignment is used to compare two or more sequences and determine their degree of similarity. When two symbolic sequence representations of DNA or protein are arranged next to each other so that their most similar elements are juxtaposed they are said to be aligned. Every element in the trace of an alignment is either a gap or a match.

```
-IRASAGFDL--AGVHYYVTA
 || | ||||  |||| |||
HIRSS-GFDLLVAGVHTYVT-
```

In the above example of aligned protein sequences there are some gaps, marked with -, and several matches. The matches can either be aligning one amino acid with the same one or with a different amino acid. Different such matches will give different **scores** to the alignment. **Substitution matrices** are employed to determine what that score should be. They contain the substitution scores for all possible combination of residues. These scores are obtained by looking at how common different substitutions are through evolution. Common such matrices are from the BLOSUM series and the PAM series. An identity matrix can be used as the substitution matrix when only match with the same amino acid is allowed. The alignment score is then calculated by adding the substitution scores for all matches. When introducing a gap in the alignment a penalty score is subtracted, this is called a **gap penalty**. There can be different gap penalty depending on how long the gap is. The **optimal alignment** is the one that maximises the alignment score.

There are different algorithms for finding the optimal alignment. The Needleman-Wunch algorithm is commonly used for finding global alignments, i.e. alignment of entire sequences with as many matches as possible (64). Local alignments are used to find stretches of highly conserved motifs. The most used method for doing local alignments is the Smith-Waterman algorithm (65). Two fast methods for searching sequence databases have been devised - these are FASTA (fast alignments) (66) and BLAST (Basic local alignment search algorithm) (67). These are both available for use as web-based tools. When using these programs, success on finding distantly related sequences depends upon an appropriate scoring matrix and gap penalty settings provided by the user.

Multiple sequence alignments are used for finding similar domains in a set of sequences and for doing phylogenetic analysis. It is an extension of two sequence alignments to align several sequences, aligning the two most similar ones first and then adding the next most similar one with hierarchical extension. An often used web-based tool for multiple alignments is CLUSTAL W (68).

2.4.2. Classification and learning systems

Learning systems are adaptive methods that can adjust to and find relations in large data sets. The goal is to extract useful information from a body of data by building good probabilistic models. Learning systems automatically improve their performance through experience. They are commonly applied to classification problems.
For biological data several methods for classification can be employed. The simplest and most straightforward method is the **linear classifier** where a decision boundary, a straight line for two dimensions or a hyper plane for several dimensions, will separate the classes. The boundary tries to minimise the interclass overlap, but it is difficult to get perfect separation. In this project only the linear classifier is used, but there are several other methods such as k-nearest neighbour, Bayesian classification, multi-layer perceptrones, hidden Markov models, etc. but they will not be mentioned further in this work (69).

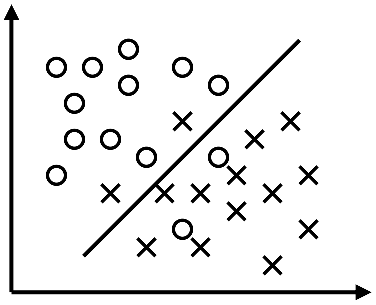


Figure 5. Example of a linear classifier in two dimensions.

2.4.3. Validation and ROC curves

When a classification method has been developed it is very important to validate its performance in an appropriate manner. The optimal method will have as good classification as possible and at the same time it will be as general as possible. It is common to do validation of the model by testing it with a set of data that has not been used when building the model, and therefore is totally independent of the classification model. Such testing is done to fine-tune parameters of the model or to decide which model is the optimal one for the problem at hand.

Unfortunately, in many cases, there are not always enough data to make both a training set and a validation set that will be sufficiently large to get the desired statistical evaluation. In these cases **cross-validation** techniques can be helpful. In k-fold cross-validation the dataset is divided into k subsets of approximately the same size. Then the model is trained k times while withholding one of the k sets each time and evaluating the performance each time with the withheld set. In the end the average performance is calculated for all k runs.

For two-class problems it is common to use **Receiver Operating Characteristics** (ROC) curves to demonstrate the accuracy of a method (70). When considering the results of a two-class classification, in the case of this project, the classification of non-allergens and allergens, there will be four possible outcomes. One case is correct classification of an allergen as an allergen (a true positive, TP), and another is incorrect classification of a non-allergen as an allergen (a false positive, FP), se Figure 6. When shifting the decision boundary for the classification to increase the number of true positives it will be at the cost of an increased number of false positives.

The desired performance can vary with the use of the model and the type of classification. In some cases it might be desirable to have a very high sensitivity, i.e. not missing any true positives, even if it is at the cost of having more false positives. In other cases it might be the opposite, that the number of false positives must be minimised. These cost-benefit characteristics can be plotted in an ROC curve where the probability of detection (pDetection), the number of TP in relation to the total number of allergens, are plotted against the probability of False Alarm (pFA), the number of FP through the total number of non-allergens.

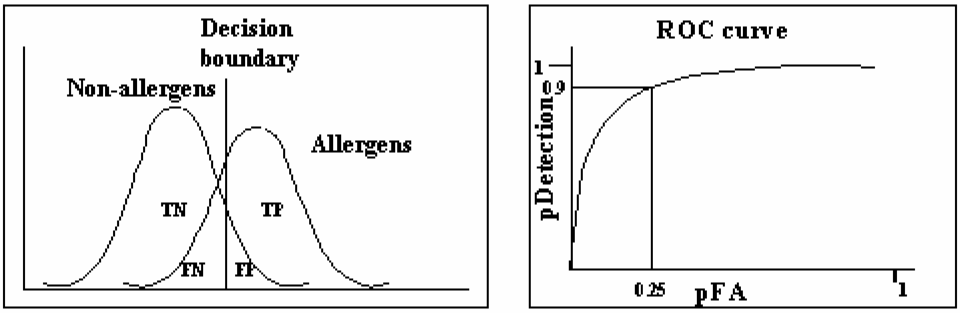


Figure 6. a) Classification of allergens and non-allergens. Where the decision boundary is drawn will determine how many allergens (true positives) and non-allergens (false positives) that will be classified as allergens. **b)** ROC curve. The probability of detection (TP) plotted as a function of the probability of false alarm (FP) for different positions of the decision boundary. Example, with a detection of 0.9 the pFA will be 0.25 as illustrated with lines in the plot.

2.4.4. Dimensionality reduction and visualisation

Some structures that can be seen with the human eye are not necessarily captured by computerised methods, but in the case of high dimensional data it is not possible to look at the data distribution. Therefore it is often practical to visualise data in two or three dimensions to find groups or structures and correlations in the data. Some techniques for dimensionality reduction are PCA, MDS and ISOMAP (see below). Clustering techniques makes it possible to test hypotheses regarding the number of distinct groups in the data and their distribution.

2.4.4.1 PCA - Principal Component Analysis

PCA is the dimensionality reduction technique most widely used. It is a linear mapping of multidimensional vectors to low dimensional vectors through projection onto the principal components of the data, i.e. the components with highest variance (the first eigenvectors of the covariance matrix). This way the dimension of the data is reduced in a manner that will preserve its variation well.

2.4.4.2 MDS - Multi-dimensional scaling

MDS finds a representation of data that will preserve the inter-point distances. It provides a visual representation of the pattern of proximities so that those data points that are close in the multidimensional space appears close to each other in the MDS plot.

2.4.4.3 ISOMAP – Isometric feature mapping

If the data sets contain non-linear structures they might be invisible with linear visualisation techniques such as PCA. For these datasets Isomap can be a more helpful tool. It builds on classical MDS, but tries to preserve the intrinsic geometry of the data, as described by the distances in a multidimensional space between all pairs of data points. For faraway points, their distance is approximated by adding up a sequence of short “hops” between neighbouring points (71).

2.4.4.4. Discriminant functions

Discriminanat functions can be used to project data in a manner that will scatter the data set to maximize class separability. One common such discriminant function is the Fisher linear discriminant. It tries to optimise the class separation, the separation of the class-middle, while at the same time keeping as low variance as possible.

2.4.4.5 Clustering

Clustering is used to identify groups or structures in the raw data. A cluster is a group of data points where all the points in the group share more similarity to the other group members than to any other data point in the set. Three of the most common clustering methods will be reviewed briefly here.

- **k-means Clustering** aims to partition the dataset into k clusters, where k is specified in advanced by the user, and then minimises the dispersion within the clusters, by reducing the distances between each data point and the cluster average.
- **Hierarchical Clustering** iteratively joins the two closest clusters starting from single clusters (bottom-up approach) or iteratively partitions clusters starting from the complete set (top-down approach).

  The hierarchical clustering process can be represented as a dendrogram, where each step in the clustering process is illustrated by a branch or the dendrogram. There are several methods for measuring the distances between the clusters. Some of the most commonly used ones are: single linkage (distance between two clusters is the shortest distance between two members from each cluster), complete linkage (the distance is the longest distance between any two cluster members) and average linkage.
- **Self-organising maps** were developed by Kohonen (72). They are considered superior to hierarchical clustering when analysing “messy data” that contains outliers, irrelevant variables and non-uniform data (73). The idea is that a partial structure is imposed on the data and then adjusted iteratively according to the data to obtain a two-dimensional grid representing its distribution.

2.4.5. Looking for protein motifs

Motifs are consensus patterns of amino acids in a protein that are associated with a known function or structural feature. Sequences of related proteins often share consensus patterns or motifs of amino acids. There are a number of databases of protein motifs such as PROSITE (74), PFAM (75), Prints (76) and BLOCKS (77). They also provide analytical tools for recognising these specific motifs. Programs that will find new motifs in families of proteins have also been developed. One such program available on the World Wide Web is the Blockmaker (78) that finds blocks in groups of related proteins. It uses two sets of algorithms, the MOTIF algorithm (79) and a Gibbs sampler (80) and returns the blocks that were found with both algorithms. The Gibbs sampler is also available on the Internet (80).
MEME (Multiple Expectation-maximization for Motif Elicitation)(81) is a program for finding protein motifs based on statistical algorithm called expectation-maximisation. It tries

to fit a statistical model to its input sequences, and for each motif MEME maximises a likelihood function. MEME gives a scoring matrix that represents each motif and which can be used to search for homologous sequences (81).

2.4.6. Useful databases and web-based tools

With the increased sea of biological data, the need for structured databases becomes vital. The three main databases for nucleotide sequences are Genbank, which is maintained by NCBI (http://www.ncbi.nlm.nih.gov/Genbank/index.html) (a), EMBL by the European Bioinformatics Institute in the United Kingdom (http://www.ebi.ac.uk/) (b), and the DNA Database of Japan (http://www.ddbj.nig.ac.jp/fromddbj-e.html) (c). They all contain more or less the same sequences but use different annotation formats. There are also several databases with the sequences of entire genomes and genome maps.

For protein sequences the most used database is SWISS-SPROT (http://www.ebi.ac.uk/swissprot/) (d). In combination with the TREMBL supplementary database it can be searched as SWALL (SWISS-PROT+TREMBL+SWISSNEW+TREMBLNEW). A good database with protein structures is the PDB (Protein Data Bank) at RCSB (http://www.rcsb.org/pdb/) (e). A useful site that contains links to several protein databases and provides sequence retrieval tools is the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) (http://us.expasy.org/) (f).

In the field of immunology there are also several databases such as the MHCPEP (http://wehih.wehi.edu.au/mhcpep/) (g), SYFPEITHI (http://syfpeithi.bmi-heidelberg.com/) (h), FIMM (http://sdmc.krdl.org.sg:8080/fimm) (i), KABAT (http://immuno.bme.nwu.edu/) (j), IMGT (http://www.ebi.ac.uk/imgt/) (k) and HIV Molecular (http://hiv-web.lanl.gov/immunology/) (l). Many of them contain the sequences of MHC binding peptides, but also many other immunologically relevant sequences such as T-cell epitopes, antibody-binding sites and immunoglobulins.

For allergens there are several databases available. The most officially accepted list is the International Union of Immunological Societies, IUIS, Allergen Nomenclature List, (http://www.allergen.org/List.htm) (m). All allergens in the IUIS list have been thoroughly characterised and named according to the widely used IUIS system. One of the most extensive databases is the Allergome database (http://www.allergome.org/) (n). Other databases are: The Allergen Database (http://www.csl.gov.uk/allergen/) (o),The Allergen Sequence Database (http://www.iit.edu/~sgendel/fa.htm) (p), ProtAll (http://www.ifrn.bbsrc.ac.uk/protall/database.html) (q) and The FARRP Protein Allergen Database (http://www.allergenonline.com/default.asp) (r).

2.5. Genetically modified organisms and safety assesment

2.5.1. GMO - Genetically modified organism

The use of GMOs in agriculture started in the early 1990's with insect resistant corn and herbicide tolerant soybeans, and has now developed to cover about 30-50 % of all crops in North America (82). Most transgenic plants have been developed to improve plant yield, but some GMOs with direct consumer benefits have been developed such as the FlavrSavr^{R} tomato with improved ripening yielding better flavour preservation attributes (83), or the so-called golden rice with increased levels of vitamin A aimed to help control wide-spread vitamin A deficiency in Asian populations (84).

There are four main approaches to achieve genetic manipulation of plants. One widely used technique is **gene insertion**, usually done with the bacterial vector *Agrobacterium tumefaciens*. **Microballistic impregnation,** where the target gene is attached to tungsten or gold particles and fired into plant tissue at high velocity, is also common. The third technique is **poration** with a pulsed electric field or chemical treatment. **Gene neutralisation** can be done using antisense technology, homologous recombination and gene replacement. These methodologies are reviewed in (53).

2.5.2. Allergenic potential of GMOs

The introduction of novel foods and proteins, with potential to elicit allergenic reactions, to our market comes from various sources. Conventional breeding, genetic manipulation, introduction of new exotic foods to our market and changes in the food handling technology are just some examples. Although there are no scientific indications to make us expect that GMO crops will more frequently lead to allergic reactions, there still might be some allergenic consequences and therefore the safety issues must be considered. One example of introcustion of an allergen is the insertion of the Brazil nut 2S albumin into soybeans to enhance their level of amino acids methionine and cysteine. When this soy was tested according to the present evaluation procedure, the IFBC/ILSI 1996 decision tree (se below), the brazil-nut derived protein was found to be allergenic and no further production was done. Later the 2S albumin was identified as a major allergen of Brazil nut (85).

Introduction of new genes into a plant genome cannot only affect the allergenicity of the derived food by introducing new allergens. A recombinant protein can have altered function in the new host due to changed fold or processing and the glycosylation pattern of the protein may be altered in the new host. This could create new allergens that were not allergenic in its original species. Random integration of the new gene into the host genome can alter the levels of endogenous allergens thereby creating a more allergenic product (53). These effects of genetic modifications cannot be easily predicted without experimental testing of the product. Much more attention has been directed towards testing if the introduced protein has any allergenic properties.

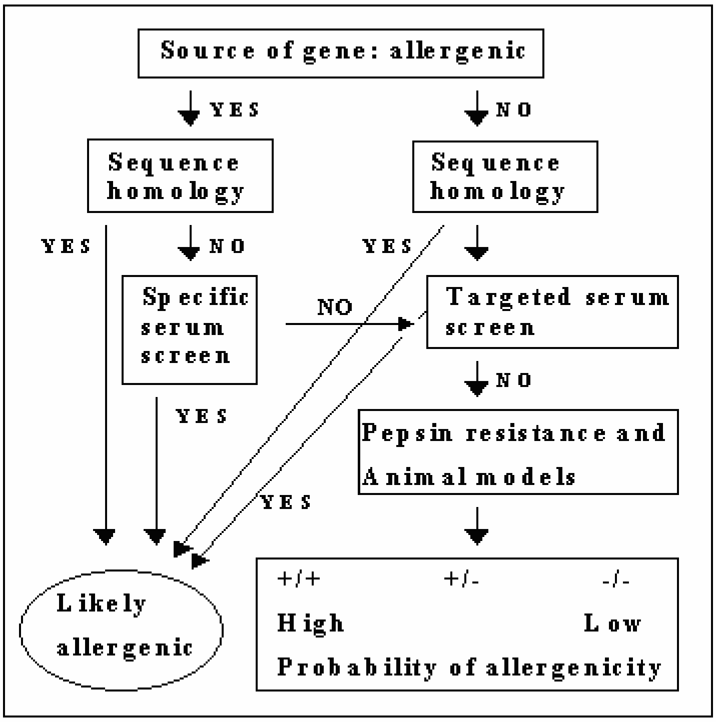


Figure 7. Schematic overview of the 2001 FAO/WHO decision tree.

2.5.3. Prediction of allergenicity

The most direct approach for detecting potential allergens is to test the response in animals. Commonly used are guinea pig, mouse and rat models (5). There are however, considerable variations to the results from these tests. Immuno assays for serum screening are also used, but there we have the problem of finding the appropriate human sera for testing each specific allergen (8). *In vivo* skin-prick tests or clinically supervised double-blind placebo-controlled food challenges are the last testing steps. In 1996 the International Food Biotechnology Council (IFBC) and the International Life Science Institute (ILSI) developed a decision tree for the evaluation of the potential allergenicity of novel gene products, which has been widely adopted in the agricultural biotechnology industry. Their strategy focuses on the source of the gene, the sequence homology of the recombinant protein to known allergens and the immunochemical binding of the introduced protein to IgE from serum of individuals with known allergies and the physiochemical properties of the protein (86). A joint World Health Organisation (WHO) and Food and Agriculture Organisation of the United Nations (FAO) consultation presented a revision of that decision tree in 2001. While the IFBC/ILSI procedure was focused on how the product should be labelled the new decision tree aimed to determine the likelihood that a new protein will produce allergic reactions (8). The 2001 FAO/WHO decision tree is presented in Figure 7.

2.5.4. In silico methods

The definition of sequence homology in the 1996 IFBC/ILSI decision tree was the identification of an identical stretch of eight amino acids or more, based on the findings that the optimal peptide length for binding T-cell epitopes appeared to be between 8 and 12 amino acids (86). Recently it has been found that small sequences of four and six amino acids can be recognised and bound by IgE from sera of allergic patients (8). Therefore in the 2001 FAO/WHO decision tree the definition of sequence homology was changed to an identical stretch of six amino acids or more than 35 % sequence identity over a 80 amino acid window (8). The minimal degree of identity is a hot topic of current discussions in the FAO/WHO.

Using 35% sequence identity over 80 amino acids is meant to detect homologous proteins. Hileman *et al.* tested 50 random maize proteins and found that a 35% limit gave too many false positives. Due to this finding they instead suggested a limit at 50% (87). But typically a sequence identity of over 70 % is required for allergenic cross-reactivity (3). There is also a debate on whether using an identity matrix is appropriate. Gendel *et al*. suggest the use of an evolutionary matrix for sequence alignment and have tested it with good results (88).

Many of the 6 amino acid identical hits are likely to be false positives. Kleter *et al.* (89) tested 33 transgenic proteins and found 22 identical stretches. These identical stretches were compared with IgE epitopes reported in literature. They also used computer algorithms to predict the most antigenic sites of the allergens and then compared if these sites coincided with some of the identical stretches. They suggest using this method to eliminate false positive by only focusing on those hits that also match sites of high antigenicity or documented IgE-binding sites (89). With this approach however, they do not take into account the potential T-cell epitopes in the allergen.

It is clear that more work has to be done before we with certainty can classify all allergens. Development of better allergen databases is in progress (8). Much effort is put into development of better experimental methods such as animal models and immunoassays. At the moment the FAO/WHO method only provides a “reasonable certainty of no evidence of allergenicity”, not a 100% certainty (85).

In our group at the National Food Administration in Sweden more sophisticated methods for prediction of allergenicity based on local alignments have been developed (90, 91). In all these methods, data descriptors were obtained from local alignments using FASTA3 (66), and three different machine learning algorithms, k-nearest neighbour (kNN) classification, linear gaussian classification (LG) and quadratic gaussian classification were applied to the problem. After extensive evaluation of key parameters and testing for optimal data descriptors a model with a LG algorithm, using a scoring matrix that combines the scores from BLOSUM50 with the protein identity matrix, was determined as the best model. This gave a prediction method with a probability of detection of allergens at 0.70 and a probability of false alarm at 0.11 (both with 95% confidence). This is a more accurate oprach compared to looking at identical matches of 6 amino acid stretches or 35% homology. Particularly important is the fact that the rate of false alarm (classifying non-allergens as allergens) was shown to be much lower without decreasing the rate of detection (91).

3. PROBLEM STATEMENT AND STRATEGY

The aim of this project is to develop computational methods for the classification of allergenic proteins by looking at their primary amino acid sequence. Previous methods for classification have been focused on finding homology with other allergens. These methods are very useful for identifying cross-reactive allergens and for identifying new allergens in already known allergen families.

Is it possible to tell in advance if a new protein will induce allergic responses when it is not homologous to a known allergen? Turning the focus towards the allergenic proteins, and what distinguishes them from their non-allergenic counterparts, might make it possible to find an answer. As reviewed in section 2.3. no one has so far been able to find any common features to all allergens. The mechanisms leading to an allergic response are very complex and not yet fully understood. It is not clear at what key point in the chain of reactions that the immune system determines that it will respond in an atopic manner to some proteins but not to others.

It has been suggested that the interaction between the epitope-MHC II complex and the T-cell receptor might play an important role in this decision. The TCE of a few allergens have been mapped, and some of them are recognised by both atopic and non-atopic individuals and in some cases similar epitopes have been found on non-allergenic proteins (16). It would be interesting to find out if there are any TCEs that are specific for allergens. Looking at these TCEs could be a good approach for identifying allergens, but since there are only a few allergens with mapped TCE available it would be a difficult task. Furthermore, the atopic response produced by the allergens may not only be because of the TCEs. There could be other structural or functional motifs affecting for example uptake by APCs, processing in APC vesicles and other unknown features of the immune system.

The working hypothesis is that there has to be some characteristics of the allergens that the non-allergens lack. The aim of this project is to find what these characteristics are. For this purpose a method for identification of allergen-specific peptides has been developed called peptide filtration; a more detailed description follows in the methods section. When these candidate motifs have been identified a classifier for identification of new allergens could be built based on homology to those candidates. With this approach the objective is not only to develop a new method for classification of allergens, the aim is also to identify certain elements that make them allergenic. This may lead to better understanding of the mechanisms leading to atopy and help in the development of immunotherapy methods.

It is important to mention that this approach is not meant for finding cross-reactive proteins, since IgE-cross reactivity only occur between highly homologous proteins and these can easily be detected with multiple alignments. We rather aim at developing a model that would asses the risk that a protein would *de novo* sensitise a susceptible individual.

4. METHODS

4.1. Methods Overview

For this project it was crucial to have amino acid sequences of allergens and non-allergens to work with. Databases were constructed, one allergen database with all certified documented allergens, and two different non-allergen databases, one with sequences from many species, and one with all the proteins in rice, construcion of databases is described further in section 4.2. These databases were divided into test and training sets, training sets for setting up the peptide filtration and test sets to evaluate classification.

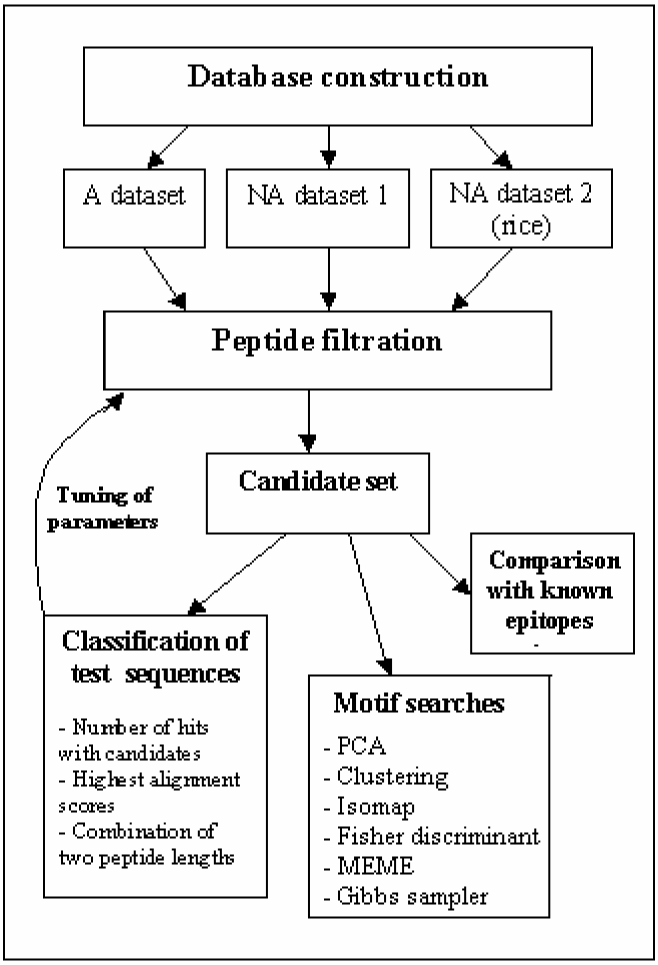


Figure 8. Overview of methods used in this project.

All sequences were chopped up into peptides, different lengths for these peptides were tested. Then the allergenic peptides were compared with the non-allergenic peptides in one or both of the non-allergen databases with a method called **Peptide filtration** (see section 4.3.) to determine the similarity between the peptides. With the information from the peptide filtration a set of candidate peptides, with low similarity to the non-allergens, were selected from the allergens. These candidates should represent motifs that are unique to allergens since they are less common among the non-allergens.

When classifying test sequences high similarity to the candidates gave high probability to be classified as an allergen. This measure of similarity was determined in two different manners, either as the number of hits between test sequence peptides and candidates over a certain similarity threshold or as a combination of the highest alignment scores between test sequence peptides and candidates. Classification counting the number of hits was also tested with the combination of two peptide lengths. All classification methods are described in section 4.4.

There were many parameters of the peptide filtration that had to be tuned, such as the optimal peptide length, the threshold for determining similarity, as well as the number of peptides that should be selected from each allergen. Optimal parameters for peptide filtration were determined as the ones that gave best classification of the test sequences.
When the best candidates for each peptide length had been determined the motif searches were done to determine if there were any particular pattern among them. The allergenic candidate peptides were compared with a set of randomly chosen non-allergenic peptides and several different visualisation techniques were applied in an attempt to find patterns

distinguishing the candidates. Two motif search algorithms available on the Internet were also used to look for motifs in the candidates; these were MEME and Gibbs motif sampler. The candidates selected were also compared with experimentally determined T-cell epitopes and IgE epitopes. Further description of motif searches is provided in section 4.6.

In addition to peptide filtration using all allergens and a broad selection of non-allergens, another test was done using more specific datasets containing only sequences from the protein-family profilins. Peptide filtration was done with an allergen set and a non-allergen set both containing only profilins, more detailed explanation is provided in section 4.5.

4.2. Database construction

4.2.1. The allergen database

An allergen database was constructed by joining annotated allergens in five publicly available databases: FARRP (92), The Allergen Database (93), The Allergen Sequence Database (94), ProtAll (95) and Allergen Nomenclature (96). Then a thorough search was done to avoid duplicates and uncertain allergens. Presumed allergens without literature documentation of being allergenic were excluded. Fragments of allergens smaller that 200 amino acids were also excluded thereby preventing short parts of the allergens, that may lack crucial information, in the dataset. To avoid a biased allergen set with too many similar sequences only one representative from each isoallergenic family was included in the database. The total amount of allergen sequences in this database was 324.

4.2.2. The non-allergen databases

Two non-allergen databases were constructed, with the aim to have a wide variety of proteins including representatives from all protein families in the dataset, but still excluding possible allergens. For the peptide filtration it is crucial to have a broad representation of all possible peptides that occur in non-allergens, since they are used to find protein motifs in allergens that do not occur in non-allergens.

Only species where thorough documentation of allergens has been done were used to build the first dataset. And all proteins non documented as allergens were used to build a non-allergen data set, with the assumption that all other proteins in these species are not allergenic. The species whose sequences were retrieved from SWALL(97) were *Lycopersicon* (tomato), *Malus* (apple), *Prunus* (peach, cherry, apricot), *Spinacia oleracea* (Spinach), *Daucus carota* (Carrot), *Salmo salar* (salmon) and *Gadus* (cod). Some proteins from cows milk and hens egg were also included. This resulted in a non-allergen dataset with 3370 sequences. More detail on the search criteria used to obtain this dataset is available in Appendix 2.

Unfortunately the representation of the different species was uneven, since we are limited only to those sequences that are available in SWALL. The ideal would be to have the whole proteome of these species in the database to get representatives from all protein families. One species with such extensive documentation was found, *Oryza sativa* (Rice). It was possible to retrieve 18812 sequences from rice that were used to build our second non-allergen dataset.

4.3. Peptide filtration

4.3.1. The peptide filtration method

The purpose of the peptide filtration is to find the peptides that are most typical for allergens, i.e. that are less common in the non-allergenic dataset. This is done by comparing short peptides from allergens with short peptides from non-allergens and assigning similarity scores to the allergenic peptides. A schematic overview of the peptide filtration method is provided in Figure 9.

In peptide filtration all protein sequences in both the allergen dataset and the two non-allergen datasets are first chopped, using a sliding window, into short peptides of sequence length $\mathbf{l_P}$ To narrow down the amount of non-allergenic peptides, all identical peptides are eliminated from the dataset leaving, only one copy of each peptide, but still counting the frequency of the peptides. Removing all peptides occurring only once in the non-allergen dataset is done to bring down the amount of data thereby making the following steps less time-consuming.

In the second step, all peptides from the allergen set were compared to all non-allergenic peptides. Measure of similarity was alignment score, $\mathbf{S_A}$, using scoring matrix BLOSUM80, not allowing for any gaps. Similarity was determined by a threshold for the alignment score denoted $\mathbf{S_{A1}}$. The number of times an allergenic peptide gives an alignment score with a non-allergenic peptide over this threshold S_{A1} is counted as the similarity score, $\mathbf{S_S}$.

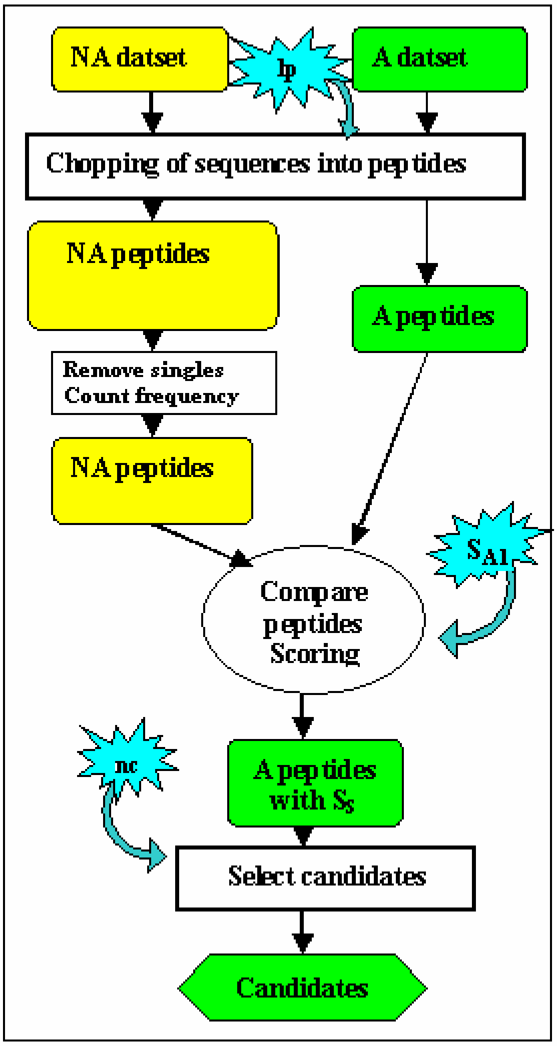


Figure 9. Schematic overview of peptide filtration, explained in 4.3.2.

When this scoring procedure had been done for all allergenic peptides a candidate set was selected. From each allergenic protein a number of peptides, $\mathbf{n_C}$, were selected to form the candidate set. This selection of candidates was done choosing the ones with the lowest similarity score (S_S) i.e. the ones that are most dissimilar to the peptides found in the non-allergen dataset. Peptide filtration was done with just one of the two non-allergen sets or with the combination of both of them.

4.3.2. Validation of peptide filtration parameters

Several parameters of this peptide filtration method had to be tuned. The threshold for the alignment score $\mathbf{S_{A1}}$, had to be optimised. The number of peptides that should be chosen from each allergen, $\mathbf{n_C}$, was looked at. It was not evident what would be the optimal peptide length $\mathbf{l_P}$ for identifying allergenic motifs, so several peptide lengths had to be tested.

4.4. Classification and validation

4.4.1. Classification

Once a set of candidate peptides had been chosen they were used for classification of the test sets. Classification is based on similarity to these candidate peptides. Before the peptide filtration step, a section of the sequences was put aside for this classification. From the non-allergen dataset 700 sequences were withdrawn randomly to create a test set, 600 from the rice non-allergens and 100 from the rest. Since there were only 324 sequences in the allergen set, a cross-validation was set up for them. Each time 50 sequences were used as test set and candidates were selected from the remaining 274. This was repeated 6 times with 6 unique test sets.

The classification was done as depicted in Figure 10. The two test sets (in this case 700 non-allergens and 50 allergens) were chopped up in peptides of same length l_P as the candidate peptides. Then each peptide was compared to all the candidates in the candidate set. Again the measure of similarity was alignment score, $\mathbf{S_A}$, using BLOSUM80, but the threshold for similarity was another, here it is called $\mathbf{S_{A2}}$. For each test protein, the number of times any of its peptides gives an alignment score with a candidate over the threshold S_{A2} was counted as the similarity score $\mathbf{S_S}$. In the end this score is divided by the number of peptides each protein has ($\mathbf{n_P}$). The last step consists of classifying all sequences with a high $\mathbf{S_s/n_P}$ ratio as allergens and the rest as non-allergens. ROC curves were done to describe how the classification results differ when the threshold for $\mathbf{S_s/n_P}$ is altered.

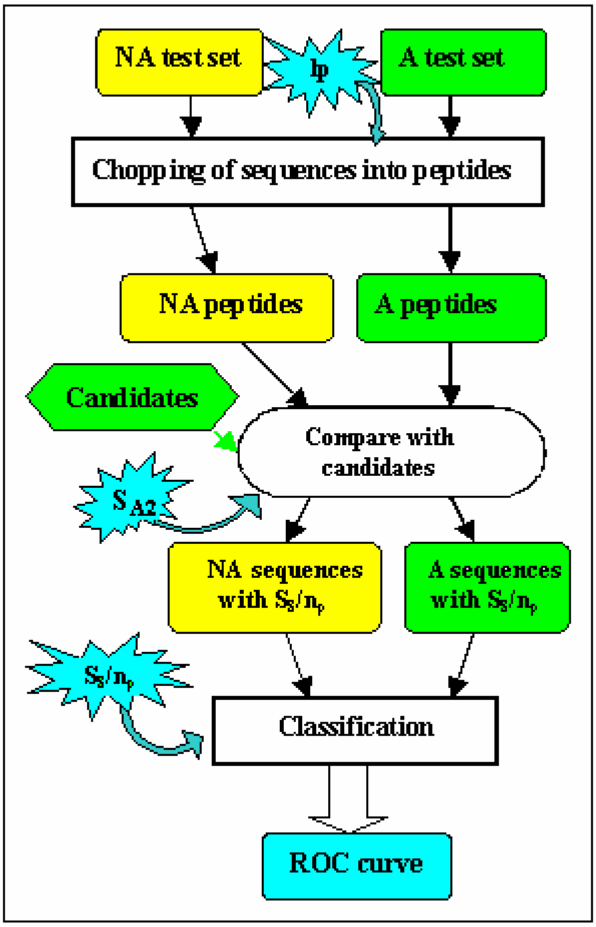


Figure 10. Schematic overview of classification using S_s/n_P ratios, detailed description in 4.4.1.

In this manner all the six allergen test sets were classified together with the non-allergen test set. The characteristics of the classifier were evaluated by plotting average ROC curves for all six classifications. This was done for peptide lengths 6, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42 and 45. The optimal S_s/n_P threshold, the threshold S_{A2} when comparing the test set peptides with the candidates, as well as all the parameters of the peptide filtration step (S_{A1} and n_C) were all determined from the ROC curves to get desired classification performance.

4.4.2. Classification with other parameters

Other methods for classification were also tested. In order to take into account data from two peptide lengths, to see if different allergens are detected with different peptide lengths, a classifier using the combination of the S_s/n_P ratios for two lengths as a two-dimensional feature vector was built. This gave two sets of thresholds, one for each S_s/n_P ratio that had to be determined. Classification as an allergen was done when one or both of the S_s/n_P ratios for a protein was above its respective threshold. To visualise the classification performance, the pDetection and the pFA were plotted in separate three-dimensional plots as functions of the two S_s/n_P thresholds. This method for classification was done combining scores from peptide length 30 with scores from peptide length 6, 12 and 24.

The classification described in section 4.4.1. counts the number of hits between candidates and test sequence peptides over a certain alignment score threshold, S_{A2}. But the values of those alignment scores are not taken into account when doing the classification. In order to use this information another classifier was built that uses the highest alignment scores, $\mathbf{S_A}$, between any pair of candidate and test sequence peptide as the features for classification. Classification with these alignment scores was done in two different manners. The first method used the **m** highest scores added together to a score $\mathbf{S_{sum}}$, and classification was based on how high this score S_{sum} was. ROC curves were plotted to see how the results differed with varying S_{sum} thresholds. This was tested with m equal to 1, 2, 5 and 10. This method for classification was done with peptide lengths 6, 12, 18, 24 and 36.

Two-dimensional classification with the two highest alignment scores were also set up similar to the one using the S_s/n_P ratios from two peptide lengths above, plotting the pDetection and the pFA as functions of the two thresholds.

All these classification methods were cross-validated with the same six allergen test sets and using one non-allergen test set as described above.

4.4.3. Comparing with classifications with randomly selected peptides and highest scoring peptides.

To evaluate how good the peptide filtration is for selection of the candidates that are important for the allergenicity of the proteins, the classification was done with the same methods as described in 4.4.1, but instead using a candidate set that was created by choosing randomly n_C peptides from each allergen. Another test classification was done for selecting, to the candidate set, the n_C peptides with the highest similarity scores, S_S, i.e. the ones that are most like the non-allergenic peptides. Classification with these “worse” candidates should be much worse than with the “best” candidates if the peptide filtration method truly finds the most allergen-specific peptides.

4.4.4. Classification using identical amino acid stretches

The FAO/WHO recommends using identity with a known allergen over 6 amino acid stretches to determine if a query protein is a potential allergen (8). In the previous 1996 IFBC/ILSI decision tree identity on 8 amino acid stretches were recommended (86). To test if the peptide filtration model is better then their recommendations classification was done with the same datasets as before, but using sequence identity over a 6, 7 and 8 amino acid-stretch as determinant of allergenicity.

4.5. Peptide filtration with profilins

One common protein group among plant allergens is profilins. Our allergen dataset contain 23 profilin sequences. There are also many profilins in plants and in other organisms without any documented allergenic properties. All these profilins were downloaded from SWALL (96), which gave a non-allergen profilin data set containing 93 sequences. Note: Even though none of the 93 profilins that were downloaded had any annotation indicating their allergenicity, the possibility that they are allergens cannot be excluded, therefore it is not a very certain non-allergen set. Peptide filtrations were set up with these datasets without removing any single occurrences or sorting the peptides in either dataset. Classification of the profilins was evaluated with a five-fold cross validation where each test set contained 4 allergens and 18 non-allergens. This was done for peptides of lengths 6, 12 and 24 amino acids and the classifications were presented as average ROC curves.

All the profilins where multiple aligned using Clustal W (68) and mapping of where the 6 amino acid candidates were located in the sequences was done. The aim was to see if any particular part of the protein seemed to differ between the allergens and the non-allergens.

4.6. Motif searching

Once a candidate peptide set had been established the aim was to find if there were any common motifs among those peptides. If such a motif is found it is possible that it could be a motif particular for allergens that plays an important role in the allergenic process.

4.6.1 Visualisation techniques

Four methods for visualisation were applied to the 6, 12 and 24 amino acid candidates that gave the best classification. These were compared with a set of randomly chosen non-allergen peptides of the same length. The methods tested were PCA, ISOMAP, hierarchical clustering with complete linkage and Fisher discriminant analysis. With PCA and ISOMAP the aim was to find differing pattern for the candidates relative to the non-allergenic peptides. With hierarchical clustering it was to find clusters specific for candidate peptides. Possibly the candidates differing a lot from the other peptides could contain some allergen-specific motifs. With Fisher discriminants the aim was to see if it is possible to distinguish the two groups of peptides, the non-allergens and the candidates.

For each method, different representations of the peptides were tested. First the sequences were represented with a n-dimensional vector with different numbers from 1 to 20 representing each amino acid, where n is the peptide lenght. Secondly each amino acid in the peptides was represented with five-dimensional ZZ-scales, where each ZZ-scale represents the “principal properties” of the amino acid. The five components of the ZZ-scales are molecular weight, hydrophobicity, isoelectric point and two combinations of several features, these properties are described by Venkatarajan *et al.* (98). The third method used seven groups of amino acids depending on their physiochemical properties, and representing them with seven-dimensional binary vectors with ones at one position depending on the group and zeros at the remaining positions. The amino acid groups were: Acidic (D,E), Aliphatic (A,G, I, L, P, V), Amidic (N, Q), Aromatic (F, W, Y), Basic (R, H, K), Hydroxylic (S, T) and Sulfur containing (C, M) as described in (48).

ISOMAP was run with 50 landmarks and 100 landmarks, 10 dimensions and tested with K, the number of neighbours/point, set to 3 ranging to 10. The Euclidian distance for the peptides, represented as described above, was used as the distance measure when performing ISOMAP. Another two tests with ISOMAP were done using, first the alignment score with BLOSUM 80 for alignments between all pairs of peptides as the distance measure. And secondly with the percent sequence identity between all peptides as the distance measure. Hierarchical clustering was done with all the same distance measures as with ISOMAP.

4.6.2. Motif search tools

There are a few protein motif search tools available on the Internet but most of them are limited to an input of 50-200 sequences. So even though the candidates are very short, it is still not possible to use many of the methods since there are too many candidates. To narrow down the number of candidates to look for motifs in, all of the candidates were first clustered with hierarchical clustering using complete linkage and distance measures calculated from alignment scores. Then the members from one cluster at a time were used to look for motifs. The two methods that were tested were MEME (81) and Gibbs Motif Sampler (80). Some of the motifs found were mapped to their positions and also searched for in homologous allergens and non-allergens.

4.6.3. Comparing with mapped epitopes

There are several T-cell epitopes and IgE-binding epitopes that have been mapped experimentally using T-cell lines and IgE antibodies extracted from sera of allergic patients or mice. The position of the candidates with peptide length 12 amino acids and 30 amino acids were manually mapped to their position in the protein sequences and compared with the position of experimentally mapped epitopes. This was done for the allergens Tri r 2 (99), Bos d 6 (100), Cry j 1 (101, 102), Cry j 2 (102, 103) and Asp f 2 (104, 105).

4.7. Implementation

All methods were implemented using MATLAB® m-code. Amino acid sequences were represented with prime numbers to enable faster implementation of sequence comparison without using for-loops. This method, developed together with Mats Gustafsson, is described in more detail in Appendix 3.

5. Results

5.1. Peptide filtration

Before peptide filtration was run the non-allergen datasets were sorted and peptides only occurring once were removed. This resulted in a dataset with 10-20% of the original number of peptides depending on the peptide length. For 6 amino acid peptides the rice non-allergen dataset contained 1 185 918 peptides and the other non-allergen set contained 165 248 peptides. With 24 amino acids the rice set contained 700 460 peptides and the other set 135 593 peptides. With other peptide lengths the sets were of similar sizes.
After chopping the allergen dataset into peptides and removing all duplicates it contained 67 394 peptides of length 6 and 75 887 peptides of length 24.

5.2. Classification

5.2.1. Classification with both non-allergen datasets

Classification of the test sets, using both of the non-allergen datasets for peptide filtration, was done as described in section 4.4.1., with several different peptide lengths and the results were plotted in ROC curves. For each peptide length the classification was done using several different settings for S_{A1}, the alignment score limit for the peptide filtration, for n_C, the number of candidates selected from each allergen and for S_{A2}, the alignment score limit for the classification of test sequences. Based on the ROC curves, the best classification for each peptide length was determined as the point where high pDetection can be obtained without having too high pFA, at least not much over 0.15. How best classification was determined is illustrated in Figure 11.

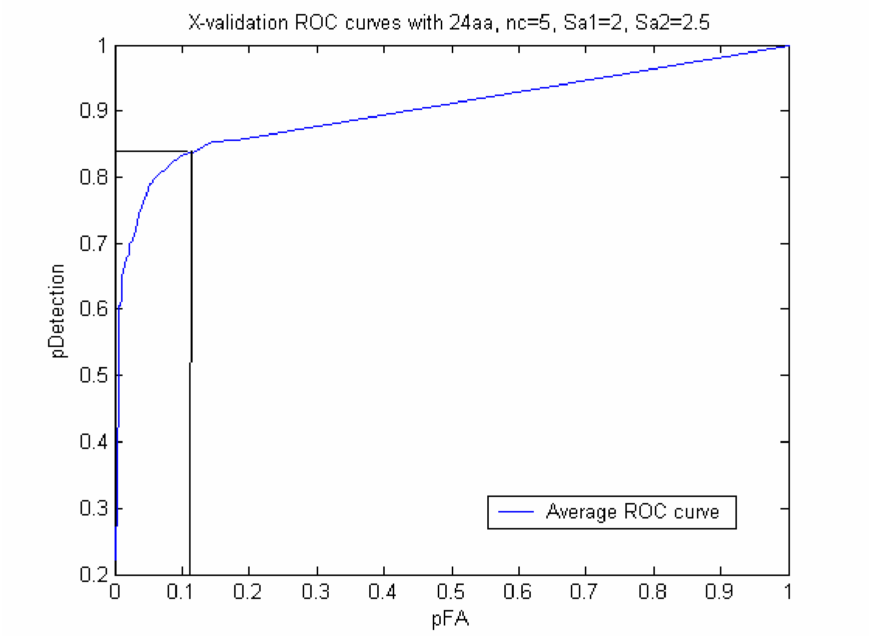


Figure 11. Classification with peptide length 24, n_C=5, S_{A1}=2 and S_{A2}=2.5.

Best results at average pDetection=0.840 and average pFA =120 as illustrated with lines.

The classification results are the averages of all six cross-validation runs and for them the standard deviations were calculated assuming normal distribution of pDetection and pFA. The results are presented with a 95.45 % confidence, i.e. ± two standard deviations. This way the highest probability of detection of allergens (pDetection) with 95 % certainty and the lowest probability of false alarm (pFA) with 95 % certainty is shown. Results are found in Table 2. Best classification, with 95% confidence, would be with 24 amino acid peptides, n_C=5, S_{A1}=2 and S_{A2}=2.5. This gave a pDetection of 0.778 and a pFA of 0.137

Settings:				**Optimal results**							
l_P	**n_C**	**S_{A1}**	**S_{A2}**	**pDetection**	**STD**	**w/ 95% c**	**pFA**	**STD**	**w/ 95% c**	**Random**	**Worse**
6	20	5	8	0.757	0.066	**0.624**	0.168	0.001	**0.171**	Worse	Worse
12	10	3	5	0.780	0.064	**0.651**	0.073	0.006	**0.084**	Worse	Worse
15	5	3	4	0.797	0.051	**0.694**	0.056	0.003	**0.062**	Better	-
18	5	1	3.5	0.847	0.059	**0.729**	0.154	0.016	**0.185**	-	-
21	7	2.5	2.5	0.830	0.058	**0.715**	0.132	0.012	**0.155**	Same	-
24	5	2	2.5	0.840	0.031	**0.778**	0.120	0.008	**0.137**	Worse	Better
27	5	0.5	2.5	0.827	0.062	**0.704**	0.064	0.003	**0.070**	-	-
30	5	1	2	0.847	0.067	**0.714**	0.144	0.010	**0.164**	Worse	-
33	5	1	2	0.823	0.064	**0.696**	0.040	0.003	**0.046**	Better	
36	5	0.5	1.5	0.857	0.066	**0.724**	0.134	0.006	**0.146**	Worse	Worse
39	5	1	1.5	0.833	0.070	**0.693**	0.083	0.003	**0.090**	Better	Worse
42	3	0.5	1.5	0.810	0.080	**0.651**	0.046	0.003	**0.052**	Worse	-
45	1	0	1	0.837	0.067	**0.702**	0.155	0.012	**0.178**	Worse	-

Table 2. Best classification results for each peptide length with average pDetection and pFA, standard deviations (STD) and with 95% confidence. The two last columns indicate if classifications selecting candidates randomly and selecting worse candidates (the ones with highest similarity to non-allergen peptides) was better or worse than these best classification results.
When using peptide lengths 42 and 45 it was not possible to select more than three respectively one candidate per allergen since the shortest allergen is 45 amino acids long.

Some of these classifications were compared with classifications using randomly selected candidates and selecting candidates with highest similarity scores (Worse candidates) as described in section 4.4.3. In most cases the random selection of candidates gave worse classification, even if it was not a large difference. In a few cases the classification was actually better. These results also presented in Table 2.

5.2.2. Classifications with the two non-allergen data sets

To determine if there was any difference in the results when doing the classification with the two different non-allergen datasets the classification with 24 amino acid peptides, n_C=5, S_{A1}=2 and S_{A2}=2.5 was tested with peptide filtrations using only the first non-allergen dataset, the rice non-allergen data set and the combination of both of them. Results presented as ROC curve in Figure 12.

Using both datasets seems to give the best classification, but the difference when using only the rice data set and using both of the data sets is very small. So in all other peptide filtrations both of the non-allergen datasets were used.

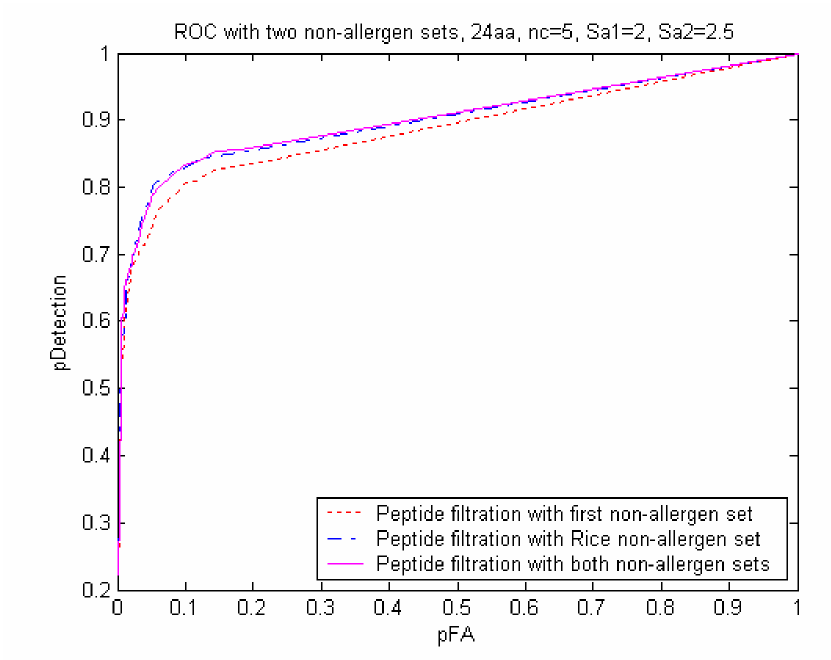


Figure 12. Average ROC curves for classification with peptide length 24, using the rice non-allergen data set , the other non-allergen set and combination of both for peptide filtration.

5.2.3. Classification with two peptide lengths

Classifications were done by combining the S_S/n_P ratios from two peptide lengths. This was tested with combination of 30 amino acid scores with 24, 12 and 6 amino acid scores. The results were visualised with pDetection and pFA as functions of the two-dimensional decision boundary for the two scores, as shown in Figure 13. The classifications with 30aa together with 24 the classification results were not improved at all, as compared to classification using only 30aa. In combination with 6aa and 12aa the result was slightly improved. When doing classifications with only 30 aa the best average classification was with pDetection= 0.847 and pFA=0.144, but together with 12aa scores it was improved to 0.86 and 0.132 respectively. Note: only average classification results were compared here. The results were still not any better than the optimal results presented in Table 2.

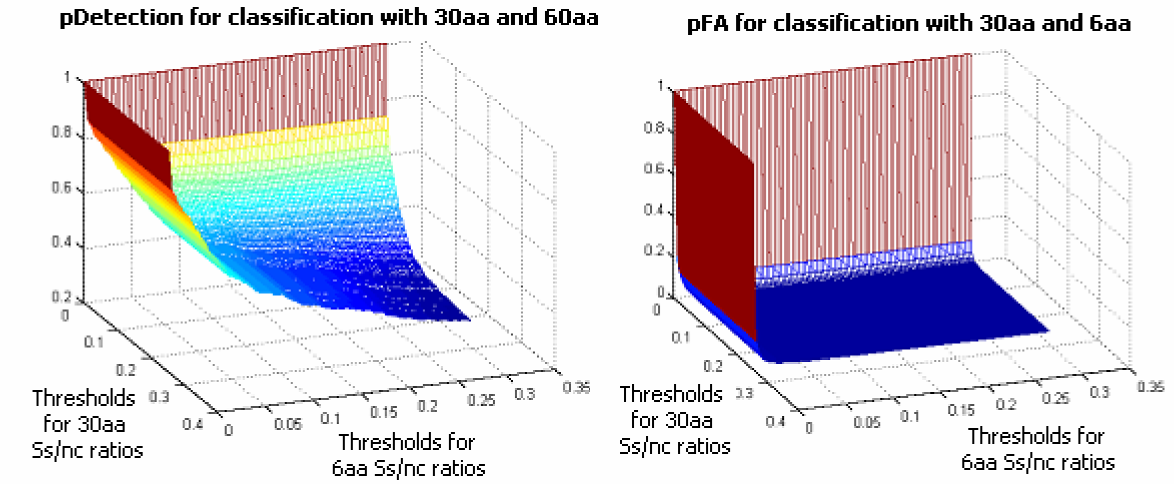


Figure 13. Classification using S_s/n_P ratios from classification with peptide lengths 30 and 6. pDetection and pFA as functions of the variation in the two thresholds for S_s/n_P ratios.

5.2.4. Classification using best alignment scores

Classification based on highest alignment scores between candidates and test sequence peptides, as described in section 4.4.2., was tested with peptide lengths 6, 12, 18, 24 and 36. The best results were with peptide length 18. Running classifications with n_C=5 and S_{A1}=1 was done with the highest alignment score and with the sums of 2, 5 and 10 alignment scores. These results are presented as average ROC curves in Figure 14. Best classification was with the sum of the two highest scores, but unfortunately there were very high standard deviations for the probability of detection. Even though the probability of False Alarm would be zero and the average pDetection=0.889, there would be a standard deviation for pDetection of 0.136, meaning that with 95% confidence we can only say that the best pDetection is 0.617.

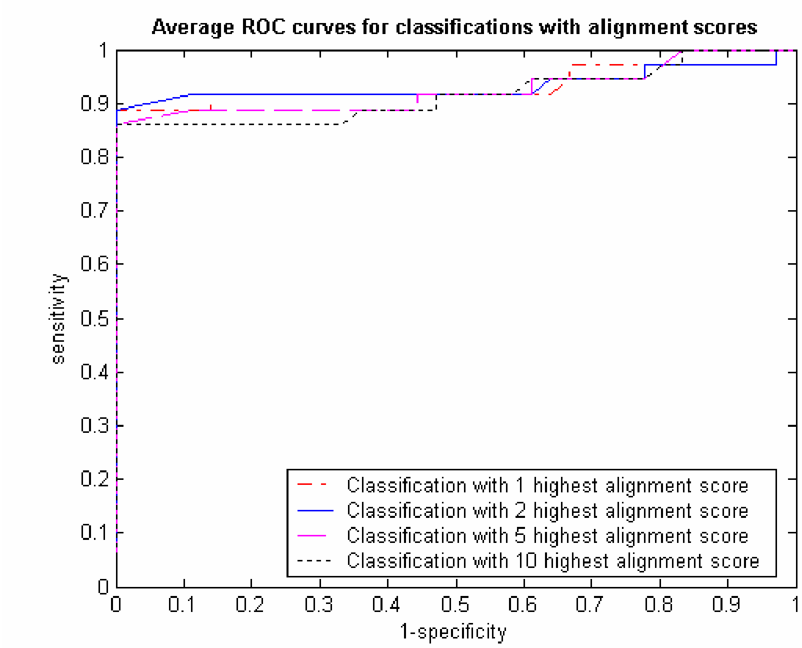


Figure 14. Average ROC curves for 6-fold cross-validation of classification using highest alignment scores. Classification using 24 amino acid candidates and adding together 1, 2, 5 and 10 highest alignment scores.

Use of the highest alignment scores for classification was also tested in two dimensions, plotting pDetection and pFA in three dimensions as functions of the two-dimensional decision boundary. Best classification obtained with this method gave, as in the example above, quite high standard deviations. Here the best classification would be with average pDetection=0.78 (0.653 with 95% confidence) and the average pFA = 0.0528 (0.0615 with 95% confidence).

5.2.5. Classification using identical amino acid stretches

Classification was done with the same datasets, using sequence identity over a 6, 7 and 8 amino acid-stretch as determinant of allergenicity as described in 4.4.4. The results are presented in Table 3.

Table 3. Results for classification using identical stretches of 6, 7 and 8 amino acids as determinant of allergenicity.

Peptide length	**pDetection**	**STD**	**w/ 95% conf.**	**pFA**	**STD**	**w/ 95% conf.**
6	0.903	0.0513	**0.801**	0.640	0.0078	**0.655**
7	0.783	0.0599	**0.664**	0.150	0.0099	**0.169**
8	0.720	0.0473	**0.625**	0.0243	0.0029	**0.030**

5.3. Peptide filtration with profilins

Peptide filtration and classification using only profilins was done with peptide lengths 6, 12 and 24. Results for the 5-fold cross-validated classifications were presented as average ROC curves. The best probabilities of detection was obtained with peptide length 6 and n_C=5, S_{A1}=5 and S_{A2}=6, see Figure 15. Same classification was tested selecting the candidates randomly or selecting the peptides with the highest scores ("worse" candidates) to compare if the peptide filtration actually had any effect and classification performance was not as good with random selection, and with "worse" candidates the classification was even lower as can be seen in Figure 15. This may indicate that peptide filtration was better at finding allergen-specific peptides than with the more general datasets.

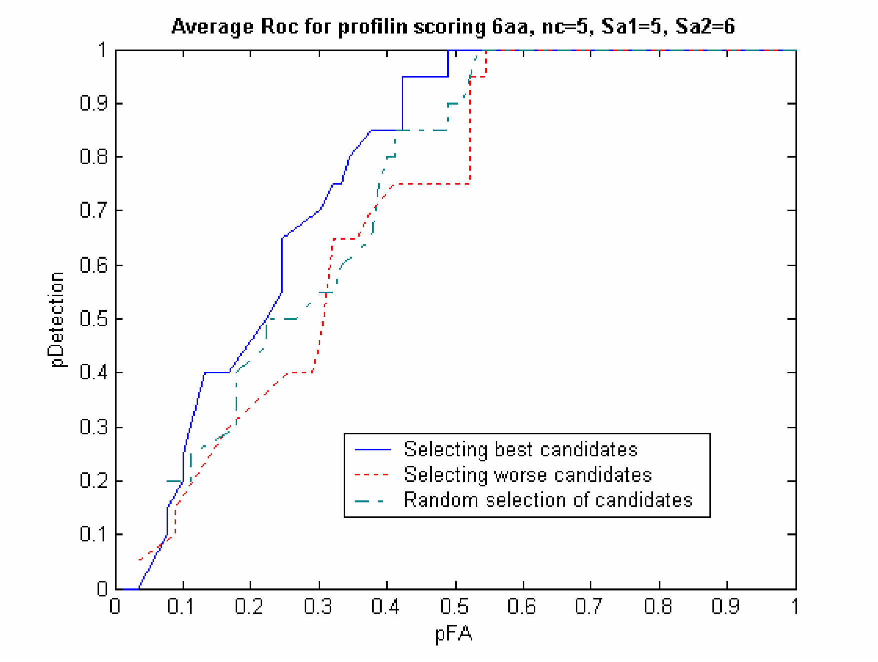


Figure 15. Average ROC curves for 5-fold cross-validation of classification of profilins. Classification using 6 amino acid candidates selected with lowest scores (full line) and highest scores (dotted line) from peptide filtration and randomly chosen candidates (dash-dot line).

When multiple alignments was done with all profilins there was no clear grouping of the allergens, even if some of them were found to be the more similar to each other than to the non-allergens, many others were not. When manually mapping where in these multiple aligned sequences the selected 6 amino acid candidates were located it was difficult to distinguish any particular pattern (results not presented). Therefore it is not possible determine if any particular profilin-allergen-specific motif has been found.

5.4. Motif searching

5.4.1. Visualisation tools

Several different visualisation tools were applied to the task of finding some allergen specific motif among the candidate peptides. The 6, 12 and 24 amino acid candidates giving best classifications, as presented in Table 2., were compared with sets of randomly chosen non-allergenic peptides of the same length.

5.4.1.1. PCA

When performing PCA with the peptides and plotting candidates and non-allergenic peptides together they looked evenly distributed. This was done for peptides represented with ZZ-scales and with binary vectors for seven amino acid groups, but no particular pattern could be seen with either methods nor with either peptide length, as can be seen in Figure 16.

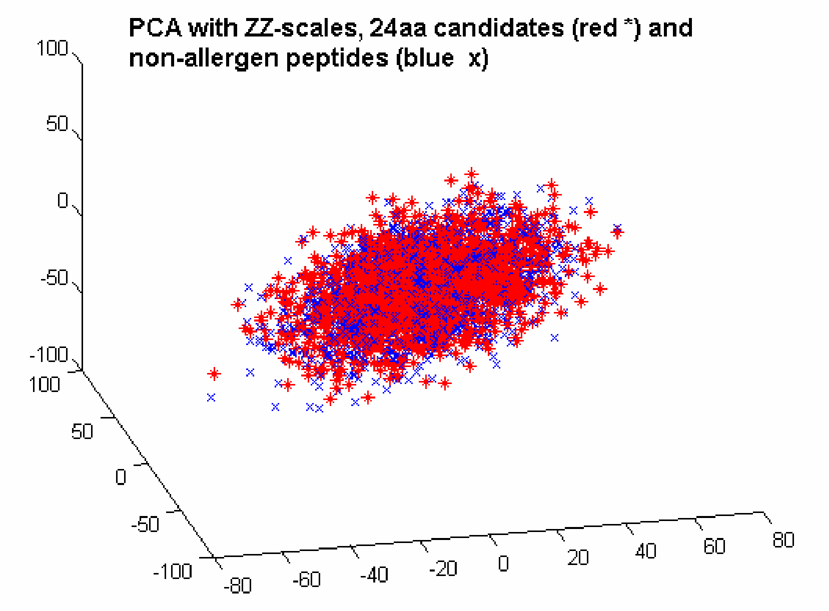


Figure 16. PCA plot for 24aa candidates (red *) and 24aa non-allergen peptides (blue x) transformed to ZZ-scales. In this example as well as with other tested methods it was not possible to distinguish any patterns

5.4.1.2. ISOMAP

ISOMAP was applied to the same peptides, plotting candidates and non-allergenic peptides together. This was done calculating the Euclidian distances for peptides represented with ZZ-scales and with binary vectors for seven amino acid groups, or with distances calculated from alignment scores and percent sequence identity. No particular pattern could be seen with either method nor with either peptide length, as can be seen in Figure 17.

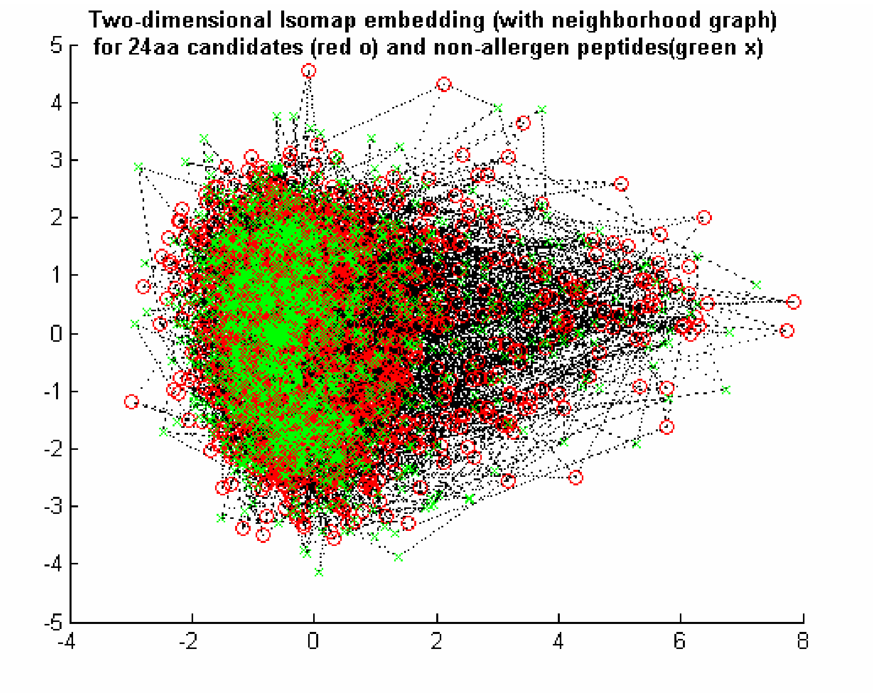


Figure 17. Isomap plot with 24aa candidates (red circle) and 24aa non-allergen peptides (green x) transformed binary vectors for seven amino acid groups. Isomap was run using 100 landmarks, 10 dimensions, K=3 candidates (red *) and 24aa non-allergen peptides transformed to binary representation of amino acids in seven groups. In this example as well as with other tested methods it was not possible to distinguish any patterns.

5.4.1.3. Hierarchical clustering

Hierarchical clustering was done with the same candidates together with non-allergen peptides using complete linkage and 50 top nodes and Euclidean distance for peptides represented with ZZ-scales and binary vectors for seven amino acid groups. It was also done with distances calculated from alignment scores and from percent sequence identity. But with all these methods the distribution of candidates and non-allergen peptides was fairly even, ranging from 63% to 38% of each kind in each cluster, so it is not possible to say that clusters particular for allergen candidates have been found.

5.4.1.4. Fishers discriminant function

An attempt to see if the candidates could be separated from the non-allergenic peptides, projecting them on Fishers discriminant function, was done for the same peptide lengths, representing the peptides with ZZ-scales and with binary vectors for seven amino acid groups. No significant separation of the two classes could be seen with either method, as can be seen in Figure 18.

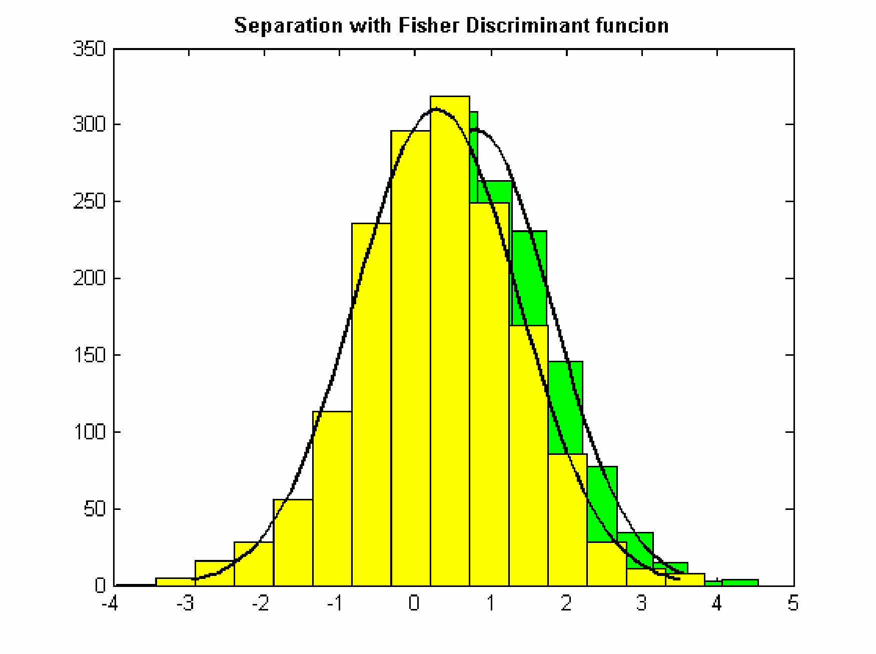


Figure 18. Separation of 24aa candidates (green) and 24aa non-allergen peptides (yellow) represented with ZZ-scales using Fisher discriminant function.

5.4.2. Motif search tools

An attempt to find motifs in the 24aa candidates giving best classification results, as presented in Table 2, was done. The candidates were first clustered with hierarchical clustering using complete linkage and 20 top nodes, distance measure was calculated with alignment scores. This gave clusters with about 50-100 peptides in each. Then three different clusters, clusters number 1 (100 peptides), 10 (59 peptides) and 15 (35 peptides), were input into MEME (81) and Gibbs motif sampler (80).

With Gibbs motif sampler only one motif is found at a time. The motifs obtained with Gibbs were also obtained with MEME, so only the MEME results were further looked at. When mapping the found motifs to their position and comparing if homologues were found in other allergens or in homologous non-allergens there were many hits. Many motifs of the same sort ended up in different clusters when performing hierarchical clustering depending on their position in the peptide and therefore not all peptides with the motifs were found.

5.4.3. Comparing with known motifs

The candidates for best classification using peptide length 12 and 30 were manually mapped to their position in the protein sequence for five allergens, Tri r 2, Bos d 6, Cry j 1, Cry j 2 and Asp f 2. These positions were compared with experimentally mapped epitopes. Results presented in Appendix 4.

A glance at the distribution of the candidates reveals that for Tri r 2 five of eight mapped T-cell epitopeswere covered by the candidates. For Bos d 6 one of three mapped TCE was covered. For Cry j 1 four of ten mapped TCE were covered. For Cry j 2 four of nine TCE were covered. Several of the mapped TCE are covered by the candidate peptides, but there are also several that are not, therfore it is difficult to draw any conclusion as to how well the peptide filtration does at finding T-cell epitopes.

5.5. Conclusion of results

Best classification using S_s/n_p ratios and both non-allergen datasets was with peptide length 24, n_C=5, S_{A1}=2 and S_{A2}=2.5. This gave a pDetection of 0.778 and a pFA of 0.137. Using just one of the non-allergen datasets did not improve these results. Classification with scores from two peptide lengths did not get much better results than using just one peptide length. Classification using highest alignment scores gave pDetection=0.653 and pFA=0.0615. The FAO/WHO method using 6 identical amino acids gave pDetection=0.801 and pFA=0.655.

Peptide filtration with profilins did not give better classification results but the difference between classification with “best” and “worse” candidates may suggest that peptide filtration was better at finding allergen-specific motifs with these datasets.

None of the tested methods for motif identification was able to find any conclusive motifs in the candidates.

6. DISCUSSION

6.1. Classification performance

Using the peptide filtration for classification of allergens gave better results then any previous method described, and with further development the method can probably be improved to render even more certain classification. Best classification was with pDetection 0.778 and pFA 0.137. The previous classification method based on alignments developed at the Swedish National Food Administration gave pDetection 0.70 and pFA 0.11 (91). That method has been evaluated with different test procedures, so the results are not directly comparable with the results in this project. Since more cross-validation tests have been performed there and they still have higher variance, it is safe to say that this method is more stable, but they had to withhold some of the allergens for a prototype set, so they had less allergen data to classify. When comparing to the FAO/WHO recommended method (8), the procedure described here gives somewhat lower detection of allergens (0.778 compared with 0.801), but shows a remarkably lower rate of false alarms (0.137 compared with 0.655).

6.2. Validation of the classifier

It is likely that the validation of the classifications would have given better results with smaller test sets and more times of cross-validation. With larger training sets, more candidates would be selected, and the classification of the test sets would be more accurate since the occurance of another allergen of the same family in the training set would be more likely. The FAO/WHO recommendation of using identical alignment over 6 amino acids were cross-validated in the same manner as the peptide filtration model, so those results are comparable.

6.3. Different peptide lengths for classification

With peptide filtration and classification using S_S/n_P ratios the best classification was rendered with peptide length 24 amino acids. When using classification with highest alignment scores the best classification was obtained with peptide length 18. The reason why longer peptides gave worse classification could be because no gaps were allowed in the alignments. If gaps are introduced it is possible that the classification with longer peptides could be just as efficient. When classifying with longer peptides it becomes more similar to using homology searches, and the similarities will better reflect the three-dimensional structure of the proteins.

But it is also possible that there are specific motifs determining if a protein is allergenic, and that they are in the size range around 18-24 amino acids or even smaller. This will have to be tested further before any such conclusions can be drawn. When doing classification of profilins the results were best with a peptide length of 6 to12 amino acids, while there were too many false positives with larger peptide lengths, since all of the proteins have high homology.

Classification with two peptide lengths was tested with the aim to see if different allergens were detected with short and long peptides. This way the allergens that were missed with one peptide length would be picked up with the other. The results with this type of classification were not very conclusive. The classification was only improved marginally when combining peptide length 30 with 12 or 6. But the way this classifier was set up can probably be improved to get better results.

6.4. Classification with highest alignment scores

The classification using highest alignment scores used here gave virtually no false classification of non-allergens as allergens. Even though the average detection rate was high, 0.78 there was too much variance in the results to make it a reliable method (standard deviation 0.0635). With 95% confidence the classification would still be worse then with the other methods tested in this project. This high variation can possibly be overcome when more allergen sequences are available to make larger allergen training and test sets.

6.5. Different types of classifiers

There are many features of the proteins that could be used for classification. One could combine alignment scores and alignment lengths, as has been used in the previous project at Swedish National Food Administration (91), with features extracted with peptide filtration methods. The features from peptide filtration could be S_S/n_P ratios using different peptide lengths, highest alignment scores from different peptide lengths and maybe the frequency of the peptides in the non-allergen and in the allergen training set. If all these features were combined in feature vectors the most important ones for correct classification could be extracted with data mining tools and different multi-dimensional classification methods could be tested for the feature vectors. But this would require setting up the test and validation methods differently since it would require selecting of candidates from one set, and then obtaining these feature vectors from another two allergen sets, one for training and one for testing. This will be easier to do once more allergenic sequences are available.

6.6. Substitution matrices

For all sequence comparison in this project, alignment scores with BLOSUM80 was used to define similarity, since it was the best substitution matrix in previous alignment methods tested at the Swedish National Food Administration. This was probably a mistake since other matrices such as the PAM30 are described as more suitable for shorter sequence alignments than BLOSUM80 (106). But there was not enough time to rerun all the peptide filtrations; testing of other substitution matrices is an interesting task for the future.

6.7. Influence of non-allergen dataset on results

With some peptide lengths classification using randomly selected candidates was just as good or even better than classification with candidates selected with peptide filtration. The reason why it is possible to get such good classification with randomly selected peptides may be because the allergen data set contains many similar proteins from the same protein families while the non-allergen test set contains all types of proteins. Due to homology between allergens in the test set and in the training set, while the non-allergens in the test set have very low homology to the training set allergens, allergens will be classified in the same class. Possibly this is not, as desired, due to similarity to some allergen specific motif that has been extracted with the candidates.

To get a really good evaluation of the method it should be tested with proteins that are homologous to the allergens but still is not allergenic. As have been tested in small scale when doing peptide filtration with profilins. Naturally the miss-classification of non-allergens would then become higher, but only when we are able to distinguish between allergens and non-allergens in a group of homologous proteins can we truly say that we have found some motifs that are important for the allergenicity of the proteins.

The non-allergen training sets used for peptide filtration were designed to get balanced representation from all protein families. But this distribution was limited by the fact that the protein sequences available in SWALL for the species selected does not cover all possible proteins. With the rice non-allergen set where a very large amount of protein sequences was available, all kinds of proteins in an organism was probably covered, but it still is biased towards proteins only occurring in the organism *Oryza Sativa* (rice), it still does not contain all sorts of proteins in all sorts of organisms.

It could be a good idea to download all proteins available in SWALL as a training set. But with such a big set the peptide filtration step would be very time consuming, therefore one would probably have to sort the peptides and only select the most abundant ones to use for peptide filtration, that way the non-allergen peptides for filtration would give a better representation of the most common peptides among all proteins, and it would non be such a big problem if some allergens happens to be included in the set.

6.8. Classification with profilins

Classification of profilins was not excellent, but it still proved better than random classification, so there must be some features in the candidates selected that is more common in allergenic profilins than in non-allergenic profilins. This could be due to bias in the allergen data set for certain kinds of profilins. Even if all the profilins are from the same group of very homologous proteins there are still some differences. All the allergenic profilins are from plants, but in the non-allergen set there were many mammalian profilins that are a bit different from the plant profilins. This could have an effect on the classification just as the bias in the non-allergen test set discussed in 6.7. There is also the possibility that some of the plant profilins in the non-allergen set are actually allergens, but have not been documented as such. Some of them share very high homology to some of the allergens, and are very likely to be allergens.

The only type of classification used for the profilins was with S_S/n_P ratios. Using highest alignment scores should be tested here as well and possibly other classification methods to see if it is possible to get better results. Even good classification using only profilins is difficult, it might still be a good method for finding allergen-specific motifs. The classification with "best" candidates and "worst" candidates differed more for the profilins than with the other datasets, this could indicate that the peptide filtration was actually better at finding allergen-specific peptides. If the non-allergens could be certified, multiple alignments, peptide filtration or other methods could be used to determine where the major differences between allergens and non-allergens are.

When looking at the position of the 6 amino acid candidates in these profilins, it was hard to see any particular pattern, but it should be further investigated with other methods to determine if there is some specific motif in the candidates, and if the candidates could be chosen in a different manner for better classification and motif finding. It would also be interesting to look at allergens from some other protein families together with their non-allergenic homologues to see if some motifs can be distinguished.

6.9. Motif searches

Whith the methods tested so far it has not been possible to find any specific motifs that are only present in allergens. It could be because the peptide filtration is not good enough at selecting the best candidates or because the methods applied for finding the motifs in the candidates were not optimal. It is also impractical to generalise the motifs to one peptide length at a time since different motifs might have different lengths.

New methods for motif searches, similar to the algorithms used in MEME should be designed more adapted to finding motifs in short peptides, and which allow the represenation of all candidates at once. This way clustering before motif searches would not be necessary and distribution of motifs in different clusters could be avoided.

When doing peptide filtration and classification the candidates that give the best classification are not necessarily the ones that are best for finding motifs. In this project classification performance has been used to evaluate how well the candidates are selected. But if motif searches should be evaluated properly there must be some experimental methods to determine their importance in the pathways leading to allergic diseases.

6.10. Future work

There are many improvements to this project to make in the future. Improved non-allergen and allergen sets must be established to obtain better results. New allergens are being identified and new sequences are continuously being added in the databases that will have to be included in the allergen-sets. Different methods for establishing non-allergen datasets, for example homology searches or using all available proteins in SWALL, should be tested. The methods should also be evaluated with other substitution matrices and other methods for classification could be tried. More work needs to be done on specific allergen groups, similar to the work that has been done with profilins.

7. ACKNOWLEDGEMENTS

I would like to thank my supervisors Ulf Hammerling, Daniel Soeria-Atmadja and Mats Gustafsson for their inputs and support during this work. I would also like to thank Tomas Olofsson for examining the report. A big “thank you” to all the staff at the Department of Toxicology at the Swedish National Food Agency for providing a friendly working atmosphere. The computer calculations performed were sponsored by a research grant form Göran Gustafssons stiftelse.

8. REFERENCES

1. Kay AB. Overview of 'allergy and allergic diseases: with a view to the future'. Br Med Bull 2000;56(4):843-64

2. Feijen M, Gerritsen J, Postma DS. Genetics of allergic disease. Br Med Bull 2000;56(4):894-907

3. Albersee RC. Structural biology of allergens. J Allergy Clin Immunol 2000;106:228-238

4. Bredehorst R, David K. What establishes a protein as an allergen? J Chromatogr B Biomed Sci Appl 2001 May 25;756(1-2):33-40

5. Huby RD, Dearman RJ, Kimber I. Why are some proteins allergens? Toxicol Sci 2000 Jun;55(2):235-46

6. Aalberse RC, Stapel SO. Structure of food allergens in relation to allergenicity. Pediatr Allergy Immunol 2001;12 Suppl 14:10-4

7. Breiteneder H, Ebner C. Atopic allergens of plant foods. Curr Opin Allergy Clin Immunol 2001 Jun;1(3):261-7

8. FAO/WHO Expert Consultation on ”Evaluation of Allergenicity of Genetically Modified Foods”. Food and Agricultural Organization or the United Nations, Rome, Italy, 22-25 January 2001

9. Campbell NA, Reese JB, Biology 4th ed. Chapter 39, The Body's Defences, Benjamin Cummings Publishers 1996

10. Roitt I, Brostoff J, Male D, Immunology, 4th ed. Times Mirror International Publishers Limited 1996

11. Fairchild PJ. Presentation of antigenic peptides by products of the major histocompatibility complex. J Pept Sci 1998 May;4(3):182-94

12. Kimballs Biology Pages, http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/ (17 April 2003)

13. Bishop GA, Hostager BS. B lymphocyte activation by contact-mediated interactions with T lymphocytes. Curr Opin Immunol 2001 Jun;13(3):278-85

14. von Pirquet C. Allergie. Münch med Wochenstr 1906; 30: 1457 (Translated from the German by Prausnitz C. In: Gell PGH, Coombs RRA. (eds) Clinical Aspects of Immunology. Oxford: Blackwell Scientific, 1963)

15. Gould HJ, Sutton BJ, Beavil AJ, Beavil RL, McCloskey N, Coker HA, Fear D, Smurthwaite L. The biology of ige and the basis of allergic disease. Annu Rev Immunol 2003;21:579-628

16. van Neerven RJ. The role of allergen-specific T cells in the allergic immune response: relevance to allergy vaccination. Allergy 1999 Jun;54(6):552-61

17. Kimber I, Dearman RJ. Factors affecting the development of food allergy. Proc Nutr Soc 2002 Nov;61(4):435-9

18. Van Bever HP. Early events in atopy. Eur J Pediatr 2002 Oct;161(10):542-6

19. Warner JA, Warner JO. Early life events in allergic sensitisation. Br Med Bull 2000;56(4):883-93

20. Greenberger PA. Immunotherapy update: mechanisms of action. Allergy Asthma Proc 2002 Nov-Dec;23(6):373-6

21. Larche M. Anti-T-cell strategies in the treatment of allergic disease. Allergy 2002;57 Suppl 72:20-3

22. Krieg AM. CpG motifs in bacterial DNA and their immune effects. Annu Rev Immunol 2002;20:709-60

23. Allsopp CE, Plebanski M, Gilbert S, Sinden RE, Harris S, Frankel G, Dougan G, Hioe C, Nixon D, Paoletti E, Layton G, Hill AV. Comparison of numerous delivery systems for the induction of cytotoxic T lymphocytes by immunization. Eur J Immunol 1996 Aug;26(8):1951-9

24. MacGlashan D Jr. Anti-IgE antibody therapy. Clin Allergy Immunol 2002;16:519-32

25. Astwood JD, Leach JN, Fuchs RL. Stability of food allergens to digestion in vitro. Nat Biotechnol 1996 Oct;14(10):1269-73

26. Sanchez-Monge R, Blanco C, Perales AD, Collada C, Carrillo T, Aragoncillo C, Salcedo G. Class I chitinases, the panallergens responsible for the latex-fruit syndrome, are induced by ethylene treatment and inactivated by heating. J Allergy Clin Immunol 2000 Jul;106(1 Pt 1):190-5

27. Soler-Rivas C, Wichers HJ. Impact of (bio)chemical and physical procedures on food allergen stability. Allergy 2001;56 Suppl 67:52-5

28. Davis PJ, Smales CM, James DC. How can thermal processing modify the antigenicity of proteins? Allergy 2001;56 Suppl 67:56-60

29. Ortolani C. Ispano M, Pastorello E, Bigi A, Anasaloni R The oral allergy syndrome. Ann Allergy 1988; 61(6 Pt 2): 47-52

30. Urisu A, Ando H, Morita Y, Wada E, Yasaki T, Yamada K, Komada K, Torii S, Goto M, Wakamatsu T. Allergenic activity of heated and ovomucoid-depleted egg white. J Allergy Clin Immunol 1997 Aug;100(2):171-6

31. Brenna O, Pompei C, Ortolani C, Pravettoni V, Farioli L, Pastorello EA. Technological processes to decrease the allergenicity of peach juice and nectar. J Agric Food Chem 2000 Feb;48(2):493-7

32. Malanin K, Lundberg M, Johansson SG. Anaphylactic reaction caused by neoallergens in heated pecan nut. Allergy 1995 Dec;50(12):988-91

33. Rosen JP, Selcow JE, Mendelson LM, Grodofsky MP, Factor JM, Sampson HA. Skin testing with natural foods in patients suspected of having food allergies: is it a necessity? J Allergy Clin Immunol 1994 Jun;93(6):1068-70

34. Maleki SJ, Chung SY, Champagne ET, Raufman JP. The effects of roasting on the allergenic properties of peanut proteins. J Allergy Clin Immunol 2000 Oct;106(4):763-8

35. Berrens L. Neoallergens in heated pecan nut: products of Maillard-type degradation? Allergy 1996 Apr;51(4):277-8

36. Banerjee B, Greenberger PA, Fink JN, Kurup VP. Conformational and linear B-cell epitopes of Asp f 2, a major allergen of Aspergillus fumigatus, bind differently to immunoglobulin E antibody in the sera of allergic bronchopulmonary aspergillosis patients. Infect Immun 1999 May;67(5):2284-91

37. Beezhold DH, Hickey VL, Slater JE, Sussman GL. Human IgE-binding epitopes of the latex allergen Hev b 5. J Allergy Clin Immunol 1999 Jun;103(6):1166-72

38. Aalberse RC, Akkerdaas J, van Ree R. Cross-reactivity of IgE antibodies to allergens. Allergy 2001 Jun;56(6):478-90

39. Spitzauer S. Allergy to mammalian proteins: at the borderline between foreign and self? Int Arch Allergy Immunol 1999 Dec;120(4):259-69

40. Van Ree R, Driessen MN, Van Leeuwen WA, Stapel SO, Aalberse RC. Variability of crossreactivity of IgE antibodies to group I and V allergens in eight grass pollen species. Clin Exp Allergy 1992 Jun;22(6):611-7

41. van Neerven RJ, Larsen JN, Wissenbach M, Würtzen PA. T cell responses to pollen allergens – Reactivity patterns and clinical relvance. ACI International, 11/2 1999: 37-42

42. Schirle M, Weinschenk T, Stevanovic S. Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. J Immunol Methods 2001 Nov 1;257(1-2):1-16

43. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. Bioinformatics 2001 Dec;17(12):1236-7

44. Borras-Cuesta F, Golvano J, Garcia-Granero M, Sarobe P, Riezu-Boj J, Huarte E, Lasarte J. Specific and general HLA-DR binding motifs: comparison of algorithms. Hum Immunol 2000 Mar;61(3):266-78

45. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 1999 Nov;50(3-4):213-9

46. Davenport MP, Ho Shon IA, Hill AV. An empirical method for the prediction of T-cell epitopes. Immunogenetics 1995;42(5):392-7

47. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics 1998;14(2):121-30

48. Mallios RR. Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. Bioinformatics 1999 Jun;15(6):432-9

49. Mallios RR. Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. Bioinformatics 2001 Oct;17(10):942-8

50. Holzhutter HG, Frommel C, Kloetzel PM. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. J Mol Biol 1999 Mar 5;286(4):1251-65

51. Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. Protein Eng 2002 Apr;15(4):287-96

52. Nussbaum AK, Kuttler C, Hadeler KP, Rammensee HG, Schild H. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. Immunogenetics 2001 Mar;53(2):87-94

53. Lack G, Chapman M, Kalsheker N, King V, Robinson C, Venables K; BSACI working party. Report on the potential allergenicity of genetically modified organisms and their products. Clin Exp Allergy 2002 Aug;32(8):1131-43

54. Petersen A, Schramm G, Schlaak M, Becker WM. Post-translational modifications influence IgE reactivity to the major allergen Phl p 1 of timothy grass pollen. Clin Exp Allergy 1998 Mar;28(3):315-21

55. Fotisch K, Altmann F, Haustein D, Vieths S. Involvement of carbohydrate epitopes in the IgE response of celery-allergic patients. Int Arch Allergy Immunol 1999 Sep;120(1):30-42

56. Garcia-Casado G, Sanchez-Monge R, Chrispeels MJ, Armentia A, Salcedo G, Gomez L. Role of complex asparagine-linked glycans in the allergenicity of plant. Glycobiology 1996 Jun;6(4):471-7

57. Longoni D, Piemonti L, Bernasconi S, Mantovani A, Allavena P. Interleukin-10 increases mannose receptor expression and endocytic activity in monocyte-derived dendritic cells. Int J Clin Lab Res 1998;28(3):162-9

58. Pomes A. Intrinsic properties of allergens and environmental exposure as determinants of allergenicity. Allergy 2002 Aug;57(8):673-9

59. Gough L, Schulz O, Sewell HF, Shakib F. The cysteine protease activity of the major dust mite allergen Der p 1 selectively enhances the immunoglobulin E antibody response. J Exp Med 1999 Dec 20;190(12):1897-902

60. Tomee JF, van Weissenbruch R, de Monchy JG, Kauffman HF. Interactions between inhalant allergen extracts and airway epithelial cells: effect on cytokine production and cell detachment. J Allergy Clin Immunol 1998 Jul;102(1):75-85

61. Ivanciuc O, Schein CH, Braun W. Data mining of sequences and 3D structures of allergenic proteins. Bioinformatics 2002 Oct;18(10):1358-64

62. Roos DS. Computational biology. Bioinformatics--trying to swim in a sea of data. Science 2001 Feb 16;291(5507):1260-1

63. Goodman N. Biological data becomes computer literate: new advances in bioinformatics.Curr Opin Biotechnol 2002 Feb;13(1):68-71

64. Needleman SB, Wunch CD. J. Mol. Biol. 1970;48:443

65. Smith TF, Waterman MS, Fitch WM. J. Mol. Evol. 1981;18:38-46

66. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 2000;132:185-219

67. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997 Sep 1;25(17):3389-402

68. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994 Nov 11;22(22):4673-80

69. Baldi P Brunak S. Bioinformatics-The Machine Learning Approach, 2nd edition, A Bradford Book, The MIT Press, Cambridge, England, 2001

70. Vining DJ, Gladish GW. Receiver operating characteristic curves: a basic understanding. Radiographics 1992 Nov;12(6):1147-54

71. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000 Dec 22;290(5500):2319-23

72. Kohonen T. Self-Organizing Maps. Springer. Berlin. 1997

73. Kaminski N. Bioinformatics. A user's perspective. Am J Respir Cell Mol Biol 2000 Dec;23(6):705-11

74. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002 Sep;3(3):265-74

75. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002 Jan 1;30(1):276-80

76. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003 Jan 1;31(1):400-2

77. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers, Nucl. Acids Res. 2000; 28:228-230

78. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 1995 Oct 3;163(2):GC17-26

79. Smith HO, Annau TM, Chandrasegaran S. Finding sequence motifs in groups of functionally related proteins. Proc Natl Acad Sci U S A 1990 Jan;87(2):826-30

80. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Science 1993 Oct 8;262(5131):208-14

81. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994

82. Taylor SL, Helfe SL, Will genetically modified foods be allergenic? J Allergy Clin Immunol. 2001 May;107(5):765-71. Review

83. Redenbaugh K, Hiatt W, Martineau B, et al. Safety assessment of genetically engineered fruits and vegetables. A case study of the Flavr Savr tomato. Boca Raton (FL): CRC Press, Inc; 1992. p. 267

84. Underwood BA. Vitamin A in human nutrition: public health considerations. In: Sporn MB, Roberts AB, Goodman DS, eds. The retinoids: biology, chemistry and medicine. 2nd ed. New York: Raven Press; 1994. p. 217-27

85. Flanders Interuniversity Institute for Biotechnology, “Safety of Genetically Engineered Crops” March 2001 VIB Publication

86. Metcalfe DD, Astwood JD, Townsend R, Sampson HA, Taylor SL, Fuchs RL. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. Crit Rev Food Sci Nutr. 1996;36 Suppl:S165-86. Review

87. Hileman RE, Silvanovich A, Goodman RE, Rice EA, Holleschak G, Astwood JD, Hefle SL. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int Arch Allergy Immunol 2002 Aug;128(4):280-91

88. Gendel SM. Sequence analysis for assessing potential allergenicity. Ann N Y Acad Sci 2002 May;964:87-98

89. Kleter GA, Peijnenburg AA. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE - binding linear epitopes of allergens. BMC Struct Biol 2002 Dec 12;2(1):8

90. Zorzet A, Gustafsson M, Hammerling U. Prediction of food protein allergenicity: a bioinformatic learning systems approach. In Silico Biol 2002;2(4):525-34

91. Soeria-Atmadja D, Zorzet A, Gustafsson M, Hammerling U. Statistical evaluation of classification methods for potential allergenicity based on local alignment. May 2003 (In press)

92. The FARRP Protein Allergen Database, http://www.allergenonline.com/default.asp (12 february 2003)

93. The Allergen Database, The Central Science Laboratory is an Executive Agency of the Department for Environment Food & Rural Affairs, http://www.csl.gov.uk/allergen/ (12 february 2003)

94. The Allergen Sequence Database, The National Center for Food Safety and Technology, http://www.iit.edu/~sgendel/fa.htm (12 february 2003)

95. The ProtAll database for food allergens of plant origin, The British Institute of Food Research, http://www.ifrn.bbsrc.ac.uk/protall/database.html (12 february 2003)

96. Allergen Nomenclature, the IUIS list of allergens, http://www.allergen.org/List.htm (12 february 2003)

97. SWISSPROT www sequence retrieval service at the Swiss Institute of Bioinfromatics(SIB),http://us.expasy.org/sprot/, 2003-03-04

98. Venkatarajan MS, Braun W, New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties, J. Mol. Model. 2001, 7:445-453

99. Woodfolk JA, Sung SJ, Benjamin DC, Lee JK, Platts-Mills TAE. Distinct Human T Cell Repertoires Mediate Immediate and Delayed-Type Hypersensitivity to the Trichophyton Antigen, Tri r 21 The Journal of Immunology, 2000, 165: 4379-4387.

100. Tanabe S, Kobayashi Y, Takahata Y, Morimatsu F, Shibata R, Nishimura T. Some human B and T cell epitopes of bovine serum albumin, the major beef allergen. Biochem Biophys Res Commun. 2002 May 24;293(5):1348-53

101. Ohno N, Ide T, Sakaguchi M, Inouye S, Saito S. Common antigenicity between Japanese cedar (Cryptomeria japonica) pollen and Japanese cypress (Chamaecyparis obtusa) pollen, II. Immunology. 2000 Apr;99(4):630-4

102. Sone T, Morikubo K, Miyahara M, Komiyama N, Shimizu K, Tsunoo H, Kino K. T cell epitopes in Japanese cedar (Cryptomeria japonica) pollen allergens: choice of major T cell

epitopes in Cry j 1 and Cry j 2 toward design of the peptide-based immunotherapeutics for the management of Japanese cedar pollinosis. J Immunol. 1998 Jul 1;161(1):448-57

103. Tamura Y, Kawaguchi J, Serizawa N, Hirahara K, Shiraishi A, Nigi H, Taniguchi Y, Toda M, Inouye S, Takemori T, Sakaguchi M, Analysis of sequential immunoglobulin E-binding epitope of Japanese cedar pollen allergen (Cry j 2) in humans, monkeys and mice, Clin. Exp. Allergy 2003, 33, 211-217

104. Banerjee B, Greenberger PA, Fink JN, Kurup VP, Conformational and linear B-cell epitopes of Asp f 2, a major allergen of *Aspergillus fumigatus*, bind differently to immunoglobulin E antibody in the sera of allergic bronchopulmonary aspergillosis patients, Infect. Immun. 1999, *67*, 2284-2291

105. Svirshchevskaya EV, Alekseeva L, Marchenko A, Viskova N, Andronova TM, Benevolenskii SV, Kurup VP. Immune response modulation by recombinant peptides expressed in virus-like particles. Clin Exp Immunol. 2002 Feb;127(2):199-205

106. The Joint Center for Structural Genomics BLAST information pages, http://www.jcsg.org/blast/docs/matrix_info.html, (25 May 2003)

List of useful Internet resources

All web-page addresses accurate in May 2003

a. Genbank, NCBI (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
b. EMBL, the European Bioinformatics Institute (http://www.ebi.ac.uk/)
c. The DNA Database of Japan (http://www.ddbj.nig.ac.jp/fromddbj-e.html)
d. SWISS-SPROT (http://www.ebi.ac.uk/swissprot/)
e. PDB (Protein Data Bank) at RCSB (http://www.rcsb.org/pdb/)
f. ExPASy (Expert Protein Analysis System), the Swiss Institute of Bioinformatics (SIB) (http://us.expasy.org/)
g. MHCPEP (http://wehih.wehi.edu.au/mhcpep/)
h. SYFPEITHI (http://syfpeithi.bmi-heidelberg.com/)
i. FIMM (http://sdmc.krdl.org.sg:8080/fimm)
j. (http://immuno.bme.nwu.edu/)
k. IMGT (http://www.ebi.ac.uk/imgt/)
l. HIV Molecular (http://hiv-web.lanl.gov/immunology/)
m. Allergen Nomenclature List, IUIS, (http://www.allergen.org/List.htm)
n. the Allergome database (http://www.allergome.org/)
o. The Allergen Database (http://www.csl.gov.uk/allergen/)
p. The Allergen Sequence Database (http://www.iit.edu/~sgendel/fa.htm)
q. ProtAll (http://www.ifrn.bbsrc.ac.uk/protall/database.html)
r. The FARRP Protein Allergen Database (http://www.allergenonline.com/default.asp)

APPENDIX 1- ABBREVIATIONS USED IN THIS REPORT

A list with some of the abbreviations frequently used in this report:

BCR	B-cells receptor
TCR	T-cells receptor
MHC	Major histocompatibility complex
HLA	Human leukocyte antigen (name of MHC in humans)
TCE	T-cells epitope
T_H	T-helper cell
IL	Interleukin (a signal molecule)
IFN	Interferon (a signal molecule)
APC	Antigen Presenting Cell
LPR	Late phase reactions (in allergy)
SIT	Specific immunotherapy
ROC	Reciever operating characteristics
PCA	Principal component analysis
GMO	Genetically modified organism
IFBC	International Food Biotechnology Council
ILSI	International Life Science Institute
FAO	Food and Agriculture Organinsation of the United Nations
WHO	World Health Organisation
STD	Standard Deviation

APPENDIX 2 – SELECTION OF THE NON-ALLERGEN DATA SET

When selecting the non-allergen datasets searches in SWALL(67) were done using the following search criteria:

Set 1, vegetables

- organism : *Lycopersicon* (tomato), *Malus* (apple), *Prunus* (peach, cherry, apricot), *Spinacia oleracea* (Spinach) or *Daucus carota* (Carrot).
- sequence length: NOT 0:50
- all text: NOT allergy, atopy or allergen
- Gave 2965 sequences

Set 1, animals

- Milk fraction
  - organism: *Bos Taurus* (cow)
  - sequence length: NOT 0:50
  - all text: NOT allergy, atopy or allergen
  - all text: milk, casein, lactalbumin, lactoferrin, proteose-peptone, lactoperoxidase or “xanthine dehydrogenase”
  - Gave 54 sequences
- Egg fraction
  - organism: *Gallus Gallus* (chicken)
  - sequence length: NOT 0:50
  - all text: NOT allergy, atopy or allergen
  - all text: egg
  - Gave 11 sequences
- Salmon fraction
  - Organism: *Salmo salar* (salmon)
  - sequence length: NOT 0:50
  - all text: NOT allergy, atopy or allergen
  - Gave 313 sequences
- Cod fraction
  - Organism: *Gadus* (cod)
  - sequence length: NOT 0:50
  - all text: NOT allergy, atopy or allergen
  - Gave 326 sequences

in total set 1 contains 3370 sequences

Set 2

- organism: *Oryza sativa* (Rice).
- sequence length: NOT 0:50
- all text: NOT allergy, atopy or allergen
- Gave 18812 sequences

The number of sequences available in SWALL from each species, and the number of sequences retrieved to our database:

Species	**No proteins in SWALL**	**No sequences to data set**
Gadus * (cod)	350	326
Salmo salar (salmon)	334	313
Gallus Gallus (chicken)	3015	11
Bos Taurus (cow)	3500	54
Lycopersicon (tomato)	1309	1262
Malus (apple)	325	296
Prunus (peach, cherry, apricot)	595	570
Spinacia oleracea (Spinach).	422	316
Daucus carota (Carrot).	234	222

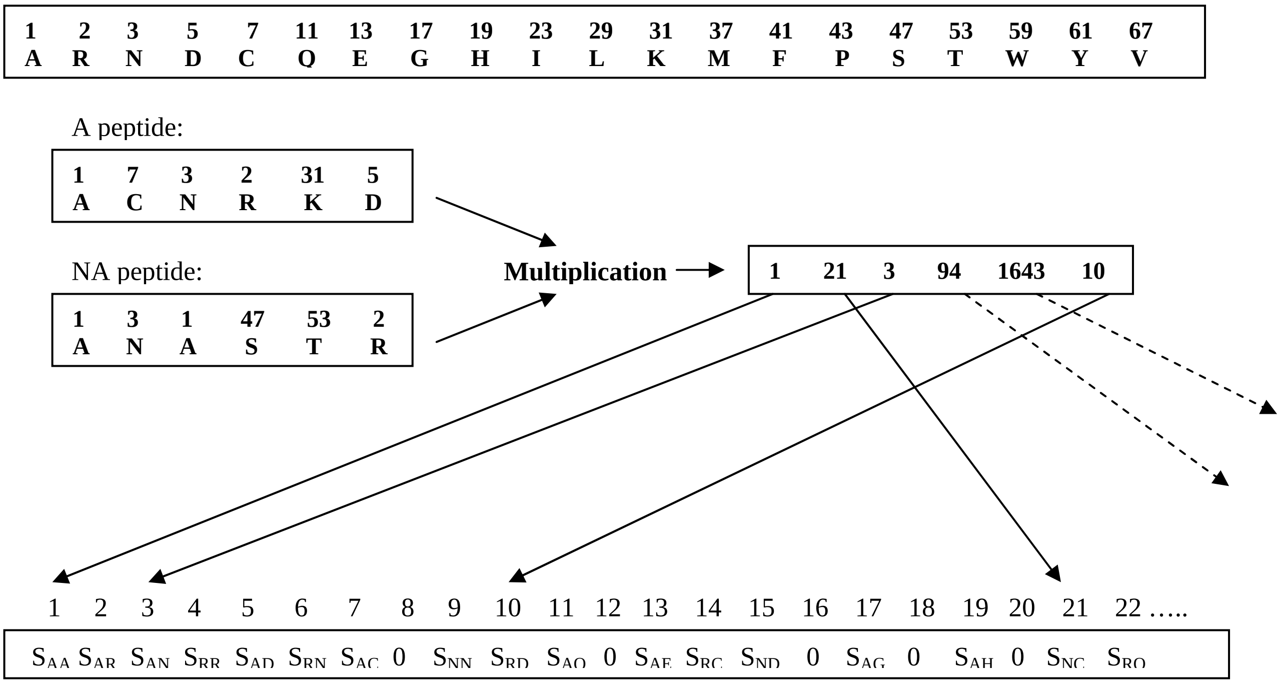
**Gadus* gave: *Gadus morhua* (Atlantic cod) mostly and a few from *Gadus callarias* (Baltic cod).

APPENDIX 3 - IMPLEMENTATION OF SEQUENCE COMPARISONS

Amino acid sequences were represented with prime numbers to enable faster implementation of sequence comparisons without using for-loops. Each allergenic peptide could thereby be comopared simultaneously with a set of non-allergenic peptides.

Calculation of alignment scores between two peptides has been demonstrated in the figure below. First the prime number representation of peptides were multiplied rendering unique numbers, each number indicating the substitution between two specific amino acids. The substitution matrix was reassembled as a vector with the scores for substitutions at positions with indices representing these unique numbers. This enabled calculation of the score for each alignment by adding the scores retrieved from this vector.

Transformation from amino acid code into prime number code:



1	**2**	**3**	**5**	**7**	**11**	**13**	**17**	**19**	**23**	**29**	**31**	**37**	**41**	**43**	**47**	**53**	**59**	**61**	**67**
A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

A peptide:

1	**7**	**3**	**2**	**31**	**5**
A	**C**	**N**	**R**	**K**	**D**

NA peptide:

1	**3**	**1**	**47**	**53**	**2**
A	**N**	**A**	**S**	**T**	**R**

1	**21**	**3**	**94**	**1643**	**10**

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22 …..
S_{AA}	S_{AR}	S_{AN}	S_{RR}	S_{AD}	S_{RN}	S_{AC}	0	S_{NN}	S_{RD}	S_{AQ}	0	S_{AE}	S_{RC}	S_{ND}	0	S_{AG}	0	S_{AH}	0	S_{NC}	S_{RQ}

Substitution matrix in vector form, where the positions representing multiplication of two prime numbers, contain the substitution scores for the substitution of two amino acids.

Alignment score (S_A) for comparison of the two peptides above is:
$S_{AA}+ S_{NC}+ S_{AN}+ S_{RS}+ S_{KT}+ S_{RD}=\mathbf{S_A}$

When comparing one allergenic peptide with several non-allergenic peptides the multiplication described above was done with the NA-peptides in one matrix and several copies of the A-peptide in another and the two matrices were multiplied. Then the substitution scores were retrieved, from the substitution vector above, for all multiplications simulatneously.

This improved the speed of the calculations with a tenfold as compared to alignment using for-loops and regular substitution matrices. The method was developed together with Mats Gustafsson.

APPENDIX 4 – COMPARING MOTIFS AND CANDIDATES

Asp f 2,Swiss prot name: ALL2_ASPFU

```
MAALLRLAVL LPLAAPLVAT LPTSPVPIAA RATPHEPVFF SWDAGAVTSF PIHSSCNATQ

                                                     F PIHSSCNATQ-

RRQIEAGLNE AVELARHAKA HILRWGNESE IYRKYFGNRP TMEAVGAYDV IVNGDKANVL
                               E IYRKYFGNRP  MEAVGAYDV IVN
R   EAGLNE AVELAR
RRQIEAGLNE AVELARHAKA HILRWGNESE
                RHAKA HILRWGNESE IYRKYFGNRP TMEAV         GDKANVL-

FRCDNPDGNC ALEGWGGHWR GANATSETVI CDRSYTTRRW LVSMCSQGYT VAGSETNTFW
                  HWR                YTTRR
                                      TTRRW LVSMCSQ       SETNTFW-
                                        RRW LVSMCSQGY       TNTFW-
FRCDNPDGNC ALEGWGGHWR GAN         DRSYTTRRW LVSMCSQGYT VAGSETNTFW-

ASDLMHRLYH VPAVGQGWVD HFADGYDEVI ALAKSNGTES THDSEAFEYF ALEAYAFDIA
ASDLM   YH VP       D HFAD                             ALEAYA
ASDLM         VGQGWVD HFADG
ASDLMHR
     HRLYH VPAVGQG
A

APGVGCAGES HGPDQGHDTG SASAPASTST SSSSSGSGSG ATTTPTDSPS ATIDVPSNCH
                                       SGSG ATTTPTDSPS A
                                       SGSG ATTTPTDS
                                              TTPTDSPS ATID
                           ASTST SSSSSGSGSG ATTTPTDSPS ATIDV

THEGGQLHCT
```

Orange: Known IgE epitopes (104)

ATQRRQI
RKYFG
HWR
YTTRR
ASDLM
YHVP
DHFAD
ALEAYA
THEGGQ

Green: Known T-cell epitopes (mapped in mouse) (105)

MEAVGAYDVIVN
SGSGATTTPTDSPSA
EIYRKYFGNRP

Red: 30aa candidates

GDKANVLFRCDNPDGNCALEGWGGHWRGAN
DRSYTTRRWLVSMCSQGYTVAGSETNTFWA
ASTSTSSSSSGSGSGATTTPTDSPSATIDV
RRQIEAGLNEAVELARHAKAHILRWGNESE
RHAKAHILRWGNESEIYRKYFGNRPTMEA

Blue: 12aa candidates

RRWLVSMCSQGY
SGSGATTTPTDS
TTPTDSPSATID
TTRRWLVSMCSQ
SETNTFWASDLM
HRLYHVPAVGQG
VGQGWVDHFADG
FPIHSSCNATQR
TNTFWASDLMHR
EAGLNEAVELAR

```
Bos d 6, Swiss prot name:  ALBU_BOVIN

MKWVTFISLL LLFSSAYSRG VFRRDTHKSE IAHRFKDLGE EHFKGLVLIA FSQYLQQCPF

DEHVKLVNEL TEFAKTCVAD ESHAGCEKSL HTLFGDELCK VASLRETYGD MADCCEKQEP
                                                RETYGD MADCCEKQEP-

ERNECFLSHK DDSPDLPKLK PDPNTLCDEF KADEKKFWGK YLYEIARRHP YFYAPELLYY
           DDSPDLPKLK PDPNTLC
ERNECFLSHK DDSP

ANKYNGVFQE CCQAEDKGAC LLPKIETMRE KVLASSARQR LRCASIQKFG ERALKAWSVA

ANKYNGVFQE CC                                           RALKAWSVA-

RLSQKFPKAE FVEVTKLVTD LTKVHKECCH GDLLECADDR ADLAKYICDN QDTISSKLKE

RLSQKFPKAE FVEVTKLVTD L                                    SSKLKE-
       KAE FVEVTKLVT                          LAKYICDN QDTI

CCDKPLLEKS HCIAEVEKDA IPENLPPLTA DFAEDKDVCK NYQEAKDAFL GSFLYEYSRR

CCDKPLLEKS HCIAEVEKDA IPEN  PLTA DFAEDKDVCK NYQEAKDAFL GSFLYE
           HCIAEVEKDA I                DVCK NYQEAKDA

HPEYAVSVLL RLAKEYEATL EECCAKDDPH ACYSTVFDKL KHLVDEPQNL IKQNCDQFEK
                              PH ACYSTVFDKL KHLVDEP
                                       FDKL KHLVDEPQ

LGEYGFQNAL IVRYTRKVPQ VSTPTLVEVS RSLGKVGTRC CTKPESERMP CTEDYLSLIL
                                                            LSLIL-
               TRKVPQ VSTPTL

NRLCVLHEKT PVSEKVTKCC TESLVNRRPC FSALTPDETY VPKAFDEKLF TFHADICTLP
NRLC
            VSEKVTKCC TES                        DEKLF TFHADICTLP
DTEKQIKKQT ALVELLKHKP KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPKLVV

DTEKQIKKQT ALVEL                          K CCAADDKEAC F
            LVELLKHKP KAT
STQTALA
```

Green: Known T-cell epitopes (100)
DDSPDLPKLKPDPNTLC
PHACYTSVFDKLKHLVDEP
LSLILNRLC

Red: 30aa candidates
DEKLFTFHADICTLPDTEKQIKKQTALVEL
RALKAWSVARLSQKFPKAEFVEVTKLVTDL
SSKLKECCDKPLLEKSHCIAEVEKDAIPEN
PLTADFAEDKDVCKNYQEAKDAFLGSFLYE
RETYGDMADCCEKQEPERNECFLSHKDDSP

Blue: 12aa candidates
DVCKNYQEAKDA
KCCAADDKEACF
SHCIAEVEKDAI
VSEKVTKCCTES
LVELLKHKPKAT
FDKLKHLVDEPQ
LAKYICDNQDTI
TRKVPQVSTPTL
ANKYNGVFQECC
KAEFVEVTKL

Cry j 1, Swiss prot name: **SBP_CRYJA**

```
MDSPCLVALL VLSFVIGSCF SDNPIDSCWR GDSNWAQNRM KLADCAVGFG SSTMGGKGGD
                                      QNRM KLADCAVGFG S
       ALL VLSFVIGSCF SDNPIDSCWR GDSNWAQ
                       DNPIDSCWR GDS                FG SSTMGGKGGD
                                 GDSNWAQNRM KL

LYTVTNSDDD PVNPAPGTLR YGATRDRPLW IIFSGNMNIK LKMPMYIAGY KTFDGRGAQV
                            RPLW IIFSGNMNIK LKMPMYIAGY KTFDGR
                PGTLR YGATRDRPLW IIFSGNMNIK LKMPM       TFDGRGAQV-
                                    SGNMNIK LKMPM
                                       MNIK LKMPMYIA

YIGNGGPCVF IKRVSNVIIH GLHLYGCSTS VLGNVLINES FGVEPVHPQD GDALTLRTAT
YIGNGG      KRVSNVIIH GLHLYG
      PCVF IKRVSNVIIH G
            KRVSNVIIH GLHLYGCSTS V                            TAT-

NIWIDHNSFS NSSDGLVDVT LSSTGVTISN NLFFNHHKVM LLGHDDAYSD DKSMKVTVAF
                                                        KSMKVTVAF-
                                                        KSMKVTVAF-
NIWIDHNSFS NSSDGLVDVT LSSTGVT                                  AF-
                                                DDAYSD DKSMKVTVAF-

NQFGPNCGQR MPRARYGLVH VANNNYDPWT IYAIGGSSNP TILSEGNSFT APNESYKKQV
NQFGPN
NQFGPNCGQR M
NQFGPNCGQR
NQFGPNCGQR MPRA        ANNNYDPWT IYAIGGSSNP TILSEGNSFTA

TIRIGCKTSS SCSNWVWQST QDVFYNGAYF VSSGKYEGGN IYTKKEAFNV ENGNATPQLT
                                 VSSGKYEGGN IYTKKEAFNV E
                                  SSGKYEGGN IYTKKEAFNV E
            CSNWVWQST QDV                             V ENGNATPQLT-
                                                        NGNATPQLT-
                                                           ATPQLT-
KNAGVLTCSL SKRC
K
KNA
KNAGVL
```

Green: Known T-cell epitopes (mouse) Ohno *et al.*(101)

KRVSNVIIHGLHLYGCSTSV
KSMKVTVAFNQFGPNCGQRM
SSGKYEGGNIYTKKEATNVE

Turquoise. Known T-cell epitopes (human), Sone *et al.* (102)

QNRMKLADCAVGFGS
KMPMYIAGYKTFDGR
PCVFIKRVSNVIIHG
KRVSNVIIHGLHLYG
KSMKVTVAFNQFGPN
DVFYNGAYFVSSGKY
RPLWIIFSGNMNIKL
TFDGRGAQVYIGNGG
EGGNIYTKKEAFNVE

30aa candidates

ALLVLSFVIGSCFSDNPIDSCWRGDSNWAQ
PGTLRYGATRDRPLWIIFSGNMNIKLKMPM
TATNIWIDHNSFSNSSDGLVDVTLSSTGVT
DDAYSDDKSMKVTVAFNQFGPNCGQRMPRA
ANNNYDPWTIYAIGGSSNPTILSEGNSFTA

Blue: 12 aa candidates

ATPQLTKNAGVL
DNPIDSCWRGDS
GDSNWAQNRMKL
FGSSTMGGKGGD
CSNWVWQSTQDV
SGNMNIKLKMPM
MNIKLKMPMYIA
AFNQFGPNCGQR
VENGNATPQLTK
NGNATPQLTKNA

Cry j 2, Swiss prot name: MPA2_CRYJA

```
MAMKFIAPMA FVAMQLIIMA AAEDQSAQIM LDSDIEQYLR SNRSLRKVEH SRHDAINIFN
              MQLIIMA AAEDQSAQIM LDSDIEQYLR SNR

VEKYGAVGDG KHDCTEAFST AWQAACKKPS AMLLVPGNKK FVVNNLFFNG PCQPHFTFKV
                                      PGNKK FVVNNLFFNG PCQPHF

DGIIAAYQNP ASWKNNRIWL QFAKLTGFTL MGKGVIDGQG KQWWAGQCKW VNGREICNDR
DGIIAAYQNP ASWKNNRIWL QFAKLTGFT                  GQCKW VNGREICNDR-
 GIIAAYQNP ASW            LTGFTL MGKGVI                VNGREICNDR
                                     VIDGQG KQWWAGQCKW

DRPTAIKFDF STGLIIQGLK LMNSPEFHLV FGNCEGVKII GISITAPRDS PNTDGIDIFA
DRPTA          IIQGLK LMNSPEFHL                                 A-
DRPT
SKNFHLQKNT IGTGDDCVAI GTGSSNIVIE DLICGPGHGI SIGSLGRENS RAEVSYVHVN
SKNFHLQKNT IGTG                                         AEVSYVHVN-
                                                     S RAEVSYVHVN-
                 CVAI GTGSSNIV                   GRENS RAEVSYV  N-

GAKFIDTQNG LRIKTWQGGS GMASHIIYEN VEMINSENPI LINQFYCTSA SACQNQRSAV
GAK                      SHIIYEN VEMINSENPI LINQFYCT
GAKF                                             YCTSA SACQNQRSAV-
GAKFIDTQNG L                  EN VEMINSENPI
                              EN VEMINSENPI LINQFYCTSA SACQNQRS
                                          I LINQFYCTSA SACQNQRSAV-

QIQDVTYKNI RGTSATAAAI QLKCSDSMPC KDIKLSDISL KLTSGKIASC LNDNANGYFS
                               C KDIKLSDISL KLTSGKIASC LNDNANGYF
QIQDV                                     L KLTSGKIASC LNDN
QIQDVTYKN                    MPC KDIKLSDIS                    YFS-

GHVIPACKNL SPSAKRKESK SHKHPKTVMV KNMGAYDKGN RTRILLGSRP PNCTNKCHGC
GHVIPACKN  SPSAKRKESK SH
                   SK SHKHPKTVMV KNMGAYDKGN RTRILLGS

SPCKAKLVIV HRIMPQEYYP QRWMCSRHGK IYHP
           HRIMPQEYYP QR

Green: Known T-cell epitopes (102)
VDGIIAAYQNPASWK            LKLTSGKIASCLNDN            SRAEVSYVHVNGAKF
IIQGLKLMNSPEFHL            GKIASCLNDNANGYF            CKDIKLSDISLKLTS
ASKNFHLQKNTIGTG            NNRIWLQFAKLTGFT

Orange: Known IgE-epitopes (103)
PGNKKFVVNNLFFNGPCQPHF                    SHIIYENVEMINSENPILINQFYCT
GQCKWVNGREICNDRDRPTA                     YCTSASACQNQRSAVQIQDV

Red: 30 aa candidates
VIDGQGKQWWAGQCKWVNGREICNDRDRPT           ILINQFYCTSASACQNQRSAVQIQDVTYKN
SKSHKHPKTVMVKNMGAYDKGNRTRILLGS           MQLIIMAAAEDQSAQIMLDSDIEQYLRSNR
ENVEMINSENPILINQFYCTSASACQNQRS

Blue: 12 aa candidates
SPSAKRKESKSH                             GIIAAYQNPASW
HRIMPQEYYPQR                             CVAIGTGSSNIV
GRENSRAEVSYV                             LTGFTLMGKGVI
NGAKFIDTQNGL                             ENVEMINSENPI
MPCKDIKLSDIS                             YFSGHVIPACK
```

Tri r 2, Swiss prot name: Q9UW97

```
MGFITKAIPI VLAALSTVNG ARILEAGPHA EAIPNKYIVV MKREVSDEAF NAHTTWLSQS
                                        IVV MKREVSDEAF NAHTTWLSQS-

LNSRIMRRAG SSKPMAGMQD KYSLGGIFRA YSGEFDDAMI KDISSHDDVD FIEPDFVVRT
LNSRIMR                          YSGEFDDAMI KD   SHDDVD FIEPDF


TTNGTNLTHQ DNVPSWGLAR VGSKKPGGTT YYYDPSAGKG VTAYIIDTGI DIDHEDFQGR
                                    DPSAGKG VTAYIIDTGI DID
                                               YIIDTGI DIDHEDFQGR-
                                                          HEDFQGR-
     NLTHQ DNVPSWGLAR VGSKKPGGTT YYYDP
                                     PSAGKG VTAYII


AKWGENFVDQ QNTDCNGHGT HVAGTVGGTK YGLAKGVSLV AVKVLDCDGS GSNSGVIKGM
              DCNGHGT HVAGTVGGTK YGL
AKW                                 AKGVSLV AVKVLDCDGS GSN
AKWGENFVDQ QNT
      FVDQ QNTDCNGHGT HVAGTVGGTK YGLAKG          DCDGS GSNSGVI
   GENFVDQ QNTDCNGHGT HVAGTVGGTK YGL         VKVLDCDGS GSN

EWAMRQASGG GNGTAKAAGK SVMNMSLGGP RSEASNQAAK AISDAGIFMA VAAGNENMDA
                                    ASNQAAK AISDAGIFMA VAA
                        MNMSLGGP RSEA
                            LGGP RSEASNQA
QHSSPASEPS VCTVAASTKD DGKADFSNYG AVVDVYAPGK DITSLKPGGS TDTLSGTSMA
                                                          LSGTSMA-
                                                        DTLSGTSMA-
                                                 KPGGS TDTLSGT

SPHVCGLGAY LIGLGKQGGP GLCDTIKKMA NDVIQSPGEG TTGKLIYNGS GK
SPHVCGLGAY LIG
   VCGLGAY LIGLGKQGGP GLC
SPHVCGLGAY LIGLGKQGGP             DVIQSPGEG TTG
              LGKQGGP GLCDT
```

Green: Known T-cell epitopes (99)

DPSAGKGVTAYIIDTGIDID
YIIDTGIDIDHEDFQGRAKW
HEDFQGRAKWGENFVDQQNT
DCNGHGTHVAGTVGGTKYGL
AKGVSLVAVKVLDCDGSGSN
ASNQAAKAISDAGIFMAVAA
LSGTSMASPHVCGLGAYLIG
VCGLGAYLIGLGKQGGPGLC

Red: 30aa candidates

NLTHQDNVPSWGLARVGSKKPGGTTYYYDP
IVVMKREVSDEAFNAHTTWLSQSLNSRIMR
GENFVDQQNTDCNGHGTHVAGTVGGTKYGL
DTLSGTSMASPHVCGLGAYLIGLGKQGGPG
FVDQQNTDCNGHGTHVAGTVGGTKYGLAKG

Blue: 12aa candidates

SHDDVDFIEPDF
VKVLDCDGSGSN
LGKQGGPGLCDT
MNMSLGGPRSEA
DVIQSPGEGTTG
YSGEFDDAMIKD
DCDGSGSNSGVI
KPGGSTDTLSGT
PSAGKGVTAYII
LGGPRSEASNQA