UPTEC X06 033
Examensarbete 20 p
Augusti 2006

Evaluation of pattern recognition methods applied to in vitro IgE measurements

Eva Schreil

UPPSALA UNIVERSITET

Bioinformatics Programme

Uppsala University School of Engineering

UPTEC X 06 033	**Date of issue 2006-08**

Author

Eva Schreil

Title (English)

Evaluation of Pattern Recognition Methods Applied to In Vitro IgE Measurements

Title (Swedish)

Abstract

Food allergens from the plant kingdom are an important source of allergic reactions which are difficult to diagnose. Methods that can visualise relationships between these allergens are therefore needed. The main aim of this project was to evaluate pattern recognition methods for visualisation of multidimensional measurements of immunoglobulin E (IgE) in blood sera. Multidimensional scaling (MDS), a method for visualisation of multidimensional data in a reduced space, was evaluated and tested on IgE data from three patient groups with different IgE reactivity to cereals and grass in order to reveal relationships between food allergens from the plant kingdom. The results show that MDS is a useful and robust method for visualisation of IgE data.

Keywords

allergy, IgE, pattern recognition, multidimensional scaling

Supervisors

Annica Önell Ingvar Edlert

Phadia AB, Uppsala

Scientific reviewer

Mats Gustafsson

Department of Engineering Sciences, Uppsala University

Project name	Sponsors
Language **English**	Security **Secret until 2007-08-31**
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

Evaluation of Pattern Recognition Methods Applied to In Vitro IgE Measurements

Eva Schreil

Sammanfattning

Den vanligaste typen av allergi orsakas av att ett främmande ämne (allergen) framkallar en respons från immunförsvaret och antikroppen immunoglobulin E (IgE) frisätts i blodet. Phadia är ett företag inom allergologisk diagnostik som tillverkar och säljer tester och testinstrument för att mäta halten av IgE-antikroppar i blod.

Födoämnen från växtriket är en viktig källa till allergiska reaktioner som är svåra att diagnostisera. Många födoämnen är botaniskt närbesläktade och har liknande proteiner som kan orsaka en så kallad korsreaktivitet. Detta innebär att proteiner med liknande aminosyrasekvens eller struktur kan binda till samma typ av antikropp och orsaka en IgE-respons. En av svårigheterna med att diagnostisera allergi mot födoämnen ligger i att förstå mekanismerna bakom denna korsreaktivitet, dvs. att avgöra vilka proteiner som korsreaktiviteten orsakas av samt om IgE-responsen orsakar symtom hos patienten. För att kunna ställa säkrare födoämnes-diagnoser behövs därför metoder som kan kartlägga och visualisera samband mellan allergen.

Målet med detta examensarbete var att identifiera och utvärdera metoder inom mönsterigenkänning som kan appliceras på stora mängder IgE-data från en intern databas på Phadia. Med hjälp av dessa metoder undersöktes mönster och samband i IgE-data från tre olika grupper av patienter som hade olika kombinationer av IgE-reaktivitet mot vete och gräspollen. Den främsta metoden som användes var multidimensional scaling (MDS) och arbetet med denna metod skedde i programmeringsverktyget Matlab. Resultaten visade att MDS är en användbar och robust metod för visualisering av IgE-data. Utvärderingen av metoden ledde till rekommendationer och en applikation som kan användas på Phadia. En slutsats av att studera de tre patientgrupperna var att patienter med IgE-reaktivitet mot både vete och gräs även har IgE-reaktivitet mot många andra födoämnen. En större mängd data som även inkluderar exempelvis kliniska symtom skulle möjliggöra en djupare och mer fullständig analys av IgE-data i framtiden.

Examensarbete 20 p
Civilingenjörsprogrammet i Bioinformatik
Uppsala universitet augusti 2006

Table of contents

1. Introduction

The most common type of allergy is associated with elevated levels of the antibody immunoglobulin E (IgE), directed to a specific allergen (foreign substance causing an immune response) in the blood. This type of allergy is an increasing health problem afflicting millions of patients. Food allergies are believed to afflict between 5 % and 7.5 % of children and between 1 % and 2 % of adults (19). Plant-origin foods can be considered the most important sources of food allergic reactions in adults (27). Cereal grains are important sources of food allergies because they constitute the staple food for most of the world's population (18). Wheat is the cereal that causes most allergic reactions. Elevated levels of immunoglobulin E directed to wheat are common among patients with grass pollen allergy and food related symptoms. However, in these patients, elevated IgE levels to wheat do not always correlate with allergic symptoms. Proteins in plant-derived food and grasses are commonly similar in structure and sequence (26, 27, 28, 32), which can cause the antibodies to bind non-specifically to proteins that are not allergenic. Allergy tests based on measurements of IgE in blood sera are therefore unreliable for diagnosis of cereal grain allergy, and wheat in particular. This, in addition to diffuse symptoms, makes it difficult to diagnose patients with an IgE reactivity to wheat grain.

Phadia is an allergologic diagnostics company that develops and sells test reagents and test instruments for allergy testing on blood sera. The test systems are based on the measurements of the level of IgE directed to a specific allergen in the blood. It is desirable for a company like Phadia to increase the specificity of the tests for cereal grain allergy. In order to do this, the mechanisms behind binding of IgE antibodies to proteins of cereal grains, grass pollens and foods of plant origin need to be studied. Different experimental studies have aimed at clarifying the relationships between IgE reactivity to cereal grains, grass pollens and plant-derived food (3, 9, 12, 18, 28). However, traditional experiments performed in laboratories are time-consuming and normally, only a few patient samples can be analysed at the same time. In addition, it is difficult to visualise the results in a way that provides an overview of IgE reactivity patterns. Therefore, new approaches are needed to study the IgE reactivity patterns in patients with sensitisation to cereal grains and grass pollens.

Bioinformatics, a cross-disciplinary area in biology, mathematics and computer science, is widely used to analyse data within molecular biology (33). So far, the role of bioinformatics has been limited in allergy diagnostics. At Phadia, a large amount of data on IgE levels in blood sera is stored in an internal database. The database is probably unique in its kind since such large number of IgE measurements on so many allergens per blood test rarely has been collected. Therefore, it provides a unique opportunity to study relationships and patterns in IgE data. Methods that can be applied on this data in order to reveal relationships and patterns are desirable to identify since they can function as a complement to experimental studies, and in the end support the clinical diagnosis of allergies. In a previous study conducted by Phadia, Uppsala University and National Food Administration, it has been shown that methods within pattern recognition, an area related to bioinformatics, could provide novel ways to visualise IgE reactivity patterns. One advantage of using a bioinformatical approach is that it demands less resources and a larger amount of patient sera can be analysed simultaneously. This degree project is focused on identifying and evaluating bioinformatical methods within pattern recognition that can be applied on IgE data in Phadia's internal database, addressing the problem of revealing IgE reactivity patterns in patients sensitised to cereal grains and pollens.

The background chapter of this report deals with mechanisms behind allergy, diagnostic methods and an introduction to pattern recognition with a short theoretical background to some of the methods presented in the methods and data chapter. Methods and data describe how the data used in this study was retrieved and structured and how the methods were implemented. The results section presents the results from this study. First are the results from the study of IgE reactivity patterns in patients with different IgE reactivity to grass pollen and plant-derived food. The second part of the result section is an evaluation of the methods used and the last part deals with improvements of the method. The results are followed by a discussion of the results in the discussion.

In this report, allergens are annotated with a code which corresponds to Phadia's product code of ImmunoCAP^{TM} allergens. The annotation is built up of one letter and one number. The letter refers to the type of allergen and the number is an identifier. For example, in 'f4' which is the wheat allergen, f denotes a food allergen. Another example is 'g6', timothy grass, in which g denotes a grass pollen allergen. Appendix A contains all allergen codes presented in the text.

Other abbreviations used:

BPCA Bayesian principal components analysis

CCD Cross-reactive carbohydrate determinants

CV Coefficient of variance

DBPCFC Double blind placebo-controlled food challenge

IgE Immunoglobulin E

LLS Local least squares

MDS Multidimensional scaling

NRMSE Normalised root mean squared error

PCA Principal components analysis

SPT Skin prick test

2. Aim

The overreaching goal of this degree project was to identify, implement and evaluate useful pattern recognition methods for visualisation and analysis of IgE data. The methods were applied on an excerpt from Phadia's internal database containing data on patients' IgE responses to several allergens. Activities in this project aimed at:

- Comparing the IgE reactivity patterns in groups of patients with different profiles with respect to their IgE response to wheat and grass allergen extracts. The following groups were compared:
  A. Patients with positive IgE responses to wheat and grass pollens.
  B. Patients with positive IgE responses to wheat and negative IgE responses to grass pollens.
  C. Patients with negative IgE responses to wheat and positive IgE responses to grass.
- Studying IgE responses to components (proteins) of allergen extracts in order to explain the resulting IgE reactivity patterns for allergen extracts. Due to lack of data, this study could only be conducted at group A.
- Evaluating and validating the robustness of the pattern recognition methods when applied to IgE data from Phadia's internal database.
- Developing a ready-to-use application for analysis and visualisation of patterns in IgE data at Phadia.

Two long-term goals associated with the aims of this degree project are:

- To improve the diagnostics of food allergy by:
  - Identifying unknown relationships between allergens from different sources
  - Increasing the specificity of the test instruments
- To develop a ready-to-use toolbox with pattern recognition methods for usage in various projects at Phadia.

3. Background

3.1. Mechanisms behind allergy

The human immune system protects the body from foreign molecules belonging to viruses, bacteria, fungi and parasites. Foreign molecules are also found on surfaces of foreign materials such as pollen. When the human body is exposed to certain foreign molecules, the immune system mounts an immune response, generated by lymphocytes circulating in the blood and lymph. A foreign molecule that triggers a response by a lymphocyte is called an antigen. The two main types of lymphocytes are B cells and T cells, which both recognize specific antigens by plasma membrane-bound antigen receptors. There are two types of immune responses to antigens (7): humoral (antibody-mediated) immune response and cell-mediated immune response.

The humoral response is initiated when an antigen binds to an antigen receptor on a B cell. As the B cell is stimulated by the antigen, it proliferates and differentiates into a clone of plasma cells and memory B cells. This type of B cell response can only be induced by so called T-dependant antigens which stimulate antibody production with help from T cells. Typically, proteins of foreign substances such as bee venom or pollen belong to this type of antigens that induce an allergic, humoral response (7).

Plasma cells, originating from B cells, secrete antibodies which constitute a group of globular serum proteins called immunoglobulins. The antibody binds to a small part of the antigen protein called epitope. One antigen protein can have many epitopes. An antibody consists of four polypeptide chains forming a Y-shaped molecule. At the two tips of the molecule are variable regions unique to each antigen, which bind to the epitope of the antigen. There are five major classes of antibodies: IgG, IgM, IgA, IgD and IgE.

What we in everyday language call allergy, is associated with immunoglobulin E (IgE). This type of allergy is sometimes called type I hypersensitivity (16). IgE has the same general features as all immunoglobulins, but is of the lowest concentration of all immunoglobulins in blood serum (17). The tail regions of IgE bind with high affinity to receptors on the surface of mast cells and basophils called FcεRI (30). When allergens (antigens) enter the body, they attach to the antigen-binding sites on two cross-linking IgE molecules on the mast cell. This induces the mast cell to degranulate which involves a release of inflammatory agents such as histamine (Figure 1) from vesicles called granules on the mast cell (also called mediator release). The released substances give rise to allergic symptoms such as sneezing, runny nose and tearing eyes.

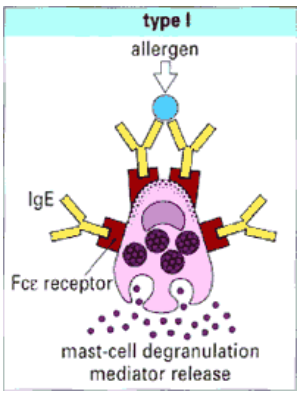


Figure 1. Allergen binding to IgE antibodies on the surface of a mast cell which causes a mediator release of inflammatory agents. (Used with permission from Phadia AB.)

3.2. Diagnosing allergy

The study and diagnosis of allergy is conducted at two levels: *in vitro* or serologic, which means that the IgE levels in blood sera are studied and *in vivo* or clinical, which means that the symptoms of the patient are studied. This section gives an overview of different *in vivo* and *in vitro* techniques for diagnosing allergy. In many cases, these methods complement each other before a diagnosis is made.

3.2.1. Some available *in vivo* and *in vitro* methods

Two commonly used *in vivo* methods for the diagnosis of allergy is the skin prick test (SPT) and the double blind placebo-controlled food challenge (DBPCFC).

Skin prick tests are quick, inexpensive and easy to use (19, 32). A small amount of allergen is introduced with a small puncture into the skin of the allergic patient. If the skin mast cells are activated, histamine is released which induce a reaction in the skin. The SPT has a good sensitivity and prediction of negative results. However, it has been found that positive reactions are not always correlated to symptoms (19, 32). This is especially the case with food allergens, which are normally absorbed into the body by ingestion (32). Therefore, skin prick tests alone cannot confirm food allergy when they show a positive result (19).

The DBPCFC is described in the literature (19, 32) as the "gold standard" for the diagnosis of food allergy. Patients receive the suspected food allergen hidden in an inert matrix and a placebo preparation without the hidden allergen, and the symptoms are subsequently observed. The risks and safety issues have limited the utility of the DBPCFC (32). A well-known problem with SPTs and DBPCFCs is that the procedures vary between clinics and countries which make the results difficult to compare.

In vitro assays are used to detect IgE in serum and include specific IgE immunoassays (to which Phadia's test systems (section 3.2.2) belong), SDS-PAGE (Sodium dodecylsulfate-polyacrylamide gel electrophoresis) immunoblotting and allergen microarrays (32). The general principle of immunoassays is to detect IgE that binds to a specific allergen fixed to a surface (30). Some of the advantages of *in vitro* testing over *in vivo* methods are that they offer quantitative measurements of IgE, higher safety and a long-term storage of samples (11). The standardised *in vitro* procedure facilitates world-wide comparisons of test results. Furthermore, Johansson (17) argues that the *in vitro* allergy tests have greatly improved the quality of allergy diagnosis.

3.2.2. Phadia's *in vitro* test principle

Phadia develops test systems to support the clinical diagnosis and monitoring of allergy. The company develops and sells reagents and instruments for *in vitro* testing on blood serum. Phadia's latest technology is the ImmunoCAP™ technology which consists of immunoassay reagents, instrumentation, and information management software developed for the measurement of total and specific IgE in serum or plasma. The detected level of allergen specific IgE antibodies in the blood serum when exposed to a specific allergen is called a specific IgE (sIgE). Allergens are bound to a solid phase called an ImmunoCAP^{TM} which consists of a cellulose derivative enclosed in a capsule (15) (Figure 2).

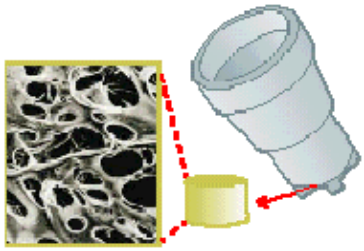


Figure 2. Structure of the solid-phase. (Used with permission from Phadia AB.)

The allergen of interest is covalently coupled to the ImmunoCAP™ and is allowed to react with the specific IgE in the patient sample (Figure 3 a). Non-specific IgE antibodies that have not reacted with the allergen are washed away and enzyme-labelled antibodies against IgE (anti-IgE) are added to form a complex (Figure 3 b). Again, unbound anti-IgE is washed away and a reagent is added to the complex (Figure 3 c). The reagent will recognise the enzyme-labelled antibodies and cause the complex to emit fluorescence. After incubation, the fluorescence of the complex is measured and the higher the fluorescence, the higher the concentration of specific IgE in the blood sample (14).

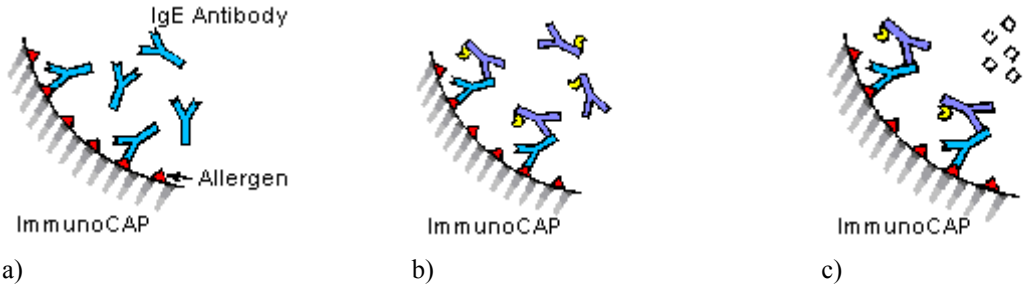


Figure 3. The ImmunoCAP™ test procedure in three steps. a) IgE antibodies in the patient sample react with the allergen bound to the CAP and unbound IgE antibodies are washed away. b) Enzyme-labelled antibodies against IgE (anti-IgE) are added to form a complex with the allergen-bound IgE. Unbound anti-IgE is washed away. c) A reagent is added to the complex and the fluorescence of the complex is measured.
(Used with permission from Phadia AB.)

Values are expressed in the unit kU_A/l (kilo units of IgE per litre), where A denotes allergen-specific antibodies. ImmunoCAP™ detects specific IgE antibodies in blood serum in the range of 0.1 - 100 kU_A/l. In clinical practice, 0.35 kU_A /l has commonly been used as a cut-off. The healthy individual has a very low level of specific IgE in the blood, normally below 0.35 kU_A/l. Patients with a sensitisation show elevated levels, i.e. above 0.35 kU_A/l. This can also be called an IgE reactivity. Generally, the higher the kU_A/l value, the more exposed the patient is to the allergen and the more likely the risk of symptoms (13, 30). A sensitisation with allergic symptoms is defined as an allergy. Multi-sensitisation occurs when a patient has elevated IgE levels to many independent allergens.

In the literature (19, 32), the ImmunoCAP™ system is described as a reliable and popular method with higher sensitivity for food allergens than the skin prick test (19). However, the relationship between specific IgE and clinical relevance is an ongoing discussion (32). Many individuals are sensitised to allergens but show no clinical symptoms to these allergens (30). Therefore, in *vitro tests* alone cannot confirm allergy and need to be complemented with other methods such as DBPCFC. One cause of clinically irrelevant positive tests is cross-reactivity, which is described in the next section.

3.3. Cross-reactivity: mechanisms and common sources

The term allergen sometimes refers to a mix of a number of different components or proteins coming from the same allergenic source, e.g. birch pollen, and sometimes to the allergenic protein. A more correct term for a mix of proteins from an allergic source is allergen extract. The allergen extracts bound to the solid phase in an immunoassay are of a complex nature because of a high heterogeneity (11). An allergen extract often contain several allergenic and non-allergenic proteins and the exact composition and amount of protein components in allergen extracts is often unknown.

Cross-reactivity involves an IgE response to proteins from different sources that share sequence homology or have similar three-dimensional structures (5). The binding between an antigen and an antibody takes place in the antibody's binding site and the epitope on the antigen. Cross-reactivity occurs when an antibody's binding site, directed to an original epitope, also recognises epitopes that have the same three-dimensional structure or a high degree of similar amino acid sequence. Cross-reactivity can have clinical relevance, which means that the IgE response to cross-reactive protein gives rise to symptoms. Clinically irrelevant cross-reactivity occurs when the IgE response to cross-reactive proteins is not related to symptoms.

Clinically irrelevant cross-reactivity causes false positive in vitro test results (8). Therefore, it is desirable to exclude cross-reactive proteins that lack clinical relevance from the allergen extracts in the solid phase of the immunoassay. By studying IgE reactivity patterns, possible cross-reactive relationships between allergens can be revealed. Once the causing protein of the cross-reactivity is identified, and the clinical relevance is determined, the allergen extract can be modified and a higher specificity of the test can be obtained.

In the study of IgE reactivity patterns, the identification of patients with multi-reactive patterns is one step towards finding possible cross-reactive relationships. According to Ebo et al. (8), multi-reactive patterns can be explained in three ways. First, true independent sensitisation for different allergens account for some of the results. Second, a high total serum IgE level can cause non-specific binding of the IgE to the solid phase. Third, cross-reactivity due to homologous sequences or structures between allergens from different sources. Three sources of cross-reactivity are particularly common in plants and plant-derived foods: carbohydrate determinants, profilin and Bet v 1, all described in the next section. Consequently, these sources are of interest in this study of IgE reactivity patterns in patients with sensitisation to grass pollen and/or plant-derived foods.

3.3.1. Common cross-reactive components

There are many examples of IgE cross-reactivity between similar allergenic proteins (components) (28). Carbohydrate determinants are a common source of cross-reactivity named cross-reactive carbohydrate determinants (CCD). Cross-reactive carbohydrate determinants are carbohydrate structures that originate from pollen and plant food glycoproteins and have a wide distribution among plant-derived proteins. Patients with plant-derived allergies and multiple pollen sensitisations have a higher prevalence of IgE to CCD (23). CCDs are capable of inducing IgE antibodies but the clinical relevance is controversial (8). Therefore, the presence of CCD-IgE complicates the serologic diagnosis of allergy. Bromelain is a protease that contains CCD (8) and is therefore often used as a marker of the presence of IgE to CCD. Bromelain can be extracted from pineapple.

Profilin is another widespread cross-reactive protein and IgE reactions to profilin occur quite frequently (28) in a wide range of plant allergen extracts. The profilins exist in

eukaryotic organisms' cytosol and take part in the formation of the cytoskeleton (22). However, the protein sequence of profilin in different organisms differs much. It is found though, that even distantly related species with a profilin homology at low 30%, have a highly conserved tertiary structure (28), explaining the high cross-reactivity. Profilins cause a wide range of cross-reactivity among pollens and plant foods. Even though they are capable of inducing IgE antibodies, the clinical role of profilin is not clear (8). K. Andersson and J. Lidholm (1) suggest that they are minor allergenic components of grass pollen and plant foods with IgE reaction in 15-30% of individuals with pollen allergy.

An example of a clinically relevant cross-reactive protein is Bet v 1, which has high sequence homology with many proteins in other food allergen extracts. Bet v1 is the major allergenic component in birch pollen and it cross-reacts with homologous proteins in hazelnut, apple, soya bean, bell pepper and celery (26). This cross-reactivity may cause symptoms.

3.4. Food allergy and allergens

Food allergy affects between 5 % and 7.5 % of children and between 1 % and 2 % of adults (19). Food can cause allergic reactions by several mechanisms (4), but the most studied and best characterised are those that are type I hypersensitivity (IgE mediated) (19). Symptoms associated with IgE mediated food allergy usually begin within an hour after ingestion (19) and involve flushing, hives, wheeze and gastrointestinal symptoms. Plant-origin foods are considered the most important sources of allergic reactions, particularly in adults (27).

A thorough investigation (diagnosis) of food allergy generally begins with a case history, followed by a specific IgE test, performed with a skin prick test or an *in vitro* IgE test. A combination of these inputs is used when making the diagnosis. One important problem for diagnosis of plant food allergy is clinically relevant and irrelevant unknown cross-reactivity of allergen extracts. Many plant foods come from closely related botanical families and have structurally homologous proteins, which can cause cross-reactivity. For example, IgE directed towards epitopes on grass pollen, can also bind to wheat proteins without any clinical relevance of the finding (4). It is often difficult to determine the clinical relevance, i.e. the connection to symptoms of cross-reactivity between plant food allergens(27).

Nuts and seeds, especially peanut, as well as fresh fruits and vegetables are common sources of food allergy. Cereals can also cause allergic reactions and is an important group of allergens because they are the main alimentary source in the world, constituting the staple food for most of the world's population (18, 27). In addition, cereal grains cause adverse reactions in some human beings (18). Cereals belong to the grass species and are, together with grasses, monocotyledons classified in the Poaceae family. Due to the close botanical relationships, cross-reactivity can occur between cereals and grasses. Studies have indeed shown that patients with cereal grain specific IgE have increased positive IgE to grasses (18). However, the clinical relevance of these findings have been questioned, suggesting that cross-reactivity gives rise to false positive *in vitro* test results to grass pollens (18).

3.4.1. Wheat allergy

Eight common foods are responsible for over 90 % of food allergies and among them is wheat (19, 4). Diseases associated with wheat exposure are gluten sensitive enteropathy (celiac disease), baker's asthma and food hypersensitivity. Celiac disease is not mediated by IgE and is caused by the gliadin fraction of wheat. Baker's asthma is an allergic reaction to inhaled wheat flour and the food hypersensitivity is related to ingestion of wheat. Both of the latter are mediated by IgE and symptoms include respiratory and gastrointestinal symptoms (18). Gliadin, which is the protein responsible for celiac disease, has also been shown to induce

allergic reactions at IgE level. In addition, allergens in rye and barley are cross-reactive with wheat gliadin at IgE level (18).

Cross-reactivity among cereal grains is more common than in other food families and it has been shown that patients with wheat allergy show an extensive *in vitro* cross-reactivity to other grains (18). In addition, patients with grass pollen allergy show an extensive *in vitro* cross-reactivity to cereal grains (18). These factors make it difficult to diagnose wheat allergy with *in vitro* IgE tests. Jones (18) argues that the problem is lack of specificity of *in vitro* testing in the diagnosis of cereal grain hypersensitivity.

The extensive cross-reactivity within cereal grains and between cereal grains and grasses addresses the need for methods that can reveal cross-reactivity patterns between these allergen extracts. A higher specificity of the *in vitro* tests for wheat allergy can be obtained by identifying the proteins responsible for clinical irrelevant cross-reactivity and exclude them from the allergen extract bound to the solid phase.

3.5. Grass pollen allergy and allergens

Grass pollens are a common source of IgE mediated allergy (10, 24). Since they are very widespread and produce large amounts of pollen grains, grass pollens are one of the most important allergen sources worldwide (1, 10). Grass species of the subfamily Pooideae dominate the temperature regions of the northern hemisphere (24). Timothy grass is one of the major allergenic grasses which belong to this subfamily and its allergenic proteins (components) are well-studied. Because of the high homology among grass pollen proteins, patients allergic to grass pollen will often react to many species (Petersen). In the following parts of this section, the term allergen refers to allergenic proteins or components and not complete extracts.

The identification of components in grass pollen extracts has led to a classification of 13 allergen grass groups (1). Each group contains similar proteins from grasses of different species. The most important grass pollen allergens belong to the groups 1 and 5 (24). These allergens are called major allergens since they account for most of the immune responses to grass pollen allergen extracts (1). In this project, group 4 and group 12 grass pollen allergens are also of great importance because their cross-reactivity with other proteins of plant origin. Therefore, allergens of group 4 and 12 can cause multi-reactive patterns among plant allergens important to consider in the IgE reactivity patterns of patients sensitised to grass pollens and plant-derived food. Table 1 summarizes the grass pollen groups of interest for this study.

Grass pollen allergen group number	Features of the components in the group	Timothy grass pollen component
1	Major allergens	Phl p 1
5	Major allergens	Phl p 5
4	Glycoproteins and major allergens. Cross-reactivity with plant foods.	Phl p 4
12	Profilins, cross-reactivity with plant foods.	Phl p 12

Table 1. Summary of grass pollen allergen groups of relevance for this project, their features and the corresponding timothy grass pollen component belonging to the group.

About 90% of individuals allergic to grass pollen show an IgE reactivity to the allergens of *group 1* (1) and the homology among these allergens is high. The major timothy grass pollen allergen Phl p 1 belongs to this group and cross-reacts with most group 1 allergens in grass, corn, and monocots (31).

Group 5 grass pollen allergens causes IgE reactions between 65 and 85% among individuals with grass pollen allergy (1). The group 5 allergen Phl p 5 is the dominant allergen in allergen extracts of timothy grass (31).

Group 4 grass pollen allergens are glycoproteins to which the major allergen of timothy grass, Phl p 4, belongs. Allergens in this group are classified as major allergens since up to 80 % of grass pollen sensitised individuals show IgE reactivity to them (1).Group 4-related allergens occur in plant food and can cause cross-reactivity between pollens and plant food (10). It has been suggested that the glycan structures of Phl p 4 cause cross-reactivity between Phl p 4 and other glycoproteins of plant origin (24).

Group 12 grass pollen allergens are profilins (1) and they account for a large part of cross-reactivity between pollen and vegetable foods. It is suggested that patients who are sensitised to pollen profilins cross-react with a wide range of fruits and vegetables (26). This cross-reactivity may not be associated with symptoms of food allergy (1). Timothy grass contains the group 12 allergen Phl p 12, a profilin.

Pollen-sensitised patients often suffer from clinical allergic reactions after intake of plant food (20, 26). The symptoms are termed oral allergy syndrome (OAS) and occur in the mouth and throat when it comes in contact with the allergen. Patients with OAS experience more severe symptoms of food allergy during and after the pollen season (20). Cross-reactive epitopes in pollens and plant derived food are responsible for sensitisation in patients with OAS (20). Profilin and Bet v 1 in birch pollen are such epitopes (see section 3.3.1). In addition, a 60 kD protein present in grass pollens has been found to share epitopes with allergens in fruits and vegetables (12). Tomato is one well-studied vegetable that is believed to share epitopes with grass pollen allergens (9, 28).

3.6. Pattern recognition

Pattern recognition aims at classifying multidimensional data. Sub-areas within these fields aim at visualising underlying patterns in data of high dimensionality in a reduced dimension space. The general idea is to find the minimum number of dimensions needed to represent the data and, if reasonable, visualise the data in two or three dimensions.

Principal components analysis (PCA) and multidimensional scaling (MDS) are dimension-reduction techniques that can be used for visualizing large data sets. These methods compress the data into a new space of a reduced dimension.

3.6.1. Principal components analysis (PCA)

The widely used visualisation technique principal components analysis projects data along the directions of maximal variance (6). The covariance matrix A is calculated from the original data matrix with samples having measurements on several variables (for example allergens). Subsequently, the covariance matrix's eigenvalues and eigenvectors are found. By ordering the eigenvectors in the order of descending eigenvalues, an ordered orthogonal basis is obtained, with the first eigenvector capturing the largest variance of the data. The eigenvectors are also known as loading vectors. A matrix product of the eigenvectors and the original data generates scores that provide information about how the original samples relates to the new orthogonal basis.

When using the first two principal components, the result of a PCA can be visualised in two-dimensional plots. The original samples will be projected in a so called score plot in which the samples are projected along the two directions of maximal variance. A so called loading plot gives information about how the variables are related.

3.6.2. Multidimensional scaling (MDS)

The aim of multidimensional scaling (MDS) projection techniques is to preserve the distances among data points. Data that are close in the original data set should be mapped in the new space so that that they are still close (6). Classical scaling is one type of MDS that takes a symmetrical n*n distance matrix, consisting of pair-wise distances between all n variables, as input and constructs a matrix of dimension n*p where p<n. The distances between the n variables of the original distance matrix are reconstructed in the reduced space of the smallest possible dimension p. In the reconstruction of the distance matrix to a reduced dimension, an eigenvector problem is solved. The eigenvectors associated with the largest eigenvalues are used to obtain a distribution of the coordinates in the reduced space that best capture the original distances in the distance matrix.

3.6.3. Missing values

Missing values occur frequently in data due to unreliable or absent measurements (21, 25, 34). Since many pattern recognition methods, including PCA and MDS, require a complete matrix with no missing elements, methods for imputing missing data are needed. Several automated methods for estimating missing data have been proposed (21, 25, 34). The first commonly used techniques for dealing with missing data were all based on models and model assumptions had to be made (34). Two recently proposed missing value estimation techniques, Bayesian Principal Components Analysis (BPCA) and Local Least Squares imputation (LLS), estimate the parameters automatically. In most cases, these methods outperform earlier proposed missing value estimation methods in accuracy (21, 25). It has been argued that LLS is better than BPCA for data with local similarities among samples (21). However, it is proposed that BPCA has an advantage for a larger numbers of samples (21).

3.6.4. Allergen maps

Allergen maps provide a novel way to visualise IgE reactivity patterns in large data sets comprising many IgE responses to several allergens. The degree to which two allergens are related, the correlation, is transformed into pair-wise distances and a pattern recognition algorithm projects the pair-wise distances between all allergens in two or three dimensions. In the resulting allergen map, correlating allergens group together.

In a recent study preceding this degree project, Phadia, National Food Administration and Uppsala University, conducted a study of IgE responses to 89 allergens from 1127 individuals (36). These allergens belonged to the following groups: foods of plant origin, foods of animal origin, grass pollens, weed pollens, tree pollens, house dust mites, epidermals, moulds, invertebrates and venoms (Appendix A). The visualisation of patterns in the IgE data of these allergens provided an overview over cross-reactivity and relationships between the allergens (Figure 4). In this study, the MDS algorithm was used to visualise the data. As can be seen in Figure 4, allergens of the same origin group together. The grouping of pollens and foods from plants (green area in Figure 4) verified the extensive cross-reactivity between allergens from the plant kingdom.

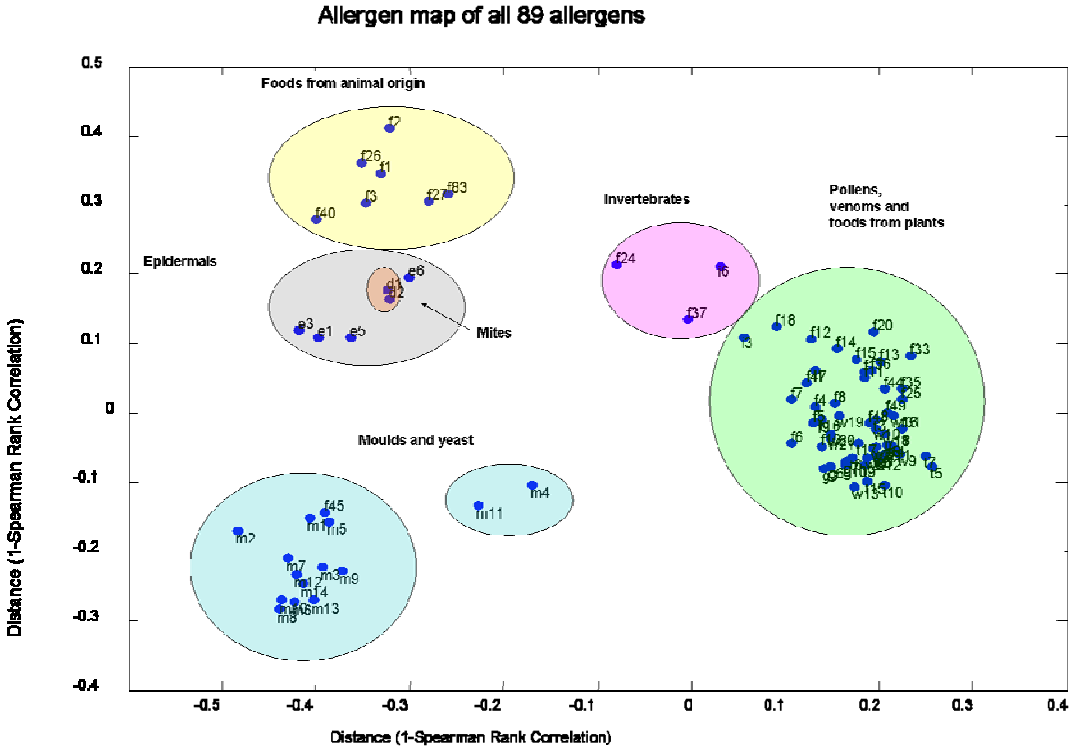


Figure 4. Allergen map of 89 allergens based on data from 1127 blood sera samples. (Used with permission from Phadia).

This degree project is a continuation of the allergen map study and aims at evaluating the method thoroughly.

3.7. Summary and outlook

The contents of the background chapter can be boiled down into one possible scenario for Phadia:

Extensive cross-reactivity between allergens from the plant kingdom gives rise to a difficulty of diagnosing plant food allergies. The allergen extracts from the plant kingdom often contain an unknown composition of proteins with high homology between different species. By visualising relationships among these allergen extracts using pattern recognition methods, possible cross-reactive relationships can be discovered. In a next step, it would be desirable to identify the protein component that is responsible for the cross-reactivity patterns and determine its clinical relevance. Once the component is determined as clinically irrelevant, it can, if possible, be excluded from the solid phase of the ImmunoCAP^{TM} and the specificity of the test can be increased. Avoiding IgE binding to non-allergic proteins in the allergen extract will result in a reduced number of false positive tests and a higher reliability for the diagnosis of plant food allergy can be obtained. To be able to determine the clinical relevance of the cross-reactivity, clinical data on symptoms must be collected. Unfortunately, clinical data is not included in this project.

This project is focused on the first step in this possible scenario by identifying and evaluating the pattern recognition methods that can be used to visualise IgE reactivity patterns. Some of the strong correlations between allergens that can be discovered with these methods might be caused by cross-reactivity.

4. Methods and data

4.1. Extract IgE data

Phadia possesses a blood serum bank in which blood serum from donors around the world is collected and stored. The blood sera have been collected since the beginning of the 1980s, mainly from the US and Northern Europe. It is collected to facilitate quality control of the production as well as research on IgE levels and allergens in order to improve the products. Blood sera are preferably bought from individuals who are multi-sensitised, which means that their blood contains a wide range of IgE antibodies directed to different allergens. Specific IgE (sIgE) responses of several allergen extracts have been detected in the bio-bank blood sera with the ImmunoCAP^{TM} technology and the data is stored in an internal database comprising about 49 000 samples. The clinical information of the samples is very limited.

4.1.1. Data retrieval

The data used for the data analysis was retrieved from Phadia's internal blood sera database. A number of 93 allergens were included within the following groups: foods of plant origin, foods of animal origin, grass pollens, weed pollens, tree pollens, house dust mites, epidermals, moulds, invertebrates and venoms (Appendix A). These allergens constitute the main part of a standard screen panel used to screen blood sera as they arrive to Phadia. Thus, we could expect to have many measurements on these allergens in the database. Individuals with at least one positive test on one of these 93 allergens were included in the resulting data set, which comprised 8855 samples.

4.1.2. Structure of data

The raw data retrieved from the database search was subsequently transformed into an excel sheet with a structure shown in Figure 5.

LABEL	DON_ID	CATE_CODE	COUNTRY	DATES	SYST_CODE	algE	f10	f11	f12	...
43205	40854	A	TYSKLAND	2002-01-15	UNICAP	3020	31,3	23,3	18,6	...
48100	42657	A	USA	2004-10-13	UNICAP	1027	100	100	41,3	...
33143	6782	H	USA	1996-07-19	CAP	3260	100	95,9	76,5	...
35000	11300	A	SVERIGE	1996-09-26	CAP	1705	0,64	0,6	0,45	...
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...

Figure 5. Structure of blood sera data in Excel

The columns of the table contain the following data:

LABEL identifies the sample ID since one donor can have more than one sample in the database.

DON_ID identifies the donor.

CATE_CODE contain additional information about the donor such as gender and age.

COUNTRY refers to the country in which the sample was collected.

DATES give information about at which time the samples were collected.

SYST_CODE contain information on which test instrument the sample was analyzed.

algE refers to the total level of IgE antibodies in the blood serum sample.

Columns labelled with blue contain the measured IgE response in the blood when exposed to the allergen extract of that column. The levels are given in kU/l.

In the following sections, rows of the data will occasionally be called samples and similarly, the columns with allergens will sometimes be called variables.

4.1.3. Subsets

In order to study relationships within plant food allergens and the relationships between food allergens and grass pollen allergens, 30 allergens were extracted from the original 93. These 30 allergens included 21 foods of plant origin and 9 grasses (Table 2). The chosen food allergens from the plant kingdom had shown high correlations to wheat in a previous allergen map study (see section 1.5). Initially, samples with no measurements on these 30 allergens were removed as well as samples with no measurements above 0.35 kU/l.

Allergen code	Name	Allergen code	Name
f4	Wheat	g2	Bermuda grass
f5	Rye	g3	Cocksfoot
f6	Barley	g6	Timothy grass
f7	Oat	g7	Common reed
f8	Maize	g8	Meadow grass
f9	Rice	g10	Johnson grass
f10	Sesame	g12	Rye pollen
f11	Buckwheat	g14	Oat pollen
f12	Pea	g15	Wheat pollen
f13	Peanut		
f14	Soya bean		
f15	White bean		
f20	Almond		
f25	Tomato		
f31	Carrot		
f33	Orange		
f35	Potato		
f36	Coconut		
f44	Strawberry		
f47	Garlic		
f48	Onion		

Table 2. The 30 allergens included in the study.

In order to survey the relationship between grass pollen allergy, wheat allergy and plant food allergy, the data was further reduced and divided into two main groups:

Group A. **Patients with positive IgE responses to wheat and grass pollens**
Samples with specific IgE (sIgE) responses >0.35 kU/l on at least one of the allergens wheat (f4), rye (f5), barley (f6) or oat (f7) and sIgE responses >1 kU/l on all of the grasses.

Group B. **Patients with positive IgE responses to wheat and negative IgE responses to grass pollens**
Samples with sIgE responses >0.35 kU/l on at least one of the allergens wheat (f4), rye (f5), barley (f6) or oat (f7) and sIgE responses ≤ 1 kU/l on all of the grasses.

Rye, barley and oat formed a basis for selection of samples together with wheat because of their close biological relationship to wheat. Therefore, in this report, individuals with a sensitisation to one of the cereal grains are regarded as sensitised to wheat. Furthermore, the threshold for a positive grass pollen test was set to 1 kU/l because IgE responses directed to grass pollens are generally higher. Data analyses were performed at these two subsets separately and the results were subsequently compared. The two groups were also used in the evaluation of methods such as missing values, measurement noise and the error of reconstruction.

As the project proceeded and the results of group A and B were obtained, the need to study a third group came up.

Group C. Patients with negative IgE responses to wheat and positive IgE responses to grass
Samples with sIgE responses ≤ 0.35 kU/l on all of the allergens wheat (f4), rye (f5), barley (f6) or oat (f7) and sIgE responses >1 kU/l on all of the grasses.

The aim of studying this group was to further clarify the relationships between grass pollens and plant food allergens. Results of data analyses were compared to the results of group A and B. Since this group was included in the project at a later stage, it was not used in the evaluation of the method including for instance the missing value and measurement noise studies.

4.2. Component IgE data

The IgE data stored in the internal database contain specific IgE responses to allergen extracts which contain several protein components. Individuals that have a positive IgE response could have reacted to one or more components in the extracts. IgE data that contains specific IgE responses to components can clarify what component(s) in the extracts that those individuals are sensitised to.

4.2.1. Data retrieval

The component data of this study came from a research group at Phadia who had studied the IgE responses to pollen components in blood sera from 81 individuals. Together with timothy grass pollen components, the cross-reactive components CCD (contained in bromelain), rBet v 2 (a Birch pollen component) and profilin, were included in the study. In addition, the group measured the specific IgE directed to wheat extract. Table 3 shows the components included in the study.

Component code	Component name
g205	Phl p 1
Rg208	Phl p 4
g215/g207	Phl p 5
g210	Phl p 7
Rg212	Phl p 12 (profilin)
k202	Bromelain
t216	rBet v 2 (recombinant)

Table 3. Components of the component data. Component code refers to the Phadia's product code and an 'R' means that the protein component is recombinant.

4.2.2. Structure of data

The structure of the component IgE data corresponded to the structure of the extract IgE data. Each row in the data contained one sample and its level of IgE against each component in

columns. The first column contained the sample ID, which corresponds to the LABEL column in the extract data.

4.2.3. Preparation of data set

By relating the sample IDs in the component data with the LABELs in the data set that were used to study group A, B and C, the IgE responses to 93 allergen extracts could be retrieved. This was done in order to relate the IgE responses to different components to each individual's IgE response to allergen extracts. Of the 81 individuals in the component study, 58 could be found in the data set with IgE responses to allergen extracts. Among these, 34 filled the criteria of group A and the rest did not have a sufficient amount of measurements and were excluded from further studies. The resulting data set comprised 34 samples with 93 measured IgE responses to allergen extracts together with 7 measurements on components. The IgE response to the wheat extract was measured both in the individual component study and in the internal database. In the following studies, the IgE response to wheat from the database was used.

4.3. Exploring the data

The IgE data of the three groups A, B and C was explored by means of multi-sensitisation, IgE levels and the prevalence of particularly high IgE responses to certain allergens. This exploration aimed at investigating the general allergic profile of the three groups.

Here, all of the 93 allergens from the original database search were used even though the criteria for forming the groups A, B and C were the same. The 93 allergens were grouped in ten groups in accordance to Appendix A and the percentage of samples with positive measurements within four different, arbitrarily chosen intervals (0.36-1 kU/l, 1-5 kU/l, 5-15 kU/l, >15 kU/l respectively) was determined for each of the groups A, B and C. Diagrams that visualised the amount of positive IgE responses in each allergen group facilitated a comparison between the general allergy profiles of the three groups.

4.4. Visualisation of data

IgE data has a high dimensionality with measurements on several allergens. In this study, the number of dimensions corresponds to the number of allergens. In order to reveal patterns and interrelationships in IgE data, it was desirable to visualise the multidimensional data in a reduced-dimension space.

4.4.1. Correlations

As a first step in studying the relationships between the allergens, correlations between each pair of allergens were calculated. In the mathematical descriptions below, consider an M*N IgE data matrix where M is the number of samples (patients' blood sera) and N the number of variables (allergens).

The correlation coefficients between all pairs of allergens were calculated with the Spearman rank order correlation (29). This correlation measure takes both linear and non-linear relationships between two variables into account. The idea is to rank all measurements within a variable and convert the data into rank order. The M measured IgE responses of one allergen are ranked according to their level and compared with the rankings of the allergen to which the correlation is calculated. The degree of similarity between the rankings of two allergens is translated into the correlation coefficient.

Spearman correlation values range from -1 to +1, where +1 reflects perfect correlation, 0 no correlation and -1 perfect negative correlation. The calculation of Spearman correlation coefficients on the IgE data matrix resulted in a symmetrical N*N matrix where N

is the number of variables (allergens) and the diagonal contains ones (see Appendix B, C or D for examples). Allergens with IgE measurements that co-vary to a large extent will obtain a high correlation coefficient.

4.4.2. Multidimensional scaling (MDS)

The objective of the data analysis was to visualise and capture as much as possible of the original distances between the allergens, modelled by correlations between them. Thus, the data reduction and visualisation was mainly performed with multidimensional scaling (MDS). The input to MDS is a matrix of distances or dissimilarities. The distance matrix was obtained by translating the correlations between the allergens into distances by calculating 1-(Spearman correlation rank coefficient). Thus, allergens that co-vary to a large extent and have a high correlation coefficient will obtain a small distance and consequently will be located close to each other in the resulting visualisation. The output of the MDS was a matrix where the original distances in N (the number of allergens) dimensions were reconstructed in two or three dimensions.

MDS was performed at the three subsets A, B and C respectively. The main study involved data sets with IgE responses to 30 allergen extracts from the plant kingdom (Table 2) containing no missing values. MDS was also used in the small study of component IgE data.

4.4.3. Evaluation of the MDS procedure

This section describes how the performance of the MDS procedure was evaluated. Classical MDS produces a set of coordinates in a reduced dimension, reconstructed from the distance matrix. The eigenvectors corresponding to the largest eigenvalues are used to reconstruct the data (35). Therefore, the performance of the reconstruction is dependant on the eigenvalues. If the eigenvalues are only positive, the classical scaling provides an exact reconstruction of the distance matrix. A distance matrix can generate negative eigenvalues. If the negative eigenvalues are small enough, a useful representation of the data is still obtained (35). However, if there is a large number of negative eigenvalues, or if some of them are large in magnitude, then the method may not suit the problem (35). If there are two or three eigenvalues that are much larger than the rest, it is possible to find a good reconstruction of the original distance matrix in two or three dimensions. When the first two eigenvalues constitute the major part of the total sum of all eigenvalues, they possess a good ability to reconstruct the distance matrix by themselves. This was measured by a simple calculation as follow:

$\frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i}$ where λ_i is the i:th eigenvalue

The resulting number can easily be translated into a percentage and can be interpreted as the degree to which the current reconstruction captures the original distances between the data points. A corresponding calculation for a 3D plot reveals if a third dimension is necessary for obtaining a useful representation in a reduced-dimensional space:

$$\frac{\lambda_1 + \lambda_2 + \lambda_3}{\sum_i \lambda_i}$$

If this number increases significantly when adding a third eigenvalue, it might imply that three dimensions are necessary to obtain a good reconstruction of the data.

The error of reconstructing the distance matrix by classical scaling was estimated by subtracting the Euclidean distances of the reconstructed coordinates from the original distance matrix and taking the maximum value:

$error = \max_{i,j} \left| D_{ij} - D_{ij\,recon}^{\,eucl} \right|$ where D is the original distance matrix and D_{recon}^{eucl} is the matrix of Euclidean distances between the reconstructed coordinates. When calculating the error of reconstructing the original data in two dimensions, the matrix D_{recon}^{eucl}, reconstructed with the first two eigenvectors, was subtracted from the original distance matrix. Similarly, the three-dimensional error was calculated by using the Euclidean distances reconstructed with the first three eigenvectors. The maximal error should be interpreted in relation to the original distance between the variables where the maximal error occurs.

4.4.4. Principal components analysis (PCA)

Principal components analysis was performed at IgE data in order to evaluate if this method could be useful for identification and visualisation of data for patients with different IgE response profiles. Using this method, the different groups of patients preferably group together. Three different approaches for pre-processing the data were used: logarithmic normalised raw data, logarithmic raw data and normalised raw data. The normalisation was carried out by subtracting the mean of each row from all values and subsequently dividing the values by the standard deviation of the row. Principal components analysis was performed at a data set containing samples of group A and B. Since PCA projects data along axes with maximal variance, the hope was to be able to separate the two groups in the resulting score plot, under the assumption that there was a difference between the groups with respect to their allergy profiles.

4.5. Missing values

IgE data contains missing values because the specific IgE response of some allergens was occasionally not measured in each blood sera sample. The missing values represent an information loss which is desirable to overcome. In addition, methods like MDS require a complete matrix. A simple way of dealing with missing values is to remove the entire rows with missing values. However, this results in a loss of useful information. A more sophisticated way to deal with missing values is to make use of a method that can predict their true values.

There are a few missing value estimation methods described in the literature which are widely used in the field of gene expression microarray data. The microarray data is usually in the form of large matrices of expression levels where rows are levels of genes and columns are different experimental conditions (34). In this project, these missing value techniques were applied at IgE data. Rows in the data are blood sera samples corresponding to genes in microarray data and the columns are different allergens corresponding to different experimental conditions in microarray data.

Different methods of filling missing values may lead to different results. Thus, two different imputation methods were tested on the IgE data in order to evaluate the usage of both methods: Bayesian principal component analysis (BPCA) and Local least squares imputation (LLS).

4.5.1. Bayesian principal components analysis (BPCA)

In the BPCA methodology, missing values are initialized with the row-wise average. Subsequently, a repetitive algorithm reestimates the missing values and model parameters using probabilistic models (25). Reestimation of the missing values involves principal components analysis performed at the observed values. The algorithm is repeated until it reaches a locally optimal solution. According to Oba et al. (25), the algorithm almost always converges to a single solution. There is no need to estimate model parameters separately which makes the algorithm easy to use.

4.5.2. Normalised root mean squared error (NRMSE)

The performance of missing value estimation is evaluated by normalised root mean squared error (NRMSE), calculated with the following formula (21):

$NRMSE = \frac{\sqrt{mean\left[\left(y_{guess} - y_{ans}\right)^2\right]}}{std\left[y_{ans}\right]}$ where y_{guess} is the vector with estimated values and y_{ans} is the vector with the known values.

The performance of the estimation is measured by using non-missing, known values and comparing them with the result of an estimation of them. The closer the NRMSE value is to 0, the more accurate is the missing value estimation. With a poor estimation or when the noise level is too high, NRMSE approaches a value of 1.0 (25). The NRMSE value was obtained as an output parameter from the Matlab functions used to estimate the missing values.

4.5.3. Local least squares imputation (LLS)

Local least squares imputation is widely used to estimate missing values in gene expression data. Missing values of a target sample are calculated using values from a set of similar genes. The similar genes are chosen as the K nearest neighbours with respect to their correlation coefficients or the L_2-norm (Euclidean distances) (21). With IgE data, the K most similar genes correspond to the K most similar samples or in practice, the K individuals having the most similar allergy profile. The parameter K is chosen by repeating the estimation using several K-values, and the one that maximizes the performance of the estimation is chosen (21), i.e., the K-value that yields the minimum NRMSE. After choosing the K most similar genes, the second step is regression and estimation.

4.5.4. Simulation of missing values

Missing values were simulated in order to evaluate the LLS and BPCA method for estimating missing values and study the behaviour of the MDS method as response to different levels of missing values in input data. As a starting-point, the subsets A and B were formed and all samples with any missing values were removed. Different percentages of missing values were studied and a certain percentage of missing values was obtained by removing a corresponding number of measurements randomly from the subsets. For each percentage of missing values, the same data points were removed and set to missing as both missing value estimation methods were evaluated. Since the same data sets were used as starting-points for the simulation of missing values, the behaviour of the MDS plots based on data with different amounts of missing values could be compared. Even though the missing values can be filled with an estimated value, the amount of missing values should not be too high to achieve a valid statistical analysis The aim of simulating different levels of missing values was to come up with some guidelines as to which amount of missing values that can be permitted in order to achieve a valid statistical analysis. These guidelines are presented in the result section.

4.6. Simulation of measurement noise

Due to variability in the allergy tests, each IgE level measurement contains a measurement noise. At Phadia, the noise is usually described with the coefficient of variance, CV, and is used to assess the precision of the measurements. The coefficient of variance can be described as the degree to which a set of data points varies. It is often displayed in percentage and the lower percentage, the lower variability between measurements. The CV (in percentage) is calculated as follow:

$CV = \frac{\sigma}{\bar{x}} \cdot 100$, where σ is the standard deviation and $\bar{x}$ is the mean (average) of data points

Phadia's test instruments usually have a total over-all CV less than 16 %. This includes errors and measurement noise from all kinds of sources. However, this figure varies between different allergens, that is, the measurement of different allergens can have different CVs due to, for instance, a variation in binding affinity of the ImmunoCAP^{TM}.

Different levels of measurement noise were simulated by varying the size of the CV in a range between 4 % and 24 % and adding a randomly drawn number from a normal distribution based on the CV.

The IgE data in the internal database does not contain any replicates; each sample has only one measurement for each allergen. Therefore, the assumption that the one measurement represents the mean value of a number of replicates was made. Under this assumption, the estimated standard deviation of the observed value was calculated by multiplying the CV with the observed value x:

$\sigma_{est} = CV \cdot x$

Subsequently, a random number r was drawn from a normal distribution with mean m = 0 and the standard deviation σ_{est} and added to the observed value x to simulate the measurement noise:

$x_{noise} = x + r \qquad r \sim N(0, \sigma_{est})$

Multidimensional scaling was applied on the data with introduced measurement noise in order to study the impact of different levels of measurement noise on the results.

4.7. Simulation of data loss

In order to evaluate if the size of the data sets were sufficient to obtain a valid data analysis, 20% of the samples in the data were randomly removed. The results of the reduced data sets were subsequently compared to the results of the original data set. If the results changed significantly, the conclusion that the data set is too small could be drawn.

4.8. Software

The data analysis was performed in Matlab 7.1 with Statistics toolbox. Statistics toolbox contains ready-to-use functions for statistical analyses such as MDS and PCA. Both of the missing value estimation algorithms; BPCA and LLS, were available as Matlab toolboxes on the internet. The packages were downloaded from links given in the articles presenting the algorithms (21, 25).

Functions in Matlab facilitated the import of the Excel input files containing the IgE data. The statistical methods were then applied at the data and the results were displayed in plots generated by Matlab.

5. Results

5.1. Differences between the groups

One of the main aims of this degree project was to study the IgE reactivity patterns in the three allergy groups A, B and C with different combinations of sensitisations to grass pollen and wheat. This was done mainly by applying the visualisation technique MDS on IgE data from these groups. The IgE data contained IgE responses to allergen extracts. All samples containing at least one missing value were removed from the data sets before the data analysis was carried out. The results presented here aim at revealing what relationships that exist between allergen extracts in each of the three groups, as well as their over-all sensitisation profile. This will hopefully help to clarify the relationships between IgE reactivity to grass pollens, cereal grains and plant-origin foods.

5.1.1. Group A: Patients with positive IgE responses to wheat and grass pollens

Individuals of this group can be regarded as sensitised to grass pollen and wheat. When the samples were selected according to the criteria for this group and samples with missing values were removed, 465 samples ended up in the resulting data set. First, multidimensional scaling (MDS) was performed at this data set, including the selected 30 allergens of plant foods and grass pollens (Table 2). The resulting two-dimensional plot of the first two eigenvalues displays the interrelationships among the allergens, captured in the two-dimensional distances between them. A clear separation between the plant foods and the grass pollens can be seen in Figure 6. This indicates that the correlation coefficients within the grasses and plant foods respectively are relatively higher than the correlation coefficients between the two groups. All pair-wise correlation coefficients for the 30 allergens can be found in Appendix B.

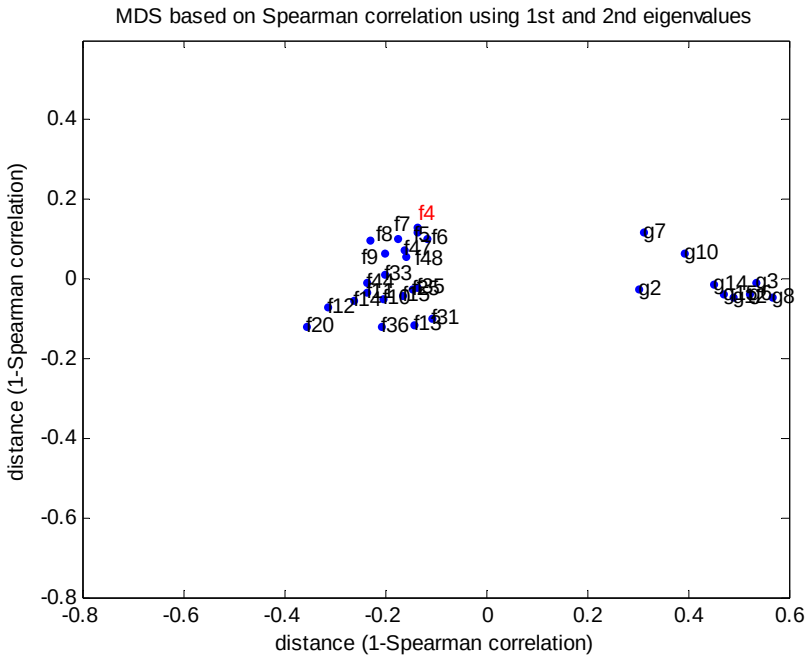


Figure 6. MDS plot of 21 food allergens (f) and 9 grass pollen allergens (g) for group A, based on samples from 465 individuals. Wheat (f4) is marked with red.

In order to obtain a higher resolution of the plant food group, MDS was performed on the same data set again, including the 21 plant food allergens only. The allergens in the resulting MDS plot of the 21 plant food allergens are strongly correlated (Figure 7).

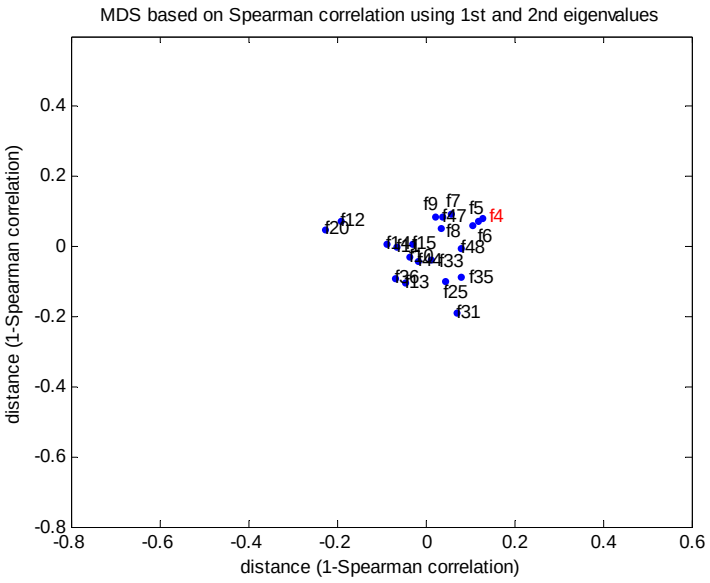


Figure 7. MDS plot of 21 food allergens for group A, based on samples from 465 individuals.

Allergen code	Common name	Spearman correlation to wheat
f4	Wheat	1,00
f5	Rye	0,95
f6	Barley	0,93
f7	Oat	0,89
f9	Rice	0,86
f8	Maize	0,85
f48	Onion	0,84
f47	Garlic	0,84
f33	Orange	0,80
f35	Potato	0,79
f11	Buckwheat	0,79
f44	Strawberry	0,78
f14	Soya bean	0,78
f25	Tomato	0,77
f10	Sesame	0,75
f15	White bean	0,74
f36	Coconut	0,71
f31	Carrot	0,71
f13	Peanut	0,68
f12	Pea	0,67
f20	Almond	0,63

Table 4. Spearman correlation coefficients between each of the food allergens and wheat in group A.

All of the included plant food allergens have high correlation coefficients to wheat (Table 4), which is illustrated by the short distances between the plant foods and wheat. The correlation coefficients correspond well to the distances in the plot. Onion (f48) and garlic (f47) with the largest correlation coefficients to wheat among the plant foods are located close to wheat (f4), whereas pea (f12) and almond (f20) have the largest distance to wheat and the smallest correlation coefficients. Onion (f48), garlic (f47), orange (f33), potato (f35), tomato (f25) and carrot (f31) are not biologically close related to wheat and therefore their high correlation to wheat was somewhat unexpected. The unexpectedly high correlation between onion and wheat is further illustrated in a plot in which the samples of group A were sorted by the IgE level of wheat and the logarithms of the values were plotted together with the IgE levels of onion for all 465 samples (Figure 8).

The patterns of group A, i.e. the high correlation between IgE levels of all plant food allergens, reflect that the patients in this group are sensitised to more allergens than wheat and grass. However, the cause of these multi-reactive patterns cannot be revealed so far.

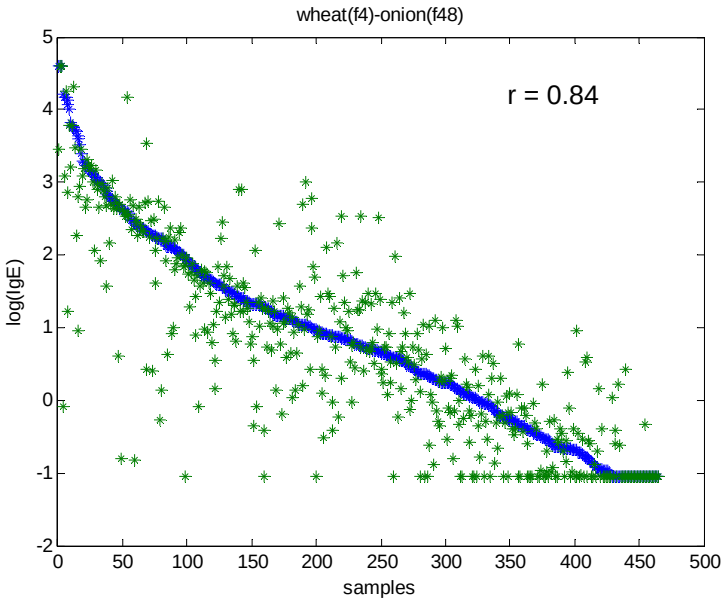


Figure 8. The IgE levels of onion (green) and wheat (blue) plotted in the same diagram. The y-axis shows the logarithm of the IgE level and the x-axis the 465 samples.

5.1.2. Group B: Patients with positive IgE responses to wheat and negative IgE responses to grass pollens

Individuals of this group can be regarded as sensitised to wheat, but not to grass pollens. When the samples were selected according to the criteria for this group and samples with missing values were removed, 53 samples ended up in the resulting data set. The patterns in the MDS plot including the chosen 30 allergens of plant foods and grass pollens, using the first two eigenvalues, differed significantly from that of group A. In contrast to group A, the correlations to wheat were generally much lower for allergens not closely biologically related to wheat (Figure 9). The plant food allergens that group closest to the grasses are those which had unexpectedly high correlations to wheat in group A (in light green Figure 9). All pair-wise correlation coefficients for the 30 allergens can be found in Appendix C.

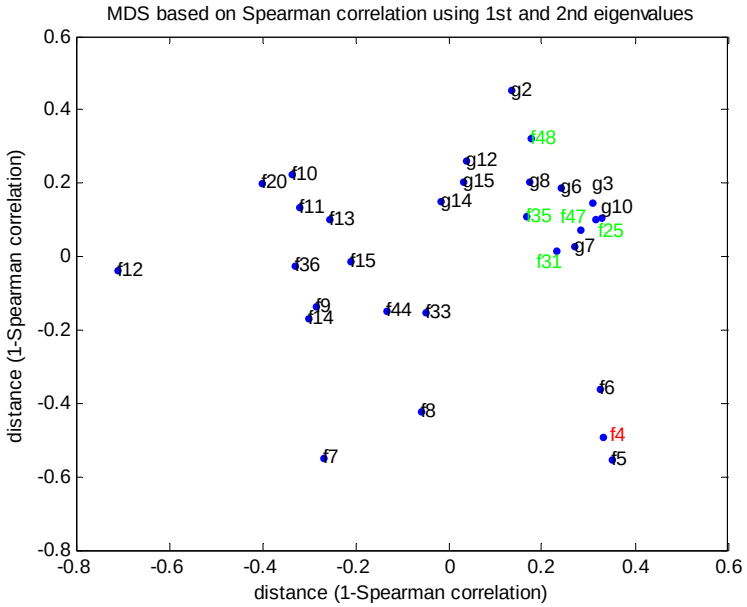


Figure 9. MDS plot of group B with 21 plant food allergens and 9 grass pollen allergens. The six food allergens that had unexpectedly high correlations to wheat in group A (compare with Figure 7) are coloured in light green.

When MDS was performed on the same data set, including only the 21 plant food allergens, the allergens ended up scattered in the resulting two-dimensional plot of the first two eigenvalues. Using the same axes in the two-dimensional plot, it is clear that the MDS plot of the 21 plant food allergens shows a different pattern than the MDS plot of group A (Figure 7 and Figure 10).

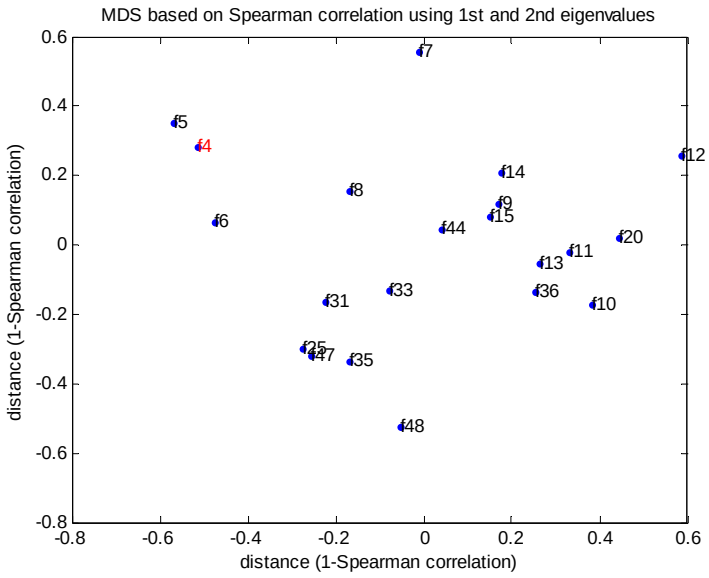


Figure 10. MDS plot of 21 food allergens for group B, based on samples from 53 individuals.

Allergen code	Common name	Spearman correlation to wheat
f4	Wheat	1,00
f5	Rye	0,78
f8	Maize	0,52
f6	Barley	0,47
f7	Oat	0,36
f31	Carrot	0,35
f44	Strawberry	0,33
f25	Tomato	0,29
f14	Soya bean	0,21
f13	Peanut	0,17
f9	Rice	0,17
f35	Potato	0,16
f47	Garlic	0,13
f33	Orange	0,11
f15	White bean	0,10
f11	Buckwheat	0,08
f48	Onion	0,05
f36	Coconut	0,04
f10	Sesame	-0,02
f20	Almond	-0,08
f12	Pea	-0,09

Table 5. Spearman correlation coefficients between each of the food allergens and wheat in group B.

The correlations to wheat are weak in general (Table 5). This is well reflected in the plot since all the allergens are scattered. For example, onion and garlic that had strong correlations to wheat in group A have a much weaker correlation to wheat in this group. Interestingly, the cereal grains group somewhat together in the upper left corner whereas the six food allergens that had a particularly and unexpectedly high correlations in group A group somewhat together in the lower mid-part of the plot (Figure 10).

The generally low correlation coefficients between allergens in this group could reflect that these patients are mono-sensitised, i.e. sensitised to only one or a few allergens.

5.1.3. Group C: Patients with negative IgE responses to wheat and positive IgE responses to grass

Individuals of this group can be regarded as sensitised to grass pollen but not to wheat. Like group B, this was a small group of individuals with only 74 samples. MDS could not be performed when one or more of the columns of the matrix contained values all equal to 0.35. The reason for this is that the values cannot be ranked and the correlation cannot be calculated which results in an invalid distance matrix. Due to this, the cereal grain allergens together with orange and coconut could not be included in the plot of group C since all of those IgE responses were equal to 0.35.The patterns in the MDS plot including the chosen 30 allergens of plant foods and grass pollens except the four cereal grains, orange and coconut, show a

clear separation between the plant food allergens and the grass pollen allergens, similar to that of group A (Figure 11). However, the plant food allergens were scattered in contrast to group A (Figure 11). All pair-wise correlation coefficients can be found in Appendix D.

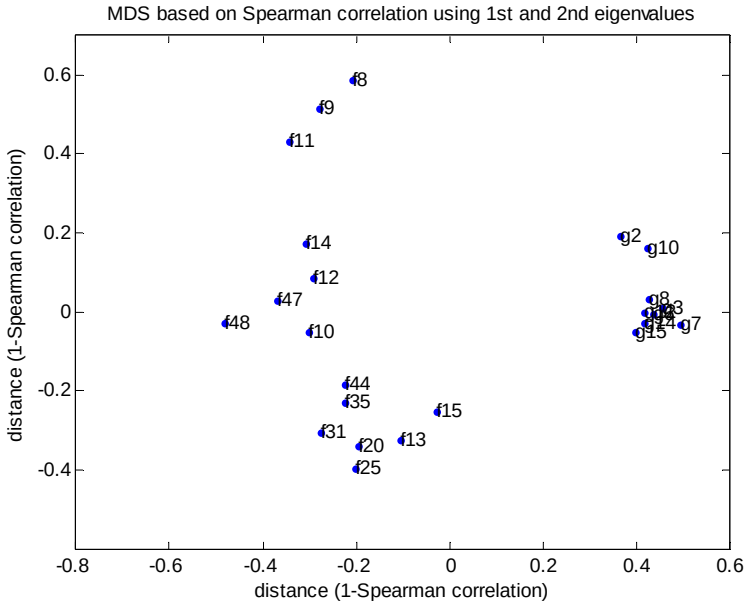


Figure 11. MDS plot of group C (74 samples) including 15 plant food allergens and 9 grass pollen allergens.

The MDS plot including the 21 plant food allergens except the four cereal grains, orange and coconut show a similar, scattered pattern of group B (Figure 12).

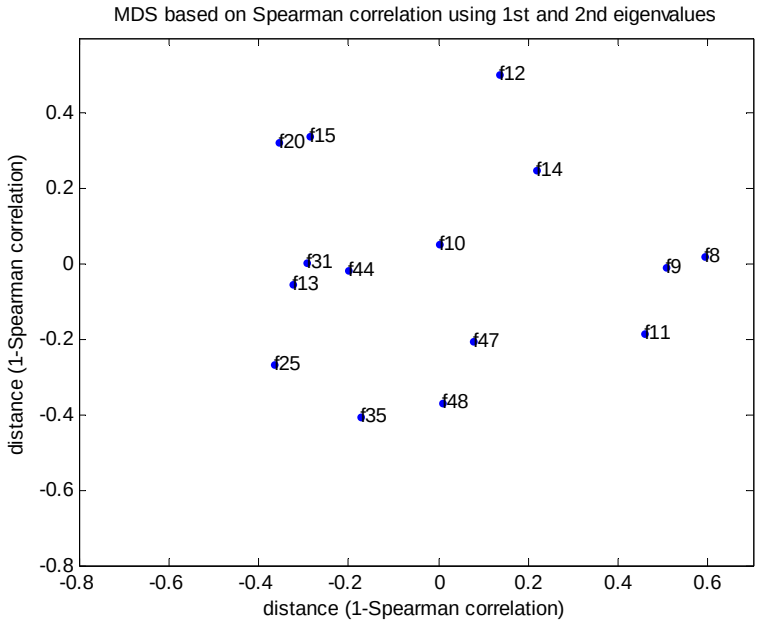


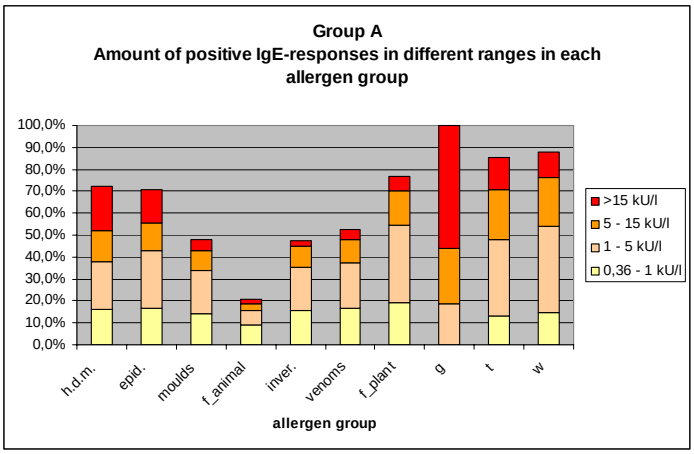
Figure 12. MDS plot of group C (74 samples) including 15 plant food allergens.

In conclusion, group C show similarities to both group A and group B.

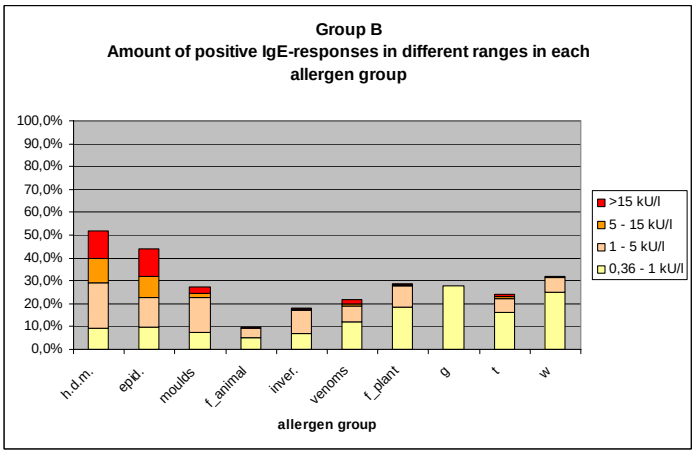
5.1.4. Allergy profile of the groups

Data of all 93 allergens of group A, B and C was explored in order to illustrate general differences between the groups. In particular, one aim was to investigate if the IgE levels and distribution of positive tests could explain the differences in the patterns seen in the MDS plots in section 5.1.1-5.1.3. In the resulting diagrams, the allergens are grouped and the IgE responses divided into four intervals. The diagrams show that the IgE levels of group A are generally higher than of group B and C (Figure 13).

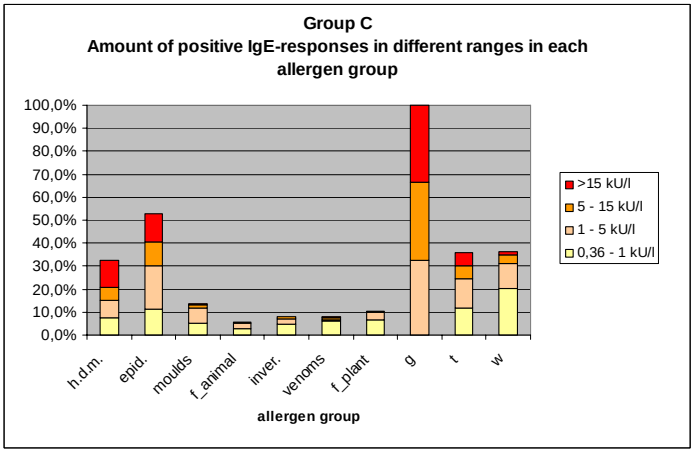
a)



b)



c)



Allergen groups

h.d.m.	house dust mites
epid.	epidermals
moulds	moulds
f_animal	foods of animal origin
inver.	invertebrates
venoms	venoms (insect poison)
f_plant	foods of plant origin
g	grass pollens
t	tree pollens
w	weed pollens

Figure 13. General allergy profile of the three groups represented by staple diagrams with positive IgE responses in four intervals for ten allergen groups. a) Group A. b) Group B. c) Group C.

In general, the total percentage of positive IgE responses in group A is high for all of the allergen groups except foods of animal origin (Figure 13 a) which implies that the individuals in this group could be multi-sensitised or react to cross-reactive components. Group B and C have a much lower total amount of positive IgE responses compared to group A (Figure 13 b and c). However, the amount of positive IgE responses in the groups of dust mite and epidermals in group B and C are not as low as in the other allergen groups. This means that sensitisation to these allergen groups is not affected neither by grass pollen nor cereal grains sensitisation. In the plant food group, there is an approximately 20 % difference in the amount of positive IgE responses in group B compared to group C. This could imply that sensitisation to cereal grains is more likely to give rise to a higher prevalence of plant food sensitisation than grass pollens.

5.1.5. The impact of IgE levels

Exploring the data (Section 5.1.4) showed that the IgE levels of group B were generally lower than the IgE levels of group A. For example, onion was one food allergen where the both groups differed particularly much in the levels of IgE responses (Figure 14). The question arose if the differences seen in the patterns of the MDS plots of group A and B were due to general differences in IgE levels rather than the selection criteria of the groups. Was the real difference between group A and B that group A had much higher IgE levels in general rather than group A being sensitised to grass and B not being sensitised to grass? In order to answer this question group A was partitioned in two parts: one with onion-levels above 2 kU/l (green square Figure 14) and the other with levels below 2 kU/l (yellow square Figure 14). Onion was chosen as a basis for forming new subsets because the IgE responses to onion differed much between group A and B (Figure 14). Besides, onion had a strong correlation to wheat in group A and a weak correlation to wheat in group B. Due to lack of time, no other food allergens were studied. MDS was performed at the two subgroups with IgE responses above and below 2kU/l respectively and the plots were compared with the MDS plot of group A (Figure 15).

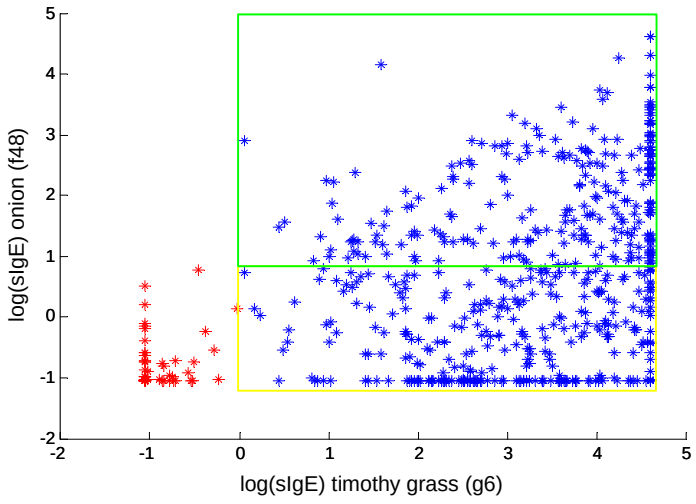


Figure 14. Samples of group A (red) and group B (blue) plotted in the same diagram with respect to their logarithmic IgE response to timothy grass and onion. The green square encloses individuals in group A with IgE responses to onion above 2kU/l. The yellow square encloses individuals in group A with IgE responses to onion below 2kU/l.

The MDS plots in Figure 15 illustrates that a lower IgE level of onion implies a slightly more scattered plot. However, the overall patterns of group A remain, namely the dense grouping of all the food allergens. Therefore, the conclusion is that higher IgE levels in group A compared to group B cannot explain the general differences between the MDS plots.

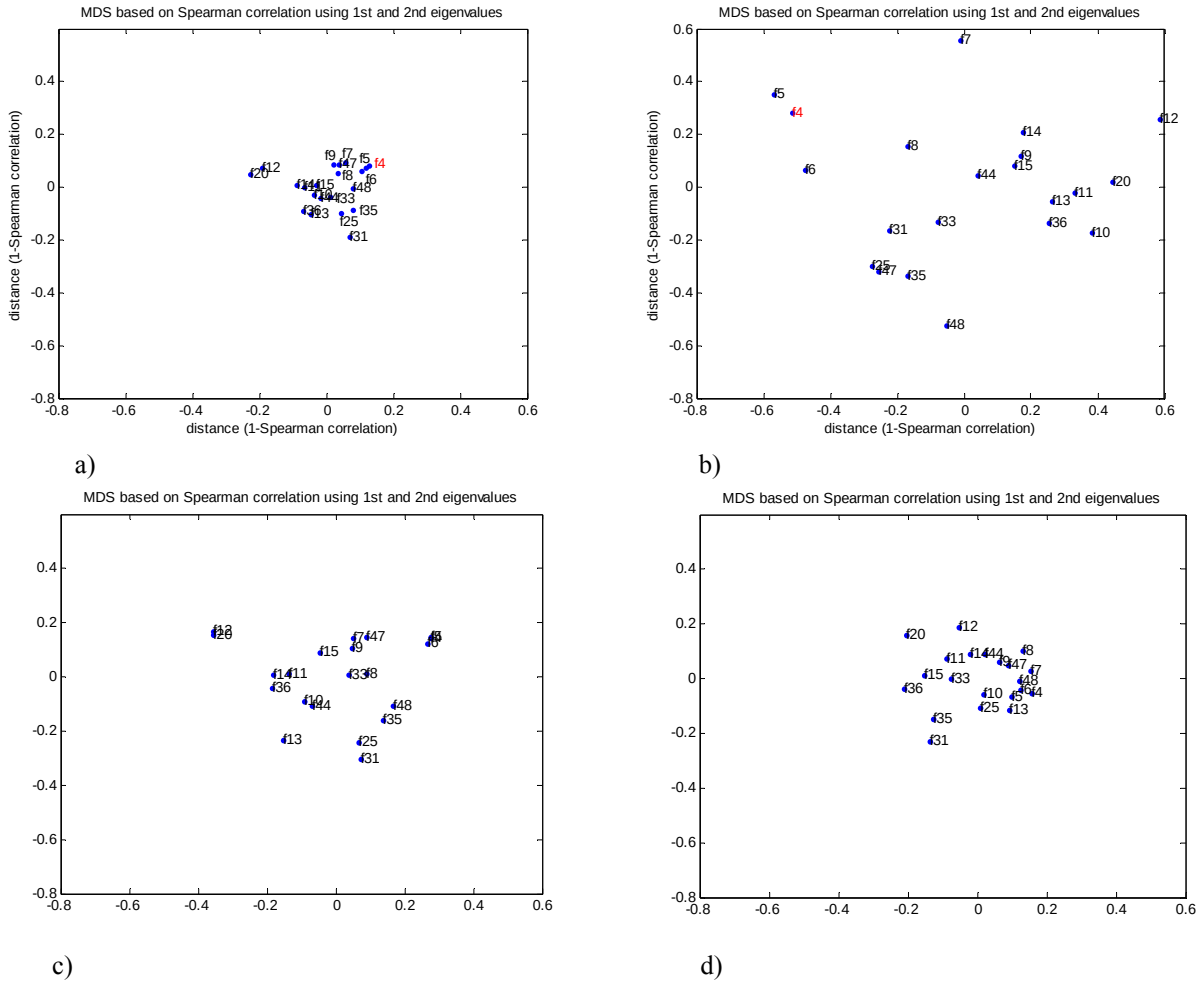


Figure 15. MDS plots of group A divided in two parts with respect to the IgE levels of onion. a) MDS plot of group A. b) MDS plot of group B. c) MDS plot of group A including only individuals with IgE responses to onion below 2kU/l. d) MDS plot of group A including only individuals with IgE responses to onion above 2kU/l.

5.2. Results of component study

The results presented here aim at identifying the component that might be involved in the multi-reactivity patterns seen in group A. MDS was performed at the data set comprising 34 samples with IgE responses to extracts as well as components in order to reveal what components that correlated mostly to the allergen extracts. Since all of the 34 samples belonged to group A, these results could explain what component that might be responsible for the patterns seen in this group.

The 21 plant food allergens and the 9 grass pollen allergens used in the study of extract data, and four allergen extracts rich in CCD were included in the MDS plots together with the 7 components. As seen in Figure 16, bromelain (denoted 'bromelin') groups together with the plant food allergens. This reflects a high correlation between CCD and the plant foods, i.e. that the IgE levels of CCD and the plant foods co-vary, suggesting that CCD could be involved in the multi-reactive patterns seen in group A. As expected the major allergen components in timothy grass, Phl p 1, Phl p 4 and Phl p 5 group together with the grasses (Figure 16 c). All correlation coefficients can be found in Appendix E.

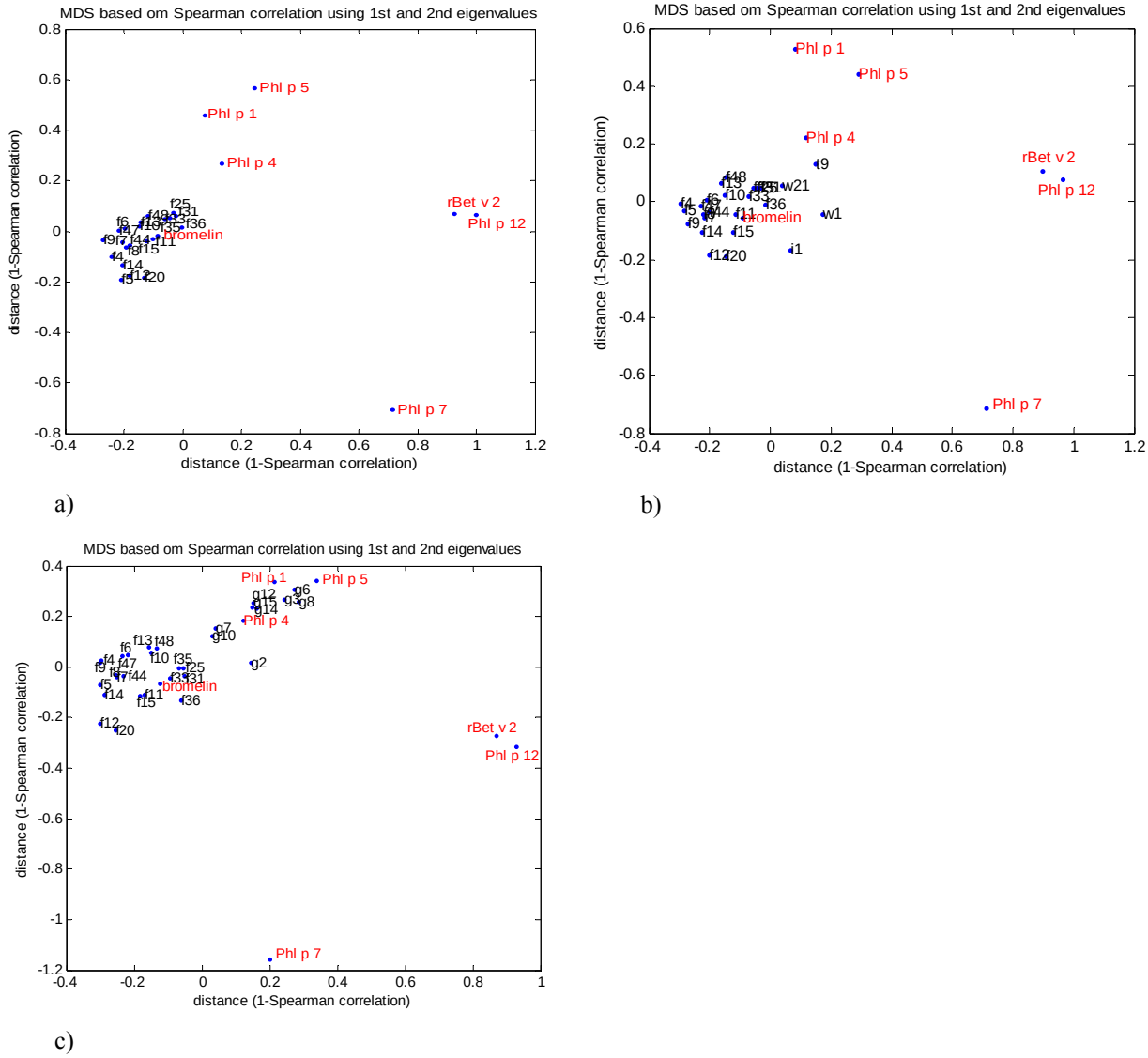


Figure 16. MDS plots of group A including 34 individuals. a) 21 plant foods and 7 components. b) 21 plant foods and the four CCD rich allergens t9 (olive tree), i1 (bee), w1 (common ragweed) and w21 (wall pellitory) together with the 7 components. c) 21 plant foods, 9 grass pollens together with the 7 components.

5.3. Evaluation of methods

One of the main goals of this work was to identify and evaluate methods in pattern recognition that could be applied on IgE data in Phadia's internal database. PCA and MDS were tested as possible methods for visualising IgE reactivity patterns. It turned out that MDS was the most useful method for visualising patterns in IgE data and therefore, the main part of the evaluation is focused on MDS.

5.3.1. Principal components analysis (PCA)

In the initial phase of the project, PCA was tested as a possible method to visualise discrepancies between the groups A and B. Since PCA projects the samples along the axis of maximal variance in the data set (Section 3.6.1), the expectation was to obtain a score plot in which the both groups were separated from each other. Three different approaches to pre-process data were tested, and the best separation of the groups was obtained using normalised raw data. The resulting score plot shows a certain separation in the first component axis (horizontal), but not as clear as one could have expected. The loading plot displays the relations between the allergens and provides information about what allergens that have large loadings that is, those allergens that have a significant influence on the scores. As can be seen in Figure 17, the grasses far to the right in the loading plot account for a significant influence on the horizontal separation of the groups. This result is expected since the difference between the groups is their IgE responses to grass pollens.

However, this method was not used in further studies since no new information or any clear patterns were obtained.

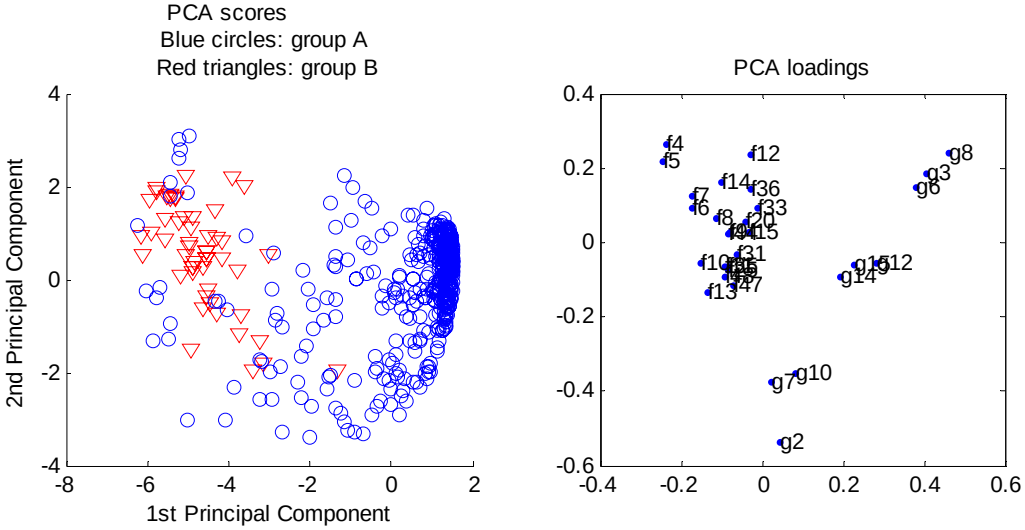


Figure 17. Results of principal components analysis performed at a data set containing samples of group A (red triangles) and B (blue circles). Normalised raw data was used.

5.3.2. Missing values

Two methods for estimating missing values; BPCA and LLS, were evaluated and the influence of missing values on the MDS method was studied.

Subsets A and B formed the basis for simulation of missing values. Different percentages of missing values were randomly introduced into the subsets and the behaviour of the MDS plots was studied to reveal the critical level of missing values for both of the groups. Since the size of subset A was larger than the size of subset B, the impact of missing values on the MDS plots was more prominent at lower levels for subset B than for subset A.

Subset A (465 samples) was evaluated at the levels 5, 10, 15 and 20 % of missing values. Since the missing values were introduced randomly, three runs were performed at each percentage and the mean NRMSE of these three runs was calculated. The NRMSE values and the mean of the two missing value estimation methods when applied at data with varying percentage of missing values can be seen in Figure 18.

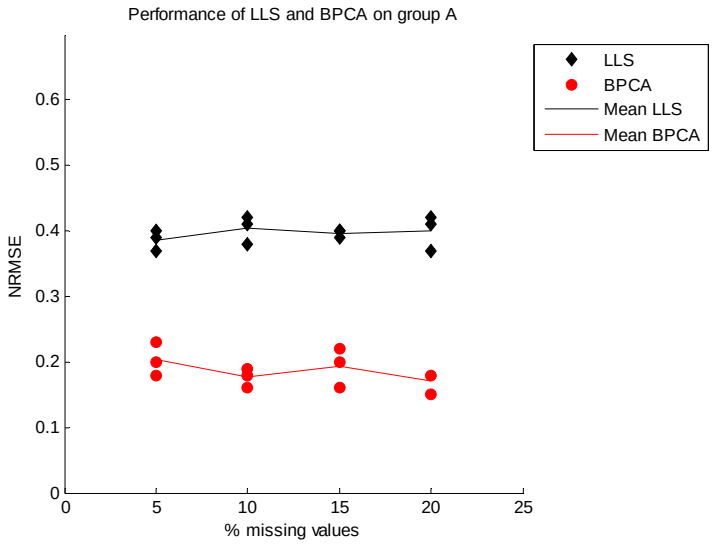


Figure 18. The NRMSE value of LLS and BPCA imputation at different levels of missing values in group A.

Figure 18 shows clearly that the NRMSE of BPCA is lower than of LLS at all levels of missing entries in subset A. However, the behaviour of the NRMSE value did not reveal what the critical percentage of missing values was. In addition, it appeared a bit strange that the NRMSE decreased for BPCA as the amount of missing values increased. Thus, it was better to evaluate the MDS plots in order to find a critical level for the amount of missing values. When the appearance of the plot changed significantly, the critical level of missing values was regarded as reached.

The MDS plots of subset A with a high percentage of simulated missing values did not differ much from the original MDS plot of subset A. In Figure 19 it is clear that at a level of 20 % missing entries in subset A, the MDS plots did not change their appearance to a large extent compared to the original plot.

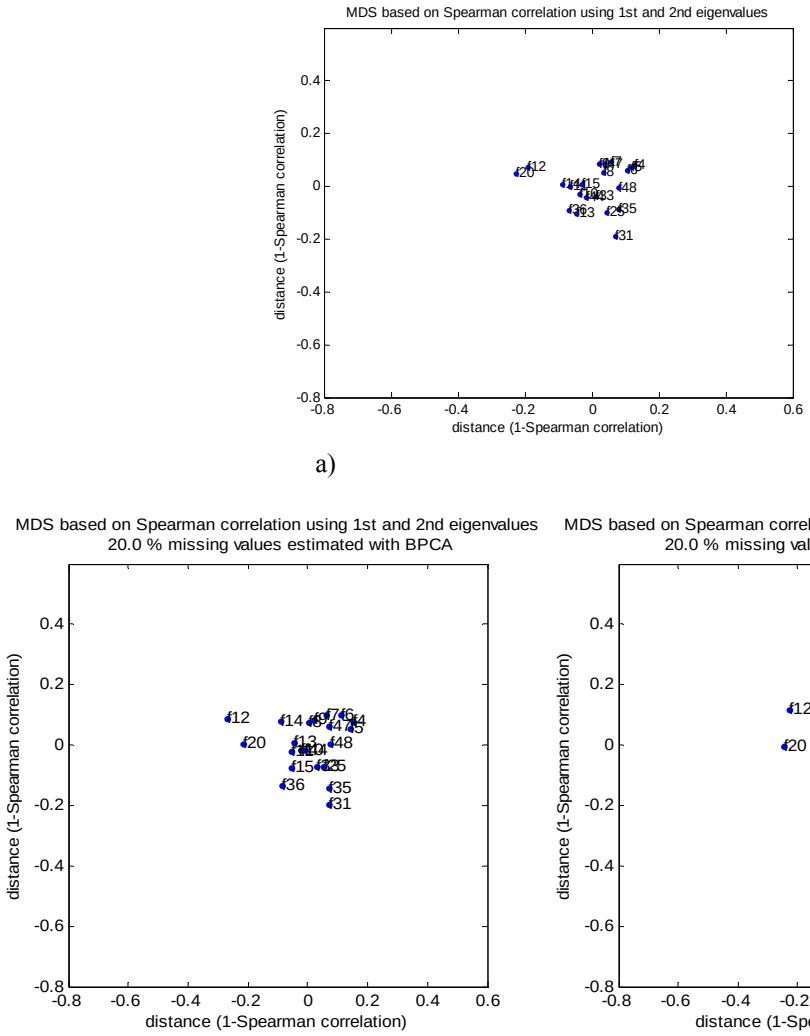


Figure 19. Missing values introduced in subset A and estimated with Local Least Squares (LLS) and Bayesian Principal Component Analysis (BPCA). a) Original plot of group A based on data with no missing values. b) Group A with 20 % artificial missing values estimated with BPCA and LLS respectively. NRMSE for this particular run was 0.18 for BPCA and 0.38 for LLS.

Compared to the MDS plot with no missing values, the plots based on data with 20 % missing values are very similar regardless of the missing value estimation method used. These results indicate that a level of 20 % missing values is acceptable in a subset of the same size as subset A.

Subset B (53 samples) was evaluated at the missing value levels 2.5, 5, 7.5 and 10 %. Since the NRMSE value varied much between individual runs within each percentage, six runs were performed for each percentage and the mean NRMSE was calculated. The NRMSE of the two missing value estimation methods when applied at this subset can be seen in Figure 20.

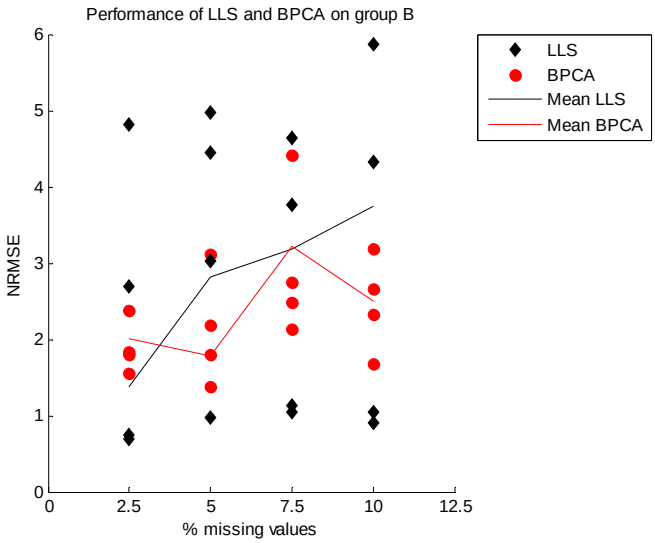


Figure 20. The NRMSE value of LLS and BPCA imputation at different levels of missing values in group B.

The NRMSE values for the missing value estimation of subset B are extremely high and variable for both methods. The reason why the NRMSE vales are high above 1 could not be revealed although it is likely that the small size of data set B is one cause. Even though the results imply that the NRMSE values are unreliable, there is a weak tendency towards a lower NRMSE for the BPCA method. However, a higher accuracy of estimation with BPCA was not observed in the MDS plots and any discrepancy between the two methods could not be revealed. At a level of 2.5 % missing values, the MDS plots based on both methods are similar to the MDS plot based on no missing values (Figure 21).

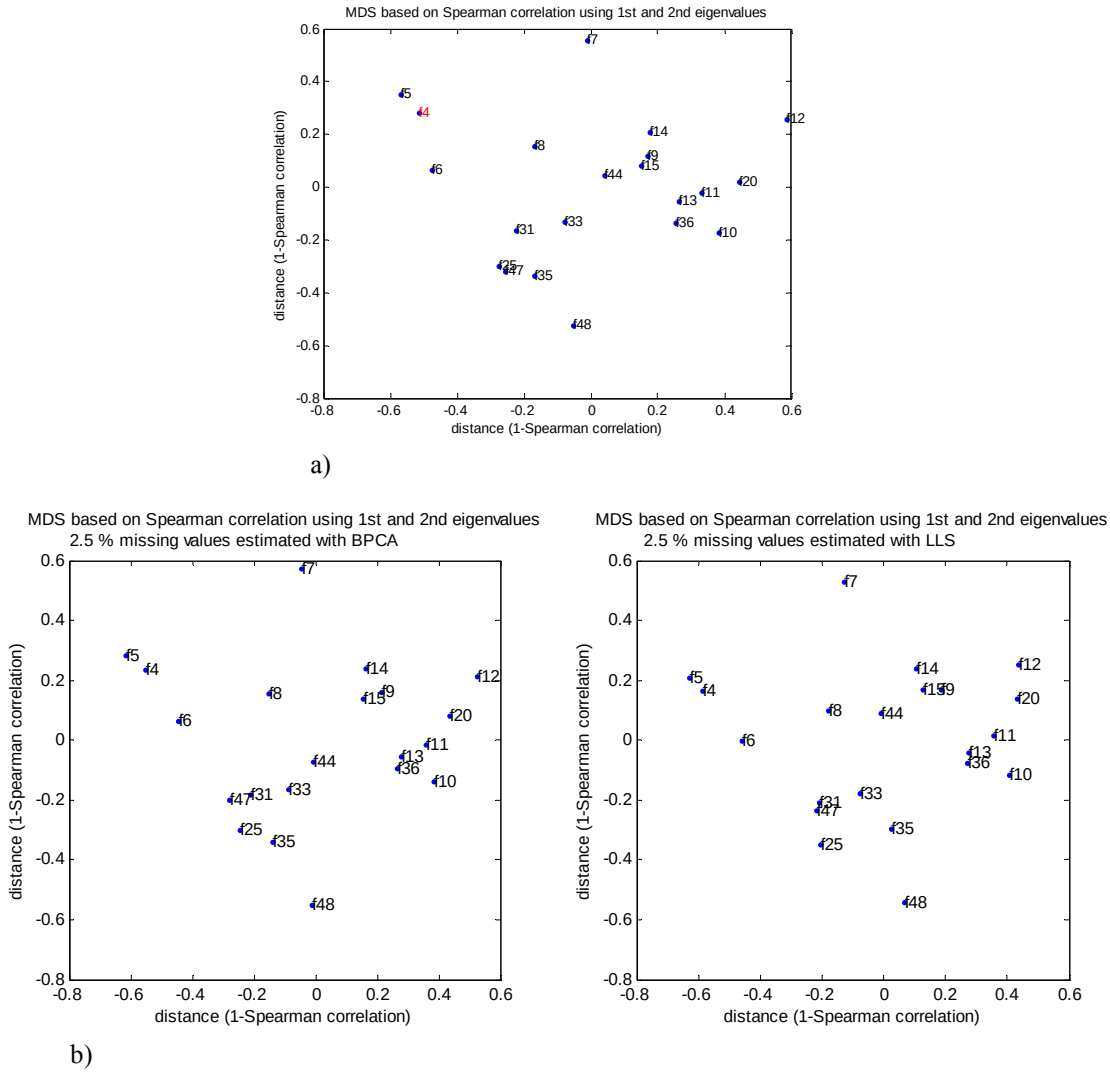


Figure 21. Missing values introduced in subset B and estimated with Local Least Squares (LLS) and Bayesian Principal Component Analysis (BPCA). a) Original plot of group B based on data with no missing values. b) Group B with 2.5 % artificial missing values estimated with BPCA and LLS respectively. NRMSE for this particular run was 1.55 for BPCA and 4.82 for LLS.

The results of both estimation methods did not differ significantly for group B at any of the levels of missing values chosen to study. At the level of 5 % missing values the plots started to change their appearance slightly compared to the plot based on no missing values, but the performance of the both methods was very similar. At the level of 7.5 % missing values, the appearance of the plots varied to a larger extent between individual runs and there was often a difference between the two plots based on the both methods. However, the overall pattern of scattered allergens in group B did not change at any level of missing values.

The difference in performance between the two missing value estimation methods was little, but the implementation in Matlab differed between them. First of all, the execution time of BPCA was much longer than for LLS, especially for the large data set of group A. One advantage of BPCA over LLS was that the Matlab functions were easier to call and to implement.

In conclusion, a level up to 20% of missing entries is acceptable for group A and a level up to 5 % for group B. However, 20% missing values is intuitively very much even though the MDS procedure displays a robustness at this level.

5.3.3. Measurement noise

Measurement noise was simulated in the subsets A and B in order to evaluate the impact of variability of the test instruments on the MDS method. The measurement noise was introduced according to the method described in section 4.6 at the two subsets without any missing values present. The changes in the MDS plots were studied at five different levels of CV for both subsets. The results for group A show that the MDS plots does not change much up to a CV at 24 % (Figure 22).

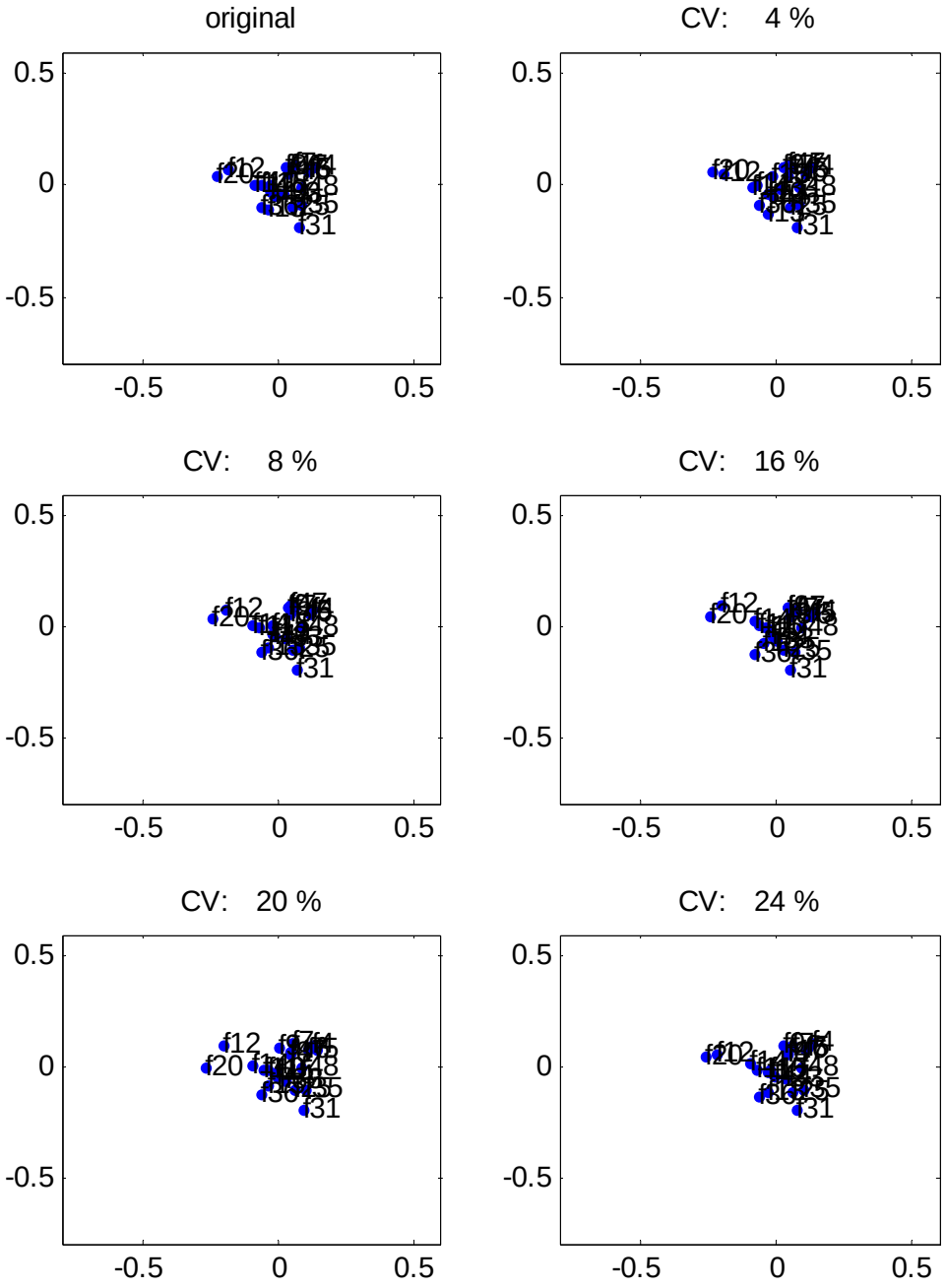


Figure 22. Simulation of measurement noise on group A at five levels of CV. Due to lack of space, the axes are not labelled. The axes represent the distance 1 – Spearman correlation and the MDS plots are based on the first two eigenvalues. The original MDS plot without introduced noise is located in the top left corner.

Measurement noise had different impact on the two groups. In group B, the impact of noise was more prominent. The appearance of the MDS plot changed significantly at a CV of 8 %. However, the over-all patterns with scattered allergens remained for all five levels of CV (Figure 23).

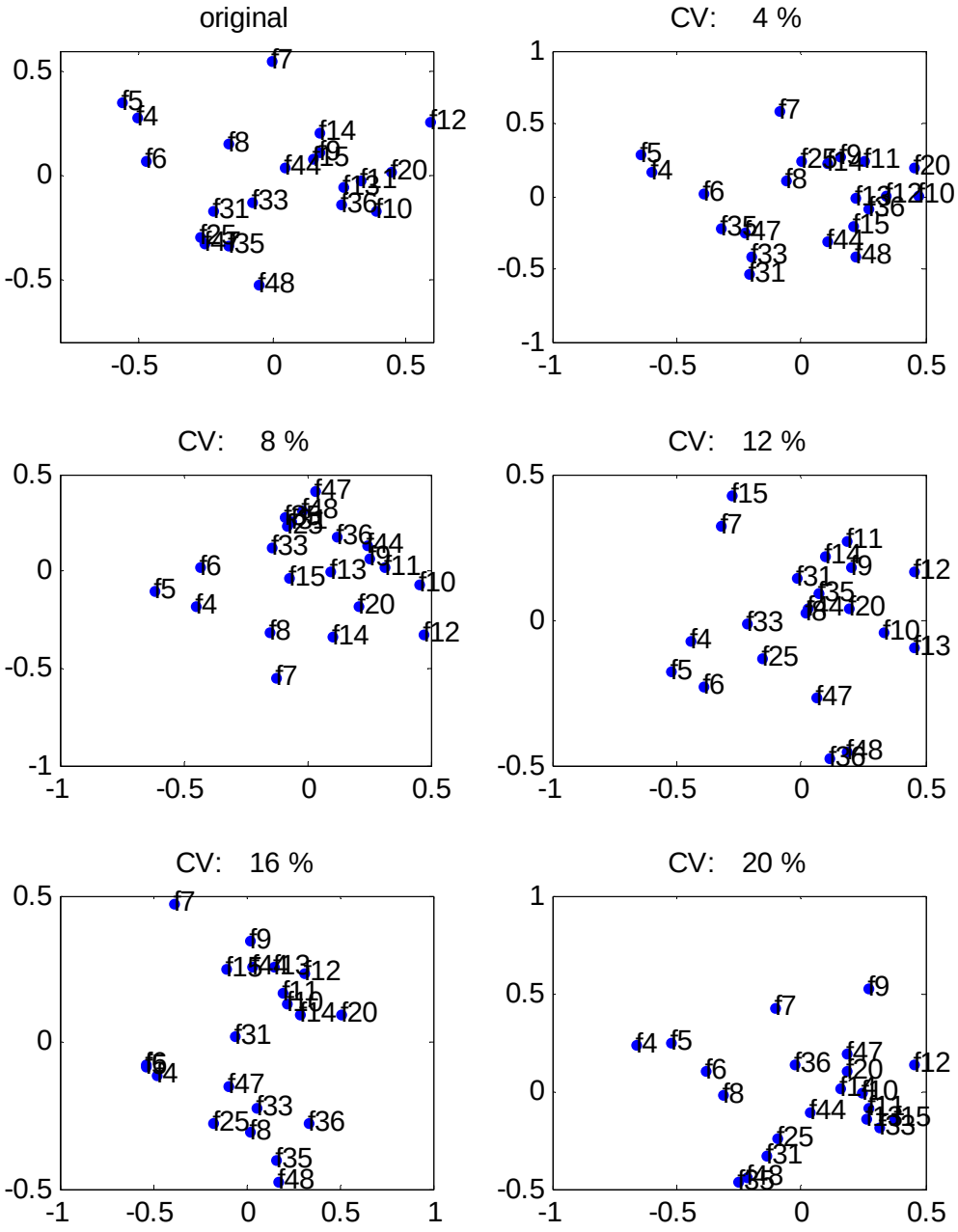
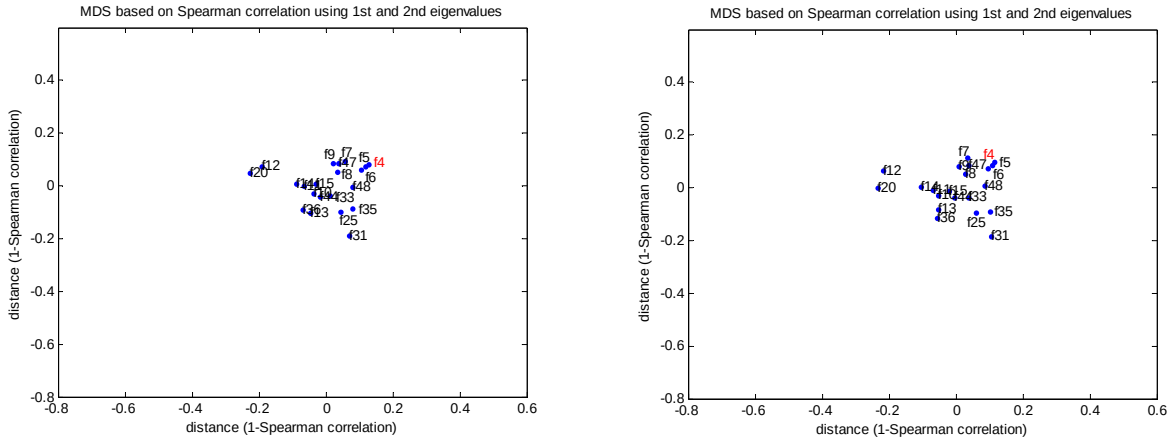


Figure 23. Simulation of measurement noise on group B at five levels of CV. Due to lack of space, the axes are not labelled. The axes represent the distance 1 – Spearman correlation and the MDS plots are based on the first two eigenvalues. The original MDS plot without introduced noise is located in the top left corner.

5.3.4. Simulating loss of data

In order to explore if the size of the subsets was sufficient to obtain a valid data analysis, 20 % of the data was removed and MDS was performed on the reduced data set. Subsequently, the MDS plots were compared with the original MDS plots in order to see if they changed their appearance.

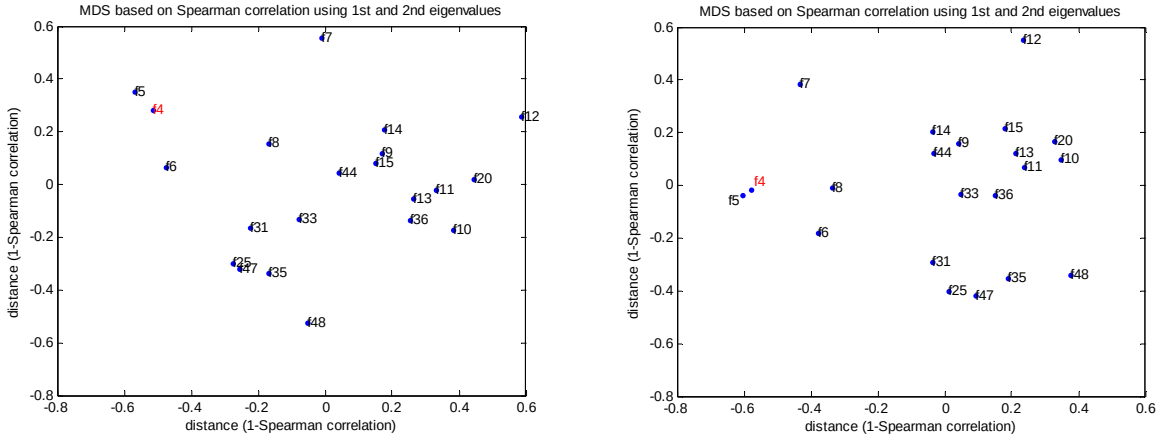
Subset A comprised 465 samples when rows with missing values were removed. This subset turned out to be robust to changes and it was therefore not surprising that removing 20 % of data changed the appearance of the plot only to a slight extent (Figure 24).



a) b)

Figure 24. MDS plots of group A. a) No missing values and 465 samples. b) 20 % of data removed, resulting in 372 samples.

Subset B comprised 53 samples when rows with missing values were removed. The mere size of this data set was a strong incentive to study the changes occurring when removing 20 % of the samples. In spite of the small size of the data set, the MDS plot did not change as much as expected when 20% of the samples were removed (Figure 25). Even though some inter-relationships were changed, the plot did not change its over-all appearance with scattered allergens.



a) b)

Figure 25. MDS plots of group B. a) No missing values and 53 samples. b) 20 % of data removed, resulting in 43 samples.

5.3.5. Eigenvalues and error of reconstruction

The corresponding eigenvalues to all of the MDS plots were studied as well as the reconstruction errors. In all cases, the negative eigenvalues were small in magnitude, which implies that the method suited the problem and that a useful representation of the data was obtained. If the two or three largest eigenvalues are much larger than the rest, it is possible to find a good reconstruction of the original distance matrix in two or three dimensions (MATLAB). Two examples illustrate this nicely. The first example is the MDS plot of group A with 21 plant food allergens included together with 9 grass pollen allergens. The eigenvector problem is solved in 30 dimensions and results in 30 eigenvectors. The magnitudes of the eigenvalues can be seen in Figure 26. In this case, the first eigenvalue is very large in magnitude and the magnitude of first two eigenvalues account for as much as 91 % of the sum of all eigenvalues (Figure 26, Table 6). When adding a third eigenvalue this percentage does not increase more than 4 % because the second and third eigenvalue have similar magnitudes. This implies that in this case, the ability to reconstruct the distances does not increase when a third eigenvalue or dimension is used. The 2D and 3D errors respectively reflects the ability to reconstruct the distances and in this example, the reconstruction error does not decrease when plotting in three dimensions (Table 6).

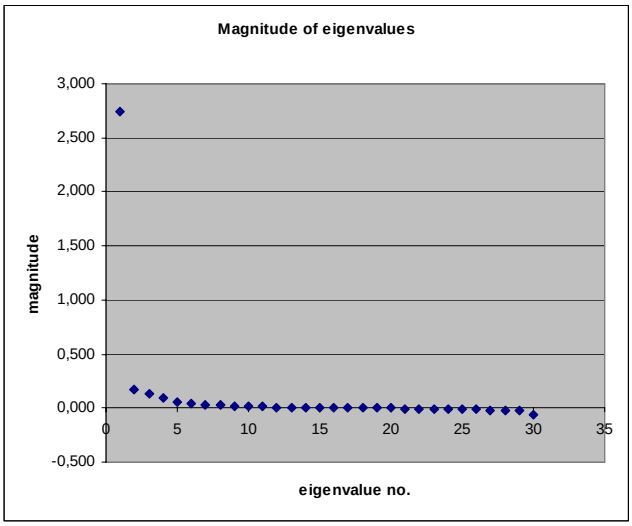


Figure 26. Magnitude of eigenvalues of MDS on group A based on 21 plant food allergens and 9 grass pollen allergens.

$\frac{\lambda_1+\lambda_2}{\sum_i \lambda_i}$	$\frac{\lambda_1+\lambda_2+\lambda_3}{\sum_i \lambda_i}$	Reconstruction errors		Maximum original distance
		2D	3D	
91 %	95 %	0.25	0.25	0.93

Table 6. Magnitude of the first two and three eigenvalues in relation to the sum of all magnitudes, reconstruction errors and the maximum original distance when performing MDS on group A with 30 allergens.

The second example is the MDS plot of group B with 21 plant food allergens included together with 9 grass pollen allergens. In this example (Figure 27), the magnitude of the second and third eigenvalue differs significantly, and the third eigenvalue has a relative magnitude which is half of the first eigenvalue. Here, the ability to reconstruct the distances, reflected in the percentage of magnitude that the first two or three eigenvalues constitute, increases with a third eigenvalue. The error of reconstruction decreases as a third dimension is

used. The percentage of the sum of all magnitudes that the first three eigenvalues constitute is only 72 % though (Table 7). This is explained by the fact that as many as the ten first eigenvalues seem to be large in magnitude (Figure 27). A consequence of this is that the reconstruction error in three dimensions is still quite large when compared to the maximum original distance between two allergens.

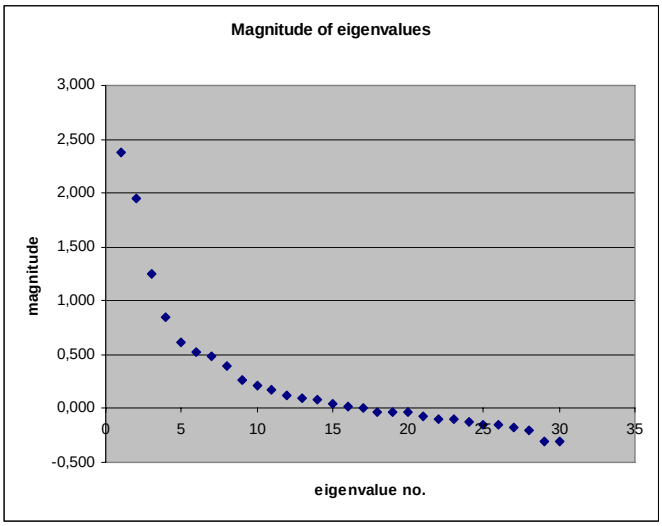


Figure 27. Magnitude of eigenvalues of MDS on group B based on 21 plant food allergens and 9 grass pollen allergens.

$\frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i}$	$\frac{\lambda_1 + \lambda_2 + \lambda_3}{\sum_i \lambda_i}$	Reconstruction errors		Maximum original distance
		2D	3D	
56 %	72 %	0.70	0.58	1.2

Table 7. Magnitude of the first two and three eigenvalues in relation to the sum of all magnitudes, reconstruction errors and the maximum original distance when performing MDS on group A with 30 allergens.

5.4. Improvement of the method

MDS had been tested briefly and found to be potentially useful as a tool to generate allergen maps with before this degree project started. In previous sections of this report, an extensive evaluation of this method has been described. This section describes some improvements of the method that have been discovered and implemented during the project and aims at presenting possible future usages of the method.

5.4.1. Higher resolution of allergen maps

In the previous allergen map study at Phadia, including 1127 samples and 89 allergens, ten allergen groups were visualised with the MDS procedure. Pollens, venoms and foods of plant origin grouped together and no sub-groups could be resolved. By restricting the MDS procedure to include allergens exclusively from pollens, venoms and foods of plant origin, a higher resolution of the green area was obtained (Figure 28). In this visualisation, sub-groups of the pollens are resolved, as well as venoms and foods from plant origin. This result implies that the resolution of allergen maps depends on what allergens are included in the analysis.

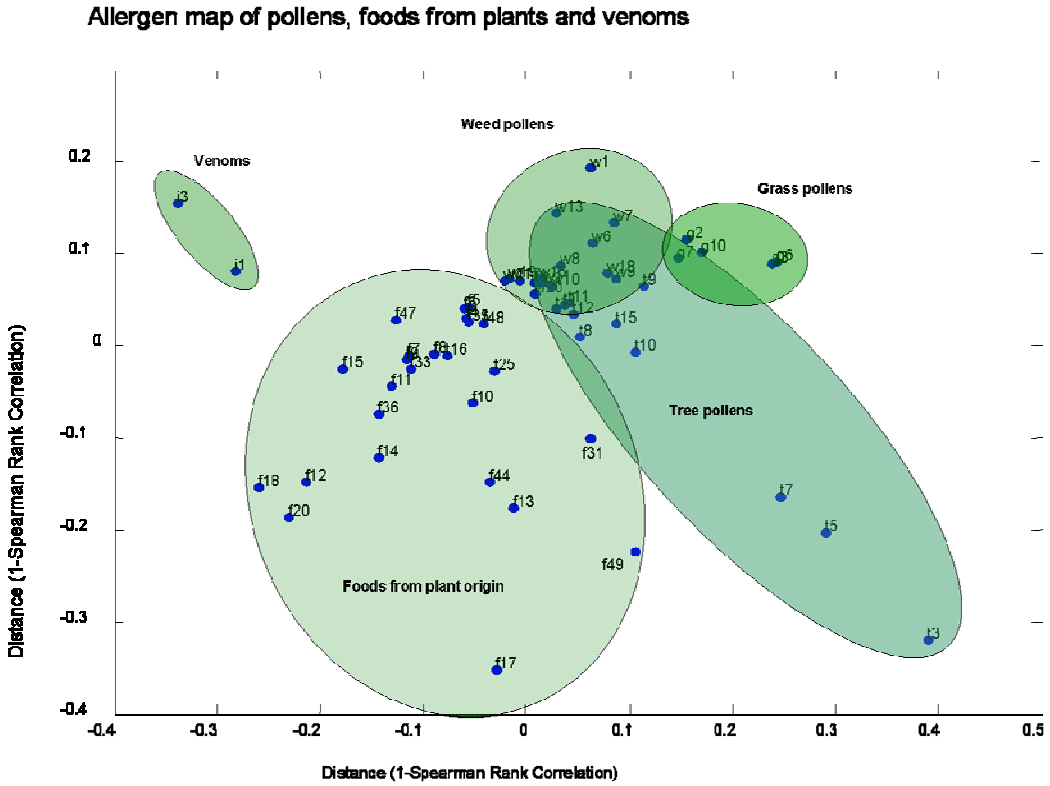


Figure 28. Allergen map based on data from the same 1127 blood sera samples, but with a reduced number of allergens included, resulting in a higher resolution. (Used with permission from Phadia).

5.4.2. Visualizing the results in 3D-plots

The results of section 5.3.5 suggest that it is sometimes useful to plot the result of the MDS in three dimensions. A possible case could be when the reconstruction error for the MDS plot in two dimensions is large, and the reconstruction error decreases when the result is plotted in three dimensions. A good example of the usefulness of 3D plots is the allergen map presented in section 3.6.4, created at a previous study at Phadia. In the two dimensional allergen map including 89 allergens and 10 allergen groups, mites ended up inside the allergen group epidermals due to a projection error (Figure 29). This error was easily detected since the correlation coefficient between mites and epidermals was low.

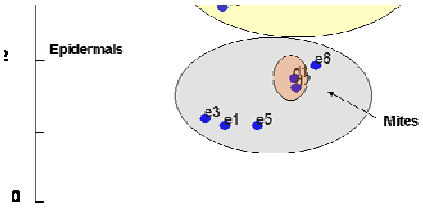


Figure 29. Zoom-in of the mites and epidermal group in allergen map of 89 allergens (see Figure 4).

However, a plot in three dimensions of the same data showed that the mites were separated from the epidermals in the 3D space even though the two groups still remained close to each other (Figure 30).

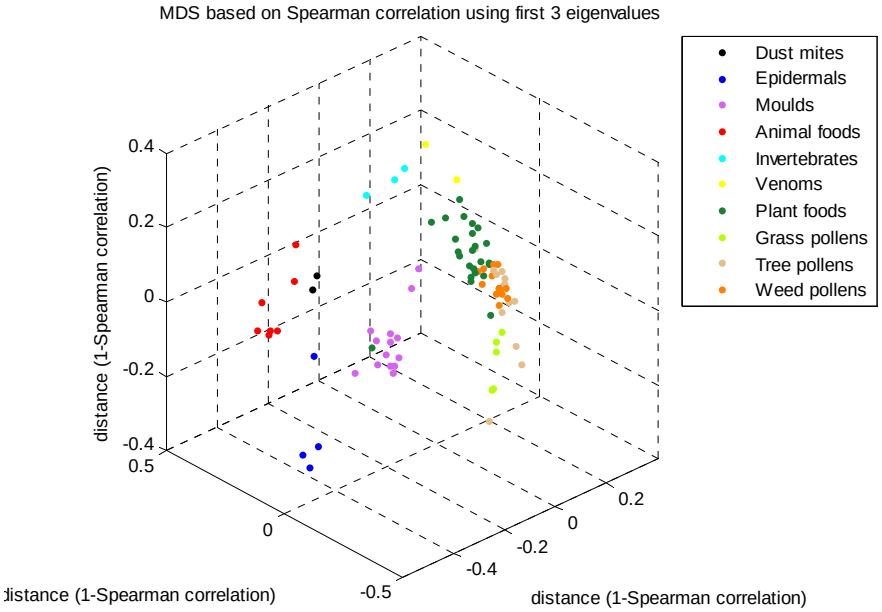


Figure 30. 3D plot of allergen map from previous study at Phadia. 1127 samples with specific IgE responses to 89 allergens included. The 3D plot shows that the dust mites can be separated from the epidermals.

In this case, the second and the third eigenvalue were almost equal in magnitude and the reconstruction error in two dimensions was fairly large in relation to the maximum original distance between any two allergens. The reconstruction error of the three dimensional plot was somewhat smaller. Even the fourth eigenvalue was quite large in magnitude in this case, which could explain why the reconstruction error in three dimensions still was not as small as one could desire.

5.4.3. Application

In order to make the method available for employees at Phadia in the future, an application with a graphical user interface was developed in the Matlab environment (Figure 31). The program connected to this interface performs MDS on a user defined input data file and presents the results in graphics defined by the user. The user starts with specifying the input Excel file, chooses missing value estimation method and finally how to display the results. The results can be displayed in 2D and/or 3D plots together with the eigenvalues and the reconstruction errors. A nice feature is that the user can choose to plot the results in a three dimensional plot with coloured allergen groups and specify the names of the allergen groups and colours himself. A manual for the application was written and is available at Phadia.

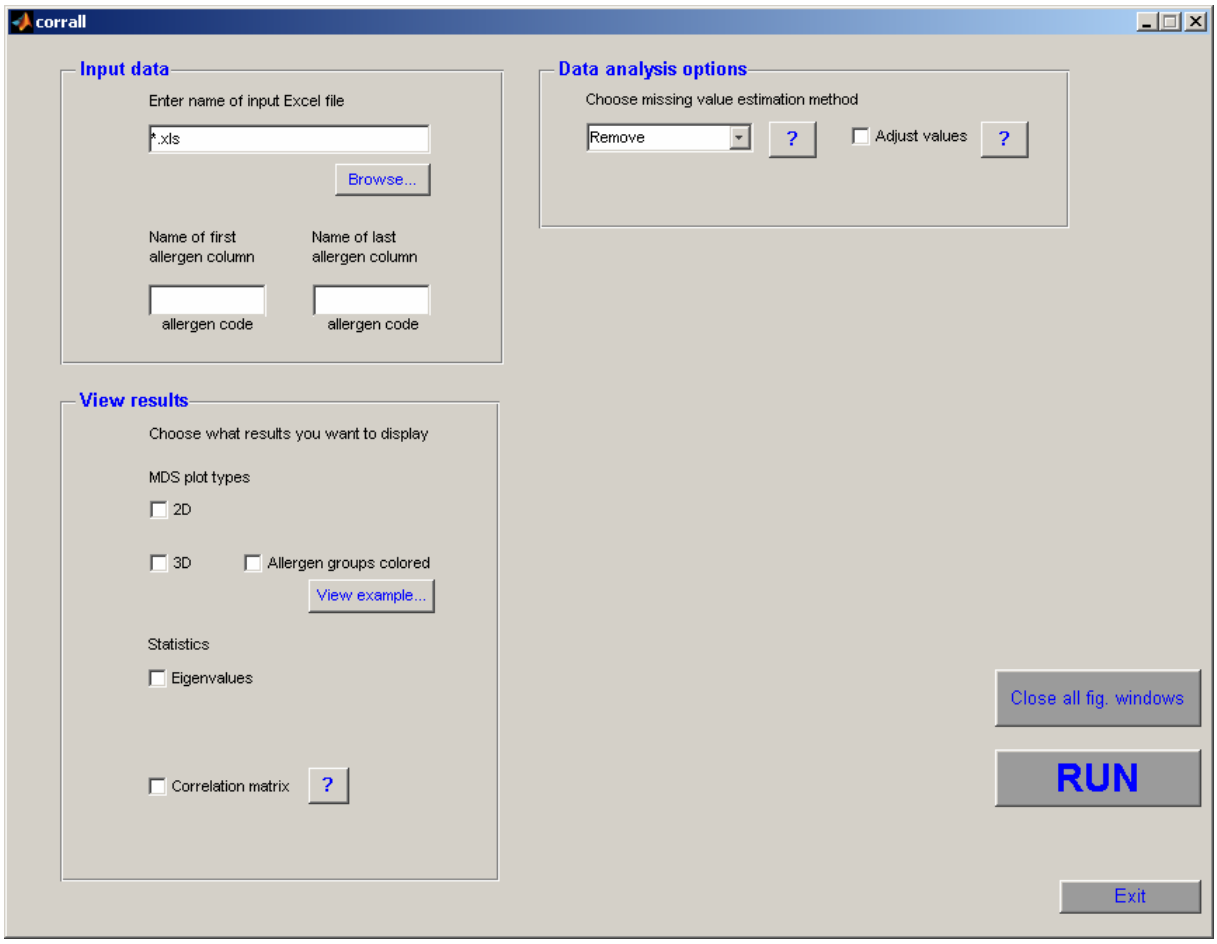


Figure 31. Graphical user interface for using MDS on IgE data sets.

6. Discussion

The aim of this project was to explore IgE data with methods within pattern recognition with the objective to visualise the IgE reactivity patterns in patients from three different patient groups. Revealing relationships between allergens of grass pollens, cereal grains and other foods of plant origin was one sub-goal. This report is mainly a description of methods and their performance when applied on IgE data. The biological interpretation of the results is limited and a deeper understanding of the biology, together with experiments that can verify the results, will be left to experts and researchers. However, concentrating on the biological problem of studying the relationships between the allergen groups of grass and plant food made it possible to evaluate the performance and robustness of the methods.

The exact and concrete use of the results in this project is not yet clear. Further investigations and analyses of the found IgE reactivity patterns, accompanied by experimental and clinical studies is a necessary continuation in order to make use of the results. The methodology and evaluation presented here constitute a first step towards a full understanding the relationships between plant food allergen extracts. The results have clearly shown that multidimensional scaling (MDS) is a robust and useful method for visualising IgE data. Using this method to visualise relationships between allergens in so called allergen maps, provide a novel and useful approach which can give researchers at Phadia clues about IgE reactivity patterns of different patient groups. In possible future steps, experiments and clinical studies can identify clinically irrelevant cross-reactive proteins which can be excluded from allergen extracts, facilitating the work towards a higher specificity of Phadia's *in vitro* test instruments.

Even though the exact interpretation of the IgE reactivity patterns of the three groups cannot be made yet, the methodology has provided Phadia with visualisations that give an overview of the sensitisation profiles that has not been available before. In addition to visualisations, the study of the three groups A, B and C, have generated useful information about the content of Phadia's sera bank. The number of samples ending up in each of the subsets (groups) gave an idea about what kind of samples are contained in the sera bank. Group A turned out to be the undoubtedly largest group with over 400 patients compared to around 50 and 70 respectively for group B and C. Moreover, group A had high IgE responses in most of the ten allergen groups comprising the 93 allergens included in the database search. In contrast, group B and C had low IgE levels in general. This implies that the sera bank mainly contains sera from multi-sensitised patients, which is not particularly surprising since the goal is to acquire such samples. However, in studies like this one, it would be desirable to have access to a larger number if patients in group B and C, i.e. patients that are only allergic to a few allergens. This could facilitate a higher accuracy of the comparing study. It would also be interesting to know how the distribution of the patient groups looks like in a normal population.

The difference in number of samples in each group gave rise to the question if the difference in patterns is due to the difference in number of samples. In order to investigate this, a few number of samples were selected randomly in group A and the results of MDS performed on the reduced subset A were compared to group B and C having the same number of samples. The resulting patterns were the same, which was expected since the allergy profile of the three groups, reflected in the IgE levels in different allergen groups, differed significantly. The difference in IgE levels gave rise to another question though: were the differences in patterns seen in the MDS plots of group A and B due to differences in IgE levels rather than the criteria for forming the groups? IgE responses of onion were used as a

test example to divide group A in two parts of which one had the same levels as group B and the MDS plots of each of the two parts were compared with the MDS plot of group A in its whole. The results showed indeed that a somewhat more scattered plot, like the plot of group B, was obtained for the samples in group A with the low IgE levels for onion. However, the scattered patterns were not similar to that of group B, implying that there is an actual difference between the groups, independently of the IgE levels. Performing the same analysis for all of the plant food allergens might have given some information about the actual cause of the difference between the groups. Due to lack of time and the fact that this was just a small side-study, this was not performed.

The study of the three groups' IgE reactivity patterns showed that there is a difference between the groups. By visualising the correlations between allergens in MDS plots, the differences between the groups were easily detected. In the MDS plot of group A, including both plant food allergens and grass pollen allergens, plant food allergens grouped densely together, separate from the grass pollen allergens, which also formed a separate group. The literature (20, 26, 27) reports cross-reactivity between grass pollens and foods of plant origin. In group A, the plant foods grouped together with the cereals, which cross-react to a large extent with grass pollens (18). This could imply that there is some component present in both grass pollens and cereals in group A, causing the plant foods to correlate closely to the cereals. The known cross-reactive relationship between tomato and grass pollens could be confirmed by this hypothesis, since tomato grouped particularly close to the cereal grains. One possible conclusion is that patients in group A are sensitised to a component present in grass pollens and cereals as well as plant foods. This component could be the cause of the multiple IgE reactivity in group A. When the grass pollen allergens were removed from the analysis, the plant food allergens still grouped densely together. Removing 20 % of the data and introducing missing values to group A, indicated that the amount of data was sufficient and the group was robust to changes. One can also draw the conclusion that it is a homogenous group of patients with similar allergy profiles since the patterns are conserved to a large extent even when 20 % of the samples were removed.

Group B and C, on the other hand, both had MDS plots with scattered plant food allergens. In group B, some plant food allergens grouped with the grasses; the same plant food allergens that grouped close to the cereals in group A. The cereal grains did not correlate significantly with any other plant foods, suggesting that individuals in group B react to wheat and cereal grains alone, perhaps to a few other food allergens. Possibly, these individuals react to one or more components that are present in wheat, but not in any other plant foods. If they experience symptoms, individuals in group B can be regarded as 'true' wheat allergic. Group B was more sensitive to missing values, removal of samples and measurement noise. Considering the mere size of this data set, this is expected. Another possible explanation is that this group is more heterogeneous than group A, containing individuals with varying allergy profiles. When samples are removed from such a group, the changes should be more prominent compared to removing samples from a homogenous group where all samples are more or less equal. Each sample in a heterogeneous group contributes to the general pattern since it differs from the rest of the samples.

In the MDS plots of group C, the grasses grouped together, separate from the plant food allergens. At the same time, the plant food allergens were scattered in the plot. This could possibly suggest that this group is sensitised to components in grasses, but the sensitisation to food allergens differ within the group. The study of group C was carried out in a late stage of this project and therefore, the characteristics if this group with respect to missing values, measurement noise and removal of data could not be evaluated. However, one can guess that the results would be similar to those of group B because of the similar size of the data sets.

Jones et al. (18) studied the cross-reactivity among cereal grains and grasses in three groups of patients corresponding to the three groups studied in this project. The study involved skin prick tests and food challenges as well as serological tests with immunoassays and immunoblotting. Their study showed extensive cross-reactivity among cereal grains and grasses in patients with both grain and grass sensitisation (group A in this study) or in patients with grass allergy alone (group C in this study). Different results were found in the group of patients with wheat allergy alone (group B in this study) though. In this group, extensive cross-reactivity was seen among cereal grains but none among related grasses (18). Similar patterns can be seen in this study: in group A, the cereal grains and grasses group closely together in the MDS plot. In group B though, the cereal grains are located next to each other and separate from the other allergens, but the grasses are scattered and mixed with plant food allergens. In group C, it is not possible to visualise the cereal grains, but the grasses group closely together. Even though the group of patients differ, it is interesting to see that some similarities are found in the results in spite of different methodology. This indicates that results from experimental and clinical studies can and should be combined with results from bioinformatical studies like this. Patterns can be confirmed and the methodology validated.

By studying the specific IgE responses to components in group A, one can identify what component that may cause the multiple IgE reactivity to grass pollens, cereal grains and plant foods in group A. The component study revealed that the CCD-containing bromelain had high correlation to the plant foods, implying that CCD could be the component that causes cross-reactivity within this group. This is rather expected since patients with plant-derived allergies have a higher prevalence of IgE to CCD (23). The prevalence increases in patients with multiple pollen sensitisation from trees, grasses and weeds (8, 23). Probably, many of the patients in group A had a multiple pollen sensitisation considering the elevated IgE levels to pollens in general, reflected in the staple diagram with IgE levels within ten allergen groups. Even though CCD is highly suspected as the responsible component for cross-reactivity within group A, the clinical relevance of group A's cross-reactivity can still not be determined since the clinical relevance of CCD is discussed (8, 23). Clinical data containing symptoms of these patients could relate symptoms to the serological patterns seen in this study and determine the clinical relevance.

In another experimental study conducted at R&D, Phadia, the degree of possible cross-reactivity between wheat and common timothy grass pollen components was investigated (3). It was found that the major cross-reactive component between wheat grain and grass pollen was Phl p 4 (3). This result could not really be seen in the component study of this project. Even though Phl p 4 was the component of timothy grass that grouped closest to the grains, it was not located as close to the plant food allergens as bromelain. However, Phl p 4 is a glycoprotein, like bromelain which is a plant glycoprotein (2). Bencúrová et al. (2) write about antiglycan IgE antibodies, which might be involved in this case. Bromelain was not studied as a potential cross-reactive compound in the study conducted at R&D at Phadia.

It is discussable whether one can draw conclusions from analyses on extract IgE data and component IgE data that have not been measured at the same occasion. In this case, the IgE responses to extracts in their whole were measured on an occasion preceding the measurement of IgE responses to the components on the same blood sera. Depending on the time in-between, one cannot exclude the risk of changes in the blood sera sample. The optimal data had contained both IgE measurements on extract level as well as IgE measurements on component level, measured at the same occasion, in order to relate these measurements correctly to each other. Furthermore, the number of samples included in the component study was mere; only 34 patients. One can question the validity of results based on such a small data set. However, the methodology demonstrates its utility on a combination of extract IgE data and component data. This methodology makes it possible to reveal the

mechanism behind multi-reactive patterns in an allergy group like group A. Of course, it had also been desirable to study component IgE data from group B and C. Unfortunately, no such data was available.

In this study, the validity of the method was evaluated by means of missing values, measurement noise and removal of samples. The over-all conclusion is that the multidimensional scaling (MDS) method is robust to all of these evaluating measures, even when small data sets are used.

Both missing value estimation methods used; BPCA and LLS, performed equally well. According to the literature (21, 25), BPCA performs better when the number of samples is large and the data set has global covariance structure. These are features of group A, and it is therefore not surprising that the NRMSE value for BPCA is lower than for LLS for this group. It is strange that the NRMSE value decrease as the percentage of missing values increase, a phenomenon not reported in the literature. LLS, should in contrast to BPCA, according to the literature, perform better on small data sets with local similarity structures (21). However, no difference in performance of missing value estimation could be seen in group B which is a small, heterogeneous data set with local similarity structures. The NRMSE values were extremely high for both methods, with a weak tendency for a lower NRMSE of BPCA. This indicates that the NRMSE value seems to be dependent of the number of samples rather than the amount of missing values. Indeed, Kim et al. (21) reports a peak in NRMSE for both methods on data sets with 50-100 samples and a decreasing value as the number of samples increase. One can also argue that the NRMSE value is not a useful or representative measure of the estimation accuracy, especially for small data sets. In conclusion, considering both the literature (21, 25) and the results of this study, BPCA is recommended for large data sets with global structures (homogeneous patient group) whereas LLS is recommended for small data sets with local similarity structures. Yet, the difference in performance between the methods was scarce.

The evaluation of missing values resulted in guidelines for appropriate levels of missing values for subset A and B respectively. An important remark is that the study of missing values was performed on randomly introduced missing values, assuming that missing values occur randomly and independently in the data matrix. This assumption may not be valid in real IgE data. Therefore, the guidelines for appropriate levels of missing values should be interpreted with a certain caution.

Among other methods that possibly can be used to discover patterns in IgE data is principal component analysis (PCA). This method was tested, but no useful results were obtained even though a range of different approaches to pre-process data were tried. Another method that could have been tested is clustering. Clustering techniques organize multidimensional data into groups with similar patterns (36) and are commonly used for gene expression data (33).

Unfortunately, the data used in this study have been insufficient in order to carry out a deeper biological analysis of the results. Yet, another important goal with this project was to come up with methods and guidelines that can be used in future studies of multidimensional IgE data at Phadia, dealing with other biological problems. The recommendations and conclusions from this work might not be applicable to all kinds of data sets within this field, but they point out what issues that are important to address in future pattern recognition studies of IgE data. In addition, this work provides methodologies for evaluating pattern recognition methods and data sets and for interpreting results of studies on IgE data. These methodologies are to be used for similar problems dealing with pattern discovery in IgE data at Phadia in the future. Together with access to more data, future studies will probably be more focused on the biological interpretation of the results. Clinical records, together with more data on IgE responses to components from all kinds of allergens will facilitate this.

Additional information about the patients such as age, gender, native country and family history of allergies will further facilitate the possibility to do large-scale studies of multidimensional IgE data with clinical relevance.

The application developed in connection with this project can be used to perform similar studies in various projects at Phadia, with other biological problems. In the future, this application may be further developed to make it more user-friendly. Moreover, this application will hopefully only be a part of a platform containing various interfaces that can handle several methods for analysis of IgE data.

In the future, the pattern recognition approach can perhaps also support the diagnosis of allergies by visualising serological relationships between allergens. Allergen maps for many different patient groups based on symptoms, and other information about the patients such as age, can provide an additional input in the diagnosis of an allergy. Given already diagnosed allergies together with symptoms and other information about the patient as input, each patient will fit into a special patient group with its own allergen map. The allergen map can reveal possible cross-reactive relationships that might occur in the particular patient group that are important to consider in the diagnosis. Suppose that a future study has indeed shown that group A has no symptoms of wheat allergy, while group B has. If, for example, a patient with a possible wheat allergy is tested positive for grass pollen, wheat, tomato and onion, he or she will belong to group A and the allergen map will show that it is most likely that the positive results are due to cross-reactivity. A positive test for wheat but no sensitisation to for example onion would indicate that the patient belongs to group B and has a true wheat allergy.

Allergen maps can function as an additional input in the diagnosis, and perhaps increase the accuracy of the test results and diagnosis. Possibly, with an additional input, one will not have to perform food challenges that are risky for the patient. This could be an opportunity for Phadia to provide doctors not only with test instruments, but also to provide a test interpretation support. However, to generate allergen maps that cover most patient groups require a large amount of accessible data.

Finally, it has been a pleasure to carry out this degree project. Taking part of this pioneering work involving a novel approach to study IgE data *in silico*[1], has been exciting. I am convinced that this is just the beginning of bioinformatics entering the field of allergy diagnostics, and perhaps the diagnostics field in general.

[1] performed on a computer or via computer simulation

7. Acknowledgements

I would like to thank my supervisor Annica Önell at Phadia for great supervision and engagement in my work. I would also like to thank Per Matsson and Ingvar Edlert at Phadia for making this degree project possible. Other persons at Phadia who have helped me to carry out this project are the Biometrics group with whom I have been able to discuss statistical topics and Jörgen Dahlström at R&D who has provided me with data, ideas and feedback. Finally, I would like to thank my scientific reviewer Mats Gustafsson at the Department of Engineering Sciences, Uppsala University, for input and feedback, and Daniel Soeria-Atmadja and Ulf Hammerling at National Food Administration for ideas, feedback and articles.

8. References

1. Andersson, K., Lidholm, J. (2003) Characteristics and Immunobiology of Grass Pollen Allergens. *Allergy and Immunology,* vol. 130, 87-107.
2. Bencúrová, M. et al. (2004) Specificity of IgG and IgE antibodies against plant and insect glycoprotein glycans determined with artificial glycoforms of human transferring. *Glycobiology* vol. 14, 457-466.
3. Bernhardsson, F., et al. (2006) IgE reactivity to grass components in wheat sensitized subjects. In: *XXV Congress of the European Academy of Allergology and Clinical Immunology 10 - 14 June 2006, Vienna, Austria.*
4. Bindslev-Jensen, K. (1998) ABC of allergies. Food allergy. *British Medical Journal,* vol. 316, 1299-1302.
5. Breitender, H., Mills, C. (2006) Structural bioinformatic approaches to understand cross-reactivity. *Mol. Nutr. Food Res.,* vol. 50, 628-632.
6. Camastra, F. (2003) Data dimensionality estimation methods: a survey. *Pattern recognition,* vol. 36, 2945-2954.
7. Campbell, et al. (1999) *Biology*, 5th edition. Benjamin/Cummings, San Francisco, USA.
8. Ebo, D. G., et al. (2004) Sensitisation to cross-reactive carbohydrate determinants and the ubiquitous protein profilin: mimickers of allergy. *Clin Exp Allergy,* vol. 34, 137-144.
9. Foetisch, K., et al. (2001) Tomato (*Lycopersicon escelentum)* allergens in pollen-allergic patients. *Eur Food Res Technol,* vol. 213, 259-266.
10. Grote, M., et al. (2002) Identification of an Allergen Related to Phl p 4, a Major Timothy Grass Pollen Allergen, in Pollens, Vegetables, and Frutis by Immunogold Electron Microscopy. *Biol. Chem.,* vol. 383, 1441-1445.
11. Hamilton, R., et al. (2004) *In vitro* assays for the diagnosis of IgE-mediated disorders. *J Allergy Clin Immunol,* vol. 114, 213-225.
12. Heiss, S., et al. (1996) Identification of a 60 kd cross-reactive allergen in pollen and plant-derived food. *J Allergy Clin Immunol,* vol. 98, 938-947.
13. ImmunoCAP™ InVitroSight™, 2006: Expected Values. (19.6.2006) http://www.immunocapinvitrosight.com/templates/Page.asp?id=1934
14. ImmunoCAP™ InVitroSight™, 2006: Product Description. (19.6.2006) http://www.immunocapinvitrosight.com/templates/Page.asp?id=1937
15. ImmunoCAP™ InVitroSight™, 2006: The new generation of allergy testing - Pharmacia CAP System™. (19.6.2006) http://www.immunocapinvitrosight.com/templates/Page.asp?id=2054
16. ImmunoCAP™ InVitroSight™, 2006: Type I Hypersensitivity (atopic allergy). (19.6.2006) http://www.immunocapinvitrosight.com/templates/Page.asp?id=1892
17. Johansson, S.G.O (2006) The discovery of immunoglobulin E. *Allergy and Asthma Proceedings,* vol. 27, S3-6.
18. Jones, S., et al. (1995) Immunologic cross-reactivity among cereal grains and grasses in children with food hypersensitivity. *J Allergy Clin Immunol,* vol. 96, 341-351.
19. Kagan, R.S. (2003) Food Allergy: An Overview. *Environmental Health Perspectives,* vol. 111, 223-225.
20. Kazemi-Shirazi, L., et al. (1999) Quantitative IgE inhibition experiments with purified recombinant allergens indicate pollen-derived allergens as the sensitizing agents responsible for many forms of plant food allergy. *J Allergy Clin Immunol,* vol. 105, 116-125.
21. Kim, H., et al. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics,* vol. 21, 187-198.
22. Lodish, et al. (2000) *Molecular Cell Biology*, 4th edition. W. H. Freeman and Company, England.
23. Mari, A. (2002) IgE to Cross-Reactive Carbohydrate Determinants: Analysis of the Distribution and Appraisal of the *in vivo* and *in vitro* Reactivity. *Allergy and Immunology,* vol. 129, 286-295.
24. Marknell DeWitt, Å., et al. (2006) Cloning, expression and immunological characterization of full-length timothy grass pollen allergen Phl p 4, a berberine bridge enzyme-like protein with homology to celery allergen Api g 5. *Clin Exp Allergy,* vol.36, 77-86.
25. Oba, S., et al. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics,* vol. 19, 2088-2096.
26. Osterballe, M., et al. (2005) The clinical relevance of sensitisation to pollen-related fruits and vegetables in unselected pollen-sensitised adults. *Allergy,* vol. 60, 218-225.
27. Pastorello, E.A., et al. (2002) New plant-origin food allergens. *Allergy,* vol. 57, 106-110.

28. Petersen, A., et al. (1996) Ubiquitous structures responsible for IgE cross-reactivity between tomato fruits and grass pollen allergens. *J Allergy Clin Immunol,* vol. 98, 805-815.
29. Press, W.H., et al. (1992) Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd edition. Cambridge University Press, Cambridge, England.
30. Prussin, C., et al. (2006) IgE, mast cells, basophils, and eosinophils. *J Allergy Clin Immunol,* vol. 117, S450-456.
31. Ruden, S. & Steinmann, H. (2004) *Grass pollens. Allergy – Which allergens?* (Handbook). Phadia, Uppsala, Sweden.
32. Sanchez-Monge, R., et al. (2005) Analytical methodology for assessment of food allergens: Opportunities and challenges. *Biotechnology Advances,* col. 23, 415-422.
33. Slonim, D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nature genetics supplement,* vol. 32, 502-508.
34. Troyanskaya, O., et al. (2001) Missing value estimation methods for DNA micro arrays. *Bioinformatics,* vol. 17, 520-525.
35. Webb, A. (2002*) Statistical Pattern Recognition*, 2nd edition. Wiley, Hoboken, USA.
36. Önell, A., et al. (2006) Allergen maps – a statistically valid visualozation of correlations among allergens using pattern recognition methods on IgE data. In: *XXV Congress of the European Academy of Allergology and Clinical Immunology 10 - 14 June 2006, Vienna, Austria*.

Appendix A – List of 93 allergens included in database search

CODE	NAME	CODE	NAME	CODE	NAME
Foods of plant origin		**Grass pollens**		**House dust mites**	
f4	wheat	g2	bermuda grass	d1	D.pteronys.
f5	rye	g3	cocksfoot	d2	D.Farinae
f6	barley	g6	timothy grass	**Epidermals**	
f7	oat	g7	common reed	e1	cat
f8	maize	g10	johnson grass	e3	horse
f9	rice	g8	meadow grass	e5	dog
f10	sesame	g12	rye pollen	e6	guinea pig
f11	buckwheat	g14	oat pollen	**Moulds**	
f12	pea	g15	wheat pollen	m1	penicillium noyatum
f13	peanut	**Tree pollens**		m2	cladosporium herbarum
f14	soya bean	t1	box elder	m3	aspergillus fumigatus
f15	white bean	t3	birch	m4	mucor racemosus
f17	hazel nut	t5	beech	m5	candida albicans
f18	brazil nut	t6	mountain juniper	m6	alternaria alternata
f20	almond	t7	oak	m7	borytis cinerea
f25	tomato	t8	elm	m8	helminthosporium halodes
f31	carrot	t9	olive	m9	fusarium moniliforme
f33	orange	t10	walnut	m10	stemphylium botryosum
f35	potato	t11	maple leaf sycamore	m11	rhizopus nigricans
f36	coconut	t12	willow	m12	aureobasidium pullulans
f44	strawberry	t14	cottonwood	m13	phoma betae
f45	yeast	t15	white ash	m14	epicoccum purpurascens
f47	garlic	t16	white pine	**Invertebraes**	
f48	onion	t17	japanese cedar	i6	cockroach
f49	apple	**Weed pollens**		f37	blue mussel
Foods of animal origin		w1	common ragweed	f24	shrimp
f1	egg white	w6	mugwort	**Venoms**	
f2	milk	w7	marguerite	i1	beef
f3	cod	w8	dandelion	i3	wasp
f26	pork	w9	plantain		
f27	beef	w10	goosefoot		
f40	tuna	w11	saltwort		
f83	chicken	w13	cocklebur		
		w15	scale		
		w18	sheep sorrel		
		w19	wall pellitory		
		w20	nettle		
		w21	wall pellitory		

In the previous allergen map study including 89 allergens, g8, g12, g14 and g15 were not included in the statistical analysis.

Appendix B – correlation coefficients group A

	f10	f11	f12	f13	f14	f15	f20	f25	f31	f33	f35	f36	f4	f44	f47	f48	f5	f6	f7	f8	f9	g10	g12	g14	g15	g2	g3	g6	g7	g8
g8	1,00	0,84	0,76	0,74	0,84	0,71	0,73	0,79	0,77	0,77	0,77	0,80	0,75	0,80	0,75	0,80	0,74	0,78	0,82	0,81	0,82	0,36	0,29	0,32	0,31	0,44	0,27	0,26	0,45	0,23
g7	0,84	1,00	0,84	0,75	0,88	0,81	0,82	0,83	0,74	0,83	0,83	0,84	0,79	0,85	0,84	0,83	0,79	0,84	0,80	0,89	0,88	0,37	0,27	0,31	0,29	0,46	0,22	0,24	0,45	0,19
g6	0,76	0,84	1,00	0,71	0,86	0,73	0,85	0,88	0,73	0,86	0,88	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,83	0,89	0,89	0,42	0,31	0,36	0,34	0,45	0,24	0,26	0,47	0,22
g3	0,74	0,88	0,71	1,00	0,82	0,67	0,78	0,88	0,73	0,80	0,86	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,83	0,89	0,89	0,42	0,31	0,36	0,34	0,45	0,23	0,26	0,47	0,22
g2	0,77	0,81	0,73	0,67	0,79	1,00	0,77	0,76	0,75	0,85	0,81	0,81	0,74	0,81	0,81	0,84	0,81	0,84	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
g15	0,77	0,81	0,73	0,67	0,79	0,77	1,00	0,76	0,75	0,85	0,81	0,81	0,74	0,81	0,81	0,84	0,81	0,84	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
g14	0,79	0,81	0,73	0,67	1,00	0,79	0,81	0,79	0,73	0,81	0,81	0,81	0,74	0,81	0,81	0,84	0,81	0,84	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
g12	0,86	0,88	0,71	1,00	0,82	0,67	0,79	0,77	0,74	0,86	0,88	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,83	0,89	0,89	0,42	0,31	0,36	0,34	0,45	0,23	0,26	0,47	0,22
g10	0,79	0,83	0,70	0,74	0,81	0,66	0,66	0,88	0,80	0,82	0,88	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
f9	0,73	0,82	0,85	0,66	0,79	0,77	0,66	0,88	0,83	0,86	0,88	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
f8	0,71	0,81	0,73	0,66	0,79	0,76	0,75	0,87	0,79	0,85	0,86	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
f7	0,71	0,81	0,73	0,67	0,79	0,76	0,75	0,87	0,77	0,87	0,86	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
f6	0,78	0,80	0,70	0,70	0,78	0,76	0,78	0,88	0,77	0,87	0,86	0,85	0,79	0,85	0,84	0,87	0,80	0,85	0,84	0,84	0,84	0,39	0,27	0,30	0,29	0,46	0,23	0,26	0,47	0,22
f5	0,74	0,79	0,68	0,68	0,78	0,75	0,78	0,80	0,72	0,81	0,80	0,73	0,95	0,78	0,82	0,84	0,80	0,94	0,87	0,85	0,84	0,37	0,27	0,31	0,31	0,52	0,25	0,26	0,52	0,25
f48	0,80	0,83	0,70	0,75	0,84	0,76	0,77	0,87	0,77	0,87	0,87	0,74	0,84	0,86	0,88	1,00	0,84	0,85	0,82	0,89	0,87	0,36	0,28	0,31	0,31	0,52	0,25	0,26	0,54	0,26
f47	0,75	0,81	0,73	0,70	0,84	0,84	0,72	0,78	0,69	0,85	0,78	0,73	0,84	0,81	1,00	0,88	0,82	0,84	0,84	0,84	0,89	0,38	0,33	0,35	0,31	0,47	0,29	0,31	0,52	0,26
f44	0,80	0,85	0,77	0,81	0,84	0,75	0,74	0,84	0,80	0,79	0,79	0,79	0,78	1,00	0,81	0,86	0,78	0,80	0,79	0,85	0,82	0,29	0,26	0,28	0,23	0,44	0,23	0,23	0,47	0,19
f4	0,75	0,79	0,67	0,68	0,78	0,74	0,63	0,77	0,71	0,80	0,79	0,71	1,00	0,78	0,84	0,84	0,95	0,93	0,89	0,85	0,86	0,38	0,32	0,35	0,31	0,52	0,31	0,32	0,55	0,27
f36	0,80	0,84	0,74	0,70	0,81	0,81	0,78	0,81	0,79	0,85	0,81	1,00	0,71	0,79	0,73	0,74	0,73	0,74	0,75	0,75	0,75	0,32	0,26	0,29	0,25	0,48	0,25	0,26	0,43	0,22
f35	0,77	0,83	0,65	0,71	0,80	0,81	0,66	0,88	0,83	0,86	1,00	0,81	0,79	0,79	0,78	0,87	0,80	0,80	0,75	0,79	0,78	0,44	0,36	0,39	0,38	0,57	0,31	0,32	0,51	0,28
f33	0,77	0,74	0,73	0,74	0,81	0,85	0,75	0,87	0,80	1,00	0,86	0,85	0,80	0,87	0,85	0,87	0,81	0,82	0,79	0,85	0,83	0,41	0,29	0,35	0,31	0,51	0,25	0,26	0,51	0,21
f31	0,77	0,74	0,61	0,73	0,75	0,75	0,61	0,82	1,00	0,80	0,83	0,79	0,80	0,80	0,69	0,77	0,72	0,74	0,70	0,73	0,72	0,43	0,37	0,40	0,38	0,57	0,34	0,35	0,50	0,31
f25	0,79	0,83	0,70	0,77	0,79	0,76	0,76	1,00	0,82	0,88	0,88	0,81	0,84	0,84	0,78	0,87	0,78	0,79	0,77	0,81	0,79	0,43	0,34	0,37	0,37	0,56	0,30	0,31	0,55	0,27
f20	0,73	0,82	0,85	0,66	0,81	0,77	1,00	0,66	0,61	0,75	0,66	0,78	0,72	0,74	0,72	0,67	0,65	0,65	0,68	0,72	0,75	0,23	0,15	0,19	0,17	0,30	0,09	0,11	0,28	0,07
f15	0,71	0,81	0,73	0,67	1,00	1,00	0,77	0,76	0,75	0,85	0,81	0,81	0,74	0,75	0,84	0,76	0,75	0,77	0,76	0,76	0,84	0,42	0,33	0,37	0,34	0,50	0,28	0,29	0,48	0,25
f14	0,84	0,88	0,86	0,82	1,00	0,79	0,81	0,79	0,73	0,81	0,77	0,77	0,78	0,84	0,78	0,79	0,75	0,78	0,81	0,81	0,82	0,33	0,25	0,28	0,27	0,41	0,20	0,22	0,41	0,17
f13	0,74	0,75	0,71	1,00	0,82	0,67	0,66	0,77	0,71	0,74	0,71	0,70	0,68	0,81	0,70	0,75	0,68	0,70	0,71	0,74	0,70	0,40	0,34	0,35	0,35	0,47	0,30	0,32	0,47	0,28
f12	0,76	0,84	1,00	0,71	0,86	0,73	0,85	0,70	0,61	0,73	0,65	0,74	0,67	0,77	0,73	0,70	0,68	0,70	0,73	0,75	0,76	0,27	0,19	0,22	0,21	0,32	0,14	0,15	0,35	0,11
f11	0,84	1,00	0,84	0,75	0,88	0,81	0,82	0,83	0,74	0,86	0,83	0,84	0,79	0,85	0,81	0,83	0,79	0,80	0,80	0,84	0,82	0,37	0,27	0,31	0,29	0,45	0,22	0,24	0,45	0,19
f10	1,00	0,84	0,76	0,74	0,84	0,71	0,73	0,79	0,77	0,77	0,77	0,80	0,75	0,80	0,75	0,80	0,74	0,78	0,82	0,81	0,82	0,36	0,29	0,32	0,31	0,44	0,27	0,26	0,45	0,23

Appendix C – correlation coefficients group B

	f10	f11	f12	f13	f14	f15	f20	f25	f31	f33	f35	f36	f4	f44	f47	f48	f5	f6	f7	f8	f9	g10	g12	g14	g15	g2	g3	g6	g7	g8
f10	1,00	0,58	0,59	0,52	0,44	0,33	0,56	0,28	0,31	0,23	0,14	0,55	-0,02	0,47	0,21	0,46	-0,17	0,04	0,20	0,23	0,53	0,25	0,53	0,21	0,50	0,33	0,27	0,28	0,32	0,35
f11	0,58	1,00	0,65	0,65	0,54	0,47	0,65	0,34	0,26	0,11	0,32	0,42	0,08	0,39	0,16	0,24	0,00	0,03	0,25	0,25	0,54	0,13	0,29	0,27	0,39	0,46	0,35	0,37	0,29	0,42
f12	0,59	0,65	1,00	0,64	0,61	0,43	0,64	-0,06	-0,03	0,23	0,00	0,50	-0,09	0,40	-0,19	0,04	-0,12	-0,10	0,37	0,20	0,37	-0,18	0,17	0,21	0,23	0,09	0,01	0,08	-0,05	0,17
f13	0,52	0,65	0,64	1,00	0,58	0,39	0,63	0,29	0,33	0,26	0,33	0,53	0,17	0,66	0,14	0,44	0,03	0,12	0,23	0,34	0,44	0,36	0,27	0,27	0,32	0,51	0,44	0,46	0,21	0,50
f14	0,44	0,54	0,61	0,58	1,00	0,34	0,56	0,17	0,42	0,21	0,35	0,42	0,21	0,66	0,09	0,09	0,24	0,21	0,42	0,35	0,52	0,23	0,28	0,39	0,35	0,14	0,22	0,29	0,20	0,31
f15	0,33	0,47	0,43	0,39	0,34	1,00	0,51	0,25	0,16	0,20	0,24	0,25	0,10	0,43	0,30	0,14	0,10	0,27	0,35	0,13	0,53	0,23	0,42	0,40	0,44	0,23	0,21	0,32	0,25	0,32
f20	0,56	0,65	0,64	0,63	0,56	0,51	1,00	0,17	0,11	-0,10	0,17	0,40	-0,08	0,35	0,19	0,25	-0,04	-0,04	0,29	0,14	0,50	0,24	0,35	0,43	0,39	0,33	0,25	0,34	0,16	0,33
f25	0,28	0,34	-0,06	0,29	0,17	0,25	0,17	1,00	0,49	0,15	0,54	0,21	0,29	0,26	0,45	0,50	0,25	0,36	0,09	0,26	0,21	0,51	0,32	0,27	0,52	0,56	0,52	0,51	0,53	0,48
f31	0,31	0,26	-0,03	0,33	0,42	0,16	0,11	0,49	1,00	0,18	0,33	0,16	0,35	0,57	0,22	0,38	0,22	0,29	0,08	0,21	0,28	0,44	0,33	0,30	0,44	0,36	0,46	0,57	0,46	0,43
f33	0,23	0,11	0,23	0,26	0,21	0,20	-0,10	0,15	0,18	1,00	0,41	0,44	0,11	0,39	0,15	0,33	0,07	0,15	0,16	0,33	0,16	0,15	0,28	0,02	0,16	0,15	0,06	0,17	0,24	0,05
f35	0,14	0,32	0,00	0,33	0,35	0,24	0,17	0,54	0,33	0,41	1,00	0,23	0,16	0,24	0,35	0,44	0,16	0,16	-0,03	0,26	0,19	0,31	0,36	0,23	0,24	0,56	0,29	0,42	0,31	0,25
f36	0,55	0,42	0,50	0,53	0,42	0,25	0,40	0,21	0,16	0,44	0,23	1,00	0,04	0,59	0,29	0,27	-0,08	0,15	0,10	0,47	0,41	0,33	0,14	0,14	0,10	0,26	0,10	0,08	0,15	0,08
f4	-0,02	0,08	-0,09	0,17	0,21	0,10	-0,08	0,29	0,35	0,11	0,16	0,04	1,00	0,33	0,13	0,05	0,78	0,47	0,36	0,52	0,17	0,30	0,04	0,03	0,16	0,02	0,42	0,36	0,27	0,33
f44	0,47	0,39	0,40	0,66	0,66	0,43	0,35	0,26	0,57	0,39	0,24	0,59	0,33	1,00	0,35	0,23	0,20	0,35	0,28	0,41	0,54	0,53	0,24	0,23	0,35	0,26	0,28	0,31	0,29	0,37
f47	0,21	0,16	-0,19	0,14	0,09	0,30	0,19	0,45	0,22	0,15	0,35	0,29	0,13	0,35	1,00	0,30	0,16	0,38	-0,02	0,15	0,40	0,67	0,31	0,28	0,33	0,38	0,28	0,22	0,65	0,27
f48	0,46	0,24	0,04	0,44	0,09	0,14	0,25	0,50	0,38	0,33	0,44	0,27	0,05	0,23	0,30	1,00	-0,07	0,28	-0,11	0,27	0,12	0,39	0,59	0,13	0,35	0,62	0,44	0,44	0,57	0,38
f5	-0,17	0,00	-0,12	0,03	0,24	0,10	-0,04	0,25	0,22	0,07	0,16	-0,08	0,78	0,20	0,16	-0,07	1,00	0,46	0,36	0,31	0,07	0,22	0,00	0,10	0,08	-0,09	0,32	0,28	0,20	0,21
f6	0,04	0,03	-0,10	0,12	0,21	0,27	-0,04	0,36	0,29	0,15	0,16	0,15	0,47	0,35	0,38	0,28	0,46	1,00	0,29	0,41	0,22	0,33	0,25	0,12	0,10	0,07	0,35	0,27	0,43	0,24
f7	0,20	0,25	0,37	0,23	0,42	0,35	0,29	0,09	0,08	0,16	-0,03	0,10	0,36	0,28	-0,02	-0,11	0,36	0,29	1,00	0,41	0,43	0,03	0,04	0,29	0,14	-0,20	0,06	0,05	0,17	0,08
f8	0,23	0,25	0,20	0,34	0,35	0,13	0,14	0,26	0,21	0,33	0,26	0,47	0,52	0,41	0,15	0,27	0,31	0,41	0,41	1,00	0,44	0,16	0,09	-0,02	0,13	0,12	0,13	0,11	0,30	0,05
f9	0,53	0,54	0,37	0,44	0,52	0,53	0,50	0,21	0,28	0,16	0,19	0,41	0,17	0,54	0,40	0,12	0,07	0,22	0,43	0,44	1,00	0,33	0,37	0,30	0,41	0,12	0,09	0,15	0,35	0,15
g10	0,25	0,13	-0,18	0,36	0,23	0,23	0,24	0,51	0,44	0,15	0,31	0,33	0,30	0,53	0,67	0,39	0,22	0,33	0,03	0,16	0,33	1,00	0,38	0,45	0,42	0,40	0,61	0,57	0,43	0,57
g12	0,53	0,29	0,17	0,27	0,28	0,42	0,35	0,32	0,33	0,28	0,36	0,14	0,04	0,24	0,31	0,59	0,00	0,25	0,04	0,09	0,37	0,38	1,00	0,45	0,58	0,30	0,42	0,54	0,39	0,45
g14	0,21	0,27	0,21	0,27	0,39	0,40	0,43	0,27	0,30	0,02	0,23	0,14	0,03	0,23	0,28	0,13	0,10	0,12	0,29	-0,02	0,30	0,45	0,45	1,00	0,45	0,30	0,44	0,54	0,24	0,49
g15	0,50	0,39	0,23	0,32	0,35	0,44	0,39	0,52	0,44	0,16	0,24	0,10	0,16	0,35	0,33	0,35	0,08	0,10	0,14	0,13	0,41	0,42	0,58	0,45	1,00	0,37	0,45	0,59	0,55	0,58
g2	0,33	0,46	0,09	0,51	0,14	0,23	0,33	0,56	0,36	0,15	0,56	0,26	0,02	0,26	0,38	0,62	-0,09	0,07	-0,20	0,12	0,12	0,40	0,30	0,30	0,37	1,00	0,58	0,55	0,42	0,56
g3	0,27	0,35	0,01	0,44	0,22	0,21	0,25	0,52	0,46	0,06	0,29	0,10	0,42	0,28	0,28	0,44	0,32	0,35	0,06	0,13	0,09	0,61	0,42	0,44	0,45	0,58	1,00	0,89	0,40	0,91
g6	0,28	0,37	0,08	0,46	0,29	0,32	0,34	0,51	0,57	0,17	0,42	0,08	0,36	0,31	0,22	0,44	0,28	0,27	0,05	0,11	0,15	0,57	0,54	0,54	0,59	0,55	0,89	1,00	0,36	0,80
g7	0,32	0,29	-0,05	0,21	0,20	0,25	0,16	0,53	0,46	0,24	0,31	0,15	0,27	0,29	0,65	0,57	0,20	0,43	0,17	0,30	0,35	0,43	0,39	0,24	0,55	0,42	0,40	0,36	1,00	0,37
g8	0,35	0,42	0,17	0,50	0,31	0,32	0,33	0,48	0,43	0,05	0,25	0,08	0,33	0,37	0,27	0,38	0,21	0,24	0,08	0,05	0,15	0,57	0,45	0,49	0,58	0,56	0,91	0,80	0,37	1,00

Appendix D – correlation coefficients group C

	f10	f11	f12	f13	f14	f15	f20	f25	f31	f35	f44	f47	f48	f8	f9	g10	g12	g14	g15	g2	g3	g6	g7	g8
f10	1,00	0,23	0,21	0,29	0,47	-0,08	0,22	0,24	0,33	0,29	0,17	0,21	0,20	0,06	0,10	0,08	0,11	0,13	0,13	0,01	0,13	0,16	0,03	0,14
f11	0,23	1,00	-0,03	0,05	0,41	-0,04	-0,06	0,12	0,12	0,23	0,17	0,23	0,31	0,49	0,57	0,09	0,06	0,07	0,04	0,18	0,05	0,10	-0,05	0,11
f12	0,21	-0,03	1,00	-0,09	0,42	-0,02	-0,03	-0,04	0,11	-0,02	-0,05	-0,03	-0,04	-0,03	-0,02	-0,10	0,02	-0,05	0,00	-0,08	-0,04	-0,04	-0,13	-0,04
f13	0,29	0,05	-0,09	1,00	0,15	0,10	0,36	0,29	0,52	0,26	0,39	0,22	0,18	-0,07	0,09	0,15	0,28	0,22	0,31	0,08	0,19	0,23	0,17	0,27
f14	0,47	0,41	0,42	0,15	1,00	-0,04	0,20	0,09	0,30	0,20	0,16	0,19	0,09	0,19	0,23	0,05	0,11	0,18	0,15	0,15	0,09	0,15	0,00	0,10
f15	-0,08	-0,04	-0,02	0,10	-0,04	1,00	0,33	0,26	0,13	-0,03	0,16	-0,04	-0,06	-0,04	-0,03	0,06	0,12	0,13	0,13	0,02	0,12	0,07	0,07	0,10
f20	0,22	-0,06	-0,03	0,36	0,20	0,33	1,00	0,14	0,44	-0,05	0,41	0,20	0,09	-0,06	-0,05	0,12	0,08	0,12	0,11	0,07	0,04	0,22	0,16	0,02
f25	0,24	0,12	-0,04	0,29	0,09	0,26	0,14	1,00	0,54	0,65	0,33	0,31	0,36	-0,08	-0,07	0,03	0,19	0,20	0,23	0,08	0,20	0,22	0,09	0,18
f31	0,33	0,12	0,11	0,52	0,30	0,13	0,44	0,54	1,00	0,35	0,54	0,25	0,27	-0,02	-0,04	0,16	0,16	0,15	0,23	0,12	0,12	0,09	0,15	0,18
f35	0,29	0,23	-0,02	0,26	0,20	-0,03	-0,05	0,65	0,35	1,00	0,23	0,26	0,39	-0,05	-0,04	0,02	0,18	0,20	0,20	-0,02	0,17	0,22	0,05	0,20
f44	0,17	0,17	0,39	0,16	0,09	-0,06	0,41	0,33	0,54	0,23	1,00	0,22	0,18	0,05	0,24	0,17	0,11	0,10	0,15	0,20	0,13	0,11	0,21	0,11
f47	0,21	0,23	0,22	0,22	0,19	-0,04	0,20	0,31	0,25	0,26	0,22	1,00	0,29	0,23	0,27	0,03	0,07	0,08	0,09	-0,01	0,04	0,05	0,03	0,04
f48	0,20	0,31	-0,04	0,18	0,09	-0,06	0,09	0,36	0,27	0,39	0,18	0,29	1,00	0,11	0,15	-0,02	-0,02	-0,05	-0,03	0,08	-0,09	-0,06	-0,10	-0,04
f8	0,06	0,49	-0,03	0,22	0,19	-0,04	-0,06	-0,08	-0,02	-0,05	0,05	0,23	0,11	1,00	0,58	0,21	0,04	0,05	0,06	0,09	0,05	0,04	0,05	0,08
f9	0,10	0,57	-0,02	0,09	0,23	-0,03	-0,04	-0,07	-0,08	-0,04	0,24	0,27	0,15	0,58	1,00	0,18	0,09	0,01	0,06	0,21	0,06	0,00	-0,02	0,08
g10	0,08	0,09	-0,10	0,15	0,23	-0,03	0,12	0,03	0,16	0,02	0,17	0,03	-0,02	0,21	0,18	1,00	0,72	0,63	0,70	0,73	0,73	0,67	0,87	0,75
g12	0,11	0,06	0,02	0,28	0,11	0,12	0,03	-0,04	-0,02	-0,02	0,11	0,07	-0,02	0,04	0,09	0,72	1,00	0,91	0,94	0,60	0,91	0,92	0,73	0,94
g14	0,13	0,07	-0,05	0,22	0,18	-0,03	0,12	0,20	0,15	0,20	0,10	0,08	-0,05	0,05	0,01	0,63	0,91	1,00	0,95	0,56	0,85	0,89	0,69	0,86
g15	0,13	0,04	-0,03	0,31	0,18	-0,04	0,11	0,23	0,16	0,20	0,15	0,09	-0,03	0,06	0,06	0,70	0,94	0,95	1,00	0,53	0,85	0,88	0,70	0,88
g2	0,01	0,18	-0,08	0,08	0,15	0,02	0,07	0,08	0,12	-0,02	0,20	-0,01	0,08	0,09	0,21	0,60	0,60	0,56	0,53	1,00	0,63	0,59	0,74	0,60
g3	0,13	0,05	-0,04	0,19	0,09	0,12	0,04	0,20	0,12	0,17	0,13	0,04	-0,09	0,05	0,06	0,73	0,91	0,85	0,85	0,63	1,00	0,95	0,79	0,94
g6	0,16	0,10	-0,04	0,23	0,15	0,11	0,03	0,22	0,09	0,22	0,11	0,05	-0,06	0,04	0,00	0,67	0,92	0,89	0,88	0,59	0,95	1,00	0,73	0,93
g7	0,03	-0,05	-0,13	0,17	0,00	0,07	0,16	0,09	0,15	0,05	0,21	0,03	-0,10	0,05	-0,02	0,87	0,73	0,69	0,70	0,74	0,79	0,73	1,00	0,75
g8	0,14	0,11	-0,04	0,27	0,10	0,16	0,02	0,18	0,18	0,20	0,11	0,04	-0,04	0,08	0,08	0,75	0,94	0,86	0,88	0,60	0,94	0,93	0,75	1,00

Appendix E – correlation coefficients between allergen extracts and components

	bromelin	Phl p 12	rBet v 2	Phl p 1	Phl p 7	Phl p 5	Phl p 4
f10	0,66	-0,12	-0,06	0,54	-0,20	0,35	0,67
f11	0,75	-0,08	-0,02	0,36	-0,04	0,29	0,65
f12	0,67	-0,19	-0,13	0,24	0,05	0,14	0,46
f13	0,69	-0,11	-0,05	0,44	-0,26	0,42	0,66
f14	0,73	-0,20	-0,14	0,30	-0,08	0,21	0,54
f15	0,95	-0,12	-0,06	0,29	0,01	0,32	0,68
f20	0,77	-0,16	-0,09	0,26	0,10	0,11	0,52
f25	0,83	0,02	0,08	0,45	-0,11	0,49	0,81
f31	0,80	0,03	0,09	0,54	-0,09	0,38	0,75
f33	0,88	-0,01	0,05	0,40	-0,07	0,44	0,76
f35	0,92	0,00	0,06	0,46	-0,12	0,45	0,80
f36	0,84	0,04	0,10	0,43	0,02	0,33	0,69
f4	0,54	-0,25	-0,19	0,31	-0,22	0,27	0,53
f44	0,68	-0,15	-0,10	0,40	-0,14	0,28	0,60
f47	0,83	-0,22	-0,16	0,35	-0,16	0,39	0,70
f48	0,79	-0,09	-0,04	0,45	-0,20	0,50	0,78
f5	0,76	-0,24	-0,18	0,29	-0,09	0,25	0,58
f6	0,83	-0,17	-0,11	0,41	-0,20	0,37	0,70
f7	0,81	-0,20	-0,14	0,37	-0,08	0,33	0,64
f8	0,58	-0,16	-0,11	0,32	-0,20	0,24	0,54
f9	0,83	-0,27	-0,21	0,32	-0,16	0,34	0,65
g10	0,76	0,05	0,10	0,69	-0,24	0,58	0,89
g12	0,65	0,13	0,18	0,72	-0,38	0,82	0,94
g14	0,65	0,17	0,23	0,71	-0,37	0,80	0,95
g15	0,68	0,15	0,20	0,73	-0,37	0,78	0,97
g2	0,77	0,20	0,26	0,61	-0,11	0,56	0,88
g3	0,55	0,23	0,27	0,78	-0,41	0,82	0,92
g6	0,55	0,24	0,28	0,70	-0,45	0,83	0,92
g7	0,70	0,05	0,11	0,73	-0,27	0,62	0,91
g8	0,55	0,28	0,32	0,74	-0,39	0,82	0,91
w1	0,84	0,20	0,25	0,42	0,17	0,36	0,71
w21	0,81	0,09	0,15	0,52	-0,02	0,49	0,81
i1	0,80	-0,02	0,04	0,27	0,19	0,33	0,67
t9	0,64	0,16	0,21	0,66	-0,03	0,38	0,71
bromelin	1,00	-0,11	-0,05	0,35	0,00	0,39	0,74
Phl p 12	-0,11	1,00	0,96	0,02	0,14	0,16	0,16
rBet v 2	-0,05	0,96	1,00	0,09	0,09	0,23	0,22
Phl p 1	0,35	0,02	0,09	1,00	-0,45	0,80	0,70
Phl p 7	0,00	0,14	0,09	-0,45	1,00	-0,45	-0,33
Phl p 5	0,39	0,16	0,23	0,80	-0,45	1,00	0,74
Phl p 4	0,74	0,16	0,22	0,70	-0,33	0,74	1,00