

Prediction and Mapping of IgE Epitopes in the Allergenic Egg Proteins using Computational Approaches

Nitish Sharma

Delhi Technological University, Delhi, India

1.ABSTRACT

Food allergy is an emerging public health problem that is most prevalent during infancy, affecting up to 6% of young children. Food allergy denotes an immunologic mechanism represented almost exclusively by IgE-mediated reactions. Rapid advances have been made in the past few years on allergen characterization and sequence determination by biochemical and molecular biological approaches. However, the last decade has seen rapid progress in identification of allergenic proteins and prediction of both linear and conformational epitope based on sequence or structure information using bioinformatics software. This study aims at identifying potential allergenic proteins in the egg proteome and also predicting and mapping IgE epitopes in the predicted allergenic proteins in egg using *in silico* approaches. We have used the support vector machine module of AlgPred, based on amino acid and dipeptide composition, to predict highly allergenic proteins in egg proteome. The structures of some allergenic proteins that lack crystal structure information were predicted using ab-intio methods. The prediction of IgE epitopes were carried out on all the predicted allergenic proteins using SPADE, EPITOPIA, SEPPA and ELLIPRO. The potential allergenic proteins in the egg proteome were identified. We then predict linear and conformational epitopes in these allergenic proteins by various softwares using different approaches to compare the predicted IgE epitopes. We also characterized the epitope in terms of properties like solvent accessibility, electrostatic potential, hydrophobicity and total area of epitope. The results obtained have been correlated with experimental studies reported in the literature. The study for the first time reports a consensus report of the epitope patches for each allergenic protein in egg proteome predicted using different approaches. It is hoped that these results will be useful for epitope identification and characterization based on a given protein sequence and structure information and pave way for vaccine development for allergic patients in future.

2.INTRODUCTION

Food allergy is defined as the immunologic response against a particular food or its component. About 25% of population in industrialized countries suffer from various types of allergic reactions such as allergic asthma, rhinitis, food allergy, skin allergy and anaphylactic shock. These reactions are majorly caused due to Type 1 Hypersensitive reactions. There are three basic components of Type 1 hypersensitive reactions 1) Allergens 2) Immunoglobulin E and 3) Basophils/Mast cells (Anderson *et al.*, 1984). The IgE antibodies cross react with various allergens through fab region and binds their Fc arm to FcεR1 receptors on the surface of mast cells (Sampson *et al.*, 1999). The above series of events causes degranulation of mast cells and stimulates them to release mediators such as histamines, leukotriene's and prostaglandins which are powerful compounds that cause allergic reactions. Basophils are the type of granulocytes that are phagocytic in nature. Mast cells are located in the lining of skin, GI tract and lungs. Both mast cells and basophils release histamines and other mediators that cause allergic reactions (Janeway *et al.*, 2001).

In the last 10 years the prevalence of food allergy has increased significantly. Recent data suggest that about 8% of children and 4% of adults are affected with food allergy. In US, prevalence of food allergy has increased by 18%. Food allergies alone have affected 2.4% and 3.24% of Dutch and French population, respectively. The eight most common food allergens are milk, eggs, fish, soy, shellfish, wheat, peanuts and tree nuts.

Egg allergy is one of the most prevalent food allergies. 3.2% of Australian and South Asian population are affected with egg allergy. It affects 2% of the children and can cause severe allergic reactions. It can even lead to severe anaphylactic shock. 12% of children show remission of type 1 hypersensitive reactions in adolescence which is later carried into adulthood (Wood *et al.*, 2003). The inclusion of egg in various foods as binding and emulsifying agent prevents complete egg avoidance. The white portion of egg which is often referred as 'Egg White' is the main cause of egg allergy in atopic individuals. This portion is rich in proteins such as ovalbumin, ovotransferrin, lysozyme, ovomucoid etc. which are major cause of egg allergy.

In this study, we have used the support vector machine module of AlgPred (Raghava *et al.*, 2006) to predict allergenic proteins in egg proteome. The structures of some allergenic proteins that lack crystal structure information were predicted using *ab-intio* methods. The IgE epitopes were predicted and mapped on the 3D structure of allergenic using SPADE (Dall'Antonia *et al.*, 2011), EPITOPIA (Rubinstein *et al.*, 2006), SEPPA (Sun *et al.*, 2009) and ELLIPRO (Ponomarenko *et al.*, 2008). We have predicted allergenic proteins in egg proteome. We have also predicted and mapped linear as well as conformational epitopes using different approaches. The epitopes are characterized in terms of various properties such as solvent accessibility, electrostatic potential, hydrophobicity and total area of epitope. The results are well correlated with experimental studies. For the first time, we have reported a consensus report of the epitope patches for each allergenic protein in egg proteome predicted using different approaches.

3. REVIEW OF LITERATURE

3.1 FOOD ALLERGY

Food allergy is defined as the adverse immunologic response to components present in food (Anderson *et al.*, 1984). A single food can contain multitude of food allergens. The allergens can be carbohydrate moieties, lipids or fats, and more generally proteins. These allergic reactions are responsible for variety of symptoms and affect the skin, gastrointestinal tract, and respiratory tract (Sampson *et al.*, 1999). They are caused either by IgE mediated reactions or Non-IgE mediated mechanisms. Our understanding about the food allergens and how they suppress our tolerance mechanism is evolving. Any food can cause allergy but there are few foods that are highly allergic. The foods such as milk, egg, peanuts, tree nuts, fish, and shellfish causes vast majority of allergic reactions (Burks *et al.*, 1999). The food constituents that cause adverse reactions during food intolerance are categorized as toxins (e.g., food poisoning), pharmacologic agents (e.g., caffeine or tyramine), and host factors such as metabolic disorders (e.g., lactase deficiency).

3.1.1 EPIDEMIOLOGY

In United States, 6% of young children and 3.7% of adults suffer from food allergy (Sicherer *et al.*, 2004). The incidence of food allergens in young children is enlisted below (Wood *et al.*, 2003)-:

Allergen	Incidence in Young Children
Cow's milk	2.5%
Egg	1.3%
Peanut	0.8%
Wheat	Approx. 0.4%
Soy	Approx. 0.4%
Tree nuts	0.2%
Fish	0.1%
Shellfish	0.1%

Table 1:- Percentage of children affected by various allergens.

Approximately 80% early childhood Allergy towards milk, egg, soy, and wheat subside by school age (Hourihane JO *et al.*, 1998). Allergy towards peanut, tree nut and sea food remain permanent and young children show remission of symptoms by the age of 5. Adults are more prone to shell fish, peanut, tree nut and fish allergens. The incidence of food allergens in adults is enlisted below (Fleischer *et al.*, 2003)-:

Allergen	Incidence in Adult
Shellfish	2%
Peanut	0.6%
Tree nut	0.5%
Fish	0.4%
Fruits and Vegetable	Approx. 5%

Table 2-: Percentage of adults affected by various allergens.

3.1.2 CASE STUDY

- Cow's milk is responsible for majority of allergic reactions in infants. 3 separate studies were carried out in Sweden (Jakobsson *et al.*, 1979), Denmark (Host A *et al.*, 1990) and Netherlands (Schrandt *et al.*, 1994) respectively on the basis of 'oral food challenge' to detect allergic reactions due to cow's milk. These studies reported the prevalence of milk allergy of 1.9%, 2.2%, and 2.8%, respectively.
- In another study, 165 children with a mean age of 4 years, 7 foods accounted for 89% of the positive challenges: milk, egg, peanut, soy, wheat, fish, and tree nuts. During these challenges, 27% responded with gastrointestinal symptoms, and 7% of the total group experienced isolated gastrointestinal symptoms (Burks *et al.*, 1998).

For majority of allergic reactions, these studies suggest that food allergy is more common among infants and young children as compared to adults.

3.1.3 PATHOPHYSIOLOGY

Allergic reactions are hyperactive responses of the immune system to generally innocuous substances. There are basically 2 types of response-:

3.1.3.1 ACUTE RESPONSE

This response is mediated by immunoglobulin (Ig) E antibodies specific to particular food proteins. These food-specific IgE antibodies bind high-affinity receptors on the surfaces of mast cells and basophils (Janeway *et al.*, 2001). When the food protein penetrates mucosal barriers, binds, and cross-links these antibodies, the cells are activated and release mediators such as histamine, prostaglandins, and leukotrienes that initiate vasodilatation, mucous secretion, smooth muscle contraction, and influx of other inflammatory cells. The symptoms include vomiting, abdominal pain, diarrhea, and oropharyngeal pruritus, skin symptoms such as urticaria, angioedema, upper and lower airway symptoms (rhinitis or wheezing), and cardiovascular symptoms, including anaphylactic shock.

3.1.3.2 LATE-PHASE RESPONSE

This response occurs after the dwindling of active response. It is not mediated by IgE. This phase is governed by the secretion of various cytokines. The cytokines are secreted by antigen presenting cells or T cells after recognition of food antigenic proteins (Holt *et al.*, 2007). The process is accompanied by the migration of leukocytes such as neutrophils, lymphocytes, eosinophils and macrophages at the site of action. The reaction occurs after 2-24 hours of acute response. Late phase responses seen in asthma are slightly different from those seen in other allergic responses, although they are still caused by release of mediators from eosinophils, and are still dependent on activity of T_H2 cells (Grimbaldeston *et al.*, 2006).

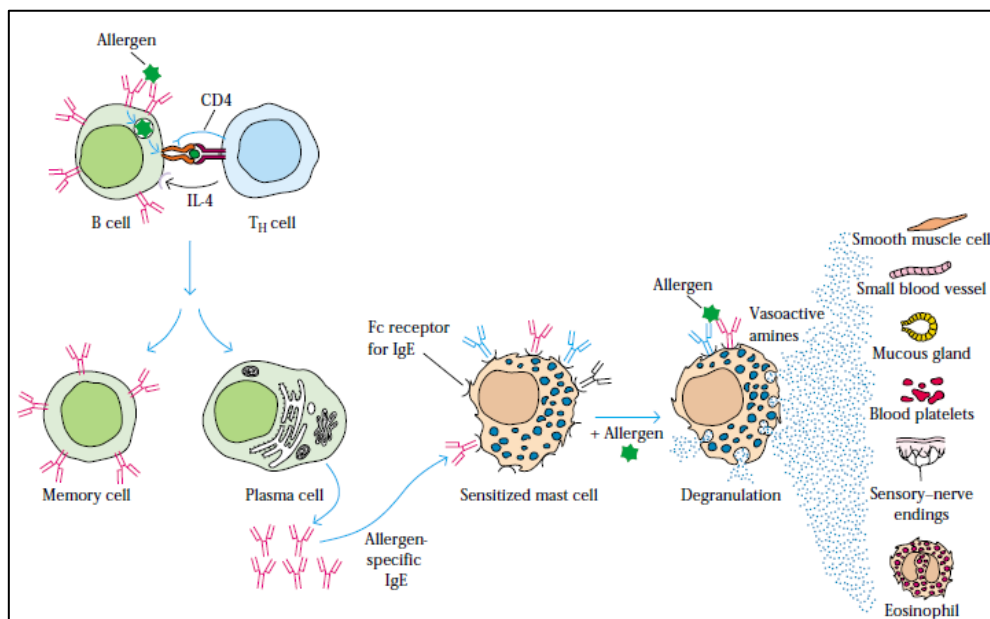


Figure1-: General mechanism of type 1 hypersensitivity (Kindt *et al.*, 2006).

MEDIATORS	PRIMARY EFFECTS
Histamine, heparin	Increased vascular permeability; smooth-muscle contraction
Serotonin	Increased vascular permeability; smooth-muscle contraction
Eosinophil chemotactic factor	Eosinophil chemotaxis
Neutrophil chemotactic factor	Neutrophil chemotaxis
Proteases	Bronchial mucus secretion; degradation of blood-vessel basement membrane, generation of complement split products
SECONDARY EFFECTS	
Platelet-activating factor	Platelet aggregation and degranulation; contraction of pulmonary smooth muscles
Leukotrienes	Platelet aggregation and degranulation; contraction of pulmonary smooth muscles
Prostaglandins	Vasodilation; contraction of pulmonary smooth muscles; platelet aggregation

Bradykinin	Increased vascular permeability; smooth-muscle contraction
Cytokines	Systemic anaphylaxis; increased expression of CAMs on venular endothelial cells

Table 3-: The effect of various mediators in hypersensitive reactions.

3.1.4 DIAGNOSIS

A recent consensus workshop (Workshop on the Classification of Gastrointestinal Diseases of Infants and Children, November 1998, Washington, DC) (Sampson *et al.*, 2000) considered a variety of factors in establishing a diagnosis of food allergy:

- History of an allergic or allergic-like hypersensitivity reaction to food ingestion.
- Exclusion of anatomic, functional, metabolic, or infectious causes.
- Pathologic findings consistent with an allergic cause (usually eosinophilia).
- Confirmation of a relationship between the ingestion of specific food to the development of symptoms by clinical challenges or repeated, inadvertent exposures.
- Evidence of the food-specific IgE antibody in settings of IgE-mediated disease.
- Failure to respond to conventional therapies aimed at anatomic, functional, metabolic, or infectious causes.
- Improvement in symptoms with elimination of the causal dietary proteins.
- Clinical response to treatments of allergic inflammation (i.e., corticosteroids).
- Similarities to clinical syndromes either proven or presumed to be caused by immunologic mechanisms.
- Lack of other explanations for the clinical allergic-like reaction.

3.1.5 LABORATORY TESTS

There are specific tests to identify foods causing allergic reactions.

3.1.5.1 Skin Prick Test-: A device such as bifurcated needle or lancet is used to puncture the skin. The device is filled with the food of interest or small amount of food sample is directly applied to the skin. The skin is punctured through these devices. This puts small amount of allergen under the skin. A wheal-and-flare response at the site indicates the presence of food-specific IgE antibodies. A wheal .3 mm is considered positive. The negative predictive value of skin prick test is most specific (Sampson *et al.*, 1984). Thus, negative skin prick tests are widely used as diagnostics. The predictive value of positive skin prick test is 50%. Thus, they cannot be used to assure hypersensitivity. Intradermal skin testing is prohibited because they give high false positive response and can cause fatal anaphylactic reactions (Bock *et al.*, 1978).

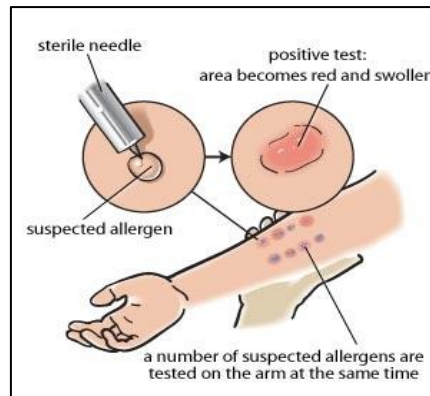


Figure 2-: Skin prick test for diagnosing allergy.

3.1.5.2 RAST-: Radio Allergosorbent Test. The suspected allergen is bound to an insoluble material and the patient's serum is added. If the serum contains antibodies to the allergen, those antibodies will bind to the allergen. Radiolabeled anti-human IgE antibody is added where it binds to those IgE antibodies already bound to the insoluble material. The unbound anti-human IgE antibodies are washed away. The amount of radioactivity is proportional to the serum IgE for the allergen. In vitro tests for a specific IgE RAST are also helpful in the evaluation of IgE-mediated food allergy. Like skin tests, a negative result is very reliable in ruling out an IgE-mediated reaction to a particular food, but a positive result has low specificity (Sampson *et al.*, 1997).

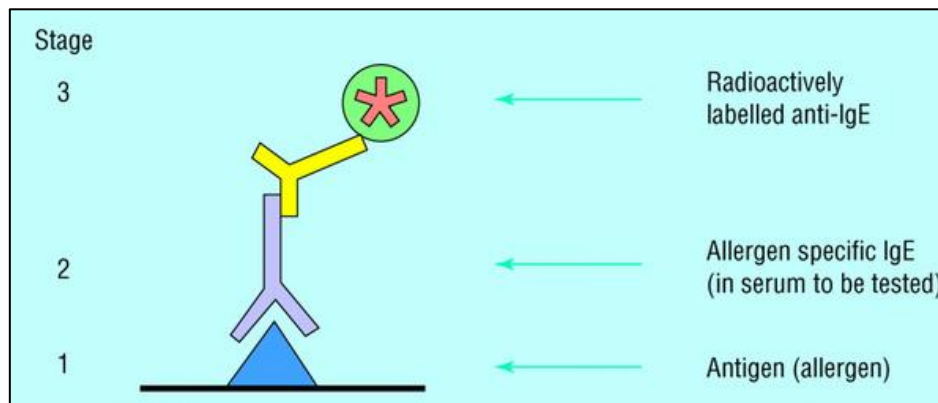


Figure 3-: RAST test for diagnosing allergy.

3.1.5.3 ELIMINATION DIETS

It involves the elimination of all forms of suspected food and observing the subsidence of symptoms. There are 3 types of elimination diets:

- The food-specific elimination diet.
- The oligoantigenic diet.
- The elemental diet.

The first type involves the elimination of those foods that may provoke an acute response and there is positive test for IgE to that food. This would also represent a therapeutic intervention.

In oligoantigenic diet, large numbers of foods that are suspected of eliciting immunologic response are eliminated from the diet. Patient is given a list of allowed foods. An example of such a diet may be one that includes lamb, rice, corn, cooked apple, broccoli, asparagus, spinach, lettuce, sweet potato, salt, sugar, vinegar, and olive oil. Individualization for this type of elimination diet is almost always needed. The advantage of this diet is that a nutritionally balanced, palatable diet is maintained while most of the likely causal foods are removed. If symptoms persist then cause is implicated on the foods left in the diet.

In elimination diet, calories are obtained from amino acids. A variation including a few foods is likely to be tolerated. This diet is generally difficult to maintain in patients beyond infancy. This diet may be required when the less restrictive diets fail to resolve symptoms, but suspicion for food-related illness remains high. Elemental diets are generally required in disorders associated with multiple food allergies, such as EG (Eosinophilic gastroenteritis).

3.1.5.4 DBPCFC-: Double-blind placebo-controlled food challenges (DBPCFC), are the gold standard for diagnosis of food allergies, including most non-IgE mediated reactions. Blind food challenges involve packaging the suspected allergen into a capsule, giving it to the patient, and observing the patient for signs or symptoms of an allergic reaction. Due to the risk of anaphylaxis, food challenges are usually conducted in a hospital environment in the presence of a doctor.

Additional diagnostic tools for evaluation of eosinophilic or non-IgE mediated reactions include endoscopy, colonoscopy, and biopsy.

3.1.6 MANAGEMENT

The basic therapy for food allergy is to avoid the suspected food. Medical identification jewellery is recommended to patients and instructions are given to them regarding its usage. They are also taught about self-injectable epinephrine (Bindslev-Jensen *et al.*, 1991). Comprehensive educational materials are available through organizations such as the Food Allergy & Anaphylaxis Network. The available treatments are-:

3.1.6.1 EPINEPHRINE-: It is also called as adrenaline. It is generally used to treat allergy symptoms. It improves blood circulation by tightening blood vessels. This increases the heart rate and rate of blood circulation in the body increases. Epinephrine is also prescribed by a physician in a form that is self-injectable. This is what is called an epi-pen (Sicherer, H. 2006).

3.1.6.2 ANTIHISTAMINES-: They block the action of histamine. Histamines causes the dilation of blood vessels and they became leaky to plasma proteins. The widely used antihistamine is diphenhydramine, also known as Benedryl. In case of anaphylaxis, they do not completely improve the dangerous symptoms that affect breathing (Nowak-Wegrzyn *et al.*, 2004).

3.1.6.3 STEROIDS- they alleviate the immune system cells that are triggered by chemicals released during an allergic reaction. It is not used to treat anaphylaxis because steroids relieve symptoms only in the area of contact. It takes long time to start its action. Steroids can also be taken orally or through injection. By taking a steroid in these manners, every part of the body can be reached and treated, but a long time is usually needed for these to take effect (Leung *et al.*, 2003).

3.1.6.4 OTHER THERAPIES- Novel Injections of anti-IgE antibodies (TNX- 901) have been developed. They show improvement in patients allergic to peanuts but 25% group shows no improvement (Nowak-Wegrzyn *et al.*, 2004). Traditional Chinese herbs showed efficacy in a murine model of peanut- induced anaphylaxis (Leung *et al.*, 2003). Engineered proteins are developed that lack IgE binding sites and overlapping peptides. These sites cause binding of allergens to IgE.

3.2 MOLECULAR PROPERTIES OF FOOD ALLERGENS

Nowadays, *in silico* approaches are explored for identification of potential allergens. Various methods have been proposed for identifying allergens computationally.

- A query protein is regarded as potentially cross-reactive if it has an identity of at least 6 continuous amino acids or more than 35% sequence similarity over a window of 80 amino acids when compared with known allergens.
- 59 proteins were analysed using FASTA algorithm and it was suggested that an 8-amino-acid window was more appropriate because it reduced the frequency of spurious matches (Hileman *et al.*, 2002).
- 52 sequence motifs were identified from a comprehensive allergen database and integrated with pairwise sequence alignment approach. This strategy is used to predict allergenic proteins (Stadler *et al.*, 2003).
- In another strategy, homology searching is used to identify short contiguous sequences between allergens and query proteins, followed by an analysis of their potential antigenicity (Kleter *et al.*, 2002).
- Another approach combines sequence and structural information to identify whether query sequences match known IgE epitopes. This has been used to identify potential cross-reactive IgE epitopes for the thaumatin-like pathogenesis-related protein (PR) protein allergen of cedar pollen (Jun a 3) (Pomes, A. 2002).

No single criterion can sufficiently predict allergenicity. There are some biochemical properties associated with food allergens such as presence of multiple, linear IgE binding epitopes and the resistance of the protein to digestion and processing. Some important physiochemical and functional properties shared by food allergens will be discussed here.

3.2.1 LIGAND BINDING

Majority of food allergens are ligand binding proteins. Their substrate ranges from metal ions to lipids. Sometimes metal ions get integrated with the protein and are buried deep within the molecule. The loss of metal ions has many detrimental effects such as improper protein folding and transition into partially folded forms. Binding of ligand can occur in many ways:-

- Some proteins form a cavity into which a ligand fits. This might be a metal ion, steroids, or a variety of lipid molecules.
- Proteins possess a tunnel into which ligands fit.
- Some proteins bind ligands through superficial surface interactions.

Ligand binding reduces mobility of polypeptide backbone. It increases thermal stability and resistance to proteolysis. Proteins such as the lipocalins and nonspecific lipid-transfer proteins (nsLTPs), which possess a lipid-binding pocket, show increased stability when the pocket is occupied. Thus the thermostability of beta lactoglobulin (bLg) increases on lipid binding (Creamer, LK. 1995).

3.2.2 INTERACTION WITH MEMBRANES AND LIPIDS

Many food allergens are able to associate with cell membranes and other types of lipid structures formed in foods. An example of this mode of action occurs when proteins protect plants against microbial pathogens through destabilization of bacterial or fungal membranes resulting in leakage (Selitrennikoff *et al.*, 2001). Proteins acting in this way include thionins, thaumatin-like proteins (TLPs), 2 types of prolamin superfamily members (2S albumins and nsLTPs), and some defensins.

3.2.3 PROTEIN STABILITY AND MOBILITY

The term stability describes the ability of a protein to retain its original 3-dimensional structure after treatments with chemicals (urea), physical stress (temperature) and resistance to degradation by proteases. There are some indications that thermostable proteins might have a greater propensity to adopt beta-structures. This is because they have low heat capacities. The small loops in thermostable proteins lead to a smaller difference in entropy between the folded and unfolded states (Chakravarty *et al.*, 2000). This leads to stabilization of protein.

3.2.3.1 DISULFIDE BONDS

Disulfide bonds are majorly responsible for the stability of proteins. The protein structure is stabilized by intrachain or interchain disulphide bond. If the structure is perturbed by heat or chemicals, disulphide bond reverses the perturbation and helps protein to retain its native state. There are some allergens that are highly disulphide bonded such as the prolamin

superfamily (nsLTP, 2S albumin, and inhibitors of trypsin and α -amylase found in cereals, together with the TLPs).

3.2.3.2 RHEOMORPHIC PROTEINS

This class of proteins contains large regions of disordered structure. There is high probability that domains and motifs of globular proteins have disordered structure. Such proteins are dynamic and their polypeptide chains adopt a series of secondary structures that are in equilibrium with unfolded, denatured and partially folded proteins (Dunker *et al.*, 2001). Such proteins are termed as rheomorphic proteins. These proteins are highly thermostable and don't undergo sharp transition under any stress. They also possess many thermostable epitopes. In addition to caseins, the seed storage prolamins can also be considered rheomorphic (Shewry *et al.*, 1999).

3.2.4 GLYCOSYLATION

Majority of food allergens undergo glycosylation after passing through endoplasmic reticulum. A high prevalence of anti-carbohydrate IgE is reported in patients with multiple pollen sensitizations. Glycosylation affects the biological properties of allergens. N-glycosylation can have a significant stabilizing effect on protein structure. There is evidence that it increases the stability of, for example, the 7S globulin of pea and its resistance to chemical denaturation (Pedrosa *et al.*, 2000).

3.2.5 REPETITIVE STRUCTURES, AGGREGATES AND GLYCATION

The sensitization and aggregation under any physiologic condition are affected by presence of repetitive structures and ability to form aggregates. This enhances the immunogenicity of the allergen. It also elicits the histamine release by mast cells. One major epitope site recognized by parasite-neutralizing antibodies in malaria corresponds to a serine-rich repeat sequence region. Repetitive structures are also a characteristic feature of many rheomorphic proteins, with the seed storage prolamins probably exhibiting the most degenerate repeat sequences. These are based on several different short motifs, ranging from 4 to 8 residues in length, which are rich in proline and glutamine (Foetisch *et al.*, 2003).

3.3 EGG ALLERGY AND EGG PROTEOME

3.3.1 EGG ALLERGY

Allergy towards egg is more common in early childhood. It affects 1-2% of preschool children. In most cases, children show resolution of symptoms as they enter adolescent stage. But in some cases remission of type 1 hypersensitivity to hen egg occurs and hypersensitivity

may persist through adolescence into adulthood (Boyano-Martínez *et al.*, 2002). Resolution of egg allergy may first manifest with tolerance to cooked egg products despite continued reaction to raw egg, whilst in others, allergy may persist to egg in any form (Urisu *et al.*, 1997). The use of egg in many foods as binding and emulsifying agent prevents complete egg avoidance. The sooner tolerance to egg is ascertained, the sooner a child can enjoy a normal, unrestricted diet.

3.3.1.1 DIAGNOSIS

Diagnosis is generally made through a combination of skin prick testing or blood testing i.e. RAST and detailed records of all foods and drink the person regularly ingests.

3.3.1.2 TREATMENT

There is currently no cure for egg allergy.

- Most people who are allergic to eggs avoid eating any form of egg or egg component.
- For people with a more serious allergic reaction to eggs, urticaria (hives) and inflammation can occur and as such, doctors suggest that the person carries around an EpiPen.

3.3.1.3 EGG WHITE INTOLERANCE

Egg white causes release of histamines and sometimes provoke a non-allergic response in some people. The proteins in egg white directly interact with the mast cells and they release histamines. This mechanism is called as pseudoallergy (Arnaldo, C. 2008). This is because no IgE is triggered. Hence, it is also called as food intolerance.

The response is localized, mainly affects gastrointestinal tract. Symptoms include:-

- Abdominal pain
- Diarrhoea
- Symptoms of histamine release

If sufficiently strong, it can result in an anaphylactoid reaction, which is clinically indistinguishable from true anaphylaxis (Joris *et al.*, 2004).

Some people with this condition tolerate small quantities of egg whites. They are more often able to tolerate well-cooked eggs, such as found in cake or dried egg-based pasta than loosely cooked eggs, such as fried eggs or uncooked eggs.

3.3.2 EGG PROTEOME

Egg represents a major raw material for the food industry because of its technological properties. It is generally used for foaming and gelling.

- The structure and functionality of major egg proteins have been widely studied in various physicochemical conditions (Li-Chan *et al.*, 1989).
- A new way to increase the value of egg products could be the extraction of biologically active molecules, especially proteins.

- The proteins have very different molecular masses (12.7-8000 kDa) and pI values (Rabilloud *et al.*, 2000).
- Their concentration differs highly from one protein to another.
- Ovalbumin represents more than 50% of total proteins.
- In addition to well-known proteins such as lysozyme or ovotransferrin, there are antimicrobial or antiviral proteins, transport proteins, or growth factors.

The proteins present in egg are classified as-:

3.3.2.1 SERPIN FAMILY

This protein family is essentially represented by the major hen egg white protein, i.e. ovalbumin (Nisbet *et al.*, 1981).

- It comprises of 54% of total proteins.
- It is a glycoprotein and has an isoelectric point of 4.5.
- It is 385 amino acids long.
- It has four cysteine residues and a single cysteine disulphide bridge.
- Ovalbumin has two further sites of modification: the N-terminus is acetylated, and the carbohydrate moiety is linked through asparagine 292.
- Two polymorphic forms of ovalbumin are known, ovalbumin A and ovalbumin B, and these differ in having asparagine and aspartic acid respectively at position 311.
- It shares homology with a group of proteinase inhibitors known as serpins. It was found to have 30% sequence homology with the archetype member of the family α_1 antitrypsin.

3.3.2.2 TRANSFERIN FAMILY

Ovotransferrin is the only member of this family.

- Ovotransferrin is a glycoprotein which occurs in egg white, egg yolk and in plasma.
- The proteins from all three sources have the same amino acid sequence, but there are slight differences in the glycosylation (Williams *et al.*, 1982).
- The protein has molecular mass of 80,000 and is made up of two domains with a short linking region (Williams, 1982).
- The two domains can be separated after proteolysis of the linking region.
- The protein is rich in disulphide bridges, having six in the N-domain and nine in the C-domain, giving the protein high stability.
- The function of ovotransferrin is iron transport. It binds two atoms of Fe, one in each domain.

3.3.2.3 KAZAL FAMILY

It includes two proteins-: ovomucoid and ovomucoid inhibitor. Ovomucoid makes up 10% of the protein in egg white.

- It is a heat stable glycoprotein of 185 amino acid residues and nine disulphide bridges.
- It is the disulphide bridges that account for its heat stability.
- The sequence comprises three homologous tandem domains which are believed to have arisen through two gene duplications (Kato *et al.*, 1978).
- Domains I and II are referred to as a-type domains and show greater similarity to each other than to domain III, which is known as a b-type domain.
- The mechanism of inhibition occurs in two steps:- The inhibitor is bound by the enzyme (E); afterwards cleavage of a single peptide bond occurs to form a modified inhibitor. A stable inhibitory complex is formed which only dissociates very slowly.

Ovoinhibitor is also an inhibitor of serine proteinases, and is similar to ovomucoid in its properties.

- It is larger than ovomucoid, having Mol. mass, value of 49,000.
- It comprises of seven domains and possesses a similar arrangement of disulphide bridges to that of ovomucoid.
- Six of the domains are of the a-type, and the seventh, which occupies the C-terminus, is a b-type.
- One molecule of ovoinhibitor is able to inhibit two molecules of trypsin and two of chymotrypsin, each binding to different domains.

3.3.2.4 GLYCOSYL HYDROLASES

Lysozyme C represents this family.

- Lysozyme is unusual among the major egg white proteins in having an alkaline pI, which means that it can form complexes with ovomucin, ovalbumin and ovotransferrin.
- It has a total of 129 amino acid residues and contains four disulphide bridges.
- Its enzyme activity is that it is able to cleave peptidoglycans, such as are found in the cell walls of bacteria.
- Its role in egg is that of protection from invading bacteria.
- The amino acid residues involved in the catalysis are aspartate-52 and glutamate-35.

3.3.2.5 LIPOCALIN FAMILY

Lipocalins are transporters for small hydrophobic molecules, such as lipids, steroid hormones, and retinoids. This family is represented by 3 proteins in hen's egg. Extra fatty acid binding protein (Ex-FABP), also called Ch21 protein or quiescence specific protein. CAL- γ is second major representative of this family. It takes part in endochondral bone formation. Third member is ovoglycoprotein. Less information is available about this protein.

3.3.2.6 BPI FAMILY-: (Bactericidal Permeability-Increasing Protein)

Tenp is a major representative of this family.

- Tenp protein sequence shares homology with the BPI2 domain.
- The biological activity assumed for such a BPI protein is the binding to the Lipid A component of lipopolysaccharide from the outer envelope of Gram-negative bacteria.
- The toxic action of BPI against Gram-negative bacteria occurs in two stages: The binding of BPI causes immediate bacterial growth arrest linked to alterations in the outer membrane, followed later by bactericidal events coincident with damage to the inner membrane.
- Tenp could then participate in the antibacterial activity of hen egg white.

3.3.2.7 CLUSTERIN FAMILY

Clusterin is the major representative of this family.

- Clusterin is a ubiquitous and highly conserved secreted glycoprotein.
- It has been found in numerous biological fluids including semen, urine, and human plasma. Clusterin is present in hen egg white, as already immunodetected in several chicken tissues such as magnum, egg shell and egg white.
- Clusterin is a member of the chaperone proteins, which interact and stabilize unfolded or partly folded proteins, preventing their aggregation or precipitation.

3.3.2.8 CYSTEINE PROTEASE FAMILY

Cystatin is the major representative of this family.

- Cystatin is known as ficin inhibitor.
- There are two major forms of cystatin having pI values of 6.5 and 5.6 referred to as A and B (Turk *et al.*, 1983).
- Each of the two forms exists in short and long forms, the former lacking the first eight amino acid residues present in the 116 residue polypeptide chain of the latter.
- The two major forms are immunologically identical and neither contains any carbohydrate.
- Cystatin inhibits a number of cysteine proteinases including ficin, papain, cathepsin B, cathepsin H, cathepsin L and dipeptidyl peptidase I, but not clostripain or streptococcal proteinase, and it only weakly inhibits bromelain.

3.3.2.9 VMO-1 FAMILY

The VMO-1 protein is one of the proteins identified in the outer layer of egg vitelline membranes.

- VMO-1, VMO-2, and lysozyme bind tightly to ovomucin and participate in the vitelline membrane structure.
- The molecular mass of VMO-1 is 17 k Da.
- The molecular mass of VMO-2 is 8 kDa.

3.3.2.10 FOLATE RECEPTOR FAMILY

Riboflavin binding protein is the most abundant vitamin binding protein in egg white, making up approximately 1% of the protein content.

- It has nine disulphide bridges, and this probably accounts in part for its high thermal stability.
- Solutions of RFBP can be boiled for 30 min without denaturation.
- The protein has a total of eight phosphate groups which together with the acidic amino acid residues and sialic acid account for its low pI of 4.0.
- It has two oligosaccharide groups attached to asparagine 36 and 147.

3.3.2.11 OVOSTATIN

Ovostatin (formerly known as ovomaeroglobulin) is a large molecule having a tetrameric structure

- Its mol. mass is $780,000 = 4 \times 195,000$.
- It inhibits a wide range of endoproteinases including thermolysin (a metal-ion requiring proteinase) and collagenase (Nagase *et al.*, 1983).
- Its structure and mechanism of action is like that of the serum proteinase inhibitor, α_2 macroglobulin.
- The proteinases first cleave a bond within ovostatin, which then undergoes a conformational change so as to hinder the access of large, but not small substrate molecules to the catalytic site.
- Ovostatin shows 40% homology with α_2 macroglobulin.

3.3.2.12 THIAMIN BINDING PROTEIN

- Thiamin binding protein has been purified from egg white by affinity (Munniyappa and Adiga, 1979).
- It has Mol. Mass of 38,000 and is not a glycoprotein.
- A similar protein has been purified from egg yolk (Munniyappa *et al.*, 1981) which cross reacts with monospecific antiserum to egg white thiamin binding protein, suggesting that both are products of the same gene, although they may differ in posttranslational modification.

3.4 ALLERGEN DATABASES

3.4.1 NEED FOR SPECIALIZED DATABASES

Allergen databases derive their information from primary databases and also provide additional features for classifying the data. Primary databases are repositories for biological data and don't provide specific data (Schönbach; Ranganathan; Brusica. 2008). For example the keywords used in GenBank are not specific and will yield large number of false positives (Malandain, H. 2004).

- Specialized databases will collect and validate allergen specific data from primary databases. For example the data in GenBank is of low quality because they are dependent on submitters to check the record of data to be submitted. Thus requirement of specialized, manually curated database is pertinent.
- Primary databases are biased towards data. They contain data of only one type. For example GenBank contains only nucleotide related data. Thus, allergen specific databases will contain information and data related to allergens and will act as one stop shop for researchers.
- The allergen information in primary databases is not classified. Classification helps the researchers to comprehend the data meaningfully. The most common form of classifying an allergen is on the basis of source, for example, food allergen.
- The specialized databases should contain better tools. They should have search tools pertaining to allergies. The use of such search fields will allow researchers to extract data quickly and accurately. Some additional bioinformatics application should be integrated with the databases that will aid in analysis of allergens.

3.4.2 FEATURES OF ALLERGEN DATABASES

The database should have following features (Schönbach; Ranganathan; Brusica. 2008):-

- The database should be as comprehensible as possible. It is difficult to create a comprehensive database that can act as one stop shop for allergen data. Databases that provide specific information about allergens are created.
- The database should contain non-redundant entries. Redundancy generates under-represented, over-represented data and can lead to errors in allergen analysis. Sequence similarity methods like Blast are used to remove redundancy.
- Each source database should contain different type of data. The databases should have common data format.
- The fields contained should be useful for allergen researchers. For example the common fields should include nucleotide sequence, protein sequence, literature references and 3-D structure of protein.
- The names of allergen should comply with standard nomenclature set out by Allergen Nomenclature subcommittee of IUIS (International Union of Immunological Societies). This will help avoiding the name conflicts.
- Manual curation should be applied to avoid conflict arising from multiple source databases.
- The database should be updated on timely basis. This ensures that the information is up to date.
- Some information is only present in the literature. Such information should be extracted manually although it is time consuming.
- The source databases contain errors. Such errors are resolved manually. The task is time and effort consuming.

3.4.3 ALLERGEN DATABASES

Given below is the list of some existing allergen databases.

Name	URL
Allallergy	http://www.allallergy.net/
Allergome	http://www.allergome.org/
BIFS (Biotechnology Information for Food Safety)	www.iit.edu/~sgendel/fa.htm
CSL (Central Science Laboratory) allergen database	http://allergen.csl.gov.uk/
FARRP (Food Allergen Research and Resource Program) allergy database	http://www.allergenonline.org/
IUIS list	http://www.allergen.org/
Protall	http://www.ifr.ac.uk/protall/
SDAP	https://fermi.utmb.edu/SDAP/
Swiss-prot allergen list	http://web.expasy.org/cgi-bin/unavailable.cgi?type=redirect&query=lists?allergen.txt

Table 4-: The list of allergen databases with their respective url's.

3.4.3.1 IUIS

A list of allergens and isoallergens is maintained by Allergen Nomenclature subcommittee of IUIS. Example Bet v 1, first three characters represent allergen name derived from genus name (bet = betula). The next character denotes species name (v = verrucosa). The number at the end indicates order in which the allergen is identified. The isoallergens are represented as Bet v 1.0101. First two numbers refer to isoallergen. Third and fourth variant indicate variant of isoallergen. The list is updated regularly and is available on internet. New allergens can be submitted by allergens and they should satisfy the prevalence of at least 5 % IgE reactivity or minimum of 5 patients showing IgE reactivity. Allergens are classified according to allergen source and each record contains species name, allergen name, protein name, molecular weight, type of sequence, database accession and literature references.

3.4.3.2 SWISS-PROT

Swiss-Prot maintains a list of allergens that currently number 347 entries. Each allergen is linked to Swiss-Prot record. The names of allergen are in accordance with the nomenclature set out by IUIS.

3.4.3.3 SDAP (Structural Database of Allergenic Proteins)

It is a specialized allergen database that incorporates information obtained from the IUIS list of allergens, Swiss-Prot, PIR(Protein Information Resource), GenBank, Genpept and literature. It contains 1526 allergens and isoallergens, 1312 protein sequences, and 29 allergens with IgE epitopes. Each record contains the name of allergen, species of origin,

protein sequences, nucleotide sequences, protein domains, 3-D protein structure, and IgE epitopes. It contains IgE epitope information that is extracted from the literature.

Records can be searched by their names, allergen source, description, and allergen type.

Allergens are compiled in-:

- Alphabetical order
- Containing PDB structures
- Containing 3D model
- Containing epitopes
- Class of allergens

It is integrated with computational tools that aid the researchers to analyse data efficiently. It is integrated with-:

- FASTA
- Sequence similarity search
- Allergen analysis
- Allergenicity prediction

The allergenicity test computes the allergenicity of given protein against the dataset present in SDAP. It has two data searching tool-:

- Exact matching tool for searching a query protein sequence against SDAP. This method is useful if the query protein sequence is an epitope. Any SDAP allergen having the same sequence as subsequence will be retrieved. The result is used to determine cross reactivity between query sequence and matched allergen. Link is provided if SDAP allergen has matched epitope. It only detects allergens with identical epitopes.
- The other method employs property distance function to score the similarity between two peptides. This PD function employs 5 descriptors E_1 - E_5 that are derived from 237 amino acid properties. For a given protein sequence the PD function is used to determine the similarity measure of novel protein sequence against all same length subsequence in SDAP. The results are ranked and displayed in the form of histogram. If a match is detected that has much lower similarity measure than the rest of the matches, match is considered significant and is analysed further. This method is used to detect cross-reactivity among the allergens on the basis of similar epitopes.
- Downloading of data is not supported.

3.4.3.4 ALLERGOME

It was started in 2000 and released in February, 2003. All records are manually curated. The data for allergome is literature published since 1960s. All the allergens in allergome are not found in IUIS list. Allergens that are not in IUIS are carefully checked before they are published in database.

It is very informative database as data is derived from literature. Each record contains allergen name

- Common name
- Biological function

- Link to primary sequence information
- Link to PDB structure
- Sequence motifs
- Source of allergen
- Route of exposure
- Allergen isoforms
- Prevalence of allergy
- References
- Molecular weight
- Sequence homologues
- Post translational modifications
- Test of allergenicity
- Recombinant forms
- Literature References

It has user friendly search facilities. A quick search using keywords enable user to display result in several ways. Advance search allows users to search using Boolean modifiers. Allergome provides list of allergens sorted by categories. Download facilities are lacking. Large amount of data cannot be exploited for bioinformatics analysis. It doesn't have any integrated bioinformatics tools.

3.4.3.5 ALLFAM

This database is based on evolutionary and structural relationships between allergens from different sources.

- A novel method of classification of allergens is proposed in which protein family databases that are linked to protein sequence databases.
- Studies have revealed that most allergens can be found in a limited number of protein families.
- AllFam is a database of allergen families.
- The data is extracted from AllFam to determine the protein family distribution of allergens and to elucidate common structural and biochemical features of allergens.
- The AllFam database is freely accessible at <http://www.meduniwien.ac.at/allergens/allfam/>.
- It can be queried for lists of allergen families filtered by source and route of exposure.
- In addition, for each family, the database contains a list of allergens and an allergen family fact sheet with information on biochemical properties and the allergologic significance of its allergenic members.
- AllFam is cross-linked with the Allergome database and regularly updated.

3.4.3.5.1 CONSTRUCTION

- Data is extracted from allergome database. (Mari et al., 2006).
- Data is categorized as inhalation, ingestion, sting/bite, contact, iatrogenic, and autoallergen.
- UniProt accession numbers were compared with Pfam. Pfam is a database of precomputed protein domain architectures.
- For entries that yielded no results, sequences were downloaded and compared with Pfam by using the hmmpfam program from the HMMER 2.3 package.
- This hmmpfam program compares a query sequence with all Pfam protein families.

Domain architectures of allergens were translated into AllFam allergen families by using the following criteria.

- For single-domain proteins, each Pfam family corresponded to an AllFam family.
- Pfam domains constituting multidomain proteins were merged into single AllFam families if the constituting domains exclusively occurred in members of a single protein family. Otherwise, each domain was treated as a separate AllFam family.

3.4.3.5.2 STRUCTURAL AND FUNCTIONAL CLASSIFICATION OF ALLERGENS

- Structures of allergens and allergen homologues were classified by using the Structural Classification of Proteins (SCOP).
- AllFam families and SCOP families were matched by using the links to SCOP embedded in the Pfam database.
- For a functional classification of allergens using standardized descriptions of biologic functions, all UniProt accession numbers of allergen sequences in AllFam were compared with the Gene Ontology (GO) Annotation Database.

3.4.3.5.3 SEQUENCE CONSERVATION WITHIN FAMILIES OF ALLERGENS

- Sequences of representative allergens from the 4 most important families of allergens were aligned by using ClustalX.
- Sequence identity matrices and neighborjoining phylogenetic trees were generated from these alignments with ClustalX and visualized with TreeView 1.6.6.

3.4.3.5.4 RESULTS

- The AllFam database contained 847 allergens with known partial or total sequences.
- 707 allergens were classified into 134 AllFam families that contained 184 different Pfam domains. Thus allergens were found in only 2% of all 9318 families in the Pfam database. The distribution of allergens was highly biased toward a few protein families.
- The protein family with the highest number of allergens, the prolamin superfamily, contained 59 allergens (8% of all allergens with known protein family) and the 10

most abundant families contained 300 allergens (42%), there were 53 families that contained only a single allergen.

- Most allergen families were confined to a single source kingdom, such as prolamins, profilins, and cupins from plants and tropomyosins, lipocalins, and caseins from animals.
- Minority of protein families contained allergens from multiple kingdoms such as, the EF-hand family and the pathogenesis-related proteins (PR-1).
- A comparison of the protein family distribution of allergens with the distribution of random UniProt entries confirmed that the number of protein families among allergens was much smaller than expected from a random sample.
- Allergens were found in all structural classes, as defined by SCOP.
- All members of protein families that contained allergens could be grouped into only 138 structural families.
- All 3012 families in the SCOP database were grouped into 1639 superfamilies and 978 folds, whereas the 138 structural families that contained allergens were grouped into 108 superfamilies and 97 folds.

	Sequences	Sequences from known protein families	AllFam families	AllFam families with >1 allergen
All Allergens	847	707	134	81
Sources				
Plants	369	338	58	34
Animals	305	268	60	36
Fungi	163	91	37	16
Bacteria	10	10	5	1
Routes of Exposure				
Inhalation	479	377	99	59
Ingestion	257	240	48	29
Sting bite	66	52	14	7
Contact	58	50	35	10
Autoallergen	14	14	14	0
Iatrogenic	11	10	7	2

Table 5-: Number of sequences and protein families of allergens in Allfam (Radauer *et al.*, 2008).

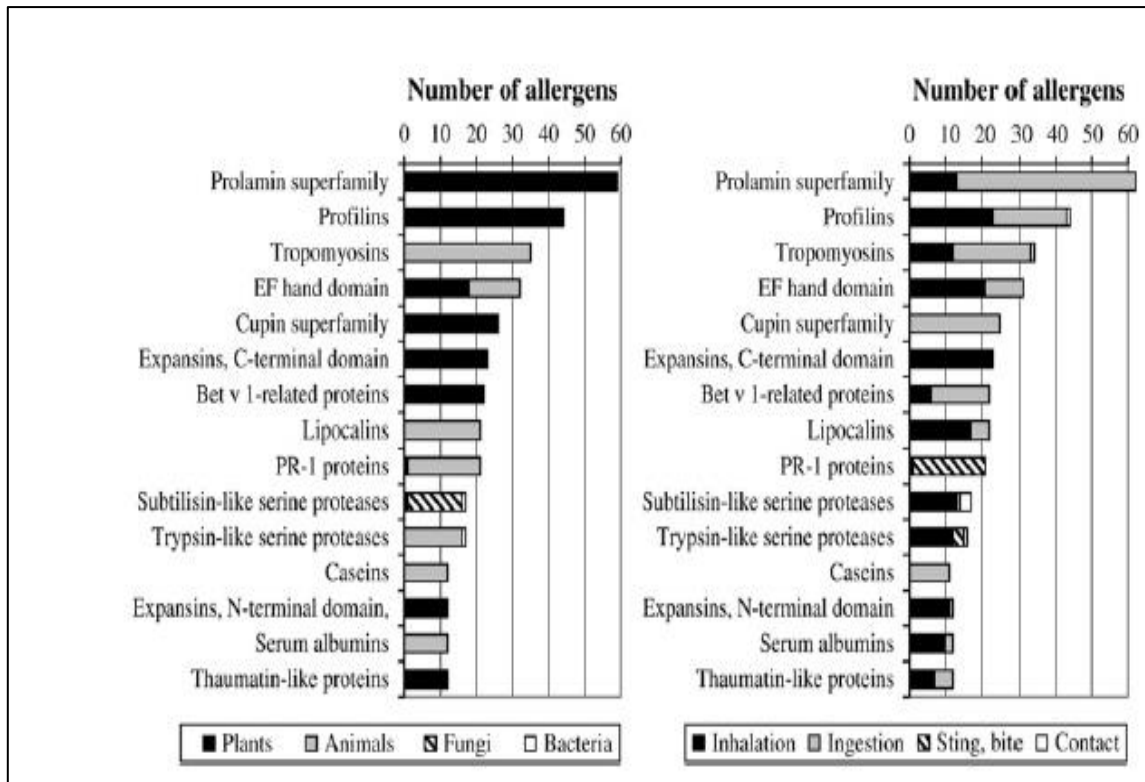


Figure 4:- The classification of allergens according to source and route of exposure (Radauer *et al.*, 2008).

SCOP class	All Structure in SCOP			Structure of allergens and Allergen homologs		
	Folds	Superfamilies	Families	Folds	Superfamilies	Families
All α Proteins	226	392	645	19 (8%)	20 (5%)	25 (4%)
All β Proteins	149	300	549	22 (15%)	24 (8%)	36 (7%)
α/β Proteins	134	221	661	14 (10%)	18 (8%)	29 (4%)
$\alpha+\beta$ Proteins	286	424	753	28 (10%)	29 (7%)	31 (4%)
Multidomain Proteins	48	48	64	2 (4%)	2 (4%)	2 (3%)
Membrane And Cell Surface Proteins	49	90	101	2 (4%)	2 (2%)	2 (2%)
Small Proteins	79	114	186	8 (10%)	9 (8%)	9 (8%)
Coiled Coil Proteins	7	50	53	2 (29%)	4 (8%)	4 (8%)
Total	978	1639	3012	97 (10%)	108 (7%)	138 (5%)

Table 6:- Classification of allergens into protein families (Radauer *et al.*, 2008).

3.5 ALLERGENICITY PREDICTION

3.5.1 AIM

The aim of bioinformatics in allergen research is allergenicity prediction. Accurate prediction will improve the allergenicity assessment of proteins. This will further reduce the cost of allergenicity testing. The impact of prediction methods is considered to be huge. Prediction method uses two criterions-:

- Precision-: it is expressed as percentage of correctly predicted allergens over all predicted allergens.
- Recall-: it is the ability of method to detect allergens in the test set. It is expressed as percentage of correctly predicted allergens over all predicted allergens.

High precision means that any predicted allergen is likely to be a true allergen. High recall means that method is able to correctly predict a large portion of allergens in a test set. A trade-off is required to get high precision and recall.

Precision= $tp/(tp + fp)$, Recall= $tp/(tp + fn)$

Tp= true positive (correctly predicted allergen)

Fp= false positive (wrong allergen that is predicted as an allergen)

Fn= false negative (an allergen predicted as non-allergen)

3.5.2 SEQUENCE SIMILARITY SEARCH METHOD

These methods are very useful in predicting allergenicity. If two proteins are highly similar and one of them is allergen then the probability of other being an allergen is very high. These methods are easy to implement. The main cause of allergenicity is binding of epitopes. BLAST and FASTA algorithms are very useful in implementing this type of prediction search. SDAP and FARRP use this method to query the content.

It is used for identifying cross reactive allergens. CR allergens are usually 70% identical. Thus, local alignment methods are useful. The performance is limited to linear epitopes. Conformational epitopes doesn't consist of continuous amino acids. Thus, this method is not useful in predicting conformational epitopes (Aalberse *et al.*, 1996).

This method depends upon the coverage if dataset against which the query is searched. The detection of novel allergens becomes difficult. Hence, requirement of comprehensive database increases (Aalberse, C. 2000).

3.5.3 SUPERVISED CLASSIFICATION APPROACHES

These methods are also used for allergenicity prediction. The supervised algorithms employed are-:

- KNN classifier
- Bayesian Linear Gaussian classifier
- Bayesian Quadratic Gaussian classifier

These methods are trained on a set of local alignments produced by FASTA. Training data includes both positive and negative datasets (Soeria *et al.*, 2004).

Results of these algorithms are-:

- Bayesian Linear Gaussian classifier-: able to detect 77% of allergens and false positive rate is 10%
- Bayesian Quadratic Gaussian classifier-: able to detect 77% of allergens and false positive rate is 11%.
- KNN classifier-: able to detect 78% of allergens and false positive rate is 13%.

The algorithms can be tuned for high precision and recall. By integrating feature methods obtained using different scoring matrices better results are obtained for Bayesian Linear Gaussian classifier (able to detect 77% of allergens and false positive rate is 8%). This method relies on local alignments so prediction of conformational epitopes is again a challenge.

3.5.4 EXPECTATION MAXIMIZATION

MEME, a motif discovery system is employed for allergenicity prediction. It employs expected minimization technique (Bailey *et al.*, 1994). The aim is to find common motifs among allergens and then predict allergenicity. These motifs act as indicators of allergenicity.

- MEME is employed in an iterative manner.
- A dataset of 779 non-redundant allergens is created from public databases.
- MEME is applied to this database.
- Most significant motif is extracted and converted into a profile.
- This profile is used to search the database for existing allergens, which are then removed from the database.
- The remaining allergens are submitted to next round of motif discovery and removal.

52 motifs were discovered and 644 allergens in the dataset contain one or more of 52 motifs. 135 allergens didn't yield motifs because of incomplete sequence information. The 52 motifs can be used to obtain a significant match with any novel protein. An e-value of 10^{-8} is used as an indicator of allergenicity. The results of this methods are-:

- On a synthetic dataset method shows recall=100% and precision=95.5%.
- In practical scenario method shows recall=100% and precision=8.6%.

A high recall prevents the slipping of any potential allergen and high precision reduces the number of false positives.

3.5.5 WAVELET TRANSFORM

This method is based on extraction of motifs from allergens and allergenicity prediction. It converts the aligned amino acid sequences into signals (Li *et al.*, 2004). The conserved motifs are detected on different scales.

- 664 allergens are collected from IUIS list, BIFS and FARRP.
- Allergens are clustered into groups.

- Clustering is done by computing the distance between every pair of allergens. Clustal W is used for this purpose.
- ‘Partitioning around medoids’ method is used to cluster allergens into groups.
- ClustalW is used to generate multiple aligned amino acids in each group.
- Conserved motifs are extracted from multiple sequence alignments using wavelet transform approach.
- HMMER package is used to create HMM profiles from these motifs.
- These profiles are used to predict the allergenicity of novel proteins.
- 20% allergens in database didn’t contain any motifs and are stored separately for BLAST search.
- The novel protein is searched using hmmpfam against all discovered motifs. If any motif is found, it is predicted as an allergen.
- If not, BLAST search is carried out and if similarity exists, then protein is predicted as an allergen. Otherwise, non-allergen.
- The e-value for BLAST is 0.001.

The results are as follows-:

- Results of 10 fold cross-validation indicates precision=99.77% and recall=70.61.
- Inclusion of BLAST causes an increase of 7% in precision.

The performance of this method is far better.

3.5.6 ALGPRED

The allergenicity is predicted on the basis of several approaches (Raghava *et al.*, 2006) -:

- First approach, a standard method has been developed for predicting allergens based on amino acid and dipeptide composition of proteins using support vector machine (SVM).
- Second approach, motif-based technique has been used for predicting allergens using the software MEME/MAST.
- Third approach, a protein is assigned as an allergen, if it has a segment similar to allergen representative proteins (ARPs).
- Fourth approach, a protein is assigned allergen if it have segment identical to known IgE epitopes.

3.5.6.1 DATASET USED-:

The dataset used in this study were obtained from

- http://www.slv.se/templatesSLV/SLV_Page_9343.asp (Bjorklund *et al.*, 2005), which contains 578 allergens and 700 non-allergens.
- The epitopes were obtained from various sources that include 56 IgE epitopes from Bcipep database and 157 IgE epitopes from SDAP database.
- 178 epitopes after removing redundant epitopes and epitopes having less than five amino acids. These IgE epitopes were scanned against dataset of allergic and non-allergic proteins.

3.5.6.2 EVALUATION

The performance of methods has been evaluated on a blind or independent dataset obtained from Li et al, 2001. The dataset have 664 allergens where allergens obtained from various sources that include-:

- 238 allergens from International union of immunological societies (IUIS)
- 270 from Swiss-Prot's Allergen Index,
- 1171 from the biotechnology information for food safety database (BIFS)
- 752 from food allergy research and resource program (FARRP).

In this dataset no two sequence have identity >95%.

3.5.6.3 ARPs COLLECTION

The dataset of ARPs consists of 2890 ARPs (24 amino acid peptides) obtained from Bjorklund et al., 2005.

- High-quality repositories of amino acid sequences of proteinaceous allergens (allergen database) and non-allergens are collected.
- It was based on the global similarity scores of each allergen peptide, a set containing 2890 ARPs was created which had high similarity in allergenic proteins but not in non-allergenic proteins.

3.5.6.4 MEME/MAST

MEME/MAST is a tool for discovering motifs in a group of related protein sequences (Timothy et al., 1994).

- A motif is a sequence pattern that occurs repeatedly in a group of related protein sequences. MEME represents motifs as position-dependent letter probability matrices.
- Matrix describes the probability of each possible letter at each position in the pattern.
- MEME takes as input a group of protein sequences (the training set) and output is many motifs.
- MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences and description for each motif.

MAST (Motif Alignment and Search Tool) is a tool for searching biological sequence databases for sequences that contain one or more of a group of known motifs.

- MAST takes as input a file containing the descriptions of one or more motifs and searches sequence databases that have been created that match the motifs.

3.5.6.5 SUPPORT VECTOR MACHINE

The SVM has been implemented using SVM_light (Joachims,T. 1995) which allow users to select various parameters and various kernel functions like radial basis function (RBF), polynomial.

3.5.6.6 PROTEIN FEATURES

- **Amino acid composition-**: Amino acid composition is the fraction of each amino acid in a protein. The fraction of all 20 natural amino acids was calculated using the following equation-:

$$\text{Amino acid composition} = \frac{\text{Total number of amino acid}}{\text{total number of amino acids in protein}}$$

- **Dipeptide composition-**: It is used to represent global information about each protein sequence. It gives a fixed pattern length of 400 (20*20). The fraction of each dipeptide was calculated using following equation:

$$\text{Dipeptide composition} =$$

$$\frac{\text{Total number of dipeptide}}{\text{total number of all possible dipeptides}}$$

3.5.6.7 RESULTS

3.5.6.7.1 SVM BASED METHOD

Accuracy of 85% is achieved using this approach. This method correctly predicted 95% of allergens at specificity of 61%. It also correctly predicted 34% of allergens at specificity around 98.

3.5.6.7.2 MOTIF BASED PREDICTION

- Sensitivity of this method ranges from 7% to 94%.
- Specificity of this method ranges from 2.85 to 66.86%.

This method has low sensitivity.

3.5.6.7.3 PREDICTION USING ARPS

- The sensitivity ranges from 52.71 to 94.28%.
- Specificity of this method is 83.58%.

3.5.6.7.4 HYBRID APPROACH

- The sensitivity ranges from 33.74–44.52%.
- Specificity of this method is 89.28%.

3.6 EPITOPE

Epitope is also known as an antigenic determinant. It is the part of antigen that is recognized by the immune system. Majorly they are recognized by antibodies, B cells and T cells. Paratope is the part of antibody that recognizes the epitope.

Epitopes are of two types:-

- Linear Epitopes
- Conformational Epitopes

3.6.1 LINEAR EPITOPES

In this type of epitope, the linear sequence of amino acid is recognized by the antibody. Majority of antibodies recognize conformational epitopes that has specific 3-D structure.

Proteins are composed of amino acids. Primary structure of protein is defined as the linear sequence of amino acids. The antigen is broken down in lysosome and it yields small peptides that are linear in nature and are recognized by antibodies. They are called as linear epitopes.

In laboratory, while performing western blot analysis, the protein is treated with beta-mercaptoethanol, and run in SDS-PAGE for the Western blot. The protein is unable to regain its native state. Thus, antibody directed against these small peptides will only recognize linear epitopes.

3.6.2 CONFORMATIONAL EPITOPES

The antigen comes directly in contact with the receptor. The residues of intact antigen that makes contact with the receptor are termed as conformational epitopes.

- Proteins exist in the form of folded helices and sheets which are connected by loops, turns or coils.
- A conformational epitope is a sequence of subunits (usually, amino acids) composing an antigen that come in direct contact with a receptor of the immune system.
- Whenever a receptor interacts with an undigested antigen, the surface amino acids that come in contact may not be continuous.
- Such discontinuous amino acids that come together in three dimensional conformations and interact with the receptor's paratope are called conformational epitopes.

1-50	NRSLILVLC FLPLAALGKV FGRCEIaAAM KRHGIDNYRG YSLgHwCaa
51-100	KFESNFNTQA TNRNTDGS TD YgIIQInSRN McNDGRTPGS RNL CNIPCSA
101-150	ILSSDITASv nCaKKIVSDG NGMIAWVAWR NRCKGTDVQA WIRGCRIL
Predicted result format: EPITOPE RESIDUE NON-EPITOPE RESIDUE core residue A	

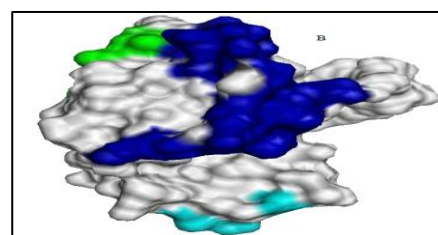


Figure 5:- Linear epitopes (A) and conformational epitopes (B).

3.7 IgE EPITOPE PREDICTIONS AND MAPPING

3.7.1 SPADE (Dall'Antonia *et al.*, 2011)

The tool takes into account 2 considerations (Furmonaviciene *et al.*, 2004)-:

- It is based on structure based methods instead of sequence based methods.
- It makes use of cross-reactive (CR) data.

This tool present a novel approach that is independent from training data approaches.

CR of IgE depends upon sequential and structural conservation among homologous allergen proteins (Kaneta *et al.*, 2002). The basic aim of this method is to compare protein surfaces on the basis of structural similarity and correlate it with the concept of cross-reactivity (Vieths *et al.*, 2001). In addition to conserved residues, the geometric and physicochemical properties of superimposed surfaces are analysed at atomic level. Thus, this approach is related to macromolecular surface comparison and Lawrence and Colman's vector-based method for the determination of shape complementarity (Binkowski *et al.*, 2004).

3.7.1.1 EPITOPE PREDICTION STEPS

3.7.1.1.1 PROTEIN COMPARISON

- A pair of structurally related proteins was chosen. Their coordinates are obtained from PDB.
- Pairwise structural protein alignments were carried out with the program MultiProt (Shatsky *et al.*, 2004) by using a block root mean SD (RMSD) limit of 3.0 Å. Based on the Calcium atom coordinates of the aligned amino acid residue subsets, the models were superimposed by using the Kabsch algorithm (Kabsch, W. 1997).
- Amino acid side chains were standardized by using an existing rotamer library (Lovell SC *et al.*, 2000) to reduce crystal packing artifacts.
- The calculation of electrostatic potentials was performed with the programs PDB2PQR (Baker *et al.*, 2001) by using the amber force field parameters and Adaptive Poisson-Boltzmann Solver (APBS) (Sanner *et al.*, 1996) by using a nonlinear Poisson-Boltzmann equation.
- Triangulated solvent-accessible molecular surfaces and solvent-excluded molecular surfaces (SEs) were calculated with the program Maximal Speed Molecular Surface (MSMS) by using a 1.5 Å probe radius.
- Hydrophobicity and hydrogen bond capacity were analysed by using reference tables and then mapped onto the surfaces.
- Novel computational algorithms were developed for the quantitative pairwise surface comparison involving multiple local surface superposition and geometric SES triangle matching, subsequently analysing the agreement of physicochemical properties

3.7.1.1.2 EPITOPE PREDICTION

- The allergen on which the epitopes have to be mapped is called as reference. This is done on the basis of surface similarity as obtained from previous pairwise comparison module.
- The compared proteins are first divide into highly cross-reactive and weakly cross-reactive proteins.
- Similarity difference (D-sim) values are then calculated for every residue by summation of surface similarity scores from the highly cross-reactive allergens and subtraction of scores from weakly cross-reactive allergens. This step is weighted by CR.
- The resulting D-sim values were mapped onto the reference allergen surface
- Residues selected by values above a filter threshold are clustered by spatial proximity to obtain contiguous surface regions termed as patches.
- Patches were accepted as likely epitopes if their solvent-accessible molecular surface area values exceeded a size threshold of 400\AA^2 .

3.7.1.2 ALGORITHM IMPLEMENTATION

The program language C was used to create the source code for

- Structural superposition
- Side chain standardization
- Surface comparison, including feature mapping
- Similarity filtering
- Cluster recognition

The graphic user interface was developed by using the Tool Command Language with the graphic toolkit. For all spatial surface feature and similarity score visualizations, the program PyMol is used.

Epitopes are predicted for Bet v 1 allergen.

Reference	Compared Protein	Sequence Identity (%)	Surface Similarity (%)
Bet v 1a (1BV1)	Api g 1 (2BK0)	41	47.7
	Bet v 1d (3K78)	95	56.7
	Bet v 11 (1FM4)	92	58.4
	LLPR10.1A (1ICX)	42	37.8
	LLPR10.1B (1IFV)	43	38.5
	Pru av 1 (1E09)	56	45.5
	VRCSBP (2FLH)	30	34.5

Table 7-: Bet v 1 as reference protein and proteins cross-reactive to reference protein. The results are obtained after running comparison module.

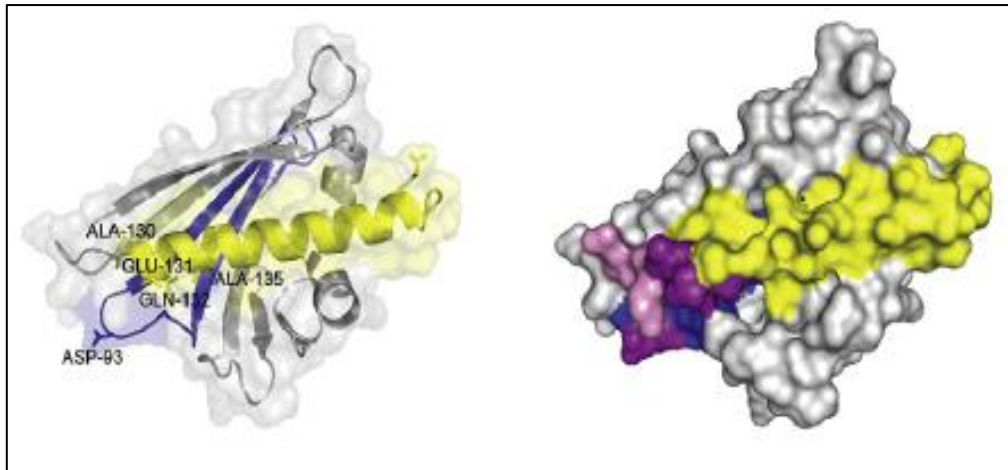


Figure 6-: The epitopes that are predicted and mapped on the surface of Bet v 1. The epitopes are labelled in different colors.

3.7.2 ELLIPRO (Ponomarenko *et al.*, 2008)

Ellipro is a web-tool that implements a modified version of Thornton's method. Thornton proposed a method for identifying continuous epitopes in the protein regions protruding from the protein's globular surface. Regions with high protrusion index values were shown to correspond to the experimentally determined continuous epitopes in myoglobin, lysozyme and myohaemerythrin (Thornton *et al.*, 1986). ElliPro is available at <http://tools.immuneepitope.org/tools/ElliPro>.

- It is integrated with a residue clustering program, MODELLER program and Jmol viewer.
- ElliPro has been tested on a dataset of epitopes obtained from 3D structures of antibody-protein complexes (Ponomarenko *et al.*, 2007) and compared with six structure-based methods.

The methods are-:

- CEP (Kulkarni-Kal *et al.*, 2005) and DiscoTope (Haste Andersen *et al.*, 2005)-: epitope prediction methods.
- DOT (Mandell *et al.*, 2001) and PatchDock (Schneidman-Duhovny *et al.*, 2003)-: protein-protein docking methods.
- PPIRED, PPIRED and ProMate-: structure-based methods for protein-protein binding site prediction.

3.7.2.1 INPUT

ElliPro accepts two types of input data-:

- Protein sequence
- Protein structure

In the first case-:

- The user may input either a protein SwissProt/UniProt ID or a sequence in either FASTA format or single letter codes.
- User can select threshold values for BLAST e-value and the number of structural templates from PDB that will be used to model a 3D structure of the submitted sequence.

In the second case-:

- The user may input either a four-character PDB ID or submit his own PDB file in PDB format.
- If the submitted structure consists of more than one protein chain, ElliPro will ask the user to select the chain(s) upon which to base the calculation.
- The user can change threshold values on the parameters used by ElliPro for epitope prediction.
- The minimum residue score (protrusion index), denoted here as S , between 0.5 and 1.0 and the maximum distance, denoted as R , in the range 4 – 8Å.

3.7.2.2 STRUCTURE MODELLING

- If a protein sequence is used as input, ElliPro searches for the protein or its homologues in PDB, using a BLAST search.
- If a protein cannot be found in PDB that matches the BLAST criteria, MODELLER is run to predict the protein 3D structure.
- The user may change the threshold values for BLAST e-value and a number of templates that MODELLER uses as an input.

3.7.2.3 ELLIPRO METHOD

ElliPro implements three algorithms performing the following tasks:

- Approximation of the protein shape as an ellipsoid
- Calculation of the residue protrusion index (PI)
- Clustering of neighboring residues based on their PI values.

The protein surface is considered as an ellipsoid, which can vary in sizes to include different percentages of the protein atoms.

- If 90% of ellipsoid includes 90% of the protein atoms then for each residue, a protrusion index (PI) is computed.
- PI is defined as percentage of the protein atoms enclosed in the ellipsoid at which the residue first becomes lying outside the ellipsoid. For example, all the residues that are outside 90% ellipsoid will have $PI = 9$ (or 0.9 in ElliPro).

In implementing the first two algorithms, ElliPro differs from Thornton's method by considering each residue's centre of mass rather than its $C\alpha$ atom.

The third algorithm for clustering residues defines a discontinuous epitope based on the threshold values for the protrusion index S and the distance R between each residue's centers of mass. All protein residues with a PI values greater than S are considered when calculating discontinuous epitopes.

Clustering separate residues into discontinuous epitopes involves three steps that are recursively repeated until distinct clusters with no overlapping residues are formed.

- First, primary clusters are formed from single residues and their neighbouring residues within the distance R .
- Second, secondary clusters are formed from primary clusters where at least three centers of mass are within the distance R from each other.
- Third, tertiary clusters are formed from secondary clusters which contain common residues. These tertiary clusters of residues represent distinct discontinuous epitopes predicted in the protein.
- The score for each epitope is defined as a PI value averaged over epitope residues.

In comparison with six other structure-based methods that can be used for epitope prediction, ElliPro performed the best, AUC value of 0.732, when the most significant prediction was considered for each protein.

3.7.3 SEPPA-: SPATIAL EPITOPE PREDICTION OF PROTEIN ANTIGENS (Sun *et al.*, 2009)

SEPPA means spatial epitope prediction of protein antigens.

- In this method a novel concept of 'unit patch of residue triangle' has been introduced.
- It defines local spatial region in protein antigen surface.
- Spatial clustering coefficient parameter is also integrated to represent 3D characteristic of epitopes.
- A comprehensive training dataset is retrieved from PDB.
- SEPPA is trained by 82 antigen–antibody protein complexes, which contained 84 unique epitopes.
- One hundred and nineteen independent spatial epitopes of protein antigens were collected as testing dataset.

3.7.3.1 DATASET

- Antigen–antibody complexes were extracted from PDB database.
- Only those with resolution better than 3.0Å and protein antigen length with more than 25 residues were retained.
- Redundant epitopes were removed by 60% similarity.
- Eighty two structures were finally retained as the training data which included 84 unique epitopes.

3.7.3.2 ALGORITHM PARAMETERS

3.7.3.2.1 UNIT PATCH OF RESIDUE TRIANGLE

- Solvent accessible surface areas (SASA) were calculated for each residue in antigen proteins.
- Surface residues were those with more than 1\AA^2 SASA.
- The residues with SASA having more than 1\AA^2 area are classified as epitope residues.
- For any three surface residues, if the distance for every two of them was within 4\AA atom distance then they are called as unit patch of residue triangle.

3.7.3.2.2 PROPENSITY INDICES

- It is assumed that the residues have similar functional moieties of R-groups in antigen-antibody interaction.
- The 20 residues were divided into 13 functional subgroups according to the conformational epitope research (Erez et al., 2007).
- Four hundred and fifty five combination patterns of subgroups were observed out of $13*13*13$ unit patches.
- Propensity index of the unit patch pattern i is calculated as the ratio of the number of pattern among all epitope unit patches compared with that ratio in the non-epitope unit patches.
- For a certain surface residue r , the propensity score of it is predominantly determined by its local neighboring environment. Thus (avg r) is calculated as the averaged propensity indices of all possible unit patches around residue.

3.7.3.2.3 RESIDUE NEIGHBOR AND CLUSTERING COEFFICIENT

Clustering coefficient is introduced to describe the compactness of the neighbouring residues around one residue.

- It reflects the probability that the neighbours of residue (r) are also neighbours with each other (Huang *et al.*, 2007).
- For one residue r , all residues within 15\AA of (r) are defined as residue neighbours of r .
- k_r is the total number of residues neighbours for r .

3.7.3.2.4 ALGORITHM IMPLEMENTATION

- Determine all the surface residues in the protein antigen.
- All possible unit patches within 15\AA atom distance of residue (r) are searched.
- Pre-calculated propensity indices are mapped on unit patches.
- Clustering coefficient for residue (r) is calculated.
- Antigenicity score for each residue is displayed.
- Residues with scores higher than a threshold are highlighted.

- Visualize the subsets of predicted epitope area graphically.

3.7.3.3 INPUT

- SEPPA requires a 3D protein structure in PDB format as input.
- Users can submit the query with a released PDB ID or upload a file in PDB format.
- Chain ID does have to be specified.

3.7.3.4 OUTPUT

- The results of prediction are displayed in html format.
- The sequence of submitted protein antigen is displayed in single letter code in result window.
- The core residues are shown in lowercase and surface residues in uppercase.
- The residues predicted as epitope are highlighted with yellow colour background.
- The scores of prediction are recorded in another file, which lists the antigenicity scores for individual residue and this file is downloadable.
- A link to visualize the prediction result is also provided in the result page.
- The visualization of result is displayed with Jmol.
- Tints from blue to red represent a rising propensity for a residue to be in the epitope.

3.7.3.5 PERFORMANCE

- SEPPA achieved the average AUC value of 0.742 on the 119 independent testing dataset.
- A sensitivity of 0.580 and a specificity of 0.707, with threshold of 1.80 were obtained on this testing dataset.

3.7.4 EPITOPIA (Rubinstein *et al.*, 2006)

Epitopia server implements a machine-learning based algorithm to predict immunogenic regions either on 3D structure or the sequence of a given protein.

- The algorithm determines the immunogenic potential at a resolution of single amino acid.
- Epitopia computes an immunogenicity score for each solvent accessible residue if a 3D structure was provided as input or a score for every amino acid if a sequence input was provided.
- A powerful visualization tool is integrated that color-codes the immunogenicity scores on either the protein sequence or the 3D structure.

3.7.4.1 DATASET

The classifier was trained to recognize immunogenic properties using a dataset of 66 non-redundant validated epitopes derived from antibody-antigen co-crystal structures (Ponomarenko *et al.*, 2006) and 194 non-redundant validated epitopes derived from antigen sequences.

3.7.4.2 ALGORITHM IMPLEMENTATION

The Epitopia algorithm (Rubinstein *et al.*, 2009) uses a Naïve Bayes classifier to predict the immunogenic potential of protein regions.

- An antigen is divided into overlapping surface patches in case of 3-D structure and stretches in case of a linear sequence input.
- Epitopia computes for each patch or stretch the probability that it was drawn from the population of epitopes on which the classifier has been trained.
- The classifier is trained on the basis of physico-chemical and structural-geometrical properties of patches or stretches.
- The immunogenicity score is the sum of logs of these probabilities and is assigned to the central residue of the patch or to the middle residue in the linear stretch.
- The immunogenicity score reflects the immunogenic potential of a certain residue relative to all residues in the antigen.

3.7.4.3 PROBABILISTIC SCORE

- Site-specific immunogenicity scores in the training data are divided into quantiles.
- For each quantile, the fraction of validated epitope residues out of the total number of residues in the quantile is computed.
- This number approximates the probability that a residue with a given immunogenicity score that falls in this quantile is an epitope residue.

3.7.4.4 EPITOPIA INPUT

- For a protein 3D structure input, Epitopia requires a protein data bank file. File can also be uploaded by user in PDB format. It is compulsory for the structural file to have seqres portion.
- Chains have to be specified. If all of the chains in the model have to be selected then the term "all" should be specified and if only a subset of chains in the model should be related to, the corresponding chain identifiers should be specified.
- For a protein sequence input, the amino-acid sequence may either be pasted or a local sequence file can be uploaded. In either case, the sequence should be in Fasta format and should contain only standard amino acids.

3.7.4.5 EPITOPIA OUTPUT

3.7.4.5.1 STRUCTURE OUTPUT

- The immunogenicity and corresponding probability scores are computed by Epitopia for each surface residue for a 3D structure input or for every amino-acid for a sequence input. These scores are given as a text file link.
- The immunogenicity scores are color-coded and projected onto the protein.
- The visualization tool that is used for the 3D structure case is Jmol.
- Epitopia provides a RasMol command script for viewing the results locally with the RasMol program.

3.7.4.5.2 SEQUENCE OUTPUT

- The clustering procedure divides the sequence to stretches and assigns each stretch a corresponding p-value, which is defined as the probability of randomly obtaining an equally-sized stretch with such a score or higher.
- The score of a stretch is the sum of immunogenicity scores of the amino acids comprising it.
- The p-value is computed by shuffling all the scores in the sequence and repeating the search procedure a large number of times.
- The clusters are ranked according to their statistical significance are given as a text file link.

4.METHODOLOGY

4.1 RETRIEVAL OF EGG PROTEOME

There are no specific databases for hen egg proteome. The sources for retrieving information about proteins in egg are primary databases and literature (Guérin-Dubiard et al., 2006). The databases used for this purpose are:-

- Uniprot-KB
- NCBI protein database
- PIR (PROTEIN INFORMATION RESOURCE)

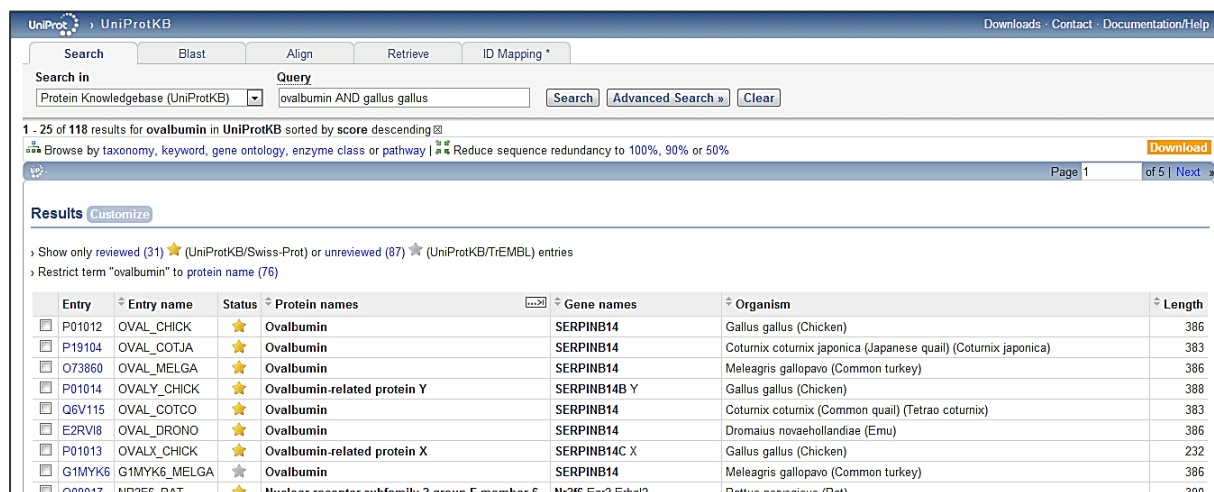
4.1.1 Uniprot-KB

The UniProt Knowledgebase aims at collecting functional information on proteins. It provides accurate, consistent and rich annotation data (Magrane, M. 2011). It provides large amount of information about proteins such as:-

- Amino acid sequence
- Protein name or description
- Taxonomic data
- Citation information
- Biological ontologies
- Classifications
- Cross-references

It provides highly annotated and high quality data. The UniProt Knowledgebase consists of two sections:-

- A section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, named as "UniProtKB/Swiss-Prot".
- A section with computationally analyzed records that await full manual annotation, named as "UniProtKB/TrEMBL".



The screenshot shows the UniProtKB search results page for the query 'ovalbumin AND gallus gallus'. The results are sorted by score descending and show 118 results. The table below displays the first 10 results.

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P01012	OVAL_CHICK	★	Ovalbumin	SERPINB14	Gallus gallus (Chicken)	386
P19104	OVAL_COTJA	★	Ovalbumin	SERPINB14	Coturnix coturnix japonica (Japanese quail) (Coturnix japonica)	383
O73860	OVAL_MELGA	★	Ovalbumin	SERPINB14	Meleagris gallopavo (Common turkey)	386
P01014	OVALY_CHICK	★	Ovalbumin-related protein Y	SERPINB14B Y	Gallus gallus (Chicken)	388
Q6V115	OVAL_COTCO	★	Ovalbumin	SERPINB14	Coturnix coturnix (Common quail) (Tetrao coturnix)	383
E2RVI8	OVAL_DRONO	★	Ovalbumin	SERPINB14	Dromaius novaehollandiae (Emu)	386
P01013	OVALX_CHICK	★	Ovalbumin-related protein X	SERPINB14C X	Gallus gallus (Chicken)	232
G1MYK6	G1MYK6_MELGA	★	Ovalbumin	SERPINB14	Meleagris gallopavo (Common turkey)	386
Q09017	NR2F6_RAT	★	Nuclear receptor subfamily 2 group F member 6	Nr2f6, Ear2, Frbal2	Rattus norvegicus (Rat)	390

Figure 7-: Screenshot of Uniprot-KB with Ovalbumin as query sequence (Magrane, M. 2011).

4.1.2 PROTEIN INFORMATION RESOURCE (PIR)

The Protein Information Resource (PIR) is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies (Wu *et al.*, 2003).

- PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information.
- NBRF compiled the first comprehensive collection of macromolecular sequences in the Atlas of Protein Sequence and Structure, published from 1965-1978 under the editorship of Margaret O. Dayhoff.

For over four decades the Protein Information Resource (PIR) has provided databases and protein sequence analysis tools to the scientific community. PIR major activities include:

- UniProt (Universal Protein Resource) development.
- iProClass protein data integration and ID mapping.
- PRO protein ontology.
- iProLINK protein literature mining and ontology development.



Figure 8-: Screenshot of PIR studies (Wu *et al.*, 2003).

4.1.3 NCBI-PROTEIN DATABASE

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health.

- The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper.
- It is a collection of 44 databases.

The Protein database is a collection of sequences from several sources such as-:

- GenBank

- RefSeq
- TPA
- SwissProt
- PIR
- PRF
- PDB

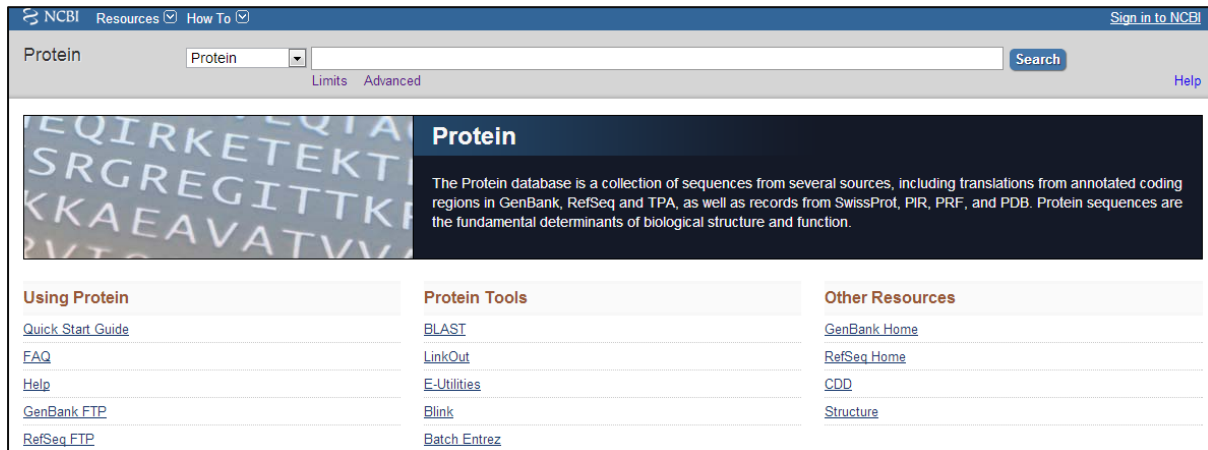


Figure 9-: Representing screenshot of NCBI protein database.

4.1.4 LITERATURE

Proteome analysis is done by various strategies such as SDS-PAGE, 2D SDS-PAGE, LC-MS etc. The results of such analysis are submitted in various journals. These journals are specific for data related to proteome research. Some of these journals are:-

- Proteome Science (<http://www.proteomesci.com/>).
- Journal of Agriculture and Food Chemistry (<http://pubs.acs.org/journal/jafcau>).
- Journal of Proteome research (<http://pubs.acs.org/journal/jprobs>).

Text based search can be performed in these journals. The papers related to egg proteome can be downloaded. The valuable information is extracted from them. In this way, data is made more comprehensible and reliable.

4.1.5 RETRIEVAL OF SEQUENCES IN FASTA FORMAT

The sequence of proteins in egg proteome can be downloaded from aforementioned databases in FASTA format. Uniprot-KB offers a retrieval tool for retrieving data in bulk. The list of proteins in form of uniprot identifiers can be pasted. The results can easily be downloaded in FASTA format.

4.2 PREDICTION OF ALLERGENICITY

AlgPred is used for prediction of allergenicity (Raghava *et al.*, 2006).

- The protein sequences are submitted to Algpred in plain format.
- A perl script can be designed for bulk submission of sequences and retrieval of results.
- Support Vector modules based on amino acid composition and dipeptide composition are used to predict allergenicity.
- The sensitivity and specificity of these two modules is high and reliable.
- It classifies the proteins into potential allergens, allergens and non-allergens.
- Potential allergens have very high probability of being an allergen and their score is also very high.
- The allergens and non-allergens have very low score of being an allergen. Thus, such proteins are not considered as predicted allergens.
- Only, potential allergens will be used in further steps.

Figure 10:- The screen shot of AlgPred home page.

Figure 11:- The screen shot of submission page with a protein sequence in submission window.

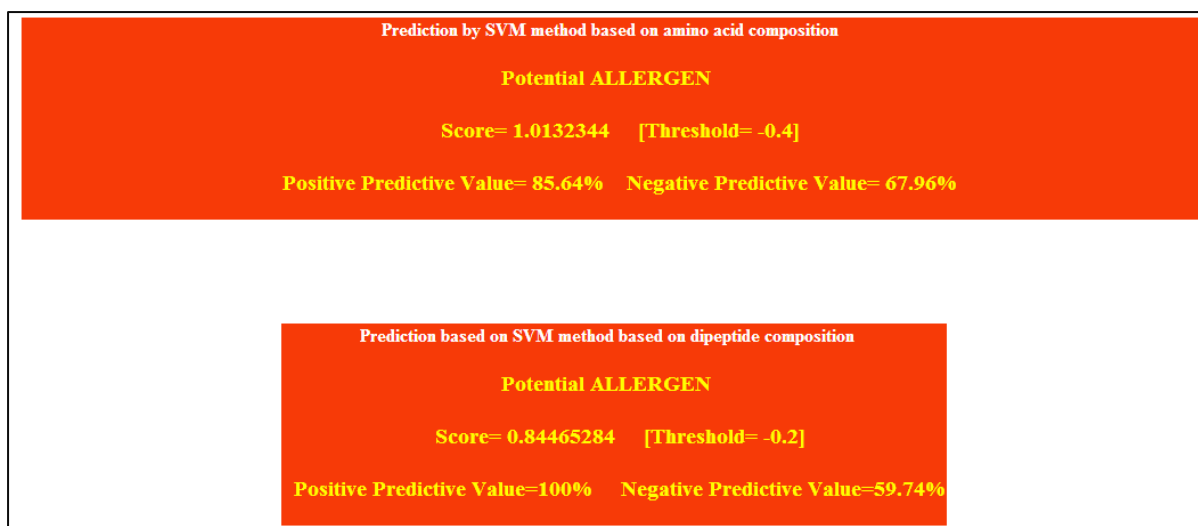


Figure 12:- The result page after submitting the protein sequence. The protein is predicted as potential allergen.

4.3 PROTEIN STRUCTURE PREDICTION OF POTENTIAL ALLERGENS

Text based search is performed in Protein Data Bank to retrieve structural file for each Potential allergen protein. Uniprot ID mapping tool can also be used for this purpose. It has the facility to map Uniprot ID's with available structural files in Protein Data Bank.

The structures are predicted for those allergen proteins that lack crystal structural data in PDB. Two servers are mainly used for this purpose:-

- I-TASSER
- Phyre2



Figure 13:- Uniprot ID mapping tool. Ovalbumin Uniprot ID is mapped with structures in PDB.

4.3.1 I-TASSER

It is a protein structure prediction tool which is available at (Zhang *et al.*, 2010) <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>. It is free for academic purposes but requires an institutional email ID for sequence submission. It predicts the structure on the basis of Homology modelling, threading and *ab-initio* approaches.

- First, it searches for structures that are homologous to query sequence and predicts model on the basis of homology modelling.
- Secondly, if sufficient homologous structures are not found then it uses fold recognition method called as threading to predict the structure.
- Thirdly, if above 2 approaches fails, then it uses *ab-initio* approach to predict the structure.

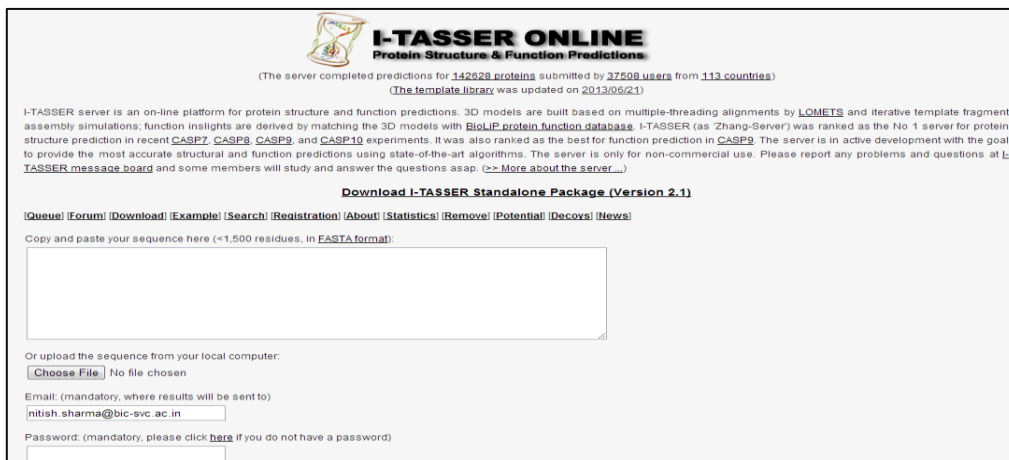


Figure 14:- The screenshot of home page of I-TASSER.

4.3.2 PHYRE-2

It predicts the protein structure by threading. Phyre2 uses a fold library that is updated weekly as new structures are solved (Kelley *et al.*, 2009). It allows user to select own templates and modelling is done on the basis of that template. It is also helpful in predicting topology of transmembrane proteins.



Figure 15:- The screenshot of home page of PHYRE-2.

4.4 EPITOPE PREDICTION AND MAPPING

Four different tools are used to predict IgE epitopes on the surface of allergenic proteins. These are SPADE, ELLIPRO, EPITOPIA and SEPPA.

4.4.1 ELLIPRO (Ponomarenko *et al.*, 2008)

The query can be submitted either in sequence format or in structure format. As, we have modelled the structure of Potential allergen proteins therefore we will submit the query in structure format. Structure file of each protein is uploaded and submit tab is entered. The results will be displayed in few minutes. ElliPro is available at <http://tools.immuneepitope.org/tools/ElliPro>.

IEDB Analysis Resource

ElliPro Prediction | Example Data | Tutorial | Download | Contact

ElliPro: Antibody Epitope Prediction

Step 1. Input type

Choose an input type: Protein sequence (Go to step 2a) Protein structure (Go to step 2b)

Step 2a. Protein sequence

Enter a protein swiss-prot ID: (example: P02185)

Or enter a protein linear sequence in PLAIN or FASTA format:

Blast expectation value: 1 (Default is 10)

Maximum number of 3D structural template(s): 5 (Default is 5)

Step 2b. Protein structure

Enter a 4 letter code PDB ID: (example: 5LYM)

Or enter a protein structure PDB file: No file chosen

Step 3. Epitope prediction parameters

Minimum score: 0.5 (Default is 0.5)

Maximum distance (Angstrom): 6 (Default is 6)

Figure 16:- The screenshot of home page of ElliPro.

ElliPro: Antibody Epitope Prediction Results

Protein Sequence(s):

Chain	Sequence
A	1 AVTVDTICKN GQLPQRNMF KRCRMEGLVM LSEITCEERK ECKKTLQKA CGEPQGLIK 61 PDPFAQVNNYK CGCEGYTLK EDTCVLVVCG YRNGKSGEC IVEYLRERIG APCSALGRY 121 PPFEDERCKT KTGETAQLK CHTDREYVCG YSPYKRCQCR EGFTRRQRK VCL

Predicted Linear Epitope(s):

No.	Chain	Start Position	End Position	Peptide	Number of Residues	Score	3D Structure
1	A	57	70	CIENPPAQVNNYK	14	0.834	View
2	A	158	173	QCMEGFTDFEKNVCL	16	0.805	View
3	A	117	135	IGKVPNPEDEKCKTKTGET	19	0.705	View
4	A	38	49	EKNECKKTLGK	12	0.680	View
5	A	139	149	LKCNTDNEVCK	11	0.645	View
6	A	90	97	QYKNGGES	8	0.551	View

Predicted Discontinuous Epitope(s):

No.	Residues	Number of Residues	Score	3D Structure
1	A:E38, A:K39, A:N40, A:E41, A:C42, A:K43, A:K44, A:E45, A:T46, A:L47, A:G48, A:K49, A:C57, A:I58, A:E59, A:N60, A:P61, A:D62, A:P63, A:A64, A:Q65, A:V66, A:N67, A:M68, A:Y69, A:K70	26	0.763	View
2	A:T3, A:V4, A:D5, A:I5, A:L139, A:K140, A:C141, A:N142, A:T143, A:D144, A:N145, A:E146, A:V147, A:K149, A:Q158, A:C159, A:M160, A:E161, A:G162, A:P163, A:T164, A:F165, A:D166, A:K167, A:E168, A:C169, A:N170, A:V171, A:C172, A:L173	30	0.701	View
3	A:K92, A:N93, A:C94, A:G95, A:E96, A:S97, A:Q98, A:Y104, A:L105, A:E106, A:I107, A:I108, A:Q109, A:S110, A:I111, A:C115, A:A116, A:I117, A:G118, A:K119, A:V120, A:P121, A:N122, A:P123, A:E124, A:D125, A:E126, A:K127, A:K128, A:C129, A:T130, A:K131, A:T132, A:G133, A:E134, A:T135	36	0.645	View
4	A:K80, A:E81, A:D82, A:T83, A:Q90, A:Y91	6	0.508	View

ElliPro: Epitope 3D Structures

No.	Epitope Residues	Number of Residues	Score
1	A:E38, A:K39, A:N40, A:E41, A:C42, A:K43, A:K44, A:E45, A:T46, A:L47, A:G48, A:K49, A:C57, A:I58, A:E59, A:N60, A:P61, A:D62, A:P63, A:A64, A:Q65, A:V66, A:N67, A:M68, A:Y69, A:K70	26	0.763

[View](#)

[[GLU]H5:A:CB #6850] Right Click to improve your experience

Figure 17:- The result page of ElliPro. Linear and Discontinuous epitopes are predicted. Jmol is also integrated to visualize the epitopes (Ponomarenko *et al.*, 2008).

4.4.2 SEPPA (Sun *et al.*, 2009)

The query can only be submitted in the structure format. The structure file should be in .pdb format. It is essential to specify chains.

SEPPA server | Batch query | Example | Help | Contact information

Please choose one submission method: ?

1. Enter an existing PDB ID and chain(s):
PDB ID: Chain(s):

2. Or upload a local file in [PDB format](#):
* A local file without chain ID column could also be uploaded for prediction.
PDB File: No file chosen
Chain(s):

Please specify a threshold: ?
Threshold:

Figure 18-: The screenshot of home page of SEPPA.

Antigenic Prediction for 1OVA:

Chain: A
Threshold: 1.80
Number of total residues: 383
Number of predicted epitope residues: 19

[View 3D structure in Jmol](#)

1-50	GSIGAA sMEf cFDVFKELKVHHANENi fYCp i aImSaLaMvYLGa KDS TR
51-100	TQiNKVvRfD KLPFGDIEA QCGTSVNvhS SIRDILNQIT KPNDVYSFSL
101-150	aSRlYa EERY PILPEYLQCV KELYRGGLEP INFQ TAAD QARELiNSWvES
151-200	QtNGIIRNVL QPSSVDSQTA mVLvna iVFKGLWEKAFKDE DTQ AMP FRVT
201-250	EQESKE VQMm YQIGLFRvaS MASEMKi lE LPFASGTMsm Lv l LPDEVSG
251-300	LEQLESI iNF EKLTEWTSSNVMEERKi KvY lpRMKMEEKY NLTSVlMAMG
301-350	iTDvFSS SAN LSGi S AE SL KISQAVhAaH aEINEAGREVVG A EAGVDAA
351-400	SVSEEF RaDH Pfl fciKHIA TNavl ffgRc VSP

Predicted result format: **EPITOPE RESIDUE** | NON-EPITOPE RESIDUE | core residue

[Download the score file](#)

Figure 19-: The result for 1OVA (ovalbumin). Highlighted residues are predicted as antigenic.

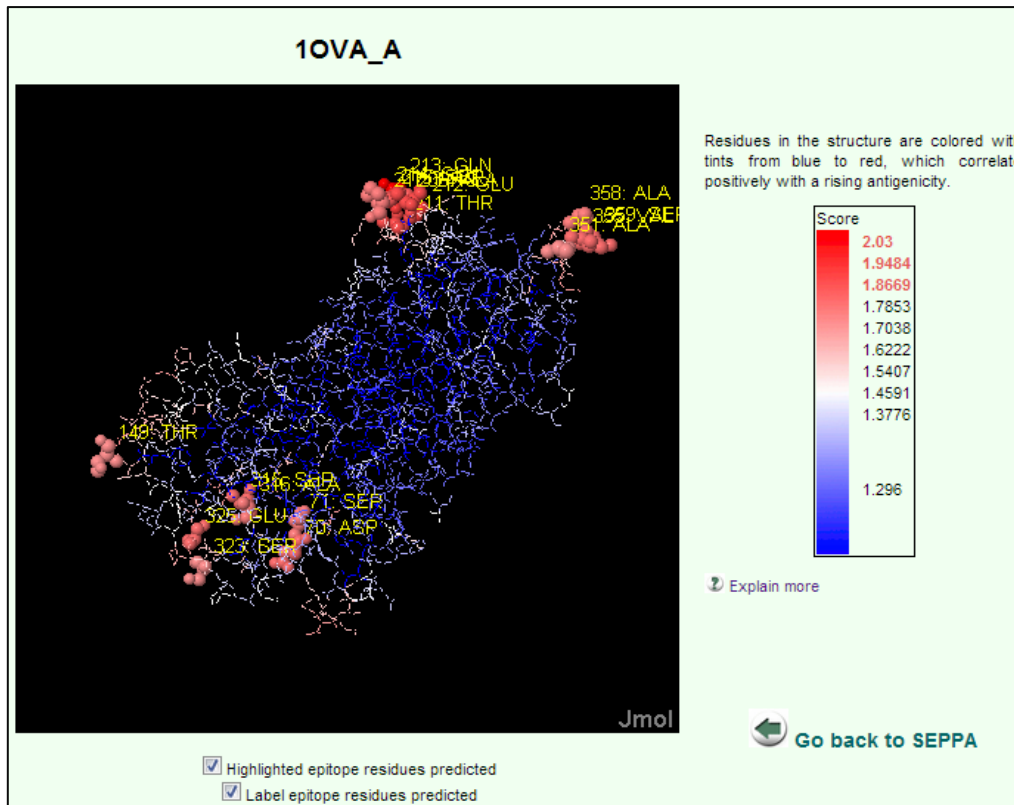


Figure 20:- The visualization predicted of antigenic residues in Jmol.

Antigenic Prediction for 10VA:

Chain: A
Threshold: 1.80
Number of total residues: 383
Number of predicted epitope residues: 19

chainID	resSeq	resName	score
A	24	GLY	1.67
A	25	SER	1.58
A	26	ILE	1.47
A	27	GLY	1.52
A	28	ALA	1.53
A	29	ALA	1.50
A	30	SER	0.00
A	31	MET	1.38
A	32	GLU	1.48
A	33	PHE	0.00
A	34	CYS	0.00
A	35	PHE	1.44
A	36	ASP	1.49
A	37	VAL	1.47
A	38	PHE	1.41
A	39	LYS	1.52
A	40	GLU	1.58
A	41	LEU	1.51
A	42	LYS	1.53
A	43	VAL	1.65
A	44	HIS	1.70

Figure 21:- The score file for each residue of 10VA.

4.4.3 EPITOPIA (Rubinstein *et al.*, 2006)

The query can either be submitted in sequence format or structure format. The structure file should have seqres part. It is essential to specify chains. It predicts both linear as well as conformational epitopes.

The screenshot shows the EpiToPIA web interface. At the top, it says "Please select one of the following options:". Below this, there are three main sections:

- If the protein structure is available:**
 - Option 1: Enter the PDB ID and the chain identifiers. Fields for "PDB ID:" and "Chains:". A checkbox for "Keep non-selected chains".
 - Option 2: Upload the PDB file and enter the chain identifiers. Field for "PDB file:" with a "Choose File" button and "No file chosen" text. Field for "Chains:". A checkbox for "Keep non-selected chains".
- If the protein structure is unavailable:**
 - Option 3: Enter the protein sequence. A large text input area.
- Option 4: Upload the sequence file. Field for "Sequence file:" with a "Choose File" button and "No file chosen" text.

At the bottom, there is a note: "An EpiToPIA run may take a while. For your convenience enter your e-mail address and the results will be sent to you." followed by an empty input field.

Figure 22:- The screenshot of home page of EpiToPIA.

4.4.3.1 STRCUTURE BASED PREDICTION

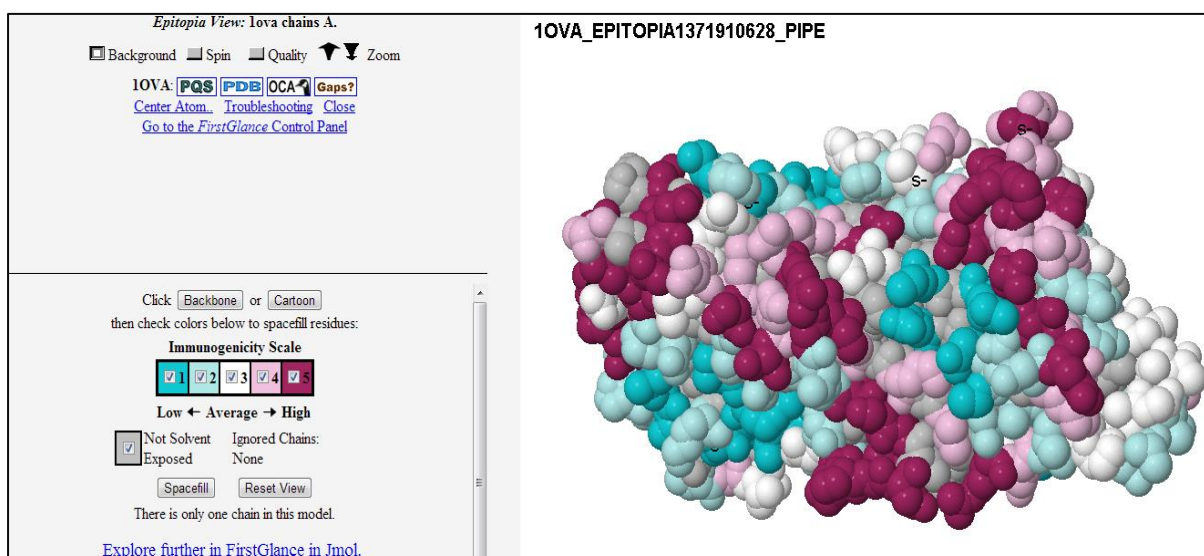


Figure 23:- The epitopes mapped on the surface of ovalbumin. The colour coding is provided according to the degree of antigenicity.

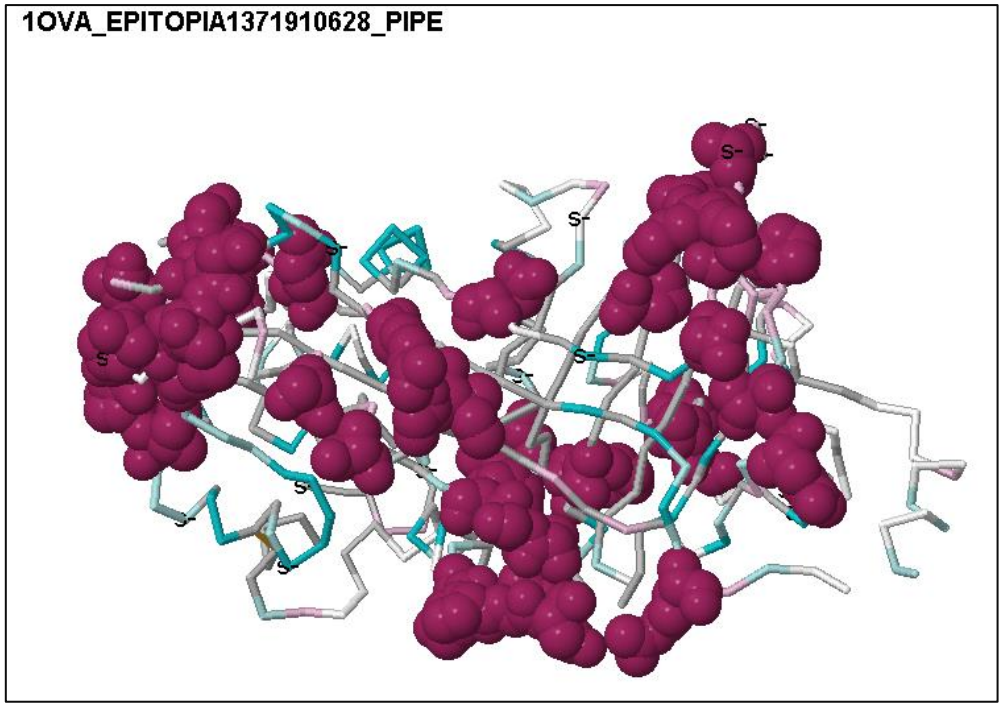


Figure 24:- Highly immunogenic residues predicted for ovalbumin.

Residue	Immunogenicity_score	Probability_score	B/E
LYS375:A	-21.295	0.188	E
ARG123:A	-22.238	0.188	E
GLN148:A	-22.288	0.188	E
TYR124:A	-22.378	0.188	E
GLY167:A	-22.675	0.188	E
ALA28:A	-22.724	0.188	E
SER323:A	-22.983	0.188	E
ASN380:A	-23.064	0.188	E
ASP179:A	-23.078	0.188	E
GLU214:A	-23.756	0.188	E
ASP108:A	-23.833	0.188	E
PRO125:A	-23.875	0.188	E
SER111:A	-23.929	0.188	E
TYR110:A	-23.985	0.188	E
ILE145:A	-24.075	0.188	E
ARG170:A	-24.223	0.188	E
GLN179B:A	-24.415	0.188	E
GLU274:A	-24.593	0.188	E
SER245A:A	-24.675	0.188	E
ASN146:A	-24.676	0.188	E
ALA324:A	-24.705	0.188	E
PRO207:A	-24.712	0.188	E
SER25:A	-24.774	0.188	E
GLU295:A	-24.969	0.188	E
LYS328:A	-25.121	0.188	E
THR180:A	-25.145	0.188	E
ILE126:A	-25.175	0.188	E
PRO217:A	-25.221	0.188	E
GLN210:A	-25.255	0.188	E

Figure 25:- Immunogenicity score and probability score for each antigenic residue in ovalbumin.

4.4.3.2 SEQUENCE BASED PREDICTION



Figure 26:- The epitopes mapped on the sequence of ovalbumin. The colour coding is provided according to the degree of antigenicity.

Residue	Immunogenicity_score	Probability_score	B/E
LEU130	-15.783	0.064	B
SER69	-15.842	0.064	E
GLY129	-15.887	0.064	E
ARG111	-15.997	0.064	E
ASP362	-16.729	0.064	E
GLU110	-16.898	0.064	E
ASP68	-17.384	0.064	E
PRO64	-17.445	0.064	E
GLY65	-17.445	0.064	E
GLY67	-17.658	0.064	E
LYS62	-17.739	0.064	E
TYR112	-18.306	0.064	E
GLU358	-18.465	0.064	E
PHE66	-18.495	0.064	E
GLY349	-18.55	0.064	E
LEU145	-19.362	0.064	E
GLU249	-19.541	0.064	E
VAL250	-19.541	0.064	E
GLU109	-19.551	0.064	E
GLU341	-19.735	0.064	E
ARG340	-19.921	0.064	E
GLU131	-20.108	0.064	E
ARG143	-20.161	0.064	E
SER356	-20.249	0.064	E
ILE260	-20.453	0.064	E
SER251	-20.471	0.064	E
SER321	-20.514	0.064	E
PRO113	-20.581	0.064	E
GLU71	-20.672	0.064	E
LEU63	-20.745	0.064	E
GLU334	-20.777	0.064	E
ASN336	-20.801	0.064	E

Figure 27:- Immunogenicity score and probability score for each antigenic residue in ovalbumin.

```

Region rank: 1
P-value: 3.40391e-05
Number of residues: 18

LYS62
LEU63
PRO64
GLY65
PHE66
GLY67
ASP68
SER69
ILE70
GLU71
ALA72
GLN73
CYS74
GLY75
THR76
SER77
VAL78
ASN79

+++++
Region rank: 2
P-value: 0.000168061
Number of residues: 22

LYS187
ALA188
PHE189
LYS190
ASP191
GLU192

```

Figure 28-: The predicted epitope patches in decreasing order of p-value.

4.4.4 SPADE-: Surface comparison based Prediction of Allergenic Discontinuous Epitopes (Dall'Antonio *et al.*, 2011)

SPADE is entirely based on structural data and cross reactivity data.

4.4.4.1 COMPARISON MODULE

- Reference protein represents allergen of interest. Allergen of interest is uploaded.
- Comparison protein represents the protein which is cross-reactive to reference protein. Comparison protein is uploaded.
- Comparison module is executed. This module is executed for every CR protein with reference protein as allergen of interest.

4.4.4.2 EPITOPE PREDICTION MODULE

- In this module the compared proteins are categorized into highly cross reactive and weakly cross reactive categories.
- The surface properties for highly CR proteins will be added and subtracted for weakly CR proteins on reference protein.
- The epitopes are mapped on the surface of reference protein.

4.4.4.3 CRITERION FOR CROSS REACTIVITY (Aalberse *et al.*, 2001)

The cross reactivity between two proteins is based on phylogenetic relationship between those two proteins. High homology in primary sequence results in homologous 3-D structure. This leads to high cross reactivity. Epitopes are surface structures. The most conserved part of structure is in the core. Majority of mutations doesn't occur at random. Some features of protein molecule are important for its stability and function. Higher the structural homology higher is the cross-reactivity and structural homology depends upon sequence homology. To determine CR proteins blast search can be performed against the reference protein. The highly homologous sequences will be highly cross reactive proteins.

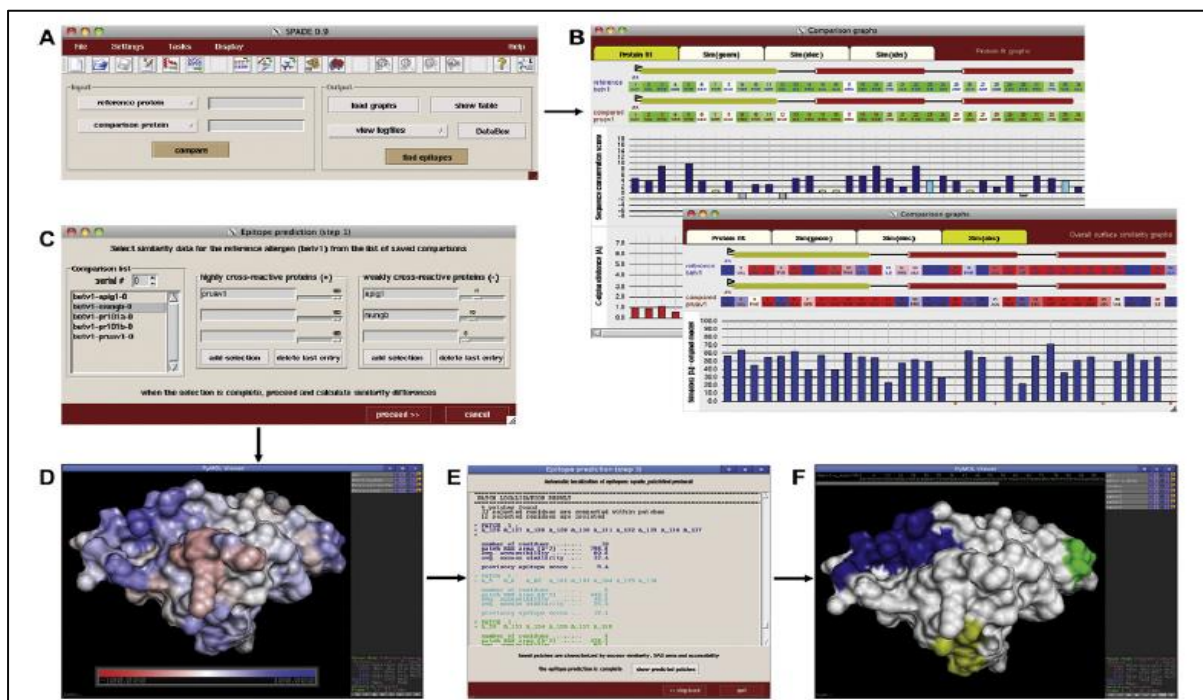


Figure 29-: A) The SPADE window. B) Representing the compared properties of reference and comparison proteins. C) Representing the classification of proteins according to cross-reactivity. D) Representing the mapping of properties on the surface. E) Representing Text output with predicted epitope residues listed E) Display of epitopes in pymol (Dall'Antonio *et al.*, 2011).

4.4.5 EPIPOPE ANALYSIS

- The predicted epitopes, linear and conformational, from all tools are mapped separately on the protein.
- These epitopes are compared with epitopes predicted experimentally.
- This step also gives us an idea about the approach that gives best result and also about the limitation of epitope prediction approaches.
- The proteins that lack experimental epitope information, a consensus result of all the tools is considered. The overlapping regions were considered highly immunogenic.

5.RESULTS

5.1 EGG PROTEOME

- Total number of proteins in hen's egg = 104
- We have removed all the redundant entries and uncharacterized proteins.
- Each protein has a unique Uniprot ID and is highly annotated.

Sr. No.	ENTRY NAME	PROTEIN NAME
1	P19121	Serum albumin
2	O57579	Aminopeptidase N
3	P18908	Natriuretic peptides A
4	P02659	Apovitellenin-1
5	P0DJJ2	Astacin-like metalloendopeptidase
6	P02701	Avidin
7	Q3V6R6	Cholesteryl ester transfer protein
8	P01038	Cystatin
9	P21760	Extracellular fatty acid-binding protein
10	Q90964	Forkhead box protein G1
11	O42220	Growth/differentiation factor 8
12	Q6IV20	Gallinacin-11
13	P46156	Gallinacin-1
14	P46158	Gallinacin-2
15	Q9DG58	Gallinacin-3
16	P0C1H3	Histone H2B
17	P10184	Ovoinhibitor
18	P01005	Ovomucoid
19	P00698	Lysozyme C
20	Q98UI9	Ovomucin
21	F1NBL0	Mucin-6
22	P29616	Myosin heavy chain
23	F1NSM7	Ovocleidin-116
24	Q9PRS8	Ovocleidin-17 (OC-17)
25	Q9YHY9	Ornithine carbamoyltransferase
26	P01014	Ovalbumin-related protein Y
27	P01012	Ovalbumin
28	P20740	Ovostatin
29	Q2VRL0	1-phosphatidylinositol 4,5-bisphosphate
30	P02752	Riboflavin-binding protein (RBP)
31	Q9YH85	Alpha-tectorin
32	P02789	Ovotransferrin
33	P87498	Vitellogenin-1
34	P02845	Vitellogenin-2
35	Q91025	Vitellogenin-3

36	P41366	Vitelline membrane outer layer protein 1
37	E1BTE1	Wee1-like protein kinase 2
38	P79762	Zona pellucida sperm-binding protein 3
39	M1RMG9	Very low density lipoprotein
40	Q49MC0	Vimentin
41	Q8QGU2	Peptidyl-prolyl cis-trans isomerase
42	Q8AV77	Hep21 protein
43	Q9IBC9	CD9 antigen
44	O42288	Interleukin 2
45	Q6LEL2	Egg white lysozyme
46	A5HIN3	Bone morphogenetic protein receptor type 1B
47	F8U4V7	Cytochrome c oxidase subunit 1
48	B6V1G0	Ovomucoid
49	Q90ZG0	Peptidyl-prolyl cis-trans isomerase
50	I0J171	OvoglobulinG2
51	F1NGS3	Protein-tyrosine sulfotransferase 2
52	D3KYT5	Ovocalycin-32
53	Q9PRR7	OVOFACTOR-1
54	Q766V2	Zona pellucida protein D
55	E0A2T5	Heme oxygenase 1
56	Q9DER4	Zona pellucida protein 1
57	P01013	Ovalbumin-related protein X
58	Q6E6M8	Extracellular fatty acid-binding protein
59	Q8QFM7	Chondrogenesis associated lipocalin
60	Q8JIG5	Alpha 1-acid glycoprotein
61	O42273	Protein TENP
62	Q9YGP0	Clusterin
63	Q9YHT1	Succinate dehydrogenase
64	P53478	Actin, cytoplasmic type 5
65	P49702	ADP-ribosylation factor 5
66	Q10751	Angiotensin-converting enzyme (ACE)
67	P08250	Apolipoprotein A-I
68	Q5G8Y9	Apolipoprotein D
69	P19204	Calsequestrin-2
70	Q703P0	Corticotropin releasing hormone
71	Q9PSS4	C-SKI protein
72	Q90839	Dickkopf-related protein 3
73	Q90844	Follistatin (FS)
74	Q90593	78 kDa glucose-regulated protein
75	P20136	Glutathione S-transferase 2
76	P15505	Glycine dehydrogenase
77	Q02391	Golgi apparatus protein 1
78	O73840	Heparin cofactor II
79	P09987	Histone H1
80	P35062	Histone H2A-III

81	Q9PUK9	High mobility group protein HMG1
82	Q90890	Lymphocyte antigen 86
83	Q92062	Melanotransferrin/EOS47
84	O42146	Metalloproteinase inhibitor 2
85	P26652	Metalloproteinase inhibitor 3
86	Q8AXY6	Muscle, skeletal receptor tyrosine protein kinase
87	O57596	CEPU-Se alpha 2
88	Q5ZJH2	Nicalin (Nicastrin-like protein)
89	Q25C36	Olfactomedin-like protein 3
90	Q90YI1	Ovocalyxin-32 (OCX-32)
91	Q9PRS8	Ovocleidin-17 (OC-17)
92	P24367	Peptidyl-prolyl cis-trans isomerase B
93	Q91348	6-phosphofructo-2-kinase
94	P32760	Pleiotrophin (PTN)
95	P26446	Poly [ADP-ribose] polymerase 1
96	P24802	Procollagen-lysine,2-oxoglutarate 5-dioxygenase
97	Q1XIH7	Renin/prorenin receptor
98	Q8JGM4	Sulfhydryl oxidase 1
99	P10039	Tenascin (TN)
100	P0CG62	Polyubiquitin-B
101	P25022	V(D)J recombination-activating protein
102	P47990	Xanthine dehydrogenase/oxidase
103	Q98T82	Zinc-finger transcription factor KROX20
104	A0AVX7	Calcineurin B homologous protein 3

Table 8-: Proteins in hen's egg. Column 2 represents Uniprot ID's and Column 3 represents protein name.

5.2 ALLERGENICITY PREDICTION-: ALGPRED

- Algpred classifies the protein either as Potential Allergen, Allergen or Non-Allergen.
- We have used SVM based on amino acid composition and dipeptide composition.
- Alpred has predicted 17 proteins as potential allergens. These proteins have very high probability of being an allergen.
- Remaining 87 proteins are predicted either as allergens or non-allergens. They have either negative score or score below threshold.
- The results for potential allergens are enlisted in table.
- Protein name, length and presence of unique domains in these potential allergens are tabulated in subsequent tables.

Sr. No.	ENTRY NAME	SVM1	SVM2	ALLERGENICITY
1	Q9PSS4	0.81	0.43	PA
2	Q90ZG0	0.91	0.57	PA
3	O42288	0.86	0.48	PA
4	P00698	0.84	0.66	PA
5	Q90890	0.88	0.93	PA
6	P24367	0.58	0.57	PA
7	P01005	1.29	0.99	PA
8	Q49MC0	0.94	0.83	PA
9	Q90844	0.87	0.52	PA
10	P01012	0.87	0.49	PA
11	P01014	0.61	0.59	PA
12	P19204	1.01	0.84	PA
13	P10184	0.89	0.43	PA
14	P19121	0.93	0.69	PA
15	Q90593	0.88	0.54	PA
16	P02789	0.98	0.65	PA
17	Q98UI9	0.94	0.62	PA

Table 9-: The proteins predicted as potential allergens out of complete egg proteome. SVM1 is support vector machine based on amino acid composition and SVM2 is based on dipeptide composition. PA represents Potential Allergen.

ENTRY NAME	PROTEIN NAME	LENGTH
Q9PSS4	C-SKI protein	101
Q90ZG0	Peptidyl-prolyl cis-trans isomerase	108
O42288	Interleukin 2	143
P00698	Lysozyme C	147
Q90890	Lymphocyte antigen 86	160
P24367	Peptidyl-prolyl cis-trans isomerase	207
P01005	Ovomucoid	210
Q49MC0	Vimentin	258
Q90844	Follistatin	343
P01012	Ovalbumin	386
P01014	Ovalbumin-related protein Y	388
P19204	Calsequestrin-2	406
P10184	Ovoinhibitor	472
P19121	Serum albumin	615
Q90593	78 kDa glucose-regulated protein	652
P02789	Ovotransferrin	705
Q98UI9	Ovomucin	2108

Table 10-: The description of potential allergens. Name and length of each predicted allergen is specified.

ENTRY NAME	Domains
Q9PSS4	Domain Of Unknown Function
Q90ZG0	Fkbp-Isomerase Domain
O42288	Il-15 Domain
P00698	Alpha Lactalbumin Family
Q90890	Ml Domain
P24367	Cyclophilin Type Domain
P01005	Kazal Like Domain
Q49MC0	Intermediate Filament Domain
Q90844	Kazal Like Domain
P01012	Serpin Domain
P01014	Serpin Domain
P19204	Calsequestrin
P10184	Kazal Like Domain
P19121	Albumin
Q90593	Hsp 70 D0main
P02789	Transferrin Domain
Q98UI9	Cysteine Rich Domain, Von Willebrand Factor Type Domain

Table 11-: The classification of potential allergens into functional domains.

5.3 PROTEIN STRUCTURE PREDICTION

Out of 17 proteins, structure file in PDB is available for only three proteins, namely ovalbumin, ovotransferrin and lysozyme C.

- Ovalbumin structural file lacks information about many residues. Therefore the structure has to be modelled.
- Ovotransferrin crystal structure lacks information about first 20 residues. Therefore the structure has to be modelled.
- Lysozyme C structure lacks information about first 25 residues. Therefore the structure has to be modelled.
- No structural information is available for remaining 14 proteins.
- For remaining 14 allergen proteins the 3-D structure is modelled.
- Ovomucin is composed of 2108 residues. Servers have a limit of predicting structures for proteins whose length is less than 1500 residues. Thus, no analysis can be carried out on this protein due to lack of 3-D structure.
- The structures are predicted using I-TASSER and PHYRE2.

Sr. No.	ENTRY NAME	PROTEIN NAME	PDB STRUCTURE
1	Q9PSS4	C-SKI protein	NA
2	Q90ZG0	Peptidyl-prolyl cis-trans isomerasE	NA
3	O42288	Interleukin 2	NA
4	P00698	Lysozyme C	2YVB
5	Q90890	Lymphocyte antigen 86	NA
6	P24367	Peptidyl-prolyl cis-trans isomerase	NA
7	P01005	Ovomucoid	NA
8	Q49MC0	Vimentin	NA
9	Q90844	Follistatin	NA
10	P01012	Ovalbumin	1OVA
11	P01014	Ovalbumin-related protein Y	NA
12	P19204	Calsequestrin-2	NA
13	P10184	Ovoinhibitor	NA
14	P19121	Serum albumin	NA
15	Q90593	78 kDa glucose-regulated protein	NA
16	P02789	Ovotransferrin	1OVT
17	Q98UI9	Mucin-5B	NA

Table 12-: The availability of structural file in PDB for allergen proteins. NA represents not available in PDB database.

5.4 EPITOPE PREDICTION AND MAPPING

Epitopes are predicted for each potential allergen protein using four tools-:

- ELLIPRO
- SPADE
- SEPPA
- EPITOPIA

The epitopes predicted are compared with experimental predicted epitopes. This is done for those proteins for which the experimental information is available. For remaining proteins a consensus regions from all 4 approaches are considered as highly antigenic.

5.4.1 EPITOPE PREDICTION FOR OVALBUMIN

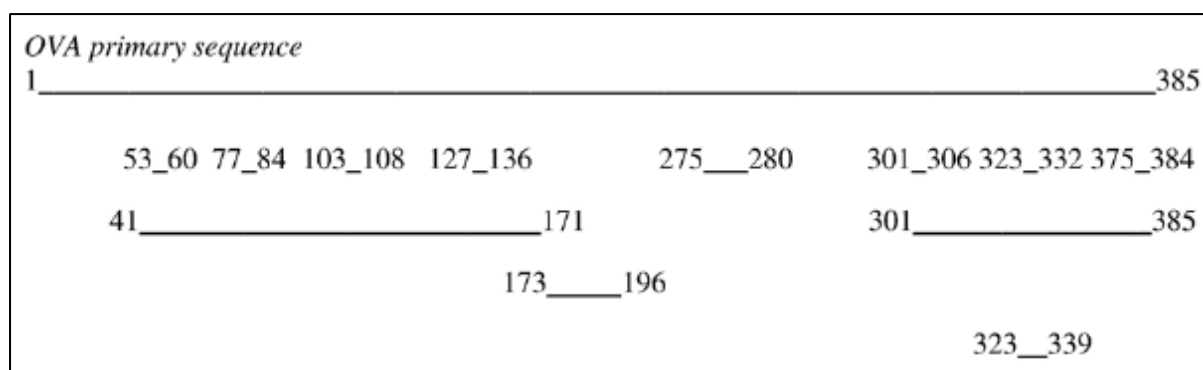


Figure 30-: The IgE epitopes determined experimentally (Yoshinori et al., 2006).

5.4.1.1 CONFORMATIONAL EPITOPES

5.4.1.1.1 SEPPA RESULTS

Total number of predicted epitope residues = 14.

Sr. No.	POSITION	RESIDUE	SCORE
1	191	GLU	1.89
2	65	PHE	1.88
3	72	GLN	1.88
4	195	ALA	1.88
5	66	GLY	1.87
6	73	CYS	1.85
7	136	THR	1.85
8	137	ALA	1.85
9	64	GLY	1.83
10	192	ASP	1.83
11	194	GLN	1.83
12	207	PRO	1.83
13	67	ASP	1.82
14	205	SER	1.81
15	204	GLU	1.8

Table 13-: Antigenic residues predicted by SEPPA. Blue colour represents those predicted residues that are in concordance with experimentally determined residues.

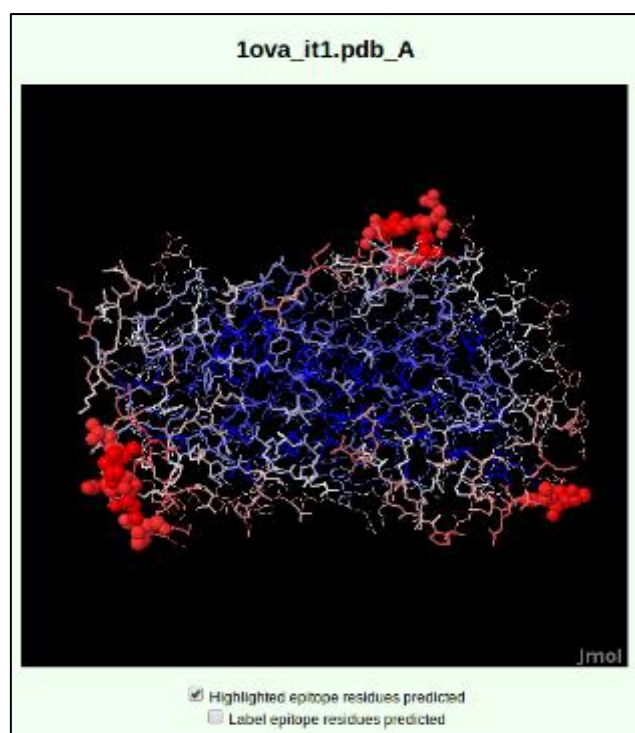


Figure 31:- The antigenic residues mapped on the protein.

5.4.1.1.2 EPITOPIA

Sr No.	RESIDUE	POSITION	I.S.	P.S.	B/E
1	ASP	190	-20.333	0.188	E
2	SER	250	-22.29	0.188	E
3	LYS	189	-22.474	0.188	E
4	SER	269	-22.496	0.188	E
5	GLU	319	-22.628	0.188	E
6	GLU	204	-22.72	0.188	E
7	ASN	133	-22.964	0.188	E
8	ASN	24	-23.006	0.188	E
9	ALA	223	-23.379	0.188	E
10	GLN	203	-23.458	0.188	E
11	GLU	191	-23.568	0.188	E
12	THR	136	-23.64	0.188	E
13	GLU	262	-23.938	0.188	E
14	THR	268	-24.068	0.188	E
15	GLU	266	-24.124	0.188	E
16	PRO	112	-24.173	0.188	E
17	SER	270	-24.185	0.188	E
18	LYS	206	-24.222	0.188	E
19	GLU	109	-24.231	0.188	E
20	PRO	63	-24.321	0.188	E
21	ALA	318	-24.352	0.188	E
22	GLU	248	-24.358	0.188	E

23	LEU	321	-24.359	0.188	E
24	THR	265	-24.414	0.188	E
25	ASN	271	-24.522	0.188	E
26	TYR	212	-24.528	0.188	E
27	GLY	237	-24.551	0.188	E
28	GLN	140	-24.778	0.188	E
29	ASP	192	-24.78	0.188	E
30	VAL	272	-24.799	0.188	E
31	ASP	247	-24.816	0.188	E
32	GLU	275	-24.822	0.188	E
33	PHE	65	-24.838	0.188	E
34	ARG	158	-24.888	0.188	E
35	GLU	274	-24.908	0.188	E
36	GLN	89	-24.986	0.188	E
37	SER	165	-25.053	0.188	E
38	MET	222	-25.093	0.188	E
39	THR	201	-25.114	0.188	E
40	GLN	254	-25.154	0.188	E
41	ARG	110	-25.194	0.188	E
42	ALA	23	-25.218	0.188	E
43	ILE	113	-25.244	0.188	E
44	LYS	263	-25.245	0.188	E
45	VAL	96	-25.261	0.188	E
46	ARG	199	-25.312	0.188	E
47	GLU	225	-25.318	0.188	E
48	PRO	93	-25.534	0.188	E
49	ALA	137	-25.632	0.188	E
50	VAL	249	-25.644	0.188	E
51	ALA	138	-25.683	0.188	E
52	SER	313	-25.739	0.188	E
53	ALA	337	-25.747	0.188	E
54	PRO	197	-25.877	0.188	E
55	SER	236	-25.883	0.188	E
56	SER	205	-25.901	0.188	E
57	GLY	155	-25.984	0.188	E
58	TYR	111	-25.987	0.188	E
59	ASP	95	-26.017	0.188	E
60	THR	91	-26.077	0.188	E
61	ARG	218	-26.172	0.188	E
62	ALA	351	-26.25	0.188	E
63	ILE	258	-26.289	0.188	E
64	LEU	114	-26.293	0.188	E
65	SER	224	-26.451	0.188	E
66	VAL	166	-26.561	0.188	E
67	GLU	143	-26.581	0.188	E

68	ASN	146	-26.584	0.188	E
69	GLU	150	-26.638	0.188	E
70	GLY	314	-26.701	0.188	E
71	SER	308	-26.844	0.188	E
72	GLY	64	-26.875	0.188	E
73	GLN	135	-26.934	0.188	E
74	GLU	202	-26.943	0.188	E
75	SER	147	-26.953	0.188	E
76	ALA	235	-27.022	0.188	E
77	GLU	253	-27.024	0.188	E
78	GLU	130	-27.113	0.188	E
79	LYS	92	-27.115	0.188	E
80	ASN	94	-27.133	0.188	E
81	ASN	159	-27.263	0.188	E
82	PRO	131	-27.301	0.188	E
83	ARG	359	-27.347	0.188	E
84	HIS	22	-27.353	0.188	E
85	ALA	220	-27.439	0.188	E
86	ILE	90	-27.457	0.188	E
87	ALA	195	-27.487	0.188	E
88	CYS	73	-27.575	0.188	E
89	GLU	116	-27.577	0.188	E
90	ILE	156	-27.583	0.188	E
91	SER	320	-27.611	0.188	E
92	PRO	163	-27.739	0.188	E

Table 14-: Antigenic residues predicted by Epitopia. Blue colour represents those predicted residues that are in concordance with experimentally determined residues.

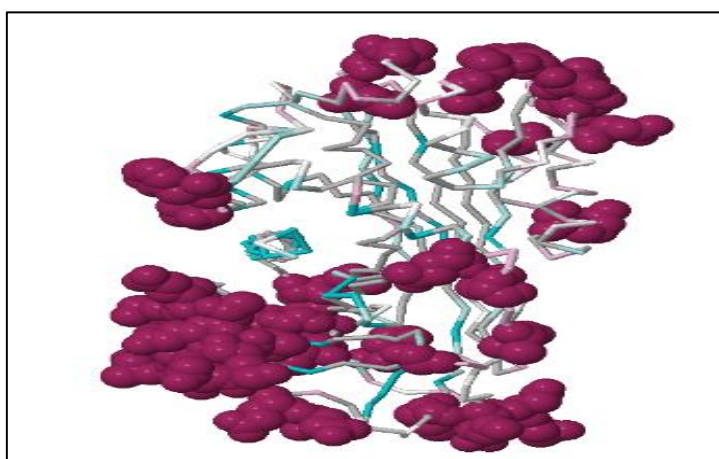


Figure 32-: The predicted epitopes mapped by Epitopia.

5.4.1.1.3 SPADE

HCR PROTEIN	PROTEIN NAME/ORGN	PERCENTAGE IDENTITY
G1MYK6	Ovalbumin (<i>Meleagris gallopavo</i>)	90.9%
E2RVI8	Ovalbumin (<i>Dromaius novaehollandiae</i>)	71.8%

Table 15-: The Highly cross reactive proteins (HCR) chosen for prediction.

WCR PROTEIN	PROTEIN NAME/ORGN	PERCENTAGE IDENTITY
P36952	SERPIN B5 (<i>Homo sapiens</i>)	30.9%
P01009	ANTITRYPSIN (<i>Homo sapiens</i>)	26.7%

Table 16-: The Weakly cross reactive proteins (WCR) chosen for prediction.

PATCH	RES NAME	RES NUMB
PATCH1	SER	2
	GLY	4
	ALA	5
	LYS	55
	ARG	58
	ASP	60
	LYS	61
	LEU	62
	PRO	63
	PATCH2	ALA
SER		352
VAL		353
SER		354
GLU		356
PATCH3	ASP	360
	HIS	361
	PRO	362
PATCH4	LYS	46
	ASP	47
	ARG	50

Table 17-: Epitopes predicted by SPADE. Blue colour represents those predicted residues that are in concordance with experimentally determined residues.

PATCH	MA	MEP	MH	TA
PATCH1	40.7	19.8	0.41	494.8
PATCH2	48.8	-27.8	0.33	266.2
PATCH3	57.3	-50.3	0.37	206.9
PATCH4	50.4	52.5	0.63	321.5

Table 18-: The value of various parameters predicted for each patch by SPADE. MA= mean accessibility, MEP= mean electrostatic potential, MH= mean hydrophobicity and TA= total area.

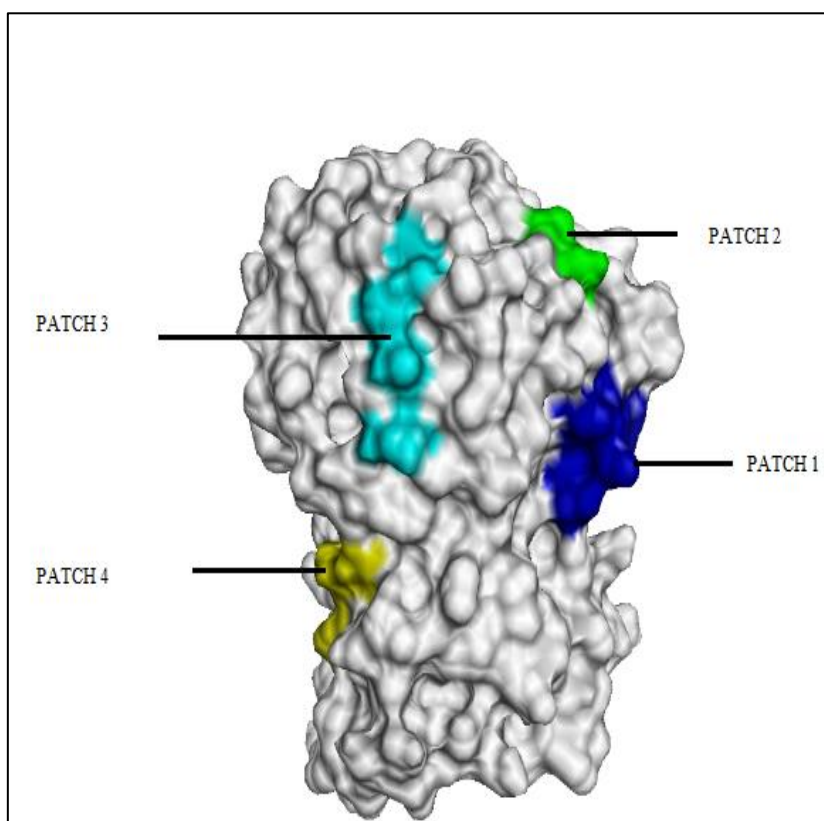


Figure 33-: The epitopes mapped on ovalbumin by SPADE.

5.4.1.2 LINEAR EPITOPES

5.4.1.2.1 EPITOPIA

PATCH	RESIDUE	POSITION	P-VALUE
PATCH 1	LYS	62	3.40E-05
	LEU	63	
	PRO	64	
	GLY	65	
	PHE	66	
	GLY	67	
	ASP	68	

SER	69		
ILE	70		
GLU	71		
ALA	72		
GLN	73		
CYS	74		
GLY	75		
THR	76		
SER	77		
VAL	78		
ASN	79		
PATCH 2	LYS	187	0.000168061
	ALA	188	
	PHE	189	
	LYS	190	
	ASP	191	
	GLU	192	
	ASP	193	
	THR	194	
	GLN	195	
	ALA	196	
	MET	197	
	PRO	198	
	PHE	199	
	ARG	200	
	VAL	201	
	THR	202	
	GLU	203	
	GLN	204	
	GLU	205	
	SER	206	
	LYS	207	
	PRO	208	
PATCH 3	GLU	249	0.000618399
	VAL	250	
	SER	251	
	GLY	252	
	LEU	253	
	GLU	254	
	GLN	255	
	LEU	256	
	GLU	257	
	SER	258	

ILE	259		
ILE	260		
ASN	261		
PHE	262		
GLU	263		
LYS	264		
LEU	265		
THR	266		
GLU	267		
TRP	268		
THR	269		
SER	270		
SER	271		
ASN	272		
VAL	273		
MET	274		
GLU	275		
GLU	276		
ARG	277		
PATCH 4	ALA	352	0.00188343
	ALA	353	
	SER	354	
	VAL	355	
	SER	356	
	GLU	357	
	GLU	358	
	PHE	359	
	ARG	360	
PATCH 5	LYS	93	0.00610143
	PRO	94	
	ASN	95	
	ASP	96	
	VAL	97	
	TYR	98	
	SER	99	

Table 19 -: Linear epitopes predicted by Epitopia. Blue colour represents those predicted residues that are in concordance with experimentally determined residues.

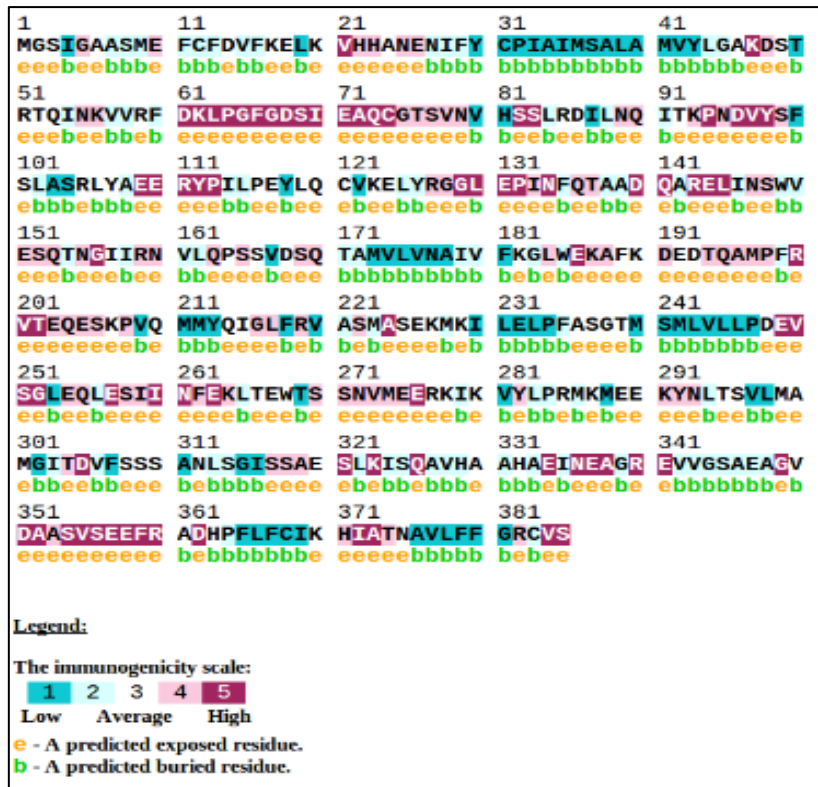


Figure 34:- Linear epitopes mapped on the sequence. Color coding represents the immunogenicity scale of the predicted residues.

5.4.1.2.2 ELLIPRO

PATCH	Res No.	Res Name	Score
PATCH1	182	GLY	0.794
	183	LEU	
	184	TRP	
	185	GLU	
	186	LYS	
	187	ALA	
	188	PHE	
	189	LYS	
	190	ASP	
	191	GLU	
	192	ASP	
	193	THR	
	194	GLN	
	195	ALA	
	196	MET	
	197	PRO	
	198	PHE	
	199	ARG	
	200	VAL	

201	THR		
202	GLU		
203	GLN		
204	GLU		
205	SER		
206	LYS		
207	PRO		
208	VAL		
209	GLN		
210	MET		
211	MET		
212	TYR		
213	GLN		
214	ILE		
215	GLY		
216	LEU		
217	PHE		
218	ARG		
219	VAL		
220	ALA		
221	SER		
222	MET		
223	ALA		
224	SER		
225	GLU		
226	LYS		
227	MET		
228	LYS		
PATCH2	266	GLU	0.774
	267	TRP	
	268	THR	
	269	SER	
	270	SER	
	271	ASN	
	272	VAL	
	273	MET	
	274	GLU	
	275	GLU	
	276	ARG	
	277	LYS	
	278	ILE	
	279	LYS	
	280	VAL	
	281	TYR	

PATCH3	59	PHE	0.745
	60	ASP	
	61	LYS	
	62	LEU	
	63	PRO	
	64	GLY	
	65	PHE	
	66	GLY	
	67	ASP	
	68	SER	
	69	ILE	
	70	GLU	
	71	ALA	
	72	GLN	
	73	CYS	
	74	GLY	
	75	THR	
	76	SER	
	77	VAL	
	78	ASN	
PATCH4	107	ALA	0.729
	108	GLU	
	109	GLU	
	110	ARG	
	111	TYR	
	112	PRO	
	113	ILE	
	114	LEU	
	115	PRO	
	116	GLU	
	117	TYR	
	118	LEU	
	119	GLN	
PATCH5	161	LEU	0.716
	162	GLN	
	163	PRO	
	164	SER	
	165	SER	
	166	VAL	
	167	ASP	
	168	SER	
	169	GLN	

	170	THR	
PATCH6	122	LYS	0.692
	123	GLU	
	124	LEU	
	125	TYR	
	126	ARG	
	127	GLY	
	128	GLY	
	129	LEU	
	130	GLU	
	131	PRO	
	132	ILE	
	133	ASN	
	134	PHE	
	135	GLN	
	136	THR	
	137	ALA	
	138	ALA	
	139	ASP	
	140	GLN	
	141	ALA	
	142	ARG	
	143	GLU	
	144	LEU	
PATCH7	45	ALA	0.683
	46	LYS	
	47	ASP	
	48	SER	
PATCH8	91	THR	0.678
	92	LYS	
	93	PRO	
	94	ASN	
	95	ASP	
	96	VAL	
	97	TYR	
PATCH9	302	ILE	0.598
	303	THR	
	304	ASP	
	305	VAL	
	306	PHE	

307	SER		
308	SER		
309	SER		
310	ALA		
311	ASN		
312	LEU		
313	SER		
314	GLY		
315	ILE		
316	SER		
317	SER		
318	ALA		
319	GLU		
320	SER		
321	LEU		
322	LYS		
PATCH10	350	ASP	0.593
	351	ALA	
	352	ALA	
	353	SER	
	354	VAL	
	355	SER	
	356	GLU	
	357	GLU	

Table 20-: Linear epitopes predicted by Ellipro. Blue colour represents those predicted residues that are in concordance with experimentally determined residues.

5.4.1.3 COMPARISON WITH EPITOPES PREDICTED EXPERIMENTALLY

Conformational epitopes predicted by SPADE, SEPPA and Epitopia are overlapped on experimentally determined epitopes. Epitopes predicted by SPADE has majorly overlapped with experimentally determined patches. Majority of patches predicted by Epitopia and SEPPA overlapped with experimentally determined patches but area and length of predicted patches is small as compared with SPADE.

Linear epitopes predicted by Ellipro and Epitopia are overlapped on experimentally determined epitopes. Patches predicted by Ellipro covers maximum surface of protein and therefore major portion of it overlaps with experimentally determined region. Five linear patches are predicted by Epitopia. The patches are shorter in length than patches predicted by ellipro. At least, three patches predicted by epitopia are successful in overlapping with experimentally determined region.

COLOUR	EPITOPES
YELLOW	Experimental epitopes
DARK GREEN	Conformational epitopes predicted by SPADE
DARK BLUE	Conformational and linear epitopes predicted by Epitopia
MAGENTA	Conformational epitopes predicted by SEPPA
SPRING GREEN	Linear epitopes predicted by Ellipro

Table 21-: The colour codes for epitopes mapped by different approaches.

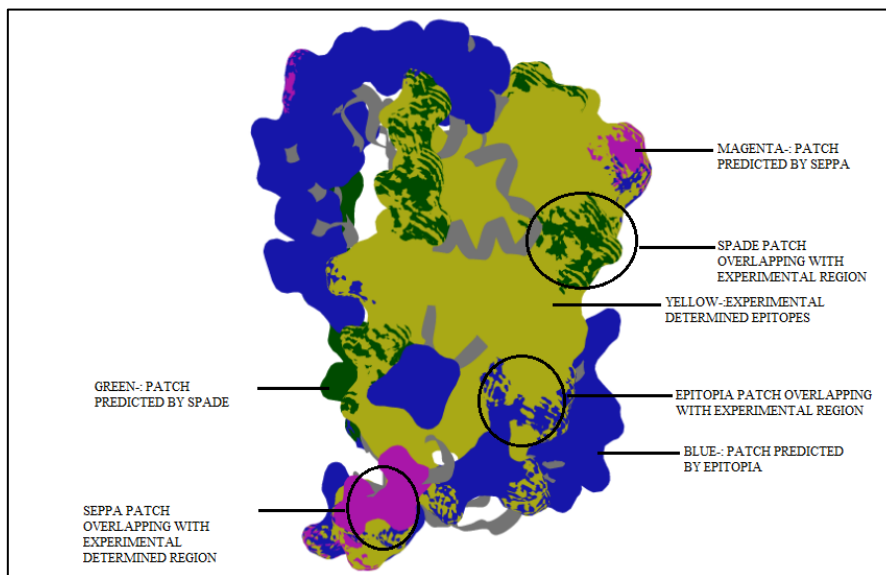


Figure 35-: The overlapping of conformational epitopes predicted by SPADE, Epitopia and SEPPA on epitopes determined experimentally.

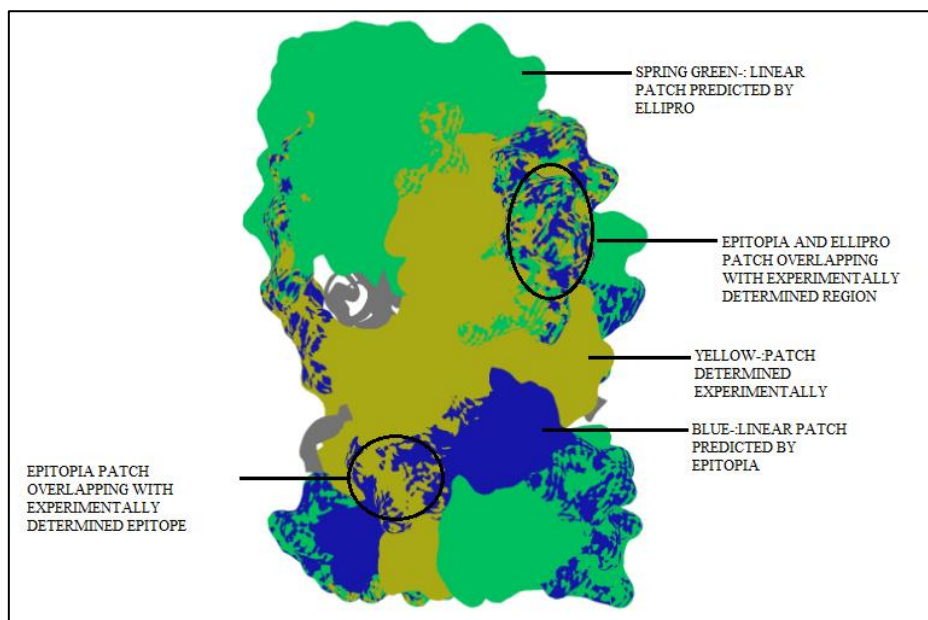


Figure 36-: The overlapping of linear epitopes predicted by Epitopia and Ellipro on epitopes determined experimentally.

5.4.2 EPITOPE PREDICTION FOR OVOMUCOID

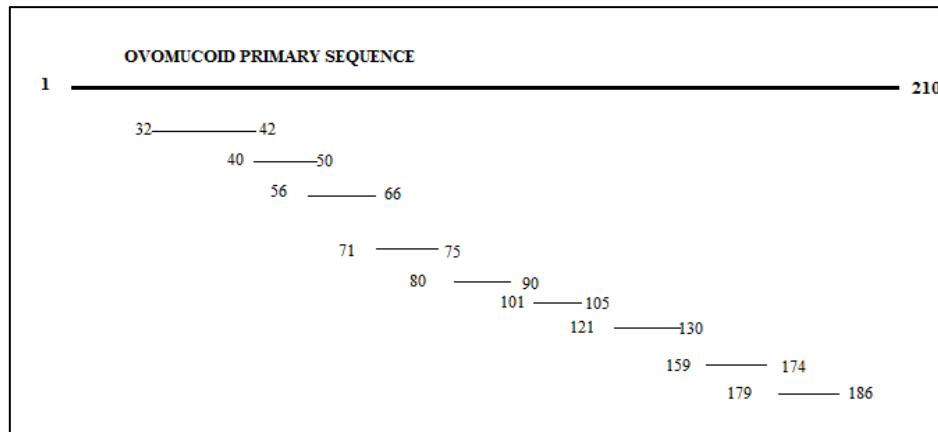


Figure 37-: The IgE epitopes determined experimentally.

5.4.2.1 COMPARISON WITH EPITOPES PREDICTED EXPERIMENTALLY

Conformational epitopes predicted by SPADE, SEPPA and Epitopia are overlapped on experimentally determined epitopes. Epitopes predicted by SPADE has majorly overlapped with experimentally determined patches. Majority of patches predicted by Epitopia and SEPPA overlapped with experimentally determined patches but area and length of predicted patches is small as compared with SPADE.

Linear epitopes predicted by Ellipro and Epitopia are overlapped on experimentally determined epitopes. Patches predicted by Ellipro covers maximum surface of protein and therefore major portion of it overlaps with experimentally determined region. Five linear patches are predicted by Epitopia. The patches are shorter in length than patches predicted by ellipro. At least, three patches predicted by epitopia are successful in overlapping with experimentally determined region.

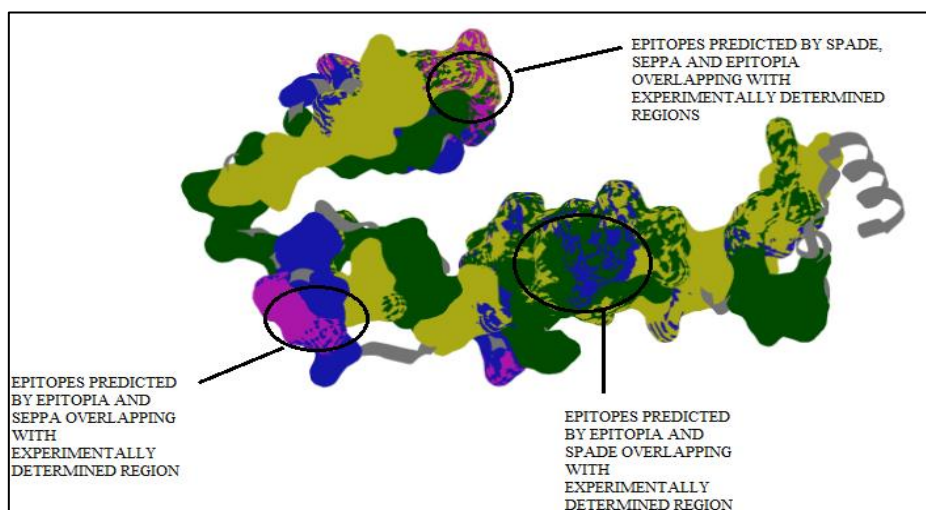


Figure 38-: The overlapping of conformational epitopes predicted by SPADE, Epitopia and SEPPA on epitopes determined experimentally.

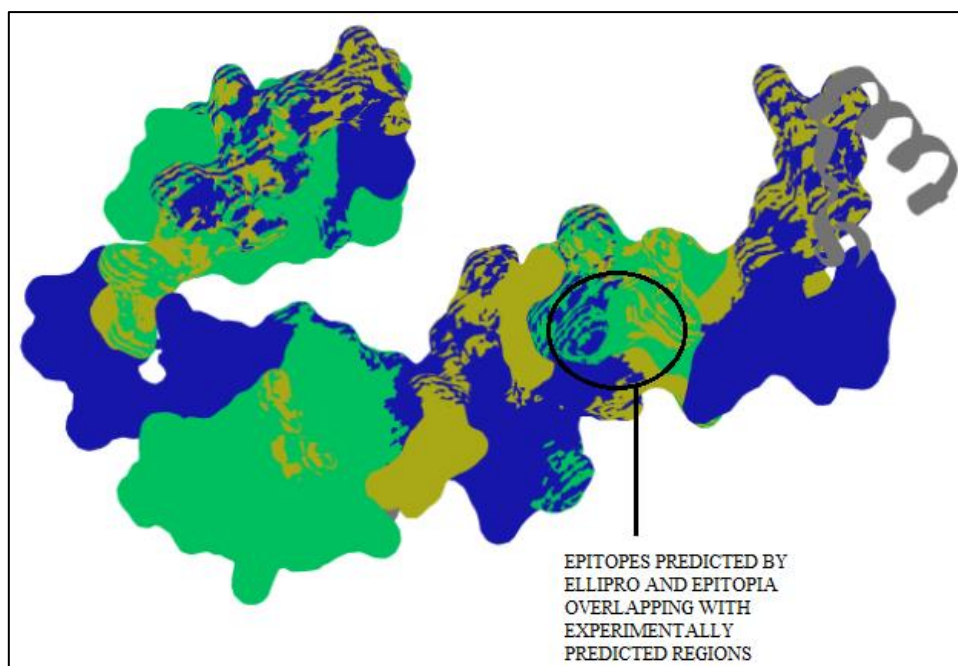


Figure 39-: The overlapping of linear epitopes predicted by Epitopia and Ellipro on epitopes determined experimentally.

5.4.3 EPITOPE PREDICTION FOR LYSOZYME

5.4.3.1 CONFORMATIONAL EPITOPES

5.4.3.1.1 SEPPA

Predicted epitope residues=71

Sr. No.	POSITION	RESIDUE	SCORE
1	34	GLY	2.5
2	39	ARG	2.42
3	118	SER	2.35
4	38	TYR	2.33
5	37	ASN	2.32
6	120	GLY	2.27
7	40	GLY	2.24
8	33	HIS	2.23
9	119	ASP	2.23
10	122	GLY	2.2
11	117	VAL	2.18
12	36	ASP	2.16
13	121	ASN	2.16
14	31	LYS	2.14
15	88	PRO	2.14
16	115	LYS	2.14
17	65	THR	2.12
18	123	MET	2.11

19	124	ASN	2.11
20	142	ILE	2.11
21	147	LEU	2.11
22	89	GLY	2.1
23	134	LYS	2.1
24	67	GLY	2.08
25	32	ARG	2.07
26	135	GLY	2.07
27	139	GLN	2.07
28	41	TYR	2.05
29	66	ASP	2.05
30	87	THR	2.05
31	114	LYS	2.05
32	111	ASN	2.03
33	90	SER	2.02
34	86	ARG	2.01
35	133	CYS	2.01
36	85	GLY	2
37	112	CYS	2
38	42	SER	1.99
39	91	ARG	1.99
40	116	ILE	1.98
41	129	TRP	1.98
42	136	THR	1.98
43	137	ASP	1.97
44	132	ARG	1.95
45	43	LEU	1.94
46	83	ASN	1.94
47	138	VAL	1.93
48	45	ASN	1.92
49	48	CYS	1.92
50	93	LEU	1.92
51	131	ASN	1.91
52	29	ALA	1.89
53	92	ASN	1.89
54	28	ALA	1.88
55	30	MET	1.88
56	80	TRP	1.88
57	68	SER	1.87
58	140	ALA	1.87
59	79	ARG	1.86
60	94	CYS	1.86
61	125	ALA	1.85
62	95	ASN	1.83
63	130	ARG	1.83

64	63	ARG	1.82
65	141	TRP	1.82
66	24	CYS	1.81
67	64	ASN	1.81
68	69	THR	1.8
69	84	ASP	1.8
70	128	ALA	1.8
71	143	ARG	1.8

Table 22-: Antigenic residues predicted by SEPPA. Blue colour represents those predicted residues that are in concordance with experimentally determined residues.

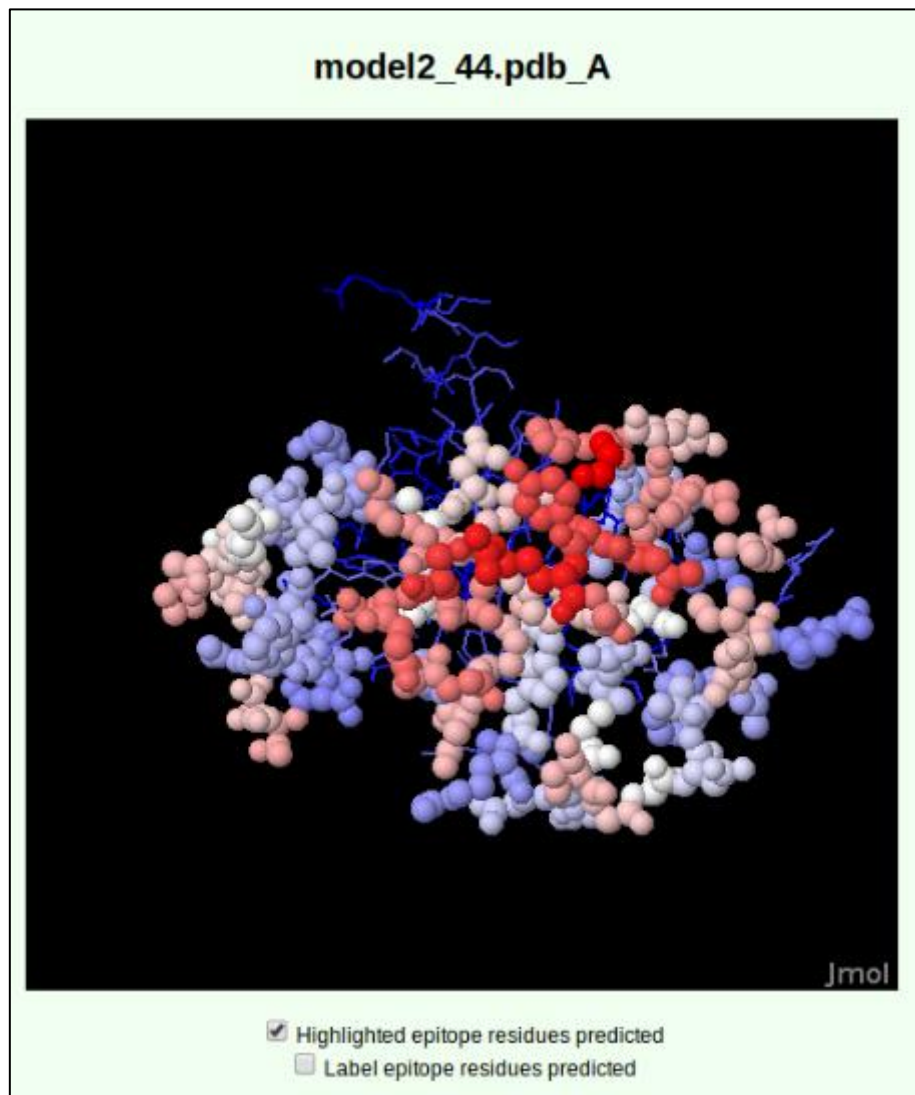


Figure 40-: Antigenic residues mapped by SEPPA.

5.4.3.1.2 EPITOPIA

RESIDUE	POSITION	I.S.	P.S.	B/E
ASP	137	-25.23	0.188	E

Table 23-: Antigenic residues predicted by Epitopia. IS= immunogenicity score and PS= probability score.

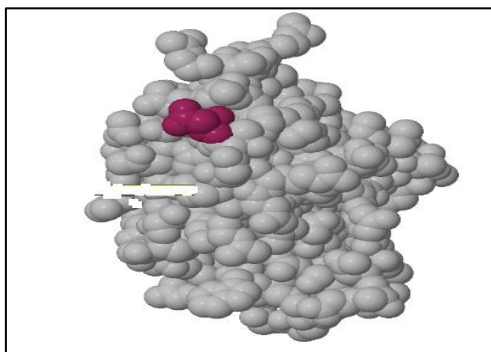


Figure 41-: The antigenic residues mapped on the protein.

5.4.3.1.3 SPADE

HCR PROTEIN	PROTEIN NAME/ORGN	PERCENTAGE IDENTITY
P00699	Lysozyme (<i>Lophortyx californica</i>)	85%
P00701	Lysozyme (<i>Coturnix japonica</i>)	95%
B8YK69	Lysozyme (<i>Bambusicola thoracicus</i>)	96%

Table 24-: The Highly cross reactive proteins (HCR) chosen for prediction.

WCR PROTEIN	PROTEIN NAME/ORGN	PERCENTAGE IDENTITY
P08334	Alpha-lactalbumin A (<i>Equus caballus</i>)	32.4%

Table 25-: The Weakly cross reactive proteins (WCR) chosen for prediction.

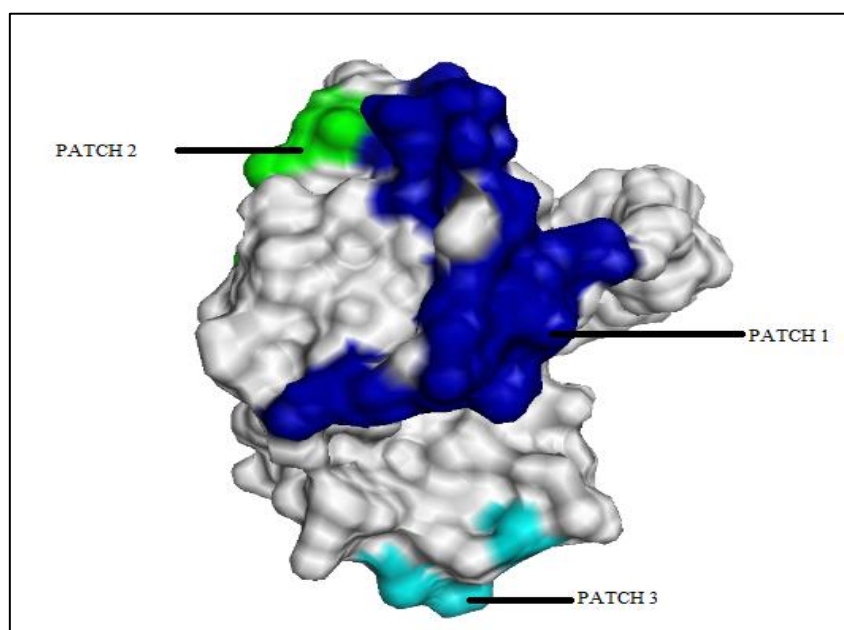


Figure 42-: The epitopes mapped on lysozyme by SPADE.

EPITOPES	RES. NAME	RES. NUM
PATCH1	GLY	120
	ASN	121
	ALA	125
	TRP	126
	VAL	127
	ALA	128
	ARG	130
	ASN	131
	ARG	132
	LYS	134
	THR	136
	ASP	137
	VAL	138
	GLN	139
	ALA	140
	ILE	142
ARG	143	
GLY	144	
PATCH2	ASP	66
	GLY	67
	ASN	83
	GLY	85
	ARG	86
	THR	87
	PRO	88
	GLY	89
	SER	90
	PRO	97
ALA	100	
PATCH 3	GLY	22
	CYS	24
	ALA	28
	ALA	29
	LYS	31
	ARG	32
	HIS	33
	GLY	34
LEU	147	

Table 26-: Epitopes predicted by SPADE.

PATCH	MA	MEP	MH	TA
PATCH1	44.3	15.5	0.29	1375.2
PATCH2	60.1	18.1	0.35	729.7
PATCH3	58.2	12.1	0.39	709.3

Table 27-: The value of various parameters predicted for each patch by SPADE. MA= mean accessibility, MEP= mean electrostatic potential, MH= mean hydrophobicity and TA= total area.

5.4.3.2 LINEAR EPITOPES

5.4.3.2.1 EPITOPIA

PATCH	RESIDUE	POSITION	P-VALUE		
PATCH1	ARG	32	0.00223547		
	HIS	33			
	GLY	34			
	LEU	35			
	ASP	36			
	ASN	37			
	TYR	38			
	ARG	39			
	GLY	40			
	TYR	41			
	SER	42			
	PATCH2	ASP		84	0.00359499
GLY		85			
ARG		86			
THR		87			
PRO		88			
GLY		89			
SER		90			
ARG		91			
ASN		92			
LEU		93			
PATCH3		PHE	52	0.00761835	
		GLU	53		
	SER	54			
	ASN	55			
	PHE	56			
	ASN	57			
	THR	58			
	GLN	59			
	ALA	60			
	THR	61			

	ASN	62	
	ARG	63	
	ASN	64	
	THR	65	
	ASP	66	
	GLY	67	
	SER	68	
	THR	69	
	ASP	70	
PATCH4	ASN	131	0.00970471
	ARG	132	
	CYS	133	
	LYS	134	
	GLY	135	
	THR	136	
	ASP	137	
	VAL	138	
	GLN	139	
	ALA	140	
	TRP	141	
	ILE	142	
	ARG	143	
	GLY	144	
	CYS	145	
	ARG	146	
	LEU	147	
PATCH5	SER	118	0.0283375
	ASP	119	
	GLY	120	
	ASN	121	
	GLY	122	
	MET	123	
	ASN	124	

Table 28-: Linear epitopes predicted by Epitopia.



Figure 43:- The linear epitopes mapped on the sequence.

5.4.3.2.2 ELLIPRO

PATCH	POSITION	RESIDUE	SCORE
PATCH1	130	ARG	0.771
	131	ASN	
	132	ARG	
	133	CYS	
	134	LYS	
	135	GLY	
	136	THR	
	137	ASP	
	138	VAL	
	139	GLN	
	140	ALA	
	141	TRP	
	142	ILE	
	143	ARG	
	144	GLY	
	145	CYS	
	146	ARG	
147	LEU		
PATCH2	1	MET	0.72
	2	ARG	
	3	SER	
	4	LEU	
	5	LEU	
	6	ILE	
	7	LEU	
	8	VAL	

9	LEU		
10	CYS		
11	PHE		
12	LEU		
13	PRO		
14	LEU		
15	ALA		
16	ALA		
17	LEU		
18	GLY		
19	LYS		
20	VAL		
21	PHE		
PATCH3	61	THR	0.624
	62	ASN	
	63	ARG	
	64	ASN	
	65	THR	
	66	ASP	
	67	GLY	
	68	SER	
	69	THR	
PATCH4	78	SER	0.602
	79	ARG	
	80	TRP	
	81	TRP	
	82	CYS	
	83	ASN	
	84	ASP	
	85	GLY	
	86	ARG	
	87	THR	
	88	PRO	
	89	GLY	
	90	SER	
	91	ARG	
	92	ASN	
	93	LEU	
	94	CYS	
	95	ASN	
	96	ILE	

Table 29-: Linear epitopes predicted by Ellipro.

5.4.3.3 CONSENSUS RESULT

Experimental epitopes are not available for lysozyme. The overlapping residues from all four approaches are reported. These residues or patches have high propensity of being immunogenic.

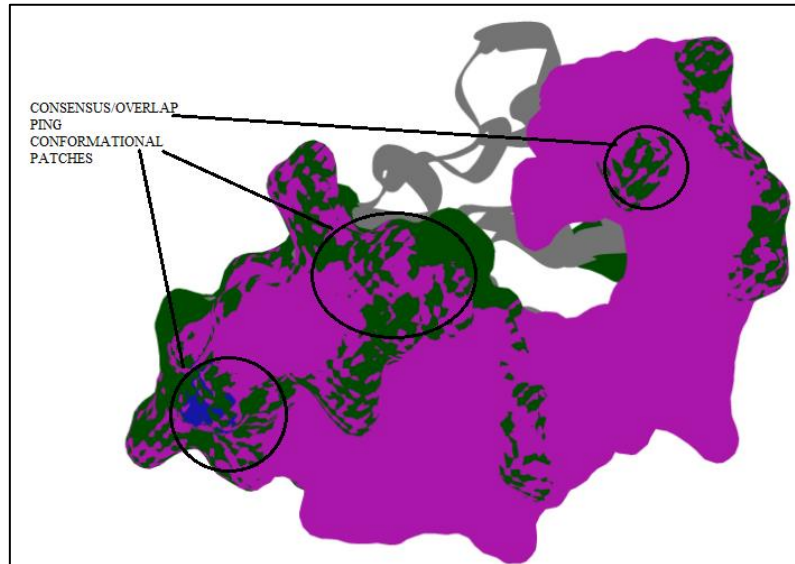


Figure 44-: The consensus conformational epitopes predicted by SPADE, Epitepia and SEPPA for lysozyme.

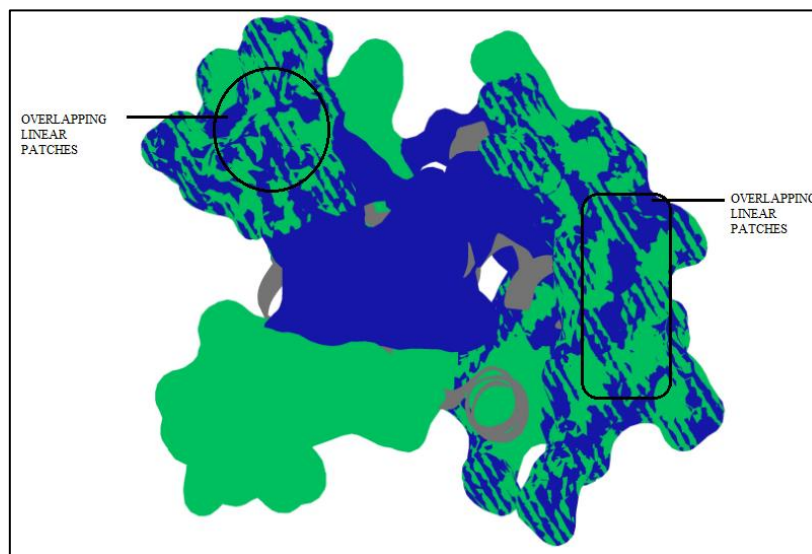


Figure 45-: The consensus linear epitopes predicted by Epitepia and Ellipro for lysozyme.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
24,28,29,31,32,33,34,66,67,83,85,86,87,88,89,90, 120,121,125,128,130,131,132,134,136,137,139,140, 142,143,147	32-41,61-69,84-93,131-147

Table 30-: The consensus conformational and linear epitope positions in lysozyme.

5.4.4 EPITOPE PREDICTION FOR INTERLEUKIN-12

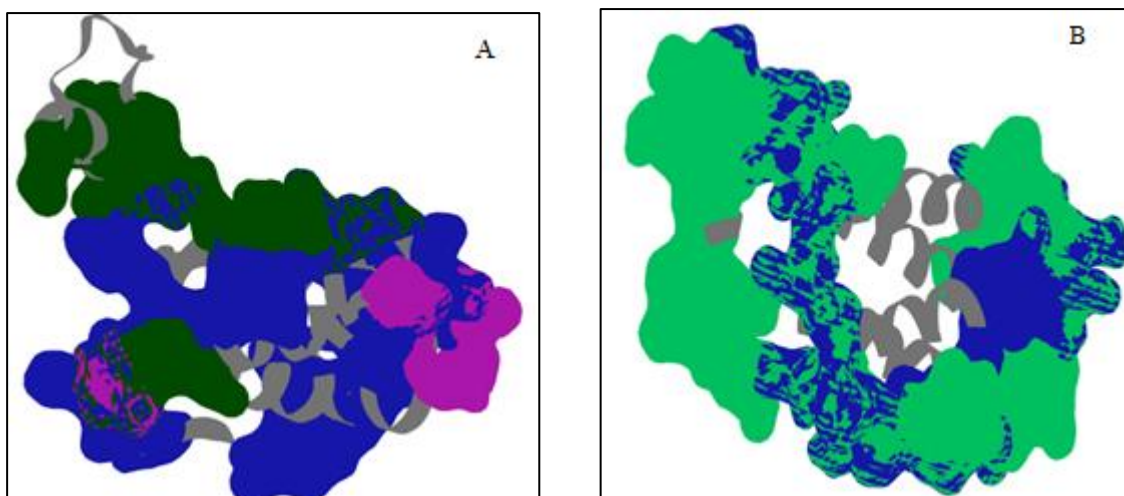


Figure 46:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for interleukin 12.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
39,41,42,44,45,48,49,50,85,106,107,108,109	23-31,108-117,80-87,121-125

Table 31:- The consensus conformational and linear epitope positions in interleukin 12.

5.4.5 EPITOPE PREDICTION FOR PEPTIDE PROLYL CIS TRANS ISOMERASE

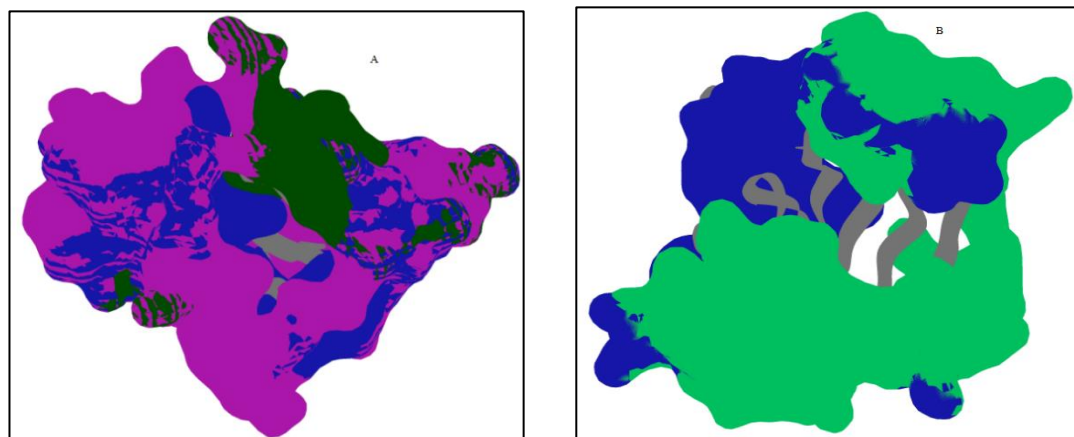


Figure 47:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for Peptide Prolyl Cis Trans Isomerase.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
1,6,8,9,10,11,12,13,14,15,16,17,18,19,31,32,33,35,37,41,46,48,50,51,52,53,54,55,69,70,72,83,86,90,91,93,94,96,97,105,107	4-21,45-46,84-91

Table 32:- The consensus conformational and linear epitope positions for Peptide Prolyl Cis Trans Isomerase.

5.4.6 EPITOPE PREDICTION FOR CALSEQUESTRIN

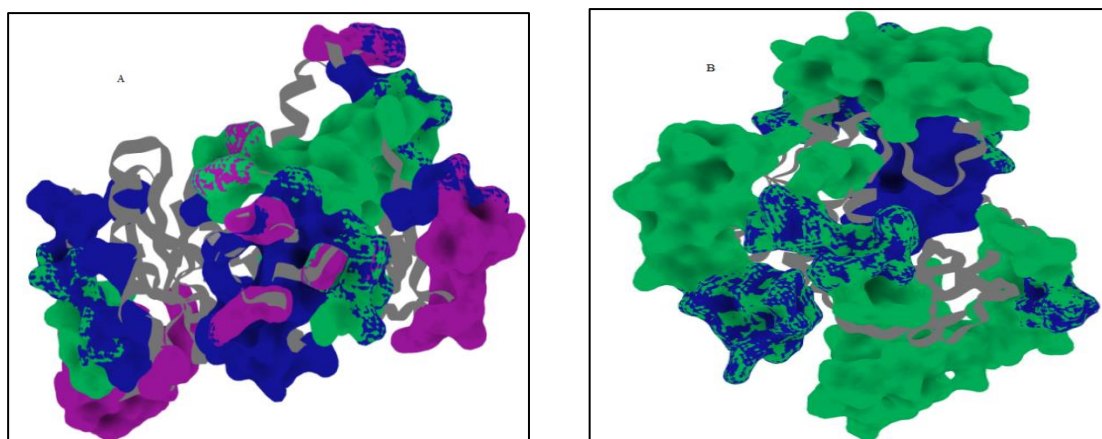


Figure 48:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for calsequestrin.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
21,22,29,30,31,33,39,40,42,43,44,47,48,49,50 64,82,100,101,104,105,148,150,152,153,157 159,172,173,175,176,177,179,180,210,211 229,351,352,355,356,357,359,363,364,366, 367,368,369	21-33,61-71,208-236,348-354

Table 33:- The consensus conformational and linear epitope positions for calsequestrin.

5.4.7 EPITOPE PREDICTION FOR 78 KDA GLUCOSE REGULATED PROTEIN

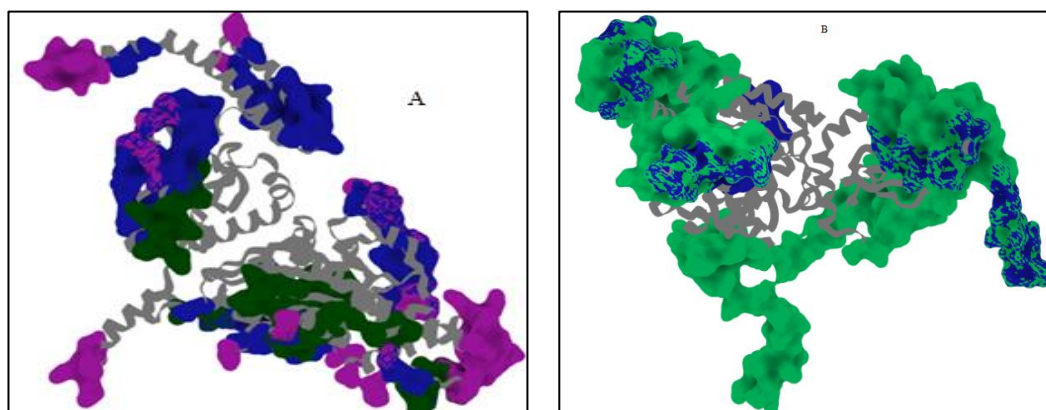


Figure 49:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for 78 KDA Glucose Regulated Protein.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
23,24,44,45,46,47,48,67,68,69,70,84, 102,104,107,122,131,132,133,134,135 306,308,309,486,487,488,489,514 515,516,581	120-138,270-281,377-382,572-591, 235-652

Table 34:- The consensus conformational and linear epitope positions for 78 KDA Glucose Regulated Protein.

5.4.8 EPITOPE PREDICTION FOR LYMPHOCYTE ANTIGEN 86

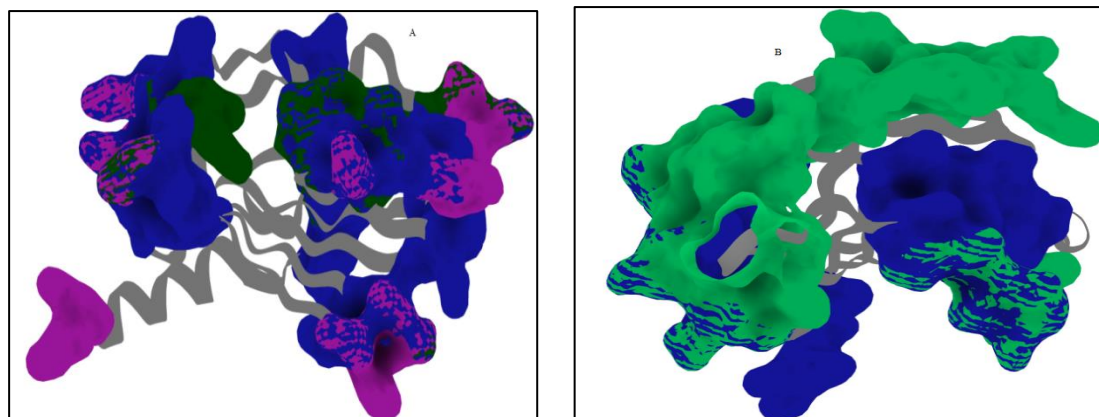


Figure 50-: The consensus conformational epitopes (A) and linear epitopes (B) predicted for Lymphocyte Antigen 86.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
22,23,30,31,58,59,60,74,85,86,89,98,99,100,108 110,112,126,159	18-25,123-132,156-160

Table 35-: The consensus conformational and linear epitope positions for Lymphocyte Antigen 86.

5.4.9 EPITOPE PREDICTION FOR PEPTIDE PROLYL CIS TRANS ISOMERASE B

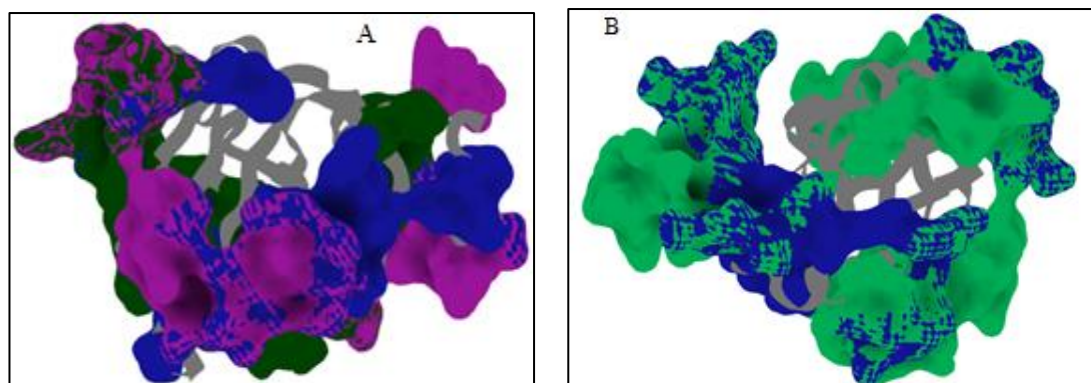


Figure 51-: The consensus conformational epitopes (A) and linear epitopes (B) predicted for Peptide Prolyl Cis Trans Isomerase B.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
24,25,26,28,38,72,99,101,103,104,107,111 112,113,115,116,134,135,136,137,177,178 179,180,181,182,183,184,185,186,187, 199,200,201,206	23-36,73-81,116,175-187,196-207

Table 36-: The consensus conformational and linear epitope positions in Prolyl Cis Trans Isomerase B.

5.4.10 EPIOTOPE PREDICTION FOR OVALBUMIN RELATED PROTEIN Y

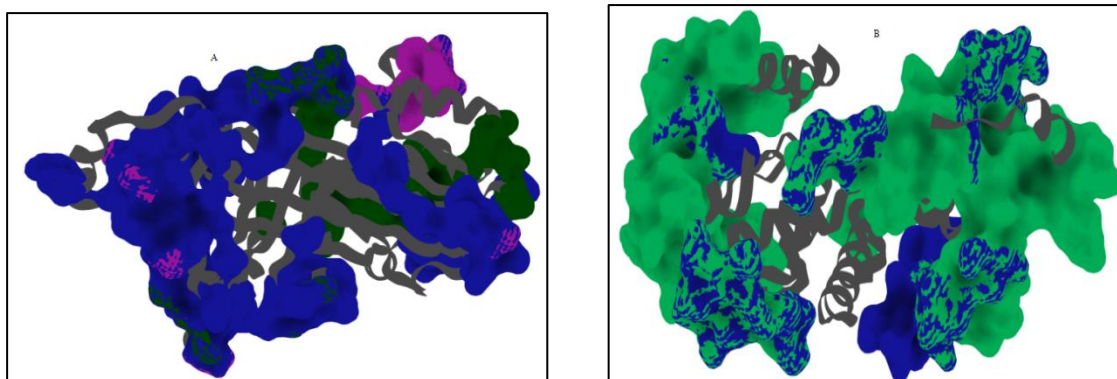


Figure 52:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for Ovalbumin Related Protein Y.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
23,24,71,73,92,93,94,95,96,137,192,198 202,205,206,271,272,278	62-79,94-98,128-134,187-208,267-277

Table 37:- The consensus conformational and linear epitope positions in Ovalbumin Related Protein Y.

5.4.11 EPIOTOPE PREDICTION FOR OVOTRANSFERRIN

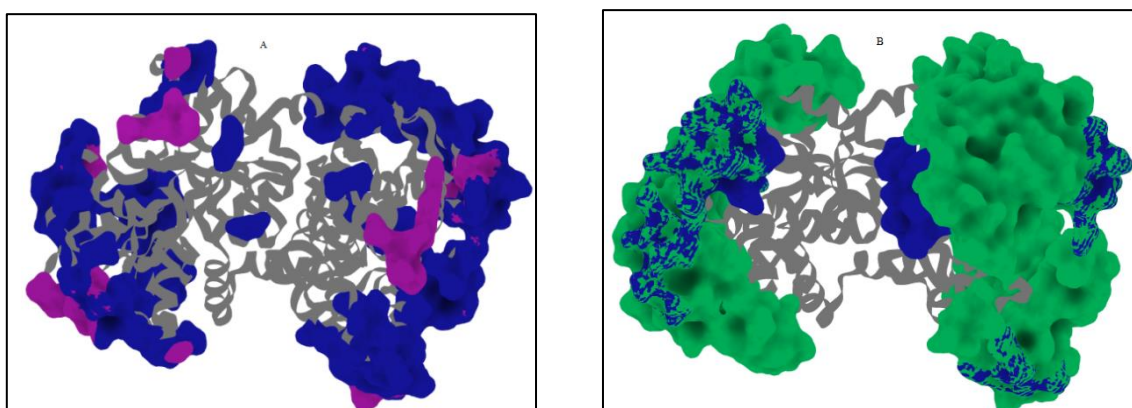


Figure 53:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for Ovotransferrin.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
33,34,48,106,160,182,194,195,196,197, 198,199,200,202,203,204,206,298,299 300,307,308,309,314,465,525,526,527, 529,569,570,573,575,578,579,580,591 592,593	195-207,291-300,434-448,591-602

Table 38:- The consensus conformational and linear epitope positions in Ovotransferrin.

5.4.12 EPITOPE PREDICTION FOR VIMENTIN

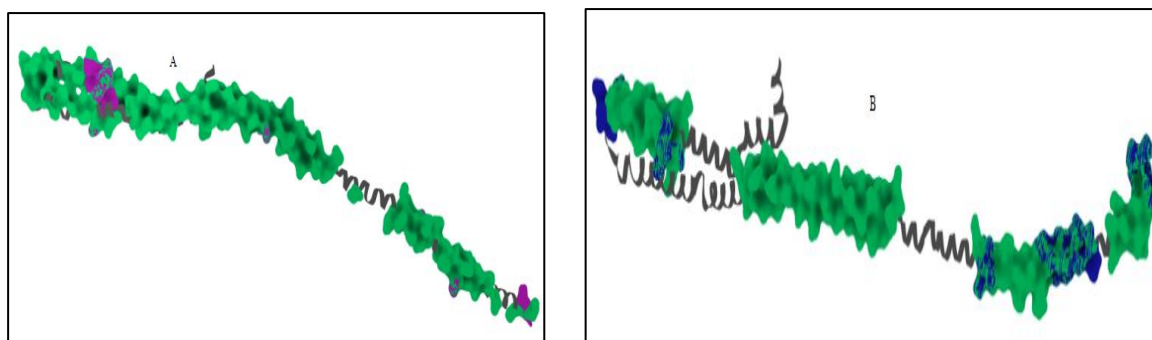


Figure 54-: The consensus conformational epitopes (A) and linear epitopes (B) predicted for Vimentin.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
50,51,52,53,54,55,56,59,74,150,151, 199,200,201,201,215,217,218,219, 220,221,222,223,224,225,226,227, 228,229,230,231,232,244,249,250, 251,252,253,254,255,256,257,258	52-56,74,199-232,249-258

Table 39-: The consensus conformational and linear epitope positions in Vimentin.

5.4.13 EPITOPE PREDICTION FOR C-SKI PROTEIN

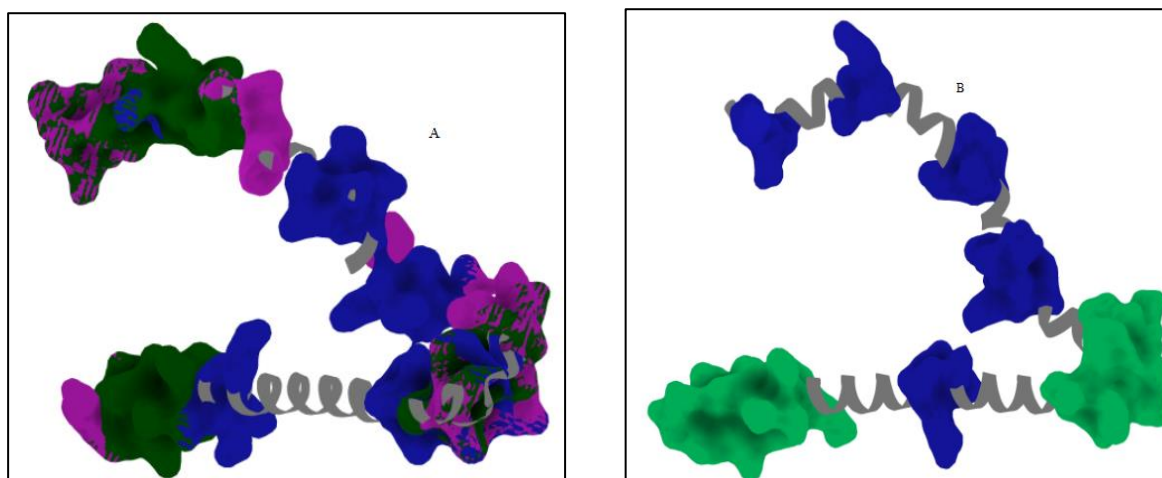


Figure 55-: The consensus conformational epitopes (A) and linear epitopes (B) predicted for C-SKI protein.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
1,2,6,7,8,9,10,11,12,18,43,49,50,54,55, 57,58,59,60,89,100	No consensus epitopes

Table 40-: The consensus conformational and linear epitope positions in C-SKI protein.

5.4.14 EPITOPE PREDICTION FOR SERUM ALBUMIN

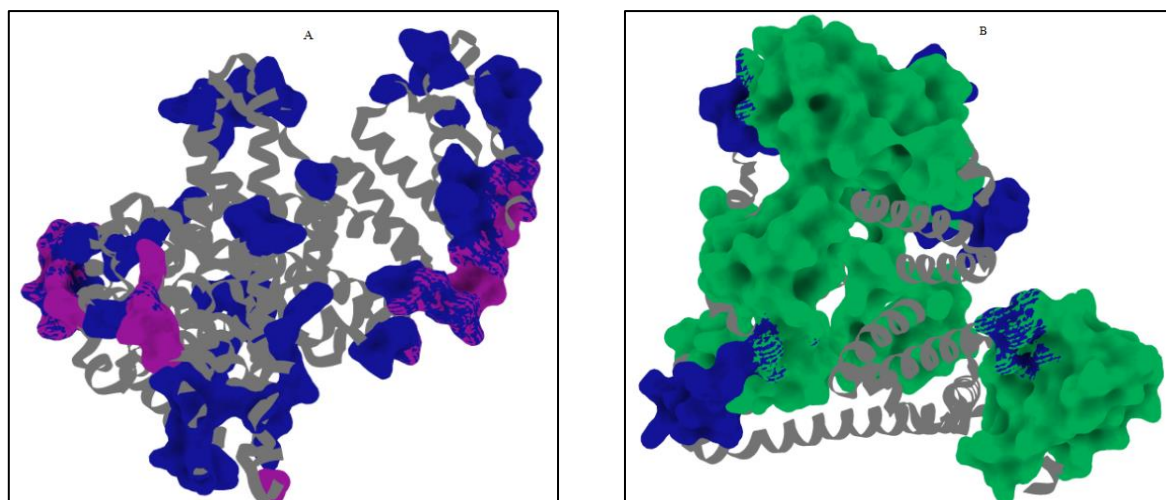


Figure 56:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for Serum albumin.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
39,79,81,82,83,84,85,86,327,328,499 520,521,522,523,527,528,564,565,566,571, 608,609,610,612	30-33,113,323-326,544-553

Table 41:- The consensus conformational and linear epitope positions in Serum albumin.

5.4.15 EPITOPE PREDICTION FOR OVOINHIBITOR

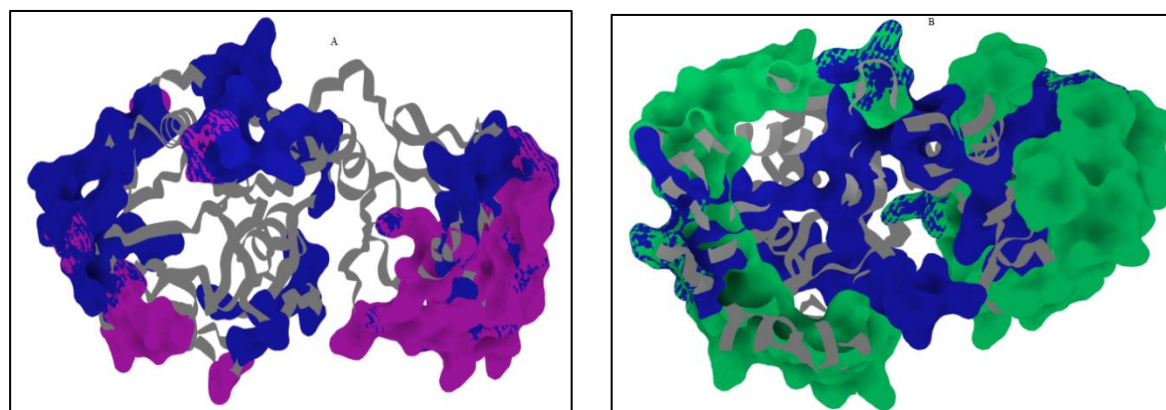


Figure 57:- The consensus conformational epitopes (A) and linear epitopes (B) predicted for Ovoinhibitor.

Consensus Conformational Epitopes Position	Consensus Linear Epitopes Position
31,42,50,51,52,54,55,56,65,70,80,81,84,93 94,95,96,99,101,111,119,140,219,221,222 352,366,367,442,443	23-30,147-166,220-229,342-363

Table 42:- The consensus conformational and linear epitope positions in Ovoinhibitor.

6. DISCUSSION

Total number of proteins in egg proteome was 104. The 17 proteins were predicted as potential allergens by support vector machine based methods. The structures were available for three proteins namely, ovalbumin, ovotransferrin and lysozyme. The structures available for three proteins have many residues missing from the structural file. Therefore, the structures were predicted for all 17 proteins by *ab-initio* approaches. We were successful in predicting and mapping IgE epitopes on 16 potential allergens. The allergenic protein ovomucin is 2108 amino acids long and lacks structural information in PDB. The huge size of this protein limits its structure prediction by eminent prediction servers. Therefore, we were unable to perform our study on this protein. The epitopes, linear and conformational, were mapped on 16 allergens using 4 different approaches (SPADE, SEPPA, ELLIPRO and EPITOPIA).

The information about experimental epitopes was available for 3 proteins namely, ovalbumin, ovomucoid and riboflavin binding protein. The predicted results were compared with available experimental epitopes. We have reported that SPADE gives best conformational patches. SEPPA and EPITOPIA predict residues that have high allergenicity. The predicted patches were smaller than that of SPADE. SPADE showed maximum overlapping on experimental epitopes as compared with SEPPA and EPITOPIA. SPADE had a limitation. It was unable to predict the epitopes if cross reactive structures are not available. The cross reactive information (sequence and structure) was not available for ovotransferrin, vimentin, ovoinhibitor and serum albumin proteins. We only considered SEPPA and EPITOPIA for predicting conformational epitopes in these 4 allergens.

In case of linear epitopes, both ELLIPRO and EPITOPIA gave equally good patches. The patches predicted by EPITOPIA were small as compared with ELLIPRO and ELLIPRO showed maximum overlapping with experimental epitopes.

The experimental epitope information was not available for remaining 13 allergens. Therefore, we have reported consensus results of all the approaches. The conformational epitope patches were predicted by taking overlapping residues predicted by SPADE, Epitopia and SEPPA. The linear epitope patches were predicted by taking overlapping residues predicted by Epitopia and Ellipro. These residues have high propensity to trigger severe hypersensitive and anaphylactic reactions.

7.CONCLUSION AND FUTURE PERSPECTIVE

Egg proteome consisting of 104 proteins were compiled from extensive literature survey and database searches. The proteins compiled were analysed for allergenicity using different softwares. A total of 17 proteins were predicted as potential allergens by support vector machine based methods. The three dimensional crystal structures were available for only three proteins namely, ovalbumin, ovotransferrin and lysozyme. The structures available for three proteins have many residues missing from the structural file. Therefore, all 17 proteins were modelled by a combination of threading and *ab-initio* approaches. Allergenic proteins react to its IgE antibodies through their specific epitopes – both linear and conformational. We have predicted and mapped, conformational and linear epitopes on the potential allergens using four different approaches. The experimentally determined epitopes were available for only three proteins namely, Ovalbumin, Ovomuroid and Riboflavin Binding Protein. We have compared our predicted results with available experimental information. The predicted region overlapping with experimentally determined epitopes is seen in results obtained for SPADE, followed by EPITOPIA and SEPPA. Thus, SPADE is best at predicting conformational epitopes followed by EPITOPIA and SEPPA. In case of linear epitope prediction, the overlapping with experimentally determined epitopes was best for ELLIPRO. EPITOPIA, also worked well with linear epitopes but the predicted region, overlapping with experimentally determined epitopes, is smaller as compared with ELLIPRO.

Furthermore, there were 13 proteins that lack information about epitopes predicted experimentally. We have presented a consensus result from all the 4 approaches for 13 allergic proteins. The epitopes, conformational and linear, are mapped on the proteins as molecular surface. They are represented with different colour codes. The region or residues that are overlapping are considered as highly immunogenic or allergenic. The positions for overlapping residues are enlisted in respective tables. These regions are considered as highly immunogenic and have large potential for triggering hypersensitivity and anaphylactic reactions.

In the future, docking studies can be carried out with human IgE antibody. The binding potential of the predicted epitope regions can be determined by protein-protein docking studies. The results will be useful for epitope identification and characterization based on a given protein sequence and structure information and pave way for vaccine development for allergic patients in future.

8. REFERENCES

- Aalberse RC. (2007). Assessment of allergen cross-reactivity. *Clin. Mol. Allergy*. **5**, 2-18.
- Aalberse, R. C; Akkerdaas, J. and Van Ree, R. (2001). Cross-reactivity of IgE antibodies to allergens. *Allergy*. **56**, 478-490.
- Aalberse, RC. (2000). Structural biology of allergens. *J. Allergy Clin. Immunol.* **106**, 228-238
- Aalberse, RC. and van Ree, R. (1997). Crossreactive carbohydrate determinants. *Clin. Rev. Allergy Immunol.* **15**, 375-387.
- Arnaldo, Cantani. (2008). *Pediatric Allergy, Asthma and Immunology*. Springer. **34**, 710–713.
- Bailey, TL. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell Syst. Mol. Biol.* **2**, 28-36.
- Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. (2002). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA*. **98**, 10037-10041.
- Binkowski, TA; Freeman, P. and Liang, J. (2004). pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic. Acids Res.* **32**, 555-558.
- Bock, S; Buckley, J; Holst, A. and May, C. (1978). Proper use of skin tests with food extracts in diagnosis of food hypersensitivity. *Clin. Allergy*. **8**, 559–564.
- Bock, SA. (1987). Prospective appraisal of complaints of adverse reactions to foods in children during the first 3 years of life. *Pediatrics*. **79**, 683–688.
- Bock, SA. and Atkins, FM. (1990). Patterns of food hypersensitivity during sixteen years of double-blind, placebo-controlled food challenges. *J. Pediatr.* **117**, 561–567.
- Bolhaar, ST; Tiemessen, MM; Zuidmeer, L; van Leeuwen, A; Hoffmann- Sommergruber, K. and Bruijnzeel-Koomen, CA. (2004). Efficacy of birch-pollen immunotherapy on cross-reactive food allergy confirmed by skin tests and double-blind food challenges. *Clin. Exp. Allergy*. **34**, 761-769.
- Boyano-Martinez, T; Garcia-Ara, C; Diaz-Pena, T. and Martin-Esteban, M. (2002). Prediction of tolerance on the basis of quantification of egg white-specific IgE antibodies in children with egg allergy. *Journal of Allergy and Clinical Immunology*. **110**, 304–309.

Bradford, WD. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*. **21**, 1487-1494.

Bruijnzeel-Koomen C; Ortolani, C; Aas, K; Bindslev-Jensen, C; Bjorksten, B; Noneret-Vautrin, D. and Wuthrich, B. (1995). Adverse reactions to food. *Allergy*. **50**, 623–635.

Brusic, V; Petrovsky, N; Gendel, SM; Millot, M; Gigonzac, O. and Stelman, SJ. (2003). Computational tools for the study of allergens. *Allergy* **58**, 1083-1092.

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

Burks, AW; James, JM; Hiegel, A; Wilson, G; Wheeler, JG; Jones, SM. and Zuerlein, N. (1998). Atopic dermatitis and food hypersensitivity reactions. *J. Pediatr*. **132**, 132–136.

Creamer LK. (1995). Effect of sodium dodecyl sulfate and palmitic acid on the equilibrium unfolding of bovine beta-lactoglobulin. *Biochemistry*. **34**, 7170-7176.

Dall'Antonia, F; Gieras, A; Devanaboyina, SC; Valenta, R. and Keller, W. Prediction of IgE-binding epitopes by means of allergen surface comparison and correlation to cross-reactivity *J. Allergy Clin. Immunol*. **128**, 872-879.

DeLano, WL. (2002) *The PyMOL Molecular Graphics System*. Palo Alto (CA): DeLano Scientific.

Derby, CJ; Gowland, MH. And Hourihane, JO. (2005). Sesame allergy in Britain: a questionnaire survey of members of the Anaphylaxis Campaign. *Pediatr. Allergy Immunol*. **16**, 171-175.

Dolinsky, TJ; Czodrowski, P; Li, H; Nielsen, JE; Jensen, JH. and Klebe, G. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res*. **35**, 522-525.

Fleischer, DM; Conover-Walker, MK; Christie, L; Burks, AW. and Wood, RA. (2003). The natural progression of peanut allergy: resolution and the possibility of recurrence. *J. Allergy Clin. Immunol*. **112**, 183-189.

Foetisch, K; Westphal, S; Lauer, I; Retzek, M; Altmann, F. and Kolarich, D. Biological activity of IgE specific for cross-reactive carbohydrate determinants. *J. Allergy Clin. Immunol*. **111**, 889-896.

Furmonaviciene, R; Sutton, BJ; Glaser, F; Laughton, CA; Jones, N. and Sewell, HF. (2005). An attempt to define allergen-specific molecular surface features: a bioinformatics approach. *Bioinformatics*. **21**, 4201-4204.

Grimbaldeston, MA; Metz, M; Yu M, Tsai M. and Galli, SJ. (2006). Effector and potential immunoregulatory roles of mast cells in IgE-associated acquired immune responses. *Curr. Opin. Immunol.* **18**, 751–760.

Guérin-Dubiard, C; Pasco, M; Mollé, D; Désert, C; Croguennec, T. and Nau, F. (2006). Proteomic analysis of hen egg white. *J. Agric. Food Chem.* **54**, 3901-3910.

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

Hileman, RE; Silvanovich, A; Goodman, RE; Rice, EA; Holleschak, G. and Astwood, JD. (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* **128**, 280-291.

Holt, C. and Sawyer, L. (1993). Caseins as rheomorphic proteins: interpretation of the primary and secondary structures of alpha S1-beta- and kappa caseins. *J. Chem. Soc. Farad Trans.* **89**, 2683-2692.

Holt, PG. and Sly, PD. (2007). TH2 cytokines in the asthma late-phase response. *Lancet.* **370**, 1396–1398.

Host, A. and Halcken, S. (1990). A prospective study of cow milk allergy in Danish infants during the first 3 years of life. *Allergy.* **45**, 587–596.

Hourihane, JO; Roberts, SA. and Warner, JO. (1998). Resolution of peanut allergy: case-control study. *BMJ.* **316**, 1271-1275.

Huang, J; Kawashima, S. and Kanehisa, M. (2007). New amino acid indices based on residue network topology. *Genome Informatics.* **18**, 152–161

Jakobsson, I. and Lindberg, T. (1979). A prospective study of cow's milk protein intolerance in Swedish infants. *Acta. Pediatr. Scand.* **68**, 853–859.

Jenkins, JA; Griffiths, Jones; Shewry, PR; Breiteneder, H. and Mills, ENC. (2005). Structural relatedness of plant food allergens with specific reference to cross-reactive allergens-an in silico analysis. *J. Allergy. Clin. Immunol.* **115**, 163-170.

Joris, Isabelle. and Majno, Guido (2004). Cells, tissues, and disease: principles of general pathology. *Oxford J.* **3**, 538-578.

Kabsch W. (1976). Solution for best rotation to relate 2 sets of vectors. *Acta. Crystallogr.* **32**, 922-923.

Kaneta, Y; Shoji, N; Ohkawa, T. and Nakamura, H. (2002). A method of comparing protein molecular surface based on normal vectors with attributes and its application to function identification. *Inform Sci.* **146**, 41-54.

Kato, I; Kohr, W. J. and Laskowski, M. (1978). Evolution of ovomucoids. *Proc. 11th FEBS Meeting.* **47**, 197-206.

Kelly, LA and Sternberg, MJE. (2009). Protein structure prediction on web: a case study using phyre server. *Nature protocols* **4**, 363-371.

Kulkarni-Kale, U; Bhosle, S. and Kolaskar, AS. (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res.* **33**, 168-171.

Leung, DY; Sampson, HA; Yunginger, JW; Burks, AW; Schneider, LC. and Wortel, CH. (2003). Effect of anti-IgE therapy in patients with peanut allergy. *N. Engl. J. Med.* **348**, 986-993.

Li, KB; Issac, P. and Krishnan, A. (2004). Predicting allergenic proteins using wavelet transform *Bioinformatics.* **20**, 2572-2578.

Li, XM; Zhang, TF; Huang, CK; Srivastava, K; Teper, AA. and Zhang, L. (2001) Food Allergy Herbal Formula-1 (FAHF-1) blocks peanut-induced anaphylaxis in a murine model. *J. Allergy Clin. Immunol.* **108**, 639-646.

Li-Chan, E. and Nakai, S. (1989). Biochemical basis for the properties of egg white. *Crit. Rev. Poult. Biol.* **2**, 21-58.

Magrane, M. and the UniProt consortium. (2011). UniProt Knowledgebase: a hub of integrated protein data Database.

Mandell, JG; Roberts, VA; Pique, ME; Kotlovyy, V; Mitchell, JC; Nelson, E; Tsigelny, I. and Ten Eyck, LF. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng.* **14**, 105-113.

Mari, A; Scala, E; Palazzo, P; Ridolfi, S; Zennaro, D. and Carabella, G. (2006). Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. *The Allergome platform as a model. Cell Immunol.* **244**, 97-100.

Munniyappa, K. and Adiga, P. R. (1979). Isolation and characterization of thiamin-binding protein from chicken egg white. *Biochem. J.* **177**, 887-894.

Nagase, H; Harris, E D; Woessner, J. F. and Brew, K. (1983) Ovostatin: a novel proteinase inhibitor from chicken egg white. Purification, physicochemical properties, and tissue distribution of ovostatin. *J. biol. Chem.* **258**, 7481-7489.

Neuvirth, H; Raz, R and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338**, 181-199.

Nisbet, A. D; Saundry, R. M; Moir, A. J. G; Fothergill, L. A. and Fothergill, J. E. (1981). The complete amino-acid sequence of hen ovalbumin. *Eur. J. Biochem.* **115**, 335-345.

Nowak-Wegrzyn, A. and Sampson, HA. (2004) Food allergy therapy. *Immunol. Allergy Clin.* **24**, 705-725.

Oksanen, E; Jaakola, VP; Tolonen, T; Valkonen, K; Akerstrom, B and Kalkkinen N. (2006). Reindeer beta-lactoglobulin crystal structure with pseudo-body-centered noncrystallographic symmetry. *Acta. Crystallogr. D Biol. Crystallogr.* **62**, 1369-1374.

Pearson WR. (2000). Flexible sequence similarity searching with the FASTA program package. (2000). *Methods Mol. Biol.* **132**, 185-219.

Ponomarenko, J; Bui, HH; Li, W; Fusseder, N; Bourne, PE; Sette, A. and Peters, B. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics.* **9**, 514-522.

Ponomarenko, J; Bui, HH; Li, W; Fusseder, N; Bourne, PE; Sette, A. and Peters, B. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics.* **9**, 514-522.

Ponomarenko, JV. and Bourne, PE. (2007). Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC. Struct. Biol.* **7**, 64-76.

Rabilloud, T. and Charmont, S. In *Proteome Research: Two- Dimensional Gel Electrophoresis and Identification Methods*. Springer. **45**, 107-126.

Radauer, C; Bublin, M; Wagner, S; Mari, A. and Breiteneder, H. Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *J Allergy Clin. Immunol.* **121**, 847-852.

Rothenberg, ME. (2004). Eosinophilic gastrointestinal disorders (EGID). *J. Allergy Clin. Immunol.* **113**, 11-28.

Roy, A; Kucukural A. and Zhang Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols.* **5**, 725-738.

Rubinstein, ND; Mayrose, I. and Pupko, T. (2009). A machine-learning approach for predicting B-cell epitopes. *Mol. Immunol.* **46**, 840-847.

Rubinstein, ND; Mayrose, I; Martz, E. and Pupko, T. (2009). Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinformatics.* **10**, 287-295.

Saha, S. and Raghava, GP. (2006). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, 202-209.

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

Sampson, H. and Ho, D. (1997). Relationship between food-specific IgE concentrations and the risk of positive food challenges in children and adolescents. *J. Allergy Clin. Immunol.* **100**, 444- 451.

Sampson, HA. (1999) Food allergy and immunopathogenesis and clinical disorders. *J. Allergy Clin. Immunol.* **103**, 717-728.

Sampson, HA. (2004). Update on food allergy. *J. Allergy Clin. Immunol.* **113**, 805-19.

Sampson, HA. and Albergo, R. (1984) Comparison of results of skin tests, RAST, and double-blind, placebo-controlled food challenges in children with atopic dermatitis. *J. Allergy. Clin. Immunol.* **74**, 26–33.

Sampson, HA. and Anderson, JA. (2000). Summary and recommendations: classification of gastrointestinal manifestations due to immunologic reactions to foods in infants and young children. *J. Pediatr. Gastroenterol. Nutr.* **30**, 87–94.

Sanner, MF; Olson, AJ. and Spehner, JC. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers.* **38**, 305-320.

Schneidman-Duhovny, D; Inbar, Y; Polak, V; Shatsky, M; Halperin, I; Benyamini, H; Barzilai, A; Dror, O; Haspel, N. and Nussinov, R. (2003). Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins.* **52**, 107-112.

Schönbach, Christian; Ranganathan, Shoba. and Brusica, Vladimir (2008). *Immunoinformatics.* Springer books. **1**, 1-200.

Schrander, JJP; van den Bogart, JPH; Forget, PP; Schrander-Stumpel, CTRM; Kuijten, RH. and Kester, ADM. (1993). Cow's milk protein intolerance in infants under 1 year of age: a prospective epidemiological study. *Eur. J. Pediatr.* **152**, 640–644.

Shatsky, M; Nussinov, R. and Wolfson, HJ. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins.* **56**, 143-156.

Shewry, PR. and Tatham, AS. (1999). The characterisation, structures and evolutionary relationships of prolamins. Kluwer Academic Publishers. **16**, 717-728.

Sicherer, SH. (2002). Food allergy. *Lancet.* **360**, 701-710.

Sicherer, SH. and Teuber, S. (2004). Current approach to the diagnosis and management of adverse reactions to foods. *J. Allergy Clin. Immunol.* **114**, 1146-1150.

Sicherer, SH; Mun˜oz-Furlong, A. and Sampson, HA. (2004). Prevalence of seafood allergy in the United States determined by a random telephone survey. *J. Allergy Clin. Immunol.* **114**, 159-65.

Sicherer, SH; Mun˜oz-Furlong; A. and Sampson; HA. (2003). Prevalence of peanut and tree nut allergy in the United States determined by means of a random digit dial telephone survey: a 5-year follow-up study. *J. Allergy Clin. Immunol.* **112**, 1203-1207.

Soeria-Atmadja, D; Zorzet, A; Gustafsson, MG. and Hammerling, U. (2004). Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int. Arch. Allergy Immunol.* **133**, 101-112.

Stadler, MB. and Stadler, BM. (2003). Allergenicity prediction by protein sequence. *FASEB J.* **17**, 1141-1143.

Stern, DA; Riedler, J; Nowak, D; Braun-Fahrlander, C; Swoboda, I. and Balic, N. (2007) Exposure to a farming environment has allergen-specific protective effects on TH2-dependent isotype switching in response to common inhalants. *J. Allergy Clin. Immunol.* **119**, 351-358.

Sun, J; Wu, D; Xu, T; Wang, X; Xu, X; Tao, L; Li, YX. and Cao, ZW. (2009). SEPPA: a computational server for spatial epitope prediction of protein antigens. *Nucleic Acids Res.* **37**, 612-616.

Thornton, JM; Edwards, MS; Taylor, WR. and Barlow, DJ. (1986). Location of continuous antigenic determinants in the protruding regions of proteins. *EMBO. J.* **5**, 409-413.

Turk, V; Brzin, J; Longer, M; Ritonia, A. and Eropkin, M. (1983). Protein inhibitors of cysteine proteinases. III. Amino acid sequence of cystatin from chicken egg white. *Hoppe-Seyler's Z. physiol. Chemic.* **364**, 1487-1496.

Urisu, A; Ando, H. and Morita, Y. (1997). Allergenic activity of heated and ovomucoid-depleted egg white. *Journal of Allergy and Clinical Immunology.* **100**, 171-176.

Vadas, P. and Perelman, B. (2003). Activated charcoal forms non-IgE binding complexes with peanut proteins. *J. Allergy Clin. Immunol.* **112**, 175-179.

Valenta, R; Steinberger, P; Duchene, M and Kraft, D (1996). Immunological and structural similarities Among allergens: prerequisite for a specific and component-based therapy of allergy. *Immunol. Cell. Biol.* **74**, 187-194.

Vieths, S; Scheurer, S. and Ballmer-Weber, B. (2002). Current understanding of cross-reactivity of food allergens and pollen. *Ann NY. Acad. Sci.* **964**, 47-68.

Williams, J. (1982). The evolution of transferrin. *Trends biochem. Sci.* **7**, 394-397.

Wood, RA. (2003). The natural history of food allergy. *Pediatrics.* **III**, 1631-1637.

Wu, CH; Yeh, LS; Huang, H; Arminski, L; Castro-Alvear, J; Chen, Y; Hu, Z; Kourtesis, P; Ledley, RS; Suzek, BE; Vinayaka, CR; Zhang, J. and Barker, WC. (2003) The Protein Information Resource. *Nucleic Acids Res.* **31**, 345-347.

Young, E. and Stoneham, MD; Petruckevitch, A; Barton, J; Rona, R. (1994). A population study of food intolerance. *Lancet.* **343**, 1127– 1130.

9.APPENDIX

Appendix I: List of Egg proteins that are predicted as allergens by Algpred. SVM1 is support vector machine based on amino acid composition and SVM2 is based on dipeptide composition.

UNIPROT ID	SVM 1	SVM 2	ALLERGENICITY
O57579	-0.168	-0.52	A
O42220	-0.31	-0.32	A
Q6IV20	-0.32	-0.45	A
P20740	-0.24	-0.133	A
Q9YH85	0.26	0.14	A
Q8QGU2	-0.27	-0.13	A
Q8AV77	0.28	0.11	A
A5HIN3	0.38	-0.06	A
Q9PRR7	-0.26	-0.188	A
Q9YGP0	-1.08	-1.33	A
P08250	-0.26	-0.3	A
Q90839	0.16	-0.46	A
P15505	-1.12	-0.52	A
Q02391	-0.25	-0.38	A
Q9PUK9	-0.12	0.28	A
Q92062	0.47	0.13	A
O42146	-0.08	-0.25	A
P26652	0.09	-0.52	A
O57596	0.09	-0.2	A
P32760	0.29	0.42	A
P26446	0.33	0.13	A
P10039	0.31	0.06	A
P0CG62	-0.16	-0.55	A
P47990	-0.157	-0.35	A
A0AVX7	-0.2	-0.68	A


```
        $flag=0;
        $counter++;
        say "done !";
    }
    else
    {
        $seq.=$ln;
    }
}
}
```