UPTEC X06 043
Examensarbete 20 p
Oktober 2006

Analysis of evolutionary co-variation of amino acid positions to discover features typical of allergens

Jonas Hagberg

UPPSALA UNIVERSITET

Bioinformatics Engineering Program

Uppsala University School of Engineering

UPTEC X 06 0043	**Date of issue 2006-10**

Author

Jonas Hagberg

Title (English)

Analysis of evolutionary co-variation of amino acid positions to discover features typical of allergens

Title (Swedish)

Abstract

In this study two protein families, both holding allergens and non-allergens, were investigated with regard to amino acid sequence features that may be attributed to allergenicity. With this purpose in mind, various computational biology operations were conducted, *e.g.* investigation on pair-wise co-variation of amino acids across the sequences. Intriguing patterns of co-varying pairs in and near known IgE epitopes were seen. The findings show that evolutionary co-variation analysis is a powerful method that can give valuable information on protein segments of potential importance to allergenicity.

Keywords
Allergy, Evolutionary co-variation, ELSC

Supervisors

Ulf Hammerling and Daniel Soeria-Atmadja
Department of Toxicology, National Food Administration

Scientific reviewer

Mats Gustafsson
Department of Engineering Sciences, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **40**

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

Analysis of evolutionary co-variation of amino acid positions to discover features typical of allergens

Jonas Hagberg
hagberg.jonas@gmail.com

November 10, 2006

Sammanfattning

Under senare år har förekomsten av allergier ökat, främst i västvärlden. Detta orsakar stor belastning på hälsovården. Allergi är relaterat till exponering av en grupp ämnen, benämda *allergener*, vilka huvudsakligen utgörs av proteiner. Allergener finns i vitt spridda ämnen såsom livsmedel, pollen, kvalster och pälsdjur.

Syftet med detta projekt är att undersöka två proteinfamiljer, båda innehållande kända allergena och icke-allergena proteiner, för att försöka finna allergen-specifika särdrag i aminosyrasekvenserna.

Flera bioinformatiska analysmetoder har använts, såsom multipel sekvensanalys, fylogenetisk analys, och analys av evolutionärt samvarierade parvisa positioner i aminosyrasekvenserna. Den sistnämnda metoden har möjliggjort påvisande av intressanta relationer mellan samvarierade positioner hos vissa allergena proteinsekvenser och kända områden där immunoglobulin E binder. Resultaten visar att analys av evolutionärt samvarierande positioner kan ge värdefull information, vilken kan vara viktig för förståelsen av allergenicitet, hos proteiner.

Examensarbete 20p i Civilingenjörsprogrammet för Bioinformatik
Uppsala universitet Oktober 2006

Contents

1 Introduction

The occurrence of allergy increases in the Western society and is a great health-care concern. Many environmental factors are believed to contribute to this increase in the prevalence of allergic diseases, such as urban living, Western life-style, reduced breast-feeding, allergen exposure, smoking, smaller families, fewer childhood infections and higher hygiene standards. Allergens are almost exclusively proteins and why they induce allergic responses is not yet fully understood, although, much progress has been made in recent years. Unintentional introduction of an allergen in genetically modified organisms (GMO) is a key aspect to consider in the risk assessment of new GMOs. Several bioinformatics methods that can predict protein allergenicity with reasonable accuracy, using a proteins Amino Acid (AA)-sequence, have been reported [1, 2, 3]. None of them, however, incorporate any information on protein structure.

In this study two protein families, both holding allergens and non-allergens, were investigated with regard to amino acid sequence features that may be attributed to allergenicity. With this purpose in mind, various computational biology operations were conducted, broadly involving multiple sequence alignment (MSA), phylogenetic analysis and investigating on pair-wise co-variation of amino acids across the sequences. A clear correlation in the tropomyosin-family between known IgE epitopes and discovered position is established. The findings in this study show that evolutionary co-variation analysis is a powerful method that can give valuable information on protein segments of potential importance to allergenicity.

In section 2 of this report the allergy concept is introduced and the risk of inadvertently introducing allergens in GMOs are presented. Moreover, information about protein families and most of the algorithms and bioinformatic methods used in this project are explained. Section 3 presents the aims of the project. In the materials and method part in section 4 information and creation of datasets used in this project are presented and several procedures and computer aid used to achieve the aims of the projects are outlined. Section 5 presents the results and, finally, the results are discussed in section 6.

2 Background

2.1 Allergy

Allergy is a fairly recently described disease. A hundred years ago the term allergy had not yet been defined, and typical symptoms of allergic disease, such as hay-fever, asthma and food intolerance, were rarely reported. In 1906 the term allergy was introduced by Clemens Von Pirquet and during the twentieth century allergy has emerged as a major global problem and the fraction of people affected has lately mounted to 20-25% of the population in some industrial nations [4].

Food allergens are mainly found in eight groups: milk, fish, eggs, crustaceans, peanuts,

soybeans, tree nuts, and wheat. These eight foods are reported to cover more than 90 % of all IgE (see section.2.1.1) mediated food allergies [5].

2.1.1 What are the mechanisms behind allergy

Allergy can be defined as an abnormal immunological reaction to certain exogenous substances, typically proteins. A person who is allergic develops symptoms when exposed to such, otherwise harmless, substances called allergens. Allergens can be divided into two groups, *major* and *minor* allergens. They are designated major if more than 50% of patients relative to the particular source have the corresponding allergen-specific IgEs, otherwise as minor [6]. IgE molecules recognize particular areas on the surface of proteins, commonly named B-cell epitopes [7].

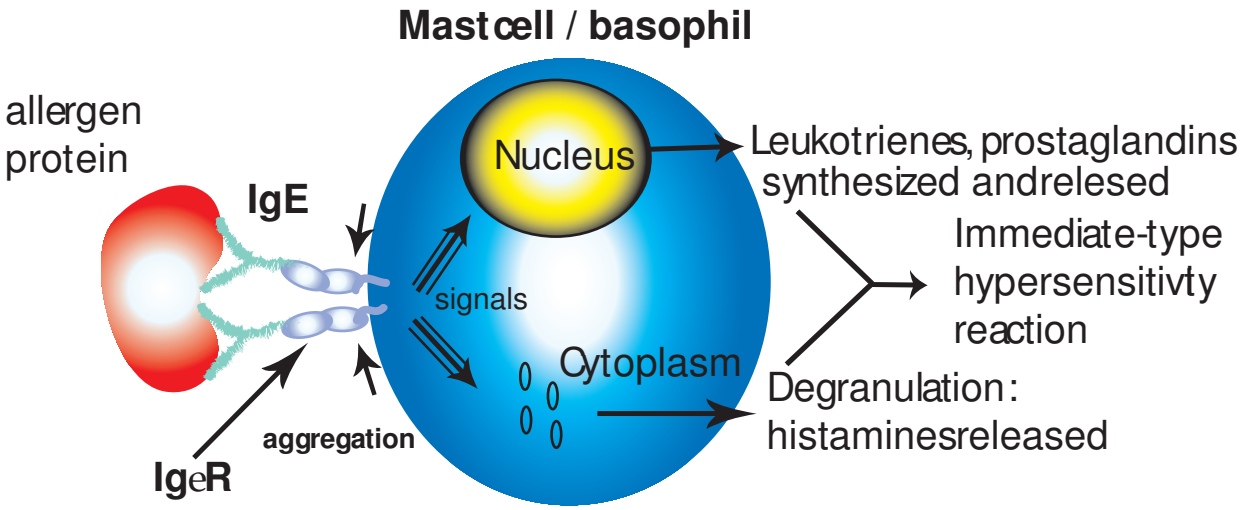

Figure 1: A sensitized mast cell with two IgEs on its surface has bound to an allergen protein. The bound between IgE and allergen is called cross-linking and is a necessity for an allergic response, and triggers degranulation of mast cells which leads to the release of inflaming mediators such as histamine etc.

Allergic people, who are sensitized to allergic proteins, have immunoglobulin E (IgE) bound to mast cells or basophils. When such mast cell-IgE antibody complexes react with an allergen the release of mediators, such as histamine, leukotrines, prostaglandins, cytokines and others is triggered (see figure 1 for an schematic view of an allergic reaction). The mediators then induce allergic symptoms in various target organs, typically the skin, the nose, the eyes, the chest etc. This kind of reactions are generally known as a type I hypersensitivity responses that occur due to an inappropriate immunoglobulin E synthesis. Hypersensitivity reactions can be divided into four types: type I through IV. This study is focusing on type I hypersensitivity, i.e. the IgE mediated reactions, and how they interact with protein molecules. They should not be confused with other sensitivity reactions, such as lactose or gluten intolerance. The structure of a protein is of great importance for a proteins allergenic ability/potency and is an important background to the study.

2.2 Genetically Modified Organism - GMO

A genetically modified organism harbours genetic material, which has been altered using various molecular genetic techniques. An outline of these techniques is beyond the scope of this report, but the methodology can be used to introduce highly specific changes of the phenotype. This is commonly achieved by altering expression levels of certain proteins produced by the organism or, more commonly by introducing entire genes that enable the production of xenogenic proteins. A major concern connected with genetically modified foods, with a particular relevance to this study, is the inadvertent introduction of novel allergenic proteins in food crops. This happened in 1996 when a protein from Brazil nut was transferred into soybean. The xeno-protein (2S albumin) increased the level of cysteine and methionine, which occur at relatively low levels in soybean. The modified crop would thus be a nutritionally improved feed to meat-producing livestock, such as poultry. As it turned out, however, the 2S albumin is also a major allergen in Brazil nut and this property was accordingly transferred to the recipient, i.e. the soybean acquired Brazilian nut allergenicity. Thus, patients that were allergic to Brazil nut, but not soybean, now had positive reaction upon exposure to transgenic soybean using skin prick test and immunoblotting on subject sera [8]. Based on these findings, further development of the GM soybean was discontinued.

The risk of unintentional introduction of an allergen in genetically modified organisms is an essential aspect to consider in the risk assessment of new GMOs. Several international regulatory bodies have proposed specific guidelines on procedures for the assessment of potential allergenicity of GM crops. [9, 10, 11]

2.3 Protein families

In this project two protein families were selected for analysis. Many proteins of these groups have known AA-sequences, and among them there are both defined allergens and nonallergenes.

As will be explained below, this makes a evolutionary co-variation analysis (sec.2.6.1) of the different groups Multiple Sequence Alignments (MSA) (sec.2.4) very well suited.

2.3.1 Tropomyosin

The tropomyosin group of proteins was discovered in 1948 by Bailey [12]. The members of this family are closely related and the proteins are present in muscle as well as in certain non-muscle cells. The evolutionary highly conserved tropomyosins bind to the sides of actin filaments and, in association with troponin, regulate the interaction of the filaments with myosin in response to Ca^{2+} [13]. Tropomyosins attain an alpha-helical configuration, which enables a coiled-coil structure of two parallel helices containing two sets of seven alternating acting binding sites [14]. The repeat pattern reads *a-b-c-d-e-f-g* wherein positions *a* and *d* are hydrophobic AA. Salt bridges between AA *e* and *g* of adjacent helices are assumed to

stabilize the coiled-coil structure [13]. Figure 2 shows how a tropomyosin is arranged as a head-to-tail linked polymer. The head-to-tail link is a central assumption in ideas about the interaction of tropomyosin with actin [13], thereby being special, and presumably particularly important regions of the protein.

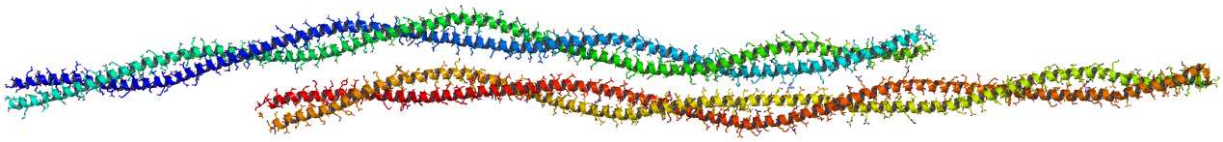

Figure 2: Head-to-tail linked Pig tropomyosin polymers, generated from PDB [15] structure 1C1G [13] via Pymol [16].

Tropomyosin is a key muscle protein in numerous vertebrate and invertebrate species [17] and is also present in yeast [18].

One of the proteins, the major shrimp allergen Pen a 1 has well characterized IgE epitopes [19]. Pen a 1 is the only known major allergen identified in shrimp and at least 80% of shrimp-allergic subjects react to tropomyosin [17]. Vertebrate tropomyosins are considered nonallergenic even though the degree of sequence similarity is high among tropomyosins and they are belived to shore a common function [17]. Invertebrate tropomyosins, on the other hand, are more likely to be allergenic and are important allergens in lobster, crabs, mollusks, house dust mites, cockroaches etc [17] (see figure 3 for some example species). The reason for differences in allergenicity between the two subgroups has not yet been explained.

No defined crystal 3D structure of allergen tropomyosin is available in the Protein Data Bank (PDB) [15], but several non-allergen tropomyosin structures occur in this repository.

(a) *Lepisma saccharina* (b) *Metapenaeus ensis* (c) *Charybdis feriatus*

Figure 3: Three species that have allergenic tropomyosin. Pictures from [20].

2.3.2 Parvalbumin

Parvalbumin, the major fish allergen, is a Ca^{2+} binding protein and is expressed at high levels in white muscle tissue of lower vertebrates, less abundantly in skeletal muscles of higher vertebrates as well as in a variety of non-muscle tissues, including testis, endocrine glands, skin and certain neurons [21]. There are two phylogenetic distinct lineages: the *alpha*-group, with less acidic parvalbumins and the *beta*-group holding more acidic parvalbumins. The allergenic parvalbumin from Cod belongs to the *beta*-lineage. Most muscles contains parvalbumin of either *alpha* or *beta* origin [22, 23]. Allergen parvalbumins can belong to either lineage.

Parvalbumins have only been recognized as allergen in fish and frog, despite the similar features of parvalbumin from other species [21]. Parvalbumin from fish is a major allergen; actually more than 90% of all fish-allergic patients react to this antigen. Allergen parvalbumin from fish is a very stable protein: Drastical changes of pH, temperature or exposure to dissociating agents do not significantly change its allergenicity [24].

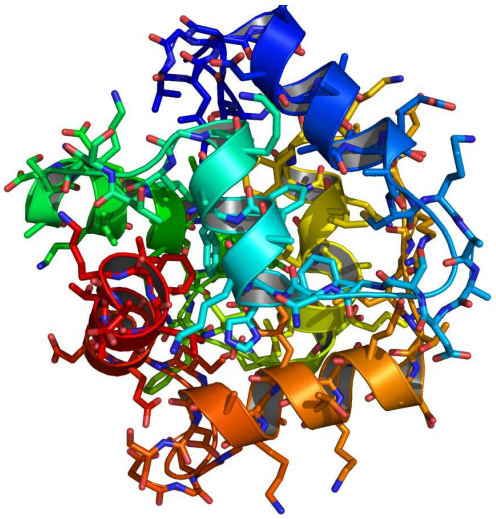

Figure 4: Carp parvalbumin generated from PDB-structure 4CPV [25] via PyMOL [16].

Parvalbumin is characterized by helix-loop-helix (HLH) binding motifs (two helices pack together at an angle of ~ 90 degrees, separated by a loop region where calcium binds) [23]. A single allergenic parvalbumin the Allergen Cyp C 1 from the common fish Carp, is structurally determined and occur in the PDB.

Studies have demonstrated dramatic conformational changes, not only in the Ca^{2+}-binding region, but also in distant parts of the structure upon Ca^{2+}binding [26]. With this feature in mind, it is not surprising that the capacity of IgE to bind parvalbumin is substantially reduced after Ca^{2+} depletion. Presumably, IgE bind to parvalbumin directly on the Ca^{2+}-binding sites or to an epitope located at a region that is affected by conformational changes, induced by Ca^{2+}. Three epitope regions have been identified on

parvalbumin, one of the epitopes being part of the Ca^{2+}binding domain [21].

2.4 Multiple Sequence Alignment - MSA

Sequence alignment is a way of arranging biomolecular sequences such as DNA, RNA, or AA-sequences, typically to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotides or amino acid residues are regularly represented as rows within a matrix. Gaps are inserted between the residues so that those with identical or similar characters are aligned in successive columns. The most widely used strategy to create an MSA is the progressive-alignment approach:

1. Calculated pairwise distances between the sequences
2. Constructed a guide tree from the distances
3. Gradually build up the alignment, following the order in the tree

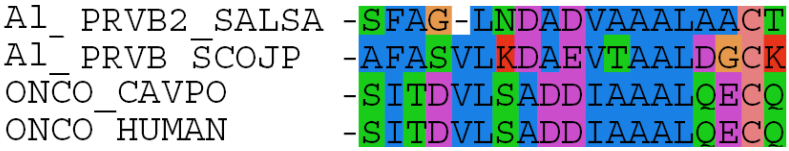

Figure 5: MSA of 20 starting AA of 2 allergen, which have Al_ first in there name, and 2 non-allergen parvalbumin sequences.

If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations, whereas gaps stem from *indels* (i.e. insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In AA-sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of the degree of conservation in a particular region or sequence motif among lineages. A MSA can reveal structures that are homologous i.e. characteristics shared by related species due to a common ancestor.

Alignment of multiple sequences is a fundamental step in the analysis of biological data. A MSA can reveal subtle similarities among large groups of proteins information that later can be used in several different ways.

2.5 Phylogenetic tree

Phylogeny is the evolution of species or higher taxonomic grouping of organisms, i.e. the history of organismal lineages as they change through time. Thus, a phylogenetic tree shows evolutionary relationship amongst various species. Each node with descendants indicates the most common ancestor of the descendants, and the branch lengths usually

corresponds to the number of changes that have occurred in that branch. There are many ways to represent such trees, e.g. a cladogram that displays the evolutionary propinquity of the displayed organisms, a phylogram that takes branch length into account, and radial that draws the tree as an unrooted tree radiating from a central point. An example tree is shown in figure 6.

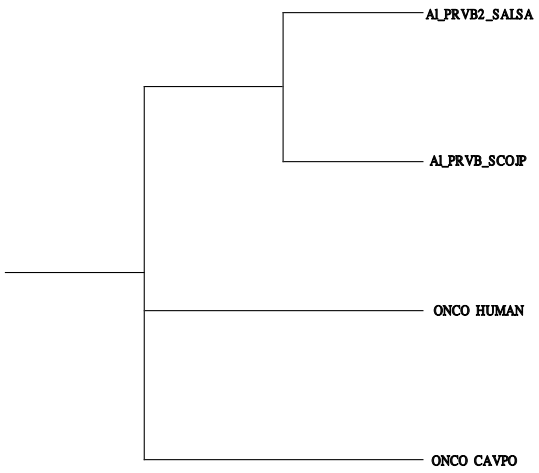

Figure 6: A rectangular cladogram plot of a simple tree created by Phyml of the 4 sequences from fig.5.

2.6 Analysis of Evolutionary co-variation

MSA of protein families can give a wealth of information, e.g. *conservation* and *correlation* (Coupling) of AA-position. AA-sequence conservation is related to the direct evolutionary pressure to retain physico-chemical characteristics of key positions in order to maintain a given function. In MSA sequence conservation is seen as the appearance of either the same or a functionally (roughly) equivalent AA in a particular position (column). AA-sequence correlation is attributed to the typically small sequence adjustments needed to maintain protein stability against constant mutational drift. Correlation or coupling refers to concerted changes of different positions in MSAs (co-variation).

In recent years, several reports describe that correlated mutations can provide information on protein structure information [27, 28, 29, 30]. A fundamental assumption behind correlated mutations or couplings is that if two columns in an MSA show high degree of correlation, the corresponding positions in that protein should be linked either energetically, functionally or by being physically close in some important conformation of the protein [27]. A main incentive for this study was to elucidate whether correlated pairs can give information about the proteins that are important for allergenicity, or that the pairs are directly linked to allergenicity.

2.6.1 Explicit Likelihood of Subset Co-variation - ELSC

Explicit likelihood of subset co-variation (ELSC), developed by Dekker *et al.* is an perturbation-based method for quantifying evolutionary co-variation (correlation) [27]. The perturbation method works by choosing subsets of sequences in an MSA, followed by comparing the AA of the subset with the AA of the full alignment. ELSC is, according to the authors, a refinement of another perturbation-based method called SCA (Statistical coupling analysis), reported by Lockless *et al.* [31]. ELSC allows for a more straightforward statistical interpretation of the resulting score values. In this study, ELSC has been extensively employed.

According to the authors ELSC seeks a score for a pair of columns in an MSA (i and j).

- A subset of the MSA is chosen where the subset is holding the n_{total}[1] sequences that have the AA that is conserved (most frequent) in pos i (interpreted from authors Java-code).

- The effects of the subset is then examined on each other position j.

- The observed AA composition of the subset at pos j is calculated.

- Then, given the AA composition at pos j in the full MSA, ELSC checks how many possible subsets of size n_{total} that would occur at pos j, exactly the observed composition of $n_{\text{ala},j}$ alanines, $n_{\text{asn},j}$ asparagines and all other AAs. The number of such subsets is given exactly by $\Omega_j^{<i>}$:

$$\Omega_j^{<i>} = \binom{N_{\text{ala},j}}{n_{\text{ala},j}} \cdot \binom{N_{\text{asn},j}}{n_{\text{asn},j}} \cdots = \prod_r \binom{N_{r,j}}{n_{r,j}} \tag{1}$$

  $N_{r,j}$ is the number of AA of type r at pos j in the full MSA and $n_{r,j}$ is the corresponding number for the subset. The combinatorial factor is given by eq.2

$$\binom{N_{r,j}}{n_{r,j}} = \frac{N_{r,j}!}{n_{r,j}!(N_{r,j} - n_{r,j})!} \tag{2}$$

  and is the number of ways to choose the exact number of sequences containing AA of type r in the subset ($n_{r,j}$) from the total number in the full MSA ($N_{r,j}$) Because every combinatorial factor in eq.1 is independent of each-other, the total number of possible subsets is simply given by the products of the factors.

- $\Omega_j^{<i>}$ is divided by the total number of possible subsets of size n_{total}, which gives the exact probability that a random selection of a subset of size n_{total} from the MSA will

[1] This is the notation used by the original authors where capital N describe properties of the full MSA and small n for subset

give the observed AA-composition at pos j in the subset. The probability is given by $L_j^{<i>}$:

$$L_j^{<i>} = \frac{\prod_r \binom{N_{r,j}}{n_{r,j}}}{\binom{N_{\text{total}}}{n_{\text{total}}}} \tag{3}$$

- ELSC calculates a normalized statistic that gives the probability of drawing the observed composition at random, relative to the probability of drawing the most likely composition. This is needed because MSAs and subsets will differ in size and combinatorial complexity. The normalization needs an ideally representative subset denoted $m_{r,j}$ created from a set of integers where $m_{r,j} \approx (\frac{N_{r,j}}{N_{\text{total}}}) \cdot n_{\text{total}}$. The author's implementation is by calculating the decimal value of $m_{r,j}$ and then rounding that to integer value with the constraint that $\sum_r m_{r,j} = \sum_r n_{r,j}$ so the subset is equal in size.

- The probability of drawing the subset $m_{r,j}$ from MSA at random is given by $L_{j,\max}^{<i>}$:

$$L_{j,\max}^{<i>} = \frac{\prod_r \binom{N_{r,j}}{m_{r,j}}}{\binom{N_{\text{total}}}{n_{\text{total}}}} \tag{4}$$

- The normalization is calculated by $\frac{L_j^{<i>}}{L_{j,\max}^{<i>}}$ and the authors denote it $\Lambda_j^{<i>}$:

$$\Lambda_j^{<i>} \equiv \frac{L_j^{<i>}}{L_{j,\max}^{<i>}} = \prod_r \frac{\binom{N_{r,j}}{n_{r,j}}}{\binom{N_{r,j}}{m_{r,j}}} = ELSC(i,j) \tag{5}$$

- The authors then takes $-\ln \Lambda_j^{<i>}$ just to be able to compare their ELSC score with the old SCA score. An overview of ELSC applied on a simple alignment can bee seen in Table 1.

ELSC discards gaps when counting sequences in the MSA. In ELSC there is a constrained relationship between i and j that it's always true that $j > i$ and co-variation is only calculated for that pair. In other words for columns 1 and 10 in a MSA ELSC only use the most conserved residue in column 1 to form the subset and report the score for the pair $(1, 10)$ but not for the pair $(10, 1)$. The JavaELSC implementation, provided by the authors, was used in this study.

2.7 WRABL - Groups of Amino Acid

Amino acids can be categorized according to features of importance to protein function and/or to evolutionarily relatedness. James O. Wrabl *et al.* has described a way of grouping AA types using variance maximization of the weighted residue frequencies in columns taken from a large alignment database [32]. In that work a range of such clusters was presented

AMF C W
A N G G W
A Q C A W
A G V Q W
C G A L W
C T A M M
C T A Y M
D T A M K
D T A M K
D T A L K
↑ ↑ ↑ ↑ ↑
i 1 2 3 4
$= j$

(a) MSA

r	N	n	m	$\binom{N}{n}$	$\binom{N}{m}$
A	0	0	0	1	2
C	0	0	0	1	1
D	0	0	0	1	1
E	0	0	0	1	1
F	0	0	0	1	1
G	0	0	0	1	1
H	0	0	0	1	1
I	0	0	0	1	1
K	3	0	1	1	3
L	0	0	0	1	1
M	2	0	1	1	2
N	0	0	0	1	1
P	0	0	0	1	1
Q	0	0	0	1	1
R	0	0	0	1	1
S	0	0	0	1	1
T	0	0	0	1	1
V	0	0	0	1	1
W	5	4	2	5	10
Y	0	0	0	1	1
	$N_{total} = 10$	$n_{total} = 4$		$\prod = 5$	$\prod = 60$

(b) ELSC details when $j = 4$

$$-\ln\left(\prod \frac{\binom{N}{n}}{\binom{N}{m}}\right) = 2.4849$$

(c) Result for $j = 4$

Table 1: Overview of ELSC applied on a simple alignment. Consider the two columns i and $j = 4$, ELSC first choose a subset in column i from in this case a hypothetical MSA fig.1(a) the subset is holding the 4 conserved Alanine (A) above the double horizontal line at column i. Next the degree of bias in the distributions of AA in column j is quantified in this subset. If substitutions at position i and j occur independently through the sequences sampled by the MSA, the distribution of AAs at position j in the subset should be similar to the distribution position j in the full MSA. If the two positions co-vary, the AAs at position j in the subset may be biased by the chosen subset in column i. 1(b) Detailed ELSC calculations of the given subset for column $j = 4$. Where r is the 20 different AA possible. N denotes number of AA of type r in the full MSA. Moreover n denotes the same but for the subset MSA. m is the count of AA of type r in the idealized MSA, created by calculating $m_r \approx (\frac{N_r}{N_{\text{total}}}) \cdot n_{\text{total}}$. The combinatorial term $\binom{N}{X}$ is calculated as stated in equation2. 1(c) The resulting ELSC score for pair $(i, j = 4)$, calculated by the $-\ln$ of equation 5.

and the one composed of 8 functional groups was identified as optimal. Hence, this sort of amino acid aggregation was selected to the study outlined in this work. The resulting 8 optimal groups correspond fairly well to AA physical properties. In this study the aggregation of the 20 letter AA alphabet to only 8 letters is denoted WRABL after the first author. The translation is as follows (Letters within parenthesis represent AA in their original form):

Aromatic W = (WFY)

Aliphatic M = (MLIV)

"Small" A = (ATS)

Polar/acidic N = (NDE)

Polar/basic H = (HQRK)

3 unique groups (C), (G), (P)

2.8 Protein structure prediction

Protein structure prediction involves computational techniques aiming at deriving 3D structures of proteins from their AA-sequences. 3D-protein structures can provide valuable information on protein function. In an allergenicity context knowledge on protein structure is important when considering if and where immunoglobulin E molecules are binding to proteins. The amount of experimentally verified structures available is, however limited because it is hard and very time-consuming to derive new structures by X-ray crystallography or nuclear magnetic resonance spectroscopy. This is where structure prediction comes in. Structure prediction *in silico* is fast and relatively inexpensive and can give good results in some cases.

Structure prediction can be divided into three areas: ab initio prediction, fold recognition, and homology modeling. Ab initio or de novo protein prediction methods are based on the laws of physics and chemistry to predict the structure of a protein, rather than using other proteins as templates. Fold recognition attempt to detect similarities between protein 3D structure that doesn't have any significant sequence similarity, i.e attempts to find folds that are compatible with a target sequence and predict how well a fold will fit. Homology modeling can, at the current stage of development, give the most accurate models and uses a single template from PDB that has a high level of sequence similarity to the target [33].

In this project SWISS-MODEL [34] by SIB[2] being of the homology modeling type, is used to predict structures of proteins with known AA-sequences. SWISS-MODEL is a freely available web-server application that can predict structures from templates. Results are sent as a PDB-file to a given e-mail address. The global SWISS-MODEL steps are:

1. Search for suitable templates in a 3D database
2. Check sequence identity with target
3. Generate models
4. Minimize energy

To verify outputs from the SWISS-MODEL, the 3D-JIGSAW [35] being another homology prediction web-tool was used. The modeling steps are similar to those of SWISS-MODEL.

3 Aims

The over-all aim of this degree study is to apply bioinformatics methods to identify and evaluate features that may separate allergen proteins from non-allergen proteins, belonging to the same family. To accomplish this, evolutionary co-variation/coupling analysis was applied to both allergens and non-allergens of two distinct families, tropomyosin and parvalbumin. Activities in this study were aimed at:

[2]Swiss Institute of Bioinformatics, http://www.isb-sib.ch

- Discovering possible differences in co-variation patterns between allergens and non-allergens. This is performed by applying the algorithm Explicit Likelihood of Subset Co-variation (ELSC see sec.2.6.1) to tropomyosins and parvalbumins.
- Testing the robustness of the ELSC algorithm regarding the number of sequences used in the analyse. This is performed by ELSC-sample-size-test, as described in section 4.5.3.
- To examine whether grouping of amino acids can reveal co-variation in positions across functional AA groups, which in turn may point out key positions as regards function/structure.
- Examining whether co-variation analysis may be used to retrieve information about allergens, such as identifying epitopes or other motifs important for allergenicity. This is carried out by comparing best resulting ELSC pairs from allergens with known epitopes.
- Examining if a homology structure prediction can be applied to detect allergen specific structure difference.

4 Materials and Methods

4.1 Datasets

A variety of allergy-dedicated databases, each holding a subset of AA-sequences, were consulted to create sets of both allergen and non-allergen sequences. Apart from the in-house database of the *National Food Administration* [2], the following repositories were mined: *Allergome* [20], *SDAP* [36], *UniProt* [37] Excerpts from the various datasets were complied into text files and formatted according to the standard FASTA format [38]. For clarity, allergen sequences are tagged with "Al_" upstream of the actual name.

4.1.1 Tropomyosin

One of the composite tropomyosin data-sets, created for this project, contains 106 presumed non-allergens and 23 allergen tropomyosins. This family was considered as particularly appropriate for this study because both allergen and non-allergen tropomyosin AA-sequence are known and the protein family displays high sequence conservation. Allergen proteins are showed in table 2.

4.1.2 Parvalbumin

The parvalbumin data-set used in this study holds 16 non-allergens (all being mammalian parvalbumins mined from UniProt) and 13 allergen sequences, as listed in table 3. This is

UniProt-Entry	Protein information
TPM4_DROME	Isoforms 33/34 (Tropomyosin II) -*Drosophila melanogaster* (Fruit fly)
TPM2_DROME	(Tropomyosin I) - *Drosophila melanogaster* (Fruit fly)
Q2WBI0_9ACAR	*Dermanyssus gallinae* (Chicken mite)
TPM_CHAFE	Allergen Cha f 1 (Fragment) - *Charybdis feriatus* (Crab) *see fig 3(c)*
TPM1_DROME	Isoforms 9A/A/B (Tropomyosin II) (Cytoskeletal tropomyosin) - *Drosophila melanogaster*
Q3Y8M6_9EUCA	Pen a 1 allergen - *Farfantepenaeus aztecus* (brown shrimp).
TPM_ANISI	(Allergen Ani s 3) - *Anisakis simplex* (Herring worm).
TPM_PERAM	(Major allergen Per a 7) - *Periplaneta americana* (American cockroach).
TPM_BLAGE	*Blattella germanica* (German cockroach).
TPM_LEPDS	(Allergen Lep d 10) - *Lepidoglyphus destructor* (Fodder mite).
TPM_HALDV	*Haliotis diversicolor* (Abalone).
TPM_PERVI	*Perna viridis* (Tropical green mussel).
TPM_MIMNO	*Mimachlamys nobilis* (Noble scallop) (Chlamys nobilis).
Q95WY0_CRAGI	(Fragment) - *Crassostrea gigas* (Pacific oyster).
TPM_PERFU	*Periplaneta fuliginosa* (Smokybrown cockroach) (Dusky-brown cockroach).
TPM_LEPSA	*Lepisma saccharina* (Silverfish) *see fig 3(a).*
TPM_METEN	(Allergen Met e 1) (Met e I) - (Greasyback shrimp) (Sand shrimp) *see fig 3(b).*
TPM_HELAS	(Allergen Hel as 1) - *Helix aspersa* (Brown garden snail).
TPM_CHIKI	(Allergen Chi k 10) - *Chironomus kiiensis* (Midge).
TPM_PANST	(Allergen Pan s 1) (Pan s I) - *Panulirus stimpsoni* (Spiny lobster).
TPM_HOMAM	(Allergen Hom a 1) - *Homarus americanus* (American lobster).
TPM_DERPT	(Allergen Der p 10) - *Dermatophagoides pteronyssinus* (House-dust mite).
TPM_TURCO	(Major allergen Tur c 1) (Fragments) - *Turbo cornutus* (Horned turban) (Battilus cornutus).

Table 2: Allergens in the tropomyosin protein family.

a considerably smaller data-set, relative to that of tropomyosins (29 sequences versus 129), but with a higher ratio between allergens and non-allergens. Moreover, allergens and non-allergens are not bifurcated into phylogenetically distinct categories like the tropomyosins of which all known allergens stem from invertebrate organisms. This makes parvalbumins a better candidate set to spot differences between the sequences which are attributed to allergenicity without relation to phylogeny.

UniProt-Entry	Protein information
PRVB_THECH	*beta* (Allergen The c 1) - *Theragra chalcogramma* (Alaska pollock).
Q90YK8_THECH	*Theragra chalcogramma* (Alaska pollock).
PRVA_RANES	*alpha* - *Rana esculenta* (Edible frog).
Q8JIU1_RANES	*beta* protein - *Rana esculenta* (Edible frog).
Q8UUS3_CYPCA, Q8UUS2_CYPCA	*beta Cyprinus carpio* (Common carp).
PRVB_GADCA	*beta* (Allergen Gad c 1) (Allergen M) - *Gadus callarias* (Baltic cod).
90YL0_GADMO	*beta* - *Gadus morhua* (Atlantic cod).

Table 3: Allergens in the parvalbumin protein family.

4.2 Bioinformatic methods

4.2.1 Kalign

In this study a rather new MSA methods was used. The Kalign algorithm is accurate and fast and is based on a strategy similar to that of the standard progressive method for

sequence alignment [39]. Kalign enhance this method by taking advantage of an approximate string-matching algorithm, that allows string matching with mismatch for distance calculation and by incorporating local matches into the otherwise global alignment. This renders Kalign estimates of distance more accurate and with a throughput not inferior to other leading methods, such as ClustalW [40], Muscle [41] or T-Coffe [42]. Kalign is as accurate as the best among other methods on small alignments, but significantly more accurate when aligning large and distantly related sets of sequences [39].

4.2.2 Phylogenies by Maximum Likelihood - Phyml

Phyml is a fast and accurate maximum likelihood algorithm to estimate phylogenies. The core is a simple hill-climbing algorithm that adjusts tree topology and branch lengths simultaneously [43]. Phyml starts by creating an evolutionary distance matrix from the sequences, by a fast distance-based method. An initial tree is built from this matrix, using the BIONJ [44] algorithm. The Phyml algorithm then modifies, at each iteration, this tree to improve the probability of observing the sequences available under an underlying statistical model that depends on the tree structure (tree parameters). Phyml search for the most likely tree (thus maximum likelihood). Phyml reaches optimum after a few iterations due to the simultaneously approach. It is a freely available program that was used in this study to create phylogenies of the protein families to visualize and detect relationships between the protein sequences.

4.3 Creation of MSAs and Phylogenetic trees

Kalign was used to align all the dataset FASTA files. Regularly the default parameters of 6.0 for gap open penalty and 0.9 on gap extension penalty was used. Other settings are clearly stated. Manual inspection of the computed MSA was performed to identify whether improvements could readily be made prior to further processing. Mostly Kalignvu [45] was used; it is a web tool for visualizing and running Kalign on given MSAs. The aligned datasets were then loaded into Matlab for further analyze and testing.

Phyml was performed on the kaligned subsets; the following parameter settings was used:

- Model of amino acids substitution : DCMut [46]
- Initial tree : [BIONJ] [44]
- Discrete gamma model : Yes
  -Number of categories : 4
  -Estimate Gamma shape parameter : YES (2.011 for parvalbumin and 1.003 for tropomyosin)
- Estimate proportion of invariant: YES (0.121 for parvalbumin and 0.000 for tropomyosin)

4.4 Computer aid

Most of the analyzes and algorithm development was performed in the MATLAB [47] programming environment. Several special scripts were, however, created in Perl and Bash. Most computer calculations where performed on one PC (AMD64 dual core 2010 MHz with 2GB RAM) with Gentoo Linux X86_64 with kernel 2.6.16-gentoo-r9 [48].

4.4.1 Computer Cluster

Due to demanding computations required for 3D-structures and 3D-structure comparison and other heavy computations conceivable within the project, the intention was originally to construct a computer cluster. Several different cluster softwares where considered and two of them, both being of the load-balancing kind, were tested. They are designated OpenSSI [49] and openMosix [50]. The *National Food Administration* provided 5 AMD64 dual-core computers equipped with latest hardware features, that later were found to be both a benefit and a disadvantage.

Firstly, an OpenSSI environment was implemented on a single computer. OpenSSI is a a kernel extension to Fedora Core 3, Debian Sarge or Red Hat 9 (three different Linux distributions). Debian was chosen because of the similarity to Gentoo Linux [48] a preferred distribution because of its unique adaptability. OpenSSI is, at the time of this work, only stable with a 2.4 kernel, which caused the main problem. The new hardware, such as SATA-II hard-disc-drives and controller, are not well supported in the old 2.4 kernel. Despite the hardware/software compatibility problems a base system with the OpenSSI kernel extension was installed. Work with graphics needs an X-windows system, and the installation of X-windows didn't, however, work with the OpenSSI/Debian combination. A Gentoo Linux base system was then chosen as cluster base system to extend its kernel with openMosix, which is a Linux kernel extension for single-system image clustering. The kernel extension can turn a network of ordinary computers into a supercomputer. openMosix is balancing the workload even on the different nodes in the cluster and continuously attempts to optimize the allocation of resources. The main advantage of this type of cluster is that there is no need to program an application to run on openMosix, in contrast to a cluster that needs programs that are implemented in a parallel fashion to use the power of the cluster. Because of the process migration openMosix is also appropriate if the required computations are based on many different processes rather than one time consuming algorithm. The cluster behaves much like a symmetric Multi-Processor which is a multiprocessor computer architecture where two or more identical processors are connected to a single shared main memory. The problems that aroused with Gentoo Linux and openMosix kernel were of the same kind as those of OppenSSi and Debian i.e. openMosix was just stable with kernel 2.4. A workaround was created, which enable installation of a fully functional X-windows system with the only drawback that full support for the graphics-card wasn't possible. Several tests were performed on the cluster, such as existing stress-test and some newly self made tests. The openMosix cluster was clearly operational but unfortunately somewhat unstable and several unpredictable crashes oc-

curred. Moreover the slow behavior of the graphics drivers wasn't satisfactory. Another main drawback is that none of the tested clustering softwares supports 64-bits processors. Because of the instability and the poor graphics performance, the Cluster installation was rejected. A plain Gentoo AMD64 version with the latest kernel was subsequently installed on all five computers.

The processor power of all computers was subsequently made conveniently available by installation of the distcc [12] was installed. Distcc helps to compile new software for the computers by dividing the code to the different nodes. Accordingly programs were run on different nodes manually.

4.5 Procedures

4.5.1 20/80-method

One method that was evaluated at the outset of the study is denoted 20/80. 20/80 is pinpointing the columns in the MSA that have a degree of conservation over 20% and below 80%. The lower limit was set to ignore columns devoid of conservation, whereas the upper limit was set to avoid perfect conservation. Columns with great conservation are important to structure and may serve the purpose of separation of allergens from non-allergens. The MSA was divided in to two sets; one containing allergens and the other non-allergens. The subsets will, however, still have the same alignment, in that way no change in "coordinates" are made and the columns can easily be compared between the sets. The main idea with 20/80 is to separate the sets based on values in the different 20/80-columns, and if the sets have the same columns, the saved conserved AA can be compared, and if the conserved AA differs between the sets we have still a way to separate allergens from non-allergens.

A simple Matlab script was created that could load a MSA and then pin-point the 20/80-columns in the two subsets separately and eventually compute the number of columns that where unique for the separate set and how many they share.

The 20/80-method can be regarded as a simple preface to ELSC, but without the co-variation analysis because that every 20/80 column exists in on or more $ELSC(i, j)$. Since ELSC was discovered and fully incorporated results obtained with the 20/80-method were not further analyzed.

4.5.2 ELSC

Several Matlab scripts were created to load and run the java implementation of ELSC within Matlab. Some Perl-scripts were also produced to help seamless integration of javaELSC in Matlab. The script mostly used to examine ELSC-scores on different MSAs includes the following steps:

1. Load Kalign-created MSA file in FASTA-format
2. Run Perl-script to format MSA-file for javaELSC and to create allergen and non-allergen subset.

3. Run Perl-script to delete sequence name and import subset MSAs to Matlab matrix.
4. Save the columns in MSA that have a conservation ratio above 90%.
5. Run javaELSC on the two subsets. And import the outputted score matrix.
6. Plot the 20 $ELSC(i,j)$ (ELSC-pairs) with highest score.
7. Plot the $ELSC(i,j)$ that have a score that are within a given percent of the max ELSC-score of that subset. (20% mostly used)

The aforementioned procedure were performed on both tropomyosin and parvalbumin datasets.

4.5.3 ELSC - sample-size-test

Due to considerable fewer allergens in the tropomyosin dataset, relative to non-allergens, a test to examine sensitivity of ELSC to the number of sequences was performed. Several subset MSAs were created from the original non-allergen MSA, containing 23 sequences (the same numbers as the numbers of allergens in the original MSA).

Hypothesis: If ELSC is not sensitive to the number of sequences, ELSC-pairs with the highest scores are the same across subset-MSAs. To test the hypotheses the following algorithm was created:

1. random generate a ELSC-INPUT-file with 23 unique non-allergen sequences.
2. run javaELSC with random INPUT-file.
3. import ELSC-OUTPUT to Matlab
4. save ELSC-score-matrix
5. do step 1-4 50 times
6. calculate statistics based on all ELSC score-matrices

A help script was created to calculate the statistics on ELSC-score matrices. Every ELSC run generates an ELSC-score matrix with pairs i and j and their corresponding scores. All unique pairs obtained are firstly summarized, and thereafter the mean score value over all ELSC-runs and the number of times each pair is present in all performed ELSC runs, is calculated. These statistics were later used to visualize the test result (see fig.9).

4.5.4 WRABL

As outlined in the Background section 2.7 (WRABL - Groups of Amino Acid), AAs can be grouped into eight functional categories [32]. A simple Perl-script was created to aid the ELSC analysis on WRABL-translated MSAs. The procedure was executed as follows:

1. Translate a MSA with a Perl-script from AAs into WRABL groups
2. Run javaELSC on the translated MSA
3. Import ELSC-score matrix to Matlab for the same analysis as mentioned in section 4.5.2

5 Results

5.1 Phylogenetic trees

Figure 7 shows a circular phylogenetic tree of parvalbumins. One of the two major branches holds most of the allergens, whereas a minor fraction (two proteins) appears on a relatively distant location that derives from the second major branch. The two separated allergens are branched together in between several non-allergen sequences. The small cluster represents allergen parvalbumins that belong to the α-lineage, whereas the large cluster holds all sequences that are either designated β or without specific lineage designation. This indicates that all non-designated sequences may belong to the β-linage.

To summarize, no clear phylogenetic clustering between allergens and non-allergens can be spotted, but the allergens appear in two distinct clusters.

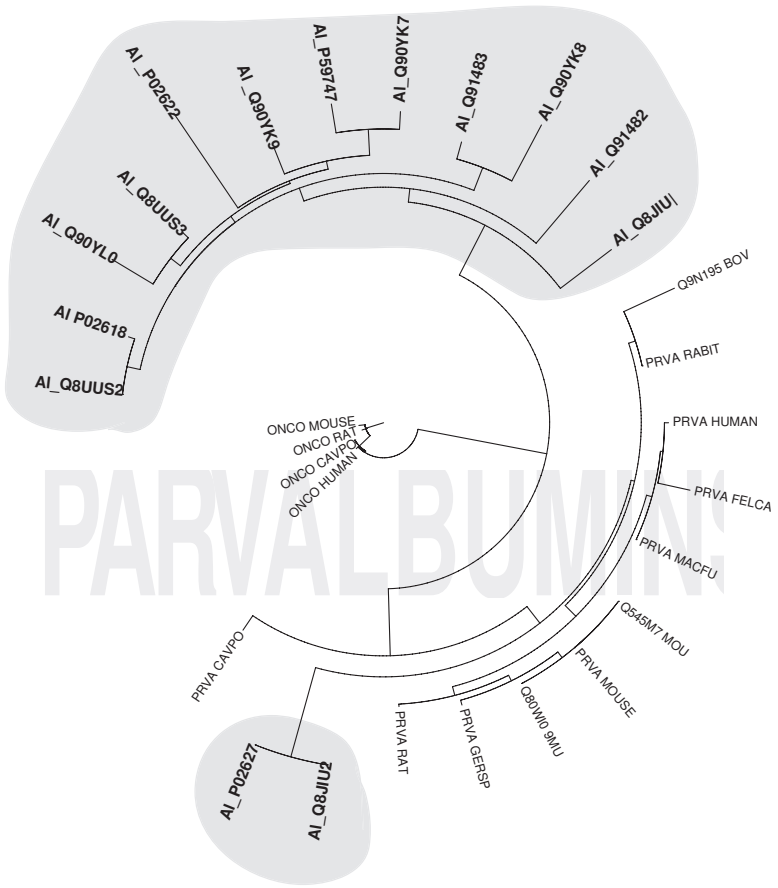

Figure 7: Circular phylogenetic tree of Parvalbumin with real branch lengths, created by Phyml. The gray shading indicates allergens, occurring in two separate groups.

Figure 8 shows a circular phylogenetic tree of the tropomyosin dataset. Since some distances between sequences are quite extended, several shorter branches become compact

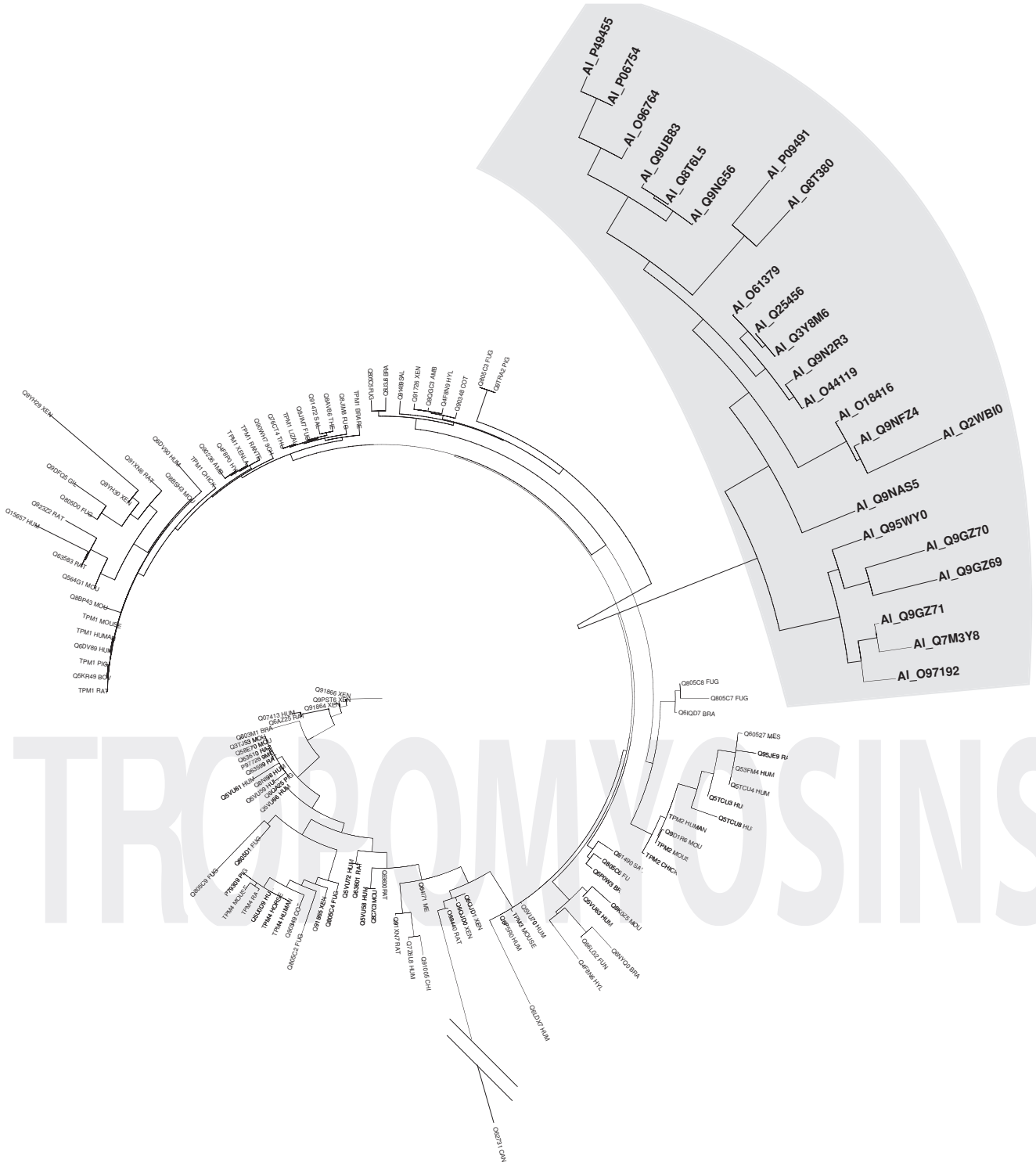

Figure 8: Circular phylogenetic tree of Tropomyosin with real branch lengths, created by Phyml. The gray shading indicates allergens.

and the nodes thereby difficult to spot. All tropomyosins allergens are located on a separate branch, with an appreciable extension from other tropomyosins. Although allergens are on a single branch the phylogenetic distance between them is relatively wide.

5.2 20/80-method

Both protein families where tested; below is a raw dump from Matlab:

```
Tropomyosin
For file /home/jonas/MatlabWork/Alignments/Tropomyosins_New_add_PenA1+.out

The following units have been calculated:
The MSA is holding a total of 129 sequences with a aligned length of 684
23 of them are Allergen and 106 is Non Allergen
126 Conserved 20/80-columns in Allergen and 139 in Non-Allergen-set

Equal sites :                                          77
Unique Conserved columns in Allergen subset :          49
Unique Conserved columns in NON Allergen subset :      62
Equal Conserved columns with equal conserved AminoAcid :     23
Equal Conserved columns with Non-equal conserved AminoAcid : 54

Total different Columns :                              165

************************************************************

Parvalbumin
For file /home/jonas/matlabWork/Alignments/Parvalbumins.out

The following units have been calculated:
The MSA is holding a total of 29 sequences with a aligned length of 114
13 of them are Allergen and 16 is Non Allergen
58 Conserved sites in Allergen and 62 sites in Non-Allergen-set

Equal sites :                                          43
Unique Conserved columns in Allergen subset :          15
Unique Conserved columns in NON Allergen subset :      19
Equal  Conserved columns with equal conserved AminoAcid :     10
Equal  Conserved columns with Non-equal conserved AminoAcid : 33

Total different Columns :                              67
```

These screen dumps illustrate how the different data-sets are organized. Since it turned out that all obtained 20/80-columns did also appear among the best-ranked correlated pairs (column i and j) , as calculated by ELSC, focus was moved to the latter algorithm.

5.3 ELSC

A variety of graphical representations were made to promote visualization and analysis of $ELSC(i, j)$ (ELSC pairs) and the corresponding scores. A particularly suitable plot-type for this purpose is the one showing $ELSC(i, j)$ as points in two dimensions. High scores reflects strong correlation and pairs with high scores are likely to actually reveal information on pertinent structure or function. To avoid blurred images many pairs of presumed low relevance were excluded from accordingly created plots. Hence, in the majority of the plots only pairs with a score within 20% of the max score are shown.

5.3.1 ELSC - sample-size-test

ELSC -sample-size-tests were performed on the non-allergen tropomyosin subset. Results were visualized as a contour plot. A contour displays isolines of a matrix (X,Y,Z) where Z

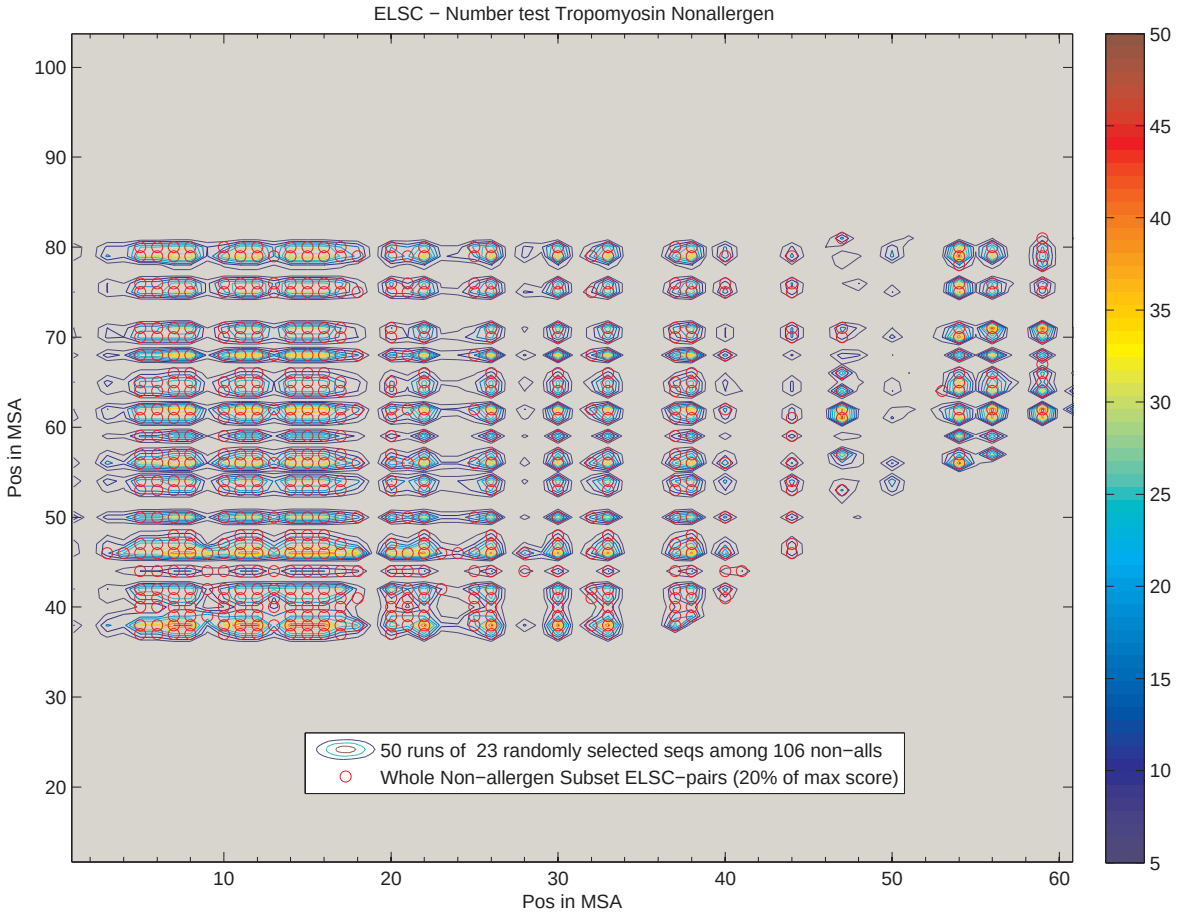

Figure 9: Part of a Matlab contour plot of ELSC-sample-size-test performed on the non-allergen tropomyosin dataset to study the stability against variation in the number of samples.

is interpreted as heights with respect to the $x-y$ plane. Fig.9 shows a typical contour and a "ring"- plot of non-allergen tropomyosins. In Fig.9, X, Y-axis correspond to indicated positions in the MSA. ELSC pairs (i, j) that have a score value within 20% of the max ELSC score of that subset, are indicated as red circles. The height Z of every point i, j in the contour corresponds to the statistics of the ELSC-sample-size-test, the height is the number of times that pair (i, j) have an ELSC value that are within 20% of the max ELSC score of all 50 ELSC runs.

Figure 9 clearly shows that even if only a few of the sequences are chosen, the pairs (i, j) with highest scores correspond to the whole subset. The contours are clearly aggregating around the red-circles, suggesting that any subset of sequences would give an output that essentially concurs with the entire set.

5.3.2 Tropomyosin

Figure 10 shows the result of two representative ELSC runs, where a pair (i, j) from the allergen subset having ELSC-score within 20% of the maximum ELSC-score is indicated

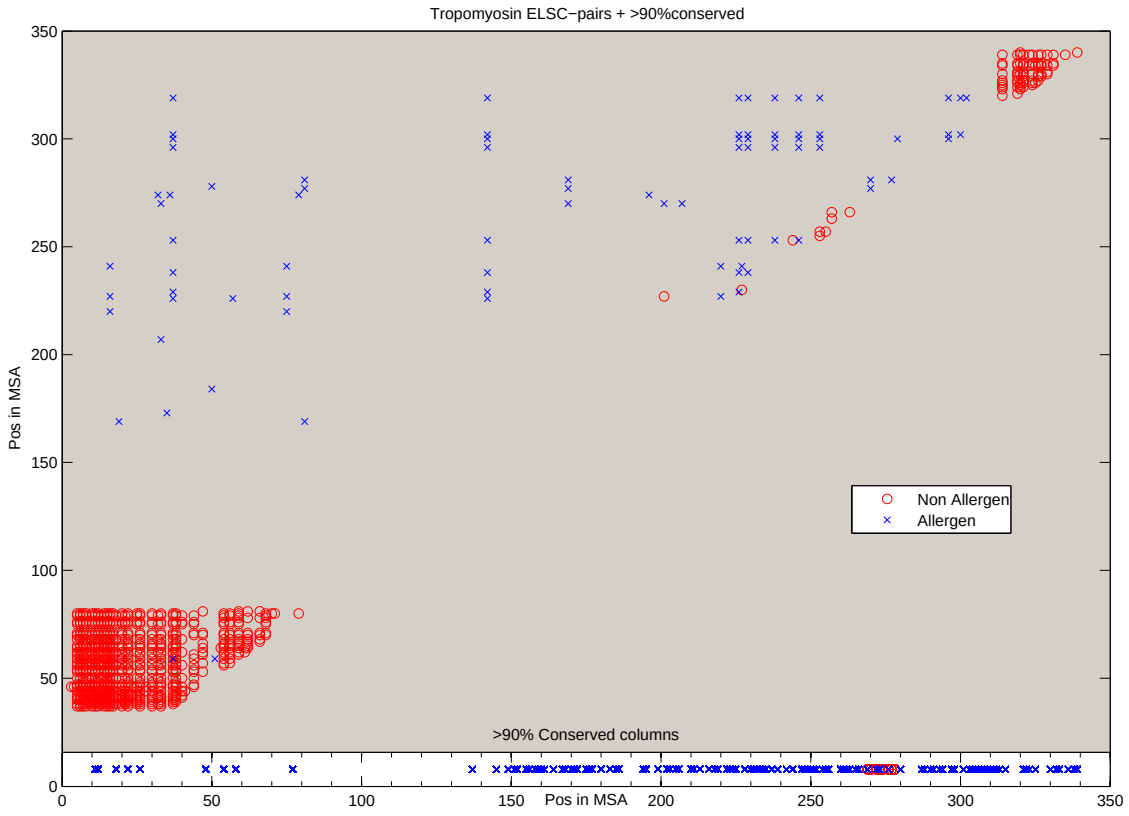

Figure 10: An i, j-plot of ELSC pairs from the two tropomyosin subset, plus a subplot that shows the columns that have a conservation larger than 90%, Columns are indecated as blue X for the allergen subset and Red circles for the non-allergen subset. Clearly the allergen subset is more conserved.

as a blue X. Analogously, a red ring (O) indicates such a pair but for the non-allergen subset. Clearly there is a difference between the distributions of the two subsets of strong ELSC-pairs. Notably, those of non-allergens appear as two big clusters in the amino- and carboxy-termini of the proteins, whereas the allergen counterparts are fewer and scattered. ELSC-pair clusters of non-allergens suggest that the head and tail areas of the coild-coil tropomyosin are functionally important. There are, however, barely any allergen ELSC-pairs in the head-tail areas. Allergen tropomyosins show, though, high degree of conservation in the carboxy terminus region. This pattern extends more than halfway across the protein.

With the aid of an experimental scanning approach, including IgE binding/peptide competition assay, Reese *et al.* has identified several epitopes of an allergen[3] tropomyosin [17]. To examine whether the distribution of ELSC pairs in any way coincides with that of epitopes, the latter were translated to MSA-space, and plotted as ribbons in Figure 11.

[3]Pen a 1

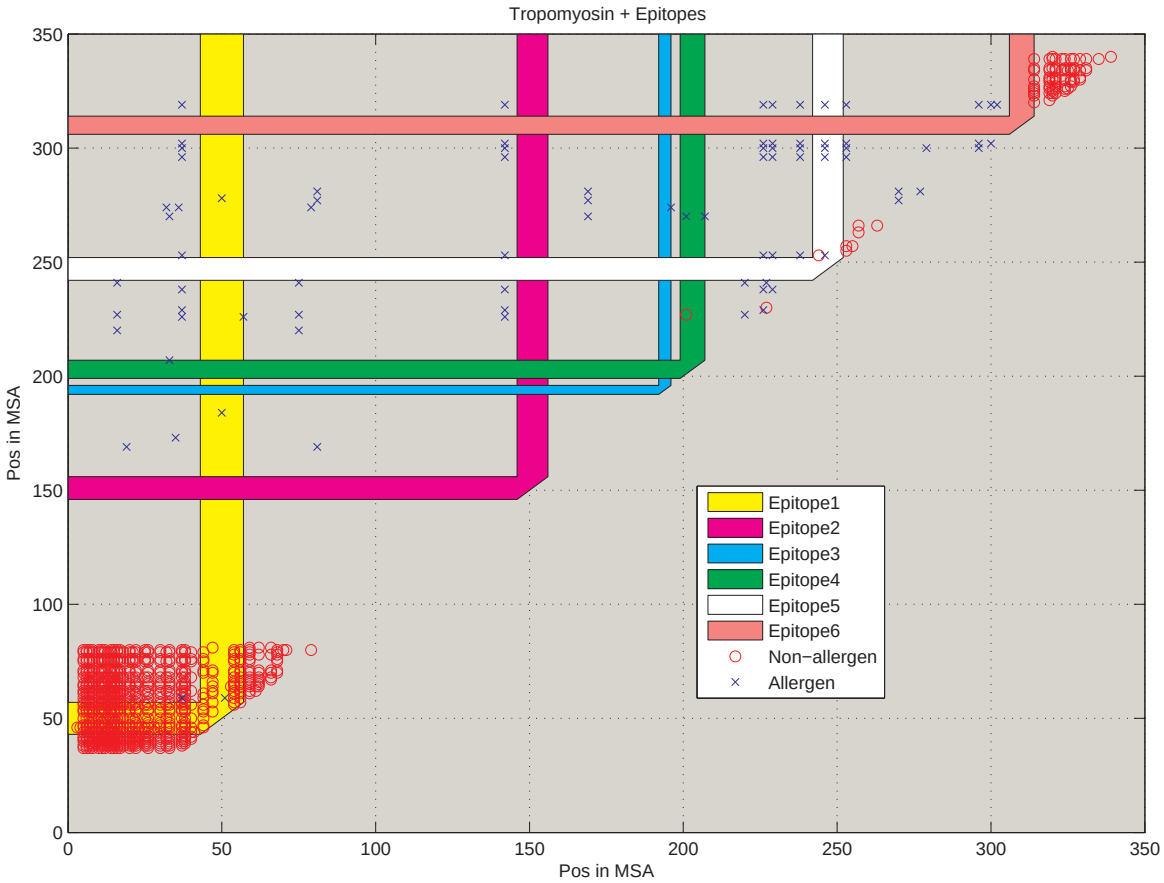

Figure 11: Epitope-ribbon plot of tropomyosin.

Each epitope has a given extension in the AA-sequence and the corresponding ribbon-area in the plot visualize pairs that have any column i or j within that area of the protein. Accordingly if any pair of AA-positions (O or X) lies within a ribbon, either one or both positions is in the epitope area. According to Figure. 11 the allergen ELSC-pairs seem to occur well outside the epitope representations, but some coupled pairs appear within and proximal to the ribbons. The most interesting areas are those, wherein two ribbons are overlaid, since a pair positioned within a ribbon-overlay indicates that it corresponds to correlated positions in two separate epitopes. Such a match is not found in Figure. 11, but many allergen pairs are located near several such ribbon-overlays. This pattern is especially prominent around the area shared by epitopes 5 and 6.

5.3.3 Parvalbumin

Figure 12 shows the distribution of ELSC-pairs in an epitope-ribbon plot of the parvalbumin family. No distinct groups of coupled pairs can be spotted and, additionally, it is hard

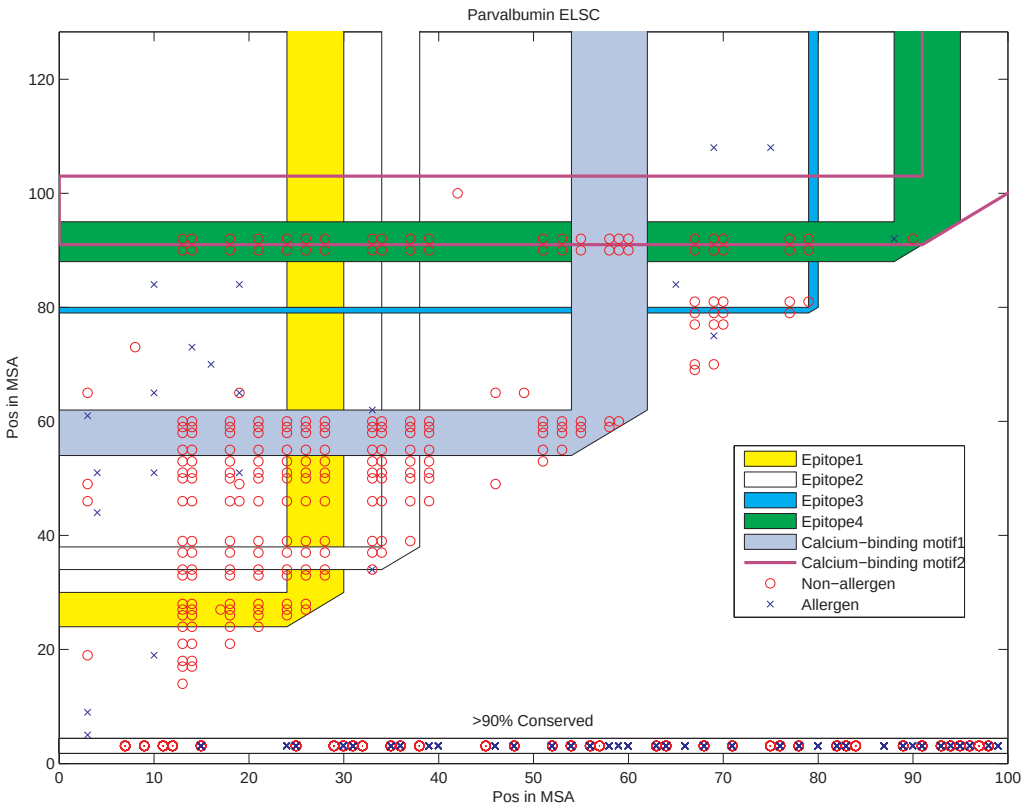

Figure 12: A i, j-plot of ELSC pairs from the two parvalbumin subsets, with Epitope-Ribbon and one Ca^{2+}-binding motif as a ribbon and the other as a bold line.

to see any clear separation between the two datasets. The allergen subset has low ELSC-score values. In order to enhance detection of allergen pairs, 22% of max-score was set as boundary to the allergen subset. The two subsets have two pairs with a perfect match. Apart from the allergen epitopes the two Ca^{2+}-binding-motifs are also shown as ribbons (Fig. 12). No clear co-occurrence between epitope ribbons and allergen ELSC-pairs is seen. A rather clear overlap between nonallergen ELSC pairs and the Ca^{2+}-binding motifs is, however, evident in the plot.

5.4 ELSC + WRABL

The WRABL-translation procedure was conducted on both of the major protein datasets used in this study. ELSC results of the tropomyosins MSA based on standard AA representation, as well as that of WRABL-encoded AA are shown in Figure 13. The plot, being of the same type as shown earlier, depicts WRABL-ELSC pairs plotted in green color. Among the top 100 non-allergen $ELSC(i, j)$-pairs based on regular AA representation and

the top 100 $ELSC(i,j)$-pairs using the WRABL translation, 35 pairs overlapped. Analogously, in the allergen set, the number of overlapping best pairs were 15. Thus, among the 100 highest ELSC scores from analysis of both standard AA MSA and that based on WRABL translation the ratio of overlap was 35% and 15 % for non-allergens and allergens, respectively. The corresponding results for parvalbumin dataset are similar, (data not shown).

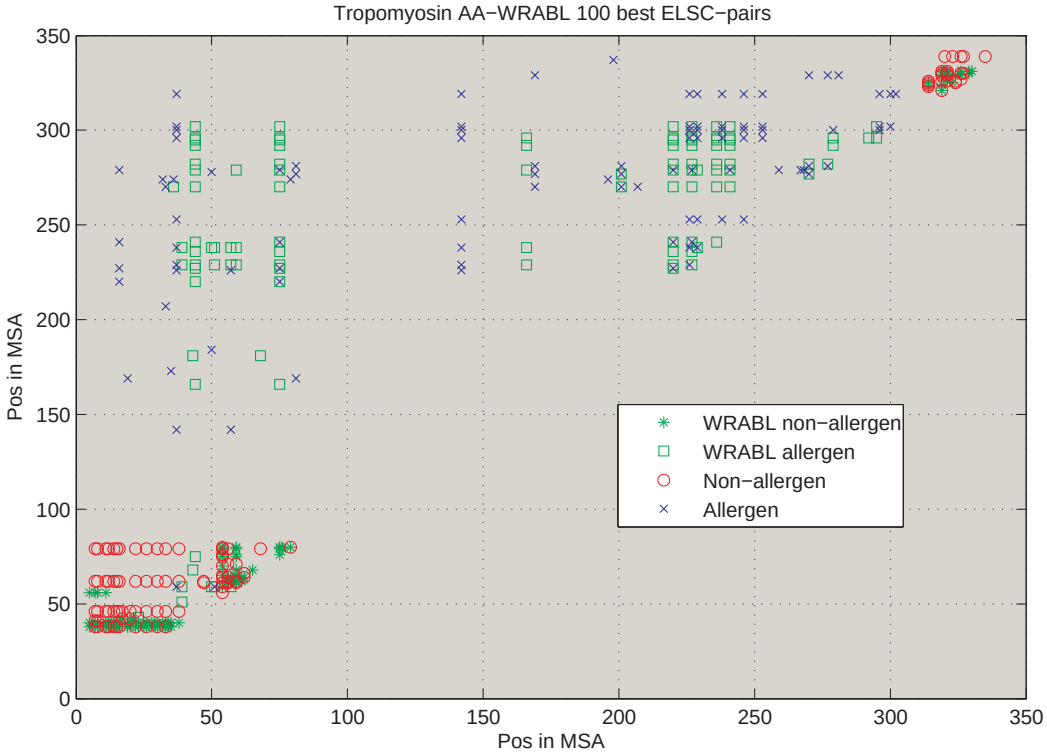

Figure 13: Ordinary ELSC and WRABL translation, 100 best scored ELSC-pairs.

5.5 Structure prediction of mutant Pen a 1

Gerald Reese *et.al.* has in a study created a mutant variant of the major shrimp allergen Pen a 1 whose allergic potency thereby was reduced by 90-98% [19]. Based on this and other tropomyosin sequence data, a goal in this study was to make two structure predictions, one of the original Pen a 1 and one of the mutated variant and then compare the two structures to see if any obvious differences would be revealed. SWISS-MODEL was able to predict a single stranded alpha-helix structure from the sequences, but the underlying template was the same for the two sequences. No significant structural differences in the two models are apparent. The 3D-JIGSAW predictions for the same sequences was almost identical to those of SWISS-MODEL and revealed only positional discrepancy where the mutation was applied, i.e. the exact location of AA dissimilarity.

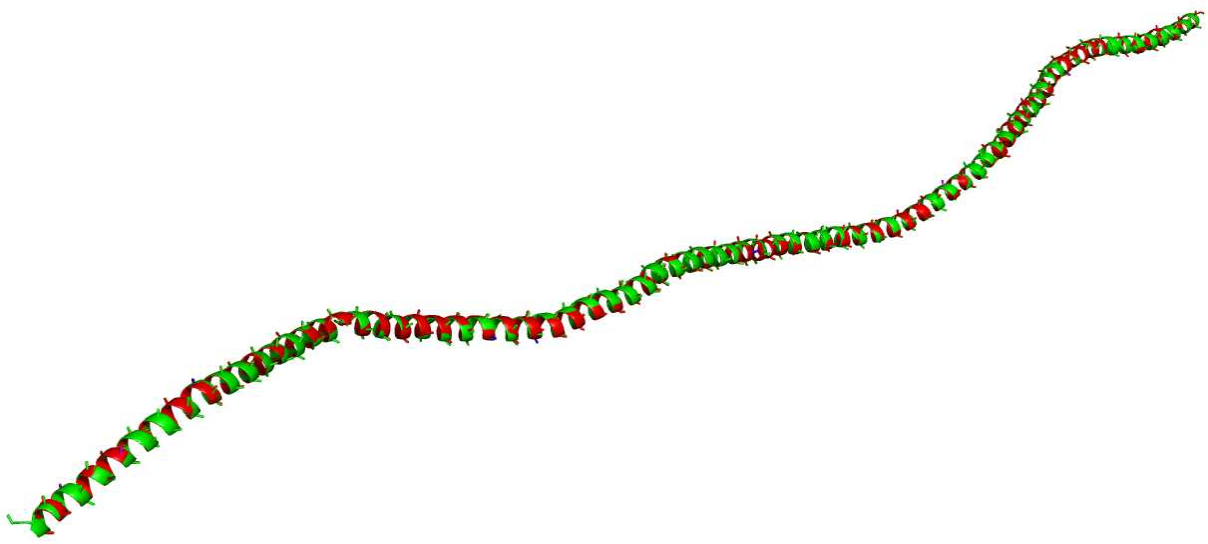

Figure 14: Pymol superimposed render of SWISS-MODEL prediction of Pen a 1 (RED) and Mutant VR9-1 (GREEN), The color Magenta indicates mutated position in the VR9-1 structure and the blue is the same position for Pen a 1.

6 Discussion

6.1 Robustness of ELSC

The ELSC-sample-size-test clearly prove that ELSC still give reliable pairs even if only 23 sequences of the non-allergen tropomyosin dataset were used. This moreover indicates that the allergen subset holds enough data for a reliable ELSC analysis. According to a personal communication with Anthony Fodor, a major developer of the ELSC-algorithm, ELSC can provide good results with as few as 15-20 sequences. On the other hand, several hundred sequences may give spurious outputs. The outcome is dependent on the phylogenetic relationship of sequences selected to the analysis. A dataset holding hundreds of sequences, that share high sequence similarity, is not as good as tens of sequences that cover a wide phylogenetic space. Although Fodor hasn't looked at the relationship between large number of sequences and performance of ELSC extensively, he has not yet seen any clear-cut correlation. The possibility that diversity is more important than size of the dataset advises the user to remove almost identical sequences from the alignment. This gives further strength to the ELSC approach with regards to the allergen subset, while it holds more sequences than the minimal number suggested by Fodor and since the phylogenetic analysis conducted on the tropomyosin dataset shows a wide span in the allergen branch. Also the non-allergen set, encompassing 106 sequences, has a relatively wide phylogenetic span. In compliance with the above guidance, clustering of tropomyosins sequences with 90% identity was performed, in addition to analyses based on the entire sets of AA sequences. In total, 49 clusters were created and all sequences, except for the representative sequence of each cluster, were deleted. Analysis by ELSC was conducted as described above. This

operation didn't appreciably change the identification of co-varying pairs, but the ELSC scores were more evenly distributed in the allergen subset, and more pairs appeard whitin the 20% max-score border. Many sequences were deleted from the non-allergen dataset due to many clusters in that set. Still, the pairs were generally positioned on the same spots. The conlusion is that filtering by percentage identity seems to be a good idea. Additional testing is, however, needed to draw firm conclusions.

ELSC can be very helpful to give positions that perhaps are important for allergenicity to test experimentally by site mutagenesis, to see if allergenicity is altered in one protein after a change in a AA in one or two position given by ELSC.

6.2 *In silico* analysis of tropomyosins

6.2.1 Phylogeny

As seen in figure 8 the tropomyosin dataset spans a wide phylogenetic space. Even though all allergens branch them self together under one node there is still quite an evolutionary distance between them. This suggests that ELSC should give positions that actually are relevant.

6.2.2 ELSC

As evident from figure 10 the allergen ELSC-pairs appear at a greater distance from the diagonal of the plot. Intuitively, co-variation involving proximal amino acids seems more likely than remote counterparts to signify a conformational connection, but folded proteins can bring together regions that are far apart in a linear amino acid sequence representation. Tropomyosins, though, consistently attain a rod like shape, which would rather favor a preference for proximal amino acids as regards functional relevance. Nonetheless, these proteins are inter-twined α-helix dimers that may confer special requirements on the overall dimer configuration, leading to possible dependencies also over considerable distances.

This suggests that, on the assumption that co-varying AA-positions in tropomyosins are functionally important, the connection between the allergen columns (i, j) indicate a different functional association, relative to that of non-allergens, whose correlation pattern mostly involves physically close positions. There are, however, also several co-varying positions among allergens on or close to the diagonal in the plot, where the i and j in a pair are close to each other. It is also important to keep in mind that the allergen subset is much smaller than that of non-allergens, which influences on the ELSC scores. The max score of the allergen subset is roughly 22% of that of the nonallergen subset ($\frac{12.7949}{58.5115}$). The two clusters of pairs with high ELSC-scores, present in the head-tail areas of the protein, justifies the assumption that these segments are important, presumably to the interaction with actin. This pattern was, however, not seen among allergen tropomyosins. This finding may appear unexpected, but the allergen subset is very conserved (>90%) in these regions, particularty in the carboxy terminus, and highly conserved columns give low ELSC-scores.

Hence the lack of allergen pairs in those areas should be seen in the content of conserved AA positions in the same regions.

In figure 11 ribbon segments indicate experimentally verified IgE epitopes in allergen tropomyosins. This is to depict whether ELSC-pairs (Circles or X-marks) fall outside or within such stretches. The ELSC pairs that are located near ribbons may tentatively indicate that sites just outside some of the epitopes are important for the maintenance of structural integrity of those immunoglobulin E-binding sites, or segments otherwise related to allergenicity. It is important to realize that epitope data are derived from one allergen only, but the ELSC pairs are constructed from all the entire set of 23 allergens. Accordingly, some discrepancies between the distribution of epitopes may occur across allergen tropomyosins. Nevertheless, extensive IgE cross-reactivity suggests modest differences in this respect. The spotted ELSC-pair locations in the vicinity of certain epitopes pairs suggest that changes of co-varying AA pairs in such regions may confer drastic modification of epitope structure. Consequently, this may enable construction of hypoallergenic variants through subtle and well targeted amino acid replacements.

The analysis of co-varying AA positions, in conjunction with tropomyosin epitope data, for a single protein only, support the conjecture that ELSC pairs can show allergen important sites.

An appreciably larger number of AA-sequence-characterized allergen tropomyosins is likely to improve the ELSC analysis and give more strength to potentially important sites. Conversely, more data about the proteins, such as epitope etc. could help support the utility of ELSC.

6.2.3 Structure prediction no good at all

As learnt from this study, The homology prediction of tropomyosin variants, as accomplished by SWISS-MODEL and 3D-JIGSAW and based on defined structural templates, is simply insufficient to disclose small differences in the tertiary structure. The few AA changes between the original Pen A1 and the mutated counterpart give the same template in two prediction runs, each based on a distinct algorithm. Moreover, there are a limited number of templates in PDB to chose from. Furthermore, the predicted structures consist of only one tropomyosin α-helix hence the important coiled-coil structure does not appear in the predictions. If also allergen tropomyosin was available as an experimentally defined structure subtle structural differences between allergen and non-allergen tropomyosins would be available to analysis, including in silico homology modeling.

6.3 *In silico* analysis of parvalbumins

6.3.1 Phylogeny

As seen in figure 7 the parvalbumin dataset spans over a relatively wide evolutionary distance, although most of the allergens cluster on one branch of the tree. This includes

both mammalian and fish sequences, which is to be considered as good for ELSC studies. The allergen subset is also quite diverse. Notably two separate sequences appears to bee more related to several non-allergen sequences.

6.3.2 ELSC

The parvalbumin data set, used in this study is much smaller than that of tropomyosins. In total the repository encompassed 29 parvalbumins. As mentioned earlier, Fodor states that at least 15-20 sequences are needed to produce reliable ELSC results provided that a wide phylogenetic relationship is covered. The allergen subset did unfortunately not fulfill this requirement. Still, some accumulation of co-varying positions occurred in and on the border of one of the Ca^{2+}-binding regions that overlap the segment of epitope 4. This pattern was, however, seen in non-allergen parvalbumins only. The accumulation of non-allergen pairs in the epitope 4 segment is probably due to to the overlapping Ca^{2+}-binding region. A larger dataset, especially on allergen parvalbumins, may reveal insight in the significance of the possible importance of the vaguely identified regions.

6.4 ELSC + WRABL

There are several other ways to group AAs, but the WRABL approach was chosen because it is founded on a meticulous analysis and is one among the newest methods available for this purpose.

A major rationale for using WRABL aggregation was to investigate whether any of the correlating AA-$ELSC(i,j)$-pairs, showing a mutation from one functional group to another (in a WRABL contex) in the first position i that is accompanied by a change of functional group of the amino acid in position j. A mutation leading to a change of functional group of that position should be less common than those occurring within such a group. An $ELSC(i,j)$-pair yielding high scores in ELSC procedures, using both the regular AA-MSA as well as after WRABL-translation, could indicate a particular functional or structural importance of such a pair. The results showing the ratio of overlapping ELSC pairs indicate that the non-allergen set has more sites that includes cross-over mutations between functional groups than the allergen data-set. A possible explanation for this observation is that the smaller allergen dataset is more conserved and all sequences being of the invertebrate kind, and less diverse than the vertebrate non-allergen data-set.

A browse through several individual $ELSC(i,j)$-pairs, identified one that drew special attention: The allergen $ELSC(229,238)$-pair appears in both WRABL- and regular AA-ELSC runs. In both positions ($i = 229$ and $j = 238$) a change from one functional group to another had occurred in 6 sequences, excerpt from MSA are shown in Table 4. An inspection of the circular phylogenetic tree of tropomyosin (figure 8) the 6 sequences belongs to the lower right group/branch of the allergens (the group that have the allergen seq Al_Q95WY0). Hence, increased frequency of cross-over mutations and thereby also overlapping $ELSC(i,j)$-pairs is seen in a set of related sequences of diverse phylogenetic

relationship. The example allergen $ELSC(229, 238)$-pair, also strengthen the idea that

Al_O18416	A	D	L	E	R	A	E	E	R	A
Al_O44119	A	D	L	E	R	A	E	E	R	A
Al_O61379	A	D	L	E	R	A	E	E	R	A
Al_O96764	A	D	L	E	R	A	E	E	R	A
Al_Q25456	A	D	L	E	R	A	E	E	R	A
Al_Q8T6L5	A	D	L	E	R	A	E	E	R	A
Al_Q7M3Y8	V	D	L	E	R	A	E	A	R	L
Al_O97192	V	D	L	E	R	A	E	A	R	L
Al_Q95WY0	V	D	L	E	R	A	E	A	R	L
Al_Q9GZ69	V	D	L	E	R	A	E	T	R	L
Al_Q9GZ70	V	D	L	E	R	A	E	A	R	L
Al_Q9GZ71	V	D	L	E	R	A	E	A	R	L
Position	229									238

Table 4: Excerpt from tropomyosin MSA, showing AA at position $i = 229$ to $j = 238$. Color indicates functional groups where Green is the "small" group and the Aquamarine indicates the Aliphatic (The gray shading is intermediate AAs). In the full MSA the major AA on position 229 and 238 is the Alanine (A) this indicates that a change from A to V or L has occurred. The 6 sequences holding the Aquamarine Aliphatic mutations is the ones that are branched together in the phylogenetic tree.

overlapping $ELSC(i, j)$-pair could pin-point positions of special functional importance. The appearance of the 6 sequences on a common major branch indicates that the changes in AA position 229 and 238 roughly coincided in (the evolution) time and, thus, causally linked events. Thus, on this provision, the positions are of such importance for the function of the protein that a mutation at pos $i = 229$ will induce a mutation at pos $j = 238$, or vice versa.

The configuration of the tropomyosin protein encompassing two parallel helices in a coiled-coil structure makes it difficult to intuitively identify correlations of great importance to protein function. Is the $ELSC(i, j)$-pair an intra helix correlation or a a helix-helix correlation? Pairs that have i and j very close to each other may indicate correlations that are important to the coil-coiled structure. Due to the parallel helix conformation those pairs are also close between the helices.

ELSC runs with both regular AAs and WRABL could give deepened insight in positions of particular relevance to protein structure. Identification of such positions could further improve construction of hypoallergenic variants.

7 Acknowledgments

I would like to thank my supervisor Daniel Soeria-Atmadja and Ulf Hammerling for great support and all the different inputs on this project, and my scientific reviewer Mats

Gustafsson at the Department of Engineering Sciences, Uppsala University. My opponents Jonathan Alvarsson and Christan Rutemark for reading and helping me improve this report. Anthony Fodor for being so helpful and when given me fast E-mail replays. Thank you *National Food Administration* for free Coffee and fruit and friendly colleagues! Moreover I thank HKF for putting up with me this stressful months "*sempre fistel*"! Finally my wonderful Marie for supporting me with lunch boxes and for bringing me home to bed at late nights!

7 References

[1] S. Saha and G. P. Raghava. Algpred: prediction of allergenic proteins and mapping of ige epitopes. *Nucleic Acids Res*, 34(Web Server issue), 2006.

[2] D. Soeria-Atmadja, T. Lundell, M. G. Gustafsson, and U. Hammerling. Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. *Nucleic Acids Res*, 34(13):3779–3793, 2006.

[3] R. C. Aalberse and B. M. Stadler. In silico predictability of allergenicity: from amino acid sequence via 3-d structure to allergenicity. *Mol Nutr Food Res*, 50(7):625–627, 2006.

[4] M. Jackson. Allergy: the making of a modern plague. *Clinical & Experimental Allergy*, 31(11):1665–1671, 2001.

[5] S. L. Hefle, J. A. Nordlee, and S. L. Taylor. Allergenic foods. *Crit Rev Food Sci Nutr*, 36 Suppl, 1996.

[6] T. P. King, D. Hoffman, H. Lowenstein, D. G. Marsh, T. A. Platts-Mills, and W. Thomas. Allergen nomenclature. who/iuis allergen nomenclature subcommittee. *Int Arch Allergy Immunol*, 105(3):224–233, 1994.

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

[8] J. A. Nordlee, S. L. Taylor, J. A. Townsend, L. A. Thomas, and R. K. Bush. Identification of a brazil-nut allergen in transgenic soybeans. *N Engl J Med*, 334(11):688–692, 1996.

[9] FAO/WHO. Evaluation of allergenicity of genetically modified foods. Technical report, Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, 2001.

[10] Codex Alimentarius. Codex alimentarius commission. report of the fourth session of the codex ad hoc intergovernmental task force on foods derived from biotechnology (alinorm 03/34a). Technical report, Codex Alimentarius Commission, 2003.

[11] EFSA. Guidance document of the scientific panel on genetically modified organisms for the risk assessment of genetically modified plants and derived food and feed. *The EFSA Journal*, 99:1–94, 2004.

[12] K. Bailey. Tropomyosin: a new asymmetric protein component of the muscle fibril. *Biochemical J*, 43:271–287, 1948.

[13] F. G. Whitby and G. N. Phillips. Crystal structure of tropomyosin at 7 angstroms resolution. *Proteins*, 38(1):49–59, 2000.

[14] Wg Lewis and Lb Smillie. The amino acid sequence of rabbit cardiac tropomyosin. *J. Biol. Chem.*, 255(14):6854–6859, 1980.

[15] RCSB PDB Research Collaboratory for Structural Bioinformatics The Protein Data Bank. http://www.pdb.org/, 3 August 2006.

[16] W.L. DeLano. The pymol molecular graphics system, http://www.pymol.org, 27 August 2006.

[17] R. Ayuso, S. B. Lehrer, and G. Reese. Identification of continuous, allergenic regions of the major shrimp allergen pen a 1 (tropomyosin). *Int Arch Allergy Immunol*, 127(1):27–37, 2002.

[18] B. M. Wolska and D. M. Wieczorek. The role of tropomyosin in the regulation of myocardial contraction and relaxation. *Pflugers Arch*, 446(1):1–8, 2003.

[19] Gerald Reese, Julia Viebranz, Susan M. Leong-Kee, Matthew Plante, Iris Lauer, Stefanie Randow, Mar S. Moncin, Rosalia Ayuso, Samuel B. Lehrer, and Stefan Vieths. Reduced allergenic potency of vr9-1, a mutant of the major shrimp allergen pen a 1 (tropomyosin). *J Immunol*, 175(12):8354–8364, 2005.

[20] Allergome database. http://www.allergome.org, 12 July 2006.

[21] Eva Untersmayr, Krisztina Szalai, Angelika B. Riemer, Wolfgang Hemmer, Ines Swoboda, Brigitte Hantusch, Isabella Scholl, Susanne Spitzauer, Otto Scheiner, and Reinhart Jarisch. Mimotopes identify conformational epitopes on parvalbumin, the major fish allergen. *Molecular Immunology*, 43(9):1454–1461, 2006.

[22] C. D. Lindstrm, T. van D, I. Hordvik, C. Endresen, and S. Elsayed. Cloning of two distinct cdnas encoding parvalbumin, the major allergen of atlantic salmon (salmo salar). *Scand J Immunol*, 44(4):335–344, 1996.

[23] S. Elsayed and H. Bennich. The primary structure of allergen m from cod. *Scand J Immunol*, 4(2):203–208, 1975.

[24] C. Hilger, F. Grigioni, L. Thill, L. Mertens, and F. Hentges. Severe ige-mediated anaphylaxis following consumption of fried frog legs: definition of alpha-parvalbumin as the allergen in cause. *Allergy*, 57(11):1053–1058, 2002.

[25] V. D. Kumar, L. Lee, and B. F. Edwards. Refined crystal structure of calcium-liganded carp parvalbumin 4.25 at 1.5-a resolution. *Biochemistry*, 29(6):1404–1412, 1990.

[26] Agnes Bugajska-Schretter, Lena Elfman, Thomas Fuchs, Sonja Kapiotis, Helmut Rumpold, Rudolf Valenta, and Susanne Spitzauer. Parvalbumin, a cross-reactive fish allergen, contains ige-binding epitopes sensitive to periodate treatment and ca2+ depletion. *Journal of Allergy and Clinical Immunology*, 101(1):67–74, 1998.

[27] J. P. Dekker, A. Fodor, R. W. Aldrich, and G. Yellen. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, 20(10):1565–1572, 2004.

[28] E. Neher. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*, 91(1):98–102, 1994.

[29] Werner Terhalle and Andreas Dress. Positional dependence, cliques, and predictive motifs in the bhlh protein domain. *Journal of Molecular Evolution*, 48(5):501–516, 1999.

[30] S. M. Larson, A. A. Di Nardo, and A. R. Davidson. Analysis of covariation in an sh3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J Mol Biol*, 303(3):433–446, 2000.

[31] Steve W. Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

[32] James O. Wrabl and Nick V. Grishin. Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. *Proteins: Structure, Function, and Bioinformatics*, 61(3):523–534, 2005.

[33] D. Petrey and B. Honig. Protein structure prediction: inroads to biology. *Mol Cell*, 20(6):811–819, 2005.

[34] SWISS-MODEL. http://swissmodel.expasy.org, 27 August 2006.

[35] P. A. Bates, L. A. Kelley, R. M. MacCallum, and M. J. Sternberg. Enhancement of protein modeling by human intervention in applying the automatic programs 3d-jigsaw and 3d-pssm. *Proteins*, Suppl 5:39–46, 2001.

[36] Structural Database of Allergenic Proteins SDAP. http://fermi.utmb.edu/sdap/index.html, 10 July 2006.

[37] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue), 2006.

[38] FASTA format description. http://www.ncbi.nlm.nih.gov/blast/fasta.shtml, 12 August 2006.

[39] T. Lassmann and E. L. Sonnhammer. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6, 2005.

[40] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, 1994.

[41] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

[42] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, 2000.

[43] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, 2003.

[44] O. Gascuel. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol Biol Evol*, 14(7):685–695, 1997.

[45] T. Lassmann and E. L. Sonnhammer. Kalign, kalignvu and mumsa: web servers for multiple sequence alignment. *Nucleic Acids Res*, 34(Web Server issue), 2006.

[46] Carolin Kosiol and Nick Goldman. Different Versions of the Dayhoff Rate Matrix. *Mol Biol Evol*, 22(2):193–199, 2005.

[47] MATLAB The MathWorks. http://www.mathworks.com/, 20 August 2006.

[48] Gentoo Linux. on world wide web http://www.gentoo.org, 5 August 2006.

[49] OpenSSI (Single System Image) Clusters for Linux. http://openssi.org/, 27 August 2006.

[50] The openMosix Project. http://openmosix.sourceforge.net/, 27 June 2006.