СУ „Св. КлиментОхридски”
Факултет по Химия и Фармация
Катедра „Аналитична химия”

Хемометричен подход за класифициране на хранителни протеини в категории „Алергени“ и „Неалергени“

Людмила Христова Нанева

АВТОРЕФЕРАТ

на дисертационен труд за присъждане на образователна и научната степен „ДОКТОР”

научна област: 4.“Природни науки, математика и информатика“
професионално направление: 4.2. „Химически науки“
научна специалност: 01.05.04 „Аналитична химия“

Научен ръководител: Проф. д-р Васил Симеонов, дхн
Научен консултант: Д-р Мирослава Недялкова

София, 2018

Благодарности:

Дължа голяма благодарност на научния си ръководител проф. д-р Васил Симеонов, дхн и на научния си консултант д –р Мирослава Недялкова за подкрепата и съветите, благодарение на които беше реализирана тази дисертация.
Изказвам благодарност и на проф. дхн Ирини Дойчинова за помощта и насоките в начюлният етап на подготовката на дисертационния труд.
Изказвам благодарност на проекта за подкрепа на докторанти към СУ: Проект BG05M2OP001-2.009-0028 въз основа на който приносите в дисертационния труд бяха популязирани.

I. Увод

Един от най – актуалните здравни проблеми на съвременния живот са алергиите. Хранителните алергии засягат 8% от новородените и подрастващи деца.С навлизането на ГМО хранителните продукти и все по широкото използване на пестициди този процент непрекъснато се увеличава. Алергиите биват предизвикани от различни източници: мляко, яйца, фъстъци, соя, миди, плодове, прахови частици и др. [1-3]. Предизвикана от външни и вътрешни фактори, алергията включва серия сложни реакции, допринасящи за развитие на заболявания с характерна симптоматика като хрема, астма, атипичен дерматит, кожна сенсибилизация. Може да се появят и тежки реакции като остър и фатален анафилактичен шок [4-5]. Първият алерген (Jun a 3) е бил изолиран от цветен прашец на кедър. По-късно са били изолирани подобни алергени от пипер (Cap a 1), череши (Pru av 2), киви (Act c 2), домати (Lyc e NP24) и ябълка (Mal d 2) [6-9]. Алергенни протеини са изолирани също от храни като мляко [10-11] (казеин [12-14] и лактоглобулин), яйца [15] (овомукоид и лизозим), риба [16-18] (парвалбумин), бобови албумини [19-21] и др. Разпознаването на алергенните протеини е важно, поради все по-широкото използване на модифицирани протеини в храни, лекарства, битова химия и др [22-24].

Всичко това налага изучаване на свойството алергенност и индентифицирането на алергенните протеини. В настоящата дисертация е създаден хемометричен подход за разпознаване на алергени от хранителен произход и опит за тяхното класифициране по отношение на биологичното им свойство- алергенност.

Получените модели са валидирани и могат да се използват при скрининг за алергенност на нови протеини.

Актуалността на темата се обуславя от:

- Значителния и увеличаващ се процент на заболяваемост по отношение на алергиите сред новородени и подрастващи деца.
- Биоинформатичните подходи за предсказване на алергенни протеини следвайки насоките на FAO и HWO за търсене на прилика в аминокиселинната последователност, генерира голям брой фалшиви резултати. Неизвестни протеини не могат да бъдат предсказани като алергени [25].
- Повишен интерес към разработване на алтернативни подходи (in vitro, in silico) за предсказване на алергени (Stadler и Stadler 2003).

Обект на изследване в на настоящия труд са алергенни протеини с хранителен произход.

Обхват: Проучването е фокусирано върху разработването на модел чрез ***in silico*** подход за идентифициране на алергенни протеини, търсейки връзка между физико – химични свойства на аминокиселините изграждащи протеините и свойството алергенност.

2. Литературен обзор

В литературния обзор е изяснен терминът „алергия“. Хранителната алергия се различава от други реакции на организма към храна, като хранителния интолеранс (непоносимост), лекарствената непоносимост, токсин-медиираните реакции[26]. Интолерансът към храни е невъзможността на организма да обработва правилно храната, което обикновено се дължи на липсата на някои ензими, а при хранителната алергия имунната система генерира отговор с антитела към резорбираната в организма храна [27-29] . Нейният механизъм се обуславя от създаването на алергична реакция която се получава, когато податливият организъм е изложен на действие на специфичен протеин. Тъй като организмът възприема този протеин (алерген) като заплаха, той започва да произвежда Т-хелперни лимфоцити (Th2), които освобождават интерлевкини. Интерлевкините повишават продукцията на антитела, наречени имуноглобулини

Е (IgE), от В-клетките. Организмът реагира като произвежда голямо количество от тези антитела. Последните се свързват към специфични клетки в кръвта – мастоцити. При повторно постъпване в организма на същия алерген, той се свързва с антителата, разположени върху мастоцитите. В резултат на реакцията антиген – антитяло, мастоцитите освобождават хистамин, който причинява алергичните симптоми: зачервяване, подуване, сърбеж [30-32].

***In silico* методи за разпознаване на алергени**

Въпреки, че няма консенсус за структура на алергените, Организацията на Обединените нации по прехрана и земеделие (FAO) и Световната здравна организация (WHO) имат насоки за оценка на потенциално-алергизиращото действие на нови протеини [33-34]. Според тези насоки, изследваният протеин е потенциален алерген, ако има в структурата си идентични от 6 до 8 последователни аминокиселини (ак) или 35% подобие в рамките на 80 последователни остатъка в сравнение с вече известни алергени [35].

В момента има два вида биоинформатични подхода за предсказване на алергени. Първият подход следва насоките на FAO и WHO и търси прилика в аминокиселинна последователност. Създадени са база данни, съдържащи обширна информация на известни алергени, които се използват при търсене на подобие в ак последователност на изследвания протеин. Такива бази данни са Structural Database of Allergenic Proteins (SDAP) [36], Allermatch [37] и AllerTool [38]. Този подход има добра разпознавателна способност, но генерира и голям брой фалшиви алергени [39-44]. Освен това, откриването на нови структурно различни алергени е ограничено от липсата на подобие с известни вече алергени.

Вторият подход се основава на идентифицирането на линейни мотиви за алергенност. *Мотивът е последователност от ак, отговорна за определена активност на протеина.* Stadler и Stadler (2003) дефинират 52 алергенни мотива чрез сравняването на алергени и неалергени с програмата *MEME* [45]. Li и съавтори (2004) идентифицират мотив за алергенност чрез кластериране на известни алергени по методите на вълновия анализ и скрит Марков модел (HMM) [46-47]. Bjorklund и съавтори (2005) [48] разработват метод за идентифициране на алергени чрез разпознаване на алергенни пептиди (allergen-representative peptides, ARP). Подробен анализ на отделните семейства алергени показва колко специфични мотиви могат да корелират с IgE епитопи[49]. Те могат да бъдат използвани за разграничаване на алергени от неалергенни протеини, дори когато те имат сходни ензимни мотиви като пектатни лиази от човешки микроби[50]. AlgPred е сървър за предсказване на алергенен протеин, който съчетава 4 метода за търсене на мотив: машини с поддържащи вектори (SVM), програма MEME/MAST, епитопи за IgE и алергенни пептиди ARP [51].

И двата подхода приемат, че алергенната активност на протеините е линейно кодирано свойство. За да действа като алерген, един протеин трябва да съдържа епитопи за свързване както с Th-2 клетките, така и с В-лимфоцитите. *Епитопът е тази част от протеина, която взаимодейства с друг протеин*[52]. Епитопите за Th-2 са линейни, но епитопите за В-клетките са краткотрайни конформационни епитопи, образувани на повърхността на протеина [53]. Furmonaviciene и съавтори (2005) [54] са открили алерген-специфични участъци, състоящи се от необичайно висок процент хидрофобни остатъци на повърхността. Това откритие е в съгласие с факта, че имунната система при човека има способността да открива хидрофобни части от имуногенни протеини, съставени от алифатни или ароматни ак [55].

Очевидно свойството „алергичност", също като свойствата „имуногенност" и „антигенност" са кодирани линейно и нелинейно и следователно разгледаните подходи не винаги са в състояние еднозначно да открият това свойство.

Като дял от аналитичната химия, хемометрията може да се използва за решаване на редица проблеми. Тя може да помага на изследователя при избор на аналитичен метод и да бъде включена във всеки един етап от вземане на проба до получаване на данни, да оптимизира и валидира избрания метод и накрая да обработва получените данни като ги интерпретира във вид

на ясни отговори. С други думи *хемометрията обхваща математически и статистически методи, които регулират потока от химическа информация и извличат знание от тези данни.*

Централна идея в хемометрията е *концепцията за модела.* Сложността на модела диктува условията за провеждане на експеримента и избора на средства, с които да се извлича информация от получените данни. Моделът пресъздава с определена точност реалния обект. Математическите модели поради универсалността на математическия език дават възможност за количествени оценки и качествени преобразувания на моделираните страни на обекта. В зависимост от това, между кои фактори и обекти се установява функционална зависимост, могат да бъдат построени различни математически модели. При построяването на тези модели основните задачи, които стоят пред изследователя, се свеждат до избора на входните данни и на вида на модела, до определянето на коефициентите и до статистическата им оценка. Основен подход при различни методи в хемометрията е *машинното обучение*. За провеждането му е необходимо едно множество от *дескриптори*, които да описват свойствата на изследваните обекти.

Регресионният анализ е основен статистически метод, който се използва в хемометрията. Този метод е популизиран от Hansch [57], който е търсил съотношение между данни за биологична активност и липофилните, електронните и стеричните свойства за серия вещества – структурни аналози. Най-често използвания регресионен анализ в хемометрията е множествения линеен регресионен анализ (Multiple Linear Regression, MLR)

Съществено ограничение в MLR е дескрипторите да не са взаимнозависими, т.е. да не корелират помежду си. В съвременните изследвания обкновенно се използва голямо множество дескриптори, някои от които са взаимнозависими. За да се реши този проблем се използват два метода: анализ на главните компоненти (Principal Components Analysis, PCA) и метод на частично най-малките частични квадрати (Partial Least Squares, PLS) [58-60].

Анализът на главните компоненти (Principal Component Analysis, PCA) се използва за намаляване броя на променливите в един модел като ги групира по подобие в главни компоненти (PC) (Eriksson *et al.*, 2001) [61]. PCA показва кои променливи допринасят в най-голяма степен за дисперсията в групата и групира в клъстери съединения с близки свойства.

Главната задача на PCA е разлагането на (*m* x *n*) матрица от данни ***X*** с *m*– характеристики и *n*-обекта на следните части:

- факторни натоварвания или тегла (factor loadings) ***A;***
- факторни коефициенти или резултати (factor scores) ***F;***

Необходимо е да се спомене за още една характеристика на *PCA*. Теоретично при анализа на ***n*** обекта, характеризирани чрез ***m*** променливи, PCA може да извлече от данните точно ***m*** главни компонента. Първият от тях (PC1) представлява онова направление в набора от данни, което съдържа най-голямата вариация. Втори главен компонент (PC2) е ортогонален спрямо (PC1) и представлява направлението на най-голямата остатъчна вариация около (PC1). Трети главен компонент (PC3) е ортогонален спрямо първите два и определя направлението на най-голямата остатъчна вариация около равнината образувана от PC1 и PC2.

.*Методът на частично най-малките квадрати (Partial Least Squares, PLS)* е регресионен метод с редица предимства пред обикновената множествена линейна регресия (МЛР) (Eriksson *et al.*, 2001) [61]. Той може да обработва матрици, съдържащи повече променливи (дескриптори), отколкото съединения, без да се получават предефинирани модели. В моделите могат да участват дескриптори, които взаимно корелират. PLS е в състояние да обработва матрици, в които 10-20% от данните липсват или има експериментална грешка. PLS може да се разглежда като регресионно продължение на PCA, в което една от независимите променливи X е заменена със зависимата Y. Аналогично на PCA, при PLS се извеждат главни компоненти, като линейна комбинация от началните променливи, които обясняват дисперсията в разглежданата група. За разлика от PCA, където по-големият брой PC дава по-добро обяснение на дисперсията, при PLS

се постига оптимален брой РС. Под оптимален брой РС се разбира броят РС, при който моделът има едновременно най-висока обяснителна и най-висока предсказваща способност. Това се постига чрез вътрешно кръстосано валидиране.

Целта на метода на частично най-малките квадрати е да намери малък брой *A* от подходящи фактори, които могат:

- да предсказват стойности на **Y** (променливите в матрицата от изходни данни).
- да използват ефективно информацията за **X.**

В този аспект регресията по частично най-малките квадрати прилича на тази по главни компоненти, като разликата е само в критериите за избор на фактори.

Факторите, подбрани при частично най-малките квадрати могат да се разглеждат като модифицирани главни компоненти. Отклонението от скритите фактори на анализа на главни компоненти PCA води до подобряване на корелацията за сметка на понижение на процента на обяснена вариация на системата. Алгоритъмът комбинира практически две изчисления на PCA, едно за матрицата на обектите **X** и едно за матрицата на променливите **Y.**

Кластерният анализ (КА) е общо название на множество изчислителни процедури, използвани при създаването на класификации на обекти. Той позволява да се открият нови зависимости и свойства в дадено множество от данни, които не биха могли да се установят по други известни теоретични и експериментални методи. Класифицираните обекти могат да бъдат както наблюдения (случаи), така и променливи. [62-63].

Целта на КА е да се намери оптимално групиране на наблюденията (или на променливите), при което елементите от даден кластер са подобни, но кластерите ясно се отличават един от друг. Обикновено в КА броят на групите предварително е неизвестен и се определя от изследователя в процеса на изследването.

Да приемем, че разполагаме с *n* обекта, които трябва да бъдат кластерирани чрез стойностите на *m* характеристики. Това води до съставяне на матрица от данни *X* с размерност *n* x *m*, имаща следния вид:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & x_{ij} & \dots \\ x_{n_1} & x_{n_2} & \dots & x_{nm} \end{pmatrix}$$

Всеки обект може да се представи с вектора X_i, който се нарича вектор на обекта *i*. Целта на кластерирането е да групира всичките *n* на брой обекта в съответствие с характерните им признаци.

Друга алтернатива е да се кластерират променливите в зависимост от обектите, които те описват. Така, че едно важно правило при КА е следното: както обектите се описват чрез стойностите на характеристиките, така и характеристиките се описват чрез регистрираните стойности за обектите, върху които се извършва измерването на характеристиките.

За да се намери структурата на данните по отношение на групирането на обектите или променливите, е необходимо да се дефинира мярка за подобие между тях. Подобието между два обекта може да се измерва по различни начини. Най-съществени за целите на кластерирането са разстоянията между обектите, корелацията между тях или информационното съдържание на обектите, описани чрез съответните данни.

Формираните групи (кластери) трябва да бъдат еднородни (хомогенни) вътре и разнородни (хетерогенни) помежду си по зададени характеристики. Установените сходства и различия между точките се определят в съответствие с разстоянията между тях. Два обекта са

идентични, ако разстоянието между тях е нула. Колкото по-голямо е разстоянието между два обекта, толкова по-голямо е несходството между тях. За оценка на сходството на обектите, попадащи в даден кластер, се използват метрики, които обикновено се основават на евклидовото разстояние между точките или подобни на него характеристики[64]. Познати са две основни групи кластерни алгоритми: йерархични и нейерархични.

При йерархичното кластериране се използват два основни подхода. Първият се нарича агломеративен. В началото всеки обект се разглежда като отделен кластер. В отделните обекти се формират най-напред малки кластери, после по-големи, които се получават от обединяването на по-малките. Другият подход е този на разделянето. Всички обекти се разглеждат като един голям кластер, от който се получават по-малки чрез последователно разделяне на началната структура. При йерархичните методи за кластериране от цялата съвкупност от данни се получават няколко отделни кластера, разположени по такъв начин, че малките кластери се включват в големите. Най-типичният резултат е т.н. дендрограма (йерархично дърво), която представлява схема за взаимното свързване на отделните групи обекти в кластери. Важно е да се отбележи, че в зависимост от нивото, където ще прекъснем връзките получаваме серия от класификации на обектите с различен брой кластери. При нейерархичните методи за кластериране се цели разделяне на съвкупността от обекти в *K* на брой (предварително зададен) кластера по такъв начин, че обектите, принадлежащи на един и същ кластер са разположени близо един до друг, а отделните кластери са добре разделени. Тъй като всички *K* кластера се получават едновременно, получената класификация няма йерархичен характер и не може да бъде представена графично. Изчислителната работа по този тип класификация е по-дълга и сложна и почива на някои евристични алгоритми[65-66].

Биоинформатика

Биоинформатиката обработва и анализира експериментални данни, отнасящи се до структурата на биологичните макромолекули (белтъци и нуклеинови киселини) с цел да се получи нова значима информация. Тя е сравнително нова интердисциплинарна наука, включваща знания от областта на биологията, химията, математиката, информатиката и др. Задачите на биоинформатиката могат да се дефинират както следва:

- моделиране на биологични системи и функции;
- анализиране на лабораторни данни;
- изследване на нови данни с помощта на математически модели;
- генериране на модели на базата на натрупани данни от експерименти;
- разпознаване на мотиви в експериментални данни;
- предсказване на функции на гени и белтъци;
- провеждане на *in silico* експерименти.

В зависимост от наличната изходна информация за анализ, методите в биоинформатиката се разделят на секвенционални и структурни. Изходната информация при секвенционалните методи са данни за първичната структура на протеини и нуклеинови киселини и техните функции. Чрез анализ на последователности на молекули с или без дадена функция, се идентифицират мотиви за тази функция. За анализ се използват различни статистически методи и методи за машинно обучение като регресионен анализ [67], невронни мрежи (ANN), машини с поддържащи вектори (SVM)[68-70], йерархични и нейерархични кластърни методи и др.

Изходната информация при структурните методи са данни за вторичната, третичната и четвъртичната структура на биомакромолекулите и комплексите между тях, получени чрез кристалография или ЯМР. Анализът на интерфейсната повърхност и симулирането на взаимодействията дават възможност да се прогнозират функциите на молекулите. Най-често използваните структурни методи са молекулна динамика и молекулен докинг. Молекуланата

динамика прилага законите на класическата механика, за да симулира движението на молекулите и промените в конформациите им, като изчислява атомните координати във функция от времето. Молекулният докинг възпроизвежда само последната стъпка от процеса на взаимодействие, а именно образувания вече комплекс между лиганда и макромолекулата [71].

3. Цел и задачи

Целта на настоящата дисертация е да се създаде хемометричен модел за разпознаване на алергени с хранителен произход, базиращ се на свойствата на аминокиселинните остатъци, изграждащи алергенните протеини.

За постигане на целта трябваше да бъдат изпълнени следните задачи:

1. Да се създаде база данни от алергени и неалергени с хранителен произход.
2. Да се опише структурата на протеините чрез аминокиселинни дескриптори.
3. Да се представят протеините като вектори с еднаква дължина.
4. Да се изведе модел за разпознаване на алергени.
5. Моделът да бъде валидиран чрез външна тестова група алергени и неалергени.
6. Да се дефинира мотив за алергенност.
7. Да се извърши опит за експресно разделяне на алергенни от неалергенни протеини използвайки кластерен анализ.

4. Материали и методи

4.1. Алергени и неалергени, използвани в дисертацията

Първоначално бе създадена база данни от 702 алергена и 702 неалергена от хранителен произход. Алергените бяха компилирани от базите данни CSL (Central Science Laboratory) (http://allergen.csl.gov.uk), FARRP (Food Allergen Research and Resource Program) (http://www.allergenonline.org) и SDAP (Structural Database of Allergenic Proteins) (http://fermi.utmb.edu/SDAP/sdap_man.html)[96]. Неалергените бяха избрани от същите организми, от които произхождаха алергените, след BLAST търсене по признак минимално структурно подобие[97].

В последствие бе създадена втора база данни от 700 алергена и 700 неалергена от хранителен произход. Някой от съществуващите протеина в първата база данни бяха заменени с нови , а други бяха изключени. Целта на втората група данни е да се създаде подобна матрица, с цел прилагане на методологията и проверка на модела за неговата валидност.

4.2. Дескриптори на химичната структура

физико-химично свойство като температура на кипене, парно налягане и някои други емпирични или изчислителни молекулни свойства.

. *z-дескриптори и E- дескриптори*

Z-дескрипторите описват свойства на аминокиселините [102]. Те са получени чрез прилагане на анализ на главните компоненти върху набор от 237 молекулни дескриптора на аминокиселините (Hellberg *et al.*, 1987) [103]. Първоначално са били изведени три *z*-дескриптора, описващи хидрофобността, размера и електронните свойства на аминокиселините (Eriksson *et al.*, 1989)[104] (Таблица 1.а). Те са използвани за анализ на активността на пептапептиди, потенциращи брадикинина.

По-късно Sandberg и колеги (1998)[64] увеличават броя на *E*-дескрипторите до 5 (Таблица 1.б) и ги използват за класификацията на 89 синтетични субстрати на еластазата и 29 пептидни аналози на невротензина. Останалите *E*-дескрипторите описват поредността (номера) на триплетния кодон, склонността към образуване на β – вторични структури и намират широко приложение в QSAR на пептиди (Siebert, 2001) [105] и в протеохемометрията (Lapinsh *et al.*, 2001) [106].

В настоящата дидертация структурата на протеините, алергени и неалергени, са описани първо чрез три *z*-дескриптора на изграждащите аминокиселини(първа група), а след това с петте-*E*-дескриптора(втора група).

Таблица 1.а. *Z-дескриптори* (Eriksson *et al.*, 1989).

ак	z_1	z_2	z_3
Ala	0.07	-1.73	0.09
Arg	2.88	2.52	-3.44
Asn	3.22	1.45	0.84
Asp	3.64	1.13	2.36
Cys	0.71	-0.97	4.13
Gln	2.18	0.53	-1.14
Glu	3.08	0.39	-0.07
Gly	2.23	-5.36	0.30
His	2.41	1.74	1.11
Ile	-4.44	-1.68	-1.03
Leu	-4.19	-1.03	-0.98
Lys	2.84	1.41	-3.14
Met	-2.49	-0.27	-0.41

Phe	-4.92	1.30	0.45
Pro	-1.22	0.88	2.23
Ser	1.96	-1.63	0.57
Thr	0.92	-2.09	-1.40
Trp	-4.75	3.65	0.85
Tyr	-1.39	2.32	0.01
Val	-2.69	-2.53	-1.29

Таблица 1.б. *E*-дескриптори (Sandberg et. Al.,1998)

ак	E_1	E_2	E_3	*E4*	*E5*
Ala	0.008	0.134	-0.475	-0.039	0.181
Arg	0.171	-0.361	0.107	-0.258	-0.364
Asn	0.255	0.038	0.117	0.118	-0.055
Asp	0.303	-0.057	-0.014	0.225	0.156
Cys	-0.132	0.174	0.07	0.565	-0.374
Gln	0.149	-0.184	0.03	0.035	-0.112
Glu	0.221	-0.28	-0.315	0.157	0.303
Gly	0.218	0.562	-0.024	0.018	0.106
His	0.023	-0.177	0.041	0.28	-0.021
Ile	-0.353	0.071	-0.088	-0.195	-0.107
Leu	-0.267	0.018	-0.265	-0.274	0.206
Lys	0.243	-0.339	-0.044	-0.325	-0.027
Met	-0.239	-0.141	-0.155	0.321	0.077
Phe	-0.329	-0.023	0.072	-0.002	0.208
Pro	0.173	0.286	0.407	-0.215	0.384
Ser	0.199	0.238	-0.015	-0.068	-0.196

Thr	0.068	0.147	-0.015	-0.132	-0.274
Trp	-0.296	-0.186	0.389	0.083	0.297
Tyr	-0.141	-0.057	0.425	-0.096	-0.091
Val	-0.274	0.136	-0.187	-0.196	-0.299

ACC-трансформация

За да преобразуваме протеините във вектори с еднаква дължина използвахме метода на авто- и кръстосаната ковариация (auto- and cross-covariance, ACC) [107].

Авто-ковариацията (A_{jj}) и кръстосаната ковариация (C_{ji}) бяха изчислени по следните формули:

$$A_{jj}(l) = \sum_{i}^{n-l} \frac{Z_{j,i} \times Z_{j,i+1}}{n-l} \quad (14)$$

$$C_{jk}(l) = \sum_{i}^{n-l} \frac{Z_{j,i} \times Z_{k,i+1}}{n-l} \quad (15)$$

За първия сет от 702 алергена и 702 неалергена - индексът j се отнася за z-дескриптора (j = 1, 2, 3); индексът i указва позицията на ак (i = 1, 2, 3 … n); n е броят на аминокиселините в протеина; l е лаг (l = 1, 2, …, L). Лагът е дължината на рамката от ак, за които се изчисляват A_{jj} и C_{ji}. Бяха избрани само къси лагове (L = 5), тъй като се изследва влиянието на разположени в съседство аминокиселини. Всеки протеин беше трансформиран в низ от 45 променливи (3^2 x 5) [108].

За втория сет от 700 алергена и 700 неалергена - индексът j и k се отнася за E-дескриптора (j = 1-5, k = 1-5, $j \neq k$), индексът i указва позицията на ак (i = 1, 2, 3 … n); n е броят на аминокиселините в протеина; l е лаг (l = 1, 2, …, L). Лагът е дължината на рамката от ак, за които се изчисляват A_{jj} и C_{ji}. Бяха избрани отново само къси лагове (L = 5 ÷ 20), тъй като се изследва влиянието на разположени в съседство аминокиселини. Всеки протеин беше трансформиран в низ от 200 променливи (5^2 x L) [108].

4.3. Дискриминантен анализ

Дискриминантният анализ (ДА) е метод за класификация на данни в класове въз основа на предварително изведена зависимост (Ligand-based design manual, Sybyl 6.6). Зависимостта се извежда на базата на обучаваща група, съдържаща съединения от всички разглеждани класове. ДА, използван в настоящия труд, е PLS-базиран, т.е. извеждат се главни компоненти за всеки клас въз основа на дескрипторите, които описват ХС на съединенията.

В настоящата дисертация дискриминантният анализ беше проведен чрез програма SIMCA P-8.0 [109]. Програмата дава възможност за анализ на данни по няколко статистически метода: PCA, многопараметров PLS и PLS-базиран на ДА(PLS-2 и PLS-ДА).

4.3.1. ROC статистика

Резултатите от ДА се оценяват чрез Receiver Operating Characteristic (ROC) крива[110]. За да се генерира ROC крива на един модел, класифициращ съединенията в два класа А и В, е необходимо да се установи броят на верните положителни TP (съединения, които принадлежат

на клас А и са предсказани като принадлежащи на този клас), верните отрицателни TN (съединения, които принадлежат на клас В и са предсказани като принадлежащи на този клас), грешните положителни FP (съединения, които принадлежат на клас В, а са предсказани като принадлежащи на клас А) и грешните отрицателни FN (съединения, които принадлежат на клас А, а са предсказани като принадлежащи на клас В) (Таблица 2).

Таблица 2. Категории съединения в дискриминантния анализ.

Действителни Предсказани	Клас А	Клас В
Клас А	TP верни положителни	FP грешни положителни
Клас В	FN грешни отрицателни	TN верни отрицателни

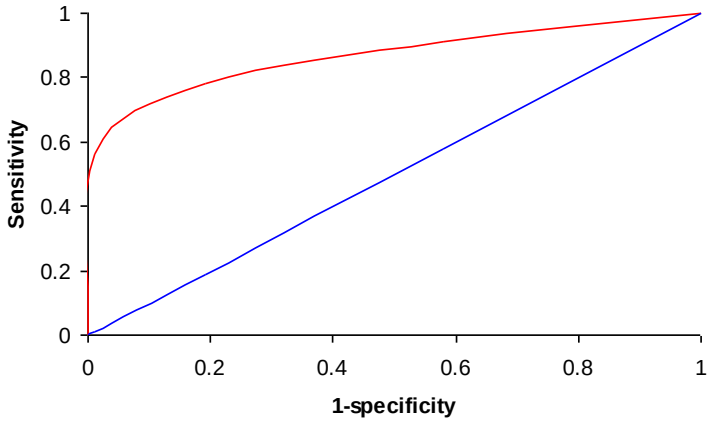
Въз основа на тези данни се изчисляват параметрите *чувствителност*, *специфичност* и *точност* на модела по формулите:

$$чувствителност = \frac{TP}{TP + FN} \quad (16)$$

$$специфичност = \frac{TN}{TN + FP} \quad (17)$$

$$точност = \frac{TP + TN}{TP + TN + FP + FN} \quad (18)$$

Тези параметри се изчисляват при различни разделителни прагове от 0 до 1, най-често със стъпка от 0.1. ROC кривата се генерира в координати *чувствителност*/*1- специфичност* (**Фиг.5**). Изчислява се площта под кривата AUC_{ROC} (или само A_{ROC}), която е индикатор за предсказващата способност на модела (Bradley, 1997)[111-112]. Стойността на AUC_{ROC} може да варира от 0.5 (моделът няма предсказваща способност) до 1.0 (моделът има идеална предсказваща способност). Модели с AUC_{ROC} от 0.6 до 0.8 имат добра предсказваща способност, а такива с AUC_{ROC}, по-високо от 0.8 – отлична предиктивност.



Фиг. 5. ROC крива на модел с координати *чувствителност*/*1- специфичност* ,без предсказваща способност (права) и на модел с отлична предсказваща способност (крива).

4.3.2. VIP променливи

Параметърът VIP (variable influence on projection) е въведен от Wold през 1993 г. (Eriksson *et al.*, 2001)[113] и сумира важността на всяка независима променлива за зависимата. Той представлява сумата от квадратите на теглата на PLS компонентите, отнесена към обяснената от всеки компонент дисперсия. Изчислява се по формулата:

$$VIP_{ak} = \sqrt{\left(\sum_{a=1}^{A} (w_{ak}^2 (SSY_{a-1} - SSY_a)) \frac{K}{SSY_0 - SSY_A} \right)} \qquad (19)$$

където w_{ak} е теглото (коефициента) на променливата k върху компонента a, а SSY_a е обяснената от компонента a дисперсия на Y променливата. Променливи със стойности на VIP, по-високи от 1, са от много съществено значение за изведения модел. Параметърът VIP се изчислява чрез програмата SIMCA-P 8.0 (Umetrics Ltd.).

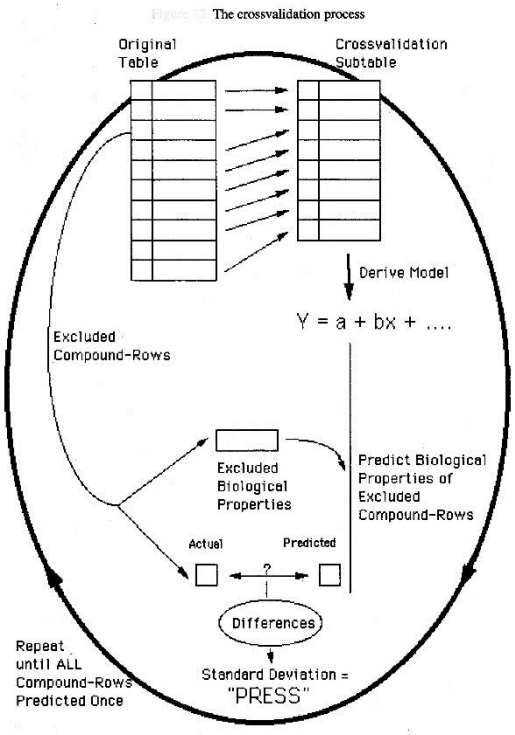
4.4. Валидиране на моделите

4.4.1. Вътрешно валидиране

Вътрешното или кръстосано валидиране (cross validation, CV) е процедура за определяне броя на оптималните PC въз основа на оценка на предсказващата способност на модела (Wold, 1995)[114-115]. Валидирането може да бъде в групи или поединично (leave-one-out, LOO-CV). Валидирането в групи е илюстрирано на **фиг.6.** Съединенията от обучаващата група се разделят произволно на няколко групи с приблизително равен брой съединения във всяка група. Една от групите се дефинира като тестова, а останалите образуват нова обучаваща група. Обучаващата група се използва за извеждане на модела, а тестовата група – за неговото валидиране. По модела, изведен въз основа на обучаващата група, се изчисляват Y променливите на съединенията от тестовата група (Y_{pred}) и се сравняват с експерименталните стойности чрез ROC статистика.

При поединичното кръстосано валидиране (leave-one-out, LOO-CV) се процедира както при валидирането в групи, но броя на групите е равен на броя на съединенията. При всяка итерация се изключва едно съединение, а моделът се извежда въз основа на останалите (n-1) съединения. Изчислява се Y_{pred} на изключеното съединение. Процедурата се повтаря и за останалите

съединения. Броят на повторенията е равен на броя на съединенията. Изчислените и експерименталните стойности се сравняват чрез ROC статистика.



Фиг.6. Поединичното кръстосано валидиране (leave-one-out) в група.

Оптималният брой главни компоненти на PLS, определен чрез CV, е броят на PC, при който AUC_{ROC} има максимална стойност[116].

4.4.2. Външно валидиране

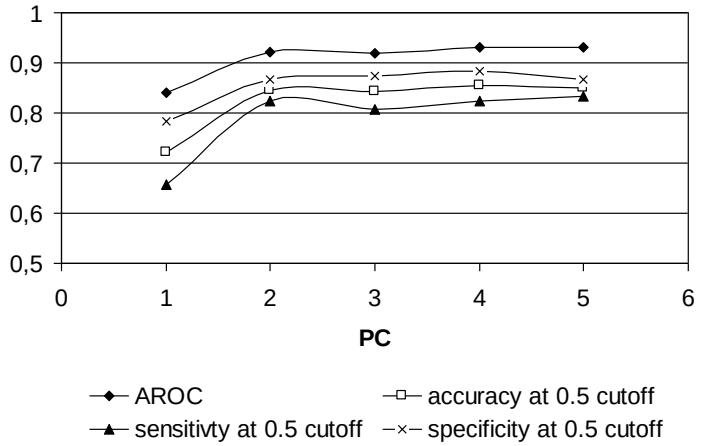
При външното валидиране на един модел се използва външна тестова група. Тя може да се състои от предварително избрани и отделени от обучаващата група съединения или да включва новосинтезирани и тествани съединения. И в двата случая, тестовата група трябва да отговаря на няколко изисквания. Първо, активността на съединенията в тестовата група не трябва да е по-ниска или по-висока от активността на съединенията в обучаващата група. Второ, съединенията в тестовата група трябва да имат активности, равномерно разпределени по целия интервал на активността. Трето, съединенията в тестовата група не трябва да съдържат заместители или структурни фрагменти, непредставени в обучаващата група. Предсказващата способност на модела се оценява с параметрите на дискриминанатния анализ: *чувствителност, специфичност, точност* и *площ под ROC кривата*.

5. Резултати

5.1. Резултати за първоначалната група от 702 алергена и 702 неалергена

5.1.1. Начален модел за разпознаване на алергени

За извеждането на начален, ориентировъчен модел за разпознаване на алергени беше създадена малка база данни от 120 алергена и 120 неалергена, избрани на случаен принцип сред всичките 1404 протеина, използвани в настоящата дисертация (Приложение 1). Структурата на протеините беше описана чрез трите *z*-дескриптора на изграждащите ги аминокиселини. Всеки протеин беше трансформиран в низ от 45 променливи, прилагайки *ACC*-трансформацията, както е описано в Глава 4 "Материали и методи". Така беше получена изходна матрица от два класа протеини (алергени и неалергени) и 45 променливи, която беше подложена на дискриминантен анализ по метода PLS. Изведени бяха няколко модела с вариращ брой на главните компоненти от 1 до 5. Моделите бяха оценени чрез параметрите *чувствителност*, *специфичност* и *точност* при праг 0.5. Отчетена беше и *площта под кривата* A_{ROC}. Резултатите са показани на **фиг.7.**



Фиг.7. *Чувствителност*, *специфичност* и *точност* при праг 0.5 и *площ под кривата* A_{ROC} за началния модел за разпознаване на алергени от първа група, при различен брой главни компоненти .

Резултатите показват, че добавянето на втори главен компонент значително повишава всички параметри на модела. По-нататъшното добавяне на главни компоненти първоначално слабо понижава параметрите, след което слабо ги повишава.

Полученият модел е даден в Таблица 3. Означението на променливите е както следва:

- първата цифра се отнася за номера на *z*-дескриптора на *i*-тата аминокиселина в протеина;
- втората цифра се отнася за номера на *z*-дескриптора на *j*-тата аминокиселина в протеина;
- третата цифра се отнася за лага (дължината на рамката от аминокиселини, за които се изчисляват ACC).

Например, ACC324 означава, че това е сумата от ACC стойностите, пресметнати за всяка двойка z_3- и z_2-дескриптори на аминокиселина, разположени на разстояние 4 аминокиселини (1-ва и 4-та, 2-ра и 5-та, 3-та и 6-та и т.н.).

Константата на модела е близка до 1. Променливите с положителен коефициент повишават вероятността за алергенност на протеина, тези с отрицателен – я намаляват.

Таблица 3. Начален модел за разпознаване на алергени.

променлива	коефициент	променлива	коефициент	променлива	коефициент
Const	0,998				

ACC111	0,083	ACC211	-0,047	ACC311	0,116
ACC112	0,030	ACC212	-0,024	ACC312	-0,009
ACC113	0,064	ACC213	-0,017	ACC313	-0,115
ACC114	-0,071	ACC214	0,043	ACC314	-0,142
ACC115	0,029	ACC215	-0,078	ACC315	-0,073
ACC121	-0,078	ACC221	0,003	ACC321	0,007
ACC122	-0,036	ACC222	0,084	ACC322	-0,031
ACC123	0,002	ACC223	0,052	ACC323	-0,081
ACC124	-0,045	ACC224	0,065	ACC324	0,132
ACC125	-0,055	ACC225	0,099	ACC325	0,001
ACC131	0,050	ACC231	0,032	ACC331	-0,125
ACC132	0,092	ACC232	-0,020	ACC332	-0,097
ACC133	-0,051	ACC233	0,103	ACC333	-0,162
ACC134	-0,085	ACC234	-0,096	ACC334	-0,178
ACC135	-0,021	ACC235	-0,014	ACC335	-0,083

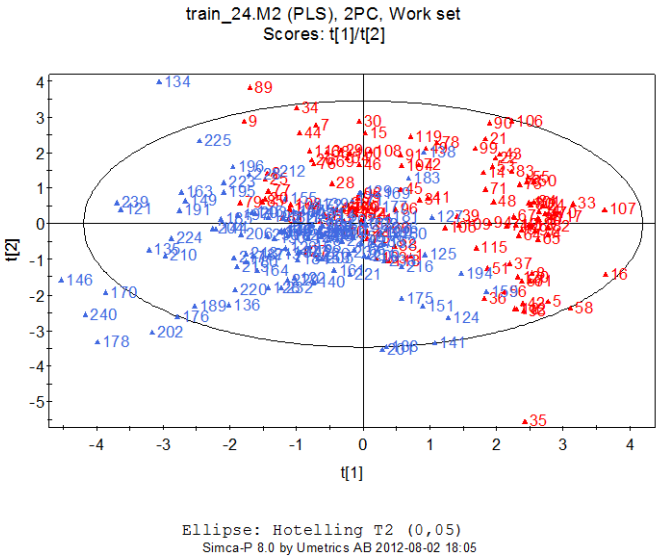
Променливите в модела бяха подредени по стойност на параметъра VIP (Таблица 4). Променливи с VIP > 1 имат съществено значение за модела. Деветнадесет променливи (42%) в модела имат VIP > 1. За да се отдиференцират най-съществените, прагът за VIP беше повишен на 1,5. Само 4 променливи имат VIP > 1,5. Това са ACC334, ACC333, ACC324 и ACC223. Две от тях са с положителни коефициенти (ACC324 и ACC223), другите две са с отрицателни (ACC334 и ACC333). Това означава, че протеини, имащи отрицателни стойности на променливите ACC334 и ACC333 и положителни на ACC324 и ACC223, е много вероятно да действат като алергени. Тази първоначална хипотеза по-нататък ще бъде потвърдена или отхвърлена при вътрешното и външното валидиране на модела.

Таблица 4. VIP стойности на променливите в модела. Променливи с VIP > 1,5 са дадени в получерен шрифт.

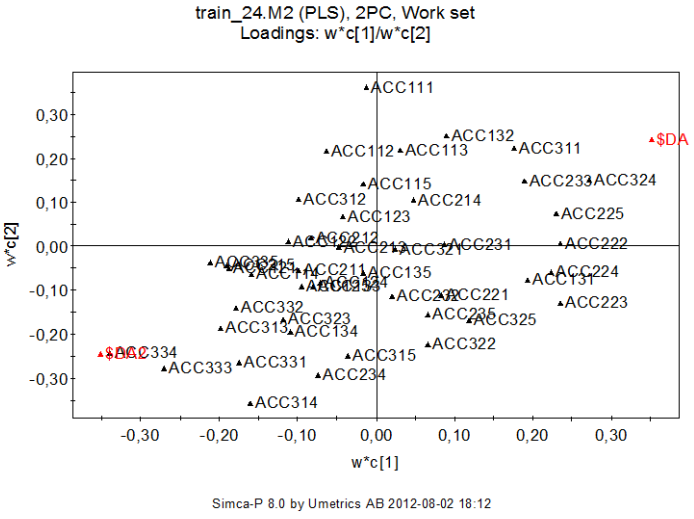
променлива	*VIP*	*коеф.*	*променлива*	*VIP*	*коеф.*	*променлива*	*VIP*	*коеф.*
ACC334	**2,003**	**-0,178**	ACC215	1,121	-0,078	ACC221	0,693	0,003
ACC333	**1,656**	**-0,162**	ACC121	1,101	-0,078	ACC122	0,685	-0,036
ACC324	**1,586**	**0,132**	ACC332	1,061	-0,097	ACC211	0,575	-0,047
ACC223	**1,578**	**0,052**	ACC325	1,011	0,001	ACC125	0,573	-0,055
ACC222	1,413	0,084	ACC234	0,950	-0,096	ACC212	0,526	-0,024
ACC224	1,413	0,065	ACC114	0,928	-0,071	ACC231	0,520	0,032
ACC225	1,344	0,099	ACC322	0,924	-0,031	ACC115	0,510	0,029
ACC314	1,315	-0,142	ACC112	0,896	0,030	ACC133	0,503	-0,051
ACC131	1,252	0,050	ACC132	0,867	0,092	ACC124	0,447	-0,045
ACC335	1,248	-0,083	ACC315	0,793	-0,073	ACC232	0,418	-0,020
ACC111	1,210	0,083	ACC134	0,792	-0,085	ACC214	0,393	0,043
ACC313	1,200	-0,115	ACC323	0,779	-0,081	ACC123	0,381	0,002

ACC331	1,185	-0,125	ACC312	0,770	-0,009	ACC213	0,284	-0,017
ACC311	1,128	0,116	ACC235	0,725	-0,014	ACC135	0,201	-0,021
ACC233	1,123	0,103	ACC113	0,699	0,064	ACC321	0,153	0,007

На **фиг.8а** са показани баловете (scores) на изследваните протеини на първа група, а на **фиг.8б** – товарите (loadings) на ACC променливите.



(а)



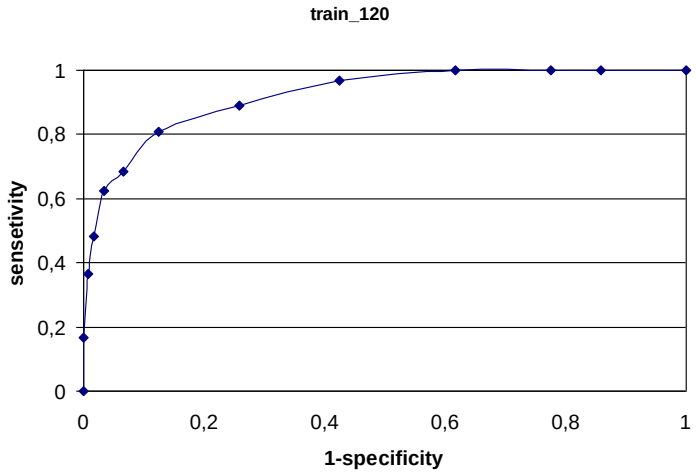
(б)

Фиг. 8. Балове (scores) на изследваните протеини (а) и товари (loadings) на ACC променливите (б). Алергените са показани в червено, неалергените – в синьо.

Моделът разграничава сравнително добре алергените (горе дясно, **Фиг.8а**) от неалергените (долу ляво), въпреки липсата на ясна граница между двата кластера. Променливите ACC324 и ACC223 са най-отдалечени в горния десен квадрант (отнасящ се за алергените, **Фиг.8б**), а променливите ACC334 и ACC333 са най-отдалечени в долния ляв квадрант (отнасящ се за неалергените).

Моделът беше тестван за *чувствителност*, *специфичност* и *точност* на предикциите при праг 0,5. Той разпознава 83% от алергените, 87% от неалергените и общо 85% вярно разпознати

алергени и неалергени в изследваната група протеини. Кривата *чувствителност/1-специфичност* е показана на **фиг. 9** и има площ 0,922.



Фиг. 9. Крива с координати *чувствителност/1-специфичност* на началния модел за разпознаване на алергени.

5.1.2. Вътрешно валидиране на модела

Началният модел за разпознаване на алергени с хранителен произход беше валидиран кръстосано в 6 групи. За целта групата от 120 алергена и 120 неалергена беше разделена на 6 групи по 20 алергена и 20 неалергена. Пет от групите бяха обединени в обучаваща група, шестата група беше тестова. Обучаващата група беше използвана за извеждане на модел, който беше валидиран с тестовата група. Процедурата беше повторена 6 пъти, така че всеки протеин участваше 5 пъти в обучаващата група и 1 път в тестовата. Получените модели бяха изведени при два главни компонента и бяха оценени по *чувствителност*, *специфичност* и *точност* на предикциите при праг 0,5 и по площ под ROC-кривата. Резултатите са показани в Таблица 5.

Таблица 5. Кръстосано валидиране в 6 групи на началния модел.

No.	*Обучаваща група* n = 100 *праг 0,5*				*Тестова група* n = 20 *праг 0,5*			
	чувстви-телност	*специфи-чност*	*точ-ност*	A_{ROC}	*чувстви-телност*	*специфи-чност*	*точ-ност*	A_{ROC}
1	0,840	0,890	0,865	0,930	0,700	0,800	0,750	0,828
2	0,850	0,890	0,875	0,924	0,650	0,850	0,750	0,875
3	0,820	0,890	0,855	0,937	0,800	0,600	0,700	0,780
4	0,820	0,860	0,840	0,921	0,750	0,850	0,800	0,854

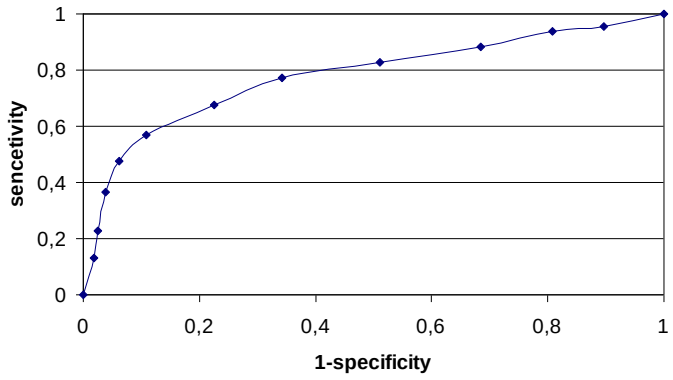
5	0,820	0,840	0,830	0,910	0,850	0,950	0,900	0,923
6	0,830	0,870	0,850	0,929	0,850	0,700	0,775	0,878
средна	**0,830**	**0,873**	**0,853**	**0,925**	**0,767**	**0,792**	**0,779**	**0,856**

Средните стойности на параметрите на моделите от вътрешното валидиране съвпадат с тези на първоначалния модел: 83% *чувствителност*, 87% *специфичност* и 85% *точност* и A_{ROC} = 0,925. Малко по-ниски са параметрите на тестовите групи: 77% *чувствителност*, 79% *специфичност* и 78% *точност* и A_{ROC} = 0,856.

Вътрешното валидиране на модела показа, че той има добра предсказваща способност, независима от броя и състава на обучаващата група.

5.1.3. Външно валидиране на модела

По-нататък началният модел беше използван за предсказване на алергенност на външна тестова група от 582 алергена и 582 неалергена (Приложение 1). Моделът разпозна 68% от алергените, 77% от неалергените и общо 73% вярно разпознати алергени и неалергени при разделителен праг 0,5 и два главни компонента. Кривата *чувствителност/1-специфичност* е показана на **фиг. 10** и има площ 0,785.



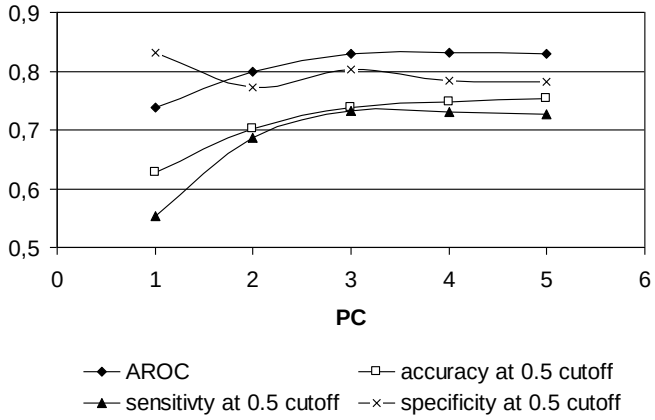
Фиг. 10. Крива с координати *чувствителност/1-специфичност* на началния модел, използван за предсказване на алергенност на външна тестова група от 582 алергена и 582 неалергена.

5.1.4. Разширен модел за разпознаване на алергени

За извеждането на разширен модел за разпознаване на алергени беше използвана цялата база данни от 702 алергена и 702 неалергена (Приложение 1). Структурата на протеините беше описана чрез трите *z*-дескриптора и трансформирана чрез ACC в низ от 45 променливи. Получената матрица беше подложена на дискриминантен анализ с вариращ брой на главните компоненти от 1 до 5. Моделите бяха оценени чрез параметрите *чувствителност*, *специфичност* и *точност* при праг 0.5. Отчетена беше и *площта под кривата* A_{ROC}**.** Резултатите са показани на **фиг.11**.

Резултатите показват, че най-високи стойности на параметрите се получават при 3 главни компонента. Моделът с 3 PC и VIP-стойностите на променливите са показани в Таблица 6. Три променливи имат ***VIP* > 1,2.** Това са **ACC333, ACC214 и ACC334**. Една от тях е с положителен коефициент (ACC214), другите две са с отрицателни (ACC333 и ACC334). Значимостта на

променливите ACC333 и ACC334, установена в началния модел, се потвърждава и тук. Появява се нова значима променлива ACC214, на мястото на променливите от предишния модел ACC324 и ACC223.

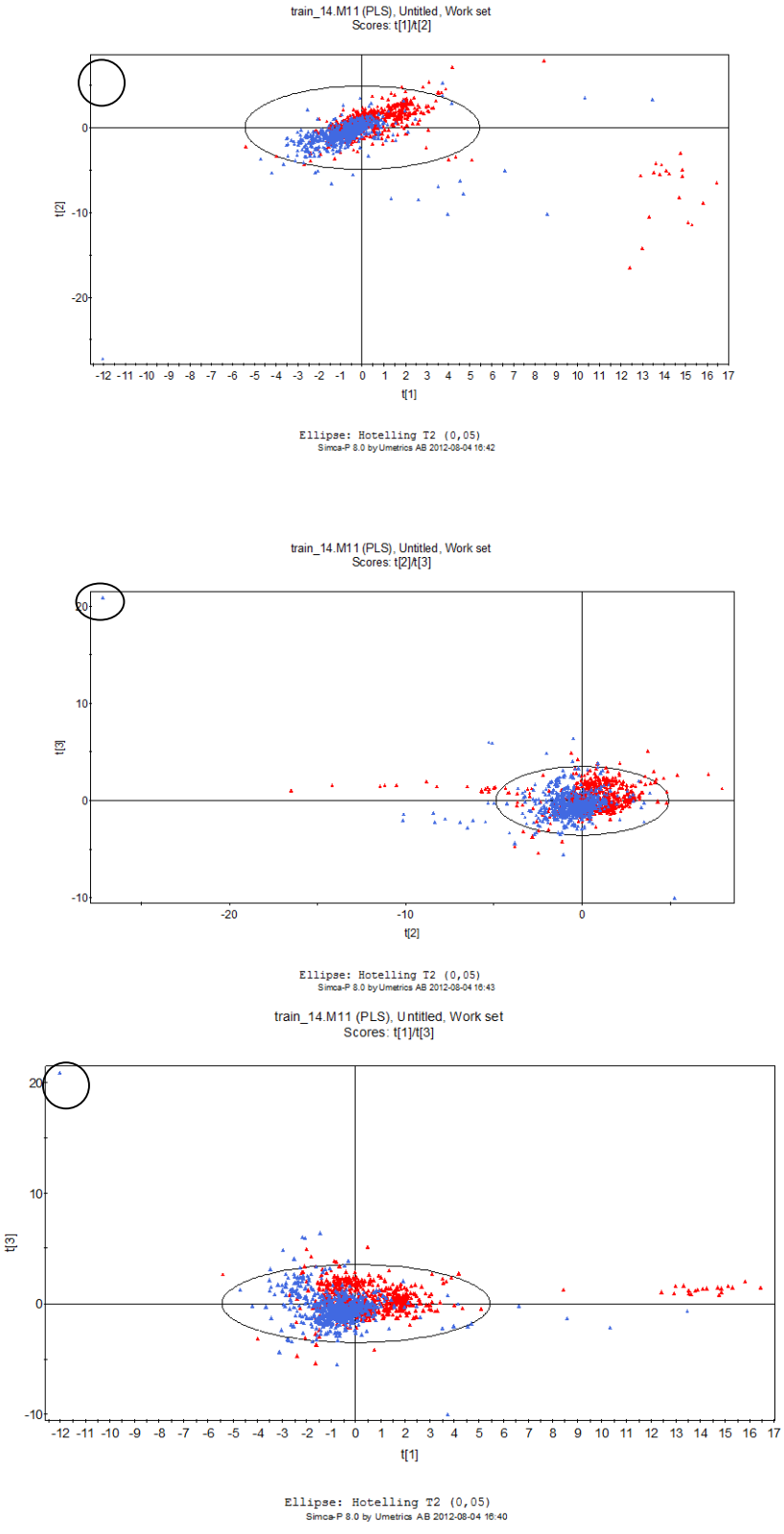


Фиг. 11. *Чувствителност*, *специфичност* и *точност* при праг 0.5 и *площ под кривата* A_{ROC} за разширения модел за разпознаване на алергени.

Таблица 6. Коефициенти и VIP стойности на променливите в разширения модел. Променливи с VIP > 1,2 са дадени в получерен шрифт. Константата на модела е 1.

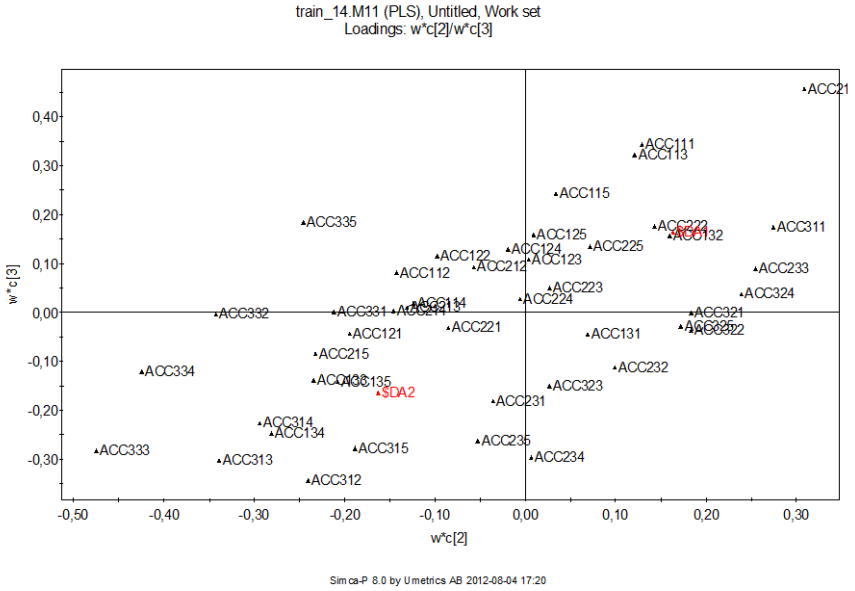
променлива	*VIP*	*коеф.*	*променлива*	*VIP*	*коеф.*
ACC333	**1,505**	**-0,158**	ACC211	1,044	-0,058
ACC214	**1,499**	**0,129**	ACC224	1,032	0,030
ACC334	**1,387**	**-0,124**	ACC311	1,015	0,107
ACC335	1,236	-0,030	ACC235	1,008	-0,074
ACC222	1,227	0,092	ACC233	1,000	0,059
ACC332	1,215	-0,085	ACC324	0,983	0,048
ACC215	1,190	-0,093	ACC124	0,979	-0,007
ACC313	1,149	-0,125	ACC314	0,964	-0,103
ACC121	1,105	-0,077	ACC221	0,963	-0,001
ACC225	1,101	0,066	ACC134	0,958	-0,102
ACC234	1,078	-0,062	ACC212	0,934	-0,020
ACC122	1,077	-0,028	ACC111	0,930	0,086
ACC312	1,054	-0,104	ACC125	0,929	0,007
ACC213	1,050	-0,054	ACC321	0,924	0,026
ACC223	1,049	0,042	ACC322	0,918	0,022

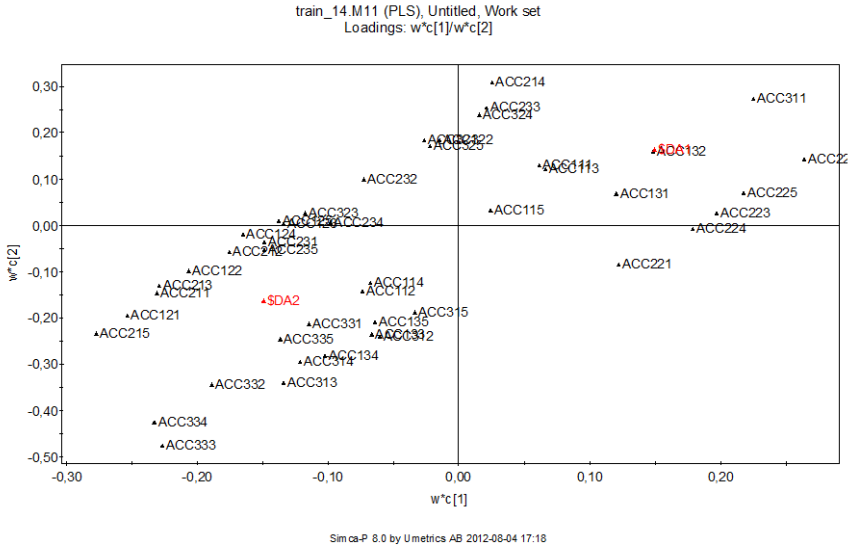
На **фиг. 12** са показани баловете (scores) на протеините в групата. Първите два компонента не успяват да обяснят дисперсия в групата (горе), което налага включването на трети компонент (в средата и долу). Много ясно се очертана протеин-беглец. Това неалергенът с GI: 315113274 (неалерген 609, Приложение 1). Който за удобство ще бъде махнат от втората група хранителни протеини подлежащи на изследване.

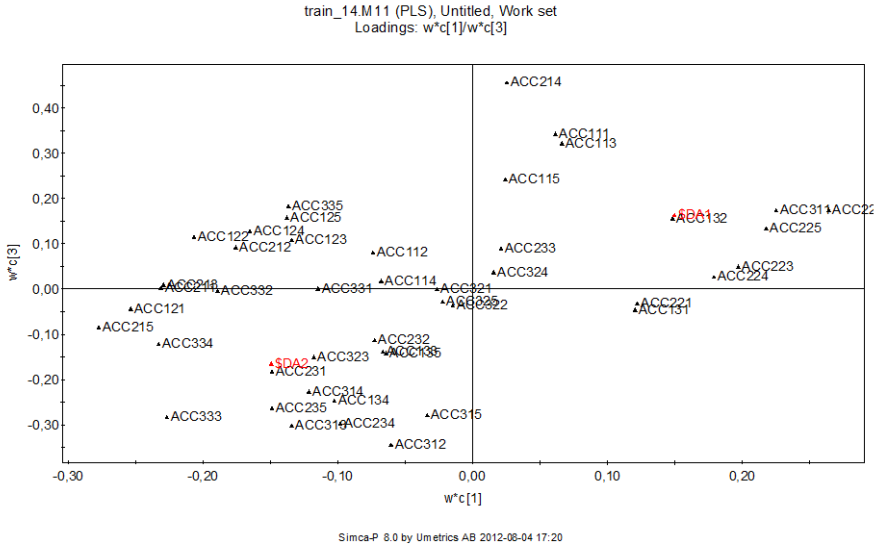


Фиг. 12. Балове (scores) на изследваните протеини. Горе с координати: PC2/PC1; в средата: PC3/PC2; долу: PC3/PC1. Алергените са показани в червено, неалергените – в синьо. Беглецът е заграден.

На **фиг. 13** са показани товарите (loadings) на ACC променливите.

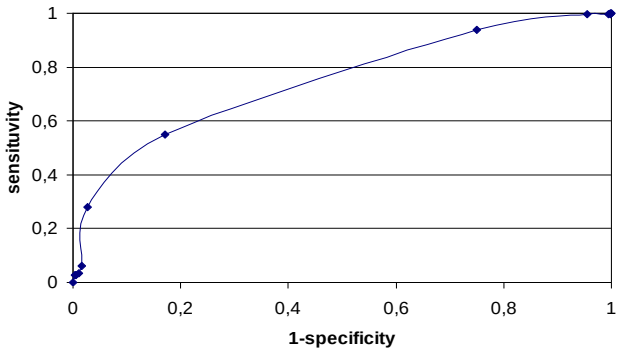






Фиг. 13. Товари (loadings) на АСС променливите. Горе с координати: PC2/PC1; в средата: PC3/PC2; долу: PC3/PC1.

Моделът разпознава 73% от алергените, 80% от неалергените или общо 77% вярно разпознати алергени и неалергени в изследваната група протеини. На **фиг.14** е показан модела с координати *чувствителност/1-специфичност* и площ под кривата A_{ROC} 0,830. Изключването на беглеца не подобри парапараметрите на модела.



Фиг.14. Крива *чувствителност/1-специфичност* на разширения модел за разпознаване на алергени.

5.1.5. Вътрешно валидиране на разширения модел

Разширеният модел за разпознаване на алергени с хранителен произход беше валидиран кръстосано в 7 групи. За целта групата от 702 алергена и 702 неалергена беше разделена на 7 групи по 100 (101) алергена и 100 (101) неалергена. Шест от групите бяха обединени в обучаваща група, седмата група беше тестова. Обучаващата група беше използвана за извеждане на модел, който беше валидиран с тестовата група. Процедурата беше повторена 7 пъти, така че всеки протеин участваше 6 пъти в обучаващата група и 1 път в тестовата. Получените модели бяха изведени при три главни компонента и бяха оценени по *чувствителност*, *специфичност* и *точност* на предикциите при праг 0,5 и по площ под ROC-кривата. Резултатите са показани в Таблица 7.

Таблица 7. Кръстосано валидиране в 7 групи на разширения модел

Получените резултати от вътрешното валидиране на разширения модел показват, че средните стойности на параметрите на моделите, получени от обучаващите подгрупи са близки до тези на

No.	*Обучаваща група n = 1202 (1203) праг 0,5*				*Тестова група n = 100 (101) праг 0,5*			
	чувстви-телност	*специфи-чност*	*точ-ност*	A_{ROC}	*чувстви-телност*	*специфи-чност*	*точ-ност*	A_{ROC}
1	0.554	0.827	0.691	0.741	0.525	0.820	0.672	0.729
2	0.687	0.774	0.731	0.795	0.762	0.720	0.741	0.792
3	0.561	0.810	0.686	0.730	0.530	0.832	0.682	0.722
4	0.551	0.824	0.687	0.735	0.530	0.802	0.667	0.692
5	0.699	0.779	0.739	0.811	0.680	0.720	0.700	0.758
6	0.581	0.824	0.703	0.750	0.560	0.800	0.680	0.760
7	0.686	0.782	0.734	0.809	0.630	0.810	0.720	0.772
средна	**0.617**	**0.803**	**0.710**	**0.767**	**0.602**	**0.786**	**0.695**	**0.746**

изходната група. *Специфичността* остава 80%, *чувствителността* намалява от 73% на 62% и оттам и *точността* намалява от 77% на 71%. Малко по-ниски са параметрите на тестовите групи: 60% *чувствителност*, 79% *специфичност*, 70% *точност* и A_{ROC} = 0,746.

Вътрешното валидиране на разширения модел показа, че той има добра и възпроизводима предсказваща способност, независеща от броя и състава на обучаващата група.

5.1.6. Дефиниране на мотив за алергенност

Доминирането на променливите ACC333 и ACC334 в двата модела показа, че те кодират свойства, особено важни за алергенността на протеините. И двете променливи имат отрицателни коефициенти в модела, което означава, че протеини с високи отрицателни стойности на двете променливи ще проявяват силна алергенност. Двете променливи са производни на z_3-дескриптора на аминокиселините в протеините. ACC333 е сума от произведенията на z_3-дескрипторите на ак с лаг 3, т.е. през една позиция (AxA). ACC334 е сума от произведенията на z_3-дескрипторите на ак с лаг 4, т.е. през две позиции (AxxA). За да се получи сума с отрицателен знак, произведенията трябва да бъдат с отрицателен знак, което означава, че z_3-дескрипторите на двете ак трябва да са с противоположни знаци (Таблица 8).

В Таблица 9 са групирани ак по знак на z_3-дескрипторите си. Колкото повече двойки ак от двете колони присъстват в даден протеин, толкова по-високи отрицателни стойности ще има този протеин за ACC333 и ACC334 и по-голяма вероятност да действа като алерген.

Таблица 8. Знаци на ACC333 и ACC334 в зависимост от знаците на z_3-дескрипторите на ак. Комбициите за алергенност са заградени.

i	+ z_3	- z_3

j		
+ z_3	+	-
- z_3	-	+

Таблица 9. Аминокиселини с положителни и отрицателни знаци на z_3-дескрипторите.

+ z_3	- z_3
C	R
D	K
P	T
H	V
W	Q
N	I
S	L
F	M
G	E
A	
Y	

За да илюстрираме с примери значението на променливите ACC333 и ACC334 за алергенността, избрахме двата протеина: единият с най-високи отрицателни, а вторият с най-високи положителни стойности на тези променливи. Както се очакваше, първият протеин беше алерген (Таблица 9), а вторият – неалерген (Таблица 10). От таблиците се вижда, че в структурата на алергена преобладават ACC333 и ACC334 с отрицателни знаци – 52% за ACC333 и 78% за ACC334 за алергена и съответно 13% и 18% за неалергена.

Комбинациите между аминокиселини с противоположни знаци на разстояние една и/или две позиции може да се разглежда като *мотив за алергенност*. Многократното присъствие на такива комбинации в структурата на един протеин е предпоставка за действието му като алерген.

Таблица 9. Знаци на ACC333 и ACC334 за алергена Q9S8D7 (*Triticum aestivum*).

SQQQQPPFSQQQPPFSQQQQPPFSQQQPPF			
ACC333		ACC334	
AxA	знак	AxxA	знак
SxQ	-	SxxQ	-
QxQ	+	QxxQ	+
QxQ	+	QxxP	-
QxQ	+	QxxP	-
QxP	-	QxxF	-
QxP	-	PxxS	+
PxF	+	PxxQ	-
PxS	+	FxxQ	-
FxQ	-	SxxQ	-
SxQ	-	QxxP	-

QxQ	+	QxxP	-
QxP	-	QxxF	-
QxP	-	PxxS	+
PxF	+	PxxQ	-
PxS	+	FxxQ	-
FxQ	-	SxxQ	-
SxQ	-	QxxQ	+
QxQ	+	QxxP	-
QxQ	+	QxxP	-
QxP	-	QxxF	-
QxP	-	PxxS	+
PxF	+	PxxQ	-
PxS	+	FxxQ	-
FxQ	-	SxxQ	-
SxQ	-	QxxP	-
QxQ	+	QxxP	-
QxP	-	QxxF	-
QxP	-		
PxF	+		
% +	38%		22%
% -	62%		78%

Таблица 10. Знаци на ACC333 и ACC334 за неалергена GI: 315113274 (*Triticum aestivum*).

MRAKWKKKRMRRLKRKRRKMRQRSK			
ACC333		ACC334	
AxA	знак	AxxA	знак
MxA	-	MxxK	+
RxK	+	RxxW	-
AxW	+	AxxK	-
KxK	+	KxxK	+

WxK	-	WxxK	-
KxK	+	KxxR	+
KxR	+	KxxM	+
KxM	+	KxxR	+
RxR	+	RxxR	+
MxR	+	MxxL	+
RxL	+	RxxK	+
RxK	+	RxxR	+
LxR	+	LxxK	+
KxK	+	KxxR	+
RxR	+	RxxR	+
KxR	+	KxxK	+
RxK	+	RxxM	+
RxM	+	RxxR	+
KxR	+	KxxQ	+
MxQ	+	MxxR	+
RxR	+	RxxS	-
QxS	-	QxxK	+
RxK	+		
% +	87%		82%
% -	13%		18%

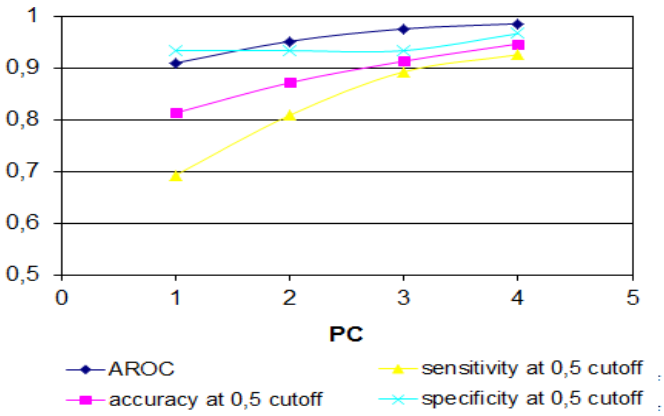
Анализът на други протеини обаче показа, че и сред алергените има представители с високи положителни стойности за ACC333 и ACC334, както и че сред неалергените също има протеини с високи отрицателни стойности за двете променливи. Това означава, че макар и важни за алергенността, тези две променливи не са достатъчни за да разпознаят един алерген. Участието и на другите променливи в модела също допринася за добрата му предсказваща способност.

В заключение може да се каже, че присъствието на мотив за алергенност е необходимо, но не достатъчно условие за проявяването на един протеин като алерген. Използването на целия модел е много по-надеждно при оценка на алергенността.

5.2 Резултати от втора група, съдържаща 700 алергена и 700 неалергена

5.2.1. Начален модел за разпознаване на алергени

За извеждането на начален за разпознаване на алергени отново беше създадена малка база данни от 120 алергена и 120 неалергена, избрани на случаен принцип сред всичките 1400 протеина, използвани в настоящата дисертация (Приложение 2). Структурата на протеините беше описана чрез пет *E*-дескриптора на изграждащите ги аминокиселини. Всеки протеин беше трансформиран в низ от 200 променливи, прилагайки *ACC*-трансформацията, както е описано в Глава 4 “Материали и методи”. Така беше получена изходна матрица от два класа протеини (алергени и неалергени) и 200 променливи, която беше подложена на дискриминантен анализ по метода PLS. Изведени бяха няколко модела с вариращ брой на главните компоненти от 1 до 4. Моделите бяха оценени чрез параметрите *чувствителност*, *специфичност* и *точност* при праг 0.5. Отчетена беше и *площта под кривата* A_{ROC}**.** Резултатите са показани на **фиг. 15.**



Фиг. 15. *Чувствителност*, *специфичност* и *точност* при праг 0.5 и *площ под кривата* A_{ROC} за началния модел за разпознаване на алергени от втора група, при различен брой главни компоненти .

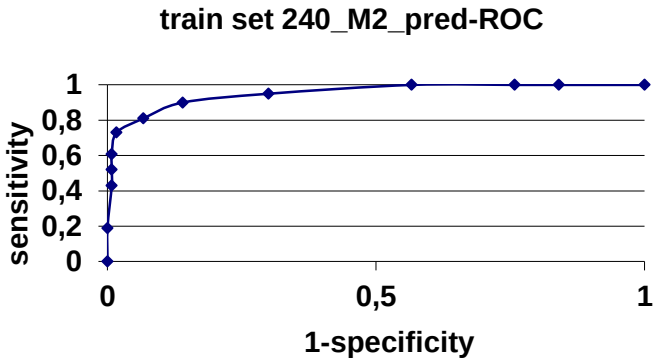
Резултатите показват, че добавянето на втори главен компонент значително повишава всички параметри на модела. По-нататъшното добавяне на главни компоненти също повишава параметрите. Изключение прави параметъра специфичност, който първоначално слабо се понижава при добавянето на втори компонент, след което слабо се повишава.

Променливите в модела бяха подредени по стойност на параметъра *VIP* (Таблица 12). Променливи с *VIP* > 1 имат съществено значение за модела. Шестдесет и пет променливи (32,5%) в модела имат *VIP* > 1. За да се отдиференцират най-съществените, прагът за *VIP* беше повишен на 2,0. Само 8 променливи имат *VIP* > 2,0. Това са **ACC121, ACC447, ACC444, ACC228, ACC222, ACC141, ACC243** и **ACC246** . Пет от тях са с положителни коефициенти (ACC121, ACC228, ACC222, ACC141 и ACC246), другите три са с отрицателни (ACC447, ACC444 и ACC243). Това означава, че протеини, имащи отрицателни стойности на променливите ACC447, ACC444 и ACC243 и положителни на ACC121, ACC228, ACC222, ACC141 и ACC246 е много вероятно да действат като алергени. Тази първоначална хипотеза по-нататък ще бъде потвърдена или отхвърлена при вътрешното и външното валидиране на модела.

Таблица 12. VIP стойности на променливите в модела. Променливи с VIP > 2,0 са дадени в получерен шрифт.

променлива	*VIP*	*коеф.*	*променлива*	*VIP*	*коеф.*
ACC121	**2,759**	**0,0229**	ACC518	1,635	-0,016
ACC447	**2,554**	**-0,0298**	ACC344	1,639	-0,016
ACC444	**2,448**	**-0,0304**	ACC335	1,605	0,0156
ACC228	**2,289**	**0,0383**	ACC427	1,589	-0,015
ACC222	**2,198**	**0,0383**	ACC353	1,587	-0,015
ACC141	**2,136**	**0,0401**	ACC118	1,583	-0,015
ACC243	**2,115**	**-0,020**	ACC147	1,566	-0,015
ACC246	**2,093**	**-0,0336**	ACC442	1,552	0,030
ACC323	1,961	-0,0191	ACC523	1,543	0,0149
ACC122	1,929	0,0187	ACC342	1,519	-0,015
ACC144	1,897	-0,0184	ACC418	1,492	-0,014
ACC128	1,799	-0,0175	ACC438	1,484	-0,014
ACC211	1,748	0,0169	ACC114	1,482	-0,014
ACC544	1,716	-0,0167	ACC251	1,468	-0,014
ACC324	1,710	-0,0166	ACC443	1,466	-0,014

Моделът беше тестван за *чувствителност*, *специфичност* и *точност* на предикциите при праг 0,5. Той разпознава 70% от алергените, 93% от неалергените и общо 82% вярно разпознати алергени и неалергени в изследваната група протеини. Моделът с кривата *чувствителност/1-специфичност* е показана на **фиг.17** и има площ 0,91.



Фиг. 17. Крива *чувствителност/1-специфичност* на началния модел за разпознаване на алергени

Моделът също беше валидиран вътрешно и външно. Получените резултати от вътрешното валидиране на разширения модел показват, че средните стойности на параметрите на моделите, получени от обучаващите подгрупи са близки до тези на изходната група. *Специфичността* остава 91%, *чувствителността* намалява от 71% на 69% и оттам и *точността* намалява от 82% на 79%. Малко по-ниски са параметрите на тестовите групи: 60% *чувствителност*, 83% *специфичност*, 75% *точност* и A_{ROC} = 0,806.

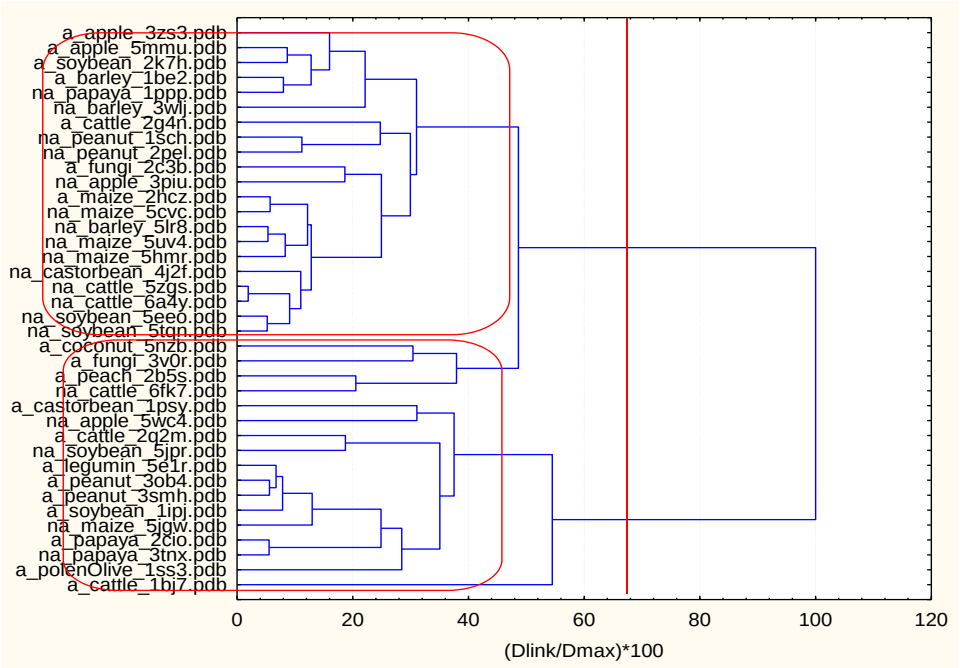
Вътрешното валидиране на разширения модел показа, че той има добра и възпроизводима предсказваща способност, независеща от броя и състава на обучаващата група.

5.3 Прилагане на кластерен анализ за разделяне на алергенни от неалергенни протеини

Беше извършен първоначален опит за експресно разделяне на алергенни от неалергенни протеини чрез КА. За целта бе подбран нов набор от дескриптори, съобразени с хидрофобността на аминокиселинните остатъци в протеините.

Бяха използвани данни от 20 експериментално определени скали за хидрофобност (s1 – s20, преизчислени с оригинална софтуерна техника до дескрипторни стойности). Бе създадена входна матрица от обучаваща серия от 38 протеина (19 алергена и 19 неалергена), описани от 98 дескриптора (избрани от общо 400) чрез редукция на променливите с анализ на главни компоненти. С набора от изходни данни бе проведен йерархичен и нейерархичен кластерен анализ. На **фиг. 23** е представена йерархична дендрограма за кластериране на 38 протеина с растителен произход, като нейерархичното кластериране с предварително условие за формиране на 2 кластера води до същия краен резултат.

Резултатите са показани на **фиг.23**.



Фиг. 23. Схема за алергични (a) / неалергични протеини с дескриптори за хидрофобност (HCA, метод на Ward, стандартизирани стойности, квадратни евклидови разстояния)

На **фиг.23** се вижда, че се постига задоволително разделяне между двете групи. Оформят се два големи кластера: K1 12(a) и 5 (na): "алергенен кластер": правилно класифициран 12/17 или 70%; неправилно класифицирани 5/17 или 30% (na като a)

K2 7(a) и 14(na) - "неалергенен кластер": правилно класифицирани 14/21 или 67%; неправилно класифицирани 7/21 или 33% (a като na).

Получените резултати почти достигат ефективността на класификацията чрез дискриминантен анализ – PLS.

6. Обсъждане на получените резултати

Алергенността на хранителните протеини е проблем от изключителна важност, свързан с все по-широкото навлизане в световната кухня на нови храни, хранителни добавки и подправки, много от които са с известен или неизвестен генно-модифициран произход. Алергенността не е явно, линейно кодирано в структурата на протеините, свойство, а конформационно-детерминирано и трудно предсказуемо свойство. Повечето от съществуващите в литературата и практиката методи за предсказване на алергенност се базират на структурното подобие на новите потенциални алергени с вече известни такива. Така появата на нов алерген не е възможно да бъде предсказана с тези методи.

В настоящата дисертация предлагаме подход за разпознаване на алергени, базиращ се на свойствата на аминокиселините, изграждащи протеини-алергени и неалергени с растителен хранителен произход. Протеините са трансформирани в линейни вектори и са анализирани с дискриминантен анализ по метода на частично най-малките парциални квадрати (PLS). Първоначално беше изведен модел върху малка група алергени и неалергени, след което базата данни беше разширена до 1404 протеина и беше изведен разширен модел за разпознаване на хранителни алергени. Моделът има добра предсказваща способност, разпознавайки 73% от алергените, 80% от неалергените или общо 77% вярно разпознати алергени и неалергени в изследваната група протеини. Въз основа на значението на променливите в модела е дефиниран мотив за алергенност, включващ аминокиселини с противоположни по знак z_3-дескриптори, разположени през една и две позиции в протеина. Процедурата беше повторена и при втора група

от протеини. Същата процедура бе приложена на 1400 протеини, за да се провери валидността на метода. Полученият модел също има добра предсказваща способност, разпознавайки 69% от алергените, 91% от неалергените или общо 79% вярно разпознати алергени и неалергени. Не беше дефиниран мотив за алергенност на втората група протеини тъй като тя бе избрана за тестване на точността на модела и процедурата му по извеждането му. Извършен бе и първоначален опит за разделяне на алергенни от неалергенни протеини чрез КА като възможност за експресно класифициране. Получените резултати почти достигат ефективността на класификацията чрез дискриминантен анализ – PLS.

Хранителните алергени имат разнообразна структура, състав и произход, което обуславя голямата дисперсия в групата. С увеличаване на броя на протеините в обучаващата група се увеличава и броя на главните компоненти, необходими за обяснение на тази дисперсия. В малката група протеини, използвана за извеждането на началния модел, два компонента бяха достатъчни за да се получи модел с добра предсказваща способност. В голямата група протеини, използвана при извеждането на разширения модел, се наложи включването на трети компонент. Моделът с трите главни компоненти имаше най-високи стойности за *чувствителност*, *специфичност* и *точност* на предикциите.

Отчитайки VIP-стойностите на първата група от хранителни протеини на 45-те променливи в модела, 42% от тях има значение за разпознаването на алергените. Това още веднъж доказва комплексността на свойството алергенност и невъзможността то да бъде предсказано еднозначно само с няколко структурни дескриптора.

Графиките с баловете (scores) на изследваните протеини (**Фиг. 12**) показват два кластера с ясно обособени ядра и дифузна периферия. Това означава, че моделът разграничава ясно алергените от неалергените, но има доста изключения. Общата точност на модела е 77%, т.е. 77% от протеините са вярно разпознати като алергени или неалергени, докато 23% са невярно разпознати.

Статистическите параметри на изведените модели варират в граници от 65% до 85% за *чувствителността*, от 60% до 95% за *специфичността* и от 70% до 90% за *точността*. Площите под кривите *чувствителност/1-специфичност* на всички изведени модели са над 0,828. Тези количествени параметри показват, че предсказващата способност на моделите е много добра, но не и отлична. Най-добрите публикувани в литературата модели за предсказване на алергенност работят с 90% и над 90% *точност* [117].

В дисертационния труд е направен опит да бъде дефиниран мотив за алергенност, базиращ се на доминиращото влияние на две от променливите в модела. Тези променливи отчитат електронните свойства на аминокиселините, кодирани в z_3-дескриптора. Комбинацията от ак с противоположни по знак z_3-дескриптори, намиращи се през една и/или две позиции в протеина е необходимо, но не и достаъчно условие за алергенност на изследвания протеин. Това, в съчетание с високите VIP-стойности на голяма част от променливите в модела, отново показва, че алергенността е скрито, комплексно свойство, зависещо от множество фактори, една част от които са кодирани в структурата на протеините.

Прилагането на КА също може да се използва за класифициране на протеините спрямо тяхната алергенност или неалергеност чрез използване на дескриптори от нов тип, базирани на скалите на хидрофобност на ак. Съответствие на получените резултати чрез йерархично кластериране се получава и при прилагане на нейерархичен подход за кластериране (K-means).

7. Приноси в дисертацияонния труд

Въз основа на резултатите, получени в дисертационния труд, могат да се дефинират няколко основни приноса:

1. Постигнато е ефективно класифициране на протеини с растителен произход на клас алергени и клас неалергени чрез модел, изведен в дисертацията, притежаващ добра предсказваща способност, разпознавайки 73% от алергените, 80% от неалергените или общо 77% вярно разпознати алергени и неалергени в изследваната група протеини.
2. Методът на прилагане на извеждане на модел за класифицирзне на протеини и проверка на неговата валидност е приложен и вурху втора група протеини с хранителен произход. Постигнато е също ефективен модел разпознавайки 69% от алергените, 91% от неалергените или общо 80% вярно предсказани алергени и неалергени в изследваната група.
3. Дефиниран е мотив за алергенност, който има важно значение за предсказване на алергенността, макар, че не е единственият показател на това свойство и не може да бъде използван самостоятелно.
4. Доказано е, че три главни компонента описват адекватно дисперсията (над 75%) в групата изследвани протеини, състояща се от 702 алергена и 702 неалергена.
5. Кластерният анализ също може да послужи за експресно разделяне на протеините в две групи на подобие – алергени и неалергени.

8. Публикации във връзка с дисертационния труд

В специализирани списания:

1. **L. H. Naneva,** I. D. Dimitrov**,** I. P. Bangov, I. A. Doytchinova, "Allergenicity prediction by artificial neural networks" J. Chemometrics 2014; 28:282-286, 2014. IF=1.5
2. **L. H. Naneva**, I. D. Dimitrov, I. P. Bangov, I. A. Doytchinova Title: Allergenicity Prediction by Partial Least Squares – based Discriminant Analysis Bulgarian Chemical Communications, 46 (2), 389-396, 2013. IF=0.31
3. **L. H. Naneva,** M.Nedyalkova, S.Madurga, F.Mas and V.Simeonov, Applying Discriminat and Cluster Analysis to Separate Allergenic from Non-allergenic Proteins, submitted to speical issue Open Chemistry, IF=1.425

Забелязани цитати на публикации:

1. Kyaw Z. Myin, Xiang-Qun X., "Ligand Biological Activity Predictions Using Fingerprint-Based Artificial Neural Networks (FANN-QSAR)" Artificial Neural Networks pp 149-164, 2014.

2. Scott McClain., "Bioinformatic screening and detection of allergencross-reactive IgE-binding epitopes" Molecular nutrition and food research Vol.61 61,8, 2017, 1600676

3. Alessandro Bitetto, Annarosa Mangone, Rosa Maria Mininni, Lorena C. Giannossa., "A nonlinear principal component analysis to study archeometric data" J. Chemometrics, Volume30, Issue7 Pages 405-415, 2016.

4. Kiran Kadam, Sangeeta Sawant, V.K. Jayaraman and Urmila Kulkarni-Kale., "Databases and Algorithms in Allergen Informatics" DOI: 10.5772/63083, 2016

5. Manju Bhardwaj ; Debasis Dash ; Vasudha Bhatnagar., "Accurate Classification of Biological Data Using Ensembles", IEEE Xplore, DOI: 10.1109/ICDMW.2015.229

Участия в научни форуми:

2018. 20th Nanoscience and Nanitechnoligy worckshop, Sofia 8-10/11/2018, „Relating Discriminant and Cluster Analysis to Predict Allergenicity of Food Proteins" poster

2018. BioCompChem, Банско 24–28/09/2018

2018. 22nd European Symposium on Quantitative Structure-Activity Relationships Translational and Health Informatics: Implications for Drug Discovery- Tessaloniki „Applying Discriminant and Cluster Analysis to Predict Allergenicity of Food Proteins", Poster

2018. 2nd International Conference & Expo on Green Chemistry and Engineering Barselona, Spain "Applying Discriminant and Cluster Analysis to Predict Allergenicity of Food Proteins"; Poster

2015. KONSTANTIN PRESLAVSKY UNIVERSITY OF SHUMEN FACULTY OF NATURAL SCIENCES DEPARTMENT OF BIOLOGY SECOND STUDENT SCIENTIFIC CONFERENCE"ECOLOGY AND ENVIRONMENT; Ivelina Stoqnova, L. H. Naneva; presentation"Predicting Biological Activity- Allergenicity Using Discriminant Analysis

2015 - XIII Национална конференция с международно участие „Природни науки '2015", Варна, постер, Marina Moskovkina, Ljudmila Naneva, Ivan Bangov "Prediction of Gas Chromatographyc retention behavior for saturated esters by QSRR approach"

2014 - XII Национална конференция с международно участие „Природни науки ' 2014", Варна, постер, Lyudmila Naneva1, Ivan Dimitrov2, Ivan Bangov1, Irini Doychinova2 "Application of machine learning techniques for allergenicity prediction in the QSAR Studies"

2014 - Oxford Drug Discovery Workshop July21-25, 2014-poster Lyudmila Naneva,Ivan Dimitrov, Ivan Bangov, Irini Doychinova "Application of machine learning techniques for allergenicity prediction in the QSAR Studies"

2013 - 5ти Международен Симпозиум „Methods and Applications of Computational Chemistry", 1-5.07.2013, Украина, гр. Харков, M. Moskovkina, l. Naneva, I. Bangov, „Gas Chromatographic Retention Modeling by QSRR Approach" (ПОСТЕР)

2013- Conferenciq Chemometrica, Sopron, Hungary 8-11.09, 2013 (Постер)

Lyudmila Naneva, Ivan Dimitrov, Ivan Bangov, Irini Doychinova, Marina Moskovkina , "Employment of Descriptor Fingerprints in the QSAR Studies"

2013 - Conferencia Modeling Interactions in Biomolecules VI, MIB'13, MariansKe Lazne, Chechen, 16-19.09, (Доклад), Ivan Bangov, Irini Doychinova, Nikolaj Kochev, Marina Moskovkina , Ivan Dimitrov, Lyudmila Naneva, Veselina Paskaleva, Use of the Descriptor Fingerprints in the QSAR and QSPR Investigations"

Участия в международни и национални научни проекти

1. Университетски проект договор № РД-05-321/9.03.2010, финансиран от фонд “Научни изследвания” на ШУ“Еп. Константин Преславски”:"Научни и приложни аспекти при изследване на веществата и процеса на обучение по Химия и опазване на околната среда"
2. Университетски проект №РД-05-134/24.02.2011, финансиран от фонд “Научни изследвания” на ШУ“Еп. Константин Преславски”: "Интегративни връзки наука - образование при изследване на веществата и обучението по химия и опазване на околната среда".
3. Национален проект проект вх. № ДФНИ- 401 /7 / от 28.11.2012:„Разработка на нови методи и софтуер за химическата и фармацевтичната промишленост за компютърно предсказване на нови химически съединения с определени физикохимични свойства или с биологична активност базирани на нови алгоритми и паралелна обработка на информацията”
4. РД-08-264/14.03.2013г. Университетски проект "Научни и практикоприложни подходи при изследване на химичните обекти и процеса на обучение", 2013
5. Университетски проект № BG051PO001-3.3.06 “Подкрепа за развитието на докторанти, постдокторанти, специализанти и млади учени”
6. Университетски проект договорРД-08-218/10.03.2014.финансиран от фонд “Научни изследвания” на ШУ“Еп. Константин Преславски”
7. Проект BG-051 PO 001/3.3.07-0002 Студентски практики
8. ПРОЕКТBG51PO001-3.3.06-0003“Подкрепа за развитието на докторанти, постдокторанти, специализанти и млади учени”
9. Проект BG05M2OP001-2.009-0028 „ Постигане на оптимална среда за развитие, научни изследвания, иновациии устойчиво развитие на човешкия капитал в сферата на химическите науки: адаптиране на образованието днес за утрешния ден“
10. ФНИ-НИС – докторантски проекти, 80-10-89/2018, Оценка на хранителни алергии чрез хемометричен и QSAR анализ

Литература:

1. Sampson HA. Food allergy. Part 2: diagnosis and management. J Allergy Clin Immunol

1999;103(6):981–9.
2. Sampson HA. Food allergy. Part 1: immunopathogenesis and clinical disorders. J Allergy Clin Immunol 1999;103(5):717–28.
3. Sampson H. Food allergy: when mucosal immunity goes wrong. J Allergy Clin Immunol 2005;115:139–41.
4. Broadfield E., McKeever T.M., Scrivener S., Venn A., Lewis S.A., Britton J. Increase in the prevalence of allergen skin sensitization in successive birth cohorts. J. Allergy Clin. Immunol. 2002;109:969–974.
5. Goodman R.E., Hefle S.L., Taylor S.L., Ree R.V. Assessing genetically modified crops to minimize the risk of increased food allergy: a review. Int. Arch. Allergy Immunol. 2005;137:153–166.
6. Midoro-Horiuti T, Goldblum R, Kurosky A, et al. Isolation and characterization of the mountain cedar (Juniperus ashei) pollen major allergen, Jun a 1. J Allergy Clin Immunol 1999;104:608–12.
7. Scheurer S, Son DY, Boehm M, et al. Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. Mol Immunol 1999;36(3):155–67.
8. Rabjohn P, Helm EM, Stanley JS, et al. Molecular cloning and epitope analysis of the peanut allergen Ara h 3. J Clin Invest 1999;103(4):535–42.
9. Midoro-Horiuti T, Goldblum R, Kurosky A, et al. Isolation and characterization of the mountain cedar (Juniperus ashei) pollen major allergen, Jun a 1. J Allergy Clin Immunol 1999;104:608-22.
10. Wal JM. Bovine milk allergenicity. Ann Allergy Asthma Immunol 2004;93(5):S2–11.
11. Wulfert F, Sanyasi E, Watanabe LA. Prediction of tolerance in children with IgE mediated cow's milk allergy by microarray profiling and chemometric approach; Journal of Immunological Methods, Volume 382, Issues 1–2, 31 August 2012, Pages 48-57
12-14. Natale M, Bisson C, Monti G, et al. Cow's milk allergens identification by two-dimensional immunoblotting and mass spectrometry. Mol Nutr Food Res 2004;48(5):363–9.
15. Mine Y, Rupa P. Immunological and biochemical properties of egg allergens. Worlds Poult Sci J 2004;60(3):321–30.
16. Van Do T, Hordvik I, Endresen C, et al. Characterization of parvalbumin, the major allergen in Alaska pollack, and comparison with codfish Allergen M. Mol Immunol 2005;42(3):345–53.
17. Swoboda I, Bugajska-Schretter A, Verdino P, et al. Recombinant carp parvalbumin, the major cross-reactive fish allergen: a tool for diagnosis and therapy of fish allergy. Allergy 2002;57:80.
18. Swoboda I, Bugajska-Schretter A, Valenta R, et al. Recombinant fish parvalbumins: candidates for diagnosis and treatment of fish allergy. Allergy 2002;57:94–6.
19. Beardslee TA, Zeece MG, Sarath G, et al. Soybean glycinin G1 acidic chain shares IgE epitopes with peanut allergen Ara h 3. Int Arch Allergy Immunol 2000;123(4): 299–307.
20. Helm RM, Cockrell G, Connaughton C, et al. A soybean G2 glycinin allergen-1. Identification and characterization. Int Arch Allergy Immunol 2000;123(3):205–12.
21. Rabjohn P, Burks AW, Sampson HA, et al. Mutational analysis of the IgE-binding epitopes of the peanut allergen, Ara h 3: a member of the glycinin family of seed-storage proteins. J Allergy Clin Immunol 1999;103(1):S101.
22. Taylor S.L. Protein allergenicity assessment of foods produced through agricultural biotechnology. Annu. Rev. Pharmacol. Toxicol. 2002;42:99–112.
23. Lee Y.H., Sinko P.J. Oral delivery of salmon calcitonin. Adv Drug Deliv Rev. 2000;42:225–238.
24. Soltero R., Ekwuribe N. The oral delivery of protein and peptide drugs. Innovat. Pharmaceut. Technol. 2002;1:106–110.
25. FAO/WHO Codex Alimentarius Commission: Codex Principles and Guidelines on Foods Derived from Biotechnology. Joint FAO/WHO Food Standards Programme. Rome, Italy; 2003
26. Rinkesh Kumar Gupta, K.Gupta2018 Maillard reaction infood allergy:Pros and cons, Critical Reviews in Food Science and Nutririon, 58:2,208-226

27. Gendel SM. Bioinformatics and food allergens. J AOAC Int 2004;87:1417–22.
28. Glaspole IN, de Leon MP, Rolland JM, et al. Characterization of the T-cell epitopes of a major peanut allergen, Ara h 2. Allergy 2005;60:35–40.
29. BreitenederH,Mills ENC. Molecular properties of food allergens. J Allergy Clin Immunol 2005;115:14–23.
30. Cooper PJ: Intestinal worms and human allergy. Parasite Immunol 2004, 26:455-467.
31. Janeway CA, Travers P, Walport M, Capra JD: Immunobiology: the immune system in health and disease. London: Current Biology Publications; 1999
32. Rusznak C, Davies RJ: ABC of allergies. Diagnosing Allergy. BMJ 1998, 316:686-689.
Huby RDJ, Dearman RJ, Kimber I: Why are some proteins allergens. Toxicological Sci 2000, 55:235-246.
33. FAO/WHO Agriculture and Consumer Protection: Evaluation of Allergenicity of Genetically Modified Foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. Rome, Italy; 2001.
34. FAO/WHO Codex Alimentarius Commission: Codex Principles and Guidelines on Foods Derived from Biotechnology. Joint FAO/WHO Food Standards Programme. Rome, Italy; 2003.
35.Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. FASEB Journal 2003;17(6):1141–3.
36. Ivanciuc O, Schein CH, Braun W: SDAP: database and computational tools for allergenic proteins. Nucleic Acids Res 2003, 31:359–362.
37. Fiers MWEJ, Kleter GA, Nijland H, Peijnenburg AACM, Nap JP, vanHam RCHJ: Allermatch, a webtool for the prediction of potential allergenicity according to current fao/who codex alimentarius guidelines. BMC Bioinformatics 2004, 5:133.
38. Zhang ZH, Koh JL, Zhang GL, Choo KH, Tammi MT, Tong JC: AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. Bioinformatics 2007, 23:504-506.
39. Hileman,R.E., Silvanovich,A., Goodman,R.E., Rice,E.A.,Holleschak,G., Astwood,J.D. and Hefle,S.L. (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int. Arch. Allergy Immunol., 128, 280–291.
40. Bjorklund,A.K., Soeria-Atmadja,D., Zorzet,A., Hammerling,U. andbGustafsson,M.G. (2005) Supervised identification of allergenrepresentative peptides for in silico detection of potentially allergenicbproteins. Bioinformatics, 21, 39–50.
41. Gendel,S.M. (2002) Sequence analysis for assessing potential allergenicity. Ann. NY Acad. Sci., 964, 87–98.
42. Kleter,G.A. and Peijnenburg,A.A. (2002) Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE-binding linear epitopes of allergens. BMC Struct. Biol., 2, 8.
43. Li,K.B., Issac,P. and Krishnan,A. (2004) Predicting allergenic proteins using wavelet transform. Bioinformatics, 20, 2572–2578.
44. Silvanovich,A., Nemeth,M.A., Song,P., Herman,R., Tagliani,L. And Bannon,G.A. (2006) The value of short amino acid sequence matches for prediction of protein allergenicity. Toxicol. Sci., 90, 252–258.
45. Bailey,T.L.,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.
46. Kaufman,L., and Rousseeuw,P.J.(1990). Finding Groups in Data:An Inntroduction to Claster Analysis ,John Wiley-Sons,Brussels,Belgium.
47. Kumar M., Verma R., Raghava G.P.S. Prediction of mitochondrial proteins using support vector machine and hidden markov model. J. Biol. Chem. 2005;281:5357–5363
48. Björklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG: Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. Bioinformatics 2005, 21: 39–50.
49. S.Negi, S.Werner(2017).Cross – React:Anew structural bioinformatics method for predicting allergen cross – reactivity. Bioinformatics(Oxford, England).33.10.1093/bioinformatics/btw767

50. Lu, W., Negi, S. S., Schein, C. H., Maleki, S. J., Hurlburt, B. K., & Braun, W. (2018). Distinguishing allergens from non-allergenic homologues using Physical–Chemical Property (PCP) motifs. Molecular Immunology, 99, 1–8. doi:10.1016/j.molimm.2018.03.022.
51. Saha S, Raghava GPS: AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res 2006, 34:W202-W209.
52. Mirsha, Ankita, Neveen2015. Mappinig B – cell epitopes of major and minor peanut allergens and identifying resideus contributing to IgE binding. Jornal of the Science of Food and Agriculture96,10.1002/jsfa.7121
53. Doytchinova,I.A. and Flower,D.R. (2003) Towards the in silico identification of class
II restricted T-cell epitopes: a partial least squares iterative selfconsistent algorithm for affinity prediction. Bioinformatics, 19, 2263–2270.
54. Furmonaviciene R, Sutton BJ, Glaser F, Laughton CA, Jones N, Sewell HF, Shakib F: An attempt to define allergen-specific molecular surface features: a bioinformatic approach. Bioinformatics 2005, 21:4201-4204.
55. Seong SY, Matzinger P: Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses. Nat Rev Immunol 2004, 4:469.
57. Hansch C., Quantitative structure-activity relationships in drug design, in E.J.ariens (Ed.), Drug Design, Vil.1, academic press, New York, 1971, p.271.
58.Wold, S.: PLS for Multivariate Linear Modeling. In: Chemometric Methods in Molecular Design. (Ed. H. van der Waterbeemd), VCH, Weinheim, 195-218, 1995.
59. Wold, S., M. Sjöström: Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In Chemometrics: Theory and Application, vol. 52, (Ed. B.R. Kowalski), ACS Symposium Series, 243-282, 1977.
60. Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S: DNA and Peptide Sequences and Chemical Processes Multivariately Modelled by Principal Components Analysis and Partial Least Squares Projections to Latent Structures. Anal Chim Acta 1993, 277:239-253.
61. L.Eriksson,E.Johansson, N.Kettaneh-Wold , J.Trygg, C.Wikstrom, S.Wold,-Multi-and Megavariate Data Analtsis Umetrics AB,Umea, Sweden, 2006 pp.85-87.
62. Simeonov V. Classification: Encyclopedia of environmetrics. – J.Wiley & Sons, New York, 2002.
63. Aggarwal Charu C., Data Mining: The Textbook, Springer-Verlag, 2015
64. Feldman Ronen, James Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, Dec 11, 2006.
65. Leskovec Jure, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Datasets, Stanford InfoLab, 2014.
66. Alam,S., Dobbie, G., & Rehman, S.U. (2015). Analysis of particle swarm optimization
based hierarchal data clustering approaches. Swarm and Evolutionary
Computation, 25, 36-51.
67. K.H.Esbensen.,Multivariate Data analysis I Practice 4-th Ed.CAMO.2000
68. Cortes, Corinna; and Vapnik, Vladimir N.;"Support-Vector Networks", Machine Learning, 20, 1995.
69. A.I. belousov, S.A.Verzakov, J.von Frese. Aplicational aspect of support vector machines. J.Chemom. 16,428(2002).
70. Muh HC, Tong JC, Tammi MT: AllerHunter: A SVM-Pairwise System for Assessment of Allergenicity and Allergic Cross-Reactivity in Proteins. PLoS ONE 2009, 4:e5861.
71. Jorgensen WL (1991). "Rusting of the lock and key model for protein-ligand binding". Science 254 (5034): 954–5.
96. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic
proteins. Nucleic Acids Res 2003;31(1):359–62.
97. Altschul, S. F.,Gish, W., Miller,W.,Myers,E.W., and Lipman,D.J.1990-Basic local alignment search tool.J.Mol.Biol.215:403-410.
102. Lu, Wenzhe,B.Werner(2018). Distingishing allergens from non-allergenic homologues using Phisical – chemical Propertiy motifs. Molecular imunilogy.99.1-8.10.1016/j.molimm.2018.03.022

103. Hellberg S, Sjöström M, Skagerberg B, Wold S: Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem 1987, 30:1126-1135.
104. SandbergM, Eriksson L, Jonsson J, Sj¨ostr¨omM,Wold S. Newchemical descriptors relevant for the design of biologically active peptides A multivariate characterization of 87 amino acids. J Med Chem 1998;41:2481
105. Siebert KJ: Quantitative structure-activity relationship modelling of peptide and protein behavior as a function of amino acid composition. J Agr Food Chem 2001, 49:851-858.
106. Lapinsh,M. et al. (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim. Biophys. Acta, 1525, 180–190.
107. Collis AVJ, Brouwer AP, Martin ACR. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. J Mol Biol 2003;325:337.
108. Nyström Å, Andersson PM, Lundstedt T: Multivariate data analysis of topographically modified á-melanotropin analoques using auto and cross auto covariances (ACC). Quant Struct-Act Relat 2000, 19:264-269.
109. SIMCA-P 8.0. Umetrics UK Ltd., Wokingham Road, RG42 1PL, Bracknell, UK.
110. Floyd RW: Algorithm 97 Shortest Path. Commun ACM 1969, 12:345-346
111. Bradley AP: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 1997, 30:1145-1159.
112. Baldi P., Brunak S., Chauvin Y., Andersen C.A., Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16:412–424.
113. Eriksson,E.Johansson, N.Kettaneh-Wold , J.Trygg, C.Wikstrom, S.Wold,-Multi-and Megavariate Data Analtsis Umetrics AB,Umea, Sweden, 2006 pp.110--112.
114. Monti S. A Computational Method Used 3d-QSAR: Comparative Molecular Field Analysis (CoMFA)Molecular design and combinatorial chemistry: Selcted methods and applications, 1998
115. SAS Institute Inc. ftp://ftp.sas.com/pub/neural/FAQ3.html#A_cross.
116. Soeria-Atmadja D., Wallman M., Bjorklund A.K., Isaksson A., Hammerling U., Gustafsson M.G. External cross-validation for unbiased evaluation of protein family detectors: application to allergens. Proteins. 2005;61:918–925.
117. Dimitrov I., Flower D.R., Doytchinova, I. AllerTOP – a bioinformatic tool for allergenicity prediction. Bioinformatics, in press, 2012.
118. James, D. W.; Armishaw, R. F.; Frost, R. L. J. Phys. Chem. 1976, 80,1346.
119. Hallenga, K.; Grigera, J. R.; Berendsen, H. J. J. Phys. Chem. 1980, 84, 2381.
120. Nakanishi, K.; Ikari, K.; Okazaki, S.; Touhara, H. J. Chem. Phys. 1984,80, 1656.
121. Maeda, Y.; Tsukida, N.; Kitano, H.; Terada, T.; Yamanaka, J. J. Phys.Chem. 1993, 97, 13903.
122. Ide, M.; Maeda, Y.; Kitano, H. J. Phys. Chem. B 1997, 101, 7022.
123. Hechte, D.; Tadesse, F.; Walters, L. J. Am. Chem. Soc. 1993, 115, 3336.
124. Bagno, A.; Lovat, G.; Scorrance, G.; Lijuen, J. W. J. Phys. Chem. 1993,97, 4601.
125. Hirata, F. Bull. Chem. Soc. Jpn. 1998, 71, 1483.
126. Kitano, A.; Hirata, F.; Go, M. J. Phys. Chem. 1993, 97, 10223.
127. Nozaki, N.; Tanford, C. J. J. Biol. Chem. 1971, 246, 2211.
128. Levitt, M. J. Mol. Biol. 1976, 104, 59.
129. Hopp, S.; Woods, K. S. Proc. Natl. Acad. Sci. U.S.A. 1981, 78, 3824.
130. Kyte, J.; Doolittle, R. J. Mol. Biol. 1982, 157, 105.
131. Sweet, R. M.; Eisenberg D. J. Mol. Biol. 1983, 171, 479.
132. Fauchere, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Int. J. Pept. Protein Res. 1988, 32, 269.
133. Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Zehfus, M. H. Science 1985, 229, 834.
134. Miyazawa, S.; Jernigan, R. L. Macromolecules 1985, 18, 534.
135. Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. J. Protein Chem.1985, 4, 23.
136. Parker, J. M. R.; Guo, D.; Hodges, R. S. Biochemistry 1986, 25, 5425.
137. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.

138. Engelman, D. M.; Steitz, T. A.; Goldman, A. Annu. ReV. Biophys. Chem. 1986, 15,321.
139. Skagerberg, B.; Wold, S.; Andrews, P. Int. J. Pept. Protein Res. 1991, 37, 414.
140. Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Proc. Natl. Acad.Sci. U.S.A. 1987, 84, 3086.

Използвани съкращения

ак – аминокиселина
ДА – дискриминантен анализ
КА – кластерен анализ
ANN – изкуствени невронни мрежи
AROC – площ под кривата чувствителност/1-специфичност
FN – положителни съединения, предсказани като отрицателни
FP – отрицателни съединения, предсказани като положителни
PC – главен компонент
PCA – анализ на главните компоненти
PLS – метод на частично най-малки квадрати
ROC – оценка на предсказващата способност на модели, получени чрез ДА
SVM – машини с поддържащи вектори
TN – отрицателни съединения, предсказани като отрицателни
TP – положителни съединения, предсказани като положителни
VIP – влияние на променливата върху проекцията (модела)