# ПРЕДСКАЗАНИЕ АЛЛЕРГЕННОСТИ БЕЛКОВ С ИСПОЛЬЗОВАНИЕМ ИНФОРМАЦИИ О КОНФОРМАЦИОННЫХ ПЕПТИДАХ

**А.О. Брагин, П.С. Деменков, В.А. Иванисенко**

Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия, e-mail: ibragim@bionet.nsc.ru

В настоящее время аллергия является одной из наиболее часто встречаемых проблем со здоровьем в развитых странах. Предсказание аллергенности белков по их аминокислотным последовательностям обладает ограниченной точностью в связи с пространственной организацией поверхностных районов белка, несущих антигенные эпитопы. Предложен новый метод предсказания белков-аллергенов на основе анализа конформационных пептидов, моделирующих поверхность белка.

**Ключевые слова:** аллергены, конформационные пептиды, пространственная структура белка, предсказание аллергенности белков.

## Введение

Аллергия встречается более чем у 20 % населения развитых индустриальных стран (Casolaro *et al*., 1996). Предсказание потенциальных аллергенов по аминокислотным последовательностям и пространственным структурам белков является важной задачей биоинформатики. Одним из первых *in silico*-методов предсказания аллергенов был предложен всемирной организацией здравоохранения (WHO) и продовольственной и сельскохозяйственной организацией (FAO) (FAO/WHO, 2003). Согласно этому методу, белок считается аллергеном, если его последовательность из 80 аминокислотных остатков имеет гомологию выше 35 % с одним из известных аллергенов, или участок анализируемого белка протяженностью как минимум 6 аминокислотных остатков идентичен участку аминокислотной последовательности известного аллергена.

Кроме метода, предложенного FAO/WHO, существует еще ряд методов анализа аллергенности белков с использованием информации по аминокислотной последовательности. Например, A. Zorzet разработал подход предсказания аллергенности с использованием алгоритма выравнивания анализируемой последовательности с последовательностями аллергенов при помощи FASTA3 и последующей классификации при помощи метода k-ближайших соседей (kNN) (Zorzet *et al*., 2002). Дальнейшее улучшение метода было достигнуто в работе D. Soeria-Atmadja, были предложены новые методы классификации, основанные на байесовском подходе (Bayesian linear Gaussian classifier, Bayesian quadratic Gaussian classifier) (Soeria-Atmadja *et al*., 2004).

В дополнение к методу FAO/WHO было предложено использовать поиск мотивов в анализируемом белке, характерных для белков-аллергенов (Stadler M., Stadler B., 2003). Позже W. Kong с коллегами показали, что поиск множественных мотивов в анализируемом белке увеличивает точность предсказания аллергенности (Kong *et al*., 2007).

Для предсказания аллергенности была создана база паттернов, характерных для белков-аллергенов, полученных с использованием методов вейвлетовых преобразований и скрытых марковских моделей (Li *et al*., 2004). В дополнение к предсказанию аллергенности, основанному на распознавании мотивов, этими же авторами было предложено использовать поиск сходства последовательностей по наборам белков-аллергенов с помощью BLAST.

В работе (Saha, Raghava, 2006a) использовались метод опорных векторов и поиск эпитопов IgE. H. Muh с коллегами (Muh *et al*., 2009) создали компьютерную программу AllerHunter, которая для предсказания аллергенности использует выравнивание анализируемой последовательности с известными последовательностями белков-аллергенов совместно с методом опорных векторов.

Как известно, медиаторы воспаления выделяются, когда IgE, расположенные на поверхности тучных клеток или базофилов, контактируют с аллергеном (Sutton *et al*., 1993). Участки поверхности антигена (эпитопы), с которыми взаимодействует IgE, могут быть как линейные, так и конформационные (Schramm *et al*., 2001; Takagi *et al*., 2005). Линейные эпитопы представляют собой непрерывный участок в аминокислотной последовательности, в то время как конформационные эпитопы формируются аминокислотными остатками, распределенными в разных местах последовательности. При этом конформационные эпитопы, так же как и линейные, образуют компактные области на поверхности белка, с которыми могут взаимодействовать антитела.

Большинство описанных выше методов явно или неявно используют информацию о линейных антигенных эпитопах. Например, такие эпитопы могут входить в состав мотивов или консервативных участков аминокислотных последовательностей белков, используемых при предсказании аллергенности. Однако конформационные эпитопы учесть при анализе первичной структуры значительно сложнее в силу того, что они могут быть распределены вдоль протяженных участков последовательности. Для их анализа требуется использование данных о пространственной структуре белков.

В настоящее время предложено несколько методов предсказания конформационных эпитопов при наличии известной пространственной структуры антигена (Kulkarni-Kale *et al*., 2005; Ponomarenko *et al*., 2008; Liang *et al*., 2009; Sun *et al*., 2009). Необходимость наличия пространственной структуры белков является серьезным ограничением для широкого использования этих методов. Для многих белков пространственная структура не известна и не может быть предсказана по гомологии, в частности для мембранных белков. Однако задача предсказания конформационных эпитопов по данным только первичной структуры белка остается пока нерешенной.

Один из возможных путей решения проблемы учета информации о конформационных эпитопах в методах предсказания аллергенности белков только по их первичной структуре может состоять в выявлении линейных участков в последовательности, способных мимикрировать конформационные эпитопы различных белков.

Такого рода мимикрия конформационных эпитопов линейными пептидами была показана при использовании фагового дисплея (Smith, 1985). Было обнаружено, что многие моноклональные антитела, связывающиеся с конформационными эпитопами антигена, также обладают способностью связываться с искусственными линейными пептидами. Оказалось, что аминокислотный состав таких линейных пептидов и соответствующих им конформационных эпитопов в значительной степени совпадал. Кроме того, последовательности аминокислот линейных пептидов представляли собой последовательные цепочки сближенных в третичной структуре белка аминокислот конформационных эпитопов. Такие цепочки аминокислот, сближенных в третичной структуре, но удаленных друг от друга в первичной структуре белка, можно назвать конформационными пептидами по аналогии с конформационными эпитопами. Ранее нами был разработан подход к предсказанию конформационных эпитопов в белках на основе поиска сходства между конформационными пептидами и пептидами, полученными методом фагового дисплея, обладающими способностью специфично связываться с моноклональными антителами. Данный подход был применен нами для идентификации конформационных эпитопов в белках ряда вирусов (Локтев и др., 2002; Туманова и др., 2002).

В настоящей работе нам было интересно выяснить, может ли информация о сходстве между аминокислотными последовательностями анализируемого набора белков и конформационными пептидами, представленными на поверхности различных белков аллергенов, увеличить точность предсказания аллергенности белков этого анализируемого набора.

Полученные результаты показали различимое увеличение точности предсказания аллергенности по сравнению с методом, использующим только сходство фрагментов последовательностей между известными белками-аллергенами и анализируемыми белками.

Можно предположить, что наличие совпадений между конформационными пептидами у разных аллергенов может свидетельствовать о том, что данные конформационные пептиды являются частью потенциальных конформационных эпитопов, существенных для проявления аллергенных свойств белков. Предложенный подход может быть использован для увеличения точности методов предсказания аллергенности белков, а также создания новых методов предсказания потенциальных конформационных эпитопов.

## Метод предсказания аллергенности белков

Для предсказания аллергенности белков использовался подход, основанный на расчете меры сходства между анализируемым белком и известными белками-аллергенами (обучающая выборка). Сходство между белками рассчитывалось двумя способами. В первом случае сравнение производилось только на основе линейных фрагментов последовательностей белков, во втором случае для проверки гипотезы о том, что информация о конформационных пептидах, представленных на поверхности известных белков-аллергенов, может увеличить точность предсказания аллергенности; дополнительно к линейным фрагментам рассматривались конформационные пептиды.

**Расчет линейных и конформационных пептидов.** Линейные пептиды рассчитывались путем сдвига подвижной рамки длиной 8 аминокислотных остатков вдоль последовательности белка. Такая длина часто используется в методах предсказания аллергенности (Saha, Raghava, 2006b; Silvanovich *et al*., 2006; Herman *et al*., 2009).

Конформационные пептиды рассчитывались в пространственных структурах белков-аллергенов по следующим правилам:

1) два аминокислотных остатка считались связанными в конформационном пептиде, если расстояние между их C-α-атомами в пространственной структуре белка было не более 5 Å;

2) требовалось, чтобы аминокислотные остатки, формирующие конформационный пептид, располагались на поверхности белка. Для этого рассматривались только те конформационные пептиды, для которых усредненная доступность для растворителя была не менее 50 % от максимальной доступности аминокислот для растворителя;

3) длина конформационных пептидов так же, как и в случае с линейными пептидами, задавалась равной 8 аминокислотным остаткам.

Пример конформационных пептидов, рассчитанных для белка-аллергена Zea m 1, изображен на рис. 1.

**Расчет сходства между анализируемым белком и белком-аллергеном.** Для расчета сходства между анализируемыми белками и белками-аллергенами было создано два набора пептидов: LP, линейные пептиды и LCP, линейные и конформационные пептиды. Известные белки-аллергены были взяты из базы данных аллергенов SDAP (Ivanciuc *et al*., 2003). Среди представленных в базе данных последовательностей белков-аллергенов с помощью программы PISCES (Wang, Dunbrack, 2003) были отобраны 586 белков, между которыми уровень сходства не превышал 90 %. Наборы LP и LCP строились на основе анализа первичных и пространственных структур данных белков соответственно. Набор LCP был расширением LP путем включения конформационных пептидов, рассчитанных по пространственным структурам белков. В базе данных PDB было обнаружено 16 экспериментально расшифрованных пространственных структур для разных белков-аллергенов. Пространственные структуры других белков извлекались из репозитория предсказанных пространственных структур (Kiefer *et al*., 2009). Всего было собрано 345 пространственных структур белков-аллергенов.

Пептиды включались в соответствующие наборы при условии, если они встречались не менее чем в двух белках-аллергенах. Сравнение пептидов проводилось с учетом их сходства по физико-химическим свойствам аминокислот. Для этого их последовательности представлялись в вырожденном алфавите согласно

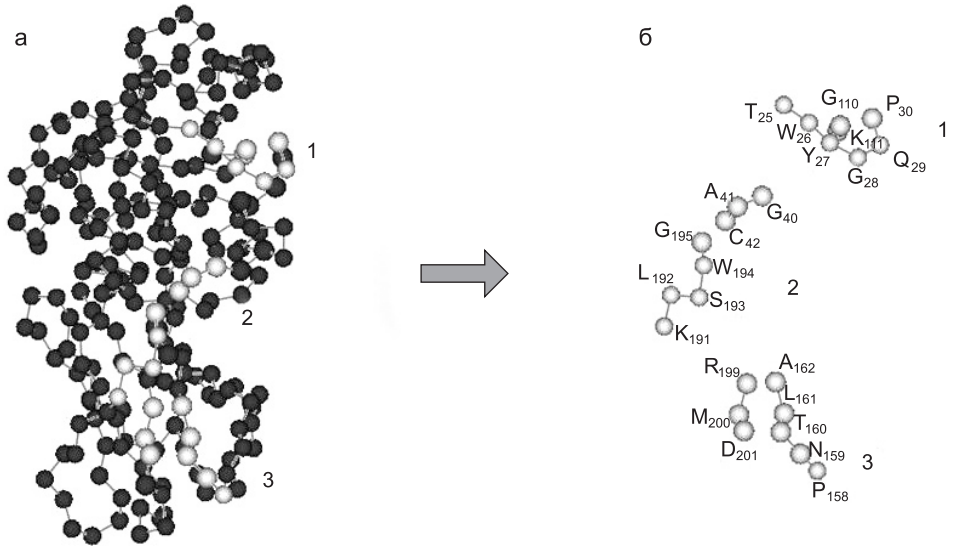


**Рис. 1.** Пространственная структура белка-аллергена Zea m 1 (PDB ID 2HCZ) (а) с примерами найденных конформационных пептидов (б).

Темными кружками изображены аминокислоты, линии между ними обозначают пептидные связи. Светлыми кружками отмечены аминокислоты, входящие в состав конформационных пептидов 1, 2 и 3.

группировке аминокислот по близости физико-химических свойств (табл. 1).

Аминокислоты, принадлежащие к одной группе, заменялись в последовательностях пептидов на соответствующий идентификатор группы. Два пептида считались одинаковыми, если их последовательности полностью совпадали с учетом вырожденности алфавита.

Таким образом, в набор LP было помещено более 44 тыс. последовательностей линейных пептидов, а набор LCP содержал более 99 тыс. линейных и конформационных пептидов.

Далее для каждой аминокислотной последовательности из LP и LCP рассчитывалась характеристика специфичности встречи пептидов в белках-аллергенах (*SA*). *SA* рассчитывался как отношение частот встречаемости пептида в белках-аллергенах к частоте встречаемости пептида в белках неаллергенах

$$SA = \ln\left(\frac{Va}{Vh}\right), \qquad (1)$$

где *Va* – частота встречаемости пептида в белках-аллергенах; *Vh* – частота встречаемости пептидов в белках неаллергенах.

**Таблица 1**

Группировка аминокислот по близости физико-химических свойств

| Группы аминокислот | Аминокислоты |
|---|---|
| Гидрофобные | V, M, I, L, F, Y, W |
| Положительно заряженные | R, K, H |
| Отрицательно заряженные | E, D |
| Полярные | S, T, N, Q |
| Малые аминокислоты | A, G |
| Цистеин | C |
| Пролин | P |

В качестве набора белков неаллергенов нами были взяты белки человека. Известно, что белки человека редко являются аллергенами, так в базе данных UniProt приведено только 5 таких белков (Apweiler *et al*., 2004). Для расчета показателя *SA* из SWISS-Prot случайным образом были выбраны около 3,5 тыс. белков человека. Сравнение пептидов из белков-аллергенов с последовательностями белков человека также проводилось с использованием вырожденного алфавита. Отрицательные значения *SA* приравнивались к нулю.

Предсказание аллергенности белка строилось на расчете значения решающей функции (DF)

$$DF = \frac{\sum_{i=1}^{N} SA_i}{L}, \qquad (2)$$

где $SA_i$ – значения $SA$ у пептидов из LP или LCP, совпадающих с пептидами анализируемого белка; $L$ – длина анализируемого белка.

Для расчета значения решающей функции последовательность анализируемого белка разбивали на линейные пептиды длиною 8 аминокислотных остатков путем сдвига подвижной рамки. Полученные таким образом пептиды сравнивались с пептидами из LP или LCP. Сравнение с LP проводилось в случае предсказания аллергенности анализируемого белка только по аминокислотной последовательности. Набор LCP использовался для предсказания с учетом конформационных пептидов. Полученные при сравнении пептидов значения $SA_i$ суммировались и нормировались на длину последовательности анализируемого белка. Все сравнения пептидов и в этом случае проводились с использованием вырожденного алфавита. Белок считался аллергеном, если значение его решающей функции $DF$ было выше заданного порога. Значение порога для $DF$ задавалось в зависимости от требований к ошибкам перепредсказания и недопредсказания.

**Оценка точности метода.** Для оценки точности метода предсказания был использован тестовый набор белков-аллергенов и неаллергенов из работы H. Muh с соавт. (2009). Из тестового набора была удалена одна последовательность длиной короче 8 аминокислотных остатков. Таким образом, тестовая выборка состояла из 140 аллергенов и 497 неаллергенов. Кроме того, было проведено сравнение белков обучающей выборки с белками тестовой выборки. При оценке точности из обучающей выборки был удален 41 белок, встречающийся в тестовом наборе. Таким образом, после их удаления обучающая выборка аллергенов составила 545 белков.

Для оценки точности метода использовались ошибки недопредсказания и перепредсказания.

$$E1 = \frac{FN}{(TP + FN)}, \quad (3)$$

$$E2 = \frac{FP}{(TN + FP)}, \quad (4)$$

где $TP$ – аллергены, предсказанные как аллергены; $FN$ – аллергены, предсказанные как неаллергены; $TN$ – неаллергены, предсказанные как неаллергены; $FP$ – неаллергены, предсказанные как аллергены; $E1$ – ошибка недопредсказания; $E2$ – ошибка перепредсказания.

## Результаты и обсуждение

Зависимость ошибок недопредсказания и перепредсказания метода LCP, использующего конформационные пептиды, от значения порога для решающей функции $DF$ изображена на рис. 2. Из рис. 2 можно видеть, что при пороге функции $DF$, равном 0,244, достигаются минимальные значения ошибок пере- и недопредсказания, соответствующие пересечению кривых, равные 0,128.

Нами также был построен аналогичный график для метода LP, основанного на анализе

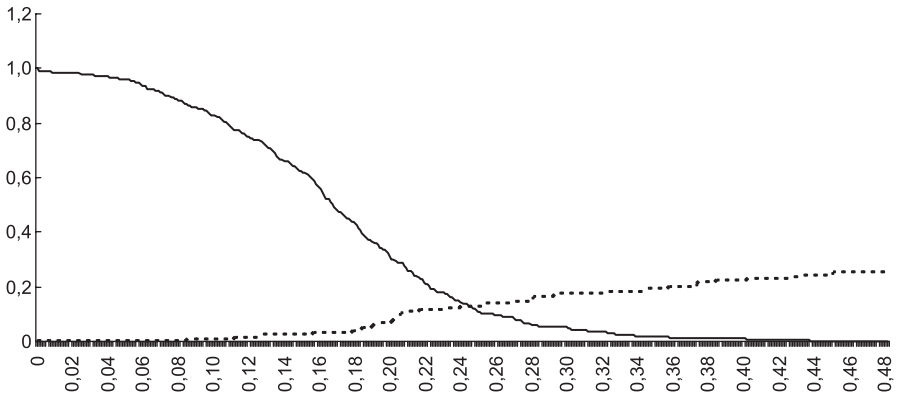


**Рис. 2.** График зависимости ошибок недопредсказания (пунктирная линия) и перепредсказания (непрерывная линия) от значения порога для решающей функции $DF$ метода LCP.

По оси ординат отложены значения ошибок, по оси абцисс – значение порога для DF.

только линейных пептидов (график не показан). Пересечение кривых для ошибок пере- и недопредсказания наблюдалось при пороге, равном 0,191, и соответствовало ошибкам, равным 0,15. Таким образом, метод, использующий конформационные пептиды, показал более высокую точность, чем метод, основанный только на линейных пептидах, что указывает на эффективность использования информации о трехмерной структуре белков в методах предсказания аллергенности.

Для сравнения разработанного нами метода с существующими методами предсказания аллергенности белков была зафиксирована ошибка перепредсказания, равная 0,068 (табл. 2). Такое значение ошибки перепредсказания приводится авторами хорошо известной программы AllerHunter (Muh *et al*., 2009).

При заданном значении ошибки перепредсказания ошибка недопредсказания метода LCP была ниже примерно на 2 % по сравнению с методом AllerHunter (табл. 2). Точность метода линейных пептидов LP оказалась наиболее низкой среди рассматриваемых методов.

Таким образом, предложенный нами подход, рассматривающий конформационные пептиды, позволяет улучшить точность методов предсказания аллергенности белков, основанных на анализе только аминокислотных последовательностей. Можно ожидать, что использование данных о конформационных пептидах может быть применено для улучшения точности других существующих методов предсказания аллергенности, поскольку такого рода информация ранее не использовалась для решения этой задачи. Важно заметить, что для расчета конформационных пептидов требуется наличие пространственной структуры только белков известных аллергенов, а не анализируемых белков. Это позволяет применять разработанный метод для предсказания аллергенности при массовом анализе белков, в том числе и целых протеомов.

**Таблица 2**

Значения ошибок недо- и перепредсказания у различных методов поиска аллергенных белков

| Название метода | Значение ошибки недопредсказания, E1 | Значение ошибки перепредсказания, E2 |
|---|---|---|
| Метод конформационных пептидов LCP | 0,143 | 0,068 |
| AllerHunter* | 0,163 | 0,068 |
| Метод линейных пептидов LP | 0,171 | 0,068 |

* Метод AllerHunter (Muh *et al*., 2009).

**Благодарности**

Работа частично была поддержана грантами EU-FP7 PATHOSYS project № 260429, Министерством науки и образования РФ 14.740.11.0001, междисциплинарными интеграционными проектами СО РАН № 119 и 26, программой РАН № 19, Министерством образования и науки Российской Федерации, ГК 07.514.11.4003.

**Литература**

Локтев А.В., Кувшинов В.Н., Меламед Н.В. и др. Локализация антигенной детерминанты белка Е вируса клещевого энцефалита, узнаваемой антигем-агглютинирующими моноклональными антителами, с помощью пептидной фаговой библиотеки // Вопр. вирусологии. 2002. Т. 47. № 2. С. 31–34.

Туманова О.Ю., Кувшинов В.Н., Ильичев А.А. и др. Локализация конформационного эпитопа гликопротеина gpl20 ВИЧ-1, узнаваемого вируснейтрализующими моноклональными антителами 2G12 // Молекуляр. биология. 2002. Т. 36. № 4. С. 657–663.

Apweiler R., Bairoch A., Wu C.H. *et al*. UniProt: the Universal Protein knowledgebase // Nucl. Acids Res. 2004. 32 (Database issue). D115–119.

Casolaro V., Georas S.N., Song Z., Ono S.J. Biology and genetics of atopic disease // Curr. Opin. Immunol. 1996 V. 8. N 6. P. 796–803.

FAO/WHO Codex Principles and Guidelines on Foods Derived from Biotechnology. 2003.

Herman R.A., Song P., Thirumalaiswamysekhar A. Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation // Clin. Mol. Allergy. 2009. V. 7. P. 9.

Ivanciuc O., Schein C.H., Braun W. SDAP: database and computational tools for allergenic proteins // Nucl. Acids Res. 2003. V. 31. N 1. P. 359–362.

Kiefer F., Arnold K., Künzli M. *et al*. The SWISS-MODEL Repository and associated resources // Nucl. Acids Res. 2009. 37 (Database issue). D387–392.

Kong W., Tan T.S., Tham L., Choo K.W. Improved prediction of allergenicity by combination of multiple sequence motifs // In Silico Biol. 2007. V. 7. N 1. P. 77–86.

Kulkarni-Kale U., Bhosle S., Kolaskar A.S. CEP: a conformational epitope prediction server // Nucl. Acids Res. 2005. 33 (Web Server issue). W168–171.

Li K.B., Issac P., Krishnan A. Predicting allergenic proteins using wavelet transform // Bioinformatics. 2004. V. 20. N 16. P. 2572–2578.

Liang S., Zheng D., Zhang C., Zacharias M. Prediction of antigenic epitopes on protein surfaces by consensus scoring // BMC Bioinformatics. 2009. V. 10. P. 302.

Muh H.C., Tong J.C., Tammi M.T. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins // PLoS One. 2009. V. 4. N 6. e5861.

Ponomarenko J., Bui H.H., Li W. *et al*. ElliPro: a new structure-based tool for the prediction of antibody epitopes // BMC Bioinformatics. 2008. V. 9. P. 514.

Saha S., Raghava G.P. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes // Nucl. Acids Res. 2006a. 34 (Web Server issue). W202–209.

Saha S., Raghava G.P. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network // Proteins. 2006b. V. 65. N 1. P. 40–48.

Schramm G., Bufe A., Petersen A. *et al*. Discontinuous IgE-binding epitopes contain multiple continuous epitope regions: results of an epitope mapping on recombinant Hol l 5, a major allergen from velvet grass pollen // Clin. Exp. Allergy. 2001. V. 31. N 2. P. 331–341.

Silvanovich A., Nemeth M.A., Song P. *et al*. The value of short amino acid sequence matches for prediction of protein allergenicity // Toxicol. Sci. 2006. V. 90. N 1. P. 252–258.

Smith G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface // Science. 1985 V. 228. N 4705. P. 1315–1317.

Soeria-Atmadja D., Zorzet A., Gustafsson M.G., Hammerling U. Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms // Int. Arch. Allergy Immunol. 2004. V. 133. N 2. P. 101–112.

Stadler M.B., Stadler B.M. Allergenicity prediction by protein sequence // FASEB J. 2003. V. 17. N 9. P. 1141–1143.

Sun J., Wu D., Xu T. *et al*. SEPPA: a computational server for spatial epitope prediction of protein antigens // Nucl. Acids Res. 2009. 37 (Web Server issue). W612–6.

Sutton B.J., Gould H.J. The human IgE network // Nature. 1993. V. 366. N 6454. P. 421–428.

Takagi K., Teshima R., Sawada J. Determination of human linear IgE epitopes of Japanese cedar allergen Cry j 1 // Biol. Pharm. Bull. 2005. V. 28. N 8. P. 1496–1499.

Wang G., Dunbrack R.L. Jr. PISCES: a protein sequence culling server // Bioinformatics. 2003. V. 19. N 12. P. 1589–1591.

Zorzet A., Gustafsson M., Hammerling U. Prediction of food protein allergenicity: a bioinformatic learning systems approach // In Silico Biol. 2002. V. 2. N 4. P. 525–534.

## PROTEIN ALLERGENICITY PREDICTION ON THE BASE OF DISCONTINUOUS PEPTIDES

**A.O. Bragin, P.S. Demenkov, V.A. Ivanisenko**

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: ibragim@bionet.nsc.ru

### Summary

Nowadays, allergy is one of the most common health problems for developed countries. The accuracy of prediction of protein allergenicity from their amino acid sequences is limited due to the spatial organization of protein patches containing allergenic epitopes. A new method of prediction of allergenic proteins by representation of their surfaces with a set of discontinuous peptides was proposed. It has been shown that the use of information on conformational peptides of a protein improves the accuracy of the method for allergenicity prediction.

**Key words:** allergen, discontinuous peptide, protein 3D structure, protein allergenicity prediction.