РЕЦЕНЗИЯ

от проф. дмн Евгения Стоименова

по конкурс за заемане на академичната длъжност „доцент" в професионално направление: 4.5 Математика, научна специалност: „Теория на вероятностите и математическа статистика", за нуждите на Института по математика и информатика при БАН

В съответствие със заповед № 99/31.03.2015 г. на директора на ИМИ – БАН и решение на научното жури с Протокол № 1 от 10.04.2014 г. съм избрана за рецензент по конкурс за доцент, обявен в Държавен вестник, бр. 8, от 30.01.2015 г. Документи за участие в конкурса е подал д-р Деян Йорданов Палежев.

1. Общо описание на представените материали

За участие в конкурса кандидатът е представил 13 научни статии. Всички те са публикувани в специализирани международни списания, 12 от които с импакт фактор, както следва: две от статиите са в списания с много висок импакт фактор, Science (IF 2013: 31.48) и Nature (IF 2013: 42.351); три статии са в Proc. Natl. Acad. Sci. U. S. A.; една статия е в Serdica J. Computing; една статия е в Stat. Appl. Genet. Molec. Biol.; останалите статии са в специализирани научни списания от областта на генетиката, биологията и медицината.

Една от публикациите е самостоятелна, 1 е с един съавтор, останалите с повече от трима съавтори.

2. Обща характеристика на научната дейност на кандидата

Деян Палежев е завършил магистърска програма във ФМИ на СУ „Климент Охридски" през 1995 г. Бил е редовен докторант в периода 2001-2006 г. в Университета на Иейл, САЩ, като през 2003 е получил образователна степен магистър, а през 2006 – научна степен Ph.D. по статистика. Защитил е дисертация на тема „Тест за бимодалност в многомерни пространства", която е призната в БАН с решение на НС на ИМИ от 29.03.2013 г. От 2012 г. е асистент в ИМИ, секция „Изследване на операциите, вероятности и статистика".

Д-р Деян Палежев е утвърден специалист в областта на математическата статистика и по-специално в областта на статистическите и биоинформатични методи, приложими в генетиката. Автор е на 13 научни публикации, изнасял е доклади на престижни научни конференции. Основната преподавателска дейност на д-р Палежев е била досега в университета в Йейл, 2004-2005 г., с курсове по случайни процеси, теория на вероятностите и теория на статистиката; ФМИ на СУ, 2012-2013 г., с курс по статистика и емпирични данни; и в съвместната програма на ИМИ и Нов български университет, 2014 г., с курс по статистически методи.

Деян Палежев е участвал в няколко научно-изследователски проекта, в периода от 2006 до 2011 в университета в Иейл и понастоящем участва в проект

на ИМИ към ФНИ и двустранен проект с МАНИ. Деян Палежев е отличен с награди: John F. Enders Fellowship, Yale University (2005) и John Perry Miller Fund Award, Yale University (2004).

3. Съдържателен анализ на научните постижения

Научните и научно-приложни интереси на д-р Деян Палежев напълно съответстват на научната специалност „Теория на вероятностите и математическа статистика". Те са в областта на математическата статистика и наи-общо се отнасят до нови математически и статистически методи в областта на биостатистиката, биоинформатиката, статистическата генетика, изчислителната биология, многомерения анализ, класификации и клъстериране.

Основните научни интереси на Деян през последните години са новите технологии за генетично секвениране (next generation sequencing или NGS) и свързаните с тях методи за анализ на данни. Това са най-бързо развиващите се технологии в сферата на биомедицинските изследвания. Те генерират голямо количество качествени данни от ДНК и РНК, много по-бързо от предходните технологии. Поради огромното количество данни, търсенето на математически, статистически и биоинформатични методи за обработка на данни от тези технологии се увеличава. Изследванията на Деян са както теоретични (от областта на математическата статистика), така и приложни, и намират реализация в конкретни задачи на генетиката и медицината.

Ще се спра по конкретно на някои от резултатите, съдържащи се в представените 13 научни статии. Те могат условно да се разделят на 4 групи: 1) Статистически и биоинформатични методи за анализ на данни от секвениране на ДНК и РНК, както и данни от микрочипове; 2) Методи за анализ на данни в случаи, в които извадката с обем, много по-малък от размерността на извадъчното пространство; 3) Анализ на данни от индуцирани плурипотентни стволови клетки (iPS cells); 4) Анализ на фенотипни признаци за географски региони [10 и 11].

NGS технологиите, които се появиха от средата на първото десетилетие на този век, позволяват на учените да генерират генетични данни, десетки пъти по-големи от тези, генерирани от предишни технологии. През последните години се разработват нови статистически методи за анализ на данни, приложими към основните биологични въпроси, на които може да се отговори с изследвания с NGS, използвайки ДНК или РНК секвениране. Няколко от публикациите на д-р Палежев се отнасят до секвениране на ДНК, публикации [1, 8, 5]. Статията [1], публикувана в Science през 2007 г., е една от първите в света по тази тема и има много висока цитируемост. В нея са идентифицирани 1300 структурни варианта чрез метод PEM (paired-end mapping) при което двата края на относително дълги ДНК сегменти се секвенират и се намират координатите им в референтния геном. Определянето е с непараметричен статистически критерий като точните области на отхвърляне се определят чрез горният и долният 0.00135 квантил на разпределението на разстоянията между двата края на сег-

ментите. Установено е, че броят на структурните варианти е много по-голям от предварително очакваният. Предложен е и параметричен модел за разпределението на дължите между двата края сегментите като смес от нормални разпределения, който дава по-точни оценки за квантилите. Статията има още 22 съавтори и е публикувана в Science през 2007 г. Участието на д-р Палежев в тази публикация е съществено.

Няколко от статиите съдържат нови статистически и биоинформатични методи за анализ на данни от секвениране на РНК [13, 12, 7, и частично 8]. При анализ на такива данни се цели сравняване на генната експресия (gene expression) между две групи, най-често между здрави хора и пациенти. Сравнението е на векторите на генната експресия при голям брой (хиляди) гени. Основен проблем при тези експериментите е, че често извадката е малка, с по-малко от 10 индивида във всяка група, а размерността на данните е в хиляди (хората имат над 20000 гени), ситуация известна като проблема на размерността (curse of dimensionality). В статията [13] се оценява грешката от I род при множествени сравнения. Това е случай когато обичайните оценки се базират на консервативни статистически критерии. Те надценяват вероятностните стойности, което води до загуба на статистическа мощност и до оскъпяване на експериментите. За подобряване качествата на критериите, в тази статия е представена изчислително ефективна техника за калибриране на вероятностните стойности, която съществено увеличава статистическата мощност при по-малък обем на извадката. Методът е приложен към метода DESeq. Тази публикация бе избрана за научно-приложно постижение на ИМИ за 2014 г.

В статията [12] се изследва разлика между генната експресия при пациенти от Синдрома на Турет и контролна група от здрави хора. Анализирани са нивата на генна експресия на около 36000 гени. Чрез метода edgeR са намерени 309 гени, в които пациентите имат статистически значими по-ниски нива на генна експресия от контролите, и 822 гени, в които пациентите имат статистически значими по-високи нива. В статията [7] са изследвани промените в генната експресия, свързани с възрастта. Използвани са данни от РНК секвениране на префронталния кортекс. Чрез метода DESeq, както и чрез статистически методи за намиране на мрежи от гени със свързани нива на експресия (coexpression networks), е показано, че най-големите промени в генната експресия, свързани с възрастта, са в клетъчните сигнални системи и съответно в преноса на нервни импулси.

Статията [9] е базирана на резултати от дисертацията на Деян Палежев. В нея е разгледан статистически критерий за идентифициране на клъстери в многомерно пространство, базиран на k-means метода за клъстеризиране. Съществена особеност на данните е, че размерността на пространството на променливите е много по-голяма от броя на наблюденията. Разгледани са недостатъците на стандартния метод, който използва проекциите върху права или равнина. Намерено е асимптотично уравнение за вероятността на опашката на максимума на много независими χ^2 случайни величини и е изведено неравенство

за асимптотичното разпределение на максимума на независими и положително корелирани χ^2 случайни величини. Предложен е консервативен статистически критерий за проверка на хипотезата за многомерно нормално разпределение срещу алтернатива за разпределение, което е смес от две многомерни нормални разпределения. Предложен е алтернативен критерий по-малко консервативен,със същите асимптотични свойства. Резултатите могат да се приложат към биологични данни, свързани със секвениране.

Третата група статии е свързана с анализ на данни от индуцирани плурипотентни стволови клетки (iPS cells) [8, 6 и 5]. Статията [8], публикувана в Nature през 2012 г., е посветена на две модерни теми на научни изследвания (iPS клетки и секвениране). Изследван е процеса на репрограмирането: дали могат да се появят структурни варианти , които не са съществували в началните клетки, но съществуват в iPS клетките (т.нар. di novo структурни варианти). Анализирани са данни от ДНК секвениране на iPS клетки, получени от фибробласт от кожата на няколко човека и е показано, че всички варианти в iPS клетките или съществуват в оригиналния фибробласт, или в родителите на съответния индивид, т.е. не се наблюдават di novo варианти. Събрани са допълнително данни от РНК секвениране, за да се покаже, че iPS клетките са по-близо от гледна точка на генна експресия до стандартна линия от индуцирани плурипотентни стволови клетки, отколкото до оригиналните фибробластни клетки, т.е. процесът на репрограмирането работи. В [6] се разглежда друг проблем, свързан с iPS клетките. Показано е, че клетки от вида hiPSCs, отгледани по специален начин, имат профил на генна експресия, подобен на този на ембрионния теленцефалон, но не и на други региони на централната нервна система. За изследването за използвани РНК микрочипове. Резултатите показват, че най-голяма близост се наблюдава със стените на ранния церебрален кортекс. Този модел може да се използва за изследване на развитието на човешкия мозък и на болести, свързани с церебралния кортекс. В [5] се разглеждат проблеми по намирането на генетични варианти (SV/CNV) с новите технологии за генетично секвениране (high-throughput sequencing, или Next Generation Sequencing). Резултатите са свързани с използването на iPS клетките за изследване на невро-психиатрични болести.

В статиите [2, 3 и 4] са анализирани данни от микрочипове (microarrays). В [4] са идентифицираме 55 гени, които се предполага, че са свързани със способността за четене. В [3] се изследва ефекта на EBV nuclear antigen 1 върху генната експресия. Идинтифицирани са гени, които са със статистически значими по-високи средни нива на експресия в присъствието на EBNA1, отколкото без него. Тези гени са свързани с растежа и миграцията на клетките, имунната система, стабилността на ДНК и др.

В двете статии [10 и 11] са правени анализи на фенотипни признаци за географски региони. В [11] е анализирано генетичното разнообразие на 52 сорта пшеница от западния и североизточния черноморски региони. Чрез клъстерен анализ е показано, че сортовете от западния регион са генетично по-близки

помежду си, отколкото до тези от североизточния регион, а също така, че клъстерирането на сортовете от североизточния регион според генетичната им информация до голяма степен съвпада с това по географски подрегиони.

4. Отражение на научните публикации на кандидата

В документите на кандидата е представен списък от 722 забелязани цитирания в научни списания с импакт фактор. Ще отбележа специално три от работите с по-голям брой цитирания от други автори: Статия [1], която е цитирана поне 557, статия [8] – цитирана повече от 80 пъти и статия [6] – цитирана повече от 50. Участието на д-р Палежев в тези публикации е съществено.

5. Оценка на личния принос на кандидата

Значителна част от представените по настоящата процедура публикации са в съавторство. Смятам, като обща оценка, че в много от тях д-р Деян Палежев има равностойно участие, сравнено с това на другите съавтори, специалисти по анализ на данни. В някои от публикации, с много съавтори, д-р Палежев е единствен статистик и е видна неговата съществена роля в изследването.

6. Критични бележки и препоръки

Нямам съществени критични бележки. Документите са подготвени старателно и не затрудняват оценката. Някои от статиите са представени в препринт версии, кое затруднява четенето им [12].

7. Лични впечатления

Личните ми впечатления за Деян са отлични като за колега и специалист по вероятности и статистика. Работи активно по проектите и научните мероприятия на ИМИ.

Заключение

Казаното дотук ми позволява да твърдя, че д-р Деян Палежев е високо квалифициран специалист, който има съществени приноси в областта на вероятностите и статистиката. Смятам, че са удовлетворени съвкупността от критерии и показатели за придобиването званието „доцент", съгласно ЗРАСРБ.

Предлагам Научното жури да препоръча на Научния съвет на ИМИ да присъди на д-р Деян Йорданов Палежев научното звание „доцент" в професионално направление 4.5. Математика, специалност „Теория на вероятностите и математическа статистика".

София, 14 юни 2015 г. Подпис:

Евгения Стоименова