

Авторска справка За научните приноси на д-р Румяна Кирова Йорданова

Основната част от изследователска ми дейност е в областта на информационни методи в геномиката. По конкретно, тя е свързана с разработването и прилагането на биоинформационни и статистически методи за анализ и интегриране на микробиологични данни от тип "omics" включващи ДНК, РНК, генни експресии и секвенции на целия геном. Тези изследвания често са проведени в колектив от молекулярни биолози и биохимици и целта им е свързана с намирането на нови направления за лечение на различни заболявания като сърдечно съдови, атеросклероза, онкологични, редки имунологични, автоимунни, на нервната система и други.

Биоинформатиката е интердисциплинарна област за разработване на методи и софтуерни програми за анализ на биологични данни. Тук се включват методи за обработване на образи, данни от геномиката като секвенционни данни и генни експресии, анотации на биологични обекти като гени, протеини, пост генетични модификации и много други. През последните години също така се разработват програми за анализ на секвенции на целия геном и корелацията му с различни често срещани или редки заболявания.

В молекулярната биология „Big Data“ обикновено се асоциира с така наречените „omics“ данни, които включват геномни данни като секвенции на генна информация, обикновено целия геном или екзом, РНК и ДНК експресии, SNP (единични нуклеотични полиморфизми) и други мутационни данни. Един от основните проблеми при анализа на такива данни е, че те са с огромен размер и имат различни нива на сложност. Тяхната интеграция и правилното интерпретиране са от съществено значение. Дизайнът на експериментите, като големината на извадката и контролирането на потенциални фактори влияещи върху грешката от анализа, е много важен за да може да се отговори смислено на поставените биологични въпроси. При анализа на тези данни често се правят голямо количество тестове и също така трябва да се контролира глобалната грешка за да се избегнат фалшиво позитивни резултати.

Голяма част от публикуваните ми резултати са свързани с изследователската ми дейност във фармацевтичната индустрия (Bristol-Myers Squibb) и Alexion Pharmaceuticals), където се занимавах с разработването на софтуерни програми за анализ и визуализация на геномни данни, биоинформатични методи за конструиране на мрежи от асоциации и причинно следствени връзки между генетичната информация, РНК, протеинова експресия и комплексни фенотипи. В Bristol-Myers Squibb (7 години) бях старши изследовател I и II и консултант (1 година). Някои от основните проекти с които се занимавах са:

- Софтуерна интерактивна програма XPLOr за анализ и визуализация на протеинови данни от Mass Spectrometry. Резултатите са представени на няколко вътрешни конференции. Системата беше написана на R (<https://cran.r-project.org>) използвайки интерактивни пакети като R Commander и Tcl/Tk. Беше използвана от биолози и физици при анализа на клинични и преди-клинични експерименти.
- Системен биологичен и генетичен анализ (Systems biology and genetics analysis) на “omics” данни от геномиката – WGS (секвенции на целия геном), WES (секвенции на целия екзом), SNP (единични нуклеотични полиморфизми), CNV (брой на мутационните вариации като включвания и изтривания на редици от нуклеотиди), microarray (микрочипова) expression; протеинови (mass spec proteomics, metabolomics) свързани с метаболитни процеси, сърдечни заболявания, атеросклероза, колит, меланом и други автоимунни заболявания. Конструирание на високо резолюционни карти на асоциативното мапиране за дисекция на комплексни финотипове. Системен генетичен анализ на интеракциите между гени и околна среда. Системен генетичен анализ на различни популации на хора и мишки от различни тъкани. Сравнителен анализ. Pathway analysis използвайки множество анотационни бази данни както и класификационни и клъстерни модели. Използване на смесени и други линейни модели за корекция на популационна структура включително и методи за контролиране на грешката. Част от резултатите са публикувани в [1,2,3,4].
- Софтуерна програма BINGO за конструирание и визуализации на мрежи от асоциации и причинно следствени връзки между “omics” данни включващи “partial correlations”, “Likelihood methods”, “Bayesian approaches”, WGCNA. Резултатите са представени на няколко вътрешни конференции. Програмата беше част от по-обща система за съхранение, обработване, визуализация и интеграция на биологични данни във фирмата. Използваше се от изследователите в R&D за търсене на цели.
- Внедряване на методи за анализ на данни с корелационна структура (използване на смесени модели; Баесови модели) [1].
- Програми за GSEA (enrichment analysis) [5].
- Комбинаторни методи за biomarker анализ.

В Alexion Pharmaceuticals, като консултант 2 години и половина, се занимавах с биоинформатичен анализ на генетични редки заболявания, като например Atypical hemolytic uremic syndrome (Атипичен хемолитично-уремичен синдром). Това включваше анализ на протеинови структури използвайки <https://www.schrodinger.com/> в търсене на мутации свързани с имунната система, които са патологични за заболяването, както и data mining на различни бази данни свързани с нарушения на така наречената complement system (система на комплемента). Работех върху различни методи за приоритизация на потенциалните патогенни мутации.

Изследователската ми дейност е осъществена в мултидисциплинарна група от изследователи поради което статиите в които участвам са предимно в биологични списания. Моят основен принос е в информационното, статистическото и биологическото моделиране: обработване, интегриране и анализ на много размерни „omics” данни започвайки от суровите данни; развитие на методи за комбинация на данни от различно естество; дизайн на експериментите с цел оптимизация на извадката за постигане на необходимата мощност на методите, визуализационни методи и програми за представянето на данните и резултатите. Също така, като информатик/статистик бях свързващо звено между биоинформатиците и биолозите/биохимиците и една част от дейността ми се състоеше в самата координация и организация на някои от проектите.

Методи за “Systems Biology Analysis” и асоциации между различни „omics” данни.

Тук се включват резултати свързани с информационни подходи за анализ на глобални геномни данни с цел намиране на специфични полиморфизми свързани с комплексни фенотипове (напр. HDL, LDL нива на холестерола); методи за изграждане на високо резолюционни мрежи от асоциации между полиморфизми, РНК експресии и геномни данни като микрочипова експресия; методи за намиране на взаимодействията между гени и околна среда. Една от целите на такъв анализ е да получим по-ясна представа за процесите на контролиране на експресията от генетичната информация както и на епигенетичните промени. Друга важна задача е да се намерят нови направления за лечения като например гени чиито полиморфизми регулират техните експресии, а те от своя страна водят до ефекти върху конкретен фенотип или заболяване.

В [1] се използват така наречените “systems genetics” методи за получаването на високо резолюционни карти от асоциации между голямо количество полиморфизми и РНК експресии с комплексни фенотипове, които са свързани със сърдечно съдовите заболявания и атеросклерозата (например HDL, LDL, триглицериди и други липиди). В случая, моделният организъм са особен вид мишки създадени чрез процес на размножаване на братя и сестри. Асоциативният анализ между генетичната структура (отделни полиморфизми, полиморфизми които са в групи като така наречени хаплотипи) и РНК експресиите се наричат eQTL (expression quantitative trait loci), докато асоциациите между полиморфизмите и фенотипите за атеросклероза са QTL (quantitative trait loci). В тази работа се използва нова стратегия, така наречения смесен модел за асоциации, за да се отчете корелационната структура на данните, тъй като мишките са биологично свързани, и да се намали глобалната статистическа грешка. Резултатите от модела представляват набор от милиони тестове (милион полиморфизми умножено по хиляди експресии). Използвани са симулации за да се регулира глобалното ниво на статистическата грешка. Крайният резултат е карта с висока резолюция от статистически съществени eQTLs и QTLs,

която е много по прецизна от традиционния генетичен анализ. Тази карта може да послужи за намиране на потенциални таргети за лечението на атеросклерозата. Основна ми роля в тази статия е при анализа на РНК експресиите, генетичните данни, паралелната реализация на модела за корекция на популационната структура, сравняването му с други такива, използването на симулации и други методи за мултивариантна корекция на глобалната статистическа грешка. Идентифицираните гени бяха допълнително анализирани използвайки различни бази данни с помощта на хомологови методи между различни организми (мишкови срещу човешки), както и “pathway” анализ на механизмите и биологичните процеси, в които участват тези гени с цел допълнително филтриране и избор на таргети.

[2] е пример на системен геномен и сравнителен анализ между протеома и транскриптома, в който се изследва връзката между нивата на експресия и нивата на протеини, които те кодират в модел от 97 мишки от вида “inbred” (мишки получени чрез продължителен процес на размножаване на братя и сестри) както и “recombinant” (рекомбинантни видове мишки), които се създават чрез последователно кръстосване на различни “inbred” видове, което води до фиксация на рекомбиниращите събития. В статията се показва, че нивата на експресиите и протеините зависят много от локацията и функцията на съответните гени. Глобална генна асоциации между съвкупността от полиморфизми с РНК експресии и протеинови нива води до различни групи от гени, които нямат много общо покритие. Също така, групата от гени със съществена корелация между тяхната РНК експресия и фенотипите свързани с липидните нива и групата от гени с корелация между съответните протеинови нива и фенотипите нямат голямо покритие. Допълнително се наблюдава, че корелацията между полиморфизми и РНК експресии е по-голяма отколкото корелацията между полиморфизми и протеинови нива. Основният ми принос в тази статия е анализа, интеграцията и визуализацията на данните включващи различни методи за нормализация на суровите данни, методи за сканиране за потенциални confounding ефекти, като например SNPs (единични полиморфизми) в транскрипционните редици, които увеличават грешката, както и осъществяването на глобален геномен анализ. Също така, сравнявам различни статистически методи за генериране на асоциациите между полиморфизми, РНК, протеинови нива и метаболити, както и методи за построяване на мрежи от такива асоциации и тестване на причинно-следствени връзки прилагайки различни вероятностни модели.

В [3] се прави системен геномен анализ за да се изследват генетичните вариации асоциирани с отговора на макрофагите при наличие на инфламационни стимули като например бактериални липополизахариди (LPS) и оксидирани фосфолипиди. В статията се конструират eQTL модели, които разграничават различните инфламационни стимули от несмутената система. Идентифицират се така наречените горещи места (ДНК полиморфизми на хромозома 8), които регулират голямо количество експресии свързани с инфламационния отговор на макрофагите. Тези горещи места и гените намиращи са в тях са мощен

регулатор на макрофагите при инфламация. Базата данни с всички асоциации е публично достъпна. Основният ми принос е в анализа, генерирането и визуализацията на така наречените горещи места, както и в симулациите за контролиране на грешката и “pathway” анализа, при които се конструират схематични мрежи от взаимодействия между различни биологични структури на базата на групата от гени активирани от инфламацията породена от LPS. Използвам различни бази данни, като например Ingenuity Pathway Analysis [Ingenuity Pathways Analysis \(IPA\) | NIH Library](#), за анотация на биологични обекти.

Взаимодействията между гени и околна среда (GxE) са важни за много човешки заболявания, но е трудно да бъдат изследвани на молекулярно ниво. В [4] правим системен геномен анализ на хиляди характеристики на експресирани транскрипти в човешки първични ендотелни клетъчни линии (ЕК) в отговор на провъзпалителни окислени фосфолипиди, участващи в сърдечносъдови заболявания. Около една трета от 59 най-регулирани транскрипционни експресии показват взаимодействие между гени и околна среда (GxE). Резултатите допринасят за разбирането на цялостната архитектура на сложните човешки фенотипове и са в съответствие с хипотезата, че GxE взаимодействията са отчасти отговорни за неуспеха на асоциативните проучвания да обяснят по-пълно вариациите при често срещаните заболявания. Основната част от моя принос е в системния геномен анализ и статистическите модели за GxE взаимодействия.

В [5] е разработен нов метод за анализ на множества от гени който динамично оптимизира избора на граница и подобрява чувствителността и селективността на анализа на обогатяването на генни множества. Процедурата превръща експерименталните резултати в поредица от списъци с регулирани гени при различни гранични стойности на коефициента на фалшиво откриване (FDR) и изчислява Р-стойността на свръхпредставянето на даден набор от гени с помощта на точния тест на Фишер (FET). Комбинирането на FDR с множество гранични стойности ни позволява да контролираме грешката, като същевременно запазваме гени, които увеличават информационното съдържание. Нашият метод може да се използва за всеки генериран от потребителя списък с гени в областта на геномиката.

Статиите [1,2,3,4,5] са написани по времето на работата ми в Bristol-Myers Squibb в тясно сътрудничество с UCLA основно с групата на проф. А. J. Lusis. Публикуваният анализ е само част от моята изследователска дейност по този проект. Голяма част от намерените гени в горните анализи бяха допълнително анализирани използвайки информатични подходи: търсене в най-различни бази данни за биологична характеристика и анотация; “pathway” (схематичен биологичен) анализ, GSEA “enrichment” анализ, разработване на нови методи за причинно

следствен анализ. Крайният резултат е филтриран лист от гени с детайлна анотация и взаимодействия, които в последствие се анализират от биолозите за откриване на нови насоки при лечението на атеросклероза. Като основен биоинформатик в проекта, главната ми задача освен анализа беше свързана със систематизирането, интерпретацията и визуализацията на резултатите и представянето им пред другите изследователи.

Методи за анализ на микромасивни геномни данни включително динамични данни (time series microarrays)

Следващите статии са резултат от научната ми работа като PostDoc в Oak-Ridge National Lab под ръководството на проф. Elissa Chessler. Тук се включват анализ на динамични микромасивни данни от мишки използвайки статистически методи [6], графични модели [7], обобщени логични мрежи (generalized logical networks) [8] както и онтологична база данни за фенотипно-центрирани геномни асоциации [9].

В [6] се анализират динамични експресионни данни от ембрионалното развитие (поредица от ембрионни дни) на малкия мозък при Sey/Sey мишки за да се тества генната регулацията на малкия мозък във времето и пространството. Използват се методи на контрастите от най-различни видове за да се тестват разликите в ембрионалното развитие и влиянието на гените Pax6 и Math1. Използват се полиномиални контрасти за да се намерят специфични патерни. Моят основен принос включва целия процес от обработването на суровите експресии, нормализацията, анализа на динамичните редици чрез модели с контрасти, построяването на подходящите контрасти, визуализацията на данни, намирането на диференциално различни експресии в ембрионалното развитие; анализа на различни множества от гени използвайки GSEA (enrichment анализ) както и “pathway” анализ, т.е. схематично представяне на множества от гени в биологичния процес. В [7] отново се анализират динамични геномни експресии на ембрионални данни от малкия мозък при мишки с помощта на три аналитични подхода. Два от тях са свързани с диференциални уравнения а един е базиран на теорията на графите, където е и основният ми принос.

В [8] е разработен алгоритъм базиран на обобщени логически мрежи на транскрипционната регулация в мозък на мишка за анализ на динамични експресии, който използва статистическа значимост за да се намали грешката от първи вид. Методът използва тестване на мултиномиални хипотези и по този начин дава по-прецизни резултати от динамичния Баесов модел. Данните, които ние анализирахме, са динамични РНК експресии от мозъка на третирани с алкохол мишки. Моделът идентифицира гени от няколко основни невронни механизми както и потенциални нови генни интеракции. Основният ми принос е в избора на статистически критерии за прилагане на модела, изследване на мощността на модела както и генериране на симулации за тестване.

В [9] е разработена онтологична откриваща система (ODE) за съхранение, споделяне, извличане и анализ на фенотипно-центрирани геномни асоциации. Системата е изградена на базата на естествени ендеогенни процеси и експериментално наблюдавани биологични мрежи, механизми и системи, а не на външно дефинирани конструкции. Всеки тип набор от данни, представящ взаимоотношенията между ген и фенотип, като позиционни кандидати за количествени признаци (eQTL, QTL), рецензии на литература, експерименти с микрочипове, онтологични или дори метаданни, могат да служат като входна информация. Наборите от гени са анотирани с няколко нива на метаданни, включително онтологии на общността. Изчислените генетични набори са интегрирани в йерархични дървета на основата на генно-извлечени фенотипни взаимозависимости. Автоматизираните идентификационни набори се допълват от статистически инструменти, които позволяват на потребителите да интерпретират доверието към моделираните резултати. Този подход позволява интегриране на данни и откриване на хипотези в множество експериментални контексти. Основният ми принос е в разработването на методи за инкорпориране на eQTL, QTL в онтологията като се използват различни дистанционни метрики, в разработката на концептуално ниво на онтологичната система и в прилагането на различни статистически метрики за оценка на резултатите.

Анализ на микробиологични геномни данни за моделиране на антимикробиаланата резистентност

В последните две години, работата ми е насочена основно към биоинформатични методи за анализ на геномни данни от бактерии и моделирането на антимикробиаланата резистентност. За конкурса е включена една излязла тази година публикация, като още две са в процес на печат.

В [10] анализирахме секвенции от бактерии взети от транспортната система (основно метро) на различни градове по света. Основната задача е да се предскаже произхода на данните по броя на различните видове бактерии в тези секвенции. Новото в нашия подход е, че използваме както стандартни методи за машинно обучение така и Баесов пространствен модел за да отчетем пространствената колерация между данните и да оценим относителния риск на различни бактерии и бактериофаги потенциално свързани с антимикробиална резистентност. Моделът е Баесов йерархичен с т.н. CAR prior и е особено подходящ за данни с различно от нормалното разпределение и такива с по-голяма дисперсия. Също така, ние сравняваме различни методи за машинното моделиране като този на случайните дървета (random forest) показва най-добра прецизност.

Последните няколко години също така преподавах Статистика на студенти по математика и Биостатистика на студенти по стоматология в университета на Хокайдо в Япония.

Публикации включени в конкурса

Статиите са описани с повече подробности в таблица (файл Table-4-docent-IMI-2.2_RY.doc) и са приложени копия на всяка от тях както и информация за IF, SJR.

1. Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, Truong A, Yang WP, He A, Kayne P, Gargalovic P, Kirchgessner T, Pan C, Castellani LW, Kostem E, Furlotte N, Drake TA, Eskin E, Lusk AJ, A high-resolution association mapping panel for the dissection of complex traits in mice, *AJ. Genome Res.* 2010, 20(2): 281-290
2. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, Farber CR, Sinsheimer J, Kang HM, Furlotte N, Park CC, Wen PZ, Brewer H, Weitz K, Camp DG 2nd, Pan C, Yordanova R, Neuhaus I, Tilford C, Siemers N, Gargalovic P, Eskin E, Kirchgessner T, Smith DJ, Smith RD, Lusk AJ, Comparative Analysis of Proteome and Transcriptome Variation in Mouse, *PLoS Genet.* 2011, 7(6): e1001393
3. Orozco LD, Bennett BJ, Farber CR, Ghazalpour A, Pan C, Che N, Wen P, Qi HX, Mutukulu A, Siemers N, Neuhaus I, Yordanova R, Gargalovic P, Pellegrini M, Kirchgessner T, Lusk AJ, Unraveling Inflammatory Responses using Systems Genetics and Gene-Environment Interactions in Macrophages, *Cell* 2012, 151(3): 658-670
4. Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, Tilford C, Guan B, He A, Gargalovic PS, Kirchgessner TG, Berliner JA, Lusk AJ, Systems genetics analysis of gene-by-environment interactions in human cells, *Am J Hum Genet.* 2010, 86(3): 399-410
5. Ji RR, Ott KH, Yordanova R, Brucoleri RE, FDR-FET – an optimizing gene set enrichment analysis method, *Advances and Applications in Bioinformatics and Chemistry* 2011, 4: 37-42
6. Ha TJ, Swanson DJ, Kirova R, Yeung J, Choi K, Tong Y, Chesler EJ, Goldowitz D, Genome-wide microarray comparison reveals downstream genes of Pax6 in the developing mouse cerebellum, *Eur J Neurosci* 2012, 36(7): 2888-2898
7. Ha T, Swanson D, Larouche M, Glenn R, Weeden D, Zhang P, Hamre K, Langston M, Phillips C, Song M, Ouyang Z, Chesler E, Duvvuru S, Yordanova R, Cui Y, Campbell K, Ricker G, Phillips C, Homayouni R, Goldowitz D, CbGRITS: Cerebellar gene regulation in time and space, *Developmental biology* 2015, 397(1): 18-30

8. Song MJ, Lewis CK, Lance ER, Chesler EJ, Yordanova RK, Langston MA, Lodowski KH, Bergeson SE, Reconstructing generalized logical networks of transcriptional regulation in mouse brain from temporal gene expression data. *EURASIP Journal on Bioinformatics and Systems Biology* 2009, 1: 545176-89
9. Baker EJ, Li Z, Jay J, Philip V, Zhang Y, Kirova R, Langston M, Chesler EJ, Ontological discovery environment: a system for integrating gene-phenotype associations, *Genomics* 2009, 94(6): 377-387
10. Zhelyazkova, M., Yordanova, R., Mihaylov, I., Kirov, S., Tsonev, S., Danko, D., & Vassilev, D. (2021). Origin Sample Prediction and Spatial Modeling of Antimicrobial Resistance in Metagenomic Sequencing Data. *Frontiers in Genetics*, 12, 213
11. Putman AH, Wolen AR, Harenza JL, Yordanova RK, Webb BT, Chesler EJ, Miles ME, Identification of quantitative trait loci and candidate genes for an anxiolytic-like response to ethanol in BXD recombinant inbred strains, *Genes Brain Behav* 2016, 15(4): 367-381

Дата

Подпис

20 декември 2021 г

Румяна Йорданова