Софийски университет „Св. Климент Охридски“
Факултет по математика и информатика

Изкуствен интелект в биоинформатиката: автоматизиран анализ и класификация на данни от паралелно секвениране

Милко Красномиров Крачунов

Автореферат на дисертация

за присъждане на образователна и научна степен „доктор“
в професионално направление 4.6 „Информатика и компютърни науки“ научна специалност 01.01.12 „Информатика“ (Изкуствен интелект)

Научни ръководители:

проф. д-р Мария Нишева
доц. д-р Димитър Василев

София, 2014 г.

Където не е указано, всички препратки в текста на автореферата към страници, глави, секции, цитирана литература, таблици или фигури, се отнасят за текста на дисертацията. Всички фигури и таблици в автореферата използват същата номерация като текста на дисертацията.

Съдържание

Обща характеристика на дисертацията **2**
Обем и структура на дисертацията 2
Цели и задачи на дисертацията 3

Глава 1: Увод **4**
Описание на проблема 5
Значимост, актуалност и сложност на проблема 7

Глава 2: Теоретични основи и състояние на изследванията по проблема **8**
Размити множества 9
Изкуствени невронни мрежи 9
Анализ на данни от паралелно секвениране 11
Откриване и корекция на грешки 12

Глава 3: Формализация и решение на проблема за анализ и класификация на данни от паралелно секвениране **13**
Постановка на задачите 13
Входни данни и предварителна обработка 14
Аналитичен подход за корекция на грешките 15
Индиректни подходи за валидация 16
Симулация на грешки 16
Валидация чрез неколкократна корекция на грешки . . . 17
Размит индикатор на увереност в анализа на метагеномни данни 18
Използване на невронни мрежи и други средства на изкуствения интелект 18
Вход 19
Обучение с виртуални грешки 19
Използване на получените модели 19

Глава 4: Софтуерна реализация **20**

Глава 5: Анализ на резултатите от проведените експерименти **21**
Изследване на разпределението на грешките 21
Експериментални резултати за аналитичния подход 22
Приложение на невронни мрежи и други инструменти на ИИ . 23
Приложение на методи на ИИ върху кандидати, избрани с аналитичния подход 23
Сравнение със съществуващи решения 24

Глава 6: Приноси и перспективи **25**
Теоретико-методични приноси на дисертационния труд 25
Експериментални приноси на дисертационния труд 26
Перспективи за бъдещо развитие 27
Публикации по темата на дисертационния труд 28

Обща характеристика на дисертацията

Обем и структура на дисертацията

Пълният обем на дисертацията е 166 страници, в които се включват 6 глави, 16 фигури и 15 таблици, списък с използвана литература и речник на термините. Използваната литература включва 170 заглавия, а списъкът с публикации по темата на дисертацията включва 6 статии.

Глава 1 въвежда проблема за анализ на данни от паралелно секвениране и важността на откриване на добри подходи за корекция на грешките при решаването на задачите от областта.

Глава 2 представя подробен обзор на областта на проблема. Разгледани са използваните в дисертационния труд средства на информатиката, направен е обзор на извършваните изчислителни изследвания в геномиката и метагеномиката, заедно с използваните средства на биоинформатиката. Представено е и състоянието на изследванията във връзка със задачата за откриване и корекция на грешки.

Глава 3 описва предложените в този дисертационен труд инструменти за решаване на проблема. Това включва аналитичен подход за откриване на грешки и модел за обучение на невронни мрежи (и други средства на изкуствения интелект), който да класифицира потенциалните грешки. Разработен е индиректен подход за валидация на средствата за корекция на грешки. Предложен е също и размит индикатор на увереност, който в комбинация с метод за оценка на грешките, предлага алтернатива на процедурата по корекция.

Глава 4 съдържа подробно описание на разработения в рамките на този дисертационен труд софтуер, който включва език за описание на управляеми работни потоци за извършване на автоматизирани изчислителни експерименти в областта на геномиката.

Глава 5 представя анализ на проведените експерименти и получените резултати.

Глава 6 представя приносите на дисертационния труд, както и възможностите за бъдещо развитие.

Цели и задачи на дисертацията

Основната цел на този труд е да се разработят и изследват нови подходи за откриване и коригиране на грешки в метагеномни данни от паралелно секвениране, използвайки средствата на информатиката и изкуствения интелект.

Аналитичен подход за оценка на грешките. Основна задача в този труд е разработването на аналитичен подход за откриване на грешките, който разчита на ясно формулиран метод, основаван на локалните сходства между последователностите, с помощта на които се групират близките секвенции, използвайки тегла. Тя може да се раздели на следните конкретни задачи:

1. Да се предложи подход за използване на локалното сходство за оценяване на грешките.
2. Този подход да се включи във функция, която генерира числова оценка в интервала $[0, 1]$, която участва в класификатор, основан на числови прагове.
3. Да се разработят и изследват различни мерки за оценка на локалното сходство.

Приложение на оценките на грешките. Има повече от един начин една числова оценка да участва като инструмент за справяне с грешките. Естеството на данните води до очаквана несигурност в оценките, което обосновава изследването на повече от една алтернатива, и в частност изпълнението на следните задачи:

1. Да бъде предложен подход за корекция на грешките, основан на изчислените оценки и класификатора, използващ числови прагове.
2. Да бъде разработена алтернатива, при която оценките се прилагат направо в алгоритмите за обработка на генетичните данни като размит индикатор на увереност чрез въвеждане на претеглени функции за разстояние.

Валидация. Важно условие, за да може да бъдат оценени и сравнени методите за оценка и корекция на грешките, е да бъдат налични средства за валидация. Използването на преки подходи за валидация е затруднено от липсата на референтни множества от данни, което води до

разширяване на списъка със задачи:

1. Да се разгледа възможността за симулиране на грешки за целите на валидацията.
2. Да се предложат различни подходи за индиректна валидация.
3. Да се сравнят всички предложени методи за откриване и корекция с помощта на предложените методи за валидация.

Оценяване на грешките с невронни мрежи. Апаратът на невронните мрежи дава много добра алтернатива за решаване на такъв вид проблеми. Оценяването и приложението му за класификация на търсените грешки включва изпълнението на следните задачи:

1. Да се предложи индиректен подход за обучение на невронната мрежа, основан на симулация на грешки.
2. Да се построи и обучи невронна мрежа на негова база, която да се използва за класификация на нуклеотидните бази на погрешно и правилно прочетени.
3. Да се комбинира резултатът, получен от аналитичния подход, с резултата, получен от невронната мрежа.
4. Да се оцени резултатът от прилагането на невронната мрежа в сравнение с резултата, получен само с аналитичния подход.
5. Да се използват и други средства за машинно самообучение, като например дървета на решенията, и резултатът да се сравнят с този от невронната мрежа.

Софтуер. Да бъде разработена гъвкава софтуерна система за автоматизирано изпълнение на разгледаните в труда процеси, чийто дизайн да позволява лесно разширяване към други цели и задачи.

Глава 1: Увод

Анализът на данни от паралелно секвениране в геномните изследвания е актуален проблем, който има значение за широк кръг от научни и практически области – от естествената история на живите организми до здравеопазването. В частност, метагеномиката, върху която е съсредоточен този труд, се занимава с изследване и класификация на големи

групи микроорганизми, и със задачи като изучаване на биоразнообразието им, еволюционната им история и оценяване на ефекта им върху околната среда, добива на земеделска продукция и човешкото здраве.

Различните метагеномни изследвания са силно затруднени от липсата на добри решения за елиминация на шума в данните, предизвикан от грешки на апарата за секвениране. Това е най-видимо в задачите за автоматизирано откриване на структурни варианти, които разчитат на възможността откритите точки на вариация да могат да бъдат ефективно класифицирани като грешки или мутации. Информатиката предлага широк набор от инструменти, които имат потенциала да спомогнат за решаване на този проблем. По-специално, в този труд е разгледан резултатът от прилагането на мощните средства за класификация на изкуственият интелект при различаването на грешки от мутации.

Описание на проблема

Геномното секвениране (genome sequencing) е процес, при който генетична информация, съхранена биологически под формата на вериги от *нуклеотиди* (nucleotides), се прочита от *секвенатор* и се прехвърля в паметта на компютърна система като последователности над четирибуквена азбука, съответстваща на четирите *нуклеотидни бази* (nuleobases), за да може да бъде обект на изчислителни изследвания [75]. *Масовото паралелно секвениране* или секвенирането от ново поколение (massively parallel sequencing, next-generation sequencing (NGS), second-generation sequencing) обхваща редица технологии за секвениране, които позволяват бърз прочит на много голям обем от кратки генетични фрагменти [3].

След извършване на процеса на секвениране, прочетените данни биват подложени на различни алгоритми за обработка преди те да станат обект на изследвания. Определени изчислителни задачи, като асемблирането на геноми на макроорганизми [164] и ДНК експертизите в съдебната медицина [64], стъпват на вече развити и утвърдени подходи. В същото време, все по-голяма популярност набират области като *метагеномиката* (metagenomics), занимаваща се с изследване на генетичен материал от групи микроорганизми, където задачите не винаги се решават оптимално от наличните средства [158, 61, 144]. Един от значимите проблеми, с които метагеномните изследвания се сблъскват, е голямото количество грешки при секвениране, при които погрешни бази са били заместени, изпуснати или добавени в прочетените последователности.

Проблемът за елиминация на ефекта на грешките в процеса на едно

метагеномно изследване е комплексен. Налице е необходимост както от разработване на по-подходящи методи за оценката и откриването им, така и от разработване на подход, при който тази оценка да бъде включена в самото изследване – било то чрез подход за корекция на самите грешки или чрез модификация на използваните алгоритми за обработка и изследване. В настоящия труд са разгледани и двете възможности. Освен това поради спецификата на метагеномните данни, се налага и разработване на оригинален подход за валидация на предложените методи.

Поради необходимостта от запазване на редки различия в генетичните вериги като *точкови мутации* (single nucleotide polymorphisms, SNPs), които са важен обект на изследване, изследователите в областта са много консервативни към средствата за премахване на шум. В същото време, поради чувствителността към качеството на данните, всякакво подобрение е от много голямо значение за тях. Всеки разработван подход за справяне със зашумени данни трябва не само да покаже, че води до подобрение, но и да покаже, че премахва възможно най-малко действителна вариация. Всякакъв напредък в качеството на класификация на откритите разлики като грешки или мутации (естествена вариация) е от голямо значение за този род изследвания.

Изкуственият интелект предлага средства, които са обещаващи при търсенето на решение на този проблем. Данните съдържат неструктурирани и хаотични връзки, които носят скрит биологичен смисъл, но не се подават на формализиране. Това предразполага към извършване на експерименти с апарата на изкуствените невронни мрежи [39], чийто потенциал се крие именно в способностите му да прави изводи точно на базата на такива неявни зависимости. От друга страна, апаратът на размитите множества [162] представя една възможност несигурните резултати, получени при оценка на грешките, да бъдат внедрени в алгоритмите за обработка, без да бъдат премахвани естествените вариации.

За потвърждение на качествата на разработените като част от този труд методи за справяне с грешките е необходимо да бъдат извършени комплексни изчислителни експерименти. В същото време самите процедури и работни потоци за изследване имат необходимост от гъвкавост, и подлежат на модификация за приложение на предложените тук средства. Това предразполага разработването на софтуерна система за описание на работни потоци, която позволява лесно и гъвкаво изпълнение на изчислителни експерименти.

Значимост, актуалност и сложност на проблема

При изследване на биоразнообразието на съобществата от микроорганизми в дадена среда, грешките при секвениране могат да доведат до значително завишени предсказания [74, 62] или да променят значително структурата на реконструираното *филогенетично* (еволюционно) дърво [117]. При търсене на редки точкови мутации, наличието на грешки води до значително повишаване на погрешните предсказания [23]. Честа практика при секвениране е да се отстраняват *прочити* (reads), в които има съмнения за грешки [63], както и прочити, които се срещат само по веднъж [144].

Наличните подходи, разгледани в 2.3.4 и 2.4, не успяват ефективно да разрешат проблема с шума в данните. Създаването на усъвършенствани средства за справяне с грешките в данните може да спомогне за подобряването на изследванията в различни ключови области по редица начини.

- Подобряване на качеството ще помогне за редуцирането на количеството отстранени данни, което ще увеличи количеството полезна информация, която може да се изведе от една проба, и ще намали цената на процеса на секвениране.

- Намаляване на броя грешки и количеството загубена информация в следствие на отстраняване на прочити ще спомогне за увеличаване на броя на открити редки точкови мутации.

- Дори малък принос за подобрение на средствата за справяне с грешките дава основа за бъдещ труд, който е необходим, за да се елиминират напълно отклоненията при оценяване на биоразнообразието, както и несъответствията и неточностите при откриване на точкови мутации.

- Промяна на начина на използване на оценките за качество може да отложи отстраняването на данните, считани за погрешни, или да ги остави достъпни за по-късен анализ.

- Промяна на алгоритмите за обработка, като разгледаните в 2.3, така че да вземат предвид предполагаемите грешки, може да позволи селективен избор на местата, където да се прилага премахване на шум – стъпките, които са силно уязвими при наличие на грешки, могат да бъдат настроени да заобикалят зашумените участъци, докато стъпките, които разчитат на запазени вариации, могат да работят с непроменените данни.

Метагеномните проби съдържат множество микроорганизми, и по тази причина след клъстериране отделните групи в секвенираните данни са хетерогенни и съдържат сходни последователности от множество геноми, за разлика от секвенирането на отделни видове, където пробите съдържат прочити на един и същ отрязък от ДНК веригата на даден геном. Това затруднява откриването на грешки, допуснати при процеса на секвениране – те включват подменени, вмъкнати и изпуснати от секвенатора бази, които съответстват на съществуващата естествена вариация.

Разработката на средства за справяне с грешките в метагеномни данни от паралелно секвениране може да срещне трудности, които могат да бъдат разделени в няколко категории.

- Хетерогенната структура на данните.
- Очакванията на изследователите за запазване на естественото вариране и други характеристики на данните.
- Трудностите при валидация и сравнение на различните подходи.
- Липсата на добри обучаващи множества за прилагане на методи на машинното самообучение.

Глава 2: Теоретични основи и състояние на изследванията по проблема

Навлизането на масово паралелно секвениране в метагеномиката води до засилено изследване на биоразнообразието на микроорганизмите, обитаващи различни среди, както и еволюционната им класификация. Анализът на данните при тези изследвания е силно затруднен от наличието на грешки и практическото отсъствие на добри средства за справянето с тях. Предвид значението им за широк набор от проблеми, включително добива на земеделска продукция и човешкото здраве, качеството на получените резултати има съществена значимост. Макар и в по-малка степен, същият проблем съществува и при изследването на някои определени растителни видове като пшеницата.

Информатиката и изкуственият интелект предлагат инструменти, които имат значителен потенциал да бъдат приложени към задачата за подобряване на качеството. Относителната комплексност на данните и отсъствието на директни методи за изчистване на грешките подчертано обосновават използване на апарата на изкуствения интелект.

В тази глава е направен както обзор на средствата на информатиката, които ще бъдат приложени, така и подробно представяне на този вид метагеномни изследвания, и наличните средства за справяне с грешките.

Размити множества

Размитото множество (fuzzy set) A описва обобщено множество, чиито елементи имат степен на принадлежност към това множество, дефинирана с помощта на функция на принадлежност, приемаща стойности в интервала $[0, 1]$. В този труд, степента на сходство на елемента x към размитото множество A ще се бележи с $A(x)$. Функцията на принадлежност $A(x)$ също е обобщение на индикаторната функция на множество.

Всички операции върху множество, като допълнение c, обединение $\cup$ и сечение $\cap$, имат разширения за размити множества, и следният набор от операции се използват най-често:

$$cA(x) = 1 - A(x) \tag{1}$$
$$(A \cup B)(x) = \max(A(x), B(x)) \tag{2}$$
$$(A \cap B)(x) = \min(A(x), B(x)) \tag{3}$$

За простота там, където се налага дефиниране на размити множества чрез прекалено сложни конструкции, същите означения в текста на настоящия труд ще се използват и за размита логика. Ще се приема, че при прилагането на логически операции върху функциите на принадлежност, резултатът е същият като на еквивалентната операция между множества. Така например, за стандартните операции ще бъде в сила следното:

$$\neg A(x) = cA(x) \tag{4}$$
$$A(x) \vee B(x) = (A \cup B)(x) \tag{5}$$
$$A(x) \wedge B(x) = (A \cap B)(x) \tag{6}$$

Размитите множества и размитата логика са разгледани в повече подробности в 2.1.1.

Изкуствени невронни мрежи

Изкуствените невронни мрежи (ИНМ) са инструмент на изкуствения интелект, взаимстван от невронните мрежи, които стоят в основата на

човешките мозък и нервна система. Те подлежат на самообучение и са най-често използваният инструмент, прилаган в разпознаването на образи и други сходни задачи като разпознаването на реч.

Всяка ИНМ представлява мрежа от възли, наречени неврони. Всеки неврон притежава своя активационна функция, която съответства на прага на активация на естествените неврони. Входът им се пресмята чрез линейна комбинация, която е аналог на комбинацията от възбуждащи и подтискащи синапси в естествените невронни мрежи. Невроните се обособяват в слоеве (един входен, един изходен и нула или повече скрити – виж 2.1.2.3), като сигналът се предава от всеки слой към следващия, макар и да съществуват и по-сложни модели като рекурентните ИНМ, където подобно обособяване не може да се направи или остава само условно. Това е илюстрирано на фигура 2.1 на страница 28.

В една изкуствена невронна мрежа всеки неврон получава сигналите от невроните, с които е свързан, и ги обобщава, като прилага тяхна линейна комбинация с адаптивни тегла, чиито стойности се избират на базата на обучаващо правило. След като се пресметне линейната комбинация, върху резултатът се прилага активационна функция, която определя изходния сигнал на неврона. Тази функция най-често е или стъпаловидна функция, приемаща стойности измежду $\{0, 1\}$, или сигмоидна функция, приемаща стойности в $(-1, 1)$ [122].

ИНМ са разгледани в повече подробности в 2.1.2.

Дърветата на решенията (decision trees) [116] са друг мощен инструмент за машинно самообучение. Те се строят за предсказване на стойността на даден целеви атрибут по стойностите на множество от числови или номинални атрибути. При обхождане на такъв вид дърво, тръгвайки от корена, всеки възел разделя множеството от възможни решения по стойностите на един от атрибутите, докато се стигне до листо, което носи предсказаната стойност на целевия атрибут [123].

Съществуват дървета за извършване на класификационен анализ, които предсказват номинален целеви атрибут, както и дървета за извършване на регресионен анализ, които се опитват да предскажат числовата стойност на целевия атрибут. За разлика от невронни мрежи, които моделират сложни линейни взаимовръзки с участие на всички входни стойности, дърветата на решенията се строят на основата на прости правила върху отделни атрибути, които разделят решенията и създават отделните разклонения. Това спомага на това те да бъдат изследвани и разбрани от човек.

Случайните гори, наричани още случайни гори на решенията [20], представят *ансамблов метод за самообучение* (ensemble method), при

който едновременно се конструират множество от дървета на решенията.

Дърветата на решенията и случайните гори са разгледани с повече подробности в 2.1.3 от текста на дисертацията, където на фигура 2.2 на страница 36 е илюстрирано едно опростено дърво.

Анализ на данни от паралелно секвениране

Паралелното секвениране е съвременен технологичен процес, при който от всяка ДНК или РНК проба могат да се прочетат в електронен вид значителен брой къси нуклеотидни последователности, които впоследствие с помощта на изчислителните средства на биоинформатиката биха могли да бъдат навързани така, че да бъде възстановена тяхната оригинална последователност. За нов вид, който не е бил секвениран, това е значително скъп и трудоемък процес. При някои видове този процес може да бъде допълнително усложнен от биологични фактори, както например е при пшеницата, чийто геном е мултиплициран в три сходни, но различни свои копия.

Метагеномиката (metagenomics) е област от геномиката, която се занимава с изследване на генетичен материал от микроорганизми, взет от проби, събрани в хетерогенни биологични среди. Такива среди могат да включват почва, водни басейни, както и вътрешностите на различни по-висши в еволюционно отношение организми. Съобществата от микроорганизми, които ги обитават, са слабо изучени, като дори оценката за разнообразието на видовете подлежи на задълбочено и всеобхватно изследване.

Основните задачи на метагеномиката са сравнителен анализ на микробиалните общества за оценката на биоразнообразието и проучване на филогенетичните (phylogenetic) взаимовръзки, които имат значение за различни научни сфери, свързани с човешкото здраве [105], земеделието [77], бактериалната и вирусната еволюция [73] и еволюцията на видовете като цяло.

След използването на процедурата по паралелно секвениране, всяко метагеномно изследване е свързано с обработката на голям обем от данни без известни взаимозависимости. Рядко може да се разчита на допълнителни помощни знания от генетични бази от данни, в които присъстват само малка част от изследваните видове. Налага се работа с наличните към момента софтуерни инструменти, които не са достатъчно развити, и се основават на алгоритми, които изискват значителни изчислителни ресурси. За отделните фази на изследването се налага избора измежду различни инструменти, чиято пригодност към извър-

швания експеримент е не винаги предварително известна [133, 34].

Същността на паралелното секвениране е разгледана в повече подробности в 2.2 от текста на дисертацията, а приложението на средства на информатиката при изследванията на получените данни е разгледано в 2.3.

Откриване и корекция на грешки

Преди извършването на същински анализ върху метагеномни данни от паралелно секвениране се налага те да бъдат изчистени от грешки. Съществуват различни практики, включващи премахване на прочити с очаквано ниско качество и подрязване на всички прочити до определена дължина, които се прилагат винаги преди извършването на анализ, но дори след тях остават голям брой неоткрити грешки в данните. Този труд е съсредоточен върху разработката на методи за откриване на такива грешки.

Различни изследователи се спират на отстраняването на прочити като решение на проблема с грешките. Това могат да бъдат както прочити, за които е установена голяма вероятност за наличие на грешки, така и прочити, които се срещат един единствен път [63]. Такава практика е подкрепена от неравномерното разпределение на грешките сред прочитите, но не гарантира елиминацията на всички грешки, като в същото време намалява количеството на данни, върху които могат да се правят изследвания.

Предложените в настоящия труд подходи едва ли ще елиминират напълно тази практика, но предлагането на един нов критерий с по-висока точност може да намали относителния дял на отстранени данни, както и дяла на неоткрити грешки. В същото време, предложеният подход за директно включване на размит индикатор на увереност в етапите за обработка на данните може да избегне част от негативния ефект на грешките без нуждата от тяхното пълно отстраняване.

В дисертационния труд са разгледани няколко алтернативни средства за премахване на шум. Алгоритъмът за корекция на грешките SHREC [134] използва *обобщено суфиксно дърво* (generalised suffix trie) за откриване на грешките при прочит на нови геноми. V-Phaser[87] използва статистически подход за откриване на грешки, който разчита не само на оценка на отделните потенциални грешки, но и на двойките грешки, които се срещат едновременно. Това има за цел да засече прочити на еволюционни разделени гени по съвпаденията на мутациите в тях. Coral [130] поправя грешките в прочитите, като използва честотите на срещане на базите върху специално построено тяхно съпоставяне.

Подходите за откриване на грешки са разгледани в повече подробности в 2.3.4 и 2.4

Глава 3: Формализация и решение на проблема за анализ и класификация на данни от паралелно секвениране

Постановка на задачите

Основната задача на този дисертационен труд е разработването на нови средства за откриване на грешки за метагеномни данни от паралелно секвениране, които са по-надеждни и по-подходящи за извършване на корекция на откритите грешки от наличните.

Разработен е оригинален аналитичен подход за откриване и корекция на грешките. Той се основава на претеглено изчисляване на честотите на срещане на всяка база по колони, като за тегла са използвани локалните сходства между двойките секвенции. Те имат за цел да неутрализират отрицателните ефекти на хетерогенната структура на данните, като приближат сходните подпоследователности и отдалечат различните, като по този начин елиминират голяма част от неправилно откритите грешки в редки секвенции, обособявайки ги локално в отделна група.

Поради трудното намиране на тестово множество данни за валидация са предложени два индиректни подхода за валидация – един, основан на неколкократна корекция на грешки, която използва симулирани грешки, и един, основан на корекция на грешките в голямо множество от данни и достатъчно малко негово случайно подмножество. За симулацията се изгражда статистически профил на грешките върху множество от данни с известни грешки, който е специфичен за използвания апарат за секвениране, и който съответства на разпределението на истинските грешки.

Предложен е метод за симулация на „виртуални“ грешки при конструирането на обучаващо множество от примери, което да се използва при създаване на класификационни модели за откриване на грешки, основани на изкуствени невронни мрежи и други средства на машинното самообучение като случайни гори. Неструктурираните зависимости и хаотичната структура на данните ги правят подходящи за приложение на инструменти като ИНМ, поради техните способности за правене на изводи. Когато аналитичните решения не постигат достатъчно

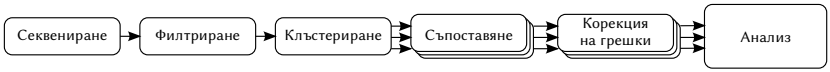
добър резултат, апаратът на ИНМ може да бъде използван за създаване на по-прецизен класификатор на грешките, който да подобри допълнително качеството. Това би било за сметка на по-малката яснота за начина, по който те биват идентифицирани.

Вторична задача на дисертационния труд е изграждането на система за изпълнение на управляеми биоинформатични работни потоци. Разработена е софтуерна библиотека, която освен реализацията на алгоритмите за откриване и корекция на грешки, за симулация на грешки и валидация, поддържа всички необходими операции, използвани за обработка на данните. Това включва множествено съпоставяне на секвенции, клъстериране на секвенции. Разработена е и система, която позволява тези операции да бъдат включени в описания на работни потоци.

Входни данни и предварителна обработка

За целите на този труд са използвани примерни множества от данни, съдържащи прочити с дължина между 300 и 500 бази. Те са разделени в проби, съдържащи между 30000 и 50000 секвенции, след извършване на предварително филтриране (от тях са отстранени прочитите с непрочетени позиции, както и прекалено късите и прекалено дългите прочити, за да се елиминират проблеми, свързани със статистически отклонения). Пробите са секвенирани с 454 платформата [146] на Roche, която е подходяща за метагеномни експерименти поради това, че генерира по-дълги прочити от платформите, използвани при секвениране на геноми на отделни организми.

За да могат да бъдат приложени разработените подходи за корекция на грешките, данните подлежат на няколко предварителни обработки, както е илюстрирано на фигура 3.2.



Фигура 3.2: Анализ на данни от метагеномно секвениране

Първоначално върху данните се извършва филтриране, при което се отстраняват прекалено къси прочити, а прекалено дългите се подрязват. Това премахва възможни статистически отклонения и намалява регионите от прочититите, където честотата на грешки е прекалено висока [47].

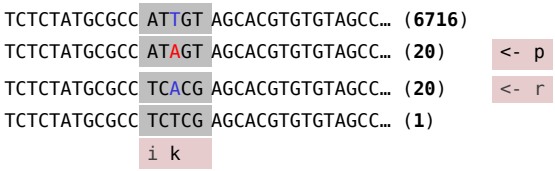
След филтрирането е необходимо секвенциите да бъдат клъстерирани, което в нашия случай е извършено с помощта на готово софтуер-

но решение – CD-HIT [80]. Клъстерирането осигурява работа с подмножества, в които не се съдържат прекалено далечни секвенции. Тъй като всички предложени методи за корекция на грешки разчитат на съпоставени прочити, върху всеки клъстър е изпълнена процедура на съпоставяне по описания в 3.3 комбиниран подход. Използването на специална процедура за съпоставяне се наложи, тъй като в хода на проведените експерименти беше установено, че готовите софтуерни решения не се справят добре с големия обем входни данни.

Аналитичен подход за корекция на грешките

Мутациите имат голямо значение за еволюционните изследвания и носят значима информация за развитието и размножаването на видовете микроорганизми. За разлика от грешките при секвениране, които са разпределени случайно в пробите, мутациите претърпяват различно развитие според това дали са били полезни за развитието на организма. Полезните се унаследяват и размножават, докато вредните убиват организма и прекъсват еволюцията му, поради което не се очаква те да имат същия дял на участие в секвенционните данни. Това от една страна прави оцелелите мутации интересни с информацията, която те носят, а от друга страна представя една възможност за класификацията на двете категории разлики. Мутациите, които са били наследени, ще се отличават по своите измерими характеристики като честота на срещане.

Директното използване на честота на срещане за откриване на грешки в този дисертационен труд е наречено „наивен“ подход и е разгледано в 3.4.1.1. Основен недостатък на използването ѝ е, че тя не взема предвид хетерогенното съдържание на данните, които включват прочити от множество различни организми. За да бъдат взети тези характеристики предвид, наивният подход е разширен чрез въвеждането на честота на срещане, претеглена по локалното сходство между двойките прочити.



Фигура 3.4: Сходство в прозорец около оценяваната позиция

Въведените тегла в честотата имат за цел тя да бъде ограничена в близките прочити, като тези, които съвпадат напълно, имат най-голямо значение, а тези, които се отличават напълно, се изключват. Сходството се пресмята локално, в процорец около базата, чиято честота бива пресмятана. Това е илюстрирано на фигура 3.4, където прочитите явно се обособяват в две групи, според съдържанието на прозореца около оценявата позиция. Използваната функция за пресмятане на претеглена честота е показана във формула (3.2), като в дисертацията са предложени две функции за локално сходство – стандартна (3.4) на страница 72 и „остра“ (3.5) на страница 73 от дисертацията.

$$s_{\text{weighted}}(r,k) = \frac{\sum_{p\in R}^{p\neq r} \text{sim}(r,p,k)[r_k = p_k]}{\sum_{p\in R}^{p\neq r} \text{sim}(r,p,k)} \tag{3.2}$$

Разработеният аналитичен подход за корекция на грешките използва непретеглени и претеглени честоти на срещане е представен в пълни подробности в текста на дисертацията в 3.4.1. Използването на фиксиран и променлив праг за откриване на грешки, както и методът използван, за да бъде извършена корекция с помощта на въведените честоти, са разгледани в 3.4.1.4.

Индиректни подходи за валидация

За да е възможно използването на предложените в 3.4.1 подходи за откриване и корекция на грешки е необходим приложим и обективен подход за валидация. Валидацията е необходима както да се оцени подобрението от прилагане на метода на сходството от 3.4.1.2 и качеството от прилагането му, така и за избор на подходящ праг, при който откриването и корекцията на грешки работят най-добре. Освен това, възможността за валидация е ключова за разработването на подход за откриване на грешки, основан на методи на изкуствения интелект, който е засегнат в предложения в 3.6 подход за обучение на невронни мрежи.

В текста на дисертацията, в 3.4.2, са предложени два потенциални подхода, по които валидацията може да бъде извършена индиректно, без да е наличен еталон, като предпочетеният от тях разчита на симулацията на грешки.

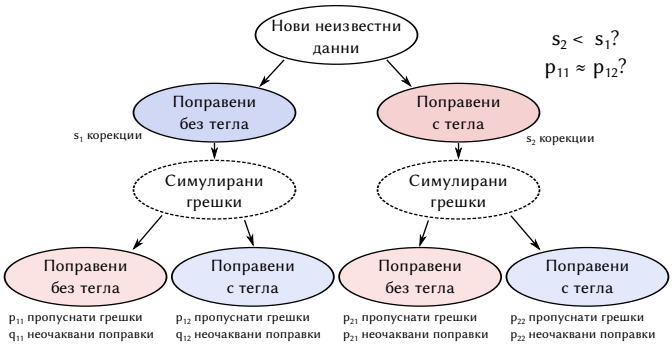
Симулация на грешки

Разработен е симулатор на грешки, който разчита на предварително оценена честота на грешките. За откриването на тези честоти многок-

ратно бяха секвенирани проби от известна повтаряща се последователност, за която наивният подход от 3.4.1.1 е достатъчно добър. След съпоставяне и корекция на грешките бяха изградени консенсусни последователности, които бяха използвани за еталон за оценка на всички прочити. За всяка колона k бяха измерени честотите на заместванията, честотата на инсерции и честотата на делеции. Симулаторът включва софтуерен генератор на грешки, който използва тези честоти, за да внесе случайни нови грешки в множество от ДНК или РНК секвенции. Използваната процедура е описана в детайли в 3.4.2.2 от текста на дисертацията.

Валидация чрез неколкократна корекция на грешки

На базата на симулираните грешки е предложен индиректен подход за сравнение на инструменти за корекция на грешките, както и за сравнение на различни стойности за прага на грешка. Целта да се намери *по-добър* подход отговаря на целта да се *намали броя на нежелани корекции при запазване на броя пропуснати грешки.* При естественото предположение, че увеличаване на прага е достатъчно, за да се намали броя на пропуснати грешки, то намаляване на броя на нежелани корекции е достатъчно, за да очакваме, че подходът е бил подобрен.



Фигура 3.5: Валидация чрез симулация на грешки и двукратна корекция

На първия етап от валидацията се сравнява количеството на извършени корекции от двата сравнявани подхода за корекция. След това в получените коригирани данни се внасят грешки с помощта на разработения симулатор, и на втория етап от валидацията двата подхода за корекция на грешките се изпълняват отново и се сравнява количество-

то на изпуснати грешки. Това е илюстрирано на фигура 3.5 и е описано в детайли в 3.4.2.3 от текста на дисертацията.

Размит индикатор на увереност в анализа на метагеномни данни

Много от алгоритмите, прилагани за обработка на данни от паралелно секвениране, както в метагеномиката, така и в други геномни изследвания, разчитат на използването на матрици на разстоянието. В 3.5 от текста на дисертацията е въведен размит индикатор на увереност, който да се включи в тези функции на разстояния, за да бъде използвана оценката $s(r, k)$, получена след откриване на грешките без да бъде извършвана корекция.

За тази цел са въведени размитите множества R и E – на правилните и на погрешните бази в съответния прочит r.

$$R_r(i) = s(r, i), \quad E_r(i) = cR_r(i) = 1 - s(r, i) \tag{3.17}$$

С тяхна помощ е въведено претеглено разстояние на Хеминг (3.18).

$$H^f(r, p) = \frac{\sum_i (R_r \cap R_p)(i)[r_i \neq p_i]}{\sum_i (R_r \cap R_p)(i)} \tag{3.18}$$

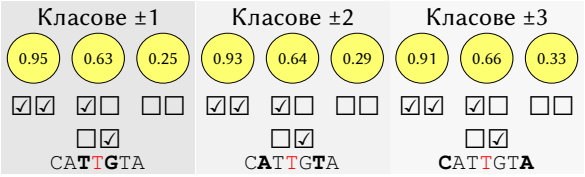
По аналогичен начин е въведено и по-комплексното претеглено разстояние на Левенщайн (3.19), което може да бъде видяно на страница 85 от дисертацията.

Използване на невронни мрежи и други средства на изкуствения интелект

В основата си, задачата е за откриване на зависимост между данните и наличието на грешки. Тази зависимост разчита на много фактори, които са или неизвестни, или които не могат да бъдат лесно формализирани. В същото време, самите изследвани данни съдържат множество хаотични, неструктурирани и трудни за формализиране зависимости. По тази причина апаратът на невронните мрежи е използван в този дисертационен труд за тази задача, а апаратът на случайните гори е бил изпробван като алтернатива на невронните мрежи. Начина на приложението им е описан в детайли в 3.6.

Вход

На изхода си обучените невронна мрежа и случайна гора трябва да класифицират дадена база от даден оценяван прочит като правилна и погрешна. В използвания тук модел, всички входни стойности са честоти на срещане на оценяваната база на съответната позиция в различни подмножества на пълното множество от прочити. Използвани са групирания с по три групи по сходство. Вземайки двойки съседни бази, прочитите се разделят на такива, които съвпадат с оценявания прочит в две, една или нула от тях. Такива групирания се правят за поредица от съседни двойки на различни отстояния, като всяка двойка осигурява три входни стойности. Това е илюстрирано във фигура 3.6



Фигура 3.6: Примерен вход за невронната мрежа

Обучение с виртуални грешки

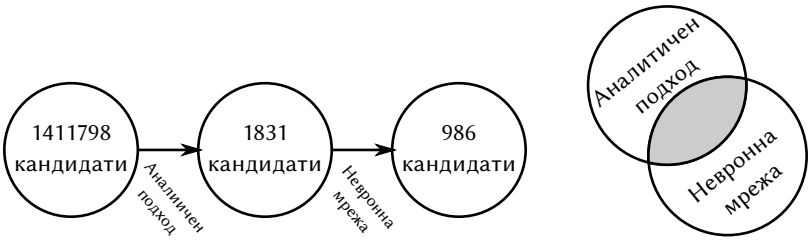
Тъй като дори с помощта на симулация на грешки не могат да се получат достатъчно голям обем обучаващи данни за невронната мрежа или за случайната гора, в 3.6.2 е предложен подход за обучение, който използва „виртуални“ грешки. За всяка възможна позиция се симулира „виртуална“ грешка – пресмята се допълнителен вектор от входни стойности, които биха се получили, ако само тази база бъде подменена с друга. На обучавания модел се подават два обучаващи примера – един с виртуалната грешка, и един без нея.

Използване на получените модели

С помощта на виртуални грешки са обучени модели на невронни мрежи и случайни гори, използвайки софтуерната програма Weka [156]. Тези модели не са използвани направо, а като допълнителен класификатор.

Те са прилагани след като е бил използван аналитичния подход за избор на потенциални грешки, които да бъдат подложени на по-

прецизна класификация. Получената комбинация е илюстрирана на фигура 3.7, където се вижда как аналитичния подход подбира ограничен набор от потенциални грешки сред множеството на всички бази в прочитите, след което невронната мрежа бива използвана за допълнителното му ограничаване (до сечението между множествата грешки предсказани от двата подхода).



(3.7а) Верига на филтриране на потенциалните грешки

(3.7б) Използвано сечение

Фигура 3.7: Комбинация на аналитичния подход с невронна мрежа

Глава 4: Софтуерна реализация

Както предварителната обработка, така и извършеният експериментален анализ за оценка на предложените в този дисертационен труд методи, наложиха разработването на множество от софтуерни инструменти за тяхната автоматизация. Беше направен извод, че обединяването на тези инструменти в обобщена система за изпълнение на работни потоци значително би улеснило бъдещи изследвания. За целите на това обединение беше създаден език за описание на работни потоци на основата на езика YAML, който е реализиран като част от софтуерен пакет, наречен untangle.

Описанието на един работен поток представлява колекция от декларации на целеви множества от данни (цели), които се конструират чрез изпълнение на операция върху част от останалите множества данни (целеви или входни). Като операции са достъпни както набор от вградени в софтуерния пакет такива, така и произволни външни инструменти, които могат да бъдат свободно декларирани. Освен това всеки работен поток може да бъде използван на мястото на операция в други работни потоци. Въведен е и език за описание на кратки изрази, подобни на изразите в езика C, които след компилация се представят като

работни потоци.

Системата за изпълнение на работни потоци е реализирана като част от софтуерния пакет untangle, използващ Python версия 2.7. Използвана е асинхронната софтуерна рамка Twisted [163], която позволява както паралелно изпълнение на различни инструменти за разработка, така и бъдещо развитие, включващо разпределено изпълнение, отдалечено управление и интеграция с мрежови услуги. Включени са инструменти за команден ред, които изпълняват описаните работните потоци.

Използвана е модулна структура, която позволява бъдещо разширение, използване на инструментите включени в нея от външен софтуер, както и модел за изграждане на алтернативни реализации на предложените в този труд методи и решения.

В текста на дисертацията е представено подробно описание на създадения език, както и информация за софтуерната му реализация.

Глава 5: Анализ на резултатите от проведените експерименти

За да бъдат оценени предложените в този труд подходи бяха проведени поредица от експерименти, целящи да потвърдят техните качества. Първоначално беше измерено разпределението на грешките, както за нуждите на симулатора на такива, така и за потвърждение на предварителните очаквания и избор на параметрите на методите. В следствие на това, че всеки от предложените методи бе прилаган под формата на разширение или подобрение на друг метод, в същинските експерименти бяха сравнени резултатите, постигнати със и без всяко отделно подобрение. Посочената градация на методи е илюстрирана на фигура 5.1, където се вижда как всеки следващ метод прави по-консервативен избор на потенциалните грешки.

Изследване на разпределението на грешките

За да бъде използван методът за симулиране на грешки от 3.4.2.2, както и за нуждите на пресмятане на променлив праг, бяха извършени експерименти за оценка на количеството грешки в данните. Беше установено, че разпределението съвпада с очакването и оценките, получени от други автори [47]. Това е разгледано по-подробно в 5.1.

(5.1а) Сходство и невронна мрежа

(5.1б) Остро сходство и случайна гора

Фигура 5.1: Градация на методите

Експериментални резултати за аналитичния подход

Бяха проведени редица от експерименти, които потвърждават ползата от използване на функция за локално сходство. В тях бяха сравнени резултатите, получени с наивния подход, и с подхода, използващ претегляне по сходство, като за валидация бяха използвани симулирани грешки по подхода от 5.3, показан на фигура 3.5. Бяхе сравнени и двете представени функции за локално сходство – стандартната и острата. Пълното сравнение е представено в 5.2.1 от текста на дисертацията.

Корекции				Изпуснати грешки			
Без тегла	Със тегла	Разлика		Без тегла	Със тегла	Разлика	
1548	1451	-97	(-6%)	392	391	-1	(-0.2%)
813	698	-115	(-14%)	386	391	+5	(+1.3%)
699	613	-86	(-12%)	415	412	-3	(-0.7%)
391	250	-141	(-35%)	445	448	+3	(+0.7%)
610	518	-92	(-15%)	70	69	-1	(-1.4%)
287	232	-55	(-19%)	73	74	+1	(+1.4%)
280	204	-76	(-27%)	81	82	+1	(+1.2%)

Таблица 5.1: Отношение на броя корекции и пропуснатите грешки

В таблица 5.1 са показани разликите между двата метода в направените корекции и изпуснатите симулирани грешки. Вижда се, че резултатът потвърждава очакването подобрение от използването на локалното сходство като тегла. На всеки един ред в таблицата се наблюдава намаляване на броя на корекции при използване на степента на сходство, и не се забелязва значително увеличаване на броя на пропуснатите грешки в следствие от използването ѝ. От това може да се заключи, че използване на модификацията, предложена в 3.4.1.2, води до предвидимо подобряване на получените резултати спрямо наивния подход. Тук

е използвана стандартната функция за сходство. Този резултат е онагледен и на фигура 5.3 на страница 115 от дисертацията.

В 5.2.2 са представени проведените отделни експерименти, които по подобен начин потвърждават ползата от използване на променлив праг, което е показано в таблица 5.2 на страница 118. В 5.2.3 са проведени експерименти, които потвърждават ползата от използване на острата функция за сходство, което е показано в таблица 5.3 на страница 119.

Приложение на невронни мрежи и други инструменти на ИИ

С помощта на софтуерната система Weka [156] беше обучена изкуствена невронна мрежа и още 19 модела на изкуствения интелект. За тази цел бяха използвани виртуални грешки по начина, описан в 3.6.2, като за входни данни бяха използвани честоти на срещане, групирани по начина, описан в 3.6.1.2. Преди да бъдат комбинирани с аналитичния подход, отделните модели бяха тествани върху деление на обучаващото множество, както и върху отделно генерирано множество, получено от друг отрязък от данните. Така бяха сравнени както различните модели, така и някои от техните параметри.

Всички модели постигнаха относително добра класификационна точност, но при нито един от моделите тя не беше достатъчно висока, за да бъде използван той независимо, без комбинация с аналитичния подход. След пълния набор от експерименти беше избрано да бъде използвана случайна гора като алтернатива на невронната мрежа за извършване на по-задълбочено сравнение на получените резултати. В 5.3.1.1 от текста на дисертацията са обсъдени експериментите, извършени с модела на изкуствени невронни мрежи, в 5.3.1.2 и 5.3.1.3 са обсъдени експериментите са сравнение на параметрите на модела, а в 5.3.1.4 е направено пълно сравнение на всички 20 разгледани модели.

Приложение на методи на ИИ върху кандидати, избрани с аналитичния подход

Обучените невронни мрежи и случайни гори бяха подложени на пълна валидация с вече използваната процедура от 3.4.2.3. При валидацията моделите не се използват направо, а комбинирани с аналитичния подход по начина, описан в 3.6.5 (и показан на фигура 3.7 на страница 92).

Резултатите от сравнението на аналитичния подход с комбинацията му с модели на изкуствения интелект са показани в таблица 5.9. Както се

Корекции						Изпуснати грешки					
Наив.	Сходство		ИНМ		Случ. гора	Наив.	Сходство		ИНМ		Случ. гора
1908	1655	-13.2%	896	-53.0%	875 -54.1%	398	385	-3.3%	388	-2.5%	450 +13.1%
						406	388	-4.4%	392	-3.4%	450 +10.8%
723	581	-19.6%	372	-48.5%	385 -46.7%	70	67	-4.3%	69	-1.4%	67 -4.3%
						56	54	-3.5%	54	-3.5%	54 -3.5%
2374	2107	-11.2%	1132	-52.3%	1093 -53.9%	419	395	-5.7%	400	-4.5%	462 +10.3%
						353	332	-5.9%	336	-4.8%	378 +7.0%
892	754	-15.4%	470	-47.3%	478 -46.4%	49	47	-4.1%	47	-4.1%	48 -2.0%
						35	34	-2.8%	35	=	35 =
4534	3882	-14.3%	1827	-59.7%	1748 -61.4%	351	320	-8.8%	330	-6.0%	380 +8.3%
						338	309	-8.5%	315	-6.8%	363 +7.3%
1529	1347	-11.9%	798	-47.8%	785 -48.6%	40	39	-2.5%	41	+2.5%	40 =
						51	49	-3.9%	48	-5.8%	49 -3.9%

Таблица 5.9: Сравнение на резултата с използване на ИНМ и случайни гори

вижда, използването на модел на ИИ като допълнителен класификатор води до повишаване на класификационната точност във всеки един от случаите. Забелязва се двойно намаляване на броя на извършените корекции, без да се забелязва повишаване на броя изпуснати симулирани грешки. При изкуствените невронни мрежи този резултат е по-изразен. Това е разгледано в повече подробности в 5.3.2.

Тези резултати са постигнати с използване на стандартната мярка за локално сходство. В 5.3.2.2 от текста на дисертацията са обсъдени и експерименталните резултати, постигнати с острата мярка за сходство, където се забелязва, че до голяма степен ползата от нейното прилагане се припокрива с ползата от използването на модел на ИИ, но въпреки това тя има потенциал за бъдещо развитие.

Сравнение със съществуващи решения

Прилагайки използваната процедура за валидация, предложените подходи бяха сравнени със SHREC [134].

Корекции					Изпуснати грешки				
SHREC	ИНМ		Случайна гора		SHREC	ИНМ		Случайна гора	
2321	1827	-21.284%	1748	-24.688%	374	330	-11.765%	380	+1.604%
					364	315	-13.462%	363	-0.275%
1370	920	-32.847%	897	-34.526%	52	52	=	52	=
					47	35	-25.532%	39	-17.021%

Таблица 5.12: Сравнение на комбинирания с ИИ метод и SHREC

В таблица 5.12 са показани резултатите от сравнението на SHREC с предложения в този дисертационен труд комбиниран подход. Както

може да се види, при тези условия аналитичния метод, комбиниран с модел на ИИ, успява да постигне чувствително по-добър резултат при търсене на грешките от SHREC. При сходен или по-малък брой пропуснати грешки, SHREC прави чувствително повече корекции. При използването на невронни мрежи, такъв категоричен резултат беше успешно постигнат във всички изпробвани случаи.

Бяха проведени експерименти, сравняващи SHREC директно с аналитичния подход, без използване на невронни мрежи, но тогава преимуществото не беше толкова явно видимо. Всички резултати са обсъдени по-подробно в 5.4 от текста на дисертацията.

Глава 6: Приноси и перспективи

Предложен е набор от средства за откриване и корекция на грешките в метагеномни данни от паралелно секвениране, с помощта на използване на инструментите на информатиката и изкуствения интелект. Качества на предложените средства са изследвани в детайли с помощта на разработения индиректен подход за валидация.

Теоретико-методични приноси на дисертационния труд

1. Разработен е аналитичен подход за корекция на грешки в метагеномни данни от паралелно секвениране. Той въвежда честота на срещане, претеглена по сходство, както и две различни мерки за сходство, локални за региона, изследван за потенциални грешки (3.4.1.2, 3.4.1.3).

2. Разработен е индиректен подход за валидация на различни методи за корекция на грешки върху данни от паралелно секвениране, в които грешките не са предварително известни. (3.4.2.3) За целите му е разработен и подход за симулация на грешки, който да бъде използван при процедурата на валидация (3.4.2.2).

3. Въведен е размит индикатор на увереност за приложение на оценка на грешките при анализ на секвенционни данни без корекция на грешките (3.5). С негова помощ са въведени претеглено разстояние на Хеминг H^f и претеглено разстояние на Левенщайн L^f.

4. Предложен е модел на невронни мрежи, който да може да бъде обучен да класифицира потенциалните грешки с по-голяма точност от прякото приложение на аналитичния подход (3.6). Той е

основан на разбиване на честотата на срещане в отделни групи по позиционно сходство (3.6.1.2).

5. Предложен е подход за обучение на модели на машинното самообучение за откриване на грешки без да са налични обучаващи данни с действителни грешки. Той използва виртуални грешки на всички разглеждани позиции (5.3.1). Предложеният подход е независим от данните и по начина си приложение наподобява самообучение без учител, което предразполага към пълната автоматизация на обучението и приложението му.

6. Създаден е модел на система за извършване на сложен анализ на секвенционни данни, като е въведен език за описание на управляеми работни потоци, който е реализиран в софтуерния пакет untangle (4.1).

Експериментални приноси на дисертационния труд

1. Постигнато е експериментално потвърждение на ползата от въведената претеглена честота на сходство (5.2.1).

2. Постигнато е експериментално потвърждение на ползата от прилагане на променлив праг при откриване на грешките с честоти (5.2.2).

3. Обучени са невронна мрежа и 19 други модела на машинното самообучение. Извършен е експериментален анализ на възможните параметри на невронната мрежа (5.3.1.2), на потенциала на другите модели (5.3.1.4), както и на някои възможности за подобрение. В процеса на анализ е потвърдена ползата от прилагане на локална мярка за сходство (5.3.1.2).

4. Постигнато е експериментално потвърждение на ползата от допълване на аналитичния подход за корекция на грешки с модел на невронна мрежа или случайна гора (5.3.2).

5. Постигнати са експериментални резултати, потвърждаващи потенциала за използване на предложените методи за други задачи като откриване на точкови мутации в генома на пшеницата (5.5).

Списък с авторски разработки и предложения

1. Претеглена честота на срещане. (3.4.1.2)

2. Локална мярка за сходство на символни низове, претеглена с отстоянието от дадена позиция. (3.4.1.2)

3. Индиректен метод за валидация, използващ многократна корекция и размити множества. (3.4.2.3)

4. Симулатор на грешки. (3.4.2.2)

5. Претеглено разстояние на Хеминг и претеглено разстояние на Левенщайн. (3.5)

6. Предложение за разбиване на честотата на срещане по позиционно сходство за обучение на невронни мрежи. (3.6.1.2)

7. Предложение за използване на виртуални грешки за обучение на невронни мрежи при тяхната класификация. (5.3.1)

Перспективи за бъдещо развитие

Представените в този труд методи за откриване и корекция на грешки представляват добра основа за бъдеща работа, в която да се обхване по-широк кръг от задачи, както и да се навлезе в по-задълбочени изчислителни изследвания.

1. Разгледаните методи имат потенциала да бъдат приложени за класификационни задачи за откриване на структурни варианти като точкови мутации. Откритите вариации, които след използването на допълнителните подобрения (сходство, променлив праг или невронни мрежи) са рекласифицирани като правилни бази, са кандидати за точкови мутации. При доближаване на броя на пропуснати грешки до 0, било то и с цената на голям брой нежелани корекции, останалите вариации могат да бъдат категорично определени като точкови мутации.

2. Разгледаните методи могат да намерят приложението си към данни, различни от метагеномните. Добър пример е геномът на пшеницата. При него усложненията, предизвикани от наличието на три подгенома, които съдържат различни версии на отделните гени, могат да бъдат разрешени с използването на методи, разчитащи на мерки за сходството.

3. Могат да се приложат алтернативни методи на валидация, основани на сравнения на филогенетични дървета. Например, може

да бъде използвана оценката за правдоподобие на филогенетично дърво, построено по метода на максимално правдоподобие.

4. Може да се разработи симулатор на грешки, който използва много по-усъвършенстван статистически модел.

Публикации по темата на дисертационния труд

[C1] Milko Krachunov and Dimitar Vassilev. "An approach to a metagenomic data processing workflow". In: *Journal of Computational Science* 5 (2014), pp. 357–362. DOI: 10.1016/j.jocs.2013.08.003.

[C2] Milko Krachunov. "Denoising of Metagenomic Data from High-Throughput Sequencing". In: *Advanced Research in Mathematics and Computer Science*. Sofia, 2013, pp. 67–76.

[C3] Milko Krachunov. "Hierarchy and expressions for automated workflows for NGS data processing". In: *Proceedings of the 8th International Conference on Information Systems & Grid Technologies (ISGT)*. [in press]. Sofia, May 30-31, 2014.

[C4] Milko Krachunov, Ognyan Kulev, Valeriya Simeonova, Maria Nisheva, and Dimitar Vassilev. "Manageable Workflows for Processing Parallel Sequencing Data". In: *Serdica Journal of Computing* 8.1 (2014), pp. 1–14. ISSN: 1312-6555.

[C5] Milko Krachunov, Maria Nisheva, and Dimitar Vassilev. "Intelligent Approach for Automated Error Detection in Metagenomic Data from High-Throughput Sequencing". In: *Proceedings of the 7th International Conference on Information Systems & Grid Technologies (ISGT)*. Sofia, June 31-1, 2013, pp. 160–168.

[C6] Milko Krachunov, Peter Petrov, Ivan Popov, and Dimitar Vassilev. "Computational challenges in a metagenomics processing pipeline". In: *Proceedings of the 6th International Conference on Information Systems & Grid Technologies (ISGT)*. Sofia, June 1-2, 2012, pp. 302–311.