На правах рукописи

Тарасов Денис Станиславович

Компьютерный метод систематики микроорганизмов на основе алгоритмической теории информации и его приложение к таксономии и номенклатуре микроскопических грибов рода Trichoderma

03.00.07- 03 микробиология

Автореферат
диссертации на соискание ученой степени
кандидата биологических наук

Казань, 2007

Работа выполнена на кафедре генетики ГОУ ВПО “Казанский государственный университет им В.И. Ульянова-Ленина», г. Казань.

Научный руководитель: *кандидат биологических наук Акберова Наталья Иванована*

Официальные оппоненты: *доктор биологических наук Наумова Римма Павловна*

доктор ветеринарных наук Алимов Азат Миргасимович

Ведущая организация: *Казанский институт биохимии и биофизики КазНЦ РАН, г.Казань*

Защита состоится 29 ноября **2007 г.** в **13 ч 00** мин. на заседании диссертационного совета Д. 212.081.08 при Казанском государственном университете по адресу г. Казань Кремлевская 18

С диссертацией можно ознакомиться в Научной библиотеке им. Н.И. Лобачевского Казанского государственного университета

Автореферат разослан “29” октября 2007 года

Ученый секретарь диссертационного совета, доктор биологических наук Абрамова З.И.

Тарасов Денис Станиславович
Казанский государственный университет, биолого-почвенный факультет
420008, Казань, Кремлевская 18
Факс: (843)238721
E-mail: dtarasov@mntech.ru

Актуальность темы

Систематика организмов имеет две цели:

1. теоретическую - установление взаимосвязей между признаками различных организмов
2. практическую - идентификация организмов, предсказание свойств идентифицированных организмов на основании их принадлежности к группе

Систематика включает в себя три большие области:

1. Номенклатура занимается вопросами выбора имен для систематических групп.
2. Таксономия занимается таксонами и взаимосвязями между ними
3. Идентификация - отнесение организма к конкретному таксону

Современная систематика микроорганизмов сталкивается со значительными проблемами.

В области номенклатуры. Современная номенклатура в систематике регулируется с помощью устоявшихся наборов правил, ведущих свое начало со времен Линнея. Существует Зоологический кодекс номенклатуры, Ботанический кодекс, Бактериологический кодекс и Вирусный кодекс. Эти кодексы номенклатуры вызывают много нареканий. Указывается, например, на то, что при работе в рамках задаваемых ими правил систематик микроорганизмов вынужден тратить значительное время (до 20% всего рабочего времени) на номенклатурные/историко-библиографические изыскания, вместо того, чтобы заниматься предметом своих исследований.

В области таксономии. Систематика микроорганизмов сталкивается с рядом проблем, обусловленных спецификой изучаемого объекта:

- большое разнообразие микроорганизмов;
- отсутствие полового процесса у многих микроорганизмов не позволяет использовать определение вида по признаку скрещиваемости;
- горизонтальный перенос генов размывает границы видов и усложняет реконструкцию филогении;
- высокая скорость мутационных изменений

В систематике микроорганизмов используются различные группы признаков, такие как морфологические, физиологические, биохимические, молекулярно-генетические. Широко признается тот факт, что данных одной группы признаков недостаточно для установления взаимосвязей между таксонами. Но в связи с тем, что на сегодняшний день отсутствуют действенные методы интеграции морфологических, физиологических, биохимических и молекулярно-генетических данных в рамках единого подхода, на практике часто 1данные одной группы признаков.

В отчете 2006 года Американской Академии Микробиологии сделан вывод о том, что используемые сегодня средства систематики не способны адекватно учитывать существующее разнообразие микроорганизмов, что является обоснованием необходимости разработки принципиально новых подходов к систематике микроорганизмов.

Одним из перспективных подходов к систематике является применение алгоритмической теории информации, которая использовалась ранее для создания метода построения филогенетических деревьев на основании сравнения целых геномов.

Предполагается, что использование алгоритмической теории информации можно распространить и на другие группы признаков, а также на другие вопросы систематики, такие как описание свойств микроорганизмов и вопросы номенклатуры.

Цель работы

Целью данной работы было создание компьютерного метода систематики микроорганизмов на основе алгоритмической теории информации

Для выполнения работы были поставлены следующие основные задачи:

1. Разработка способа унифицированного описания морфологических, физиологических, биохимических и молекулярно-генетических признаков, пригодный для использования в компьютерных классификационных процедурах
2. Разработка компьютерных классификационных процедур для построения таксономических деревьев, использующих унифицированные описания признаков и метрику расстояния, основанную на алгоритмической сложности
3. Проверка разработанных процедур на практике

Научная новизна

Впервые создан метод, позволяющий использовать универсальную меру расстояния (нормализованное информационное расстояние) при анализе таксономии микроорганизмов с использованием морфологических, физиологических и биохимических признаков. Разработанный метод сочетает в себе преимущества нумерического и традиционного (интуитивного) подходов к систематике и обладает следующими преимуществами:

- Исключаются проблемы связанные с произвольным выбором меры расстояния и сводится к минимуму эффект от выбора алгоритма кластеризации.

28. Izotova E.D. Virtual Machine for Analyzing Living Systems/E.D. Izotova, D.S. Tarasov //International Moscow conference on computational molecular biology.- M.- 2007.- p. 128-130
29. Tarasov D.S. Object orientation and biological taxonomy: applying programming concepts to species classification/D.S. Tarasov, E.D Izotova, N.I. Akberova//International Moscow conference on computational molecular biology.- M.- 2007.- p. 290-292

- Метод позволяет включать знания и личный опыт систематика посредством выбора средств кодирования признака в программе-описании. В отличие от матрицы признаков, используемой в других нумерических методах, программа-описание способствует более обдуманному подходу к процессу выбора и кодирования признаков, сохраняет логику принятых в этом процессе решений для последующего анализа другими исследователями.
- Разработанный метод не требует использования строго независимых признаков. В описания-программы могут включаться связанные признаки, *одновременно* с информацией о способах их взаимодействия и развития процесса во времени.

Практическая значимость работы

Разработанный метод может использоваться во всех задачах таксономии микроорганизмов, где обычно используется нумерическая систематика и кластерный анализ.

Разработанное программное обеспечение и язык ConceptSystem может быть применен в практической работе исследователя-микробиолога, а также в учебном процессе.

Предложенные улучшения к микробиологической номенклатуре, основанные на использовании пространств имен и псевдонимов, могут быть использованы в практической работе с систематикой микроорганизмов, поскольку являются совместимыми с существующими номенклатурными правилами, что одновременно упростит работу.

Методы исследования

Программы для синтаксического разбора формализованных описаний микроорганизмов, графический интерфейс пользователя для работы с описаниями, программа, вычисляющая приблизительное значение алгоритмической сложности и программа построения матрицы расстояний были написаны с помощью языка программирования F#. Для сжатия данных описаний использовался алгоритм gzip. Классификационные деревья строились с помощью алгоритмов UPGMA, Neighbor-Joining, и метода минимального эволюционного расстояния, реализованных в программном пакете PHYLIP, и визуализировались с помощью пакета PhyloDraw. При разработке языка ConceptSystem использовалась технология объектно-ориентированного программирования. Для построения объектно-ориентированных классификаций *Trichoderma* использовался графический язык UML (Unified Modeling Language), для создания UML диаграмм использовалась программа UMLet.

Апробация работы

Результаты работы докладывались на международных конференциях Bioinformatics of Genome Regulations and Structure'2002 и 2004, Moscow Conference on Computational Molecular Biology, 2003 и 2007. Кроме того, результаты работы докладывались на 6-ой (2002), 7-ой (2003), 8-ой (2004) и 9-ой (2005) Пущинских школах-конференциях молодых ученых «Биология - наука XXI века», секция «Математическая биология», и XII Международной конференция студентов, аспирантов и молодых ученых «Ломоносов 2005».

Публикации

По теме диссертации опубликовано 29 печатных работ, в том числе 19 тезисов конференций, 4 трудов международных конференций, 6 статей в научных журналах , в. т. ч. 4 - в изданиях, рекомендованных ВАК для публикации результатов кандидатских диссертаций, 1 зарегистрированная программа для ЭВМ.

Объем и структура диссертации

Диссертация состоит из введения, трех глав, заключения и списка литературы, изложена на 110 страницах. Работа включает 25 рисунков и 8 таблиц.

амплификаторов/Д.С. Тарасов, Н.И. Акберова //III Научная конференция молодых ученых, аспирантов и студентов научного-образовательного центра КГУ "Материалы и технологии XXI века".- Казань.- 2003. - с. 84

20. Тарасов Д.С. Язык описания клеточных программ CDPL-1 и его применение/Д.С. Тарасов, Н.И. Акберова , А.Ю. Леонтьев//6-ая Пущинская школа-конференция молодых ученых "Биология - наука XXI века".- Пущино.- 2002. - с. 195-196
21. Tarasov D.S. The model of molecular biological computational device and its application to automatic genome annotation/D.S. Tarasov, N.I. Akberova, A.Y. Leontiev//International Moscow conference on computational molecular biology.- M..- 2003. - p. 225-226
22. Тарасов Д.С. Компьютерное моделирование структурно-функциональной огранизации ori-сайтов бактерий/Д.С. Тарасов, Н.И. Акберова//12-я международная конференция "Ферменты микроорганизмов", Казань 2001. - с 83-84
23. Тарасов Д.С. Регуляция и контроль инициации репликации: лингвистический подход/Д.С. Тарасов, Н.И. Акберова, А.Ю. Леонтьев//Материалы XL международной научной студенческой конференции «Студент и научно-технический прогресс»:Биология.- Новосибирск.- 2002.- с. 149-150
24. Леонтьев А.Ю. Алгоритм построения образа функциональных областей генома/А.Ю. Леонтьев, Д.С. Тарасов//Материалы международной научной конференции, посвященной 70-летию образования зооиженерного факультета. Казанская государственная академия ветеринарной медицины.- 2000.- с. 203-205
25. Тарасов Д.С. Язык представления описания морфологии грибов на примере Trichoderma/Д.С. Тарасов, Р.И. Тухбатова//9-ая Пущинская школа-конференция молодых ученых "Биология - наука XXI века".- Пущино.- 2005. - с. 331.
26. Шишкин А.В. Построение нетривиальной классификации грибов рода Trichoderma/А.В. Шишкин, Р.И. Тухбатова, Д.С. Тарасов// 9-ая Пущинская школа-конференция молодых ученых "Биология - наука XXI века".- Пущино.- 2005. - с. 223.
27. Тарасов Д.С. Интерпретатор языка CONCEPTSYSTEM// Программа для ЭВМ. Зарегистрирована в Реестре программ для ЭВМ 18.01.2007. Свидетельство о регистрации № 2007610350

Internet-конференция "Компьютерное и математическое моделирование в естественных науках.- Тамбов.- 2001. - с. 23
11. Акберова Н.И.Компьютерный дизайн ПЦР-праймеров различной специфичности/Н.И. Акберова, Д.С. Тарасов //Четвертая всероссийская Internet-конференция "Компьютерное и математическое моделирование в естественных науках".- Тамбов.- 2002. - с. 32
12. Акберова Н.И. Метод симметрийного моделирования структуры ДНК-текстов/Н.И. Акберова, А.Ю. Леонтьев, Д.С. Тарасов//Первая всероссийская Internet-конференция "Компьютерное и математическое моделирование в естественных науках.- Тамбов.- 2001. - с. 24
13. Тарасов Д.С. Архитектура клеточного устройства и гибридные биокибернентические системы/Д.С. Тарасов, Н.И. Акберова//7-ая Пущинская школа-конференция молодых ученых.- Пущино.- 2003. - с. 256
14. Тарасов Д.С.Компьютерный дизайн праймеров для ПЦР/Д.С. Тарасов, Н.И. Акберова//II научная конференция молодых ученых, аспирантов и студентов научно-образовательного центра КГУ.- Казань.- 2001. - с. 91
15. Тарасов Д.С. Молекулярно-биологическое вычислительное устройство и клеточное киберпространство/Д.С. Тарасов, Н.И. Акберова//8-ая Пущинская школа-конференция молодых ученых "Биология - наука XXI века".- Пущино.- 2003. - с. 24
16. Тарасов Д.С. Объекто-ориентированная система описания, классификации и моделирования биологических объектов и ее применение к грибам рода Trichoderma/Д.С. Тарасов//XII Международная конференция студентов, аспирантов и молодых ученых "Ломоносов".- М.- 2005. - с. 40-41
17. Тарасов Д.С. Организация базы знаний для молекулярно-биологических исследований/Д.С. Тарасов, Н.И. Акберова //IV Научно-практическая конференция молодых ученых и специалистов Республики Татарстан.- Казань.- 2001. - с. 110
18. Тарасов Д.С. Применение новых достижений молекулярной биологии при проектировании современных устройств микроэлектроники/Д.С. Тарасов, Н.И. Акберова//Новые методологии проектирования устройств микроэлектроники.- Владимир.- 2002. - с. 75-76
19. Тарасов Д.С. Технологии молекулярной биокибернетики: использование системы CDPL/CDS в конструировании ПЦР-

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Введение

Во введении формулируется проблема, дается краткое описание состояния современной микробиологической систематики, ставится цель работы и задачи. Обосновывается формулировка цели и задач.

Первая глава. Обзор литературы

Современная биологическая систематика имеет длительную историю. Существовало и существует множество различных направлений, часто называемых по-разному в различных источниках. Первоначально целью биологической систематики было построение некоторой «естественной» системы живых организмов.

Систематика в это время опиралась на философскую доктрину, известную как «*эссенциализм*». Под эссенциализмом обычно понимается точка зрения, согласно которой для любого предмета или существа возможно перечислить набор характеристик, наличие которых необходимо и достаточно для определения его принадлежности к определенной группе. Эти характеристики выражают «сущность» этого предмета. Эссенциализм можно видеть в учении древнегреческого философа Платона об идеях.

Позднее Аристотель впервые вводит иерархический принцип в систематику, говоря о том, что каждая вещь характеризуется родом (то, что есть у нее общего с другими вещами) и видом – конкретной спецификой данной вещи.

В XVIII веке Карл Линней вводит во всеобщее употребление биномиальную номенклатуру. Кроме того, он вводит схему классификации по принадлежности организма к таксонам различных рангов (от конкретного к более общему), выделяя царство, класс, порядок, семейство, род и вид. Таким образом оформляются два из трех основных принципов номенклатуры современной систематики – иерархический принцип и биномиальные названия. Также оформляется и деление таксонов по рангам и названия самих рангов.

В 1867 году де Кандоль вводит третий основной принцип современной номенклатуры – принцип приоритета. Согласно ему за названием (родом и видом) организма закрепляется первое данное ему название, которое впоследствии не меняется. Необходимость этого принципа была обусловлена тем, что до этого общая практика состояла в том, что названия родов и видов постоянно пересматривались, в поисках «наиболее лучшего» названия, отражающего новые знания об этом виде,

что естественно создавало массу проблем для общения систематиков. Чтобы решить эту проблему «наиболее лучшее» название было принесено в жертву стабильности в названиях.

В рассмотренное время номенклатура продолжает совершенствоваться, а способ построения классификаций остается по сути неизменным. Для построения классификации исследователь на основании своей интуиции и личного опыта выбирает «существенные» признаки, т.е. признаки, характеризующие сущность организма, таким же образом выбирает наиболее важные из этих признаков и уже на основании этих признаков (как правило, небольшого их числа) строит классификацию.

По мере того, как среди биологов распространяется убеждение о том, что виды не являются неизменными, цель систематики смещается от открытия «естественной системы», которой видимо, не существует, к удовлетворению практических потребностей. С другой стороны, с развитием эволюционных представлений был выдвинут тезис о том, что систематика должна отражать родство живых организмов. Соответственно в систематике оформляются различные направления

Методы фенетики

В противоположность идее о наличии «существенных» признаков нумерическая систематика основана на количественном учете всех признаков (точнее, большого числа различных признаков).

Возникновение идеи нумерической систематики связывают с именем французского ботаника Адансона, жившего в XVIII веке. Он впервые предположил, что отдельные признаки не имеют устойчивого значения, и только совокупность всех признаков может являться целостной систематической характеристикой. Точный учет большого количества признаков, остается, однако, во времена Адансона непосильной задачей.

В 60-х годах XX века П. Снит и Р. Сокал, работая над проблемой систематики бактерий, разрабатывают принципы и методы количественной фенетики, опирающиеся на использование компьютеров для проведения вычислений.

В фенетическом анализе используется очень большое (200-300 или столько, сколько можно практически определить) число признаков. Первоначально предполагалось использование только невзвешенных и независимых признаков, однако последующие работы рассматривали также применение взвешенных признаков. По степени сходства организмов (т.е. по количеству совпадающих признаков) строится матрица, по которой затем производится кластерный анализ и строится фенограмма.

Сторонники фенетики считают этот метод объективным, поскольку он теоретически не зависит от субъективной оценки «важности» признаков

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Тарасов Д.С., Формат представления биологических описаний гриба и его применение на примере рода Trichoderma/Д.С. Тарасов, Р.И. Тухбатова, Н.И. Акберова, Ф.К. Алимова//Вестник Татарстанского отделения российской экологической академии. – 2005 .- №2(24).- С. 44-49
2. Тарасов Д.С. Молекулярно-биологическое устройство – принципы организации//Д.С. Тарасов, Н.И. Акберова//Ученые записки КГУ, серия естественные науки.- 2005. - т.147. - кн. 2. - с 180-196
3. Тарасов Д.С. Увеличение интеллектуальных ресурсов научного исследования в биологических областях/Д.С. Тарасов, Н.И. Акберова//Ученые записки КГУ, серия естественные науки.- 2006 - т.148. - кн 1.- с.138-155
4. Тарасов Д.С. Применение принципов объектно-ориентированного программирования к описанию и классификации биологических объектов на примере грибов рода *Trichoderma/* Д.С. Тарасов, Н.И. Акберова, Р.И. Тухбатова, Ф.К. Алимова//Ученые записки КГУ, серия естественные науки.- 2006.- т. 148.- кн. 3. - с. 125-148
5. Тарасов Д.С. Виртуальные машины для исследования молекулярно-биологических процессов/Д.С. Тарасов, Н.И. Акберова//Георесурсы.- 2006.- №4(21).- с.45-48
6. Тарасов Д.С.Молекулярно-биологическое вычислительное устройство: архитектура и язык управления/Д.С. Тарасов, Н.И. Акберова//Новая Геометрия Природы.- Казань.- 2003. - с. 341-250
7. Тарасов Д.С. Применение концепции молекулярно-биологического устройства для создания современных обучающих программ нового поколения/Д.С. Тарасов, Н.И. Акберова//Новая Геометрия Природы.- Казань.- 2003. - с. 332-334
8. Tarasov D.S.A Language for modeling generic regulation in prokaryotes/D.S. Tarasov, A.Y. Leontiev , N.I. Akberova //4 th International conference of bioinformatics and genome regulation and structure.- Novosibirsk.- 2004. - p. 146-149
9. Tarasov D.S.Architecture of cell device/D.S. Tarasov, A.Y. Leotiev, N.I. Akberova//Third International conference of bioinformatics and genome regulation and structure.- Novosibirsk.- 2002. - p. 216-218
10. Акберова Н.И.Исследование симметрийной структуры геномов вирусов HSV/Н.И. Акберова, Д.С. Тарасов //Первая всероссийская

ВЫВОДЫ

1. Разработан метод кодирования морфологических и физиологических признаков микроорганизмов для применения универсальной информационной нормализованной меры расстояния и метод нахождения приблизительного значения этой меры. Метод заключается в представлении признаков организма в форме программы-модели его онтогенетического развития, записанной на специальном языке программирования.
2. Разработана программа-интерпретатор языка программ-моделей для описания признаков микроорганизмов, а также интерактивный графический интерфейс пользователя.
3. Использование новых методов продемонстрировано на примере описания и классификации грибов рода *Trichoderma,* что позволило получить описания, интегрирующие морфологические и физиолого-биохимические признаки, а также построить более компактный по сравнению с принятым вариант систематики.

или от выбора ограниченного набора признаков. Реально, однако, набор исследуемых признаков всегда ограничен, имеются сложности с выбором исключительно независимых признаков. Существует огромное количество разнообразных мер расстояний – формул, по которым вычисляется степень сходства организмов и, кроме того, существует большое число алгоритмов построения фенограмм, дающих разные результаты. Поэтому фенетика не смогла принести в систематику ту ясность и однозначность, на которую надеялись ее сторонники.

Методы кладистической (филогенетической) систематики

Кладистическое направление в систематики возникло благодаря работам В. Хеннига. Слово «кладистика» происходит от греческого слова *κλάδος*, *klados – ветвь* (Хенниг для обозначения своего подхода употреблял термин «филогенетическая таксономия»). Сущность кладистического направления можно определить как классификацию организмов исключительно по их порядку ветвления на эволюционном дереве, а не в соответствии с морфологическим сходством.

Кладистика соответственно признает только монофилитические (происходящие от одного общего предка) систематические группы.

Монофилитические группы выделяются путем анализа признаков, которые присутствовали до появления последнего общего предка группы (плезиоморфные признаки) и признаки, появившиеся у последнего общего предка (синапоморфные признаки).

Разделение признаков на плезиоморфные и синапоморфные производится путем сравнения организмов некоторой группы с внешней группой (родственной группой, но не происходящей от последнего общего предка исследуемой группы).

Кладистический анализ можно производить по любым признакам, однако в последнее время часто используются данные о последовательности ДНК и РНК. Для построения кладограмм используются компьютерные алгоритмы, такие как метод максимальной парсимонии (MP) и максимального правдоподобия (ML). Эти методы часто требуют чрезмерно больших вычислительных ресурсов.

Современное состояние методов систематики

В настоящее время применяются как количественные методы (фенетика, кладистика), так и построение классификации на основании личного опыта систематика. Количественные методы часто критикуются за то, что при их использовании «настоящая» систематическая работа подменяется необдуманным использованием компьютерных программ. Количественные методы часто требуют использования более или менее произвольных числовых коэффициентов, метрик расстояния и.т.п. Кроме

того, компьютерная программа, как правило, не может объяснить, почему был получен тот или иной результат. В свою очередь сторонники количественных методов указывают на субъективность и невоспроизводимость результатов традиционных подходов.

Вторая глава. Разработка автоматизированного метода использования нормализованного информационного расстояния для таксономии микроорганизмов

Алгоритмическая теория информации в биологической систематике

Относительно недавно в работах рада авторов для использования в классификации биологических объектов была предложена «универсальная мера расстояния», основанная на понятии алгоритмической сложности. Данная мера выражается следующей формулой:

$$NID(x,y) = \frac{\max\{K(x \mid y), K(y \mid x)\}}{\max\{K(x), K(y)\}}$$

NID – нормализованное информационное расстояние между *x* и *y*; *x,y* – некоторые строки символов, содержащие информацию; *K(x|y)* – алгоритмическая сложность x при данном y.

Алгоритмическая сложность *K(x|y)* - это длина самой короткой двоичной программы для машины Тьюринга, которая, имея на входе *x*, выдаст на выходе *y*. Нормализованное информационное расстояние является универсальной мерой, которая отражает любое сходство *x* и *y*. В этом смысле *NID* является лучшей из возможных мер расстояний вне зависимости от природы сравниваемых объектов.

Проблема заключается в том, что *K(x|y)* является в общем случае невычислимой величиной, и для практических целей были предложены методы нахождения ее приблизительного значения. Данные методы оказались пригодными для вычисления *NID* между последовательностями ДНК (включая целые геномы) и дали хорошие практические результаты при построении филограмм.

Однако для морфологических и физиологических признаков микроорганизмов методы аппроксимации *NID* не были разработаны, а методы, предложенные для других групп признаков, не являются адекватными.

Кодирование морфологических и физиологических признаков микроорганизмов для применения информационной меры расстояния

Для того чтобы применить нормализованное информационное расстояние для морфологических и физиологических признаков, их надо

2. Номенклатура. В качестве эксперимента при построении ОО-систематики вместо микробиологических правил номенклатуры были использованы правила номенклатуры, принятые в программировании. Нами было произведено сравнение подходов к проблеме номенклатуры в биологической систематике и в программировании. В ходе проведенного сравнения нами был выдвинут ряд предложений касающихся того, как известные в программировании и информатике принципы могли бы быть использованы для решения проблем номенклатуры в биологии.

Выдвинутые предложения направлены на решение противоречия между необходимостью иметь уникальные и неизменные имена таксонов и потребностями в существовании простых, легко запоминающихся имен, отражающих существенные признаки таксона. Противоречие разрешается путем разделения имен на две группы. В качестве уникального и неизменного имени предлагается использовать GUID таксона. На GUID распространяется правило приоритета. В то же время GUID является лингвистически и культурно нейтральным идентификатором, который может генерироваться как локально (на компьютерах пользователей), так и центрально (на специальном сервере). GUID предназначен прежде всего для компьютерной обработки.

Для использования людьми GUID может быть с помощью технологии псевдонимов сопоставлен с несколькими удобными *локальными* именами.

Для предотвращения конфликта локальных названий предлагается использовать технологию пространств имен. Благодаря этому каждый исследователь или группа сможет использовать собственную предпочитаемую систему наименований для часто используемых объектов, без риска возникновения путаницы.

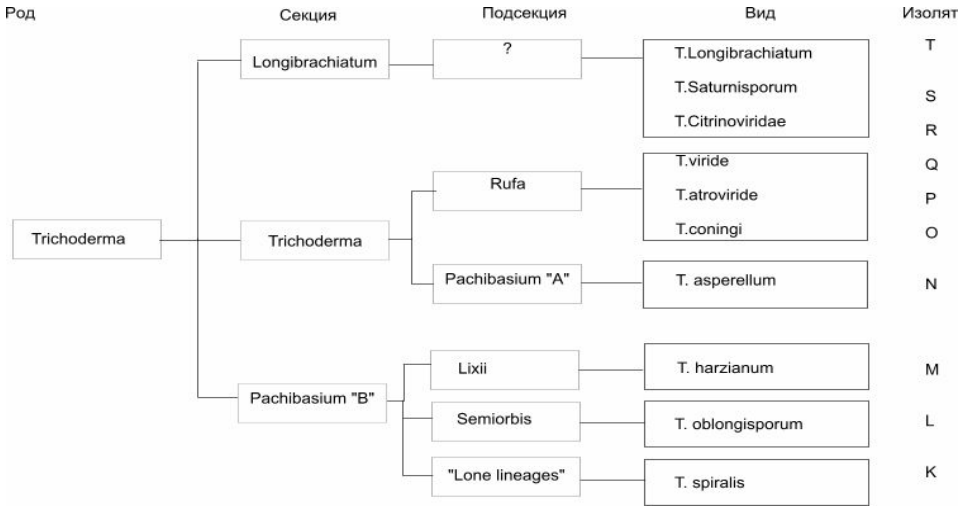

Рис 4. Фрагмент одной из «традиционных» классификаций рода Trichoderma. Справа приведено систематическое положение изолятов, которые использовались при построении ОО-систематики.

ОО-версия классификации, полученная в результате применения разработанного метода (рис 3), содержит 11 абстракций на 10 экземпляров, из которых имеется 9 видов 1 род и 1 промежуточный таксон. Между тем, как можно видеть из рисунка, ОО-классификация описывает взаимоотношения между рассматриваемыми организмами более подробно. В частности, в «стандартной» версии виды к которым относятся изоляты T, S и R поставлены в один ряд и дальнейшая информация об их взаимоотношениях отсутствует, в то время как в ОО-версии класс TExPig (R) является наследником класса TPigmented (S) и TExudated (T), из чего сразу следует что R имеет признаки как S так и T.

Следует особо отметить две особенности объектно-ориентированного подхода к систематики.

1. Использование множественного наследования. У бактерий распространенным явлением является горизонтальный перенос генов, а у микроскопических грибов часто имеется несколько ядер (в результате слияния клеток мицелия), содержащих гетерогенную генетическую информацию. В результате возможна ситуация когда штамм, вид или таксон более высокого ранга имеет нескольких предков одного уровня. Существующие схемы классификации не предусматривают такой ситуации. В ООП же имеется понятие множественного наследования, что позволяет расширить выразительные средства систематики без ущерба их строгости.

вначале представить в виде строки символов. Поскольку обычно морфологические и физиологические признаки представляются в виде текстовых описаний на естественном языке (русском, английском и.т.п.), на первый взгляд задача кажется очень простой.

Простое решение, однако, оказывается неадекватным. Метод приблизительного вычисления NID, описанный ранее, основан на использовании алгоритмов сжатия информации (используемых обычно для сжатия файлов на компьютере). Нельзя рассчитывать на то, что компьютерный алгоритм сжатия сумеет обнаружить в текстовых описаниях на естественном языке закономерности, отражающие структуру описываемого объекта.

Предлагаемое в настоящей работе решение состоит в следующем. Вместо текстовых описаний возможно использовать программы, записанные на специальном языке программирования. Каждая программа будет при запуске моделировать процесс развития морфологических и физиологических признаков данного организма. Модель может быть как очень приблизительной, так и достаточно детальной, при условии, что уровень детализации одинаков для всех сравниваемых групп организмов. Длина полученной программы будет приближением к значению $K(x|\varepsilon)$. После чего автоматический компьютерный алгоритм может быть использован для нахождения приблизительного значения $K(x|y)$.

Для целей записи программ-моделей организмов разработан специальный язык BMLIDA. (Biological Modeling Language for Information Distance Approximation).

Применение информационной меры расстояния к систематике *Trichoderma*

Существует множество групп живых организмов, систематика которых находится в процессе постоянного изменения. Одной из таких групп являются грибы рода *Trichoderma*. Грибы рода *Trichoderma* представляют ценность для применения в различных областях, в частности, для биологической защиты растений против грибных возбудителей болезней, а также как продуценты различных биологических активных веществ. Точное определение систематического положения изолятов необходимо для оценки их практических свойств.

До настоящего времени отсутствует естественная система, позволяющая выявить однозначные связи между определенными видами этого рода. В литературе отмечены большие изменения в систематике, вызванные, во-первых, пересмотром объема рода, во-вторых, делением его на секции и описанием большого числа новых видов.

На разрабатываемом языке BMLIDA были описаны модели развития 43 видов *Trichoderma*. Для этого сначала был отобран ряд морфологических и физиологических признаков, обычно используемый при описании данного вида. Список включал в частности, такие параметры как рост колонии и зависимость скорости роста от среды (КГА, SNA) и температуры, форма колонии, запах, цвет, вид воздушного мицелия, размеры и форма спор и ряд других признаков. Была проанализирована последовательность проявления признаков, их возможные состояния, а также возможные взаимосвязи признаков между собой.

Далее было изучено, какие структуры данных и процедурные конструкции существующих языков программирования являются наиболее подходящими для представления используемых признаков. В результате анализа 16 различных языков программирования, представляющих каждую из четырех основных парадигм программирования (императивное программирование, функциональное программирование, логическое программирование, объектно-ориентированное программирование) были выбраны наиболее компактные и эффективные средства представления признаков. Эти средства были включены в разрабатываемый язык BMLIDA. Были выработаны правила, гарантирующие одинаковое представление одинаковых признаков в описаниях различных видов.

Для удобства работы была создана графическая программная оболочка и система автоматического поиска ошибок в описаниях.

После этого было произведено собственно написание программ для видов *Trichoderma*. Полученные описания были использованы для вычисления *NID* и построения матрицы расстояний *NID* между видами с помощью специально написанной программы. На основании матрицы *NID* с помощью алгоритма Neighbor-Joining была построена дендрограмма, иллюстрирующая результаты кластеризации (Рисунок 1).

Для сравнения на рисунке 4 приведен один из существующих вариантов «стандартной» биологической классификации *Trichoderma*, который содержит 19 абстракций на 10 экземпляров, из них 10 видов и 1 род и 8 промежуточных таксонов. Различные по смыслу таксоны имеют одинаковые названия (Trichoderma – одновременно род и секция). При этом рисунок не содержит никакой информации о смысле различий между таксонами.

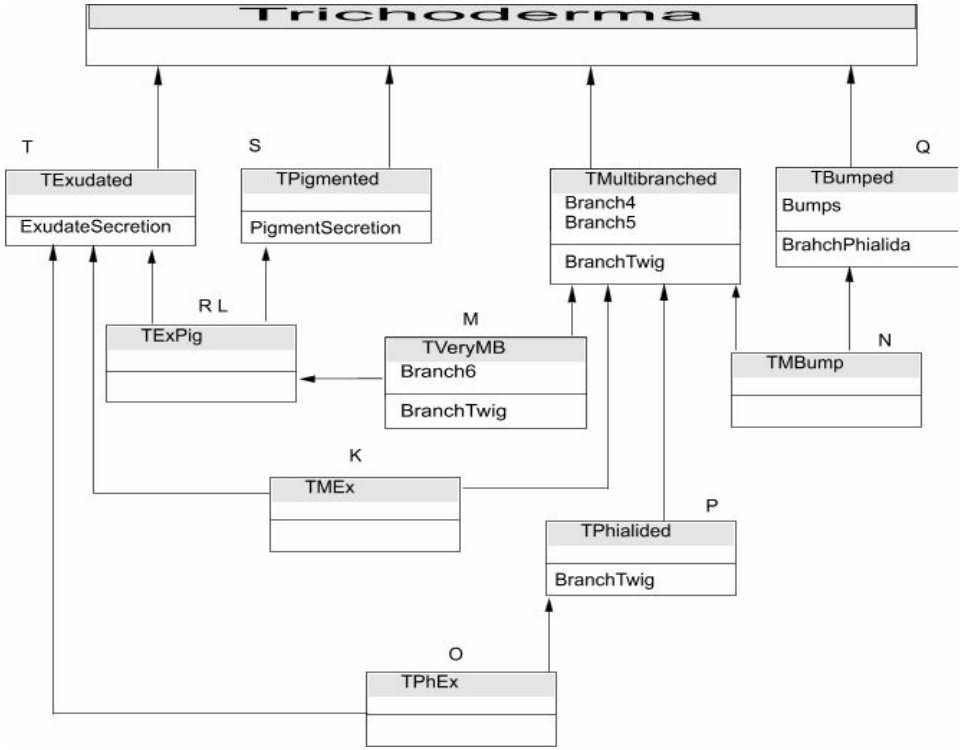

Рис 3. ОО-версия фрагмента классификации Trichoderma. Рядом с каждым классом обозначены буквами (K-T) соответствующие экземпляры (изоляты) Trichoderma

Таблица 1. Соответствие понятий биологической классификации и ООП

Биологическая классификация	ООП
Таксон	Класс
Вид	Неабстрактный класс, который может иметь экземпляры.
Таксон высших порядков	Абстрактный класс
Организм	Экземпляр класса

В тоже время ОО-подход предоставляет классификационные средства, для которых в биологических систематиках нет аналогов (например, множественное наследование). Такие классификационные средства позволяют в явном виде указать известные закономерности признаков, либо использовать автоматический алгоритм поиска закономерностей. При этом в конце можно будет получить отчет о том, какие именно закономерности использовались при вычислении *NID*.

Для представления описаний и классификации биологических объектов был разработан специальный объектно-ориентированный язык ConceptSystem, который был применен для описания и классификации микроскопических грибов рода Trichoderma.

Применение метода явного задания закономерностей для классификации грибов рода Trichoderma

Использовался способ описания, базирующийся на рассмотрении развития организмов рода *Trichoderma* как последовательности деления и дальнейшей специализации отдельных клеток

Описание представляет собой программу, которую необходимо задать исходной клетке (споре), чтобы получить ее развитие сначала в колонию *Trichoderma,* а в итоге снова в спору. Такое описание является естественным, поскольку отражает реальное положение вещей и позволяет разделить признаки на *свойственные единичным клеткам* и *появляющиеся в результате взаимодействия большого числа клеток*. Описание-программа следует логике развития самого организма, а не логике проведения наблюдений, особенностей экспериментов и исторических обстоятельств, нередко оказывающих влияние на обычные текстовые описания.

С использованием предложенных методов были построены программы описания и классификация для 10 изолятов грибов рода *Trichoderma*. На Рис.3. приведен пример фрагмента систематики, полученной с использованием описаний-программ и принципа множественного наследования. Буквами K,L,M,N,O,P,Q,R,S,T обозначены конкретные изоляты *Trichoderma.*

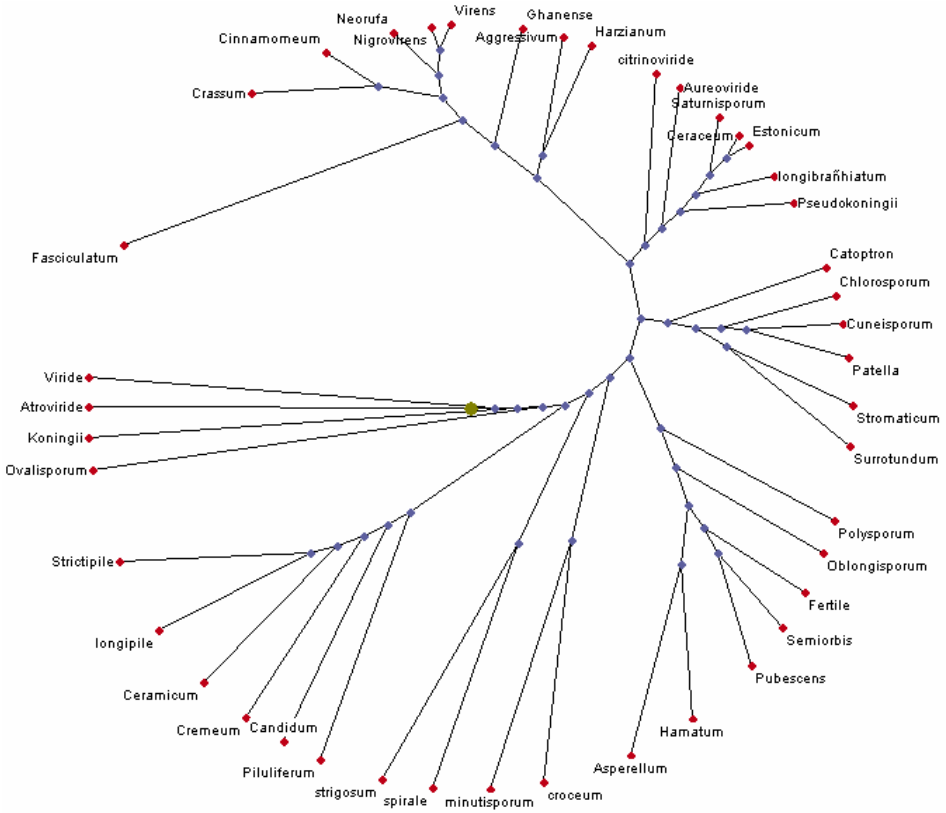

Рис. 1. Кластерный анализ 43 видов рода *Trichoderma* с использованием меры расстояния *NID*.

Выяснилось, что выбор алгоритма кластеризации не влияет на полученный результат. Так топологии деревьев полученных по матрице *NID* с помощью алгоритмов UPGMA и метода минимальной эволюции не отличались существенно от дерева, полученного с помощью Neighbor-Joining. Также результат не чувствителен к порядку видов в матрице.

Полученное дерево (Рис. 1) согласуется в ряде пунктов с известными данными о взаимосвязи видов внутри рода *Trichoderma.* В частности выделяются отдельно виды секции Trichoderma, подсекции Rufa – *T.viride, T. atroviride, T. koningi u T. ovalisporum.* Вместе оказываются виды секции Pachibasium “A” – *T. Hamatum u T. Asperellum.* Также близко расположены виды секции Longibrachiatum – *T. longibraciatum, T.citrinoviride, T. pseudokoningi.* В то же время имеется и ряд отличий, не согласующихся с принятой моделью классификации рода (следует

отметить, что во многих случаях систематическое положение видов Trichoderma является спорным вопросом, по которому отсутствует единое мнение).

Разработанный метод сочетает в себе преимущества нумерического и традиционного (интуитивного) подходов. С одной стороны, используется точный количественный метод, при этом исключаются проблемы связанные с произвольным выбором меры расстояния и сводится к минимуму эффект от выбора алгоритма кластеризации. С другой стороны, метод позволяет включать знания и личный опыт систематика посредством выбора средств кодирования признака в программе-описании. При этом, в отличии от матрицы признаков, используемой в других нумерических методах, программа-описание не поощряет формальный, недостаточно обдуманный подход к процессу выбора и кодирования признаков. Программа-описание сохраняет логику принятых в этом процессе решений, в то время как матрица признаков эффективно скрывает эту информацию. Наконец, разработанный метод не требует использования строго независимых признаков. В описания-программы могут включаться связанные признаки *одновременно* с информацией о способах их взаимодействия и развития процесса во времени. Эта информация адекватно учитывается при построении дерева благодаря использованию *NID*-метрики.

В то же время в процессе проводимой работы были выявлены некоторые недостатки предложенного метода:

- Трудно понять причину получения тех или иных результатов кластеризации. Этот недостаток является общим для всех нумерических методов.
- Автоматический алгоритм компрессии все же может давать неправильные результаты для $K(x|y)$, при этом нет возможности ручной коррекции результатов.

Третья глава. Использование технологии объектно-ориентированного программирования для представления закономерностей в признаках микроорганизмов

Отмеченные во второй главе недостатки послужили стимулом для создания варианта разработанного метода, в котором основная работа по оценки $K(x|y)$ возложена на человека-систематика, а автоматические алгоритмы только оказывают помощь, при этом любое их решение может быть прокомментировано компьютером. Данный вариант является более трудоемким, зато дает больше возможностей для оптимизации таксономии.

Основной задачей было создать способ, который позволил бы в явном виде описать закономерности и сходства программ-моделей. Для этого были использованы техники объектно-ориентированного программирования.

Если мы сопоставим объектно-ориентированное программирование (ООП) и биологическую систематику, то можем идентифицировать некоторые соответствия. Любая существующая биологическая система может быть представлена средствами ООП. Чтобы понять эти соответствия, обратимся к рисунку 2, на котором представлен фрагмент биологической классификации в форме диаграммы классов. Любому таксону может быть сопоставлен класс, определяющий общие характеристики данного таксона (табл 1).

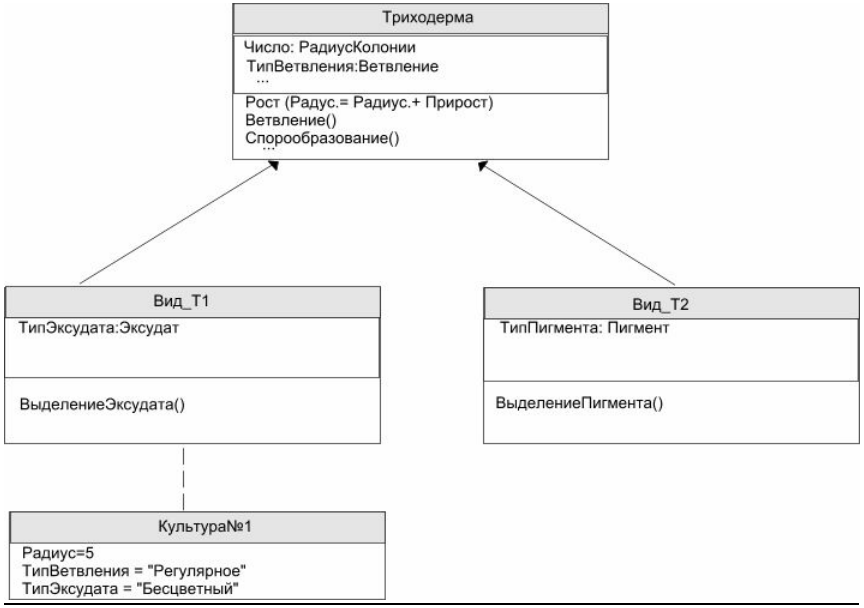

Рис 2. Наследование классов и биологическая систематика.

«Триходерма» является абстрактным классом (не может иметь экземпляров), в котором определены атрибуты и методы, свойственные всем организмам рода Trichoderma. Классы Вид_Т1 и Вид_Т2 являются подклассами «Триходерма», определяющими дополнительные атрибуты. Классы Вид_Т1 и Вид_Т2 являются *видами* с точки зрения биологической систематики, поскольку из них могут образовываться конкретные экземпляры (Культура№1)