(19) Евразийское патентное ведомство

(11) 046521 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2024.03.22

(21) Номер заявки
202390243

(22) Дата подачи заявки
2018.04.18

(51) Int. Cl. ***C40B 40/10*** (2006.01)
G16B 15/30 (2019.01)
G16B 20/30 (2019.01)
G16B 50/00 (2019.01)
G16C 20/50 (2019.01)
G16C 20/64 (2019.01)
G16C 20/70 (2019.01)

(54) СПОСОБЫ ИДЕНТИФИКАЦИИ СОЕДИНЕНИЙ

(31) **62/486,692**
(32) **2017.04.18**
(33) **US**
(43) **2023.05.31**
(62) **201992476; 2018.04.18**

(71)(73) Заявитель и патентовладелец:
ИКС-ЧЕМ, ИНК. (US)

(72) Изобретатель:
Сайджел Эрик Алан, Сюэ Лин, Малхерн Кристофер Джеймс, Моччиа Деннис Джозеф (US)

(74) Представитель:
Гизатуллина Е.М., Угрюмов В.М., Строкова О.В., Костюшенкова М.Ю., Гизатуллин Ш.Ф., Джермакян Р.В. (RU)

(56) ZHENG Xiliang et al. Pocket-Based Drug Design: Exploring Pocket Space. The AAPS Journal, vol. 15, N 1, January 2013, p. 228-241

HOUK K.N. et al. Grails for Computational Organic Chemistry and Biochevistry. Acc. Chem. Res., 2017, 50, p. 539-543

CN-A-102224255

(57) Изобретение относится к способам виртуального скрининга, в которых используют массивы данных, полученных с применением нуклеотид-кодируемых библиотек (например, ДНК-кодируемых библиотек). Такие способы позволяют с высокой степенью достоверности предсказывать связывающие взаимодействия между кандидатными соединениями и представляющими интерес белками, пригодными для разработки терапевтических препаратов.

046521 B1

046521 B1

046521 B1

Уровень техники

С помощью способов виртуального скрининга можно расширить доступные возможности скрининга для данной мишени, и можно увеличить вероятность успешной оптимизации. Виртуальный скрининг может представлять собой быстрый и недорогой способ для идентификации множества скаффолдов, которые будут применять в качестве отправных точек для оптимизации. Возможности виртуального скрининга обычно ограничены размером экспериментально определенного массива данных, применяемого для сравнения с известными экспериментальными данными с целью получения виртуальных данных. Таким образом, существует потребность в способах, в которых объединены надежные вычислительные методы с очень большими массивами данных для обеспечения достаточной достоверности в отношении вычислительных предсказаний, для замены традиционных способов высокопроизводительного скрининга.

Сущность изобретения

Настоящее раскрытие относится к способам идентификации соединений, пригодных в качестве терапевтических средств и/или пригодных в качестве исходных точек для оптимизации разработки терапевтических средств. В таких способах объединены вычислительные способы, пригодные для предсказания связывания соединений и белков, с крупными массивами экспериментальных данных, полученных с применением нуклеотид-кодируемых библиотек (например, ДНК-кодируемых библиотек). Объединение данных, полученных с использованием нуклеотид-кодируемых библиотек и вычислительных способов, позволяет с высокой степенью достоверности предсказывать связывающие взаимодействия между кандидатными соединениями и представляющими интерес белками.

Соответственно, согласно одному аспекту настоящее раскрытие относится к способу, предусматривающему стадии: (a) обеспечения множества установленных фактов о связывающих взаимодействиях (например, по меньшей мере 250000 установленных фактов) для белка-мишени в физическом вычислительном устройстве, имеющем определенное представление набора кандидатных соединений (например, низкомолекулярных соединений), причем по меньшей мере 50% (например, по меньшей мере 60%, по меньшей мере 70%, по меньшей мере 80%, по меньшей мере 90%, по меньшей мере 95%, по меньшей мере 99%) установленных фактов о связывающих взаимодействиях в пределах указанного множества представляют связывающее взаимодействие между белком-мишенью и соединением, содержащим нуклеотидную метку, кодирующую идентичность соединения (например, компонент ДНК-кодируемой библиотеки); (b) применения вычислительного устройства для получения предполагаемых связывающих взаимодействий для кандидатных соединений с применением множества установленных фактов о связывающих взаимодействиях; и (c) вывода перечня кандидатных соединений с возможностью отображения и ранжирования по наиболее предполагаемым связывающим взаимодействиям.

Согласно некоторым вариантам осуществления множество установленных фактов о связывающих взаимодействиях включает по меньшей мере 250000 (например, по меньшей мере 500000, по меньшей мере один миллион, по меньшей мере два миллиона, по меньшей мере пять миллионов, по меньшей мере десять миллионов, по меньшей мере двадцать пять миллионов) установленных фактов о связывающих взаимодействиях.

Согласно некоторым вариантам осуществления по меньшей мере 50% множества установленных фактов о связывающих взаимодействиях были определены путем приведения множества (например, по меньшей мере 250000, по меньшей мере 500000, по меньшей мере одного миллиона, по меньшей мере двух миллионов, по меньшей мере пяти миллионов, по меньшей мере десяти миллионов) соединений, содержащих нуклеотидную метку, кодирующую идентичность соединения, в контакт с белком-мишенью одновременно (например, в одном и том же реакционном сосуде в одно и то же время). Например, согласно некоторым вариантам осуществления по меньшей мере 50% установленных фактов о связывающих взаимодействиях для компонентов ДНК-кодируемой библиотеки, используемой для получения предполагаемых связывающих взаимодействий, определяли в одном эксперименте в одном реакционном сосуде.

Согласно некоторым вариантам осуществления способ дополнительно предусматривает обеспечение одного или более дополнительных множеств установленных фактов о связывающих взаимодействиях для одного или более дополнительных белков-мишеней, причем по меньшей мере 50% установленных фактов о связывающих взаимодействиях в пределах одного или более дополнительных множеств представляют связывающее взаимодействие между дополнительным белком-мишенью и соединением из множества установленных фактов о связывающих взаимодействиях с белком-мишенью из стадии (a). Согласно некоторым вариантам осуществления способ дополнительно предусматривает обеспечение одного или более дополнительных множеств установленных фактов о связывающих взаимодействиях для одного или более экспериментов с использованием отрицательного контроля, причем по меньшей мере 50% установленных фактов о связывающих взаимодействиях в пределах множества представляют отрицательный контроль для соединения из множества установленных фактов о связывающих взаимодействиях с белком-мишенью из стадии (a). Согласно некоторым вариантам осуществления способ дополнительно предусматривает обеспечение одного или более дополнительных множеств установленных фактов о связывающих взаимодействиях для одного или более контрольных экспериментов, причем

множество установленных фактов о связывающих взаимодействиях включает установленные факты о связывающих взаимодействиях для соединения, характеризующегося известными связывающими взаимодействиями с белком-мишенью из стадии (a) (например, с известными ингибиторами или естественными лигандами). Согласно некоторым вариантам осуществления способ предусматривает получение показателя избирательности путем сравнения связывания или предполагаемого связывания соединения или кандидатного соединения с белком-мишенью со связыванием или предполагаемым связыванием соединения или кандидатного соединения с одним или более дополнительными белками-мишенями и/или отрицательным контролем. Согласно некоторым вариантам осуществления перечень кандидатных соединений может быть отображен и ранжирован по показателю избирательности. Согласно некоторым вариантам осуществления один или более дополнительных белков-мишеней включают мутант белка-мишени.

Согласно некоторым вариантам осуществления предполагаемые связывающие взаимодействия получают с применением сравнений химической структуры, например, с использованием представлений молекул. Представления молекул включают без ограничения топологические представления на основе атомов, элементов топологии или функциональных групп и их связности (например, фингерпринты, таблицы связности, молекулярная связность и/или представления в виде молекулярных графов), электростатические представления (например, электронные свойства поверхностей), геометрические представления (например, фармакофоры, фармакофорные фингерпринты, фингерпринты на основе формы и/или молекулярные координаты в трехмерном пространстве на основе атомов, элементов топологии или функциональных групп) или квантово-химические представления. Согласно некоторым вариантам осуществления предполагаемые связывающие взаимодействия получают с применением топологических представлений на основе атомов, элементов топологии или функциональных групп и их связности (например, фингерпринтов, таблиц связности, молекулярной связности и/или представлений в виде молекулярных графов). Согласно некоторым вариантам осуществления предполагаемые связывающие взаимодействия получают с применением электростатических представлений (например, электронных свойств поверхностей).

Согласно некоторым вариантам осуществления предполагаемые связывающие взаимодействия получают с применением геометрических представлений (например, фармакофоров, фармакофорных фингерпринтов, фингерпринтов на основе формы и/или молекулярных координат в трехмерном пространстве на основе атомов, элементов топологии или функциональных групп). Согласно некоторым вариантам осуществления предполагаемые связывающие взаимодействия получают с применением квантово-химических представлений. Согласно некоторым вариантам осуществления предполагаемые связывающие взаимодействия получают с применением химических фингерпринтов.

Химические фингерпринты можно применять для объединения информации о структуре соединений и данных по связывающему взаимодействию для идентификации структурных паттернов, определяющих связывание с белком-мишенью. Соответственно, согласно некоторым вариантам осуществления способ дополнительно предусматривает: (i) обеспечение множества химических фингерпринтов множества соединений (например, химических фингерпринтов, таких как ECFP6, FCFP6, ECFP4, MACCS или фингерпринты Моргана/кольцевые фингерпринты с изменяющимся числом битов (например, 166, 512, 1024)); и (ii) использование множества химических фингерпринтов в получении предполагаемых связывающих взаимодействий. Согласно некоторым вариантам осуществления, например, в обучающих наборах данных, множество химических фингерпринтов включает химические фингерпринты одного или более соединений, содержащих нуклеотидную метку, кодирующую идентичность соединения, например, химический фингерпринт является представлением структуры соединения без нуклеотидной метки. Согласно некоторым вариантам осуществления, например, в наборах данных для предсказания, множество химических фингерпринтов включает химические фингерпринты одного или более кандидатных соединений. Согласно некоторым вариантам осуществления химические фингерпринты представляют собой фингерпринты ECFP6.

Согласно некоторым вариантам осуществления способ дополнительно предусматривает обеспечение одного или более установленных фактов о свойствах (например, молекулярная масса и/или clogP) для набора кандидатных соединений. Согласно некоторым вариантам осуществления один или более установленных фактов о свойствах используют для получения предполагаемых связывающих взаимодействий. Согласно некоторым вариантам осуществления перечень кандидатных соединений может быть отображен и ранжирован по одному или более установленным фактам о свойствах.

Согласно некоторым вариантам осуществления способ дополнительно предусматривает передачу перечня кандидатных соединений через Интернет или на устройство отображения. Согласно некоторым вариантам осуществления работа с физическим вычислительным устройством и доступ к нему осуществляется через Интернет.

Согласно некоторым вариантам осуществления способ дополнительно предусматривает получение показателя правдоподобия для каждого из предполагаемых связывающих взаимодействий кандидатных соединений, причем показатель правдоподобия получают с применением сравнений химической структуры (например, метод главных компонент) между кандидатным соединением и одним или более соеди-

нениями из множества связывающих взаимодействий для белка-мишени из стадии (a). Например, согласно некоторым вариантам осуществления показатель правдоподобия получают путем сравнения кандидатного соединения с химическим пространством, определяемым соединениями из множества связывающих взаимодействий из стадии (a), путем определения удаленности, такой как евклидово расстояние в измерениях, определяемых методом главных компонент, кандидатного соединения до химического пространства. Согласно некоторым вариантам осуществления перечень кандидатных соединений может быть отображен и ранжирован по показателю правдоподобия для предполагаемого связывающего взаимодействия для кандидатного соединения.

Согласно некоторым вариантам осуществления способ дополнительно предусматривает (d) синтезирование одного или более кандидатных соединений из перечня кандидатных соединений.

Согласно некоторым вариантам осуществления способ дополнительно предусматривает (e) приведение одного или более синтезированных кандидатных соединений в контакт с белком-мишенью для определения одного или более экспериментальных связывающих взаимодействий.

Согласно одному аспекту настоящее раскрытие относится к машиночитаемому носителю с хранящимися на нем выполняемыми командами для управления физическим вычислительным устройством с целью выполнения способа, предусматривающего стадии:

(a) обеспечения множества установленных фактов о связывающих взаимодействиях для белка-мишени в физическом вычислительном устройстве, имеющем определенное представление набора кандидатных соединений,

причем по меньшей мере 90% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением, содержащим нуклеотидную метку, кодирующую идентичность соединения;

(b) применения вычислительного устройства для получения предполагаемых связывающих взаимодействий кандидатных соединений с применением множества установленных фактов о связывающих взаимодействиях; и

(c) вывода перечня кандидатных соединений с возможностью отображения и ранжирования по наиболее предполагаемым связывающим взаимодействиям.

Согласно одному аспекту настоящее раскрытие относится к устройству, имеющему определенное представление набора кандидатных соединений, и запрограммированному посредством выполняемых команд для управления устройством с целью выполнения способа, предусматривающего стадии:

(a) обеспечения множества установленных фактов о связывающих взаимодействиях для белка-мишени в физическом вычислительном устройстве, имеющем определенное представление набора кандидатных соединений,

причем по меньшей мере 90% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением, содержащим нуклеотидную метку, кодирующую идентичность соединения;

(b) применения вычислительного устройства для получения предполагаемых связывающих взаимодействий кандидатных соединений с применением множества установленных фактов о связывающих взаимодействиях; и

(c) вывода перечня кандидатных соединений с возможностью отображения и ранжирования по наиболее предполагаемым связывающим взаимодействиям.

Определения.

"Показатель правдоподобия" в контексте настоящего документа относится к вычислению, которое указывает на достоверность предполагаемого связывающего взаимодействия для кандидатного соединения на основе структурного подобия между кандидатным соединением и одним или более соединениями в массиве данных, используемом для получения предположения.

Термин "связывающее взаимодействие" в контексте настоящего документа относится к связи (например, нековалентной или ковалентной) двух или более объектов или между двумя или более объектами. "Прямое" связывание включает физический контакт между объектами или их фрагментами; непрямое связывание включает физическое взаимодействие посредством физического контакта с одним или более промежуточными объектами. Связывание двух или более объектов, как правило, можно оценивать в любом из множества контекстов, в том числе в тех случаях, когда взаимодействующие объекты или их фрагменты изучают изолированно или в контексте более сложных систем (например, при ковалентном связывании или другом типе связывания с объектом-носителем и/или в биологической системе или в клетке).

Аффинность молекулы X к ее партнеру по связыванию Y обычно может быть представлено константой диссоциации (K_D). Аффинность можно измерять с помощью обычных способов, известных в данной области, включая таковые, описанные в настоящем документе. Подразумевается, что термин "K_D" в контексте настоящего документа относится к равновесной константе диссоциации для конкретного взаимодействия соединение-белок или комплекс-белок. Как правило, соединения по настоящему изобретению связываются с презентирующими белками с равновесной константой диссоциации (K_D) менее чем приблизительно 10^{-6} М, как, например, менее чем примерно 10^{-7} М, 10^{-8} М, 10^{-9} М или 10^{-10} М или

даже меньше, например, при определении с помощью технологии поверхностного плазмонного резонанса (SPR) с применением презентирующего белка в качестве анализируемого вещества, и соединения в качестве лиганда. Согласно некоторым вариантам осуществления соединения по настоящему изобретению связываются с белками-мишенями (например, белком-мишенью эукариот, таким как белок-мишень млекопитающих или белок-мишень грибов, или белком-мишенью прокариот, таким как белок-мишень бактерий) с равновесной константой диссоциации (K_D) менее чем приблизительно 10^{-6} М, как, например, менее чем примерно 10^{-7} М, 10^{-8} М, 10^{-9} М или 10^{-10} М или даже меньше, например, при определении с помощью технологии поверхностного плазмонного резонанса (SPR) с применением белка-мишени в качестве анализируемого вещества, и соединения в качестве лиганда.

"Установленный факт о связывающем взаимодействии" в контексте настоящего документа относится к связывающему взаимодействию между соединением и белком (например, белком-мишенью), или его отсутствию, которое было определено экспериментально, например, с помощью SPR. Например, согласно некоторым вариантам осуществления установленный факт о связывающем взаимодействии относится к определению того, что соединение не взаимодействует с белком (например, белком-мишенью).

Термин "представления молекул" относится, например, к топологическим представлениям, электростатическим представлениям, геометрическим представлениям или квантово-химическим представлениям соединений. Представления молекул включают, например, химические фингерпринты.

Термин "электростатические представления" относится к типу представлений молекул, включающему такую информацию, как электронные свойства поверхностей.

"Предполагаемое связывающее взаимодействие" в контексте настоящего документа относится к связывающему взаимодействию, которое было предсказано с применением вычислительного анализа. Согласно некоторым вариантам осуществления предполагаемое связывающее взаимодействие кандидатного соединения с белком-мишенью получают путем сравнения химической структуры кандидатного соединения с химической структурой одного или более соединений, для которых связывающее взаимодействие с белком-мишенью было определено экспериментально.

В контексте настоящего документа термин "химический фингерпринт" относится к машиночитаемым представлениям молекул для соединений, таким как битовая строка, т.е. перечень двоичных значений (0 или 1), которые характеризуют двух- и/или трехмерную структуру молекулы. Иллюстративные способы получения химических фингерпринтов известны в данной области, включая без ограничения MACCS, фингерпринты с расширенной связностью (ECFP), фингерпринты функциональных классов (FCFP), фингерпринты Моргана/кольцевые фингерпринты и фингерпринты на основе химических хешей.

В контексте настоящего документа термин "clogP" относится к расчетному коэффициенту разделения молекулы или участка молекулы. Коэффициент разделения представляет собой соотношение концентраций соединения в смеси двух несмешиваемых фаз в равновесном состоянии (например, октанол и вода), и с помощью него можно определять гидрофобность или гидрофильность соединения. В данной области доступно множество способов определения clogP, например, согласно некоторым вариантам осуществления clogP можно определять с применением алгоритмов для определения количественной связи структура-свойство, известных в данной области (например, с применением способов предсказания на основе использования фрагментов, с помощью которых можно предсказывать logP соединения путем определения суммы его неперекрывающихся фрагментов молекулы). Некоторые алгоритмы для расчета clogP известны в данной области, включая таковые, применяемые в пакете программного обеспечения для молекулярного редактирования, такого как CHEMDRAW® Pro, версии 12.0.2.1092 (Cambridgesoft, Кембридж, Массачусетс) и MARVINSKETCH® (ChemAxon, Будапешт, Венгрия).

Термин "сопоставимый" в контексте настоящего документа относится к двум или более соединениям, объектам, ситуациям, совокупностям условий и т.д., которые могут не быть идентичными относительно друг друга, но которые являются достаточно схожими для обеспечения возможности сравнения их между собой, так что на основе наблюдаемых различий или сходств могут быть сделаны обоснованные выводы. Согласно некоторым вариантам осуществления сопоставимые совокупности условии, случаев, индивидуумов или популяций характеризуются множеством практически идентичных признаков и одним или небольшим числом различных признаков. Специалистам в данной области будет понятно, в контексте, какая степень идентичности требуется в любом данном случае для того, чтобы два или более таких соединений, объектов, ситуаций, совокупностей условий и т.д. считались сопоставимыми. Например, специалистам в данной области будет понятно, что совокупности случаев, индивидуумов или популяций сопоставимы друг с другом, когда они характеризуются достаточным числом и типом практически идентичных признаков для гарантии получения обоснованного вывода о том, что различия в полученных результатах или наблюдаемых явлениях при различных совокупностях случаев или с различными индивидуумами или популяциями свидетельствуют о разностях в таких отличительных признаках, или обусловлены ими.

Многие описанные в настоящем документе методологии включают стадию "определения". Специалисты в данной области, читая настоящее описание, поймут, что в случае такого "определения" может быть использована, или оно может быть выполнено посредством применения любой из ряда методик,

доступных для специалистов в данной области, включая, например, конкретные методики, явно упомянутые в настоящем документе. Согласно некоторым вариантам осуществления определение включает осуществление манипуляций с физическим образцом. Согласно некоторым вариантам осуществления определение включает рассмотрение и/или осуществление манипуляций с данными или информацией, например, с использованием компьютера или другого устройства обработки данных, выполненного с возможностью осуществления соответствующего анализа. Согласно некоторым вариантам осуществления определение включает получение соответствующей информации и/или материалов из определенного источника. Согласно некоторым вариантам осуществления определение включает сравнение одного или более признаков образца или объекта с сопоставимым эталоном.

Термин "геометрические представления" относится к типу представления молекулы. Геометрические представления могут включать в себя информацию, касающуюся, например, фармакофоров, фармакофорных фингерпринтов, фингерпринтов на основе формы и/или молекулярных координат в трехмерном пространстве на основе атомов, элементов топологии или функциональных групп.

В контексте настоящего документа термин "библиотека" относится к группе из 2, 5, 10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9 или более разных молекул. Согласно некоторым вариантам осуществления по меньшей мере 10% (например, по меньшей мере 20%, по меньшей мере 30%, по меньшей мере 40%, по меньшей мере 50%, по меньшей мере 60%, по меньшей мере 70%, по меньшей мере 80%, по меньшей мере 90%, по меньшей мере 95%, по меньшей мере 99% или 100%) соединений в библиотеке являются соединениями, включающими нуклеотидную метку, кодирующую их идентичность, такими как ДНК-кодируемые соединения.

В контексте настоящего документа термин "отрицательный контроль" относится к эксперименту для определения связывающего взаимодействия в отсутствие белка-мишени.

Термин "площадь полярной поверхности" относится к суммарной поверхности по всем обеспечивающим полярность атомам молекулы или участка молекулы, включая их присоединенные атомы водорода. Площадь полярной поверхности определяется вычислительным путем с применением такой программы, как CHEMDRAW® Pro, версии 12.0.2.1092 (Cambridgesoft, Кембридж, Массачусетс).

В контексте настоящего документа термин "положительный контроль" относится к эксперименту для определения связывающего взаимодействия, причем аффинность связывания соединения, приводимого в контакт с белком-мишенью, является известной.

"Установленный факт о свойстве" в контексте настоящего документа относится к рассчитанному или экспериментально определенному свойству (например, clogP, площадь полярной поверхности, молекулярная масса) конкретного соединения.

Термин "избирательный", при применении по отношению к соединению, обладающему определенной активностью, специалистам в данной области следует понимать как означающий, что соединение распознает потенциальные объекты-мишени или состояния. Например, согласно некоторым вариантам осуществления считается, что соединение "избирательно" связывается со своей мишенью, если оно связывается предпочтительно с такой мишенью в присутствии одной или более конкурирующих альтернативных мишеней. Согласно множеству вариантов осуществления избирательное взаимодействие зависит от наличия конкретного структурного признака объекта-мишени (например, эпитопа, кармана, сайта связывания). Следует понимать, что избирательность не обязательно должна быть абсолютной. Согласно некоторым вариантам осуществления избирательность может быть оценена относительно таковой для связывающего средства в отношении одного или более других потенциальных объектов-мишеней (например, конкурирующие вещества). Согласно некоторым вариантам осуществления избирательность оценивают относительно таковой эталонного избирательного связывающего средства. Согласно некоторым вариантам осуществления избирательность оценивают относительно таковой эталонного неизбирательного связывающего средства. Согласно некоторым вариантам осуществления средство или объект выявляемо не связываются с конкурирующей альтернативной мишенью в условиях связывания с его объектом-мишенью. Согласно некоторым вариантам осуществления связывающее средство связывается с более высокой скоростью прямой реакции, более низкой скорость обратной реакции, повышенной аффинностью, пониженной степенью диссоциации и/или повышенной стабильностью в отношении его объекта-мишени по сравнению с конкурирующей(ими) альтернативной(ыми) мишенью(ями).

"Показатель избирательности" в контексте настоящего документа относится к расчету специфичности соединения в отношении белка-мишени. Согласно некоторым вариантам осуществления показатель избирательности может быть рассчитан путем сравнения связывания соединения с белком-мишенью и связывания соединения с другим белком (например, мутантом белка-мишени или неродственным белком). Согласно другим вариантам осуществления показатель избирательности может быть рассчитан путем сравнения связывания соединения с белком-мишенью и связывания с отрицательным контролем.

Термин "малая молекула" означает органическое и/или неорганическое соединение с низкой молекулярной массой. В целом, "малая молекула" представляет собой молекулу, размер которой составляет менее чем приблизительно 5 килодальтон (кДа). Согласно некоторым вариантам осуществления размер малой молекулы составляет менее чем приблизительно 4 кДа, 3 кДа, приблизительно 2 кДа или приблизительно 1 кДа. Согласно некоторым вариантам осуществления размер малой молекулы составляет ме-

нее чем приблизительно 800 дальтон (Да), приблизительно 600 Да, приблизительно 500 Да, приблизительно 400 Да, приблизительно 300 Да, приблизительно 200 Да или приблизительно 100 Да. Согласно некоторым вариантам осуществления вес малой молекулы составляет менее чем приблизительно 2000 г/моль, менее чем приблизительно 1500 г/моль, менее чем приблизительно 1000 г/моль, менее чем приблизительно 800 г/моль или менее чем приблизительно 500 г/моль. Согласно некоторым вариантам осуществления малая молекула не является полимером. Согласно некоторым вариантам осуществления малая молекула не включает полимерный фрагмент. Согласно некоторым вариантам осуществления малая молекула не является белком или полипептидом (например, не является олигопептидом или пептидом). Согласно некоторым вариантам осуществления малая молекула не является полинуклеотидом (например, не является олигонуклеотидом). Согласно некоторым вариантам осуществления малая молекула не является полисахаридом. Согласно некоторым вариантам осуществления малая молекула не содержит полисахарид (например, не является гликопротеином, протеогликаном, гликолипидом и т.д.). Согласно некоторым вариантам осуществления малая молекула не является липидом. Согласно некоторым вариантам осуществления малая молекула представляет собой модулирующее соединение. Согласно некоторым вариантам осуществления малая молекула является биологически активной. Согласно некоторым вариантам осуществления малая молекула является выявляемой (например, содержит по меньшей мере один выявляемый фрагмент). Согласно некоторым вариантам осуществления малая молекула является терапевтической.

Специалисты в данной области, читая настоящее раскрытие, поймут, что описанные в настоящем документе определенные низкомолекулярные соединения могут быть обеспечены и/или использованы в любой из множества форм, таких как, например, солевые формы, защищенные формы, формы в виде пролекарств, сложноэфирные формы, изомерные формы (например, оптические и/или структурные изомеры), изотопные формы и т.д. Согласно некоторым вариантам осуществления упоминание конкретного соединения может относится к конкретной форме такого соединения. Согласно некоторым вариантам осуществления упоминание конкретного соединения может относится к такому соединению в любой форме. Согласно некоторым вариантам осуществления, если соединение является соединением, которое существует или встречается в природе, такое соединение может быть обеспечено и/или использовано, в соответствии с настоящим изобретением, в форме, отличающейся от таковой, в которой соединение существует или встречается в природе. Специалистам в данной области будет понятно, что препарат соединения, предусматривающий другой уровень, количество или соотношение одной или более отдельных форм, по сравнению с эталонным препаратом или источником (например, естественным источником) соединения, может рассматриваться как отличающаяся форма соединения, описанного в настоящем документе. Таким образом, согласно некоторым вариантам осуществления, например, препарат одного стереоизомера соединения может рассматриваться как отличающаяся форма соединения по сравнению с рацемической смесью соединения; конкретная соль соединения может рассматриваться как отличающаяся форма по сравнению с другой солевой формой соединения; препарат, содержащий один конформационный изомер ((Z) или (E)) положения двойной связи, может рассматриваться как отличающаяся форма по сравнению с таковым, содержащим другой конформационный изомер ((E) или (Z)) положения двойной связи; препарат, в котором один или более атомов являются изотопами, отличающимися от изотопов, которые присутствуют в эталонном препарате, может рассматриваться как отличающаяся форма; и т.д.

В контексте настоящего документа термины "специфическое связывание", или "специфический для", или "специфический в отношении" относятся к взаимодействию между связывающим средством и объектом-мишенью. Как будет понятно специалистам в данной области, взаимодействие считается "специфическим", если оно является предпочтительным в присутствии альтернативных взаимодействий, например, имеет место связывание с K_D менее 10 мкМ (например, менее 5 мкМ, менее 1 мкМ, менее 500 нМ, менее 200 нМ, менее 100 нМ, менее 75 нМ, менее 50 нМ, менее 25 нМ, менее 10 нМ или от 10 нМ до 100 нМ, от 50 нМ до 250 нМ, от 100 нМ до 500 нМ, от 250 нМ до 1 мкМ, от 500 нМ до 2 мкМ, от 1 мкМ до 5 мкМ). Согласно множеству вариантов осуществления специфическое взаимодействие зависит от наличия конкретного структурного признака объекта-мишени (например, эпитопа, кармана, сайта связывания). Следует понимать, что специфичность не обязательно должна быть абсолютной. Согласно некоторым вариантам осуществления специфичность может быть оценена относительно таковой для связывающего средства в отношении одного или более других потенциальных объектов-мишеней (например, конкурирующие вещества). Согласно некоторым вариантам осуществления специфичность оценивают относительно таковой эталонного специфического связывающего средства. Согласно некоторым вариантам осуществления специфичность оценивают относительно таковой эталонного неспецифического связывающего средства.

Термин "структурное подобие" относится к подобию в отношении расположения в двух- или трехмерном пространстве и/или ориентации атомов или фрагментов относительно друг друга (например, расстояние между ними и/или углы, образуемые ними, у представляющего интерес средства и эталонного средства) в одном или более разных соединениях.

Термин "практически" относится к качественному состоянию, характеризующемуся полной или по-

чти полной мерой или степенью проявления представляющих интерес характеристики или свойства. Специалисту в области биологических наук будет понятно, что биологические и химические явления редко, если вообще когда-либо, доходят до окончания и/или протекают до завершения, или достигают или избегают абсолютного результата. Термин "практически", следовательно, применяют в настоящем документе для охвата потенциального отсутствия завершенности, присущего многим биологическим и химическим явлениям.

Термин "практически не связывается" с конкретным белком в контексте настоящего документа может относиться, например, к молекуле или участку молекулы с K_D для мишени 10^{-4} М или больше, в качестве альтернативы, 10^{-5} М или больше, в качестве альтернативы, 10^{-6} М или больше, в качестве альтернативы, 10^{-7} М или больше, в качестве альтернативы, 10^{-8} М или больше, в качестве альтернативы, 10^{-9} М или больше, в качестве альтернативы, 10^{-10} М или больше, в качестве альтернативы, 10^{-11} М или больше, в качестве альтернативы, 10^{-12} М или больше или K_D в диапазоне от 10^{-4} М до 10^{-12} М, или от 10^{-6} М до 10^{-10} М, или от 10^{-7} М до 10^{-9} М.

Термин "белок-мишень" относится к белку, который связывается с малой молекулой. Согласно некоторым вариантам осуществления белок-мишень участвует в биологическом пути, ассоциированном с заболеванием, нарушением или состоянием. Согласно некоторым вариантам осуществления белок-мишень представляет собой встречающийся в природе белок; согласно некоторым таким вариантам осуществления белок-мишень в природе встречается в определенных типах клеток млекопитающих (например, белок-мишень млекопитающих), клетках грибов (например, белок-мишень грибов), клетках бактерий (например, белок-мишень бактерий) или клетках растений (например, белок-мишень растений). Согласно некоторым вариантам осуществления белок-мишень характеризуется естественной способностью к взаимодействию с одним или более комплексами естественный презентирующий белок/естественная малая молекула. Согласно некоторым вариантам осуществления белок-мишень характеризуется естественными взаимодействиями с множеством разных комплексов естественный презентирующий белок/естественная малая молекула; согласно некоторым таким вариантам осуществления в некоторых или во всех из комплексов используется один и тот же презентирующий белок (и разные малые молекулы). Белки-мишени могут быть встречающимися в природе, например, дикого типа. В качестве альтернативы, белок-мишень может отличаться от белка дикого типа, но сохранять биологическую функцию, например, будучи аллельным вариантом, сплайс-мутантом или биологически активным фрагментом. Иллюстративные белки-мишени млекопитающих представляют собой ГТФазы, белок, активирующий ГТФазу, фактор обмена гуаниновых нуклеотидов, белки теплового шока, ионные каналы, белки, имеющие структуру в виде суперспирали, киназы, фосфатазы, убиквитин-лигазы, факторы транскрипции, модификаторы/реконструкторы хроматина, белки с классическими доменами и мотивами, отвечающими за белок-белковые взаимодействия, или любые другие белки, которые участвуют в биологическом пути, ассоциированном с заболеванием, нарушением или состоянием.

Термин "топологические представления" относится к типу представления молекулы, который зависит от топологии молекулы, и который указывает на положение отдельных атомов и связей между ними. Топологические представления могут основываться на атомах, элементах топологии или функциональных группах и их связности (например, фингерпринты, таблицы связности, молекулярная связность и/или представления в виде молекулярных графов). Топологические представления могут быть вычислены на основе графического представления молекул.

Термин "квантово-химические представления" относится к типу представления молекулы. Квантово-химические представления могут включать в себя информацию, касающуюся, например, энергетических показателей или электрических свойств соединения.

Краткое описание чертежей

Фиг. 1 представляет собой график, на котором представлены предсказания относительно связывающих взаимодействий с возрастающим числом библиотек.

Фиг. 2 представляет собой график, на котором представлены множества выполнений предсказаний с течением времени по мере улучшения моделей предсказания.

Подробное описание изобретения

Настоящее раскрытие относится к способам виртуального скрининга для идентификации соединений, пригодных в качестве терапевтических средств и/или пригодных в качестве исходных точек для оптимизации разработки терапевтических средств. В этих способах используют крупные массивы экспериментальных данных, полученных с применением ДНК-кодируемых библиотек для получения предсказаний относительно связывающих взаимодействий между кандидатными соединениями и представляющими интерес белками с высокой степенью достоверности.

Кодируемые соединения.

В настоящем изобретение представлены способы с использованием кодируемых химических объектов, включая химический объект, одну или более меток и головной фрагмент, функционально ассоциированный с первым химическим объектом и одной или более метками. Химические объекты, головные фрагменты, метки, связи и бифункциональные спейсеры дополнительно описаны ниже.

Химические объекты.

Кодируемые соединения (например, малые молекулы), используемые в способах по настоящему изобретению, могут включать один или более структурных элементов, и необязательно включают один или более скаффолдов.

Скаффолд S может представлять собой отдельный атом или скаффолд в виде молекулы. Иллюстративные скаффолды в виде отдельного атома включают атом углерода, атом бора, атом азота или атом фосфора и т.д. Иллюстративные многоатомные скаффолды включают циклоалкильную группу, циклоалкенильную группу, гетероциклоалкильную группу, гетероциклоалкенильную группу, арильную группу или гетероарильную группу. Конкретные варианты осуществления гетероарильного скаффолда включают триазин, такой как 1,3,5-триазин, 1,2,3-триазин или 1,2,4-триазин; пиримидин; пиразин; пиридазин; фуран; пиррол; пирролин; пирролидин; оксазол; пиразол; изоксазол; пиран; пиридин; индол; индазол или пурин.

Скаффолд S может быть функционально связан с меткой с помощью любого пригодного способа. Согласно одному примеру, S представляет собой триазин, который связан непосредственно с головным фрагментом. Для получения такого иллюстративного скаффолда проводят реакцию трихлортриазина (т.е. хлорированного предшественника триазина, имеющего три атома хлора) с нуклеофильной группой головного фрагмента. За счет применения такого способа, S содержит атом хлора в трех положениях, доступных для замещения, где два положения являются доступными узлами разнообразия, а по одному положению присоединен головной фрагмент. Затем структурный элемент A_n добавляют к узлу разнообразия скаффолда, и метку A_n, кодирующую структурный элемент A_n ("метка A_n"), лигируют с головным фрагментом, причем эти две стадии можно осуществлять в любом порядке. Далее структурный элемент B_n добавляют к оставшемуся узлу разнообразия, и метку B_n, кодирующую структурный элемент B_n, лигируют с концом метки A_n. Согласно другому примеру, S представляет собой триазин, который функционально связан с меткой, где трихлортриазин вступает в реакцию с нуклеофильной группой (например, аминогруппой) PEG, алифатического или ароматического линкера метки. Могут быть добавлены структурные элементы и ассоциированные метки, как описано выше.

Согласно еще одному примеру, S представляет собой триазин, который функционально связан со структурным элементом A_n. Для получения такого скаффолда проводят реакцию структурного элемента A_n, имеющего два узла разнообразия (например, электрофильную группу и нуклеофильную группу, как, например, Fmoc-аминокислота), с нуклеофильной группой линкера (например, концевой группой PEG, алифатического или ароматического линкера, который присоединен к головному фрагменту). Далее проводят реакцию трихлортриазина с нуклеофильной группой структурного элемента A_n. За счет применения такого способа, все три положения хлора S применяют в качестве узлов разнообразия для структурных элементов. Как описано в настоящем документе, могут быть добавлены дополнительные структурные элементы и метки, а также дополнительные скаффолды S_n.

Иллюстративные структурные элементы A_n включают, например, аминокислоты (например, альфа-, бета-, гамма-, дельта- и эпсилон-аминокислоты, а также производные естественных аминокислот и аминокислот не природного происхождения), реакционноспособные химические вещества (например, азид или алкиновые цепи) с амином, или тиольный реагент, или их комбинации. Выбор структурного элемента A_n зависит, например, от природы реакционноспособной группы, применяемой для линкера, природы фрагмента скаффолда и растворителя, применяемого для химического синтеза.

Иллюстративные структурные элементы B_n и C_n включают любую пригодную структурную единицу химического объекта, как, например, необязательно замещенные ароматические группы (например, необязательно замещенный фенил или бензил), необязательно замещенные гетероциклильные группы (например, необязательно замещенный хинолинил, изохинолинил, индолил, изоиндолил, азаиндолил, бензимидазолил, азабензимидазолил, бензизоксазолил, пиридинил, пиперидил или пирролидинил), необязательно замещенные алкильные группы (например, необязательно замещенные линейные или разветвленные C_{1-6}-алкильные группы или необязательно замещенные C_{1-6}-аминоалкильные группы) или необязательно замещенные карбоциклильные группы (например, необязательно замещенный циклопропил, циклогексил или циклогексенил). Особенно пригодные структурные элементы B_n и C_n включают таковые с одной или более реакционноспособными группами, такими как необязательно замещенная группа (например, любая описанная в настоящем документе) с одним или необязательными заместителями, которые представляют собой реакционноспособные группы, или могут быть химически модифицированы с образованием реакционноспособных групп. Иллюстративные реакционноспособные группы включают одно или более из амина ($-NR_2$, где каждый R представляет собой, независимо, H или необязательно замещенный C_{1-6}-алкил), гидрокси, алкокси (-OR, где R представляет собой необязательно замещенный C_{1-6}-алкил, такой как метокси), карбокси (-COOH), амида или химически-реакционноспособных заместителей. Можно вводить сайт рестрикции, например, в метку B_n или C_n, причем комплекс может быть идентифицирован путем осуществления ПЦР и расщепления рестриктазами с использованием одного из соответствующих ферментов рестрикции.

Головной фрагмент.

В кодируемом химическом объекте головной фрагмент функционально связывает каждый химический объект с его кодирующей олигонуклеотидной меткой. Обычно головной фрагмент представляет

собой начальный олигонуклеотид по меньшей с двумя функциональными группами, которые дополнительно могут быть дериватизированы, причем первая функциональная группа функционально связывает первый химический объект (или его компонент) с головным фрагментом, а вторая функциональная группа функционально связывает одну или более меток с головным фрагментом. Необязательно можно применять бифункциональный спейсер в качестве разделительного фрагмента между головным фрагментом и химическим объектом.

Функциональные группы головного фрагмента можно применять для образования ковалентной связи с компонентом химического объекта и другой ковалентной связи с меткой. Компонент может представлять собой любую часть малой молекулы, такую как скаффолд с узлами разнообразия или структурным элементом. В качестве альтернативы, головной фрагмент может быть дериватизирован с обеспечением спейсера (например, разделительного фрагмента, отделяющего головной фрагмент от малой молекулы, которая должна быть получена в библиотеке), заканчивающегося функциональной группой (например, гидроксильной, амино-, карбоксильной, сульфгидрильной, алкинильной, азидо-или фосфатной группой), которая используется для образования ковалентной связи с компонентом химического объекта. Спейсер может быть присоединен к 5'-концу, в одном из внутренних положений, или к 3'-концу головного фрагмента. Если спейсер присоединен по одному из внутренних положений, то спейсер может быть функционально связан с дериватизированным основанием (например, по С5-положению уридина) или помещен внутри олигонуклеотида с применением известных в данной области стандартных методик. Иллюстративные спейсеры описаны в настоящем документе.

Головной фрагмент может характеризоваться любой пригодной структурой. Длина головного фрагмента может составлять, например, 1-100 нуклеотидов, предпочтительно 5-20 нуклеотидов и наиболее предпочтительно 5-15 нуклеотидов. Головной фрагмент может быть однонитевым или двунитевым, и может состоять из естественных или модифицированных нуклеотидов, как описано в настоящем документе. Например, химический фрагмент может быть функционально связан с 3'-концом или 5'-концом головного фрагмента. Согласно конкретным вариантам осуществления головной фрагмент включает шпилечную структуру, образуемую комплементарными основаниями в пределах последовательности. Например, химический фрагмент может быть функционально связан по внутреннему положению, с 3'-концом или 5'-концом головного фрагмента.

Обычно головной фрагмент включает не являющуюся самокомплементарной последовательность на 5'- или 3'-конце, которая обеспечивает возможность связывания олигонуклеотидной метки посредством полимеризации, ферментативного лигирования или химической реакции. Головной фрагмент может обеспечивать возможность лигирования олигонуклеотидных меток и необязательной очистки, а также стадий фосфорилирования. После добавления последней метки к 5'-концу последней метки можно добавлять дополнительную адаптерную последовательность. Иллюстративные адаптерные последовательности включают связывающую праймер последовательность или последовательность, содержащую метку (например, биотин). В случаях, когда применяют множество структурных элементов и соответствующие метки (например, 100), в ходе стадии синтеза олигонуклеотидов для получения необходимого числа меток можно использовать комбинаторную стратегию. Такие комбинаторные стратегии для синтеза ДНК известны в данной области. Компоненты полученной библиотеки могут быть амплифицированы с помощью ПЦР с последующим отбором связывающих объектов в сравнении с представляющей(ими) интерес мишенью(ями).

Головной фрагмент или комплекс может необязательно включать одну или более связывающих праймер последовательностей. Например, головной фрагмент содержит последовательность в петлевой области шпильки, которая служит в качестве области связывания праймера для амплификации, причем область связывания праймера характеризуется более высокой температурой плавления для комплекса с его комплементарным праймером (например, который может включать фланкирующие области для идентификации), чем для комплекса с последовательностью в составе головного фрагмента. Согласно другим вариантам осуществления комплекс включает две связывающие праймер последовательности (например, для обеспечения ПЦР-реакции) на любой из сторон одной или более меток, которые кодируют один или более структурных элементов. В качестве альтернативы, головной фрагмент может содержать одну связывающую праймер последовательность на 5'- или 3'-конце. Согласно другим вариантам осуществления головной фрагмент представляет собой шпильку, и петлевая область образует связывающий праймер сайт, или связывающий праймер сайт вводят посредством гибридизации олигонуклеотида с головным фрагментом с 3'-стороны петли. Олигонуклеотид для связывания праймера, содержащий область, гомологичную 3'-концу головного фрагмента и несущий связывающую праймер область на его 5'-конце (например, для обеспечения ПЦР-реакции), может быть гибридизирован с головным фрагментом, и может содержать метку, которая кодирует структурный элемент или добавление структурного элемента. Олигонуклеотид для связывания праймера может содержать дополнительную информацию, например, область рандомизированных нуклеотидов, например, длиной 2-16 нуклеотидов, которая включена для биоинформатического анализа.

Головной фрагмент может необязательно включать шпилечную структуру, причем такая структура может быть получена с помощью любого пригодного способа. Например, головной фрагмент может

включать комплементарные основания, которые образуют пары, вовлеченные в межмолекулярное спаривание оснований, такое как спаривание оснований ДНК по Уотсону-Крику (например, аденин-тимин и гуанин-цитозин) и/или неоднозначное спаривание оснований (например, гуанин-урацил, инозин-урацил, инозин-аденин и инозин-цитозин). Согласно другому примеру, головной фрагмент может включать модифицированные или замещенные нуклеотиды, которые могут образовывать дуплексные структуры с более высокой аффинностью по сравнению с немодифицированными нуклеотидами, причем такие модифицированные или замещенные нуклеотиды известны в данной области. Согласно еще одному примеру, головной фрагмент включает одно или более поперечно-связанных оснований для образования шпилечной структуры. Например, основания в пределах одной нити или основания в разных двух нитях могут быть поперечно связаны, например, с применением псоралена.

Головной фрагмент или комплекс могут необязательно включать одну или более меток, которые обеспечивают возможность выявления. Например, головной фрагмент, одна или более олигонуклеотидных меток и/или одна или более последовательностей праймера могут включать изотоп, радиоконтрастное средство, маркер, меченый элемент, флуоресцентную метку (например, родамин или флуоресцеин), хемилюминесцентную метку, квантовую точку и репортерную молекулу (например, биотин или гис-метку).

Согласно другим вариантам осуществления головной фрагмент или метка могут быть модифицированы для обеспечения растворимости в полуводных средах, неводных средах или средах с пониженным количеством воды (например, органических). Основания нуклеотидов головного фрагмента или метки можно сделать более гидрофобными путем модификации, например, С5-положений оснований Т или С алифатическими цепями без существенного нарушения их способности к образованию водородной связи с комплементарными им основаниями. Иллюстративные модифицированные или замещенные нуклеотиды представляют собой 5'-диметокситритил-N4-диизобутиламинометилиден-5-(1-пропинил)-2'-дезоксицитидин, 3'-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит; 5'-диметокситритил-5-(1-пропинил)-2'-дезоксиуридин, 3'-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит; 5'-диметокситритил-5-фтор-2'-дезоксиуридин, 3'-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит и 5'-диметокситритил-5-(пирен-1-илэтинил)-2'-дезоксиуридин или 3'-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит.

Кроме того, в головной олигонуклеотид могут быть внесены модификации, которые способствуют растворимости в органических растворителях. Например, с помощью азобензоламидофосфита в структуру головного фрагмента может быть внесен гидрофобный фрагмент. Такие вставки гидрофобных амидитов в головной фрагмент могут иметь место в любом участке молекулы. Однако вставка не должна препятствовать последующему введению метки с применением дополнительных ДНК-меток в ходе синтеза библиотеки или обеспечению ПЦР после завершения отбора или микроматричному анализу, в случае применения деконволюции метки. Такие добавления в структуру головного фрагмента, описанного в настоящем документе, будут придавать головному фрагменту растворимость, например, в 15%, 25%, 30%, 50%, 75%, 90%, 95%, 98%, 99% или 100% органическом растворителе. Таким образом, добавление гидрофобных остатков в структуру головного фрагмента обеспечивает повышенную растворимость в полу- или неводных (например, органических) средах, в то же время делая головной фрагмент пригодным для мечения олигонуклеотидами. Кроме того, ДНК-метки, которые затем вводят в библиотеку, также могут быть модифицированы по С5-положению Т или С оснований, так что они также делают библиотеку более гидрофобной и растворимой в органических растворителях для последующих стадий синтеза библиотеки.

Согласно конкретным вариантам осуществления головной фрагмент и первая метка могут быть одним и тем же объектом, т.е. может быть сконструировано множество объектов головной фрагмент-метка, все из которых характеризуются общими частями (например, связывающей праймер областью) и все из которых при этом отличаются по другой части (например, кодирующей области). Их можно использовать на стадии "разделения" и объединять после того, как произошло событие, которое они кодируют.

Согласно конкретным вариантам осуществления в головном фрагменте может быть закодирована информация, например, путем включения последовательности, которая кодирует первую стадию(и) разделения, или последовательность, которая кодирует идентичность библиотеки, например, с применением конкретной последовательности, относящейся к конкретной библиотеке.

Олигонуклеотидные метки.

Описанные в настоящем документе олигонуклеотидные метки (например, метку или участок головного фрагмента или участок хвостового фрагмента) можно применять для кодирования любой пригодной информации, такой как молекула, участок химического объекта, добавление компонента (например, скаффолда или структурного элемента), головной фрагмент в библиотеке, идентичность библиотеки, применение одного или более компонентов библиотеки (например, применение компонентов в аликвоте библиотеки) и/или происхождение компонента библиотеки (например, путем применения исходной последовательности).

Для кодирования любой информации можно применять любую последовательность в олигонуклеотиде. Таким образом, одна олигонуклеотидная последовательность может служить более чем для одной цели, например, для кодирования двух или более типов информации или для обеспечения начального

олигонуклеотида, который также кодирует один или более типов информации. Например, первая метка может кодировать добавление первого структурного элемента, а также возможность идентификации библиотеки. Согласно другому примеру, головной фрагмент можно применять для обеспечения начального олигонуклеотида, который функционально связывает химический объект с меткой, причем головной фрагмент дополнительно включает последовательность, которая кодирует идентичность библиотеки (т.е. идентифицирующую библиотеку последовательность). Соответственно, любая описанная в настоящем документе информация может быть закодирована в отдельных олигонуклеотидных метках или может быть объединена и закодирована в одной и той же олигонуклеотидной последовательности (например, олигонуклеотидная метка, такая как метка или головной фрагмент).

Последовательность структурного элемента кодирует идентичность структурного элемента и/или тип реакции связывания, осуществляемой со структурным элементом. Такая последовательность структурного элемента включена в метку, причем метка необязательно может включать один или более типов последовательности, описанной ниже (например, идентифицирующая библиотеку последовательность, последовательность применения и/или исходная последовательность).

Идентифицирующая библиотеку последовательность кодирует идентичность конкретной библиотеки. С целью обеспечения возможности смешивания двух или более библиотек, компонент библиотеки может содержать одну или более идентифицирующих библиотеку последовательностей, как, например, в составе идентифицирующей библиотеку метки (т.е. олигонуклеотида, включающего идентифицирующую библиотеку последовательность), в составе лигированной метки, в части последовательности головного фрагмента или в последовательности хвостового фрагмента. Такие идентифицирующие библиотеку последовательности можно применять для выведения кодируемых взаимосвязей, где последовательность метки переводится в хронологическую информацию о химическом превращении (синтезе), или коррелирует с ней. Соответственно, такие идентифицирующие библиотеку последовательности обеспечивают возможность смешивания двух или более библиотек вместе для отбора, амплификации, очистки, секвенирования и т.д.

Последовательность применения кодирует изменение со временем (т.е. применение) одного или более компонентов библиотеки в отдельной аликвоте библиотеки. Например, отдельные аликвоты можно обрабатывать с использованием разных условий реакции, структурных элементов и/или стадий отбора. В частности, такую последовательность можно применять для идентификации таких аликвот и определения их изменения со временем (применения), и, следовательно, обеспечения возможности смешивания вместе аликвот одной и той же библиотеки с другими изменениями со временем (применениями) (например, другими экспериментами по отбору) в целях смешивания вместе образцов для отбора, амплификации, очистки, секвенирования и т.д. Такие последовательности применения могут быть включены в головной фрагмент, хвостовой фрагмент, метку, метку применения (т.е. олигонуклеотид, включающий последовательность применения) или любую другу метку, описанную в настоящем документе (например, идентифицирующую библиотеку метку или исходную метку).

Исходная последовательность представляет собой вырожденную (случайную, стохастически сгенерированную) олигонуклеотидную последовательность любой пригодной длины (например, приблизительно шесть нуклеотидов), которая кодирует информацию о происхождении компонента библиотеки. Эта последовательность служит для стохастического разделения компонентов библиотеки, которые в ином случае являются идентичными во всех отношениях, на объекты, различимые по информации о последовательности, так что результаты по продуктам амплификации, полученным на основе уникальных исходных матриц (например, отобранных компонентов библиотеки) можно отличить от результатов по множеству продуктов амплификации, полученных на основе той же исходной матрицы (например, отобранного компонента библиотеки). Например, после образования библиотеки и перед стадией отбора каждый компонент библиотеки может включать отличающуюся исходную последовательность, как, например, в исходной метке. После отбора, отобранные компоненты библиотеки могут быть амплифицированы с получением продуктов амплификации, и участок компонента библиотеки, который, как ожидается, включает исходную последовательность (например, в исходной метке) можно обнаружить и сравнить с исходной последовательностью в каждом из других компонентов библиотеки. Поскольку исходные последовательности являются вырожденными, каждый продукт амплификации каждого компонента библиотеки должен характеризоваться отличающейся исходной последовательностью. Однако обнаружение одной и той же исходной последовательности в продукте амплификации может указывать на наличие множества ампликонов, происходящих из одной и той же молекулы-матрицы. Если требуется получение статистических и демографических данных относительно популяции кодирующих меток до амплификации, в отличие от ситуации с получением данных после амплификации, можно применять исходную метку. Эти исходные последовательности могут быть включены в головной фрагмент, хвостовой фрагмент, метку, исходную метку (т.е. олигонуклеотид, включающий исходную последовательность) или любую другую метку, описанную в настоящем документе (например, идентифицирующую библиотеку метку или метку применения).

Любые типы описанных в настоящем документе последовательностей могут быть включены в головной фрагмент. Например, головной фрагмент может включать одно или более из последовательности

структурного элемента, идентифицирующей библиотеку последовательности, последовательности применения или исходной последовательности.

Любые из этих описанных в настоящем документе последовательностей могут быть включены в хвостовой фрагмент. Например, хвостовой фрагмент может включать одно или более из идентифицирующей библиотеку последовательности, последовательности применения или исходной последовательности.

Любые описанные в настоящем документе метки могут включать соединительный фрагмент с фиксированной последовательностью на 5'- или 3'-конце или вблизи него. Соединительные фрагменты способствуют образованию связей (например, химических связей) за счет обеспечения реакционноспособной группы (например, химически-реакционноспособной группы или фотореакционноспособной группы) или за счет обеспечения сайта для средства, которое обеспечивает образование связи (например, средства для интеркалирующего фрагмента или обратимо-реакционноспособной группы в соединительном(ых) фрагменте(ах) или образующем поперечные связи олигонуклеотиде). Каждый 5'-соединительный фрагмент может быть одинаковым или отличающимся, и каждый 3'-соединительный фрагмент может быть одинаковым или отличающимся. В иллюстративном неограничивающем комплексе более чем с одной меткой, каждая метка может включать 5'-соединительный фрагмент и 3'-соединительный фрагмент, где каждый 5'-соединительный фрагмент имеет одинаковую последовательность и каждый 3'-соединительный фрагмент имеет одинаковую последовательность (например, где последовательность 5'-соединительного фрагмента может быть такой же, как последовательность 3'-соединительного фрагмента, или может отличаться от нее). Соединительный фрагмент обеспечивает последовательность, которая может применяться для образования одной или более связей. Для обеспечения связывания релейного праймера или гибридизации образующего поперечные связи олигонуклеотида, соединительный фрагмент может включать одну или более функциональных групп, обеспечивающих образование связи (например, связи, в отношении которой полимераза обладает пониженной способностью к прочтению или перемещению через нее, такой как химическая связь).

Такие последовательности могут включать любую модификацию, описанную в настоящем документе для олигонуклеотидов, такую как одна или более модификаций, которые способствуют растворимости в органических растворителях (например, любых описанных в настоящем документе, таких как применяемые для головного фрагмента), которые обеспечивают образование аналога естественной фосфодиэфирной связи (например, тиофосфатного аналога) или которые обеспечивают один или более искусственных олигонуклеотидов (например, 2'-замещенных нуклеотидов, таких как 2-О-метилированные нуклеотиды и 2'-фторнуклеотиды или любые описанные в настоящем документе).

Такие последовательности могут включать любые характеристики, описанные в настоящем документе для олигонуклеотидов. Например, такие последовательности могут быть включены в метку, размер которой составляет менее 20 нуклеотидов (например, как описано в настоящем документе). Согласно другим примерам, метки, включающие одну или более таких последовательностей, имеют приблизительно одинаковую массу (например, каждая метка имеет массу, которая составляет приблизительно +/- 10% от средней массы в пределах конкретного набора меток, которые кодируют определенную переменную); не содержат связывающей праймер (например, константной) области; не содержат константной области; или содержат укороченную константную область (например, длина составляет менее 30 нуклеотидов, менее 25 нуклеотидов, менее 20 нуклеотидов, менее 19 нуклеотидов, менее 18 нуклеотидов, менее 17 нуклеотидов, менее 16 нуклеотидов, менее 15 нуклеотидов, менее 14 нуклеотидов, менее 13 нуклеотидов, менее 12 нуклеотидов, менее 11 нуклеотидов, менее 10 нуклеотидов, менее 9 нуклеотидов, менее 8 нуклеотидов или менее 7 нуклеотидов).

Стратегии секвенирования для библиотек и олигонуклеотидов такой длины необязательно могут включать стратегии на основе конкатенации или сцепления для повышения точности прочтения или глубины секвенирования соответственно. В частности, отбор в отношении кодируемых библиотек, в последовательностях которых отсутствуют связывающие праймер области, были описаны в литературе для SELEX, как описано в Jarosch et al., Nucleic Acids Res. 34: e86 (2006), которая включена в настоящий документ посредством ссылки. Например, компонент библиотеки может быть модифицирован (например, после стадии отбора) так, чтобы он включал первую адаптерную последовательность на 5'-конце комплекса и вторую адаптерную последовательность на 3'-конце комплекса, причем первая последовательность является практически комплементарной второй последовательности, и в результате образуется дуплекс. Для дополнительного увеличения выхода к 5'-концу добавляют два фиксированных свисающих нуклеотида (например, CC).

Связи.

Связи по настоящему изобретению имеют место между олигонуклеотидами, которые кодируют определенную информацию (как, например, между головным фрагментом и меткой, между двумя метками или между меткой и хвостовым фрагментом). Иллюстративные связи включают фосфодиэфирные, фосфонатные и тиофосфатные связи. Согласно некоторым вариантам осуществления полимераза обладает пониженной способностью к прочтению или перемещению через одну или более связей. Согласно определенным вариантам осуществления в химические связи вовлечены одна или более из химически-

реакционноспособных групп, таких как монофосфатная и/или гидроксильная группа, фотореакционноспособная группа, интеркалирующий фрагмент, образующий поперечные связи олигонуклеотид или обратимо-кореакционноспособная группа.

Связь может быть проверена для определения того, обладает ли полимераза пониженной способностью к прочтению или перемещению через такую связь. Такая способность может быть проверена с помощью любого пригодного способа, такого как жидкостная хроматография с масс-спектрометрией, RT-PCR-анализ, демографический анализ последовательностей и/или ПЦР-анализ.

Согласно некоторым вариантам осуществления химическое лигирование включает применение одной или более химически-реакционноспособных пар для обеспечения связи, таких как пара монофосфат и гидроксил. Как описано в настоящем документе, читаемые связи могут быть синтезированы путем химического лигирования, например, с помощью реакции монофосфата, монотиофосфата или монофосфоната на 5'- или 3'-конце с гидроксильной группой на 5'- или 3'-конце в присутствии цианоимидазола и источника двухвалентного металла (например, $ZnCl_2$).

Другие иллюстративные химически-реакционноспособные пары представляют собой пару, включающую необязательно замещенную алкинильную группу и необязательно замещенную азидо-группу, образующие триазол посредством реакции 1,3-диполярного циклоприсоединения Хьюсгена; необязательно замещенный диен, обладающий 4π-электронной системой (например, необязательно замещенное 1,3-ненасыщенное соединение, такое как необязательно замещенный 1,3-бутадиен, 1-метокси-3-триметилсилилокси-1,3-бутадиен, циклопентадиен, циклогексадиен или фуран), и необязательно замещенный диенофил или необязательно замещенный гетеродиенофил, обладающий 2π-электронной системой (например, необязательно замещенная алкенильная группа или необязательно замещенная алкинильная группа), образующие циклоалкенил посредством реакции Дильса-Альдера; нуклеофил (например, необязательно замещенный амин или необязательно замещенный тиол) с напряженным гетероциклилом в качестве электрофила (например, необязательно замещенным эпоксидом, азиридином, азиридиний-ионом или эписульфоний-ионом), образующие гетероалкил посредством реакции раскрытия кольца; тиофосфатную группу с йодсодержащей группой, как, например, при мостиковом лигировании олигонуклеотида, содержащего 5'-йод-dT, с олигонуклеотид-3'-тиофосфатом; необязательно замещенную аминогруппу с альдегидной группой или кетогруппой, как, например, при реакции 3'-альдегид-модифицированного олигонуклеотида, который необязательно может быть получен путем окисления коммерчески доступного 3'-глицерил-модифицированного олигонуклеотида, с 5'-амино-модифицированным олигонуклеотидом (т.е. при реакции восстановительного аминирования) или 5'-гидразидо-модифицированным олигонуклеотидом; пару, включающую необязательно замещенную аминогруппу и карбоксильную группу или тиольную группу (например, с применением или без применения сукцинимидил-транс-4-(малеимидилметил)циклогексан-1-карбоксилата (SMCC) или 1-этил-3-(3-диметиламинопропил)карбодиимида (EDAC)); пару, включающую необязательно замещенный гидразин и альдегид или кетогруппу; пару, включающую необязательно замещенный гидроксиламин и альдегид или кетогруппу; или пару, включающую нуклеофил и необязательно замещенный галоидный алкил.

Комплексы на основе платины, алкилирующие средства или фуран-модифицированные нуклеотиды также можно применять в качестве химически-реакционноспособной группы для образования меж- или внутринитевых связей. Такие средства можно применять между двумя олигонуклеотидами, и они необязательно могут присутствовать в образующем поперечные связи олигонуклеотиде.

Иллюстративные неограничивающие комплексы на основе платины включают цисплатин (цис-диамминдихлорплатину (II), например, для образования внутринитевых связей GG), трансплатин (транс-диамминдихлорплатину (II), например, для образования межнитевых связей GXG, где X может представлять собой любой нуклеотид), карбоплатин, пиколатин (ZD0473), ормаплатин или оксалиплатин для образования, например, связей GC, CG, AG или GG. Любые из таких связей могут представлять собой меж- или внутринитевые связи.

Иллюстративные неограничивающие алкилирующие средства включают азотистый иприт (хлорметин, например, для образования связей GG), хлорамбуцил, мелфалан, циклофосфамид, циклофосфамид в форме пролекарства (например, 4-гидропероксициклофосфамид и ифосфамид)), 1,3-бис-(2-хлорэтил)-1-нитрозомочевину (BCNU, кармустин), азиридин (например, митомицин C, триэтиленмеламин или триэтилентиофосфорамид (тиотепа) для образования связей GG или AG), гексаметилмеламин, алкилсульфонат (например, бусульфан для образования связей GG) или нитрозомочевину (например, 2-хлорэтилнитрозомочевину для образования связей GG или CG, как, например, кармустин (BCNU), хлорозотоцин, ломустин (CCNU) и семустин (метил-CCNU)). Любые из таких связей могут представлять собой меж- или внутринитевые связи.

Фуран-модифицированные нуклеотиды также можно применять для образования связей. При окислении in situ (например, с использованием N-бромсукцинимида (NBS)), фурановый фрагмент образует реакционноспособное оксо-энальное производное, которое вступает в реакцию с комплементарным основанием с образованием межнитевой связи. Согласно некоторым вариантам осуществления фуран-модифицированные нуклеотиды образуют связи с комплементарным нуклеотидом A или C. Иллюстра-

тивные неограничивающие фуран-модифицированные нуклеотиды включают любой 2'-(фуран-2-ил)пропаноиламино-модифицированный нуклеотид; или ациклические модифицированные нуклеотиды 2-(фуран-2-ил)этилгликолевой нуклеиновой кислоты.

Фотореакционноспособные группы также можно применять в качестве реакционноспособной группы. Иллюстративные неограничивающие фотореакционноспособные группы включают интеркалирующий фрагмент, производное псоралена (например, псорален, HMT-псорален или 8-метоксипсорален), необязательно замещенную циановинилкарбазольную группу, необязательно замещенную винилкарбазольную группу, необязательно замещенную циановинильную группу, необязательно замещенную акриламидную группу, необязательно замещенную диазириновую группу, необязательно замещенный бензофенон (например, сукцинимидиловый сложный эфир 4-бензоилбензойной кислоты или бензофенонизотиоцианат), необязательно замещенную 5-(карбокси)винил-уридиновую группу (например, 5-(карбокси)винил-2'-дезоксиуридин) или необязательно замещенную азидную группу (например, арилазид или галогенированный арилазид, такой как сукцинимидиловый сложный эфир 4-азидо-2,3,5,6-тетрафторбензойной кислоты (ATFB)).

Интеркалирующие фрагменты также можно применять в качестве реакционноспособной группы. Иллюстративные неограничивающие интеркалирующие фрагменты включают производное псоралена, производное алкалоида (например, берберин, пальматин, коралин, сангвинарин (например, их иминиевые или алканоламиновые формы) или аристололактам-β-D-глюкозид), катион этидия (например, бромистый этидий), производное акридина (например, профлавин, акрифлавин или амсакрин), производное антрациклина (например, доксорубицин, эпирубицин, даунорубицин (дауномицин), идарубицин и аклярубицин) или талидомид.

В случае образующего поперечные связи олигонуклеотида любую пригодную реакционноспособную группу (например, описанную в настоящем документе) можно применять для образования меж- или внутринитевых связей. Иллюстративные реакционноспособные группы включают химически-реакционноспособную группу, фотореакционноспособную группу, интеркалирующий фрагмент и обратимо-кореакционноспособную группу. Образующие поперечные связи средства для применения с образующими поперечные связи олигонуклеотидами включают без ограничения алкилирующие средства (например, как описано в настоящем документе), цисплатин (цис-диамминдихлорплатину(II)), транс-диамминдихлорплатину(II), псорален, HMT-псорален, 8-метоксипсорален, фуран-модифицированные нуклеотиды, 2-фтор-дезоксиинозин (2-F-dI), 5-бром-дезоксицитозин (5-Br-dC), 5-бром-дезоксиуридин (5-Br-dU), 5-йод-дезоксицитозин (5-I-dC), 5-йод-дезоксиуридин (5-I-dU), сукцинимидил транс-4-(малеимидилметил)циклогексан-1-карбоксилат, SMCC, EDAC или сукцинимидилацетилтиоацетат (SATA).

Олигонуклеотиды также можно модифицировать так, чтобы они содержали тиольные фрагменты, которые могут вступать в реакцию с различными тиольными реакционноспособными группами, такими как малеимиды, галогены и йодамцетамиды и, таким образом, их можно применять для поперечного связывания двух олигонуклеотидов. Тиольные группы могут быть связаны с 5'- или 3'-концом олигонуклеотида.

Для межнитевого поперечного связывания олигонуклеотидов дуплекса в положении пиримидина (например, тимидина) может быть выбран интеркалирующий фотореакционноспособный фрагмент псорален. Псорален интеркалирует в дуплекс и образует ковалентные межнитевые поперечные связи с пиримидинами, преимущественно в сайтах 5'-ТрА, при облучении ультрафиолетовым светом (приблизительно 254 нм). Псораленовый фрагмент может быть ковалентно присоединен к модифицированному олигонуклеотиду (например, с помощью алкановой цепи, такой как C_{1-10}-алкил, или полиэтиленгликолевой группы, такой как $-(CH_2CH_2O)_nCH_2CH_2-$, где n представляет собой целое число от 1 до 50). Также можно применять иллюстративные производные псоралена, где неограничивающие производные включают 4'-(гидроксиэтоксиметил)-4,5',8-триметилпсорален (HMT-псорален) и 8-метоксипсорален.

Для введения связи можно модифицировать различные участки образующего поперечные связи олигонуклеотида. Например, концевые тиофосфаты в составе олигонуклеотидов также можно применять для связывания двух смежных олигонуклеотидов. Галогенированные урацилы/цитозины также можно применять в качестве обеспечивающих поперечное связывание модификаций в олигонуклеотиде. Например, можно проводить реакцию модифицированных 2-фтор-дезоксиинозином (2-F-dI) олигонуклеотидов с дисульфид-содержащими диаминами или тиопропиламинами с образованием дисульфидных связей.

Как описано ниже, обратимо-кореакционноспособные группы включают таковые, выбранные из циановинилкарбазольной группы, циановинильной группы, акриламидной группы, тиольной группы или сульфонилэтиловых тиоэфиров. Необязательно замещенную циановинилкарбазольную (CNV) группу также можно применять в олигонуклеотидах для поперечного связывания с пиримидиновым основанием (например, цитозином, тимином и урацилом, а также их модифицированными основаниями) в комплементарных нитях. CNV-группы способствуют [2+2] циклоприсоединению со смежным пиримидиновым основанием при облучении при 366 нм, что обуславливает в результате межнитевое поперечное связывание. Облучение при 312 нм устраняет поперечную связь, и, таким образом, предусматривается способ

обратимого поперечного связывания олигонуклеотидных нитей. Неограничивающей CNV-группой является 3-циановинилкарбазол, который может быть включен в виде карбоксивинилкарбазольного нуклеотида (например, в виде 3-карбоксивинилкарбазол-1'-β-дезоксирибозид-5'-трифосфата).

CNV-группа может быть модифицирована с заменой реакционноспособной цианогруппы другой реакционноспособной группой для получения необязательно замещенной винилкарбазольной группы. Иллюстративные неограничивающие реакционноспособные группы для винилкарбазольной группы включают амидную группу $-CONR_{N1}R_{N2}$, где каждый R_{N1} и R_{N2} может быть одинаковым или может отличаться и представляет собой независимо H и C_{1-6}-алкил, например, $-CONH_2$; карбоксильную группу $-CO_2H$; или C_{2-7}-алкоксикарбонильную группу (например, метоксикарбонил). Кроме того, реакционноспособная группа может быть расположена у альфа- или бета-углерода винильной группы. Иллюстративные винилкарбазольные группы включают циановинилкарбазольную группу, как описано в настоящем документе; амидовинилкарбазольную группу (например, амидоовинилкарбазольный нуклеотид, такой как 3-амидовинилкарбазол-1'-β-дезоксирибозид-5'-трифосфат); карбоксивинилкарбазольную группу (например, карбоксивинилкарбазольный нуклеотид, такой как 3-карбоксивинилкарбазол-1'-β-дезоксирибозид-5'-трифосфат); и C_{2-7}-алкоксикарбонилвинилкарбазольную группу (например, алкоксикарбонилвинилкарбазольный нуклеотид, такой как 3-метоксикарбонилвинилкарбазол-1'-β-дезоксирибозид-5'-трифосфат). Дополнительные необязательно замещенные винилкарбазольные группы и нуклеотиды с такими группами представлены в химических формулах патента США № 7972792 и Yoshimura and Fujimoto, Org. Lett. 10:3227-3230 (2008), оба из которых включены в настоящий документ посредством ссылки во всей своей полноте.

Другие обратимо-реакционноспособные группы включают тиольную группу и другую тиольную группу для образования дисульфида, а также тиольную группу и винилсульфоновую группу для образования сульфонилэтиловых тиоэфиров. Тиол-тиольные группы необязательно могут включать связь, образуемую в результате реакции с бис-((N-йодацетил)пиперазинил)сульфонродамином. Другие обратимореакционноспособные группы (например, такие как некоторые фотореакционноспособные группы) включают необязательно замещенные бензофеноновые группы. Неограничивающим примером является бензофенонурацил (BPU), который можно применять для избирательного в отношении сайта и последовательности образования межнитевой поперечной связи BPU-содержащих олигонуклеотидных дуплексов. Такая поперечная связь может быть устранена при нагревании, что обеспечивает способ обратимого поперечного связывания двух олигонуклеотидных нитей.

Согласно другим вариантам осуществления химическое лигирование предусматривает введение аналога фосфодиэфирной связи, например, для проводимого после отбора ПЦР-анализа и секвенирования. Иллюстративные аналоги фосфодиэфирной связи включают тиофосфатную связь (например, вводимую с применением тиофосфатной группы и уходящей группы, такой как йодогруппа), фосфорамидную связь или дитиофосфатную связь (например, вводимую с применением дитиофосфатной группы и уходящей группы, такой как йодогруппа).

Для любой из описанных в настоящем документе групп (например, химически-реакционноспособной группы, фотореакционноспособной группы, интеркалирующего фрагмента, образующего поперечные связи олигонуклеотида или обратимо-кореакционноспособной группы) группа может быть встроена на конце олигонуклеотида, или вблизи него, или между 5'- и 3'-концами. Кроме того, в каждом олигонуклеотиде могут присутствовать одна или более групп. Если требуются пары реакционноспособных групп, то олигонуклеотиды могут быть сконструированы так, чтобы они содействовали протеканию реакции между парой групп. Согласно неограничивающему примеру циановинилкарбазольной группы, которая совместно реагирует с пиримидиновым основанием, первый олигонуклеотид можно сконструировать так, чтобы он включал циановинилкарбазольную группу на 5'-конце или вблизи него. Согласно этому примеру, второй олигонуклеотид можно сконструировать так, чтобы он был комплементарным первому олигонуклеотиду и включал кореакционноспособное пиримидиновое основание в положении, которое совпадает с положением циановинилкарбазольной группы, когда первый и второй олигонуклеотид гибридизируются. Любые из описанных в настоящем документе групп и любые олигонуклеотиды с одной или более группами могут быть сконструированы так, чтобы они содействовали протеканию реакции между группами для образования одной или более связей.

Бифункциональные спейсеры.

Бифункциональный спейсер между головным фрагментом и химическим объектом можно изменять для обеспечения соответствующего разделительного фрагмента и/или для повышения растворимости головного фрагмента в органическом растворителе. Коммерчески доступным является широкое разнообразие спейсеров, которые могут соединять головной фрагмент с элементами библиотеки малых молекул. Спейсер, как правило, состоит из линейных или разветвленных цепей, и может включать C_{1-10}-алкил, гетероалкил из 1-10 атомов, C_{2-10}-алкенил, C_{2-10}-алкинил, C_{5-10}-арил, циклическую или полициклическую систему из 3-20 атомов, фосфодиэфир, пептид, олигосахарид, олигонуклеотид, олигомер, полимер или полиалкилгликоль (например, полиэтиленгликоль, такой как $-(CH_2CH_2O)_nCH_2CH_2-$, где n представляет собой целое число от 1 до 50) или их комбинацию.

Бифункциональный спейсер может обеспечивать соответствующий разделительный фрагмент между головным фрагментом и химическим объектом библиотеки. Согласно определенным вариантам осуществления бифункциональный спейсер включает три части. Часть 1 может представлять собой реакционноспособную группу, которая образует ковалентную связь с ДНК, например, представлять собой карбоновую кислоту, предпочтительно активированную N-гидроксисукцинимидным (NHS) эфиром для реакции с аминогруппой ДНК (например, амино-модифицированным dT), амидит для модификации 5'- или 3'-конца однонитевого головного фрагмента (что достигается с помощью стандартных способов химии олигонуклеотидов), химически-реакционноспособные пары (например, азид-алкиновое циклоприсоединение в присутствии в присутствии катализатора Cu(I) или любые описанные в настоящем документе способы) или тиольные реакционноспособные группы. Часть 2 также может представлять собой реакционноспособную группу, которая образует ковалентную связь с химическим объектом, либо структурным элементом A_n, либо скаффолдом. Такая реакционноспособная группа может представлять собой, например, амин, тиол, азид или алкин. Часть 3 может представлять собой химически инертный разделительный фрагмент варьирующей длины, встроенный между частью 1 и 2. Такой разделительный фрагмент может представлять собой цепь из звеньев этиленгликоля (например, PEG разной длины), алкановую, алкеновую, полиеновую цепь или пептидную цепь. Спейсер может содержать ветви или вставки с гидрофобными фрагментами (такими как, например, бензольные кольца) для улучшения растворимости головного фрагмента в органических растворителях, а также флуоресцентные фрагменты (например, флуоресцеин или Cy-3), применяемые для целей обнаружения в отношении библиотеки. Гидрофобные остатки в структуре головного фрагмента можно изменять в зависимости от структуры спейсера для облегчения синтеза библиотеки в органических растворителях. Например, комбинацию головной фрагмент и спейсер конструируют так, чтобы она характеризовалась соответствующими остатками, причем коэффициент разделения октанол:вода (P_{oct}) должен составлять, например, 1,0-2,5.

Спейсеры могут быть выбраны эмпирическим путем для данной конструкции библиотеки малых молекул так, чтобы библиотека могла быть синтезирована в органическом растворителе, например, в 15%, 25%, 30%, 50%, 75%, 90%, 95%, 98%, 99% или 100% органическом растворителе. Спейсер можно изменять с применением модельных реакций до синтеза библиотеки для выбора соответствующей длины цепи, при которой головной фрагмент растворяется в органическом растворителе. Иллюстративные спейсеры включают таковые, характеризующиеся увеличенной длиной алкильной цепи, увеличенным количеством звеньев полиэтиленгликоля, наличием разветвленных молекул с положительными зарядами (для нейтрализации отрицательных зарядов фосфатов головного фрагмента) или повышенной степенью гидрофобности (например, за счет добавления структур с бензольным кольцом).

Примеры коммерчески доступных спейсеров включают амино-карбоксильные спейсеры, как, например, таковые, представляющие собой пептиды (например, Z-Gly-Gly-Gly-Osu (N-альфа-бензилоксикарбонил-(глицин)$_3$-N-сукцинимидиловый сложный эфир) или Z-Gly-Gly-Gly-Gly-Gly-Gly-Osu (N-альфа-бензилоксикарбонил-(глицин)$_6$-N-сукцинимидиловый сложный эфир, SEQ ID NO: 1)), PEG (например, Fmoc-амино-ПЭГ2000-NHS или амино-PEG (12-24)-NHS) или цепи алкановых кислот (например, Boc-ε-аминокапроновая кислота-Osu); спейсеры химически реакционноспособной пары, как, например, химически реакционноспособные пары, описанные в настоящем документе, в сочетании с пептидным фрагментом (например, азидогомоаланин-Gly-Gly-Gly-OSu (SEQ ID NO: 2) или пропаргилглицин-Gly-Gly-Gly-OSu (SEQ ID NO: 3)), PEG (например, азидо-PEG-NHS) или фрагмент цепи алкановой кислоты (например, 5-азидопентановая кислота, (S)-2-(азидометил)-1-Boc-пирролидин, 4-азидоанилин или сложный N-гидроксисукцинимидный эфир 4-азидо-бутан-1-овой кислоты); тиол-реактивные спейсеры, как, например, таковые, представляющие собой PEG (например, SM(PEG)n NHS-PEG-малеимид), алкановые цепи (например, 3-(пиридин-2-илдисульфанил)-пропионовая кислота-Osu или сульфосукцинимидил-6-(3'-[2-пиридилдитио]пропионамидо)гексаноат)); и амидиты для синтеза олигонуклеотидов, такие как модификаторы для введения аминогруппы (например, 6-(трифторацетиламино)-гексил-(2-цианоэтил)-(N,N-диизопропил)-амидофосфит), модификаторы для введения тиольной группы (например, S-тритил-6-меркаптогексил-1-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит или модификаторы химически-реакционноспособных пар (например, 6-гексин-1-ил-(2-цианоэтил)-(N,N-диизопропил)-амидофосфит, 3-диметокситритилокси-2-(3-(3-пропаргилоксипропанамидо)пропанамидо)пропил-1-O-сукциноил, длинноцепочечный алкиламино-CPG или сложный N-гидроксисукцинимидный эфир 4-азидо-бутан-1-овой кислоты)). Дополнительные спейсеры известны в данной области, и спейсеры, которые можно применять в ходе синтеза библиотеки, включают без ограничения 5'-O-диметокситритил-1',2'-дидезоксирибоза-3'-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит; 9-O-диметокситритилтриэтиленгликоль, 1-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит; 3-(4,4'-диметокситритилокси)пропил-1-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит; и 18-O-диметокситритилгексаэтиленгликоль, 1-[(2-цианоэтил)-(N,N-диизопропил)]амидофосфит. Любые из спейсеров, описанных в настоящем документе, могут быть добавлены в тандеме друг с другом в различных комбинациях для получения спейсеров различной требуемой длины.

Спейсеры также могут быть разветвленными, причем разветвленные спейсеры являются хорошо известными в данной области, и примеры их могут включать симметричные или асимметричные удвои-

тели или симметричный утроитель. См., например, Newcome et al., Dendritic Molecules: Concepts, Synthesis, Perspectives, VCH Publishers (1996); Boussif et al., Proc. Natl. Acad. Sci. USA 92:7297-7301 (1995); и Jansen et al., Science 266:1226 (1994).

Способы определения нуклеотидной последовательности комплекса.

В настоящем изобретение представлены способы, которые предусматривают определение нуклеотидной последовательности комплекса так, чтобы можно было установить взаимосвязи кодирования между последовательностью подвергнутых сборке меток и структурными единицами (или структурными элементами) химического объекта. В частности, идентичность и/или изменение со временем химического объекта могут быть выведены из последовательности оснований в олигонуклеотиде. С применением такого способа библиотеку, включающую разнообразные химические объекты или компоненты (например, малые молекулы или пептиды), можно определить с использованием конкретной последовательности метки.

Любые из связей, описанных в настоящем документе, могут быть обратимыми или необратимыми. Обратимые связи включают вовлеченные в фотореакции связи (например, образуемые циановинилкарбазольной группой и тимидином) и вовлеченные в окислительно-восстановительные реакции связи. Дополнительные связи описаны в настоящем документе.

Согласно альтернативному варианту осуществления "нечитаемую" связь можно подвергнуть ферментативной репарации для получения читаемой или по меньшей мере связи, через которую может перемещаться полимераза. Способы ферментативной репарации хорошо известны специалистам в данной области, и они включают без ограничения механизмы репарации пиримидиновых (например, тимидиновых) димеров (например, с применением фотолиазы или гликозилазы (например, гликозилазы пиримидиновых димеров (PDG) TT4)), механизмы эксцизионной репарации оснований (например, с применением гликозилазы, апуриновой/апиримидиновой (AP) эндонуклеазы, эндонуклеазы Flap или поли(АДФ-рибоза)-полимеразы (например, апуриновой/апиримидиновой (AP) эндонуклеазы человека, APE 1; эндонуклеазы III (Nth); эндонуклеазы IV; эндонуклеазы V; формамидопиримидин[fapy]-ДНК-гликозилазы (Fpg); 8-оксогуанин-гликозилазы 1 человека (@-изоформы) (hOGG1); эндонуклеазы 1 человека, подобной эндонуклеазе VIII (hNEIL1); урацил-ДНК-гликозилазы (UDG); монофункциональной урацил-ДНК-гликозилазы человека, избирательной в отношении однонитевых участков (SMUG1); и алкиладенин-ДНК-гликозилазы человека (hAAG)), которую необязательно можно комбинировать с одной или более эндонуклеазами, ДНК-или РНК-полимеразами и/или лигазами для репарации), механизмы репарации метилированных нуклеотидов (например, с применением метилгуанин-метилтрансферазы), механизмы AP-репарации (например, с применением апуриновой/апиримидиновой (AP) эндонуклеазы (например, APE 1; эндонуклеазы III; эндонуклеазы IV; эндонуклеазы V; Fpg; hOGG1 и hNEIL1), которую необязательно можно комбинировать с одной или более эндонуклеазами, ДНК- или РНК-полимеразами и/или лигазами для репарации), механизмы эксцизионной репарации нуклеотидов (например, с применением белков эксцизионной репарации кросс-комплементарной группы или эксцизионных нуклеаз, которые необязательно можно комбинировать с одной или более эндонуклеазами, ДНК- или РНК-полимеразами и/или лигазами для репарации) и механизмы репарации ошибочно спаренных нуклеотидов (например, с применением эндонуклеазы (например, T7 эндонуклеазы I; MutS, MutH и/или MutL), которую необязательно можно комбинировать с одной или более эндонуклеазами, ДНК- или РНК-полимеразами и/или лигазами для репарации). Коммерческие смеси ферментов доступны для обеспечения таких типов механизмов репарации, например, смесь для репарации PreCR® Repair Mix (New Enlgand Biolabs Inc., Ипсвич, Массачусетс), которая включает Taq ДНК-лигазу, эндонуклеазу IV, Bst ДНК-полимеразу, Fpg, урацил-ДНК-гликозилазу (UDG), T4 PDG (T4 эндонуклеазу V) и эндонуклеазу VIII.

Способы кодирования химических объектов в пределах библиотеки.

В способах по настоящему изобретению можно использовать библиотеку с различным числом химических объектов, кодированных олигонуклеотидными метками. Примеры структурных элементов и кодирующих ДНК-меток можно найти в публикации заявки на патент США № 2007/0224607, структурные блоки и метки из которой включены в настоящий документ посредством ссылки.

Каждый химический объект образован одним или более структурными элементами, и необязательно скаффолдом. Скаффолд служит для обеспечения одного или более узлов разнообразия в конкретной геометрии (например, триазин для обеспечения трех узлов, пространственно расположенных вокруг гетероарильного кольца или линейной геометрии).

Структурные элементы и кодирующие их метки можно добавлять прямо или опосредованно (например, посредством спейсера) к головному фрагменту с образованием комплекса. Если головной фрагмент включает спейсер, структурный элемент или скаффолд добавляют к концу спейсера. Если спейсер отсутствует, структурный элемент можно добавлять прямо к головному фрагменту, или структурный элемент сам по себе может включать спейсер, который вступает в реакцию с функциональной группой головного фрагмента. Иллюстративные спейсеры и головные фрагменты описаны в настоящем документе.

Скаффолд можно добавлять любым пригодным способом. Например, скаффолд можно добавлять к концу спейсера или головного фрагмента, и последовательные структурные элементы можно добавлять к

доступным узлам разнообразия скаффолда. Согласно другому примеру, структурный элемент A_n сперва добавляют к спейсеру или головному фрагменту, а затем проводят реакцию узла разнообразия скаффолда S с функциональной группой структурного элемента A_n. Олигонуклеотидные метки, кодирующие конкретный скаффолд, необязательно можно добавлять к головному фрагменту или комплексу. Например, S_n добавляют к комплексу в n реакционных сосудах, где n представляет собой целое число больше единицы, и метка S_n (т.е. метку S_1, S_2, ..., S_{n-1}, S_n) связывается с функциональной группой комплекса.

Структурные элементы можно добавлять на нескольких стадиях синтеза. Например, аликвоту головного фрагмента, необязательно содержащего присоединенный спейсер, разделяют на n реакционных сосудов, где n представляет собой целое число, составляющее два или больше. На первой стадии структурный элемент An добавляют в каждый реакционный сосуд n (т.е. структурный элемент A_1, A_2, ... A_{n-1}, A_n добавляют в реакционный сосуд 1, 2,... n-1, n), где n представляет собой целое число, и каждый структурный элемент A_n является уникальным. На второй стадии скаффолд S добавляют в каждый реакционный сосуд с образованием комплекса A_n-S. Необязательно скаффолд S_n можно добавлять в каждый реакционный сосуд с образованием комплекса A_n-S_n, где n представляет собой целое число, составляющее более двух, и каждый скаффолд S_n может быть уникальным. На третьей стадии структурный элемент B_n добавляют в каждый реакционный сосуд n, содержащий комплекс A_n-S (т.е. структурный элемент B_1, B_2,... B_{n-1}, B_n добавляют в реакционный сосуд 1,2,... n-1, n, содержащий комплекс A_1-S, A_2-S,... A_{n-1}-S, A_n-S), где каждый структурный элемент B_n является уникальным. На дополнительных стадиях структурный элемент C_n можно добавлять в каждый реакционный сосуд n, содержащий комплекс B_n-A_n-S (т.е. структурный элемент C_1, C_2,... C_{n-1}, C_n добавляют в реакционный сосуд 1, 2,... n-1, n, содержащий комплекс B_1-A_1-S... B_n-A_n-S), где каждый структурный элемент C_n является уникальным. Полученная в результате библиотека будет содержать n^3 число комплексов с n^3 метками. Таким образом, дополнительные стадии синтеза можно применять для связывания дополнительных структурных элементов для внесения дополнительного разнообразия в библиотеку.

После образования библиотеки полученные в результате комплексы можно необязательно очищать и подвергать реакции полимеризации или лигирования, например, с хвостовым фрагментом. Эту общую стратегию можно расширить с включением дополнительных узлов разнообразия и структурных элементов (например, D, E, F и т.д.). Например, проводят реакцию первого узла разнообразия со структурными элементами и/или S и кодируют с помощью олигонуклеотидной метки. Затем проводят реакцию дополнительных структурных элементов с полученным комплексом, и последующий узел разнообразия дериватизируют с помощью дополнительных структурных элементов, который кодируется праймером, применяемым для реакции полимеризации или лигирования.

Для получения кодируемой библиотеки олигонуклеотидные метки добавляют к комплексу после или перед каждой стадией синтеза. Например, перед или после добавления структурного элемента A_n в каждый реакционный сосуд метка A_n связывается с функциональной группой головного фрагмента (т.е. метку A_1, A_2,...A_{n-1}, A_n добавляют в реакционный сосуд 1, 2,... n-1, n, содержащий головной фрагмент). Каждая метка A_n характеризуется отличающейся последовательностью, которая коррелирует с каждым уникальным структурным элементом A_n, и определение последовательности метки A_n дает информацию о химической структуре структурного элемента A_n. Таким образом, дополнительные метки применяют для кодирования дополнительных структурных элементов или дополнительных скаффолдов.

Кроме того, последняя добавленная к комплексу метка может либо включать связывающую праймер последовательность, либо обеспечивать функциональную группу для обеспечения возможности связывания (например, путем лигирования) связывающей праймер последовательности. Связывающую праймер последовательность можно применять для амплификации и/или секвенирования олигонуклеотидных меток комплекса. Иллюстративные способы амплификации и секвенирования включают полимеразную цепную реакцию (ПЦР), полимеразную цепную реакцию с линейной амплификацией (LCR), амплификацию по типу катящегося кольца (RCA) или любой другой известный в данной области способ амплификации или определения последовательностей нуклеиновых кислот.

С применением таких способов можно получать крупные библиотеки, содержащие большое число кодируемых химических объектов. Например, проводят реакцию головного фрагмента со спейсером и структурным элементом A_n, который предусматривает 1000 разных вариантов (т.е. n = 1000). Для каждого структурного элемента A_n проводят лигирование ДНК-метки A_n или удлинение праймера в сторону головного фрагмента. Такие реакции можно осуществлять в 1000-луночном планшете или 10 × 100-луночных планшетах. Все продукты реакций могут быть объединены, необязательно очищены и распределены по планшетам второго набора. Затем такую же процедуру можно осуществлять для структурного элемента Bn, который также предусматривает 1000 разных вариантов. ДНК-метку B_n можно лигировать с комплексом A_n-головной фрагмент, и все продукты реакции могут быть объединены. Полученная библиотека включает 1000 × 1000 комбинаций $A_n \times B_n$ (т.е. 1000000 соединений), меченых 1000000 разных комбинаций меток. Такой же подход можно расширить с добавлением структурных элементов C_n, D_n, E_n и т.д. Созданную библиотеку затем можно применять для идентификации соединений, которые связываются с мишенью. Структуру химических объектов, которые связываются с библиотекой, необязатель-

но можно оценивать с помощью ПЦР и секвенирования ДНК-меток для идентификации соединений, в отношении которых было осуществлено обогащение.

Такой способ можно модифицировать во избежание мечения после добавления каждого структурного элемента или во избежание объединения (или смешивания). Например, способ можно модифицировать путем добавления структурного элемента A_n в n реакционных сосудов, где n представляет собой целое число больше единицы, и добавления идентичного структурного элемента B_1 в каждую реакционную лунку. В данном случае B_1 является идентичным в отношении каждого химического объекта, и, следовательно, олигонуклеотидная метка, кодирующая такой структурный элемент, не требуется. После добавления структурного элемента комплексы можно объединять, или можно не объединять. Например, библиотеку не подвергают объединению после завершения конечной стадии добавления структурного элемента, и пулы подвергают скринингу отдельно для идентификации соединения(й), которое(ые) связывается(ются) с мишенью. Во избежание объединения всех продуктов реакций после синтеза, для отслеживания связывания на поверхности сенсора в высокопроизводительном формате (например, 384-луночных планшетах и 1536-луночных планшетах), например, можно применять анализ связывания, например, ELISA, SPR, ITC, анализ сдвига Tm, SEC или им подобные. Например, структурный элемент A_n может кодироваться ДНК-меткой A_n, а структурный элемент Bn может кодироваться его положением в пределах лунки планшета. Кандидатные соединения затем можно идентифицировать с применением анализа связывания (например, ELISA, SPR, ITC, анализа сдвига Tm, SEC или им подобных) и путем анализа A_n-меток с помощью секвенирования, микроматричного анализа и/или анализа расщепления рестриктазами. Такой анализ позволяет идентифицировать комбинации структурных элементов A_n и B_n, которые образуют требуемые молекулы.

Способ амплификации необязательно может предусматривать образование эмульсии типа "вода в масле" для создания множества водных микрореакторов. Условия реакции (например, концентрация комплекса и размер микрореакторов) могут быть скорректированы для обеспечения, в среднем, микрореактора, содержащего по меньшей мере один компонент библиотеки соединений. Каждый микрореактор также может содержать мишень, одну гранулу, способную к связыванию с комплексом или участком комплекса (например, одной или более метками) и/или связыванию с мишенью, и раствор для реакции амплификации с одним или более необходимыми реагентами для осуществления амплификации нуклеиновых кислот. После амплификации метки в микрореакторах, амплифицированные копии метки будут связываться с гранулами в микрореакторах, и покрытые гранулы можно идентифицировать с помощью любого пригодного способа.

После идентификации структурных элементов из первой библиотеки, которые связываются с представляющей интерес мишенью, может быть получена вторая библиотека путем повторения способа. Например, можно добавлять один или два дополнительных узла разнообразия, и вторую библиотеку создают и генерируют из нее выборки, как описано в настоящем документе. Этот процесс можно повторять столько раз, сколько необходимо для получения молекул с требуемыми молекулярными и фармацевтическими свойствами.

Для добавления скаффолда, структурных элементов, спейсеров, связей и меток можно применять различные методики лигирования. Соответственно, любые из стадий связывания, описанных в настоящем документе, могут включать любые пригодные методику или методики лигирования. Иллюстративные методики лигирования включают ферментативное лигирование, как, например, применение одной или более РНК-лигаз и/или ДНК-лигаз, как описано в настоящем документе; и химическое лигирование, как, например, применение химически реакционноспособных пар, как описано в настоящем документе.

Способы скрининга.

Существует множество известных технических способов определения связывания соединений с белками, например, путем определения Kd. Способы выявления или количественного определения связывания соединения с белком-мишенью предусматривают, например, анализы поглощения, флуоресценции, рамановского рассеяния, фосфоресценции, люминесценции, люциферазной активности и радиоактивности. Иллюстративные методики включают метод поверхностного плазмонного резонанса (SPR) и поляризации флуоресценции (FP). С помощью SPR измеряют изменение показателя преломления поверхности металла при связывании соединения с белком, который иммобилизован на такой поверхности металла, тогда как с помощью FP измеряют изменение скорости молекулярных колебаний для соединения при его связывании с белком с применением потери поляризации падающего света. Согласно некоторым вариантам осуществления такие способы можно применять для экспериментального определения связывания кандидатного соединения, для которого с применением способов по настоящему изобретению было предсказано связывание с белком-мишенью.

В качестве альтернативы, соединения, которые связываются с белками-мишенями, можно идентифицировать с применением способов на основе аффинности. Например, белки-мишени с аффинными метками (например, поли-His-метками) можно предварительно инкубировать с насыщающей концентрацией одного или более кандидатных соединений. Последующая аффинная очистка и идентификация соединений (например, за счет использования метки-идентификатора) будут обеспечивать идентификацию соединений, которые связываются с белком-мишенью.

Белки-мишени.

Белок-мишень (например, белок-мишень эукариот, такой как белок-мишень млекопитающих или белок-мишень грибов, или белок-мишень прокариот, такой как белок-мишень бактерий) представляет собой белок, который опосредует развитие болезненного состояния или симптома болезненного состояния. Таким образом, требуемый терапевтический эффект может быть достигнут путем модулирования (подавления или повышения) его активности.

Белки-мишени могут быть встречающимися в природе, например, дикого типа. В качестве альтернативы, белок-мишень может отличаться от белка дикого типа, но сохранять биологическую функцию, например, будучи аллельным вариантом, сплайс-мутантом или биологически активным фрагментом.

Согласно некоторым вариантам осуществления белок-мишень представляет собой фермент (например, киназу). Согласно некоторым вариантам осуществления белок-мишень представляет собой трансмембранный белок. Согласно некоторым вариантам осуществления белок-мишень имеет структуру в виде суперспирали. Согласно определенным вариантам осуществления белок-мишень представляет собой один из белков димерного комплекса.

Согласно некоторым вариантам осуществления белок-мишень представляет собой ГТФазу, такую как DIRAS1, DIRAS2, DIRAS3, ERAS, GEM, HRAS, KRAS, MRAS, NKIRAS1, NKIRAS2, NRAS, RALA, RALB, RAP1A, RAP1B, RAP2A, RAP2B, RAP2C, RASD1, RASD2, RASL10A, RASL10B, RASL11A, RASL11B, RASL12, REM1, REM2, RERG, RERGL, RRAD, RRAS, RRAS2, RHOA, RHOB, RHOBTB1, RHOBTB2, RHOBTB3, RHOC, RHOD, RHOF, RHOG, RHOH, RHOJ, RHOQ, RHOU, RHOV, RND1, RND2, RND3, RAC1, RAC2, RAC3, CDC42, RAB1A, RAB1B, RAB2, RAB3A, RAB3B, RAB3C, RAB3D, RAB4A, RAB4B, RAB5A, RAB5B, RAB5C, RAB6A, RAB6B, RAB6C, RAB7A, RAB7B, RAB7L1, RAB8A, RAB8B, RAB9, RAB9B, RABL2A, RABL2B, RABL4, RAB10, RAB11A, RAB11B, RAB12, RAB13, RAB14, RAB15, RAB17, RAB18, RAB19, RAB20, RAB21, RAB22A, RAB23, RAB24, RAB25, RAB26, RAB27A, RAB27B, RAB28, RAB2B, RAB30, RAB31, RAB32, RAB33A, RAB33B, RAB34, RAB35, RAB36, RAB37, RAB38, RAB39, RAB39B, RAB40A, RAB40AL, RAB40B, RAB40C, RAB41, RAB42, RAB43, RAP1A, RAP1B, RAP2A, RAP2B, RAP2C, ARF1, ARF3, ARF4, ARF5, ARF6, ARL1, ARL2, ARL3, ARL4, ARL5, ARL5C, ARL6, ARL7, ARL8, ARL9, ARL10A, ARL10B, ARL10C, ARL11, ARL13A, ARL13B, ARL14, ARL15, ARL16, ARL17, TRIM23, ARL4D, ARFRP1, ARL13B, RAN, RHEB, RHEBL1, RRAD, GEM, REM, REM2, RIT1, RIT2, RHOT1 или RHOT2. Согласно некоторым вариантам осуществления белок-мишень представляет собой белок, активирующий ГТФазу, такой как NF1, IQGAP1, PLEXIN-B1, RASAL1, RASAL2, ARHGAP5, ARHGAP8, ARHGAP12, ARHGAP22, ARHGAP25, BCR, DLC1, DLC2, DLC3, GRAF, RALBP1, RAP1GAP, SIPA1, TSC2, AGAP2, ASAP1 или ASAP3. Согласно некоторым вариантам осуществления белок-мишень представляет собой фактор обмена гуаниновых нуклеотидов, такой как CNRASGEF, RASGEF1A, RASGRF2, RASGRP1, RASGRP4, SOS1, RALGDS, RGL1, RGL2, RGR, ARHGEF10, ASEF/ARHGEF4, ASEF2, DBS, ECT2, GEF-H1, LARG, NET1, OBSCURIN, P-REX1, P-REX2, PDZ-RHOGEF, TEM4, TIAM1, TRIO, VAV1, VAV2, VAV3, DOCK1, DOCK2, DOCK3, DOCK4, DOCK8, DOCK10, C3G, BIG2/ARFGEF2, EFA6, FBX8 или GEP100. Согласно определенным вариантам осуществления белок-мишень представляет собой белок с доменом, отвечающим за белок-белковые взаимодействия, таким как ARM; BAR; BEACH; BH; BIR; BRCT; BROMO; BTB; C1; C2; CARD; CC; CALM; CH; CHROMO; CUE; DEATH; DED; DEP; DH; EF-рука; EH; ENTH; EVH1; F-box; FERM; FF; FH2; FHA; FYVE; GAT; GEL; GLUE; GRAM; GRIP; GYF; HEAT; HECT; IQ; LRR; MBT; MH1; MH2; MIU; NZF; PAS; PB1; PDZ; PH; POLO-Box; PBT; PUF; PWWP; PX; RGS; RING; SAM; SC; SH2; SH3; SOCS; SPRY; START; SWIRM; TIR; TPR; TRAF; SNARE; TUBBY; TUDOR; UBA; UEV; UIM; VHL; VHS; WD40; WW; SH2; SH3; TRAF; бромодомен или TPR. Согласно некоторым вариантам осуществления белок-мишень представляет собой белок теплового шока, такой как Hsp20, Hsp27, Hsp70, Hsp84, альфа-B-кристаллин, TRAP-1, hsf1 или Hsp90. Согласно определенным вариантам осуществления белок-мишень представляет собой ионный канал, такой как Cav2.2, Cav3.2, IKACh, Kv1.5, TRPA1, NAv1.7, Nav1.8, Nav1.9, P2X3 или P2X4. Согласно некоторым вариантам осуществления белок-мишень представляет собой белок, имеющий структуру в виде суперспирали, такой как геминин, SPAG4, VAV1, MAD1, ROCK1, RNF31, NEDP,1 HCCM, EEA1, виментин, ATF4, Nemo, SNAP25, синтаксин 1a, FYCO1 или CEP250. Согласно определенным вариантам осуществления белок-мишень представляет собой киназу, такую как ABL, ALK, AXL, BTK, EGFR, FMS, FAK, FGFR1, 2, 3, 4, FLT3, HER2/ErbB2, HER3/ErbB3, HER4/ErbB4, IGF1R, INSR, JAK1, JAK2, JAK3, KIT, MET, PDGFRA, PDGFRB, RET RON, ROR1, ROR2, ROS, SRC, SYK, TIE1, TIE2, TRKA, TRKB, KDR, AKT1, AKT2, AKT3, PDK1, PKC, RHO, ROCK1, RSK1, RKS2, RKS3, ATM, ATR, CDK1, CDK2, CDK3, CDK4, CDK5, CDK6, CDK7, CDK8, CDK9, CDK10, ERK1, ERK2, ERK3, ERK4, GSK3A, GSK3B, JNK1, JNK2, JNK3, AurA, ARuB, PLK1, PLK2, PLK3, PLK4, IKK, KIN1, cRaf, PKN3, c-Src, Fak, PyK2 или AMPK. Согласно некоторым вариантам осуществления белок-мишень представляет собой фосфатазу, такую как WTP1, SHP2, SHP1, PRL-3, PTP1B или STEP. Согласно определенным вариантам осуществления белок-мишень представляет собой убиквитин-лигазу, такую как BMI-1, MDM2, NEDD4-1, бета-TRCP, SKP2, E6AP или APC/C. Согласно некоторым вариантам осуществления белок-мишень представляет собой модификатор/реконструктор хроматина, такой как модификатор/реконструктор хроматина, кодируемый геном BRG1, BRM, ATRX, PRDM3, ASH1L, CBP,

KAT6A, KAT6B, MLL, NSD1, SETD2, EP300, KAT2A или CREBBP. Согласно некоторым вариантам осуществления белок-мишень представляет собой фактор транскрипции, такой как фактор транскрипции, кодируемый геном EHF, ELF1, ELF3, ELF4, ELF5, ELK1, ELK3, ELK4, ERF, ERG, ETS1, ETV1, ETV2, ETV3, ETV4, ETV5, ETV6, FEV, FLI1, GAVPA, SPDEF, SPI1, SPIC, SPIB, E2F1, E2F2, E2F3, E2F4, E2F7, E2F8, ARNTL, BHLHA15, BHLHB2, BHLBHB3, BHLHE22, BHLHE23, BHLHE41, CLOCK, FIGLA, HAS5, HES7, HEY1, HEY2, ID4, MAX, MESP1, MLX, MLXIPL, MNT, MSC, MYF6, NEUROD2, NEUROG2, NHLH1, OLIG1, OLIG2, OLIG3, SREBF2, TCF3, TCF4, TFAP4, TFE3, TFEB, TFEC, USF1, ARF4, ATF7, BATF3, CEBPB, CEBPD, CEBPG, CREB3, CREB3L1, DBP, HLF, JDP2, MAFF, MAFG, MAFK, NRL, NFE2, NFIL3, TEF, XBP1, PROX1, TEAD1, TEAD3, TEAD4, ONECUT3, ALX3, ALX4, ARX, BARHL2, BARX, BSX, CART1, CDX1, CDX2, DLX1, DLX2, DLX3, DLX4, DLX5, DLX6, DMBX1, DPRX, DRGX, DUXA, EMX1, EMX2, EN1, EN2, ESX1, EVX1, EVX2, GBX1, GBX2, GSC, GSC2, GSX1, GSX2, HESX1, HMX1, HMX2, HMX3, HNF1A, HNF1B, HOMEZ, HOXA1, HOXA10, HOXA13, HOXA2, HOXAB13, HOXB2, HOXB3, HOXB5, HOXC10, HOXC11, HOXC12, HOXC13, HOXD11, HOXD12, HOXD13, HOXD8, IRX2, IRX5, ISL2, ISX, LBX2, LHX2, LHX6, LHX9, LMX1A, LMX1B, MEIS1, MEIS2, MEIS3, MEOX1, MEOX2, MIXL1, MNX1, MSX1, MSX2, NKX2-3, NKX2-8, NKX3-1, NKX3-2, NKX6-1, NKX6-2, NOTO, ONECUT1, ONECUT2, OTX1, OTX2, PDX1, PHOX2A, PHOX2B, PITX1, PITX3, PKNOX1, PROP1, PRRX1, PRRX2, RAX, RAXL1, RHOXF1, SHOX, SHOX2, TGIF1, TGIF2, TGIF2LX, UNCX, VAX1, VAX2, VENTX, VSX1, VSX2, CUX1, CUX2, POU1F1, POU2F1, POU2F2, POU2F3, POU3F1, POU3F2, POU3F3, POU3F4, POU4F1, POU4F2, POU4F3, POU5F1P1, POU6F2, RFX2, RFX3, RFX4, RFX5, TFAP2A, TFAP2B, TFAP2C, GRHL1, TFCP2, NFIA, NFIB, NFIX, GCM1, GCM2, HSF1, HSF2, HSF4, HSFY2, EBF1, IRF3, IRF4, IRF5, IRF7, IRF8, IRF9, MEF2A, MEF2B, MEF2D, SRF, NRF1, CPEB1, GMEB2, MYBL1, MYBL2, SMAD3, CENPB, PAX1, PAX2, PAX9, PAX3, PAX4, PAX5, PAX6, PAX7, BCL6B, EGR1, EGR2, EGR3, EGR4, GLIS1, GLIS2, GLI2, GLIS3, HIC2, HINFP1, KLF13, KLF14, KLF16, MTF1, PRDM1, PRDM4, SCRT1, SCRT2, SNAI2, SP1, SP3, SP4, SP8, YY1, YY2, ZBED1, ZBTB7A, ZBTB7B, ZBTB7C, ZIC1, ZIC3, ZIC4, ZNF143, ZNF232, ZNF238, ZNF282, ZNF306, ZNF410, ZNF435, ZBTB49, ZNF524, ZNF713, ZNF740, ZNF75A, ZNF784, ZSCAN4, CTCF, LEF1, SOX10, SOX14, SOX15, SOX18, SOX2, SOX21, SOX4, SOX7, SOX8, SOX9, SRY, TCF7L1, FOXO3, FOXB1, FOXC1, FOXC2, FOXD2, FOXD3, FOXG1, FOXI1, FOXJ2, FOXJ3, FOXK1, FOXL1, FOXO1, FOXO4, FOXO6, FOXP3, EOMES, MGA, NFAT5, NFATC1, NFKB1, NFKB2, TP63, RUNX2, RUNX3, T, TBR1, TBX1, TBX15, TBX19, TBX2, TBX20, TBX21, TBX4, TBX5, AR, ESR1, ESRRA, ESRRB, ESRRG, HNF4A, NR2C2, NR2E1, NR2F1, NR2F6, NR3C1, NR3C2, NR4A2, RARA, RARB, RARG, RORA, RXRA, RXRB, RXRG, THRA, THRB, VDR, GATA3, GATA4 или GATA5; или C-myc, Max, Stat3, андрогеновый рецептор, C-Jun, C-Fox, N-Myc, L-Myc, MITF, Hif-1-альфа, Hif-2-альфа, Bcl6, E2F1, NF-каппа B, Stat5 или ER (коакт.). Согласно определенным вариантам осуществления белок-мишень представляет собой TrkA, P2Y14, mPEGS, ASK1, ALK, Bcl-2, BCL-XL, mSIN1, RORγt, IL17RA, eIF4E, TLR7 R, PCSK9, IgE R, CD40, CD40L, Shn-3, TNFR1, TNFR2, IL31RA, OSMR, IL-12 бета 1, 2, Tau, FASN, KCTD 6, KCTD 9, Raptor, Rictor, RALGAPA, RALGAPB, члены семейства аннексина, BCOR, NCOR, бета-катенин, AAC 11, PLD1, PLD2, Frizzled 7, RaLP, MLL-1, Myb, Ezh2, RhoGD12, EGFR, CTLA4R, GCGC (коакт), адипонектин R2, GPR 81, IMPDH2, IL-4R, IL-13R, IL-1R, IL-2-R, IL-6R, IL-22R, TNF-R, TLR4, Nrlp3 или OTR.

Способы виртуального скрининга.

Сбор данных и получение статистики.

Согласно некоторым вариантам осуществления стадия способов виртуального скрининга по настоящему изобретению включает сбор данных, полученных в результате эксперимента по отбору в отношении ДНК-кодируемой библиотеки (например, эксперимент на основе аффинности), в отношении белка-мишени. Данные по отбору считываются в виде последовательностей ДНК, которые затем объединяются в статистические данные, например, подсчет последовательностей. Объединение в статистические данные основывается на группировании совокупно кодируемых соединений, например, предполагаемой химической структуры, кодируемой ДНК (уровень экземпляра) или частичной подструктуры такого кодируемого химического вещества (уровень моно-, ди- или трисинтона). Определение того, связывается ли соединение или частичное соединение с мишенью (связующая молекула), выполняют с применением граничных значений для статистики, полученной по результатам секвенирования, из одного или более условий отбора. От миллионов до десятков (или даже сотен) миллионов последовательностей применяют на условие отбора с целью сбора значимых статистических значений, отражающих истинное лежащее в основе связывание малая молекула/белок.

Машинное обучение.

Способы машинного обучения известны в данной области, например, неограничивающие способы машинного обучения включают подходы на основе применения наивного байесовского классификатора, Random Forest, дерево принятия решений, метод опорных векторов, подходы на основе применения нейронных сетей и глубокое обучение.

Согласно некоторым вариантам осуществления каждую единицу данных из стадии сбора данных применяют для обучения алгоритма машинного обучения. Каждая единица данных включает информацию, полученную на основе молекулярной структуры соединения (полного или частичного) из ДНК-

кодируемой библиотеки и ассоциированных статистических данных по результатам одного или более экспериментов по отбору. Структура применяется для получения числовых входных данных (вычисленные химические свойства, например, молекулярная масса, cLogP) и двоичных строк (например, химические фингерпринты, которые соответствуют атомам, группам атомов и связности в пределах структуры). Такие вычисленные данные о молекулах применяют в качестве входных столбцов для обучения алгоритма машинного обучения и выполнения ним предсказания. Согласно некоторым вариантам осуществления модель построена так, что требуются только те входные данные, которые получены непосредственно на основе структуры молекулы. Согласно некоторым вариантам осуществления с помощью любой структуры, для которой могут быть вычислены такие фингерпринты и свойства, можно получить предсказание.

Согласно некоторым вариантам осуществления дополнительные структурные производные соединения (например, основной анализ, при котором удаляются боковые цепей) можно применять для получения дополнительных расчетов фингерпринтов и свойств, или альтернативных структурных фингерпринтов, применяемых для обучения и предсказания.

Согласно некоторым вариантам осуществления данные, полученные в результате одного или более отборов в отношении ДНК-кодируемой библиотеки, применяют для оценки того, считается ли молекула примером связующей молекулы (положительный результат), молекулы, не являющейся связующей (отрицательный результат) или не специфической связующей молекулы (отрицательный результат). Хотя оценка (положительная или отрицательная) основывается на поведении кодируемых молекул по меньшей мере в одной выборке из ДНК-кодируемой библиотеки, для оценки положительных и отрицательных классификаций, применяемых для обучения, можно применять дополнительную информацию из других источников. Следует отметить, что структура молекул, о которых известно, что они были синтезированы в библиотеке, но не характеризуются какими-либо подсчетами в результате секвенирования, рассматривается в качестве отрицательных примеров при обучении. Согласно некоторым вариантам осуществления в массивы данных включены положительные контроли. Например, могут быть включены данные по связывающим взаимодействиям для соединений с известными показателями аффинности связывания (например, известных ингибиторов или естественных лигандов) в отношении белка-мишени.

Согласно одному варианту осуществления оценка связывания для входной молекулы осуществляется посредством выявления статистически значимого обогащения (увеличенный подсчет последовательностей) в выборке, содержащей белок-мишень. Обогащение в контрольном условии, где белок-мишень не включен, также применяют для оценки специфичности связывания. Такое условие будет обычно включать смолу, применяемую для захвата белка в ходе отбора, но без добавления белка. Дополнительную информацию можно применять при определении того, что конкретная молекула или частичная молекула помечена как положительная, например, обогащение или отсутствие обогащения при дополнительных условиях или при отборе в отношении родственных белков. Также можно применять информацию, полученную в результате отбора в отношении ряда нецелевых белков, например, подсчет общего количества белков, в отношении которых было показано, что данная молекула или частичная молекула характеризуется обогащением в выборке. Например, выявление обогащения данной молекулы в отношении нескольких дополнительных мишеней в базе данных может обуславливать отрицательное определение ввиду отсутствия специфичности.

Представления молекул.

Согласно некоторым вариантам осуществления настоящего изобретения представления молекул применяют для получения расчетов в отношении предполагаемого связывания. Представления молекул включают, например, топологические представления, электростатические представления, геометрические представления или квантово-химические представления. Топологические представления могут основываться на атомах, элементах топологии или функциональных группах и их связности (например, фингерпринты, таблицы связности, молекулярная связность и/или представления в виде молекулярных графов). Электростатические представления включают, например, электронные свойства поверхностей. Геометрическими представлениями являются, например, фармакофоры, фармакофорные фингерпринты, фингерпринты на основе формы и/или молекулярные координаты в трехмерном пространстве на основе атомов, элементов топологии или функциональных групп. Согласно некоторым вариантам осуществления применяют квантово-химические представления. Согласно некоторым вариантам осуществления электронными представлениями молекул являются химические фингерпринты.

Согласно некоторым вариантам осуществления стадия способов виртуального скрининга по настоящему изобретению включает получение химических фингерпринтов как для соединений, для которых были получены данные по связывающим взаимодействиям, так и для кандидатных соединений. Химические фингерпринты могут быть получены с применением любого известного в данной области способа, например, ECFP6, FCFP6, ECFP4, MACCS или способов получения фингерпринтов Моргана/кольцевых фингерпринтов. Химические фингерпринты затем анализируют для идентификации паттернов, например, идентификации структурных признаков, которые обуславливают повышенную или пониженную способность к связыванию с белком-мишенью. Информация, полученная в результате сравнения химических фингерпринтов для большого числа соединений, например, по меньшей мере

250000 молекул, можно применять для увеличения точности в отношении полученных предполагаемых связывающих взаимодействий, в отличие от сравнения химических фингерпринтов для меньшего числа соединений, например, до 100000 соединений. Согласно некоторым вариантам осуществления химические фингерпринты применяют в этом способе в качестве первичной информации для машинного обучения.

Например, пример обучающего набора входных данных для 8-битного фингерпринта может включать следующее.

	Биты фингерпринта								Столбец для обучения
ID	1	2	3	4	5	6	7	8	Связывает белок?
Соединение 1	1	0	0	1	1	0	1	1	T
Соединение 2	1	1	0	0	0	0	1	1	F
Соединение 3	1	1	0	1	1	0	0	0	F
Соединение 4	0	0	1	1	1	0	1	0	F
Соединение 5	1	1	1	1	1	0	0	1	T
Соединение 6	1	0	0	0	0	0	0	1	F

Фингерпринт является представлением химических объектов. Машинное обучение выполняется путем задания строк для обучения, т.е. каждого соединения со столбцами, т.е. битами фингерпринтов и столбцов для обучения, в которых указано, что этот пример является положительным или отрицательным.

Алгоритмы (RF, наивный байесовский классификатор, глубокое обучение, нейронные сети и т.д.) работают путем поиска паттернов, которые коррелируют с истинными или ложными определениями. Такие паттерны могут включать один или более битов. Они могут быть обнаружены путем прямого анализа статистики (например, наивный байесовский классификатор, Random Forest) или посредством эмпирической обратной связи от изменяющихся параметров модели (например, нейронная сеть).

Другой поход, который можно применять, заключается в добавлении столбцов вычисленных свойств (например, MW, cLogP, tPSA) в дополнение к фингерпринтам. В этом случае алгоритм машинного обучения может использовать такие дополнительные столбцы в своем статистическом анализе или в своем поиске параметров модели. Применение свойств в анализе может обеспечить повышение точности предсказаний по сравнению с предсказанием, выполненным без применения свойств.

Молекулы, в отношении которых впоследствии будет выполняться предсказание в рамках этого подхода, представлены точно так же, как и молекулы, представленные в обучающем наборе, с тем основным отличием, что столбец для обучения, показанный выше, теперь неизвестен. Модель генерирует предсказанное значение для заполнения в столбце характеристик связывания (например, столбце предсказания относительно связывания). Согласно некоторым вариантам осуществления столбец является логическим (T/F), категорийным (например, молекула, не являющаяся связующей, конкурентная связующая молекула, неконкурентная связующая молекула) или числовым (например, показатель, отражающий вероятность для связующей молекулы).

	Биты фингерпринта								Свойства			
ID	1	2	3	4	5	6	7	8	MW	cLogP	tPSA	Связывает белок?
Соединение 1	1	0	0	1	1	0	1	1	504	3,2	160	T
Соединение 2	1	1	0	0	0	0	1	1	612	5,3	94	F
Соединение 3	1	1	0	1	1	0	0	0	453	4,6	112	F
Соединение 4	0	0	1	1	1	0	1	0	476	1,7	185	F
Соединение 5	1	1	1	1	1	0	0	1	598	7,1	131	T
Соединение 6	1	0	0	0	0	0	0	1	485	3,3	87	F

Молекулы для предсказания, включая только столбцы фингерпринтов, можно применять с моделью, сгенерированной первым примером выше.

	Биты фингерпринта								
ID	1	2	3	4	5	6	7	8	Связывает белок?
Соединение 1	0	1	0	0	1	0	1	1	?
Соединение 2	1	0	1	1	0	0	0	1	?
Соединение 3	1	1	1	1	1	1	0	0	?

Ниже представлено иллюстративное предсказание с входной информацией, расширенной с включением свойств, которые можно применять с моделью, созданной вторым примером выше.

	Биты фингерпринта								Свойства			
ID	1	2	3	4	5	6	7	8	MW	cLogP	tPSA	Связывает белок?
Соединение 1	0	1	0	0	1	0	1	1	467	5,4	135	?
Соединение 2	1	0	1	1	0	0	0	1	534	1,5	173	?
Соединение 3	1	1	1	1	1	1	0	0	399	4,5	97	?

Вывод.

Согласно некоторым вариантам осуществления сгенерированные модели будут давать либо двоичную оценку, указывающую, что кандидатное соединение является положительным или отрицательным, либо оценку вероятности (например, от 0 до 1), указывающую присвоенную моделью вероятность того, что кандидатное соединение является положительным или отрицательным в отношении активности/связывания. Такое значение можно затем применять для принятия решения о продолжении или прекращении эксперимента в отношении данной молекулы (бинарный случай) или для назначения приоритетов в отношении кандидатных соединений (показатель вероятности).

Примеры

Пример 1.

Данные по отбору для растворимой эпоксидгидролазы (sEH), полученные из набора библиотек, применяли для обучения одной из нескольких моделей машинного обучения (Random Forest, наивный байесовский классификатор или нейронная сеть), а затем применяли для предсказания поведения выборки молекул из библиотек, которые не были включены в обучающий набор, в отношении одной и той же мишени. Библиотеки, применяемые в обучающем наборе, включали библиотеку линейных пептидов с 25844065 соединениями, библиотеку 3-циклических соединений пиразола с 3976320 соединениями, библиотеку 2-циклических соединений пиридина с 5079459 соединениями и библиотеку 4-циклических макроциклических соединений с 1511399304 соединениями. Библиотеки, применяемые в наборе для предсказания, включали библиотеку линейных пептидов с 221580000 соединениями, библиотеку 3-циклических соединений пиридина с 285917292 соединениями и библиотеку 2-циклических соединений бензимидазола с 1622820 соединениями.

Как показано на фиг. 1, обогащение связующих молекул наблюдали в наборе, в отношении которого выполняли предсказание. В 4 квадрантах на графике представлено предсказание относительно положительных дисинтонов с применением возрастающего числа библиотек (слева направо, сверху вниз). По оси Y представлено обогащение положительных соединений в наборе, в отношении которого выполняли предсказание, по сравнению со случайной выборкой из исходной популяции. По оси Y представлена процентная доля положительных соединений в исходном наборе, которые наблюдались в наборе, в отношении которого выполняли предсказание. Из результатов видно, что для обучающего и тестового наборов (исключенные дисинтоны не в обучающем наборе, но из одних и тех же библиотек), набор, в отношении которого выполняли предсказание, характеризовался последовательным 2-2,5 кратным обогащением в сравнении с исходной популяцией. Набор для предсказания включает дисинтоны из библиотек, не применяемых в обучении. В этом случае для возрастающего числа библиотек, применяемых в обучении, наблюдали возрастающую долю положительных соединений в популяции для предсказания по сравнению с исходной популяцией.

Пример 2.

Данные по отбору из тех же библиотек, что и в примере 1 для sEH, применяли с использованием алгоритма машинного обучения (RF, MLP, глубокое обучение) для обучения и получения модели, которую применяют для предсказания активности молекул, не обнаруживаемых в ДНК-кодируемой библиотеке. Например, вводят данные и получают модель, с помощью которой можно предсказать активность молекул, тестируемых в стандартном эксперименте в рамках высокопроизводительного скрининга (HTS) (т.е. при автоматизированном тестировании от десятков тысяч до миллиона молекул). Предсказание с помощью указанной модели применяют в качестве фильтра для получения перечня (например, сотен тысяч соединений) из начального перечня, включающего от 10000 до 100000 или более молекул. Цель состоит в том, чтобы идентифицировать молекулы в этом коротком списке так, чтобы окончательный список был значительно обогащен (от 10X до 100X) по сравнению с базовой долей активных молекул, обнаруженных в начальном наборе.

Как показано на фиг. 2, наблюдали обогащение предсказанных молекул >40X по сравнению со случайной выборкой. На фиг. 2 представлены множества выполнений предсказаний с течением времени по мере улучшения моделей предсказания. Наблюдается тенденция к возрастающему обогащению как в отношении совпадений на основе первичного HTS, так и более строго подтвержденных активных соединений в наборе, в отношении которого выполняли предсказание, по сравнению со случайной выборкой. Подтвержденные активные соединения подвергали вторичному подтверждающему биохимическому анализу, и для них была продемонстрирована активность. Из наилучшего результата видно, что полученный набор, в отношении которого выполняли предсказание, улучшался в >40 раз по сравнению со случайной выборкой молекул из исходной популяции.

Пример 3. Оптимизация предсказаний.

Для данной мишени или мишеней существует известный массив данных HTS. Условия с несколькими параметрами тестировали с целью достижения высокого уровня предсказания. По сути, высокий уровень предсказания является результатом настройки предсказания на результаты HTS. Используя HTS для подтверждения применимости, модель затем можно применять для предсказания в отношении новых соединений или существующих соединений (например, коммерчески доступных или из уже существующей частной библиотеки соединений). Эти молекулы затем можно тестировать с ожиданием большей доли активных соединений, например, более 1% или 10% активных молекул в пределах набора, в отношении которого выполняется предсказание, независимо от базовой доли активных соединений случайной выборки.

Пример 4. Оптимизация предсказаний.

Данные по отборам в отношении данной мишени, но в других условиях (например, с применением фрагментов других белков, мутантов, изоформ, с применением близкородственных мишеней, с применением известных конкурирующих малых молекул и т.д.) применяют для дополнительного уточнения определения положительных данных в обучающем наборе, применяемом для проверки модели.

Пример 5. Оптимизация предсказаний.

Данные по отборам в отношении от десятков тысяч до сотен тысяч белков-мишеней, мутантов, изоформ и т.д. применяли в качестве серий колонок с дополнительными данными с целью определения положительного или отрицательного примера для проверки модели, определяемой с помощью машинного обучения.

Другие варианты осуществления

Различные модификации и вариации описанных способа и системы по настоящему изобретению будут очевидны для специалистов в данной области без отступления от объема и сущности настоящего изобретения. Несмотря на то, что настоящее изобретение было описано в сочетании с конкретными вариантами осуществления, следует понимать, что заявленное изобретение не должно неправомерным образом ограничиваться такими конкретными вариантами осуществления. В действительности, подразумевается, что различные модификации описанных способов осуществления настоящего изобретения, очевидные для специалистов в области медицины, фармакологии или в смежных областях, включены в объем настоящего изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Реализуемый на компьютере способ для идентификации и ранжирования связывающих взаимодействий между белком-мишенью и набором кандидатных соединений, включающий стадии:

(a) обеспечения множества установленных фактов о связывающих взаимодействиях для белка-мишени в физическом вычислительном устройстве, где устройство дополнительно содержит представление набора кандидатных соединений,

причем по меньшей мере 90% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением из обучающего набора, где каждое соединение из обучающего набора содержит нуклеотидную метку, кодирующую идентичность соединения, и где дополнительно множество содержит по меньшей мере 25000 установленных фактов о связывающих взаимодействиях;

(b) обучения модели с использованием алгоритма машинного обучения и множества установленных фактов о связывающих взаимодействиях со стадии (a);

(c) применения вычислительного устройства и модели на стадии (b) для получения предполагаемых связывающих взаимодействий между белком-мишенью и набором кандидатных соединений;

где кандидатные соединения отличны от соединений из обучающего набора; и

(d) вывода перечня кандидатных соединений отображенных и ранжированных по наиболее предполагаемым связывающим взаимодействиям.

2. Способ по п.1, в котором множество установленных фактов о связывающих взаимодействиях включает по меньшей мере один миллион установленных фактов о связывающих взаимодействиях.

3. Способ по п.1, в котором множество установленных фактов о связывающих взаимодействиях включает по меньшей мере два миллиона установленных фактов о связывающих взаимодействиях.

4. Способ по п.1, в котором множество установленных фактов о связывающих взаимодействиях включает по меньшей мере пять миллионов установленных фактов о связывающих взаимодействиях.

5. Способ по п.1, в котором множество установленных фактов о связывающих взаимодействиях включает по меньшей мере десять миллионов установленных фактов о связывающих взаимодействиях.

6. Способ по п.1, в котором множество установленных фактов о связывающих взаимодействиях включает по меньшей мере двадцать пять миллионов установленных фактов о связывающих взаимодействиях.

7. Способ по любому из пп.1-6, в котором стадия (b) включает анализ соединения дисинтона.

8. Способ по любому из пп.1-6, в котором стадия (c) включает анализ соединения дисинтона.

9. Способ по любому из пп.1-8, в котором каждый установленный факт о связывающих взаимодействиях на стадии (b) представляет собой связывающее взаимодействие или отсутствие связывающего взаимодействия между целевым белком и соединением обучающего набора, определенное экспериментально в эксперименте по выбору библиотеки соединений.

10. Способ по любому из пп.1-9, в котором по меньшей мере 95% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением из обучающего набора, содержащим нуклеотидную метку, кодирующую идентичность соединения.

11. Способ по любому из пп.1-10, в котором по меньшей мере 99% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением из обучающего набора, содержащим нуклеотидную метку, кодирующую идентичность соединения.

12. Способ по любому из пп.1-11, в котором по меньшей мере 50% множества установленных фактов о связывающих взаимодействиях были определены путем приведения множества соединений из обучающего набора, содержащих нуклеотидную метку, кодирующую идентичность соединения, в контакт с белком-мишенью одновременно.

13. Способ по любому из пп.1-12, где способ дополнительно предусматривает обеспечение одного или более дополнительных множеств установленных фактов о связывающих взаимодействиях для одного или более дополнительных белков-мишеней, причем по меньшей мере 50% установленных фактов о связывающих взаимодействиях в пределах дополнительного множества представляют связывающее взаимодействие между дополнительным белком-мишенью и соединением из обучающего набора, и где дополнительный целевой белок представляет собой мутант или изоформу целевого белка.

14. Способ по п.13, в котором перечень кандидатных соединений может быть отображен и ранжирован по избирательности кандидатного соединения в отношении белка-мишени среди одного или более дополнительных белков-мишеней.

15. Способ по п.13 или 14, в котором один или более дополнительных белков-мишеней представляют собой мутант белка-мишени.

16. Способ по любому из пп.1-15, где способ дополнительно предусматривает обеспечение одного или более дополнительных множеств установленных фактов о связывающих взаимодействиях для одного или более экспериментов с использованием отрицательного контроля, причем по меньшей мере 50% установленных фактов о связывающих взаимодействиях в пределах дополнительного множества представляют эксперимент с использованием отрицательного контроля для соединения из обучающего набора с белком-мишенью.

17. Способ по любому из пп.1-16, где способ дополнительно предусматривает передачу перечня кандидатных соединений через Интернет или на устройство отображения.

18. Способ по любому из пп.1-17, в котором работа с физическим вычислительным устройством и доступ к нему осуществляется через Интернет.

19. Способ по любому из пп.1-18, в котором предполагаемые связывающие взаимодействия получают с применением сравнений химической структуры.

20. Способ по п.19, в котором для сравнения химической структуры используют представления молекул.

21. Способ по п.20, в котором представления молекул включают химические фингерпринты.

22. Способ по п.21, в котором метод химических фингерпринтов представляет собой ECFP6, FCFP6, ECFP4, MACCS или метод фингерпринтов Моргана/кольцевых фингерпринтов.

23. Способ по любому из пп.1-22, где способ дополнительно предусматривает получение показателя правдоподобия для каждого из предполагаемых связывающих взаимодействий кандидатных соединений, причем показатель правдоподобия получают с применением сравнений химической структуры кандидатного соединения и одного или более соединений из множества связывающих взаимодействий для белка-мишени.

24. Способ по п.23, в котором показатель правдоподобия генерируется с использованием метода главных компонент.

25. Способ по п.23 или 24, в котором перечень кандидатных соединений может быть отображен и ранжирован по показателю правдоподобия для предполагаемого связывающего взаимодействия для кандидатного соединения.

26. Способ по любому из пп.1-25, где способ дополнительно предусматривает обеспечение одного или более установленных фактов о свойствах для набора кандидатных соединений.

27. Способ по п.26, в котором один или более установленных фактов о свойствах включают молекулярную массу и/или clogP.

28. Способ по п.26 или 27, в котором один или более установленных фактов о свойствах используют для получения предполагаемых связывающих взаимодействий.

29. Способ по любому из пп.26-28, в котором перечень кандидатных соединений может быть отображен и ранжирован по одному или более установленным фактам о свойствах.

30. Способ по любому из пп.1-29, где способ дополнительно предусматривает (d) синтезирование одного или более кандидатных соединений из перечня кандидатных соединений.

31. Способ по п.30, где способ дополнительно предусматривает приведение одного или более синтезированных кандидатных соединений в контакт с белком-мишенью для определения одного или более экспериментальных связывающих взаимодействий.

32. Машиночитаемый носитель с хранящимися на нем выполняемыми командами для управления физическим вычислительным устройством с целью выполнения способа для идентификации и ранжирования связывающих взаимодействий между белком-мишенью и набором кандидатных соединений, включающего стадии:

(a) обеспечения множества установленных фактов о связывающих взаимодействиях для белка-мишени в физическом вычислительном устройстве, где устройство дополнительно содержит представление набора кандидатных соединений,

причем по меньшей мере 90% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением из обучающего набора, где каждое соединение из обучающего набора содержит нуклеотидную метку, кодирующую идентичность соединения, и где дополнительно множество содержит по меньшей мере 25000 установленных фактов о связывающих взаимодействиях;

(b) обучения модели с использованием алгоритма машинного обучения и множества установленных фактов о связывающих взаимодействиях со стадии (a);

(c) применения вычислительного устройства и модели на стадии (b) для получения предполагаемых связывающих взаимодействий между белком-мишенью и набором кандидатных соединений;

где кандидатные соединения отличны от соединений из обучающего набора; и

(d) вывода перечня кандидатных соединений отображенных и ранжированных по наиболее предполагаемым связывающим взаимодействиям.

33. Машиночитаемый носитель по п.32, в котором каждый установленный факт о связывающем взаимодействии на стадии (b) представляет собой связывающее взаимодействие или отсутствие связывающего взаимодействия между целевым белком и соединением из обучающего набора, определенное экспериментально в эксперименте по выбору библиотеки соединений.

34. Физическое вычислительное устройство, имеющее представление набора кандидатных соединений и запрограммированное посредством выполняемых команд для управления устройством с целью выполнения способа для идентификации и ранжирования связывающих взаимодействий между белком-мишенью и набором кандидатных соединений, включающего стадии:

(a) обеспечения множества установленных фактов о связывающих взаимодействиях для белка-мишени в физическом вычислительном устройстве, где устройство дополнительно содержит представление набора кандидатных соединений,

причем по меньшей мере 90% установленных фактов о связывающих взаимодействиях в пределах множества представляют связывающее взаимодействие между белком-мишенью и соединением из обучающего набора, где каждое соединение из обучающего набора содержит нуклеотидную метку, кодирующую идентичность соединения, и где дополнительно множество содержит по меньшей мере 25000 установленных фактов о связывающих взаимодействиях;

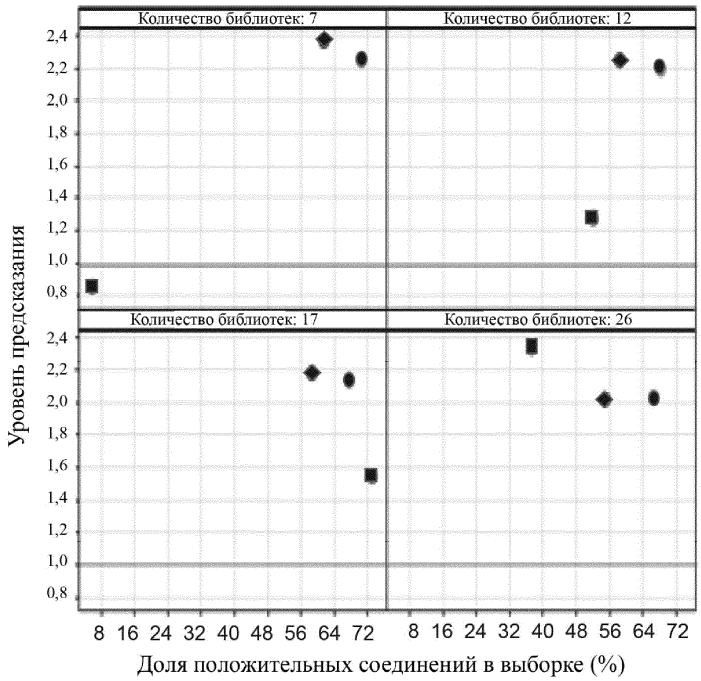
(b) обучения модели с использованием алгоритма машинного обучения и множества установленных фактов о связывающих взаимодействиях со стадии (a);

(c) применения вычислительного устройства и модели на стадии (b) для получения предполагаемых связывающих взаимодействий между белком-мишенью и набором кандидатных соединений;

где кандидатные соединения отличны от соединений из обучающего набора; и

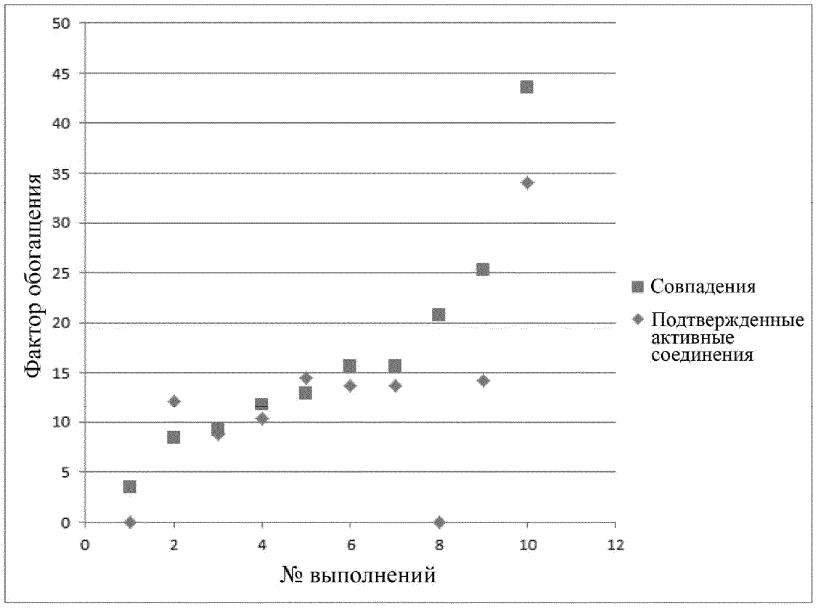
(d) вывода перечня кандидатных соединений отображенных и ранжированных по наиболее предполагаемым связывающим взаимодействиям.

35. Физическое вычислительное устройство по п.34, в котором каждый установленный факт о связывающем взаимодействии на стадии (b) представляет собой связывающее взаимодействие или отсутствие связывающего взаимодействия между целевым белком и соединением из обучающего набора, определенное экспериментально в эксперименте по выбору библиотеки соединений.



	Дисинтоны
Обучающий набор	40000
Тестовый набор	90000
Набор для предсказаний	90000

Фиг. 1



Фиг. 2