Доклад на тема

„Бази данни за биологични секвенции“

за курса

ООП с Java

Investigation of Gene sequences

Търсенето в бази данни на генетични секвенции (последователности) и изследването им революционизира биологията и създава нови научни методи за изследването на биологичните видове и има много медицински приложения.

Биологията отдавна е наблюдателна и сравнителна наука. Например сравнителната анатомия намира точки на сходство в биологични единици на два различни организма, като например кости или органи и предполага, че те могат да изпълняват подобни функции и в двата организма. Въз основа на тези съответствия откриваме, че можем да правим експерименти или да правим наблюдения върху един организъм, които можем да приложим към съответните структури в другия организъм, и да заключим, че подобни резултати ще получат при подобни условия, дори ако всъщност не правим експеримента и двата организма. Когато открием много подобни експериментални резултати върху сходни структури в различни организми, можем да изведем по-общи принципи за редица организми. Интегрирането на този вид работа на компютъра е трудно поради редица причини. Често е трудно да се получи достатъчен брой проби за статистически анализ. Може да е трудно да се изберат или моделират съответните свойства на сложна биологична форма или функция, тъй като има много голям брой параметри, които могат или не могат да имат отношение към функцията. Например това може да бъде не само форма, но и гъвкавост, състав, близост до други структури и други.

Последователността на протеин (или ДНК на ген, който кодира протеина) може да бъде моделирана като обикновен низ от букви, като всяка буква представлява определена аминокиселина или нуклеинова киселина във веригата. Докато протеинът може да се разглежда в триизмерна форма с много от параметрите, описани за анатомичните структури, изглежда, че простата линейна верига от аминокиселини съдържа голяма част от информацията, необходима, за да стане възможна окончателната форма. Така че вместо да сравняваме крайната форма или разпределението на заряда на сгънатия протеин, можем да сравним директното отчитане на гена като низ от аминокиселинни букви. Това е много лесно за моделиране на компютър и има много алгоритми за сравняване на низове и извличане на статистическа информация от тях. Такъв например е BLAST алгоритъма.

В биоинформатиката BLAST (basic local alignment search tool) е алгоритъм и програма за сравняване на информация за първична биологична последователност, като аминокиселинни последователности на протеини или нуклеотиди на ДНК и / или РНК последователности. Търсенето с BLAST дава възможност на изследователя да сравнява субект на протеинова или нуклеотидна последователност (наречена query) с библиотека или база данни от последователности и да идентифицира последователности, които приличат на последователността на заявката над определен праг. Например, след откриването на неизвестен досега ген в мишката, учен обикновено извършва BLAST търсене на човешкия геном, за да види дали хората носят подобен ген. BLAST ще идентифицира последователности в човешкия геном, които приличат на миши ген, въз основа на сходство на последователността.

Друг пример, този път в контекста на медицината, показва резултата от BLAST търсене на протеин, замесен в човешки рак на дебелото черво в база данни от протеинови последователности. Има значителни прилики с протеина от дрожди (малък организъм, участващ в производството на хляб, наред с други неща) и друг протеин от бактерия, наречена Ешерихия коли, която се намира в дебелото черво. Обърнете внимание, че нито едно от нещата, които сме знаели за човешкия протеин, не се отнасят за другите два организма. Нито е човек, нито има дебело черво, нито е болен от рак. Съществуват обаче множество експериментални резултати, описващи функциите на тези два протеина. Оказва се, че и двата са ДНК ензими за възстановяване. Това веднага ни дава представа защо генетичното увреждане на този протеин при някои хора може да ги направи по-податливи на рак.

Списък от някои бази данни с нуклеотидни секвенции:

1. GenBank

GenBank е изчерпателна публична база данни за нуклеотидни секвенции и поддържаща библиографска и биологична анотация. GenBank е изградена и разпространявана от Националния Център за Биотехнологична Информация (NCBI), подразделение на Националната Медицинска Библиотека (NLM), разположено в кампуса на Националния Здравен Институт на САЩ (NIH) в Бетесда, MD, САЩ.

1. ENA (Европейски Нуклеотиден Архив)

Европейският нуклеотиден архив (ENA) улавя и представя информация, свързана с експериментални работни потоци, които се основават на секвенцията на нуклеотидите. Типичният работен поток включва изолиране и подготовка на материал за секвениране, цикъл на машина за секвениране, в която се създават данни за секвениране и последващ тръбопровод (pipeline) за биоинформатичен анализ. ENA записва тази информация в модел на данни, който обхваща входна информация (проба, експериментална настройка, конфигурация на машината), извежда данни за машината (следи от последователности, оценки на качеството) и интерпретирана информация (сглобяване, картографиране, функционална анотация). ENA е изработена и се поддържа от Европейския Институт по Биоинформатика и е член на Международното Сътрудничество за Бази данни от Нуклеотидни Последователности (INSDC) заедно с ДНК Банката Данни на Япония (DDBJ) и GenBank.

1. DDBJ (ДНК Банката Данни на Япония)

DDBJ събира данни за нуклеотидни последователности като член на INSDC (International Nucleotide Sequence Database Collaboration) и предоставя свободно достъпни данни за нуклеотидни последователности и суперкомпютърна система, за да подпомогне изследователските дейности в науката за живота. Той се намира в Националния Институт по Генетика (NIG) в префектура Сидзуока в Япония.

*(И трите бази данни са членове на Международното Сътрудничество на Базите от данни за Нуклеотидни Последователности или INSDC. Те ежедневно обменят данните към Националния Център за Биотехнологична Информация* NCBI*. По този начин тези три бази данни съдържат едни и същи данни по всяко време. NCBI предоставя информация за бази данни за ДНК и РНК, бази данни за статии научна литература и таксономична информация, обезпечава търсене на данни за конкретни биологични видове, а също така съдържа различни програми по биоинформатика. Базите данни са достъпни чрез търсачката Entrez.)*

1. RefSeq

Базата данни с референтна последователност (RefSeq) представлява колекция от отворен достъп, анотирана и подбрана от публично достъпни нуклеотидни последователности (ДНК, РНК) и техните протеинови продукти. Тази база данни е изградена от Националния Център за Биотехнологична Информация (NCBI) и, за разлика от GenBank, предоставя само един запис за всяка естествена биологична молекула (т.е. ДНК, РНК или протеин) за основните организми, вариращи от вируси до бактерии до еукариоти.

За всеки примерен организъм RefSeq има за цел да осигури отделни и свързани записи за геномната ДНК, генните транскрипти и протеините, произтичащи от тези транскрипти. RefSeq се ограничава до основни организми, за които са налице достатъчно данни (повече от 66 000 различни „наименовани“ организми към септември 2011 г.), докато GenBank включва последователности за всеки представен организъм (приблизително 250 000 различни наименовани организми).

Други бази данни:

* Базата данни на 23andMe
* HapMap
* OMIM (Online Mendelian Inheritance in Man)
* други

*Изготвил: Виктор Василев, 2 курс Информатика, ф.н. 45676*