Курсовая работа

Формирование непротиворечивых множеств прецедентов для задачи распознавания вторичной структуры белка

Выполнил студент 317 группы
Солодкин Дмитрий Леонидович

МГУ, 2010

Содержание

Аннотация

В данной работе решена задача построение **представительной выборки** белков из исходной экспериментальной базы. Таким образом, что из множества одинаковых белков, записанных в исходной базе несколько раз с некоторыми *изменениями и неточностями*, в представительную выборку отобран ровно 1 белок.

ŋqdqȺeɎ bɆ dədɆbɎзə d ɕɔɑbɔȺ ɈqddɎqȺè

Введение

Любой белок имеет 4 уровня структуры: первичная (последовательность аминокислот), вторичная структура (последовательность локальных конформаций), третичная (трехмерная), четвертичная структура (совокупность трехмерных структур). Биологами предложена гипотеза о том, что все 4 структуры белка определяются первичной структурой.

Для проверки этой гипотезы была поставлена задача распознования вторичной структуры белка по первичной, т.е. определения того какова будет вторичная структура белка, сформированного из заданной последовательности аминокислот.

Распознавание вторичной структуры белка по его первичной структуре --- одна из фундаментальных задач вычислительной биологии и биоинформатики. Известные методы существенно ограничены как по точности распознавания, так и по вычислительной эффективности [1]. Одной из причин низкой точности является произвол в формировании обучающей выборки. Имеющиеся в свободном доступе данные из PDB (Protein Data Bank) естественно содержат неточные и противоречивые данные, так как PDB – репозиторий всех экспериментальных данных по структуре белка. До 60..70% записей являются «неточными дубликатами», содержащими несколько отличающиеся вторичные структуры для одинаковых первичных структур. Представленность белков в базе существенно неравномерна: некоторые имеют десятки и сотни «неточных дубликатов», другие представлены единственной записью. Всё это существенно затрудняет применение методов поиска закономерностей и классификации.

В связи с этим ставится задача формирования представительной базы белков таким образом, что запись о каждом белке в представительной базе будет присутствовать ровно 1 раз, причем эта будет запись о самом представительном прецеденте.

Данная задача решена в представленной работе.

При решении данной задачи встает ряд проблем: неформализуемость понятия того, что 2 записи в базе являются записями об одном и том же белке, а также неформализуемость понятия самый представительный прецедент.

Эти проблемы успешно решены при построении математической модели задчи, в работе предложены эффективные алгоритмы построения множества представительных прецедентов.

Особенностями задачи является большой объем входных данных и большая вычислительная сложность алгоритмов.

Обозначения

1. **Алфавит A** = $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$

2. **Алфавит B** = {H,S,L}

3. **Первичная структура** – строка из алфавита A.

4. **Вторичная структура** – строка из алфавита B.

5. **X – прецедент** – это объект, имеющий следующую структуру:
   - $n \in N$ - номер прецедента
   - Первичная структруа (P1)
   - Вторичная структура (P2)

   Прецедентом в исходной базе является запись о структурах некоторого белка.

6. **M** – множество прецедентов.

7. *Пусть S1,S2* – произвольные строки из алфавита A.
   Расстояние Левенштейна d(S1,S2) между двумя строками — это минимальное количество операций вставки одного символа, удаления символа и замены символа на другой, необходимых для превращения одной строки в другую.

8. *Пусть length(S)* – количество символов в строке S.
   10<length(S)<5000
   Пусть $A, B \in M$. S1, S2 – их первичные структуры.
   Расстояние между прецедентами $\delta(A, B)$ = 200*d(S1,S2)/
   (length (S1)+length(S2))

9. *Пусть* $A, B \in M$.
   A «похож» на B, если $\delta(A, B) < R_0$ **,** где R_0 **-** числовая константа.

   Гипотеза 1.

   С биологической точки зрения A «похож» на B означает, что эти прецеденты являются описанием одного и того же белка, если A не «похож» на B, то прецеденты являются описаниями разных белков.

10. *Пусть* $\forall A, B \in M$.
    n(A) – номер прецедента A.
    n(B) – номер прецедента B.
    $U_A = \{X \in M : n(X) < n(A), X$ *похож* на A$\}$
    $U_B = \{X \in M : n(X) < n(B), X$ *похож* на B$\}$
    M – непротиворечиво, если $\forall A, B \in M$ n(A) < n(B) =># $U_A <¿ U_B$.
    M – представительно, если $\forall A \in M$ # $U_A = 0$.

    Гипотеза 2.