Программа Distance-to-Structs-GGL

для перевода решения задачи ЦЛП в геномные структуры

Руководство пользователя

(http://lab6.iitp.ru/ru/chromoggl)

(авторы: В.А. Любецкий, К.Ю. Горбунов, Р.А. Гершгорин)

2019 год

Программа **distance-to-structs-ggl** является вспомогательной для задачи поиска кратчайшего преобразования одной геномной структуры в другую, где структуры содержат паралоги. Эта задача решается с помощью пакета целочисленного линейного программирования (ЦЛП), в качестве которого используется IBM CPLEX (https://www.ibm.com/analytics/cplex-optimizer). Программа distance-to-structs-ggl преобразует найденное решение задачи ЦЛП в исходный формат геномных структур с расставленной нумерацией паралогов.

Входные данные программы distance-to-structs-ggl

1) Файл, содержащий входные структуры, на основании которых программой **distance-circular-ggl** или **distance-common-ggl** был построен исходный файл для пакета ЦЛП. Например, файл **data\input_circular.txt** из контрольного примера в составе дистрибутива:

```
5
1 2
2 3
3 2
4 2
5 2
2
Struct_a; 0; 3; C2: +1.1+3.1; C3: +1.2+2.1+2.2; C4: +3.2+5.1+2.3+4.1;
Struct_b; 1; 3; C2: +4.1+2.1; C3: +1.1+2.2+1.2; C4: +4.2+5.1+5.2+3.1;
```

Здесь в первой строке указано число номеров генов в данных структурах. Номера играют роль имён генов, причем ортологичные или паралогичные гены обозначаются одним и тем же номером. Следующие пять строк указывают для каждого из этих номеров максимальное число паралогов в одной структуре. В следующей строке указывается, что всего структур две.

Две последние строки содержат описания структур *a* и *b*, соответственно. Описание структуры включает ее имя, индекс (здесь 0 и 1), число хромосом и описание каждой хромосомы. Последнее включает: тип хромосомы (линейная L или кольцевая C), число генов в ней, указание цепи ДНК (знак плюс или минус) и двойной номер гена: первый номер задан по условию, второй – начальная нумерация паралогов, которая берется произвольно.

2) Выходной файл пакета ЦЛП с решением в формате XML, содержащий, среди прочего, значения переменных *z*, задающих соответствие между паралогами данных структур. Например, файл **data\ilp_solution.sol** из контрольного примера в составе дистрибутива содержит эти переменные в строках 303...318:

```
<?xml version = "1.0" encoding="UTF-8" standalone="yes"?>
<CPLEXSolution version="1.2">
 <header
  problemName="output_circular.lp"
  ...
<variables>
  ...
 <variable name="z_1.1.1" index="60" value="1"/>
 <variable name="z_1.1.2" index="61" value="0"/>
 <variable name="z_1.2.1" index="62" value="0"/>
 <variable name="z_1.2.2" index="63" value="1"/>
 <variable name="z_2.1.1" index="64" value="-0"/>
 <variable name="z_2.1.2" index="65" value="-0"/>
 <variable name="z_2.2.1" index="66" value="0"/>
 <variable name="z_2.2.2" index="67" value="1"/>
 <variable name="z_2.3.1" index="68" value="1"/>
 <variable name="z_2.3.2" index="69" value="-0"/>
 <variable name="z_3.1.1" index="70" value="0"/>
 <variable name="z_3.2.1" index="71" value="1"/>
 <variable name="z_4.1.1" index="72" value="0.99999999999999978"/>
 <variable name="z_4.1.2" index="73" value="2.2204460492503131e-16"/>
 <variable name="z_5.1.1" index="74" value="2.2204460492503131e-16"/>
 <variable name="z_5.1.2" index="75" value="0.99999999999999978"/>
  ...
 </variables>
</CPLEXSolution>
```

Командная строка запуска программы

Утилита имеет три параметра запуска:

-s [имя файла с исходными структурами],

-i [имя файла с решением задачи ЦЛП, выданным CPLEX],

-o [имя файла, в который будут записаны структуры с оптимальной расстановкой паралогов, формат в точности соответствует формату входного файла для ChromoGGL].

Пример командной строки:

```
distance-to-structs-ggl.exe -s data/input_circular.txt -i data/ilp_solution.sol -o results/resolved_structures.txt
```

Выходные данные программы distance-to-structs-ggl

Формируемый программой выходной файл содержит те же структуры, но с итоговой нумерацией паралогов. Например, файл **results\resolved_structures.txt** из контрольного примера в составе дистрибутива имеет вид:

2;
Struct_a; 3; C2: +1.1+3.2; C3: +1.2+2.3+2.2; C4: +3.1+5.2+2.1+4.1;
Struct_b; 3; C2: +4.1+2.1; C3: +1.1+2.2+1.2; C4: +4.2+5.1+5.2+3.1;

Видно, что в нём нумерация паралогов отличается от начальной.