**Программа для расчета значений совместной встречаемости пар биологических объектов из онтологии ЭНДСистем в текстах рефератов научных публикаций (ЭНДМатч) / Program for pairwise scoring of co-occurrences between biological objects from the ontology of ANDSystem in the unstructured texts of abstracts of scientific publications (ANDMatch)**

Авторы: Иванисенко В.А., Деменков П.С., Иванисенко Т.В.

Разработанная программа позволяет проводить автоматический расчёт значений совместной встречаемости пар объектов в неструктурированных текстах научных публикаций. В основе реализованной программы лежит модификация подхода, используемого в хорошо известной системе STRING (https://string-db.org). Разработанная программа производит оценку значений совместной встречаемости объектов в неструктурированных текстах рефератов научных публикаций для всех пар объектов из уникальной онтологии ANDSystem (https://www-bionet.sscc.ru/andvisio/#!/app/about). Расчет значений со-встречаемости между всеми парами биологических единиц из онтологии (всего более 18 млн, разделенных на 13 типов: белки, гены, клеточные линии, клеточные компоненты, заболевания, лекарства, метаболиты, микроРНК, молекулярные функции, организмы, фенотипы, и биологические пути) осуществляется по следующему принципу: $\left(O(i_1 j_1), O(i_2 j_2)\right)$, где $i_1 = (1, NT)$, $j_1 = (1, N(i_1))$, $i_2 = (1, NT)$, $j_2 = (1, N(i_2))$, $NT$ – количество типов объектов (концепций) в онтологии ANDSystem, $N(i)$ – число всех объектов $i$-го типа в онтологии ANDSystem, $j_1 \neq j_2$ в случаях когда $i_1 = i_2$.

Для каждой пары объектов разработанная программа рассчитывает взвешенный показатель их совстречаемости на уровне всех текстов рефератов, где данные объекты упоминаются, а также на уровне отдельных n предложений текстового корпуса:

$C\left(O(i_1 j_1), O(i_2 j_2)\right) = \sum_{k=1}^{n} \left[\omega_s \delta_{sk}\left(O(i_1 j_1), O(i_2 j_2)\right) + \omega_a \delta_{ak}\left(O(i_1 j_1), O(i_2 j_2)\right)\right]$, где $\omega_a$=3 и $\omega_s$=0.2 являются весами для со-встречаемости в пределах одного абстракта и одного предложения, соответственно, $\delta_{sk}$ и $\delta_{ak}$ принимают значения 1 или 0 в зависимости от того со-встречаются ли объекты $O(i_1 j_1)$ и $O(i_2 j_2)$ в тексте реферата k или в одном из его предложений.

Коэффициент со-встречаемости рассчитываются программой по формуле:

$$S\left(O(i_1 j_1), O(i_2 j_2)\right) = C\left(O(i_1 j_1), O(i_2 j_2)\right)^{\alpha} \left(\frac{C\left(O(i_1 j_1), O(i_2 j_2)\right) C\left(O(i_1), O(i_2)\right)}{C\left(O(i_1 j_1), O(i_2)\right) C\left(O(i_1), O(i_2 j_2)\right)}\right)^{1-\alpha},$$

где $C\left(O(i_1 j_1), O(i_2)\right)$ является суммой числа всех объектов $O(i_2)$, имеющих тип $i_2$, в паре с объектом $O(i_1 j_1)$, $C\left(O(i_1), O(i_2 j_2)\right)$ сумма всех объектов $O(i_1)$, типа $i_1$, в паре с объектом $O(i_2 j_2)$, коэффициент нормализации $C\left(O(i_1), O(i_2)\right)$ является суммой всех пар объектов с типами $i_1$ и $i_2$, а $\alpha$ = 0.6 является значением весового коэффициента.