Использование KC-грамматики для распознавания доменов 16s

Semyon Grigorev

20 июля 2017 г.

1 Введение

Вторичная структура достаточно богата. Более того, известно, что некоторые участки обладают достаточно консервативной вторичной структурой.

Грамматика позволяет минимизировать знания о первичной структуре. Поиск структурного шаблона. Грамматика задаётся вручную, но возможен и вывод грамматики, но это тема для отдельного исследования.

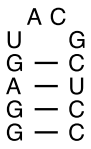
2 Грамматика

Используемая грамматика приведена в приложении A. Язык описания — YARD. Четыре терминальных символа-нуклеотида: A, U, C, G.

Грамматическая конструкция	Описание
any	Один из нуклеотидов
$any^*[n..m]$	Цепочка нуклеотидов длины от n до m
$stemN<s>$	Стем высоты N со свободной частью s (последовательность любых конструкций грамматики)
$mk\_stem<s>$	Стем произвольной высоты (от 0 до N) со свободной частью s

Таблица 1: Базовые конструкции грамматики

$stem4<any^*[3..5]>$ $mk\_stem<any^*[1..2]$ $stem2<any^*[3..4]>$ $any^*[2..5]>$



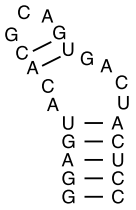


Таблица 2: Примеры описания структур

3 Эксперименты

Два эксперимента: обработка баз известных 16s, обработка полноразмерных геномов.

Домен	Стартовый нетерминал	Бактерии		Эукариоты		Археи	
		p	нр	p	нр	p	нр
Центральный	h19	17878	335	2153	3165	306	13
5'M	h3	10298	7915	50	5268	55	264

Таблица 3: Результаты анализа базы организмов

Базы размеченных полноразмерных геномов с информацией о 16s: оценить точность, полноту и т.д. (сколько из отмеченных найдено, сколько из отмеченных не найдено, сколько найдено неотмеченных). Проанализировать ложные срабатывания и пропущенных кандидатов.

Приложение

A Грамматика 16S

```
inline any: A | U | G | C
inline any_1_2: any*[1..2]
inline any_1_3: any*[1..3]
inline any_2_3: any any_1_2
inline any_2_4: any*[2..4]
inline any_3_4: any*[3..4]
inline any_3_5: any any_2_4
inline any_5_7: any any any_3_5
inline any_4_6: any any_3_5
inline any_6_8: any any_5_7
inline any_9_11: any*[9..11]
inline any_4 : any any any any

stem1<s>:
      A s U
    | U s A
    | C s G
    | G s C
    | G s U
    | U s G
    | A s G
    | G s A

stem2<s>: stem1<stem1<s>>
stem4<s>: stem2<stem2<s>>
stem6<s>: stem4<stem2<s>>
stem8<s>: stem4<stem4<s>>

mk_stem<s>:
      A mk_stem<s> U
    | U mk_stem<s> A
    | C mk_stem<s> G
    | G mk_stem<s> C
    | G mk_stem<s> U
    | U mk_stem<s> G
    | G mk_stem<s> A
    | A mk_stem<s> G
    | s
```

```
stem<s>: mk_stem<stem4<s>>
stem_2<s>: mk_stem<stem2<s>>

stem_e1<s> : stem_2<(any stem_2<s> | stem_2<s> any)> | stem<s>
stem_e2<s> : stem_2<(any stem_e1<s> any | any stem_e1<s>
             | stem_e1<s> any)> | stem<s>
stem_4: stem_2<any_4>

[<Start>]
full: middle_part_root

head_part_root: h3
middle_part_root:  h19
tail_part_root: h28 any_3_5 h44 any_3_5 h45

head_middle_folded: stem2<(any_6_8 h3 any_9_11 h19 any_1_2 h27 any_2_4)>
full_size_root: h3 any_9_11 h19 any_1_2 h27 any*[7..9] tail_part_root

(* 5'M domain *)
h3: stem_e2<(any_1_2 h4 any_1_3 h16 any_3_5
    (h17 | any*[1..6]) any*[2..5] h18 any_1_2)>
h4: stem_e1<(h5 h15 any?)>
h5: any_5_7 stem_e2<(any_1_3 h6 any_5_7
    stem_2<(any_5_7 h7 any? h11 any_1_3 h12 any?)>
    any_1_2 h13 any_1_2 h14 any_2_4)> any_3_5

h6: stem_e2<stem_e2<stem_e2<stem_e2<any_3_4>>>>
h7: stem_e2<(any_2_4 stem<(any_1_2 h8 any_4_6 h9 any_3_5 h10 any_1_2)>
             any_1_3)>
h8: stem_2<(any_3_5 stem_4 any_3_5)>
h9: stem_2<any_3_5>
h10: stem_e2<any_3_5>
h11: stem_2<(any_2_4 stem_e2<any_6_8> any_3_5)>
h12: stem<(any? stem_2<any_3_5> any_2_4)>
h13: stem<any_9_11>
h14: stem_2<any_3_5>

h15: stem_e1<(any_2_4 stem2<any_4> any?)>
h16: stem_2<(any_5_7 stem_2<any_2_4> any_4_6)>
h17: stem<(any*[6..9] stem_2<any*[7..11]> any_6_8)>
h18: stem<(any_5_7 stem<(any_4_6 stem_2<any_3_5> any_6_8)>)>

(* Central domain *)
h19: stem_2<(any_5_7 h20 any_3_5 h25 any*[9..12] h26 any_1_2)>
h20: stem_2<( any_3_4 stem_2<( any_1_2 h21 any_2_4 h22 any_2_4 )> any_3_4 )>
```

```
h21: stem_e2<( any_3_5 stem_e2<(any_3_5 stem_e1<any*[5..6]> any_2_4)> any_3_5 )>
h22: stem_e2<( any_1_3 stem<(any_3_4 h23 any*[10..12] stem_2<( any any A any )>
               any_1_2)> any_1_3 )>
h23: stem<(any_2_4 stem_2<any*[5..6]> any_5_7)>
h25: stem<(any*[7..11] stem<any*[8..10]> any*[4..7])>
h26: stem_e1<(any_1_2 stem_e2<any_4_6> any_3_5 stem_4 any_3_5 )>
h27: stem_2<(any_5_7 stem_4 any_3_5)>

(* 3'M domain *)
h28: stem_e2<(any h28_a any_2_4)>
h28_a: stem<(any_1_3 h29 any_4_6 h43 any_4_6)>
h29: stem<(h30 any_2_4 h41 any_5_7 h42 any_4_6)>
h30: stem_e1<(any_3_5 h31 any*[7..9] h32 any_2_4)>
h31: stem<any*[7..9]>
h32: stem<(any_4_6 h33 any_1_2 h34 any_3_5)>
h33: stem<(any_1_3 stem<any_4> any_1_3 stem<any_4> any_1_3)>
h34: stem_e1<(any_1_2 stem<(stem_e2<(any_2_4 h35
     any_4_6 h38 any_3_5)> any_2_4)>)>
h35: stem<(h36 any_2_3 h37 any_2_3)>
h36: stem<any_4>
h37: stem<any_5_7>
h38: stem<(any_1_2 h39 any_1_3 h40 any_4_6)>
h39: stem<(any_2_4 stem<(any_1_3 stem<any_4_6>)> any_2_4)>
h40: stem<any_4>
h41: stem<(any_4_6 stem<(any_1_3 stem<(any_2_4 stem<any_4> any_2_4)>
           any_3_5)> any_4_6)>
h42: stem<(any_3_4 stem<any*[7..9]> any_3_4)>
h43: stem<any*[7..9]>


(* 3'm domain *)
h44: stem<(any_1_3 stem<(any_2_4 stem<(any_1_3 stem<(any_3_5
     stem_e1<(any_1_3 stem<any_4>)> any_2_4)> any_1_3)> any_3_5)> any_2_3)>
h45: stem<any_4>
```