Использование КС-грамматики для распознавания 16s рРНК

Semyon Grigorev, Dmitry Kovalev

8 сентября 2017 г.

1 Введение

Задача поиска и классификации цепочек — важна. Некоторые из них используются как маркерные для обнаружения и классификации организмов. Одна из таких последовательностей — 16s rRNA.

Вторичная структура достаточно богата. Более того, известно, что некоторые участки обладают достаточно консервативной вторичной структурой. Ещё Эдди и коллеги стали использовать информацию о вторичной структуре для классификации.

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

Данный отчёт описыывает эксперимент по распознованию 16s только на основе вторичной структуры, описанной контекстно-свободной граммтикой.

2 Грамматика

Используемая грамматика приведена в приложении **??**. Язык описания — YARD. В правых частях можно использовать конструкции регулярных выражений. Четыре терминальных символа-нуклеотида: A, U, C, G.

3 Эксперименты

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

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.

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

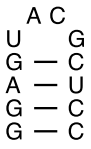
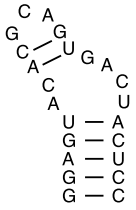
$stem4<any^*[3..5]>$	$mk\_stem<any^*[1..2]\ stem2<any^*[3..4]>\ any^*[2..5]>$
	

Таблица 2: Примеры описания структур

Домен	Стартовый нетерминал	Бактерии		Эукариоты		Археи	
		P	HP	P	HP	P	HP
Центральный	h19	17878	335	2153	3165	306	13
5'M	h3	11498	6715	64	5254	81	238

Таблица 3: Результаты анализа базы организмов

NCBI ID	Name	Expected	Covered	FP-intervals	Length(avg.)	Length SD
NC_014640.1	Achromobacter xylosoxidans A8	3	2	4261	430.9	234.8
NZ_CP009448.1	Achromobacter xylosoxidans C54	3	2	4157	446.6	241.6
NZ_CP014060.1	Achromobacter xylosoxidans strain FDAARGOS_147	3	1	4770	469.3	284.2
NZ_CP012046.1	Achromobacter xylosoxidans strain MN001	3	2	4114	457.5	270.7
NZ_LN831029.1	Achromobacter xylosoxidans genome assembly NCTC10807	3	2	4441	442.2	239.0
NZ_CP007618.1	Bacillus anthracis strain 2000031021	11	1	2384	665.6	485.8
NZ_CP012475.1	Bacillus clausii strain ENTPro	7	0	1862	453.2	241.3
NC_006582.1	Bacillus clausii KSM-K16 DNA	7	7	1744	451.9	230.0
NZ_CP010052.1	Bacillus subtilis subsp. subtilis str. 168	10	9	1610	450.4	236.3
NZ_CP016852.1	Bacillus subtilis subsp. subtilis strain 168G	10	9	1616	450.4	240.1
NZ_CP017763.1	Bacillus subtilis strain 29R7-12	10	1	1721	434.1	208.7
NZ_CP010314.1	Bacillus subtilis subsp. subtilis strain 3NA	10	9	1596	448.9	238.2
NC_020507.1	Bacillus subtilis subsp. subtilis 6051-HGW	10	9	1607	451.2	239.3
NZ_CP008698.1	Bacillus subtilis subsp. subtilis str. AG1839	10	9	1613	450.4	237.1
NZ_CP009748.1	Bacillus subtilis strain ATCC 13952	7	6	1427	427.7	224.4
NZ_CP009749.1	Bacillus subtilis strain ATCC 19217	7	6	1533	431.1	233.3
NC_011835.1	Bifidobacterium animalis subsp. lactis AD011	2	1	1134	438.1	243.9
NC_017834.1	Bifidobacterium animalis subsp. animalis ATCC 25527	4	0	1044	436.0	237.1
NC_022523.1	Bifidobacterium animalis subsp. lactis ATCC 27673	4	0	1076	427.4	226.4
NC_017866.1	Bifidobacterium animalis subsp. lactis B420	4	0	1050	429.8	236.9
NC_017214.1	Bifidobacterium animalis subsp. lactis BB-12	4	0	1033	434.4	236.3
NZ_CP009045.1	Bifidobacterium animalis subsp. lactis strain BF052	4	0	1051	430.0	236.9
NC_017867.1	Bifidobacterium animalis subsp. lactis Bi-07	4	0	1049	429.9	237.3
NC_021593.1	Bifidobacterium animalis subsp. lactis Bl12	4	0	1047	430.3	237.6
NZ_CP017037.1	Dialister pneumosintes strain F0677	5	2	570	721.2	542.6
NZ_CP008740.1	Haemophilus influenzae 2019	6	2	654	420.2	196.3
NZ_CP007470.1	Haemophilus influenzae strain 477	6	2	564	421.9	189.3
NZ_CP007472.1	Haemophilus influenzae strain 723	6	4	751	425.2	200.5
NC_007146.2	Haemophilus influenzae 86-028NP	6	2	711	428.0	198.2
NZ_AP012334.1	Scardovia inopinata JCM 12537 DNA	2	0	736	413.7	186.0
NC_022238.1	Streptococcus constellatus subsp. pharyngis C1050	4	2	784	541.9	349.8
NC_022236.1	Streptococcus constellatus subsp. pharyngis C232	4	2	767	531.3	342.6
NC_022245.1	Streptococcus constellatus subsp. pharyngis C818	4	2	765	532.5	341.9
NC_022246.1	Streptococcus intermedius B196	4	2	755	527.5	314.9
NC_022237.1	Streptococcus intermedius C270	4	2	733	524.3	308.4
NZ_CP020433.1	Streptococcus intermedius strain FDAARGOS_233	4	2	865	561.0	350.5
NC_018073.1	Streptococcus intermedius JTH08 DNA	4	2	917	503.5	305.2
NZ_AP013044.1	Tannerella forsythia 3313 DNA	2	0	1310	445.9	252.6
NC_016610.1	Tannerella forsythia 92A2	2	0	1398	442.5	246.2
NZ_AP013045.1	Tannerella forsythia KS16 DNA	2	0	1537	448.5	258.1

Таблица 4: Результаты анализа полноразмерных геномов (центральный домен)

NCBI ID	Name	Expected	Covered	FP-intervals	Length(avg.)	Length SD
NC_014640.1	Achromobacter xylosoxidans A8	3	2	530	596.8	246.4
NZ_CP009448.1	Achromobacter xylosoxidans C54	3	1	732	630.2	273.1
NZ_CP014060.1	Achromobacter xylosoxidans strain FDAARGOS_147	3	0	952	673.1	361.3
NZ_CP012046.1	Achromobacter xylosoxidans strain MN001	3	0	722	664.4	414.9
NZ_LN831029.1	Achromobacter xylosoxidans genome assembly NCTC10807	3	1	752	624.2	266.1
NZ_CP007618.1	Bacillus anthracis strain 2000031021	11	1	662	663.9	287.4
NZ_CP012475.1	Bacillus clausii strain ENTPro	7	0	101	548.6	153.9
NC_006582.1	Bacillus clausii KSM-K16 DNA	7	7	112	567.0	182.2
NZ_CP010052.1	Bacillus subtilis subsp. subtilis str. 168	10	9	85	531.1	161.0
NZ_CP016852.1	Bacillus subtilis subsp. subtilis strain 168G	10	9	85	536.3	168.8
NZ_CP017763.1	Bacillus subtilis strain 29R7-12	10	1	82	509.4	93.0
NZ_CP010314.1	Bacillus subtilis subsp. subtilis strain 3NA	10	9	80	534.8	175.0
NC_020507.1	Bacillus subtilis subsp. subtilis 6051-HGW	10	9	85	530.7	161.2
NZ_CP008698.1	Bacillus subtilis subsp. subtilis str. AG1839	10	9	82	529.3	138.6
NZ_CP009748.1	Bacillus subtilis strain ATCC 13952	7	6	63	547.5	168.1
NZ_CP009749.1	Bacillus subtilis strain ATCC 19217	7	6	65	535.9	198.6
NC_011835.1	Bifidobacterium animalis subsp. lactis AD011	2	1	139	613.9	242.0
NC_017834.1	Bifidobacterium animalis subsp. animalis ATCC 25527	4	0	101	658.5	287.7
NC_022523.1	Bifidobacterium animalis subsp. lactis ATCC 27673	4	0	110	645.3	297.3
NC_017866.1	Bifidobacterium animalis subsp. lactis B420	4	0	123	615.0	263.4
NC_017214.1	Bifidobacterium animalis subsp. lactis BB-12	4	0	115	627.9	279.6
NZ_CP009045.1	Bifidobacterium animalis subsp. lactis strain BF052	4	0	116	631.0	281.4
NC_017867.1	Bifidobacterium animalis subsp. lactis Bi-07	4	0	123	614.4	263.8
NC_021593.1	Bifidobacterium animalis subsp. lactis Bl12	4	0	117	629.5	279.2
NZ_CP017037.1	Dialister pneumosintes strain F0677	5	2	178	648.5	300.3
NZ_CP008740.1	Haemophilus influenzae 2019	6	2	25	508.9	75.1
NZ_CP007470.1	Haemophilus influenzae strain 477	6	2	34	536.1	99.0
NZ_CP007472.1	Haemophilus influenzae strain 723	6	4	41	559.5	165.5
NC_007146.2	Haemophilus influenzae 86-028NP	6	2	57	526.1	119.7
NZ_AP012334.1	Scardovia inopinata JCM 12537 DNA	2	0	38	533.8	177.3
NC_022238.1	Streptococcus constellatus subsp. pharyngis C1050	4	2	102	558.1	165.7
NC_022236.1	Streptococcus constellatus subsp. pharyngis C232	4	2	84	589.2	172.0
NC_022245.1	Streptococcus constellatus subsp. pharyngis C818	4	2	91	582.0	169.4
NC_022246.1	Streptococcus intermedius B196	4	2	82	589.1	193.6
NC_022237.1	Streptococcus intermedius C270	4	2	105	558.8	160.1
NZ_CP020433.1	Streptococcus intermedius strain FDAARGOS_233	4	2	137	572.5	204.6
NC_018073.1	Streptococcus intermedius JTH08 DNA	4	2	88	590.8	236.9
NZ_AP013044.1	Tannerella forsythia 3313 DNA	2	0	93	597.5	250.1
NC_016610.1	Tannerella forsythia 92A2	2	0	107	580.6	209.0
NZ_AP013045.1	Tannerella forsythia KS16 DNA	2	0	122	555.5	212.1

Таблица 5: Результаты анализа полноразмерных геномов (5'M домен)

Приложение

A Грамматика 16S на языке YARD, использовавшаяся в эксперименте

```
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>
```