Информационные процессы, Том 6, № 1, 2006, стр. 24–32.

Алгоритм определения белка, согласованного с заданным филогенетическим профилем

Л.А. Леонтьев, В.А. Любецкий

Институт проблем передачи информации РАН, 127994, Россия, Москва, Большой Каретный переулок, 19, e-mail: maravan@yandex.ru, lyubetsk@iitp.ru

Поступила в редколлегию 19.12.2005

Аннотация—Предложен новый алгоритм для поиска белков по заданному филогенетическому профилю, соотвествующая программа была разработана одним из авторов (Л. Леонтьевым). А также приведены результаты счета для различных наперед заданных филогенетических профилей.

1. ВВЕДЕНИЕ

Разработан алгоритм для поиска белков по заданному филогенетическому профилю. В заметке рассматривается случай, когда филогенетический профиль определяется непосредственно по двум спискам геномов. В общем случае профиль определяется сходством искомого белка с представителем из каждого генома. Итак, ищется белок, наиболее соответствующий этим двум спискам, т.е. белок гомологи которого присутствуют во всех геномах из первого списка (“плюс-список”), и в тоже время лучший гомолог в каждом геноме из второго списка (“минус-список”) имеет меньшее сходство с данным белком, чем лучший гомолог из любого генома, принадлежащего первому списку. На самом деле, ищутся несколько лучших (“субоптимальных”) белков, удовлетворяющих этому условию.

Например, так могут искаться регуляторные белки по их потенциальным сайтам связывания с ДНК или РНК, когда плюс-список состоит из геномов, содержащих хотя бы один регуляторный сайт рассматриваемого типа, а минус-список состоит из геномов, не содержащих такого сайта. Другой пример: алгоритм применим для поиска белков, кодирующих характерные признаки организма (наличие/отсутствие жгутика или фотосистем и т.д. у организма).

2. АЛГОРИТМ

Алгоритм состоит в специально организованном последовательном просмотре двух входных списков (“плюс” и “минус” списки) с вычислением минимакса. Для этого из плюс-списка выбирается геном (называемый “первым”): самый короткий по количеству белков в нем или самый длинный в этом смысле или просматриваются все геномы из этого списка в качестве “первых”.

Для каждого генома (называемого “текущий геном”) из плюс-списка, кроме первого, и каждого белка (называемого “текущий белок”) из первого генома выполняется следующая процедура.

- В текущем геноме определяется самый близкий в смысле качества выравнивания белок к текущему белку.
- Каждому текущему белку сопоставляется минимакс качеств выравниваний по плюс-списку, где максимум берется по белкам текущего генома, и минимум — по геномам из плюс-списка.

– Теперь для каждого текущего генома из минус-списка и каждого текущего белка ищем белок из текущего генома, качество выравнивания которого с текущим белком превышает минимакс. Если находим такой белок, то переходим к следующему текущему белку. Иначе переходим к следующему геному из минус-списка. В результате текущий белок со всеми соответствующими ему максимумами качеств относительно всех геномов из плюс и минус списков заносим в итоговый список. Эти максимумы и являются филогенетическим профилем каждого текущего белка, попавшего в итоговый список.
– Вычисляем тот из этих профилей, который наилучшим образом согласуется с исходным разбиением на плюс- и минус-списки. Он и соответствующий белок являются результатом работы алгоритма.

Филогенетический профиль любого белка — функция, сопоставляющая каждому геному из плюс и минус списков число, равное качеству выравнивания этого белка с ближайшим гомологом в данном геноме. Качество выравнивания двух белков определяется, например, как значение E-value, выдаваемое программой BLAST. Вместо E-value можно рассматривать, по крайней мере для плюс-списка некоторому среднему двух E-value в прямую и обратную стороны (как в случае вычисления ортолога). В нашем случае, филогенетический профиль двух списков (и приблизительный профиль искомого белка) — функция, принимающая два значения, одно из них (например, равное 0) на геномах из плюс списка, другое (равное 1) — на геномах из минус-списка. Мера сходства профилей, понимаемых как векторы, определяется здесь как косинус $\cos\varphi$ угла φ между векторами профилей.

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

Приведены результаты только для случаев, когда плюс- и минус-списки геномов организмов брались из базы данных Genome Properties (www.tigr.org/tigr-scripts/CMR2/genome_ properties), а аминокислотные последовательности, соответствующие генам, — из базы данных Genebank.

В таблицах 1-6 в первом столбце указано имя белка (protein ID в Genebank), во втором столбце указана аннотация к этому белку, в третьем — его длина в аминокислотах, в четвертом — мера сходства филогенетического профиля этого белка с филогенетическим профилем исходной пары плюс-минус-списков.

Ниже перечислены плюс и минус списки для последующих шести таблиц.

1. Плюс-список состоит из: *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Yersinia pestis*, *Brucella suis*, *Porphyromonas gingivalis*, *Clostridium tetani*, *Vibrio vulnificus*, *Enterococcus faecalis*, *Treponema denticola*, *Yersinia pestis*, *Streptococcus agalactiae*, *Burkholderia pseudomallei*, *Campylobacter jejuni*, *Staphylococcus epidermidis*.
Минус-список состоит из: *Aquifex aeolicus*, *Pyrobaculum aerophilum*, *Pyrococcus furiosus*, *Shewanella oneidensis*, *Pseudomonas putida*, *Pyrococcus abyssi*, *Thermoplasma acidophilum*, *Thermoplasma volcanium*, *Oceanobacillus iheyensis*, *Methanococcus maripaludis*, *Streptococcus thermophilus*, *Dehalococcoides ethenogenes*, *Silicibacter pomeroyi*, *Acinetobacter* sp.

2. Плюс-список состоит из: *Helicobacter pylori*, *Treponema pallidum*, *Aquifex aeolicus*, *Pseudomonas aeruginosa*, *Escherichia coli*, *Caulobacter crescentus*, *Agrobacterium tumefaciens*, *Geobacter sulfurreducens*, *Treponema denticola*, *Bacillus subtilis*, *Listeria monocytogenes*, *Methanocaldococcus jannaschii*.
Минус-список состоит из: *Streptococcus agalactiae*, *Enterococcus faecalis*, *Dehalococcoides ethenogenes*, *Staphylococcus epidermidis*, *Staphylococcus aureus*.

3. Плюс-список состоит из: *Helicobacter pylori*, *Pseudomonas putida*, *Treponema denticola*, *Vibrio vulnificus*, *Campylobacter jejuni*, *Silicibacter pomeroyi*.

Минус-список состоит из: *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Chlorobium tepidum*, *Porphyromonas gingivalis*, *Enterococcus faecalis*, *Corynebacterium efficiens*, *Dehalococcoides ethenogenes*, *Staphylococcus epidermidis*.

4. Плюс-список состоит из: *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Yersinia pestis*, *Escherichia coli*, *Brucella suis*, *Porphyromonas gingivalis*, *Clostridium tetani*, *Vibrio vulnificus*, *Enterococcus faecalis*, *Treponema denticola*, *Streptococcus agalactiae*, *Burkholderia pseudomallei*, *Campylobacter jejuni*, *Staphylococcus epidermidis*.
Минус-список состоит из: *Chlorobium tepidum*, *Agrobacterium tumefaciens*, *Pyrobaculum aerophilum*, *Pyrococcus furiosus*, *Pseudomonas putida*, *Bacteroides thetaiotaomicron*, *Staphylococcus epidermidis*, *Mannheimia succiniciproducens*, *Pyrococcus abyssi*, *Nitrosomonas europaea*, *Thermoplasma volcanium*, *Corynebacterium efficiens*, *Methanococcus maripaludis*, *Streptococcus thermophilus*, *Dehalococcoides ethenogenes*, *Silicibacter pomeroyi*, *Desulfotalea psychrophila*, *Acinetobacter*,sp, *Azoarcus* sp.

5. Плюс-список состоит из: *Helicobacter pylori*, *Neisseria meningitides*, *Chlorobium tepidum*, *Yersinia pestis*, *Escherichia coli*, *Brucella suis*, *Pseudomonas putida*, *Porphyromonas gingivalis*, *Bacteroides thetaiotaomicron*, *Vibrio vulnificus*, *Haemophilus ducreyi*, *Treponema denticola*, *Rhodopseudomonas palustris*, *Burkholderia pseudomallei*, *Borrelia garinii*, *Campylobacter jejuni*, *Dehalococcoides ethenogenes*, *Silicibacter pomeroyi*.
Минус-список состоит из: *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Clostridium tetani*, *Enterococcus faecalis*, *Lactobacillus johnsonii*, *Mycoplasma mobile*, *Mycoplasma hyopneumoniae*, *Streptococcus agalactiae*, *Corynebacterium efficiens*, *Geobacillus kaustophilus*, *Methanococcus maripaludis*, *Bacillus licheniformis*, *Streptococcus thermophilus*, *Staphylococcus epidermidis*.

6. Плюс-список состоит из: *Clostridium acetobutylicum*, *Bacillus cereus*, *Bacillus anthracis*, *Bacillus subtilis*, *Streptomyces coelicolor Bacillus halodurans*, *Clostridium perfringens*, *Oceanobacillus iheyensis*.
Минус-список состоит из: *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Chlorobium tepidum*, *Streptococcus pyogenes*, *Yersinia pestis*, *Brucella suis*, *Pseudomonas putida*, *Porphyromonas gingivalis*, *Enterococcus faecalis*, *Treponema denticola*, *Streptococcus agalactiae*, *Corynebacterium efficiens*, *Burkholderia pseudomallei*, *Campylobacter jejuni*, *Dehalococcoides ethenogenes*, *Staphylococcus epidermidis*, *Silicibacter pomeroyi*.

Авторы благодарят М.С. Гельфанда и А.В. Селиверстова за обсуждение и помощь при выполнении работы.

Работа поддержана грантом МНТЦ 2766.

Имя белка	Комментарий к описанию белка	Длина белка	Мера сходств двух профилей
AAK44233.1	Пептидил-пролил цис-транс изомераза	182	0.685
AAK47907.1	Рибосомальный белок L36	37	0.681
AAK48043.1	spoU рРНК метилаза	253	0.669
AAK46547.1	Семейство глицерат киназы	381	0.657
AAK44336.1	Рибосомальный белок L28	94	0.641
AAK45654.1	ABC транспортер	859	0.614
AAK48188.1	Неопределенный BCR, YbaB	133	0.594
AAK45284.1	Пара-аминобензоат-синтаза	547	0.580
AAK48409.1	Рибосомальный белок L34	47	0.529
AAK47393.1	Белок, связывающий ДНК HU	214	0.522
AAK46790.1	Белок, ассоциированный с IojAP	126	0.520
AAK44424.1	ABC транспортер, пермеаза	1241	0.487
AAK48405.1	Белок, содержащий R3H домен	187	0.485
AAK45135.1	Белок холодового шока	135	0.484
AAK44559.1	Деоксицитидин трифосфат деаминаза	190	0.465
AAK44964.1	Рибосомальный белок L22	197	0.462
AAK45207.1	Фосфатный ABC траспортер	276	0.447
AAK44976.1	Рибосомальный белок S14	61	0.439
AAK47421.1	Глутамил-тРНК субъединица C аминотрансферазы	99	0.419
AAK46854.1	ABC транспортер, белок, связывающий АТФ	558	0.394
AAK46833.1	Хомоцистеин S-метилтрансфереаза	302	0.391
AAK45308.1	Регулятор связи с ДНК KdpE	226	0.372
AAK47806.1	Оксидоредуктаза, связаная с FMN	396	0.368
AAK44817.1	Хем-тиолатный протеин P450	472	0.351
AAK47102.1	Оксидоредуктаза	468	0.316
AAK47726.1	Домен, связаный с Fe-S метаболизмом	143	0.316
AAK46029.1	Предположительный белок MT1757	386	0.315
AAK45872.1	Фумарат редуктаза, 15 kDa белок	126	0.313
AAK45257.1	Регулятор связи с ДНК	230	0.311
AAK47292.1	Неопределенный белок UPF0102	141	0.309

Таблица 1. Первые 30 белков, отобранных нашим алгоритмом по плюс-минус-спискам (1) и характерных для организмов, являющимся патогенными для животных.

Имя белка	Комментарий к описанию белка	Длина белка	Мера сходств двух профилей
AAG03801.1	Метилтрансфераза PilK	291	0.780
AAG07039.1	Фосфатидат цитидилтрансфераза	271	0.769
AAG07324.1	Вероятный компонент ABC тауринового транспортера	263	0.752
AAG06642.1	Вероятный компонент ABC транспортера	331	0.733
AAG06995.1	Белок PotA, траспортер полиаминов.	363	0.733
AAG08955.1	Рибосомальный белок L34	44	0.720
AAG07630.1	Рибосомальный белок L36	38	0.720
AAG04423.1	SEC-C мотив	66	0.718
AAG08921.1	Прокариотический dksA/traR C4 цинковый палец	134	0.658
AAG07302.1	Белок биосинтеза B1	185	0.628
AAG07866.1	Белок, похожий на Maf.	201	0.628
AAG07429.1	Малая субъединица эксодеоксирибонуклеазы	80	0.620
AAG07278.1	Вероятный компонент ABC транспортера	387	0.620
AAG03527.1	Вероятная пермеаза ABC транспортера	365	0.620
AAG05121.1	Суперсемейство, схожее с транглютаминазой	266	0.616
AAG08114.1	2-амино-4-гидрокси-6-гидроксиметилдегидроптеридин пирофосфокиназа	162	0.613
AAG04858.1	Неизвестный белок YfiH семейства КОГ1496	240	0.609
AAG06850.1	Вероятный регуляторный гибрид	919	0.592
AAG08270.1	Двухкомпонентный регулятор ответной реакции	229	0.585
AAG04294.1	RsmA, регулятор вторичных метаболитов	61	0.577
AAG06033.1	NADH дегидрогиназа I цепочка J	166	0.573
AAG05275.1	ДНК полимераза II	787	0.564
AAG04865.1	Белок CcmB	223	0.554
AAG07931.1	Гипотетический белок	242	0.554
AAG06772.1	Вероятный компонент ABC фосфатного транспортера	278	0.554
AAG08203.1	Долицил-фосфат-манноза манносилтрасфереаза	478	0.554
AAG04741.1	Суперсмейство фацилитатор	403	0.554
AAG04840.1	Домен с неизвестной функцией (DUF477)	447	0.550
AAG04826.1	Вероятный двухкомпонентный регулятор	229	0.538
AAG06592.1	Вероятный двухкомпонентный регулятор	225	0.531

Таблица 2. Первые 30 белков, отобранных нашим алгоритмом по плюс-минус-спискам (2) и характерных для организмов, обладающих свойством хемотаксиса.

Имя белка	Комментарий к описанию белка	Длина белка	Мера сходств двух профилей
AAN69932.1	Белок синтеза жгутика FliQ	89	0.911
AAN66924.1	Аминокислотный траспортер	254	0.707
AAN66643.1	ABC транспортер сахаров, АТФ-связывающий	384	0.679
AAN68382.1	ABC транспортер глицина бетаина/L-пролина	698	0.679
AAN66429.1	Белок системы выделения, пермеаза	722	0.672
AAN68581.1	Диацилглицерол киназа	117	0.667
AAN68067.1	ABC транспортер рибозы	524	0.655
AAN65872.1	ABC транспортер сульфоната, субъединица SsuB	270	0.654
AAN69732.1	ABC транспортер белков	534	0.654
AAN67930.1	Вероятный ABC транспортер	228	0.654
AAN70413.1	Вероятный ABC транспортер разветвленных аминокислот	285	0.654
AAN68367.1	Вероятный ABC транспортер рибозы	512	0.654
AAN70502.1	Вероятный белок системы выделения, пермеаза	602	0.654
AAN69797.1	Система экспорта пиовердина, пермеаза	552	0.654
AAN67853.1	Вероятный транспортер, пермеаза	626	0.654
AAN68028.1	Вероятный ABC транспортер железа	262	0.654
AAN70772.1	Вероятный ABC транспортер, пермеаза	906	0.654
AAN70702.1	ABC транспортер	352	0.654
AAN68200.1	ABC транспортер сидерофора	258	0.654
AAN69331.1	ABC транспортер	221	0.654
AAN68375.1	Вероятный ABC транспортер разветвленных аминокислот	277	0.654
AAN68949.1	ABC транспортер никеля	256	0.654
AAN67399.1	Система экспорта липополисахаридов	405	0.654
AAN68356.1	Вероятный ABC транспортер разветвленных аминокислот	266	0.654
AAN65775.1	Вероятный ABC транспортер	263	0.654
AAN65914.1	ABC транспортер аминокислот	257	0.654
AAN66578.1	Вероятный ABC транспортер	241	0.654
AAN68361.1	Вероятный ABC транспортер разветвленных аминокислот	257	0.654
AAN70744.1	ABC транспортер путресцина	380	0.654
AAN65752.1	ABC транспортер цинка, белок ZnuC	257	0.654

Таблица 3. Первые 30 белков, отобранных нашим алгоритмом по плюс-минус-спискам (3) и характерных для организмов, имеющих жгутик.

Имя белка	Комментарий к описанию белка	Длина белка	Мера сходств двух профилей
AAN82013.1	Вероятный регулятор ответной реакции	240	0.728
AAN79496.1	Транспортный белок, cydD	588	0.728
AAN83905.1	Транскрипционный реглуяторный белок creB	229	0.712
AAN83058.1	50S Рибосомальный белок L34	46	0.682
AAN79132.1	Транскрипционный регуляторный белок cusR	227	0.666
AAN83715.1	50S Рибосомальный белок L9	149	0.634
AAN79689.1	АТФ связывающий белок	704	0.586
AAN78986.1	Фосфатный регулон. Регуляторный белок phoB	229	0.565
AAN82395.1	Вероятный ABC-транспортер, yrbF	269	0.509
AAN79350.1	Вероятный ABC-транспортер, ybhF	583	0.441
AAN80846.1	Вероятный транскрипиционный регуляторный yedW	260	0.417
AAN80127.1	Пептидил-тРНК гидролаза	194	0.416
AAN82906.1	Вероятный белок yicG	223	0.375
AAN79495.1	Транспортный белок, cydC	573	0.363
AAN82445.1	Формообразующий белок mreC	367	0.362
AAN79161.1	Вероятный белок yjiX	65	0.356
AAN79198.1	Формообразующий белок rodA	370	0.352
AAN79590.1	Белок холодного шока cspH	70	0.339
AAN83848.1	Вероятный белок yjiX	67	0.330
AAN80880.1	Вероятный ABC транспортер	609	0.327
AAN81349.1	Активатор Д-серин деаминазы	317	0.318
AAN78649.1	Белок yadF	220	0.317
AAN83755.1	Вероятный белок protein	136	0.308
AAN81387.1	Регуляторный белок ксантозинового оперона	299	0.305
AAN82309.1	Вероятный белок yhaL	56	0.304
AAN81531.1	Вероятный белок yfhC	178	0.300
AAN81536.1	Holo-[acyl-carrier] синтаза	126	0.296
AAN78650.1	Вероятный ABC-транспортер, yadG	308	0.292
AAN79118.1	Пептидил-пролил цис-транс изомераза B	164	0.277
AAN79030.1	Вероятный белок	433	0.273

Таблица 4. Первые 30 белков, отобранных нашим алгоритмом по плюс-минус-спискам (4) и характерных для организмов, являющихся патогенными для человека.

Имя белка	Комментарий к описанию белка	Длина белка	Мера сходств двух профилей
AAN82012.1	Гистидин киназа, ДНК гираза	490	0.526
AAN83058.1	50S Рибосомальный белок L34	46	0.526
AAN78892.1	Вероятный транскрипционный регулятор, схожий с LysR	329	0.503
AAN78842.1	Вероятный цитоплазматический мембранный белок	715	0.494
AAN80557.1	Вероятный белок ydiD	566	0.470
AAN80906.1	Транскрипционный регулятор cbl	316	0.453
AAN79132.1	Транскрипционный регуляторный белок cusR	227	0.436
AAN78649.1	Белок yadF	220	0.424
AAN82013.1	Вероятный регулятор ответной реакции	240	0.390
AAN83081.1	Транспорт фостфата, белок pstB	257	0.360
AAN78544.1	Вероятная кротобетаин/карнитин-CoA лигаза	522	0.351
AAN80839.1	ДНК-цитозиновая метилтрансфераза	472	0.328
AAN82618.1	Белок envZ	450	0.310
AAN78650.1	Вероятный ABC транспортер yadG	308	0.308
AAN82462.1	Вероятный белок yhdT	86	0.308
AAN79190.1	Sec-независимый белок tatE	67	0.308
AAN80098.1	Дисульфид связывающий белок	178	0.308
AAN80524.1	Гомолог белка PhsC	261	0.300
AAN82952.1	Вероятный белок	145	0.273
AAN79030.1	Вероятный белок	433	0.273
AAN82906.1	Вероятный белок yicG	223	0.263
AAN81986.1	Вероятный белок yggT	188	0.243
AAN81244.1	Вероятный транскрипционный регулятор yfaX	260	0.242
AAN79161.1	Вероятный белок yjiX	65	0.236
AAN82616.1	Вероятный белок	50	0.228
AAN83376.1	Биосинтез тиамина, возможный донор серы	66	0.226
AAN82525.1	50S Рибосомальный белок L23	100	0.225
AAN79689.1	АТФ связывающий белок	704	0.225
AAN79258.1	Вероятный предшественник ybfA	68	0.218
AAN79311.1	Вероятный предшественник ybhT	51	0.218

Таблица 5. Первые 30 белков, отобранных нашим алгоритмом по плюс-минус-спискам (5) и характерных для организмов, имеюших внешнюю мембрану.

Имя белка	Комментарий к описанию белка	Длина белка	Мера сходств двух профилей
CAB94648.1	Хитиназа	244	0.690
CAB52067.1	Вероятный транспортер разветвленных аминокислот	309	0.661
CAA16436.1	Ацилтрансфераза	295	0.627
CAB88882.1	Вероятный белок внутренней мембраны	110	0.612
CAB51427.1	Вероятный мембранный белок	275	0.608
CAC04234.1	Вероятный трансмембранный белок	511	0.589
CAB88472.1	ABC транспортер фосфатов	258	0.583
CAA20004.1	Гистоноподобный ДНК-связывающий белок	218	0.579
CAB90891.1	Двухкомпонентная система обратной реакции	233	0.566
CAB70633.1	Вероятная двухкомпонентная система обратной реакции	222	0.547
CAB94591.1	Репрессор, чувствительный к перекиси водорода	138	0.538
CAB42663.1	Пирролин-5-карбоксилат редуктаза	284	0.533
CAC01592.1	Вероятный сигма фактор	221	0.530
CAB66241.1	30s Рибосомальный белок S20.	88	0.528
CAB94530.1	50S Рибосомальный белок L31	74	0.516
CAB42783.1	Вероятный 30S Рибосомальный белок S14	101	0.516
CAA15876.1	RecX, вероятный регуляторный белок	188	0.507
CAB82083.1	30S Рибосомальный белок S14	61	0.505
CAD55531.1	Возможный транскрипционный регулятор семейства LysR	316	0.503
CAB77409.1	50S Рибосомальный белок L33	54	0.500
CAB94062.1	Вероятный белок SC7A12.15.	206	0.500
CAC37897.1	Вероятный ABC транспортер	662	0.500
CAC14501.1	Ацетилтрансфераза	174	0.499
CAC32286.1	Вероятный белок внутренней мембраны	324	0.499
CAB42698.1	Вероятный белок	124	0.498
CAB92674.1	Вероятный белок внутренней мембраны	558	0.494
CAB61184.1	Вероятный белок	293	0.485
CAB92836.1	Вероятный мембранный белок	293	0.478
CAC04497.1	Вероятный белок	375	0.478
CAB93378.1	Шикимат 5-дегидрогиназа	255	0.472

Таблица 6. Первые 30 белков, отобранных нашим алгоритмом по плюс-минус-спискам (6) и характерных для организмов, способных к спорообразованию.