Аннотация фрагмента генома археи класса Thermoplasmata

Пунько Анастасия
Факультет компьютерных наук, ВШЭ, Москва

Введение.

Археи класса Thermoplasmata являются ацидофилами и обитают в кислой среде с оптимальным pH < 2,0. Большинство организмов этого класса не имеют клеточной стенки, а также являются термофилами [1,2]. Благодаря своим особенностям и широкому спектру мест обитания, Thermoplasmata могут служить отличной моделью для изучения адаптации к окружающей среде и эволюционных процессов у архей.

Аннотация геномов новых видов архей может быть сложной задачей из-за отсутствия достаточного количества сходных геномов для сравнения и ограниченной доступности биологических данных для этой группы организмов, что затрудняет понимание метаболического потенциала, экологии и эволюционной истории этого класса архей.

В данной работе мы провели аннотацию участка генома археи класса Thermoplasmata (35.fasta, 21646 bp) и предоставили максимальное количество биологически значимой информации. В результате нашей работы мы получили ценную информацию о метаболическом потенциале и экологии археи класса Thermoplasmata. Наша аннотация может служить основой для дальнейших исследований этого уникального класса архей и помочь расширить наше понимание их адаптации к окружающей среде и эволюционных процессов.

Методы.

Предсказание белок-кодирующих генов

Для первичной аннотации фрагмента генома археи использовались биоинформатические инструменты RAST [3], PROKKA [4] и GeneMarkS [5]. В целом, перечисленные инструменты предоставляют быстрый и автоматизированный способ аннотации геномов бактерий и архей, что позволяет исследователям быстро получать информацию о функциях генов и метаболических путях. Анализ был выполнен с параметрами по умолчанию. Между выводами трех программ обнаружены небольшие отличия, но они не существенны. Для дальнейшего предсказания функций белков использовались координаты кодирующих последовательностей участка генома и fasta-файл аминокислотных последовательностей из вывода программы RAST.

Предсказание некодирующих РНК

Для поиска и аннотации некодирующих РНК во фрагменте генома археи использовались программы tRNAscan-SE [6] и RFAM [7]. Они предоставляют информацию о функциях ncRNA, которые могут играть важную роль в регуляции генной экспрессии и метаболизме клетки.
Для предсказания транспортной РНК была использована программа tRNAscan-SE v. 2.0 с параметрами по умолчанию. tRNAscan-SE использует скрытые Марковские модели для предсказания местоположения транспортной РНК в геноме. Также она может определять не только классические тРНК, но и нестандартные варианты, такие как тРНК с измененными антикодонами.
Для предсказания других некодирующих РНК использовалась программа RFAM. RFAM используется для идентификации и классификации различных типов ncRNA, таких как рибосомные РНК, транспортные РНК, сигнальные РНК и другие.

Предсказание функций белков

Для предсказания функций белков использовалась программа InterPro [8], которая обеспечивает функциональный анализ белков, классифицируя их по семействам и предсказывая домены и важные участки. Для этого InterPro использует прогностические модели (сигнатуры) из нескольких баз данных (CATH-Gene3D, CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, PROSITE profiles, PROSITE patterns, SFLD, SUPERFAMILY, SMART, TIGRFAMs), входящих в консорциум InterPro. Сигнатуры белков из этих баз данных объединены в единый ресурс с возможностью поиска, используя их индивидуальные преимущества для создания мощной интегрированной базы данных и диагностического инструмента.

Pfam, SUPERFAMILY и SCANPROSITE используются для интерпретации функций белков на основе сходства. SMART и CATH — для поиска функций белков на основе архитектуры домена и для категоризации доменов в структурной иерархии соответственно. Для поиска консервативных доменов используется база данных CDD. Анализ был выполнен с параметрами по умолчанию.

Прогнозирование местоположения белков в клетке, трансмембранных спиралей и сигнальных белков

Для оценки функции белка важно знать его местоположение в клетке. Программа PSORTb [9] использует как экспериментальные данные, так и прогнозы in silico для обнаружения локализации белков в клетке.
Для предсказания наличия трансмембранных спиралей и сайтов расщепления сигнального пептида использовались программы TMHMM [10] и SignalP [11]. Анализ был выполнен с параметрами по умолчанию.

Поиск по BLAST для предсказания функций белков и для выявления горизонтального переноса генов

Для поиска областей локального сходства между нашими белковыми последовательностями и белковыми последовательностями из баз данных использовали программу BLAST [12]. В данном случае использовалась база данных NCBI nonredundant data-base и хиты с идентичностью >90%. Хиты с низким значением e-value из филумов, отличных от изучаемой археи, рассматривали как следствие возможного горизонтального переноса генов для последующего подтверждения с помощью построения филогенетического дерева.
Множественное выравнивание последовательностей осуществлялось с помощью MUSCLE [13] с настройками по-умолчанию. Выравнивание сохраняли в формате Phylip.
Для построения филогенетического дерева использовался пакет PhyML [14] с параметрами по умолчанию. PhyML использует современные статистические подходы для анализа выравнивания нуклеотидных или аминокислотных последовательностей в филогенетических исследованиях. Он основан на критерии максимального правдоподобия и реализует большое количество моделей замещения в сочетании с эффективными опциями для поиска топологий филогенетических деревьев.

Предсказание оперонов

Оперон выявляли в тех случаях, если два гена на одной нити имели межгенное расстояние < 150 bp, а функции белков были похожи.

Результаты и обсуждение.

Технологии секвенирования ДНК продолжают развиваться, и высокопроизводительные методы секвенирования позволяют получать все большее количество последовательностей геномов архей. Программы для аннотации не всегда могут точно предсказать функции генов, поэтому необходимо применять несколько биоинформатических инструментов. В данном исследовании мы использовали ряд эффективных инструментов и баз данных для аннотации кодирующих и некодирующих последовательностей фрагмента генома археи класса Thermoplasmata.

Предсказание белок-кодирующих генов

При первичной аннотации программы RAST и PROKKA обнаружили по 25 CDS (8 на прямой цепи, 17 на обратной), GeneMarkS обнаружил 26 CDS (8 на прямой цепи, 18 на обратной). Результаты программных пакетов представлены в Приложении 1, Приложении 2 и Приложении 3 соответственно. Таблица 1 содержит сравнение результатов всех трех программ.

Между выводами обнаружены небольшие отличия, но они не существенны. Есть небольшие отклонения в значения координат, однако диапазон значений в выводе программы RAST включал значения из PROKKA и GeneMarkS, поэтому для дальнейшего предсказания функций белков использовались координаты кодирующих последовательностей участка генома и fasta-файл аминокислотных последовательностей из вывода этой программы. В GeneMarkS была обнаружена дополнительная кодирующая последовательность на обратной цепи, однако в анализ мы ее не включали, так как в предыдущих двух базах она не была обнаружена, а в базе данных BLAST не было найдено ни одного схожего белка.

Предсказание некодирующих РНК

Программа tRNAscan выявила две транспортные РНК (Таблица 2). Последовательность для tRNA-Ile2-CAT расположена на прямой цепи, для tRNA-Gly-TCC — на обратной. Дополнительные детали приведены в Приложении 4.

RFAM не обнаружила других некодирующих РНК.

Таблица 1 - Сравнение результатов программ для аннотирования генома

	RAST			**PROKKA**			**GeneMarkS**		
№CDS	Start	Stop	Strand	Start	Stop	Strand	Start	Stop	Strand
1	1229	81	-	1229	81	-	1229	81	-
2	1847	1278	-	1847	1278	-	1847	1278	-
3	1942	2883	+	1942	2883	+	1942	2883	+
4	4077	2884	-	4077	2884	-	4077	2884	-
5	4502	4071	-	4502	4071	-	4502	4071	-
6	5639	4503	-	5639	4503	-	5639	4503	-
7	5734	7179	+	5734	7179	+	5734	7179	+
8	8009	7299	-	8009	7299	-	8009	7299	-
9	8485	8006	-	8485	8006	-	8485	8006	-
10	8641	10989	+	8641	10989	+	8641	10989	+
11	11025	11918	+	11025	11918	+	11331	11918	+
12	12585	11884	-	12585	11884	-	12585	11884	-
13	12663	13163	+	12714	13163	+	12663	13163	+
14	13226	13618	+	13226	13618	+	13226	13618	+
15	13615	14844	+	13615	14844	+	13615	14844	+
16	15418	14837	-	15418	14837	-	15418	14837	-
17	15527	15886	+	15527	15886	+	15527	15886	+
18	17261	16017	-	17261	16017	-	17273	16017	-
19	17875	17270	-	17875	17270	-	17875	17270	-
20	18497	17868	-	18497	17868	-	18518	17868	-
21	19038	18502	-	19038	18502	-	19038	18502	-
22	20060	19044	-	20060	19044	-	20060	19044	-
23	20361	20107	-	20361	20107	-	20361	20107	-
24	21145	20354	-	21145	20354	-	21145	20354	-
25	21645	21142	-	21645	21142	-	21645	21142	-
26							11332	11132	-

Желтым цветом выделены координаты, которые отличаются.

Таблица 2 - тРНК, найденные во фрагменте генома

№tRNA	tRNA Begin	Bounds End	tRNA Type	Anti Codon	Intron Begin	Bounds End	Inf Scope	Isotope CM	Isotope Score
1	5	78	Ile2	CAT	0	0	88.9	Ile2	102.9
2	15983	15911	Gly	TCC	0	0	77.5	Gly	88.9

Предсказание функций белков

Результаты программы InterPro представлены в Приложении 5 и кратко описаны в Таблице 3. Для кодирующих последовательностей 3, 13, 17 функция белка может быть неточной, так как достоверность предсказания низкая. Для последовательностей 16 и 23 функции белков не были обнаружены.

Таблица 3 - Предсказанные функции белков

№CDS	**InterPro**
1	Aminoacyl-tRNA synthetase, class Ic (IPR002305)
2	Amino acid exporter protein, LeuE-type (IPR001123)
3	Winged helix-turn-helix DNA-binding domain
4	GTP-binding protein
5	Uncharacterised protein family UPF0179 (IPR005369)
6	Glycerol dehydrogenase (IPR016205)
7	Phenylalanyl-tRNA synthetase, class IIc, alpha subunit (IPR004529)
8	Queuosine biosynthesis protein QueC (IPR018317)
9	6-pyruvoyl tetrahydropterin synthase/QueD family (IPR007115)
10	DNA-directed DNA polymerase, family B (IPR006172)
11	Phosphate transporter (IPR001204)
12	Putative phosphate transport regulator (IPR018445)
13	Region of a membrane-bound protein predicted to be embedded in the membrane.
14	Ribosomal protein S6e (IPR001377)
15	Translation initiation factor 2, gamma subunit (IPR022424)
16	None predicted
17	Region of a membrane-bound protein predicted to be embedded in the membrane.
18	Dihydroorotase (IPR004722)
19	AMMECR1 (IPR023473)
20	7-carboxy-7-deazaguanine synthase-like (IPR024924)
21	3-hexulose-6-phosphate isomerase (IPR017552)
22	AP endonuclease 2 (IPR001719)
23	None predicted
24	Arsenical pump ATPase, ArsA/GET3 (IPR016300)
25	Nicotinamide-nucleotide adenylyltransferase, archaeal type (IPR006418)

Прогнозирование местоположения белков в клетке, трансмембранных спиралей и сигнальных белков

Результаты прогнозирования представлены в таблице 4. Сигнальные белки для нашего участка генома не обнаружены. Белки кодирующих последовательностей 2, 11, 13 и 17 расположены в цитоплазматической мембране. Для белков CDS 5, 6, 18, 19, 24 локализация в клетке не установлена.

Таблица 4 - Список предсказанных локализаций в клетке, трансмембранных спиралей и сигнальных белков

	Sub-cellular localization	**Transmembrane helice**	**Signal Peptide**
No.CDS	**PSORT B**	**DeepTMHMM**	**SignalP**
1	Cytoplasmic	0	No
2	CytoplasmicMembrane	6	No
3	Cytoplasmic	0	No
4	Cytoplasmic	0	No
5	Unknown	0	No
6	Unknown	0	No
7	Cytoplasmic	0	No
8	Cytoplasmic	0	No
9	Cytoplasmic	0	No
10	Cytoplasmic	0	No
11	CytoplasmicMembrane	8	No
12	Cytoplasmic	0	No
13	CytoplasmicMembrane	1	No
14	Cytoplasmic	0	No
15	Cytoplasmic	0	No
16	Cytoplasmic	0	No
17	CytoplasmicMembrane	3	No
18	Unknown	0	No
19	Unknown	0	No
20	Cytoplasmic	0	No
21	Cytoplasmic	0	No
22	Cytoplasmic	0	No
23	Cytoplasmic	0	No
24	Unknown	0	No
25	Cytoplasmic	0	No

Поиск по BLAST для предсказания функций белков и для выявления горизонтального переноса генов

Результаты поиска схожих последовательностей в Blastp представлены в таблице 5. Указаны ID белка, вид организма, покрытие, e-value, идентичность и функция белка. Для всех CDS обнаружены те же функции, что и в InterPro. Для 16 и 23 CDS не было найдено значимых последовательностей схожих с нашими. Также стоит отметить, что чаще всего схожие последовательности с низким e-value и высокой идентичностью встречаются у Thermoplasmatales archaeon, вероятно, к этому виду принадлежит наш участок генома.

Таблица 5 - Результаты поиска в Blastp

No. CDS	Protein ID	Organism	Query cover	e-value	Identity	Product
1	MBD6955104.1	Thermoplasmatales archaeon	100 %	0.0	100.00	tryptophan-tRNA ligase
	HEU13153.1	Euryarchaeota archaeon	77 %	0.0	100.00	TPA: tryptophan-tRNA ligase
2	MBD6955103.1	Thermoplasmatales archaeon	100 %	5E-125	98.94	LysE family translocator
	PMP74867.1	Aciduliprofundum sp.	73 %	1E-82	95.65	lysine transporter LysE
3	MBD6955102.1	Thermoplasmatales archaeon	100 %	0.0	99.68	winged helix-turn-helix transcriptional regulator
	HEU13152.1	Euryarchaeota archaeon	20 %	1E-33	98.44	TPA: hypothetical protein
4	MBD6955101.1	Thermoplasmatales archaeon	99 %	0.0	99.75	redox-regulated ATPase YchF
	PMP73506.1	Aciduliprofundum sp.	99 %	0.0	99.49	redox-regulated ATPase YchF
5	MBD6955100.1	Thermoplasmatales archaeon	98 %	1E-99	100.00	UPF0179 family protein
6	MBD6955099.1	Thermoplasmatales archaeon	92 %	0.0	100.00	NAD(P)-dependent glycerol-1-phosphate dehydrogenase
	HEU12571.1	Euryarchaeota archaeon	73 %	0.0	99.64	TPA: iron-containing alcohol dehydrogenase
	PMP73503.1	Aciduliprofundum sp.	92 %	0.0	99.43	NAD(P)-dependent glycerol-1-phosphate dehydrogenase
7	MBD6955098.1	Thermoplasmatales archaeon	100 %	0.0	100.00	phenylalanine--tRNA ligase subunit alpha
	PMP73502.1	Aciduliprofundum sp.	100 %	0.0	99.79	phenylalanine--tRNA ligase subunit alpha

8	MBD6955097.1	Thermoplasmatales archaeon	100 %	2E-170	100.00	7-cyano-7-deazaguanine synthase QueC
	HEU12569.1	Euryarchaeota archaeon	55 %	8E-90	100.00	TPA: 7-cyano-7-deazaguanine synthase
9	MBD6955096.1	Thermoplasmatales archaeon	100 %	3E-113	100.00	6-pyruvoyl tetrahydropterin synthase family protein
10	MBD6955095.1	Thermoplasmatales archaeon	100 %	0.0	99.87	DNA polymerase II
	PMP73499.1	Aciduliprofundum sp.	100 %	0.0	99.62	DNA polymerase II
	MCI4435017.1	Thermoplasmata archaeon	100 %	0.0	90.55	DNA polymerase II
11	MBD6955093.1	Thermoplasmatales archaeon	100 %	1E-162	100.00	DUF47 family protein
12	MBD6955094.1	Thermoplasmatales archaeon	100 %	0.0	99.66	hypothetical protein
	PMP73498.1	Aciduliprofundum sp.	100 %	0.0	99.33	hypothetical protein C0180_06585
	HEU12837.1	Euryarchaeota archaeon	100 %	0.0	98.99	TPA: hypothetical protein
13	MBD6955092.1	Thermoplasmatales archaeon	89 %	4E-97	100.00	hypothetical protein
	PMP73437.1	Aciduliprofundum sp.	100 %	2E-108	98.80	hypothetical protein C0180_06730
14	MBD6955091.1	Thermoplasmatales archaeon	100 %	6E-86	100.00	30S ribosomal protein S6e
	MCI4434244.1	Thermoplasmata archaeon	100 %	1E-78	90.00	30S ribosomal protein S6e
15	MBD6955090.1	Thermoplasmatales archaeon	99 %	0.0	100.00	translation initiation factor IF-2 subunit gamma
	PMP73447.1	Aciduliprofundum sp.	99 %	0.0	99.75	translation initiation factor IF-2 subunit gamma
	HEU12675.1	Euryarchaeota archaeon	83 %	0.0	99.71	TPA: translation initiation factor IF-2 subunit gamma
	MCI4434245.1	Thermoplasmata archaeon	99 %	0.0	93.84	translation initiation factor IF-2 subunit gamma
16	MBD6955089.1	Thermoplasmatales archaeon	100 %	1E-136	100.00	hypothetical protein
17	MBD6955088.1	Thermoplasmatales archaeon	100 %	7E-78	100.00	hypothetical protein
	HEU12677.1	Euryarchaeota archaeon	80 %	2E-61	100.00	TPA: hypothetical protein
	PMP73440.1	Aciduliprofundum sp.	100 %	4E-77	98.32	hypothetical protein C0180_06750

18	MBD6955087.1	Thermoplasmatales archaeon	100 %	0.0	99.76	amidohydrolase family protein
	PMP73441.1	Aciduliprofundum sp.	100 %	0.0	99.52	amidohydrolase
19	PMP73449.1	Aciduliprofundum sp.	100 %	3E-142	100.00	TIGR00296 family protein
	MBD6955086.1	Thermoplasmatales archaeon	100 %	8E-142	99.50	TIGR00296 family protein
20	MBD6955085.1	Thermoplasmatales archaeon	100 %	2E-150	100.00	radical SAM protein
	HEU12589.1	Euryarchaeota archaeon	75 %	4E-110	100.00	TPA: radical SAM protein
21	MBD6955084.1	Thermoplasmatales archaeon	100 %	7E-126	100.00	SIS domain-containing protein
22	MBD6955083.1	Thermoplasmatales archaeon	100 %	0.0	100.00	AP endonuclease
23	MBD6955082.1	Thermoplasmatales archaeon	84 %	1E-42	100.00	hypothetical protein
24	MBD6955081.1	Thermoplasmatales archaeon	100 %	0.0	100.00	ArsA family ATPase
	PMP73444.1	Aciduliprofundum sp.	100 %	0.0	99.24	hypothetical protein C0180_06790
25	MBD6955080.1	Thermoplasmatales archaeon	100 %	1E-117	99.40	nicotinamide-nucleotide adenylyltransferase
	PMP73445.1	Aciduliprofundum sp.	100 %	7E-117	98.20	nicotinate-nucleotide adenylyltransferase

Для построения филогенетических деревьев и анализа возможного горизонтального переноса генов брали схожие последовательности из результатов BLAST для CDS 19 для видов Thermoplasmatales archaeon, Euryarchaeota archaeon и Aciduliprofundum sp.
На филогенетическом дереве (Рисунок 1) показано, что CDS19 кластеризуется с Thermoplasmatales archaeon, а значит, горизонтального переноса нет.

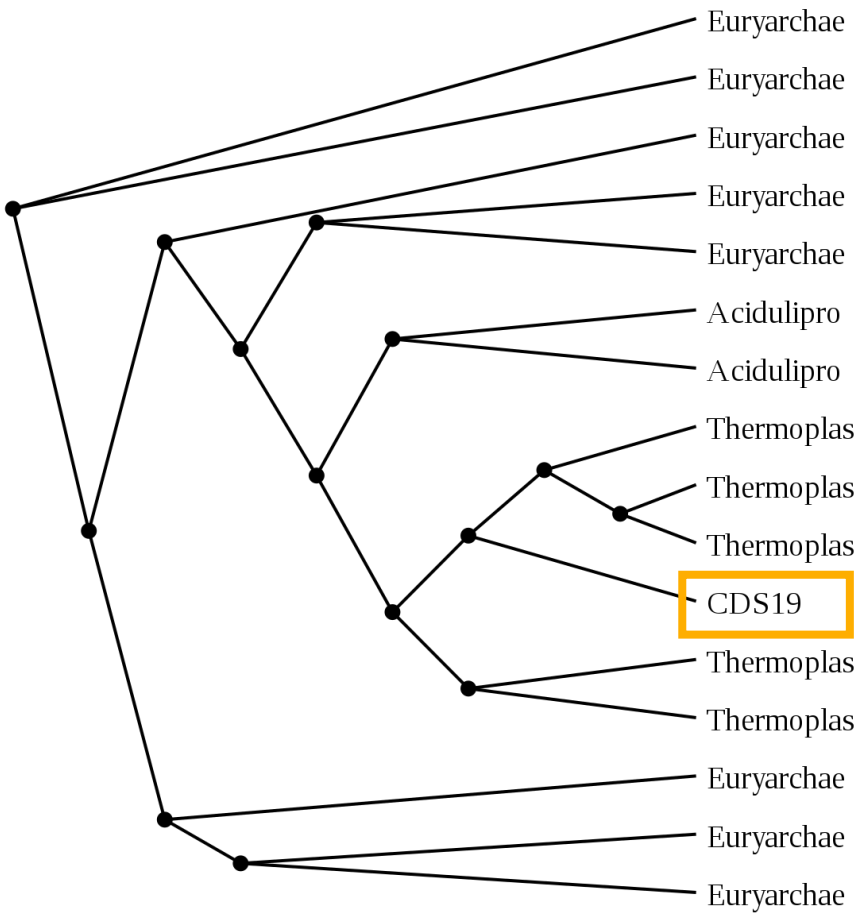

Рисунок 1 - Филогенетическое дерево видов

*Предсказание опероно*в

CDS8 и CDS9 могут оказаться опероном из белков QueC и QueD (ферменты биосинтеза Queuosine-гипермодифицированное основание в положении wobble некоторых тРНК). Ферменты QueD, QueE и QueC последовательно участвуют в образовании промежуточного продукта preQ0, который в конечном итоге преобразуется через ряд реакций в Queuosine.
Белки CDS5, CDS6 и CDS18, CDS20, CDS22 связывают ионы металлов, поэтому кластеры генов в пределах CDS5-CDS6 и CDS18-CDS22 также могут оказаться оперонами.

Таблица 6 - Участки генов с межгенным расстоянием < 150 bp

№CDS	Start	Stop	Strand	Intergenic regions	Function
10	8641	10989	+		DNA-directed DNA polymerase, family B (IPR006172)
11	11025	11918	+	36	Phosphate transporter (IPR001204)
13	12663	13163	+		None predicted
14	13226	13618	+	63	Ribosomal protein S6e (IPR001377)
15	13615	14844	+	-3	Translation initiation factor 2, gamma subunit (IPR022424)
1	1229	81	-		Aminoacyl-tRNA synthetase, class Ic (IPR002305)
2	1847	1278	-	49	Amino acid exporter protein, LeuE-type (IPR001123)
4	4077	2884	-		GTP-binding protein
5	4502	4071	-	-6	Uncharacterised protein family UPF0179 (IPR005369)
6	5639	4503	-	1	Glycerol dehydrogenase (IPR016205)
8	8009	7299	-		Queuosine biosynthesis protein QueC (IPR018317)
9	8485	8006	-	-3	6-pyruvoyl tetrahydropterin synthase/QueD family (IPR007115)
18	17261	16017	-		Dihydroorotase (IPR004722)
19	17875	17270	-	9	AMMECR1 (IPR023473)
20	18497	17868	-	-7	7-carboxy-7-deazaguanine synthase-like (IPR024924)
21	19038	18502	-	5	3-hexulose-6-phosphate isomerase (IPR017552)
22	20060	19044	-	6	AP endonuclease 2 (IPR001719)
23	20361	20107	-	47	None predicted
24	21145	20354	-	-7	Arsenical pump ATPase, ArsA/GET3 (IPR016300)
25	21645	21142	-	-3	Nicotinamide-nucleotide adenylyltransferase, archaeal type (IPR006418)

Функции некоторых предсказанных белков участка генома археи

CDS 1 и 7 - связывают ATP, нуклеотиды, осуществляют аминоацилирование тРНК для трансляции белка. CDS 2 - осуществляет транспорт аминокислот. CDS 4 и 15 - связывают гуанозин трифосфат (GTP). CDS 8 и 9 - ферменты биосинтеза Queuosine. CDS 10 - осуществляет функции ДНК-полимеразы. CDS 14 - структурный компонент рибосомы. CDS 19 - высокий уровень консервативности домена AMMECR1 указывает на базовую клеточную функцию, потенциально связанную с механизмами транскрипции, репликации, репарации или трансляции. CDS 24 - связывает ATP. CDS 25 - участвует в процессе биосинтеза NAD.

CDS 11 - обеспечивает перенос неорганического фосфата с одной стороны мембраны на другую, вверх по градиенту концентрации.
CDS 12 - может играть роль в транспорте ортофосфата.

CDS 5 - функция этого семейства неизвестна, однако белки содержат два цистеиновых кластера, которые могут быть железо-серными redox-центрами. CDS 6 - связывает ионы тяжелых металлов. CDS 18 - связывает ионы металлов. CDS 20 - связывает комбинации атомов железа и серы. CDS 21 - осуществляет связывание с углеводными производными. CDS 22 - связывает ионы цинка.

Заключение.

Функциональная аннотация белков имеет важное значение, так как белковые макромолекулы участвуют в многочисленных биологических процессах. В данном исследовании был использован подход in silico для функциональной аннотации участка генома археи класса Thermoplasmata.
Мы предсказали функции 23 белков. В нескольких белках обнаружена функция связывания ионов металлов (CDS 5, 6, 18, 20, 22), два белка выполняли функцию транспорта фосфатов (CDS 11, 12). Также мы показали места локализации белков в клетке для понимания специфических свойств аннотированных белков, попытались выявить горизонтальный перенос генов и наличие оперонов. Изучение представленной информации позволяет понять, что обеспечивает адаптацию археи к экстремальным условиям окружающей среды.
Сочетание такого in-silico анализа с соответствующими лабораторными экспериментами позволит получить функциональные аннотации белков из различных организмов. Кроме того, полученные результаты открывают перспективы для дальнейшего изучения этой археи в биотехнологических целях.

Список использованных источников

1-Yuan Y, Liu J, Yang TT, Gao SM, Liao B, Huang LN. Genomic Insights into the Ecological Role and Evolution of a Novel *Thermoplasmata* Order, "*Candidatus* Sysuiplasmatales". Appl Environ Microbiol. 2021 Oct 28;87(22):e0106521. doi: 10.1128/AEM.01065-21. Epub 2021 Sep 15. PMID: 34524897; PMCID: PMC8552897.

2-Hu, Wenzhe; Pan, Jie; Wang, Bin; Guo, Jun; Li, Meng; Xu, Meiying (2020). Metagenomic insights into the metabolism and evolution of a new Thermoplasmata order (Candidatus Gimiplasmatales). Environmental Microbiology, 1462-2920.15349–. doi:10.1111/1462-2920.15349.

3-The RAST Server: Rapid Annotations using Subsystems Technology.
Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.
BMC Genomics, 2008.

4-Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nature Biotechnology. 2018;36: 566. doi: 10.1038/nbt.4163.

5-John Besemer, Alexandre Lomsadze and Mark Borodovsky. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Research (2001) 29, pp 2607-2618.

6-Lowe, T.M. and Chan, P.P. (2016) tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucl. Acids Res. 44:W54-57.

7-Rfam 14: expanded coverage of metagenomic, viral and microRNA families Kalvari et al., Nucleic Acids Research (2020)

8-Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunić I, Marchler-Bauer A, Mi H, Natale DA, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A. InterPro in 2022. *Nucleic Acids Research, Nov 2022, (doi: 10.1093/nar/gkac993).*

9-PSORTb v3.0: N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26(13):1608-1615.

10-Jeppe Hallgren, Konstantinos D. Tsirigos, Mads D. Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh and Ole Winther (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. https://doi.org/10.1101/2022.04.08.487609.

11-Nielsen H, Tsirigos KD, Brunak S, von Heijne G. A brief history of protein sorting prediction. The pro- tein journal. 2019; 38(3):200–16. https://doi.org/10.1007/s10930-019-09838-3 PMID: 31119599.
12-Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic acids research. 2008; 36(suppl_2):W5–W9. https://doi.org/10.1093/nar/gkn201 PMID: 18440982.
13-Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Apr:gkac240. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731.
14-"New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. Systematic Biology, 59(3):307-21, 2010.

Приложение

Приложение 1 - результаты программы RAST.
Приложение 2 - результаты программы PROKKA.
Приложение 3 - результаты программы GeneMarkS.
Приложение 4 - результаты программы tRNAscan.
Приложение 5 - результаты программы InterPro.
Приложение 6 - результаты программы BLAST.
Приложение 7 - результаты программы MUSCLE.