doi 10.18699/bgrs2026-60

Классификация редко встречающихся митохондриальных генов различных групп эукариот и их сопоставление с ядерными генами человека

Иванов М.М.[1]*, Диброва Д.В.[1, 2]
[1] *Факультет биоинженерии и биоинформатики, МГУ имени М.В. Ломоносова, Москва, Россия*
[2] *НИИ ФХБ имени А.Н. Белозерского, МГУ имени М.В. Ломоносова, Москва, Россия*
* *ivamikem@gmail.com*

Ключевые слова: митохондриальный геном; редкие гены; кластеры ортологичных групп

Мотивация и цель: Ключевую роль в возникновении эукариот сыграл эндосимбиоз с предком митохондрий – α-протеобактерией. После этого события многие гены из митохондриального генома (митогенома) были либо утрачены, либо перенесены в ядро. Обратный процесс практически не описан и сопряжен с рядом трудностей, поэтому закодированные в митогеномах гены представляют особый интерес для изучения эволюции эукариот. Крайняя степень редукции митогенома характерна для Metazoa (у человека в нем закодировано лишь 13 белков), тогда как митогеномы растений, грибов и протистов намного более разнообразны. Цель работы – поиск и филогеномный анализ новых и редких белок-кодирующих генов в митохондриях различных групп эукариот с их последующей классификацией и сопоставлением с ядерными геномами человека.
Методы и алгоритмы:
1. Составление выборки митохондриальных геномов. Из базы NCBI Datasets (раздел Organelles) загружены 2106 полных митохондриальных геномов, относящихся к грибам (851 вид), зеленым растениям (683 вида), красным водорослям и протистам (572 вида). Metazoa исключены из анализа ввиду однообразия их митогеномов. Для обработки файлов GenBank и извлечения белок-кодирующих последовательностей использованы собственные скрипты на Python 3.
2. Сравнение белковых последовательностей с профильными HMM для MitoCOGs. Ранее были описаны MitoCOGs (митоКОГи) – кластеры ортологичных генов, описывающие типичные митохондриальные белки [1]. Мы построили на их основании 140 профильных HMM и с помощью пакета HMMER v3.4 провели сравнение этих моделей с полученными белковыми последовательностями (порог E-value = 1e-5). Белки без значимых совпадений классифицированы как потенциально редкие. Профильные HMM разделены на функциональные категории согласно оригинальной статье.
3. Кластеризация и функциональный анализ редких белков. Для всех белков, не имеющих соответствий с профильными HMM митоКОГов, был проведен поиск сходства с прокариотическими КОГами с использованием программы hmmsearch (порог E-value = 1e-5). Оставшиеся белки были подвергнуты процедуре кластеризации *de novo* на основе алгоритма EdgeSearch для построения новых КОГов [2]. Алгоритм строит граф симметричных лучших совпадений BLAST (с пороговым E-value = 1e-5) и выделяет максимальные треугольно-связные подграфы, формирующие новые кластеры ортологичных групп (addCOGs). Для полученных кластеров проводился поиск консервативных мотивов и доменов с использованием баз данных NCBI CDD, InterProScan и инструмента DomainAnalyser (https://boabio.belozersky.msu.ru/DomainAnalyser). Для оптимизации кластеров, полученных алгоритмом EdgeSearch, сначала были получены профильные HMM на основе получившихся addCOGs и было проведено сравнение данных профилей со всеми известными последовательностями с помощью hmmsearch. Для визуализации распределения находок новых профильных HMM использовался подход, аналогичный реализованному в инструменте COGcollator (https://boabio.belozersky.msu.ru/COGcollator), позволяющий оценить степень родства между различными кластерами. Для объединения кластеров использовались собственные скрипты на языке Python 3.
Результаты: Из 2106 митохондриальных геномов получены последовательности 65752 белков, аннотированных в них. Из них 55481 белок (84,4 %) имел соответствия с митоКОГами, тогда как для 10271 белка (15,6 %) значимых совпадений не обнаружено (такие белки были нами названы «редкими»). Наибольшее число редких белков выявлено у зеленых растений (5397, ~19 % от их общего числа), затем у грибов (3506, ~17,5 %) и протистов с красными водорослями (1368, ~7,8 %).
Анализ таксономического распределения показал, что среди грибов максимум редких белков наблюдается у представителей класса Agaricomycetes и некоторых аскомицетов (до 266 белков у *Morchella brunnea*). У зеленых растений, при стабильном размере митогеномов (30–40 белков), у отдельных видов отмечено резкое увеличение числа редких белков (у *Capsicum annuum* 127 из 193 белков не имеют соответствия с митоКОГами). В трибе Brassiceae стабильно высокое число редких белков (в среднем 35 из 75). Среди протистов наибольшее разнообразие редких белков выявлено у инфузорий класса Spirotrichea (до 43 % от общего числа генов).
Функциональный состав белков с митоКОГами закономерно различается: у грибов преобладают белки дыхательной цепи и LAGLIDADG-эндонуклеазы, у растений – сбалансированный набор белков дыхательной цепи, рибосомальных белков и белков домашнего хозяйства, у протистов – промежуточный вариант.

Сравнение редких белков с прокариотическими КОГами дало совпадения для 258 последовательностей (2,5 %). Среди них обнаружены 116 белков, характерных для хлоропластов: компоненты фотосистем I и II, рубиско-подобные белки, субъединицы цитохромного *b6f*-комплекса. Вероятно, эти белки – части белок-кодирующих генов пластид, которые называются MTPT (mitochondrial plastid DNA) [3].

Кластеризация оставшихся 10013 редких белков позволила сгруппировать 3276 последовательностей в 555 новых кластеров (addCOGs). После оптимизации на основе анализа веса находок HMM и множественного выравнивания получено 380 кластеров addCOGs, включающих 3711 белков (т.е. были добавлены 435 ранее некластеризованных последовательностей). Пример оптимизации кластера в сторону расширения показан на Рис. 1. Как можно видеть, изначально разделенные на более мелкие кластеры последовательности объединяются в более крупные кластеры ортологов, в зависимости от веса находок и таксономической принадлежности. Анализ доменной структуры показал, что 802 последовательности содержат домен Pfam субъединицы 8 АТФ-синтазы растений (PF02326), а 472 — домен Pfam субъединицы 8 АТФ-синтазы грибов (PF05933). «Каноническая» субъединица 8, которая встречается у животных, описывается MitoCOG0140, и наши данные указывают на значительные отличия найденных вариантов этого белка, требующие более детального изучения.

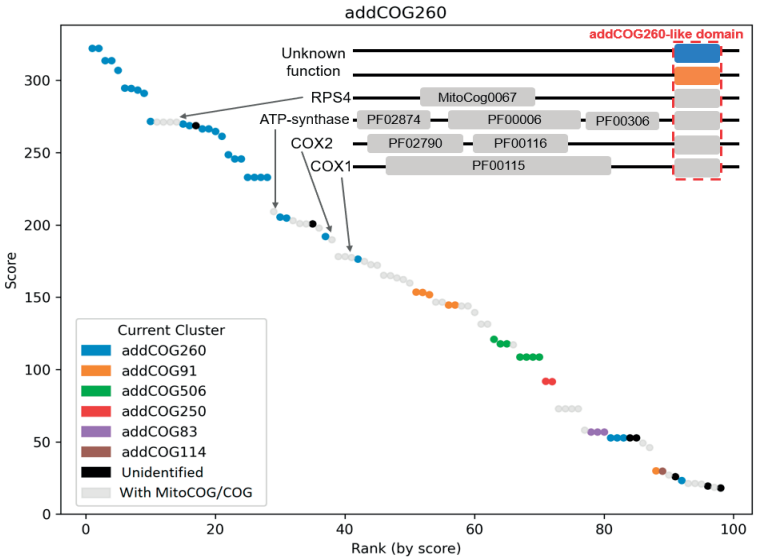


Рис. 1. График зависимости веса находок hmmsearch от их ранга для профильной HMM полученного кластера addCOG260 в наборе из 2106 митогеномов. Каждая точка на графике – отдельный белок, цветом обозначено к какому из кластеров addCOG каждая находка принадлежит. Также показана доменная архитектура некоторых находок и их общий домен, характерный для белков из addCOG260 (красный пунктир).

Мы также провели сравнение представителей отобранных новых кластеров и ядерного генома человека используя как подход, основанный на последовательностях аннотированных белков (blastp), так и подход, основанный на поиске по нуклеотидным базам данных (tblastn).

Выводы: В ходе работы показано, что во всех рассмотренных группах эукариот присутствует значительное количество белков (более 10 тысяч), не относящихся к известным семействам митохондриальных белков. Наибольшее число таких редких белков обнаружено у зеленых растений. Часть из них (2,5 %) имеет гомологи среди прокариотических КОГов, включая белки хлоропластного происхождения. Кластеризация позволила выделить 380 новых кластеров (addCOGs), объединяющих 3711 редких белков, для части которых предсказаны функции (варианты субъединицы 8 АТФ-синтазы). Дальнейший анализ этих кластеров и оставшихся 6302 некластеризованных белков, включая сопоставление с ядерными геномами человека (учитывая миграцию митохондриальных генов в ядро в ходе эволюции находки гомологичных белков могут присутствовать в геноме), позволит уточнить их функцию и эволюционное происхождение. Построенные профильные HMM для митоКОГов и для новых, относительно редких белков, закодированных в митогеномах, доступны на сайте: https://boabio.belozersky.msu.ru/tools.

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 25-24-00757, https://rscf.ru/project/25-24-00757/.

Classification of rare mitochondrial genes from various groups of eukaryotes and their comparison with human nuclear genes

Ivanov M.M.[1]*, Dibrova D.V.[1, 2]

[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

[2] *A. N. Belozersky Research Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia*

* *ivamikem@gmail.com*

Key words: mitochondrial genome; rare genes; clusters of orthologous groups

Motivation and Aim: Endosymbiosis with the ancestor of mitochondria – an α-proteobacterium – played a key role in the origin of eukaryotes. Following this event, many genes from the mitochondrial genome (mitogenome) were either lost or transferred to the nucleus. The reverse process has been described only rarely and is associated with multiple constraints; therefore, genes encoded in mitogenomes are of particular interest for studying eukaryotic evolution. An extreme degree of mitogenome reduction is typical for Metazoa (only 13 proteins are encoded in the human mitogenome), whereas the mitogenomes of plants, fungi, and protists are much more diverse. The aim of this work is to identify and perform a phylogenomic analysis of novel and rare protein-coding genes in the mitochondria of various eukaryotic groups, followed by their classification and comparison with the human nuclear genome.
Methods and Algorithms:
1. Compilation of a dataset of mitochondrial genomes. A total of 2106 complete mitochondrial genomes were downloaded from NCBI Datasets (Organelles section), including fungi (851 species), green plants (683 species), and red algae together with protists (572 species). Metazoa were excluded from the analysis due to the high uniformity of their mitogenomes. Custom Python 3 scripts were used to process GenBank files and extract protein-coding sequences.
2. Comparison of protein sequences with profile HMMs for MitoCOGs. Previously described MitoCOGs (mitochondrial clusters of orthologous genes) represent clusters of orthologous genes corresponding to typical mitochondrial proteins [1]. Based on these clusters, 140 profile HMMs were constructed and the HMMER v3.4 package was used to compare these models with the obtained protein sequences (E-value threshold = 1e-5). Proteins without significant matches were classified as potentially rare. Profile HMMs were grouped into functional categories according to the original publication.
3. Clustering and functional analysis of rare proteins. For proteins without matches to MitoCOG profile HMMs, similarity searches against prokaryotic COGs were performed using hmmsearch (E-value threshold = 1e-5). The remaining proteins were subjected to *de novo* clustering using the EdgeSearch algorithm to construct new COG-like clusters [2]. The algorithm builds a graph of symmetric best BLAST hits (E-value threshold = 1e-5) and identifies maximal triangle-connected subgraphs that form new clusters of orthologous groups (addCOGs). For the resulting clusters, conserved motifs and domains were identified using the NCBI CDD database, InterProScan, and the DomainAnalyser tool (https://boabio.belozersky.msu.ru/en/DomainAnalyser). To optimize clusters obtained by the EdgeSearch algorithm, profile HMMs were first constructed from the resulting addCOGs, after which these profiles were compared against all known sequences using hmmsearch. Visualization of the distribution of hits for the new profile HMMs was performed using an approach analogous to that implemented in the COGcollator tool (https://boabio.belozersky.msu.ru/en/COGcollator), allowing the degree of relatedness between clusters to be evaluated. Custom Python 3 scripts were used to merge clusters where necessary.
Results: From the 2106 mitochondrial genomes, 65,752 proteins annotated in these genomes were obtained. Of these, 55,481 proteins (84.4 %) had matches to MitoCOGs, while 10,271 proteins (15.6 %) showed no significant matches and were therefore classified as rare proteins. The largest number of rare proteins was identified in green plants (5,397, approximately 19 % of the total), followed by fungi (3,506, approximately 17.5 %) and protists together with red algae (1,368, approximately 7.8 %).
Analysis of taxonomic distribution showed that among fungi the highest number of rare proteins occurs in representatives of the class Agaricomycetes and in some Ascomycetes (up to 266 proteins in *Morchella brunnea*). Among green plants, despite relatively stable mitogenome sizes (30–40 proteins), some species exhibit a sharp increase in the number of rare proteins (127 of 193 proteins in *Capsicum annuum* lack matches to MitoCOGs). Within the tribe Brassiceae, the number of rare proteins is consistently high (on average 35 out of 75). Among protists, the greatest diversity of rare proteins was observed in ciliates of the class Spirotrichea (up to 43 % of all genes).
The functional composition of proteins with MitoCOG matches differs systematically between taxa. In fungi, respiratory chain proteins and LAGLIDADG endonucleases dominate; in plants, a balanced set of respiratory chain proteins, ribosomal proteins, and housekeeping proteins is observed; protists show an intermediate pattern.
Comparison of rare proteins with prokaryotic COGs produced matches for 258 sequences (2.5 %). Among these, 116 proteins characteristic of chloroplasts were identified, including components of photosystems I and II, Rubisco-like proteins, and subunits of the cytochrome b6f complex. These proteins likely correspond to fragments of plastid protein-coding genes known as MTPT (mitochondrial plastid DNA) [3].
Clustering of the remaining 10,013 rare proteins grouped 3,276 sequences into 555 new clusters (addCOGs). After optimization based on the analysis of HMM hit weights and multiple sequence alignments, 380 addCOG clusters were obtained containing 3,711 proteins (including 435 previously unclustered sequences). An example of cluster expansion during optimization is shown in Figure 1. As illustrated, sequences initially separated into smaller clusters are merged into a larger orthologous clusters depending on hit weights and taxonomic affiliation. Domain structure analysis revealed that 802 sequences contain the Pfam domain of the plant ATP synthase subunit 8 (PF02326), while 472 sequences contain the fungal ATP synthase subunit 8 domain (PF05933). The canonical subunit 8 found in animals corresponds to MitoCOG0140, and our results indicate substantial differences between the newly identified variants of this protein, requiring further investigation. We also compared representatives of selected new clusters with the human nuclear genome using both a protein-based approach (blastp) and a nucleotide-database search approach (tblastn).

Conclusion: Our results demonstrate that all analyzed groups of eukaryotes contain a substantial number of proteins (more than 10,000) that do not belong to known families of mitochondrial proteins. The largest number of such rare proteins was detected in green plants. A small fraction of these proteins (2.5 %) has homologs among prokaryotic COGs, including proteins of chloroplast origin.

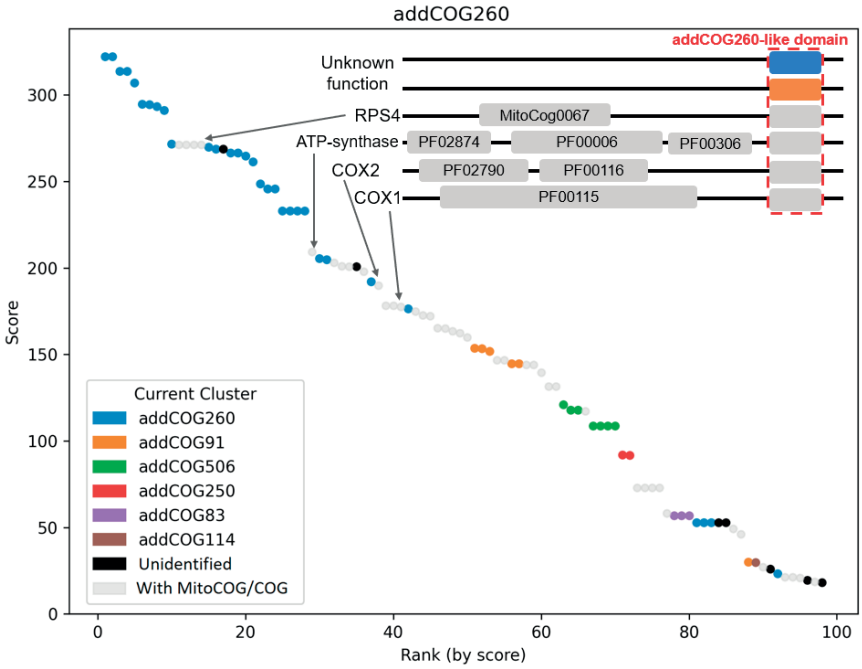


Fig. 1. Graph showing score of hmmsearch findings against their rank for the profile HMM of the addCOG260 cluster as well as domain architecture of some proteins. Each point on the graph represents a separate protein, with the color indicating which addCOG cluster each finding belongs to. addCOG260-like conserved protein domain is shown with red-dotted line.

Clustering analysis identified 380 new clusters (addCOGs) comprising 3,711 rare proteins, for some of which functions were predicted (variants of ATP synthase subunit 8). Further analysis of these clusters and the remaining 6,302 unclustered proteins, including comparison with the human nuclear genome (considering that mitochondrial genes have migrated to the nucleus during evolution and homologous proteins may therefore be present in nuclear genomes), will help clarify their function and evolutionary origin. Profile HMMs constructed for MitoCOGs and for newly identified rare mitogenome-encoded proteins are available at: https://boabio.belozersky.msu.ru/en/tools.

Funding: This work was supported by the Russian Science Foundation, project No. 25-24-00757, https://rscf.ru/en/project/25-24-00757/.

Список литературы/References

1. Kannan *et al.*: MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. *BMC Evolutionary Biology* 2014, 14:237. doi: 10.1186/s12862-014-0237-5
2. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, Mushegian A. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* 2010, 26(12):1481-7. doi: 10.1093/bioinformatics/btq229
3. Ellis, J. Promiscuous DNA—chloroplast genes inside plant mitochondria. *Nature* 1982, 299:678–679. doi: 10.1038/299678a0