Реферат по биоинформатике на тему:

**«Какие белковые домены были у последнего общего предка (LUCA)?»**

Выполнила студентка 4 курса ФББ Калинина Марина

Содержание

[Введение 3](#_Toc406969994)

[Обсуждение 5](#_Toc406969995)

[LUCApedia 13](#_Toc406969996)

[Выводы 15](#_Toc406969997)

[Список литературы 16](#_Toc406969998)

# Введение

Проблема определения общего предка встала после того, как была предложена трехдоменная классификация организмов (Bacteria, Archaea, Eukaryota). Carl Woese изначально предположил, что предками бактерий являются археи [1], однако позже пришел к выводу, что обе группы должны иметь общего предка. Если построить филогенетическое дерево на основании данных рРНК современных организмов, общий предок будет до разделения ветвей Bacteria и Archaea+Eucaryota (Рисунок 1). Для этого общего предка было предложено название LUCA – Last Universal Common Ancestor.

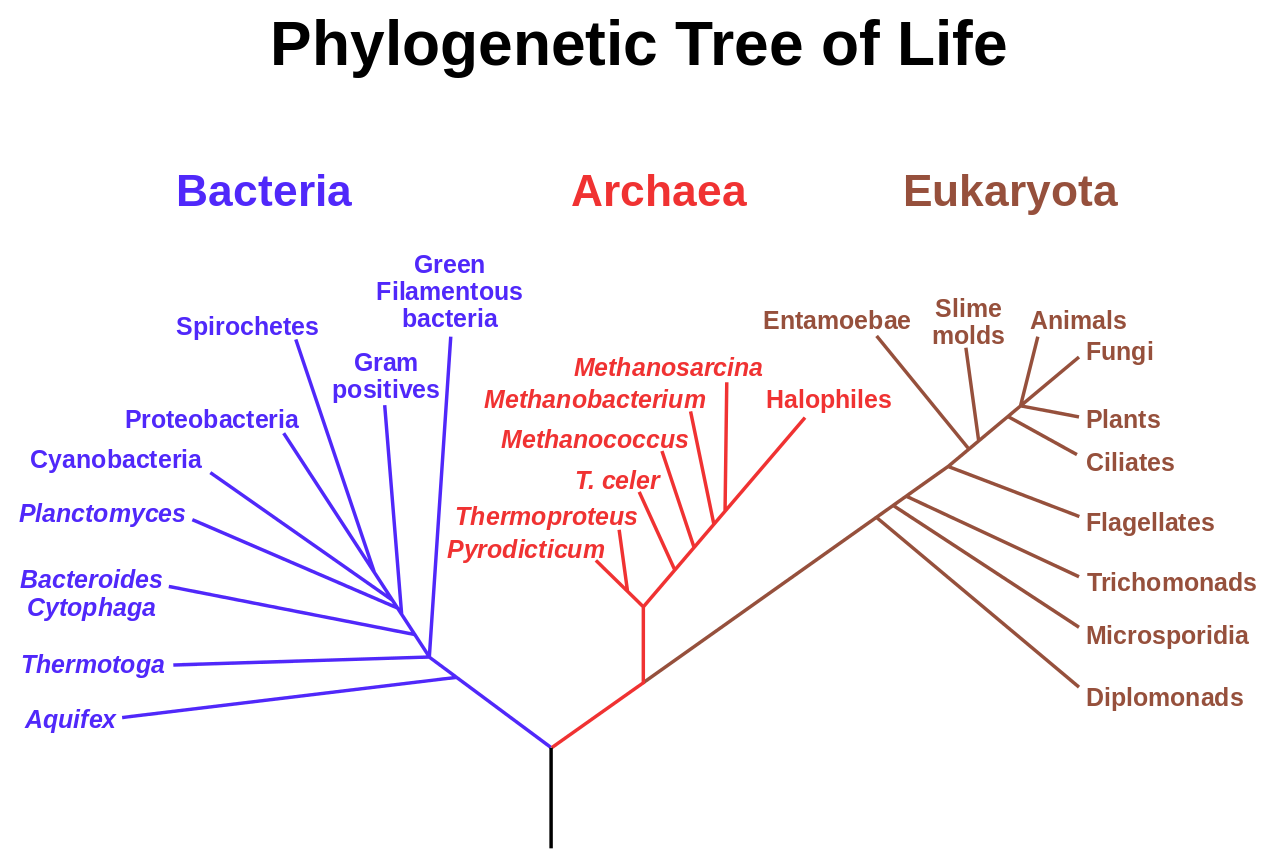


Рисунок 1. Кладограмма, построенная на основе анализа рРНК. Показывает связь бактерий, архей и эукариот с последним общим предком (показан черным цветом внизу дерева)

Возможно, предковая форма вместе с ее современниками вымерла. В 1998 году тот же Carl Woese предположил, что LUCA – не индивидуальный организм, а сообщество организмов, и что генетическое наследство всех современных организмов получено путем горизонтального переноса генов среди предкового сообщества организмов [2].

Современное научное сообщество пришло к согласию, что LUCA был ДНК-содержащим организмом. Определение генетического состава предкового организма – сложная задача, так как требуется каким-то образом оценить вклад горизонтального переноса генов. Были попытки оценить минимальный набор генов, присутствующих у этого организма. При сравнительном анализе около 100 геномов обнаружилось 60 генов, присутствующих У ВСЕХ них [3]. Однако, большинство этих генов кодируют компоненты системы трансляции. Этот набор генов явно недооценен, так как сложно представить организм, у которого присутствует сложный аппарат трансляции, однако, слишком мало генов, чтобы осуществлять метаболизм и другие клеточные процессы.

Были также и более поздние работы по оценке генетического состава LUCA, где оценивались семейства генов по базе данных COGs. По одним оценкам – это 80 COGs (присутствуют у всех организмов) [4], по другим оценкам – 571 COGs (не обязательно присутствуют у всех организмов, но должны присутствовать функциональные аналоги) [5].

Предположение, что структуры белков более консервативны с течением эволюции, чем нуклеотидные последовательности, легло в основу исследований, посвященных определению протеома (а если быть более точным, то состава белковых доменов) общего предка.

# Обсуждение

В данном разделе будут обсуждаться результаты статьи [6], в основу которой было положено предположение о том, что структура белков более консервативна в ходе эволюции, чем нуклеотидные последовательности. Для того, чтобы произвести структурный анализ, была использована база данных Gene3D [7], которая соотносит суперсемейства доменов CATH с миллионами белковых последовательностей. Последовательность каждого домена из CATH выравнивается в BLAST с целью найти ближайших гомологов. Затем обучается некая скрытая Марковская модель (HMM) на соответствие белковых последовательностей доменам CATH. На вход подается геном, и при помощи HMM в нем ищутся во всех открытых рамках считывания все возможные домены (Рисунок 2).

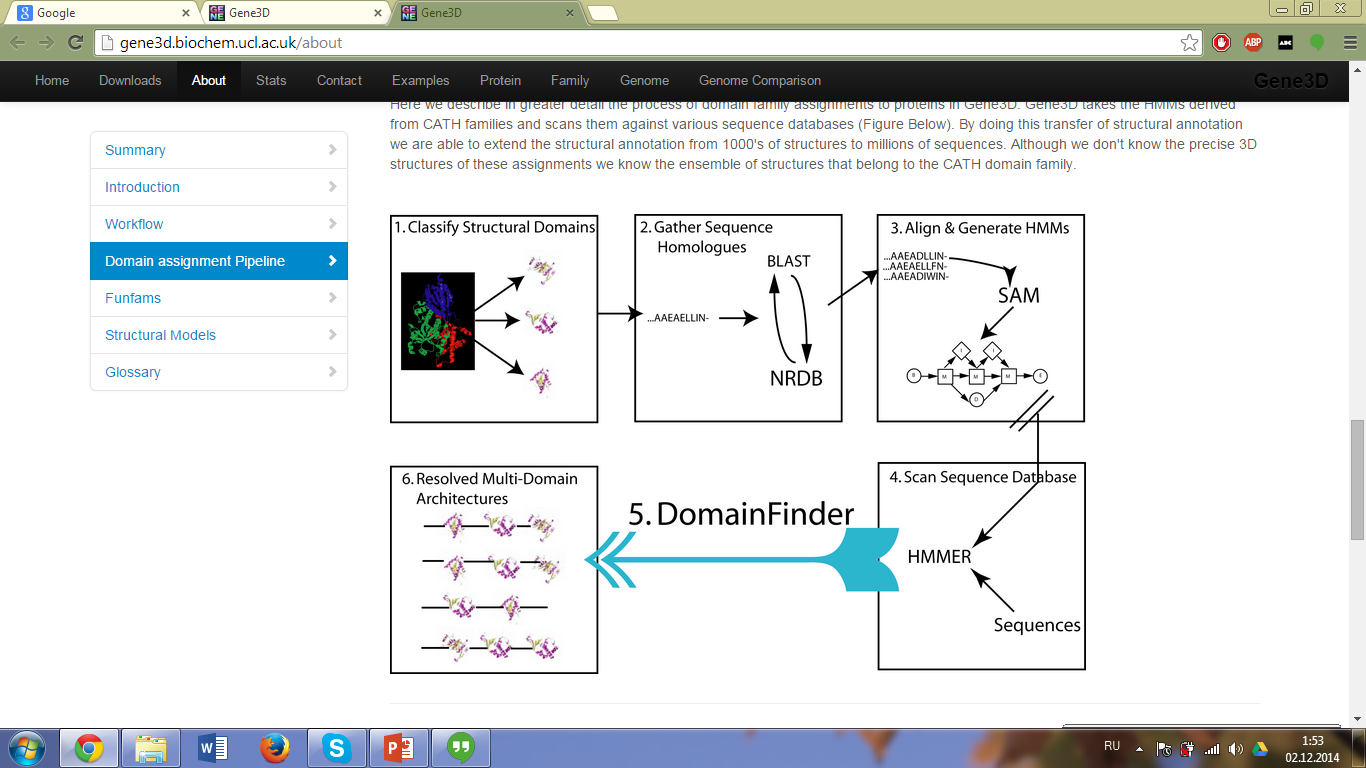


Рисунок 2. Алгоритм в основе базы данных Gene3D

А как определяется структурная схожесть доменов? Для этого использовался алгоритм SSAP [8], который с помощью двойного динамического программирования может оценить структурную гомологию доменов. Для этого для каждой пары Сβ атомов (для глицина используется мнимый Сβ атом) из сравниваемых доменов строятся векторы до других (видимо, близких в пространстве) Cβ атомов (считается, что сравнивая взаимное расположение Cβ атомов можно получить лучшие результаты при структурном выравнивании, чем для Сα атомов). Затем строится матрица расстояний, в которой находится оптимальный путь методом динамического программирования. Это повторяется для всех пар аминокислотных остатков из сравниваемых доменов. Происходит сравнение некоторого значения оптимального пути с порогом, и если значение выше порога, путь переносятся в суммарную матрицу. Наконец, в суммарной матрице методом динамического программирования находится оптимальный путь. Этот путь и есть структурное выравнивание доменов. При этом считается некоторый Score, который сам по себе ничего не говорит, а служит лишь для сравнения различных структурных выравниваний. Данный алгоритм реализуется в базе данных CATH.

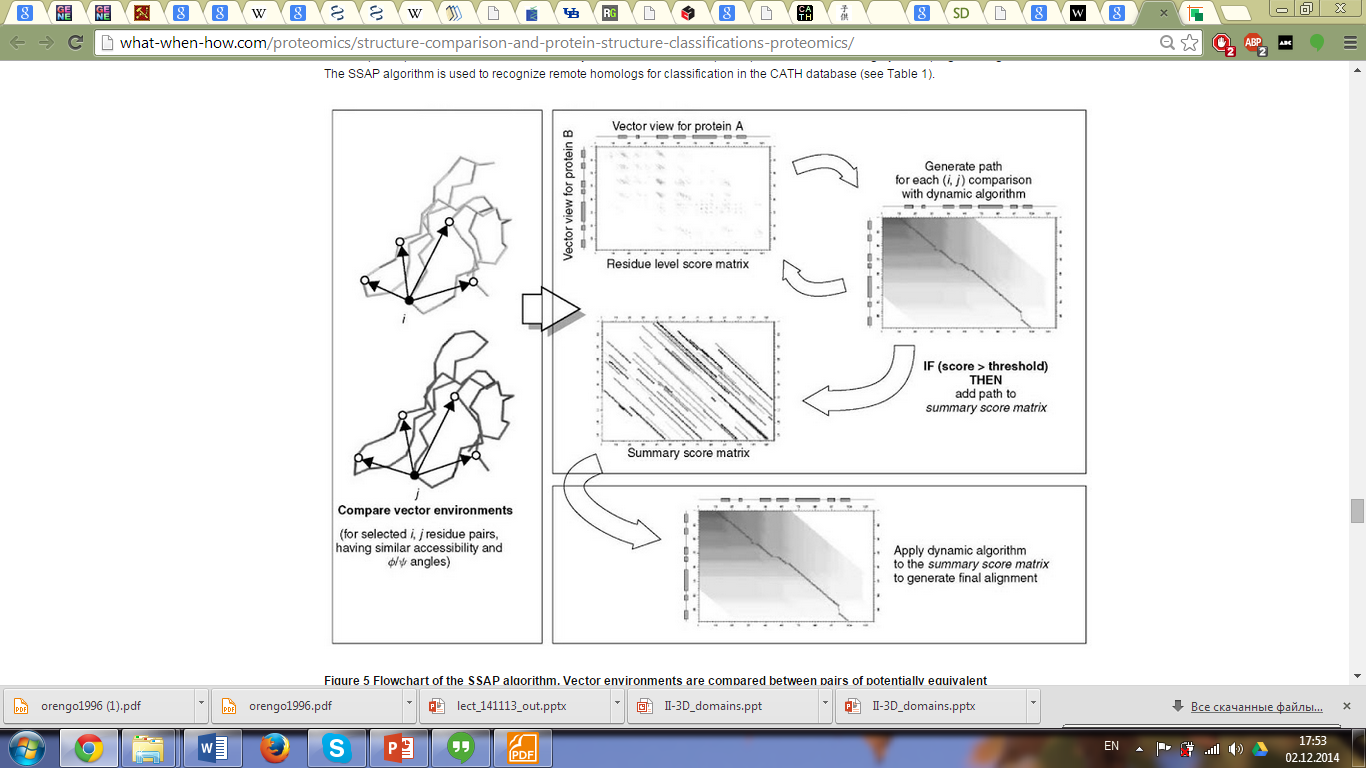


Рисунок 3. Алгоритм SSAP

Также использовался алгоритм CORA [9]. В этом алгоритме для каждого суперсемейства доменов CATH строился некоторый консенсус по последовательностям, где отмечались остатки, важные для функционирования этого домена (а также их окружение, конформация и т.д.). Затем через этот «шаблон» пропускается интересующий нас домен. При этом определяется соответствие интересующего домена консенсусу и структурная гомология.

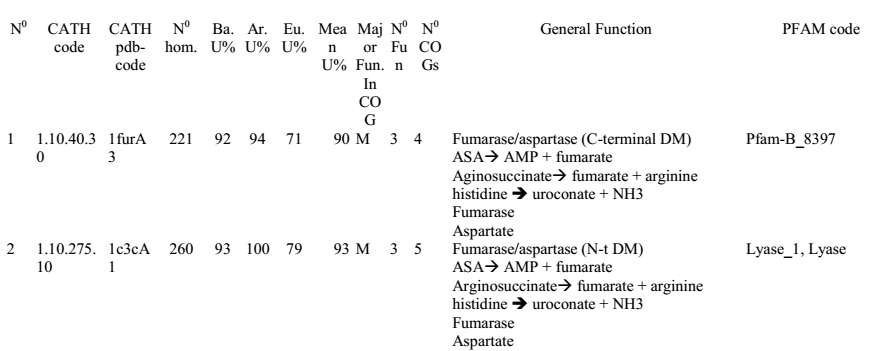
Однако, окончательное решение о гомологии доменов принималось экспертами. При этом учитывались данные о функциональности доменов из различных источников – баз данных (COGs, KEGG и др.), и из статей.

А как определить «предковость» домена? Во всех работах на определение генома или протеома LUCA пытались как-либо оценить горизонтальный перенос генов. Но так как никто не знает среднего уровня горизонтального переноса генов, в статье попытались избежать каких-либо его оценок за счет очень консервативного отбора доменов, претендующих на звание «предковых». Для этого домен должен присутствовать более, чем в 90% всех видов, и более, чем в 70% видов архей и эукариот (последним пытались нормализовать недопредставленность эукариотических геномов и геномов архей по сравнению с бактериальными геномами). Такие жесткие оценки, на взгляд авторов, помогут устранить влияние горизонтального переноса генов, хотя и приведут к недооценке сложности протеома LUCA.

При анализе 114 полных геномов (85 бактериальных, 15 архей и 14 эукариотических) выделили 140 предковых суперсемейств доменов (15% от всех суперсемейств, 55% от бактериальных, 18% от эукариотических). Каждое суперсемейство было отнесено к какой-либо функциональной группе из COG (трансляция, репликация, метаболизм, клеточные процессы, транскрипция, либо плохо охарактеризованные).

Затем попытались определить, как же функционировал организм LUCA, у которого был бы такой набор структурных доменов. Некоторые домены, однако, представлены в большом количестве белков, где выполняют различные функции (например, АТФ-связывающий домен, или NADH-связывающий домен), что ничего не говорит о том, какую же функцию они выполняли в предковом организме. Специалистами для каждого суперсемейства доменов были выбраны функции, которые затем были наложены на схему клеточных процессов (Таблица 1, Рисунок 4).

Таблица 1. Пример отобранных «предковых» доменов с указанием процента организмов, в которых они присутствуют, перечислением функций и соотнесением с доменами из Pfam.



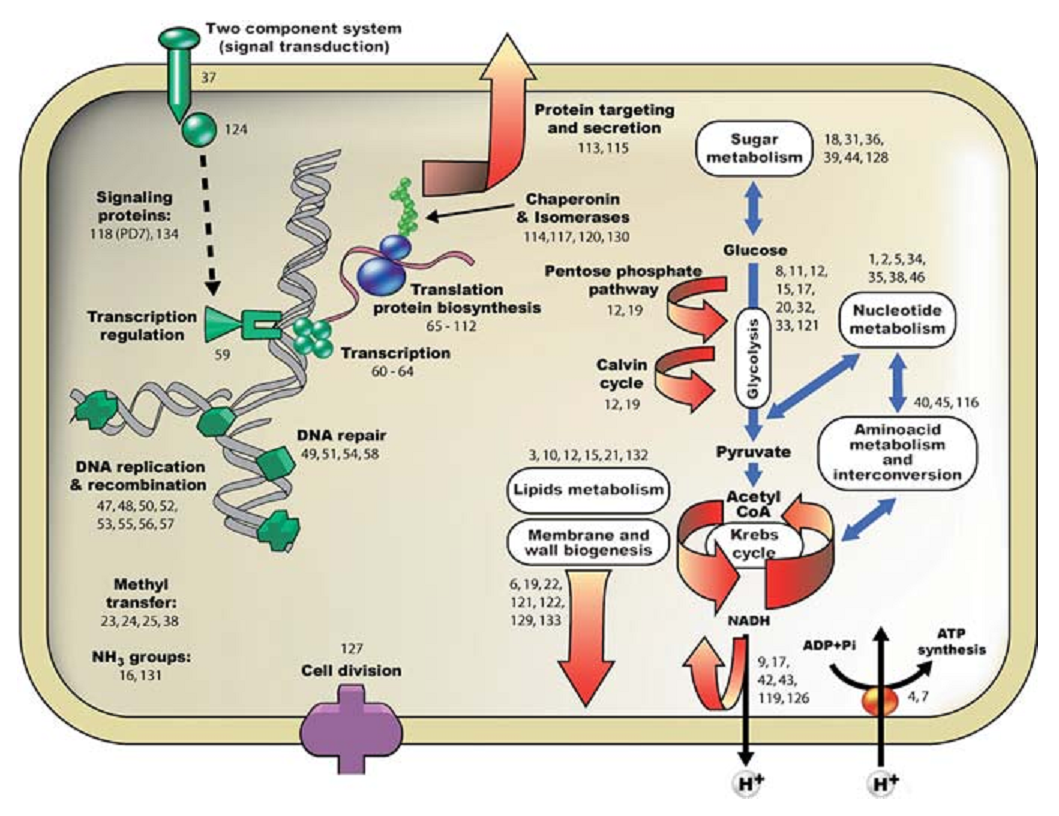


Рисунок 4. Карта функциональных процессов в клетке с номерами суперсемейств доменов, задействованных в этих процессах.

Видим, что у LUCA присутствовал аппарат транскрипции и трансляции, репликации и рекомбинации ДНК, репарации ДНК. Также у него присутствуют ключевые ферменты гликолиза (практически всех реакций) и цикла трикарбоновых кислот (присутствуют домены, присущие α-кетоглуторат-дегидрогеназе и фумаразе). Также присутствуют ферменты трансальдолаза и транскетолаза, которые катализируют обратимые реакции в цикле Кальвина и в пентозофосфатном пути. Цикл Кальвина позволил бы LUCA ассимилировать CO2 и синтезировать гексозы. Однако, ему бы потребовалась дополнительная система для улавливания света и синтеза NADPH и ATP.

Вероятно, что аминокислоты и азотистые основания были в «первичном супе» вокруг LUCA, и он был бы в состоянии их захватывать, нежели синтезировать самостоятельно. Также есть вероятность того, что различные части этих сложных путей биосинтеза были разделены между разными организмами в гипотетической предковой экосистеме.

***«Предковость» и эволюционная температура***

Универсальность всех функциональных групп суперсемейств, представленных в CATH для бактерий (726 суперсемейств), приведена на Рисунке 5. При этом добавление эукариот к анализу не ведет к существенному изменения картины. Видим, что среди функциональной группы «трансляция» наибольший пик наблюдается в интервале наибольшей универсальности (90-100%), в то время как функциональная группа «метаболизм» показывает высокую репрезентативность среди всех классов универсальности (Рисунок 5, а). При переводе этой гистограммы в процентное соотношение суперсемейств и распределения (Рисунок 5, b), видим, что группа «трансляции» имеет ярко выраженное смещение в сторону классов с высокой универсальностью, в то время как распределение группы «метаболизма» больше похоже на равномерное.

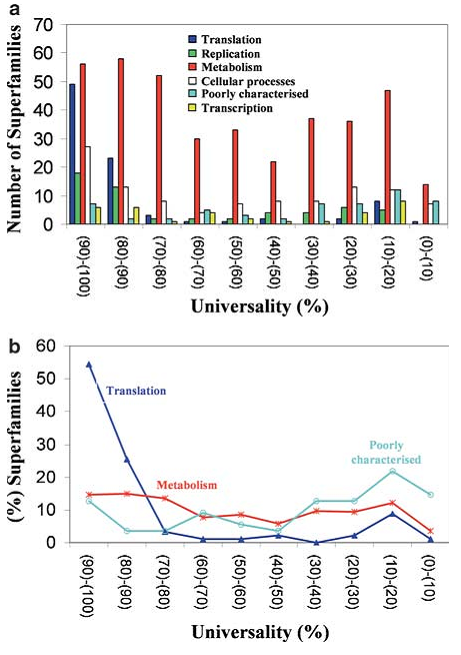


Рисунок 5. Показана универсальность суперсемейств доменов различных функциональных групп для бактериальных видов, a) в абсолютном количестве суперсемейств доменов, b) в процентном соотношении среди доменов данной функциональной группы

Количество гомологов в суперсемействе может быть использовано для оценки уровня дупликации суперсемейства в ходе эволюции. Следует сказать, что Carl Woese ввел понятие эволюционной температуры как количества изменений, примененных к системе для получения в результате разнообразия вариантов. Если построить зависимость количества гомологичных доменов в суперсемействе от кластеров различных ортологов в COG для 726 прокариотических суперсемейств (Рисунок 6, а), видим хорошую корреляцию между уровнями дупликации семейств и их функциональным разнообразием. Это говорит о том, что дупликация генов хорошо коррелирует со снабжением прокариотической клетки новыми функциональными вариантами, и, тем самым, с эволюционной температурой суперсемейств.

Количество различных COG подкатегорий функций показано для доменов в суперсемейств используется как оценка функционального разнообразия в суперсемействе. Была построена зависимость этого количества подкатегорий от количества гомологичных доменов в каждом из 726 суперсемейств прокариот (Рисунок 6, b). Видим, что суперсемейства с высокой универсальностью показывают широких спектр функционального разнообразия, в то время как менее универсальные (или более специфичные) суперсемейства показывают низкие уровни функционального разнообразия.

Если разделить все суперсемейства, отобранные как предковые (140 суперсемейств доменов), по фунциональности, образуются две большие группы – трансляции (46 суперсемейств) и метаболизма (48 суперсемейств). Построив зависимость количества суперсемейств от количества гомологичных доменов (Рисунок 6, c и d), видим четкую разницу между ними. Предковые суперсемейства метаболизма показывают значительный разброс, выявляя, что у них уровень дупликации значительно выше, чем в трансляционных суперсемействах. Около 80% суперсемейств трансляции имеют только одну функциональную подкатегорию COG, ассоциированную с ними. Это контрастирует с метаболическими суперсемействами, где 73% имеет две и более функциональных подкатекорий.

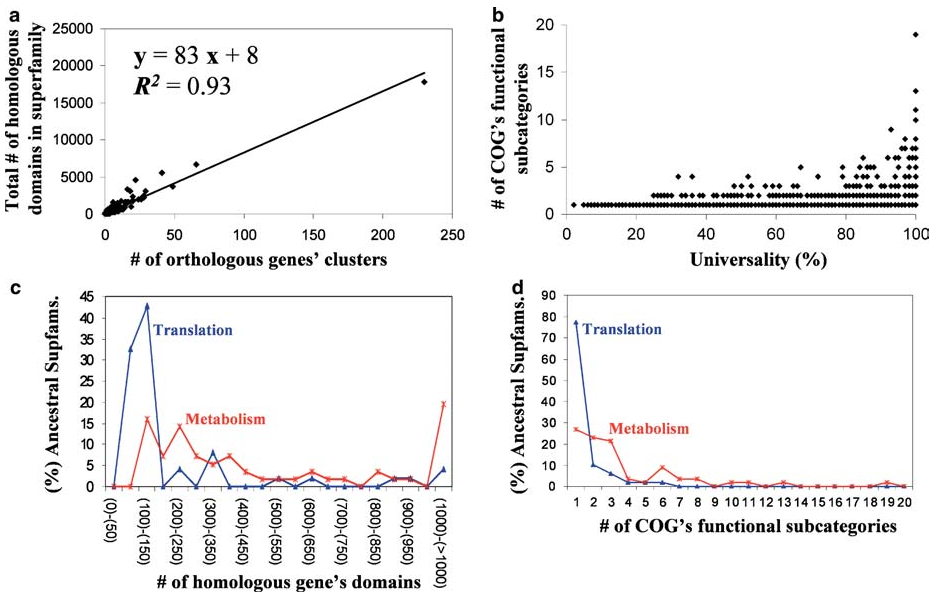


Рисунок 6. a) Зависимость количества гомологов в суперсемействе от количества кластеров ортологов в COG для 726 бактериальных суперсемейств, b) зависимость количества гомологичных доменов в суперсемействе от универсальности суперсемейства, c) распределение предковых суперсемейств функциональных групп трансляции и метаболизма в зависимости от количества гомологичных доменов, d) то же, только в зависимости от функциональных подкатегорий COG.

Далее была проанализирована корреляция между встречаемостью профилей суперсемейств и размером бактериального генома с использованием коэффициента корреляции Пирсона (Рисунок 7, а). Самые большие суперсемейства, с большим количеством гомологов в них, имеют тенденцию к хорошей корреляцией с размером генома, в то время как другие суперсемейства изменяются независимо от размера генома, показывая низкий уровень дупликации в суперсемействах.

На Рисунке 7 (b) изображено распределение частот суперсемейств в зависимости от коэффициента корреляции с размером генома. Наиболее случайное распределение имеет функциональная группа трансляции, в то время как группа метаболизма наиболее коррелирует с изменениями размера генома. Различия между группами трансляции и метаболизма еще четче видны на подобном графике для предковых семейств (Рисунок 7, c).

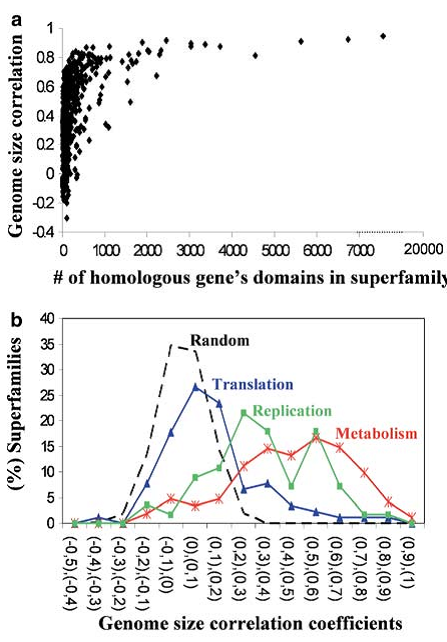
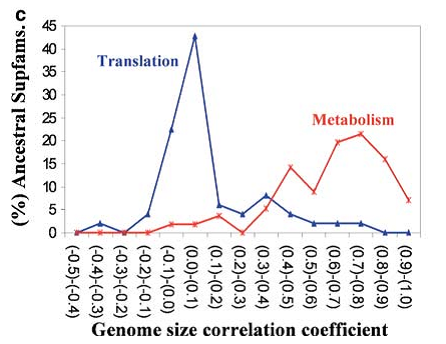


Рисунок 7. a) Корреляция между размером генома и количеством гомологичных доменов в суперсемействе среди бактериальных видов, b) распределения суперсемейств относительно коэффициента корреляции по функциональным группам, и их отклонения от случайного распределения, c) то же самое, но для предковых суперсемейств с функциями трансляции и метаболизма

Далее было важным рассмотреть распределение гомологичных генов среди различных видов. Для этого использовался коэффициент межвидовой вариации генов (CIGV), который рассчитывался делением стандартного отклонения среди всех значений профиля данного суперсемейства на среднее значение этого суперсемейства. Суперсемейства с высоким значением CIGV, вероятно, более адаптивны (имеют бОльшую эволюционную температуру) (Рисунок 8, a), в то время как суперсемейства с константным числом гомологов на вид (низкое значение CIGV) (Рисунок 8, b).

Опять же, в группе трансляции среди предковых суперсемейств распределение сильно смещено в сторону низких значений CIGV (Рисунок 8, c).

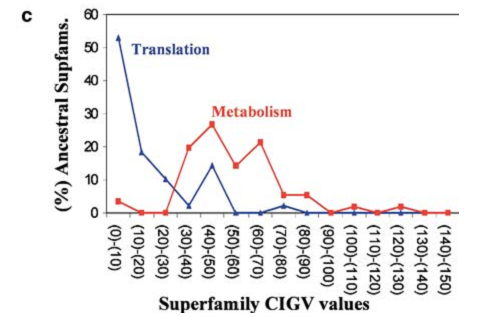
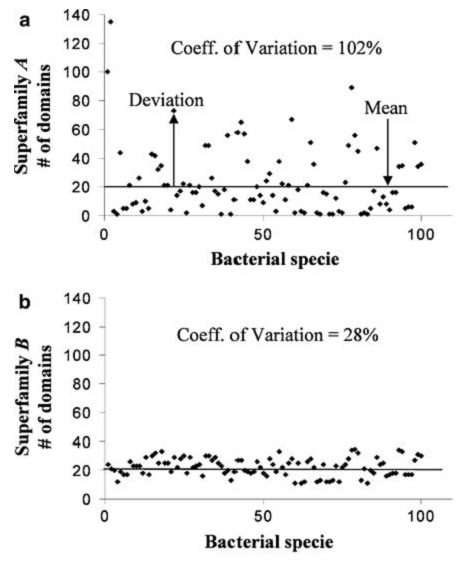


Рисунок 8. a) различные CIGV для суперсемейства доменов шаперонов, b) то же самое для суперсемейства ДНК – связывающих доменов, c) распределения функциональных групп метаболизма и трансляции в предковом наборе суперсемейств относительно значений CIGV.

На Рисунке 9 изображена кумулятивная доля суперсемейств в каждой функциональной группе в различных интервалах универсальности. Диагональная линия в данном случае предполагает, что образование суперсемейства пропорционально образованию новых видов в ходе эволюции. Расположение распределения для функциональной группы трансляции показывает, что новообразование в этих суперсемействах идет медленнее, чем пропорциональное (располагается выше диагональной линии). Метаболизм и клеточные процессы, кажется, имеют константную и важную роль в видообразовании (расположены достаточно близко от диагональной линии).

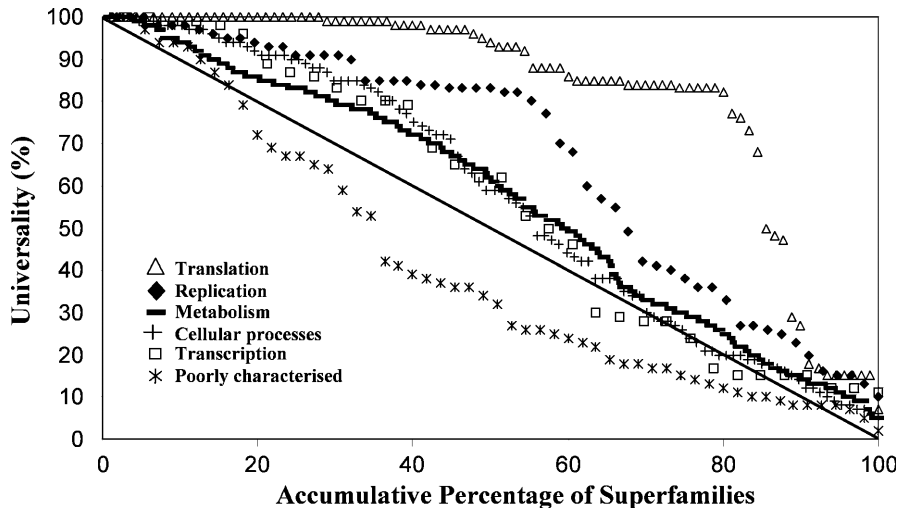


Рисунок 9. Кумулятивный процент суперсемейств в различных интервалах универсальности для разных функциональных групп.

В завершении обзора к этой статье хотелось бы заметить, что различные статистические, проведенные как для всего набора суперсемейств, так и для предковых доменов, не показывают чего-либо нового. Они указывают на очевидные вещи, что белки трансляции более консервативны по отношению к белкам метаболизма, и разнообразие функций у белков в разделе «метаболизм» значительно выше, чем для белков из прочих разделов. Заслуга этой статьи в использовании нового способа для структурной аннотации геномов (Gene3D), и выбор набора из 140 суперсемейств базы данных CATH, претендующих на звание «предковых». Дальнейший статистический анализ нагляден, прост в понимании, но, на мой взгляд, является довольно бесполезным для дальнейшего использования.

# LUCApedia

В 2013 году был создан сервер LUCApedia [10], в котором собраны воедино все наборы белков, генов и др., которые разные авторы относили к LUCA. Было удивительным не увидеть здесь статью [6], на основе которой был написан данный реферат. Зато здесь можно увидеть следующие наборы данных, основанные на работах, посвященных организму LUCA:

1. Набор данных от Harris 2003 [4] - содержит 80 COGs. База данных COGs была использована для определения 80 семейств генов, которые имеются у всех организмов.
2. Набор данных от Mirkin 2003 [5]– содержит 571 COGs. В данной работе требования, предъявленные к геному LUCA, были менее строгие, и отбирались семейства генов, которые необязательно присутствуют во всех организмов, но обязательно имеют функциональные аналоги.
3. Набор данных от Delaye 2005 [11] – содержит 115 мотивов Pfam. Для 20 таксономически разных организмов был проведен поиск «all-against-all» BLAST в обоих направлениях. В результате обнаружились наиболее консервативные гены, и по ним были идентифицированы 115 Pfam доменов.
4. Набор данных от Yang 2005 [12] – 66 SCOP суперсемейств. Был произведен анализ 174 таксономически различных организмов на состав белковых доменов. Метод идентифицировал 66 суперсемейств из SCOP, присутствующих у всех проанализированных организмов.
5. Набор данных от Wang 2007 [13] – 165 укладок из базы данных SCOP. Выбирались укладки, присутствующие у трех таксономических доменов.
6. Набор данных от Srinvasan и Morowitz [14] – 286 EC кодов. Составлен при сравнении метаболических путей трех хемоавтотрофных бактерий и одной хемоавтотрофной археи.

Также содержится информация о рибозимах (33, из экспериментальных литературных данных), использовании в качестве кофакторов нуклеотидов, аминокислот, Fe-S кластеров и ионов цинка (из аннотации Uniprot).

Поисковое окно сервиса выглядит следующим образом (Рисунок 10). При введении описания белка (гена) или его кода (из Uniprot, например), напротив различных наборов данных ставится «+», если белок или ген в данном наборе присутствует, и «-» в противном случае.

An external file that holds a picture, illustration, etc.
Object name is gks1217f2p.jpg

Рисунок 10. Внешний вид сервиса LUCApedia (и его выдача на определенный запрос).

# Выводы

Нельзя однозначно сказать, какой протеом имел общий предок всех ныне живущих организмов, однако, можно оценить сложность этого организма. Из предположения, что белковые домены, которые присутствуют у абсолютного большинства всех организмов, были и у LUCA, можно сделать вывод, что это был довольно сложный организм, в котором, в частности, имелась машина трансляции почти в таком же виде, как у современных организмов. Но при этом набор белковых доменов, выбранных авторами основной статьи, является явно недооцененным, так как слишком малым количеством доменов представлены такие функциональные группы, как метаболизм и клеточные процессы.

Преимущество работы в основе реферата в использовании большого количества геномов (по сравнению с аналогичными работами), а также в попытке предсказать функциональность получившегося организма. Наверно, стоило бы сравнить, насколько пересекаются наборы данных о доменах LUCA разных авторов, но из-за недостатка времени это не представляется возможным.

# Список литературы

1. Woese, C. R. [Bacterial evolution](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC373105/) (англ.) // Microbiol Rev. — 1987. — Т. 51. — № 2. — С. 221–271.
2. Woese C (June 1998). ["The universal ancestor"](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC22660). Proceedings of the National Academy of Sciences of the United States of America
3. Koonin EV (2003) Comparative genomics minimal gene-sets and the last universal commonancestor. Nat Rev Microbiol 1:127–136
4. Harris JK, Kelley ST, Spiegelman GB, Pace NR. The genetic core of the universal ancestor. Genome Res. 2003;13:407
5. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol.
6. Ranea J. A. G. et al. Protein superfamily evolution and the last universal common ancestor (LUCA) //Journal of molecular evolution. – 2006. – Т. 63. – №. 4. – С. 513-525.
7. Buchan D. W. A. et al. Gene3D: structural assignments for the biologist and bioinformaticist alike //Nucleic acids research. – 2003. – Т. 31. – №. 1. – С. 469-473.
8. Orengo C. A., Taylor W. R. SSAP: sequential structure alignment program for protein structure comparison //Computer methods for macromolecular sequence analysis. – 1996
9. Orengo,C. (1999) CORA—topological fingerprints for protein structural families. Protein Sci., 8, 699–715.
10. Goldman A. D. et al. LUCApedia: a database for the study of ancient life //Nucleic acids research. – 2012. – С. gks1217.
11. Delaye L, Becerra A, Lazcano A. The last common ancestor: what's in a name?
12. Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. Proc. Natl Acad. Sci. USA. 2005;102:373–378
13. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res.2007;17:1572–1585
14. Srinvasan V, Morowitz HJ. The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. Biol. Bull. 2009;216:126–130.