Тема работы:
Исследование молекулярной эволюции регуляторных генных сетей развития растений.

Состав коллектива:
Гунбин Константин Владимирович, к.б.н., с.н.с. ИЦиГ СО РАН
Дорошков Алексей Владимирович, к.б.н., с.н.с. ИЦиГ СО РАН
Константинов Дмитрий Константинович, инженер ИЦиГ СО РАН

Информация о гранте:
Грант РНФ 18-14-00293. Название гранта: “Широкомасштабный анализ транскриптомов сельскохозяйственных растений: идентификация новых генов устойчивости к биотическому и абиотическому стрессу и оценка потенциала альтернативной трансляции мРНК.”
Руководитель: Афонников Д.А. 2018 – 2020гг.
Грант РФФИ 20-04-01112. Название гранта “ Системное изучение механизмов морфогенеза растений на основе данных секвенирования транскриптомов одиночных клеток”.
Руководитель: Зубаирова У.С. 2020 – 2022гг.

Научное содержание работы:

1. *Постановка задачи.*

Различия в структуре и функции регуляторных генных сетей (РГС), участвующих в развитии организмов, являются ключом к пониманию видо- и кладо-специфических стратегий эволюции. Даже самая крошечная модификация РГС, контролирующих развитие организма, может привести к существенному изменению сложного морфогенеза организма в целом. Большое разнообразие трихом и их доступность для изучения делает их полезной моделью для изучения молекулярных процессов детерминации судьбы клетки, контроля клеточного цикла, клеточного морфогенеза и дифференцировки. В настоящее время описано большое количество генов, регулирующих морфогенез трихом *A. thaliana*. В данной работе мы исследовали эволюцию РГС, контролирующей формирование трихом, что позволило также оценить ее важность в других процессах развития организма.

2. *Современное состояние проблемы.*

(I) Ряд исследований показал, что ортологи генов образования трихомом *A. thaliana* участвуют в формировании нитей хлопка [1, 2, 3, 4, 5]. Однако ранее было высказано предположение, что у филогенетически удаленных видов трихомы могут развиваться через разные молекулярно-генетические механизмы [6]. Действительно некоторые данные говорят в пользу функциональной диверсификации отдельных регуляторных путей развития трихом. Например, было показано, что эктопическая экспрессия транскрипционного фактора OsTCL1 риса в геноме *A. thaliana* влияет на формирование трихом; однако изменения в экспрессии OsTCL1 в рисе не приводят к каким-либо связанным с трихомами фенотипическим эффектам [7]. Сверхэкспрессия гена GL1, не смотря на то что он является ключевым белком в трихом-специфичном инициаторном комплексе у *A. thaliana*, в табаке не влияет на развитие трихом.

Кроме того, табак имеет пять типов трихом, что также отражает различия в генетических механизмах трихомообразования [6].

С другой стороны, известно, что наросты эпидермальных клеток широко распространены и представляют собой чрезвычайно древние образования. Простые наросты встречаются у водорослей - Chara (Charophytales) и Spirogyra (Zygnematales) [8]. Ризоиды во мхах имеют характерную пространственную закономерность и выполняют функции фиксации в субстрате, участвующем в поглощении воды и питательных веществ [9]. Выявлено, что гены Physcomitrella patens PpRSL1 и PpRSL2 влияют на количество ризоидов на растении [10, 11]. Мутанты Arabidopsis, лишенные функции RHD6 (одного из ключевых генов развития волосков), развивают корневые волоски, если они трансформированы генами PpRSL1 из Physcomitrella. Это указывает на то, что функция белков семейства RSL не была потеряна в течение 420 миллионов лет после расхождения видов [10].

Таким образом, чтобы понять процессы развития и эволюции морфогенеза трихом, нам необходимо объединить данные о белках и их функциях в топологии РГС, контролирующей формирование трихом, для дивергенции каждого из основных таксонов растений, после чего, нам нужно будет связать изменения в топологии РГС с изменениями компонентов этой РГС (отдельных белков).

(II) Функции любого белка являются прямым следствием его химических и физических свойств, которые, в свою очередь, определяются стерическими и физико-химическими требованиями для сворачивания в трехмерную глобулу. Следовательно, ожидается, что замена аминокислоты, взаимодействующей с большим количеством других аминокислот в белковой глобуле, тесно связана с изменениями контекста эпистатических взаимодействий аминокислот в глобуле. Исследования эволюции белков выявили несколько очевидных признаков эпистаза, например, это переключение толерантности к мутациям по мере эволюции белка или, другими словами, вредные мутации в одно эволюционное время становятся безвредными или наоборот [12], аналогичным примером может быть постепенное появление все большего и большего количества ограничений эпистатических взаимодействий в процессе эволюции белков [13, 14].

Несмотря на эти факты, вплоть до настоящего времени, подавляющее большинство доступных процедур реконструкции предковых биологических последовательностей [15, 16] основано на эволюционной обратимости единственной эмпирической матрицы аминокислотных замен (которая применяется ко всем сайтам белка). В сложившейся ситуации, кажется очевидным, что новые программные средства для реконструкции предковых белков (например, ProtASR [17]), адаптированные к 3D структуре белка и стабильности её сворачивания, должны быть наиболее подходящими для анализа. Однако к сожалению, все еще катастрофически не хватает экспериментально разрешенных трехмерных структур белков, что особенно касается растительных организмов. Другим способом учета эпистаза в реконструкции предковых белков является конструирование библиотек предков [18]. Этот подход учитывает известную проблему реконструкции предков - нет никакой гарантии, что алгоритмически реконструированные предковые белки являются

биологически функциональными белками. Недавнее экспериментальное исследование искусственной эволюции белка mRFP1 показывает, что предковые последовательности, полученные с использованием метода максимального правдоподобия, наиболее напоминают естественных предковых белков mRFP1, в то время как лучшие белки, реконструированные с использованием байесовского метода не настолько похожи на этих предков [19]. Тем не менее, с использованием метода максимального правдоподобия может быть получен только один “лучший” предковый белок, что очевидно недостаточно для создания библиотеки предков. Чтобы сделать генерацию библиотек предков достаточно точной, недавно было предложено использовать метод реконструкции «AltAll». Этот подход объединяет все вероятные альтернативные состояния, введенные в один белок, а затем функционально характеризует этот белок набором этих состояний [20, 21]. Было показано, что этот подход значительно улучшает несовершенство отдельных предковых последовательностей, генерируемых байесовским подходом.

Таким образом, лучшее, что мы можем сделать в случае отсутствия трехмерных белковых структур, - это использовать основанный на AltAll подход для создания библиотек предков для последующих исследований эволюции этих белков.

1. Li Y, Shan X, Gao R, Yang S, Wang S, Gao X, Wang L. Two IIIf clade-bHLHs from Freesia hybrida play divergent roles in flavonoid biosynthesis and trichome formation when ectopically expressed in Arabidopsis. Sci Rep. 2016;6:30514.
2. Jaffé FW, Tattersall A, Glover BJ. A truncated MYB transcription factor from Antirrhinum majus regulates epidermal cell outgrowth. J Exp Bot. 2007;58(6):1515–24.
3. Pu L, Li Q, Fan X, Yang W, Xue Y. The R2R3 MYB transcription factor GhMYB109 is required for cotton fiber development. Genetics. 2008;180(2):811–20.
4. Guan X, Yu N, Shangguan X, Wang S, Lu S, Wang L, Chen X. Arabidopsis trichome research sheds light on cotton fiber development mechanisms. Chin Sci Bull. 2007;52(13):1734–41.
5. Guan XY, Li QJ, Shan CM, Wang S, Mao YB, Wang LJ, Chen XY. The HD-zip IV gene GaHOX1 from cotton is a functional homologue of the Arabidopsis GLABRA2. Physiol Plant. 2008;134(1):174–82.
6. Serna L, Martin C. Trichomes: different regulatory networks lead to convergent structures. Trends Plant Sci. 2006;11(6):274–80.
7. Zheng K, Tian H, Hu Q, Guo H, Yang L, Cai L, Wang X, Liu B, Wang S. Ectopic expression of R3 MYB transcription factor gene OsTCL1 in Arabidopsis, but not rice, affects trichome and root hair formation. Sci Rep. 2016;6:19254.
8. Lewis LA, McCourt RM. Green algae and the origin of land plants. Am J Bot. 2004;91(10):1535–56.
9. Sakakibara K, Nishiyama T, Sumikawa N, Kofuji R, Murata T, Hasebe M. Involvement of auxin and a homeodomain-leucine zipper I gene in rhizoid development of the moss Physcomitrella patens. Development. 2003;130(20):4835–46.

10. Menand B, Yi K, Jouannic S, Hoffmann L, Ryan E, Linstead P, Schaefer DG, Dolan L. An ancient mechanism controls the development of cells with a rooting function in land plants. Science. 2007;316(5830):1477–80.
11. Jang G, Yi K, Pires ND, Menand B, Dolan L. RSL genes are sufficient for rhizoid system development in early diverging land plants. Development. 2011;138(11):2273–81.
12. Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. A model of substitution trajectories in sequence space and long-term protein evolution. Mol Biol Evol. 2014;32(2):542–54.
13. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016;25(7):1204–18.
14. Goldstein RA, Pollock DD. Sequence entropy of folding and the absolute rate of amino acid substitutions. Nature ecology & evolution. 2017;1(12):1923.
15. Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AF. Ancestral reconstruction. PLoS Comput Biol. 2016;12(7):e1004763.
16. Merkl R, Sterner R. Ancestral protein reconstruction: techniques and applications. Biol Chem. 2016;397(1):1–21.
17. Arenas M, Weber CC, Liberles DA, Bastolla U. ProtASR: an evolutionary framework for ancestral protein reconstruction with selection on folding stability. Syst Biol. 2017;66(6):1054–64.
18. Gumulya Y, Gillam EM. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the ‘retro’approach to protein engineering. Biochem J. 2017;474(1):1–9.
19. Randall RN, Radford CE, Roof KA, Natarajan DK, Gaucher EA. An experimental phylogeny to benchmark ancestral sequence reconstruction. Nat Commun. 2016;7:12847.
20. Anderson DP, Whitney DS, Hanson-Smith V, Woznica A, Campodonico-Burnett W, Volkman BF, King N, Thornton JW, Prehoda KE. Evolution of an ancient protein function involved in organized multicellularity in animals. elife. 2016;5:e10147.
21. Eick GN, Bridgham JT, Anderson DP, Harms MJ, Thornton JW. Robustness of reconstructed ancestral protein functions to statistical uncertainty. Mol Biol Evol. 2017;34(2):247–61.

3. *Подробное описание работы, включая используемые алгоритмы.*

(I) Реконструкция РГС, контролирующей формирование трихом. По анализу терминов GeneOntology (GO) [1], связанных с образованием трихом у *A. thaliana* (отрицательная регуляция формирования паттернов трихом, разветвление трихом, регуляция морфогенеза трихом, морфогенез трихом, дифференцировка трихом, формирование паттернов трихом) было экстрагировано 90 генов. Этот набор генов и их взаимодействий был дополнен на основе информации из баз данных биологических взаимодействий (например, STRING [2], GeneMania [3]). В ходе работы отбирались гены, которые имели наивысшее количество связей с исходной выборкой генов.

(II) Экстракция белковых последовательностей и филогенетический анализ. Филогенетические отношения между всеми гомологами в каждом семействе белков исследовались на основе полностью секвенированных геномов растений из базы данных PLAZA 3.0 [4]. Идентификация доменов в белках проводилась с использованием базы данных PFam. Белки, не содержащие доменов, были исключены из анализа. Проводилось итеративное выравнивание белков используя mafft 7 [5] с параметрами «--add» «--auto» и «--keeplength». Автоматическая очистка множественных выравниваний белков от неинформативных сайтов (сайт, в котором более 80% белков имеют делецию) была сделана с использованием Python-скрипта. Кроме этого, короткие белки, имеющие более 75% делеций в итоговом выравнивании, были также удалены из анализа. Анализ молекулярной эволюции проводился с помощью конвейера SAMEM v. 0.82 [6]. Построение специфической для семейства белков эволюционно обратимой модели аминокислотных замен на основе множественного выравнивания осуществлялось при помощи Modelestimator [7]. Программа FastTree 2.1.1 [8] использовалась для оценки первичной топологии дерева. Построение конечного филогенетического дерева на основе созданной обратимой модели аминокислотных замен было выполнено при помощи Phyml 3 [9] путем оптимизации топологии первичного дерева и длин ветвей. Чтобы проверить статистическую устойчивость точек ветвления дерева использовалась процедура aLRT. На деревьях, согласно их топологии и OTU (оперативные таксономические единицы), была проведена оценка биоразнообразия, таксон-специфическая оценка устойчивости функций белков и их доменного состава, идентификация поддеревьев разных ортологов и качественная оценка времен диверсификации функций белков-ортологов.

(III)Углубленный анализ эволюции белков РГС, контролирующей формирование трихом. Для углубленного анализа эволюции этой РГС были отобраны 4 семейства белков: EGL3, GLABRA, CPC, TTG1. Множественные выравнивания этих семейств были уточнены при помощи PROMALS [10]. Лучшие обратимые модели аминокислотных замен были выбраны при помощи IQTree 1.5.4 [11]. Первоначальные топологии белкового дерева были скорректированы с использованием видового дерева Viridiplantae из TimeTree DB [12] с использованием программного обеспечения TreeFix v1.1.10 [13], после чего была проведена повторная оптимизация длин ветвей с помощью IQTree и лучших обратимых моделей аминокислотных замен. Байесовский сэмплинг предковых последовательностей в каждом внутреннем узле четырех филогенетических деревьев проводился с использованием PhyloBayes 4.1 [14], эволюционной модели CAT [15] и 6 категорий скоростей эволюции сайтов. Исходная выборка предковых последовательностей в каждом внутреннем узле дерева использовалась для создания полных и усеченных (с использованием модифицированного нами подхода, называемого ‘AltAll * N’) библиотек предковых последовательностей. Наша процедура ‘AltAll * N’ - это итеративное переписывание всех вероятных (с апостериорной вероятностью> 0,1) альтернативных состояний в предковых последовательностях в каждом внутреннем узле дерева. Например, если есть 3 альтернативных состояния в множественном выравнивании в сайте A и 4 альтернативных состояния в сайте B предкового узла X, то мы должны переписать предковую последовательность узла X 4 раза, чтобы получить 4 альтернативных предка в узле X: a)

последовательность, состоящая из лучших состояний в A- и B-сайтах, б) последовательность со вторыми по вероятностью состояниями сайтов A и B, в) последовательность с третьими по вероятности состояниями сайтов A и B и г) последовательность с третьим по вероятности состоянием в сайте A и четвертым - в сайте B.

Эпистатическая конверсия белков в ходе их эволюции детектировалась при помощи анализа отклонений в скорости эволюции белка от белок-специфической эволюционно обратимой модели аминокислотных замен на каждой из ветвей белкового дерева (1) и простым сравнением длин внутренних ветвей с учетом данных о структуре белка (2). (1) Чтобы сравнить специфические для каждой внутренней ветви скорости аминокислотных замен с общими по дереву эволюционно обратимыми скоростями мы использовали полные библиотеки предковых последовательностей (см. выше). Для этого мы: а) реконструировали специфическую для белка обратимую во времени модель замен аминокислот при помощи Modelestimator [7]; б) для каждой возможной замены каждого внутреннего узла дерева рассчитали, d = PPa * PPb * 2 * NC, где PPa и PPb - апостериорные вероятности аминокислот a и b, a не равно b, NC = 1 / (1 + e ^ (200 * RFab)), RFab - относительная скорость замены ab в специфичной для белка обратимой модели аминокислотных замен; в) суммировали d по всем сайтам выравнивания в каждом внутреннем узле дерева и вычислили натуральные логарифмы этих сумм; г) провели непараметрическое сравнение (по процентилям) лог-сумм (см (в)) по всему дереву с целью выявления ветвей с максимальной суммой (ветвей с признаками эпистатических преобразований). (2) Чтобы сравнить ветвь-специфические скорости структурных изменений, мы использовали усеченные библиотеки ‘AltAll * N’. Для этого, мы: а) реконструировали вторичные структуры, меры близости аминокислоты к поверхности глобулы и меры неупорядоченности (disorder) для каждого аминокислотного остатка каждой альтернативной предковой последовательности в каждом внутреннем узле дерева с использованием конвейера RaptorX_Property Fast [16]; б) вычислили частоты изменений этих мер (см. (а)) между всеми альтернативными предковыми последовательностями соседних узлов внутренних ветвей дерева; в) провели непараметрическое сравнение (по процентилям) вышеупомянутых частот изменений по всему дереву с целью выявления ветвей с максимальными структурными изменениями (ветвей с признаками эпистатических преобразований).

1. Gene Ontology Consortium. Gene ontology consortium: going forward. Nucleic Acids Res. 2014;43(D1):D1049–56.
2. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids res. 2016;45(D1):D362–68.
3. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214–W220.

4. Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K. PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. 2014;43(D1):D974–81.
5. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
6. Gunbin KV, Suslov VV, Genaev MA, Afonnikov DA. Computer system for analysis of molecular evolution modes (SAMEM): analysis of molecular evolution modes at deep inner branches of the phylogenetic tree. In silico biology. 2012;11(3, 4):109–23.
7. Arvestad L. Efficient methods for estimating amino acid replacement rates. J Mol Evol. 2006;62(6):663–73.
8. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
9. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.
10. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 2007;23(7):802–808.
11. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–274.
12. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34(7):1812–1819.
13. Wu YC, Rasmussen MD, Bansal MS, Kellis M. TreeFix: statistically informed gene tree error correction using species trees. Syst Biol. 2013;62(1):110–120.
14. Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 2007;7 Suppl 1(Suppl 1):S4.
15. Quang le S, Gascuel O, Lartillot N. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 2008;24(20):2317–2323.
16. Ma J, Wang S, Zhao F, Xu J. Protein threading using context-specific alignment potential. Bioinformatics. 2013;29(13):i257–65.

4. *Полученные результаты*

В генной сети развития трихом были выявлены функциональные блоки: гормон-чувствительные регуляторы, инициаторный комплекс и его ингибиторы, гены цитоскелета, гены клеточного цикла и др. Мы представили список генов-кандидатов, ответственных за развитие трихом в широком спектре видов растений.

Было показано, что основные гены инициаторного комплекса РГС, контролирующей формирование трихом являются эволюционно старыми и они, вероятно, имели одну функцию в предке всех сосудистых растений, и отвечали за образование простого

одномерного паттерна опушения растения. Генные сети растений с более сложным паттерном опушения прошли через огромное количество событий дупликации отдельных генов, которые, вероятно, сыграли решающую роль в формировании сложных паттернов опушения. Тем не менее, паттерны опушения однодольных и двудольных покрытосеменных растений образуются равными по сложности генными сетями.

На основе результатов углубленного филогенетического анализа мы предположили, что дивергенция и/или специализация компонентов РГС, контролирующей формирование трихом, связаны напрямую с эволюционным появлением крупных таксонов растений. Полученная информация об основных белковых компонентах РГС дала возможность предсказать точки переключения в эволюции и функционировании генных сетей.

5. *Иллюстрации, визуализация результатов.*

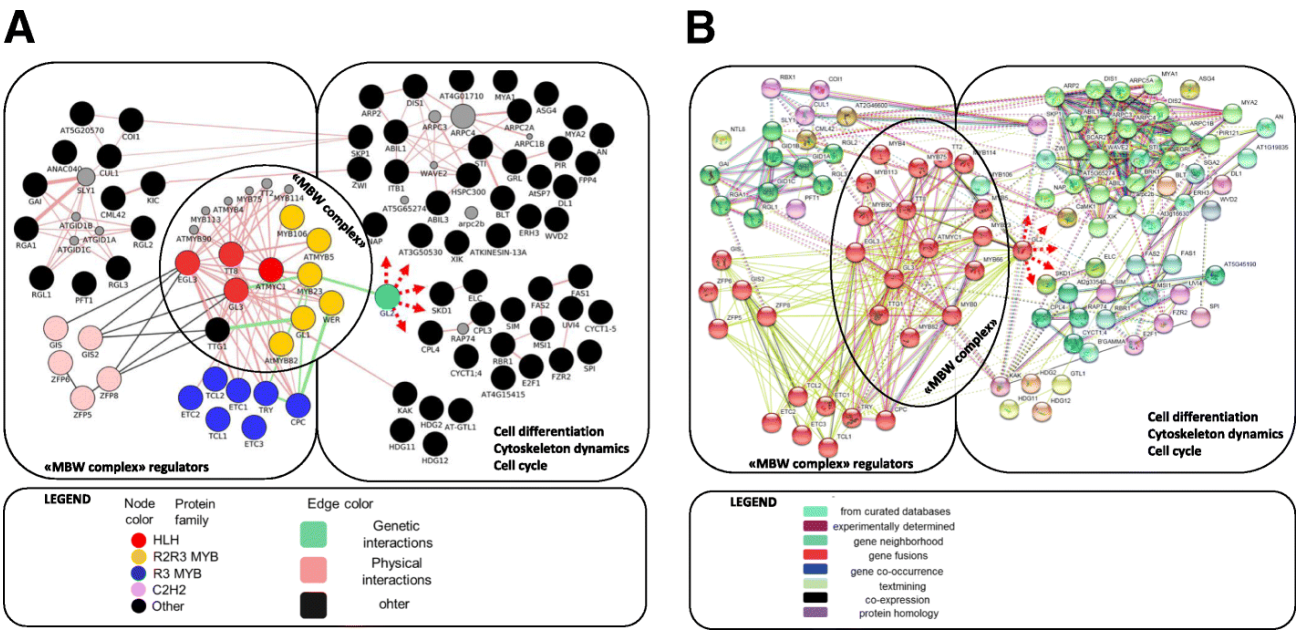

Рис. 1. РГС, контролирующая формирование трихом, реконструирована с использованием плагина GeneMANIA (a) и STRING (b).

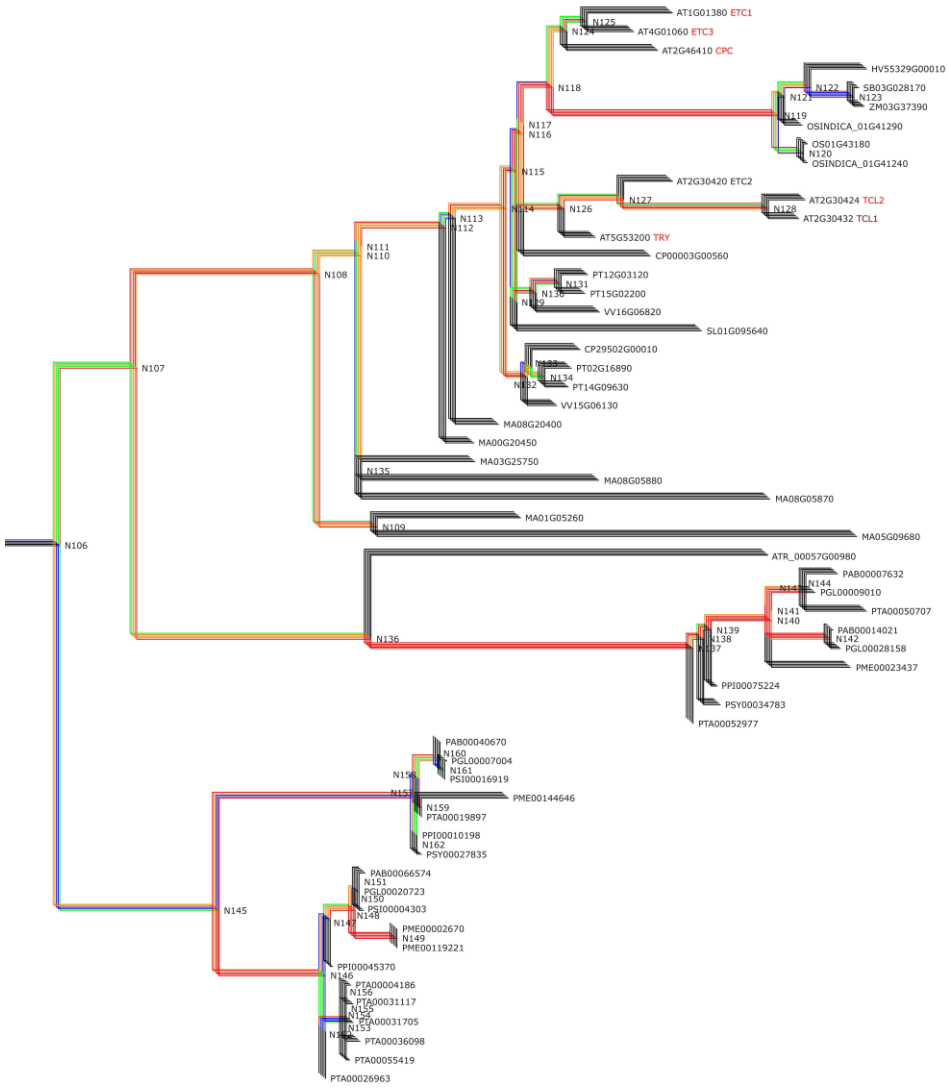

Рис. 2. Эволюционные изменения структуры семейства белков R3-MYB – линии (на ветвях дерева) справа налево: неупорядоченность (2 типа остатков), вторичная структура (3 типа), вторичная структура (8 типов), поверхность глобулы (3 типа остатков), редкие (по сравнению с обратимой моделью аминокислотных замен) аминокислотные замены. Цветовая схема: черный - внешняя ветвь (не анализируются); цвета внутренних ветвей определяют квартиль скоростей эволюции: синий - Q1; зеленый - Q2; оранжевый –Q3; красный –Q4.

6. *Эффект от использования кластера в достижении целей работы.*

Проведение компьютерного анализа эволюции белков с использованием рабочих станций невозможно как вследствие большого объема требуемой оперативной памяти, так и по времени счета. Поэтому использование кластера является определяющим для успешного достижения целей работы.

Перечень публикаций, содержащих результаты работы

Doroshkov AV, Konstantinov DK, Afonnikov DA, Gunbin KV. The evolution of gene regulatory networks controlling Arabidopsis thaliana L. trichome development. BMC Plant Biol. 2019 Feb 15;19(Suppl 1):53. doi: 10.1186/s12870-019-1640-2. IF=3.670; Q1 in Plant Science; Scopus percentile - 92%