БИОФИЗИКА, 2020, том 65, № 2, с. 263–275

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.32

АНАЛИЗ ЭКСПРЕССИИ ГЕНОВ ЦВЕТЕНИЯ В СОРТЕ НУТА CDC FRONTIER МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

© 2020 г. Б.С. Подольный*, В.В. Гурский, М.Г. Самсонова***

**Санкт-Петербургский политехнический университет Петра Великого, 195251, Санкт-Петербург, Политехническая ул., 29*

***Физико-технический институт им. А.Ф. Иоффе, 194021, Санкт-Петербург, Политехническая ул., 26*

E-mail: m.samsonova@spbstu.ru

Поступила в редакцию 16.12.2019 г.
После доработки 16.12.2019 г.
Принята к публикации 24.12.2019 г.

Представлены результаты анализа динамики экспрессии генов, контролирующих переход к цветению у сорта нута CDC Frontier. Построено несколько версий модели, предсказывающей динамику экспрессии пяти генов цветения по данным экспрессии их регуляторов. Модели обучены с помощью метода случайных лесов на опубликованных ранее данных экспрессии десяти генов цветения для условий короткого и длинного дня выращивания растений. Полученные модели правильно предсказывают динамику средних уровней экспрессии в условиях длинного дня. Показано, что модели для CDC Frontier в целом воспроизводят регуляторные взаимодействия между ключевыми генами, описанные для модельного растения *Arabidopsis thaliana*. На основе анализа моделей высказано предположение о том, что данные короткого дня и данные длинного дня содержат качественно разную информацию, что может определяться функционированием разных регуляторных модулей в разных условиях. Среди регуляторов генов идентичности цветковых меристем *AP1* и *LFY* модели предсказывают лидирующую роль гена *FTa3* как активатора и гена *TFL1c* как репрессора в условиях длинного дня.

Ключевые слова: нут, CDC Frontier, генная сеть цветения, алгоритм случайного леса.

DOI: 10.31857/S0006302920020076

Исследование процессов, лежащих в основе принятия растением решения о переходе к цветению, и, в частности, контроль времени инициации цветения сельскохозяйственных культур имеет важное значение для разработки сценариев эффективной адаптации культур к вариабельности сезонов выращивания [1–3]. Эта задача особенно актуальна в условиях меняющегося климата [4, 5]. В число важных сельскохозяйственных культур входит нут (*Cicer arietinum*), представитель семейства бобовых. К настоящему времени собрано большое количество информации о сходствах и отличиях в механизмах зацветания у бобовых и у модельного организма *Arabidopsis thaliana* [6–12]. Основные гены, контролирующие время перехода к цветению и наиболее полно описанные в *Arabidopsis*, консервативны и присутствуют во многих растениях [6, 13]. Одним из существенных отличий бобовых от *Arabidopsis* является более сложная (составная) архитектура соцветия, включающая в себя вторичные меристемы соцветия, что подразумевает наличие дополнительных генов, контролирующих идентичность этих меристем [7, 8]. Другим отличием является то, что геномы представителей семейства бобовых содержат несколько гомологов ключевых генов из *Arabidopsis*. В частности, активатор цветения *FLOWERING LOCUS T* (*FT*) и репрессор цветения *TERMINAL FLOWER 1* (*TFL1*) присутствуют в нуте в виде соответственно пяти и двух гомологов (гены *FTa1*, *FTa2*, *FTa3*, *FTb*, *FTc* и гены *TFL1a*, *TFL1c*) [14].

Согласно классической схеме регуляции перехода к цветению в модельном растении *Arabidopsis thaliana*, продукты экспрессии *FT*-генов в листьях доставляются в апикальные меристемы растения, где они активируют экспрессию генов идентичности цветковых меристем *LEAFY* (*LFY*) и *APETALA1* (*AP1*) [15]. Экспрессия этих генов подавляется белками *TFL1*-генов с целью поддержания локального вегетативного состояния. Белки FT и TFL1 осуществляют свою регуляторную функцию, образуя комплекс с транскрипционным фактором FD или его гомологом в случае бо-

Сокращения: LD – длинный световой день, SD – короткий световой день, НСКО – нормализованная среднеквадратичная ошибка.

бовых [15, 16]. Высокий уровень AP1 можно считать маркером инициации цветения. Были разработаны несколько динамических моделей этой генной сети в *Arabidopsis*, предсказывающих время инициации цветения по уровням экспрессии ключевых генов [15, 17].

Экспрессия десяти генов (пять *FT*-генов, два *TFL1*-гена, а также гены *FD*, *LFY*, *AP1*), вовлеченных в процесс зацветания, была измерена в двух сортах нута (ICCV 96029 и CDC Frontier) в процессе выращивания растений в условиях короткого и длинного дня [14]. ICCV 96029 является слабо чувствительным к фотопериоду и самым раннецветущим сортом нута. CDC Frontier представляет собой чувствительный к фотопериоду сорт, для которого был построен референсный геном [18]. На полученных данных по экспрессии генов цветения в этих сортах была построена динамическая модель на основе дифференциальных уравнений, которая в случае сорта ICCV 96029 корректно воспроизводила динамику экспрессии всех генов и позволила протестировать несколько гипотез о функционировании гомологов *FT* и *TFL1* в этом сорте [19]. Однако эта модель не смогла воспроизвести динамику экспрессии в сорте CDC Frontier, из-за чего оставалось не ясным, является ли эта неудача в моделировании следствием специфики модельного подхода или же связана с особенностями данных для этого сорта нута.

В представленной работе используется другой подход к моделированию динамики экспрессии генов цветения в CDC Frontier, основанный на машинном обучении. Используя результаты обучения методом случайного леса, предложен способ построения динамических предсказаний уровней экспрессии в каждый день на основе данных или предсказаний модели в предыдущий день. Полученная таким способом динамика экспрессии в модели исследуется при разных условиях обучения модели, а также при выключении отдельных генов.

ДАННЫЕ И МОДЕЛЬ ЭКСПРЕССИИ

Модель обучалась на опубликованных ранее динамических данных экспрессии генов цветения в сорте нута CDC Frontier [14] (данные в виде csv-файлов можно найти по адресу: http://doi.org/10.5281/zenodo.1451748). Рассмотренные гены цветения включают в себя пять гомологов гена *FT* (*FTa1*, *FTa2*, *FTa3*, *FTb*, *FTc*), два гомолога гена *TFL1* (*TFL1a*, *TFL1c*), ген *FD*, а также гены идентичности цветковых меристем *LFY* и *AP1*. Данные представляют собой средние значения и стандартные отклонения уровней экспрессии, измеренных с 9-х по 43-и сутки после посева с двух- или трехдневным интервалом в условиях длинного (LD) или короткого (SD) светового дня. Появление развивающихся цветковых бутонов впервые было детектировано на 31-е сутки развития в условиях короткого дня и на 32-е сутки в условиях длинного дня [14].

Поскольку переход к цветению должен быть связан с ростом экспрессии *AP1*, несколько поздних наблюдений, соответствующих резкому падению уровня экспрессии *AP1*, были исключены. Данное падение может быть связано с изменением механизмов регуляции после начала цветения. Таким образом, для дальнейшего анализа из данных длинного дня был исключен один последний день наблюдений, а из данных короткого дня — три последних дня.

Общая схема моделирования описана на рис. 1. Далее детально описаны этапы моделирования, проиллюстрированные на рисунке.

Для построения модели методами машинного обучения необходимы обучающая и тестовые выборки данных. Для этого на первом этапе исходные данные экспрессии мультиплицировались и создавались искусственные выборки значений уровней экспрессии всех генов для каждого дня развития растения. Такие выборки генерировались множественным сэмплированием из усеченных нормальных распределений, определенных на положительной вещественной полуоси, где исходные средние значения и стандартные отклонения уровней экспрессии генов были использованы в качестве параметров распределений.

Были рассмотрены два типа моделей: модели первого типа обучались совместно по LD- и SD-данным, модели второго типа – только по LD-данным. Для моделей первого типа в каждый день всего было сгенерировано по 325 значений уровней экспрессии для LD- и SD-условий с целью обучения и по 125 значений для тестирования. Для обучения и тестирования моделей второго типа было сгенерировано соответственно по 750 и 250 значений.

В исходных данных измерения для LD и SD не всегда проводили в одинаковые дни, при этом для построения моделей, учитывающих оба типа условий, необходимо, чтобы обучающие и тестовые выборки в каждый день содержали значения для обоих условий выращивания. Для заполнения пропущенных значений была проведена линейная интерполяция LD-значений на те дни, когда измерения проводили только для SD, и наоборот. Для интерполяции из выборки искусственных данных случайным образом выбирали пары измерений из соседних дней.

Уровни экспрессии *FT*-генов в данных соответствуют экспрессии этих генов в листьях, тогда как уровни экспрессии остальных генов, включая потенциальные гены-мишени *FT*-генов (*AP1* и *LFY*), соответствуют экспрессии в апикальной

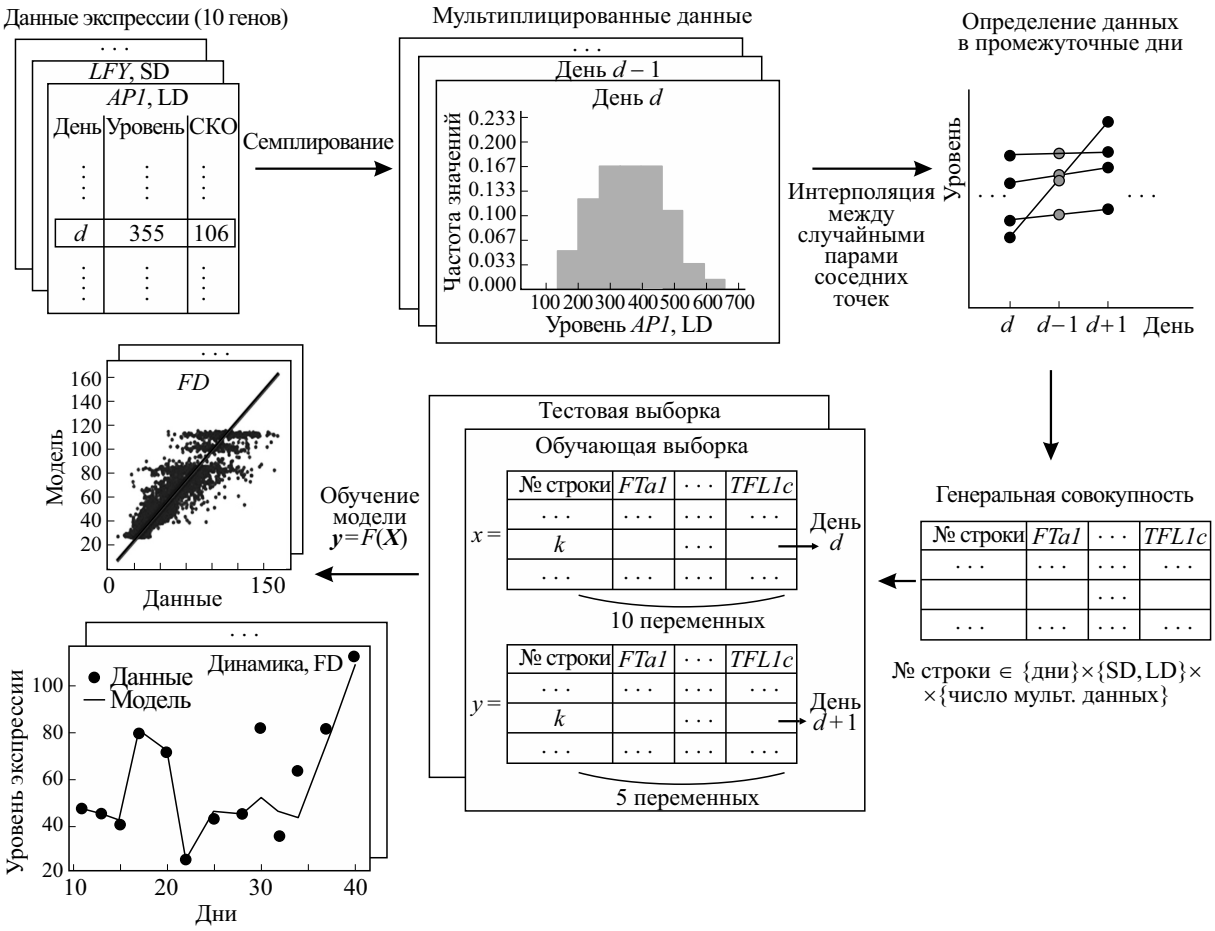

Рис. 1. Схема моделирования. Детальное описание отдельных этапов приведено в тексте.

меристеме. Поскольку модель описывает регуляцию в апикальной меристеме, значения экспрессии *FT*-генов сдвигались с помощью линейной интерполяции на фиксированное время τ, необходимое для транспорта белков FT из листьев в апекс. Исходя из полученной ранее оценки этого времени для *Arabidopsis*, было принято значение τ = 0.5 суток [17]. Для сдвига первого дня измерений (9-е сутки после посева) был добавлен нулевой день с нулевыми значениями уровней экспрессии для всех генов и проведена интерполяция между нулевым и первым днем наблюдений.

В результате такой обработки данных была собрана генеральная совокупность данных, содержащая уровни экспрессии десяти генов (в мультиплицированной форме) для каждого дня и для двух условий выращивания. Из этой совокупности случайным образом делали выборки для обучения и тестирования модели (рис. 1). Модель строилась как регрессионная модель, предсказывающая уровни экспрессии пяти целевых генов (т. е. генов, экспрессирующихся в апикальной меристеме) в текущий день по значениям уровней экспрессии всех десяти генов (т. е. дополнительно с учетом *FT*-генов, экспрессирующихся в листьях) в предыдущий день: $\boldsymbol{y} = F(\boldsymbol{X})$, где $\boldsymbol{X}$ – вектор, содержащий уровни экспрессии генов *FTa1*, *FTa2*, *FTa3*, *FTb*, *FTc*, *AP1*, *LFY*, *FD*, *TFL1a*, *TFL1c* в день d; $\boldsymbol{y}$ – вектор, содержащий уровни экспрессии генов *AP1*, *LFY*, *FD*, *TFL1a*, *TFL1c* в день $d + 1$. Таким образом, обучающая и тестовая выборки представляли собой наборы факторов – десятимерных векторов уровней экспрессии генов в определенный день и откликов – пятимерных векторов уровней экспрессии целевых генов на следующий день.

АЛГОРИТМ И ПАРАМЕТРЫ ОБУЧЕНИЯ МОДЕЛИ

Для построения и обучения модели использовалась реализация алгоритма случайного леса «*RandomForestRegressor*» из пакета *sklearn* [20]. Для нахождения оптимальной модели производилась настройка следующих гиперпараметров по соответствующим наборам значений:

– количество деревьев: {20, 50, 100, 300},

Таблица 1. Взаимодействия, учтенные в усеченных моделях, согласно генной сети зацветания у *Arabidopsis* [15]

Регуляторы	*FT*-гены	*AP1*	*LFY*	*TFL1*-гены	*FD*
Мишени	*AP1*, *LFY*	*LFY*, *TFL1*-гены	*AP1*, *FD*	*AP1*, *LFY*	*AP1*, *LFY*

– минимальное число значений в листе дерева: {10, 20, 50, 100},

– максимальная глубина дерева: {5, 8, 20, без ограничений}.

Для каждого фиксированного значения каждого гиперпараметра проводилось обучение модели на обучающей выборке данных с пятикратной кроссвалидацией, в ходе которого минимизировалось среднеквадратичное отклонение (СКО) вектора откликов y в модели от данных. В результате перебора всех значений гиперпараметров выбиралась модель, дающая минимальное СКО на валидационных выборках.

Были построены два типа моделей вида $\boldsymbol{y} = F(\boldsymbol{X})$. В моделях первого типа (далее — «полные» модели) не делалось никаких предположений о характере регуляции между генами, являющимися входными факторами в модели и входящими в вектор $\boldsymbol{X}$, и целевыми генами, входящими в вектор откликов $\boldsymbol{y}$ модели. Математически это выражается в том, что отображение F априори (т. е. до обучения) включает в себя влияние каждого компонента вектора $\boldsymbol{X}$ на каждый компонент вектора $\boldsymbol{y}$. При построении моделей второго типа (далее – «усеченные» модели) использовалось предположение о том, что исследуемая генная сеть подобна соответствующей генной сети у *Arabidopsis thaliana* [15, 19]. В усеченных моделях отображение F связывает только те пары фактор-отклик из векторов $\boldsymbol{X}$ и $\boldsymbol{y}$, которые соответствуют взаимодействиям из генной сети у Arabidopsis (табл. 1), при этом характер взаимодействия (активация или репрессия) априори не постулируется и определяется в результате обучения модели.

Полная и усеченная модели обучались на данных двух типов – либо на полной выборке данных, включающей в себя данные по экспрессии для двух условий (LD+SD), либо на данных только для длинного дня (LD). Таким образом, всего было построено и обучено четыре модели: полная модель на данных LD+SD, полная модель на данных LD, усеченная модель на данных LD+SD и усеченная модель на данных LD. В табл. 2 приведены значения гиперпараметров, найденные в ходе обучения каждой из моделей.

СПОСОБЫ ВЫЧИСЛЕНИЯ ДИНАМИЧЕСКИХ ПРЕДСКАЗАНИЙ МОДЕЛИ

После обучения модели F на выборке, содержащей данные сразу для всех дней выращивания, необходимо создать процедуру генерации предсказаний модели в динамике. Были рассмотрены два способа генерации динамики экспрессии целевых генов в модели. В первом способе предсказание $\boldsymbol{y}(d)$ для экспрессии пяти целевых генов в любой день d вычисляли по данным экс-

Таблица 2. Значения гиперпараметров, соответствующие лучшим результатам обучения рассмотренных моделей

Модель		Количество деревьев	Минимальное число значений в листе	Максимальная глубина дерева
Полная LD+SD		100	10	Без ограничений
Полная LD		300	10	Без ограничений
Усеченная LD + SD	*AP1*	300	10	Без ограничений
	FD	100	100	5
	LFY	300	10	Без ограничений
	TFL1a	100	100	5
	TFL1c	100	100	5
Усеченная LD	*AP1*	300	10	Без ограничений
	FD	300	100	5
	LFY	300	10	Без ограничений
	TFL1a	300	100	5
	TFL1c	100	100	5

Таблица 3. Результаты обучения и тестирования полной модели на выборке LD + SD

	НСКО		Коэффициент корреляции	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
AP1	0.038	0.062	0.97	0.92
FD	0.062	0.074	0.92	0.89
LFY	0.058	0.084	0.89	0.76
TFL1a	0.051	0.062	0.94	0.91
TFL1c	0.064	0.073	0.91	0.88

Примечание. Для каждого гена отдельно приведены оценки качества предсказаний модели на обучающей и тестовой выборках, в терминах нормализованной среднеквадратичной ошибки (НСКО) и коэффициента корреляции Пирсона. НСКО вычисляли нормализацией СКО на максимальный уровень экспрессии соответствующего гена.

прессии $\boldsymbol{X}(d-1)$ всех десяти генов в предыдущий день $d-1$: $\boldsymbol{y}(d) = F(\boldsymbol{X}(d-1))$, $d = 2, 3, \ldots$. Во втором способе динамика предсказаний $\boldsymbol{y}(d)$ строилась согласно следующему алгоритму:

1. По исходным данным $\boldsymbol{X}(1)$ первого дня ($d = 1$) вычисляли предсказания экспрессии пяти целевых генов для второго дня ($d = 2$): $\boldsymbol{y}(2) = F(\boldsymbol{X}(1))$.

2. Во второй день ($d = 2$) строился вектор $\boldsymbol{X}_{\boldsymbol{y}}(2)$, равный вектору $\boldsymbol{X}(2)$ данных экспрессии для этого дня, в котором уровни экспрессии пяти целевых генов заменялись на соответствующие уровни экспрессии из предсказания $\boldsymbol{y}(2)$ из предыдущего пункта. Таким образом, в векторе $\boldsymbol{X}_{\boldsymbol{y}}(\mathrm{d})$ из данных берутся только уровни экспрессии пяти *FT*-генов.

3. Предсказания для третьего дня ($d = 3$) вычисляли по вектору $\boldsymbol{X}_{\boldsymbol{y}}(2)$ из предыдущего пункта. Таким образом, для любого дня, начиная с третьего, динамику экспрессии пяти целевых генов вычисляли по формуле: $\boldsymbol{y}(d) = F(\boldsymbol{X}_{\boldsymbol{y}}(d-1))$, $d \geqslant 3$.

Описанные два способа вычисления динамических предсказаний модели можно сравнить с решением модели на основе дифференциальных уравнений. Первый из описанных способов генерации динамических предсказаний модели аналогичен использованному ранее подходу к моделированию динамики экспрессии генов цветения в сорте нута ICCV 96029 [19]. В этом подходе динамические предсказания модели получались путем решения дифференциальных уравнений модели, в правой части которых вместо динамических переменных (уровней экспрессии) подставлялись экспериментальные данные. Такие решения должны рассматриваться лишь как аппроксимация истинного решения модельных уравнений. Второй способ генерации динамики в модели, описанный выше, соответствует такому истинному решению модели, сформулированной в терминах дифференциальных уравнений.

Для выяснения характера влияния отдельных генов на генную сеть с помощью модели симулировались нокауты этих генов. При симуляции нокаутов вычислялись динамические предсказания модели вторым способом, однако при этом уровни экспрессии нокаутированных генов во все дни фиксировали на нулевых значениях.

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПОЛНОЙ МОДЕЛИ

На первом этапе была обучена полная модель на совместной выборке, включающей данные для условий длинного светового дня и короткого светового дня (модель «Полная LD+SD» из табл. 2). Результирующая модель примерно одинаково хорошо описывает уровни экспрессии всех генов, и качество предсказаний модели примерно одинаково на обучающей и тестовой выборках (табл. 3).

Однако полученная таким образом модель имеет ряд недостатков. Динамические предсказания модели на усредненных входных данных, вычисленные с учетом данных на предыдущем дне (первый способ, описанный выше), хорошо соответствуют данным длинного дня (черные кривые на рис. 2). Средний по генам коэффициент корреляции между предсказаниями модели и данными в динамике равен $r = 0.84$ для LD и $r = 0.74$ для SD. Однако динамика усредненной экспрессии в модели при условиях LD значительно ухудшается при вычислении с учетом откликов модели на предыдущем дне (второй способ, описанный выше) (красные кривые на рис. 2; средний $r = 0.23$). В частности, уровень экспрессии *AP1* растет недостаточно сильно (рис. 2), и в целом падает корреляция между предсказаниями модели и данными. В условиях SD недостаток активации *AP1* проявляется в модели при любом способе вычисления динамики экспрессии, а для второго способа для всех генов, кроме *AP1*, в модели наблюда-

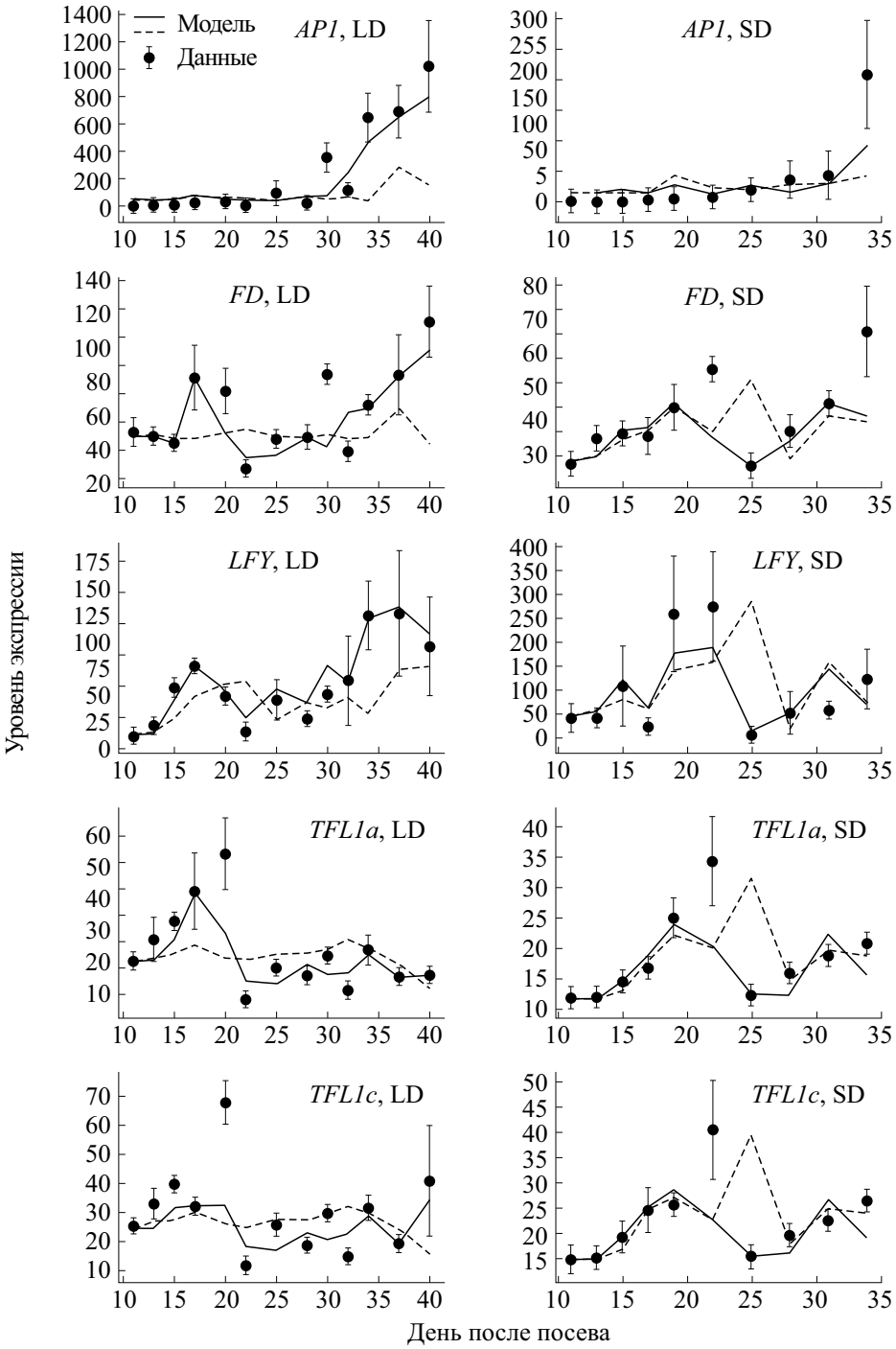

Рис. 2. Динамические предсказания полной модели, обученной на выборке LD+SD, в сравнении со средними уровнями экспрессии в данных. Черная кривая описывает динамику экспрессии в модели, вычисленную первым способом, описанным в тексте (предсказания в каждый день зависят от данных в предыдущий день). Пунктирная кривая показывает динамику экспрессии в модели, вычисленную вторым способом (предсказания в каждый день зависят от данных FT-генов и предсказаний модели для остальных генов в предыдущий день).

ется аномальное повышение экспрессии на 25-й день (рис. 2); также для второго способа корреляция между предсказаниями модели и данными низкая (средние $r = 0.74$ для первого способа и $r = 0.24$ для второго способа вычисления динамики в условиях SD).

Из этого анализа можно сделать предположение о том, что модель не восприимчива к данным

Таблица 4. Результаты обучения и тестирования полной модели на LD-выборке

	НСКО		Коэффициент корреляции	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
AP1	0.042	0.073	0.96	0.90
FD	0.061	0.075	0.89	0.86
LFY	0.084	0.111	0.85	0.81
TFL1a	0.058	0.071	0.91	0.89
TFL1c	0.063	0.081	0.90	0.88

для условий короткого дня, и поэтому обучение модели на данных для LD и SD одновременно привело к недостаточно качественной динамике в модели и для условий длинного дня. В рамках такого предположения полная модель была обучена только на данных длинного дня (модель «Полная LD» из табл. 2). Результирующая модель показала хорошую степень близости к данным на обучающей и тестовой выборках (табл. 4). При этом, в отличие от случая обучения на объединенной выборке LD+SD, теперь модель хорошо предсказывает динамику усредненных уровней экспрессии всех генов для обоих способов вычисления такой динамики (рис. 3; средний $r = 0.99$ для первого способа и $r = 0.89$ для второго способа вычисления динамики).

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ УСЕЧЕННОЙ МОДЕЛИ

На следующем этапе была обучена усеченная модель, в которой явно учтены (как описано выше) только те взаимодействия между генами, которые присутствуют в генной сети зацветания у *Arabidopsis thaliana* (табл. 1) [15, 19]. Такая задача обусловлена общим предположением о консервативности этой генной сети у растений и, следовательно, направлена на проверку того, могут ли исследуемые данные экспрессии генов цветения у сорта нута CDC Frontier быть результатом функционирования в точности такой генной сети. Как и полная модель, усеченная модель обучалась отдельно на выборке LD+SD и только на LD-выборке.

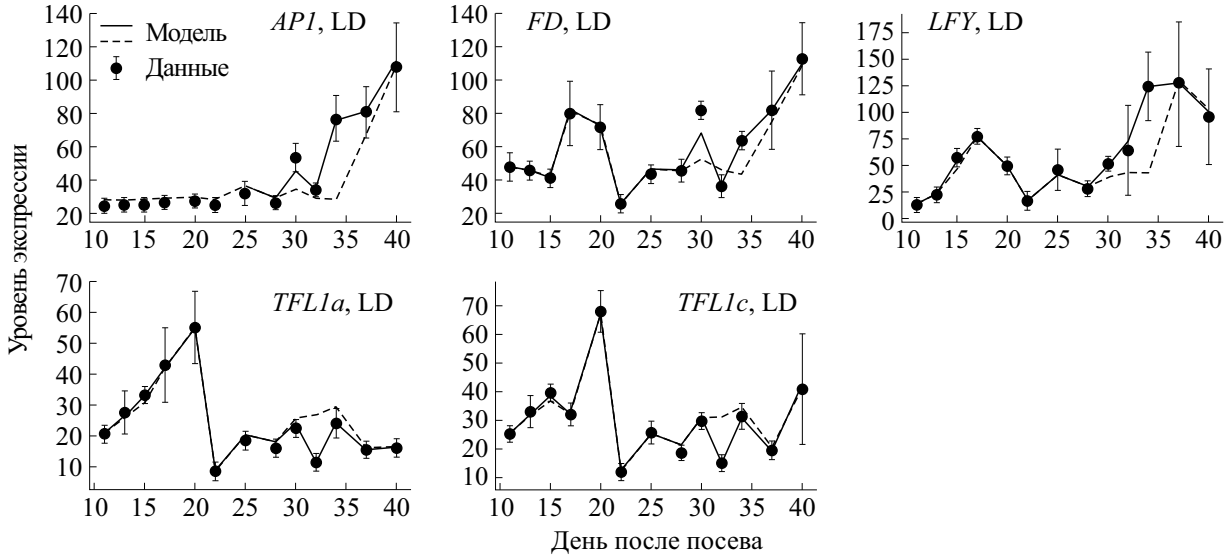

Рис. 3. Динамические предсказания полной модели, обученной на LD-выборке, в сравнении со средними уровнями экспрессии в данных. Все обозначения соответствуют рис. 2.

Таблица 5. Результаты обучения и тестирования усеченной модели на выборке LD+SD

	НСКО		Коэффициент корреляции	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
AP1	0.050	0.066	0.94	0.91
FD	0.141	0.145	0.44	0.44
LFY	0.067	0.086	0.85	0.74
TFL1a	0.138	0.143	0.30	0.30
TFL1c	0.148	0.151	0.27	0.25

Усеченная модель, обученная на данных для условий длинного светового дня и короткого светового дня (модель «Усеченная LD+SD» из табл. 2) показала достаточно хорошие корреляции на тестовой выборке для генов *AP1* и *LFY* и плохие корреляции для остальных генов (табл. 5). Анализ динамических предсказаний модели на усредненных данных показал значительные дефекты в динамике экспрессии (рис. 4). В частности, в отличие от данных, экспрессия *AP1* в модели падает в последний день наблюдений при условиях LD и демонстрирует повышенный уровень при условиях SD. В случае гена *LFY* динамика средней экспрессии достаточно хорошая в обоих условиях роста, но только в случае первого способа вычисления динамики в модели, тогда как для второго способа динамические предсказания в модели значительно ухудшаются (рис. 4). Остальные гены демонстрируют еще более значительные дефекты.

Гипотеза о плохой репрезентативности данных для условий SD была исследована также в контексте усеченной модели, по аналогии с тем, как это было сделано с полной моделью. Для этого усеченная модель была обучена только на LD-данных (модель «Усеченная LD» из табл. 2). Так же, как и в случае обучения на выборке LD+SD, в результате были получены достаточно высокие коэффициенты корреляции между предсказаниями модели и данными на тестовой выборке только для экспрессии генов *AP1* и *LFY* (табл. 6). Анализ динамических предсказаний для средних уровней экспрессии в модели выявил дефекты в динамике, аналогичные описанным для усеченной модели, обученной на выборке LD+SD (рис. 5). Таким образом, в отличие от полной модели, в случае усеченной модели исключение данных короткого дня из обучающей выборки качественно не улучшило результаты обучения.

АНАЛИЗ НОКАУТОВ ГЕНОВ В МОДЕЛЯХ

Для определения характера влияния отдельного гена на экспрессию других в моделях вычисляли предсказания в условиях, когда этот ген выключен («нокаутирован»). Для этого сравнивались площади под графиком динамики экспрессии гена-мишени в модели для дикого типа и при нокауте. Увеличение этой площади при нокауте гена-регулятора соответствует репрессии (прямой или через другие регуляторные пути в генной сети) со стороны гена-регулятора, а уменьшение соответствует активации. Визуализация результатов таких вычислений для всех генов показывает, что модели воспроизводят большую часть регуляторных взаимодействий из генной сети для *Arabidopsis* (рис. 6).

Основной интерес представляет влияние потенциальных регуляторов на гены идентичности цветковых меристем *AP1* и *LFY*, поскольку экспрессия этих генов является маркером инициации цветения. Во всех моделях *TFL1*-гены оказывают репрессивное действие на эти гены в условиях LD, в то время как в условиях SD их влияние не столь однозначно (рис. 6). Среди *FT*-генов только *FTa3* проявляет себя как активатор *AP1* и *LFY* во всех моделях и в обоих условиях роста (LD и SD). Кроме того, *FTa3* является сильнейшим активатором для этих генов-мишеней среди всех *FT*-генов в условиях LD. Поскольку полная модель, обученная на LD-выборке, демонстрирует наиболее адекватную динамику средней экспрессии (рис. 3), для этой модели на рис. 7 показаны примеры динамики экспрессии гена *AP1* при выключении некоторых генов. В частности, *AP1* в этой модели очень слабо меняется при выключении всех *FT*-генов, кроме *FTa3*, и практически перестает экспрессироваться при выключении *FTa3* (рис. 7а). Также из двух репрессоров (*TFL1a* и

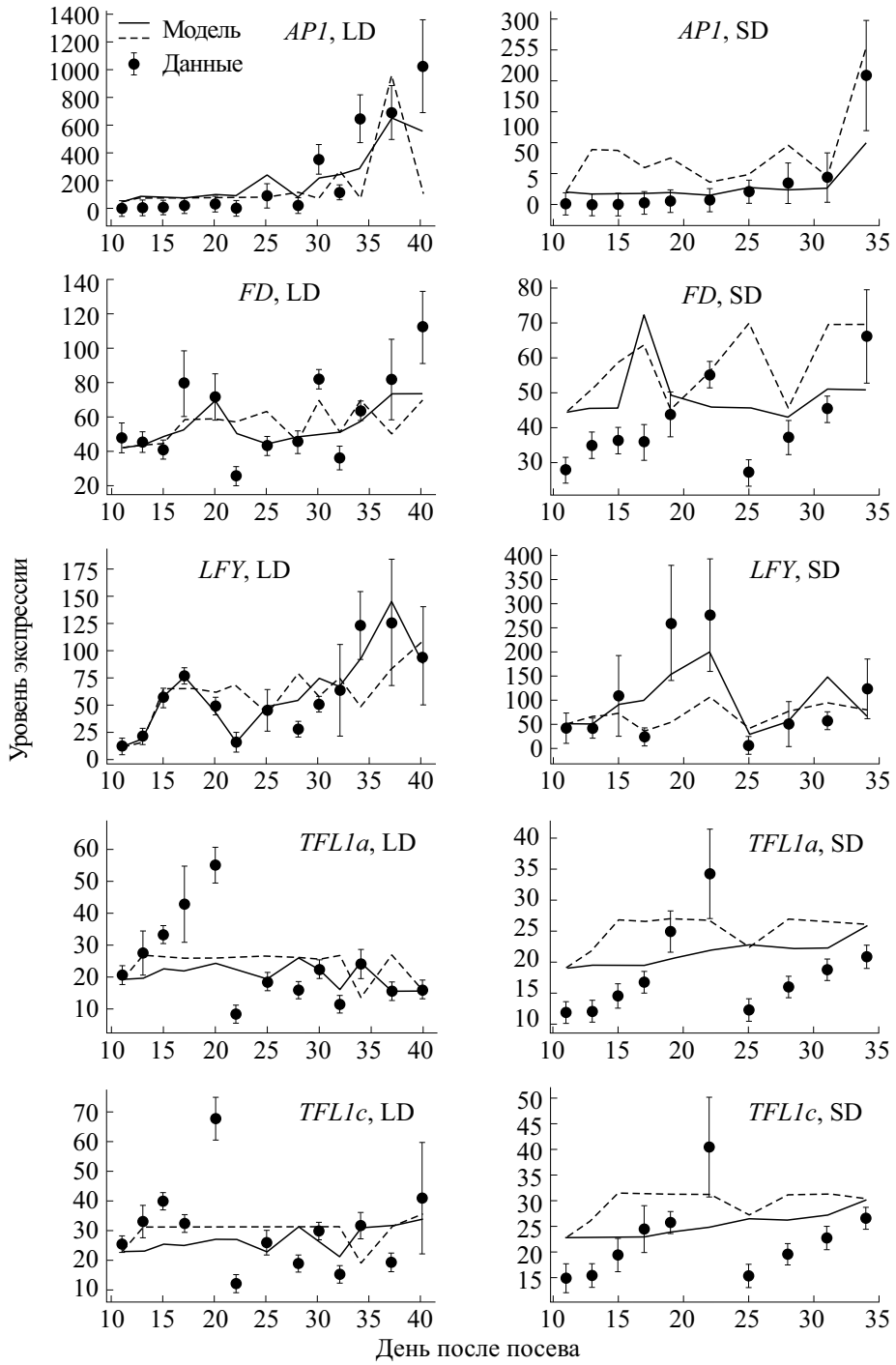

Рис. 4. Динамические предсказания усеченной модели, обученной на выборке LD+SD, в сравнении со средними уровнями экспрессии в данных. Все обозначения соответствуют рис. 2.

TFL1c) наиболее сильное влияние на *AP1* оказывает *TFL1c* (рис. 7в).

В соответствии с результатами, полученными ранее в рамках моделирования экспрессии генов цветения в сорте нута ICCV 96029 [19], представленные модели также предсказывают нарушение в сорте CDC Frontier регуляторного модуля, состоящего из взаимной активации между *AP1* и *LFY*. Например, в полной модели, обученной на LD+SD выборке, в условиях LD *AP1* слабо репрессирует *LFY*, а в условиях SD *LFY* является сильным репрессором *AP1* (рис. 6). В полной мо-

Таблица 6. Результаты обучения и тестирования усеченной модели на LD-выборке

	НСКО		Коэффициент корреляции	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
AP1	0.050	0.075	0.94	0.90
FD	0.108	0.117	0.59	0.60
LFY	0.077	0.113	0.88	0.80
TFL1a	0.125	0.141	0.41	0.39
TFL1c	0.139	0.163	0.29	0.26

дели, обученной на LD-выборке, *LFY* является активатором *AP1* (рис. 7б), однако *AP1* слабо репрессирует *LFY* (рис. 6).

ЗАКЛЮЧЕНИЕ

Полученные результаты демонстрируют применимость методов машинного обучения для анализа экспрессии генов цветения в нуте. Алгоритм случайного леса показывает хорошее качество обучения и тестирования на данных экспрессии в сорте нута CDC Frontier в условиях длинного дня. Построенные на этом обучении модели позволяют вычислять динамику средних уровней экспрессии, которая хорошо соотносится с наблюдаемой динамикой в данных [14] в случае, когда данные используются как входные значения для предсказаний уровня экспрессии на следующий день. Этот результат контрастирует с плохими результатами предсказания такой динамики для сорта CDC Frontier, полученными ранее с помощью модели на основе дифференциальных уравнений [19].

Из четырех построенных моделей только полная модель, обученная на LD-выборке, позволяет корректно воспроизводить динамику средней

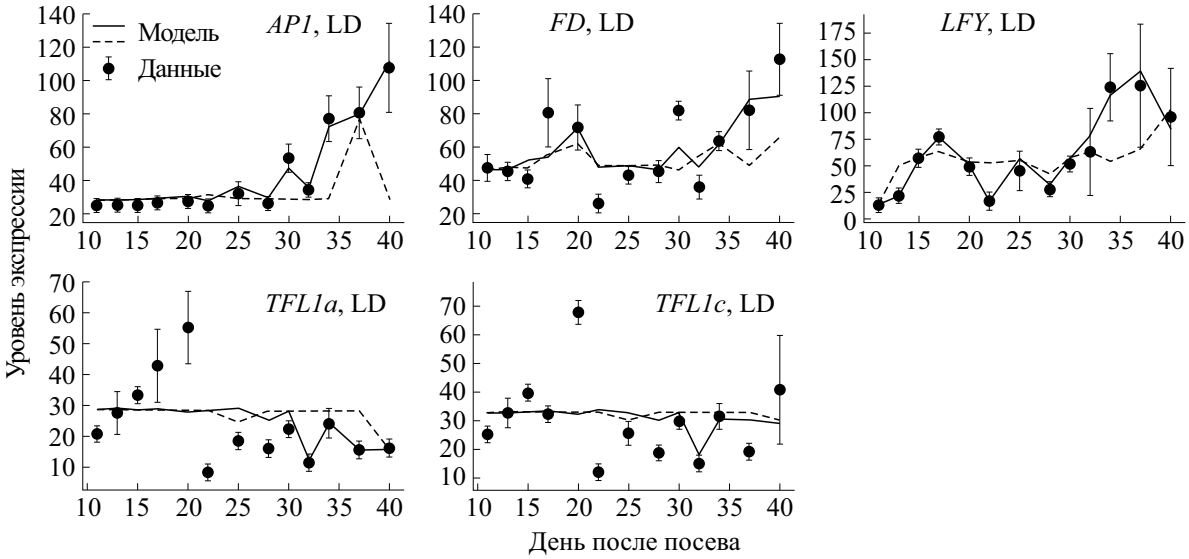

Рис. 5. Динамические предсказания усеченной модели, обученной на LD-выборке, в сравнении со средними уровнями экспрессии в данных. Все обозначения соответствуют рис. 2.

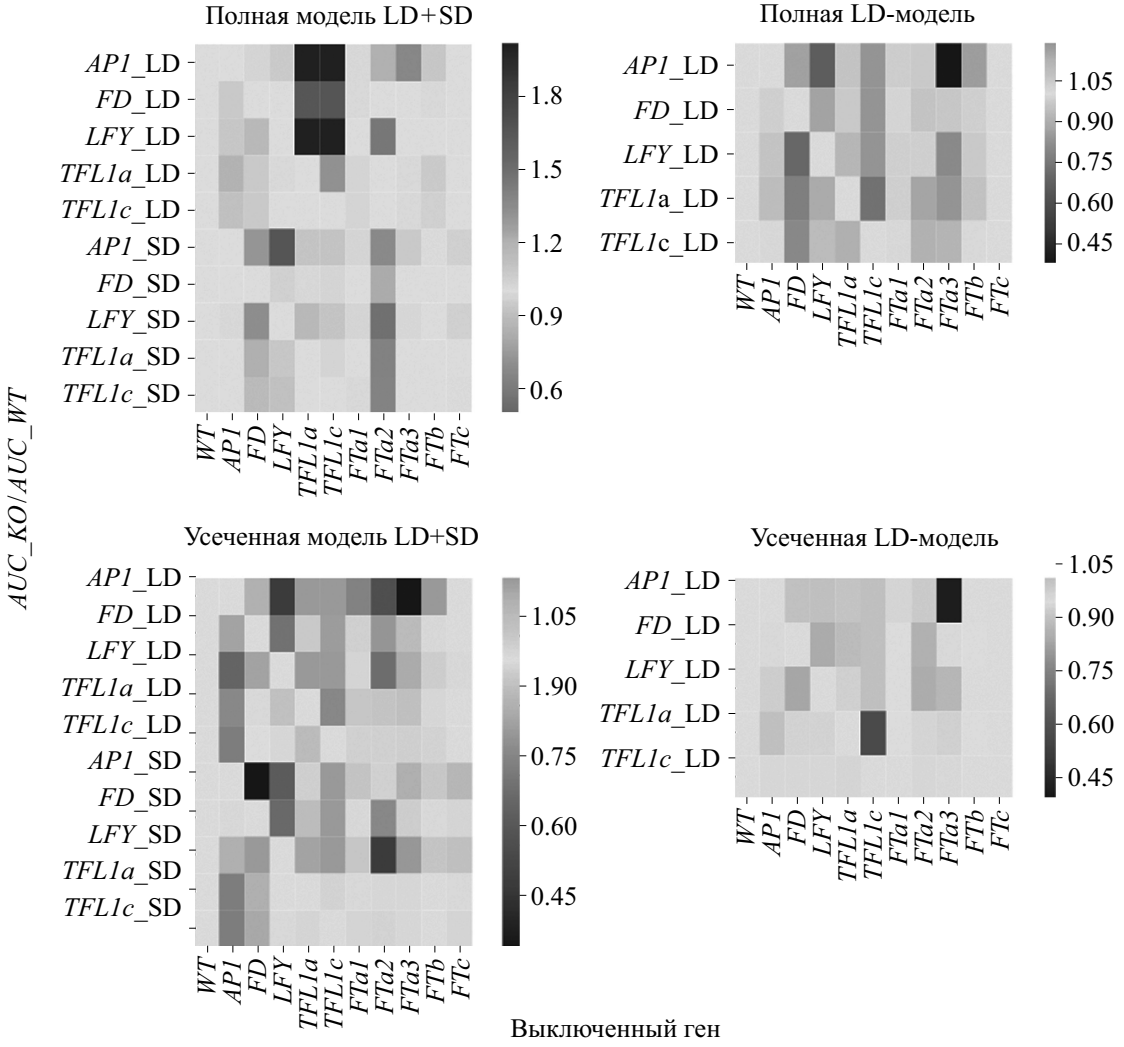

Рис. 6. Результаты моделирования нокаутов генов. Для каждой из четырех моделей показано отношение площади под кривой динамики среднего уровня экспрессии гена-мишени в модели при одном выключенном гене-регуляторе (AUC_KO) к такой же площади для условий дикого типа, т. е. когда все регуляторы присутствуют в модели (AUC_WT), отдельно для условий SD и LD. Динамику экспрессии в модели вычисляли вторым способом, описанным в тексте. Значение AUC_KO/AUC_WT > 1 при выключенном регуляторе соответствует репрессии гена-мишени этим регулятором, а значения AUC_KO/AUC_WT < 1 – активации.

экспрессии в случае, когда такая динамика вычисляется не на основе данных в предыдущий день, а используя выходы самой модели в предыдущий день (кроме данных экспрессии *FT*-генов, являющихся внешними факторами в модели). Этот результат свидетельствует, что использование SD-данных в обучении скорее ухудшает качество предсказаний модели. Также модели, обученные с учетом и без учета данных короткого дня, предсказывают разный характер взаимодействий между генами. Таким образом, моделирование показывает, что SD- и LD-данные содержат качественно разную информацию, что, возможно, связано с наличием разных регуляторных взаимодействий в условиях короткого и длинного дня.

Наличие нескольких *FT*-генов и нескольких *TFL1*-генов в нуте ставит вопрос о функциональной роли отдельных гомологов и о характере потенциальной активации и репрессии генов цветения отдельными *FT*- и *TFL1*-генами соответственно. Из представленных результатов моделирования можно сделать предположение о ведущей роли *FTa3* в активации и *TFL1c* в репрессии генов идентичности цветковых меристем в условиях длинного дня. Как и для сорта нута ICCV 96029 [19], в случае CDC Frontier моделирование также выделяет взаимную активацию между *AP1* и *LFY* в качестве одного из регуляторных модулей, нарушение которого отличает нут от *Arabidopsis*. Общая изменчивость характера взаимодействий, наблюдающаяся между моделями,

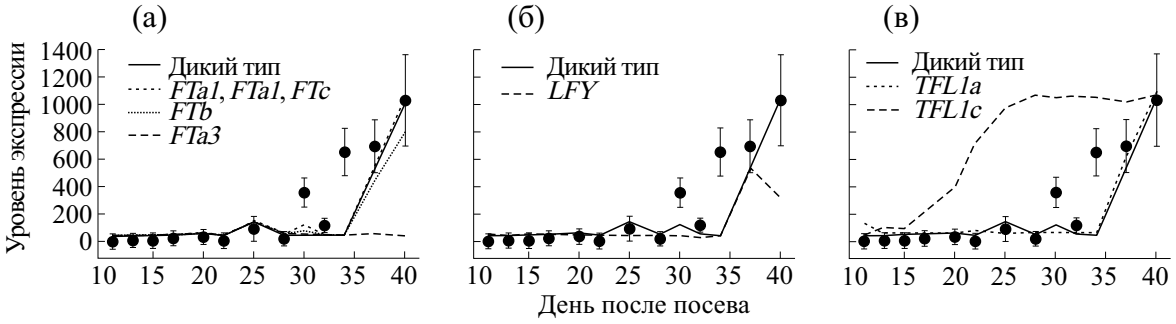

Рис. 7. Динамика экспрессии *AP1* в полной модели, обученной на LD-выборке, в условиях, когда один из следующих генов выключен: (а) – один из *FT*-генов, (б) – *LFY*-ген, (в) – один из *TFL1*-генов. Для сравнения на каждом графике также показана динамика экспрессии в диком типе (черная кривая). Название выключенных генов приведены на графиках рядом с цветом соответствующей кривой. Кривые, соответствующие нокауту каждого из генов *FTa1*, *FTa2* и *FTc*, показаны одной пунктирной линией в силу слабого различия между этими кривыми. Кривая дикого типа сливается с этими кривыми на графике (а).

построенными с учетом и без учета генной сети из *Arabidopsis*, вероятно, свидетельствует о недостаточности взаимодействий, лежащих в основе этой сети, для адекватного описания данных экспрессии генов цветения в CDC Frontier.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации, задание № 1.8697.2017/БЧ.

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания каких-либо исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. C. Jung and A. E. Müller, Trends Plant Sci. **14**, 563 (2009). DOI: 10.1016/j.tplants.2009.07.005
2. A. Kanth and M. Schmid, Cell Mol. Life Sci. **68**, 2013 (2011). DOI: 10.1007/s00018-011-0673-y
3. M. Khan, X Ai, and J. Zhang, Wiley Interdiscip. Rev. RNA **5**, 347 (2014). DOI: 10.1002/wma.l215
4. J. A. Banta, I. M. Ehrenreich, S. Gerard, et al., Ecol. Lett. **15**, 769 (2012). DOI: 10.1111/j.1461-0248.2012.01796.x
5. F. Andrés and G. Coupland, Nat. Rev. Genet. **13**, 627 (2012). DOI: 10.1038/nrg3291
6. V. Hecht, F. Foucher, C. Ferrándiz, et al., Plant Physiol. **137**, 1420 (2005). DOI: 10.104/pp.l04.057018
7. R. Benlloch, A. Berbel, L. Ali, et al., Front. Plant Sci. **6**, 543 (2015). DOI: 10.3389/fpls.2015.00543
8. J. L. Weller and R. Ortega, Front. Plant Sci. **6**, 207 (2015). DOI: 10.3389/fpls.2015.00207
9. C.-H. Jung, C. E. Wong, M. B. Singh, and P. L. Bhalla, PloS One **7**, e38250 (2012). DOI: 10.1371/journal.pone.0038250
10. H. D. Upadhyaya, D. Bajaj, S. Das, et al., Plant Mol. Biol. **89**, 403 (2015). DOI: 10.1007/sl1103-015-0377-z
11. J. L. Weller and R. C. Macknight, Methods Mol. Biol. **1822**, 261 (2018). DOI: 10.1007/978-1-4939-8633-0_17
12. R. Ortega, V. F. Hecht, J. S. Freeman, et al., Front. Plant Sci. **10**, 824 (2019). DOI: 10.3389/fpls.2019.00824
13. M. Blümel, N. Dally, and C. Jung, Curr. Opin. Biotech. **32**, 121 (2015). DOI: 10.1016/j.copbio.2014.11.023
14. S. Ridge, A. Deokar, R. Lee, et al., Plant Physiol. **175**, 802 (2017). DOI: 10.1104/pp.l7.00082
15. K. E. Jaeger, N. Pullen, S. Lamzin, et al., Plant Cell. **25**, 820 (2013). DOI: 10.1105/tpc. 113.109355
16. F. C. Sussmilch, A. Berbel, V. Hecht, et al., Plant Cell. **27**, 1046 (2015). DOI: 10.1105/tpc. 115.136150
17. F. Valentim, S. van Mourik, D. Posé, et al., PloS One **10** (2), e0116973 (2015). DOI: 10.1371/journal.pone.0116973
18. R. K. Varshney, C. Song, R. K. Saxena, et al., Nat. Biotechnol. **31**, 240 (2013). DOI: 10.1038/nbt.2491
19. V. V. Gursky, K. N. Kozlov, S. V. Nuzhdin, and M. G. Samsonova, Front. Gen. **9**, 547 (2018). DOI: 10.3389/fgene.2018.00547
20. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., J. Machine Learning Res. 12, 2825 (2011).

Machine-Learning Analysis of Flowering Gene Expression in CDC Frontier Chickpea Cultivar

B.S. Podolny*, V.V. Gursky, and M.G. Samsonova***

**Peter the Great St. Petersburg Polytechnic University, Polytekhnicheskaya ul. 29, St. Petersburg 195251 Russia*

***Ioffe Physical Technical Institute, Polytekhnicheskaya ul. 26, St. Petersburg, 194021 Russia*

We analyze gene expression dynamics in floral transition in the CDC Frontier chickpea cultivar. We provide a model, in several versions, to predict the expression dynamics of five flowering genes taking the expression of their regulators as the input. The models were trained using the random forest method on previously published expression data for ten flowering genes under the short- and long-day growing conditions. The resulting models correctly predict the dynamics of the average expression levels under long days. We show that the models for CDC Frontier mainly reproduce regulatory interactions between the key genes described for the model plant Arabidopsis thaliana. Based on the analysis, we hypothesize that the short-day data and the long-day data contain qualitatively different information, which can be related to different regulatory modules functioning in different conditions. For the regulators of the flower meristem identity genes AP1 and LFY, our models predict FTa3 as the main activator and TFL1c as the main repressor under long days.

Keywords: chickpea, CDC Frontier, flowering gene network, random forest algorithm