

Руководитель: к.б.н., Левицкий Виктор Георгиевич, levitsky@bionet.nsc.ru
([Researchgate](#), [Scopus](#))

Темы дипломных работ:

1. Анализ типов прямых взаимодействий транскрипционных факторов с геномной ДНК с помощью данных полногеномного картирования их сайтов связывания
2. Анализ организации промоторов генов домашнего хозяйства и генов развития плодовой мушки *D. melanogaster* на основе поиска сайтов связывания транскрипционных факторов
3. Анализ специфичности экспрессии генов по тканям и стадиям развития модельного растения *A. thaliana* на основе предсказания сайтов связывания транскрипционных факторов разных структурных типов в промоторах генов

Выполнение всех тем дипломных работ будет проходить с поддержкой вычислительных ресурсов Центра коллективного пользования «Биоинформатика» ИЦиГ СО РАН.

В зависимости от выбранной темы знания и навыка, которые получит студент в ходе выполнения работы:

1. Освоение базовых форматов данных в биоинформатике, способов их автоматического преобразования;
2. Работа с интернет-доступными базами полногеномных данных;
3. Освоение построения стандартной и альтернативных моделей распознавания сайтов связывания транскрипционных факторов *de novo* и оценка точности этих моделей на основе полногеномных данных картирования;
4. Методы первичной и вторичной обработки данных полногеномного картирования сайтов связывания транскрипционных факторов;
5. Методы извлечения и анализа полногеномных разметок генов, структурно-функциональных районов генома;
6. Методы анализа регуляции транскрипции генов с помощью данных полного транскриптома;
7. Методы анализа полногеномных данных с помощью подхода терминов геномной онтологии;
8. Анализ полногеномных данных массового секвенирования сайтов связывания транскрипционных факторов для модельных организмов из разных таксонов (растения, млекопитающие, насекомые, дрожжи и т.д.);
9. Применение методов математической статистики в биоинформатике.

Требования и пожелания к соискателям:

1. Развитие способностей к разработке и применению новых алгоритмов анализа данных в биоинформатике;
2. Средний балл не ниже 4;
3. Желание и способность эффективно и творчески работать, осваивать современные методы полногеномного анализа;
4. Приветствуется опыт программирования (C++, Python, R), либо желание его быстрого приобретения;
5. Знание английского языка (уровень, достаточный для свободного чтения научной литературы), как и способность к структурному и лаконичному

представлению результатов работы на английском и русском языках могут быть существенно улучшены в ходе выполнения дипломного проекта при должной целеустремленности.

План действий:

1. Прислать своё резюме (CV) на levitsky@bionet.nsc.ru
2. Очное или дистанционное собеседование;
3. Решение тестовой задачи (1-2 недели);
4. Обсуждение результатов выполнения тестовой задачи, принятие решения и начало работы над дипломным проектом.

Краткое описание направлений исследования Тема 1:

Транскрипционные факторы (ТФ) это белки, контролирующие процесс транскрипции генов путём связывания со специфичными участками ДНК, сайтами связывания ТФ (ССТФ). Как правило, ТФ выполняют свою функцию в комплексе с другими белками, включая другие ТФ. Для изучения функций ТФ широко используется метод иммунопреципитации хроматина с последующим массовым секвенированием (ChIP-seq), который позволяет находить локусы ДНК в составе хроматина, где находится ТФ. Поиск ССТФ в данных ChIP-seq производят с помощью de novo методики, которая позволяет выявлять обогащенные мотивы в ДНК, предположительно соответствующие ССТФ целевого белка ТФ ChIP-seq эксперимента. Стандартная модель описания мотивов ССТФ, позиционная весовая матрица (ПВМ), обычно описывает не менее половины всех выявленных локусов ChIP-seq. Отчасти, оставшиеся локусы можно объяснить наличием мотивов альтернативной структуры. Эта возможность обусловлена недостатком модели ПВМ, предполагающей аддитивность вкладов отдельных позиций в ССТФ. К настоящему времени показано, что альтернативные модели, которые учитывают зависимости разных позиций в пределах ССТФ, могут существенно дополнять и уточнять предсказания модели ПВМ. Согласно классификации структуры ТФ, выделяют несколько суперклассов ТФ, которые в свою очередь делятся на классы и семейства родственных ТФ. В рамках дипломного проекта, на основе данных полногеномного картирования ССТФ ChIP-seq, будут изучены вклады стандартной и альтернативных моделей распознавания ССТФ в общие прямые взаимодействия ТФ с геномной ДНК. Будет произведен сравнительный анализ ТФ различных семейств ТФ, а также произведен сравнительный анализ вкладов стандартной и альтернативных моделей в наблюдаемое разнообразие природных ССТФ для таких удалённых таксонов как млекопитающие, растения, насекомые и дрожжи.

Краткое описание направлений исследования Тема 2:

Гены развития у плодовой мушки *D. melanogaster* специфически активны на определённых стадиях и в отдельных тканях. Эти гены расположены в геноме в составе закрытого (плотно упакованного) хроматина. Гены домашнего хозяйства - это гены, необходимые для поддержания общих функций организма. Такие гены экспрессируются в большинстве тканей и стадий развития организма. Промоторы генов домашнего хозяйства чаще располагаются в областях более открытого (менее упакованного) хроматина. По результатам полногеномного картирования белков открытого хроматина для генома дрозофилы ранее была разработана модель его фрагментации на несколько типов хроматина с присутствием сайтов определенного набора белков. Среди этих типов был выделен как тип открытого хроматина, так и

другие типы, соответствующие в том числе и наиболее плотно упакованному хроматину. В рамках дипломного проекта будет изучена специфичность распределения сайтов связывания транскрипционных факторов (ССТФ) определённой структуры с учётом классификации генов на гены домашнего хозяйства и гены развития, а также более детальной их классификации по уровням экспрессии в тканях и на стадиях развития (на основе данных полного транскриптома RNA-seq). ССТФ будут предсказываться на основе экспериментальных данных их полногеномного массового картирования с помощью технологии ChIP-seq. Для описания структуры ССТФ будет применяться как стандартная модель описания мотивов ССТФ, позиционная весовая матрица (ПВМ), так и альтернативные модели распознавания ССТФ, учитывающие зависимости разных их позиций. С помощью этого подхода будет производится классификация ССТФ с учётом типа хроматина и паттерна экспрессии гена для разных тканей и стадий развития организма.

Краткое описание направлений исследования Тема 3:

Транскрипционные факторы (ТФ), являясь главными регуляторами транскрипции генов, определяют ткане- и стадие-специфичную экспрессию генов. ТФ взаимодействуют с геномной ДНК в составе многокомпонентных комплексов с другими ТФ. Успехи массового секвенирования сайтов связывания ТФ (ССТФ) последних 5-10 лет, позволили начать изучение молекулярных механизмов взаимодействий ТФ с геномной ДНК и друг другом *in vivo* для различных условий тканей и стадий развития. Как правило, наиболее консервативный участок ССТФ представляется мотивом, длина которого составляет обычно 8-15 нуклеотидов. Технология полногеномного картирования ССТФ ChIP-seq даёт тысячи геномных локусов (пиков) в которых предположительно располагаются функциональные ССТФ. Последующая дифференциация различных структурных типов ССТФ в регуляторных районах генов, активных в различных тканях и на разных стадиях развития, может позволить выявить разнообразие мотивов ССТФ и существенно продвинуться в исследовании механизмов регуляции экспрессии генов. Для решения задачи поиска мотивов ССТФ применяются разные модели, самой распространённой из них является позиционная весовая матрица (ПВМ). ПВМ строится по выравниванию известных ССТФ на основе частот встреч нуклеотидов в его позициях. Несмотря на то, что модель ПВМ в целом удовлетворительно объясняет большинство известных ССТФ, данная модель не учитывает зависимости между разными позициями в сайтах, так как, общая аффинность сайта в модели оценивается как сумма вкладов всех позиций сайта. В ходе выполнения дипломного проекта будет проведён анализ структурной гетерогенности ССТФ по нескольким ТФ с учетом разных условий тканей и стадий развития модельного растения *A. thaliana*. Структурная гетерогенность ССТФ будет оцениваться с помощью применения стандартной (ПВМ) модели, а также ряда альтернативных моделей мотивов ССТФ.