Основнные результаты лаборатории за 2011 год

1. **Модель транскрипции в пластидах и митохондриях**.

Транскрипция – основной процесс в жизнедеятельности клетки. Нами впервые выдвинута количественная концепция конкуренции РНК-полимераз в этом процессе. Впервые предложена математическая модель этого процесса, включая регуляцию транскрипции на основе конкуренции РНК-полимераз. Модель реализована эффективными алгоритмами, распараллелена и компьютерно реализована для многопроцессорных вычислительных устройств. Результаты вычислений, проведенных на кластере MVS100K МСЦ с использованием до 2048 процессоров, позволили:

1.1. получить объяснение сложного изменения уровней транскрипции генов в изолированных хлоропластах ячменя на основе моделирования конкуренции РНК-полимераз при различных температурах;

1.2. получить количественное изменение транскрипции генов при различных мутациях на основе моделирования транскрипции кольцевых митохондриальных ДНК у человека, крысы и лягушки;

1.3. предложить и обосновать гипотезу о природе митохондриальных генетических болезней: изменение уровня фенилаланиновой тРНК, что можно использовать при выборе компенсирующих лекарственных препаратов.

Поясним пункт 1.3. Транзиция A→G в середине сайта связывания белка-терминатора mTERF вызывает тяжелые генетические заболевания человека, известные как МЕЛАС-болезни. В опытах влияние мутаций, разрушающих этот сайт, на интенсивность транскрипции митоходриальных генов у человека таково: для рРНК изменение незначительное, как и среднее отклонение от нормы для генов, кодирующих белки, для tRNA-Leu отклонение не более 20% , для tRNA-Lys отклонение не более 40-50% . В тоже время медицинские наблюдения указывают, что экспрессия генов существенно нарушается и возникают тяжелые заболевания. Почему при незначительном изменении транскрипции всех генов возникает такой результат? Модель показала, что ответ состоит в существенном снижении связи mTERF с сайтом и соответственно в изменении уровня транскрипции фенилаланиновой тРНК. А именно, моделирование показывает, что в результате мутации в 1.5 раза снижается связь mTERF с сайтом и одновременно в 1.4 раза – связь РНК-полимераз с промотором HSP1 в связи с ролью mTERF как белка-активатора (кооперативное связывание). При этом уровень транскрипции фенилаланиновой тРНК уменьшается в 1.4 раза, тогда как уровни транскрипции всех без исключения белковых

генов и большинства остальных тРНК остаются практически на уровне нормы (изменения не превышают 10%). Что касается рРНК, модель предсказывает небольшое уменьшение уровня транскрипции, приблизительно на 16%. Таким образом, наиболее существенное изменение – снижение уровня tRNA-Phe, что никак не вытекало из наблюдений.

2. Модель согласования деревьев и ее алгоритм кубической сложности.

2.1. Одна из самых старых задач в филогеномике состоит в построении дерева *S*, согласующего данный набор деревьев G_i; листьям дерева G_i приписаны родственные последовательности, образующие *i*-е семейство. Результат является безусловно пионерским: в литературе отсутствовали всякие указания на такой подход, алгоритм и даже на саму возможность полиномиальной сложности корректного алгоритма в явно экспоненциальной ситуации этой задачи.

В типичных случаях *S* является деревом видов или деревом регуляторной системы вместе с регулируемым геном или отдельно от него. Лист дерева видов отождествляется с именем приписанного ему вида, лист дерева генов обозначается парой ген и вид, из которого получен этот ген; допускаются несколько генов из одного вида. Традиционный подход к решению этой задачи таков: дерево *S* строится так, чтобы суммарная близость к *S* всех данных деревьев G_i была максимальной. Однако фундаментальная проблема состояла в том, что математически доказана NP-трудность этой задачи, т.е. любой алгоритм ее решения обязан иметь экспоненциальное время работы. Многочисленные эвристические алгоритмы ее решения работают более или менее быстро, но заведомо не находят глобальный максимум суммарной близости. Нами предложена новая постановка задачи поиска супердерева, которая допускает вычислительно простой детерминированный алгоритм решения и вместе с тем адекватна биологическим задачам. А именно, ищется супердерево *S*, большинство клад которого представлено среди клад исходных деревьев генов G_i. Нами математически доказаны корректность алгоритма и время его работы, которое для худших данных имеет порядок $n^3 \cdot m^3$, где n – число деревьев генов, а m – число видов в них. Соответствующая программа построения дерева *S* свободно доступна на сайте http://lab6.iitp.ru/en/super3gl/.

2.2. Другая столь же старая задача филогеномики состояла в согласовании одного дерева генов $G=G_i$ с деревом видов *S*. Известны многочисленные эвристические алгоритмы ее решения. Нами впервые предложен математически корректный алгоритм кубической сложности, который решает эту задачу. Нами предложена концепция временных слоев в

дереве S с тем, чтобы горизонтальные переносы разрешались только внутри одного слоя. Этого можно достичь, разбивая ребра в S новыми вершинами (получается новое дерево видов S_0), тогда слой состоит из ребер, одинаково удаленных от корня; в частности, S_0 может совпадать с S. Наш алгоритм согласует деревья G и S_0 (формально: определяет вложение G в S_0) за время $O(|G|\cdot|S_0|)$, что дает $O(|S|^3)$. Здесь $|\cdot|$ означает число вершин в соответствующем дереве.

3. **Совместная эволюция промоторов и РНК-полимераз пластид**.

До недавнего времени РНК-полимеразы и промоторы в пластидах подробно исследованы у небольшого числа модельных организмов: практически только у резушки Таля, табака и кукурузы. Биоинформатический анализ ядерных геномов, в которых кодируются РНК-полимеразы фагового типа и σ-субъединицы РНК-полимераз бактериального типа, стал возможен после секвенирования достаточно большого количества локусов хромосомной ДНК и кДНК. После такого секвенирования и на основе оригинального алгоритма нами впервые изучены РНК-полимеразы и промоторы в пластидах видов, которые охватывают все разнообразие растений и водорослей.

Показана корреляция между эволюцией промотора и эволюцией σ-субъединицы РНК-полимеразы бактериального типа. Исследовано возникновение паралогов σ-субъединиц РНК-полимераз бактериального типа. На этой основе построен эволюционный сценарий для σ-субъединиц РНК-полимераз. Полученные результаты говорят, в частности, о неэффективности в борьбе с Piroplasmida антибиотиков, ингибирующих РНК-полимеразу бактериального типа. Напротив, такие антибиотики могут применяться против *Plasmodium* spp, *T. gondii* и *Neospora caninum*.

3.1. Проведена кластеризация белков, кодируемых в пластидах багрянок и видов с родственными им пластидами (родофитная ветвь). Это позволяет эффективно проводить поиск белков, кодируемых в пластидах, по их филогенетическому профилю.

3.2. Проведен поиск РНК-полимеразы фагового типа в ядерных геномах споровиков, водорослей и наземных растений. Показано хорошее согласие между деревьями видов и белков.

3.3. Проведен поиск σ-субъединиц РНК-полимераз бактериального типа в ядерных геномах споровиков, водорослей и наземных растений. Много σ-субъединиц найдено у багрянок, диатомовых, золотистой *Aureococcus anophagefferens* и бурой *Ectocarpus*

siliculosus водорослей. Напротив, в нуклеоморфах криптофитовых водорослей (*Hemiselmis andersenii* и *Guillardia theta*), у золотисто-бурой водоросли *Heterosigma akashiwo*, у кокцидий и *Plasmodium* spp найдено лишь по одной σ-субъединице. Две σ-субъединицы из *Cyanophora paradoxa* значительно различаются между собой, особенно на N-конце, хотя имеют несколько консервативных участков; первая близка к одной из таковых у диатомовых водорослей, а вторая существенно отличается от σ-субъединиц всех водорослей. У *Bigelowiella natans* и у большинства зелёных водорослей группы Chlorophyta найдено лишь по одной σ-субъединице; исключениями являются *Micromonas pusilla* и *Chlorella variabilis*, имеющие по две σ-субъединицы, которые в каждом из этих видов близки друг другу. Это позволяет предположить их независимое и недавнее возникновение в результате дупликаций. Выполнен поиск минорных σ-субъединиц в геномах наземных растений, включая мхи. За основу поиска принята классификация σ-субъединиц у арабидопсиса. Выполнен поиск ортологов Sig1, который подтвердил, что эта субъединица имеет наиболее широкое распространение у растений. Показана корреляция между присутствием субъединицы Sig5 и транскрипционного фактора, необходимого для светозависимой регуляции соответствующего промотора.
3.4. Исследованы промоторы в пластидах диатомовых водорослей. У диатомовых водорослей и видов с близкими пластомами перед некоторыми генами, включая *psaA*, промоторы имеют два нуклеотида G в составе -35 бокса, что противоречит как бактериальному, так и другому характерному для этих видов консенсусу. Сравнительно низкое число предсказанных промоторов у диатомовых, золотистых и бурых водорослей можно объяснить разнообразием σ-субъединиц и соответствующих промоторов, которые не были учтены. Напротив, у *Heterosigma akashiwo* найдена только одна σ-субъединица и для нее предсказано наибольшее число промоторов (103), больше чем для багрянок. Вероятно, отсутствие у *H. akashiwo* минорных σ-субъединиц привело к стабилизации всех промоторов вблизи одного консенсуса.

4. **Поиск и эволюция регуляторных систем у бактерий и пластид**. **Гиббсовский подход для реконструкции структур РНК**. В отличие от эволюции генов и белков, эволюция регуляторных систем является новой областью исследований. Здесь на основе предложенных нами оригинальных алгоритмов получен ряд пионерских результатов.
4.1. Совместно с лаб. М.С. Гельфанда реконструирована эволюционная история сайтов транскрипционных факторов, в частности, семейств NrdR, MntR, LacI, FNR, Irr, Fur и

Rrf2. В большинстве семейств изменения сайтов связывания ограничивались несколькими ребрами эволюционного дерева. Показано: изменение консенсусных нуклеотидов проходит через промежуточную стадию, в которой соответствующая позиция сайта не является консервативной.

4.2. Найден консервативный мотив перед генами *proA* и *proB* у многих γ-протеобактерий. Предположено, что соответствующим транскрипционным фактором является белок из семейства TetR, ортологичный белку NP_249058 из *P. aeruginosa* PAO1, что затем нашло подтверждение в проведенной нами экспериментальной работе. Получены сценарии совместной эволюции генов, белковых факторов регуляции и сайтов связывания.

4.3. Показана регуляция лейцина на уровне транскрипции с помощью различных вариантов аттенюаторных структур. Для этого проведен широкомасштабный поиск потенциальных аттенюаторных регуляторных структур у бактерий, который опирается на две оригинальные компьютерные программы – модель процесса аттенюаторной регуляции и множественное выравнивание по данному филогенетическому дереву. Найдены различные типы аттенюаторной регуляции у бактерий из таксонов α-, β-, γ-, δ-proteobacteria, Actinobacteria, Bacteroidetes/Chlorobi, Firmicutes и Thermotogae и Cloroflexi перед генами *hisG, hisZ, hisS*, *pheA*, *pheST, trpEG*, *trpA*, *trpB*, *trpE*, *trpS*, *thrA, thrS*, *leuA*, *leuS*, *ilvB*, *ilvI, ilvA, ilvC, ilvD, ilvG*. Показано, что для регуляции весьма существенны РНК-триплексы (при формировании эффективных антитерминатора и терминатора) и псевдоузлы (в процессе терминации). Показано, что аттенюаторная регуляция гена *lysQ* у *Lactobacillus lactis* зависит от концентрации гистидил-тРНК. Среди новых типов аттенюаторной регуляции особенно выделяются регуляции на основе найденных нами специфических для актинобактерий и протеобактерий псевдоузлов (LEU и LEU1). Изучена эволюция аттенюаторной регуляции.

4.4. Впервые на основе гиббсовского подхода предложена модель эволюцию нуклеотидных последовательностей с учетом вторичной структуры в них; она позволяет реконструировать структуры РНК у предков видов. Подход основан на оптимизации функционала гиббсовского типа, в котором помимо стандартной эволюционной динамики первичной структуры заложено требование консервативности вторичной структуры.

4.5. Найдены предковые регуляторные сигналы, вовлекающие вторичную структуру РНК, для основных типов аттенюаторной регуляции у бактерий: классической аттенюа-

торной (биосинтез треонина и лейцина у гамма-протеобактерий), Т-боксовой (у актинобактерий), RFN-регуляции (у эубактерий), LEU-регуляции (у актинобактерий и протеобактерий).

4.6. Предсказаны регуляторные структуры и консервативные сайты перед пластомными генами, имеющими интроны. Перед инициирующими кодонами генов *clpP*, *petB*, *psaA* и *rpl16* предсказаны консервативные структуры РНК. Подробно изучена структура РНК 5'-лидерной области ряда генов, включая *rpl16*. В частности, для этого гена предсказана регулируемая задержка инициации трансляции посредством перекрытия сайта связывания рибосомы вторичной структурой РНК. Также изучена вторичная структура 3'-трейлерных областей генов, имеющих интроны, в которых также предсказаны шпильки РНК.

4.7. Исследованы вставки прямых повторов в пластомах семенных растений. Предположено, что в ходе эволюции прямые повторы часто возникают одномоментно, возможно, в результате репликативных ошибок, состоящих в удвоении участка ДНК. Наиболее обычны повторы одного и того же нуклеотида, однако локальный максимум частоты приходится на длину 5 п.н повторяемого слова.

5. **Регуляция на уровне трансляции и процессинга в пластидах**. Для ряда генов здесь получены существенно новые результаты.

5.1. Предсказана регуляция инициации транскрипции посредством связывания транскрипционного фактора с сайтом ДНК вблизи промотора перед опероном *chlLN*. Сайт перекрывает промотор и является тандемным повтором с консенсусом GATCTAT-11п.н.- GATCTAT. Сайт найден у единственной водоросли *Chara vulgaris*, мохообразного *Anthoceros formosae*, плауна *Huperzia lucidula*, папоротника *Adiantum capillus-veneris* и голосеменных растений *Cycas taitungensis*, *Keteleeria davidiana*, *Picea abies*, *Pinus spp*. У этих видов перед другими генами сайт не определился. Искаженный сайт с промотором низкого качества найден перед геном *chlN* у некоторых видов Coniferopsida и перед псевдогеном *chlL* у *Gnetum gnemon*. Сайты другого типа являются инвертированными повторами с консенсусом GATCTAT-11п.н.-ATAGATC и перекрывают промоторы у голосеменных *Chamaecyparis lawsoniana*, *Ch. obtusa*, *Cunninghamia lanceolata*. Инвертированный повтор у *Cunninghamia* не имеет аналогов у других исследованных видов.

5.2. У трёх водорослей из таксономической группы Chlorophyta обнаружены длинные вставки, не нарушающие рамку считывания *chlL*. Такая вставка есть у *Ch. reinhardtii*,

Volvox carteri (принадлежащих Chlamydomonadales) и у *Pycnococcus provasolii* (принадлежащих Pseudoscourfieldiales), но отсутствует у других исследованных видов. В частности, вставка отсутствует у близких видов *Ch. moewusii* и *Dunaliella salina* из Chlamydomonadales и *Nephroselmis olivacea* из Pseudoscourfieldiales. Поскольку аминокислотная последовательность белка ChlL черезвычайно консервативна за исключением указанной вставки вблизи С-конца у перечисленных видов, можно думать, что эта вставка удаляется в ходе процессинга.

5.3. Предсказан консервативный сайт перед геном с неизвестной функцией в пластидах споровиков Piroplasmida. Ген расположен между генами *rpl14* и *rps8*, кодирующими рибосомные белки. У *Babesia bovis* и *Babesia bigemina* сайты располагаются внутри кодирующей области гена *rpl14*. Однако в окрестности сайта у белка произошла вставка (вида TSYSIDDRNRFKD у *B. bovis*), отсутствующая у ортологичных белков L14. У *Theileria parva* сайт не перекрыт кодирующими областями.