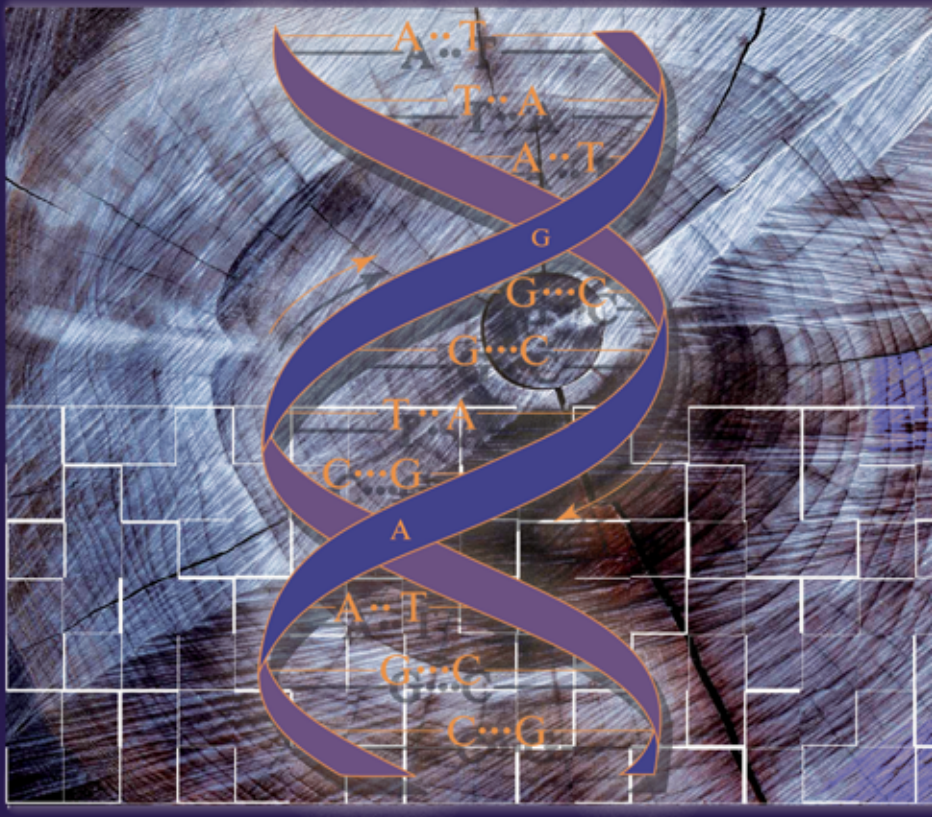
Н.Н. Козлов

МАТЕМАТИЧЕСКИЙ АНАЛИЗ ГЕНЕТИЧЕСКОГО КОДА



ИЗДАТЕЛЬСТВО БИНОМ

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Н.Н. Козлов

МАТЕМАТИЧЕСКИЙ АНАЛИЗ ГЕНЕТИЧЕСКОГО КОДА

Москва
БИНОМ. Лаборатория знаний
2010

УДК 575+573+519.8
ББК 28.04+22.18
К59

Козлов Н. Н.

К59 Математический анализ генетического кода / Н. Н. Козлов. — М. : БИНОМ. Лаборатория знаний, 2010. — 215 с. : ил., [8] с. цв. вкл. — (Математическое моделирование).

ISBN 978-5-9963-0119-5

В монографии на основе изучения генов установлены новые свойства генетического кода и вычислены важнейшие его интегральные характеристики; выделены две группы таких характеристик. Установлена взаимосвязь полученных характеристик в этих группах. Проанализирован известный к настоящему времени набор генов, в том числе человеческого генома; получен ряд неизвестных ранее эффектов.

Для научных работников, преподавателей и студентов, специализирующихся в области математического моделирования в науках о живом.

УДК 575+573+519.8
ББК 28.04+22.18

Первый тираж издания осуществлен при финансовой поддержке Российского фонда фундаментальных исследований по проекту № 09-01-07047

Научное издание

Серия: «Математическое моделирование»

Козлов Николай Николаевич

МАТЕМАТИЧЕСКИЙ АНАЛИЗ ГЕНЕТИЧЕСКОГО КОДА

Ведущий редактор *М. С. Стригунова*
Художественный редактор *Н. А. Новак*
Технический редактор *Е. В. Денюкова*
Корректор *Н. Н. Ектова*
Оригинал-макет подготовлен *М. Ю. Копаницкой* в пакете LaTeX 2_ε

Подписано в печать 18.05.10. Формат $60 \times 90/16$.
Усл. печ. л. 14. Тираж 600 экз. Заказ

Издательство «БИНОМ. Лаборатория знаний»
125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272, e-mail: binom@Lbz.ru, http://www.Lbz.ru

ISBN 978-5-9963-0119-5

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Монография Н. Н. Козлова «Математический анализ генетического кода» посвящена оригинальным исследованиям в области биоматематики. Круг научных интересов автора чрезвычайно широк. Ранее мы совместно выполнили ряд работ по анализу космических траекторий и эволюции структур, начиная с изучения движения ИС Луны и галактик до математического моделирования процесса структуризации вторичных структур РНК. Последняя тема относится к биоматематике, автор приступил к работе над ней после того, как я предложил ему обратиться к задачам молекулярной биологии.

Задача, о которой идет речь в монографии, была впервые поставлена автором позднее и активно мною поддерживалась. На основе 15 работ автора в ДАН, которые были опубликованы вплоть до 2008 г., была написана данная монография. Используя большой опыт исследования сложных природных дискретных систем различных типов, автор нашел свой оригинальный подход к решению поставленных задач. Исследование начиналось с анализа перекрывающихся генов, которые представлялись одним из типов сложных взаимосвязанных систем. Однако исследование показало, что такие гены являются хорошей площадкой для анализа свойств генетического кода. Было доказано (теорема для генетического кода), что структура генетического кода содержит феноменальные возможности для построения генетических перекрытий различных типов. Полученные результаты привели к постановке задачи о взаимосвязи генетических перекрытий и вариабельности кода, а также к исследованию неперекрывающихся генов. В конечном итоге возник новый подход в изучении больших геномов, в том числе генома человека. Была изучена также установленная автором математическая аналогия между генетическими перекрытиями и стеблями вторичной структуры матричных РНК. Интегральные характеристики генетического кода, введенные автором, позволяют с новых позиций изучать структуру кода. Открытие новых, неизвестных ранее свойств генетического кода с неожиданной стороны осветили проблему происхождения кода и его эволюции. Сказанное свидетельствует о глубине проникновения автора в суть

рассматриваемых фундаментальных проблем. При этом полученные оригинальные результаты не имеют аналогов в зарубежных исследовательских работах.

В связи со сказанным выше считаю исключительным важной публикацию данной монографии.

Академик Т. М. Энеев,
9 сентября 2009 г.

ПРЕДИСЛОВИЕ АВТОРА

В каждой естественной науке заключено
столько истины, сколько в ней математики.

И. Кант

Автора могут упрекнуть в том, что, как математик, он не вправе указывать биологам, что же таится в структуре генетического кода. Однако, оглядываясь назад, на всю историю открытия такого феномена, как генетический код, можно сказать, что именно неспециалисты внесли решающий вклад в постановку задачи о существовании кода, его структуре и свойствах. У истоков проблемы стоял Г. Мендель (ученик Доплера!), который в 1866 г. установил, что передача наследственных признаков потомству определяется независимыми факторами, которые позже получили название генов. Практически через 80 лет физик Э. Шредингер, а позднее астрофизик Г. Гамов и физик Ф. Крик внесли решающий вклад в постановку задачи и прояснение принципиальных аспектов, связанных с проблемой генетического кода (см. работы [1–5]). Было выяснено, что ДНК является носителем генетической информации, в 1953 г. определена пространственная структура ДНК, доказано существование первичной структуры у белка (см. п. 4 из монографии [6], а также работы [7, 8]), после чего был экспериментально установлен генетический код. Практически в ходе проведения всех названных исследований возникала новая наука — молекулярная биология. После завершения гигантского международного проекта по геному человека (1990–2003 гг.) в ней наступила постгеномная эпоха. Как известно, стоимость этого проекта оценивалась в 3 млрд долларов, а его завершение было приурочено к 50-летию классической работы [7]. Впервые объединенная коллекция статей по человеческому геному была представлена в журнале «Nature» в 2006 г.; она составлена по публикациям 2001 г. для отдельных хромосом, с включением последующих комментариев [9].

К настоящему времени расшифровано относительно небольшое число других больших геномов. Ситуация существенным образом изменится в ближайшие годы в связи с внедрением новейших методов расшифровки. В конце 2006 г. был объявлен конкурс на премию в 10 млн долларов, которая будет вручена создателям быстрого и дешевого метода расшифровки (см. статью [10]). Среди условий конкурса — возможность

расшифровать 100 любых человеческих геномов за 10 дней. Создание такого метода приведет к резкому возрастанию числа геномов больших размеров, которые будут расшифрованы уже в ближайшие годы. Принципиально важно, что создание такого метода откроет возможность расшифровать геномы, которые по оценкам генетиков на порядок и более превосходят человеческий. Математический анализ огромных объемов подобной информации приобретает особую актуальность. При этом наиболее значимыми становятся исследования, связанные с поиском новых постановок, которые ранее не обсуждались. Именно такие постановки позволят по-новому подойти к изучению громадной по объему и невероятно таинственной информации, которой мы уже обладаем. Результаты одного из таких исследований представлены в данной монографии.

Итак, к настоящему времени молекулярная биология уже стала производителем гигантских по объему экспериментальных данных, осмысление которых невозможно без математических методов и алгоритмов. Многолетний опыт математического моделирования с применением ЭВМ от легендарной «Стрелы» вплоть до самых современных супер-ЭВМ позволяет сформулировать важный вывод. Принципиально новые результаты при работе с гигантской генетической информацией могут быть получены только при использовании новых подходов, которые в своей основе учитывают сущность генетической информации, ее отличие от сугубо физической или химической информации. Именно такой подход оказался с успехом примененным к решению задач, о которых идет речь в данной монографии. Кратко опишем его.

В основе исследования лежат экспериментальные данные по необычным способам записи генетической информации, так называемым перекрывающимся генам, когда один и тот же участок ДНК кодирует два белка. К началу данного цикла исследований уже были экспериментально установлены все пять случаев парных генетических перекрытий, которые разрешаются структурой ДНК. Пониманию этого феномена во многом способствовало то, что уже к 1992 г. объем публикаций по перекрытиям генов был значительным и все более и более увеличивался. Это были публикации в журналах Nature, Cell, J. Mol. Biol., J. Virol., Genetics, The J. Biological Chemistry, J. Vol. Evol. и др. Первоначально в исследовании речь не шла о свойствах генетического кода. Вопрос был поставлен только о потенциальных позициях молчащих мутаций, которые могут иметь место в областях, занятыми перекрывающимися генами. Было установлено значительное (на порядок и более) сужение спектра подобных позиций по сравнению с генами без перекрытий. При анализе этого спектра для сотен генетических перекрытий были обнаружены ряд геномов, в которых перекрытия требовали участия всех смысловых кодонов. Стало ясно, что из перекрытий можно каким-то

образом выявить неизвестное свойство структуры генетического кода. Была поставлена задача изучения потенциала генетического кода, который использован природой для всех пяти случаев перекрытий. Главный результат был представлен теоремой для стандартного (первоначальное название — универсального) генетического кода (см. работы [11, 12]). Биологические следствия из нее позволили обратиться к анализу экспериментальных данных по всем девиантным генетическим кодам, или кодам, отклоненным от стандартного. Однако в рамках генетических перекрытий не удалось объяснить функциональную значимость всех переосмысленных кодонов, или кодонов, которые изменили свой смысл по сравнению со стандартным кодом. Путей дальнейших исследований было несколько. Требуемое решение было найдено при исследовании областей ДНК, где гены не перекрываются, а подчиняются принципу, сформулированному как предположение в 1941 г.: один ген отвечает за один белок (см. работу [13]). Таких генов — подавляющее большинство; на сегодня это миллиардные нуклеотидные последовательности больших геномов (в том числе человека).

Как видим, на всех этапах исследования наши математические утверждения подтверждались или дополнялись разнообразными экспериментальными данными, а именно: многими случаями перекрытий пар генов, в том числе записанными нестандартными кодами, полным набором природных нестандартных кодов, полными кодирующим областями больших геномов, в том числе генома человека. В ходе проведения данного исследования был использован наш опыт проведения более ранних работ по изучению эволюции и структуры сложных природных и технических дискретных систем с большим числом взаимодействующих элементов.

Представим кратко эти исследования, в которых я принимал участие. Из этого представления станет ясно, что перекрывающиеся гены, с которых были начаты наши исследования, есть еще одна достаточно сложная система, состоящая из большого числа взаимозависимых элементов. Перед представлением первой из таких задач следует отметить, что именно это исследование в итоге привело нас к биологической проблематике. Речь идет об изучении гравитационного взаимодействия галактик по компьютерной программе, созданной первоначально для изучения движения искусственного спутника Луны в поле тяготения, где помимо центрального тела учитывались масконы — вкрапления некоторого количества тел, которые были установлены экспериментально в статье [15]. Результаты изложены в работе [14] по космическим исследованиям, в которых заметное место занимали исследования по оптимизации процесса траекторных измерений в случае ИС Марса (см. также статьи [16–20] и рис. В.1).

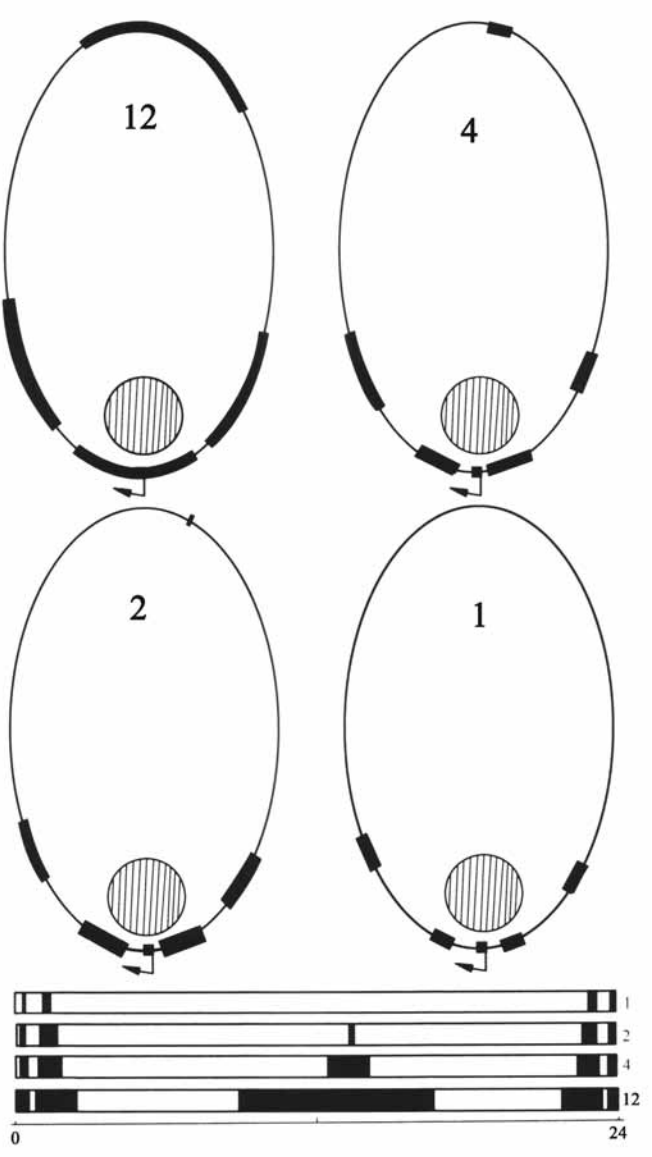


Рис. В.1. Оптимальные сеансы измерений радиальной скорости на одном обороте ИС Марса [17]. Приводятся решения для наилучшего определения минимального расстояния до Марса. Сеансы располагаются по времени (спектры) и по истинной аномалии (на орбитах). Время 1, 2, 4, 12 — допустимое время измерений в часах, период обращения спутника — 24 часа

ГРАВИТАЦИОННОЕ ВЗАИМОДЕЙСТВИЕ ГАЛАКТИК

При близком пролете массивного тела мимо галактики, как показали численные эксперименты, возникают специфические приливные эффекты, появляются спиральные ветви, значительные отклонения от плоскости диска, искажения поля скоростей вещества, падение газа на плоскость галактик. Качественный анализ и численный эксперимент позволили выявить основные эффекты, возникающие при характерных вариантах гиперболического пролета массивного тела относительно галактик: перпендикулярно ее плоскости, над плоскостью и в плоскости по направлению и против направления вращения галактики. Анализировалось поведение до 2000 невзаимодействующих между собой точек-спутников, двигавшихся первоначально по круговым кеплеровским орбитам вокруг центральных областей галактики и возмущаемых при близком пролете массивного тела (см. статьи [21–26]). На основе расчетов в 1973 г. был создан компьютерный кинофильм продолжительностью около 15 минут. Оценка этой работы дана Президентом АН СССР, академиком М. В. Келдышем на торжественном заседании, посвященном двадцатилетию созданного им ИПМ. Приводим выдержку из его доклада[27]. «Ярким примером успешного применения машинных расчетов к классической задаче является работа по гравитационному взаимодействию галактик. Изготовленный вычислительной машиной кинофильм наглядно показывает образование у галактик спиральной структуры. По-видимому, именно таким путем возникла спиральная структура нашей собственной галактики» На рис. В.2 представлены 6 кадров из около 2000 кадров одного из вариантов такого пролета.

Все 7 эпизодов пролета (по 6 кадров в каждом) представлены в статье [25]. Создание наиболее полной версии кинофильма было приурочено к чрезвычайной сессии МАС, посвященной 500-летию Коперника, проходившей в 1973 г.. Первоначальным местом проведения этой сессии бала выбрана Австралия, где в заседала галактическая секция. Однако затем некоторые секции были перенесены на родину Коперника, где на секциях небесной механики и астрофизики, и был впервые показан этот кинофильм. Лишь много лет спустя зарубежным специалистам была представлена только цифровая копия фильма на конференции Dynamics of Galaxies. (Санкт-Петербург, 2007 г.)

МОДЕЛЬ АККУМУЛЯЦИОННОГО ПРОЦЕССА ФОРМИРОВАНИЯ ПЛАНЕТНЫХ СИСТЕМ

Рассматривалась эволюция плоского протопланетного облака, состоящего из большого числа гравитационно взаимодействующих и объединяющихся при контактах тел (протопланет), движущихся в поле цен-

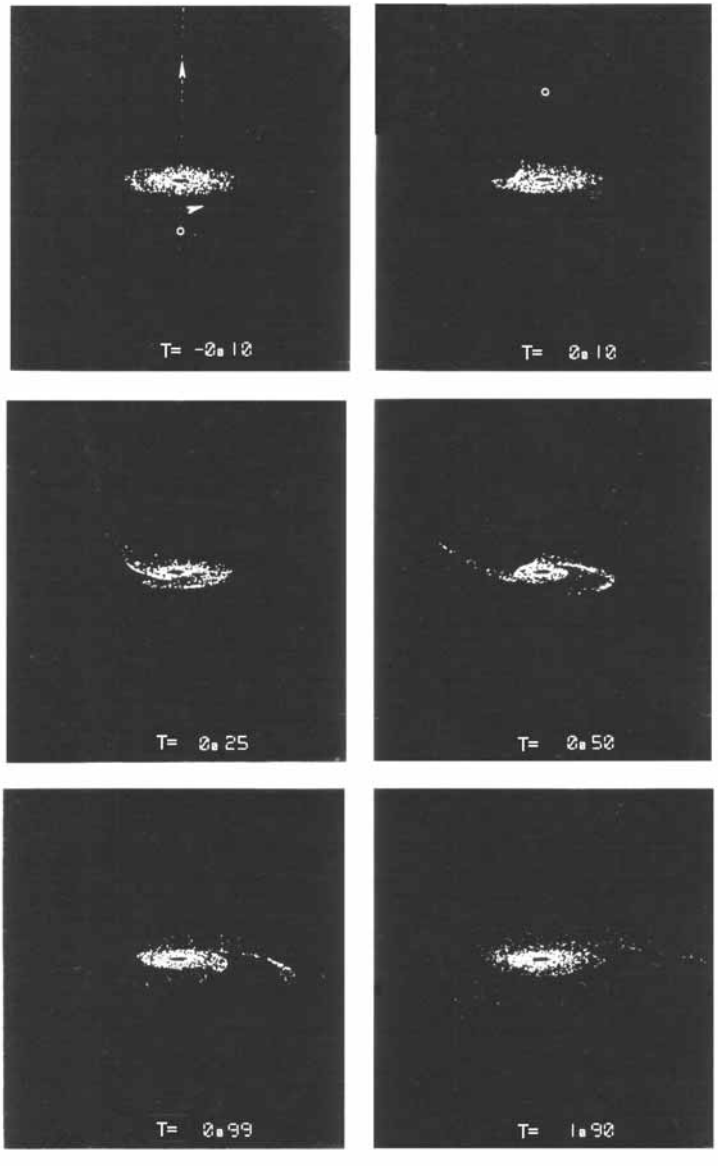


Рис. В.2. Фрагменты кинофильма (ИПМ, 1973 г.) относятся к варианту пролета тела с массой равной массе галактики, пролетающего вблизи галактического диска (с параметрами порядка нашей галактики), перпендикулярно его плоскости с удвоенной параболической скоростью. Время T дается в миллиардах лет, момент $T = 0$ соответствует моменту наибольшего сближения (см. статьи [24, 25])

трального массивного тела (Солнца или планеты). Предполагалось, что гравитационное взаимодействие между телами имеет место лишь при их бинарном тесном сближении. Предполагалось также, что от одного тесного сближения до другого тела движутся по кеплеровским орбитам, причем орбиты всех тел в начальный момент эволюции облака являются круговыми. Рассматривались так называемая предельная модель процесса аккумуляции, в которой каждое тесное сближение тел заканчивается их объединением. Показывается, что в ходе эволюции такой модели в ней появляются кольцевые зоны уплотнения вещества, последующее развитие которых приводит к образованию планет (см. работы [28–34], а также рис. В.3).

Основные численные результаты работы были получены с помощью моделирования процесса аккумуляции планет на БЭСМ-6. (см. также следующий пункт) Кроме того, одновременно с образованием самих планет изучался механизм формирования вращательного движения планет. Показывается, что подавляющее большинство крупных тел, образующихся в конце аккумуляционного процесса, приобретают прямое (т. е. такое же, как и орбитальное) вращение вокруг своих осей. Одним из важнейших результатов численных экспериментов является установление возможности обратного вращения протопланет Венеры и Урана к моменту образования планет из протопланетного облака. Поскольку указанные протопланеты обладали достаточно большими размерами, то большую роль должна играть приливная эволюция вращательного движения. Член-корреспондент В. В. Белецкий исследовал этот вопрос и показал различие в результате указанной эволюции в случае Урана, по сравнению с Венерой (см. статью [35]).

На основе анализа формирования вращения планет была установлена связь между предельной моделью аккумуляционного процесса и теорией гравитационной неустойчивости допланетного газопылевого облака. Этот анализ позволил также дать четкую физическую интерпретацию и смысл предельной модели процесса аккумуляции. В итоге проведенных исследований академик Т. М. Энеев создал новую модель процесса аккумуляции планетных систем.

По нашим данным проведенное исследование и прежде всего численные эксперименты не были до сих пор повторены за рубежом. Ссылка на английский перевод нашей первой работы по данному исследованию [28], датированной 1977 г., представлена на сайте NASA (см. ссылку на эту работу). Причина этого таится по-видимому в достаточно нестандартной методике расчета (см. следующий пункт), так как расчет, основанный на простом переборе не представляется возможным за приемлемое время даже на самых современных супер-ЭВМ.

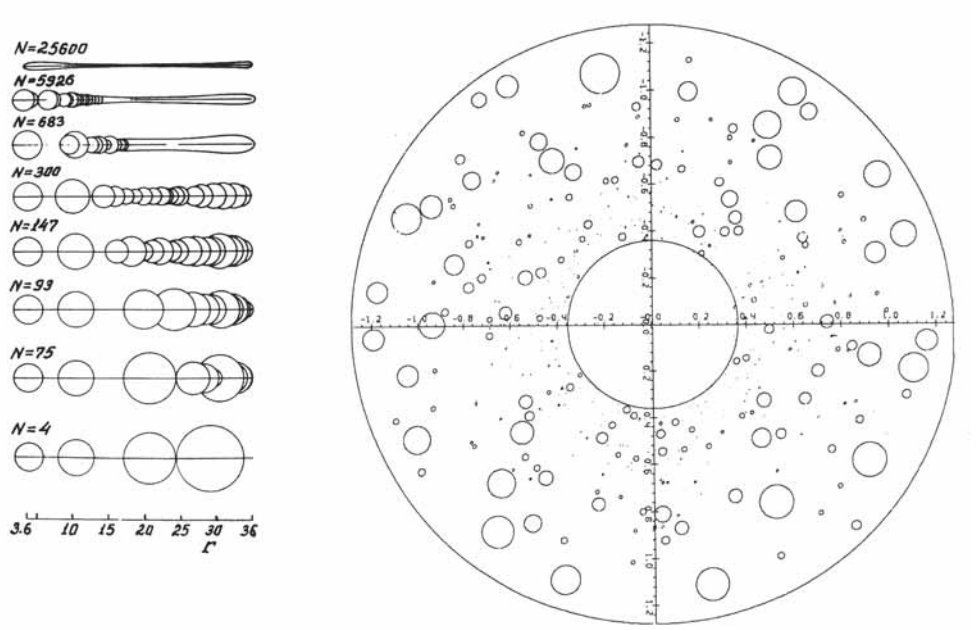


Рис. В.3. *Слева*: фрагменты образования планет-гигантов в одной из моделей [31]. Число тел N изменяется в ходе аккумуляционного процесса от исходного (25 600) до финального (4). Приводится радиальная проекция, по оси абсцисс — расстояние от Солнца в астрономических единицах. *Справа*: фрагмент одного из вариантов образования планет в узком кольце. Все исходные тела имеют одинаковые размеры. Приводится состояние системы для 500 тел или когда исходное число тел — 10^6 уменьшилось в 2000 раз. Программа составлена на БЭСМ-6 в 1981 г. В. Н. Торопцевой на основе алгоритма из главы 6 работы [41]. Публикуется впервые

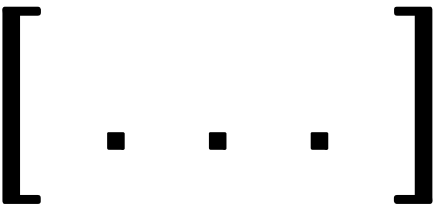
МЕТОД ВИРТУАЛЬНЫХ КОНТАКТОВ

Рассматривается новый подход к исследованию с помощью ЭВМ эволюции сложных дискретных систем, состоящих из большого числа $N \gg 1$ контактирующих элементов. На основе этого подхода был разработан метод, получивший название метода виртуальных контактов. Показывается, что использование этого метода приводит к затратам времени ЭВМ порядка N^2, в отличие от подхода, основанного на полном переборе, когда эти затраты имеют порядок N^3. Метод в своей основе может использоваться для широкого круга столкновительных и коагуляционных процессов (см. работы [36–46]).

Непосредственно созданный метод разрабатывался и был использован при проведении численных экспериментов по имитации формирования планетных систем в новой модели в случае $N = 25\,600$, при этом была выявлена весьма высокая его эффективность, характерное время расчета оказалось порядка $N^{3/2}$; уменьшение времени счета по сравнению с N^2 было достигнуто за счет учета специфики изучаемой модели. В ходе проведения численных экспериментов была выявлена специфика протекания процесса формирования планетных систем в рассматриваемой модели, что позволило разработать методику расчета такой модели за время порядка N в диссертации [41]. Методика была опробована на единственном экспериментальном варианте расчета в котором исходное число прототел было равно 1 млн (фрагмент расчета приведен на рис. В.3).Фактически эта работа завершилась в 1981 г. В настоящее время в связи с появлением новейших супер-ЭВМ нами предпринимаются попытки повторить расчет указанной модели.

ТРАССИРОВКА БИС

Предлагается новый подход к решению некоторых задач, возникающих при конструкторском проектировании двухслойных БИС. Этот подход основывается на математическом анализе множественных конфликтов в препринте [47]. На основе этого подхода было создано несколько версий алгоритма трассировки (см. работы [48–50]). Наиболее общий из таких алгоритмов позволяет определять такие сочетания всех трасс на двухслойной плате, при которых число точек межслойных переходов становится минимальным. На ПК был опробован такой алгоритм для известного типа двухслойных БИС содержащего около 1500 трасс с более чем 3250 выводами, так как не все трассы были двухконцевыми (см. рис. В.4). Оказалось, что число точек межслойных переходов более чем вдвое сокращается по сравнению со случаем ортогонального расслоения (см. отчет [50]).



Козлов Николай Николаевич – доктор физико-математических наук, главный научный сотрудник Института прикладной математики им. М.В. Келдыша РАН.

Круг научных интересов автора чрезвычайно широк, о чем свидетельствует его участие в работах по анализу космических траекторий ИС Луны, ИС Марса, по эволюции сложных дискретных структур: изучение гравитационного взаимодействия галактик и моделирование процесса формирования планет из протопланетного облака. Также исследовалась задача компьютерного проектирования БИС. В ходе математического моделирования процесса структуризации вторичных структур РНК автором была поставлена задача анализа необычных способов записи генетической информации – перекрывающихся генов. Оказалось, что такие гены исключительно важны для выявления и анализа новых свойств генетического кода, которые могут быть получены только математическими методами.

Используя большой опыт изучения сложных природных и технических дискретных систем различных типов, автор нашел свой оригинальный подход к решению поставленных задач. Главные положения такого подхода излагаются в данной монографии.