БИООРГАНИЧЕСКАЯ ХИМИЯ
том 14 * № 4 * 1988

УДК 577.113.5.088 : 519.85

РЕКОНСТРУКЦИЯ ДЛИННЫХ ПОЛИНУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ИЗ ФРАГМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ПЕРСОНАЛЬНОЙ ЭВМ «ИСКРА-226»

Костецкий П. В., Доброва И. Е.

Институт биоорганической химии им. М. М. Шемякина Академии наук СССР, Москва

Предложен алгоритм реконструкции ДНК из нуклеотидных последовательностей фрагментов. Алгоритм реализован на языке БЭЙСИК применительно к отечественной персональной ЭВМ «Искра-226». Программа работает в режиме, не требующем вмешательства экспериментатора. Результат работы программы – один или несколько контигов, составленных из фрагментов ДНК, прочитанных с электрофоретических гелей. С помощью программы можно реконструировать последовательность ДНК длиной до 10 т.п.о.

Процесс установления последовательностей нуклеиновых кислот постоянно совершенствуется [1, 2], что ведет к быстрому увеличению накопленной структурной информации, объем которой составляет в настоящее время более 5 млн. нуклеотидов [3, 4]. Необходимым этапом при изучении линейной структуры нуклеиновых кислот является первоначальное расщепление молекулы исходного биополимера на относительно короткие фрагменты, нуклеотидная последовательность которых прочитывается целиком с помощью гель-электрофореза. Поскольку длина фрагментов достигает 200 и более нуклеотидов, а число исследуемых фрагментов обычно превышает 100, их хранение и сравнение выполняют с помощью ЭВМ. Для этой цели применяют как ЭВМ средней мощности [5–10], так и персональные ЭВМ [11–13].

При сборке нуклеиновых кислот из фрагментов важным этапом является нахождение всех возможных пар перекрывающихся фрагментов. Для такого поиска наиболее часто применяется способ, предложенный Гингерасом с сотр. [6]. Перекрывающимися считаются последовательности, в которых имеются идентичные участки длиной не менее 10 нуклеотидов. В сравнении участвуют как прямые, так и инвертированные последовательности. В связи с тем что один и тот же участок нуклеиновой кислоты часто прочитывается по 3 и более раз в виде соответствующих фрагментов, важно уметь автоматически находить непрерывные блоки взаимно перекрывающихся фрагментов, или, согласно принятой номенклатуре, контиги [7].

При установлении первичной структуры ДНК длиной более 5 т.п.о. число исследуемых фрагментов превышает 100 и сравнение каждой новой серии нуклеотидных последовательностей с уже имеющимися последовательностями требует больших затрат машинного времени даже при использовании быстрых ЭВМ [5–10]. В связи с этим Стаден [8] предложил вместо группы перекрывающихся последовательностей, образующих контиги, использовать соответствующие обобщенные последовательности, называемые консенсусами. Стаден, однако, не сообщает в своих работах подробности формирования контигов из исходного «базового» набора фрагментов ДНК, ограничиваясь главным образом описанием функций программ, обрабатывающих контиги, консенсусы и новые олигонуклеотидные последовательности. Вместе с тем очевидно, что при использовании произвольных ЭВМ с иной операционной системой важно располагать алгоритмом образования контигов из первоначального набора прочитан-

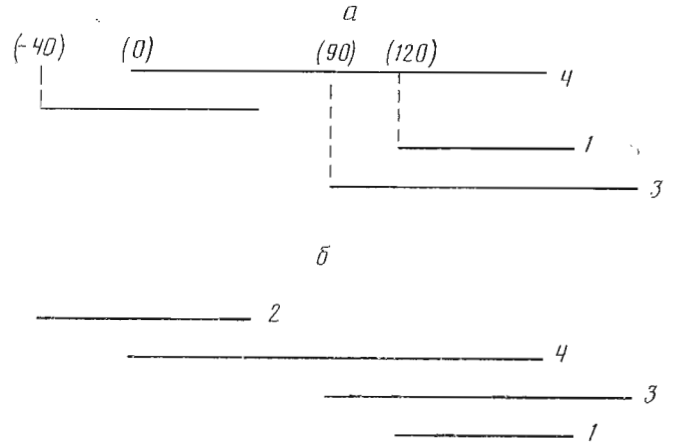

Рис. 1. Начало формирования контига. *а* – фрагменты, перекрывающие опорный фрагмент «4»; нахождение координат 5′-концов фрагментов; *б* – определение правильного чередования фрагментов. Номера фрагментов указаны справа; цифры в скобках над фрагментом «4» отвечают координатам 5′-концов фрагментов

ных гелей. В настоящей работе сообщается один из таких возможных алгоритмов, реализованный на отечественной персональной ЭВМ «Искра-226» на языке БЭЙСИК.

В начале формирования контигов один из наиболее длинных и надежно прочитанных фрагментов ДНК (далее — опорный фрагмент) сравнивается с остальными (рис. 1). Таким образом выявляются все последовательности, перекрывающие опорный фрагмент. Результатом попарного сравнения являются координаты 5′-концов выявленных последовательностей относительно 5′-конца опорного фрагмента. Имея значения координат 5′-концов, несложно упорядочить расположение всех перекрывающихся последовательностей таким образом, чтобы 5′-конец каждого последующего из них в контиге оказался правее 5′-конца предыдущего (рис. 1*б*).

Продолжение контига в обе стороны формируется аналогично, но при этом в качестве опорных выбираются левый верхний фрагмент и тот из фрагментов, чей 3′-конец располагается наиболее далеко вправо (рис. 1). При наращивании контига вправо, опорным становится не обязательно самый нижний фрагмент. Парные перекрывания находят с прямым или инвертированным опорным фрагментом в зависимости от его направления в уже сформированной части контига.

Все последовательности, вошедшие внутрь растущего контига, исключаются из базового набора, что значительно ускоряет процесс сборки (см. ниже). Возможно, что какая-нибудь из них войдет внутрь контига не на данном, а на следующем этапе сборки. В этом случае все последовательности займут свои места в контиге в момент упорядочивания расположения 5′-концов. Сборка считается законченной, когда его 5′- и 3′-концевые участки последовательностей не дают перекрытий с оставшимися фрагментами. Для построения следующего контига в качестве опорного фрагмента выбирают наиболее длинный из числа не вошедших в предыдущие контиги. Интересно, что даже случайное перекрытие двух нуклеотидных последовательностей (вероятность этого события крайне мала) не обязательно помешает формированию контига, поскольку сбившаяся последовательность может оказаться внутренним, а не опорным фрагментом. Если же некоторые нуклеотидные последовательности из-за невысокого качества интерпретации секвенируемых гелей или по другим причинам не дадут нужных парных перекрытий и не войдут ни в один из контигов, процесс сборки полной структуры не должен нарушиться по причине избыточности общего числа всех последовательностей.

Естественно, что число правильно найденных парных перекрытий зависит от минимально допустимой длины идентичных участков. Из-за опасности возникновения случайных перекрытий эту длину нецелесообразно назначать менее 10 нуклеотидов [6, 8]. Более того, для успешной сборки контигов минимальную длину идентичных участков обычно повышают до 15 нуклеотидов [8], что может привести к потере части правильных перекрытий. Для поиска перекрываний олигонуклеотидных фрагментов мы применили другой критерий, согласно которому подсчитывалась сумма длин идентичных участков протяженностью не менее 10 нуклеотидов. Перекрывание сравниваемой пары фрагментов считалось правильным, если указанная сумма составляла не менее 15 нуклеотидов, что означает наличие одного общего безошибочного участка из 15 нуклеотидов или двух таких участков длиной по 10 нуклеотидов. Естественно, оба участка не должны располагаться слишком далеко друг от друга, что характерно при случайном перекрытии. В нашем алгоритме допускалось расстояние между общими участками до 20 нуклеотидов.

Для обоснования выбранного критерия перекрывания нами моделировалось возникновение случайных перекрытий. С этой целью при помощи ЭВМ генерировались пары произвольных нуклеотидных последовательностей длиной по 200 оснований, которые сравнивались между собой. Оказалось, что случайные перекрытия происходят с заметной частотой, лишь когда длины обоих идентичных участков меньше 10 нуклеотидов. Напротив, случайных перекрытий не наблюдали вовсе на десятках тысяч сравнений, если эти длины составляли 10 и более нуклеотидов.

На основе описанного выше алгоритма была разработана программа «КОНТИГ». Программа обрабатывает фрагменты реконструируемой ДНК, хранящиеся на гибком магнитном диске в виде отдельных файлов в формате, предложенном Александровым [13].

Емкость одного диска вполне достаточна для хранения 300 и более фрагментов реконструируемой ДНК, причем длина некоторых из них может достигать 5000 нуклеотидов, что дает возможность использовать программу и для опознания олигонуклеотидов с помощью известной гомологичной последовательности. Перед началом работы программы указываются номера обрабатываемых фрагментов, после чего автоматически составляется список номеров в соответствии с убыванием длины фрагментов.

На формирование контига, отвечающего гену α-субъединицы Na,K-зависимой АТР-азы [14] длиной 3420 нуклеотидов из набора 70 фрагментов средней длины в 200 нуклеотидов с помощью программы «КОНТИГ», требуется ~1,5 ч. Следует иметь в виду, что время работы программы зависит как от числа обрабатываемых последовательностей, так и от их длины. Это время можно заметно сократить, если при сравнении двух фрагментов для более короткого из них ограничиться анализом только концевых участков по 50–60 нуклеотидов. В любом случае время работы программы гораздо меньше времени, необходимого для ввода исходных последовательностей в ЭВМ.

Результатом работы программы образования контигов является таблица, содержащая всю необходимую информацию для последующей распечатки контигов. Таблица сохраняется на магнитном диске и может быть использована для повторной печати контигов.

К аналогичной таблице контигов приводит также алгоритм для персонального компьютера IBM PC, предложенный американскими авторами [11]. В этом алгоритме при формировании контигов из исходных фрагментов с учетом их инверсии необходимо произвести $n(n-1)$ парных сравнений, что при наличии 100 нуклеотидных последовательностей составит почти 10 000 сравнений, каждое из которых в свою очередь требует выполнения большого числа операций. Предлагаемый в настоящей работе алгоритм требует в несколько раз меньшего количества парных сравнений. Действительно, первый из опорных фрагментов сравнивается со всеми остальными, тогда как на последующих этапах сборки контига только очередной опорный фрагмент участвует в парных сравнениях, а фрагмен-

Данные о структуре контига гена α-субъединицы Na, K-АТР-азы, сформированного из набора фрагментов
Знак «–» является признаком инверсии фрагмента

Номер фрагмента	Координата 5′-конца фрагмента в контиге	Длина фрагмента	Номер фрагмента	Координата 5′-конца фрагмента в контиге	Длина фрагмента
38	1	166	11	485	118
14	1	112	23	508	126
32	1	127	5	561	200
–20	1	132	36	587	172
26	1	170	12	603	246
21	132	166	30	639	207
2	138	122	37	759	164
3	260	178	43	763	238
28	307	130	25	808	193
40	335	108	19	822	101
–34	347	108	31	846	123
17	429	174	13	849	117
35	455	132	7	865	136

ты, вошедшие внутрь растущего контига, выбывают из дальнейших парных сравнений. Экономия в вычислениях будет, очевидно, тем заметней, чем выше избыточность числа фрагментов в формируемом контиге. Так, в рассмотренном выше примере сборки гена α-субъединицы Na,K-зависимой АТР-азы из набора в 70 фрагментов со средней длиной 200 нуклеотидов (соответствует 4-кратному разбиению исходной цепи ДНК) число всех возможных парных сравнений составляет 4900. Однако при построении контига по предлагаемому нами алгоритму понадобилось только 1600 сравнений, т. е. втрое меньше. При 6-кратном разбиении исходной цепи того же гена экономия в числе парных сравнений, как и ожидалось, оказалась еще выше: 2000 вместо 10 000. В целом сборка контига происходит гораздо быстрее, так как время, затрачиваемое на парные сравнения фрагментов, составляет не менее 90% от полного времени работы программы «КОНТИГ».

При реконструкции ДНК длиной до 10 000 оснований исследователю достаточно примерно 200 фрагментов со средней длиной 200 нуклеотидов. Для обработки этих данных на ЭВМ «Искра-226» потребуется ~12,5 ч непрерывного счетного времени, что представляется вполне допустимым, так как программа работает в режиме, не требующем вмешательства экспериментатора. Эта возможность является полезной особенностью примененного алгоритма, тогда как в настоящее время в основном используются интерактивные программы [5–10, 12, 13, 15]. Следует иметь в виду, что на ранних этапах секвенирования имеющиеся олигонуклеотидные фрагменты не обязательно образуют в совокупности всю исходную последовательность ДНК. В этом случае результатом работы программы будет набор не связанных между собой контигов.

Эффективность работы описанного выше алгоритма была проверена на модели, основой которой явилась многократная сборка нуклеотидной последовательности, кодирующей α-субъединицу Na,K-зависимой АТР-азы. Предварительно одиночная цепь ДНК с помощью программного датчика псевдослучайных чисел произвольным образом разбивалась на фрагменты длиной от 100 до 250 нуклеотидов. Поскольку реальные нуклеотидные последовательности, читаемые с гелей, содержат ошибки в виде неверно расшифрованных, пропущенных или вставленных оснований, нами также вносились дефекты во фрагменты, участвующие в модельных сборках гена Na,K-зависимой АТР-азы, а именно: в каждый фрагмент случайным образом вносилось 7% ошибок — 4% замен и 3% вставок и делеций. В результате получали один набор пересекающихся фрагментов одиночной цепи исходной ДНК. Аналогичным образом выполняли еще одно разбиение полинуклеотидной цепи, что обеспечивало получение полного набора перекрывающихся фрагментов. Однако, если точки щепления ДНК в двух наборах окажутся расположенными слишком близко друг к другу, пере-

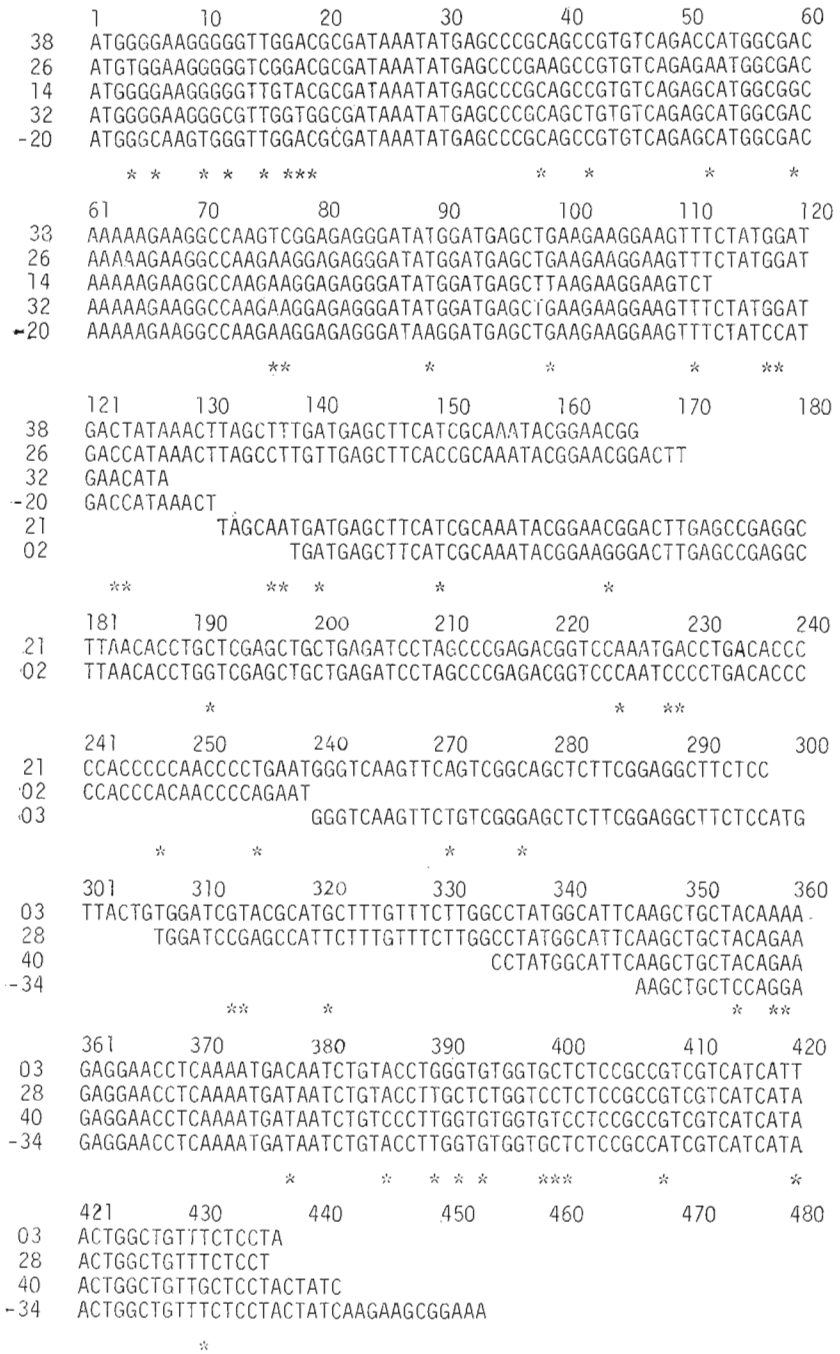

Рис. 2. Контиг, полученный в результате модельной сборки последовательности 5′-концевой части гена α-субъединицы Na,K-зависимой АТР-азы почек свиньи на основе большого числа олигонуклеотидных фрагментов. Левая колонка цифр составлена из номеров фрагментов, образующих контиг. Знак «–» является признаком инверсии фрагмента. Случаи несовпадения нуклеотидов в какой-либо позиции отмечаются символом «*» в нижней строке

крытие соответствующих фрагментов может быть неочевидным. Поэтому, чтобы увеличить вероятность надежного перекрывания фрагментов по всей длине реконструируемой ДНК, нами так же, как и в лабораторной практике, применялось 4–5-кратное разбиение исходной полинуклеотидной цепи. В реальных «базовых» наборах встречаются фрагменты обеих комплементарных цепей ДНК, поэтому некоторые фрагменты полученного набора инвертировались.

Оказалось, что при уровне ошибок до 7% в нуклеотидных последовательностях исходных фрагментов процесс модельной сборки ДНК осуществляется в основном безошибочно и результатом является получение одного непрерывного контига. При этом в отдельных случаях контиг длиной 3420 нуклеотидов распадался на два-три более коротких контига. При уровне ошибок более 7% резко возрастало число случаев распадения контига. Контиг распадался в тех участках исходной цепи, где точки щепления одновременно во всех разбиениях располагались слишком близко, и перекрытие фрагментов становилось неочевидным.

Интересно, что при использовании для поиска парных перекрытий традиционного критерия (наличие общего безошибочного участка из 15 нуклеотидов [8]) число случаев распадения единого контига при 7% ошибок возрастает в 1,5 раза. Это свидетельствует в пользу того, что примененный в настоящей работе комбинированный критерий более эффективен, т. е. наличие вспомогательного условия перекрытия в виде двух близкорасположенных идентичных участков длиной 10 нуклеотидов существенно повышает стабильность контига. В случае заметного числа ошибок, неизбежных при чтении последовательностей длинных олигонуклеотидов, для некоторых перекрывающихся фрагментов наличие двух идентичных участков длиной по 10 нуклеотидов более вероятно, чем присутствие одного общего участка из 15 нуклеотидов. Можно надеяться, что предлагаемый нами критерий, включающий в себя и традиционный как один из возможных вариантов, окажется вполне приемлемым для реконструкции цепей ДНК при использовании данных невысокого качества.

На рис. 2 представлена 5′-концевая часть контига, полученного в результате сборки гена α-субъединицы Na,K-зависимой АТР-азы почек свиньи из большого числа фрагментов. Следует отметить, что для приближения к реальным условиям в данной модельной сборке ~30% фрагментов считались потерянными и не участвовали в реконструкции исходной цепи ДНК. Время, необходимое для распечатки контига, зависит от его длины и числа входящих в него последовательностей. Для распечатки контига гена потребовалось ~40 мин. Это время, конечно, включает в себя и обработку таблицы контигов, которая хранится на магнитном диске и содержит в компактном виде всю необходимую промежуточную числовую информацию (таблица). Важной особенностью описываемой программы «КОНТИГ» является детекция тех позиций контига, в которых имеет место несовпадение нуклеотидов. В примере сборки, приведенном на рис. 2, отсутствуют дефекты в виде делеций или вставок нуклеотидов. Естественно, что наличие такого дефекта в какой-либо позиции резко увеличит число несовпадений в прилегающей части контига. Для устранения подобного рода дефектов нами разрабатывается программа редакции контига, позволяющая в диалоговом режиме вставлять, заменять и удалять символы в отдельных позициях.

Программа «КОНТИГ» рассчитана на использование специалистами по определению последовательности ДНК, обладающими минимальными навыками применения ЭВМ в лабораторных исследованиях, и распространяется в виде объектного модуля, записанного на дискете размером 203 мм.

Авторы признательны Г. С. Монастырской и Ю. В. Смирнову за участие в постановке задачи и ценные рекомендации, высказанные ими в процессе обсуждения отдельных этапов ее выполнения, и Ю. П. Козьмину за помощь в разработке программы.

ЛИТЕРАТУРА

1. *Messing J.* // Meth. Enzymol. 1983. V. 101. P. 20–78.
2. *Sanger F., Coulson A. R., Barrel B. G., Smith A. J. H., Roe B. A.* // J. Mol. Biol. 1980. V. 143. № 2. P. 161–178.
3. *Bilofsky H. S., Burks C., Fickett J. W., Goad W. B., Lewitter F. I., Rindone W. P., Swindell C. D., Tung C.-S.* // Nucl. Acids Res. 1986. V. 14. № 1. P. 1–4.
4. *Hamm G. M., Cameron G. N.* // Nucl. Acids Res. 1986. V. 14. № 1. P. 5–10.
5. *Grymes R. A., Travers P., Engelberg A.* // Nucl. Acids Res. 1986. V. 14. № 1. P. 87–98.
6. *Gingeras T. R., Milazzo J. P., Sciaky D., Roberts R. J.* // Nucl. Acids Res. 1979. V. 7. № 2. P. 529–545.

7. *Staden R.* // Nucl. Acids Res. 1980. V. 8. № 1. P. 3673–3693.
8. *Staden R.* // Nucl. Acids Res. 1982. V. 10. № 15. P. 4731–4751.
9. *Staden R.* // Nucl. Acids Res. 1986. V. 14. № 1. P. 217–231.
10. *Peltola H., Soderlung H., Ukkonen E.* // Nucl. Acids Res. 1984. V. 12. № 1. P. 307–321.
11. *Johnston R. E., Mackenzie J. M., Jr., Dougherty W. G.* // Nucl. Acids Res. 1986. V. 14. № 1. P. 517–528.
12. *De Bansie J. S., Steeg E. W., Lis J. T.* // Nucl. Acids Res. 1984. V. 12. № 1. P. 619–625.
13. *Александров А. А.* // Журн. Всес. хим. о-ва им. Д. И. Менделеева. 1984. № 2. С. 64–69.
14. *Овчинников Ю. А., Арсения С. Г., Броуде Н. Е., Петрухин К. Е., Гришин А. В., Алданова Н. А., Арзамазова Н. М., Аристархова Е. А., Мелков А. М., Смирнов Ю. В., Гурьев С. О., Монастырская Г. С., Модянов Н. Н.* // Докл. АН СССР. 1985. Т. 285. № 6. С. 1490–1495.
15. *Миронов А. А., Александров Н. Н., Дюновская-Гурова Л. В., Кистер А. Э.* // Молекуляр. биология. 1987. Т. 21. № 3. С. 672–677.

Поступила в редакцию
29.IV.1987
После доработки
2.X.1987

RECONSTRUCTION OF LONG NUCLEOTIDE SEQUENCES FROM THE FRAGMENTS USING PERSONAL COMPUTER «ISKRA-226»

KOSTETSKY P. V., DOBROVA I. E.

M. M. Shemyakin Institute of Bioorganic Chemistry, Academy of Sciences of the USSR, Moscow

An algorithm for reconstructing long DNA sequences, i. e. arranging all overlapping gel readings in the contigs, and the corresponding BASIC programme for personal computer «Iskra-226» (USSR) are described. The contig construction begins with the search for all fragments overlapping the basic (longest) one follower by determination of coordinates of 5′ ends of the overlapping fragments. Then the gel reading with minimal 5′ end coordinate and the gel reading with maximal 3′ end coordinate are selected and used as basic ones at the next assembly steps. The procedure is finished when no gel reading overlapping the basic one can be found. All gel readings entered the contig are ignored at the next steps of the assembly. Finally, one or several contigs consisted of DNA fragments are obtained. Effectiveness of the algorithm was tested on a model based on the multiple assembly of the nucleotide sequence, encoding the Na, K-ATPase α-subunit of pig kidney. The programme does not call for user's participation and can comprise contigs up to 10 000 nucleotides long.