Rectangle Graph

May 10, 2011

1 From k-mer pairs to edge pairs

Define a k-mer as a string of length k. Given a circular string $S = s_1 \dots s_n$, $S_k(i)$ is the k-mer $s_i \dots s_{i+k-1}$. A (k,d)-*mer* of S is a pair of k-mers $S_k(i)$ and $S_k(i+d)$. A set of all possible (k,d)-mers of S is called a (k,d)-spectrum of S. The paired de Bruijn graph utilizes this set directly by representing each (k,d)-mer as a directed edge $u \to v$ where u is labeled by the $(k-1,d)$-mer prefix and v is labeled by the $(k-1,d)$-mer suffix. Nodes with the same labels are merged (glued) in this graph. This direct use of mate-pair leads to cases where the contigs produced by the paired de bruijn graph are more fragmented than the ones produced by the normal de Bruijn graph (Fig. 1).

Given an additional parameter Δ with a nonnegative integer value, a pair of strings (a,b) is called a (k,d,Δ)**-mer** of S that **maps** to position i if there exists x such that $S_k(i) = a$ and $S_k(i+d+x) = b$ where $0 < i < n$ and $-\Delta \le x \le \Delta$. A set P of pairs of strings are called (k,d,Δ)-spectrum of S iff every element in this set is a (k,d,Δ)-mer of S and for every $i \in [0,n]$ there exists at least one (k,d,Δ)-mer in P that maps to this position. The approximate paired de Bruijn graph assemble the genome from this

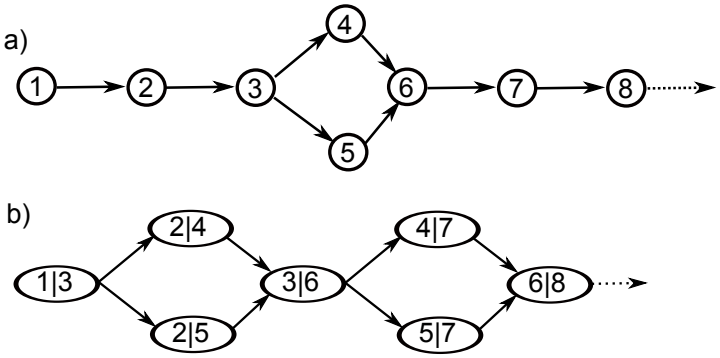


Figure 1: A case where paired de Bruijn graph (b) is more complicated than the traditional de Bruijn graph (a).

data set by modifying the gluing rule used for the (k, d)-spectrum dataset slightly. For the approximated distance dataset, two node $(a|b)$ and $(c|d)$ are glued if a = c and the distance between b and d in the de Bruijn graph is not greater than Δ. In general, the approximated paired de Bruijn graph performs better with small values of Δ. This leads to a requirement to decrease the variance of the insert size either from the biotechnology or algorithmic side. While we are not aware of the feasibility of this problem in the biotechnology side, we believe that the information of the vicinity mate-pairs is valuable to tighten the variance of the insert size. The use of single (k, d, δ) mers in the approximated de bruijn graph make it difficult to utilize this information. Below, we define a different set of data: edge-pairs and describe an approach to assemble the genome from this type of dataset.

1.1 From (k, d)-mers to edge-pairs

Consider the (k, d)-spectrum of S and let B be a condensed de Bruijn graph constructed from the k-mers spectrum of S. Every (k, d)-mer $(a|b)$ maps to edges e_a, e_b at positions p_a, p_b correspondingly. For each (k, d)-mer $(a|b)$, we form an *edge-pair*, represented by a triple $(e_a, e_b, d + p_a - p_b)$ where $d + p_a - p_b$ represents the genomic distance between the starts of e_a and e_b. As a result, the (k, d)-spectrum is transformed into a set of edge-pairs. The set of edge-pairs generated from the (k, d)-spectrum of a string S is called an *d-edge-pair spectrum* of S.

Problem Construct contigs from a d-edge-pair spectrum of an unknown string S.

2 Grid Graphs

In this section, I define graphs on grids when S is **known** with the goal to prepare a ground to assemble the contigs from a d-edge-pair spectrum of an **unknown** string. Let $S = e_1 \dots e_n$ be a cyclic string where e_i is a condensed edge in the de Bruijn graph of S. Without loss of generality, assume that the (genomic) distance between the starts of e_1 and e_m is d (Otherwise, we can decompose e_m into two edges e_{m1} and e_{m2} such that distance between the starts of e_1 and e_{m2} is d). We defined the edge-pair spectrum of S as a result of the transformation from (k, d)-spectrum.

String S can be presented by a sequence of n **different** vertices $S = v_1, \dots, v_n$ where v_i is the start of edge e_i and also the end of edge e_{i+1}. Analogously, $S_d = v_m, \dots, v_n, \dots, v_{m-1}$. Let's define a *grid graph* G to be the cartesian product of two paths S and S_d: $G \triangleq S \square S_d$. The vertices of G are ordered pairs (u, v) where u and v are vertices in S and S_d correspondingly; the edges are $(u, v) \to (u', v)$ when $u \to u'$ in S, together with $(u, v) \to (u, v')$ when $v \to v'$ in S_d. Each edge of type $(u, v) \to (u', v)$ is labeled with the label of the edge $u \to u'$ in S and colored by red. Each edge of type $(u, v) \to (u, v')$ is labeled with the label of the edge $v \to v'$ in S_d and colored by blue.

In the grid G, any indirected cycle of length 4 is called a rectangle. Each rectangle corresponds to a pair of edges in S and S_d (Fig.2). For each node $x = (u_i, v_i)$ in G, define $d(x)$ as the genomic distance between u_i and the corresponding node of v_i in S. Given a non-negative integer value d, a rectangle is called a d-rectangle if and only if it corresponds to a d-edge pair (defined above as a result of the transformation from a (k, d)-mer). Using Bolzano-Cauchy theorem, it's trivial to see that a rectangle is a d-rectangle if there exists a pair of nodes x and y among 4 nodes of the rectangle such that $d(x) < d < d(y)$.

Theorem 1 *Given any vertex v in S, let R be a set of all d-rectangles that have $(v, *)$ as one of their vertices. There always exists an ordering of R such that adjacent elements (in this order) share a vertex $(v, *)$ in the grid graph G.* [1]

Given the grid graph G (as a cartesian product of two paths S and S_d), it is trivial to factorize G into its prime graphs S and S_d. In particular, S is obtained by contracting all blue edges into single nodes and considering parallel red edges as single edges. S_d can also be obtained in an analogous manner: contracting all red edges into single nodes and considering parallel blue edges as single edges (If it is not clear for you, refer to the final subsection of the Appendix). It is also clear that using the above approach, we don't need the whole grid G to find S or S_d. A subgraph of G is called **red-consistent** if contracting all blue edges in this graph results in a single path S . Similarly, we call it **blue-consistent** if S_d can be obtained by contracting all of its red edges. We now characterize the property of the red-consistent and blue consistent graphs.

Fact 1: Let $G' = (V', E')$ be a vertex-induced subgraph of G. G' is red-consistent if for every vertex u in S, the set $C_u = \{(u_i, v_j) \in V' \mid u_i = u\}$ is nonempty and forms a weakly connected component by using only blue edges in G'.

Fact 2: Let $G' = (V', E')$ be a vertex-induced subgraph of G. G' is blue-consistent if for every vertex v in S_d, the set $\{(u_i, v_j) \in V' \mid v_j = v\}$ is nonempty and forms a weakly connected component by using only red edges in G'.

Theorem 2 *Let $G' = (V', E')$ be a vertex-induced subgraph of G. If V' is a vertices set of all d-rectangles in G, G' is both red and blue-consistent.*

Proof 1 *Let $G' = (V', E')$ be a vertex-induced subgraph of G where V' is a vertices set of all d-rectangles in G. Let's first prove that G' is red-consistent. It's clear that for every vertex u in S, the set $C_u = \{(u, v_i) \in V'\}$ is non-empty. It remains to prove that this set forms a weakly connected component using the blue edges. This is an immediate result from theorem 1. We call G': the **d-grid graph**.*

[1] The order of R is the order that the diagonal lines intersect with these rectangles (in the language of the previous document). The proof for this theorem using the framework I define here is not short and I omit it in this write-up.

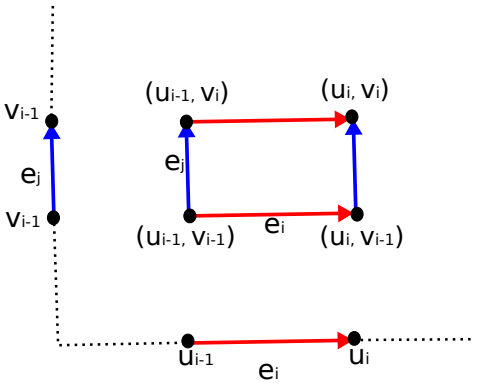


Figure 2: A rectangle in the grid graph G is a cartesian product of two edges

2.1 Quotient Graph

Let $G(V, E)$ be a graph with vertices set V and edges set E. $\bar{G}(\bar{V}, \bar{E})$ is a quotient graph of G iff $\bar{G}$ can be obtained from G by gluing some vertices in V . A gluing rule R on the graph G can be given as a set $R = (R_1, R_2, \ldots, R_t)$ where $R_1 \cup R_2 \cup \ldots \cup R_t = V$. The quotient graph of G is obtained by gluing all nodes in each R_i into a single vertex (for a detailed description of the gluing operation, see Pevzner 2004 or Munkres 2000).

The gluing operations can be operated in any orders. In particular, an order of R is a partition of each $R_i \in R$ into non-intersecting subsets:

$\{\{R_1^1, R_1^2, \ldots, R_1^{i_1}\}, \ldots, \{R_t^1, R_t^2, \ldots, R_t^{i_t}\}\}$ where $R_j = R_j^1 \cup R_j^2 \cup \ldots \cup R_j^{i_j}$.

For each R_i, we first glue all nodes in each of its subset R_i^j into a single vertex r_{ij}. As a result of this step, graph G is transformed into $\tilde{G}$ where all nodes in R_i^j are glued in to a single node r_{ij}. Finally, all r_{ij} is glued into a single node r_i. The final graph is the same, regardless of different choices of gluing orders. We note that the final graph is also a quotient graph of the intermediate graph $\tilde{G}$ under a different gluing rule.

Note that the contracting operation described above can also be described as a gluing operation. Contracting edge (u, v) in to a single node is equivalent to gluing u, v and ignoring the loop edge.

Theorem 3 *Let $G'(V, E)$ be a red-consistent subgraph of the grid graph G. $\bar{G}'$ be a quotient graph of G'. Contracting all blue edges in $\bar{G}'$ results in a quotient graph of S.*

Proof 2 *As previously noticed, the contracting operation of the blue edges is equivalent to the gluing operation and the final quotient graph does not depend on the gluing order. Therefore, we can firstly apply the contracting operation first and obtain an intermediate graph as S, the rest of the gluing operations will be apply on S and therefore, the final graph is a quotient graph of S.*

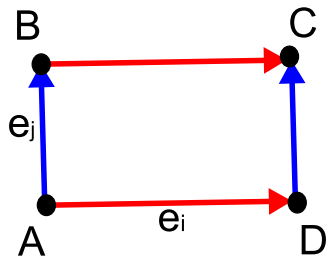


Figure 3: Rectangle.

2.2 From separated rectangles to the quotient graph

Let RT be a set of all *rectangles* in the d-grid graph. It's clear that this set can be constructed from the d-edge pair spectrum of the **unknown** string S. For each d-edge pair, (e_i, e_j, d_{ij}), construct a rectangle $ABCD$ as shown in Fig. 3. We color edges e_i with red and edges e_j with blue color and label them by $lb(e_i)$ and $lb(e_j)$ correspondingly. The vertices of the rectangle is labeled by a triple:

$$\begin{aligned} A &: (S(e_i), S(e_j), d_{ij}) \\ B &: (S(e_i), E(e_j), d_{ij} + l(e_j)) \\ C &: (E(e_i), E(e_j), d_{ij} + l(e_i) - l(e_j)) \\ D &: (S(e_i), E(e_j), d_{ij} - l(e_i)) \end{aligned} \tag{1}$$

where $S(e_i)$, $E(e_i)$ is the start and end vertex of edge e_i in the **condensed de Bruijn graph** correspondingly. $l(e_i)$ is the length of edge e_i. We note that there is a relation between the vertices in this rectangle and the nodes in the d-grid graph. The distance function on a node x of the d-grid graph $d(x)$ corresponds to the third component of the vertex in this representation.

Given the set of rectangles and the de Bruijn graph of sequence S (constructed from k-mers spectrum), we construct the following graph by gluing to the vertices of the rectangles if they have the same labels (While it's not the best gluing rule in practice, we choose to present it here for simplicity - See the Appendix section for a more advanced gluing rule).

Theorem 4 *Given the set of rectangles RT, the graph Γ results from gluing vertices with the same labels is a quotient graph of the d-grid graph.*

Proof 3 *It's clear that if two rectangles share the same node in the d-grid graph, they will have the same label. However, not all vertices with the same labels correspond to the sharing nodes of the rectangles in the d-grid graph. For a given label lb_i of the vertex, let N_i be a set of all vertices with the same label lb_i. There exists a partition of N_i into non-intersecting subsets $N_i = \{N_i^1, N_i^2, \ldots\}$ such that all vertices in each N_i^j correspond to the*

*same node (sharing nodes) in the **d-grid graph**. In another word, there exist an order of gluing such that the intermediate graph obtained by merging each N_i^j into a single node is the d-grid graph. Therefore, the final graph is the quotient graph of the d-grid graph.*

Since Γ is the quotient graph of the diag graph, contracting blue/red edges of Γ results in the quotient graph of S/S_d and therefore is useful in assembling the original sequence.

3 From rectangles to hyper-rectangles

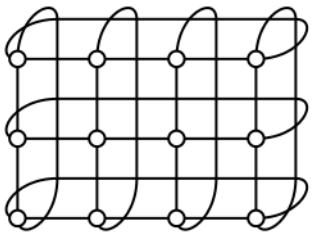
Given a non-negative integer value k, a vector of positive integer elements $\mathbf{d} = (d_1, \ldots, d_{t-1}) \in R^{t-1}$ and a cyclic string S, define a $(k, \mathbf{d})$-strobe of S as $(S_k(i), S_k(i+d_1), S_k(i+d_2), \ldots, S_k(i+d_{t-1}))$ where $i \in [1, \ldots, n]$. The $(k, \mathbf{d})$-strobe spectrum is a set of all possible $(k, \mathbf{d})$-strobes of S.

Problem: *Assemble S from its $(k, \mathbf{d})$-strobe spectrum.*

Consider the $(k, \mathbf{d})$-strobe spectrum of S and let B be a condensed de Bruijn graph constructed from all k-mers in the spectrum. For each edge e_i in the condensed de Bruijn graph, $l(e_i)$ is the length of edge e_i. Every $(k, \mathbf{d})$-strobe $(a_1, \ldots, a_t)$ maps to edges $e_1, \ldots, e_t$ at positions $p_1, \ldots p_t$ correspondingly (note that multiple k-mers in the same strobe can map to a single edge). For each $(k, \mathbf{d})$-strobe $(a_1, a_2, \ldots, a_t)$, we form an **d**-edge strobe, represented by a $2t-1$-tuple $(e_1, \ldots, e_t, d_1 + p_1 - p_2, \ldots, d_{t-1} + p_1 - p_t)$ where $d_1 + p_1 - p_j$ represents the distance between the starts of e_1 and e_j. As a result, the $(k, \mathbf{d})$-spectrum is transformed into a set of **d**-edge strobes. The set of all **d**-edge strobes generated from the (k, d)-spectrum of a string S is called an **d**-edge-strobes spectrum of S.

Similar to the previous section, instead of considering each (k, d)-strobe separately, we assemble S from a set of d-edge strobes. In difference from the previous section where we considered the cartesian product of $S \square S_d$, it is actually simpler to consider the cartesian product of S and itself: $S \square S$. The only reason we preferred to use two strings S and S_d in the previous section was to keep the **grid** form instead of using the *torus* that may confuse readers(Fig. 4). All theorems in the previous section can be easily generalized to the **d**-edge strobes data. We just present the framework and reintroduce the theorems in this section without proofs to prepare a ground for the next section: Multiple Libraries.

Assume that S is known in advanced, therefore, it can be presented as $S = e_1, \ldots, e_n$ where e_i is the condensed edge of the de Bruijn graph. Alternately, S can be also be presented as a sequence of n **different** vertices $S = v_1, \ldots, v_n$. Let $C = \{c_1, \ldots, c_t\}$ be a set of t different colors and consider the t-color graph: $G = S \square S \square \ldots S$ (t times) with vertices set are all possible t-tuples $\{(v_{i_1}, v_{i_2}, v_{i_t})\}$ together with edges: $(v_{i_1}, \ldots, v_{i_{k-1}} v_{i_k}, v_{i_{k+1}}, \ldots, v_{i_t}) \rightarrow (v_{i_1}, \ldots, v_{i_{k-1}} v_{i_k+1}, v_{i_{k+1}}, \ldots, v_{i_t})$ and with color $c_k, \forall k \in [1, t]$. Given a node $v = (v_{i_1}, v_{i_2}, \ldots, v_{i_t})$, define $\mathbf{d}(v)$ as: $\mathbf{d}(v) \triangleq (d_{i_1}, \ldots d_{i_{t-1}})$ where d_{i_j} is the genomic distance between v_{i_1} and v_{i_j}. The graph obtained is called a **hypertorus**. For any suitable value of $(i_1, i_2, \ldots, i_t)$, the vertex-induced subgraph of 2^t vertices $(\{v_{i_1}, v_{i_1+1}\}, \{v_{i_2}, v_{i_2+1}\}, \ldots, \{v_{i_t}, v_{i_t+1}\})$ forms



$C_3 \square C_4$

$(3 \times 4$ torus$)$

Figure 4: The cartesian product of two cycle is a torus (Figure reproduced from Knuth 2011).

a hyper-rectangle. For $t = 2$, it's simply a rectangle described as in the previous section. A hyper-rectangle corresponds to the dot product of t edges $e_{i_1} \square e_{i_2} \ldots \square e_{i_t}$. A hyper-rectangle is called a **d**-hyper-rectangle iff it corresponds to a **d**-edge strobe.

Theorem 5 *A hyper-rectangle is a* **d***-hyper-rectangle iff there exist a pair of nodes x, y among 2^t of its nodes, such that $\mathbf{d}(x) \preceq d \preceq \mathbf{d}(y)$.*

It's clear that given the hypertorus, the sequence S can be easily obtained by choosing an arbitrary set of $t-1$ colors and contracting all edges that have colors belong to this set (If it is not clear for you, please see section Appendix). We don't need the whole hyper-torus to infer S. Let G' be a subgraph of G and a color c_i. We call G': c_i-consistent if by contracting edges of all colors but c_i results in a cyclic sequence S. G' is called **all-color-consistent** if it is c_i-consistent for all $1 \leq i \leq t$.

Let V' be a set of vertices of all **d**-hyper-rectangles in the hyper-torus, $G'(V', E')$ be an vertex-induced subgraph of the hyper-torus G.

Theorem 6 *$G'(V', E')$ is all-color-consistent.*

Theorem 7 *If $\bar{G}'$ is a quotient graph of the graph G' and let C_i be any arbitrary set of $t-1$ colors. The graph obtained by contracting all edges having colors in C_i is the quotient graph of the string S.*

When sequence S is **unknown**, from the **d**-strobe spectrum and the de Bruijn graph of S (constructed from k-mers, we can construct the set of all hyper-rectangles. Namely, for each **d**-edge strobe $(e_1, \ldots, e_t, d*_1, \ldots, d^*_{t-1})$ where d^*_j represents the genomic distance between the starts of e_1 and e_j, we construct a hyper-rectangle by getting the cartesian

product of these t edges. Note that since S is unknown, each edge e_i is presented by two end vertices in the *de Bruijn graph* of S. A node of the rectangle is a $(2t-1)$-tuple: $(v_{i_1}, \ldots, v_{i_t}, d'_1, \ldots, d'_{t-1})$ where d'_j is the genomic distance between v_{i_1} and v_{i_j}. It's trivial to prove that the graph obtained by gluing all vertices of the hyper-rectangles having the same labels is the quotient graph of the graph G' and therefore, can be used to assemble S.

4 Multiple Libraries

I choose to present this section in a separate document when it's already polished!

5 Appendix

This section is merely a discussion and some small details that I feel that it's benificial to write it down.

5.1 Beyond Rectangles

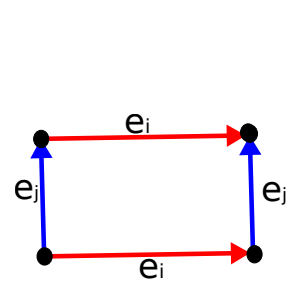
Let consider a non-square rectangle in Fig. 5(a), since e_i is longer than e_j, we can further extend e_j by continuing to map the (k,d)-mers until the end of e_i is reached. As a result, edge e_j can be further extended by e_t (Fig. 5(b)). We call it the extended part of the rectangle. We can use this extended part to restrict the gluing. For instance, vertices A and B in Fig. 5(c) may have the same label, but should not be glued, since e_t cannot be **embedded** into the rectangle on the right (The term *embedded* should be defined rigorously). We note that the extended parts just play the role of restricting the gluing. The contigs are spelled only by edges in the main rectangles.

5.2 Variation of the insert distances is benificial

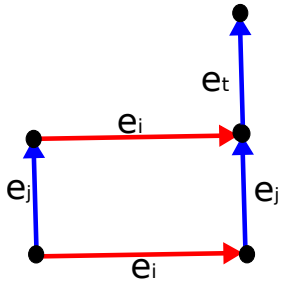
While the variation in the insert size was the source of difficulties in the paired de Bruijn graph. In the grid graph, the distance of between pair of edges in the de Bruijn graph can be estimated more exactly. From another side, the variation of the insert size can help to increase the extended parts of the rectangle, even for the square case. When the extended parts become larger, it's easier to restrict the gluing and therefore, improve the assembly.

5.3 How a hyper-rectangle is created from a singe edge

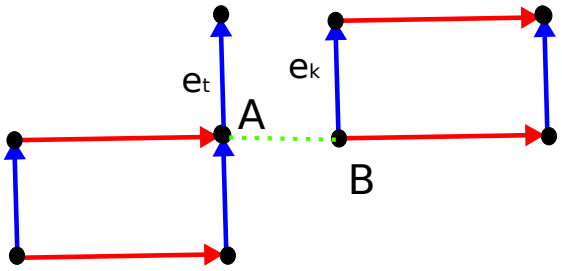
This part will be removed



(a) A non-square rectangle



(b) Non-square rectangle can be extended



(c) Although A and B may have the same label, they can not be glued since edge e_t can not be embedded in to the rectangle on the right

Figure 5: Beyond rectangles

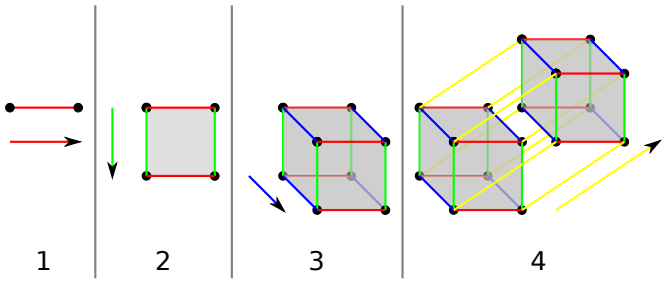


Figure 6: hyper-rectangle. Figure reproduced from wiki

Given a set of directed edges $\{e_1, e_2, e_3, \ldots, e_n\}$, one can create an hyper-rectangle in the following way: Starting from e_1 (or any arbitrary edge in the set), move it in the direction of e_2. It will sweep out a rectangle (Fig. 6). Next, move this rectangle in the direction of e_3, it will sweep out a box. Continue the procedures until we obtain a hyper-rectangle that have all the edges in the set. Therefore, the original edge e_1 can be obtained by contracting all other edges $e_2, \ldots, e_n$.