DNA Dizilerinin De Bruijn Grafları ile İncelenmesi

Analysis of DNA Sequences with De Bruijn Graphs

İrfan KILIÇ[1], Doç.Dr.Ali KARCI[2]

[1]Bilgisayar Mühendisliği
İnönü Üniversitesi
ikilic23@hotmail.com

[2]Bilgisayar Mühendisliği
İnönü Üniversitesi
ali.karci@inonu.edu.tr

Özet

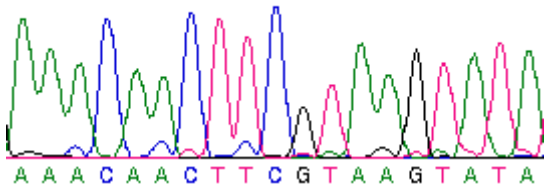
Genel olarak DNA verisi olarak adlandırılan genetik veriler biyoinformatiğin en temel konularından birini teşkil etmektedir. Bu verilerin gösterimi yapılırken çeşitli dizilime yöntemleri kullanılmaktadır. Bu bildiride bu dizileme yöntemlerine değinilecek ve De Bruijn grafları kullanılarak basit DNA verileri üzerinde baştan kısa okuma dönüşümünün (de novo short read assembly) nasıl yapıldığı gösterilecektir.

Abstract

The purpose of this study is to analysis the genetics data, in general named as DNA, which is the very main issue of bioinformatics. It is aimed to use the methods of various sequencing to demonstrate the aforementioned data. In this study, it will be mentioned about the methods of sequencing and will be used "De Bruijn Graphs" to show how to read the de novo short read assembly on the DNA data..

1. Giriş

DNA dizisi veya **genetik dizi**, gerçek veya hayalî bir DNA molekülü veya ipliğinin birincil yapısına karşılık gelen harfler dizisidir.



Şekil 1: Bir otomatik dizileme aracının çıktısından okunabilen DNA dizisi

Bu dizide bulunan harfler *A*, *C*, *G*, ve *T* 'dir, bunlar DNA ipliğinde bulunan adenin, sitozin, guanin, ve timin adlı dört bazı temsil eder. Tipik olarak bu diziyi oluşturan harfler birbirine bitişik olarak, aralarda boşluk olmaksızın yazılır, örneğin AAAGTCTGAC gibi; bu dizinin soldan sağa okunuşu 5'-3' doğrultusuna karşılık gelir.
Fonksiyona göre bir DNA dizisine *anlamlı* veya *anti-anlamlı* ve *kodlayan* veya *kodlamayan* olarak değinilebilir.

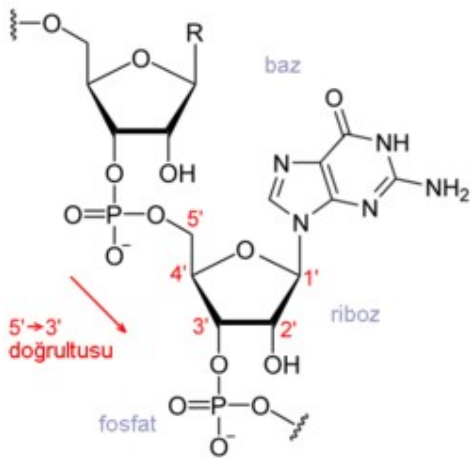
Bir DNA molekülünün baz dizisinin okunmasına DNA dizilemesi denir.

2. Dizi formatları

DNA dizilerinin biyoenformatik programları tarafından okunması için belli standart formatlar oluşmuştur. Örneğin bunların en yaygını olan FASTA formatında birinci satır bir ">" sembolünü takibeden bir başlık içerir, onu izleyen satırlarda ise DNA dizisi yer alır. Örneğin:

>gi|311883|emb|X72213.1| H.sapiens genomic DNA with integrated fragment of HBV DNA
AATTCTGTGCGTCTCTGATGCTTTGTCCTCCAGCCCCC
ACCCCCAGAAATCTCCCCCTNNNGCAAAGA……
……[1].

2.1. Doğrultu



Şekil 2: Bir RNA zincirinde 5' -> 3' doğrultusu

Moleküler biyolojide **doğrultu**, bir nükleik asit ipliğini oluşturan nükleotidlerin uç uca eklenme yönüyle ilişkilidir. Kimyasal adlandırma konvansiyonu gereği, bir nükleotid şeker halkasındaki karbon atomları 1', 2', 3', 4' ve 5' olarak adlandırılır ("1 üssü" vb. olarak okunur). Nükleik asitlerin doğada sentezlenmeleri sırasında büyüyen zincirin bir ucundaki şeker grubunun serbest bir 3' hidroksil (-OH) grubu vardır, öbür ucundaki şekerin ise serbest bir 5'-OH grubu vardır. Bu iki uca, sırasıyla 3' ve 5' uçları denir. Nükleik asidin sentezi sırasında polimeraz enzimi 3'-OH grubuna bir fosfodiester bağı ile yeni bir nükleotid bağlar. Konvansiyon olarak bir iplikli DNA ve RNA dizileri yazılırken bazların kısaltmaları 5'-3' doğrultusunda yazılır[2].

3. DNA dizileme

DNA dizilemesi, bir DNA molekülündeki nükleotid bazlarının (adenin, guanin, sitozin ve timin) sırasının belirlenmesidir.
DNA dizilerinin bilinmesi temel biyoloji, biyoteknoloji, adli bilim, tıbbi tanı koyma gibi pek çok sahada vazgeçilmez hâle gelmiştir. DNA dizilemesi biyolojik araştırma ve keşifleri çok hızlandırmıştır. Modern DNA dizileme teknolojilerin mümkün kıldığı hızlı DNA dizileme sayesinde İnsan Genom Projesi'nde insan genomunun dizilenebilmiştir. Benzer projelerle pekçok hayvan, bitki ve mikrop genomunun tam dizisi üretilebilmiştir.
İlk DNA dizileri 1970'lerin başlarında üniversite araştırmacıları tarafından iki-boyutlu kromatografiye dayanan zahmetli yöntemlerle elde edilmiştir. Otomatik analizle çalışan boya-tabanlı dizileme yöntemlerinin gelişimiyle DNA dizilemesi çok daha kolaylaşmış ve birkaç büyüklük mertebesi hızlanmıştır[3][4].
1970'lerin başlarında hızlı DNA dizileme yöntemlerinin geliştirilene kadar, DNA dizilemesi için çeşitli zahmetli yöntemler kullanılmaktaydı.[5][6] Örneğin, 1973'te Gilbert ve Maxam, dolaşan benek analiz (*wandering-spot analysis*)

olarak adlandırılan bir yöntemle 24 nükleotitin dizisini yayımladılar[7].
Sanger ve arkadaşlarının 1975'te geliştirdiği zincir sonlandırma yöntemi göreceli olarak daha kolay ve güvenilir olmasından dolayı kısa sürede en tercih edilen yöntem oldu[8][9].

3.1. Maxam - Gilbert dizilemesi

1976-1977'de Harvard Üniversitesi'nden Allan Maxam and Walter Gilbert, DNA'nın kimyasal modifikasyonu ve ardından onun spesifik bazlarda kesilmesi esasına dayanan bir DNA dizileme yöntemi geliştirdi.[5] Maxam ve Gilbert kimyasal dizileme yöntemi hakkındaki makale, Sanger ve Coulson'un artı-eksi dizilemesi hakkındaki makalesinden iki yıl daha sonra yayınlamasına rağmen, Maxam-Gilbert dizilemesi daha popüler oldu[8][10]. Bunun nedeni, Maxam-Gilbert yönteminde saflaştırılmış DNA'nın doğrudan dizilenebilmesi, buna karşın ilk Sanger yönteminde tek iplikli DNA üretilebilmesi için okunacak her DNA'nın ayrıca klonlanmasının gerekmesiydi. Ancak, zincir sonladırma yönteminin zaman içinde iyileştirilmesiyle M axam-Gilbert dizilemesi gözden düştü, zira teknik karmaşıklığı onun standart moleküler biyoloji kitlerinde kullanılmasına olanak vermiyordu, ayrıca zararlı kimyasallara gerek gösteriyordu ve ölçeklenmesinde zordu[11].

3.2. Zincir sonlandırma yöntemleri

Zincir sonlandırma yöntemi (veya Sanger yöntemi, onu geliştiren Frederick Sanger'e atfen) Maxam ve Gilbert yöntemine kıyasla daha verimli olduğu, daha az toksik kimyasal ve radyoaktivite gerektirdiği için kısa sürede hızla yaygınlaştı. Sanger yönteminin ana ilkesi zincir sonladırıcı olarak dideoksinükleotit trifosfatlar (ddNTP'ler) kullanılmasıdır.
Zincir sonlandırması DNA dizilemesini son derece kolaylaştırmıştır. Örneğin, ticarî olarak elde edilebilen zincir sonladırma kitlerinde dizileme yapmak için gerekli olan reaktantlar, taksim edilmiş ve kullanıma hazır şekilde bulunur. Yöntemin sınırlı kalabileceği durumlar, (1) primerin DNA'ya non-spesifik bağlanıp DNA dizisinin doğru okunmamasına neden olması ve (2) DNA'daki ikincil yapıların dizinin sadakatine etki etmesidir.

3.3. Yüksek hacimli dizileme

Düşük masraflı dizilemeye olan büyük talep, yüksek hacimli dizileme teknolojilerinin geliştirilmesine yol açmıştır. Bu teknolojilerde dizileme süreci paralelleştirilerek binlerce veya milyonlarca dizi aynı anda üretilir[12][13]. Yüksek hacimli dizileme teknolojilerinin amacı, standart boya sonlandırma yöntemleri ile mümkün olan DNA dizileme masrafını azaltmaktır[14].

3.4. Dizi analizi

Biyolojide dizi analizi bir DNA veya protein dizisini bir bilgisayar kullanarak, biyoenformatik yöntemlerle incelenmesidir. Dizi analizi ana hatlarıyla aşağıdaki konuları kapsar:

1. İki veya daha çok dizinin karşılaştırılarak benzerliklerini bulunması (dizi hizalama)
2. Bir dizinin içinde belli özelliğe sahip nokta veya bölgelerin bulunması.
3. Örneğin, dizi motifleri, okuma çerçeveleri, intron ve ekson sınırları, ve gen düzenleyici elemanlar.
4. Genom haritalaması
5. Homolog dizilerin karşılaştırılarak filojenik ağaç kurmak.
6. Protein yapı tahmini

Kimyada dizi analizi, bir polimeri oluşturan monomerlerin dizisini çözmek anlamında kullanılır. Moleküler biyoloji ve genetikte bu işleme sadece dizileme denir.

4. De Bruijn grafları

Graf teorisinde, m sembollü, n boyutlu bir De Bruijn grafı sembol dizileri arasında örtüşmeleri gösteren yönlü bir graftır. Bu grafın verilen sembollerin olası tüm n uzunluğundaki dizisinden oluşan m^n adet düğümü vardır. Aynı sembol bir dizide birden çok görülebilir. $S:=\{s_1,s_2,\ldots,s_m\}$ şeklinde sembol kümemiz olsun, buna göre düğümlerin kümesi;
$V=S^n=\{(s_1,\ldots,s_1, s_1), (s_1,\ldots,s_1, s_2),\ldots, (s_1,\ldots,s_1, s_m), (s_1,\ldots,s_2, s_1),\ldots, (s_m,\ldots,s_m, s_m)\}$ şeklinde olur.

Düğümlerden biri, tek bir yerde tüm sembollerin sola kayması ve bu düğümün sonuna yeni bir sembol ekleyerek başka bir düğüm olarak ifade edilebilir. Bundan sonra elde edilen düğümün önceki düğümden yönlü bir kenarı vardır. Böylece yönlü kenarların kümesi;

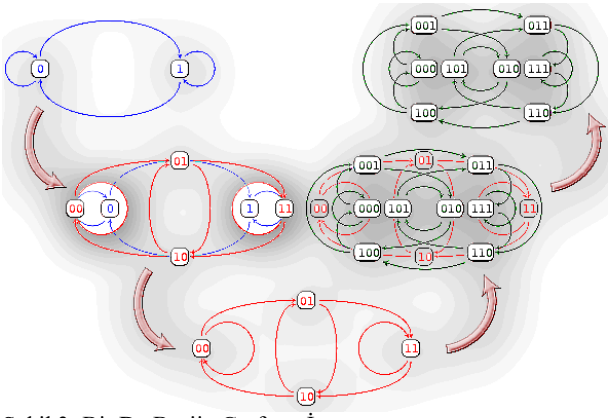
$E=\{((v_1,v_2,\ldots,v_n),(w_1,w_2,\ldots,w_n)):v_2=w_1, \quad v_3=w_2,\ldots, \quad v_n=w_{n-1}\}$ şeklinde olur.

De Bruijn grafları Nicolaas Govert De Bruijn'den ismini almasına rağmen, De Bruijn ve I. J. Good her ikisi tarafından bağımsız olarak geliştirilmiştir[15][16]. Çok daha öncesinde, Flye Sainte- Marie tarafından bu grafların özellikleri dolaylı olarak kullanılmıştır[17].

Özellikler

- N=1 ise herhangi iki düğüm arasında boş kenar olmaması şartıyla tüm düğümler toplam m^2 kenar ile bağlıdır.
- Her düğümün m adet giriş ve çıkış kenarı vardır.[18]
- Her n boyutlu De Bruijn grafı aynı sembollere sahip n-1 boyutunda yönlü line(yol) grafıdır.
- Her De Bruijn grafı Euler ve Hamilton grafıdır. Bu grafın Euler ve Hamilton çevrimleri (line grafın inşası yoluyla bir diğerinin eşdeğeri) De Bruijn dizisidir.

N=3 boyutlu ikili De Bruijn grafının çizgi graf inşası Şekil 3.' de gösterilmiştir. Örnekte görebildiğimiz gibi, n-1 boyutunda De Bruijn grafının bir kenarı n boyutlu De Bruijn grafının bir düğümüne karşılık gelir. n-1 boyutlu De Bruijn grafında iki kenarlı bir yola, n boyutlu De Bruijn grafında bir kenara karşılık gelir.



Şekil 3: Bir De Bruijn Grafının İnşası

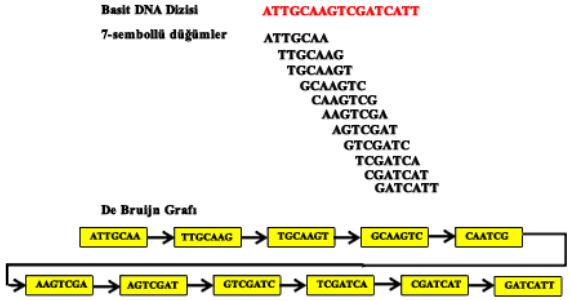
4.1. Kullanım alanları

- Bazı grid ağ topolojilerinde
- Dağıtık hash tablolarındaki Koorde protokolü, De Bruijn graflarını kullanır.
- Biyoinformatikte bir genomun okunan dizilerinin de novo yerleşimi için de Bruijn grafları kullanılır[19][20][21].

5. Basit DNA verisinin De Bruijn grafları ile oluşturulması[22]

Kısa okuma dönüştürücü(short read assembly) için yeni algoritmalarda çoğunlukla gen dizisi verisini gösteren ve saklayan De Bruijn grafları kullanılır. De Bruijn grafları nedir ve niçin kısa okuma dizileri için bu kadar popülerdir. Burada bunu açıklayacağız.

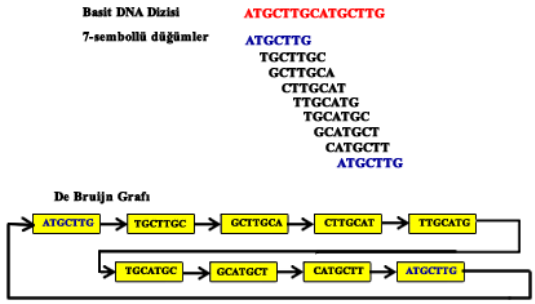
De Bruijn grafı k-sembol bileşenlerle bir diziyi göstermenin etkili bir yoludur. De Bruijn grafları geniş ölçekte problemler için kullanılmasına rağmen tartışmamızı nükleotid dizileriyle sınırlı tutacağız. Çoğunlukla makaleler De Bruijn graflarından gen dizisini elde etme ve kısa okumalardan (short read) De Bruijn graflarını oluşturma üzerinedir. Burada ilk önce bir gen dizisinin De Bruijn grafı ile başlayalım sonrasında bu graftan kısa okumalar nasıl yapılır onu açıklayalım.
Bir De Bruijn grafı, uzun veya kısa herhangi bir gen dizisi için oluşturulabilir. Şekil 4'te basit örnek verirsek



Şekil 4: K=7 için De Bruijn Grafı

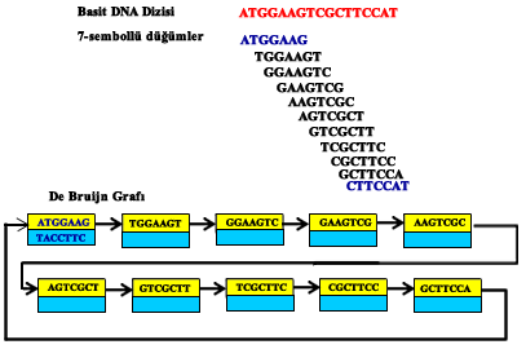
Şekil 4'teki örnekte ATGGAAGTCGCGGAATC gen dizisi, içinde örtüşme olan (örnekte K=7) K- sembole bölünmüş yönlü bir grafın düğümleri her biri 7-sembolle oluşturulmuştur. Kenarlar orijinal dizi üzerinde 7-sembollü komşular (düğümler) arasında çizilmiştir. Bu metot bağlı düğümlerdeki 6 (=K-1) nükleotidin örtüşmesini sağlar.

Orijinal gen dizisinde tekrar eden 7-sembollü (düğüm) olmadığından yukarıdaki örnek basittir. Şekil 5'teki örnekte bazı tekrar eden düğümler var. Bu örnekte en yüksek 5' açılı 7-sembol (düğüm), en yüksek 3' açılı 7-sembol (düğüm)'de görülmüştür (her ikisi de mavi olarak gösterilmiştir).



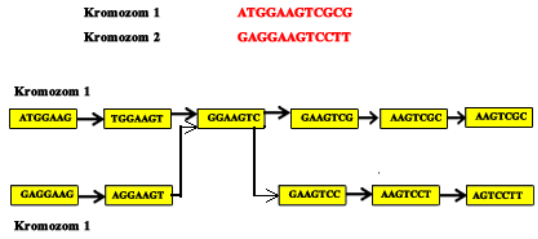
Şekil 5: K=7 için De Bruijn Grafında Döngü

Bu durumda De Bruijn grafında bir döngü oluşur. Şekil 5'teki örnekte gösterilen düğümler gen dizisinin her iki sarmalı için gösterilmemesine rağmen, gerçekte her düğüm Şekil 6'da gösterildiği gibi çift sarmallıdır.



Şekil6: K=7 için çift sarmallı De Bruijn grafı

Şekil 6'daki örnekte en büyük 3' açılı 7- sembollü (düğüm) en büyük 5' açılı 7- sembollünün (düğümün) tamamlayıcısının tersidir. Çift sarmallı De Bruijn graflarının bağlantıları buna göre oluşturulur.

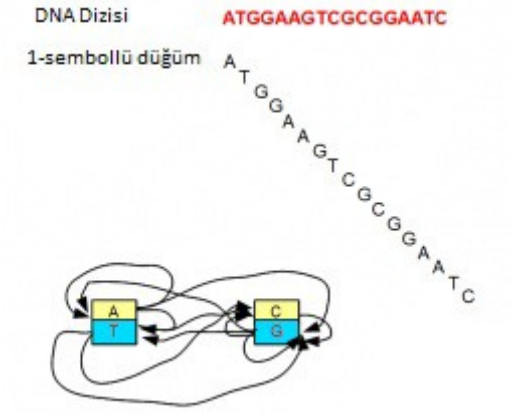


Şekil 7: K=7 için 2 ayrı kromozomlu De Bruijn grafı

Şekil 7'de görülen adımlar herhangi boyutta büyük bir gen dizisinin De Bruijn graflarıyla oluşturulması tekrar edilebilir. Bir gen dizisi 2 ayrı kromozoma sahip olsada De Bruijn grafları bu kromozomların Şekil 7'de görüldüğü gibi K-sembollüleri örtüşürse ayrı kalmayabilir.

Yukarıdaki örneklerin hepsinde K=7 alındı, fakat K küçük veya büyük bir tamsayı olabilir. K oldukça küçük 1 olabilir.

Ancak Şekil 8'den görebileceğiniz gibi K=1 için De Bruijn grafları çok kullanışlı değildir.



Şekil 8: K=1 için De Bruijn grafı

6. Sonuç

Amacımız için yukarıda De Bruijn graflarını açıkladık. Buradan hareketle bu graflar ile ilgili şu sonuçlara varabiliriz.

1. Verilen herhangi bir gen dizisi ve K- sembollü ile basit yapıda bir De Bruijn grafını oluşturabiliriz.
2. Daha büyük K-sembollüler için tek bir gen dizisini de Bruijn graflarına dönüştürmek daha kolaydır.
3. Çok yüksek boyutlu K-sembollüler için genellikle, grafı saklamak ve işlem yapmak için daha fazla bilgisayar belleğine gereksinim duyulur. Ayrıca K değeri için üst sınırı ayarlanabilen ne kadar belleğe sahip bir bilgisayara ihtiyaç duyulması bir tartışma konusudur.

Örneklerde gösterildiği gibi herhangi bir gen dizisini De Bruijn graflarına kolaylıkla dönüştürebiliriz.

7. Kaynaklar

[1] "FASTA format description". 20 Mayıs 2012 tarihinde erişilmiştir.

[2] Lodish et al., *Molecular Cell Biology*, 5th edn., 2004, W.H. Freeman and Company, New York. ISBN 0-7167-4366-3.

[3] Olsvik O, Wahlberg J, Petterson B, *et al.* (January 1993). "Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains". *J. Clin. Microbiol.* 31 (1): 22–5.

[4] Pettersson E, Lundeberg J, Ahmadian A (February 2009). "Generations of sequencing technologies". *Genomics* 93 (2): 105–11.

[5] Maxam AM, Gilbert W (February 1977). "A new method for sequencing DNA". *Proc. Natl. Acad. Sci. U.S.A.* 74 (2): 560–4.

[6] Gilbert, W. DNA sequencing and gene structure. Nobel lecture ,8 December 1980.

[7] Gilbert W, Maxam A (December 1973). "The nucleotide sequence of the lac operator". Proc. Natl. Acad. Sci. U.S.A. 70 (12): 3581–4.

[8] Sanger F, Coulson AR (May 1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". J. Mol. Biol. 94 (3): 441–8.

[9] Sanger F, Nicklen S, Coulson AR (December 1977). "DNA sequencing with chain-terminating inhibitors". *Proc. Natl. Acad. Sci. U.S.A.* 74 (12): 5463–7

[10] Sanger F. Determination of nucleotide sequences in DNA. Nobel lecture, 8 December 1980.

[11] Graziano Pesole; Cecilia Saccone (2003). *Handbook of comparative genomics: principles and m ethodology*. New York: Wiley-Liss. pp. 133. ISBN 0-471-39128-X.

[12] Hall N (May 2007). "Advanced sequencing technologies and their wider impact in microbiology". *J. Exp. Biol.* 210 (Pt 9): 1518–25.

[13] Church GM (January 2006). "Genomes for all". *Sci. Am.* 294 (1): 46–54.

[14] Schuster, Stephan C. (2008). "Next-generation sequencing transforms today's biology". *Nature methods* (Nature Methods) 5 (1): 16–18.

[15] de Bruijn, N. G. (1946). "A Combinatorial Problem". Koninklijke Nederlandse Akademie v. Wetenschappen 49: 758–764.

[16] Good, I. J. (1946). "Normal recurring decimals". *Journal of the London Mathematical Society* 21 (3): 167–169.

[17] Flye Sainte-Marie, C. (1894). "Question 48". *L'Intermédiaire Math.* 1: 107–110.

[18] Zhang, Fu Ji; Lin, Guo Ning (1987). "On the de Bruijn-Good graphs". *Acta Math. Sinica* 30 (2): 195–205.

[19] Pevzner, Pavel A.; Tang, Haixu; Waterman, Michael S. (2001). "An Eulerian path approach to DNA fragment assembly". *PNAS* 98 (17): 9748–9753.

[20] Pevzner, Pavel A.; Tang, Haixu (2001). "Fragment Assembly with Double-Barreled Data". *Bioinformatics/ ISMB* 1: 1–9.

[21] Zerbino, Daniel R.; Birney, Ewan (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". *Genome Research* 18 (5): 821–829.

[22] http://www.homolog.us/blogs/2011/07/28/de-bruijn-graphs-i