TÜBİTAK

RNA İkincil Yapılarının Çok Boyutlu Gösterimi ve Pre-Mirna Tespiti İçin Uygulamaları

Program Kodu:1002

Proje No:120E042

Proje Yürütücüsü:

Dr. Öğr. Üyesi MÜŞERREF DUYGU SAÇAR DEMİRCİ

Araştırmacı(lar)

Dr. Öğr. Üyesi YILMAZ MEHMET DEMİRCİ

Bursiyer(ler)

FURKAN BÜYÜKGÖL

MAYIS 2021

ANKARA

Önsöz

“RNA İkincil Yapılarının Çok Boyutlu Gösterimi ve Pre-Mirna Tespiti Için Uygulamaları” başlıklı TÜBİTAK destekli (proje no: 120E042) proje kapsamında herhangi bir RNA sınıfında (rRNA, mRNA, miRNA vb.) uygulanabilirliği olan RNA ikincil yapısı gösterim yöntemi geliştirilmiştir. Biyoenformatik, matematik ve istatistik tekniklerinin kullanılmasıyla ortaya çıkarılan metot, miRNA moleküllerinin *in silico* tahmini için önemli bir kaynak oluşturacaktır.

İçindekiler

Tablo ve şekil listeleri

Özet

MikroRNA'lar (miRNA'lar), transkripsiyon sonrası gen ekspresyonu düzenleyicileridir. Bir miRNA yüzlerce haberci RNA'yı (mRNA'lar) hedefleyebildiği gibi, bir mRNA farklı miRNA'lar tarafından hedeflenebilir, üstelik tek bir miRNA bir mRNA sekansında çeşitli bağlanma bölgelerine sahip olabilir. Bu nedenle miRNA'ları deneysel olarak araştırmak oldukça karmaşıktır. Bu tür zorlukları aşabilmek için makine öğrenimi (ML) sıklıkla kullanılmaktadır. ML analizinin temel kısımları büyük ölçüde giriş verilerinin kalitesine ve verileri tanımlayan özelliklerin kapasitesine bağlıdır. Daha önce miRNA'lar için 1000'den fazla özellik önerilmişti. Bu projede, RNA ikincil yapısını temsil eden yeni özellikler ve yüksek doğruluk değerleri sağlayan, dinamik, çok boyutlu grafik gösterimini tanımlamayı hedeflemiştik. Bu çalışmada, ML tabanlı miRNA tahmini için yeni ve kolayca güncellenebilir bir yaklaşım geliştirilmiştir. Bilinen insan miRNA'larının ve sözde saç tokalarının random forest (RF), support vector machine (SVM) ve multilayer perceptron (MLP) gibi çeşitli sınıflandırıcılarla sınıflandırılmasıyla binlerce model oluşturulmuştur. Yöntem insan verilerine dayanarak oluşturulmuş olsa da en iyi model miRBase ve MirGeneDB gibi kamu veri tabanlarından insan olmayan saç tokaları üzerinde test edilmiş ve yüksek skorlar üretilmiştir. Ayrıca, yöntemin farklı veriler üzerindeki etkinliğini göstermek için ekspresyon farkları tahmini (differential expression prediction) analizinde de kullanılmıştır. Bu aşamada SARS-CoV-2 enfeksiyonunun etkisini ölçen bir veri setinin analizinden elde edilen sonuçlar yayınlanmıştır.

Anahtar kelimeler: miRNA, tahmin, makine öğrenmesi, model

Abstract

MicroRNAs (miRNAs) are posttranscriptional regulators of gene expression. While a miRNA can target hundreds of messenger RNA (mRNAs), an mRNA can be targeted by different miRNAs, not to mention that a single miRNA might have various binding sites in an mRNA sequence. Therefore, it is quite complicated to investigate miRNAs experimentally. Thus, machine learning (ML) is frequently used to overcome such challenges. The key parts of a ML analysis largely depend on the quality of input data and the capacity of the features describing the data. Previously, more than 1000 features were suggested for miRNAs. In this project, we aim to define new features representing the RNA secondary structure and its dynamic multidimensional graphical representation providing high accuracy values. In this study, a new and easily updateable approach for ML-based miRNA prediction has been developed. Thousands of models have been created by classifying known human miRNAs and pseudo hairpins with various classifiers such as random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP). Although the method was created based on human data, the best model was tested on non-human hairpins from public databases such as miRBase and MirGeneDB and high scores were produced. It has also been used in differential expression prediction analysis to show the effectiveness of the method on different data sets. At this stage, the results obtained from the analysis of a data set measuring the impact of SARS-CoV-2 infection have been published.

Keywords: miRNA, prediction, machine learning, model

1. Giriş

MikroRNA'lar (miRNA), hedef mRNA'ların translasyonel engellenmesi veya destabilizasyonunu kullanarak transkripsiyon sonrası düzenleme yoluyla gen ekspresyonunu kontrol eden yaklaşık 22 nükleotid (nt) uzunluğunda tek iplikli RNA'lardır (Piast vd., 2005; Filipowicz vd., 2008). MiRNA'ların ilk örneği, gelişimsel zamanlamanın bir düzenleyicisi *olarak C. elegans*'ta keşfedilmiştir (Lee vd., 1993). Virüslerden yüksek ökaryotlara kadar değişen çeşitli organizmalarda, önemli süreçler miRNA'ların etkisi altındadır. MiRNA'lar ile kanser ve nörodejeneratif hastalıklar gibi insan hastalıkları arasında birçok bağlantı gösterilmiştir. Ayrıca, miRNA'ların memelilerde tüm protein kodlayan genlerin yaklaşık %30'unun aktivitelerini kontrol ettiği tahmin edilmektedir (Filipowicz vd., 2008). MiRNA temelli gen düzenlemesi sadece yüksek ökaryotlarda değil, bazı basit çok hücreli organizmalarda da gözlemlenmiştir (Kim vd., 2009).

Kendi kendine katlanarak ikincil yapılar oluşturan tek sarmallı RNA yapısına ek olarak, miRNA'lar, miRNA biyogenez makine elemanları tarafından tanınabilmeleri ve değiştirilebilmeleri için karakteristik bir saç tokası yapısına sahiptir (Kozlowski vd., 2008). Bu nedenle miRNA tahmin analizleri genellikle birincil ve ikincil yapılardan bilgi gerektirir. Ne yazık ki, bu saç tokası yapısı sadece miRNA'lara özgü tamamıyla ayırt edici bir özellik değildir (Roden vd., 2017). Belirli bir dizinin miRNA olup olmadığını belirlemek için tasarlanan araçların çoğu makine öğrenimi (ML) uygulamasına dayanmaktadır (Saçar Demirci vd., 2017). Makine öğrenimi miRNA çalışmaları için oldukça güçlü ve avantajlı olsa da veri kalitesi, parametre seçimi ve makine öğrenmesi algoritması seçimi gibi bazı önemli noktaların verimli bir analiz için dikkate alınması gerekmektedir (Saçar Demirci ve Allmer, 2017).

Proje kapsamında, bilinen miRNA öncüllerinin saç tokası yapılarının 3B temsiline dayalı miRNA tahmini için bir ML yaklaşımı geliştirilmiştir. Yöntem insan miRNA verilerine dayalı olarak geliştirilmiş ve test edilmiş olsa da ancak diğer organizmalar için de uygulamak ve/ veya genişletmek mümkündür.

2. Literatür özeti

Ribonükleik asit (RNA) birçok hücresel süreçte önemli bir oyuncudur ve bazı viral organizmalar için genetik bilginin kaynağıdır. Sadece RNA moleküllerinin dizileri değil, yapıları da büyük önem taşımaktadır. Üç ana RNA yapısı seviyesi vardır: birincil (baz sekansı), ikincil (baz çiftlerine dayalı, örn., Saç tokaları veya transfer RNA'nın (tRNA) yonca yaprağı yapısı) ve üçüncül (ikincil yapı elemanları arasındaki etkileşimler). RNA sekonder yapısı, A-U ve G – C baz çiftleri arasındaki hidrojen bağları ile oluşturulur (G – U eşleşmesi de sıklıkla gözlenir) (Varani ve McClain, 2000). Ancak, bu bazlar ve eşleşmeler aynı güce sahip değildir. Dört baz birkaç özelliklerine göre farklı sınıflara ayrılabilirler. Örneğin;

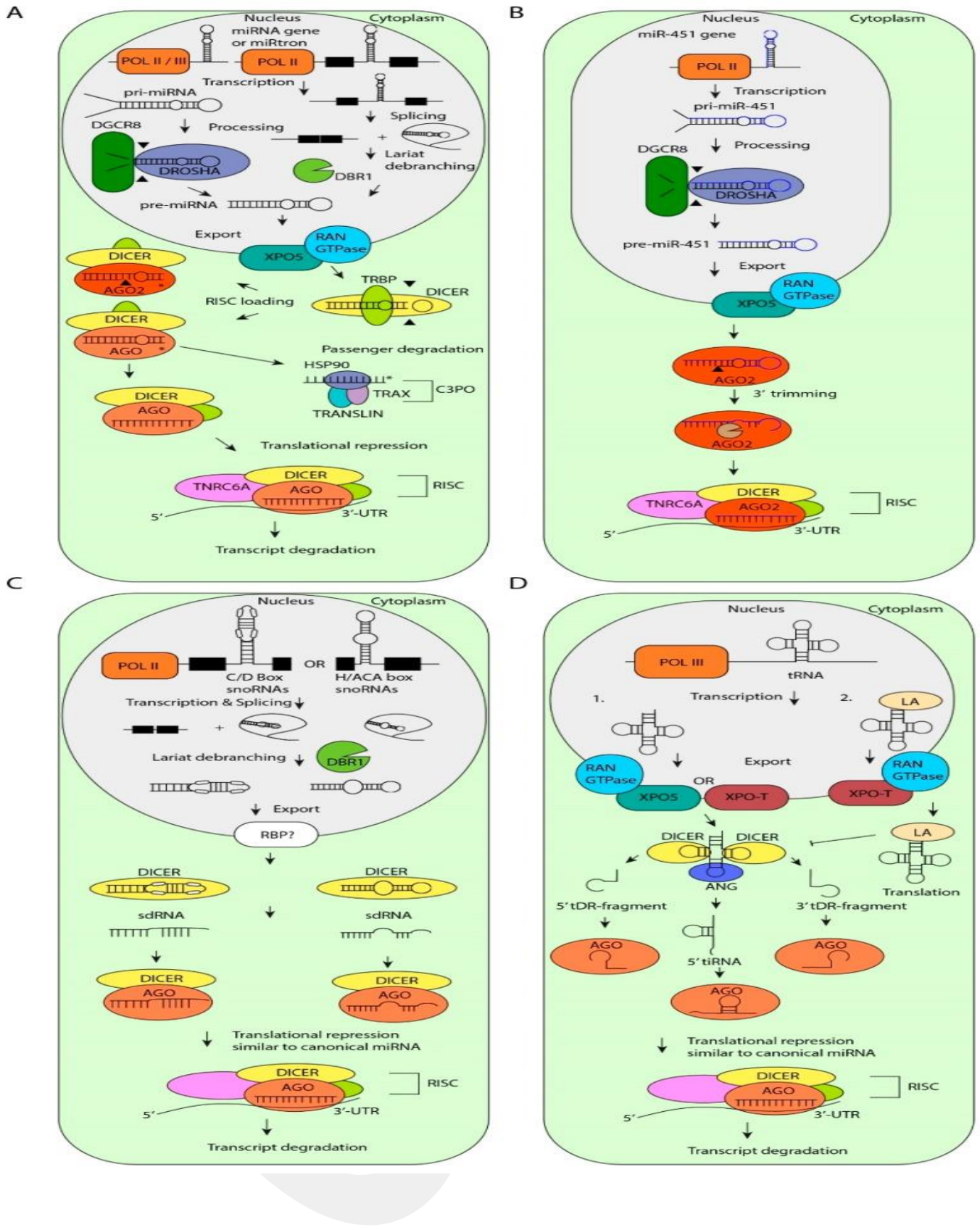
- Hidrojen bağının gücüne göre
    - zayıf H-bağları (A, U)
    - güçlü H-bağları (G, C)
- Amino grubuna göre (A, C)
- Keto grubuna göre (G, U)
- Kimyasal yapılarına göre
    - pürin (A, G)
    - pirimidin (C, U)

Bazların özellikleri ve eşleştirme bilgilerinin kullanılmasıyla, RNA benzerliğini ölçmeyi amaçlayan çeşitli yöntemler önerilmiştir. Bu yaklaşımlardan bazıları, bilgi kaybından muzdarip olabilecek RNA 2B yapısının grafiksel temsiline dayanmaktadır (Zhang vd., 2016). Öte yandan, RNA ikincil yapılarının 3B grafik temsili için geliştirilen yöntemler, baz dizisi, kimyasal ve yapısal bilgileri kullanır.

Son yıllarda, transkripsiyon sonrası gen ekspresyonunu düzenleyen mikroRNAlar (miRNA'lar) olarak bilinen küçük, kodlanmayan RNA'lar kapsamlı bir şekilde incelenmektedir. MiRNA'ların popülerliğinin çeşitli nedenleri vardır. Örneğin, çok çeşitli organizmalar miRNA'lar üretir ve bunların konak-parazit tedavilerine katılımları hakkında bazı raporlar vardır (Saçar Demirci vd., 2016; Acar vd., 2018). Ayrıca, birçok hastalık fenotipi miRNA'lar ile ilişkilidir ve miRNA'ları hastalık belirteçleri ve yeni terapötik ajanlar olarak kullanmak mümkündür (Avci ve Baran, 2014; Tüfekci vd., 2014). Bununla birlikte, bir ökaryotik genomun miRNA öncüleri üretme kapasitesi göz önüne alındığında, yeni miRNA'ları deneysel olarak ayırt etmek zor bir görevdir. Örneğin, önceki çalışlarımızdan elde ettiğimiz sonuçlara göre insan genomundan 200 milyondan fazla olası saç tokası yapısı üretilebilir ve bunların yaklaşık 60 milyonu genel miRNA özelliklerine sahiptir (Saçar Demirci vd. 2017). Tüm bu miRNA adaylarının ıslak – laboratuvar deneysel yöntemleriyle test edilmesi mümkün

değildir. Ayrıca bilinen miRNA'lardan elde edilen bilgilere göre, miRNA'lar genomda herhangi bir yerde bulunabilmektedirler (Kim vd. 2009). Bu nedenle miRNA tahmin yöntemlerinin tüm genomu aramaya elverişli tasarlanması gerekmektedir. Sonuç olarak güvenilir, hızlı ve doğru sonuçlar üreten biyoinformatik temelli miRNA tahmin yaklaşımlarının geliştirilmesi gereklidir.

Sonuç olarak, miRNA analizi için hesaplamalı yaklaşımların tasarlanması ve kullanılması önemli bir araştırma alanı haline gelmiştir. Bilişimsel yöntemlerle miRNA'ların tespitinin yapılabilmesi için bazı parametrelerin belirlenip hesaplanması gerekmektedir. Tek zincirli RNA dizilerinin kendinden katlanarak oluşturduğu ikincil yapılarının yanı sıra, miRNA'ların biyogenezi esnasında bazı enzimler tarafından tanınması ve değiştirilmesi gereklidir (Şekil 1) (Kozlowski vd., 2008). Bu nedenle, miRNA tahmin analizleri genellikle birincil ve ikincil yapılardan elde edilen bilgileri kullanır. Ne yazık ki, miRNA'ların karakteristik özelliklerinde biri olan saç tokası yapısı sadece miRNA'lara özgü bir özellik değildir (Roden vd., 2017). Belirli bir sekansın miRNA olup olmadığını belirlemek için tasarlanan araçların çoğu makine öğrenimi (ML) uygulamasına dayanmaktadır (Saçar Demirci vd., 2017). ML, miRNA çalışmaları için oldukça güçlü ve avantajlı olmasına rağmen, veri kalitesi, özellik seçimi ve ML algoritması seçimi gibi verimli bir analiz için dikkate alınması gereken bazı temel noktalar vardır (Saçar Demirci ve Allmer, 2017).



Şekil 1. MikroRNA'ların biyogenezi.

(A) Kanonik miRNA biyogenezi, miRNA genlerinin RNA Polimeraz II (POL II) veya POL III tarafından transkripsiyonu ile başlar. Daha sonra, birincil (pri) -miRNA'lar DROSHA / DiGeorge sendromu kritik bölge 8 (DGCR8) tarafından işlenir. Elde edilen pre-miRNA'lar Exportin-5 (XPO-5) ile sitoplazmaya gönderilir. MiRtronlar da intron ayrılması (splicing) sonucunda pre-miRNA olabilir. Sitoplazmada, pre-miRNA'lar DICER/trans-aktivasyon-duyarlı RNA bağlayıcı protein (TRBP) ile ayrılır (cleaved). Daha sonra, yolcu dizisi (passenger strand), (C3PO) kompleksi tarafından ayrıştırılır (degraded). RNA ile indüklenen susturma kompleksine (RISC) yüklenen kılavuz dizisi (guide strand), translasyonel baskı ve müteakip

transkript bozulmasında rol oynar. (B) MiR-451, DICER'den bağımsız bir şekilde işlenir. DROSHA / DGCR8 tarafından işlendikten ve sitoplazmaya ihraç edildikten sonra, yolcu dizisi Argonaute 2 (AGO2) aracılı kesilme ve düzeltme ile ayrıştırılır. (C) Küçük nükleolar RNA'ların (snoRNA'lar) kanonik olmayan işlenmesi, snoRNA türevi RNA'lara (sdRNA'lar) yol açar. SnoRNA'lar genlerden eklenir (spliced) ve DBR1 tarafından ayrılır. Daha sonra, snoRNA'lar bilinmeyen bir mekanizma ile sitoplazmaya ihraç edilir ve DICER tarafından RISC'ye yüklenen sdRNA'lara işlenir. (D) Transfer RNA'larının (tRNA'lar) kanonik olmayan işlenmesi, tRNA türevi miRNA'lar ile sonuçlanır. (1) Transkripsiyondan sonra, tRNA'lar XPO-5 veya XPO-T ile sitoplazmaya taşınır. 5p-döngüsü (5p-loop) ve 3p-döngüsü (3p-loop) DICER tarafından ayrılır, bu da sırasıyla 5p-tRNA türevi RNA (tDR)-parçaları ve 3 t-tDR-parçaları ile sonuçlanır. Antikodon döngüsü (anticodon loop), Angiogenin (ANG) tarafından ayrılır ve 5p-tRNA stresiyle indüklenen fragmanlar (tiRNA'lar) oluşur. Tüm tDR-fragmanları daha sonra kanonik miRNA'lara benzer şekilde RISC'ye yüklenir. (2) Transkripsiyondan sonra, tRNA'lar Lupus otoantijen (LA) ile stabilize edilebilir ve XPO-T ile sitoplazmaya ihraç edilebilir. LA, DICER tarafından tRNA'ların işlenmesini engeller ve translasyon için tRNA stabilitesini korur. (Stavast ve Erkeland 2019)

Genel olarak, çeşitli skorlama fonksiyonlarına sahip dinamik programlama tabanlı algoritmalar, RNA ikincil yapıları arasındaki benzerlikleri ölçmek için yaygın olarak kullanılmaktadır (Bafna vd., 1996; Dowel ve Eddy, 2006). Ancak, dinamik programlamaya dayanan yöntemler hesapsal olarak verimsizdir, bu da sahte düğüm (pseudo-knot) gibi karmaşık ikincil yapılara sahip RNA'ları tahmin etmeyi zorlaştırır (Zhang vd. 2016).

RNA benzerliğini daha verimli ölçmek için çeşitli alternatif teknikler önerilmiştir. Örneğin, RNA ikincil yapısının yeni bir 2B grafik temsili Yao vd. tarafından geliştirilmiştir (Yao vd. 2005). Ancak bu yöntem bir RNA dizisini benzersiz bir şekilde temsil edememesi nedeniyle bilgi kaybına neden olabilmektedir. Bilgi kaybı problemini çözmek için DNA dizilerinin kaos oyunu temsiline dayanan, RNA ikincil yapısının dejeneratif olmayan 2B grafik temsili önerildi (Li vd. 2008). Dizi ve baz kimyasal bilgilerini temel alan benzer iki 3B gösterim yöntemi de önerilmiştir (Jeffrey, 1990; Zhu vd., 2005). Bununla birlikte, bu yöntemlerin en önemli dezavantajlarından biri özellikle uzun RNAlar için alan gerektirmesidir (space demanding). Yüksek boyutlu bir temsil şeması olarak, yapı dejenerasyonu ve bilgi kaybı sorununu da çözmek için 4B bir yöntem geliştirilmiştir, ancak bu yaklaşım görselleştirme için iyi değildir (Liao vd., 2007). Ayrıca, kodlayıcı olmayan RNA ikincil yapılarını sınıflandırmak için yeni bir dalgacık tabanlı grafik gösterim yöntemi (wavelet-based graphical representation method) kullanılmıştır (Li vd., 2012). Bununla birlikte, bu yöntemle elde edilen veriler gereksizdir (redundant) çünkü her baz üç vektör ile karakterize edilir.

Literatürde RNA yapılarının temsiline dair geliştirilen ve uygulanan yöntemlerin hemen hepsinde RNA yapılarından benzerlik hesaplayarak filogenetik ağaçlar oluşturulması amaçlanmıştır. Örneğin 6 farklı çalışmada, 9 virüse ait RNA dizilerini içeren aynı veri seti kullanılarak filogenetik ağaçlar elde edilmiştir (Zhang vd. 2016; Liao vd. 2005; Liao ve Wang 2004; Yao vd. 2005; Li vd. 2008; Bai vd. 2005).

Bu proje kapsamında, bilinen miRNA öncüllerinin ve psödo-saç tokası yapılarının çok boyutlu temsiline dayanan miRNA tahmini için bir ML çerçevesi geliştirilmiştir. Öncü çalışmalarımızdan elde ettiğimiz bulgulara göre, RNA yapılarının gösterimine dayanan parametrelerle eğitilecek bir sınıflandırma algoritmasının etkili olma potansiyeli bulunmaktadır (Saçar Demirci, 2019). Ancak bu aşamada doğruluğu etkileyen en önemli unsurlardan biri parametrelerin seçimi ve hesaplanmasıdır. Bu nedenle parametre çıkarımı safhasında matematik alanından bir araştırmacının proje ekibinde bulunması son derece faydalı olmuştur. Proje kapsamında elde edilen yöntem, insan miRNA verilerine dayanılarak geliştirilmiş olmasına rağmen diğer organizmalar için de uygulanması ve / veya genişletilmesi mümkündür.

3. Gereç ve yöntem

MiRNA saç tokalarının tanımlanması genellikle 2 sınıflı sınıflandırma tabanlı ML yaklaşımları kullanılarak gerçekleştirilir. Model oluşturmak ve bu modellerin etkisini test etmek için farklı giriş veri setleri elde edilmiş ve bir iş akışı sisteminde çeşitli sınıflandırma algoritmaları kullanılacaktır.

3.1 Veri setleri

Eğitim ve testte kullanılan veri setleri aşağıdaki gibidir:

İnsan miRNA dizileri MiRBase'den (Sürüm 22.1), insan cirRNA veri seti circAtlas 2.0'dan, SARS-CoV-2 CDS, verileri ise NCBI RefSeq_NC_045512.2'den elde edildi. Farklı olarak ifade edilen miRNA (differentially expressed miRNA) listesi Chow ve Salmena'nın (2020) çalışmasında bulunan olgun miRNA'ların saç tokası öncü dizilerinden oluşmaktadır (Tablo 1).

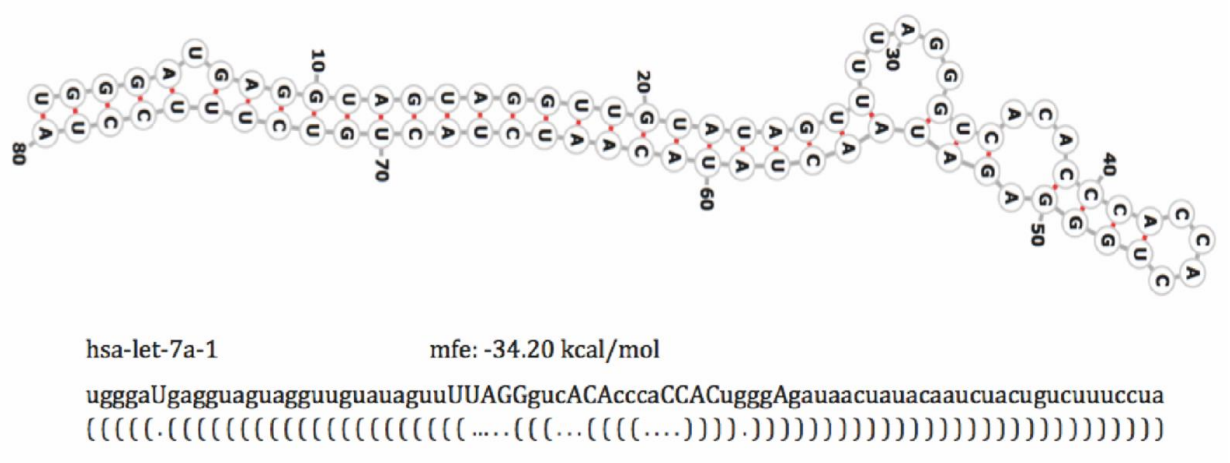
Tablo 1. Farklı ifade tahmini eğitimi için kullanılan miRNA'ların listesi.

Regülasyon	MiRNA
Artan	hsa-mir-4485, hsa-mir-483, hsa-mir-6891, hsa-mir-4284, hsa-mir-4463, hsa-mir-155, hsa-mir-107, hsa-mir-29b-2, hsa-mir-139, hsa-mir-299, hsa-mir-501, hsa-mir-4745, hsa-mir-12136

Azalan	hsa-let-7a-1, hsa-let-7a-2, hsa-let-7a-3, hsa-mir-374a, hsa-mir-194-1, hsa-mir-194-2, hsa-mir-4454, hsa-mir-135b, hsa-mir-16-2, hsa-mir-23b, hsa-mir-21, hsa-let-7f-1, hsa-mir-429, hsa-mir-5701-1, hsa-mir-5701-2, hsa-mir-5701-3, hsa-mir-450b, hsa-mir-7-1, hsa-mir-26b, hsa-mir-23c, hsa-mir-374c, hsa-mir-374b, hsa-mir-26a-1, hsa-mir-365a, hsa-mir-365b, hsa-mir-940, hsa-mir-362, hsa-mir-1275, hsa-mir-1296, hsa-mir-126, hsa-mir-548d-2

3.2. RNA ikincil yapılarının 3B grafik gösterimi

RNA sekanslarının ikincil yapıları, varsayılan ayarlarla RNAfold (Hofacker, 2003) kullanılarak elde edilmiştir. Her sekans için en iyi yapı, minimum serbest enerji değerlerine göre seçilmiştir (Şekil 2). Nokta yapıların (sırasıyla bağlanma ve bağlanma bazları) 2B yapıların temsiline göre, dizideki bazlar işaretlenmiştir (büyük ve küçük harf olarak gösterim, A, A', G, G' şeklinde ifadesi vb.). Bu sekanslar daha sonra RNA ikincil yapılarını karakterize eden vektörleri üretmek için KNIME platformunda kullanılacaktır (Şekil 2).



Şekil 2. hsa-let-7a-1 dizi ve ikincil yapısının temsili. mfe: Minimum free energy (minimum serbest enerji) (Saçar Demirci 2019).

3.3. Parametre analizi

Veri setlerinin sınıflandırma ve kümeleme gibi makine öğrenmesi yöntemleriyle analiz edilebilmesi için bu veri setleri için parametrelerin belirlenmesi ve hesaplanması gereklidir. Literatürde yer alan çalışmalara göre, bazı parametreler birbiriyle yüksek korelasyona sahip, bazıları ise herhangi bir bilgi kazancı sağlamamaktadır. Parametrelerin hesaplanması için gereken zaman ve hesaplama gücü göz önünde bulundurulduğunda, etkin bir özellik seçim metodolojisi bütün analizin önemli bir bileşeni haline gelir. Bu nedenle RNA ikincil yapılarını temel alan etkin parametrelerin cebirsel yöntemlerle belirlenip uygulanması ve elde edilecek veri matrislerinin değerlendirilmesi projenin en önemli aşamalarından biridir. Parametrelerin hesaplanması için KNIME platformu kullanılmış, gerekli durumlarda R kodları yazılmıştır.

Sınıflandırma analizini büyük ölçüde zorlaştıran en büyük engellerden biri, verilerin boyutunun artmasıdır. Ayrıca, veriler yalnızca büyüklük değil aynı zamanda kapladığı alanda da seyrek (sparse) olabilir. Bu fenomen aynı zamanda boyutsallık laneti (curse of dimentionality) olarak da bilinir ve durum makine öğrenmesi için büyük sorunlara neden olabilir (Powell, 2011). Sonuç olarak, sınıflandırma analizi için birçok parametreyi kullanmak, yüksek bir hesaplama maliyetiyle daha düşük sınıflandırma doğruluğu ile sonuçlanabilir.

Gerçek biyolojik veri setleri ile uğraşırken, önemli parametreler çoğunlukla bilinmemektedir. Dolayısıyla, veri setlerini doğru bir şekilde korumak ve temsil etmek için çok sayıda parametre tasarlanmış ve hesaplanmıştır. Ancak bu parametrelerin çoğu bilgilendirici / yararlı olmama eğilimindedir. Veri seti boyutunun büyük olabildiği miRNA analizi gibi durumlarda, daha az zamanda daha iyi bir öğrenime sahip olmak için ilgisiz / gereksiz parametrelerin çıkarılması önemlidir.

3.4 Parametre tanımlama

KNIME'de oluşturulan iş akışında, dizinin nükleotidlerini büyük harf ve küçük harf karakterleri olarak değiştirmek için RNA dizisini ve ikincil yapının nokta-parantez temsilleri kullanıldı.

Zhang vd. bazların kimyasal özelliklerine dayalı olarak RNA yapısı için dinamik bir 3B grafik gösterimi oluşturmuştu:

(i) amino grubu M = {A, C} ve keto grubu K = {G, U},

(ii) pürin grubu R = {A, G} ve pirimidin grubu Y = {C, U}

(iii) zayıf grup H-bağları W = {A, U} ve güçlü H-bağları grubu S = {C, G}.

Benzer şekilde RNA ikincil yapısı için 3 nokta setimiz vardır (x1i, y1i, z1i), (x2i, y2i, z2i), ve (x3i, y3i, z3i), i = 1, 2, …, n, burada n dizinin/yapının uzunluğunu göstermektedir. Buradan 36 boyutlu bir vektör oluşturmak mümkündür (Tablo 2).

Tablo 2. Tanımlar (α1, α2 ve α3)

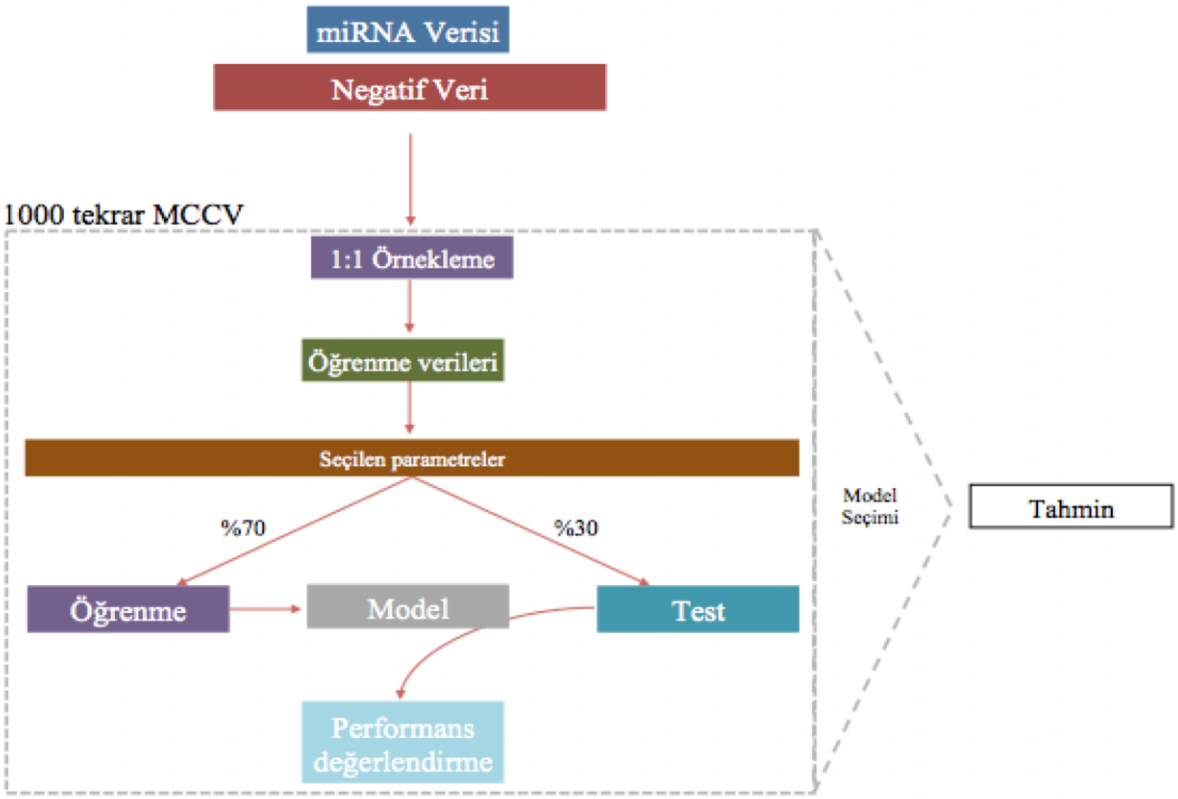
g_i	$\alpha_1(g_i)=(x_{1i},y_{1i},z_{1i})$			g_i	$\alpha_2(g_i)=(x_{2i},y_{2i},z_{2i})$			g_i	$\alpha_3(g_i)=(x_{3i},y_{3i},z_{3i})$		
	x_{1i}	y_{1i}	z_{1i}		x_{2i}	y_{2i}	z_{2i}		x_{3i}	y_{3i}	z_{3i}
$\{A \text{ or } C\}$	$\sin(\frac{2\pi i}{n})$	$\cos(\frac{2\pi i}{n})$	2^i	$\{A \text{ or } G\}$	$\sin(\frac{2\pi i}{n})$	$\cos(\frac{2\pi i}{n})$	3^i	$\{A \text{ or } U\}$	$\sin(\frac{2\pi i}{n})$	$\cos(\frac{2\pi i}{n})$	5^i
$\{G \text{ or } U\}$	$\sin(\frac{2\pi i}{n})$	$-\cos(\frac{2\pi i}{n})$	3^i	$\{C \text{ or } U\}$	$\sin(\frac{2\pi i}{n})$	$-\cos(\frac{2\pi i}{n})$	5^i	$\{C \text{ or } G\}$	$\sin(\frac{2\pi i}{n})$	$-\cos(\frac{2\pi i}{n})$	7^i
$\{A' \text{ or } C'\}$	$-\sin(\frac{2\pi i}{n})$	$\cos(\frac{2\pi i}{n})$	5^i	$\{A' \text{ or } G'\}$	$-\sin(\frac{2\pi i}{n})$	$\cos(\frac{2\pi i}{n})$	7^i	$\{A' \text{ or } U'\}$	$-\sin(\frac{2\pi i}{n})$	$\cos(\frac{2\pi i}{n})$	2^i
$\{G' \text{ or } U'\}$	$-\sin(\frac{2\pi i}{n})$	$-\cos(\frac{2\pi i}{n})$	7^i	$\{C' \text{ or } U'\}$	$-\sin(\frac{2\pi i}{n})$	$-\cos(\frac{2\pi i}{n})$	2^i	$\{C' \text{ or } G'\}$	$-\sin(\frac{2\pi i}{n})$	$-\cos(\frac{2\pi i}{n})$	3^i

Bu vektörün hesaplanması için de şu ve benzeri işlemler yapılabilir:

$$x_1^1=\frac{1}{n}\sum_{i=1}^{n}x_{1i}^{A,C},\ y_1^1=\frac{1}{n}\sum_{i=1}^{n}y_{1i}^{A,C},\ z_1^1=\frac{1}{2^n}\sum_{i=1}^{n}z_{1i}^{A,C},\ x_1^2=\frac{1}{n}\sum_{i=1}^{n}x_{1i}^{G,U},\ y_1^2=\frac{1}{n}\sum_{i=1}^{n}y_{1i}^{G,U},\ z_1^2=\frac{1}{3^n}\sum_{i=1}^{n}z_{1i}^{G,U},$$
$$x_1^3=\frac{1}{n}\sum_{i=1}^{n}x_{1i}^{A',C'},\ y_1^3=\frac{1}{n}\sum_{i=1}^{n}y_{1i}^{A',C'},\ z_1^3=\frac{1}{5^n}\sum_{i=1}^{n}z_{1i}^{A',C'},\ x_1^4=\frac{1}{n}\sum_{i=1}^{n}x_{1i}^{G',U'},\ y_1^4=\frac{1}{n}\sum_{i=1}^{n}y_{1i}^{G',U'},\ z_1^4=\frac{1}{7^n}\sum_{i=1}^{n}z_{1i}^{G',U'},$$
$$x_2^1=\frac{1}{n}\sum_{i=1}^{n}x_{2i}^{A,G},\ y_2^1=\frac{1}{n}\sum_{i=1}^{n}y_{2i}^{A,G},\ z_2^1=\frac{1}{3^n}\sum_{i=1}^{n}z_{2i}^{A,G},\ x_2^2=\frac{1}{n}\sum_{i=1}^{n}x_{2i}^{C,U},\ y_2^2=\frac{1}{n}\sum_{i=1}^{n}y_{2i}^{C,U},\ z_2^2=\frac{1}{5^n}\sum_{i=1}^{n}z_{2i}^{C,U},$$
$$x_2^3=\frac{1}{n}\sum_{i=1}^{n}x_{2i}^{A',G'},\ y_2^3=\frac{1}{n}\sum_{i=1}^{n}y_{2i}^{A',G'},\ z_2^3=\frac{1}{7^n}\sum_{i=1}^{n}z_{2i}^{A',G'},\ x_2^4=\frac{1}{n}\sum_{i=1}^{n}x_{2i}^{C',U'},\ y_2^4=\frac{1}{n}\sum_{i=1}^{n}y_{2i}^{C',U'},\ z_2^4=\frac{1}{2^n}\sum_{i=1}^{n}z_{2i}^{C',U'},$$
$$x_3^1=\frac{1}{n}\sum_{i=1}^{n}x_{3i}^{A,U},\ y_3^1=\frac{1}{n}\sum_{i=1}^{n}y_{3i}^{A,U},\ z_3^1=\frac{1}{5^n}\sum_{i=1}^{n}z_{3i}^{A,U},\ x_3^2=\frac{1}{n}\sum_{i=1}^{n}x_{3i}^{C,G},\ y_3^2=\frac{1}{n}\sum_{i=1}^{n}y_{3i}^{C,G},\ z_3^2=\frac{1}{7^n}\sum_{i=1}^{n}z_{3i}^{C,G},$$
$$x_3^3=\frac{1}{n}\sum_{i=1}^{n}x_{3i}^{A',U'},\ y_3^3=\frac{1}{n}\sum_{i=1}^{n}y_{3i}^{A',U'},\ z_3^3=\frac{1}{2^n}\sum_{i=1}^{n}z_{3i}^{A',U'},\ x_3^4=\frac{1}{n}\sum_{i=1}^{n}x_{3i}^{C',G'},\ y_3^4=\frac{1}{n}\sum_{i=1}^{n}y_{3i}^{C',G'},\ z_3^4=\frac{1}{3^n}\sum_{i=1}^{n}z_{3i}^{C',G'}.$$

3.5. Veri madenciliği

Veri madenciliği analizi için Konstanz Bilgi Madencisi (KNIME) (Berthold vd., 2008) platformu kullanılmıştır. Öğrenme için en az 3 sınıflandırıcı; Rastgele Orman (Random Forest), Karar Ağacı (Decision Tree), Saf Bayes (Naive Bayes) vb., miRBase'den insan miRNA'larını pozitif, psödo dizileri ise negatif örnek olarak kullanacak şekilde eğitilmiştir (Şekil 3). Sınıf dengesizliğini (imbalanced data) önlemek için, her iki veri kümesinden de eşit boyutlu örnekler rastgele seçilip ve %70 öğrenme - %30 test setleri 1000 kat Monte Carlo çapraz doğrulaması ile uygulanmıştır (Xu ve Liang, 2001). Her bir sınıflandırıcıdan en yüksek doğruluk skoruna sahip modeller kaydedilip sonraki test analizi için kullanılmıştır.



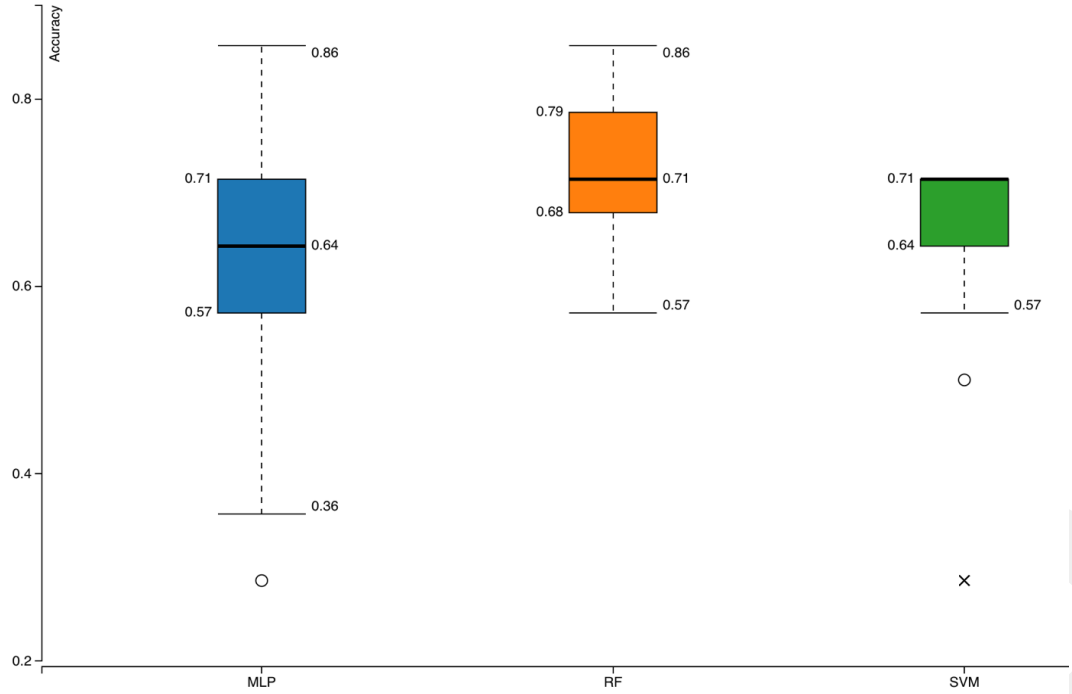
Şekil 3. Öğrenme iş akış diyagramı.

Farklı ifade tahmin iş akışı benzer şekilde %70 öğrenme ve %30 test oranları, 100 kat MCCV ve üç farklı sınıflandırıcı kullanılarak oluşturuldu; Rastgele Orman (RF), Destek Vektör Makinesi (SVM) ve Çok Katmanlı Algılayıcı (MLP) ile oluşturuldu.

MiRNA hedef tahmini işlemleri için psRNATarget yazılımı kullanıldı.

4. Bulgular

Kısıtlı SARS-CoV-2 farklı ifade edilen miRNA veri setiyle eğitilen sınıflandırıcıların doğruluk (accuracy) değerleri beklendiği gibi farklılıklar göstermiştir (Şekil 4). Tamamıyla aynı verilerle ve ayarlarla yapılan analizlere göre bu farkın nedeni sınıflandırıcıların algoritmalarındaki farklardır.



Şekil 4. Sınıflandırıcıların doğruluk değerlerinin kutu grafikleri.

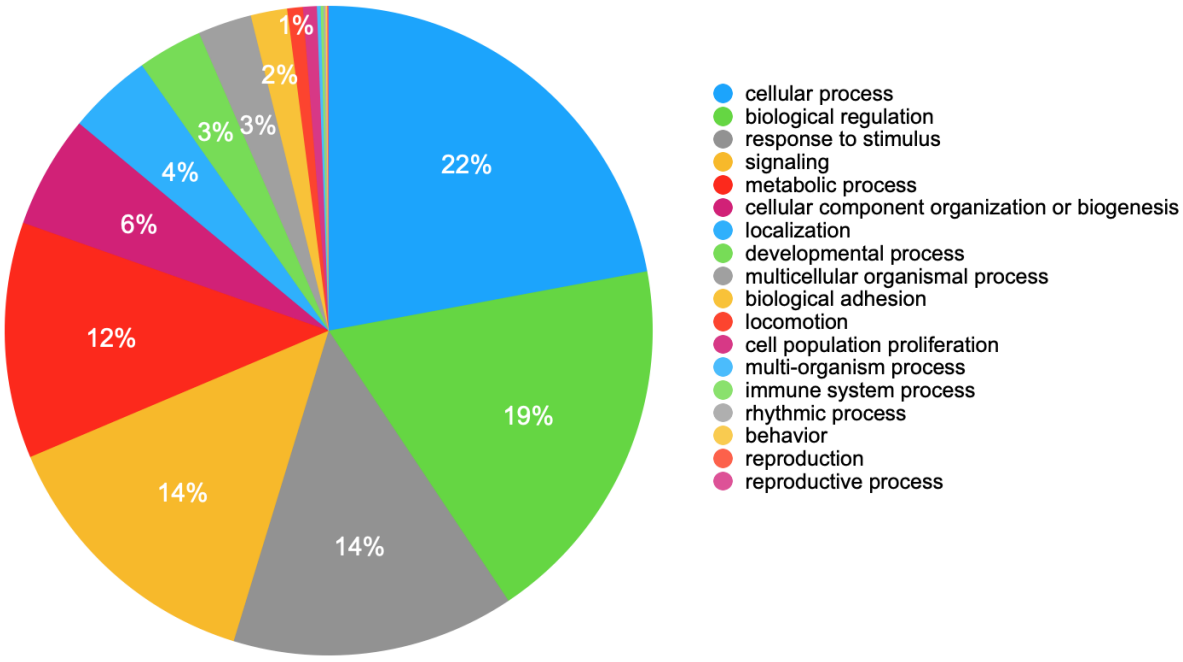
MiRBase'de bulunan 2,654 olgun insan miRNAsından 2,498'i, 18,950 insan geni ile toplamda 272,822 hedefleme olayına dahil olmuştur. Benzer şekilde 2,498 miRNA 208.642 circRNA ile toplam 393.877 hedefleme olayında, 484 miRNA ise 11 SARS-CoV-2 genini hedeflemede rol almıştır (Tablo 3).

Tablo 3. İnsan genleri (Gen), insan dairesel RNA'ları (CircRNA) ve SARS-CoV-2 kodlama dizileri üzerinde farklı şekilde ifade edilen miRNA'ların hedef sayısı.

MiRNA	Gen	CircRNA	SARS-CoV-2	Regülasyon
hsa-miR-6891-5p	197	200	1 (ORF3a)	Artan
hsa-miR-4284	-	-	-	Artan
hsa-miR-4463	-	-	-	Artan
hsa-miR-12136	-	-	-	Artan
hsa-miR-181-5p	-	-	-	Artan
hsa-miR-126-5p	130	193	1 (ORF1ab)	Azalan
hsa-miR-194-5p	76	132	1 (ORF1ab)	Azalan
hsa-miR-374a-3p	100	155	2 (ORF1ab, S)	Azalan

hsa-miR-181-3p	-	-	-	Azalan
hsa-miR-1275	-	-	-	Azalan

Regülasyonu artan insan miRNA'sı hsa-miR-6891-5p sadece insan genlerini ve cirRNA'ları değil aynı zamanda SARS-CoV-2'nin ORF3a genini de hedefleyebilir (Tablo). İnsan gen hedeflerinin PANTHER Gene Ontoloji analizi, çeşitli biyolojik süreçlerin bu miRNA'nın eylemlerinden potansiyel olarak etkilenebileceğini göstermiştir (Şekil 5).



Şekil 5. hsa-miR-6891-5p tarafından hedeflenebilecek insan genlerinin biyolojik süreçleri için pasta grafiği. Sağ kısımdaki etiketler, grafik yönünü saat yönünde olacak şekilde azalan sırada sıralanır.

5. Tartışma

MiRNA tanımlaması için geliştirilen yöntemlerin çoğu makine öğrenmesine dayanmaktadır; dolayısıyla, makine öğrenmesinin zorluklarından da etkilenirler. Örneğin başarılı bir sınıflandırma sistemi için en önemli kriterlerden biri yüksek kaliteli veri setleri kullanabilmektir (Saçar Demirci ve Allmer, 2017). MiRNA analizi için, miRBase ve MirGeneDB gibi halka açık veri tabanlarında bulunan bilinen miRNA'lar pozitif veri seti olarak kullanılır. Kaliteli negatif veri setlerinin oluşturulabilmesi için, dizilerin bilinen miRNA'lara benzer özelliklere sahip olması gereklidir ancak benzerlik seviyesi algoritmaların pozitif ve negatif örenkleri doğru bir şekilde ayırt edebilmesi için çok yüksek olmamalıdır. Bu zorluklar nedeniyle, şu anda gerçek

ve doğrulanmış negatif veri setine sahip olmak imkansızdır. Bu çalışma için sözde saç tokası olarak bilinen en popüler negatif veri kümesi seçildi ve kullanıldı.

Yalnızca negatif veri kümeleri değil, aynı zamanda pozitif veri kümelerinin de daha fazla iyileştirmeye ihtiyacı var gibi görünüyor. Daha önce, miRBase'deki bazı girdilerin gerçek miRNA'lar olma ihtimalinin düşük olduğu gösterilmişti (Saçar vd., 2013). Proje esnasında elde edilen ön sonuçlara göre de kalite açısından MirGeneDB'de listelenen insan miRNA'larının miRBase'deki insan miRNA girdilerinden daha iyi olduğu gözlemlenmiştir. Bununla birlikte miRBase, 286 organizmadan miRNA dizi bilgisi sağlayan standart kaynak (Sürüm 22) olduğu için miRNA çalışmalarında sürekli olarak kullanılmaktadır.

Makine öğrenmesi analizleri için, verileri matematiksel olarak açıklayan bazı parametreler gereklidir. MiRNA'lar için önerilen ve kullanılan çeşitli özellikler; yapısal, dizi tabanlı, olasılık tabanlı ve termodinamik parametreler olarak gruplandırılabilir. Daha önceki çalışmalarımızda, bu tür yüzlerce özelliği uyguladık, ancak yaklaşık 50 özelliğin genellikle etkili bir makine öğrenimi modeli oluşturmak için yeterli olduğunu bulduk (Saçar ve Allmer, 2013). Bununla birlikte, bu tür özelliklerin hesaplanması, özellikle genom çapında bir miRNA araştırması için hesaplama açısından oldukça pahalıdır. Ayrıca, bilgilendirici özelliklerin seçimi, genel model performansı üzerinde büyük etkisi olan önemli bir adımdır (Yousef vd., 2016). Bu nedenle, RNA ikincil yapılarının 3B grafik temsilini miRNA'ları tanımlayan özellikler olarak kullanmak gibi alternatif bir yaklaşım umut verici bir yaklaşımdır.

RNA dizilerinin 2D ve 3D gösterimleri, yapısal bilgilere dayalı bir veri matrisi oluşturur. Bu tür temsiller RNA benzerliklerini ölçmek ve virüsleri sınıflandırmak için kullanılmış olsa da (Yao vd., 2005; Li vd., 2012), bunlar nadiren miRNA öncesi analiz için uygulanır (Fu vd., 2018). Bu çalışmada geliştirilen iş akışı, ML tabanlı miRNA tahmini için RNA'ların 3B temsillerinin uygulanmasının ilk örneklerindendir. Proje kapsamında elde edilen sonuçlar, 3B parametrelerinin yüksek kaliteli bir veri setinde kullanıldığında, miRNA analizi için başarılı bir model oluşturmak için yeterli olduklarına işaret etmektedir.

6. Sonuç ve Öneriler

RNA'ların aracılık ettiği türler arası iletişim mekanizmaları, çeşitli virüsler, *Toxoplasma gondii*, *Histoplasma capsulatum* (enfeksiyöz mantar) dahil olmak üzere çeşitli organizmalar için araştırılmıştır. Virüsler, işlemlerinin çoğu için konaklarına bağlı olan parazitlerdir. Genellikle viral enfeksiyonlar, viral gen ekspresyonunu modüle etmek ve/veya virüsü uygun bir ortamda barındırmak için hücresel yollarda değişikliklere neden olur. SARS-CoV-2 enfeksiyonu gibi durumlarda, miRNA'lar gibi konakçı transkripsiyon sonrası gen düzenleme elemanları da enfeksiyon sırasında farklı ifade seviyeleri gösterebilir (Chow ve Salmena, 2020). Bu proje kapsamında geliştirilen iş akışı, SARS-CoV-2 enfeksiyonu sırasında miRNA'ların ifade

değişikliklerinin tahmininde kullanılmıştır (Tablo 1). Oluşturulan 300 model arasında en yüksek doğruluk değeri RF sınıflandırıcı ile gözlenmiştir (Şekil 4).

MiRNA veri setlerine makine öğrenmesi yaklaşımlarını uygularken, genel performansı etkileyecek çeşitli öğeler vardır. Bunların arasında, parametre setleri ve verilerin kalitesi en önemli parçalar olabilir. Daha fazla veri kümesi mevcut olduğunda, geliştirilen iş akışı yeni verileri içerecek şekilde kolayca güncellenebilecektir.

Kaynaklar

Acar, İ.E., Saçar Demirci, M.D., Groß, U., Allmer, J., 2018. The Expressed MicroRNA—mRNA Interactions of Toxoplasma gondii. Front. Microbiol. 8.

Avci, Ç.B., Baran, Y., 2014. Use of MicroRNAs in Personalized Medicine, in: Methods in Molecular Biology (Clifton, N.J.). pp. 311–325.

Bafna, V., Muthukrishnan, S., Ravi, R., 1995. Computing similarity between RNA strings. Springer, Berlin, Heidelberg, pp. 1–16.

Bai, F., Zhu, W., Wang, T., 2005. Analysis of similarity between RNA secondary structures. Chem. Phys. Lett. 408, 258–263.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2008. KNIME: The Konstanz Information Miner, in: SIGKDD Explorations. pp. 319–326.

Chow, J.T.-S., Salmena, L., 2020. Prediction and Analysis of SARS-CoV-2-Targeting MicroRNA in Human Lung Epithelium. Genes (Basel). 11, 1002.

Dowell, R.D., Eddy, S.R., 2006. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics 7, 400.

Filipowicz, W., Bhattacharyya, S.N., Sonenberg, N., 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. 9, 102–14.

Fu, X., Liao, B., Zhu, W., Cai, L., 2018. New 3D graphical representation for RNA structure analysis and its application in the pre-miRNA identification of plants. RSC Adv. 8, 30833–30841.

Hofacker, I.L., 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31, 3429–3431.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.

Kim, V.N., Han, J., Siomi, M.C., 2009. Biogenesis of small RNAs in animals. Nat. Rev. Mol. Cell Biol. 10, 126–39.

Kozlowski, P., Starega-Roslan, J., Legacz, M., Magnus, M., Krzyzosiak, W.J., 2008. Structures of MicroRNA Precursors, in: Current Perspectives in MicroRNAs (MiRNA). Springer Netherlands, Dordrecht, pp. 1–16.

Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–54.

Li, C., Xing, L., Wang, X., 2008. Analysis of similarity of RNA secondary structures based on a 2D graphical representation. Chem. Phys. Lett. 458, 249–252.

Li, Y., Duan, M., Liang, Y., 2012. Multi-scale RNA comparison based on RNA triple vector curve representation. BMC Bioinformatics 13, 280.

Liao, B., Zhu, W., Li, P., 2007. On a four-dimensional representation of RNA secondary structures. J. Math. Chem. 42, 1015–1022.

Piast, M., Kustrzeba-Wójcicka, I., Matusiewicz, M., Banaś, T., 2005. Molecular evolution of enolase. Acta Biochim. Pol. 52, 507–513.

Powell, W.B., Wiley InterScience (Online service), 2011. Approximate dynamic programming : solving the curses of dimensionality. Wiley.

Roden, C., Gaillard, J., Kanoria, S., Rennie, W., Barish, S., Cheng, J., Pan, W., Liu, J., Cotsapas, C., Ding, Y., Lu, J., 2017. Novel determinants of mammalian primary

microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. Genome Res. 27, 374–384.

Saçar Demirci, M.D., 2019. MicroRNA prediction based on 3D graphical representation of RNA secondary structures. Turkish J. Biol. 43, 274–280.

Saçar Demirci, M.D., Allmer, J., 2017. Delineating the impact of machine learning elements in pre-microRNA detection. PeerJ 5, e3131.

Saçar Demirci, M.D., Bağci, C., Allmer, J., 2016. Differential expression of toxoplasma gondii microRNAs in murine and human hosts, Non-coding RNAs and Inter-kingdom Communication.

Saçar Demirci, M.D., Baumbach, J., Allmer, J., 2017. On the performance of pre-microRNA detection algorithms. Nat. Commun. 8.

Saçar, M.D., Allmer, J., 2013. Comparison of Four Ab Initio MicroRNA Prediction Tools, in: Bioinformatics 2013, Barcelona, Spain. SciTePress - Science and and Technology Publications, Barcelona, pp. 190–195.

Saçar, M.D., Hamzeiy, H., Allmer, J., 2013. Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? J. Integr. Bioinform. 10, 215.

Stavast, C.J., Erkeland, S.J., 2019. The Non-Canonical Aspects of MicroRNAs: Many Roads to Gene Regulation. Cells 8.

Tüfekci, K.U., Öner, M.G., Meuwissen, R.L.J., Genç, Ş., 2014. The Role of MicroRNAs in Human Diseases, in: Methods in Molecular Biology (Clifton, N.J.). pp. 33–50.

Varani, G., McClain, W.H., 2000. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. EMBO Rep. 1, 18–23.

Xu, Q.-S., Liang, Y.-Z., 2001. Monte Carlo cross validation. Chemom. Intell. Lab. Syst. 56, 1–11.

Yao, Y.-H., Nan, X.-Y., Wang, T.-M., 2005. A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them. J. Comput. Chem. 26, 1339–1346.

Yao, Y.H., Liao, B., Wang, T.M., 2005. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. J. Mol. Struct. THEOCHEM 755, 131–136.

Yousef, M., Saçar Demirci, M.D., Khalifa, W., Allmer, J., 2016. Feature selection has a large impact on one-class classification accuracy for microrna s in plants. Adv. Bioinformatics 2016.

Zhang, Y., Huang, H., Dong, X., Fang, Y., Wang, Kejing, Zhu, L., Wang, Ke, Huang, T., Yang, J., 2016. A dynamic 3D graphical representation for RNA structure analysis and its application in non-coding RNA classification. PLoS One 11, 1–15.

Zhu, W., Liao, B., Ding, K., 2005. A condensed 3D graphical representation of RNA secondary structures. J. Mol. Struct. THEOCHEM 757, 193–198.

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Dr. Öğr. Üyesi MÜŞERREF DUYGU SAÇAR DEMİRCİ
Proje No:	120E042
Proje Başlığı:	RNA İkincil Yapılarının Çok Boyutlu Gösterimi ve Pre-Mirna Tespiti Için Uygulamaları
Proje Türü:	1002 - Hızlı Destek
Proje Süresi:	10
Araştırmacılar:	YILMAZ MEHMET DEMİRCİ
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	ABDULLAH GÜL Ü. YAŞAM VE DOĞA BİLİMLERİ F.
Projenin Başlangıç ve Bitiş Tarihleri:	15/06/2020 - 15/04/2021
Onaylanan Bütçe:	20714.68
Harcanan Bütçe:	15027.97
Öz:	MikroRNA'lar (miRNA'lar), transkripsiyon sonrası gen ekspresyonu düzenleyicileridir. Bir miRNA yüzlerce haberci RNA'yı (mRNA'lar) hedefleyebildiği gibi, bir mRNA farklı miRNA'lar tarafından hedeflenebilir, üstelik tek bir miRNA bir mRNA sekansında çeşitli bağlanma bölgelerine sahip olabilir. Bu nedenle miRNA'ları deneysel olarak araştırmak oldukça karmaşıktır. Bu tür zorlukları aşabilmek için makine öğrenimi (ML) sıklıkla kullanılmaktadır. ML analizinin temel kısımları büyük ölçüde giriş verilerinin kalitesine ve verileri tanımlayan özelliklerin kapasitesine bağlıdır. Daha önce miRNA'lar için 1000'den fazla özellik önerilmişti. Bu projede, RNA ikincil yapısını temsil eden yeni özellikler ve yüksek doğruluk değerleri sağlayan, dinamik, çok boyutlu grafik gösterimini tanımlamayı hedeflemiştik. Bu çalışmada, ML tabanlı miRNA tahmini için yeni ve kolayca güncellenebilir bir yaklaşım geliştirilmiştir. Bilinen insan miRNA'larının ve sözde saç tokalarının random forest (RF), support vector machine (SVM) ve multilayer perceptron (MLP) gibi çeşitli sınıflandırıcılarla sınıflandırılmasıyla binlerce model oluşturulmuştur. Yöntem insan verilerine dayanarak oluşturulmuş olsa da en iyi model miRBase ve MirGeneDB gibi kamu veri tabanlarından insan olmayan saç tokaları üzerinde test edilmiş ve yüksek skorlar üretilmiştir. Ayrıca, yöntemin farklı veriler üzerindeki etkinliğini göstermek için ekspresyon farkları tahmini (differential expression prediction) analizinde de kullanılmıştır. Bu aşamada SARS-CoV-2 enfeksiyonunun etkisini ölçen bir veri setinin analizinden elde edilen sonuçlar yayınlanmıştır.
Abstract:	MicroRNAs (miRNAs) are posttranscriptional regulators of gene expression. While a miRNA can target hundreds of messenger RNA (mRNAs), an mRNA can be targeted by different miRNAs, not to mention that a single miRNA might have various binding sites in an mRNA sequence. Therefore, it is quite complicated to investigate miRNAs experimentally. Thus, machine learning (ML) is frequently used to overcome such challenges. The key parts of a ML analysis largely depend on the quality of input data and the capacity of the features describing the data. Previously, more than 1000 features were suggested for miRNAs. In this project, we aim to define new features representing the RNA secondary structure and its dynamic multidimensional graphical representation providing high accuracy values. In this study, a new and easily updateable approach for ML-based miRNA prediction has been developed. Thousands of models have been created by classifying known human miRNAs and pseudo hairpins with various classifiers such as random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP). Although the method was created based on human data, the best model was tested on non-human hairpins from public databases such as miRBase and MirGeneDB and high scores were produced. It has also been used in differential expression prediction analysis to show the effectiveness of the method on different data sets. At this stage, the results obtained from the analysis of a data set measuring the impact of SARS-CoV-2 infection have been published.
Anahtar Kelimeler:	miRNA, tahmin, makine öğrenmesi, model
Fikri Ürün Bildirim Formu Sunuldu Mu?:	Hayır
Projeden Yapılan Yayınlar:	1- Circular RNA-MicroRNA-MRNA interaction predictions in SARS-CoV-2 infection (Makale - Diğer Hakemli Makale),