Mikro-RNA Metabolik Ağ Kontrol Analizi için Veri Ambarı

Proje No: 113E326

Doç. Dr. Jens ALLMER
Müşerref Duygu SAÇAR DEMİRCİ
İlhan Erkin ACAR

Eylül 2017
İzmir

ÖNSÖZ

MikroRNA, yaklaşık 18-22 nükleotit uzunluğundaki, kodlama yapmayan, düzenleyici RNA'lardır. Birçok mikroRNA günümüzde keşfedilmiş olsa da, farklı sebeplerden dolayı henüz deneysel yöntemlerle bulunamamış milyonlarca mikroRNA olduğu tahmin edilmektedir. MikroRNA'lar, hedefleri olan mRNA'ları baskılayarak düzenleyici rollerini gerçekleştirirler. Bir mikroRNA, birden çok mRNA'yı düzenleyebilir. Tüm olası mikroRNA – mRNA etkileşimlerinin deneysel olarak bulunması masraflı ve oldukça uzun süren bir iştir. Bu yüzden, hem yeni mikroRNA'ların bulunmasında, hem de bu mikroRNA'ların hedeflerinin tahmin edilmesinde, bilişimsel yollar tercih edilmeye başlanmıştır. Makine öğrenimi içeren bu yöntemlerle alınan sonuçlar, deneysel uygulamalara aktarılarak mikroRNA'ların daha verimli şekilde anlaşılmasını sağlamıştır. Bu proje kapsamında, bir genomdaki olası tüm mikroRNA'ları çıkarıp, bu mikroRNA'ların olası tüm hedeflerini belirleyecek bir paket yazılım geliştirilmesine çalışılmıştır. Makine öğrenim algoritmalarını hali hazırda keşfedilmiş mikroRNA'lar ile birlikte uygulayarak, yüksek güven aralıklarında, yeni mikroRNA'ların tahmin edilmesi sağlanmıştır. Bu proje sayesinde geliştirilen mikroRNA tahmin iş akışımız, birçok farklı çalışmada kullanıldığı gibi önemli bilimsel dergilerde de yayınlanmıştır. Ne yazık ki mikroRNA hedefleme algoritmalarının, deneysel olarak kanıtlanmış gerçek hedefleri bulma yüzdesi, yapılan denemelerimiz sonucunda yeterli bulunmadığından, hedefleme tahminleri, proje kapsamında planladığı şekilde tamamlanamamıştır. Bu yüzden, yapılan çalışmalarda hedefleme tahmini için mevcut olan araçlar kullanılarak elde edilen önemli sonuçlar başarıyla yayınlanmıştır. Bu araçların birlikte kullanımıyla, mikroRNA'ların düzenleyici ağlarının incelenmesi, *Toxoplasma gondii* çalışmamızda görülebileceği üzere, mümkün olmuştur. Ayrıca, mikroRNA ağlarının, gen yolaklarıyla birlikte incelenmesini ve bunların bir arada gösterilmesinde görselleştirme kolaylıkları sağlayan, Alman işbirlikçilerimizin geliştirdiği, VANESA aracıyla kızamık hastalığındaki mikroRNA etkileşimlerinin gösterimi yapılabilmiştir.

Bu çalışmanın yürütülmesini, 111E326 numaralı proje ile destekleyen TÜBİTAK - Elektrik, Elektronik ve Enformatik Araştırma Destek Grubuna (EEEAG) teşekkür ederiz.

İÇİNDEKİLER

TABLO VE ŞEKİL LİSTELERİ

ÖZET

MikroRNA’lar (miRNA) uzunluğu yaklaşık 22 nükleotid olan, tek diziden oluşan ve kodlama özelliği olmayan küçük RNA'lardır ve gen ekspresyonunu transkripsiyon sonrası seviyede hedefleri olan mRNA'ların translasyonel baskılanması ve istikrarsızlaştırılması yoluyla kontrol ederler. Çeşitli türlerde yüzlerce miRNA tespit edilmesine rağmen, miRNA’ların büyük bir miktarı hala bilinmemektedir. Bu nedenle, yeni miRNA genlerinin keşfi, miRNA aracılığıyla düzenlenen transkripsiyon sonrası düzenleme mekanizmalarının anlaşılması için önemli bir adımdır. Konvansiyonel ileri genetik tarama, klonlanmış ürünleri domine eden, yüksek miktarda sentezlenen ve/veya her yerde görülen miRNA’lara karşı yanlı bir yöntemdir. Fakat bu tarz biyolojik yöntemler nadir miRNA’ların saptanmasında etkisiz kalmaktadır. İncelenen doku ve organizmanın içinde bulunduğu gelişimsel dönemlerin farklılıkları gibi sınırlamalar, olası miRNA’ları in silico olarak bulmak için karmaşık bilgisayar programlarının geliştirilmesine yol açmıştır. Ancak bir genomdaki muhtemel miRNA’ları tahmin etme amacıyla oluşturulan bu programlar, tahminlerini deneysel olarak doğrulamak için yeterli güveni garanti edebilecek kadar hassas ya da kesin olmaktan çok uzaklardır. Bu nedenle, bu proje kapsamında miRNA analizinde daha güvenilir sonuçlar elde etmek için yeni ve daha etkili bir araç geliştirdik. Proje kapsamında geliştirdiğimiz yöntem sayesinde artık miRNA’lar organizmaların genom dizilerinden yüksek güven aralıklarında bulunabilmektedir. MiRNA’ların potansiyel hedeflerini tespit edebilmek için kullanılması planlanan algoritmaların yeterli doğruluk seviyesinde olmadığı denemeler sonucu görüldükten sonra, hedef tahminlemesi için psRNATarget gibi özelleştirilebilir tahmin araçlarının kullanımı tercih edilmiştir. Bu araçlar birlikte kullanarak farklı organizmalarda önemli miRNA etkileşimleri bulunmuştur. VANESA’nın verilerini aldığı DAWIS-M.D.’ye, tüm bilinen miRNA’lar ve bunların hedeflerini içeren bir veritabanı entegre edilmiştir. Böylece, düzenleyici yolakların görselleştirilmesi (örn: KEGG, Reactome) ve miRNA etkileşimleriyle zenginleştirilmesi mümkün hale gelmiştir. Ek olarak, tahmini yapılan miRNA’lar ve hedefleri, yerel olarak VANESA’ya eklenebilmektedir. Bu özellik, kızamık yolaklarının daha iyi anlaşılmasına ve ALS için yeni potansiyel ilaç hedeflerinin tanımlanmasına olanak sağlamıştır.

Anahtar Kelimeler: mikroRNA, miRNA regülasyonu, ağ, miRNA gen tahmini, miRNA hedef tahmini, ağ görselleştirme, yolak analizi

ABSTRACT

MicroRNAs (miRNAs) are single-stranded, small, non-coding RNAs of about 22 nucleotides in length, which control gene expression at the posttranscriptional level through translational inhibition, degradation, adenylation, or destabilization of their target mRNAs. Although hundreds of miRNAs have been identified, many more remain unknown. Therefore, discovery of new miRNA genes is an important step for understanding miRNA mediated post transcrptional regulation mechanisms. It appears that, such biological approaches might be limited in their ability to detect rare miRNAs due to their biased applications. Limitations, such as developmental stage of the organism under examination, have led to the development of sophisticated computational approaches attempting to identify possible miRNAs *in silico*. However, the programs designed to predict possible miRNAs in a genome are not sensitive or accurate enough to warrant sufficient confidence for further investigation. Therefore, to be able to obtain more accurate results, izMiR, a new and more effective tool, was developed within the scope of the project. With this new method, it is now possible to detect miRNAs from the genomes of organisms with high confidence levels. The algorithms that were originally planned to use to detect potential targets of miRNAs were seen to be lacking in accuracy, therefore, highly customizable prediction tools like psRNATarget were preferred for targeting predictions. Important miRNA interactions were identified by using these tools in combination. Such interactions define regulation and can be combined with gene regulatory pathways. A database containing all known miRNAs and their targets was integrated into DAWIS-M.D. from which VANESA draws its data. Thereby, it became possible to visualize regulatory pathways (e.g.: KEGG, Reactome) and enrich them with miRNA interactions. Additionally, predicted miRNAs and targets can be added to VANESA locally. This ability led to a better understanding of the Measels pathway and identification of new potential drug targets in ALS.

Keywords: microRNA, miRNA regulation, network, miRNA gene prediction, miRNA target prediction, network visualization, pathway analysis

1. GİRİŞ

MiRNA’lar pek çok hücresel olayda ve hatta kanser gibi çeşitli hastalık durumlarında düzenleyici olarak görev alan küçük RNA’lardır. Yüzlerce türde binlerce farklı miRNA tespit edilmiş olmasına rağmen, büyük bir çoğunluğu hala bilinmemektedir. Laboratuvar ortamında hiç bir ön bilgi olmadan miRNA keşfi yapmak hem çok vakit alan hem de çok pahalı bir işlemdir. Bu nedenle, şu ana kadar değişik araştırma grupları tarafından farklı özelliklerde miRNA tahmin programları geliştirilmeye çalışılmıştır. Ne yazık ki, bu programlar kesin sonuçlar vermekten oldukça uzaktır. Olgun miRNA dizileri küçük boyutta olduğu ve mRNA içinde yer alan hedef bölgelerine yüzde yüz tamamlayıcı olarak bağlanmadığı için, tamamen yeni miRNA’ların ve hedeflerinin tespiti oldukça zordur ve sadece dizi bilgisi veya yapısal bilgilere dayalı metotlar esas alınarak başarılamaz. Karşılaştırmalı genomik yaklaşımlardan farklı olarak, *ab initio* yaklaşımlar, bilinen dizi benzerliklerine ihtiyaç duymadan türe özgü miRNA'ları keşfedilme olanağı sağlar. Bu nedenle, bu proje kapsamında geliştirilen yeni miRNA tahmin programı ab initio bir methoddur. MiRNA’ların önemi düşünüldüğünde, etkili bir miRNA tahmin programının gerekliliği açıkça anlaşılmaktadır. Bu proje kapsamında, verilen bir RNA ya da DNA (genom) dizisi dışında herhangi bir bilgiye ihtiyaç duymadan bu dizilerde yer alan olası miRNA’ları tahmin edebilmek için doğru sonuçlar verecek etkili bir program geliştirilmiştir. Geliştirilen bu program, Bielefeld Üniversite’sinde tasarlanmış olan VANESA programıyla birlikte kullanılarak, miRNA’lar ile diğer düzenleyici elementlerin arasındaki ilişkiyi görme şansı sağlayan sonuçlar vermiştir. Geliştirilen programda hedef tahmini özelliği, entegrasyonu yapılan hedef tahmini algoritmasının yetersizliği dolasıyla, bulunmamaktadır. Hedefleme tahminin öneminden ötürü, bu tahminler, başka araçlarla sağlanmıştır. Bu proje tamamıyla insan miRNA’ları üzerine tasarlanmış olsa da, oluşturulan yöntemler diğer türlere ait miRNA için de uygulanmış ve yüksek doğruluk ile tahminler yapılmıştır.

2. LİTERATÜR ÖZETİ

2.1 MikroRNA

MikroRNA (miRNA) yaklaşık 18–24 nükleotit uzunluğunda, kodlayıcı özelliği olmayan, hedef mRNA'larının translasyonel inhibisyonu ve istikrarsızlaştırması yoluyla posttranskripsiyonel seviyede gen ekspresyonunu kontrol eden RNA çeşididir (Ambros ve ark. 2003, Filipowicz ve ark. 2008). İlk olarak *Caenorhabditis elegans*'da gelişimsel zamanlamayı modüle eden düzenleyici moleküller olarak keşfedilmişlerdir (Lee ve ark, 1993). Çok hücreli organizmalarda yapılan genetik, biyokimyasal ve bilişimsel çalışmalar, miRNA'ların temel ve çok yönlü rollerini göstermiştir. Ökaryotik türlerin çoğunda binlerce miRNA keşfedilmiş olup, aldıkları çeşitli görevlerde hızla aydınlatılmaktadır (Bushati ve Cohen 2007, Jones-Rhoades ve ark. 2006). Ayrıca miRNA ile, kanser ve nörodejeneratif hastalıklar dahil olmak üzere insan hastalıklar arasındaki bağlantılar da ortaya çıkarılmıştır (Bushati ve Cohen, 2007, Satoh, 2010, Wang ve ark. 2008a). MiRNA biyogenezi ve eylem şekillerini He ve Hannon (He ve Hannon 2004) ve Bartel (Bartel 2004) tarafından yazılmış derleme makalelerde bulmak mümkündür. Hücre içi düzenleyici yolakların önemli bir parçasını oluşturan miRNA'ların tespiti için miRNA gen tahmini ve miRNA hedef tahmini sağlayan bilişimsel metotlar oldukça önemlidir.

1.1 MiRNA Gen Tahmini

MiRNA genlerin belirlenmesi transkripsiyon sonrası gen regülasyonunun anlaşılabilmesi için önemli ve oldukça zorlayıcı bir sorundur (Ding ve ark. 2010). MiRNA genlerinin belirlenmesi için ilk girişimlerin neredeyse tamamı endojen küçük RNA'ların yönlü klonlanmasına ve çok sayıdaki cDNA klonlarının yüksek verimli sıralanmasına (high-throughput sequencing) dayalıdır (Lagos-Quintana ve ark. 2001, Lau ve ark. 2001, Lee ve Ambros 2001, Berezikov ve ark. 2006). Ancak konvansiyonel ileri genetik tarama (conventional forward genetic screening), klonlanmış ürünleri domine eden yüksek miktardaki miRNA'lara karşı yanlı bir yöntemdir (Lagos-Quintana ve ark. 2003). Görünen o ki, bu tarz biyolojik yöntemler nadir miRNA'ların saptanmasında etkisiz kalmaktadır. İncelenen doku ve organizmanın içinde bulunduğu gelişimsel dönemlerin farklılıkları da bu yöntemleri sınırlandıran faktörlerdendir. Bu faktörler ve öncü miRNA (pre-miRNA) dizilerinin ortak bir saç tokası ikincil yapısı paylaşması gibi sebepler muhtemel miRNA'ları tanımlamak için karmaşık bilişimsel yaklaşımların geliştirilmesine yol açmıştır (Berezikov ve ark. 2006). Bilişimsel miRNA gen tahmininde kullanılan yaklaşımlar çeşitli gruplara ayrılabilir. Genel olarak, bir genomdan olası miRNA'ları elde etmek için homoloji modelleme veya *ab initio*

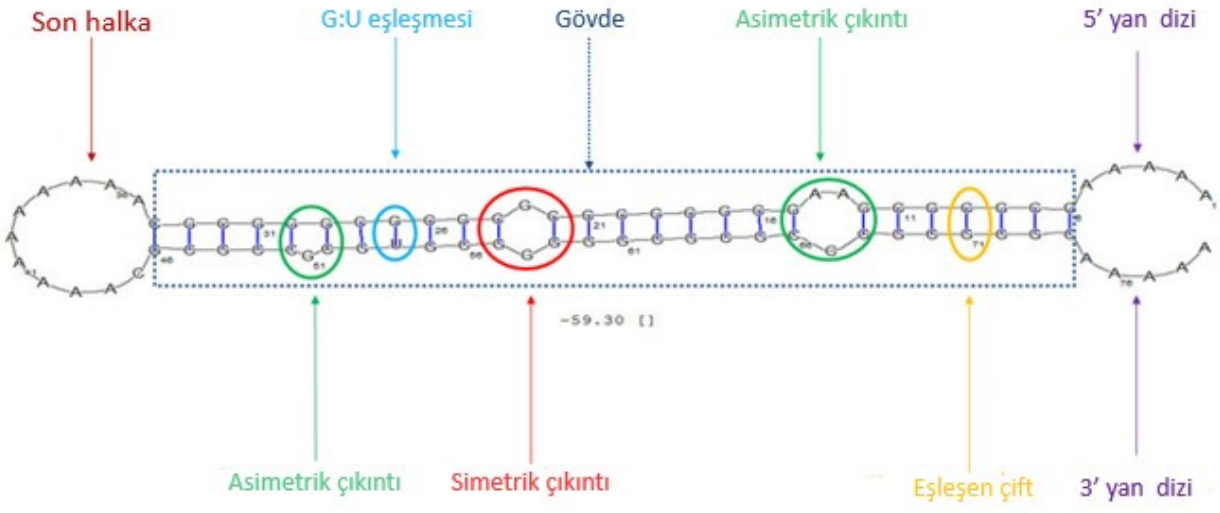
yöntemler uygulanır. Fakat bu yöntemlerde de en azından iki bilişimsel sorun vardır: 1) herhangi bir genomdaki miRNA'ların tahmini ve 2) miRNA'ların muhtemel hedeflerini eşleştirilmesi.

2.2.1 Homolojiye Dayalı miRNA Gen Tahmini

Homoloji modelleme, bir organizmanın genomunda başarılı bir şekilde belirlenmiş olan miRNA bilgisini kullanarak benzer miRNA'ları diğer türlerde bulabilmesine rağmen, tamamen yeni miRNA'ların bu şekilde tespit edilememesi nedeniyle sınırlıdır. Bu yaklaşımdaki ilk girişimlerin büyük bir çoğunluğu yayınlanmış olan pre-miRNA'ların yakın homologlarının tespitine dayanmaktadır (Pasquinelli ve ark. 2000). Bu yöntemde bir çok uyumsuzluklara (mismatch) ve boşluklara (gap) birbirleriyle aralarındaki filogenetik mesafeye bağlı olarak izin verildiğinden, NCBI BlastN ile dizileri hizalamak gibi doğrudan olarak görülebilir (McGinnis ve Madden, 2004). Yoğun bilişimsel çoklu genom hizalamalarına (intensive multiple genome alignments) dayalı türler arası dizi korunumu, yakın türler arasında filogenetik açıdan iyi korunmuş pre-miRNA'ların tüm genom düzeyinde taranması için geliştirilen güçlü bir yaklaşımdır, ancak, özellikle farklı evrimsel mesafelerde bu yaklaşım daha düşük duyarlılık gösterir (Berezikov ve ark. 2005, Boffelli ve ark. 2003). Üstelik, evrimsel olarak korunan saç tokası yapılarını koruyarak, önemli ölçüde farklılık gösteren ya da dizi düzeyinde hızlı evrim geçiren pre-miRNA'ların belirlenmesi de problem yaratabilir. Bir başka önemli konu ise cins-spesifik desenleri olan korunmamış pre-miRNA'ların saptamadan kaçma olasılığıdır. *E.Barr* virüs, *K.sarcomaassociated herpes* virus, *M.herpesvirus* 68, *H.cytomegalo* virus ve *Simian virus* 40'da tespit edilmiş olan patojenik viral kodlanmış pre-miRNA'lar, bilinen konak pre-miRNA'larıyla ve hatta kendi aralarında bile önemli bir dizi benzerliği paylaşmamaktadır (Ng ve Mishra 2007). Homoloji tabanlı haritalama yöntemleri deneysel olarak doğrulanmış miRNA'lar üzerine inşa edilebilir ve böylece ilgili türlerin benzer yapıları ve dizileri bulunabilir. MiRNA'ların çoğu türe özgü olduğundan bu yöntemle aranan yeni miRNA genleri bulunamayacağı için tandem stratejilere ihtiyaç duyulmaktadır. Ayrıca, miRNA genlerinin çok hızlı evrim geçirmesi, homolojiye tabanlı yöntemlerin uygulanabilirliğini sınırlayan faktörlerden bir diğeridir (Liang ve Li 2009). Homoloji bilgisine dayalı miRNA gen tahmini gerçekleştiren programlara dair kapsamlı bilgi Lindow ve Gorodkin (Lindow ve Gorodkin 2007), Sun (Sun ve ark. 2010), Mendes (Mendes ve ark. 2009) ve Li (Li ve ark. 2010) derleme makalelerinden elde edilebilir. Homoloji tabanlı programların eksiklerinden dolayı miRNA gen tahmini için *ab initio* yöntemler geliştirilmiştir.

2.2.2 Ab initio miRNA Gen Tahmini

Homolojiye dayalı yöntemler çoğunlukla karşılaştırmalı genomik bilgi kullanırken, *ab initio* miRNA gen tahmini, eldeki RNA dizisinin gerçek bir miRNA olup olmadığını belirlemek için birincil dizisi dışında hiç bir bilgiye ihtiyaç duymaz. Bu, bilinen hiç bir yakın homoloğu olmayan yeni miRNA'ların tanımlanması için tek yoldur (Brameier ve Wiuf, 2007). MiRNA'lar üzerinde *ab initio* metotları kullanarak çalışırken karşılaşılan en büyük zorluk, verilen bir dizinin özelliklerine göre miRNA olup olmadığını ayırt etmekte kullanılacak uygun parametreleri seçmektir. Seçilen parametre iyi bir hassasiyet veya doğruluk potansiyeline sahip değilse çok bilgilendirici olmayacağı gibi yanlış pozitif sonuçlar üretme potansiyeline sahip olabilir. Örneğin, saç tokası yapısı (Şekil 1) miPred gibi miRNA tahmin araçlarında yaygın olarak kullanılan özelliklerden birisidir (Ng ve Mishra 2007). Öncül miRNA'lar olgun miRNA biyogenezinin erken dönemleri için çok önemli olan ve evrimsel olarak korunmuş RNA saç tokası yapısına sahip olmalıdır ancak bu yapıya dayalı miRNA tahmini gerçekleştirmek istendiğinde bazı sorunlar ortaya çıkmıştır; a) saç tokası yapısı miRNA'lara özgü değildir (Ding ve ark. 2010), b) insan genomunda yaklaşık olarak 11 milyon saç tokası yapısının bulunabileceği gösterilmiştir (Bentwich 2008).



Şekil 1. MiRNA saç tokası yapısı. Saç tokası yapısının bazı elemanları: G-U baz-eşlenmesi, asimetreik ve simetrik çıkıntılar, gövde (stem) vb. İkincil RNA yapısı ve minimum serbest enerji değeri (-59.30) RNAShapes (Steffen ve ark. 2006) kullanılarak hesaplanmıştır.

Tüm mevcut miRNA tahmin yöntemleri, RNAfold (Hofacker 2003) veya mfold (Zuker 2003) gibi verilen bir ya da daha fazla RNA dizisinin ikincil yapısını *ab initio* tahmin etmek için

geliştirilmiş, enerji minimizasyonu tabanlı algoritmaların ürettiği saç tokası yapısı tahminlerine dayalıdır.

MiRNA gen tahmininde *ab initio* metotların kullanımı farklı avantajlar sağlar:

a) İnsan genomunda bulunan miRNA sayısının daha önce inanılandan daha büyük olduğu ve bu miRNA'ların türler arası korunmuş dizilerle sınırlı olmadığı gösterilmiştir (Bentwich ve ark. 2005).
b) *Ab initio* tahmin yöntemleri kullanılarak, karşılaştırmalı dizi analizi yöntemine ya da korunmuş bir dizi veya yapıya ihtiyaç duymadan, bir genomdaki miRNA'lar tahmin edebilir (Brameier ve Wiuf, 2007). Bu sayede, herhangi bir yakın homoloğu bilinmeyen, tamamen yeni miRNA'lar belirlenebilir.
c) Düşük oranda ekspresyonu yapılan insan miRNA genlerinin daha hızlı evrim geçirdiği gösterilmiştir (Liang ve Li 2009). Bu durumda karşılaştırmalı genomik bilgiye dayalı metotlar işlevsiz kalırken, *ab initio* yaklaşım kullanılarak bu tarz miRNA'lar da sorunsuzca bulunabilir.

MiRNA çalışmalarının çoğunluğunda, miRNA tahmin programları tarafından bulunan sonuçlar daha sonra laboratuvar ortamında kullanılır. Eğer kullanılan miRNA tahmin programının duyarlılık ve özgüllük değerleri %80 ya da daha düşük ise, insan gibi büyük genoma sahip organizmalarda bu durum ciddi sorunlara yol açar. Oluşan büyük hata payı, laboratuvar ortamında miRNA tahmin programı kullanılarak elde edilen sonuçların denenmesini neredeyse imkansız hale getirir. Bu nedenle, hassas ve doğruluk değeri yüksek bir miRNA tahmin programı geliştirilmesi son derece önemlidir. *Ab initio* miRNA gen tahmini gerçekleştiren programlara dair kapsamlı bilgi Lindow ve Gorodkin (Lindow ve Gorodkin 2007), Sun (Sun ve ark. 2010), Mendes (Mendes ve ark. 2009) ve Li (Li ve ark. 2010) derleme makalelerinden elde edilebilir. MiRNA çalışmalarında miRNA genlerinin bulunmasından sonraki adım bulunan miRNA'ların hedeflerinin tespitidir.

1.2 MiRNA Hedef Tahmini

Daha çok sisteme odaklı yaklaşımlarda tahmin edilen olgun miRNA'ların hedeflerine bakılarak doğruluğu onaylanabilir. Örneğin mRNA'ların 3'UTR'ları (UnTranslated Region – Translasyonu yapılmayan bölge) birçok miRNA için potansiyel hedefler içerir ve miRNA başına hedefleri ya da regüle edilen mRNA başına hedef alanların çokluğu değerlendirilebilir (Liu ve ark. 2010). MiRNA'lara dair daha iyi bir anlayış sahibi olabilmek ve onları muhtemel tedavilerde kullanabilmek için, ilk olarak miRNA'ları genomda etkili bir şekilde bulabilen bir yöntem belirlemek ve sonra bu miRNA'ların potansiyel hedeflerini ve eylem biçimlerini belirleyebilmek

gerekmektedir. Farklı grupların tasarladığı birçok miRNA tahmini algoritmaları olmasına rağmen, bu tür programların verimi çok düşüktür (Brodersen ve Voinnet 2009, Lindow ve Gorodkin 2007). Bu performans düşüklüğünün başlıca nedenlerinden biri, miRNA – hedef ilişkilerinin ardında yatan mekanizmaların tam anlamıyla bilinmemesidir.

Hedef tahmini için ilk adım, hedef araması yapılacak dizileri elde etmektir. Hayvan miRNA hedeflerinin büyük bir çoğunluğu UTR'larda yer aldığı için cDNA dizi bilgilerini kullanmak arama alanını azaltacağından, geliştirilecek olan algoritmanın hızına olumlu bir katkı yapacaktır (Lindow ve Gorodkin 2007). Hayvan miRNA hedef tahmini için temel ilke, miRNA'ların baz eşleşmesi yapabileceği UTR segmentlerini tespit etmektir. Bulunan eşleşmelerin gerçekten miRNA hedefleri olup olmadığını değerlendirmek için farklı yaklaşımlar kullanılabilir. Miranda (Enright ve ark 2003) üç temel özelliğe dayanarak hedef seçer: (i) dizi hizalama, (ii) miRNA ve mRNA arasındaki hibridizasyon enerjisinin belirlenmesi ve (iii) ilgili genomlardaki dizi korunması. Diğer bir yaklaşım ise, kuralları baştan belirleyip bu kurallara uyan hedefleri aramaktansa, kuralları otomatik olarak örneklerden öğrenmeye dayalı olan makine öğrenimi metodudur. Bu yaklaşımda, onaylanmış hedef – miRNA etkileşimlerinin negatif ve pozitif örneklerinden yararlanılarak hedef kuralları belirlenir (Lindow ve Gorodkin 2007). Hedef-miRNA etkileşimi, yapısal, termodinamik veya pozisyonel özellikler kullanılarak değerlendirilir. Bu özellikler dışında, mevcut programlarda kullanılmayan ama miRNA hedef tahmininde kullanılabilecek başka parametreler de geliştirilecek olan programlara entegre edilebilir (Allmer, 2011).

MiRNA hedef tahmini gerçekleştiren programlara dair kapsamlı bilgi Lindow ve Gorodkin (Lindow ve Gorodkin 2007), Sun (Sun ve ark. 2010), Mendes (Mendes ve ark. 2009) ve Li (Li ve ark. 2010) derleme makalelerinden elde edilebilir. MiRNA ve miRNA hedef tahmini programları geliştirilirken en önemli aşamalardan birini, geliştirilen programı uygun veri setleri üzerinde test etmek oluşturur. Bu tarz programlarda yaygın olarak kullanılan makine öğrenimi yaklaşımları için de doğru veri setlerinin kullanılması son derece önemlidir.

1.3 Veri Setleri ve Makine Öğrenimi

Yeni miRNA'ların tahmini için, pre-miRNA'ların birincil dizi bilgisi ve katlanma yapısını kullanan pek çok algoritma geliştirilmiştir (Ritchie ve ark. 2012, Xue ve ark. 2005). Bu algoritmalar, yaygın olarak kullanılan veritabanlarında mevcut olmayan, doku ya da türlere özgü miRNA'ları bulmak için çok önemlidir. Bu tahmin araçlarını değerlendirmek ve yeni algoritmaların oluşturulmasına yardımcı olmak için, yüksek güven değerlerine sahip pozitif ve negatif kontrol veri setleri tanımlanmalı ve kullanılmalıdır (Ding ve ark. 2010, Ritchie ve ark. 2012). Bu veriler algoritma

geliştirilmesinde kullanılabileceği gibi, miRNA'ların ayırt edici özelliklerinin tanımlanması için de kullanılabilir.

MiRNA çalışmalarında en sık kullanılan pozitif referans veri seti miRBase veritabanında bulunmaktadır (Kozomara ve Griffiths-Jones 2010) ancak daha önceki çalışmalarımızda bu veritabanından elde edilen pozitif veri setinin güvenilirliğinin yeterli olmadığını göstermiştik (Saçar ve ark, 2013). Negatif veri seti oluşturmak oldukça zor bir işlemdir. İnsan genomunda bulunan miRNA'ların sayısı kesin olarak bilinmediğinden, saç tokası yapısı gösteren dizileri rastgele seçip bu amaç için kullanmak için uygun değildir (Ding ve ark. 2010). Yeni miRNA'ların tahmininde yüksek spesifite elde edebilmek için negatif örneklerin miRNA'lara mümkün olduğunca benzer olması gereklidir (Ding ve ark. 2010).

Uygun negatif veri kümesi seçimi iyi eğitilmiş bir makine öğrenimi (machine learning) sınıflandırıcı için önemlidir (Wu ve ark. 2011). MiRNA çalışmalarında SVM (Support Vector Machine) (Tyagi ve Prasad 2012) ya da HMM (Hidden Markov Model) (Agarwal ve ark. 2010) gibi makine öğrenimi yöntemleri sıklıkla kullanılmaktadır. Ancak kullanılan veri setindeki diziler çok yapay ise, (örneğin tamamen rastgele bir şekilde hazırlanmış), SVM gibi makine öğrenimine dayalı yaklaşımların, gerçek biyolojik dizilerin farklı kategorilerini ayırt edebilecek kadar iyi eğitilmemiş olma riski vardır (Yousef ve ark. 2008, Wu ve ark. 2011). Aksine, negatif veri seti ve pozitif veri seti birbirine çok benzer ise, SVM'in bu iki veri arasında yeterli derecede bir ayrım gerçekleştirebilmek için bir yol bulması mümkün olmayacaktır (Wu ve ark. 2011). Negatif ve pozitif veri setine ihtiyaç duyan 2-sınıflı (2-class classifier) makine öğrenimi algoritmalarının yanı sıra, negatif veri seti ile ilgili problemlerden kurtulmak için, yalnızca pozitif veri setine dayalı 1-sınıflı (1-class classifier) yaklaşımlar da miRNA tahmin çalışmalarında kullanılmıştır (Yousef ve ark. 2011).

MiRNA ve miRNA hedef tahmini programlarında makine öğrenimine dair kapsamlı bilgi Lindow ve Gorodkin (Lindow ve Gorodkin 2007), Sun (Sun ve ark. 2010) ve Mendes (Mendes ve ark. 2009) derleme makalelerinden elde edilebilir.

3. GEREÇ VE YÖNTEM

1.4 Program Geliştirme Ortamı

Bu projede yer alan yazılım, nesneye dayalı en yaygın programlama dillerinden biri olan JAVA ile yazılmıştır. Java Development Kit Standart Sürümü 7 (JDK SE 7) işletim sistemine yüklüdür. Programın kodlarının yazımı, en iyi geliştirme ortamı ödülünü de içeren çeşitli dallarda ödüller kazanmış bir yazılım olan NetBeans IDE – 7u5 versiyonu kullanılmıştır. Yazılan her kod için, NetBeans IDE ortamında JUnit 4x test kodları yazılarak, programın planlandığı gibi çalışıp çalışmadığı test edilmiştir.

Program, Java Runtime Environment (JRE)'nin kurulu olduğu her ortamda, çalışılan platformdan bağımsız (Windows, Linux, MacOS) olarak çalışabilmektedir. Ayrıca, program geliştirme aşamasında bitbucket (https://bitbucket.org/) ve GIT (http://git-scm.com/) gibi kaynak kodu barındırma imkanı sağlayan programlar ortak kod geliştirme (collaborative code development) ve sürüm oluşturmak için kullanılmıştır.

1.5 MikroRNA Tahmini

Ön çalışmalarımızda daha önce geliştirilmiş olan ab initio miRNA tahmin programlarından bazılarını inceleyerek kullandıkları parametreleri, doğruluk ve hassasiyet değerlerini hesaplamıştık. Literatürde yer alan miRNA tahmin programlarının hiçbiri kullanılabilir bir program (software) sağlamadığı için söz konusu çalışmalarda kullanılan parametrelerin her birini JAVA ortamında kendimiz programlayarak çeşitli veri setleri üzerinde denedik (Tablo 1). Yeni geliştirdiğimiz parametrelerle birlikte yaklaşık olarak 800 parametre bu proje dâhilinde JAVA ortamında programlanmış ve test kodları yazılmıştır.

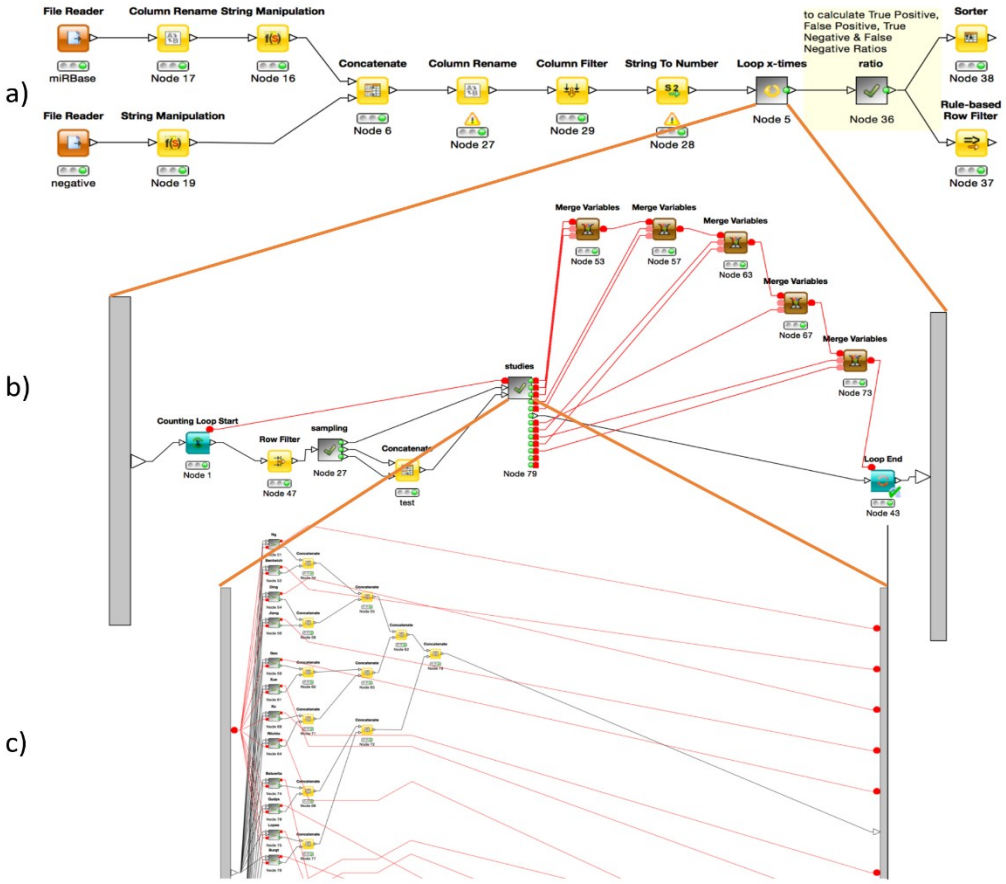
Tablo 1. "Computational Prediction of MicroRNAs from Toxoplasma gondii Potentially Regulating the Hosts' Gene Expression" başlıklı çalışmamızda kullanılan 32 parametre ve açıklamaları.

Acronym	Type	Definition/Explanation	Reference
bpd/hpl	Structure	Base pairing distance normalized to hairpin length	Ding et al. 2010
st(G-C)/hpl	Structure	The number of G-C bonds in stem normalized to hairpin length	Unpublished
l(lsr)/hpl	Structure	Length calculated over the longest symmetrical region (lsr) of the stem loop normalized to hairpin length	Sewer et al. 2005
bpp/hpl	Structure	Base pairing propensity normalized to hairpin length	Ding et al. 2010
Sl	Structure	Stem length	Burgt et al. 2009

Hpl	Structure	Hairpin length	Bentwich 2008
#nial_h/hpl	Structure	Number of nucleotides in asymmetrical loops computed over the entire hairpin structure normalized to hairpin length	Sewer et al. 2005
lscm/nl	Structure	Longest continious stretch of matches in the hairpin divided by number of loops	Burgt et al. 2009
adal/hpl	Structure	Absolute length difference between both arms normalized to hairpin length	Burgt et al. 2009
#gih/saln	Structure	Total gaps in alignment divided by stem alignment length	Burgt et al. 2009
st(A-U)/sl	Structure	The number of A-U bonds in stem normalized to stem length	Unpublished
mwm/sl	Structure	Highest number of matches in 24 positions in the stem alignment string of the hairpin structure normalized to stem length	Nam et al. 2005
saln/hpl	Structure	Stem alignment length normalized to hairpin length	Unpublished
mbs/sl	Structure	Maximum bulge size normalized to stem length	Bentwich 2008
hpmfe_rf/sl	Thermodynamic/ structure	Hairpin minimum free energy calculated by rnafold normalized to stem length	Jiang et al. 2007
hpmfe_rf_I1	Thermodynamic/ structure	Hairpin minimum free energy index 1	Zhang 2006
hpmfe_rf/hpl	Thermodynamic/ structure	Hairpin minimum free energy calculated by rnafold normalized to hairpin length	Bentwich 2008
dH/sl	Thermodynamic	Enthalpie normalized to stem length (calculated using Mfold)	Ding et al. 2010
dS/sl	Thermodynamic	Entropy normalized to stem length (calculated using Mfold)	Ding et al. 2010
Q	Thermodynamic	Shannon Entropy (calculated using RNAFold)	Nam et al. 2005
Tm	Thermodynamic	Melting temperature (calculated using Mfold)	Ding et al. 2010
dG/sl	Thermodynamic	Gibbs free energy normalized to stem length	Ding et al. 2010
Efe	Thermodynamic	Ensemble Free Energy (provided by RNAFold)	Ding et al. 2010
Efq	Thermodynamic	Ensemble Frequency (provided by RNAFold)	Ding et al. 2010
Ediv	Thermodynamic	Ensemble Diversity (provided by RNAFold)	Ding et al. 2010
dscs/nl	Statistical	Dinucleotide sequence complexity score measuring the relative occurrence of the two most occurring dinucleotides normalized to number of loops	Burgt et al. 2009
lsr(%bp)/hpl	Structure/ sequence	Proportion of base pairs in the longest region without any asymmetrical loop normalized by hairpin length	Sewer et al. 2005
#C.../sl	Structure/ sequence	Proportion of "...", "(((", "(.("… triplet-strcutures normalized to stem length	Xue et al. 2005
#A..(/sl	Structure/ sequence	Proportion of "...", "(((", "(.("… triplet-strcutures normalized to stem length	Xue et al. 2005
#U(((/sl	Structure/ sequence	Proportion of "...", "(((", "(.("… triplet-strcutures normalized to stem length	Xue et al. 2005

#G(../sl	Structure/ sequence	Proportion of "…", "(((", "(.("… triplet-structures normalized to stem length	Xue et al. 2005
%C++%G	Sequence	Sum of G and C nucleotide frequencies	Ng & Mishra 2007

Literatürde yer alan makine öğrenimine dayalı ab initio miRNA tahmini sağlayan makalelerde bulunan tüm parametreler tarafımızdan programlanmış, bu parametreler çeşitli veri setleri üzerinde denenerek güvenilir bir karşılaştırma sağlanmaya çalışılmıştır. Şekil 2'de bu işlem için Konstanz Information Miner (KNIME, http://www.knime.org/) kullanılarak oluşturulan iş akışı (workflow) görülebilir.

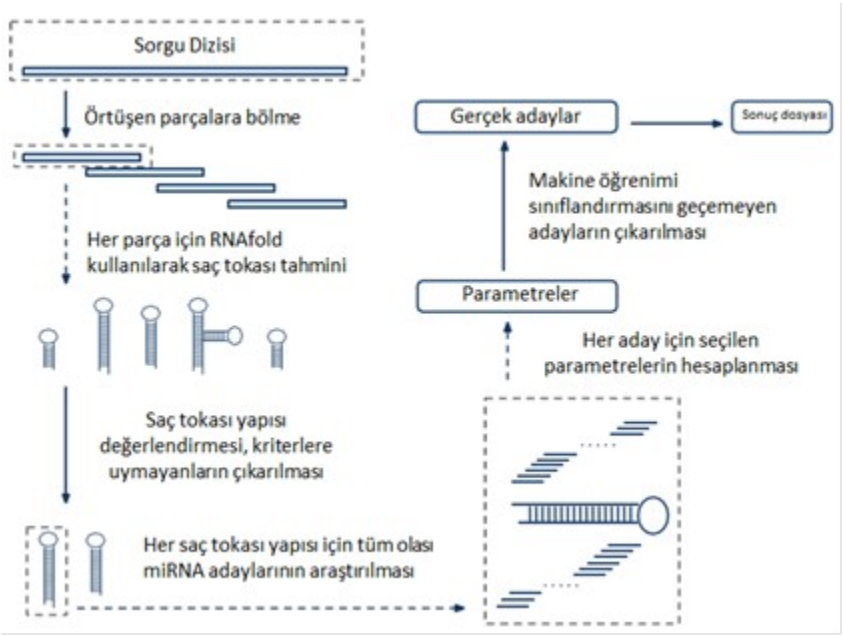


Şekil 2. 12 miRNA tahminine dayalı makalenin karşılaştırması için hazırlanan KNIME iş akışı. a)genel şema, b) 1000 kez tekrar edilen bölüm, c) her bir çalışma için hazırlanan metotlar görülmektedir. Not: Şekillerde yer alan turuncu çizgiler iş akışı hiyerarşisini göstermektedir. A'da yer alan loop x-times'ın içeriği B'de, B'de bulunan studies ise C'de geniş haliyle gösterilmiştir. Bu şekillerdeki en alt seviye olan C bile daha alt seviyelerde pek çok ağdan (node) oluşmaktadır ve daha da genişletilebilir.

Ayrıca, KNIME kullanılarak, oluşturulan 750 parametrenin bilgi kazanım değerleri (information gain) ve bağıntı analizleri yapılmıştır.

1.6 MikroRNA Hedef Tahmini

Bilinen ya da bizim tarafımızdan tahmin edilen miRNA öncü yapılarının hedeflerini bulmak için öncelikle saç tokası yapısı içerisinde bulunan olgun (mature) miRNAların belirlenmesi gereklidir (Şekil 3). Bu işlem için, iki sınıf sınıflandırma metoduyla Random Forest sınıflandırıcısı kullanılarak bilinen olgun miRNAlar pozitif veri olarak kullanılmış ve her bir pozitif değer için alternatif negatif değer yaratılmıştır. Elde edilen model, tahmin edilen saç tokası yapıları üzerinde uygulanmış ve her bir olgun miRNA için BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) ve RNAHybrid (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) programları kullanılarak olası tahminler hesaplanmıştır. İlk aşamada hedef veritabanı olarak insan 3' UTR (untranslated region) kullanılmıştır.



Şekil 3. Saç tokası yapısı ve olgun miRNA tahmini işlemlerinde kullanılan iş akışının sadeleştirilmiş hali.

Bunların yanı sıra, makine öğrenimi kullanılması için literatür araştırması yapılmış ve makine öğrenmesine dayalı hedef tahmini yapan araçların kullandıkları parametreler incelenmiştir ve benzer parametreler kendi içinde gruplandırılmıştır (Tablo 2).

Tablo 2. Hedefleme parametreleri grupları ve hangi çalışmalarda kullanıldıkları.

Authors	Seed Region	Conservation of Sequence	Sequence Context	Thermodynamic Features	Site Access
Ding et al. (2016)	✓	✓	✓	✓	✓
Ghoshal et al. (2015)	✓	✓	✓	✓	✓
Karathanou et al. (2015)	✓	✓	✓	✓	✓
Mousavi et al. (2015)	✓	✓	✓	✓	✓
Rabiee-Ghahfarrokhi et al. (2015)	✓	✓	✓	✓	✓
Varghese et al. (2015)	✓	✓	✓	✓	✓
Wang (2016)	✓	✓	✓	✓	✓
Lu & Leslie (2016)	✓	✓	✓	✓	✓

Yapılan incelemelerde, daha eski hedefleme tahmini yapan araçların kullandığı parametrelerin, yeni araçlar tarafından da kullanıldığı ve yeni araçlarda bu parametrelere ekleme yapıldığı görülmüştür. Bu yüzden, yakın zamanda çıkarılmış araçların makalelerinde anlatılan parametrelere öncelik verilmiştir. Parametrelerin tam listesi, Tablo 3'te görülebilir.

Tablo 3. Hedefleme parametreleri. Farklı çalışmalarda kullanılan hedefleme tahminine dayalı parametrelerin tamamı bir tabloda özetlenmiştir. Tabloda görülebileceği üzere, tablonun çok uzun olmaması adına, aynı hesaplama çeşidine ait parametreler gruplanarak belirtilmiştir (örn. A ve G miktarı farklı iki parametre olmasına rağmen aynı maddede bulunmaktadır).

Features	
Total Seed Match	Match in 8th position
Total number of matches in the seed region	Number of bulges
A, G, U and C Content	Number of loops
AU Content	Number of Stems
GC Matches	Max Stem Length
AU Matches	Max Loop Length
Total Bases / Total Length	Max number of consecutive free bases on target
Dinucleotide frequencies	Positional Score

Target Site Length	Out of seed score
Length of the largest consecutive pairs	Total Mismatches
Position of the largest consecutive pairs relative to the miRNA 5'	Target Site Location
Length of the largest consecutive pairs allowing 2 mismatches	2mers Count
Position of the largest consecutive pairs allowing 2 mismatches	Target site > 800 n.t. to UTR end
Number of matches at the miRNA 3' end	Bases in position 1,9,10,11,12,13 on target site
Total number of matches in the seed region and the miRNA 3' end	Binary representation for each pair
Difference between the number of matches in the seed region and that in the miRNA 3' end	Folding Energy
Exon preference	Accessibility
Pair type / Length	Target site <200 n.t. to UTR end
GC content	Stem Conservation
Total match	Flanking conservation
GU Matches	Difference between stem and flanking conservation
m/e motif	

1.7 Veritabanı

Makine öğreniminde kullanılması ve miRNA'ların etkileşimlerinin daha kolay ve eksiksiz görülebilmesi adına miRBase, miRTarBase ve TarBase veritabanları, tüm mevcut organizmalar için indirilmiştir. Elde edilen verilerin birbirleriyle bağlantıları, yeni bir MySQL veritabanında sağlanmıştır. Bu yeni veritabanına erişimin daha kullanıcı dostu olması amacıyla erişim, phpmyadmin aracıyla sağlanmıştır.

1.8 Ağ analizi

MiRNA'ların genleri transkripsiyon sonrası düzenlemelerine bağlı olarak, mirna'ların genlerle olan etkileşimlerini ağ şeklinde incelemek mümkündür. Bunun için, geliştirilmiş olan mirna tahmini ve mirna hedef tahmini gereklidir. Yaptığımız *Toxoplasma gondii* çalışmasında, proje kapsamında, genom dizisinden mirna tahmini yapmamıza olanak sağlayan tahmin aracımız kullanılarak belirli güven aralığında olan miRNA'lar ve olası gen hedefleri, psRNATarget

kullanarak belirlenmiştir. Bu çıkarımlarımızı, gen ekspresyonu analizimizle birleştirerek bir miRNA düzenleyici ağı oluşturulmuştur. Ağ oluşumunda, *Toxoplasma gondii* gen yolaklarının henüz belirlenmemiş olmasından ötürü VANESA kullanılamamış, bunun yerine Cytoscape adlı araç kullanılarak gerçekleştirilmiştir.

2. BULGULAR

2.1 MikroRNA Tahmini

MiRNA tahmini için incelenen çalışmalar (Tablo 4) doğruluk, hassasiyet, Youden's index, F-measure gibi değerler hesaplanarak karşılaştırılmıştır (Tablo 5, Şekil 4).

Tablo 4. İncelenen tüm çalışmaların karşılaştırmalı listesi (Saçar Demirci, Baumbach, and Allmer 2017).

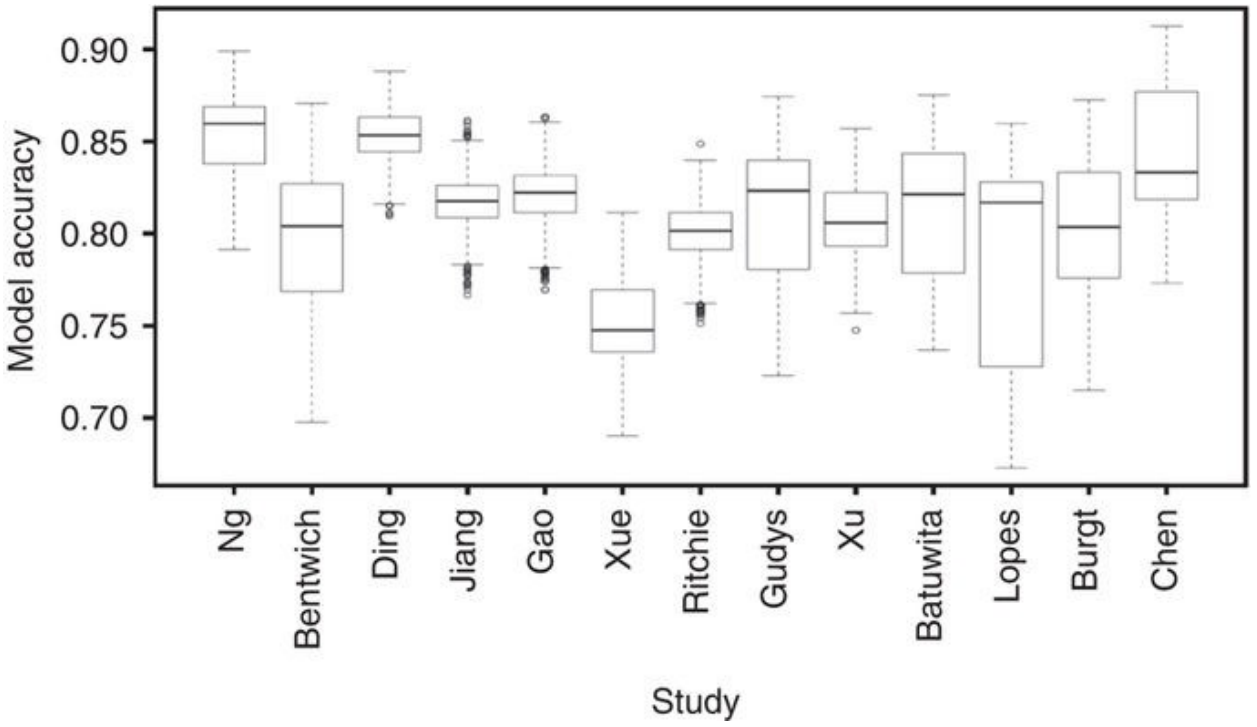
Table 1 Available pre-miRNA detection tools

Study	ML algorithm	Feature number	Positive data	Negative data	Sampling	Implementation	Number of citations (Google Scholar)
Xue[46]	SVM	32	MiRBase 5.0	CODING dataset (Pseudo)	Random selection (approx. 1:1 positive negative ratio)	*	412 (34)
Jiang[47]	RF, SVM	34	MiRBase 8.2	pseudo	Random sampling (approx. 1:1 positive negative and 1:1.5 training testing ratio)	*	376 (48)
Ng[37]	SVM	29	MiRBase 8.2	pseudo	Random selection without replacement (1:2 positive negative ratio)	*	203 (19)
Batuwita[48]	SVM	21	MiRBase 12	pseudo & Human other ncRNAs	Outer-5-fold-cv	+	172 (16)
Xu[49]	A novel ranking algorithm based on random walks & SVM	35	MiRBase (September 1, 2007)	Random, non-overlapping 90nt fragments from the human genome	Random selection (1:2 positive to negative ratio)	*	80 (4)
Ding[50]	SVM	32	Known miRNAs	UTRdb & ncRNA from Rfam 9.1	Outer 3-fold cross-validation	–	61 (11)
Chen[41]	LibSVM	99	miRBase (2013)	pseudo & Zou	Leave-one-out	+	31 (24)
Burgt[51]	L score classifier	18	non-plant miRNA hairpin sequences (miRBase version 9.0)	-	10-fold cross-validation	*	31 (4)
Gudys[40]	NB, MLP, SVM, RF, APLSC	28	MiRBase 17	From genomes and mRNAs of ten animal and seven plant species as well as 29 viruses	Stratified 10-fold CV	+	27 (5)
Ritchie[52]	SVM	36	Murine miRBase v17	Transcripts without evidence of processing by Dicer	-	–	20 (5)
Bentwich[53]	-	26	Hairpins from Human Genome	10000 hairpins found in non-coding regions		–	20 (2)
Lopes[54]	SVM, RF, G^2 DE	13	MiRBase 19	pseudo	Non-standard training and testing scheme.	*	16 (6)
Gao[55]	SVM	57	MiRBase v20	Exonic regions of our some available genomes and ncRNAs from rFam	1:1 positive to negative ratio	*	11 (1)

SVM support vector machine, *NB* Naïve Bayes, *MLP* Multi-Layered Perceptron, *RF* Random Forest, *APLSC* Asymmetric Partial Least Squares Classification, *G2DE* Generalized Gaussian Density Estimator, + implementation exists, – no implementation, * experienced problems with the implementation
Previously published studies performing ab initio pre-miRNA detection using machine learning (ML). Listed are the number of features that were effectively used, the training data that was employed and whether an implementation is available
The negative data (see Online Methods) "pseudo" was generated by Xue[46] but downloaded from Ng[37]. The Table is sorted by the number of citations in Google Scholar (please note that there is a relationship between year of publication and number of citations, therefore, the number of citations in 2016 is provided in parentheses, as well)

Tablo 5. İncelenen tüm çalışmaların farklı değerlere göre karşılaştırılması. 1000 tekrar içerisinden sadece ilk tekrarın (iteration) sonuçları gösterilmiştir.

Study	Classif...	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accur...	Cohen...	Iteration	Youde...	p+	p-
Ng	DT	484	85	484	85	0.882	0.877	0.882	0.882	0.866	0.863	0.727	0	0.763	7.446	0.134
Ng	SVM	498	87	498	87	0.907	0.901	0.907	0.907	0.878	0.874	0.749	0	0.814	9.765	0.102
Ng	NB	466	102	466	102	0.849	0.843	0.849	0.849	0.834	0.832	0.663	0	0.698	5.614	0.178
Bentwich	DT	459	107	459	107	0.836	0.831	0.836	0.836	0.823	0.821	0.641	0	0.672	5.1	0.196
Bentwich	SVM	502	228	502	228	0.914	0.872	0.914	0.914	0.785	0.75	0.499	0	0.829	10.681	0.094
Bentwich	NB	507	150	507	150	0.923	0.905	0.923	0.923	0.841	0.825	0.65	0	0.847	12.071	0.083
Ding	DT	470	114	470	114	0.856	0.846	0.856	0.856	0.83	0.824	0.648	0	0.712	5.949	0.168
Ding	SVM	464	87	464	87	0.845	0.845	0.845	0.845	0.844	0.843	0.687	0	0.69	5.459	0.183
Ding	NB	491	81	491	81	0.894	0.89	0.894	0.894	0.876	0.873	0.747	0	0.789	8.466	0.118
Jiang	DT	462	95	462	95	0.842	0.839	0.842	0.842	0.835	0.834	0.668	0	0.683	5.31	0.188
Jiang	SVM	462	124	462	124	0.842	0.83	0.842	0.842	0.814	0.808	0.616	0	0.683	5.31	0.188
Jiang	NB	518	163	518	163	0.944	0.926	0.944	0.944	0.842	0.823	0.647	0	0.887	16.71	0.06
Gao	DT	439	111	439	111	0.8	0.799	0.8	0.8	0.799	0.799	0.597	0	0.599	3.991	0.251
Gao	SVM	463	91	463	91	0.843	0.842	0.843	0.843	0.84	0.839	0.678	0	0.687	5.384	0.186
Gao	NB	507	135	507	135	0.923	0.908	0.923	0.923	0.851	0.839	0.678	0	0.847	12.071	0.083
Xue	DT	410	152	410	152	0.747	0.741	0.747	0.747	0.738	0.735	0.47	0	0.494	2.95	0.339
Xue	SVM	537	278	537	278	0.978	0.958	0.978	0.978	0.787	0.736	0.472	0	0.956	44.75	0.022
Xue	NB	516	205	516	205	0.94	0.912	0.94	0.94	0.813	0.783	0.566	0	0.88	15.636	0.064
Ritchie	DT	433	126	433	126	0.789	0.785	0.789	0.789	0.782	0.78	0.559	0	0.577	3.733	0.268
Ritchie	SVM	448	106	448	106	0.816	0.814	0.816	0.816	0.812	0.811	0.623	0	0.632	4.436	0.225
Ritchie	NB	510	164	510	164	0.929	0.908	0.929	0.929	0.834	0.815	0.63	0	0.858	13.077	0.076
Gudys	DT	460	113	460	113	0.838	0.83	0.838	0.838	0.82	0.816	0.632	0	0.676	5.169	0.193
Gudys	SVM	452	152	452	152	0.823	0.804	0.823	0.823	0.784	0.773	0.546	0	0.647	4.66	0.215
Gudys	NB	487	99	487	99	0.887	0.879	0.887	0.887	0.858	0.853	0.707	0	0.774	7.855	0.127
Xu	DT	438	116	438	116	0.798	0.796	0.798	0.798	0.794	0.793	0.587	0	0.596	3.946	0.253
Xu	SVM	497	131	497	131	0.905	0.889	0.905	0.905	0.845	0.833	0.667	0	0.811	9.558	0.105
Xu	NB	507	167	507	167	0.923	0.901	0.923	0.923	0.829	0.81	0.619	0	0.847	12.071	0.083
Batuwita	DT	461	102	461	102	0.84	0.836	0.84	0.84	0.829	0.827	0.654	0	0.679	5.239	0.191
Batuwita	SVM	440	147	440	147	0.801	0.787	0.801	0.801	0.775	0.767	0.534	0	0.603	4.037	0.248
Batuwita	NB	484	92	484	92	0.882	0.875	0.882	0.882	0.86	0.857	0.714	0	0.763	7.446	0.134
Lopes	DT	461	107	461	107	0.84	0.834	0.84	0.84	0.825	0.822	0.645	0	0.679	5.239	0.191
Lopes	SVM	447	206	447	206	0.814	0.771	0.814	0.814	0.744	0.719	0.439	0	0.628	4.382	0.228
Lopes	NB	483	123	483	123	0.88	0.866	0.88	0.88	0.836	0.828	0.656	0	0.76	7.318	0.137
Burgt	DT	441	116	441	116	0.803	0.8	0.803	0.803	0.797	0.796	0.592	0	0.607	4.083	0.245
Burgt	SVM	466	177	466	177	0.849	0.818	0.849	0.849	0.782	0.763	0.526	0	0.698	5.614	0.178
Burgt	NB	469	82	469	82	0.854	0.854	0.854	0.854	0.853	0.852	0.705	0	0.709	5.862	0.171
Chen	DT	451	101	451	101	0.821	0.821	0.821	0.821	0.819	0.819	0.638	0	0.643	4.602	0.217
Chen	SVM	507	79	507	79	0.923	0.918	0.923	0.923	0.893	0.89	0.78	0	0.847	12.071	0.083
Chen	NB	490	151	490	151	0.893	0.871	0.893	0.893	0.824	0.809	0.617	0	0.785	8.305	0.12



Şekil 4. İncelenen çalışmaların toplam doğruluk değerlerinin karşılaştırması (Saçar Demirci, Baumbach, and Allmer 2017).

Çalışmalarımızda elde ettiğimiz sonuçlara göre miRNA tahmininde veri setlerinin kalitesi ve güvenilirlik derecesi, tahmin performansını önemli ölçüde etkilemektedir (Saçar Demirci and Allmer 2017). Bu sebeple, daha başarılı ve kapsamlı bir miRNA tahmin yöntemi geliştirdiğimizi gösterebilmek için mevcut tüm veri setlerine ek olarak tamamıyla yeni ve farklı pozitif ve negatif veri setleri tarafımızdan oluşturulmuştur (Tablo 6).

Tablo 6. Kullanılan tüm mevcut ve yeni veri setlerinin türleri, boyutları, özellikleri ve kaynaklarının listesi (Saçar Demirci, Baumbach, and Allmer 2017).

Table 2 Data sets

Dataset	Type	Size	Property	Source
hsa	Positive	1881	All human miRNAs in miRBase	http://www.mirbase.org
mirbase	Positive	28596	All miRNAs available in miRBase	http://www.mirbase.org
mmu	Positive	1193	All mouse miRNAs in miRBase	http://www.mirbase.org
mmu*	Positive	380	Mouse miRNAs in miRBase (RPM > =100)	http://www.mirbase.org
mirgenedb	Positive	1434	All miRNAs available in MirGeneDB	http://www.mirgenedb.org
hsa+	Positive	523	All human miRNAs available in MirGeneDB	http://www.mirgenedb.org
mmu+	Positive	395	All mouse miRNAs available in MirGeneDB	http://www.mirgenedb.org
gga+	Positive	229	All chicken miRNAs available in MirGeneDB	http://www.mirgenedb.org
dre+	Positive	287	All zebra fish miRNAs available in MirGeneDB	http://www.mirgenedb.org
NegHsa	Negative	68046	Extracted from genome and mRNAs of H. sapiens	http://adaa.polsl.pl/agudys/huntmi/huntmi.htm
Zou	Negative	14246	Extracted from coding regions	http://datamining.xmu.edu.cn/main/-leyiwei/mirnaDetect.html
pseudo	Negative	8492	Popular, used in many studies, constructed by using the protein coding sequences (CDSs) of human RefSeq genes with no known alternative splice events	http://web.bii.a-star.edu.sg/archive/stanley/Publications/Supp_materials/06-002-supp.html
Chen	Negative	3054	Excerpt of the combination of Zou and Pseudo	http://bioinformatics.hitsz.edu.cn/iMiRNA-SSF/Material.jsp
NotBestFold	Negative	1881	Created by not using the best fold proposed by RNAFold for human hairpins from miRBase	http://jlab.iyte.edu.tr/software/izmir
Shuffled	Negative	1423	Created by shuffling hsa data	http://jlab.iyte.edu.tr/software/izmir
hsa_{FR}	Positive	5000	Created by random number generation between minimum and maximum for all features (hsa)	http://jlab.iyte.edu.tr/software/izmir
hsa_{BQ}	Positive	5000	Created by random number generation between lower and upper quartile for all features (hsa)	http://jlab.iyte.edu.tr/software/izmir
hsa_{AM}	Positive	5000	Created by random number generation between 40th and 60th percentile for all features (hsa)	http://jlab.iyte.edu.tr/software/izmir
$pseudo_{FR}$	Negative	5000	Created by random number generation between minimum and maximum for all features (pseudo)	http://jlab.iyte.edu.tr/software/izmir
$pseudo_{BG}$	Negative	5000	Created by random number generation between lower and upper quartile for all features (pseudo)	http://jlab.iyte.edu.tr/software/izmir
$pseudo_{AM}$	Negative	5000	Created by random number generation between 40th and 60th percentile for all features (pseudo)	http://jlab.iyte.edu.tr/software/izmir

List of positive and negative data sets used to create and evaluate pre-miRNA detection tools. The first 13 rows refer to previously available data sets whereas the latter 8 are created for this study

Elde edilen yaklaşık 800 parametrenin çoğunun yüksek düzeyde bilgi verici olmadığı ve birbirleriyle etkileşimlerinin oldukça yüksek olduğu gözlemlenmiştir. Önceki çalışmalarımızda da fazla sayıda parametre kullanımının, duyarlılığı azalttığı açıkça gösterilmiştir (DOI: http://dx.doi.org/10.1109/HIBIT.2013.6661685). Tablo 7'de elimizdeki parametreler için gerçekleştirdiğimiz bilgi kazanımı (information gain) değerleri ilk 25 örnek için gösterilmiştir.

Tablo 7. En yüksek bilgi çıkarımı değerine sahip ilk 25 parametre.

featureName	informationGain
hpmfe_rf_l1	0.381
dns_z(efe)	0.354
dns_z(hpmfe_rf)	0.35
dns_z(hpmfe_rf/hpl)	0.343
dns_p(efe)	0.341
dns_p(hpmfe_rf/hpl)	0.34
dns_p(hpmfe_rf)	0.337
subu/sl	0.268
bpp/sl	0.268
subu/hpl	0.257
hpmfe_rf_l1/hpl	0.249
lsr(%bp)	0.247
hpmfe_rf_l1/sl	0.245
lsr(%bp)/hpl	0.238
lsr(%bp)/sl	0.237
bpp/hpl	0.227
dns_z(bpp)	0.213
lscm	0.213
dG/sl	0.212
dns_z(bpp/hpl)	0.209
dns_p(bpp/hpl)	0.208
mwmF	0.207
dG/hpl	0.2
hpmfe_rf/sl	0.199
efe/sl	0.199

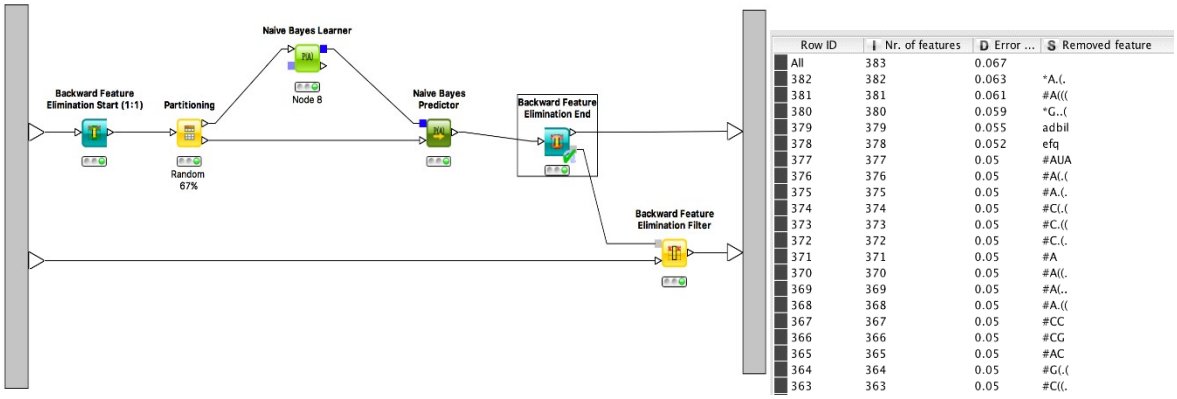
Ayrıca parametrelerin önemini daha iyi gözlemleyebilmek için Lineer Bağıntı hesaplaması gibi yöntemler denenmiş, bilgi çıkarımı işlemi cross-validation metoduyla test edilmiştir. Bulunan sonuçlar Tablo 7'yi destekler niteliktedir.

"Gereç ve Yöntem" bölümünde yer alan çalışmaların karşılaştırılması, bu miRNA saç tokası tahmin programlarının kullandığı parametrelerin farklı veri setleri üzerinde kullanılması esasına dayalıdır. Bulduğumuz sonuçlara göre bazı parametre grupları belirli pozitif ya da negatif veri setleri üzerinde oldukça başarısızdır. Örneğin, tamamıyla miRNA saç tokası yapısal özelliklerine dayalı parametre kullanımı, bizim geliştirdiğimiz NotBestFold; her bir miRNA dizisi için en ideal olan değil, alternatif ikincil yapıların seçilmesi ile oluşturulmuş negatif veri setiyle kullanıldığında, beklediğimiz üzere çok düşük bir performans sergilemiştir. Bu sebeple farklı parametrelerin kullanımıyla elde edilen modellerin karşılaştırılması ve kapsayıcı bir tahminde bulunulabilmesi için aynı veri setleri üzerinde test edilmesi önemlidir (Tablo 8).

Tablo 8. Kullanılan tüm mevcut ve yeni veri setlerinin türleri, boyutları, özellikleri ve kaynaklarının listesi (Saçar Demirci, Baumbach, and Allmer 2017).

	Negative										Positive													
Model	NegHsa	Zou	Pseudo	NotBestFold	Shuffled	Chen	$Pseudo_{TR}$	$Pseudo_{IBQ}$	$Pseudo_{AM}$	Neg Rank	hsa	mmu	mmu*	mirbase	hsa_{FR}	hsa_{IBQ}	hsa_{AM}	mirgenedb	hsa+	mmu+	gga+	dre+	Pos Rank	Total Rank
$Average_{DT}$	82	56	93	31	93	77	30	100	100	99	97	83	95	91	91	100	100	97	98	96	98	96	52	151
$Consensus_{NB}$	89	52	86	24	96	77	53	100	100	97	86	82	93	89	100	100	100	96	96	93	98	97	84	181
$Consensus_{DT}$	74	44	90	20	88	72	16	100	100	155	99	87	96	93	97	100	100	98	99	97	100	97	31	186
$Ding_{NB}$	93	47	84	9	96	73	30	100	100	127	88	84	94	90	100	100	100	97	97	96	97	97	59	186
$Average_{NB}$	92	58	89	95	97	82	86	100	100	50	83	77	91	87	99	100	100	94	95	91	96	95	148	198
Ng_{DT}	74	64	89	13	91	77	31	100	100	118	89	80	93	88	85	100	100	96	96	94	98	97	100	218
Consensus-Model	84	69	96	69	89	81	58	100	100	70	97	76	94	87	33	100	100	92	95	89	93	92	157	227
$Batuwita_{NB}$	90	53	83	11	97	76	45	100	100	114	86	79	92	87	98	100	100	96	96	93	97	97	120	234
$Bentwich_{NB}$	37	23	71	9	69	52	21	100	100	222	92	92	98	95	99	100	100	99	99	97	100	100	26	248
Ng_{NB}	74	42	81	9	87	63	36	100	100	187	86	83	95	91	99	100	100	96	97	94	98	98	63	250
$Gudys_{NB}$	87	48	81	14	96	76	27	100	100	140	87	80	92	88	100	100	100	95	95	93	97	96	113	253
$Lopes_{NB}$	89	54	80	9	96	73	39	100	100	134	86	79	93	86	98	100	100	93	94	91	96	93	146	280
$Ding_{DT}$	58	47	87	13	75	66	72	94	98	178	93	83	96	91	20	97	100	94	94	94	95	94	111	289
$Jiang_{NB}$	94	68	94	99	98	90	100	100	100	21	72	65	84	78	21	100	100	85	86	81	85	87	274	295
Gao_{NB}	90	55	90	94	96	85	100	100	100	59	77	71	85	82	41	100	100	89	90	86	89	91	239	298
$Gudys_{DT}$	82	54	85	17	83	71	26	91	100	160	93	82	93	90	90	96	100	92	93	90	93	93	142	302
$Jiang_{DT}$	89	51	85	33	92	67	37	97	97	142	92	76	91	89	69	100	100	93	94	90	93	95	160	302
$Bentwich_{DT}$	73	49	87	13	76	69	22	100	100	172	92	82	94	89	88	98	100	93	93	93	91	93	131	303
$Chen_{DT}$	60	52	85	34	81	68	44	98	100	158	93	78	93	89	61	92	100	93	94	91	94	94	150	308
ConsensusRule	94	65	94	95	98	86	87	100	100	28	76	59	84	84	3	100	100	83	83	80	83	87	281	309
Xu_{DT}	78	64	83	97	89	75	45	94	100	112	93	75	91	83	58	100	100	90	91	89	92	91	199	311
$Batuwita_{DT}$	76	53	85	21	88	71	30	100	100	145	90	78	93	88	91	98	100	91	93	89	93	90	167	312
$Lopes_{DT}$	57	51	84	17	75	70	27	90	97	197	90	80	91	89	87	96	100	95	96	92	97	95	133	330
Gao_{DT}	73	63	84	90	87	70	48	94	100	137	93	81	92	84	58	100	100	88	88	86	89	88	195	332
$Chen_{NB}$	73	54	88	90	89	79	9	100	100	124	78	76	89	86	99	100	100	88	89	84	90	92	208	332
$Ritchie_{NB}$	91	69	91	98	95	90	100	100	100	42	70	63	80	75	28	100	100	83	84	80	83	85	292	334
Xu_{NB}	92	81	92	100	96	92	100	100	100	26	71	62	81	66	2	100	100	81	81	78	79	85	308	334
Xue_{NB}	92	81	92	100	96	92	100	100	100	32	65	56	75	66	2	100	100	81	81	78	79	85	313	345
$Burgt_{NB}$	89	56	81	88	98	72	79	100	100	94	80	73	90	82	88	100	100	66	68	64	69	64	276	370
$Ritchie_{DT}$	76	49	83	66	83	72	67	94	100	150	92	76	91	86	36	98	100	84	85	82	86	85	236	386
$Burgt_{DT}$	77	51	85	61	79	73	49	92	99	147	90	76	90	85	57	93	99	86	88	83	86	86	243	390
Xue_{DT}	73	63	78	95	81	73	47	82	98	150	92	71	83	79	71	95	100	87	89	84	87	87	244	394

Etkin bir parametre seçimi için farklı yaklaşımlar denenmiştir. KNIME platformunda yer alan “backward feature elimination” metoduyla her bir parametrenin toplam hata oranına etkisi hesaplanarak, parametreler karşılaştırılmıştır (Şekil 5). Ayrıca bilgi çıkarımı ve bağıntı analizi yöntemleri kullanılarak filtreleme sistemiyle parametre seçimi uygulanmıştır.



Row ID	Nr. of features	Error ...	Removed feature
All	383	0.067	
382	382	0.063	*A.(.
381	381	0.061	#A(((.
380	380	0.059	*G..(
379	379	0.055	adbil
378	378	0.052	efq
377	377	0.05	#AUA
376	376	0.05	#A(.(
375	375	0.05	#A.(.
374	374	0.05	#C(.(
373	373	0.05	#C.((
372	372	0.05	#C.(.
371	371	0.05	#A
370	370	0.05	#A((.
369	369	0.05	#A(..
368	368	0.05	#A.((
367	367	0.05	#CC
366	366	0.05	#CG
365	365	0.05	#AC
364	364	0.05	#G(.(
363	363	0.05	#C((.

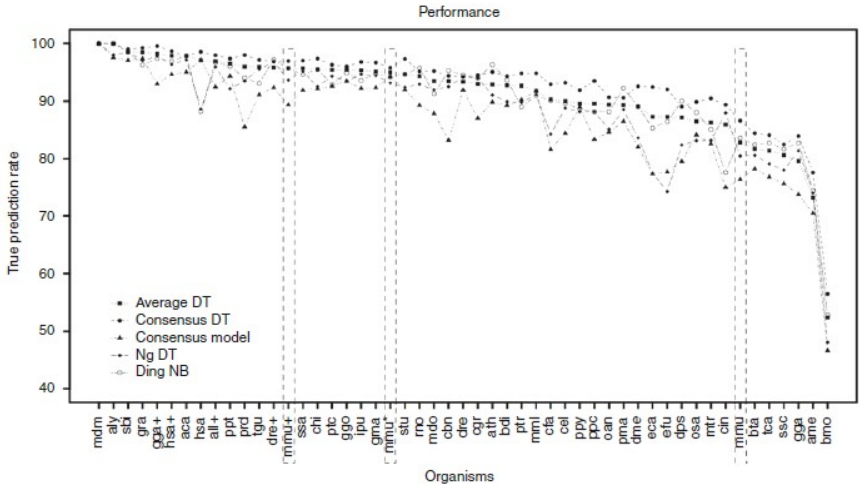
Şekil 5. Geriye dönük parametre elenmesi (Backward feature elimination) iş akışı. Sağ bölümde sonuç tablosunun ilk 20 elementi gösterilmiştir.

Geliştirilen parametre seçimi yöntemleri çeşitli veriler üzerinde denenmiştir ve elde edilen sonuçlar yayınlanmıştır (Yousef ve ark. 2015, 2016a, 2016b ve 2016c).

İki sınıf sınıflandırma (2 class-classification) yöntemiyle oluşturulan, her bir sınıflandırıcı (classifier; SVM, Decision Tree, Naive Bayes) için Monte Carlo cross validation metoduyla 1000 kez yinelenen ve en yüksek doğruluk (accuracy) değerine sahip modellerin doğruluğunu değerlendirilmesi için miRBase (http://www.mirbase.org/) veri tabanında yer alan organizmaların saç tokası dizileri kullanılmıştır. Model oluşturma işlemi için insan miRNA'ları kullanılmasına rağmen, farklı organizmalarda yüksek doğruluk değerleri elde edilmiştir (Tablo 9, Şekil 6).

Tablo 9. Elde edilen modellerin farklı türlere ait miRNAlar üzerinde test edilmesi. DT: Decision Tree, NB: Naive Bayes, mmu: filtrelenmiş fare miRNA veri seti. Filtrelenmiş fare miRNA'ları daha yüksek doğruluk değerlerine sahip olduğu için kalın karakterlerle belirtilmiştir*

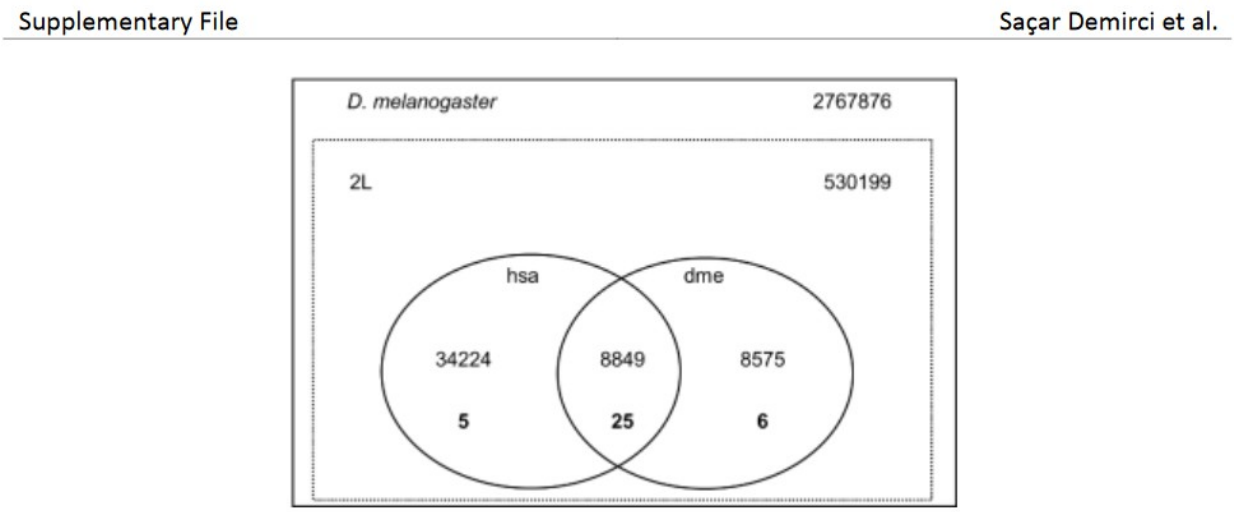
Kingdom	Species		Acronym	DT	NB
Animalia	*Homo*	*sapiens*	hsa	98,56	86,12
Animalia	*Mus*	*musculus*	mmu	86,59	82,23
Animalia	**Mus**	**musculus**	mmu*	**95,79**	**93,42**
Animalia	*Macaca*	*mulatta*	mml	94,83	88,85
Plantae	*Glycine*	*max*	gma	96,68	94,76
Animalia	*Rattus*	*norvegicus*	rno	95,15	93,54
Animalia	*Gorilla*	*gorilla*	ggo	96,02	90,63
Plantae	*Gossypium*	*raimondii*	gra	99,26	94,80
Plantae	*Malus*	*domestica*	mdo	100,00	99,51
Chromalveolata	Ectocarpus	siliculosus	esi	97,83	95,65
Virus	Rhesus	lymphocryptovirus	rlcv	97,22	97,22
Protozoa	Dictyostelium	discoideum	ddi	100,00	100,00



Şekil 6. Elde edilen modellerin farklı organizmalara ait miRNAlar üzerinde test edilmesi (Saçar Demirci, Baumbach, and Allmer 2017).

Geliştirdiğimiz izMiR yönteminin insan miRNA verisi kullanılarak dizayn edilmiş olmasına rağmen bitkilerden virüslere kadar geniş bir alanda etkili olması üzerine, *Drosophila melanogaster* gibi bir model organizmanın genomundan yeni miRNA'ların tahmininde de kullanılıp kullanılamayacağını test ettik (Saçar Demirci, Baumbach, and Allmer 2017). Bunun için küçük parçalara böldüğümüz (500 nükleotid, 250 nükleotid örtüşme) 2L kromozomundaki her bir dizi için gerekli hesaplamaları yapıp izMiR'i uyguladık. Elde ettiğimiz sonuçlara göre, beklediğimiz

gibi insan verisiyle oluşturulan izMiR modelleri oldukça başarılı bir performans sergilese de, Drosophila verisiyle oluşturulan modeller kendi genomlarından miRNA tahmininde daha etkili olmuştur (Şekil 7). Daha detaylı bilgi ve saç tokası yapılarının incelemesi için bakınız: http://jlab.iyte.edu.tr/software/izmir.



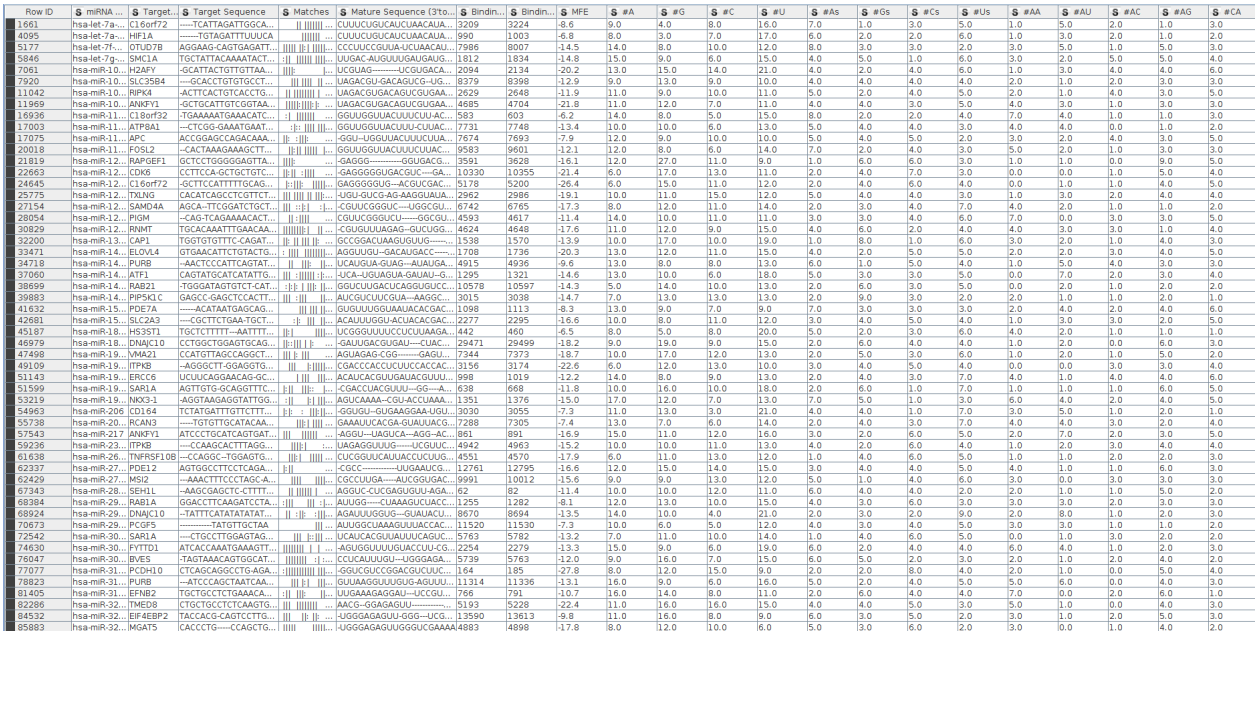
Şekil 7. Drosophila melanogaster genomundan tahmin edilen miRNA'lar. Çember içindeki sayılar belirlenen tahmin skorunu geçen saç tokası yapısı sayısını belirtir. Kalın karakterle yazılan sayılar tahmin edilen ve miRBase'de yer alan miRNA saç tokası sayısını gösterir (Saçar Demirci, Baumbach, and Allmer 2017)

2.2 MiRNA Hedef Tahmini

Hedef tahmininde, "Gereç ve Yöntem" bölümünde anlatılan araçların yazılan tahmin programımıza entegre olarak çalışması sağlanmıştır. BLAST aracılığıyla muhtemel hedef bölgeleri, programa sağlanan diziler arasından kesilerek RNAHybrid programıyla miRNA – hedef

düpleksi (çifti) oluşturulmuştur. RNAHybrid, termodinamik olarak düpleks oluşumunun en mümkün olduğu yere göre dizinin içindeki bölgeyi seçmiştir.

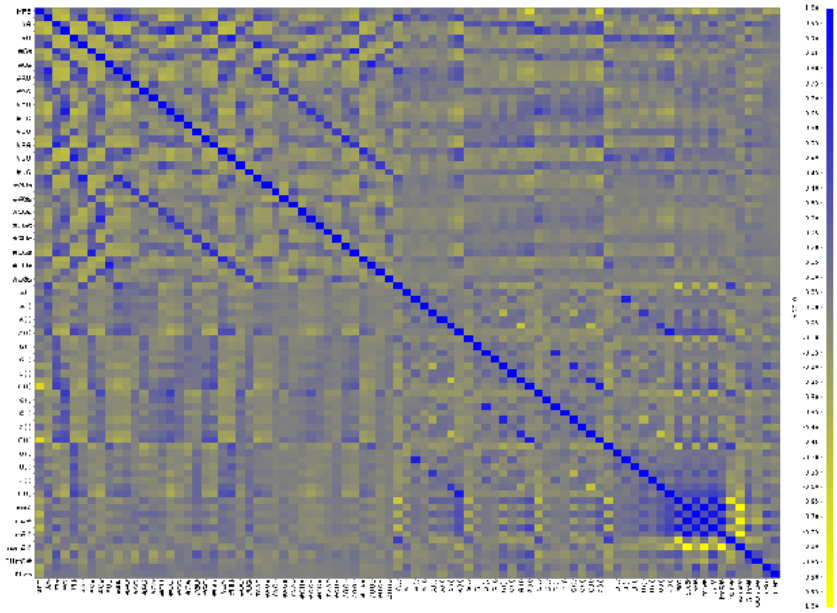
Tablo 10. Hedef tahmini için hesaplanan bazı parametreler ve sonuçları.



Row ID	miRNA ...	Target...	Target Sequence	Matches	Mature Sequence (3'to...	Bindin...	Bindin...	MFE	#A	#G	#C	#U	#As	#Gs	#Cs	#Us	#AA	#AU	#AC	#AG	#CA
1661	hsa-let-7a-...	C16orf72	-----TCATTAGATTGGCA...	...	CUUUCUGUCAUCUAACAUA...	3209	3224	-8.6	9.0	4.0	8.0	16.0	7.0	1.0	3.0	5.0	1.0	5.0	2.0	1.0	3.0
4095	hsa-let-7a-...	HIF1A	-------TGTAGATTTUUUCA	...	CUUUCUGUCAUCUAACAUA...	990	1003	-6.8	8.0	3.0	7.0	17.0	6.0	2.0	2.0	6.0	1.0	3.0	2.0	1.0	2.0
5177	hsa-let-7f-...	OTUD7B	AGGAAG-CAGTGAGATT...	...	CCCUUCCGUUA-UCUAACAU...	7986	8007	-14.5	14.0	8.0	10.0	12.0	8.0	3.0	3.0	2.0	3.0	5.0	1.0	5.0	3.0
5846	hsa-let-7g-...	SMC1A	TGCTATTACAAAATACT...	...	UUGAC-AUGUUUGAUGAUG...	1812	1834	-14.8	15.0	9.0	6.0	15.0	4.0	5.0	1.0	6.0	3.0	2.0	5.0	5.0	4.0
7061	hsa-miR-10...	H2AFY	-GCATTACTGTTGTTAA...	...	UCGUAG----------UCGUGACA...	2094	2134	-20.2	13.0	15.0	14.0	21.0	4.0	2.0	4.0	6.0	1.0	3.0	4.0	4.0	6.0
7920	hsa-miR-10...	SLC35B4	----GCACCTGTGTGCCT...	...	UAGACGU-GACAGUCG--UG...	8379	8398	-12.9	9.0	13.0	9.0	10.0	4.0	4.0	4.0	4.0	2.0	1.0	2.0	3.0	3.0
11042	hsa-miR-10...	RIPK4	-ACTTCACTGTCACCTG...	...	UAGACGUGACAGUCGUGAA...	2629	2648	-11.9	11.0	9.0	10.0	11.0	5.0	2.0	4.0	5.0	2.0	1.0	4.0	3.0	5.0
11969	hsa-miR-10...	ANKFY1	-GCTGCATTGTCGGTAA...	...	UAGACGUGACAGUCGUGAA...	4685	4704	-21.8	11.0	12.0	7.0	11.0	4.0	4.0	3.0	5.0	4.0	3.0	1.0	3.0	3.0
16936	hsa-miR-11...	C18orf32	-TGAAAAATGAAACATC...	...	GGUUGGUUACUUUCUU-AC...	583	603	-6.2	14.0	8.0	5.0	15.0	8.0	2.0	2.0	4.0	7.0	4.0	1.0	1.0	3.0
17003	hsa-miR-11...	ATP8A1	---CTCGG-GAAATGAAT...	...	GGUUGGUUACUUU-CUUAC...	7731	7748	-13.4	10.0	10.0	6.0	13.0	5.0	4.0	4.0	3.0	4.0	4.0	0.0	1.0	2.0
17075	hsa-miR-11...	APC	ACCGGAGCCAGACAAA...	...	-GGU--UGGUUACUUUCUUA...	7674	7693	-7.9	12.0	9.0	10.0	10.0	5.0	4.0	5.0	2.0	3.0	2.0	4.0	3.0	5.0
20018	hsa-miR-11...	FOSL2	--CACTAAAGAAAGCTT...	...	GGUUGGUUACUUUCUUAC...	9583	9601	-12.1	12.0	8.0	6.0	14.0	7.0	2.0	4.0	3.0	5.0	2.0	1.0	3.0	3.0
21819	hsa-miR-12...	RAPGEF1	GCTCCTGGGGGAGTTA...	...	-GAGGG----------GGUGACG...	3591	3628	-16.1	12.0	27.0	11.0	9.0	1.0	6.0	6.0	3.0	1.0	1.0	0.0	9.0	5.0
22663	hsa-miR-12...	CDK6	CCTTCCA-GCTGCTGTC...	...	-GAGGGGGUGACGUC----GA...	10330	10355	-21.4	6.0	17.0	13.0	11.0	2.0	4.0	7.0	3.0	0.0	0.0	1.0	5.0	4.0
24645	hsa-miR-12...	C16orf72	-GCTTCCATTTTTGCAG...	...	GAGGGGGUG---ACGUCGAC...	5178	5200	-26.4	6.0	15.0	11.0	12.0	2.0	4.0	6.0	4.0	0.0	1.0	1.0	4.0	5.0
25775	hsa-miR-12...	TXLNG	CACATCAGCCTCGTTCT...	...	-UGU-GUCG-AG-AAGGUAUA...	2962	2986	-19.1	10.0	11.0	15.0	12.0	5.0	4.0	4.0	3.0	1.0	3.0	2.0	4.0	4.0
27154	hsa-miR-12...	SAMD4A	AGCA--TTCGGATCTGCT...	...	-CGUUCGGGUC----UGGCGU...	6742	6765	-17.3	8.0	12.0	11.0	14.0	2.0	3.0	4.0	7.0	4.0	2.0	1.0	1.0	2.0
28054	hsa-miR-12...	PIGM	--CAG-TCAGAAAACACT...	...	CGUUCGGGUCU-----GGCGU...	4593	4617	-11.4	14.0	10.0	11.0	11.0	3.0	3.0	4.0	6.0	7.0	0.0	3.0	3.0	5.0
30829	hsa-miR-12...	RNMT	TGCACAAATTTGAACAA...	...	-CGUGUUUAGAG--GUCUGG...	4624	4648	-17.6	11.0	12.0	9.0	15.0	4.0	6.0	2.0	4.0	4.0	3.0	3.0	1.0	4.0
32200	hsa-miR-13...	CAP1	TGGTGTGTTTC-CAGAT...	...	GCCGGACUAAGUGUUG------...	1538	1570	-13.9	10.0	17.0	10.0	19.0	1.0	8.0	1.0	6.0	3.0	2.0	1.0	4.0	3.0
33471	hsa-miR-14...	ELOVL4	GTGAACATTCTGTACTG...	...	AGGUUGU-GACAUGACC-----...	1708	1736	-20.3	13.0	12.0	11.0	15.0	4.0	2.0	5.0	5.0	2.0	2.0	3.0	4.0	5.0
34718	hsa-miR-14...	PURB	--AACTCCCATTCAGTAT...	...	UCAUGUA-GUAG---AUAUGA...	4915	4936	-9.6	13.0	8.0	8.0	13.0	6.0	1.0	5.0	4.0	1.0	5.0	4.0	3.0	3.0
37060	hsa-miR-14...	ATF1	CAGTATGCATCATATTG...	...	-UCA--UGUAGUA-GAUAU--G...	1295	1321	-14.6	13.0	10.0	6.0	18.0	5.0	3.0	3.0	5.0	0.0	7.0	2.0	3.0	4.0
38699	hsa-miR-14...	RAB21	-TGGGATAGTGTCT-CAT...	...	GGUCUUGACUCAGGUGUCC...	10578	10597	-14.3	5.0	14.0	10.0	13.0	2.0	6.0	3.0	5.0	0.0	2.0	1.0	2.0	2.0
39883	hsa-miR-14...	PIP5K1C	GAGCC-GAGCTCCACTT...	...	AUCGUCUUCGUA---AAGGC...	3015	3038	-14.7	7.0	13.0	13.0	13.0	2.0	9.0	3.0	2.0	2.0	1.0	1.0	2.0	1.0
41632	hsa-miR-15...	PDE7A	-----ACATAATGAGCAG...	...	GUGUUUGGUAAUACACGAC...	1098	1113	-8.3	13.0	9.0	7.0	9.0	7.0	3.0	3.0	3.0	2.0	4.0	2.0	4.0	6.0
42681	hsa-miR-15...	SLC2A3	----CGCTTCTGAA-TGCT...	...	ACAUUUGGU-ACUACACGAC...	2277	2295	-16.6	10.0	8.0	11.0	12.0	3.0	4.0	5.0	4.0	1.0	3.0	3.0	2.0	5.0
45187	hsa-miR-18...	HS3ST1	TGCTCTTTTT---AATTTT...	...	UCGGGUUUUCCUCUUAAGA...	442	460	-6.5	8.0	5.0	8.0	20.0	5.0	2.0	3.0	6.0	4.0	2.0	1.0	1.0	1.0
46979	hsa-miR-18...	DNAJC10	CCTGGCTGGAGTGCAG...	...	-GAUUGACGUGAU----CUAC...	29471	29499	-18.2	9.0	19.0	9.0	15.0	2.0	6.0	4.0	4.0	1.0	2.0	0.0	6.0	3.0
47498	hsa-miR-19...	VMA21	CCATGTTAGCCAGGCT...	...	AGUAGAG-CGG--------GAGU...	7344	7373	-18.7	10.0	17.0	12.0	13.0	2.0	5.0	3.0	6.0	1.0	2.0	1.0	5.0	2.0
49109	hsa-miR-19...	ITPKB	--AGGGCTT-GGAGGTG...	...	CGACCCACCUCUUCCACCAC...	3156	3174	-22.6	6.0	12.0	13.0	10.0	3.0	4.0	5.0	4.0	0.0	0.0	3.0	3.0	4.0
51143	hsa-miR-19...	ERCC6	UCUUCAGGAACAG-GC...	...	ACAUCACGUUGAUACGUUU...	998	1019	-12.2	14.0	8.0	9.0	13.0	2.0	4.0	3.0	7.0	4.0	1.0	4.0	4.0	6.0
51599	hsa-miR-19...	SAR1A	AGTTGTG-GCAGGTTTC...	...	-CGACCUACGUUU---GG----A...	638	668	-11.8	10.0	16.0	10.0	18.0	2.0	6.0	1.0	7.0	1.0	1.0	1.0	6.0	5.0
53219	hsa-miR-19...	NKX3-1	-AGGTAAGAGGTATTGG...	...	AGUCAAAA--CGU-ACCUAAAA...	1351	1376	-15.0	17.0	12.0	7.0	13.0	7.0	5.0	1.0	3.0	6.0	4.0	2.0	4.0	5.0
54963	hsa-miR-206	CD164	TCTATGATTTGTTCTTT...	...	-GGUGU--GUGAAGGAA-UGU...	3030	3055	-7.3	11.0	13.0	3.0	21.0	4.0	4.0	1.0	7.0	3.0	5.0	1.0	2.0	1.0
55738	hsa-miR-20...	RCAN3	-----TGTGTTGCATACAA...	...	GAAAUUCACGA-GUAUUACG...	7288	7305	-7.4	13.0	7.0	6.0	14.0	2.0	4.0	3.0	7.0	4.0	4.0	3.0	2.0	4.0
57543	hsa-miR-217	ANKFY1	ATCCCTGCATCAGTGAT...	...	-AGGU---UAGUCA--AGG--AC...	861	891	-16.9	15.0	11.0	12.0	16.0	3.0	2.0	6.0	5.0	2.0	7.0	2.0	3.0	5.0
59236	hsa-miR-23...	ITPKB	----CCAAGCACTTTAGG...	...	UAGAGGUUUG-----UCGUUC...	4942	4963	-15.2	10.0	10.0	11.0	13.0	4.0	2.0	6.0	4.0	1.0	2.0	3.0	4.0	4.0
61638	hsa-miR-26...	TNFRSF10B	---CCAGGC--TGGAGTG...	...	CUCGGUUCAUUACCUCUUG...	4551	4570	-17.9	6.0	11.0	13.0	12.0	1.0	4.0	6.0	5.0	1.0	1.0	2.0	2.0	3.0
62337	hsa-miR-27...	PDE12	AGTGGCCTTCCTCAGA...	...	-CGCC----------UUGAAUCG...	12761	12795	-16.6	12.0	15.0	14.0	15.0	3.0	4.0	4.0	5.0	4.0	1.0	1.0	6.0	3.0
62429	hsa-miR-27...	MSI2	---AAACTTTCCCTAGC-A...	...	CGCCUUGA----AUCGGUGAC...	9991	10012	-15.6	9.0	9.0	13.0	12.0	5.0	1.0	4.0	6.0	3.0	0.0	3.0	3.0	3.0
67343	hsa-miR-28...	SEH1L	--AAGCGAGCTC-CTTTT...	...	AGGUC-CUCGAGUGUU-AGA...	62	82	-11.4	10.0	10.0	12.0	11.0	6.0	4.0	4.0	2.0	2.0	1.0	1.0	5.0	2.0
68384	hsa-miR-29...	RAB1A	GGACCTTCAAGATCCTA...	...	AUUGG-----CUAAAGUCUACC...	1255	1282	-8.1	12.0	13.0	10.0	15.0	4.0	3.0	6.0	3.0	3.0	3.0	2.0	3.0	3.0
68924	hsa-miR-29...	DNAJC10	--TATTTCATATATATAT...	...	AGAUUUGGUG---GUAUACU...	8670	8694	-13.5	14.0	10.0	4.0	21.0	2.0	3.0	2.0	9.0	2.0	8.0	1.0	2.0	3.0
70673	hsa-miR-29...	PCGF5	----------TATGTTGCTAA	...	AUUGGCUAAAGUUUACCAC...	11520	11530	-7.3	10.0	6.0	5.0	12.0	4.0	3.0	4.0	5.0	3.0	3.0	1.0	1.0	2.0
72542	hsa-miR-30...	SAR1A	----CTGCCTTGGAGTAG...	...	UCAUCACGUUAUUUCAGUC...	5763	5782	-13.2	7.0	11.0	10.0	14.0	1.0	4.0	6.0	5.0	0.0	1.0	3.0	2.0	2.0
74630	hsa-miR-30...	FYTTD1	ATCACCAAATGAAAGTT...	...	-AGUGGUUUUGUACCUU-CG...	2254	2279	-13.3	15.0	9.0	6.0	19.0	6.0	2.0	4.0	4.0	6.0	4.0	1.0	2.0	3.0
76047	hsa-miR-30...	BVES	-TAGTAAACAGTGGCAT...	...	CCUCAUUUGU--UGGGAGA...	5739	5763	-12.0	9.0	16.0	7.0	15.0	6.0	5.0	2.0	3.0	2.0	1.0	2.0	4.0	2.0
77077	hsa-miR-31...	PCDH10	CTCAGCAGGCCTG-AGA...	...	-GGUCGUCCGGACGUCUUC...	164	185	-27.8	8.0	12.0	15.0	9.0	2.0	2.0	8.0	4.0	2.0	1.0	0.0	5.0	4.0
78823	hsa-miR-31...	PURB	--ATCCCAGCTAATCAA...	...	GUUAAGGUUUGUG-AGUUU...	11314	11336	-13.1	16.0	9.0	6.0	16.0	5.0	2.0	4.0	5.0	5.0	6.0	0.0	4.0	3.0
81405	hsa-miR-31...	EFNB2	TGCTGCCTCTGAAACA...	...	UUGAAAGAGGAU---UCCGU...	766	791	-10.7	16.0	14.0	8.0	11.0	2.0	6.0	4.0	4.0	7.0	0.0	2.0	6.0	1.0
82286	hsa-miR-32...	TMED8	CTGCTGCCTCTCAAGTG...	...	AACG--GGAGAGUU----------...	5193	5228	-22.4	11.0	16.0	16.0	15.0	4.0	4.0	5.0	3.0	5.0	1.0	0.0	4.0	3.0
84532	hsa-miR-32...	EIF4EBP2	TACCACG-CAGTCCTTG...	...	-UGGGAGAGUU-GGG---UCG...	13590	13613	-9.8	11.0	16.0	8.0	9.0	6.0	3.0	5.0	2.0	3.0	1.0	2.0	5.0	3.0
85883	hsa-miR-32...	MGAT5	CACCCTG-----CCAGCTG...	...	-UGGGAGAGUUGGGUCGAAAA	4883	4898	-17.8	8.0	12.0	10.0	6.0	5.0	3.0	6.0	2.0	3.0	0.0	1.0	4.0	2.0

Oluşturulan düpleksler için yaklaşık 120 parametre kodlanmış (Tablo 10) ve KNIME ortamında bu parametrelerin bilgi kazanımları ile bağıntı değerleri hesaplanmıştır (Şekil 8).

featureName	informationGain
scoS	0.684
mS#	0.675
nomS#	0.557
tmS#	0.557
U...	0.272
U(((	0.267
m#	0.181
U..(	0.123
G((.	0.103
U.((	0.098
U.(.	0.09
MFE	0.089
A(((	0.084
sco	0.084
C((.	0.071
nom#	0.071
U(.(	0.066
GUmS#	0.066
tm#	0.058
nBul	0.051
U((.	0.048
A.(.	0.047
BLen	0.047
G.((	0.046
G...	0.044

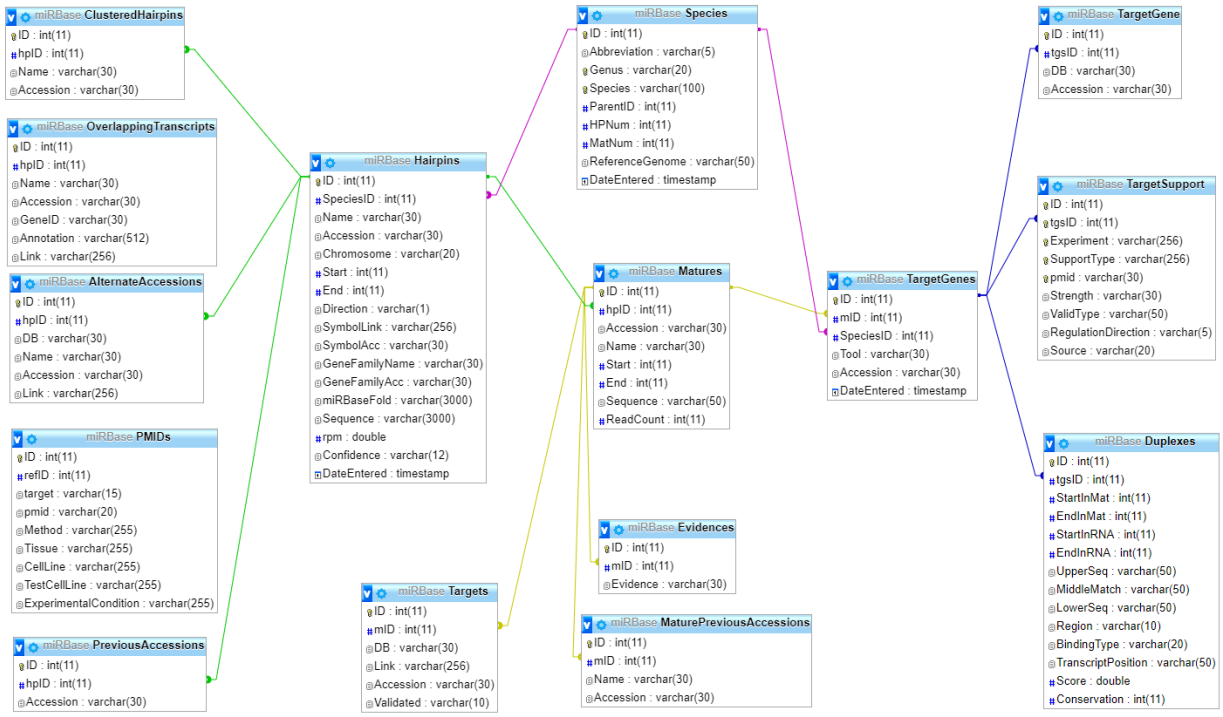


Şekil 8. Hedef tahmini parametrelerinin bilgi kazanımı ve bağıntı analizi. Oluşturulan parametrlerden en yüksek bilgi kazanım değerlerine sahip olanlar solda gösterilmiştir. Sağ tarafta, bağıntı değerlerinden oluşuturulmuş ısı haritası görülebilir.

İki sınıf sınıflandırma yöntemiyle hedefleme tahmini yapmak için gerekli negatif veri, bizim tarafımızdan, düplekslerin rastgele hale getirilmesiyle oluşturulmuştur. Oluşturulan bu veri kullanılarak hedef tahmini yapıldığında, pozitif verilerin %95'ten fazla doğruluk oranıyla tahmin edildiği görülmüştür. Ne yazık ki, bu doğruluk oranını sağlayan tahmin modelimiz, gerçek hedefleri bulmada başarısız kalmıştır. Bunun sebebi incelendiğinde, programımızda kullanılan entegre BLAST ve RNAHybrid araçlarının, hedef tahmininde başarısız kaldığı görülmüştür. RNAHybrid'in oluşturduğu düplekslerin gerçekliğe yakınlığını test etmek amacıyla, oluşturduğumuz veritabanından yüksek güvenilirlikte (8'den çok makale tarafından deneysel yöntemlerle kanıtlanmış) olan miRNA hedef dizileri kullanılmıştır. Yüksek güvenirlikteki bu miRNA – hedef düplekslerinin sadece yaklaşık %4 kadarı RNAHybrid tarafından tahmin edilebildiği için bu yaklaşımdan vazgeçilmiştir. Bunun yerine, miRNA hedefleme tahminine dayalı çalışmalarımızda, psRNATarget (http://plantgrn.noble.org/v1_psRNATarget/) isimli, yüksek miktarda özelleştirme parametrelerine sahip hedefleme tahmini aracı kullanılmıştır.

2.3 Veritabanı

Elde edilen tahminlerin bilinen miRNA öncü, saç tokası, olgun yapıları ve tahminleriyle benzerliklerini ve etkileşimlerini inceleyebilmek için, internette yer alan miRBase, TarBase ve miRTarBase gibi veri tabanları geliştirmekte olduğumuz sisteme entegre edilmiştir. Bu yapıdaki veri tabanı, bizim tahminlerimizin analizi için de oldukça kullanışlıdır. Bu nedenle belirtilen veri tabanları elde edilip birleştirilerek, hedef verilerini de kapsayacak şekilde genişletilmiş ve yeni bir veri tabanı oluşturulmuştur (Şekil 9).

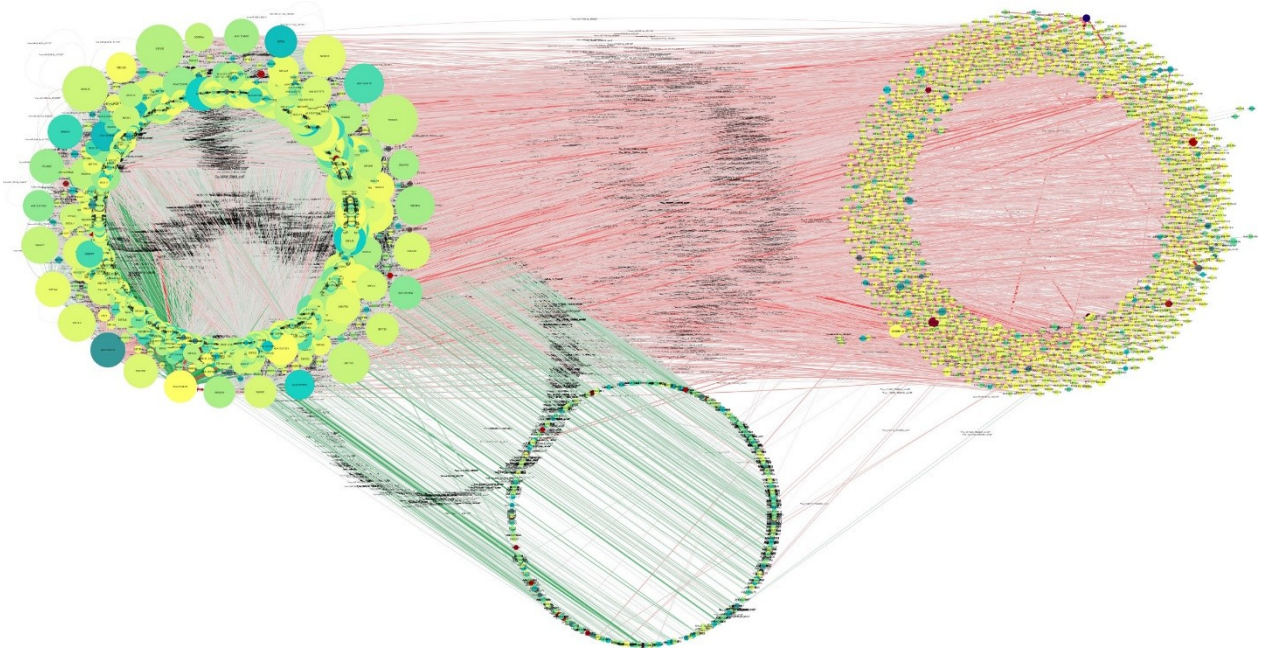


Şekil 9. Oluşturulmuş veritabanı. Veritabanımızda miRBase, miRTarBase ve TarBase veritabanlarında mevcut bulunan tüm miRNA ve etkileşimleri bulunmaktadır.

2.4 Ağ Analizi

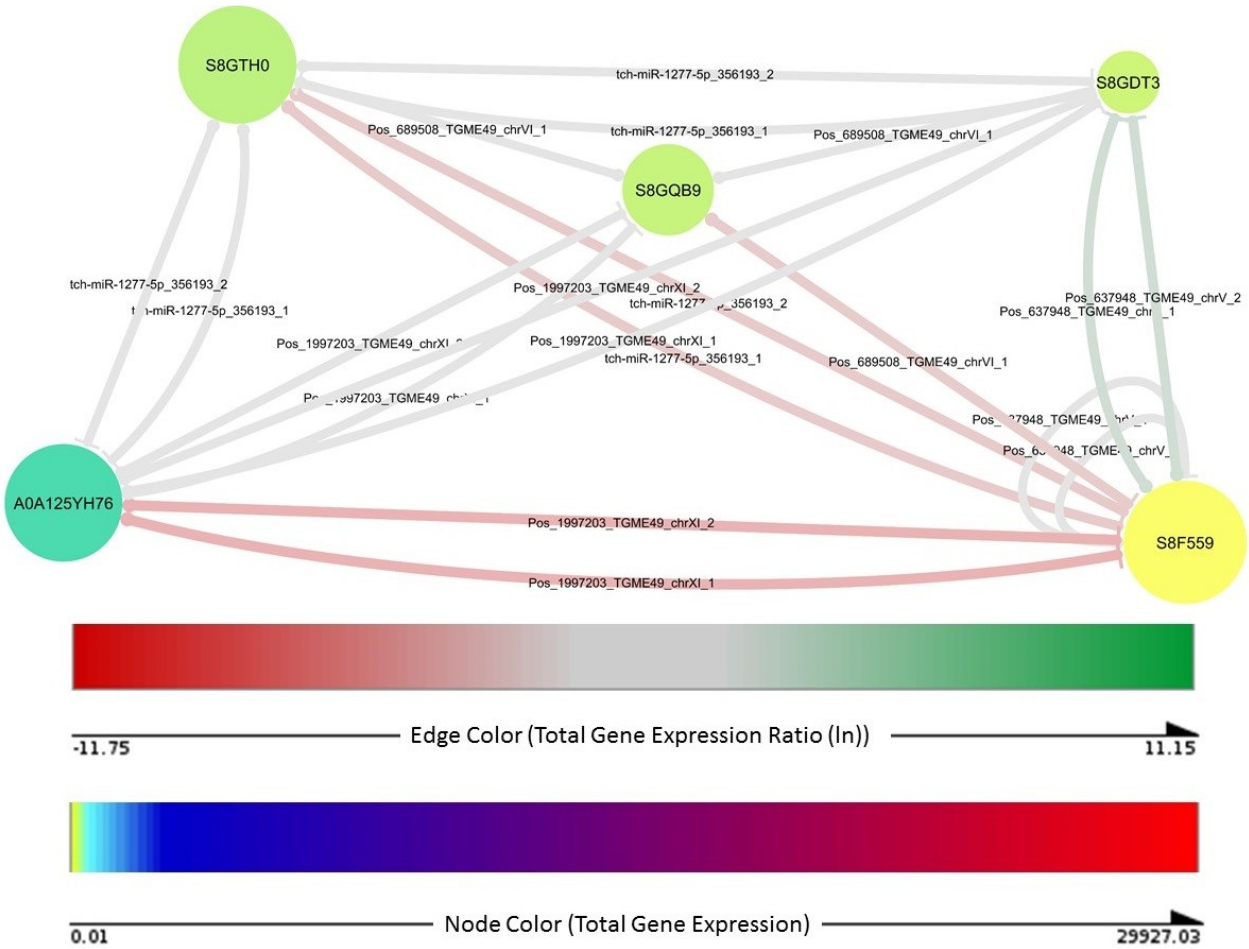
Yaptığımız çalışmalar sonucunda *Toxoplasma gondii*'nin miRNA düzenleyici ağı, proje kapsamında oluşturduğumuz aracın yardımıyla kurulabilmiştir. *Toxoplasma gondii*'nin gen yolakları henüz net belirlenmemiş ve bir çok geni hipotetik olarak belirtilmektedir. Bu yüzden, bu çalışmada sadece muhtemel miRNA genleri üzerinde durulmuştur. Yapılan gen expresyonu değerlerine göre filtrelenen genlere göre, eksprese olan genlerden diğer eksprese olan genleri hedefleyen miRNA'lar ağ üzerinde gösterilmiştir. Bu bağlamda, toplam yaklaşık 65.500 gen – miRNA etkileşimi bulunmuştur (Şekil 10).

Şekil 10. Toxoplasma gondii'nin miRNA düzenleyici ağı. Yapılan gen ekspresyonu çalışmasıyla da birleştirilen ağda, yuvarlaklar genleri, çizgiler miRNA'ları göstermektedir. Yuvarlakların büyüklüğü, o genin miRNA etkileşim sayısına bağlı olarak artmaktadır. Aynı zamanda toplam gen ekspresyonu yeşilden kırmızı renge doğru artmaktadır. MiRNA'lara atanan renkler ise miRNA kaynağı gen ile hedeflenen gen arasındaki ekspresyon oranını göstermektedir. Eğer kaynak ekspresyon daha büyükse kırmızı, hedef gen ekspresyonu daha büyükse yeşil renk kullanılmıştır.



Ağ içinde bulunan kliklerin, potansiyel yeni biyobelirteç ve yeni ilaç hedefleri için önemli bilgiler içermesinden ötürü, ağımızda bulunan ve 2 genden daha fazla gen içeren tüm klikler çıkarılmıştır. Yaklaşık 64.300 klik bulunan ağımızda, 3 ile 11 gen arasında, farklı büyüklüklerde klikler bulunmuştur (Şekil 11). Oluşturulan ağın, rastgele bağlanmalarla oluşmadığını ve biyolojik anlamlar içerdiğini göstermek adına, Cytoscape aracına eklenebilen ağ rasgeleleştirici

(NetworkRandomizer) aracı kullanılmıştır. Bu araçla 10 kere rastgeleleştirilen ağımızda bulunan klikler incelenmiştir. İncelemeler sonucunda, sadece 2 rastgele ağda 4 genli bir kliğin oluştuğu, bunun dışında rastgele oluşumların 3 genden daha öteye geçmediği gözlemlenmiştir. Bu yüzden, oluşturulan bu miRNA ağının biyolojik olarak anlamlı olduğu sonucuna varılmıştır.



Şekil 11. Örnek Toxoplasma gondii kliği. Bu klikte, 5 gen ve 19 miRNA görülmektedir. Bazı miRNA'ların, kendilerini sentezleyen genleri hedefleyerek regülasyon sağladığı da görülebilir.

3. SONUÇ VE TARTIŞMA

Proje kapsamında, planlanmış olan miRNA tahmin programı yapılmış ve birçok farklı çalışmada kullanılmıştır. Bu program, KNIME programıyla birlikte çalışarak tahminlerin doğruluk değerlerini hesapladığı gibi aynı zamanda tahmin modellerinin de saklanmasına olanak sağlamaktadır. Ayrıca, KNIME programının başlıksız (headless) çalışma imkanı, programın entegrasyonuna ve ardışık düzen (pipeline) kurulmasına kolaylık sağlamaktadır. Geliştirilen miRNA tahmininin, mevcut tahmin sistemleri arasında en gelişkin (state of the art) araç olduğu, çalışmalarımızca gösterilmiştir.

Projede ilk planlanan şekilde hedef tahmininin yapılamamasından dolayı ağ analizlerinde aksaklıklar yaşanmıştır. Planlanan hedef tahmini yöntemimiz, BLAST ve RNAHybrid araçlarının birlikte kullanımına dayalıyken, bu araçlarda kullanılan algoritmaların gerçeğe yakın hedefler bulmada yetersiz kaldıkları görülmüştür.

MiRNA tahmin aracımız ve psRNATarget hedef tahminleme aracının bir arada kullanımıyla birçok farklı organizma için yayın çıkartılması başarılmıştır. Alman ikili işbirlikçilerimizin sistem biyolojisi alanında ağ analizi yapan programı VANESA'ya, miRNA etkileşimleri eklenmiştir. Hedef tahmin programının, kendi oluşturduğumuz programın bir parçası olmamasından ötürü tahmin sistemimiz VANESA'ya entegre edilememiştir. Ancak, miRNA ve hedef tahminleri yaptıktan sonra bu sonuçların VANESA programında kullanılması sağlanmıştır. Bu proje kapsamında programımız ve VANESA programı kullanılarak kızamık miRNA'larının insan genlerine ve yolaklarına etkisi incelenmiş, uluslararası, hakemli bir bilimsel dergide yayınlanmıştır.

Oluşturduğumuz araçlar ve kullandığımız metotlar ile *Drosophila melonagaster* ve *Toxoplasma gondii* genom dizilerinden miRNA tahminleri başarıyla yapılmıştır. Bu tahminlerde, tahmin doğruluk oranı olarak sırasıyla %99 ile %100 gibi oldukça yüksek değerler seçilmesine rağmen, hali hazırda deneysel olarak ispatlanmış miRNA'ların bulunmasıyla birlikte daha önce bulunmamış bir takım miRNA'lar bulunmuştur. *Toxoplasma gondii* çalışmamızda gerçekleştirdiğimiz tahminler sayesinde miRNA'ların genlerle olan etkileşimlerinin de incelenmesi mümkün olmuştur. Bu gibi etkileşimlerin incelenmesi, gelecekte yapılacak olan çalışmaların odak noktası olmalıdır.

Yapılacak sonraki çalışmalarda, Smith – Waterman algoritmasının puanlama matrisi, miRNA – hedef dizi oluşumuna uygun hale getirilerek gerçeğe daha yakın hedefleme tahmini yapılabilir.

Bu doğrultuda, hedefleme tahmininin VANESA'ya entegre edilmesi mümkün hale gelerek dış kaynaklı yazılımların iş akışından çıkarılması sağlanabilir. Tek bir iş akışı kullanılarak genomdan düzenleyici ağ sonucuna ulaşılması, bilgisayar becerileri ileri düzeyde olmayan araştırmacıların da bu araçları kullanabilmesine olanak sağlayacaktır.

Ek olarak, pre-miRNA parametrelerinin hesaplanması, özellikle dizilerdeki bazların sırasının rastgele değiştirilip istatistiksel değerlerin hesaplandığı parametreler için hala çok uzun süreler almaktadır. Yaklaşık 3 milyar baz çiftinden oluşan insan genom dizisinin, tüm olası pre-miRNA'lar için parametre hesaplamasının standart donanımdaki bir bilgisayar kullanımıyla 12 aydan fazla süreceği öngörülmüştür. Yaptığımız *Toxoplasma gondii* çalışmasında hesaplanan pre-miRNA'lar, yaklaşık 69 milyon baz çiftinden oluşan toxoplasma genomu için 1 ay kadar süre almıştır. Bunlar göz önünde bulundurularak, zamandan kazanmak adına sonraki çalışmalarda parametre seçimi yapılması gerekmektedir. Parametre seçimine alternatif olarak IBM Watson ya da Google TensorFlow gibi derin öğrenme (deep learning) yöntemleri, pre-miRNA tahmini için kullanılabilir. Bu yöntemlerle parametre tahmin basamağı tamamen öğrenme algoritmasına bırakılarak verimli bir iş akışı kurulabilir. Ancak, derin öğrenme yöntemlerinin kullanılabilmesi için algoritmanın muazzam sayıda pre-miRNA örneğiyle beslenmesi gerekmektedir. Bu yüzden, türlere özgü miRNA tahminleri yapmak için daha çok miRNA'nın keşfedilmesi ve deneysel olarak doğrulanmasıyla bu yöntemler gelecekte denenebilir.

4. REFERANSLAR

Agarwal S, Vaz C, Bhattacharya A, Srinivasan A, Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM), BMC Bioinformatics, 11 Suppl 1:S29 (2010).

Allmer J, MikroRNA Analizi ve Saklanmasında Hesaplamaya Dayalı Yaklaşımlar, MikroRNA ve Sinir Sistemi, ed: Şermin Genç, Şenol Form Matbaacılık, Ankara, (2011).

Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D., MicroRNAs and other tiny endogenous RNAs in C. elegans, Curr Biol., 13(10):807-18. (2003).

Bartel DP, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, 116(2):281-97 (2004).

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z, Identification of hundreds of conserved and nonconserved human microRNAs, Nat Genet., 37(7):766-70 (2005).

Bentwich I, Identifying human microRNAs, Curr Top Microbiol Immunol., 320:257-69 (2008).

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk R.H.A, Cuppen E, Phylogenetic shadowing and computational identification of human microRNA genes, Cell 120: 21-24 (2005).

Berezikov E, Cuppen E, Plasterk RH., Approaches to microRNA discovery, Nat Genet., 38 Suppl:S2-7 (2006).

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM, Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome, Science, Vol. 299 no. 5611 pp. 1391-94, (2003).

Brameier M, Wiuf C., Ab initio identification of human microRNAs based on structure motifs, BMC Bioinformatics, 8:478, (2007).

Brodersen P, **Voinnet** O., Revisiting the principles of microRNA target recognition and mode of action, Nat Rev Mol Cell Biol., 10(2):141-8, (2009).

Bushati N, Cohen SM., microRNA functions, Annu Rev Cell Dev Biol., 23:175-205 (2007).

Ding J, Zhou S, Guan J, MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features, BMC Bioinformatics,11 Suppl 11:S11 (2010).

Ding J, Li X, Hu H, TarPmiR: a new approach for microRNA target site prediction, *Bioinformatics*, *32*(May), 1–8. https://doi.org/10.1093/bioinformatics/btw318 (2016).

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS., MicroRNA targets in Drosophila, Genome Biol., 5(1):R1 (2003).

Filipowicz W, Bhattacharyya SN, Sonenberg N., Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?, Nat Rev Genet., 9(2):102-14, (2008).
Ghoshal A, Shankar R, Bagchi S, Grama A, Chaterji S, MicroRNA target prediction using thermodynamic and sequence curves, *BMC Genomics*, *16*(1), 999 (2015). https://doi.org/10.1186/s12864-015-1933-2 (2015).
He L, **Hannon** GJ, MicroRNAs: small RNAs with a big role in gene regulation, Nat Rev Genet., 5(7):522-31 (2004).
Hofacker, I.L., Vienna RNA secondary structure server, Nucleic Acids Res., 31, 3429–3431 (2003).
Jones-Rhoades MW, Bartel DP, Bartel B., MicroRNAS and their regulatory roles in plants, Annu Rev Plant Biol., 57:19-53 (2006).
Karathanou K, Theofilatos K, Kleftogiannis D, Alexakos C, Likothanassis S, Tsakalidis A, Mavroudi S, NcRNAclass: A web platform for non-coding RNA feature calculation and MicroRNAs and targets prediction, *International Journal on Artificial Intelligence Tools*, *24*(1), 1–17 https://doi.org/10.1142/S0218213015400023 (2015).
Kozomara, A. ve **Griffiths-Jones**, S., miRBase: integrating microRNAannotation and deep-sequencing data, Nucleic Acids Res., 39, D152–D157 (2010).
Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T., Identification of novel genes coding for small expressed RNAs, Science, 294(5543):853-8 (2001).
Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T., New microRNAs from mouse and human, RNA, 9(2):175-9 (2003).
Lau NC, Lim LP, Weinstein EG, Bartel DP., An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans, Science, 294(5543):858-62 (2001).
Lee RC, Feinbaum RL, Ambros V., The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell, 75(5):843-54 (1993).
Lee RC, Ambros V., An extensive class of small RNAs in Caenorhabditis elegans, Science, 294(5543):862-4 (2001).
Li L, Xu J, Yang D, Tan X, Wang H, Computational approaches for microRNA studies: a review, Mamm. Genome, 21(1-2):1-12 (2010).
Liang H, Li WH., Lowly expressed human microRNA genes evolve rapidly, Mol Biol Evol., 26(6):1195-8 (2009).
Lindow M., **Gorodkin** J., Principles and Limitations of Computational MicroRNA Gene and Target Finding, DNA and Cell Biology, 26(5): 339-351 (2007).

Liu H, Yue D, Chen Y, Gao SJ, Huang Y., Improving performance of mammalian microRNA target prediction, BMC Bioinformatics, 11:476 (2010).

Lu Y, **Leslie** CS, Learning to Predict miRNA-mRNA Interactions from AGO CLIP Sequencing and CLASH Data, *PLoS Computational Biology*, *12*(7), 1–18 https://doi.org/10.1371/journal.pcbi.1005026 (2016).

McGinnis S, **Madden** TL., BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res., 32(Web Server issue):W20-5 (2004).

Mendes ND, Freitas AT, Sagot MF, Current tools for the identification of miRNA genes and their targets, Nucleic Acids Res., 37(8):2419-33 (2009).

Mousavi R, Eftekhari M, Haghighi MG, A new approach to human MicroRNA target prediction using ensemble pruning and rotation forest, *Journal of Bioinformatics and Computational Biology*, *13*(6), 1550017 https://doi.org/10.1142/S0219720015500171 (2015).

Ng KL, **Mishra** SK, De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures, Bioinformatics, 23(11):1321-30, (2007).

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G., Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA, Nature, 408(6808):86-9 (2000).

Rabiee-Ghahfarrokhi B, Rafiei F, Niknafs AA, Zamani B, Prediction of microRNA target genes using an efficient genetic algorithm-based decision tree, *FEBS Open Bio*, *5*, 877–884 https://doi.org/10.1016/j.fob.2015.10.003 (2015).

Ritchie W, Gao D, Rasko JE, Defining and providing robust controls for microRNA prediction, Bioinformatics, 28(8):1058-61 (2012).

Saçar MD, Hamzeiy H, ve Allmer J, Can MiRBase Provide Positive Data for Machine Learning for the Detection of MiRNA Hairpins? Journal of Integrative Bioinformatics (2013), doi:10.2390/biecoll-jib-2013-215.

Saçar Demirci MD, Allmer J, Delineating the Impact of Machine Learning Elements in Pre-microRNA Detection, *PeerJ* 5: e3131. doi:10.7717/peerj.3131. (2017)

Saçar Demirci MD, Baumbach J, Allmer J, On the Performance of Pre-microRNA Detection Algorithms, *Nature Communications* 8 (1). Nature Publishing Group: 330. doi:10.1038/s41467-017-00403-z. (2017)

Satoh J, MicroRNAs and their therapeutic potential for human diseases: aberrant microRNA expression in Alzheimer's disease brains, J Pharmacol Sci., 114(3):269-75 (2010).
Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R, RNAshapes: an integrated RNA analysis package based on abstract shapes, Bioinformatics, 22(4):500-3 (2006).
Sun W, Julie Li YS, Huang HD, Shyy JY, Chien S, microRNA: a master regulator of cellular processes for bioengineering systems, Annu Rev Biomed Eng., 12:1-27 (2010).
Tyagi V, **Prasad** CS, RAmiRNA: Software suite for generation of SVMbased prediction models of mature miRNAs, Bioinformation, 8(12):581-5 (2012).
Varghese S, Salim A, Vinod Chandra SS, MicroRNA binding site scoring model, *International Conference on Control Communication & Computing India (ICCC)*, Trivandrum, 2015, pp. 628-633 doi: 10.1109/ICCC.2015.7432972 (2015).
Wang G, van der Walt JM, Mayhew G, Li YJ, Züchner S, Scott WK, Martin ER, Vance JM., Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein, Am J Hum Genet., 82(2):283-9 (2008).
Wang X, Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies, *Bioinformatics*, *32*(9), 1316–1322. https://doi.org/10.1093/bioinformatics/btw002 (2016).
Wu Y, Wei B, Liu H, Li T, Rayner S., MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences, BMC Bioinformatics, 12:107 (2011).
Xue C, Li F, He T, Liu GP, Li Y, Zhang X, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinformatics, 6:310 (2005).
Yousef M, Jung S, Showe LC, Showe MK., Learning from positive examples when the negative class is undetermined--microRNA gene identification, Algorithms Mol Biol., 3:2 (2008).
Yousef M, Allmer J, Khalifa W, Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant microRNA Detection, Journal of Biomedical Science and Engineering, 8, 684-694 doi: 10.4236/jbise.2015.810065 (2015).
Yousef M, Saçar Demirci MD, Khalifa W, Allmer J, Feature Selection Has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants, Advances in Bioinformatics, vol., Article ID 5670851, 6 pages, 2016. doi:10.1155/2016/5670851 (2016).
Yousef M, Allmer J,Khalifa W, Feature Selection for MicroRNA Target Prediction - Comparison of One-Class Feature Selection Methodologies, Proceedings of the 9th International Joint

Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS, (BIOSTEC 2016), ISBN 978-989-758-170-0, pages 216-225. DOI: 10.5220/0005701602160225 (2016).

Yousef M, Allmer J, Khalifa W, Accurate Plant MicroRNA Prediction Can Be Achieved Using Sequence Motif Features, Journal of Intelligent Learning Systems and Applications, **8**, 9-22. doi: 10.4236/jilsa.2016.81002 (2016).

Zuker M, Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31 3406–3415 (2003).

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Doç. Dr. JENS ALLMER
Proje No:	113E326
Proje Başlığı:	Micro-RNA Metabolik Ağ Kontrol Analizi için Veri Ambarı
Proje Türü:	Uluslararası
Proje Süresi:	36
Araştırmacılar:	
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	İZMİR YÜKSEK TEKNOLOJİ ENSTİTÜSÜ FEN FAKÜLTESİ MOLEKÜLER BİYOLOJİ VE GENETİK BÖLÜMÜ
Projenin Başlangıç ve Bitiş Tarihleri:	01/10/2014 - 01/10/2017
Onaylanan Bütçe:	352929.0
Harcanan Bütçe:	293329.0
Öz:	MikroRNA?lar (miRNA) uzunluğu yaklaşık 22 nükleotid olan, tek diziden oluşan ve kodlama özelliği olmayan küçük RNA'lardır ve gen ekspresyonunu transkripsiyon sonrası seviyede hedefleri olan mRNA'ların translasyonel baskılanması ve istikrarsızlaştırılması yoluyla kontrol ederler. Çeşitli türlerde yüzlerce miRNA tespit edilmesine rağmen, miRNA?ların büyük bir miktarı hala bilinmemektedir. Bu nedenle, yeni miRNA genlerinin keşfi, miRNA aracılığıyla düzenlenen transkripsiyon sonrası düzenleme mekanizmalarının anlaşılması için önemli bir adımdır. Konvansiyonel ileri genetik tarama, klonlanmış ürünleri domine eden, yüksek miktarda sentezlenen ve/veya her yerde görülen miRNA?lara karşı yanlı bir yöntemdir. Fakat bu tarz biyolojik yöntemler nadir miRNA?ların saptanmasında etkisiz kalmaktadır. İncelenen doku ve organizmanın içinde bulunduğu gelişimsel dönemlerin farklılıkları gibi sınırlamalar, olası miRNA?ları in silico olarak bulmak için karmaşık bilgisayar programlarının geliştirilmesine yol açmıştır. Ancak bir genomdaki muhtemel miRNA?ları tahmin etme amacıyla oluşturulan bu programlar, tahminlerini deneysel olarak doğrulamak için yeterli güveni garanti edebilecek kadar hassas ya da kesin olmaktan çok uzaklardır. Bu nedenle, bu proje kapsamında miRNA analizinde daha güvenilir sonuçlar elde etmek için yeni ve daha etkili bir araç geliştirdik. Proje kapsamında geliştirdiğimiz yöntem sayesinde artık miRNA?lar organizmaların genom dizilerinden yüksek güven aralıklarında bulunabilmektedir. MiRNA?ların potansiyel hedeflerini tespit edebilmek için kullanılması planlanan algoritmaların yeterli doğruluk seviyesinde olmadığı denemeler sonucu görüldükten sonra, hedef tahminlemesi için psRNATarget gibi özelleştirilebilir tahmin araçlarının kullanımı tercih edilmiştir. Bu araçlar birlikte kullanarak farklı organizmalarda önemli miRNA etkileşimleri bulunmuştur. VANESA?nın verilerini aldığı DAWIS-M.D.?ye, tüm bilinen miRNA?lar ve bunların hedeflerini içeren bir veritabanı entegre edilmiştir. Böylece, düzenleyici yolakların görselleştirilmesi (örn: KEGG, Reactome) ve miRNA etkileşimleriyle zenginleştirilmesi mümkün hale gelmiştir. Ek olarak, tahmini yapılan miRNA?lar ve hedefleri, yerel olarak VANESA?ya eklenebilmektedir. Bu özellik, kızamık yolaklarının daha iyi anlaşılmasına ve ALS için yeni potansiyel ilaç hedeflerinin tanımlanmasına olanak sağlamıştır.
Anahtar Kelimeler:	mikroRNA,miRNA regülasyonu,miRNA gen tahmini, miRNA hedef tahmini,ağ görselleştirme,yolak analizi
Fikri Ürün Bildirim Formu Sunuldu Mu?:	Hayır

Projeden Yapılan Yayınlar:	1- The impact of feature selection on one and two-class classification performance for plant microRNAs (Makale - İndeksli Makale), 2- Development of genomic simple sequence repeat markers in faba bean by next generation sequencing (Makale - İndeksli Makale), 3- Feature Selection has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants (Makale - Diğer Hakemli Makale), 4- One Step Forward, Two Steps Back; Xeno-MicroRNAs Reported in Breast Milk Are Artifacts (Makale - Diğer Hakemli Makale), 5- Feature Selection for MicroRNA Target Prediction (Makale - Diğer Hakemli Makale), 6- Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant microRNA Detection (Makale - Diğer Hakemli Makale), 7- Accurate Plant MicroRNA Prediction Can Be Achieved Using Sequence Motif Features (Makale - Diğer Hakemli Makale), 8- Computational Prediction of MicroRNAs from Toxoplasma gondii Potentially Regulating the Hosts? Gene Expression (Makale - Diğer Hakemli Makale), 9- Integrative Multi-Omics Analysis of Toxoplasma gondii MicroRNAs and their Effects; Powered by KNIME (Bildiri - Uluslararası Bildiri - Poster Sunum), 10- A Machine Learning Approach for MicroRNA Precursor Prediction in Retro-Transcribing Virus Genomes (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 11- VANESA Provides a Platform for the Visualization and Analysis of MicroRNAs within KEGG Pathways (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 12- Next Generation Sequencing?s Sensitivity is both Boon and Bane (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 13- Analysis of Features Describing pre-microRNAs (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 14- How to Avoid Pitfalls in Next Generation Sequencing Data Analysis (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 15- How to Avoid Pitfalls in Next Generation Sequencing Data Analysis (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 16- Cost and Benefit Analysis of Features Used in Machine Learning Based Pre-miRNA Detection (Tez (Araştırmacı Yetiştirilmesi) - Yüksek Lisans Tezi), 17- Characterization of world spinach genetic collection by using molecular markers (Tez (Araştırmacı Yetiştirilmesi) - Yüksek Lisans Tezi), 18- Development of genomic simple sequence repeat markers in faba bean by next-generation sequencing (Makale - İndeksli Makale), 19- MicroRNA categorization using sequence motifs and k-mers (Makale - İndeksli Makale), 20- Categorization of species based on their microRNAs employing sequence motifs, information-theoretic sequence feature extraction, and k-mers (Makale - İndeksli Makale), 21- Delineating the impact of machine learning elements in pre-microRNA detection (Makale - İndeksli Makale), 22- On the performance of pre-microRNA detection algorithms (Makale - İndeksli Makale), 23- COMPUTATIONAL ESTABLISHMENT OF MICRORNA METABOLIC NETWORKS (Tez (Araştırmacı Yetiştirilmesi) - Doktora Tezi),